

9/21/99 FCmtg.

1- NCBI/Celera MOU
Celera data → NCBI DB. (not GenBank)
longman re: our use of data.
Wellcome lawyers have seen it & will not block
it. Morgan sees some value.
Julstra has seen it - but Durbin has seen
it. - & doesn't like it.
Needs to go to GS.

2- ASHG Abstract - emphasize next next events
since June - 1 pg.

3- WI response to pipeline (2)

4- Inosiphila ^{Celera} frontation @ Miami - 6x but
not 10x. A couple of Cortiz ~ 10x
There were not ~~the~~ "mini scaffolds".

5. Mtgs re: Internat'l participants.
to Listsev.

Mouse mtg - FC on 1st or Mon AM.
Agendas. Hsu mtg.

Oct 14- end by 3³⁰ - 8 AM - start.

- Agenda ^{mouse}
- Biol identity
- present PI prog report

1
Linsaven, Jones.

Quinn, James

Katrychke -

5/6/99 FC Mtg -

Yasmin

List of seq. opp - Check Abstract's Issues
UK - none
Japan - full length cDNA.

Server - Call from McPherson -

NCBI - back on track.

Working Ace DB program that implements

Clone Traffick Control -

→ St. Louis then May 13.

⌘

Call Eric Green re: 65 Abstract.

4/14/99

FC mtg



Press events:

- TSC
- Chr 29 Adam is looking into what the status is
calling Don Durham
- Craig @ Council

Preparing for Council session -
instead invite TSC - + ask Craig + Jerry
for Sept.

Mygrand & FC -

wants to play important role.

Conts - 1) Inverte model

2) focus on fishing - Phil Green

- figure out how much needed to keep his critical
mass.

- Build w/ EL to be sure it can be exported

- not \Rightarrow Inverte.

Phil G - turning 3700

Phil is very irritated -

MO will accept a ship without 3700 from Herkopolis

FC e-mailed Herkopolis

DCox - renewal?

ful out DC re: 1 yr ext.

see how long we can put off

Rutin - Ultra - 5 x June 15

5 x July 15

Probably not see data until mid - July

machines (Mar 25) 17 machines running

50 in Bldg.

PCR in Droppids -

Is food expensive w/d.



Software approaches for antifishing?

He can get good coverage @ lower field coverage.

3Kb Plasmid units.

do 5x shotgun - if full path the OK

if not \Rightarrow 7x - if still for funds by

PCR.

Cost funding in $1/2$ 1.5 Mb / row.

more spnt in very light (0.7x) coverage

QC step -

pushing hard to put phys wrap in place

1x cov = 1,000 BACS

@ .7x cov. / BAC

\sim 10,000 reads / whr.



will help the Celera sequence assemble

BAC ends - Wessimbach

9, 656. 7870 paired end cov.

Boyer is doing fingerprinting -

done 8,000 runs, but not in RPC yet

2/25/99

Waterston - call (FC)

Used of left over funds? \$2.5M
let them keep savings from leasing
÷ \$4M from Waterston w/ jobs

Finishing - 8-8 1/2 M needs in 12 Mo.

2.2 M needs on finishing Chr 7 & 9.

12 Mo 550 MB @ 5K

? 100 MB in Gen Bank already

150 MB more to finish

750 MB of total done in some form.

WATERSTON do this w/ 1/2 of comp. cost.

Add - Budget + 000 for sending.

Add Qual items - plans.

Deadline for sending = May 1

BAC - end seg - \$650K - ??

KH - Monday of Nov 8 - press conf. press release.

Page - 0.5K coverage → lots of ex. no.

\$2-2.5M/yr ~~for 2 yrs?~~

600 MB.

NCF interested; also OOR. also Wellcome

2/12/99

Agenda next week

Mapping - of BACS

Council - call ahead of time -

Table in terms of roads + mb.
roads + 70 completion @ 5% of ~~total~~ total given
2 boards.

Call Rick Wathis - MC.

B. Wolf

B. Hritz

J. Lawrence

J. Rodman - get them on time.

A. Wilson

A. Chakravarti

"handwriting as using credible merit in rest in
Council, only by scale & economy." Dennis
w/ delay in decision - can't wait.

Clear Clerk will make much noise abt
covering process by end of yr. + CV already
made that decision in Cognis + W.H.
Exlcity sci opportunity.

"The majority" of funds.



Airlie Center

1/26

Jan W. Kirby -

Bartony mtg - his fr send out invitations
because they don't want it out.

Then they really do know what is going on

Enc 23rd done one of 24th.

Room for 40. -

over 40. Conference room is problem.

In terms of reg. we will ~~invite~~ ^{contact} people, & then they
send out invite.

FC Mtg Summary of funds for reg. for all agencies
pull out pilot project statement.

Table of projects coming in May.

QA info for all by then.

Giffis - $\approx 20M$, + reg.

Lander - 75% of reg.

Waterman - fully funded.

Ask each for what they
would do at diff out
of income
for Sci Staff mtg.

time for PE person.

Ask all mtg on Sunday with - R6



4/2/99

9AM-CV. @ Lipsitt

Get April dates for FC.

RFA - group that takes on BAC counts to finish
- need to demonstr. by end of 1st year that
they can finish.

Already learned most effective way to learn to
is to prod. by final. seg.

Existing cts - should justify why are advised
not to finish or go along.
find interests again.

Keep it vague - report of finding.

Give plan to demonstr ability to finish + pay that prepare finding
w/o un. cost.

Human - mapping evidence.

Mouse - produce STS @ ~100Kb intervals

for seg. ID BAC addresses.

Indiv. BACs clearly ID by seg.

"Internal." Service to help w/ building contig.

Mtg w/ G. Skuler -

- invited to Baylar.

✓ if Ice Hood is doing more fingerprinting.

In mouse, will we need to support binning?

Next FC acts - give rich regions

Trace data for mouse + human so we can finish?

Mouse issues:

- 1 - how will decide who does what. core control? lab based
- 2 - trace data
- 3 - target gene-rich & haplogroups eq.

CS - has started already + hope to form a system of people to look @ by Jan 8.
Will try to have 2nd round by Feb 11.

Ask ST if anyone from Europe (Sanger?) coming.

12/17/98 FC mtg

- Jimmy letter - copyright
- * get Maria, Lambert, Tynan together.
- Letter to James, Patricia, Megan, Freir

- DOE - meeting w/ Mon.

1 - what is next step w/ MOU.

Q: understand concerns, HV talked to M. Krebs.

FC asked Q: to initiate discussion of what would be agreeable to Welton, DOE + NCH.

2 - what are they doing w/ process?

are they going to Princeton mtg? Marv?

- PI mtg - evening night 10th for ribbon

Cutting - FC cannot do 12th

SPIs, Marv, Michael, res + Greg L.

Agenda -

? 'Are members getting copies of all apps. -?

FC invite Sulata. - 4 minutes

Email or call fbb

Pipeline / Issues

for Transport issues

Collab effort?

Sanitized vers. of gc data.

Princeton sop
ok to 10/11 - but
better by 3 or 4 PM
5-7 PM. (LHM)

*
No but will
send fbb

FC Mtg 12/8

- NCR R EN - are we interested?

Leaving to Patrick + we'll see.

- not - are we interested in net

- Dissipider - Should enforce principle that Jerry should not deal w/ seg. info that is not pub available.

Trace files? Not resolved. -

AF tried to get Jerry to see that any data he gets should be released.

Model is: Jerry write a letter to Craig, re: what he will do. P1 - Planned Colloq in May + what they agreed upon. Want to do that plan if he has backed away from commitment principles are:

- 1- no seg files Alex → Patrick w/o pub release
- 2- no gap closure on any seg. not pub avail
- 3- Jerry do BAC - and seg - will do more.
- 4- Craig will agree to no more than 6 mo. from start time to release.
- 5- In meantime Jerry proceed as if Craig not there.

Jimmy needs to write 2 letters - 1 to us; 1 to CV
FC will talk to Morgan + Patrino re: trace data
Make award when data release ltr is OK.

Harold has talked to Annie Levine who is confused.

Turner - MG make changes ^{to summary, to 1000 pgs (12/2)} & send out carefully.

What now? wait 'til after rev. & then get tog.
Wash U will go ahead w/ their mapping.
(will it be part of the app? if not,
we need to be sure to get a Supplement to
fund it.)

Dietz - MG. talked to Schuler. Good to get underway.
Major topic for Jan mts. - geting now.
Have e-mail conv. w/ Greg. -

May Internat'l mts. -
Jan mts @ Houston - Sec. Cnd. - dates & arrangements.

Mouse 17% transfer -

- \$15M w/ \$6M NHGRI

3 new + 3 new exp

5% version of plan.

outside panel.

future years - built in to NHGRI.

- NCBI - \$1.7M stat - covers database

e-mail to Council re: mouse RFA.

EN - anticipatory - and is a trans NIH initiative

Mouse + MRC + Wellcome -

Can say we want to coordinate w/ them.

Dec 2 + 3rd Mtg

Sulston is vague abt when he could come.

asked him to come on 1st - still don't know.

What does it cost to screen a library w/ an
organism. - more sequencing.

Mtg w/ FC - 2 PM. Monday. ✓ w/ Peggy re: 9 PM plane?

Mtg room for 2nd & 3rd. Ask Blag 1 re: hotel.

Meynard?

11/24 FC Mtg

Air breakfast -

MG talk & Max. Clear vs. Data → BOE

Revenue Canada

" "None" - but don't insist

Disruptive -

work w/10 seg that put avail.

Contact advisors!

write clear end + terms of assoc.

~~John~~ John Jay will consult w/us.

with dept Y+C for a mtg

w/ HV, MF, next week.

by Fri e-mail agreement.

Dec 2 mtg - 4-4:30 to 6 or 6³⁰

About minutes? Mon or F.

Agenda.

Start w/ Sanger plan -

what more data do we need in the next 6 mos.

possibility of centralized mapping - later in agenda.

regulators need to come prepared to describe

sig ready maps.

EL as well.

5 PIs.

FC + YNHGRI

MM, MF

8 @ table - rest on sides.
bring dinner in.

DAC - agenda?

FC talking.

Substr 4?

Disc 4th - Mtg w/MM.

Who is down for sponsored library?
ask Pdb.

Talk w/MF re: frequency, BAC - end seq.
Mussio closely track frequency.

J

Sanger >875 Mb
 (finished 32 Mb 6/97-8/98)
 (finished 55.5 Mb to date)
 1 248 Mb
 6 166 Mb
 9 131 Mb
 10 132 Mb
 13 107 Mb
 20 62 Mb
 (parts of 22, X)

Wash U. 555 Mb
 (Finished 25 Mb 3/97-98)
 (finished 39.2 Mb to date)
 2 237 Mb
 7 154 Mb
 14 104 Mb
 Y 60 Mb

Baylor >315 Mb
 (finished 6 Mb 3/97-98)
 (finished 14.9 Mb to date)
 3 192 Mb
 12 123 Mb
 (parts of X)

JGI >344 Mb
 (Finished 20 Mb 10/97-98)
 (finished 23.3 Mb to date)
 5 174 Mb
 16 84 Mb
 19 68 Mb
 (plus parts of 4)

The four "regionally-based centers" have put in claims for 2089 Mb (plus most of X).

Others 812 Mb

Stanford	4	184Mb	(Fin. .8 Mb 3/97-98; 2.8 Mb to date)
Keio	parts of 8	{136 Mb}	(Fin. 1.8 Mb 6/97-8/98; 6 Mb to date)
Jena	rest of 8	{136 Mb}	(Finished 6 Mb to date)
UTSW	11	131 Mb	(Finished 5.3 Mb to date)
	15	100 Mb	
Genoscope	14	{104 Mb}	(Finished 300 kb to date)
Hood	parts of 14	{104 Mb}	(finished 2.6 Mb to date)
Whitehead	17	81 Mb	(Fin. 5.5 Mb 3/97-98; 21 Mb to date)
	18	76 Mb	

chromosomes 21 is not claimed on the HGSI

11/16/98 Mtg w/FC

Character for de Jong & CIT libraries
Letter to Tony @ Wash U.

Rich W needs to let PIS know all this

Dec 3 - Holiday Inn.

Agenda - discussion of survey after dinner
1st - Let John talk. - why so unhappy.

Need to tie together. (do?)

What do we want to see in bars -

what data do we need?

What about other interat'l groups?

Sumada mtg?

maybe next Sumada mtg.



Human Genome Program Operations

459 Donner Laboratory
Lawrence Berkeley Laboratory
Berkeley, CA 94720

Phone: [REDACTED] Fax: [REDACTED]

AGENDA

DOE BIOTECHNOLOGY FORUM
TYSONS CORNER MARRIOTT
8:30 AM-5:00 PM, JUNE 4, 1996

1. *Budget realities*
2. *NIST experience in developing standards for sequencing*
Dennis Reeder and Keith McKenney, NIST
3. *BAC sequencing*
4. *Human subjects*
5. *ELSI review*
6. *High through-put sequencing*

Budget reality - '97 + beyond.

Very serious - agreement emerging for bipartisan support
\$7M increase - to \$81M - language to protect genome.

Krebs says that they may target increases

Expect a \$11M cut + earmarks. - Expect CR for 80% of '96#

Structural bias 2nd priority. Out-years could be as much
as \$100M by 2000. Bipartisan.

NIST - dev of Stds for sequencing -

To: The files

From: Dan Drell

Date: 5/22/96

Re: BAC library human subjects issues

The purpose of this note is to document and, it is hoped, to clarify the events and issues surrounding the development of BAC and PAC libraries from human materials (including sperm) for use by various NCHGR and/or DOE funded genome sequencing projects.

DNA Libraries

Background: DNA libraries can be developed using a variety of vectors, each with different properties, advantages, and disadvantages.

Historically, Yeast Artificial Chromosomes (YACs) came first, largely through the work of Maynard Olson; YACs can contain exogenous DNA (the "insert") from human, mouse, or any other source, up to a length of about 1,000,000 base pairs (a megabase). It has been the experience of many scientists that in the process of making YACs, the source DNA can become fragmented and that more than one fragment can be incorporated into a single YAC (resulting in a DNA insert that is "chimeric", i.e. not representative of the natural order of the DNA in the source genome). Another issue that has arisen with YACs is their stability over time; during propagation in yeast (which treats the YAC as simply another chromosome), the insert can be degraded, lost, or recombined with yeast DNA and its integrity disturbed. Due to the work of Mel Simon (CalTech), Pieter de Jong (LLNL, now at Roswell Park) and others, new vectors (Bacterial Artificial Chromosomes or BACs; P1-derived artificial chromosomes, or PACs) have been developed that display far less chimerism (and are thus much more "true" to the natural organization of the genome) and greater stability. These vectors, while not perfect, are a major improvement over YACs and are currently the preferred vector for working with human DNA in the size range of 100-200 Kb. (While BACs and PACs are distinct entities, the issues are nearly identical and I will speak of BACs herein.)

In current BACs and PACs, the size of the insert is usually in the range of 150,000 to 200,000 base pairs which means that a minimal set of BACs (or PACs) that could "cover" the 3 billion base pair human genome would consist of from 15,000 to 20,000 elements, assuming (ideally) exact, end-to-end matching of consecutive BACs. Due to the fact that the shearing and size selection processes used to get DNA of about 150-200 Kb are statistically random, it is necessary to increase the "depth" of "coverage" to about 20-fold in order to ensure that among the

(now) 300,000 - 400,000 BACs in the "library" is at least one which contains every piece of the original human genome. Statistically, most parts of the human genome will have as many as 10 - 15 representations in the library, but with 20-fold coverage, every bit of genome sequence should be in at least one BAC.

Propagation of BACs is straightforward; E. coli will treat a BAC as a plasmid and all one needs to do to study an individual BAC is grow (as a colony) the E. coli which contain it; the bacteria itself will do all the hard work of reproducing ("cloning") the BAC. The major challenges are twofold: to array the BACs and to sequence the DNA insert. Much of the arraying process is done in order to determine the approximate placement of an individual BAC on the original human genome, i.e. which chromosome, which arm, potentially which band or map position the BAC's insert "comes from." This is important so that the BAC insert can be correlated with other information about the chromosome, segment, etc. of its origin (e.g. is the gene for XYZ, a restriction site, an STS, an EST, or a given marker there?) as well as placed properly with respect to any neighboring BACs so that when actual sequencing is begun, only the minimal set of BACs that contain inserts covering the human genome ONCE, with minimal overlaps at each end, will be sequenced.

There has been discussion recently about using a suitably deep (e.g. 20 fold) BAC library to "end-sequence" the BACs. In this process, each end of each BAC insert is quickly sequenced so that about 500 base pairs are determined. By computational analysis, overlaps of one end with others can be easily determined, based on some 50 base pairs of sequence (more or less) that two BAC ends share. This will, when each sequence is compared against all the others, permit large blocks of original genome sequence to be correlated with the BACs, or, to put this another way, it allows many (if not most) of the BACs to be relatively positioned along the entire human genome so that a "tiling path" of BACs, each positioned and oriented correctly with respect to the original genome, can be determined. Since all the BACs will be positioned in this way, there will be two outcomes. The first is that islands of several Kb will be located, at from 100-200 Kb intervals (often less) across the human genome. The second outcome is that a set of BACs, in correct order and showing minimal sequence overlap, can be identified and used for the next step in sequencing. This next step might involve sequencing one BAC insert and then, with the adjacent BACs on either side previously identified, going straight to them for continued sequencing. Since their relative positions have already been determined, there is no mapping problem. In this way, the entire human genome could be efficiently sequenced with minimal wasted effort resequencing what has already been done. I am deliberately making this a bit simpler than it probably would work out to be; there are concerns about possible genomic regions that may not clone in BACs, as well as areas of repetitive sequence where the problem of correctly assembling the order and

orientation of the inserts would be difficult.

There is one scientific issue that is not entirely resolved. The exact organization of human genomic DNA is not precisely identical from cell to cell within any one human. That is, the chromosomal DNA in a white blood cell is detectably different from the chromosomal DNA in a liver or brain cell. This is a consequence of the biology of white blood cells, vs. liver vs. brain cells. It is not clear yet how general this observation is, i.e. whether there are natural rearrangements in the chromosomal DNA of other cells (there is an indication that brain cell DNA might also have rearrangements, but this awaits further work.) It is clear that the chromosomal DNA in at least one type of cell differs from the germ line chromosomal DNA that is in either sperm or unfertilized eggs. This has resulted in the view that the best source of the human DNA from which BAC libraries can be made is sperm since this is ground state, unmodified, germ line material. Others have already made libraries from cell lines or white blood cells; other than the specific regions associated with immunoglobulin production, it is not known that any other regions of the genome are modified relative to the DNA in germ cells. Thus it is arguable how much impact, if any, using one cell type vs. another (potentially easier to obtain) cell type may have. An additional variation on this theme is that the frequency of structural polymorphism (the rate at which the organization of one person's genome may slightly differ from another person's) is unknown.

The overall advantage of BACs and PACs at this time is their stability, integrity, relative ease of manipulation, and, before too long, depth of coverage which will give high assurance that the entire human DNA sequence is included.

Issues: The principal issue facing us today is the source(s) of the DNA to be used in BAC library construction. Here, some recent history is worth reviewing. The DOE Human Genome Program has, for a number of years, funded Mel Simon at CalTech to explore BACs as a better vector for manipulating human DNA and to begin to develop some libraries, both of human and mouse genetic material. Simon's success has been noted and his human BACs have been used by a variety of labs (including the Skolnick lab at Utah for the final localization of the long-sought BRCA-1 gene) in their gene hunts. The BAC virtues of stability, lack of chimerism, and ease of growth have made them attractive reagents. To date, the Simon lab has made a number of BAC libraries, at least two of them from human DNA. One of these, made in 1995, was developed by Simon's colleague Dr. Hiroaki Shizuyu. This BAC library, currently about 6-fold coverage, was made from the sperm of a living human donor, whose identity is known to Dr. Shizuyu. (Besides the fact that sperm represents germ-line human DNA, it is also easy to store in a normal freezer from the time of collection to the time of processing for DNA isolation, not requiring excessively expensive facilities.) Prior to collection of sperm from this donor, Drs. Shizuyu and Simon explored the

issue of the need for Human Subjects Approvals from CalTech's Institutional Review Board (IRB), and a signed Informed Consent statement from the sperm donor. Simon and Shizuyu were advised that sperm was a "waste tissue" for which IRB approval and informed consents were not necessary and thus none were obtained. Both of these libraries (the sperm library and another BAC library generated from DNA of a cell line obtained from ATCC in Rockville) have been distributed to Research Genetics (which partially supported the libraries' development) for distribution to researchers.

NCHGR recently awarded 6 grants to sequencing groups to begin pilot sequencing efforts on regions of the human genome. Each of these is slated to run for two years before a "re-review" would determine those groups (presumed to be 2 - 3 at most of the original 6) that would be ramped up with additional funding to continue sequencing. (This gives some of the urgency to the issue.) At least one of the groups (TIGR) has indicated that they plan to use BAC materials developed by the Simon lab as their sequencing "substrates." As part of the routine paperwork prior to making an award, NCHGR asked them for assurances that guidelines regarding IRB clearance and signed informed consent statements had been followed. When NCHGR learned that these guidelines, which have the force of law and without which NIH cannot make an award, were not followed, they raised these concerns at their May Council meeting.

Much of the discussion to date has been on how to resolve the situation so that the pilot sequencing projects can go forward with as little interruption or delay as possible, using the current BAC/PAC libraries. Since the amount of sequence to be generated will absolutely define the donor's DNA, "true" anonymity is not possible since (if enough effort, time and money were devoted to it) it would be possible eventually to link the DNA in any given library with an individual. More disturbing is that at least one of the present libraries (one of the de Jong PAC libraries, maybe both) almost certainly comes from an individual associated with the de Jong lab which raises the issue of the suitability of obtaining samples from lab personnel.

In thinking about the questions that arise regarding BAC libraries, it may be useful to keep the following points in mind.

- 1) The human genome is 3 billion base pairs in size and it is hard to envisage that all 3 billion of the first "reference" sequence will come from the same resource. At least 5 BAC libraries are out there already, others are under development, LLNL and LANL have nearly complete cosmid libraries of chromosomes 19 and 16 respectively, and new technologies are likely to make new libraries even easier to generate. (It is currently estimated that it takes about \$200,000 and on the order of 4 months to "make" a BAC library).

- 2) There is absolutely nothing scientifically to be gained from doing the same individual (or library) for each chromosome or even chromosome arm. Normal recombination causes enough heterozygosity over these distances that it is unthinkable that any serious linkage relationships would survive.
- 3) By current sequencing methods, the sequence error rate can be adjusted but, beyond a rate of about 1/1000, the cost of higher accuracy goes up rapidly. The normal human polymorphism rate is about 1/1000. What does this imply for the proper assembly of sequence fragments derived from different libraries versus a single library?
- 4) Science does not happen in a vacuum. DOE is the successor agency to the AEC that, to exaggerate the public perception only slightly, fed radioactive breakfast cereal to mentally handicapped children at Fernald and we need to visibly take the high road, even if it slows down the science a bit.
- 5) Permanent anonymity is unachievable since DNA is the ultimate identifier. Rather, the barriers to protect the anonymity of BAC library donor(s) should be as high as possible.
- 6) There is an operational difference between obtaining a human sample, whether sperm, blood, whatever, for the sole use of the lab soliciting it and obtaining a sample for the explicit purpose of developing a resource that can be distributed to others for other purposes. Informed consent and IRB clearance are absolutely required for both. The ethical (and political) burden is much more serious for tangible materials and products which will be disseminated to others.
- 7) in sum: both politically and scientifically, the first "reference" human sequence ought to be a mosaic of many donors' DNA.

Outstanding questions:

- 1) is a sperm DNA library substantially preferable to libraries developed from other cell sources? Why?
- 2) Is it necessary to have a permanent somatic cell line from the donor of the DNA used for BAC library construction? What opportunities would be lost if an immortalized donor cell line was not made?
- 3) Should lab personnel be excluded as potential cell/sperm/tissue donors? Lab personnel may be more cognizant of the potential uses of a resource, but they also may be more vulnerable to coercion to participate in a project that is ongoing in the lab to which they belong. (Regardless, proper human subjects protection procedures must be followed.)

4) Can truly "informed" consent ever be given when not all the downstream uses for a BAC library are defined yet?

5) Intellectual property issues have not been addressed at all; how should these be dealt with? It is illegal (and unenforceable) to waive ones rights. Thus any statements, signed or otherwise, that "I waive any future rights to intellectual property derived from my cells, tissues, or DNA", no matter how sincere, are without legal force and would in no way preclude a DNA donor from later on aggressively pursuing financial reward for anything discovered from the library of "his/her" DNA.

6) Is "2-way" anonymity an option? (For example, the lab would not know where the sample came from, and the donor would not know that his/her sample was used for the library.)

7) Political issues: Does the HGP want to revisit the "whose genome are you sequencing?" question? This is not a scientific question; there is no science in this issue, only public perception. Is there a potential threat to the HGP from this?

Potential Options:

1a) withdraw ALL support from the current BAC sperm library. Stipulate that DOE will not fund any new projects, nor continue to support any current projects, which utilize any library that cannot document satisfactory IRB approval and satisfactory informed consent. Retroactive informed consent is not an acceptable option and, since there were enough violations of human subjects regulations to "poison" the current libraries for good, their use should be discontinued immediately.

1b) Get Informed Consent "for continued use" from the donor of the sperm used in the sperm BAC library. Understand that this course is not without risk; from the point of view of some critics out there, it may not be enough. However, it should be done anyway and, if carefully worded and acceptable documentation can be obtained, continue the use of these libraries but expand coverage ONLY with new libraries from different donors.

2) In as reasonable a way as possible, e.g. through a letter over the Associate Director's signature to ALL DOE HGP grantees and contractors, clarify that no compromises in adherence to human subjects guidelines will be tolerated.

3) Allocate additional resources to generate new BAC libraries, done in strict compliance with applicable regulations and guidelines. This would have to be monitored more closely than we have been comfortable with in the past. In particular, try to encourage limited BAC library development, e.g. monochromosomal BAC libraries. Make the distribution conditions favorable enough so that the Research Genetics BAC libraries will be seen to be less attractive, e.g. if necessary, provide support to make

enough copies of a new library so that it can be distributed, at cost or for an extremely modest cost, to those who might otherwise want to use the older ones.

4) Put as a condition in ALL future HGP awards for sequencing, both to the labs and to off site projects, that ANY DNA to be sequenced must come from a library meeting tough human subject regulations. This condition must be in writing. At a minimum, these conditions should include signed informed consent statements from any donors, IRB clearance with consent from DOE, and a clear description of all measures designed to preserve confidentiality.

5) Stress that the use of lab personnel as donors is absolutely unacceptable.

Genome ethics panel comes under the microscope at NIH

The US National Institutes of Health (NIH) have launched a review of the Ethical, Social and Legal Issues (ELSI) Working Group of the US Human Genome Project, following a change of leadership, disagreement over the control of its budget and a call from at least one working group member for extra funds. The review will be led by Mark A. Rothstein, professor of law and director of the Health Law and Policy Institute at the University of Houston, Texas, and M. Anne Spence, a geneticist in the department of pediatrics at the University of California, Irvine.

According to Rothstein, the nine-member committee is to concentrate on structural and funding issues, reviewing the working group's relationship with the ELSI program funded by the NIH and the Department of Energy (DOE). It will also consider its future role in the light of the planned creation by the Clinton administration of the National Bioethics Advisory Commission. Rothstein says that the goal is not to develop a "report card" on the ELSI panel, but to suggest how it might operate most effectively. The setting up of the review panel follows the recent resignation of the chair of the working group, Lori Andrews, professor of law at the Chicago-Kent College of Law in Chicago. Since then, at least two other members have threatened to resign unless the board is given more autonomy and greater control over its own budget. But Francis Collins, the director of the National Center for Human Genome Research (NCHGR), points out that the working group's funds come out of the center's own administrative budget, and that these have already been substantially cut by Congress, to enable more money to be spent directly on research. The working group is funded jointly by the NCHGR and the DOE. Its goal is to inform the public about ethical, legal and social issues raised by the genome project, and it is funded separately from the ELSI program, which awards grants for research on ELSI issues. The review committee will look at where financial support for the working group should come from, and how the money should be distributed.

Tensions arose between the working group and the NCHGR earlier this year when the group was told that the center's budget was sufficient to cover only one meeting this year, compared with three to four a year in earlier years (see *Nature* 380, 96; 1996). Some panel members are also unhappy about disagreements with NCHGR staff over matters such as whether to comment on the timing of the marketing of genetic tests. Dorothy Nelkin, professor of sociology and law at New York University,

and a member of the working group's executive committee, says that members feel they are sometimes being used to legitimate the genome project, rather than to explore critically issues relating to its social impact. Collins denies that the working group has been prevented from expressing opinions on controversial topics. But he argues that there has to be "some limits to its autonomy because it is not a free-standing commission." Collins says that the review of its activities will help define the panel's role within the Human Genome Project.

NIH officials say they are particularly upset at claims that moves to commission a set of papers on behavioral genetics had been suppressed because geneticists in the field had been concerned that the panel would put together an attack on their work. "We are not aware of any evidence that such factors played any role whatsoever in the discussions about carrying out this project," says one.

Dan Drell, a biologist in the DOE's Office of Health and Environmental Research who is responsible for the department's liaison with the ELSI program, said an analysis of behavioral genetics could be an important initiative for the working group, and that it would be of major interest to the general population. But he says that any such analysis would require peer review, and an extremely sensitive and scientifically rigorous approach. The general review is due to be completed by the end of this year. Drell points out that such external reviews are standard practice for government bodies. "There is a sense that there is a little bit of a lack of focus right now," he said.

Troy Duster, professor of sociology at the University of California, Berkeley, and director of the Institute for the Study of Social Change, is acting director of the working group. He says that, in addition to an explicit mandate, he would like to see the review consider whether ELSI issues overall deserve far more than the three per cent of funding allocated to the Human Genome Project by DOE, and the five per cent set aside by NCHGR. "The growing gap between diagnostic information and therapeutic capacity is a time-bomb," says Duster. "In this context, the formula for 95 per cent for the mapping and sequencing versus the five per cent for the social consequences seems particularly absurd. What about 50:50?" he suggests.

Sally Lehrman

--
From: M. Simon
re: update on Human Subjects application to the Cal Tech IRB

I wanted to provide a progress report on our current application to the Caltech IRB with regard to the human sperm DNA BAC library. We have met once with the IRB and they have begun to consider our application to extend the library and to discuss the already existing library. A number of initiatives have been taken in this connection:

- 1) The committee is drafting a letter to the Federal oversight committee (OPRR) describing the past and current situation with regard to the human sperm library.
- 2) Mr. Pool of Caltech General Council Office has been in contact with a number of individuals in the Federal Agencies who are knowledgeable with respect to current regulations regarding human subjects (e.g. Susan Rose in the DOE) and is developing language for informed consent. In addition he is developing an outreach program for Caltech to educate investigators with regard to human subjects.
- 3) The IRB is reviewing our proposed informed consent form and the procedures that we are using with regard to confidentiality in the case of our current sperm donor. The IRB will limit its review to the current donor and to the continuing use of the clones generated from this individual. Any other donor arrangement is subject to an entirely separate review.
- 4) We are developing plans and protocols for future library construction that

will include provisions for donor anonymity.

We are staying in contact with a number of people and consulting about these issues as you know they remain quite complex.

I am enclosing some of the provisional documents that we are generating for this process. PLEASE BE AWARE THAT THESE ARE DRAFTS AND INTERMEDIATE DOCUMENTS AND ARE NOT NECESSARILY THE FINAL WORD

I would also like to apologize to those of you who have inadvertently provided me with some specific wording that I thought most appropriate. You may find that I borrowed it from your Email to draft some of this material. Unfortunately I will not be able to be in Washington for the DOE meeting on Tuesday. However I will be available by speakerphone and I would be happy to be

involved in the discussion of this or of other relevant issues electronically

M. Simon

DRAFT DRAFT DRAFT DRAFT DRAFT

Date: May 21, 1996

To: Dr. Charles Plott
and Members of the IRB

From: M. Simon

Re: Human Subject Derived DNA Libraries

Enclosed is a copy of the revised consent form which we have amended in accordance with the suggestions from the IRB. We understand that approval of our protocol to extend the sperm DNA libraries that we generate applies only to this individual donor and that in the future, protocols involving other individuals will require separate approval. In line with the suggestions of the committee, the Biology Division will be developing new procedures to screen grant proposals for human subject related experiments that require IRB approval. We will also make appropriate announcements to the Division to help ensure compliance and we will cooperate with Mr. Pool in an outreach program to educate investigators about the IRB and the requirements of the program.

With regard to this particular sperm donor, I will sign the consent form first. After the donor and Dr. Shizuya comply it will be placed in a locked file in my office. We will ask those members of the laboratory that may be familiar with the donor's identity to assure us that it will remain confidential. Furthermore, payments will be made to the donor directly by Dr. Shizuya who will then be reimbursed from a Caltech special fund so that records of these payments with the name of the recipient are not specific parts of the laboratory protocol.

We want to thank the IRB for their help and consideration and we will keep the IRB informed during all phases of this work.

cc: Earl Freise
Sandy Pool

DRAFT DRAFT DRAFT

Background to the Application to the IRB for Approval of Development of Extensive Genome Libraries from Single Human Donors

The scientific establishment led by the National Institutes of Health and the Department of Energy has undertaken a project designed to result in the knowledge of the complete nucleotide sequence of the human genome. In order for this project to succeed, it is necessary to prepare cloned libraries of human DNA. A cloned library is generated by obtaining human DNA from human cells, either cells grown in tissue culture, human blood cells, or sperm. This DNA is then digested to generate random fragments and these fragments are attached to "vector" DNA. The vector is a small unique sequence of DNA with properties that allow it to act as a guide to the attached human DNA fragment when it is introduced into a bacterial cell. The vector endows the attached human DNA fragment with the ability to replicate in step with bacterial growth. Thus, each bacterium is the repository of a certain segment of human DNA, a clone. The ensemble of bacteria contains fragments that represent the entire human genome and this ensemble is called a genome library. Investigators around the world determine the sequence of each of these fragments. If the library is extensive enough so that the entire human genome is represented and if it includes enough redundancy so that there is slight overlap between the clones that have been sequenced then the sequence can be reassembled so that the position of each nucleotide in the entire array of 3 billion nucleotides that make up the human genome is known.

There are individual differences between human beings that are recorded in their genome sequence. On the average, human beings differ from each other by one nucleotide in 500 and by the distribution of small stretches of sequence at different places in the genome. It was originally thought that the entire human genome sequence would be assembled from DNA libraries made from a number of individuals around the world and thus the final genomic sequence would be a composite sequences of clones taken from multiple sources.

In 1991 our laboratory at the California Institute of Technology invented a method to generate libraries from large fragments of human DNA. This technique has allowed us to produce libraries that provide clones that are very useful for sequencing. They are relatively stable and can be used to accurately determine human DNA sequence. We initially prepared this library by using human cell lines that are grown in tissue culture as a source of DNA. These tissue culture cell lines are anonymous and are available in the public domain. In 1994 a number of scientists raised questions about the source of DNA. They suggested that the cells that we were using could have sustained mutations, deletions, or other aberrations which would still allow them to grow in tissue culture and artificial laboratory conditions but would not represent a totipotential human genome. We proposed therefore to make a library from human sperm, since sperm cells have the potential to give rise to a whole organism. The general notion at that time was that many different libraries would be generated from a variety of different sources and different individuals and when all of these DNA sequence data were arranged, a composite human genome sequence would result.

In 1994 we contracted to receive sperm from a single donor and succeeded in generating a library of approximately 75,000 clones that would provide single source material to the sequencing community. At the time we saw no possible risk or harm that could emerge from the use of this material and we proceeded to obtain sperm samples based on the following reasons and assurances.

1. The samples were delivered to us by the donor, on his own, at suitable intervals and did not involve risk of physical harm. This material essentially represents human excreta and the donor was compensated for his time, effort, and cost of transportation of the material and suffered no financial harm.

2. The donor was informed as to the use of the material and was in a position to obtain ongoing information about the libraries and the uses that were made of the material supplied to us.

3. Consultation with colleagues who used clinical sperm samples led us to believe that accepted practice was to adopt strict confidentiality in dealing with the donor in order to avoid any possibility of risk of embarrassment or of sociological or psychological harm that might result from disclosure of his identity and his association with the contribution of sperm samples. Thus, there is only one person in our laboratory who deals directly with the donor. Furthermore, since confidentiality can be broken, the donor was made aware of that risk. During 1994 and 1995 this material was used to develop a library that contains approximately 75,000 individual clones. In December of 1995 this library was deposited with Research Genetics a company that distributes libraries for use in the genome community. Our intention is to increase the size of this library to approximately 100,000 clones. The library has not yet been extensively distributed though it has been used to search for individual DNA fragments that might correspond to specific genes.

This single human donor library is one of four or five such libraries that are currently extant. None of the existing libraries are sufficiently "deep" to generate a significant fraction of the human genome sequence. In order to assemble large stretches of human genome sequence from a single individual, highly redundant libraries are required. Thus, we are currently contemplating the extension of this library and its expansion to include approximately 400,000 clones. The development of a new extensive library could pose some risk to the initial donor since an extensive library could be used as a general source for large scale sequencing. It is possible that significant fractions of the human genome prototype sequence may be developed from this library. These sequences would then be associated with the library source and thus associated with a single individual. It is thus possible that significant (more than 1%) portions of the sequence of the genome of this individual could be in the public domain. It is further possible if it becomes difficult to assemble sequence from many different sources that single source large libraries will be used by many laboratories to determine very large portions of the human genome sequence. Thus, it is possible that libraries derived from this individual could become a major source of information about the human genome. Sociological and psychological risks then become a conceivable element and therefore we are approaching the IRB for approval and oversight during this process.

DRAFT DRAFT DRAFT

Amendment to Human Genome Library Consent Form

As you know, we have used samples of sperm that you provided to generate DNA from which human genome libraries have been constructed. Thus far, these libraries have been relatively small and they have been used together with libraries from many other sources to study human genes. As we collect more sperm DNA samples and increase the size of the human genome library above 100,000 clones, the possibility arises that the library that we are building with your DNA could be used as a major source for sequencing the human genome. If the library gets large enough and becomes the major source for sequencing the human genome then a large portion of the sequence of your DNA will exist in the public domain. The Institute (Caltech) cannot guarantee that the source of this DNA will remain confidential. Indeed, confidentiality can be breached in a variety of ways and there is a risk that you may be identified as the source of the DNA for the new large library. At present, we do not know the extent of the information about a person that could be read from his DNA sequence, therefore, we cannot know all of the potential harm or risks that this situation might pose for you. It is clear that sequencing of the human genome will lead to enormous increases in our understanding of human biology, however, it is not at all clear that there will be any particular individual benefit to the person whose DNA is the first to be completely sequenced. There are some potential risks: If you could be identified with this sequence, for example, it is possible that the sequence may reveal some disease causing form of a gene that you may not want to know about or that you may not want revealed to an employer or insurance company. Another possibility is that you might assume unwanted celebrity status as the first person to have a major portion of his genome sequenced if your identity is ever revealed to the public. At this time it is not clear whether we will be able to succeed in generating a very large library with your DNA nor is it clear that this will become a primary resource for human genome sequencing and even if that is the case it is certainly possible and likely that most of the human genome sequence will be derived from many different sources and that it will be difficult to identify individual donor regions. Thus, it becomes difficult to precisely define possible sociological or psychological risk. Your contact, Dr. Hiroaki Shizuya, will be available at all times to answer any further questions that you might have about the project (818-395-4154). Any information that is derived from the project will be available to you. Any significant information that might bear upon the possibility of risk will be communicated to you directly through Dr. Hiroaki Shizuya. You will receive a copy of the this entire consent form and IRB application and you are free to contact the chairman of the IRB or Dr. Hiroaki Shizuya during any point in the experiment. Your consent does not take away any legal rights in case of negligence or any other legal fault of anyone who is involved in this study. Furthermore, nothing in this consent form is intended to preempt any applicable federal, state, or local laws regarding informed consent.

I have read and understand this consent form and have reviewed the attached original consent form and I volunteer to participate in this research study. This study has been explained to me by Dr. Hiroaki Shizuya and my questions have been satisfactorily answered.

Participant's Name

Date

Melvin I. Simon, Co-Principal Investigator

Date

Hiroaki Shizuya, Co-Principal Investigator

Date

DOE Biotech Forum
June 6, 1996
Committee Members

Elbert W. Branscomb
Human Genome Center/BBRP
Lawrence Livermore National
Laboratory
P. O. Box 808, L-452
7000 East Avenue
Livermore, CA 94550

[REDACTED]

Charles Cantor
Dir., Center for Advanced
Biotechnology
Boston University
36 Cummington Street
Boston, MA 02215

[REDACTED]

Anthony Carrano
Director, Human Genome Center/BBRP
Lawrence Livermore National
Laboratory
BBRP, L-452
P. O. Box 808
Livermore, CA 94550

[REDACTED]

Chris Fields
National Center for Genome Resources
1800 Old Pecos Trail
Santa Fe, NM 87505

phone: [REDACTED]

David Kingsbury
Genome Data Base
Johns Hopkins University
2024 E. Monument Street

[REDACTED]

Robert K. Moyzis
Director, Human Genome Center
Los Alamos National Laboratory
CHGS - MS M885
Los Alamos, NM 87545

[REDACTED]

Lloyd M. Smith
Department of Chemistry; Analytical
Division
University of Wisconsin-Madison
1101 University Avenue
Madison, WI 53706-1396

[REDACTED]

DOE Biotech Forum
June 6, 1996
Ex Officio

Benjamin J. Barnhart
ER-70 GTN
U.S. Department of Energy
Office of Health and Energy Research
19901 Germantown Road
Germantown, MD 20874-1290

[REDACTED]

Daniel W. Drell
Human Genome Program
OHER, Department of Energy
ER-72 GTN/DOE
19901 Germantown Road
Germantown, MD 20874-1290

[REDACTED]

Marvin Frazier
Office of Health and Environmental
Research
U.S. Department of Energy
ER72, GTN
19901 Germantown Road,
Germantown, MD 20874-1290

[REDACTED]

Gerald Goldstein
Medical Application and Biological
Research,
OHER, U.S. Department of Energy
ER 73 GTN
19901 Germantown Road,
Germantown, MD 20874

[REDACTED]

Roland F. Hirsch
ER-73, MS F240-GTN
U.S. Department of Energy
19901 Germantown Road
Germantown, MD 20874

[REDACTED]

Aristides Patrinos
Associate Director, Health and
Environmental Research
U.S. Department of Energy
ER-70 GTN
19901 Germantown Road
Germantown, MD 20874-1290

[REDACTED]

Melvin I. Simon
Biology Division
California Institute of Technology
147-75
1201 East California Boulevard
Pasadena, CA 91125

[REDACTED]

Hamilton O. Smith
502 PCTB,
Johns Hopkins University School of
Medicine
725 N. Wolfe Street
Baltimore, MD 21205-2185

[REDACTED]

Jay Snoddy
Human Genome Task Group
U.S. Department of Energy
ER72 GTN
19901 Germantown Road,
Germantown, MD 20874-1290

[REDACTED]

DOE Biotech Forum
June 6, 1996
Ex Officio

Sylvia Spengler
Human Genome Program
Lawrence Berkeley National
Laboratory
1 Cyclotron Road
MS Donner 459
Berkeley CA 94720

[REDACTED]
[REDACTED]
[REDACTED]
S [REDACTED]

Marvin Stodolsky
Human Genome Task Group
HELSD, U.S. Department of Energy
ER-72 GTN
19901 Germantown Road
Germantown, MD 20874-1290

[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]

David Thomassen
HELSD, OHER
U. S. Department of Energy
ER-72
19901 Germantown Road,
Germantown, MD 20874-1290

[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]

John C. Wooley
U.S. Department of Energy
ER-1/7B-084
1000 Independence Ave., SW
Washington, DC 20585

[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]

DOE Biotech Forum
June 6, 1996
Invitees

Machi Dilworth
Integrative Plant Biology
National Science Foundation
4201 Wilson Blvd.
Arlington, VA 22230

[REDACTED]

David Galas
Vice President of Research and
Development
Darwin Molecular, Inc.
1631 220th Avenue SE
Bothell, WA 98021

[REDACTED]

Stephen R. Heller
ARS, BARC-West, Bldg. 005, Room 337
United States Department of
Agriculture
10300 Baltimore Blvd.
Beltsville, MD 20705-2350

[REDACTED]

Elke Jordan
Deputy Dir., Center for Human
Genome Rsch.
National Institutes of Health
Bldg. 38A, Rm. 605
38 Library Drive MSC 6050
Bethesda, MD 20892-6050

[REDACTED]

Keith McKenney
Biotechnology Division
NIST
8353 Chemistry Bldg.
Gaithersburg, MD 20878

[REDACTED]

Dennis Reeder
Biotechnology Division
NIST
8353 Chemistry Bldg.
Gaithersburg, MD 20878

[REDACTED]

Lisa Stubbs
Biology Division
Oak Ridge National Laboratory
Bldg 9210, MS 8077
P.O. Box 2009
Oak Ridge TN 37831-8077

[REDACTED] gov or

DOE Human Genome Coordinating Committee
5 April 1996
Livermore, CA
Minutes

Attending: E. Branscomb, C. Cantor, A. Carrano, D. Kingsbury, R. Moyzis, M. Narla, M. Palazzolo, H. Smith, L. Smith, M. Simon, M. Frazier, A. Patrinos, S. Spengler.
(D.Drell and G. Goldstein by telephonenumber)

Patrinos announced the renaming of this committee to the Biotechnology Forum, a name that indicates the broader scope of the OHER interests, including bioremediation and sustainable development.

In considering the budget for basic energy research in FY97, there is a decrease of 5.6% from the FY 96 budget. Part of this is a decrease in construction costs, so the overall effect is that operating funds are approximately flat. OHER has requested an additional genome increment of \$2M for large scale sequencing, added to an increased request of \$6M over FY 96. Budget control imposes on ER a significant decrease in the three years 1998-2000. If this level is sustained, substantial cuts would be necessary and a major question is where the cuts should occur. Patrinos indicated a need to have these cuts applied to the more entropic programs. All this is difficult in the face of the competitions within ER.

Patrinos has begun a personal dialogue with Francis Collins (NCHGR). The genome community is facing some major decisions, and a unified front, with interagency cooperation, is needed. There are specific areas where substantial cooperation is possible and the counsel, help and input of this committee is solicited. Narla asked about the Congressional view of program merging. Patrinos responded that while there is no sound bite explanation, it is always necessary to explain why DOE is doing genome work. ELSI on the hill has some potential show stoppers, key issues that mandate a unified position. For example, the perception of a disconnect with various ELSI components might lead to "pauses" until specific issues are resolved. Carrano noted that HUGO does well on ELSI and intellectual property and has taken Bartha Knoppers' position paper in *Genomics* as a standard. Patrinos indicated that there is a sensitivity within Congress such that international acceptance is not as important as a unified front.

The ELSI working group, as Patrinos described, has viewed itself as independent of the scientific part of the Genome Program and has wanted its own budget control. He indicated that a review of ELSI was now being established. Branscomb agreed that the ELSI group saw itself as the moral watchdog of the

genome Project, a Project that in their eyes was seen as giving credence to *Bell Curve* ideas.

Cantor questioned the specific area of biotechnology that DOE viewed as important, given the very broad range of biotechnology and its lack of a well-defined funding base. Patrinos responded that the relevant areas included targeted health effects for workers, bioremediation, sustainable development and global climate change. Patrinos indicated he believed that DOE had high hopes and a good position with great applications and good synergy. Frazier indicated that there were some positive signs at DOE about these linkings. Kingsbury agreed that OHER could articulate a position and then get linkages to other groups. Simon also indicated that even a conservative Congress recognized the need for involvement in new technologies, and that it was a legitimate role of government to do the basic investment.

Fields raised the issue of a disconnect between the alliance with NCHGR and the broader biotechnology development. He said that the challenge was to get the idea across that the genome program was not merely to find disease genes, but to be a foundation for broadly applicable technologies. The perception was that 3 billion base pairs and all the genes are the goals of the Genome Project, period. Patrinos responded that a coherent strategy with NIH is immediate issue-driven: large scale sequencing, ELSI, data annotation, --issues that are operational and functional. Simon agreed that the issue is coordination of efforts and data unification, with an 18-24 month window to decide how to do things.

H. Smith argued that if DOE were to focus into a "virtual" mega Center, it could easily do mass production. Fields argued that the conclusions from the Bermuda meeting were incorrect or incoherent--99.99% accurate chimeric sequence is incoherent. His alternative is a moderately accurate, rapidly produced framework. Cantor argued for the utility of any sequence data. Patrinos indicated that one legacy from the earlier days was the promise of complete sequence by a given date. That commitment is always restated. Carrano reported that in Bermuda, Watson had said that the sequence should be done 3-5 years ahead of time to appease Congress. Branscomb indicated that even without "us", there would be 200 Mb/yr and that we have to produce, say 25% of the genome, at a reasonable cost. Simon indicated the NIH has accepted this argument and that the question is how to coordinate and setup, then worry about partitioning the effort. He proposed a planning group with the labs to agree on goals, technologies, techniques and substrates and be concerned with generalizing of technologies. Branscomb indicated that although many virtuous things had been done, it was still necessary to have management, organization, and accountability.

Libraries and Library Uses

Simon spoke of the wide acceptance of BAC libraries for sequencing. The issues include (1) Big enough? Deep enough? Adequate coverage? (2) What is the best way to use for sequencing? (3) What is the source?

There is currently a question of informed consent and appropriate use in part of the CalTech library. In some ways, according to Drell (phone link), consent is an endless tunnel, with no easy way to say when to stop. Moyzis said that it was because of the incredible amount of information and wondered if this could be a test case with federal review. Simon responded that privacy was a concern.

H. Smith indicated that with a library of 300,000 clones and an average insert of 130 kb, there would be 15X coverage and, if random, the proportion not present would be very small. If we obtained end sequences from each clone, it would cost, he estimated, about \$10M, a reasonable goal and a resource at the end. An alternative is to use STS/EST content and bin, pick one, sequence, and maybe hit another STS. Palazzolo suggested a pilot project. Carrano agreed and Moyzis indicated that it should be done now to maximize the impact.

Moyzis didn't think that STSs would be dense enough or good enough to do the integration. Fields raised the issue of benefit, not just cost. The end sequencing has a strategic benefit, producing a tiling path. At the same time, it changes the project from being 3 billion to doing only the remaining gaps, estimated to be on the order of 5,000-20,000 bp.

Accuracy Issues

Patrinos indicated that it was time to be more rigorous in addressing QA/QC issues in genome sequencing. One alternative is to have someone, NIST for example, develop a QA/QC approach. This has been used in climate modeling prediction and has paid off, even though there was originally some skepticism. Then there is the issue of someone to serve as the watchdog. Cantor queried who the customer was and Branscomb asked just what was quality and that this was separate from institutional mechanisms to assure quality. Moyzis said that, given the reality of many dollars in a few places, doing QA/QC was a political necessity. He supported having outside people.

Fields distinguished QA mechanisms that look only at output and QA that focuses on process accountability, as demonstrated by documentation and records. Patrinos is looking for something that applies across the system, not just DOE, although he is not necessarily advocating an independent agency. He is looking for research activity into quality and accuracy with cost estimates, all of this

from beginning to end. He indicated that it was possible that such research would show that it would not be feasible.

There are three levels of quality: the quality of the sequence itself, the quality of the sequence with respect to the clone (do they match?) and the genome versus the sequence. The last part is seen as very difficult. Goldstein indicated that QA programs were being misunderstood and will have someone address QA at the next meeting.

Annotation

Kingsbury raised a concern about sequence that went into the databases with only the barest annotation, that it was human and from chromosome....., no ORFs, ESTs, STS, clone ID, library documentation. Branscomb agreed that the critical requirement was to identify the source of the materials. On the other hand, the database needs to be ready and able to accept and extend annotation. Fields responded that annotation re function is not time critical, but sequence coming in has to have the origin at the level of BAC, cosmid and PAC id, identifiable clonal source and an identifier and at least the library source. Moyzis wanted STSs, ESTs, etc. all put in. Branscomb noted that annotate could mean a narrative blah-blah in a text field versus a queryable description with controlled vocabularies in machine searchable forms. He indicated this was unlikely to happen until and unless data comes in by machine, annotated by machine. Kingsbury also asked about the definition of an entry and how it was aggregated--1 kb? 3 kb? 30 kb?

Fields responded that in the "Bermuda Accord (Triangulated)" the major sequencing labs would release "immediately", prior to anything, including annotation. There was also concern about what counted as an entry and what happens to it. GSDB decided to get ride of the "entry" concept. The GSDB annotator in a demonstration version is available. It does not have editing functionality but does sequence features by class. He indicated that the greatest resistance has been to annotating by source. Carrano asked was was the best pointer to use from sequence back to GDB. Kingsbury responded that clone name is most reliable.

It was decided that Simon would report on the effectiveness of Bac-end sequencing technology (BEST) at the next meeting in June. The date currently scheduled, 6 June, will not work for several members.

Session Name: [REDACTED] 1

Page 1

rom whatsnew@aps.org Fri May 31 12:20:05 1996

Received: from [REDACTED] (8.7.5/1.33-960227)

id MAA21048; Fri, 31 May 1996 12:20:02 -0700

Received: by aps.org (8.6.12/1.35)

id TAA20910; Fri, 31 May 1996 19:21:26 GMT

Date: Fri, 31 May 1996 19:21:26 GMT

To: [REDACTED]

What's New for May 31, 1996

Status: RO

WHAT'S NEW by Robert L. Park Friday, 31 May 96 Washington, DC

1. BUDGET RESOLUTION: HOUSE CUTS THREATEN BASIC RESEARCH AT DOE. Based on the House version of the Budget Resolution (WN 17 May 96), the House Appropriations Committee proposes to slash \$1.3B from the FY 97 allocation for Energy and Water--on top of the \$0.5B cut last year. The Office of Management and Budget warns that such a cut translates into big trouble for DOE programs in basic physics and biology at universities. Maybe the Senate can help. Last week, the Senate approved a budget that includes \$5B more for domestic discretionary programs than the House version; House and Senate Budget Committees will meet Wednesday to resolve their differences. Meanwhile, John Myers (R-IN), chair of the House Energy and Water Appropriations Subcommittee, delayed mark up of the FY 97 spending bill in hopes that the House will agree to an increase. Last night, the House rejected an amendment by Martin Sabo (D-MN) to instruct House conferees to agree to the extra \$5B, but House Budget Committee chair John Kasich (R-OH), did not rule out a compromise. The House does not oppose basic research--members just need to be better informed of DOE's role.

TYSONS CORNER
Marriott

Lander - OK

Olson - OK

~~XXXXXXXXXXXXXXXXXXXX~~

de jure sep be can end seg from Colvin
10-15% of BAs are empty in base
Small units.

Venter said he could do for 1¢/bp! They
agree more like 10¢/bp

~~XXXXXXXXXXXXXXXXXXXX~~

Integrate the DOE centers -
HERAC sub committee bless & provide
guidance re: future integration.
Wants proposal from 3 ctrs how
to integrate for virtual ctr. Incl
sig, informatics & tech dev.
Need to bring GDB + GSDB
into this. - 1 mo from award

As you have noted, our data have been deposited with the Genome Database (GDB). However, with respect to GenBank, our informatics specialist (Ron Lundstrom, Ph.D.) was told that they did not want such single pass sequences. Perhaps, there was some misunderstanding on some individual's part along the way. Regardless, the data were deposited in a public database. Furthermore, we will make sure that the sequences are deposited in GenBank.

We may scientifically disagree on the relative merits of further converting these unique markers to PCR amplifiable loci. We have provided rationale supporting obtaining multiple sequences from each clone without the additional cost and effort of developing PCR amplifiable loci, per se. Is there not more to be gained by allowing us to further our large scale sequencing studies (which are going extremely well), by developing sequence ready maps for test regions, by increasing the overall number of well positioned sequenced sites especially that have a high likelihood of representing genes and by continual progress in other aspects of Genomics that will be important to the progress of chromosome 3. However, there is nothing to be gained from any impasse on either side. There is much important work to be done and we wish to press on. Therefore, given the above discussion we ask you to reconsider whether or not the NCHGR wants us spend the time and effort to develop PCR primers to amplify each locus or whether our efforts are best spent on other items mentioned above. If PCR confirmation of the STSs we have generated is the highest priority, please let us know as soon as possible as this will take some time to accomplish, perhaps a few months or so.



Harry A. Drabkin, M.D.
Associate Professor of Medicine

Sincerely,



Robert Gemmill, Ph.D.
Assistant Professor of Medicine

(PS: As I mentioned to you, I will be out of town all of next week and then have MGN study section. Please let us know as soon as possible.)

NATIONAL CENTER FOR HUMAN GENOME RESEARCH, NIH



FAX TRANSMITTAL SHEET

TO: Dr. Jane Peterson

FAX NUMBER: [REDACTED]

FROM: Bettie J. Graham, Ph.D.

DATE: June 13, 1996

of pages including cover sheet: 3

Return FAX [REDACTED]

Comments:

Participant in Molecular Genetics Profiling Conference.

Will pick up fax at front desk.



University of Colorado Health Sciences Center

Dr. Harry A. Drabkin
Biomedical Research Bldg, Rm 502/512
Division of Medical Oncology
University of Colorado Cancer Center

Box B171 phone
4200 E. Ninth Avenue
Denver, CO 80262 FAX

June 12, 1996

Dr. Jane Peterson
Center for Human Genome Research
NIH, Bldg. 38A, Room 610
9000 Rockville Pike
Bethesda, MD 20892

Dear Dr. Peterson,

In response to your letter of today, we wish to make the following points:

1. Our hybridization-based probes used to construct the chromosome 3 YAC contig maps represent unique loci in the genome. These probes and their mapping positions have been previously published (Drabkin et al., *Genomics* 8:435-446, 1990). The clones have been used successfully by many investigators for several years attesting to their validity and uniqueness in the genome. Without this latter feature, it would have been impossible to map them.
2. The clones have also been mapped to unique YACs or overlapping YACs isolated from a total human YAC library. The YAC contig map is also richly integrated with the polymorphic STSs from Genethon which were independently mapped by genetic linkage studies. The concordance of our physical mapping studies and the genetic linkage assignments was very high (Gemmill et al., *An Integrated Yac Contig Map for Human Chromosome 3. Nature* 377: 299-320, 1995).
3. In sum, the probes were mapped to chromosome 3-specific regions using somatic cell hybrids and to YACs identified from a total human library. Along with the additional caveat that the YAC positions based on genetically linked markers and the probe positions are in agreement, there can be little doubt that the probe set represents unique loci in the genome.
4. Multiple sequences are being obtained from each of the larger insert NotI clones. Thus, it is very unlikely that the positions of these hybridization-based probes will not be correctly integrated into the developing sequence of the chromosome. Similarly, our sequences will certainly allow correct positioning of random or otherwise obtained DNA sequence to fixed points on the chromosome. The fact that BLAST searches of our NotI sequences are identifying many genes and other loci previously mapped to chromosome 3, as pointed out in the progress report, attest to the validity of our conclusions. Thus, it is hard to argue that these are not sequence-tagged sites (STSs).
5. To begin with an unknown, unmapped sequence and to develop PCR primers in order to position loci makes good scientific sense. In contrast, it seems less compelling from a scientific standpoint to develop PCR primers for loci that are already well mapped and positioned and fully integrated with other PCR-based markers in the contigs.



THE RITZ-CARLTON

Finally, we have the question of what to do about next year? He was scheduled to make these STSs + continue sequences ^{of a 200 kb region}. We could put another condition on the award, but this is the terminal year so it wouldn't have much effect. We could give him funds ^{on a 1/4 by basis} ~~per year~~ to ensure that he does STSs, but at this point I am not sure it is worth the effort - Do we really need more Chr 3 STSs & are the sequencers going to use them?

I will call Ellie tomorrow to discuss these issues. I will be at home tonight if either of you want to talk to me there, or you can respond to this e-mail & I'll call back.

So what should we do?

I plan to write him a letter (perhaps give him a meeting)
making the ^{same} points as above (slightly expanded) and telling him that our

goal is to make his NOT1 clones as useful as possible by making them

STSS. We could offer a compromise

apparently the 100 ESTs came from

Crafty, & ~~I guess he thought they have~~
been tested ^{because he mapped them by PCR} ~~(I guess they were mapped)~~.

So that leaves 150 STSS from the NOT1

clones. We could require that he test

& make sure he generates successful

STSS from 100 of these. That one well
spread (will have to define well spread -
100 Kb or more?). ^{This 9 clones may be difficult because he is making multiple probes for each gene} & deposit in GDB + GenBank.

He should annotate those that he does
not make into STSS.

6/14/96



- 1- mapped by Lyle to YACS.
- 2- make 100 STSS as well speeded as possible.

Spoke w/ Duabkin -

He was quite angry & wanted money to cover salaries for the work. He said he was tired of being threatened. He asked why this didn't come up in the interim progress report - I told him I said that but was busy & wanted to see what was in the final report. I told him he could write a letter asking for lines of support. He asked about how he gets the first to pay for this. I said he could prepay. He said that he puts together a map of a large classroom on an ROI budget & we treat him like a 7th class citizen. Encrypted I tried to make the print

that he knew the definition of STS
and purposely did not do it, he
went to about how it does it help
to argue about it & come to an
impasse. He finally just got to
the same, going to make 100 STSs
& deposit them.

Up

5/10/99

got to go early June 17-18
week of 20th -
27th.

Supernovas Rev @ Jostbrooky Hrd. Sam
looked at all BAC ends.

Hrd could put Phred 20 gene scans in db.
scans to be primary step.

Analysis by looking @ how reads.

Cal Tech vs. RPC - 11 - 50, 50.

Plan: not to affect what's going on - would put
behind if steps for goal.
for next phase 20 cont off at 150 bp.

Mouse BAC end seq. - will fund some 70 of lines.

Mouse sequencing? - 10 x 10⁶ bp. 5x hi final 0+0.

Next year - a lot of effort in mouse.
Selected regions. - sending the regions 1st.
system w/ human.

Mouse del. mutant for the regions.

JGI - getting 24 MD's right away of 100 by end
of year.

prod. capacity 300-400 M bp. 11-12 M lanes
plan for this yr = 6M lanes.

lib -

Fog - @ Berkeley - Lead wrapper -

name of lib in Gu Bank.

JGI - Paul Richards - R+D.

some new disputes.

Chris Morten's job → Ken's pen

prod. hit a lag.

SUMMARY - DOE REVIEW 99-04 HUMAN GENOME PROGRAM ACTIONS

Sequencing technology:

Ger van de Engh, UWash - Replacement of front end culture preparation for sequencing, by flow sorting single "PCR ready" bacteria into microwells.

Stanley Tabor, Harvard Med. - Renewal for continuing improvement of polymerase complexes for DNA sequencing

cDNA sector:

M. Bento Soares - Renewal to improve cDNA libraries including selective cloning of rarer species.

Richard Gibbs - Following a successful pilot on full length cDNA sequencing, his cDNA concatenation strategy will be used to sequence cDNAs from a high quality brain cDNA library provided by collaborator **Claudio Snyder**.

Michael Altherr with LANL colleagues - Importation of CAP trapping technology for selection of full length cDNAs initially, to be followed by full length cDNA sequencing using the Gibbs concatenation strategy and/or L. Ulanovsky's primer walking strategy.

Difficult chromosomal regions: *Shive*

Robert Moyzis, UC Davis - Sequencing of near telomeric regions of chromosomes 16, 19 and 5.

Ann Olson (LLNL) and Evan Eichler (Case Western) - Contiging of the centromeric region of chromosome 19, wherein genes are interspersed with centromeric repeats.

BAC library technology:

Hiroaki Shizuya, CalTech

Pieter de Jong, Roswell Park

Six months support to demonstrate BAC construction with sheared DNAs, with probable follow on support for genome scale library constructions for mouse and man.

BAC Sequencing Tag Connector (STC) technologies:

Claire Frazer team, TIGR

Leroy Hood & Greg Mahairas, UWash.

1) Following completion of STC acquisition for the human genome in July, there will be support for technology development to further lower costs and improve quality. At the request of NIH, support for mouse or rat STC generation could be rapidly initiated through DOE channels.

2) Representatives of the major genome sequencing teams, UWash, TIGR and NIH will be invited to participate in a July/August, 1999 Press Conference marking the completion a major milestone toward Human Genome Sequencing, the completion of STC acquisition providing an average of one STC every 3 kb across the entire human genome.

#2 M

15 x 10⁶ bp

Mouse & human

#1

Can extract costs too hi

A cell sorter with tape conveyer system for the generation of sequencing samples
PI: Ger van den Engh, Department of Molecular Biotechnology, U. of Washington, Seattle.

Summary

The next generation of instruments for DNA sequencing will have a greatly increased throughput capacity and will soon render the preparation of DNA samples as the major rate-limiting phase of genome sequencing. We propose to streamline the process that feeds the sequencing machines in order to resolve this looming bottleneck of the human genome project. We have constructed a high-copy plasmid that expresses Green Fluorescent Protein after integration of a DNA insert. This vector makes it possible to select bacteria with DNA inserts by fluorescent cell sorting. The use of a cell sorter as the "clone-picking tool" integrates clone selection, preparation of sequencing templates, and generation of dye-terminated sequencing ladders into one streamlined process. We have demonstrated feasibility of several key steps. In addition to developing a suitable GFP-expressing vector, we have demonstrated that insert amplification and template preparation can be done with single sorted bacteria as a starting material. We now want to exploit these principles in building an assembly-line for conducting sequencing reactions on several tens-of-thousands of templates per day. Such a facility will be established, evaluated, and made ready for routine operation within the next 24 months. The proposed process is scalable. An increase in the size of the operation will yield additional economies of scale. We will put the new process to the test in the large-scale sequencing facility of the Department of Molecular Biotechnology (Hood/Mahairas).

The project will be executed in two phases. We will optimize clone selection and amplification methods using multi-well trays as the sample carrier. After suitable procedures have been established, we will explore the use of linear tapes with sealed pockets. Such tapes, containing thousands of pockets, are standard in the electronics industry, which has established an economic, automated technology for handling these tapes. The conveyer tapes are particularly advantageous for interfacing to (capillary) sequencing instruments.

We expect that the process integration combined with an efficient conveyer format will greatly increase the throughput rate of the clone-preparation phase. We envision that our template preparation and sequencing process will provide a significant boost in overall throughput capacity. After optimization, the method should allow a team of 3 or 4 technicians to prepare a hundred thousand or more sequencing samples per day.

Stanley Tabor

Characterization and Modification of DNA Polymerases and Their Accessory Proteins for Use in DNA Sequencing and Amplification

The goal of this project is to continue to develop DNA polymerases for use in DNA sequencing and amplification applications. The focus of our research is to understand the basic mechanisms of DNA replication, using the enzymes from bacteriophage T7 as a paradigm. The knowledge gained from these studies will likely lead to modified enzymes that will be useful for the Human Genome Project. Related work achieved during previous granting periods led to the commercial development of the sequencing enzymes Sequenase, ThermoSequenase, AmpliTaq FS and Omnibase.

In the past granting period we determined the crystal structure of T7 DNA polymerase in a complex with a primer-template, an incoming dNTP, and the processivity factor thioredoxin. This structure illustrates how nucleotide substrates are selected in a template-directed manner and it provides a structural basis for the metal-catalyzed mechanism by which polymerization occurs. Knowledge of the specific contacts between the polymerase and the primer-template, dNTP and the two metal ions critical for catalysis will allow for the rationale design of new mutations in T7 DNA polymerase and the homologous Taq DNA polymerase that have altered specificities for nucleotide analogs. For example, the structure identifies two specific residues that make critical contacts with the C2' and C3' positions of the ribose moiety of the incoming dNTP. Mutation of one of these residues alters the ability of these polymerases to incorporate chain terminating dideoxynucleotides by several orders of magnitude. Based on this structure, we have constructed over one hundred specific mutations in T7 and Taq DNA polymerases. We propose to continue to characterize these mutant enzymes to modify the active sites of these DNA polymerases to incorporate more efficiently nucleotide analogs modified in the ribose, base and triphosphate moieties. We plan to complement these enzymatic studies with a continued collaboration with the structural biology laboratory headed by Dr. Thomas Ellenberger to determine the structure of other polymerase-DNA complexes using both mutant polymerases and modified DNAs.

In the past granting period we developed an extremely efficient isothermal amplification system based on the T7 replication proteins. This system is capable of producing 15 μ g of product DNA from one picogram of input DNA (i.e., a 15 million fold amplification) in a 15 min reaction at 37 °C. The reaction does not require any added primers, and is nonspecific for the DNA template; all plasmid and BAC DNAs amplify equally well. We believe that this system will be an attractive alternative to current methods used for the automated preparation of plasmid and BAC DNA templates for DNA sequencing. We propose to further develop this system by (1) optimizing the overproduction and purification of the required enzymes, taking particular care that they are free of contaminating DNA, (2) optimizing the conditions for lysis and preferential release of plasmid and BAC DNA from E. coli cells so that DNA templates can be prepared directly from a very small number of cells and (3) collaborate with genome centers to test the use of this system in large scale DNA sequencing projects. We also propose to investigate the use of this amplification system for other purposes such as an *in vitro* alternative for subcloning DNA fragments and for the general amplification of old or rare genomic DNA samples.

One property of DNA polymerases that we have spent much time investigating is pyrophosphorolysis, the reversal of the polymerase reaction. If not prevented, this reaction can cause significant variability in band intensities on DNA sequencing. Since this reaction requires a fully base-paired primer-template, we have been investigating its use both to locate and to select for heteroduplex DNAs that have mismatches. We propose to continue to investigate this system as a means of identifying single nucleotide polymorphisms (SNPs) in genomic DNA, and to construct genomic libraries which consist exclusively of fragments that contain one or more SNPs.

PI name: Marcelo Bento Soares, Ph.D.

Institutional affiliation: University of Iowa

Title: Technology Development for Gene Discovery and Full-length sequencing Any URLs beneficial to

One of the goals of the Genome Project is to identify and determine the full-length sequence of all human and mouse mRNAs. Although much progress has been made towards the goal of identifying all human and mouse genes, we have just begun to tackle the problems associated with cloning and sequencing of full-length cDNAs. This proposal has three primary objectives. First, to develop novel strategies to facilitate the cloning of rare mRNAs, likely not to be represented in the existing libraries being used for gene discovery in the large-scale EST programs. This will be done in an attempt to expedite completion of the ongoing gene discovery efforts. Second, to generate an arrayed set of 10,000 full-length cDNA clones, to be made available for full-length sequencing programs. This collection of full-length clones will be identified by a two-step procedure involving end-sequencing and PCR-screening of a collection of libraries enriched for full-length cDNAs currently being constructed in my laboratory. Third, to construct libraries enriched for full-length cDNAs in a new vector developed by Dr. John Dunn and to conduct a pilot full-length sequencing project for comparative assessment of alternative approaches for construction of shotgun libraries for full-length sequencing, and for quality assessment of the collection of full-length cDNA clones generated in this project.

Towards a Complete Set of Full Length Virtual Human cDNA Clones

Richard A. Gibbs

A complete set of full length human cDNA clone sequences is one aim of the Human Genome Project. Ideally, the data would be generated from a comprehensive set of full length clones that could also be used for gene expression studies, but such clones have not been forthcoming. While efforts are underway to improve cloning techniques, assemblies of partial data have been effectively used to build longer cDNA sequences. This "clustering" approach is hampered by the low quality of EST data, the large bias in EST abundance, and the preponderance of 3'-EST sequences in public databases. We have found that both high quality, finished sequence and partial or "draft" sequence from complete inserts of non-full length clones can greatly improve clustering. In the first year of this proposal we therefore aim to generate 250,000 DNA sequence reads from approximately 10.0 Mb of unique cDNA inserts in order to complete the sequence of approximately 12 Mb of full length clone clusters. The 10.0 Mb of templates will be analyzed by Concatenation cDNA Sequencing, that will generate a mixture of finished and draft sequence. The CCS data will be melded to current clusters, and high quality EST traces from the public resources will be identified to speed the editing and finishing process. The emerging human genomic sequence will also be used for this purpose. The resulting "virtual" cDNAs will be annotated with features including polymorphisms and splice variants that are revealed because of the heterogeneous clone sources. We estimate that with the concurrent human genomic sequencing, this method can yield a virtual sequence of more than 90% of all human expressed sequences within two years, and a cost of <5 cents/virtual base.

Michael Altherr
Los Alamos National Laboratory

Full Length cDNA Sequencing

The work proposed in this application is meant to address sub topic 3 described in that call: Protocols and Reagents for full-length messenger RNA to cDNA production and sequencing. Toward this end, the application has been divided into two distinct subsections. First, the construction of new cDNA libraries, that more faithfully represent the protein encoding segments of genes (ORFs), is described. In collaboration with Dr. Jerry Pelletier at Mc Gill University through a subcontract being conducted under a LANL LDRD project the applicant is constructing cDNA libraries enriched in 5 prime sequences as described below. Strategies for the construction of "full length" cDNAs using both the 5' Capture technique and size selection are described. In addition, we will explore the possibility of using 5' Capture to generate 5 prime enriched cDNA to complete ORF sequencing in areas that have undergone extensive genomic sequencing, particularly those genomic regions targeted by the DOE's JGI (ie. chromosomes 5, 16 and 19). These completed segments of genomic sequence have already undergone extensive high quality sequencing and have been annotated with numerous listings to ESTs generated by oligo-dT or random priming. The depth of sequence coverage and associated annotation should result in a considerable reduction in the effort required, in conjunction with 5 prime sequence, to complete the identification of the ORFs encoded by these regions. The second major subsection of this project deals with strategies for full-length cDNA sequencing. The applicant's laboratory has gained considerable experience in building sequencing libraries generated for concatenation cDNA sequencing. This process will continue to serve as a major method to generate full-length cDNA sequence in the laboratory. However, the recent description of the complete sequencing of several cDNAs by DENS has resulted in a collaborative effort between the applicant and Dr. Levy Ulanovsky at the Argonne National Laboratory. In this application, we propose to will evaluate both methods relative to one another.

Human Telomere Mapping and Sequencing.

Robert. K. Moyzis

The Human Genome Project has undergone a dramatic shift this past year to the goal of obtaining a "working draft" sequence of human DNA in just a few years. Such a framework sequence will catalyze gene discovery and functional analysis, and allow finished sequencing to be focused on regions of the highest biomedical priority. While perhaps 80% of human DNA can be rapidly sequenced in the next few years by highly automated, high throughput sequencing centers, a significant fraction of the human genome will not, we believe, be sequenced to completion by such approaches. These are regions that contain: 1) a high percentage of repetitive DNA sequences; 2) internal tandem duplications, including multigene families; and/or 3) are unstable in all current sequencing vectors. This would be irrelevant if such regions were rare, or contained little of intrinsic informational value. Such is not the case. The mapping phase of the Human Genome Project has clearly indicated that such regions represent a significant fraction of human DNA (perhaps as high as 20%). This includes such critical regions as centromeres and telomeres, as well as a greater abundance of low-copy repeats and multigene families than previously anticipated. Producing quality DNA sequence of these regions, which faithfully represents genomic DNA, will be a continuing challenge.

We propose that a focused, yet distributed, "boutique" approach to sequencing such regions is warranted, where individual laboratories specialize in genomic regions they have special expertise in investigating. Such efforts would complement and integrate with the few truly large-scale sequencing centers that are emerging, such as the DOE Joint Genome Institute sequencing center. Further, such "boutique" efforts will clearly multiply over the next few years, as first-pass draft sequence becomes publicly available. One such "boutique" market is telomeric regions, which exhibit both high levels of repetitive DNA composition and cloning instability. Indeed, great heterogeneity exists in these regions between various individuals. Following the discovery of the human telomere 11 years ago, numerous investigations have implicated genes near telomeres as likely targets for alterations during aging and cancer progression. Through the efforts of my laboratory and those of my collaborator, Dr. Harold Riethman, nearly all human telomeres have now been cloned by functional complementation in yeast. My laboratory has finished three telomere sequences (7q, 9q, 11q), the first RARE cleavage confirmed telomere regions to be sequenced directly up to the terminal (TTAGGG)_n repeat. Greater than 4 Mb of confirmed telomeres are currently available for sequencing. We propose to conduct framework sample sequencing (SASE) on 20 confirmed human telomeres in the next three years, as well as produce finished sequence of the telomeres of chromosomes 5, 16 and 19. In the process we will "cap" the sequence of these chromosomes and identify numerous important genes. An important QC/QA aspect of this proposal is that all sequences will be extensively confirmed against genomic DNA by PCR-sequencing. Polymorphisms in these regions, including SNPs, VNTRs and large-scale deletions will be efficiently determined by pooled DNA PCR/sequencing.

Evan Eichler, Ph.D.
Assistant Professor
Case Western Reserve University
Department of Genetics, BRB 720
Cleveland, OH 44106

[REDACTED]

Anne Olsen, Ph.D.
Staff Scientist
Human Genome Center, L-452
Lawrence Livermore National Laboratory
Livermore, CA 94550

URLs:

<http://www-bio.llnl.gov/bbrp/genome/genome.html>
<http://www.genome.cwru/eichler/P12/>

Sequence-Ready Characterization of the Pericentromeric Region of Chromosome 19

Current mapping and sequencing strategies have been inadequate within the proximal portion of 19p12 due, in part, to the presence of a recently expanded ZNF (zinc-finger) gene family and the presence of large (25-50 kb) inverted beta-satellite repeat structures which bracket this tandemly duplicated gene family. The virtual absence of classically defined "unique" sequence within the region has hampered efforts to identify and characterize a suitable minimal tiling path of clones which can be used as templates required for finished sequencing of the region. The goal of this proposal is to develop and implement a novel sequence-anchor strategy to generate a contiguous BAC map of the most proximal portion of chromosome 19p12 for the purpose of complete sequence characterization. The target region will be an estimated 4.0 Mb of DNA extending from STS marker D19S450 (the beginning of the ZNF gene cluster) to the centromeric (alpha-satellite) junction of 19p11. The approach will entail 1) pre-selection of 19p12 BAC and cosmid clones utilizing both 19p12 -unique and 19p12-SPECIFIC repeat probes 2) the generation of a BAC/cosmid end-sequence map across the region with a density of one marker every 4kb; 3) the development of a second-generation of STS (sequence tagged sites) which will be used to identify and verify clonal overlap at the level of the sequence; 4) incorporation of these sequence-anchored overlapping clones into existing cosmid/BAC restriction maps developed at Livermore National Laboratory; and 5) validation of the organization of this region utilizing high-resolution FISH techniques (extended chromatin analysis) on monochromosomal 19 somatic cell hybrids and parental cell lines of source material. The data generated will be used in the selection of the most parsimonious tiling path of BAC clones to be sequenced as part of the JGI effort on chromosome 19 and should serve as a model for the sequence characterization of other difficult regions of the human genome.

Hiroaki Shizuya

California Institute of Technology

Construction and Characterization of Human and Mouse BAC Libraries from Sheared DNA

We have developed a bacterial F-factor based cloning system (BAC) for cloning large complex DNA fragments of mammalian origin in *Escherichia coli*. BACs can maintain human DNA ranging in size from 80 kb to 350 kb with a high degree of stability. Over the years, extensive work with human BAC libraries constructed by us and others has established their usefulness for physical and genetic mapping, gene discovery, and large scale sequencing.

We have thus far prepared four human and one mouse BAC libraries, and deposited them with Research Genetics for general distribution. The most recent human library (library D) is compliant with the NIH/DOE Guidance to protect donor privacy and confidentiality. The library has been extensively used for mapping and sequencing, including BAC-end sequencing.

In this proposal we plan to construct one human BAC library from sheared sperm DNA. The library is expected to represent human genome much better than the libraries based on partial digestion by restriction enzymes. The DNA sources for this library have been obtained from 20 individuals who have been informed fully about the nature of the experiment, and signed the consent forms conformed with Caltech IRB and the NIH-DOE Guidance. Only one sample from these anonymous individuals will be used for the library construction.

In addition, we plan to construct one mouse BAC library. The library will be made initially using partially digested ES DNA (C57Bl/6J) by HindIII or EcoRI, and after the technique for the construction of the human BAC library from sheared DNA is well established, we will then use sheared mouse DNA to complete the library construction.

A number of new methods will be developed for preparation of the sheared DNA and vector DNA to accomplish blunt-end and T/A ligation. We plan to use ATP dependent nuclease (RecBC nuclease) in addition to more traditional enzymes such as T4 DNA polymerase and mung bean nuclease to generate properly terminated ends of the sheared DNA. Furthermore, in order to increase the ligation efficiency we plan to develop a ligation method based on *Vaccinia* topoisomerase.

A series of quality control tests will be done to maintain the average size of inserts within the specified range, and minimize the "empty" clones. To examine the representation of the new libraries, we plan to screen them with probes specific to the centromeres and telomeres. Furthermore, in order to examine the coverage and representation of the general structure of the human and mouse genome, we plan to map 15 mb of human and mouse chromosomes using a variety of probes and a newly developed multiplexed fluorescence fingerprinting method. The libraries will be distributed through Genome Systems for distribution to the general research community.

Bacterial Artificial Chromosome Libraries for the Human and Mouse Genomes Using Sheared DNA as a Cloning Source

Pieter de Jong

We have previously constructed BAC vectors, improved BAC cloning procedures and have generated BAC libraries, which now serve as templates for large-scale sequencing of the human and mouse genomes (<http://bacpac.med.buffalo.edu>). Although the current BAC libraries appear to represent most of the sequences in the human and mouse genomes, there nevertheless exist uncloneable and unstable sequences and sequences under-represented due to cloning bias. Cloning bias depends in part on the choice of the restriction enzyme used for partial digestion of the genomic DNA. To reduce this likely source of representation bias, we propose to prepare BAC libraries from DNA sheared to the desirable size range or reduced in size by any other way excluding restriction enzyme treatment. Hydrodynamic shearing, very low levels of sonication, or low levels of DNaseI will be used to create fragments in the desirable size range. Prior to attempted cloning, the fragment ends will be polished using Klenow polymerase. We will initially explore cloning of blunt-ended fragments directly into a blunt-end cloning site in our BAC vector, pTARBAC1. In addition, we propose to ligate partially double-stranded adapters to the blunt-ended DNA to create overhangs compatible with EcoRI or BamHI cohesive ends, or to create 12-base overhangs. The adapter-modified genomic DNA will be ligated to compatible ends at the vector and then transformed into electro-competent *E.coli* DH10B cells. Once we have established conditions for cloning blunt-ended DNA into the BAC vector, we will prepare a BAC library for *Drosophila melanogaster*. Only 120 Mbp of the 150 Mbp *Drosophila melanogaster* genome is present in the 17-fold redundant BAC library (http://bacpac.med.buffalo.edu/drosophila_bac.htm) previously prepared in our laboratory. The missing or under-represented regions mainly belong to the heterochromatin. We will use the lower (than human) complexity of the *Drosophila* to explore the advantage of our new cloning approach. Once robust conditions for creating BACs with at least 100 kb average inserts have been established, we will prepare ten-fold redundant BAC libraries for the human genome and the mouse (C57BL/6J) genome.

End Sequencing and Fingerprinting of Human and Mouse BAC Libraries
Fraser, Claire M.
INSTITUTE FOR GENOMIC RESEARCH

The Human Genome Project's new 5-year goals have incorporated the generation of a "working draft" of the human genome by 2001 and the completion of a highly accurate reference sequence by 2003. High throughput sequencing is now the major focus of the human genome effort but needs for supporting resources and technologies remain in several areas. As the human genome project shifts into the large-scale sequencing phase, one of the overwhelming technical challenges is development of an efficient method for producing minimum tiling paths of sequence-ready clones across the entire genome. Libraries constructed in Bacterial artificial chromosomes (BACs) vectors have become the choice for high throughput genomic sequencing projects because of their higher stability as compared to their YAC or cosmid counterparts. A whole-genome approach has been proposed to use BAC end sequences in genome sequencing, in which the complete sequence of a seed BAC is searched against a BAC end database to select the minimally overlapping clones in each direction. This map-as-you-go strategy saves substantial time and effort in constructing sequence ready maps, particularly the process of contig 'walking'. Funded by Department of Energy, we have been end sequencing BAC clones from BAC libraries developed in Dr. Mel Simon's laboratory at CalTech and in Dr. Pieter de Jong's laboratory at the Roswell Park Cancer Institute. These libraries are currently being used for high-throughput human production sequencing. To date, we have generated 203,605 sequences, 92,671,900 bases from 118,446 clones. Together with University of Washington, we will soon reach a total of 600,000 sequences from 300,000 BAC clones, which represents 15X clone coverage and 10% base coverage, providing one sequence marker every 5kb across the genome. BAC end sequences are available from GenBank/dbGSS and through our website. We have conducted comprehensive quality assessments and sequence analyses on BAC ends from both TIGR and University of Washington. By all measures of quality: sequencing accuracy, read length adjusted for quality, quality values of sequences, and cost per base, the TIGR end sequences are superior. The fact that twice as many TIGR BAC end sequences match the finished sequences indicates that TIGR BAC ends are twice as useful in building minimum tiling paths of sequence-ready clones across the genome. As the human project shifts into production sequencing, a plan is developing for sequencing the mouse genome. The current plan employs a "sequence first, map second" strategy which, like the human project, is based on shotgun sequencing of BAC clones comprising contiguous regions of the genome. The strain of mouse to be sequenced (C57BL6/J) has been selected and solicitations for mouse proposals have been announced. We propose to continue end sequencing of existing and/or new BAC libraries constructed to support human sequencing as well as to initiate BAC end sequencing from the mouse BAC libraries constructed to support mouse sequencing. In collaboration with the Clemson University Genomics Institute, we will develop restriction fingerprints of the end sequenced BACs. The clones, the sequences, and the fingerprints will be an available resource for those sequencing the mouse and human genomes, and the community at large. We will continue our focus on quality in this BAC clone end sequencing and restriction fingerprinting project while developing and implementing automation and new methodologies for reducing costs and increasing throughput.

Relevant URLs:

http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_intro.html
http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_search.html
http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_search.htm#clone
http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_anno.html

Construction of a Genome-Wide, Highly Characterized Clone Resource for Mouse Genome Sequencing

Leroy E. Hood

We propose to create a random and dense sequence map with 500 base pair (bp) tags scattered, on average, every kb across the murine genome by sequencing the ends of bacterial artificial chromosome (BAC) clones. This sequence map allows: (1) a minimum overlapping tiling path of BAC clones to be identified from any point in the genome; (2) a physical map to be created from BAC-end sequences matching mapped chromosomal markers (e.g. expressed sequence tags [ESTs], cDNAs, sequence tagged sites [STSs], Unigene clusters, etc); and (3) potential genetic markers to be identified from BAC-end sequences containing simple sequence repeats. Over the past year, the University of Washington High-Throughput Sequencing Center (HTSC) has developed a high-throughput BAC-end sequencing process that includes BAC DNA purification, BAC-end sequencing, BAC clone restriction fingerprinting, and the ability to displace these data on the world wide web through dbGSS and our web site (orcas.htsc.washington.edu). We propose to initially characterize the C57 Bl6 BAC library PPCI-23 produced by Pieter de Jong (200 kb average insert size). Thus, we propose to create a sequence tagged connector (STC) or BAC-end sequence resource from C57 Bl6 DNA containing: (1) 300,000 arrayed BAC clones, (2) 600,000 STCs, and (3) 300,000 restriction digest fingerprints.



file
DOE notes

~~Budget Est for E. Green.~~

My term RFA.

2/5/99 mts w/ Newv. F

QE of BAC end seg - don't know yet.

D libr org is finished

Send website address to Newv.

QA disc @ Baylor - Adam tell them RFP schedule.

~~☞~~ e-mail Ori re: phone hookup.

Juvor locate out of LANSL (paid by)

@ JCF 90% of time

Return cutting part of info.

Next RFA - will put me dates.

DOE pass on West Coast SV (Hond for sure)

Galexo + Lytle have approached DOE re: Collab.

11/2-3/98 - next date

10/02/98

OMB Briefing yesterday
Budget done \$50M
pres 7 \$2.5M

PSF dead 4/12 Cert Metz 12-16
Leyman Res - of PSF.
Leyman is a Scientist -
Nov 17-20.

PSF is running in Nov.
19M+ LLNL + Berkeley 9M each 3M/bp/yr.
LANL 3M.

next yr. 30M guided } hope to do
40M draft. } better.
35bp → for Benb \$10/line.
priority to drive costs ↓.

Cost = 60¢/bp.

JGI Adv. Board Nov. 3 - Bay Area.
James Scott Deputy Dir. Candidate

5-yr Plan + JGI press release 10/23.
in Bay Area.

Int Rev 7-9 Dec.
ELSI 10 Dec

Bevac 5-6 Nov. need be disc of genome.
Can peopley Union.

Hand no. - vent well. -

Quality? not very good. - length $\frac{1}{4}$ phed 20
Acres.

Is this best you can do?

Now collecting TIER + Hand 7gms -

+ these QA, group + compare.

The data is useful but need to compare.

Probably estimate for 400,000 end now
(200,000) by end 250,000 + fragmented.
fragmenting - no - thought it was good.
data re website.

Mouse - Pieter making 20x BAC 129.

Mouse strategy

orig - 100% w.o. 5x plasmid

040.

Chromosomes synteny w/ her.

target - olfactory, Zn. fingers + cyto lines
important to bird - Oak Ridge group

prox, Chr 4 (Kovach 19g)

distal 15 (" 22g, 12g + P_g)

10 m bp - JGI - not P5F.

DOE wants to target Biol.

CONA Aug. full length of these topics.

About can come for Oct 20th Mtg.
send e-mail re: plans.

Mtg w/ Venter on Wed.

MOU w/ Ed Celera

Informatics - using computer power.

Craig has \$40M for computers +
env + need DOE

Will have 1 sci/chemo for a team.

"Team leader" for Gen Chemo.

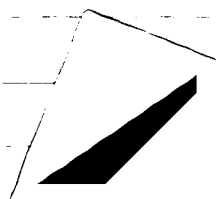
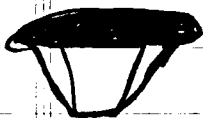
Called all DOE doing their chemos.

Random shared library - DOE make library
for them, make BAC ends from it.

5 individuals. identified ethnic prov

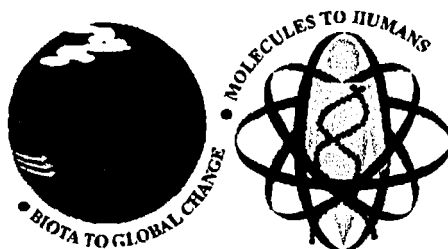
Put out every 3 mo.

No trace files. - only consensus. seq.



DOE Office of Biological and Environmental Research

Genome Research



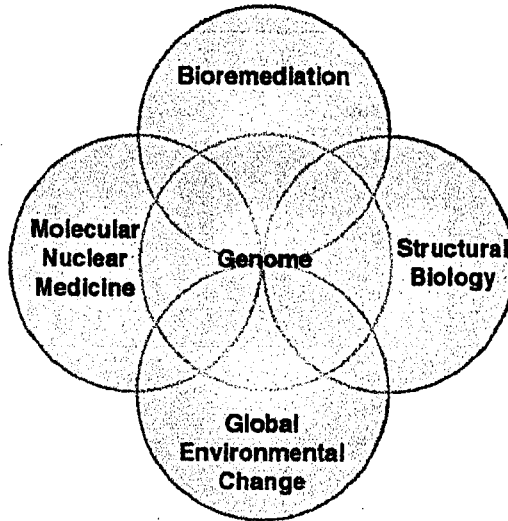
Dr. Ari Patrinos
Associate Director
Office of Biological and Environmental Research
Office of Energy Research
OMB/OSTP Briefing - September 25, 1998

Commencement Address by President Clinton at Morgan State University

May 18, 1997

**"...if the last 50 years were the age of physics, the
next 50 years will be the age of biology."**

BER Program



A diverse research portfolio driving science at the interface

BER Program



DNA \longleftrightarrow **RNA & PROTEIN** \longrightarrow **PROTEIN**
SEQUENCE implies STRUCTURE implies FUNCTION

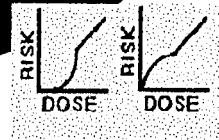
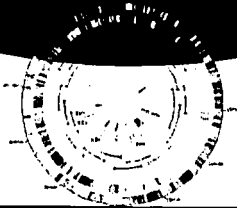


Understanding and exploiting the structure-function relationship will have far-reaching applications, e.g., in health effects research, sustainable development, and possible climate-change mitigation

*The Human Genome Project
Gateway to Tomorrow's Biology*



**Human Genome
Program**



Human Genome Program



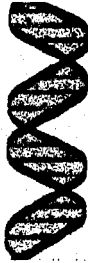
Why DOE?

- Innovative radiation biology research
- Multipurpose National Laboratories
- Unique resources/infrastructure
- Diversity of application
 - health effects research/susceptibility
 - bioremediation
 - sustainable development

HGP web site:

http://www.er.doe.gov/production/ober/hug_top.html

DOE Research Underpins U.S. Biotechnology



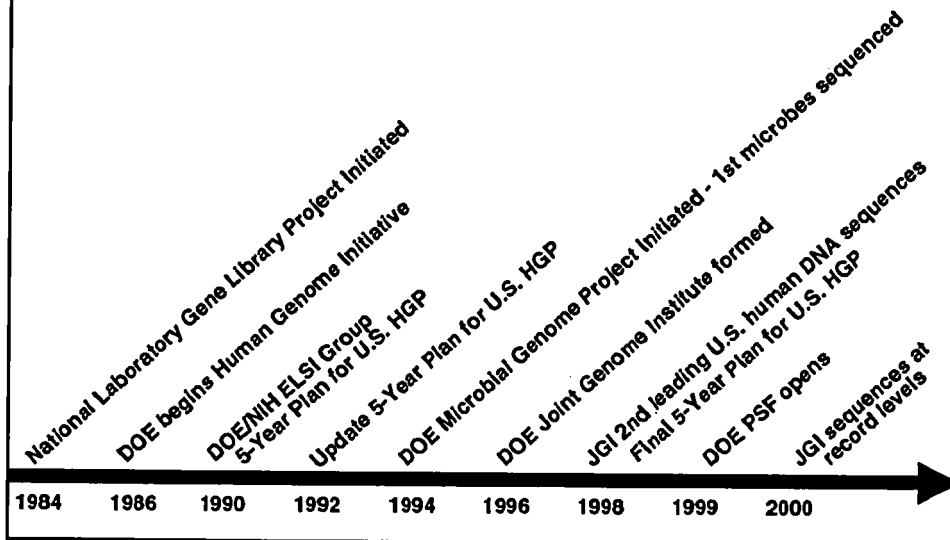
Human Genome
Microbial Genome
Structural Biology
Computational Biology

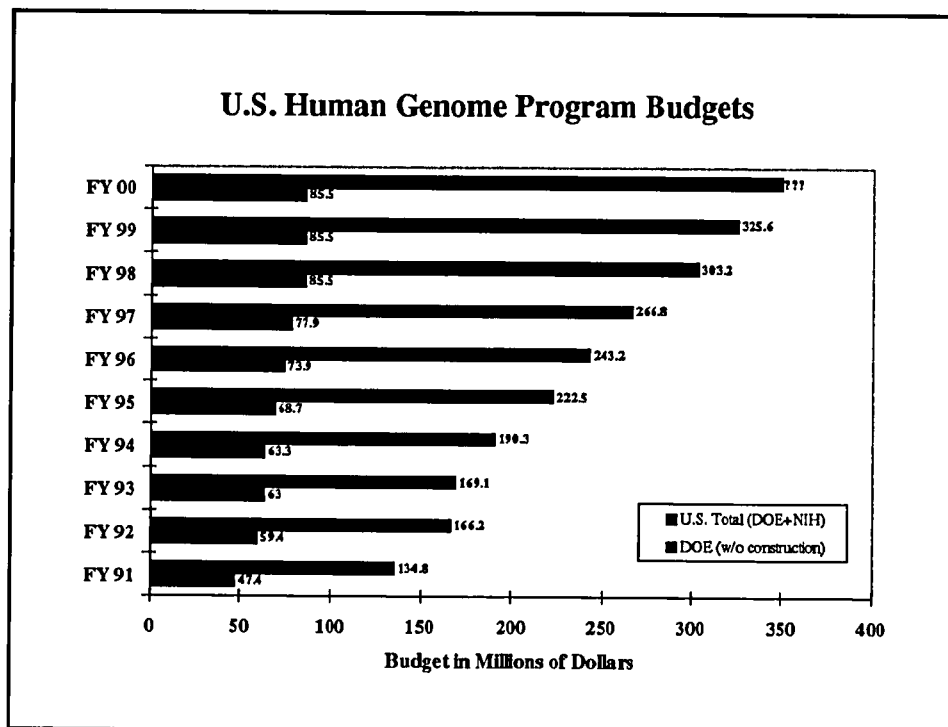
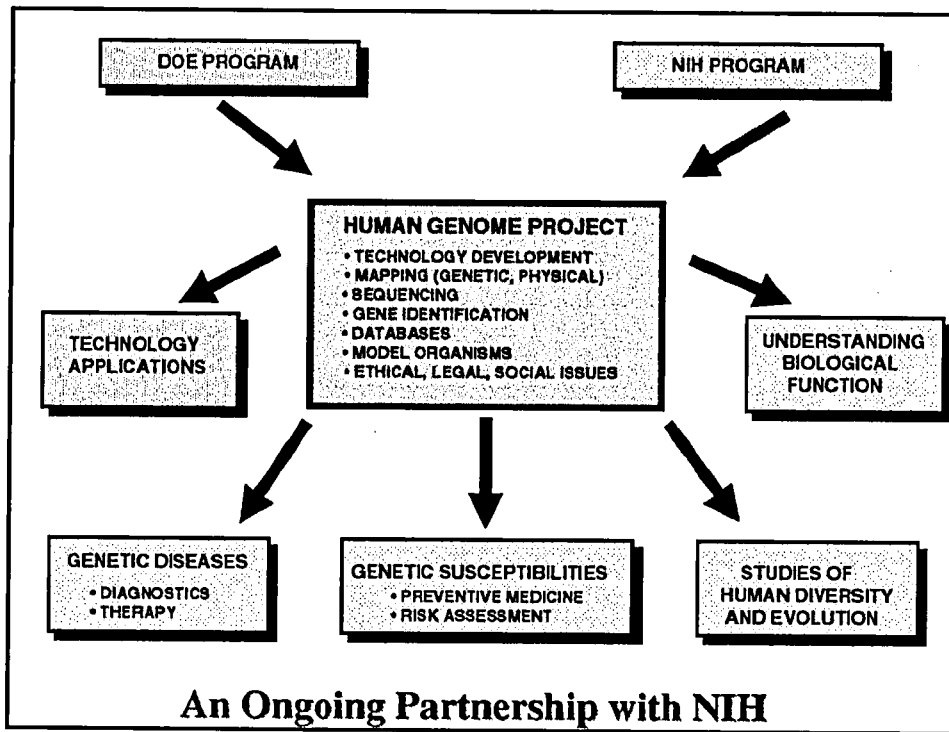


- Improved and cheaper health care
- Sustainable development
- Biofuels
- Bioremediation
- Improved industrial processes
- Increased/improved agricultural output



Highlights in DOE Genome Research





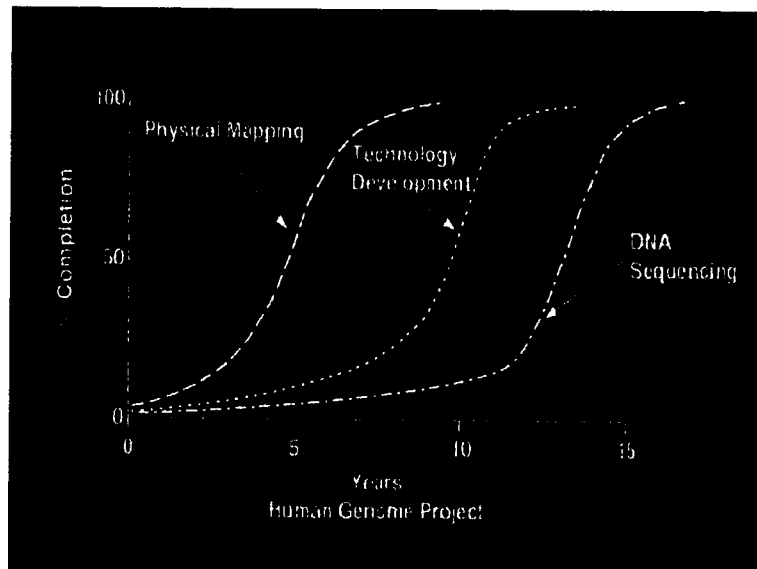
Human Genome Program

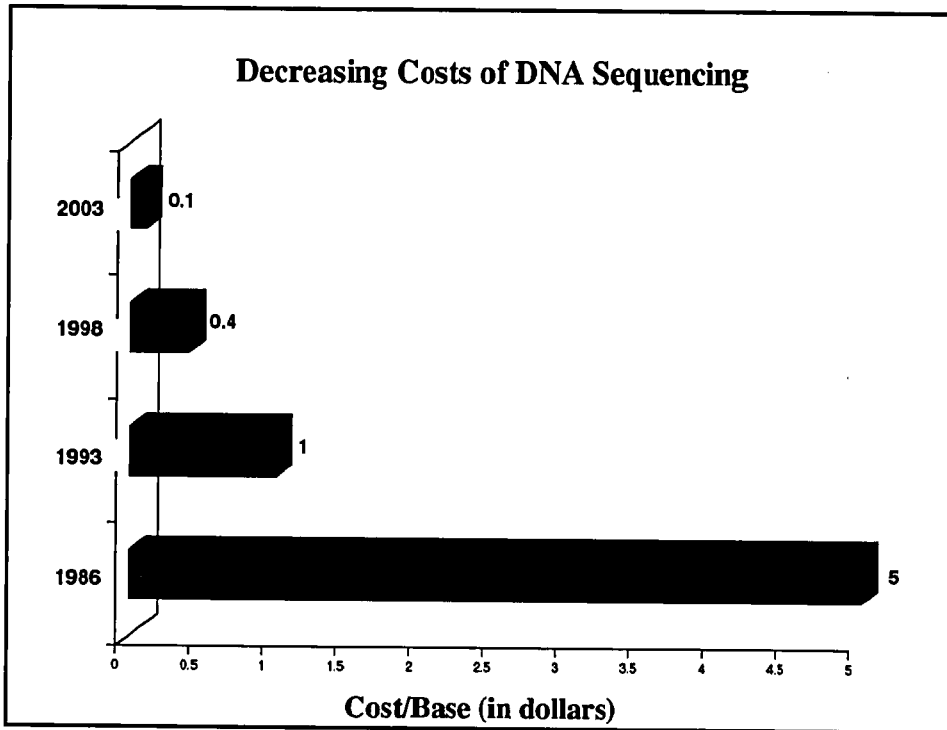


Current Challenges--

- **High Throughput Sequencing**
- **Sequencing technologies**
- **Sequencing QA/QC**
- **Resource distribution (clones, technologies, etc.)**
- **Informatics**
- **Functional Genomics ("Biology")**
- **ELSI**

The Time for Large Throughput Sequencing is Now!





Human Genome Program Third 5-Year Plan

- **Working Groups**
 - NHGRI
 - NHGRI Council
 - OBER
 - BERAC
 - ERPEG
- **Elements**
 - Sequencing
 - Informatics
 - Technology Development
 - ELSI
 - Functional Genomics
 - Comparative Genomics
 - Training
 - Sequence Variation

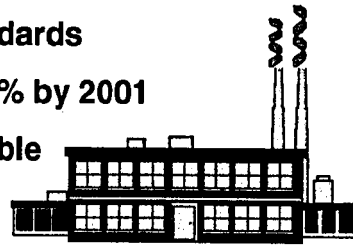
HGP Five-year Plan: October 1998 Science



New DOE/NIH Five-Year Plan

Human Sequencing Goals

- Complete by 2003
- 1/3 of Human Genome completed by 2001
 - emphasize gene rich regions
 - develop prioritization process
 - apply rigorous quality standards
- “Working Draft” coverage - 90% by 2001
- Data freely and totally accessible



New DOE/NIH Five-Year Plan (continued)



Sequencing - Related Goals

- Model Organisms
 - C. Elegans - 1998
 - Drosophila 2002
 - Mouse - 2008
- Full length cDNAs - 2003
- Continued technology development
- Sustained sequencing capacity



Standards for Sequence Produced by the Human Genome Project



Accurate

Error rate 10^{-4} or better

Assembled

Contiguous over 500kb or more

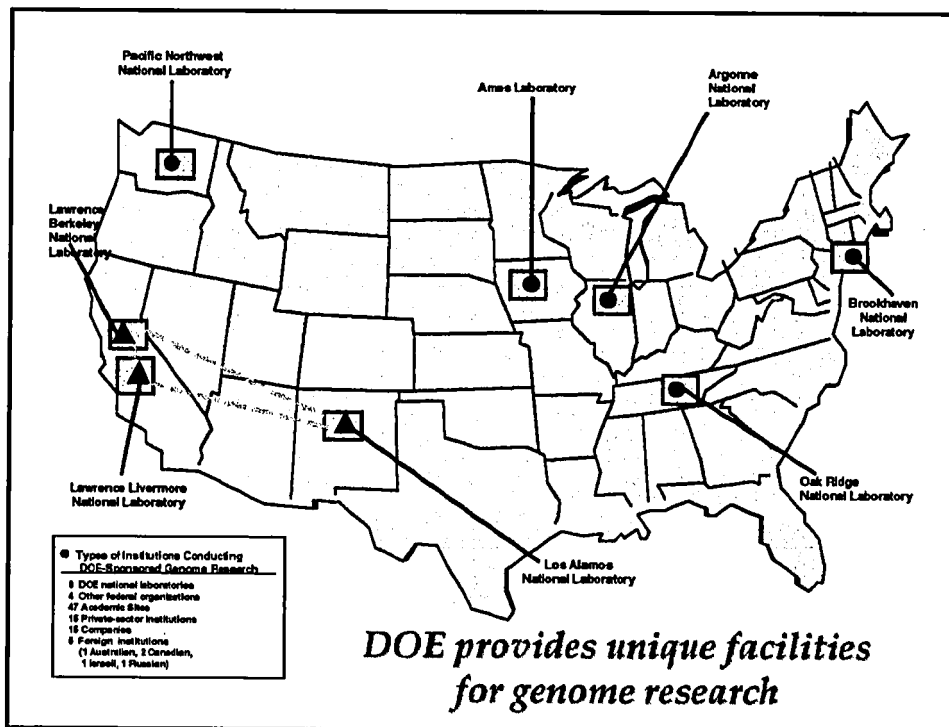
Affordable

Cost-efficient, \leq \$0.50/bp

Accessible

Public availability of sequence in <24 hours

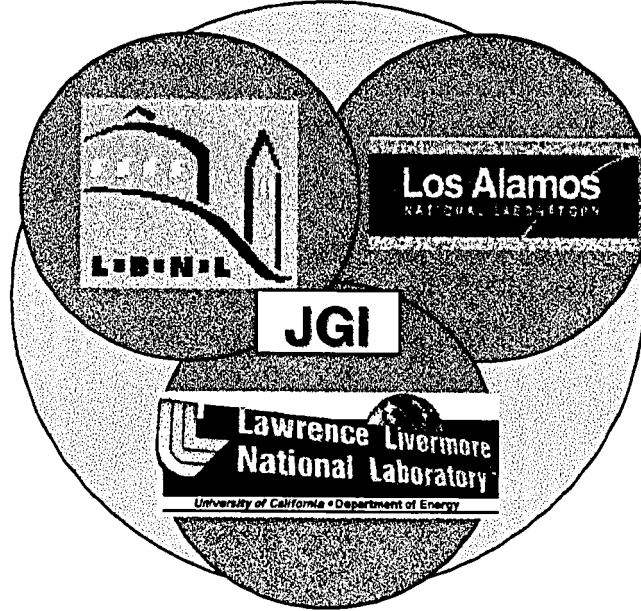
Human
Genome Project
Goals



Human Genome Project Goals: 1993-2003

Area	Goals 1993-1998	Status 10/98	Goals 1998-2003
Genetic Map	Avg. 2-5 cM resolution	1cM map published 9/94	Completed
Physical Map	Map 30,000 STSs	52,000 STSs mapped	Completed
DNA Sequence	Complete 80 Mb all organisms by 1998	180 Mb human 80 Mb <i>C. elegans</i> 14 Mb <i>Drosophila</i> 12 Mb yeast 5 Mb <i>E. Coli</i> 1 Mb mouse	Finish 1/3 of human sequence by 2001. Working draft of remainder by 2001. Complete human sequence by 2003. Sustained sequence capacity of <.5 Bbp.
Human Sequence variation	Not a goal	—	100,000 mapped SNPs Develop technology
Gene identification	Develop technology	30,000 unique ESTs mapped	Full length cDNAs
Functional analysis	Not a goal	—	Develop genomic scale technologies
Model organisms	<i>E. coli</i> : complete sequence Yeast: complete sequence <i>C. elegans</i> : most of sequence <i>Drosophila</i> : begin sequencing Mouse: map 10,000 STSs	Published 9/97 Released 4/96 80% done 9% done 12,000 STSs mapped	— — Complete 12/98 Sequence by 2002 Develop extensive genomic resources. Lay basis to sequence by 2008

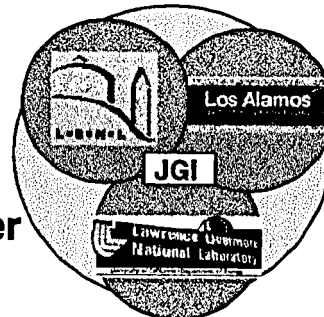
DOE Joint Genome Institute



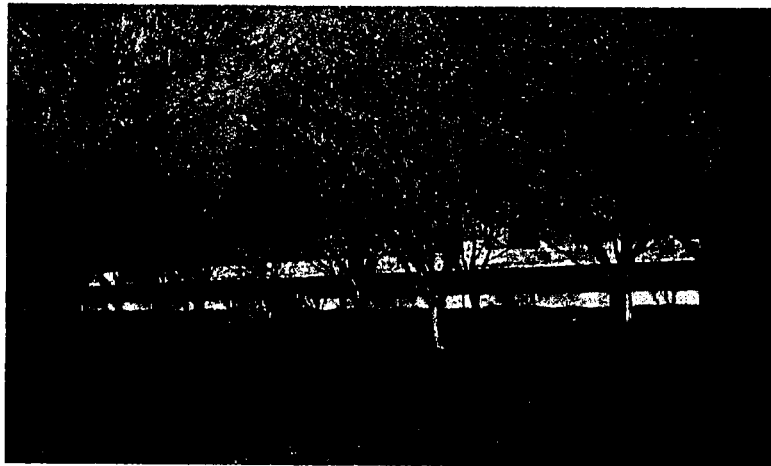
DOE Joint Genome Institute

Principles of Interaction--

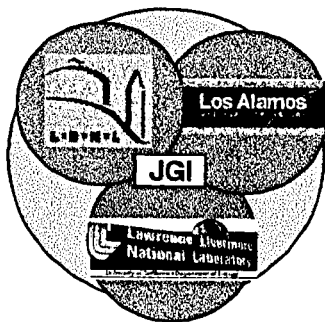
- **Virtual Center**
- **Minimize redundancy**
- **Joint planning**
- **Accountability through peer review**
- **Meeting goals/milestones**



Production Sequencing Facility Walnut Creek, California



DOE Joint Genome Institute



Genomic Focus Areas--

- **High throughput DNA sequencing**
- **Technology development**
- **Functional genomics**
- **Informatics**

Production Sequencing Facility



- First building renovation complete June 1998
- Outfitting complete October 1998
- Staff & production equipment move November/December 1998
- Dedication January 12, 1999

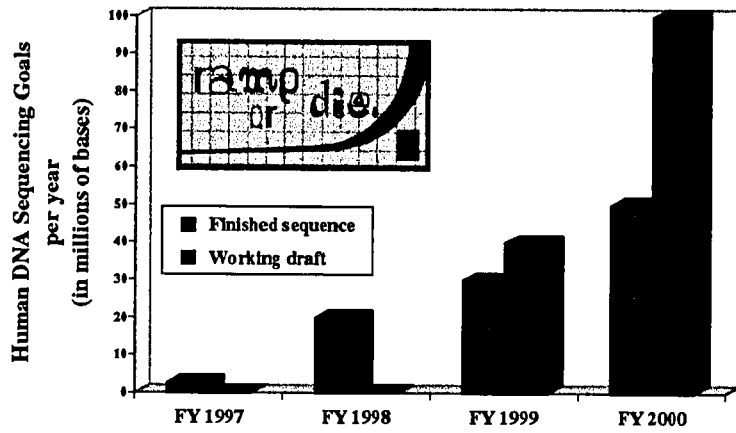
- Second building design complete
- Projected move-in summer 1999
- Additional space may be needed

Production Sequencing Facility

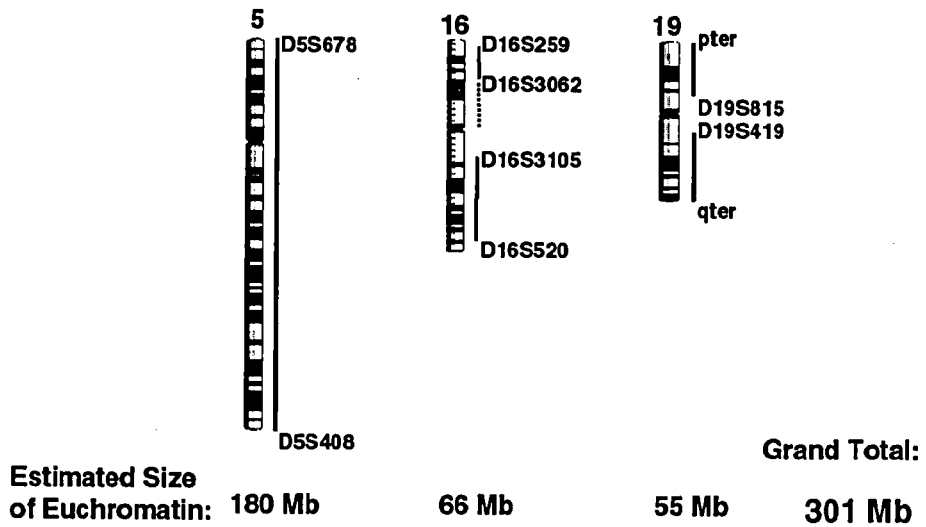


- State-of-the-art robotics factory for DNA sequencing
- Community resource for testing/implementing new sequencing methodologies/tools
- Sequencing costs competitive with best-in-the-business (\$ in sequence out)
- Collaboration with universities - \$17 million over three years to import state-of-the-art technology

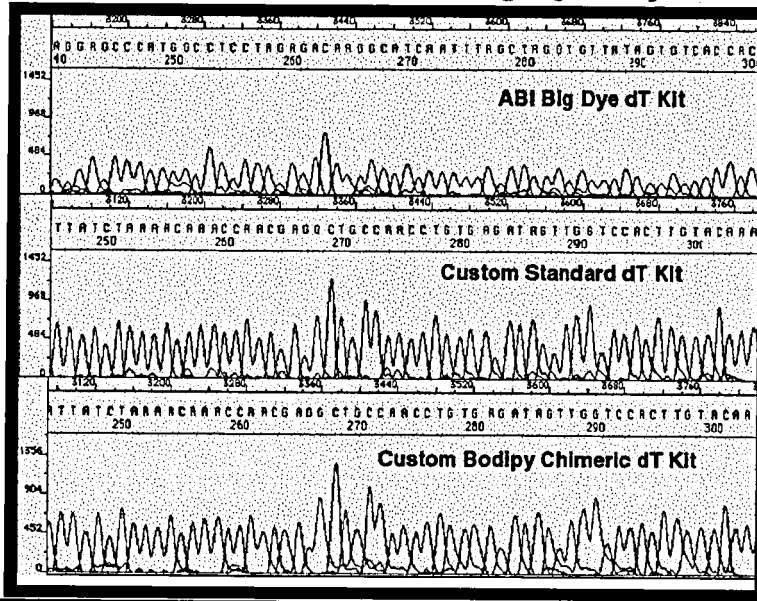
Ambitious JGI Sequencing Goals



JGI Human Sequencing Targets



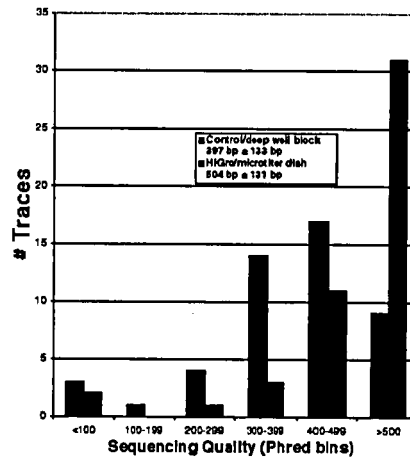
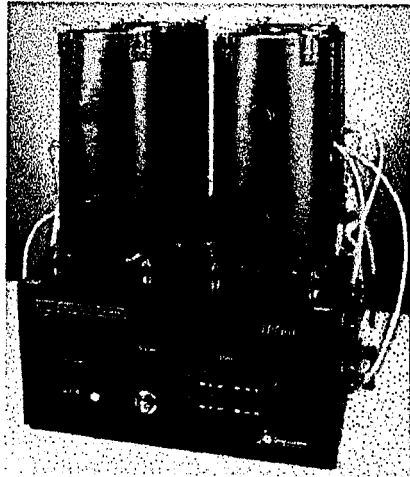
**Custom Terminator Kit with
a Bodipy-Labeled ddNTP: Highly Purified M13**



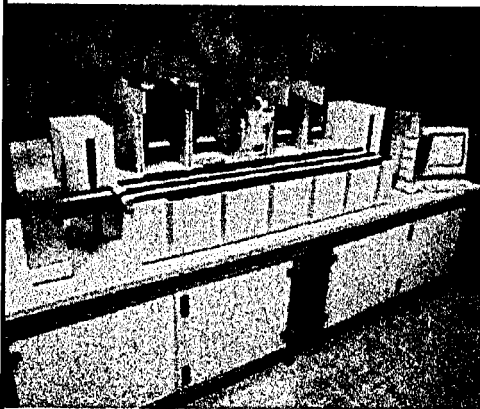
**Development of Microchannel
Sequencer Technology**



HiGro Incubator for Cell Growth in Microtiter Format



Prep Track: A Flexible, Microtiter-Compatible Liquid Transfer Robot



- Equipped with four Hydra 96 well heads
- Bulk fill station
- Conveyor belts transfer plates
- In production:
 - PCR set up
 - Cell dilution
 - Library copying
- In development:
 - Sequencing reactions
 - Plasmid template preps
- Coming soon: *Prep Track II*
 - 384 well capable
 - Carousel transfer



University JGI Interactions



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

The University of Texas
Southwestern Medical Center
at Dallas

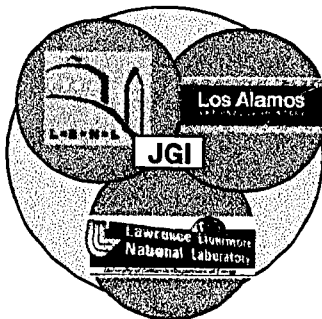
University
of Florida

STANFORD
UNIVERSITY



JGI Functional Genomics

Revealing the biological
content of sequence data



- FY 1998 pilot projects at BNL, LANL, LLNL, ORNL
- Genome-wide resources/tools
 - Mouse resources - cDNA, deletions
 - Structure / function determination
 - Analytic tools

Sequence → Function

Fxn Genomics (ORNL)
Panels of Mice with
Decreased Expression
(Knockout Mice)



JGI "Genes"

A
B
C
D
E
F
G

Fxn Genomics (LBNL)
Panels of Mice with
Increased Expression
(Transgenic Mice)



Biomedical researchers can
study mice to decipher and
rapidly access the function of
JGI "genes" whose expression
is altered

Private Sector Human Genome Initiatives



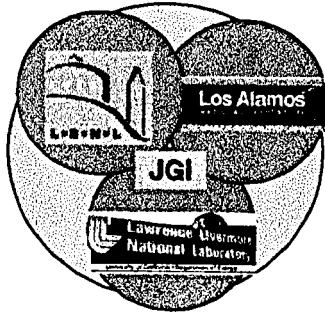
- Sequence the genome in 3 years
- DOE supported methodology
- Untested, genome-wide approach
- Many gaps will remain
- Clone location may be difficult
- Value unknown until complete?
- Computational challenges will remain
- Quarterly data release
- Some data kept as proprietary

- Sequence the genome in 2 years
- No public data release





Synergy of Public/Private Initiatives

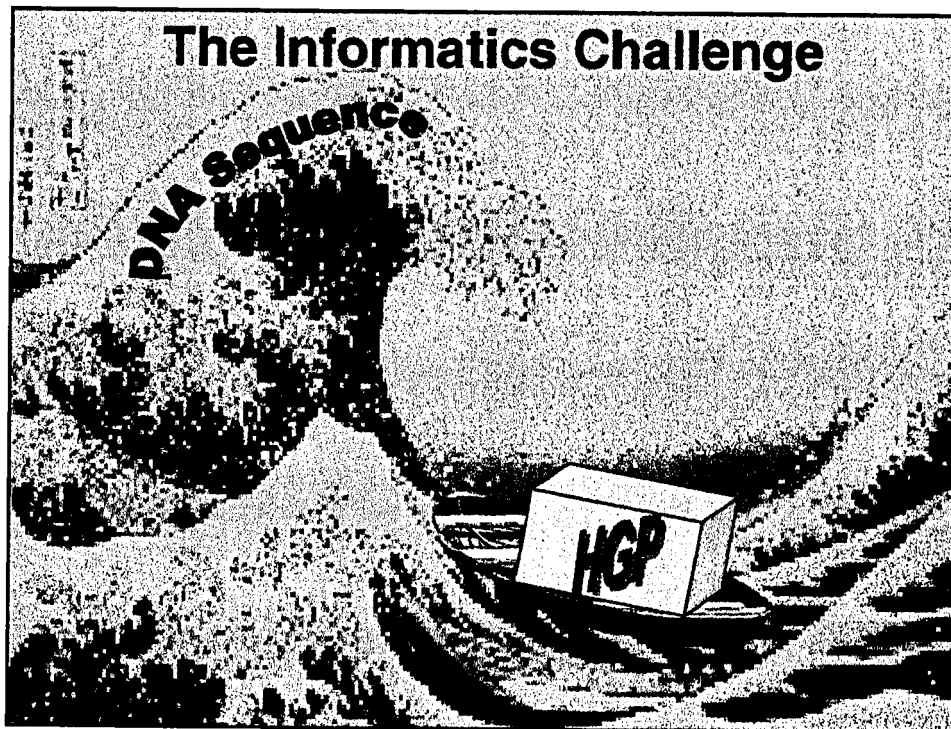


Assumptions

- Large private sector data releases
- Private sector data can be “anchored” to chromosomes using BAC end sequences

Benefits

- Reduction in overall error rates
- Remaining gaps filled
- Identification of additional human variation
- Accelerate completion of sequence





DOE Human Genome Informatics Eyes on the Prize

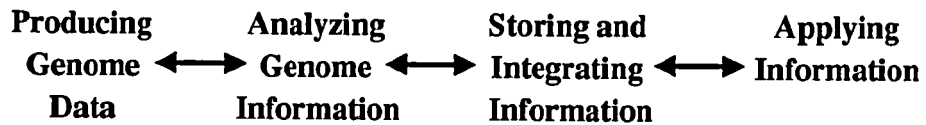
GOALS:



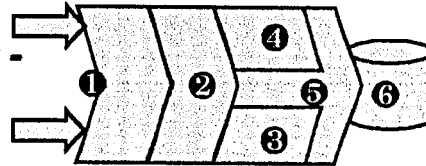
- Sequence DNA to “Bermuda standards”
- Develop tools and resources
- Integrate efforts across agencies, databases, communities
- Responsive to users



Genome Informatics... Cradle to Grave



***Genome Data Management -
From DNA to Function***

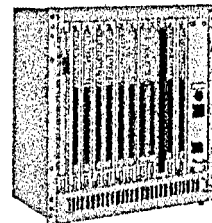


- 1 - Retrieving data and assembling genomes
- 2 - Computing genes, RNAs, proteins, features
- 3 - Computing homology, function, and other relationships
- 4 - Genome-wide structure modeling of gene products
- 5 - Analyzing and modeling pathways and systems
- 6 - Data management, access, and visualization

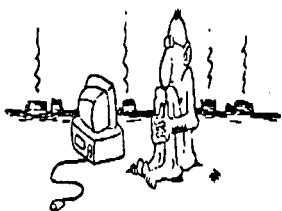


Informatics Challenge

- Currently need to process 2 million assembled base pairs per day
- New strategy will radically change the landscape
- Data generation January 1, 1999 - 30 million per day.
Mid 1999 - 100 million bases per day
- Most comprehensive analyses will be beyond capabilities of all but a few sites



Projected Computer Requirements

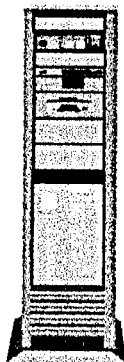


- Assembly - Estimated at 1.5 TeraOPS/day
- Gene modeling - Estimated at 100 GigaOPS/day

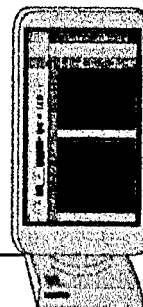
- Protein threading - Estimated at 1 TerOP/day
- Homology, protein classification, etc... - ??



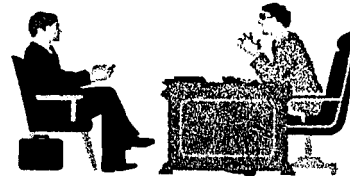
DOE-HGP Near-term Computing & Hardware Needs



- Routine access to high-performance computing facilities and servers
- Large local clusters
- Terabyte-scale local disk storage in 1999
- Access to high-performance storage



DOE ELSI Program

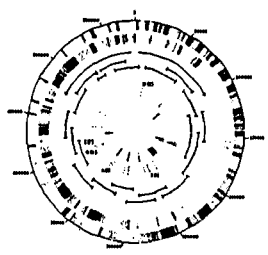
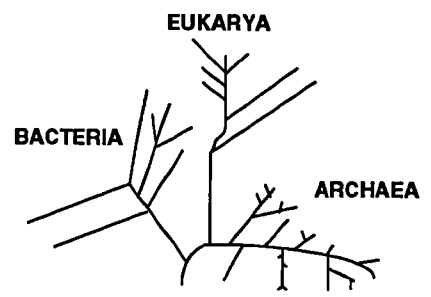


DOE ELSI Program Today



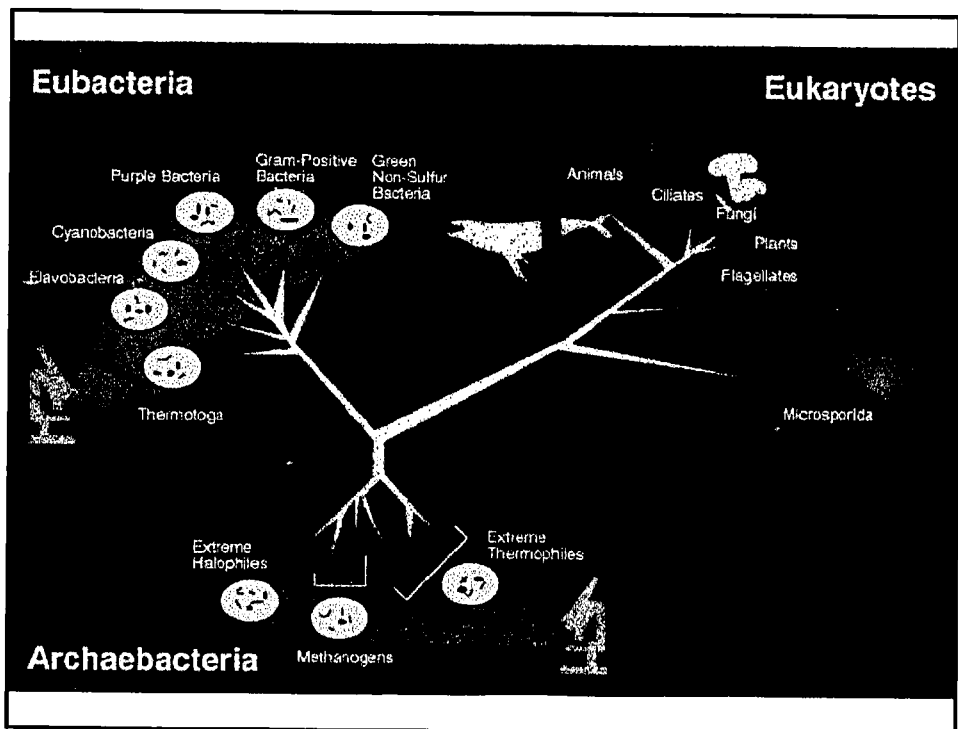
- **Privacy and confidentiality of genetic information**
 - workplace
 - databases
- **Intellectual property and commercialization issues**
- **Professional and public education**
 - IRBs
 - specific groups, e.g., judges, etc.
- **Societal issues of research on complex traits**

BER Program



**Pioneering and exploiting
Microbial Genome research for DOE needs**

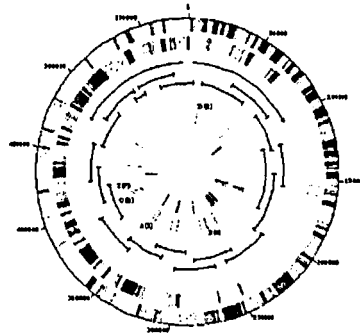
HGP web site: http://www.er.doe.gov/production/ober/hug_top.html



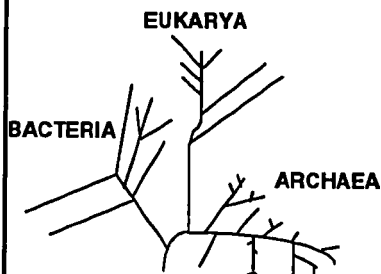


Microbial Genome Program

- Capitalizes on advances in human genome program
- Mapping/sequencing microbes with
 - environmental/energy relevance
 - phylogenetic significance
 - commercial value
- Prediction of gene function

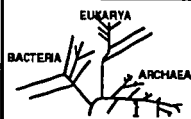
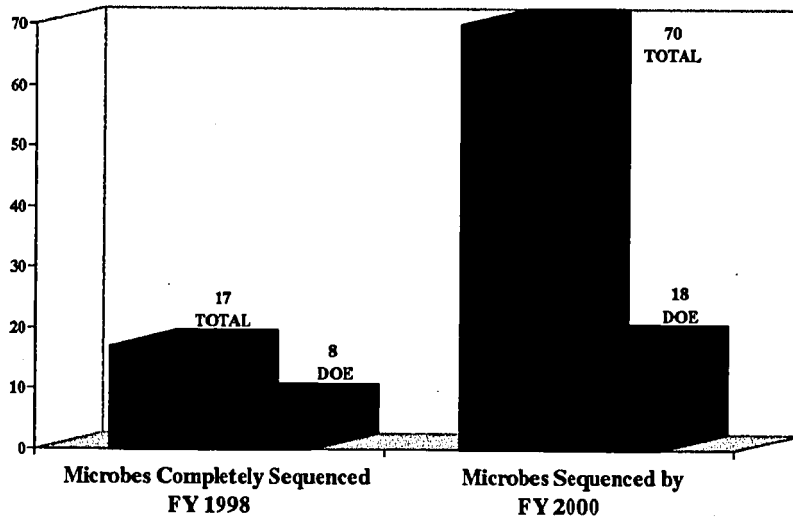


Microbial Genome Program Selection Criteria



- Energy/Environmental relevance
 - energy use/production
 - carbon cycle
 - bioremediation
 - waste cleanup
- DNA obtainable
- Genome size (<8 Mb)
- Genetically manipulable
- Nonpathogenic
- Scientifically interesting

**Accomplishments
DOE Initiated Microbial Genome Program -
The Revolution Continues**

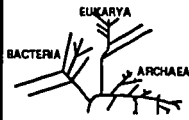


**DOE Microbial Genome
Program**



Sequencing Completed --

- *Mycoplasma genitalium* -- free living, smallest genome
- *Methanococcus jannaschii* -- methane producer, 3rd life form
- *Archaeoglobus fulgidus* -- oil well souring
- *Thermotoga maritima* -- energy from plant biomass
- *Deinococcus radiodurans* -- radiation resistant, bioremediation
- *Methanobacterium thermoautotrophicum* -- methane producer
- *Pyrobaculum aerophilum* -- thermophile (100° C)
- *Aquifex aeolicus VF5* -- deep branching lineage



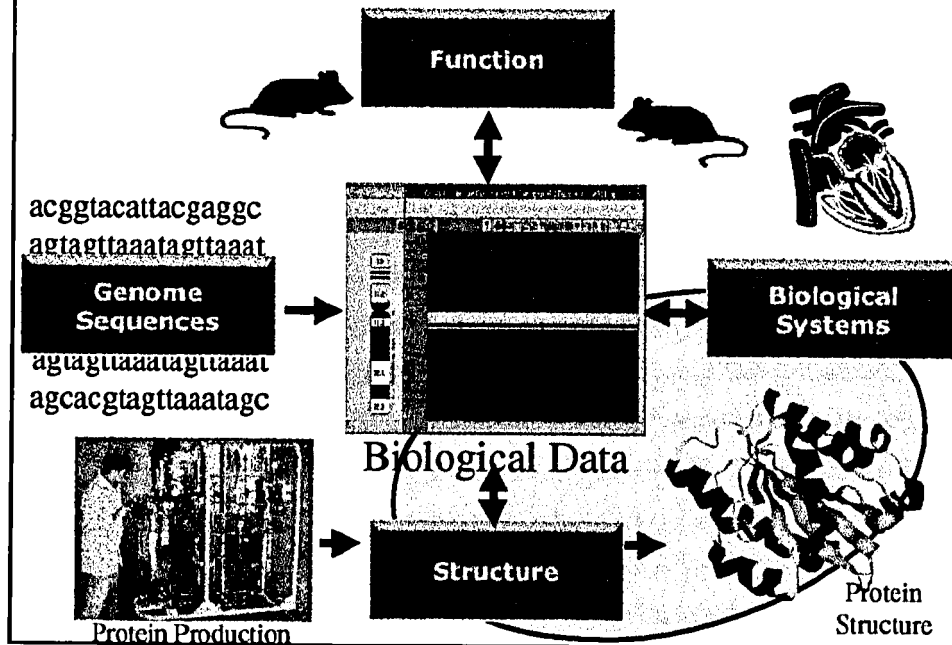
DOE Microbial Genome Program



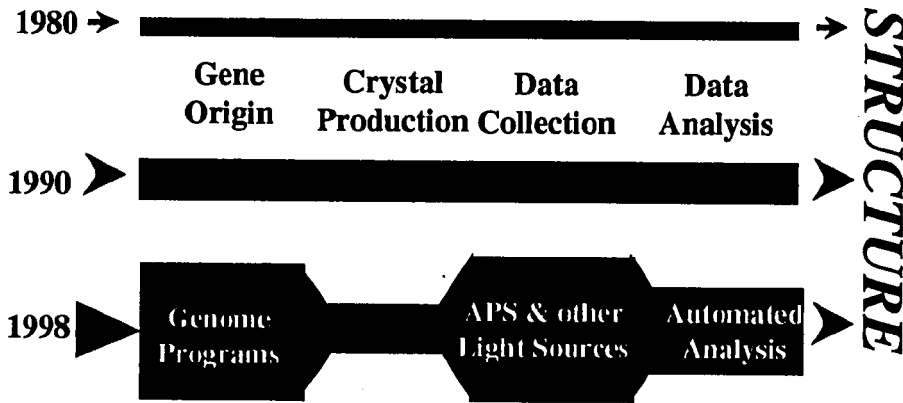
Sequencing in progress--

- *Pyrococcus furiosus* -- model hypothermophile
- *Clostridium acetobutylicum* -- biotech & waste remediation
- *Shewanella putrefaciens* -- bioremediation
- *Pseudomonas putida* -- bioremediation
- *Thiobacillus ferroxidans* -- CO₂ fixation
- *Desulfovibrio vulgaris* -- bioremediation
- *Caulobacter crescentus* -- bioremediation
- *Chlorobium tepidum* -- carbon management
- *Dehalococcoides ethenogenes* -- bioremediation
- *Carboxydotherrmus hydrogenoformans* -- H₂ production

Planning and Launching the "Proteome" Project



Evolution of the Protein Structure Pipeline



DNA Patenting



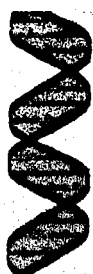
“The Tragedy of the Anticommons”

US Patent and Trademark Office

BER Program



DNA \longleftrightarrow **RNA** \longrightarrow **PROTEIN**
SEQUENCE implies STRUCTURE implies FUNCTION



Understanding and exploiting the structure-function relationship will have far-reaching applications, e.g., in health effects research, sustainable development, and possible climate-change mitigation

Responses to Notices 98-16 and LAB98-16

Proposal	Title	Principal Investigator	Co-Investigator(s) and Collaborators	Venue
66840	High Speed High Thruput Mutation Detection	Yeung, Edward S.	Oefner, Peter	AMES LABORATORY, Ames, IA
66855	DNA Sequence Ladder Readout by Massive Cluster Impact Mass Spectrometer	Williams, Peter	Mahoney, John	ARIZONA STATE UNIVERSITY, Tempe, AZ
66875	Reactive Sequencing of DNA	Williams, Peter	Bloom, Linda; Hayes, Mark A.; Rose, Seth D.; Reba-Krantz, Linda J.; Towe, Bruce C.; Pizzieoni, Vincent B.	ARIZONA STATE UNIVERSITY, Tempe, AZ
066899	A Novel Approach for Identification of Specific Regions in Cloned Genomes	Lyubchenko, Yuri	Sinden, Richard R.; Stasiak, Audrzej	ARIZONA STATE UNIVERSITY, Tempe, AZ
066883	A Hand-Held Capacitive Genosensor Chip	Whitaker, Tom J.	Jacobson, K. Bruce; Willey, Kenneth; Larimer, Frank W.; Doktycz, Michael J.; Egholm, Michael	ATOM SCIENCES, INC., Oak Ridge, TN
066884	High Throughput Single DNA Molecule Sizing and Sorting for Genomic Analysis	Quake, Stephen	Simon, Mel	CALIFORNIA INST OF TECHNOLOGY, Pasadena, CA
066849	High Performance Labeling, Separation and Detection Methods for Genome Analysis	Mathies, Richard A.	Glazer, Alexander N.; Scherer, James; Seusabaugh, George F.	CALIFORNIA, UNIV. OF BERKELEY, Berkeley, CA
066898	New Methods for Sequencing Individual DNA Molecules	Webb, Watt W.	Craighead, Harold G.	CORNELL UNIVERSITY, Ithaca, NY
066887	A Novel Device for DNA Sequencing Based on DNA-Mediated Electron Transfer	Grinstaff, Mark W.	Khan, Shoeb; Beilstein, Amy E.; Hu, Xi	DUKE UNIVERSITY, Durham, NC
066851	Automated DNA Sequencing Using Continuous Flow IR-MALDI Mass Spectrometry	Murray, Kermit K		EMORY UNIVERSITY, Atlanta, GA
066897	Integrated Automation for Large Scale Sequencing	Smith, James H.	Engelstein, Marcy; Madan, Deepika; Sietz, Bruce Robert	GENOME THERAPEUTICS CORP., Waltham, MA

066889	Development of a Cryogenic Based Mass Spectrometer for DNA Sequencing	Gillevet, Patrick M.	Eckenrode, Brian A.; Christenson, Jeffrey; Huang, Yuchi; Barry L. Bruber; Eric Z. Qiu	GEORGE MASON UNIVERSITY, Fairfax, VA
066894	Optimization and Validation of Two Novel High-Throughput DNA Sequencing Instruments in Production Settings	Seto, Donald	Tibbetts, Clark	GEORGE MASON UNIVERSITY, Fairfax, VA
066918	Cell-Based Assay for High Throughput Gene Function Analysis	Hickman, James J.	Krauthamer, Victor; Ravenscroft, Melissa; Wheeler, Bruce; Quackenbush, John	GEORGE WASHINGTON UNIVERSITY, Washington, DC
066895	Carbon Nanotube Probes for Rapid DNA Sequencing	Lieber, Charles M.		HARVARD UNIVERSITY, Cambridge, MA
066853	Externally Controllable Sample Capture and Cleanup for DNA in Micromachined Integrated DNA Analysis Systems	Swedler, Jonathan	Bohn, Paul	ILLINOIS, UNIVERSITY OF, Champaign, IL
066904	Instrumentation for DNA Fiber Mapping	Weier, Heinz-Ulrich	Lersch, Robert Alan; Kim, Ung-Jim; Pedersen, Roger A., Jan-Fang Cheng; Christopher H. Martin	LAWRENCE BERKELEY NATIONAL LAB, Berkeley, CA
066906	Mass Spectrometry for DNA Sequencing Verification	Benner, W. Henry	Frank, Mathias; Labos, Simon	LAWRENCE BERKELEY NATIONAL LAB, Berkeley, CA
066907	Improved Coatings & Sieving Media for DNA Sequencing	Madabhushi, Ramki	Balch, Joseph W.; Carrano, Anthony; Goldberg, Eugene; Thico Hogen-Ereh	LAWRENCE LIVERMORE NATL LAB, Livermore, CA
066859	Single Molecule DNA Sequencing	Keller, Richard A.	Jett, James H.; Goodwin, Peter M.; Cai, Hong; Katrin Kneipp; Linda Rebakrantz; Richard Smith	LOS ALAMOS NATIONAL LABORATORY, Los Alamos, NM
066912	Sample Handling for High Throuput Genomic Analysis	Nolan, John P.	White, P. Scott; Jett, James H.; Bengelsdijk, Tony; Daryl Ricke	LOS ALAMOS NATIONAL LABORATORY, Los Alamos, NM
066913	A Novel Technology for High Throughput DNA Sequence Validation: Stable-Isotope Assisted Mass Spectrometry	Chen, Xian	Majidi, Vahid; Duan, YiXiang; Smith, Lloyd M.; Norman Doggett; Larry Deaven	LOS ALAMOS NATIONAL LABORATORY, Los Alamos, NM

066881	Development of High Throughput, Micro-Systems for DNA Genotyping and Diagnostic Applications	Soper, Steven A.	Liubach, Patrick; Murphy, Michael C.; Kelly, Kevin W.; Nikitopoulos, Dimitris E.; Batzer, Mark A.	LOUISIANA STATE UNIVERSITY, Baton Rouge, LA
066850	An Intelligent System for Accurate Identification of DNA Bases	Musavi, Mohamad	Van Beneden, Rebecca	MAINE, UNIVERSITY OF, Orono, ME
066920	Ultrafast Sequencing of DNA and Other Polymers	Bension, Rouvain M.	Thundat, Thomas G.	NEOTECH DEVELOPMENT CO. LLC, Marlborough, MA
066852	Biosensor Array for Detecting Multiple DNA Sequences	Wang, Joseph	Chiu, C.	NEW MEXICO STATE UNIVERSITY, Las Cruces, NM
066921	Development of Advanced Systems for Optical Mapping	Schwartz, David C.	Miohra, B.; Anantharaman, Thomas S.; Aston, Christopher W.; Huff, Edward J.; Porter, Brent E.A.; Wang, Weining	NEW YORK UNIVERSITY, New York, NY
066917	Electrical Sequencing of DNA	Sachs, Frederick	Gottlieb, Philip A.	NEW YORK, STATE U. OF BUFFALO, Amherst, NY
066879	Advances in DNA Sequencing by Capillary Array Electrophoresis: Extended Sequence Read Length, Micro-fluidic Sample Preparation & on Expert System Basecaller	Karger, Barry L.	Foret, Frantisek; Kotler, Lev; Miller, Arthur W.; Schander, Eric	NORTHEASTERN UNIVERSITY, Boston, MA
066880	Microchannel DNA Sequencing by End-Labelled Free Solution Electrophoresis (ELFSE): Development of Polymeric End-Labels, Microchannel Coatings, & Electrophoresis	Barron, Annelise E.	Slater, Gary W.; Letsinger, Robert L.; Zuckermann, Ronald N.; J. William Efeavitch	NORTHWESTERN UNIVERSITY, Evanston, IL
066892	New High Resolution Method for Separation of DNA Fragments	Updyke, Timothy	Burlatsky, Sergei F.; Eason, Bruce H.; Bogoev, Roumen A.; Amshey, Joseph W.	NOVEL EXPERIMENTAL TECHNOLOGY, San Diego, CA
066900	Novel Detection System for DNA Hybridization Without Extrinsic Labeling	Thundat, Thomas G.	Doktycz, Mitchel J.; Warmack, R.J.; Mlcak, Richard	OAK RIDGE NATIONAL LABORATORY, Oak Ridge, TN
066901	Flowthrough Genochips	Beattie, Kenneth L.	Doktycz, Mitchel J.; Zhan, Ming; Stubbs, Lisa; Rugan, William L.	OAK RIDGE NATIONAL LABORATORY, Oak Ridge, TN

066902	High-Speed DNA Sequencing in Microdevices by Sequential Single Base Extension	Ramsey, J. Michael	Jacobson, Stephen C.; Foote, Robert S.; Waters, Larry C.	OAK RIDGE NATIONAL LABORATORY, Oak Ridge, TN
066903	Integrated Microdevices for High-Speed DNA Sequencing	Ramsey, J. Michael		OAK RIDGE NATIONAL LABORATORY, Oak Ridge, TN
066905	Laser Desorption Mass Spectrometry for Fast Human Genome Sequencing	Chen, Chung-Hsuan Winston	Pinnaduwege, Lal; Beattie, Kenneth L.; Valerie V. Golovlev	OAK RIDGE NATIONAL LABORATORY, Oak Ridge, TN
066908	A Novel Multiplex Labeling Technique Based on Surface-Enhanced Raman for Genomics Analysis	Vo-Dinh, Tuan	Griffin, Guy D.; Michaud, Edward J.; Wintenberg, Alan L.; Ung-Jin Kim; Melvin I. Simon	OAK RIDGE NATIONAL LABORATORY, Oak Ridge, TN
066909	High Speed DNA Sequencing by Electrospray and Ion/Ion Chemistry	McLuckey, Scott A.	Stephenson, J.L.; Hurst, G.B.	OAK RIDGE NATIONAL LABORATORY, Oak Ridge, TN
066911	Direct Genomic Sequencing by Hybridization	Doktycz, Mitchel J.	Beattie, Kenneth L.; Britton, Charles L.; Britton, William L.	OAK RIDGE NATIONAL LABORATORY, Oak Ridge, TN
066837	The Development of Electrospray Ionization-Mass Spectrometry for High Speed DNA Sequencing and Ultra-High Sensitivity Characterization of Modified DNA	Smith, Richard	Bruce, James E.; Anderson, Gordon; Winsehel, David S.; Binglaing Feng	PACIFIC NORTHWEST NATIONAL LAB, Richland, WA
066838	Application of Elctrospray Ionization-Mass Spectrometry to Single Molecule DNA Sequencing	Smith, Richard	Keller, Richard A. ; Anderson, Gordon	PACIFIC NORTHWEST NATIONAL LAB, Richland, WA
066839	The Development of Proteome Characterization for Broad Genomic Surveys of Gene Function	Smith, Richard	Bruce, James E. ; Weir-Lipton, Mary S.; Anderson, Gordon	PACIFIC NORTHWEST NATIONAL LAB, Richland, WA
066856	A Novel, Rapid Approach to DNA Sequencing by Surface Plasmon Resonance	Natan, Michael J.	Benkovic, Stephen J.	PENNSYLVANIA STATE UNIVERSITY, University Park, PA
066919	Complex Gradients in Continuously Cast DNA Gels--An Unexplored Approach to Maximum Gel Information Density in High Speed DNA Analysis for the Human Genome	Champagne, James T.	Meyer, Stephen	PROTEOTOOLS, INC., Seattle, WA
066874	Integrated Optics for Chip-Based Sequencing of DNA	Lytel, Fred E.	Regnier, Fred	PURDUE RESEARCH FOUNDATION, West Lafayette, IN

066910	High-Throughput, Long-Read-Length Single-Molecule DNA Sequencing	Trautman, Jay K.	Du, Mei; harris, Timothy D.; Linford, Matthew R.; Macklin, John J.; Mitsis, Paul G.; Nicklaw, Neal; Perkins, Thomas T.	SEQ, LTD., Lawrenceville, NJ
066890	Peptide Nucleic Acid-Based Diagnostic Screening for Genetic Mutations	Zare, Richard N.		STANFORD UNIVERSITY, Stanford, CA
066896	High Molecular Density Immobilization of DNA and Micro-Array Scanning by Kelvin Microprobe for Sequencing by Hybridization	Thompson, Michael		TORONTO, UNIVERSITY OF, Ontario, CAN
066882	Time Resolved Sequence Analysis on High Density Fiberoptic DNA Probe Arrays	Walt, David R.	Chee, Mark S.; Czarnik, Anthony	TUFTS UNIVERSITY, Medford, MA
066885	DNA Sequence Analysis and Data Reduction	Kepart, Thomas		VANDERBILT UNIVERSITY, Nashville, TN
066873	Manipulation of Small Quantities of DNA with Induced-dipole Traps	van den Engh, Gerrit J.	Ashury, Charles L.	WASHINGTON, UNIVERSITY OF, Seattle, WA
066888	A Cell Sorter with Tape Conveyor System for the Generation of Sequencing Samples	van den Engh, Gerrit J.	Esposito, Richard; Fey, Carran; Choe Juno	WASHINGTON, UNIVERSITY OF, Seattle, WA
066891	Fundamental Physics of Spin Relaxation in Magnetic Resonance Force Microscopy	Sidles, John A.		WASHINGTON, UNIVERSITY OF, Seattle, WA
066929	Addressing the Issue of 'Problem Sequences' in the Human Genome Project	SantaLucia, Jr., John	Deuereux, John; Hofacker, Iuo; Wilson, Richard K.; Neri, Bruce P.; Zuker, Michael; Friend, Stephen H.	WAYNE STATE UNIVERSITY, Detroit, MI

ID/Org	Proposer/Principal Investigator	Received	Title	Requested	Type/Action	Manager/Detailee
067117 ER-72	FOUNDATION FOR GENETIC MED INC Manassas, VA Alpert, Sheri A.	09/17/98	Machine-Tractable Human Tissues: Policy Implications for Medical Privacy	\$389,058 24 months	New Under Review	Drell
067129 ER-72	GEORGE WASHINGTON UNIVERSITY Washington, DC Schaffner, Kenneth	09/17/98	Genes and Environments in Behavioral and Psychiatric Genetics: Ethical Implications of Complex Trait Genetics	\$497,953 24 months	New Under Review	Drell
067135 ER-72	INSTITUTE OF GENETICS ED. Santa Fe, NM Dillingham, Clay	09/17/98	Raising African-American Awareness of the Ethical, Legan, & Social Issues of the Human Genome Project: A Pilot Project Targeted Toward Rural & Urban African-Am	\$480,700 36 months	New Under Review	Drell
067145 ER-72	KANSAS, UNIVERSITY OF Kansas City, KS Collins, Debra L.	09/18/98	GeneNet: Human Genome Teacher Education Program	\$608,648 36 months	New Under Review	Drell
067121 ER-72	MIAMI, UNIVERSITY OF Miami, FL Goodman, Kenneth W.	09/17/98	Ethical Issues in Bioinformatics	\$446,718 36 months	New Under Review	Drell
067119 ER-72	MICHIGAN, UNIVERSITY OF Ann Arbor, MI Eisenberg, Rebecca S.	09/17/98	Private Appropriation, Public Dissemination, and Commercial Product Development in Genomics	\$113,387 18 months	New Under Review	Drell
067115 ER-72	NOELEYE DOCUMENTARIES San Francisco, CA Schwerin, Noel	09/17/98	TRUTH & JUSTICE: Science and its Appeals - A Three-Hour Documentary Television Special for National PBS	\$497,811 18 months	New Under Review	Drell
067113 ER-72	OREGON HEALTH SCIENCES UNIV. Portland, OR Bevan, Leslie	09/17/98	Impact of Genetic Privacy Legislation on Research	\$400,793 36 months	New Under Review	Drell
067133 ER-72	RAND CORPORATION Santa Monica, CA Eiseman, Elisa	09/17/98	Genetic Research and the Education of Institutional Review Boards	\$249,598 18 months	New Under Review	Drell
067112 ER-72	RHODE ISLAND, UNIVERSITY OF Kingston, RI Pasquerella, Lynn Rothstein, Lawrence E.	09/16/98	An Investigation of the Ethical Concepts that Inform the Laws Limiting Genetic Screening in Employment Decisions: Privacy, Dignity, Nondiscrimination, Autonomy	\$286,179 24 months	New Under Review	Drell
067132 ER-72	SELF RELIANCE FOUNDATION Santa Fe, NM Salazar, Roberto	09/17/98	Hispanic Role Model and Science Education Outreach Project: HumanGenome Project Education and Outreach Component	\$451,386 36 months	New Under Review	Drell
067118 ER-72	SOUNDVISION PRODUCTIONS Berkeley, CA Scott, Barinetta	09/17/98	Science Training Initiative	\$247,031 5 months	New Under Review	Drell
067125 ER-72	SOUNDVISION PRODUCTIONS Berkeley, CA Scott, Barinetta	09/17/98	The DNA Files	\$406,073 12 months	New Under Review	Drell
067130 ER-72	SOUTH CAROLINA, MED. UNIV. OF Charleston, SC Musham, Catherine	09/17/98	The Potential Role of Biomarkers in Occupational Prevention: Stakeholder Perspectives on Ethical, Legal, and Social Issues	\$172,238 12 months	New Under Review	Drell
067114 ER-72	STANFORD UNIVERSITY Stanford, CA Koenig, Barbara A.	09/17/98	Dilemmas in Commercializing Human Genome and Biotechnology Products: Developing a Case-based Business Ethics Curriculum for Industry	\$256,323 18 months	New Under Review	Drell
067120 ER-72	STANFORD UNIVERSITY Stanford, CA Cho, Mildred	09/17/98	Case Studies in Patented Genetic Tests	\$884,230 36 months	New Under Review	Drell

ID/Org	Proposer/Principal Investigator	Received	Title	Requested	Type/Action	Manager/Detailee
067128 ER-72	STANFORD UNIVERSITY Stanford, CA Tobin, Sara L.	09/17/98	Getting the Word Out on the Human Genome Project: A Course for Physicians	\$350,498 24 months	New Under Review	Drell
067124 ER-72	TEXAS TECH UNIVERSITY Lubbock, TX Knaff, David B.	09/17/98	Journalism and the Human Genome Initiative: A Conference on Science and the Media	\$330,778 36 months	New Under Review	Drell
067136 ER-72	TUSKEGEE UNIVERSITY Tuskegee, AL Smith, Edward J.	09/17/98	Raising African-Am Awareness of the Ethical, Legal & Social Issues of the Human Genome Project: A Pilot Project Targeted Toward Rural & Urban African-Am Comm	\$331,477 36 months	New Under Review	Drell
067134 ER-72	WASHINGTON, UNIVERSITY OF Seattle, WA Faustman, Elaine M.	09/17/98	Medical Effectiveness, Socioeconomic Costs, and Legal and Ethical Issues in Development of Occupational Risk Management Programs Using Genetic Biomarkers	\$311,673 24 months	New Under Review	Drell
067131 ER-72	WORLDVIEW PICTURES LTD. Saratoga Springs, NY Trombley, Stephen	09/17/98	Bad Luck: The Gene Lottery -- a 52' documentary film	\$224,827 5 months	New Under Review	Drell

Count: 21

067143 Amer Soc. Microbiology
Washington, DC
Needham, Cynthia

11/13/98

Microbial Literacy
Collaborative: Intimate
Strangers: Unseen Life on Earth

\$200,000
12 months

Drell

<u>ID/Org</u>	<u>Proposer/Principal Investigator</u>	<u>Received</u>	<u>Title</u>	<u>Requested</u>	<u>Type/Action</u>	<u>Manager/Detailee</u>
067214 ER-72	EINSTEIN INSTITUTE FOR SCIENCE Bethesda, MD Zweig, Franklin M.	10/01/98	The Genetics Adjudication Resource Project	\$1,347,450 36 months	Renewal Under Review	Drell
067126 ER-72	FRED HUTCHINSON CANCER RES CTR Seattle, WA Robbins, Robert J.	09/17/98	Electronic Scholarly Publishing: Foundations of Genetics	\$553,792 36 months	Renewal Under Review	Drell

Count: 2

ID/Org	Proposer/Principal Investigator	Received	Title	Requested	Type/Action	Manager/Detailee
067162 ER-72	LAWRENCE LIVERMORE NATL LAB Livermore, CA Carrano, Anthony	09/17/98	Genomics and Society: Preparing our Nation's Leaders for the 21st Century	0 months	FWP (DOE Under Review	Drell
067160 ER-72	OAK RIDGE NATIONAL LABORATORY Oak Ridge, TN Bjornstad, D. J. Stewart, Steven	09/17/98	An Economic Analysis of Intellectual Property Rights Issues Concerning the Human Genome Program	0 months	FWP (DOE Under Review	Drell
067161 ER-72	OAK RIDGE NATIONAL LABORATORY Oak Ridge, TN Wolfe, Amy K. Cain, Linda C. Melear, Claudia	09/17/98	Educating Educators: An Inquiry-Based, Integrated Approach to the Human Genome Program Ethical, Legal, and Social Implications	0 months	FWP (DOE Under Review	Drell
067163 ER-72	OAK RIDGE NATIONAL LABORATORY Oak Ridge, TN Greeley, Leigh G.	09/17/98	Computer-aided Instruction for Human Consent in Genetic Research	0 months	FWP (DOE Under Review	Drell

Count: 4

*** TX REPORT ***

TRANSMISSION OK

TX/RX NO	2804
CONNECTION TEL	[REDACTED]
SUBADDRESS	
CONNECTION ID	ER 622 GTN
ST. TIME	07/13 08:13
USAGE T	01'22
PGS.	4
RESULT	OK

FAX TRANSMITTAL SHEET
NATIONAL HUMAN GENOME RESEARCH INSTITUTE



National Institutes of Health
Building 38A, Room 605
Bethesda, MD 20892

TO: Lauren Harris, AO, DOE

FAX NUMBER: [REDACTED]

FROM: Jane L. Peterson, Ph.D.

DATE: July 13, 1998

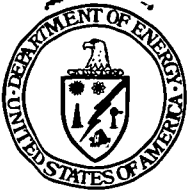
of pages including cover sheet: 4

Return FAX number: [REDACTED]

If there are problems, call [REDACTED]

7/13-Original mailed

Comments: CONFIDENTIAL. PLEASE DELIVER DIRECTLY TO ADDRESSEE.



Department of Energy
Germantown, MD 20874-1290

June 25, 1998

Dr. Jane Peterson
National Institutes of Health
38 Library Drive
MSC 605
Bldg. 38A, Room 614
Bethesda, MD 20892

Jane
Dear Dr. Peterson:

The Department of Energy's appraisal system for senior executive service executives includes feedback from our customers and stakeholders, as well as peers and subordinates. We are now conducting our mid-year progress reviews, and I would appreciate it if you would complete the survey form for **Marvin E. Frazier, Director, Health Effects & Life Sciences Research Division, Office of Biological & Environmental Research, Office of Energy Research**. To assist me in analyzing the feedback, I encourage you to provide additional narrative comments to accompany each rating you select for the ranking elements.

I will use the information gathered in this process in my discussions with **Marvin**. I request that you complete the survey form enclosed and return it to **Lauren Harris, Administrative Officer**, by **July 10, 1998**, to the following address: U. S. Department of Energy, ER-62, 19901 Germantown Road, Germantown, Maryland 20874-1290. You do not need to sign nor put your name on the form. If you are unable to meet this date or have any questions, please contact **Lauren** directly on **301-903-3137**.

Thank you for taking the time to respond to this request. We are making every effort to improve our efficiency and effectiveness in managing the Office of Energy Research. Your assessment of our performance is an important ingredient in our efforts.

Sincerely,

A handwritten signature in black ink that reads "Ari".

Ari Patrinos
Associate Director for Biological
and Environmental Research
Office of Energy Research

Enclosure



**360 DEGREE FEEDBACK SURVEY
FOR**

**Marvin E. Frazier, Director, Health Effects & Life Sciences Research Division,
Office of Biological & Environmental Research, ER**

Scale: Weak 1-----2-----3-----4-----5-----6-----7-----8-----9-----10 Strong			
COMPETENCY	"LEADERSHIP" DESCRIPTION	POINTS	COMMENTS
QUALITY ADVOCACY	<ul style="list-style-type: none"> -Incorporates customer/quality management principles and programs as tools for improving organization. -Implements appropriate process improvements in a timely manner. -Ensures subordinate staff are trained in quality principles and techniques. 	10	<i>I have interacted w/ Dr Frazier monthly as a means to establish & improve communication between</i>
CUSTOMER ORIENTATION	<ul style="list-style-type: none"> -Establishes methods to identify customers, their needs and expectations. -Continuously fosters customer's participation, feedback and satisfaction. -Initiates changes based on customer needs/input. -Meets commitment to customers/clients. 	10	<i>N/H + OBE. Overall, I have found him to be open-minded & extremely eager to make the program succeed.</i>
COMMUNICATIONS	<ul style="list-style-type: none"> -Active listener and fosters an open, candid, and two-way information exchange. -Provides written and oral information in a clear, concise and timely manner. -Establishes criteria to promote communications within the organization and within the organization's customers. 	10	<i>Many of the elements listed are needed N/A as I have no knowledge of Dr. Frazier's work in the office</i>
EXTERNAL RELATIONSHIPS/ NETWORKING	<ul style="list-style-type: none"> -Effectively articulates and promotes the organization's purposes and programs to outside groups. -Meets regularly with clients to foster their cooperation and support. 		<i>although I believe he is well liked & manages the program well. Overall,</i>
NEGOTIATING/ INFLUENCING	<ul style="list-style-type: none"> -Language and behavior promote "win-win" solutions when differences or opportunities between groups/individuals arise. -Makes timely, thoughtful and practical proposals to resolve impasses or reach consensus, taking stakeholders' interests into account. 	10	<i>Dr. Frazier is doing an outstanding job of leading the OBE game program. He has the vision</i>
MANAGING DIVERSITY	<ul style="list-style-type: none"> -Creates atmosphere of equal opportunity as evidenced by training, promotions and job enhancing assignments. -Demonstrates progress in building a diverse workforce. -Develops a plan and continuously monitoring progress toward achieving workforce diversity. 	N/A	<i>needed to ensure success.</i>
INTERPERSONAL RELATIONSHIPS	<ul style="list-style-type: none"> -Conveys respect and trust for clients and employees. -Take positive steps to build trust, morale, and esprit de corps within the organization. 	10	

NOTE: Please give an overall point (1-10) for each Competency by taking into consideration the bullets under "Leadership" Description.

FOR
Marvin E. Frazier, Director, Health Effects & Life Sciences Research Division,
Office of Biological & Environmental Research, ER

Scale: Weak 1-----2-----3-----4-----5-----6-----7-----8-----9-----10 Strong			
COMPETENCY	"LEADERSHIP" DESCRIPTION	POINTS	COMMENTS
HUMAN RESOURCE MANAGEMENT	<ul style="list-style-type: none"> -Effectively allocates resources to meet organizational goals. -Apply resources to achievement organizational priorities. -Adjust resource allocations to meet changing requirements. 	N/A	
TEAM BUILDING/ TEAMWORK	<ul style="list-style-type: none"> -Actively participates in team process. -Values, fosters, and makes constructive contributions to teamwork to improve programs and operations. -Integrates team goals and personal goals. 	N/A	
DEVELOPING TALENT/MANAGING PERFORMANCE	<ul style="list-style-type: none"> -Delegates tasks and empowers organizations to maximize effectiveness. -Serves as coach/mentor. -Recognizes, compliments, and awards achievements of the staff. 	N/A	
STRATEGIC PLANNING	<ul style="list-style-type: none"> -Identifies priorities critical to organizational success. -Identifies appropriate measures for goals and objectives. -Holds self and others accountable for achieving program and mission goals and objectives. 	N/A	
VISION/CHANGE AGENT	<ul style="list-style-type: none"> -Works with others to develop a shared vision of the organization aligned with DOE's mission, vision, and values. -Is forward thinking and encourages new concepts and ideas. 	N/A	
FINANCIAL MANAGEMENT	<ul style="list-style-type: none"> -Effectively budgets and manages fiscal resources. -Acts to avoid or correct potential fraud, waste and abuse. 	N/A	

(Optional) Name and Title of Evaluator: Jane L. Petersen

Signature and Date Jane L. Petersen 3/9/98

NOTE: Please give an overall point (1-10) for each Competency by taking into consideration the bullets under "Leadership" Description

INSTRUCTIONS FOR COMPLETING THE 360 DEGREE SURVEY FORM

- An overall point should be provided for each Competency by taking into consideration the “Leadership” Description.
- The scale numbers to be used are 1 - 10 and have been defined below.
- If you cannot provide an overall point for a particular Competency, please insert N/A.
- Putting your name and signature are optional. The feedback on the individual is more important than knowing who completed the survey.
- The survey should be forwarded to the Administrative Officer noted in the memorandum according this document, or it can be faxed to 301-903-2481.

DEFINING THE 360 DEGREE FEEDBACK RATING SCALE

Scale: Weak 1 -- 2 -- 3 -- 4 -- 5 -- 6 -- 7 -- 8 -- 9 -- 10 Strong

- Please consider the entire scale when rating.
- The scale numbers are meant to be relative.
 - A rating of "1" should be considered to mean that the person being evaluated is performing the competency element but that he/she is perceived by you to be a weak advocate of it and there is little or no evidence of positive outcomes.
 - A "5" or other midscale number, should be used when the person being evaluated is performing the element well but that considerable improvement in advocacy and outcomes could be achieved.
 - A "10" rating should mean that the person being evaluated is proactively performing within the element and that there is little or no room for improvement in outcomes.

It is likely that in an organization like DOE (or ER) that the distribution of ratings would form a bell shaped curve around midscale or lower. However, it is recognized that in some individual instances the average of the ratings will be higher. In general, the lowest and highest numbers should be rarely used.

FAX TRANSMITTAL SHEET
NATIONAL HUMAN GENOME RESEARCH INSTITUTE



National Institutes of Health
Building 38A, Room 605
Bethesda, MD 20892

TO: Lauren Harris, AO, DOE

FAX NUMBER: [REDACTED]

FROM: Jane L. Peterson, Ph.D.

DATE: July 13, 1998

of pages including cover sheet: 4

Return FAX number: [REDACTED]

If there are problems, call [REDACTED]

Comments: CONFIDENTIAL. PLEASE DELIVER DIRECTLY TO ADDRESSEE.

4/17/98 DOE mtg. @ Germantown.

DOE Supplements -

Leudar

Hankins

R. Davis

J. Gamm - Oligonucle

J. Putni

J. Bunker

Setting up adv. group for Quantitative group @ ORNL
SV-

JGI? -

finalize Bldg by early Oct -

move people in late Oct/Nov.

Developing a common process.

Still an issue for submitted

5.5M.

Instrumentation RFA, Fed Request back.

#2-3M.

As it works looking @ data release policy in say before
ask MG. is we can send new NIT policy.

\$2.5M/yr BAC-end sup

Invite Elbert to PI mtg.

Drosophila - might be a bit of

but it's tight next yr.

Need to tell them re: dates

to do $\$5M$

Remain - protein caps groups - May 11-12 $\$3M$

all via
genome +
5th bird
 $\$40M$

$\$15M$ Modeling - study for Lee - 5/12-13 Dec

6/17-18/78 section

$\$12M$ Microbial genomes

density + seq -

Modeling - take adv. of current investment

understand seq as it comes out, e.g. gene

systems.

U.S. Gen. Kern - that do have EST map/BACs.

Peter de Jong - 8th then Glen Evans.

deeply human library w/ another lib. medicine

NCI might need it for.

main res - w/seq 1 use.

12/22/97 DOE Mtg

JASONS - delay to July.
Plan in Jan / Feb.
5 - yr planing

Olson

McPherson

Sum (ETP?)

Cot

Gibbs

JGI

Invite JASON to
discuss there?

Health Effects Restructuring approved -

focus up \$12M - for genome $\rightarrow \frac{1}{2} + \frac{1}{2}$

some in microbial diversity, genome biology
extramuscular $\frac{1}{3}$; Intramuscular $\frac{2}{3}$

Radiation to dose - a comparative genomes. (fly + zebra)

Will do an RFA

JGI 2 regions - 1 on the Chr 5. (growth factor genes)
Chr 19 for fungal proteins.

JGI - RFA review -

Parts of F proposal -

Louder's the lg. project.

Set up review in JGI - met w/ them in Jan -

a second mtg. - decision by 1st of Feb.

fund in March.

Internal RFA for functional genomics < \$500K

- genomic scale + add value to seq.
- provide resources for community.

23 proposals from all Nat'l labs -
pilots / 1 yr. - \$3M

Reviewed externally. (Edey Rubin - leader)

DOE Budget for costs - Morvin will send to us.

Microbial Genes - 5 new organisms.

C5DB review in April - (in NM?)

Instrumentation RFA - "blue skies" - extra-instrumental.
FY 98? funds. < \$5M

Informatics - Model Workshop - 5-yr planning
figure out scope of what DOE should do.
Trying to get details

Mouse - will have a workshop.

Leaders - will have contig limits on size - 300 kb?

Meet again 1/26 + 27.

JGI
Status Report

Elbert Branscomb

BERAC MEETING
WASHINGTON DC
DEC. 16-17, 1997

Purpose and Topics

- Purpose:
 - brief status summary
 - set context for activities of Genome Subcommittee
 - ... and for the advice and help we are seeking from BERAC
- Topics:
 - chronology
 - results of review and advisory panels
 - new organization and goals
 - current status/progress
 - management issues
 - critical strategic problems on which we need advice

Chronology

DATE	LOCATION	EVENT
January 1, 1997		Official start date
January 9		Preliminary plan completed
June 12	Wash DC	HERAC Meeting
July 18		UC Regents authorize PSF lease
July 20		Formal JGI proposal submitted to DOE/OBER
August 20	Wash DC	JGI Reverse Site Visit Peer Review
September 24-25	ILNL	UC President's S&I Panel Presentation
September 30	ILNL	Reorganization plan presented to staff
October 9-10	ILNL	Advisory Board Meeting
October 24	LBNL	PSF lease signed
November 11	Santa Fe	Informal Advisory Board Meeting
December 15		Advisory Board Conference Call

JGI Reverse Site Visit Review: August 20, 1997

Reviewers:

- Charles Cantor, Boston University
- Carol Dahl, National Cancer Institute
- Maynard Olson, University of Washington
- Mark Adams, The Institute for Genomic Research
- Richard McCombie, Cold Springs Harbor Laboratory
- Clark Tibbetts, George Mason University

JGI Advisory Board

David Cox
Stanford University

Ronald W. Davis *
Stanford University Medical
School

David J. Galas *
Darwin Molecular Corp.

Ray Gesteland *
University of Utah

Richard A. Gibbs - CHAIR
Baylor College of Medicine

David Housman
Massachusetts Institute of
Technology

Dave Kingsbury *
Chiron Corp.

David L. Nelson
Baylor College of Medicine

Melvin I. Simon *
California Institute of Technology

Allan C. Spradling
Carnegie Institution of
Washington

Michael Waterman
University of Southern California

* Members of the Kitchen Cabinet

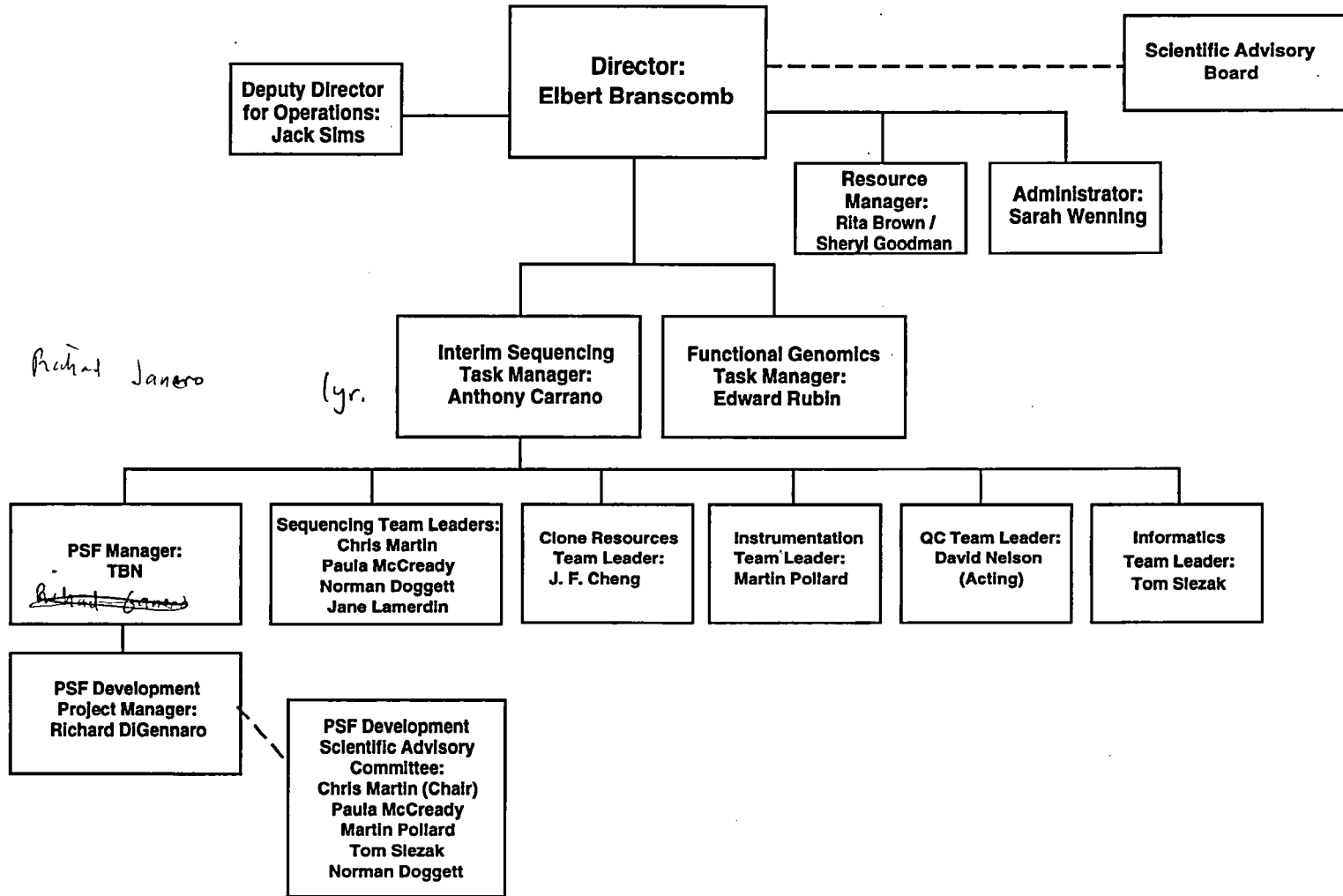
Advisors 10/9/97 Advice

- Focus all efforts on two “must achieve” goals in FY 98:
 - 20 MBb deposited in the public databases
 - a fully operational PSF
- Postpone Functional Genomics component
- Be very conservative in strategy
- Acquire more administrative support for the JGI, and more support for PSF development to ease load on the scientists leading the production effort
- Develop a single sequencing strategy plan by December 1997 based on the conventional M13-based “shotgun” already in place and using plamid-Tn methods as part of the closure phase
- Articulate clearly a central contributing role for LANL

Major Changes Instituted October 1

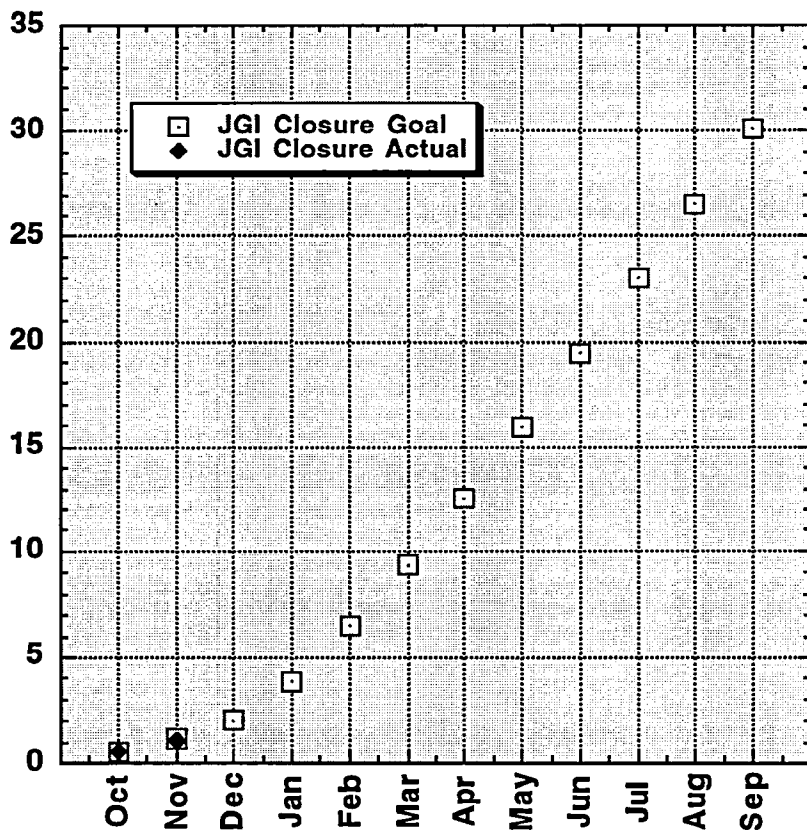
- Reorganization
- Redesign of scientific plan
 - 20MBb in FY98 (2/10/10 -- LANL/LBNL/LLNL)
 - operational PSF
 - postpone Functional Genomics

JGI FY98 ORGANIZATIONAL CHART

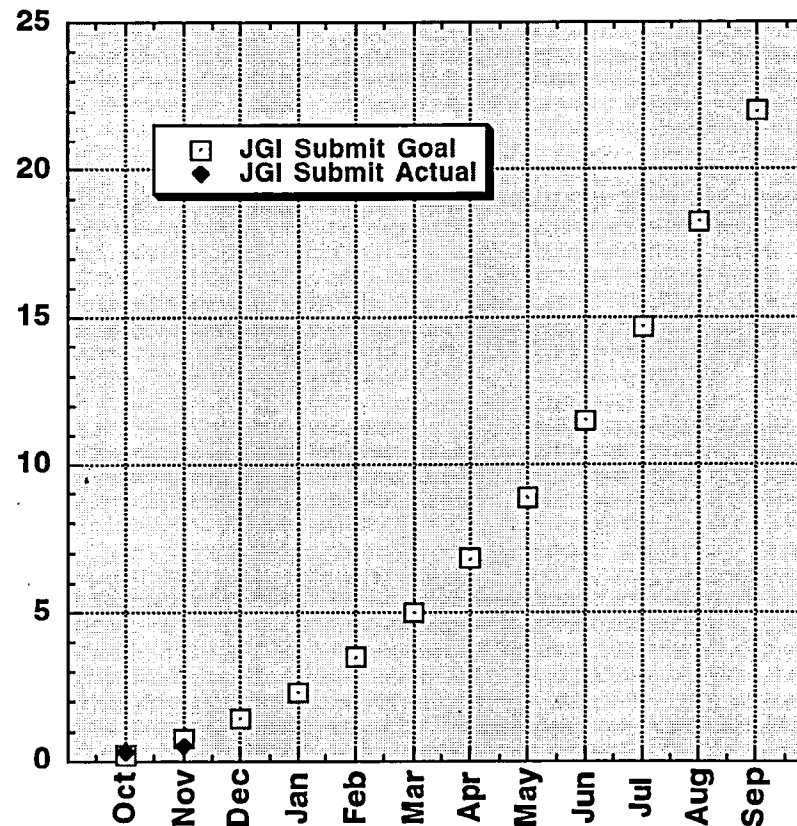


JGI Sequencing Progress in FY98

Megabases closed



Megabases submitted



Quality Summary of Recently Completed Projects

This form displays a summary of recently completed projects. The quality values are generated by phrap (Phil Green, University of Washington).

Notes:

- Phrap values greater than 90 are assumed to be user edits.
- Averages exclude values over 90.

40 ⇒ $\frac{1}{10^4}$

not 95% more than

11 / 97

Clone	Cytog. Band	Region	Acc. Num	Insert (kb)	Unique (kb)	Total Bases	PHRAP VALUES												
							0-9		10-19		20-29		30-39		40-90		>90	Avg.	
F20191	19q13.4	q13.4ZNF		41.6	37.77	41601	95	0.23%	14	0.03%	65	0.16%	127	0.31%	40029	96.22%	1271	3.06%	83.35
R28830	19q13.4	q13.4ZNF		43.1	40.44	43351	0	0.00%	0	0.00%	55	0.13%	330	0.76%	42966	99.11%	0	0.00%	81.71
R28253	19q13.4	q13.4ZNF		41.0	38.9	36807	361	0.98%	119	0.32%	211	0.57%	932	2.53%	34619	94.06%	565	1.54%	76.77
Totals				117.11	121759	456	0.37%	133	0.11%	331	0.27%	1389	1.14%	117614	96.60%	1836	1.51%	81.49	

10 / 97

Clone	Cytog. Band	Region	Acc. Num	Insert (kb)	Unique (kb)	Total Bases	PHRAP VALUES												
							0-9		10-19		20-29		30-39		40-90		>90	Avg.	
BAC33152	19p12	pZNF		165.0	165.0	165819	266	0.16%	191	0.12%	474	0.29%	2026	1.22%	162862	98.22%	0	0.00%	80.53
R34078	19p13.1	OLFR		41.3	40.0	46327	2	0.00%	0	0.00%	187	0.40%	531	1.15%	45570	98.37%	37	0.08%	81.90
R27945	19q13.4	q13.4ZNF		37.4	36.0	37400	33	0.09%	44	0.12%	156	0.42%	353	0.94%	36763	98.30%	51	0.14%	83.02
R33485	19p12	COMP		42.9	40.0	42667	3	0.01%	2	0.00%	52	0.12%	419	0.98%	42191	98.88%	0	0.00%	79.50
Totals				281	292213	304	0.10%	237	0.08%	869	0.30%	3329	1.14%	287386	98.35%	88	0.03%	81.05	

Functional Genomics

- Reduce first year Functional Genomics efforts to:
 - 1-2 Mb comparative mouse sequencing done to Bermuda standards
 - cDNA sequencing to the extent compatible with genome sequencing goals
 - Mouse physical mapping in preparation for increased mouse comparative genomic sequencing in years 2 and later
- Pilot projects in Functional Genomics technology development

3:1
MH : DOE

JGI Financial Management

- The team of Rita Brown and Sheryl Goodman appointed JGI financial managers October 1
 - This team also acts as resource managers for LLNL JGI effort
- Their role is to provide integrated central reporting and monitoring which unifies their LLNL management with that of their counterparts at LANL (Cheryl Straub) and LBNL (Wendell Hom)
- A single account structure unifying all three labs has been established in which production and R&D tasks are separated

JGI Financial Management (continued)

- Cost per base pair can be calculated monthly and cumulatively
- A final budget and staffing plan has been formulated and distributed to all players
- A monthly detailed cost report has been developed that will be used to collect cost data

*U.S. Department of Energy
Office of Energy Research
Biological and Environmental Research Advisory Committee (BERAC) Meeting
December 16-17, 1997
American Geophysical Union
2000 Florida Avenue, N.W.
Washington, D.C. 20009
(202) 462-6900*

Agenda

Tuesday, December 16, 1997

- 8:30 *Welcoming Remarks-Logistics, Agenda, Federal Advisory Committee Act Comments*
Dr. David Thomassen, Ms. Shirley Derflinger, Designated Federal Officers, Office of Biological and Environmental Research, Office of Energy Research
- 8:45 *Introduction of BERAC Members and Guests*
Dr. Keith Hodgson, BERAC Chair, Stanford University
- 9:00 *Remarks to BERAC from Director of Office of Energy Research*
Dr. Martha Krebs, Director, Office of Energy Research (ER)
- 10:00 *Biological and Environmental Research Program Status Report*
Dr. Ari Patrinos, Associate Director, Office of Biological and Environmental Research (OBER), ER
- 10:45 *Break*
- 11:00 *Science Talk: "Opening the Black Box of Soil Microbial Diversity"*
Dr. James Tiedje, Michigan State University
- 12:15 *Lunch*
- 1:45 *Subcommittee Report on the Natural and Accelerated Bioremediation Research Program*
Dr. W. Franklin Harris, University of Tennessee
- 2:15 *Genome Programs Update: Human and Microbial*
Dr. Elbert Branscomb, Director, Department of Energy Joint Genome Institute
Dr. Ari Patrinos, Associate Director, OBER, ER
Dr. Marvin Frazier, Director, Health Effects and Life Sciences Research Division (HELSD), OBER
- 3:00 *Alexander Hollaender Distinguished Postdoctoral Fellowships Briefing*
Dr. Benjamin Barnhart, Program Coordinator, OBER

Gestland - JGI

large scale case

informatics

Jason summer bioint.

user viewpoint → should be focus

$\frac{1}{10^4}$ Lee Hood - $\frac{1}{1000}$ should be ok, + cheaper

informatics - quality problem

lab - failed to use informatics = place to another

**U.S. Department of Energy
Office of Energy Research
Biological and Environmental Research Advisory Committee (BERAC) Meeting
December 16-17, 1997
American Geophysical Union
2000 Florida Avenue, N.W.
Washington, D.C. 20009
(202) 462-6900**

Agenda

Tuesday, December 16, 1997 (Continued)

- 3:15 *Break*
- 3:35 *Subcommittee Report on the Human Genome Program and Joint Genome Institute*
Dr. Raymond Gesteland, University of Utah
- 4:20 *The Washington Advisory Group Report on Atmospheric Radiation Measurement*
Dr. Robert White, National Academy of Engineering
- 4:50 *Public Comment*
- 5:00 *Closing Remarks and Adjourn*
- 6:30 *Evening Function: Reception (Cash Bar) and Dinner (Ruth's Chris Steakhouse)*
Guest Speaker: Dr. Robert Malone, Los Alamos National Laboratory
"Progress & Prospects in High Resolution Ocean Modeling"

Wednesday, December 17, 1997

- 8:30 *Environmental Management Science Program Update*
Dr. Roland Hirsch, Acting Director, Medical Applications and Biophysical Research
Division (MABRD), OBER
- 8:45 *Workshop Reports: "BNCT" and "Isotope Based Medical Research in the Post Genome
Era"*
Dr. Roland Hirsch, Acting Director, MABRD, OBER
Dr. Ludwig Feinendegen, Program Manager, MABRD, OBER
- 9:15 *Subcommittee Report on Structural Biology*
Dr. Jonathan Greer, Abbott Laboratories
- 9:30 *Birgeneau Report, Biosync Report, FASEB Activities*
Dr. Keith Hodgson, BERAC Chair, Stanford University

**U.S. Department of Energy
Office of Energy Research
Biological and Environmental Research Advisory Committee (BERAC) Meeting
December 16-17, 1997
American Geophysical Union
2000 Florida Avenue, N.W.
Washington, D.C. 20009
(202) 462-6900**

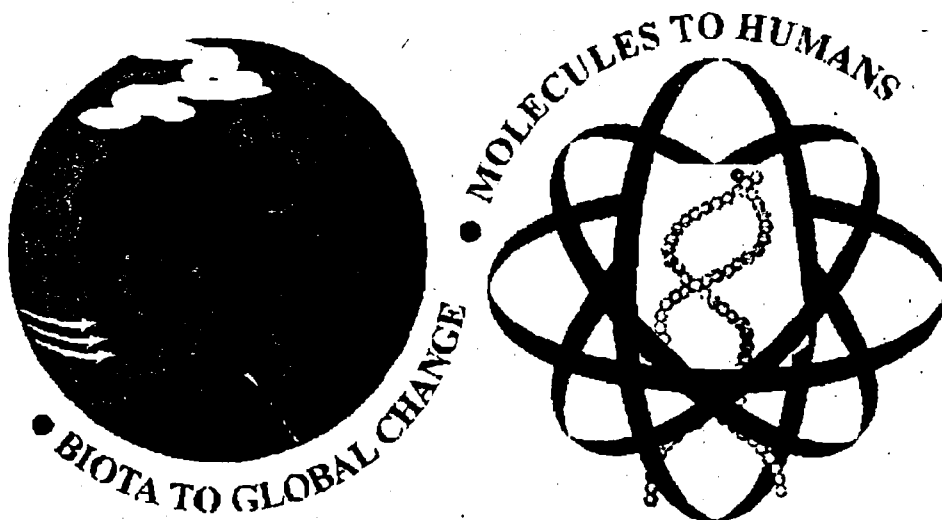
Agenda

Wednesday, December 17, 1997 (Continued)

- 10:00 *Basic Energy Sciences Advisory Committee Synchrotron Report*
Dr. Keith Hodgson, Chair, BERAC
Dr. Jonathan Greer, Abbott Laboratories
- 10:30 *Facilities Roadmap and Program Themes*
Dr. Ari Patrinos, Associate Director, OBER, ER
- 11:15 *Other BERAC Business (Schedule Future Meetings, etc.)*
- 11:50 *Public Comment*
- 12:00 *Closing Remarks and Adjourn*

BERAC MEETING

ARI PATRINOS
OBER



DECEMBER 16-17, 1997

• PEOPLE / ORGANIZATION



• BUDGETS



• PROGRAM ELEMENTS



• MAJOR ISSUES



PEOPLE / ORGANIZATION

- Ben Barnhart
- Jim Beall
- Charles Edmonds
- Michelle Levy
- Lana Pearl
- Joyce Rohek



OBER



New DOE Deputy Secretary: Betsy Moler

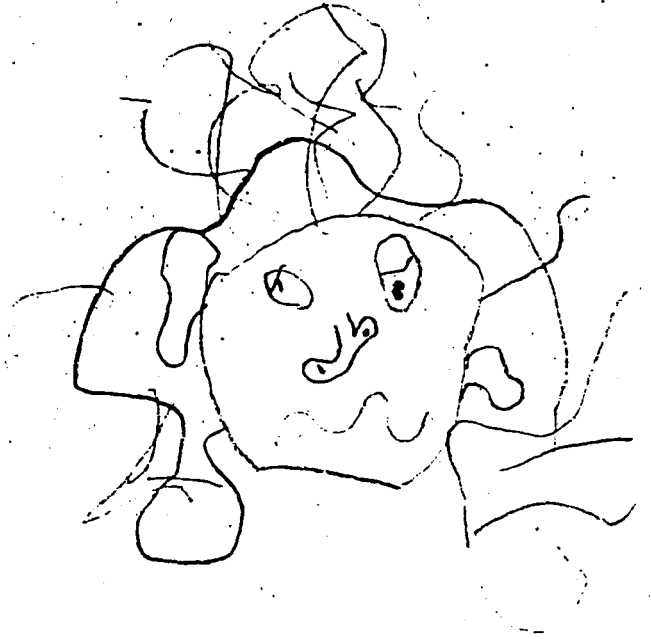
New DOE Undersecretary: Ernie Moniz

OHER → OBER

- Cuts in program direction funds
- Restrictions in support services
- New rules for IPA's & detailees



BUDGETS



- Not as bad as feared
- Many earmarks but with added funds
- Net shortfall : a few \$ M

BER BUDGET SUMMARY

	<u>FY 1997</u>	<u>FY 1998</u>
Life Sciences	\$140,366	\$159,354
Environmental Processes	\$106,968	\$104,457
Environmental Remediation	\$34,229	\$64,036
Medical Applications & Measurement Science	\$55,795	\$63,509
SBIR AND STTR	\$8,602	\$10,151
General Reduction	\$6,702	\$5,203
BNL Reprogramming	\$300	\$0
Total Operating	\$352,962	\$406,710
Construction	\$36,113	\$0
TOTAL BER	\$389,075	\$406,710

	FY 1997 AMOUNT	FY 1997 BUDGET REQUEST	FY 1998 AMOUNT	FY 1998 BUDGET REQUEST
PRESCRIPTION				
CONFERENCE				
INDIANA UNIVERSITY SCHOOL OF MEDICINE	\$3,000			
HEALTH EFFECTS PLAN			\$3,000	\$3,000
UC-DAVIS BNCT			\$4,000	\$0
MCCLELLAN NUCLEAR RADIATION CENTER - BNCT	\$1,000			
MEDICAL U OF S. CAROLINA			\$7,500	\$0
LOMA LINDA MED. CENTER			\$3,000	\$0
ROCHESTER MEDICAL CENTER CANCER			\$3,000	\$0
ENGLEWOOD HOSPITAL, NJ BREAST CANCER			\$2,000	\$0
NORTHEAST REGIONAL CANCER INSTITUTE, MICROBIAL GENETICS			\$10,000	\$0
HIGHLANDS UNIV. LAS VEGAS, NM SCIENCE AND ENGINEERING CENTER			\$2,500	\$0
NIGEC			\$8,200	\$8,200
SUBTOTAL	\$4,000	\$0	\$43,200	\$11,200

PRESCRIPTION	FY 1997 AMOUNT	FY 1997 BUDGET REQUEST	FY 1998 AMOUNT	FY 1998 BUDGET REQUEST
SENATE				
U OF NEVADA LAS VEGAS FOR BIODIVERSITY and INDOOR AIR QUALITY			\$450	\$0
OCEAN SCIENCES	\$6,539	\$6,539		
OREGON HEALTH SCIENCES UNIV.	\$10,000			
UCLA			\$3,930	\$2,780
SUBTOTAL	\$16,539	\$6,539	\$4,380	\$2,780
HOUSE				
NIGEC	\$9,000	\$9,000		
SUBTOTAL	\$9,000	\$9,000	\$0	\$0
TOTAL	\$29,539	\$15,539	\$47,580	\$13,980
NET		\$14,000	\$33,600	
CONGRESSIONAL "plus up"	\$10,000		\$30,000	
NET REDUCTION	\$4,000		\$3,600	

	FY 1997 AMOUNT	FY 1997 BUDGET REQUEST	FY 1998 AMOUNT	FY 1998 BUDGET REQUEST
PRESCRIPTION				
OTHER				
GENERAL REDUCTIONS	\$6,702	\$0	\$5,203	\$0
SBIR and STTR	\$8,602	\$8,602	\$10,151	\$10,151
BNL REPROGRAMMING	\$300			
SUBTOTAL	\$15,604	\$8,602	\$15,354	\$10,151
NET REDUCTION	\$7,002		\$5,203	
TOTAL NET REDUCTION IN RESEARCH	\$11,002		\$8,803	
TOTAL OPERATING	\$352,962		\$406,710	
CONSTRUCTION	\$36,113		\$0	
TOTAL BER APPROPRIATED BUDGET	\$389,075		\$406,710	

HUMAN GENOME PROGRAM

- Year of restructuring and review
- Joint Genome Institute (JGI)
- Sequencing factory
- JASON involvement
- LBNL building



ISSUES OF THE HUMAN GENOME PROGRAM

- Focus on sequencing
- QA/QC
- Data bases / Informatics
- Rigorous peer review / competition
- Clone resources / ELSI
- Functional genomics
- Interactions with NIH

ex. cooperative

5 yr. plan

~~DOE et al~~

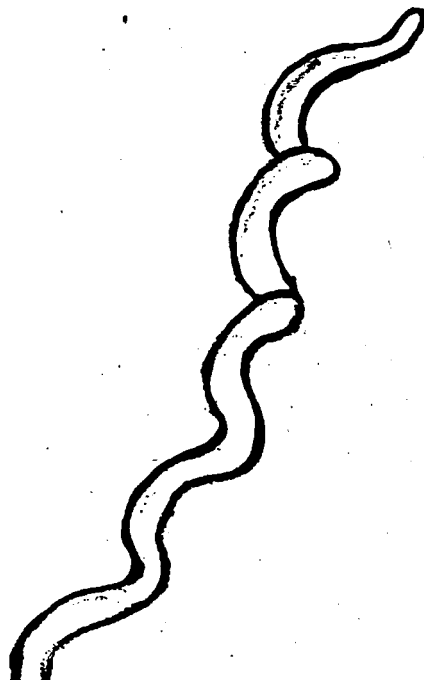
Gesteland
head

need to be coherent (DOE + NIH) about ELSI issues



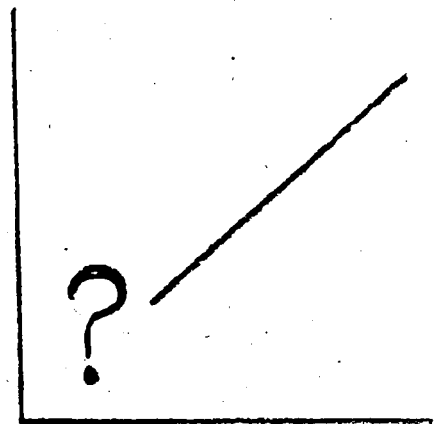
MICROBIAL GENOME PROGRAM

- Continuing success story
- Recent publication in "Nature":
completed sequence of "Archaeoglobus fulgidus"
- New microbes selected for sequencing
- Emphasis on energy & environmental relevance



RESTRUCTURING THE BER HEALTH EFFECTS PROGRAM

- Advice from BERAC
- Workshops
- Letter to Lab Directors
- Elements :
 - Microbial genome
 - Model organisms
 - Low dose, low dose rate exposures
 - Technology applications & development
- Peer review & competition



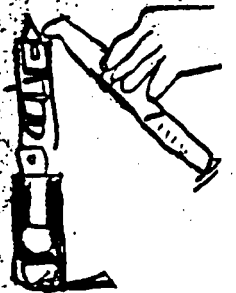
MEDICAL APPLICATIONS & BIOPHYSICAL RESEARCH

- BNCT workshop / clinical aspects



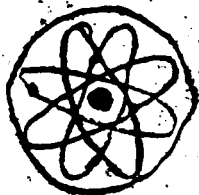
- Workshop with NE & NIH :

Medical research with isotopes / postgenome



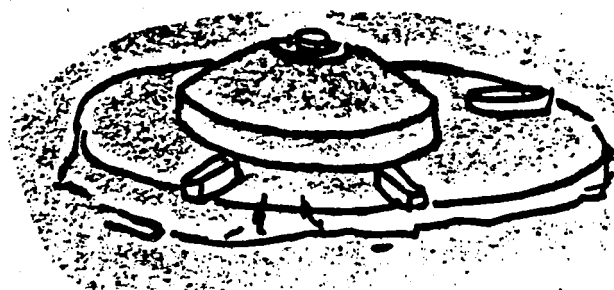
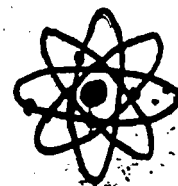
- Environmental Management Science Program :

Another continuing success story



STRUCTURAL BIOLOGY

- High Flux Beam Reactor (HFBR) debacle
- Birgeneau report (BESAC panel)
- FASEB, Biosync
- NIGMS, NCRR, NSF, BER initiative

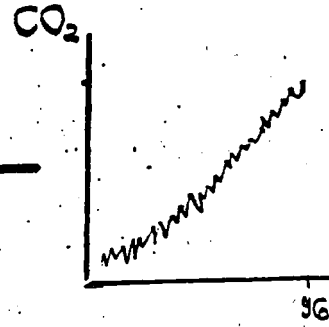


ENVIRONMENTAL MOLECULAR SCIENCES LABORATORY

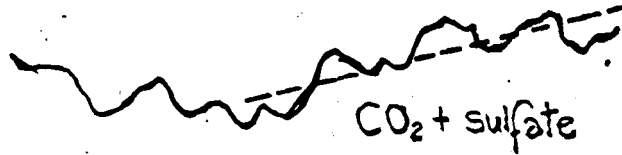
- Fully operational on 10/1/97
- Symposium on molecular sciences for the environment
- Recruiting a new director



GLOBAL CHANGE

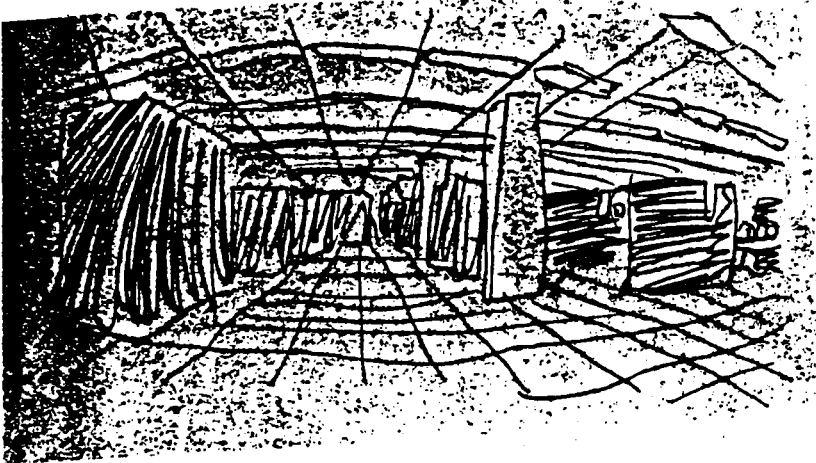


- ARM
- Climate modeling
- National assessment
- Carbon management
- Atmospheric science



CLIMATE MODELING

- Hearings
- OMB/OSTP Request
- Leveraging DP assets



NATIONAL ASSESSMENT

- Mandated by P.L. 101-606 USGCRP Act

- Regions and sectors

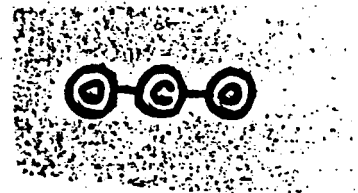


- Workshops and Forum

- Products by 2000

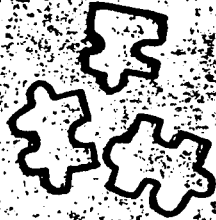
CARBON MANAGEMENT

- Major initiative in FY 99
- Collaborations with BES & DOE Technologies
- Emphasis on carbon cycle, carbon sequestration
- Multilaboratory reports, PCAST
- USGCRP context

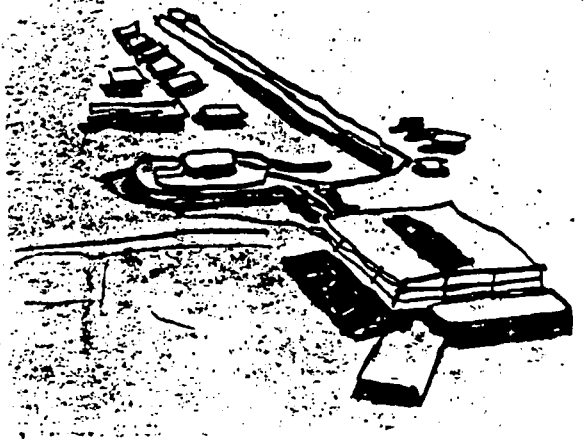


MAJOR ISSUES

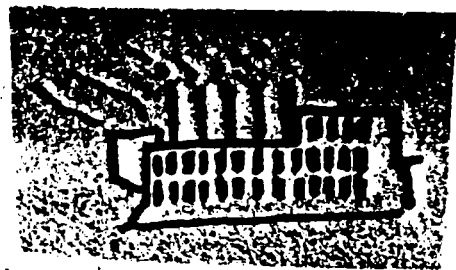
- Post. Kyoto activities



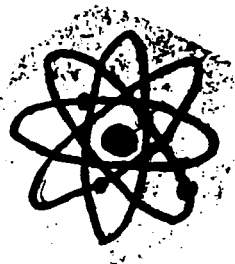
- Spallation Neutron Source



- New BNL Contractor / HFBR



- Comprehensive National Energy Strategy



- Themes & Roadmaps

- GPRA

- New faces on House Approp. Subcommittee



*Man Frasier gives his
report on 10/30.
Jane P*

*Copies to FC
EJ
Circulated to
Program +
Review Staff*

October 11, 1997

To: JGI Members, JGI Advisory Panel Members, DOE Staff

Dear Colleagues

Please find attached a copy of the summary of the advisory committee's review of the JGI plans. This has been circulated by the Panel members, and should be considered a final and accurate representation of the visit.

I would like to add two postscripts: First, it is easy for an external panel to recommend sweeping changes, when they are not the ones who have to actually carry out the plan. The JGI group has certainly earned the respect of the panel so far, and in successfully carrying out further restructuring will continue to do so. Second, the entire panel recognized the degree to which they want this group to succeed. The HGP is depending on it!

Sincerely

Richard A. Gibbs, Ph.D.
On Behalf of the JGI Advisory Committee

members of the panel

Mel Anon

Ron Davis

Roy Gustafson

Daniel Galas

Daniel Nelson

Shirley J. Johnson } - not present

David Hausman }

David Cox

October 9, 1997

JOINT GENOME INSTITUTE ADVISORY PANEL SUMMARY:

The advisory group for the Joint Genome Institutes (JGI) met on Oct. 9 and 10 to discuss the status of human genome sequencing activities at the Lawrence Berkeley Laboratories (LBL), Lawrence Livermore National Laboratories (LLNL) and at the Los Alamos National Laboratories (LANL). The role of the group was to discuss details of the ongoing operations and the restructuring that is being executed in order to form the consolidated JGI.

At the outset it was recognized that the goal of a total 20 Mb production of Bermuda Bases' by the end of the 97-98 fiscal year would remain. The Walnut Creek sequence production facility (PSF) site would not be occupied until after June 1998, so that this 20 Mb would need to be done at the current sites. The proposed split is 10 Mb (LBL), 10 Mb (LLNL) and 2-3 Mb (LANL). While this is going on the PSF design would need to be executed, so that it could be operational very soon after occupancy, and the 1998-1999 goals of 40-50 Mb could be carried out there.

It was clear to the advisory group that many advances in organizational structure and direction setting had been achieved by the group over the last six months. The recent review of the JGI proposal had identified many weaknesses and shown the seriousness of some deficiencies. Some clear messages had been sent. It appeared many of these had been heard, while others still had not. Overall, it appeared that the corner had been turned, but there was still a very tough road ahead.

The combined aims of the highly increased throughput and the design of, and move into the PSF provided a source of conflict that formed a major point of discussion. The panel thought that planning to avoid achieving one goal at the expense of the other was lacking. The group will be measured at the end of year one by its achievements in both areas, and to fail in either would mean overall failure.

The other major point of discussion was the diversity of current sequence strategies, and the urgent need to resolve questions about the use of the LBL- transposon based methods. Enough has been said about these problems in the previous review to serve as an adequate introduction.

The following summarizes specific comments and recommendations of the committee.

COMMENT ON SENIOR STAFF AND LEADERSHIP

Elbert Branscomb: The panel affirmed its confidence in Elbert Branscomb to lead this effort. He is more than the best choice of available talent - he is ideally suited to the task. It was agreed that he urgently needs more administrative support, to deal with the increased demands of his office.

Tony Carrano: The panel was very impressed with Dr. Carrano's effort and role. He is clearly defining the path for the sequencing effort and tackling the important questions. The short term of his appointment is noted and not thought to be overly negative - once the JGI is underway the demands and needs of the sequencing leader will change and it will be appropriate for this position to be reviewed. It is understood there is ongoing recruitment.

GENERAL ADVICE FROM THE PANEL:

1. In order to manage the dual and sometimes conflicting demands of the immediate year one goals of completing 20 Mb of sequence, from the PSF design, the scientific elements of the evolution of sequencing methods should be separated from the demands of the factory design. The model recommended is to use the next three to six months to iron-out all the wrinkles in the sequencing activities at the current sites while achieving the required levels of production. The PSF design can be delegated to a new Project Officer' who will use the advice of the Scientific Staff to make sure the design is appropriate. This advice is particularly directed at Dr. Chris Martin, who is recognized for his key role in bringing the Berkeley process in line with the new unified strategy that is required. At this time he is supposed to oversee design and building of the PSF while producing 10 Mb of finished sequence. He simply cannot do both jobs. However his input in the design and his role once it is built should not be undermined by this change. His high skill level and critical role in defining the Berkeley activities should be the immediate priority.

2. The panel was unanimous on the issue of the urgency with which a single unified sequencing strategy must be found. The PSF occupancy must have only a single strategy to deal with. This means that hard decisions will have to be made in the very near future, regarding the current activities at LBL and LLNL. After discussion and listening to the progress it was clear that the LLNL process was in much more robust health than the LBL process. This was true from many perspectives, however it was not to say that the LBL team do not have enormous talent. Indeed the committee felt that the issue was not whether the LBL effort competed well with the LLNL effort. Instead it was clear that the single opportunity for success in this overall demanding endeavor requires the immediate crafting of a strategy that takes all the talent from the LBL and the LLNL sites and melds it into one focused effort. It is vital that both teams drop issues that relate to their previous identities as separate sequencing groups, and concentrate on this unified task.

To achieve this the committee has specific technical recommendations. First, the LLNL process should remain largely the same, which means that they continue with M13 based shotgun sequencing, and the use of transposon mediated closure

as an experimental tool for speeding the sequence finishing. Their focus should be on streamlining these methods to allow for the new activities in the JGI, and providing coherent data for the LBL team to compare their efforts.

The committee's advice is motivated largely by the concern that the LBL group will carry out 'Business as usual' if left to their own devices. This will be fatal. Reliance on the Tn method must be curtailed and viewed as a potential finishing strategy only. To achieve this the LBL group should immediately increase the shotgun content of their projects, using plasmids as the sequencing template. This will test the use of plasmids in a high throughput fashion that will ask if these are an appropriate substrate for the PSF. The LBL efforts will then be comparable with the tests that the LLNL group is undergoing - with the essential difference of plasmids vs. M13. This should provide the basis of a meaningful comparison in short order. The panel suggested three months to resolve fully the default strategy for the JGI.

The committee recognizes that there are many elements of the sequencing strategy that are not mentioned here that are potentially very influential. For example the transposon efforts in each site are not the same. In addition the LBL group has to bring about some fundamental alterations in data handling in order that the enhanced shotgun effort is not simply supplementing the existing directed strategy, but that the shotgunning is introduced in a way that makes best use of the random data and leaves the transposon methods only when absolutely needed for closure. The panel would like to see all these kinds of factors built into the single unified sequencing plan that is to be complete in three months. Therefore the two groups should be meeting to discuss these, to figure out the critical details and tests of the data, so as to plan the next months activities.

An important aspect of the driving together of the LBL and LLNL strategies is that a unified informatics platform can be built. It is inconceivable that any more than one platform could be created, given the time and resources. The LLNL system is the most robust and well developed, and the committee was very impressed with Tom Slezak as the leader of this plan. He has an excellent opportunity to encourage the LBL team to become enthusiastic about an enhanced shotgun component to their effort, by providing the tools that are needed to use their data in this way. He requires the authority to direct the LBL informatics effort if needed to ensure that the data flow infrastructure is appropriate.

Overall this issue of the need to select a unified strategy as the default to proceed with the PSF design is recognized to be of paramount importance. The group should have an articulated plan completed within three months. This means that preliminary plans and short term tests of critical elements should be planned immediately. The onus is on this team to have this plan articulated properly, and to be able to convince outside reviewers it is viable.

3. The role of Los Alamos in the development of the focused sequencing plan is conspicuously weak. This was discussed at length at the meeting and it was

recognized that the retention of the LANL group in the pre-JGI plan was thought out and justified by various criteria. Nevertheless the rationale is weak at this time. In order for the LANL group to make a contribution to the JGI it must more fully justify its role. Simply finishing the 2-3 Mb that are planned is inadequate. If the LANL group cannot satisfy an external committee that it is an essential, or at least very highly desirable, ingredient of the JGI plan, then support should be phased out. To facilitate the LANL involvement key individuals should be in close communication with the LBL and LLNL members who are planning the unified strategy. It is unfortunate that the LANL group does not have an additional articulated role, such as in mapping or annotation, to bring to this effort, as it is thought that the strengths there may exceed the possible contributions to the sequencing effort.

4. Overall the mapping effort is not in as urgent a need for re-organization as some other elements of the JGI. This is in part due to the reserve of mapped cosmid clones, but also recognizes Dr. Fong's planning and efforts. He has some hard decisions to make about the relative roles of hybridization versus STS mapping by PCR, and fingerprinting. In addition there were obviously critical database issues that need to be addressed. However Drs. Fong and Slezak appeared to know what the requirements were, and The Panel was optimistic that in subsequent visits the mapping would show great progress.

5. Instrumentation planning appeared lacking, although there were assurances that thought had been given to the matter. The problem of an undefined process was noted, which added to the difficulty of the planned automation. In addition there is a call for proposals from external groups to identify some prospective solutions. Nevertheless the group should have a plan in place very soon for both this years solutions, and the PSF. While the LBL group has been given the lead role for this activity, it has yet to show that it is up to this task.

6. In summary, the groups efforts show considerable progress and the next few months will hopefully demonstrate that the faith in this team is well placed. The difficult task of performing in both the short and long term requires some hard decisions now, and a demonstration of the willingness to press these to completion.

10/23/97

Status of libraries?

QA of interim seq - Hord, JGI, McCordie -

Sequencing -

DOE - Stroganov does infrastructure of factory
Mutter putting factory together
Conans in charge of LLNL sequencing
Jane - - finishing @ LLNL
Paula McCordie - seq @ LLNL
Who should come?

JGI - led to 20 MB.

Mouse - 10 To B seq - LLNL + LBL
systemic regions

Had mtg w/ NCI re: mouse -

Klausner, Bob + Carol

met w/ NCI, DOE + H#MI + it

down to discuss what to go forward

re: what to w/ mouse - ESTs, mapping

& seq.

DOE - cDNAs *Drosophila*, (full length) } Comparative
DNA repair genes in *Drosophila* } Genomics
found some work to understand the repair
mutant.

Cytogenetic number resource w/ DOE -

DOE will send sequences - RH has them.

Nov 3 - Israel game wtz

Invite in to Strickers
by pointing.

Nov 7 -

Regene - Mysbale N. not present
Allopathy train

B' team - rest ed del → JGI
Cop → command a JGI
pencil walking - train

JGI - go in w/ 1 approach.
B'ly testing cycle train on 15 BACs
streamlined report -
1st Train. is in lab -

FY98 Goals - 9 B'ly, 2 LANK; 9 LLWL

Drosophila - Comparative genomics -
Neutral effects - DNA repair + cell cycle.

Univ. prep - Oct 16th → Nov 3 - Res. Dec 6th.

LLWL - Tony Brown - consult libr.

follow up w/ Peter / Wook U.

Nov 5th of town sites - Hood.

10
cc: EJ
MG
Sci. Staff

Fakunding, Patricia

From: Francis Collins/D [REDACTED]
Sent: Thursday, June 05, 1997 1:06 PM
To: Fakunding, Patricia
Subject: JASON schedule

Please print out.
FC

To: [REDACTED]
cc: (bcc: Francis Collins/DIR/NCHGR)
From: Ari.Patrinis @ [REDACTED] ("Ari.Patrinis") @ INTERNET
Date: 06/04/97 09:20:00 AM
Subject: JASON schedule

----- Forwarded with Changes

From: [REDACTED]
Date: 6/3/97 12:56PM
To: Ari Patrinis at ER-GTN
***To:** [REDACTED]
Subject: JASON schedule

--


ATT01.txt

Francis:

Below is the draft agenda for the JASON summer study (it was part of a message that Koonin sent to Moniz and me). You may consider having somebody from your Institute attend.

I'll send you the HERAC agenda in a separate e-mail.

Enjoyed our get-together, and stay healthy!

Ari

Draft Agenda (a few names to be added and minor tweaking still likely)

Tuesday July 1 - QA/QC

8:00 AM Motorola

10:00 AM Dick McCombie Cold Spring Harbor Laboratory

[Redacted]

1:00 PM Maynard Olsen University of Washington

[Redacted]

3:00 PM Clark Tibbets George Mason University

[Redacted]

Wednesday July 2 - Technology

8:00 AM Charles Cantor Boston University

[Redacted]

10:00 AM Mike Ramsey Oak Ridge National Laboratory

[Redacted]

1:00 PM George Church Harvard University

[Redacted]

3:00 PM Affymetrix?

Thursday July 3 - Technology

8:00 AM Lloyd Smith University of Wisconsin

[Redacted]

10:00 AM David Allison Oak Ridge National Laboratory

[Redacted]

1:00 PM Dick Keller Los Alamos National Laboratory

[Redacted]

3:00 PM Radomir Crkvenjakov HySeq
[redacted]
crk@sbh.com

Friday July 4

OFF

Monday July 7 - Informatics - software/analysis/databases

8:00 AM overview

10:00 AM Doug Bassett The Johns Hopkins University
[redacted] v

1:00 PM Phil Green University of Washington
[redacted]

3:00 PM David Lipman GenBank
[redacted]

Tuesday July 8 - Informatics - software/analysis/databases

8:00 AM Bob Cottingham Genome Data Base
[redacted] 0

10:00 AM David Searls SmithKline Beecham
[redacted] pm

1:00 PM Peter Schad Genome Sequence Database
[redacted]

3:00 PM Rainer Fuchs Glaxo Wellcome
[redacted].com

Wednesday July 9

"Hands-on" lab work at Scripps (Joyce lab)

Thursday July 10 - Functional genomics

8:00 AM David Galas Darwin Molecular
[redacted]

10:00 AM David Botstein Stanford University
[redacted]

1:00 PM Craig Venter The Institute for Genomic Research
3 [redacted]

3:00 PM Joe Gray University of California - San Francisco

[REDACTED]

Friday July 11 - Functional genomics

8:00 AM Sid Suggs

Amgen

[REDACTED]

9:45 AM Eddy Rubin

Lawrence Berkely National Laboratory

[REDACTED]

11:30 AM National Cancer Institute Initiatives?

1:15 PM Greg Lennon?

Lawrence Livermore National Laboratory

3:00 PM Lon Cardon

Sequana

[REDACTED]

```

*****
** Steven E. Koonin **
** Vice President and Provost/Professor of Theoretical Physics **
** California Institute of Technology **
** **
** e-mail: [REDACTED] [REDACTED] [REDACTED] **
** [REDACTED] [REDACTED] [REDACTED] **
** [REDACTED] **
*****

```

List of Upcoming Genome-Related Events for OHER Staff

<u>Dates</u>	<u>Event</u>	<u>Possible Attendees</u>
6/9	GDB Site Visit, Baltimore	D. Drell, M Frazier
6/11-12	HERAC	All Staff
6/12	Pete Domenici at NAS	All Staff
6/16-17	Arabidopsis Workshop at NSF	M. Stodolsky
6/26-27	Human Subjects Mtg at NLM	S. Rose, D. Drell, D. Thomassen
7/1-11	JASON Genome Summer Study	D. Thomassen, A. Katz
7/15-16	Bioethics Workshop	D. Thomassen, D. Drell, S. Rose
7/24	BAC End Sequencing Workshop	M. Frazier, D. Thomassen, M. Stodolsky, D. Drell, A. Katz
July/Aug??	Microbial Genome Panel Review	All Staff
8/19-20	Genome Project Review <i>JGI, Argonne and Brookhaven local meetings</i>	M. Frazier
9/1-3	Fungal Genome Conference	M. Frazier
9/8	ELSI Panel Review, Ritz-Carlton at Pentagon City	All Staff
Aug	DOE Genomic Informatics Workshop GSDB, GDB, GenBank, Ann. Cons.	Frazier, Drell, Katz
Sept	Hilton Head Sequencing MTG	Frazier, Drell, Stodolsky
Oct <i>Nov/Dec</i>	RFA Genome Sequencing Panel Review	All Staff
11/9-13	Human Genome Contractors Mtg Santa Fe, NM	M. Frazier, D. Drell, M. Stodolsky, D. Thomassen, A. Katz

CDB - 3 yrs - is problem
and not RFA. ✓ on program

Arthur Katz - role in
reming IGE facility
fungal abyd. - worked w/
large group.

Jim Beall - Health effects
aspects of functional genomics

Donald Thomson - right
hand person. no trial,

ford 1st Jan.

6/2/94

Finite DOE - 3?

BAC end seq - continue
@ current level, no extra
digest + focus,
move to new libraries.
Qual - can make good STS.

More telling alt strategy -
pilot studies 1st in
systemic regions - just sel.
regions.

JBI - doing "genome scale"
functional genomics.

Office of Energy Research**Notice 97-10**
Microbial Genome Program

Department of Energy
Office of Energy Research

Energy Research Financial Assistance Program Notice 97-10; Microbial Genome Program

AGENCY: U. S. Department of Energy

ACTION: Notice inviting grant applications

SUMMARY: The Office of Health and Environmental Research (OHER) of the Office of Energy Research, U.S. Department of Energy (DOE), hereby announces its interest in receiving applications for grants in support of the Microbial Genome Program (MGP). The MGP focus is on developing and using high-throughput microbial genome sequencing that will provide functional genomic sequence and mapping information on microorganisms: with environmental or energy relevance; of phylogenetic significance; and of potential commercial importance and application. Bioinformatics tools relating to complete genomic sequences are also of importance to the MGP.

DATES: Preapplications referencing Program Notice 97-10 should be received by March 24, 1997. Formal applications in response to this notice should be received by 4:30 p.m., E.D.T., June 9, 1997, to be accepted for merit review and funding in early FY 1998.

ADDRESSES: Preapplications referencing Program Notice 97-10 should be sent to Dr. Marvin E. Frazier, Office of Health and Environmental Research, ER-72, Office of Energy Research, U.S. Department of Energy, 19901 Germantown Road, Germantown, MD 20874-1290; e-mail is acceptable for submitting preapplications using the following address: iana.ahalt@oer.doe.gov. Formal applications referencing Program Notice 97-10 should be forwarded to: U.S. Department of Energy, Office of Energy Research, Grants and Contracts Division, ER-64, 19901 Germantown Road, Germantown, MD 20874-1290, ATTN: Program Notice 97-10. This address must be used when submitting applications by U.S. Postal Service Express Mail or any commercial mail delivery service, or when hand-carried by the applicant.

FOR FURTHER INFORMATION CONTACT: Dr. Marvin E. Frazier, ER-72, Office of Health and Environmental Research, Office of Energy Research, U.S. Department of Energy, 19901 Germantown Road, Germantown, MD 20874-1290, telephone: (301) 903-5468, e-mail: iana.ahalt@oer.doe.gov.

SUPPLEMENTARY INFORMATION: Molecular biological research on industrially important microorganisms and on microorganisms that live in extreme environments (including the deep subsurface, geothermal environments, hypersaline environments, frozen environments, and toxic waste sites) is a developing area of great scientific promise that will impact many DOE missions, other federal agency missions, and U.S. industry. The Microbial Genome Program supports key DOE business areas by providing microbial DNA sequence information that will further the understanding and application of microbial biology relating to energy production, chemical and materials production, and environmental cleanup. The exploration of microbial genomic sequence diversity is a natural outgrowth of past and current Biological and Environmental Research (BER) Programs, including chromosome mapping and

DNA sequencing from the Human Genome Program, structural biology studies utilizing BER-supported facilities and synchrotrons located at DOE laboratories, and molecular microbiological research supported by BER environmental programs. The MGP benefits directly from capabilities at DOE national laboratories, DOE and National Institutes of Health Human Genome Centers, the DOE and NIH Genome Data Base (GDB), and university capabilities, including the DOE-sponsored Subsurface Microbial Culture Collection and the DOE Genome Sequence Data Base (GSDB). The MGP represents a considerable interdisciplinary effort and will contribute to and draw from a wide variety of public and private programs.

Applications are being sought in three complementary areas: genomic sequencing, functional analysis, and bioinformatics. Each application must clearly state which area is being addressed; if an applicant wishes to address more than one area, the application must clearly describe the expected advantages of an integrated approach.

1. Genomic Sequencing. The DOE intends to continue its support of one or two laboratories that will completely sequence carefully selected microbial genomes. Applicants must demonstrate that they can apply the most recent, high-throughput technology cost-effectively to the production of sequence data and show that they can adequately and efficiently accumulate, store and disseminate those data for future interpretation and application. A commitment to and a plan for making the sequence data publicly available by deposition into an accessible sequence database (GenBank and GSDB) within three months of data acquisition and annotation must be included in the Project Description. Preference will be given to those applicants that demonstrate well developed plans for selecting candidates for DNA sequencing. Candidate microorganisms may include, but are not limited to, bacteria and archaea that mediate or catalyze metabolic events of energy or environmental importance. Strict pathogens or parasites will not be considered. Applicants are encouraged to create process- and cost-effective partnerships that will maximize sequence data production and analysis, data dissemination, and progress towards understanding basic biological mechanisms that can further the development of biotechnology. It is anticipated that one or two major awards will be made to conduct microbial genome sequencing for a total of \$3 to 4 million in FY 1998.

Many microorganisms that are closely related by means of phylogenetic measures (e.g., 16S rRNA comparisons) display dramatic differences in phenotypic characteristics. Such differences can be chromosomal in origin, or they can be due to extrachromosomal genetic elements. DOE is interested in technologies that could exploit the completed sequence of one microorganism to efficiently determine the sequence of a related taxon, without resequencing the entire genome of the related organism *de novo*. New technologies up to the proof-of-principle stage are eligible for support, and it is estimated that between two and four awards for a total of \$500,000 to \$1 million could be available in FY 1998.

2. Functional Analysis. It is presently difficult, and in many instances impossible, to predict biological function from genomic sequence data. Better methods are needed to identify open reading frames and predict their function. This is especially true for environmental isolates and for environmental microorganisms that cannot yet be cultured. Accordingly, applications are requested that will address these and related needs in the area of predicting biological function. It is estimated that between two and four awards for a total of \$1 to 2 million could be available for this area in FY 1998.

3. Bioinformatics. It is estimated that by June, 1997, completed genomic sequences of five or six archaea and bacteria (*Mycoplasma genitalium*, *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum*, *Archaeoglobus fulgidus*, *Pyrococcus furiosus*, and *Aquifex* sp. strain VF5) will be publicly available, as a direct result of DOE Microbial Genome Program funding. In addition, completed

sequences for *Haemophilus influenzae*, *Saccharomyces cerevisiae*, and *Synechocystis* sp. strain PCC6803 are also now publicly available, and by June, 1997, *Escherichia coli*, *Helicobacter pylori*, and *Borrelia burgdorferii* genomic sequences should also be publicly available (all funded by other sources). This unprecedented explosion of genetic information, along with the anticipated increase in other genomic sequence data that will occur over the next year, has underscored the need for better approaches and tools for comparing and analyzing this rapidly increasing volume of data. Accordingly, applications are requested that will propose ways in which data from all databases can be accessed, analyzed, compared, updated, verified, and annotated. It is anticipated that between two and four awards for a total of \$1 to 2 million could be available for this area in FY 1998.

Potential applicants are strongly encouraged to submit a brief preapplication that consists of two to three pages of narrative describing the research objectives and method of accomplishment. Preapplications will be reviewed relative to the scope and research needs of the BER Microbial Genome Program, as outlined in the summary paragraph and in the SUPPLEMENTARY INFORMATION. Principal investigator telephone number, FAX number, and e-mail address are required as part of the preapplication. A response to each preapplication discussing the potential programmatic relevance of a formal application will be communicated to the Principal Investigator within 14 to 21 days of receipt.

It is anticipated that approximately \$7 million will be available for all MGP awards, five to ten awards are anticipated, contingent on availability of appropriated funds in FY 1998. Multiple year funding is expected, also contingent on availability of funds and progress of the research. Previous awards have ranged from \$200,000 to \$2 million per year with terms of one to three years.

Applications will be subjected to formal merit review (peer review) and will be evaluated against the following evaluation criteria which are listed in descending order of importance codified at 10 CFR 605.10(d):

1. Scientific and/or Technical Merit of the Project;
2. Appropriateness of the Proposed Method or Approach;
3. Competency of Applicant's personnel and Adequacy of Proposed Resources;
4. Reasonableness and Appropriateness of the Proposed Budget.

The evaluation will include program policy factors such as the relevance of the proposed research to the terms of the announcement and an agency's programmatic needs. Note, external peer reviewers are selected with regard to both their scientific expertise and the absence of conflict-of-interest issues. Non-federal reviewers will often be used, and submission of an application constitutes agreement that this is acceptable to the investigator(s) and the submitting institution.

The Office of Energy Research (ER), as part of its grant regulations, requires at 10 CFR 605.11(b) that a grantee funded by ER and performing research involving recombinant DNA molecules shall comply with the National Institutes of Health "Guidelines for Research Involving Recombinant DNA Molecules" (51 FR 16958, May 7, 1986), or such later guidelines as may be published in the Federal Register. The Project Description must be 30 pages or less, exclusive of attachments. It must contain an abstract or project summary, letters of intent from collaborators, and short curriculum vitae consistent with NIH guidelines.

To provide a consistent format for the submission, review and solicitation of grant applications submitted under this notice, the preparation and submission of grant applications must follow the guidelines given in the Application Guide for the Office of Energy Research Financial Assistance Program 10 CFR Part 605. Access to ER's Financial Assistance Application Guide is possible via the World Wide Web at: <http://www.er.doe.gov/production/grants/grants.html>.

Other useful web sites include:

MGP Home Page - http://www.er.doe.gov/production/oher/EPR/mig_top.html

GenBank Home Page - <http://www.ncbi.nlm.nih.gov/>

GSDB Home Page - <http://www.ncgr.org/gsdb/>

Human Genome Home Page - <http://www.ornl.gov/hgmis>

The Catalog of Federal Domestic Assistance Number for this program is 81.049, and the solicitation control number is ERFAP 10 CFR Part 605.

John Rodney Clark
Associate Director
for Resource Management
Office of Energy Research

Published in the Federal Register February 19, 1997, Vol. 62, No. 33, pages 7443-7445.

Office of Energy Research

Notice 97-11

Human Genome Program - Ethical, Legal, and Social Implications

Department of Energy
Office of Energy Research

Energy Research Financial Assistance Program Notice 97-11: Human Genome Program - Ethical, Legal, and Social Implications

AGENCY: U.S. Department of Energy

ACTION: Notice inviting grant applications

SUMMARY: The Office of Health and Environmental Research (OHER) of the Office of Energy Research (ER), U.S. Department of Energy (DOE), hereby announces its interest in receiving applications in support of the Ethical, Legal, and Social Implications (ELSI) subprogram of the Human Genome Program (HGP). The HGP is a coordinated, multi disciplinary, directed research effort aimed at obtaining a detailed understanding of the human genome at the molecular level. This particular research notice invites research grants that address ethical, legal, and social implications from the use of information and knowledge resulting from the HGP.

DATES: Preapplications referencing Program Notice 97-11 should be received by April 17, 1997. Formal applications submitted in response to this notice must be received by 4:30 p.m., E.D.T., July 10, 1997, to permit timely consideration for awards in Fiscal Year 1998.

ADDRESSES: Preapplications referencing Program Notice 97-11 should be sent to Dr. Daniel W. Drell, Health Effects and Life Sciences Research Division, ER-72, Office of Health and Environmental Research, Office of Energy Research, U.S. Department of Energy, 19901 Germantown Road, Germantown, MD 20874-1290. Formal applications referencing Program Notice 97-11 should be forwarded to: U.S. Department of Energy, Office of Energy Research, Grants and Contracts Division, ER-64, 19901 Germantown Road, Germantown, MD, 20874-1290, ATTN: Program Notice 97-11. This address also must be used when submitting applications by U.S. Postal Service Express Mail or any commercial mail delivery service, or when hand carried by the applicant.

FOR FURTHER INFORMATION CONTACT: Dr. Daniel W. Drell, Health Effects and Life Sciences Research Division, ER-72, Office of Health and Environmental Research, Office of Energy Research, U.S. Department of Energy, 19901 Germantown Road, Germantown, MD 20874-1290, phone: (301) 903-6488 or E-mail: daniel.drell@oer.doe.gov.

SUPPLEMENTARY INFORMATION: The DOE encourages the submission of applications that will address, analyze, or anticipate ELSI issues arising from advances in the scientific understanding of genetically influenced susceptibilities/sensitivities, complex or multi-genic characteristics and conditions, and human polymorphisms. This may include research on privacy and confidentiality issues (as well as ownership and commercialization issues) arising from the creation, use, maintenance, and disclosure of genetic information relevant to such complex or multi-genic conditions. This may also include research on the privacy implications of the development of HGP materials, resources, databases and technologies, as well as the privacy implications of the use of genetic information obtained in the workplace. Issues to be

examined may also include (but are not limited to) implications of advances in the genetic characterization of complex traits and susceptibility/sensitivity genes and the impacts of advances in knowledge about polygenic conditions for individuals and communities potentially faced with these impacts (e.g. courts, schools, etc).

All applications should demonstrate knowledge of the relevant literature, and should include detailed plans for the gathering and analysis of factual information and the associated ethical, legal, and social implications. All applications should include, where appropriate, detailed discussion of human subjects protection issues; e.g., storage of, manipulation of, and access to data. Provisions to ensure the inclusion of women, minorities, and potentially disabled individuals must be described, unless specific exclusions are scientifically necessary and justified in detail. All proposed research applications should address the issue of efficient dissemination of results to the widest appropriate audience. All applications should include letters of agreement to collaborate from potential collaborators; these letters should specify the contributions the collaborators intend to make if the application is accepted and funded.

The DOE also solicits applications for the preparation and dissemination of educational materials in any appropriate medium that will enhance understanding of the ethical, legal, and social aspects of the HGP among the public or specified groups; a particular interest of this notice is Institutional Review Boards (IRBs) and genome investigators who work with patients. This may include (but is not limited to) implications of disease predispositions, susceptibility genes, increased knowledge of polygenic conditions, informed consent issues or Human Genome Project materials- and resources-development and dissemination projects (e.g. the creation of a human DNA library, etc.). If an educational effort for a specific group is proposed, the value to the Human Genome Program of that group or community should be explained in detail. In addition, the DOE encourages applications for the support of novel and innovative conferences focusing on the concerns addressed in this notice (e.g. susceptibility/sensitivity genes, polymorphisms, and education of IRBs and investigators).

Educational and conference applications should demonstrate awareness of the relevant literature, and include detailed plans for the accomplishment of project goals. In applications that propose the production of series for broadcast, audio-visuals or other educational materials, the DOE requests that samples of previous similar work by the producers and writers be submitted along with the application. In applications for the support of educational activities, the DOE requests inclusion of a plan for assessment of the effectiveness of the proposed activities. For conference applications, a detailed and largely complete roster of speakers is necessary. At the completion of the conference, a summary or report is required. Educational and conference applications must also demonstrate awareness of the need to reach the widest appropriate audience, and not be focused exclusively on a local community or group.

Possible outcomes of these research and/or educational efforts may include (but are not limited to): model guidelines for research practices for studies of polygenic conditions and susceptibility genes; consensus documents on implications or significance of the genetic bases for complex conditions; privacy and confidentiality studies of genetic information pertinent to complex conditions; model policies for genetic information about polygenic conditions for various settings (e.g. the workplace); exploration of worker/workplace issues; and materials for IRBs.

In all applications, a clear description of expected products or "deliverables" should be included, as well as a time line for their production and dissemination. In the absence of tangible products, rigorous assessments must be included to facilitate evaluation of progress.

DOE does not encourage applications dealing with issues consequent to the initiation or implementation of genetic testing protocols. Also, DOE does not encourage survey-based research, unless a compelling

case is made that this methodology is critical to address an issue of uncommon significance. For applications which propose the development of college-level curricula, DOE requests both detailed justification of the need for external support, beyond normal departmental and college resources, evidence of commitment from the parent department or college, and a dissemination plan. Applications for the writing of scholarly publications or books should include justifications for the relevance of the publications or book to the goals of the Human Genome Project as well as discussion of the estimated readership and impact. DOE ordinarily will not provide unlimited support for a funded program and thus strongly encourages the inclusion of plans for transition to self-sustaining status.

The dissemination of materials and research data in a timely manner is essential for progress towards the goals of the DOE Human Genome Program. The OHER requires the timely sharing of resources and data. Applicants should, in their applications, discuss their plans for disseminating research results and materials that may include, where appropriate, publication in the open literature, wide-scale mailings, etc. Once OHER and the applicant have agreed upon a distribution plan, it will become part of the award conditions. Funds to defray the costs of disseminating results and materials are allowable; however, such requests must be sufficiently detailed and adequately justified. Applicants should also provide timelines projecting progress toward achieving proposed goals.

Potential applicants are strongly encouraged to submit a brief preapplication that consists of two to three pages of narrative describing the research project objectives and methods of accomplishment. These will be reviewed relative to the scope and research needs of the DOE's Human Genome Program. Principal investigator address, telephone number, FAX number and E-mail address are required parts of the preapplication. A response to each preapplication discussing the potential program relevance of a formal application generally will be communicated within 20 days of receipt. ER's preapplication policy for submitting preapplications can be found on ER's Grants and Contracts Web Site at: <http://www.er.doe.gov/production/grants/preapp.html>.

It is anticipated that approximately \$1,500,000 will be available for grant awards in this area during FY 1998, contingent upon availability of appropriated funds. Multiple year funding of grant awards is expected, and is also contingent upon availability of funds. Previous awards have ranged from \$50,000 per year up to \$500,000 per year with terms from one to three years; most awards average about \$200,000 per year for two or three years. Similar award sizes are anticipated for new grants.

Applications will be subjected to formal merit review (peer review) and will be evaluated against the following evaluation criteria which are listed in descending order of importance codified at 10 CFR 605.10(d):

1. Scientific and/or Technical Merit of the Project;
2. Appropriateness of the Proposed Method or Approach;
3. Competency of Applicant's personnel and Adequacy of Proposed Resources;
4. Reasonableness and Appropriateness of the Proposed Budget.

The evaluation will include program policy factors such as the relevance of the proposed research to the terms of the announcement and an agency's programmatic needs. Note, external peer reviewers are selected with regard to both their scientific expertise and the absence of conflict-of-interest issues. Non-federal reviewers will often be used, and submission of an application constitutes agreement that this is acceptable to the investigator(s) and the submitting institution.

To provide a consistent format for the submission, review and solicitation of grant applications submitted under this notice, the preparation and submission of grant applications must follow the guidelines given in the Application Guide for the Office of Energy Research Financial Assistance Program 10 CFR Part 605. Access to ER's Financial Assistance Application Guide is possible via the World Wide Web at: <http://www.er.doe.gov/production/grants/grants.html>.

DOE policy requires that potential applicants adhere to 10 CFR 745 "Protection of Human Subjects", or such later revision of those guidelines as may be published in the Federal Register.

The Catalog of Federal Domestic Assistance Number for this program is 81.049, and the solicitation control number is ERFAP 10 CFR Part 605.

John Rodney Clark Associate Director for Resource Management Office of Energy Research

Published in the Federal Register March 3, 1997, Vol. 62, No. 41, pages 9419-9420.

[6450-01-P]

**Department of Energy
Office of Energy Research**

Energy Research Financial Assistance Program Notice 97-17; Human Genome Program - Technologies in support of the DOE Joint Genome Institute

AGENCY: U.S. Department of Energy

ACTION: Notice inviting grant applications

SUMMARY: The Office of Health and Environmental Research (OHER) of the Office of Energy Research (ER), U.S. Department of Energy (DOE), hereby announces its interest in receiving applications for Special Research Grants in support of the Human Genome Program. This Program is a coordinated multidisciplinary research effort to develop creative, innovative resources and technologies that lead to a molecular level understanding of the human genome. As one aspect of this program, the DOE is establishing a "Joint Genome Institute" (JGI) a DNA sequencing factory. The JGI will oversee a central sequencing facility that will initially have parallel production lines that use shotgun and transposon-based directed sequencing approaches. This dual approach is intended to evolve into an optimized and unified sequencing strategy within two to three years. This unified strategy will take advantage of technologies and expertise at the JGI and in the broader research community. An important aspect of developing this automated facility will be the establishment of external collaborations and partnerships aimed at technology development. The JGI's genomic sequencing program will also be coupled to a collection of experimental functional genomics approaches designed to provide a partial functional characterization of the genes as they are revealed by the sequencing. Here, the primary goal will be to develop cost-effective approaches that can yield worthwhile functional information. A related goal is to develop improved ways of integrating human genomics with the information coming from model organism genomics.

DATES: Preapplications referencing Program Notice 97-17 should be received by August 15, 1997. Formal applications in response to this notice must be received by 4:30 p.m., E.D.T., October 23, 1997, to be accepted for merit review and to permit timely consideration for award in FY 1998.

ADDRESSES: Preapplications referencing Program Notice 97-17 should be sent to Dr. Marvin E. Frazier, Office of Health and Environmental Research, ER-72, Office of Energy Research, U.S. Department of Energy, 19901 Germantown Road, Germantown, MD 20874-1290; e-mail is acceptable for submitting preapplications using the following address: joanne.corcoran@oer.doe.gov. Formal applications referencing Program Notice 97-17 should be forwarded to: U.S. Department of Energy, Office of Energy Research, Grants and Contracts Division, ER-64, 19901 Germantown Road, Germantown, MD 20874-1290, ATTN: Program Notice 97-17. This address must be

used when submitting applications by U.S. Postal Service Express Mail or any commercial mail delivery service, or when hand-carried by the applicant. An original and seven copies of the application must be submitted; however, applicants are requested not to submit multiple application copies using more than one delivery or mail service.

FOR FURTHER INFORMATION CONTACT: Dr. Marvin E. Frazier, ER-72, Office of Health and Environmental Research, Office of Energy Research, U. S. Department of Energy, 19901 Germantown Road, Germantown, MD 20874-1290, telephone: (301) 903-6488, e-mail: joanne.corcoran@oer.doe.gov.

SUPPLEMENTARY INFORMATION: The goal of this notice is to support technology development that serves the needs of the JGI. With respect to the JGI genomic sequencing task, the specific goals are: (1) to establish a cooperative technology development project with the JGI that will produce, within two years, an automated DNA sequencing production line based on either shotgun or directed strategies; and (2) to develop and implement technologies for automated and advanced high-throughput DNA sequencing that can be integrated into the unified sequencing production strategy that is identified and implemented at the JGI.

In support of the first goal, the grantee will form a close collaboration with the JGI aimed at technology co-development and transfer for high throughput production DNA sequencing. A critical success factor for this effort will be the construction of a new, highly automated pilot DNA sequencing production line at the JGI within 6 to 9 months of the project's start. The grantee, working in conjunction with the JGI, will help build and maintain automated devices as appropriate for this pilot line (e.g., those for DNA purification, DNA sequencing, and automated finishing). It is anticipated that this pilot DNA sequencing production line may use, in significant part, technology supplied by the grantee. The second phase of the project, to be completed within two years, will be the development of a high throughput DNA sequencing production line. It is anticipated that this production line will lead current technology in automation and the minimization of human labor and will ultimately produce 100-200 Mb of finished human genomic sequence per year. It is also expected that, in close cooperation with the JGI, the grantee will use the technology being supplied to perform a significant amount of DNA sequencing on targets that support the DOE effort. This would be designed to drive the technology development and to permit modifications in technology between the pilot and production phases to be evaluated and validated under high throughput conditions. It is estimated that one major award, for a total of approximately \$4 million in FY 1998, will be made.

In support of the second sequencing goal, technology developments aimed at improving the constituent technologies and overall performance of the JGI DNA sequencing production line are sought. These could include: innovative instrumentation and automated systems that offer the potential for rapid, cost-effective sequencing of approximately a million bases per day; for non-gel techniques and direct imaging approaches; for development of applied genome informatics software for use in

DNA sequencing and functional interpretation, including information retrieval; for user interfaces compatible with Genome Data Base (GDB), Genome Sequence DataBase (GSDB), and GenBank; and for communications, software engineering, and data management. Improved algorithms and hardware for DNA sequence annotation, including identification of homologies, regulatory sites, and protein coding regions can also be included. It is anticipated that between 2-4 awards for a total of up to \$1 million could be made in FY 1998.

With respect to the functional genomics and model organism goals, projects in the following program areas are solicited: 1) strategies for full-length cDNA clone generation and sequencing and for economically and accurately determining transcript lengths and types; 2) strategies for expression mapping, sub-cellular localization, and pathway tracing; 3) economical approaches for revealing single base pair polymorphisms and for characterizing their haplotypes; and 4) affordable approaches for using model organisms to systematically relate phenotype information to anonymous genes discovered in the human genome. It is anticipated that between 2-4 awards for pilot and proof-of-principle studies, for a total of up to \$1 million could be made in FY 1998.

Potential applicants are strongly encouraged to submit a brief preapplication that consists of two to three pages of narrative describing the research objectives and methods of accomplishment. Preapplications will be reviewed relative to the scope and research needs of the DOE Human Genome Program, as outlined in the summary paragraph and in the SUPPLEMENTARY INFORMATION. Principal investigator address, telephone number, FAX number, and e-mail address are required as part of the preapplication. A response to each preapplication discussing the potential programmatic relevance of a formal application generally will be communicated to the Principal Investigator within 21 days of receipt. ER's preapplication policy can be found on ER's Grants and Contracts Web Site at:

<http://www.er.doe.gov/production/grants/preapp.html>.

It is anticipated that approximately \$6 million will be available for grant awards during FY 1998, contingent upon availability of appropriated funds. Multiple year funding of grant awards is expected, with out-year funding also contingent upon the availability of appropriated funds, progress of the research, and programmatic needs. It is expected that most awards will be from one to three years and that there will be one award for approximately \$4 million per year (total costs) with the remaining 4-6 awards in the \$200 thousand to \$400 thousand per year (total costs) range.

Applications will be subjected to formal merit review (peer review) and will be evaluated against the following evaluation criteria which are listed in descending order of importance codified at 10 CFR 605.10(d):

1. Scientific and/or Technical Merit of the Project;
2. Appropriateness of the Proposed Method or Approach;
3. Competency of Applicant's personnel and Adequacy of Proposed Resources;
4. Reasonableness and Appropriateness of the Proposed Budget.

The evaluation will include program policy factors such as the relevance of the proposed research to the terms of the announcement and an agency's programmatic needs. Note, external peer reviewers are selected with regard to both their scientific expertise and the absence of conflict-of-interest issues. Non-federal reviewers will often be used, and submission of an application constitutes agreement that this is acceptable to the investigator(s) and the submitting institution.

Information about development and submission of applications, eligibility, limitations, evaluation, selection process, and other policies and procedures may be found in the ER Application Guide for the Office of Energy Research Financial Assistance Program 10 CFR Part 605, which is available on the World Wide Web at: <http://www.er.doe.gov/production/grants/grants.html>. The ER, as part of its grant regulations, requires at 10 CFR 605.11(b) that a grantee funded by ER and performing research involving recombinant DNA molecules and/or organisms and viruses containing recombinant DNA molecules shall comply with the National Institutes of Health "Guidelines for Research Involving Recombinant DNA Molecules" (51 FR 16958, May 7, 1986), or such later revision of those guidelines as may be published in the Federal Register. The dissemination of materials and research data in a timely manner is essential for progress towards the goals of the DOE Human Genome Program. OHER requires the timely sharing of resources and data. Applicants should, in their applications, discuss their plans for disseminating research data and materials which may include, where appropriate, putting cell lines, probes, sequence data, etc., into public repositories. Funds to defray the costs of disseminating materials or submitting data to repositories are allowable; however, such requests must be adequately justified.

The Catalog of Federal Domestic Assistance Number for this program is 81.049 and the solicitation control number is ERFAP 10 CFR Part 605.

Issued in Washington, DC on _____.

John Rodney Clark
Associate Director
for Resource Management
Office of Energy Research

cc:Mail for: Marvin Frazier

Subject: JASON genome summer study - July 1-11, 1997

From: David Thomassen 05/30/1997 12:44 PM



Thank you all for agreeing to participate in what should be a very interesting and enjoyable series of discussions with the JASONS on aspects of the human genome project. As a starting point I will give a very brief overview of the JASONS, the goals of this summer's study, and previous interactions the JASONS have had on this study. I am also providing a draft agenda that still has one or two holes in it that will, hopefully, be filled within the next few days. As I indicated to each of you on the telephone, the goal of your presentations should be (1) to provide information in your area of expertise, (2) to identify opportunities, needs, issues, problems, concerns, etc that affect the community's ability to get the most out of the genome project, and (3) to expect an interactive and wide ranging discussion with the JASON genome study group.

You will be contacted by Email by the JASON program office with logistical details for your participation in this study. You might go

[REDACTED]; cbennett@mitre.org) or Diane Huth ([REDACTED])

[REDACTED] and your time.

David Thomassen

Phone: [REDACTED]

Fax: [REDACTED]

Email: [REDACTED]

Broad study goals:

Informatics

What are the current database issues, including data integrity, submission, and usability? What is the current state of algorithm development for finishing and annotating sequence?

- database management issues (pro/con)
- annotation (how, issues, etc.)
- assembly programs (algorithms)

QA/QC in the genome project

What levels of sequence quality are required by various users of genome data and what steps can be taken to ensure various levels of quality? What are the "dust to dust" QA/QC issues/needs in the genome project?

- how are we going to implement this?

Technology

What alternatives are there for DNA sequencing? What strategies should be used for inserting new technologies into production sequencing? What are the broader uses of these technologies? What are the technology needs in this area beyond those of the genome project?

- state of the art and unorthodox approaches

Functional genomics

- model organisms

- biotech applications
- end users of genomic information

The JASONS

The JASONS are an elite group of physicists, mathematicians, computer scientists, and, recently, biologists, who have interacted with the Department of Energy for more than 30 years on a variety of topics including nuclear stewardship, arms control, climate change, atmospheric radiation measurements, and the computer hardware, advanced mathematics and model physics component of the global climate change program. This year they were very interested in getting involved with the the genome program. Typically, the JASONS study a topic during their winter and summer meetings and then issue a report to the appropriate assistant secretary in the department for which they are doing the study, in this case Dr. Martha Krebs. The JASONS genome study group is chaired by Steve Koonin (provost of Cal Tech) and are currently chaired overall by Bill Press (Harvard).

During January winter study, the JASONS met with Lee Hood, Francis Collins, Eric Lander, Ron Davis, Mel Simon, Elbert Branscomb, and Mike Palazzolo for an overview of genomics and the US and DOE genome programs. This spring they visited three genome centers at the Whitehead Institute, Washington University, and Lawrence Berkeley National Laboratory. This summer they will focus on QA/QC issues associated with high throughput sequencing, data management, and data analysis; informatics, including databases and data analysis; technology development for sequencing; and functional genomics.

Draft Agenda (a few names to be added and minor tweaking still likely)

Tuesday July 1 - QA/QC

8:00 AM Motorola

10:00 AM Dick McCombie Cold Spring Harbor Laboratory

[REDACTED]

1:00 PM Maynard Olsen University of Washington
206-685-7366
mvo@u.washington.edu

3:00 PM Clark Tibbets George Mason University

[REDACTED]

Wednesday July 2 - Technology

8:00 AM Charles Cantor Boston University

[REDACTED]

10:00 AM Mike Ramsey Oak Ridge National Laboratory

[REDACTED]

1:00 PM George Church Harvard [REDACTED] u

3:00 PM Affymetrix?

Thursday July 3 - Technology

8:00 AM Lloyd Smith University of Wisconsin
[REDACTED]

10:00 AM David Allison Oak Ridge National Laboratory
[REDACTED] gov

1:00 PM Dick Keller Los Alamos National Laboratory
[REDACTED]

3:00 PM Radomir Crkvenjakov HySeq
[REDACTED]

Friday July 4

OFF

Monday July 7 - Informatics - software/analysis/databases

8:00 AM overview

10:00 AM Doug Bassett The Johns Hopkins University
[REDACTED]

1:00 PM Phil Green University of Washington
[REDACTED]

3:00 PM David Lipman GenBank
[REDACTED]

Tuesday July 8 - Informatics - software/analysis/databases

8:00 AM Bob Cottingham Genome Data Base
[REDACTED]

10:00 AM David Searls
[REDACTED]

1:00 PM Peter Schad Genome Sequence Database
[REDACTED]

3:00 PM Rainer Fuchs Glaxo Wellcome
[REDACTED].com

Wednesday July 9

"Hands-on" lab work at Scripps (Joyce lab)

Thursday July 10 - Functional genomics

8:00 AM David Galas Darwin Molecular
[REDACTED]

10:00 AM David Botstein Stanford University
[REDACTED]

1:00 PM Craig Venter The Institute for Genomic Research
301-838-3500
jcventer@tigr.org

3:00 PM Joe Gray University of California - San Francisco
[REDACTED]

Friday July 11 - Functional genomics

8:00 AM Sid Suggs Amgen
[REDACTED]

9:45 AM Eddy Rubin Lawrence Berkely National Laboratory
[REDACTED]

11:30 AM National Cancer Institute Initiatives?

1:15 PM Greg Lennon? Lawrence Livermore National Laboratory

3:00 PM Lon Cardon Sequana
[REDACTED]

Health and Environmental Research Advisory Committee (HERAC)

Subcommittee Recommendation on Future Directions for the OHER Health Effects Research Program

June 6, 1997

MISSION

To identify critical research opportunities, bottlenecks, and needs for emerging "new biology" and recommend opportunities in keeping with the objectives of the OHER Health Effects Research Program.

INTRODUCTION (Leroy Hood)

A series of paradigm changes over the last five to ten years have profoundly changed the framework of biology, medicine and environmental sciences.

- **Biology is an Information Science.** There are three general types of biological information: the one-dimensional (1°) or digital information of DNA; the three-dimensional (3°) information of proteins; and the four-dimensional (4°) information of complex biological systems and networks. Major biological challenges are associated with each of these types of information: 1° - defining the gene and regulatory components provides fundamental insights into development and triggering by environmental signals; 2° - the protein folding problem and how does the 3° shape of individual proteins permit them to execute their functions; and 4° - how do we define the elements of systems, their inter connections, and come to understand their systems or emergent properties. There are two alternative modes for deciphering each type of biological information: 1° - determine the genome sequence (the objective of the Human Genome Project) and decipher the biological information 3.7 billion years of evolution has inscribed on our chromosomes; 3° - characterize the shapes of proteins and determine how their shape permits them to execute function; and 4° decipher the elements and connections of systems verses understanding how these lead to systems properties. Computer scientists and applied mathematicians will play a critical role in deciphering biological information.

DRAFT

- **High-Throughput Tools Are Changing Our Fundamental Approaches to Biology.** The advent of tools such as large-scale DNA sequencing, genome-wide genotyping, high-density DNA arrays, high-throughput mass spectrometry for the analysis of proteins, high speed fluorescence activated cell sorting, and the computational tools associated with the acquisition, storage, analysis, modeling, and distribution of biological data--have had a profound impact on biology and medicine. The development of these tools and their next generation prototypes reflects the urgent need for bringing to biology the leading-edge technologies of applied mathematics, applied physics, chemistry, computer science, and engineering. How to facilitate scientists from different disciplines to learn to understand the languages of other disciplines is one of the major challenges of contemporary science.
- **The Human Genome Project is Defining the Periodic Table of Biological Elements at the DNA Level.** The objective of the Human Genome project is to map and sequence the human genome and the genomes of five model organisms: *E coli*, yeast, *Drosophila*, nematode, and mouse. This information as illustrated by the complete genome sequence of yeast, will facilitate the identification of all genes, regulatory regions, and the many other functions of chromosomes (e.g. many sites of DNA replication, sites of protein/DNA interactions, etc.) The sequence of the human (and other) genome already opens the possibility of large-scale analysis of polymorphisms (e.g. to correlate with susceptibility to diseases and environmental agents--that is, molecular epidemiology), comparative genome analysis, functional genomics (understanding how genes execute their functions), protein folding (the basic lexicon of 1,000-1,500 protein motifs), *cis* and trans-regulatory control of genes, evolution, and ultimately powerful new approaches to systems analysis (e.g. monitor the global analyses of mRNA transcription and protein expression, analyze how the secondary modifications of proteins alter their functions, look at the interactions of all proteins, etc.). From these opportunities, one in particular is critical to emphasize. Genomics has given us the tools to revolutionize the systems study of gene regulation--this is a striking opportunity that is ripe for exploitation.
- **Model Systems Permit the Deciphering of Informational Pathways.** Yeast, *Drosophila*, and to a lesser extent the nematode, have demonstrated the power of genetics (e.g. transgenics, knock-outs, over expression, etc.) coupled with genomics to begin defining basic informational pathways. The striking single observation is the profound conservation of informational pathways and strategies across species. Thus, we can readily learn how many genes (and systems) work in the genetically

manipulable organisms such as *E. coli*, yeast, *Drosophila*, and nematode and these insights will have applications to human biology. Moreover, the mouse provides the opportunity to analyze mammalian (or vertebrate) specific information pathways for more evolved systems such as immunity, the nervous system, and how the liver deals with environmental toxins.

- **Microbial Genome Analysis Has Opened Up New Worlds.** Microbiology has been truly revolutionized by DOE's pioneering microbial genome project. The possibility of taking genome-wide approaches to the analysis of microbial function could revolutionize our ability to manipulate microorganisms to deal with human environmental challenges (e.g. toxic waste). The same may be true of new approaches to producing energy. The possibility of genome engineering--modifying microbial genomes to execute completely new types of functions is incredibly exciting.
- **The "New Biology" Has Given Us Powerful New Approaches to the Systems Analysis of Complex Biological Systems and Networks.** The analysis of complex systems requires an identification of their elements, their interconnections, and ultimately their systems properties. Clearly the genome sequences potentially identify all genes for humans and model organisms. Genetics can identify genes predisposing to particular traits. Model organisms can be used to define fundamental informational pathways. The high-throughput tools have the capacity to break very complex systems into simpler subsystems that are analyzable and still exhibit systems properties (e.g. for immunity an analyzable subsystem is the families of genes encoding T cell receptors). As we move into the 21st century, systems analysis will be the dominant theme in biology.

In what follows below, various members of the committee have elaborated on various of the themes mentioned in the discussion.

COMPARATIVE GENOMICS (David Galas)

A major area of opportunity for the BER program is the area of comparative genomics. Not only is this area now just becoming ripe for exploration, but the benefits to health and environmental understanding and, therefore, to medicine and environmental protection, will be enormous. This area consists of two general components: first, the comparison of the genomic organization and

DRAFT

gene sequences of human with several model organisms, including mouse, *Drosophila*, *C. elegans* and yeast; and second, the comparison among the organization and sequences of microbial genomes. The first area should focus initially on: 1) a systematic comparison of the apparently homologous regions of the worm and fly genomes with human sequence; and 2) the parallel sequencing and systematic comparison of corresponding regions of the mouse and human as part of the new plan for the Joint Human Genome Institute. As discussed elsewhere, the comparisons will need to include especially the regulatory regions of the genome. Future biological understanding of the human will depend not only on the structure and function of the encoded proteins and RNA's, but particularly on the nature of the regulatory networks that control the batteries of genes in their expression in space and time in the organism.

The BER program has made a significant and long-lasting contribution to the nation's scientific and technical potential by its innovative and prescient initiative in establishing the microbial genome program. This program has transformed both the fundamental microbiological sciences and the applied sciences of environmental and medical microbiology. The BER program should make sure that it does not lose this leading role by insuring that the applications of this fundamental program's results are addressed. One entirely new opportunity for the application of the vast amount of microbial genomic information that is becoming available is an area that can be termed Genome Engineering. For the first time the new knowledge permits us to conceive of transferring entire large batteries of genes, segments of genomes, from one microbe to another in order to engineer new properties of these organisms. This is the next natural step beyond single gene transfers and the recombinant organisms we are familiar with today. For example, the properties of tolerance of extreme environments (temperature, radiation etc.) with the ability to carry out complex metabolic processes could make novel and very useful organisms.

REGULATORY GENOMICS (Eric Davidson)

If we want to know how the genome works as a system, it is essential to understand the hard-wired regulatory information encoded in it. The heritable regulatory processes that control development are largely constituted of the *cis*-regulatory modules that govern the expression of genes. These determine when in the life cycle and where in the organism shall be expressed structural genes encoding the huge diversity of proteins, and regulatory genes encoding transcription factors, the gene control machines. Changes in *cis*-regulatory wiring have certainly contributed the major driving force of organismal evolution. Thus, fundamental insight into the two most profound problems in metazoan biology, development and evolution, depend on progress in understanding genomic regulatory hard-wiring.

DRAFT

The DOE could have a unique and global impact on the course of genomic science, were it to support a program directed specifically at regulatory genomics. Four areas which such a program might include are as follows:

- **Experimental cis-Regulatory Analysis.** The program should support experimental analyses, the objective of which is to characterize genomic *cis*-regulatory modules, by mapping and identifying transcription factor binding sites and determining their function *in vitro* by means of gene transfer. Research should be supported in any metazoan system in which direct molecular-level analysis is possible, as the knowledge gained will often be transferable to other genomes because of conservation of regulatory interactions amongst animal genomes.
- **Comparative Regulatory Genomics.** The program should support comparative explorations of given *cis*-regulatory systems of interest, designed to determine conservation, and differences--at the DNA level--amongst animal phyla or classes. Three forms of research that would fall in this category are: 1) identification and analysis of conserved *cis*-regulatory sequence elements; 2) functional tests of *cis*-regulatory sequences of one species by gene transfer into eggs or appropriate cells of another species; and 3) direct analysis of *cis*-regulatory interactions within homologous *cis*-regulatory sequences of different species.
- **Cis-Regulatory Informatics.** A concerted effort is required to make use of current knowledge to set up databases and algorithms that would permit recognition in DNA sequences of likely *cis*-regulatory site clusters; and permit inferences to infer the nature of the regulatory functions of these clusters from their identity, complexity, and organization. Key data such as site spacing, linkage amongst given sites, and functional evidence should be utilized. This effort will require engagement of molecular biologists who possess expert knowledge of *cis*-regulatory systems together with informatics groups and their resources.
- **Gene Network Systems.** This program would support research on functional connections between *cis*-regulatory systems within the genome, both upstream and downstream. Such research would include attempts to go upstream by isolating genes that encode transcription factors interacting in a given *cis*-regulatory system, and characterizing their *cis*-regulatory systems, and to go downstream by finding sets of genes that are the regulatory targets of given transcription factors, or combinations of factors.

DRAFT

MODEL SYSTEMS AND GENE PATHWAYS (Gerald Rubin)

The degree of evolutionary conservation--not just of gene sequence, but also of gene function--that has emerged from the study of genes and pathways in different organisms over the last few years has surprised even the most avid proponents of model organism research. It is striking how many of the most intensively studied and important genes in mammalian development--HOX genes, engrailed, and hedgehog to cite a few examples--were originally identified by screening mammalian DNA with probes derived from the corresponding *Drosophila* gene (whose importance in *Drosophila* had already been revealed by genetic and molecular studies). In the study of human genetic diseases, there are numerous cases where information from model organism has been crucial. The biochemical function of the NF1 gene was revealed by its homology to the yeast IRA2 gene. Knowledge of the sequence and biochemical function of genes involved in repairing DNA damage in yeast has provided critical insights into human colon cancer genes, not simply by providing a way to interpret the sequence of one gene, but by also illuminating a path to rapidly isolate other genes involved in the same pathology. For example, researchers are now actively looking for polymorphisms associated with the human homologues of all known yeast DNA repair genes in families with increased occurrence of cancer. Similarly, knowledge of the Patched and Notch receptor pathways in *Drosophila* has been important in understanding diseases such as skin cancer and stroke.

The recognition that fundamental developmental and physiological mechanisms are so similar between organisms is one of the most profound insights to come from all of biological science over the past several years. The intellectual excitement, interest and understanding that this revelation has generated will only increase as the genome projects of the model organisms continue. Just as the Rosetta stone gave the world the key to the long-forgotten language of ancient Egypt, so the biologically-annotated sequences of the genomes of the model organisms will provide an invaluable key to understanding the sequence of the human genome.

Among the model organisms, *Drosophila* is particularly well-suited for the role of determining human gene function. In terms of evolutionary conservation of sequence similarity, *Drosophila* is the closest of the genetically tractable model organisms (that is, bacteria, yeast, *Arabidopsis*, *C. elegans* and *Drosophila*) to humans. Moreover, in terms of morphological, physiological, and behavioral complexity, *Drosophila* is by far the closest to humans of these model organisms, yet its genome is not substantially bigger than the least complex metazoans. Finally, the large *Drosophila* research community has provided a wealth of information and understanding unusual in its depth and intellectual breadth. While the mouse is clearly more closely related to humans than *Drosophila*, one cannot do the kinds of sophisticated genetics in mice, such as a saturation genetic screen for all new mutations that effect a particular phenotype, that one can in the other

DRAFT

model organisms. While gene knock-out by homologous recombination in mice is a very powerful method, its expense, as well as the difficulties in interpretation of phenotypes caused by the high degree of apparent gene redundancy in mammals, limit its utility on a genome-wide scale.

CREATING AND SEARCHING A DATABASE OF PREDICTED PROTEIN STRUCTURES (George Rose)

For historical reasons, the preceding four decades of molecular biology have focused on template-driven processes--replication, transcription, and translation--the *trinity*. However, unlike the trinity, most biological organization arises via spontaneous, iterative self-assembly from components. Indeed, a principal benefit of current genome projects will be to provide a "parts list" of such components.

Fortuitously, during the timeframe of these genome initiatives, we have also witnessed progress in solving the protein folding problem. Obtaining predictive understanding of the chemical reaction in which a globular protein folds from a disordered state to its unique native conformation has been one of the century's major scientific challenges. A solution to this problem at the level of a high quality X-ray structure (~2 Å resolution) may take a while. However, a less accurate solution (~5 Å resolution) may be feasible in the immediate future.

It seems likely that even a 5-Å solution would be sufficient to recognize a protein fold and distinguish it from other possible folds. If so, then the utility of current genome initiatives can be amplified enormously by transforming sequence into structure.

Structure Is Far Better Conserved Than Sequence. In comparison to structural degeneracy, the genetic code is only modestly degenerate. For example, approximately Avogadro's number of conceivable nucleotide sequences can code for the amino acid sequence of ribonuclease A. But, the number of distinct protein folds can be reliably estimated to be $\sim 1,000 \pm 500$, and any amino acid sequences that folds at all will map to one of them.

The conservation of structure can be used to underwrite a more effective search paradigm. In the current paradigm, a target sequence of interest is compared against entries in a database of known sequences, with similarity scores computed following optimal pairwise sequence alignment. Similar sequences are recognized readily when the sequence identity is high (>30%), but all too often, closely related

DRAFT

sequences have diverged sufficiently to evade detection. This problem is both familiar and frustrating.

In a structure-based paradigm, the structure of a target sequence would be predicted and used to search either a database of known structures (e.g. the PDB) or of predicted structures (to be generated). The extreme degeneracy of structure assures that a structure-based paradigm would be far more effective, assuming that structure can be predicted from sequences with sufficient accuracy.

Therefore, a promising project for the DOE is to develop the means to conduct structure-based searches using current prediction methods. Conceptually, the approach is simple. A database of predicted structures or structural fragments would be compiled from all open reading frames of a test genome (e.g. bacterial or yeast). Then, the predicted structure of unknown targets would be used for structure searches against the database. In addition, all-against-all searches would be conducted among all database entries to classify the entire complement of proteins encoded by the genome.

This database project can also be used to identify proteins of particular interest for solution by X-ray crystallography and NMR spectroscopy. It is important to have at least one experimentally determined representative from every structural class. Predicted classes that are either unrepresented or underrepresented in the database of known structures could be detected and suitable examples chosen for subsequent structure determination.

TECHNOLOGY (Leroy Hood)

The need for technology development in biology has never been greater. Technology development fall into several categories: development of new instruments (e.g. high-throughput DNA arrays); the development of new strategies (e.g. the BAC-end sequencing strategy for high-throughput sequencing); the development of automated and integrated procedures (e.g. large-scale DNA sequencing from clone to finished sequence); and software (e.g. laboratory information management systems/databases/analytic software tools).

In keeping with biological opportunities described above, there are striking opportunities for technology development.

DRAFT

DNA

- High-throughput DNA sequencer: capillary, microchannel, single molecule, or improvement of standard machine)
- Integrated sequencing assembly line
- High-density DNA arrays: gene expression, identification, and typing of single base polymorphisms, protein/DNA interactions with double-stranded oligonucleotides
- Use of microfabrication to miniaturize, parallelize, high-sensitivity and high-throughput standard DNA procedures: PCR amplification, genotyping, sequencing, restriction mapping, etc.
- Better robots for picking plaques, arraying, pipetting, etc.
- Implementation of BAC-end sequencing strategy in human and mouse genomes
- Strategies for making full-length cDNA libraries
- Strategies for identifying rare message cDNAs

Protein

- Applications of mass spectrometry: protein identification, protein sequencing, protein modifications, protein interactions
- Synthesize 1,000s of peptides on a chip
- New tools for determining 3° structure (microscopy)
- Application of microfabrication to isolation and characterization of proteins with high-throughput approaches

Cells

DRAFT

- Rapid, multiparameter cell sorting
- Representative cDNA libraries from single cells
- *in situ* hybridization with rare message genes

Mouse

- Cheaper, more rapid knock-out, overexpression, and transgenic procedures
- Develop use of embryonic stem (ES) for genetic manipulation

Software

- Robust and extensible LIMS/databases for large-scale DNA sequencing
- Annotation of DNA sequences (putting biology into the sequences)
- Improved gene-finding algorithms
- Improved regulatory motif-finding algorithms
- Find protein motifs (1°+3°)

RECOMMENDATIONS

These recommendations are centered on opportunities for DOE to make unique contributions, and not to merely follow the lead of other funding agencies.

- Develop mouse as model system BAC-end sequencing for mapping syntenic comparison of human and mouse sequences
- Develop a few model systems for analyzing on a systems basis *cis*- and trans-regulatory control of gene expression; *Drosophila* and sea urchin provide the best

DRAFT

opportunities for the application of modern molecular techniques

- Expand the microbial genome program to include comparative genomics of microbes
- Create a database of predicted protein structures and the capacity to search it against new proteins
- Continue to emphasize technology development, e.g. microfabrication, high-density DNA arrays, mass spectrometry for protein analysis, and algorithmic development to mine the information of DNA and proteins for other particular opportunities

NHGRI ROUTE SLIP

Please Circulate

Division of Extramural Research	<i>DISTRIBUTION</i>		Message
<p><i>Copy to:</i> <i>Dr Callin</i> <i>Dr Jordan</i></p>	<i>Dr. Mark Guyer</i>	----- ✓	<p><i>This is a draft distributed at the latest HERAC meeting. It is a recommendation for the future of the DOE program.</i></p>
	<i>Dr. Bettie Graham</i>	----- ✓	
	<i>Dr. Elise Feingold</i>	----- ✓	
	<i>Dr. Jane Peterson</i>	-----	
	<i>Dr. Lisa Brooks</i>	----- ✓	
	<i>Dr. Adam Felsenfeld</i>	----- ✓	
	<i>Dr. Jeff Schloss</i>	----- ✓	
	<i>Dr. Eric Meslin</i>	----- ✓	
	<i>Ms. Elizabeth Thomson</i>	----- ✓	
	<i>Ms. Joy Boyer</i>	----- ✓	
	<i>Ms. Anita Allen</i>	-----	
<i>Ms. Peggy Whittington</i>	-----		
<i>Ms. Charlotte Quinn</i>	-----		
<i>Ms. Stephanie Reeves-Walker</i>	-----		
<p style="text-align: center;"><i>Office of Scientific Review</i></p>	<i>Dr. Ken Nakamura</i>	----- ✓	
	<i>Dr. Rudy Pozzatti</i>	----- ✓	
	<i>Dr. Jerry Roberts</i>	----- ✓	
	<i>Ms. Gwendolyn Williams</i>	-----	
	<i>Ms. Marchelle Dickerson</i>	-----	
<p style="text-align: center;"><i>Grants Management</i></p>	<i>Ms. Jean Cahill</i>	-----	
	<i>Ms. Sally York</i>	-----	
	<i>Ms. Linda Hall</i>	-----	
	<i>Ms. Diane Patterson</i>	-----	
	<i>Ms. Tara Mowery</i>	-----	
	<i>Ms. Monika Yakovich</i>	-----	
<p style="text-align: center;"><i>Office of Information Systems Management</i></p>	<i>Ms. Carol Martin</i>	-----	
	<i>Francis Collins</i>	-----	<i>Date: June 23, 1997</i>
	<i>Elke Jordan</i>	-----	<i>Return to:</i>
	<i>Kathy Hudson</i>	-----	<i>From: Jane Peterson</i>

Peterson, Jane

From: [REDACTED]
Sent: Wednesday, May 07, 1997 4:24 AM
To: Jordan, Elke; Guyer, Mark; Peterson, Jane; Schloss, Jeff
Subject: Re[2]: -No Subject-

FYI, update on DOE plans, we should discuss today at Program Staff.
FC

To: Francis Collins/DIR/NCHGR
cc:
From: [REDACTED]
Date: 05/06/97 04:30:00 PM
Subject: Re[2]: -No Subject-

I'm pushing the system as much as I can.
Hopefully, we'll have preapplications by 7/1; full proposals by 9/1.
Awards during the start of the fiscal year.
The RFA will be for one major collaboration on shotgun sequencing at our sequencing factory and a for a variety of smaller grants on associated collaborations.
I'll forward you a draft of the RFA via e-mail tomorrow. Critical comments welcome, please.
I'm impressed that you also garden. The only thing we sometimes plant is, you guessed it, hot peppers!

Reply Separator

Subject: Re: -No Subject-
Author:
RFC: [REDACTED]
X400PO
Date: 5/5/97 9:10 AM

Glad to hear that things are proceeding with the RFA. What is the expected timing of release of the RFA and due date for applications? We are considering the possibility of pushing back the decision about the third year of funding

for our pilot projects by one Council cycle (to May 1998 instead of February)

-- if we do this (and we will try to decide very soon, please keep this confidential for now) it would mean that the Hood application could come in this fall instead of this summer. Would that help?

We're still planning to go to Motorola on Friday, I'll try to find out whatever I can. We're scheduled for lunch with Galvin.

Hope you enjoyed this beautiful weekend. I finally got my backyard planting done!

Francis

To: [REDACTED]
[REDACTED]

Date: 05/05/97 08:54:00 AM
Subject: -No Subject-

Francis:

Sorry I didn't get to you earlier, but things always take longer than one hopes..

In any case, we're close to crafting an RFA for the academic involvement in our sequencing factory operations that I would like you to take a look at. I'm confident that it will accommodate a Hood proposal that can be jointly reviewed by both NIH and DOE.

Regarding Motorola, I would be curious to know whether they've written off any serious collaboration with the DOE effort.

Regards,

Ari

CONFIDENTIAL

cc: Fc
EJ

Francis:
Attached is a rough draft of the RFA.
Please review, dissect, and criticize.
Good luck tomorrow and I'm eager to hear your impressions.
Regards,
Ari

cc: Mark
Jane P
Jeff S.

OPTIONAL FORM 99 (7-90)

FAX TRANSMITTAL

of pages ▶ 4

To <i>Ari</i>	From <i>For Mark</i>
Dept./Agency	Phone # <i>Jane</i>
Fax #	Fax # <i>Jeff</i>

NSN 7540-01-317-7368 5099-101 GENERAL SERVICES ADMINISTRATION

Department of Energy
Office of Energy Research
Special Research Grant Program Notice: Human Genome Program - Technologies in support of
the DOE Joint Genome Institute

ACTION: Notice inviting grant applications.

SUMMARY: The Office of Health and Environmental Research (OHER) of the Office of Energy Research (ER), U.S. Department of Energy hereby announces its interest in receiving applications for Special Research Grants in support of the Human Genome Program. This Program is a coordinated multidisciplinary research effort to develop creative, innovative resources and technologies that lead to a molecular level understanding of the human genome. As one aspect of this program, the Department of Energy is establishing a Joint Genome Institute (JGI). The JGI will oversee a central sequencing facility that will initially have parallel production lines that use shotgun and transposon-based directed sequencing approaches. This dual approach is intended to evolve into an optimized and unified sequencing strategy within two to three years. This unified strategy will take advantage of technologies and expertise at the JGI and in the broader research community. An important aspect of developing this automated facility will be the establishment of external collaborations and partnerships aimed at technology development.

The goal of this notice is to support technology development that serves the needs of the JGI. Specific goals are: (1) to establish a cooperative technology development project with the JGI that will produce, within two years, an automated DNA sequencing production line based on either shotgun or directed strategies and (2) to develop and implement technologies for automated and advanced high-throughput DNA sequencing that can be integrated into the unified sequencing production strategy that is identified and implemented at the JGI.

In support of the first goal, the grantee will form a close collaboration with the JGI aimed at technology co-development and transfer for high throughput production DNA sequencing. A critical success factor for this effort will be the construction of a new, highly automated pilot DNA sequencing production line at the JGI within 6 months of the project's start. The grantee, working in conjunction with the JGI, will build and maintain automated devices for DNA purification, DNA sequencing, and automated finishing. It is anticipated that this pilot DNA sequencing production line will use, in significant part, technology supplied by the grantee. The second phase of the project, to be completed within two years, will be the development of a high throughput DNA sequencing production line. It is anticipated that this production line will lead current technology in automation and the minimization of human labor and will ultimately produce 100-200Mb of finished human genomic sequence per year. The grantee will sequence DNA, in support of the DOE effort, to evaluate and validate any modifications in sequencing technology required between the pilot and production phases. It is estimated that one major award, for a total of \$4.5 million in FY 1998, will be made.

In support of the second goal, technology developments aimed at improving the constituent technologies and overall performance of the JGI DNA sequencing production line are sought. These could include, innovative instrumentation and automated systems that offer the potential for

rapid, cost-effective sequencing of approximately a million bases per day; for non-gel techniques and direct imaging approaches; for development of applied genome informatics software for use in DNA sequencing and functional interpretation, including information retrieval; for user interfaces compatible with Genome Data Base (GDB), Genome Sequence DataBase (GSDB), and GenBank; and for communications, software engineering, and data management. Improved algorithms and hardware for DNA sequence annotation, including identification of homologies, regulatory sites, and protein coding regions can also be included. It is anticipated that between 3-6 awards for a total of \$1.5 million could be made in FY 1998.

Potential applicants are encouraged to submit a brief preapplication in accordance with 10 CFR 600.10(d)(2), consisting of a two to three page narrative describing the research project objectives and methods of accomplishment. These will be reviewed relative to the scope and research needs of the DOE Human Genome Program. Preapplications referencing Program Notice 97-XX should be received by June 15, 1997, and sent to Dr. Marvin E. Frazier, Office of Health and Environmental Research, ER-72 (GTN), Washington, D.C. 20585, (301) 903-6488. Telephone, telefax numbers, and Electronic mail addresses are required parts of the preapplication. A response to the preapplications discussing the potential program relevance of a formal application generally will be communicated generally within 15 days of receipt.

DATES: Formal applications submitted in response to this notice must be received by 4:30 p.m., E.D.T., August 1, 1997, to be accepted for merit review in and to permit timely consideration for award in Fiscal Year 1998.

ADDRESS: Formal applications referencing Program Notice should be forwarded to: U.S. Department of Energy, Office of Energy Research, Acquisition and Assistance Management Division, ER-64, Room G-236, Washington, D.C. 20585, ATTN: Program Notice 97-xx. The following address must be used when submitting applications by U.S. Postal Service Express, any commercial mail delivery service, or when handcarried by the applicant: U.S. Department of Energy, Acquisition and Assistance Management Division, ER-64, 19901 Germantown Road, Germantown, MD 20874, attention: Ms. Debbie Greenawalt.

FOR FURTHER INFORMATION CONTACT: Dr. Marvin E. Frazier, Office of Health and Environmental Research, [REDACTED]

SUPPLEMENTARY INFORMATION: It is anticipated that \$5-6 million will be available for grant awards during FY 1998, contingent upon availability of funds. Multiple year funding of grant awards is expected, and is also contingent upon availability of funds. It is expected that most awards will be from 1 to 3 years and that there will be one award for \$4-5 million per year (total costs) with the remaining 3-6 awards in the \$200 thousand to \$400 thousand per year (total costs) range. Information about development and submission of applications, eligibility, limitations, evaluation, selection process, and other policies and procedures may be found in the ER Application and Guide for the Special Research Grants Program and 10 CFR Part 605, which is available on the World Wide Web at: <http://www.er.doe.gov/production/grants/grants.html>. The OER, as part of its grant regulations, requires at 10 CFR 605.11(b) that a grantee funded by OER and performing research involving recombinant DNA molecules and/or organisms and viruses

containing recombinant DNA molecules shall comply with the National Institutes of Health "Guidelines for Research Involving Recombinant DNA Molecules" (51 FR 16958, May 7, 1986), or such later revision of those guidelines as may be published in the Federal Register. The dissemination of materials and research data in a timely manner is essential for progress towards the goals of the DOE Human Genome Program. OHER requires the timely sharing of resources and data. Applicants should, in their applications, discuss their plans for disseminating research data and materials which may include, where appropriate, putting cell lines, probes, sequence data, etc., into public repositories. Funds to defray the costs of disseminating materials or submitting data to repositories are allowable; however, such requests must be adequately justified.

The application kit and guide is available from the U.S. Department of Energy, Acquisition and Assistance Management Division, Office of Energy Research, ER-64, Washington, D.C. 20585 and is available on the World Wide Web at: <http://www.er.doe.gov/production/grants/grants.html>. The Catalog of Federal Domestic Assistance Number for this program is 81.049.

Issued in Washington, D.C. on .

John Rodney Clark
Associate Director for Management
Office of Energy Research

5/1/97 Friday

Chr 19 - The donor has now signed a new
confid. statement for cont'd use.

Chr. 16 - call line from ATCC - going to IRB
w/ info that there is no info out of it +
see what they say
Consent for any research.

Chr 22 - he will call LLNL.

DOE will not stop using old libraries over new
ones on line. D + some PAC will be used
for BAC end seq - but C will also
be used.

July 24 tentative date. BAC end w/ stop?

May 19 - CDNA seq.

how this affects
also see
how to go fwd?

Microbial Genome Panel July.

✓ w/ Mike if he is coming to CSH mtg.

Rev site visit 6/19-20 Argonne + Brookhaven

J-G1 - ^{new} plans late July for '98
leaders proposal not approved
RFA to open up to community

Trying to bring techs to bear a factory-
emphasis is on automation.

& shotgun assembly line.

factory will produce, test & improve techs also.

Factory space - 6 sites. Select a site by early July
Could be 1 year before produce seg @ Factory.

Stopped from all 3 labs - + vicinity.

97 goal = 10 Mb. - achieve? 98 = 20 Mb.

LLNL 5 Mb funded in '97.

Ari & Max S. send invite to mtg.

Summer Jason mtg in LaBolla. - visited
which & Berkeley - looking @ seg tech; migration
94 + 95.

JGE active - Seg has subtasks - mtg.

DOE funds will be focused @ Labs -

Maynard - still interested in aspects of Cornell
boutique org of problematic BACS

Hard - if no - but falls outside JGE -

Ari is committed to an appl for Hard.

Send any letter we get to DOE

Would still consider funding because
made commitment

~~DOE~~ Ask Elle re: 5-year Plan-DOE involvement

DOE RFA for next fall - 5 yrs away

6/2/97.

Peterson, Jane

From: Guyer, Mark
Sent: Tuesday, June 03, 1997 10:36 AM
To: Collins, Francis <fc23a>; Jordan, Elke; 'Hudson Kathy'; Boyer, Joy; Brooks, Lisa; Feingold, Elise; Felsenfeld, Adam; Graham, Bettie; Meslin, Eric M; Nakamura, Ken; Peterson, Jane; Pozzatti, Rudy O.; Roberts, Jerry; Schloss, Jeff; Thomson, Elizabeth
Subject: report on BAC end workshop

DOE BAC End Sequencing Workshop May 29, 1997

The DOE held a meeting that was a combination workshop about the proposed BAC end sequencing strategy and review of the two pilot projects set up in 1996. The attendees were:

Reviewers:	Pilot Project Folk:
Elbert Branscomb	Lee Hood
Lisa Stubbs	Greg Mahairis
Mike Palazzolo	(2 others from U. Washington whose names I didn't get)
Stan Letovsky	Mel Simon
Trevor Hawkins	Ung-jin Kim
Rick Myers	Mark Adams
David Nelson	Ham Smith
Bob Cottingham	(2 others from TIGR)
Norman Doggett	Glen Evans
Larry Deaven	Skip Gamer
DOE Staff:	Pieter de Jong
Marvin Frazier	Julie Korenberg
Marvin Stodolosky	
Dan Drell	
Dave Thomassen	
Arthur Katz	
Jim Beall	
NIH staff	
Adam Felsenfeld	
Jane Peterson	
Jeff Schloss	
Mark Guyer	

In an initial executive session (for which only JP among the NHGRI staff arrived early enough to sit in on), the following were given for the level of funding for the BAC end sequencing pilots:

Hood (/Simon/Adams) -- \$3,000,000
Evans (/deJong/Simon) -- \$1,200,000 (Evans -- \$500,000; De Jong -- \$500,000; Korenberg -- \$200,000)

We assumed that, in each case, these were total cost figures for the first year of a two year pilot.

The meeting began with presentations from the two pilot project groups:

A. Hood/Simon/Adams

Lee Hood began by describing his group's strategy, which involves analysis of a deep (15X) genomic BAC library by arraying 300,000 clones, and sequencing both ends of all clones as well as fingerprinting them all. They refer

to the sequences as STCs (sequence tagged connectors) and propose using the data for sequence walking; they also argue that the data will be very useful for constructing deep genetic and physical maps.

The goals of their pilot are to establish the approaches for purifying BAC DNA for end sequencing, developing automation to allow very high throughput, and generation of data of useful quality at low cost. Specific scientific concerns to be addressed included the randomness of the BAC library, the fraction of STCs with unique sequence, the dependence on a single BAC library, and the quality (representativeness) of BAC clones.

In summary, to date they have gotten sequence representing about 15,000 STCs (a total of 6.4 megabases of sequence, or 0.21% of genome). They have an average read length of 384 bp per STC, and estimate the error frequency at 0.1 to 1%. 89% of the STCs have greater than 30 bp of unique sequence. Lee argued that they have good evidence that the STCs they have determined represent random sequence; the evidence includes the frequency with which the STC sequences have hits known genomic sequence and ESTs, the average GC content of the STCs is the same as the average for the human genome, FISH analysis, and the finding that they have obtained few, if any, exact repeats (i.e. duplicates). With respect to fingerprinting, they currently can fingerprint 576 BACs per day (with a single enzyme, I believe).

Mei Simon then described the Cal Tech component of the collaboration, the objectives of which are to disseminate BAC clones to TIGR and U. Washington, to develop efficient methods to prepare BAC DNA for end sequencing and characterization, and to analyze BAC ends in a 20 Mb region of chromosome 16 and a 40 Mb region of chromosome 22 as demonstrations of the utility of end sequencing for tiling path determination.

The current status of the Cal Tech libraries are:

Library A: 96000 clones, 4X, human male fibroblasts

Libraries B & C: 300,000 clones , 15X, human sperm, continued use approval

Library D: currently under construction, 2 randomly selected sperm samples from 6-8 anonymous donors, IRB approved protocol

In the chromosome 16 project, they have screened clones equivalent to 14X coverage and have picked over 1600 candidate chromosome 16 BACs. They have more than 2000 end sequences and about 500 bases/sequence. On chromosome 22, they have 1000 fingerprinted BACs (so far, their fingerprint analysis has been done with radioactive label, not on an ABI sequencer).

TIGR has gotten end sequences on about half of the chromosome 22 BACs so far.

Finally, Mark Adams reported on the TIGR group's effort. He noted that they are doing BAC end sequencing for both human and *Arabidopsis* (the *Arabidopsis* community has embraced the BAC end strategy for sequencing; so far they have sequenced 21 BACs, 9 of which were selected by BAC end criteria. The largest contig is 450 kb, of which only 17 kb is overlap). With human DNA, their results are as follows:

Chromosome 22 BACs:

data from both ends were obtained from 321 clones

data from one end only were obtained from 142 clones

128 clones yielded no data

Thus, they have analyzed 591 clones and have obtained 835 sequences.

For "random" (whole genome library) BAC clones:

data from both ends were obtained from 1643 clones

data from one end only were obtained from 727 clones

857 clones yielded no data (295 of those had no insert)

Thus, 3227 clones have been analyzed and 4414 sequences have been obtained.

Their current throughput is 360 lanes per day. The average trimmed length for the BAC end reads is 450 bases and the average phred quality value is 26.1

All three of the P.I.'s of this group were very enthusiastic about the potential contribution that the BAC end strategy could make in human genomic DNA sequencing.

B. Evans/de Jong/Korenberg

The second group had undertaken a smaller pilot effort, with the goal of the Evans lab being to generate 3800 PAC end sequences and that of the de Jong lab to generate 5000. As expressed by Evans, this group also has a somewhat different view of the role that BAC end sequencing could play, namely that end sequencing is a valuable tool, but it is not sufficient alone to generate the kinds of high quality maps that are needed for sequencing. In other words, BAC/PAC end sequencing could simplify the problem of rapidly building maps and therefore be a valuable contributor to sequence ready mapping.

To date, the Evans lab has generated 3683 end sequences, 707 from chromosome 11 clones, the rest from a whole genome library. The average useful read length is 164 bp [Jeff and I both got that number, but it seems really low – I got the same number]. FISH analysis has given "extremely low" level of multiple signals, from which they conclude that clone distribution is "acceptably random."

Skip Gamer then discussed some of their hardware/software developments: they have used Sagian robot for production end sequencing of PACs; they have developed a new primer selection program (PRIMO) that uses phred quality data and gives a 20% increase in the success rate for primer selection; and they are "starting to use" the Astral sequencer for production sequencing at 144 lanes.

The focus of Pieter de Jong's talk was on library construction. He reported that the construction of the new libraries is well under way. They obtained blood samples from 20 donors and already have a 30X library from a single male (but what this means is that all the ligations are done, the DNA electroporated, and they have picked about 3X redundancy; they plan to finish the picking in "a couple of weeks," after which there will still remain the hard work of replicating the library). Pieter claimed they should be ready to distribute at least 10X within a couple of months. He also noted that his lab has new libraries from mouse (10X, not arrayed yet) and rat, 8X for the dog genome, and is also working on baboon and chimp libraries.

As for end sequencing, Pieter has about 3-4 people working on it and they have done about 7000 end sequences with about a 30% failure rate (mostly PCR failures). The average read length is 384 bases.

Finally, Julie Korenberg spoke. She started by describing the BAC end strategy in terms of producing an "integrated resource" that will be used, not only for sequencing, but also eventually for downstream applications. She then went on at some length about the downstream applications, but didn't illuminate the BAC end strategy issues very much.

A short discussion period followed, in which a few generic questions were considered – how accurate do the BAC end sequences have to be for this strategy to work? How random are the BAC libraries really and will the BAC end strategy address this (no seemed to be the answer)? How will the BAC end strategy address the clone fidelity issue (deep fingerprinting, which was not a component of the original BAC end proposal, would seem to be necessary)?

The pilot groups then left, and the remainder of the day was spent in executive session with the reviewers. As usual, no consensus was articulated but the general sense I took away was that the reviewers were not as convinced about the value of the BAC end approach as the proponents were. After the presentations, there were still questions remaining about whether the data will be good enough to allow detection of the proper matches when doing the comparisons on a genome-wide scale, as well as the issue of what the useable read lengths are. The reviewers were not convinced that the pilot projects have yet answered the question of whether they can produce data that will be useful for the proposed purpose, and they recommended that DOE have the current pilot project data made available on the Web for electronic analysis.

They also reiterated a previously expressed concern, namely will the sequencers use this information and how? They noted that the BAC end sequence data will really only have value if the entire sequencing community buys into it, and they questioned whether that will happen. Finally, the reviewers were skeptical of the estimates (\$0.04-0.05 per base) that had been given (primarily by Lee Hood) for the cost of BAC end sequencing; they noted that, at that cost, the -02 year commitments were large enough so that a significant number of additional BAC end sequences could be obtained.

On the other hand, several of the reviewers also made it quite clear that, at the least, generation of a large

number of BAC end sequences would be valuable as a source for new STSs that could be used in standard mapping approaches. The meeting ended with a request from DOE for individual reviewer comments.

Peterson, Jane

From: Graham, Bettie
Sent: Tuesday, July 01, 1997 4:50 PM
To: Jordan, Elke; Boyer, Joy; Brooks, Lisa; Feingold, Elise; Felsenfeld, Adam; Graham, Bettie; Guyer, Mark; Meslin, Eric M; 'Nakamura, Ken'; 'Peterson, Jane'; Pozzatti, Rudy O.; Roberts, Jerry; 'jeff schloss'; 'Thomson, Elizabeth'; 'Collins, Francis <fc23a>'
Subject: FW: DOE HGP Competition announced

FYI

Bettie

From: [REDACTED]
Sent: Tuesday, July 01, 1997 3:10 PM
Subject: DOE HGP Competition announced

This announcement is copied from:
http://www.er.doe.gov/production/grants/fr97_17.html

Office of Energy Research

Notice 97-17
Human Genome Program
Technologies in Support of the DOE Joint Genome Institute

Department of Energy
Office of Energy Research

Energy Research Financial Assistance Program Notice 97-17; Human Genome Program

-
Technologies in Support of the DOE Joint Genome Institute

AGENCY: U.S. Department of Energy

ACTION: Notice inviting grant applications

SUMMARY: The Office of Health and Environmental Research (OHER) of the Office of Energy Research (ER), U.S. Department of Energy (DOE), hereby announces its interest in receiving applications for support of the Human Genome Program. This Program is a coordinated multidisciplinary research effort to develop creative, innovative resources and technologies that lead to a molecular level understanding of the human genome. As one aspect of this program, the DOE is establishing a "Joint Genome Institute" (JGI) to develop a DNA sequencing factory. The JGI will oversee a central sequencing facility that will initially have parallel production lines that use shotgun

and transposon-based directed sequencing approaches. This dual approach is intended to evolve into an optimized and unified sequencing strategy within two to three years. This unified strategy will take advantage of technologies and expertise at the JGI and in the broader research community. An important aspect of developing this automated facility will be the establishment of external collaborations and partnerships aimed at technology development. The JGI's genomic sequencing program will also be coupled to a collection of experimental functional genomics approaches designed to provide a partial functional characterization of the genes as they are revealed by the sequencing. Here, the primary goal will be to develop cost-effective approaches that can yield worthwhile functional information. A related goal is to develop improved ways of integrating human genomics with the information coming from model organism genomics.

DATES: Preapplications referencing Program Notice 97-17 should be received by August 1, 1997. Formal applications in response to this notice must be received by 4:30 p.m., E.D.T. October 16, 1997, to be accepted for merit review and to permit timely consideration for award in FY 1998.

ADDRESSES: Preapplications referencing Program Notice 97-17 should be sent to Dr. Marvin E. Frazier, Office of Health and Environmental Research, ER-72, Office of Energy Research, U.S. Department of Energy, 19901 Germantown Road, Germantown, MD 20874-1290; e-mail is acceptable for submitting preapplications using the following address: joanne.corcoran@oer.doe.gov. Formal applications referencing Program Notice 97-17 should be forwarded to: U.S. Department of Energy, Office of Energy Research, Grants and Contracts Division, ER-64, 19901 Germantown Road, Germantown, MD 20874-1290, ATTN: Program Notice 97-17. This address must be used when submitting applications by U.S. Postal Service Express Mail or any commercial mail delivery service, or when hand-carried by the applicant. An original and seven copies of the application must be submitted; however, applicants are

requested not to submit multiple application copies using more than one delivery or mail service.

FOR FURTHER INFORMATION CONTACT: Dr. Marvin E. Frazier, ER-72, Office of Health and Environmental Research, Office of Energy Research, U. S. Department of Energy, 19901 Germantown Road, Germantown, MD 20874-1290, telephone: (301) 903-6488, e-mail: joanne.corcoran@oer.doe.gov.

SUPPLEMENTARY INFORMATION: The goal of this notice is to support technology development that serves the needs of the Department of Energy's (DOE) Joint Genome Institute (JGI). The DOE JGI is developing a high throughput DNA sequencing factory. This factory will take advantage of the complementing strengths of each of the three current DOE Genome Centers: Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory (LANL). The JGI Sequencing Factory will be physically located in proximity to LLNL and LBNL. The Scientific Director of the DOE Human Genome Program, Dr. Elbert Branscomb, is the leader of the JGI. With respect to the JGI genomic sequencing task, the specific goals are: (1) to establish a cooperative technology development project with an outside entity that will produce, within two years, an automated DNA sequencing production line based on either shotgun or directed strategies; and (2) to develop and implement technologies for automated and advanced high-throughput DNA sequencing that can be integrated into the unified sequencing production strategy that is identified and implemented at the JGI.

In support of the first goal, the grantee will form a close collaboration with the JGI aimed at technology co-development and transfer for high throughput production DNA sequencing. A

critical success factor for this effort will be the construction of a new, highly automated pilot DNA sequencing production line at the JGI within 6 to 9 months of the project's start. The grantee, working in

conjunction with the JGI, will help build and maintain automated devices as appropriate for this pilot line

(e.g., those for DNA

purification, DNA sequencing, and automated finishing). It is anticipated that this pilot DNA sequencing production line may use, in significant part, technology supplied by the grantee. The second phase of the project, to be completed within two years, will be the development of a high

@VGI?

throughput DNA sequencing production line. It is anticipated that this production line will lead current technology in automation and the minimization of human labor and will ultimately produce 100-200 Mb of finished human genomic sequence per year. It is also expected that, in close cooperation with the JGI, the grantee will use the technology being supplied to perform a significant amount of DNA sequencing on targets that support the DOE effort. This would be designed to drive the technology development and to permit modifications in technology between the pilot and production phases to be evaluated and validated under high throughput conditions. It is estimated that one major award, for a total of approximately \$4 million in FY 1998, will be made.

by when

how much?

+ 99?

In support of the second sequencing goal, technology developments aimed at improving the constituent technologies and overall performance of the JGI DNA sequencing production line are sought. These could include: innovative instrumentation and automated systems that offer the potential for rapid, cost-effective sequencing of approximately a million bases per day; for non-gel techniques and direct imaging approaches; for development of applied genome informatics software for use in DNA sequencing and functional interpretation, including information retrieval; for user interfaces compatible with Genome Data Base (GDB), Genome Sequence DataBase (GSDB), and GenBank; and for communications, software engineering, and data management. Improved algorithms and hardware for DNA sequence annotation, including identification of homologies, regulatory sites, and protein coding

regions can also be included. It is anticipated that between 2-4 awards for a total of up to \$1 million could be made in FY 1998.

+ 99?

With respect to the functional genomics and model organism goals, projects in the following program areas are solicited: 1) strategies for full-length cDNA clone generation and

sequencing and for economically and accurately determining transcript lengths and types; 2) strategies for expression mapping, sub-cellular localization, and pathway tracing; 3) economical approaches for revealing single base pair polymorphisms and for characterizing their haplotypes; and 4) affordable approaches for using model organisms to systematically relate phenotype information to anonymous genes discovered in the human genome. It is anticipated that between 2-4 awards for pilot and proof-of-principle studies, for a total of up to \$1 million could be made in FY 1998.

Potential applicants are strongly encouraged to submit a brief preapplication that consists of two to three pages of narrative describing the research objectives and methods of accomplishment. Preapplications will be reviewed relative to the scope and research needs of the DOE Human Genome

Program, as outlined in the summary paragraph and in the SUPPLEMENTARY INFORMATION. Principal investigator address, telephone number, FAX number, and e-mail address are required as part of the

preapplication. A response to each preapplication discussing the potential programmatic relevance of a formal application generally will be communicated to the Principal Investigator within 21 days of receipt.

ER's preapplication policy can be found on ER's Grants and Contracts Web Site at: <http://www.er.doe.gov/production/grants/preapp.html>.

It is anticipated that approximately \$6 million will be available for grant awards during FY 1998, contingent upon availability of appropriated funds. Multiple year funding of

grant awards is expected, with out-year funding also contingent upon the availability of appropriated funds,

progress of the research, and programmatic needs. It is expected that most awards will be from one to three years and that there will be one award for approximately \$4 million per year (total costs) with the remaining 4-6 awards in the \$200 thousand to \$400 thousand per year (total costs) range. The dissemination of

materials and research data in a timely manner is essential for progress towards the goals of the DOE Human

Genome Program.

OHER requires the timely sharing of resources and data. Applicants should, in their applications, discuss their plans for disseminating research data and materials which may include,

where appropriate, putting cell lines, probes, sequence data, etc., into public repositories. Funds to defray the costs of disseminating materials or submitting data to repositories are allowable; however, such requests must be adequately justified.

Applications will be subjected to formal merit review (peer review) and will be evaluated against the following evaluation criteria which are listed in descending order of importance codified at 10 CFR 605.10(d):

1. Scientific and/or Technical Merit of the Project;
2. Appropriateness of the Proposed Method or Approach;
3. Competency of Applicant's personnel and Adequacy of Proposed Resources;
4. Reasonableness and Appropriateness of the Proposed Budget.

The evaluation will include program policy factors such as the relevance of the proposed research to the terms of the announcement and an agency's programmatic needs. Note, external

peer reviewers are selected with regard to both their scientific expertise and the absence of conflict-of-interest issues. Non-federal reviewers will often be used, and submission of an application constitutes agreement that this is acceptable to the investigator(s) and the submitting institution.

Information about development and submission of applications, eligibility, limitations, evaluation, selection process, and other policies and procedures may be found in the ER Application Guide for the Office of Energy Research Financial Assistance Program 10 CFR Part 605, which is available on the World Wide Web at: <http://www.er.doe.gov/production/grants/grants.html>. The ER, as part of its grant regulations, requires at 10 CFR 605.11(b) that a grantee funded by ER and performing research involving recombinant DNA molecules and/or organisms and viruses containing recombinant DNA molecules

shall comply with
the National Institutes of Health "Guidelines for Research Involving
Recombinant
DNA Molecules" (51 FR
16958, May 7, 1986), or such later revision of those guidelines as may be
published in the Federal
Register.

The Catalog of Federal Domestic Assistance Number for this program is 81.049
and
the solicitation
control number is ERFAP 10 CFR Part 605.

John Rodney Clark
Associate Director
for Resource Management
Office of Energy Research

Published in the Federal Register July 1, 1997, Vol. 62, No. 126, pages
35476-35478.

DOE Planning Meeting; December 1-2, 1997
Jane Peterson

DOE held a meeting of its five-year planning committee in Alta, Utah. The committee is chaired by Ray Gesteland and the members are Elbert Branscomb, Mario Capecchi, Skip Garner, Richard Gibbs, Phil Green, Trevor Hawkins, Keith Hodgson (not present), Mike Knotek, Mirian Meisler, Lloyd Smith, Randy Smith (not present), Monte Westerfield, Gerry Rubin (not present), Mike Waterman (not present). Marv Frazier and Ari Patrinos from DOE also attended.

DOE staff presented information about the DOE programs to set the context for the meeting. A few items of interest were:

- The HGP budget is split nearly evenly between the JGI (\$44 M/yr) and non-JGI (\$43.5 M/yr) activities
- Only \$7.4 M of the JGI funds are given directly to the JGI. The rest of the funds are in to National Lab's budgets with the understanding that they are to be redistributed to the JGI. This raised concern about Elbert's ability to have direct control of the money. He, however, believes that the Labs are committed to seeing the JGI succeed and will not cause a problem.
- The DOE said that they are "embarrassed" by the funds they have invested in informatics and the lack of results from it. They want to reassess this area and reinvest their funds in ways that will produce useful tools
- The JASONS will continue their interest in genomics with a meeting in January 1998 on QA/QC and another during their summer retreat on informatics.
- Applications in response to the RFA for JGI-partnerships will be reviewed next week. The JGI will not be involved in the scientific review, but will have a chance to comment on which projects will be most useful to its mission before projects are funded. Skip Garner and Lloyd Smith argued that the DOE should not be so concerned about peer review and more concerned about what will be useful to the JGI. Ari responded that the DOE's poor reputation in genomics is due to a lack of attention to peer review and that the JGI will have to make a very strong argument in order to convince headquarters that a project that reviewed poorly should be funded.
- The JGI goals for 1998 remain 20 Mb (Oct 1, 97 to Oct 1, 998).
- There will be no functional genomics in the JGI in the first year but Elbert expects to add it in year 2
- Once the JGI facility is occupied and producing data, DOE expects that they will be able to attract capable leadership for sequencing (Tony Carrano is currently acting in this capacity)

Each member of the committee made a presentation discussing directions that the DOE program could go in the next 5 years.

Informatics:

Phil Green took issue with many of the points in the JASONS report. He believes that software for data management within a center will continue to be specific to the center and not portable. He also thinks that there is only marginal (20% or so) improvement to be had in base calling and assembly software. He believes that finishing is fairly automated now in several centers and that the problem with finishing is that finishers need a clear protocol. They spend too much time on the data. His general philosophy is that there is no trade off in quality that is worth it in cost. He is not pessimistic about getting sequencing costs down to \$0.25/bp; \$0.10/bp will be more difficult. He does not believe that there are major investments needed in the lab based software area. Phil believes that the DOE supported Genome Channel is good and making important contributions to annotating the data for community use. He is currently working on a program called "Plan" that will process raw ABI traces so that it can be used to evaluate data from centers that don't use Phred/Phrap. It is Phil's opinion that mutational information should be part of annotation.

A fuller discussion of informatics was held the next day after Phil had left. There were several members of the committee who were not able to attend the meeting whose input was needed on the topics of databases and user software. It was agreed that there is a lot of information available to users and a lack of tools to use it fully. However, the group was cautioned that the advice of users, not just informatics experts, is needed to fully discuss what tools are needed for making genomic data useful and accessible. There was a lot of concern expressed about GDB's failure to make mapping data easily available and GSDB's contributions and role were not clear.

Sequencing: Richard Gibbs presented the total number of bases completed as reported at Bermuda and the projections for the next year. There was a good deal of discussion about whether the human genome sequence could be finished by 2005. Mike Knotek, who has been involved in a number of large physics projects, was particularly concerned about optimism that a 6 to 20-fold increase in efficiency and decrease (depending upon how you calculate the increase needed!) in cost can be realized. Those familiar with the issues were confident that such reductions can be made, but cautioned that it is still too early to tell. Miriam Meisler asked whether HGP could ask Congress for an extension in time (!).

Technology: Lloyd Smith, Skip Garner and Trevor Hawkins presented their visions of what technology is needed to first complete the sequence and second to provide for continually evolving technology for the future of the project. Their thoughts were primarily that the current systems could accomplish the project if necessary, the next step will be more integrated systems and eventually miniaturized systems. There was not much discussion about technology for functional genomics.

Functional Genomics:

Miriam Meisler presented her view of resources needed for the mouse. Although she passed out the "manifesto", she seemed to have backed down from the demand for the sequence ready map and the sequence of the entire mouse genome on a short time scale. (I asked her about it later and she said that she realized that the mouse genome will be

done eventually and that the cost to do it now was very high.) She talked primarily about the value of the mouse for annotating the human sequence and the value of a mutagenesis program. Specifically, she would like to see DOE support a generation of mutagenized panel of mice and put in place a “sequence on request” facility so that scientists can get BACs from a region of interest sequenced quickly. She also is interested in development of an insertional mutagenesis system that gives mutants at a high rate.

Mario Capecchi took the opposite view to Miriam. He urged DOE to consider making a mutation in each gene using recombination systems for targeting. He argued that the mutagenesis panels that are reliant upon phenotype are useful, but will only tell you about certain systems and most likely involve mutations in many genes. He stated that in his lab, knocking out one gene is about \$5,000 (mostly cage costs) so doing them all would be about \$0.5 billion. He stressed that the sequence of the mouse genome will be needed and that DOE should not look for cheap solutions, such as cDNA sequencing; that, in his opinion, is too little information for the investment. In his opinion, most of the differences between humans and mouse will be in the cis elements and you need genomic sequence to find them. He did suggest that a starting point for sequencing the mouse genome would be to start with the gene-rich regions. I found his comments to be very insightful and provocative and suggest that he be invited to the NIH mouse workshop.

Monte Westerfield talked about Zebrafish. Primarily he reviewed the status in the field with respect to genomic resources. He urged DOE to set up a high throughput BAC sequencing service. There was general agreement that it would be worthwhile to sequence syntenic regions of human, mouse and zebrafish.

At the end of the meeting, Ray Gesteland summed up and asked what further information was needed. It was agreed that much more discussion of informatics was needed and they would like to hear from Gerry Rubin, someone working on the expression database (Martin Ringwald, Jonathan Bard or Janan Eppig), Randy Smith, GDB and GSDB at the next meeting. It was agreed that a “chip” person was needed and Mark Chee and Pat Brown were suggested, if they were available.

I found the meeting useful and was surprised at how similar in thinking DOE and NHGRI have become. The one area that DOE is actively discussing that so far has not been fully discussed by NHGRI, is informatics. I think there are important issues in this area that need discussing although we may believe that some of the research, such as tools for users to mine genomic data, should be supported by the NIH as a whole. This may be an important future NIH-wide initiative. I was relieved that the mouse community (at least as represented by Miriam) seems to have had a reality check and I hope this means the discussion at our workshop will be useful. Even though DOE sees itself as taking a lead in technology (and informatics), I saw no unique vision as to what should be done to stimulate new technology.

①

12/1 DOE Planning mtg.

APB - DOE + NIH plans.

Partners - they had an earmark for the 1st time - \$10M in Secret to do microbial seg. Univ - DeBuccio 1/2 medical tumor samples - 1/2 microbial seg. Med Apps \$40M - Hirsch Health effects \$150M - Frojier } doubled in 5 yrs. Envr Sci - \$150M - Brando }

Seg - \$60M - Prod + \$10M - ^{SP}Seg. Ctrs + 1 GC center based on previous 50% (TR) 1bp in '98 + decision w/ 1/2 like every 4 yrs. (25% by 2002) Discipline in addn (10-12M) Tech dev. separate genome var. functional genomics - RFA's - deciding what to do on a genome scale. SWP map.

Part is ~ 10% of budget.

Outline for OER is physics.
 OER = $2.5B$

40% of budget goes to academic inst - Rest to Nat'l labs

Expectations?

- making changes in HGP @ DOE.
- Need good solid advice.

HGP budget - \$7.3M

Univ - \$21.8M

Labs - \$4.8M

Not for Prof. \$6.8M (TIGR, whitehead, NCGR)

JGI - \$44M

non JGI - \$43.5M

From JGI - get copy from Mar F.
~~ELSI 25~~

Lab	Group B =	LBNL	# 13.9	} JGI + instrumentation
		LLNL	19.2	
		LANL	6.7	
		JGI	7.4	

ANL 1.5
 BNL 1.4
 ORNL 2.6
 AMES 0.3
 PNNL 0.4
 ORISE - 0.9

because of JGI?

- Are jobs @ risk -? to some extent
- partnership of Academic community - Right way to do?
 forming the partnership in a more structured way.

Informatis - embraced w/ what they are funding in informatis. Management of DB + future work.

JASONS will continue interest in genome full blown summer study on informatis

Very little international funding - a little Russian + China

Frazier -

Whole program will be reviewed -

JGI is LLNL, LBNL, LANL

JGI an umbrella over whole program
want integrated program.

ANL + BNL will get "back in" to HGP by
competing

7.4 M in JGI is lab facility

* for JGE in labs are admin, etc.

LLNL ≈ \$12 M JGI

LLNL ≈ \$17 M JGI
16

JGI \$ are dictated by JGE - not lab

RFA-97-17 HGP

Techon Dev serving JGI

Autonomous

Hi throughput

large scale fuel genies

*6 M in FY198 - 3 yrs

how will JGI choose partners?

1- scientific review & pick best apps

2- go to JGI & pick out ones they think they
can work w/ & is relevant to them.

Expect to make 1 major \$4M award

Current

- Health Effects - \$40M
- Biol effects - \$20M
- Radiation Biol, str/func
- Mol Biol \$15M
- Microbial genome
- Cell cycle
- Cell Biol \$8M
- DNA repair

Restructured Program ~ \$38M

also 3 genomes

- low-level effects \$10M
- radiation Biol
- Susceptibility - DNA repair, polynucle
- Microbial genome - \$12M
- Comparative Genomics - \$10M
- mouse + Drosophila (tie to DNA repair)
- Biotechnology - \$6M (facilitate looking @ structures)
- hi risk; hi payoff.

Microbial Genomes - FY1998 - completed 6 genomes
 RFA - 8 more genomes funded.

- Citrium - TIGR use single seq ctr for microwg
- energy/energy release
 - Dis remediation
 - Carbon cycle
 - Waste Cleanup
 - DNA obtainable
 - Genome size < 8 Mb
 - genetically manipulable
 - non pathogenic
 - Sci interesting

(5)

TIGR will be doing most of seq -
#bp for a certain # of years
\$4.8M new awards \$5.5M total program
also new technologies & informatics

Elbort - JGI FY98
Gras for 98
- 20 Mb sequence
- functioning PSF

"Barnes base" -

- >90% in >1 Mb size; posit wrt public markers
- Coverage - <1 gap/200 kb; Sum of all gaps <1%; gaps measured.
- accuracy < 1 error per 10^4 bp
- participate in community g.c.

In '97 ~ 3 Mb to similar criteria

quality, productivity & cost in that order of priority.
In future, cost has to move to #2.

Maps? - enough to handle 20 Mb (~40 Mb)
Cost of mapping ↓

Quality - for all clones have quality values on web site

Occupied by 6 or 7/98 30,000 sq. ft
at least 80% of prod. seq.

no R+D @ UGI

how + when to show they work @ labs.

Sequencing strategy Guidelines

- conservative strategy decision
- 1st yr ramp up based on in-place expertise
- 12/97 making decision

Ramping in FY98

- 2/10/10 MB (LANL/LBL/LLNL)
- automatic grant & prod monitoring

Seq. strategy

- 6-7% shotgun front-end
mixed M13 + plasmid end seq.
- role of plasmids is closed in shotgun approach - based on transposon bombing

Subsequent Strategy

- what is best M13: plasmid ratio?
- low nucleotide shotgun?
- which plasmid system?
- which methods for plasmid insert seq?
- what seq. platforms, automation

what authority does Elbert have in other labs?

Elbert believes he has authority.

It has been tested. - he stopped a budget reduction
LANL has lost 35% of budget

meets directly w Lab directors. - they are cooperating
UC is also helping

Functional Genomics

- FY98 - • 10% comparative seq in mouse
- physical mapping to support subseq

RFA to prepare for next years (already used will be used)

- several small (~\$500k) pilot proj
 - criteria: illuminate known seq
- Infra & review of existing values to be applied to all proj.
- idea is want to do expression mapping of all genes.
- full cDNA seq in mouse?
 - need to "tag" phenotypes on genes

Ramping plans -

LANL - 0.7 mb / mo

LBL + LLNL will reach 1.5 mb a time by Oct months.

Biggest issues are hiring.

#1's of personnel? - 40 people in yr 1
more from labs - 60 people

How do results of RFA or seq support effect hiring?

liberty really real collob in yr 2.

tension between reaching some & concept of blowing for longer time.

A big issue about not doing "business as usual"

Management, Judgment, stds, budget, time scale, performance assessment

8

Will DOE labs look to the outside for expertise?
do they plan to do that? The facility is not just
for neutrality but also to encourage
reaching out.

Have you got the best people signed up? About not
sure. 1st yr OK, but beyond that need
better quality people.
If they succeed it will be an attractive
place to come.

Dranscomb 1@LLNL.gov

9

① Green - p 44 of Jasni's Report

Software portability? - for Green Ctr. Report -
This is really unique for each ctr.

Base calling - gains to be made are to extend
read length. - there are fundamental data quality
issues w/ that data. 30% more would be useful
(@ above accuracy level).

Assembly - need better methods for repeats -
on downside of curve in terms of gain
> 95% contigs assemble w/no errors

Better instruments would also give better data -
but not software problem.

Automated finishing - several groups have done it.
feels that a lot of problem w/ finishers is that
they don't have a defined protocol.

General Philosophy that no trade off in quality is worth
it in cost.

Not pessimistic to get prices down to 25-30¢/bp. 10¢ have
difficult.

Long tracking can be improved -

Bottom line that there are not major investments needed
in this area.

DB management -

CSDB disappeared. Not really adding anything

NCBI is making everyone pretty happy

Annotation is a problem

Umbach is a good step -

expects that NCBI ~~will~~ will probably also do.

Gene prediction - ~~is~~ in a consistent way would like to see better org -

Should be a collection of seq + evolutionary relationships.

Gene trees + functional information

GDB - has content that not in NCBI -

lots of polymorph data still unique -

genetic disease -

type to annotation effort.

tighter relationship between GDB + annotation consistent ^{the} _{messy?}

Should be quality data submitted along w/ seq.

Relevance of Phred on ABI software?

"Phred" process ABI raw traces - Could put other into it. Not distributed yet. - will need to do some tuning of Phred w/ new instrument but not much.

Mutational DB - where should data end up?

Phil thinks part of annotation.

Should we be set standards? - should be starting

Amotster is still in flux - a "fine" issue

Search launcher? useful.

Context organized pg - mostly map coord?

people are using NCBI - for mapping info.

Gibbs - the old model -

in other organisms - how will we do these if map tech doesn't improve?

Current db doesn't really have the ability to

~~support this~~ Support this.

database to handle the maps? 10 - BDB really has tried but inconsistent data.

Ship - real end user tools + annotation

Sequencing - Gibbs -

He thinks a factor of 3-6 is needed & current technology can do it. There will be continued improvements

Comparative sequencing

cDNA seq

Character of polymorphic seq.

Technology - Smith

future -

now -

- next phase - • more lanes → 192+
- obtain • Improved automation of loading
- internal markers for tracking
- more dyes
- larger + faster runs



Next 2 - • pumpable gels - robustness more automatable
- lane length
- cost


when? - no problem w/ read length.
3-5 yrs.

Next 3 - microfabricated systems -
↓ reagent consumption
↑ speed
more automated

Issues to bring, stuff on line - surface, fluid flow,
cost, robustness, lane length
10 yrs.

Next 4 - microelectronic systems
• seq by hydr - probably not do now
• single wave - skeptical - ^{finishing + auto} detect size bp. ^{end.}
• was spec MALDI.
Baker to 100 bp.
1 more break thru 2-300 bp
2 " " " 800+
most probably but a long way out

  finishing automation & software would help a lot

 Ship - agree - near term & long range.
Is ABI 95K? (Ship thought so)



Turn - production line for genomics -
Integrated systems
shrink everything

Smith - would like to see genome held hostage
to commercial vendors

Functional Genomics -

Missler

Functional Genomics + Mouse

- 1- Comparative genomics: mouse - best annotation of
human genome sequences.
holds up in conserved regions
- 2- Mutagenesis
about 500 mutants arise over the past -
hi mutation rate w/ ENU $1/1000$ /base.
random, whole genome
large scale mutagenesis
Bully in Jevon - resin
UK Steve Brown 4,000 looking for behavior
mutants
JAX - Skintie + Besser
ORNL - Justice + Johnson

What's needed?

- Can't get BAC seq where you need it
- Insertional mutagenesis systems at a hi rate

DOE could do a pilot project -
take adv of stocks @ ORNL

The system for mutagenesis is to insert TK next to a gene that you want to target & then select for TK to find deletions 1-2 Mb is a good size.

Schematic approach is #2 M

Copochi - mutagenesis by recomb - he thinks vertebrate genomes are 4X of inverte - 20,000 - 80,000 genes.

because of duplication probably better to make the mutations in the gene you want.

you're no longer at the mercy of phenotype.

You really don't want a phenotype. You really want to see minimal effects. Chip technology will be useful here to look at a lot of genes.

You can make any mouse you want for \$5,000. Big cost is analyzing the mouse.

Need to sequence the genomes & compare to human. All the differences will be cis elements.

You could seq @ the level that at least 10's the gene regions - have to seq > the structural genes.

Stressed don't look for short term solution - don't try to do it cheap - CDNAS won't work that well -

In real terms not that expensive.

\$5,000 - mostly ^{orig} cost - takes 4-5 hrs.
#.5B give you brock wts in all years.

Westfield -

Zebrafish -

2 outstanding features
embryology + genetics

optically transparent for 1st few days -
can be manipulated - do cell lineage

Analysis

Dev. is rapid - 10 hrs thru gastr.

end of day 1 - NS + organs by day 2

Can also transplant the cells. Also used
to make genetic stocks

Dev is ext. fertl -

Genetic analysis - Streisinger -

- can make homozygous

- + haploids

Clutch size is 100s.

Can do saturation mutagenesis screens. ~ 1/100 hits/locus
inad induces del.

25-30% of genome mapped in deletions

Seems - IP 7,000 mutations. 800 - 1,000 genes

Starting all Apic + ab Apic screens

Transgenics -

- Can insert transgene stably.
- Can use them to rescue mutant phenotypes

Transposon system - can be mobilized by transposase activity.

Homologous recomb is missing. Trying to make ES-type cells.

Genomics - Postelwait & Eisenman

- 1st map had a couple hundred markers
- currently 1900 congenous
- 460 cloned genes

3000 cM - 1.4 cM resolution = 750 Kb.
haploids makes it very easy to map.

Why Zebrafish?

2 examples -

Cloning gene in Zebrafish

"no-tail" mutant - ortholog of a mouse mutation
expression pattern is same as mouse.

Similarity extended to other genes.

Could bootstrap study of genes by looking @ other systems.

Postelwait - looked at Septoria - is fairly hi
looked @ 80 genes cloned in all 3 org

36 of these 50 were in synteny Mouse/Human
38 " " " " " Zebra/uman

Map mutants to regions in Zebra fish - if conserved w/ human - look @ human genes. find candidate genes.

Resources needed?

Have analyzed PAC library assemble these robotics -

Access to hi throughput seq of BACs

Genome size ~ 1/2 of human
25 chromosomes.

Surprising that you could PCR from human back to Zebrafish. Hydro is full back prob.

Synteny -

repeats? - probably less the repeats - may account for difference in genome size

Fugu - Brunner has MRC grant to seq 1,000 cosmid smaller community (Brunner's lab)

Zebra fish -> 100 labs in 25 countries - Can really be done in small labs.

PAC library done? didn't know.

Could seq - mouse/Zebra/Human in a region

Gibbs - what is missing is Ki throughput

Copriaki - could do mutation in big genome - but want to make it useful to expression studies later on.

Smith - 540. Glen

Argue for mouse + Zebrafish sequencing - opens things up for technology dev.

What other information?

Needs a chip person -

Informatics - Waterman + Randy Smith couldn't come alot of projects of questionable value.

Ship stated JGI's informatics is inadequate

Branscomb asked how to proceed that

Conclusion.

Issues:

1) internal working of JGI

2) support for sequencing production

3) DB - tools to query db

4) need tools. Comparative tools.

Janave.

Martin Ringwald + Jonathan Bard expression db. not cross species comparisons - no organized effort

Smith - could JGI be a mediating center to dev software that is transportable to community. Ship didn't agree - not sure you need common system.

Certain tools are adopted & free to put together
Hawkins agrees w/ Smith - put tog a system
that is modular. Torbenter did this.

10 Gibbs - incentive for standardization & exportability.

#2 databases - bring in Rungwald & Espig & GDB.
Getting a good articulation of what biologists
need & what is not being provided.
The dbs are very frustrating & cannot get the
info. - advice must be generic

Is there any need to do more than NCBI?

Ship - dbs - D benefit:

- 1) Tools to extract value for end-user utilization
- 2) Tool to better support gene hunters
- 3) Informatics research.
Integrate & coordinate better.
Improved interface/ user finders
- 4) preparing for big data

JASONs are going to talk about computer needs.

Branscomb - what should the dbs be?

An archive or interpretive?

Are fact that it must be interpretive.

However agreed that this is a big problem -

doing all searches & retrieve doing so. is big

Is the type of data known yet?

- That's being worked on.

When do you draw the line for annotation.

How to find out user needs:

Browsers - useful to hear from GDB, Annotation Count

CSDB? - need to see the problems.

Rubin has thought about this in *Drosophila*

Chip technology - Mizalchuk - (eye rolling!)

Pat Brown + Mark Chee -

Ask Mark Chee - he will talk about it.

Maybe commercial firms are getting experience
pushing technology is important.

Microfabrication?

Should look @ proteins? - 10 yr plan - but need to start now.
bring up @ HERAC.

Next meeting:

More input on the stuff they ID.

Come up w/ 5-yr plan.

Molecular Combing -

Stretching is uniform
not dep on length
pH dep

bind on end to optical fiber, other to bead
if bead is magnetic, can use magnet

Stretching is uniform

1 μm = 2 Kb - once calibrated is "forever".

Can be used for DNA + reconstruction -

drop method -

- no 1 direction of combing
- need small vol.
- " low conc
- long decay time

Have limited ability to use genomic DNA

Put genomic DNA in res. - put lower slip in
& pull out
get 700 human genes on 1 slide

Colloids w/ graphene -

use fluorescence hyper to look @ order.

Can look @ differences between presence of 2 probes

lower threat limit is 200bp
practically = 500-700bp

time frame -

Schwartz -

10M $2\frac{1}{2}$ weeks
300

Human genome - whole

10 Mb molecules - (3 mins long)

2-10 Mb - restr map + order contig
PAC I.

1500 molecules in 2 wks -

1.5 - 2 tu genome eq. - $3\frac{1}{2}$ wks - 10 people

PAD is program.

Want to make a reference map.

What they want is for NCI to support tumor cell maps -

May have to go to 10x to get whole genome covered.

Reference DNA a conglomerate -

- tumor DNA a 2x ref

1, 6, 7, 20, 22 + X -

aligned among 8 BACs - order clones.

current RFA 1 clone/Mb -

The program they are talking about
is finding the physical map
BAC map covers genome in 2 hrs
a close respiratory for the entire thing.

Completely like the map to cytogenetic -
find clones + analyze overlaps.

Schwartz - w/ end seq + PAC map could do
hi res. mapping onto restriction map.
Reference map for PAC1 - 1 yr.
~~reference~~

1997 DOE HUMAN GENOME MEETING
U. S. Department of Energy
Alta Lodge - Alta, Utah
November 30 - December 2, 1997

Schedule

Sunday, November 30

2:00 - 3:00 pm	Check in at <i>Alta Lodge</i>
5:00 - 6:00 pm	Social - Sitzmark Room
6:00 - 8:00 pm	Dinner

Monday, December 1

7:30 - 8:30 am	Breakfast
8:30 - 10:00 am	Meeting - Deck Room
10:00 - 10:15 am	Break
10:15 - 12:00 n	Meeting
12:00 - 3:00 pm	Lunch
3:00 - 5:30 pm	Meeting
5:30 - 6:30 pm	Social - Sitzmark Room
6:30 - 8:00 pm	Dinner
8:00 pm	Open for meeting if desired

Tuesday, December 2

7:30 - 8:30 am	Breakfast
8:30 - 11:00 am	Meeting
11:00 am	Afternoon Skiing (if you wish)

1997 DOE Human Genome Meeting

U.S. Department of Energy December 1-2, 1997 Participant List

Elbert Branscomb
DOE Joint Genome Institute
Lawrence Livermore National Laboratory
7000 East Avenue, L452
P.O. Box 55C7
[REDACTED]

Mario Capecchi
Department of Human Genetics
University of Utah
Room 2100, Eccles Bldg
Salt Lake City, UT 84112-5330
[REDACTED]

Dorothy Dart
Department of Human Genetics
University of Utah
Room 2160, Eccles Bldg
Salt Lake City, UT 84112-5330
[REDACTED]

Harold R. (Skip) Garner
UT Southwestern Medical Center
5323 Harry Hines Blvd.
MS NB10.2C4
Dallas, TX 75235-8591
[REDACTED]

Raymond Gesteland
Department of Human Genetics
University of Utah
Room 6160, Eccles Bldg
Salt Lake City, UT 84112-5330
[REDACTED]

Richard Gibbs
Baylor College of Medicine
Molecular and Human Genetics
One Baylor Plaza
Houston, TX 77030
[REDACTED]

Phillip Green
University of Washington
Department of Molecular Biotechnology
Fluke Hall on Mason Road, 3rd Floor
Seattle, WA 98195
[REDACTED]

Trevor L. Hawkins
185 Morrison Avenue
Apartment 201
Somerville, MA 02144
[REDACTED]

Keith Hodgson
Keck Building
Roth Way, RM327
Stanford University
Stanford, CA 94305-5080
[REDACTED]

Mike Knotek
Argonne National Laboratory
Advanced Photon Source, Bldg 401
9700 S. Cass Avenue
Argonne, IL 60439
[REDACTED]

Miriam H. Meisler
Department of Human Genetics
University of Michigan
2806 Medical Science 2
Ann Arbor, MI 48109-0618
[REDACTED]

Jane Peterson
NIH/National Human Genome Research Institute
38 Library Drive, MSC 6050
Room 614, Building 38-A
Bethesda, MD 20892-6050
[REDACTED]

Lloyd M. Smith
Department of Chemistry; Analytical Division
University of Wisconsin-Madison
1101 University Avenue
Madison, WI 53706-1396
[REDACTED]

Randall F. Smith
Director, Bioinformatics Research
SmithKline Beecham Pharmaceuticals
Mail Stop UW2230
709 Swedeland Rd
King of Prussia, PA 19406
[REDACTED]

Monte Westerfield
1245 University of Oregon
Institute of Neuroscience
Eugene, OR 97403
[REDACTED]

U.S. Department of Energy Staff

Ari Patrinos
U.S. Department of Energy
19901 Germantown Road
Germantown, MD 20874
[REDACTED]

Marvin Frasier
U.S. Department of Energy
19901 Germantown Road
Germantown, MD 20874
[REDACTED]

Marvin Stodolsky
U.S. Department of Energy
ER-72 GTN
Germantown, MD 10874-1290
[REDACTED]

A New Five-Year Plan for the U.S. Human Genome Project

Francis Collins and David Galas*

The U.S. Human Genome Project is part of an international effort to develop genetic and physical maps and determine the DNA sequence of the human genome and the genomes of several model organisms. Thanks to advances in technology and a tightly focused effort, the project is on track with respect to its initial 5-year goals. Because 3 years have elapsed since these goals were set, and because a much more sophisticated and detailed understanding of what needs to be done and how to do it is now available, the goals have been refined and extended to cover the first 8 years (through September 1998) of the 15-year genome initiative.

In 1990, the Human Genome programs of the National Institutes of Health (NIH) and the Department of Energy (DOE) developed a joint research plan with specific goals for the first 5 years [fiscal year (FY) 1991-95] of the U.S. Human Genome Project (1). It has served as a valuable guide for both the research community and the agencies' administrative staff in developing and executing the genome project and assessing its progress for the past 3 years. Great strides have been made toward the achievement of the initial set of goals, particularly with respect to constructing detailed human genetic maps, improving physical maps of the human genome and the genomes of certain model organisms, developing improved technology for DNA sequencing and information handling, and defining the most urgent set of ethical, legal, and social issues associated with the acquisition and use of large amounts of genetic information.

Progress toward achieving the first set of goals for the genome project appears to be on schedule or, in some instances, even ahead of schedule. Furthermore, technological improvements that could not have been anticipated in 1990 have in some areas changed the scope of the project and allowed more ambitious approaches. Earlier this year, it was therefore decided to update and extend the initial goals to address the scope of genome research beyond the

completion of the original 5-year plan. A major purpose of revising the plan is to inform and provide a new guide to all participants in the genome project about the project's goals. To obtain the advice needed to develop the extended goals, NIH and DOE held a series of meetings with a large number of scientists and other interested scholars and representatives of the public, including many who previously had not been direct participants in the genome project. Reports of all these meetings are available from the Office of Communications of the National Center for Human Genome Research (NCHGR) and the Human Genome Management Information System of DOE (2, 3). Finally, a group of representative advisors from NIH and DOE drafted a set of new, extended goals for presentation to the National Advisory Council for Human Genome Research of NIH and the Health and Environmental Research Advisory Committee of DOE. These bodies have approved this document as a statement of their advice to the two agencies, and the following represents the goals for FYs 1994-98 (1 October 1993 to 30 September 1998).

General Principles

Several general observations underlie the specific goals (Fig. 1) described here. The first observation is that successful development of new technology for genomic and genetic research has been essential to the achievements of the project to date and will continue to be critical in the future. It was clearly recognized, both in the 1988 National Research Council (NRC) report (4) and in the first NIH-DOE plan, that attainment of the ambitious goals originally set for the genome project would require significant technological advances in all areas, such as mapping, sequencing, informatics, and gene identification. As the genome project has proceeded, progress along a broad range of technological fronts has been conspicuous. Among the most notable of these developments have been (i) new types of genetic markers, such as microsatellites, that can be assayed by polymerase chain reaction (PCR); (ii) improved vector systems for cloning large DNA fragments and better experimental strategies and computational methods for assembling those clones into large, overlapping sets (contigs) that compose useful

physical maps; (iii) the definition of the sequence tagged site (STS) (5) as a common unit of physical mapping; and (iv) improved technology and automation for DNA sequencing. Further substantial improvements in technology are needed in all areas of genome research, especially in DNA sequencing, if the project is to stay on schedule and meet the demanding goals that are being set.

A second general observation concerns an evolution in the levels of biological organization at which genomic research will likely function over the next few years. Initially, attention was focused on the chromosome as the basic unit of genome analysis. Large-scale mapping efforts, in particular, were directed at the construction of chromosome maps. The sophisticated genetic linkage maps now available and the detailed physical maps that are being produced are clear measures of the success of that approach. However, other units of study for the Human Genome Project will also have increasing usefulness in the future. Therefore, further mapping efforts directed at both larger and smaller targets should be encouraged. At one end of the scale, "whole genome" mapping efforts, in which the entire genome is efficiently analyzed, have become feasible with developments in PCR applications and robotics. These approaches generally produce relatively low-resolution maps with current technology. At the other end of the scale, increasing attention needs to be paid to detailed mapping, sequencing, and annotation of regions on the order of one to a few megabases in size. Although small in comparison with the whole genome, a megabase is still large in comparison with the capabilities of conventional molecular genetic analysis. Thus, development of efficient technology for approaching detailed analysis of several-megabase sections of the genome will provide a useful bridge between conventional genetics and genomics, and provide a foundation for innovation from which future methods for analysis of larger regions may arise.

Third, a goal for identifying genes within maps and sequences, implicit in the original plan, has now been made explicit. The progress already made on the original goals, combined with promising new approaches to gene identification, allow this element of genome analysis to be given greater visibility. This increased emphasis on gene identification will greatly enrich the maps that are produced.

It must also be noted that, as in the original 5-year plan, these goals assume a funding level for the U.S. Human Genome Project of \$200 million annually, adjusted for inflation. As the detailed cost analysis for the first 5-year plan was performed in

F. Collins is the director of the National Center for Human Genome Research, National Institutes of Health, Bethesda, MD 20892.

D. Galas was associate director, Office of Health and Environmental Research, Department of Energy, Washington, DC 20585.

* Present address: Darwin Molecular, 2405 Carillon Point, Kirkland, WA 98033.

1991, a cost of living increase must be added for all years beyond FY 1991. This funding level has not yet been achieved (Table 1).

International Aspects

The Human Genome Project is truly international in scope, as the original planners envisioned it. Its success to date has been possible because of major contributions from many countries and the extensive sharing of information and resources. It is hoped and anticipated that this spirit of international cooperation and sharing will continue. This coordination has been achieved largely by scientist-to-scientist interaction, facilitated by the Human Genome Organization (HUGO), which has taken on responsibility for some aspects of the management of the international chromosome workshops in particular. These workshops have served to encourage collaboration and the sharing of information and resources and to facilitate the expeditious completion of chromosome maps.

Several notable individual international collaborations have marked the genome project so far. One is the United States-United Kingdom collaboration on the sequencing of the *Caenorhabditis elegans* genome. Scientists at the Los Alamos National Laboratory are collaborating with Australian colleagues to develop a physical map of chromosome 16, and investigators at the Lawrence Livermore National Laboratory are working with Japanese scientists on a high-resolution physical map of chromosome 21. Other joint efforts include the collaboration between NIH and the Centre d'Etude du Polymorphisme Humain (CEPH) on the genetic map of the human genome and the Whitehead/Massachusetts Institute of Technology-Généthon collaboration on the whole-genome approach to the human physical map. These are but examples of the myriad interrelationships that have formed, generally spontaneously, among participating scientists.

Specific Goals

Genetic map. The 2- to 5-cM human genetic map of highly informative markers called for in the original goals is expected to be completed on time. However, improvements to make the map more useful and accessible will still be needed. If the field develops as predicted, there will be an increasing demand for technology that allows the nonexpert to type families rapidly for medical research purposes. In addition, to study complex genetic diseases, there is a need to be able to easily test large numbers of individuals for many markers simultaneously. In the long run, polymorphic

Table 1. The budget for the Human Genome Project for NIH and DOE (in millions of dollars). Budgets for 1994 and 1995 have not yet been determined.

Fiscal year	NIH	DOE	Total	1991 Projection of Needs
1991	87.4	47.4	134.8	135.1
1992	104.8	61.4	166.2	169.2
1993	106.1	64.5	170.6	218.9
1994				246.8
1995				259.9

markers that can be screened in a more automated fashion, and methods of gene mapping that obviate the need for a standard set of polymorphic markers are also desirable.

Goals

- (i) Complete the 2- to 5-cM map by 1995.
- (ii) Develop technology for rapid genotyping.
- (iii) Develop markers that are easier to use.
- (iv) Develop new mapping technologies.

Physical map. An STS-based physical map of the human genome is expected to be available in the next 2 to 3 years, with some areas mapped in more detail than others and an average interval between markers of about 300 kb. However, such a map will not likely be sufficiently detailed to provide a substrate for sequencing or to be optimally useful to scientists searching for disease genes. The original goal of a physical map with STS markers at intervals of 100 kb remains realistic and useful and would serve both sequencers and mappers. Using widely available methods, a molecular biologist can isolate a gene that is within 100 kb of a mapped marker, and a sequencer can use such a map as the basis for preparing the DNA for sequencing. To the extent that they do not introduce statistical bias, the use of STSs with added value (such as those derived from polymorphic markers or genes) is encouraged because such markers add to the usefulness of the map.

Goal

- (i) Complete an STS map of the human genome at a resolution of 100 kb.
- Physical maps of greater than 100-kb resolution are needed for DNA sequencing, for the purpose of finding genes and for other biological purposes. Although a variety of options are being explored for creating such maps, the optimal approach is by no means clear. There is a need to develop new strategies for high-resolution physical mapping as well as new cloning systems that are well integrated with advanced sequencing technology. Technology for se-

quencing is evolving rapidly. Therefore, preparation of sequence-ready sets of clones should be closely associated with an imminent intent to sequence.

There is a pressing need for clone libraries with improved stability and lower chimerism and other artifacts and a need for better technology for traveling from one STS to the next. A greater accessibility to clone libraries should also be encouraged.

DNA sequencing. Although the goal of sequencing DNA at a cost of \$0.50 per base pair may be met by 1996 as originally projected, the rate at which DNA can be sequenced will not be sufficient for sequencing the whole human genome. Priority should be given during the next 5 years to increasing sequencing capacity by increasing the number of groups oriented toward large-scale production sequencing. Substantial new technology that will allow sequencing at higher rates and lower costs is also needed: evolutionary technology developed from improvements in current gel-based approaches and revolutionary technology developed on the basis of new principles. These developments will only occur if significantly greater financial resources can be invested in this area. It is estimated that an immediate investment of \$100 million per year will be needed for sequencing technology alone, to allow the human genome to be sequenced by the year 2005.

Goals

- (i) Develop efficient approaches to sequencing one- to several-megabase regions of DNA of high biological interest.
- (ii) Develop technology for high throughput sequencing, focusing on systems integration of all steps from template preparation to data analysis.
- (iii) Build up a sequencing capacity to a collective rate of 50 Mb per year by the end of the period. This rate should result in an aggregate of 80 Mb of DNA sequence completed by the end of FY 1998.

The standard model organisms should be sequenced as rapidly as possible, with *Escherichia coli* and *Saccharomyces cerevisiae* completed by 1998 or earlier and *C. elegans* nearing completion by 1998. It is often advantageous to sequence the corresponding regions of human and mouse DNA side by side in areas of high biological interest. The sequencing of full-length, mapped complementary DNA molecules is useful, especially if it is associated with technological innovation applicable to genomic sequencing.

The measurement of the cost of sequencing is complex and fraught with many uncertainties due to the diversity of approaches being used. However, we need to continue to reduce costs, as well as im-

prove our ability to assess the accuracy of the sequence produced. This latter point must be addressed in future sequencing efforts. Cost will be highly dependent on the level of accuracy achieved.

Gene identification. Identification of all the genes in the human genome and in the genomes of certain model organisms is an implicit part of the Human Genome Project. Although the previous 5-year plan did not explicitly identify this activity with a specific goal, progress in mapping and in technology now makes it desirable to do so. With both genetic and physical maps of the human genome and the genomes of certain model organisms becoming available and large amounts of sequence data beginning to appear, it is important to develop better methods for identifying all the genes and incorporating all known genes onto the physical maps and the DNA sequences that are produced. This information will make the maps most useful to scientists studying the involvement of genes in health and disease. While many promising approaches are being explored, more development is needed in this area.

Goals

- (i) Develop efficient methods of identifying genes and for placement of known genes on physical maps or sequenced DNA.

Technology development. The development of new and improved technology is vital to the genome project. Certain technologies, such as automation and robotics, cut across many areas of genome research and need particular attention. Cooperation in technology development should be encouraged where possible because it is likely to be more effective and efficient than competition and duplication. The technology developed must be expandable and exportable, the long-term goal being to create technology that will be available in many basic science laboratories and allow the efficient sequencing of other genomes. Technology development is costly and has not been sufficiently funded.

Goal

- (i) Substantially expand support of innovative technological developments as well as improvements in current technology for DNA sequencing and to meet the needs of the Human Genome Project as a whole.

Model organisms. Excellent progress has

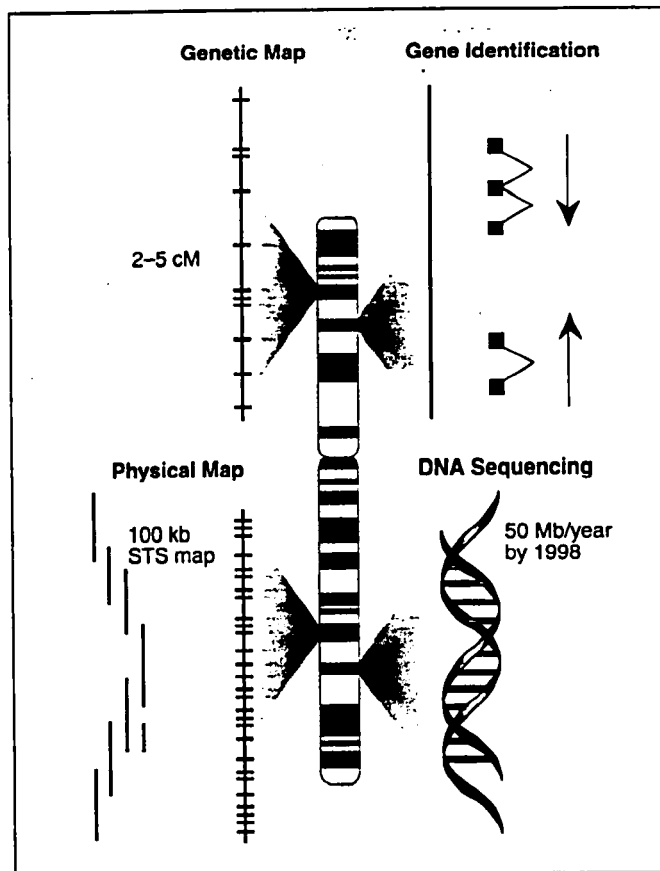


Fig. 1. Graphic overview of the new goals for the human genome. A 2- to 5-cM genetic map is expected to be completed by 1995 and a physical map with STS markers every 100 kb by 1998. Efficient methods for gene identification need to be developed and refined. The DNA sequencing goal of 50 Mb per year by 1998 includes all DNA, both human and model organisms, and assumes an exponential increase in sequencing capacity over time. Other important goals involving model organisms are not shown here, but are described in the text.

been made on the mouse genetic map and the *Drosophila* physical map, as well as the sequencing of the DNA of *E. coli*, *S. cerevisiae*, and *C. elegans*. Many of the original goals for this area are likely to be exceeded. Completion of the mouse map and sequencing of all the selected model organism genomes continue to be high priorities. The current emphasis for sequencing of mouse DNA should be placed on the sequencing of selected regions of high biological interest side by side with the corresponding human DNA.

Goals

- (i) Finish an STS map of the mouse genome at 300-kb resolution.
- (ii) Finish the sequence of the *E. coli* and *S. cerevisiae* genomes by 1998 or earlier.
- (iii) Continue sequencing *C. elegans* and *Drosophila* genomes with the aim of bringing *C. elegans* to near completion by 1998.
- (iv) Sequence selected segments of mouse DNA side by side with corresponding human DNA in areas of high biological interest.

Informatics. In order to collect, organize, and interpret the large amounts of complex mapping and sequencing data produced by the Human Genome Project, appropriate algorithms, software, database tools, and operational infrastructure are required. The success of the genome project will depend, in large part, on the ease with which biologists can gain access to and use the information produced. Although considerable progress has been made in this area since the beginning of the genome project, there is a continuing need for improvements to stay current with evolving requirements. As the amount of information increases, the demand for it and the need for convenient access increase also. Thus, data management, data analysis, and data distribution remain major goals for the future.

Goals

- (i) Continue to create, develop, and operate databases and database tools for easy access to data, including effective tools and standards for data exchange and links among databases.
- (ii) Consolidate, distribute, and continue to develop effective software for large-scale genome projects.
- (iii) Continue to develop tools for comparing and interpreting genome information.

Ethical, legal, and social implications (ELSI). The ELSI components of the Human Genome programs of NIH and DOE are strongly connected with genomic research so that policy discussions and recommendations are couched in the reality of the science. To date, the focus of the ELSI programs has been on the most immediate potential applications in society of genome research. Four areas were identified by advisers to the ELSI program for initial emphasis: privacy of genetic information, safe and effective introduction of genetic information in the clinical setting, fairness in the use of genetic information, and professional and public education. The program gives strong emphasis to understanding the ethnic, cultural, social, and psychological influences that must inform policy development and service delivery. Initial policy options for genetic family studies, clinical genetic services, and health care coverage have been developed, and reports on a range of urgent issues are expected by 1995.

As the genome project progresses, the need to prepare for even broader public impact becomes increasingly important. Poli-

cies are needed to anticipate the potential consequences of widespread use of genetic tests for common conditions, such as genetic predisposition to certain cancers or genetic susceptibility to certain environmental agents. In addition, as the genetic elements of behavioral and other nondisease-related traits are better understood, increased educational efforts will be needed to prevent stigmatization or discrimination on the basis of these traits. Continued emphasis on public and professional education at all levels will be critical to achieving these goals. Mechanisms for developing policy options that build on the current research portfolio and actively involve the public, the relevant professions, and the scientific community need to be developed.

Goals

- (i) Continue to identify and define issues and develop policy options to address them.
- (ii) Develop and disseminate policy options regarding genetic testing services with potential widespread use.
- (iii) Foster greater acceptance of human genetic variation.
- (iv) Enhance and expand public and professional education that is sensitive to sociocultural and psychological issues.

Training. There is a continuing need for individuals highly trained in the interdisciplinary sciences related to genome research. The original goal of supporting 600 trainees per year proved to be unattainable, because the capacity to train so many individuals in interdisciplinary sciences did not exist. However, now that a number of genome centers have been established, it is anticipated that training programs will expand. Although no numerical goal is specified, expansion of training activities should be encouraged, provided standards are kept high. Quality is more important than quantity.

Goal

- (i) Continue to encourage training of scientists in interdisciplinary sciences related to genome research.

Technology transfer. Technology transfer is already occurring to a remarkable extent, as evidenced by the number of genome-related companies that are forming. Many interactions and collaborations have been established between genome researchers and the private sector. In addition to the need to transfer technology out of centers of genome research, there is also a need to increase the transfer of technology from other fields into the genome centers. Increased cooperation with industry, as well as continued cooperation between the agencies, is highly desirable. Care must be taken, however, to avoid conflicts of interest.

Goal

- (i) Encourage and enhance technology transfer both into and out of centers of genome research.

Outreach. It is essential to the success of the Human Genome Project that the products of genome research be made available to the community. However, only a subset of the total information is likely to be of interest at any one time, with the nature of that subset changing over time. Therefore, it is desirable to have flexible distribution systems that respond quickly to user demand. The private sector is best suited to this situation and has begun to play an active and highly valued role. This should be encouraged and facilitated where possible, including the provision of seed funding in some instances.

The NIH and DOE genome programs have adopted a rule for sharing of information: Newly developed data and materials are to be released within 6 months of their creation. This policy has been well accepted. In many instances, information has been released before the end of the 6 months.

Goals

- (i) Cooperate with those who would establish distribution centers for genome materials.
- (ii) Share all information and materials within 6 months of their development. The latter should be accomplished by submission of information to public databases or repositories, or both, where appropriate.

Conclusion

To date, the Human Genome Project has experienced gratifying success. However, enormous challenges remain. The technology that will lead to the sequencing of the entire human genome at reasonable cost must still be developed. Major support of research in this area is essential if the genome project is to succeed in the long run. The new goals described here are designed to address the long- and short-term needs of the project.

Although there is still debate about the need to sequence the entire genome, it is now more widely recognized that the DNA sequence will reveal a wealth of biological information that could not be obtained in other ways. The sequence so far obtained from model organisms has demonstrated the existence of a large number of genes not previously suspected. For example, almost half of the open reading frames identified in the genomic DNA of *C. elegans* appear to represent previously unidentified genes. Similar results have been observed

in both *S. cerevisiae* and *E. coli* genomic DNA. Comparative sequence analysis has also confirmed the high degree of homology between genes across species. It is clear that sequence information represents a rich source for future investigation. Thus, the Human Genome Project must continue to pursue its original goal, namely, to obtain the complete human DNA sequence. At the same time, it is necessary to assure that technologies are developed that will allow the full interpretation of the DNA sequence once it is available. In order to increase emphasis on this area, an explicit goal related to gene identification has been added.

The genome project has already had a profound impact on biomedical research, as evidenced by the isolation of a number of genes associated with important diseases, such as Huntington's disease, amyotrophic lateral sclerosis, neurofibromatosis types 1 and 2, myotonic dystrophy, and fragile X syndrome. Genes that confer a predisposition to common diseases such as breast cancer, colon cancer, hypertension, diabetes, and Alzheimer's disease have also been localized to specific chromosomal regions. All these discoveries benefitted from the information, resources, and technologies developed by human genome research. As the genome project proceeds, many more exciting developments are expected including technology for studying the health effects of environmental agents; the ability to decipher the genomes of many other organisms, including countless microbes important to agriculture and the environment; as well as the identification of many more genes involved in disease. The technology and data produced by the genome project will provide a strong stimulus to broad areas of biological research and biotechnology. Exciting years lie ahead as the Human Genome Project moves toward its second set of 5-year goals.

REFERENCES AND NOTES

1. U.S. Department of Health and Human Services and Department of Energy, *Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years* (April 1990).
2. National Institutes of Health, National Center for Human Genome Research, Office of Communications, Bethesda, MD 20892. Phone, (301)402-0911; Fax, (301)402-4570.
3. U.S. Department of Energy, Human Genome Management Information System, Oak Ridge National Laboratory, PO Box 20008, Oak Ridge, TN 37831-6050. Phone, (615) 576-6669; Fax, (615) 574-9188.
4. National Research Council, Committee on Mapping and Sequencing the Human Genome, *Mapping and Sequencing the Human Genome* (National Academy Press, Washington, DC, 1988).
5. M.V. Olson, L. Hood, C. Cantor, D. Botstein, *Science* **245**, 143 (1989).

Human Genome Program

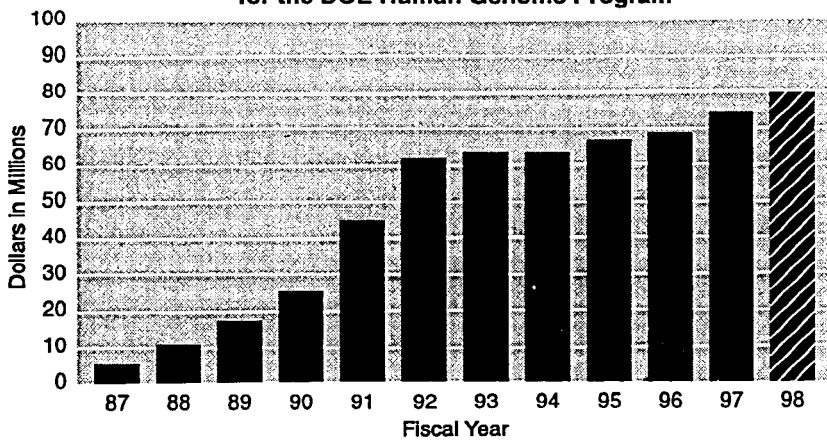
Coordination and Resources

Program coordination is the responsibility of the Human Genome Task Group (see box, p. 60), which, beginning in 1997, includes Elbert Branscomb, the Joint Genome Institute's Scientific Director. The task group is aided by the Biotechnology Consortium (which succeeded the former Human Genome Coordination Committee; see box, p. 60) to foster information exchange and dissemination. The task group administers the DOE Human Genome Program and its evolving needs and reports to the

Associate Director for Biological and Environmental Research (currently Aristides Patrinos). The task group arranges periodic workshops and coordinates site reviews for genome centers, the Joint Genome Institute, databases, and other large projects. It also coordinates peer review of research proposals, administration of awards, and collaboration with all concerned agencies and organizations.

The Biotechnology Consortium provides the OBER Associate Director with external expertise in all aspects of genomics and informatics and a mechanism by which OBER can keep track of the latest developments in the field. It facilitates development and dissemination of novel genome technologies throughout the DOE system, ensures appropriate management and sharing of data and resources by all DOE contractors and grantees, and promotes interactions with other national and international genomic entities.

Operating Expenditures and FY 1998 Projected Budget for the DOE Human Genome Program



Human Genome Program Fiscal Year Expenditures (\$M)

Year	Operating	Capital Equipment	Construction	Total
1996	68.3	5.6	5.7	79.6
1997	73.9	6.0	1.0	80.9
1998*	79.9	5.2	0.0	85.1

*Projected expenses.

Human Genome Program Operating Funds Distribution in FY 1996 (\$K)

FY 1996	Mapping	Sequencing	Sequencing Technology	Informatics	ELSI	Administration	Totals	%
DOE Laboratories	8,980	11,015	11,128	6,840	313	2,783*	41,059	60.1
Academic	6,671	4,368	3,257	6,178	642	4	21,120	30.9
Nonprofit	563	0	467	2,783	1,311	38	5,162	7.5
Federal	0	0	0	0	0	1,000**	1,000	1.5
Total	16,214	15,383	14,852	15,801	2,266	3,825	68,341	
% of Total	23.8	22.5	21.7	23.1	3.3	5.6	100	

*Includes DOE laboratories' nonresearch costs but not U.S. government administration or SBIR.

**DOE contribution to the international Human Frontiers Neurosciences Program.



Human Genome Project

JSR 97-315

Study leader: S. Koonin

JASON contributors include:

S. Block	H. Kimble
J. Cornwall	N. Lewis
W. Dally	C. Max
F. Dyson	T. Prince
N. Fortson	R. Schwitters
G. Joyce	P. Weinberger
W. Woodin	

September, 1997

1. BACKGROUND, CHARGE, AND RECOMMENDATIONS	3
1.1 OVERVIEW OF THE HUMAN GENOME PROJECT.....	3
1.2 CHALLENGES FOR THE PROGRAM.....	3
1.2.1 <i>The complexity of genomic data</i>	4
1.2.2 <i>The state of technology</i>	4
1.2.3 <i>The pace of sequencing</i>	5
1.2.4 <i>The cost of sequencing</i>	6
1.2.5 <i>Project coordination</i>	7
1.3 STUDY CHARGE	7
1.4 RECOMMENDATIONS	8
1.4.1 <i>General recommendations</i>	8
1.4.2 <i>Technology recommendations</i>	8
1.4.3 <i>Quality recommendations</i>	9
1.4.4 <i>Informatics recommendations</i>	10
2. TECHNOLOGY	10
2.1 IMPROVEMENTS OF PRESENT GENOMICS TECHNOLOGY	11
2.1.1 <i>Electrophoresis improvements and an ABI Users Group</i>	11
2.1.2 <i>Algorithms</i>	13
2.1.3 <i>A method to bypass assembly</i>	15
2.2 DOE'S MISSION FOR ADVANCED SEQUENCING TECHNOLOGY	17
2.2.1 <i>Institutional barriers to advanced technology development</i>	18
2.2.2 <i>Purposes of advanced sequencing technology</i>	19
2.3 SPECIFIC ADVANCED TECHNOLOGIES	20
2.3.1 <i>Single-molecule sequencing</i>	20
2.3.2 <i>Mass-spectrometric sequencing</i>	24
2.3.3 <i>Hybridization arrays</i>	26
3. QUALITY	28
3.1 QUALITY REQUIREMENTS.....	29
3.1.1 <i>The diversity of quality requirements</i>	30
3.1.2 <i>Accuracy required for assembly</i>	31
3.2 VERIFICATION PROTOCOLS.....	34
3.2.1 <i>Restriction enzyme verification of sequence accuracy</i>	34
3.2.2 <i>Hybridization arrays for sequence verification</i>	37
3.2.3 <i>Implementation of verification protocols</i>	40
3.3 ASSESSING AND IMPROVING PRESENT TECHNIQUES.....	40
3.3.1 <i>A systems approach is required</i>	41
3.3.2 <i>"Gold standards" for measuring sequence accuracy</i>	42
3.3.3 <i>Quality issues pertaining to sequencing templates</i>	43
4. GENOME INFORMATICS	44
4.1 INTRODUCTION.....	44
4.2 DATABASES.....	47
4.2.1 <i>User issues</i>	48
4.2.2 <i>Modularity and standards</i>	49
4.2.3 <i>Scaling and storage</i>	50
4.2.4 <i>Archiving raw data</i>	51
4.2.5 <i>Measures of success</i>	52
4.3 SOCIOLOGICAL ISSUES.....	53

1. Background, charge, and recommendations

1.1 Overview of the Human Genome Project

The US Human Genome Project (the "Project") is a joint DOE/NIH effort that was formally initiated in 1990. Its stated goal is

"...to characterize all the human genetic material--the genome--by improving existing human genetic maps, constructing physical maps of entire chromosomes, and ultimately determining the complete sequence... to discover all of the more than 50,000 human genes and render them accessible for further biological study."

The original 5-year plan was updated and modified in 1993 [F. Collins and D. Galas, "A new five-year plan for the US Human Genome Project," *Science* 262, 43-46 (1993)]. The Project's goals to be achieved by the end of FY98 that are relevant for this study are:

- To complete an STS (Sequence Tagged Site) map of the entire genome at 100 kb resolution
- To develop approaches for sequencing Mb regions
- To develop technology for high-throughput sequencing, considering the process as integral from template preparation to data analysis.
- To achieve a large-scale sequencing capacity of 50 Mb/yr and to have completed 80 Mb of human sequence
- To develop methods for identifying and locating genes
- To develop and disseminate software to archive, compare, and interpret genomic data

Congress has authorized funding through the planned completion of the Project in FY05. The funding in FY97 is \$189M for the NIH activity and \$78M for the DOE. Thus the total US effort is \$267M this year. This amounts to more than half of the worldwide effort, with France, UK, the EU, and Japan being the other major partners.

The DOE program in FY97 included \$29 M for production sequencing, \$15 M for the startup of the Joint Genome Institute (a "factory scale" sequencing facility to be operated jointly by LLNL, LANL, and LBNL), \$13 M for technology development, \$11 M for informatics, and \$3M for applications (construction of cDNA libraries, studying gene function, etc.)

1.2 Challenges for the Program

There are a number of challenges that the program faces if it is to meet its stated goals. We briefly describe several of them in this section as a background to our charge.

1.2.1 The complexity of genomic data

One of the challenges to understanding the genome is the sheer complexity of genomic data. Not all sequence is equivalent. The 3-5% of the genome that is coding consists of triplet codons that specify amino acid sequence. The control regions are binding sites for regulatory proteins that control gene expression. The functions of the introns within a gene and the intergenic regions are largely unknown, even though they comprise the bulk of the genome. There are also special structural elements (centromeres and telomeres) that have characteristic base patterns.

Even given the sequence, the genes are not manifest. And the function and control of a particular gene (When and where is it expressed? What is the function of the protein it encodes?) generally must be determined from the biological context, information beyond the bare sequence itself.

Yet another challenge is that the genomes of any two individuals (except of identical twins) are different (10^{-3} in the coding region; unknown in the non-coding regions), and that the homologies between organisms are invariably less than perfect.

Many of these difficulties arise because we don't yet understand the language of the genome. A good metaphor for the state of genetic information is "It's like going to the opera." That is, it's clear something substantial is happening and oftimes it's quite beautiful. Yet we can't really know what's going on because we don't understand the language.

1.2.2 The state of technology

Another hurdle for the project is the state of technology. The present state of the art is defined by Sanger sequencing, with fragments labeled by fluorescent dyes and separated in length by gel electrophoresis (EP). A basic deficiency of the present technology is its limited read-length capability (the number of contiguous bases that can be read); best current practice can achieve 700-800 bases, with perhaps 1000 bases being the ultimate limit. Since interesting sequence lengths are much longer than this (40 kb for a cosmid clone, 100 kb or more for a gene), the present technology requires that long lengths of DNA be fragmented into overlapping short segments (~1 kb long) that can be sequenced directly. These shorter reads must then be assembled into the final sequence. Much of the current effort at some sequence centers (up to 50%) goes into the assembly and finishing of sequence (closing gaps, untangling compressions, handling repeats, etc.). Hence, longer read lengths would greatly step up the pace and quality of sequencing.

However, it is important to realize that, beyond the various genome projects, there is little pressure for longer read lengths. The 500-700 base reads allowed by the current

technology are well-suited to many scientific needs (pharmaceutical searches, studies of some polymorphisms, studies of some genetic diseases). Thus, the goal of the entire sequence implies unique technology needs, for which there are no medical or pharmaceutical drivers.

Other drawbacks of the present technology include the time- and labor-intensive nature of gel preparation and running and the comparatively large sample amounts required to sequence. This latter influences the cost of reagents involved, as well as the necessity for extra PCR steps.

1.2.3 The pace of sequencing

One regularly updated “score card” of the Human Genome Project is maintained at http://weber.u.washington.edu/~roach/human_genome_progress2.htm. This site regularly updates its tallies from the standard human genome databases. As of 5/1/97, there was some 39 Mb of human sequence in contigs of 10 kb or longer; this has been accumulated over the past 20 years. Although 98.7% of the genome thus remains to be sequenced, 15 Mb have been added in the past year. Figure 1 below shows the progress in the past few years.

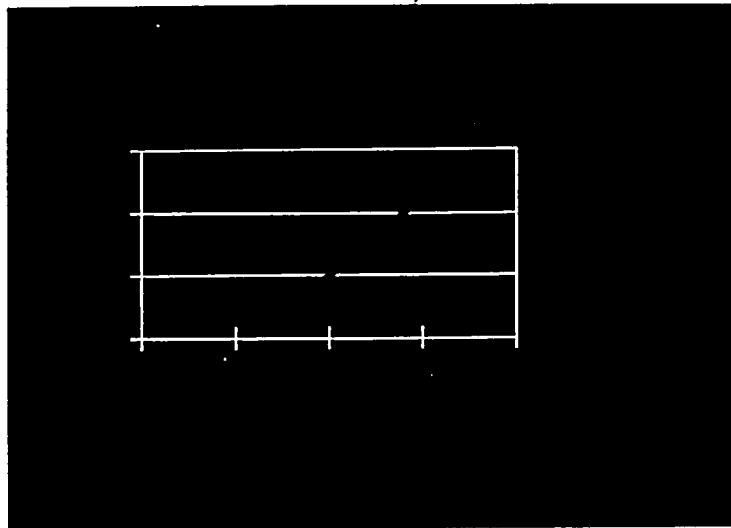


Figure 1: Fraction of the human genome in contigs longer than 10 kb that is deposited in publically accessible databases.

The world’s large-scale sequencing capacity is estimated to be roughly 20 Mb/yr; although not all of this resource is applied to the human genome. The Joint Genome Institute is projected to have a sequencing capacity of 57 Mb/yr in FY98, growing to 250 Mb/yr in FY01. These capacities are to be compared with the Project’s 9/98 goal of 50 Mb/yr.

It is sobering to contemplate that an average daily production of 400 Mb/yr is required to complete the sequence "on time" (i.e., by FY05); this corresponds to a daily generation of 50,000 samples and 15 Gbytes of raw data (if the EP traces are archived). Alternatively, if a geometric increase in production is assumed, the doubling time must be 15 months. These figures correspond to a scale-up of the present capacity by a factor of 30-100. Most observers believe that significant technology advances will be required to meet the FY05 goal.

The length of the known human sequences is also important. The Project's goal is the contiguous sequence of the entire genome. The table below (taken from http://weber.u.washington.edu/~roach/human_genome_progress2.htm) shows the number of known contiguous segments that are equal to or greater than a specified cut-off length. Note that only 1/3 of the known sequence is in lengths of 100 kb or greater, and that the longest human contig is about 1 Mb. It should also be noted that there are many known sequences of several hundred bases or less, for cDNA fragments of this size are generated at a prodigious rate in the pharmaceutical industry. (We heard of one company, Incyte, which produces 8 Mb of raw sequence each day, albeit in small fragments.)

Length cutoff (kb)	Contigs longer than cutoff	Sequence in contigs (Mb)
100	112*	16.15
50	191	22.06
40	302	26.82
30	494	33.72
20	579	35.85
10	782	38.66
5	1227	41.72
1	5283	50.50
0.1	very many	---

1.2.4 The cost of sequencing

The cost of sequencing is also a major consideration. If funding continues at the present rate over the next 8 years, the US project will spend some \$2.5B. If all of this funding were devoted to production sequencing, a cost of roughly \$1 per base would suffice. But only a fraction of it is.

Several cost benchmarks are available. The tenth complete microbial genome (*Bacillus subtilis*) has just been announced. It consists of 4000 genes in 4.2 Mb of sequence. This joint European/Japanese project cost something over \$2 per base sequenced. Best practice in the Human Genome Project is currently \$0.5/base, and the project goal is less than \$0.10/base. Specific plans for the Joint Genome Center project an initial (FY97) cost of \$0.60 per base, falling to \$0.10 per base by FY01. It should be noted that there is difficulty in comparing the costs claimed across laboratories, and across the different funding systems in different nations.

1.2.5 Project coordination

The Human Genome Project presents an unprecedented set of organizational challenges for the biology community. Success will require setting objective and quantitative standards for sequencing costs (capital, labor, and operations) and sequencing output (error rate, continuity, and amount). It will also require coordinating the efforts of many small laboratories supported by multiple funding sources in the US and abroad.

A number of diverse scientific fields have successfully adapted to a "Big Science" mode of operation (nuclear and particle physics, space and planetary science, astronomy, and oceanography being among the prominent examples). Such transitions have not been easy on the scientists involved. However, in essentially all cases the need to construct and allocate scarce facilities has been an important organizing factor. No such centripetal force is apparent (or likely) in the genomics community, although the Project is very much in need of the coordination it would produce.

1.3 Study charge

Our study was focused on three broad areas:

- **Technology:** Survey the state-of-the-art in sequencing. What are the alternatives beyond gel electrophoresis? What strategies should be used for inserting new technologies into production sequencing? What are the broader uses of sequencing technologies? What are the technology needs beyond those of the Human Genome Project?
- **Quality Assurance and Quality Control:** What are the "lust to dust" QA/QC issues and needs of the Human Genome project? What levels of sequence quality are required by various users of genome data? What steps can be taken to ensure these various levels of quality?
- **Informatics:** Survey the current database issues, including data integrity, submission, annotation and usability? What is the current state of algorithm development for finishing and annotating sequence?

Beyond briefings focused on these specific topics, we also heard a variety of speakers on functional genomics, in order to better get a sense of the needs, standards, and expectations of the consumers of genomic information.

Our recommendations in response to this charge are given in the following section. The balance of this report provides the necessary context and detail, dealing successively with Technology (Section 2), Quality (Section 3), and Informatics (Section 4).

1.4 Recommendations

1.4.1 General recommendations

We begin with two recommendations pertinent to many aspects of the Human Genome Project.

“Know thy system”

It is important to have a comprehensive, intimate, and detailed understanding of the sequencing process and the uses of genomic data. Gaining such understanding is a quite different exercise from sequencing itself. Answers to questions such as “What are the pacing factors in production sequencing?” (cloning? gel prep? run time?, lane count?, read length?, ...) or “What is the sequence error budget?” or “What quality of sequence is required?” are essential to optimizing the Project’s utility and use of resources.

Couple users/providers of technology, sequence, data

The Human Genome Project involves technology development, production sequencing, and sequence utilization. Greater coupling of these three areas can only improve the Project. Technology development should be coordinated with the needs and problems of production sequencing, while sequence generation and informatics tools must address the needs of data users. Promotion of such coupling is an important role for the funding agencies.

1.4.2 Technology recommendations

Technology development should be emphasized as a DOE strength

Technology development is essential if the Human Genome Project is to meet its cost, schedule, and quality goals. DOE technology development leverages traditional and extensive Department expertise in the physical sciences, engineering, and the life sciences. These are, in many ways, complementary to NIH strengths and interests. If the DOE does not continue to play a leading role in technology development for high-throughput, high-capacity sequencing, it is not clear to us who will.

Continue work to improve present technologies

Although a number of advanced sequencing technologies look promising, none are sufficiently mature to be candidates for the near-term major scale-up needed. Thus, it is important to support research aimed at improving the present Sanger/EP effort. There are clear hardware and software opportunities for improving gel reading capabilities; formation of an ABI user group might accelerate the realization and dissemination of these improvements. There are also software opportunities to improve the crucial assembly and finishing processes. And, as we demonstrate in Section 2.1.3, there is still room to invent promising sequencing protocols.

Enhance long-term technology research

The present sequencing technology leaves much to be desired and *must* be supplanted in the long term if the potential for genomic science is to be fully realized. Promising directions at present for advanced technology development include single-molecule sequencing, mass spectrometric methods, hybridization arrays, and micro-fluidic capabilities. The total annual funding for advanced technology (in FY97 only \$1.7M of the roughly \$11M total technology funding in the overall \$78M DOE Human Genome Project), should be increased by approximately 50%.

Retain technology flexibility in production sequencing facilities

Because sequencing technology should (and is likely to) evolve rapidly (ideally, both evolutionary and revolutionary changes will occur before FY05) it is important to retain the flexibility to insert new technologies into the large-scale sequencing operations now being created (e.g., the Joint Genome Center). The decisions of when to freeze technology and how much upgrade flexibility to retain are faced in most large scientific projects (i.e., spacecraft, accelerators, ...) and, unfortunately we have no magic prescription for dealing with them. However, the common sense steps of building in modularity and of thoroughly and frequently scanning the technology horizon are well worth remembering.

1.4.3 Quality recommendations

Work to make quality considerations an integral part of the Project

Quality issues must be brought to the fore in the sequencing community, since measures of sequence quality will greatly enhance the utility of the Human Genome Project's "product." Among the top-level steps that should be taken are allocating resources specifically for quality issues and establishing a separate QA/QC research program (perhaps a group at each sequencing center).

Quantify QA/QC issues

Promote research aimed at quantifying (through simulation and other methods) the accuracy required by various end uses of genomic data. Further, since accuracy is a full-systems issue, there is the need for a comprehensive, end-to-end analysis of the error budget and error propagation in the sequencing process, from clone library development through sequencing to databases and analysis software. "You can't discuss it if you can't quantify it."

Develop and implement QA/QC protocols

Develop, distribute, and use "gold standard" tests of sequencing centers. Support research aimed at developing, validating, and implementing useful verification protocols, along the lines discussed in Section 3.2. Make quality assessments an integral part of all database sequence. A good start would be to require that all database entries include quality scores for each base call. Existing sequencing software tools such as PHRED,

PHRAP, and CONSED produce figures of merit for base calls and DNA assembly. While there is room for innovative research aimed at improving the basis for these figures of merit, the existing confidence indicators are nevertheless quite informative and should be made available to users of sequence data.

1.4.4 Informatics recommendations

Listen to the customers

Adhere to a “bottom-up”, “customer” approach to informatics efforts supported by DOE. Encourage forums, including close collaborative programs, between the users and providers of DOE-supported informatics tools, with the purposes of determining what tools are needed and of training researchers in the use of new tools and methods. Further, critically evaluate DOE-supported informatics centers with regards to the actual use of their information and services by the community.

Encourage standardization

Encourage the standardization of data formats, software components and nomenclature across the community. Invest in translators if multiple formats exist. Modularize the functions of data archiving, data retrieval, and data manipulation. Distribute the effort for development across several groups. Standardization of data formats allows more than one group to work in each area.

Maintain flexibility

Do *not* demand that “one-size” (in databases) fits all. Make it easy to perform the most common operations and queries, but do not make it impossible for the expert user to execute complicated operations on the data. The community should be supporting several database efforts and promoting standardized interfaces and tools among those efforts.

2. Technology

The technology to sequence the human genome is now in hand. Indeed, this was true when the project was formulated and initiated in 1990, and there have been significant improvements in the intervening 7 years. Nevertheless, as we have noted in Sections 1.2.2-4, there are ample reasons to improve the present technology, particularly if the Project’s cost, schedule, and quality goals are to be achieved. Further, improvements in sequencing technology will accelerate genomics research and applications beyond human biology and medicine.

The Project faces the classic dilemma inherent in any large technological project: when to freeze the technology available, to declare “good enough” at the risk of not pursuing the “better.” We believe that the likely inadequacy and ease of improvement of the present technology and the future importance and relative inexpense of developing

radically different technology all argue for pursuing both tracks simultaneously. Our rationale is presented in the following sections.

2.1 Improvements of present genomics technology

In the course of our study, we identified three aspects of the present sequencing technology where improvements that could have a significant impact seemed possible. These are

- Electrophoresis
- Algorithms for base calling, assembly, and finishing
- Bypassing assembly by ordering the sequence of read-length fragments

We consider each of these in turn.

2.1.1 Electrophoresis improvements and an ABI Users Group

The Applied Biosystems Inc. (ABI) automated DNA sequencers are the *de facto* standard for sequencing and will almost certainly carry the brunt of the sequencing load for the Project. These are “closed-box” instruments that utilize proprietary technology owned exclusively by ABI. The company has both the responsibility and the financial incentive to ensure reliable, standardized operation of its instruments, even if this results in sequencing that is less than optimal. On the other hand, the desire of many end users, especially those at major genome sequencing centers, is to push the performance of these instruments to the limit.

This tension raises both issues of technology *per se* and of how new technology can be inserted in ABI machines to the satisfaction of all. We first discuss possible technology improvements, then propose a users group.

It is clear that modifications could be made to the hardware, and especially the software, of the ABI sequencers without sacrificing accuracy of base calling or reliability of operation; one of our briefers spoke convincingly to this issue [C. Tibbets, briefing to JASON, July 1, 1997]. These instruments use the Sanger sequencing method to sample automatically molecules labeled with any of four (ABI-proprietary) fluorescent dyes. The samples undergo gel EP in 36 lanes. The lanes are scanned with an argon laser and bases are “called” by a combination of hardware and software.

Errors can (and do) arise from a number of sources, including lane tracking; differential migration of the four dyes; overlapping emission spectra of the dyes; and variable oligomer separations, due, for example, to secondary sources. There are a number of efforts underway to improve the software packages used for interpreting the (trace) data stream produced by the sequencing instrument. It is important to note that specific improvements might have a dramatic impact of the Project, but be of marginal significance for broad classes of commercial applications. One example is attaining longer read lengths.

Specific areas with clear potential for significant improvement include:

- increasing the lateral scan resolution, thus allowing for more lanes;
- indexing the lateral scan in space (instead of time) for greater trace precision and reproducibility;
- adding a fifth dye for enhanced lane tracking;
- allowing access to the raw (preprocessed) trace data, thus enabling improved base calling algorithms.

ABI has no obligation to respond to users' requests for modifications such as those suggested above, nor are they required to make available detailed specifications that would allow users to make such modifications themselves. As a result, advanced users are taking matters into their own hands through reverse engineering, even if this risks invalidating the manufacturer's warranty or service agreement. For both legal and sociological reasons these aftermarket modifications tend to be made at the level of individual genome centers. This may result in fragmentation of the standards of practice for acquisition of sequence data, complicating the establishment of quality-control measures across the entire genomics community.

It would be desirable to unify the genomics community's efforts to enhance the performance of ABI instruments, without infringing on ABI's right to control its products and to guard its proprietary technology. We recommend that DOE take an active role in setting up an ABI "Users Group" that would serve as a sounding board for issues pertaining to the operation of existing instruments, the modification of existing instruments for enhanced performance, and the development of next-generation instruments. The group would include members from each of the major genome centers, various private genomics companies that choose to participate, and a sampling of small-scale users who receive federal support for DNA sequencing activities. The group should also include a representative from DOE, NIH, and (if it wishes to participate) ABI itself.

The activities of the users' group should be self-determined, but might include in-person or electronic meetings, generation of reports or recommendations concerning the operation and potential improvement of the ABI instruments, and distribution of information to the scientific community via journal articles or the World Wide Web. DOE should provide principal funding for these activities, although industry members and ABI should pay expenses related to their own participation. It must be understood by all participants that ABI is under no obligation to consider or follow the recommendations of the users' group. We would expect, however, that by finding common ground and speaking with one voice, the users will have substantial impact on the improvement of automated DNA sequencing technology, while maintaining common standards of practice across the genomics field and respecting the proprietary rights to sequencing technology.

2.1.2 Algorithms

Algorithms, and the software packages in which they are embodied, for lane tracking, base calling, assembly, and finishing appear to be in a formative stage. Research into new algorithms, and development and dissemination of software packages containing them, can return significant dividends in terms of both productivity and accuracy.

2.1.2.1 *Base calling:*

The base calling problem involves converting a four-channel record of dye fluorescence intensity to a sequence of bases along with a confidence value for each base. Several factors make this a challenging problem. Spreading of the intensity function along the lane leads to inter-symbol interference. Overlap in the spectral response of the four dyes leads to cross-talk. The spacing between bases may be non-uniform, certain sequences of bases distort the record, and the signal levels are very low toward the end of a read.

All of the problems present in base calling are also present in the demodulation of signals in communication and magnetic recording systems. As a result, there is a rich literature of methods for dealing with these problem. For example, inter-symbol interference can be reduced, by employing linear equalization or decision-feedback equalization. Clock-recovery methods can be applied to keep the base calls properly centered. Sequences can be decoded as multi-base symbols to compensate for sequence-dependent distortion. A trellis decoder or a hidden Markov model can be employed to exploit knowledge about expected sequences to compute the most likely sequence to be generated by a particular intensity record. It would be worthwhile to consider implementing new (or improving present) base calling algorithms on the basis of these techniques.

2.1.2.2 *Assembly:*

Assembly algorithms stitch together a set of sequences (of perhaps 500 bases each) that are subsequences of a clone (of perhaps 30 kb in length) to generate the (hopefully) complete sequence of the clone. The process is similar to assembling a linear puzzle where the pieces are allowed to overlap arbitrarily. We saw considerable variability in the methods used for assembly. The PHRAP program uses a *greedy* algorithm where the segments with the closest matches are assembled first and the program builds out from this initial start. The group at Whitehead, on the other hand, uses an algorithm based on tags to find overlapping segments. All of these algorithms are heuristic and approximate, as a complete search for the optimum map is perceived to require excessive computation.

There are many directions for research on assembly algorithms. To start, better methods for comparing two sequences to determine if they match can be employed. The PHRAP program achieves more accurate assembly by using base-call confidence values in grading matches. This corresponds exactly to the use of soft-decision decoding in a communication system. One can further improve the accuracy of matching by taking into account the sequence-dependent probability of erasures and insertions, computing, for example, the probability of a compression based on the surrounding GC-rich sequence. Similar techniques can be used to handle assembly in the presence of repeats.

Better methods for searching the space of possible assemblies can also be developed. For example, the greedy algorithm employed by PHRAP can get stuck if it makes a wrong choice early in its processing. One should benchmark such algorithms against a complete branch-and-bound search on representative difficult sequences to determine how often such failures occur. If there is a significant advantage to a full search, one can construct special-purpose assembly computers to perform this computation in a reasonable amount of time. For example, one could use an ASIC or a few FPGAs to build an accelerator that plugs into a standard workstation that will compute (in less than a microsecond) matching scores for all shifts of two segments through an algorithm that employs confidence values and sequence-dependent insertions and deletions. Even with a complete search, the use of heuristics is important to guide the search to explore the most likely assemblies first, so that large parts of the search space can be pruned.

2.1.2.3 Finishing:

The finishing process involves taking an assembled sequence and filling in the gaps through a combination of manual editing and directed sequencing. At some sequencing centers we saw that finishing accounted for roughly half of the entire sequencing effort. Yet the software available to assist finishing consisted of no more than simple sequence editors. Research into finishing software has the potential to automate much of this labor-intensive process.

The first step toward automated finishing is to improve assembly software. Generating a correct assembly without manual intervention would eliminate much of the need for manual editing, leaving only the genuine gaps to be filled using directed sequencing.

The directed sequencing process involves ordering new reads of the clone using primers designed to extend the ends of sections that have already been sequenced. Much of this process can be automated using a rule-based expert system. Such a system is built by having a knowledge engineer observe an expert finisher at work and capture the finisher's thought process in a set of rules: e.g.,

when a contig of a particular length is terminated in a particular way at each end, order a set of primers that match part of the sequence and order new reads taken using these primers and dye-terminator sequencing.

By combining the approaches taken by several finishers from different centers, the system could, in some cases, outperform a single human finisher. At the very least, a set of a few hundred of these rules would be likely to cover most of the common finishing cases. This would allow the human experts to focus their effort only on the most difficult cases.

2.1.3 A method to bypass assembly

In this section, we propose a prototypical strategy for a particular aspect of technology development: to order the fragments of appropriate read lengths that are created from a cosmid. The aim of this approach is to increase the information throughput, and to increase the effective read length of a cloned DNA strand, while staying within the constraints on an individual sequencing run that are dictated by existing gel EP technology (or, for that matter, by such advanced technologies as mass spectrometry).

In shotgun sequencing, the key problem is that the lack of information about the spatial origin of any individual sequenced region dictates that overlapping 500 base length reads must be obtained to enable successful assembly. This typically requires 7-fold redundancy in coverage of a given area. Longer effective reads would clearly be useful in reducing this redundancy and would thus increase throughput. In addition, because much (up to 60% in many cases) of the sequencing cost in current genome center operations is incurred in the assembly and finishing processes, elimination of assembly could allow a significant reduction in the cost per base pair of the overall sequencing effort.

The scheme proposed here outlines a prototypical process that retains the positional information of any 500 base sequence read with conventional EP technology. One would start with, for example, a 40 kb cosmid clone and bind one end (in this example, the 5' terminus) uniformly to a solid support. The bound DNA would then be exposed to an exonuclease. There are exonucleases that can digest DNA from either the 3' or the 5' terminus, but since the 5' end in our example is bound to the support, we require an exonuclease that digests from the 3' terminus. The goal of the digestion step is to obtain an ordered, binned distribution of lengths of DNA, with some strands being systematically longer than others. One possible approach is to utilize a time or temperature gradient in the digestion, so that more digestion occurs for strands that are located further down the solid support. In another implementation, no support is needed and the digestion can be done in solution, with aliquots withdrawn at various times; those samples subjected to more digestion time will naturally be shorter in length. In either implementation, one would adjust digestion conditions such that each successive

binned sample contains DNA strands that are progressively shorter in length by approximately 450 base pairs. If needed, the sizes of the DNA to be placed in these bins can also be obtained by a sizing gel electrophoresis step, with bands of the appropriate length physically transferred into separate solutions.

In the second phase of the process, a common end tag (for instance, TTTTTTTTTTTT) is attached to each of the various binned fragments. Since all of the strands then have a common end sequence, a common primer (in our example, AAAAAAAAAAAA) can be used to amplify all of the differently sized, binned, strands by PCR. The strands in each bin would be amplified in parallel and then sequenced from the 3' end using the Sanger dideoxy sequencing methodology. Each individual read will only be 500 bases in length, but because each read bin is, by design, progressively 450 bases shorter than the last read bin, we know the ordering on the original strand of each 500 base read (so that no assembly is required), and we obtain an estimate of the error rate of the sequencing step based on the base calling agreement observed in the 50-base regions common to the digested DNA located in adjacent bins.

This procedure resembles primer walking, except that it uses a common primer for each 500 base region to be sequenced. Furthermore, all of the amplification and extension steps can be performed in parallel instead of in series, as is required for primer walking. This likely significantly reduces the cost and complexity of the process as compared to primer walking, which is a method of last resort. The binning procedure is the key to retaining the positional information regarding where an individual 500 base pair sequence resides on the original 40 kb clone.

This is clearly only one of several possible binning strategies. Another strategy would involve digestion from one end, binning of fragments, digestion from the other end, and then sequencing only the small 500 base pair sequences that remained in each bin. This method also might be useful as a directed sequencing method in which one wants to target specific regions that are not readily assembled using shotgun strategies.

We have not developed precise details of any binning strategy here, but believe that methods could be devised to increase the effective read length of a clone while still using currently available gel electrophoresis sequencing methods for individual reads. The common theme of such approaches is to combine the overarching strategies of binning protocols, whose main advantage is that they retain the positional information of an individual read sequence with respect to a longer contiguous piece of DNA, with the capabilities and familiarity of conventional gel EP methods.

A remaining issue is the cost of such a process relative to conventional shotgun sequencing. Experimental data on the efficiency of digestion, the length distribution of the digested binned fragments, and other related variables would be required to accurately estimate the cost per base pair of any positionally-binned sequencing method. However, it seems that the strategy is worth exploring on a pilot basis to obtain such cost estimates.

We note that this is but one strategy developed in the course of a one-month study by non-experts. It is likely that other such ideas would be developed as the importance, methods, and problems of sequencing become known to a broader community of scientists.

2.2 DOE's mission for advanced sequencing technology

We heard briefings from nine experts, describing various technologies that might bring radical improvements to the art of sequencing DNA. These are discussed in some detail below. They are all different, but they have several features in common. They are small-scale, and currently absorb a small fraction of the DOE genome project budget (some \$1.7 M of the \$13 M DOE technology development budget); unfortunately, they are scheduled to receive even less in the future. These projects are long-range, aimed at developing technologies whose greatest use will be come in the sequel of applications following the initial sequencing of the human genome. They are, to some extent, high-risk, exploring ways to overcome obstacles that could prove to be insuperable. But they also are high-promise, offering a real possibility of new sequencing methods that would be significantly faster and cheaper than gel EP.

How much money should DOE spend on high-risk, high-promise ventures? This is one of the important questions addressed by our study. We recommend a gradual increase of funding for technology development by about 50% (to \$20 M per year) with a substantial fraction of this money going to projects other than improvements in current gel EP techniques. One should be prepared to increase this level rapidly in case one or more of the new technologies becomes ripe for large-scale operation.

In making this recommendation for increased support for advanced technologies, we are well aware of the need for the DOE to play a significant role in the current stage of the Project. We also know of, and approve of, the technology goals of vastly improving current EP techniques by such means as high-voltage capillaries, ultrathin gels, and use of resonance ionization spectroscopy. It is likely that such improvements in gel EP are essential to completing the genome project on time, and we have commented in Section 2.1 on improving gel EP throughput in the near term. However, we believe that in the long run DOE's greatest impact will be in support of the development of advanced technology for various sequencing tasks that go beyond the current goals of faster gel EP.

There are two main reasons for DOE to support these high-risk technologies. First, this is the part of the Project that DOE can do better than NIH. The history and traditions of DOE make it appropriate for DOE to explore new sequencing technologies based on the physical sciences. Second, existing gel EP technology is barely adequate for sequencing a single human genome, and new technologies will be required to satisfy the future needs of medicine, biological research, and environmental monitoring. The new ventures supported by DOE are the seed-corn of sequencing efforts, for a crop to be reaped far beyond the Project itself.

2.2.1 Institutional barriers to advanced technology development

Most of the current attention in the Project is currently focused on rapid, low-cost sequencing of a representative human genome, to be finished by FY05. As a result, there has been a tendency to freeze technology at a fairly early level of development, sometimes not much past the proof-of-principle level, in order to cut down lead times. This tendency is exacerbated by the subsequent commercialization of the technology, making it difficult, for the usual property-rights reasons, to incorporate improvements found by those outside the commercial sector. Even this would not be so bad if it were not that the majority of genome researchers are not oriented toward technology development *per se*, but to the biological research that the technology enables. There is a vicious circle in which lack of technology support by agencies such as NIH leads to an insufficient technology knowledge base among the supported researchers, while this lack of knowledge among peer reviewers leads to a reluctance to support technology development.

2.2.1.1 A parallel in ultrasound technology development

Three years ago, a JASON study sponsored by DARPA [H. Abarbanel *et al.*, Biomedical Imaging (JASON Report JSR-94-120, August 1995)] looked at the maturity and sophistication of technology both for ultrasound and for MRI. In both cases the study found concrete examples of the institutional barriers discussed in the previous section. Ultrasound was further behind in advanced technology than MRI, and we will comment only on ultrasound here. The problems of ultrasound are well-known to all who work in it: The transmission medium (flesh and bones) is so irregular that images have very poor quality, interpretable only by those devoting their lifetime to it. In-principle improvements were known, especially the construction of two-dimensional ultrasound arrays to replace the universally-used one-dimensional arrays (which severely degrade the resolution in the direction transverse to the array). But this was a difficult technological challenge, requiring sophisticated engineering beyond the reach of much of the ultrasound community, and not representing an obvious profit potential for the commercial suppliers.

The JASON study found that members of the ultrasound research community were largely limited by the pace of commercial technology development, which was conservative and market-oriented, not research-oriented. In some cases there were ultrasound researchers quite capable of making advances in the technology, but frustrated by the lack of NIH funding. The study recommended that DARPA occupy, at least temporarily, the niche of technology development for ultrasound, which existed because agencies like the NIH were not filling it.

In response to this study, DARPA put a considerable amount of money into advancing ultrasound technology, with emphasis on using (two-dimensional) focal-plane array techniques developed by defense contractors for infrared and other electrooptical arrays. While it is too early to foresee the ultimate impact, it appears that this funding will significantly improve ultrasound technology.

2.2.2 Purposes of advanced sequencing technology

The goal of sequencing 3 billion base pairs of a representative human genome requires a limited amount of redundancy (perhaps a factor of 10) to insure complete coverage and improve accuracy. However, further developments in genomics will have to address questions of diversity, rarity, and genomic function, which may make this sequencing effort seem small.

One can imagine the need to go from sequencing 3×10^9 base pairs per decade to sequencing this many or more per year, as diversity becomes the issue. Diversity arises from individual variation (RFLPs, VNTRs, and other manifestations of introns, mutations in genes, etc.) and from the desire to compare human genomes with those of other species, or to compare (parts of) one individual's genome with another's. If it is ever to become possible for MDs and laboratory technicians outside biotechnology laboratories to do sequencing routinely, the sequencing process itself will have to become much simpler, and not subject, for example, to fluctuations in the artistry of the experts who nowadays prepare gels. (Not everyone subscribes to such a goal, the alternative being large sequencing centers to which samples are submitted.). The databases that keep track of this diversity will grow correspondingly, as will the search engines needed to mine the databases. It is not out of the question to anticipate computing needs increasing even faster (a pairwise correlation search of a ten times larger database may require up to one hundred times more searching, for example).

The hunt for rare genes (associated perhaps with rare genetic diseases or obscure functions) may call for advanced technology for constructing and searching cDNA libraries, perhaps massively-parallel machinery built on a considerably smaller unit scale than is now common.

Functional genomics (to oversimplify, the understanding of the roles and interactions of the proteins coded for by DNA) presents difficulties so specific to each individual case study that it is nearly impossible to summarize here, and we will not attempt to do so. But it is clear that many functional genomics activities will require a total sequencing rate substantially beyond that of the present Project.

Advanced technologies also have a role to play in quality assurance and quality control. The chemical and physical bases of current sequencing technology result in intrinsic limitations and susceptibility to errors. Alternative sequencing methodologies at

least as accurate and efficient as the present one would allow independent verification of sequence accuracy. An example is given in Section 3.2.2 below.

Some advanced technology development will be done by commercial companies, to be sure, and that is to be welcomed, but if ultrasound or even the current state of the Project is a guide for the future, there is a most important role for DOE advocacy and support of advanced technology beyond the goals of initial sequencing of the human genome.

2.3 Specific advanced technologies

One cannot, of course, confidently predict the future of advanced technologies in any area. Instead, we comment in the following subsections on three directions that seem particularly promising:

- Single-molecule sequencing (by STM, AFM, flow cytometry, etc.)
- Mass-spectrometric sequencing
- Massively-parallel sequencing (hybridization arrays)

2.3.1 Single-molecule sequencing

For at least thirty years, some molecular biologists have been dreaming that it might be possible to sequence DNA molecules one at a time. To do this, three steps would need to be taken:

- **Step 1:** stretch out a molecule of DNA in a known orientation, with one end attached to a solid surface and the other end free.
- **Step 2:** detach nucleotides one at a time from the free end.
- **Step 3:** count and identify the nucleotides in order as they are released.

Before any of these three steps were mastered, the technique of sequencing DNA by gel EP was invented and the three steps became unnecessary - gel EP became the standard method of sequencing. A significant disadvantage of this method was the requirement for a macroscopic quantity of identical molecules as input. This requirement initially limited its application to viral genomes and other small pieces of DNA that could be obtained in pure form. A few years later, the invention of PCR made the preparation of pure macroscopic quantities of identical molecules routine and gel EP could then be applied to all kinds of DNA. Thus, the technology was ready for large-scale development when the Project began (indeed, its availability was one of the factors in initiating the project) and the technology of single-molecule sequencing was left far behind. [Single-molecule spectroscopy and related fields are nevertheless very active areas of research; see, for example, the symposium on Single Molecule Spectroscopy: New Systems and Methods, held last year in Ascona, Switzerland.]

The Human Genome Project has given only token support to single-molecule sequencing efforts. We heard about only two serious programs to develop single-molecule sequencing. One, at LANL, was described to us in a briefing by Richard Keller. The other, a proprietary program at seQ Ltd. in Princeton, was mentioned but not described in detail. Neither program is now supported by the Project. Details of the LANL program have been published [P. M. Goodwin, W. P. Ambrose, and R. A. Keller, "Single-molecule Detection in Liquids by Laser-Induced Fluorescence", *Accounts of Chemical Research*, **29**, 607-613 (1996); R. A. Keller *et al.*, "Single-Molecule Fluorescence Analysis in Solution", *Applied Spectroscopy*, **50**, 12A-32A (1996)]

Why should anybody be interested in single-molecule sequencing? There are two main reasons. First, each of the three steps required for single-molecule sequencing has recently been demonstrated to be feasible. Second, single-molecule sequencing, if all goes well, might turn out to be enormously faster and cheaper than EP. The following paragraphs explain the factual basis for these two statements.

The first step in single-molecule sequencing is the attachment of one end of a molecule to a solid surface and the stretching out of the rest of the molecule in a controlled manner. This has been done by the LANL team, using flow cytometry, a standard technique of microbiology. A single molecule of single-stranded DNA is attached by the covalent bonds of the biotin-avidin protein system to a plastic microsphere. The microsphere is held in an optical trap in a cylindrical fluid flow, which pulls the molecule straight along the cylinder's axis. The second step is the detachment of nucleotides in sequence from the end of the molecule. This has also been demonstrated by the LANL team, using standard microbiological techniques. Exonucleases are dissolved in the flowing fluid. A single exonuclease molecule attaches itself to the free end of the DNA and detaches nucleotides, one at a time, at a rapid rate (many per second).

The third step, the identification of bases in the detached nucleotides, is the most difficult. It might be done in at least three different ways. The LANL team identifies the bases by passing the flowing fluid through a laser-beam. As each base passes through the beam, the molecule fluoresces at a wavelength that is different for each of the four bases. Because the passage through the beam is rapid, the fluorescence must be intense if it is to be detected reliably. To intensify the fluorescence, the DNA molecule is initially prepared for sequencing by attaching a fluorescent dye residue to each base, with four species of dye marking the four species of base. The four types of base can then be identified unambiguously during roughly one millisecond that each nucleotide spends in the laser beam. Unfortunately, the LANL team has not succeeded in eliminating spurious detections arising from unwanted dye molecules in the fluid. They expect to be able to reduce the background of spurious events to a level low enough to allow accurate sequencing, but this remains to be demonstrated.

The seQ Ltd. team accomplishes the first two steps in the same way as the LANL team, but addresses the third step differently. The bases are not modified by addition of

dye residues. Instead, the unmodified nucleotides are detected by fluorescence in an ultraviolet laser-beam. Since the fluorescence of the unmodified bases is relatively weak, they must be exposed to the laser for a longer time. This is achieved by depositing each nucleotide, immediately after it is detached from the DNA, onto a moving solid surface. The surface is then scanned by ultraviolet lasers at a more leisurely pace, so that each nucleotide is exposed to the lasers long enough to be identified unambiguously. The details of this technique are proprietary, and we were not told how well it is actually working.

A third possible way to do the third step in single-molecule sequencing is to use mass spectrometry. The state of the art of mass spectrometry is discussed in Section 2.3.2. Mass-spectrometric identification of the detached nucleotides would require their transfer from the liquid phase into a vacuum. This might be done by ejecting the flowing liquid into a spray of small droplets, letting the droplets evaporate on a solid surface, and then moving the solid surface into a vacuum. Molecules sticking to the surface could then be detached and ionized by MALDI. Once ionized, they could be detected and identified in a mass-spectrograph, since the four species of nucleotide have different masses. (As noted in the next subsection, it is considerably more difficult to differentiate the four base pairs by mass than to distinguish their presence or absence, as in Sanger sequencing.) However, none of the mass-spectrograph projects that we heard about has addressed the problems of single-molecule sequencing.]

To summarize the present situation, each of the steps of single-molecule sequencing has been shown to be feasible, but no group has yet succeeded in putting all three together into a working system. The programs at LANL and seQ Ltd. are on a modest scale. Dr. Keller informs us that he is exploring the possibility of collaboration with a larger German-Swedish consortium headed by Manfred Eigen and Rudolf Rigler. The latter have published a plan for single-molecule sequencing essentially identical to the LANL program [M. Eigen and R. Rigler, *Proc. Nat. Acad. Sci. (USA)* 91, 5740 (1994)], although LANL is ahead of the consortium in the implementation of their plan. If the collaboration goes ahead, using the skills of LANL and supported by the larger resources of the consortium, there is a good chance that the plan can be developed into a practical system.

We turn now from the present situation to the future promise of single-molecule sequencing. The promise is that it might become radically faster and cheaper than gel electrophoresis. The claim that single-molecule sequencing might be extremely cheap stands or falls with the claim that it might be extremely fast. Sequencing by any method is likely to be a labor-intensive operation, with costs roughly proportional to the number of person-years devoted to it. The costs of machines and materials are likely to be comparable with the costs of wages and salaries. When we are concerned with large-scale operations, the number of bases sequenced per dollar will be roughly proportional to the number of bases sequenced per hour. The main reason why gel electrophoresis is expensive is that it is slow. If single-molecule sequencing can be a hundred times faster than gel electrophoresis, then it is also likely to be a hundred times cheaper.

The claim that single-molecule sequencing might be a hundred times faster than gel electrophoresis rests on a comparison of the inherent speeds of the two processes. The process of gel electrophoresis requires about eight hours to separate molecules with resolution sufficient to sequence 500 bases per lane. The inherent speed of gel electrophoresis is thus less than one base per minute per lane. In contrast, the elementary steps in single-base sequencing might have rates of the order of a hundred bases per second. The digestion of nucleotides in sequence from the end of a DNA molecule by exonuclease enzymes has been observed to occur at rates exceeding one hundred bases per second. And the discrimination of bases in ionized molecules detected by a mass-spectrometer can certainly be done at rates of hundreds of molecules per second. These facts are the basis for hoping that the whole process of single-molecule sequencing might be done at a rate of a hundred bases per second. That would imply that an entire human genome could in principle be sequenced by a single machine operating for a year.

Needless to say, this possibility is very far from being demonstrated. The three steps of single-molecule sequencing have not yet been integrated into a working process. And the rate of sequencing in a large-scale operation is limited by many factors beyond the rates of the elementary process involved. With either single-molecule or gel electrophoresis separation, the production of sequence will be slowed by the complicated manipulations required to prepare the molecules for sequencing and to assemble the sequences afterwards. Until single-molecule sequencing is developed into a complete system, no realistic estimate of its speed and cost can be made. The most that can be claimed is that single-molecule sequencing offers a possibility of radically increasing the speed and radically reducing the cost.

Two other potential advantages of single-base sequencing are longer reading-lengths and superior accuracy. The reading-length in gel EP is limited to about a thousand bases (roughly half of this in conventional practice). The LANL group has demonstrated attachment and suspension of single DNA molecules with many thousand bases. It is likely that DNA molecules with tens of thousands of bases could be handled, so that a single-molecule sequence could have a read length of tens of thousands of bases. As the short read length of gel EP makes final assembly and finishing an elaborate and costly process, these longer reads could greatly simplify the process of assembly..

One of the major obstacles to accurate sequencing is the prevalence in the genome of repeated sequences of many kinds. Repeated sequences are a frequent cause of ambiguities and errors in the assembly process. Since the single-molecule system will have longer read lengths, it will be less vulnerable to effects of repetition. Repeated sequences will usually be displayed, without ambiguity, within the compass of a single consecutive read. As a result, it is possible that single-base sequencing may be not only faster, but also more accurate than gel EP.

In addition to the LANL and seQ Ltd. programs and the mass-spectroscopy programs described in the following subsection, there are some efforts directed towards

single-molecule sequencing by non-destructive methods using microscopes. The idea of these efforts is to discriminate bases by scanning a DNA molecule with an Atomic Force Microscope or a Scanning Tunneling Microscope. These efforts are much further from practicality than the LANL and seQ Ltd. programs; we have not examined them in detail. Since the art of microscopy is advancing rapidly, it is possible that some new invention will make it possible to visualize individual bases in DNA with enough resolution to tell them apart. However, without a new invention, it appears that the existing microscope technology cannot do the job.

In conclusion, this study's recommendation is that DOE give modest support to single-molecule sequencing in general, and to the LANL program in particular. With modest support, there is a finite probability that single-molecule sequencing will be developed into a practical system within a few years. There is a smaller, but still finite, probability that it will prove to be superior to gel EP by a wide margin.

One can look at the support of single-molecule sequencing from two points of view. On the one hand, it is a gamble that DOE can afford to take, offering an opportunity to win a large pay-off by betting a small fraction of the genome budget. On the other hand, it is a premium that DOE can afford to pay for insurance against the possibility that the electrophoresis-based sequencing program might fail to reach its schedule, budget, and accuracy goals. From both points of view, modest support of single-molecule sequencing appears to be a prudent investment.

2.3.2 Mass-spectrometric sequencing

In the simplest terms, mass spectrometry (MS) in DNA sequencing replaces the gel EP step in Sanger sequencing. Instead of measuring the lengths of various dideoxy-terminated fragments by observing their rate of diffusion in a gel, one measures their mass with one of several possible MS techniques, including time-of-flight (TOF) and Fourier-transform ion cyclotron resonance (FTICR) spectroscopy. Presently, MS techniques are usable on fragments of about the same length as those used in gel EP (that is, several hundred bases), although this is not a fundamental limitation. The real advantage of MS sequencing is speed, since reading the output of the MS instrument is virtually instantaneous, compared to eight hours or so needed for the gel lanes to evolve to readable length. Many other techniques can be used, in principle, for sequencing with MS, and we will not go into all of them here. Some of these require a mass resolution capable of distinguishing all of the four base pairs by mass; this is a difficult job, since A and T differ by only 9 Da. (Sanger sequencing needs only to resolve one whole base pair, or about 300 Da.)

In early investigations into MS DNA sequencing, the methods for preparing and ionizing DNA (or protein) fragments were fast-atom bombardment or plasma ionization. (There are recent review articles on DNA MS, including references to the work described below [K. K. Murray, *J. Mass Spect.* 31, 1203 (1996); P. A. Limbach, *Mass Spectrometry*

Reviews 15, 297 (1996)]; the discussion here is based on these articles and on remarks from several experts.) But spectroscopy was limited to oligonucleotides of ten or fewer bases.

One significant step forward is the use of MALDI (Matrix-Assisted Laser Desorption/Ionization) to prepare ionic fragments of DNA for MS. The general idea is to embed the DNA in a matrix, which can be as simple as water ice, and to irradiate the complex with a laser of carefully-chosen frequency. This can both vaporize the complex and ionize the DNA, possibly by first ionizing the matrix followed by charge transfer to the DNA. There is a great deal of art in applications of MALDI, which is considerably more difficult to use with DNA than with proteins and peptides. For example, problems arise with unwanted fragmentation of the (already-fragmented) DNA during the MALDI process. Moreover, this MALDI fragmentation process is different for different bases. It is now possible to generate DNA fragments up to 500 bases long with MALDI, with resolution at about the 10 base level (compared to the needed resolution of 1 base). Typically MALDI DNA fragments have one unit of charge for every several hundred base pairs.

Another promising method for ionization is electrospray ionization (ESI). Here the charge produced is much higher (but can be varied by changing the chemistry of the solution containing the DNA). For example, experiments using T4 phage DNA fragments up to 10^8 Da have shown charges up to 3×10^4 . It is then necessary to determine both the mass per unit charge (as in conventional TOF MS) *and* the charge, in order to determine the mass. One potentially-important method introduces the accelerated ions into an open metal tube, where they induce an image charge that is measured; the charge-to-mass ratio is then measured by TOF.

MALDI-based methods are generally best for Sanger sequencing, but improvements are needed in the mass resolution and sensitivity (equivalently, DNA ion yield). ESI techniques lead to both higher mass resolution and higher mass accuracy, but because a great many charge states are created, it is not well-suited to analysis of a mixture of a large number of fragments (as is required in Sanger sequencing).

Looking toward the future, there are two ideas in MS that might someday reach fruition.

Arrays and multiplex MS sequencing Several briefers discussed ideas for using large arrays of DNA fragments with MS. One scheme [Charles Cantor, briefing to JASON, July 3, 1997] involves using arrays with various laydowns of DNA fragments, for subsequent MALDI-MS, with the fragments on the MALDI array designed to have properties desirable for MS. Another [George Church, briefing to JASON, July 2, 1997] points out that multiplexing with arrays is feasible for MS sequencing at rates of possibly 10^3 b/sec. One uses large (~65000) arrays with electrophore-tagged primers on the DNA fragments, with each primer having an electrophore of unique mass attached. DNA primed with these primers is grown

with dideoxy terminators, just as in Sanger sequencing. The four varieties are electrophoretically separated, then collected as droplets on an array. Finally, MALDI-TOF is used to remove the electrophores, ionize them, and identify them by MS. Each of the 400 different varieties of DNA is thus identified, yielding a multiplex factor which is the number of different electrophores (400 in this case). (Electrophore tagging of primers has been suggested as a means of increasing the ion yield from MALDI [P. F. Britt, G. B. Hurst, and M. V. Buchanan, abstract, Human Genome Program Contractor-Grantee Workshop, November, 1994].)

Single-molecule detection It is not obvious that MS-DNA sequencing requires single-molecule detection, but it in any case can be cited as the ultimate in MS sensitivity. It has already been shown [R. D. Smith *et al.*, *Nature* **369**, 137 (1994)] that a single ESI-DNA ion (up to 25 kb long) can be isolated for many hours in an FTICR mass spectrometer cell, making it available for measurements during this time. In another direction, detecting a single DNA molecule after acceleration should be possible, thus increasing the sensitivity of MS methods. Methods used for detection might involve bolometric arrays of detectors similar to those used for searches for cosmic dark matter. Such bolometric arrays are made on a pitch of $\sim 25 \mu\text{m}$ for use as sensitive IR focal-plane arrays. An ESI-ionized 30 kDa DNA fragment of charge 100 in a 30 keV potential drop will deposit some 3 MeV in a pixel, the same as 3×10^6 optical photons. The $25 \mu\text{m}$ spatial resolution can be used for resolving the mass and charge of the ion. It is intriguing to note that a single charged DNA fragment is something like the hypothesized magnetic monopoles of particles physics; both have masses of tens of kDa and large charges (of course, magnetic charge for the monopole). Considerable effort has gone into methods for detection of single monopoles, which are known to be very rare.

2.3.3 Hybridization arrays

A new technology that has progressed considerably beyond the stage of laboratory research is the construction of large, high density arrays of oligonucleotides arranged in a two-dimensional lattice. ["DNA Sequencing: Massively Parallel Genomics," S. P. A. Fodor, *Science* **277**, 393 (1997)] In one scheme (termed *Format 1*), DNA fragments (e.g., short clones from cDNA libraries) are immobilized at distinct sites on nylon membranes to form arrays of 10^4 - 10^5 sites with spot-to-spot spacing of roughly 1 mm. ["DNA Sequence Recognition by Hybridization to Short Oligomers: Experimental Verification of the Method on the *E. coli* Genome," A. Milosavljevic *et al.*, *Genomics* **37**, 77 (1996)] In a second scheme (termed *Format 2*), techniques of modern photolithography from the semiconductor industry have been adapted to generate arrays with 400,000 total sites [Fodor, *op cit.*] and densities as high as 10^6 sites/cm² ["DNA Sequencing on a Chip," G. Wallraff *et al.*, *Chemtech*, (February, 1997) 22], although the commercial state of the art appears to be perhaps 10 times smaller. For *Format 2* arrays, distinct oligomers (usually termed the *probes*) are lithographically generated *in situ* at

each site in the array, with the set of such oligomers designed as part of an overall objective for the array.

In generic terms, operation of the arrays proceeds by interacting the probes with unknown *target* oligonucleotides, with hybridization binding complementary segments of target and probe. For Format 2 arrays, information about binding of target and probe via hybridization at specific sites across an array is obtained via laser excited fluorescence from intercalating dyes which had previously been incorporated into either probe or target, while for Format 1 arrays, readout can be by either phosphor imaging of radioactivity or by fluorescence. Interrogation of the array via changes in conductivity is a promising possibility with potential for both high specificity and integration of the readout hardware onto the array itself.[T. Meade, private communication]

Typical probe oligomers are of length 7-20 base pairs, with single base-pair mismatches between target and probe having been detected with good fidelity. ["Mapping Genomic Library Clones Using Oligonucleotide Arrays," R. J. Sapolsky and R. J. Lipshutz, *Genomics* 33, 445 (1996); "Accessing Genetic Information with High-Density DNA Arrays," M. Chee *et al.*, *Science* 274, 610 (1996)]. For lithographically generated arrays, an important point is that all possible oligomers of length L (of which there are 4^L) can be generated in of order $4L$ processing steps, so that large search spaces (the number of probes) can be created efficiently.

Such large-scale hybridization arrays (with commercial names such *SuperChips* [Hyseq Inc., 670 Almanor Ave., Sunnyvale, CA 94086.] or *GeneChips* [Affymetric, <http://www.affymetric.com/research.html>]) bring a powerful capability for parallel processing to genomic assaying. The list of their demonstrated applications is already impressive and rapidly growing, and includes gene expression studies and DNA sequence determination. While hybridization arrays are in principle capable of *de novo* sequencing ["DNA Sequence Determination by Hybridization: A Strategy for Efficient Large-Scale Sequencing," R. Drmanac *et al.*, *Science* 260, 1649(1993)], the combinatorics make this a formidable challenge for long segments of DNA, since an unknown string of length N base pairs is one of $p=4^N$ possibilities (e.g., for $N=10^3$, $p\sim 10^{600}$).

Some sense of the probe resource requirements for *de novo* sequencing can be understood by the following "reverse" strategy applied to an array of Format 2 type. Consider an array containing oligomers of total length J with nondegenerate cores of length L that is exposed to an unknown fragment of length N . *A posteriori* one must be left with a sufficient number of probes that have matched the target so that a tiling pattern of probes can be assembled to span the entire target. As a lower bound on the number of required probes, imagine butting a set of N/L probes representing the nondegenerate cores end to end to cover the target, with $p=N/4^L \ll 1$ so that the conditional probability for two probes to match identical but disjoint regions of the target is small. For $(L, N) = (7, 10^3)$, $p\sim 0.06$, while for $(L, N) = (10, 10^4)$, $p\sim 0.01$. Since each probe has as its nondegenerate segment an arbitrary combination of base pairs, 4^L distinct oligomers are required in the original array, which for $L=7$ is 2×10^4 elements (well within the realm of

current capabilities), while $L=10$ requires about 10^6 elements (an array with 400,000 sites is the largest of which we are aware).

Unfortunately, this simple strategy does not allow one to deduce the ordering of the matching oligomer segments, of which there are approximately $(N/L)!$ permutations. Hence, imagine augmenting the above strategy so that the matching probes are arranged one after the other with the nondegenerate regions overlapping but offset by k base pairs. That is, adjacent probes are identical to each other and to the target in their overlapping regions, but differ by k base pairs in the nondegenerate regions at each end to provide sufficient redundancy to determine the ordering of the segments with high confidence. The number of probe segments needed to tile the target is then $1+(N-L)/k$. With the assumption of only pair-wise probe overlaps (i.e., $k>L/2$), the requirement for uniqueness in sorting then becomes $r=4^{(L-k)}/[1+(N-L)/k] \gg 1$, which cannot be satisfied for $(L, N)=(7, 10^3)$, while for $(L, N)=(10, 10^3)$, r is at most 5. On the other hand, for sequencing applications with $N=10^4$, L must be increased ($L=14$ gives $r\sim 10$ for $k=7$), with a concomitant explosion beyond current capabilities in the number of array elements required ($4^{14}=3\times 10^8$).

Note that these simple limits assume that target-probe hybridization and identification at each site are perfect and that N is a “typical” random sequence without perverse patterns such as multiple repeats. Certainly in practice a number of processes are encountered which complicate the interpretation of the hybridization patterns presented by arrays (e.g., related to complexity of the thermodynamics of hybridization, of patterns from multiple mismatches, etc.) and which are currently being addressed in the research literature, with promising demonstrations of fidelity. Clearly in any real application somewhat larger arrays than those based upon simple combinatorics will be needed for *de novo* sequencing to maintain accuracy and robustness in the face of errors, with an optimum array size lying somewhere between the limits discussed above.

While there are undoubtedly many “niche” applications for high density hybridization arrays to *de novo* sequencing (e.g., increasing the read length from 500-700 bases to beyond 1 kb would be important in the assembly process), such arrays seem to be better suited to comparative studies that explore differences between probe and target. Indeed, for Format 1 arrays, previously non-sequenced biological materials can be employed. It is clear that hybridization arrays will profoundly impact comparative genetic assays such as in studies of sequence polymorphism [M. Chee *et al.*, *op cit.*] and of gene identification and expression, as well as for understanding the relationship between genotype and phenotype. Beyond the research environment, one can imagine biochemical micro-laboratories for clinical applications [G. Wallraff *et al.*, *op cit.*] with hybridization arrays as essential elements for (differential) sequence analysis.

3. Quality

A project with the stated goal of sequencing the entire human genome must make data accuracy and data quality integral to its execution. It is clear that much of the genome will later be re-sequenced piece -by-piece. But a high-quality database can reduce the need for such resequencing, provide useful and dense markers across the genome, and enable large-scale statistical studies. A quantitative understanding of data quality across the whole genome sequence is thus almost as important as the sequence itself.

Technology for large-scale DNA sequencing is relatively new. While current sequencing tools and protocols are adequate at the lab-bench level, they are not yet entirely robust. For generic DNA sequence, the mainstream techniques are straightforward and can be carried out with low error rates. However problems and errors occur more frequently when sequencing particular portions of the genome or particular sequence patterns, and resolving them requires expert intervention. Phenomena such as deletions, unremoved vectors, duplicate reads, and chimeras are often the consequence of biological processes, and as such are difficult or impossible to eliminate entirely. Base-call accuracy tends to degrade toward the end of long sequence reads. Assembly of complete genomic sequences remains a challenge, and gaps are sometimes difficult to fill. In this situation, quality assurance and quality control (QA/QC) are essential. In particular it is crucial to understand quantitatively the accuracy of information going into the genome data base. The present section of this report discusses the coupled issues of quality assurance, quality control, and information about data quality, as they impact the Project, as well as other national and international sequencing efforts.

The following three steps provide a useful framework for analyzing and address in QA/QC issues for the Project (indeed, for any large-scale sequencing effort):

1. Quantify the quality requirements of present and future uses of genomic information
2. Develop assays that can accurately and efficiently measure sequence quality
3. Take steps to ensure that present and evolving sequencing methods and data meet the prescribe level of quality.

The following subsections consider each of these issues in turn. We then follow with some summary recommendations on QA and QC. Following the conclusion of our study, we became aware of a report of an NHGRI Workshop on DNA Sequence Validation held in April, 1996 [http://www.nhgri.nih.gov/HGP/Reports/dna_sequence_workshop.html] that independently examined some of the same issues and, in some cases, came to similar conclusions.

3.1 Quality requirements

Our briefers reflected a wide range of opinions on the magnitude of the required error rates for sequence data. This has clearly been a controversial issue and, at times, it has been used as a surrogate for other inter-Center disputes. We believe that the debate

on error rates should focus on what level of accuracy is needed for each specific scientific objective or end-use to which the genome data will be put. The necessity of “finishing” the sequence without gaps should be subject to the same considerations. In the present section, we stress the need for developing quantitative accuracy requirements.

3.1.1 The diversity of quality requirements

Genomic data will be (indeed, are being) put to a variety of uses and it is evident that the quality of sequence required varies widely among the possible applications. If we quantify accuracy requirements by the single-base error, ϵ , then we can give some representative estimates:

<u>Application</u>	<u>Error requirement</u>
Assemble long contigs	$\epsilon \sim 10^{-1}$
Identify a 20-mer sequence	$\epsilon \sim 10^{-1}$
Gene finding	$\epsilon \sim 10^{-2}$
Construct 20-mer STS primer	$\epsilon = 5 \times 10^{-4}$ (99% confidence) $\epsilon = 5 \times 10^{-3}$ (90% confidence)
Polymorphism	$\epsilon \sim 10^{-4}$ (coding regions) $\epsilon \sim ?$ (non-coding regions)
Studies of genomic evolution, statistics	???
Genetic defects	$\epsilon = "0"$

Although these are only rough order-of-magnitude estimates; we justify each as follows.

- The surprisingly low accuracy we estimate to be required to assemble long contigs and to identify the presence of a precisely known 20-mer in a sequence is discussed in the following subsection
- Our estimate for the gene finding requirement is based on the observation that pharmaceutical companies engaged in this activity seem satisfied with short sequences (400 bases) at this level of accuracy.
- The required accuracy to construct a 20-mer STS primer is based on straightforward probabilistic calculations.
- The polymorphism entry simply repeats the common statement that accuracy 10 times better than the observed polymorphism rate is sufficient.
- The requirements for evolutionary or statistical studies of the genome have not been quantified
- Our value for genetic defects stems from the single-base errors causing some genetic diseases (e.g., sickle cell anemia)..

More precise estimates for each of these uses (and others) can surely be generated by researchers expert in each of the various applications. Beyond qualitative judgment, one useful technique would be to run each of the applications with pseudodata in which a test sequence is corrupted by artificially generated errors. Variation of the efficacy of each application with the error level would determine its error requirement and

robustness. Such exercises, carried out in software, cost little, yet would go a long way toward setting justifiable quality goals. We recommend that the DOE encourage the genomics community to organize such exercises.

With this kind of data in hand, one could establish global quality requirements for the final sequence (perhaps different for coding and non-coding regions). It is likely that arbitrarily high accuracy could be achieved by expending enough effort: multiple sequencing with alternative technologies could guarantee high accuracy, albeit at unacceptable cost. In the real world, accuracy requirements must be balanced between what the users need, the cost, and the capability of the sequencing technology to deliver a given level of accuracy. Establishing this balance requires an open dialog among the sequence producers, sequence users, and the funding agencies, informed by quantitative analyses.

3.1.2 Accuracy required for assembly

A probabilistic analysis of the assembly problem shows that (in an ideal case) assembly requires relatively little accuracy from the raw sequence data. These data are the sequences of base calls derived from the individual reads. An accuracy as low as 0.9 (per base call) is sufficient to ensure reliable assembly. A high degree of coverage is required, however, to have any chance of assembling the entire clone without gaps.

We first consider the problem of assembling k fragments of length L with left endpoints uniformly distributed over a clone of length M . Requiring overlaps above a given threshold does not really complicate the gap problem. The point is that a tiling of the sequence of length M with fragments of length L overlapping with subsegments of length at least x is ensured by a tiling with no gaps with fragments of length $L-x$.

We can compute an approximate lower bound for the probability of success as follows. The probability that for a given region of length L^* , some fragment has its left endpoint somewhere in the given region is

$$1 - (1 - L^*/M)^k$$

where k is the number of fragments considered.

We now suppose that the clone length is 30,000 and that the fragments have length 1300. The probability that with 450 fragments there exists a sequence of 150 distinct fragments starting at the left end of the clone such that each successive fragment starts in the left-justified 1200-length subfragment of the previous fragment (thereby ensuring overlaps of 100) is at least

$$\left[1 - \left(1 - \frac{1200}{30000} \right)^{300} \right]^{150} > 0.99928,$$

which is conservative since the inner exponent is really varying from 449 to 300.

Randomly selecting such a *walk* across the clone, the probability that the walk reaches the other end of the clone is greater than

$$2 \binom{150}{50} \left(\frac{1}{2}\right)^{150} > 3 \times 10^{-5}.$$

This conservatively estimates the probability that at least 50 of the successive overlaps begin in the right-justified half of the 1200 length region of the previous fragment (and so extend the walk by at least 600 bases). Thus the probability that the selected walk covers the clone is greater than 0.999.

Sequencing the fragments from both ends yields the sequence, assuming read lengths of 650. The advantage of longer reads is that longer fragments can be used and hence for a desired probability for coverage, fewer fragments can be used. A distinct possibility is that merely improving the *percentage* of long reads has a significant effect.

We emphasize that these are simply lower bounds which are rather conservative, computed for this idealized case.

We next consider the probability that a complete tiling can be constructed and correctly assembled given a specific error rate in the base calls. Suppose that G is a sequence of bases of length x , G^* is a probabilistic garbling of G with an error rate $1-E$ and that R is a random sequence of length x . For each $m < x$, the probability that G and G^* disagree in at most m places is

$$p_m = \sum_{k=0}^m \binom{x}{k} E^k (1-E)^{x-k}.$$

The probability that G^* and G disagree in at most m places is

$$q_m = \sum_{k=0}^m \binom{x}{k} (0.75)^k (0.25)^{x-k},$$

which is dominated by the last term for the relevant values of x and m .

We examine the case when $x=100$ and $E=0.1$. In the assembly problem, p_m should be calculated with a smaller error rate since one is considering matches between two garbled sequences. For an error rate of $E=0.1$, the effective error rate is approximately 0.186. Typical values for varying choices of m are

$$p_{39}=0.99999996; p_{40}=0.999999987; p_{41}=0.999999995.$$

The corresponding values for q_m are

$$q_{39}=2.87 \times 10^{-14}; q_{40}=1.33 \times 10^{-13}; q_{41}=5.90 \times 10^{-13}.$$

At each stage of the construction of the walk and with a threshold of m , the probability that there is an assembly error which passes the threshold requirement is at most

$$1 - (1 - q_m)^{1200 \times 450}.$$

The probability that a correct fragment will pass, correctly placed, is at least p_m (in the worst case of there only being one such fragment). Thus, if there is a walk across the clone, the probability of constructing a valid walk across the clone is at least

$$P_m = (1 - q_m)^{1200 \times 450 \times 150} \times P_m^{150}.$$

With values as above, we have

$$P_{39}=0.99993; P_{40}=0.99997; P_{41}=0.99996.$$

With a threshold of 40 the probability of constructing a correct walk across the clone is essentially the same (0.999) as the probability that there exists such a walk across the clone.

The analysis here makes several (important) simplifying assumptions. For example, it assumes that the fragments are uniformly distributed across the clone and that the clone itself is a random sequence of base pairs. While in some regions of the genome the latter may be a good assumption, there are certainly areas where it is not. Even somewhat limited partial repeats within the clone will have a possibly significant impact on the analysis. This can be explored experimentally via computer simulations using known stretches of the sequence (Section 3.3.1).

Further, with fragments produced using sets of restriction enzymes, the fragments may well not be uniformly distributed and we only considered pointwise garbling (not insertions or deletions). However the intent of this analysis is simply to illustrate the relative importance of base-calling accuracy and coverage (number of fragments) in the sequencing process.

Another important point is that attention should be paid to examining the relative merits of:

- Having the sequence of the genome at relatively low accuracy, together with a library of fragments mapped to the sequence;
- Having the sequence of the genome at high accuracy.

There are sequencing strategies in which the order of the fragments is essentially known in advance; one such is discussed in Section 2.1.3. The assembly of such a library of fragments is easier (significantly easier for the idealized random genome). It is possible that for sequencing certain regions of the genome these approaches coupled to accepting higher error rates in the reads, are superior.

A final point concerning accuracy is the placement of known sequences against the garbled genome sequence. Suppose that, as above, the garble rate is 0.1; *i.e.*, the accuracy is 0.9. Then given a sequence of length 50 from the true sequence, the probability that the sequence is correctly, and uniquely, placed is 0.999 using a threshold of 12 errors. Again, the assumptions are that the genome sequence is random or at least that the given segment is from a portion of the genome which is random. However if a significant fraction of the genome *is* random then (with high probability) false placements will only happen in the remaining fraction of the genome. This could be used to produce interesting kinds of maps, using a small library of target fragments. Again some simulations can easily test these various points known sequence data and allowing errors of insertion and deletion.

3.2 Verification protocols

Since the “proof of the pudding” lies in the actual accuracy of the output, absolute accuracy can be determined only by physical testing of the sequence output. That is, given the putative sequence of base pairs for a certain contig (which we term the “software sequence”), independent protocols should be established to verify this software sequence relative to the physical contig. Such “verification” is a different task from *de novo* sequencing itself, and should be accomplished by means as independent as possible from those employed to obtain the initial sequence.

An ideal verification method would be:

- **Sequence blind:** requires no *a priori* knowledge of the sequence
- **Sequence independent:** efficacy independent of the sequence being verified
- **Reliable:** a high probability of detecting errors, with low probability of false alarms
- **Economical:** cost (labor, materials, time) a small fraction of the cost of sequencing
- **Capable:** long sequences easily verified
- **Specific:** provides further information about the errors beyond “Right or Wrong”

One obvious strategy is to resequence the DNA by a method different than that used by the original researcher. Unfortunately, this fails on the grounds of economy and the fact that today there is really only one large-scale sequencing technique.

In this section, we describe two possible verification protocols, and close with a discussion of the implementation of *any* protocol.

3.2.1 Restriction enzyme verification of sequence accuracy

We propose Multiple Complete Digestions (MCD) as a verification protocol satisfying most of the criteria above. It will allow statements like “With 90% probability, this sequence is accurate at the 10^{-3} level” or, more generally, “With confidence C , the sequence is accurate at the ϵ level.” It may also be used to localize and characterize errors in the sequence.

MCD has been developed and used as a method for generating high-quality physical maps preparatory to sequencing [G. K.-S. Wong *et al.*, PNAS 94, 5225-5230, 1997]. Here, we quantify the ability of this technique to provide probabilistic sequence verification.

The basic idea is that the putative sequence unambiguously predicts the fragment lengths resulting from digestion by any particular endonuclease, so that verification of the

fragment lengths is a necessary (but not sufficient) check on the sequence. Multiple independent digestions then provide progressively more stringent tests. Of course, if the putative sequence has been generated by MCD with one set of enzymes, a completely different set must be used for verification.

Let us assume that ϵ is the single-base error rate, that only single-base substitutions or deletions can occur, and that we are using restriction enzymes specific to a b -base pattern (most commonly, $b = 6$ for the enzymes used in sequencing, although enzymes with $b = 4, 5, 7,$ and 8 are also known).

A digestion will give an error (i.e., fragments of unexpected length) when an error has destroyed a restriction site or created a new one from a “near-site” of b -bases whose sequence differs from the target sequence by one base (we ignore the probability of two or more errors occurring simultaneously within a restriction site or near-site). Then the probability of any one restriction site being destroyed is $b\epsilon$ (since the error can occur in any one of the b positions), while the probability of a near-site being converted is $\epsilon/3$ (since only one of the three error possibilities for the “wrong base” leads to a true site).

Then the expected number of errors in a sequence containing S sites and N near sites is

$$\langle E \rangle = \epsilon b S + \epsilon N / 3 \equiv \epsilon \sigma$$

where $\sigma = bS + N / 3$ is the effective number of sites.

3.2.1.1 Probabilistic estimate

Let us now consider a sequence of length L bases. Assuming that bases occur at random, we expect $S=L/4^b$ sites for a single restriction enzyme and $N=3bL/4^b$ near sites, since there are 3 ways each of the b bases at a site can differ from the target pattern. Hence, for D different digestions, we expect

$$\sigma = 2DbL / 4^b$$

Since the number of fragments expected if there are no errors is $S=L/4^b$ and a convenient number of fragments to separate is $S=10$, taking $b=6$ implies a sequence length of $L=40$ kb (the size of cosmid clones) and $\sigma= 120D = 600$ if $D = 5$.

3.2.1.2 Real DNA

The probabilistic estimate of σ above assumed that all b -mers were equally likely, or more precisely, that the recognized b -mers were uniformly distributed. However, there is no need to make that assumption when DNA is presented for checking. Instead one

can scan the proposed sequence and count the number of sites where errors could make a difference in how the sequence is cleaved. The calculation mimics exactly the random model above: each recognized site contributes 1 to σ and each near site contributes 1/3. The total for the sequence is then the contribution of that endonuclease to σ .

The table below shows the results of this counting for $D=5$ restriction enzymes for three pieces of human sequence from the Whitehead Center: L10 of length 48 kb, L8 of length 47 kb, and L43 of length 44 kb. (The first two are on 9q34, while the third is on the Y chromosome). Also considered is a completely random sequence of 40 kb.

Site \ Fragment	L10 (48 kb)	L8(47 kb)	L43(44 kb)	Random (40 kb)
GGATCC (<i>Bam</i> I)	126	117	112	137
GATATC (<i>Eco</i> RV)	49	40	105	94
AAGCTT (<i>Hind</i> III)	66	112	134	121
TCTAGA (<i>Bgl</i> II)	84	79	190	145
TGGCCA (<i>Msc</i> I)	295	377	109	122
σ	620	725	650	619

These results agree with the probabilistic estimate of $\sigma \sim 600$ for $D=5$ and $L \sim 40$ kb. However, while the probabilistic model is true on average, it is not true in detail and some restriction enzymes give more meaningful tests of a given sequence than others (i.e., contribute more to σ). For example, digestion of L10 with *Eco*RV does not add very much information, while digestion with *Msc*I does. Hence, for a given DNA sequence, it is possible to choose the most meaningful set of restriction enzymes to be used in the test.

3.2.1.3 Judging the results

When a particular sequence is digested with a particular set of enzymes, the number of errors actually observed will be given by a Poisson distribution, in which the probability of observing E errors is

$$P(E) = \frac{\langle E \rangle^E}{E!} e^{-\langle E \rangle}$$

What can be learned from a MCD test that shows E errors? Let us assume that the tests are arranged so that $\sigma=700$, that $\epsilon=10^{-3}$ the quality goal, and that we declare that any sequence showing $E < 2$ errors in an MCD test is "good." In that case, there is a false

alarm probability of $P_{FA}=0.16$ that an $\epsilon=0.001$ sequence will be rejected, and will have to be redone. However, if the sequence has $\epsilon=0.01$, there is only a $P_A=0.007$ probability that it will be accepted. Hence, this simple operational definition (at most one error) implies only slightly more work in resequencing, but gives high confidence (>99%) in a sequence accuracy at the level of $\epsilon=0.01$ and 90% confidence in the sequence at the $\epsilon\sim 0.005$ level. The implications of other choices for the maximum acceptable number of errors or for different values of $\langle E \rangle$ follow straightforwardly from the properties of the Poisson distribution; some representative values for $\sigma=700$ are given in the table below.

	$E<1$	$E<2$	$E<3$	$E<4$
$P_{FA}(\epsilon=0.001)$	0.50	0.16	0.035	0.006
$P_A(\epsilon=0.01)$	0.0009	0.007	0.03	0.08
$\epsilon(P_A=0.1)$	0.003	0.005	0.008	0.010

Note that the estimates above assume both perfect enzyme specificity; and sufficient fragment length resolution (1% seems to be achievable in practice, but one can imagine site or near-site configurations where this would not be good enough, so that a different set of restriction enzymes might have to be used). The extent to which these assumptions hinder MCD verification can best be investigated by trials in the laboratory.

3.2.2 Hybridization arrays for sequence verification

As we have discussed in Section 2.3.3, the combinatorics make *de novo* sequencing a formidable challenge for present-day hybridization arrays. However, beyond the differential sequencing applications we have discussed, one potentially important application of hybridization arrays is to the problem of sequence quality control and verification, particularly since it is extremely important to employ means independent of those used to derive the putative sequence of a particular contig.

Hybridization arrays could provide a method for sequence verification independent of the present Sanger sequencing. The strategy would be to construct a Format 2 array based upon the candidate sequence for the contig. This array would then be challenged by the physical contig, with the goal being to detect differences between the “software” sequence as determined by a previous sequencing effort and the “hardware” sequence of the contig itself. For this protocol the “software” sequence would be represented by the oligomer probes of the array. Since the objective is to detect differences between two very similar sequences, the requirements on the number of distinct probes and hence on the size of the array are greatly relaxed as compared to the previous discussion of *de novo* sequencing. More explicitly, to scan a target contig of length N bases for single-base mismatches relative to a “known” (candidate) sequence, an array of $4N$ probes is required, which would increase to $5N$ if single site deletions were included. The array might include as well sets of probes designed to interrogate specific

“problem” sections of the target. For $N \sim 40$ kb, the required number of probes is then of order 2×10^5 , which is within the domain of current commercial capability.

Note that relative to the proposal in Section 3.3.2 to establish “gold standards” of DNA sequence, this strategy could also play an important role in helping to verify independently the standards themselves.

A case study relevant to the objective of sequence verification and error detection by hybridization is the work of M. Chee *et al.* [op cit.], for which an array with 135,000 probes was designed based upon the complete (known) 16.6 kb sequence of human mitochondrial DNA. As illustrated in Figure 2, this work detected sequence polymorphisms with single-base resolution, with 15-mer probes. Note that the total number of probes (135,000) is considerably smaller than the total possible set for a 15-mer ($4^{15} \sim 10^9$), allowing considerable flexibility in the design of the probes. In terms of an overall figure of merit for accuracy, the simplest possible procedure was employed whereby a scan to detect the highest fluorescent intensity from among the four possible base substitutions was made and led to 99% of the target sequence being read correctly. While this accuracy is not overwhelmingly, considerable improvement could presumably be achieved by incorporating more sophisticated analysis algorithms which take into account the overall pattern of mismatches, such as the were in fact employed by Chee *et al.* in their studies of polymorphisms for mitochondrial DNA from various populations. Of course since mDNA is eubacterial in character, many of the more challenging sequence pathologies are absent relative to eukaryotic DNA. Still, Chee *et al.* provides a useful benchmark against which to assess the potential of hybridization arrays for sequence verification.

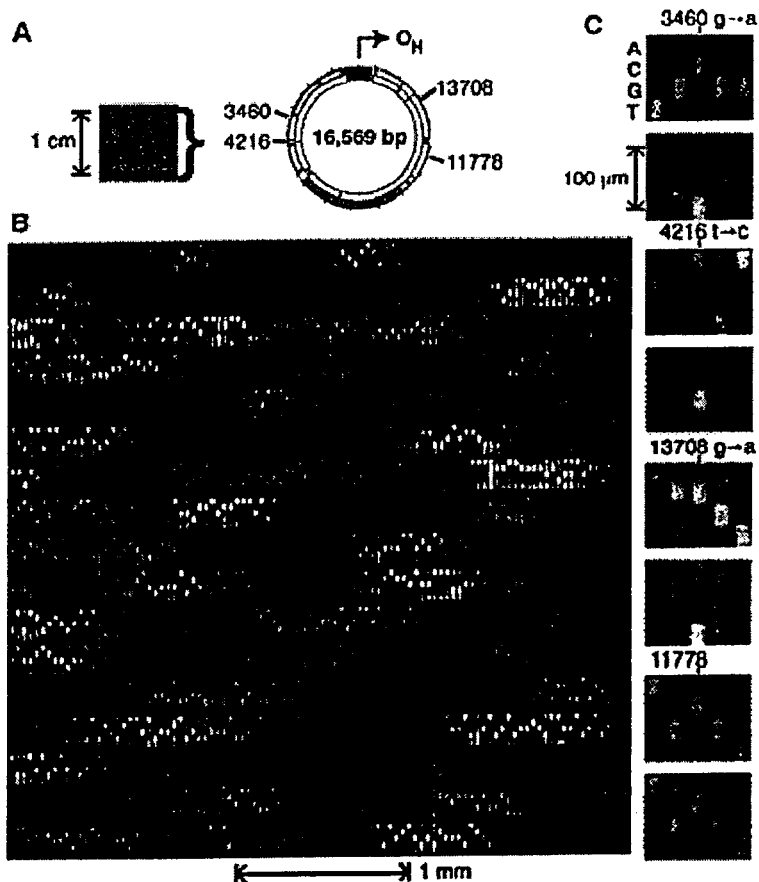


Figure 2: Human mitochondrial genome on a chip. (A) An image of the hybridized array with 135,000 probes designed to interrogate the 16.6 kb mitochondrial target RNA. (B) A magnified portion of the array. (C) Illustration of the ability to detect single base-pair differences. [from M. Chee *et al.*, *op cit.*]

Perhaps the most important motivation for suggesting this strategy for verification is that the “mistakes” associated with sequence determination from target-probe interactions in a massively parallel fashion may well be sufficiently different from those arising from the gel-based procedures so as to give an independent standard for accuracy. Of course there are a host of issues to be explored related to the particular kinds of errors made by hybridization arrays (including the fidelity with which the original array is produced, hybridization equivalents, etc.). For the purpose at hand, attention should be focused on those components that most directly impact the accuracy of the comparison.

Particular suggestions in this regard relate to the readout and image processing for the array, tasks which are often accomplished site by site via scanning confocal microscopy. It would seem that alternate readout strategies should be explored, including (perhaps image-intensified) CCDs. Since the target sequence is known with small uncertainty as are the set of errors associated with single-base substitutions and deletions as well as with other “typical” errors in sequence reconstruction, image templates could be pre-computed and cross-correlated with the actual image by adapting algorithms from

the image processing community to improve the accuracy with which information is extracted from the array.

The value of such a strategy for sequence verification extends beyond that of providing an independent avenue for error checking. It would also couple the traditional gel-based effort to emerging advanced technologies, with potential benefit to both. Moreover, it could be part of a broader attempt to define a longer-term future for the sequencing centers as new technologies come on line to supplant gel-based sequencing and as emphasis shifts from *de novo* sequencing to comparative studies such as related to polymorphisms.

3.2.3 Implementation of verification protocols

Any verification protocol must require significantly less effort than resequencing, and so there will be considerable latitude in its implementation. In one limit, sequencing groups might be required to perform and document verification protocols for all finished sequence that they wish to deposit in a database. Alternatively, a "verification group" could be established to perform "spot" verifications of database entries selected at random. A third possibility is to offer a "bounty" for identifying errors in a database entry.

Clearly, economic, sociological, and organizational factors must be considered in choosing among these, and other, possible implementations. We recommend that the funding agencies promote a dialog within the sequencing communities about possible verification protocols and their implementation.

3.3 Assessing and improving present techniques

Our emphasis on quantitative metrics for accuracy is made against the backdrop of inadequate understanding of the quality of the "end product" in the current Human Genome sequencing effort. While the level of competence and effort devoted to "doing the job right" in the sequencing centers is commendable, there is a clear need to implement a comprehensive program of quality assurance and quality control. Here we suggest some ways to provide more quantitative measures of the errors in the end product, and to understand how the various steps in sequencing contribute to the overall error budget.

Quality assurance and quality control (QA/QC) are of sufficient importance to be made integral requirements in the sequencing project. Each sequencing center should invest a fraction of its own budget to characterize and understand its particular accuracy and error rates. This should be part of a continuing effort, rather than a one-time event. Quality control within the Centers should not be externally micro-managed, but each Center should be required to develop its own credible plan for QA/QC.

We further urge that the effort to develop new QA/QC technology be tightly coupled to the sequencing centers. In particular, new technologies such as large scale hybridization arrays or single-molecule sequencing are not currently competitive with gel-based electrophoresis for high throughput sequencing and long base reads, but they could be exploited in the short term for "niche" applications such as sequence verification for QA/QC. In the longer term, the Centers must integrate new technical advances into their operations, and the avenue of QA/QC is an important mechanism to help to accomplish this goal. From a longer-term perspective it seems clear that after the human genome has been sequenced once, emphasis will shift toward differential sequencing relevant to the study of polymorphism and homologies, and to the genetic origins of disease. QA/QC can thus be viewed as part of a broader effort to define a long-term future for the sequencing Centers, with technological leadership at the forefront as a prime component.

3.3.1 A systems approach is required

As we have noted, the issues of accuracy and error rates in reconstructed genomic information are crucial to the value of the "end-product" of the Human Genome Project, yet requirements for accuracy are complex and detail-dependent. DOE should support a portfolio of research in genome quality assurance and quality control issues.

One of the elements of this research should be computer simulation of the process of sequencing, assembly, and finishing. We believe that research into the origin and propagation of errors, *through the entire system*, are fully warranted. We see two useful outputs from such studies: 1) more reliable descriptions of expected error rates in final sequence data, as a companion to database entries, and 2) "error budgets" to be assigned to different segments of mapping and sequencing processes to aid in developing the most cost-effective strategies for sequencing and other needs.

DOE should solicit and support detailed Monte Carlo computer simulation of the complete mapping and sequencing processes. The basic computing methods are straightforward: an ideal segment of DNA is generated and subjected to models of all steps in the sequencing process; individual bases are randomly altered according to models of errors introduced at the various stages; the final, reconstructed segment or simulated database entry is compared with the input segment; and errors are noted.

Results from simulations are only as good as the models used for introducing and propagating errors. For this reason, the computer models must be developed in close association with technical experts in all phases of the process being studied so that they best reflect the real world. This exercise will stimulate new experiments aimed at the validation of the error-process models, and thus will lead to increased experimental understanding of process errors as well.

Useful products of these types of simulations are "error budgets" for different steps in the measurement and analysis chain. Such budgets reflect the contributions of individual steps and their effect on the accuracy of the final result. This information can

be used, for example, to establish quality criteria for the various stages of the sequencing process, so that emphasis and funds can be devoted to improving the accuracy of those steps which have the strongest influence on the accuracy of the final sequence product.

Error budgets will depend on the final accuracy required for a specific use of the end-product, which is analyzed sequence information. By comparing specific end-product needs for accuracy and quantity of information with error budgets and costs of individual steps in the overall process from DNA to database, it should be possible to perform cost/benefit analyses for developing optimum sequencing strategies.

3.3.2 "Gold standards" for measuring sequence accuracy

DOE should take the lead in developing "gold standards" for human DNA sequence. Standard DNA sequences could be used by the whole sequencing community for assessing the quality of the sequence output and sequencing protocol through "blind" experiments within the various centers. These gold standards should be designed to highlight quality assessment in "hard" DNA-sequencing regions and in potential problem areas, as well as in "ordinary" coding regions. They would consist of cloned DNA molecules of two types:

- a cosmid vector containing an insert of ~40 kb of human DNA that has been sequenced with high accuracy and assembled without any unresolved discrepancies;
- a phagemid vector containing an insert of ~1 kb of synthetic DNA including both human-derived sequences and contrived sequences that are known to cause common artifacts in DNA sequence acquisition.

The standard cosmid will have been transduced and propagated in bacterial cells, then stored as individual aliquots kept at -70 °C. Upon request, one or more of these aliquots would be made available to a sequencing group. All of the subsequent steps, including further propagation of the cosmid, restriction mapping, subcloning, sequencing, assembly, and finishing would be carried out by the sequencing group. Performance could be assessed based on a variety of indices such as PHRED and PHRAP scores, number of sequencing errors relative to the known standard, type of sequencing errors, time required to complete the sequencing, and cost of sequencing. The cosmid standard might also be used to compare alternative sequencing protocols within a sequencing center or to conduct pilot studies involving new instrumentation.

The standard phagemid will have been produced in large quantity, purified, and stored as individual aliquots kept at -70 °C. After thawing, the DNA will be ready for sequencing, employing "universal" primers that either accompany the template DNA or are provided by the sequencing group. The purpose of this standard is to assess the quality of DNA sequencing itself, based on indices such as PHRED score, read length, and the number and type of sequencing errors relative to the known standard. The target

sequence will have been designed to elicit common sequencing artifacts, such as weak bands, strong bands, band compressions, and polymerase pauses.

Although the cosmid standard is expected to have greater utility, the phagemid standard will be used to control for variables pertaining to DNA sequencing itself within the overall work-up of the cosmid DNA. It is likely that the sequencing groups will be on their “best behavior” when processing a gold standard, resulting in enhanced performance compared to what might be typical. This cannot be avoided without resorting to cumbersome procedures such as surprise examinations or blinded samples. Thus it will be important to examine not only the output of the sequencing procedures, but also the process by which the data is obtained. The extent to which it is possible to operate in a “best behavior” mode will itself be instructive in assessing DNA sequencing performance.

We recommend that the DOE provide funding, on a competitive basis, to one or two individual investigators who will construct and maintain the DNA standards. It might be appropriate to construct a small family of cosmid and phagemid standards that would be made available sequentially. The experience of the sequencing groups in processing these gold standards will suggest ways in which they could be improved to better assess critical aspects of the sequencing process.

3.3.3 Quality issues pertaining to sequencing templates

While most of our discussion has involved QA/QC issues in the sequencing and assembly process, it is useful to consider also quality issues in the processes used to prepare DNA for sequencing. We do so in this subsection.

There are many steps involved in construction of a human genomic DNA library and subcloning of that library into a form suitable for automated DNA sequencing. These include:

1. fragmentation of chromosomal DNA by mechanical shearing or partial enzymatic digestion;
2. size fractionation of the DNA fragments by gel electrophoresis or centrifugation;
3. cloning of ~1 Mb fragments into high-capacity vectors, such as YACs or BACs;
4. propagation of YACs or BACs within host cells;
5. enzymatic digestion of YAC or BAC inserts to obtain fragments of ~40 kb;
6. cloning into medium-capacity cosmid vectors;
7. propagation of cosmids within bacterial cells;
8. enzymatic digestion of cosmid inserts to obtain fragments of ~1 kb;
9. cloning into low-capacity plasmid or phagemid vectors;
10. preparation of purified plasmid or phagemid DNA.

Although each of these steps can introduce artifacts that make sequencing more difficult, they are not the most critical with respect to the overall quality of the sequencing process.

The subsequent steps of dideoxy sequencing, base calling, assembly, and finishing are all more prone to error.

The steps involved in the preparation of templates for sequencing are made error tolerant by the exponential amplification that is inherent in these procedures. Errors do occur, such as empty vectors, poor transformation efficiency, insufficient vector amplification, and inadequate purity of the template DNA. These problems usually result in clones that drop out of the process. Provided that there is redundant coverage of the DNA among the successful clones, the failed clones can essentially be ignored. However three quality control issues pertaining to template preparation merit special attention:

1. There may be incomplete representation of the genomic DNA at the level of the BAC/YAC, cosmid, or plasmid/phagemid libraries. This may be due to insufficient redundancy in construction of the library, but more often they are due to regions of the chromosome that are either difficult to clone or difficult to propagate within host cells. The genomics community is well aware of these problems and has taken appropriate countermeasures. Unlike the yeast genome, which has been sequenced successfully in its entirety, there may be regions within the human genome that cannot be cloned and therefore cannot be sequenced. At present the best course of action is to press ahead and deal with the problem of “unsequenceable” DNA if and when it arises.
2. There may be spurious DNA sequences intermixed with the desired genomic DNA. The two most common sources of contamination are vector-derived DNA and host cell DNA. Vector sequence can be recognized easily by a suitable sequence-matching algorithm. Incredibly, there are many entries in the genomic databases today that are either partly or completely derived from vector sequence. Host cell DNA is more difficult to recognize, but these too can be identified with the complete genomic sequences of yeast and *E. coli* available. Although spurious sequences can be eliminated after the fact, it should be made incumbent on the sequencing centers to do this prior to database submission.
3. There are challenges in maintaining proper inventory control over the vast number of clones and subclones that are being generated by the human genome project. Current procedures at the major genome centers are adequate in this regard. A physical inventory should be maintained for all BAC/YAC and cosmid clones, but this is not critical for the plasmid/phagemid clones. An electronic inventory, with secure back-up copies, should be maintained for all clones and subclones that are generated.

4. Genome informatics

4.1 Introduction

In a statement of research goals of the US Human Genome Project [F. Collins and D. Galas, "A new five-year plan for the US Human Genome Project," *Science* 262: 43-46 (1993)], the project's leaders define "informatics" as:

... the creation, development, and operation of databases and other computing tools to collect, organize, and interpret data.

Their goals for the current 5-year period are:

- Continue to create, develop, and operate databases and database tools for easy access to data, including effective tools and standards for data exchange and links among databases.
- Consolidate, distribute, and continue to develop effective software for large-scale genome projects.
- Continue to develop tools for comparing and interpreting genome information.

While similar in purpose and style to other major scientific cataloging efforts of the past and present—for example, Handbook of Chemistry and Physics, Chart of the Nuclides, to name two familiar resources—the Human Genome Project's informatics task is strikingly unified in that its focus is solely on translating and disseminating the information coded in human chromosomes. Genome informatics differs from earlier scientific catalogs also because it is a "child" of the information age, which brings clear advantages and new challenges, some of which are related to the following:

- the large amount of information to be assembled in meaningful ways, while the semantic content of that information is still largely not understood.
- the reliance on software algorithms at all stages from assembling pieces of information to interpreting results
- the large, globally distributed and diverse provider/user base
- the broad range of quality of information being processed and accessed, with uncertainties in even the measures of quality
- the rapidity with which the quantity and quality of information is increasing

Within the Human Genome Program, technical challenges in the informatics area span a broad range. Genome informatics can be divided into a few large categories: data acquisition and sequence assembly, database management, and genome analysis tools. Examples of software applications within the three categories include:

Data acquisition and sequence assembly:

- Process management and inventory control within Genome Centers
- Tools to track the pedigree of raw input data sources
- Servo control systems for individual robotic processes
- Software environments for coordinated distributed computing (e.g. robotic control systems) within a Genome Center

- Physical mapping software tools
- Base-calling software
- Sequence assembly tools
- Annotation tools; software for automatic sequence annotation
- Tools for automated submission of information to database centers

Database management:

- Local special-purpose databases
- Community-wide relational databases
- Software for database curation and quality control
- User “front ends” and interfaces for complex database queries
- “Middleware” for integration between separate databases
- Software to resolve semantic and nomenclature conflicts

Genome analysis:

- Data-mining tools
- Homology searches
- Identification of coding regions and genes
- Comparative genomics tools
- Placing proteins into gene families
- Tools for lineage analysis
- Tools for combinatorial analysis of hybridization array data

Managing such a diverse informatics effort is a considerable challenge for both DOE and NIH. The infrastructure supporting the above software tools ranges from small research groups (e.g. for local special-purpose databases) to large Genome Centers (e.g. for process management and robotic control systems) to community database centers (e.g. for GenBank and GDB). The resources which these different groups are able to put into software sophistication, ease of use, and quality control vary widely. In those informatics areas requiring new research (e.g. gene finding), “letting a thousand flowers bloom” is DOE’s most appropriate approach. At the other end of the spectrum, DOE and NIH must face up to imposing community-wide standards for software consistency and quality in those informatics areas where a large user community will be accessing major genome data bases.

The need for genome quality assurance enters the informatics field at several different levels. At the earliest level, both policies and tracking software are needed that will preserve information about the pedigree (origin and processing history) of data input to the sequencing process. This potentially includes information on the origins of clones and libraries, image data of gel runs, and raw data of ABI-machine traces. Policies need to be developed concerning minimum standards for archiving the raw data itself, as well as for the index that will allow future users to find raw data corresponding to the heritage of a specific DNA sequence.

At the level of sequencing and assembly, DOE and NIH should decide upon standards for the inclusion of quality metrics along with every database entry submitted (for example PHRED and PHRAP quality metrics, or improvements thereon).

At the level of database quality control, software development is needed to enhance the ability of database centers to perform quality checks of submitted sequence data prior to its inclusion in the database. In addition, thought needs to be given towards instituting an ongoing software quality assurance program for the large community databases, with advice from appropriate commercial and academic experts on software engineering and quality control. It is appropriate for DOE to insist on a consistent level of documentation, both in the published literature and in user manuals, of the methods and structures used in the database centers which it supports.

At the level of genome analysis software, quality assurance issues are not yet well posed. Many of the current algorithms are highly experimental and will be improved significantly over the next five years. Tools for genome analysis will evolve rapidly. Premature imposition of software standards could have a stifling effect on the development and implementation of new ideas. For genome analysis software, a more measured approach would be to identify a few of the most promising emerging analysis tools, and to provide funding incentives to make the best of these tools into robust, well-documented, user-friendly packages that could then be widely distributed to the user community.

4.2 Databases

Currently, there are many, diverse resources for genomic information, essentially all of which are accessible from the World Wide Web. Generally, these include cross references to other principal databases, help-files, software resources, and educational materials. The overall impression one gets after a few hours of browsing through these web sites is that of witnessing an extraordinarily exciting and dynamic scientific quest being carried out in what is literally becoming a world-wide laboratory.

Web tools and the databases are also changing how the biology community conducts its business. For example, most journals now require a “receipt” from one of the standard databases indicating that reported sequence data have been filed before a paper is published. The databases are finding ways to hold new entries private pending review and publication. The databases contain explicit reference to contributors—there is probably no better way to exercise real quality control than the threat of exposure of incorrect results. We view all these developments as being very positive.

With so much information coming available, considerable effort goes into staying current. Many institutions conduct daily updates of information from the database centers. This works because such updates can be performed automatically off of peak working hours. The resources needed to update and circulate information are likely to

increase as volume increases. The effort in learning how to use relevant database tools represents an important investment for individual scientists and group leaders.

Maintenance of databases is an important resource question for the Project. Currently, DOE supports two major efforts:

1. **Genome Sequence DataBase (GSDB)** (www.ncgr.org) operated by the National Center for Genome Resources which was established in Santa Fe in July, 1994. GSDB is described in its Web information as "one of the key components of the emerging federated information infrastructure for biology and biotechnology."
2. The **Genome Database (GDB)** (gdbwww.gdb.org) was established at Johns Hopkins University in Baltimore, Maryland in 1990. GDB is the official central repository for genomic mapping data resulting from the Human Genome Initiative. In support of this project, GDB stores and curates data generated worldwide by those researchers engaged in the mapping effort of the Human Genome Project.

GenBank (www.ncbi.nlm.nih.gov/Web/Genbank/index.html) is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. There are approximately 967,000,000 bases in 1,491,000 sequence records as of June 1997. GenBank is part of the International Nucleotide Sequence Database Collaboration, which also includes the DNA Data Bank of Japan (DDBJ) and the European Molecular Biology Laboratory (EMBL/EBI) Nucleotide Sequence Database.

4.2.1 User issues

The major genomic databases serve broad communities, whose users have vastly differing needs. In this situation several levels of user input and management review are called for.

To assure that all the database centers are "customer oriented" and that they are providing services that are genuinely useful to the genome community, each database center should be required to establish its own "Users Group" (as is done by facilities as diverse as NSF's Supercomputer Centers and NASA's Hubble Space Telescope). Membership in these "Users Groups" should be on a rotating basis, and should represent the full cross-section of database applications (small academic groups, large genome centers, pharmaceutical companies, independent academic laboratories, etc.). The "Users Groups" should be convened by each Center Director and should meet several times a year, with written reports going to the Center Directors as well as to the sponsoring Federal agencies.

Several briefers from database centers expressed concern that the "average user" was not well-informed about appropriate ways to query the databases, and that search

tools (e.g. BLAST) frequently were not being used in a sound fashion. To address this type of issue, DOE should encourage the database centers in consultation with their "Users Groups" to organize appropriate tutorials and workshops, to develop "crib sheets" and other instructional documentation, and to take further steps to educate targeted user communities in techniques for sound database use appropriate to their applications.

At a higher management level, DOE and NIH should continue the process of constituting independent panels every few years, to review the health of the entire suite of genomic database centers. These panels should provide independent peer review of every community database, including input from "Users Groups" as well as technical and management review of center operations. Inclusion of Computer Science database experts on the review panels will help facilitate exchange of information with the Computer Science community.

4.2.2 Modularity and standards

Too often database efforts attempt to "do it all"; i.e., they attempt to archive the data, provide mechanisms for cataloging and locating data, and develop tools for data manipulation. It is rare that a single data base effort is outstanding in all three areas, and linking the data too closely to the access and analysis methods can lead to premature obsolescence. For reference, the following functions can be identified:

Authoring: A group produces some set of data, e.g. sequence or map data.

Publishing and archiving: The data developed by individual authors is "published" electronically (i.e. put into some standard format) and accumulated in a network accessible location. This also involves some amount of "curation", i.e. maintenance and editing of the data to preserve its accessibility and accuracy.

Cataloging (metadata): This is the "librarian" function. The primary function of a library is not to store information but rather to enable the user to determine what data is available and where to find it. The librarian's primary function is to generate and provide "metadata" about what data sets exist and how they are accessed (the electronic analog of the card catalogue). Other critical functions include querying, cross-referencing, and indexing.

Data access and manipulation: This is the "user interface". Because the data volumes are typically large, computerized methods for data access and manipulation must be provided, including graphical user interfaces (GUIs).

The key point is that the various functions should be modularized, rather than tangled together in a single monolithic effort. The reason is obvious: computer technology, storage technology, data base technology, networks, and GUIs are evolving on a time scale much shorter than the projected lifetime of the data. Each technology

evolves on its own time scale and schedule. Therefore, the functions must be modularized to allow separate upgrading. Modularization also allows multiple approaches, e.g. to user access: simple, intuitive GUIs for some users, powerful search and combinatoric engines for others.

Data format standards are a key to successful modularity. The community should invest in developing a common "language" which includes definition of certain basic data types (e.g., "classes" or "objects" in object-oriented terminology). Data format conventions should be defined for sequence data, map data, etc. Where multiple standards already exist, investment should be made in translators. Some standardization of methods to operate on data objects is also desirable, particularly for the most frequent operations and queries. However, the user should be able to develop powerful customized methods and manipulation techniques.

Currently, neither standards nor modularity are very much in evidence in the human genome project. The DOE could contribute significantly by encouraging standards. Database groups should be encouraged to concentrate on the "librarian" functions, and leave the publishing and archival functions to other groups. Development of user interfaces and manipulation tools may also be tackled by database efforts, but it is not obvious that the best librarians are also the best GUI developers.

As part of the librarian function, investment should be made in acquiring automatic engines that produce metadata and catalogues. With the explosive growth of web-accessible information, it is unlikely that human librarians will be able to keep pace with the ancillary information on the genome, e.g. publications and web-sites. The technology for such search engines is well-developed for the web and needs to be applied specifically to genomic information for specificity, completeness, and efficiency.

Indexing and cross-referencing are critical database functions. It is often the case that the indexes which encapsulate the relationships in and between data bases constitute a far larger data set than the original data. Significant computer resources should go into pre-computation of the indexes that support the most frequent queries.

Consideration should be given by the database efforts to development of shell programs for genome database queries and manipulation. A shell is a simple interactive command-line interface that allows the user to invoke a set of standard methods on defined objects, and lists of objects. In the numerical world, Mathematica, Maple, and IDL are examples of such approaches. The shell typically has a simple syntax with standard if-then constructs, etc.

4.2.3 Scaling and storage

About 40 Mb of sequence data exists in the genome databases today, using a storage capacity of 60 GB (NCGR). By the time the Human Genome Project is complete,

these databases can be expected to hold at least 3 Gb of sequence, along with annotations, links, and other information. If today's ratio of 1.5 KB per sequence-base is maintained, 4.5 TB of storage will be required. At the very least, a comparable 100-fold increase in submission/transaction rates will occur, but we expect the transaction rates to grow even faster as genomic data are more complete and searches become more sophisticated. While these capacities and transaction rates are well within the bounds of current database technology, careful planning is required to ensure the databases are prepared for the coming deluge.

4.2.4 Archiving raw data

As the Project proceeds it is reasonable to expect improvements in the analysis of the raw data. Therefore *a posteriori* processing could be quite valuable, provided that *the trace data are archived*.

One of the algorithms used currently has been developed by P. Green. His base calling algorithm, PHRED, takes as input the trace data produced by the ABI instrument (chromatogram files). Quality parameters are developed based on qualitative features of the trace. Currently 4 such (trace) parameters are used. These are converted to quality thresholds through calibration on known sequence data.

Experiments conducted by Green, involving 17259 reads in 18 cosmids yielded the following results, comparing the error rates of the actual ABI software calling package to those of PHRED.

Method	Sub	Del	Ins	Total
ABI	4.26%	0.38%	1.47%	6.10%
PHRED	2.55%	0.58%	0.47%	3.60%

Of course, the distribution of errors should also be compared, error clusters have potentially serious implications for the assembly problem, more so than well isolated errors. Another potentially important consideration is the location of errors within the read.

It is not unreasonable to expect that the actual conversions, used in the PHRED algorithm, might be improved as the library of known sequence increases. Further, more than one conversion table might be required, depending on the general region of the genome one is attempting to sequence.

C. Tibbetts of George Mason University has developed a based calling algorithm based upon a neural network architecture. He has also worked to maximize the quality of the base calls through an engineering analysis of, for example, the ABI PRISM™ 377.

Whatever algorithms are used it is important that the called sequence of bases have *associated confidence values* together with an interpretation of what these values are supposed to mean. For example confidence values could be pairs of numbers, the first representing the confidence that the base call is correct and the second representing the confidence that the base called is the *next* base. One might also consider adding a third coordinate representing the confidence that the called base corresponds to one base as opposed to more than one. These values should continually be checked for *internal consistency*; *every* read should be compared to the assembled sequence. This comparison involves the alignment of the read against the assembled sequence minimizing an adjusted error score.

Finally, there are currently several degrees of freedom in sequencing. Two, that could yield different (and hopefully independent) processes are:

1. Using dye labeled primers versus dye labeled terminators;
2. Sequencing the complementary strand.

Correlated errors define an upper bound in the accuracy of base calling algorithms that *cannot* be surmounted by repeated sequencing using the same chemistry. Ideally the confidence values assigned to individual base calls would closely correspond to these intrinsic errors. This can (and should) be tested experimentally.

There are two final points on the issue of archiving the raw data. More powerful algorithms (enhanced by either a growing body of knowledge about the genome or by better platforms) could improve the reads, and hence enhance overall accuracy. Such developments could also enable re-assembly in some regions (if they exist) where errors have occurred.

4.2.5 Measures of success

Databases are crucial tools needed for progress in the Human Genome project, but represent large direct costs in capital equipment and operations and potentially large hidden costs in duplication of effort and training. We believe the *only* true measure of success will be whether or not these tools are used by researchers making scientific discoveries of the first rank. That a given database installation is “better” than another in some theoretical sense is not sufficient. There are examples in consumer electronics where the “best” technology is not the one chosen by the majority—a similar situation could easily occur with databases in the Human Genome project. We urge DOE to critically evaluate the “market impact” of the database efforts it supports by regularly surveying users and comparing with other efforts, supported outside DOE. Fundamentally, the operation of a major database is a service role—of very great importance and with real technical challenges—that may not be in the long-term interests

of DOE, assuming other satisfactory database tools are available to its researchers at reasonable cost.

4.3 Sociological issues

Until recently the biological sciences have been based upon relatively free-standing bench-top experimental stations, each with its own desk-top computer and local database. However a “sequencing factory” with high throughput faces new informatics needs: inventory management, a coordinated distributed computing environment (e.g. EPICS), automated tools for sequence annotation and database submission, and tools for sequence analysis. In addition the national and international Human Genome Projects must integrate the genomic information into a common and accessible data structure.

The broadly distributed nature of the Project presents a challenge for management of the informatics effort. In particular, across-the-board imposition of standards for software engineering and data quality will be difficult. The best course is for DOE to “choose its battles”, emphasizing the development of common standards in areas of highest priority such as database centers, while tolerating a diversity of approaches in areas such as advanced algorithm development for genomic analysis. In addition to standards, consistent “User Group” input and peer review are needed for all of the genome database centers.

It will be helpful to increase the level of participation of the Computer Science community in genome-related informatics activities. While the human genome sequence database is not among the largest databases being developed today, the diverse nature of genome applications and the need to combine information from several different database sources provide real Computer Science challenges. Reaching out to the academic Computer Science community to engage the interest of graduate students and faculty members has not been easy to date. The genome community continues to debate whether it might be more fruitful to educate biologists in computer sciences, rather than educating computer scientists in biology. In our view, both approaches should continue to be pursued. DOE’s informatics program should include outreach activities such as workshops, short courses, and other support which will familiarize Computer Scientists with the challenges of the genome program, *and* which will educate young biologists in those areas of Computer Science which are of importance to the genome effort.



12/5/97 U of Utah.

Exploring larger plasmid inserts - simpler upshot locally
& easier f'ed + more reads than pUC
larger insert gives more coverage.

Lyr. tech very good @ transposon mapping.
could scale & add 2-4k cuts. as receive
automatic & 1k/bp - for megas - would have to add
personal cuts

Could they become a mapping resource?

Sequencing @ a level of dev where they could obtain
good data but not 80% of time - more like 20%
Further dev of whole bird, invertebrates + trees.
Question of whether to continue @ ext.

Mapping capacity of Charter - 1800 lanes -
30 Mb/yr. of seq ready clones from 1 machine.

Pyrococcus genome - 600 fragments → 2.1 Mb.

NFI region - 3 overlapping BACs.

Have learned how to generate med-vicint plasmid like

Con. still have to bootstrap contigs.

need less effort prep BAC DNA.

BAC based mapping

Could make finger prints of see based bases (90,000 clones)
in 2 weeks.

looking to see if they could do this.

Doing hi density grids from Leturcki's lab -
making the grids w/ multiple pubs on each one.
100 grids - could look @ in 8 hrs.

Think continuation of rapid fingerprinting + STS content
mapping.

Applied for SGI collaboration - to provide mapping
resource. transfer plasmid technology + provide
transport mapped clones.

Clone retrieval - existing technology + infrastructure is
transferable.

Still believe that their fingerprinting is easier.

Need to work on:

1 - need to convince people they can seq on plasmids
best sequencing of ~~minichloroplast~~ chloroplasts. - all however seq
from their ABI. 40Kb contiguous - 370 rows.
of 230Kb BAC 265Kb assembly.

Continuing to demonstrate that seq of plasmid is accurate

2 - Need to make more robust. - consistent reads.
Mostly their big dyes & Chem. Their chem compet
w/ dye primer. big dyes make a big diff in ~~seq~~ successful
runs + can delete sample.

Feel they could provide in lg. scale transp mapped
& could transfer clone tracking.

Have a web based tool to see the contigs.

Will have 5 machines by Spring - 150 MB/yr.

3 have been running for 2 yrs.

Group not interested to make seg + take care

Plans to fully automate is 'just' next level of dev -
Automatic is Cigarettes aka Stuffed.



Division of
Clinical Sciences

Medicine Branch

National Naval Medical Center
8901 Wisconsin Avenue
Building 8, Room 5101
Bethesda, MD 20889-5105
(301) 496-0901
(301) 496-0017 FAX

U.S. Department of Health
and Human Services
Public Health Service
National Institutes of Health

November 10, 1997

Francis Collins, M.D., Ph.D.
NHGRI OD
31 Center Drive MSC 2152
Building 31 4B09
Bethesda, MD 20892-2152

Dear Dr. Collins:

We would like to invite you to attend a "think tank" on the use of "molecular combing" technology to expedite the completion of the physical maps of the human and murine genomes. We are envisioning a half-day meeting to take place on November 21, 1997 in Conference Room 10 of Bldg 31 C on the main NIH campus. The format of the meeting would be as follows:

12:30-2:30 PM--"Back to back" seminars by Drs. David Schwartz of New York University and Aaron Bensimon of the Pasteur Institute (These seminars would be advertised in the yellow sheet and open to the NIH community. The focus and data presented would be at the discretion of the speakers.)

2:30-3:00 PM-break

3:00-5:00 PM-focused discussion on the use of this technology for physical mapping of the genome

Dr. Klausner has specifically requested this type of meeting in order to be able to appropriately place this technology in the context of NCI and NIH initiatives to link physical and cytogenetic genomic maps. Please let me know if you will be able to attend this afternoon meeting and whether there are any other details that I can provide for you with regard to these sessions.

Thank you for your consideration of this invitation.

Yours sincerely,

A handwritten signature in black ink, appearing to read "Ian R. Kirsch".

Ian R. Kirsch, M.D.
Genetics Department, Medicine Branch
Division of Clinical Sciences
National Cancer Institute
kirsch@helix.nih.gov (email)

Adv. Committee for 5-yr. Plan

Aravinda Chakravarti

Barbara Wald

Rich Mathis

Chuck Ingley

Tom Pallord

Eric Ferson

Several Adv. Mtgs.

- Dec 2+3 - Functional Genomics Workshop

- Dec 18-19 Sequencing

- Mar - meeting re: mouse

- Feb des mtg?

May - Airlie House mtg - to bring together

Published in genome issue of Science.