2/28/97

Intl. Sequencing Mtg. - Bermuda

I. Introduction - Morgan

Hinxton - Pathogen sequencing facility
Area for courses
Conf. facility

II. Presentations - Cox, chair

A. Sulston

Chr. $\underbrace{22, X}_{1/2}$, 20, 6, 1 — GB4, use PACs for ordering

665 Mb

Strategy RH → PAC screening → Fingerprinting /STS analysis
→ GAP closure
→ Sequencing M13, some pUC
Phrap

Need to focus on finishing - software, some hardware

May have to do YACs - Xfer into windows

Status 10-20 markers/Mb
97 Mb covered in clones

Seq: 14.6 out in GenBank] publicly available
(Mb) 11.9 unfinished

17.5 ready to go

Total output 1996: 34 Mb
Ever: 52 Mb

Chr. 22 - substantial coverage now
X - less covered. Example on Xq22

Plan 30-40 Mb this year, 80 next yr.,
100 Mb/ yr after that

B. Waterston: 7, 22, X

Finished 1.85 mb

Submitted 2.95 Mb

In Finishing 11.6 mb   Bottleneck? Clones were limiting

In Shotgun 15.1 mb

In Library 3.5 mb

Strategy — STSs on chr. 7 (1/79 kb)

→ Bacterial clones (BACs by EG, Hyb. to PACs by Wash U)

→ MD fluorimager of BACs, Hind III

Use Sanger software

Minimal end-walking

600 STSs across 50 mb

128 BAC/PAC contigs

~250 kb avg size → 32 mb

175 clones underway, 21 finished

Shotgun/directed → M13, pUCs → phred, phrap → finish → consed

Q.C. Restriction digest ×3

Reassemble c̄ alt. versions of phred/phrap

Complete contiguity

Annotate

1:10,000 error rate sought   Anderson

Tech — Want 64-72 lanes/377. Gel loaders. Dye term. Tn transposons

Future - 96 lanes, pipetting station, Lloyd Smith sequencer

Projection 1 Mb/month now → aim for >2

C. Hudson

Screen 20X BACs, PACs c̄ STSs q 100kb

Contigs ~350kb

Areas of contigs → will need to be walked across
Use BAC end sequencing

Isolated STS → 12 clones
Fingerprint - pick one where every band is represented
in other clones HindIII, EcoRI
Know which go furthest → STS, screen, walk
? end sequence

Re-made RG pools

Use Genomatron - 100 STSs/day

Hawkins - Seq.
2.1 Mb finished — 11 gaps (size known by PCR)
— mostly chr 17 BACs
0.65 in finishing
Aim for 5Mb by 5/97, 20mb the next year, then 80

Sequatron II ↓ automate front end

QC/QA - Gel to gel fluctuation, track to well

Finishing! The bottleneck - Needs to be a production line

Human - mouse synteny

Fbaron has joined them. GDB is decimated

Assembles using "Alewife" - overlapping segments of 25-mers. Also provides a std.

Almost none of this is in GenBank!

C. Adams
Seq. non-overlapping BACs from 60 STSs on 16p
Using BACs as probes against human genome DNA
Arabidopsis scaling to 10,000 reads/month
Same team is doing human
Goal was 2.7Mb → have 2.6 mb, submitted to GenBank
735kb in closure
1.9 mb ready for random seq. (<5% E.coli, <5% pK)

Library Team, Random Team, Closure Team

Scale up of finishing is a challenge – 90% of it is
a software issue

Uses phrap & TIGR Assembler

Goal is 11 mb for next year "very ambitious"

Robot will help

12 genes per 2,363,073

1 gene/196 kb

There are 200 kb BACs c̄ no ESTs, no GRAIL hits

D. Gibbs

Progress – 3mb in GenBank (1.2mb in previous 4 mos)

ABI → BODIPY x̄ for walking (very small Fx)

Power of full length cDNA seq.

Concatenation

Have done 180 F.L. cDNAs

One expt 78 cDNAs → 100 Kb concatemer

Human vs. mouse also very helpful

Want to reach 15mb next year

Xpter, chr.12

Expand Dallas collab.

E. Cox – Goal is to do 200mb of chr.4 by 2005

Last year target: 2.5mb – won't make that

chr.21 EPM1 1.2 mb

D3 0.3 mb

chr.4 4q25 5mb

In GenBank 100Kb finished

1.2mb of clones > 3kb

Whole genome radiation hybrid maps — G3, in press

Can map to 240 kb theoretically (300-500 more realistic)

GB4 1 mb (1.2-1.5)

But if coverage isn't random, won't know there are gaps c̄ larger bin sizes

Transposon method — vulnerable to bad libraries ←

Chip: 140 × 25bp standard design

<1% false ⊖ <2% false ⊕

* Need to understand this

PCR up STSs, 800 at a time, hyb. to chip

Use to determine tiling path, check assembly

Cost is 1.5¢/bp

200 3kb clones — end sequence, then design chip to get tiling

F. Fiona Francis (Lehrach)

Planning 6 Mb over 3 years 1-2-3

3 groups — Rosenthal, Lehrach, Max Planck

chr. 21 — seq, ready maps

Hyb screening c̄ cosmid lib & PAC (BAC later)

Some FISH, Restr digest like Wash U for minimal tiling path

Shotgun into pUC (no picking robot), Phred/Phrap

Xp22 PAX 243 kb, 9 cosmids

In progress 21q, Xq, 17p

Using oligos to preselect shotgun clones (8-mers) by bar-coding → more even spreading

No data

G. Weissenbach - Haven't started yet. Announced by Minister of Research,

$14M/yr.

In Evry, near Genethon   30-35 people from Genethon will move over   joint venture c̄ CNRS, private Co (tech, xfer) to allow hiring

Start summer 1997

Sign lease in a couple of weeks -

office bldg, will need 4-5 mos. to renovate

Projects - In house

Collaborative - eval. by scientific committee. Academic

Ratio?

There is also "Steering Committee" which could change priority

Data release & I.P. will be decided by Steering Com.

In house → release more likely

Collabs → different

Will do some Arabidopsis, probably some microorganisms

Also tetraodon

TGS is Gen Set's private facility (5' ends of cDNAs, 30-50K)

H. Mattick Australia

$8M/yr. Voted by Fed. govt.

Facility to begin function mid-year

Melbourne - Simon Foote Dick Cotton

Genotyping, mutation data. 8M genotypes/yr.

Queensland - sequencing, Mattick

Expect ~30 ABIs   1500 templates

Have $ for infrastructure, not projects -

will need to draw on other sources - a problem - funding agencies [illegible] +

Service sequencing? - ESTs for plants

In house? Pathogens. Human [illegible] - clones to be provided by suppliers

I. Rosenthal

1.5 mb in GenBank now

Targets – Xq28 3Mb

Xp11 2.5mb

X-PABA 1mb

21q 25mb

7q 7q22 7mb Scherer/Tsui

7q32 0.5mb

Mouse syntenic region of 3mb – Xq28

1300 reads/day → 3000 by 5/97?

20 ABI's (16 bought by industry)

Bloecker

German Human Genome Project – work ē

IMB Rosenthal	Lehrach	Brocade	
4	1	1	← start 5/97
9	2	2	
15	3	3	

Have 6 mb available

Doing comparative Seq. in Fugu; Disease gene; rhizobium

Zebrafish – no organized effort

Very interested in methylation

J. Green (Olson)

Fidelity - 2X validation of all sequence-ready clones, using methods adequate to detect small (<1kb) colligations, deletions, Xposon

→ use as start point

Accuracy: <1/10Kb

Submit base-specific error probs.

Independent test of assembly accuracy

Contiguity – All gap sizes estimated, all contigs oriented
and ordered within the chromosome

MCD mapping

Chr. 7   2mb mapped   7q31.3

HLA   "   700kb seq

Mouse TCRα

340kb submitted

Bottleneck in editing

Expect to meet 2mb goal for year 1

Doesn't state 2nd year goal – waiting for $

Discrepancies –

Chr. 7   0 in 2x38802bp

HLA   2 in 2x43084 bp

1 was a phrap error

1 cosmid mutation 12bp ins/del

K. Chen – ACGT, div. of ABI – collab. c̄ Schlessinger

20 people, 4 groups – ~~10~~ 11 ABIs

New institute in Shanghai (ABI, Sequenom)

55% of budget is govt. grants

X   2.4mb, at a rate of 3mb/yr → 0.5Mb in GenBank

Micro – Ureaplasma 760kb, 99% done

Arabidopsis 0.4mb/yr   3 in 1998

Ordered shotgun   30 by 2000

→ See NAR

10kb clones (λ)   0.5 mb/tech/yr

Mapping done by Schlessinger → BACs

New dye primers – lower background (better spectral sep.),
equal mobilities.   4 mos.

Sakaki not here - broke shoulder skiing

L. Fujiyama - Japan

4 groups - JICSD → JSC

Nakamura - chr. 3, 8, 9

Sakaki - chr. 21 }

Shimizu - chr. 21/22 } 21q

Fujiyama - }

Sakaki

2.7 mb finished in 3 contigs

500 kb to be finished by end of March

Next FY 3.4 mb in 4 regions - have contigs of part

Directed deletion method

Testing Hitachi capillary sequencer (96)

Not sure if it will be commercial

Expanded facility - scientists agree, govt. slow to respond.

Start FY98?

Mb: 15, 30, 60 ——→ (7 yrs)

(98) (99) (00)

chr. 21 h21/m21 m21/h11 m = mouse syntenic region

Budget - economic decline is affecting

$60M will be severely cut ($20M?)

Data release by JST

900kb available

Sakaki has his own Web site

M. Evans Chr 11, 15

Chr 11 - 905 STSs

17,965 end sequences from cosmids

Chr. 15 - harder, less well mapped

High density grid hyb. c̄ pooled STS-specific oligos

4 restriction enzyme fingerprints of each PAC

Chr. 11   11p → PAC contig of >3.5 Mb

2465 STSs, screened against 465

3185 PACs

467 fingerprints

216 PACs → 3 c̄ mixed signals (1.3%)

Seq. strategy

Sagian

Auto-finishing – use phred/phrap output and high capacity oligo synthesizer

Accuracy – want Phrap scores > 40

Phase I] 29 → >1 Kb ordered (II) or unordered (I)

II

III Closed – 10^{-3} to 10^{-4}] → GenBank

IV Uniform $>10^{-4}$ accuracy

Annotation

155 Kb 11p13.3 color coded output, showing overlap

Available on Web

Per base sequence displayed

Automation

<10¢/nt small scale

MerMade oligo 96/192 300/day → Avantec, Inc., c̄ no-cost license from UT

Sagian robot (Beckman purchased) – 3 m rail

DNA seq'r – Astral. 7 months. A lot like ABI. Uses hyperspectral imaging

Chr. 11 – needs coordination, designate by STS, not band

N. Palazzolo

Won't present JGI

800 Kb/month

Physical map – random light shotgun, build paths, transposon

Quality – all double stranded

Hardware - New space needed

colony pickers, oligos ...

Partnership c̄ Motorola: Chicago group designs their 95 factories, does their tech transfer

Volume, quality, cycle time, cost

Need precise goal definition - we don't have it

Peer review is impossible

Benchmarking - metrical tools

Bottleneck analysis - predicts where to put R&D

Process model - must have predictive value. Looks only at volume

Cost model (Motorola paid)

Cost accounting

Pick-a-mix - cost models.

Predicts effects of changing volume on a spreadsheet

Did an LBNL review - cost $250K, 3 mos.

Chr. 22

O. Roe - Chr. 22 3.8 Mb in Gen Bank

He doesn't do mapping

Chr. 9 (bcr-abl) → Rowley collab.

Interested

Aspergillus

N. gonorrhoeae 2.2 mb] 95% on website

Strep. pyogenes 1.9 mb]

Sees 2 genes/100 Kb

"40% of the human genome is sequenced" - The Alens

III. Data Quality

Day 2

IV. Cost - Palazzolo, chair

Value/Danger

Methods

Validation

Need to collect data in a serious way

Methods - separate out R&D ?

1) Cost model extrapolations - easiest, but prone to error

Ex oligo synthesizer, miss cost of reagents that had to be thrown out

2) Cost accounting

Separate budgets for each activity

Estimates turned out to be 2-3X low

3) Cost models

Define product, establish process flow model, fixed protocols, databases on cost [materials, equipment, stock solns., labor

→ Identify & manage R&D opportunities

Genome Cooperative Purchasing Group ?

Govt. can't take a leadership role

4) Output-based

NHGRI to take a role ?

Do audits in a couple of places

Then send around an MBA to instruct the rest

Rosenthal - unhappy c̄ generality

Aim for 30¢/bp

"Game of liar's poker" - MP

	$ in/out	other $
Gibbs	50¢	60¢

V. Data Release

VI. Etiquette – John/Bob

Mapping
Sequencing] Claims may be different

Mapping doesn't entitle Sequencing

Sanger Center has gotten into conflict on chr. 1 c̄ TIGR
Their mapping strategy focuses on whole chromosome

X chromosome – different mapping resources were very helpful

Mapping can be redundant, Sequencing shouldn't be

Sequencing – claim no more than a year

HUGO site
HSM Index – Flat text file
Don't need to make the link explicit

FC proposes giving it to NCBI

Lipman: GenBank postdoc could curate

Cameron: EBI could support too – be careful about not calling it GenBank

NCBI/EBI/DDBJ – May Advisory Mtg. Put in other organisms too?

HUGO Council will meet next week

Genethon markers as the boundaries

Minimum size – Megabase? Between Genethon markers agreed to

Concern that ~~[illegible]~~ small scale efforts not be injured by claims

↘ Is this happening?

Maximum - a year's worth
No more than 3-5X what you did last year
Sequence-ready map is a significant investment -
it's tacky for someone else to move in on it
Specific issue of chr. 1
Should Sanger be expected to turn over maps?
To TIGR?
Ex: Chr. 11 Peter Little wanted to do 11p13 and 11p15
Overlap c̄ Evans?
End up c̄ 2 sequence-ready maps

VII. Annotation
Standards? What should be submitted?
"Electronic BSE"
Can look at 1° data to check for Ψgene; frameshifts -
producing centers are in a better position to
do this than users
Should all traces be made available on the internet?
Storage of traces? Tape → optical disk

* [John Spouge, NCBI MD PhD
Sen. Sci. - assist c̄ data exchange
plan] *

What about non in-silico methods?
Software: What option is best? Algorithm to synthesize?
Lipman agrees its database better, in fact, shouldn't
even report unless you have real exptl. data
"Suspected gene" is helpful - exon structure isn't reliable

ESTs → through end of 1997 from NCI, Merck, Genentech, BMS

8000/week being asked by NCI

3' ends + sizing

Subtracted libs / normalized libs ?

Lifetech, Stratagene → 20 libs each

Soares → 15,000 used to subtract a pool of libs →

4x ↓ in those clones

Cluster algorithm to find all > 1 rep.

singletons is rising at a slower rate

than clusters now (28% → 21%)

Mapping

Cox urges high resolution panels

MIT 3000
Sanger ~~6~~ 6,000
Genethon 6,000
Stanford 2,000

most on GB4.

17,000 more by June!

Update Web then → no, sooner!

RHdb

Can get data now but have to go to 4 webs

Schuler

↓

WWW/NCBI

Full clone seq.? NCI will fund ~15,000

Mouse: 1-2 mb comparisons beginning to appear

1 Mb of chr. 11 in Germany (won't say where)

Xq28 2.5-3 mb Steve Brown IDS

1 mb Rosenthal

Mouse IDS /
12p13 (CD4) ~~PGK~~
Xq (PGK)

Gibbs

each ~0.2 Mb

MIT - 1 mb mouse nu / human 17

"

Roe - 500kb Drbearge 2 BACs chr. 22 + Reeves grant

→ ~1 mb

Sanger 1-2mb BRCA2

Bruce – useful to find genes missed by ESTs

10% of human! (André)

FC – no more than that!

Mouse ESTs

Aim for 30,000 full clone seqs. in next 2 yrs.

EC consortium to map mouse ESTs to

Goodfellow RH panel

Oxford ESTs

Genethon will do 3000 – ? TOTAL?

RH panel is low resolution

Not much enthusiasm for higher resolution because it wouldn't coalesce

Feb. 27-28 – March 1

Evening session? Free afternoon?

Statement:

Needs more explanation of rationale?

And moderation of statement re Germany

Michael will work c̄ Ursula to re-word