Press Conference 10/29/02

Welcome

Today we launch the IHMC

Nine groups from 5 countries

Supported by multiple funding agencies, including support from the private & charitable sector

$100M, three years, all data goes into public domain

Introduce the group

What is the haplotype map?

When completed in three years, it will ~~be~~ provide the critical missing link to allow researchers all over the world to uncover the hereditary factors in common Dz, and drug response

In regard to its consequences for medicine, this is in the same category of ambitious, groundbreaking, foundational projects as the sequencing of the human genome

personalization of the genome

10 million common variants - SNPs

Simplest model for gene discovery: case-control

But can't afford it

Hap Map is a powerful and conceptually elegant short cut

SNPs are correlated with their neighbors

Defining how those neighborhoods are organized will save a factor of 30 - 40

~~[illegible] [illegible] [illegible] - [illegible] of [illegible] ?~~

→ Gasleaks in NYC or DC

Sample Set of DNAs - see ELSI

Differing Technologies

International participation

The initiation of the project brings us a ~~quantum~~ major step closer to a new era in medicine, where disease can be prevented, treated, or even cured in a powerful new way, based on a rational understanding of the human genome.

10/27/02

FC Comments - Hap Map

Sun PM Historic gathering on a beautiful Sunday in the fall
Our goals are no less ambitious than to ...
International collaboration a defining aspect
This will be much harder than sequencing the human genome
But we can build on that experience
We will all get to know each other very well over the next 2 years

Mon AM Principles: You represent $100M
Openness, collegial spirit
Joint decision-making about science
Explicit measurable milestones - timetables, deliverables
High emphasis on QC/QA
Expectations of some adjustments & reassignments along the way - we must not allow part of the genome to fall far behind
Importance of ELSI concerns about samples - which may inspire some frustrations.
Remember the HGDP!
Early data release is critical
Objective external oversight
Funding agencies
External Scientific Advisors
Failure is not

Competition in technologies is good

Different definitions of the map in different parts of the genome is bad

10/27/02

HapMap

I. Welcomes

A. FC

B. Tanaka

C. Kennedy

D. Hua Han

E. Holden not here

II. 10 minute Overviews

7:25 A. Nakamura

SNP Discovery → 190,562 SNPs as of 7/17/02

78,000 have allele frequencies

Invader assay + multiplex

Have done 60M SNP types

82% successful

Blocks correlate ī recombination frequency

7:37 B. Bentley

Technology: ~50% of SNPs provide robust assay (m.a.f. > 0.1)

Illumina 288-plex 98.7% data has 99.87% accuracy

Chr. 22

Ask → Chr. 20 - seq'd to 8.8x, does get nicely even coverage

maf 0.2 or 0.1 has half D' at 65kb

0.04 → stretches out

LD Chromosome profiles: robust 20 - 100% of SNPs

19 → 9 → 4 kb See some breaking up of blocks

??? Didn't expect this

Need more SNPs

C. Hudson

5 years old. Expansion 1/02

Crohn's Dz

ABI TaqMan, Orchid, ---

→ Mass spec was the plan

Moving to new bldg & Illumina?

D. Yang

Latecomer, but prepared   1% sequence → 10% HapMap

As great as the HGP

~~Many~~ Funding agencies → $6M seed

China HapMap Consortium - Henry & Board are responsible

Beijing GI

NHGRC of North China - Beijing

NHGRC of South China - Shanghai

Also 1-2 groups in Hong Kong

Probably a group in Taiwan

All have done pilots

SNuPE on MegaBASE ⎫
Mass (Bruker) ⎭ Exploring

Also sample collection

E. Chee

1152-plex. 1M genotypes/day. Very pretty data shots

F. Altshuler

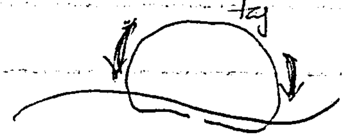
8^{10} Sequenom for launch

G. Gibbs / Willis

ParAllele Tested 1500 probes, including

90 locally verified & tested markers (>80%)

Padlock



$\xrightarrow[\text{single dNTP (4 reactions)}]{\text{polymerase}}$ $\xrightarrow{\text{ligation}}$ $\xrightarrow{\text{exo}}$ $\xrightarrow[\text{(uracil base)}]{\text{cleave}}$ $\xrightarrow{\text{amplify}}$ $\xrightarrow[\text{(Affy)}]{\text{chip}}$

1500 multiplex

all in dbSNP

Collab c̄ Rich Lifton → 1500 SNPs on ch. 6 (10 Mb)

85% conversion rate (no repeats)

by BLAST — threw out ~10%

Accuracy good

H. Pui Kwok (UCSF/WU)

Primer extension c̄ FRET

1.7M SNPs Assay (all not in repeats) designed

Conversion >90%

X, 7, 5, 17

↓ only 07

Molecular haplotyping — US Genomics collab. + padlock

I. Zwick

10/29/02

HapMap Issues

Sample availability   Dec. 15   Mark Leppert will be here Tues.

Possible to start on CEPH before consent is complete (as a pilot?)

Possibility of further slippage?

Press conference – agenda? Who speaks? Do they know? FC's remarks?

Sun. PM – Order of presenters? Who chairs?

Regional assignments – is there an easy solution?

Sample sizes?

MEXT? Extra copy?

Plan for TSC mtg?

FC comments – both Sun PM & Mon AM?

Advisory structure – Steering Committee
Oversight Committee
External Advisors

Plan A:

1) 3Ø trios CEPH = 93
2) Asia + Africa = 93 } 192 wells

Plan B:

Double → 384 wells

10/28/02

HapMap

Questions - Issues

① How to start?
CEPH + what? PDR? Coriell African samples? Nakamura Asian samples?

② Primate samples - how many and what kind?
DNA - need permit to ship? Huh?

③ Chromosome 20 data - analyze by multiple methods!

④ For validated SNPs seen by Perlegen, vs.
those not seen, what is MAF? → done by Steve

⑤ Would Perlegen continue the practice of identifying
public SNPs they have seen, even after original
release? (But only samples 1/4 of the genome)

⑥ Consequence of shifting to individual samples in W. Africa?

⑦ How to set up a test set of SNPs that everyone
will try to convert

⑧ Is it worth doing 3-generation pedigrees on <u>all</u> of the CEPH?
Needs modelling to see if error-checking and
definition of long-range haplotypes is worth the redundancy
of chromosome sampling

⑨ What motifs define recombination and determine block boundaries?
Analysis

Questions (cont.)

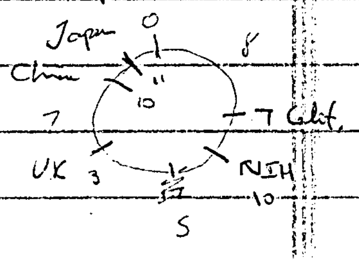
~~(10) Will Perlegen continue this gift?~~

(10) What's an acceptable error rate for genotyping?
Simulate errors by degrading a good data set

(11) Standards for internal QA/QC. How many replicates?
How to organize the external QA/QC

(12) DCC selection of SNPs across the genome, with priorities noted - getting that started

(13) How to pay for another 2x WGS coverage?
Celera plans to investigate?



(14) Times for phone calls
10 AM Eastern or 6 PM Eastern?

(15) Can Celera data be used for confirmation? (No genotype data, just a yes/no)

(16) Data release - what's the plan?

10/28/02

HapMap

I. Intro - FC

II. Samples

CEPH - Existing trios - Utah only

Japanese - 100 individuals (collect 130-140)

Han Chinese - ditto

Yoruba - 50 trios

A. CEPH - Consent form wasn't finalized until late June (TSC)

Priority list - 84

28 deceased

56 → 46 have agreed by phone → 18 have signed, rest in 2-3 wks.

Need to find those 10!

mid-Nov. 2002 → one plate of trios

About 65 trios ready by early December

B. Nigeria - Howard

Community consent is NOT the idea

Community engagement - but what's the community?

HIV - will need to screen

Expect sample collection April - August

Cell lines → Oct. 2003 to labs

Why are we doing trios? (That's what IRB ~~[illegible]~~ currently has approved)

PDR? Use a plate to cover diversity

→ find common haplotypes, do SNP validation

C. Japan

D. China

Primate samples?

Chimps c̄ bonobos are available now from Coriell

But chimp sequence will be available very soon – 1 year

III. Chromosome Allocations – Nakamura

Revisit after 6 mos?

IV. dbSNP

2.2M mapped to the genome (out of 3M unique)

507,696 Validated (Perlegen + others)

↳ 136,960 c̄ MAF >10%

Mean intermarker spacing 1.146 kb

N5	9899	
N10	5728	90%
N20	3000	

SNPs per 10 kb window

Variation by chromosome – is it more than ~~of~~ expected?

What is MAF spectrum of Perlegen ⊕ SNPs

Do we capture non-polymorphic SNPs?

V. Scientific strategy – Altshuler

Overall goal → tag SNPs (hierarchical plan?) – depending on GRR

Optimize power

Optimize efficiency

Blocks are a means to the end

Common? Is that important?

Another goal is software/methods to interpret data

Uniform properties across the genome

Previously discussed

Wanted to have 99% power to detect 5% allele freq. in any of the populations (at least once)

Is that the right number

TSC pilot – look at rare alleles and see what haplotypes look like

Cover the entire genome

Haplotype blocks vary in size <5kb to >100kb

Need to model the difference in 1% vs. 10% cutoff ?

In 2-4 years will we want to do all the SNPs ?

Proposal:

Start c̄ ≥1 SNP / 5–10 kb (want at least ~~4~~ markers/block)

↳ Some will drop out   one working / 10 kb

→ useful intermediate product for the scientific community

→ Phase I   (must include non-CEPH)

Working polymorphic SNP every 4-5kb (Bentley!)

Being sure we can "finish"

Each group to push 10Mb to the next phase?

FC: How could most groups do this?

SNP Selection:

90% of HGP & TSC assayable SNPs are "real" (and 90% are assayable)

~60% have maf >0.1 in any one population

"Double-hit" – Both alleles seen ≥ 2x

Validation 91% (vs. 90%)

maf > 0.1 88% (vs. 55%) ←

Mark Daly → resequencing saves money?

Looks like it

Even more so if Perlegen continues to validate

Propose:

Prioritize validated SNPs (say cSNPs)

Also pick a common set for all platforms? Worry that these will be tinkered with more

DCC to <u>serve</u> the SNPs?

– For test of conversion success, Yes

– For the entire project? Yes, but basically serve <u>all</u> of the SNPs and let centers decide

Selection of SNPs

Should we include all known putative functional SNPs?

Sanger project is at early stage

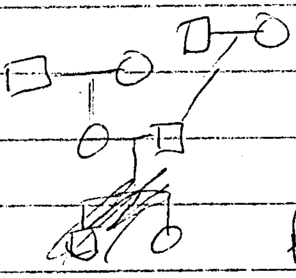
Yusuke 78,000 have allele frequencies → 60,000 $\bar{c}$ MAF >0.1

# of samples

88 chromosomes to see 5%

Trios vs. individuals

→ error checking



Lon wants 3-generation pedigrees!

Needs more

Population samples – proposal

Phase I – Genotyping of an even map across the genome in CEPH and in a multiethnic panel

Phase II – Finish each region in the complete & newly collected & consented samples

CEPH, Nigeria, China, & Japan

Chinese
Japanese samples would be added in as soon as possible – i.e
– IA
– IB (add consortial samples)

What's the multiethnic panel?
PDR?
African-American
Asian
? Samples available from Coriell + Nakamura
all 3?

VI. Bentley – SNP Discovery

Argues we want to map all blocks > 10 kb
Need 5 SNPs → 10 selectable (assumes 50% conversion)
Recognize that ~10% of the genome may be intractable

Chr. 20 is 1/2 kb now

Current: 2.2M SNPs (1.3M TSC random + 1.5M lumpy overlap)
+ 0.4M Mullikin

53% of genome has 5+ SNPs in 10kb window
76% 2+

Getting more SNPs:
Targeted: PCR
mapped clones – inventory problem
Random: ↑'d double-hit SNPs → ↓ cost of assay development

Correlation of SNP density & recombination may make SNP poor regions also those c̄ long LD
⇒ maybe only a modest effect

TSC SNPs are potentially biased

Hap goals is 90% of genome c̄ 10 SNPs/10 kb

Would take another 5x (added to existing 2x)

But what does this look like wrt M allele freq.

of SNPs you discover at each increment

Assume $50 per PCR discovery of a SNP

Seq 6-12 DNAs c̄ ~~[illegible]~~ 500bp amplicons

This model does not account for:

1) Skipping 10kb blocks that are already in LD, where targetted PCR is not needed

2) double hit SNPs will translate into reduced costs of genotyping (by 30%) — Daly

55% have MAF s̄ this %

85% double-hit

Celera SNPs

Connection to TaqMan? Allowable to use for confirmation?

ELSI test is failed

Even 2x gives you another ~1M double hit SNPs

Is ~~[illegible]~~ conversion rate higher for targetted? Maybe — but depends on how many DNAs were seq'd in targeted phase

12-1~~4~~M reads? → $20M Could be done in 3 mos?

Who would do this?

BCM
WIBR
SI

Has to be done immediately

New $?

Redirected $?

VII. Scope of QA Discussion – Chee

A. Leave aside conversion rate
→ Genotyping success
(DNA Quality as an issue)

Accuracy
Concordance – results from 2 technologies (need a 3rd to judge)
Strand correlation
Reproducibility
Non-inheritance rate
(H-W)

PCR 1st
FP
Invader
Mass Spec
} low/medium plexy

PCR 2nd
Illumina/GoldenGate
ParAllele/Padlock
} high plexy

B. What are acceptable error rates?
Lon Cardon: ≤ 0.3%
But has anyone done the simulation?

C. Cross checking – TSC
Discovery PCC randomly selected SNPs from previous Q
→ 2 Genome Centers reseq'd (some outside of TSC)
Several thousand

2) Allele detn. TSC

5% were tested by 2 centers

Do we want round robin or external checking?

① Internal QA/QC: (After a std. filter to weed out horrendous assays)

Reproducibility – x% of samples should be ~~blinded~~ duplicates

Inheritance pattern → Should be posted by each site

② DCC

③ External

Start = round robin – DCC can organize that

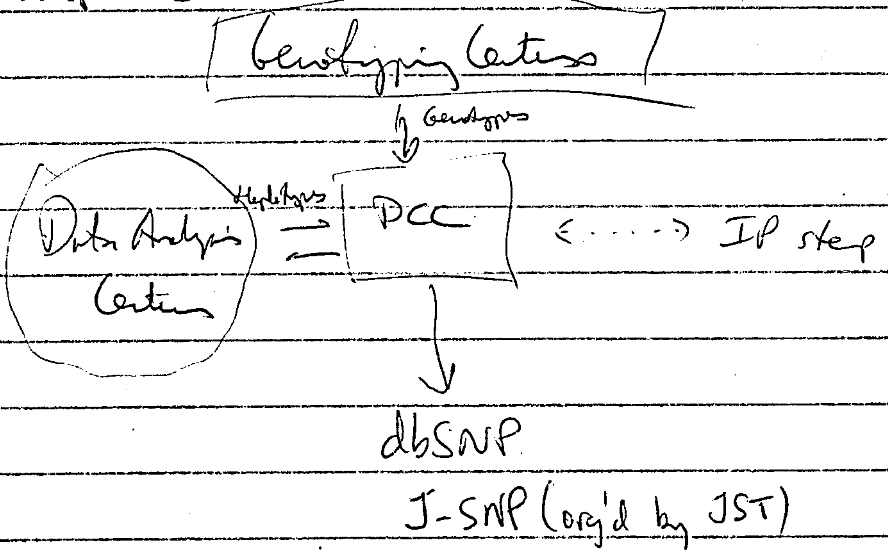
Can you apply quality standards to each genotype?

It would be a shame to lose this information, for methods that can measure it / estimate it

VIII DCC – Stein

Why have a DCC?

How it operates



Mirror web site of DCC

IX. Intellectual Property - NOT   Holden

Data release   ~~Dat~~ SIRs; TSC was filer

Every quarter? with protective IP

How to deal with arbitrariness of haplotype block construction?

What about SNPs that don't fall into blocks?

File as with TSC I

a. SNP discovery

b. Genotyping

Quarterly release?

Principle - don't let this be rate-limiting

Is this only a US issue? - Oversight Group

Genotypes as an uncertain part of this

Hale & Dorr advised this so needed

Need more opinions!

Japan wants to be listed in the filing

X. Analysis - Chakravarti

Resampling existing large data sets?

Haplotype blocks

1) Block properties and definitions

2) Inference of block structure - interblocks

3) Stopping rules (depends on #1)

4) Sampling properties

Genomic properties

How will this all happen?

Collaboration c̄ funded ACs

Cooperation c̄ DCC

Who are the analysts?   Jamborees? GAW-like

~30 people 2½d mtg

in the next 2-3 mos.

Who wants to participate?

Nakamura - 2 candidates

Bentley - Lon Cardon

Hudson - "recruiting"

Yang -

f/u needed

How do we decide <u>what</u> definition to use for the project

Who decides? Steering Committee?

Analysis group

Production component

Development component

International HapMap Project

Press Conference
Tues., Oct. 29, 2002

Speakers

Francis Collins
National Human Genome Research Institute

Yusuke Nakamura
University of Tokyo

David Bentley
The Wellcome Trust Sanger Institute

At Table

David Altshuler
Whitehead Institute for Biomedical Research

Aravinda Chakravarti
Johns Hopkins School of Medicine

Mark Chee
Illumina, Inc.

Ellen Wright Clayton
Vanderbilt University

Georgia Dunston
Howard University

Richard Gibbs
Baylor College of Medicine

Arthur Holden
The SNP Consortium

Thomas Hudson
McGill University

Pui-Yan Kwok
University of California, San Francisco

Huanming Yang
Beijing Genomics Institute

Also Attending

~~**Hua Han**~~
~~Chinese Academy of Sciences~~

Karen Kennedy
The Wellcome Trust

Mark Leppert
University of Utah School of Medicine

Ichiro Matsuda
Kumamoto University

Satoshi Tanaka
Japanese Ministry of Education, Culture,
Sports, Science and Technology (MEXT)

Changqing Zeng
Beijing Genomics Institute

Genome Canada

CAS

NIH NEWS RELEASE

NATIONAL INSTITUTES OF HEALTH National Human Genome Research Institute

EMBARGOED:
Hold for Release at 1 p.m. EST, Oct. 29, 2002

For additional information, contact:
Geoff Spencer, NHGRI
[redacted]

International Consortium Launches Genetic Variation Mapping Project

HapMap Will Help Identify Genetic Contributions to Common Diseases

WASHINGTON, Oct. 29, 2002 – An international research consortium today launched an approximately $100 million public-private effort to create the next generation map of the human genome. Called the International HapMap Project, this new venture is aimed at speeding the discovery of genes related to common illnesses such as asthma, cancer, diabetes and heart disease.

Expected to take three years to complete, the HapMap will chart genetic variation within the human genome. By comparing genetic differences among individuals, consortium members believe they can create a tool to help researchers detect the genetic contributions to many diseases. Where the Human Genome Project provided the foundation on which researchers are making dramatic genetic discoveries, the HapMap will begin to make the results of genomic research applicable to individuals.

"The HapMap promises to accelerate medical research around the globe in many different ways," said Yusuke Nakamura, M.D., Ph.D., director of the University of Tokyo's Human Genome Center and leader of the RIKEN SNP Center and the Japanese group working on the HapMap. "Not only will it lead to the identification of genes related to disease, it should help to pinpoint genes that influence how individuals react to various medications – discoveries that could improve drug design and lead to the development of diagnostic tools aimed at preventing adverse drug reactions."

To create the HapMap, DNA will be taken from blood samples collected by researchers in Nigeria, Japan, China and the United States. Initially, researchers will work with samples from between 200 and 400 people in widely distributed geographic regions. Samples will be collected from the Yorubas in Nigeria, Japanese, Han Chinese and U.S. residents with ancestry from northern and western Europe. A very careful sampling strategy has been developed to ensure that participants can give full informed consent. No medical or personal identifying information will be obtained from the people providing the samples. The samples, however, will be identified by the population from which they were collected.

"Studies like this must be done as ethically and transparently as we can," said Ellen Wright Clayton, M.D., J.D., of Vanderbilt University, who is chair of the group that is addressing the project's ethical and social issues. "For the HapMap project, we have devoted a lot of effort to achieving both these goals in order to do truly responsible science."

The samples will be processed and then stored at the Coriell Institute for Medical Research in Camden, N.J., a non-profit biomedical research center that specializes in storing living cells and making them available to scientists for further study.

Researchers from academic centers, non-profit biomedical research groups and private companies in Japan, the United Kingdom, Canada, China and the United States will analyze the samples to create the HapMap. The results will be made quickly and freely available on the Internet in keeping with the data release approach of the Human Genome Project.

Public funding for the effort will be provided by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Tokyo; Genome Canada in Ottawa and Genome Quebec in Montreal; the Chinese Academy of Sciences, the Chinese Ministry of Science and Technology, and the Natural Science Foundation of China, all in Beijing; and the U.S. National Institutes of Health (NIH) in Bethesda, Md. The SNP Consortium (TSC) in Deerfield, Ill., will coordinate private funding, while The Wellcome Trust in London will provide charitable funding for the United Kingdom portion of the project.

Understanding Variation

The International HapMap Project builds on the freely available sequence of the human genome produced by the International Human Genome Sequencing Consortium. Although research shows that any two people are 99.9 percent identical at the genetic level, the 0.1 percent difference is important because it helps explain why one person is more susceptible to a specific disease – say diabetes – than someone who is less susceptible. By studying the patterns of these genetic differences   or genetic variation   in many people, researchers expect to identify which differences are related to disease.

"The goal of studying the human genome has always been to provide health benefits to all humankind. This project should be seen in that grand tradition," said Francis S. Collins, M.D., Ph.D., director of the National Human Genome Research Institute, which is part of NIH, U.S. Department of Health and Human Services. "The HapMap will provide a powerful tool to help us take the next quantum leap toward understanding the fundamental contribution that genes make to common illnesses like cancer, diabetes and mental illness."

Genetic information is physically inscribed in a linear molecule called deoxyribonucleic acid (DNA). DNA is composed of four chemicals, called bases, which are represented by the four letters of the genetic code: A, T, C and G. The Human Genome Project determined the order, or sequence, of the 3 billion A's, T's, C's and G's that make up the human genome. The order of genetic letters is as important to the proper functioning of the body as the order of letters in a word is to understanding its meaning. When a letter in a word changes, the word's meaning can be lost or altered. Variation in a DNA base sequence – when one genetic letter is replaced by another – may similarly change the meaning.

More than 2.8 million examples of these substitutions of genetic letters – called single nucleotide polymorphisms or SNPs (pronounced snips) – are already known and described in a public database called dbSNP (http://www.ncbi.nlm.nih.gov/SNP/), operated by NIH. The major source of this public SNP catalog was work done by The SNP Consortium (TSC), a collaborative genomics effort of major pharmaceutical companies, the Wellcome Trust and academic centers.

The human genome is thought to contain at least 10 million SNPs, about one in every 300 bases. Theoretically, researchers could hunt for genes using a map listing all 10 million SNPs, but there are major practical drawbacks to that approach.

Instead, the HapMap will find the chunks into which the genome is organized, each of which may contain dozens of SNPs. Researchers then only need to detect a few tag SNPs to identify that unique chunk or block of genome and to know all of the SNPs associated with that one piece. This strategy works because genetic variation among individuals is organized in "DNA neighborhoods," called haplotype blocks. SNP variants that lie close to each other along the DNA molecule form a haplotype block and tend to be inherited together. SNP variants that are far from each other along the DNA molecule tend to be in different haplotype blocks and are less likely to be inherited together.

"Essentially, the HapMap is a very powerful shortcut that represents enormous long-term savings in studies of complex disease," said David Bentley, Ph.D., of the UK's Wellcome Trust Sanger Institute.

Since all humans descended from a common set of ancestors that lived in Africa about 100,000 years ago, there have been relatively few generations in human history compared to older species. As a result, the human haplotype blocks have remained largely intact and provide an unbroken thread that connects all people to a common past and to each other. Recent research indicates that about 65-85 percent of the human genome may be organized into haplotype blocks that are 10,000 bases or larger.

The exact pattern of SNP variants within a given haplotype block differs among individuals. Some SNP variants and haplotype patterns are found in some people in just a few populations. However, most populations share common SNP variants and haplotype patterns, most of which were inherited from the common ancestor population. Frequencies of these SNP variants and haplotype patterns may be similar or different among populations. For example, the gene for blood type is variable in all human populations, but some populations have higher frequencies of one blood type, such as O, while others have higher frequencies of another, such as AB. For this reason, the HapMap consortium needs to include samples from a few geographically separated populations to find the SNP variants that are common in any of the populations.

Charles Rotimi, Ph.D., leader of the Howard University group collecting the blood samples in Nigeria, said, “We need to be inclusive in the populations that we study to maximize the chance that all people will eventually benefit from this international research effort.”

Because of the block pattern of haplotypes, it will be possible to identify just a few SNP variants in each block to uniquely mark, or tag, that haplotype. As a result, researchers will need to study only about 300,000 to 600,000 tag SNPs, out of the 10,000,000 SNPs that exist, to efficiently identify the haplotypes in the human genome. It is the haplotype blocks, and the tag SNPs that identify them, that will form the HapMap.

Haplotypes and Health

To date, most of the known disease-causing genetic variations have been for relatively rare disorders, such as Huntington’s disease and cystic fibrosis. These diseases are caused by variants in single genes that tend to have a big impact on health, making the genetic contributions to the illnesses relatively easy to find using current methods that rely on gathering family information, or pedigrees.

Researchers face a much tougher challenge when it comes to uncovering the genetic contributors to more common diseases, such as Alzheimer’s disease, arthritis, cancer, diabetes, schizophrenia and stroke. These disorders are caused by many genetic variants that individually have a relatively weak contribution to the disorder, but together can increase the risk of illness. Environmental and other non-genetic factors also contribute to the disease process, making it even harder to find the genetic factors.

Researchers emphasize that the HapMap is not meant to minimize the role of environmental factors in disease development. “In fact, studying genetic factors may greatly increase the likelihood of our understanding the environmental contribution to illness, since these influences often interact,” said Thomas Hudson, M.D., leader of the HapMap group at McGill University in Canada.

Once the HapMap is constructed, researchers around the globe will use it to study the genetic risk factors underlying a wide range of diseases and conditions. For any given disease, researchers would use the HapMap tag SNPs to compare the haplotype patterns of a group of people known to have the disease to a group of people without the disease, a method known as an association study. If the association study finds a certain haplotype more often in the people with the disease, researchers would then zero in on that genomic region in their search for the specific genetic variant. The tag SNPs would serve as signposts indicating that a genetic variant involved in the disease may lie nearby.

“Even with the human sequence in hand, linking small changes in the genome to changes in health is tedious work,” said Huanming Yang, Ph.D., director of the Beijing Genomics Institute and coordinator of The China HapMap Consortium. “The HapMap project will create a powerful tool for linking differences in the genome to differences in health, including increased risk for common illnesses.”

Mapping an individual patient's haplotypes also may be used in the future to help customize medical treatment. Genetic variation has been shown to affect the response of patients to drugs, toxic substances and other environmental factors. Some already envision an era in which drug treatment is customized, based on the patient's haplotypes, to maximize the effectiveness of the drug while minimizing side effects.

In addition, the HapMap may eventually help pinpoint genetic variations that may contribute to good health, such as those protecting against infectious diseases or promoting longevity.

Technology and Cooperation

Carrying out such a complex project will depend on the application of robust technologies to analyze individual SNP variants. The technologies must be capable of high throughput, high quality, and low cost. Different groups will be using different technologies, providing the scientific community a chance to test which approaches work best. That experience is likely to speed the process of technology development, so that once the HapMap is available, the tools to use it will be much better developed.

In addition to its pioneering approach towards developing the HapMap and related technologies, the international consortium continues the strategy of pulling together a wide range of public and private partners from around the globe to both conduct and fund the research.

TSC chairman Arthur Holden said, "We are very positive about the chance to work collaboratively with the HapMap effort to support the informatic aspects of the program, as well as to ensure that the resulting HapMap will be useful in both disease and pharmacogenomic research."

###

International HapMap Project

Participants: Sample Collection

COUNTRY	RESEARCH GROUPS	FUNDING AGENCY
Nigeria & United States	Charles Rotimi and a team based at Howard University, in collaboration with colleagues at University College Hospital in Ibadan, Nigeria	U.S. National Institutes of Health
Japan	Ichiro Matsuda of Kumamoto University, in collaboration with colleagues at the Eubios Ethics Institute at University of Tsukuba	Japanese Ministry of Education, Culture, Sports, Science and Technology
China	Changqing Zeng of Beijing Genomics Institute, in collaboration with colleagues at Beijing Normal University	Chinese Ministry of Science and Technology
United States	Mark Leppert of University of Utah School of Medicine	U.S. National Institutes of Health

International HapMap Project

Participants: HapMap Construction

COUNTRY	RESEARCH GROUP LEADERS	FUNDING AGENCY	ROLE
Japan	Yusuke Nakamura, RIKEN/University of Tokyo	Japanese Ministry of Education, Culture, Sports, Science and Technology	Analyze 25% of genome
United Kingdom	David Bentley, The Wellcome Trust Sanger Institute	The Wellcome Trust	Analyze 24% of genome
Canada	Thomas Hudson, McGill University	Genome Canada Genome Quebec	Analyze10% of genome
China	Huanming Yang, Beijing Genomics Institute and The China HapMap Consortium	Chinese Ministry of Science and Technology Chinese Academy of Sciences Natural Science Foundation of China	Analyze10% of genome

COUNTRY	**RESEARCH GROUPS**	**FUNDING AGENCY**	**ROLE**
United States	Mark Chee, Illumina, Inc.	U.S. National Institutes of Health	Analyze 31% of genome
	David Altshuler, Whitehead Institute for Biomedical Research		
	Richard Gibbs, Baylor College of Medicine		
	Pui-Yan Kwok, University of California, San Francisco		
	Aravinda Chakravarti, Johns Hopkins School of Medicine		
Multinational	The SNP Consortium (TSC)	TSC	Data coordination center, analysis groups and more HapMap construction

NIH BACKGROUNDER

National Institutes of Health | National Human Genome Research Institute

EMBARGOED:
Hold for Release at 1 p.m. EST, Oct. 29, 2002

For additional information, contact:
Geoff Spencer, NHGRI
[redacted]

Background on Ethical and Sampling Issues Raised by the International HapMap Project

Members of the research consortium working on the International HapMap Project have taken steps to try to ensure that the map will be designed, developed and used in a manner that is sensitive to the ethical, legal and social concerns raised by this type of genomics research.

Since its inception in 1990, the Human Genome Project has paid special attention to the complex ethical, legal and social implications of this kind of research. In keeping with this practice, the National Human Genome Research Institute (NHGRI) established a group consisting of geneticists, social scientists and experts on the ethical and societal implications of genetic research to address a number of aspects of the project. These included how to design the most scientifically valid sampling strategy, how to engage the individuals and communities that will be asked to provide samples, how to describe the populations, and how to minimize the chance of misunderstanding or misuse of the results of future studies that will rely on the HapMap.

This group, which was co-chaired by David Valle, M.D., of Johns Hopkins University, Ellen Wright Clayton, M.D., J.D., of Vanderbilt University and Lynn Jorde, Ph.D., of the University of Utah, proposed a strategy for the HapMap project designed to meet both the need for high quality, scientific research and the need for the project to adhere to the highest ethical standards to protect participants.

Sampling Strategy

The HapMap will be a new tool to speed the discovery of genetic contributions to diseases. The HapMap will describe the common patterns of human genetic variation and will be used in future studies that compare the patterns of genetic variation (haplotypes) in people with a specific disease to patterns in people who do not have the disease. By identifying regions in those genomes showing differences in the haplotype patterns, researchers can focus their studies on genomic regions to more efficiently find the particular genetic variants that contribute to the disease.

The HapMap project will begin with sample collection. Research groups will collect blood samples from a total of 200 to 400 people from four large, geographically distinct populations. These populations are: the Yorubas in Nigeria; the Japanese; the Han Chinese; and U.S. residents with ancestry from northern and western Europe. Except for the U.S. samples, all of the samples will be newly collected for this project. The U.S. samples, which already exist, will be used only after the donors provide a new and specific consent for the HapMap project.

These four populations were selected to include people with ancestry from widely separate geographic regions. Researchers have found that most human populations share the common haplotype patterns. Research already suggests that the overall organization of genetic variation is similar in all four populations, but that there will be enough differences in haplotype frequencies to justify genome-wide studies of samples from these populations.

Because populations have similar haplotype patterns, the project will not have to examine all of the world's thousands of populations to make the HapMap useful for studies relating genetic variation to disease in any population. Additional research is underway to confirm whether the common haplotypes in other populations really will be found in the four populations being studied for the HapMap. If needed, more populations could be added to the HapMap to ensure that the map is broadly useful.

The four populations chosen to develop the HapMap initially are neither typical nor well-defined. None of the populations should be considered representative of all populations on the same continent. For example, the Yoruba samples studied for the HapMap are not representative of all Africans, or even of all West Africans.

The purpose of the HapMap and its sampling strategy make this project very different from the Human Genome Diversity Project (HGDP), an anthropologically oriented effort proposed more than a decade ago that was designed to learn about human population history and the biological relationships among human populations. The HGDP would have studied genetic variation "to see if, for example, the Irish are more closely related to the Spaniards or to the Swedes," according to the project's material. A number of groups representing indigenous peoples were concerned that the project would exploit vulnerable individuals and populations. They also objected to the HGDP's potential intrusion into cultural beliefs about population origins. Ultimately, and in large measure because of the criticisms, the HGDP was never carried out.

Unlike the HGDP, the HapMap's goal is biomedical: to create a resource that can be used in many future studies of health and disease. In addition, unlike the HGDP, which would have studied primarily small, isolated populations, the International HapMap Project will study only large, less vulnerable populations.

Informed Consent and Privacy

Obtaining meaningful informed consent from people who are donating DNA samples for the HapMap project raises complex challenges. The international researchers collecting the samples are devoting considerable effort to figuring out how best to translate complex information about genetics and haplotypes into language that ordinary people can understand. Researchers must be sensitive to cultural norms surrounding decision-making within families and communities, and to beliefs about the relationships among genetics, kinship and group identity.

All donors will be asked to give consent for their samples to be used not just for the HapMap itself, but also in many types of future genetic variation studies. Such studies may examine how genes are regulated, the biology of DNA, how new variations arise and the genetic history of human groups. Researchers will explain to donors that the benefits of the HapMap and of other genetic variation research may not become apparent for some time and that the donors themselves may not directly benefit from participating.

Before obtaining consent from the individual donors, researchers will initiate a process of community engagement. People in these communities will provide advice about the informed consent process, as well as how samples from their community will be collected, described and used. A community advisory group will be established for each sampled community to serve as liaison between the people in that community and the repository where the samples will be stored. These groups will monitor future uses of the samples to make certain that these future uses are consistent with the informed consent form.

The blood samples used to make the HapMap will be collected without any medical or personally identifying information about the donors. In a further step to ensure the complete anonymity of the donors, more samples will be collected than will actually be used, which means that no one, not even the donors themselves, will ever know for sure whose samples were used to develop the HapMap.

Genetic Discrimination and Determinism

Because researchers will not collect medical or personally identifying information, there is virtually no risk that the HapMap itself will lead to discrimination against any of the individual sample donors. However, in future studies, some genetic variants will be identified that promote wellness and protect against disease, while other variants will be identified that increase the risk for particular diseases. When researchers use the HapMap and find that a disease is associated with a genetic variant that is common in a particular population, some people may mistakenly generalize that all individuals in that population have increased risk for the disease or that the population as a whole is somehow genetically inferior.

Another problem with the interpretation of genetic variation is assuming that "genetic" means "unchangeable," and that because someone has a particular genetic variant they are "doomed" to get the disease. These incorrect assumptions are called genetic determinism. Genetic determinism overlooks the strong contributions that environmental factors make to diseases and that there may be ways to reduce the risk of getting those diseases. So, even though people may have genetic variants contributing to their risk of a disease, many of them will never get the disease.

Genetic discrimination and genetic determinism are both potential problems that can arise from any association study in which researchers relate genetic variation to disease risk. These potential problems are not unique to studies that will use the HapMap. Nevertheless, the HapMap consortium intends to make concerted efforts to reduce the risk of such problems. Among the steps the group plans to take are:

- Educating the public and researchers about what the results of genetic studies in general, and association studies in particular, mean and do not mean — with the focus on differences in genetic risk among individuals within a population, not among populations. An association study compares a haplotype pattern in individuals with a disease to individuals who do not have the disease to find the genes directly associated with the condition.

- Educating researchers to design their studies and describe their results carefully. For example, researchers should describe the studied population accurately; they should also report how much of the risk for a disease can be attributed to genetic variants and how such variants interact with environmental factors. Where these matters are not well understood, uncertainty should be acknowledged.

####

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
NATIONAL INSTITUTES OF HEALTH

Strategy Meeting for the International HapMap Project

Washington Marriott
Washington, DC

October 27-29, 2002

① Japan
② UK
③ Canada
④ China

⑤ Chee
⑥ Alt
⑦ Gibbs
⑧ Kwok
⑨ Chak (Michael Zwick)

AGENDA

Alpha	Size
⑤ Hudson	
⑦ Nakamura	①
① Altshuler	
③ Chee	
④ Gibbs	
~~② Chakravarti~~	
⑥ Kwok	
⑧ Young (1b) Bentley	②
~~Canada~~	
~~China~~	

Sunday, October 27

6:30 a.m. – 7:00 a.m.	Buffet Dinner
7:00 a.m. – 7:15 a.m.	**Welcome:** Francis Collins, Satoshi Tanaka, Karen Kennedy, Hua Han, Arthur Holden
7:15 a.m. – 9:00 a.m.	**Introduction by the Nine Genotyping and Analysis Groups:** *Members, technologies, prior projects (10 minutes each)*

Monday, October 28

8:00 a.m. – 8:30 a.m.	Continental Breakfast
8:30 a.m. – 8:40 a.m.	**Goals for the Meeting:** Francis Collins
8:40 a.m. – 9:30 a.m.	**Samples - Timeframe and Numbers:** Jean McEwen
9:30 a.m. – 10:15 a.m.	**Chromosome Allocations:** Yusuke Nakamura, discussion leader
10:15 a.m. – 10:30 a.m.	Break
10:30 a.m. – 11:00 a.m.	**SNP Distributions and dbSNP:** Stephen Sherry
11:00 a.m. – 1:00 p.m.	**Scientific Strategy:** David Altshuler, discussion leader *Number of samples, primate samples, SNP density and numbers, SNP allele frequencies*
1:00 p.m. – 2:00 p.m.	Lunch
2:00 p.m. – 3:30 p.m.	**SNP Discovery:** David Bentley, discussion leader *Cost models of random vs. targeted SNP discovery, SNP sources*
3:30 p.m. – 4:30 p.m.	**Quality Measures:** Mark Chee, discussion leader
4:30 p.m. – 4:45 p.m.	Break
4:45 p.m. – 5:30 p.m.	**Data Coordination Center:** Lincoln Stein

Monday, October 28 (continued)

5:30 p.m. – 6:00 p.m.	**Intellectual Property:** Arthur Holden *Data flow and IP protection of SNP, genotype, and haplotype information for complete and rapid release to the public domain*
6:00 p.m. – 7:30 p.m.	Dinner
7:30 p.m. – 9:00 p.m.	**Analysis:** Aravinda Chakravarti, discussion leader *Organization of how analysis will be done* Description of the plan

Tuesday, October 29

7:30 a.m. – 8:45 a.m.	Breakfast for the Funding Agency Representatives
8:30 a.m. – 9:00 a.m.	Continental Breakfast for Others
9:00 a.m. – 11:45 a.m.	**Issues To Be Addressed and Organization of Working Groups:** Francis Collins, discussion leader
11:45 a.m. – 12 noon	**Group Photograph**
12 noon – 1:00 p.m.	**Press Briefing** (Funders and PIs of the sample collection, genotyping, and analysis groups)

David Bentley DRAFT 10/24/02

THE HAPMAP PROJECT – SNP DISCOVERY

1. Factors affecting SNP requirement

(a) Definition of the problem

We need enough SNPs to detect LD and build the map of common haplotypes in most or all of the genome. Most studies involving reasonable numbers of DNA samples have been SNP-limited, and more SNPs are needed. LD varies considerably with physical distance. Some regions of the genome therefore need a much higher density of SNPs than others. LD also varies in different populations. More SNPs are likely to be needed for some populations than others in some regions. Because of the regional and population variability, the overall average range of LD in the genome is not enough to establish the criteria we will use to declare each part of the genome done. We therefore need to make some initial assumptions based on the available data from existing studies, and to adjust our projections as we go along.

(b) Existing studies (please suggest expansions/amendments)

Jeffreys et al.: Using 1 SNP (m.a.f. > 0.15) on average every 1.2kb, blocks of LD were obtained spanning 94% of the 216kb region analysed and were separated by short intervals (1-2kb) where observed present-day recombination occurs.
Johnson et al.: Using 1 SNP (m.a.f. > 0.1) on average approx every 2kb within 9 genes, similar characterization of common haplotypes observed.
Gabriel et al.: Using 1 SNP (m.a.f. > 0.1) on average every 7.8kb, average LD block size of 22kb (Caucasians) and 11 kb (Africans) was identified in 51 regions of ~0.25Mb each.
Dawson et al.: Using 1 SNP (m.a.f. > 0.2) on average every 20kb, LD blocks were detected covering 20% of the chromosome. LD inversely correlated with recombination frequency.
All studies: Within an LD block, typically up to 5 common haplotypes account for most of the variation. These can be captured using a subset of haplotype tag SNPs (htSNPs).

(c) Initial assumptions (we can discuss and change these)

That all LD blocks (criteria to be defined) of at least 10kb will be detected in this study.
That blocks have up to 5 common haplotypes (>5%)(although further work on practical block definition is required)
That 5 SNPs are needed to analyse each 10kb region for LD and characterize blocks.
That 50% of untested candidate SNPs in the map are sufficiently polymorphic for this study and provide a robust genotyping assay (based on Sanger and Whitehead experience using Sequenom or Illumina platforms).
That ~10% of the genome may be intractable to the present analysis (e.g. 5% is in recent duplications).

2. SNPs required

We need to discuss what SNPs are needed to *finish* the project, and what SNPs are needed to *start* the project. Clearly to start now, we need to explore if there are enough SNPs for the optimal start point (which will emerge from the strategy discussion) or whether we are SNP-limited to a start point that is suboptimal strategically.

(a) What do we need to finish the project?

To complete the HapMap to the level defined in the initial assumptions (above), each 10kb window might need up to 10 candidate SNPs in the map, of which 5 would be sufficiently polymorphic provide data to define the extent of the block and the common haplotypes within each block (confirmation of this may be required). Fewer SNPs may be required in some sections of the genome to reach the specified endpoint, if for example LD is more extensive than the example of Jeffries in the MHC.

(b) What do we have to start the project?

The published map has ~2.2M SNPs with unique map position (average 1 SNP per 1.3kb), comprising 1.3M TSC SNPs (average 1 SNP per 2.2kb, randomly distributed) and 1.5M SNPs from overlaps (lumpy distribution - includes some duplication with TSC SNPs). Re-analysis of the TSC data by Jim Mullikin has resulted in detection of up to 0.4M more 'TSC' SNPs (subject to checking and validation). Below is an excerpt of a table (analysis by Steve Sherry and Sarah Hunt) showing the distribution of these SNPs in 281,000 adjacent 10kb windows (covering 2.8Gb of the genome). It shows that 76% of the genome has 2 SNPs in 10kb, whereas only 29% of the genome has 10 SNPs in 10kb. *(Note that this analysis illustrates how average SNP density can be misleading; for example the average density of the random TSC SNPs in 1 SNP per 2.2kb across the genome, but in this set only 64% of the genome has a local SNP density of 1 SNP / 5kb or more)*[a].

SNPs per 10kb window	TSC only	OVRLP only	ALL SNPs
2 or more	64% [a]	38%	76%
5 or more	24%	28%	53%
10 or more	10%	17%	29%

On this basis, we can choose to start now by selecting for example ~500,000 evenly spaced SNPs, of which 250,000 (1 SNP per 10kb on average) will provide data. (Sanger can cite practical experience of this exercise). Progressively less of the genome has enough SNPs for starting (or follow up) at higher densities, and only 29% of the genome currently has enough SNPs for optimal analysis according to the initial assumption of 10 SNPs per 10kb.

3. How do we obtain more SNPs?

(a) Strategy

New SNPs may be detected by random or targeted approaches, or a combination of both. The choice of strategy, or the point to switch from a random to a targeted approach, will be governed by the cost-benefit ratio, which alters as the project progresses.

Random shotgun sequencing (of whole genomes or individual chromosomes) has the advantage that it follows established large-scale protocols for data generation and analysis (as used by TSC). The process is cheap, involving universal primed sequencing on cloned DNA templates, SNP discovery by alignment of haploid sequence data. Without targeting, SNPs accumulate all over the genome, both filling in gaps where there are no SNPs and providing much more choice for the HapMap project in all other areas.

Targeted SNP discovery is more expensive (on a per SNP basis) as it requires more prior investment (in custom PCRs or pre-mapping clones), and in the case of PCR requires analysis of diploid sequence trace data, where many SNPs are in heterozygous form. The effort is targeted to exactly where the SNP is needed but does not contribute any additional SNPs elsewhere. The present model is based on doing a minimum investment to identify at least one SNP per assay.

Additional features of random shotgun versus targeted SNP discovery:

Population specificity: More work needs to be done to define the need for SNPs to determine the HapMap in different populations. SNP discovery may be directed towards different ethnic groups to balance the SNPs available in the map. A random shotgun strategy would support the greater choice of SNPs in multiple populations, and double-hit SNPs (see above) in different populations would identify SNPs which are polymorphic in multiple populations before assay development.

Long-term considerations: Refinement or extension of the study may be beyond the scope of the HapMap project as designed, but is a real possibility in the longer term. The availability of many more SNPs from additional random shotgun would add value to the SNP map and underpin long term studies.

(b) Cost

A cost modeling exercise carried out to examine this by members of Sanger, Whitehead, Seattle, GSK and Genaissance, provided the following results. For this discussion, we assume the endpoint for SNP discovery is to obtain 90% of the genome in 10kb windows containing 10 SNPs or more. It does *not* take account of prior LD mapping (see below). Alternative endpoints, with or without LD mapping, could be built in (for discussion).

The upper part of the table describes the anticipated progression due to shotgun sequencing alone and is based on a recent chromosome 20 sequencing study at Sanger.

The middle part of the table adds in a PCR-based targeted sequencing component at each stage to provide the same endpoint by different combinations of the two approaches. The table shows a reasonably flat minimum point between 2x and 5x shotgun. The lower part of the table provides a combined cost, new SNP total and average cost per SNP added to the map. Maximum value is obtained with 4x additional random sequencing and a limited amount of additional targeting.

New shotgun sequence:	0	1x	2x	3x	4x	5x	6x
% windows done:	29	36	58	73	82	88	91
Cost @ $1.5 per read ($M)	0	6	12	18	24	30	36
Minimum new SNPs (M):	0	1.1	2.0	2.7	3.3	3.8	4.3
New targeted PCRs (M):	1.1	0.64	0.34	0.16	0.04	0	0
Cost @$50 per region ($M):	55	32	17	8	2	0	0
Minimum new SNPs:	1.1	0.64	0.34	0.16	0.04	0	0
Combined cost ($M):	55	38	29	26	26	30	36
Total minimum new SNPs:	1.1	1.74	2.34	2.86	3.34	3.8	4.3
Average cost per SNP ($):	50	22	12.4	9.1	7.8	7.9	8.4

The projected cost may be reduced in a number of ways (for further discussion):

Different start point:	More SNPs available free from other sources
Lower operational cost:	Cost savings in sequencing etc
Different end point:	e.g. less than 90% of the genome at 10 SNPs / 10kb e.g. 90% of the genome at less than 10 SNPs / 10kb
Use LD hierarchy:	avoid sequencing for more SNPs in areas where LD endpoint is reached. This would benefit from further discussion, in conjunction with defining when such an endpoint is reached.

Other factors which may reduce the cost of the project as a whole include SNPs hit twice (which accrue as a major product of the random shotgun sequencing option); SNPs with known allele frequency; SNPs with known working assay. The success rate for assay conversion of SNPs in any of these categories is likely to be much higher.

TSC - what to focus on?

SNP discovery?
More samples?

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
NATIONAL INSTITUTES OF HEALTH

Strategy Meeting for the International HapMap Project

Washington Marriott
Washington, DC

October 27-29, 2002

PARTICIPANT LIST

David Altshuler, M.D., Ph.D.
Whitehead Institute/MIT Center for Genome Research
[redacted]

John W. Belmont, M.D., Ph.D.
Associate Professor
Baylor College of Medicine
[redacted]

David R. Bentley, M.Phil.
Head of Human Genetics
Sanger Institute
Wellcome Trust Genome Campus
Hinxton
[redacted]

Vence L. Bonham, Jr., J.D.
Senior Consultant to the Director on Health Disparities
Office of the Director
[redacted]

Lisa D. Brooks, Ph.D.
Program Director
Genetic Variation Program
National Human Genome Research Institute
National Institutes of Health
[redacted]

Aravinda Chakravarti, Ph.D.
Henry J. Knott Professor and Director
Institute of Genetic Medicine
Johns Hopkins University School of Medicine
[redacted]

Mark Chee, Ph.D.
Research Fellow
[redacted]

Francis S. Collins, M.D., Ph.D.
Director
National Human Genome Research Institute
National Institutes of Health
[redacted]

David J. Cutler, Ph.D.
Research Associate
Institute of Genetic Medicine
Johns Hopkins University School of Medicine
[redacted]

Mark J. Daly
Whitehead Fellow
Whitehead Institute/MIT Center for Genome
Research
[redacted]

Panos Deloukas, Ph.D.
Senior Group Leader
Sanger Institute
Wellcome Trust Genome Campus
[redacted]

Michael L. Feolo, M.S.
Staff Scientist
National Center for Biotechnology Information
National Library of Medicine
[redacted]

Lance Fors, Ph.D.
Chief Executive Officer
[redacted]

Stacey Gabriel, Ph.D.
Scientific Director, Human Haplotype Map
Medical and Population Genetics Program
[redacted]

Richard Gibbs, Ph.D.
Director
Human Genome Sequencing Center
[redacted]

Alan E. Guttmacher, M.D.
Deputy Director
National Human Genome Research Institute
[redacted]

Mark Guyer, Ph.D.
Director
Division of Extramural Research
National Human Genome Research Institute
[redacted]

Hua Han
Chinese Academy of Science
[redacted]

Arthur Holden, M.B.A.
Chairman
[redacted]

Thomas J. Hudson, M.D.
Director
Montreal Genome Centre
The Research Institute
McGill University Health Centre
[redacted]

Bronya Keats, Ph.D.
Professor and Chair
Department of Genetics
[redacted] Sciences

Karen Kennedy, Ph.D.
Coordination Manager
The Wellcome Trust
[redacted]

Semyon Kruglyak, Ph.D.
Senior Scientist
Informatics
[redacted]

Pui-Yan Kwok, M.D., Ph.D.
Henry Bachrach Distinguished Professor
and Investigator
Cardiovascular Research Institute
[redacted]

Eric H. Lai, Ph.D., M. Phil., M.A.
Vice President, SNP Capabilities
GlaxoSmithKline
[redacted]

Ichiro Matsuda, M.D., Ph.D.
Professor
Ezu Institution for Developmental Disabilities
[redacted]

Jean E. McEwen, Ph.D., J.D.
Program Director
Ethical, Legal, and Social Implications Program
[redacted] Institute

Raymond Miller, Ph.D.
Research Assistant Professor
[redacted] icine

Alexandre Montpetit, Ph.D.
Postdoctoral Fellow
Montreal Genome Centre
The Research Institute
McGill University Health Centre
[redacted]

Koichi Morimoto
Science Counselor
[redacted] W

Richard M. Myers, Ph.D.
Professor and Chair
Department of Genetics
[redacted] dicine

Yusuke Nakamura, M.D., Ph.D.
Director
Human Genome Center
Institute of Medical Science
[redacted]

Debbie Nickerson, Ph.D.
Associate Professor of Genome Sciences
University of Washington
[redacted]

Arnold Oliphant, Ph.D.
Vice President
Scientific Operations
[redacted]

Allison Peck, M.S.
Scientific Program Analyst
National Human Genome Research Institute
[redacted]

Jane L. Peterson, Ph.D.
Associate Director of Extramural Research
National Human Genome Research Institute
[redacted]

Zhaoxia Ren, Ph.D.
National Institute on Alcohol Abuse
and Alcoholism
[redacted]

John Rice, Ph.D.
Professor of Mathematics in Psychiatry
[redacted]

Stephen T. Sherry, Ph.D.
Staff Scientist
National Center for Biotechnology Information
[redacted]

Koki Sorimachi
Chief
Planning Section
Yokohama Research Promotion Division
[redacted]

Lincoln Stein, Ph.D.
Associate Professor
[redacted]

Wenching Sung, Ph.D.
Department of Science and Research
Beijing Genomics Institute
[redacted]

Danilo A. Tagle, Ph.D.
National Institute of Neurological Disorders
and Stroke
National Institutes of Health
[redacted]

Satoshi Tanaka, M.S.
Director, Life Science Division
Promotion Bureau
Ministry of Education, Culture, Sports, Science,
and Technology
[redacted]

Toshihiro Tanaka, M.D., Ph.D.
Laboratory Head
Laboratory for Cardiovascular Diseases
SNP Research Center
RIKEN
[redacted]

Tatsuhiko Tsunoda, Ph.D.
Laboratory Head
[redacted]

Andrei Verner, M.S.
Manager, Genotyping Platform
Montreal Genome Centre
The Research Institute
McGill University Health Centre
[redacted]

Wendy Wang, Ph.D.
National Cancer Institute
[redacted]

Alan R. Williamson, Ph.D.
Maywood
[redacted]

Tom Willis, Ph.D.
Chief Executive Officer and President
Par*Allele* BioScience, Inc.
[redacted]

Huanming Yang, Ph.D.
Director
Beijing Genomics Institute
B5 Datun Road
[redacted]

Changqing Zeng, Ph.D.
Head
Department of Science and Research
Beijing Genomics Institute
[redacted]

Michael E. Zwick, Ph.D.
John Wasmuth Postdoctoral Fellow
Institute of Genetic Medicine
Johns Hopkins University School of Medicine
[redacted]

David Altshuler DRAFT 10/22/02

HAPMAP – INITIAL SCIENTIFIC STRATEGY

Below are listed a set of issues related to scientific design and analysis of the Hap Map project. In order to launch the project in the near term, we will have to reach agreement on many of these topics. In some cases (for example, deciding when we will decide we are done with a region), we have a bit more time before we need to decide, but will want to map out a process towards reaching agreement.

Please consider this outline an initial attempt to frame the issues and stimulate discussion. I ask that each group assign a single individual to respond on behalf of the group (to limit the email traffic), responding with comments and proposals related to what is described below, and adding areas that aren't covered but perhaps should be. I will synthesize/summarize the responses and send out an updated version, which will allow us one more round of emails in preparation for the discussion in Washington next week.

The outline below is based on past discussions with many of you, and was jointly prepared by our group and the Hopkins group. The format is a brief statement summarizing a proposed strategy or goal, followed by specific questions and issues for comment and discussion. There are six topics covered:

1) **Overall strategy**
2) **Criteria for SNP selection**
3) **Sample selection**
4) **Quality control and quality assessment**
5) **Criteria for deciding we have completed a region**
6) **Data release**

1) **Overall strategy** — nearly all discussions have focused on a hierarchical approach, with three steps: an initial round of genotyping at a fixed spacing, analysis of LD/haplotypes, and additional rounds of genotyping to increase density where finishing rules are not yet reached. Stopping rules are described in #5, below. This strategy is based on the assessment that it is currently impractical to achieve more complete genotyping everywhere (for example, typing all available SNPs, or by complete resequencing) with available technology and resources. Questions and issues:

   a. Does everyone agree that this is the high-level strategic plan?

      i. We assume that the group will jointly determine the goals and strategies, and that each group will apply them uniformly across the genome. It should not be the case that the haplotype map does not have a predictable set of characteristics, or that the HapMap constructed by one of us would not be equivalent to that prepared by a different group at a different chromosomal location

   b. What is the correct density for the initial spacing of markers?

      i. The spacing for an initial round of genotyping should be based on current estimates of LD in the genome. The goal should be to capture large blocks first (aiming for an initial screen with one SNP every 5-20kb) and then analyze the initial data. We will need to discuss the optimal spacing.
      ii. Should we try to set spacing based on genetic distance (correcting for cM/Mb) rather than physical distance? This would be unlikely to have a major impact, but all available data suggest that cM/Mb is correlated with the extent of LD. It would also require each group to integrate the map data into their SNP selection rules
      iii. We assume that our goal is to cover the entire genome without bias according to current estimates of where the most "interesting" places are. That is, we will not bias coverage towards genes or other annotated regions.

   c. It would seem most valuable to the community if we completed and rapidly released this initial screen across the entire genome before increasing density in any one region. This is because:

      i. We would immediately provide a useful tool for researchers working in any region
      ii. We haven't yet defined stopping rules, and this would give us a bit of time to establish them.

2) **Criteria for SNP selection** — as above, we will agree to a uniform set of criteria for selecting SNPs for genotyping. The discussion of how to achieve complete genome-wide coverage of whatever number of SNPs are needed falls under David Bentley's SNP Discovery topic.

   a. The vast majority of SNPs on the public map are discovered by low-pass resequencing of genomic regions by either the HGP or the TSC. Some of these have public frequency data provided by TSC AF project or other large-scale genotyping projects in academia or industry. The cost of developing assays that are not useful (either because they fail, or are too rare for our purposes) will be a major cost driver. On the other hand, we want a uniform ascertainment procedure for SNPs, because otherwise the map produced will not have predictable properties (vis-à-vis the true underlying haplotype structure). A possible strategy for picking SNPs might include:

      i. Prioritizing SNPs discovered by two or more independent efforts to avoid spending too much money genotyping SNPs that will be rare or monomorphic. These might include both alleles being seen more than once in:
         1. TSC
         2. HGP
         3. Allele frequency projects
         4. directed resequencing efforts

      ii. If SNPs have known allele frequency characteristics, choosing a single set of rules for their inclusion
         1. frequency >5% in any single population studied?
            a. Do we limit this to major continental groups?

      iii. Prioritizing SNPs for which a successful genotyping assay has previously been developed
         1. although this may not be predictive for any of our particular platforms, it at least means that someone could amplify and confirm the alleles

      iv. If none of the criteria are met (and a SNP is needed in the location), using an unconfirmed SNP from the public map

      v. Avoiding SNP sets that have been shown to have lower rates of validation or high frequency

      vi. Not everyone agrees we should prioritize SNPs – we will need to discuss the empirical data each group has already collected to help decide this issue.

   b. Do we want to have as a priority to obtain and genotype on the map (regardless of the distribution of LD) all putative functional SNPs — those that alter protein-coding regions or fall on regions with strong interspecies conservation.

        i. This requires a comprehensive source of such SNPs based on targeted resequencing
            1. should discuss ongoing projects in Japan, Sanger, etc.

c. An important issue is the prefilters that each group may want/need to apply to achieve success with their chosen technology platform. On the one hand, this is obviously a good and important part of optimal laboratory management. On the other hand, the true cost of each method includes the fraction of SNPs it can genotype, because the project is going to have to pay to discover additional SNPs if some methods are highly restrictive in their SNP selection.
    i. We need to compare and collect empirical data about the fraction and properties of SNPs that can be genotyped.
    ii. We cannot reject categories of SNPs lightly (such as those near repeats), since the total project cost is based on completion of the map, not the fraction of SNPs successfully genotyped.

3) **Sample selection** — much work has already gone into the collection of appropriately consented samples for HapMap. Thus, we have limited flexibility (at least to start). Moreover, our understanding is that only the CEPH will be complete and consented in 2002.

a. Do we all agree to start genotyping in available samples (CEPH), repeating all markers (except, if logistically feasible, those that are pure technical failures) in the additional samples as they become available?

i. Are there any options whereby we could start with all the populations sooner? That would certainly be optimal from a logistical and political point of view.

b. How many people do we genotype in each population? The project has discussed two scales: 192 samples (total) and 384. This decision will be driven by economic arguments to a significant degree, since I think we all agree that more samples is better, if achievable.

i. How many samples from each group

1. do we balance evenly across all population groups?

ii. Trios or unrelated individuals.

1. trios provide internal genotyping quality checks

2. trios provide phase information

a. whether information is needed depends on the goals for the project — to define haplotypes in blocks (regions with low rates of historical recombination), or across long regions with little LD (which, to our minds, has little utility or biological meaning).

3. trios are less efficient. Thus, we should only use them to the extent that they provide valuable information.

c. How will we evaluate comprehensiveness of a map produced using solely samples from three or four population samples (one European in origin, one West African, and another East Asian)?

i. Discuss ongoing projects to more broadly define diversity, haplotype and LD properties across the globe.

d. Do we want to reserve some samples (even if we can afford them) for the results of such studies, or assume that any additional genotyping that might be required will be paid for by another mechanism?

4) **Genotyping quality control and ongoing assessment** — it is obviously critical that the map have uniform quality and performance. Given our diversity of technology platforms, this requires ongoing efforts to demonstrate quality with both internal and external validation.

   a. It will be critical to demonstrate the characteristics of SNPs that can be successfully genotyped by each method, since any platform-specific filtering procedures will bias the SNPs on the map (for one region relative to another). For this reason, we propose that the initial phase of SNP genotyping be performed under a uniform set of SNP selection rules, with each group attempting to genotype these SNPs in their laboratories. This will allow us to inform users as to the effect of platform and group on the properties of the map.
      i. An initial set of SNPs will be selected by a uniform procedure and served to each group by a Data Coordinating Center (DCC). These SNPs will be genotyped in each laboratory and results deposited in the DCC.
   b. Internal quality control should be provided either by use of trios (see above, which allows estimates of error rates) or by duplicate genotyping of a small fraction (5%?) of individual samples.
   c. External quality control should be evaluated by having the DCC routinely select a modest collection of SNP genotypes deposited by each group (1%?) and having them duplicated by at least two other groups using different platforms. These quality control activities should continue throughout the project to evaluate quality in an ongoing manner (since methods and personnel will certainly be in flux throughout the project.
   d. These QC/QA activities should be entirely public, with data from each summarize and released by the DCC.
      i. All categories of data (not designed, failed assay, monorphic, genotypes) should be deposited

5) **Deciding when we're done with a region** — this is clearly the most complex and least well defined aspect of the project. Our goal is fairly clear: to develop a resource of haplotype-based association studies that provides as much power and completeness as is practical given the resources and methods available. A number of concrete implementations can be discussed:

a. Defining haplotype "blocks" as regions over which there is little history of recombination among the common alleles represented in the population. Such regions have the property that measures of LD are fairly constant with distance, and that the diversity of common haplotypes is low. If we select this as our goal, we will want to define and uniformly apply methods that
    i. Are strongly predictive that a region defined as "block" behaves as such if a greater density of genotyping (up to complete resequencing) is acheived
    ii. Have sufficient number of markers that haplotype diversity is largely captured
    iii. support the selection of comprehensive "tag" SNPs that have a supra-threshold correlation coefficient (r^2) to a fraction of all SNPs in the region
b. Our recent paper (Gabriel et al) proposed one such set of methods, although much more work would be needed. Many other groups have developed independent or updated versions of these methods. We need to agree on a process to evaluate and compare methods and to decide which ones will be applied to guide the construction of the map.
    i. What other proposed end-points should we discuss?
c. What will be our gold standard to evaluate the performance of whatever metrics are selected
    i. Ideally would have regions that are oversampled (perhaps by complete resequencing) to define performance across all variants in a region

6) **Data release** — we presume that our collective data release policy will be to release all raw data in as rapid as is practical and result in data that is freely available and unencumbered by any intellectual property restrictions. As it is not clear if the details of this policy fall under the current document's scope, we will not provide any detailed proposals.

Collins, Francis (NIH/NHGRI)

From: [redacted]
Sent: [redacted]
To: [redacted]
Cc: [redacted]
Subject: HapMap: Allocation of Chromosome Regions

CH. Allocation.xls ATT60967.txt

Dear Members of the chromosomal allocation group,

Please find the updated information of the chromosomal allocations that I have heard from all of the members.

As you can find in the attached sheet, there are some conflicts. A total proportion of chromosomes proposed from the US groups is now 37.1%. I have heard from Francis that US contributes to 30% of the chromosomes and would like to know the proportion assigned to each group prior to the meeting. (Lisa or Francis, could you let me know?)

Chromosomes that were proposed by two groups are
Chromosome 6 ; UK, BCM
Chromosome 12; Illumina, BCM
Chromosome X; Illumina, UCSF/WashU
Chromosome 8p; MIT, China

Chromosome 4; none

We need to discuss to solve this issue in the meeting and hope you to compromise for establishment of the good international partnership.

Yusuke Nakamura,M.D.,Ph.D.

Yusuke Nakamura,M.D.,Ph.D.

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X		
	###	###	###	###	###	###	###	###	###	###	###	###	###	###	###	###	###	###	###	###	###	###	###		
Illumina				■								▨											▨	15.2%	37.1%
UCSF/WashU				■																			▨	5.0%	
MIT				■																				7.1%	
BCM				■		▨						▨												9.8%	
Canada				■																				7.7%	
UK				■		▨																		24.0%	
China				■																				8.2%	8p
Japan				■																				25.1%	
				■																				102.1%	

■ no one selected

▨ selected by two groups

□ second possibility

Collins, Francis (NIH/NHGRI)

From: [redacted]
Sent: [redacted]
To: [redacted]
Cc: [redacted]

Subject: Re: HapMap: Allocation of Chromosome Regions

Yusuke: Genome Canada provided funding for 10% of the genome/150,000 SNPs. I would be pleased if my allocation could increase to reflect that goal, by adding chromosome 4p (with the caveat suggested below).

I remain flexible as to the allocation. I think that the issue of expected coverage and number of DNAs (200 vs 400) must be clarified early in the process. I think that there has been incredible support from many countries to generate the Haplotype Map. Expectations will be high for a quality product that can be used in all populations.

David Altshuler raised important questions regarding sample sixes in his text sent earlier this week. As I am on the road this week, I do not have the "samples" report that was produced in 2001 by the HapMap working group. I ask Lisa Brooks if she can send the report and appendices of sample size simulations done my Cardon, Abecassis, Daly and others. My opinion at the time (which I still believe) is that we would generate a better HapMap for ~ 200 chromosomes per population (instead of ~120 as would be done with a sample of 192 DNAs. This is particularly more important for the AA population which has more haplotypes per block. The doubling of DNAs will probably not increase the project by more than 30%, since oligo costs etc will already be incurred.

I am not sure what impact this would have on various members.

Regarding David Bentley's report on the need for more SNPs, I hope that a solution can be found soon. The Montreal Genome Centre could produce random SNPs, but it would quite inefficient compared to having them done by the big sequencing centres. I certainly do plan to do targetted SNP discovery in Montreal at a late phase of the project to close gaps on all regions assigned to us.

Best regards to all,

Tom

>Dear Members of the chromosomal allocation group,
>
> Please find the updated information of the chromosomal
>allocations that I have heard from all of the members.
>
> As you can find in the attached sheet, there are some
>conflicts. A total proportion of chromosomes proposed from the US
>groups is now 37.1%. I have heard from Francis that US contributes
>to 30% of the chromosomes and would like to know the proportion

>assigned to each group prior to the meeting. (Lisa or Francis,
>could you let me know?)
>
>Chromosomes that were proposed by two groups are
>Chromosome 6 ; UK, BCM
>Chromosome 12; Illumina, BCM
>Chromosome X; Illumina, UCSF/WashU
>Chromosome 8p; MIT, China
>
>Chromosome 4; none
>
>We need to discuss to solve this issue in the meeting and hope you
>to compromise for establishment of the good international
>partnership.
>
>Yusuke Nakamura,M.D.,Ph.D.
>
>
>
>
>
>
>Yusuke Nakamura,M.D.,Ph.D.
>Attachment converted: Macintosh HD:CH. Allocation.xls 1 (XLS8/XCEL) (00033719)
>--

[redacted]

**!

Please note new e-mail: [redacted]

Thomas J. Hudson

[redacted]

[redacted]

[redacted]

**!

Collins, Francis (NIH/NHGRI)

From: [redacted]
Sent:
To:
Cc:

Subject: Re: HapMap: Allocation of Chromosome Regions

I will reiterate my opinion that we might not want to establish "squatters rights" over the entire genome before any work has been done. Over time it will become clear how well each group can establish the estimated throughput and cost model for the project, and it may be that some shifts need to take place in responsibility over time. Moreover, as Yusuke points out, the fraction of money allocated to different groups - and the fraction of genome claimed -- are quite different ratios for the different groups.

Perhaps it would make sense to consider the following:

1) establish allocations over the entire genome (as Yusuke has coordinated) for an initial phase of evenly-spaced, genome-wide genotyping. (The strategy document I sent out this week discusses some of the issues around initial strategy)

2) allocate a fraction of the genome (30-50%?) for finishing in the near term (once #1 is complete). The ratio of genome per group would be based on a fixed ratio of Mb per $ committed to each group by their funders.

3) Based on the first six months of the project -- and what we learn about the process and performance of each method and group - the rest of the genome will be allocated for finishing. This doesn't have to be all at once - it could be on a chromosome by chromosome basis as each group finishes their initial allocations.

It is worth considering that some technologies may be better suited to the initial screen as opposed to finishing. Finishing will require converting a very high fraction of all SNPs into working assays - when you get down to the smaller blocks, you need every SNP on the map (and more!) to define the entire region in blocks of strong LD. Some methods may fly along where there are many SNPs to choose from (because they are fast, if selective), and others may do better at finishing (because they have high conversion rates). If we don't think about this carefully, we may end up in a situation where finishing requires different methods (or, expensive SNP discovery - which some groups have already pointed out is not availabe in their budgets). Let's not put stakes down everywhere now, when we don't know what it will take for each of us (for any of us!) to actually finish a single chromosome. After we've done that once, we'll have a much better idea of how much of the genome can be done by each of us.

I believe that it makes more sense to do this than to place boundaries over the entire genome now - it may look quite different in a year or so, and it will be hard for any group to give up a part of the genome they claimed initially.

Best,

David

On Friday, October 25, 2002, at 02:07 AM, Yusuke Nakamura wrote:

> Dear Members of the chromosomal allocation group,
>
> Please find the updated information of the chromosomal allocations that I have heard from all of the members.
>
> As you can find in the attached sheet, there are some conflicts. A total proportion of chromosomes proposed from the US groups is now 37.1%. I have heard from Francis that US contributes to 30% of the chromosomes and would like to know the proportion assigned to each group prior to the meeting. (Lisa or Francis, could you let me know?)
>
> Chromosomes that were proposed by two groups are
> Chromosome 6 ; UK, BCM
> Chromosome 12; Illumina, BCM
> Chromosome X; Illumina, UCSF/WashU
> Chromosome 8p; MIT, China
>
> Chromosome 4; none
>
> We need to discuss to solve this issue in the meeting and hope you to compromise for establishment of the good international partnership.
>
> Yusuke Nakamura,M.D.,Ph.D.
>
>
>
>
>
>
> Yusuke Nakamura,M.D.,Ph.D.<CH. Allocation.xls>--
> Yusuke Nakamura,M.D.,Ph.D.
> Director, Human Genome Center
> Prof., Laboratory of Molecular Medicine
> [redacted]

****
David Altshuler, MD/PhD

Assistant Professor of Genetics and Medicine
Massachusetts General Hospital
Harvard Medical School

Director, Medical and Population Genetics
Whitehead/MIT Center for Genome Research
****

Fakunding, Patti (NIH/NHGRI)

From: [redacted]
Sent: [redacted]
To: [redacted]
Subject: FW: HapMap: Scientific Strategy

Unknown Document

Please print for me and FAX to the Four Seasons in Toronto
-----Original Message-----
From: [redacted]
Subject: Re: HapMap: Scientific Strategy

Attached is a draft discussion memo about scientific strategy in HapMap.

This was prepared based on discussions with many of you, and incorporates comments by our group at Whitehead/MGH (myself,
Mark Daly, Stacey Gabriel) and the group at Hopkins (Aravinda, Mike Zwick and David Cutler).

I look forward to comments and reactions. If we can have a round of email this week,
it will help us flesh out the issues for our meeting next week.

Best regards,

David

"Frampton, Lynn (NIH/NHGRI)" wrote:

> Dr. Altshuler,
>
> The names I have for the e-mail discussion group on the scientific strategy
> of the HapMap are listed below and in the cc field of this message. The
> leader of each discussion group should send an initial e-mail to get things
> started.
>
> All PIs will be included on the discussion groups. If anyone would like
> additional names added to this list, please let me know.
>
> Thanks,
> Lynn
>
> Group Leader
> Altshuler [redacted].edu
>
> PIs
> Nakamura y[redacted]

```
> Bentley          [redacted]
> Hudson
> Yang
> Chee
> Gibbs
> Chakravarti
> Kwok
>
> Additional Representatives
> Toshihiro Tanaka          [redacted]
[redacted]
>
> ____________________________
> Lynn Frampton, MPH
> Science Program Analyst
> National Human Genome Research Institute
> National Institutes of Health
> [redacted]
```

--
****
David Altshuler, MD/PhD

Assistant Professor of Genetics and Medicine
Massachusetts General Hospital
Harvard Medical School

Director, Medical and Population Genetics
Whitehead/MIT Center for Genome Research
****

Fakunding, Patti (NIH/NHGRI)

From: [redacted]
Sent:
To:
Subject: RE: Promotion Committee Meeting Confirmed

I really should go hear Jim Mullikin's presentation. So yes, cancel Program Staff.

-----Original Message-----
[redacted]
Subject: FW: Promotion Committee Meeting Confirmed
Importance: High

Do you want to cancel Program Staff on this date so you can attend this meeting?

-----Original Message-----
From: Claire Rodgaard [mailto:claire@nhgri.nih.gov]
Sent: Tuesday, October 22, 2002 2:36 PM
To: pmeltzer@nhgri.nih.gov; rlnuss@nhgri.nih.gov; jpuck@nhgri.nih.gov; tedyaz@nhgri.nih.gov; egreen@nhgri.nih.gov; of6@nhgri.nih.gov; Fakunding, Patti (NIH/NHGRI); mckoyp@nhgri.nih.gov; Graham, Bettie (NIH/NHGRI); rjk@nhgri.nih.gov; Adams, Linda (NIH/NHGRI); Rolfes, Ellen (NIH/NHGRI); andy@nhgri.nih.gov; mmuenke@nhgri.nih.gov; dhandon@nhgri.nih.gov; dwilson@nhgri.nih.gov; bpavan@nhgri.nih.gov; okalli@nhgri.nih.gov; jebw@nhgri.nih.gov; afw@nhgri.nih.gov; lesb@nhgri.nih.gov; pliu@nhgri.nih.gov; Willits, Victoria (NIH/NHGRI)
Subject: Promotion Committee Meeting Confirmed
Importance: High

Hi Everyone,

Thanks for your responses regarding a promotion committee meeting to review Dr. Jim Mullikin for appointment as a Senior Research Fellow in NISC.

The meeting is now confirmed as follows:

Wednesday, October 30, 12 noon
Building 50, 5th floor conference room

I will distribute a package for your review as soon as possible.

Joan and Alec, I will arrange a video connection for you through Jack Moore.

Claire

--
Claire Rodgaard
National Institutes of Health
National Human Genome Research Institute
Office of the Scientific Director
50 South Drive
[redacted]

SNP Discovery

Collins, Francis (NIH/NHGRI)

From: [redacted]
Sent:
To:

Cc: [redacted]

SNP Discovery Pre3.doc ATT25362.txt

Dear All,

Here is a draft document on SNP discovery to stimulate discussion both before (by email) and at the meeting. This document includes excerpts of information taken from several extensive analyses carried out by multiple laboratories, some stretching back a number of months. In the interests of clarity they are kept short and there may be insufficient explanation in some places. I would like to ask for one responder per group to take responsibility for emailing feedback before the meeting.

Many thanks to all contributors to the recent and earlier discussions.

Best wishes

David B.

Collins, Francis (NIH/NHGRI)

From: [redacted]
Sent: [redacted]
To: [redacted]
Cc: [redacted]

Subject: Re:

David -

I think that the document clearly articulates the key issues in SNP discovery for HapMap - that we clearly need more SNPs to complete a high-quality haplotype map over the entire genome.

While the detailed discussion will take place next week, I would add one point to the discussion about whole genome shotgun vs. targeted discovery.

One approach (WGS) discovers SNPs all over the entire genome - not just where you "need them" based on current ideas about haplotype map. The other approach (targeted resequencing of PCR products or BACs) is more focused.

A driving issue in evaluating the cost/yield of each strategy is the extent to which we value the discovery of SNPs everywhere (as opposed to just at the sites that we want to target). My own view is that SNPs everywhere have real value - they provide a resource for researchers who want to exceed the density provided by HapMap (or the population breadth), and they will support any future views of what a haplotype map may be. I think it would be short-sighted to cost out each approach based only on the yield of SNPs in the "gaps" of the map we are creating right now.

Best,

David

On Thursday, October 24, 2002, at 06:05 PM, David Bentley wrote:

> Dear All,
>
>
> Here is a draft document on SNP discovery to stimulate discussion both before (by email) and at the meeting. This document includes excerpts of information taken from several extensive analyses carried out by multiple laboratories, some stretching back a number of months. In the interests of clarity they are kept short and there may be insufficient explanation in some places. I would like to ask for one responder per group to take responsibility for emailing feedback before the meeting.
>
> Many thanks to all contributors to the recent and earlier discussions.
>
> Best wishes
>
> David B.<SNP Discovery Pre3.doc>

****

Collins, Francis (NIH/NHGRI)

From:
Sent:
To:
Cc:
[redacted]

Subject: Re:

Dear all the participants,

In response to the messages below, Japanese team would like to describe our situtation that we cannot afford the cost of SNP discovery; we obtained our budget only for SNP genotyping. In addition, our SNP discovery project has ended this March, and discovery team does not exist any more. We may not be able to do something for SNP discovery.

Sincerely,

Toshihiro Tanaka

Yusuke Nakamura

> David -
>
> I think that the document clearly articulates the key issues in SNP
> discovery for HapMap - that we clearly need more SNPs to complete a
> high-quality haplotype map over the entire genome.
>
> While the detailed discussion will take place next week, I would add
> one point to the discussion about whole genome shotgun vs. targeted
> discovery.
>
> One approach (WGS) discovers SNPs all over the entire genome - not just
> where you "need them" based on current ideas about haplotype map. The
> other approach (targeted resequencing of PCR products or BACs) is more
> focused.
>
> A driving issue in evaluating the cost/yield of each strategy is the
> extent to which we value the discovery of SNPs everywhere (as opposed
> to just at the sites that we want to target). My own view is that SNPs
> everywhere have real value - they provide a resource for researchers
> who want to exceed the density provided by HapMap (or the population
> breadth), and they will support any future views of what a haplotype
> map may be. I think it would be short-sighted to cost out each
> approach based only on the yield of SNPs in the "gaps" of the map we
> are creating right now.
>
> Best,
>
> David
>
> On Thursday, October 24, 2002, at 06:05 PM, David Bentley wrote:

```
>
>> Dear All,
>>
>>
>> Here is a draft document on SNP discovery to stimulate discussion both
>> before (by email) and at the meeting.   This document includes
>> excerpts of information taken from several extensive analyses carried
>> out by multiple laboratories, some stretching back a number of months.
>>  In the interests of clarity they are kept short and there may be
>> insufficient explanation in some places.  I would like to ask for one
>> responder per group to take responsibility for emailing feedback
>> before the meeting.
>>
>> Many thanks to all contributors to the recent and earlier discussions.
>>
>> Best wishes
>>
>> David B.<SNP Discovery Pre3.doc>
>
> ****************************
> David Altshuler, MD/PhD
>
> Assistant Professor of Genetics and Medicine
> Massachusetts General Hospital
> Harvard Medical School
>
> Director, Medical and Population Genetics
> Whitehead/MIT Center for Genome Research
> ****************************
>
>


--
Toshihiro Tanaka MD, PhD

Laboratory Head
Laboratory for Cardiovascular Diseases
SNP Research Center
RIKEN (The Institute of Physical and Chemical Research)

located at
Institute of Medical Science, University of Tokyo
```

[redacted]

[redacted]

Collins, Francis (NIH/NHGRI)

From: [redacted]
Sent:
To:
[redacted]

Subject: HapMap - sample availability update

Francis--

Attached are the Powerpoint slides for my Monday a.m. presentation on the status of the samples, in case you want a preview.

The most important info for the near-term relates to the re-consent of the CEPH samples. Mark Leppert is giving priority to the samples that David, Eric, et al. were already working on. So far there have been no refusals - just 2 families who have not yet been located - but Mark thinks he can find those families in the next couple of weeks and that they will also agree.

The signed re-consent forms from the rest of the "priority" CEPH trios should all be in hand in the next 2-3 weeks, plus forms from at least a few more families from the "non-priority" list, This should give us enough samples to fill one plate by mid to late November. By early to mid December, barring unforeseen problems, around 65 trios should be available.

I know that people will be frustrated that not all the CEPH samples are already re-consented and in-hand - but we did not get TSC's final revisions to the consent form until the end of June, and we have been pushing as fast and as hard as we can since then. The University of Utah IRB took longer than expected to act on this - something we should also be prepared for with the other IRBs.

You may also want to take a look at the timeline for the Yoruba samples (slide #15) - Charles is still set to collect these in the spring, but as you can see, this assumes that everything else that needs to happen between now and then proceeds on course (e.g., getting the other IRB approvals & State Dept. clearance, finalizing the subcontract with Univ. of Ibadan). Also, as we've discussed, the samples also need to be tested for HIV, which will complicate the process & inevitably cause some delay. So Charles projects that it will be August by the time all the samples arrive at Coriell (though many will come in before that as shipments will be staggered). Then, of course, Coriell needs a couple of months to make the cell lines.

Let me know if you have any questions about this before the meeting.

Sample Collection
for the HapM...

Jean E. McEwen, J.D., Ph.D.
Program Director
Ethical, Legal, and Social Implications Program
National Human Genome Research Institute
National Institutes of Health
[redacted]

Collins, Francis (NIH/NHGRI)

From: [redacted]
Sent:
To:

Subject: HapMap agenda question

AgendaOct02Meeting.doc

Attached is the current (slightly revised) HapMap meeting agenda.

Steve Sherry will discuss information on SNPs, such as number and distribution, as well as say that dbSNP will accept haplotypes. Would it make sense to have Steve talk (30 minutes) before David discusses strategy?

David, Is there any additional information that Steve should present?

I have asked Yusuke whether he needs all of the 45 minutes to discuss chromosome allocations.

Thanks, Lisa.

Lisa D. Brooks, Ph.D.
Program Director
Genetic Variation Program
Computational Genomics Program
[redacted]

Collins, Francis (NIH/NHGRI)

From: [redacted]
Sent: [redacted]
[redacted]

Subject: Re: HapMap agenda question

I think it would make sense to have Steve go before me - it is important information for both the strategy section and the SNP discovery discussion.

In addition, I think that perhaps strategy and analysis should go back to back (rather than separated). It is hard to discuss strategy without figuring out how and what you are going to analyze. I believe (if Aravinda agrees) that he and I could share these two sections - they are likely to be closely linked.

David

On Thursday, October 24, 2002, at 06:54 PM, Brooks, Lisa (NIH/NHGRI) wrote:

> <<AgendaOct02Meeting.doc>>
>
> Attached is the current (slightly revised) HapMap meeting agenda.
>
> Steve Sherry will discuss information on SNPs, such as number and
> distribution, as well as say that dbSNP will accept haplotypes. Would
> it
> make sense to have Steve talk (30 minutes) before David discusses
> strategy?
>
> David, Is there any additional information that Steve should present?
>
> I have asked Yusuke whether he needs all of the 45 minutes to discuss
> chromosome allocations.
>
> Thanks, Lisa.
>
> ____
>
> Lisa D. Brooks, Ph.D.
> Program Director lisa_brooks@nih.gov
> Genetic Variation Program
> www.nhgri.nih.gov/About_NHGRI/Der/variat.htm
> Computational Genomics Program
> [redacted]

****
David Altshuler, MD/PhD

Assistant Professor of Genetics and Medicine
Massachusetts General Hospital
Harvard Medical School

Director, Medical and Population Genetics

From: [redacted]
[redacted]

[redacted]

China is proliferating HapMap groups!
We already knew about Shanghai.

From C. Zeng:

We would like to put ~{!0~}The China HapMap Consortium~{!1~} in addition to BGI. The consortium has been established for the purpose of coordinating efforts by member groups in order to guarantee full accomplishment of the project. Henry is the coordinator and my institute will take full responsibility for the 10 % contribution by China. Also, there will be more groups, such as Hong Kong group, to join this consortium and contribute to the HapMap.

Prof. Qiang Boqin from the North Center, another group which would join
the HapMap, may not be able to attend the meeting.

Thanks, Lisa.

Lisa D. Brooks, Ph.D. [redacted]
Program Director
[redacted] htm
n.htm

From: Brooks, Lisa (NIH/NHGRI)
Sent: Monday, October 07, 2002 6:00 PM
To: [redacted]
Cc: [redacted]
Subject: Agenda Oct. 27-29

Attached is the agenda for the HapMap meeting on Oct. 27-29.
Please send me any items that should be on Tuesday morning under issues to be addressed.

AgendaOct02Meeting.doc

Thanks, Lisa.

Lisa D. Brooks, Ph.D.
[redacted] RI/Der/variat.htm
RI/Der/ginform.htm