Planning the First Large-Scale Phase of the Haplotype Map Project

Washington A Conference Room
Hyatt Regency Crystal City
2799 Jefferson Davis Highway
Arlington, VA 22202
703-418-1234
January 16, 2002

The purpose of this meeting is to prepare for the large-scale studies in the first three populations, by considering the populations to include, the scientific strategy, the project organization, and the timetable and costs of the project.

8:00 – 8:30	Continental Breakfast	
8:30 – 8:45	**Introduction:**	Francis Collins, Arthur Holden
	Populations	
8:45 – 9:00	**Data on haplotypes in populations:**	David Altshuler
9:00 – 9:15	**Discussion: Which populations should be included in the first phase of the large scale project?:**	Francis Collins
9:15 – 9:45	**Processes for collecting samples:**	David Valle, Ellen Clayton, Charles Rotimi
9:45 – 10:00	Break	
	Scientific strategy	
10:00 – 10:45	**Obtaining SNPs: How many more are needed, and how can they be obtained?:**	Eric Lander
10:45 – 11:30	**Genotyping and haplotyping:**	David Bentley
11:30 – 12:00	**Project milestones and work division:**	Eric Lander, David Bentley
12:00 – 12:15	**Data handling:**	Lon Cardon, Gudmundur Thorisson, Steve Sherry
12:15 – 12:30	Break to get working lunch	
	Project organization and related issues	
12:30 – 1:00	**Legal issues: Public access IP strategy, and primer access:**	Wayne Keown, Jorge Contreras
1:00 – 1:30	**Project organization: NIH, TSC, international:**	Francis Collins, Arthur Holden
1:30 – 1:40	**Management structure:**	Francis Collins, Arthur Holden
1:40 – 2:00	**Summary and next steps:**	Francis Collins, Arthur Holden

International collaboration in the HapMap project.

Draft for Discussion 1/14/02

A publicly available Haplotype Map of the human genome is a high priority resource needed by disease gene mappers as well as others interested in studying the function and evolution of the human genome. Over the last few months, NHGRI has supported an intensive planning effort to explore how a HapMap might best be developed so that it will provide optimal utility to researchers. From this planning activity, a number of research needs have been identified.

1. A larger number of diverse populations need to be studied to determine more systematically than has been done to date, the extent of haplotype variability in the human species. The following are needed to accomplish this:
    a) DNA samples and cell lines from 5 to 10 additional populations need to be obtained with the proper informed consent and community consultation.
    b) Samples from several populations need to be systematically analyzed in the same set of 10 to 60 chromosome regions.

2. Although a large number of SNPs are already in public databases, additional common SNPs will need to be isolated, possibly in a targeted way that focuses on regions where there are currently too few SNPS to make an accurate, useful map.

3. Technology needs to be developed for faster, cheaper SNP genotyping.

4. Efficient methods are needed for large-scale analysis of SNPs, haplotypes and their associations with genes affecting diseases and drug response.

5. A large-scale, highly coordinated effort needs to be organized and started as soon as possible to create a first generation human HapMap using two or three populations and a sufficient number of random SNPs to yield about 200,000 appropriately spaced SNPs across the genome.

6. The first generation HapMap will need to be refined and amplified, based on the results of the initial large-scale effort and the analysis of chromosome regions in multiple populations.

Developing a human HapMap is an ideal project for international participation. Many of the tasks 1 through 6 can be done in a distributed, loosely coordinated fashion. There are also many other complementary lines of research that could be pursued. The NIH welcomes and encourages participation by many groups. While NIH intends to support some research in all six areas, far more needs to be done than NIH will be able to do.

Based on the successful model of The SNP Consortium, NHGRI and TSC (including the Wellcome Trust) are planning a collaboration to address task 5 – the creation of a first generation HapMap. This project will be a direct continuation of the ongoing collaborations between TSC and NHGRI on SNP research. Other partners, such as Genome Canada, may also be involved. The project will be carried out as a centrally coordinated high throughput endeavor based at Institutions that have existing large-scale capacity. This approach will produce an initial map in the shortest feasible time, so that the benefits to human health and biomedical research can be garnered as soon as possible. However, the full utility of this map will depend on progress in tasks 1-4 being pursued in parallel and ultimately on task 6 as well.

In order to promote coordination of the entire HapMap effort and to expedite progress through exchange of information, setting of mutually agreed to standards, identification of additional needs etc., NHGRI proposes that all participants in this public effort meet approximately once a year.

Structure for NIH – TSC –collaboration on a HapMap project

Draft for discussion 1/14/02

The following outline describes a possible scenario by which NIH and TSC could collaborate to produce am initial HapMap of the human genome.

1. Additional population samples for both pilots and large scale study:
   NIH will organize and fund the collection of the additional samples needed. Funding will be through supplements to ongoing projects wherever possible. Most samples are anticipated to be collected by summer 2002 and available as cell lines and DNA by fall 2002.

2. Limited genotyping of additional population samples (pilot projects):
   NIH will organize and fund this activity to start when samples are available. NHGRI awardees known to have the capability for this activity will be invited to submit administrative supplement requests, which will be peer reviewed by an abbreviated method. The populations to be included and the regions to be studied will be specified in the request for supplements so that results can be readily compared and integrated. Funding will be in place by summer 2002, with most results available by the end of 2002.

3. Large-scale genotyping of three populations:
   TSC will start this activity in April 2002 using CEPH samples. TSC will use its own procedures for selecting the groups to do the work, including the data coordinating center. TSC will also set up whatever process is determined to be needed to protect IP.

NHGRI will issue an RFA by March 1 requesting cooperative agreement applications to do additional genotyping for CEPH samples and two other populations, potentially Yoruban and Japanese. Other NIH Institutes will be invited to join in this effort and contribute funding. The RFA will specify that the population samples and the SNPs to be genotyped will be determined by a steering committee. Applicants can apply to do a portion of the human genome, such as one or more chromosomes, and will compete on capacity, experience and cost. Applications will be due May 1 and will be reviewed in time for funding before September 30, 2002. It is anticipated that the large-scale genotyping will be completed in two years. How the results will be analyzed and presented will be decided by the participating groups.

4. Oversight:
A steering committee will be established with membership of TSC (including Wellcome?) and NIH. This group will manage and oversee the joint TSC/NIH project. Milestones will be set and progress of each funded group evaluated regularly. Groups that do not meet targets may be subject to redistribution of funds or removal. In order for this to work, it will be desirable that each funding agency has mechanisms in place to allow it to move funds between all the groups. For NIH to have this flexibility would mean that all participating groups need to apply for an NIH grant. Wellcome and TSC need to determine how they would ensure this flexibility.

5. Collaboration with other groups:
It is possible that other parties, such as Japan, may want to join in this project using their own funds. This can be accommodated if they agree to use the samples and SNPs endorsed by the steering committee and accept the data release policy that will be developed. They can then be given an allocation of chromosome(s) to work on. Coordination with such groups will have to be at a somewhat more distant level, because TSC/NIH would have no direct control over their funding. However, TSC/NIH would have to be ready to take up the slack, if such a group fails to perform.

There may be yet other groups who want to have a looser connection to the project. For example, the Chinese might want to genotype parts or all of the genome using their own samples. Coordination with such groups could be through periodic information exchange and perhaps an international gathering at some point during the project.

GENAISSANCE
PHARMACEUTICALS

XC Francis Collins
FYI,
Art

24 September 2001

cc: Barbara McGarvey
EJ
MG ✓
Lisa B

Arthur L. Holden
Chairman and CEO
First Genetic Trust, Inc.
3 Parkway North Center, Suite 150 North
Deerfield, IL 60015

Dear Arthur:

I apologize for the delay in responding to your thoughtful 14 July 2001 e-mail regarding recent comments made about the Human Genome Program and the proposed public/private consortium to create a genome-wide haplotype map (*New York Times*, *Wall Street Journal* – 13 July, 2001). I truly appreciate the important issues that you raised and would like to respond.

We applaud the accomplishments of the human genome initiative and The SNP Consortium because they have already produced information that is being used by the private sector to create products that address urgent medical needs. For example, Genaissance has a partnership with Janssen, a J&J company, to use our proprietary technology in drug development. We are committed to initiating additional agreements with pharmaceutical and/or biotechnology companies in the near term.

The goal of the human genome initiative was to produce a finished sequence of a single composite human genome. The sequence, in its current form, is already an important tool for researchers and a starting point for commercial entities to create products. The SNP Consortium was formed as a "pre-competitive" initiative to help annotate the composite genomic sequence by discovering and mapping single nucleotide polymorphisms. We at Genaissance examine a large number of individuals of diverse ancestry, and, as a result, greater than 95% of the SNPs we reported for the 313 genes in our *Science* paper are novel.

Genaissance was founded over four years ago with the mission of revolutionizing healthcare by improving and customizing treatments based upon each person's DNA. Our recent publications in *Science* and *PNAS* reflect the achievements that we have made in industrializing the process of discovering haplotypes and using these haplotypes to correlate gene variation with clinical response. In addition to Genaissance, a number of other companies currently have large-scale efforts to use the human genome sequence to identify haplotypes. The purpose of this innovative and competitive work is to meet an urgent medical need, i.e. to correlate gene variation with disease susceptibility and drug response. Clinical trials addressing the genetic basis of drug safety and efficacy are already underway at Genaissance, as evidenced by our current STRENGTH Study. STRENGTH is designed to identify gene haplotypes that can be used to improve the efficacy and safety profile of an important class of drugs, the statins.

Therefore, we believe the planned haplotype consortium would place the federal government in direct competition with genomics companies in the private sector. The government and The SNP Consortium have fulfilled their mission. It is now time for the private sector to create the innovative products that will revolutionize our healthcare system.

Arthur, I value our relationship and hope that we can continue to discuss these issues.

Sincerely,

Gualberto Ruaño, M.D., Ph.D.
Chief Executive Officer

$40M +
Sanger/Wellcome

A haplotype map of the human genome

Goal

The next key step of the Human Genome Project (following the creation of the genetic, physical and sequence maps) is the generation of a "haplotype" map of the human genome. Such a "haplotype" map consists of a high density of SNPs defining the small number of ancestral haplotypes (blocks of tightly correlated genetic variants) in each region of the human genome. Knowledge of these haplotypes will allow comprehensive and efficient testing of the association of human genes with human diseases. The haplotype map can and should be generated rapidly and should be made freely available to researchers worldwide.

Background

A haplotype map of the human genome has become both justified and practical due to significant advances over the last two years. Specifically, these include:

- **Genomic Sequence**: The development of a complete genome sequence - integrated with human genes and annotations - providing a reference framework on which to layer knowledge about allelic variation.
- **Genetic Variants**. The development of a dense map of 1.4 million human SNPs (and rapidly growing), provides a genome-wide resource of genetic variation adequate to uniquely tag the vast majority of human haplotypes.
- **Genotyping Technology**: The development of a high-throughput methods, allowing a rapid, efficient and cost-effective experimental approach to a project of the required scale.
- **Long-range LD**. The discovery that human SNPs display strong linkage disequilbrium (LD or allelic association) over large distances. LD is detectable over distances in the range of 100kb and is extremely strong over regions spanning several tens of kb (the size of typical genes). For such regions, the vast majority of chromosomes in the population carry one of a handful of highly conserved haplotypes. As a result, genetic diversity in the region can be represented by a small number of well-chosen SNPs.

Impact on biomedical research

The availability of a haplotype map of the human genome will have a substantial impact on human genetic studies. Specifically, these include:

- **Comprehensive association studies of individual genes**. The association of genes with disease has traditionally been probed by testing individuals SNPs one-at-a-time. The drawback to this approach is that the task is never-ending: one can exclude particular SNPs as playing a role, but one cannot exclude a gene. Once the haplotype structure of the genome is defined, one can (i) comprehensively test all significant haplotypes in the gene and (ii) decrease the number of SNPs needed by selecting a subset that defines the population variability. This will allow haplotype studies of individual genomic loci in an unbiased manner, without assumption about the locations of causal mutations in coding regions, promoters, or regulatory sites at significant distance away. And, it will greatly decrease the technical and financial barriers faced by laboratories in undertaking such work

- **Genome-wide association studies**. A genome-wide haplotype map will make possible whole-genome scans for association in the population. Rather than focusing only on 'candidate' genes, it will become possible to search the genome in an unbiased manner for genes whose common variation contributes to disease in the population. Routine use of genome-wide association studies will also require further decreases in genotyping costs, but such decreases are likely to be driven by the development of the haplotype map.
- **Human Population Stucture and History**. Knowledge of haplotypes will transform our understanding of human population structure and history. The LD pattern turns out to be an extremely sensitive indicator of population history, because the multi-allelic nature of haplotypes provides rich detail and because the breakdown of haplotypes follows a predictable clock set by recombination rates. In particular, LD patterns are more powerful than traditional studies of allele frequencies per se. Information about human population history is interesting in its own right, but is also very valuable in the design of medical studies (such as admixture mapping).

Technical Issues

Generating a haplotype map would involve the following components:

- **Population Samples**. Development of appropriate population samples, consisting of parent-offspring trios (to allow inference of haplotypes). We estimate that a total of about 300 samples will be needed, representing major ethnic groups in a manner appropriate for generating a map that can be used for medical studies in all populations. The population samples should be a renewable resource (i.e., immortalized cell lines).
- **Sample and Data Availability**. The samples should be made freely available so that any interested scientific group can contribute data (in the manner of the CEPH and NIH diversity panels). Conversely, all data generated by the project should be immediately released into the public domain without restrictions of any kind
- **Numbers of SNPs to be genotyped.** It is estimated that generating the haplotype map will require successful genotyping of 450,000 SNPs, which will in turn require initial testing of some 800,000-900,000 SNPs. The required scale is now well within reach: the Whitehead and Sanger Centre are each currently engaged in pilot projects involving 25,000 SNPs using automated genotyping setup and MALDI-TOF-based detection. Given the required scale and efficiencies, it is likely that the bulk of the work should be performed by a few large groups. But, all groups should be encouraged to participate in the project by analyzing genes and regions of interest.
- **Analytical Tools**. The project will require various analytical tools to readily define haplotype blocks from genotype data, software systems to aid in the hierarchical selection of SNPs to fill in blocks, and databases to make the information maximally useful to the community. Prototype systems have been developed, but focussed effort will be needed to develop mature systems.

Prepared by:
David Altshuler, *Harvard Medical School, Massachusetts General Hospital, Whitehead Institute*
Eric Lander, *Whitehead Institute and MIT*

DEPARTMENT OF HEALTH & HUMAN SERVICES Public Health Service

National Institutes of Health
National Human Genome
Research Institute
31 Center Drive MSC 2152
Building 31, Room 4B09
Bethesda, MD 20892-2152
Telephone: (301) 496-0844
Fax: (301) 402-0837

September 18, 2001

Gerald F. Vovis, Ph.D.
Chief Technology Officer
Senior Vice President
Genaissance Pharmaceuticals, Inc.
Five Science Park
New Haven, Connecticut 06511

Dear ~~Dr. Vovis~~ Gerry:

Thank you for sending a letter to the members of the National Advisory Council on Human Genome Research, expressing some concerns about the current plan for generating a public haplotype map of the human genome. Your letter refers to some of the previous published data on linkage disequilibrium (LD) in the human genome, but does not seem to take into account additional experimental results that were presented by several groups at the NIH Workshop held here in July. Accordingly, I shared your letter with Dr. Eric Lander, who has responded with the attached letter and preprint of some of his work. Note that this manuscript and additional publications on these issues are expected in the October issue of Nature Genetics.

I hope you find this information helpful.

Sincerely yours,

Fran

Francis S. Collins, M.D., Ph.D.
Director

Enclosures

cc: Eric Lander, Ph.D.
Elke Jordan, Ph.D.
National Advisory Council Human Genome Research Members

FSC/phf

WHITEHEAD INSTITUTE

Eric S. Lander, Ph.D.
Member,
Whitehead Institute for Biomedical Research
Professor of Biology,
Massachusetts Institute of Technology
Director, Whitehead Institute/MIT Center
for Genome Research

One Kendall Square, Bldg. 300
Cambridge, Massachusetts 02139-1561
617.252.1906 / 617.252.1933 fax
lander@genome.wi.mit.edu

September 17, 2001

Francis Collins
National Center For Human Genome Research
Building 31, Room 4B39
9000 Rockville Pike
Bethesda, MD 20892

Fax: (301) 402-0837

Dear Francis:

Thank you for passing on the letter from Dr. Vovis raising a scientific question about the work on haplotype maps. Dr. Vovis cites two recent papers, one of which is from my own laboratory.

Dr. Vovis's concern is based on his intuition that, "if such [haplotype] blocks existed, the LD [between pairs of SNPs] would be expected to decrease monotonically and gradually as physical distance between a pair of SNPs increases."

As it happens, contrary to Dr. Vovis's intuition, the presence of haplotype blocks is not expected to cause pairwise LD between SNPs to decrease monotonically. Broadly speaking, this is because SNPs are binary and thus cannot perfectly correspond to haplotypes. On the other hand, LD does tend to decrease monotonically when measured with respect to haplotype blocks rather than individuals SNPs.

This is described in great detail in a paper from our laboratory that will appear in the October issue of Nature Genetics. I enclose a preprint [Daly et al.], which you may wish to forward to Dr. Vovis.

In addition, I understand that the same issue of Nature Genetics has articles from two other groups reporting similar findings about large haplotype blocks. These may also be of interest to Dr. Vovis.

Finally, I am aware of three even larger studies that will be appearing over the next months that document haplotype blocks on the scale of many megabases. Much of this work was presented at the NIH meeting in July.

In closing, Dr. Vovis's letter is very helpful because it underscores the counter-intuitive nature of population genetics and suggests that it may be helpful for someone to write an accessible review on the subject of haplotypes.

Sincerely,

Eric S. Lander
Member, Whitehead Institute for Biomedical Research
Professor of Biology, MIT
Director, Whitehead Institute/MIT
Center for Genome Research

ESL/gbs
Enclosure

High-resolution haplotype structure in the human genome

Mark J. Daly[1], John D. Rioux[1], Stephen F. Schaffner[1], Thomas J. Hudson[1,2] & Eric S. Lander[1,3]

Published online: XXX 2001, DOI: 10.1038/ngXXX

Linkage disequilibrium (LD) analysis is traditionally based on individual genetic markers and often yields an erratic, non-monotonic picture, because the power to detect allelic associations depends on specific properties of each marker, such as frequency and population history. Ideally, LD analysis should be based directly on the underlying haplotype structure of the human genome, but this structure has remained poorly understood. Here we report a high-resolution analysis of the haplotype structure across 500 kilobases on chromosome 5q31 using 103 single-nucleotide polymorphisms (SNPs) in a European-derived population. The results show a picture of discrete haplotype blocks (of tens to hundreds of kilobases), each with limited diversity punctuated by apparent sites of recombination. In addition, we develop an analytical model for LD mapping based on such haplotype blocks. If our observed structure is general (and published data suggest that it may be), it offers a coherent framework for creating a haplotype map of the human genome.

In a companion project, we are studying a 500-kb region on human chromosome 5q31 that is implicated as containing a genetic risk factor for Crohn disease[1]. After high-density SNP discovery, we selected 103 common (>5% minor allele frequency) SNPs genotyped in 129 trios from a European-derived population. Our results thus describe 258 chromosomes transmitted to individuals with Crohn disease and 258 untransmitted chromosomes.

The genotype data provide the highest-resolution picture to date of the patterns of genetic variation across a large genomic region, with a marker density of 1 SNP roughly every 5 kb. For studying both disease association (marker versus disease) and LD (marker versus marker), the traditional approach has been to perform single-marker analysis. Examples of such analysis are shown in Fig. 1. Although there are clearly many strong correlations, the picture is noisy and unsatisfying, and important localization information is obscured by properties of the markers not relevant to the issues under study.

To obtain a clearer picture, we focused on identifying the underlying haplotypes. It became evident that the region could be largely decomposed into discrete haplotype blocks, each with a striking lack of diversity (Fig. 2). Our initial focus was on untransmitted control chromosomes; however, the same

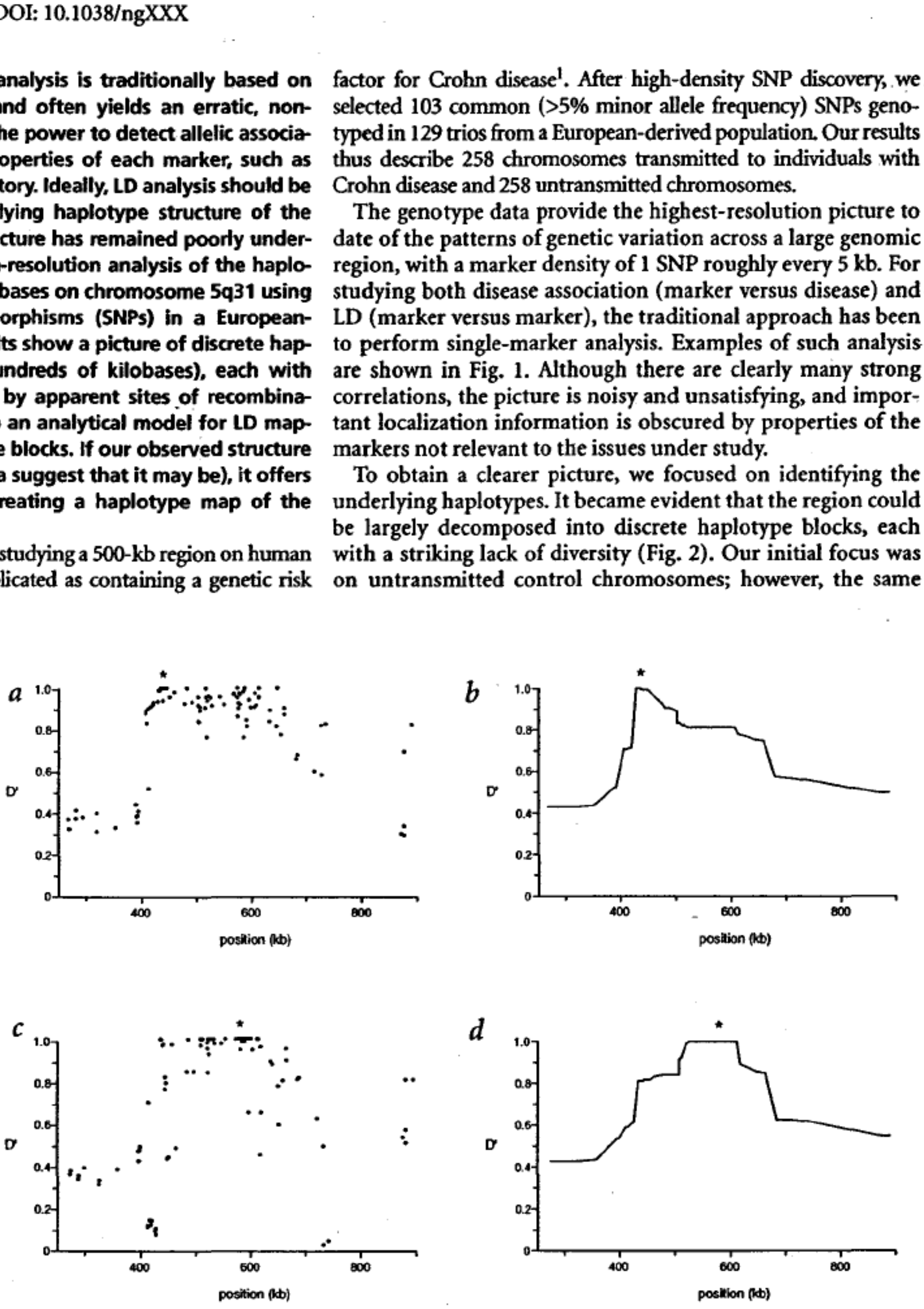


Fig. 1 Comparison of single-marker LD with haplotype-based LD. **a**, LD between an arbitrary marker (at the 26th position, indicated with an asterisk) and every other marker in the data set using *D'*. **b**, Multiallelic *D'* is used to plot LD between the maximum-likelihood haplotype group assignment at the location of the 26th marker and that assignment at the location of every other marker in the data set. **c**,**d**, Repeat of the comparison in **a** and **b** but with respect to a second marker (at the 61st position) in the map. Both pairs of graphs show the common feature that, when haplotypes rather than individual SNP alleles are considered to be the basic units of variation, the noise (presumably caused by marker history and properties of the specific statistic chosen) essentially disappears, resulting in a clear, monotonic and step-like breakdown of LD by recombination.

[1]*Whitehead Institute/Massachusetts Institute of Technology, Center for Genome Research, Cambridge, Massachusetts, USA.* [2]*Montreal Genome Center, McGill University, Montréal, Québec, Canada.* [3]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence and requests for materials should be addressed to M.J.D. (e-mail: mjdaly@genome.wi.mit.edu) or E.S.L. (e-mail: lander@wi.mit.edu).*

haplotype structure was seen in the chromosomes transmitted to individuals with Crohn disease, with the only difference being that one of the haplotypes was enriched in frequency, reflecting its association to Crohn disease[1]. Because this structure is the same in both groups, we present combined data from all chromosomes (transmitted and untransmitted).

The haplotype blocks span up to 100 kb and contain multiple (five or more) common SNPs. The blocks have only a few (2–4) haplotypes, which show no evidence of being derived from one another by recombination, and which account for nearly all chromosomes (>90%) in all cases in the sample. For example, an 84-kb block shows only two distinct haplotypes that together account for 95% of the observed chromosomes (Table 1). The lack of diversity is readily seen from the fact that the probability an individual is homozygous for all SNPs genotyped in a block ranges from 30–70%.

The discrete blocks are separated by intervals in which several independent historical recombination events seem to have occurred, giving rise to greater haplotype diversity for regions spanning the blocks. The most common recombination events are indicated in Fig. 2 by lines connecting the haplotypes. The recombination events appear to be clustered; multiple obligate exchanges must have occurred between most blocks, with little or no exchange within blocks. For example, in the 84-kb block (Table 1), not a single apparent recombinant between the two major haplotypes was observed (despite the fact that such a recombinant would be obvious because the haplotypes differ at all SNPs examined).

The clustering is suggestive of local hotspots of recombination[2–4], and the same observation of inhomogeneity of recombination is made for the class II region of the MHC elsewhere in this issue[5]. Although there is detectable recombination between blocks, it is modest enough for there to be clear long-range correlation (that is, LD) among blocks. The haplotypes at the various blocks can be readily assigned to one of four ancestral long-range haplotypes. Indeed, 38% of the chromosomes studied carried one of these four haplotypes across the entire length of the region.

Using a hidden Markov model (HMM), we developed an approach to define the block structure formally. The HMM simultaneously assigns every position along each observed chromosome to one of the four ancestral haplotypes and estimates the maximum-likelihood values of the 'historical recombination frequency' (Θ) between each pair of markers. The quantity Θ provides a convenient summary of the degree of haplotype exchange across inter-marker intervals and relates directly to the conventional measures of LD, such as D'. (An alternative measure is the joint probability of homozygosity[6].) In the case at hand, the discrete block structure is evident from the fact that Θ is estimated at less than 1% for 73 of the inter-marker intervals, 1–4% for 14 of the intervals, and more than 4% for only 9 of the intervals.

We considered whether the selection of the SNPs could have significantly influenced the results. The SNPs studied were ascertained by complete resequencing of seven individuals with Crohn disease and one control[1]. To test whether this survey failed to

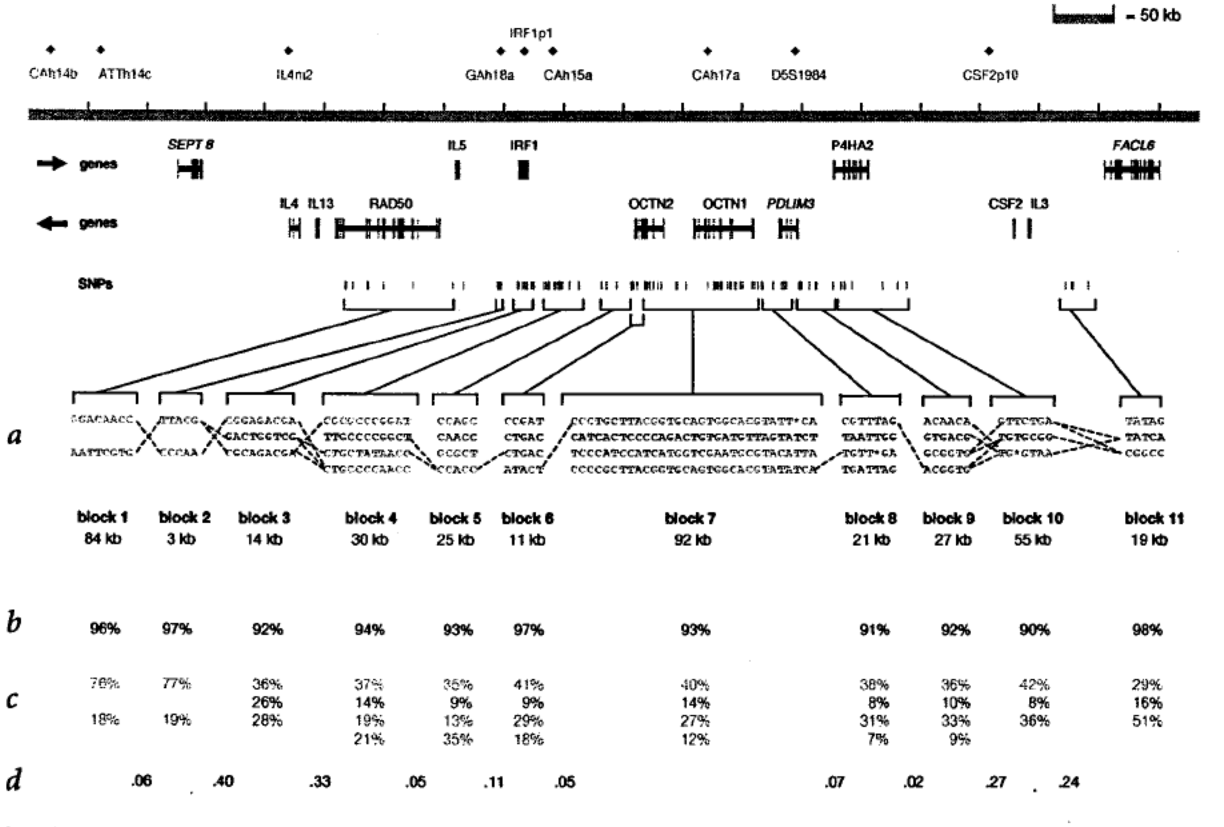


Fig. 2 Block-like haplotype diversity at 5q31. ***a***, Common haplotype patterns in each block of low diversity. Dashed lines indicate locations where more than 2% of all chromosomes are observed to transition from one common haplotype to a different one. ***b***, Percentage of observed chromosomes that match one of the common patterns exactly. ***c***, Percentage of each of the common patterns among untransmitted chromosomes. ***d***, Rate of haplotype exchange between the blocks as estimated by the HMM. We excluded several markers at each end of the map as they provided evidence that the blocks did not continue but were not adequate to build a first or last block. In addition, four markers fell between blocks, which suggests that the recombinational clustering may not take place at a specific base-pair position, but rather in small regions.

detect much of the common variation, we compared our SNPs in a 100-kb subregion to those identified by the International SNP Map Working Group (ISMWG)[7]. We detected 47 of the 54 SNPs (86%) reported by the ISMWG, a rate that exceeds the proportion of ISMWG SNPs (ascertained in a multi-ethnic panel) typically found to be polymorphic in a Caucasian population (roughly 80%; S. Bolk, personal communication). In addition, we discovered 150 SNPs in this region not reported by the ISMWG.

This analysis used SNPs with minor allele frequency greater than 5%. We genotyped six rarer SNPs and found that the rare allele fell exclusively or nearly exclusively on one of the major haplotype patterns and simply created a subtype of that pattern. This underscores that, when we refer to limited haplotype diversity, we are not implying complete sequence identity among chromosomes with the same haplotype, but rather that chromosomes fall into a small number of deep clades. Chromosomes within a clade may differ at one or a few rare SNPs, whereas chromosomes in different clades differ at many SNPs. Finally, we note that we initially eliminated SNPs at CpG sites because the higher mutation rate at such sites[8,9] might introduce recurrent mutation and thereby confound the analysis. Of the 16 high frequency CpG SNPs genotyped, 13 had alleles that aligned perfectly with the haplotype patterns in Fig. 1 and only one added significantly to the overall heterozygosity of the block in which it fell.

Our analysis of this region of chromosome 5q31 in a European-derived population indicates the following: the region may be largely divided into discrete blocks of 10–100 kb; each block has only a few common haplotypes; and the haplotype correlation between blocks gives rise to long-range LD. Determining whether these are general features of human genetic variation will require studies of other regions with similarly dense genetic maps (increasingly feasible given the availability of human genome sequence[10] and large SNP collections[7]); however, available evidence seems to be consistent with this picture. In numerous data sets, comprehensive SNP genotyping in small regions (2–5 kb upstream from candidate genes) indicates limited haplotype diversity (3 or 4 haplotypes accounting for 80–95% of all observed chromosomes[11–14]), similar to the data presented here. Together with observations of an unexpectedly long extent of LD[15–17], these reports suggest that our description of haplotype diversity in 5q31 may be, in qualitative terms, fairly general.

The structure of LD described here has important implications for the analysis of LD, for association studies to find medically relevant variation, for population genetics, and for the next steps of the Human Genome Project.

Focusing on haplotype blocks greatly clarifies LD analyses. Once the haplotype blocks are identified, they can be treated as alleles and tested for LD (for example, our simple analysis uses Hedrick's multi-allelic extension of D'[18,19], thereby reflecting the underlying population variation more accurately than any individual SNP. The power of the haplotype-based approach can be seen by comparing the noisy single-marker analyses of LD (Fig. 1*a*,*c*) with corresponding analyses performed on the underlying haplotype blocks (Fig. 1*b*,*d*). The latter analyses show that LD decays monotonically (as expected if recombination has the main role in the breakdown), with the decrease occurring in abrupt drops reflecting the sites of significant historical recombination.

In analogous fashion, the haplotype structure provides a crisp approach for testing the association of genomic segments with disease. By contrast, disease association studies traditionally involve testing individual SNPs in and around a gene. This approach is statistically weak and has no clear endpoint: true associations may be missed because of the incomplete information provided by individual SNPs; negative results do not rule out association involving other nearby SNPs; and positive results do not indicate the discovery of the causal SNP but simply a marker in LD with a true causal SNP located some distance (perhaps several genes) away. Once the haplotype blocks are defined, however, it is straightforward to examine a subset of SNPs that uniquely distinguish the common haplotypes in each block (shown elsewhere in this issue)[20]. This allows the common variation in a gene to be tested exhaustively for association with disease (given a specified level of genotype relative risk and disease allele frequency). Although this analysis, such as presented in the companion paper[1], will not always directly result in the identification of the causal gene and mutation, it focuses subsequent functional studies on the critical region of maximum haplotype distortion within which there exists insufficient historical recombination for variation studies to reduce it further. (In addition, although association studies with haplotypes are much clearer than those with individual SNPs, we note that strict monotonic decay of association is not expected, even with perfect haplotype data, for reasons described elsewhere[21].)

Table 1 • Haplotypes of SNPs in block 1 (8SNPs/84 kb)

Haplotype	Observations
G G A C A A C C	283 (83.2%) haplotype A
A A T T C G G G	40 (11.8%) haplotype B
G A T T A G C C	2 (0.6%)
G G T C A G C C	2 (0.6%)

*Another 13 chromosomes (3.8%) were observed that matched haplotype A or B at all alleles except one, and might represent gene conversion or an undetected genotyping error.

The structure and composition of the haplotype blocks have considerable implications for human population genetics. The data here are broadly consistent with coalescent simulations[17], which suggest that models, including both inhomogeneous recombination (reflecting the apparent clustering of major recombinational events) and recent bottlenecks (accounting for the limited number of distinct haplotypes over long distances seen in Canadians (non-black, non-Asian from metropolitan Toronto), may be necessary to explain modern human diversity. Detailed haplotype analysis of many genomic regions in several populations, together with comprehensive simulation studies, will be needed to determine the relative importance of these and other factors.

Finally, our approach provides a precise framework for creating a comprehensive haplotype map of the human genome. By testing a sufficiently large collection of SNPs, it should be possible to define all of the common haplotypes underlying blocks of LD. Once such a map is created, it will be possible to select an optimal reference set of SNPs for any subsequent genotyping study. Such a project is becoming feasible, and this detailed understanding of common human variation represents an important step in the Human Genome Project.

Methods

Individuals and marker selection. The individuals studied, Canadians from metropolitan Toronto (non-black, non-Asian), and the genotyping methodologies are described in the companion paper[1]. To ensure our ability to reconstruct multi-marker haplotypes, SNPs for haplotype analysis were selected from the set of markers for which full genotypes were available for all members of 85 or more trios. To eliminate markers likely to contain significant numbers of undetected genotyping errors, markers not in Hardy–Weinberg equilibrium ($P<0.05$) or those for which more than 10 Mendelian inheritance errors were detected were excluded from this analysis. SNPs at CpG sites were not included in the initial analysis to prevent potential confounding of common haplotype patterns from recurrent mutation. In addition, rare SNPs (minor allele frequency <5%) were not included in the initial analysis. The underlying data for this analysis is contained on our website (http://www-genome.wi.mit.edu/humgen/IBD5).

Haplotype counting. Haplotype percentages in Fig. 2 were computed by using haplotypes generated by the transmission disequilibrium test (TDT) implementation in Genehunter 2.0 (ref. 22), followed by use of an EM-type algorithm[23,24], to include the minority of chromosomes that had one or more markers with ambiguous phase (that is, where both parents and offspring were heterozygous) or where one marker was missing genotype data. Clark's method[25], or simply counting only fully informative phase-known haplotypes, provided essentially identical answers, because within each block most chromosomes were fully reconstructed without ambiguity from the parental data.

Regions of low-haplotype diversity were initially identified as follows: five-marker haplotypes for all consecutive sets of five markers were generated; the observed haplotypic heterozygosity ($\mathrm{HET}_{\mathrm{obs}}$) and expected haplotypic heterozygosity ($\mathrm{HET}_{\mathrm{exp}}$) (given allele frequency and assuming equilibrium) were tallied; and each five-marker window was assigned a score, $S_5=\mathrm{HET}_{\mathrm{obs}}/\mathrm{HET}_{\mathrm{exp}}$. A smaller value therefore represents lower diversity of haplotypes compared with expectation. Windows with locally minimal scores were then expanded or contracted by adding or subtracting markers to the ends to find the longest local minimum window. Boundaries between these windows (which we call 'blocks') were examined. The most common connections between haplotypes considered to be the 'ancestral haplotype class' (displayed on the same line in the same color in Fig. 2), and cases in which a high frequency (>2%) haplotype is observed that represents a connection between two different 'ancestral classes' are shown by a line connecting those classes across that interval.

Hidden Markov model. The observations that over long distances most haplotypes can be described either as belonging to one of a small number of common haplotype categories, or as a simple mosaic of those categories, suggested the use of an HMM in which haplotype categories were defined as states. We assigned observed chromosomes to those hidden states (allowing for missing/erroneous genotype data), and simultaneously estimated the transition probability in each map interval by using an EM algorithm and by making the simplifying assumption that there was one transition probability for each map interval (the aforementioned probability of historical recombination Θ) rather than allowing specific transition probabilities from each state to each state. The output of this method was a maximum-likelihood assignment to haplotype category at each position (which can be used to compute, for example, multi-allelic D' and TDT) and maximum-likelihood estimates of Θ indicating how significantly recombination has acted to increase haplotype diversity in each map interval. The use of probabilities of recombination in this context[6] has a simple relationship with the most commonly used measure of gametic disequilibrium (D'). If we consider two SNPs at a time before any recombination (or other type of event) has occurred to create a fourth haplotype (as in the following table):

	SNP 2	
	Allele 1	Allele 2
SNP 1 Allele 1	a	b
Allele 2	c=0	d

we can see that D' (which equals $(ad-bc)/[(a+c)(c+d)]$ for this table configuration) is equal to 1 (full disequilibrium). Many generations later, we can collapse all recombination that has occurred between the two markers into a single value: the probability that a modern chromosome has undergone recombination at any time between those two markers. Let $(1-\Theta)$ represent the probability that no recombination has taken place at any time between these two markers. At this time, the table of haplotype frequencies will have changed to

	SNP 2	
	Allele 1	Allele 2
SNP 1 Allele 1	$a-ad\Theta$	$b+ad\Theta$
Allele 2	$ad\Theta$	$d-ad\Theta$

And now $D'=(a-d\Theta)(d-d\Theta)-bd\Theta\ -(d\Theta)^2/\mathrm{ad}$, which reduces to $(ad-ad\Theta)/ad$, and thus $D'=(1-\Theta)$. Θ here (Θ_{real}) differs from the observed rates (Θ_{obs}) reported in Fig. 2, as some recombinations occur between chromosomes with identical local haplotypes; however, the observed values are trivially corrected by the local homozygosity to produce the real values.

Acknowledgments

The authors thank D. Reich, D. Altshuler, J. Hirschhorn, K. Lindblad-Toh and M.P. Reeve for many valuable discussions and comments on the manuscript, A. Kirby for informatics support, the technicians in the inflammatory disease research group at the Whitehead Institute Center for Genome Research for their skilled genotyping work and the anonymous referees for their helpful comments. The authors would also like to thank L. Gaffney for her help in the preparation of this manuscript. This work was supported by research grants from Bristol-Myers Squibb, Millennium Pharmaceutical, and Affymetrix.

Received 9 May; accepted 29 August 2001.

1. Rioux, J.D. *et al.* Hierarchical linkage disequilibrium mapping of a susceptibility gene for Crohn's disease to the cytokine cluster on chromosome 5. *Nature Genet.* **29**,223–228 (2001).
2. Templeton, A.R. *et al.* Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am. J. Hum. Genet.* **66**, 69–83 (2000).
3. Jeffreys, A.J., Ritchie, A. & Neumann, R. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum. Mol. Genet.* **9**, 725–733 (2000).
4. Smith, R.A., Ho, P.J., Clegg, J.B., Kidd, J.R. & Thein, S.L. Recombination breakpoints in the human β-globin gene cluster. *Blood* **92**, 4415–4421 (1998).
5. Jeffreys, A.J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222 (2001).
6. Sved, J.A. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Pop. Biol.* **2**, 125–141 (1971).
7. The International SNP Map Working Group. A map of the human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
8. Krawczak, M., Ball, E.V. & Cooper, D.N. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am. J. Hum. Genet.* **63**, 474–488 (1998).
9. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
10. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
11. Drysdale, C.M. *et al.* Complex promoter and coding region β2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc. Natl Acad. Sci. USA* **97**,10483–10488 (2000).
12. Park, H.Y. *et al.* Identification of new single–nucleotide polymorphisms in the thrombin receptor gene and their effects on coronary artery diseases in Koreans. *Clin. Exp. Pharmacol. Physiol.* **27**, 690–693 (2000).
13. Jordanides, N., Eskdale, J., Stuart, R. & Gallagher, G. Allele associations reveal four prominent haplotypes at the human interleukin-6 (IL-6) locus. *Genes Immun.* **1**, 451–455 (2000).
14. D'Alfonso, S., Rampi, M., Rolando, V., Giordano, M. & Momigliano–Richiardi, P. New polymorphisms in the IL–10 promoter region. *Genes Immun.* **1**, 231–233 (2000).
15. Bonnen, P.E. *et al.* Haplotypes at ATM identify coding–sequence variation and indicate a region of extensive linkage disequilibrium. *Am. J. Hum. Genet.* **67**, 1437–1451 (2000).
16. Moffatt, M.F., Traherne, J.A., Abecasis, G.R. & Cookson, W.O. Single nucleotide polymorphism and linkage disequilibrium within the TCR α/δ locus. *Hum. Mol. Genet.* **9**, 1011–1019 (2000).
17. Reich, D.R. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
18. Lewontin, R.C. The interaction of selection and linkage. General considerations; heterotic models. *Genetics* **49**, 49–67 (1964).
19. Hedrick, P.W. Gametic disequilibrium measures: proceed with caution. *Genetics* **117**, 331–341 (1987).
20. Johnson, G.C.L. *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29**, 233–237 (2001).
21. Kruglyak, L. & Lander, E.S. High-resolution genetic mapping of complex traits. *Am. J. Hum. Genet.* **56**, 1212–1223 (1995).
22. Daly, M.J. *et al.* Genehunter 2.0—a complete linkage analysis system. *Am. J. Hum. Genet.* **63**, A286 (1998).
23. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **39**, 1–38 (1977).
24. Excoffier L. & Slatkin M. Maximum–likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**, 921–927 (1995).
25. Clark A.G. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7**, 111–122 (1990).

THE UNIVERSITY OF TEXAS
SOUTHWESTERN MEDICAL CENTER
AT DALLAS

Michael S. Brown, M.D.
Paul J. Thomas Professor of Genetics
Director, Jonsson Center for Molecular Genetics

5323 Harry Hines Boulevard
Dallas, Texas 75390-9046
214-648-2179 Telefax 214-648-8804

September 14, 2001

Francis S. Collins, M.D., Ph.D., Director
National Human Genome Research Institute
National Institutes of Health
31 Center Drive MSC 2152
Building 31, Room 4B09
Bethesda, MD 20892-2152

cc - Jeff T
Eric G
Andy B
Elke
Kathy H
Mark G

Dear Francis:

Thank you for inviting me to join the intramural planning group. This is an important function, but unfortunately my schedule makes it impossible for me to participate. I could not make either of the initial meetings this fall, and could probably not make the subsequent meetings either.

On another note, I was disappointed to learn about Rick's leaving the NCI. I have heard that the Department has begun to micro-manage the administrative functions of the Institutes. This is a tragedy. We obviously need a strong NIH Director who can negotiate with the administration. It would be a shame if the most talented people left NIH, especially in light of the recent tragedy, which will likely have major effects on federal priorities.

Again, I'm sorry I cannot join your panel. Good luck with the Genome Research Institute.

Yours sincerely,

Michael

Michael S. Brown, M.D.

MSB:mah
B6/098

DANA-FARBER
CANCER INSTITUTE

Children's Hospital Boston

Stuart H. Orkin, M.D.
Chairman, Department of Pediatric Oncology
Dana-Farber Cancer Institute
David G. Nathan Professor of Pediatrics
Leland Fikes Professor of Pediatric Medicine
Harvard Medical School
Investigator, Howard Hughes Medical Institute

Dana-Farber Cancer Institute
44 Binney Street
Boston, Massachusetts 02115
617.632.3564 tel, 617.632.4367 fax
stuart_orkin@dfci.harvard.edu

September 17, 2001

Francis S. Collins, M.D., Ph.D.
Director
National Institutes of Health
National Human Genome
Research Institute
31 Center Drive MSC 2152
Building 31, Room 4B09
Bethesda, MD 20892-2152

cc: Jeff Trent
Eric Green
Andy Baxevanis
Elke Jordan
Kathy Hudson
Mark Guyer ✓

Dear Francis:

Unfortunately, my schedule is too full to permit me to participate in your intramural review Nov 5th and December 13-14th. Perhaps with better lead-time I will be ok next time.

Best regards,

Stuart H. Orkin, M.D.
David G. Nathan Professor of Pediatrics
Leland Fikes Professor of Pediatric Medicine
Chairman, Department of Pediatric Oncology

nature genetics
ARCHIVE SEARCH INFORMATION CLASSIFIED SUBSCRIBE
News & Views

volume 29 no. 2 pp 109 - 111

OCT 2001
FULL TEXT
PDF
REFERENCES

Islands of linkage disequilibrium

David B. Goldstein

Department of Biology, University College London, 4 Stephenson Way, London NW1 2HE, UK. e-mail: d.goldstein@ucl.ac.uk

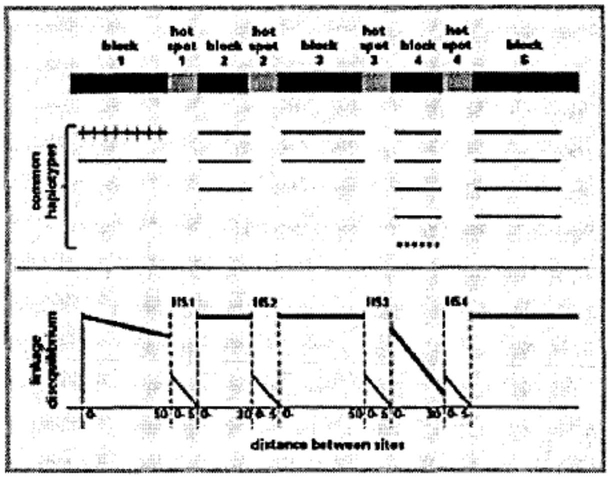
A detailed knowledge of patterns of linkage disequilibrium in human populations is widely seen as a prerequisite for effective population-based disease gene mapping. New data suggest that linkage disequilibrium is highly structured into discrete blocks of sequence separated by hot spots of recombination.

The natural world is not famous for making life easy for human geneticists. Despite application of an increasingly powerful set of tools provided by the Human Genome Project, the complexity of common diseases has made them largely refractory to genetic analysis. In the face of this complexity, geneticists agree that the family-based approaches that proved so successful for the monogenic diseases are not up to the job. Instead, most favor association studies, in which genetic and phenotypic variation is compared in large population samples in order to identify correlations implicating genetic risk factors. The classic example is the case-control study in which, for example, a group of sufferers from a condition is compared to a group of healthy controls.

Linkage disequilibrium in gene mapping

Association studies, however, introduce a new complication. Linkage disequilibrium (LD) is the non-random association between alleles at different loci, and it creates opportunities as well as difficulties in gene mapping. In the extreme case, an allele found at one locus predicts which allele would be found at the other, making one of the loci redundant for mapping purposes. Thus, it should, in principle, be possible to use knowledge of LD to design strategies that could represent most of the variation in the genome by genotyping only a small fraction of the polymorphic sites. On the other hand, it is difficult, and sometimes impossible, to pick out the causal variants in a set of sites in strong LD using only association data (see figure).

REFERENCES



The lowdown on LD. Idealized representation of block-like structure of linkage disequilibrium, with regions of low haplotype diversity separated by recombinational hot spots. Lines below the blocks represent examples of the number of common haplotypes that might be present for such blocks. SNPs distinguishing the two common haplotypes in block 1 are represented by short vertical lines. The graphs plot (idealized) LD as a function of distance, averaged across pairs of sites, either for sites within a given block or within a hot spot. The plots show that within a block LD decays only gradually with distance, or not at all. Within hot-spot areas, however, LD falls away much more rapidly with distance. If no LD-generating event, such as a bottleneck, has recently occurred in the population, then there may have been enough recombination across the hot spots that the haplotypes in adjacent blocks are randomly associated. Similarly, with sufficient time, or in blocks with higher within-block recombination rates, LD may be substantially reduced for distant sites within a block, as represented here in block 4. Note that for block 1, any of the SNPs indicated would be sufficient to represent the majority of the haplotypic variation within this block. If haplotype 1 were shown to increase the risk of a condition relative to haplotype 2, however, it would be impossible to determine from association data which of the SNPs distinguishing haplotypes 1 and 2 was the biological cause of the increased risk.

BOB CRIMI

For these reasons there has been an explosion of interest in patterns of LD in human populations. In an ideal world the patterns would be highly consistent, in which case it would be relatively straightforward to figure out how to make LD work for us and not against us. Early signs, however, were not encouraging. For example, Clark *et al.*[1] suggested that extremely variable patterns of LD over relatively short sequence stretches, such as they observed in the liboprotein lipase gene (*LPL*), might seriously complicate association studies. A practical example of the difficulty of separating causation and correlation comes from an association study that implicated a promoter variant in the gene *CYP3A5* (which encodes a member of the cytochrome P450 IIIA subfamily) as determining polymorphic activity of liver CYP3A5 (ref. 2). A subsequent sequencing study of the *CYP3A* locus, however, demonstrated that the implicated variant was in the promoter region of a *CYP3A* pseudogene (*ps1*) about 25 kb upstream from *CYP3A5*, and so was very unlikely to be the cause of the polymorphic activity[3].

REFERENCES

In coarser assessments of broad patterns of LD, both in genes and in non-coding regions, the overall picture has been one of inconsistency, with patterns differing markedly across genomic regions and across populations and failing to show a consistent dependence on genetic distance. For example, Reich *et al.*[4] studied 160 kb at each of 19 different regions, documenting consistently maximal LD across some of the regions, highly variable patterns across others, and large average differences between African and north European populations.

In fact, such complexity is not unexpected. Factors such as selection and local mutation and recombination rates can influence LD in particular genomic regions, and population factors such as changes in population size and admixture between differentiated populations can influence genome-wide patterns of LD. Although extreme variability of LD would not preclude mapping genes with association data, it would certainly complicate the job of obtaining a detailed description of genome-wide patterns of LD.

Recombination hot spots

Two lines of evidence published in this issue now suggest that human LD might not be so complicated after all. Characterizing a 500 kb stretch of chromosome 5q31, Mark Daly and colleagues[5] (page 229) note that the pattern of LD is highly structured, with stretches of consistently high LD interspersed with short intervals of rapid breakdown. In fact, the 500-kb interval can be divided into 11 blocks of consistently high LD. In spite of including more than 75% of the total sequence, the blocks show no evidence for recombination and extremely little haplotype diversity.

Daly *et al.*[5] suggest that the blocks of LD result from the localization of recombination to irregularly spaced hot spots, an explanation given dramatic support by the study of Alec Jeffreys and colleagues[6] (page 217) on LD and recombination in the HLA region. A high-resolution analysis of LD throughout 216 kb of the HLA class II region again shows clear block-like structure. Estimation of recombination frequencies in sperm reveals three clusters of recombinational hot spots, apparently accounting for at least 94% of the observed recombinations and corresponding to areas of LD breakdown. Moreover, within these areas, the fine-scale pattern of LD corresponds to the precise boundaries of hot spots 1–2 kb long.

REFERENCES

As noted by Daly *et al.*[5], this atomistic picture of LD is largely consistent with published data. For example, the plot of LD against distance averaged across the 19 regions studied by Reich *et al.*[4] shows a gradual decay with distance, but this gradual decay is not recapitulated within regions. In fact, on visual inspection, at least half of the regions show precisely the pattern noted by Daly *et al.*[5], with LD appearing to collapse at a specific distance, whereas most of the others give no clear impression. In the case of *LPL*, Templeton *et al.*[7] had already noted that recombination is apparently mainly restricted to a 1.9-kb segment and that haplotype diversity is again very simple outside this segment.

Even in this atomistic world there does remain a quantitative aspect to patterns of LD, both in inter-block disequilibrium, observed in the 5q region by Daly *et al.*[5], and in the possibility of gradual decay of LD within blocks, observed in the HLA region by Jeffreys *et al.*[6] but not in 5q, possibly as a result of the greater genealogical depth for the HLA region or greater within-block recombination rates.

The greatly reduced LD in Africa reported by Reich *et al.*[4] and others could therefore result from less inter-block LD, sharper decay of disequilibrium within blocks or some combination of these two. Indeed, at the extreme, if a population has been stable and of large size for sufficiently long, then there could be enough recombination within some blocks to reduce LD between distant markers within blocks to unmeasurable levels. In this case, the clear block structure reported in Europeans could be absent from other populations and/or from certain regions in which within-block recombination rates are sufficiently high (for example, 'block' 4 in the figure).

Implications for association studies

A largely atomistic pattern of LD would have profound implications both for the description of LD and for typing strategies and data analysis in association studies. With respect to LD, the goal would largely switch from the assessment of its quantitative dependence on genetic distance to the detection of block boundaries and the coupling strengths of associations across them. With blocks defined, a relatively small number of markers could represent most of the haplotype diversity in the genome. Covering hot-spot regions, however, would considerably increase the number of markers required relative to the amount of sequence represented by hot spots[6]. For example, Daly *et al.*[5] note that in their first block of LD, spanning 84 kb, 96% of the chromosomes sampled have one of two haplotypes. Just one single-nucleotide polymorphism (SNP) would therefore be sufficient to represent most of the diversity in this 84-kb stretch in Europeans, while in other blocks a handful at most would be required (because of the ascertainment process, everything discussed here relates to common SNPs). The observed pattern of LD also means, however, that association data will generally be insufficient to pinpoint the precise causal SNP within blocks of high LD.

REFERENCES

These implications of the block structure of LD are clearly illustrated in the study of John Rioux and colleagues[8] (see page 223) on a risk factor for Crohn disease in the 5q31 region. The simple haplotype structure throughout the region provides clear and strong evidence that one of the haplotypes is a risk factor for Crohn disease, with an increased risk of 2.0 in heterozygous form. The at-risk haplotype extends over 250 kb, however, and the association data provide no means of selecting the SNP that is responsible for increased risk out of the many SNPs that are uniquely associated with the risk-conferring haplotype.

It is worth noting that the existence of discrete blocks of consistent LD would probably reduce the need for sophisticated population-genetic inference in gene mapping. Instead of using formal models to estimate statistically the precise location of a causal variant, the focus would shift to simple assignment of detected effects to a hot spot–delimited block. Fine assignment within the block would rely on the biology of the trait or, where possible, on cross-population comparisons looking for either rare discrete recombinants or 'old' populations in which LD has had time to decay even within blocks. For example, the CYP3A mystery was resolved when an intronic splice-site polymorphism in *CYP3A5* was shown to be in complete association with the *ps1* SNP. The

CYP3A5 variant was the clear choice on biological grounds, but the inference was supported by analysis of enzymatic activity in another ethnic group in which the association between the two SNPs was incomplete[9]. This underscores the importance of describing haplotype structure in multiple populations.

Despite its limited scope, the current evidence is compelling and implies that studies should now assume that LD has a block-like structure until such a view either is confirmed or becomes unsustainable. But that still leaves difficult questions about priorities and strategies in characterizing global patterns of LD. For example, one discouraging note in the current batch of papers comes from the study of John Todd and colleagues[10] (page 233), in which 135 kb spanning 9 genes are characterized. They observed long stretches of LD with very limited haplotype diversity, but found that the SNPs currently available in dbSNP (http://www.ncbi.nlm.nih.gov/SNP) are not generally sufficient to capture that diversity. Todd and colleagues[10] propose a systematic program to search for SNPs in exons and in small upstream and downstream regions around genes (compare ref. 11). Daly *et al.*[5], on the other hand, clearly envisage a genome-wide effort.

I can see some argument for doing gene regions first. But I think it is clear that the whole genome must be characterized, given our ignorance of, for example, the distance between enhancer elements and the genes they influence. Ignoring currently available SNPs when determining the haplotype structure of genes seems wasteful, however. An inspection of dbSNP makes clear the tremendous variability of currently available SNPs per kilobase in different genes. It seems likely, therefore, that available SNPs would be sufficient to capture haplotype diversity for some blocks but not others, and that a more efficient strategy would be to characterize available SNPs in targeted regions first, followed by SNP discovery as necessary where the available SNPs appear insufficient to capture the haplotype diversity.

Association studies suddenly look much less difficult than before, and the case for an international project explicitly dedicated to describing haplotype structure in multiple populations now seems overwhelming. It is essential that such a project be as inclusive as possible to ensure that the results can be used to study conditions prevalent in any and all human populations.

REFERENCES

1. Clark, A.G. *et al. Am. J. Hum. Genet.* **63**, 595-612 (1998). | Article | PubMed | ISI |
2. Paulussen, A. *et al. Pharmacogenetics* **10**, 415-424 (2000). | Article | PubMed | ISI |
3. Finta, C. & Zaphiropoulous, P.G. *Gene* **260**, 13-23 (2000). | Article | PubMed | ISI |
4. Reich, D.E. *et al. Nature* **411**, 199-204 (2001). | Article | PubMed | ISI |
5. Daly, M.J. *et al. Nature Genet.* **29**, 229-232 (2001). | Article |
6. Jeffreys, A.J. *et al. Nature Genet.* **29**, 217-222 (2001). | Article |
7. Templeton, A.R. *et al. Am. J. Hum. Genet.* **66**, 69-83 (2000). | Article | PubMed | ISI |
8. Rioux, J.D. *et al. Nature Genet.* **29**, 223-228 (2001). | Article |
9. Kuehl, P. *et al. Nature Genet.* **27**, 383-391 (2001). | Article | PubMed | ISI |
10. Johnson, G.C.L. *et al. Nature Genet.* **29**, 233-237 (2001). | Article |
11. Stephens, J.C. *et al. Science* **20**, 489-493 (2001).

Population/ELSI Group
Report of ELSI Sub-Group

Draft 9/24/01

The ELSI sub-group of the Population/ELSI group has begun to discuss a range of ELSI issues related to both the pilot study and the main project. Discussion so far has been aimed primarily at establishing criteria for evaluating the acceptability of the informed consents for the existing samples being considered for the pilot study. The following six issues have been identified:

1. There is consensus that the consent forms should specify, at a minimum, that the samples would be used for genetic research. There should also have been explicit consent for the making of cell lines (in instances where cell lines would be used) and for sharing the samples with other researchers.

2. There is general agreement that the consent forms should specify that the samples would be used for studies of genetic variation. An unresolved issue is whether the forms should be required to go further and state specifically that samples would be used for studies looking at variation *within and between populations*. There is general agreement that samples collected for research aimed specifically at studying the genetic etiology of a particular identified disorder (or type of disorder) would *not* be appropriate for use in the pilot unless the consent form were written in more general terms. This is because individuals' assessments of the potential for benefit or harm may be influenced by whether or not they perceive the research as focusing on a disorder of particular interest to them.

3. There is general agreement that some form of community consultation is desirable before samples will be used, although this may be difficult for the pilot study with respect to samples from older sets that were collected before community consultation for genetic variation research became the accepted practice. Some pilot studies of community consultation should be undertaken simultaneously with other parts of the pilot. It was recognized that community consultation should be viewed as a process of engaging affected communities and assessing a range of responses, and does not ordinarily mean that there needs to be formal community consent (except with certain populations).

4. There is general agreement that the informed consent should have been given under appropriate conditions and with appropriate conversations (which would generally require some type of inquiry beyond examination of the consent form). It was recognized, however, that the details of actual consent processes may be difficult to evaluate with older sample sets given the length of time that has passed since the samples were collected.

The criteria listed above will also apply to any *new* samples that will have to be collected for either the pilot study or the main study. In addition, the informed consent for the main study will raise additional issues and will need to be more comprehensive.

The group has begun the process of examining individual consent forms for each of the sample sets under consideration for the pilot study. The group believes that it *may* be appropriate to use the Utah samples from the CEPH collection for the pilot even though they do not fulfill all of these criteria. The suggestion has been made that there may be a justification for treating the CEPH samples somewhat differently given that those samples have already been so widely studied (and used in other studies of LD), given that the population from which the samples were collected is a majority U.S. population, and given that samples were quite clearly given as an altruistic donation to science. The CEPH samples may *not* be appropriate for use in the *main* study, however, given the absence in the CEPH consent form of any explicit mention of genetic variation research (as distinct from genetic research more generally) and the problematic nature of the concept of "presumed consent."

Discussion is continuing on a number of other ELSI issues relevant to both the pilot and the main study. These include: the advantages and disadvantages of including particular populations (such as populations from places other the U.S., Europe, and Japan); methods to be used for designating individuals as belonging to a particular population; goals of and methods to be used for community consultation; and protections for privacy and confidentiality. The group has also discussed the need for the development of a sound communication strategy to make sure that the public accurately understands the project.

Guyer, Mark (NHGRI)

From: Pilar Ossorio [redacted]]

Sent: Friday, September 21, 2001 8:24 PM

To: [redacted]l)

Subject: inf'd consent issues

Hi All,

After a phone call with the "existing resources" working group, Morris and I agreed to write something about the consent question and circulate it. We would like something to go to Eric L. and Francis sometime this weekend, so if you can quickly chime in on the memo below with comments and questions , please do.

****

Question: If existing samples are used to conduct the pilot study for the haplotype map project, what criteria should be used to evaluate the acceptability of the existing informed consents?

Suggested criteria for evaluating existing informed consents:

1)genetic research should have been specified.

> This one is a deal breaker; if there was not consent for genetics research, then we cannot use the samples.

2)studies of genetic variation within and between populations should have been specified.

> Here Morris Foster and I have some disagreement. He sent me the following: "Samples that lack explicit approval for cross-population genetic variation research may be considered if some language in the consent talks about some groups having greater burden of disease than others because of genetic factors (this is common language in studies targeting African Americans for sickle cell trait research, hypertension research, etc.). Actually, language of that sort probably does a marginally better job at concretely communicating the idea of genetic variation as being important in the study of disease genes than more abstract language that mentions haplotypes. However, many consents for disease-specific research explicitly limit the use of the samples to studying that particular disease. I don't believe that those limitations can be overcome for a general haplotype project."
>
> I feel that it is OK to use samples for which the consent did not explicitly mention haplotyping, so long as the consent was very clear concerning the fact that genetic variation within and between populations would be studied. I am not comfortable accepting consents that talk about

different disease burdens in populations if the purpose of the original research was to study the genetic etiology of a particular disease or some diseases. People's assessments of the potential for harm/benefit are likely influenced by whether they perceive the project as focusing on a disease of interest to them or not. Therefore, I would not accept prior consent that focused on the different frequencies of a disease in different populations, unless the consent also talked about non-disease genetics research and population comparisons.

3)there should have been explicit consent for sharing the samples.

This one is a deal breaker; if it does not exist, then we cannot use the samples.

4)if we are going to use cell lines, then there should have been explicit consent for the making of cell lines (it seems basic, but I know of samples collected in the 80s that were used to make cell lines without explicit consent for cell lines. The fact that cell lines were made came as a very unwelcome surprise to some of the people who gave the original samples).

5) community consultation

In cases where there was no consultation at the time samples were taken, but other aspects of consent meet our minimum criteria, it may be possible for NHGRI to conduct a consultation before the samples are used in the pilot.

6) Consent should have been given under appropriate conditions and with appropriate conversations

I do not know how easy or difficult this will be to evaluate. Aravinda has told us about some samples that have been and are being collected in India, for which both individual consent and community consultation/consent exists. These consents were videotaped and consent was specifically given for population variation studies. This sounds like the ideal; however, I suspect that few samples will have documentation beyond a signed form. The question is, what should we accept in terms of evidence that the consent process was valid???

Ciao,
Pilar

****

Pilar N. Ossorio, Ph.D., JD
Assistant Professor of Law and Medical Ethics
University of Wisconsin Law School
975 Bascom Mall, rm. 9103
Madison, WI 53706-1399

[redacted]

"...
Waves of anger and fear
Circulate over the bright

And darkened lands of the earth,
Obsessing our private lives;
The unmentionable odour of death
Offends the September night.
...
Who can release them now,
Who can reach for the deaf,
Who can speak for the dumb?
..."

W.H. Auden, poem titled "Sept. 01, 1939"

Guyer, Mark (NHGRI)

From: Aravinda Chakravarti, Ph.D. [redacted]

Sent: Friday, September 21, 2001 9:47 PM

To: Pilar Ossorio

Cc: [redacted]

Subject: Re: inf'd consent issues

Aravinda's quick comments:

****

Question: If existing samples are used to conduct the pilot study for the haplotype map project, what criteria should be used to evaluate the acceptability of the existing informed consents?

Suggested criteria for evaluating existing informed consents:

1)genetic research should have been specified.

yes.

2)studies of genetic variation within and between populations should have been specified.

I dont agree with either of you ! I do not believe that disproportionate disease burden should be required or that the form mention within-vs-between studies. The form should mention the use in studies of genetic variation. The apportionment of variation within-vs-between can be within the sampled population as well with respect to some other criteria such as autosomal vs. x linked or high vs low recombination.

3)there should have been explicit consent for sharing the samples.

yes.

4)if we are going to use cell lines, then there should have been explicit consent for the making of cell lines (it seems basic, but I know of samples collected in the 80s that were used to make cell lines without explicit consent for cell lines. The fact that cell lines were made came as a very unwelcome surprise to some of the people who gave the original samples).

yes although older forms may use some other word such as immortalized or permanent source etc.

5) community consultation

Yes but I have problems with who speaks for the entire group ?

6) Consent should have been given under appropriate conditions and with appropriate conversations

Yes but is not crucialif the above are satisfied.

Cheers,

Aravinda

Aravinda Chakravarti, Ph.D.

Guyer, Mark (NHGRI)

From: McEwen, Jean (NHGRI)

Sent: Saturday, September 22, 2001 12:11 PM

To: [redacted]

Subject: RE: inf'd consent issues

Thanks, Pilar & Morris, for helping to take this discussion to the next level.

To give us something concrete to look at as we continue these discussions, I am attaching at the bottom of this e-mail the text of the consent form that was used for the collection of samples from the 48 Utah Mormon families in the CEPH collection (the forms that were used for the French & Venezuelan CEPH samples are not available.) I decided just to re-type the text of the form rather than fax you all copies.

As you can see, the form does not remotely resemble what would be required by today's standards, although it is fairly reflective of the consents that were used back in the 80's, when (I believe) the samples were collected. Of the criteria Pilar & Morris have listed, really only one (the specification for genetics research) is satisfied. There's no explicit mention of genetic variation research (although, on the plus side, these samples were not collected in connection with the study of any PARTICULAR genetic disorder). Apart from the reference to "the various laboratories," there's no explicit consent for sharing the samples (although these samples have, as a matter of fact, been widely shared for many years now). There's no explicit mention of cell lines. There also was (presumably) no community consultation since the very notion of community consultation for genetics research did not exist when these samples were collected. Determining the exact conditions under which the consents were given (apart from what's stated in the form) is also likely to be difficult, given the length of time that has passed since the samples were collected.

There are obviously strong arguments for not using these CEPH samples for the main study, in light of the deficiencies noted above. The more immediate issue is whether these samples (& other collections of samples with similar consents) should ALSO be "off limits" for the PILOT, given the more limited purpose of the pilot & the practical impossibility of collecting over the next 6 months all new samples in the numbers that the Methods group has determined is necessary (50 samples each from 12 populations, with trios preferred). I raise this question NOT to suggest a particular answer, but just to more precisely frame the issue for debate. (Also - I could not be on the Resources sub-group call, yesterday, so I apologize if some of what I'm saying here has already been hashed over.)

It seems that one point to keep in mind is that if it turns out that none (or almost none) of the existing samples satisfy the stringent ethical standards we'd demand for the main study, we will necessarily, as a practical matter, need to limit the number of populations we can look at in the pilot. This, in turn, could raise a whole separate set of ethical problems, because we'll have to limit ourselves to those populations that are "easiest" to collect samples from within a tight frame - which could raise fairness issues & unwittingly have the effect of making the project start to look like the creation of a "race map" (e.g., if we end up only with samples from groups of Europeans, African Americans, and/or Japanese).

Perhaps this is an argument for re-evaluating whether it's even realistic to think about doing the pilot in 6 months; we obviously do not want to be in the position of having to make unacceptable ethical tradeoffs just to meet an artificial deadline. Still, even with a longer timeframe for completion of the pilot, it's hard to imagine how we can realistically collect brand new samples (or get all the re-consents) from the number of people that we'll need if we really are going to sample from 12 (or even 3-4) populations.

Another issue to consider: Is there a basis for treating the CEPH samples somewhat differently (for purposes of the pilot) than we would treat (for example) an existing sample set collected in Africa or another developing nation? The CEPH samples have already been very widely studied & used in other linkage disequilibrium studies (e.g., Eric Lander's recent study). Given the population from which they were collected (Utah Mormon), these samples may ARGUABLY not be quite as problematic as some as the others, despite the limitations of the consent form.

The full text of the Utah CEPH consent form is set out below. I look forward to people's further input on this. (I am adding Sharon Terry and Susan Zullo to this e-mail to make sure that everyone on the Pop/ELSI group has a chance to weigh in on these issues.)

CONSENT FOR PARTICIPATION IN AN INVESTIGATIONAL STUDY OF GENETIC MAPPING OF HUMAN CHROMOSOMES

In order to study how an individual inherits genetic traits or markers from his or her parents, it is valuable to know on which chromosome those specific genes are located. Researchers have developed laboratory techniques which will alow us to locate these genetic markers and establish maps of them. Moreover, these markers can be used to locate genes which cause inherited diseases.

Participation in this study involves drawing a 30 ml sample of blood, which is equal to 1/4 cup, into vacutainer tubes. The amount of blood to be drawn will be determined by the person's age, height and weight. This amount is necessary for the various laboratories to run the numerous tests needed to establish "maps" of the human chromosomes. Risks for drawing blood are: superficial bruises, bleeding from the site of the puncture, and uneasiness associated with needles. Potential benefits of the study will be a greater understanding of genetic defects in humans.

Participation in all aspects of this study is voluntary. All records and other information obtained will be kept strictly confidential. These records will be used only by authorized people in health research. No names will be released or published in reports. Questions concerning the research project should be directed to the Eccles Institute of Human Genetics, Ray White ([redacted]; Mark Leppert ([redacted], or Leslie Jerominski [redacted]

I have read the foregoing and my questions have been answered. I will recieve a copy of this signed and dated consent form for my records. I would not object to being contacted by research personnel in the future if they feel it would help in their effeorts to better understanding [sic] human genetics.

Medical Treatment or Compensation for Injury

In the event you sustain injury resulting from your participation in the research project, the University of Utah can provide to you, without charge, emergency and temporary medical treatment not otherwise covered by your own insurance. If you believe that you have sustained an injury as a result of your participation in this research program, please contact the Office of the Vice President for Research, phone numbe[redacted]. If you have questions pertaining to your rights as a subject, you may call the Institutional Review Board Office at [redacted].

The data obtained from the present study may be used for medical and scientific purposes. A copy of this authorization shall be as valid as the original authorization.

______________ ____________________

Witness Signature

____

Date

Jean E. McEwen, J.D., Ph.D.
Program Director
Ethical, Legal, and Social Implications Program
National Human Genome Research Institute
National Institutes of Health
31 Center Drive, Room B2B07
Bethesda, MD 20892-2033
(301) 402-4997 (phone)

Guyer, Mark (NHGRI)

From: Valle, David (NIDCD)
Sent: Saturday, September 22, 2001 4:49 PM
To: [redacted]
Subject: Re: inf'd consent issues

Pilar and Morris -
Thanks, this is very helpful. Points 1, 3, 4, 5 and 6 seem straightforward and I agree with you. Point 2 is more problematic. I agree that specific limitation to a particular disease seems a deal breaker. Inclusion of language indicating "genetic variation" or "genetic differences" seems important but also seems to require a "between whom" to be fully informative. I lean towards Aravinda - if the the consent is an open-ended "genetic variation" without specifiying between whom, it would seem to give us carte blanche.
David
--
Professor of Pediatrics, Molecular Biology and Genetics
Howard Hughes Medical Institute and The Institute of Genetic Medicine
Johns Hopkins University School of Medicine
802 PCTB
725 N. Wolfe St.
Baltimore MD, 21205
email d[redacted]
phone [redacted]
FAX 410-955-7397

On Friday, September 21, 2001 8:23 PM, Pilar Ossorio <[redacted]du> wrote:
>Hi All,
>
> After a phone call with the "existing resources" working
>group, Morris and I agreed to write something about the consent
>question and circulate it. We would like something to go to
>Eric L. and Francis sometime this weekend, so if you can
>quickly chime in on the memo below with comments and questions
>, please do.
>
> *****************
> Question: If existing samples are used to conduct the pilot
>study for the haplotype map project, what criteria should be
>used to evaluate the acceptability of the existing informed
>consents?
>
>Suggested criteria for evaluating existing informed consents:
>
>1)genetic research should have been specified.This one is a
>deal breaker; if there was not consent for genetics research,
>then we cannot use the samples.
>
>2)studies of genetic variation within and between populations
>should have been specified. Here Morris Foster and I have some
>disagreement. He sent me the following: "Samples that lack
>explicit approval for cross-population genetic variation
>research may be considered if some language in the consent
>talks about some groups having greater burden of disease than
>others because of genetic factors (this is common language in
>studies targeting African Americans for sickle cell trait
>research, hypertension research, etc.). Actually, language of
>that sort probably does a marginally better job at concretely
>communicating the idea of genetic variation as being important

Guyer, Mark (NHGRI)

From: Ellen Wright Clayton, MD, JD [redacted]
Sent: Saturday, September 22, 2001 8:22 PM
To: McEwen, Jean (NHGRI)
Cc: [redacted]
Subject: Re: inf'd consent issues

Card for Ellen Wright Clayton,...

I will take Jean up on her invitation to address the concrete example of the CEPH consent form and urge that it is probably acceptable to use these samples for the pilot given the particular way that CEPH have been used over the years. The form does talk about genetics, and clearly implies sharing among qualified investigators. I am made most uncomfortable by the fact that the form talks about looking for disease genes, not understanding human variation, but I am largely consoled by the facts that CEPH was clearly an altruistic donation to science, pretty broadly writ, and that these samples have been used a lot. What do other people think?

> "McEwen, Jean (NHGRI)" wrote:
>
> Thanks, Pilar & Morris, for helping to take this discussion to the next level.
>
> To give us something concrete to look at as we continue these discussions, I am attaching at the bottom of this e-mail the text of the consent form that was used for the collection of samples from the 48 Utah Mormon families in the CEPH collection (the forms that were used for the French & Venezuelan CEPH samples are not available.) I decided just to re-type the text of the form rather than fax you all copies.
>
> As you can see, the form does not remotely resemble what would be required by today's standards, although it is fairly reflective of the consents that were used back in the 80's, when (I believe) the samples were collected. Of the criteria Pilar & Morris have listed, really only one (the specification for genetics research) is satisfied. There's no explicit mention of genetic variation research (although, on the plus side, these samples were not collected in connection with the study of any PARTICULAR genetic disorder). Apart from the reference to "the various laboratories," there's no explicit consent for sharing the samples (although these samples have, as a matter of fact, been widely shared for many years now). There's no explicit mention of cell lines. There also was (presumably) no community consultation since the very notion of community consultation for genetics research did not exist when these samples were collected. Determining the exact conditions under which the consents were given (apart from what's stated in the form) is also likely to be difficult, given the length of time that has passed since the samples were collected.
>
> There are obviously strong arguments for not using these CEPH samples for the main study, in light of the deficiencies noted above. The more immediate issue is whether these samples (& other collections of

Guyer, Mark (NHGRI)

From: Vivian Ota Wang [redacted]

Sent: Saturday, September 22, 2001 6:24 PM

To: [redacted]

Subject: Re: Inform Consent Issues

Dear All,

First, thanks Pilar and Morris for putting something together for us to further gather our collective thoughts and feelings about these issues.

Points 1, 3, and 4 seem straightforward. For Point 2, I have similar sentiments as Pilar and feel that (1) clear language and understanding is necessary about intra- and inter- genetic variation and (2) that mentioning harm/benefit issues of disease burden is not essential since these judgments are responding to more variable individual/institutional perceptions and judgments- a potentially slippery slope of "competitions" of who experiences greater or lesser harms/benefits (while extremely personally, culturally, and socially relevant) seems to distract us from the originally stated purposes of the pilot study.

In terms of Point 5 Iagree with Aravinda's point about community consultation. It seems to me that who (or what collection of people) is/are representing the community needs to be described since it still remains unclear to me about what is meant by community consultation in the sense of (a) defining what is community and (b) how can researchers ensure that issues of the full range of community representativeness has been reflected in the consultation). [In fact, this still looms as another issues probably for further discussion at another time.] To Point 6, I also unclear what the criteria for appropriate conditions and conversations are.

On a broader note, I think Jean has raised important points of the practical issues of the pilot and project including the time line and sample availability. Specificially, how are we going to address the tensions between pragmatic issues such as using convenience samples (which may not be representative) and ELSI issues (e.g., equity, representativeness, fairness, etc) with science given the broad and long ranging effects of this projects?

Just a few rambling thoughts.

Best Wishes,
Vivian

****

Vivian Ota Wang, Ph.D.
Assistant Professor, Counseling/Counseling Psychology Programs
Division of Psychology in Education
College of Education
Arizona State University
P.O. Box 870611
Tempe, AZ 85287-0611
Office: 480.727.6933
Fax: 480.965.0300
Email: [redacted]

Guyer, Mark (NHGRI)

From: Troy Duster [redacted]

Sent: Sunday, September 23, 2001 2:03 PM

To: [redacted]

Subject: Re: inf'd consent issues

Once again, thanks to Pilar and Morris for laying out the issues.
My responses are in blue.

After a phone call with the "existing resources" working group, Morris and I agreed to write something about the consent question and circulate it. We would like something to go to Eric L. and Francis sometime this weekend, so if you can quickly chime in on the memo below with comments and questions , please do.

****

Question: If existing samples are used to conduct the pilot study for the haplotype map project, what criteria should be used to evaluate the acceptability of the existing informed consents?

Suggested criteria for evaluating existing informed consents:

1)genetic research should have been specified.

This one is a deal breaker; if there was not consent for genetics research, then we cannot use the samples.
This is perhaps the most straightforward and essential matter, and I think we have consensus here, with "yes"!

2)studies of genetic variation within and between populations should have been specified.

Here I side with what I take to be Aravinda's position, namely, that what needs to be explicit is that the consent address genetic variation, *at least with respect to "existing samples"*. For future studies, I favor some language that captures between-group variation, since that is the knotty issue, with a nod to "the obvious" of within-group variation.

3)there should have been explicit consent for sharing the samples.

This one is a deal breaker; if it does not exist, then we cannot use the samples.

again, looks like a consensus so far of "yes"

4)if we are going to use cell lines, then there should have been explicit consent for the making of cell lines (it seems basic, but I know of samples collected in the 80s that were used to make cell lines without explicit consent for cell lines. The fact that cell lines were made came as a very unwelcome surprise to some of the people who gave the original samples).

yes

5) community consultation

While I understand Aravinda's "problem with who speaks for the entire group", it should be assumed that no group has complete consensus, but that the role of "community consultation" is to explicitly address a range of issues and concerns, and to get a strong sense of the nature of that consensus, or the range of disagreements, and so forth. "Community consultation" is not an attempt to achieve community consent... but a process of engagement and involvement. The issues of borders and group boundaries can be difficult, but that is part of the consultation process.

6) Consent should have been given under appropriate conditions and with appropriate conversations

I do not know how easy or difficult this will be to evaluate. Aravinda has told us about some samples that have been and are being collected in India, for which both individual consent and community consultation/consent exists. These consents were videotaped and consent was specifically given for population variation studies. This sounds like the ideal; however, I suspect that few samples will have documentation beyond a signed form. The question is, what should we accept in terms of evidence that the consent process was valid???

This may seem like a finesse move, but I do believe that this is most likely to be decided best on a local case-by-case determination, with "majority" ELSI input.

The pressure to come up with a group position over this week-end has a double-edge. On the one hand, the complexities are such that we could spend weeks on these matters and probably come to similar conclusions. On the other hand, some have not voiced their concerns in a textured manner that might have influenced how the rest of us respond. We can live with that for this preliminary go at a preliminary response to a pilot project. But it is important for everyone to acknowledge that, as I indicated in the phone conversation, a summary distillation at this juncture could only produce a "flashing

Guyer, Mark (NHGRI)

From: Aravinda Chakravarti [[redacted]du]
Sent: Sunday, September 23, 2001 4:10 PM
To: [redacted]
Subject: informed consent issues

Dear Folks:

Let me clarify my position on the consent and inclusionary issue.

I did not imply that we need community consent everywhere since I do have doubts whether we can reasonably identify the, let alone a, leader. Every population from which a sample is used is a special case. For many, community consultation AND consent will be needed given past history (Native American, Australian aborigines, etc.). For others, such as the (Utah Mormons and the Venezuelan Lake Maricaibo) CEPH samples my understanding is that community consultation and consent was obtained by not always documented on paper. The French CEPH samples were obtained with individual consent and I would presume that a community leader may have been difficult to identify. That is, for some samples, individual consent is all that can reasonably be wanted.

I would be less inclined to ask for within-between comparison language on the consent form. I do not want vague consent forms but genetics is in its essence comparative (i.e., within vs. between tests are all we do) and we cannot define all types of comparative studies we will do with these samples. So, I would ask for language that allows more than population within vs. between comparisons.

My major concern is that we are not being very proactive in inclusion of all groups and being thoughtful on how to sample human diversity. I am not assured that sampling a few well-known named human groups will capture current human diversity. There are many questions about human evolution that extant samples can answer and will answer. I, however, am less (used in a relative sense) interested in that than how current world residents are susceptible or not to different diseases and how extant variation will shape the human disease burden in the future. The haplotype map project, indeed any global human diversity project, is only one crucial step in that direction. I can understand the TSC not willing to delay experiments and needing details of the samples to be used in a pilot study. But this matter is more important than the interests of the funding body. We are always in a rush and never have the time to do things ideally. A compromise would be to design a pan-human DNA collection but proceed with samples that might be assembled (with proper consent) within 6 months. I can agree to this whatever the initial nature of this pilot might be. However, I would be disappointed if we were to drop the ball once the pilot started.

I would like to resurrect the idea of grid sampling of human diversity (standard in all other studies of biodiversity) due to the late Alan Wilson. We should perform a global grid sampling where individuals are asked a limited set of questions (birthplace, birthplace of parents, languages spoken, native language, ethnic affiliations, etc.), and density improving over time, where individuals donate their samples into an International not-for-profit Foundation. The foundation can own rights to these samples and use income for many health-related humanitarian efforts. There are many existing foundations well respected throughout the world that could achieve this. Both consent and intellectual property issues are manageable and enforceable. If such a collection could be created its value to the entire world community would be immense and would propel both the science and medicine forward.

I know we do not have the time, and it might not be the prerogative of this group, to advance this discussion. So, my pragmatic solution is to do a three part solution: start with what can be obtained in 6 months; begin a serious and committed discussion with other groups reluctant to participate and try to incorporate them in the next 12 months starting now; begin a more expanded discussion to create a global sample increasing in prospect over time and using the efforts of the entire NIH. Cheers,

Aravinda
--
Aravinda Chakravarti, Ph.D.
Henry J. Knott Professor & Director
McKusick - Nathans Institute of Genetic Medicine
Johns Hopkins University School of Medicine
600 N. Wolfe Street
Jefferson Street Building, 2-109
Baltimore, MD 21287
tel: (410) 502-7525
fax: (410) 502-7544
e-mail: [redacted]u

Guyer, Mark (NHGRI)

From: Pilar Ossorio [redacted]u]
Sent: Sunday, September 23, 2001 11:55 PM
To: [redacted]
Subject: inf'd consent

Hi All,

Thank you everybody for getting back to us so quickly! As Troy is the Chair for the ethics working group, his summary of the positions is our official one for Jean/Eric/Francis.

I have found this exchange quite interesting, and I have 3 comments for people to dwell on before our next conference call.

First, with respect to the second criterion originally proposed by Morris and me [studies of genetic variation within and between populations should have been specified]... Perhaps for the pilot the best we will be able to do with existing samples is find one's for which consent was given for studies of genetic variation; I can accept that. However, as a general matter, I do not think that it is good enough to discuss "genetic variation" with people when we intend to compare allele frequencies between groups. The reason is because I do not think that most people who are not geneticists will take language such as "genetic variation" to mean that we would be reporting out data comparing allele frequencies of their village/town/tribe/ethnicity/race (whatever) to another village/town/tribe/ethnicity/race. I fall back on the question of what we should be communicating to people so that they will not be surprised, taken aback, embarrassed, or angry by the kinds of research done with their samples. I would put good money down on the bet that most people have little idea of what kinds of experiments might be done under the rubric of genetic variation, and that some people will have concerns about between group comparisons that they would not have about within group comparisons.

Second, with respect to community consultation... Here I am in complete agreement with Troy. Not all community consultations are about consent or consensus. In most cases, I do not think we should be looking for "community representatives" in the sense of individuals who are formally acknowledged or appointed to speak on behalf of a community. Rather we should be looking for people of the sort who will likely be affected by the claims made as a result of the research.

If you find such people and then engage them in community meetings, focus groups and/or interviews, what will likely emerge are themes, considerations, issues and concerns. This is what we want out of such a consultation. The hope is that the themes raised will help improve the research protocols, and allow researchers to minimize harms and maximize benefits by taking account of things we might not otherwise have taken into account. If, during these consultations many people raised very strong concerns, this might create a "yellow light" for researchers about going forward with the project as currently conceived.

Of course, there are communities that have elected or appointed leaders who are designated to speak on behalf of the community. When such people exist, they should of course be consulted and their approval obtained. And when there is a community process for approving research protocols, of course this process must be followed.

Third: With respect to the CEPH samples... It seems to me that those samples have been used for many, many studies, including studies of genetic variation. If I were convinced that donors of those samples were aware of the genetic variation studies, and had no complaints, then I would feel that there was tacit consent for the kinds of studies that will be undertaken in the pilot. I would be convinced of this if, for instance, I knew that donors received newsletters reporting to them some things that had been discovered with the CEPH samples and that none or very few had complained about genetic variation studies.

I say this with some reluctance, however. For the pilot I think it would be OK. In general I do not think that people in the US, or in many other parts of the world, are big fans of "presumed consent," which is essentially what we are doing if we go forward and do studies for which there was no explicit consent, and which people probably did not imagine when they gave the very general consent they did give. In general, I do not think that we should put the burden on the donor to have to come to us with complaints just because we did not do a good enough job of getting informed consent the first time 'round.

Enough blabbing for now. I hope everybody had a good weekend.

Ciao,
Pilar

Guyer, Mark (NHGRI)

From: Morris W. Foster [redacted]
Sent: Monday, September 24, 2001 11:26 AM
To: [redacted])
Subject: Community Consultation

In addition to beginning to specify the minimum standards for informed consent for both the pilot and the full project, we also should begin to specify the minimum standards for community consultation. With the usual disclaimers about thinking out loud, here are some suggestions to begin that conversation:

1) The community being consulted should be that which is being labeled. This is to say, if we're going to label a set of samples as being "Yoruba" then, potentially, everyone who is known as "Yoruba" may potentially be affected by findings reported as such (a population of some 10 million). In fact, though, the actual "Yoruba" samples probably will come from a more delimited local population. If that is the case, then it makes more sense for the samples to be labeled with that locality than as representing Yorubas generally, both from a scientific point of view and as a means to facilitate community consultation. It would be much easier to design a local consultation than one that attempts to sample a cross-section of all Yorubas. Nonetheless, the possibility that findings from a local samples of Yorubas may be extrapolated to all Yorubas does remain, and those implications should be considered as well.

2) The consultation should not be limited just to recognized leaders. It should, in some manner, sample the general or "grass roots" population as well.

3) A successful consultation will be one that discovers concerns that were not anticipated by researchers in advance.

4) A successful consultation will be one that discovers a diversity of viewpoints about genetic research (i.e., both those in favor of it and those opposed to it, as well as positions in between).

5) Community consultation should include some component that provides ongoing community involvement in the research project beyond the collection of samples. That is, results should periodically be reported back to community members and community members should be periodically consulted about the management and uses of their donated samples.

6) Pre-sampling consultation should involve more than one discrete community meeting. For communal decision-making processes to operate, there needs to be some time between the introduction of the project to members and sampling their views. Moreover, internal community discussions often must take place in private (i.e., within the community) rather than in public gatherings with outside researchers present.

7) Individual interviews and randomly recruited focus groups can be used to confirm (as well as augment) findings from community consultations, at least in larger populations where random recruitment can take place.

8) As Troy pointed out, the findings of a consultation are more likely to comprise a diverse range of concerns and viewpoints than a consensus. This means that we need to put some thought into how to weigh those differing

views (with the help of community members) so that we can determine whether it is appropriate to proceed with a specific community.

Again, these are very sketchy suggestions, but a way to begin a more detailed discussion about what we mean when we say "community consultation."

Morris

Guyer, Mark (NHGRI)

From: Royal, Charmaine D. [[redacted]]
Sent: Monday, September 24, 2001 12:31 PM
To: [redacted]
Subject: RE: inf'd consent

Hi everyone,

I'm in complete agreement with Pilar on the issues summarized here, particularly the need for clarity on 'genetic variation' in the consent form. Just to reiterate the points on community consultation/consent. I firmly believe that to the extent possible, ALL 'groups' identified for research on genetic variation, such as is being proposed, should be involved in (or at least invited for) community "consultation", for the reasons articulated by Pilar. Community "consent", on the other hand, is a process required by certain groups and is necessarily commensurate with the "consultation" process for those groups. In my mind, the ever-present question of whether any one person (or subgroup) can speak for the 'group', has always been irrelevant.

Best,
Charmaine

-----Original Message-----
From: Pilar Ossorio [mailto:[redacted]u]
Sent: Sunday, September 23, 2001 11:55 PM
To: Ellen Wright Clayton, MD, JD; McEwen, Jean (NHGRI); 'Pilar Ossorio'; Valle, David (NIDCD); [redacted]; [redacted] Brooks, Lisa (NHGRI); Frampton, Lynn (NHGRI); Guyer, Mark (NHGRI)
Subject: inf'd consent

Hi All,

Thank you everybody for getting back to us so quickly! As Troy is the Chair for the ethics working group, his summary of the positions is our official one for Jean/Eric/Francis.

I have found this exchange quite interesting, and I have 3 comments for people to dwell on before our next conference call.

First, with respect to the second criterion originally proposed by Morris and me [studies of genetic variation within and between populations should have been specified]... Perhaps for the pilot the best we will be able to do with existing samples is find one's for which consent was given for studies of genetic variation; I can accept that. However, as a general matter, I do not think that it is good enough to discuss "genetic variation" with people when we intend to compare allele frequencies between groups. The reason is because I do not think that most people who are not

Guyer, Mark (NHGRI)

From: Vivian Ota Wang [redacted]]
Sent: Monday, September 24, 2001 3:56 PM
To: 'Ellen Wright Clayton'; Aravinda Chakravarti
Cc: [redacted]; Brooks, Lisa (NHGRI); Frampton, Lynn (NHGRI); Guyer, Mark (NHGRI)
Subject: RE: informed consent issues

Dear All,

I fall into line with Aravinda who I think thoughtfully spoke to the ELSI issues of representativeness and the pragmatic realities faced by both the pilot and full study. It seems to me that to be pressed by sources (beit funding, political, etc) the use of convenience samples is more of a practical issue at this time that is competing with the science of what and how to investigate a truly representative pan-human sample. Therefore, it seems to me that if the pilot is to proceed with existing samples, there must first be some honest recognition to the narrow representativeness of the pilot study. Second, a serious discussion is needing to take place regarding how pan-human representation is envisioned to be and how will it be implemented in the full study. I'm fearful that if we don't take the time to be thoughtful at this juncture, our hast may have unforeseen and unintended consequences on both those included and excluded in the sampling. As one way of organizing all of this, I also like Wilson's sampling grid since it provides a nice framework to systematically identify samples so investigators would then know who are and are not represented.

In the meantime, it seems that we are left with the immediate practical matters of the pilot which I think should start with what can be obtained in the 6 month pilot period. However, I do confess that I do find myself in a dilemma - I'm not a big fan of vague language and feel that merely having the words "genetic variation" in the consent form is not enough given we do not know how broadly or narrowly genetic variation was used...(I suspect to the general lay person, the term genetic variation is a bunch of scientific mumbo-jumbo and may not be understood in the manner in which investigators are using it).

Enough for now. Vivian

-----Original Message-----
From: Ellen Wright Clayton [redacted]]
Sent: Monday, September 24, 2001 11:46 AM
To: Aravinda Chakravarti
Cc: McEwen, Jean (NHGRI); 'Pilar Ossorio'; Valle, David (NIDCD); [redacted] Lisa (NHGRI); Frampton, Lynn (NHGRI); Guyer, Mark (NHGRI)
Subject: Re: informed consent issues

Guyer, Mark (NHGRI)

From: Frampton, Lynn (NHGRI)
Sent: Friday, October 05, 2001 3:06 PM
To: [redacted]
Cc: [redacted]
Subject: HapMap Project: Samples Update

To HapMap Working Group Members:

Please review the attached status report of our search for existing samples for use in the pilot study. It has been difficult to find samples where the investigator is willing to share and the consent is appropriate for this type of project (not disease specific). The list is divided into several categories, from those that seem quite possible, to those not available.

If you have any additional suggestions of samples that may be available, please let us know and we will follow-up.
Also, if you see someone at the upcoming ASHG or SNP meetings, who may have samples, have them give us a call.

Here is an outline of what characteristics we are looking for in a population sample:

Populations

We are not looking for indigenous, endangered, or small isolated populations. Majority and urban populations are good, although others may be fine as well. We are particularly looking for samples from Africa and Asia.

Number and type

It would be nice to have samples from 50 individuals, but as few as 20 might be OK.
Trios would be great, but individuals are fine.
Cell lines are great, but if not available then 50 micrograms of DNA will be needed, or perhaps less if the samples are used only to study haplotype variation and not used for testing methods.

Consent

We will need to see the consent form. Even though the consent was fine for the study that was done with the samples, in order for us to use the samples, these issues need to be mentioned in the form or in the consent process. Non-written consent may be OK.

--the samples will be used by other researchers
--the samples will be used for genetic variation research
(as opposed simply to research to study a particular disease).

HapProjects1.xls

Lynn Frampton, MPH
Science Program Analyst
National Human Genome Research Institute
National Institutes of Health
31 Center Drive 31 / B2B07
Bethesda, MD 20892-2033
(301) 496-7531/ (301) 480-2770 (FAX)
[redacted]

PIs	Institution	Project Title or Paper Title	Samples (number, ethnicity)	Cell lines	Families	Available for use by others?	Consent forms
Sample collections - pilot project							
Quite Possible							
CEPH Pedigrees	Coriell		48 Utah (others: 1 Amish, 10 French, 2 Venezuelan pedigrees, but not consent forms)	Yes	Yes	Yes	Received, checking community consultation
Tel Aviv University	Tel Aviv U	http://www.tau.ac.il/medicine/NLGIP/catalog.htm	Moroccan 120 unrelateds or 8 families with 63 individ., Ashkenazi 200 unrelateds or 21 families with 96 individ., all from Israel	Yes, or DNA	Yes	Yes	Fine
Samples Pending Appropriate Consent							
CEPH Panel- HGDP	CEPH		Some samples might be OK	Yes	No	Yes	Checking consent
Coriell - ADA samples	Coriell		For diabetes studies	Yes	Yes		Checking consent
Coriell - TSC pops	Coriell		42 European-American, 42 African-American, 10 Chinese and 32 Japanese	Yes		Yes	Checking consent
European Collection of Cell Cultures	ECACC	http://www.ecacc.org/	Mixed European Caucasian, maybe some others	Yes		Yes	
Kidd, Ken	Yale		recent samples with good consent, ex. Yoruban			Yes, need to think about data release of individuals	
Kittles, Ricky	Howard		75 Liberia	No			Will send consent form
Liu, Ed and David Goldstein	Genome Institute of Singapore		Chinese, Malay, and Indian samples from Singapore			They probably want to do the work themselves.	David Goldstein is checking on the consents.
Nakamura, Yusuke, and Tanaka, Toshihiro	U Tokyo	http://snp.ims.u-tokyo.ac.jp/	48 Japanese individuals; samples would be available depending on amount of DNA needed.	No	No	May be	Received- needs translation into English

PIs	Institution	Project Title or Paper Title	Samples (number, ethnicity)	Cell lines	Families	Available for use by others?	Consent forms
Samples Pending Further Follow-up							
Chakravarti, Aravinda	Hopkins		Finland (non disease)	Yes		Yes	
Hammer, Mike	U Arizona	(2001) Hierarchical patterns of global human Y-chromosome diversity	2,858 males from 50 populations- sub-Saharan Africa, North Africa, Middle East, Europe, South Asia, Central Asia, North Asia, East Asia, Oceania, Americas				
Soodyall, Himla			San and S. African Bantu				
Jin, Li	U Cinncinnati	(2001) African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes	12,127 males from 163 populations- Southeast Asia, Oceania, East Asia, Siberia, Central Asia				
Majumder, Partha			30-50 samples from each of 50 ethnic groups in India. Both tribal (not use) and caste ? How about urban?	DNA			Verbal consent on video tape
Oefner, Peter and Jin, Li	Stanford	(1999) Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations. Proc. Natl. Acad. Sci. 96, 3796-3800.	51 Africa, 96 Europe, 40 Pakistan and India, 59 East Asia, 60 America, 48 Oceania				
Stoneking, Mark	Max-Plank						
Rotimi, Charles	Howard						
Todd, John	Cambridge		US samples from HBDI	Yes	Yes	?	
Wallace, Doug	Emory						
Ward, Ryk	Oxford		urban African				
Samples Not Available							
Abecasis, Gonçalo R, and Cookson	U Michigan (was at Oxford)		European-Australians, British			No, limited DNA	
Chakravarti, Aravinda	Hopkins		Michigan, Zimbabwe, Mennonite			No, only disease consent	
Di Rienzo	U Chicago		Italians, Han Chinese, Cameroon	no	no	No, limited DNA	
Estonian (Andres Metspalu)	Estonia	http://www.genomics.ee/genome/index.html	100 Estonians	No	No	Yes	Not good enough for the pilot projects.
Goldstein, David	Univ College, London		Not enough DNA				

PIs	Institution	Project Title or Paper Title	Samples (number, ethnicity)	Cell lines	Families	Available for use by others?	Consent forms
Groop, Leif	Lund U	Reich et al.	48 Southern Swedes			Probably not but maybe willing to collect new samples.	For diabetes association study
Leonetti, Donna (Melissa Austin) 206-543-6083	U Washington		Japanese-Americans				
McGrath, Barbara Burns	U Washington		Pacific Islanders in the Seattle area.				
Sing, Boerwinkle, Clark, Nickerson	U Michigan		Rochester, Jackson MS			No	
Song, Kyuyoung	U Ulsan, Korea		Korean			No, not good consent. Interested in participating in large-scale studies.	No consents for existing samples.
Samples from Other Sources							
Lander, Eric	Whitehead		Japanese individuals (from Coriell)				
Lander, Eric	Whitehead		Yorubans (from Rotimi, Ward, Cooper)				
Possible sample collections - large-scale project							
CEPH ?			Utah				
Estonia ?			Estonia				
Licinio, Julio	UCLA		Mexian-Americans				
Foster, Morris	U Oklahoma		African-Americans				
Rotimi, Charles	Howard		Africans (Yoruba, others?) Good medical infrastructure.				
Liu, Ed	Genome Institute of Singapore		Chinese, Malay, and southern Indian, in Singapore. Good medical infrastructure.				
Japan ?							

Guyer, Mark (NHGRI)

From: Brooks, Lisa (NHGRI)
Sent: Thursday, September 27, 2001 6:00 PM
To: Collins, Francis (NHGRI); Eric Lander (E-mail); Lai Eric H (E-mail); Nickerson Deborah (E-mail); Peterson, Jane (NHGRI); Schloss, Jeff (NHGRI); Jordan, Elke (NHGRI); Guyer, Mark (NHGRI); Kwok Pui (E-mail); David Bentley (E-mail); Valle, David (NIDCD); Nussbaum, Robert (NHGRI); McEwen, Jean (NHGRI); Frampton, Lynn (NHGRI)
Subject: HapMap: Pilot project implementation

We need to think about how genotyping and haplotyping platforms are going to be compared in the pilot phase and chosen for the production phase.
Bob Nussbaum's group came up with a clear set of criteria for comparing platforms.
They did not want to restrict which technologies to actually test.

Pilot phase:
Should we let researchers and companies propose the platforms to include? Perhaps based on a certain amount of genotyping throughput having been achieved (as opposed to the claims I see in some applications). Or do we let anybody propose to accomplish a certain amount, without our specifying the platform?

Production phase:
Do we envision companies doing some of the production genotyping/haplotyping? Or just academic production labs? Both could respond to an RFA.

We have groups figuring out the experimental design and samples for the pilot phase, but I am worrying about the implementation.
What is the right group to be considering this issue?

Thanks, Lisa.

Lisa D. Brooks, Ph.D.
Program Director [redacted]
Genetic Variation Program
Genome Informatics Program
National Human Genome Research Inst.
National Institutes of Health
31 Center Dr. 31 / B2B07 301-435-5544
Bethesda, MD 20892-2033 301-480-2770 fax

Guyer, Mark (NHGRI)

From: David Bentley [redacted]
Sent: Friday, September 28, 2001 4:13 AM
To: Brooks, Lisa (NHGRI); Collins, Francis (NHGRI); Eric Lander (E-mail); Lai Eric H (E-mail); Nickerson Deborah (E-mail); Peterson, Jane (NHGRI); Schloss, Jeff (NHGRI); Jordan, Elke (NHGRI); Guyer, Mark (NHGRI); Kwok Pui (E-mail); Valle, David (NIDCD); Nussbaum, Robert (NHGRI); McEwen, Jean (NHGRI); Frampton, Lynn (NHGRI)
Subject: Re: HapMap: Pilot project implementation

At 10:59 PM 09/27/01, Brooks, Lisa (NHGRI) wrote:
>We need to think about how genotyping and haplotyping platforms are going to
>be compared in the pilot phase and chosen for the production phase.
>Bob Nussbaum's group came up with a clear set of criteria for comparing
>platforms.
>They did not want to restrict which technologies to actually test.
>
>Pilot phase:
>Should we let researchers and companies propose the platforms to include?
>Perhaps based on a certain amount of genotyping throughput having been
>achieved (as opposed to the claims I see in some applications).
>Or do we let anybody propose to accomplish a certain amount, without our
>specifying the platform?

I would suggest that any proposal to carry out a pilot should be able to state which platform will be used, and that it should carry some brief summary of experience to date. This would help ensure realistic claims based on that experience, and consequently minimise the time required for a ramp-up. I imagine that both these factors will be important to achieve a rapid and timely response to this initiative which would be valuable for its success and for the impression it creates to others.

>Production phase:
>Do we envision companies doing some of the production
>genotyping/haplotyping? Or just academic production labs? Both could
>respond to an RFA.
I imagine any company must be able to come to the terms of data release, and maintain competitive costs preferably without subsidy (no loss-leaders here?) - just thinking aloud here.

>We have groups figuring out the experimental design and samples for the
>pilot phase, but I am worrying about the implementation.
>What is the right group to be considering this issue?
>
Are you referring to implementation of the pilot, or of the main project?

David Bentley

Guyer, Mark (NHGRI)

From: Robert Nussbaum [[redacted]]
Sent: Friday, September 28, 2001 8:50 AM
To: David Bentley; Brooks, Lisa (NHGRI); Collins, Francis (NHGRI); Eric Lander (E-mail); Lai Eric H (E-mail); Peterson, Jane (NHGRI); Schloss, Jeff (NHGRI); Jordan, Elke (NHGRI); Guyer, Mark (NHGRI); Kwok Pui (E-mail); Valle, David (NIDCD); Nussbaum, Robert (NHGRI); McEwen, Jean (NHGRI); Frampton, Lynn (NHGRI); [redacted]
Subject: Re: HapMap: Pilot project implementation

>I agree with David that any group applying to do the pilot would
>need to state which platforrm and what their previous experience
>with such a platform is - that goes without saying, since we would
>require that the group to successfully convert (sorry for split
>infinitive) 1500 FASTA sequences containing SNPs into about 1000
>assays and then genotype approximately 90 people (96,000 genotypes
>total with blind duplicates) IN FOUR WEEKS. NO one can ramp up and
>do that - they have to be already ramped up.

I see no particular reason to restrict to academic versus commercial, etc. We want the best platforms, with proven track records in doing high throughput genotyping, competing with each other. We consciously chose NOT to specify which platforms. I think I am correct is saying that the Lab Methods subgroup did not feel that there is currently a consensus as to a platform that is a clear winner that combines the best of THROUGHPUT, ACCURACY, and COST, although some members of the subgroup have extensive experience with invader technology and think very highly of it, others have had experience with single-base extensions/MALDI Mass Spec and see that it has advantages. I am sure there are other technologies that the subgroup has not had as much personal, hands-on experience with that also have their adherents and admirers, The point of the pilot exercise is: anyone who can make a compelling argument that their previous experience and track record would qualify them to compete to carry out the stringent pilot we have proposed, in the brief time allotted, should have a crack at it. Transparency, however, is essential, i.e. full disclosure of software used to design the assays and organize the throughput pipeline, full disclosure of raw and finished genotyping data, and full disclosure of the cost of designing each assay and the cost of genotyping real samples.

bob nuss

>. At 10:59 PM 09/27/01, Brooks, Lisa (NHGRI) wrote:
>>We need to think about how genotyping and haplotyping platforms are going to
>>be compared in the pilot phase and chosen for the production phase.
>>Bob Nussbaum's group came up with a clear set of criteria for comparing
>>platforms.
>>They did not want to restrict which technologies to actually test.
>>
>>Pilot phase:
>>Should we let researchers and companies propose the platforms to include?
>>Perhaps based on a certain amount of genotyping throughput having been
>>achieved (as opposed to the claims I see in some applications).
>>Or do we let anybody propose to accomplish a certain amount, without our

Guyer, Mark (NHGRI)

From: Pui-Yan Kwok [redacted]u]
Sent: Friday, September 28, 2001 12:33 PM
To: Brooks, Lisa (NHGRI); Collins, Francis (NHGRI); Eric Lander (E-mail); Lai Eric H (E-mail); Nickerson Deborah (E-mail); Peterson, Jane (NHGRI); Schloss, Jeff (NHGRI); Jordan, Elke (NHGRI); Guyer, Mark (NHGRI); David Bentley (E-mail); Valle, David (NIDCD); Nussbaum, Robert (NHGRI); McEwen, Jean (NHGRI); Frampton, Lynn (NHGRI)
Subject: Re: HapMap: Pilot project implementation

I agree with David and Bob in that both academic and commercial groups should participate in the pilot study to assess the platforms' performance in a realistic test. Cost accounting is going to be tricky because companies and large groups already have the infrastructure and may have an edge over other groups. One way forward is to place a cap on the amount of money each group will get for the project, with the amount set by the HapMap Group 2 that includes everything from assay development to personnel to reagent cost but exclude instrumentation (say $50-60 per marker). Here we are talking about $50K to $60K for each pilot. The groups coming forward for the test will then either use existing equipment in the lab or negotiate with the platform company or the NIH for instrument support. If a company wants to subsidize the project, that's fine as long as they'll subsidize the main project in the future (if we are taking contributions from the TSC members, there is nothing wrong if the genotyping companies want to contribute too). What we don't want is companies using the pilot study as a publicity thing but do not perform or contribute to the main project.

I agree that as long as the groups (commercial or academic) comply with the free and immediate data release policy set by the project, anyone with the capacity and can stay WITHIN THE COST STRUCTURE should be included. If we are doing 2 populations of 150 samples each (50 trios) at a 20 kb resolution, we are talking about 150,000 SNPs typed on 300 samples = 45 million genotypes. At 96,000 genotypes per month (as the pilot project throughput requires), we are talking about 39 years for the winning platform when it is done by one group. In reality, when we fold in the work to identify the common SNPs (minor allele frequency >0.2) and to deal with regions requiring even better resolution, it will be many more SNPs and genotypes before we are done. So I don't think that it's going to be a winner-takes-all situation and I am pretty sure that the main project will be done by several groups with more than one platform.

Pui
--
Pui-Yan Kwok, M.D., Ph.D.
Washington University
660 S. Euclid Ave, Box 8123
St. Louis, MO 63110
Voice: [redacted]
Fax: 314-362-8159
Mobile: [redacted]
Email: [redacted]

> We need to think about how genotyping and haplotyping platforms are going to
> be compared in the pilot phase and chosen for the production phase.
> Bob Nussbaum's group came up with a clear set of criteria for comparing
> platforms.
> They did not want to restrict which technologies to actually test.
>
> Pilot phase:

Haplotype Map Project Working Group Reports Draft 9/25/01

Group 1: Methods: Experimental Design

Primary Aims

1) To provide a comprehensive characterization of human genome sequence variation to guide population-based association studies of disease and drug response.

2) To provide a reference set of common SNPs that are freely available for performing efficient and powerful association studies on a genome-wide scale.

Strategy

Where likely functional variants can be identified based on sequence context (for example, by altering the amino acid sequence of a protein), the most statistically powerful approach is to discover and directly test such variants for association to disease or drug response. However, only 2% of the human genome sequence is occupied by coding regions, and the remaining 98% will certainly harbour a substantial proportion of functional (presumably regulatory) mutations. Given this distribution of putative functional mutations, we propose a two-pronged approach as follows.

1. Genomic Characterization: Identification of conserved haplotype blocks throughout the human genome and a reference set of SNPs that define these blocks for genome-wide studies.
2. Gene Characterization: Identification of putative functional variants and other SNPs by focused resequencing of gene regions.

These two steps will then be integrated into a single map:

3. Definition of a comprehensive "human SNP set": Based on (1) and (2), selection of a minimal set of maximally informative SNPs for genome-wide scans.

Specific Goals

1) Map haplotype "blocks" using a hierarchical approach. This will involve genotyping of previously discovered SNPs (for example, from the TSC and BAC overlap projects) to define regions of high to medium intermarker association, and to outline the frequent haplotypes within each such "block". Initial typing will be at a low density [≈1 polymorphic (m.a.f. >0.2) SNP every 20kb] to identify blocks of 60kb or greater in size. Those regions not yet falling into "blocks" will be analyzed by genotyping an increasingly higher density of markers in an iterative manner. Achievement of this goal will require:

   a) Guidelines for automated definition, detection, and refinement of haplotype 'blocks'. Key questions include the variability among genomic regions, ultimate marker density, relationship to SNP allele frequencies, and sampling design (phase-known or unknown).
   b) Assessment of sequence diversity in regions that do not fall into blocks. Key questions include how much of the genome is contained within blocks (of a given size or greater) and how to capture haplotype variation in regions that do not fall into blocks.

2) Identify a comprehensive collection of putative functional variants by targeted re-sequencing within the coding regions of genes for integration with the genomic map.

3) Select a maximally efficient and comprehensive set of SNPs for association studies across any genome region.

4) Develop a mechanism and structures for data dissemination and presentation to facilitate usage of the haplotype map in disease association studies.

Pilot Studies

In order to realistically define the scope, scale and cost of the study, we need further understanding of the nature of haplotype blocks in terms of the distributions of block length and diversity, variability within and between regions, conservation across populations, and degree of genomic coverage for given marker and gene densities. These factors require careful consideration of pilot studies, which will involve re-sequencing of specific regions and extensive genotyping of the variants therein to guide specific aims 2-4 above. We have constructed a detailed outline of the overall pilot model, comparing what we know with what we do not know but need to know to meet the primary aims above. We will use the next month to refine the details of this general perspective. A draft outline of pilot study is appended below.

Assumptions and Issues:

- Assumption: 5% haplotype frequency is minimal of practical interest
- Assumption: There exist blocks of conserved haplotypes in humans. Issue: We do not know how best to define them.
- Issue (critical): There is an immense need for new statistical approaches, algorithms and software applications for automated genotyping, haplotype definition, block assessment, instrumentation support, and a number of other areas. Automated genotyping is particularly crucial in this regard. Unless addressed directly and individually, these alone could prevent achievement of the broadest aims of the entire mapping project.

Group 1: Methods: Experimental Design: Pilot Studies Outline

Established data:

Our group agreed on the following empirical observations (and unanswered questions):

1) Putative functional variants (missense cSNPs)
    a) The typical gene contains 2-3 missense SNPs (> 1% frequency[1-3]).
        i) Of these, fewer than half [2] have allele frequencies >5%.
        ii) Thus, the universe of missense cSNPs (>5% frequency) is $\lesssim$1 per gene.
    b) There exist robust protocols for PCR resequencing (by Sanger chemistry) in individual samples [1,2,4-6].
        i) Human review of trace data is required to discover heterozygous SNPs with high sensitivity and specificity
            (1) Genotyping by sequencing (for haplotype analysis) requires, at present, human review of each trace.
            (2) This would be a major bottleneck in a high-throughput project.
        ii) Improvements in automated analysis of sequence data would be required.
        iii) The relationship between coverage in sequencing and sensitivity in SNP discovery is not well established by most users, and requires use of second method to determine false-negative rates.
            (1) At Whitehead, DA reports something like 75% sensitivity with a single-pass of sequencing (using PolyPhred), increasing to 85% if forward and reverse sequencing is used. To get sensitivity >90% requires design of overlapping PCR assays, since there are end-effects and template-specific secondary structure that cause regions of low quality on both strands. (These comparisons are made using comparison to VDA[2], or by multi-fold resequencing of a given gene (unpublished).) We await comparable data from Debby Nickerson and others.
        iv) Other methods (VDA "chips" [2,3,7-9]) exist and should be considered.

2) Analysis of ancestral segments
   a) There exist regions in the human genome that have experienced low rates of historical recombination. In such regions, haplotype diversity is often modest, allowing identification of ancestral segments that are shared with highly conserved sequence in a substantial proportion of the population. SNPs can be selected that "tag" these ancestral haplotypes, summarizing variation across the region in an efficient manner.

      However, there is as yet no consensus on:
      i) How to define these "blocks"
      ii) The genome-wide distribution of block sizes
      iii) The haplotype diversity of such blocks
      iv) How well haplotype diversity can be defined by typing a subset of common SNPs across each block
      v) How to set objective thresholds for when a "region of low recombination" has been delimited
      vi) How to set objective thresholds for when the haplotypes in a block have been defined.
      vii) How to select the most efficient set of maximally informative SNPs for subsequent study
      viii) How i-vii might vary across populations.

Proposed pilots (designed to address the questions outlined above).

1) Putative functional variants (missense cSNPs)
   a) We do not feel that additional pilots are needed to determine the spectrum of amino-acid altering SNPs in human genes.
   b) Useful pilots might address:
      i) Improved methods for automated detection and scoring of heterozygous SNPs
      ii) Detection of SNPs in pooled DNA, requiring fewer sequencing lanes per gene
   c) As discussed by the lab methods group, demonstrated scale, LIMS and quality control would be required of any laboratories proposing to perform this work.
2) Analysis of ancestral segments
   a) How to define blocks?
      i) a proposed definition of a block: a segment of the human genome that has been robustly determined to have experienced low rates of historical recombination, such that ancestral segments are inherited unmodified by recombination from common ancestors.
      ii) This is an analytic question, requiring evaluation of multiple definitions in representative empirical datasets. We suggest that these efforts begin with existing resequencing and genotyping data, and then apply these tools to additional empirical datasets generated during the pilot phase (below). Key issues include finding definitions that are robust and stable, and that correspond to regions with strong allelic associations for association studies.
   b) Genome-wide distribution of block sizes.
      i) There is a significant amount of genotyping data that can be used to measure the large-scale structure (>10kb or so). These data should be analyzed in a consistent way across multiple datasets, and a representative distribution confirmed.
      ii) For fine-scale (<10kb), dense analysis of contiguous regions is required. Due to directed resequencing studies and inhomogeneity in the map density of genotyping studies, there may already be considerable data at a fine scale that can be extracted and compared across laboratories.
   c) Diversity of haplotypes within regions of low historical recombination. (It may be less meaningful to speak of "haplotype diversity" when there is evident recombination across the region, since this confounds the diversity of ancestral segments with that due to recombination. Further, if the region becomes long enough, all "haplotypes" become unique. Thus, haplotype diversity can be discussed

in the context of regions without evident recombination.)

i) The only way to fully define the haplotype structure of a region is to completely resequence it in a sample of adequate size. Additional data of this sort will clearly be needed, and can be obtained through two approaches:

(1) Complete resequencing of a contiguous genomic region. This can be comprehensive, but is likely inefficient (given the existence of blocks of significant size) and expensive. In addition, this would not (unless linked to additional genotyping of regions) address the technical and substantive issues regarding the genotyping approach using TSC and overlap SNPs.

(2) Targeted resequencing of blocks (and inter-block regions) defined by previous genotyping studies. This should answer the same question, and can be designed to check the inferences about historical recombination and haplotype diversity obtained through existing genotyping approaches.

(a) Such a targeted study should resequence segments of adequate length to robustly define all haplotypes present at 5% frequency (within "blocks"), as well as confirming the block structure, and evaluating regions that fall between previously defined blocks.

(3) Either approach should analyze a single set of regions by both approaches (resequencing and genotyping) so that the results and costs can be directly compared. The number of regions to be examined is not yet known, but expected to be on the order of ~20. There are additional tradeoffs between the scope of resequencing and the number of populations to be studied and the sampling design (phase known vs unknown) which require further evaluation.

d) Objective thresholds and selection of "maximally informative" SNPs

i) These analytic pilots will use existing data and that developed in c) above. Participants should propose methods and strategies to develop automated tools for these purposes, and validating the proposed methods.

e) Populations.

i) Each pilot will need to be undertaken for multiple populations, since the performance of any given approach or analytic method will vary based on the underlying haplotype patterns.

References

1. Cambien, F. et al. Sequence Diversity in 36 Candidate Genes for Cardiovascular Disorders. *Am J Hum Genet* **65**, 183-191 (1999).

2. Cargill, M. et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* **22**, 231-8 (1999).

3. Halushka, M. K. et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet* **22**, 239-47 (1999).

4. Nickerson, D. A. et al. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene [see comments]. *Nat Genet* **19**, 233-40 (1998).

5. Nickerson, D. A., Tobe, V. O. & Taylor, S. L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res* **25**, 2745-51 (1997).

6. Stephens, J. C. et al. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489-93. (2001).

7. Chee, M. et al. Accessing genetic information with high-density DNA arrays. *Science* **274**, 610-4 (1996).

8. Hacia, J. G., Brody, L. C., Chee, M. S., Fodor, S. P. & Collins, F. S. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis [see comments]. *Nat Genet* **14**, 441-7 (1996).

9. Wang, D. G. et al. Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome. *Science* **280**, 1077-82 (1998).

Group 2: Methods: Laboratory Methods

Executive Summary

The working group recommends that a minimum of ~96,000 SNP genotypes be generated in a pilot, using ~96 DNA samples (of which 4-6 are blind duplicates) and 1000 SNP assays. The coordinating group for the pilot studies should make 1200 SNPs sequences available as FASTA files for each group to use to design a minimum of 1000 assays. The Laboratory Information Management System (LIMS) needs to be available to other laboratories and its capabilities clearly defined and tested under high throughput conditions. Raw genotyping data should be available for review by TSC coordinators and groups participating in the pilot testing of methods. Costs need to be clearly and accurately reported, both per assay developed and then per genotype performed on an established assay. Accuracy should be determined at a minimum by comparing blind duplicates and comparison of genotyping results for the same SNP between platforms, although additional methods (Mendel-checking) may be appropriate depending on the structure of the DNA samples. Throughput should be measured by the amount of time it takes to develop and genotype – a total of 4 weeks should be set as the maximal amount of time a group can take to complete the pilot.

The critical issues are:

1. Choice of SNPs to be used for assay design and pilot genotyping
2. Choice of DNA samples to be used that would best allow assessment of accuracy
3. Efficiency of assay design
4. Robustness of Laboratory Information Management System (LIMS)
5. Cost
6. per assay developed
7. per genotype once assay developed
8. Accuracy
9. Throughput

1. Choice of SNPS

A set of 1200-1500 SNPs should be identified that meet the minimum criterion of having been discovered independently at least twice. Vanishingly rare variants are unlikely to be part of a set that has been independently discovered at least twice. .Dr. Nakamura has offered to provide sequence information for 35,000 SNPs discovered through re-sequencing of an ~110-Mb genomic region that were identical to SNPs that had already been discovered by others and deposited in NCBI. It is important to characterize each platform based on how well it can turn uncharacterized SNPs into real assays rather than use SNPs that have already been converted into working assays.

TSC should provide FASTA files containing a few hundred bases around the SNP for these 1200-1500 SNP loci to participants in the pilot studies. It is advisable for the choice of SNPs to be coordinated with the Experimental design working subgroup in case they want certain types (cDNA versus genomic) or distribution (clusters or evenly distributed) of SNPs they would like tested.

2. Choice of DNA samples

The primary decision here is whether trios will be used, which would allow additional error checking, and/or if the samples will be ethnically diverse. The Methods Working subgroup awaits recommendations of the others subgroups on this topic. For reach plate of DNAs, however, 6 blind duplicates should be included. In addition, if samples are from families (trios, for examples), genotyping labs need to be blinded to family structure.

3. Assay Design

Participating labs should report their efficiency of assay design. Which SNPs yielded robust assays for which platforms? Are design issues dependent or independent of genotyping platform?

Please note that no particular genotyping method is being proposed for pilot testing. There are many methods, some based on hybridization, others on enzymatic mismatch detection, and still others using primer extension, as well as, perhaps, others still under development.

4. LIMS

The capacity of any LIMS system needs to be clearly spelled out and its ability to perform the pilot study assured. Can it handle both the assay design and genotyping phases of the pilot? LIMS needs to be exportable and available to other laboratories interested in using a particular SNP method.

5. Cost

The costs should be all-inclusive and broken down into two categories: cost for assay set up and cost per genotype in the production setting. Cost for assay set up will include the cost for the design and synthesis of oligos, or other SNP-specific reagents. Assay set up would also include any steps taken to validate the assay before production genotyping (i.e., genotyping DNA pools or a very small number of samples initially). This is critical since in large-scale genotyping some percentage of assays are lost to failure, and the cost of oligo design could be an important "sunk cost". Cost per genotype in the production setting should be broken down into:

hardware (assume 3-year amortization)
the number of people required for daily production genotyping
reagents and disposables
licensing and software

6. Accuracy

Accuracy should be calculated for each genotyping pilot by comparing the blind duplicate error rate. An error rate less than 0.1-0.2% is to be expected for a reasonably accurate system. Genotypes for the same individuals at the same SNP should be compared between all groups using different platforms to provide additional, critical information on accuracy

Raw genotyping data may come in many forms, i.e. optical or fluorescent measurements, mass sizes, etc. In any case, raw genotyping results should be available for review by TSC coordinators and groups participating in the pilot testing of methods.

Finally analysis of accuracy should include a within-platform measurement of reproducibility in blind duplicates as well as between-platform comparisons of genotypes performed on the same DNA samples but by different methods.

7. Throughput

Four weeks should be allotted for assay design and genotyping the entire set of DNA samples. Such a tight time schedule will require that laboratories have already established pipelines for high throughput assay design and genotyping.

Group 3: Methods: Samples

Key Questions:

1. What samples should be studied in the long term? Number? Type?
2. What are the tradeoffs in terms of statistical power, logistics, etc.
3. What samples should be used in the short term ? What samples should be used to gain information about experimental design issues and population selection?

Main Conclusions:

1. Pilot studies should ideally study 50 trios per population. Sample collections for the main project should also target 50 trios. If the analysis of the pilot show that singletons are sufficient to generate accurate haplotypes, then we can drop the offspring DNA and save on genotyping costs.
2. More comparisons of trios versus singletons are needed.
3. Conversions to haploid cell lines should be done, followed by extensive validation of the platform before considering the creation of large collections.
4. An Analysis subgroup should be created in the next phase of the project for a more detailed discussion on analytical tools that need to be generated for the pilot and main project.

General Issues:

Minimum haplotype frequency: The project must reliably detect 99% of haplotypes with frequencies of 5%, in each sample population. The choice of this arbitrary cutoff was made in order to assure that future genetic studies of common variants will include functional SNPs of frequency greater than 5%.

Number of chromosomes to sample to detect haplotypes: The number of chromosomes to sample will determine the probability that a given haplotype will be sampled. For 99% probability of sampling a haplotype of 1, 5 and 10%, we should need about 458, 90 and 44 chromosomes. Appendix 1 includes tables of probabilities of sampling at least N copies of a particular haplotype for haplotypes of difference frequencies.

Number of chromosomes to estimate haplotype frequency: Sample size requirements are considerably larger for accurately estimating haplotype frequencies versus simply detecting a haplotype. Appendix 1 includes a table with estimates of population allele frequency with sample sizes of 100, 200 and 500 chromosomes.

Haplotype accuracy: The sampling methods have a huge impact on the accuracy of the haplotypes and their frequency estimates. Mark Daly has provided the group with the results of computational analyses comparing TRIO and SINGLETON analyses for 5 SNP haplotypes using the EM algorithm on simulated and real data [See appendix 1]. Briefly, the results are fairly consistent with the Fallin and Shork paper in that even low frequency haplotypes were detected almost as well with singletons as well as trios. There is however a shared concern among the group that the singleton samples could fare less well if the LD between markers is low (see analyses by Goncalo Abecasis in Appendix 1), if haplotypes with 10 or more SNPs are reconstructed, and if genotyping errors are present.

Haploid cell lines should allow haplotypes for every sampled chromosome to be deduced.

Genotyping errors can cause false haplotypes to be inferred as well as true haplotypes to be missed. Quality control measures will be essential in the pilot and main projects.

Number of populations to sample: The sample group is aware that the population groups have suggested that the pilot project should attempt to oversample populations – and that a target of 12 populations is being considered. Although this number is somewhat larger than we discussed in our preliminary discussions (ie 4 – 8), we agree that the pilot should gather as much empirical data as possible, not only to determine the optimal number of populations, but also guide the choice of populations for the main project. The pilot project in several populations could serve to determine the usefulness of ethnic identifiers for the main project and in future genomic studies.

Use of isolated populations: There was a consensus that (although interesting), the SIMMAP project should NOT include populations with recent bottlenecks, as this could be less informative for the determination of the common haplotypes shared across populations.

Relative cost of sample preparation: The cost of setting up the resources (ie. trios, singletons or haploids) to be genotyped is likely to be so outweighed by the genotyping cost as to be largely ignorable. [Other issues such as consent, time constraints, etc. are more critical than cost].

Ratio of chromosomes tested per DNA sample genotyped.

Singletons:	1 sample = 2 chromosomes tested
Trios:	1 sample = 1.33 chromosomes tested.
Haploids:	1 sample = 0.67 chromosomes tested.

I. TRIOS

Pros:

- Partial phasing allows more reliable inference of haplotypes in trio family than with unrelated individuals, particularly for rarer haplotypes.
- Ability to detect a subset of genotyping errors via Mendelian checks. (The percentage can often be as low as 13% for even four-person nuclear families with common SNPs, i.e., frequency greater than 20%.) Models of error rates that account for allele frequency and the rate of Mendelian errors could be included in the project, and used as a way to monitor the quality of the genotyping data.
- Allows the flexibility of analyzing the trios or just the unrelated parents.

Cons:

- Harder to recruit trios than singletons.
- Compared to singletons, 3 DNA samples need to be genotyped to obtain information on 4 (parental) chromosomes: 1 sample = 1.33 chromosomes tested.

How many would we need?

- 50 trios (per population) is currently proposed for the pilot project. 200 chromosomes would be sampled, giving a 99% probability that 5% haplotypes will be sampled at least 4 times. [33 trios would allow 5% haplotypes to be sampled at least 2 times. The 50 trios in the pilot would allow a more thorough comparison of Trios versus Singletons (the offspring could always be dropped in the main project if the value of that extra DNA is shown to be negligible). Perhaps fewer trios would be sufficient for the main project, but again, this needs to be tested in the pilot project.

What exists for short term tests?

- Lisa Brooks and Lynn Frampton are gathering this information. The table is being updated as this early report is being generated. The current list suggests that there may not be 50 trios available in many populations for the pilot project. The committee will discuss suitable choices after seeing the reports of the other subgroups and have an opportunity to discuss with the entire SIMMAP group.

- CEPH trios are suggested as a useful and practical choice for one of the pilot samples.

II. RANDOM SINGLETON

Pros:

- Ease of collection of samples.
- More efficient than trios and hybrids in terms of the chromosome number that is tested per sample: 1 sample = 2 chromosomes tested.

Cons:

- Haplotype frequencies are estimated for the group of individuals as opposed to determined for each individual.
- Inability to detect genotyping errors using Mendelian checks.
- The statistical inference of haplotypes may miss rare haplotypes (although this can also be due to a sampling problem - which will affect all methods
- The statistical inference of haplotypes may cause errors in haplotype determination if the sample is not in Hardy-Weinberg Equilibrium.

How many would we need?

- 100 singletons is the number currently being discussed. This is 200 chromosomes and equivalent to 50 trios. In the pilot experiment, the parental chromosomes will be studied independently from the offspring, to represent the singleton scenario.

What exists for short term tests?

- Lisa Brooks and Lynn Frampton are gathering this information.

III. HAPLOID CELL LINES

Pros:

- High accuracy of haplotype, since the haplotype is directly constructed from the genotypes in the haploid cells. Particular advantage for rarer haplotypes, but also more generally true. Long-range haplotypes can be obtained.

Cons:

- The use of somatic cell hybrids for haplotype construction has not been tested on the scale being considered here. A pilot study is a must if this avenue is to be pursued.
- Lack of an adequate resource for short-term pilots. [see below].
- Some chromosome rearrangements (including deletions) reported.
- Among cell lines containing a given chromosome, the relative amount of the human chromosome will vary, increasing the chance of equivocal genotypes.
- Need to construct and characterize somatic cell hybrids from every participant - prior to the genotyping phase of the project: This is a minor point, as the cost of sample preparation is small compared to the genotyping costs for the project.
- Compared to singletons, 3 DNA samples need to be genotyped to obtain information on 2 (donor) chromosomes: 1 sample = 0.67 chromosomes tested. [The 3 DNAs are the donor DNA and 2 hybrids characterized as containing the chromosome to be tested]

What exists for short term tests?

- Nick Papadopolous was contacted by Mike Boehnke. 2-3 months would likely allow enough time to do conversions, get DNA, do some QC genotyping on approximately 50 samples. Less than that is likely not possible.
- Pui Kwok has also been in contact with GMP – and sent a message that March 2002 is a likely timeline to get sufficient number of cell lines. Pui also suggested that Jeff Trent and Bert Vogelstein should be contacted.
- The subgroup suggests that the cell lines from CEPH trios be used for this pilot, as it will allow direct comparisons between trios, singletons and haploids.

How would we establish a more appropriate collection within 4-6 months?

- More pilot work is needed before considering the creation of large collections from different populations.

Group 3: Methods: Samples: Appendix 1. Preliminary simulations and data analysis

Sections:

A. Haplotype reconstruction issues (Mark Daly and Steve Schaffner)

B. Probability of sampling at least N copies of a particular haplotype for haplotypes of difference frequencies (Goncalo Abecasis)

C. Haplotype reconstructions in the absence of disequilibrium: Trios versus Singletons (Goncalo Abecasis)

A. Haplotype reconstruction (Mark Daly and Steve Schaffner)

We've begun several sets of analyses and simulations in an attempt to strengthen our understanding in the following areas: understanding the differences between a SINGLETON approach, a TRIO approach, and perfect PHASE knowledge. In all cases describing here, haplotype frequencies are computed for both TRIO and SINGLETON approaches using the EM algorithm. The simulations and real data analysis are both done assuming trios (2 parents and an offspring) are typed in an attempt to recover the four phased parental haplotypes. SINGLETON analysis is done on the parental genotypes alone and for TRIO analysis, phase is inferred using the offspring before the call to EM to resolve any remaining ambiguity.

(with the TRIO approach, as much phase as possible is determined from the offspring)

Question 1 -

How well does a sample of a given size estimate the population allele frequency:

		s.d. given number of chromosomes		
		100	200	500
f	25%	4.3%	3.1%	1.9%
r	10%	3.0%	2.1%	1.3%
e	5%	2.2%	1.5%	1.0%
q	1%	1.0%	0.7%	0.4%

(remember 25 trios = 100 chromosomes)

These numbers are simply s.d. = sqrt(f*(1-f)/N) but the last is better addressed through Poisson - with only 100 samples, we fail to sample a 1% chromosome 36.7% of the time, with 200 we still miss it 13.5% of the time. This puts a strong bound on how well we can detect chromosomes that exist at low frequency in only one population. This is trivial background information but important for two reasons:

1) it places a bound on how well we can estimate data with perfect phase information and gives us a yardstick by which to compare the additional variance added by NOT having perfect phase information
2) it's necessary to understand how limited we're going to be to make statements about 1% haplotypes and as such may focus our approach (as we discussed over the phone) on chromosomes that either:
   a) exist at 5% or higher in a single population -OR-
   b) exist at 1 or 2% in the overall sample

Question 2 –

In practice how well can we detect rare chromosomes (~1%)? I've set up simulations in which haplotypes of 5 SNPs are drawn from a region containing 4 common haplotypes with a 5th at 1% and inquire of each replicate whether that 5th haplotype was detected and how accurately its frequency was estimated in the SAMPLE (perfect phase information), TRIOS, and SINGLETON design:

25 trios (100 chromosomes) – 10,000 replicates

	SAMPLE	TRIO	SINGLETON
% of times 1% haplotype detected	63.8	60.8	52.3
s.d. of estimated frequency	.0099	.0100	.0100
% additional variance compared to SAMPLE	-	3.8%	20.6%

50 trios (200 chromosomes) – 2,500 replicates

	SAMPLE	TRIO	SINGLETON
% of times 1% haplotype detected	87.8	85.1	76.5
s.d. of estimated frequency	.0070	.0071	.0074
% additional variance compared to SAMPLE	-	1.8%	16.0%

So despite needing larger samples to reliably detect these haplotypes (shown in Tom's earlier message and the earlier charts), we actually do a reasonable job on average of estimating their frequency even in SINGLETON scenarios. Assumptions relevant to these simulations:

1) In these simulations, the rare chromosome always resembled a common chromosome except at one SNP position. This is largely based on the observations I've made from real data so I believe it to be the most common scenario (and for what it's worth these could be the result of gene conversion or undetected genotyping error). When I changed the simulations to make the 1% haplotype differ from the most common chromosomes at two or more sites, all estimations improved so the tables above represent an upper bound with respect to the composition of the rare chromosome.
2) Since, as reported in Fallin/Schork and elsewhere, EM reconstruction gradually degrades as the number of haplotypes in the population increases, I also tried increasing the number of common chromosomes but have seen no increase in any of the above numbers in a few limited attempts. Even in the most unfavorable models in that paper (5 equifrequent SNPs in total equilibrium), the EM reconstruction from straight genotype data introduces less variance than the sampling itself.

Question 3 –

How well do we estimate the frequency of common haplotypes in situations in which offspring are used to determine phase (TRIOS) versus if they are not (SINGLETONS). To address this we've examined data from 24 CEPH grandparents (96 chromosomes) that have been typed for a high density of markers in ~40 regions (150 kb each) of the genome and comparing the frequency estimate for various haplotype frequencies with and without use of a single offspring for phase inference.

The data are broken down into bins of haplotype frequency (as measured with phase information), and calculated for each haplotype is the difference between the frequency calculated with phase and without phase info. Shown are the mean and variance of this difference:

```
2.5% - 7.5% n: 58 mean: 0.0027 var: 0.00020
7.5% - 12.5% n: 26 mean: -0.0037 var: 0.00045
12.5% - 17.5% n: 11 mean: -0.0001 var: 0.00070
17.5% - 22.5% n: 6 mean: 0.0063 var: 0.00005
25.0% - 35.0% n: 8 mean: -0.0039 var: 0.00025
35.0% - 45.0% n: 7 mean: -0.0010 var: 0.00015
45.0% - 55.0% n: 5 mean: 0.0013 var: 0.00005
55.0% - 65.0% n: 0
65.0% - 75.0% n: 0
```

Notes: This only include haplotypes that were identified by both procedures. Markers were only used if they had 20% minor allele frequency and fell into a (loosely-defined) block of high linkage disequilibrium.

In regions of high LD, EM reconstruction works just about as well (i.e. small errors and no bias) as phased chromosomes in determining haplotype frequencies, regardless of haplotype frequency. This does not address the question of how well you can do at finding moderately rare haplotypes without phase info, nor does it say anything about low LD regions.

(very) PRELIMINARY CONCLUSIONS (Daly and Schaffner)

What have we learned thus far? We see that for relevant allele frequencies and sample sizes, the additional variance introduced by having no phase information (SINGLETONS vs. TRIOS or SAMPLE) is usually quite modest compared to the sampling variance. One concludes from this that, from a strictly mathematical standpoint regarding "how well can we estimate population haplotype frequency", the variance in our haplotype frequency estimates would be lower if we used, for example, 150 unrelated individuals (300 unphased chromosomes) instead of 100 trios (200 phased chromosomes). This is essentially the conclusion reached by Fallin and Schork (who performed much more detailed simulation studies comparing SINGLETON inference to perfect knowledge of the SAMPLE) and by Tishkoff et. al. who performed this analysis on real data at the CD4 locus so in this light it's not particularly surprising.

Of course there are significant advantages to a TRIO design over and above the ability to reconstruct phase. Most notably, these analyses and simulations do not take genotyping error into account and we must consider how such a large project, spread over many centers and laboratory techniques, will deal with the reality of errors in the raw data. Unless rigorous quality control, duplicate samples, etc. are in place, a strategy that uses only unrelated individuals will be unable to screen out individual markers or techniques that have very high error rates and will thus disrupt the construction of haplotypes. A TRIO design at least offers the ability to detect a reasonable fraction of errors as being Mendelian inheritance violations and thus can identify (given enough samples) the markers and techniques with the highest error rates.

B. Probability of sampling at least N copies of a particular haplotype for haplotypes of difference frequencies (Goncalo Abecasis)

The attached tables show the 99% confidence intervals for sample frequencies and the probability of sampling at least N copies of a particular haplotype for haplotypes of difference frequencies.

Results are listed for samples of 200, 132 and 100 chromosomes (50, 33 and 25 trios or 100, 66 and 50 unrelateds) and were calculated from a binomial distribution.

In short, with 33 trios or more (132 chromosomes) there is 99% probability of sampling 2 or more copies of 5% haplotypes. With 200 chromosomes there is 99% probability of sampling 4 or more copies of a 5% haplotype.

				99% Confidence Interval	
		Expected F	StDev	Lower	Upper
25 trios or 50 unrelateds (100 chromosomes)					
Freq	0.01	0.01	0.01	0.00	0.05
	0.05	0.05	0.02	0.00	0.12
	0.10	0.10	0.03	0.03	0.19
	0.20	0.20	0.04	0.10	0.32
	0.50	0.50	0.05	0.37	0.64
	0.80	0.80	0.04	0.69	0.91
33 trios or 66 unrelateds (132 chromosomes)					
Freq	0.01	0.01	0.01	0.00	0.05
	0.05	0.05	0.02	0.01	0.11
	0.10	0.10	0.03	0.04	0.18
	0.20	0.20	0.03	0.11	0.30
	0.50	0.50	0.04	0.39	0.62
	0.80	0.80	0.03	0.70	0.89
50 trios or 100 unrelateds (200 chromosomes)					
Freq	0.01	0.01	0.01	0.00	0.04
	0.05	0.05	0.02	0.02	0.10
	0.10	0.10	0.02	0.05	0.17
	0.20	0.20	0.03	0.13	0.28
	0.50	0.50	0.04	0.41	0.60
	0.80	0.80	0.03	0.73	0.88

				99% Confidence Interval	
		Expected N	**StDev**	**Lower**	**Upper**
25 trios or 50 unrelateds (100 chromosomes)					
Freq	0.01	1.00	0.99	0	5
	0.05	5.00	2.18	0	12
	0.10	10.00	3.00	3	19
	0.20	20.00	4.00	10	32
	0.50	50.00	5.00	37	64
	0.80	80.00	4.00	69	91
33 trios or 66 unrelateds (132 chromosomes)					
Freq	0.01	1.32	1.14	0	6
	0.05	6.60	2.50	1	15
	0.10	13.20	3.45	5	24
	0.20	26.40	4.60	15	40
	0.50	66.00	5.74	51	82
	0.80	105.60	4.60	93	118
50 trios or 100 unrelateds (200 chromosomes)					
Freq	0.01	2.00	1.41	0	7
	0.05	10.00	3.08	3	20
	0.10	20.00	4.24	10	33
	0.20	40.00	5.66	26	56
	0.50	100.00	7.07	82	119
	0.80	160.00	5.66	145	175

		Probability of Sampling N or More Chromosomes						
		1	2	3	5	10	20	50
25 trios or 50 unrelateds (100 chromosomes)								
Freq	0.01	0.63	0.26	0.08	0.00	0.00	0.00	0.00
	0.05	0.99	0.96	0.88	0.56	0.03	0.00	0.00
	0.10	1.00	1.00	1.00	0.98	0.55	0.00	-0.00
	0.20	1.00	1.00	1.00	1.00	1.00	0.54	0.00
	0.50	1.00	1.00	1.00	1.00	1.00	1.00	0.54
	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00
33 trios or 66 unrelateds (132 chromosomes)								
Freq	0.01	0.73	0.38	0.15	0.01	0.00	0.00	0.00
	0.05	1.00	0.99	0.96	0.79	0.13	0.00	0.00
	0.10	1.00	1.00	1.00	1.00	0.86	0.04	-0.00
	0.20	1.00	1.00	1.00	1.00	1.00	0.94	0.00
	0.50	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00
50 trios or 100 unrelateds (200 chromosomes)								
Freq	0.01	0.87	0.60	0.32	0.05	0.00	0.00	0.00
	0.05	1.00	1.00	1.00	0.97	0.55	0.00	0.00
	0.10	1.00	1.00	1.00	1.00	1.00	0.53	0.00
	0.20	1.00	1.00	1.00	1.00	1.00	1.00	0.05
	0.50	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00

C. Haplotype reconstructions in the absence of disequilibrium: Trios versus Singletons (Goncalo Abecasis)

Conclusions: In brief, in the absence of disequilibrium the relative efficiency per genotype of unrelateds vs. trios drops from ~1.5, to ~1.2, to ~0.7 to ~0.6 for 1, 2, 4 and 8 marker haplotypes. With disequilibrium, the relative efficiency of unrelateds increases and trios only seem advantageous for even longer haplotypes (more than 4 markers).

Random Haplotypes and equifrequent alleles:

		96 unrelateds		32 trios		
Markers	True(F)	Avg	Var	Avg	Var	Ratio
1	0.5	0.49826	0.00130	0.49826	0.00222	1.70
2	0.25	0.25045	0.00130	0.24822	0.00151	1.17
4	0.0625	0.06208	0.00081	0.06216	0.00057	0.71
8	0.00391	0.00377	6.3E-05	0.00385	3.8E-05	0.59

Columns are number of markers, true frequency of each haplotype, average estimated frequency for arbitrary haplotype and variance of estimate in 96 unrelateds, the same two values for 32 trios, and the ratio between variances.

I think if there is only one marker, trios should be 50% better, so the value of 1.7 (rather than 1.5 in the var(trios)/var(unrel) ratio gives some idea of the noise in these simulations).

The ratio seems to decrease with increasing number of markers and trios look better than unrelateds at four markers or more.

To be fair, these numbers are not that useful, because the point of the haplotype map is to estimate haplotypes where there is LD, and they are calculated under linkage equilibrium.

With LD, the ratios seem to depend on the actual pattern of LD, but the ratios seem to increase, that is the sample of unrelateds seems to benefit more from LD than the trios, in my limited experience.

One of the arbitrary models we tried was saying that haplotype frequency is proportional to k^max (#of 1 alleles, #2 of alleles in haplotype). For large k, this makes alleles with lots of 1's or lots of 2's very common and alleles with equal numbers of 1's and 2's rare.

For two markers this gives:

Freq	Haplotype
0.4404	11
0.0596	21
0.0596	12
0.4404	22

For four markers this gives:

Freq	Haplotype
0.3132	1111
0.0424	2111
0.0424	1211
0.0057	2211
0.0424	1121
0.0057	2121
0.0057	1221
0.0424	2221
0.0424	1112
0.0057	2112
0.0057	1212
0.0424	2212
0.0057	1122
0.0424	2122
0.0424	1222
0.3132	2222

For eight markers it gives two common haplotypes at 0.18 percent frequency, sixteen others at about 0.025 frequency and a bunch of rarer ones.

For the common haplotype this is the equivalent of the above table:

		Unrelateds		Trios		
Markers	True(F)	Avg	Var	Avg	Var	Ratio
2	0.440	0.441	0.00125	0.438	0.00209	1.67
4	0.313	0.314	0.00116	0.315	0.00171	1.47
8	0.182	0.207	0.00172	0.184	0.00132	0.77

So while the ratio is always larger (compared to the previous table), at eight markers trios are again better. Also we see a slight bias on estimating the common haplotype frequency in unrelateds (0.21 vs 0.18) in the eight marker case

For the rarer haplotypes, the ratios generally seem smaller than for the more common haplotypes.

Group 4: Population/ELSI: Resource Assessment and State of Current Knowledge

Overall Aims: To determine the organization of SNP variants over the entire human genome by identifying haplotypes that mark specific locations on the chromosomes, to assess the importance of variation in their frequencies in populations around the world, and to present these data so that they can be used to facilitate identification of the genes that contribute to common complex diseases.

A large scale, genome-wide study of genetic variation will be required to achieve the overall aims. It is not clear, however, how many populations should be studied and from which regions of the world these populations should be selected. A second question is whether populations differ in their haplotype patterns to an extent that it will be advantageous to identify samples by the populations that contributed them. As a first step to answering these questions, published and on-going studies that compare haplotype structure among populations are being collected. The results of these studies will be summarized to show what is currently known. Since these completed studies generally focus one or a few loci in one or a limited number of populations they are not sufficient to answer the above questions. Thus, a pilot study aimed at estimating the extent of SNP haplotype variation within and among populations around the world is required.

To this end, we have focused on design of a pilot project with the aims of:

1. Determining the number and frequency of common (>5%) SNP haplotypes in a number of genomic regions (~ 20) sufficient to make reliable predictions for the whole genome;
2. Comparing the frequencies of these haplotypes in population samples selected to optimize opportunities for identifying as many common haplotypes as possible and for determining population differences in haplotype frequency. Because of time and financial constraints, the pilot study is likely to be limited to a small number of populations, say 10 – 12.
3. Testing the value of population identification by comparing the yield of SNP haplotype information obtained with the populations identified with that obtained from the same data set with all population identifiers removed.
4. Testing the value of studying American populations of self declared origin by comparing their haplotype profiles with those of populations from geographical locations that correspond roughly with those declared by the American groups.

Strategy

Populations: In order to determine the extent of SNP haplotype variation, we recommend studying well-defined populations chosen from widely dispersed geographical locations. The theoretical studies of Fallin and Schork on the accuracy of the EM algorithm for haplotype frequency estimation and the studies of Tishkoff et al. on the haplotype variation at the CD4 locus suggest that analysis of samples from 50 unrelated individuals in each population should be sufficient to detect common (> 5%) haplotypes. Ideally, the geographical locations of the populations to be studied should be widely dispersed. This could be accomplished by obtaining populations from 3 regions of Europe, 4 regions of Africa, 2 regions of India or the near east and 3 regions of East Asia. Sampling more populations in the pilot phase than are expected to be included in the large-scale phase will provide information useful in deciding which populations to use in the large-scale study.

Samples:

Given the desire to complete the pilot in a 6 month time frame, it does not seem feasible to collect these samples prospectively. Thus, we have focused on identifying existing collections that meet the criteria of: i) adequate numbers of individuals; ii) adequate amount of DNA per individual (50 micrograms); iii) collected from appropriate regions; iv) acceptable consent. This information is still being collected and collated but initial review suggests that it may well be possible to assemble a collection that will suffice for the pilot study.

A tentative of possibilities with alternatives includes:

European samples

1. CEPH samples
2. Estonian samples
3. Southern/eastern Europe (? Sardinian samples)
4. US samples (additional discussion needed)

Alternatives Swedes (Lander)

African samples

1. West Africa (Yoruban samples ?)
2. East Africa (?)
3. North Africa – Moroccan (Tel Aviv Univ.)
4. South Africa (S. African Bantus – T. Jenkins, or samples from K. Kidd)
5. African-American samples (additional discussion needed)

Alternatives (other Kidd samples)

Indian samples

Aravinda has contacted Partha Majumder in Calcutta who appears to have adequate samples from many regions of India and is willing to contribute aliquots from them. He will need a statement from NIH indicating the importance for medical research in order to obtain permission from the Indian government to ship the samples oversees.

East Asian samples

1. Japan
2. Siberia (Yakut samples)
3. Asian-American (additional discussionneeded)

Alternative South Korea?

Lisa Brooks and Lynn Frampton are assembling specific information about available samples and this effort should be complete soon. At the same time Jean McEwen is collecting the consents under which these samples were collected.

We recognize that it is necessary to complete this initial pilot as quickly as possible in order to obtain information required for the planning of the large scale project. Some consideration should be given, however, to leaving the door open for pilot-like studies of other populations in the future. This would add to our knowledge of the extent of haplotype variation and would diminish the sense that populations not included in the initial study have been excluded from the potential benefits of the haplotype map.

The group is also discussing whether samples should be included only from populations from the US/Europe/Japan (potentially including samples from individuals with recent ancestry from other places), or whether samples from other places should be included as well.

Group 5: Population/ELSI: ELSI

The group has begun to discuss ELSI issues related to both the pilot study and the main project. Discussion so far has been primarily on establishing criteria for evaluating the acceptability of the informed consents for the existing samples being considered for the pilot study. The following issues have been identified:

1. There is consensus that the consent forms should specify, at a minimum, that the samples would be used for genetic research. There should also have been explicit consent for the making of cell lines (in instances where cell lines would be used) and for sharing the samples with other researchers.

2. There is general agreement that the consent forms should specify that the samples would be used for studies of genetic variation. An unresolved issue is whether the forms should be required to go further and state specifically that samples would be used for studies looking at variation *within and between populations.* There is general agreement that samples collected for research aimed specifically at studying the genetic etiology of a particular identified disorder (or type of disorder) would *not* be appropriate for use in the pilot unless the consent form were written in more general terms. This is because individuals' assessments of the potential for benefit or harm may be influenced by whether they perceive the research as focusing on a disorder of particular interest to them.

3. There is general agreement that some form of community consultation is desirable before samples will be used, although this may be difficult for the pilot study with respect to samples from older sets that were collected before community consultation for genetic variation research became the accepted practice. Some pilot studies of community consultation should be undertaken simultaneously with other parts of the pilot. It was recognized that community consultation should be viewed as a process of engaging affected communities and assessing a range of responses, and does not ordinarily mean that there needs to be formal community consent (except with certain populations).

4. There is general agreement that the informed consent should have been given under appropriate conditions and with appropriate conversations (which would generally require some type of inquiry beyond examination of the consent form). It was recognized, however, that the details of actual consent processes may be difficult to evaluate with older sample sets given the length of time that has passed since the samples were collected.

These criteria will also apply to any *new* samples collected for either the pilot study or the main study. The informed consent for the main study will raise additional issues and will need to be more comprehensive.

The group has begun the process of examining individual consent forms for each of the sample sets under consideration for the pilot study. The group believes that it *may* be appropriate to use the Utah samples from the CEPH collection for the pilot even though they do not fulfill all of these criteria. The suggestion has been made that there may be a justification for treating the CEPH samples somewhat differently given that those samples have already been so widely studied (and used in other studies of LD), given that the population from which the samples were collected is a majority U.S. population, and given that samples were quite clearly given as an altruistic donation to science. The CEPH samples may *not* be appropriate for use in the *main* study, however, given the absence in the CEPH consent form of any explicit mention of genetic variation research (as distinct from genetic research more generally).

Discussion is continuing on a number of other ELSI issues relevant to both the pilot and the main study. These include: the advantages and disadvantages of including particular populations from other than the U.S., Europe, and Japan; methods to be used for designating individuals as belonging to a particular population; goals of and methods to be used for community consultation; and protections for privacy and confidentiality. The group has also discussed the need for the development of a sound communication strategy to make sure that the public accurately understands the project.

Guyer, Mark (NHGRI)

From:	Brooks, Lisa (NHGRI)
Sent:	Friday, July 27, 2001 7:18 PM
To:	Collins, Francis (NHGRI); Jordan, Elke (NHGRI); Good, Peter (NHGRI); Hudson, Kathy (NHGRI); Boyer, Joy (NHGRI); Brooks, Lisa; Feingold, Elise (NHGRI); Felsenfeld, Adam (NHGRI); Frampton, Lynn (NHGRI); Graham, Bettie (NHGRI); Guyer, Mark (NHGRI); McEwen, Jean (NHGRI); Nakamura, Ken (NHGRI); Peterson, Jane (NHGRI); Pozzatti, Rudy O. (NHGRI); Roberts, Jerry (NHGRI); Schloss, Jeff (NHGRI); Thomson, Elizabeth (NHGRI); Wetterstrand, Kris (NHGRI)
Subject:	HapMap working groups and project names

HAP followup.doc

This file is the same as the text below.

HapExpertise.xls

To help you think of other names, all those who attended or were invited or who wanted to be invited to the HapMap meeting are enclosed.
I am in the middle of moving names into the appropriate categories for future reference, so the categories are not neat now.

Thanks, Lisa.

Planning process for the Haplotype Map Project

.

Funders

NIH

NHGRI, NIMH, NIDDK, …

Karen Kennedy? Wellcome

Martin GodboutGenome Canada

.

International

Thomas Meitinger	Germany
Yusuke Nakamura	Japan
Jean W eissenbach	France

.

Should these funders or researchers be on a planning committee?
Should there be another committee for the organization of the project?

.

(Names in italics volunteered.)

.

Study Design and Technology Group

What methods should be used to find haplotypes?
Are methods such as chromosome conversion, single-sperm typing, or moles ready to be used? What about long-range PCR?
For statistical methods, which samples are best: individuals, families with 1 or 3 kids?

How much flexibility should be included in the sample design, such as collecting families so that family information can be used but does not always have to be?
What types of markers should be used? What density of markers and what type of hierarchical scheme should be used?
What are the costs?
What information needs to be gathered to answer these questions, and how could it be obtained?

Large-scale SNP or haplotype discovery

Eric Lander	Whitehead (Chair)
David Altshuler	Mass General
Tom Hudson	*McGill*
David Bentley	Sanger

SNP and haplotype technologies

Pui Kwok	*Wash U*
Robert Nussbaum	*NHGRI*
Debbie Nickerson	U Wash

Population genetics

Andrew Clark	Penn State

Statistical analysis

Leonid Kruglyak	FHCRC
Richard Hudson	U Chicago
Bruce Weir	NCSU

ELSI

Charmaine Royal	Howard U

What about company reps?
David Cox
Eric Lai
Clay Stephens
David Wang
Michael Boyce-Jacino

Backup names

Mark Daly	*Whitehead*
Lon Cardon	Sanger
Steve Scherer	*Hospital for Sick Children, Toronto*
Michael Boehnke	U Michigan
Julie Douglas	*U Michigan*
Maynard Olson	U Wash
Peter Oefner	Stanford
Warren Ewens	U Penn
Charles Langley	UC Davis
Nick Schork	
Peter Donnelly	Oxford
Hongyu Zhao	*Yale*
Laura Lazzeroni	Stanford
Steve Sherry	NCBI

<u>Population and ELSI Group</u>

What are the rationales for studying identified populations?
What type or extent of population differentiation would not require the use of identified populations? What information

would address this question, and how could it be obtained? Could current projects be modified to obtain relevant data in a consistent and comparable way?
What types of populations should be sampled? (Population history, extent of LD, geography, admixture, known history, phenotyping potential)
Should the populations come from just the US/UK/Japan, or should populations in other countries be included?
How should populations be described? How should individuals be identified as being part of a population?
How can communities be involved in the process of deciding to participate?
What are the best mechanisms for consulting communities?
How should individual informed consent be obtained?
What needs to be done to ensure that the public accurately understands the project?
How can potential harms be minimized? (group stigmatization, reification of race)

Disease-gene mapping

David Valle	Johns Hopkins (Chair)
Aravinda Chakravarti	Johns Hopkins

Population studies

Ken Kidd	Yale
David Goldstein	University College London
Charles Rotimi	Howard U

Anthropology

Lynn Jorde	U Utah

Large-scale genotyping

Jim Weber	*Marshfield*

ELSI

Ellen Wright Clayton	*Vanderbilt*
Mildred Cho	*Stanford*
Troy Duster	UC Berkeley
Morris Foster	U Oklahoma
Pilar Ossorio	U Wisconsin
Marla Jasperse	*U New Mexico*
Pamela Sankar	U Penn
Vivian Ota Wang	Arizona State

Backup names

Nancy Cox	U Chicago
Anna Di Rienzo	U Chicago
Rick Kittles	Howard
Mark Shriver	*Penn State*
Li Jin	U Cincinnati
Alan Templeton	Wash U
Marty KreitmanU Chicago	
Jeff Long	NIAAA / U Michigan
Jonathan Friedlaender	Temple
Julio Licinio	UCLA (psychiatry, pharmacogenomics, community consent processes)
Wylie Burke	U Washington
Carl Elliot	U Minnesota (philosopher)
Mark Rothstein	U Kentucky
Francine Romero	Portland Health
Sharon Terry	PXE International
Nancy Press	Oregon Health Sci U
Barbara Koenig	Stanford
Bartha Knoppers	U Montreal

Patricia Marshall	Case Western
Jim Childress	U Virginia
LeRoy Walters	Georgetown
Laurie Zoloth	SF State

.Project Names

OK, we all agree that "Haplotype Map" just isn't going to cut it. If this is truly the Next Big Thing in genomics, it needs to sound like it. As Tom Murray pointed out at the meeting, even for genomicists we have hit a new low with "linkage disequilibrium" and "haplotype".

Names suggested so far by meeting attendees have been interesting but not overwhelming. We need a real zinger here. Words that might appear would include medical, health, heredity, genetics, genome, variation, innovation, people's, common, inheritance, etc. I sort of like "Map of All People's Shared Inheritance" (MAPSI), but it's, uh, not mellifluous. Another idea is the HHH project (could call this H-cubed if you'd prefer) -- Health, Humanity, Heredity. But some will think we mean Hubert Humphrey.

FC

I suspect that if we ran focus groups on this, the names people would respond to best would include both the words "shared" and "inheritance" and would NOT include words like variation, genome, genetic. We also need to stay away from words like "ancestry" and "history." Jean E. McEwen

Words not to include: linkage disequilibrium, haplotype, history, ancestry, pattern,
world populations

Shared Inheritance Map Elke Jordan
Shared Inheritance Map for Medicine Jean McEwen

Common Map Lincoln Stein
Common Threads Anne Stone

HuMAP- Humanity (mapping) Project- an international human genome program for discovery of the role of DNA sequence variation in disease in world populations. Steve Scherer

Genetic association map Pui Kwok
Common haplotype pattern map
Common haplotype map

Shared genome analysis Debbie Nickerson
Shared genome history
Shared pattern analysis (my Mom understood that one didn't like the genome term)
Shared pattern mapping (she liked that one better) mapping health related genes.
Variation pattern mapping
Shared history mapping
Shared ancestry mapping (could put variation instead of any of these)
Similarity Mapping

Common Heritage Genome Map Claude Laberge
Common Disease and Drug Response Discovery Map, For short: Common Map Morris Foster

People's Map

Lisa D. Brooks, Ph.D.
Program Director [redacted]
Genetic Variation Program
Genome Informatics Program
[redacted]
National Institutes of Health
31 Center Dr. 31 / B2B07 [redacted]
Bethesda, MD 20892-2033 301-480-2770 fax