Meeting Report: A Workshop to Plan a Deep Catalog of Human Genetic Variation

September 17-18, 2007
Cambridge, UK

Executive summary

A meeting was held to discuss the scientific rationale and design of an international consortium to develop a comprehensive catalog of sequence variants in multiple human populations. The primary purpose of the proposed project is to support the discovery and understanding of genetic variants that influence human disease. The workshop defined as specific goals (a) the discovery of single nucleotide variants at frequencies of 1% or higher in diverse populations, (b) even more comprehensive discovery (variants down to frequencies of 0.1 – 0.5%) in functional gene regions, and (c) discovery of structural variants, such as copy-number variants, other insertions and deletions, and inversions, including sequence-level understanding of breakpoints. For all variants, the frequencies and haplotype backgrounds need to be accurately established.

While these goals are not new, the project is motivated by recent proof-of-concept that genome-wide studies of human genetic variation can inform understanding of common human diseases, and has been made practical by recent advances in sequencing technology. These new methods produce shotgun coverage, and the group considered multiple design strategies to achieve the project goals. These ranged from deep coverage of a few genomes to light coverage of a larger collection. The effects of raw sequencing accuracy on design and statistical imputation were discussed in detail.

Based on these considerations, three pilot projects were designed to inform the design of the full project. The pilot projects, which are expected to take up to a year, are:

1. **To evaluate the use of low-redundancy genome sequencing to characterize single nucleotide and copy number variants, discovering all variants with frequency > 5% in the original HapMap samples.**

   This pilot will evaluate the utility of low-redundancy genomic sequence from many individuals, using the new sequencing technologies, including paired-end reads, for discovering SNP and structural variants and inferring haplotypes. These data will guide evaluation and development of methods for imputation from incomplete sequence data. In total 180 samples (60 unrelated samples from each of the HapMap CEU, YRI, and CHB+JPT populations) would be sequenced to a coverage depth of 2X of high quality mapped bases (1080 Gb total), and the resulting data analyzed to discover SNP and copy number variants.

2. **To evaluate the effect of coverage depth on project goals, based on deep sequencing of two sets of trio samples.**

   This pilot will evaluate the relationship between coverage depth and the yield of variation data, based on genomic sequence from a few individuals. A high level of redundancy will provide a solid basis for assessing the coverage needed for discovering variants, inferring haplotypes, imputing non-typed variants, and using paired-end reads for finding structural variants. Two trios (6 samples), one from each of the HapMap YRI and CEU panels, would be sequenced to a coverage depth of 20X of high quality mapped bases (360 Gb total).

3. **To develop and evaluate technologies to perform targeted sequencing of exons and other functional elements at genome-wide scale, and pilot deep sequencing in more than 1,000 DNA samples.**

   This pilot will develop and evaluate technologies to capture specific genomic regions and discover variants. It will provide data on the frequency distribution of rare variants, and in combination with other data enable the study of haplotype patterns around rare alleles. It will thus guide development of algorithms to impute less common alleles from SNP data. In total, 1000-2000 gene regions and conserved elements would be sequenced at 20X of high quality mapped bases in 1085-1536 samples (109-307 Gb).

The specific designs of these pilots may be modified based on intermediate results.

Introduction

Over the last decade, a genome-wide catalog of human genetic variation has been developed for single nucleotide variants covering most SNPS with allele frequencies > 5% in certain populations, and initial catalogs of structural variants have been developed. Based on these resources, initial genome-wide association studies (GWAS) have provided proof-of-concept that systematic, genome-wide association studies can discover new loci that contribute to common human diseases. Over 50 new loci contributing to common human diseases have been identified for diseases ranging from heart attack and diabetes, prostate and breast cancer, rheumatoid arthritis and inflammatory bowel disease, age related macular degeneration, and many others.

By identifying new disease loci, each such discovery prompts much additional genetic research. For each such locus, it is currently necessary to sequence the newly discovered region to define all common and rare variants, one or a few of which contribute to disease. At present, this sequencing must be done on a case-by-case basis, and it is both expensive and inefficient.

Moreover, it is clear that discoveries yet made explain only a modest fraction of disease risk. Some of this uncaptured risk is due to alleles of lower frequency but larger effect. If such alleles are in genes already localized by GWAS, then targeted sequencing may find them; but if the genes do not carry high frequency variants of sufficiently large effect, they will go undiscovered unless deeper catalogs of variation are obtained. Similarly, some of the uncaptured risk is due to the effects of structural variants that are not in linkage disequilibrium (LD) with common SNPs; these too will go undiscovered until a detailed and complete map of structural variants is obtained to guide systematic association studies of this important class of genetic variation.

Thus, a more complete understanding of the role of genetic variation in disease requires a deeper catalog of genetic variation, both for single nucleotide and structural variants. This hadn't previously been approachable because (a) there wasn't proof-of-concept that the general approach would succeed, and thus greater investment in the resource was unjustified, and (b) sequencing costs and throughput made it impractical to sequence more deeply and broadly to create this more comprehensive catalog of genetic variants for use in genetic research.

Recently, next generation sequencing technologies have become available that appear likely to reduce the cost of sequencing by one to two orders of magnitude, and dramatically increase throughput. These technologies make it practical to produce a database of genetic variants that is deeper, broader, and more comprehensive, and that will have immediate and substantial impact in human genetics.

Sequencing many human genomes, unselected with regard to phenotype, should provide a resource of variants to support deeper understanding of newly discovered loci influencing human disease, and a next generation of association studies that query less common variants and structural variants. More broadly, these data will illuminate many other applications and push the technologies and analytical approaches for human genetics. This report describes plans for developing such a variation resource.

Background and planning process

Catalyzed by the combination of rapid progress in population genomics and the development of new sequencing technologies, several groups have developed projects to obtain sequence information from many people. During the May 2007 Biology of Genomes meeting at the Cold Spring Harbor Laboratory, at a gathering of the International Genome Sequencing Consortium, Richard Durbin of the Sanger Institute proposed that members of the consortium plan such an effort.

After preliminary discussion among groups in the U.K., the U.S., and China, a working group was established to consider possible goals, designs, and analytical issues. During a series of five conference calls the participants discussed the goals of the project, identified issues related to samples and sequence quality, and developed several straw proposals to address those goals for discussion at the face-to-face meeting. (See the end of this document for lists of participants.) The purpose of the September meeting in Cambridge was to develop a plan for the project for the next two to three years. At the meeting, several components of the project were agreed to, but many topics need more discussion.

Project goals

The primary goal of the project was generally agreed to be the development of a public resource of genetic variation to support the next generation of association studies relating genetic variation to disease. This includes both the follow-up of whole genome scans (i.e., loci already discovered) and the discovery of loci not discoverable with previous resources. Secondary goals include supporting studies of human population genetics and evolution, which inform the biomedical goals of the project.

Primary goals:

1. Discover variants (SNPs, copy-number variants, insertions, deletions, other structural variants).

   As a genomic project the resource should provide completeness; **the resource should include almost all accessible variants with allele frequencies as low as 1% across the genome and 0.1-0.5% in gene regions**. Currently the common SNPs are mostly known; the additional sequencing will be especially valuable for the discovery and characterization of many more rare variants and structural variants.

2. Estimate the frequencies of the variant alleles, identify their haplotype backgrounds, and characterize their LD patterns.

   The variant calls and haplotype phasing should be provided for each sample, which would support imputation of non-typed variants and tag SNP choice.

Secondary goals

3. Support better SNP and probe selection for genotyping platforms.
4. Improve the human reference sequence.
5. Support studies of regions under selection.

6. Support studies of variation in multiple populations.
7. Inform understanding of the underlying processes of mutation and recombination.

Straw proposals

Participants considered several straw proposals to highlight different approaches to reaching the project goals. The straw proposals were motivated by questions that were raised on the calls:

1. For a whole-genome approach, assuming a fixed project sequencing budget of 6 terabases (Tb), what is the right balance between the number of samples sequenced and the depth of sequence coverage for each sample?
2. Focused sequencing in genomic regions of particular interest such as coding exons could provide deeper information on low frequency variants. What are the practical approaches to doing this? What are the advantages and disadvantages compared to whole-genome approaches?

Whole-genome approaches

Sequence many samples with light coverage (Richard Durbin)

Sequence 100 samples each from 10 populations (of 4 major geographic groups) at 2X coverage of high quality mapped bases = 1000 diploid genomes.

This approach would reach the minor allele frequency (MAF) goal (finding 90% of the accessible 1% frequency variants per population, 99% overall) without needing to sequence to high (20-30X) coverage per sample, which would be too expensive. Simulations and preliminary analyses of yeast sequence data were presented indicating that haplotypes are imputable by using chromosome segments shared among samples. Having data from more samples would improve imputation power when genotyping new individuals. However, there are uncertainties about the performance with human sequence data obtained on very short read platforms; the exact trade-off between sample number and coverage is sensitive to the sequencing error rate.

Sequence trios with moderate coverage (Gil McVean)

Sequence 30 trios each from 2 populations at 11X coverage of high quality mapped bases (= 11X coverage for the transmitted chromosomes and 5.5 X coverage for other chromosomes) = 990 diploid genomes.

Sequencing trios at moderate coverage would find low frequency variants with good power. The deep coverage would allow confidence in the variants detected, although the limited number of samples would not allow reaching the MAF goal (finding only 70% of the 1% frequency variants). Haplotypes would also be found with confidence and low ascertainment bias, and the data would allow good QC and potential detection of mutation events. Having well-phased chromosomes would help good imputation of variants and haplotypes in other studies. Sample selection would need to be considered carefully: using the second HapMap plates would maximize the new information gained, but using the first plates would maximize the validation data available for assessing the strategy.

Sequence using a staged approach (Gonçalo Abecasis)

Initially, sequence 60 unrelated samples from each of the HapMap YRI, CEU, and CHB+JPT panels at 2X coverage of high quality mapped bases = 180 diploid genomes. Evaluate the data as they are produced, and decide whether to add coverage or samples.

The optimal depth of coverage for finding rare variants and for combining data from multiple samples to infer haplotypes depends on the variability in coverage and the error rates for sequencing. This approach would start with sequencing a few samples at low coverage, imputing the variants and haplotypes for all the samples, and evaluating the predictions by comparison with other data, such as sequence data from the HapMap ENCODE regions. If the results are worse than expected (e.g., due to poor data quality) then add coverage; otherwise add samples. The samples should overlap with the extended HapMap set used for the HapMap ENCODE II sequencing.

Regional approaches

Sequence gene regions or the HapMap ENCODE II regions (Richard Gibbs)

Sequence the 1085 unrelated samples of the extended HapMap set and some trios, in gene regions or alternatively in the 1.3 – 2 Mb of selected ENCODE regions.

This approach aims to find almost all variants, including very rare ones (MAF $\geq$ 0.1%), in targeted regions of the genome. Finding such rare variants and their haplotypes requires deep coverage, using many more samples than studied previously. Two options were discussed. The first would be to study the HapMap ENCODE II regions to focus on allele frequencies and haplotype structures. Looking at 13-20 regions each 100 kb long already sequenced in 805 unrelated samples of the extended HapMap set would allow considerable validation because of the extensive work already done in those regions. However, even deep sequencing of these regions would not provide a resource directly applicable to association studies, since the follow-up for association studies is to examine gene regions, not the ENCODE regions. The second option would be to sequence the regions around a set of genes, including exons plus flanks, 5' and 3' regions, and conserved non-coding regions, to identify variants that are likely to have functional consequences. The data therefore would provide information about the frequency distribution of rare non-synonymous variants. Developing a complete catalog of these variants is complicated because non-synonymous SNPs differ more between populations than do other SNPs, even accounting for frequency. Because this approach would provide useful information about variation in regions that will be of high interest for disease studies, the result would be an enduring resource.

Critical issues for designing the project

The discussion of straw proposals, together with the background information provided on the planning calls, led to a discussion of several critical issues that need to be addressed in designing the project.

Data quality standards and project metrics

Several types of data quality standards will need to be developed, as well as metrics to measure the progress of the pilot and full-scale projects. The group discussed these metrics:

Discovery of variants

1. Fraction of alleles discovered as a function of frequency, and for structural variants, of their size.
2. False positive and negative rates.

Genotyping and imputation

1. Error rates for calling variants.
2. Error rates for allele frequencies.

3. Error rates for phasing haplotypes.
4. Error rates for imputed genotypes.

Some questions were raised that could form the basis for additional project metrics:

1. What will the data give us beyond what is provided by existing resources?
2. How much variation will be found that cannot currently be assessed in association studies?
3. How complete and accurate will the data be for variation in regions found to be of interest in association studies?

All samples used should first be genotyped on commercial platforms, which will aid phasing and QA. The Broad Institute has genotyped all the samples in the extended HapMap set for the Affymetrix 6.0 chip, and the Sanger Institute agreed to type these samples using the Illumina 1M SNP platform. This genotyping will provide some validation.

To validate the sequence data, a subset of the novel variants that are found should be genotyped, especially the rare ones. To distinguish very rare variants from ones seen only in single individuals, genotyping will be needed in large sample sets, such as 30,000 samples.

Technology challenges and scientific questions to address

Sequence production: The performance of the new platforms is not sufficiently well understood to know how the sequencing should be done. Technological diversity is good, but common metrics of production and quality will be needed.

Data quality: The accuracy of calls for SNPs and structural variants needs to be measured. False negative rates can be estimated based on data on the same samples from the Affymetrix 6.0 and Illumina 1M SNP platforms; false positive rates can be estimated by genotyping new variants. The quality scores need to be related to the accuracy of sequencing. The definition of "good" sequence needs to be established.

Paired-end reads and detecting structural variants: Paired-end reads improve the mapping of reads to the genome and provide information on structural variants. Paired-end reads are desirable, but the technology is not mature yet; there is some loss due to failed reads, and the process costs are not clear. However, the technology is expected to improve. The pilot projects should use at least some paired-end reads. The full project should be done using paired-end reads, unless the technology cannot be made to work.

Exon (region) capture: Methods to capture regions of the genome for regional approaches to sequencing also require more development. A few methods are being worked on now and they should be evaluated.

The frequency distribution of rare variants: The number and frequency spectrum of rare variants are unknown. There are about 3-5 common (MAF > 5%) coding SNPs per gene, and there may be ten times as many sites with rare variants. Each person may have about 3000 rare coding variants. What proportion of rare variants are unique to individuals, and which are at low frequency but polymorphic in the population? Genotyping large samples will provide information on the frequency distribution of

rare variants. Finding this allele frequency spectrum and placing the variants on haplotypes will address whether it would be useful to place all the variants that are found on haplotypes.

Phasing and imputing genotypes: What data accuracy and depth of coverage is needed to phase variants accurately? How well can rare alleles be imputed and placed on the correct haplotypes?

Sequence coverage and data quality

What genomic sequence coverage is needed for accurate diploid sequencing? Accuracy in sequencing low frequency alleles arises from:

1. Including samples with the variant: The number of times a variant is seen in a sample has a binomial distribution, which is essentially a Poisson distribution for rare alleles.
2. Including the haplotype with the variant: The probability of seeing a variant in a heterozygous individual has a binomial distribution, given an overall depth of coverage at that locus in the individual.
3. Including genomic coverage at the site of the variant: Shotgun sequencing has a broader distribution of coverage at a particular site than a Poisson; more regions have few reads.
4. Sequencing accurately: With errors, a site needs to be sequenced at least twice for accuracy. The error rate needs to be very low for most differences to be variants and not errors.

About 80-90% of the genome is sequenceable with paired end short reads. When a single sample is sequenced, high average coverage is needed to ensure that the genome has good coverage. A depth of 11X gives a 99% probability that both alleles are seen at least twice. Average coverage of 20-30X is needed for sufficient depth for finding both alleles confidently: 21X genome-wide gives a 95% probability of having 11X coverage at a site and 27X gives a 99% probability of 11X coverage. If sequencing errors are sufficiently common or correlated that a variant must be seen more than twice to be believed, then even higher average coverage would be needed.

A variant seen in multiple samples is generally real. Imputation methods allow the use of haplotype data to be shared among samples to provide deeper effective depth. This works with 2-4X coverage per sample, which has lots of missing data, although it misses a large fraction of the very rare variants (below 1%).

Samples

Based on the overall goals for the project, the samples should be chosen to provide power in populations where association studies are being done for common diseases. There was also a case made for sequencing African samples, which contain maximal founding genetic diversity. Since this will be a basic resource on human variation, the samples do not need to have associated medical or phenotype data. This project should focus on samples that are consented for open access on the web without needing approval for each use. These requirements probably limit the project to sequencing the extended set of HapMap samples, at least for the pilot projects. It may be desirable to expand this sample set for the future, especially if more power is needed in certain populations and for the deeper sampling of gene regions. It was discussed that immigrant populations did not provide ideal sampling of founder populations. Expanding the sample set is likely to require new consent. Two potential sample sets considered (the 1958 British Birth Cohort and the NIMH Gejman controls) would not be usable due to restrictive consent clauses. It was agreed that a consent framework should be made

public so that samples meeting the necessary criteria could be collected from multiple populations by diverse investigators, but the time required to obtain new samples with adequate individual consent and, in some instances, with community consultation, should not be underestimated.

Evaluating the pilot projects and understanding the properties of the resource require comparing results with those from other projects, so it will be useful to focus on samples that have the maximum additional data available (such as ENCODE sequence, genome-wide genotypes, fosmid-end sequence, structural variation assays, and gene expression). Therefore it will be useful to have overlap in samples between this project and others. In addition, because the design of the full project depends on the data from the pilot projects, the optimal sample overlap among the pilot projects needs to be considered.

Data and sample release

The meeting participants agreed that the project data should be freely accessible on the web, prior to publication, in accordance with the principles for genomic community resource projects. The participants thought that the data should be more easily available in user-friendly form than the data currently are in the Trace Archive. In addition, participants thought that cell lines should be available from the samples to allow future studies, including cellular phenotyping.

Pilot projects

The quantitative goals of the full project are

- To identify > 95% of the variants with a MAF $\geq$ 1% in the sequenceable parts of the human genome, with > 95% certainty.
- To identify > 95% of the variants with a MAF $\geq$ 0.1 - 0.5% in exons, with > 95% certainty.

There are several uncertainties in how to accomplish these goals. The participants proposed three pilot projects to resolve the issues needed to design the full project. The pilot projects should be completed within a year. They will be:

1. **To evaluate the use of low-redundancy genome sequencing to characterize single nucleotide and structural variants, discovering all variants with frequency > 5% in the original HapMap samples.**

   This pilot aims to evaluate the ability of the platforms to use light coverage to find 95% of variants across the genome with frequencies down to 1%. The U.K. and U.S. centers will focus first on quickly producing data from the CEU samples, for rapid analysis of the initial results. Whether to use the first or second HapMap sample plates still needs to be decided. The first plates would provide the maximum information for validation and imputation; the second plates would provide more new information on variants, frequencies, and LD. Some sequencing should be done by paired-end reads for assessing the potential improvement in read mapping and discovery of structural variants. A total of 180 samples will be sequenced (60 samples from each of 3 major geographic areas (HapMap CEU, YRI, CHB+JPT)) at 2X of high quality mapped bases (1080 Gb = 18% of the 6 Tb).

2. **To evaluate the effect of coverage depth on project goals, based on deep sequencing of two sets of trio samples.**

   This pilot aims to assess how well high-coverage data from the various technologies detect the variation, and will provide comprehensive variation, haplotype, and transmission data on a few samples. The 20X

(diploid) coverage in parent and child means that the transmitted chromosomes will be done at 20X. (The actual coverage may depend on the technology used; the coverage should deeply oversample to provide enough high quality data to design the project.) The trios should have the maximum amount of other data, including fosmid paired-end reads, Affymetrix and Illumina genotyping, and CNV detection by high density oligo CGH. The coverage should be split across centers so as to compare technologies. This project will also include an exploration of the use of paired-end sequence reads. Two trios (6 samples), one from each of HapMap YRI and CEU, will be sequenced at 20X of high quality mapped bases (360 Gb = 6% of the 6 Tb).

3. **To develop and evaluate technologies to perform targeted sequencing of exons and other functional elements at genome-wide scale, and pilot deep sequencing in more than 1,000 DNA samples.**

   This pilot aims to develop and compare technologies for creating a deeper catalog of variation in gene regions and other conserved regions, to provide data on the frequency distribution and LD patterns for rare variants. The genes could be chosen randomly or stratified in some way to provide guidance on dealing with different types of gene regions in the full study. Finding rare variants requires examining many samples; these will include the 1085 unrelateds in the extended HapMap sample set, hopefully augmented by other existing samples that have HapMap-level consents to reach a total of 1536, and it would be good to include many samples (800) from one population, possibly of European ancestry. The ability to obtain haplotype and LD information in conjunction with genotype data on the same samples will be explored. Some samples may be trios, which would help to phase rare variants. Multiple approaches are encouraged for isolating regions and for doing the sequencing, including ones that pool samples or deal with all genes at once. To compare the technologies, each gene and sample should be done by at least two centers. Since the proportion of rare variation shared among people and among populations vs. singleton variation only in individual samples is unknown, these data will be useful for designing the full study, which will look at all genes. In total, 1000-2000 gene regions and conserved elements will be sequenced for an average of 5 kb each at 20X of high quality mapped bases in 1085-1536 unrelated samples (109-307 Gb = 2-5% of the 6 Tb).

Implementation of the project

Meeting participants have roughly the following capacity available for the project:

	Have			Will use		
	454	Solexa	SOLiD	454	Solexa	SOLiD
Sanger	2	25	0	0	12	0
BGI	1	7	0	0	5	0
Broad	3	20	1	as directed by NHGRI		
Baylor	3	1	2	as directed by NHGRI		
Wash U	3	6	1	as directed by NHGRI		

If the whole-genome light-coverage pilot project were done using just Solexa machines, it would take 1000-2000 Solexa runs / 2 runs per week = 10-20 machine-years. The trios project would take 3-6 machine-years. If the gene region project were done using the 454 platform, it would take about 1.5 machine-years.

The participants agreed that several working groups will be needed to manage this project:

1. Steering Committee: This group will include representatives from the sequencing centers (Sanger Institute, Beijing Genomics Institute, Broad Institute, Baylor College of Medicine, Washington University, and others that may become involved) and the joint data coordination center (EBI and

NCBI), experts on analysis, human genetics, and population genetics, and representatives from the funding agencies.

2. Samples and ELSI Group: This group will decide on the criteria for choosing samples, choose appropriate samples, assess the need for new samples, obtain them if necessary, and address the ethical, legal, and social issues related to this project. An immediate need is to choose samples for the pilot projects.
3. Data Production and Technology Exchange Group: This group will exchange information on technologies being developed (such as quality scores and paired-end reads), develop methods of data production, track progress, and assess the data as they are produced. QC will be done jointly with the Analysis Group. An immediate need is to start comparing the technologies.
4. Analysis Group: This group will address the many analysis issues related to data production and quality, evaluation of the pilot projects and design of the full project, and analysis of the project data. There will be several sub-groups, including gene selection, variant calling, population genetics, structural variation, and QC. Tasks include doing the analyses related to sample selection, gene selection (from CCDS and possibly from medical sequencing projects, annotation, choice, boundaries), SNP and small variant calling (read mapping, quality scores, validation), large structural variation (counts and paired-end reads), QA/QC (round-robin, third-party, validation by overlap among projects), and statistical genetics (imputation, data simulations, project design). Several of these tasks should be started immediately.
5. Data Coordination and Flow Group: This group will address how data flow from the production centers to the databases, including coordination, QA, curation, and release. An immediate need is to agree on data exchange standards with the Data Production and Technology Exchange Group.

Future meetings may be held in conjunction with other meetings, such as the Advances in Genome Biology and Technology meeting (February 6-9, 2008) and the CSHL Genome meeting (May 6-10, 2008).

Evaluating the resource

The full variation resource is not going to be available for a couple of years, and association studies will sequence under association peaks in that time. The resource can be evaluated when it is developed by checking whether the results from the follow-up sequencing studies could have been obtained by using the resource; is the database sufficiently complete that resequencing would not have been necessary? This will allow improvements in the strategy for sequencing for association studies and in the next generation of the resource.

Participants in the planning process

Participants on the phone calls: Gonçalo Abecasis, David Altshuler, Lisa Brooks, Andrew Clark, Francis Collins, Manolis Dermitzakis, Peter Donnelly, Richard Durbin, Adam Felsenfeld, Richard Gibbs, Mark Guyer, Matt Hurles, David Jaffe, Eric Lander, Elaine Mardis, Gil McVean, Jane Peterson, Jonathan Pritchard, Chris Tyler-Smith, Jun Wang, George Weinstock, Richard Wilson, and Henry Yang.

Participants at the meeting: Gonçalo Abecasis, David Altshuler, Ewan Birney, Lisa Brooks, Aravinda Chakravarti, Francis Collins, Greg Cooper, Richard Durbin, Adam Felsenfeld, Paul Flicek, Richard Gibbs, Mark Guyer, Matt Hurles, David Jaffe, Eric Lander, Elaine Mardis, Gabor Marth, Gil McVean, Alan Michelson, Rasmus Nielsen, Steve O'Rahilly, Leena Peltonen, Jane Peterson, Jonathan Pritchard, Michael Province, Alan Schafer, Steve Sherry, Simon Tavaré, Jun Wang, George Weinstock, and Richard Wilson.