The 1000 Genomes Project:

obtaining a deep catalogue of human genetic variation with new sequencing technology
Chromosome 9p21: diabetes, coronary heart disease. Three genes, multiple SNPs


500,000 basepairs of Chr 9 (total length 109M bp)
After GWAS “hit”, what next?

(remember, these are associations, not causes)

One region (~Mb), multiple genes, or sometimes no genes (!), multiple SNPs to sort through

Which is the right gene? What is the “causal” variant?

The current SNP catalog is not complete – may not have the causal variant
After a GWAS “hit”, what next?

- One could get lucky (gene is a likely candidate based on previously known function*; a known associated SNP is a variant that prevents any gene function)
- Gene expression correlates with believed function (e.g. tissue specific, disease specific)
- Conservation of sequence between genomes of many mammals
- Get a complete list of variants in the region, and one of them will be right. Need to sequence the associated region in many people.

*CDKN: evidence for a role in islet cell growth. Also a tumor suppressor.
Chromosome 9p21: diabetes, coronary heart disease. Three genes, multiple SNPs

Good bet on the gene, but what is the cause?
1000 Genomes Project: A resource for aiding human genetics studies

• An essentially complete list of all variants in human populations

• To provide a catalog of almost all variants in regions of all possible GWAS hits (i.e., the whole genome) ahead of time, so studies do not need to sequence their samples

(Gives the complete list of candidates, but still have to follow up on all candidate variants!)
Other potential benefits for Whole Genome Association studies

- The new variants will be associated by LD context with all existing variants, increasing the power of GWAS
- Better design of future assays for variation
- Access to lower frequency variants than current designs, e.g. down to 1%. (At what frequency do disease-causing variations occur in the population?)
- Can find alternate alleles in region of interest (disease could be caused by more than one variant in a single gene)
1000 Genomes primary goals: how many more variants?

“Essentially all” (not just a lot of) common variation genome-wide: any variant occurring in the population down to 1% allele frequency.

Deeper in gene regions (0.5%-0.1%)

All variant types (SNPs, insertions/deletions, and structural variants)

Place variants in their haplotype context (what other variants are they associated with?)

Do this in multiple populations—enough people at random that “all” (medically relevant) variation will be represented
How to do?

- Sequence in three populations to start: European, Africa, East Asian*; 500 individuals each

- Need to understand exactly how much sequence needed from each individual to build haplotype information

- A one-year pilot phase to test theory and technology:
  - What will it take for the new platforms to produce data that are useful for this?
  - How much sequence from each individual is needed?
  - Do we have enough from each population?
  - Build analytical infrastructure

- Two year main project

*Samples are mostly those already collected for HapMap under appropriate consent for fully anonymous release of genomic data. Some new anonymous samples will be needed.
1000 Genomes Pilots

Started Feb 2008, ~ 300 Gb data already

- Pilot 1: 180 samples @ less sequence each:
  CEU (European) 4x, YRI (African) 2x, CHB/JPT (East Asian) 2x

- Pilot 2: CEU and YRI families (two parents, one child) @ high levels of sequence (20X)
  CEU trio mostly complete
  YRI trio in progress

- Pilot 3: 1000 genes in 1000 people
  Starting

- Test multiple platforms/protocols

- Develop and evaluate methods for data collection and analysis

  Simulations of trios, 1000 people at 2x, plus samples at 4x, 8x

~10 people done
Additional goals

Not just SNPs: structural variation (2bp to >1M bp)

Population genetic studies
- Identifying regions under selection (now or in the past)
- Studies of processes of mutation and recombination
- Population differentiation and history

Improvement of the human reference sequence
- Find and fix errors
- The current reference sequence, and any one individual, is missing sequence present in others
- Coordinate with the Human Genome Reference Consortium to represent all unique human sequence
Impractical without new sequencing technologies

• Project requires ~18,000 Gb

• “Old” tech (2006): >$1B

• New tech (2008): ~$50M
Challenges in “drinking from the firehose”

- Data handling, informatics resources: a LOT of data—the initial deposition increased the total sequence data available in the public domain by 10%, overnight

- Analysis, analysis, analysis…

- Samples, with appropriate consent for use in genomic studies and data release
1000 Genomes Consortium

Production: Sanger Institute, Beijing Genomics Institute, Baylor College of Medicine, Broad Institute, Washington University of St Louis

Analysis: many statistical and population geneticists

Data Coordination: European Bioinformatics Institute, National Center for Biotechnology Information

Samples/ELSI: expertise in ethics and population sampling

Funding: Wellcome Trust, Beijing Genomics Institute, National Institutes of Health/NHGRI
A Public Resource

- Data publicly available shortly after it is produced
  - raw sequence data in the Short Read Archive
  - SNPs and other variant data in dbSNP
- Cell lines available
1000 Genomes Project steering co-chairs:

Richard Durbin   Wellcome Trust Sanger Institute
David Altshuler   Broad Institute

NHGRI Staff:
Lisa Brooks
Jean McEwen
Adam Felsenfeld
1000 Genomes
A Deep Catalog of Human Genetic Variation

Samples and ELSI Group
Leena Peltonen (co-chair) Sanger Institute
Bartha Knoppers (co-chair) University of Montreal
Aravinda Chakravarti (co-chair) Johns Hopkins
Goncalo Abecasis University of Michigan
Richard Gibbs Baylor College of Medicine
Lynn Jorde University of Utah
Eric Juengst Case Western Reserve University
Jane Kaye Oxford University
Alastair Kent Genetic Interest Group
Rick Kittles University of Chicago
Jim Mullikin National Human Genome Research Institute
Mike Province Washington University in St. Louis
Charles Rotimi Howard University
Yeyang Su Beijing Genomics Institute
Chris Tyler-Smith Sanger Institute
Ling Yang Beijing Genomics Institute

Steering Committee
Richard Durbin (co-chair) Sanger Institute
David Altshuler (co-chair) Broad / MGH / Harvard
Goncalo Abecasis University of Michigan
Aravinda Chakravarti Johns Hopkins
Andrew Clark Cornell University
Francis Collins National Human Genome Research Institute
Peter Donnelly Oxford University
Paul Flicek European Bioinformatics Institute
Stacey Gabriel Broad Institute
Richard Gibbs Baylor College of Medicine
Bartha Knoppers University of Montreal
Eric Lander Broad Institute
Elaine Mardis Washington University in St. Louis
Gil McVean Oxford University
Debbie Nickerson University of Washington
Leena Peltonen Sanger Institute
Stephen Sherry National Center for Biotechnology Information
Rick Wilson Washington University in St. Louis
Huaming (Henry) Yang Beijing Genomics Institute

Production Group
Elaine Mardis (co-chair) Washington University in St. Louis
Stacey Gabriel (co-chair) Broad Institute
Richard Durbin Sanger Institute
Richard Gibbs Baylor College of Medicine
David Jaffe Broad Institute
Ruiqiang Li Beijing Genomics Institute
Donna Muzny Baylor College of Medicine
Chad Nusbaum Broad Institute
Aarno Palotie Sanger Institute
Dan Turner Sanger Institute

Data Flow Group (being formed)
Paul Flicek (co-chair) Euroean Bioinformatics Institute
Stephen Sherry (co-chair) National Center for Biotechnology Information
Ewan Birney European Bioinformatics Institute
Clive Brown Sanger Institute
David Dooling Washington University in St. Louis
Richard Gibbs Virginia Commonwealth University
Sol Katzman University of California, Berkeley
Neeraj Nanda Memorial Sloan Kettering Cancer Center
Hoda Khoury National Center for Biotechnology Information
Martin Shumway National Center for Biotechnology Information
Jun Wang Beijing Genomics Institute
George Weinstock Baylor College of Medicine

Funders
Alan Schafer Welcome Trust
Francis Collins National Human Genome Research Institute
Lisa Brooks National Human Genome Research Institute
Audrey Duncanson Welcome Trust
Adam Felsenfeld National Human Genome Research Institute
Mark Guyer National Human Genome Research Institute
Ruth Jamieson Welcome Trust

Analysis Group
Gil McVean (co-chair) Oxford University
Goncalo Abecasis (co-chair) University of Michigan
David Altshuler Broad / MGH / Harvard
Paul de Bakker Broad / BWI / Harvard
Brian Browning University of Auckland
Sharon Browning University of Auckland
Carlos Bustamante Cornell University
David Carter Sanger Institute
Aravinda Chakravarti Johns Hopkins
Andrew Clark Cornell University
Don Condon Sanger Institute
Mark Daly Broad / MGH / Harvard
Manolis Dermitzakis Sanger Institute
Peter Donnelly Oxford University
Richard Durbin Sanger Institute
Evan Eichler University of Washington
Paul Flicek European Bioinformatics Institute
Bryan Howie Oxford University
Matt Hurles Sanger Institute
David Jaffe Broad Institute
Lynn Jorde University of Utah
Hosta Khoat National Center for Biotechnology Information
Eric Lander Broad Institute
Charles Lee Brigham and Women's Hospital
Guoqiang Li Beijing Genomics Institute
Heng Li Sanger Institute
Ruiqiang Li Beijing Genomics Institute
Yingni Li Beijing Genomics Institute
Yan Li University of Michigan
Jonathan Mandel Oxford University
Gabor Marth Boston College
Steve McCarroll Broad Institute
Jim Mullikin National Human Genome Research Institute
Simon Myers Oxford University
Rasmus Nielsen University of California, Berkeley
Alkes Price Broad / Harvard
Jonathan Pritchard University of Chicago
Mike Province Washington University in St. Louis
Molly Przeworski University of Chicago
Shawn Parcells Broad / MGH / Harvard
Noah Rosenberg University of Michigan
Pandis Sabeti Broad / Harvard
Paul Schenkeveld University of Washington
Steven Schork National Human Genome Research Institute
Nancy Shriver California Institute of Technology
Matthew Stephens Stanford University
Simon Tavare University of California
Chris Tyler-Smith Sanger Institute
Jun Wang Beijing Genomics Institute
David Wheeler Baylor College of Medicine
Hongkun Zheng Beijing Genomics Institute
Medical Sequencing

• Finding sequence variants that underlie disease
• Ideal: Sequence whole genomes of patients vs. healthy people, identify differences
• Reality: Too expensive now
• Challenge: Too many variants to sort through
• Solution: Pick candidate regions (e.g., GWAS; by function; by other previous findings); or “exomes” (practical very soon).
Medical Sequencing

Example: Autism

• Choose candidates based on function e.g., in neuronal synapses
• Sequence those genes in multiple affected and unaffected individuals
• Follow-up all differences (will find many differences, so this step needs to be relatively easy)