The 1000 Genomes Project

Goncalo Abecasis

Project Goals

- >95% of accessible genetic variants
  with a frequency of >1%
  in each of multiple continental regions

- Extend discovery effort to lower frequency variants in coding regions of the genome

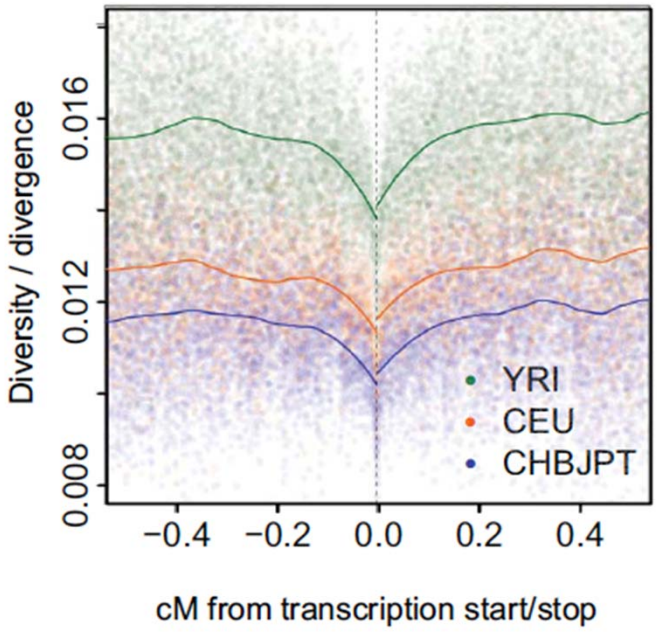
- Define haplotype structure in the genome

Pilot Projects Completed

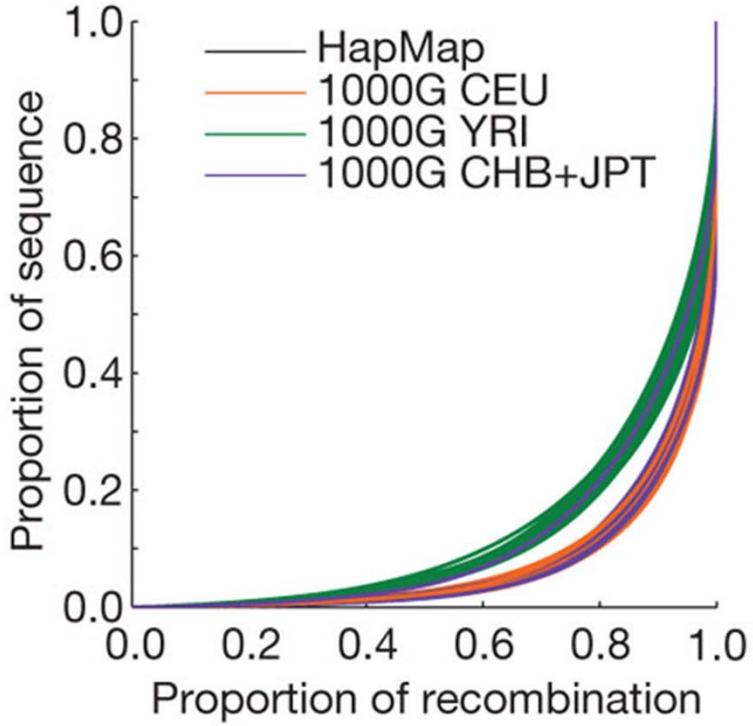
- 2 deeply sequenced trios
- 179 whole genomes sequenced at low coverage
- 8,820 exons deeply sequenced in 697 individuals

- 15M SNPs, 1M indels, 20,000 structural variants

Population Genetic Insights



Highlights Reduced Diversity
Extending ~120kb Around Genes



Better Localization
Of Recombination Hotspots

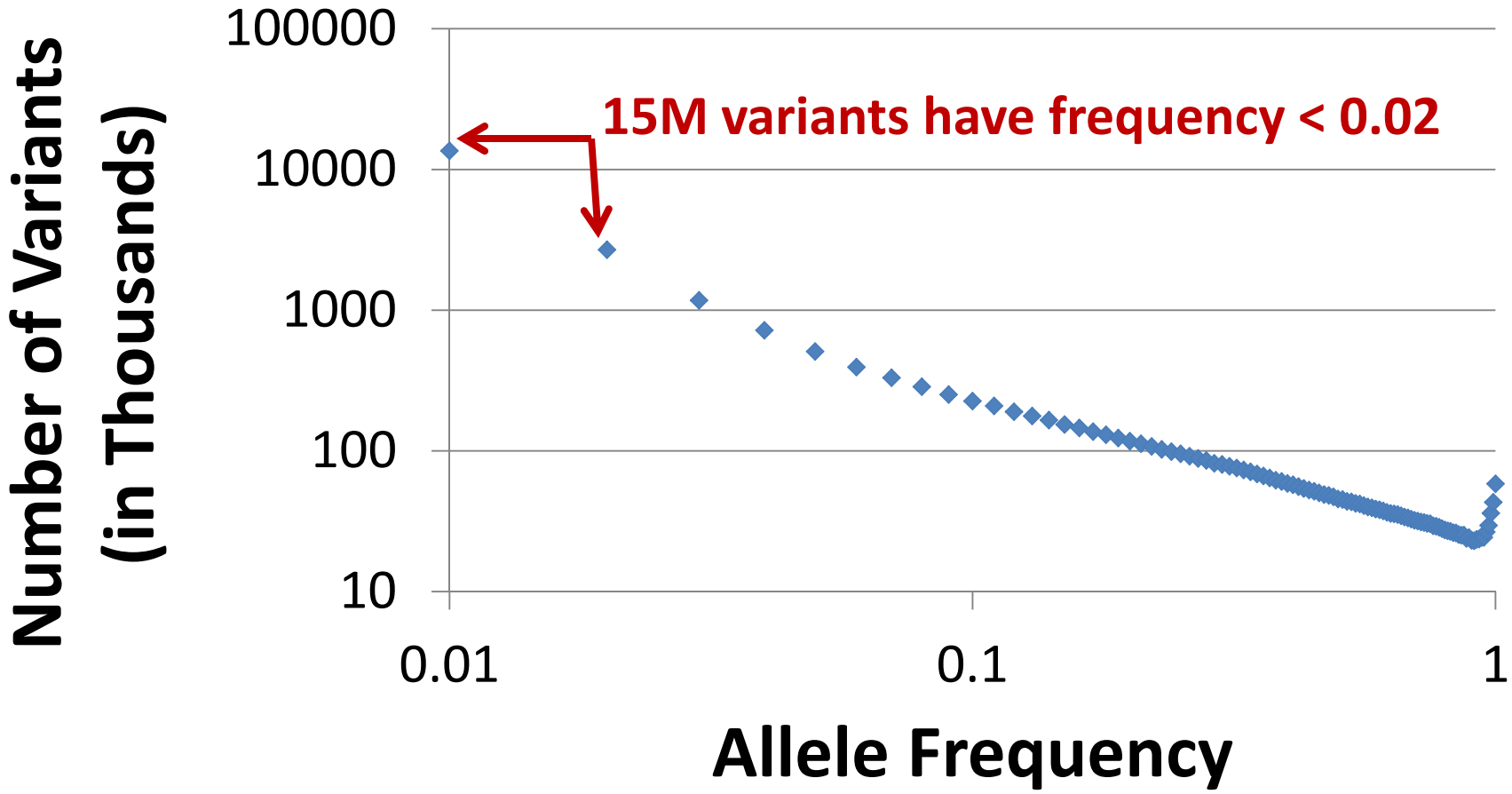
Current Status

- 25,487,060 variant sites called on 629 samples
  - 7,922,125 sites in dbSNP 129
  - 17,564,935 sites not in dbSNP 129
  - 98.8% of HapMap III sites rediscovered
  - Transition/transversion ratio of 2.21 vs 2.04 in Pilot

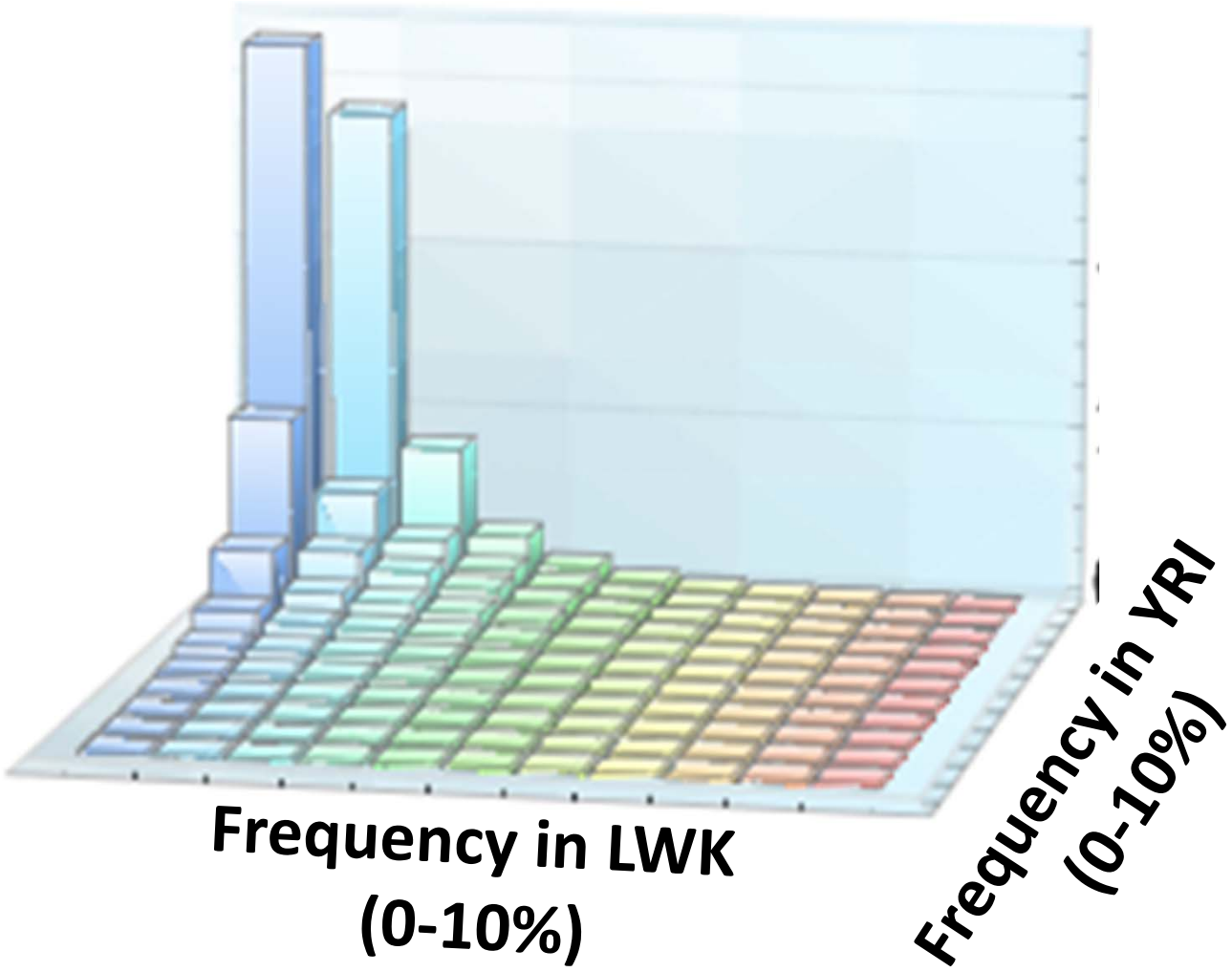
- As of November 2010:
  - 1103 sequenced samples
  - 22.6 Tb of raw sequence data

- ftp://ftp.1000genomes.ebi.ac.uk
- ftp://ftp.ncbi.nlm.nih.gov/1000genomes/

Frequency Spectrum

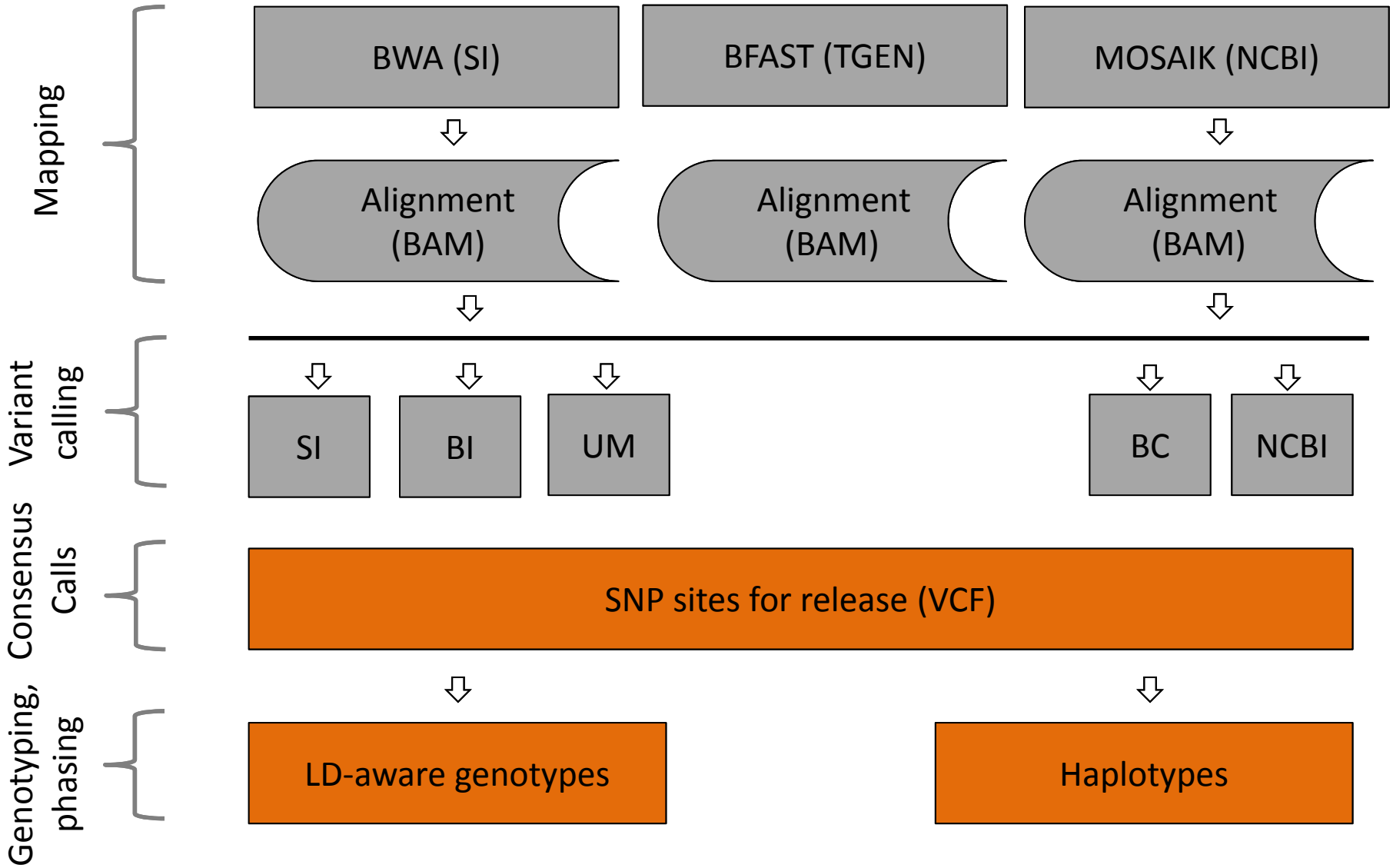


Most Low Frequency Variants Are Population Specific



Simon Gravel

Data processing / variant calling

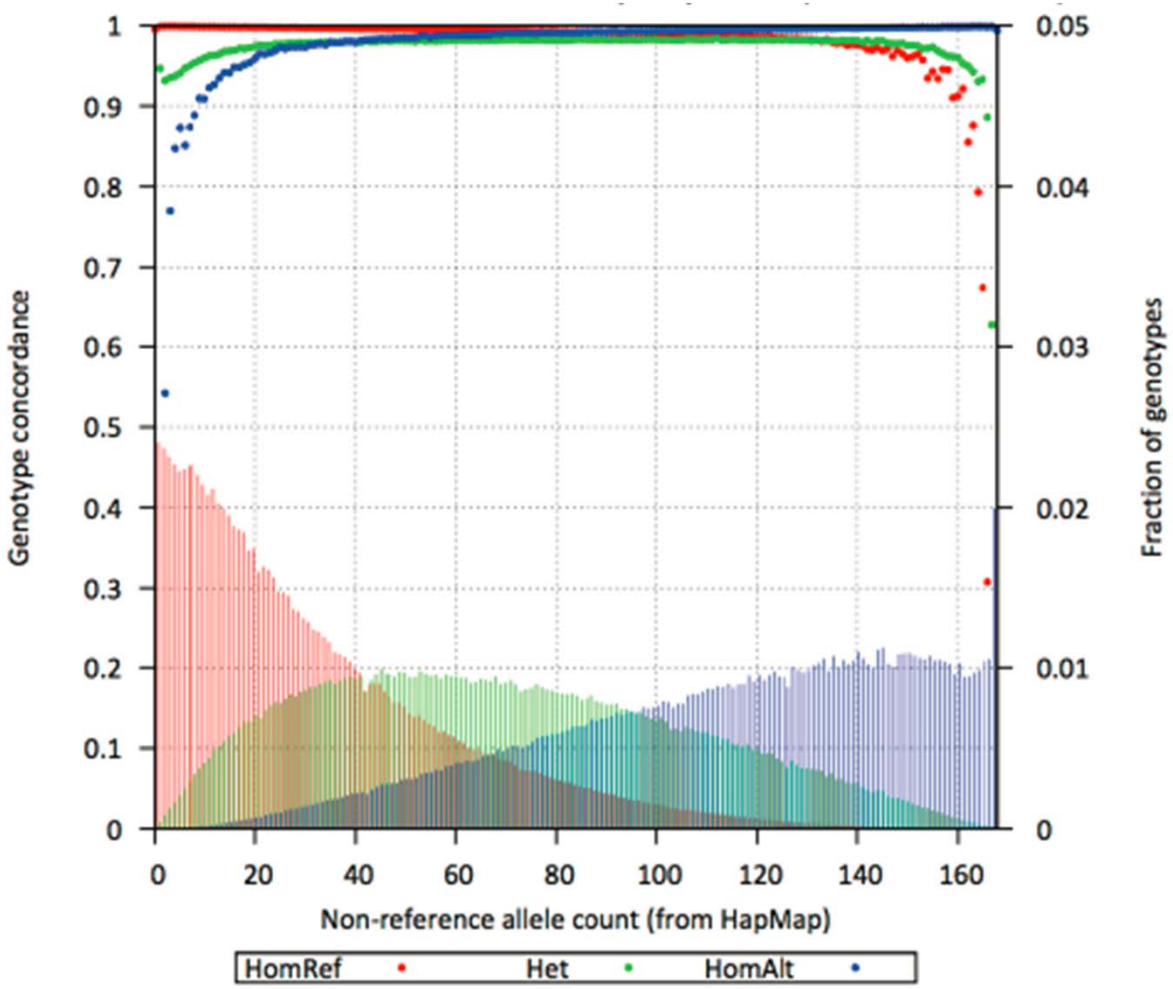


Genotype Discordance Rates

Panel	Number of Sites (All Populations)	Discordance With Chip When …		
		HapMap 3 Homozygote For Reference	HapMap 3 Heterozygote	HapMap 3 Homozygote Non-Reference
1000G August 2010	630,338 (chr. 20) 27M (genomewide)	0.43%	2.09%	0.99%
1000G November 2010	721,250 (chr. 20) 32M (genomewide)	0.20%	1.12%	0.58%

Accuracy of genotypes estimated from low coverage data for 1000 Genomes data freezes date August 2010 (~600 individuals) and November 2010 (~1000 individuals). Samples from each continental ancestry group were analyzed separately.

Accuracy of Low Pass Genotypes



Genotype accuracy for rare genotypes is lowest, but definition of rare changes as more samples are sequenced.

Hyun Min Kang

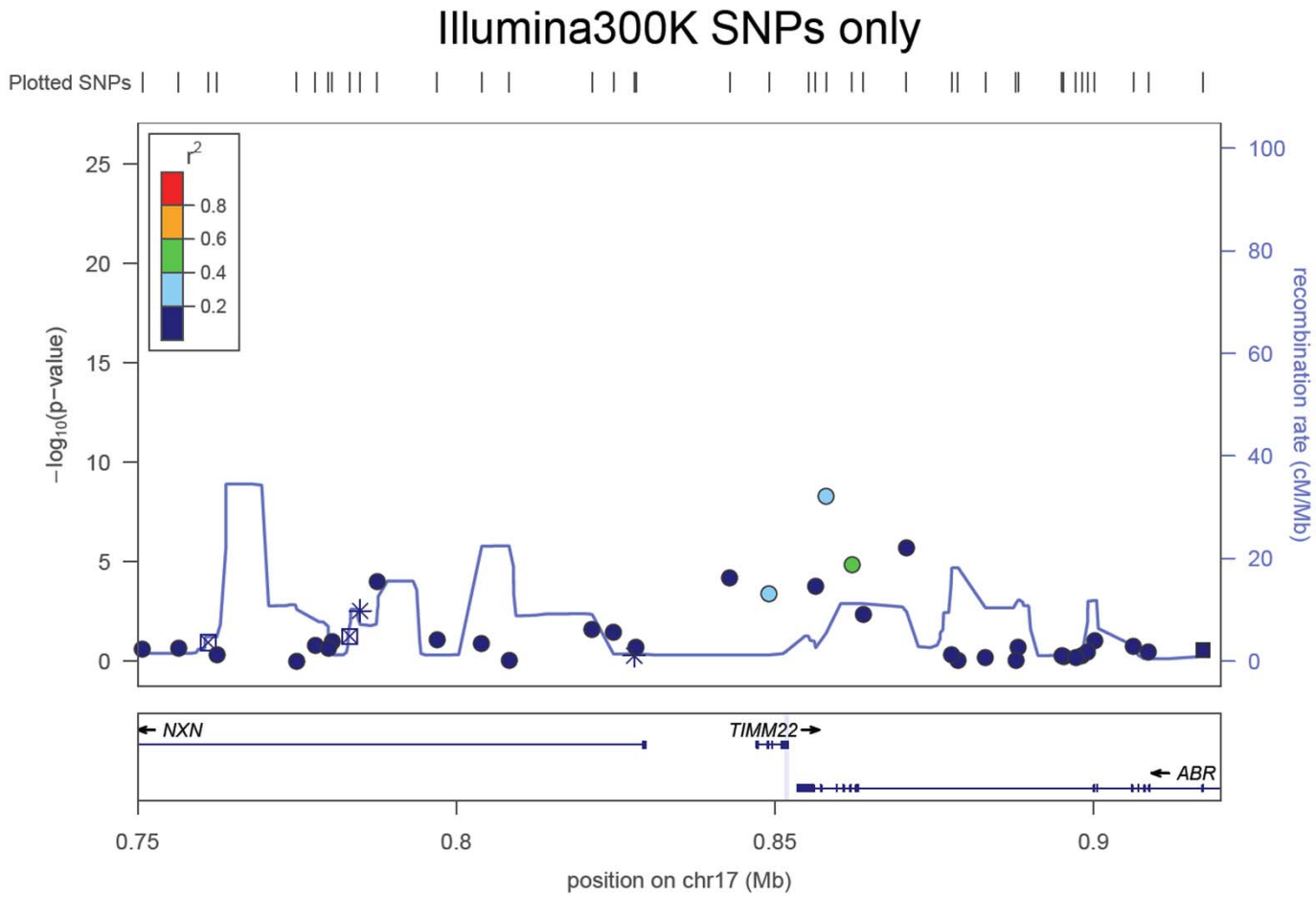
Current Plans

- Phase I Data Freeze in November 2011
  - >1,100 low coverage samples
  - Full set of variation analysis (structural variation, indels, SNPs)
- Genotyping of all Phase I Samples
  - Data generation complete in December 2010
- Exome Resequencing of Phase I Samples
  - Data generation complete in February 2011
- Phase II will result in >2,000 sequenced samples
  - Whole genome low pass sequencing
  - Deep exome resequencing
  - Genotyping of up to 10M variants
  - Planned for 2011

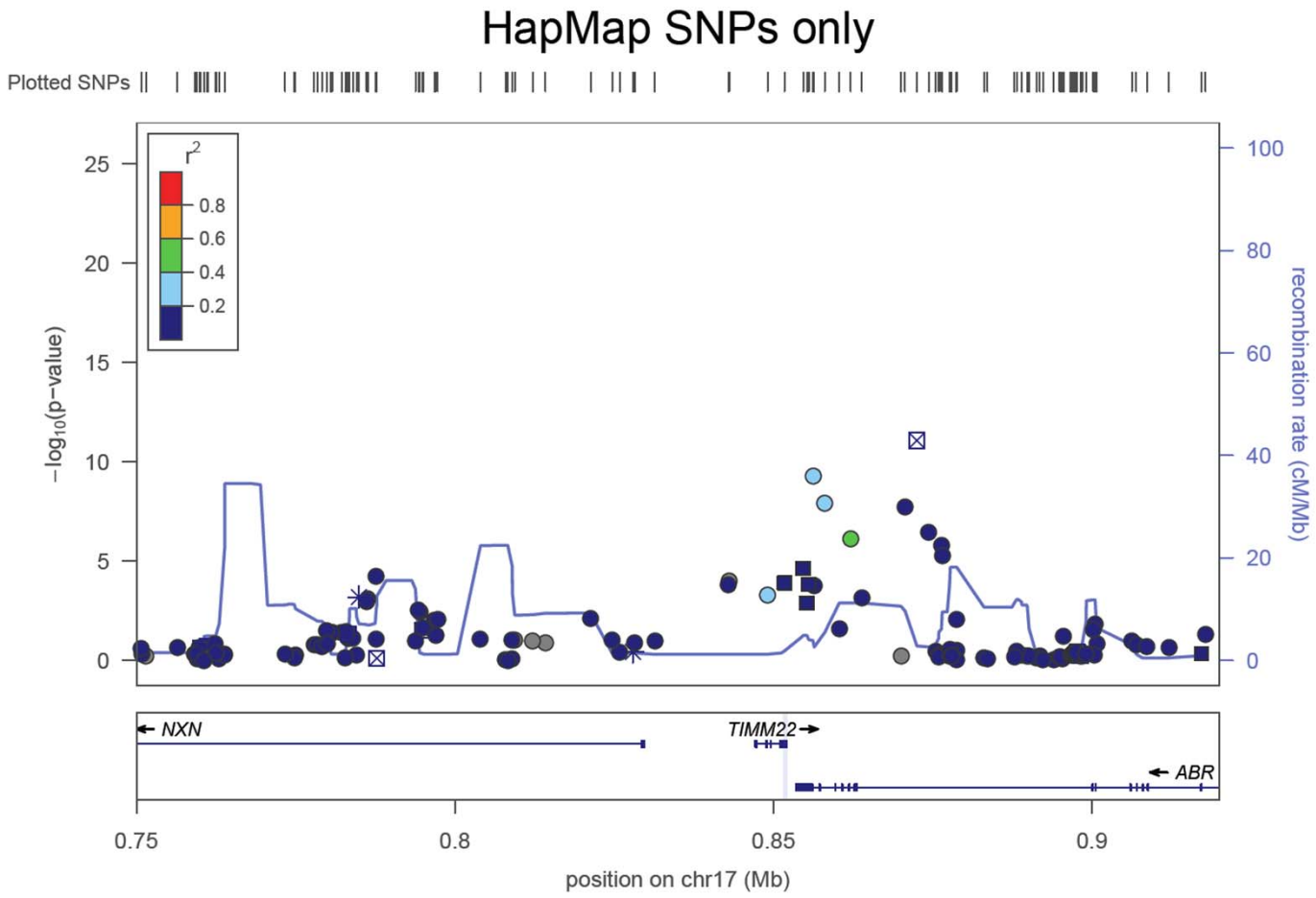
Uses of Project Data

- Identify private variants in resequencing experiments
  - Catalog includes ~95% of coding variants in any individual
- Identify interesting variants in regions of interest and genomewide
  - Supporting the design of next generation arrays
- Genotype imputation into existing association studies
- >80 abstracts directly reference project data

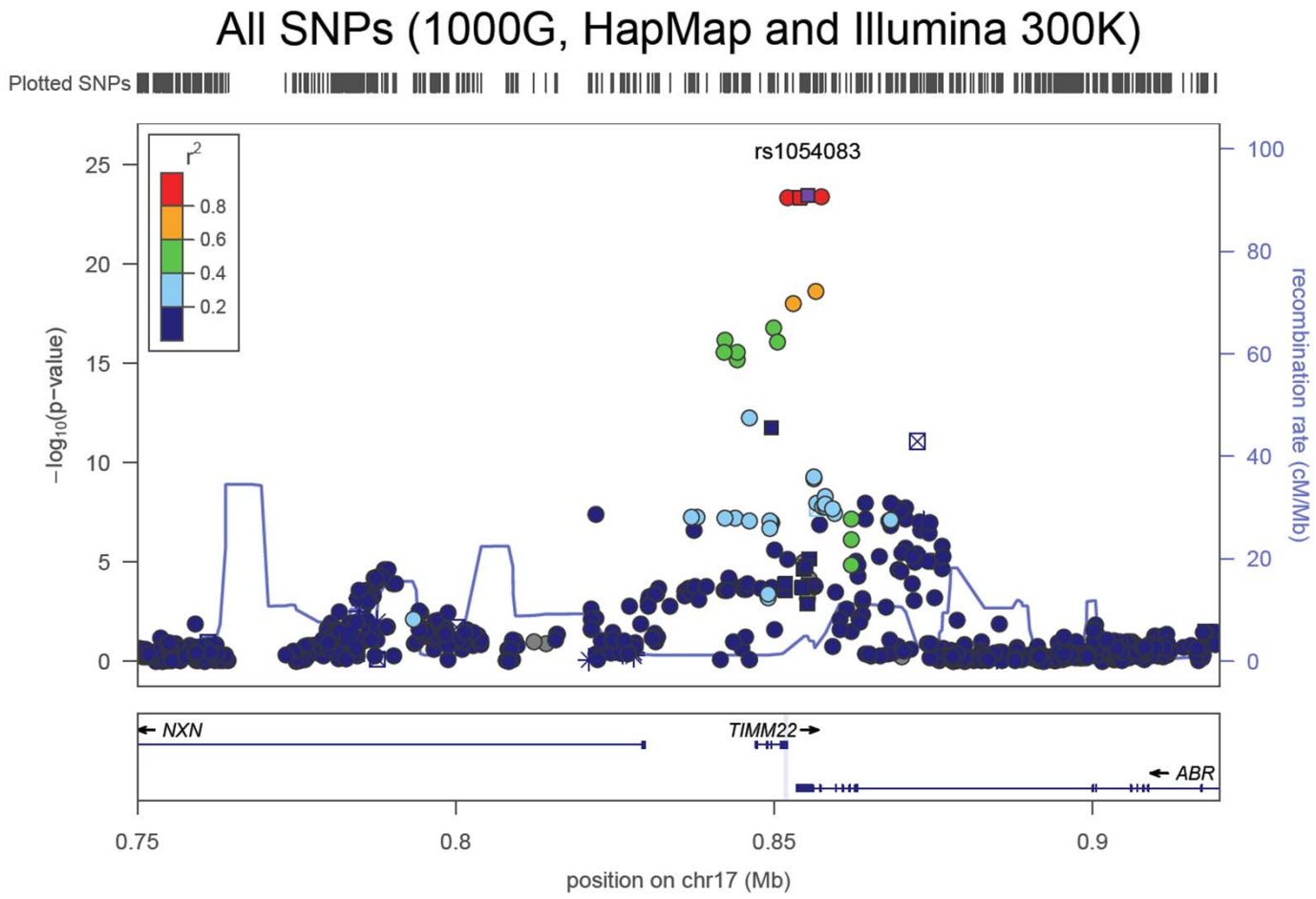
Enhance Association Studies: eQTL Imputation Example



Enhance Association Studies: eQTL Imputation Example



Enhance Association Studies: eQTL Imputation Example



Imputation Accuracy

Reference Panel	Imputation Accuracy (r^2)		
	MAF 1-3%	MAF 3-5%	MAF >5%
1000G Pilot March 2011 (60 Europeans)	0.69	0.77	0.91
1000G August 2011 (283 Europeans)	0.73	0.78	0.92
1000G November 2011 (563 Europeans)	0.83	0.85	0.94

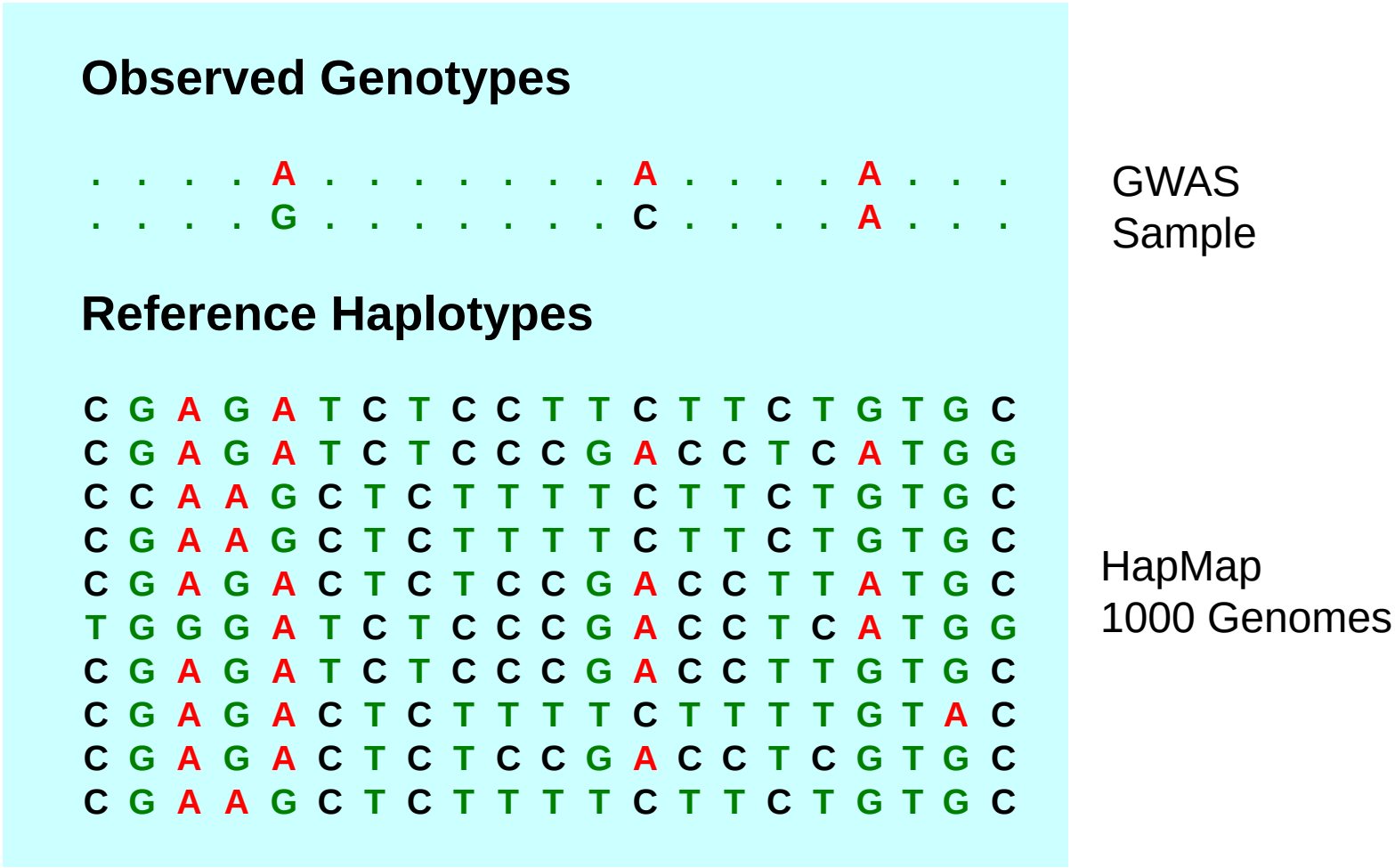
We evaluate the accuracy of imputation from low coverage sequence data using different iterations of the 1000 Genome Project haplotypes – specifically, those generated for the pilot paper published in Nature (which included 60 European individuals) and two more recent updates. It is clear that, particularly for rare alleles, imputation improves as the reference panel increases in size. Accuracy was evaluated in GWAS samples genotyped on 2 different whole genome arrays (from the Nair et al. psoriasis GWAS).

Practical Imputation In Large Samples

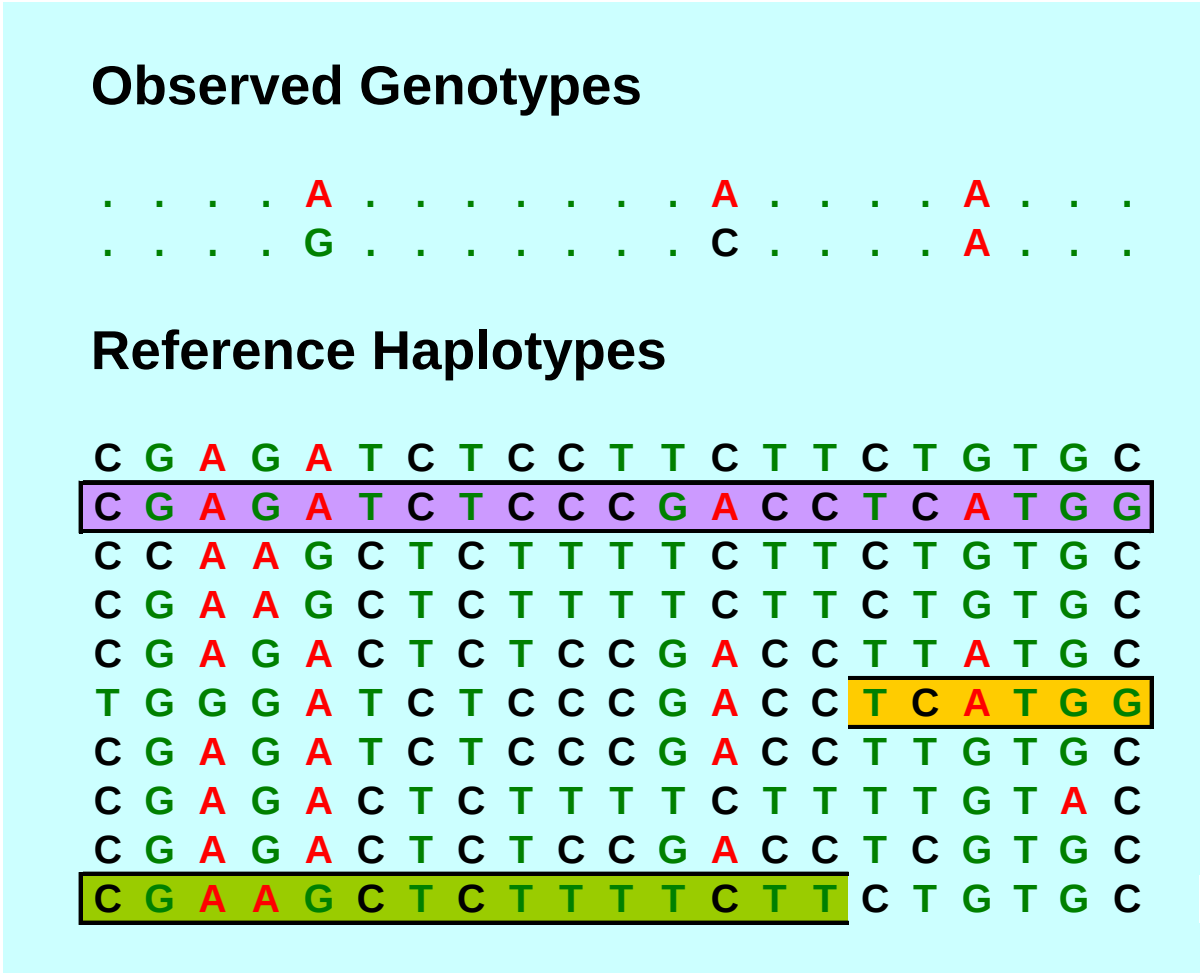
Minimac,

The Son of MaCH

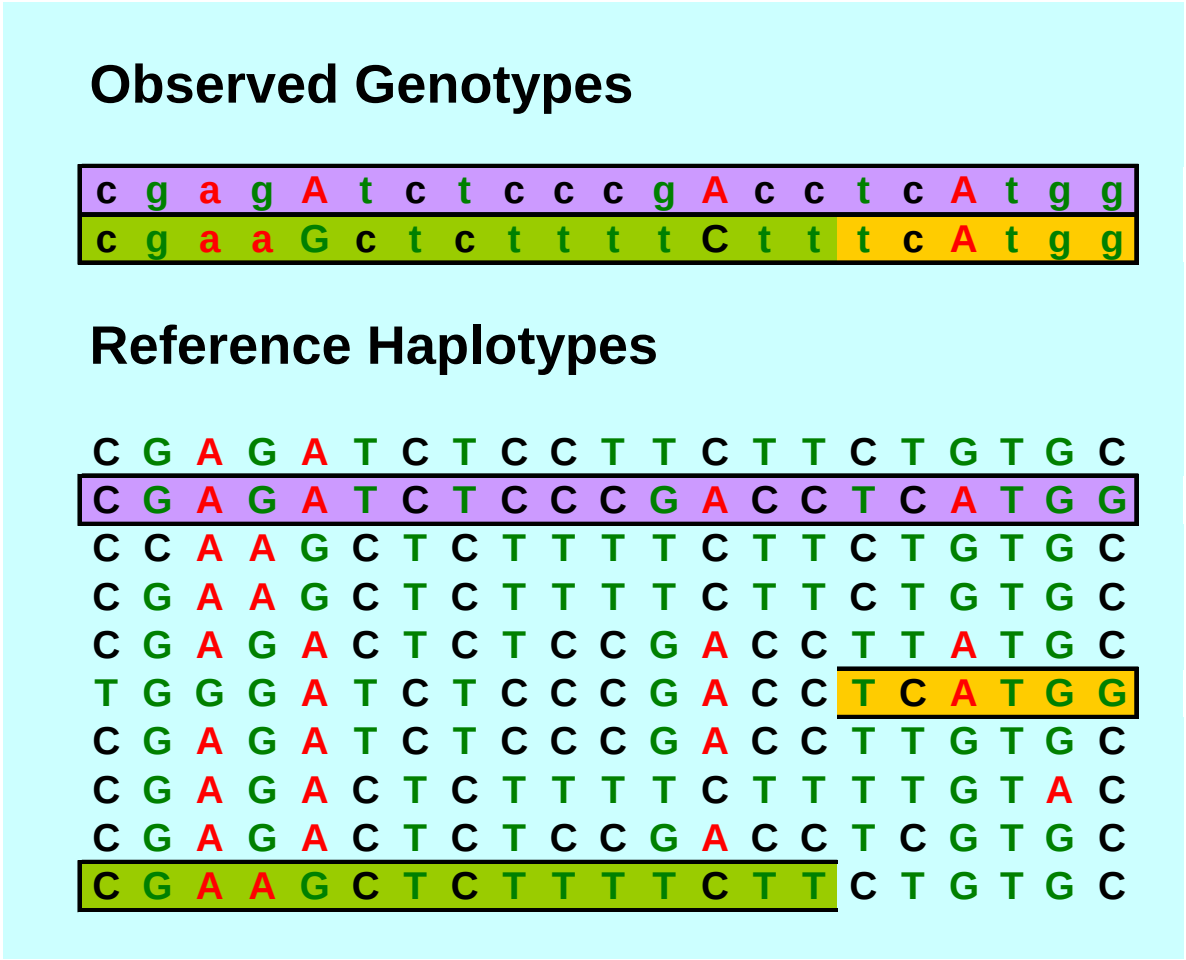
How Imputation Works



Identify Match Among Reference



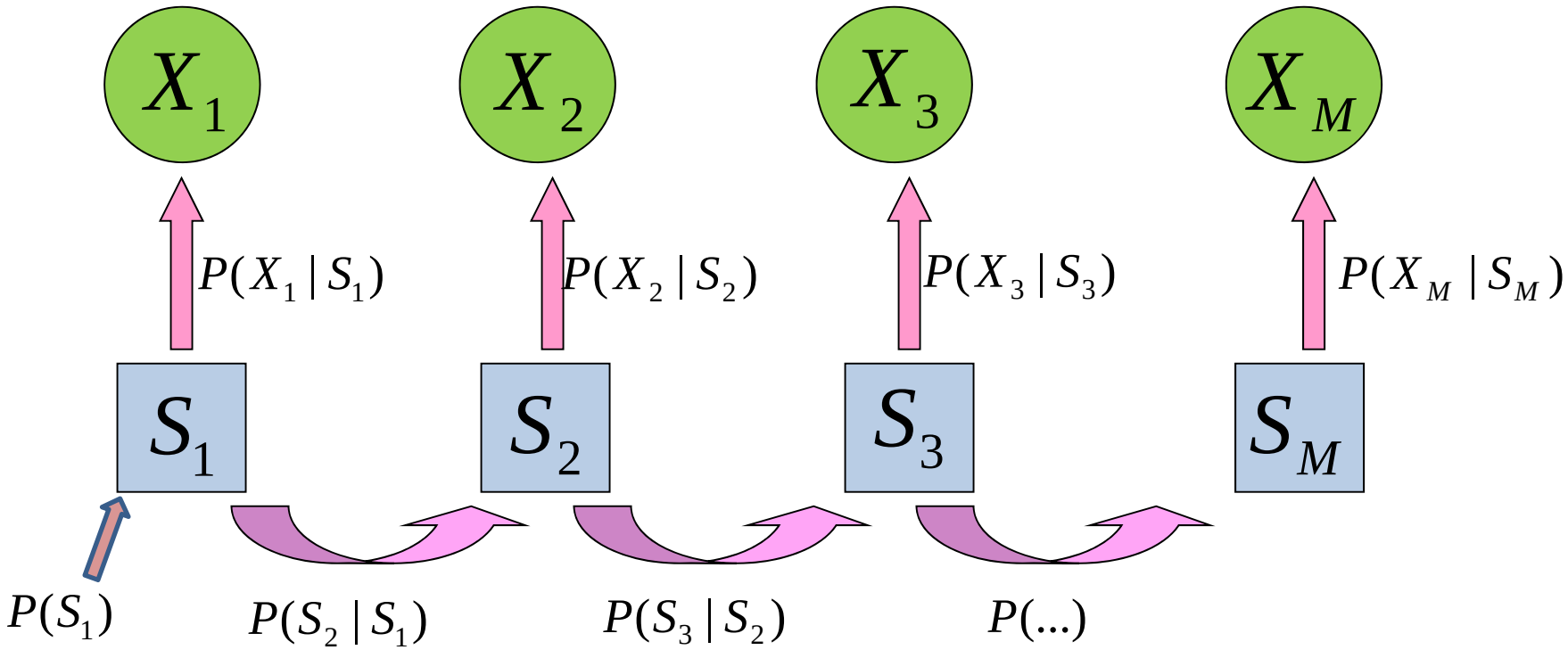
Phase Chromosome, Impute Missing Genotypes



Implementation

- Markov model is used to model each haplotype, conditional on all others
- At each position, we assume that each haplotypes being modeled copies a template haplotypes
- Each individual has two haplotypes, and therefore copies two template haplotypes
  - With H reference haplotypes …
  - … H(H+1)/2 options to consider per individual, per position

Markov Model



Number of states to be considered increases exponentially with panel size …

Does this work?

- 90 GAIN psoriasis study samples were re-genotyped for 906,600 SNPs using the Affymetrix 6.0 chip.

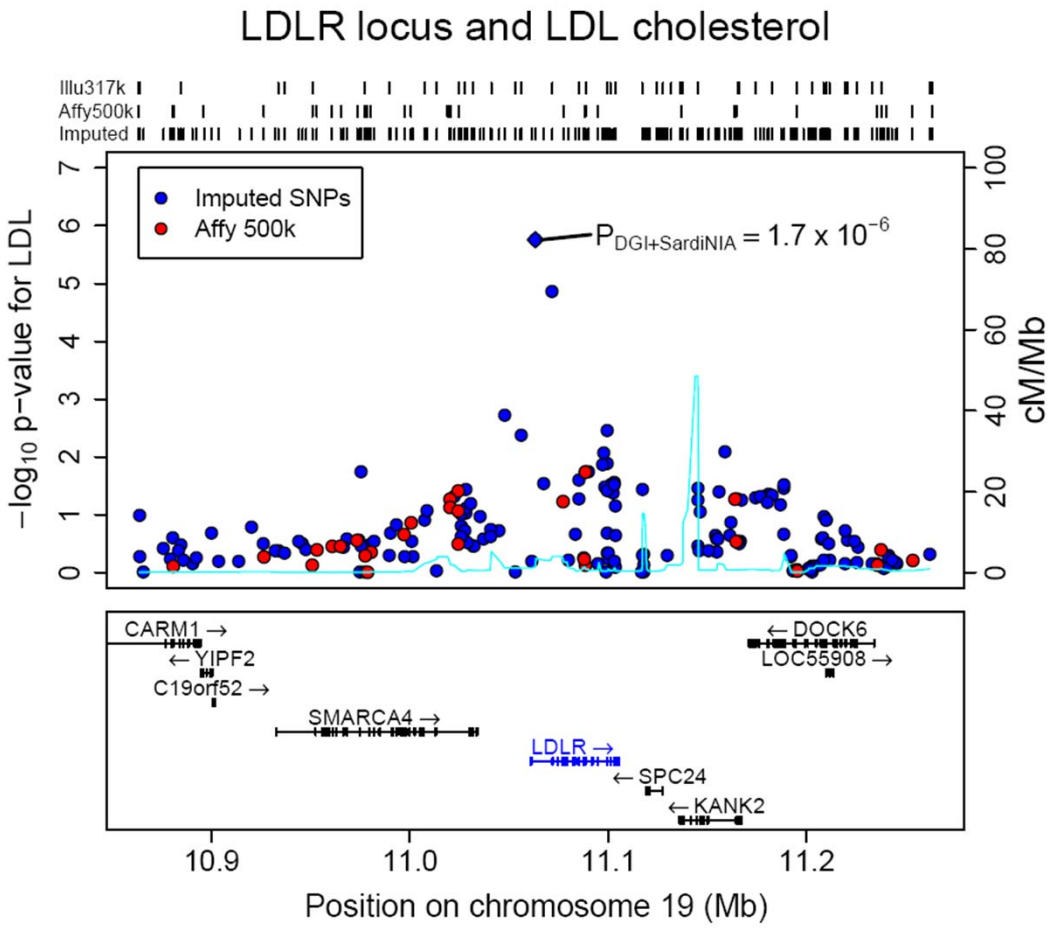
- Comparison of 15,844,334 genotypes for 218,039 SNPs that overlap between the Perlegen and Affymetrix chips resulted in discrepancy rate of 0.25% per genotype (0.12% per allele).

- Comparison of 57,747,244 imputed and experimentally derived genotypes for 661,881 non-Perlegen SNPs present in the Affymetrix 6.0 array resulted in a discrepancy rate of 1.80% per genotype (0.91% per allele).

- Overall, the average r^2 between imputed genotypes and their experimental counterparts was 0.93. This statistic exceeded 0.80 for >90% of SNPs.

Nair et al, *Nature Genetics*, 2009

Is imputation worth it? LDLR and LDL example



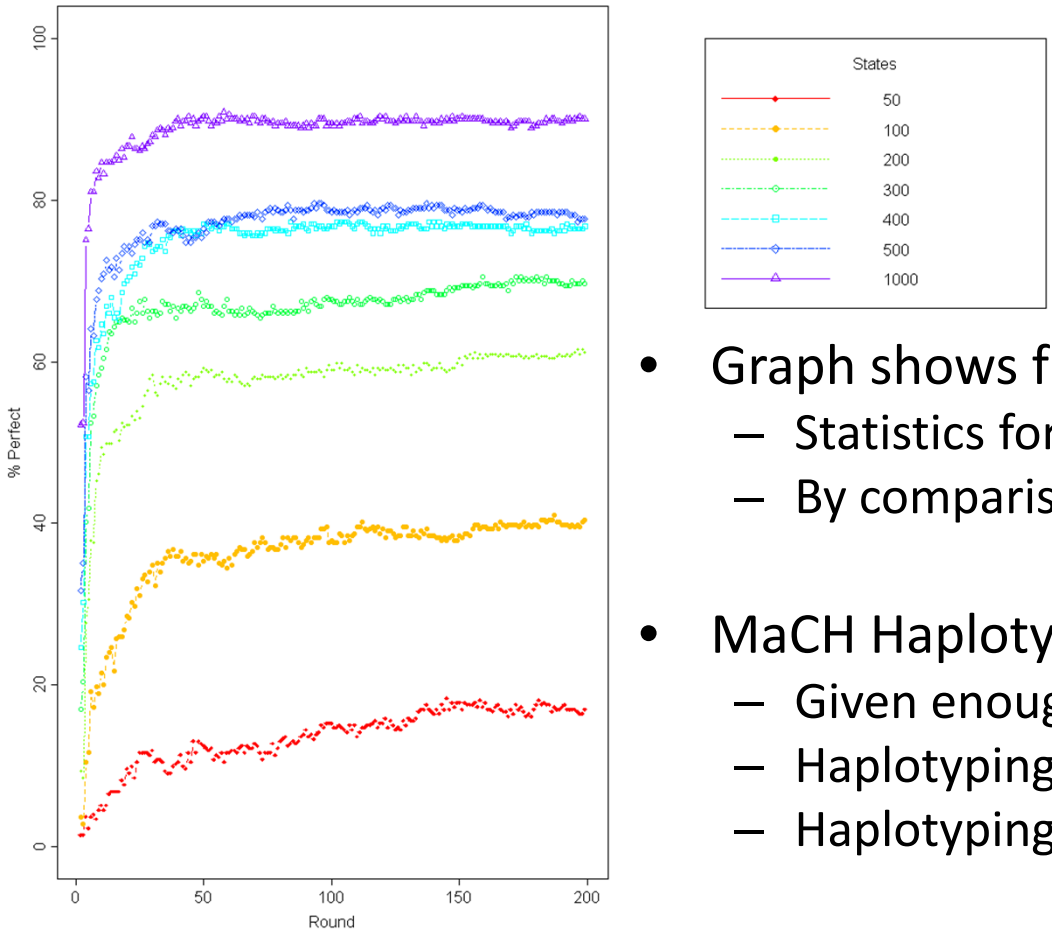
Imputing With Large Reference Panels…

What happens when we move from 100 to 1000 reference haplotypes??

How To Impute With Large Panels?

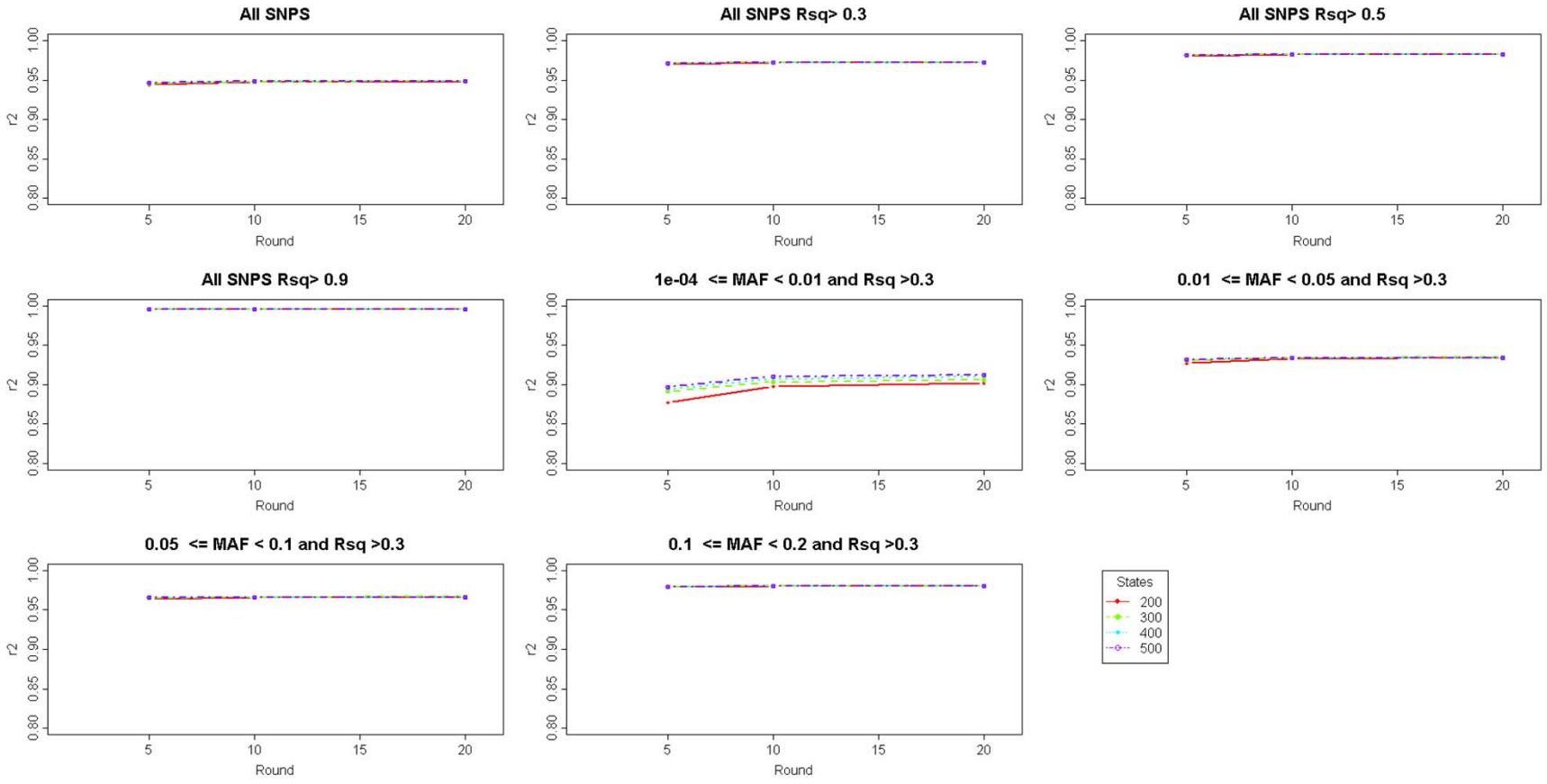
- Split imputation into two separate steps
- In the first step, we estimate haplotypes for each sample
- In the second step, we impute missing genotypes into each haplotype
  - This step now has time proportional to H (vs. H^2)

Is Phasing Accurate?



- Graph shows fraction of correct haplotypes
  - Statistics for chr. 20, in set of 2500 Finns
  - By comparison with trio phasing
- MaCH Haplotyping can be great
  - Given enough time
  - Haplotyping rounds vary along X axis
  - Haplotyping states are indicated by color

MaCH vs Minimac Accuracy



Sorry for small print, but bottom line is that all lines are very similar.

How Fast Is Minimac?

- In 1 hour, it can impute:
  - 1,000,000 markers in
  - 1,000 samples using
  - 100 reference haplotypes

- Generally, run time in hours will be about
  - (Markers x Samples x Reference Haplotypes)*10^{-11}

- http://genome.sph.umich.edu/wiki/Minimac

Samples and ELSI Group

Leena Peltonen (co-chair) Sanger Institute
Bartha Knoppers (co-chair) University of Montreal
Aravinda Chakravarti (co-chair) Johns Hopkins
Gonçalo 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
Gonçalo 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
Huanming (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
Jun Wang [illegible]
[illegible] Wang [illegible]
[illegible] Wilson [illegible]

Data Flow Group (being formed)

Paul Flicek (co-chair) European Bioinforma
Stephen Sherry (co-chair) National Center
Ewan Birney European Bioinformatics Instit
Clive Brown Sanger Institute
David Dooling Washington University in St.
Richard Gibbs [illegible]
Sol Katzman [illegible]
Hoda Khouri [illegible]
Martin Shumway National Center for Biotechnology Information
Jun Wang Beijing Genomics Institute
George Weinstock Baylor College of Medicine
(Broad representative)

Funders

Alan Schafer Wellcome Trust
Francis Collins National Human Genome Research Institute
Lisa Brooks National Human Genome Research Institute
Audrey Duncanson Wellcome Trust
Adam Felsenfeld National Human Genome Research Institute
Mark Guyer National Human Genome Research Institute
Ruth Jamieson Wellcome Trust
[illegible]
[illegible]
[illegible]
[illegible] Peterson National Human Genome Research Institute
Anne Pierson National Human Genome Research Institute
Zhiwu Ren National Planning and Development Committee
Jian Wang Beijing Genomics Institute

Analysis Group

Gil McVean (co-chair) Oxford University
Gonçalo Abecasis (co-chair) University of Michigan
David Altshuler Broad / MGH / Harvard
Paul de Bakker Broad / BWH / 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 Conrad 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
Hoda Khouri National Center for Biotechnology Information
Eric Lander Broad Institute
Charles Lee Brigham and Women's Hospital
Guoqing Li Beijing Genomics Institute
Heng Li Sanger Institute
Ruiqiang Li Beijing Genomics Institute
Yingrui Li Beijing Genomics Institute
Yun Li University of Michigan
Jonathan Marchini 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
Shaun Purcell Broad / MGH / Harvard
Noah Rosenberg University of Michigan
Pardis Sabeti Broad / Harvard
Paul Sch[illegible] [illegible]
Steven [illegible] [illegible] Institute
Jonathan Sebat [illegible]oratory
Ste[illegible] [illegible]technology Information
Matthew Stephens University of Chi[illegible]o
Simon Tavaré University of S[illegible] California
Chris Tyler-Smith Sanger Institute
Jun Wang Beijing Genomics Institute
David Wheeler Baylor College of Medicine
Hongkun Zheng Beijing Genomics Institute

www.1000genomes.org