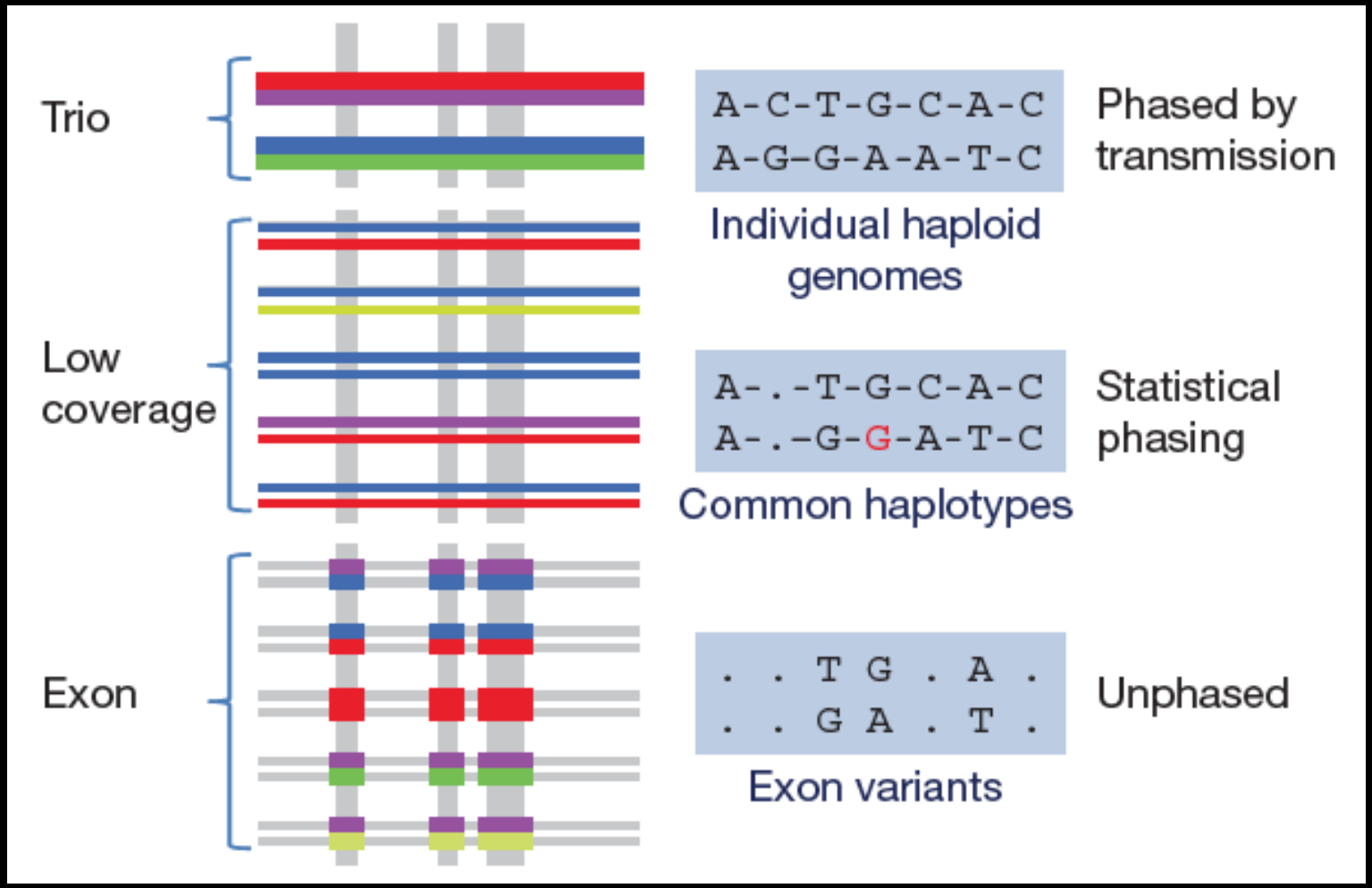
1000 Genomes Project: Datasets

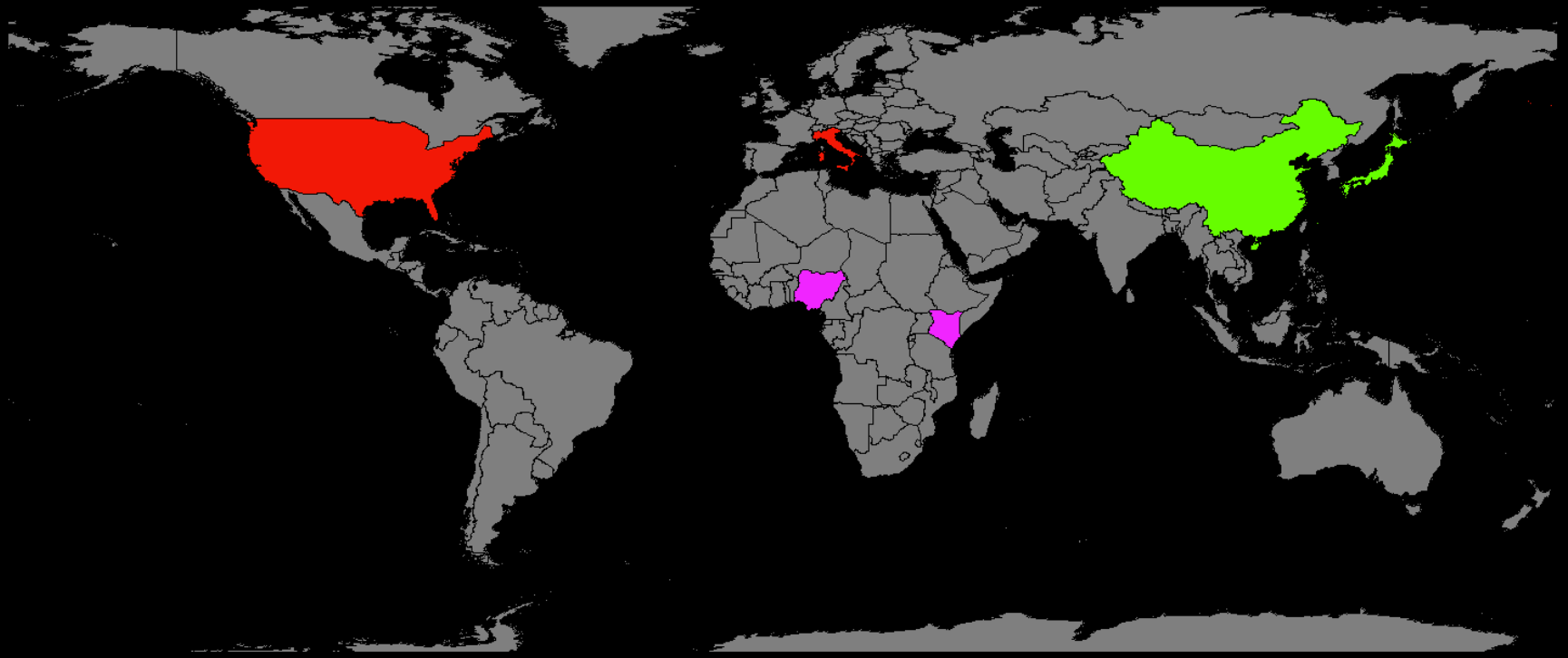
Gabor Marth
Boston College Biology Department

1000 Genomes Project Tutorial
ASHG 2010, Washington, DC
November 3, 2010

3 pilot coverage strategies



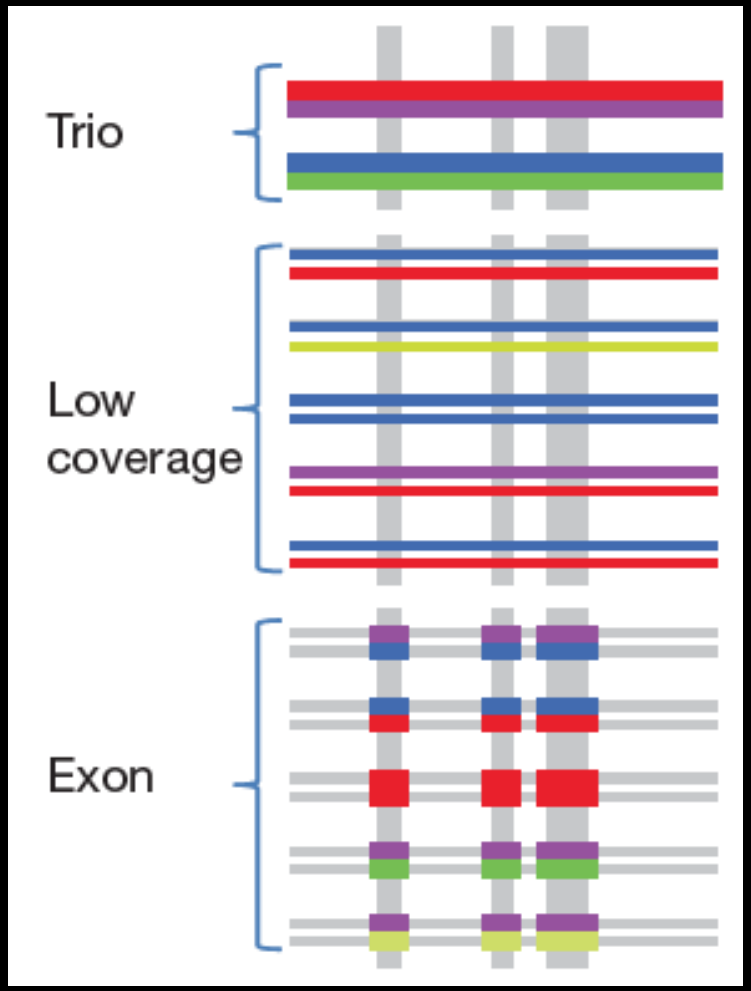
Samples



Population	YRI	LWK	CHB	CHD	JPT	CEU	TSI	All
Samples	112	108	109	107	105	90	66	697

1000 Genomes
A Deep Catalog of Human Genetic Variation

Pilot datasets



Populations	Samples	Coverage
2	6	20-40x
4	179	2-4x
7	697	20-50x

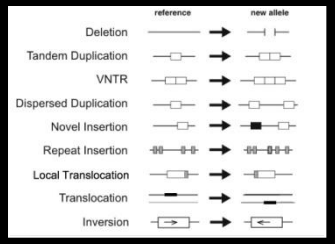
1000 Genomes
A Deep Catalog of Human Genetic Variation

Data processing / variant calling

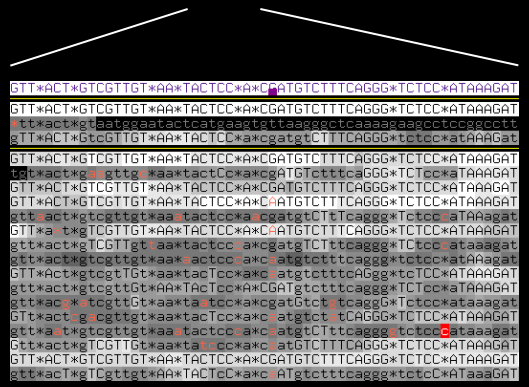
REF

IND

(ii) read mapping



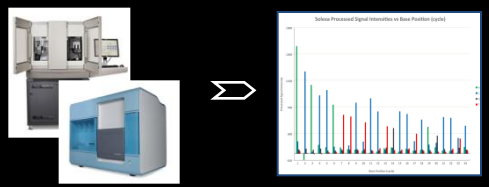
(iv) SV calling



(iii) SNP and short INDEL calling

IND

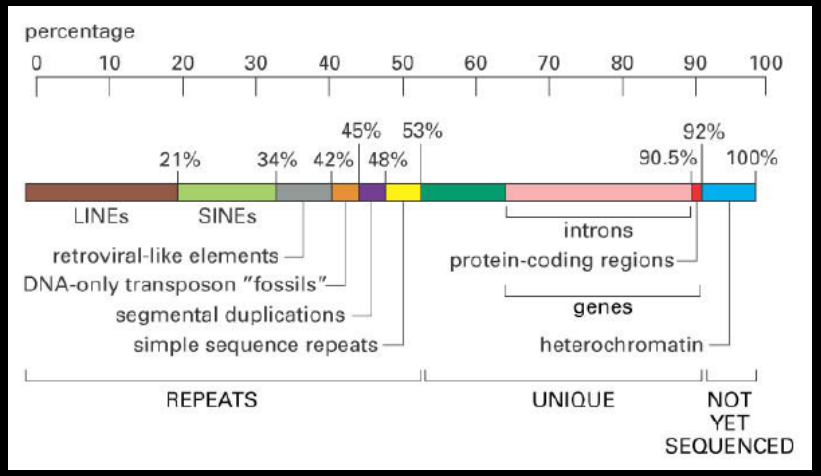
(i) base calling

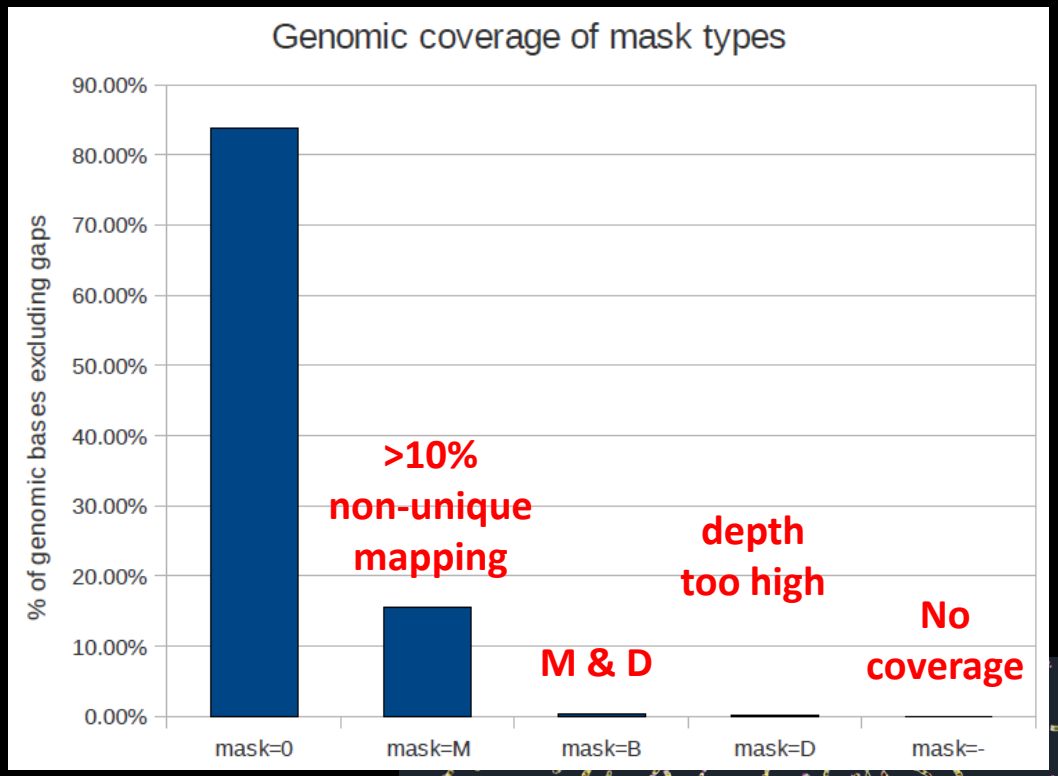


1000 Genomes

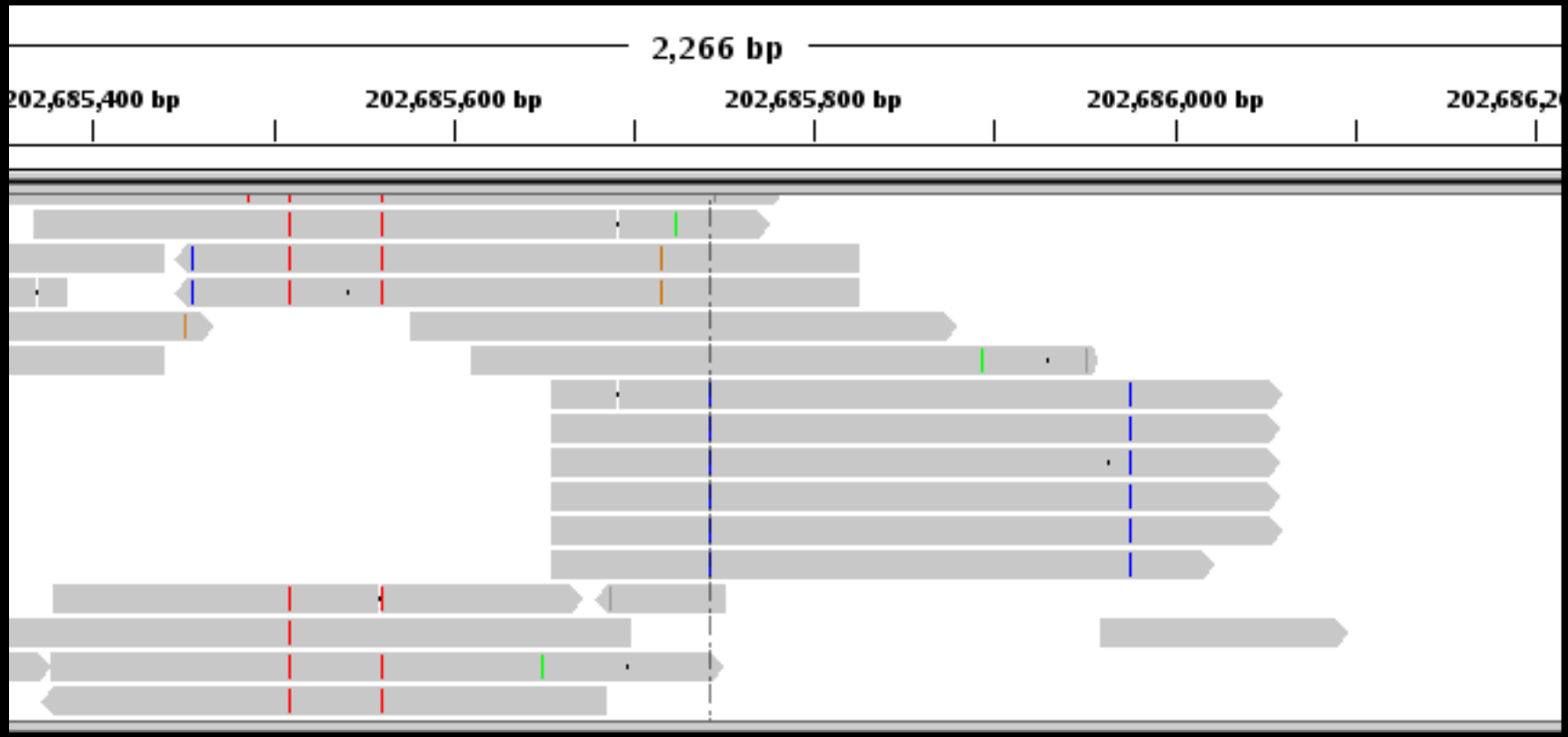
A Deep Catalog of Human Genetic Variation

>80% of the genome accessible with short reads

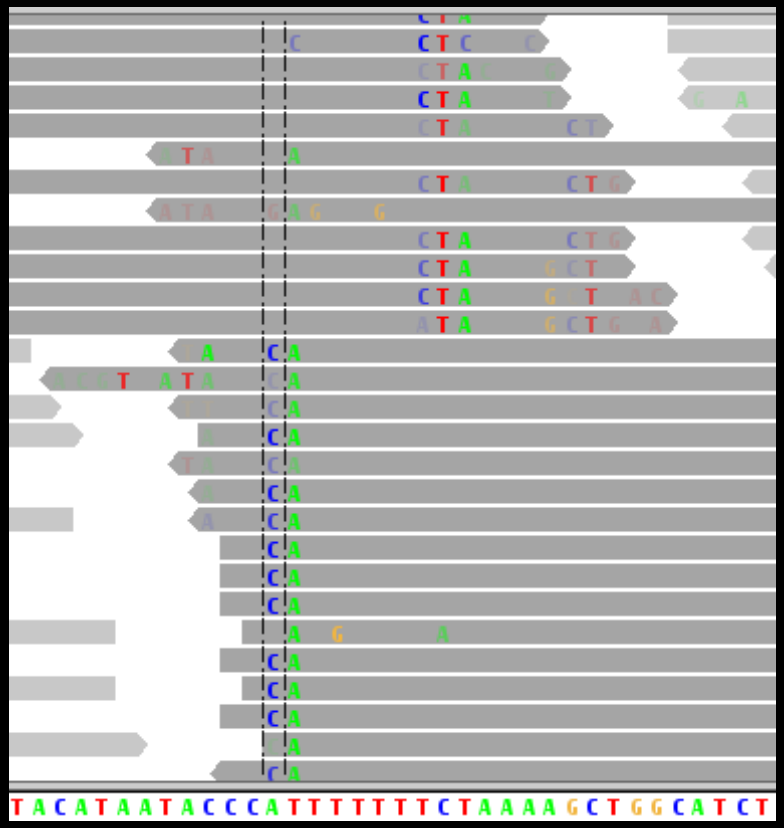


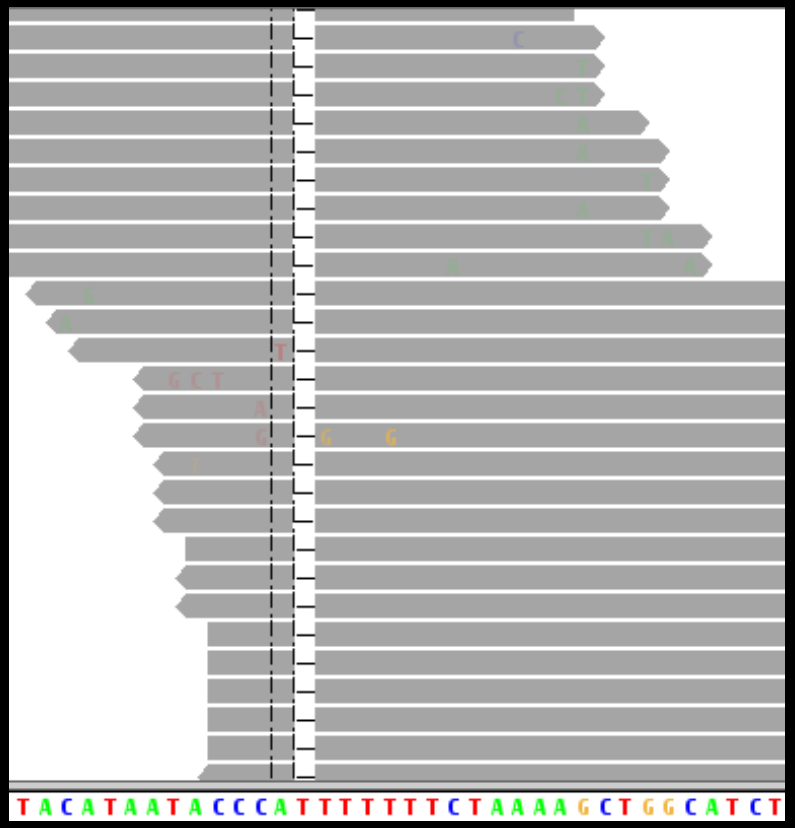


PCR-duplicate reads

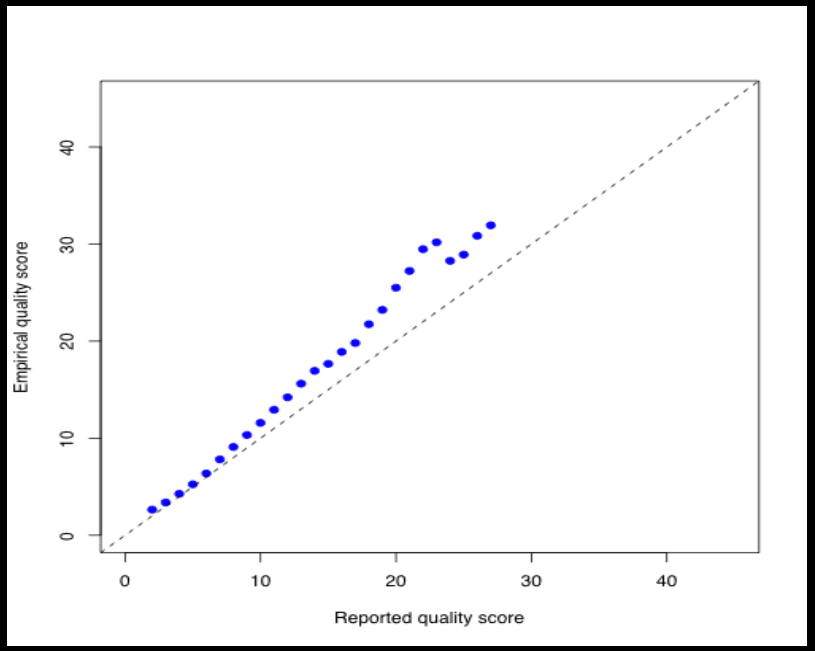


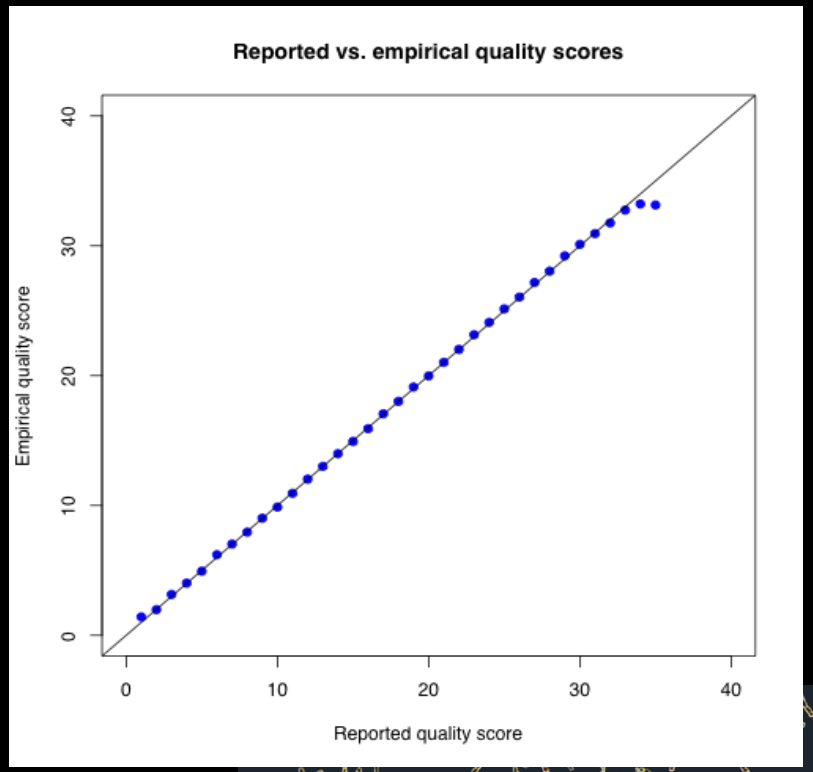
Locally misaligned bases



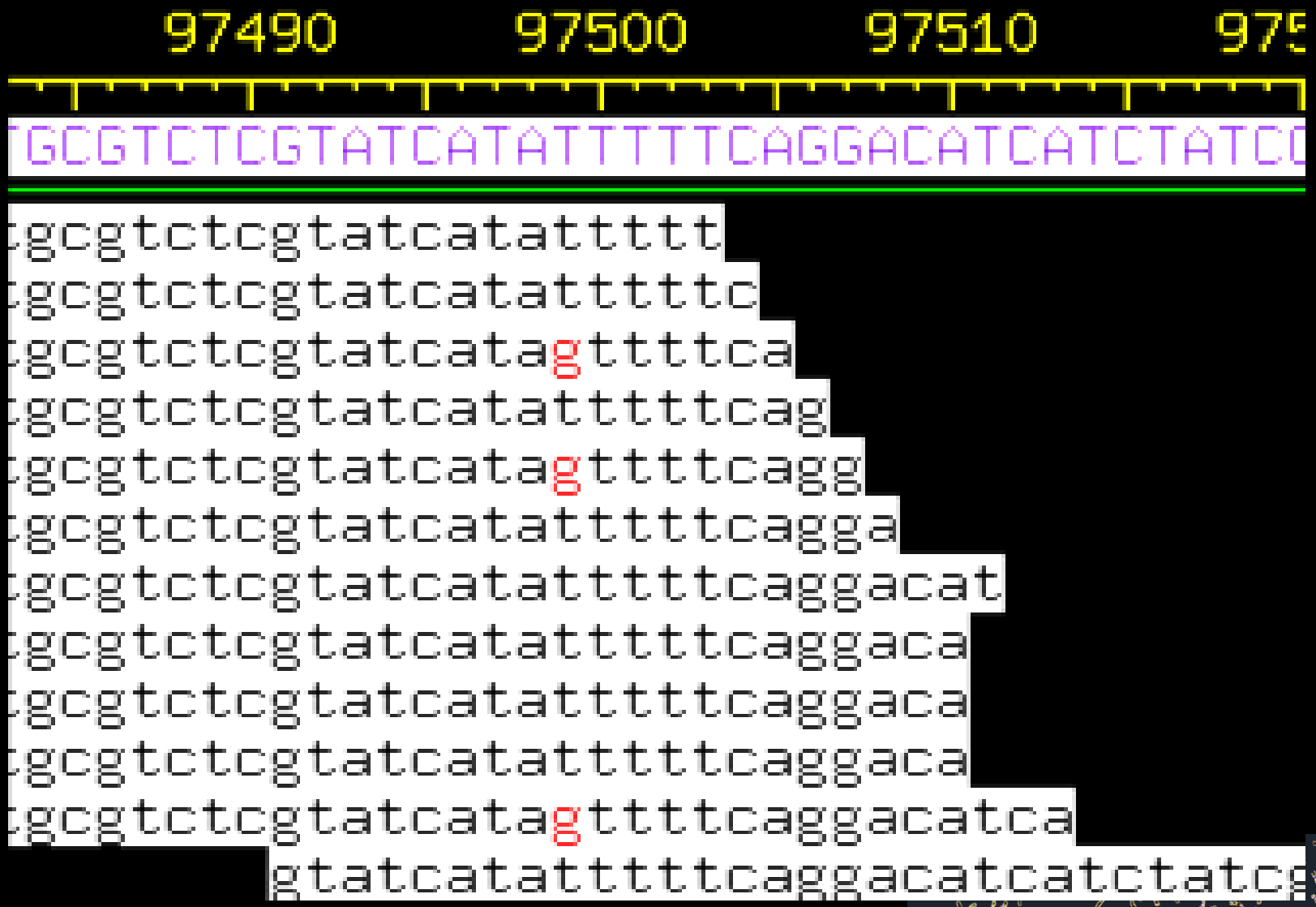


Un-calibrated base quality values





SNP calling


97490
97500
97510
97
TGCGTCTCGTATCATATTTTTCAGGACATCATCTATCG
tgcgtctcgtatcatattttt
tgcgtctcgtatcatatttttc
tgcgtctcgtatcatagttttca
tgcgtctcgtatcatatttttcag
tgcgtctcgtatcatagttttcagg
tgcgtctcgtatcatatttttcagga
tgcgtctcgtatcatatttttcaggacat
tgcgtctcgtatcatatttttcaggaca
tgcgtctcgtatcatatttttcaggaca
tgcgtctcgtatcatatttttcaggaca
tgcgtctcgtatcatagttttcaggacatca
gtatcatatttttcaggacatcatctatcg

SNP calling (continued)

a-----
a-----
c-----
-----c-----

a-----
a-----
a-----
-----a-----
-----c-----

c-----
c-----
c-----
-----c-----

$P(B_1=aacc|G_1=aa)$
$P(B_1=aacc|G_1=cc)$
$P(B_1=aacc|G_1=ac)$

$P(B_i=aaaac|G_i=aa)$
$P(B_i=aaaac|G_i=cc)$
$P(B_i=aaaac|G_i=ac)$

$P(B_n=cccc|G_n=aa)$
$P(B_n=cccc|G_n=cc)$
$P(B_n=cccc|G_n=ac)$

"genotype likelihoods"

$Prior(G_1,..,G_i,..,G_n)$

$P(G_1=aa|B_1=aacc; B_i=aaaac; B_n=cccc)$
$P(G_1=cc|B_1=aacc; B_i=aaaac; B_n=cccc)$
$P(G_1=ac|B_1=aacc; Bi=aaaac; B_n=cccc)$

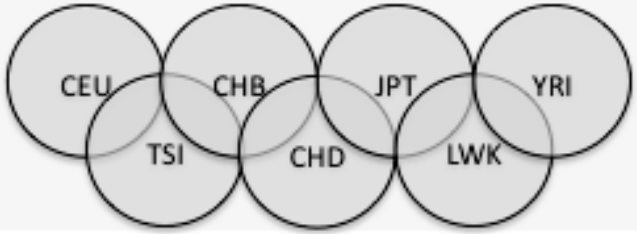
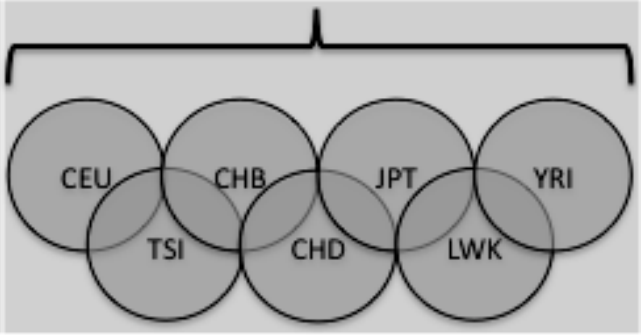
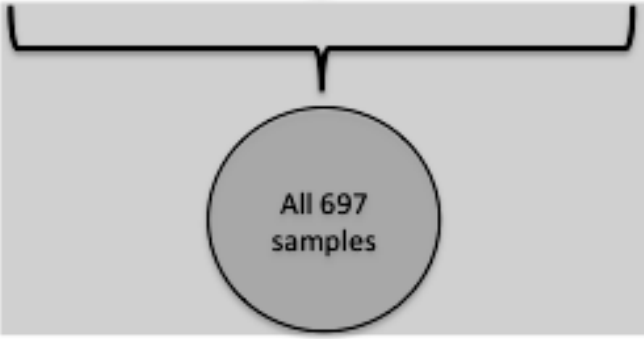
$P(G_i=aa|B_1=aacc; B_i=aaaac; B_n=cccc)$
$P(G_i=cc|B_1=aacc; B_i=aaaac; B_n=cccc)$
$P(G_i=ac|B_1=aacc; Bi=aaaac; B_n=cccc)$

$P(G_n=aa|B_1=aacc; B_i=aaaac; B_n=cccc)$
$P(G_n=cc|B_1=aacc; B_i=aaaac; B_n=cccc)$
$P(G_n=ac|B_1=aacc; Bi=aaaac; B_n=cccc)$

"genotype call"

"SNP call"

Data processing / variant calling pipeline

Processing step	BC	BI
Read mapping SW	MOSAIK	MAQ (SLX) + SSAHA2 (454)
Duplicate filtering SW	Picard MarkDuplicates (SLX) BCMMarkduplicates (454)	Picard MarkDuplicates (SLX) Picard MarkDuplicates (454)
BQ recalibration SW	GATK (SLX) None (454)	GATK (SLX) GATK (454)
SNP calling SW	GigaBayes	UnifiedGenotyper
SNP filtering	Based solely on **probabilities**: P(SNP), P(G)	Based on **probabilities**: P(SNP), QDP; **context**: Hrun, and **read counts**: AB
SNP calling		
SNP statistics		

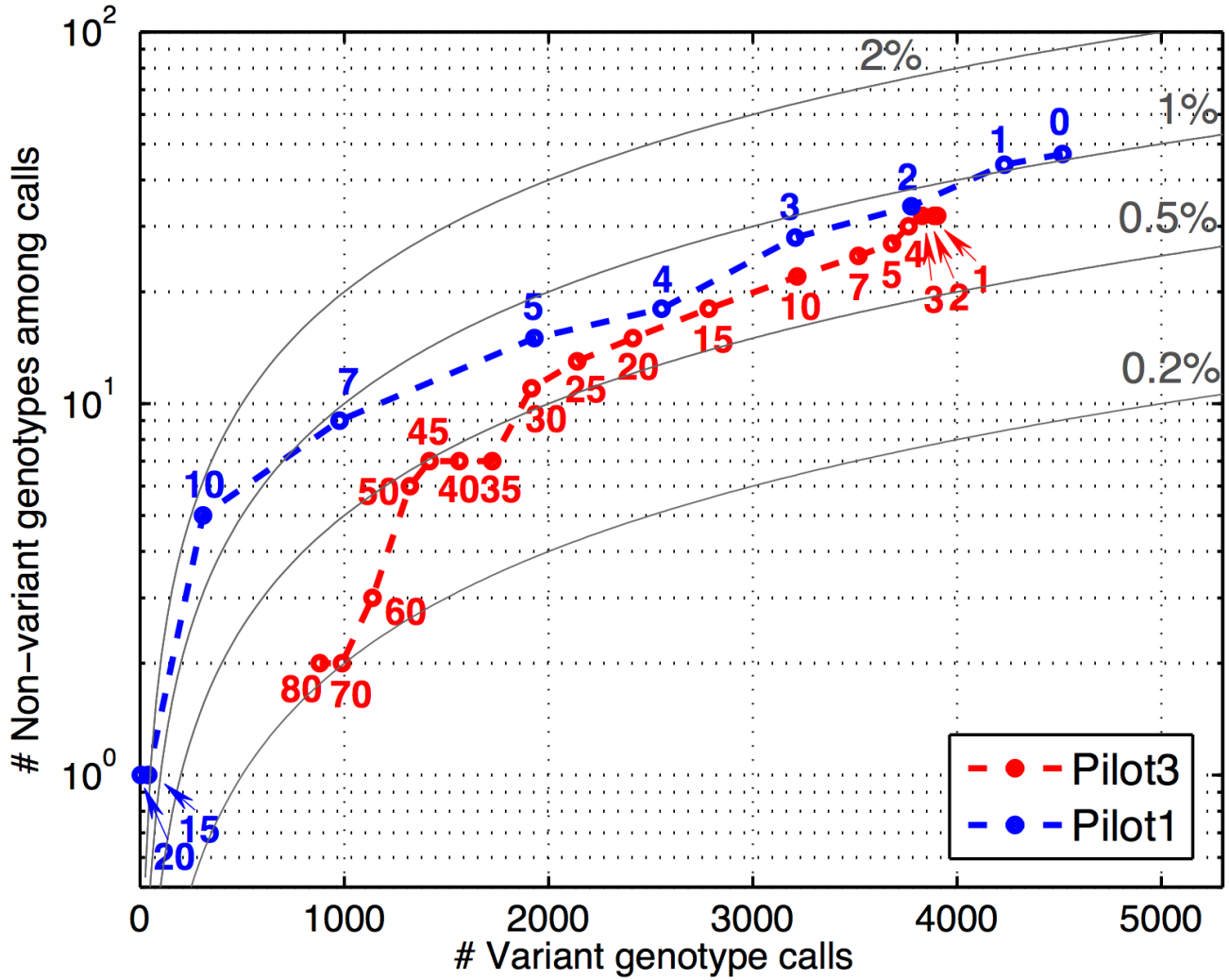
SNP calls from the 3 pilot datasets

	Trios	Low coverage	Exon pilot
Samples	6	179	697
Raw data	1.08 Tb	2.22 Tb	1.43 Tb
SNPs found	4.03M (CEU) 5.01M (YRI)	14.5M	12,761
% novel	15% (CEU) 29% (YRI)	55%	70%
Short indels	0.68 M	1.12 M	-
Deletions	~10,000	15,765	-
SV breakpts	6,169	9,092	-
Mobile element insertions	2,528	4,774	-

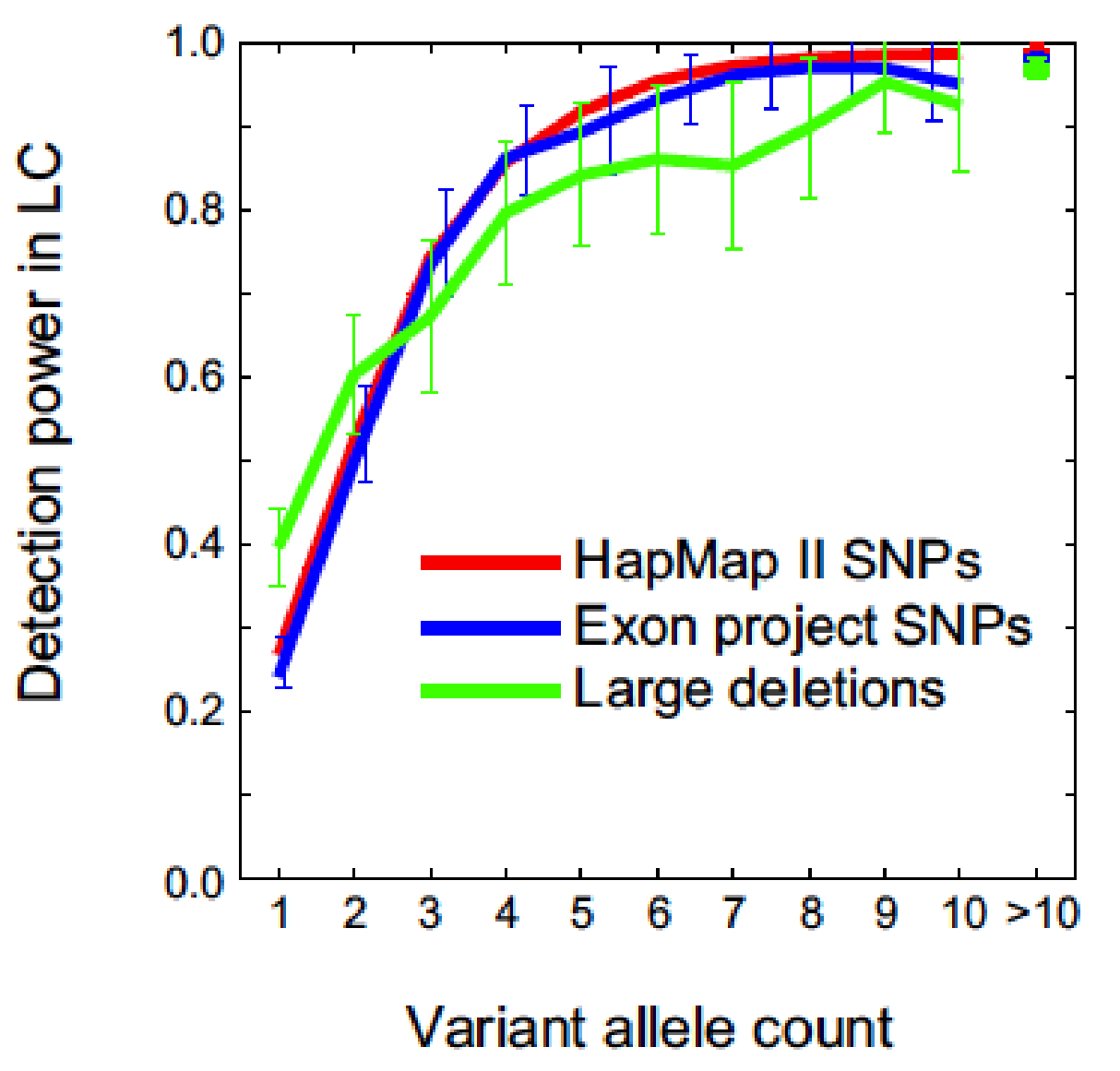
Validation by typing a random sample of novel variants

Overall FDR < 5% (10% for large SVs)

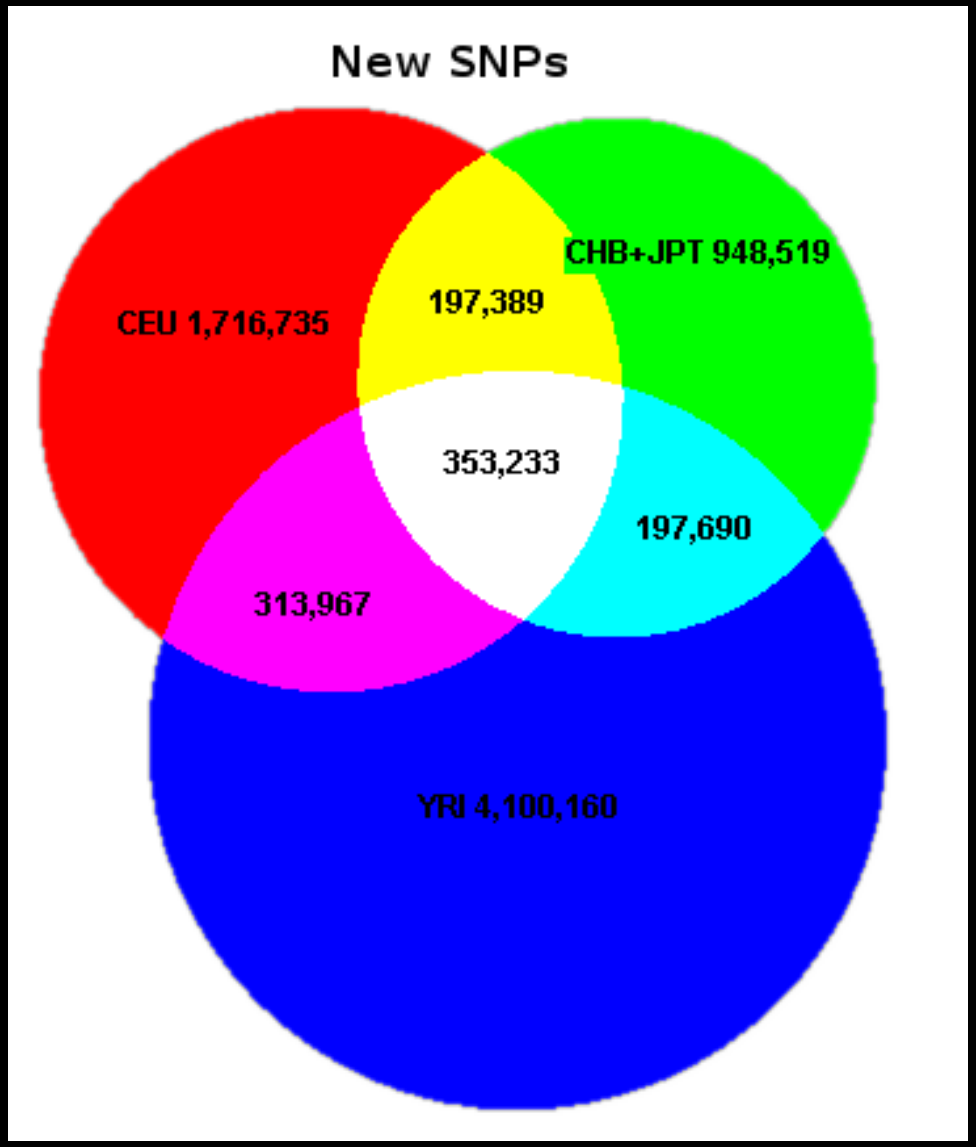
Imputation helps genotype calls



Power (sensitivity)



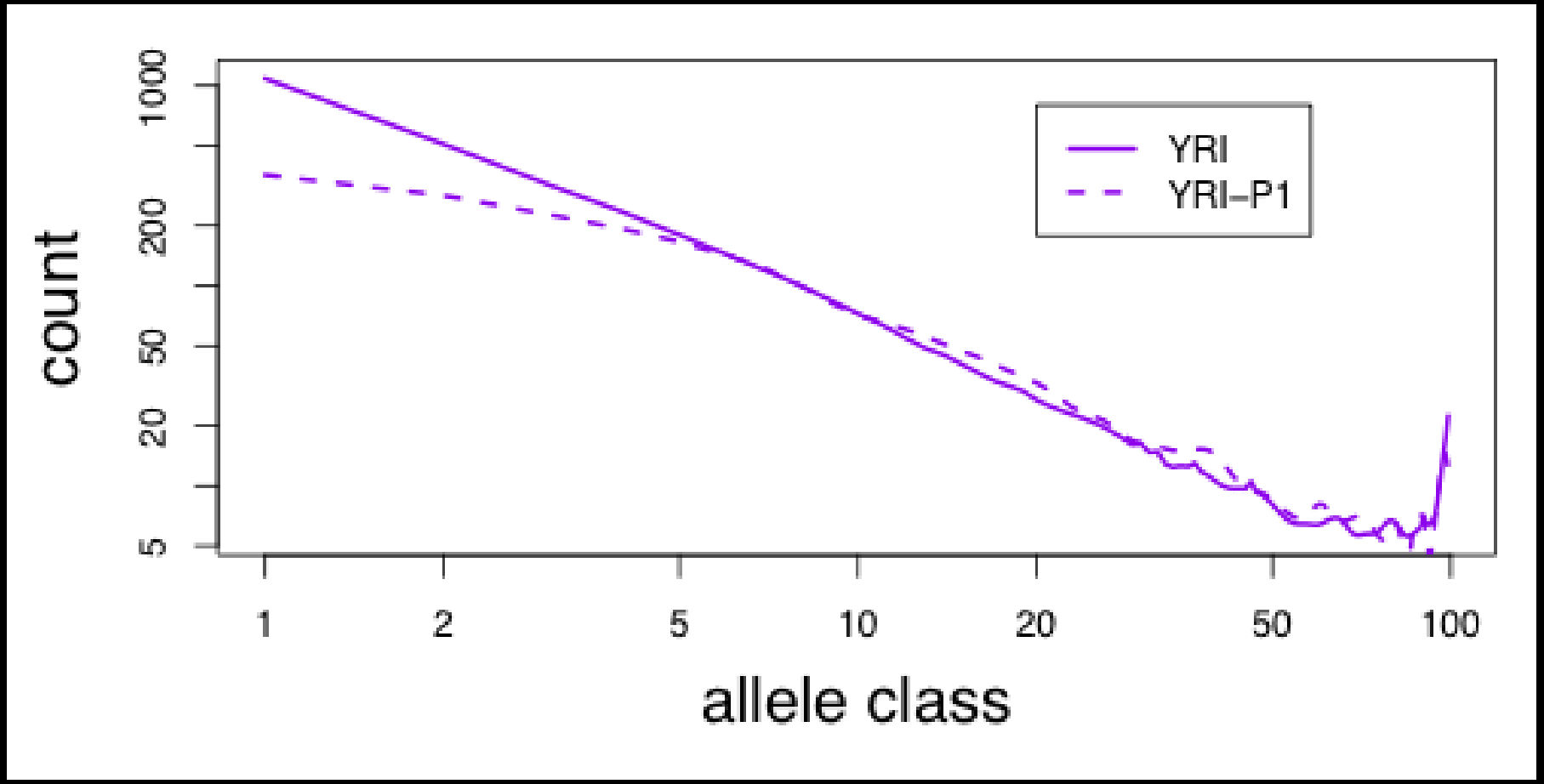
Novel variants



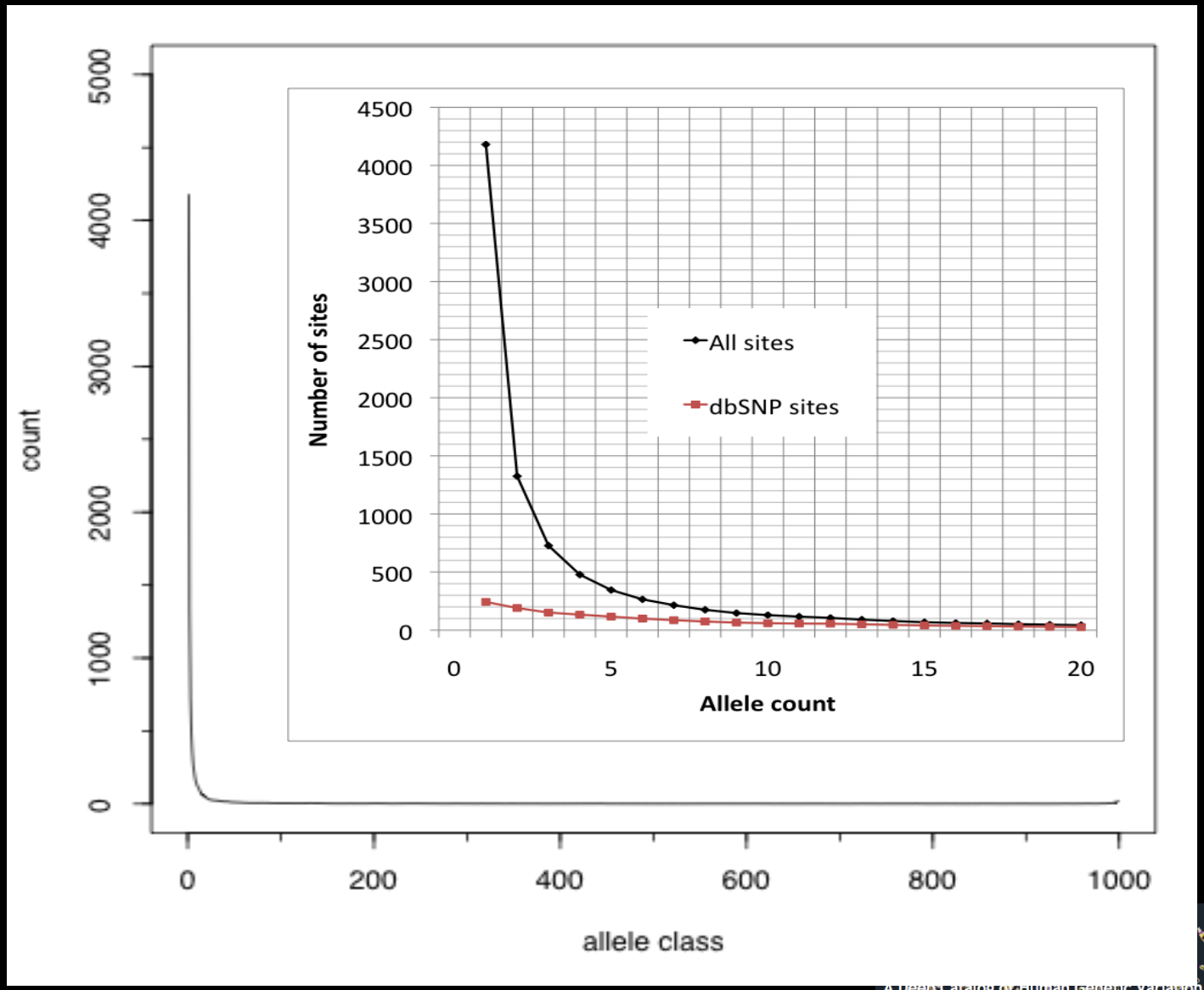
Variants per sample genome

- 3-4,000,000 variants
- 10-11,000 nonsynonymous changes
- 220-250 in-frame indels
- 80-100 premature stop codons
- 40-50 splice site disruptions
- 50-100 HGMD “recessive disease causing” mutations

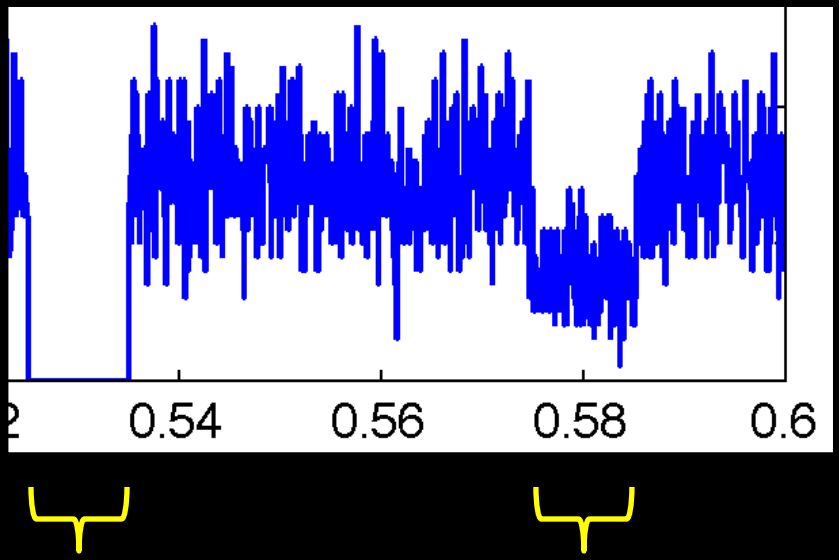
Exon Pilot: high sensitivity for rare variants

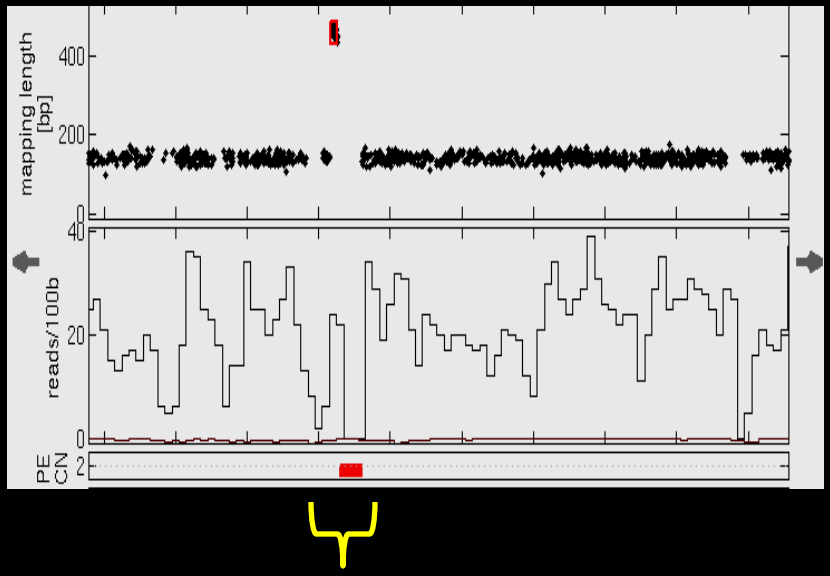


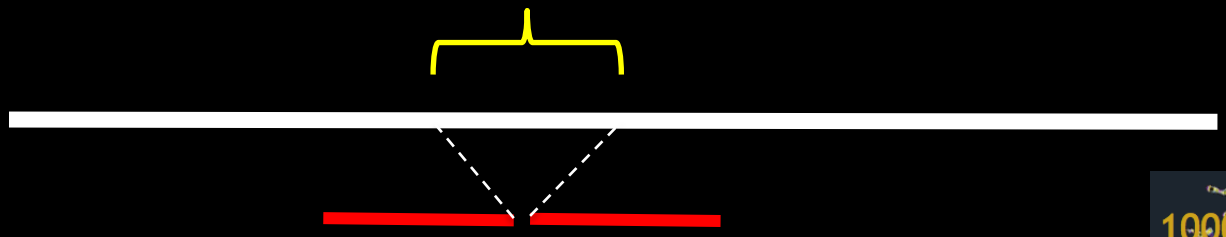
Exon Pilot: most sites low-frequency and novel



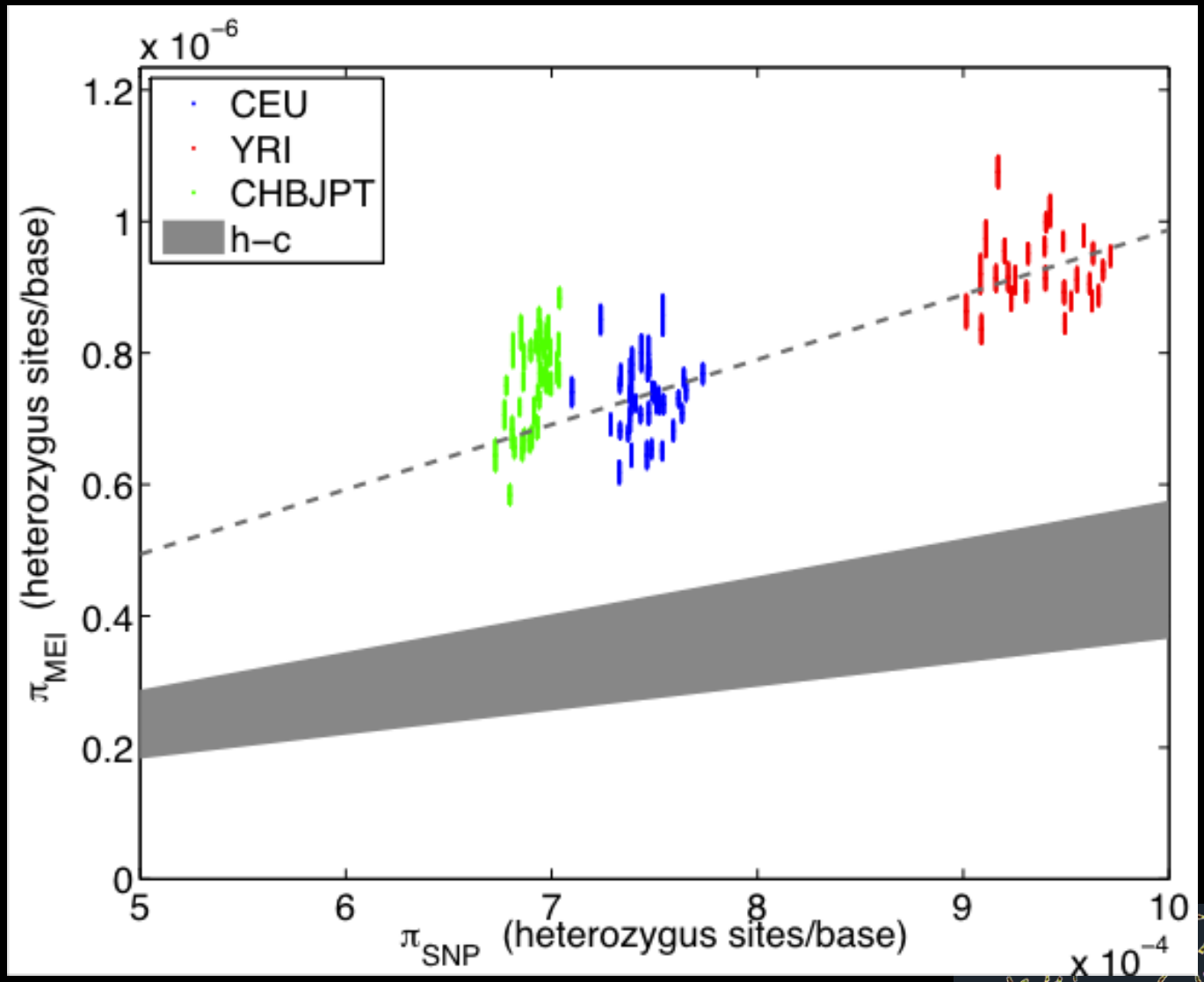
1000G data also supports structural variants







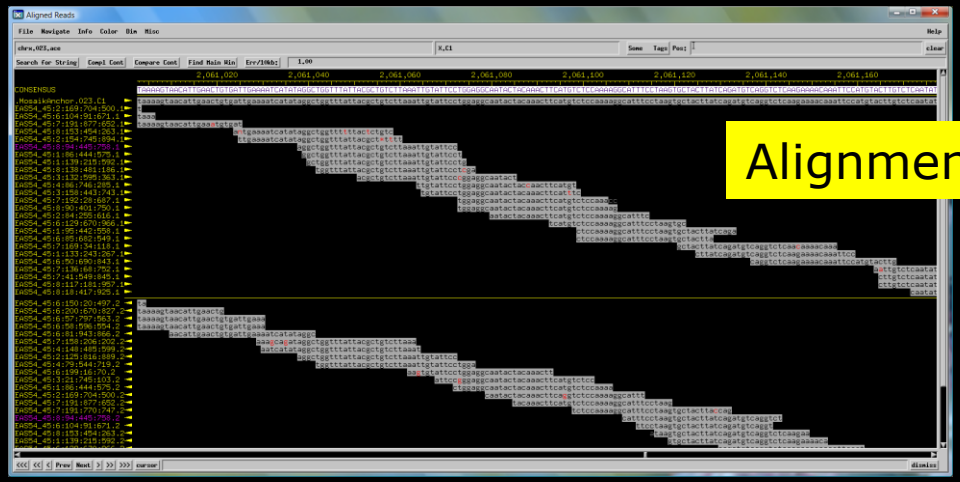
Opportunity: different variants from the same data



Data types delivered

```
@IL11_266:1:1:395:231/1
CCAACCACAACACAAAAAACACAAGCAAC
+
@@AAAAA?<>@@>?:475;A6?384,>5
@IL11_266:1:1:399:301/1
CAAAAAAAAAAAGAAGTACGAGATACGACACATCAC
+
;@AAAA>5;>@C67'&2?&7<&7&@1/1408=19::
```

Reads: FASTQ

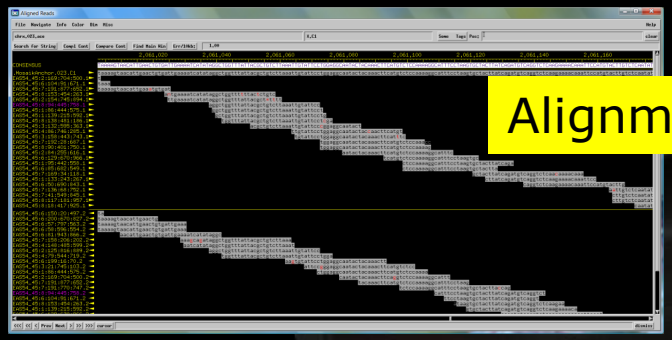


Alignments: SAM/BAM

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002
20	14370	rs6054257	G	A	29	0	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1	1|0:48:8
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3	
20	1110696	rs6040355	A	G,T	67	0	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6	
20	1230237	.	T	.	47	0	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7	
20	1234567	microsat1	G	D4,IGA	50	0	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

Variants: VCF

Tools for analyzing / manipulating 1000G data



Alignments: SAM/BAM

- samtools: http://samtools.sourceforge.net/
- BamTools: http://sourceforge.net/projects/bamtools/
- GATK: http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit

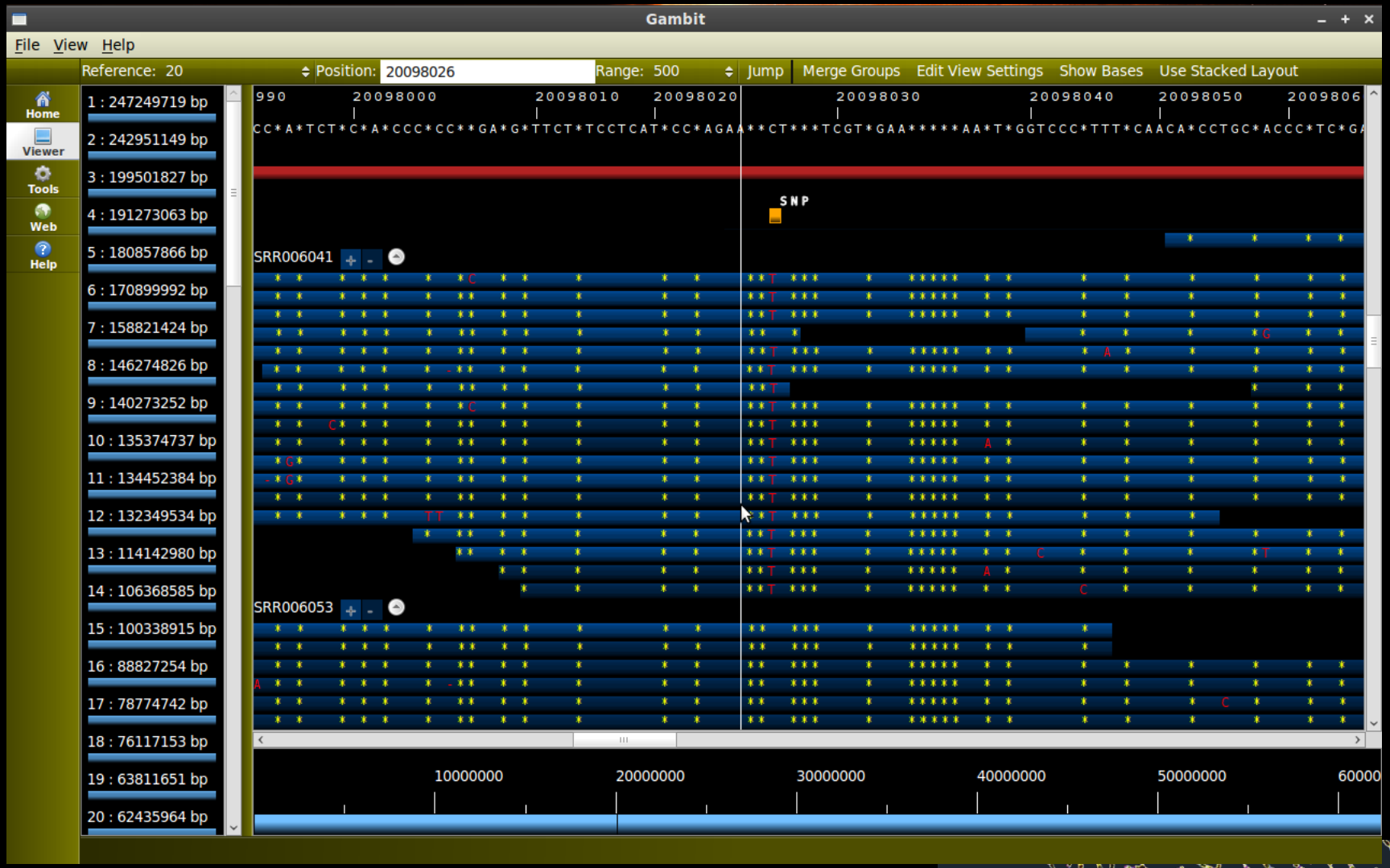
```
#CHROM  POS      ID         REF  ALT     QUAL  FILTER  INFO                              FORMAT      NA00001        NA00002
20      14370    rs6054257  G    A       29    0       NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51
20      17330    .          T    A       3     q10     NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:
20      1110696  rs6040355  A    G,T     67    0       NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:
20      1230237  .          T    .       47    0       NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:
20      1234567  microsat1  G    D4,IGA  50    0       NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2
```

Variants: VCF

- VCFTools: http://vcftools.sourceforge.net/

1000 Genomes
A Deep Catalog of Human Genetic Variation

Alignment visualization



IGV viewer, GAMBIT viewer

1000 Genomes
A Deep Catalog of Human Genetic Variation

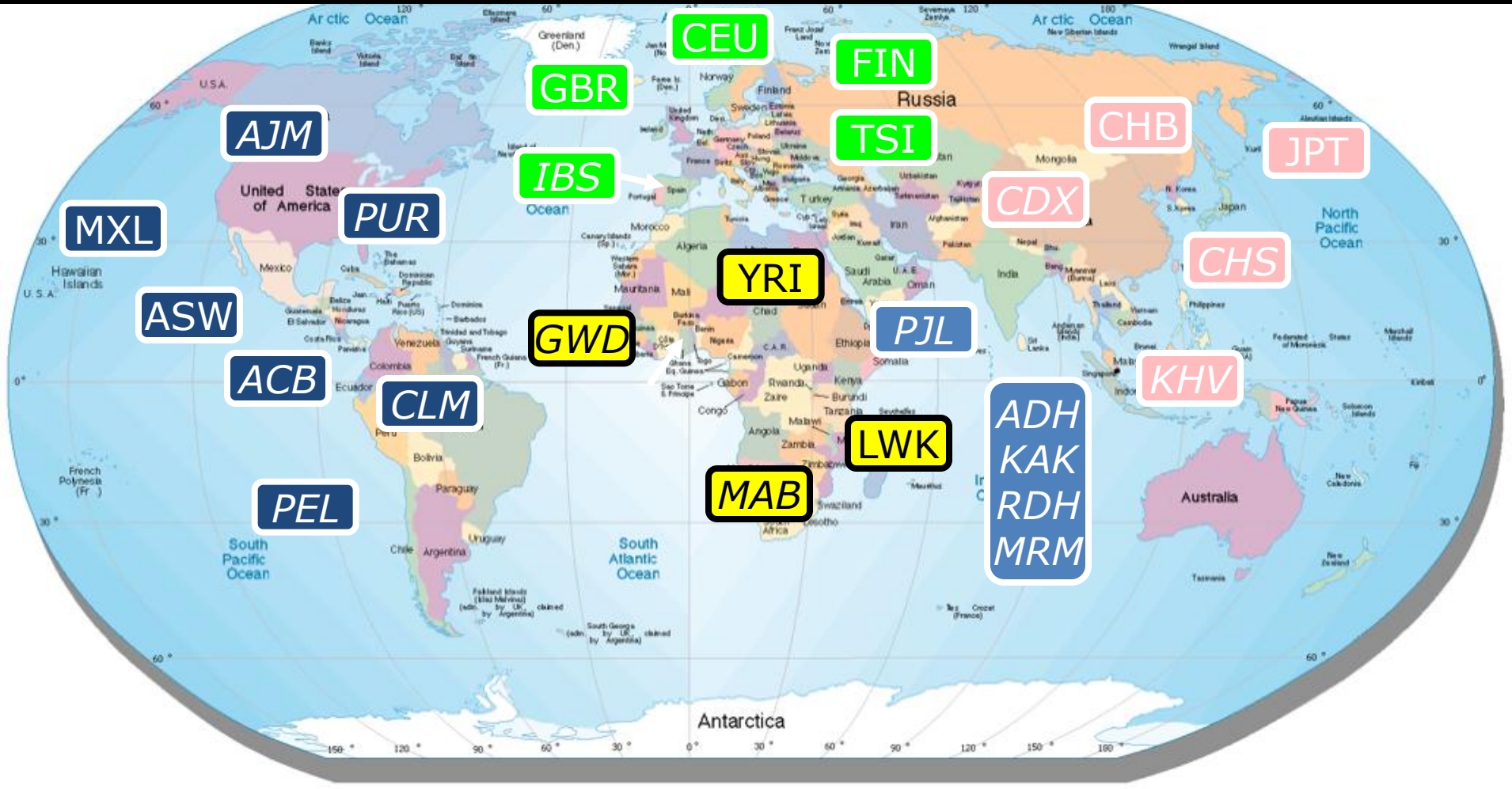
Current status based on 629 samples

Samples	# SNPs			FN metrics	
	Known	Novel	Total	dbSNP	missed HM
629	7,922,125	17,564,935	**25,487,060**	31.08%	1.21%

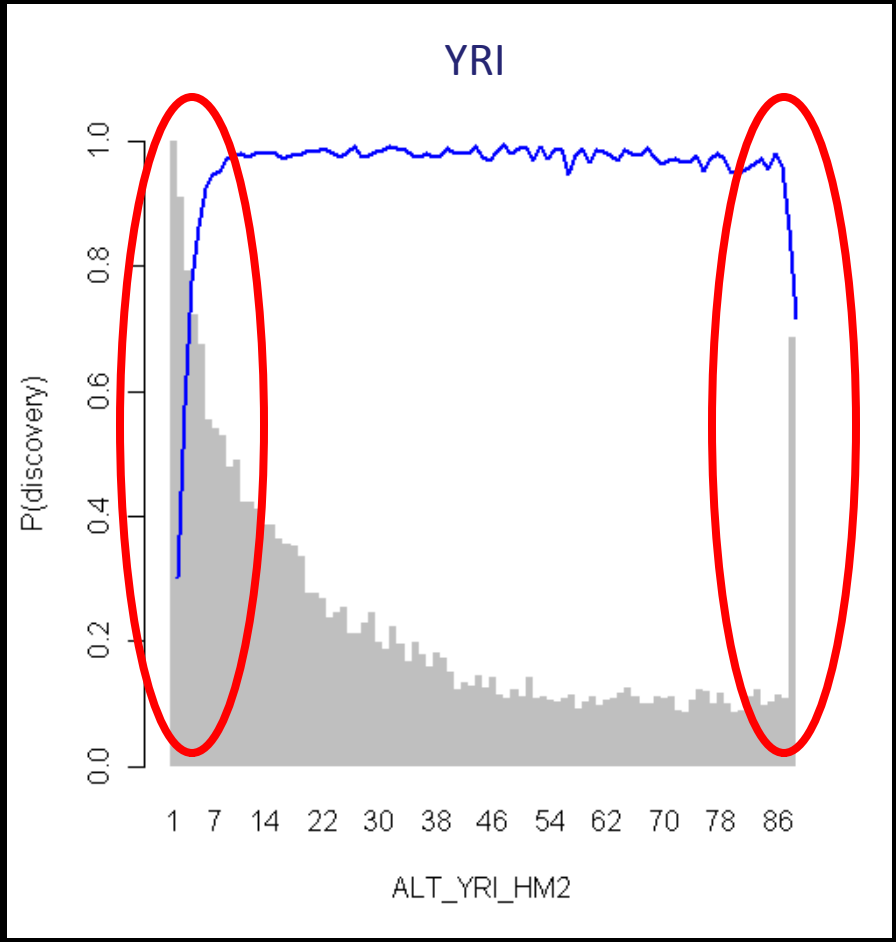
- As of 11/02/2010
- Calls present in at least 2 of Broad Institute, University of Michigan, NCBI, and Boston College call sets

The full 1000 Genomes Project data

1,100 samples early 2011; 2,500 samples 2011/12



Complementary strategies



- Low-coverage WGS (~4x per sample): a near-complete SNP catalog in the genome AF > 1%

- The deep-coverage WG exomes: rare variants, i.e. AF < 1% in genes

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
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

Data Flow Group (being formed)

Paul Flicek (co-chair) European Bioinforma[illegible]
Stephen Sherry (co-chair) National Center [illegible]
Ewan Birney European Bioinformatics Instit[illegible]
Clive Brown Sanger Institute
David Dooling Washington University in St. [illegible]
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)

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]

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

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
[illegible] **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]
Steven [illegible]
[illegible]oratory
[illegible]echnology Information
Matthew Stephens [illegible]
Simon Tavaré University of S[illegible]alifornia
Chris Tyler-Smith Sanger Institute
Jun Wang Beijing Genomics Institute
David Wheeler Baylor College of Medicine
Hongkun Zheng Beijing Genomics Institute

www.1000genomes.org