The 1000 Genomes Project

Lessons From Variant Calling and Genotyping

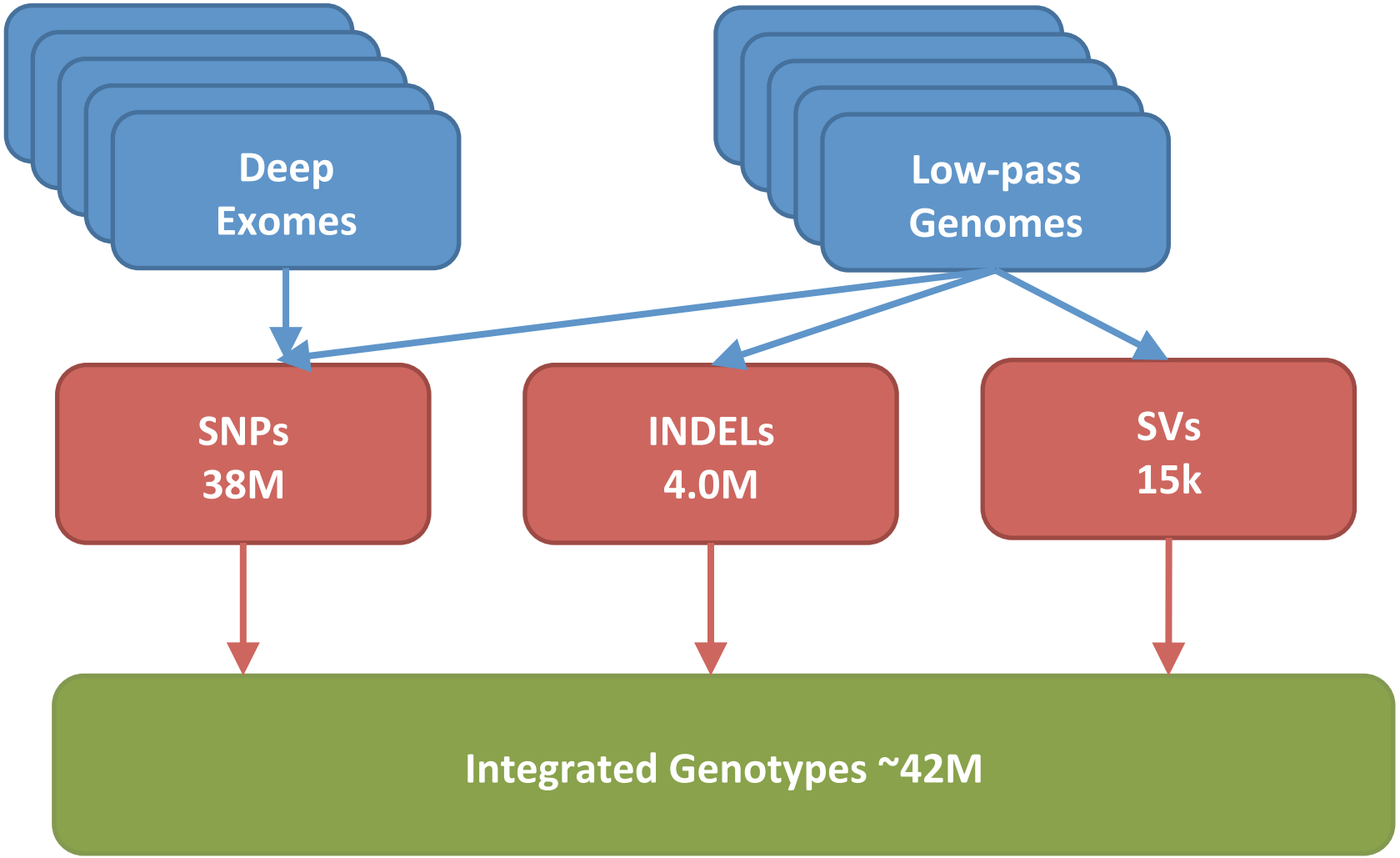
October 13th, 2011

Hyun Min Kang

University of Michigan, Ann Arbor

OVERVIEW OF PHASE 1 CALL SET

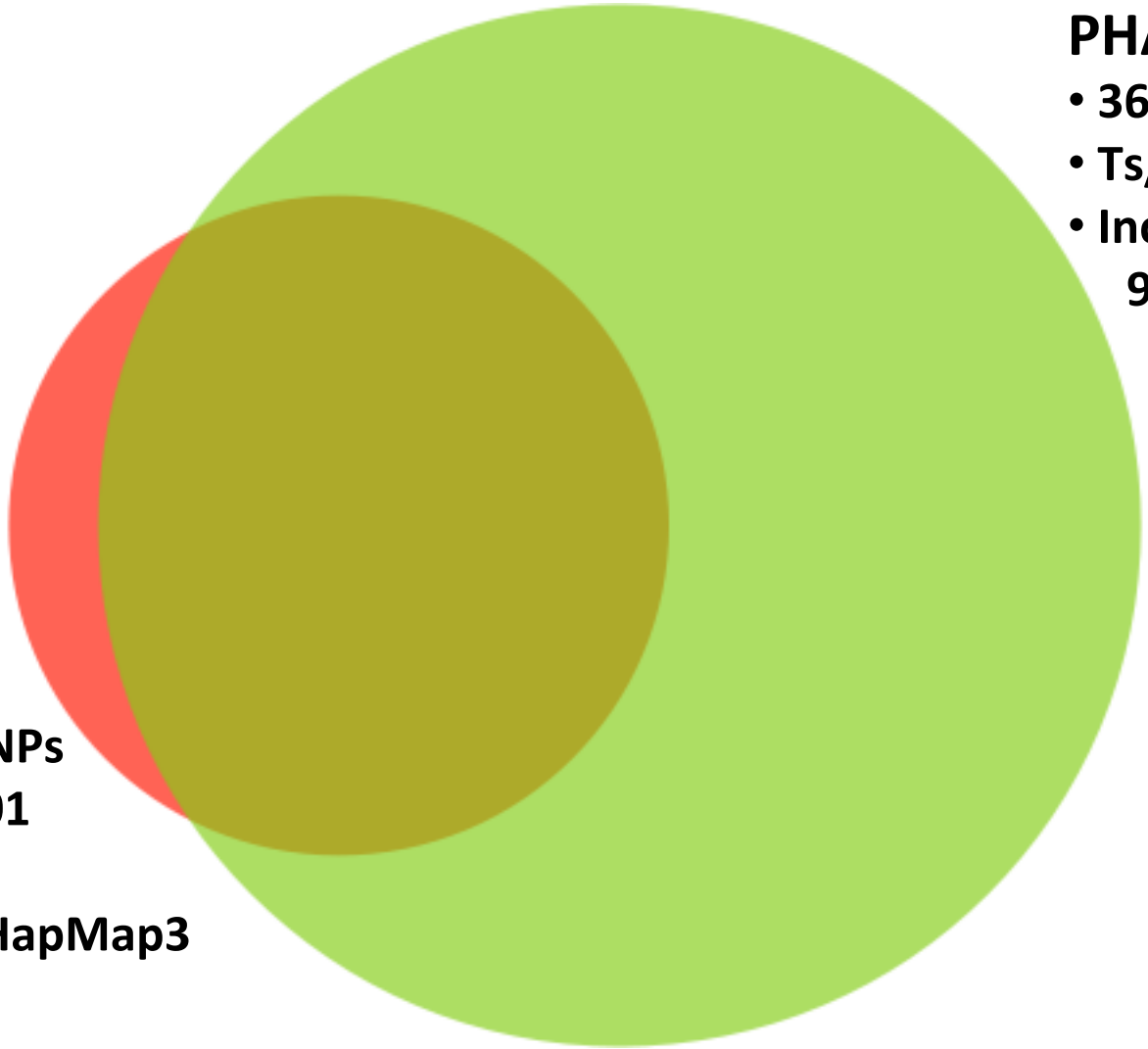
1000 Genomes integrated genotypes



Methods for integrated genotypes

Components		SNPs	INDELs	SVs
Low-Pass Genomes	**Call Sets**	BC, BCM, BI NCBI, SI, UM	BC, BI, DI OX, SI	BI, EBI, EMBL UW, Yale
	Consensus	VQSR	VQSR	GenomeSTRiP
Deep Exomes	**Call Sets**	BC, BCM, BI UM, WCMC	N/A	N/A
	Consensus	SVM	N/A	N/A
Likelihood		BBMM	GATK	GenomeSTRiP
Site Models		Variants are linearly ordered as point mutations		
Haplotyper		MaCH/Thunder with BEAGLE's initial haplotypes		

From PILOT to PHASE1

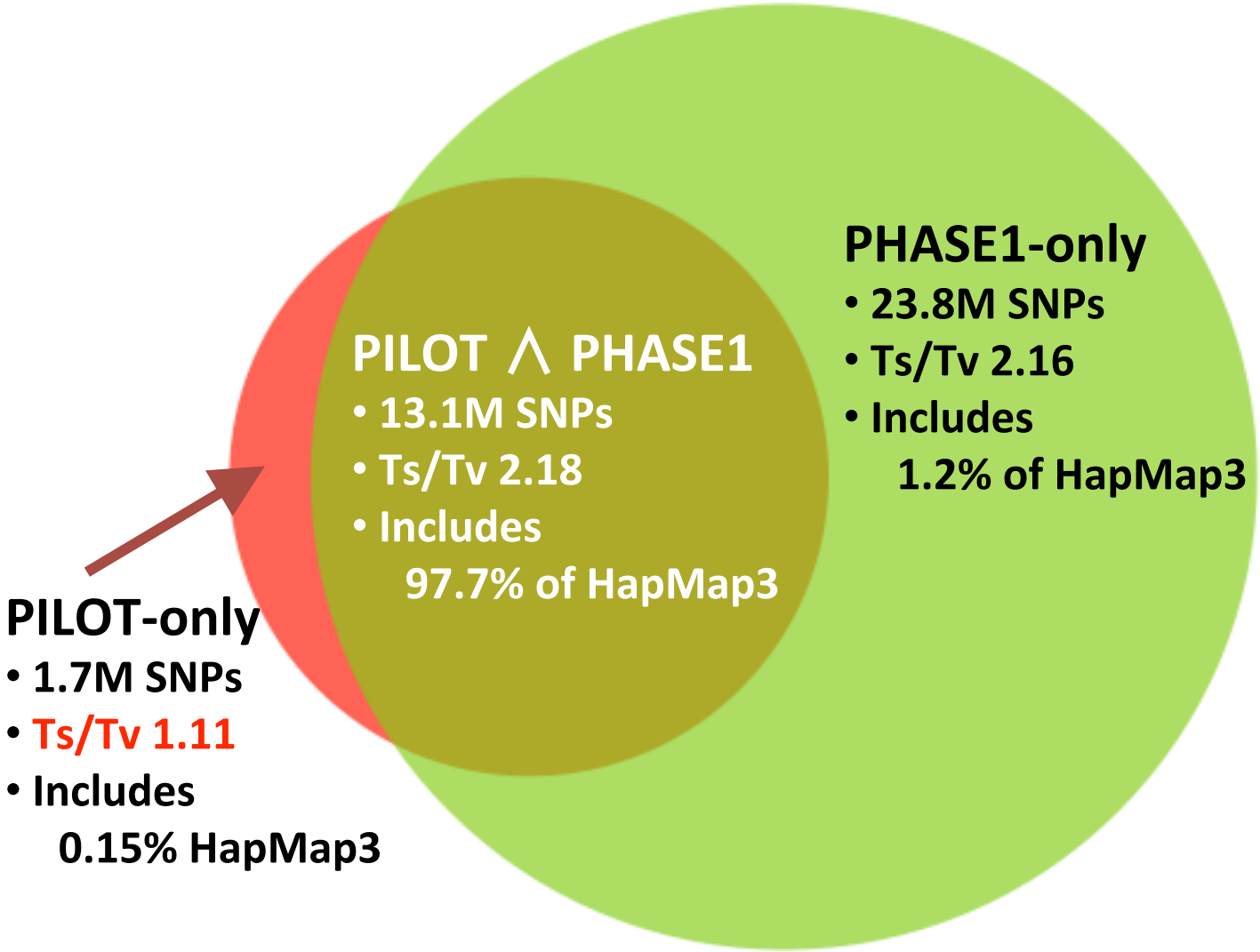


PHASE1
- 36.8M SNPs
- Ts/Tv 2.17
- Includes
  98.9% HapMap3

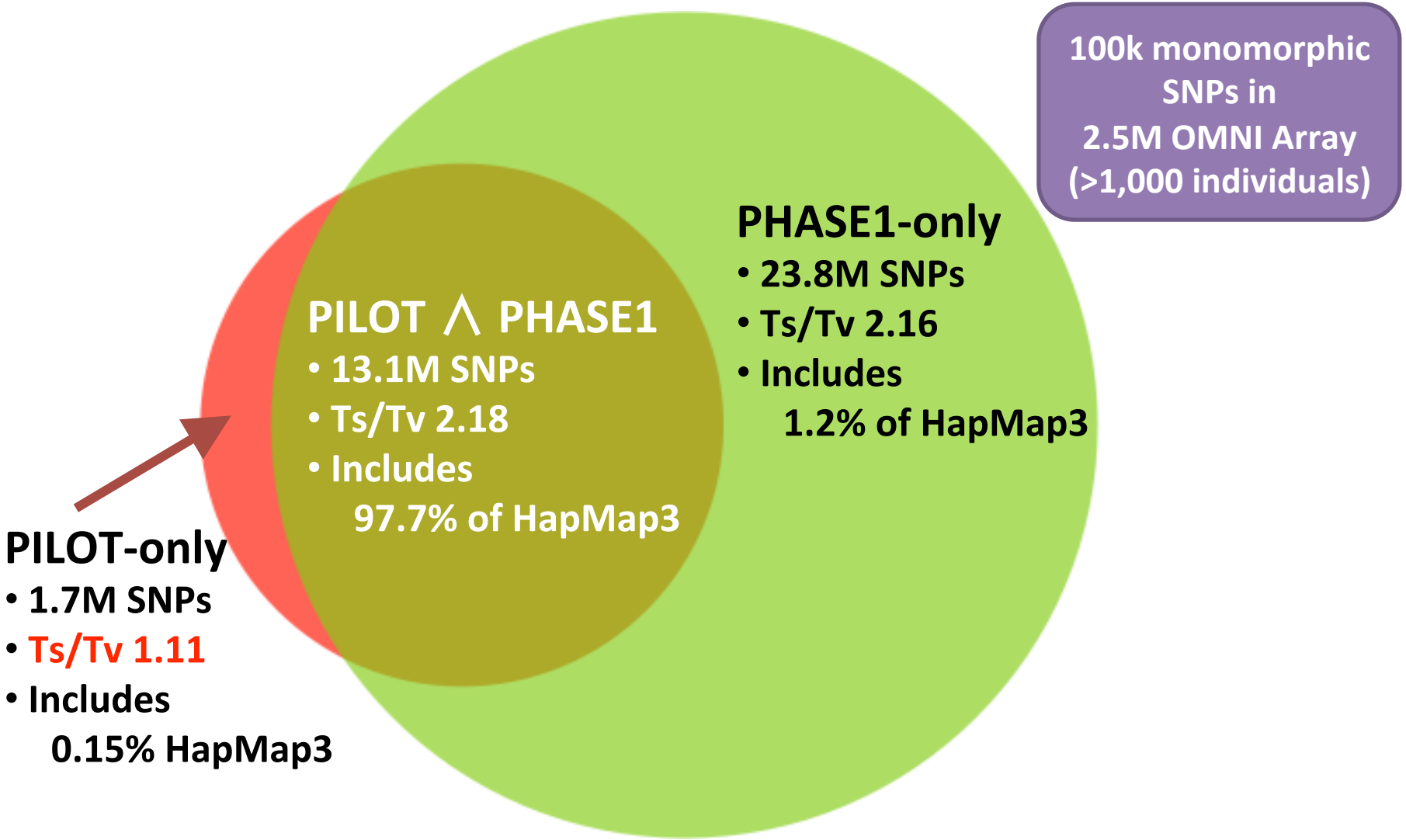
PILOT
- 14.8M SNPs
- Ts/Tv 2.01
- Includes
  97.8% HapMap3

Autosomal chromosomes only

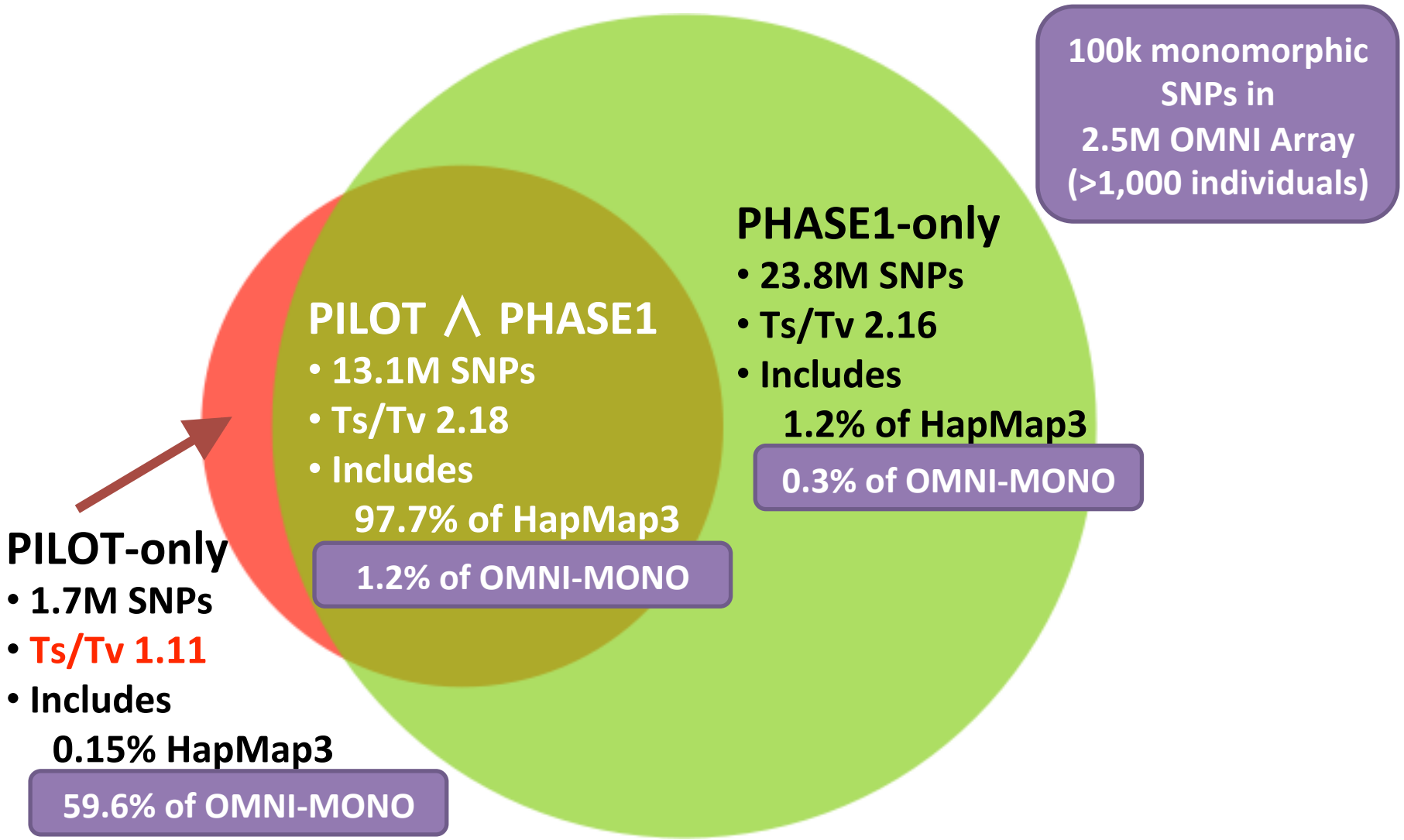
From PILOT to PHASE1



From PILOT to PHASE1



From PILOT to PHASE1 : Improved SNP calls



OMNI-MONO information was not used in making phase1 variant calls

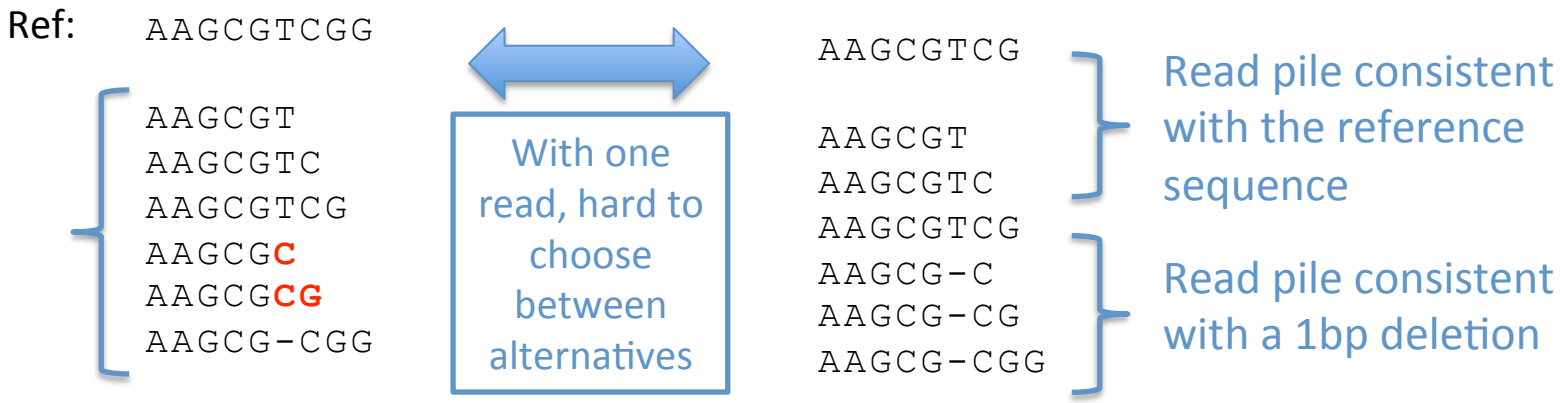
IMPROVEMENT IN METHODS SINCE PILOT

1000 Genomes' engines for improved variant calls and genotypes

- INDEL realignment
- Per Base Alignment Quality (BAQ) adjustment
- Robust consensus SNP selection strategy
  - Variant Quality Score Recalibration (VQSR)
  - Support Vector Machine (SVM)
- improved Genotype Likelihood Calculation
  - BAM-specific Binomial Mixture Model (BBMM)
  - Leveraging off-target exome reads

INDEL Realignment : How it works…

- Given a list of potential indels …
- Check if reads consistent with SNP or indel
- Adjust alignment as needed
- Greatly reduces false-positive SNP calls



Eric Banks and Mark DePristo

Per Base Alignment Qualities

Short Read

GATAGCTAGCTAGCTGATGA GCCG

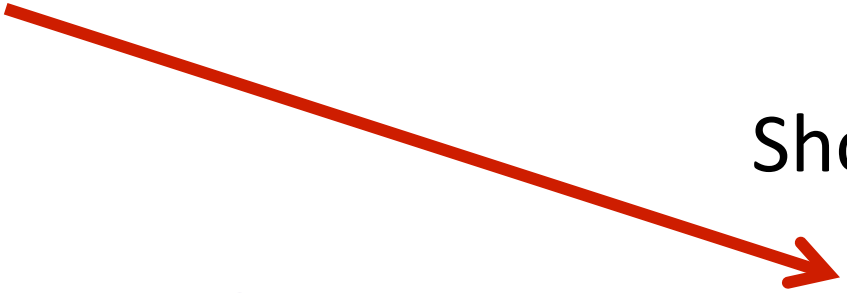
5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Heng Li

Per Base Alignment Qualities

Should we insert a gap?



Short Read

GATAGCTAGCTAGCTGATGAGCC-G

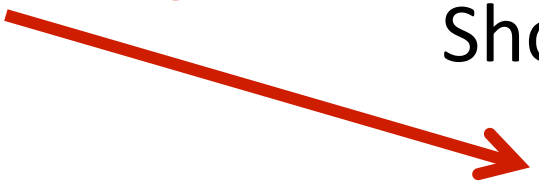
5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Heng Li

Per Base Alignment Qualities

Compensate for Alignment Uncertainty With Lower Base Quality



Short Read

GATAGCTAGCTAGCTGATGAGCCG

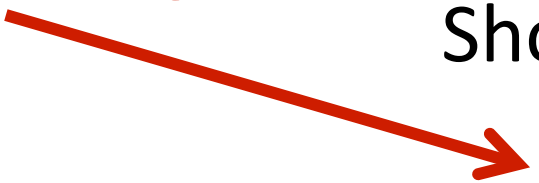
5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Heng Li

Per Base Alignment Qualities

Compensate for Alignment Uncertainty With Lower Base Quality



Short Read

GATAGCTAGCTAGCTGATGAGCCG

5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Improves quality near new indels and sequencing artifacts

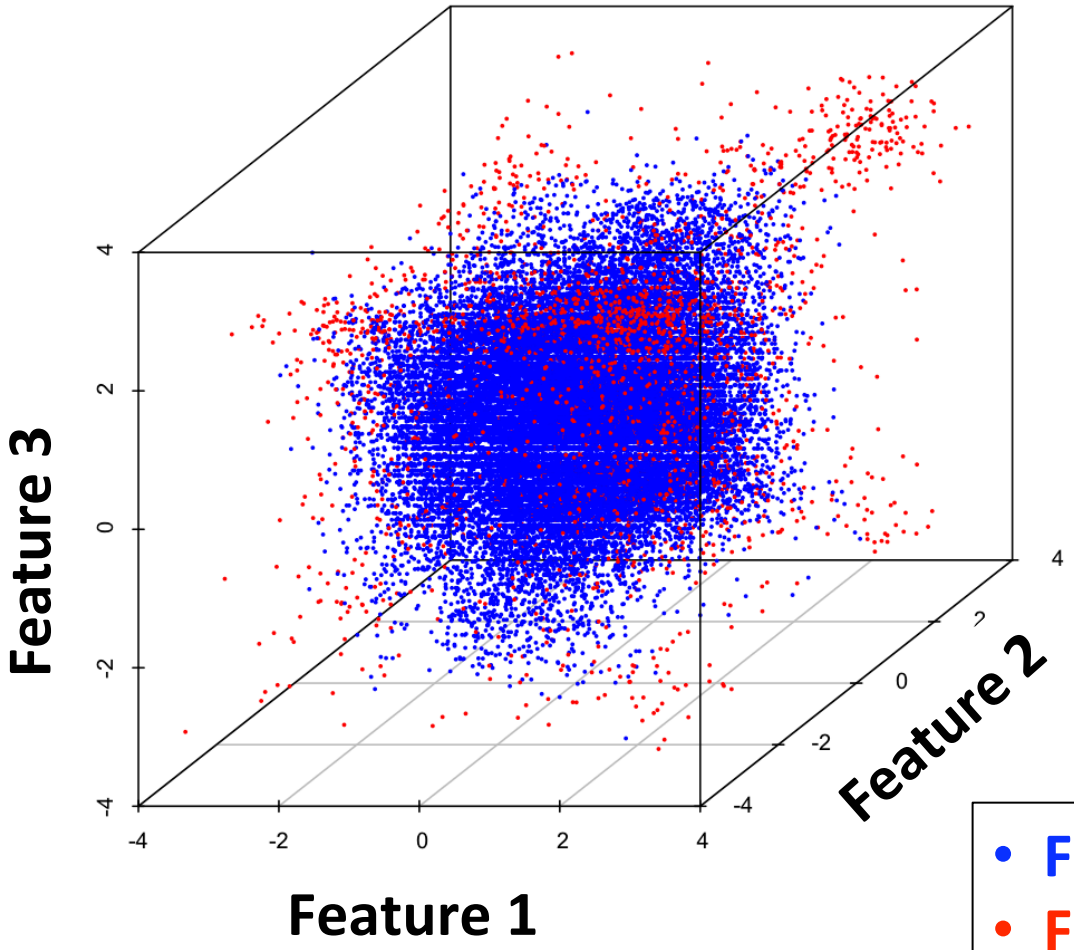
Heng Li

Producing high-quality consensus call sets

Center	Total # variants	dbSNP% (129)	Novel Ts/Tv	Omni poly sensitivity	Omni MONO false discovery
Broad	36.6M	22.7	2.17	96.5%	5.45%
Sanger	34.8M	22.9	2.18	96.1%	4.94%
UMich	34.5M	24.4	2.16	98.0%	2.77%
Baylor	34.1M	21.8	2.13	93.8%	1.43%
BC	33.3M	23.9	2.10	94.9%	9.72%
NCBI	30.7M	25.7	2.33	94.6%	10.47%
VQSR Consensus	**37.9M**	**21.7**	**2.16**	**98.4%**	**1.80%**
2 of 6	39.1M	22.2	2.15	98.6%	11.23%

Ryan Poplin

Consensus SNP site selection under multidimensional feature space

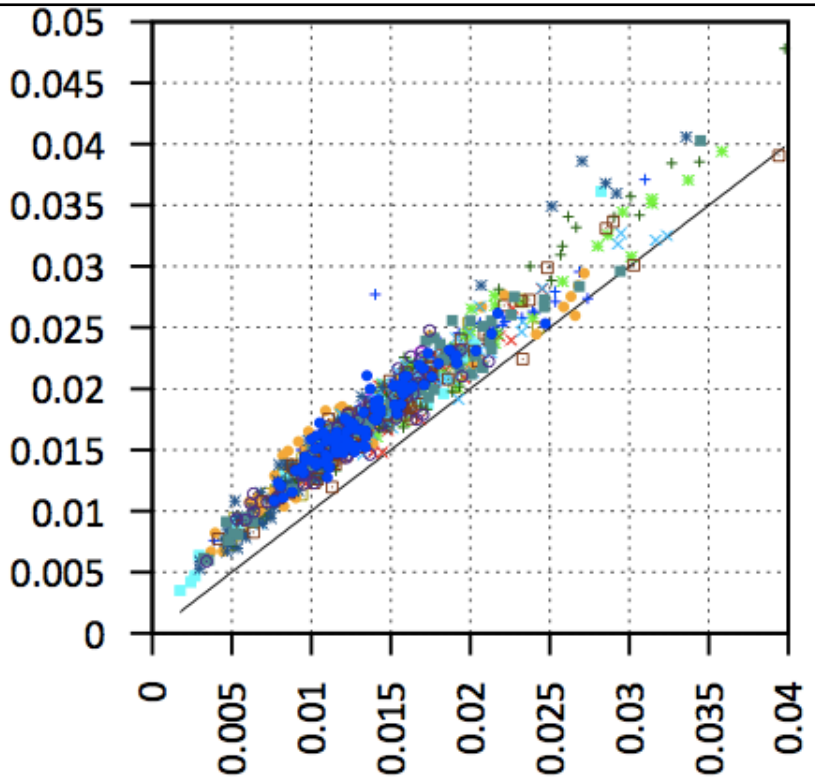


Goo Jun – Joint variant calling and … - Platform 192, Friday 5:30 (Room 517A)

Improved likelihood estimation produces more accurate genotypes

Likelihood Model	# SNPs Evaluated	HET (OMNI)	NONREF -EITHER	OVERALL
MAQ	51,002	1.86%	2.03%	0.65%
BBMM	51,002	1.49%	1.86%	0.60%

Evaluation in April 2011



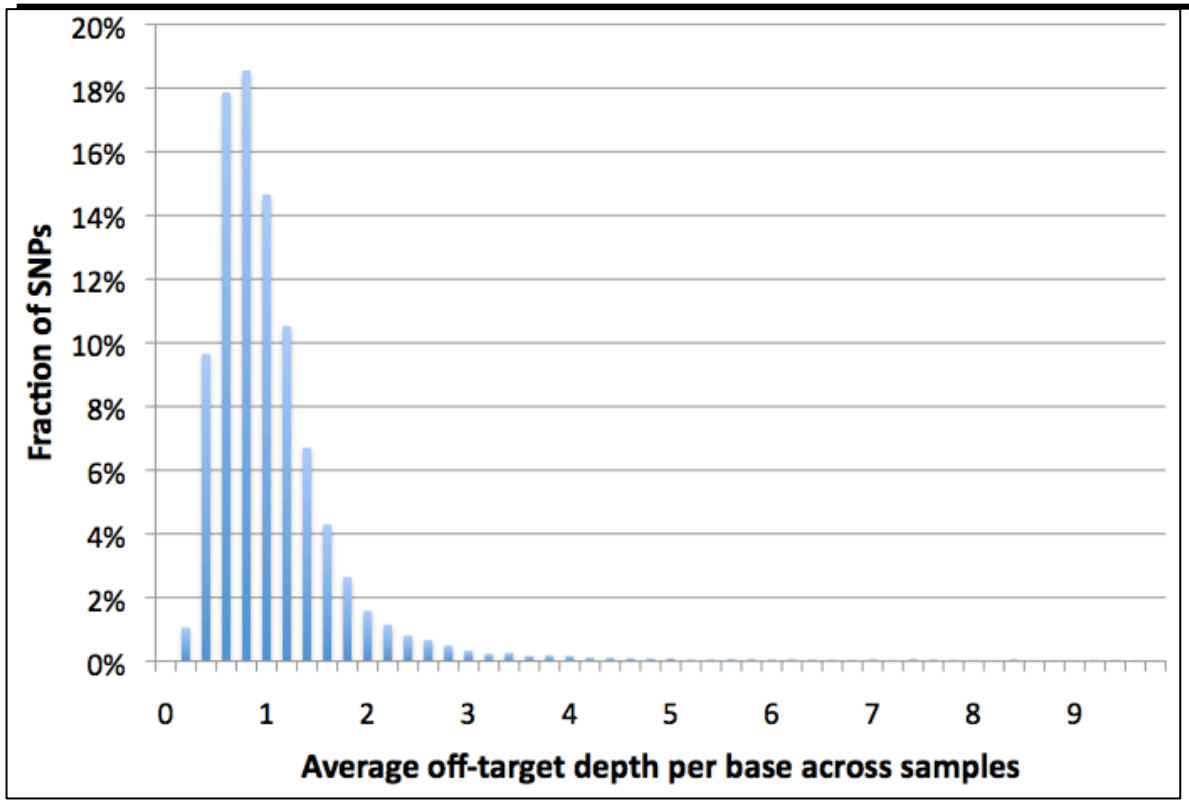
KEY IDEA in BBMM:
Re-estimate the genotype likelihood by clustering the variants based on the read distribution

GBR	+	IBS	○	LWK	□		
FIN	×	CEU	●	ASW	■		
CHS	✳	YRI	+	MXL	○		
PUR	□	CHB	×	TSI	●		
CLM	■	JPT	✳				

Fuli Yu, James Lu – An integrative variant analysis… - Poster 842W

Off-target exome reads improves genotype quality

Sites	#chr20 Variants	#OMNI Overlaps	HET (OMNI)	NREF-EITHER	OVER-ALL
Low-coverage SNPs (May 2011)	824,876	52,329	1.10%	**1.41%**	0.46%
Integrated (Nov 2011) - LC+EX/ INDELs/ SVs -	907,452	52,329	0.79%	**1.07%**	0.35%



Integrated on-target coding genotypes are also more accurate than low-coverage-only or exome-only platforms

Genotype Qualities in SVs and INDELs

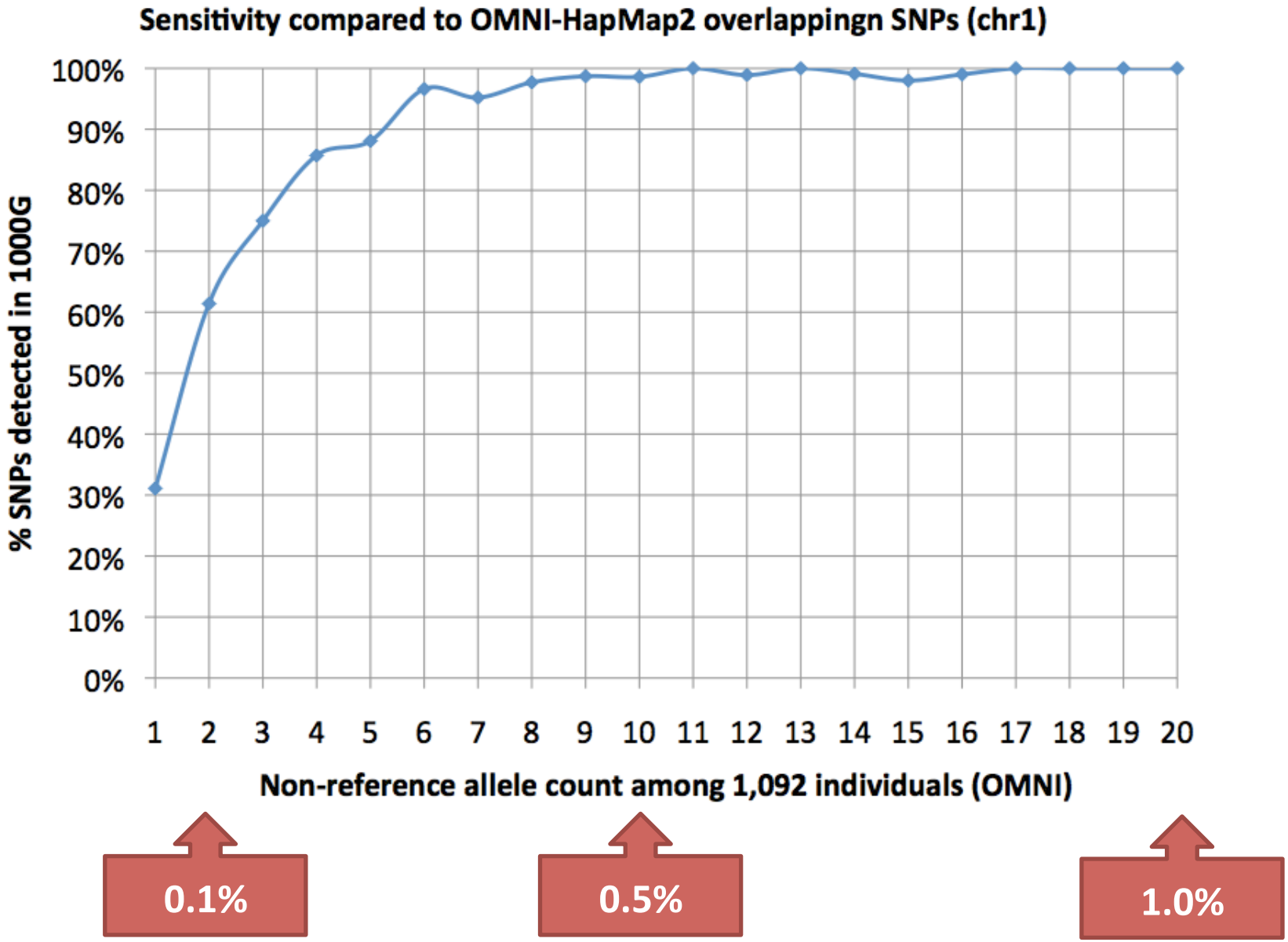
SV genotypes	Sites	Call Rate	Evaluation Data	# Sites Evaluated	HET (eval)	NONREF -EITHER	OVERALL
BEFORE Integration	13,973	95.2% [2]	Conrad (80% RO)	1962	0.61%	1.60%	0.20%
AFTER integration	13,973	100%	Conrad (80% RO)	1962	0.62%	0.93%	0.11%
IMPUTED	13,973	100%	Conrad (80% RO)	1962	4.17%	5.75%	0.74%

Bob Handsaker

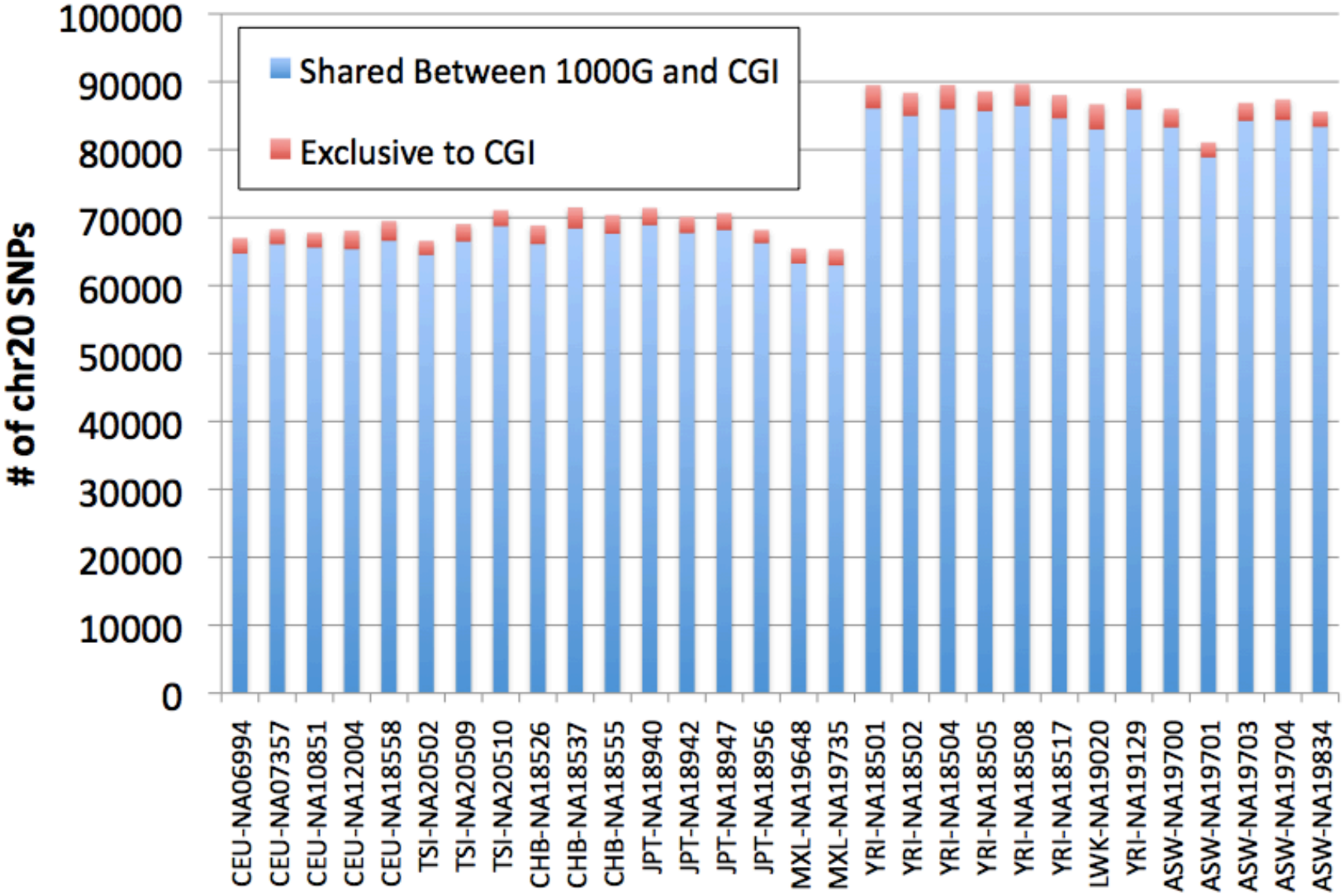
INDEL genotypes	Evaluation Data	#Sites Evaluated	HOMREF	HET	HOMALT	NREF-EITHER	OVER-ALL
1000G	CGI	1,029	0.65%	2.68%	1.24%	2.65%	1.35%
1000G	Array (Mills et al)	1,029	2.21%	7.16%	3.77%	7.56%	3.97%

MORE IN-DEPTH VIEW OF PHASE 1 INTEGRATED GENOTYPES

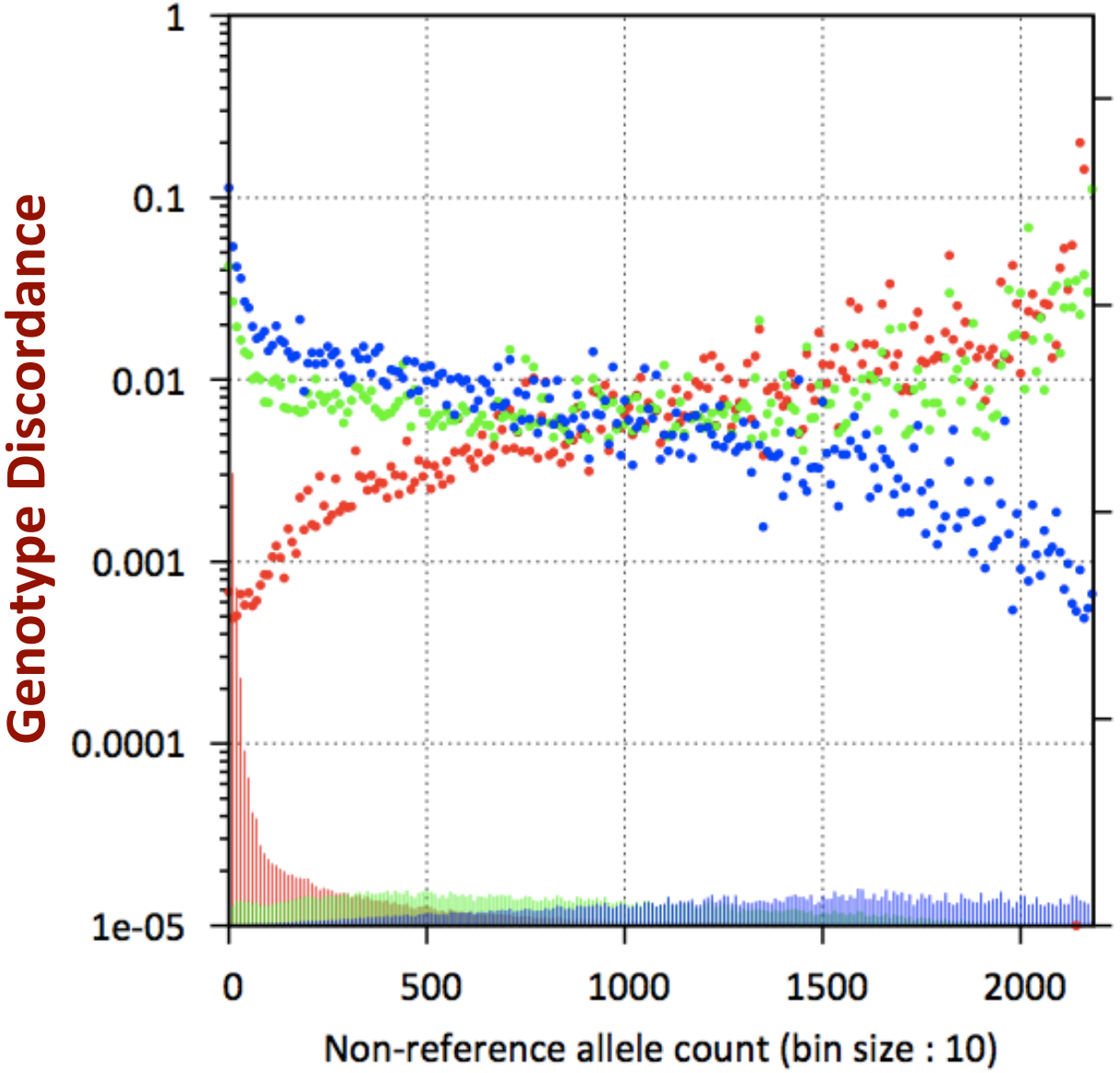
Sensitivity at low-frequency SNPs



>96% SNPs are detected compared to deep genomes

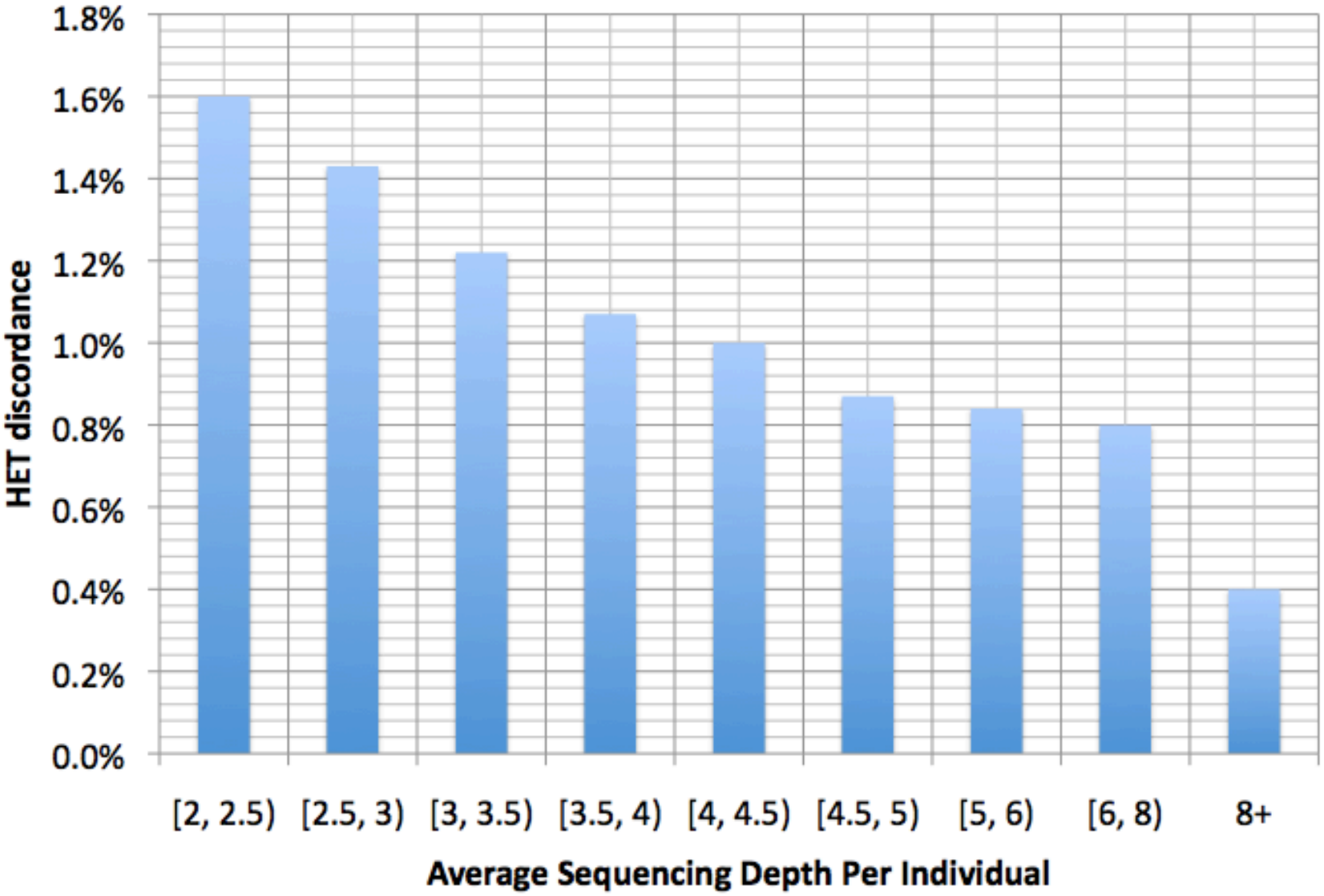


Genotype discordance by frequency



Impact of sequencing depth on genotype accuracy

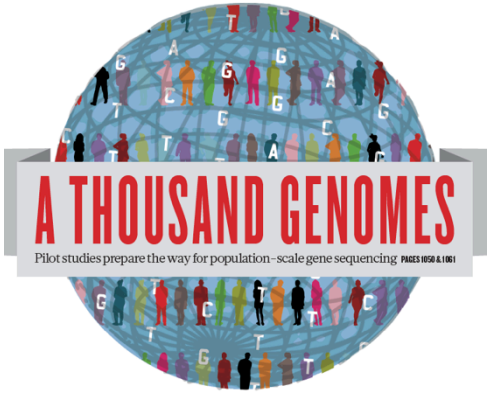
(interim integrated panel, chr20)



Highlights

- The quality of phase 1 call set is much more improved compared to pilot call set
- 1000G engines for phase1 variant calls produced high-sensitivity, high-specificity variant calls
- >99% of genotypes are concordant with array-based genotypes
- Likelihood-based integrated improves off-target & on-target genotyping qualities

Acknowledgements



The 1000 Genomes Project
1000 Genomes Analysis Group