NCBI read mapping and variant detection pipeline

Al Ward, Wen-Fung Leong, Gabor Marth (Boston College)

Hyun Min Kang, Tom Blackwell, Goncalo Abecasis (University of Michigan)

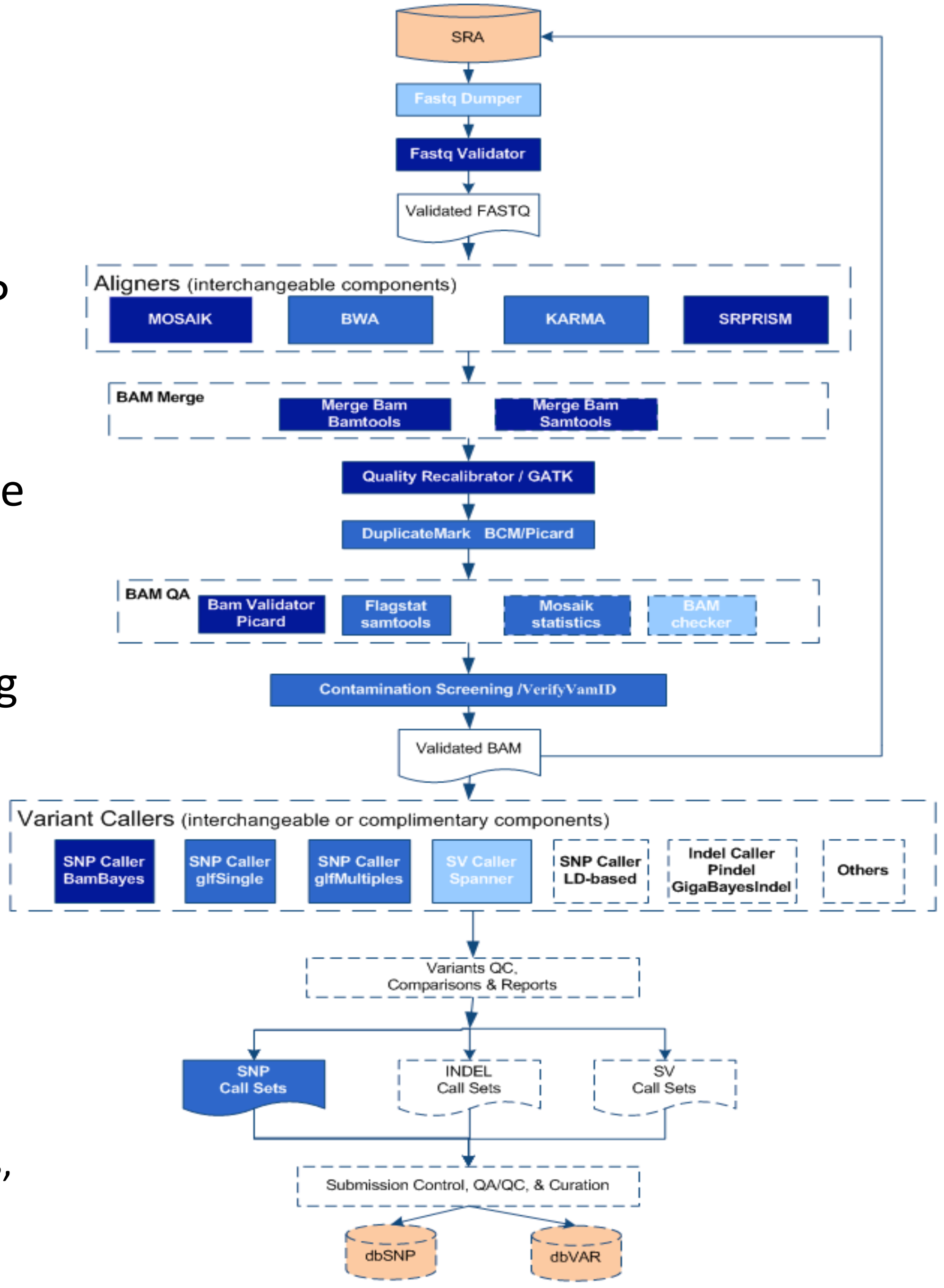
Chunlin Xiao, Anatoly Mnev, Steve Sherry (NCBI)

Project Goals / Context for 1000G

- **Goal**: develop a pipeline for variation analysis in large scale data sets deposited in NCBI’s Short Read Archive (SRA)
- **Scope**: Read mapping, SNP calling, short INDEL calling, SV calling
- **Software redundancy**: Multiple alternative components e.g. read mappers and SNP callers
- **Pipeline structure**: Modular component framework (NCBI’s GPIPE) to support primarily the BC and UM pipelines; other pipelines/components will also be supported, as needed in the future
- **What we plan to offer 1000 Genomes**: alternative pipeline for fast turn-around mapping and variant calls, on a regular schedule, based on completely automated analysis, running at the NCBI. The target turnaround is less then a month after sequence reads release.
- **Deployment**: early Fall (**now!**)

Implementation

- **Components**: Full BC SNP pipeline + many parts of UM SNP pipeline running at NCBI on a scripted basis
- **Redundancy**: Multiple alternative components e.g. read mappers and SNP callers
- **GPIPE implementation**: including multiple mappers and variant callers planned for Fall 2010
- **Next steps**:
  - SOLiD SNP calling (on TGEN BAMs?)
  - SOLiD read mapping (MOSAIK)
  - Short INDELs
  - Structural Variant calling (starting with deletions, tandem duplications, mobile element insertions) and SV genotyping



Data processing status

- **Read mapping:** Based on MOSAIK (currently on <8GB memory version).
  - June 11 1000G Data Release completely mapped (454 and Illumina reads). Turn-around time **~4 weeks for 275 samples** (198 European and 77 YRI).
  - August 4 1000G Data Release completely mapped (454 and Illumina reads). Turn-around time **~2 weeks for 383 samples** (163 Asian, 22 American, 88 African; and incremental update for 110 European genomes).
- **SNP calling**: Current call sets are produced with the BC SNP caller GigaBayes (BamBayes) and with the UM SNP caller GlfMultiples, on the June 11 Data Release (Aug. 4 release is in processing).
  - This represents our first whole genome call set
  - Compares well to equivalent call sets from the 1000G (see Hyun Min Kang's presentation)
  - Signals that our pipeline has come online and will be producing calls on a regular basis
- **Immediate to do items:** Extend our call set for SOLiD data
  - We will make calls using our SNP callers on the TGEN BAMs
  - We will map the SOLiD data (>35bp reads) with MOSAIK
  - We will make SNP calls using our own SOLiD mappings

Analysis of NCBI/BC/UM Chr20 and Whole Genome Calls

Overview

- 4 call sets not using LD information
  - MOSAIK alignments (6/11 index, ILLUMINA + LS454, 198 individuals)
    - glfMultiples (MOS-gM)
    - GigaBayes (BamBayes) (MOS-BB)
  - BWA-alignments
    - glfMultiples (BWA-gM) - (5/17 index, ILLUMINA only, 186 individuals)
    - QCALL (BWA-QC) - (5/17 index, 195 individuals, ILLUMINA+LS454)
- Uniform filter based only on genomic context
  - QUAL ≥ 10 (q10)
  - Flanking sequence (10bp) frequency ≤ 0.1% (F.1)
  - *LD-aware filter available for BWA-gM (GenoQual)*
- Comparisons on chr20 only and whole genome

Individual call set summary – chr20

#SMs	Mapper	Caller	Filter	#SNPs	%dbSNP (129)	Ts/Tv	HM3 %FNR
198	MOSAIK	bamBayes	Unfiltered	313.868	47.7	1.90	2.95
198	MOSAIK	glfMultiples	Unfiltered	330,455	46.5	2.00	2.23
186	BWA	glfMultiples	Unfiltered	332,615	44.9	1.63	2.95
195	BWA	QCALL	Unfiltered	581,774	29.7	1.42	1.92
198	MOSAIK	bamBayes	q10/F.1	295,049	49.2	1.99	3.06
198	MOSAIK	glfMultiples	q10/F.1	276,501	53.2	2.08	2.52
186	BWA	glfMultiples	q10/F.1	309,848	46.5	1.73	2.95
195	BWA	QCALL	q10/F.1	328,634	47.3	1.81	2.60
186	BWA	glfMultiples	q10/F.1/GenoQual	266,635	51.1	2.10	3.80

Consensus calls – chr20

# way	# SNPs	%UNION	%dbSNP	Ts/Tv	HM3 %FNR
UNION	409,210	100.0	39.0	1.64	2.10
2 out of 4	322,110	78.7	47.8	1.84	2.33
3 out of 4	259,317	63.4	55.9	2.08	2.58
4 out of 4	219,385	53.6	60.7	2.17	4.13
BWA-consensus	273,735	66.9	51.9	1.85	3.21
MOSAIK-consensus	257,857	63.0	55.5	2.12	3.31
glfMultiples-consensus	234,008	57.2	58.1	2.14	3.21

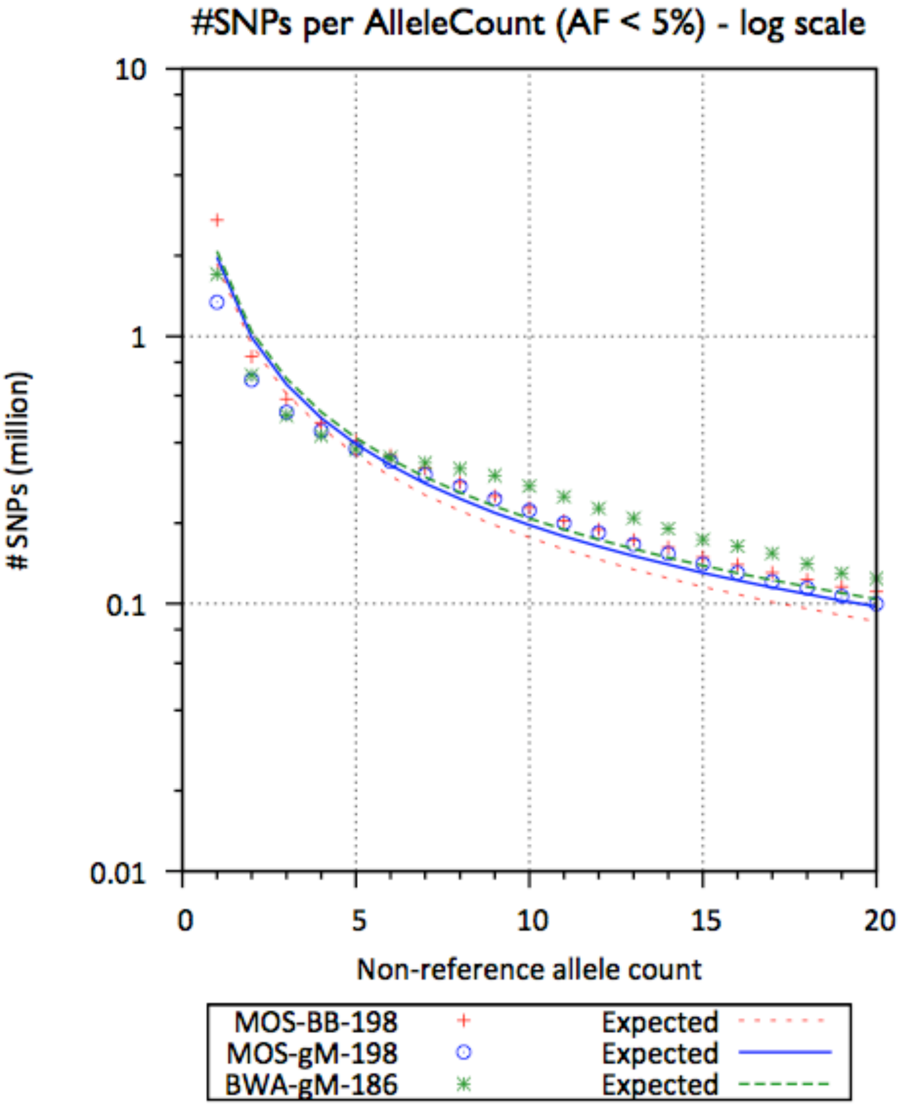
Individual call set summary – WG

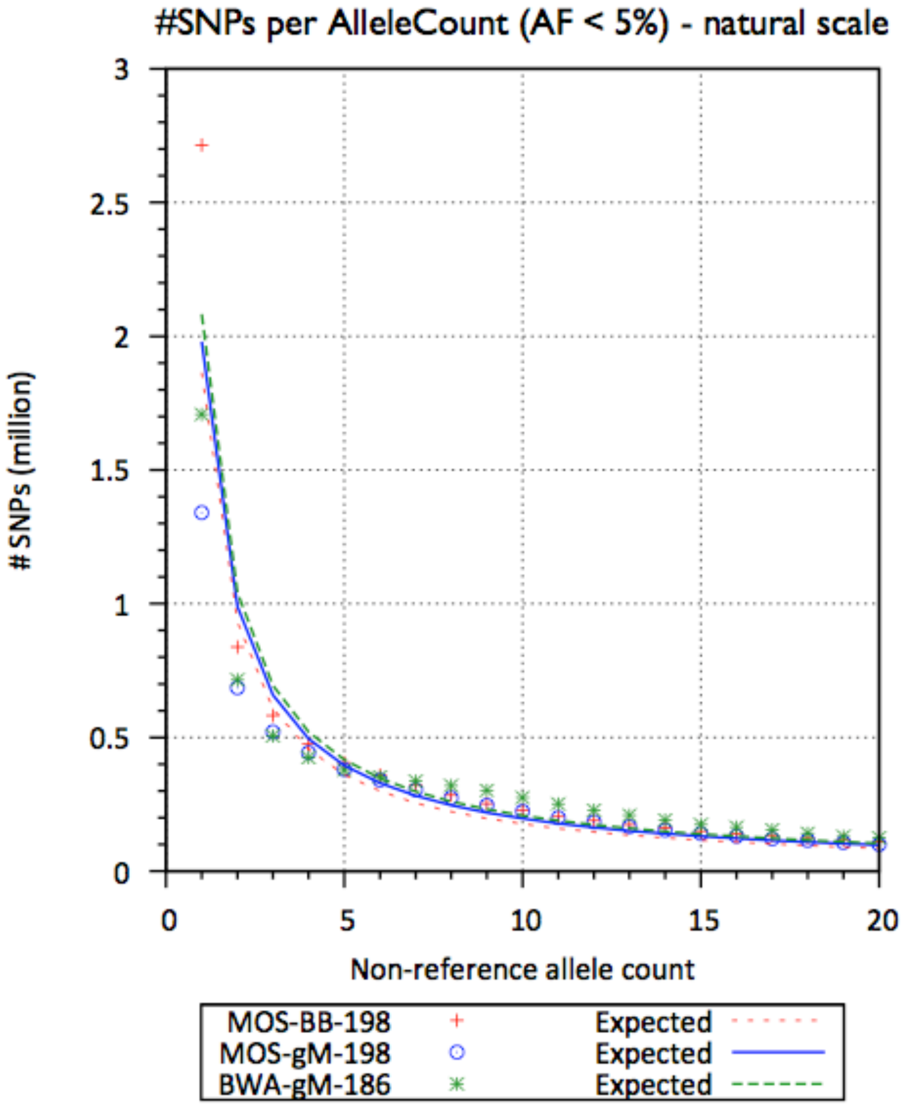
#SMs	Mapper	Caller	Filter	#SNPs	%dbSNP (129)	Ts/Tv	HM3 %FNR
198	MOSAIK	bamBayes	Unfiltered	13,263,962	48.1	1.81	3.09
198	MOSAIK	glfMultiples	Unfiltered	14,169,698	46.4	1.91	2.06
186	BWA	glfMultiples	Unfiltered	14,088,363	45.2	1.56	2.83
195	BWA	QCALL	Unfiltered	25,921,004	29.1	1.36	1.60
198	MOSAIK	bamBayes	q10/F.1	12,419,605	49.8	1.89	3.23
198	MOSAIK	glfMultiples	q10/F.1	11,697,945	53.7	1.96	2.37
186	BWA	glfMultiples	q10/F.1	13,094,933	46.9	1.65	2.91
195	BWA	QCALL	q10/F.1	14,255,682	47.8	1.71	2.34
186	BWA	glfMultiples	q10/F.1/GenoQual	11,409,996	51.1	1.98	3.90

Consensus calls – WG

# way	# SNPs	%UNION	%dbSNP	Ts/Tv	HM3 %FNR
UNION	17,151,844	100.0	39.6	1.58	1.92
2 out of 4	13,484,621	78.6	48.6	1.76	2.15
3 out of 4	10,885,760	63.5	56.8	1.97	2.46
4 out of 4	9,351,593	54.5	60.9	2.02	4.34
BWA-consensus	11,596,455	67.6	52.2	1.75	3.12
MOSAIK-consensus	10,804,544	63.0	56.4	2.01	3.44
glfMultiples-consensus	9,994,078	58.3	58.3	2.00	3.20

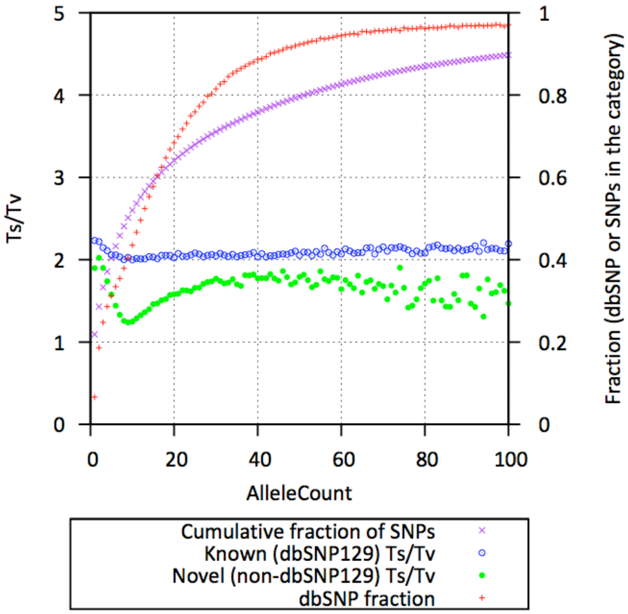
Allele frequency spectrum



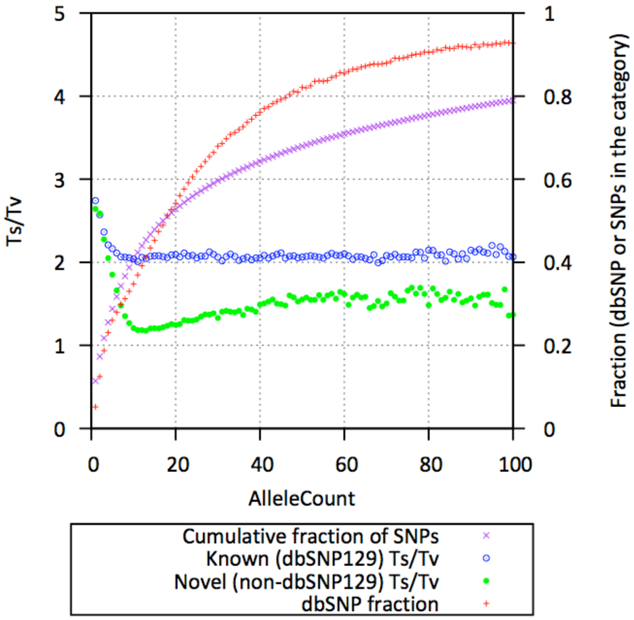


SNP Quality at low-AF variants

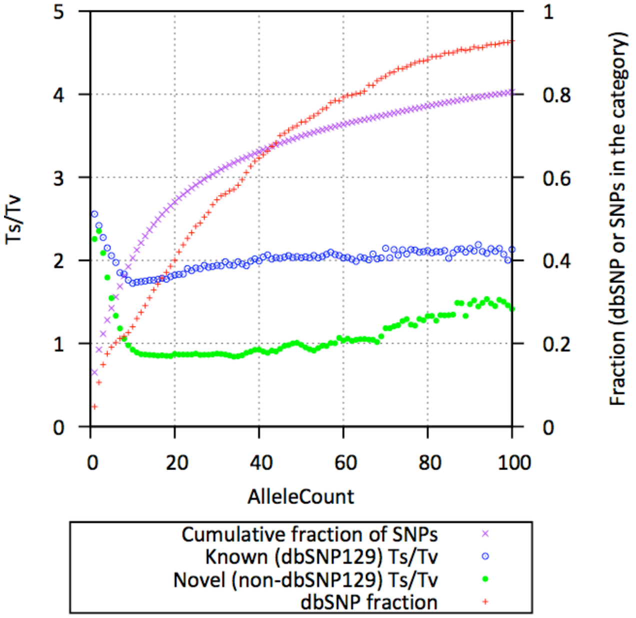
MOS-BB-198 whole-genome



MOS-gM-198 whole-genome



BWA-gM-186 whole-genome



- MOSAIK alignments show better Ts/Tv for novel SNPs
- Common novel SNPs with GigaBayes (BamBayes) show good Ts/Tv
- glfMultiples call singletons less aggressively

Summary

- Taking the consensus of calls from multiple alignments/callers improves the quality of calls.
- MOSAIK-based SNP calls appear to have higher qualities than BWA-based calls (e.g. the consensus of two MOSAIK based calls seems as good as the 3/4 intersection and is close to the 4/4 intersection).
- The allele frequency spectrum is close to expectation across call sets, with GigaBayes calling low-AF variants more aggressively.
- Based on proxy measures e.g. Ts/Tv, and on comparisons to other call sets, the NCBI whole-genome calls are of good quality, and we are happy to share them with the Analysis Group

Acknowledgements

NCBI

- Steve Sherry
- Anatoly Mnev
- Chunlin Xiao
- Richa Agarwala

U of Michigan

- Goncalo Abecasis
- Tom Blackwell
- Hyun Min Kang
- Carlo Sidore

Boston College

- Gabor Marth
- Alistair Ward
- Erik Garrison
- Wan Ping Lee
- Wen Fung Leong