Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls

Justin Zook, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide & Marc Salit

National Institute of Standards and Technology (NIST), Bethesda MD
Arpeggi Inc., Austin TX
Match 2014

Tim Farrell
BF831: Translational BF Seminar
Oct 28, 2015

Background

"The reference human genome provides a foundation... but systematic investigation of human variation requires full knowledge of DNA sequence variation across the entire spectrum of allele frequencies and types of DNA differences."

- 1000 Genomes Consortium (2010)

- Projects investigating human genome variation:
  - HapMap: catalogued allele frequencies and linkage disequilibrium (LD)
  - 1000 Genomes Project: human variation at population-scale
- Variation data sources:
  - dbSNP: small (<50 b) genomic variants
  - dbVar: large (>50 b) genomic variants

- 80 million genetic variants identified, as of June 2015[1]
- Discordance between sequencing technologies and variant callers

Motivation

- How can we be sure these variations are “true” variations?
- How can we be sure these variations are not only accurate but also “clinically actionable”?

- This study aimed to:
  - Develop methods to make high-confidence variant calls, across sequencing technologies and computational tools
  - Make results publicly available as benchmark for variant-calling tools

Data

Table 1 Description of data sets and their processing to produce bam files for our integration methods

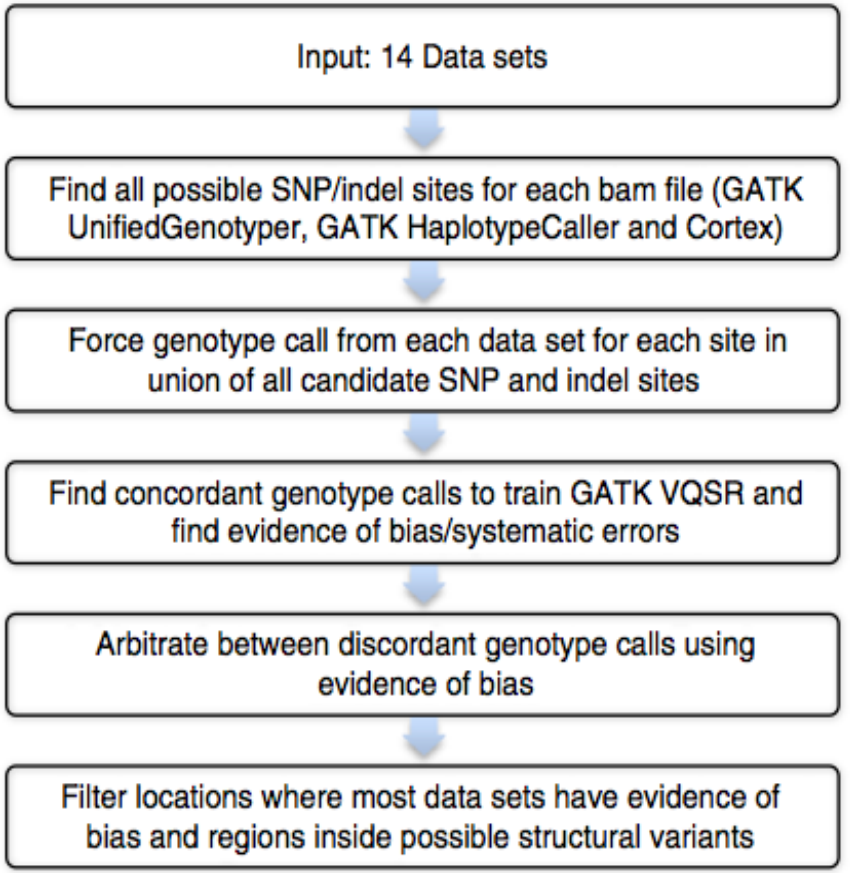
Source[a]	Platform	Mapping algorithm	Coverage	Read length	Genome/exome
1000 Genomes	Illumina GaIIx	BWA	39	44	Genome
1000 Genomes	Illumina GaIIx	BWA	30	54	Exome
1000 Genomes	454	Ssaha2	16	239	Genome
X Prize	Illumina HiSeq	Novoalign	37	100	Genome
X Prize	SOLiD 4	Lifescope	24	40	Genome
Complete Genomics	Complete Genomics	CGTools 2.0	73	33	Genome
Broad	Illumina HiSeq	BWA	68	93	Genome
Broad	Illumina HiSeq	BWA	66	66	Exome
Illumina	Illumina HiSeq	CASAVA	80	100	Genome
Illumina	Illumina HiSeq – PCR-free	BWA	56	99	Genome
Illumina	Illumina HiSeq – PCR-free	BWA	190	99	Genome
Life Technologies	Ion Torrent	tmap	80	237	Exome
Illumina	Illumina HiSeq – PCR-free	BWA-MEM	60	250	Genome
Life Technologies	Ion Torrent	tmap	12	200	Genome

[a]These data and other data sets for NA12878 are available at the Genome in a Bottle ftp site at NCBI (ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/) and are described on a spreadsheet at http://genomeinabottle.org/blog-entry/existing-and-future-na12878-datasets.

14 data sets across 5 sequencing tech and 7 mappers

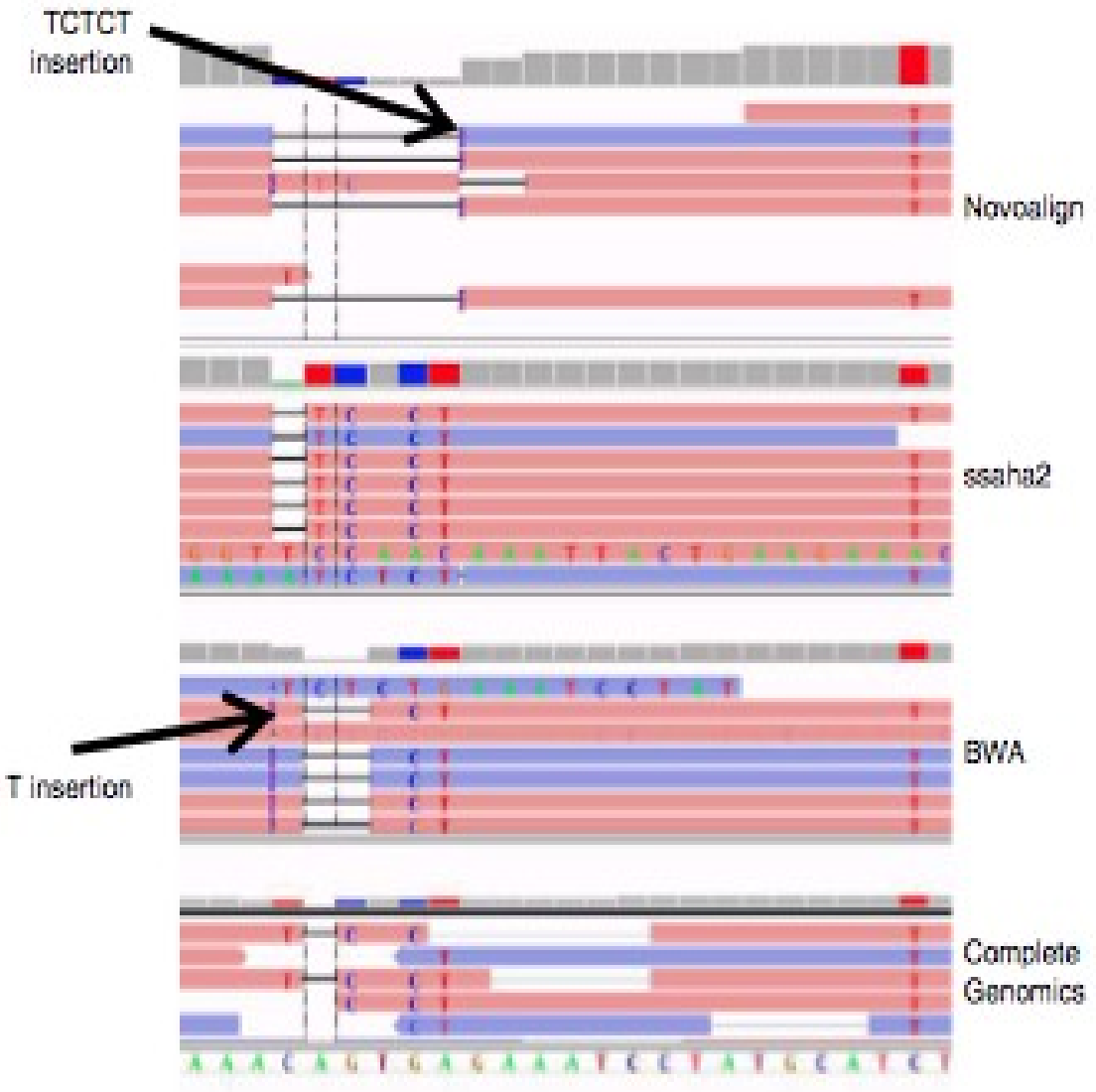
Methods

- 3 variant callers:
  - GATK UnifiedGenotyper
  - GATK HaplotypeCaller
  - Cortex
- Arbitrated between data sets that disagreed
- Filtered less confident calls



Methods

- Different variant representations make comparison difficult
- vcfallelicprimitives (from vcflib) to regularized representations



Results

Integrated variant calls are highly sensitive and specific

Table 2 Performance assessment of two individual callsets and our integrated calls vs. OMNI microarray SNPs and versus our integrated SNPs and indels

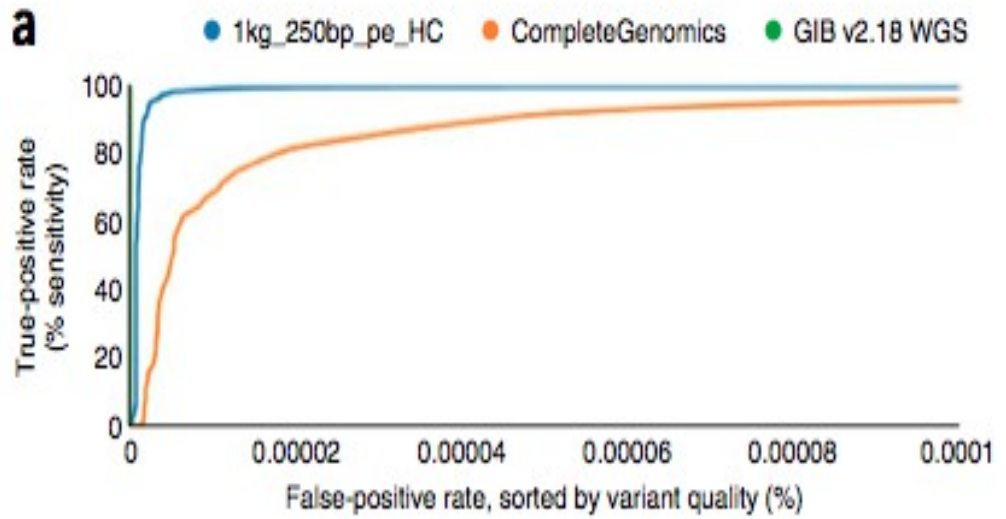
		OMNI SNPs with integrated BED file		OMNI SNPs without BED file		Integrated SNPs with BED file		Integrated indels with BED file		Common variants	Novel variants
Data set	Capture	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)	Sensitivity (%)	PR[a] (%)	Sensitivity (%)	PR[a] (%)	Ti/Tv	Ti/Tv
250bp_HC	Genome	99.49	99.97	98.47	99.93	99.90	99.73	99.55	93.11	2.04	1.43
CG	Genome	98.55	99.98	97.11	99.96	97.09	99.27	72.27	89.43	2.10	1.29
Integrated	Genome	99.54	99.98	n/a	n/a	n/a	n/a	n/a	n/a	2.14	1.94
250bp_HC	Exome	99.55	99.98	99.10	99.96	99.90	99.58	100.00	94.60	2.60	1.57
CG	Exome	98.35	99.99	97.64	99.96	99.00	99.04	90.00	85.86	2.71	1.04
Integrated	Exome	99.57	99.98	n/a	n/a	n/a	n/a	n/a	n/a	2.92	1.33[b]

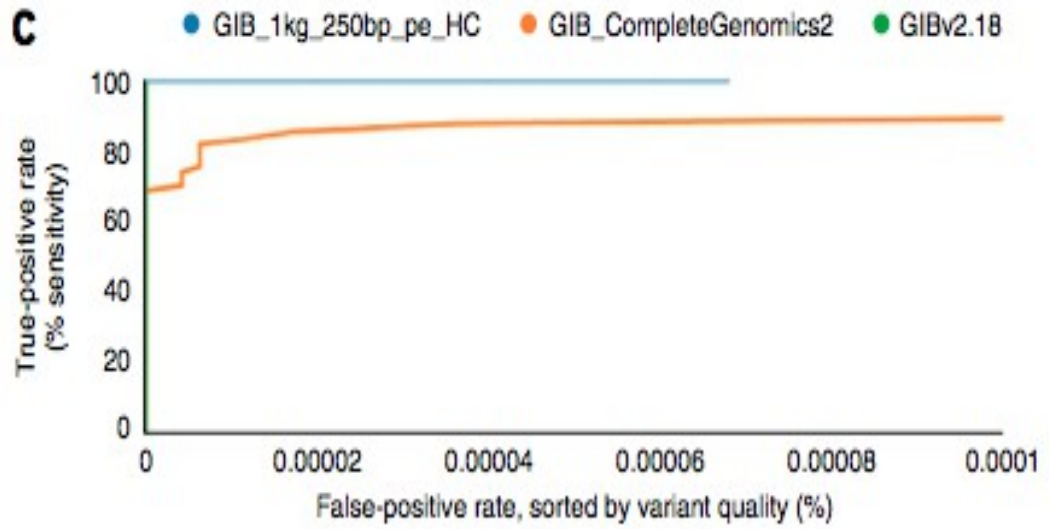
[a]Precision ratio (PR) = true positive/(true positive + false positive). The specificity of all data sets versus our integrated calls is 100.00% owing to the large number of TNs. [b]Our integrated calls only contain 30 novel variants in the exome, so the Ti/Tv has a high uncertainty. 250bp_HC is 250-bp Illumina sequencing mapped with BWA-MEM and called with GATK HaplotypeCaller v2.6. CG is Complete Genomics sequencing from 2010. n/a, not applicable.

Results

Can now quantify variant-caller performance against their benchmark

Available on Genome Comparison and Analytic Testing (GCAT) website





Conclusion

- Understanding human variation is an essential foundation for precision genomic medicine
- Here, authors develop variant-calling performan benchmarks, working toward clinical-grade reference materials
- Limitations:
  - B/c of arbitration/filter process of discordant regions 23% of genome not included in benchmark dataset
  - Homozygous v. heterozygous variant-calling

Discussion/ Reference

Assuming similar approaches confirm “true” variation, what are the best ways forward to making genomic information “clinically-actionable”?

[1] Rehm HL, et al. 2015. ClinGen – the clinical genome resource. *N Engl J Med* 372;23.

References