Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing

Jason O'Rawe[1,2], Tao Jiang[3], Guangqing Sun[3], Yiyang Wu[1,2], Wei Wang[4], Jingchu Hu[3], Paul Bodily[5], Lifeng Tian[6], Hakon Hakonarson[6], W. Evan Johnson[7], Reid J. Robison[9], Zhi Wei[4], Kai Wang[8,9], Gholson J. Lyon[1,2,9]

1) Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, NY, USA; 2) Stony Brook University, Stony Brook, NY, USA; 3) BGI-Shenzhen, Shenzhen, China; 4) Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA; 5) Department of Computer Science, Brigham Young University, Provo, UT, USA; 6) Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, USA; 7) Department of Medicine, Boston University School of Medicine, Boston MA, USA; 8) Zilkha Neurogenetic Institute, Department of Psychiatry and Preventive Medicine, University of Southern California, Los Angeles, CA, USA; 9) Utah Foundation for Biomedical Research, Salt Lake City, UT, USA.

Background

To facilitate the clinical implementation of genomic medicine by next-generation sequencing, it will be critically important to obtain accurate and consistent variant calls on personal genomes. Multiple software tools for variant calling are available, but it is unclear how comparable these tools are or what their relative merits in real-world scenarios might be. Under conditions where "perfect" pipeline parameterization is un-attainable, researchers and clinicians stand to benefit from a greater understanding of the variability introduced into human genetic variation discovery when utilizing many different bioinformatics pipelines or different sequencing platforms.

Methods

We sequenced 15 exomes from four families using the Illumina HiSeq 2000 platform and Agilent SureSelect v.2 capture kit, with ~120X coverage on average. We analyzed the raw data using near-default parameters with 5 different alignment and variant calling pipelines (*SOAP, BWA-GATK, BWA-SNVer, GNUMAP,* and *BWA-SAMTools*). Whole genome sequencing was performed on five samples using the Complete Genomics sequencing and bioinformatics pipeline v2.0 as well as with the the Illumina HiSeq2000 sequencing platform with a BWA v0.6.2-r126/GATK v2.3-9 analytical pipeline. Finally, we validate 919 SNVs and 841 indels, including similar fractions of GATK-only, SOAP-only, and shared calls, on the MiSeq platform by amplicon sequencing with ~5000X average coverage.

Results

SNV concordance between five Illumina pipelines across all 15 exomes is 57.4%, while 0.5-5.1% variants were called as unique to each pipeline. Indel concordance is only 26.8% between three indel calling pipelines, even after left-normalizing and intervalizing genomic coordinates by 20 base pairs. 2085 CG v2.0 variants that fall within targeted regions in exome sequencing were not called by any of the Illumina-based exome analysis pipelines, likely due to poor capture efficiency in those regions. Based on targeted amplicon sequencing on the MiSeq platform, 97.1%, 60.2% and 99.1% of the GATK(v.15)-only, SOAPsnp(v1.03)-only and shared SNVs can be validated, yet 54.0%, 44.6% and 78.1% of the GATK-only, SOAP-only and shared indels can be validated. Average concordance at the whole genome level across five samples between the two WGS pipelines is 71%, with 21% being called uniquely by the Illumina BWA/GATK pipeline and 8% being called uniquely by the CG pipeline.

- All Illumina exomes have at least 20 reads or more per base pair in >80% or more of the 44 MB target region.
- Concordance rates with common SNPs genotyped on Illumina 610K genotyping chips were calculated.

Sample	Software	Compared Sites	Concordance Sites	Concordance rate
Mother-1	SOAPsnp	6088	6074	99.77%
	GATK	6249	6224	99.60%
	SNVer	5723	5708	99.74%
	GNUMAP	5458	5434	99.56%
	SAMTools	5885	5848	99.37%
Son-1	SOAPsnp	6366	6353	99.80%
	GATK	6341	6323	99.72%
	SNVer	6255	6239	99.74%
	GNUMAP	5850	5828	99.62%
	SAMTools	6383	6362	99.67%
Son-2	SOAPsnp	6412	6401	99.83%
	GATK	6426	6413	99.80%
	SNVer	6336	6325	99.83%
	GNUMAP	5906	5889	99.71%
	SAMTools	6477	6450	99.58%
Father-1	SOAPsnp	6247	6238	99.86%
	GATK	6304	6288	99.75%
	SNVer	6205	6192	99.79%
	GNUMAP	5805	5786	99.67%
	SAMTools	6344	6327	99.73%

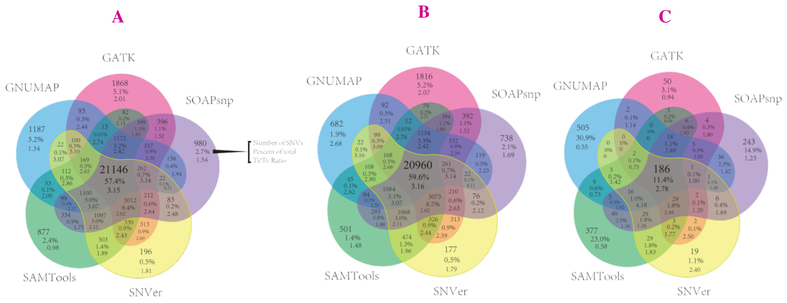
- Sensitivities and specificities were calculated for each exome pipeline using the Illumina 610k genotyping chips as a golden standard.
- All pipelines show relatively high sensitivity and specificity when detecting known and common SNPS.
- Specificity generally increases for sets of variants detected by more than a single pipeline.

	Specificity		Sensitivity		Known SNPs			Novel SNPs		
	Mean*	SD	Mean*	SD	#Total	#cSNP	Ti/Tv	#Total	#cSNP	Ti/Tv
SOAPsnp	99.82	0.039	94.53	2.287	30,022	17,409	2.77	875	419	1.94
GATK	99.72	0.085	95.33	1.161	29,620	17,306	2.8	365	206	2.34
SNVer	99.78	0.044	92.32	4.339	28,242	17,111	2.85	490	253	2.52
GNUMAP	99.64	0.065	86.67	3.286	24,893	15,144	3.03	1,091	659	1.28
SAMTools	99.59	0.158	94.45	4.221	29,577	17,449	2.78	949	539	1.33
ANY pipeline	99.62	0.113	97.72	1.215	33,947	19,638	2.68	2,163	1,182	1.23
>=2 pipelines	99.69	0.074	96.68	2.298	31,099	18,108	2.77	639	323	2.17
>=3 pipelines	99.73	0.045	95.65	3.143	29,363	17,257	2.84	416	230	2.56
>=4 pipelines	99.82	0.041	92.63	3.412	26,772	16,097	2.91	318	193	2.67
5 pipelines	99.87	0.015	80.61	5.266	21,174	13,320	3.12	234	149	2.83

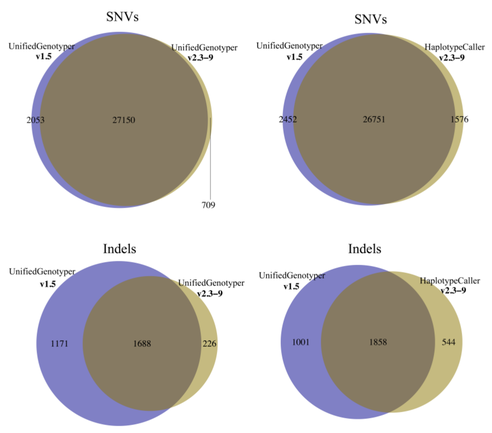
A) SNV concordance was measured between all SNV calls made by the five illumina data pipelines across all 15 exomes. Overall concordance is low: 57.4%.

B) SNV concordance is higher for already described variation (present in dbSNP135).

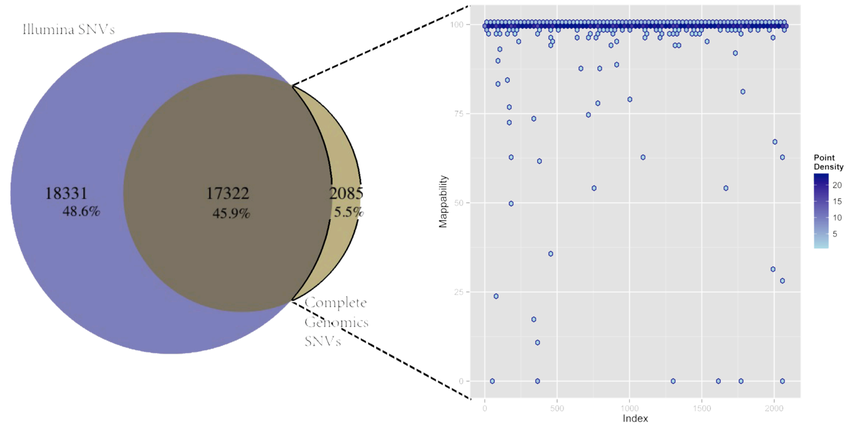
C) SNV concordance is lower for novel, un-described, human genetic variation (absent in dbSNP135).

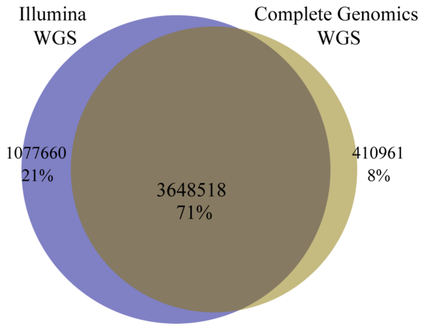


- The similarity between SNV and indel calls made between two versions of GATK, v1.5 and v2.3-9, was measured.
- SNV and indel calls were made using both the UnifiedGenotyper and HaplotypeCaller modules on the same "k8101-49685" exome.



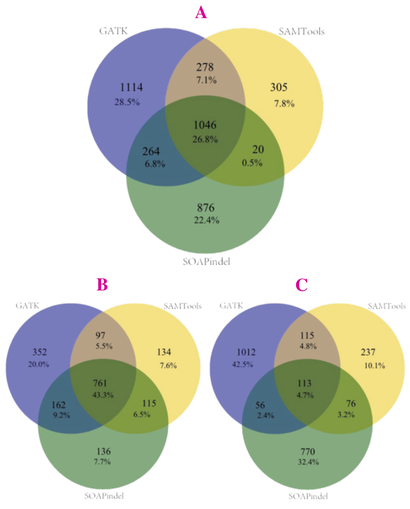
- SNV concordance was calculated for a single exome, "k8101-49685", between the illumina data calls and the Complete Genomics v2.0 calls. There are 2085 SNVs that Complete Genomics v2.0 detected but are not detected by any of the five Illumina data pipelines, despite high mappability among these variants.



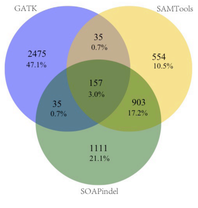


- Mean concordance across five samples between the Complete Genomics v2.0 and the Illumina HiSeq 2000 BWA/GATK whole genome sequencing and analysis pipelines is 71%.
- On average, the CG pipeline detected 410,961 variants that the Illumina BWA/GATK pipeline did not; however, the Illumina BWA/GATK pipeline detected more than double the amount of unique to pipeline variants, 1,077,660.

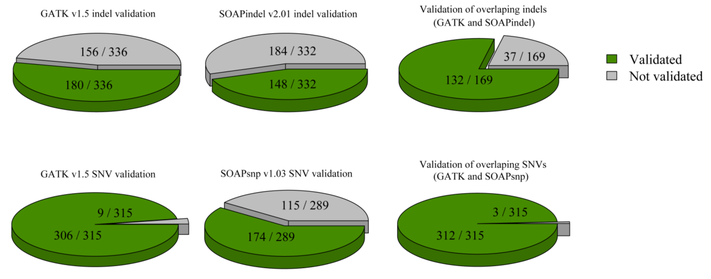
- Indel concordance across all 15 exomes between the three indel calling Illumina data pipelines (A) is low, 26.8%.
- Concordance is much better for known indels (B), and conversely much lower for novel, unknown, indels (C) (as defined by presence or absence in dbSNP135).



- Indel concordance is very low for pre-standardize calls, with only 3% agreement between the three indel calling pipelines across all 15 exomes.



- MiSeq validation was performed on a combination of SNPs and indels chosen (1756 in total) from exome sequencing data from the sample "k8101-49685".
- SNVs that were uniquely called by the SOAP-SNP v.1.03/Soap indel v2.01 and GATK v1.5 pipeline validated relatively well, with the SNVs called by both pipelines being better validated.
- Indels validated poorly for both unique to GATK(v.1.5) and SOAPindel (v2.01) calls. Overlapping indel calls validated better, though still relatively poorly.



Conclusions

We have shown that there remains significant discrepancy in SNV and indel calling between many of the currently available variant calling pipelines when applied to the same set of Illumina sequence data under near-default software parameterizations, thus demonstrating fundamental, methodological, variation between these commonly used bioinformatics pipelines. In spite of this inter-methodological variation, there exists a set of robust calls that are shared between all pipelines even under lax parameterization. However, the false negative rate may still be relatively high, even at the whole genome level, and we agree that sequencing and analyzing samples with multiple platforms and methodologies is needed to attain a high accuracy "personal genome".