Toward more accurate variant calling for “personal genomes”

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*). We additionally sequenced a single whole genome using the Complete Genomics (CG) sequencing and analysis pipeline (v2.0), with 95% of the exome region being covered by 20 or more reads per base. Finally, we attempted to 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.

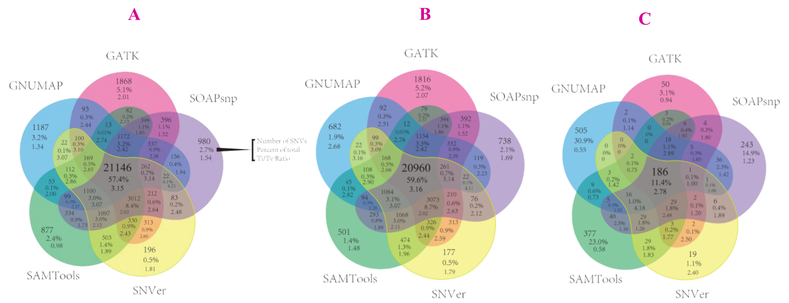
- **All Illumina exomes have at least 20 reads or more per base pair in >80% or more of the 44 MB target region.**
- **Concordance rates with common SNPs genotyped on Illumina 610K genotyping chips were calculated.**
- **All pipelines are very good with identifying already known, common SNPs.**

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%

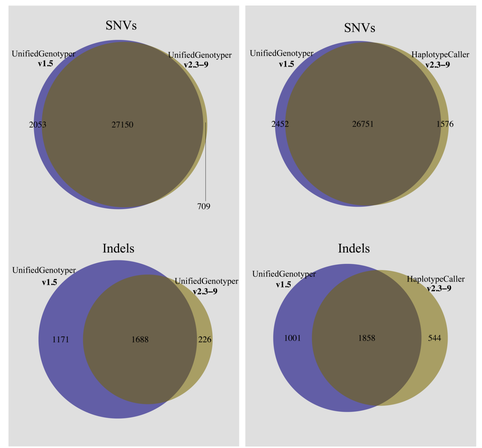
- **Sensitivities and specificities were calculated for each pipeline using the Illumina 610k genotyping chips as a golden standard.**
- **All pipelines show relatively high sensitivity and specificity when detecting known and common SNPS.**
- **Specificity generally increases for sets of variants detected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

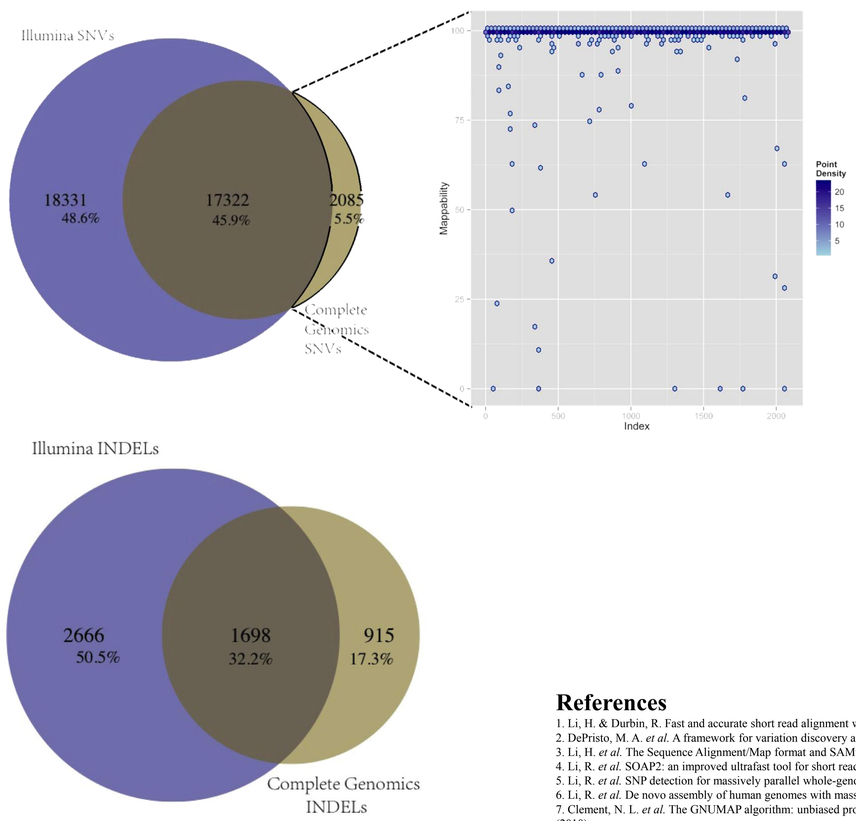
- **A) SNV concordance was measured between all SNV calls made by the five illumina data pipelines. Overall concordance is low: 57.4%.**
- **B) SNV concordance is higher for already described variation (present in dbSNP135).**
- **C) SNV concordance is lower for novel, un-described, human genetic variation (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 participant sample.

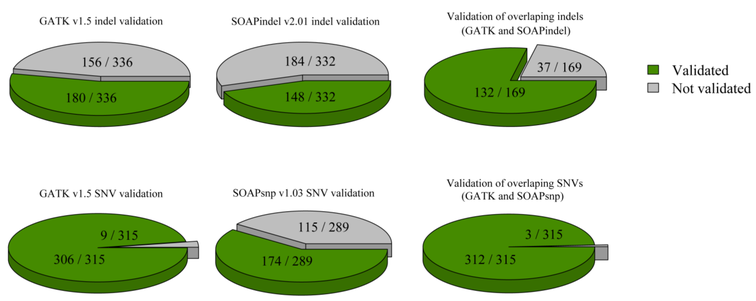


- **SNP concordance between the illumina data calls and the Complete Genomics v2.0 data calls was calculated for a single sample, “k8101-49685”.**
- **There are 2085 SNVs that Complete Genomics v2.0 detected but are not detected by any of the five Illumina data pipelines, despite high mappability among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



- **Indel concordance 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).**

- **MiSeq validation was performed on a combination of SNPs and indels chosen (1756 in total) from 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 is relatively high, and we agree that sequencing and analyzing samples with multiple platforms and methodologies is needed to attain a high accuracy “personal genome”.

References

1. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:btp324 [pii]10.1093/bioinformatics/btp324 (2009).
2. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).
3. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:btp352 [pii]10.1093/bioinformatics/btp352 (2009).
4. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967 (2009).
5. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res* **19**, 1124-1132, (2009).
6. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**, 265-272, doi:gr 097261.109 [pii]
7. Clement, N. L. *et al.* The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics* **26**, 38-45, doi:btp614 [pii]10.1093/bioinformatics/btp614 (2010).
8. Wei, Z., Wang, W., Hu, P., Lyon, G. J. & Hakonarson, H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic acids research* **39**, e132, doi:10.1093/nar/gkr599 (2011).