Next Generation Sequencing (NGS): Introduction, Data Formats, Processing & Variant Detection: Part II

Dr Stephen J Newhouse
Lead Data Scientist & Senior Bioinformatician @ NHIR BRC-MH SlaM NHS & IoPPN KCL & UCL Farr Institute
stephen.j.newhouse@kcl.ac.uk, @s_j_newhouse

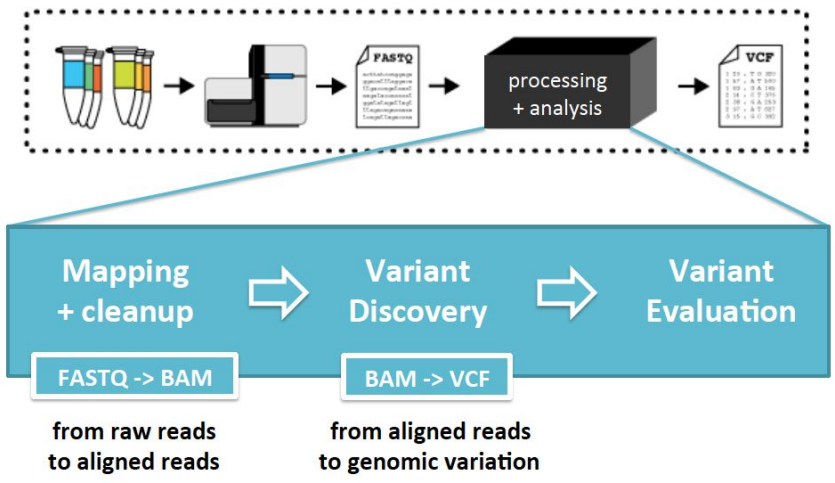
Genomic Medicine MSc
16 Feb 2016

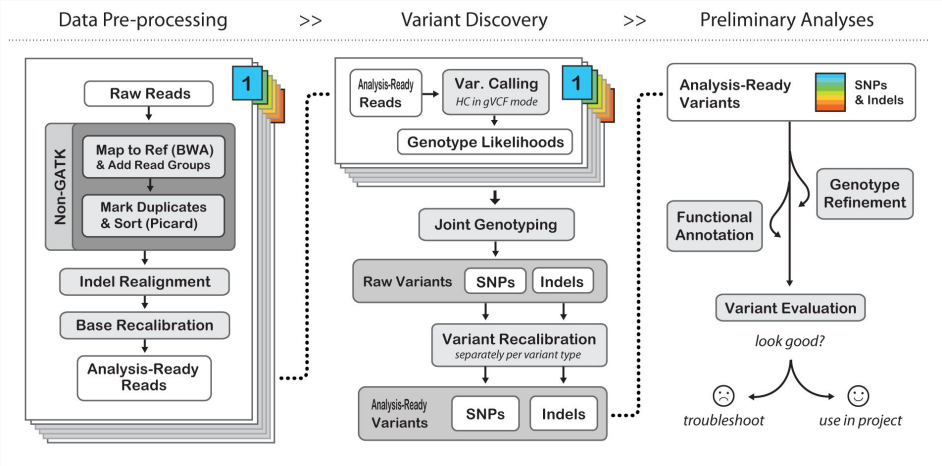
Overview

1. Variant Calling
2. Types of Variants
3. Tools for Variant Calling
4. Things to Think About (Depth, Quality, Errors?)

Genomic Medicine & NGS

- **The Ultimate goal is to detect clinically relevant variants from sequencing data**





Tools for Variant Calling

- **basics**: all look for difference between reference genome and the mapped short read
  - exact maths/algorithms beyond the scope of the course
- **Popular Tools**
  - GATK : HaplotyperCaller
  - Freebayes
  - VarDict
  - samtools
  - platypus
  - ...and many more...
- **Specialist Tools for Copy Number variation and Structural Variants**
  - Lumpy
  - cnvkit
  - ...and many more...

Types of variants detected by NGS

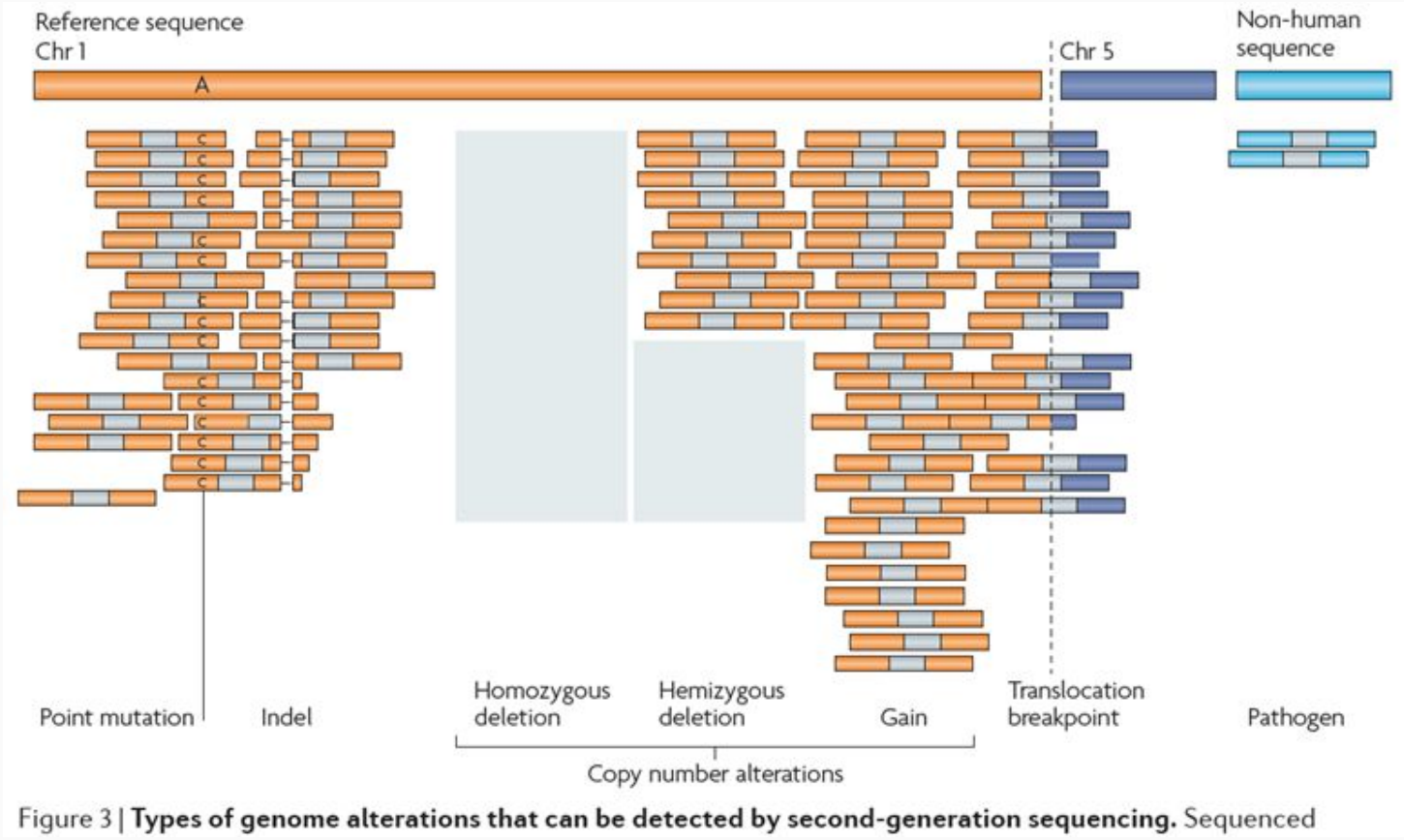
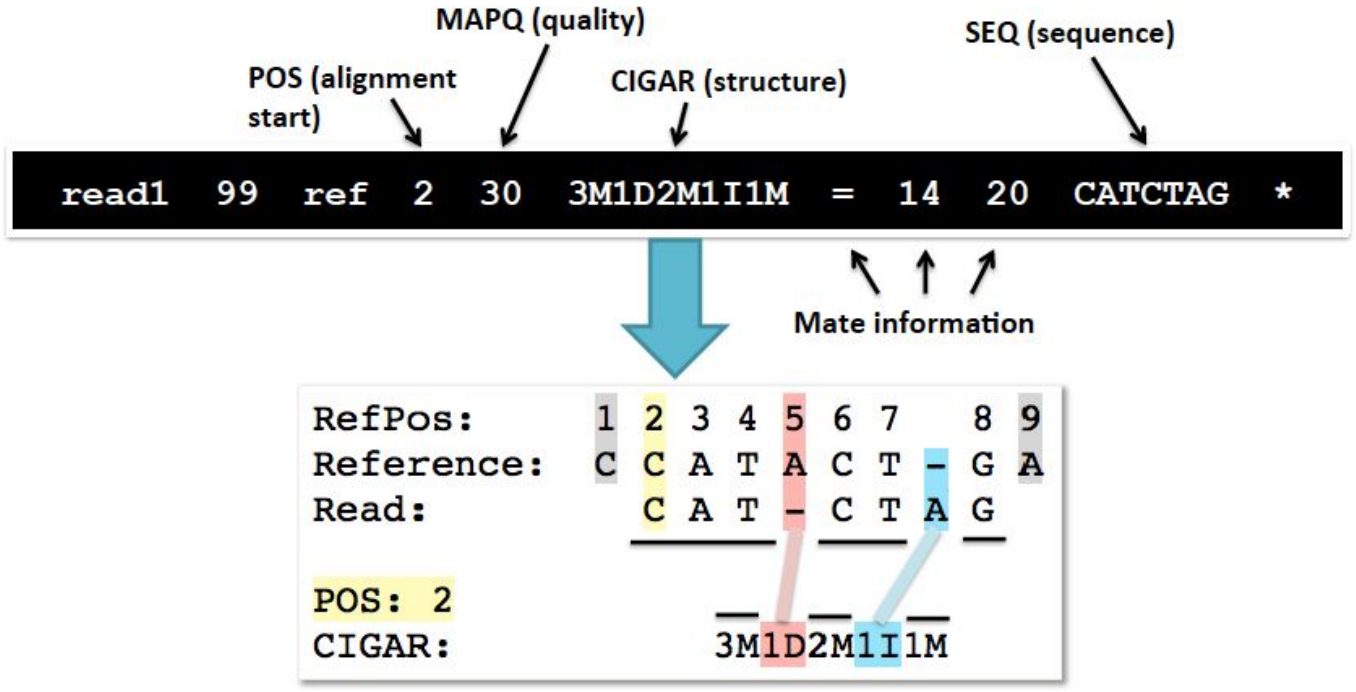


Figure 3 | **Types of genome alterations that can be detected by second-generation sequencing.** Sequenced

Mapping produces SAM summarizing position, quality, structure **AND ANY DIFFERENCES** for given sequence alignment...

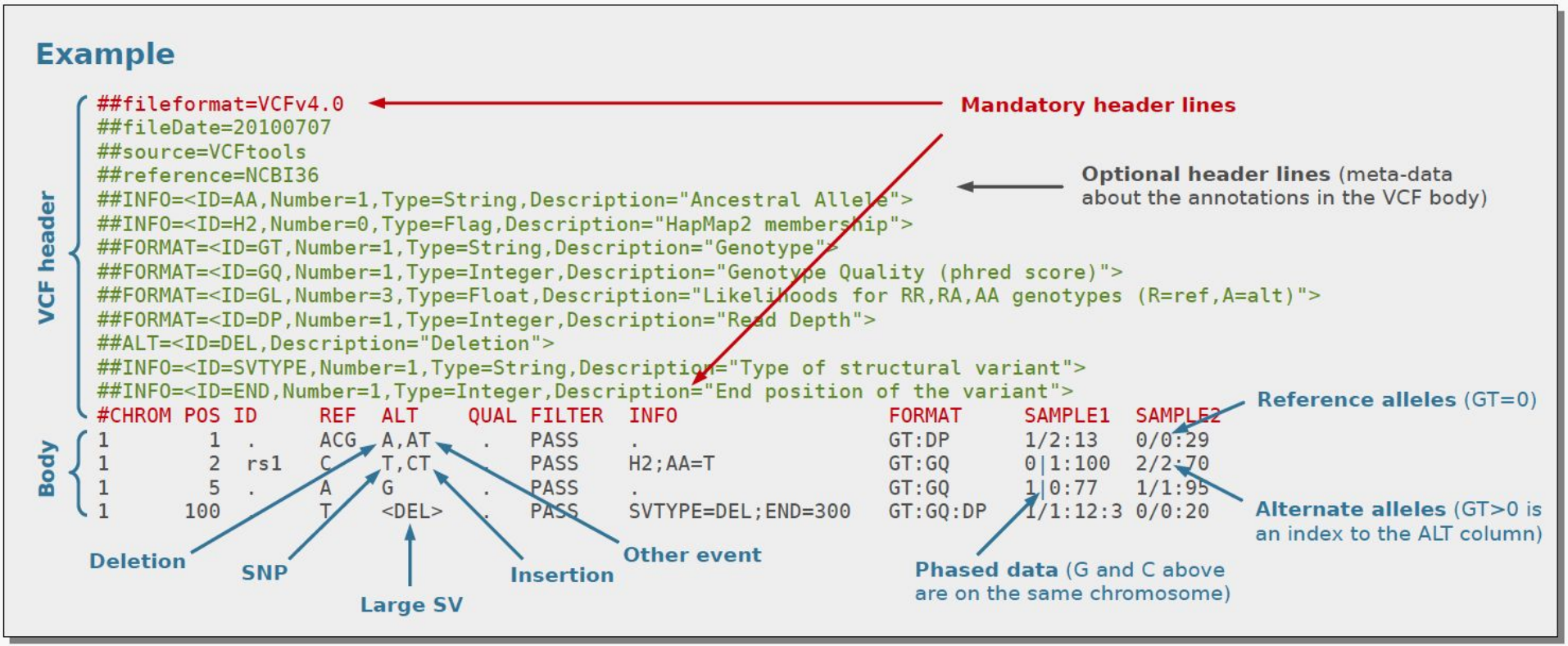


See also:

- SAM format spec: http://samtools.github.io/hts-specs/SAMv1.pdf
- Explain SAM flags: http://broadinstitute.github.io/picard/explain-flags.html

VCF Format: Stores variant information

https://vcftools.github.io/specs.html



http://vcftools.sourceforge.net/VCF-poster.pdf
https://samtools.github.io/hts-specs/VCFv4.2.pdf

VCF Format: Stores variant information
https://vcftools.github.io/specs.html

Types of variants

SNPs

Alignment
```
ACGT
ATGT
```

VCF representation

POS	REF	ALT
2	C	T

Insertions

Alignment
```
AC-GT
ACTGT
```

VCF representation

POS	REF	ALT
2	C	CT

Deletions

Alignment
```
ACGT
A--T
```

VCF representation

POS	REF	ALT
1	ACG	A

Complex events

Alignment
```
ACGT
A-TT
```

VCF representation

POS	REF	ALT
1	ACG	AT

Large structural variants

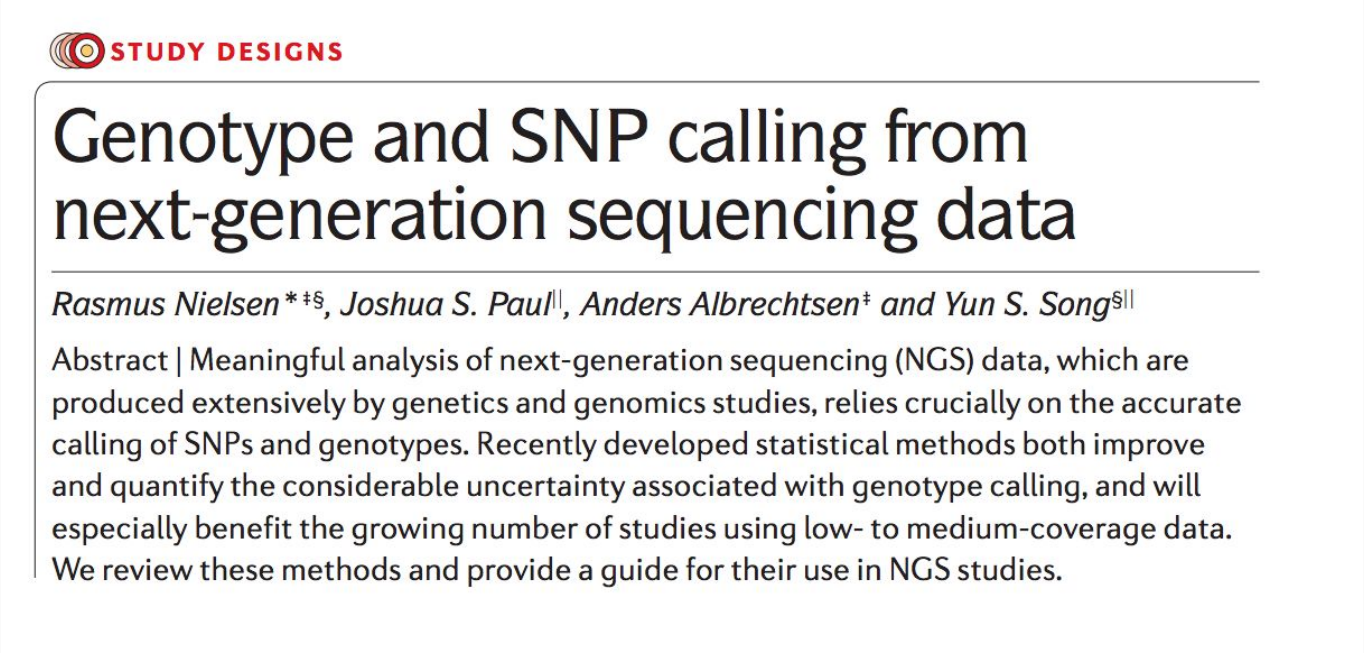
VCF representation

POS	REF	ALT	INFO
100	T	<DEL>	SVTYPE=DEL;END=300

http://vcftools.sourceforge.net/VCF-poster.pdf
https://samtools.github.io/hts-specs/VCFv4.2.pdf

Variant Calling

- SNP & Indel calling with **high coverage data** is relatively robust



STUDY DESIGNS

Genotype and SNP calling from next-generation sequencing data

Rasmus Nielsen[*‡§], *Joshua S. Paul*[‖], *Anders Albrechtsen*[‡] *and Yun S. Song*[§‖]

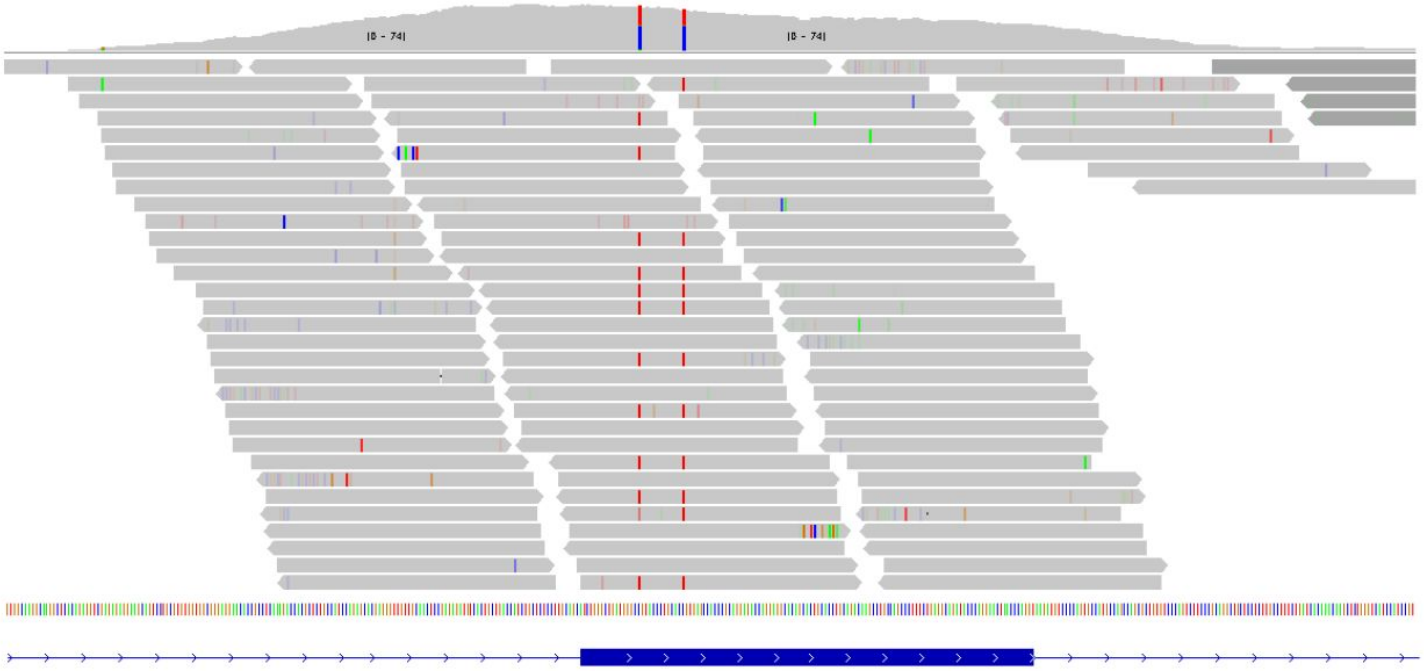
Abstract | Meaningful analysis of next-generation sequencing (NGS) data, which are produced extensively by genetics and genomics studies, relies crucially on the accurate calling of SNPs and genotypes. Recently developed statistical methods both improve and quantify the considerable uncertainty associated with genotype calling, and will especially benefit the growing number of studies using low- to medium-coverage data. We review these methods and provide a guide for their use in NGS studies.

1. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12,** 443–51 (2011).

Variant Calling: IGV Visualisation of a Heterozygous SNP

Variant Calling II

Heterozygous SNP

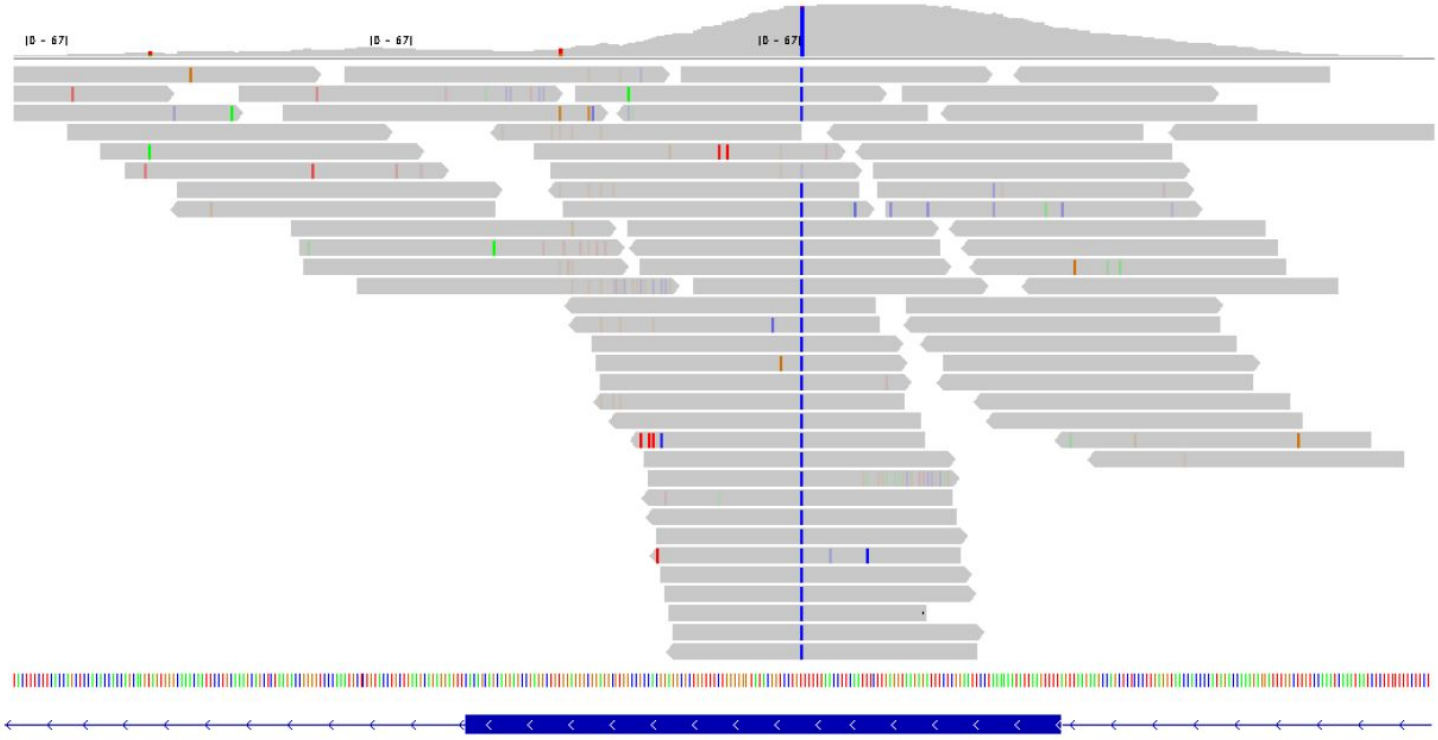


Thanks to Dr Mike Simpson (michael.simpson@kcl.ac.uk)

Variant Calling: IGV Visualisation of a Homozygous SNP

Variant Calling III

Homozygous SNP



Thanks to Dr Mike Simpson (michael.simpson@kcl.ac.uk)

Variant Calling is not perfect...

- **Post variant call filtering required!**
- Beware False Positives & False Negatives
- Need to review variant calls:
  - **Low-complexity (LC)** filter: filtering variants overlapping with low-complexity regions
  - **Maximum depth (MD**) filter: filtering sites covered by excessive number of reads.
  - **Allele balance (AB)** filter: filtering sites where the fraction of non-reference reads is too low.
  - **Double strand (DS)** filter: filtering variants if either the number of non-reference reads on the forward strand or on the reverse strand is below a certain threshold.
  - **Fisher strand filter (FS)**: filtering sites where the numbers of reference/non-reference reads are highly correlated with the strands of the reads.
  - **Quality filter (QU)**: filtering sites with the reported variant quality below a threshold

A look at improving variant call filtering in WGS...

BIOINFORMATICS REVIEW

Vol. 30 no. 20 2014, pages 2843–2851
doi:10.1093/bioinformatics/btu356

Sequence analysis

Advance Access publication June 27, 2014

Toward better understanding of artifacts in variant calling from high-coverage samples

Heng Li

Medical Population Genetics Program, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Whole-genome high-coverage sequencing has been widely used for personal and cancer genomics as well as in various research areas. However, in the lack of an unbiased whole-genome truth set, the global error rate of variant calls and the leading causal artifacts still remain unclear even given the great efforts in the evaluation of variant calling methods.

Results: We made 10 single nucleotide polymorphism and INDEL call sets with two read mappers and five variant callers, both on a haploid human genome and a diploid genome at a similar coverage. By investigating false heterozygous calls in the haploid genome, we identified the erroneous realignment in low-complexity regions and the incomplete reference genome with respect to the sample as the two major sources of errors, which press for continued improvements in these two areas. We estimated that the error rate of raw genotype calls is as high as 1 in 10–15 kb, but the error rate of post-filtered calls is reduced to 1 in 100–200 kb without significant compromise on the sensitivity.

Availability and implementation: BWA-MEM alignment and raw variant calls are available at http://bit.ly/1g8XqRt scripts and miscellaneous data at https://github.com/lh3/varcmp.

Contact: hengli@broadinstitute.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 3, 2014; revised on May 9, 2014; accepted on May 19, 2014

artifacts such as the non-random distribution of variants, dependent errors, incomplete reference genome and copy number variations. An improved version is to incorporate real variants instead of using simulated variants (Talwalkar *et al.*, 2013), but it does not address the artifacts caused by large-scale effects either. A better simulation is to take the reads sequenced from one sample with a finished genome, map them to another finished genome, call variants and then compare the calls to the differences found by genome-to-genome alignment (Li *et al.*, 2008). However, this approach is limited to small haploid genomes. There are attempts to apply a similar idea to mammalian genomes (Bolosky *et al.*, Unpublished data; Li *et al.*, 2013), but as the mammalian reference genomes are frequently incomplete and the whole-genome alignment is imperfect, such a simulation is still different from realistic scenarios.

The difficulties in simulation have motivated us to focus more on real data. One simple approach is to thoroughly sequence a small target region with mature technologies, such as the Sanger sequencing technology, and take the resultant sequence as the ground truth (Harismendy *et al.*, 2009). It does not capture large-scale artifacts, though. Another more commonly used method is to measure accuracy either by comparing variant calls from different pipelines, or by comparing calls to variants ascertained with array genotyping or in another study (Boland *et al.*, 2013; Cheng *et al.*, 2014; Clark *et al.*, 2011; Goode *et al.*, 2013; Lam *et al.*, 2012a,b; Li, 2012; Liu *et al.*, 2013; O'Rawe *et al.*, 2013;

A look at improving variant call filtering in WES...

Carson *et al. BMC Bioinformatics* 2014, **15**:125
http://www.biomedcentral.com/1471-2105/15/125

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Effective filtering strategies to improve data quality from population-based whole exome sequencing studies

Andrew R Carson[1†], Erin N Smith[1†], Hiroko Matsui[1], Sigrid K Brækkan[2,3], Kristen Jepsen[1], John-Bjarne Hansen[2,3] and Kelly A Frazer[1,4,5,6*]

Abstract

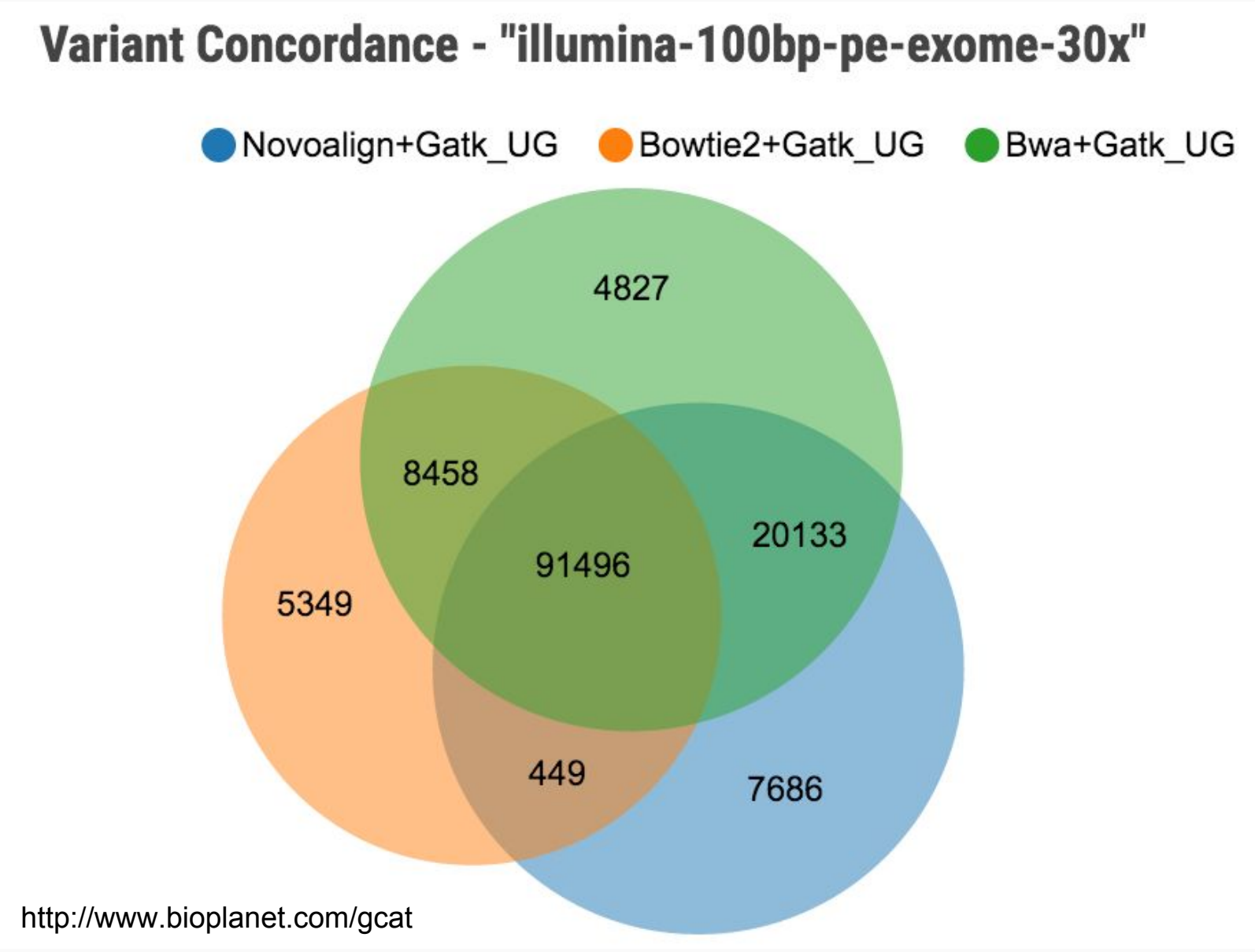
Background: Genotypes generated in next generation sequencing studies contain errors which can significantly impact the power to detect signals in common and rare variant association tests. These genotyping errors are not explicitly filtered by the standard GATK Variant Quality Score Recalibration (VQSR) tool and thus remain a source of errors in whole exome sequencing (WES) projects that follow GATK's recommended best practices. Therefore, additional data filtering methods are required to effectively remove these errors before performing association analyses with complex phenotypes. Here we empirically derive thresholds for genotype and variant filters that, when used in conjunction with the VQSR tool, achieve higher data quality than when using VQSR alone.

Results: The detailed filtering strategies improve the concordance of sequenced genotypes with array genotypes from 99.33% to 99.77%; improve the percent of discordant genotypes removed from 10.5% to 69.5%; and improve the Ti/Tv ratio from 2.63 to 2.75. We also demonstrate that managing batch effects by separating samples based on different target capture and sequencing chemistry protocols results in a final data set containing 40.9% more high-quality variants. In addition, imputation is an important component of WES studies and is used to estimate common variant genotypes to generate additional markers for association analyses. As such, we demonstrate filtering methods for imputed data that improve genotype concordance from 79.3% to 99.8% while removing 99.5% of discordant genotypes.

Conclusions: The described filtering methods are advantageous for large population-based WES studies designed to identify common and rare variation associated with complex diseases. Compared to data processed through standard practices, these strategies result in substantially higher quality data for common and rare association analyses.

Keywords: Next generation sequencing, Single nucleotide variants, Genotyping, Imputation, Genomics

not all tools find the same variants: Illumina 100bp PE 30x Whole Exome - 3 tools



http://www.bioplanet.com/gcat

not all tools find the same variants…..

RESEARCH Open Access

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], Zhi Wei[4], Kai Wang[8,9*] and Gholson J Lyon[1,2,9*]

Abstract

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.

Methods: We sequenced 15 exomes from four families using commercial kits (Illumina HiSeq 2000 platform and Agilent SureSelect version 2 capture kit), with approximately 120X mean coverage. We analyzed the raw data using near-default parameters with five different alignment and variant-calling pipelines (SOAP, BWA-GATK, BWA-SNVer, GNUMAP, and BWA-SAMtools). We additionally sequenced a single whole genome using the sequencing and analysis pipeline from Complete Genomics (CG), with 95% of the exome region being covered by 20 or more reads per base. Finally, we validated 919 single-nucleotide variations (SNVs) and 841 insertions and deletions (indels), including similar fractions of GATK-only, SOAP-only, and shared calls, on the MiSeq platform by amplicon sequencing with approximately 5000X mean coverage.

Results: SNV concordance between five Illumina pipelines across all 15 exomes was 57.4%, while 0.5 to 5.1% of variants were called as unique to each pipeline. Indel concordance was only 26.8% between three indel-calling pipelines, even after left-normalizing and intervalizing genomic coordinates by 20 base pairs. There were 11% of CG variants falling within targeted regions in exome sequencing that were not called by any of the Illumina-based exome analysis pipelines. Based on targeted amplicon sequencing on the MiSeq platform, 97.1%, 60.2%, and 99.1% of the GATK-only, SOAP-only and shared SNVs could be validated, but only 54.0%, 44.6%, and 78.1% of the GATK-only, SOAP-only and shared indels could be validated. Additionally, our analysis of two families (one with four individuals and the other with seven), demonstrated additional accuracy gained in variant discovery by having access to genetic data from a multi-generational family.

Conclusions: Our results suggest that more caution should be exercised in genomic medicine settings when analyzing individual genomes, including interpreting positive and negative findings with scrutiny, especially for indels. We advocate for renewed collection and sequencing of multi-generational families to increase the overall accuracy of whole genomes.

Variant callers : a comparison study...

OPEN ACCESS Freely available online

PLOS ONE

Variant Callers for Next-Generation Sequencing Data: A Comparison Study

Xiangtao Liu[1,2], Shizhong Han[1,2], Zuoheng Wang[3], Joel Gelernter[1,2,4], Bao-Zhu Yang[1,2*]

1 Department of Psychiatry, Division of Human Genetics, Yale University School of Medicine, New Haven, Connecticut, United States of America, **2** VA CT Health Care Center, West Haven, Connecticut, United States of America, **3** Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, United States of America, **4** Departments of Genetics and Neurobiology, Yale University School of Medicine, New Haven, Connecticut, United States of America

Abstract

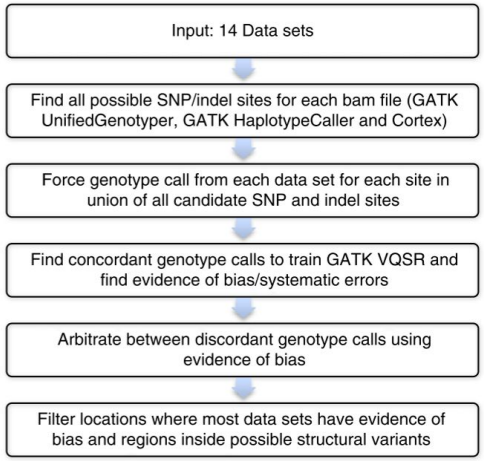
Next generation sequencing (NGS) has been leading the genetic study of human disease into an era of unprecedented productivity. Many bioinformatics pipelines have been developed to call variants from NGS data. The performance of these pipelines depends crucially on the variant caller used and on the calling strategies implemented. We studied the performance of four prevailing callers, SAMtools, GATK, glftools and Atlas2, using single-sample and multiple-sample variant-calling strategies. Using the same aligner, BWA, we built four single-sample and three multiple-sample calling pipelines and applied the pipelines to whole exome sequencing data taken from 20 individuals. We obtained genotypes generated by Illumina Infinium HumanExome v1.1 Beadchip for validation analysis and then used Sanger sequencing as a “gold-standard” method to resolve discrepancies for selected regions of high discordance. Finally, we compared the sensitivity of three of the single-sample calling pipelines using known simulated whole genome sequence data as a gold standard. Overall, for single-sample calling, the called variants were highly consistent across callers and the pairwise overlapping rate was about 0.9. Compared with other callers, GATK had the highest rediscovery rate (0.9969) and specificity (0.99996), and the Ti/Tv ratio out of GATK was closest to the expected value of 3.02. Multiple-sample calling increased the sensitivity. Results from the simulated data suggested that GATK outperformed SAMtools and glfSingle in sensitivity, especially for low coverage data. Further, for the selected discrepant regions evaluated by Sanger sequencing, variant genotypes called by exome sequencing versus the exome array were more accurate, although the average variant sensitivity and overall genotype consistency rate were as high as 95.87% and 99.82%, respectively. In conclusion, GATK showed several advantages over other variant callers for general purpose NGS analyses. The GATK pipelines we developed perform very well.

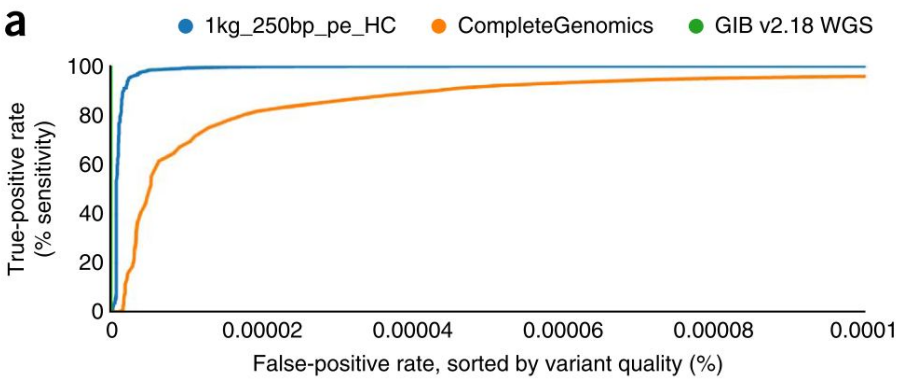
Integrating human NGS datasets for benchmarking Variant calls

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

Justin M Zook[1], Brad Chapman[2], Jason Wang[3], David Mittelman[3,4], Oliver Hofmann[2], Winston Hide[2] & Marc Salit[1]

Clinical adoption of human genome sequencing requires methods that output genotypes with known accuracy at millions or billions of positions across a genome. Because of substantial discordance among calls made by existing sequencing methods and algorithms, there is a need for a highly accurate set of genotypes across a genome that can be used as a benchmark. Here we present methods to make high-confidence, single-nucleotide polymorphism (SNP), indel and homozygous reference genotype calls for NA12878, the pilot genome for the Genome in a Bottle Consortium. We minimize bias toward any method by integrating and arbitrating between 14 data sets from five sequencing technologies, seven read mappers and three variant callers. We identify regions for which no confident genotype call could be made, and classify them into different categories based on reasons for uncertainty. Our genotype calls are publicly available on the Genome Comparison and Analytic Testing website to enable real-time benchmarking of any method.





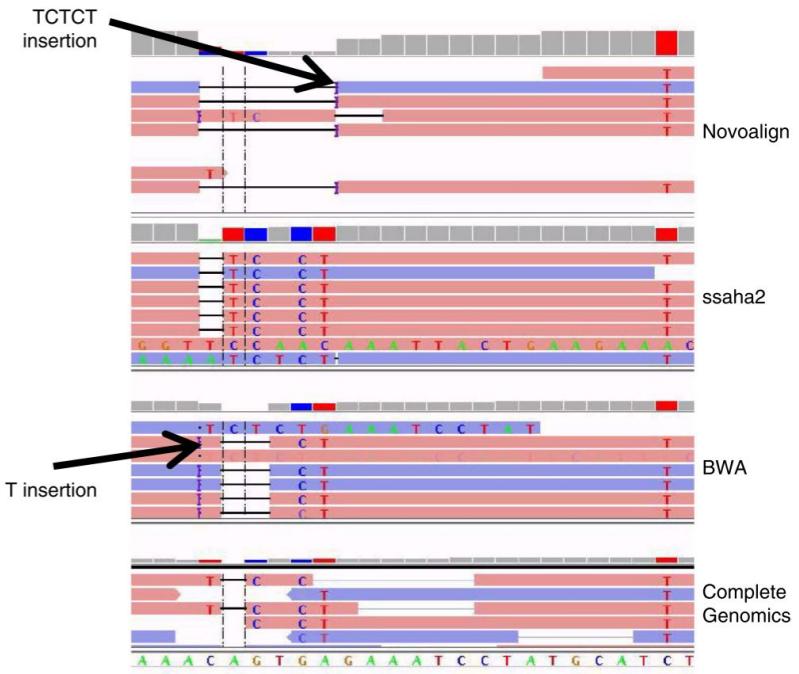


Figure 2 Complex variants have multiple representations. Example of complex variant with four different representations from four different mappers, which can cause data sets to appear to call different variants when in reality they are the same variant. In this case, the six bases CAGTGA are replaced by the five bases TCTCT at location 114841792–114841797 on chromosome 1. The four sets of reads are from Illumina mapped with BWA, 454 mapped with ssaha2, Complete Genomics mapped with CGTools, and Illumina mapped with Novoalign.

1. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32,** 246–51 (2014).

ARTICLE

Received 15 Sep 2014 | Accepted 13 Jan 2015 | Published 25 Feb 2015

DOI: 10.1038/ncomms7275 OPEN

An analytical framework for optimizing variant discovery from personal genomes

Gareth Highnam[1], Jason J. Wang[1], Dean Kusler[1], Justin Zook[2], Vinaya Vijayan[3], Nir Leibovich[1] & David Mittelman[1,3]

The standardization and performance testing of analysis tools is a prerequisite to widespread adoption of genome-wide sequencing, particularly in the clinic. However, performance testing is currently complicated by the paucity of standards and comparison metrics, as well as by the heterogeneity in sequencing platforms, applications and protocols. Here we present the genome comparison and analytic testing (GCAT) platform to facilitate development of performance metrics and comparisons of analysis tools across these metrics. Performance is reported through interactive visualizations of benchmark and performance testing data, with support for data slicing and filtering. The platform is freely accessible at http://www.bioplanet.com/gcat.

GCAT: Genome Comparison & Analytic Testing - BioPlanet

www.bioplanet.com/gcat

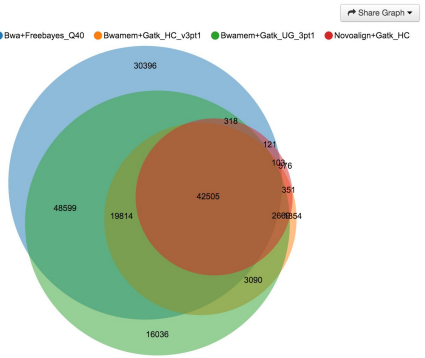
LINK TO REPORT: http://goo.gl/5P8eGQ

The variant calls in this report have been analyzed across a series of metrics including comparisons to the Genome in a Bottle (GIB) call set, consistency with genotyping array data, and concordance with other variant callers. Use the filtering options on the left sidebar to compare against other runs and to filter different views of the data.

Share Table

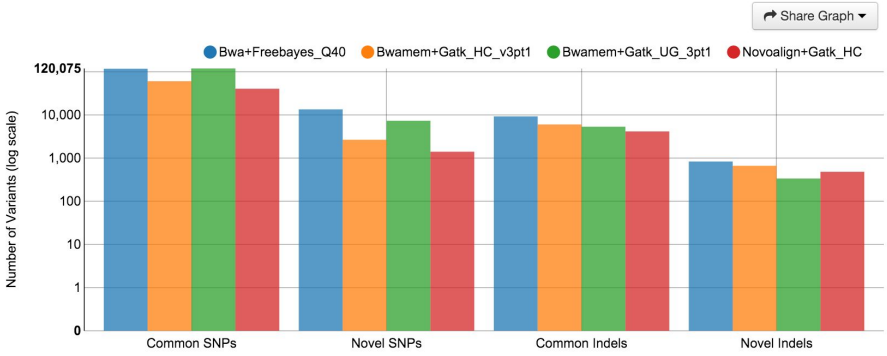
Pipeline	GIB Sensitivity	GIB Specificity	Ti/Tv	SNPs	Indels	Novel %
Bwa+Freebayes_Q40	81.34%	99.9978%	1.942	131,884 (92.70%)	10,148 (7.13%)	14,417 (10.13%)
Bwamem+Gatk_HC_v3pt1	62.77%	99.9983%	2.298	63,574 (90.43%)	6,718 (9.56%)	3,340 (4.75%)
Bwamem+Gatk_UG_3pt1	85.21%	99.9974%	2.141	127,447 (95.71%)	5,707 (4.29%)	7,710 (5.79%)
Novoalign+Gatk_HC	44.85%	99.9990%	2.324	42,141 (90.04%)	4,660 (9.96%)	1,898 (4.06%)

The venn diagram shows the concordance of variant calls between the selected runs. Use the filtering options on the left to change which runs are compared, and the subset of data you are interested in. Overlapping regions indicate calls with high concordance and thus higher confidence in the accuracy.



Variant Types

This chart shows the breakdown of the variant classes by SNPs, Insertions, and Deletions.

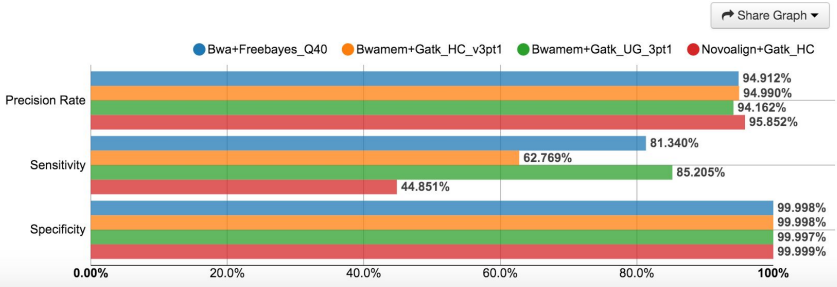


Mutation Recall: GENOME IN A BOTTLE

Variant calls are compared against version 2.18 of the NIST Genome in a Bottle highly confident SNP and indel call set. The call set was generated by integrating 14 different sequencing datasets from 5 different sequencing technologies, and removing regions with unresolved differences or where all datasets have evidence of bias. We filter for the subset of loci within the sequencing capture and regions confidently identified in the Genome in a Bottle call set, and for the comparisons on GCAT we remove regions that contain complex variants (I.e., nearby SNPs and indels) due to difficulties in regularizing the representation of complex variants. Visit Genome in a Bottle for more info and to download vcf and bed files.

Share Table

Pipeline	True+	False+	True-	False-	Het-Ref	Het-HomVar	HomVar-Het	HomVar-Ref
Bwamem+Gatk_UG_3pt1	19,760	1,225	46,467,312	3,431	213	744	249	13
Bwa+Freebayes_Q40	18,953	1,016	46,467,521	4,348	114	241	451	15
Bwamem+Gatk_HC_v3pt1	14,713	776	46,467,761	8,727	26	667	81	2
Novoalign+Gatk_HC	10,652	461	46,468,076	13,098	20	407	31	3



http://goo.gl/5P8eGQ

Optimizing variant discovery & genotyping: Take home message...

- **Essential**:LCR, Quality & Depth Filters
- **Research = Variant Discovery:** Run multiple mapping tools and variant calling tools and take a consensus
    - eg. https://bcbio.wordpress.com/2014/10/07/joint-calling/
- **Genomic Medicine = targeted genotyping:** we are only looking at specific regions and variants
- With sufficient quality & depth we will always find the variant?
- Still need QC pipeline (Genomics England QC pipeline - see lecture by Tim Hubbard)

Genomics England 100K QC Pipeline

- **Automatic stats and checks**
  - Delivery integrity – custom Pipeline Pilot protocol
  - BAM and VCF files – picard ValidateSAMFile & bcftools
  - Coverage (perc_bases_ge_15x_mapQ_ge11 > 95%) – samtools stats
  - Number of bases (GbQ30NoDupsNoClip > 85G) – custom pysam based script
  - Samtools stats – samtools stats
  - BCFtools stats – bcftools stats
  - Verifybamid
- **Semi-automatic stats and checks**
  - Sex
  - Mendelian errors
  - Inbreeding estimates
  - IBD estimation
  - Ancestry

Further Reading

- See all slide links & Handbook & https://bcbio.wordpress.com & https://www.mendeley.com/groups/8265471/genomic-medicine-msc-bioinformatics-module-kcl/papers/