Building a High-performance Pipeline for Analysis and Management of Whole Exome Sequencing Data

Riyue Bao[1], Lei Huang[1], Elizabeth Bartom[1], Wenjun Kang[1], Kyle Hernandez[1], Gang Feng[2], Hongmei Jiang[3], Jorge Andrade[1]

[1] Center for Research Informatics, University of Chicago, Chicago, IL; [2] Clinical and Translational Sciences Institute, Northwestern University, Chicago, IL; [3] Department of Statistics, Northwestern University, Evanston, IL

ABSTRACT

Whole exome sequencing (WES) has facilitated discovery of inherited and novel genetic alterations associated with diseases at low cost, high efficiency and reliability. Discovery and hypothesis -based data analyses reply on accurate identification of variants, which are compromised by limited concordance among variant callers.

We developed ExScalibur, an automated and scalable WES pipeline capable of processing hundreds of human exomes in parallel. It consists of raw read quality control (QC), preprocessing, read alignment, InDel realignment and base quality score recalibration (BQSR), multi-sample variant calling / joint genotyping (SNPs and InDels), variant annotation and filtration. We integrated three aligners (BWA aln/mem, Bowtie2, Novoalign) and six variant callers (GATK UnifiedGenotyper (UG), GATK HaplotypeCaller (HC), FreeBayes, Atlas2, SAMtools, Isaac Variant Caller (IVC)) to generate up to eighteen sets of calls. The joined list of variants is ranked through majority voting to produce a set of consensus calls. It is further filtered to remove common variants in 1000 Genomes and ESP6500. The final list of variants is prioritized based on deleterious prediction, conservation, Combined Annotation Dependent Depletion (CADD), as well as gene-level network and pathway analyses. The results are stored in a database (VariantDB), and a web interface is provided for fast retrieval and comparison.

We evaluated the performance of our pipeline with a selection of three aligners (BWA mem, Novoalign, Bowtie2) and four callers (GATK HC, FreeBayes, SAMtools, IVC) on simulated and benchmark datasets. It demonstrated high sensitivity (99.40% for SNPs, 95.37% for InDels), specificity (> 99.99% for both SNPs and InDels) and precision (99.53% for SNPs, 98.34% for InDels). We have used ExScalibur to confidently identify candidate genomic mutations in rare genetic disorders, familial diseases and cancer predisposition syndromes.

IMPLEMENTATION

High-level Modularization

- Six major modules and twenty-two sub-modules
- Modules include QC, preprocessing, alignment, postprocessing, variant analysis, as well as collection of analysis statistics and visualization of results
- Robotic for pipeline maintenance and new tool extension

Massive Parallelization

- 91% of the sub-modules support parallelization
- Two running modes: splitChrom or wholeGenome
- In-house parallization functions for single-threaded tools
- Fast turnaround time (120 human exomes in 16-24 hours on a 1024-core HPC)

Easy to Use

- Two input files: (1) metadata table with sample information (2) config file with pipeline and tool-specific parameters
- Aligners/callers specified at pipeline command line
- Minimal requirement of human intervention
- Real-time progress monitoring of pipeline runs

Comprehensive representation of results

- Data files: QC reports, alignment BAM, coverage BigWig, variant VCF, annotation, custom filtered and prioritized variant lists, etc.
- Collection of statistics: alignment and variant call stats, exome coverage, quality distribution, etc.
- Visualization: alignment, coverage and variants (UCSC Genome Browser data hub), distribution of strand-specific genome coverage, insert size, exome coverage (criViz)
- Complete record of commands and logs for reproducible analysis

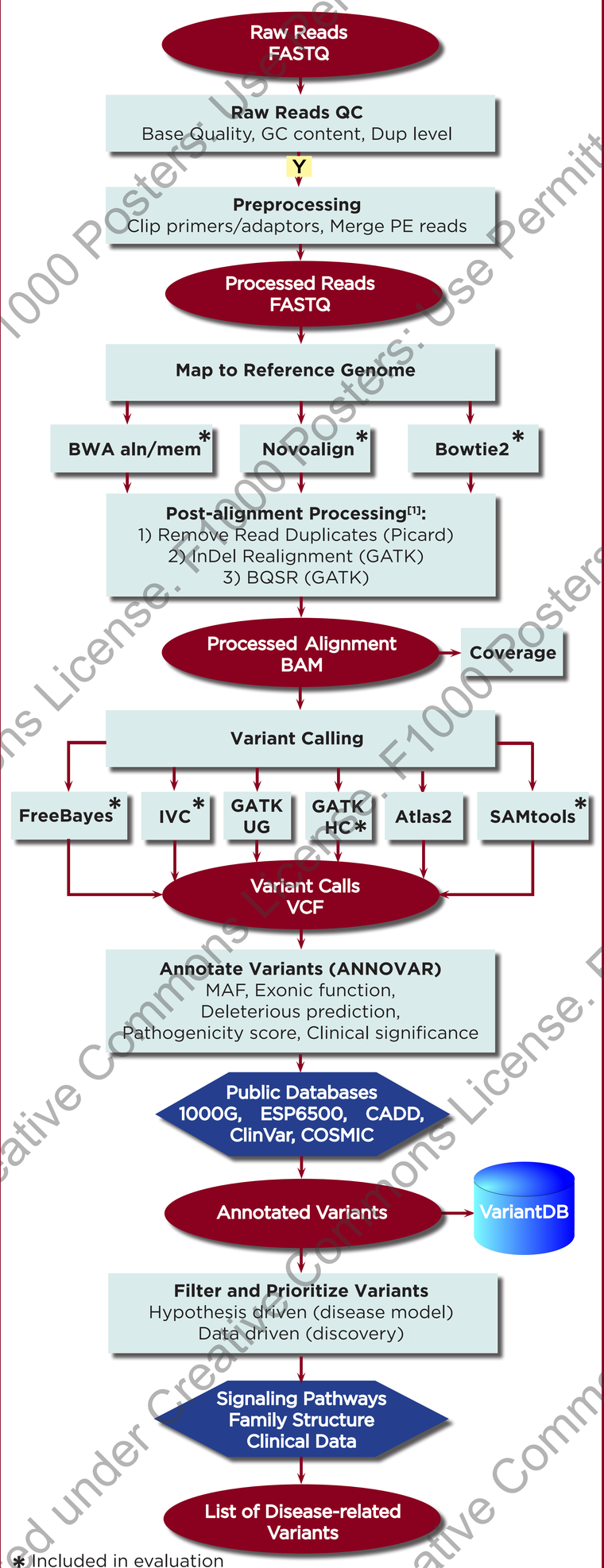
REFERENCES

[1] Van der Auwera, G., et al. (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline, Current Protocols in Bioinformatics, 43, 11.10.11-11.10.33.

[2] Zook, J.M., et al. (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls, Nature biotechnology, 32, 246-251.

[3] Li, H. (2014) Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples, Bioinformatics, pii: btu356. [Epub ahead of print].

EXSCALIBUR WORKFLOW



* Included in evaluation

METHODS

Simulation dataset

We simulated 100bp paired-end (PE) reads at 50x coverage from target regions[a] on chromosome 1 of hg19 genome assembly using DWGSIM. The parameters were set as base error rate (0.0001-0.005), mutation rate (0.001), indel fraction (0.05), and random DNA read probability (0.01). A total of 4,000,000 reads were mapped to hg19 and those with mapping quality (MapQ) < 30 were removed.

Benchmark dataset

High-confidence variant calls of NA12878 were obtained from NIST-GIAB datasets (v2.18) and used as benchmark in this study[2]. For pipeline evaluation, we downloaded 50x NA12878 WES data from the HapMap project (SRX079575). A total of 170,987,444 50bp PE reads were mapped to hg19 at 85% success rate with MapQ > 30.

Evaluation

Twelve sets of variant calls were generated with three aligners and four callers[b]. Off-target variants and those with read depth lower than 6x were removed from subsequent analysis. For NA12878, variants were further filtered to exclude those located within genomic regions where no confident calls could be made[2]. Sensitivity, specificity and precision were calculated for: (1) all united variants, (2) variants with PASS[c], (3) 2aligner x 2caller, and (4) single aligner + single caller.

[a] SeqCap EZ Human Exome Library v2.0
[b] Programs: BWA mem (v0.7.9a), Novoalign (v3.02.05), Bowtie2 (v2.2.3), GATK HC (v3.1.1), FreeBayes (v0.9.9), SAMtools (v0.1.19), IVC (v1.0.6)
[c] PASS filters include: low QUAL, low DP, strand bias, SNP cluster, etc.[3]

RESULTS

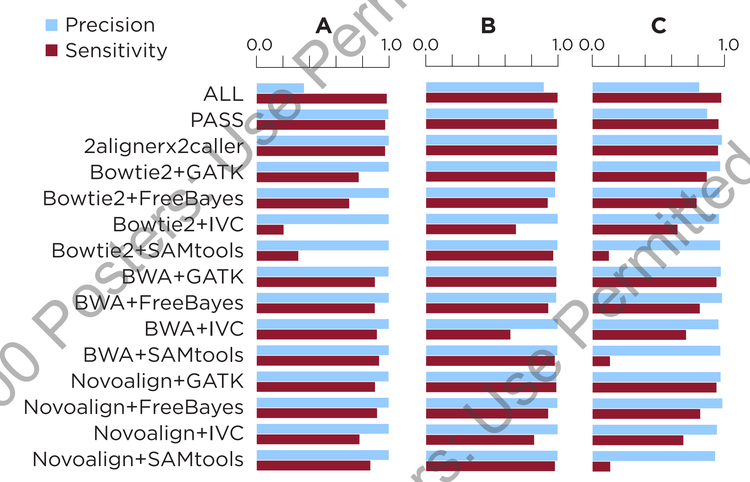


Figure 1. Precision and sensitivity of variant calls detected in the simulation dataset for SNPs (A) and in the benchmark dataset for SNPs (B) and InDels (C). ALL: all united variants. PASS: variants that pass the PASS filters. 2alignerx2caller: variants that pass the PASS filters and are detected by at least two callers and two aligners.

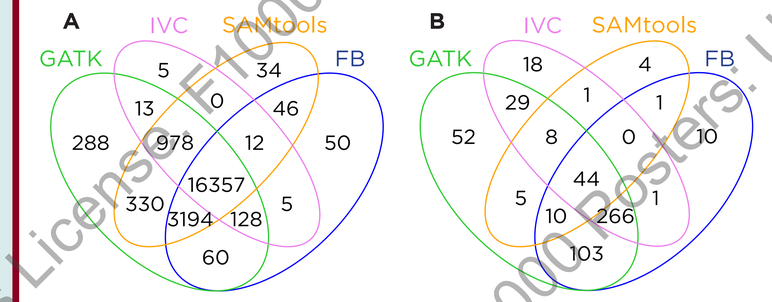


Figure 2 Concordance of calls between variant callers for SNPs (A) and InDels (B) in the benchmark dataset. Variants detected in Novoalign alignment are shown (filtered by PASS). Similar patterns were observed in other aligners (data not shown). FB represents FreeBayes.

ACKNOWLEDGEMENT

This research was supported by the Biological Sciences Division and the Institute for Translational Medicine/CTSA (NIH UL1 RR024999) at The University of Chicago.

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES

bioinformatics@bsd.uchicago.edu

CRI Bioinformatics Core | 900 East 57th Street | KCBD 5132 | Chicago, IL 60637

cri.uchicago.edu