OXFORD

Sequence analysis

FermiKit: assembly-based variant calling for Illumina resequencing data

Heng Li

Genomics Platform, Broad Institute, Cambridge, MA 02142, USA

Abstract

Summary: FermiKit is a variant calling pipeline for Illumina whole-genome germline data. It *de novo* assembles short reads and then maps the assembly against a reference genome to call SNPs, short insertions/deletions and structural variations. FermiKit takes about one day to assemble 30-fold human whole-genome data on a modern 16-core server with 85 GB RAM at the peak, and calls variants in half an hour to an accuracy comparable to the current practice. FermiKit assembly is a reduced representation of raw data while retaining most of the original information.

Availability and implementation: https://github.com/lh3/fermikit

Contact: hengli@broadinstitute.org

1 Introduction

Deep resequencing of a human sample typically results in a BAM file of 60–100 GB in size. Storing, distributing and processing many such huge files is becoming a burden for sequencing facilities and research labs. While better compression helps to alleviate this issue, it adds processing time and can barely halve the size, which does not keep up with the rapidly increasing sequencing throughput. Illumina and GATK use gVCF (Raczy *et al*., 2013) as a reduced representation of raw data. However, gVCF is reference dependent and it is nontrivial to encode both large and small variants consistently. We still need to go back to raw data for long events and when upgrading the reference genome. Another idea from the past practice is to assemble sequence reads into contigs that ideally retains all information in the raw data, but whether this approach is practical to Illumina human resequencing remains to be confirmed.

2 Methods

FermiKit uses BFC (Li, 2015) for error correction, ropeBWT2 (Li, 2014a) for BWT construction, Fermi (Li, 2012) for *de novo* assembly, BWA-MEM (Li, 2013) for mapping and HTSBox (http://bit.ly/HTSBox) for variant calling. The caller simply parses edits in the 'pileup' output for small variant calling from one or multiple BAMs, and extracts alignment break points for SV calling, though it may misclassify some SV events (Trappe *et al*., 2014). FermiKit sets

thresholds on mapping quality and the number of supporting reads without using sophisticated statistical models. In comparison to our earlier work (Li, 2012), FermiKit is faster, more sensitive (due to better error correction) and more complete as a pipeline.

FermiKit does not use paired-end information for the time being, but this does not have a great impact on its power empirically. With longer upcoming Illumina reads, it will be actually preferred to merge overlapping ends and treat them as single-end reads.

3 Results

We have run FermiKit on multiple whole-genome datasets of sample NA12878 along with GATK-HaplotypeCaller (HC in brief) and FreeBayes (Garrison and Marth, 2012). We used Genome-In-A-Bottle (GIAB; Zook *et al*., 2014) as truth data to evaluate the accuracy (Table 1). Recent Illumina data have excessive systematic errors around poly-A which HC does not handle well. It called over 4000 false INDELs from sample S1+ and S4+ with the vast majority around poly-A. We excluded these regions to avoid one simple error source greatly affecting the metrics. After this treatment, variant callers are broadly comparable when the same set of hard filters are applied. VQSR as is advised in GATK Best Practice does not work well with single-sample calling.

GIAB was generated from multiple NA12878 call sets. It is potentially biased against new callers and biased towards easier regions that can be called by the existing callers. For example, the GATK

Table 1. GIAB evaluation of SNP/INDEL accuracy for sample NA12878

Sample	Caller	SNP-FN	SNP-FP	InDel-FN	InDel-FP
PG-	FermiKit	45 700	824	2324	472
	FreeBayes	21 548	439	3858	400
	HC+hardFilter	27 010	144	943	370
	HC+VQSR	128 604	1955	1423	366
S7-	FermiKit	65 217	531	2340	549
	FreeBayes	50 796	676	2891	420
	HC+hardFilter	66 847	228	1543	457
	HC+VQSR	103 979	1508	1396	605
S11-	FermiKit	91 468	541	2973	554
	FreeBayes	52 071	903	3195	431
	HC+hardFilter	65 223	407	1502	472
	HC+VQSR	111 504	1694	1175	765
S2-	FermiKit	63 445	448	2244	568
S12-	FermiKit	74 940	501	2562	553
S1+	FermiKit	67 816	455	4051	516
	FreeBayes	63 101	902	4625	436
	HC+hardFilter	71 174	531	2376	591
	HC+VQSR	108 101	8852	2377	1827
S4+	FermiKit	71 262	452	4197	536
	FreeBayes	65 427	1061	4781	437
	HC+hardFilter	75 040	672	2477	653
	HC+VQSR	103 595	10 492	2401	1622

PCR-free Platinum Genome NA12878 (PG-; AC:ERR194147; 100 bp reads), four Illumina X10 lanes of PCR-free NA12878 (S7-, S11-, S2- and S12- under BaseSpace project ID 18475457; 150 bp) and two X10 lanes of PCR-amplified NA12878 (S1+ and S4+ under project ID 8998991; 150 bp) were acquired and called with FermiKit-0.9, FreeBayes-0.9.20 (option: '–experimental-gls –min-repeat-entropy 1') and HC-3.3 (option: '-stand_e mit_conf 10 -stand_call_conf 30'). For FreeBayes and HC, BWA-MEM was used for mapping against GRCh37 plus decoy (http://bit.ly/GRCh37d5) with duplicates marked by Samblaster (Faust and Hall, 2014). Short variant calls were hard filtered with hapdip (http://bit.ly/HapDip). GATK-VQSR was also applied to HC calls. The filtered calls were compared to GIAB-v2.18 excluding poly-A regions longer than 6 bp plus 10 bp flanking. A true variant is counted as an FN if there are no called variants within 10 bp around the truth, and a called variant is counted as an FP if it falls in GIAB trusted regions and there are no true variants within 10 bp around the called variant.

call set available from the Platinum Genome website has 13 278 FN SNPs and 46 FPs out of 2.03 Gb confident regions (i.e. one SNP error per 44 Mb), which is overly good and is worrying.

We turned to the CHM1-NA12878 dataset (Li, 2014b) for an unbiased evaluation (Table 2). In this evaluation, FermiKit produces calls of higher specificity at the cost of sensitivity. This is probably because FermiKit is less powerful in repetitive or duplicated regions or regions affected by systematic artefacts. Nonetheless, in well-behaved regions that are outside 'uniMask', the loss of sensitivity is minor. The gain in precision is significant if we consider that there may be 5–20 k real heterozygous SNPs in CHM1 (Li, 2014b), which should not be counted as FPs.

FermiKit performs well in calling long deletions (Table 3). While it does not use read pairs, it achieves comparable sensitivity and higher specificity in comparison to the popular tools. FermiKit also called ~480 insertions over 100 bp and identified multiple kb-long contigs having poor alignments to GRCh37 but nearly perfect alignment to a PacBio assembly of CHM1 (AC:GCA_001007805.1). We also mapped the CHM1 FermiKit unitigs to the PacBio assembly and called 71 long deletions, 11 insertions and 262 other events. As PacBio assemblies are generally of higher quality, these numbers

Table 2. Evaluation on SNP/INDEL accuracy with CHM1-NA12878 pair

Caller	Filter	SNP-TP	SNP-FP	InDel-TP	InDel-FP
FermiKit	hard-polyA	1 937 469	22 743	230 955	14 602
	uniMask	1 802 820	9507	127 304	1126
FreeBayes	hard-polyA	2 026 883	59 422	190 587	30 909
	uniMask	1 842 634	15 252	117 764	6329
HC	hard-polyA	2 003 655	32 030	267 870	15 541
	uniMask	1 824 658	14 912	133 458	2046

SNP/INDELs were called from the CHM1 (AC:SRR642636 through SRR642641; 100 bp) and NA12878-PG- BWA-MEM alignments used by Li (2014b). On the assumption that CHM1 is haploid, (heterozygous) FP equals the number of CHM1 heterozygotes and (heterozygous) TP equals the number of NA12878 heterozygotes minus the number of CHM1 heterozygotes. Two sets of filters were applied for filtering. 'Hard-polyA' is the same as the filter used in Table 1. 'UniMask' filters out genomic regions that tend to be repetitive, low-complexity or susceptible to copy number changes or systematic artefacts (http://bit.ly/unimask). This filter is sample independent.

Table 3. Performance on calling long deletions over 100 bp

Sample	Caller	1000 g pilot	Ensemble	LUMPY	Merged
S7-	FermiKit	0.43/0.23	0.50/0.15	0.32/0.23	0.58/0.09
S1+	FermiKit	0.43/0.22	0.51/0.15	0.33/0.23	0.58/0.10
PG-	FermiKit	0.43/0.20	0.52/0.14	0.34/0.22	0.59/0.09
	DELLY	0.47/0.34	0.50/0.22	0.31/0.28	0.58/0.16
	LUMPY	0.72/0.34	0.76/0.29	0.68/0.37	0.79/0.20

FermiKit was used to call 100 bp–100 kb deletions from the PG-, S7- and S1+ datasets. DELLY (Rausch *et al.*, 2012) and LUMPY PG- calls were acquired from http://bit.ly/bcbsval (B. Chapman, personal communication). For all call sets, overlapping events were merged and deletions longer than 100 kb were discarded. The two numbers in a cell at row *R* and column *C* give the false negative rate and false positive rate of call set *R*, assuming truth set *C* is correct and complete. In the table, truth set '1000g pilot' consists of 3142 deletions by Mills *et al.* (2011) and further validated by Layer *et al.* (2014); 'Ensemble' contains 4095 validated calls by multiple callers; 'LUMPY' consists of 2657 validated LUMPY-only deletions; 'Merged' is the union of all the three truth sets above, containing 4695 deletion calls.

give a rough estimate on the number potential false positives of Fermi on a haploid dataset.

4 Conclusions

A FermiKit assembly is about 3 GB compressed. After assembly, single-sample variants can be obtained in half an hour to high accuracy through mapping against a reference genome. Jointly calling 261 aligned whole-genome samples only took ~40 CPU hours. FermiKit is a viable option for aggressive data compression, greatly reducing the efforts and expense on data storage, distribution and re-analyses at an acceptable cost of information loss.

Funding

References

Faust,G.G. and Hall,I.M. (2014) Samblaster: fast duplicate marking and structural variant read extraction. *Bioinformatics*, 30, 2503–2505.

Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907*.

Layer,R.M. *et al*. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.

Li,H. (2012) Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*, **28**, 1838–1844.

Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997*.

Li,H. (2014a) Fast construction of FM-index for long sequence reads. *Bioinformatics*, **30**, 3274–3275.

Li,H. (2014b) Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, **30**, 2843–2851.

Li,H. (2015) BFC: correcting illumina sequencing errors. *arXiv:1502.03744*.

Mills,R.E. *et al*. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Raczy,C. *et al*. (2013) Isaac: ultra-fast whole-genome secondary analysis on illumina sequencing platforms. *Bioinformatics*, **29**, 2041–2043.

Rausch,T. *et al*. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.

Trappe,K. *et al*. (2014) Gustaf: Detecting and correctly classifying svs in the ngs twilight zone. *Bioinformatics*, **30**, 3484–3490.

Zook,J.M. *et al*. (2014) Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls. *Nat. Biotechnol.*, **32**, 246–251.