strandngs

Fast and Accurate Variant Calling in Strand NGS

A benchmarking study

Radhakrishna Bettadapura, Shanmukh Katragadda, Vamsi Veeramachaneni,
Atanu Pal, Mahesh Nagarajan and Ramesh Hariharan

Abstract

We present a workflow that generates fast, highly accurate variant callsets from paired-end DNA samples. When measured against the Genome in a Bottle Callset, precision/recall rates on the precisionFDA whole genome sample are 98.5/ 99.5% on the entire callset and 97.2/ 96.2% on indels. Against the GATK best-practices workflow, whole exome concordances are similarly excellent and persist across a variety of capture techniques and samples. Depending on the sample, Strand NGS is up to twice as fast as GATK best-practices. Additionally, each stage of the Strand NGS workflow supports in-depth visualization and interrogation tools absent from its GATK counterpart. Finally, the DNA sequence analysis workflow is also heavily storage-optimized, making only incremental demands on disk space beyond the alignment stage. The DNA-Seq workflow presented here is part of Strand NGS v3.0, our flagship bioinformatics tool.

DNA-Seq in Strand NGS

Calling variants (Figure S) from paired-end data starts with alignment, in which reads are mapped to the reference genome on a per-fragment basis. After mapping, reads are sorted by chromosome and position, and duplicates removed. Reads are pairwise duplicate if they have same start, mate start, and alignment length. Duplicate removal retains at most one copy of pairwise duplicate reads, with ties broken by mapping quality and average base quality in that order.

Local realignment follows deduplication. Candidate haplotypes are identified in heuristically determined windows, and each read in the window is aligned against the reference as well as several candidate haplotypes. The haplotype chosen for any given read maximizes a probabilistic Phred-scaled metric.

The final stage is variant calling. A Bayesian variant caller is used to determine the Phred-scaled probability that a given variant is a mutation. The Bayesian prior matrix is computed from empirically determined quantities such as the Ti/Tv ratio and the heterozygosity ratio. The resulting variant callset contains single and multi base substitutions, indels, and complex variants, each of which is assigned a Phred score proportional to the probability that it is a variant.

Comparison methods

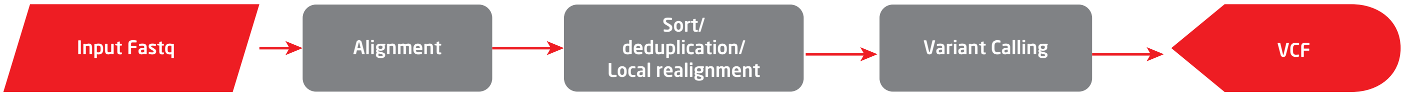

Figure S. The Strand NGS workflow.

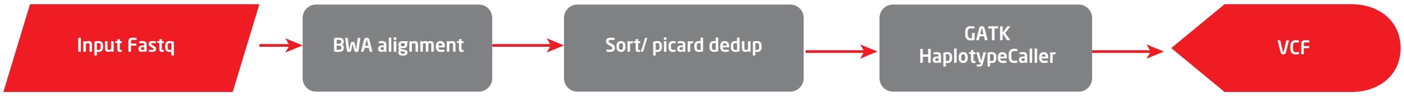

Figure G. The GATK best-practices workflow, without base quality recalibration.

We measure the accuracy of the DNA-Seq workflow in Strand NGS v3.0 (Figure S) in two ways. On the whole genome dataset (Table D), we do comparisons against the Genome in a Bottle callset [Zook et. al]; on three whole exome datasets (Table D), we use the GATK best practices workflow [Li, DePristo et. al] (Figure G). Both workflows involve paired-end read alignment, deduplication and variant calling. In lieu of the local reassembly step internal to the GATK haplotype caller, the Strand NGS workflow performs local realignment around indels.

We omit base quality recalibration from both workflows. This is due to a practical consideration: for relatively small whole exome samples, the statistical effect of recalibration is likely to be small. On the other hand, the whole genome sample on which we benchmark Strand NGS has relatively high base qualities, and calls are unlikely to be affected by recalibration. Note that Strand NGS supports base quality recalibration.

A paired-end sample subject to the workflows in Figures S and G generates a pair of variant callsets, respectively V_S and V_G. Both callsets are thresholded on locations with a minimum coverage of 10 reads. We chose a minimum cutoff of ten after considering high-confidence variants[1] in the Genome in a Bottle callset: an overwhelming majority (>=99.99%) of these have a Strand NGS read depth of 10 or greater. Further, all whole-exome samples (Table D) in this study have a mean coverage significantly greater than 10.

Using V_G as a baseline, the rtg [Cleary et. al] tool is used to compute the accuracy of V_S on a suitable subset of of the whole genome. In addition, for the whole genome precisionFDA sample, the set V_G corresponds to the Genome in a Bottle (GiaB) v3.2.2 callset, restricted as before to locations with a minimum Strand NGS coverage of 10 reads.

For a given input sample, we adopt a widely-accepted subsetting approach [see, for instance, Chapman et. al]. The subset is the intersection of regions R_m and R_h, where R_m is the manifest specified by the capture method used to sequence the sample, and R_h is a set of genomic high confidence regions. High confidence regions are those on which comparisons can be reliably made; we use v3.2.2 of the GiaB high confidence set [Zook et. al]. In the case of the precisionFDA whole genome sample (Table D), there is no manifest, so we use the high confidence set directly.

Accuracy is measured with respect to precision and recall. Precision is the ratio of the number of true positives to the number of true and false positives; recall is the ratio of true positives to the number of true positives and false negatives. Both figures are expressed as a percentage.

DNA-Seq in Strand NGS

Table D: Datasets, capture methods and manifest sizes

ID	Capture method/Type	L/C/S (Gbp)	Individual	$\|R_m\|$ (Mbp)	$\|R_m \cap R_h\|$ (Mbp)
SRR2106343 (SRA)	Agilent SureSelect Exome v6/WEX	100bp/ 100x/8.8	NA12878	74.5	67.4
SRR2106344 (SRA)	Illumina Nextera Rapid Capture/WEX	100bp/ 100x/16.3	NA12878	45	39.4
DRR039932 (SRA)	Roche Nimblegen v3.0/WEX	159bp/70x/15.3	NA18948[2]	64.19	55.2
HG001 (SRA)	Illumina TruSeq PCR-free/WGS	150bp/50x/162	NA12878	NA	2529

Table D. The paired-end whole exome datasets used in this study. L, C and S stand respectively for the length of each read in the dataset, the mean coverage and the size of the dataset. R_m and R_h stand respectively for the manifest and the high confidence regions.

Datasets

[1] High-confidence variants: variants in high-confidence regions.

[2] The only sample from an individual other than NA12878. Nonetheless, we use GiaB 3.2.2 high confidence regions for comparison: while high-confidence callsets are unique to a given individual, it is probable that high-confidence regions are conserved across a large population.

Results

Accuracy

Table A: SNP and Indel Accuracy (100% callset)

Sample	P/R (%)	S_G/I_G
Agilent SSCR/WEX	99.43/ 98.0[3]	32,579/ 2,333
Illumina Nextera/WEX	98.9/ 97.1[3]	22,738/ 1,467
Roche Nimblegen/WEX	98.6/ 96.6[3]	42,222/ 3,280
Illumina TruSeq	98.5/ 99.5	3,485,648/ 332,298

Table A. Strand NGS precision and recall on three whole exome samples and one whole genome sample, subsetted by the capture manifest restricted to the Genome in a Bottle high confidence regions. S_G and I_G stand respectively for the number of substitutions and indels in the truth callset.

Table A(i): Indel Accuracy (≈10% callset)

Sample	P/R	P^n/R^n	L_{tpi}
Agilent SSCR/WEX	97.8/ 91.9[3]	98.22/ 94.6	36
Illumina Nextera/WEX	95.3/ 86.6[3]	97.8/ 93.5	31
Roche Nimblegen/WEX	96.1/ 86.3[3]	98/ 91	49
Illumina TruSeq	97.2/96.2	99/ 97.8	53

Table A(i). Strand NGS indel precision and recall on three whole exome samples and one whole genome sample, subsetted by the capture manifest restricted to the Genome in a Bottle high confidence regions. The superscript "n" denotes comparisons after disregarding homopolymer regions and other low-complexity stretches in the reference. L_{tpi} is the length of the longest true positive indel detected by Strand NGS. Indels constitute about 10% of the entire callset.

[3] Lower-bounds on precision and recall. Since these figures use the GATK best-practices callset as a baseline, the true precision and recall are probably higher, especially for indels.

The results, in Table A and Table A(i), show that Strand NGS v3.0 produces callsets with consistently high precision and recall. On the precisionFDA whole genome sample, callset precision and recall is close to 99% overall, whereas indel precision and recall is at 97 and 96% respectively. Precision and recall remain high on the whole exome samples as well. Callset accuracy is greater than indel accuracy, suggesting that indels remain harder to call despite significant advances in variant calling.

The whole exome figures in Table A and A(i) are lower bounds on precision and recall. This is because the GATK callset used as baseline for those experiments has not itself been subject to the same type of cross-validation as the whole genome GiaB callset. This lack of baseline callset validation, combined with the significantly higher figures we achieve for the precisionFDA sample (Table A and A(i)), means that the true Strand NGS precision/recall rates are probably higher for whole exomes as well.

Speed

Strand NGS v3.0 is significantly faster than the GATK best-practices workflow. On typical whole exome samples, Strand NGS v3.0 takes less than half the time taken by its GATK counterpart (Table SE). Times for whole-genome samples are favourable, too: on the precisionFDA sample, Strand NGS v3.0 is 1.5 times as fast overall, with significant speedups in the post alignment stage (Figure SW). These fast execution times are despite the in-depth visualization and interrogation support in Strand NGS, support that is entirely absent from the command-line GATK-best practices workflow.

Table SE: Whole-Exome Timings

	Strand NGS time (min)	BWA + GATK time (min)
Alignment	22	14
View + Sort	12	24
Dedup + realign (Strand NGS only)	18	13
Variant calling	03	69
Total time	55	120
Speedup factor	2.2	1

Table SE: Strand NGS DNA-Seq execution times on whole exome sample SRR2106343. See Appendix for system configuration. Note that realignment is not performed in the GATK best-practices workflow.

Table SW: Whole-Genome Timings

	Strand NGS time (hrs)	BWA + GATK time (hrs)
Alignment	8	10
View + Sort	5.3	8
Dedup + realign (Strand NGS only)	8.93	5
SNPs	1.83	12
Total time	24	35
Speedup factor	1.45	1

Table SW: Strand NGS DNA-Seq execution times on the whole genome precisionFDA sample. See Appendix for system configuration.

Storage

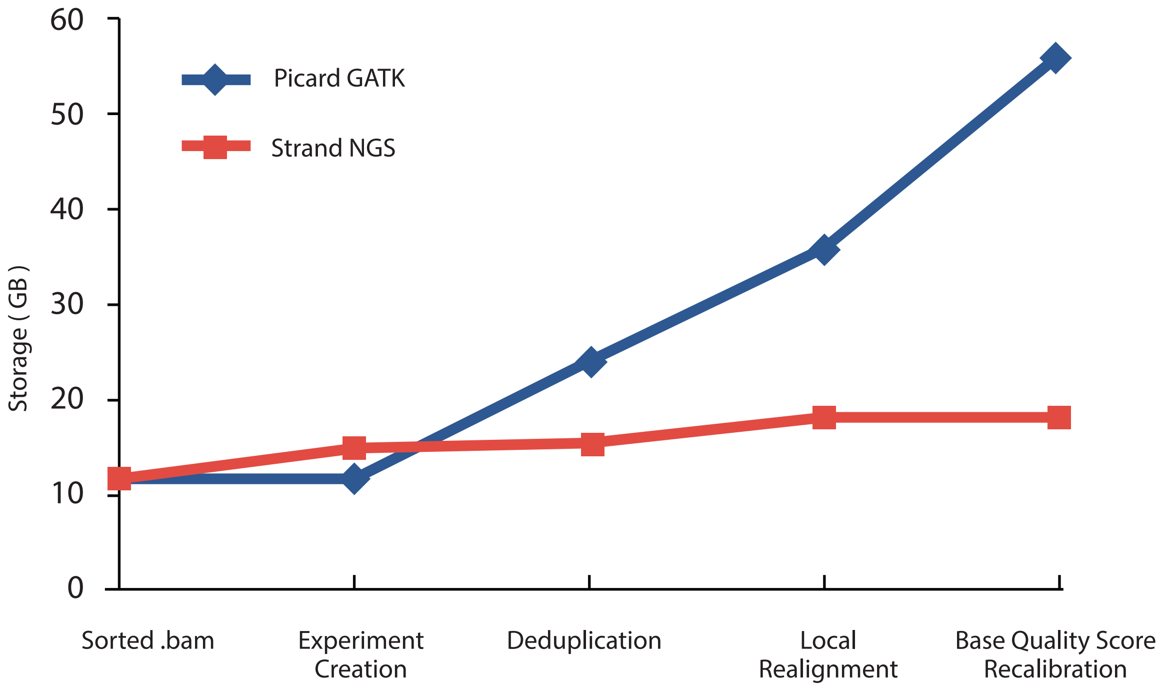

Figure ST: Strand NGS and GATK best-practices storage footprints for the post-alignment workflow.

Strand NGS is frugal with storage. On a typical whole exome sample (Figure ST), storage increases incrementally past the alignment stage. This is in contrast to open-source workflows like the GATK best practices, in which storage scales linearly with the number of stages involved.

Divergent variants

Much of the 1-3% divergence between the GATK best practices workflow and Strand NGS v3.0 is attributable to two causes: differences at the hom/het frontier and differences in repeat regions (Table A(i)).

1) **Hom/het frontier.** A high fraction of the divergent indels are called homozygous and heterozygous in GATK and Strand NGS respectively. In general, the Strand NGS workflow is more frugal with homozygous variants than GATK. The allele fraction threshold at which a Strand NGS heterozygous call turns homozygous ranges from 85 to 95%, and depends most importantly on base quality: the higher the base quality of the support, the higher the probability that Strand NGS calls a heterozygous variant.

2) **Repeat regions.** Two features of repeat regions lead to differences in called variants. The first relates to zygosity in homopolymer runs. Homopolymer runs lead to sequencing errors, which lead in turn to reads with the wrong number of repeat bases. These reads can sometimes lead to differences in zygosity, as they contribute to an allele different from the major allele at the location. A second feature of repeat stretches is that they are often covered by reads that specify multiple alleles. In these cases the alleles involved in the Strand NGS call may differ from the one in the GATK callset.

Conclusions

We have shown that Strand NGS produces fast and accurate SNP and indel calls from whole exome and whole genome samples while being frugal with storage. A majority of the remaining small discordance is due to differences in zygosity and in repeat stretches.

The benchmark we present here is significant for two reasons.

1) **Workflow-to-workflow.** The only commonality between the rival workflows evaluated here is the input sample, with accuracy measured between the output VCFs. This is in contrast to studies [Chapman et.al, some of the precisionFDA studies] where the aligner, usually BWA or bowtie, is also common, and only variant callers are compared. Our workflow-to-workflow benchmark ensures that the cumulative inaccuracy across the various complex stages in DNA-Seq is kept to a minimum.

2) **Local reassembly can be replaced by local realignment + locus-based calling.** The GATK HaplotypeCaller performs local reassembly to infer the most probable haplotype supporting a set of contiguous locations. Strand NGS replaces local reassembly with a local realignment step and a locus-based Bayesian SNP caller. Our benchmarks demonstrate the near equivalence of these approaches. More significantly, our approach is both faster and less demanding of computational resources, implying in turn a better scalability to larger as well as a greater number of samples.

References

[Chapman et. al] B. Chapman et. al, Blue-Collar Bioinformatics, bcbio-nextgen, https://github.com/chapmanb/bcbio-nextgen

[Cleary et.al] J. G. Cleary, R. Braithwaite, K. Gaastra, B. S. Hilbush, S. Inglis, S.A Irvine, A. Jackson, R. Littin, M. Rathod, D. Ware, J. M. Zook, L. Trigg, F. M. M. De La Vega. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines, bioRxiv 023754

[Li] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv preprint, 2013

[DePristo et. al] M. DePristo, E. Banks, R. Poplin, K. Garimella, J. Maguire, C. Hartl, A. Philippakis, G. Del Angel, M.A. Rivas, M. Hanna, A. McKenna, T. Fennell, A. Kernytsky, A. Sivachenko, K. Cibulskis, S. Gabriel, D. Altshuler and M. Daly, A framework for variation discovery and genotyping using next-generation DNA-Sequencing data, Nature Genetics, 43:491-498, 2011

[Zook et. al] J.M. Zook, B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide and M. Salit. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls, Nature Biotechnology, 32: 246-51, 2014

Appendix: System configuration and software versions

All experiments in this whitepaper were run on an server with a dual Intel Xeon E5-2680 (16 cores, 32 threads @ 2.7GhZ) with 64GB of 1600MhZ DDR3 RAM and RAID-5 network attached storage.

Timing and accuracy benchmarks were obtained on Strand NGS v3.0 as well as the third-party BWA, samtools, GATK and rtg software. Versions and computing resource limitations are in Table A-S.

Table A-S: Software versions and resource limitations

	Version	Subcommands	Number of threads	Memory (GB)
Strand NGS	3.0	-	16	32
bwa	0.7.12	mem	16	-
samtools	1.3.1-43	view/sort/rmdup	1/8/1	-
GATK	3.7	HaplotypeCaller	16	32
rtg	3.7.1	vcfeval	1	-

Table A-S: Versions of software used in this whitepaper.