GigaScience, 6, 2017, 1–7

doi: 10.1093/gigascience/gix091
Advance Access Publication Date: 18 September 2017
Technical Note

TECHNICAL NOTE

Lightning-fast genome variant detection with GROM

Sean D. Smith, Joseph K. Kawash and Andrey Grigoriev*

Department of Biology, Center for Computational and Integrative Biology, Rutgers University, 315 Penn St, Camden 08102, NJ, USA

*__Correspondence address.__ Andrey Grigoriev, Department of Biology, Center for Computational and Integrative Biology, Rutgers University, 315 Penn St, Camden, NJ 08102. Tel: 856-225-2960; Fax: 856-225-6312; E-mail: andrey.grigoriev@rutgers.edu

Abstract

Current human whole genome sequencing projects produce massive amounts of data, often creating significant computational challenges. Different approaches have been developed for each type of genome variant and method of its detection, necessitating users to run multiple algorithms to find variants. We present Genome Rearrangement OmniMapper (GROM), a novel comprehensive variant detection algorithm accepting aligned read files as input and finding SNVs, indels, structural variants (SVs), and copy number variants (CNVs). We show that GROM outperforms state-of-the-art methods on 7 validated benchmarks using 2 whole genome sequencing (WGS) data sets. Additionally, GROM boasts lightning-fast run times, analyzing a 50× WGS human data set (NA12878) on commonly available computer hardware in 11 minutes, more than an order of magnitude (up to 72 times) faster than tools detecting a similar range of variants. Addressing the needs of big data analysis, GROM combines in 1 algorithm SNV, indel, SV, and CNV detection, providing superior speed, sensitivity, and precision. GROM is also able to detect CNVs, SNVs, and indels in non-paired-read WGS libraries, as well as SNVs and indels in whole exome or RNA sequencing data sets.

Keywords: variant detection; GROM; SNVs; structural variants; indels; copy number variants; whole genome sequencing

Findings

Introduction

The 1000 Genomes Project [1] was launched in 2008 with the goal of producing and analyzing whole genome sequencing (WGS) for 1000 genomes. By 2016 decreasing costs and increasing sequencing throughput had led to an exponential increase in the size and scope of WGS projects from Human Longevity, Inc.'s 10 000 publicly available WGS genomes [2] to the United Kingdom's 100 000 Genomes Project [3] to even larger, though less-clearly defined, sequencing projects involving 1 000 000 participants proposed in the United States (Precision Medicine Initiative [4] and Million Veteran Program [5]) and China [6]. Such projects produce massive amounts of data, straining computational resources and requiring much faster methods than current capabilities [7].

Comprehensive analysis of genomic differences requires detection of a wide range of variants, including single nucleotide variations (SNVs), indels (insertions and deletions <50 bases), and larger copy number variants (CNVs) and structural variants (SVs), which include deletions, duplications, insertions, inversions, and translocations. Methods have been developed for each type of variant; subsequently, a typical WGS analysis workflow requires running multiple algorithms. A recent pipeline, SpeedSeq [8], focused on reducing the computational resources needed for WGS analysis, though it still employed 4 variant detection algorithms. This can be wasteful of computational resources due to repetitive input/output and analysis of the same read sequences by several algorithms.

We present our method, Genome Rearrangement OmniMapper (GROM), a novel comprehensive method of variant detection, combining mismatch, split-read, read pair, and read

Received: 7 May 2017; **Revised:** 28 July 2017; **Accepted:** 13 September 2017

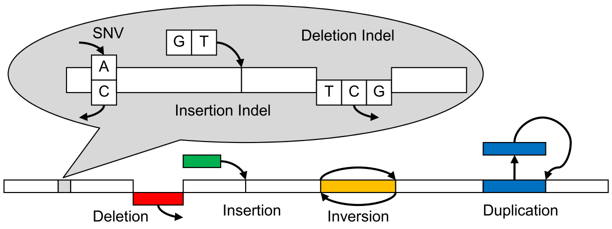

Figure 1: Examples of variants detected by GROM. GROM detects a comprehensive range of variants (SNVs, indels, deletions, insertions, inversions, and duplications). GROM also detects translocations spanning more than 1 chromosome (not shown).

depth WGS evidence. GROM boasts lightning-speed runtimes an order of magnitude faster than state-of-the-art variant detection pipelines. While drastically reducing computational time, GROM detects SNVs, indels, SVs, and CNVs in a single algorithm and provides superior overall variant detection compared with commonly employed algorithms.

Algorithm

Differences in variant types (Fig. 1) have resulted in separate algorithms designed for a limited range of variants. GROM achieves fast, comprehensive variant analysis via a compact workflow (Fig. 2), efficiently analyzing and gathering information at each reference base in 1 pass through a BAM file. Base information includes average mapping and base qualities, overlapping discordant pairs, unmapped mate reads, and split-reads, and read depth. Discordant pairs are identified based on abnormal read orientation or abnormal insert size. GROM determines abnormal insert size based on a sample of 10 million paired reads. Since insert size distributions tend to have right skewness, GROM calculates the median insert size and uses a rank-based method to determine abnormal insert size thresholds corresponding to 3 standard deviations from the median under a normal distribution (after outliers more than 5× the median insert size have been filtered). Each read with a split mapping, indel, discordant mate, or unmapped mate contributes breakpoint evidence to each potential reference base breakpoint. For simple cases such as a 2-base deletion within a read, there is 1 potential reference base start breakpoint and 1 potential reference base end breakpoint. Other cases may have less precise breakpoints, such as a read from a discordant deletion pair (abnormally large insert size). In this case, the exact breakpoint is unknown and a potential breakpoint is recorded for each reference base consistent with forming a concordant pair in the sample, where a concordant pair corresponds to insert sizes $\geq i_{min}$ and $\leq i_{max}$, where i_{min} and i_{max} represent the minimum and maximum insert size thresholds, respectively (Fig. 3). Using the deletion example in Fig. 3, a breakpoint distant from both reads would necessitate an insert size that is too large to be consistent with a concordant pair (and the source DNA fragment), and thus would not be a potential breakpoint. When soft-clipping (≥5 bases) or a split-read (each mapped split ≥20 bases)

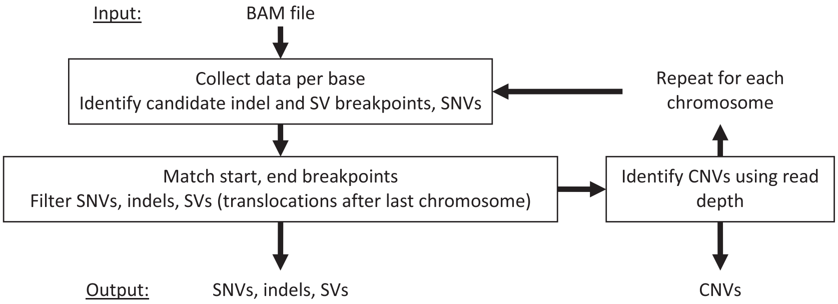
Input:
BAM file
Collect data per base
Identify candidate indel and SV breakpoints, SNVs
Repeat for each chromosome
Match start, end breakpoints
Filter SNVs, indels, SVs (translocations after last chromosome)
Identify CNVs using read depth
Output:
SNVs, indels, SVs
CNVs

Figure 2: GROM workflow. GROM simultaneously collects data for each reference base and identifies candidate breakpoints and SNVs in 1 pass through a BAM file. After each chromosome, SNVs are filtered; start and end breakpoints are matched and filtered for each indel and SV type (excluding translocations), and CNVs are identified (using read depth).

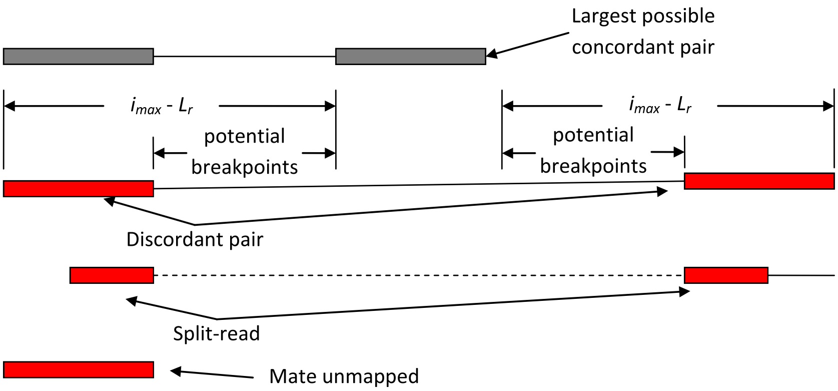

Figure 3: Example of SV evidence and potential breakpoints. GROM considers multiple input features at each reference base position to statistically determine the likelihood of an SNV, indel, SV, or CNV. Inputs in this example (discordant pairs, split-reads, and unmapped mate reads) are primarily used for SV detection. Discordant deletion pairs identified by insert size exceeding i_{max}. For discordant pairs, potential start and end breakpoints are recorded for each reference base capable of forming a concordant pair in the sample. L_r indicates read length.

occurs in the potential breakpoint region, the reference base immediately adjacent to the soft-clipping or split-read is recorded as a potential breakpoint and other potential breakpoints are recorded with half-weighting. This enables base resolution of breakpoints while limiting a single aberrant read mapping from misidentifying the true breakpoint.

Base by base of the reference, breakpoint evidence is stored for each distinct indel or SV. In some cases, it is difficult to distinguish variants. For instance, 2 heterozygous deletions may overlap and have similar start and end breakpoints and similar lengths. Thus, for each potential breakpoint, we cluster read evidence by variant type and length. Such clustering can be a computationally intensive task. We use the following efficient method.

We define a cluster or breakpoint cluster as a specific reference base location with a set of reads supporting a breakpoint at that location for a specific indel or SV type (deletion, duplication, etc.) of a certain length. A read from a discordant pair provides imprecise breakpoints and thus may be a member of multiple clusters, 1 cluster per reference location. A read is placed into an existing breakpoint cluster if the read and cluster support the same indel or SV type and the variant lengths are close, i.e.,

$$|L_{bc} - L_{disc}| \leq (i_{max} - i_{min} + i_{median} - 2L_r)\left(1 + \frac{1}{x_{bc}}\right), \tag{1}$$

where L_{bc} is the mean indel or SV length for the breakpoint cluster, L_{disc} is the length of the indel or SV pertaining to the candidate read, L_r is the read length, x_{bc} is the number of previously recorded reads supporting the breakpoint cluster, and i_{max}, i_{min}, and i_{median} are the maximum, minimum, and median concordant pair lengths, respectively. If a candidate read does not fit in any existing breakpoint clusters, a new cluster is created. If a candidate read fits in more than 1 breakpoint cluster at the same reference position, the breakpoint cluster with the most reads is chosen. This method is efficient and has the benefit of a read being considered in multiple clusters.

Additionally, the number of previously recorded reads influences whether a read is added to a breakpoint cluster because we expect our estimated (averaged) variant length to be closer to the true SV length as supporting reads are incorporated into the SV length average. For example, in Equation (1), let insert size statistics be such that $i_{max} - i_{min} + i_{median} - 2L_r = 500$, let an SV be a deletion of 1200 bases, and let our first discordant pair indicate an SV of length $L_{disc} = 1700$. One read is a poor estimate of the true SV length. Thus, in our example, the second read's SV length may differ from the first read's SV length by 1000 bases, $|1700 - L_{disc}| \leq 1000$. However, as the number of supporting reads increase, we expect the average SV length (L_{bc}) to converge to the true SV length of 1200, at which point we will not add the read as evidence unless its estimated SV length (L_{disc}) is within 500 bases of the true SV length, $|1200 - L_{disc}| \leq 500^*(1 + \varepsilon)$, where $\varepsilon \ll 1$.

For each reference base, a mismapping probability, p_{bc}, is calculated for each possible SNV, indel, and SV. p_{bc} is the binomial probability of at least x_{bc} reads supporting the breakpoint cluster given n_{bc} read depth and a mapping quality threshold m. Thus, p_{bc} indicates the likelihood that all of the supporting reads are mismappings. Read depth includes all mapped reads, unsequenced segments between concordant pairs, and potential breakpoints, and thus is an estimate of physical coverage. Physical coverage provides a more comprehensive representation of genome coverage than read coverage. It also helps GROM define deletion and duplication breakpoints when soft-clipping is unavailable as a decrease in coverage will affect breakpoint probability estimates. The mapping quality threshold m indicates the probability of a read mismapping, $p = 10^{-m/10}$. Thus, p_{bc} is given as

$$p_{bc} = \Pr(X \geq x) = 1 - \sum_{k=0}^{x-1} \binom{n}{k} p^k q^{n-k}, \tag{2}$$

where q = 1- p. To reduce computational time, binomial probability tables are precomputed and stored as data files. GROM will compute additional probability data files if the default mapping quality threshold ($m = 20$) is adjusted.

Potential indel and SV breakpoints are retained for further analysis. After processing reads for a chromosome (or the whole genome for translocations), GROM identifies indels and SVs with matching start and end breakpoints. Matching SV breakpoints must meet the following criteria:

$$|B_S + L_S - B_e| \leq c \times (i_{max} - i_{min}), \tag{3}$$

$$|B_e - L_e - B_s| \leq c \times (i_{max} - i_{min}), \tag{4}$$

where $c = 3/8$, B_s and B_e are the start and end breakpoints, respectively, and L_s and L_e are the average variant length of reads supporting the start or end breakpoints, respectively. Matching translocation breakpoints follow the same concept modified due to the start and end breakpoints occurring on different chromosomes,

$$|M_S - B_e| \leq c \times (i_{max} - i_{min}), \tag{5}$$

$$|M_e - B_s| \leq c \times (i_{max} - i_{min}), \tag{6}$$

where $c = 3/8$, B_s and B_e are the start and end breakpoints, respectively, and M_s and M_e are the average mate read reference locations of reads supporting the start or end breakpoints, respectively.

Mixed libraries/BAM files, e.g., with insert size distributions appreciably different as to affect Equations (3–6) for matching breakpoints, or libraries containing paired-end with mate-pair data, require separate runs of GROM. Also, GROM can analyze exome or RNA sequencing reads with detection limited to SNVs and indels.

GROM will also work for libraries of non-paired reads using (in addition to finding SNVs and SVs within reads) our earlier method for finding copy number variants (CNVs), GROM-RD [9]. GROM-RD also performs well compared with the standard tools such as CNVnator [10]. GROM and GROM-RD have the same foundation of collecting information for each reference base, but GROM-RD detects CNVs based on read depth, where low or high coverage is evidence of a deletion or duplication, respectively. This method is complementary to the core GROM approach described above.

GROM is able to simultaneously perform duplicate filtering; its duplicate filter is conceptually similar to Picard's MarkDuplicates [11] and SAMtools rmdup [12], which have been shown to have similar performance. Duplicate filtering may improve predictive accuracy relative to no filtering [13]. GROM provides an option to include such filtering, if necessary. GROM filters read pairs with identical orientation and external mapping coordinates, retaining the pair with highest mapping quality. Unlike SAMtools, GROM and Picard's MarkDuplicates are able to filter duplicates with reads mapping to different chromosomes and adjust external coordinates based on soft-clipping [13]. For the sake of speed optimization and 1-pass analysis, soft-clipping is not considered for a read's mate.

Results

We compared GROM's performance to 4 commonly used algorithms, GATK HaplotypeCaller (GATK-HC) [14], SAMtools [12], LUMPY [15], and Manta [16] using 2 extensively validated human WGS data sets, 51× NA12878 "platinum" genome [17] and 68× HX1, a recent Chinese genome [18]. GATK-HC, considered a gold standard in SNV/indel detection, has been shown to outperform state-of-the-art algorithms [19], and SAMtools is present in most pipelines. Because GROM integrates multiple lines of evidence, we also specifically compared it with a similar SV tool in the SpeedSeq pipeline (SpeedSeq, RRID:SCR_000469), LUMPY, shown to outperform other algorithms [15], such as DELLY (DELLY, RRID:SCR_004603) [20], Pindel (Pindel, RRID:SCR_000560) [21], and GASVPro (GASVPro, RRID:SCR_005259) [22]. As part of a 10 000 genome sequencing study, presently the largest human WGS variant study, a comparison of 7 SV detection algorithms (BreakDancer [23], DELLY [20], GenomeSTRiP [24], LUMPY [15], Manta [16], MatchClip2 [25], and Pindel [21]), showed that Manta performed the best for SV detection [2]. We evaluated SNV and indel detection with the Illumina Platinum pedigree-validated benchmark sets [17]. GROM exhibited the highest SNV and insertion indel sensitivity and precision and the highest deletion indel sensitivity when compared with GATK-HC and SAMtools for the NA12878 genome (Supplementary Table S1). SVs are notoriously difficult to reliably detect [2]. Thus, we extensively analyzed GROM's performance using 4 benchmark sets for NA12878: Database of Genomic Variants Gold Standard (DGV-GS, deletions and duplications) [26], Mills Gold Standard (Mills-GS; deletions, duplications, and insertions) [27], Genome in a Bottle (GIAB, deletions and insertions) [28]; and Pendleton PacBio (deletions and inversions) [29]. And we utilized 3 deletion and duplication benchmark sets for HX1: DGV-GS, Shi PacBio [18], and Shi IrysChip [18] (see the Methods section for a more complete description of benchmark/validation sets). A summary of the deletion and duplication comparison with LUMPY and Manta indicated superior deletion and duplication detection (Supplementary Table S2), with GROM being the highest in 10 of 14 deletion (Supplementary Table S3) and 7 of 10 duplication (Supplemental Table S4) metrics (sensitivity and precision) across the benchmark data sets. Additionally, GROM was highest in all inversion (Supplemental Table S5) and insertion (Supplemental Table S6) metrics. GROM also detected 545 and 472 translocation events in NA12878 and HX1, respectively. However, these events were not included in the benchmarking due to the lack of validated translocation data sets for either genome.

With dropping sequencing costs and growing data throughput, it is imperative to reduce the computational costs of big data analysis. GROM was 1.7× (NA12878) and 2.1× (HX1) faster than the next fastest algorithm, Manta (Supplementary Table S7). Since typical analyses involve running separate algorithms for SNV/indel and SV detection, we compared a simple 24-thread parallelized GROM version (allocating a thread per 1/24 of the genome) with the fastest and best-performing 2-algorithm workflow (GATK-HC/Manta). Strikingly, GROM ranged from 24× (HX1, no duplicate filtering) to 72× (NA12878 with duplicate filtering) faster than a combination of 22-thread GATK-HC/2-thread Manta (Supplementary Table S8), drastically reducing variant detection and duplicate filtering from 41% to <1% of a typical WGS analysis pipeline (Fig. 4). For 1000 genomes on a 24-thread server, it may literally save years of computation.

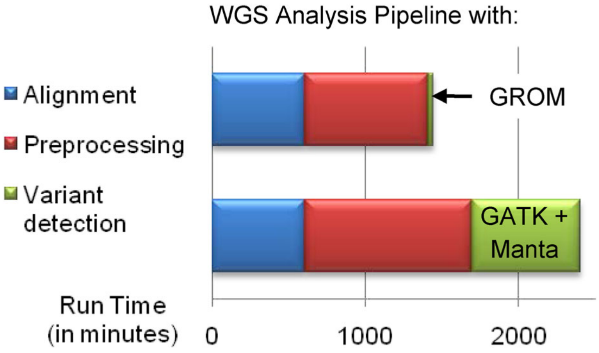

Figure 4: Total WGS pipeline timing on NA12878. GROM reduces WGS analysis time by drastically cutting run time for variant detection (green). It enables further speedup in preprocessing (red) by simultaneously performing an optional step, duplicate filtering. For visibility in the bar chart, GROM's variant detection run time was artificially increased 3-fold.

Comparing the variants predicted by different tools, we identified 33 validated NA12878 SVs detected by GROM (but unreported by LUMPY and Manta) that overlapped genes and ranked them using the number of independent validations (Supplementary Table S9). A variant was considered validated if it occurred in at least 1 of the NA12878 benchmarks corresponding to the SV type (DGV-GS, Mills-GS, GIAB, Pendleton PacBio for deletions; DGV-SV, Mills-GS for duplications; Mills-GS, GIAB for insertions; and Pendleton PacBio for inversions).

Among these variants, we noted 4 deletions with significant health-related impact for NA12878: RHD, GSTM1, IFI16, and UGT2B17 (Fig. 5). GROM predicted a deletion spanning the entire RHD gene, 1 of 2 genes responsible for Rh blood group antigens [30]. Decreased copy numbers or null genotype of GSTM1 have been associated with hepatotoxicity [31] and higher risk of many cancers including lung cancer [32], gastric cancer [33], and bladder cancer [34]. UGT2B17 copy number variation has been associated with changes in bone mineral density and risk of osteoporosis [35]. IFI16 is involved in viral defense [36] and p53-mediated apoptosis [37, 38].

Additionally, GROM provides an option to include duplicate filtering. This leads to minor accuracy gains in a number of cases (see example in Supplementary Table S10) and achieves additional speedup (Supplementary Table S8). Lastly, we have summarized GROM's relative performance in Table 1.

Methods

All timings were performed on an Intel Xeon E5–2690 v. 3 processor, 2.60 GHz, with 24 threads and 128 GB RAM.

Rankings in Table 1 and Supplementary Table S2 were based on average ranking across benchmarks (1-highest to 3-lowest). Ranking for each benchmark was based on sensitivity and precision values in Supplementary Tables S3–S6. For instance, GROM had the highest value for 10, second highest for 2, and lowest for 2 of the 14 deletion sensitivity and precision benchmarks (average benchmark rank, 1.4) Subsequently, the algorithms were ranked after sorting by their average benchmark ranking, resulting in deletion rankings of GROM, 1; LUMPY, 2; and Manta, 3 (as shown in Table 1).

Unlike most SV variant callers, GROM is able to analyze data sets with single or paired reads. However, all SV tests included

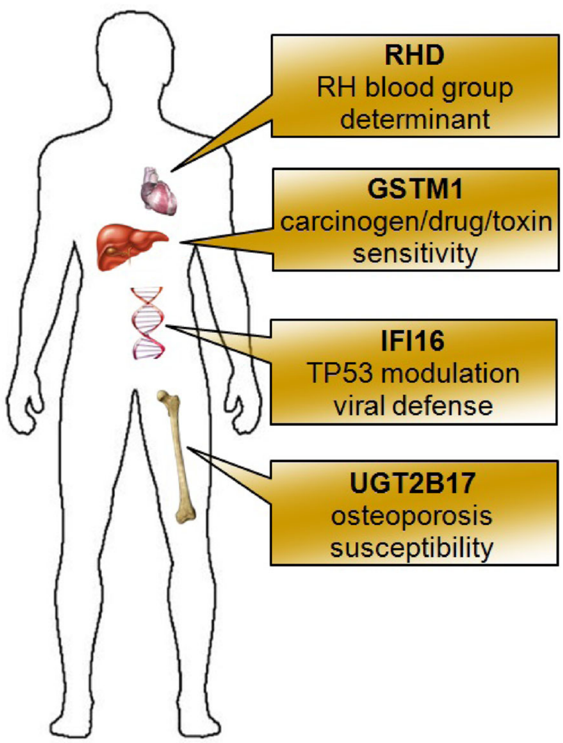
RHD
RH blood group determinant
GSTM1
carcinogen/drug/toxin sensitivity
IFI16
TP53 modulation viral defense
UGT2B17
osteoporosis susceptibility

Figure 5: Example of genes overlapped by validated GROM-specific SVs. In the example are 4 of 33 genes overlapped by validated SVs that were identified by GROM and unreported by LUMPY and Manta. Biological significance listed below gene.

only paired reads since most of the other callers operate on those.

While state-of-the-art detection methods for SNVs and indels have been deemed adequate for the clinical setting, SV detection is notably more difficult [2]. Additionally, synthetic data sets have suffered from oversimplifications and misleading conclusions [2]. Thus, we extensively analyzed GROM's SV detection performance using 4 validation benchmark sets for NA12878:

1. Database of Genomic Variants Gold Standard (deletions and duplications) [26] in Supplementary Tables S2–S4;
2. Mills Gold Standard (deletions, duplications, and insertions) [27] in Supplementary Tables S2–S5;
3. Genome in a Bottle (deletions and insertions) [28] in Supplementary Tables S2, S3, S5; and
4. Pendleton PacBio (deletions and inversions) [29] in Supplementary Tables S2, S3, S6.

Additionally, we utilized 3 deletion and duplication benchmark sets for HX1: DGV-GS (as above), Shi PacBio, and Shi IrysChip [18] in Supplementary Tables S2–S4. For NA12878 DGV-GS benchmarks, all deletions and duplications with the "NA12878" tag were extracted from the DGV-GS. The HX1 DGV-GS benchmarks were created by extracting deletions and duplications with the "Asian" tag. To obtain a benchmark set of common Asian variants, deletions and duplications with fewer than 200 "Asian"-tagged samples were filtered.

To limit potential biases, we selected benchmarks covering a range of technologies, including Illumina, PacBio, and IrysChip, and inclusive of multiple variant detection algorithms (Illumina platinum pedigree-validated, DGV-GS, Mills-GS, and GIAB). Indels were defined as deletions and insertions $<$50 bases, whereas SVs were $\geq$50 bases. To identify true positives, indel benchmarking required variant call breakpoints within 2 bases of the benchmark. Insertion SV calls within 10 bases of the benchmark were considered true. All other SV benchmarking required a 50% (10% for IrysChip due to low resolution) reciprocal overlap of a variant call and the benchmark. Some false positives may potentially be true positives not represented in the benchmark. To limit false positives due to unrepresented calls, for each SV type (excluding insertions where the length is often unknown), we ignored SV calls smaller or larger than a particular benchmark's shortest or longest SV, respectively.

NA12878 and HX1 Illumina platinum fasta files were mapped to human references hg19 and GRCh38, respectively, using BWA mem [39], version 0.7.15, with the -M parameter to mark shorter read splits as secondary. Duplicate filtering comparisons were performed using default parameters for SAMtools [12], version 1.3.1, and Sambamba [40], version 0.6.4. GATK version 3.6.0 HaplotypeCaller (GATK, RRID:SCR_001876) [14], SAMtools (SAMTOOLS, RRID:SCR_002105) [12], LUMPY (LUMPY, RRID:SCR_003253; version 0.2.11) [15], and Manta (version 1.0.1) [16] were run with default parameters.

Conclusion

Our extensive performance analysis indicates that GROM achieves superior variant detection and is significantly faster than current state-of-the-art methods by incorporating comprehensive variant detection (SNV, indel, SV, CNV), duplicate filtering, and multi-threading in 1 algorithm. GROM's superior variant detection makes it valuable for WGS analysis projects of

Table 1: Comparison of GROM and leading algorithms' variant detection accuracy and run time

		GATK-HC	SAMtools	LUMPY	Manta	GROM
SNV		2	3	-	-	**1**
Indel	Deletion	**1**	3	-	-	**1**
	Insertion	2	3	-	-	**1**
SV	Deletion	-	-	2	3	**1**
	Duplication	-	-	2	2	**1**
	Insertion	-	-	-	2	**1**
	Inversion	-	-	3	2	**1**
Run time		4	5	3	2	**1**

Performance based on sensitivity and precision rankings (1 = highest, 3 = lowest) averaged across benchmarks for NA12878 and HX1. Bold text indicates the best-performing algorithm in each category. A dash sign indicates that an algorithm does not detect variant type.

all sizes, and its "lightning"-fast speed is especially critical for keeping pace with increasingly higher sequencing throughput and larger data projects.

Availability of data and materials

NA12878 raw short-read Illumina platinum WGS data, as well as pedigree-validated SNVs and indels, supporting the results in this study are available from the Database of Genotypes and Phenotypes under accession number phs001224.v1.p1 [41]. HX1 raw short-read Illumina WGS data supporting the results in this study are available from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), study PRJNA301527 [42]. DGV-GS-validated SVs supporting the results in this study are available from the Database of Genomic Variants website [43]. Mills-GS-validated SVs supporting the results in this study are available as Supplementary Table S5 in the associated paper [27]. GIAB validation data supporting the results in this study are available from NCBI at separate locations for deletions (v. 3.3.1) [44] and insertions [45]. Pendleton PacBio–validated deletions and inversions supporting the results in this study are available as Supplementary Tables S5 and S6, respectively, in the associated paper [29]. Shi PacBio– and Shi IrysChip–validated SVs supporting the results in this study are available from the corresponding author's website [46]. Human reference genomes hg19 and GRCh38 are available from the Broad Institute [47] and UCSC [48], respectively. Snapshots of the GROM project code are available via the Open Science Framework [49] and the *GigaScience* database, *GigaDB* [50].

Availability and requirements

Project name: GROM
Project home page: https://osf.io/6rtws/
Code DOI: 10.17605/OSF.IO/6RTWS
Operating system: Linux
Programming language: C
Other requirements: see manual in the distribution
License: GNU General Public License v2

Additional files

Additional file 1: Supplementary tables. Benchmark results (Supplementary Tables S1–S6), run time comparisons (Supplementary Tables S7–S8), GROM-specific SVs overlapping genes (Supplementary Table S9), and duplicate read filtering comparison (Supplementary Table S10).

Abbreviations

CNV: copy number variant; DGV-GS: Database of Genomic Variants–Gold Standard; GATK-HC: GATK HaplotypeCaller; GIAB: Genome In A Bottle; GROM: Genome Rearrangement OmniMapper; Mills-GS: Mills–Gold Standard; SNV: single nucleotide variant; SV: structural variant; WGS: whole genome sequencing.

Competing interests

The authors declare that they have no competing financial interests.

Author contributions

A.G. and S.D.S. conceived the project. S.D.S. designed and wrote the algorithm, with contributions from J.K.K. and A.G. S.D.S., J.K.K., and A.G. analyzed results. S.D.S. and A.G. wrote the manuscript with input from all authors. A.G. supervised the project and secured funding from startup and grant funds.

Acknowledgements

The authors thank Kevin Abbey, Hui-Jou Chou, Spyros Karaiskos, and Ian Biluck (all from Rutgers) for excellent technical help throughout the development and testing process. The authors thank Tal Nevo, Benzi Galili, and Erez Valen (all from ScaleMP, Inc.) for discussions, suggestions, and code contributions used in the parallel version of GROM. This work was supported by the National Science Foundation (award DBI-1458202 to A.G.) and in part by the New Jersey Health Foundation grant to A.G.

References

1. Genomes Project Consortium, Abecasis GR, Altshuler D et al. A map of human genome variation from population-scale sequencing. Nature 2010;**467**(7319):1061–73.
2. Telenti A, Pierce LCT, Biggs WH et al. Deep sequencing of 10,000 human genomes. Proc Natl Acad Sci U S A 2016;**113**(42):11901–6.
3. Genomics England. https://www.genomicsengland.co.uk/. Accessed 30 September 2017.
4. National Institutes of Health. All of Us Research Program. 2017. https://allofus.nih.gov/. Accessed 30 September 2017.
5. U.S. Department of Veterans Affairs. Million Veteran Program. 2017. https://www.research.va.gov/mvp/. Accessed 30 September 2017.
6. David Cyranoski, Nature. Nature News Feature: China's bid to be a SNA Superpower. 2016. http://www.nature.com/news/china-s-bid-to-be-a-dna-superpower-1.20121. Accessed 30 September 2017.
7. Stephens ZD, Lee SY, Faghri F et al. Big data: astronomical or genomical? PLoS Biol 2015;**13**(7):e1002195.
8. Chiang C, Layer RM, Faust GG et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods 2015;**12**(10):966–8.
9. Smith SD, Kawash JK, Grigoriev A. GROM-RD: resolving genomic biases to improve read depth detection of copy number variants. Peer J 2015;**3**:e836.
10. Abyzov A, Urban AE, Snyder M et al. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 2011;**21**(6):974–84.
11. Broad Institute hosted by GiHub. Broad Institute Picard Releases. 2017. https://github.com/broadinstitute/picard/releases. Accessed 30 September 2017.
12. Li H, Handsaker B, Wysoker A et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;**25**(16):2078–9.
13. Ebbert MTW, Wadsworth ME, Staley LA et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics 2016;**17**(S7):239.
14. Depristo MA, Banks E, Poplin R et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;**43**(5):491–8.

15. Layer RM, Chiang C, Quinlan AR et al. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 2014;**15**(6):R84.
16. Chen X, Schulz-Trieglaff O, Shaw R et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 2016;**32**(8):1220–2.
17. Eberle MA, Fritzilas E, Krusche P et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res 2017;**27**(1):157–64.
18. Shi L, Guo Y, Dong C et al. Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun 2016;**7**:12065.
19. Yi M, Zhao Y, Jia L et al. Performance comparison of SNP detection tools with illumina exome sequencing data-an assessment using both family pedigree information and sample-matched SNP array data. Nucleic Acids Res 2014;**42**(12):e101.
20. Rausch T, Zichner T, Schlattl A et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 2012;**28**(18):i333–9.
21. Ye K, Schulz MH, Long Q et al. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 2009;**25**(21):2865–71.
22. Sindi SS, Onal S, Peng LC et al. An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biol 2012;**13**(3):R22.
23. Chen K, Wallis JW, Mclellan MD et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 2009;**6**(9):677–81.
24. Handsaker RE, Van Doren V, Berman JR et al. Large multiallelic copy number variations in humans. Nat Genet 2015;**47**(3):296–303.
25. Wu Y, Tian L, Pirastu M et al. MATCHCLIP: locate precise breakpoints for copy number variation using CIGAR string by matching soft clipped reads. Front Genet 2013;**4**:157.
26. Macdonald JR, Ziman R, Yuen RKC et al. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 2014;**42**(D1):D986–92.
27. Mills RE, Walter K, Stewart C et al. Mapping copy number variation by population-scale genome sequencing. Nature 2011;**470**(7332):59–65.
28. Zook JM, Chapman B, Wang J et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 2014;**32**(3):246–51.
29. Pendleton M, Sebra R, Pang AWC et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods 2015;**12**(8):780–6.
30. Wagner FF, Flegel WA. RHD gene deletion occurred in the Rhesus box. Blood 2000;**95**(12):3662–8.
31. Singh HO, Lata S, Angadi M et al. Impact of GSTM1, GSTT1 and GSTP1 gene polymorphism and risk of ARV-associated hepatotoxicity in HIV-infected individuals and its modulation. Pharmacogenomics J 2017;**17**(1):53–60.
32. Yang H, Yang S, Liu J et al. The association of GSTM1 deletion polymorphism with lung cancer risk in Chinese population: evidence from an updated meta-analysis. Sci Rep 2015;**5** 1:9392.
33. Lao X, Peng Q, Lu Y et al. Glutathione S-transferase gene GSTM1, gene-gene interaction, and gastric cancer susceptibility: evidence from an updated meta-analysis. Cancer Cell Int 2014;**14**(1):127.
34. Norskov MS, Frikke-Schmidt R, Bojesen SE et al. Copy number variation in glutathione-S-transferase T1 and M1 predicts incidence and 5-year survival from prostate and bladder cancer, and incidence of corpus uteri cancer in the general population. Pharmacogenomics J 2011;**11**(4): 292–9.
35. Yang T, Chen X, Guo Y et al. Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. Am J Hum Genet 2008;**83**(6):663–74.
36. Orzalli MH, Conwell SE, Berrios C et al. Nuclear interferon-inducible protein 16 promotes silencing of herpesviral and transfected DNA. Proc Natl Acad Sci U S A 2013;**110**(47):E4492–501.
37. Aglipay JA, Lee SW, Okada S et al. A member of the Pyrin family, IFI16, is a novel BRCA1-associated protein involved in the p53-mediated apoptosis pathway. Oncogene 2003;**22**(55):8931–8.
38. Johnstone RW, Wei W, Greenway A et al. Functional interaction between p53 and the interferon-inducible nucleoprotein IFI 16. Oncogene 2000;**19**(52):6033–42.
39. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;**25**(14):1754–60.
40. Tarasov A, Vilella AJ, Cuppen E et al. Sambamba: fast processing of NGS alignment formats. Bioinformatics 2015;**31**(12):2032–4.
41. National Center for Biotechnology Information. Database of Genotypes and Phenotypes. https://www.ncbi.nlm.nih.gov/gap. Accessed 24 April 2017.
42. National Center for Biotechnology Information. Sequence Read Archive. https://www.ncbi.nlm.nih.gov/sra/. Accessed 24 April 2017.
43. Database of Genomic Variants. http://dgv.tcag.ca/dgv/app/home. Accessed 30 September 2017.
44. Zook JM, Catoe D, Mcdaniel J et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 2016;**3**.
45. Parikh H, Mohiyuddin M, Lam HYK et al. svclassify: a method to establish benchmark structural variant calls. BMC Genomics 2016;**17**(1):160025.
46. Wang Genomics Lab. Shi PacBio and Shi IrysChip validated SVs. 2016. http://hx1.wglab.org/data/cnv_sv/. Accessed 30 September 2017.
47. Broad Institute. GATK Resource Bundle, hg19 Human Reference Genome. 2017. https://software.broadinstitute.org/gatk/download/bundle. Accessed 30 September 2017.
48. UCSC. Genome Browser, GRCh38 Human Reference Genome. 2017. https://genome.ucsc.edu. Accessed 30 September 2017.
49. Smith SD, Grigoriev A. hosted by Open Science Framework. GROM. 2017. http://doi.org/10.17605/OSF.IO/6RTWS. Accessed 30 September 2017.
50. Smith SD, Kawash JK, Grigoriev A. Supporting data for “Lightning-fast genome variant detection with GROM.” GigaScience Database 2017. http://dx.doi.org/10.5524/100352.