Published in final edited form as:
*Nat Biotechnol*. ; 30(3): 226–229. doi:10.1038/nbt.2134.

# Detecting and annotating genetic variations using the HugeSeq pipeline

**Hugo Y K Lam**[1,5], **Cuiping Pan**[1], **Michael J Clark**[1], **Phil Lacroute**[1], **Rui Chen**[1], **Rajini Haraksingh**[1], **Maeve O'Huallachain**[1], **Mark B Gerstein**[2,3,4], **Jeffrey M Kidd**[1], **Carlos D Bustamante**[1], and **Michael Snyder**[1]

[1]Department of Genetics, Stanford University, Stanford, California, USA

[2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA

[3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, USA

[4]Department of Computer Science, Yale University, New Haven, Connecticut, USA

## To the Editor

Deciphering genome sequences is important for the mapping of genetic diseases and prediction of their risks. Advances in high-throughput DNA sequencing technologies using short read lengths have enabled rapid sequencing of entire human genomes and unlocked the potential for comprehensive identification of their underlying genetic variations. Various computational algorithms for identifying and characterizing variants have been developed; however, most of these computational methods are neither integrated nor interoperable, making it difficult for biologists to extract all the genetic information from billions of sequences generated by these sequencing technologies. Here, we present HugeSeq, an integrated computational pipeline to fully automate the process of variant detection from alignment of these genomic sequences to detection and annotation of all types of genetic variations (single nucleotide polymorphisms (SNPs), short insertions or deletions (indels) and larger structural variations (SVs)).

Compared with other popular platforms for genome data analysis that typically analyze SNPs or a limited set of variants (Supplementary Table 1), HugeSeq covers a more complete spectrum of variant types. The complete variant detection and characterization workflow of the HugeSeq pipeline (Fig. 1) is a modular framework comprising three phases: first, a mapping phase that prepares and aligns reads; second, a sorting phase that combines and sorts alignments for parallel variant detection; and third, a reduction phase that detects and annotates different variants (SNPs, indels and structural variations). It is based on a

[5]Present address: Personalis, Inc., Palo Alto, California, USA. mpsnyder@stanford.edu

**COMPETING FINANCIAL INTERESTS**

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology.

Note: Supplementary information is available on the Nature Biotechnology website.

MapReduce[1] approach and runs in a parallel computational environment, making it highly efficient and scalable.

HugeSeq uses sequence reads (both single end and paired end) in a FASTA or FASTQ format (optionally compressed in a GZIP format) as input for alignment. Because alignment of a single read is independent of others, HugeSeq divides the reads into smaller subsets so they can be aligned in parallel. It then distributes the reads in the computer cluster and carries out a gapped alignment against the reference genome using the Burrows-Wheeler aligner[2]. The generated sequence alignment map (SAM)[3] is then converted into its binary format, BAM, using SAMtools[3] to ensure efficient storing and access of alignment information. After alignment, HugeSeq collects all the mapped reads and sorts them according to their aligned chromosomal positions with the Picard tool. The sorted reads for each chromosome are assigned to their corresponding chromosomal BAM and indexed for rapid and random access using SAMtools. Because most variant detections are intrachromosomal, the detection process can be carried out on each chromosomal BAM simultaneously. Interchromosomal translocation detection can also be enabled and run in a nonparallel mode, although it slows down the process considerably.

To enhance the quality of the alignments for more accurate variant detection, HugeSeq carries out several processing (‘cleanup’) procedures before variant calling. First, to minimize experimental artifacts, it removes potential PCR duplicates using the Picard tool. Second, it carries out a local realignment around indels and SNP clusters using the Genome Analysis ToolKit (GATK) realigner[4]. Last, based on the realignment, it recalibrates the base quality of the alignments using the GATK recalibrater[4] so that the quality scores represent the empirical probability of mismatching to the reference genome. With the processed read alignments or any user-specified BAMs, HugeSeq detects variants of different kinds in a parallel fashion.

For SNP and small indel detection, HugeSeq uses two different well-established SNP and indel calling algorithms, the GATK UnifiedGenotyper[4] and SAMtools[3]. When calling indels using GATK, it uses the Dindel[5] model for greater sensitivity. The resulting SNPs and indels are then passed through the GATK variant filtering tool with default parameters similar to those used in the 1000 Genomes Project[6]. Structural variations and copy number variants (CNVs) are often difficult to detect, largely owing to their heterogeneous nature. A variety of different methods can be used to find them but each has distinct biases. To identify as many structural variations as possible, HugeSeq uses four major approaches: first, paired-end mapping using BreakDancer[7]; second, split-read analysis using Pindel[8]; third, read-depth analysis using CNVnator[9] and fourth, junction mapping using BreakSeq[10] (a version we modified to support BAM as input for unmapped reads). Because these structural variation and CNV callers generate variant calls in different formats, HugeSeq standardizes their outputs by converting them into the standard general feature format.

The resulting SNP and indel call sets, which are in a standard variant call format (VCF), are combined and merged using VCFtools[11]. HugeSeq also uses VCFtools to concatenate variants from different BAMs for each algorithm and to merge calls from different algorithms into a single VCF. SNPs called by both GATK and SAMtools are of particularly high confidence. For the structural variation and CNV call sets, BEDtools[12] is used to intersect and merge the calls. Calls detected by two or more algorithms (50% reciprocal overlap) are regarded as high confidence[13]. Finally, HugeSeq carries out a functional annotation on the variants using Annovar[14]. Because both the input and output formats of Annovar are nonstandard variant formats, HugeSeq converts variants into the Annovar input format and converts the Annovar output into a tab-delimited format that can be easily merged and interpreted with the original call sets. The annotation includes gene intersection, exonic variations, repeat elements and mutation information (for example, SIFT score[15]). It can be expanded easily to include different databases, such as PolyPhen[16].

We initially applied HugeSeq to a single human genome sequenced (~48×) with Illumina HiSeq. Overall, HugeSeq reported >8 million raw variant calls, or ~4.5 million after merging (Table 1). For SNPs, GATK UnifiedGenotyper reported 3,570,658 calls (after filtering), and SAMtools reported 3,632,090 calls. There were 3,399,561 concordant SNP calls based on positions between GATK and SAMtools. For small indels, GATK (with the Dindel model) reported 523,445 calls and SAMtools reported 553,360 calls. There were 422,305 concordant indel calls based on matching the positions. For structural variation or CNV ( ≥50 bp), there were 11,043 paired-end calls, 11,911 read-depth calls, 1,741 split-read calls and 1,003 junction calls. There were 21,381 structural variation union calls with 1,639 calls reported by two or more algorithms (50% reciprocal overlap). Because of the ascertainment bias toward deletions, the majority (>90%) of the structural variation calls were deletions; however, duplications, insertions and inversions were also detected (Supplementary Table 2). We observed enrichments of ~300-bp and ~6,000-bp structural variation deletions, possibly owing to *Alu* and long interdispersed element 1 elements. The majority of indel calls were 1 bp, whereas there was enrichment of 4-bp and 8-bp indel calls (Supplementary Fig. 1).

We carried out benchmarking on HugeSeq to assess its performance. We measured its run time and memory usage at various sequencing coverage ranging from 6× to 96× (Supplementary Figs. 2 and 3). For run time, we found that HugeSeq spent relatively less clock time on variant detection, as it benefited from parallel processing (Supplementary Fig. 2a,b). We also found that its run time (~25 hours) was ~10 times faster than the time of processing individual steps in a nonparallel mode on a single computer (~250 hours) at 30× coverage, whereas the run time increased almost at a linear rate at different coverage (Supplementary Fig. 2c). For memory usage, most processes took at most 6 GB of physical memory. The duplicate removal, recalibration and GATK UnifiedGenotyper were bound by the maximum heap size of 6 GB preallocated to the Java Virtual Machine and the realignment and sorting were allocated a maximum of 12 GB heap size for a more efficient caching of high-coverage data (Supplementary Fig. 2d). When focusing on variant detection, we observed similar performance as in the overall process (Supplementary Fig. 3). We found that BreakDancer (read-pair mapping) and BreakSeq (junction mapping) were fastest and used the least memory, whereas GATK UnifiedGenotyper was slowest and used the most memory.

To assess the sensitivity of variant detection in HugeSeq, we detected SNPs and structural variations or CNVs using the Illumina (San Diego, CA) Human Omni1Quad genotyping array with ~1 million markers. For SNPs, we assessed the true positives by taking all 260,112 heterozygous calls from the array. We intersected these calls with the 3,399,561 SNP calls reported by both GATK and SAMtools (high confidence) and all 3,803,187 merged SNP calls (Fig. 2a). There were 254,700 and 258,654 concordant calls, corresponding to a sensitivity of 97.9% and 99.4% for the high-confidence and total sets, respectively. For the individual call sets, GATK had 257,243 and SAMtools had 256,111 concordant calls, corresponding to a sensitivity of 98.9% and 98.5%, respectively. We also observed that the GATK- and SAMtools-specific calls had a relatively lower transition-to-transversion rate (<1.6) compared with the high-confidence calls (~2.1). For structural variation or CNVs, we generated a list of true positives by taking the 482 deletion calls reported by Illumina GenomeStudio and CNVision[17] based on the Omni1Quad array. We intersected these calls with the 1,594 deletion calls reported by two or more structural variation/CNV algorithms (high confidence) and all 19,809 merged deletion calls. There were 383 and 471 concordant calls ( ≥1 bp overlapping), corresponding to a sensitivity of 79.5% and of 97.7% for the high-confidence and total sets, respectively. By requiring an overlap of 50% on the array calls, we found that there were 358 and 444 concordant calls, giving a sensitivity of 74.3% and 92.1% for the high-confidence and total sets, respectively.

We extended our sensitivity test for SNP or structural variation detection to various sequencing coverage as in the benchmarking. The test was done using the total call sets. We found that SNP calls reached saturation at a sequencing coverage of ~40× with a sensitivity of ~98.5%, whereas structural variation calls reached saturation at ~30× with a sensitivity of ~97% (Fig. 2b). We also intersected the 19,809 deletion calls detected in HugeSeq with 22,025 deletion calls reported by the 1000 Genomes Project in the pilot phase. With 50% reciprocal overlap, we found that structural variations detected by three or more algorithms had up to 93.8% (98.2% by 1 bp) concordance with the 1000 Genomes calls, whereas those detected by two or more algorithms had 88.7% (92.3% by 1 bp) and those detected by any algorithm had only 17.7% (27.6% by 1 bp; Table 2). Thus, we think that the majority of structural variations detected by two or more algorithms are probably correct.

Given the interest in genome sequencing, the availability of an integrated platform to analyze genomes is expected to be useful. The implementation of HugeSeq using widely accessible code with output in a common format could facilitate genome analysis and clinical interpretation[18]. In addition to whole-genome sequencing, HugeSeq can be used for other types of sequencing such as exome-seq (currently not applicable for CNV detection). It can also be used for Illumina's long mate-pair library. We envision that comprehensive variant detection pipelines, such as the one we present here, will facilitate the analysis of biological information extracted from the large amounts of sequencing data currently being generated from thousands of genomes.

HugeSeq is open source and available for download at http://hugeseq.snyderlab.org/. A description of the methods used in our assessment of HugeSeq is in Supplementary methods.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We acknowledge support from the US National Institutes of Health. We also thank K. Ye, K. Chen and A. Abyzov for helpful discussions.

## References

1. Dean, J.; Ghemawat, S. MapReduce: simplified data processing on large clusters. OSDI'04 Proceedings of the 6th Symposium on Operating Systems Design and Implementation; San Francisco. 2004.
2. Li H, Durbin R. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
3. Li H, et al. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]
4. McKenna A, et al. Genome Res. 2010; 20:1297–1303. [PubMed: 20644199]
5. Albers CA, et al. Genome Res. 2011; 21:961–973. [PubMed: 20980555]
6. 1000 Genomes Project Consortium. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
7. Chen K, et al. Nat Methods. 2009; 6:677–681. [PubMed: 19668202]
8. Ye K, et al. Bioinformatics. 2009; 25:2865–2871. [PubMed: 19561018]
9. Abyzov A, et al. Genome Res. 2011; 21:974–984. [PubMed: 21324876]
10. Lam HYK, et al. Nat Biotechnol. 2010; 28:47–55. [PubMed: 20037582]
11. Danecek P, et al. Bioinformatics. 2011; 27:2156–2158. [PubMed: 21653522]
12. Quinlan AR, Hall IM. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]
13. Mills RE, et al. Nature. 2011; 470:59–65. [PubMed: 21293372]
14. Wang K, Li M, Hakonarson H. Nucleic Acids Res. 2010; 38:e164. [PubMed: 20601685]
15. Ng PC, Henikoff S. Annu Rev Genomics Hum Genet. 2006; 7:61–80. [PubMed: 16824020]
16. Ramensky V, Bork P, Sunyaev S. Nucleic Acids Res. 2002; 30:3894–3900. [PubMed: 12202775]
17. Sanders SJ, et al. Neuron. 2011; 70:863–885. [PubMed: 21658581]
18. Ashley EA, et al. Lancet. 2010; 375:1525–1535. [PubMed: 20435227]

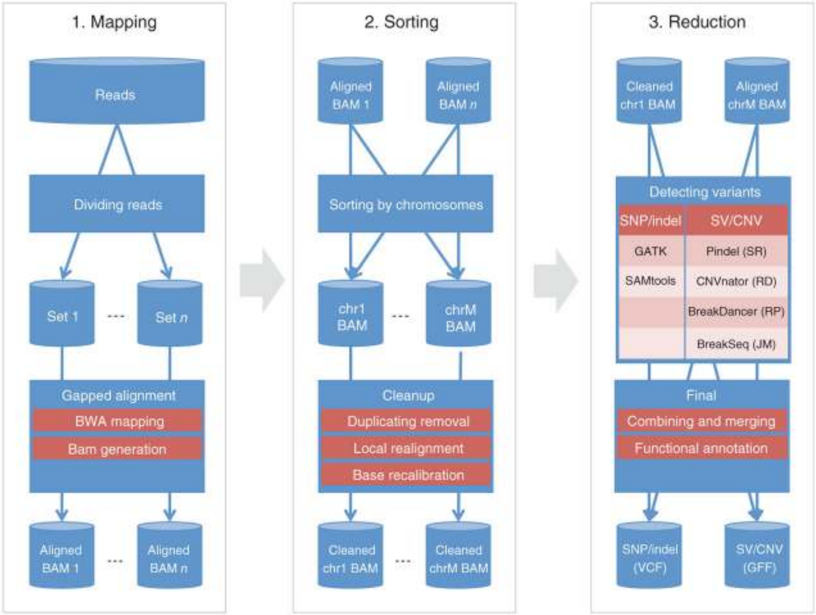
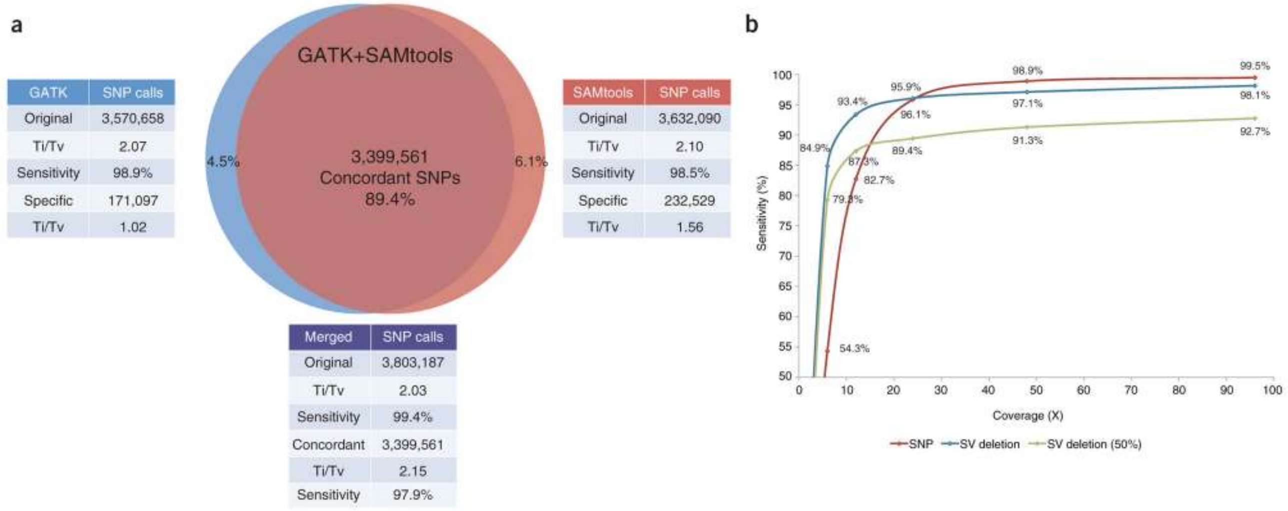


**Figure 1.**

A MapReduce approach for detecting genetic variants from high-throughput genome sequencing. Phase 1 is the mapping phase including sequence alignment. Phase 2 is the sorting phase including sorting alignments by mapped chromosomes. Phase 3 is the reduction phase including variant detection. chr1, chromosome one; chrM, chromosome M; SV, structural variation; SR, split-read analysis; RD, read-depth analysis; RP, read-pair mapping; JM, junction mapping; VCF, variant call format; GFF, general feature format.



| GATK | SNP calls |
|---|---|
| Original | 3,570,658 |
| Ti/Tv | 2.07 |
| Sensitivity | 98.9% |
| Specific | 171,097 |
| Ti/Tv | 1.02 |

| SAMtools | SNP calls |
|---|---|
| Original | 3,632,090 |
| Ti/Tv | 2.10 |
| Sensitivity | 98.5% |
| Specific | 232,529 |
| Ti/Tv | 1.56 |

| Merged | SNP calls |
|---|---|
| Original | 3,803,187 |
| Ti/Tv | 2.03 |
| Sensitivity | 99.4% |
| Concordant | 3,399,561 |
| Ti/Tv | 2.15 |
| Sensitivity | 97.9% |

**Figure 2.**

Accuracy and sensitivity of variant detection. (**a**) Concordant and specific calls in SNP detection by GATK and SAMtools. Original calls, total number of calls from methods; specific calls, total number of method-specific calls. Validated calls, number of calls validated by the array. Ti/Tv, transition-to-transversion rate. (**b**) Overall sensitivity of SNP and structural variation or CNV detection over different sequencing coverage. X is the average number of reads representing a given nucleotide in a haploid human genome. 50% indicates that detected structural variation calls overlap ≥50% the array calls.

## Table 1

Summary of detected variant calls

| Variant | Tool (algorithm) | Original | Merged (union) | Concordant (intersection) |
|---|---|---:|---:|---:|
| SNP | GATK | 3,570,658 | 3,803,187 | 3,399,561 |
| | SAMtools | 3,632,090 | | |
| Indel | GATK<br>SAMtools | 523,445<br>553,360 | 654,500 | 422,305 |
| Structural variation or CNV | BreakDancer (paired-end mapping) | 11,043 | 21,381 | 1,639 |
| | CNVnator (read-depth analysis) | 11,911 | | |
| | Pindel (split-read analysis) | 1,741 | | |
| | BreakSeq (junction mapping) | 1,003 | | |
| Total | | 8,305,251 | 4,479,068 | 3,823,505 |

## Table 2

Concordant deletion calls between structural variations detected and from 1000 Genomes Project

| Number of algorithms | Deletion calls | 1-bp overlapping | 50% reciprocal overlapping |
|---|---|---|---|
| ≥3 algorithms | 451 | 443 (98.2%) | 423 (93.8%) |
| ≥2 algorithms | 1,594 | 1,472 (92.3%) | 1,414 (88.7%) |
| Any algorithm | 19,809 | 5,468 (27.6%) | 3,516 (17.7%) |