

## A Bioinformatics Pipeline for Whole Exome Sequencing: Overview of the Processing and Steps from Raw Data to Downstream Analysis

Narendra Meena<sup>1</sup>, Praveen Mathur<sup>2</sup>, Krishna Mohan Medicherla<sup>1</sup> and Prashanth Suravajhala<sup>1</sup>. \*

<sup>1</sup>Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur 302001, RJ, India; <sup>2</sup>Department of Pediatric Surgery, SMS Medical College, Jaipur 302005, RJ, India

\*For correspondence: [prash@bisr.res.in](mailto:prash@bisr.res.in)

**[Abstract]** Recent advances in Next Generation Sequencing (NGS) technologies have given an impetus to find causality for rare genetic disorders. Since 2005 and aftermath of the human genome project, efforts have been made to understand the rare variants of genetic disorders. Benchmarking the bioinformatics pipeline for whole exome sequencing (WES) has always been a challenge. In this protocol, we discuss detailed steps from quality check to analysis of the variants using a WES pipeline comparing them with repositied public NGS data and survey different techniques, algorithms and software tools used during each step. We observed that variant calling performed on exome and whole genome datasets have different metrics generated when compared to variant callers, GATK and VarScan with different parameters. Furthermore, we found that VarScan with strict parameters could recover 80-85% of high quality GATK SNPs with decreased sensitivity from NGS data. We believe our protocol in the form of pipeline can be used by researchers interested in performing WES analysis for genetic diseases and any clinical phenotypes.

**Keywords:** Whole exome sequencing, Next generation sequencing, Bioinformatics pipeline, Variants, Genetics, Clinical phenotypes

**[Background]** Next Generation Sequencing (NGS) technologies have paved the way for rapid sequencing efforts to analyze a wide number of samples. From the whole genome to transcriptome to exome, it has changed the way we look at nonspecific germline variants, somatic mutations, structural variant besides identifying associations between a variant and human genetic disease (Singleton *et al.*, 2011). This can help understand the complex genetic disorders to get better diagnosis and assess disease risk. The analysis of exome sequencing data to find variants, however still poses multiple challenges. For example, there are several commercial and open source pipelines but configuring (Pabinger *et al.*, 2014; Guo *et al.*, 2015) them in terms of benchmarking and optimizing them is a time-consuming process. Among the steps, *viz.* quality check, alignment, recalibration, variant calling, variant annotation, one needs to reach consensus on the set of tools following which one's output should be fed as other tool's input (Stajich *et al.*, 2002; Gentleman *et al.*, 2004; Chang and Wang, 2012). While integrating, it would be appropriate to check and use the tools before reproducing and maintaining highly heterogeneous pipelines (Hwang *et al.*, 2015). In this protocol, we discuss the steps for whole exome sequence (WES) analyses and its pipeline to identify variants from exome sequence data. Our pipeline

includes open source tools that include a number of tools from quality check to variant calling (see Software section).

### **Equipment**

1. Computer  
64GB RAM with 8 core CPUs in an Ubuntu operating system (14.04 LTS machine)

### **Software**

All the software can be downloaded/used from following locations:

1. FastQC <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
2. Bowtie2 <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>
3. Samtools <http://samtools.sourceforge.net/>
4. VarScan <http://varscan.sourceforge.net/>
5. Bcftools <https://github.com/samtools/bcftools>
6. Vcftools <https://vcftools.github.io/index.html>
7. PANTHER <http://pantherdb.org/>
8. dbSNP <https://www.ncbi.nlm.nih.gov/projects/SNP/>
9. 1000 genomes dataset <http://www.internationalgenome.org/>
10. GeneMania <http://genemania.org/>

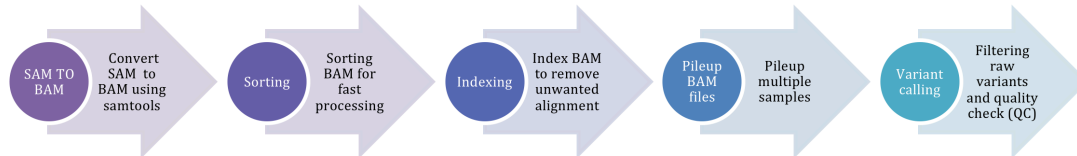
### **Procedure**

The raw file (fastq) is subjected to different steps such as quality check, indexing, alignment, sorting, duplication removal, variant calling, variant annotation and finally downstream bioinformatics annotation (Pabinger *et al.*, 2014) (Figure 1). The pipeline is integration of tools, *viz.* bowtie2 (Langmead and Salzberg, 2012), samtools (Li *et al.*, 2009), FastQC (Andrews, 2010), VarScan (Koboldt *et al.*, 2012) and bcftools (Li *et al.*, 2009), apart from necessary files containing the human genome (Venter *et al.*, 2001), alignment indices (Trapnell and Salzberg, 2009), known variant databases (Sherry *et al.*, 2001; Landrum *et al.*, 2014; Auton *et al.*, 2015). A workflow of given pipeline is shown in Figure 1. In view of the fact that the benchmark metrics for pipelines is an essential step, we have ensured that our pipeline is benchmarked on a sample fastq file taken from a human genome project. As the pipeline runs on Linux, all commands are case sensitive wherever used. Whereas this pipeline was run on a 64 GB RAM with 8 core CPUs in an Ubuntu operating system (14.04 LTS machine), this can also be run on a minimum 16 GB RAM machine based on the size of raw fastq file. A shell script (with an extension sh) was created with all the commands as detailed below.

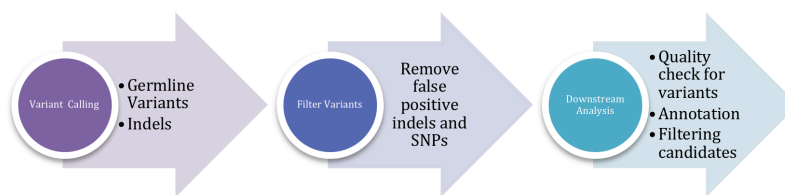
Step 1: NGS Data pre-processing



Step 2: Variant Discovery



Step 3: Variant Prioritization



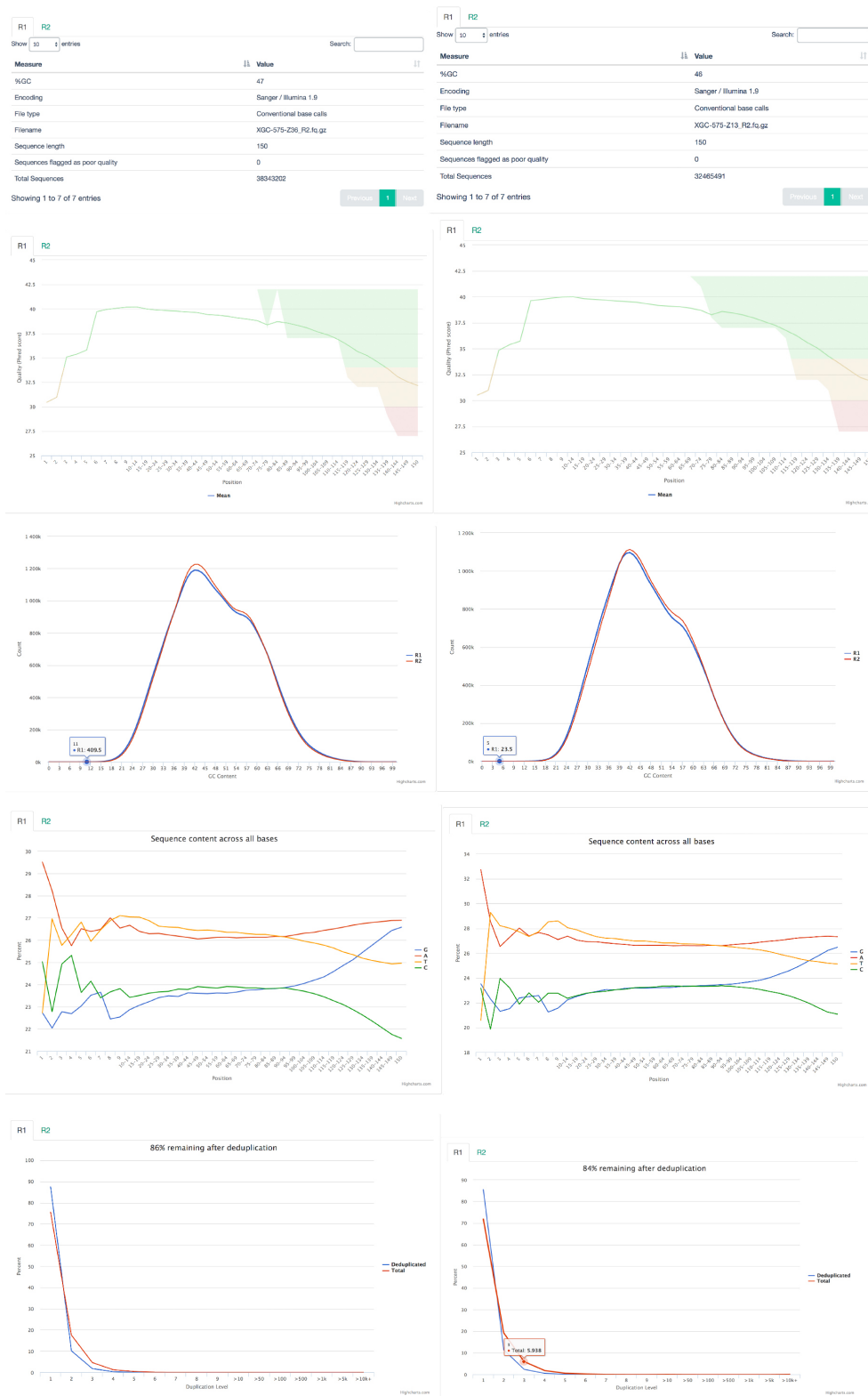
**Figure 1. The pipeline involving three important phases, viz. preprocessing, variant discovery and prioritization of variants**

A. Preprocessing the raw data

**Quality check:** NGS data analysis depends on the raw data control as it provides a quick insight into the quality of the sequences. This will potentially reduce the amount of further downstream analyses with early identification of questionable samples. The ideal base quality scores for Phred (Cock *et al.*, 2010) have paved the way for the best quality scores for GC content (ca. 50% threshold) and the nucleotide distribution across all reads. In our pipeline, we used FastQC (with default Phred = 20 value) as it plots the read depth and quality score besides a host of other statistical inferences.

1. `./fastqc ~/samples/sample1.fastq`

FastQC generates an HTML formatted report with box plots and graph plots for mean quality scores for sequences, read length and depth along with the intended coverage (see Figure 2).



**Figure 2. A pictorial representation containing the box plots and figures of FASTQC run containing information on statistics, quality, read coverage, depth, yield, based per read call, etc.**

**Indexing human genome using bowtie2:** Bowtie2-build is used to index reference genome which works at high speed and memory efficient way.

2. `./bowtie2-build -u 10 indexes/references/reference.fq reference`

When the command is expedited, the current directory will contain six new files that all start with reference and end with .1.bt2, .2.bt2, .3.bt2, .4.bt2, .rev.1.bt2, and .rev.2.bt2. While the first four files are the forward strands, the rev files indicate the reverse indexed sequences.

**Alignment and post processing of alignment:** Bowtie2 is used for short read alignment. What makes bowtie2 interesting is the use of very little RAM with accuracy and modest performance in indexing the alignment (Langmead and Salzberg, 2012). The mismatch or any sequencing errors or small genetic variation between samples and reference genome could be checked using the following commands:

3. `./bowtie2 -x reference_filename -1 path/filename1 -2 path/filename2 > filename.sam`

*Note: The -2 option may be omitted for single-end sequences.*

Bowtie2 aligns a set of unpaired reads (in fastq or .fq format) to the reference genome using the Ferragina and Manzini (FM)-index (Langmead and Salzberg, 2012). The alignment results output in SAM format (Li *et al.*, 2009) and a short alignment summary is written to the console.

**Samtools** is a collection of tools to manipulate the resulting alignment in SAM/BAM format. SAM stands for sequence alignment/map format and its corresponding format is binary mapped format (BAM). SAM could be converted into different alignment format, sort, merge, alignment, remove duplicates, call SNPs and short indel variants; whereas the BAM and following indices (.bai) are used to view the index of the binary aligned sequences. The basic necessity of having the binary mapped files is to save the memory.

4. `~/samtools view -bS sample1.sam > sample1.bam`

**Sorting BAM:** A sorted BAM file is used for streamlined data processing to avoid extra alignment when loaded into memory. Whereas it can be easily indexed for faster retrieval of alignment, overlapping is needed to retrieve alignment from a region.

5. `~/samtools sort sample1.bam sample1.sorted`

*Samtools sort* is used to convert the BAM file to a sorted BAM file and *samtools index* to index BAM file.

6. `~/samtools index sample1.sorted.bam`

**Pileup all samples:** *Samtools mpileup* step is used to analyze multiple samples across all samples thus giving coverage to all mappable reads.

7. `~/samtools mpileup -E -uf reference.fa sample1.bam > sample1.mpileup`

## B. Variant calling

To call variants from NGS data, VarScan among other tools provide heuristic statistical approaches, which give the desired threshold for reading depth, base quality, variant allele frequency and statistical confidence over other bayesian methods. VarScan uses *SAMtools mpileup* data as input and there are a number of options included for variant calling. For each position, the variants, which

don't meet the user input criteria of coverage, number of reads, variant alleles frequency and Fisher's exact test,  $P$ -value are filtered out. This step is a prerequisite to identify those candidate mutations underlying the phenotype/disease.

**Germline variants:** For germline variants, mutations that an individual inherits from their parents, or SNV calling VarScan mpileup2snp protocol is used.

1. `java -jar VarScan.jar mpileup2snp sample.mpileup > sample.varScan.snp`

**Indel calling:** Detecting of insertion and deletion (Indels) is the second most abundant source of finding genetic mutations in human population in a reliable manner. VarScan mpileup2indels protocol is used to call indels. The sensitivity and range for calling indels are determined by the respective alignments.

2. `java -jar VarScan.jar mpileup2indel sample.mpileup > sample.varScan.indel`

**Variant filter:** To get rid of false variants call and remove overlapping between SNP and indels, a filtering option is applied on the resultant variant calling files which provide SNV and indels with higher confidence. An option to generate readcount report can also be used with VarScan (see Table 1).

**Table 1. Variant calling pipelines and their respective arguments**

Pipeline	Arguments
VarScan-Default	Default
VarScan-Pvalue	-- $P$ -value 0.05
VarScan Strict	--min-coverage 10 --min-avg-qual 20 or --min-reads2 4 --min-var-freq 0.3
GATK WPP	"No Preprocessing" pipeline
GATK-PP	"Realign Only", "Recalibrate Only", and "Full Pipeline"

3. `java -jar VarScan.jar filter sample.varScan.snp --indel-file sample.varScan.indel --output-file sample.varScan.snp.filter`
4. `java -jar VarScan.jar filter sample.varScan.indel --output-file sample.varScan.indel.filter`
5. `java -jar VarScan.jar readcounts sample.mpileup.sam > sample.mpileup.readcounts`

**Contamination check:** Once the BAM files and IDs are generated, we could end up checking whether or not the BAMids are error prone or contaminated across the samples using VerifyBAMID (Jun *et al.*, 2012).

### C. Downstream processing the files

VCF, an acronym for variant call format, is a popular format to store variants calling data as it stores both SNPs and indel information succinctly. While BCF is a binary version of VCF, the format can be written and read using BCFtools tool using the following commands:

1. `samtools mpileup -uf sample.sorted.bam | bcftools view - > sample.var.raw.bcf`  
While generating BCF file from BAM using samtools, -u is used for generating uncompressed VCF which can be piped as BCFtools designed for stream data and -f for the faidx-indexed reference file in the FASTA format.
2. `bcftools view sametools.var.raw.bcf | vcfutils.pl varFilter -D100 > sample.var.flt.vcf`
3. `samtools calmd -Abr sample.sorted.bam ~/hg38/hg38.fa > sample.baq.bam`
4. `samtools mpileup -uf ~/hg38/hg38.fa sample.baq.bam | bcftools view - > sample.baq.var.raw.bcf`

**Annotation and curation:** Post processing the files, annotation and curation of the data followed by prioritizing the candidate SNPs/variants involves a great deal of user's discretion. There are a host of tools and annotation methods meant for this. Population stratification can be one such step in this process. While the 1,000 genomes dataset (Auton *et al.*, 2015) or Genome Aggregation database (gnomAD) (see Reference 5) containing the datasets are already used to summarize worldwide population, estimating individual ancestries using ADMIXTURE would help researchers to (Alexander *et al.*, 2009) project the samples. On the other hand, downstream bioinformatics annotation can then be supplemented to integrate the results with different pathway tools, *viz.* PANTHER (Mi *et al.*, 2016) which assesses the ontology/pathways affecting the 'mutated' genes. This can be further supplemented by usage of assorted databases like Clinvar and dbSNP. In addition, global enrichment analysis and association networks using GeneMania (Warde-Farley *et al.*, 2010) would allow creating and visualizing gene networks by evidence in pathways and gene-gene/protein-protein interactions (predicted and experimental).

### Data analysis

Benchmarking yielded a good recovery rate for validation of variants/ SNPs (Figures 3, 4 and 5) while VarScan with default values was found to have highest overall sensitivity with VarScan strict parameters having the lowest overall sensitivity (Figures 6 and 7). However, we observed that the preprocessing steps have little impact on the final output, with base recalibration step using GATK Unified Genotyper identifying fewer validated SNPs when compared to VarScan. On the other hand, we found that the recovery of exon variants among the exome samples was typically high when compared to the two whole genome datasets (Figure 5B). When variant lists were confined to previously observed variants as observed from the benchmark analyses between Sentieon and GATK (Weber *et al.*, 2015), we observed that the recovery of SNPs with default parameter was found to be considerably good. Whereas changing variant calling criterion especially using VarScan, for example, imposing strict coverage requirement (Figure 7) yielded less numbers of false positives giving the number of *bona fide* or *de novo* variants (Figures 5A and 5B). This subtly proves that our benchmarking the six WES and two WGS datasets (see Table 2) is variable with the capture, sequencing, processing and post-processing/analysis in the human genome and VarScan is comparable with the GATK in terms of identifying the *de novo* variants (Figures 5A and 5B). With the wet-lab components of NGS being cumbersome, analyzing the exons or for that matter intronic

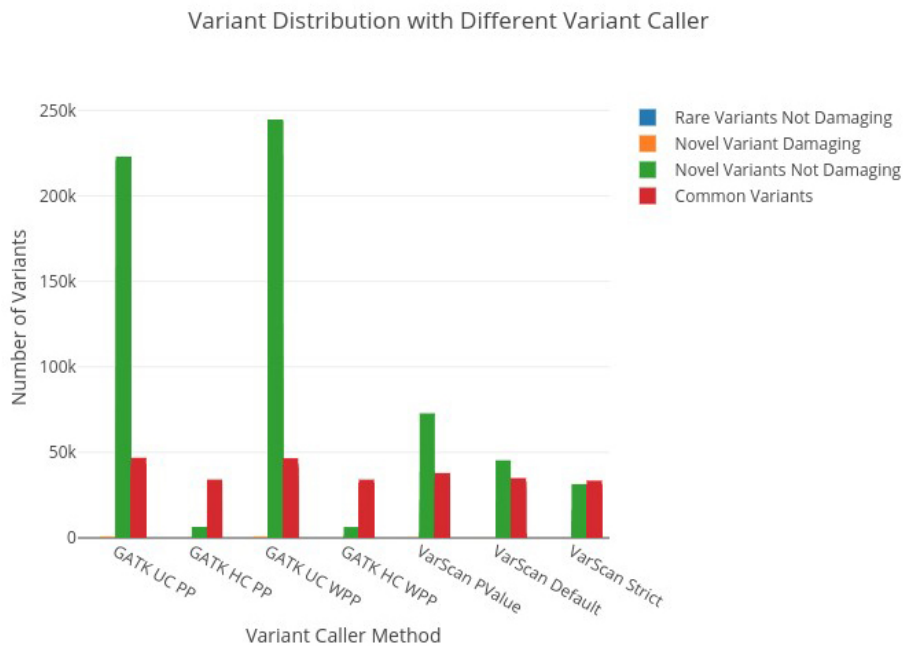
variants using bioinformatics pipeline is equally challenging. More pictorial representations such as density plots (Figure 8) are helpful for further interpretation of variants. There must be significant *in silico* hurdles and organizational steps discussed from time to time and yet at the end of the analysis, one needs to arrive at the fittest in using the discretionary tools. Although technology challenges persist in setting up certain standards and guidelines, the end-user can enhance the pipeline with further tools. In this protocol, we have essentially shown how a WES pipeline can be run using batch file process and the comparison of VarScan over GATK using benchmarked datasets.

**Table 2. 1,000 genomes samples used for benchmarking\***

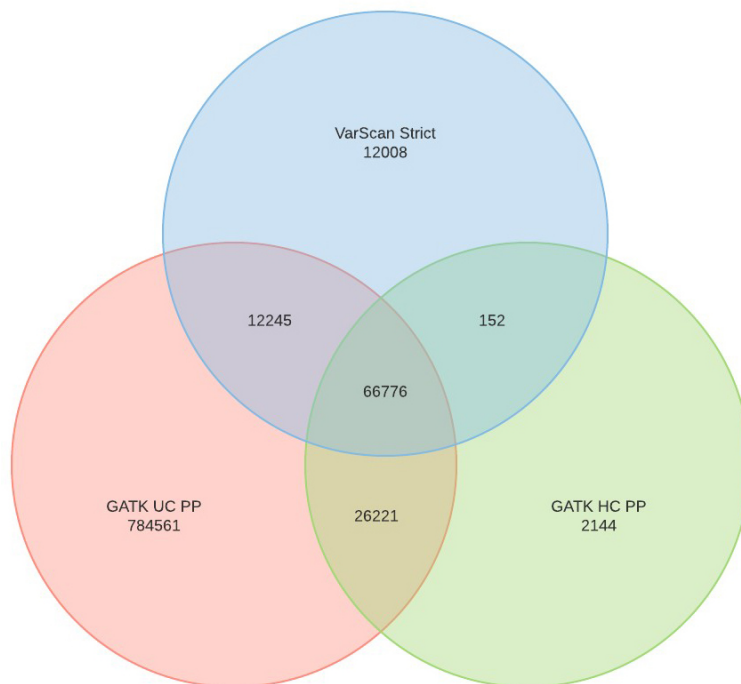
Individual	Sample	Type	Sequenced Technology	Sequencing Bases	Coverage
NA11930	SRR098416	Exome	Illumina HiSeq	145 M	3x
NA12046	SRR742200	Exome	Illumina HiSeq	5 G	102x
NA12155	SRR702068	Exome	Illumina HiSeq	7 G	140x
NA11932	ERR034544	Exome	Illumina HiSeq	11 G	251x
NA12878	SRR098401	Exome	Illumina HiSeq	16 G	341x
NA12891	SRR098359	Exome	Illumina HiSeq	16 G	347x
NA12878	SRR622461	Whole Genome	Illumina HiSeq	17 G	6x
NA12878	ERR091571	Whole Genome	Illumina HiSeq	41 G	14x

\*The sequences can be provided by the author upon request.

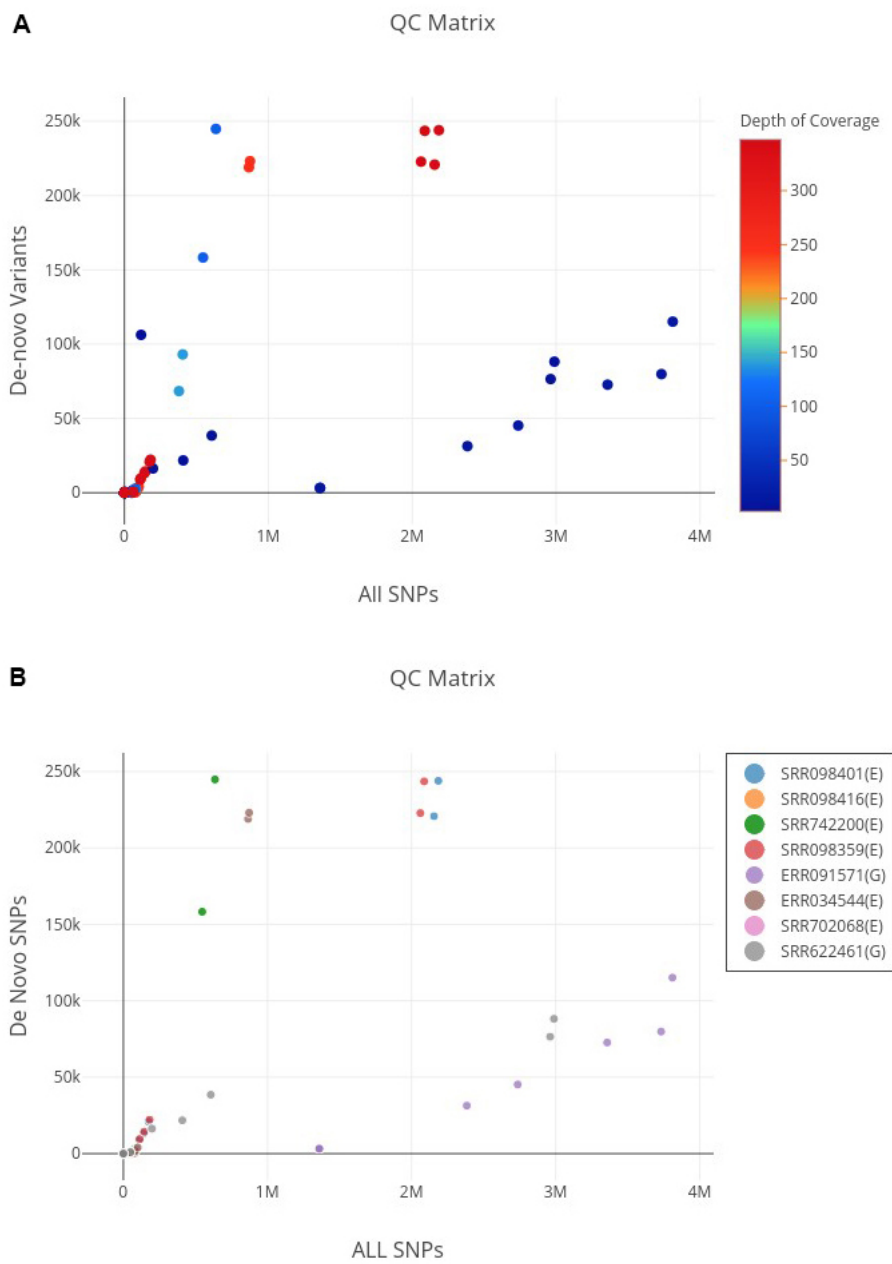




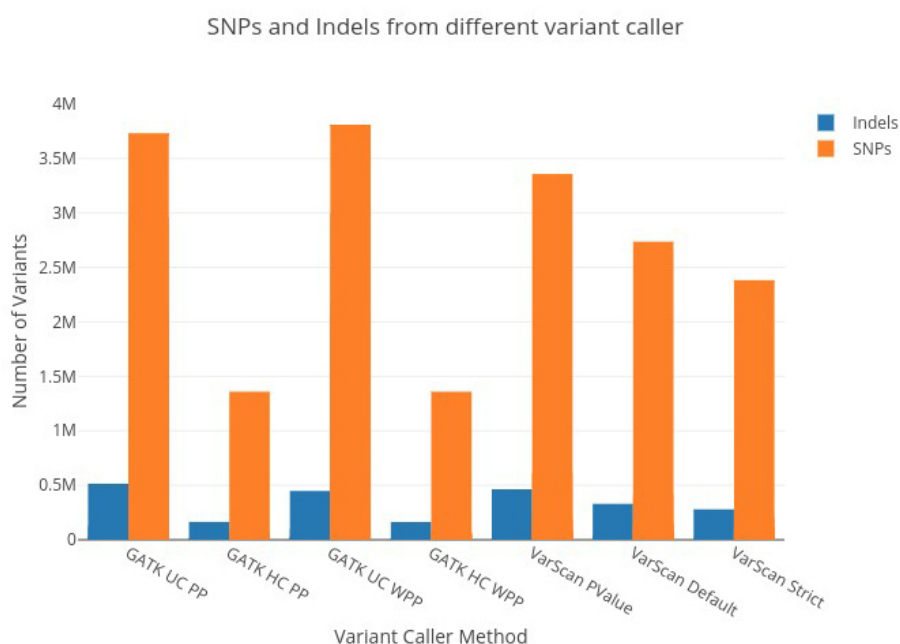
**Figure 3. Number of variants obtained from GATK and VarScan with various parameters.** We observe GATK Unified caller to have a large number of false positives while VarScan with strict parameters performed well with less number of false positives.



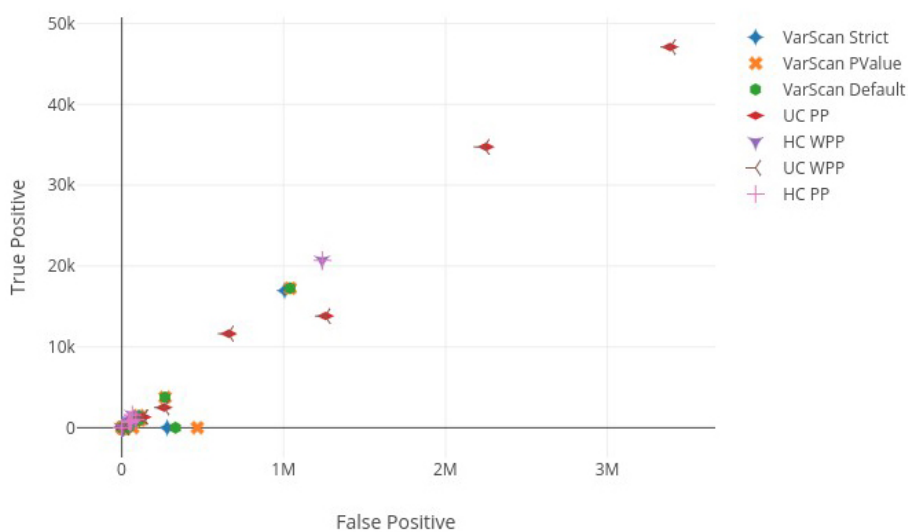
**Figure 4. Venn diagram of three methods using Haplotype caller with preprocessing (HC-PP) and Universal genotype caller with preprocessing (UC-PP) and VarScan strict on sample SRR098359.** We observed that all the three share the most true positive variants.



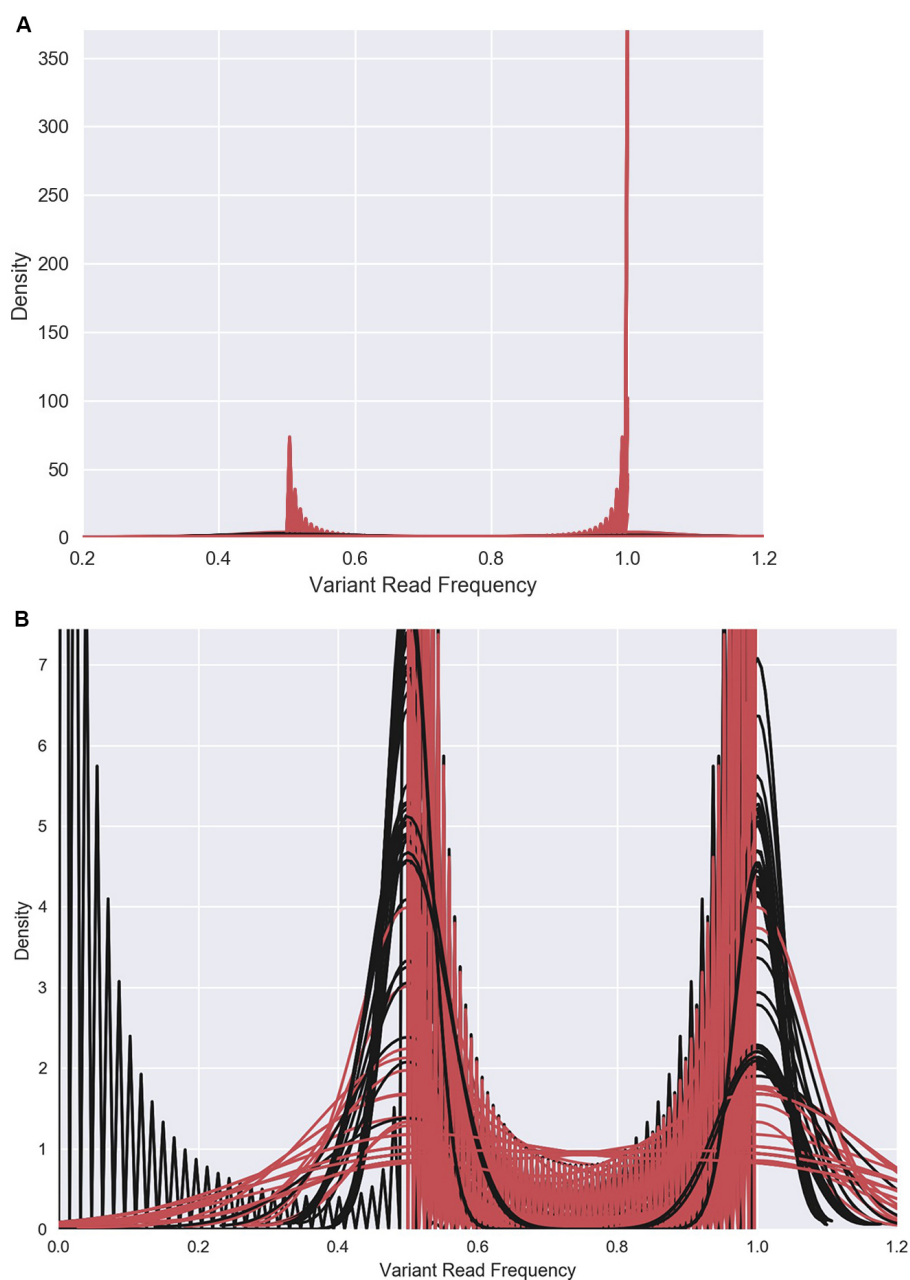
**Figure 5. Distribution of *de novo* variants with the x-axis showing million reads with depth of coverage (right in the legend) and the y-axis showing the number of *de novo* variants. A. All variants in regards to the depth of coverage of NGS run; B. *de novo* variants in regards to all SNPs against each sample.**



**Figure 6. Number of SNPs and Indels called by GATK and VarScan using all parameters against the samples.** We observed again that VarScan gave the best results with less false positive variants.



**Figure 7. Scatter plot of the number of true positives/false positives for all variant calling parameter options**



**Figure 8. Density plot of an exome NGS run for *de novo* and known variants.** The x-axis shows the variant read frequency against the density in y-axis. Panel B is the zoomed view of Panel A.

### Acknowledgments

The authors declare no conflicts of interest whatsoever. PS wants to acknowledge [biostars.org](http://biostars.org) forum which enabled him to enhance the pipeline consistently. He gratefully acknowledges the forum and immense discussions from users/researchers. The Birla Institute of Scientific Research would like to thank the Biotechnology Information System Network (BTIS), Department of Biotechnology,

Government of India for funding and providing the resources and facilities. The authors gratefully acknowledge the Indian Council Medical research towards grant # 5/41/11/2012 RMC.

## References

1. Alexander, D. H., Novembre, J. and Lange, K. (2009). [Fast model-based estimation of ancestry in unrelated individuals](#). *Genome Res* 19(9): 1655-1664.
2. Andrews, S. (2010). [A quality control tool for high throughput sequence data](#). FastQC.
3. Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A. and Abecasis, G. R. (2015). [A global reference for human genetic variation](#). *Nature* 526(7571): 68-74.
4. Chang, X. and Wang, K. (2012). [wANNOVAR: annotating genetic variants for personal genomes via the web](#). *J Med Genet* 49(7): 433-436.
5. Genome Aggregation database(gnomAD): <http://gnomad.broadinstitute.org/>
6. Hwang, S., Kim, E., Lee, I. and Marcotte, E. M. (2015). [Systematic comparison of variant calling pipelines using gold standard personal exome variants](#). *Sci Rep* 5: 17875.
7. Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. and Rice, P. M. (2010). [The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants](#). *Nucleic Acids Res* 38(6): 1767-1771.
8. Fischer, M., Snajder, R., Pabinger, S., Dander, A., Schossig, A., Zschocke, J., Trajanoski, Z. and Stocker, G. (2012). [SIMPLEX: cloud-enabled pipeline for the comprehensive analysis of exome sequencing data](#). *PLoS One* 7(8): e41948.
9. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. and Zhang, J. (2004). [Bioconductor: open software development for computational biology and bioinformatics](#). *Genome Biol* 5(10): R80.
10. Guo, Y., Ding, X., Shen, Y., Lyon, G. J. and Wang, K. (2015). [SeqMule: automated pipeline for analysis of human exome/genome sequencing data](#). *Sci Rep* 5: 14283.
11. Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., Boehnke, M. and Kang, H. M. (2012). [Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data](#). *Am J Hum Genet* 91(5): 839-848.
12. Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L. and Wilson, R. K. (2012). [VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing](#). *Genome Res* 22(3): 568-576.
13. Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M. and Maglott, D. R. (2014). [ClinVar: public archive of relationships among sequence variation and human phenotype](#). *Nucleic Acids Res* 42(Database issue): D980-985.

14. Langmead, B. and Salzberg, S. L. (2012). [Fast gapped-read alignment with Bowtie 2](#). *Nat Methods* 9(4): 357-359.
15. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009). [The sequence alignment/map format and SAMtools](#). *Bioinformatics* 25(16): 2078-2079.
16. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T. and Thomas, P. D. (2016). [PANTHER version 10: expanded protein families and functions, and analysis tools](#). *Nucleic Acids Res* 44(D1): D336-342.
17. Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J. and Trajanoski, Z. (2014). [A survey of tools for variant analysis of next-generation genome sequencing data](#). *Brief Bioinform* 15(2): 256-278.
18. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. and Sirotkin, K. (2001). [dbSNP: the NCBI database of genetic variation](#). *Nucleic Acids Res* 29(1): 308-311.
19. Singleton, A. B. (2011). [Exome sequencing: a transformative technology](#). *Lancet Neurol* 10(10): 942-946.
20. Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D. and Birney, E. (2002). [The Bioperl toolkit: Perl modules for the life sciences](#). *Genome Res* 12(10): 1611-1618.
21. Trapnell, C. and Salzberg, S. L. (2009). [How to map billions of short reads onto genomes](#). *Nat Biotechnol* 27(5): 455-457.
22. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L.,

- Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferreira, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X. (2001). [The sequence of the human genome](#). *Science* 291(5507): 1304-1351.
23. Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D. and Morris, Q. (2010). [The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function](#). *Nucleic Acids Res* 38(Web Server issue): W214-220.
24. Weber, J. A, Aldana, R, Gallagher, B. D. and Edwards, J. S. (2015). [Sentieon DNA pipeline for variant detection-Software-only solution, over 20x faster than GATK 3.3 with identical results](#). *PeerJ Pre Prints*.