New York Genome Center's Somatic Pipeline

Pipeline Version	6.0
Library	Whole Genome
Organism	Human

May 29, 2019

Table of Contents

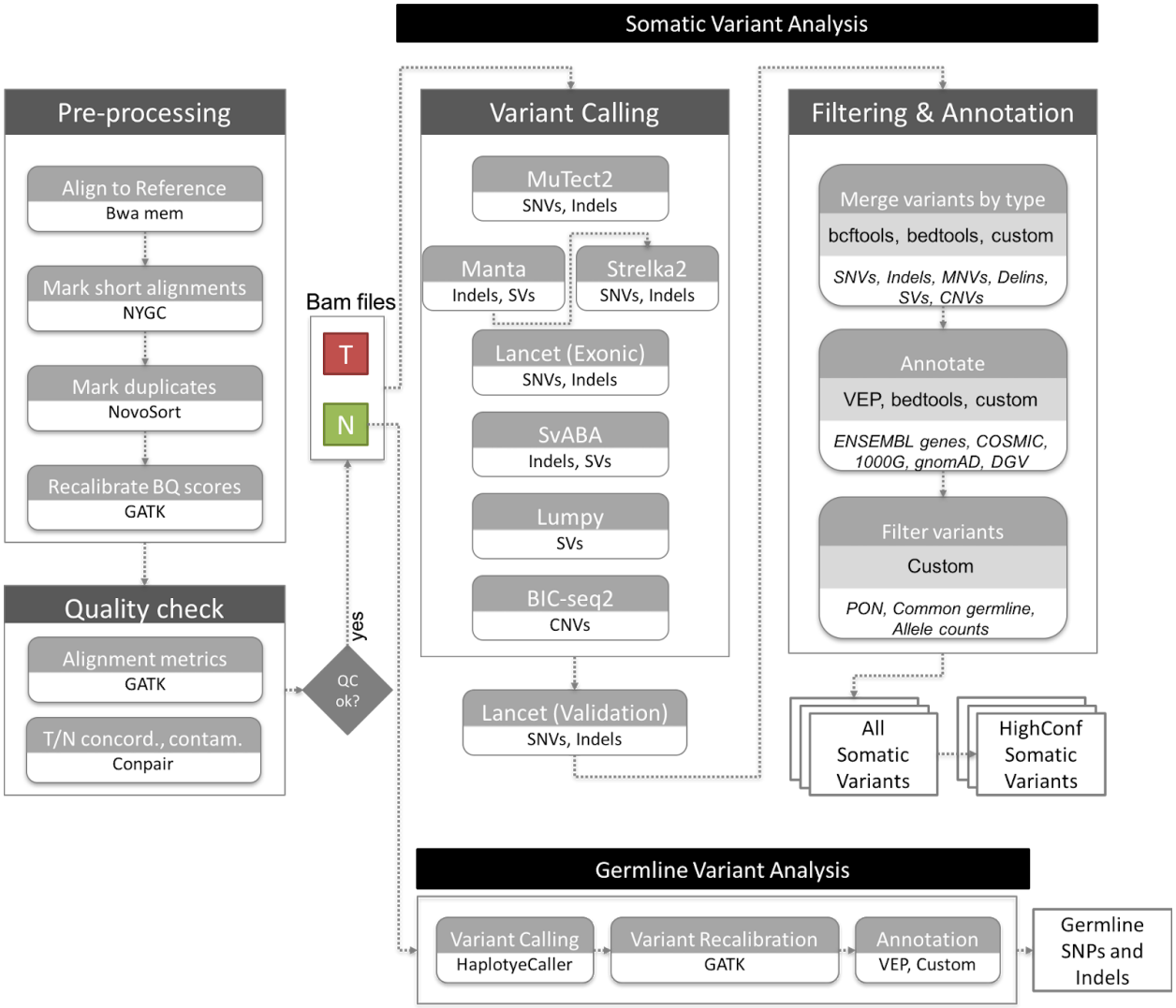

Figure1: NYGC Somatic WGS Pipeline v6

1. Pre-processing

Sequencing reads for the tumor and normal samples are aligned to the reference genome using BWA-MEM (v0.7.15) (1). NYGC's ShortAlignmentMarking (v2.1)[1] is used to mark short reads as unaligned. This tool is intended to remove spurious alignments resulting from contamination (e.g. saliva sample bacterial content) or from too aggressive alignments of short reads the size of BWA-MEM's 19bp minimum seed length. These spurious alignments result in pileups in certain locations of the genome and can lead to erroneous variant calling.

GATK (v4.1.0) (2) FixMateInformation is run to verify and fix mate-pair information, followed by Novosort (v1.03.01) markDuplicates to merge individual lane BAM files into a single BAM file

[1] https://github.com/nygenome/nygc-short-alignment-marking

per sample. Duplicates are then sorted and marked, and GATK's base quality score recalibration (BQSR) is performed. The final result of the pre-processing pipeline is a coordinate sorted BAM file for each sample.

1.1. PDX pre-processing

Patient-derived xenograft (PDX) samples undergo an additional preprocessing step. Prior to the pre-processing pipeline, mouse reads are detected and removed from the FASTQ files by aligning the data to a combined reference of mouse (GRCm38) and human (GRCh37). All read pairs with both reads mapping to mouse or one read mapping to mouse and the other unmapped are excluded from subsequent processing and analyses steps.

2. Quality control

Once preprocessing is complete, we compute a number of alignment quality metrics such as average coverage, %mapped reads and %duplicate reads using GATK (v4.1.0) and an autocorrelation metric (adapted for WGS from(3)) to check for unevenness of coverage. We also run Conpair(4), a tool developed at NYGC to check the genetic concordance between the normal and the tumor sample and to estimate any inter-individual contamination in the samples.

3. Somatic variant calling pipeline

3.1. Variant detection

The tumor and normal bam files are processed through NYGC's variant calling pipeline which consists of MuTect2 (GATK v4.0.5.1) (5), Strelka2 (v2.9.3) (6) and Lancet (v1.0.7) (7)] for calling Single Nucleotide Variants (SNVs) and short Insertion-or-Deletion (Indels), SvABA (v0.2.1) (8) for calling Indels and Structural variants (SVs), Manta (v1.4.0) (9) and Lumpy (v0.2.13) (10) for calling SVs and BIC-Seq2 (v0.2.6) (11) for calling Copy-number variants (CNVs). Manta also outputs a candidate set of Indels which is provided as input to Strelka2 (following the developers recommendation, as it improves Strelka2's sensitivity for calling indels >20nt). Due to its computing requirements, in this pipeline Lancet is only run on the exonic part of the genome. It is also run on the +/- 250nt regions around non-exonic variants that are called by only one of the other callers, to add confidence to such variants. Small SVs called by Manta are also used to add confidence to the indel calls.

3.2. Variant merging

Next, the calls are merged by variant type (SNVs, Multi Nucleotide Variants (MNVs), Indels and SVs). MuTect2 and Lancet call MNVs, however Strelka2 does not, and it also does not provide any phasing information. So to merge such variants across callers, we first split the MNVs called by MuTect2 and Lancet to SNVs, and then merge the SNV callsets across the different callers.

If the caller support for each SNV in a MNV is the same, we merge them back to MNVs. Otherwise those are represented as individual SNVs in the final callset.
Lancet and MantaSV are the only tools that can call deletion-insertion (delins or COMPLEX) events. Other tools may represent the same event as separate yet adjacent indel and/or SNV variants. Such events are relatively less frequent, and difficult to merge. We therefore do not merge COMPLEX calls with SNVs and Indels calls from other callers.

The SVs are converted to bedpe format, all SVs below 500bp are excluded and the rest are merged across callers using bedtools (12) pairtopair (slop of 300bp, same strand orientation, and 50% reciprocal overlap).

3.2. Tumor-only analysis

When a matched normal sample is not available, in its place we use a “contemporary normal”, that is, DNA from the HapMap sample NA12878 that was prepped and sequenced using the same protocol as the tumor sample. Using a contemporary normal removes some of the false positives that are due to library preparation and sequencing (that would manifest in the same way in the tumor and NA12878), as well as some germline variants that are common to the tumor sample and NA12878.

4. Somatic variant annotation

4.1. SNVs and Indels

SNVs and Indels are annotated with Ensembl as well as databases such as COSMIC (v86) (13), 1000Genomes (Phase3) (14) , ClinVar (201706) (15), PolyPhen (v2.2.2) (16), SIFT (v5.2.2) (17), FATHMM (v2.1) (18), gnomAD (r2.0.1) (19) and dbSNP (v150) (20) using Variant Effect Predictor (v93.2) (21).

4.2. CNVs and SVs

For CNVs, segments with log2 > 0.2 are categorized as amplifications, and segments with log2 < -0.235 are categorized as deletions (corresponding to a single copy change at 30% purity in a diploid genome, or a 15% Variant Allele Fraction). CNVs of size less than 20Mb are denoted as focal and the rest are considered large-scale.
We use bedtools (12) for annotating SVs and CNVs. All predicted CNVs are annotated with germline variants by overlapping with known variants in 1000 Genomes and Database of Genomic Variants (DGV) (22). Cancer-specific annotation includes overlap with genes from Ensembl (23) and Cancer Gene Census in COSMIC, and potential effect on gene structure (e.g. disruptive, intronic, intergenic). If a predicted SV disrupts two genes and strand orientations are compatible, the SV is annotated as a putative gene fusion candidate. Note that we do not check reading frame at this point. Further annotations include sequence features within breakpoint flanking regions, e.g. mappability, simple repeat content and segmental duplications.

5. Somatic variant filtering

5.1. Panel Of Normals

The Panel Of Normals (PON) filtering removes recurrent technical artifacts from the somatic variant callset (5).

5.1.1. PON generation

The Panel of Normals for SNVs, indels and SVs was created with whole-genome sequencing data from normal samples from 242 unrelated individuals. Of these, sequencing data for 148 individuals was obtained from the Illumina Polaris project[2] which was sequenced on the HiSeqX platform with PCR-free sample preparation. The remaining samples were sequenced by the NYGC. Of these, 73 individuals were sequenced on HiSeqX, 11 on NovaSeq, and 10 were sequenced on both.
We ran MuTect2 in artifact detection mode and Lumpy in single sample mode on these samples. For SNVs and indels, we created a PON list file with sites that were seen in two or more individuals.
For SVs, we used SURVIVOR (v1.0.3) (24) to merge Lumpy calls. Variants were merged if they were of the same type, had the same strand orientation, and were within 300bp of each other (maximum distance). We did not specify a minimum size. After merging SVs, we used these calls as a PON list.

5.1.2. PON filtering

For SNVs and Indels, we use the PON list to filter the somatic variants in the merged SNV and indel files. To filter our somatic SV callset, we merge our PON list with our callset using bedtools pairtopair (slop of 300bp, same strand orientation, and 50% reciprocal overlap), and filtered those SVs found in two or more individuals in our PON.

5.2. Common germline variants

In addition to the PON filtering, we remove SNVs and Indels that have minor allele frequency (MAF) of 1% or higher in either 1000Genomes (phase 3) or gnomAD (r2.0.1) (25), and SVs that overlap DGV and 1000Genomes (phase3). CNVs are annotated with DGV and 1000 Genomes but not filtered.

5.3. Allele counts

Since our variant callsets are generated by merging calls across callers, and each of them reported different allele counts, we report final chosen allele counts for SNVs and indels. For SNVs, and for indels less than 10nt in length, these are computed as the number of unique

[2] https://github.com/Illumina/Polaris

read-pairs supporting each allele using the pileup method, with minimum mapping quality and base quality thresholds of 10 each.
For larger indels and complex (deletion-insertion) events, we choose the final allele counts reported by the individual callers Strelka2, MuTect2, Lancet, in that order. For indels larger than 10nt that are only called by SvABA, we do not report final allele counts and allele frequencies because SvABA does not report the reference allele count, making it difficult to estimate the variant allele frequency.
We then use these final chosen allele counts and frequencies to filter the somatic callset. Specifically, we filter any variant for which the variant allele frequency (VAF) in the tumor sample is less than 0.0001, or if the VAF in the normal sample is greater than 0.2, or if the depth at the position is less than 2 in either the tumor sample or the normal sample. We also filter variants for which the VAF in normal sample is greater than the VAF in tumor sample.

5.4. All Somatic and High-confidence variants

Variants that pass all of the above-mentioned filters are included in our final somatic callset (hereby referred to as AllSomatic).
For SNVs, indels and SVs, we also annotate a subset of the somatic callset as high confidence.
For SNVs and indels, high confidence calls are defined as those that are either called by two or more variant callers, or called by one caller and also seen in the Lancet validation calls or in the Manta SV calls.
For structural variants, high confidence calls are taken from the somatic callset if they meet the following criteria: called by 2 or more variant callers, or called by Manta or Lumpy with either additional support from nearby CNV changepoint or split-read support from SplazerS (26), an independent tool used to calculate the number of split-reads supporting SV breakpoints. An SV is considered supported by SplazerS if it found at least 3 split-reads in the tumor only. Nearby CNV changepoints are determined by overlapping BIC-Seq2 calls with the SV callset using bedtools closest. An SV is considered to be supported by a CNV changepoint if the breakpoint of the CNV is within 1000bp of an SV breakpoint.

6. Germline variant analysis

We call germline SNPs and indels on the matched normal sample using GATK HaplotypeCaller (v3.5), which generates a single-sample GVCF. We then run GATK's GenotypeGVCF to perform single sample genotype refinement and output a VCF, followed by variant quality score recalibration (VQSR) for variant filtering (at tranche 99.6%). Next, we run Variant Effect Predictor (v93.2) to annotate the variants with Ensembl as well as databases such as COSMIC (v86), 1000Genomes (Phase3), gnomAD (r2.0.1), dbSNP (v150), ClinVar (201805), Polyphen2 (v2.2.2) and SIFT (v5.2.2).

7. MSI detection

We run MANTIS (v1.0.4) (27) for Microsatellite Instability (MSI) detection in microsatellite loci (found using RepeatFinder, a tool included with MANTIS). A sample is considered to be

microsatellite unstable if it's Step-Wise Difference score reported by MANTIS is greater than 0.4 (or 0.62[3] in absence of a matched-normal). Otherwise it is considered to be microsatellite stable (MSS).

8. HLA-typing

We run OptiType (v1.3.2) (28) and Kourami[4] (v0.9.6) (29) on the matched normal sample for Human Leukocyte Antigen (HLA)-typing. OptiType predicts major histocompatibility complex (MHC) Class I alleles (HLA-A, HLA-B, HLA-C), whereas Kourami predicts both MHC Class I and Class II alleles (HLA-DP, HLA-DQ, HLA-DR).

9. Mutational signature analysis

We run deconstructSigs (v1.8.0) (30) on the High Confidence somatic SNV callset within autosomes to estimate contribution of known COSMIC mutational signatures (v2 - March 2015)[5] in the tumor sample.

[3] Note: Threshold chosen based on internal benchmarking
[4] Note: Kourami only supports build 38 of the human genome and therefore is not run if data is aligned to GRCh37
[5] https://cancer.sanger.ac.uk/cosmic/signatures_v2

10. References

1. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*.

2. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M., *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

3. Zhang,L. and Zhang,L. (2013) Use of autocorrelation scanning in DNA copy number analysis. *Bioinformatics*, **29**, 2678–2682.

4. Bergmann,E.A., Chen,B.-J., Arora,K., Vacic,V. and Zody,M.C. (2016) Conpair: concordance and contamination estimator for matched tumor-normal pairs. *Bioinformatics*, **32**, 3196–3198.

5. Cibulskis,K., Lawrence,M.S., Carter,S.L., Sivachenko,A., Jaffe,D., Sougnez,C., Gabriel,S., Meyerson,M., Lander,E.S. and Getz,G. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.

6. Kim,S., Scheffler,K., Halpern,A.L., Bekritsky,M.A., Noh,E., Källberg,M., Chen,X., Kim,Y., Beyter,D., Krusche,P., *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, **15**, 591–594.

7. Narzisi,G., Corvelo,A., Arora,K., Bergmann,E.A., Shah,M., Musunuri,R., Emde,A.-K., Robine,N., Vacic,V. and Zody,M.C. (2018) Genome-wide somatic variant calling using localized colored de Bruijn graphs. *Commun Biol*, **1**, 20.

8. Wala,J.A., Bandopadhayay,P., Greenwald,N.F., O'Rourke,R., Sharpe,T., Stewart,C., Schumacher,S., Li,Y., Weischenfeldt,J., Yao,X., *et al.* (2018) SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.*, **28**, 581–591.

9. Chen,X., Schulz-Trieglaff,O., Shaw,R., Barnes,B., Schlesinger,F., Källberg,M., Cox,A.J., Kruglyak,S. and Saunders,C.T. (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, **32**, 1220–1222.

10. Layer,R.M., Chiang,C., Quinlan,A.R. and Hall,I.M. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.

11. Xi,R., Lee,S., Xia,Y., Kim,T.-M. and Park,P.J. (2016) Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res.*, **44**, 6274–6286.

12. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

13. Tate,J.G., Bamford,S., Jubb,H.C., Sondka,Z., Beare,D.M., Bindal,N., Boutselakis,H., Cole,C.G., Creatore,C., Dawson,E., *et al.* (2019) COSMIC: the Catalogue Of Somatic

Mutations In Cancer. *Nucleic Acids Res.*, **47**, D941–D947.

14. 1000 Genomes Project Consortium, 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., *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

15. 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 Research*, **42**, D980–D985.

16. Adzhubei,I., Jordan,D.M. and Sunyaev,S.R. (2013) Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics*, **76**, 7.20.1–7.20.41.

17. Vaser,R., Adusumalli,S., Leng,S.N., Sikic,M. and Ng,P.C. (2016) SIFT missense predictions for genomes. *Nat. Protoc.*, **11**, 1–9.

18. Shihab,H.A., Gough,J., Mort,M., Cooper,D.N., Day,I.N.M. and Gaunt,T.R. (2014) Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum. Genomics*, **8**, 11.

19. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B., *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.

20. 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**, 308–311.

21. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17**, 122.

22. MacDonald,J.R., Ziman,R., Yuen,R.K.C., Feuk,L. and Scherer,S.W. (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, **42**, D986–92.

23. Hubbard,T. (2002) The Ensembl genome database project. *Nucleic Acids Research*, **30**, 38–41.

24. Jeffares,D.C., Jolly,C., Hoti,M., Speed,D., Shaw,L., Rallis,C., Balloux,F., Dessimoz,C., Bähler,J. and Sedlazeck,F.J. (2017) Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.*, **8**, 14061.

25. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B., *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.

26. Emde,A.-K., Schulz,M.H., Weese,D., Sun,R., Vingron,M., Kalscheuer,V.M., Haas,S.A. and Reinert,K. (2012) Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS. *Bioinformatics*, **28**, 619–627.

27. Kautto,E.A., Bonneville,R., Miya,J., Yu,L., Krook,M.A., Reeser,J.W. and Roychowdhury,S.

(2017) Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. *Oncotarget*, **8**, 7452–7463.

28. Szolek,A., Schubert,B., Mohr,C., Sturm,M., Feldhahn,M. and Kohlbacher,O. (2014) OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*, **30**, 3310–3316.

29. Lee,H. and Kingsford,C. (2018) Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. *Genome Biol.*, **19**, 16.

30. Rosenthal,R., McGranahan,N., Herrero,J., Taylor,B.S. and Swanton,C. (2016) DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.*, **17**, 31.