New York Genome Center's Somatic Pipelines

January 12, 2017

Preprocessing

Before calling, tumor and matched normal DNA sequencing data go through our somatic pre-processing pipeline which includes aligning reads to the GRCh37 human reference genome using the Burrows-Wheeler Aligner (BWA) aln (Li and Durbin, 2009), marking of duplicate reads by the use of NovoSort (a multi-threaded bam sort/merge tool by Novocraft technologies http://www.novocraft.com); realignment around indels (done jointly for all samples derived from one individual, e.g. tumor and matched normal samples, or normal, primary and metastatic tumor trios) and base recalibration via Genome Analysis Toolkit (GATK) (McKenna *et al.*, 2010).

Figure 1: NYGC pre-processing pipeline.

Quality control

Basic DNA sequencing metrics. We run a battery of Picard (QualityScoreDistribution, MeanQualityByCycle, CollectBaseDistributionByCycle, CollectAlignmentSummaryMetrics, CollectInsertSizeMetrics, CollectGcBiasMetrics, CollectOxoGMetrics) and GATK (FlagStat, ErrorRatePerCycle) metrics on all DNA data. In addition, for WGS experiments we run bedToolsCoverage and custom R scripts to compute sequencing depth of coverage, and for exomes and panels we run GATK CalculateHsMetrics and DepthOfCoverage modules. We perform outlier detection to identify samples that need to be manually reviewed, and if verified not to pass QC, failed.

Sample contamination and tumor-normal concordance. We run Conpair (Bergmann *et al*., 2016) on all tumor-normal pairs to to detect cross-individual contamination and sample mix-ups.

Autocorrelation. We compute a metric called Autocorrelation (*Zhang et al.*, 2013) to give us an indication of unevenness in coverage in sequencing data. This method was originally developed for array data but we have adapted it for WGS data. We generate intervals with window size of 1kb every 10kb along the genome, calculate read depth in these windows using Picard HsMetrics and then compute Autocorrelation.

Calling SNVs and indels

We return the union of somatic SNVs called by muTect (Cibulskis *et al.*, 2013), Strelka (Saunders *et al.*, 2012) and LoFreq (Wilm *et al.*, 2012) and the union of indels called by Strelka, and somatic versions of Pindel (Ye *et al.*, 2009) and Scalpel (Narzisi *et al.*, 2014).

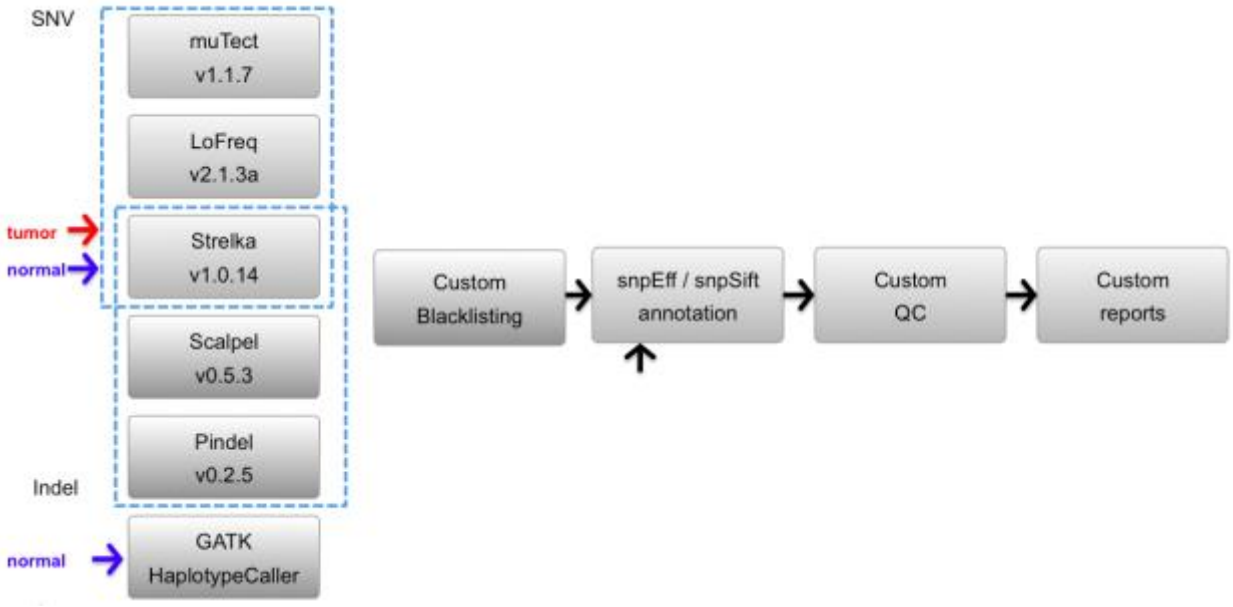


Figure 2: NYGC somatic SNV/indel pipeline.

The choice of SNV callers was based on internal benchmarking of individual and combinations of callers on a synthetic virtual tumor created by spiking reads from two HapMap samples in a way that mimics somatic variants with predefined variant allele frequencies (Cibulskis *et al.*, 2013). The choice of indel callers was based on internal benchmarking on synthetic data from the DREAM challenge (Ewing *et al.*, 2015).

For human samples, we also return germline calls in a panel of cancer risk genes (APC, ATM, BARD1, BMPR1A, BRCA1/2, BRIP1, CDH1, CDK4, CDKN2A, CHEK2, CYLD, EPCAM, IDH1/2, MEN1, MET, MLH1, MSH2/6, MUTYH, NBN, NF1/2, PALB2, PMS1/2, PRKAR1A, PTCH1, PTEN, RAD51C/D, RB1, RET, SDHAF2, SDHB/C/D, SMAD4, STK11, TP53, TSC1/2, VHL, WRN, WT1), made by the use of GATK HaplotypeCaller.

Calling CNVs and SVs [WGS data only]

Structural variants (SVs), such as deletions and amplifications as well as copy-neutral genomic rearrangements are detected by the use of multiple tools (NBIC-seq (*Xi et al.*, 2016), Crest (*Wang et al.*, 2011), Delly (*Rausch et al.*, 2012), BreakDancer (*Chen et al*., 2009)) that employ complementary detection strategies, such as inspecting read depth within genomic windows, analyzing discordant read pairs, and identifying breakpoint-spanning split reads.

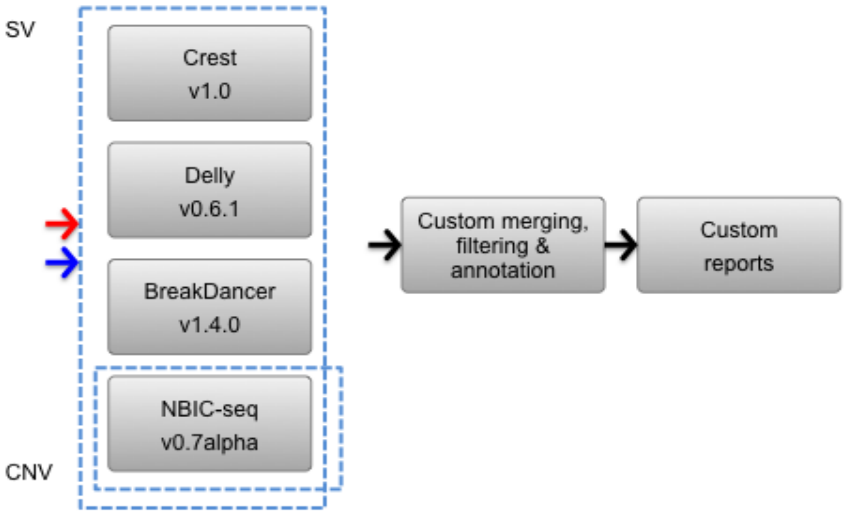


Figure 3: NYGC somatic WGS CNV/SV pipeline.

Calling CNVs [WES data only]

We use EXCAVATOR (*Magi et al*., 2013), a read depth based tool, to detect copy-number variants (CNVs) such as deletions and amplifications.

Figure 4: NYGC somatic WES CNV pipeline.

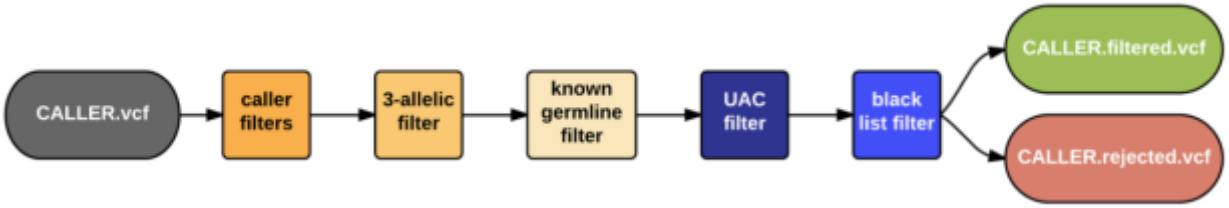
Calling variants without a matched normal [human samples only]

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 prep and sequencing (that would manifest in the same way in the tumor and NA12878), as well as (mostly common) germline variants that are common to the tumor sample and NA12878.

Processing of patient-derived xenograft (PDX) samples

PDX samples undergo an additional preprocessing step. Prior to the preprocessing pipeline, mouse reads are detected and removed by aligning the data to a combined reference genome of mouse (GRCm38/mm10) and human (GRCh37). All read pairs with both reads mapping to mouse or one read mapping to mouse and one unmapped are excluded from the subsequent processing and analyses steps.

Filtering SNVs and indels



We use a multi-step filtering process:

Figure 5: The NYGC custom multi-step SNV/indel filtering

Default caller filters. SNVs and indels are filtered using the default filtering criteria as natively implemented in each of the callers. For Pindel and Scalpel (natively germline callers) we use custom in-house scripts for filtering. For each caller we keep these variants:

- LoFreq: FILTER=PASS
- muTect: variants with "PASS" in the filter field of the VCF file, which is equivalent to "KEEP" in the text file
- Strelka: FILTER=PASS
- Pindel: FILTER=PASS
- Scalpel: FILTER=PASS

Triallelic positions. The latest revision of the pipeline removes triallelic positions. Some SNV callers (e.g. muTect) remove them by default, and our internal investigation showed that triallelic sites within a sample are by and large due to an unmatched normal, not observing the second allele in the normal because of low coverage in the normal at that locus, or mapping artifacts.

Common germline variants.

Human samples:

The resulting set of SNVs and indels is further filtered with common variants seen at MAF ≥ 5% in DNMT3A, TET2, JAK2, ASXL1, TP53, GNAS, PPM1D, BCORL1 and SF3B1 genes (see Xie *et al.*, 2014) and with MAF ≥ 1% elsewhere in the genome, as reported in the 1000 Genomes Project release 3 (1000 Genomes Project Consortium, 2012) and the Exome Aggregation Consortium (ExAC) server (http://exac.broadinstitute.org), because these are very unlikely to be important in cancer.

Mouse samples:

The resulting set of SNVs and indels is further filtered with variants seen in dbSNPv138 and Mouse Genome Project (v3).

UAC filter. Because callers often return different ref/alt allele counts for the same variant we introduced unified allele counts (UAC). Computation of UAC is based on the bam-readcount tool (Larson *et al.*, 2012). For each variant we generate 4 values that are independent of callers:

tumor-ref, tumor-alt, normal-ref, normal-alt. If the tumor_VAF < normal_VAF we discard the variant.

Artifacts [human samples only]. In addition, we remove a subset of artifactual calls by the use of an blacklist created by calling somatic variants on 16 random pairings of 80x/40x in-house sequenced HapMap WGS data.

More. If you wish to further filter the variant call set, the bam-readcount tool (https://github.com/genome/bam-readcount) will provide a list of technical co-variates (eg. mapping or base quality statistics) for each position in the tumor and normal BAM files.

Annotation and prioritization of SNVs and indels

Human samples:

Variants are annotated for their effect (non-synonymous coding, nonsense, etc.) using snpEff (Cingolani *et al.*, 2012) based on human genome annotations from ENSEMBL. We further annotate the variants via snpEff, snpSift and GATK VariantAnnotator module with information from COSMIC (Forbes *et al.*, 2012), 1000 Genomes Project, ExAC, CIViC (Clinical Interpretation of Variants in Cancer, https://civic.genome.wustl.edu), UniProt (http://www.uniprot.org), etc. We return variant prioritization scores for coding changes based on CHASM (Carter *et al.,* 2009), MutationAssessor (Reva *et al.*, 2011) and FATHMM Somatic (Shihab *et al.,* 2013).

Mouse samples:

Variants are annotated for their effect (non-synonymous coding, nonsense, etc.) using snpEff (Cingolani *et al.*, 2012) based on mouse genome annotations from ENSEMBL.

Filtering and annotation of SVs and CNVs [WGS data only]

All filtering and annotation of SVs and CNVs is done with in-house scripts, making heavy use of bedtools (http://bedtools.readthedocs.org).

SV merging. We merge and annotate SVs called by Crest, Delly and BreakDancer using BEDPE format. Two SV calls are merged if they share at least 50% reciprocal overlap (for intra-chromosomal SVs only), their predicted breakpoints are within 300bp of each other and breakpoint strand orientations match for both breakpoints. Thus, merging is done independent of which SV type was assigned by the SV caller (a classification that we found to be unreliable and variable from caller to caller).

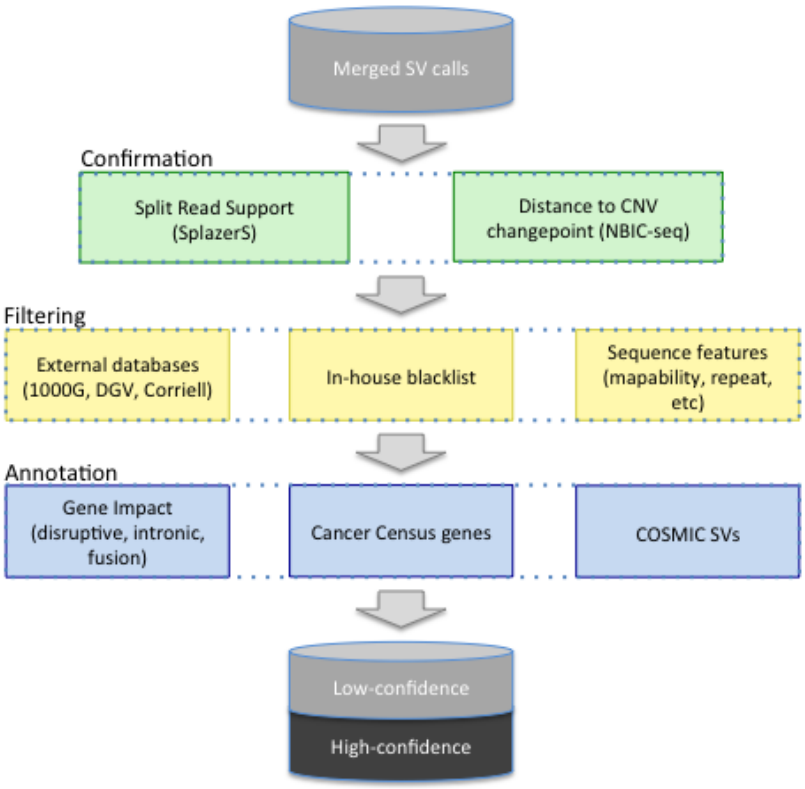


Figure 6: NYGC somatic CNV/SV filtering and annotation pipeline.

Additional SV confirmation. After merging, we annotate each SV with the closest CNV changepoint as detected by NBIC-seq from read depth signals. This adds confidence to true SV breakpoints that are not copy-neutral. Additionally, we do an independent sensitive split read check for each breakpoint using SplazerS. Apart from adding confidence and basepair precision to the breakpoint, this step also helps remove remaining germline SVs also found in the normal.

SV filtering. Some SV callers still suffer from large numbers of false positives; those are often due to germline SVs overlooked in the normal, e.g. because of low coverage or an unmatched normal, or systematic artifacts due to mapping ambiguities. We annotate and filter germline variants through overlap with known SVs (1000G call set, DGV for human; MGP for mouse) as well as through overlap with an in-house blacklist of SVs (germline SVs and artifacts called in healthy genomes). As mentioned above, also the split read check helps remove remaining germline SVs.
Finally, we prioritize SVs that were called by more than one tool, or called by only one tool but also confirmed by 1) a CNV changepoint, or 2) at least 3 split reads (in tumor only). Since we found them to be very specific, we also keep Crest-only calls in the high confidence set.

SV/CNV Annotation. All predicted copy number and structural variants are annotated with gene overlap (RefSeq, Cancer Census) 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, segmental duplications and Alu repeats.

Filtering and annotation of CNVs [WES data only]

All filtering and annotation of CNVs is done with in-house scripts, making heavy use of bedtools (http://bedtools.readthedocs.org).

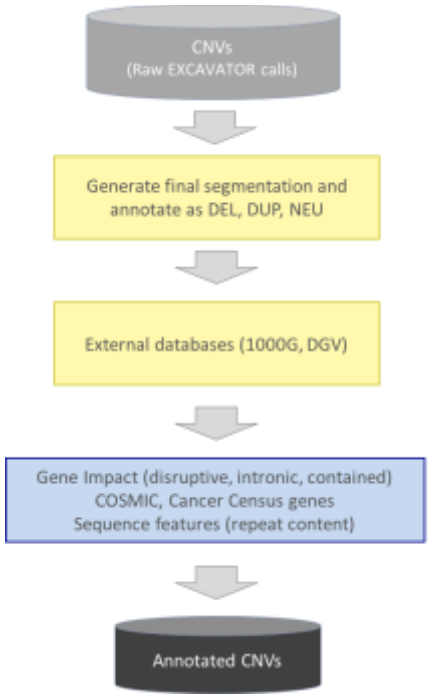


Figure 6: NYGC somatic CNV annotation pipeline.

Final Segmentation. Adjacent targets (intervals) from the same chromosome and having the same normalized mean read count are merged together to generate the final segmentation and further annotated as deletion, amplification or copy-neutral.

Annotation. All predicted CNVs are annotated with germline variants through overlap with known events (1000G call set, DGV for human). Cancer-specific annotation includes overlap with genes (RefSeq, Cancer Census) and potential effect on gene structure (e.g. disruptive, intronic, intergenic). Sequence features within breakpoint flanking regions, e.g. mappability, simple repeat content, segmental duplications and Alu repeats are also annotated. CNVs of size <20Mb are denoted as focal and the rest are large-scale.

Delivered files

We return the caller-ready BAM files (*.final.bam) for the tumor and matched normal sample.

SNVs/indels. The SNV/indel pipeline returns the raw outputs of all variant callers, in VCF format (and for muTect also in TXT format).

We in addition return the annotated union of all SNVs (*.snv.union.v*.*), union of all indels (*.indel.union.v*.*), and union of all SNVs and indels together (*.union.v*.*), in three formats:

1. VCF - union of individual caller output VCFs, combined using the GATK CombineVariants module;

2. MAF - Mutation Annotation Format, as specified by TCGA (https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+(MAF)+Specification) and modified with variant/reference counts columns to be compatible with MSKCC cBioPortal (http://www.cbioportal.org/public-portal);

3. TXT - tab-separated text file, easiest to read and parse but unlike the previous two, this is not a standard, widely accepted file format.

In the *.union.v*.annotated.txt files, the column named “CALLED_BY” indicates the tool(s) that called it. For SNVs this can be:

- mutect
- strelka_snv
- lofreq
- mutect-strelka_snv
- mutect-lofreq
- lofreq-strelka_snv
- mutect-lofreq-strelka_snv

And for indels:

- pindel
- scalpel
- strelka_indel
- pindel-strelka_indel
- scalpel-pindel
- scalpel-pindel-strelka_indel

SVs/CNVs [WGS data only]. We deliver the raw caller output which comes in a variety of formats (please refer to the individual caller documentation for details). For Delly, these are all files containing “sv.delly”, for BreakDancer “sv.breakdancer”, for Crest “sv.crest” and for NBIC-seq “sv.bicseq”.

The output of our SV processing pipeline is in extended BEDPE format (see ReadmeSV_v3.txt) and comes at two levels of confidence:

a. Merged files containing calls from all tools
b. High-confidence files

The full union of calls (a) without any filtering typically still contains many germline variants. The high-confidence variants (b), however, may miss especially low-frequency variants. For an intermediate filtering level we recommend to keep only lines with “known=;” (germline/artifact filter) from the union file.

For more details on files delivered please see the ReadmeSV_v3.txt within Documents in the project root directory.

CNVs [WES data only].

We deliver the raw caller output for EXCAVATOR “HSLMResults_*.txt”. Our CNV processing pipeline generates the final EXCAVATOR segmentation file “cnv.excavator.v2.2”, which contains deletion, amplification and copy-neutral segments, in BED format and the CNV profile. The segments are further annotated to generate an annotated file in extended BED format.

For more details on files delivered please see ReadmeCNV_v3.txt within Documents in the project root directory.

Frequently Asked Questions

Q: Which mouse genome reference is used for mouse projects?
A: GRCm38/mm10.

Q: What about human reference GRCh38?
A: We currently do not run our pipelines on GRCh38. Once all tools in our pipeline have been thoroughly tested and all annotation databases have been lifted to GRCh38, we will port our pipelines to the next human genome reference.

Q: Why are you not running BWA mem as the alignment algorithm?
A: We are in the process of evaluating the best parameter settings for BWA mem.

Q: Are duplicate reads removed from the BAM files that you deliver?
A: No, they are not removed, only marked as duplicates (0x400 bit is set in the SAM/BAM bitwise flag).

Q: Would you get different results had you removed duplicate reads?
A: The SNV and indel callers that we use recognize the duplicate bit and do not incorporate duplicate reads in the models for calling variants, so the results will be the same no matter if duplicate reads were removed or just marked. However, for calling CNVs and SVs, Crest and NBIC-seq do not recognize the duplicate bit and in this case, we ran these callers on internal BAM files from which duplicates were removed.

Q: Why are you running NovoSort instead of Picard MarkDuplicates for marking duplicates?
A: NovoSort is a multi-threaded tool which allows to mark duplicates while sorting the bam files. It significantly reduces run times because of multi-threading and by combining sort & merge in one step. We did extensive benchmark on several cancer and non-cancer samples to compare results from NovoSort and Picard MarkDuplicates, and found that the results were very similar for the two tools. We therefore switched to NovoSort to reduce the data processing time and complexity.

Q: I have SNV calls made by Virmid, but not by LoFreq? I have indel calls made by SomaticIndelDetector, but not by Pindel nor Scalpel?
A: We benchmark new tools as they become available, and Virmid (Kim *et al.*, 2013) was part of version 3 of somatic SNV calling pipelines, but was replaced by LoFreq in pipeline version 4. SomaticIndelDetector gave way to Pindel and Scalpel in pipelines version 5. If your calls were made using an older version of a pipeline, you can e-mail your project manager to get an older version of this document.

Q: Why do I see AD field in the raw muTect output file, but MUTECT_AD in the union file?
A: When we combine VCF outputs of individual callers into the union VCF, we may prepend the fields in order to avoid naming conflicts.

References

1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*. 491(7422):56-65. PMID: 23128226

Bergmann EA, *et al*. (2016) Conpair: concordance and contamination estimator for matched tumor–normal pairs. *Bioinformatics*. 32(20):3196-3198. PMID: 27354699

Carter H, *et al.* (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.* 69:6660–7. PMID: 19654296

Chen K, *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods.* 6(9):677-81. PMID: 19668202

Cibulskis K, *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*. 31(3):213-9. PMID: 23396013

Cingolani P, *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly.* 6(2):80-92. PMID: 22728672

Ewing AD, *et al.* (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature Methods*. PMID: 25984700

Forbes SA, *et al.* (2012) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* 39(suppl 1): D945-D950. 2011. PMID: 19906727

Gonzalez-Perez A, *et al.* (2012) Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med.* 4:89. PMID: 23181723

Jun G, *et al*. (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J of Human Genetics.* 91(5):839-48. PMID: 23103226

Kim S, *et al.* (2013) Virmid: accurate detection of somatic mutations with sample impurity inference. Genome Biology. 14(8):R90. PMID: 23987214

Larson DE, *et al*. (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*. PMID: 22155872

Li H, Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25(14):1754-1760. PMID:19451168

McKenna A, *et al*. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*. 20(9):1297-1303. PMID: 20644199

Narzisi G, *et al*. (2014) Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nature Methods*. 11(10):1033-6. PMID: 25128977

Rausch T, *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 15;28(18):i333-i339. PMID: 22962449

Reva B, *et al.* (2011) Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics. *Nucleic Acids Res*. 39(17):e118. PMID: 21727090

Saunders CT, *et al.* (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*. 28(14):1811-7. PMID: 22581179

Shihab HA, *et al.* (2013) Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics.* 29: 1504–1510, PMID: 23620363

Wang J, *et al.* (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nature Methods.* 8(8):652-4. PMID: 21666668

Wilm A, *et al.* (2012) LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res*. 40(22):11189-201. PMID: 23066108

Xi R, *et al.* (2011) Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci U S A*. 108(46):E1128-36. PMID: 22065754

Xie M, *et al.* (2014) Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Medicine.* 20(12):1472-8. PMID: 25326804

Ye K, *et al*. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics.* 25(21):2865-71. PMID: 19561018

Zhang L, *et al*. (2013) Use of autocorrelation scanning in DNA copy number analysis. *Bioinformatics.* 29(21):2678–2682. PMID: 24045776