Somatic mutation detection in plants: a critical evaluation through simulations and reanalyses

Sylvain Schmitt^{1,*}, Thibault Leroy^{2,3}, Myriam Heuertz⁴, Niklas Tysklind⁵

¹CNRS, UMR EcoFoG (Agroparistech, Cirad, INRAE, Université des Antilles, Université de la Guyane), Campus Agronomique, 97310 Kourou, French Guiana

²Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, A-1030 Vienna, Austria

³IRHS-UMR1345, Université d'Angers, INRAE, Institut Agro, SFR 4207 QuaSaV, 49071 Beaucouzé, France

⁴Université Bordeaux, INRAE, BIOGECO, 69 route d'Arcachon, CS 80227, 33612 Cestas Cedex, France

⁵INRAE, UMR EcoFoG (Agroparistech, CNRS, Cirad, Université des Antilles, Université de la Guyane), Campus Agronomique, 97310 Kourou, French Guiana

*Correspondence: Sylvain Schmitt, UMR EcoFoG, Campus Agronomique, 97310 Kourou, French Guiana. Email: sylvain.m.schmitt@gmail.com. ORCID: 0000-0001-7759-7106

Abstract

Mutation, the source of genetic diversity, is the raw material of evolution; however, it remains an understudied process in plants. Using simulations, we demonstrate that generic variant callers, commonly used to detect mutations in plants, are outperformed by methods developed for cancer research. Reanalysis of published data identified up to 7x more somatic mutations than initially reported, advocating the use of cancer research callers to boost mutation research in plants.

Main text

DNA sequence mutation is the raw material for evolutionary change, but, despite its crucial role, many fundamental questions around the mutation process are still open. Mutation rates are highly variable along the tree of life, differing by several orders of magnitude among species and kingdoms, and are considered as an evolvable trait *per se*¹. Mutations are assumed to be random, but the rate at which different nucleotides mutate strongly depends on the surrounding nucleotides², hereafter referred to as a mutation spectrum. The mutation spectra themselves are now believed to evolve over time³, even at relatively short evolutionary timescales⁴. The drivers of new mutations, previously thought to be simply due to DNA replication errors, are now also debated⁵.

Unlike most animals that transmit to the next generation only mutations present in their germ cells (sperm and eggs), plants are expected to produce heritable somatic mutations as they grow throughout their lives, departing from the so-called Weismann's germ plasm theory^{6,7}. As a consequence, long-lived species, such as trees, are generally assumed to accumulate more heritable mutations than herbs per generation⁸. To generate new knowledge on plant mutation processes⁹, several studies examined within-individual variation in long-lived trees, whose individuals can live for more than a thousand years¹⁰. Two studies used the pedunculate oak (*Quercus robur*), a long-lived European tree species, as a plant model to identify somatic mutations. Schmid-Siegert et al.¹¹ identified 17 mutations by comparing sequencing data from two branches of a 234-year-old individual. The authors therefore argued that their results are consistent with a low mutation rate in

pedunculate oak. Plomion et al.¹² identified 46 mutations using three branches of a younger (century-old) individual, which is an almost 10-fold higher rate after taking the tree age difference into account. A shared limitation of both studies is that the authors did not identify the detection method best suited for the features of somatic mutations.

The development of tools to detect mutations in humans is rapidly expanding in cancer research^{13,14}. Detecting mutations in cancers is conceptually similar to detecting somatic mutations in plants, *i.e.*, the aim is to detect mutations that potentially affect only a small fraction of the sequenced tissue. This specific challenge is poorly addressed in plants, where mutation detection remains based on generic variant callers, which were initially designed to detect heterozygous sites^{11-12,15-17}. Transferring mutation detection tools from cancers to plants requires evaluating their performance in a plant research context. Cancer research frequently uses very high sequencing depths (100X - 1000X), while the depth available for plants is often considerably lower (e.g., 34X¹⁶, 40X¹⁸ or 70X¹¹), bar a few exceptions (240X¹⁷, 250X¹² or 1000X¹⁵). To improve the detection of mutations for basic and applied plant research, a deep evaluation of the performance of variant callers is needed in relation to the biological features and quality of data typical of plant studies.

We performed both an *in silico* and an empirical data-based evaluation of the performance of variant callers to detect somatic mutations, using two large published datasets on the same species (pedunculate oak, *Quercus robur*) that applied different strategies for sequencing depth and mutation detection^{11,12} (see Fig. S1). We particularly explored the role of the sequencing depth and allelic frequency of the somatic mutation in plant tissues

to answer the following questions: (1) Can cancer-derived methods, both in terms of protocols (i.e. sequencing depth) and tools (i.e. callers), improve the detection of somatic mutations in plants?; and (2) Can reanalyses of within-individual sequencing data provide new insights regarding plant mutation processes?

To compare the performance of different variant callers to detect mutations, we simulated sequencing data containing new mutations at a given allele frequency (fraction of simulated reads per genomic position carrying the mutation), and using varying depths of sequencing (for variable transition/transversion ratios, see Supplementary Note S1). We then evaluated the performance of variant callers as a function of allele frequency and sequencing depth. We found marked differences in: (1) the recall, the ability to recover the simulated mutations; and (2) the precision, the proportion of truly simulated mutations among all variants detected. For allele frequencies equal to, or lower than, 0.25, cancer-specific variant callers (*Strelka2*, *Mutect2*, *MuSE*, but not *Somatic Sniper*) outperform generic variant callers such as *GATK*, *freebayes*, and *VarScan* (Fig. 1). For allele frequencies over 0.25, all variant callers perform similarly well, except for *freebayes*, which identified many false positives. Over the 80 tested parameter combinations, *Strelka2* was the best performing variant caller for various allelic frequencies and sequencing depths (57/80 with an average recall of 0.95 for a precision of 0.98, Fig. S2-4 and Table S1 and S2).

Fig. 1 | Variant caller performances to identify simulated mutations for varying allelic frequencies and sequencing depths (see Fig. S5 for all parameter combinations). The recall is the ability to detect (recover) the simulated mutations. The precision is the proportion of simulated mutations among all variants detected (i.e. including false positives). Each point represents the averaged mutation recall or precision (10 simulations) for increasing allelic frequency and sequencing depth. The shaded area represents the 90th quantiles of recall and precision rates over the 10 replicates computed for all callers, but only visible for the precision of *Muse*, *Mutect2*, and *VarScan*.

We further investigated the performance of the best performing variant caller, *Strelka2*, on two empirical datasets on pedunculate oak^{11,12} in comparison to the variant callers used in the original publications, *i.e.*, *GATK* and *Mutect2*, respectively (see Supplementary Note S2). Mapping the raw data of Schmid-Siebert et al.¹¹ and Plomion et al.¹² on the oak genome that we used as a mapping reference for our empirical study, we successfully mapped 14 and 60 of the mutations detected in the original articles, respectively. Across variant callers, we recovered 12 (86%) and 60 (100%) of these original mutations in our total list of candidate somatic mutations (Fig. 2A), strongly supporting the results shown by the two previous studies. However, our analyses were able to detect far more candidate mutations than initially reported. Applying filtering based on sequencing depth and mutated allele copies (see Supplementary Note S2), *Strelka2* produced a more reliable set of candidate mutations than *GATK* but similar to *Mutect2*, with an estimated number of mutation candidates 10- to 25-fold higher than that of the original studies (Fig. 2A). Adding *Strelka2* automatic filtering yielded the most robust and conservative dataset with a similar number of mutations between both studies and a 2- to 7-fold increase compared to the original number of mutations. The distribution of allelic frequencies of detected mutations partly explains differences among detection methods (Fig. 2A), with *Strelka2* and *Mutect2* detecting mutations with lower allelic frequencies than the candidate mutations presented in the original publications, especially for the Plomion et al.¹² study that used higher sequencing depths.

Fig. 2 | Candidate mutation spectra depending on variant callers and filtering in Schmid-Siebert *et al.*¹¹ and Plomion *et al.*¹². **A.** Allelic frequency distribution for every dataset, including the candidate mutations from the original article present in the raw data from the reanalysis (red), the results of *GATK* with Best Practices (blue), *Mutect2* (green), and *Strelka2* after filtering (purple), and the results of *Strelka2* using the automatic filtering named robust (orange). The labels indicate the number of candidate mutations in each dataset. Per caller comparisons are available in Fig. S6. **B.** Annotation of the mutations detected by *Strelka2* across chromosomes using the automatic filtering named robust for Schmid-Siebert *et al.*¹¹ (green) and Plomion *et al.*¹² (orange) compared to the genomic expectation (grey, see Supplementary Note S3). Error bars represent the standard deviation of the observed percentages across chromosomes, and the annotation above the columns

indicates the significance of the Student's t-test comparing the mean percentage of mutations to the mean genomic expectation, with *ns*, **, and *** corresponding to non-significant, $p < 0.01$, and $p < 0.001$ differences, respectively. C. Context-dependent mutation spectra depending on mutation types for the results of *Strelka2* using the automatic filtering named robust. Mutation types have been summarised into six main classes with thicker lines for transversion compared to transition, and then differentiated depending on their 5' and 3' genomic contexts, see Fig. S7-8. Pearson's correlation r measures the correlation of the mutation spectra between Schmid-Siebert *et al.*¹¹ and Plomion *et al.*¹².

Based on the set of robust mutations detected by *Strelka2*, we then explored annotations and mutation spectra in both datasets (Fig. 2B-C), which have rarely been explored in model plant species (but see first evidence based on mutation accumulation lines in *Arabidopsis*¹⁹) and never in the wild. The proportions of mutations found in different genomic regions (e.g. genic, intergenic) were highly correlated between both original studies and proportional to the representation of the genomic regions, supporting a random distribution of mutations throughout the genome (Fig. 2B). Mutation spectra of the two studies are significantly correlated (Pearson's $r = 0.49$, $p < 7.4 \times 10^{-5}$), with an enrichment in C>T transitions, particularly in some specific genomic contexts (Fig. 2C).

In conclusion, mutation research in plants still primarily uses generic variant callers and methodologies that are not developed for the specificity and complexity of mutation detection; however, plant research could benefit from the development of tools and protocols initially designed for human cancer research, which is a rapidly expanding field¹³.

Here, we found marked differences in the performance of variant callers for mutation detection based on sequencing depth and allelic frequency. We found that cancer variant callers performed better than generic variant callers for mutation detection at low or intermediate allelic frequency or with low sequencing depth, and similarly well for high allelic frequency. Therefore, plant mutation studies should make greater use of cancer variant callers such as *Strelka2* rather than generic variant callers such as *GATK* to detect mutations, especially for low frequency mutations and when using low sequencing depth. By reanalyzing the raw oak data^{11,12}, we found that the marked differences in the performance of variant callers account for the discrepancies in genome-wide plant somatic mutation rate estimates. Our reanalysis shows robust evidence for an up to 7-fold higher number of mutations than previously reported, a value closer to the expectations based on the theory^{9,20}. We argue that knowledge and methodological transfers from cancer to plant mutation detection are expected to contribute strongly to the upward trend of this field and to reconcile empirical reports with theoretical expectations.

Material and Methods

Study design. We developed two workflows: 1) to generate Illumina-like sequencing reads including mutations with varying biological and sequencing parameters; and 2) to detect mutations with multiple variant callers (Fig. S1). We used both *singularity* containers²¹ and the *snakemake* workflow engines²² to build automated, highly reproducible (FAIR), and scalable workflows. We then used both workflows to test the best performing variant caller for mutation detection *in silico* based on biological and sequencing parameters. The

workflows are publicly available at <https://github.com/sylvainschmitt>. We finally used the identified variant caller to detect mutations in pedunculate oak, *Quercus robur* L., by re-analysing data from two somatic mutation projects on oaks led by INRA Bordeaux, France¹² and the University of Lausanne, Switzerland¹¹.

Generation of mutations. The first workflow named *generateMutations* takes advantage of the two scripts included in *simuG*²³, *vcf2model.pl*, and *simuG.pl*, respectively, to build a model of heterozygous sites distribution in the haploid reference genome based on a user-defined base of known heterozygous sites in *vcf* format and to build the second reference haploid genome comprising a user-defined number of heterozygous sites to accurately represent diploidy. Typically, the user can define a number of heterozygous sites based on the product of nucleotide diversity (π) and genome length (L). The workflow uses a homemade R script named *generate_mutations* to spike randomly the reference genome with a user-defined number of mutations which are drawn in a binomial distribution using a user-defined transition/transversion ratio (**R**). Finally, the workflow takes advantage of *InSilicoSeq*²⁴ defined with the model option *hiseq* to generate datasets of mutated and non-mutated *in silico* Illumina-like sequencing reads using (1) the original reference haploid genome, (2) the reference haploid genome with heterozygous sites (pedunculate oak is a diploid species), and (3) the reference genome spiked with mutations following user-defined allelic fraction (**AF**) and depth of sequencing depth (**C**).

Detection of mutations. The second workflow named *detectMutations* aims to detect somatic mutations from mapped sequencing reads on a genome reference. Pair-end

sequencing reads of every library are quality checked using *FastQC* before trimming using *Trimmomatic*²⁵ keeping only pair-end reads without adaptors and a phred score above 15 in a sliding window of 4 bases. Reads are aligned against the reference per chromosome using BWA *mem* with the option to mark shorter splits²⁶. Alignments are then compressed using *Samtools view* in CRAM format, sorted by coordinates using *Samtools sort*, and indexed using *Samtools index*²⁷. Duplicated reads in alignments are marked using *GATK MarkDuplicates*²⁸. Finally, the workflow uses seven variant callers to detect mutations, including generic variant callers to detect variants and dedicated variant callers for mutation detection. Generic variant callers to detect variants include the combination of *GATK HaplotypeCaller* and *GATK GenotypeGVCFs*²⁸ and *freebayes*²⁹ using and reporting genotype qualities, without priors on allele balance, with a minimum alternate fraction of 0.03, a minimum repeated entropy of 1 and a minimum alternate count of 2. Cancer variant callers developed for mutation detection include *VarScan*³⁰, *Strelka2*¹³, *MuSE*³¹, *Mutect2*³² (using a panel of normal and without soft clipped bases), and *Somatic Sniper*³³ (filtering reads with mapping quality less than 25, filtering mutations with quality less than 15 with prior probability of a mutation of 0.0001). Then we only focused on the simulated mutations, and therefore excluded from the analyses the known heterozygous sites provided by the user thanks to the vcf file for *GATK*, *freebayes*, *Somatic Sniper*, and *Strelka2* using *BEDTools subtract*³⁴ or directly within the variant caller for *Mutect2* and *VarScan*.

***In silico* experiment.** We used the *generateMutations* workflow to generate mutations in the oak genome with varying biological and sequencing parameters. To ensure consistency between the *in silico* experiment and the reanalyses, we used the reference genome

"Qrob_PM1N" of *Quercus robur* 3P from Bordeaux¹², thus assessing the behavior of variant callers in the same genomic context. To reduce the computational load, we only generated mutations on the first megabase of the first chromosome of the oak assembly ("Qrob_Chr01") in order to later focus the detection on this region. We used known heterozygous sites from the reference genome¹² to simulate back one thousand heterozygous sites ($\pi = 0.01$, $L = 1$ Mb , $N = \pi \times L = 10^4$). We used varying values of transition/transversion ratio ($\mathbf{R} = [2, 2.5, 3]$), allelic fraction ($\mathbf{AF} = [0.05, 0.1, 0.25, 0.5]$), and sequencing depth ($\mathbf{C} = [25, 50, 100, 150, 200]$), resulting in 60 simulated datasets of mutated and associated base reads ($3\mathbf{R} \times 4\mathbf{AF} \times 5\mathbf{C}$). We then used the *detectMutations* workflow to detect back spiked mutations with every variant caller (*Mutect2*, *freebayes*, *GATK*, *Strelka2*, *VarScan*, *Somatic Sniper*, and *MuSe*). Using known spiked mutations, we assessed the number of true positive (TP), false positive (FP), and false negative (FN) for each variant caller to detect mutations and each combination of biological and sequencing parameters. We used the resulting confusion matrix to calculate the recall ($\frac{TP}{TP+FN}$) and the precision rates ($\frac{TP}{TP+FP}$). The recall rate represents the ability of the variant caller to detect all mutations, while the precision rate represents the ability of the variant caller to not confound other sites with mutations. We finally assessed each variant caller to detect mutations using the recall and the precision rates with varying transition/transversion ratio ($\mathbf{R}$), allelic fraction ($\mathbf{AF}$), and sequencing depth ($\mathbf{C}$) to identify the best performing variant caller based on biological and sequencing parameters.

Oak data reanalyses. We re-analyzed publicly available oak data from two projects led by Bordeaux, France¹² and Lausanne, Switzerland¹¹ (SRA PRJNA327502 and ENA PRJEB8388,

respectively). We then used the best-performing variant caller based on our *in silico* investigation, *Strelka2*, and the variant caller for mutation detection from the original publication to compare the results *i.e.* *GATK* with *Best Practices* for Swiss data¹¹ and *Mutect2* for French data¹². The Swiss data comprised 2 libraries of medium sequencing depth (60X) representing one lower and one upper branch. The French data comprised 3 libraries of high sequencing depth (160X) representing 3 branches (lower, mid, and upper). For both Swiss and French data, we compared each pair of sample points sequentially as the normal library and the mutated library to decipher mutations among branches from heterozygous sites and sequencing errors. For the French data, we further filtered out candidate somatic mutations by using a cross-validation procedure to keep a coherent temporal pattern among mutations following the original publication¹². We used these raw datasets to identify the mutations from the original studies after realigning them on the 3P genome using *BLAT*³⁵. For both datasets, we finally kept candidate mutations with (1) a read depth for both the normal and mutated samples between half and two times the mean sequencing depth (30-120X and 80-320X for Swiss and French datasets, respectively), (2) an absence of the mutated allele in the normal sample, (3) a minimum of 10 copies of the mutated allele in the mutated sample and (4) an allelic frequency <0.5. In addition, we took advantage of the Empirical Somatic Score (ESS) calculated by *Strelka2*¹³ to build a robust dataset of candidate mutations for both datasets. Finally, we measured the proportion of the genome following the sequencing depth requirement, *i.e.* half and two times the mean sequencing depth, for both studies to estimate the total number of mutations from the observed number of candidates.

Acknowledgements

This study was funded through an Investissement d’Avenir grant of the ANR: CEBA (ANR-10-LABEX-0025).

Authors’ contributions

All authors conceived the ideas; SS developed the pipelines, conducted the virtual experiment and the data reanalyses; SS and TL analyzed outputs and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data Accessibility

generateMutations and detectMutations pipelines are available on GitHub (<https://github.com/sylvainschmitt/generateMutations> and <https://github.com/sylvainschmitt/detectMutations>). Reanalyzed reads and corresponding genomes were extracted from GenBank under accession BioProject PRJNA327502 and from European Nucleotide Archive under project accession code PRJEB8388.

References

Main text

1. Lynch, M., Ackerman, M. S., Gout, J. F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11), 704–714. <https://doi.org/10.1038/nrg.2016.104>
2. Martincorena, I., & Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, 349(6255), 1483–1489. <https://doi.org/10.1126/science.aab4082>
3. Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., & Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. *Nature Communications*, 8(May), 1–8. <https://doi.org/10.1038/ncomms15183>
4. Harris, K., & Pritchard, J. K. (2017). Rapid evolution of the human mutation spectrum. *eLife*, 6, 1–17. <https://doi.org/10.7554/eLife.24284>
5. Gao, Z., Moorjani, P., Sasani, T. A., Pedersen, B. S., Quinlan, A. R., Jorde, L. B., ... Przeworski, M. (2019). Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 116(19), 9491–9500. <https://doi.org/10.1073/pnas.1901259116>
6. Weismann, A. (1893). *The germ-plasm: a theory of heredity*. Scribner's.
7. Lanfear, R. (2018). Do plants have a segregated germline? *PLoS Biology*, 16(5), 1–13. <https://doi.org/10.1371/journal.pbio.2005439>
8. Hanlon, V. C. T., Otto, S. P., & Aitken, S. N. (2019). Somatic mutations substantially increase the per-generation mutation rate in the conifer *Picea sitchensis*. *Evolution Letters*, 3(4), 348–358. <https://doi.org/10.1002/evl3.121>

9. Schoen, D. J., & Schultz, S. T. (2019). Somatic Mutation and Evolution in Plants.
Annual Review of Ecology, Evolution, and Systematics, 50, 49–73.
<https://doi.org/10.1146/annurev-ecolsys-110218-024955>
10. Schöngart, J., Bräuning, A., Barbosa, A. C. M. C., Lisi, C. S., & de Oliveira, J. M. (2017).
Dendroecology Tree-Ring Analyses Applied to Ecological Studies. Springer Nature.
11. Schmid-Siebert, E., Sarkar, N., Iseli, C., Calderon, S., Gouhier-Darimont, C., Chrast, J., ...
Reymond, P. (2017). Low number of fixed somatic mutations in a long-lived oak tree.
Nature Plants, 3(12), 926–929. <https://doi.org/10.1038/s41477-017-0066-9>
12. Plomion, C., Aury, J. M., Amselem, J., Leroy, T., Murat, F., Duplessis, S., ... Salse, J.
(2018). Oak genome reveals facets of long lifespan. *Nature Plants*, 4(7), 440–452.
<https://doi.org/10.1038/s41477-018-0172-3>
13. Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., ... Saunders,
C. T. (2018). Strelka2: fast and accurate calling of germline and somatic variants.
Nature Methods, 15(8), 591–594. <https://doi.org/10.1038/s41592-018-0051-x>
14. Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., ... Gut, I. G.
(2015). A comprehensive assessment of somatic mutation detection in cancer using
whole-genome sequencing. *Nature Communications*, 6.
<https://doi.org/10.1038/ncomms10001>
15. Watson, J. M., Platzer, A., Kazda, A., Akimcheva, S., Valuchova, S., Nizhynska, V., ...
Riha, K. (2016). Germline replications and somatic mutation accumulation are
independent of vegetative life span in Arabidopsis. *Proceedings of the National
Academy of Sciences of the United States of America*, 113(43), 12226–12231.
<https://doi.org/10.1073/pnas.1609686113>

16. Hanlon, V. C. T., Otto, S. P., & Aitken, S. N. (2019). Somatic mutations substantially increase the per-generation mutation rate in the conifer *Picea sitchensis*. *Evolution Letters*, 3(4), 348–358. <https://doi.org/10.1002/evl3.121>
17. Orr, A. J., Padovan, A., Kainer, D., Külheim, C., Bromham, L., Bustos-Segura, C., ... Lanfear, R. (2019). A phylogenomic approach reveals a low somatic mutation rate in a long-lived plant. *bioRxiv*. <https://doi.org/10.1101/727982>
18. Wang, L., Ji, Y., Hu, Y., Hu, H., Jia, X., Jiang, M., ... Tian, D. (2019). The architecture of intra-organism mutation rate variation in plants. *PLoS Biology*, 17(4), 1–29. <https://doi.org/10.1371/journal.pbio.3000191>
19. Weng, M. L., Becker, C., Hildebrandt, J., Neumann, M., Rutter, M. T., Shaw, R. G., ... Fenster, C. B. (2019). Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. *Genetics*, 211(2), 703–714. <https://doi.org/10.1534/genetics.118.301721>
20. Burian, A. (2021). Does Shoot Apical Meristem Function as the Germline in Safeguarding Against Excess of Mutations? *Frontiers in Plant Science*, 12(August), 1–9. <https://doi.org/10.3389/fpls.2021.707740>
21. Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLoS ONE*, 12(5), 1–20. <https://doi.org/10.1371/journal.pone.0177459>
22. Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19), 2520–2522. <https://doi.org/10.1093/bioinformatics/bts480>

23. Yue, J. X., & Liti, G. (2019). SimuG: A general-purpose genome simulator. *Bioinformatics*, 35(21), 4442–4444. <https://doi.org/10.1093/bioinformatics/btz424>
24. Gourel, H., Karlsson-Lindsjö, O., Hayer, J., & Bongcam-Rudloff, E. (2019). Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*, 35(3), 521–522. <https://doi.org/10.1093/bioinformatics/bty630>
25. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
26. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
27. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
28. Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., ... DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, 43(1), 483–492. <https://doi.org/10.1002/0471250953.bi1110s43>
29. Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing, 1–9. Retrieved from <http://arxiv.org/abs/1207.3907>
30. Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., ... Ding, L. (2009). VarScan: Variant detection in massively parallel sequencing of individual

and pooled samples. *Bioinformatics*, 25(17), 2283–2285.

<https://doi.org/10.1093/bioinformatics/btp373>

31. Fan, Y., Xi, L., Hughes, D. S. T., Zhang, J., Zhang, J., Futreal, P. A., ... Wang, W. (2016).

MuSE: accounting for tumor heterogeneity using a sample-specific error model

improves sensitivity and specificity in mutation calling from sequencing data.

Genome Biology, 17(1), 178. <https://doi.org/10.1186/s13059-016-1029-6>

32. Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., & Lichtenstein, L. (2019).

Calling Somatic SNVs and Indels with Mutect2. <https://doi.org/10.1101/861054>

33. Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., ... Ding,

L. (2012). Somaticsniper: Identification of somatic point mutations in whole genome

sequencing data. *Bioinformatics*, 28(3), 311–317.

<https://doi.org/10.1093/bioinformatics/btr665>

34. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for

comparing genomic features. *Bioinformatics*, 26(6), 841–842.

<https://doi.org/10.1093/bioinformatics/btq033>

35. Kent, W. J. (2002). BLAT---The BLAST-Like Alignment Tool. *Genome Research*, 12(4),

656–664. <https://doi.org/10.1101/gr.229202>

36.