



Leveraging large family analyses for more accurate de novo mutation detection

Maher Alnajjar¹ · Endre Barta^{1,2}

Received: 30 August 2025 / Accepted: 23 February 2026
© The Author(s) 2026

Abstract

De novo mutations (DNMs), which arise in the offspring and are absent in the parents, are increasingly studied in farm animals with the advent of whole-genome sequencing (WGS). Variant calling after genome sequencing is a crucial step in modern genomics, and its accuracy directly influences subsequent genetic analyses, which are vital not only in breeding and human healthcare but also in functional genomic research. Yet, using only families with trios neglects the important shared information in the family and leads to an inaccurate determination of DNMs. Here, we show that using inheritance-based Whole Genome Sequencing (WGS) data analysis on a larger family is an effective way to identify DNMs in offspring accurately, and we present the first such study in rabbits.

Keywords De novo mutations · Whole genome sequencing · Family-based analysis · Rabbit

Introduction

Apart from inheriting half of the genome from each parent, individuals are also born with a small number of novel variants (Keightley et al. 2014). Germline variants or de novo mutations (DNMs) influence the rate of genome evolution and serve as a key parameter in the evolutionary process. In other words, DNMs are a source of genomic innovation, as well as inherited traits and diseases. Therefore, having precise knowledge and understanding of spontaneous mutations (also called DNMs) is essential for gaining advanced insights into important questions about diversity and evolution (Acuna-Hidalgo et al. 2016; Lynch et al. 2016; Sakumi 2019).

The advancement of next-generation sequencing provides a new tool for directly identifying germline mutations by comparing sequences across generations (Scally and Durbin 2012). Several mechanisms have been reported as causes of DNM induction in the human germline. For example, non-allelic recombination (Stankiewicz and Lupski 2002), replication infidelity (Arana and Kunkel 2010), and genomic damage (Friedberg 2003). Consequently, there is significant interest in understanding the frequency of (DNMs) occurrence in humans (Kohailan et al. 2022; Werling et al. 2018) as well as in domesticated animals (Azevedo et al. 2024). A recent study reported the germline mutation rate in various vertebrates (Bergeron et al. 2023). The authors of the study used 68 species to discover DNMs and determine the mutation rate. However, all the used species were trios (parents and one offspring), and rabbits were not included in the study.

Today, the most widely used methods for downstream variant analysis rely heavily on variant calling based on short-read sequencing (Seah et al. 2023). Although the average depth of affordable genome sequencing is increasing over time, significant challenges remain in variant calling accuracy. Copy number variations, repetitive elements, and low-complexity regions also cause poorly mapped reads (Guan and Sung 2016) which in turn affect variant calling. Therefore, in special cases, such as detecting DNMs, a high level of fidelity is crucial (Wagner et al. 2022).

✉ Endre Barta
barta.endre@uni-mate.hu

Maher Alnajjar
alnajjar.maher@uni-mate.hu

¹ Department of Genetics and Genomics, Institute of Genetics and Biotechnology, Hungarian University of Agriculture and Life Sciences, Szent-Györgyi A. u. 4, Gödöllő 2100, Hungary

² Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen, Egyetem tér 1, Debrecen 4032, Hungary

In this study, we demonstrate the impact of using a larger nuclear family on the accuracy of detecting DNMs over the trio approach. We demonstrate that increasing the number of utilized offspring in this experiment results in a significant reduction in the number of detected false positive DNMs. We used a rabbit family consisting of the parents (the mother is a meat production breed, and the father is a dwarf) with their eight offspring to investigate and characterize DNMs. Finally, we provided Discover_DNMs.py tool to detect DNMs in a family with any size.

Materials and methods

Samples and whole genome sequencing (WGS)

Samples were obtained from a rabbit (*Oryctolagus cuniculus*) family. The family consists of two divergent breeding parents. The mother belonged to the Hycote XXL line, selected for the meat production commercial line due to large body weight over more than 20 generations, and the father is a Thuringian breed (1.5 to 2 kg), and their eight offspring. The parents were reported in our previous study (Fekete et al. 2025). DNA extraction was performed at the Institute of Genetics and Biotechnology by the Applied Wildlife and Farm Animal Genomics Group. Whole-genome sequencing (WGS) data are available on the NCBI/SRA platform, BioProject ID: PRJNA1237607. The material and methods were detailed in our previous work (Alnajjar et al. 2025). In short, WGS was performed on the family with approximately 30x coverage. After quality control steps, reads were mapped to the available OryCun3.0 reference genome sequence under the NIH BioProject (PRJNA635554). Variants were called using GATK HaplotypeCaller (gatk-package-4.4.0.0-local.jar) (DePristo et al. 2011; Van der Auwera et al. 2013), and the variant quality recalibration (VQSR) was not used, following the best practices to preserve the de novo mutations. Instead, a stringent hard filtration set was applied that reduces the false discovery rate and is most likely to keep only true variants. The variant must have a genotype quality (GQ) of at least 30. The result is a merged VCF file containing variants for the entire family. The complete pipeline for alignment and variant calling from WGS (rabbit_tfbs_pipeline21.sh) is available on the following repository (<https://github.com/Maher199/ASE-in-a-family/>).

De novo mutations (DNMs) discovery

In order to accurately discover the novel variants in the offspring, a few stringent criteria should be applied. The following criteria were set in a python script (Discover_DNMs.py) (https://github.com/Maher199/Discover_DNM

s) and applied on the merged VCF file obtained from the previous step:

- A DNM should be a heterozygous locus in the given offspring.
- The parents should be homozygous for the reference.
- The variant should not be present in the parents, i.e., no reads supporting the alternative allele should be presented.
- The variant should be present at the given offspring only, while the rest of the offspring should be homozygous for the reference. One read supporting the alternative allele is allowed, allowing for potential sequencing errors.
- The default minimum coverage threshold used is 15, and the maximum is 45 to limit the false variant calls due to inefficient reads. This parameter can be set by the user.
- $PL > 450$. PL parameter in the VCF file is the normalized likelihoods of the possible genotypes.
- At least 50% of the minimum depth (7–8) reads should support the alternative allele at the DNM.

Filtering DNMs

To ensure more accuracy and exclude repetitive regions from the analysis, RepeatMasker version 4.1.5 (Smit, 2013) was used to identify and mask the repetitive elements that carry the risk of false positively reported DNMs. Then, the DNMs discovery process was carried out using the Python script.

To compare the results of our Python script with those of other software, Bcftools +trio-dnm2 version 1.22 (Danecek et al. 2021) was used, which only considers trios. After running Bcftools on the VCF file that includes variants after the repeat masking step, the results were filtered to keep only the high-quality DNM reported by this +trio-dnm2. The following criteria were used: $FORMAT/DNM=0$ && $FORMAT/VA=0$ && $FORMAT/VAF > 20$.

Results

De novo mutations: accurate discovery and filtration

The experiment setup consists of two genetically divergent breeding parents (the mother (Hycote) as a meat-producing animal and the father, a dwarf pet) with their eight offspring. This experimental setup with Whole Genome Sequencing (WGS) for each individual helped us to investigate the de novo mutation (DNM) events. The meat-producing mother's DNA sequence had 9.5 million heterozygous (HET) loci, while the dwarf father's sequence had 10.4 million HET variations after mapping to the OryCun3.0 reference

genome (Alnajjar et al. 2025). Also, there are about 13 million sites where one of the parents is HET while the other is homozygous. About 2.9 million sites were also found where both parents are HET.

To discover DNMs in the offspring, the variant should only be present in the given offspring and not in the parents. However, due to the possible sequencing errors and the erroneous alignment of the sequencing reads, false positive DNM identifications can happen, where the same new variation exists in more than one sibling. Therefore, we excluded those cases where the same variant was also present in other siblings. Based on this idea, we set the algorithm as restricted criteria in a Python script (https://github.com/Maher199/Discover_DNMs) to demonstrate the power of using a larger family in detecting accurate DNMs that comply with these criteria. The script can be executed on a trio family, i.e., the parents and only one offspring, as well as a larger family with several offspring. The script also classifies the DNMs into Single Nucleotide Variants (SNVs) and Insertions and Deletions (INDELs). The script is also adaptable with its parameters to accommodate different read coverages. Based on that, we created a combination of all possible trios and gradually added siblings, then ran the script on the family with all combinations. Figure 1 illustrates the decline in the number of DNMs reported after adding more siblings, keeping only the most accurate set of

DNMs, especially after the addition of the 3rd sibling. An immediate drop can be noticed when at least one sibling is available. Initially, a high number of DNMs was reported when each offspring was treated as the only offspring (trio). On average, the number of DNMs was about 3651 at each offspring, and then gradually adding more siblings was accompanied by a decrease in the false positive reported variants until it became about 170 at each offspring.

Obtaining the accurate DNMs based on our script, which leveraged the entire set of siblings, the script was executed after removing the repetitive regions. Repetitive regions, such as duplications and transposons, can cause misalignments, particularly in short-read technologies, leading to the reporting of false-positive DNMs (Wagner et al. 2022). Therefore, we masked these regions and excluded the DNMs identified within them. Table 1 lists the number of potential DNMs discovered using Bcftools +trio-dnm2 and our algorithm (both as a trio and the entire set of siblings). On average, the number of remaining accurate DNMs is approximately 136 variants per offspring after applying the entire family. The latter number is divided into SNVs and INDELs, as shown in the table. In the Bcftools run, we only kept the high-quality DNM reported. The DNMs were filtered based on the DNM field, and the VA were only filtered, with the ALT allele kept, and the quality VAF > 20% of the alleles supporting the DNM allele. For the entire list of

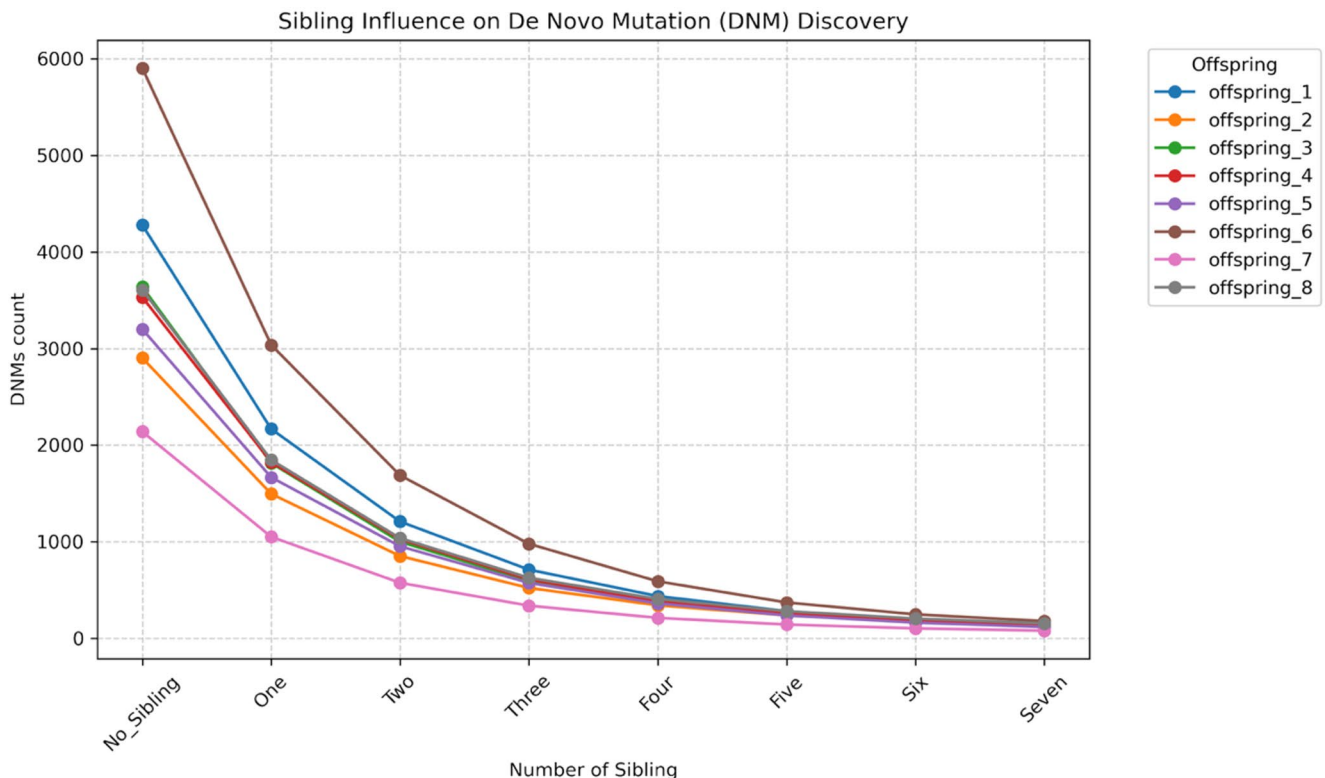


Fig. 1 The impact of a larger family size on the number of reported DNMs. The X-axis indicates the number of siblings, starting with no siblings (i.e., a trio family), while the Y-axis refers to the putative DNMs reported after repeat masking step

Table 1 Number of DNMs at each offspring before and after Repeat Masking, and after running Bcftools +trio-dnm2, the algorithm applied to each offspring as a trio, and then to the entire family. “Full family” column indicates the number of DNMs that remained after the addition of all siblings. The final number of DNMs was classified as SNVs and INDELs

Offspring	No siblings	No sibling after repeat masking	Bcftools +trio-dnm2	Full family	SNVs	INDELs
Offspring_1	5408	4280	2033	140	51	89
Offspring_2	3605	2903	1662	132	53	79
Offspring_3	4541	3638	1877	144	51	93
Offspring_4	4412	3533	1921	137	57	80
Offspring_5	3988	3201	1749	119	44	75
Offspring_6	7945	5904	3362	179	77	102
Offspring_7	2656	2140	1496	81	41	40
Offspring_8	4534	3608	2004	156	88	67

Fig. 2 The Density Distribution of DNMs along the chromosomes. A: The entire set of DNMs at all offspring along all the set of chromosomes. The plot shows the density of DNMs within a 2 Mbp window. Created Using KaryoploteR (Gel and Serra 2017)



the reported DNMs sorted by offspring and the type of the mutation, see supplementary Table S1.

We have also benchmarked the Genotype Quality (GQ) used in this study, and therefore, we used $GQ \geq 30$. We did not find differences after using the $GQ > 60$ criterion presented in the study by Bergeron et al. (2021). We repeated our experiment using $GQ = 30, 40, 50$, and 60 . The difference was negligible (Supplementary file S2). By contrast, adding only one sibling can significantly reduce the number of false positives identified as DNM, as Fig. 1 demonstrates.

De novo mutation hotspots

To further investigate the potential clustering of the final DNMs, we conducted a genomic localization analysis. Hotspots are regions in chromosomes that are prone to

having an increased number of mutations in a small chromosomal area. After filtering out the mutations that fell in the masked region, we characterized the DNMs in each chromosome. Figure 2 shows these hotspots, where these regions tend to produce more mutations than other regions on the same chromosome. We found that these variants are often clustered in hotspots, with most located near the ends of chromosomes.

Analysis of de novo mutations: base substitutions

The final step in the analysis involved investigating the distribution of substitution types across the SNVs. Base substitution in the SNVs, such as (C=>T or its complementary G=>A) and so on, is usually classified into two groups. Transition (Ts), which includes (purine=>purine

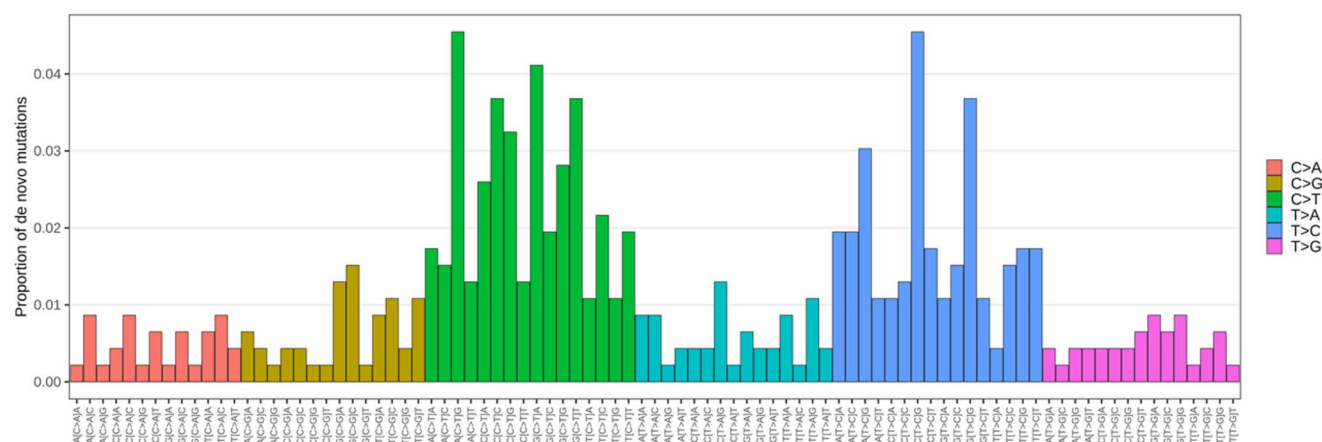


Fig. 3 Distribution of SNV Base Substitutions: Transition vs. Transversion. High-resolution mutational spectrum of de novo mutations. The plot shows the contribution of individual mutation contexts,

or pyrimidine=>pyrimidine), and Transversion (Tv), including (purine=>pyrimidine or pyrimidine=>purine). Figure 3 illustrates the counts of different types of base substitutions for the identified DNMs, providing a higher-resolution mutational spectrum. We estimated the frequency of mutations in each of the 96 trinucleotide contexts, defined by the mutated base and its flanking bases on both sides. We also found, as expected, that the most frequent type of base substitution was transition, while the least frequent was transversion, with a ratio (Ts/Tv) (315/147) of 2.14.

Discussion

In this study, we demonstrate the power of utilizing a larger family for identifying de novo mutations (DNMs) from Whole-Genome Sequencing (WGS) data. Whole-genome sequencing allows mapping the reads into the reference genome and thus identifying variations genome-wide. In special cases, such as de novo mutation (DNM) detection, a high level of fidelity is crucial, and a minimum of a trio (comprising two parents and an offspring) approach must be used (Wagner et al. 2022). We developed a script (Discover_DNMs.py) that incorporates a set of stringent criteria to preserve accurate variants for downstream analysis when the data from other siblings is also available. The script offers flexible options and can be run on various family sizes, ranging from a trio to a larger family, with a relatively decent running time. The script can serve as a complementary tool in DNM studies. To the best of our knowledge, rabbits have not been used in DNMs studies. The script was run on a large rabbit family consisting of the parents and eight offspring. The results show a significant decrease in the number of DNMs discovered when more than one offspring is added (Fig. 1 and Table 1). The results of running

grouped by the six possible point mutation types as the colors show. A small number of contexts dominate the signal, with particularly higher peaks among C: G>T:A and T: A>C:G transitions

the widely used Bcftools program on trios also show that using only trio families can lead to false-positive DNMs. Our approach is not intended to replace other methodologies using strict filtering criteria for trios, such as (Bergeron et al. 2023), but rather to serve as a complementary solution: improving the specificity of DNM when a larger family, or even one sibling, is available. Moreover, they used very stringent filtering at the beginning of the analysis, $GQ > 60$ (significantly higher than most pipelines that tolerate lower GQ thresholds), which can restrict their data to the DNM analysis. We have compared our results with this approach, and we found a negligible difference after using the $GQ > 60$ criterion. However, adding a single sibling can significantly reduce the number of false positives among detected DNMs. Additionally, Bergeron et al. had to call the variants not only using the GATK tools but also using cross-callers with bcftools. This can be computationally exhausting and time-consuming. A central advantage of our method is that it uses the consistency of inheritance across multiple offspring as an internal validation step.

Although a mutation is generally considered a random event, its occurrence and distribution along the genome are not entirely random, and certain sites have a higher probability of mutations (Supek and Lehner 2019). These sites are referred to as DNM hotspots, where new mutations cluster, and they tend to be closer to each other compared to other sites (Keightley et al. 2014). We found DNM hotspots in this rabbit family and reported their occurrence mainly at sites closer to the telomeres. However, further investigations are needed in this area. We also identified and summarized in Fig. 3 the number and ratio of the two base substitution types, transitions (Ts) and transversions (Tv). We found the ratio of (Ts/Tv) is about 2.14. It is consistent with several previous studies that found a similar ratio in humans (Conrad et al. 2011), and in monkeys (Bergeron et

al. 2021). This ratio is important since it might indicate evolutionary or environmental signatures.

In summary, we have developed a Python script with flexible options to accurately pinpoint the DNMs when data from other siblings is available, demonstrating the impact of a larger family on the analysis. We are convinced that utilizing long-read technology will significantly increase accuracy, especially in repetitive regions. Furthermore, additional research is needed in this area to investigate DNM hotspots and their associated consequences.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s42977-026-00312-7>.

Acknowledgements The authors acknowledge the support of ELIXIR Hungary (www.elixir-hungary.org).

Funding Open access funding provided by University of Debrecen. This research was funded by the National Research, Development, and Innovation Office of Hungary, grants NKFI/2017–1.3.1-VKE-2017-00026 and NKFI/OTKA K 132814. Project no. TKP2021-NKTA-34 has been implemented with support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development, and Innovation Fund, financed under the TKP2021-NKTA funding scheme.

Data availability Scripts used for de novo mutations analysis are available from GitHub at https://github.com/Maher199/Discover_DNMs. The whole-genome sequences used in this study are available at the following BioProject accession number: PRJNA1237607. Sequencing runs can be accessed individually with the SRA accession numbers starting from SRX28036935. The father's sample was submitted earlier under the PRJNA1118464/SRS21521871 accession number.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval Not applicable.

Informed consent Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Acuna-Hidalgo R, Veltman JA, Hoischen A (2016) New insights into the generation and role of de novo mutations in health and disease. *Genome Biol*. <https://doi.org/10.1186/s13059-016-1110-1>
- Alnajjar M, Fekete Z, Nagy T et al (2025) A new family-based approach for detecting allele-specific expression and for mapping possible eQTLs. *Animals*. <https://doi.org/10.3390/ani15182766>
- Arana ME, Kunkel TA (2010) Mutator phenotypes due to DNA replication infidelity. *Semin Cancer Biol* 20:304–311. <https://doi.org/10.1016/j.semcancer.2010.10.003>
- Azevedo L, Amaro AP, Niza-Ribeiro J, Lopes-Marques M (2024) Naturally occurring genetic diseases caused by de novo variants in domestic animals. *Anim Genet* 55:319–327. <https://doi.org/10.1111/age.13403>
- Bergeron LA, Besenbacher S, Bakker J et al (2021) The germline mutational process in rhesus macaque and its implications for phylogenetic dating. *Gigascience*. <https://doi.org/10.1093/gigascience/giab029>
- Bergeron LA, Besenbacher S, Zheng J et al (2023) Evolution of the germline mutation rate across vertebrates. *Nature* 615:285–291. <https://doi.org/10.1038/s41586-023-05752-y>
- Conrad DF, Keebler JEM, Depristo MA et al (2011) Variation in genome-wide mutation rates within and between human families. *Nat Genet*. <https://doi.org/10.1038/ng.862>
- Danecek P, Bonfield JK, Liddle J et al (2021) Twelve years of SAMtools and BCFtools. *Gigascience* <https://doi.org/10.1093/gigascience/giab008>
- DePristo MA, Banks E, Poplin R et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498. <https://doi.org/10.1038/ng.806>
- Fekete Z, Németh Z, Ninausz N et al (2025) Whole-Genome Sequencing-Based Population Genetic Analysis of Wild and Domestic Rabbit Breeds. *Animals* 15:775. <https://doi.org/10.3390/ani15060775>
- Friedberg EC (2003) DNA damage and repair. In feature 436 *NATURE* | (Vol. 421)
- Gel B, Serra E (2017) KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* 33:3088–3090. <https://doi.org/10.1093/bioinformatics/btx346>
- Guan P, Sung WK (2016) Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods* 102:36–49
- Keightley PD, Ness RW, Halligan DL, Haddrill PR (2014) Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. *Genetics* 196:313–320. <https://doi.org/10.1534/genetics.113.158758>
- Kohailan M, Aamer W, Syed N et al (2022) Patterns and distribution of de novo mutations in multiplex Middle Eastern families. *J Hum Genet* 67:579–588. <https://doi.org/10.1038/s10038-022-01054-9>
- Lynch M, Ackerman MS, Gout J-F et al (2016) Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* 17:704–714. <https://doi.org/10.1038/nrg.2016.104>
- Sakumi K (2019) Germline mutation: De novo mutation in reproductive lineage cells. *Genes Genet Syst* 94:3–12. <https://doi.org/10.1266/ggs.18-00055>
- Scally A, Durbin R (2012) Revising the human mutation rate: Implications for understanding human evolution. *Nat Rev Genet* 13:745–753. <https://doi.org/10.1038/nrg3295>
- Seah YM, Stewart MK, Hoogstraal D et al (2023) In Silico Evaluation of Variant Calling Methods for Bacterial Whole-Genome Sequencing Assays. *J Clin Microbiol*. <https://doi.org/10.1128/jcm.01842-22>

- Smit AHR& GP (2013) RepeatMasker Open-4.0. <http://www.repeatmasker.org>
- Stankiewicz P, Lupski JR (2002) Genome architecture, rearrangements and genomic disorders. *Trends Genet* 18:74–82. [https://doi.org/10.1016/S0168-9525\(02\)02592-1](https://doi.org/10.1016/S0168-9525(02)02592-1)
- Supek F, Lehner B (2019) Scales and mechanisms of somatic mutation rate variation across the human genome. *DNA Repair (Amst)*. <https://doi.org/10.1016/j.dnarep.2019.102647>
- Van der Auwera GA, Carneiro MO, Hartl C et al (2013) From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr Protoc Bioinf*. <https://doi.org/10.1002/0471250953.bi1110s43>
- Wagner J, Olson ND, Harris L et al (2022) Benchmarking challenging small variants with linked and long reads. *Cell Genomics*. <https://doi.org/10.1016/j.xgen.2022.100128>
- Werling DM, Brand H, An J-Y et al (2018) An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet* 50:727–736. <https://doi.org/10.1038/s41588-018-0107-y>