MEMÒRIA DEL TREBALL DE FI DE GRAU DEL GRAU (ESCI-UPF)

Variant calling in highly homologous regions

Author: Òscar Casals Morro. **NIA:** 106378

Degree: Bioinformatics.

Academic course: 3rd-year

Data: 21/6/2023

Tutors: Alesia Squitieri and Rosa Barcelona Cabeza

Project title: Variant calling in highly homologous regions.
Author: Òscar Casals Morro **NIA:** 106378
Academic course: 3rd-year
Tutors: Alesia Squitieri and Rosa Barcelona Cabeza

Keywords

➢ English:
- Pseudogene.
- Gene
- Mapping
- Variant Calling

➢ Spanish:
- Pseudogén.
- Gen.
- Mapear
- Llamada de variantes.

➢ Catalan:
- Pseudogèn
- Gen
- Mapejar
- Trucada de variants.

Abstract

➢ English

Along the human genome, there are regions known as “dark” characterized by the fact that they can not be adequately assembled or mapped via standard short-read sequencing technologies[3].

The National Human Genome Research Institute (NIH) defines pseudogenes as non-coding DNA segments that resemble a gene and are derived from one that lost its function due to a mutation during the course of evolution. This definition might make pseudogenes look like just “corpses” of true DNA but that could not be further from the truth as some mutations can reactivate them and even if they were always inactive their similarity with coding genes causes a lot of errors in the mapping step of any variant calling pipeline.

The aim of this project was to improve the variant calling such complex regions by using the approach described by Ebbert et al., 2019[3].

➢ Spanish

En el genoma humano hay regiones conocidas como “oscuras” caracterizadas por no poder ser adecuadamente ensambladas o mapeadas mediante las tecnologias de sequenciación short-read estándar[3].

El National Human Genome Research Institute (NIH) define a los pseudogenes como segmentos de ADN parecidos a genes los cuales derivaron de uno que perdió su función debido a una mutación producida durante el transcurso de la evolución. Esta definición puede hacer ver a los pseudogenes como simplemente “cadàveres” del verdadero ADN pero esto no podria estar más alejado de la verdad debido a que muchas mutaciones pueden reactivarlos e incluso si esto no pudiera suceder siguen siendo muy similares a genes que sí tienen función lo que causa bastantes errores en la etapa de mapeado de muchos programas destinados a la llamada de variantes.

El objetivo de este projecto es mejorar la llamada de variantes en estas regiones complejas usando la metodología descrita por Ebbert et al., 2019[3].

- ➢ Catalan

En el genoma humà podem trobar regions conegudes com a “fosques” caracteritzades per no poder ser adecuadament ensamblades o mapejades mitjançant les tecnologies de seqüenciació estandard short-read.

El National Human Genome Research Institute (NIH) defineix als pseudogens com a segments d’ ADN semblants a gens els quals deriven de un que va perdre la seva funció degut a una mutació produïda durant el transcurs de l’evolució. Aquesta definició pot fer veure als pseudogens com simplement “cadàvers” del verdader ADN però això no podría estar més allunyat de la veritat degut a que moltes mutacions els poden reactivar i encara que això últim no pogués passar segueixen sent molt similars a genes que sí tenen una funció el que produeix bastants errors en l’etapa de mapejat de molts programes destinats a la trucada de variants.

L’objectiu d’aquest projecte és millorar la trucada de variants en aquestes regions tan complexes utilitzant la metodologia descrita per Ebbert et al., 2019[3].

Introduction

For a long time, researchers have known of regions in our genome that could not be assembled or aligned via the most common sequencing technology, Next Generation Sequencing (NGS), one of which are curious gene-like segments known as pseudogenes[3].

Pseudogenes are segments of DNA structurally similar to genes that are unable to code for a protein and are predicted to be as abundant as coding genes, with predictions claiming there could be around 20,000 in humans[30]. They originated from coding genes that during the course of evolution suffered some kind of mutation that made them lose their function[1].

Their lack of functionality may make them look like just “junk” or “relic” DNA and even if they have been considered that way for a long time, some studies have proven that certain mutations can reactivate them giving birth to many kinds of variants that could affect us in a positive or negative way, whether it be causing important health issues or codifying a protein with a really beneficial function[1]. For example, pseudogenes are closely related to cardiovascular diseases, the leading cause of death globally taking around 18 million lives every year[4], and they are widely expressed in human cancers[5].

Pseudogenes are still considered a challenge to researchers and clinicians. In fact, the percentage of homology with their counterparts can generate misalignments of target genes. Standard NGS aligners, fail when complex regions have to be mapped, increasing the error rate in variant calling algorithms, by detecting false positives or false negatives mutations[3].

This issue is so important that many NGS mapping algorithms have tried to deal with this problem. The most used solution is the one followed by the famous Burrows-Wheeler Aligner(BWA)[15] which consists of assigning a read that matches equally well multiple regions to all of them giving it a low mapping quality, still in some pair-end sequencing scenarios BWA might think the reason why a read can be aligned to multiple zones is that the mate pair is nearby and not due to duplication, therefore, giving it a high mapping quality; the problem with this solution is that just indicates which reads are ambiguous instead of solving the problem.

Another solution is using long-read sequencing technologies such as Oxford Nanopore or PacBio since their higher read size makes identifying the correct placement of a read easier for specialized long-read mapping algorithms such as minimap2[31]. However, this comes at the cost of a higher error rate which can go up to 15% in many cases. Even with this handicap researchers have been using long reads to deal with pseudogene variants due to the lack of a better option, since for example even with the added error rate Oxford nanopore reduces a 9.6% the number of difficult-to-map regions in the genome[6,1].

The problem with having such unreliable ways to locate variants is that a lot of gaps can be found in a patient's genome which in turn can make important diseases slip through the radar of a researcher till it is too late. This is a common problem in the detection of rare diseases which affect 30 million people in the USA[7].

Finally, it is important to highlight the obvious issue of the great number of false positives caused by pseudogenes, which can complicate scientists' work, by reaching wrong conclusions in any kind of genetics-related research[32].

It is clear there is a need to adapt the traditional mapping and variant calling process to pseudogenes, which is why in this project one of the most accredited approaches will be applied: the Dark and camouflaged region[3].

The dark and camouflaged regions approach defines those regions that are hard to map due to a low mapping quality or depth as dark and those that confuse alignment algorithms due to their homology as camouflaged, masking the latter so they are not taken into account during the analysis. This algorithm consists of first locating the dark regions by comparing multiple alignments performed on the same reference genome and dividing them into two groups: dark by depth and dark by quality, those regions inside the dark-by-quality group are masked and the samples we are studying are realigned to this genome preventing the high homology of pseudogenes to cause misassignment of reads, variant calling is then performed. Even if we have masked repetitive regions so only one is taken into account when a camouflaged region is not completely identical to its counterpart it can happen that a read from it is forcibly aligned to another region creating an artificial variant (reference based artifacts), to reduce the amount of these false positives the Dark and camouflaged approach aligns the coding regions that are repeated five times or more and belong to one of the regions that have been camouflaged back to the reference genome, selects those with sequence identity higher or equal to 98% and aligns them back to the sequence, annotating every gap and mismatch as a position prone to reference artifacts; finally the inbreeding coefficients of each of the variants found are computed and the mutations with low quality that are at a zone prone to reference based artifacts are removed[3].

While this approach solves the issue of misalignments, it doesn’t allow us to distinguish the mutations that happen in genes from those that happen in pseudogenes[3].

Objectives

This project aims to provide a pipeline that improves variant calling of complex regions due to the presence of high homology regions like pseudogenes.

Methodology

To execute the pipeline the first thing needed was the data to execute it and see how precise it is, these files were: a reference genome, its annotation file to locate genes and pseudogenes, and reads of a sample with known mutations in high homology regions.

The reference genomes used was the human chromosome 1 from the GRCh38 assembly in GenBank[21], the sequence was downloaded in fasta format[29], and the annotation was in

gff3[27]. This sequence was used to simulate the reads where to look for variants as there were no publicly available real data.

The reads were simulated using dwgsim[22], an updated version of the renowned Whole Genome Sequence simulator wgsim, this tool allows the simulation of both Whole Genome Sequence (WGS) and Whole Exon Sequence (WES) reads based on a reference sequence and configured parameters, which were: a mutation rate of 0.0001(the same as humans[12]), reads of 150 bp and 50000000 read pairs; since that mutation rate did not produce many mutations it was changed to 0.1.

With the data ready it was time for building the pipeline, the main tasks to accomplish were locating the regions with high homology to a pseudogene, camouflaging them leaving only one representative of each unmasked, and performing the variant calling on this new masked genome.

The first step, locating the zones with high homology to a pseudogene, was done by first extracting the regions belonging to genes or pseudogenes using bedtools[18] getfasta and the annotation file, then a blast database of the gene sequences was created using makeblastdb[24] and all pseudogenes were queried on it keeping the regions that aligned with 90% sequence identity or more as members of the camouflaged group of the pseudogene, this alternative way of looking for genes with high homology to pseudogenes removes the need of aligning our reference with multiple reads to find out which regions have low mapping quality due to repeats. From this step, two groups of bed files were made for each ploidy: one contained the representatives of each camouflaged group(that from now on will be referred to as Align_to file) which were the genes with lower e-value in the blast alignment, and the other the regions that needed to be realigned because of their high homology(that from now on will be referred to as Realign_to).

With the bed files created, the reads of the simulated sample were aligned to the reference genome using *bwa mem*[15] so the regions in need of realigning could be extracted via samtools[10] view and bedtools *bamtofastq*. Then the regions that were not in the *Align_to bed* file were selected via bedtools complement and masked with bedtools maskfasta, the resulting genome was indexed using *bwa index, samtools faidx and picard*[20] *CreateSequenceDictionary*. This process was repeated for each ploidy.

Now that a masked genome had been created for each ploidy what was left to do was align the regions extracted from the sample to the reference genome with bwa mem, index the result using samtools index, run gatk[16] HaplotypeCaller on the alignment to locate variants and finally use gatk GenotypeGVCFs to erase those that were likely caused by sequencing biases and apply a phred-scaled confidence threshold to remove the ones that were forcely aligned to regions they didn't correspond to; the vcfs generated for each ploidy were fused using vcf-concat and the highly homologous regions present in this final vcf were flagged using the results from the earlier blast.

With the pipeline ready it was time to select the gene from which reads will be simulated to locate variants in them. The first test was done simulating 986 mutations 113 of which were in high homology regions inside the widely studied GBA1, a gene highly homologous to the pseudogene GBAP1 that encodes the lysosomal enzyme glucocerebrosidase known to have

mutations that are among the most common risk factors for Parkinsons disease and related synucleinopathies[28]. When the test finished a variant calling without any masked regions was performed on the same simulated reads to see if the pipeline made a difference, this consisted of simply aligning the reads to the reference via bwa mem and running gatk HaplotypeCaller and GenotypeGVCFs to locate mutations and remove those likely caused by sequencing errors, the improvement the pipeline made over a normal variant calling was not as substantial as expected probably because most of the simulated variants were not in regions with high homology to GBAP1(as can be seen in the supplementary material by comparing *Figure 7* and *Figure 8* with *Figure 9* and *Figure 10*), therefore another test was done in which only reads of the zones that have high sequence identity with the pseudogene were simulated with the objective of seeing if our pipeline was able to handle this specific type of duplications better than a normal variant calling would, this last case was repeated with different quality thresholds for the variants to see which brought a good balance of a good amount of true positives and false positives, in all the tests the maximum ploidy used was 10.

Once all tests finished a graph displaying the score for each variant was created in order to see if the phred-scaled confidence of the mutations in the zones of high homology was abnormally high or low(*Figures 5* and *FIgure 6*) in addition to a confusion matrix[25] (*Figure 1*, *Figure 3,* and from supplementary material *Figure 7 and Figure 9*)of each test, this matrix is a table commonly used to define the performance of a classification algorithm such as this one that contains the amount of correctly predicted gene variants(true positives), the number of variants predicted in genes that do not actually exist or belong to pseudogenes (false positive), undetected variants(false negatives) and positions that do not possess any variant and no variant was predicted in it(true negatives).

From each confusion matrix the accuracy[25], precision[25], recall[25], F-score[25], and Phred-score[26] (*Figure 2* and *Figure 4*) were computed to see if the pipeline was good or required more tweaking.

Accuracy was computed to see how many cases our pipeline predicted correctly, still, this metric is known to be highly biased therefore to see the percentage of variants predicted by the algorithm that were actually true positives the precision was computed. Additionally, Recall was computed to see how many variants the pipeline is able to identify and the F1-score was calculated to find out if there was a good balance between precision and recall.

All code used can be found in the following GitHub:
https://github.com/PhOsMi/Detect_variants_High_homology

Additionally in both the README file of the repository and in *Figure 11* inside the supplementary materials, there are instructions on how to use it.

Results and discussion

These results refer to simulated data and can not be applied to real cases.

The confusion matrix and metrics computed from the variant calling our pipeline did on the reads with mutations only in high homology regions are the following:

58/145 (variants predicted/total variants)	Variants simulated in the gene	Positions with no variants
Variants predicted in the gene	46 (TRUE POSITIVES)	12 (FALSE POSITIVES)
Positions where no variants where predicted	87 (FALSE NEGATIVES)	1390 (TRUE NEGATIVES)

Figure 1: Confusion matrix of pipeline applied on all GBA1.

Accuracy	0.92
Precision	0.79
Recall	0.35
F1-score	0.48
Phred-score	11.96

Figure 2: Accuracy, Precision, Recall, F-score and Phred-score of the pipeline.

In this case we can see that the pipeline has high ccuracy and precision which means that the variants this program detects on simulated data have a high chance of being true positives, the same can not be said about the F-score and recall since a recall of 0.35 implies that this pipeline is only capable of identifying 35% of the variants in high homology regions and an F1-score of 0.48 means there is not a good balance between precision and recall, from this results, we can say this imbalance is caused by the fact that while the program has good precision its recall is so low that it misses a good chunk of variants. The Phred score of the pipeline is also not really high with simulated data since it implies there is around a 90% chance the base we are observing is correct instead of the commonly accepted 99% (phred-score of 20).

To see how much our pipeline improves over a variant calling in which no masking was done we have to compare it with the metrics from one performed on the same reads.

75/145 (variants predicted/total variants)	Variants simulated in the gene	Variants not simulated in the gene
Variants predicted in the gene	32 (TRUE POSITIVES)	43 (FALSE POSITIVES)
Variants not predicted in the gene	70 (FALSE NEGATIVES)	1390 (TRUE NEGATIVES)

Figure 3: Confusion matrix of normal variant calling on all GBA.

Accuracy	0.91
Precision	0.43
Recall	0.31
F-score	0.36
Phred-score	11.33

Figure 4: Accuracy, Precision, Recall, F-score and Phred-score of normal variant calling on all GBA.

Comparing these metrics with the ones from the test above shows that our pipeline greatly improved the precision meaning it has fewer errors when predicting mutations in simulated data, but the accuracy, recall, F-score and Phred-score have not improved much which means that while our pipeline has fewer false positives it does not cover much more relevant variants than a normal variant calling would.

Finally, plots showing the phred-scale-confidence of each variant were made to see if there was any region in both cases that presented an abnormally high or low quality as well as seeing if the false positives had a lower score than the other mutations and therefore, could be easily noticed.

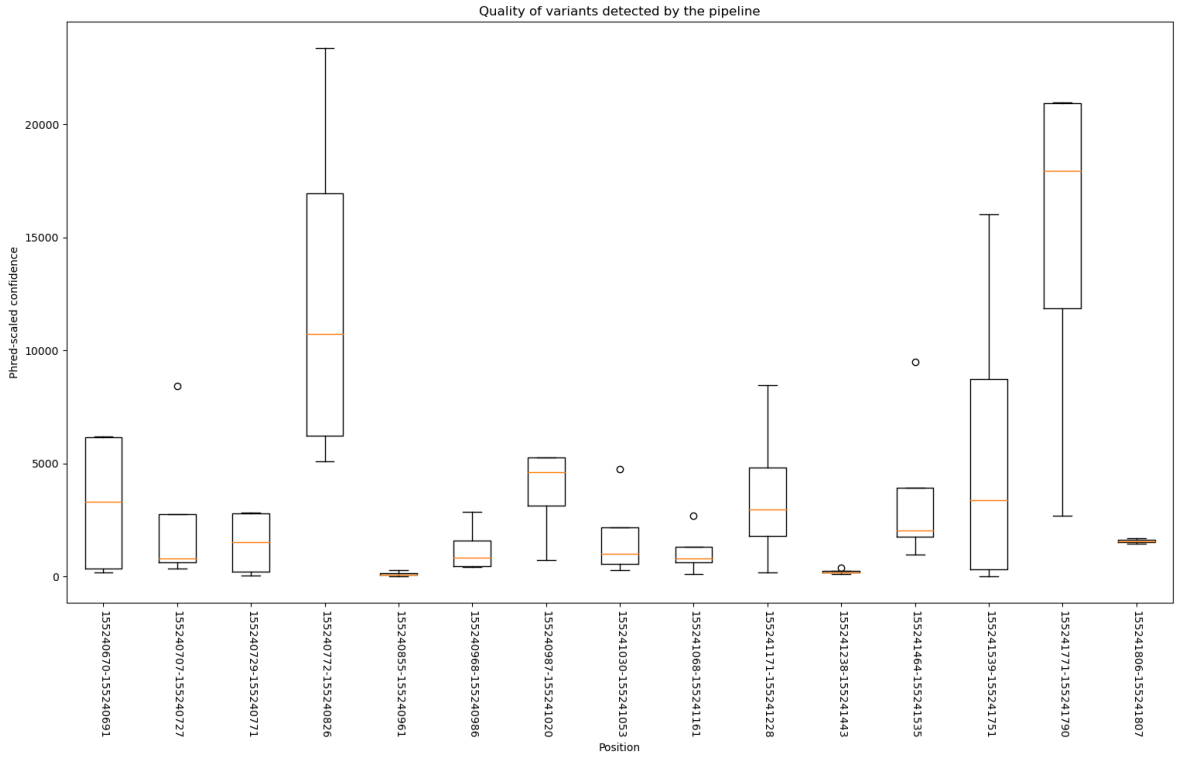

Figure 5: Phred-scale-confidence of the variants detected by the pipeline, for display convenience, they have been grouped in groups of 4. The x-axis contains the start and end positions of each group of variants and the y-axis is the phred-scale-confidence score.

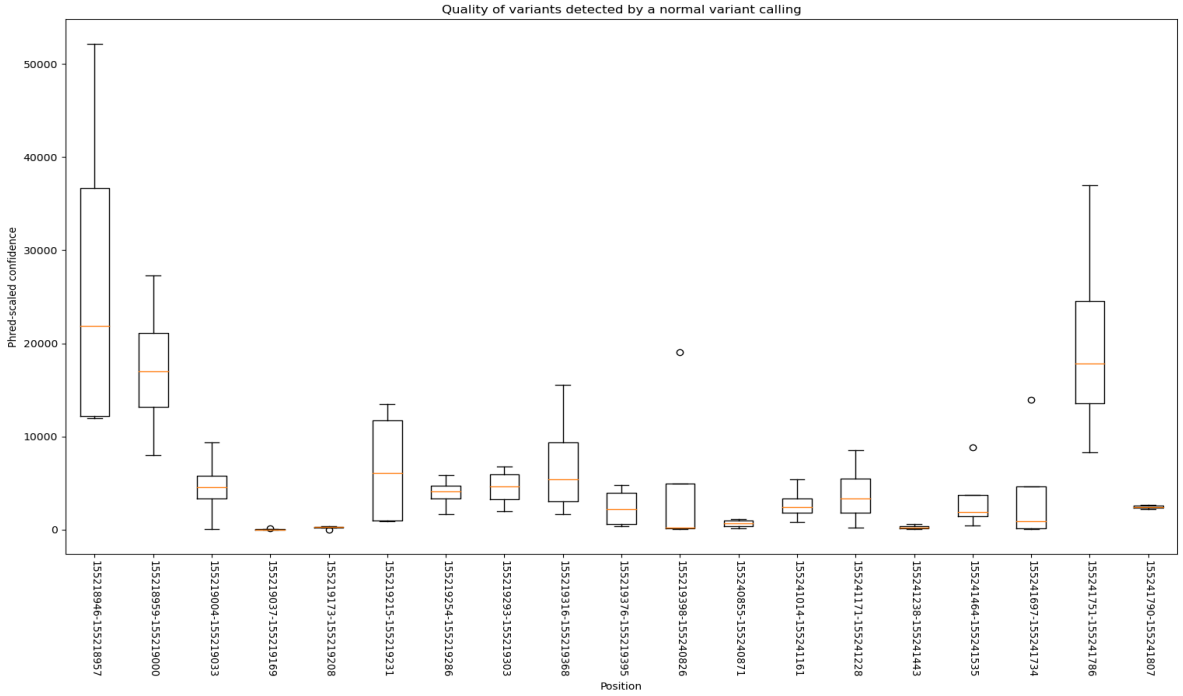

Figure 6: Phred-scaled-confidence of the variants detected by the normal variant calling, for display convenience, they have been grouped in groups of 4. The x-axis contains the start and end positions of each group of variants and the y-axis is the phred-scale-confidence score.

In *Figure 5* we can see three regions with a lower phred scale confidence than the rest, these regions are the ones containing the false positives therefore we can say that if a variant predicted by the algorithm has an outstandingly lower quality compared to the rest it is probably a false positive. Alternatively in *Figure 6* we can also see there are variants with a lower quality than the others which can also be explained by the fact that they are false positives but there are also some which are false positives and possess a pretty high quality such as the one in position 155218946 that has a phred-scaled-confidence of 52176.77(the highest quality in the normal variant calling) and therefore could be mistakenly taken as a true positive.

Conclusion

Given the results and considering that they have not been produced with real data, we can conclude that in these simulations the pipeline reduces the number of false positives which constitute one of the main problems traditional variant calling approaches have with these regions.

The algorithm though does not improve much the number of relevant variants detected compared to normal variant calling algorithms and is not capable of distinguishing reads that come from genes from those that come from pseudogenes, issues another project tackling this issue should solve.

In conclusion, the pipeline reduces the number of false positives usually found when looking for variants in regions with high homology to a pseudogene though the amount of relevant variants detected does not increase much compared to traditional variant calling methods and the pipeline can not distinguish reads from genes from those of pseudogenes. The fact the data used to test it was artificial prevents it to be a good option for projects focused on these regions as it is not known if the high precision translates to real cases.

References

1. Pink, Ryan Charles, Kate Wicks, Daniel Paul Caley, Emma Kathleen Punch, Laura Jacobs, and David Raul Francisco Carter. “Pseudogenes: Pseudo-Functional or Key Regulators in Health and Disease?” RNA 17, no. 5 (May 2011): 792–98. https://doi.org/10.1261/rna.2658311.
2. Genome.gov. “Pseudogén,” September 14, 2022. https://www.genome.gov/es/genetics-glossary/Pseudogen.
3. Ebbert, Mark T. W., Tanner D. Jensen, Karen Jansen-West, Jonathon P. Sens, Joseph S. Reddy, Perry G. Ridge, John S. K. Kauwe, et al. “Systematic Analysis of Dark and Camouflaged Genes Reveals Disease-Relevant Genes Hiding in Plain Sight.” Genome Biology 20, no. 1 (May 20, 2019): 97. https://doi.org/10.1186/s13059-019-1707-2.
4. Ebbert, Mark T. W., Tanner D. Jensen, Karen Jansen-West, Jonathon P. Sens, Joseph S. Reddy, Perry G. Ridge, John S. K. Kauwe, et al. “Systematic Analysis of

Dark and Camouflaged Genes Reveals Disease-Relevant Genes Hiding in Plain Sight." Genome Biology 20, no. 1 (May 20, 2019): 97. https://doi.org/10.1186/s13059-019-1707-2.
5. Poliseno, Laura, Andrea Marranci, and Pier Paolo Pandolfi. "Pseudogenes in Human Cancer." Frontiers in Medicine 2 (September 25, 2015): 68. https://doi.org/10.3389/fmed.2015.00068.
6. Holley, Guillaume, Doruk Beyter, Helga Ingimundardottir, Peter L. Møller, Snædis Kristmundsdottir, Hannes P. Eggertsson, and Bjarni V. Halldorsson. "Ratatosk: Hybrid Error Correction of Long Reads Enables Accurate Variant Calling and Assembly." Genome Biology 22, no. 1 (January 8, 2021): 28. https://doi.org/10.1186/s13059-020-02244-4.
7. Marwaha, Shruti, Joshua W. Knowles, and Euan A. Ashley. "A Guide for the Diagnosis of Rare and Undiagnosed Disease: Beyond the Exome." Genome Medicine 14, no. 1 (February 28, 2022): 23. https://doi.org/10.1186/s13073-022-01026-w.
8. Duret, L. (2008) Neutral theory: The null hypothesis of molecular evolution. Nature Education 1(1):218
9. Ebler, Jana, Peter Ebert, Wayne E. Clarke, Tobias Rausch, Peter A. Audano, Torsten Houwaart, Yafei Mao, et al. "Pangenome-Based Genome Inference Allows Efficient and Accurate Genotyping across a Wide Spectrum of Variant Classes." Nature Genetics 54, no. 4 (April 2022): 518–25. https://doi.org/10.1038/s41588-022-01043-w.
10. Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li. Twelve years of SAMtools and BCFtools. GigaScience, Volume 10, Issue 2, February 2021, giab008, https://doi.org/10.1093/gigascience/giab008
11. Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. "Twelve Years of SAMtools and BCFtools." GigaScience 10, no. 2 (February 16, 2021): giab008. https://doi.org/10.1093/gigascience/giab008.
12. Stern, Adi, and Raul Andino. "Viral Evolution: It Is All About Mutations." In Viral Pathogenesis: From Basics to Systems Biology: Third Edition, 233–40. Elsevier Inc., 2016. https://doi.org/10.1016/B978-0-12-800964-2.00017-3.
13. Tanner, Georgette, David R. Westhead, Alastair Droop, and Lucy F. Stead. "Simulation of Heterogeneous Tumour Genomes with HeteroGenesis and in Silico Whole Exome Sequencing." Bioinformatics (Oxford, England) 35, no. 16 (August 15, 2019): 2850–52. https://doi.org/10.1093/bioinformatics/bty1063.
14. Blueprint Genetics. "Blueprint Genetics' Approach to Pseudogenes and Other Duplicated Genomic Regions." Accessed May 14, 2023. https://blueprintgenetics.com/pseudogene/.
15. Li, Heng. "Aligning Sequence Reads, Clone Sequences and Assembly Con*gs with BWA-MEM," 2014, 0 Bytes. https://doi.org/10.6084/M9.FIGSHARE.963153.V1.
16. McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." Genome Research 20, no. 9 (September 2010): 1297–1303. https://doi.org/10.1101/gr.107524.110.

17. Lapidus, A. L. “Genome Sequence Databases: Sequencing and Assembly.” In Encyclopedia of Microbiology (Third Edition), edited by Moselio Schaechter, 196–210. Oxford: Academic Press, 2009. https://doi.org/10.1016/B978-012373944-5.00028-6.
18. Quinlan, Aaron R., and Ira M. Hall. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” Bioinformatics 26, no. 6 (March 15, 2010): 841–42. https://doi.org/10.1093/bioinformatics/btq033.
19. Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, et al. “Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics.” Bioinformatics 25, no. 11 (June 1, 2009): 1422–23. https://doi.org/10.1093/bioinformatics/btp163.
20. “Picard Tools - By Broad Institute.” Accessed May 14, 2023. https://broadinstitute.github.io/picard/.
21. “Homo Sapiens Chromosome 3, GRCh38 Reference Primary Assembly,” December 20, 2013. 568336021. NCBI Nucleotide Database. http://www.ncbi.nlm.nih.gov/nuccore/CM000665.2.
22. “Nh13/DWGSIM: Whole Genome Simulator for Next-Generation Sequencing.” Accessed May 15, 2023. https://github.com/nh13/DWGSIM.
23. Lai, Jianbo, Peifen Zhang, Jiajun Jiang, Tingting Mou, Yifan Li, Caixi Xi, Lingling Wu, et al. “New Evidence of Gut Microbiota Involvement in the Neuropathogenesis of Bipolar Depression by TRANK1 Modulation: Joint Clinical and Animal Data.” Frontiers in Immunology 12 (2021): 789647. https://doi.org/10.3389/fimmu.2021.789647.
24. Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. “BLAST+: Architecture and Applications.” BMC Bioinformatics 10, no. 1 (December 15, 2009): 421. https://doi.org/10.1186/1471-2105-10-421.
25. B, Harikrishnan N. “Confusion Matrix, Accuracy, Precision, Recall, F1 Score.” Analytics Vidhya (blog), June 1, 2020. https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd.
26. “Phred Quality Score - an Overview | ScienceDirect Topics.” Accessed May 15, 2023. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/phred-quality-score.
27. “GFF3 File Format.” Accessed June 12, 2023. https://www.ensembl.org/info/website/upload/gff3.html.
28. Do, Jenny, Cindy McKinney, Pankaj Sharma, and Ellen Sidransky. “Glucocerebrosidase and Its Relevance to Parkinson Disease.” Molecular Neurodegeneration 14, no. 1 (August 29, 2019): 36. https://doi.org/10.1186/s13024-019-0336-2.
29. Zhang, Hongen. “Overview of Sequence Data Formats.” Methods in Molecular Biology (Clifton, N.J.) 1418 (2016): 3–17. https://doi.org/10.1007/978-1-4939-3578-9_1.
30. Torrents, David, Mikita Suyama, Evgeny Zdobnov, and Peer Bork. “A Genome-Wide Survey of Human Pseudogenes.” Genome Research 13, no. 12 (December 2003): 2559–67. https://doi.org/10.1101/gr.1455503.

31. Li, Heng. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” Bioinformatics 34, no. 18 (September 15, 2018): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
32. Brandt, Débora Y C, Vitor R C Aguiar, Bárbara D Bitarello, Kelly Nunes, Jérôme Goudet, and Diogo Meyer. “Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data.” G3 Genes|Genomes|Genetics 5, no. 5 (May 1, 2015): 931–41. https://doi.org/10.1534/g3.114.015784.