Cell Biology: Research & Therapy

Review Article

A SCITECHNOL JOURNAL

Advances in Simple Sequencing, NGS and WGS: from Cancer to Oleaginous Yeast and to COVID-19

Tania Iouk*

Abstract

Objective:

Until recently, the whole-genome-Until recently, the whole-genome-sequencing (WGS) results for the new strains and organisms weren't devoid of errors, and individual coding sequences from each contig had to be revisited and re-deposited based on the mRNA sequence.

Developed to detect somatic mutations in cancer patients, the next-generation sequencing (NGS) approach substantially decreased the error rate, and it was then applied to the sequencing of new or frequently mutating genomes including the SARS-CoV-2 genome. Sequence assembly of novel oleaginous yeast *Yarrowia lipolytica* strain also showed the usefulness of NGS in the new-strain sequencing. Hence, previously, the oleic-acid utilizing yeast led to the discovery of novel transcription factors and they also proved to be useful in testing the chemical libraries.

Conclusions: The advances in genome sequencing including NGS and emerging long-read sequencing (LRS) technologies are constantly improving the characterization of genetic variants. Some alternative bioinformatics tools are described in this review.

Keywords:

DNA sequencing, NGS, cancer, oleaginous yeast, SARS-CoV-2, genome annotation

Introduction

While analyzing the new sequences of COVID-19 coronavirus isolates we often look back at the previously established error profiles in reporting the sequencing data, including the related to cancer deep NGS (next-generation sequencing) results [1-3].

In a way, the coronavirus mutations are similar to the related to cancer somatic mutations, as they develop in the patient's organism, sometimes on the background of certain other abnormalities or due to aging [4-5]. NGS was increasingly suggested as a method of choice for the analysis of the coronavirus genome [6].

Revealed by biopsy, including the blood-cell biopsy which analyzes a circulating tumor DNA, the NGS-elucidated cancer somatic mutations include the p53 tumor-suppressor gene mutations in patients with Li-Fraumeni syndrome, the somatic mutations in NRAS/KRAS GTPase genes that cause a relapse of leukemias, and the somatic mutations in DNMT3A, ASXL1, and TET2 genes that increase the likelihood of blood cancer in aging patients (65 years and older) [3, 7-8].

The detected somatic cancer mutations often exhibit mosaicism (or clonal hematopoiesis – in the case of blood cancer) when only a small percentage of identical reads (3-20% in the case of p53 for example) contains new mutations [7]. In the case of SARS-Cov2, the mosaicism could be linked to the presence of two or more virus variants in organism [9].

The somatic cancer mutations are considered the low-frequency ones, as were initially the coronavirus mutations [4-5]. Indeed, early SARS-Cov-2 isolates were treated as the products of low-frequency mutagenesis until the appearance of Delta with its 7 hot-spot mutations and then Omicron (a.k.a. B.1.1.529) variant with 30 mutations in a spike-protein RNA [10-11]. In the case of Omicron, the exceptionally high mutation rate was attributed to the presence of HIV in the immune-suppressed patient.

In an experiment, a high rate of mutagenesis is usually achieved by decreasing (10 fold) concentration of one of the nucleotides in the PCR reaction, usually ATP, and also by using a non-high fidelity polymerase such as Taq [12]. However, the shortage of ATP in organisms is not known to increase the incidence of somatic mutations. Instead, it causes energy-related dysfunctions and increased blood urate levels (after the AMP deaminase degrades AMP to IMP) [13]. IMP then becomes transformed into inosine and subsequently to purines in the hypoxanthine -> xanthine -> urate chain of biochemical reactions. The treatment with the xanthine oxidoreductase inhibitors and inosine is known to increase the pool of salvageable purines and to restore ATP levels in humans [13]. Remarkably, however, one publication suggests that SARS-CoV-2 infection causes the xanthine-oxidoreductase inhibition and that it, therefore, could cause a localized depletion of ATP, thereby increasing the likelihood of random mutagenesis during propagation of viral RNA [14].

COVID-19 studies

Originally it was the WGS (whole genome sequencing) to be applied for studying SARS-CoV-2. The mutation rate was initially estimated at 33 genomic mutations/year. The analysis of numerous (>18,000) SARS-CoV-2 sequences sampled between December 2019 and fall 2020 suggested that mutations across the genomes were due to neutral evolution and not adaptive selection [15]. Based on the presence in the viral genome mutations, the lineages were assigned (Omicron for example is a B.1.1.529) [10, 11].

*Corresponding author: Dr. Tania Iouk, Department of Biology, Concordia University, Science pavilion, room SP 532.01 7141 rue Sherbrooke H4B 1R6, Montreal, Canada, E-mail: tania.iouk@concordia.ca

Received: December 15, 2021 Accepted: December 20, 2021
Published: December 25, 2021

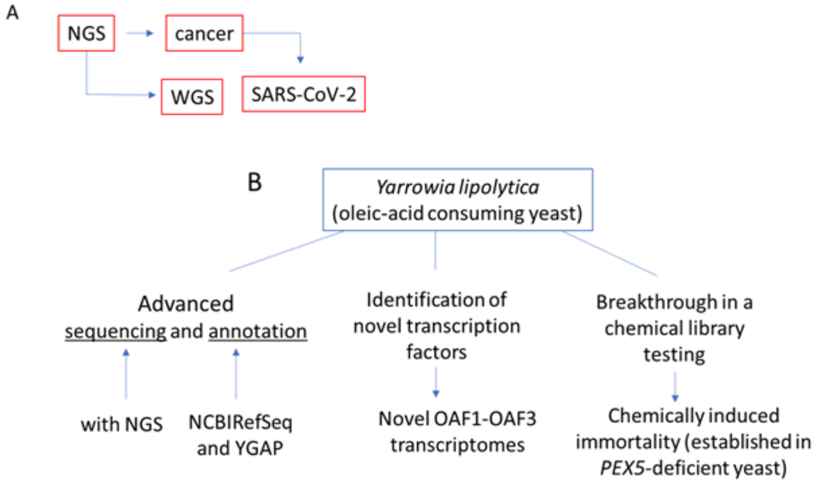
A
NGS
cancer
WGS
SARS-CoV-2
B
Yarrowia lipolytica
(oleic-acid consuming yeast)
Advanced
sequencing and annotation
Identification of novel transcription factors
Breakthrough in a chemical library testing
with NGS
NCBIRefSeq and YGAP
Novel OAF1-OAF3 transcriptomes
Chemically induced immortality (established in *PEX5*-deficient yeast)

Figure 1: (A) Originally used to characterize somatic cancer mutations, NGS was subsequently applied to the whole-genome sequencing (WGS) and more recently to identifying COVID-19 variants. (B) The genome of the new *Yarrowia lipolytica* strain was sequenced in recent years using the combination of NGS and PacBio long-read sequencing [23]. Previously the oleic-acid consumption and its involving mechanisms were used to identify new transcription factors and to chemically immortalize yeast cells.

Recently, the NGS was proposed as a method of choice to identify and track the emergence and prevalence of novel strains of SARS-CoV-2, and low- to mid-throughput NGS assay was developed specifically to combat the COVID-19 pandemics [6, 16-17]. The kit was industrially developed that includes 98 amplicons so that SARS-CoV-2 consensus could be reported with confidence if 90 or more amplicons were detected. The sequencing procedures are being standardized according to ARTIC (Advancing Real-Time Infection Control Network) guidelines that were also specifically developed to fight COVID-19 [18].

NGS-related breakthroughs in the eukaryotic genome sequencing and annotation

The available at NCBI and other databases genome sequencing data are not free from errors [3]. The deposited contig-based gene sequences from low eukaryotes may carry a 2-3% sequence error, and the genome assemblies are not devoid of errors either, as they may contain structural defects. The correcting programs were developed to analyze reads that do not map properly to the assembly (e.g. when there is a difference between the read size and insert size or when the "soft-clipped" reads are present, i.e. when one end of the read is mapped to the reference while the second is not). One program, known as NucBreak analyzes the alignments of reads that are properly mapped to an assembly, however, it is designed to elucidate the alternative read alignments, which could actually be the correct ones [19]. Several tools exist aimed to assess the genome assembly accuracy (e.g. REAPR, FRCbam, Pilon, etc.) that detect structural errors including medium to long insertions and deletions, inversions, duplications, and inter-and intra-chromosomal rearrangements [19-20].

Analyzing the whole genome using NGS and Illumina systems provides high throughput and is also cost-effective. The sequencing data is characterized by a relatively short read length (100–300 bp) and high accuracy, however, it may include the sequencing errors towards the 3'-end of the reads and it does not necessarily provide a uniform distribution of reads across the genome [21-22]. Still, despite its short read length, Illumina data is often used for *de novo* genome assembly, sometimes complemented by data generated through other platforms, such as PacBiO, which generate longer "bridging" reads [21, 23].

All input reads are used to generate a de Bruijn graph which establishes the overlap between individual reads, ultimately creating a path through a contig.

Good example of such combinational NGS Illumina/PacBio approach to the whole-genome is sequencing of industrially-relevant strain W29/CLIB89 of *Yarrowia lipolytica* that utilizes an oleic fatty acid (OA, C18) as a sole carbon source. Genome characterization of other *Y. lipolytica* strains was reported previously, hence OA-utilizing yeast cells had been extensively studied in the past. These cells and their responsible for the OA metabolism organelles became the unique experimental models in which to identify novel transcription factors, test chemical libraries, and sharpen the genome sequencing and annotation skills (Figure 1)[23-25].

In the mentioned above project, each *Y. lipolytica* chromosome was represented by a single contig. The Illumina system produced ≈14x109 of the very short (≈100bp) reads, and the third-generation sequencing technology PacBio reads (3000-5000 bp long) along with the optical Irys genome mapping system were used to evaluate the integrity of the assembly, to estimate the extent of unassembled sequence (in telomeric regions), and to localize rDNA repeats [23].

The Irys system from BioNano Genomics Inc. is designed to stretch long chromosome segments inside an array of nanochannels for genetic analysis. It should be noted that the acquisition of stretched and labeled DNA images was enabled years ago, before the appearance of NGS [26-27].

The non-optical genome mapping systems

The already mentioned NucBreak program enables the non-optical mapping of data, and it also allows the omission of long PacBio reads. Together with BreakDancer, Lumpy, and Whamit tools, it presents a computer program capable of detecting structural variants that can't be seen through the conventional NGS data compiling [19].

The NucBreak tool was created to detect structural errors in the assembly by using paired-end Illumina reads. The reads are first mapped to the assembly, and then the mapping results are analyzed to detect the assembly error locations. The error detection process starts with mapping reads to the assembly by using Bowtie2, a gap-read alignment tool.

Bowtie2 is run separately for each read file to report all local alignments with an added nucleotide match bonus. The resulting files contain all possible alignments for each read, and they do not depend on the second read in a pair. Notably, a read alignment may contain either a full read sequence or a read sequence clipped on one or both ends [19].

Back to the new *Y. lipolytica*-strain sequencing

The genome annotation pipeline included the previous *Y. lipolytica* annotations from NCBI/RefSeq, yeast YGAP [28], and SnowyOwl fungal HMM (hidden Markov model) predictions [29]. The strongest overlap in terms of the CDS prediction and identification was reported between the sequencing data and RefSeq, and the weakest between the sequencing data and SnowyOwl [29].

The SnowyOwl uses RNA-Seq data to provide hints for the generation of Hidden Markov Model (HMM)-based gene predictions and to evaluate the resulting models. The pipeline has been developed based on three manually curated gene models in fungal genomes, including among others *Aspergillus niger* (ATCC 1015). *A. niger* is the industrially relevant producer of the organic metabolites and enzymes used in numerous biotechnology applications. It secretes numerous glycoside hydrolases capable of hydrolyzing cellulose, hemicellulose, or pectin to sugars, and is famous for studying the signal-peptide- containing proteins traveling through ER (endoplasmic reticulum) to become secreted [30]. It is also used as a host in which to express the secreted enzymes from the other phylogenetically diverse and frequently pathogenic fungal species [31]. However, the *A. niger* genomic DNA with its 53.3% GC content is characterized by reduced usage of the subset of synonymous A- and T-ending codons, therefore expression of the A, T-rich sequences requires numerous codon substitutions [32]. Many GC-rich fungal cDNAs mimic the average codon usage in *A. niger*, however, the >60% GC content of those organisms could interfere with protein biosynthesis due to the limited availability of some tRNAs, plus strong GC enrichment interferes with the diversification of synonymous codon usage [33].

DNA sequencing and alternative genome browsers

The SnowyOwl label used to designate a fungal HMM prediction pipeline is the continuation of a 15-year-old tradition to name the biology-related data-engineering platforms, genome browsers, and programs after the birds, with the best-known examples being a MAGPIE data engineering platform, a Bluejay genome browser, and a Seahawk program designed to efficiently load and retrieve data-containing Web pages and text files [34-36]. All modern operating systems perform these operations today momentarily, but the early operating systems and browsers weren't always satisfying to many biologists, thus prompting the creation of additional browsing tools. Bluejay for example is a genome viewer that integrates genome annotation with the gene expression information and comparative analysis, with potentially several other genomes in the same view. Blujay was remarkable for its versatile and detailed displays, and it also offered a circular genome viewing, the prototype of CIRCOS, while also offering visualization of read alignment and the WGS error profiling [36].

The NGS error profiling and its computational suppression

The substitution error rate by conventional NGS is > 0.1%, which is higher than the FDA-authorized detection limit of 0.02 mutant allele fraction (MAF) for hotspot mutations, and 0.05 for nonhotspot mutations at a read-depth of 500–1000x [3].

Errors could be introduced at various steps of the NGS procedure, including sample handling, library preparation, enrichment PCR, and sequencing itself.

The C>A/G>T errors were reported due to DNA damage during sample processing [37, 38]. Spontaneous deamination of methylated cytosine to uracil can cause C>T/G>A errors. Additional errors, as already said, can also be introduced by target-enrichment PCR and the sequencing step [2].

Close evaluation of read-specific error distribution suggests the possibility of computational suppression by which the substitution error rate can be suppressed 10- to 100-fold. The error rates differ by nucleotide substitution types, ranging from 10−5 (for A>C/T>G, C>A/G>T, and C>G/G>C) to 10−4 (for A>G/T>C). The C>T/G>A errors exhibit strong sequence context-dependency, while C>A/G>T errors are dominated by sample-specific effects. The target-enrichment PCR leads to a ~ 6-fold increase in overall error rate, and more than 70% of hotspot variants can be detected at 0.01 ~ 0. 1%frequency with the current NGS technology, by applying in silico error suppression [3].

To determine the lowest frequency at which a true somatic mutation can be distinguished from a sequencing error and to determine site-specific sequencing error rates, the dilution experiment should be performed using cancer- and normal cells from the same patient. In such an experiment, a biopsy material from a tumor is used as well as healthy cells such as lymphocytes, which are easy to isolate from blood, using density centrifugation. The dilution is then performed by spiking 0.1% and 0.02% of cancer genomic DNA into normal DNA [3]. The already known somatic substitution mutations are targeted by amplicon sequencing, using short amplicon reads.

In the high-quality reads, there shouldn't be false-positive detections from the indicated dilution datasets, i.e. the detected mutant allele fraction (MAF) shouldn't exceed the already known number of single-nucleotide variants (SNV) [39]. The exceptions include a loss-of-heterozygosity (LOH) when a chromosome in the cancer line has 1, 2, or 4 mutant alleles. In such a case one SNV may produce several MAF numbers.The error rate in sequencing is estimated as a ratio of reads with a modified nucleotide to the total number of reads. The low-quality reads are those with an error rate of ≥ 1%. They usually contain low-quality bases (a base quality score ≤20), sometimes due to adapter contamination, or due to the low mapping quality. Identified by HiSeq data, these reads are then manipulated either by trimming the flanking 5 bp at both ends to remove adapter contamination or discarded altogether. Ideally, the reads with a base quality score ≥ 30 and the estimated error rate less than 0.1% should be used. The in silico error suppression methods are developed, to identify and filter the low-quality reads (LQReads). Among other criteria, these methods are designed to account for the concordance between forward and reverse readouts so that discordant readouts are not counted [3, 21-22]. The mutation-identifying algorithms are designed to screen the aligned reads and they also suppress error rates resulting from the nucleotide substitution, consequently MAF < 0.002 cannot be distinguished from sequencing errors. On the other hand, "forced calling" of hotspot mutations without considering error may result in false positives. For example, hotspot mutation BRAF K601E is a T>C change that was detected in > 100 tumors [39]. This site was also shown to have an allele fraction of ~ 0.0003 in the mentioned above melanoma cancer cell dilution experiment, and therefore appeared to be the mutation present also in that particular cancer. That however was not confirmed by an undiluted cancer experiment in which it had an exceptionally low MAF of 0.00002 [3].

As already mentioned, a C>T/G>A change has the highest error rate [2]. However, the C>T/G>A mutations are the most common mutation type in cancers [39]. In such cases, the signature analysis is performed that is similar to a signature analysis in disease. It was shown, for example, that the C>T errors exhibit a strong context-dependency, with elevated error rates for G(C>T)N or N(C>T)G and the highest error rate in G(C>T)G consensus [3]. The same pattern was observed for G>A errors in a reverse complement. Other substitution types do not exhibit sequence context dependency as strong as that of C>T/ G>A.

The semi-automated analysis of small-scale DNA sequencing

The routine DNA sequencing aimed to verify the molecular-cloning or lab-performed PCR result has comparable to NSG margin error (Figure 2). We used mRNA of the chemically challenged yeast to synthesize cDNA, and then to PCR amplify distinct ORFs and sequence them in search of mutations. At the same time, we analyzed the sequencing errors. Our best sequencing results contained anywhere from 0.1% to 1.1% error, however, some sequenced ORFs had up to 16% errors. In the good-quality sequencing results G>A/ C>T changes were more frequent than other changes ($p<0.05$) (Figure 2) [40-41]. Some poor-quality sequencing results, which nevertheless still enabled peptide identification, contained: (i) up to 34% of nucleotide substitutions, (ii) 0.02% of inserted triplet codons, and (iii) up to 13% of sequence loops, suggesting the amplification and not the sequencing error. The sequencing results with 40% nucleotide change error were linked to poor DNA quality and were discarded.

Conclusions and perspectives

It was frequently suggested that NGS short reads pose a limitation for the identification of structural variants, sequencing repetitive regions, phasing of alleles, and distinguishing highly homologous genomic regions. These limitations may contribute to the diagnostic gap in patients with genetic disorders. The emerging long-read sequencing (LRS) technologies may improve the characterization of genetic variants. LRS primarily used to investigate genetic disorders with previously known disease loci, and future studies will determine whether LRS technologies can be used for routine WGS to trigger further advancements in medical genetics.

Meanwhile, NGS remains the first choice in WGS, cancer- and COVID-19 sequencing. Hence the already dated shotgun WGS often introduces not only the assembly- but also the sequence errors, including ≤2% errors in deposited protein sequences. Because of this, the individual cDNA sequences or mRNAs must be re-examined and re-deposited, with an improved error margin of <0.3%.

The outlined in this study sequencing of the novel strain of *Y. lipolytica* and other mentioned projects suggest that genome annotation soon will become completely automated and computerized, and it will also be error-free.

Moreover, not only the NGS was optimized further to fully avoid errors, but the methodology also itself is now could be used in a small laboratory setting that does not require a giant hospital lab or an over-equipped university research center.

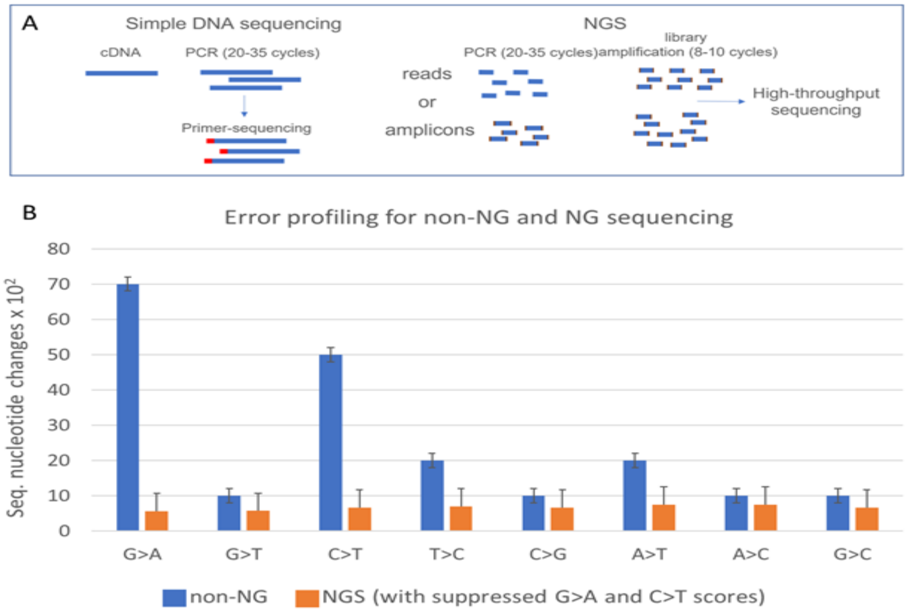

Figure 2: The error rates differ by nucleotide substitution types. (A) NGS could increase the incidence of flanking sequence error, which is suppressed computationally. (B) Error profiling of non-NG and NG sequencing, based on our data.

References

1. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17: 333-351.

2. Salk JJ, Schmitt MW, Loeb LA (2018) Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat Rev Genet 19: 269-285.

3. Ma X, Shao Y, Tian L, Flasch DA, Mulder HL et al. (2019) Analysis of error profiles in deep next-generation sequencing data. Genome Biology 20: 50.

4. Callaway E (2020) Making sense of coronavirus mutations. Nature 585: 174-177.

5. Jungreis I, Sealfon R, Kellis M (2021) SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. Nature Commun 12: 2642.

6. Chen X, Kang Y, Luo J, Pang K, Xu X, et al. (2021) Next-Generation Sequencing revels the Progression of COVID-19. Front Cell Infect Microbiol. 11: 632490.

7. Prochazkova K, Pavlikova K, Minarik M, Sumerauer D, Kodet R, et al. (2009) Somatic TP53 mutation mosaicism in a patient with Li-Fraumeni syndrome. Am J Med Genet A 149A: 206-211.

8. Ma X, Edmonson M, Yergeau D, Muzny DM, Hampton OA, et al. (2015) Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. Nat Commun 6: 6604.

9. Vitale J, Mumoli N, Clerici P, De Paschale M, Evangelista I et al. (2021) Assessment of SARS-CoV-2 Reinfection 1 Year After Primary Infection in a Population in Lombardy, Italy. JAMA Int Med 1407-1408.

10. Callaway E, and Ledford H (2o21) How bad is Omicron? What scientists know so far. Nature https://doi.org/10.1038/d41586-021-03614-z.

11. Karim SA, Karim QA (2021) Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. Lancet https://doi.org/10.1016/S0140-6736(21)02758-6.

12. Lin-Goerke JL, Robbins DJ, Burczak JD (1997) PCR-based random mutagenesis using manganese and reduced dNTP concentration. Biotechniques 23: 409-412.

13. Johnson TA, Jinna H, Kamatani N (2019) Shortage of Cellular ATP as a Cause of Diseases and Strategies to Enhance ATP. Front Pharmacol 10: 98.

14. Pratomo IP, Ariane A, Tedjo A, Heryanto R, Paramita RI (2020) Xanthine oxidase inhibition in SARS-CoV-2 infection: the mechanism and potency of allopurinol and febuxostat in COVID-19 management Med J Indonesia DOI:10.13181/mji.rev.204641.

15. Dearlove B, Lewitus E, Bai H, Li Y, Reeves D, et al. (2020) A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. Proc Natl Acad Sci USA 117: 23652-23662.

16. Hourdel V, Kwasiborski A, Balière C, Mateus S, Betéjat CF et al. (2020) Rapid Genomic Characterization of SARS-CoV-2 by Direct Amplicon-Based Sequencing Through Comparison of MinION and Illumina iSeq100TM System.Front Microbiol 11: 571328.

17. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M et al. (2020) Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. Lancet Infect Dis. 2020; 20: 1263-1271.

18. Tyson JR, James P, Stoddart D, Sparks N, Wickenhagen A et al. (2020) Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. bioRxiv https://doi.org/10.1101/2020.09.04.283077.

19. Khelik K, Sandve GK, Nederbragt AJ, Rognes T (2020) NucBreak: location of structural errors in a genome assembly by using paired-end Illumina reads. BMC Bioinformatics 21: 66.

20. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9: e112963.

21. Heydari M, Miclotte G, Demeester P, Van der Perr Y, Fostier J (2017) Evaluation of the impact of Illumina error correction tools on de novo genome assembly. BMC Bioinformatics 18: 374.

22. Minoche AE, Dohm JC, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol 12: 112.

23. Magnan C, Yu J, Chang I, Jahn E, Kanomata Y et al. (2016) Sequence assembly of Yarrowia lipolytica strain W29/CLIB89 shows transposable element diversity. PloS One 11: e0162363.

24. Smith JJ, Ramsey SA, Marelli M, Marzolf B, Hwang D, et al. (2007) Transcriptional responses to fatty acid are coordinated by combinatorial control. Mol Syst Biol 3: 115.

25. Goldberg AA, Richard VR, Kyryakov P, Bourque SD, Beach A, et al. (2010) Chemical genetic screen identifies lithocholic acid as an anti-aging compound that extends yeast chronological life span in a TOR-independent manner, by modulating housekeeping longevity assurance processes. Aging 2: 393-414.

26. Zhou S, Bechner MC, Place M, Churas CP, Pape L et al. (2007) Validation of rice genome sequence by optical mapping. BMC Genomics 8: 278.

27. Arielly R and Ebenstein Y (2018) Irys Extract. Bioinformatics 34: 134-136.

28. Proux-Wéra E, Armisén D, Byrne KP, Wolfe KH (2012) A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. BMC Bioinformatics 13, 237.

29. Reid I, O'Toole N, Zabaneh O, Nourzadeh R, Dahdouli M, et al. 2014 SnowyOwl: accurate prediction of fungal genes by using RNA-Seq and homology information to select among ab initio models. BMC Bioinformatics 15: 229.

30. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783-795.

31. Kikuchi M, Kitamoto N, Shishido K (2004) Secretory production of Aspergillus oryzae xylanase XynF1, xynF1 cDNA product, in the basidiomycete Coprinus cinereus. Appl Microbiol Technol 63: 728-733.

32. Koda A, Bogaki T, Minetoki T, Hirotsune M (2005) High expression of a synthetic gene encoding potato -glucan phosphorylase in Aspergillus niger. J Biosci Bioeng 100: 531-537.

33. Suzuki H, Saito R, Tomita R (2009) Measure of synonymous codon usage diversity among genes in bacteria. BMC Bioinformatics 10: 167.

34. Gordon PMK and Sensen CW (2007) Seahawk: moving beyond HTML in Web-based bioinformatics analysis. BMC Bioinformatics 8: 208.

35. Baldow C, Salentin S, Schroeder M, Roeder I, Glauche I (2017) MAGPIE: Simplifying access and execution of computational models in the life sciences. PLoS Comput Biol 13: e1005898.

36. Soh J, Gordon PMK, Taschuk ML, Dong A, Ah-Seng AC, et al. (2008) Bluejay 1.0: genome browsing and comparison with rich customization provision and dynamic resource linking. BMC Bioinformatics 9: 450.

37. Chen L, Liu P, Evans TC Jr, Ettwiller LM. (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science 355: 752-756.

38. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, et al. (2018) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res 41: e67.

39. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, et al. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 463: 191-196.

40. Iouk T and Glebov A (2021) Yeast-longevity surrogates obtained in response to steroid lithocholic acid. Cell Biol (Henderson, NV) 10: 5. https://doi.org/10.4172/2324-9293.1000145.

41. Mantere T, Kersten S, Hoischen A (2019) Long-read sequencing emerging in medical genetics. Front Genet 10: 426.

Author Affiliation

Department of Biology, Concordia University, Montreal, Canada