

Yeni Nesil Dizileme: Pipeline Geliştirilmesi

Faruk Barbaros Kalyoncu^{Y.}, İrem Coşkun^{E.}

2003 yılında tamamlanan **İnsan Genom Projesi** ile insan genomuna ait ilk dizinin elde edilmesi, modern biyolojik araştırmaları farklı bir boyuta taşıyarak, tıp ve biyoteknoloji alanında hızlı ve sonuç odaklı değişimlere neden olmuştur¹⁻³. Örneğin, **Maxam-Gilbert** ve **Sanger** nükleik asit (DNA, RNA) dizileme yöntemleri, gelişen bilgisayar teknolojileri ile birlikte daha ayrıntılı analiz imkânı sağlayan **yeni nesil dizileme (NGS)** platformlarının geliştirilmesine öncülük etmiştir⁴. NGS, paralel ve yüksek verimli dizileme teknolojilerini tanımlamak için kullanılan bir ifadedir. NGS, hız ve maliyet açısından Sanger'e kıyasla kazanım sağlayan bir platformdur; ancak beraberinde oldukça karmaşık bir veri analizi sürecini de gerektirmektedir. Bu yüzden, Sanger dizileme teknikleri, NGS sonuçlarını doğrulamak için günümüzde yaygın olarak kullanılmaya devam etmektedir⁵⁻⁷.

NGS Platformları

Birinci nesil dizileme (FGS) veya Sanger dizileme, dideoksinükleotid yöntemi veya sentez yöntemiyle dizileme olarak da bilinmektedir ve DNA dizilemesi için zincir sonlandırma ilkesine dayanmaktadır. Bu yöntem, çok zaman ve maliyet gerektirmesine rağmen küçük ölçekli dizileme çalışmaları için hala altın standart olarak kullanılmaktadır⁸. Dizileme maliyeti/süresini azaltmak ve

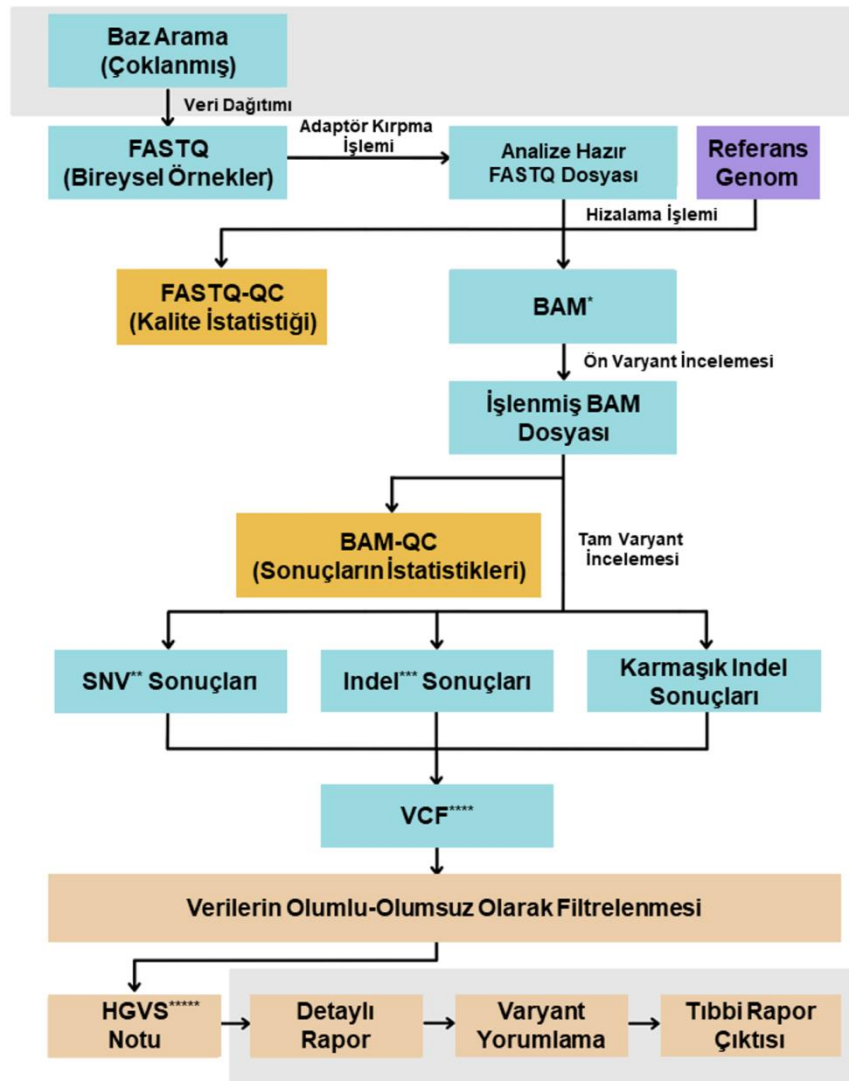
dizilemenin doğruluğunu artırmak için Sanger, **ikinci nesil dizileme (SGS)** teknolojileri ile değiştirilmiştir⁹⁻¹¹. SGS veya yüksek verimli dizileme teknolojilerinin ana adımları; örnek hazırlama, DNA fragmantasyonu, kitaplık hazırlama, dizileme ve veri analizidir. **Roche/454** (2005), **Illumina/Solexa** (2006), **ABI/SOLiD** (2007) ve **Ion Torrent** (2010) geliştirilen platformlar olmakla birlikte¹² Illumina ve Ion Torrent biyomedikal laboratuvarlarda yaygın olarak kullanılan yöntemlerdir¹³. Illumina platformu tek-uçlu (single-end) ve eşli-uçlu (paired-end) okumalar üretebilirken, Ion Torrent yalnızca tek-uçlu okumalar üretebilmektedir¹⁴. Illumina'nın en büyük sınırlaması, genomun tekrarlayan bölümlerinde hataya neden olan kısa okumalar yapmasıdır¹⁵. Ion Torrent'de ise sistemin uzun homopolimerik tekrarları yanlış ölçebildiği bilinmektedir^{16,17}. SGS'nin dezavantajları, düşük dizileme maliyetiyle uzun okumalar üretebilen **üçüncü nesil dizileme (TGS)** yöntemlerinin geliştirilmesine neden olmuştur. TGS, aynı zamanda uzun okuma sıralaması olarak da bilinmektedir; çünkü birkaç kilo bazdan daha uzun okumalar üretebilmektedir. Tek moleküllü dizileme (SMS) ve gerçek zamanlı dizileme, TGS teknolojilerinin iki temel özelliği olarak tasarlanmıştır. En yaygın olarak kullanılan

TGS platformları, Pacific Biosciences (PacBio) tarafından **Tek Molekül Gerçek Zamanlı (SMRT) dizileme** ve Oxford Nanopore (ONT) tarafından **Nanopore dizileme**¹⁸.

Biyoenformatik ve NGS Pipeline

Teknolojik gelişmelerle birlikte oluşan NGS platformları sayesinde çok büyük miktarda veri ortaya çıkmaktadır. Artan veri miktarı ile

paralel olarak bu verilerin işlenmesi ve akışının kontrol edilmesi oldukça önemli kazanmıştır. Bu yüzden dizileme ile elde edilen veriler, **pipeline (boru hattı)** yani belirli bir iş akışı çerçevesinde ele alınmaya başlanmıştır^{13,19,20}. **Şekil 1**'de genel hatlarıyla NGS pipeline bileşenleri gösterilmiştir²⁰.



Şekil 1. NGS biyoenformatik pipeline genel iş akışı²⁰. Her bir NGS platformu farklı algoritmalar ve ek adımlar içerebilmektedir. *BAM, ikili hizalama haritası; **SNV, tek nükleotid varyasyonu; ***Indel, insersiyon-delesyon; ****VCF, varyant çağrı formatı; *****HGVS, İnsan Genom Varyasyon Topluluğu.

Ortaya çıkan bu verileri analiz etmek ve gerekli modülleri geliştirmek biyoenformatik alanının konusu haline gelmiştir. Biyoenformatik yaklaşımla, bilimsel araştırmalarda elde edilen verilerin gerekli uygulamalar ile analizlerinden başarılı sonuçlar alınmaya başlanmıştır; ayrıca analiz verilerinin gelecek araştırmalara uygunluğunu sağlamak amacıyla oluşturulan iş akışlarının ve analizlerde kullanılacak modüllerin tasarlanmasında işletim sisteminin önemi ortaya çıkmıştır. Bundan dolayı biyoenformatik analizlerde **Linux** gibi açık kaynak kodlu ve üzerinde işlem yapmaya müsait işletim sistemleri kullanılmaktadır. Ek olarak, analiz amaçlı modüllerin tasarımında kullanılacak programlama dilleri de araştırmalara katkı sağlamaktadır. Günümüzde veri işlemede öne çıkmakta olan “**R**” ve “**Python**” gibi nesne yönelimli ve geliştirilmesi basit diller kullanılmaktadır. Bu programlama dillerinin avantajı, veri kütüphaneleri ile uyumlu çalışarak istatistik gibi alanlarla beraber kullanılabilmesidir²¹⁻²³. Python programlama diliyle kolay bir şekilde oluşturulan nesne yönelimli algoritmalar ile derin öğrenme ve yapay zekâ alanlarında çalışmalar yapılabilir; ayrıca **Pytorch** gibi kütüphaneleri desteklemesiyle temel düzeyde programlama bilen araştırmacılara büyük avantaj sağlamaktadır²⁴⁻²⁶. NGS platformları sayesinde elde edilen verilerde diğer programlama yöntemleri de kullanılabilir; ancak kullanım ve veri

kütüphaneleri ile uyumundaki zorluktan dolayı günümüzde pek tercih edilmemektedir^{22,27}.

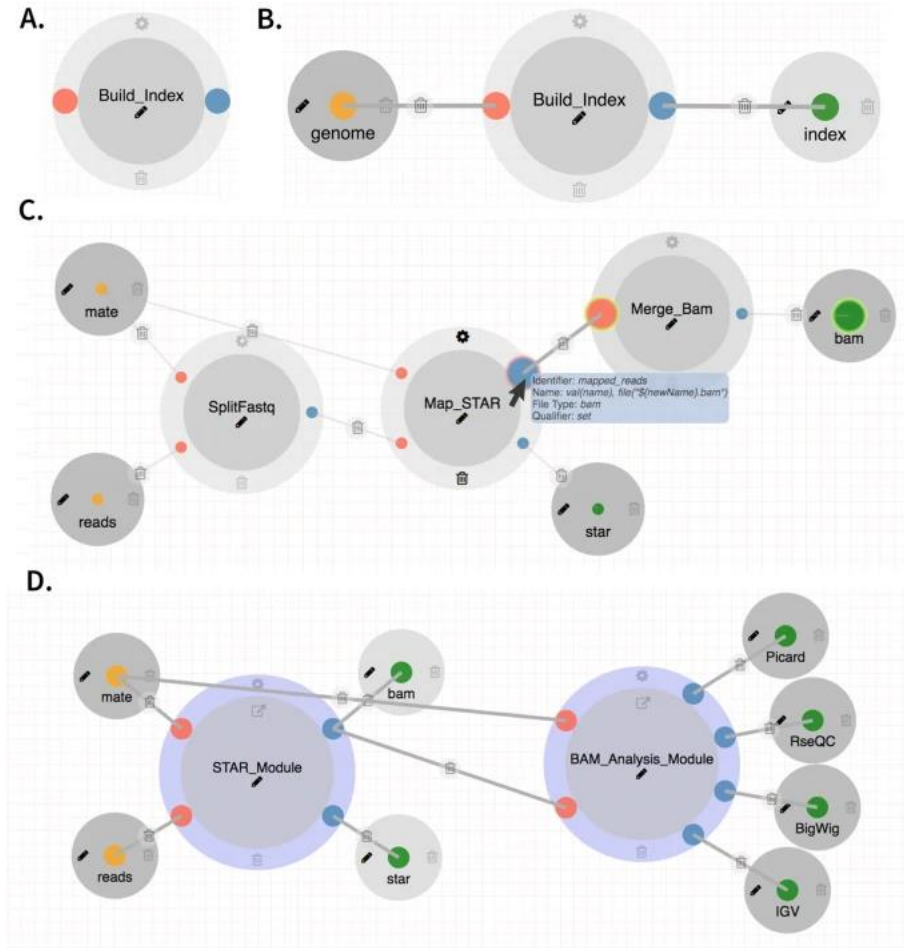
Her gün milyonlarca farklı dizilemenin yapıldığı göz önüne alındığında bütün verileri her seferinde baştan incelemek büyük bir zaman ve emek kaybına sebep olmaktadır. Bu yüzden NGS analizleri ile elde edilen verilerde biyoenformatik araçların kullanılması önemlidir. Bu araçlardan bazıları günümüz NGS çalışmalarının çoğunda kullanılmaktadır. **FastQC**, **Hisat2** ve **Cutadapt**, NGS çalışmalarında tercih edilen biyoenformatik araçlara örnektir²⁸. FastQC bir çeşit hata ayıklama ve inceleme modülüdür. NGS yardımıyla elde edilen ham verileri birden fazla kontrol işlemine sokarak her bir baz okuması için detaylı rapor oluşturulmasını sağlamaktadır²⁹. Hisat2 modülü ise NGS yardımıyla dizilemesi yapılan DNA/RNA örneğinin tek bir referans genomuna eşlemesini sağlayarak hızlı bir hizalama yapmaktadır. FastQ formatındaki NGS verilerini de paralel bir şekilde yüksek hız ve başarı ile işleyebilmesi günümüz hizalama modülleri arasında Hisat2’yi öne çıkarmaktadır³⁰. Cutadapt modülü de FastQC ile benzer bir amaçla kullanılmaktadır. Genel olarak NGS kapsamında elde edilen okumaların adaptör sekanslarını, primerlerini ve diğer istenmeyen sekans türlerini bulup uzaklaştırmakta işlevseldir³¹. Bu modül de Python ile rahatça kullanılabilir. Bu ve bunlara benzer birçok biyoenformatik araç da bulunmaktadır³².

NGS Pipeline Geliştirme

Çeşitlenen NGS yöntemleriyle birlikte ortaya çıkan devasa boyuttaki verilerin karmaşıklığı ile birlikte, ölçeklenebilirlik ve tekrarlanabilirlik sadece deneyler için değil, hesaplamalı analizler için de gerekli hale gelmiştir. Ek olarak, verileri bilgiye dönüştürmek, çok sayıda araç çalıştırmayı, parametreleri en iyi duruma getirmeyi ve dinamik olarak değişen başvuru verilerini tümleştirmeyi de içermektedir^{20,33}. **İş akışı yönetim sistemleri (WMS'ler)** bu tür zorluklara yanıt olarak geliştirilmiştir. WMS programlarının da kendi içlerinde bulunabilir, erişilebilir, birlikte çalışabilir ve yeniden üretilebilir bir protokole sahip olmaları gerekmektedir. Bu kriterler günümüzde, bir dizi hesaplama ortamında kod tanımlamak, paylaşmak ve yürütmek için kapsayıcı yazılımların ve standart **uygulama programlama arayüzünün (API)** gelişimini desteklemektedir³⁴⁻³⁹. **Nextflow**, yaygın olarak kullanılan bir WMS'dir; ancak özelinde bütün bu işlemleri basitleştiriyor olsa da yapılacak işlemin boyutunun ve karmaşıklığının artmasıyla oluşan bazı sınırlamalara sahiptir^{40,41}.

Grafik kullanıcı arayüzü (GUI), programlama bilgisi olmayanlar için NGS analizlerinin

gerçekleştirilmesinde avantaj sağlayan bir arabirimdir ve iyi tasarlanması gerekmektedir. Bazı kompleks NGS pipelineleri (Galaxy, GenePattern, GeneProf vb.) veri işleme ortamları arasında paralellik ve taşınabilirlik sağlayamamaktadır^{42,43}. Günümüz bulut sistemlerine uyumu ve eski nesil iş akış platformlarının yaşattığı sorunları gidermek üzere yakın zamanda başlatılan **Biocore: DolphinNext** projesi, araştırmacılara detaylı teknik bilgiye sahip olmadan da verilerini işleyebilecekleri bir ortam sunmaktadır. Kullanıcı dostu arayüzüyle birlikte pipeline oluşturulması ve kullanılacak modüllerin tasarlanması konusunda kolaylıklar sağlamaktadır; ayrıca tasarlanan pipeline ve modüllerin gelecek araştırmalara kaynak olabilmesi için tekrarlanabilir bir şekilde tüm aşamalar oluşturulabilmektedir. Bu durum bağımsız projelerin birbirlerinin modüllerinden girdi (input) ve çıktı (output) verilerinin kullanılabilmesine ve araştırmaların hız kazanmasına neden olmaktadır. DolphinNext'in benzer bir platform olan NextFlow'dan en büyük farkı, kullanıcıların programlama dillerini detaylı bilmesine gerek kalmadan **“sürükle bırak”** gibi basit işlemlerle istedikleri iş akış modelini oluşturabilmesidir (Şekil 2)⁴⁴.



Şekil 2. Biocore: DolphinNext platformunda oluşturulan örnek bir pipeline tasarımı⁴⁴. A. Düzenli bir hesaplama için verilerden indeks oluşturmaktadır. B. Genom dizileme verileri girdi olarak kullanılmaktadır. Veriler, indeks oluşturma modülünde işlendikten sonra indeks çıktısı (yeşil renkle belirtilen) alınmaktadır. C. Platform için verisi kullanılan Mate modülü, örnek miRNA listesini sağlamaktadır⁴⁵ ve “reads” ile dizilemeden okumalar alınmaktadır. Bu iki girdi Fastq modülü ile kalite puanlarıyla beraber Fastq formatına çevrilmektedir. STAR modülü girdileri düzenleyerek hizalama yapmaktadır. BAM çıktısı ile genom dizilemesinin kapsamlı ham verileri ve sıkıştırılmış halini ikili şekilde kaydetmektedir. D. Biyoenformatik inceleme sonucunda alınacak veriler çeşitlendirilerek çalışma metoduna göre çıktı modülleri eklenmektedir ve böylece RNA dizileme (RNA-Seq) pipeline oluşturulmaktadır.

Biyoenformatik pipelinelerde kullanıcı arabirimi, veri analizinde programlama dillerine hâkimiyet gerektiren dizileme verileri için her geçen gün daha da önemli bir özellik haline gelmektedir. DolphinNext, NGS pipeline komutları ve büyük verileri arasında iş akışının oluşturulması, ilgili modüllere dağıtımı ve kullanıcı dostu arayüzü ile veri analizinde destekleyici rol üstlenen bir veri işleme enstrümanıdır⁴⁶. Ek olarak, DolphinNext'te çalışmaların bulut sistemleri ve Docker veri

sanallaştırma aracı ile sunuculara erişimin sağlanarak yapılabilmesi, araştırmaların hızlı ve kolayca tekrarlanabilmesini sağlamaktadır; bundan dolayı da teknik olarak büyük bir desteğe sahip olmayan projelere de düşük kaynakla en iyi sonucu almalarında yardımcı olmaktadır^{44,47}.

Yaklaşık 20 yıl önce İnsan Genom Projesi'nin tamamlanması ile kapılarını açan genom çağı, günümüzde teknoloji ile birlikte hızla gelişen NGS platformları sayesinde insan dahil pek

çok organizma hakkında genetik bilginin en ince ayrıntılarına erişim imkânı sağlamıştır. Uçsuz bucaksız veri havuzundaki bilgilerin işlenmesi için programlama dilleri, hesaplamalı analizler ve biyolojik araştırmalar arasında iş birliği sağlayan biyoenformatik yaklaşımlar NGS pipelinelarının ortaya çıkmasına ve verilerin zamanla daha da karmaşık hale gelmesiyle bu pipelineların sürekli geliştirilmesine zemin hazırlamıştır. Hali hazırda kullanıma sunulmuş olan NGS pipelinelarının geliştirilmesi gereken yönlerinden biri, genetik bilgiye ulaşan biyoloji

temelli bilim insanları ile bu bilgiyi analiz eden biyoenformatik çalışan insanlara duyulan ihtiyacın bir noktada bütünleyici bir yaklaşımla değerlendirilmesi gerekliliğidir. Birkaç yıl önce geliştirilen DolphinNext, hesaplamalı analizleri deneysel araştırmacılar için tekrarlanabilir, erişilebilir ve yorumlayabilir hale getirmiştir. Benzeri araçların artması, ileride veri analizlerindeki hata oranını azaltacağı gibi özellikle hastalıklar bazında daha fazla araştırma yapılmasına da olanak sağlayacaktır.

BIOINFORANGE

Referanslar:

1. Hood, L., & Rowen, L. (2013). The human genome project: Big science transforms biology and medicine. *Genome Medicine*, 5(9), 1–8. <https://doi.org/10.1186/GM483/METRICS>
2. Cooke Bailey, J. N., Pericak-Vance, M. A., & Haines, J. L. (2014). The Impact of the Human Genome Project on Complex Disease. *Genes*, 5(3), 518. <https://doi.org/10.3390/GENES5030518>
3. Rood, J. E., & Regev, A. (2021). The legacy of the Human Genome Project. *Science*, 373(6562), 1442–1443. <https://doi.org/10.1126/SCIENCE.ABL5403>
4. Giani, A. M., Gallo, G. R., Gianfranceschi, L., & Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, 18, 9–19. <https://doi.org/10.1016/J.CSBJ.2019.11.002>
5. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012. <https://doi.org/10.1155/2012/251364>
6. Mu, W., Lu, H. M., Chen, J., Li, S., & Elliott, A. M. (2016). Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. *The Journal of Molecular Diagnostics*, 18(6), 923–932. <https://doi.org/10.1016/J.JMOLDX.2016.07.006>
7. Artech-López, A., Ávila-Fernández, A., Romero, R., Riveiro-Álvarez, R., López-Martínez, M. A., Giménez-Pardo, A., Vélez-Monsalve, C., Gallego-Merlo, J., García-Vara, I., Almoguera, B., Bustamante-Aragón, A., Blanco-Kelly, F., Tahsin-Swafiri, S., Rodríguez-Pinilla, E., Minguez, P., Lorda, I., Trujillo-Tiebas, M. J., & Ayuso, C. (2021). Sanger sequencing is no longer always necessary based on a single-center validation of 1109 NGS variants in 825 clinical exomes. *Scientific Reports* 2021 11:1, 11(1), 1–7. <https://doi.org/10.1038/s41598-021-85182-w>
8. Totomoch-Serra, A., Marquez, M. F., & Cervantes-Barragán, D. E. (2017). Sanger sequencing as a first-line approach for molecular diagnosis of Andersen-Tawil syndrome. *F1000Research*, 6. <https://doi.org/10.12688/F1000RESEARCH.11610.1>
9. Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 2016 17:6, 17(6), 333–351. <https://doi.org/10.1038/nrg.2016.49>
10. Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New Biotechnology*, 25(4), 195–203. <https://doi.org/10.1016/J.NBT.2008.12.009>
11. Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10), 1135–1145. <https://doi.org/10.1038/NBT1486>
12. Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1), 1. <https://doi.org/10.1016/J.YGENO.2015.11.003>
13. Di Resta, C., Galbiati, S., Carrera, P., & Ferrari, M. (2018). Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities. *EJIFCC*, 29(1), 4. <http://www.ncbi.nlm.nih.gov/pmc/articles/pmc5949614/>
14. Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10), 1932–1941. <https://doi.org/10.1016/J.BBADI.2014.06.015>
15. Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N., & Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 39(13), e90. <https://doi.org/10.1093/NAR/GK1344>
16. Merriman, B., Torrent, I., & Rothberg, J. M. (2012). Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis*, 33(23), 3397–3417. <https://doi.org/10.1002/ELPS.201200424>
17. Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13(1), 341. <https://doi.org/10.1186/1471-2164-13-341>
18. Xiao, T., & Zhou, W. (2020). The third generation sequencing: the advanced approach to genetic diseases.

- Translational Pediatrics, 9(2), 163.
<https://doi.org/10.21037/TP.2020.03.06>
19. Leipzig, J. (2017). A review of bioinformatic pipeline frameworks. Briefings in Bioinformatics, 18(3), 530–536.
<https://doi.org/10.1093/BIB/BBW020>
 20. Roy, S., Coldren, C., Karunamurthy, A., Kip, N. S., Klee, E. W., Lincoln, S. E., Leon, A., Pullambhatla, M., Temple-Smolkin, R. L., Voelkerding, K. V., Wang, C., & Carter, A. B. (2018). Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. The Journal of Molecular Diagnostics : JMD, 20(1), 4–27.
<https://doi.org/10.1016/J.JMOLDX.2017.11.003>
 21. Lee, D. J., Kwon, T., Kim, C. K., Seol, Y. J., Park, D. S., Lee, T. H., & Ahn, B. O. (2020). NGS_SNPAnalyzer: a desktop software supporting genome projects by identifying and visualizing sequence variations from next-generation sequencing data. Genes & Genomics, 42(11), 1311.
<https://doi.org/10.1007/S13258-020-00997-7>
 22. Shajii, A., Numanagić, I., Leighton, A. T., Greenyer, H., Amarasinghe, S., & Berger, B. (2021). A Python-based programming language for high-performance computational genomics. Nature Biotechnology, 39(9), 1062.
<https://doi.org/10.1038/S41587-021-00985-6>
 23. Sepulveda, J. L. (2020). Using R and Bioconductor in Clinical Genomics and Transcriptomics. The Journal of Molecular Diagnostics : JMD, 22(1), 3–20.
<https://doi.org/10.1016/J.JMOLDX.2019.08.006>
 24. Pellegrino, E., Jacques, C., Beaufile, N., Nanni, I., Carlioz, A., Metellus, P., & Ouafik, L. H. (2021). Machine learning random forest for predicting oncosomatic variant NGS analysis. Scientific Reports 2021 11:1, 11(1), 1–14.
<https://doi.org/10.1038/s41598-021-01253-y>
 25. Schmidt, B., & Hildebrandt, A. (2021). Deep learning in next-generation sequencing. Drug Discovery Today, 26(1), 173.
<https://doi.org/10.1016/J.DRUDIS.2020.10.002>
 26. Ekmekci, B., McAnany, C. E., & Mura, C. (2016). An Introduction to Programming for Bioscientists: A Python-Based Primer. PLoS Computational Biology, 12(6).
<https://doi.org/10.1371/JOURNAL.PCBI.1004867>
 27. Perkel, J. M. (2020). Why scientists are turning to Rust. Nature, 588(7836), 185–186.
<https://doi.org/10.1038/D41586-020-03382-2>
 28. Kuznetsova, I., Karpievitch, Y. V., Filipovska, A., Lugmayr, A., & Holzinger, A. (2017). Review of Machine Learning Algorithms in Differential Expression Analysis. International Series on Information Systems and Management in Creative EMedia , 2016(2), 11–24.
<https://doi.org/10.48550/arxiv.1707.09837>
 29. Mohideen, A. M. S. H., Johansen, S. D., & Babiak, I. (2020). High-Throughput Identification of Adapters in Single-Read Sequencing Data. Biomolecules, 10(6), 1–12.
<https://doi.org/10.3390/BIOM10060878>
 30. Guo, J., Gao, J., & Liu, Z. (2022). HISAT2 Parallelization Method Based on Spark Cluster. Journal of Physics: Conference Series, 2179(1), 012038.
<https://doi.org/10.1088/1742-6596/2179/1/012038>
 31. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, 17(1), 10–12.
<https://doi.org/10.14806/EJ.17.1.200>
 32. Didion, J. P., Martin, M., & Collins, F. S. (2017). Atropos: Specific, sensitive, and speedy trimming of sequencing reads. PeerJ, 2017(8).
<https://doi.org/10.7717/PEERJ.3720/SUPP-1>
 33. Federico, A., Karagiannis, T., Karri, K., Kishore, Di., Koga, Y., Campbell, J. D., & Monti, S. (2019). Pipeliner: A nextflow-based framework for the definition of sequencing data processing pipelines. Frontiers in Genetics, 10(JUN), 614.
<https://doi.org/10.3389/FGENE.2019.00614/BIBTEX>
 34. Bianchi, V., Ceol, A., Ogier, A. G. E., De Pretis, S., Galeota, E., Kishore, K., Bora, P., Croci, O., Campaner, S., Amati, B., Morelli, M. J., & Pelizzola, M. (2016). Integrated Systems for NGS Data Management and Analysis: Open Issues and Available Solutions. Frontiers in Genetics, 7(MAY).
<https://doi.org/10.3389/FGENE.2016.00075>
 35. Liu, J., Pacitti, E., Valduriez, P., & Mattoso, M. (2015). A Survey of Data-Intensive Scientific Workflow Management. Journal of Grid Computing 2015 13:4, 13(4), 457–493.
<https://doi.org/10.1007/S10723-015-9329-8>
 36. Fjukstad, B., & Bongo, L. A. (2017). A Review of Scalable Bioinformatics Pipelines. Data Science and Engineering, 2(3), 245–251.
<https://doi.org/10.1007/S41019-017-0047-Z/TABLES/1>
 37. Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak,

- A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 2016 3:1, 3(1), 1–9. <https://doi.org/10.1038/sdata.2016.18>
38. Schulz, W. L., Durant, T. J. S., Siddon, A. J., & Torres, R. (2016). Use of application containers and workflows for genomic data analysis. *Journal of Pathology Informatics*, 7(1), 53. <https://doi.org/10.4103/2153-3539.197197>
39. Birney, E., Vamathevan, J., & Goodhand, P. (2017). Genomics in healthcare: GA4GH looks to 2022. *BioRxiv*, 203554. <https://doi.org/10.1101/203554>
40. DI Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4), 316–319. <https://doi.org/10.1038/NBT.3820>
41. Jackson, M., Kavoussanakis, K., & Wallace, E. W. J. (2021). Using prototyping to choose a bioinformatics workflow management system. *PLoS Computational Biology*, 17(2). <https://doi.org/10.1371/JOURNAL.PCBI.1008622>
42. Milicchio, F., Rose, R., Bian, J., Min, J., & Prosperi, M. (2016). Visual programming for next-generation sequencing data analytics. *BioData Mining*, 9(1). <https://doi.org/10.1186/S13040-016-0095-3>
43. DolphinNext: A Graphical User Interface for Distributed Data Processing of High Throughput Genomics. (n.d.-aq). Retrieved March 6, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6938066/>
44. Yukselen, O., Turkyilmaz, O., Ozturk, A. R., Garber, M., & Kucukural, A. (2020). DolphinNext: A distributed data processing platform for high throughput genomics. *BMC Genomics*, 21(1), 1–16. <https://doi.org/10.1186/S12864-020-6714-X/TABLES/1>
45. Yousef, M., Abdallah, L., & Allmer, J. (2019). maTE: discovering expressed interactions between microRNAs and their targets. *Bioinformatics*, 35(20), 4020–4028. <https://doi.org/10.1093/BIOINFORMATICS/BTZ204>
46. Ahmed, A. E., Allen, J. M., Bhat, T., Burra, P., Fliege, C. E., Hart, S. N., Heldenbrand, J. R., Hudson, M. E., Istanto, D. D., Kalmbach, M. T., Kapraun, G. D., Kendig, K. I., Kendzior, M. C., Klee, E. W., Mattson, N., Ross, C. A., Sharif, S. M., Venkatakrishnan, R., Fadlelmola, F. M., & Mainzer, L. S. (2021). Design considerations for workflow management systems use in production genomics research and the clinic. *Scientific Reports* 2021 11:1, 11(1), 1–18. <https://doi.org/10.1038/s41598-021-99288-8>
47. Pittard, W. S., & Li, S. (2020). The Essential Toolbox of Data Science: Python, R, Git, and Docker. *Methods in Molecular Biology* (Clifton, N.J.), 2104, 265–311. https://doi.org/10.1007/978-1-0716-0239-3_15