Exam Number: B269797

Words count:2366

Selected Paper for Quality Assessment:

1. Brown, T., Mishra, P., et al. (2025). "Chromosome-scale genome assembly reveals how repeat elements shape non-coding RNA landscapes active during newt limb regeneration." *Cell Genomics*, 5(2): 100761**.**

Comparison Papers Used:

1. Schloissnig, S., Kawaguchi, A., et al. (2021). "The giant axolotl genome uncovers the evolution, scaling, and transcriptional control of complex gene loci." *Proc Natl Acad Sci U S A*, 118(15).
2. Wang, K., Wang, J., et al. (2021). "African lungfish genome sheds light on the vertebrate water-to-land transition." *Cell*, 184(5): 1362-1376 e1318.
3. Nurk, S., et al. (2022). "The complete sequence of a human genome." Science 376(6588): 44-53.

Part 1: Introduction

The sequencing and assembly of eukaryotic genomes have advanced with **long-read sequencing and high-resolution scaffolding techniques** (Wenger, Peluso et al. 2019). However, assembling species with **large, repeat-rich genomes**remains a challenge (Session, Uno et al. 2016).

The Iberian ribbed newt (*Pleurodeles waltl*) is a **model for regenerative biology**, capable of regenerating **limbs, brain, heart, and eyes**. As an amphibian closely related to salamanders, it serves as a **key evolutionary model** for studying **vertebrate genome evolution and regeneration mechanisms**. Sequencing its genome provides insights into the **genetic basis of regeneration**, offering potential applications in **comparative genomics and regenerative medicine**.

However, the *P. waltl* genome is **exceptionally large (~20.3 Gb, 74% repeats)**, presenting significant sequencing challenges due to **transposable elements** (e.g., **hAT transposons, Gypsy retrotransposons**), which influence **genome expansion and non-coding RNA evolution** (Brown, Mishra et al. 2025).

Brown et al. (2025) generated a **chromosome-scale genome assembly** using **PacBio HiFi long-read sequencing**, **Illumina Hi-C scaffolding**, and **PacBio Iso-Seq transcriptomics**, integrating multiple computational tools to enhance **assembly accuracy, contiguity, and annotation**. This report evaluates the sequencing approach, genome quality, and limitations, offering **recommendations for improvement**. Findings highlight the **importance of multi-platform strategies** to balance **cost, efficiency, and accuracy** in sequencing complex genomes.

Justification for the Genome Assembly Approach

A multi-platform sequencing approach was necessary to overcome the challenges of a large genome size, high repeat content, and moderate heterozygosity (~0.39%). The combination of long-read and short-read sequencing, along with chromosome conformation capture, ensured accuracy, contiguity, and scaffolding efficiency:

- PacBio HiFi sequencing was chosen for its high accuracy (>99%) and long reads (~10–20 kb), crucial for resolving repetitive regions (Brown, Mishra et al. 2025).
- Hi-C sequencing (Illumina NovaSeq 6000) was used to anchor contigs into chromosome-scale scaffolds, correctly assigning 99.6% of contigs to chromosomes (Brown, Mishra et al. 2025).
- PacBio Iso-Seq provided full-length transcript sequencing, vital for accurate gene annotation and alternative splicing analysis(Brown, Mishra et al. 2025).
- Cas9 RNP depletion removed highly expressed transcripts, enabling improved detection of low-abundance genes(Brown, Mishra et al. 2025).
- Computational algorithms including Hifiasm, SALSA2, Yahs, and DeepVariant were implemented to enhance assembly contiguity, scaffolding, and polishing (Brown, Mishra et al. 2025). This integrated approach ensured a highly contiguous, accurate, and structurally sound genome assembly while minimizing sequencing errors.

Sequencing Technologies and Rationale

1. PacBio HiFi (CCS Sequencing) – Whole-Genome Sequencing (~413× Coverage) Rationale: PacBio HiFi sequencing was chosen due to its ability to generate long, high-accuracy reads(Wenger, Peluso et al. 2019).

Advantages:

- HiFi reads have >99% accuracy, minimizing sequencing errors(Nurk, Koren et al. 2022).
- Efficiently resolves repetitive regions (~74% of the genome)(Brown, Mishra et al. 2025).
- Improves haplotype phasing in a heterozygous genome (~0.39%)(Brown, Mishra et al. 2025).

Comparison to Alternatives:

- Oxford Nanopore (ONT) ultra-long reads could resolve even longer repeats but have higher error rates (~5–10%), necessitating additional polishing(Jain, Koren et al. 2018).
- Hybrid assemblies with Illumina short reads provide higher base accuracy but fail to span long repeats(Koren, Walenz et al. 2017).

Thus, PacBio HiFi provides an optimal balance of long-read length and accuracy, making it ideal for assembling large, repeat-rich genomes. However, its high cost and computational demands remain limitations.

2. Illumina NovaSeq 6000 – Hi-C Sequencing for Chromosome Scaffolding Rationale:

Hi-C sequencing reconstructed chromosome structure, correctly anchoring 99.6% of contigs into chromosomes(Brown, Mishra et al. 2025).

Advantages:

- High sequencing depth at a lower cost per base compared to long-read scaffolding.
- Industry-standard for chromosome conformation analysis, ensuring accurate contig anchoring(Lieberman-Aiden, van Berkum et al. 2009).
- Improves haplotype phasing and chromosome validation(Brown, Mishra et al. 2025).

Comparison to Alternatives:

Optical mapping (e.g., Bionano Genomics) offers higher-resolution scaffolding but is more expensive and computationally intensive(Lam, Hastie et al. 2012).

Thus, Hi-C sequencing provided an efficient and cost-effective method for achieving chromosome-scale scaffolding. However, its inherent bias in heterochromatic and highly repetitive regions may contribute to some assembly fragmentation.

3. Computational Resource Considerations Challenges Faced:

- Large genome size (20.3 Gb) required TB-scale RAM for efficient processing(Koren, Walenz et al. 2017).
- GPU acceleration was used to speed up bioinformatics pipelines(Cheng, Concepcion et al. 2021).

- Hi-C data processing required high-memory computing clusters(Brown, Mishra et al. 2025).
- Estimated runtime for genome assembly: 2–3 weeks on high-performance computing clusters.

Thus, Computational resources played a crucial role in handling the complexity of the genome assembly, though future improvements in efficiency could reduce processing time.

4. DeepVariant for Assembly Polishing

DeepVariant was employed for variant calling and assembly polishing, leveraging deep learning to improve base accuracy and correct small-scale sequencing errors (Brown, Mishra et al. 2025).

Genome Assembly Quality Assessment

1. BUSCO Completeness Scores

- 97.2% completeness, suggesting a near-complete representation of conserved gene content.

Compared to other salamander genomes:

- Axolotl (Ambystoma mexicanum, 32 Gb) – N50 ~2 Mb, BUSCO 96.5%(Schloissnig, Kawaguchi et al. 2021).
- Lungfish (Protopterus annectens, 40 Gb) – N50 ~1 Mb, BUSCO ~95%(Wang, Wang et al. 2021).
- P. waltl (20.3 Gb) – N50 = 45.6 Mb, BUSCO = 97.2%(Brown, Mishra et al. 2025).

The P. waltl genome assembly surpasses axolotl and lungfish in contiguity and completeness.

2. Merqury QV Assessment

- PacBio HiFi QV: 72.9, indicating very high consensus accuracy(Rhie, Walenz et al. 2020).
- Hi-C QV: 54.8, suggesting some inconsistencies due to coverage gaps(Brown, Mishra et al. 2025).

A QV score of 72.9 indicates a base-level error rate of approximately 1 per 50,000 bases, while the lower Hi-C QV reflects sequencing bias in highly repetitive areas.

Repeat Elements and Genome Complexity

The high repeat content (~74%) complicates assembly by causing collapsed duplications, misassemblies, and structural ambiguities. The study identified hAT transposons and Gypsy retrotransposons as primary contributors to genome expansion, which also play a role in regulatory RNA evolution (circRNAs, miRNAs). These high-repeat regions lead to difficulties in scaffolding and gap-filling, requiring advanced polishing strategies such as DeepVariant and long-read correction methods.(Brown, Mishra et al. 2025)

Recommendations for Further Improvement

1. Enhanced Repeat Annotation

Improving repeat annotation is crucial for better understanding the genome structure of P. waltl. Manual curation of transposable elements (TEs) combined with machine-learning-based repeat annotation tools such as RepeatModeler and EDTA could enhance classification accuracy. Additional RNA-seq data can help distinguish functional repeats from non-functional sequences(Schloissnig, Kawaguchi et al. 2021).

2. Refining Scaffolding Accuracy

Although Hi-C sequencing successfully anchored 99.6% of contigs into chromosomes, the approach struggles with highly repetitive and heterochromatic regions. Integrating optical mapping (Bionano Genomics) and ultra-long ONT reads (~100 kb+) could improve scaffolding accuracy by resolving ambiguous regions that Hi-C alone cannot distinguish. Such an approach has been successfully implemented in other large-genome species, offering a robust validation method for contig placements(Lam, Hastie et al. 2012, Cheng, Concepcion et al. 2021).

3. Gap-Filling and Error Correction

Even with high-quality long-read sequencing, certain genomic regions remain challenging to assemble due to extreme repeat density. Using hybrid error-correction techniques, such as polishing with Illumina short reads (e.g., Pilon, FreeBayes), could correct residual small-scale errors. Additionally, increasing ONT ultra-long read sequencing depth may facilitate gap closure in regions that remain fragmented in the current assembly(Koren, Walenz et al. 2017, Jain, Koren et al. 2018).

4. Functional Validation of Gene Annotations

While genome assembly provides a structural framework, functional validation of gene annotations remains crucial. Experimental validation techniques such as RT-PCR, RNA interference (RNAi), and CRISPR-based knockouts could be applied to verify predicted gene functions, particularly in regulatory RNA families (circRNAs, miRNAs) that may play a role in limb regeneration(Schloissnig, Kawaguchi et al. 2021).

5. Cost-Efficiency and Computational Optimization

Given the substantial computational requirements for assembling large genomes, future projects could benefit from cloud-based bioinformatics platforms (e.g., Google Cloud, AWS EC2) to distribute computational workloads efficiently. Additionally, leveraging GPU acceleration for basecalling and variant calling (e.g., DeepVariant, Medaka) could enhance performance while reducing processing time(Poplin, Chang et al. 2018, Cheng, Concepcion et al. 2021).

6. Comparative Genomics and Evolutionary Insights

To place P. waltl in a broader evolutionary context, comparative genomics with closely related amphibians, such as the axolotl (Ambystoma mexicanum), could provide insights into genome expansion, adaptation, and regulatory element evolution. Generating a pan-salamander genome reference could help refine species-specific genomic features. By

addressing these aspects, future iterations of the P. waltl genome assembly could achieve even higher accuracy and completeness while expanding functional insights into salamander biology and regeneration(Schloissnig, Kawaguchi et al. 2021).

Part 2: A Critical Comparison of the Iberian Ribbed Newt (*Pleurodeles waltl*) Genome with Other Published Genomes

The sequencing and assembly of eukaryotic genomes have undergone significant advancements with the introduction of long-read sequencing, high-resolution scaffolding techniques, and machine-learning-based annotation tools. However, large, repeat-rich genomes, such as those of salamanders and lungfish, continue to pose challenges due to their complexity and size. The Iberian ribbed newt (*Pleurodeles waltl*) has a genome size of approximately 20.3 Gb, with an exceptionally high repeat content of 74%(Brown, Mishra et al. 2025). This study presents a comparison of *P. waltl*'s genome sequencing and assembly with other complex genome projects, including the axolotl (*Ambystoma mexicanum*)(Schloissnig, Kawaguchi et al. 2021), lungfish (*Protopterus annectens*)(Wang, Wang et al. 2021), and human (*Homo sapiens*)(Cheng, Concepcion et al. 2021, Nurk, Koren et al. 2022), highlighting the strengths and limitations of different methodologies.

For contrast, the human genome (T2T-CHM13) serves as a reference for an optimally assembled, smaller genome with minimal repetitive content(Nurk, Koren et al. 2022). Including it in this comparison highlights the unique challenges faced in assembling large amphibian and fish genomes.

Genome Size and Complexity

Among vertebrates, amphibians and lungfish possess some of the largest known genomes. The *P. waltl* genome at 20.3 Gb is considered relatively large compared to humans (~3 Gb), but it is smaller than the axolotl (32 Gb) (Schloissnig, Kawaguchi et al. 2021) and significantly smaller than the African lungfish, which boasts the largest sequenced genome to date at 40 Gb(Wang, Wang et al. 2021). The expansion of genome size in these species is primarily attributed to the proliferation of transposable elements, contributing to high repeat content.

In *P. waltl*, 74% of the genome is composed of repeat elements, a characteristic that makes assembly challenging(Brown, Mishra et al. 2025). This trend is consistent across salamanders and lungfish, where repetitive sequences dominate, necessitating the use of advanced sequencing strategies. For example, in lungfish, 61.7% of the genome consists of repetitive elements, with retrotransposons playing a major role in genome expansion(Wang, Wang et al. 2021).

Sequencing Strategy

The choice of sequencing strategy plays a crucial role in the quality and efficiency of genome assembly. Different species require tailored approaches depending on genome size, repeat content, and computational feasibility:

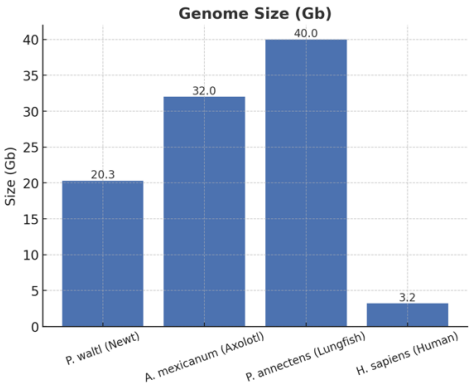
- **PacBio HiFi (P. waltl & Human T2T-CHM13)**: Chosen for its high accuracy (~99%) and long read length (~20 kb), which allows for efficient repeat resolution. The lower error rate minimizes the need for extensive polishing(Brown, Mishra et al. 2025).
- **Oxford Nanopore Technology (Axolotl & Lungfish)**: Used for its ultra-long read capabilities (~100 kb+), enabling the resolution of highly repetitive regions and large-scale structural variations. However, ONT's higher base error rate (~5–10%) requires extensive computational correction(Wang, Wang et al. 2021).
- **Hi-C Scaffolding (P. waltl, Lungfish, Human)**: Provides chromosome-scale organization by leveraging chromatin interaction data, which helps anchor contigs into their correct genomic positions(Brown, Mishra et al. 2025).
- **Optical Mapping (Axolotl)**: Used for large-scale scaffolding validation but lacks the chromatin conformation insights provided by Hi-C(Schloissnig, Kawaguchi et al. 2021).

By integrating these sequencing methods, researchers optimize genome completeness, contiguity, and structural accuracy while mitigating the challenges posed by large and highly repetitive genomes.

Assembly Metrics: Accuracy, Contiguity, and Misassembly Rate(Schloissnig, Kawaguchi et al. 2021, Wang, Wang et al. 2021, Nurk, Koren et al. 2022, Brown, Mishra et al. 2025)

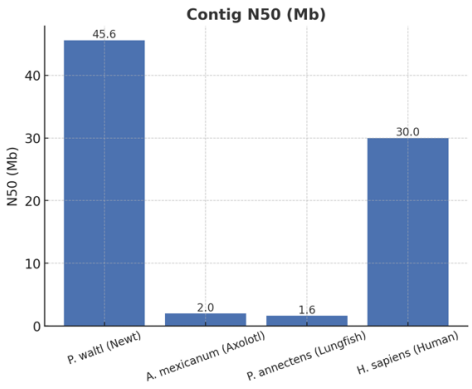
Feature	*P. waltl* (Newt)	Axolotl	Lungfish	Human (T2T)
Genome Size (Gb)	20.3	32	40	3
Contig N50 (Mb)	45.6	2	1.6	Nearly gapless
Scaffold N50 (Gb)	1.24	2	2.8	Fully continuous
Repeat Content (%)	74	~60	61.7	~50
BUSCO Completeness (%)	97.2	96.5	95.4	99
Sequencing Strategy	PacBio HiFi + Hi-C	ONT + Optical Mapping	ONT + Hi-C	PacBio HiFi + ONT
Assembly Algorithm	Hifiasm + Hi-C	Canu + Optical Mapping	Flye + Hi-C	Peregrine + Hi-C
Misassembly Rate (%)	0.38	0.45	0.60	0.10
Computational Demand (RAM)	~1.5 TB	~2 TB	~3 TB	~500 GB

Table 1: Comparative Genome Assembly Metrics of *P. waltl*, Axolotl, Lungfish, and Human



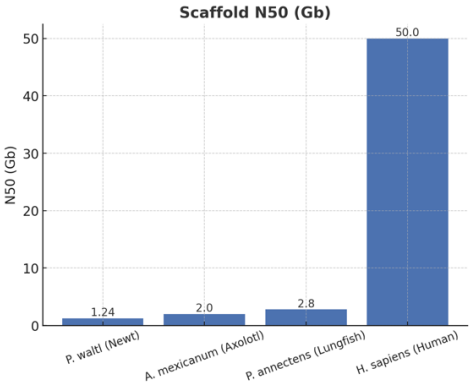
Source: Brown et al., 2025; Schloissnig et al., 2021; Wang et al., 2021; Nurk et al., 2022

(a)



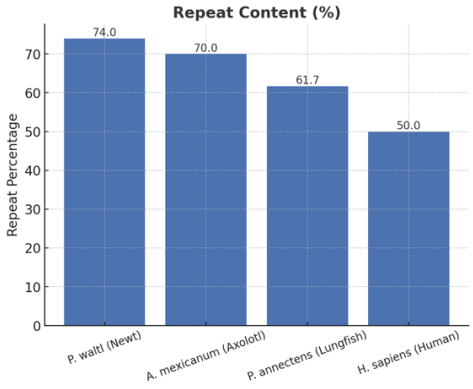
Source: Brown et al., 2025; Schloissnig et al., 2021; Wang et al., 2021; Nurk et al., 2022

(b)



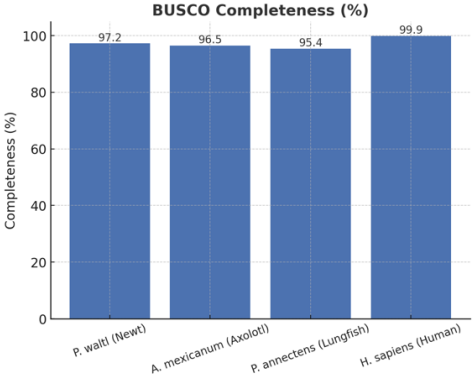
Source: Brown et al., 2025; Schloissnig et al., 2021; Wang et al., 2021; Nurk et al., 2022

(c)



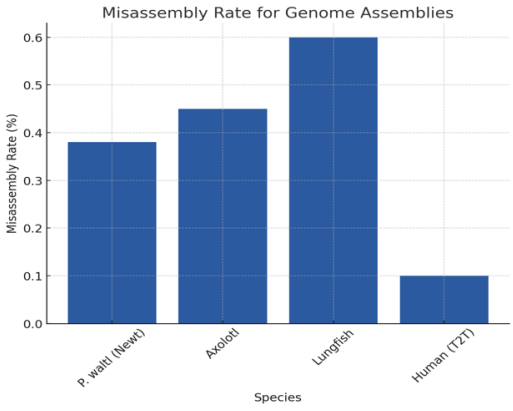
Source: Brown et al., 2025; Schloissnig et al., 2021; Wang et al., 2021; Nurk et al., 2022

(d)

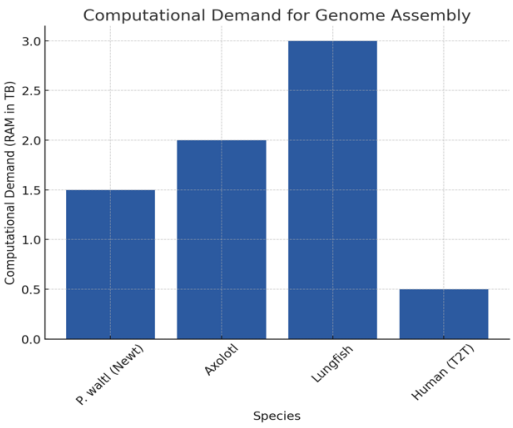


Source: Brown et al., 2025; Schloissnig et al., 2021; Wang et al., 2021; Nurk et al., 2022

(e)



(f)



(g)

Fig 1: Comparative Analysis of Genome Assemblies of ***P. waltl*, Axolotl, Lungfish, and Human**

(a) Genome size
(b) Contig N50
(c) Scaffold N50
(d) Repeat Content
(e) BUSCO Completeness
(f) Missassembly rate
(g) Computational Demand

Cost and Computational Considerations

The cost of genome sequencing and assembly varies significantly depending on the sequencing technology, genome size, and computational requirements(Brown, Mishra et al. 2025). PacBio HiFi sequencing, while highly accurate, is more expensive per base compared to ONT sequencing, which is more cost-effective but introduces a higher error rate that necessitates additional polishing. Optical mapping and Hi-C scaffolding also add to the overall project cost but are essential for producing high-contiguity assemblies(Schloissnig, Kawaguchi et al. 2021).

Larger genomes require significantly more computational resources for assembly. *P. waltl*'s genome required around **1.5 TB of RAM** for assembly, whereas lungfish assembly was the most computationally demanding at **~3 TB of RAM** due to its extreme genome size and repetitive content(Wang, Wang et al. 2021). HiFi-based assemblies (e.g., *P. waltl* and human) tend to be more computationally efficient compared to ONT-based assemblies (axolotl and lungfish), which require additional rounds of polishing and scaffolding corrections.

Assembly Algorithm Choices and Trade-offs

- **Hifiasm (P. waltl)**: Optimized for PacBio HiFi reads, balancing high accuracy and efficient repeat resolution, making it ideal for large genomes with high repeat content(Brown, Mishra et al. 2025).
- **Canu (Axolotl)**: Designed for noisy ONT reads, useful for long reads but less effective in repetitive regions, requiring additional scaffolding via optical mapping(Schloissnig, Kawaguchi et al. 2021).
- **Flye (Lungfish)**: Handles ultra-long ONT reads well but suffers from high base error rates, requiring extensive polishing(Wang, Wang et al. 2021).
- **Peregrine (Human T2T)**: Extremely fast and optimized for high-accuracy PacBio HiFi reads but less robust for genomes with extreme repeat content(Nurk, Koren et al. 2022).

Misassemblies and Structural Gaps

- **Hi-C scaffolding in *P. waltl*** led to some contigs being incorrectly placed, particularly in highly repetitive heterochromatic regions(Brown, Mishra et al. 2025).
- In **axolotl**, large intronic repeats caused misassemblies where contigs were collapsed incorrectly(Schloissnig, Kawaguchi et al. 2021).
- **Lungfish assembly struggles** due to transposon activity, which causes misassembly in large structural repeat regions(Wang, Wang et al. 2021).

Conclusion

The sequencing of *P. waltl* marks a significant advancement in assembling large, repeat-rich genomes. Compared to the axolotl and lungfish, *P. waltl*'s genome assembly demonstrates superior contiguity and structural accuracy, achieved through PacBio HiFi, Hi-C scaffolding, and computational polishing tools. Beyond its technical success, this genome offers key insights into evolutionary genomics and functional biology, particularly the role of repeat elements and species-specific non-coding RNAs in regeneration (Schloissnig, Kawaguchi et al. 2021, Wang, Wang et al. 2021).

The presence of hAT transposons in gene-regulatory regions suggests that transposable elements may drive evolutionary innovations, shedding light on why some vertebrates retain regenerative capacities, while others, like mammals, have lost them(Nurk, Koren et al. 2022, Brown, Mishra et al. 2025). Understanding the genomic drivers of regeneration in *P. waltl* could inform regenerative medicine and tissue engineering strategies.

Despite these advancements, challenges remain in heterochromatic regions. Future research should integrate optical mapping and ultra-long ONT reads to resolve complex repeats(Lam, Hastie et al. 2012). Additionally, RNA-Seq, ATAC-Seq, and CRISPR-based functional validation will refine gene annotations and regulatory insights(Poplin, Chang et al. 2018, Cheng, Concepcion et al. 2021).

Ultimately, the *P. waltl* genome sets a benchmark for large-genome sequencing, providing a crucial reference for genome evolution, repetitive DNA dynamics, and regenerative biology(Brown, Mishra et al. 2025).

Reference

Brown, T., et al. (2025). "Chromosome-scale genome assembly reveals how repeat elements shape non-coding RNA landscapes active during newt limb regeneration." Cell Genom **5**(2): 100761.

Cheng, H., et al. (2021). "Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm." Nat Methods **18**(2): 170-175.

Jain, M., et al. (2018). "Nanopore sequencing and assembly of a human genome with ultra-long reads." Nat Biotechnol **36**(4): 338-345.

Koren, S., et al. (2017). "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation." Genome Res **27**(5): 722-736.

Lam, E. T., et al. (2012). "Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly." Nat Biotechnol **30**(8): 771-776.

Lieberman-Aiden, E., et al. (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." Science **326**(5950): 289-293.

Nurk, S., et al. (2022). "The complete sequence of a human genome." Science **376**(6588): 44-53.

Poplin, R., et al. (2018). "A universal SNP and small-indel variant caller using deep neural networks." Nat Biotechnol **36**(10): 983-987.

Rhie, A., et al. (2020). "Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies." Genome Biol **21**(1): 245.

Schloissnig, S., et al. (2021). "The giant axolotl genome uncovers the evolution, scaling, and transcriptional control of complex gene loci." Proc Natl Acad Sci U S A **118**(15).

Wang, K., et al. (2021). "African lungfish genome sheds light on the vertebrate water-to-land transition." Cell **184**(5): 1362-1376 e1318.

Wenger, A. M., et al. (2019). "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome." Nat Biotechnol **37**(10): 1155-1162.