Transcriptome Assemblies

Dr Stevie Pederson (They/Them) stevie.pederson@thekids.org.au

Black Ochre Data Labs
The Kids Research Institute Australia

Acknowledgement Of Country

I'd like to acknowledge the Kaurna people as the traditional owners and custodians of the land we know today as the Adelaide Plains, where I live & work.

I also acknowledge the deep feelings of attachment and relationship of the Kaurna people to their place.

I pay my respects to the cultural authority of Aboriginal and Torres Strait Islander peoples from other areas of Australia, and pay my respects to Elders past, present and emerging, and acknowledge any Aboriginal Australians who may be with us today

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Transcript Assembly

- Beyond (or as part of) differential expression analysis $\implies$ *transcriptome assembly*
- Well annotated genomes have gene models defined
  - May only have a reference from related organism
- Biology can be messy & unexpected
  - Unexpected cryptic transcripts
  - Novel TSS, UTRs etc
  - *How well do existing annotations describe our samples*?
- We can assemble *de novo* or using *reference guided* strategies

Reference Guided Assembly

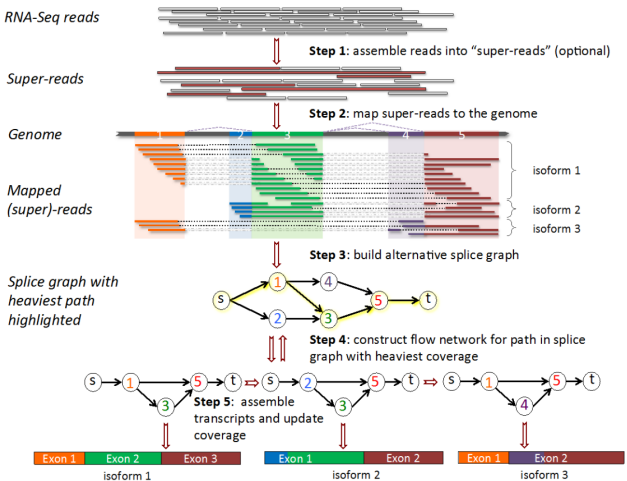

Figure 1: Image taken from (Pertea et al. 2015)

- With a good reference genome $\Longrightarrow$ *StringTie* can identify novel transcripts
- Un-annotated genes/lncRNA
- Relies on a *splice-graph* to assemble transcripts
- High expression $\Longrightarrow$ High confidence assembly
- Low expression $\Longrightarrow$ Less confidence

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Reference Guided Assembly

- StringTie returns a GTF with novel transcripts added to reference (See here)
  - Gene & Transcript annotations capture the entire transcribed region
  - Exons are annotated by transcript + gene
- Can merge assemblies across libraries/samples
  - Use merged GTF to obtain counts with `featureCounts` etc
  - Also returns counts for DE analysis

Full Transcriptome Assembly

- Complete *de novo* transcriptome assembly using Trinity (Grabherr et al. 2011)
  - Far more computationally demanding than *StringTie*
  - Can also perform a *reference-guided assembly*
- Best option where no reference genome is available
  - Or reference is low quality
- Will be tissue specific $\implies$ a subset of transcripts will be assembled
- Returns a fasta file naive to any reference genome
  - Gene/Transcript clusters in sequence header

Full Transcriptome Assembly

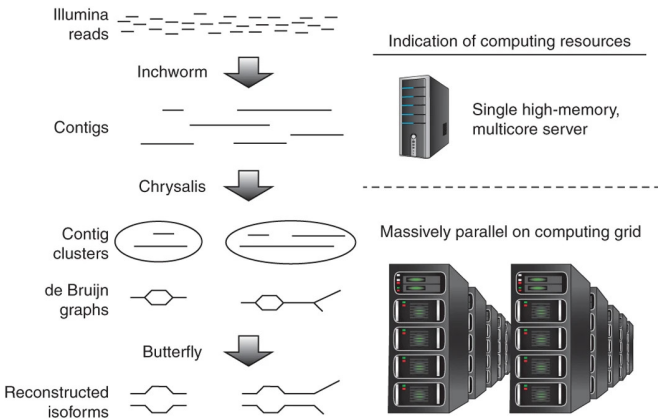

Figure 2: Figure from Haas et al. (2013)

1. Inchworm
   - Naively assembles reads into contigs
2. Chrysalis
   - Pools contigs into *de Bruijn* graph
3. Butterfly
   - Trims *de Bruijn* graph and compares against raw reads

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

De Bruijn Graphs

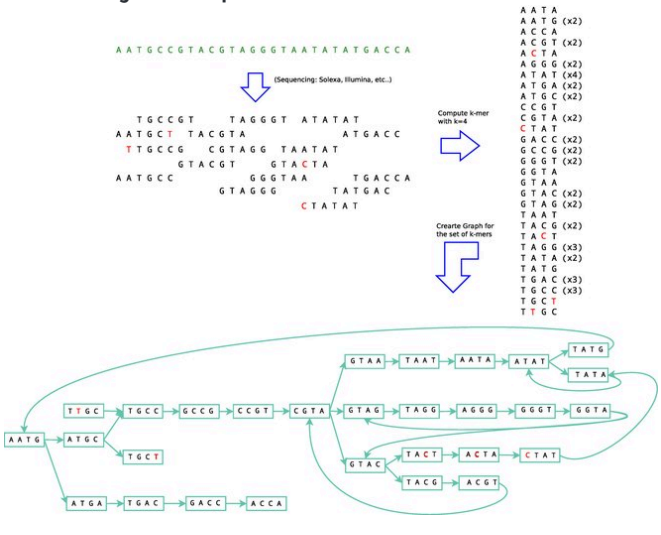

Figure 3: *An example de Bruijn graph* (Wikimedia commons)

- Reads represent fragmented transcripts
- Reads broken into shorter k-mers
- Adjacent k-mers $\rightarrow$ graph
- Wise choice of k helps extend graph well beyond read length
- Well-represented paths through graph $\rightarrow$ transcripts

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Assessing Assembly Quality

- Benchmarking Universal Single-Copy Orthologs (BUSCO) (Tegenfeldt et al. 2025)
  - Checks expected genes conserved across species
  - Assess assembly quality against expected true positives
  - Most (but not all) orthologs expressed in assembled tissue
- Example BUSCO output: **C:89.0%[S:85.8%,D:3.2%],F:6.9%,M:4.1%,n:3023**
  - **C**: Complete orthologs
    - **S**: Single copy + **D**: Duplicated copy
  - **F**: Fragmented copy
  - **M**: Missing copy
  - **n**: Total orthologs

Long Read Technology

Long Read Technology

- Most transcriptome assemblies performed using Trinity/StringTie
  - Illumina reads $\leq$ 2x150nt
- Long Reads are becoming dominant for assemblies
  - Sequence (near-)complete transcripts
- Transposon expression in relevant cell types
- Quantification approaching short-read consistency
  - Overcomes fragmentation bias in short-read technology (Chen et al. 2025)

Long Read Technology

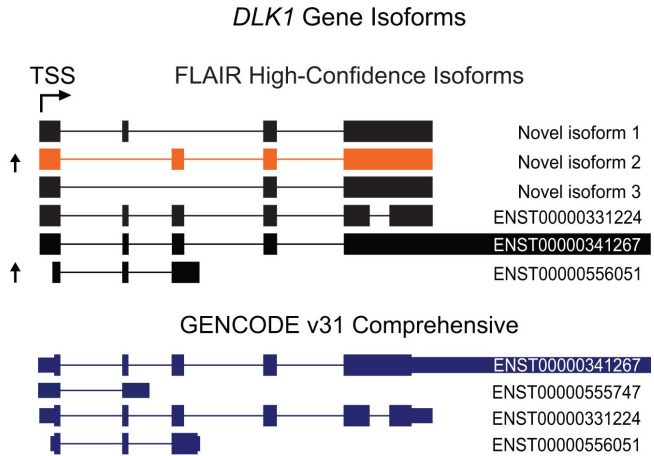

Figure 4: Image from Gleeson et al. (2021). Isoforms observed in SH-5Y5Y cell

Oxford Nanopore (ONT)

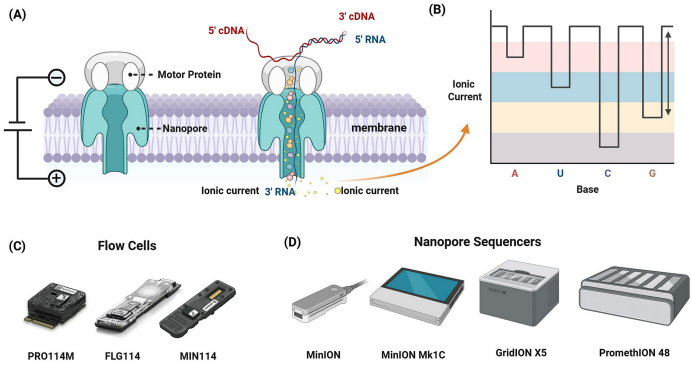

Figure 5: Image from Sun et al. (2025)

- From 50 bp to >4 Mb
- Can sequence RNA or cDNA (pre/post-PCR)
- More error prone

Pacific Biosciences (PacBio)

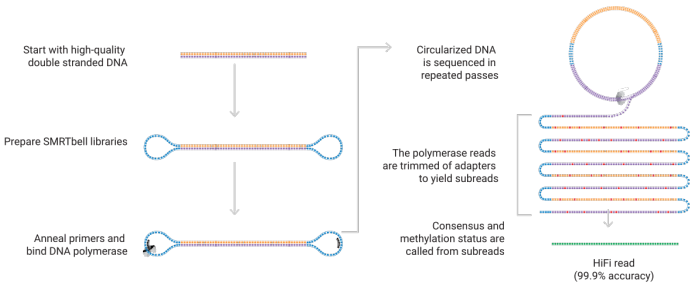

Figure 6: Image from https://www.pacb.com/technology/hifi-sequencing/how-it-works/

- PacBio IsoSeq:
  - Up to 25kb
  - Sequence cDNA only
  - Highly accurate reads

References

Chen, Ying, Nadia M Davidson, Yuk Kei Wan, Fei Yao, Yan Su, Hasindu Gamaarachchi, Andre Sim, et al. 2025. "A Systematic Benchmark of Nanopore Long-Read RNA Sequencing for Transcript-Level Analysis in Human Cell Lines." *Nat. Methods* 22 (4): 801–12.

Gleeson, Josie, Adrien Leger, Yair D J Prawer, Tracy A Lane, Paul J Harrison, Wilfried Haerty, and Michael B Clark. 2021. "Accurate Expression Quantification from Nanopore Direct RNA Sequencing with NanoCount." *Nucleic Acids Research* 50 (4): e19–19. https://doi.org/10.1093/nar/gkab1129.

Grabherr, Manfred G, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data Without a Reference Genome." *Nat. Biotechnol.* 29 (7): 644–52.

Haas, Brian J, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, et al. 2013. "De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis." *Nat. Protoc.* 8 (8): 1494–1512.

Pertea, Mihaela, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. 2015. "StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads." *Nat. Biotechnol.* 33 (3): 290–95.

Sun, Kai, Jiaxin Li, Chaohao Chen, Xin Zhou, Guofang Ma, Lingfeng Mao, Qiao Tang, et al. 2025. "Advances in Nanopore Direct RNA Sequencing and Its Impact on Biological Research." *Biotechnology Advances* 85: 108710. https://doi.org/https://doi.org/10.1016/j.biotechadv.2025.108710.

Tegenfeldt, Fredrik, Dmitry Kuznetsov, Mosè Manni, Matthew Berkeley, Evgeny M Zdobnov, and Evgenia V Kriventseva. 2025. "OrthoDB and BUSCO Update: Annotation of Orthologs with Wider Sampling of