Wellcome Trust Advance Course: Module 8: Genome assembly

2020-10-23

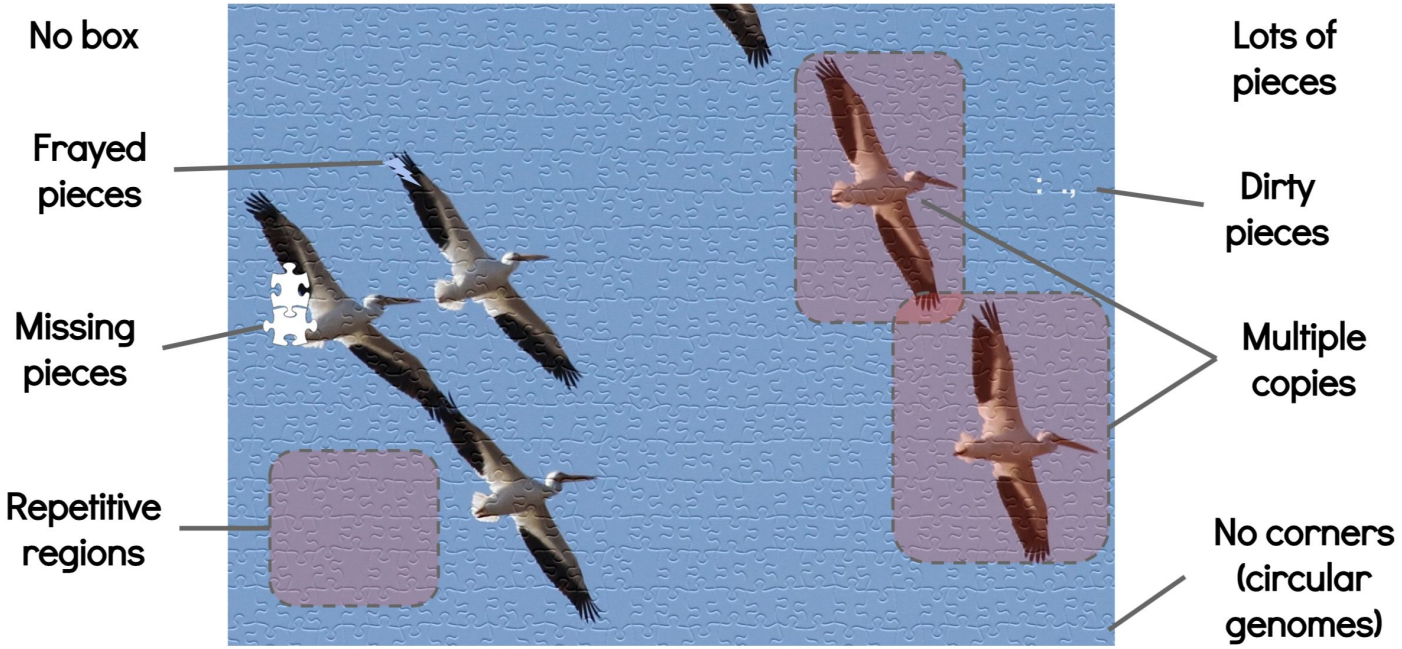
Shane A. McCarthy
University of Cambridge
Department of Genetics
sam68@cam.ac.uk

What is genome assembly?

- Sequencing technologies can only sequence short stretches of DNA
- Given many (millions or billions) of reads, produce a linear (or perhaps circular) genome
- Clone based
  - Select clones using markers
  - Sequence clones separately
- Whole-genome shotgun
  - Fragment whole-genome and sequence
- **De novo assembly**: reconstruct the genome sequence fr

Figure 1. An example of a clone tiling path. All lines represent clones and their relative positions. The red lines represent a minimal tiling path through this region.

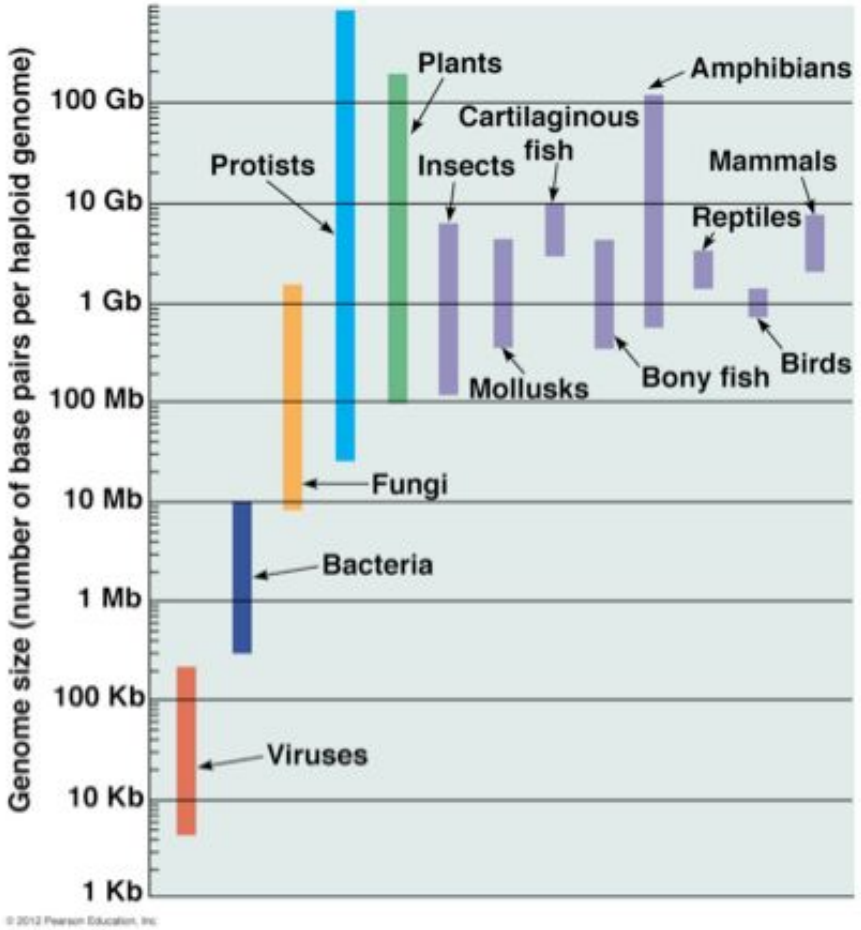
What makes a jigsaw puzzle hard?



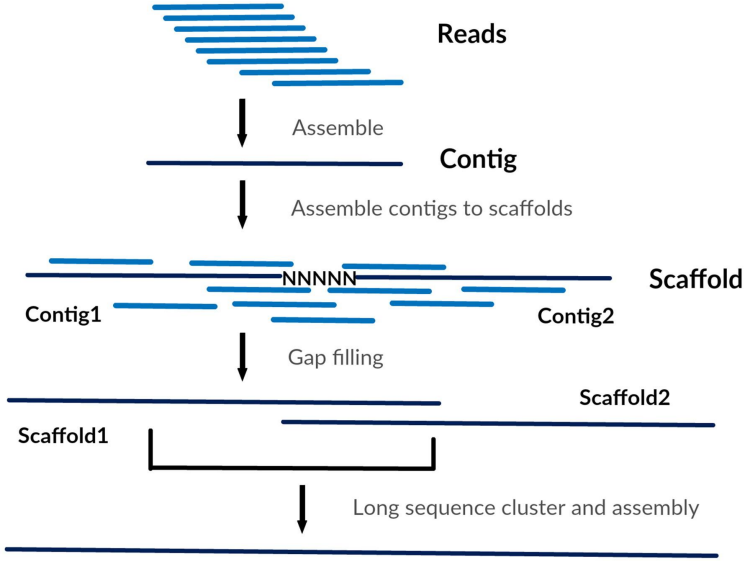
- What helps? Larger pieces (read length); fewer dirty or frayed pieces (errors in reads). fewer repeats and copies…

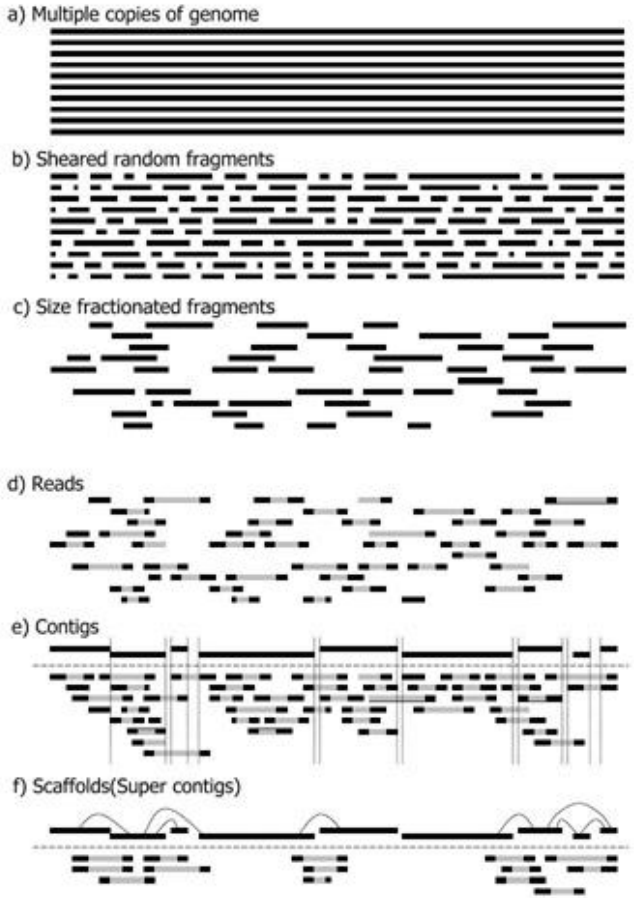
Key considerations

- sequencing coverage
- errors in reads
- reads lengths
- non-uniqueness of the genome (repeats)
- genome size
- genome heterozygosity and ploidy
- running time and memory



Terminology





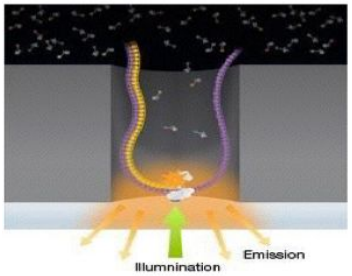
Chromosome

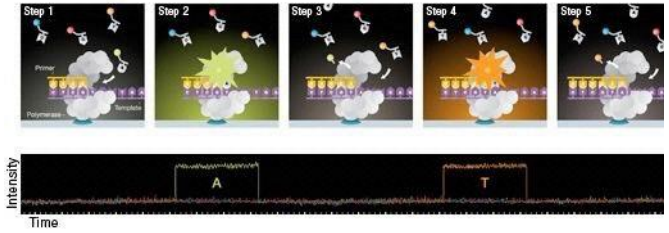
Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis

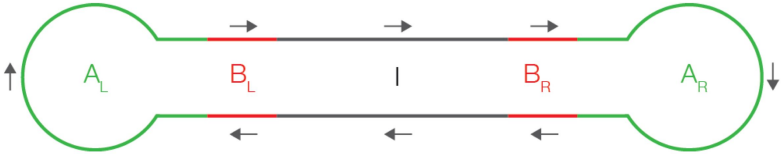
Yan Guo [a,*], Yulin Dai [a], Hui Yu [a], Shilin Zhao [a], David C. Samuels [b], Yu Shyr [c,*]

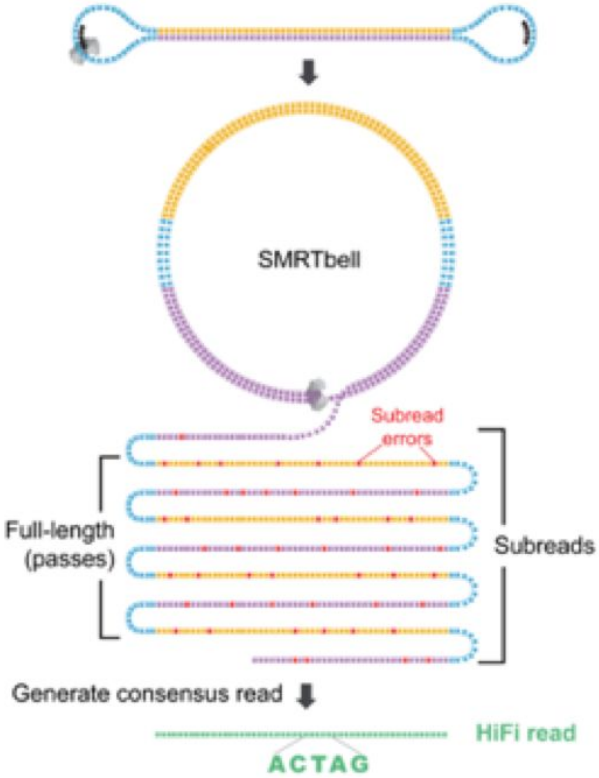
- contig: contiguous stretch of assembled sequence
- scaffold: ordered set of contigs with gaps (Ns) placed between

PacBio sequencing









Key points:

- 1 DNA molecule and 1 polymerase in each well (zero-mode waveguide, ZMW)
- 4 colours flash in realtime as polymerase acts
- Methylated bases have distinct pattern
- No theoretical limit to DNA fragment length

PacBio sequencing and assembly

Advantages:

- Long read lengths
- Few systematic errors
- Can detect some base modifications
- Can get more accuracy, if you loop over the ZMW multiple times - circular consensus sequencing (CCS)

Disadvantages:

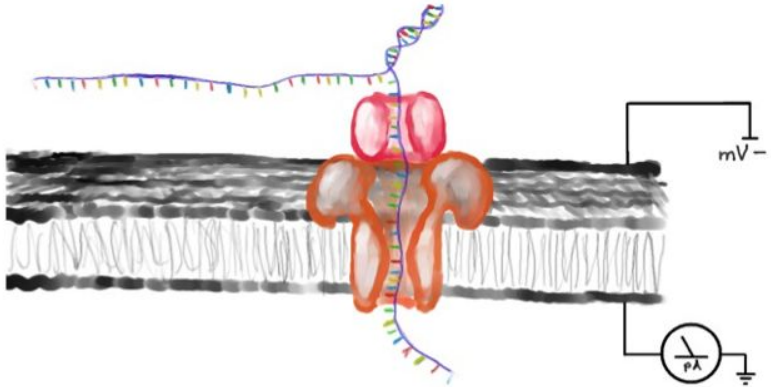
- High read error rate
- High cost per-base

2019 Update:

- Sequel II launched with 8-fold increase in yield and improvement in CCS accuracy (HiFi reads).

Pacbio Sequel
>10kbp read length
Up to 10GB yield

Nanopore sequencing



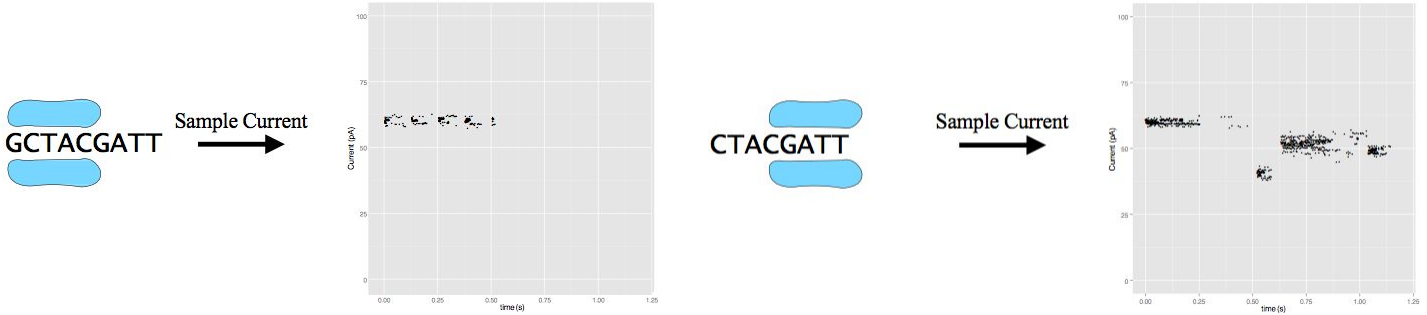


Illustration by David Eccles

Nanopore sequencing and assembly

Oxford Nanopore MinION
>10kbp read length
5-15GB yield

Oxford Nanopore PromethION
>10kbp read length
30-90GB yield per flowcell

Advantages:

- Portability and low capital cost
- Read lengths up to 1Mbp reported
- Base modification detection
- Sequence RNA directly

Disadvantages:

- High error rate
- Systematic errors around homopolymers

Read clouds or linked-reads

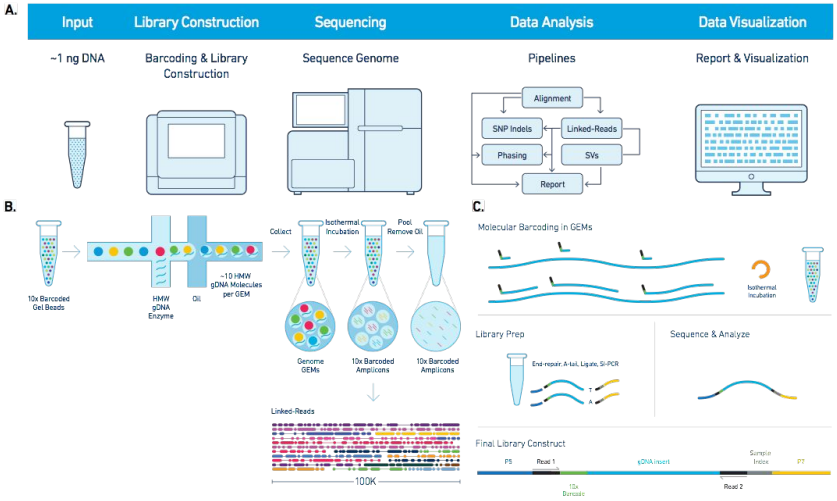
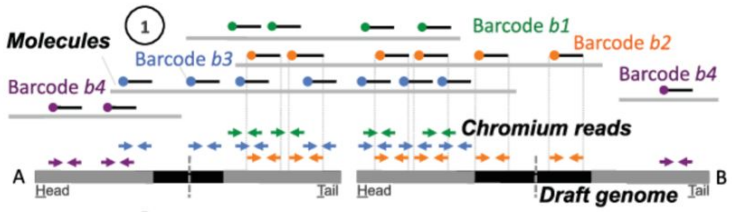


Figure 1. Chromium™ Genome Solution. (a) The Chromium Genome Solution provides a streamlined workflow and easy-to-use informatics pipeline and software for whole genome analysis. (b) An automated microfluidic system allows for functionalized gel beads to be combined with high molecular weight DNA (HMW gDNA) and oil to form a 'Gel Bead in Emulsion (GEM)'. Each GEM contains ~10 molecules of HMW gDNA and primers with unique barcodes. Isothermal incubation allows for the addition of a unique barcode to all DNA within the GEM. (c) Assay schematic overview.

10x GENOMICS®

Longer range “read clouds” e.g. from 10X Genomics Chromium, can bridge bigger repeats/gaps

Multiple reads from a long ~100kb template sharing a barcode



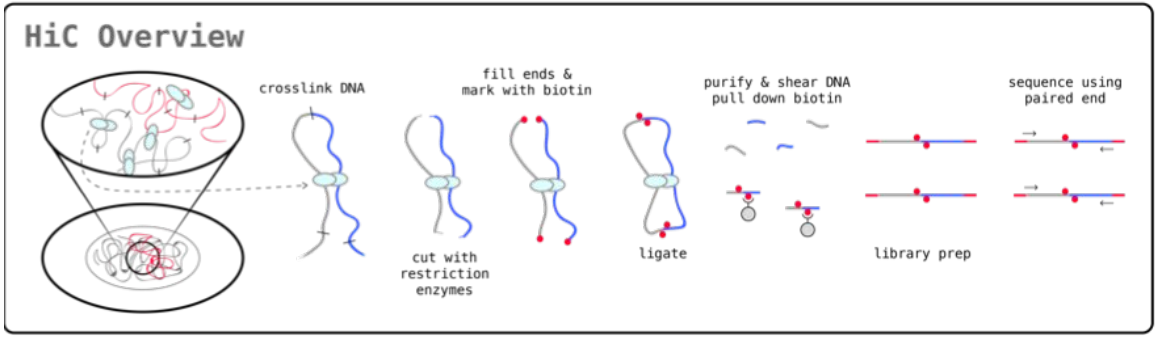
Yeo S, Coombe L, Warren RL, Chu J, Birol I. ARCS: scaffolding genome drafts with linked reads. Bioinformatics. 2018.

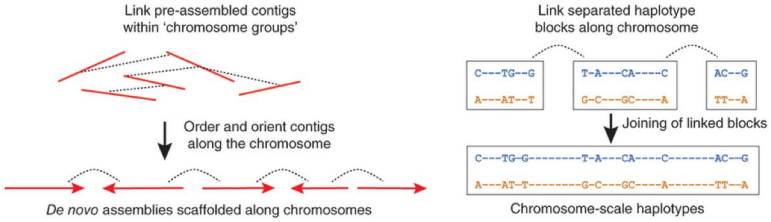
Hi-C

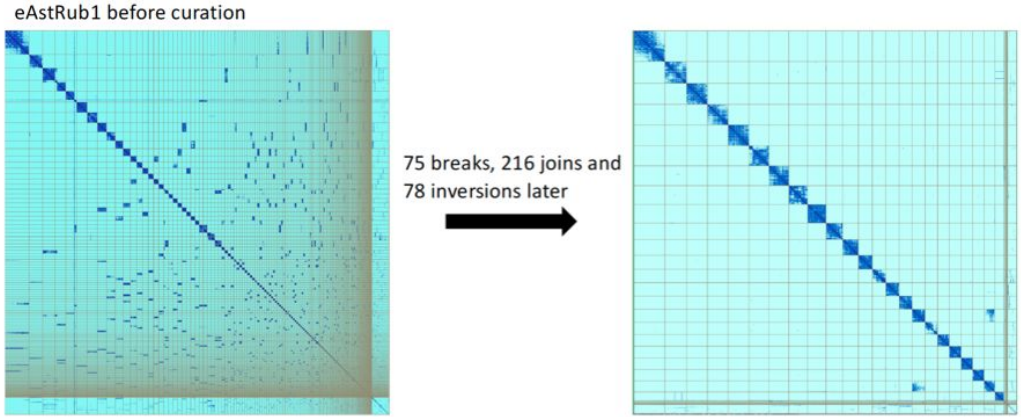
Hi-C “proximity ligation”

is used to look at local chromatin organisation, but globally most Hi-C links are on the same chromosome, and more are close than distant

- Automated tools (e.g. SALSA2, 3d-dna)
- Visualisers (e.g. Juicer, Hi-Glass)
- Editors (e.g. Juicer, Pretext)

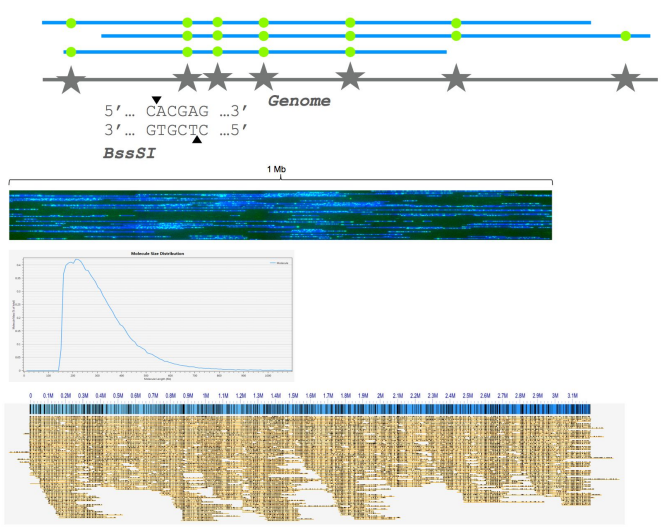






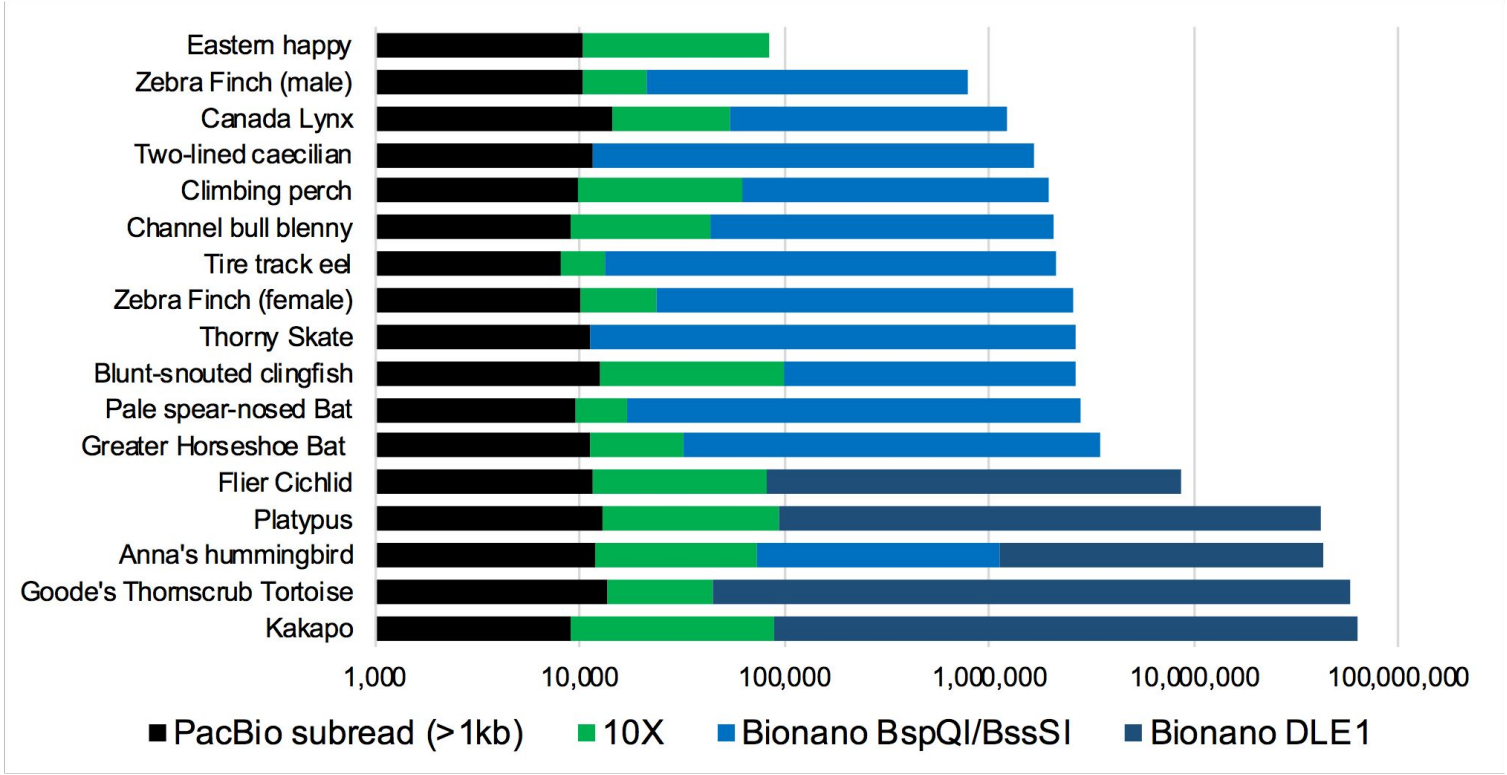
Optical mapping

- Isolate high molecular weight DNA fragments
- Label specific sequence motifs across the entire genome
- Linearise and image labelled DNA
- Restriction digest maps give approximate spacing of restriction sites (short sequence motifs), skipping over repeats
- Now scaled up by dedicated optical imaging machines, e.g. BioNano



bionano GENOMICS

Long-range technologies



Long-range technologies

Table 1 | **Long-range sequencing and mapping platforms**

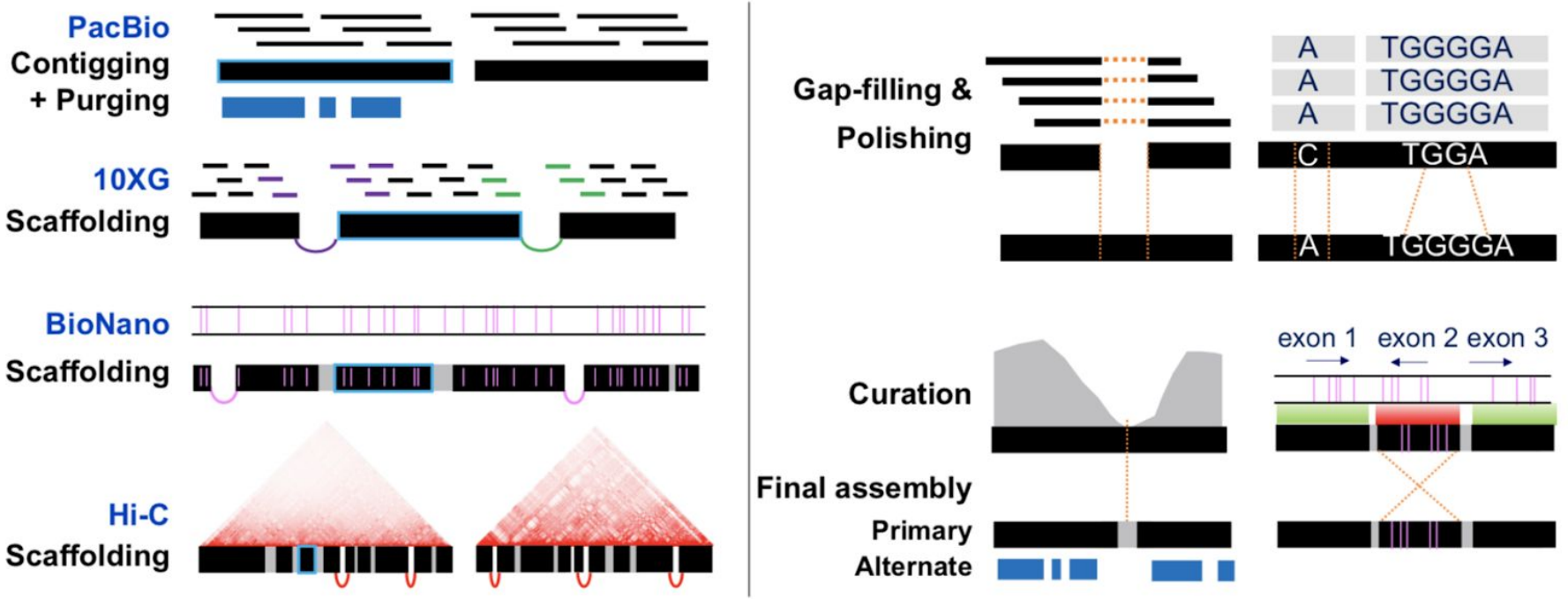
Platform	General characteristics and costs	Major applications	Bioinformatics challenges
PacBio SMRT sequencing	Single-molecule long reads averaging ~10 kb with some approaching 100 kb; several fold more expensive than short reads	De novo genome assembly, structural variant detection, gene isoform resolution and epigenetic modifications	Raw reads have high error rates dominated by false insertions; requires new alignment and error correction algorithms
Oxford Nanopore sequencing	Single-molecule long reads averaging ~10 kb with some >1 Mb; several fold more expensive than short reads	De novo genome assembly, structural variant detection, gene isoform resolution and epigenetic modifications	Raw reads have high error rates dominated by false deletions and homopolymer errors; requires new alignment and error correction algorithms
10X Genomics Chromium	Linked reads spanning ~100 kb derived from a collection of short-read sequences; moderately more expensive than short reads	De novo genome assembly and scaffolding, phasing, detection of large structural variants (>10 kb) and single-cell gene expression	Sparse sequencing rather than true long reads; more complicated to align, with poorer resolution of locally repetitive sequences
Hi-C-based analysis	Pairs of short reads (<100 bp) formed from crosslinking chromatin interactions; moderately more expensive than short reads	Genome scaffolding and phasing	Sparse sequencing with highly variable genomic distance between pairs (1 kb to 1 Mb or longer)
BioNano Genomics optical mapping	Optical mapping of long DNA molecules (~250 kb or longer) labelled with fluorescent probes; less expensive than short reads	Genome scaffolding and detection of large structural variants (>10 kb)	Limited algorithms to discover high-confidence alignment between an optical map and a sequence assembly

PacBio SMRT, Pacific Biosciences single-molecule real time.

Piercing the dark matter: bioinformatics of long-range sequencing and mapping

Fritz J. Sedlazeck[1], Hayan Lee[2], Charlotte A. Darby[3] and Michael C. Schatz[3,4]*

Example genome project workflow

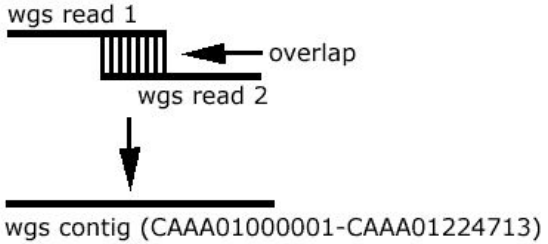


Vertebrate Genomes Project

Assembly: contig generation

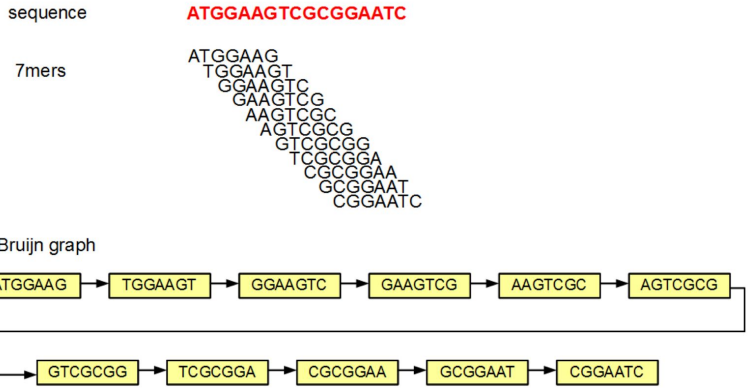
OLC (Overlap Layout Consensus):

- For all pairs x, y, of reads
  - determine if there is sufficient overlap
  - bundle stretches of overlap graph into contigs
- Computationally expensive (quadratic scaling with current approaches)
- Assembly software
  - Falcon (PacBio), Canu (PacBio, ONT), minimap/miniasm



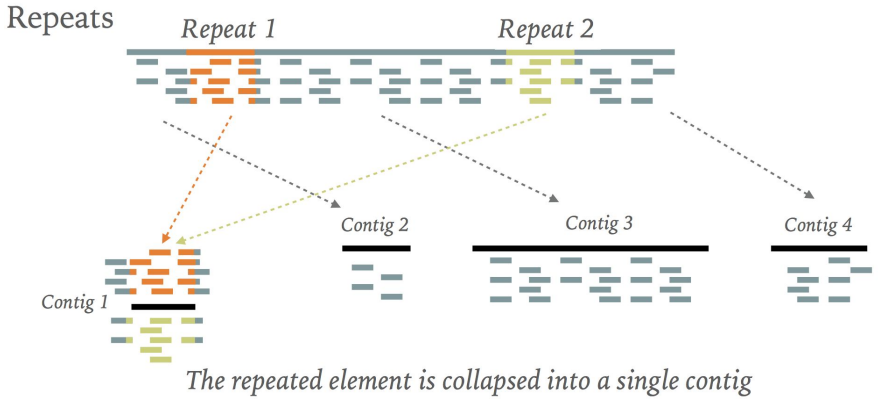
DBG: de Bruijn graph (*k*-mer):

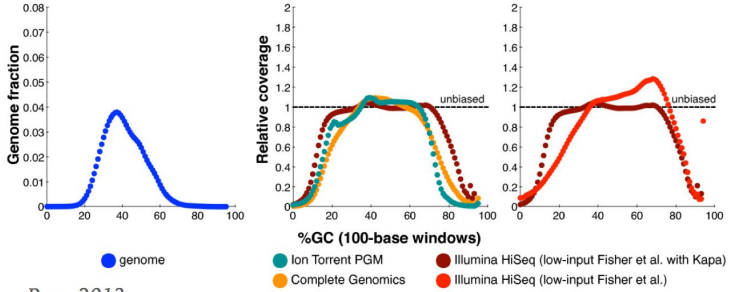
- Build a graph of all subsequences of length *k*.
- Assembly software
  - velvet, ABySS, SPAdes, wtdbg2



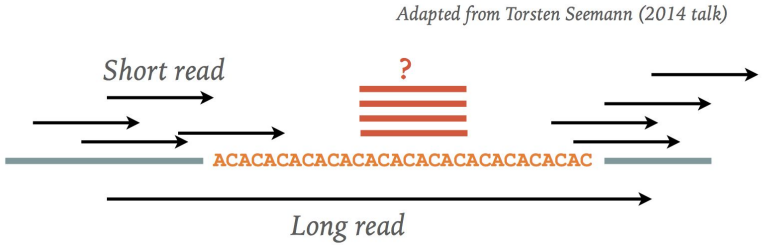
Hurdles

- Heterozygosity
- Sequencing errors
- Repeats
- Low complexity genomic regions
- Base composition and sequencing bias



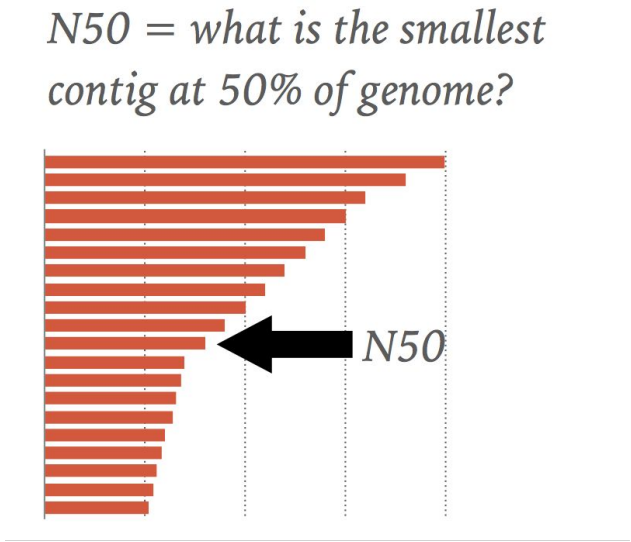


Ross, 2013



Assembly metrics

- total length
- number of sequences (contigs and scaffolds)
- average length (contigs and scaffolds)
- largest/smallest (contigs and scaffolds)
- N50 = X means 50% of the genome is in sequences larger than X
- NG50 (N50 scaled by the expected genome size)
- Gene content (% conserved core genes mapped)



Scaffolding

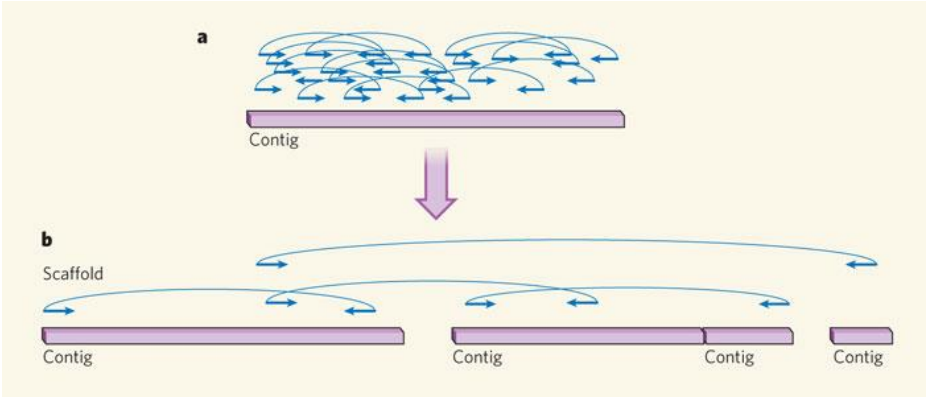
Goal: Order and orient the contigs into larger structure

Construct libraries of varying insert sizes

- Smaller size (2, 4 or 6 Kb), intermediate size (10-40 Kb), and libraries with large insert sequences (>100 Kb)
- Ends of these clones are sequenced, generating sequence reads.

Sources of evidence

- Mate-pair illumina libraries
- Fosmid ends
- Bacterial artificial chromosomes
- 10x Genomics linked reads
- Hi-C
- Optical maps



Gaps consisting of N (unknown) bases are inserted between contigs

(Pseudo) chromosome assignment

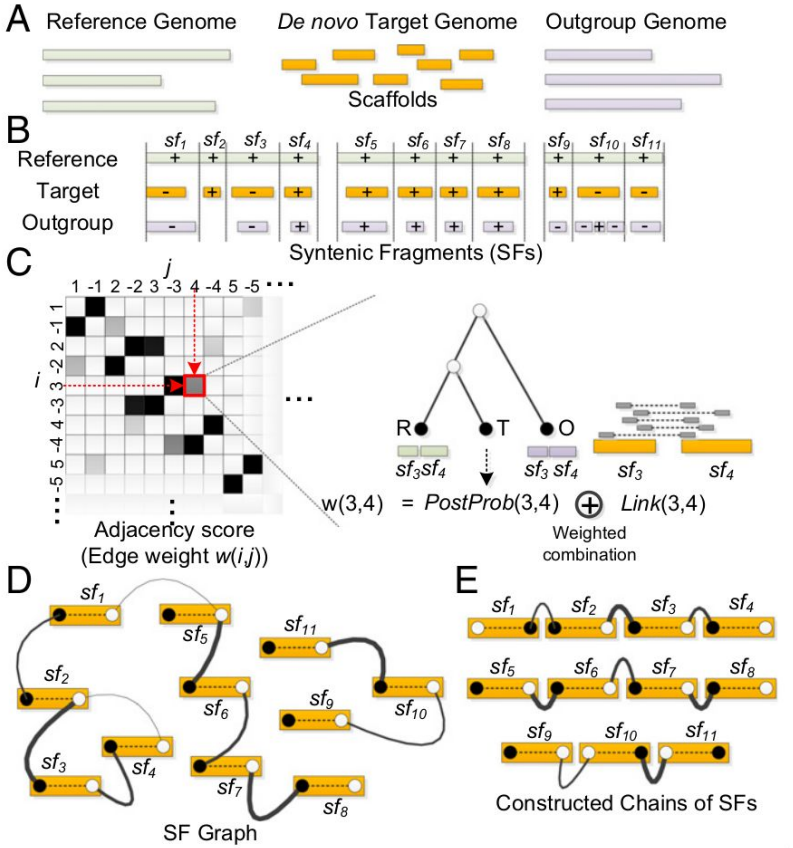
Goal: Assembly of chromosome-scale DNA fragments from scaffolds

Sources of evidence

- A close reference genome and/or outgroup genome
- Genetic markers or map

RACA software

- Input: Reference genome, de novo assembly, and one or more outgroups (A)
- Identify syntenic fragments (B)
- Calculate likelihood of adjacencies (C)



Reference-assisted chromosome assembly

Jaebum Kim[a,b,1], Denis M. Larkin[c,1], Qingle Cai[d], Asan[d], Yongfen Zhang[d], Ri-Li Ge[a,2], Loretta Auvil[f,g], Boris Capitanu[f,g], Guojie Zhang[d], Harris A. Lewin[a,h,2], and Jian Ma[a,i,2]

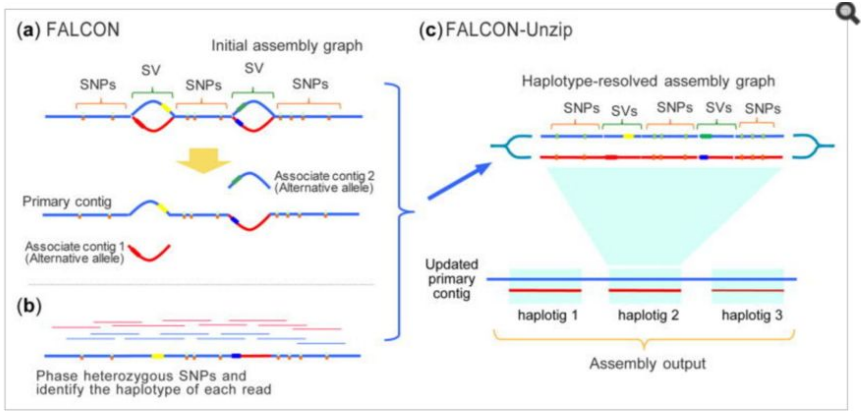
Diploid assembly

Almost all assemblers assume genome is haploid

- Ignore allelic variation between parental chromosomes

Diploid assembly: Produce two haplotype phased genomes

- e.g. Falcon-unzip

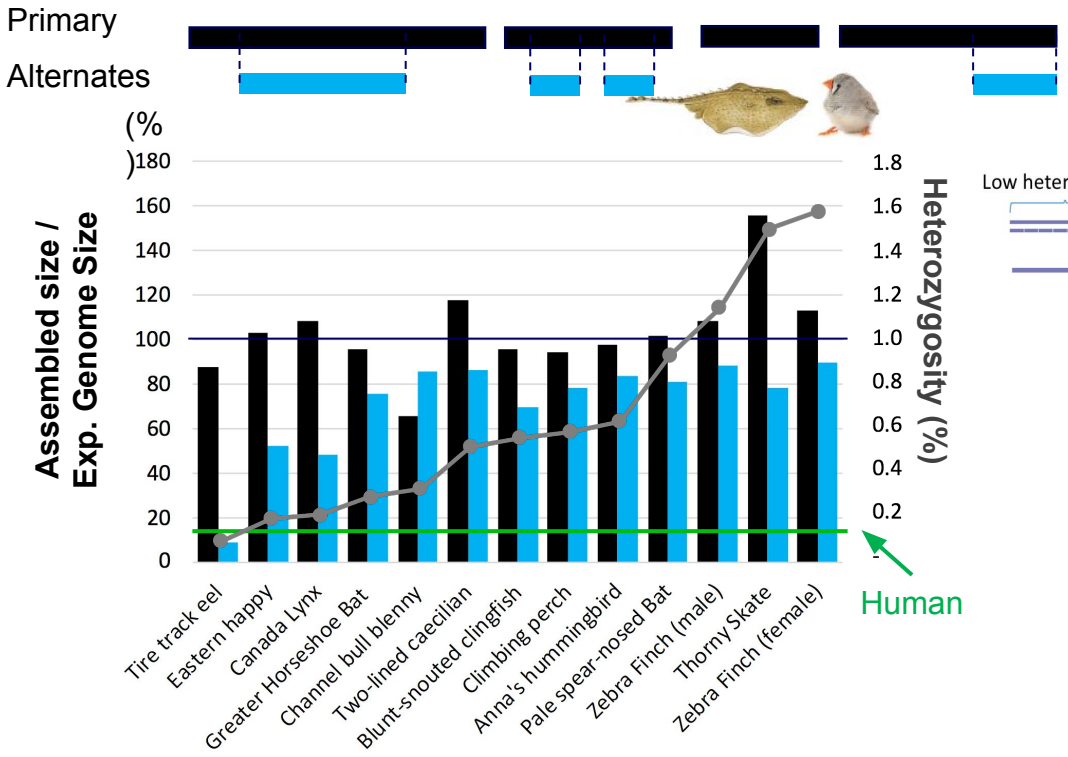


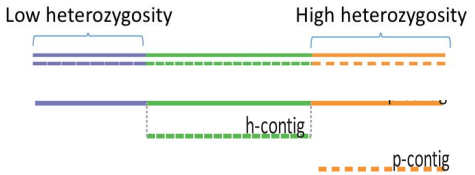
FALCON and FALCON-Unzip overview

(**a**) The initial assembly is computed by FALCON, which error corrects the raw reads (not shown) and then assembles using a string graph of the read overlaps. The assembled contigs are further refined by FALCON-Unzip into the final set of contigs and haplotigs. (**b**) Phase heterozygous SNPs and group reads by haplotype (**c**) The phased reads are used to open up the haplotype-fused path and generate as output a set of primary contigs and associated haplotigs.

Chin et al. (2016), Nat Methods

Heterozygosity and allelic duplication





Chin *et al., Nat Met.* (2016)

Dealing with haplotype duplication

For high heterozygosity genomes, Falcon-unzip leaves allelic duplicates in the primary assembly

purge_haplotigs will identify these duplicates and place them in the haplotig bin. Can remove true paralogous sequence.

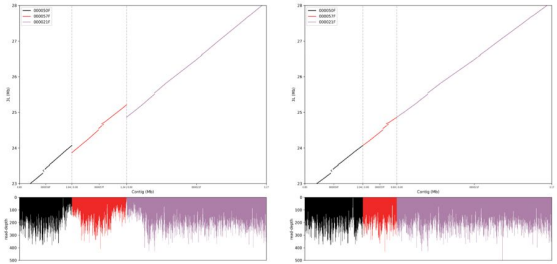
Falcon-phase will use the Hi-C data to phase, switching chunks of the primary with haplotigs to match haplotype phase

Purge Haplotigs: Synteny Reduction for Third-gen Diploid Genome Assemblies

Michael J Roach[1*], Simon Schmidt[1] and Anthony R Borneman[1]

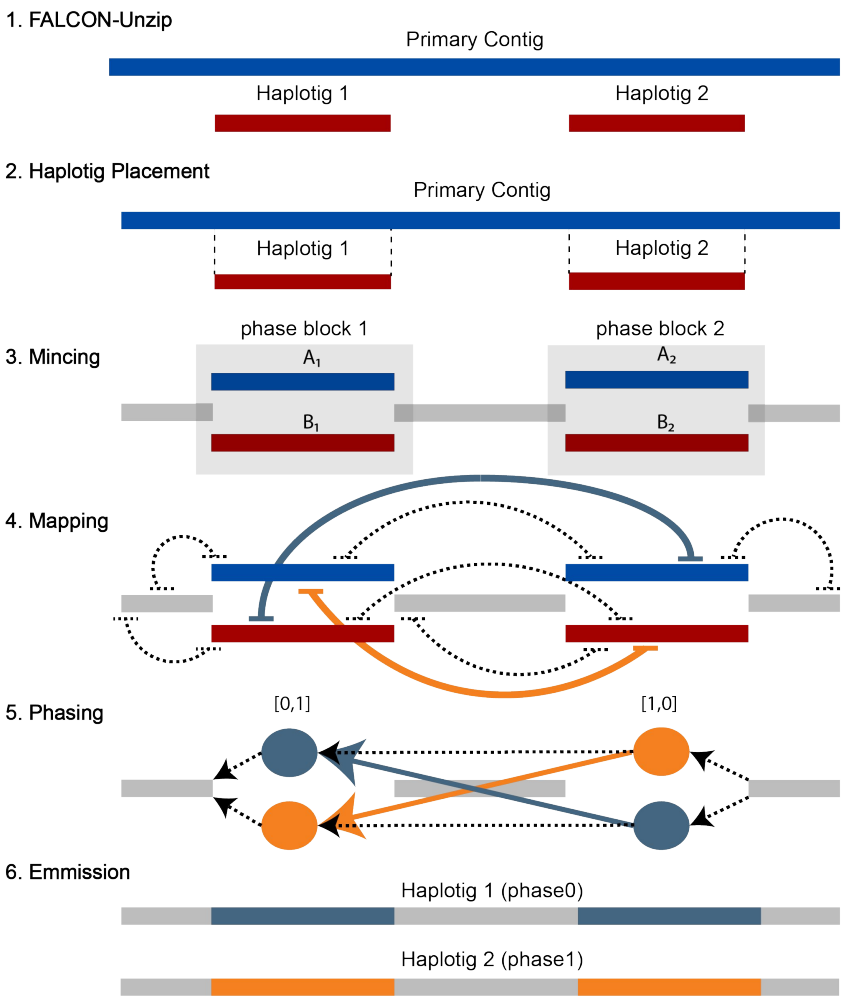
Identifying and removing haplotypic duplication in primary genome assemblies

Dengfeng Guan[1,2], Shane A. McCarthy[2], Jonathan Wood[3], Kerstin Howe[3], Yadong Wang[1,*] and Richard Durbin[2,3,*]



Falcon-phase

Use Hi-C data to help phase variants over a longer range than Falcon-unzip can do with SNP calls from PacBio reads alone.

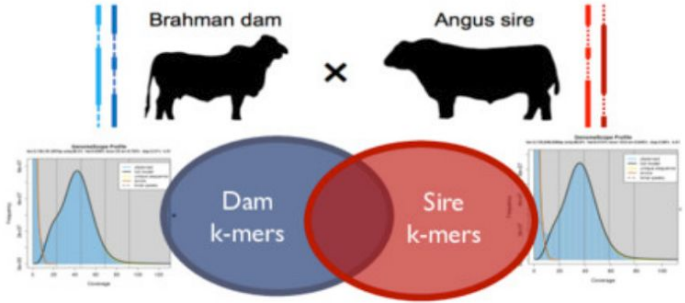


FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes
bioRxiv: https://doi.org/10.1101/327064

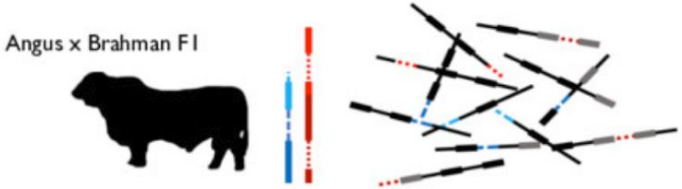
Diploid assembly

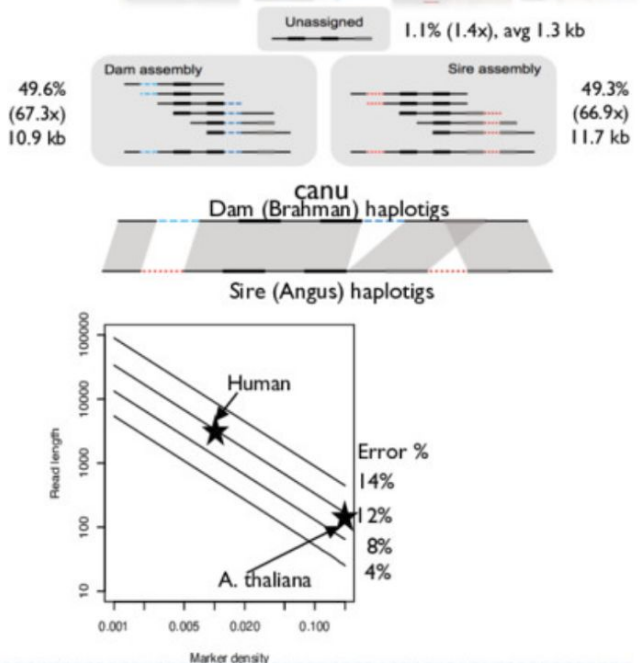
Trio Binning

- K-mer profiling of each parent (Illumina, 60x)



- K-mer profiling of the F1 (PacBio, 120x)





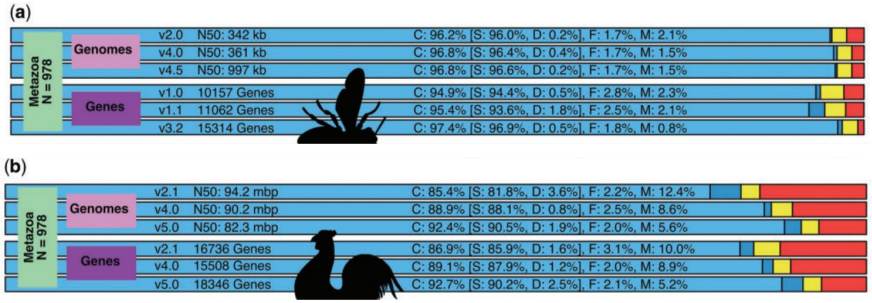
Genome assembly QC

Base Accuracy

- Realign reads from the same species
  - Identify SNPs+indels. Indels - known to be enriched in Pacbio+Nanopore assemblies
  - k-mer completeness with respect to Illumina data (KAT)

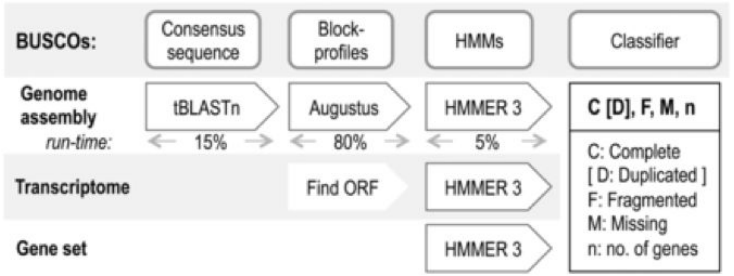
Local structure accuracy

- External evidence - e.g. sequencing data (REAPR, QUAST)
- Known adjacencies, e.g. PCR primers or structural variants



Gene content

- Order and orientation of genes/exons of known genes (e.g. housekeeping genes)
- BUSCO: Quantitative assessment of genome assembly and annotation completeness based on evolutionarily informed expectations of gene content



Benchmarking Universal Single-Copy Orthologs

- Start from set of conserved genes
- Find them in the assembly
- Assess whether they are Complete (C), Single copy (S) or Duplicated (D), Fragmented (F) or Missing (M)

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs

Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, Evgeny M. Zdobnov

Bioinformatics, Volume 31, Issue 19, 1 October 2015, Pages 3210–3212,

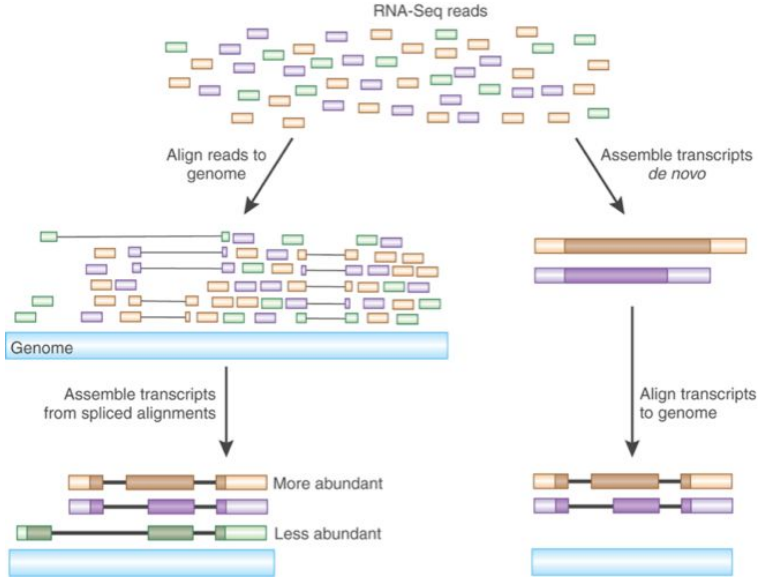
Metagenomic assembly

- What if your DNA sample is a mix of uncultured (probably microbial) organisms?
- Abundances of the different components vary
  - Need to separate divergent sequences from errors
  - Output is typically a set of contigs, with coverage
  - Still a valuable simplification of a data set
- What scaffolding approaches are suitable?

Ayling M, Clark MD, Leggett RM. New approaches for metagenome assembly with short reads. Brief Bioinform. February 2019. doi:10.1093/bib/bbz020.

Transcriptome assembly

- Another important form of data is cDNA from a sample
  - Again no single linear underlying sequence, and abundances of different components vary
- Alternative splicing generates branches



Packages: Trinity, TopHat

Tree of Life Project Sequencing Projects

GENOME 10K

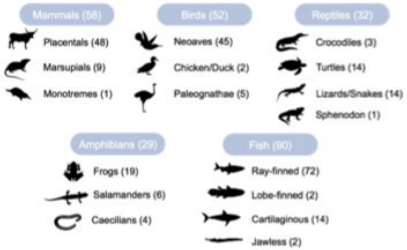
The Vertebrate Genomes Project (VGP), a project of the G10K Consortium, aims to generate near error-free reference genome assemblies of all 66,000 extant vertebrate species.

VGP VERTEBRATE GENOMES PROJECT

VertebrateGenomesProject.org

Celebrate 25th anniversary of the Sanger Institute by sequencing 25 genomes representing five key areas of biodiversity in Britain

- **Cryptic**: Species that are out of sight, or have identical forms that are different in behaviour
- **Dangerous**: Invasive and harmful species
- **Floundering**: Endangered and declining species
- **Flourishing**: Species on the up in the UK
- **Iconic**: Species that represent the British countryside

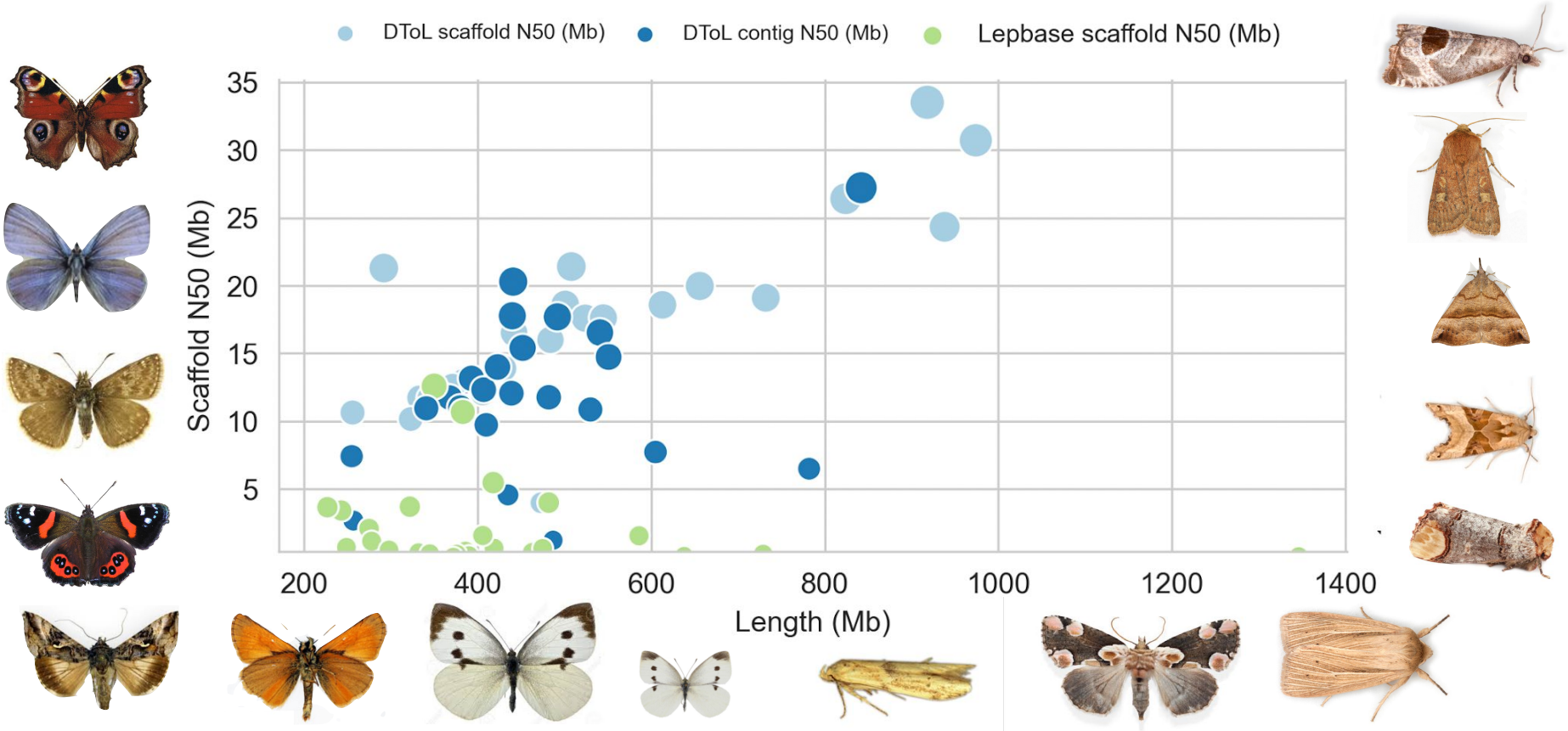


Darwin TREE of LIFE

- New UK initiative to sequence all ~66,000 eukaryotic UK species.
- Collaboration between a number of UK scientific centres, including Wellcome Sanger Institute, Earlham Institute, Natural History Museum, Kew Gardens, Royal Botanic Garden Edinburgh

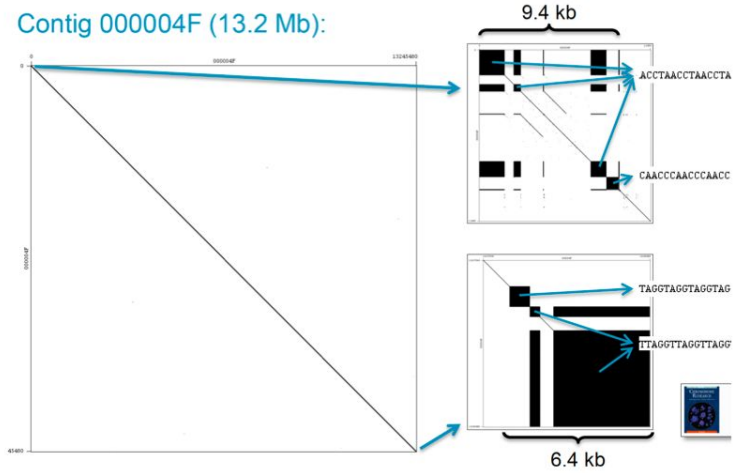
Darwin Tree of Life Project

48 New Lepidoptera assemblies



Complete chromosome contigs

Red Admiral butterfly
Darwin project + Jonas Korlach
34x PacBio CCS data



Human X chromosome
Miga et al. *bioRxiv* 2019
70x PacBio CLR data
39x Oxford Nanopore
"ultra-long" data