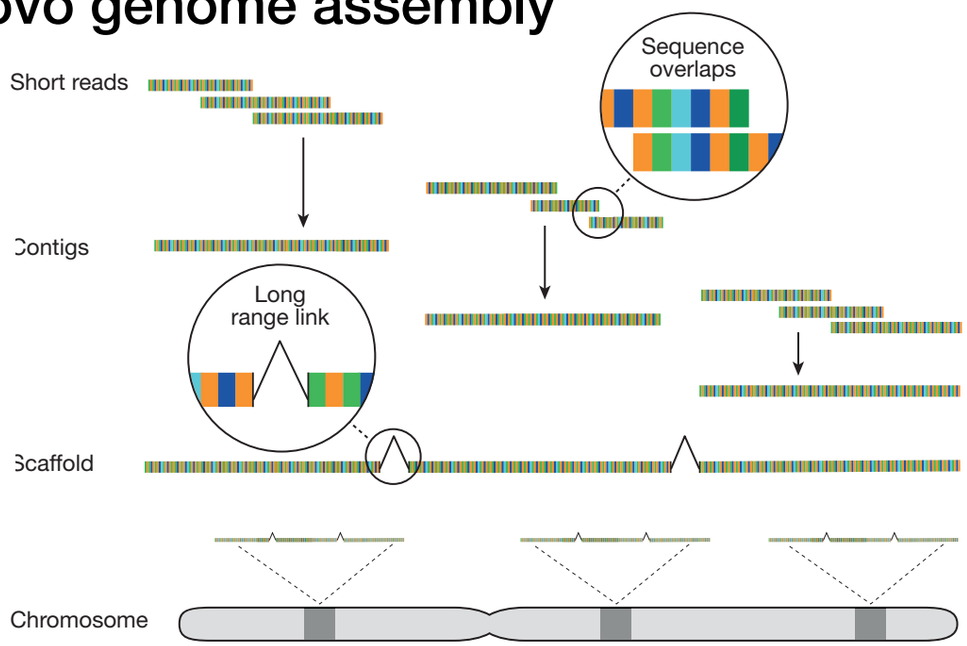
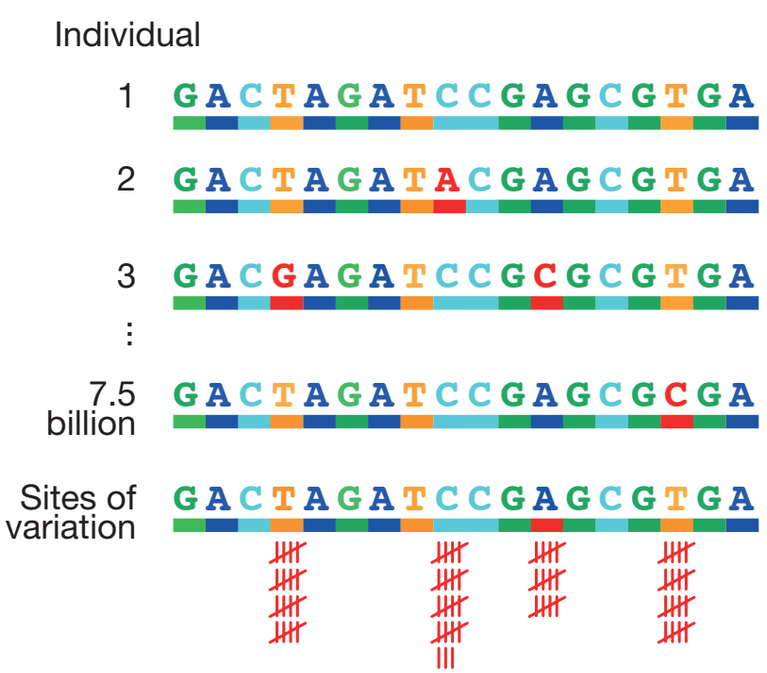
Sequencing technologies

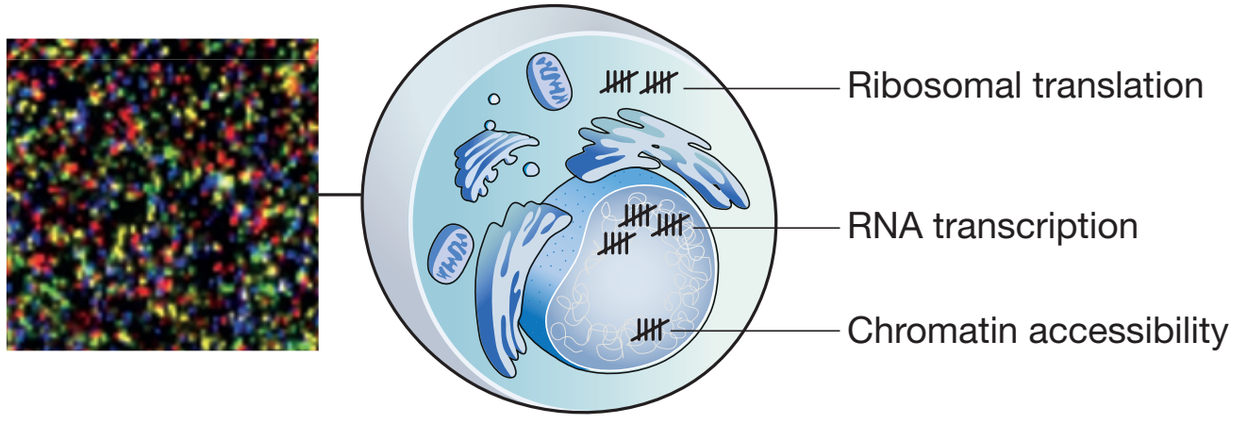
De novo genome assembly



Genome resequencing



Sequencer as counting devices



Shendure, Jay, et al. DNA sequencing at 40: past, present and future. *Nature* 550.7676 (2017):

Sequencing technologies

Escherichia coli strain ER1709 chromosome, complete genome
Sequence ID: CP030240.1 **Length:** 4582842 **Number of Matches:** 1

Range 1: 167828 to 168128 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
551 bits(298)	5e-153	300/301(99%)	0/301(0%)	Plus/Minus

```
Query  1       CACCAGCGTTAACAGCAATAACAGCAGGACGATAATCAGCAGGTGATTACGTGCCAGTCC  60
Sbjct  168128  CACCAGCGTTAACAGCAATAACAGCAGGACGATAATCAGCAGGTGATTACGTGCCAGTCC  168069
Query  61      GGCGGTAATGGTCATAAATTCGCTAAGAAAAATGTTGAAGGGCGGCATCCCTGCCAGCGC  120
Sbjct  168068  GGCGGTAATGGTCATAAATTCGCTAAGAAAAATGTTGAAGGGCGGCATCCCTGCCAGCGC  168009
Query  121     CAGCGCACCGCCGCCAAACAGCACGGCGGTAAATGGCATGATTTTGAGCATCCCACAGAC  180
Sbjct  168008  [illegible]                                                   167949
Query  181     [illegible]                                                   240
Sbjct  167948  [illegible]CGGAACCGCAGAACAGCAG                                 167889
Query  241     CGCTTTTGCCAGACTGTGGTTTAAGATGTGCAGCAGCGCGGCAAAAATTCCCAGCGGCCC  300
Sbjct  [illegible]
Query  301     G  301
Sbjct  [illegible]
```

Read length = 150bp (Max 300bp)
Average error rate = 0.01%
Human genome (30X, ILL) = ~~$1500~~ ~$600
Human genome (30X, BGI) = ~~$1000~~ ~$500

Short

Long

Escherichia coli strain ER1709 chromosome, complete genome
Sequence ID: CP030240.1 **Length:** 4582842 **Number of Matches:** 1

Range 1: 129130 to 131560 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
1661 bits(899)	0.0	2016/2506(80%)	274/2506(10%)	Plus/Plus

```
Query  94      CGCCACGGCT----ACACGTCGGTAATGCACGGTTCGCC-ACCAGACATATGGCCAGAGC  148
Sbjct  129130  CGCCACGGCTGCACACACGTCGGTAATGCACGGTTCGCCCACCGGAC--ATGGCCAGAGC  129187
Query  149     G--ATGGC-A-GCAGTCAAGGCT-A--C-ACGCGTC-GGCCAACGGTCAT-CCTGCCTGA  198
Sbjct  129188  GTCATGGCGATACCTTTAACGGTCAGGCTACGCGTCAGCCCGGCGGTCATCCCTGCCTGA  129247
Query  199     TGCAAAAAGCTGTCTGCC-TCACGAACAGATGTCTTTCAGCCCACGCGTTTGCACT-ACT  256
Sbjct  129248  TGCAAAAAGCTGTCTGCCATCACGAACAGATG--TTTCAGCCCACGCGTTTGCGCTTGCT  129305
Query  257     GT-C-CTACT--TCTCTGAAG-CGGA---CATAAGCGTCTACCGGTGGAACGCTAAATGT  308
Sbjct  129306  GTCCGCAACTGCTCACCGAGCCCGGACCGCA-AAGCGTC-ACCGGTGGAACGCTAAATGT  129363
Query  309     [illegible]C-ACAATT  367
Sbjct  129364  [illegible]TCTGCAATT  129419
Query  368     [illegible]ACGCAGCG-TCGCCAAGC-AGG  422
Sbjct  129420  CATTTTGCATATAGCCCCTCCGCTTTCTCTCGCCCACTAC-TGGCGATCGCC-AGCGGGG  129477
Query  423     [illegible]
Sbjct  129478  TCAT-AAACAAACTGTGCTGCGGT-AAATGCACCTGTAACGCCGGGAATTGTTTGC-GAA  129534
Query  476     [illegible]
```

Read length = 15kb (Max 2Mb)
Average error rate = 15%
Human genome (30X, ONT) = ~~$5000~~ ~$3000
Human genome (30X, PAC) = ~~$14000~~ ~$10000

Genome assembly

Genome

Sequencing technology

Reads

Assembler

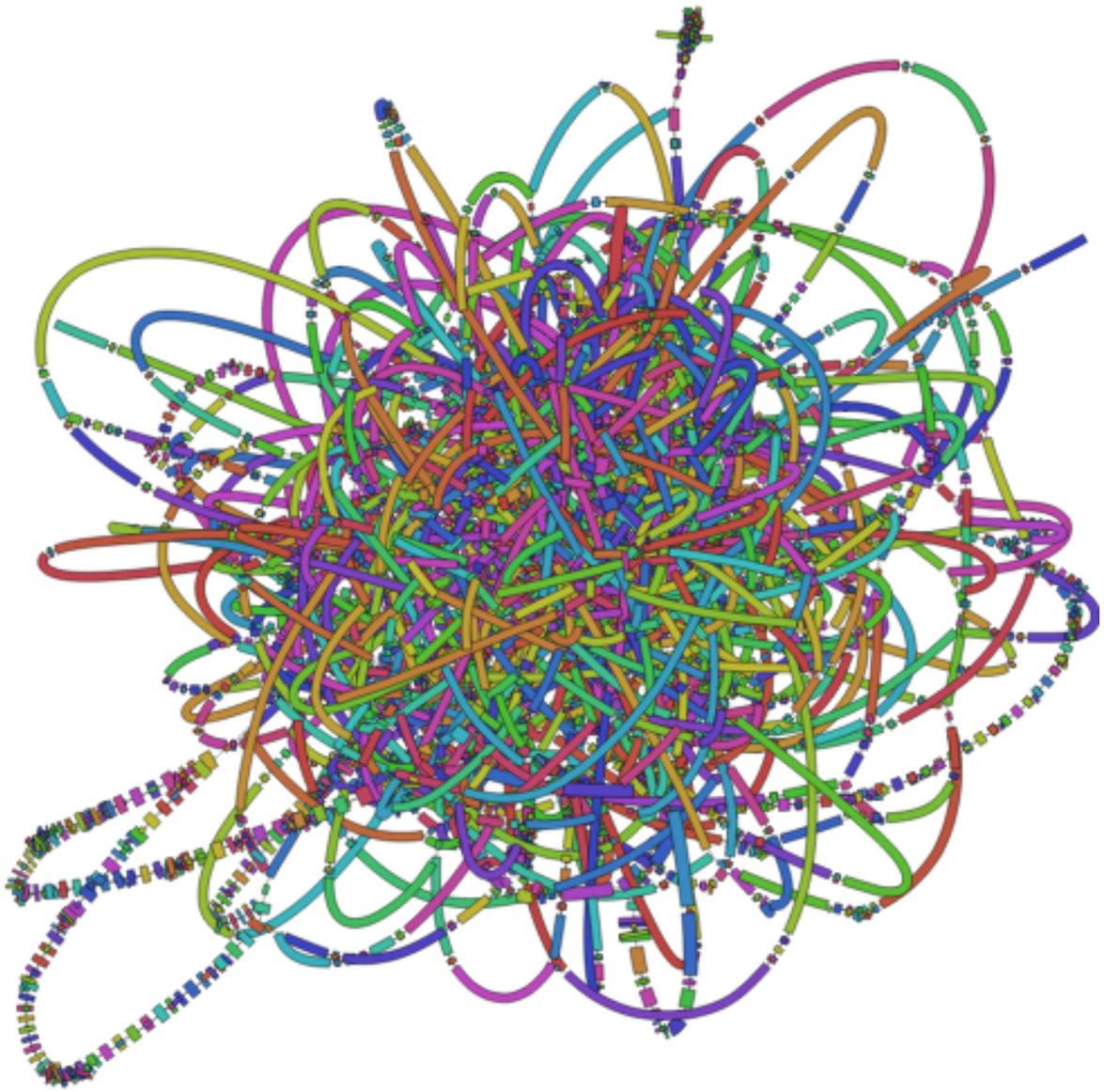
Overlaps

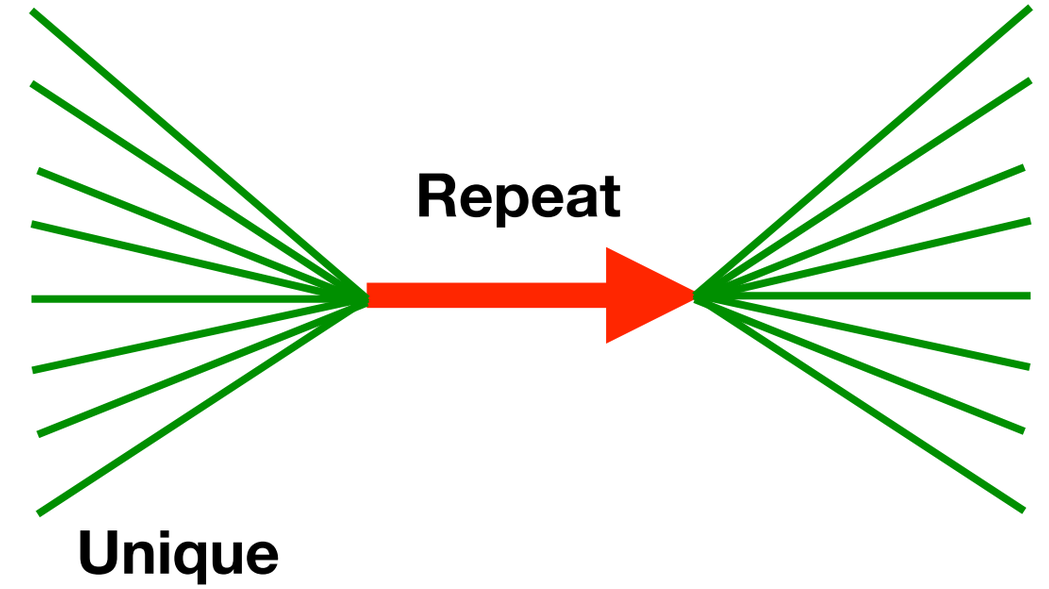
Assembly graph

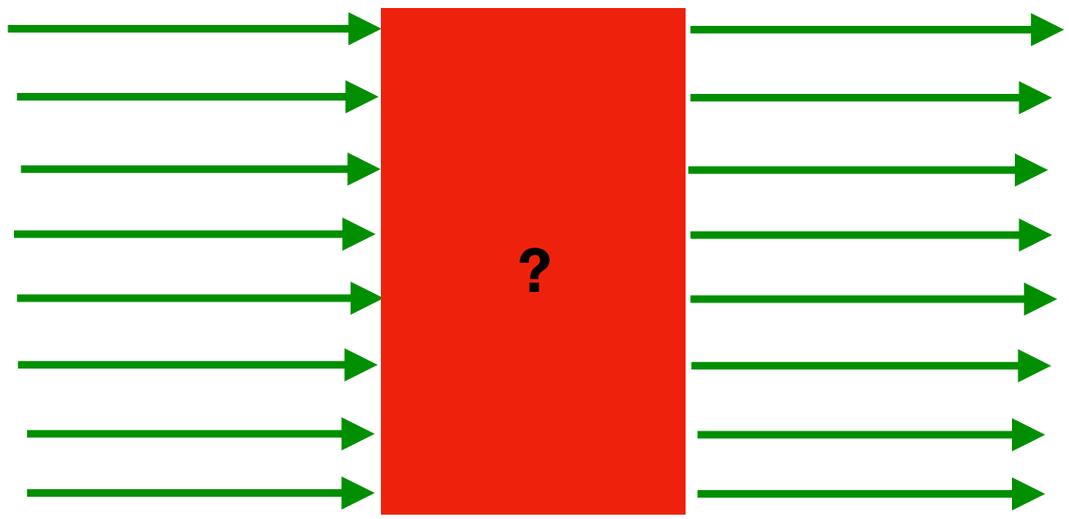
The genome path (sequence)

40 years of genome assembly

Genome assembly is complex







How do you get through?

How do we make “the” genome path?

Hybrid assembly: How can we combine short and long reads?

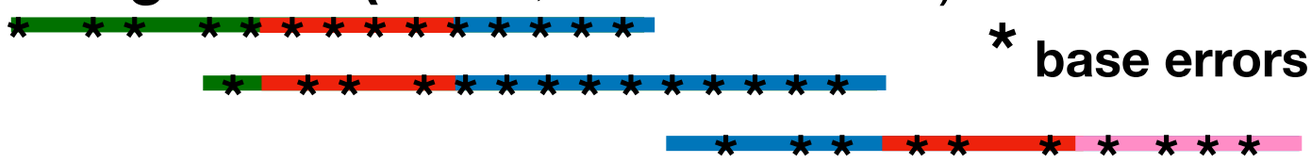
Sequencing technology

Genome

Identical repeats

Short reads (< 300 bp, base error < 0.1%**)**

Long reads (>10 kb, base error <15%)

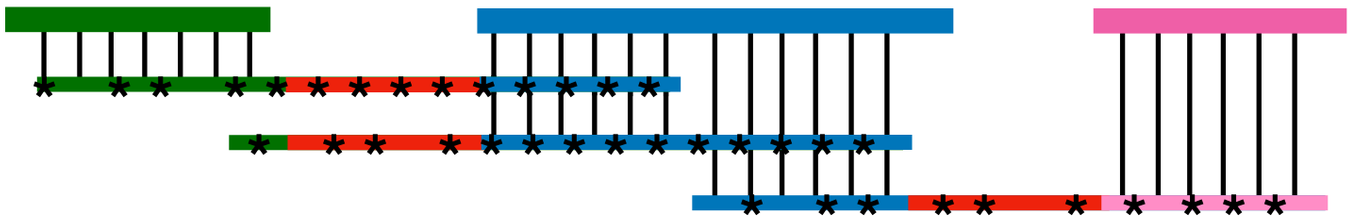


* **base errors**

Wengan Assembler

(1) Assemble short-reads

(2) Scaffold using long reads

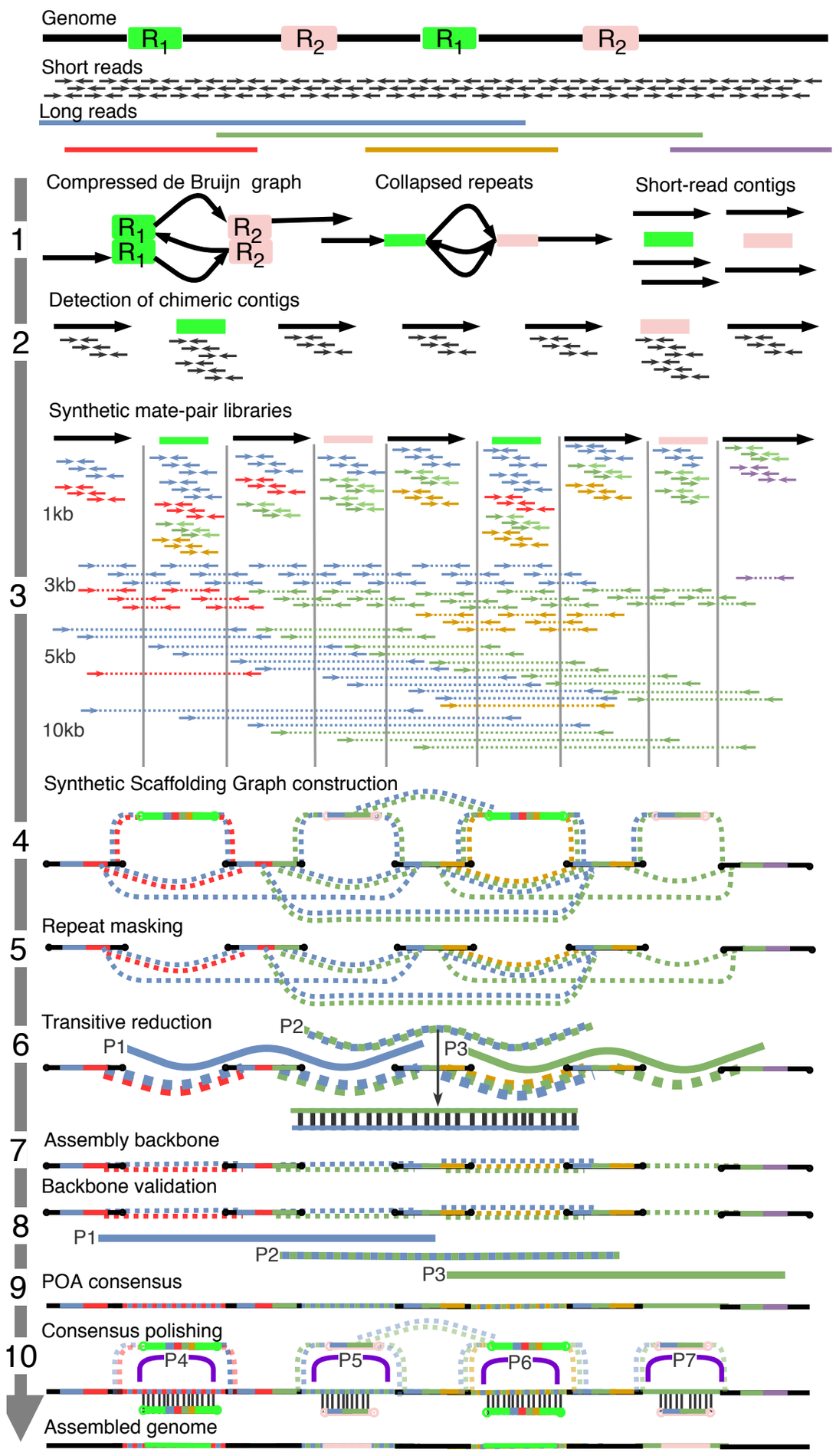


(3) Refine repetitive regions (polishing)

The resulting assembly is both *contiguous and accurate*

Wengan: a new assembly paradigm

- **Full** hybrid assembler.
- Avoids entirely all-vs-all read comparisons (**fast**).
- A new assembly graph (*GoogleMaps*).
- 1.5 years of development.
  - ~20k lines of code (C++, PERL)
- https://github.com/adigenova/wengan



- **Di Genova, A**. (2018). Fast-SG: an alignment-free algorithm for hybrid assembly. *GigaScience*, *7*(5).
- **Di Genova, A**. (2021). Wengan: Efficient and high-quality hybrid de novo assembly of human genomes. *Nature Biotechnology*.

Wengan

ARTICLES
https://doi.org/10.1038/s41587-020-00747-w
nature biotechnology

Check for updates

OPEN

Efficient hybrid de novo assembly of human genomes with WENGAN

Alex Di Genova[1,2] ✉, Elena Buena-Atienza[3,4], Stephan Ossowski[3,4] and Marie-France Sagot[1,2] ✉

Generating accurate genome assemblies of large, repeat-rich human genomes has proved difficult using only long, error-prone reads, and most human genomes assembled from long reads add accurate short reads to polish the consensus sequence. Here we report an algorithm for hybrid assembly, WENGAN, that provides very high quality at low computational cost. We demonstrate de novo assembly of four human genomes using a combination of sequencing data generated on ONT PromethION, PacBio Sequel, Illumina and MGI technology. WENGAN implements efficient algorithms to improve assembly contiguity as well as consensus quality. The resulting genome assemblies have high contiguity (contig NG50: 17.24–80.64 Mb), few assembly errors (contig NGA50: 11.8–59.59 Mb), good consensus quality (QV: 27.84–42.88) and high gene completeness (BUSCO complete: 94.6–95.2%), while consuming low computational resources (CPU hours: 187–1,200). In particular, the WENGAN assembly of the haploid CHM13 sample achieved a contig NG50 of 80.64 Mb (NGA50: 59.59 Mb), which surpasses the contiguity of the current human reference genome (GRCh38 contig NG50: 57.88 Mb).

WENGAN is a Mapudungun word.

WENGAN means "Making the path".