# The wonderful world of long-read genome assembly

Rayan Chikhi

Institut Pasteur & CNRS

ERGA workshop 2022

1

# Hello

- PI, Bioinformatics algorithms lab @ Institut Pasteur
- CV: PhD@ENS Rennes, Postdoc@PSU, CNRS

# Research:

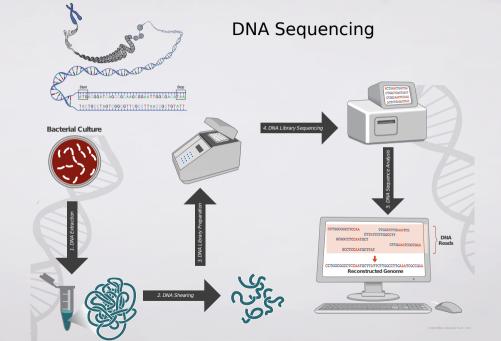
- *de novo* assembly
- k-mer methods
- metagenomics
- large-scale bioinfo



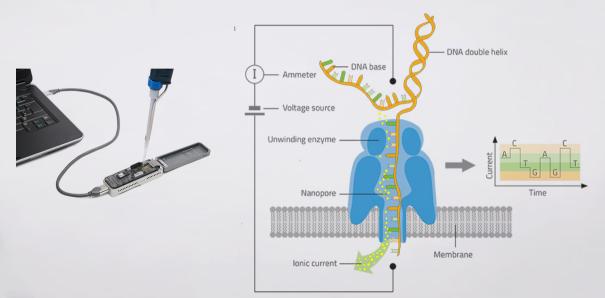
@RayanChikhi on Twitter



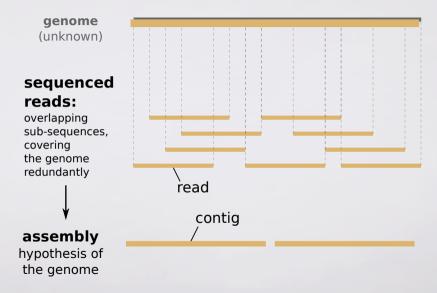
http://rayan.chikhi.name



# Long-read, portable DNA sequencing (Oxford Nanopore)

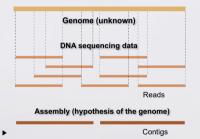


# Genome assembly



# 44 years of genome assembly

- 1977: First complete genome assembled (phi X 174)
- > 2003: Human Genome Project completed
- > 2014: First \$1,000 genome
- 2021: Truly completed (Telomere-2-Telomere)





(Staden 1979) "With modern fast sequencing techniques and suitable computer programs it is now possible to sequence whole genomes without the need of restriction maps."



The many applications of assembly

## Reconstruct genomes

- transcriptomes
- metagenomes
- genes
- Phylogeny of species
- Evolution of genes
- Find novel insertions
- SNPs in non-model organisms
- cell-free DNA struct. variants
- Pangenomics
- ...
- Any major one I'm missing?

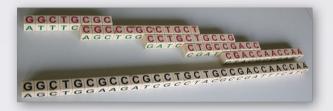
***     **	Cons 2	-	Qual 0	•	💷 Insert Edit Mod	es >> 🔟 Cutoff	s Undo Next S	earch Command	s >> Settings >>	•		Quit Help :
+11     NC_012967     gggtaccagaacatggcggcaacaggaacgccgggtlcacgcgcatatgttatggatacggtattcttag       +18     _cer_sxa_0_     gggtaccagaacatggcggcaacaggaacgccgggt       +19     _cer_sxa_10_     gggtaccagaacatggcggcaacaggaacgccgggt       +200     _cer_sxa_10_     gggtaccagaacatggcggcaacaggaacgccggg       +3663     _cer_sxa_10_     gggtaccagaacatggcggcaacaggaacgccgg       +3664     _cer_sxa_10_     gggtaccagaacatggcggcaacaggaacgcc       -33401     SRN030257.2989     CCAGAACATGGGGCGCAAACAGGAACGCGGGGTGCAC       -33402     SRN030257.1128     CAGAAACATGGGGCGCAAACAGGAACGCGGGGTGCACG       -33403     SRN030257.1128     CAGAACATGGGGCGCAAACAGGAACGCGGGGTGCACG       -33404     SRN030257.1204     AGAACATGGGGCCGGGGTGCACGG       -33405     SRN030257.1204     AGAACATGGGGCCGGGGTGCACGG       -33405     SRN030257.1204     AGAACATGGGGCCGGGGTGCACGGGTGCACGG       -33405     SRN030257.1204     AGAACATGGGGACCCGGGGTGCACGG       -33404     SRN030257.1204     AGACATGGGAACAGGAACGGGGGTGCACGGGTGCACGG       -33405     SRN030257.1403     ACATGGGAAACAGGAACGGGGGGGGGGCGCGGGTGCACGGG       -33407     SRN030257.1403     ACATGGGAAAACAGGAACGCGGGGTGCACGGGGTGCACGGG	<<	<	<b>&gt;</b>	>>								
*18     cer_sxa_0_     Eggtaccagaacatggcggcaaacaggaacgccgggt       *19     cer_sxa_0_     Eggtaccagaacatggcggcaaacaggaacgccggg       *20     cer_sxa_120_     Eggtaccagaacatggcggcaaacaggaacgccggg       *3661     cer_sxa_130_     Eggtaccagaacatggcggcaaacaggaacgccg       *3464     cer_sxa_120_     Eggtaccagaacatggcggcaaacaggaacgcc       *3405     SR030257.883     ACCAGAACATGCGGGCAAAACAGGAACCCGGGTGCAC       *3404     SR030257.1128     CAGAACATGCGGCAAAACAGGAACCCGGGTGCAC       *3403     SR030257.7128     CAGAACATGGCGGCAAAACAGGAACCCGGGTGCAC       *3404     SR030257.7128     CAGAACATGGGGCAAAACAGGAACCCGGGTGCAC       *3404     SR030257.7129     AGAACATGGGGCAAAACAGGAACCCGGGTGCACC       *3404     SR030257.7207     AGAACATGGGGCAAAACAGGAACCCGGGTGCACCGGGTGCAC       *3405     SR030257.7209     GAACATGGGGCAAAACAGGAACCCGGGTGCACCGGGTGCACCGG       *3404     SR030257.7209     GAACATGGGGCAAAACAGGAACCCGGGTGCACCGGGTGCACCGG       *3405     SR030257.1413     AACATGGGGGCAAAACAGGAACCCGGGTGCACCGGGT       *3409     SR030257.1453     ACATGGGGGCAAAACAGGAAACCGGGGTGCACCGGGTGCACCGGGA       *3410     SR030257.1453     ACATGGGGCAAAACAGGAAACCGGGGTGCACCGGGTGCACCGGGT </td <td></td> <td>1</td> <td></td>		1										
+19     cer_sxa.60     gggtaccagaacatggcggcaaacaggaacgccggg       +20     cer_sxa.130     gggtaccagaacatggcggcaacaggaacgccggg       +3663     cer_sxa.210     gggtaccagaacatggcggcaacaggaacgcc       +3664     _eer_sxa.210     gggtaccagaacatggcggcaacaggaacgcc       -33401     SR030257.2939     CLCAGAACATGCGGGCAAACAGGAACCCGGGTGCAC       >34404     SR030257.1294     CAGAAACATGCGGGCAAACAGGAACCCGGGTGCAC       *3404     SR030257.1204     CAGAAACTGGCGGCAAACAGGAACCCGGGTGCACGG       *3405     SR030257.1204     AGAAACTGGCGGCAAACAGGAACCCGGGTGCACGG       *3406     SR030257.1204     AGAACTGGCGGCAAACAGGAACCCGGGTGCACGG       *3405     SR030257.1204     AGAACTGGCGGCAAACAGGAACCCGGGTGCACGG       *3405     SR030257.1204     AGAACTGGCGGCAAAACAGGAACCCGGGTGCACGG       *3405     SR030257.1204     AGACTGGCGGCAAACAGGAACCCGGGTGCACGGG       *3405     SR030257.1403     ACATGGCGGCAAAACAGGAACCCGGGTGCACGGGT       *3408     SR030257.1403     ACATGGCGGCAAAACAGGAACCCGGGTGCACGGGTGCACGGGT       *3410     SR030257.1403     ACATGGCGGCAAACAGGAACCCGGGTGCACGGGTGCACGGGTGCACGGGTGG       *3411     SR030257.5986     GGCGGCGAAAACAGGAACCCGGGTGCACGGGTGGG <t< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>cgcatatcgt</td><td>tatggatacg</td><td>cgtatcggta</td><td>ttcttcag</td></t<>									cgcatatcgt	tatggatacg	cgtatcggta	ttcttcag
+20     cer_sxa_123     BBB taccagaacatggcggcaacaggaacgccg       +3663     cer_sxa_100     BBB taccagaacatggcggcaacaggaacgcc       +3664     cer_sxa_200     BBB taccagaacatggcgcaacaggacgcc       -33401     SRR030257.8883     ACCAGAACATGCGGGGTGCACCGGGTGCAC       -343403     SRR030257.2883     ACCAGAACATGGCGGCAAACAGGAACCGGGGTGCAC       -33405     SRR030257.1128     CAGAACATGGCGGCAAACAGGAACCGGGGTGCACC       -33405     SRR030257.2104     CAGAACATGGCGGCAAACAGGAACCCGGGTGCACCG       -33405     SRR030257.2104     AGAACATGGCGGCAAACAGGAACCCGGGTGCACCG       -33405     SRR030257.2757     GAACATGGCGGCAAACAGGAACCCGGGTGCACCGG       -33405     SRR030257.2757     GAACATGGCGGCAAACAGGAACCCGGGTGCACCGG       -33405     SRR030257.1453     ACATGGCGGCAAACAGGAACCCGGGTGCACCGG       -33405     SRR030257.1453     ACATGGCGGCAAACAGGAACCCGGGTGCACCGGGT       -33405     SRR030257.1453     ACATGGCGGCAAAACAGGAACCCGGGTGCACCGGGT       -33410     SRR030257.1453     ACATGGCGGGCAAAACAGGAACCCCGGGTGCACCGGGT       -33412     SRR030257.1821     GGCGGCGAAAACAGGAACCCGGGGTGCACCGGGTGCACCGGGTGCACCGGGTGCACCGGGTGCACCGGGTGCACCGGGTGCACCGGGTGCACCGGGTGCACCGGGTGCACCGGGTGCACCGGGTGCACCGGGGTGCACCGGGGTGCACCGGGGTGCACCGGGGTGCACCGGGGTGCACCGGGTGCACCGGGGTGCA												
+3663   _cer_sxa_190   _gggtaccagaacatggcggcaaacaggaacgc     +3664   _cer_sxa_280   _gggtaccagaacatggcggcaaacaggaacgc     -3340   SR030257.828   CACGAAAACATGGCGGCGGGGCGGGGGCAG     -33403   SR030257.2989   CCAGAAACTGGCGGCAAACAGGAACCCGGGTGCAG     -33404   SR030257.1298   CAGAAACTGGCGGCAAACAGGAACCCGGGTGCAGC     -33405   SR030257.1204   AGAAACTGGCGGCAAACAGGAACCCGGGTGCACGG     -33404   SR030257.1204   AGAAACTGGCGGCAAACAGGAACCCGGGTGCACGG     -33405   SR030257.1204   AGAACTGGCGGCAAACAGGAACCCGGGTGCACGG     -33405   SR030257.1204   AGAACTGGCGGCAAAACAGGAACCCGGGTGCACGG     -33405   SR030257.1204   AGACTGGCGGCAAAACAGGAACCCGGGTGCACGGG     -33405   SR030257.1205   GAACTGGCGGCAAAACAGGAACCCGGGTGCACGGG     -33405   SR030257.1755   ACATGGCGGCAAAACAGGAACCCGGGTGCACGGGTGCACGGGTGCACGGGTGCACGGGGGCAAACAGGAAACCGGGAACCGGGGTGCACGGGGGCGCGGGTGG     -33410   SR030257.1462   CATGGCGGCAAAACAGGAACCCGGGTGCACGGGGGGCGGGG     -33413   SR030257.1462   CAGGGCGCAAACAGGAACCCGGGTGCACGGGGGGCGGGG     -33414   SR030257.5986   GGCGGCAAAACAGGAACCCGGGTGCACGGGGGCCGGGG     -33415   SR030257.3429   GCGGCGAAAACAGGAACCCGGGTGCACGGGGGCCGGGGGGGG												
+3664												
-33401     SRR030257, 1883     CCRGARACATGÉGGÉCARACAGÉGACÉCCEGGETECA       -33402     SRR030257, 1883     CCRGARACATGÉGCÉCARACAGÉGACÉCCEGGETECAC       -33403     SRR030257, 128     CAGARACATGÉCEGGEARACAGÉGACÉCCEGGETECACC       -33404     SRR030257, 128     CAGARACATGÉCEGGEARACAGÉGACÉCCEGGETECACC       -33405     SRR030257, 120     AGARCATGÉCEGGEARACAGÉGACÉCCEGGETECACCE       -33405     SRR030257, 1204     AGARCATGÉCEGGEARACAGÉGACÉCCEGGETECACCE       -33405     SRR030257, 2602     GARCATGÉCEGCEARACAGÉGACÉCCEGGETECACCE       -33405     SRR030257, 143     ACATGÉCEGCEARACAGÉGACÉCCEGGETECACCEG       -33409     SRR030257, 1453     ACATGÉCEGCEARACAGÉGACÉCCEGGETECACCEGGET       -33410     SRR030257, 1453     ACATGÉCEGCEARACAGÉGACÉCCEGGETECACCEGGET       -33411     SRR030257, 1453     CATGÉCEGCEARACAGÉGACÉCCEGGETECACCEGGET       -33413     SRR030257, 1821     GÉCEGCECARACAGÉGACCECGÉGETEGACCEGGET       -33414     SRR030257, 1821     GÉCEGCEGARACAGÉGACCECGÉGETEGACCEGGETEG												
-33402     SRR030257.2989     CCAGGACATGEGGGCAAACAGGGACGCCGGGT ECAC       +33403     SRR030257.2089     CCAGGACATGEGGGCAAACAGGAACCGCGGGT ECACG       -33404     SRR030257.1204     AGAACATGEGCGCAAACAGGAACCGCGGGT ECACGC       -33405     SRR030257.2073     AGAACATGEGCGCAAACAGGAACCGCGGGT ECACGC       -33405     SRR030257.1204     AGAACATGEGCGCAAACAGGAACCGCGGGT ECACGC       -33406     SRR030257.2767     GAACATGEGCGCAAACAGGAACCGCGGGT ECACGCG       -33407     SRR030257.1413     AAACATGEGCGCAAACAGGAACCGCGGGT ECACGCG       -33408     SRR030257.1413     AACATGEGCGCAAACAGGAACCGCGGGT ECACGCGG       -33409     SRR030257.1425     ACATGEGCGCAAAACAGGAACCGCGGGT ECACGCGG       -33410     SRR030257.1423     CATGEGCGCAAAACAGGAACCGCGGGT ECACGCGGAT       -33413     SRR030257.1423     CGGCGGCAAAACAGGAACGCCGGGT ECACGCGCAT       -33413     SRR030257.1521     CGGCGGCAAAACAGGAACGCCGGGGT ECACGCGCAT       -33414     SRR030257.2932     GGCGGCGAAAACAGGAACGCCGGGT ECACGCGCATATC       -33415     SRR030257.3228     GCGCGCAAAACAGGAACGCCGGGT ECACGCGCATATC       -33414     SRR030257.3428     GCGCGCAAAACAGGAACCCGGGT ECACCGCGCATATC       -33415     SRR030257.3428     GCGCGCAAAACA												
+33403     SRR030257,1128     CAGAACATGEGGGCAAACAGGAACCGCGGGTCCCGGGT       +33404     SRR030257,1204     AGAACATGEGGGCAAACAGGAACCGCGGGTCCCGGGTCCCGGC       -33405     SRR030257,204     AGAACATGEGGGCAAACAGGAACCGCGGGTGCCCGGGTGCCGCG       -33405     SRR030257,204     GAACATGEGGCCAAACAGGAACCGCGGGTGCCCGGGTGCCGCG       -33405     SRR030257,2602     GAACATGEGGCCAAACAGGAACCGCGGGTGCCCGGGTGCCCGCG       -33405     SRR030257,1255     GAACATGEGGCCAAACAGGAACCGCGGGTGCCCGGGTGCCCGCG       -33405     SRR030257,1453     AACATGEGGGCAAAACAGGAACCGCGGGTGCCCGGGTGCCCGGGT       -33405     SRR030257,1453     AACATGEGGGCAAAACAGGAACCGCGGGTGCCCGGGT       -33410     SRR030257,1623     CATGEGGGCAAAACAGGAACCGCGGGTGCCCGGGTGCCCGGGT       -33411     SRR030257,1821     GGCCGGCAAAACAGGAACCGCGGGTGCCCGGGTGC       -33413     SRR030257,1821     GGCCGGCAAAACAGGAACCGCGGGTGCCCGGGTGC       -33414     SRR030257,1821     GGCGGCGAAAACAGGAACCGCGGGTGC       -33415     SRR030257,3232     GGCGGCGAAAACAGGAACCGCGGGTGCACCGCGGTGC       -33415     SRR030257,323     GGCGGCGAAACAGGAACCGCGGGTGCACCGGGTGCACCGCGGTTGC       -33415     SRR030257,3232     GGCGGCGAAACAGGAACCGCGGGTGCACCGGGGTGCACCGCGGTTGC       -33415 <td< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></td<>												
+33404     SRR030257.7073     AGAACATGEGGGCAAACAGGAACGCCGGGTTCCACGC       -33405     SRR030257.1204     AGAACATGEGGGCAAACAGGAACGCCGGGTTCCACGC       -33405     SRR030257.1204     GAACATGEGGGCAAACAGGAACGCCGGGTTCCACGC       -33405     SRR030257.1204     GAACATGEGGGCAAACAGGAACGCCGGGTTGCACGC       -33405     SRR030257.2767     GAACATGEGGGCAAACAGGAACGCCGGGTTGCACGCG       -33409     SRR030257.1413     AACATGEGGGCAAACAGGAACGCCGGGTTGCACGCG       -33409     SRR030257.1453     ACATGEGGGCAAACAGGAACGCCGGGTTGCACGCGG       -33410     SRR030257.1453     ACATGEGGGCAAACAGGAACGCCGGGTTGCACGCGGT       -33411     SRR030257.1623     CTGECGGCAAAACAGGAACGCCGGGTTGCACGCGGGTTGCACGCGGGTTGCACGCGGGTTGCACGCGGTTGCACGCGGGTTGCACGCGGTTGCACGCGGTTGCACGCGGTTGCACGCGGTTGCACGCGGTTGCACGCGGTTGCACGCGGTTGCACGCGGGTTGCACGCGGTTGCACGGGTTGCACGCGGGTTGCACGCGGGTTGCACGCGG												
-33405     SRR030257.1204     AGAACAT GEGGGCAAACAGGAACCCCGGGT ECACCE       -33405     SRR030257.2602     GAACAT GEGGGCAAACAGGAACCGCGGGT ECACCEG       -33405     SRR030257.2767     GAACAT GEGCGGCAAACAGGAACCGCGGGT ECACCEG       -33405     SRR030257.1413     AACAT GEGCGGCAAACAGGAACCGCGGGT ECACCEGG       -33405     SRR030257.1413     AACAT GEGCGGCAAACAGGAACCCCGGGT ECACCEGG       -33405     SRR030257.1453     ACAT GEGCGGCAAACAGGAACCCCGGGT ECACCEGG       -33405     SRR030257.1453     ACAT GEGCGGCAAACAGGAACCCCGGGT ECACCEGGA       -33410     SRR030257.1453     CAT GEGCGGCAAACAGGAACCCCGGGT ECACCEGGAT       +33412     SRR030257.1821     CGGCGGCAAACAGGAACCCCGGGT ECACCEGGAT       -33413     SRR030257.1821     CGGCGGCAAACAGGAACCCCGGGT ECACCEGGAT       -33414     SRR030257.1821     CGGCGCAAACAGGAACCCCGGGT ECACCCGGAT       -33415     SRR030257.1821     CGGCGCAAACAGGAACCCCGGGT ECACCCGGAT TC       -33415     SRR030257.3232     GCGCGCAAAACAGGAACCCCGGGT ECACCCGGAT TC       -33415     SRR030257.3423     GCGGCGAAACAGGAACCCCGGGT ECACCCGGAT TTC       -33415     SRR030257.3423     GCGGCGAAACAGGAACCCCGGGT ECACCEGCGAT TTCG       -33415     SRR030257.3423     GCGGCG									c			
-33406     SRR030257.2602     GAACATGGCGCAAACAGGAACCCCGGGTCCACCGG       -33407     SRR030257.2767     GAACATGGCGCAAACAGGAACCCCGGGTCCACCGG       -33408     SRR030257.1413     AACATGGCGCAAACAGGAACCCCGGGTCCACCGG       -33409     SRR030257.1453     ACATGGCGCAAACAGGAACCCCGGGTGCACCGGG       -33410     SRR030257.1453     ACATGGCGGCAAACAGGAACCCCGGGTGCACCGGGTGCACCGGGA       +33411     SRR030257.1623     CATGGCGGCAAACAGGAACCCCGGGTGCACCGGGTGCACGGGA       -33413     SRR030257.1623     CATGGCGGCAAACAGGAACCCCGGGTGCACGGGGGCACGGGA       -33414     SRR030257.1623     CATGGCGGCAAACAGGAACCCCGGGTGCACGGGGGCACGGGA       -33414     SRR030257.1821     GGCGCGCAAACAGGAACCCCGGGTGCACGGGGGGCACGGGGA       -33414     SRR030257.2932     GGCGGCGAAACAGGAACCCCGGGTGCACGGGGGGGGGGA       -33415     SRR030257.3729     GGCGGCAAAACAGGAACCGCCGGGTGCACGCGGATATCG       -33415     SRR030257.3423     GCGGCAAAACAGGAACCCCGGGTGCACCGCGGATATCG       -33418     SRR030257.3423     GCGGCAAAACAGGAACCCGGGGTGCACCGGGGTGCACCGCGATATCG       -33419     SRR030257.1401     CGGCGAAACAGGAACCCGGGGTGCACCGGGTGCACCGCGATATCG       -33420     SRR030257.3423     GCGGCAAAACAGGAACCCGGGGTGCACCGGGTGCACCGCGATATCG       -33420 <t< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></t<>												
-33407     SRR030257.2767     GAACATGEGEGECAAACAGGAACGCCGGGTGCACGCG       -33408     SRR030257.1413     AACATGEGEGECAAACAGGAACGCCGGGTGCACGCGG       -33409     SRR030257.1455     ACATGEGEGCAAACAGGAACGCCGGGTGCACGCGG       -33410     SRR030257.1453     ACATGEGEGCAAACAGGAACGCCGGGTGCACGCGG       -33411     SRR030257.1453     CATGEGEGCAAACAGGAACGCCGGGTGCACGCGGGTGCACGCGGA       -33412     SRR030257.1821     CGGCGGCAAACAGGAACGCCGGGTGCACGCGGTG       -33413     SRR030257.1821     GGGCGGCAAACAGGAACGCCGGGTGCACGCGGTG       -33414     SRR030257.1821     GGGCGGCAAACAGGAACGCCGGGTGCACGC       -33415     SRR030257.5986     GGCCGGCAAACAGGAACGCCGGGTGCACGCGGTGC       -33415     SRR030257.3229     GGCGCGAAACAGGAACGCCGGGTGCACGCGGTGCACGCGGTTGCACGCGGTTGCACGCGGGTGCACGC												
-33409     SRR030257.1413     AACATGEGGGCAAACAGGAACGCCGGGT CACGCGC       +33409     SRR030257.1453     ACATGEGGGCAAACAGGAACGCCGGGT CACGCGC       +33410     SRR030257.1453     ACATGEGGCAAACAGGAACGCCGGGT GCACGCGCA       +33411     SRR030257.1453     CATGEGCGCAAACAGGAACGCCGGGT GCACGCGCA       +33412     SRR030257.1623     CATGEGCGCAAACAGGAACGCCGGGT GCACGCGCA       -33414     SRR030257.1821     GGCGGCAAACAGGAACGCCGGGT GCACGCGGT       -33414     SRR030257.2932     GGCGGCAAACAGGAACGCCGGGT GCACGCGGA       -33415     SRR030257.3293     GGCGGCAAACAGGAACGCCGGGT GCACGCGCATATC       -33416     SRR030257.3293     GCGGCGCAAACAGGAACGCCGGGT GCACGCGCATATC       -33417     SRR030257.3423     GCGGCGCAAACAGGAACGCCGGGT GCACGCGCGATATC       -33418     SRR030257.3423     GCGGCAAAACAGGAACGCCGGGT GCACGCGCGATATC       -33419     SRR030257.3423     GCGGCAAACAGGAACGCCGGGT GCACGCGCGATATC       -33419     SRR030257.3423     GCGGCAAACAGGAACGCCGGGT GCACGCGCGATATC       -33419     SRR030257.3423     GCGGCAAACAGGAACGCCGGGT GCACGCGCGATATC       -33419     SRR030257.3423     GCGGCAAACAGGAACGCCGGGT GCACGCGCGCATATC       -33419     SRR030257.3423     GCGGCAAACAGGAACGCCGGGT GCA												
+33409     SRR030257,1755     ACATGGCGGCAAACAGGAACGCCGGGTGCACGCGCGA       +33410     SRR030257,1453     ACATGGCGGCAAACAGGAACGCCGGGTGCACGCGGGA       +33411     SRR030257,1453     CATGGCGGCAAACAGGAACGCCGGGTGCACGCGGGTGCACGCGGGTGCACGCGGGTGCACGCGGGTGGAGGCGCGGGGAAGAGGACGCCGGGTGGGGGGGG												
+33410     SRR030257.1463     ACATGGCGGCAAACAGGAACCCCGGGTGCACCGCGGA       +33411     SRR030257.1623     CATGGCGGCAAACAGGAACCGCGGGTGCACCGCGGT       -33412     SRR030257.602     TGGCGGCAAACAGGAACCGCGGGTGCACCGGGTGCACGCGGTG       -33413     SRR030257.1821     GGCCGGCAAACAGGAACCGCGGGTGCACGCGGGTGCACGCGGTGG       -33414     SRR030257.1821     GGCCGGCAAACAGGAACCGCGGG       -33415     SRR030257.1938     GGCCGGCAAACAGGAACCGCGGGTGG       -33415     SRR030257.329     GGCGGCAAACAGGAACCGCGGGTGCACCGGGTGG       -33415     SRR030257.329     GGCGGCAAACAGGAACCGCGGGTGCACCGGGTGCACGCGGTTGCACGCGGGTGCACGCGGGTGCACGCGGGTGCACGCGGGTGCACGCGGGTGCACGCGGGTGCACGCGGGTGCACGGGGTGCACGGGGTGCACGCGGGTGCACGCGGGTGCACGGGGTGGCAGGGGGGGG												
+33411   SRR030257.1623   CATGGCGGCAAACAGGAACGCCGGGTGCACGCGCGAT     +33412   SRR030257.1620   TGGCGGCAAACAGGAACGCCGGGTGCACGCGGGTG     -33413   SRR030257.1821   GGCGGCAAACAGGAACGCCGGGTGCACGCGGGTG     -33414   SRR030257.1821   GGCGGCAAACAGGAACGCCGGGTGCACGCGGGTG     -33414   SRR030257.2932   GGCGGCAAACAGGAACGCCGGGTGCACGGGTGG     -33415   SRR030257.3729   GGCGGCAAACAGGAACGCCGGGTGCACGCGGGTGCACGCGCATATC     -33416   SRR030257.3729   GGCGGCAAACAGGAACGCCGGGTGCACGCGGGTTGCACGCGGATTCG     -33419   SRR030257.3423   GCGGCAAAACAGGAACGCCGGGTGCACGCGGGTTATCG     -33419   SRR030257.1401   CGGCAAAACAGGAACGCCGGGTGCACGCGGGTTATCG     -33420   SRR030257.1401   CGGCAAAACAGGAACGCCGGGTGCACGCGCGTATTCG     -33420   SRR030257.1425   CGGCAAAACGGGAACGCCGGGTGCACGCGCGATATCG     -33420   SRR030257.1425   CGGCAAAACGGGAACGCCGGGTGCACGCGCATATCG												
4:3412     SRR030257.6602     TGGCGGCAAACAGGAACCGCGGGTGG       -33413     SRR030257.1821     GGGCGGCAAACAGGAACCGCGGGTGCACGG       -33414     SRR030257.2932     GGGCGGCAAACAGGAACCGCGGGTGCACGGG       -33415     SRR030257.5936     GGCCGGCAAACAGGAACCGCGGGTGCACGGGTGGCACGCGGTGGCACGGCGGTGCACGCGCGCG												
-33413     SRR030257.1821     GGGCGGCAAACAGGAACGCCGGGTGCACGG       -33414     SRR030257.2932     GGGCGGCAAACAGGAACGCCGGGTGCAGGG       -33414     SRR030257.2932     GGGCGGCAAACAGGAACGCCGGGTGG       -33415     SRR030257.3729     GGCGGCAAACAGGAACGCCGGGTGGCAGCGCGATATC       -33416     SRR030257.3729     GCGGGCAAACAGGAACGCCGGGTGCAGCGCGATATC       -33417     SRR030257.3423     GCGGCAAACAGGAACGCCGGGTGCACGCGCATATCG       -33418     SRR030257.1422     GCGGCAAACAGGAACGCCGGGTGCACGCGCGATATCG       -33419     SRR030257.1401     CGGCAAACAGGAACGCCGGGTGCACGCGCGATATCG       -33420     SRR030257.1425     CGGCAAACAGGAACGCCGGGTGCACGCGCATATCGT       -33420     SRR030257.1125     CGGCAAAACAGGAACGCCGGGTGCACGCGCATATCGT									CUCHI			
-33414     SRR030257.2932     GGGGGGCAAACAGGGAACGCGGGG       +33415     SRR030257.5936     GGGCGGCAAACAGGAACGCCGGGG       -33416     SRR030257.3729     GGCGGCAAACAGGAACGCCGGGG GCACCGGGGTGCACGCGCATATC       +33417     SRR030257.3423     GCGGCAAACAGGAACGCCGGGG GCACGCGCGATATCG       +33418     SRR030257.3428     GCGGCAAACAGGAACGCCGGGG GCACGCGCGATATCG       +33419     SRR030257.1401     CGGCAAACAGGAACGCCGGGG GCACGCGCGATATCG       -33420     SRR030257.3555     CGGCAAACAGGAACGCCGGGG GCACGCGCGATATCG       -33421     SRR030257.1125     CGGCAAACAGGAACGCCGGGG GCACGCGCGATATCG									r			
+33415     SRR030257.5986     GGCGGCAAACAGGAACGCCGGGTGG       -33416     SRR030257.3729     GGCGGCAAACAGGAACGCCGGGTGCACGCGCATATC       +33417     SRR030257.3423     GCGGCAAACAGGAACGCCGGGTGCACGCGCATATCG       -33418     SRR030257.2482     GCGGCAAACAGGAACGCCGGGTGCACGCGCATATCG       -33419     SRR030257.1401     CGGCAAACAGGAACGCCGGGTGCACGCGCGATATCG       -33420     SRR030257.1401     CGGCAAACAGGAACGCCGGGTGCACGCGCATATCG       -33420     SRR030257.1125     CGGCAAACAGGAACGCCGGGTGCACGCCGATATCGT									0			
-33416     SRR030257.3729     G6CG6CAAACAG6AACGCCG6GTGCACGCGCGATATC       +33417     SRR030257.3423     GCGCCAAACAG6AACGCCGGGTGCACGCCGCATATCG       +33418     SRR030257.1402     GCGGCAAACAG6AACGCCGGGTGCACGCGGTTGCACGCGATATCG       -33429     SRR030257.1401     CGGCAAACAG6AACGCCGGGTGCACGCGCATATCG       -33420     SRR030257.1401     CGGCAAACAG6AACGCCGGGTGCACGCGCATATCG       -33420     SRR030257.1125     CGGCAAACAG6AACGCCGGGTGCACGCCGATATCGT												
+33417     SRR030257.3423     GCGGCANACAGGAACCGCGGGT GCACCGCGTATTG       +33418     SRR030257.2482     GCGGCANACAGGAACCGCGGGT GCACCGCGCATATCG       +33419     SRR030257.1401     CGGCAANCAGGAACCGCCGGGT GCACCGCGCATATCG       -33420     SRR030257.1401     CGGCAANCAGGAACCGCCGGGT GCACCGCGCATATCG       -33420     SRR030257.1451     CGGCAANCAGGAACCGCCGGGT GCACCGCCGATATCG       -33420     SRR030257.1125     CGGCAANACAGGAACCGCGGGT GCACCGCGCAATCGT									CGCATATC			
+33418     SRR030257.2482     GCGGCAAACAGGAACGCCGGGTGCACGCGCGATATCG       -33419     SRR030257.1401     CGGCAAACAGGAACGCCGGGTGCACGCGCGATATCG       -33420     SRR030257.3565     CGGCAAACAGGAACGCCGGGTGCACCGCGATATCG       -33421     SRR030257.1125     CGGCAAACAGGAACGCCGGGTGCACCGCGATATCG												
-33419 SRR030257.1401 CGGCAAACAGGAACGCCGGGTGCACGCGCGTTATCG -33420 SRR030257.1355 CGGCAAACAGGAACGCCGGGTGCACGCGCATATCG -33421 SRR030257.1125 CGGCAAACAGGAACGCCGGGTGCACGCGGGTGCACGCCGCATATCG												
-33420 SRR030257.3565 CGGCAAACAGGAACGCCGGGTGCACGCGCGATATCGT +33421 SRR030257.1125 CGGCAAACAGGAACGCCGGGGTGCACGCGCGCACATCGT												
+33421 SRR030257.1125 CGGCAAACAGGAACGCCGGGTGCACGCGCACATCGT												
										т		

Screenshot: MIRA



## 1. Assembly using strings

- Shortest Common Superstring (Kececioglu, Myers 1993)
- Greedy algorithms (CAP3 from Huang, Madan 1999)

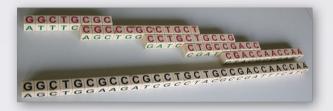


# 1. Assembly using strings

- Shortest Common Superstring (Kececioglu, Myers 1993)
- Greedy algorithms (CAP3 from Huang, Madan 1999)
- 2. Assembly using graphs: string graphs and de Bruijn Graphs (both from DIMACS'94)

A History of DNA Sequence Assembly, G. Myers, 2016

dBGs widely used across genomics (SPAdes: 13,000 citations; Trinity: 12,000 citations)



# 1. Assembly using strings

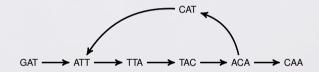
- Shortest Common Superstring (Kececioglu, Myers 1993)
- Greedy algorithms (CAP3 from Huang, Madan 1999)
- 2. Assembly using graphs: string graphs and de Bruijn Graphs (both from DIMACS'94)

A History of DNA Sequence Assembly, G. Myers, 2016

dBGs widely used across genomics (SPAdes: 13,000 citations; Trinity: 12,000 citations)

Modern genome assembly: graphs

- 1. Construct a graph
- 2. Nodes are reads (or k-mers)
- 3. Edges are overlaps



## Theory will say...

[Nagarajan 09]

4. Return a path of *minimal length* that traverses each node at least once.

# Assembly theory is somewhat unhelpful

Genome assembly is linear-time solvable.

Genome assembly is NP-hard.

If all **repeats** are **longer** than reads, Genome assembly is **polynomial**. (!)

If all repeats are either **shorter** than reads, or are **spanned** by reads, Genome assembly is **polynomial**, *and* with a **unique** solution.

[Pevzner et al, 2001]

[Medvedev, Brudno 2007]

[Nagarajan, Pop 2009]

[Nagarajan, Pop 2009] [Bresler, Bresler, Tse 2013]



# Yet, in practice..

- Illumina data: none of the theories apply
- Because graph is often disconnected
- So, can't frame the problem as finding a single path
- Contigs = all the unambiguous paths

- Long reads: "theory meets practice" appears possible

[Kamath et al, 2017]

Either way:

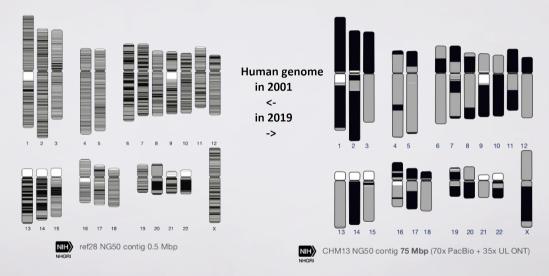
- > 100s of GB of input data (for eukaryotes)
- days/months of CPU time human: 1 CPU-month, 200 GB RAM (wtdbg2)

Recommended teaching material

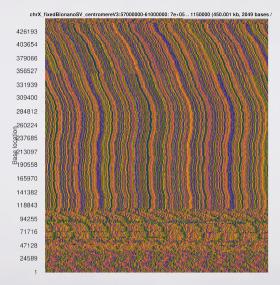
# Modeling biological problems in computer science: a case study in genome assembly @ Paul Medvedev S

Briefings in Bioinformatics, bby003, https://doi.org/10.1093/bib/bby003 Published: 30 January 2018 Article history ▼

# The human genome is challenging to assemble



# Why? centromeres, but not only



credit: D. Eccles, https://twitter.com/gringene\_bio/status/1102121757828210688

# Consensus caveats

Suppose a diploid organism with 2 haplotypes:

..AGCCTGAGTTC.. ..AGCATGATTTC..

Assembly usually results in a single consensus:

..AGCCTGATTTC..

Haplotype separation at chromosome scale is recent.

Letter | Open Access | Published: 07 December 2020

# Chromosome-scale, haplotype-resolved assembly of human genomes

Letter | Open Access | Published: 07 December 2020

Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads

# Genome assembly software is complex

- Coding:
  - PhD
  - or a team of engineers (1-2 years)
- Several not-so-independant components



Heuristics everywhere

"A genome assembly is like a good sausage, it is best to ignore how it was made"

(apocryphal) S. Gnerre, ALLPATHS assembler

# Genome assembly landscape, in 2022

- Short reads: don't, except for Hi-C
- Long reads: Oxford Nanopore, PacBio HiFi, PacBio CLR

Genome assembly landscape, in 2022

- Short reads: don't, except for Hi-C
- Long reads: Oxford Nanopore, PacBio HiFi, PacBio CLR



(Img: M. Watson)

Genome assembly landscape, in 2022

- Short reads: don't, except for Hi-C
- Long reads: Oxford Nanopore, PacBio HiFi, PacBio CLR



(Img: M. Watson)

But: short-read **methods** are making a come-back with long reads. de Bruijn graphs in Flye, rust-mdbg, LJA, Verkko & *k*-mer assembly validation.

k-mers

*k*-mer: any sequence of length *k* 

N.G. de Bruijn (1946), de Bruijn sequences <sup>1</sup>



C. Shannon (1948), information theory<sup>2</sup>

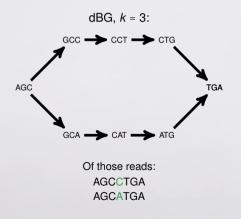


<sup>&</sup>lt;sup>1</sup>construct shortest sentence containing all *k*-mers exactly once <sup>2</sup>predict future data given past data, where past = last seen *k*-mer

# de Bruijn graph

A **de Bruijn** graph for a fixed integer *k*:

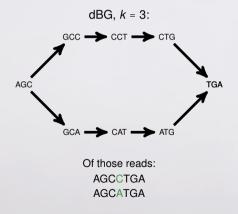
- 1. **Nodes** = all k-mers (substrings of length k) in the reads
- 2. **Edges** = all exact overlaps of length exactly (k 1)

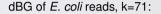


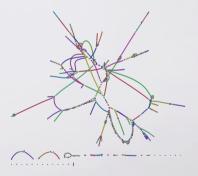
# de Bruijn graph

A **de Bruijn** graph for a fixed integer *k*:

- 1. **Nodes** = all k-mers (substrings of length k) in the reads
- 2. **Edges** = all exact overlaps of length exactly (k 1)

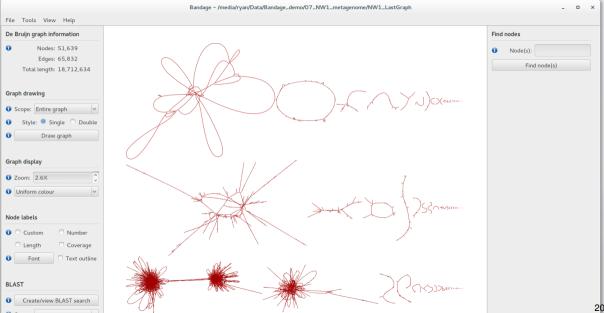


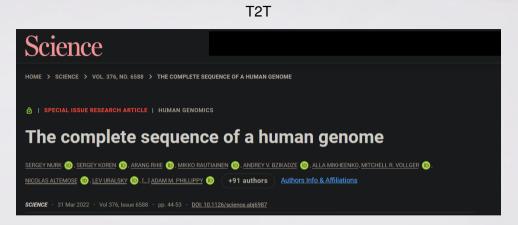






# Assembly graph visualization: Bandage and Bandage-NG





- First gap-less human genome
- 154 Mb contig N50 (retiring that metric)
- Added ≈ 200 Mbp compared to GRCh38 (centromeres mostly)
- CHM13: haploid, no Y chromosome
- 30x HiFi (HiCanu) + 120x ONT ultralong

# Han1

A second complete human genome. (\* not peer reviewed)

The first gapless, reference-quality, fully annotated genome from a Southern Han Chinese individual (KH. Chao, A. Zimin, M. Pertea, S. Salzberg) Only 4 authors!

How?:

- 39x HiFi (hifiasm), 35x ONT ultralong (Flye, discarded)
- scaffolded against CHM13 (with MaSuRCA)
- Semi-manual gap-closing
- JASPER for HiFi polishing

# Automatic near-T2T tools

#### Verkko

*Telomere-to-telomere assembly of diploid chromosomes* HiFi + UL + Hi-C, Strand-seq, trio

# - LJA

HiFi only

# Hifiasm

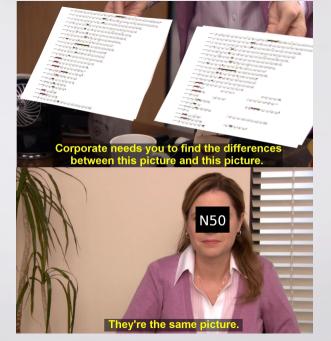
HiFi + Hi-C, trio

## **Relevance of assembly metrics**

- N50

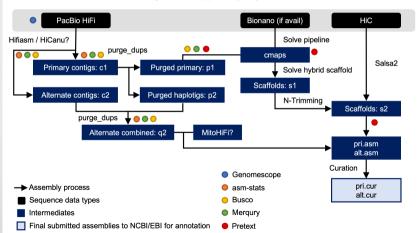
- Completeness / BUSCO

Is N50 relevant in the T2T era? less and less. BUSCO? Very important for spotting duplications



# VGP, ERGA, DToL, AfricaBP

- Sequence as much as possible
- Assemble it
- Challenges ahead (repetitiveness, ploidy, ...samples collection)



VGP assembly standard pipeline (v2.0) + essential QC

Is bacterial genome assembly solved?

In 2019, 30% of PacBio bacterial assemblies were fragmented



P. Marijon *et al*, Bioinformatics 2019 https://gitlab.inria.fr/pmarijon/knot



# Viral assembly

Serratus:

- 5M RNA-seqs aligned (10 PB)
- 50k assemblies, 28,000 vCPUs on AWS in a weekend
- Discovery of a new coronavirus species
- 10x expansion of RNA viruses species
- It remains challenging to assemble viruses from metagenomes
- Edgar et al, 2022



# 2) Fast (meta)genome assembly with accurate long reads



PacBio HiFi reads (~1% error rate)



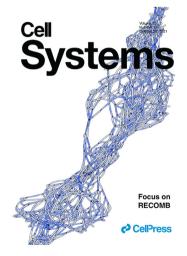
- Minimizer-space de Bruijn graphs
- Human genome assembled in 10 minutes / 10 GB RAM
- Pangenome of 661k bacteria

"Tricks": change of alphabet (ACGT->minimizers) discards 99% of the bases until last step

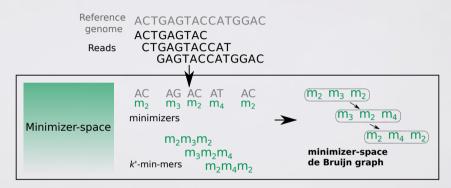








Minimizer-space de Bruijn graph

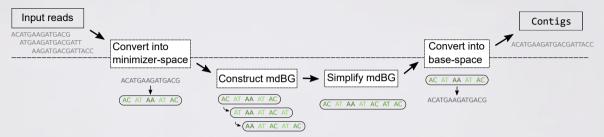


A minimizer-space de Bruijn graph is a de Bruijn graph over the minimizer alphabet. Nodes = k-min-mers,

Edges = exact overlaps between k-1 minimizers

Applied to whole-genome de novo assembly

From accurate HiFi (< 1% error-rate) reads



Whole human PacBio HiFi (HG002) 50x coverage:

Tool name	Peregrine	hifiasm	rust-mdbg
Wall-clock time	14h8m	58h41m	10m23s
Memory usage	188 GB	195 GB	10 GB
# contigs	8109	431	805
NG50 (Mbp)	18.2	88.0	16.1
Genome fraction	97.0%	94.2%	95.5%

# Conclusion



#### Take-aways:

- Genome assembly, nowadays done with long reads
- One of the two fundamental "sequence bioinformatics" problem along with alignment
- Data structures & algorithms play a huge role in methods
- Also, large amounts of Python/C++/Rust for making tools

# An open problem:

Highest-quality assembly from single-sample long reads only

# Thank you! Any questions?

Acknowledgements for this talk material: Pierre Marijon, Jean-Stéphane Varré, Adam Phillippy, Antoine Limasset, Camille Marchet, Brian Bushnell, Sergey Nurk, Marco Previtali, Paul Medvedev, Shaun Jackman, Guillaume Rizk, Ryan Wick, David Eccles, Mick Watson

# Sequence Bioinformatics @ Institut Pasteur



Y. Dufresne, R. Vicedomini, T. Lemane, C. Duitama, L. Blassel, F. Andreace

Funding: EC H2020, ANR Inception, ANR Prairie, ANR Transipedia, ANR SeqDigger



## En réponse à @ctitusbrown

# "Finding your way in life is like finding the genome in a De Bruijn graph: it is very easy to find \*a\* path, very hard to find \*the\* path".