Genome assembly

Dan Jones, Bioinformatics Institute

The Bioinformatics I

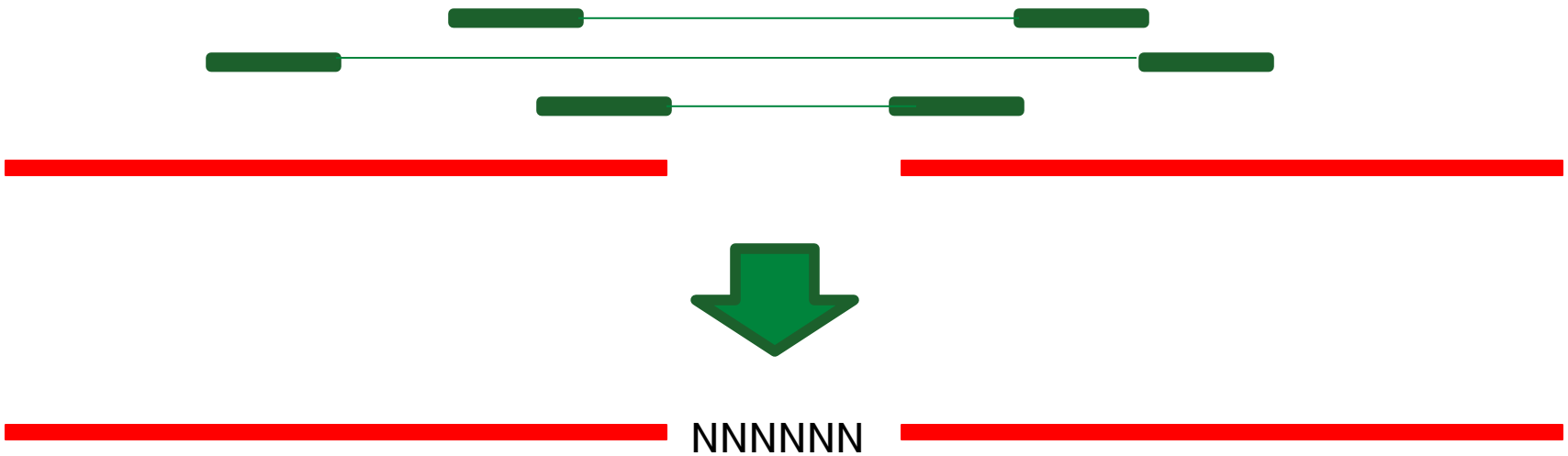
THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

Caveats

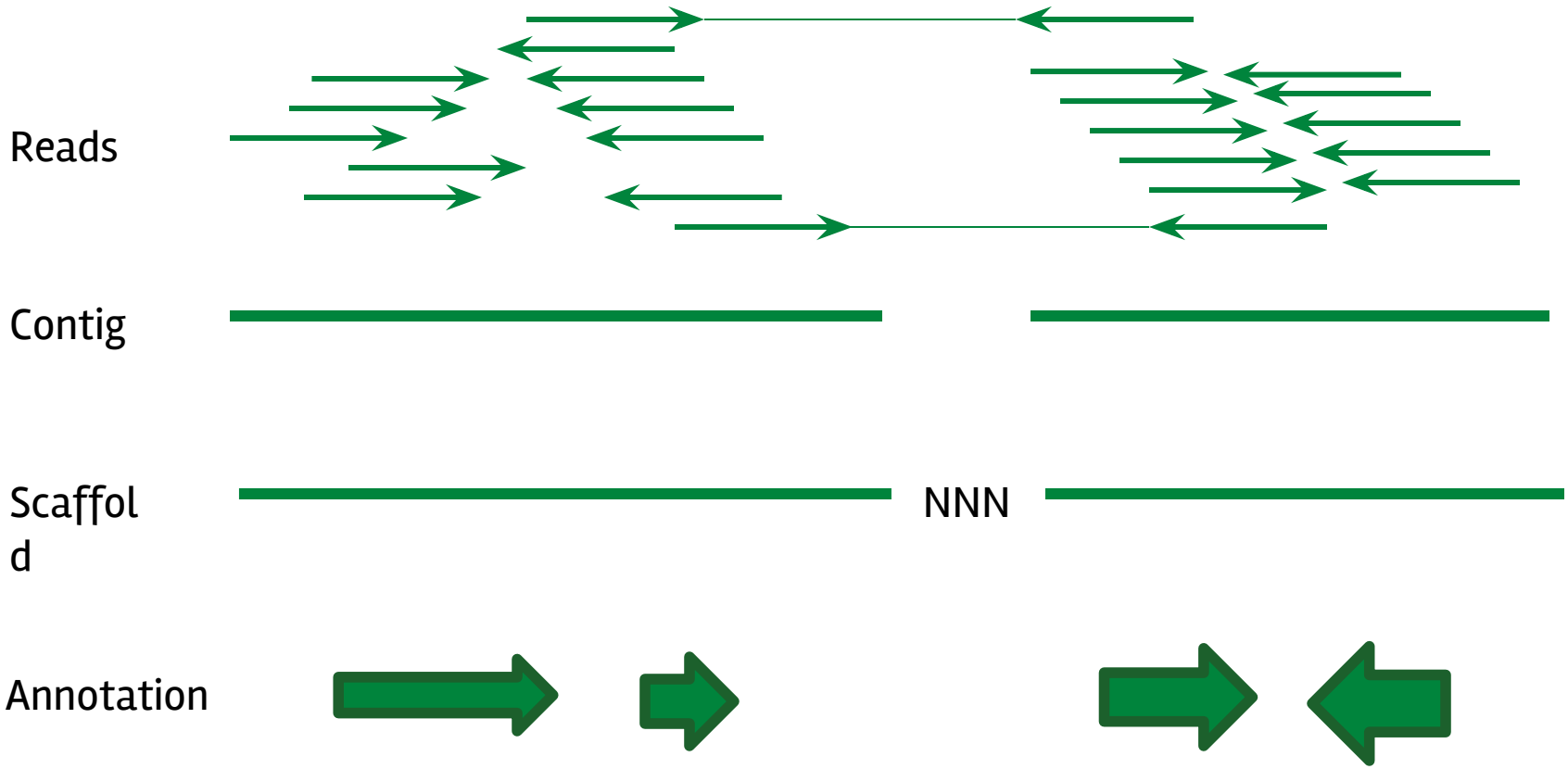
- Assembly of the genome is by far the most computationally intensive step:
  - Requires high-memory computers; CPU speed is not the limiting factor
    - Bacterial genomes: Standard desktops/laptops are fine
    - Fungal / very small eukaryote: requires a small cluster e.g. a 40 MB fungal genome required ~30 GB of RAM
    - Human-size genome: ~100 GB
    - More sequencing for greater depth = more memory
    - Adjustment of critical assembly parameters = more or less memory
    - Newer assembly programs = less memory (but there is a theoretical lower limit)

It’s more useful to have different library types (e.g. mate-pair) with a range of insert sizes, than it is to simply do more sequencing..

…as this helps with scaffolding and therefore increasing the size of your scaffolds.



De novo assembly



Assembly strategy

- De Bruijin graph method (Velvet)
  - Designed to handle the problem of aligning millions or billions of reads in a computationally sane manner
  - Reads are broken down into k-mers of a particular size (a critical parameter) and a De Bruijn graph assembled

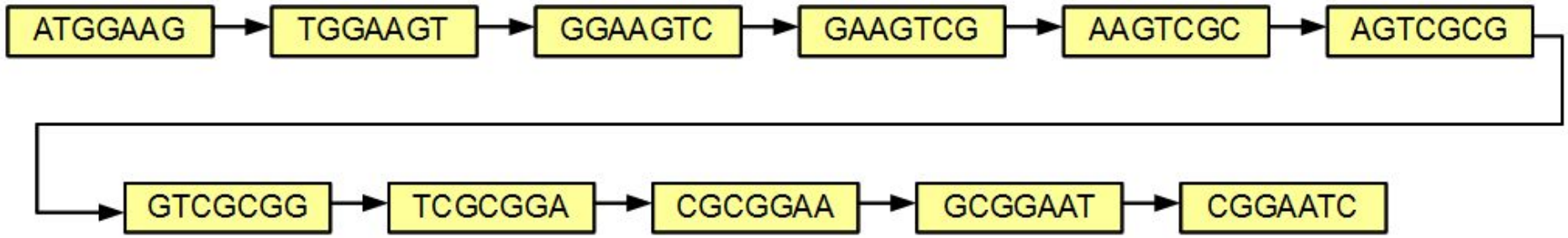
- Overlap graph method (Edena)
  - Designed for a smaller number of reads: can’t really handle larger NGS datasets
  - Some assemblers are hybrids: partially De Bruijn, partially OL

sequence **ATGGAAGTCGCGGAATC**

7mers

```
ATGGAAG
 TGGAAGT
  GGAAGTC
   GAAGTCG
    AAGTCGC
     AGTCGCG
      GTCGCGG
       TCGCGGA
        CGCGGAA
         GCGGAAT
          CGGAATC
```

de Bruijn graph



Assemble genome

Assembly strategy

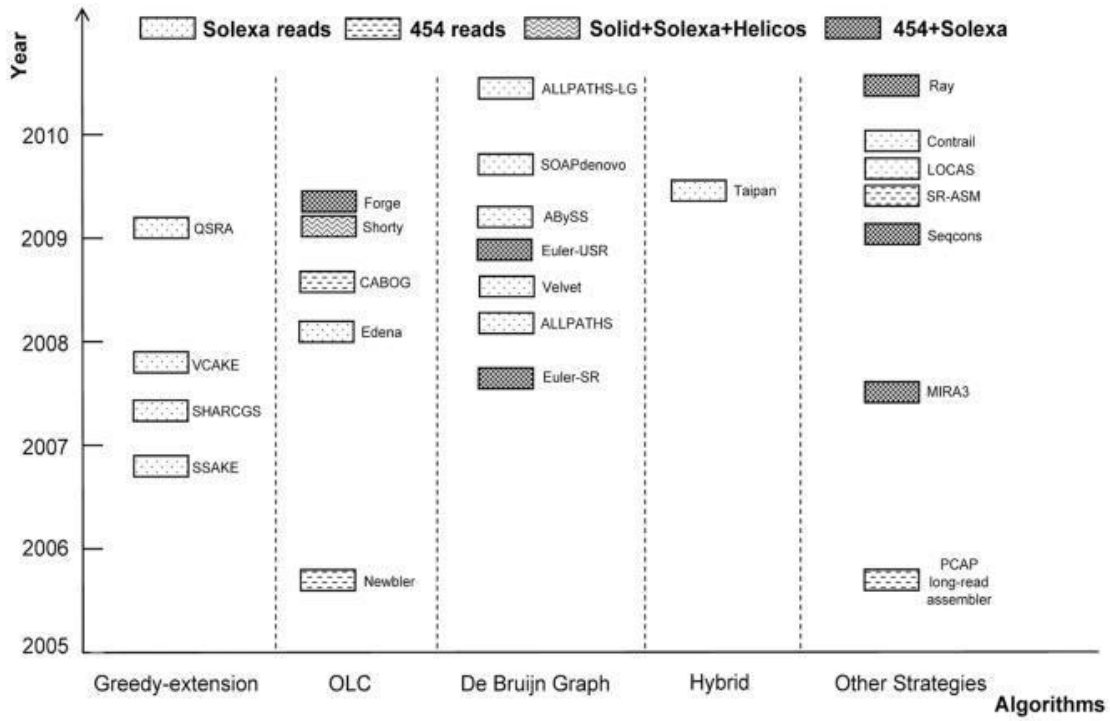
- Other approaches

MaSuRCA: Assembles reads into “super reads” using de Bruijn graphs, and uses OLC to assemble “super reads”.. used to generate the gigantic Loblolly pine genome

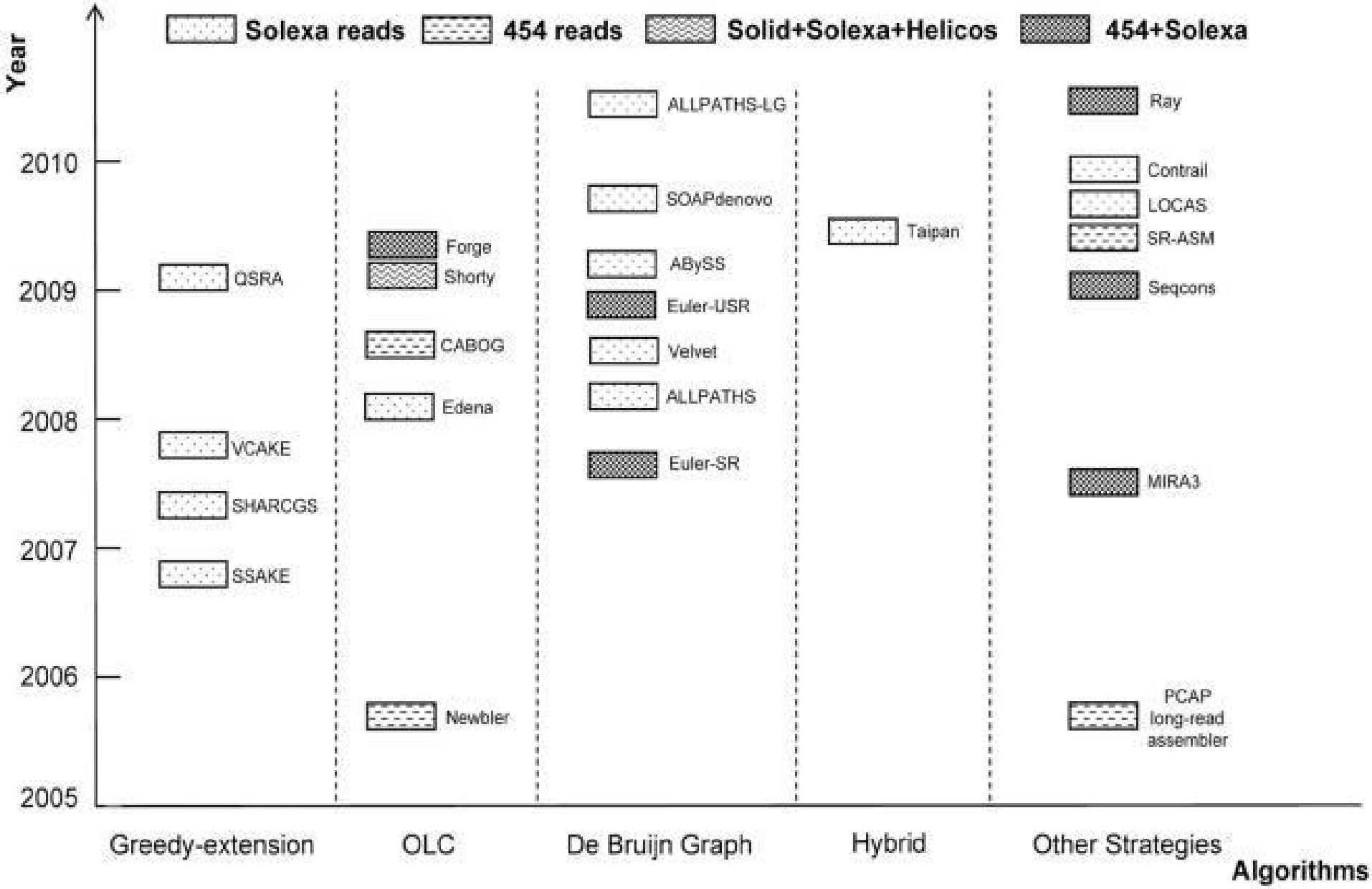
MIRA4: A very flexible and useful assembler for small genomes; can also call SNPs

Ray: Parallellisable de Bruijn assembler, which is useful for metagenomic assemblies

A bewildering array of short read assemblers



PLoS One. 2011; 6(3): e17915.


Solexa reads
454 reads
Solid+Solexa+Helicos
454+Solexa
Year
2010
2009
2008
2007
2006
2005
QSRA
VCAKE
SHARCGS
SSAKE
Forge
Shorty
CABOG
Edena
Newbler
ALLPATHS-LG
SOAPdenovo
ABySS
Euler-USR
Velvet
ALLPATHS
Euler-SR
Taipan
Ray
Contrail
LOCAS
SR-ASM
Seqcons
MIRA3
PCAP long-read assembler
Greedy-extension
OLC
De Bruijn Graph
Hybrid
Other Strategies
Algorithms

Using “mate-pair” libraries

- A “mate pair” library is sequenced in exactly the same way but library preparation is different
- Orientation of the reads is:

  < ----------<INSERT>---------->

  (as opposed to)

  >---------->INSERT←----------<

  ..for standard paired-end libaries

  Some assemblers will deal with this, some require that you reverse-complement the sequence..

Using “mate-pair” libraries

Of great importance in spanning repetitive regions (but also can use long reads)

As always, do an initial assembly and check that the mate-pairs are behaving as expected:

Object lesson: *V. pirina* mate-pair libaries from BGI

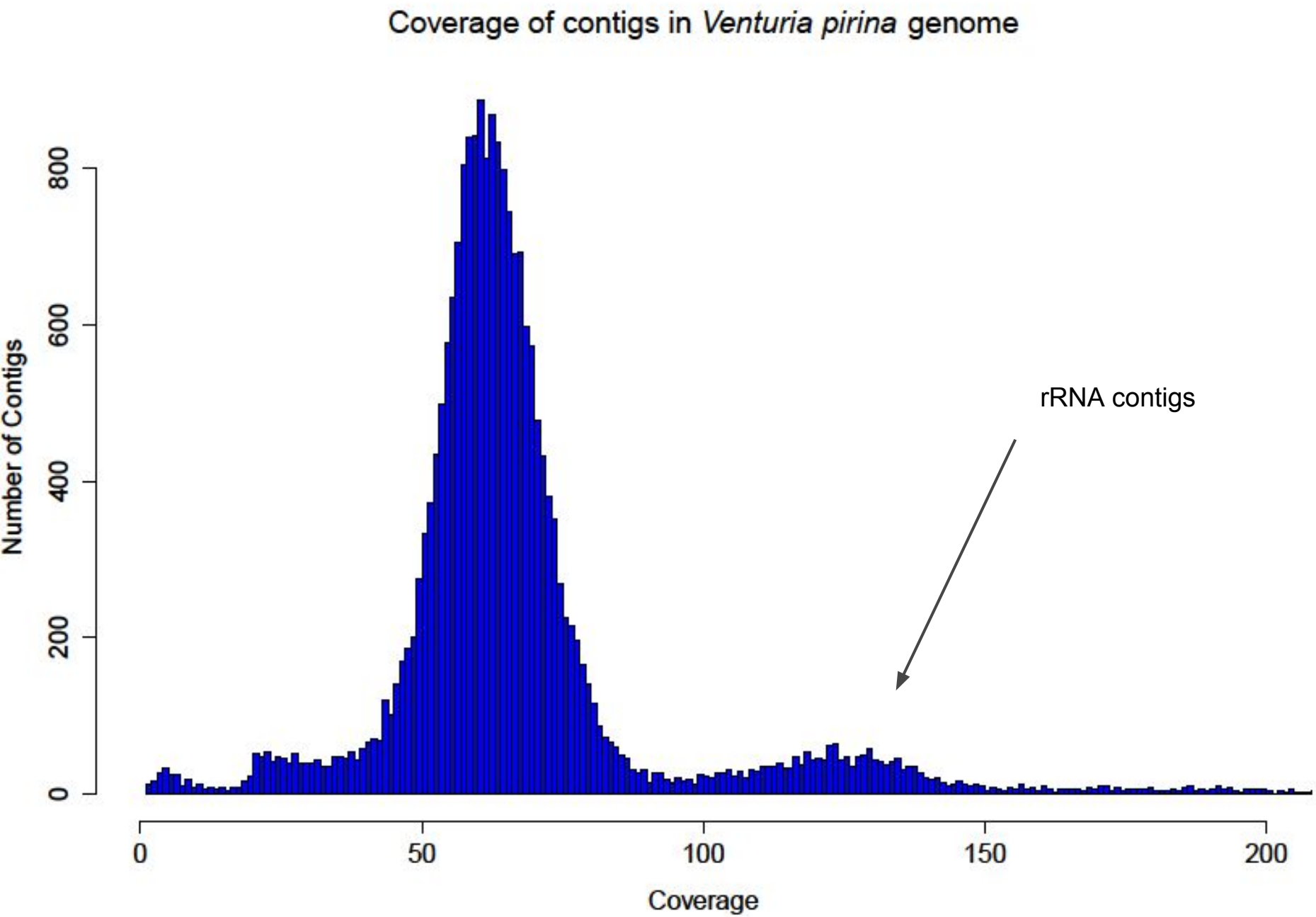
Using “long” reads

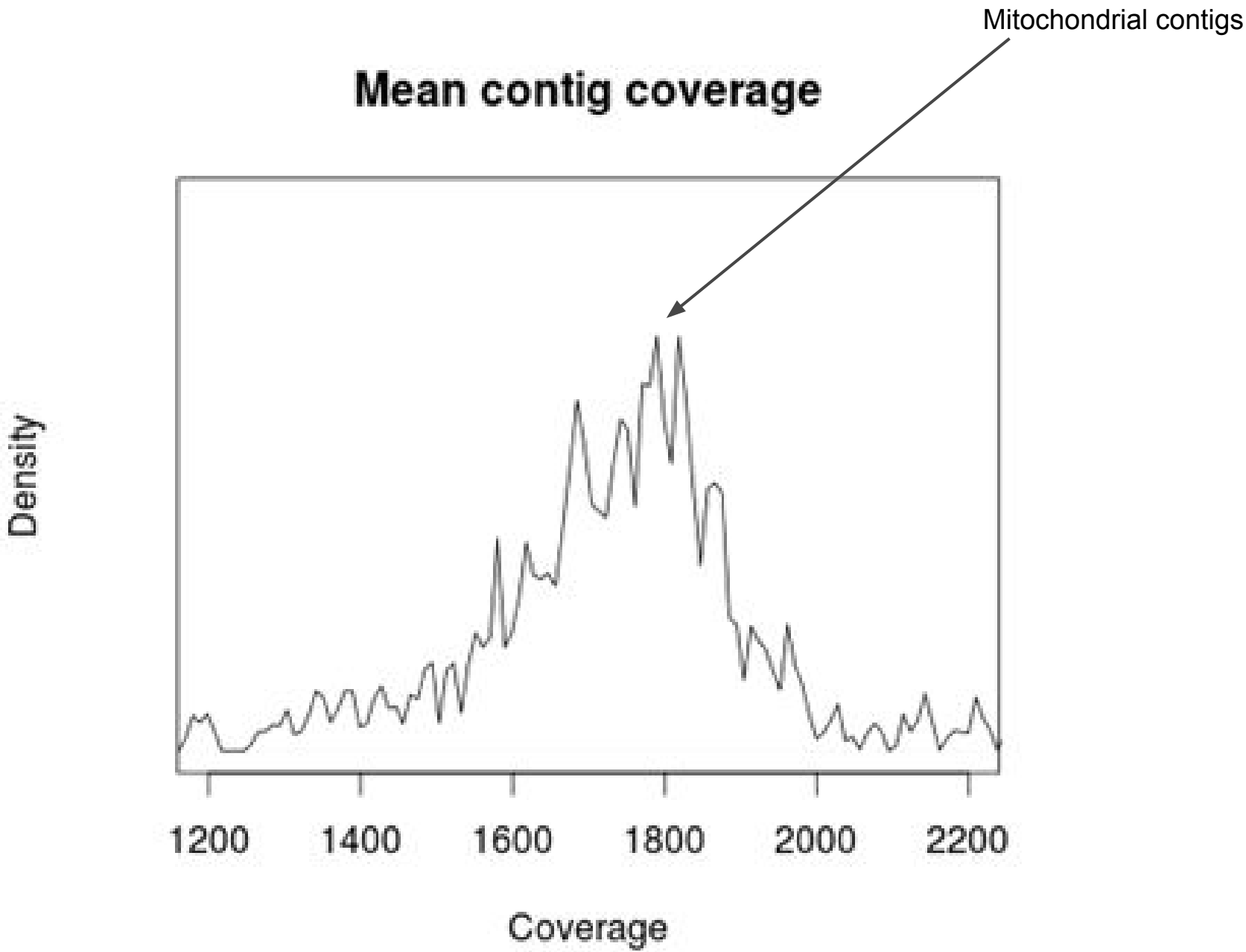
- Assemblers can use both “short” (Illumina) and “long” (PACBio or Sanger) reads, and the reads are treated differently (we will look at these options in Velvet)

- Of great importance in spanning repetitive regions
  - e.g. Pine genome

Organelles

- Organelles can be obtained from *de novo* genome sequencing projects
  - Copy number is often higher than nuclear genome and therefore coverage is often higher
  -
- Simple approach is to check if a contig has high coverage and homology to mitochondria


Coverage of contigs in *Venturia pirina* genome
Number of Contigs
0
200
400
600
800
0
50
100
150
200
Coverage
rRNA contigs

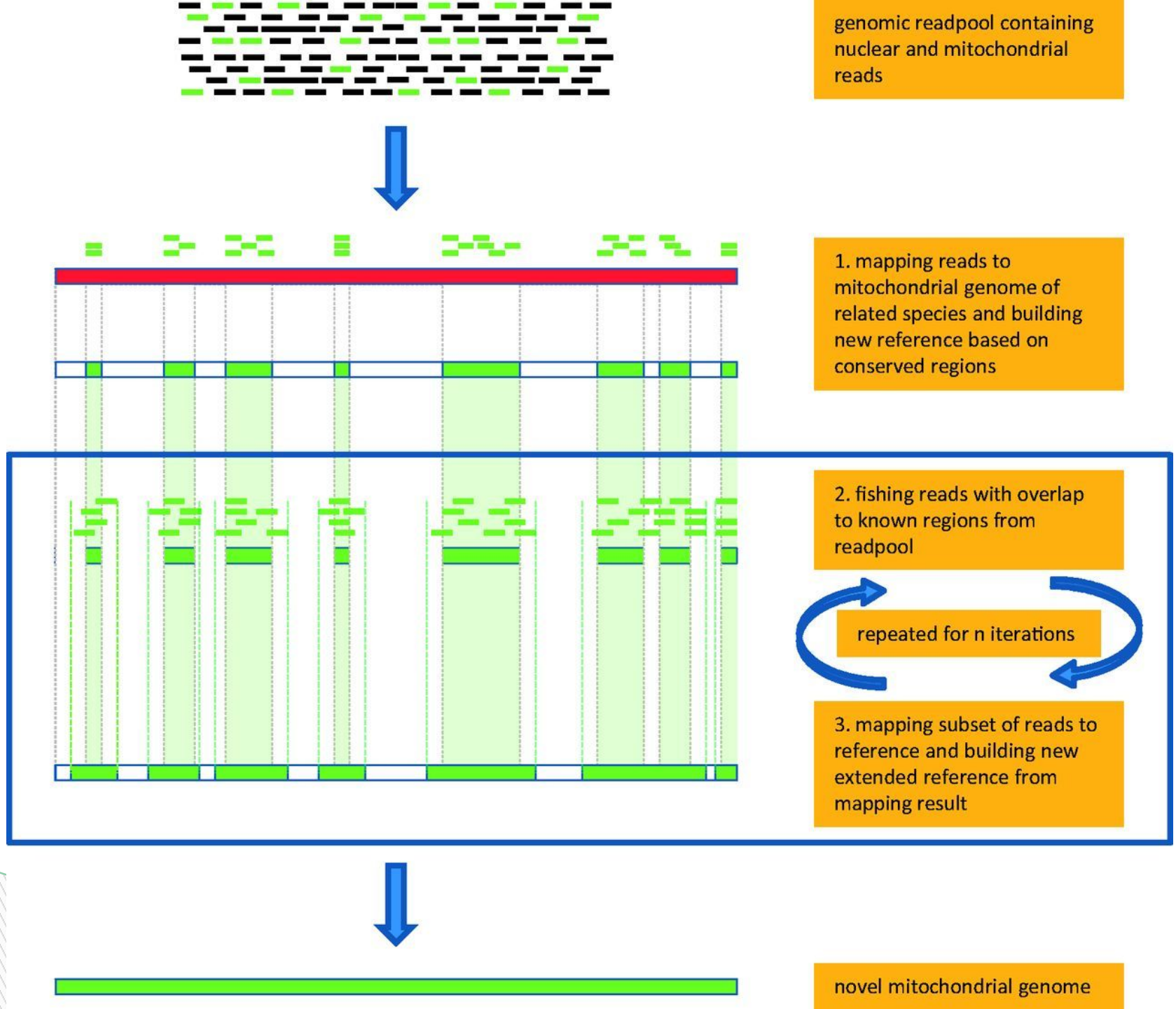

Mitochondrial contigs
Mean contig coverage
Density
1200
1400
1600
1800
2000
2200
Coverage

Organelles

- More sophisticated approach to getting organelle genomes: bait-and-capture from raw DNA reads

Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach.
Christoph Hahn, **Lutz Bachmann** **and** **Bastien Chevreux**

Nucl. Acids Res. (2013)


genomic readpool containing nuclear and mitochondrial reads
1. mapping reads to mitochondrial genome of related species and building new reference based on conserved regions
2. fishing reads with overlap to known regions from readpool
repeated for n iterations
3. mapping subset of reads to reference and building new extended reference from mapping result
novel mitochondrial genome