De novo assembly of NGS data

Sandra Smit
WUR bioinformatics

PLANT RESEARCH INTERNATIONAL WAGENINGEN UR | BioSystems GENOMICS The Netherlands Plant Genomics Network | eusol | PGSC

De novo assembly. Why?

PacBio

454 | Illumina | SOLiD

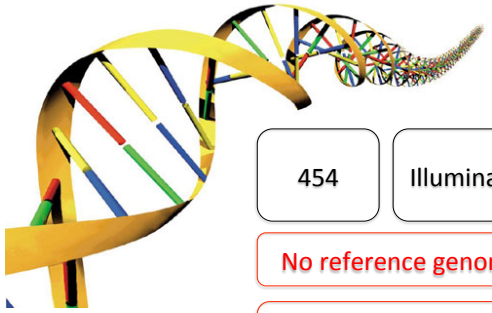
De novo assembly. Why?

PacBio

454 | Illumina | SOLiD

Read lengths are much shorter than even the smallest genome

De novo assembly. Why?



PacBio

454 | Illumina | SOLiD

No reference genome available

Not all reads map to the reference

Assemble this!

quickb
ckbrow
sove
helaz
ownfo
heq
quic

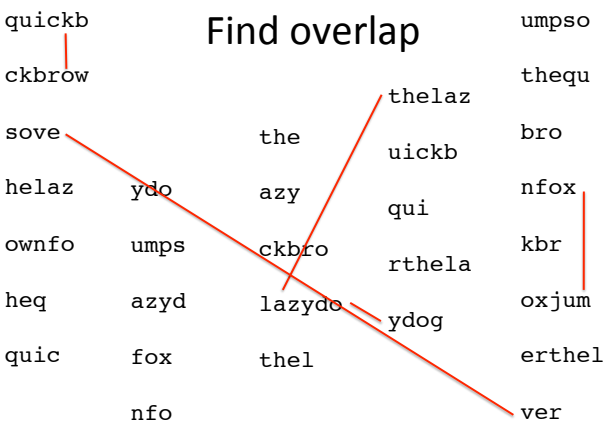
ydo
umps
azyd
fox
nfo

the
azy
ckbro
lazydo
thel

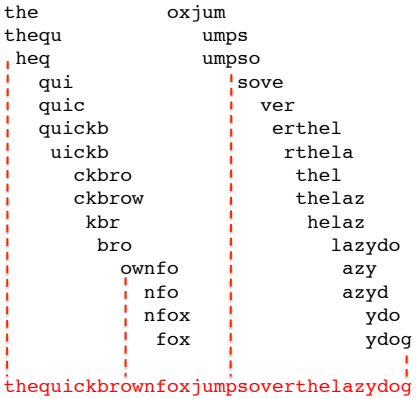
thelaz
uickb
qui
rthela
ydog

umpso
thequ
bro
nfox
kbr
oxjum
erthel
ver

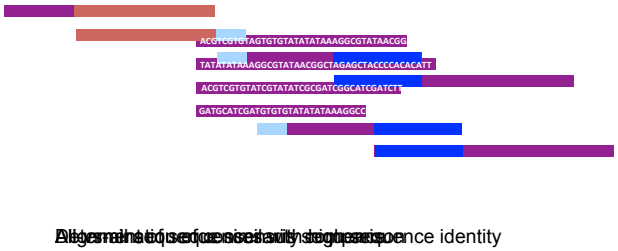
Find overlap



Layout of the reads

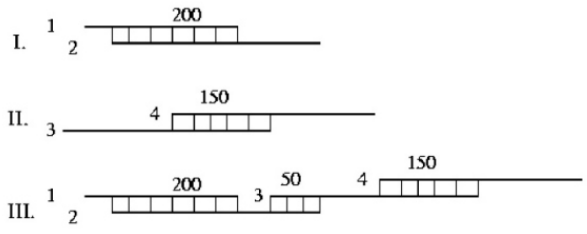


Overlap-layout-consensus approach



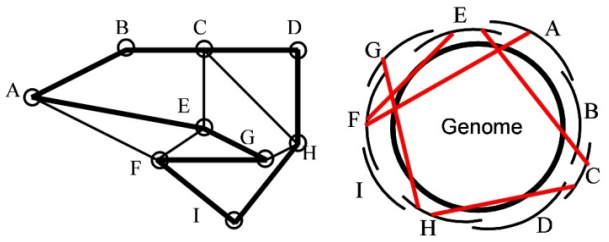
Alignment of sequences with high sequence identity

Greedy extension



Greedily join reads that are most similar to each other

Graph-based OLC



Find Hamiltonian path

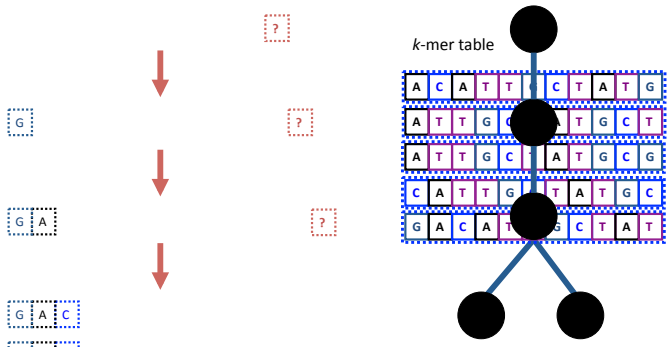
The de Bruijn graph approach

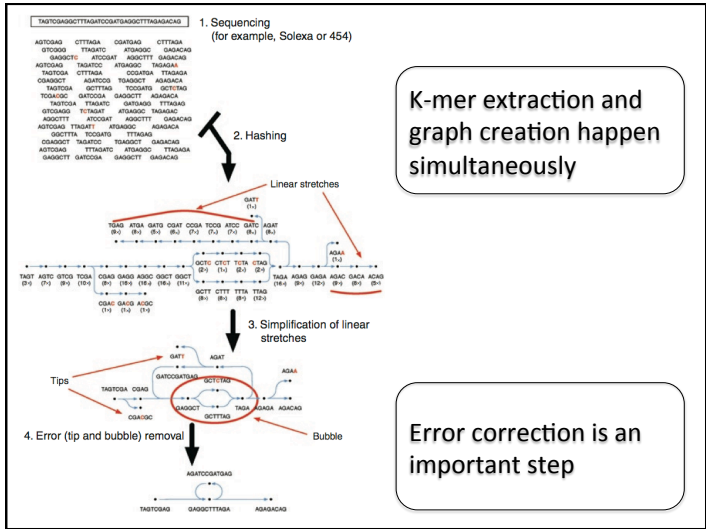
- Solving the assembly problem with De Bruijn graphs
  - Nodes are *k*-mers
  - Edges are overlaps between *k*-mers on *k*-1 positions
- A *k*-mer is a (short) substring of a read
  - A read of *n* bp consists of (*n* – *k* + 1) *k*-mers
  - Example: a read of 75 bp consists of 45 (overlapping) 31-mers

```
TGACCAGTG
TGAC
 GACC
  ACCA
   CCAG
    ....
```

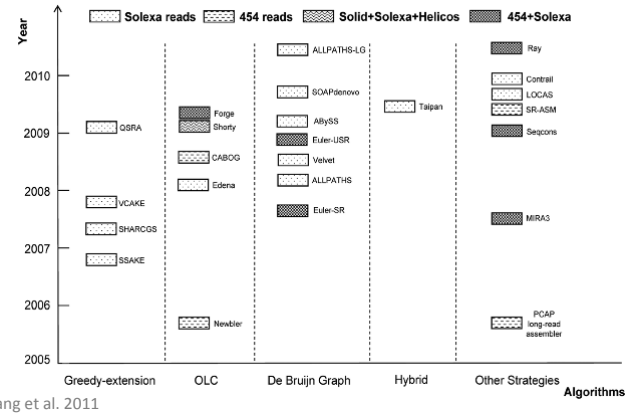
Generating *k*-mers from reads

Graph-based assembly of *k*-mers





Strategies & Assemblers



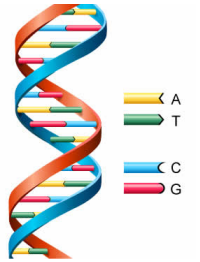
Zhang et al. 2011

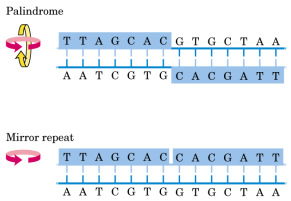
Just push the button?

Complicating factor: DNA structure

DNA is double stranded

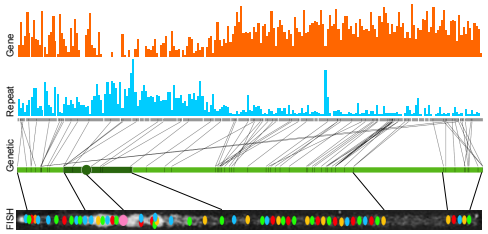
DNA contains palindromes etc.





Complicating factor: repeats

Tomato has a large, repetitive, genome
- 200-250 Mb gene-rich euchromatin
- 600-700 Mb highly repetitive heterochromatin



- Multiple overlaps in overlap-layout-consensus approach
- Multiply connected nodes in graph-based approach

Complicating factor: sequencing errors

Technologies have different properties and error profiles

Sequence-dependent coverage bias

Nonuniform error rates

	Throughput	Length	Quality	Costs
Sanger	6 Mb/day	800 nt	10^{-4}–10^{-5}	~500$/Mb
454/Roche	750 Mb/day	400 nt	10^{-3}–10^{-4}	~20$/Mb
Illumina	5,000 Mb/day	100 nt	10^{-2}–10^{-3}	~0.50$/Mb
SOLiD	5,000 Mb/day	50 nt	10^{-2}–10^{-3}	~0.50$/Mb
Helicos	5,000 Mb/day	32 nt	10^{-2}	<0.50$/Mb

Kircher & Kelso, 2010

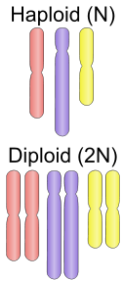
Filter your data!

Data filtering is essential to remove:
- Sequencing errors
- Homopolymer stretches
- Low-quality bases

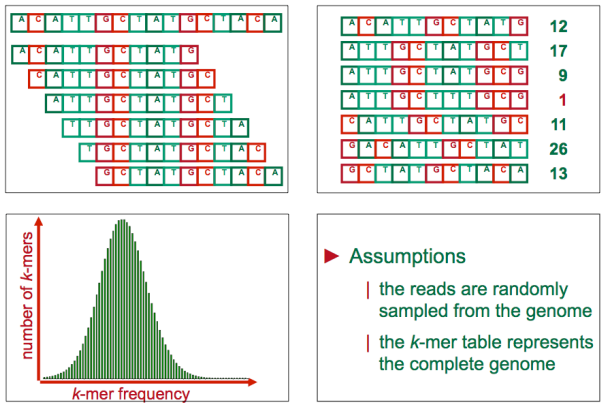
Use for example a k-mer table
- assumption: k-mer table represents the complete genome

Complicating factor: ploidy

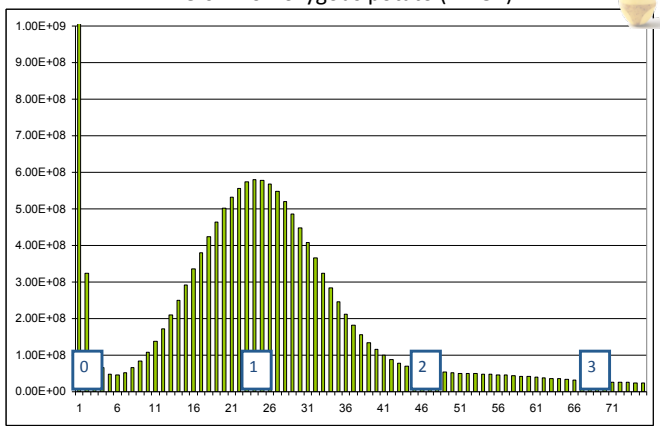
- Ploidy is the number of sets of chromosomes in a biological cell
- Diploid: two sets of chromosomes
  - Human cells
- Polyploidy means more than two sets of chromosomes per nucleus
- Tetraploid: four sets of chromosomes
  - Common in plants



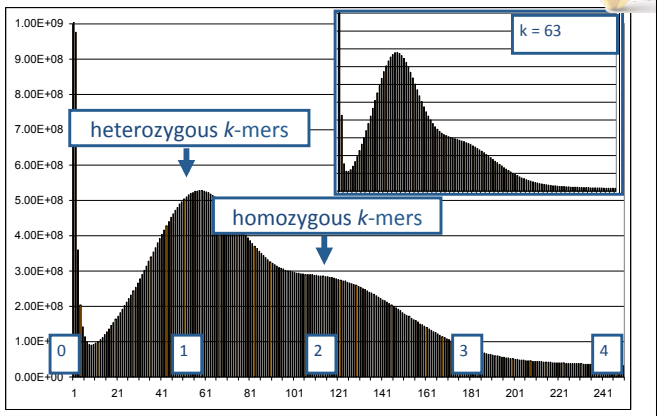
Complicating factor: heterozygosity



k-mers – homozygous potato (*k* = 31)

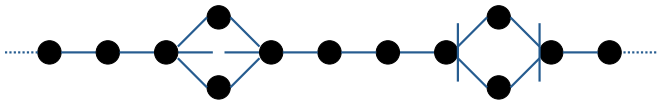


k-mers – heterozygous potato (*k* = 31)



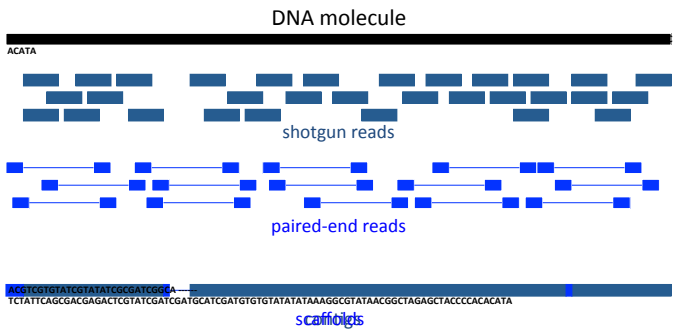
Complicating factor: heterozygosity

- Heterozygous *k*-mers create "bubbles" in the graph
  - Every SNP and in/del creates a bubble
- Solutions
  - Pinching bubbles: loss of diversity
  - Breaking graph: loss of contiguity



Assembly is not so easy

Assembly result



Assembly quality

- Number and length of contigs/scaffolds
  - N50
- Number of gaps/Ns
- GC content
- Coverage and completeness
- Structural consistency
  - Paired-end data
  - Physical and genetic maps
- Error rate
  - Known sequences (BAC clones, ESTs, etc.)

Contigs (length)

250
150
120
100
90
80
80
70
40
20

Basic stats: exercise

- Calculate
  - N50 contig size and index
  - N90 contig size and index

Contigs (length)

250
150
120 — 50%
100
90
80
80
70 — 90%
40
20

Basic stats: exercise

- Total assembly size = 1000
- Sort from large to small

- N50 lookup value = 500
- N50 contig size = 120
- N50 contig index = 3

- N90 contig size = 70
- N90 contig index = 8

Genome finishing

From initial draft to complete genome

- Closing gaps
  - IMAGE pipeline, Tsai et al. 2010
  - GapFiller, Boetzer and Pirovano 2012
- Establish order and orientation of contigs/scaffolds
  - E.g. FISH experiments
- Base error correction
  - *k*-mer correction
- Contamination removal

Chain et al., Science 2009
Genome Project Standards in a New Era of Sequencing

Gap closure

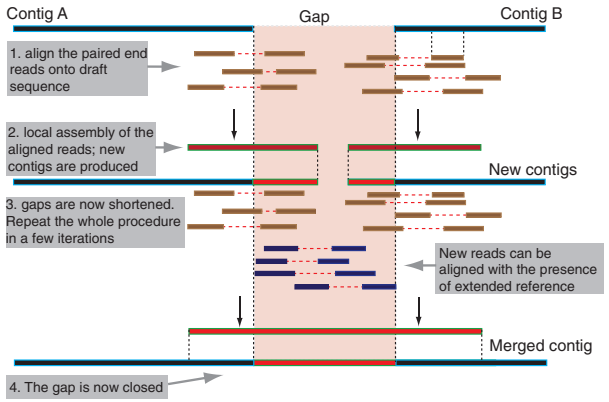


IMAGE pipeline, Tsai et al. 2010

Sequencing strategy

Sequencing strategy: which sequencing data is necessary for my project? How can we balance the benefits and the costs?

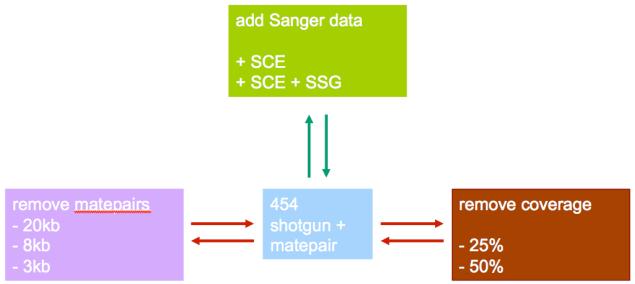
Technologies have different properties and error profiles

What is the contribution of different input data to an assembly?

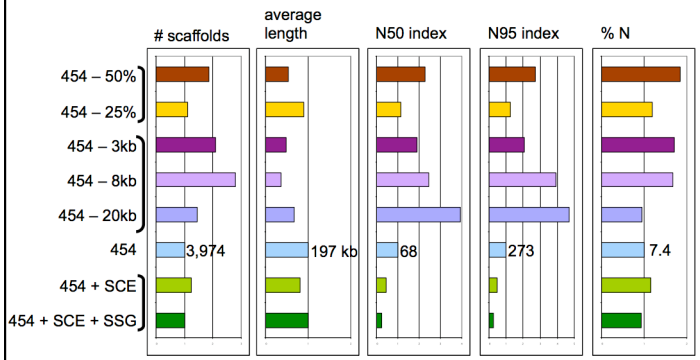
	Throughput	Length	Quality	Costs
Sanger	6 Mb/day	800 nt	10^{-4}–10^{-5}	~500$/Mb
454/Roche	750 Mb/day	400 nt	10^{-3}–10^{-4}	~20$/Mb
Illumina	5,000 Mb/day	100 nt	10^{-2}–10^{-3}	~0.50$/Mb
SOLiD	5,000 Mb/day	50 nt	10^{-2}–10^{-3}	~0.50$/Mb
Helicos	5,000 Mb/day	32 nt	10^{-2}	<0.50$/Mb

Kircher & Kelso, 2010

Sequencing strategy



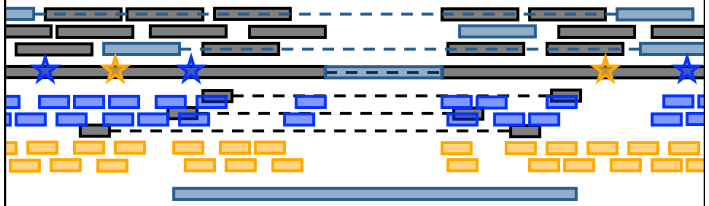
Sequencing strategy



Recommendations

- Longer reads (Sanger, 454) improve assembly quality
- Deep coverage by reads with lengths longer than common repeats
- Mixture of short and long insert sizes for paired-end data

Example



- *De novo* assembly: overlap-layout-extend
  - 454 & Sanger shotgun sequences
  - 454 & Sanger matepair sequences
- Base error correction with SOLiD reads
- Base error correction with Illumina reads
- Gap filling with Sanger BAC sequences

Summary: assembly workflow

- NGS sequencing/data production
- Data cleaning/filtering/error correction
- *De novo* assembly into contigs and (super)scaffolds
- Assembly finishing/gap filling
- Assembly validation

Literature

Assembly of large genomes using second-generation sequencing
Schatz et al.
Genome Res. 2010

Assembly algorithms for next-generation sequencing data
Miller et al.
Genomics 2010

High-throughput DNA sequencing -- concepts and limitations
Kircher and Kelso
Bioessays 2010

Acknowledgements

- Jack Leunissen
- Roeland van Ham
- Erwin Datema
- Jan van Haarst

WAGENINGEN UNIVERSITY
WAGENINGEN UR

The International Tomato Genome Sequencing Consortium

Potato Genome Sequencing Consortium