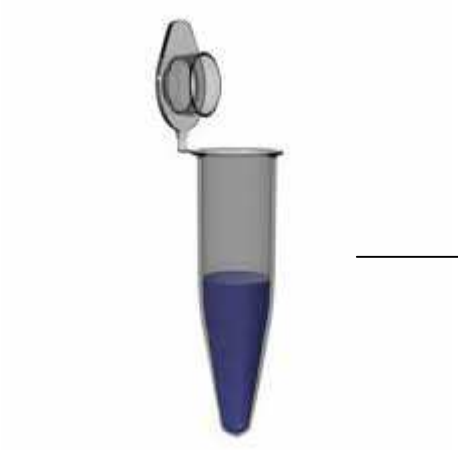
Assembling NGS data

Dr Torsten Seemann

IMB Winter School - Brisbane – Tue 3 July @ 09:45am

Ideal world

I would not need to give this talk!



Human DNA

Non-existent USB3 device

AGTCTAGGATTCGCTA
TAGATTCAGGCTCTGA
TATATTTCGCGGGATT
AGCTAGATCGCTATGC
TATGATCTAGATCTCG
AGATTCGTATAAGTCT
AGGATTCGCTATAGAT
TCAGGCTCTGATATAT
TTCGCGGGATTAGCTA

46 complete haplotype chromosome sequences

Real world

- Can’t sequence full-length native DNA
  - no instrument exists (yet)

- But we can sequence short fragments
  - 100 at a time (Sanger)
  - 100,000 at a time (Roche 454)
  - 1,000,000 at a time (Ion Torrent)
  - 100,000,000 at a time (HiSeq 2000)

De novo assembly

- *De novo* assembly is the process of reconstructing the original DNA sequences using only the fragment read sequences

- Instinctively
  - like a jigsaw puzzle
  - involves finding overlaps between reads
  - sequencing errors will confuse matters

Shakespearomics

- ***Reads***

```
ds, Romans, count
ns, countrymen, le
Friends, Rom
send me your ears;
crymen, lend me
```

- ***Overlaps***

```
Friends, Rom
      ds, Romans, count
            ns, countrymen, le
                  crymen, lend me
                        send me your ears;
```

- ***Majority rule***

```
Friends, Romans, countrymen, lend me your ears;
```

The awful truth

"Genome assembly is impossible."

He wears glasses so he must be smart

A/Prof. Mihai Pop

World leader in *de novo* assembly research.

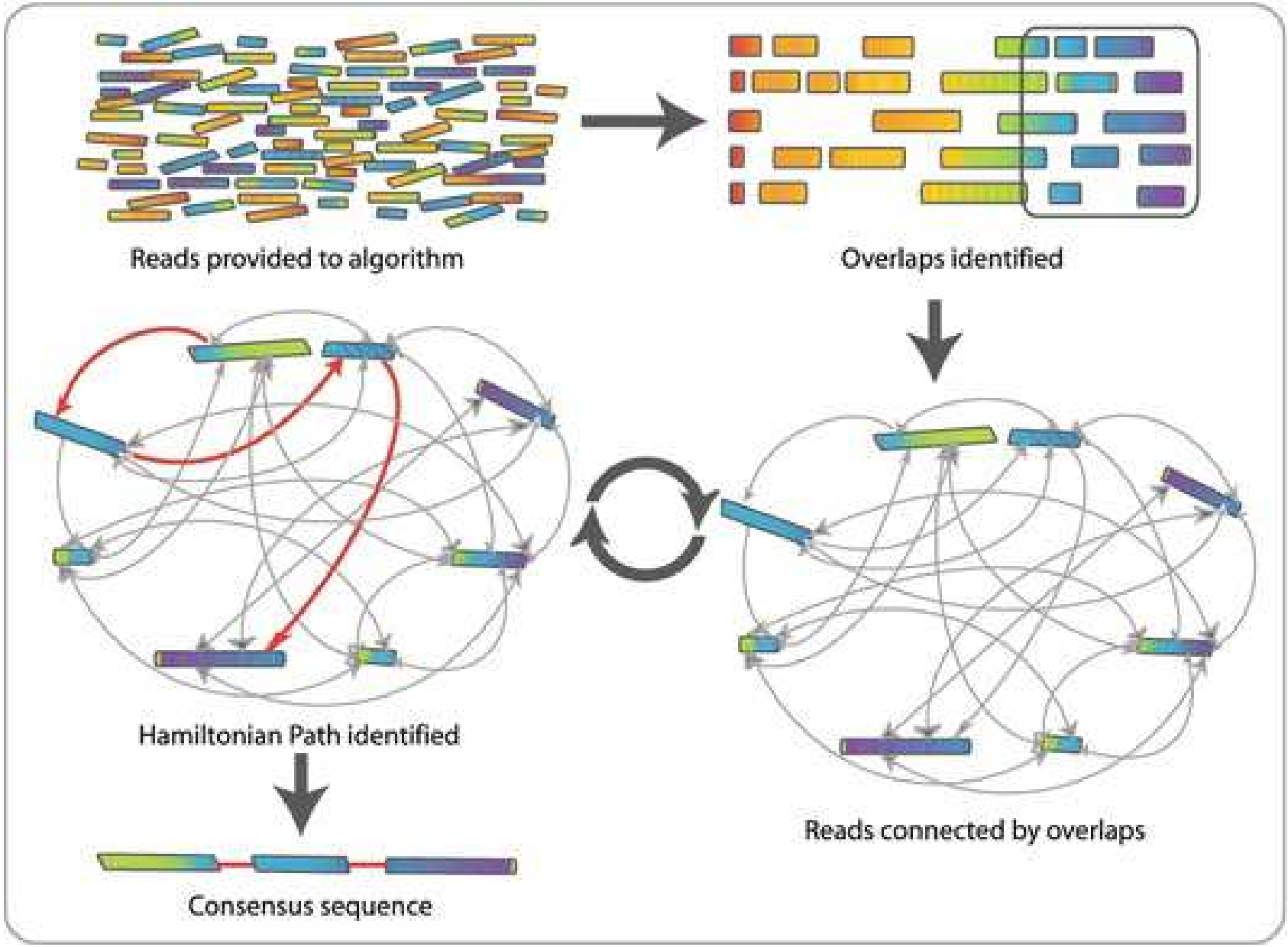
Approaches

- greedy assembly
- overlap :: layout :: consensus
- de Bruijn graphs
- string graphs
- seed and extend

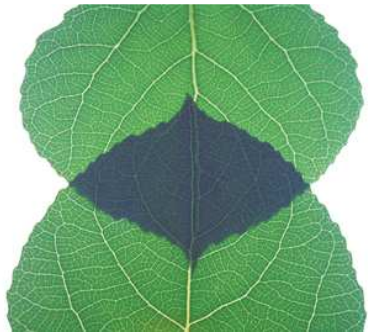
… all essentially doing the same thing, but taking different short cuts.

Assembly recipe

- Find all overlaps between reads
  - hmm, sounds like a lot of work…
- Build a graph
  - a picture of read connections
- Simplify the graph
  - sequencing errors will mess it up a lot
- Traverse the graph
  - trace a sensible path to produce a consensus


Reads provided to algorithm
Overlaps identified
Hamiltonian Path identified
Reads connected by overlaps
Consensus sequence

Find read overlaps



- If we have N reads of length L
  - we have to do ½N(N-1) ~ O(N²) comparisons
  - each comparison is an ~ O(L²) alignment
  - use special tricks/heuristics to reduce these!

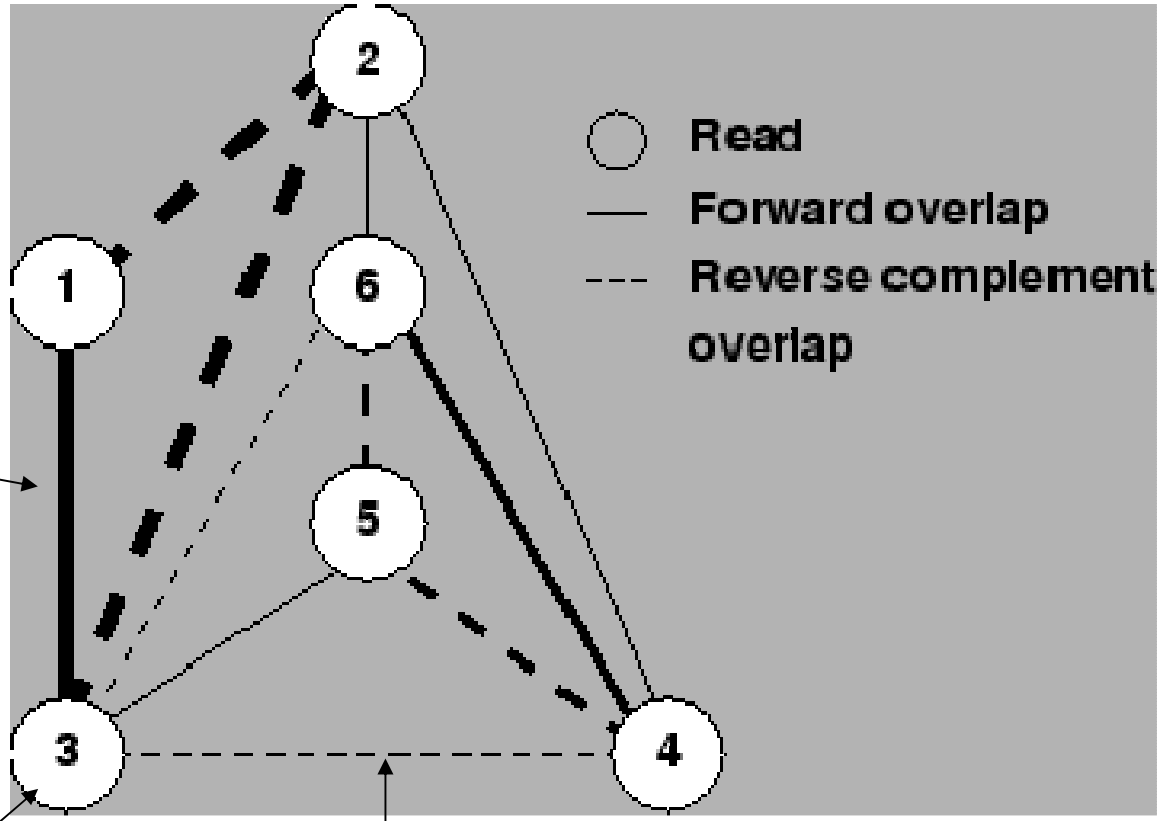
- What counts as “overlapping” ?
  - minimum overlap length eg. 20bp
  - minimum %identity across overlap eg. 95%
  - choice depends on L and expected error rate

N=6 means 15 overlap “scores”

Read#	1	2	3	4	5	6
1	-	-	-	-	-	-
2	80	-	-	-	-	-
3	95	85	-	-	-	-
4	0	30	20	-	-	-
5	0	0	25	70	-	-
6	0	35	25	60	50	-

Graph construction

Thicker lines mean stronger evidence for overlap



Node/Vertex

Edge/Arc

A more realistic graph

What ruins the graph?

- Read errors
  - introduce false edges and nodes
- Non-haploid organisms
  - heterozygosity causes lots of detours
- Repeats
  - if longer than read length
  - causes nodes to be shared, locality confusion

Graph simplification

- Squash small bubbles
  - collapse small errors (or minor heterozygosity)
- Remove spurs
  - short “dead end” hairs on the graph
- Join unambiguously connected nodes
  - reliable stretches of unique DNA
- Remove transitive edges
  - Collapse paths saying the same thing differently

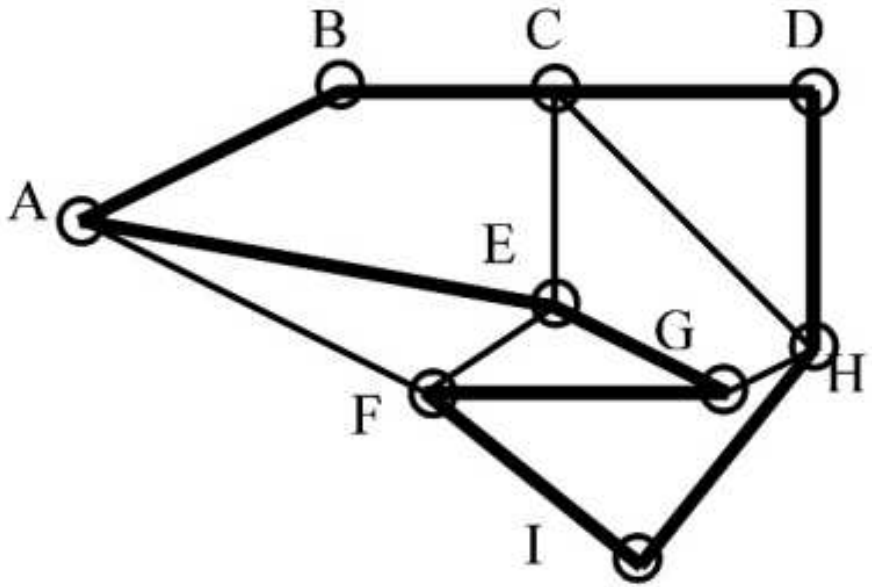
Graph traversal

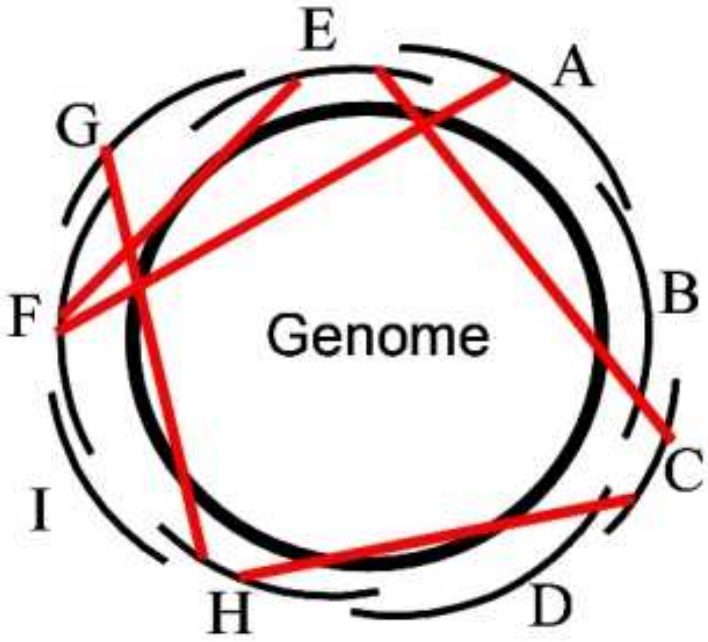
- For each unconnected graph
  - at least one per replicon in original sample

- Find a path which visits each node once
  - the Hamiltonian path (or cycle)
  - provably NP-hard (this is bad)
  - unlikely to be single path due to repeat nodes
  - solution will be a set of paths which terminate at decision points

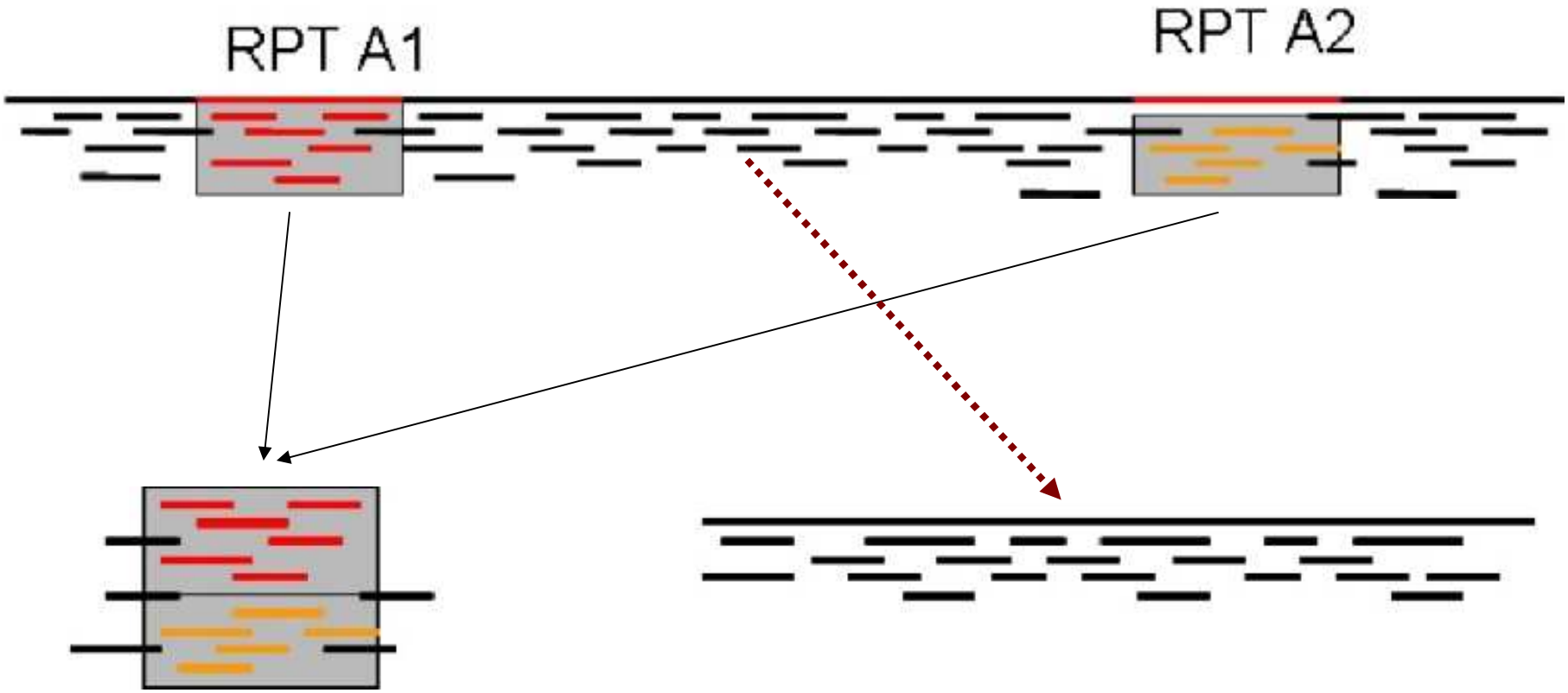
- Form a consensus sequence from path
  - use all the overlap alignments
  - each of these is a CONTIG

Graph traversal



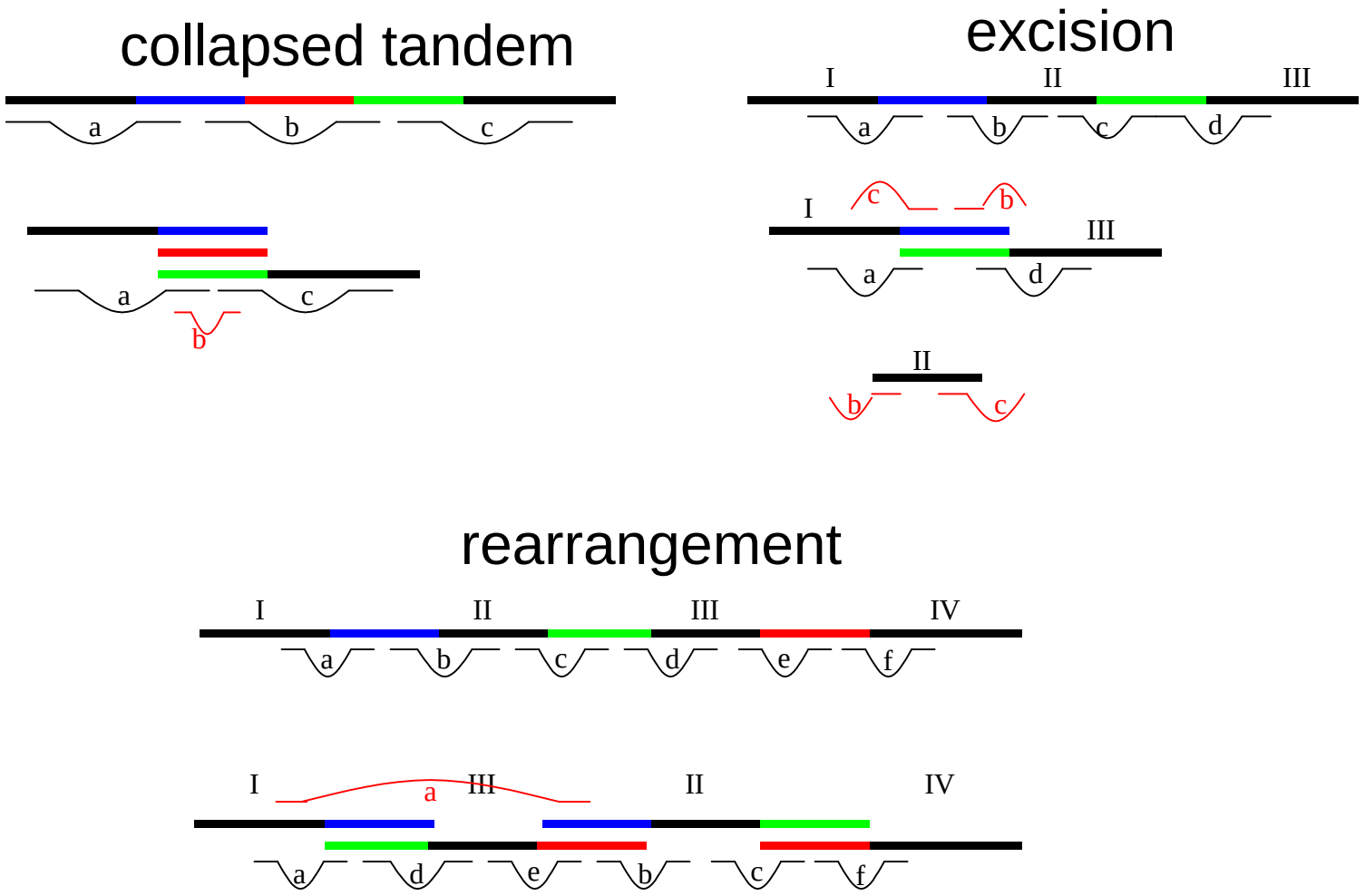


What happens with repeats?



The repeated element is collapsed into a single contig

Mis-assembled repeats



The law of repeats

- It is impossible to resolve repeats of length S unless you have reads longer than S.

- It is impossible to resolve repeats of length S unless you have reads longer than S.

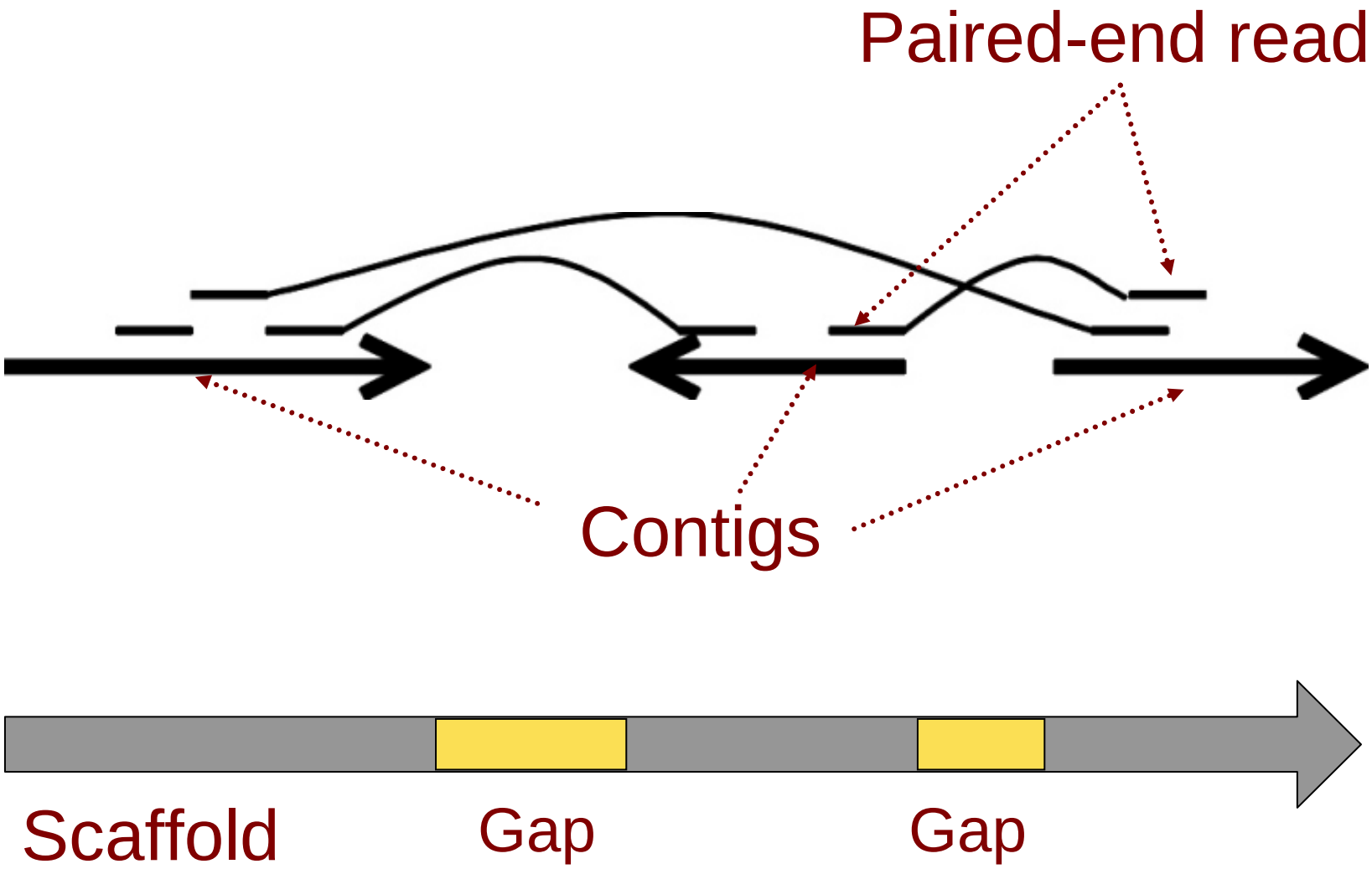
Types of reads

- Example fragment
  - **atcgtatgatcttgagattctcttcccttatagctgctata**
- “Single-end” read
  - **atcgtatg**atcttgagattctcttcccttatagctgctata
  - Sequence *one* end of the fragment
- “Paired-end” read
  - **atcgtatg**atcttgagattctcttcccttatag**ctgctata**
  - Sequence *both* ends of *same* fragment
  - we can exploit this information!

Scaffolding

- Paired-end reads
  - known sequences at either end
  - roughly known distance between ends
  - unknown sequence between ends
- Most ends will occur in same contig
  - if our contigs are longer than pair distance
- Some ends will be in different contigs
  - evidence that these contigs are linked!

Contigs to Scaffolds

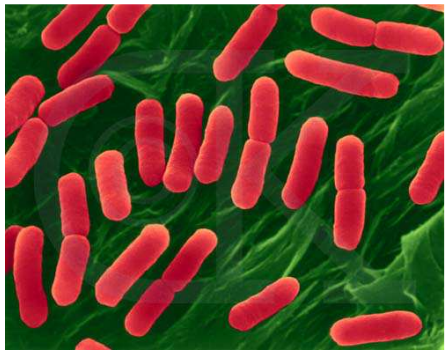



PHASE THREE: INTERPRETATION
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

What can we assemble?

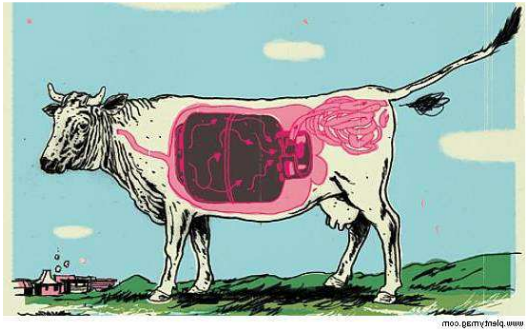
- Genomes
  - A single organism eg. its chromosomal DNA
- Meta-genomes
  - gDNA from mixtures of organisms
- Transcriptomes
  - A single organism’s RNA inc. mRNA, ncRNA
- Meta-transcriptomes
  - RNA from a mixture of organisms

Genomes



- Expect uniformity
  - Each part of genome represented by roughly equal number of reads
- Average depth of coverage
  - Genome: 4 Mbp
  - Yield: 4 million x 50 bp reads = 200 Mbp
  - Coverage: 200 ÷ 4 = 50x (reads per bp)

Meta-genomes



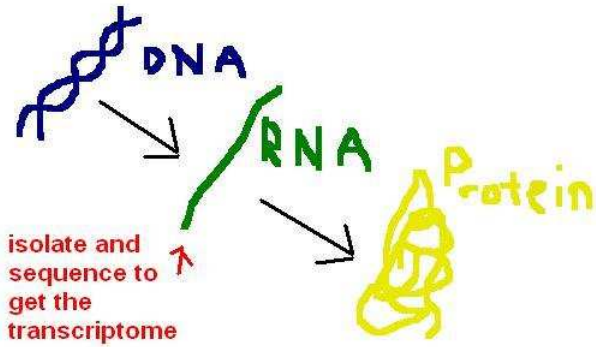
- Expect proportionality & uniformity
  - Each genome represented by proportion of reads similar to their proportion in mixture

- Example
  - Mix of 3 species: ¼ Staph, ¼ Clost, ½ Ecoli
  - Say we get 4M reads
  - Then we expect about:
    1M from Staph, 1M from Clost, 2M from Ecoli

Meta-genome issues

- Closely related species
  - will have very similar reads
  - lots of shared nodes in the graph
- Conserved sequence
  - bits of DNA common to lots of organisms
  - “hub” nodes in the graph
- Untangling is difficult
  - need longer reads

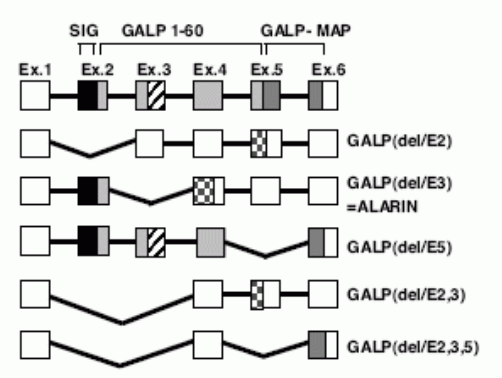
Transcriptomes



- RNA-Seq
  - first convert it into DNA (cDNA)
  - represents a snapshot of RNA activity

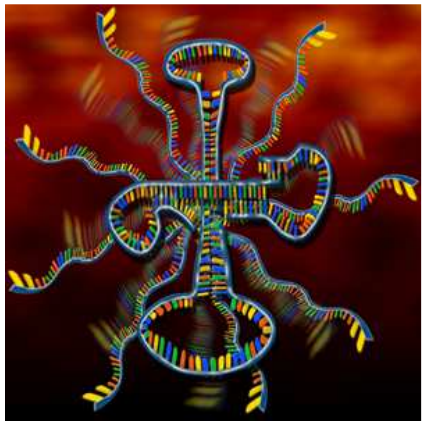
- Expect proportionality
  - the expression level of a gene is proportional to the number of reads from that gene's cDNA

Transcriptome issues



- Huge dynamic range
  - some gets lots of reads, some get none
- Splice variation
  - very similar, subtly different transcripts
  - lots of shared nodes in graph

Meta-transcriptomes



- RNA-Seq
  - on multiple transcriptomes at once
- Expect proportional proportionality
  - proportion of that organism in mixture
  - proportions due to expression levels
- Meta x transcriptome issues combined!

Assessing assemblies

- Genome assembly
  - Total length similar to genome size
  - Fewer, larger contigs
  - Correctness of contigs

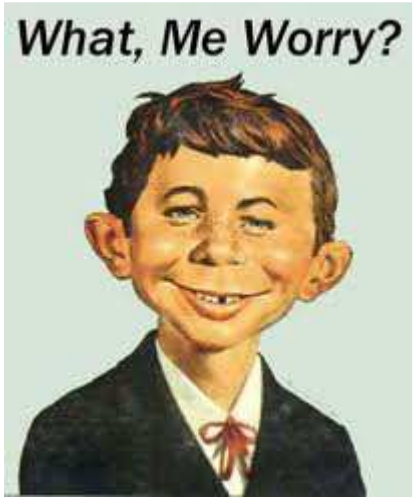
- Metrics
  - Maximum contig length
  - N50 (next slide)

The “N50”

- “The length of that contig from which 50% of the bases are in it and shorter contigs”

- Imagine we got 7 contigs with lengths:
  - 1,1,3,5,8,12,20
- Total
  - 1+1+3+5+8+12+20 = 50
- N50 is the “halfway sum” = 25
  - 1+1+3+5+8+12 = 30 (≥ 25) so **N50 is 12**

N50 concerns



- Optimizing for N50
  - encourages mis-assemblies!
- An aggressive assembler may over-join:
  - 1,1,3,5,**8,12**,20 (previous)
  - 1,1,3,5,**20**,20 (now)
  - 1+1+3+5+20+20 = 50 (unchanged)
- N50 is the “halfway sum” (still 25)
  - 1+1+3+5+20= 30 (≥ 25) so **N50 is 20**

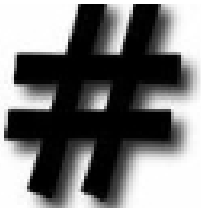
Assembly tools

- Genome
  - Velvet, Abyss, Mira, Newbler, SGA, AllPaths, Ray, Euler, SOAPdenovo, Edena, Arachne
- Meta-genome
  - MetaVelvet, SGA, custom scripts + above
- Transcriptome
  - Trans-Abyss, Oases, Trinity
- Meta-Transcriptome
  - custom scripts + above

Example

- Culture your bacterium
- Extract your genomic DNA
- Send it to AGRF for Illumina sequencing
  - 100bp paired end
- Get back two files:
  - `MRSA_R1.fastq.gz`
  - `MRSA_R2.fastq.gz`
- Now what?

Velvet: hash reads



```
velveth
Dir
31
-fmtAuto
-separate
MRSA_R1.fastq.gz
MRSA_R2.fastq.gz
```

New options

No interleaving required

Velvet: assembly


```
velvetg
Dir
-exp_cov auto
-cov_cutoff auto
```

“Signal” level

“Noise” level

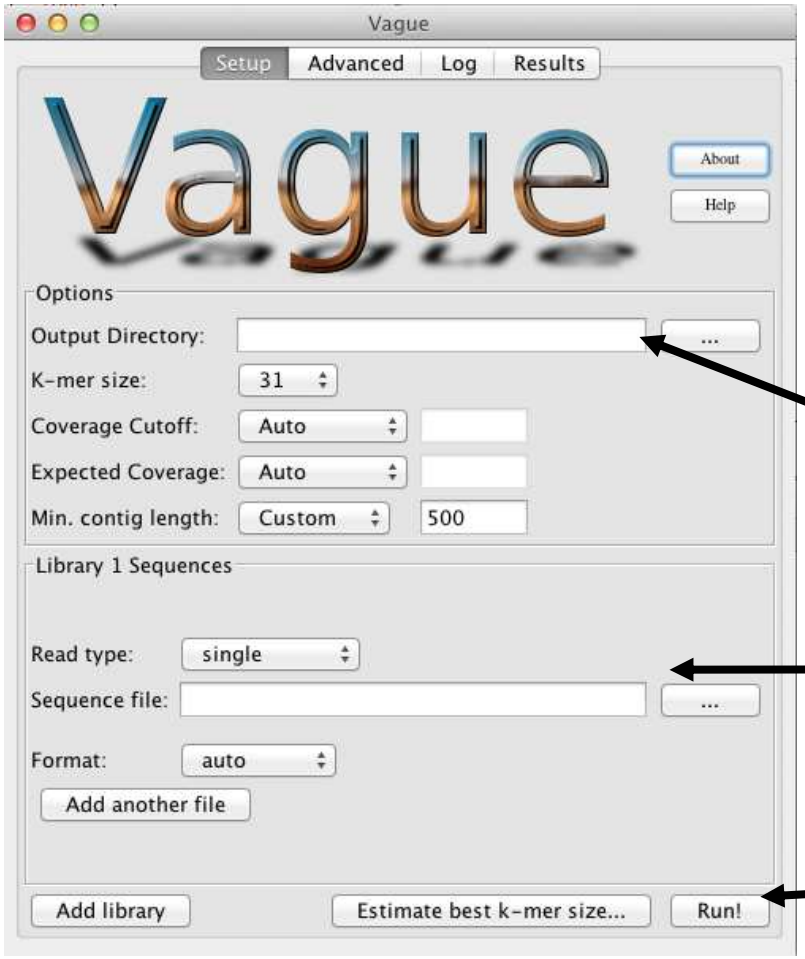
Velvet: examine results

```
less Dir/contigs.fa
```

```
>NODE_1_length_43211_cov_27.36569
AGTCGATGCTTAGAGAGTATGACCTTCTATACAAAA
ATCTTATATTAGCGCTAGTCTGATAGCTCCCTAGAT
CTGATCTGATATGATCTTAGAGTATCGGCTATTGCT
AGTCTCGCGTATAATAAATAATATATTTTTCTAATG
ATCTTATATTAGCGCTAGTCTGATAGCTCCCTAGAT
CTGATCTGATATGATCTTAGAGTATCGGCTATTGCT
AGTCTCGCGTATAATAAATAATATATTTAGTAGTCT …
```

Velvet: GUI

Velvet Assembler Graphical User Environment



Where to save

Add your reads

Click run


That's all Folks!

Contact

- Email
  - torsten.seemann@monash.edu
- Web
  - http://vicbioinformatics.com/
  - http://vlsci.org.au/
- Blog
  - http://TheGenomeFactory.blogspot.com