ALLPATHS

Assemblying Large Genomes with Short Illumina Reads

Sante Gnerre

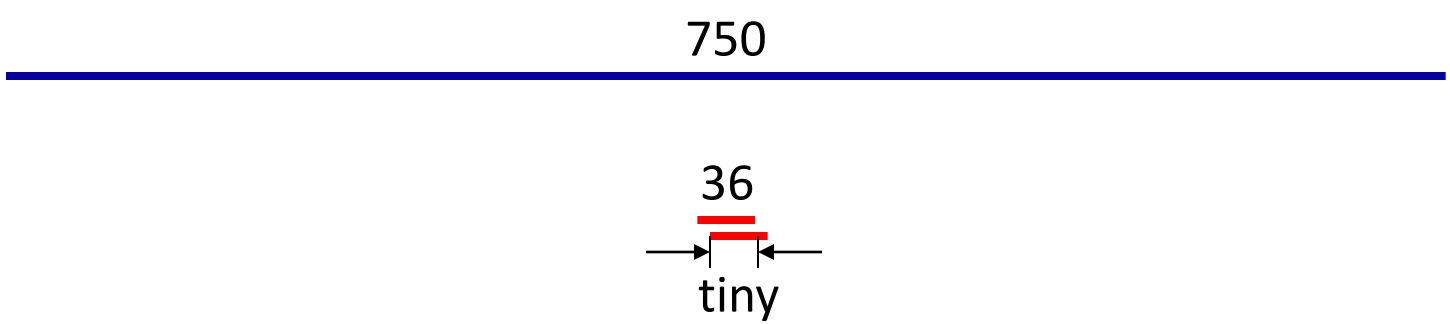
BROAD INSTITUTE

Overview

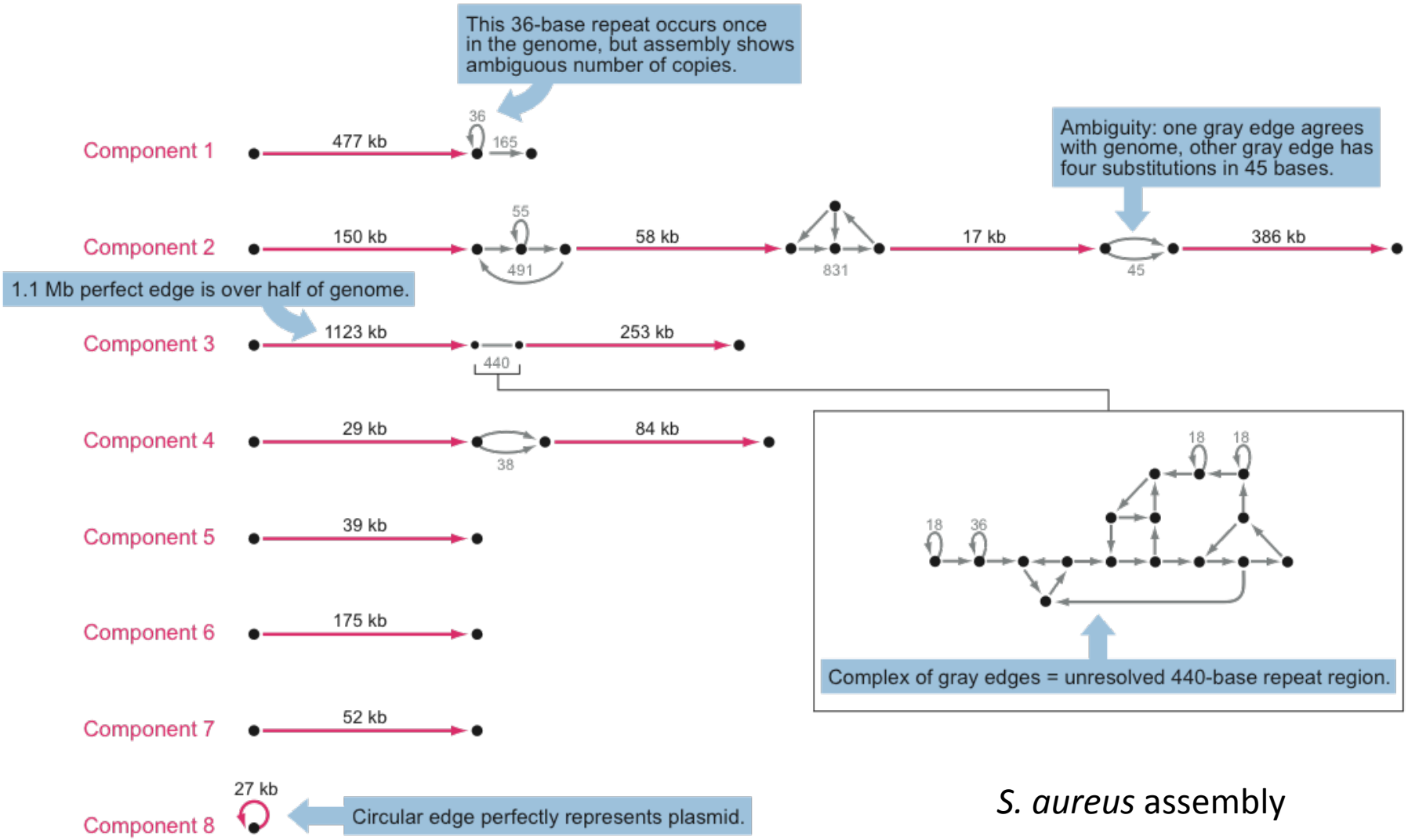
- Short read assemblies
- ALLPATHS v2: what you can do now
- ALLPATHS v3: tackling larger genomes

De-novo short read assembly

- Shorter reads are much cheaper…
- … but can traditional tools assemble them?
- Not really:
  - Too many missing / false read-read overlaps
  - Genomes are “more repetitive” at tiny scales



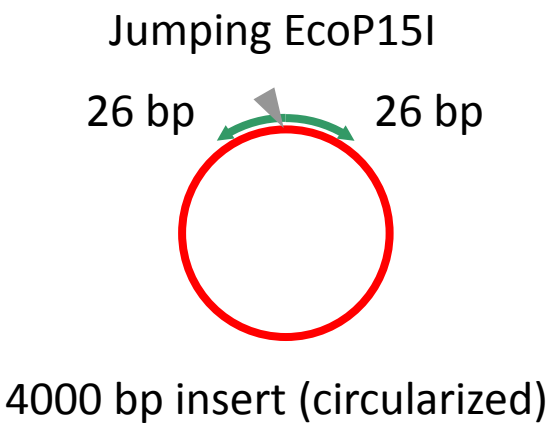
ALLPATHS assemblies are graphs

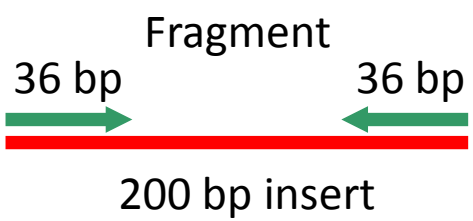


S. aureus assembly

Small genomes

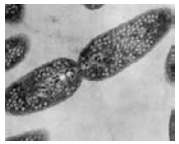
- ALLPATHS can now assemble small genomes
- Tested on various organisms:
  - Up to 40 Mb
  - From 33% to 69% GC content
  - Very high coverage with short Illumina reads





Small genomes

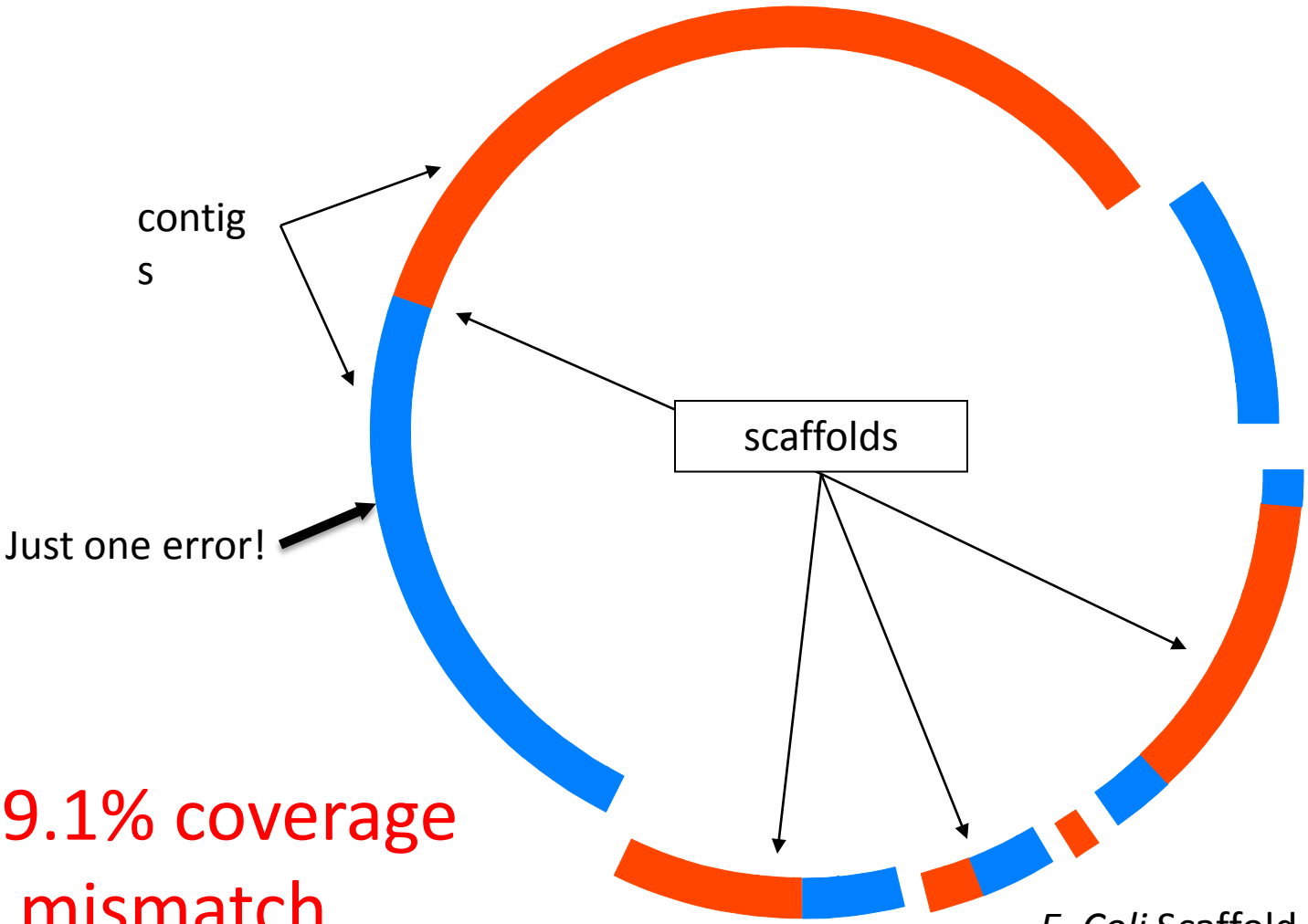




	S. aureus 2.9 Mb 33% GC	*E. coli* 4.6 Mb 51% GC	*R. sphaeroides* 4.6 Mb 69% GC	*S. pombe* 12.6 Mb 36% GC	*N. crassa* 40 Mb 49% GC
Sequence coverage	89x	139x	370x	148x	123x
Genome covered	99.1%	99.3%	98.5%	95.9%	89.5%
Contig N50	477 kb	337 kb	156 kb	51 kb	19 kb
Scaffold N50	611 kb	2680 kb	858 kb	222 kb	58 kb
Scaffold accuracy	100%	100%	100%	99.8%	99.8%
Base accuracy	~Q59	~Q67*	~Q60	~Q42	~Q39

ALLPATHS V2 source code now available:
http://www.broadinstitute.org/science/programs/genome-biology/crd

Small genomes – E.coli



99.1% coverage
1 mismatch

E. Coli Scaffold N50 = 2680 kb

ALLPATHS approach for large genomes

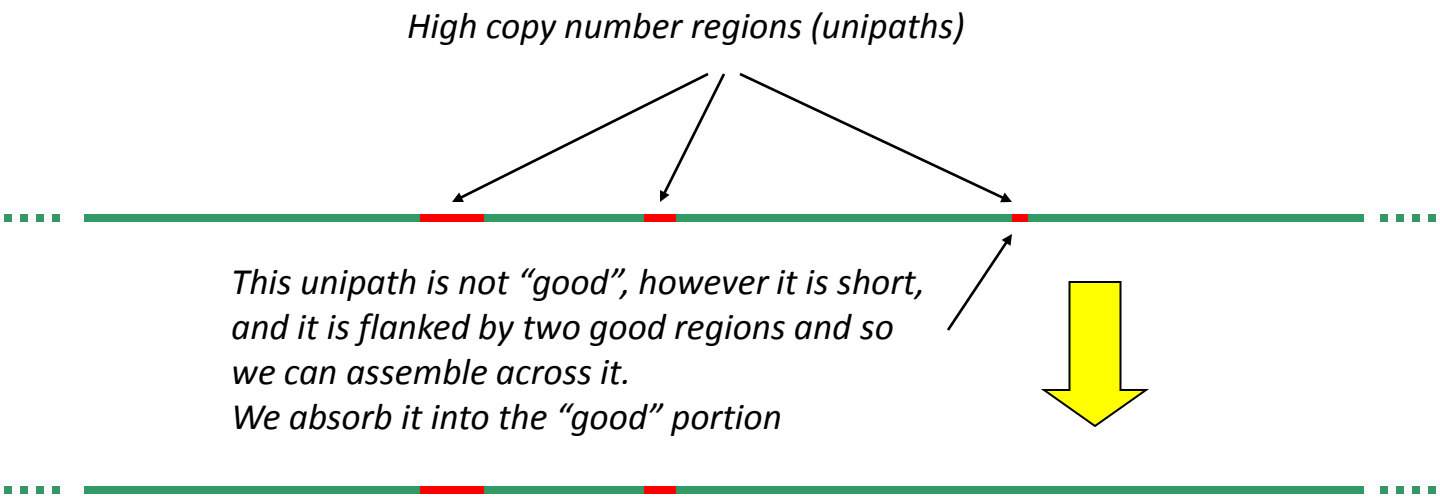
- A much harder problem:
  - Large genomes tend to be more repetitive
  - Polymorphism
  - Large data sets
- Need longer reads to allow higher k-mer size!
  - Genomes appear less repetitive with larger k
  - How large is large enough?

Larger k is better

- Read size limits the k-mer size, k, that can be used in the assembly process
- Longer reads allow larger k-mer size, which results in a better assembly
- How to assess impact of larger k-mer size
  - Start with reference genome
  - Create ideal unipaths (depends on k-mer size)
  - Tag “good” regions by using simple criteria
  - Look at statistics of “good” regions

Define "good" regions

- A "good" (easy to assemble) region:
  - has copy number 1
  - is long enough to provide a good starting point for the assembly process



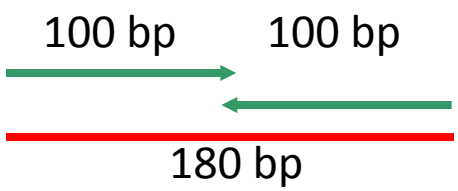
How much would be missing?

	k = 40	k = 64	**k = 100**
E. coli	0.33 %	0.33 %	**0.13 %**
N. crassa	5.95 %	1.30 %	**0.76 %**
G. aculeatus	13.46 %	9.18 %	**6.76 %**
H. sapiens	34.22 %	12.65 %	**4.19 %**

Table shows percentage of genome not contained in ‘good’ regions

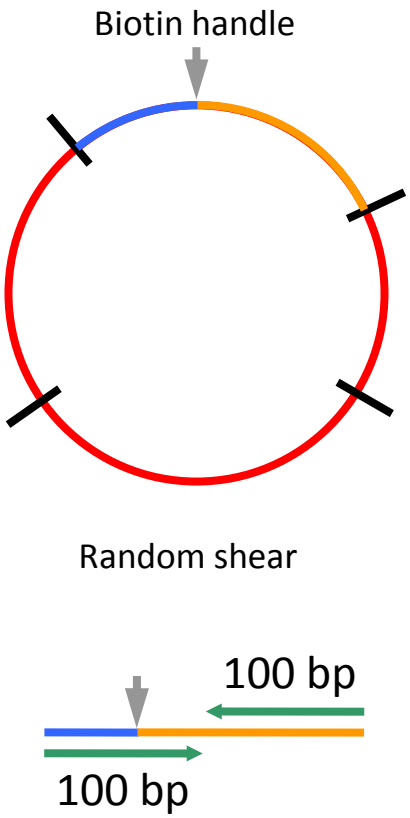
Fragment library

- Size: 180 ± 20 bp
- Most pairs will overlap
- Pairs can be merged computationally
- Used to make unipaths



Jumping library

- Size: 3000 ± 400 bp
- Random shear often causes one of the reads to contain a portion from the other end of the insert
- Provides long range linking!
- Larger insert sizes under development
- These are difficult to produce



Read quality – a further challenge

- Problems with fragment reads
  - Coverage varies greatly across the genome
  - Low coverage regions occur every ~10 Kb
  - Causes breaks in the assembly graph
- Problems with jumping reads
  - Low yield: more DNA required
  - Low yield results in duplicate molecules
  - Problem will increase for larger insert sizes

Low coverage zones

top sequence is reference

```
   TGATACATTTCAGATCAGGTCAGGTGTTCTTCTTTGATTGTTGCAGCTGTTTGTCACAGTTACAATCCACACAAAATCCTGATATATATATATATATATATATATATATATATATAATACCCTCAAAGACTTTGAACTTTGGCGACCTTTACTTTTCCGATAGAAACTTGAAGGAGCTCCTGATCATCGATGAGCCTGCA
RC  TATACATTTCAGATCAGGTCAGGTGTTCTTCTTTGATTGTTGCAGCTGTTTGTCACAGTTACAATCCACACAAAATCCTGATATATATATATATATATATATA
FW                                                                                                TATATATATATATATATATATATATATGTCACACCCTCAAACTTTTAGCAATTTGGCGCTATTTTTTTTTCTCGATAGAAACAAAATATCTCTCCAGCAC
FW                                                                                                          TATATATATATATATATATAATACCCTCAAAGACTTTGAACTTTGGCGACCTTTACTTTTCCGATAGAAACTTGAAGGAGCTCCTGATCATCGATGAGCCT
```

orientation

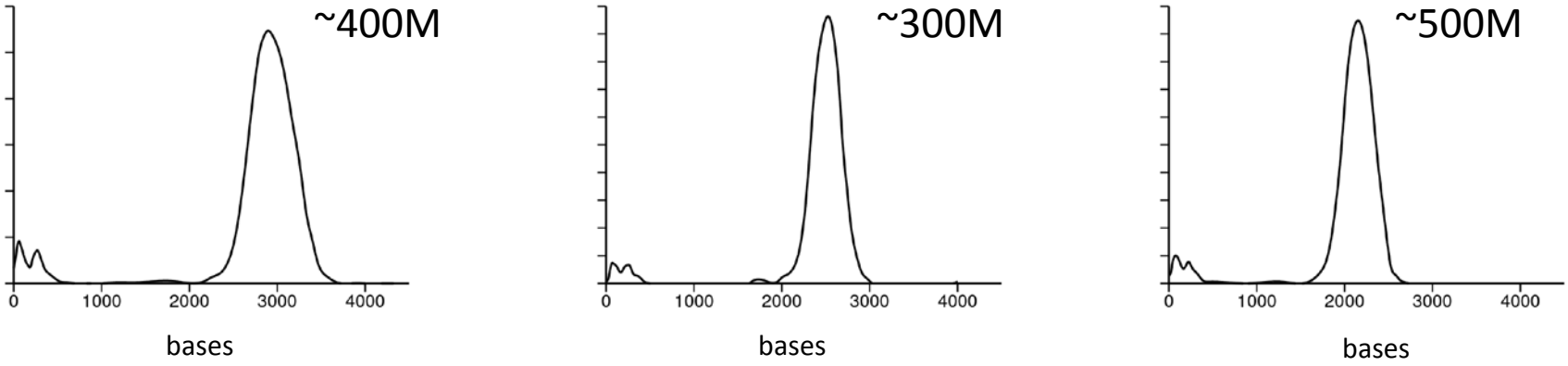
all reads covering this position shown

```
   GCTGGCCCTCAACAACGGCGCGACCAACGGCCACAAACCTCACAGCAACGGCGTCGCTGCCGTGGAGGCGTGAAGCGACTCAGAGTCAACGGTTGGGGGGGGGGGGGGGGGGATCAAAGGATTTGTGGTTGTGTTGAAAATGAGCACCAGCAAACTACGCATCCAGAGAGGGATTTTTGCTTAGATAATTCCAGTGGTTCCA
RC            TGGCCGGGGGTTGGGTTGCCAGGGGCGGCCAGACCTTGGGCGTTTTTCCCGTTGTTTTGTTAACCGTTTTTCAGTTAATGGTTCGGGGGGGGGGGGGGGG
FW                 GGCGAGACCAACGGCCACAAACCTCACAGCAACGGCGTCGCTGCCGTGGAGGCGTGAAGAGACTCAGAGTCAACGGTTGGGGGGGGGGGGGGGGGGGAGAG
FW                       CAACGGCCACAAACCCAACAGCAACGGCGTCGCTGCCGTGGAGGCGTGAAGCGACTCAGAGACATCGGTTCGGGGGGGGGGGGGGGGGAAAAAACACGAGG
FW                                                   TCGCTGCCGTGGAGGAGTGAAGCGACTCAGAGTCAACGGTTGGGCGGGGGGGGGGGGGAACAAGAGTGGTTGGGGTGGTGCACCAAAAACAAACTCAAAG
RC                                                           CGGTGGATGCATAATGGTCCCGACCACGACGGCGCAGGGGGGGGGGGGAGATCAAAGGATTTGTGGTTGTGTTGAAAATGAGCACCAGCAAACTACGCAAC
FW                                                                   AGCTCAGAGTCAACGGTTGGGGGGGGCGGGGGGTGGGTACGGTGTCCTATGTCGCGACACAGCAGGACACACAACATGATGCGCACAGACACAGCGTACAG
FW                                                                   GACTCAGAGTCAACGGTTGGGGGGGGGGGGGGGAGGTAAGAGGTTTGTGTTGTGGGTAGAAGCAACAACACAAAAACACAACACCCACAGAGGGTTTTGTT
FW                                                                         GAGTCAACGGTTGGGGGGGGGGGGGGGGGGGATACAAAGGTTTTGTGTTTTTATTAAAAAAACAAAAAACAAAAAAACCACACAAGGGGCGCGTTTTTTATTG
RC                                                                                 TGGGGGGGGGGGGGGGGGGGATCAAAGGATTTGTGGTTGTGTTGAAAATGAGCACCAGCAAACTACGCATCCAGAGAGGGATTTTTGCTTAGATAATTCCA
FW                                                                                    GGGGGGGGGGGGTGGGGATCAAACAATTTCTTGCTTGTCACACAAAGCCCCAAAGTGTGAAACTATCCCGAACGGGGCATCGCGGCTGCTATAAACACATA
FW                                                                                        GGGGGGGGGGGGGATCAAGGGGTTTGTATTTTTATTTTTTTGTTAACTTCCTGGGCCAGAAAATGGTAAAAGTAGGGAAATCAAAAGGGGTGCTGTCCCT
FW                                                                                         GGGGGGGGGGGGTTAAAGGGTTTTGGGGTGGGGTTGAAAATGAGCACCACAAAACACCGCACCCAGAGGGGAATTTTGGCTAAGATATTTCCAGTGGTTCA
```

On average, these events occur approximately every 10 Kb

Jumping libraries yield

- Yield (# of molecules) and size distribution for 3 jumping libraries
- Each required 5µg of DNA

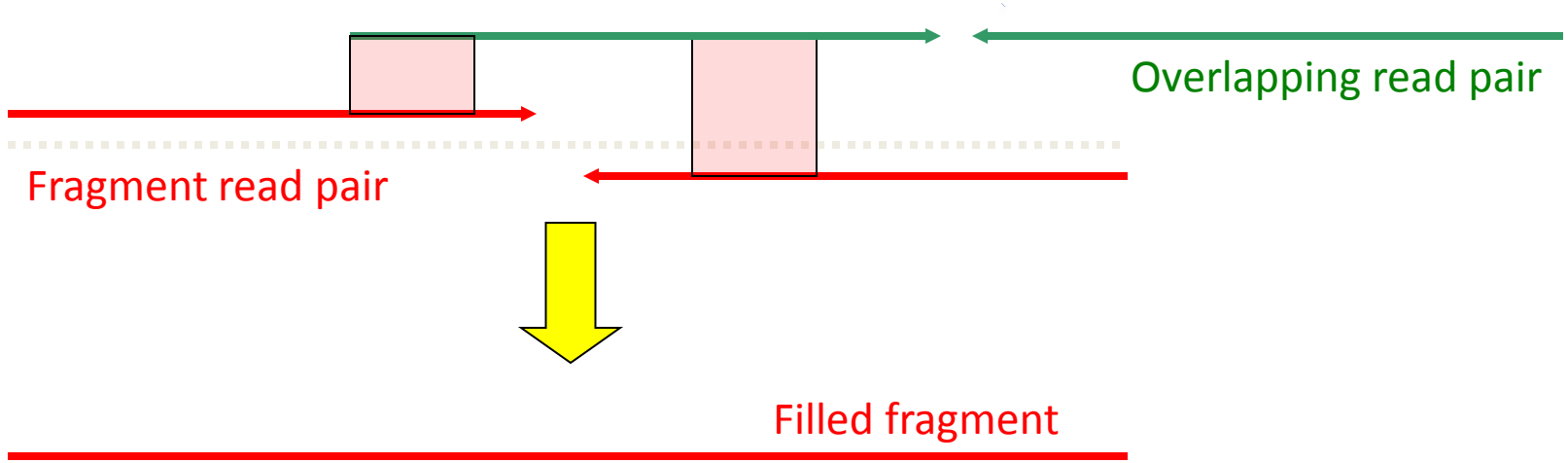


Matt Hims

Strategy to assemble large genomes

1. Merge paired fragment reads (fill fragments)
2. Generate unipaths from filled fragments
3. Error correct jumping reads
4. Localize assemblies (these run in parallel)
5. Merge local assemblies

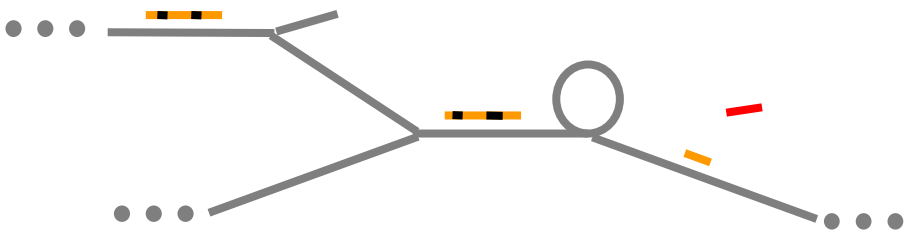
Filling fragments



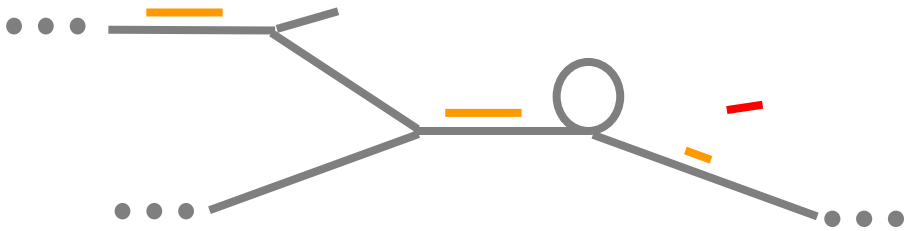
- 89% of fragments filled (Stickleback)
- Aligned to reference with good coverage and few errors

Error correct jumping reads

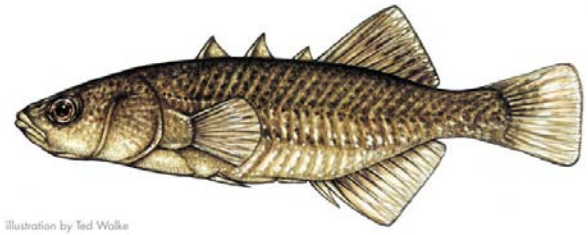
- Align jumping reads to unipaths
- Allow for errors



- Replace reads by their alignment
- Chimeric bits are discarded



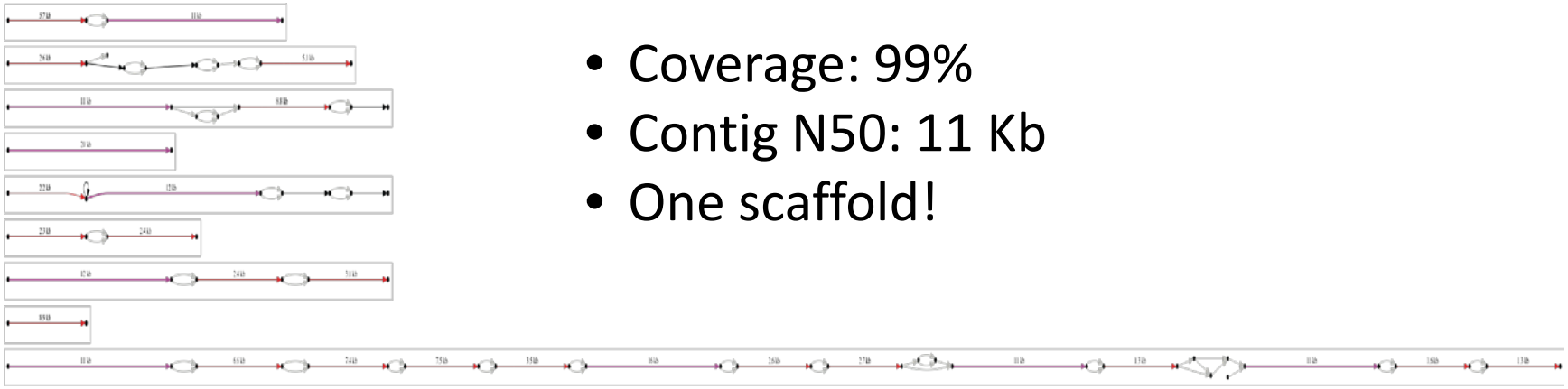
Gasterosteus aculeatus



- Three-spined stickleback
- Our first target:
  - A draft assembly (Sanger sequenced) already exists, and can be used to evaluate our assembly
  - At 450 Mb it is an appropriate stepping stone to mammalian genomes
  - It is highly repetitive and polymorphic

Stickleback – small regions

- The full assembly is in progress
- We can assemble small regions in isolation
- This is an example of a 200 Kb region:



- Coverage: 99%
- Contig N50: 11 Kb
- One scaffold!

Stickleback – whole genome

- Assembly is under way
- We have filled fragments
- 96% of reference covered

For mammalian genomes, we need to

- Improve library yield and quality
- Fix genome coverage issue
- Generate larger jumping libraries
- Reduce memory requirements
- Improve run time via parallelization
- Bushbaby next!

Acknowledgments

The ALLPATHS team

Iain MacCallum
David Jaffe
Joshua Burton
Dariusz Przybylski
Filipe Ribeiro

Chad Nusbaum
Bruce Birren

Scientific Project Management
Carsten Russ

Molecular Biology
Andreas Gnirke
Louise Williams

Broad Institute Genome Sequencing Platform

Illumina
Matt Hims