5- Scaffolding and further improvement

Bernardo J. Clavijo
Richard Smith-Unna
Gonzalo Garcia / Jon Wright





TGAC

A correct assembly has:

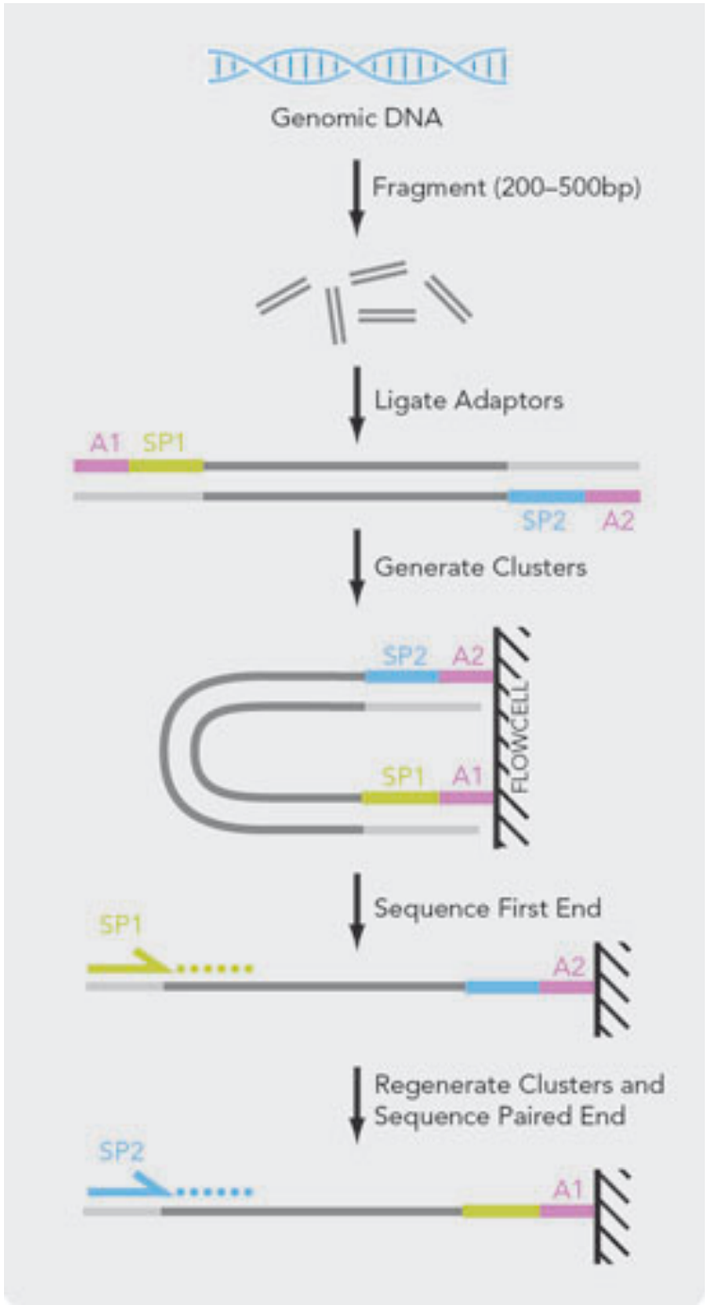
The right *motifs*

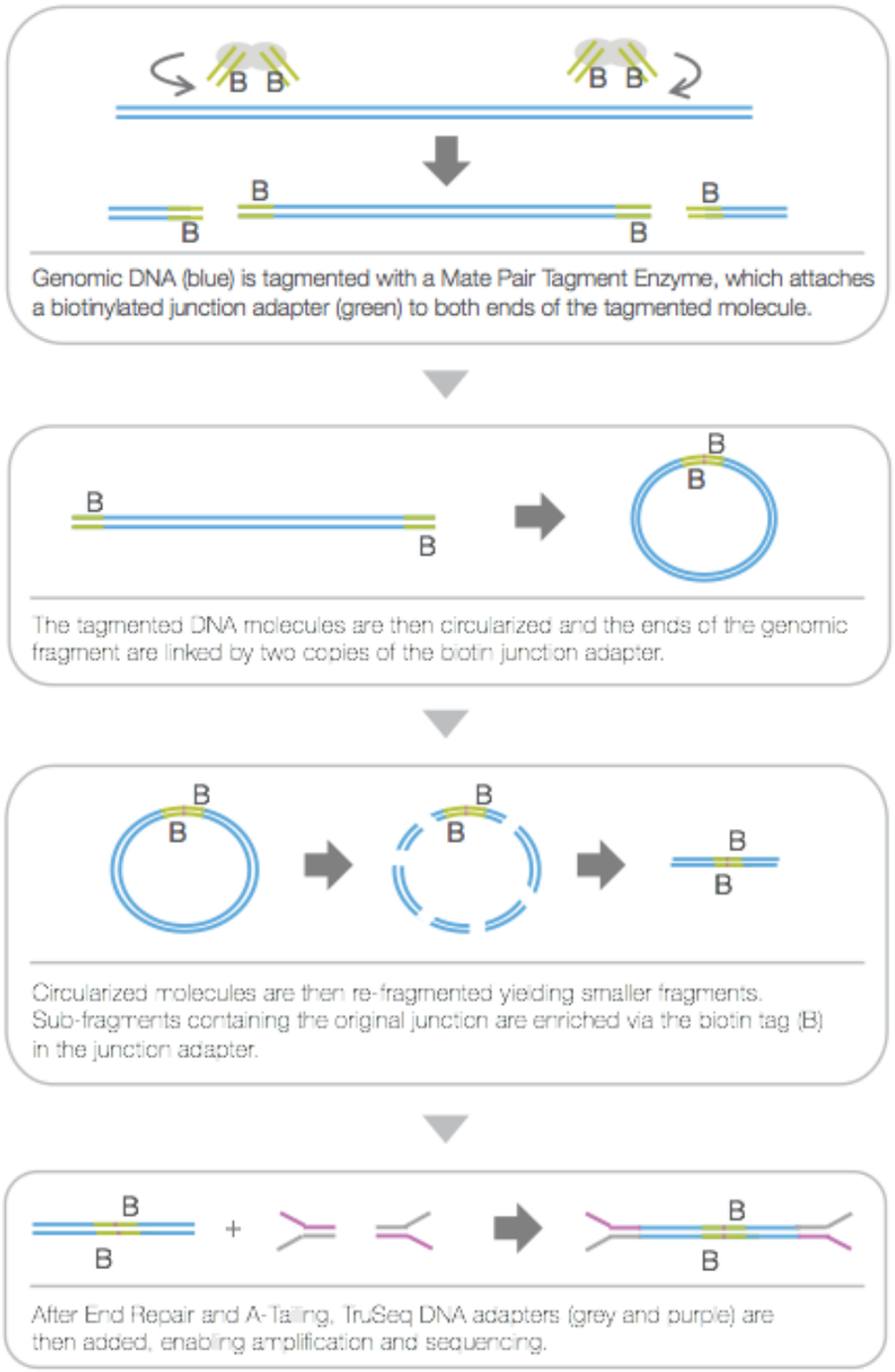
the correct number of times

in the correct order and position.

THE ART OF SCAFFOLDING

Creating and Sequencing Paired Libraries





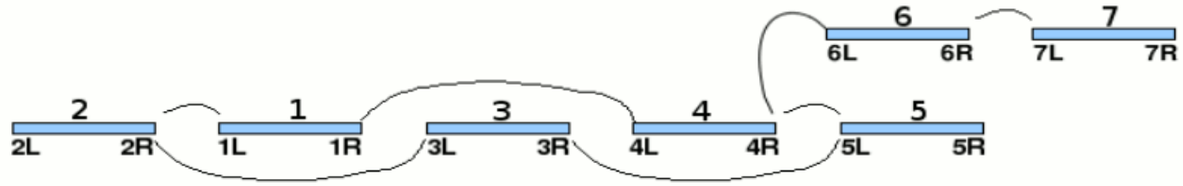
TGAC

Scaffolding with paired reads

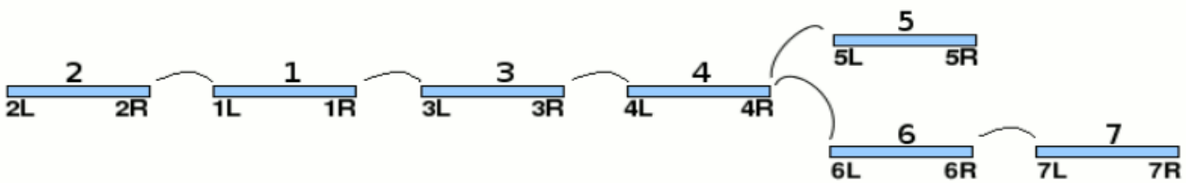
A list of contig-connecting links is calculated from mate pairs

List of contig links:
2R → 1L, 3L
1R → 4L
4R → 5L, 6L
6R → 7L
.....

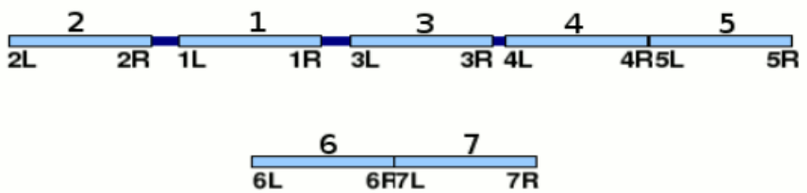
Graph of contig links →



The links are resolved in order to connect adjacent contigs with “path steps” →

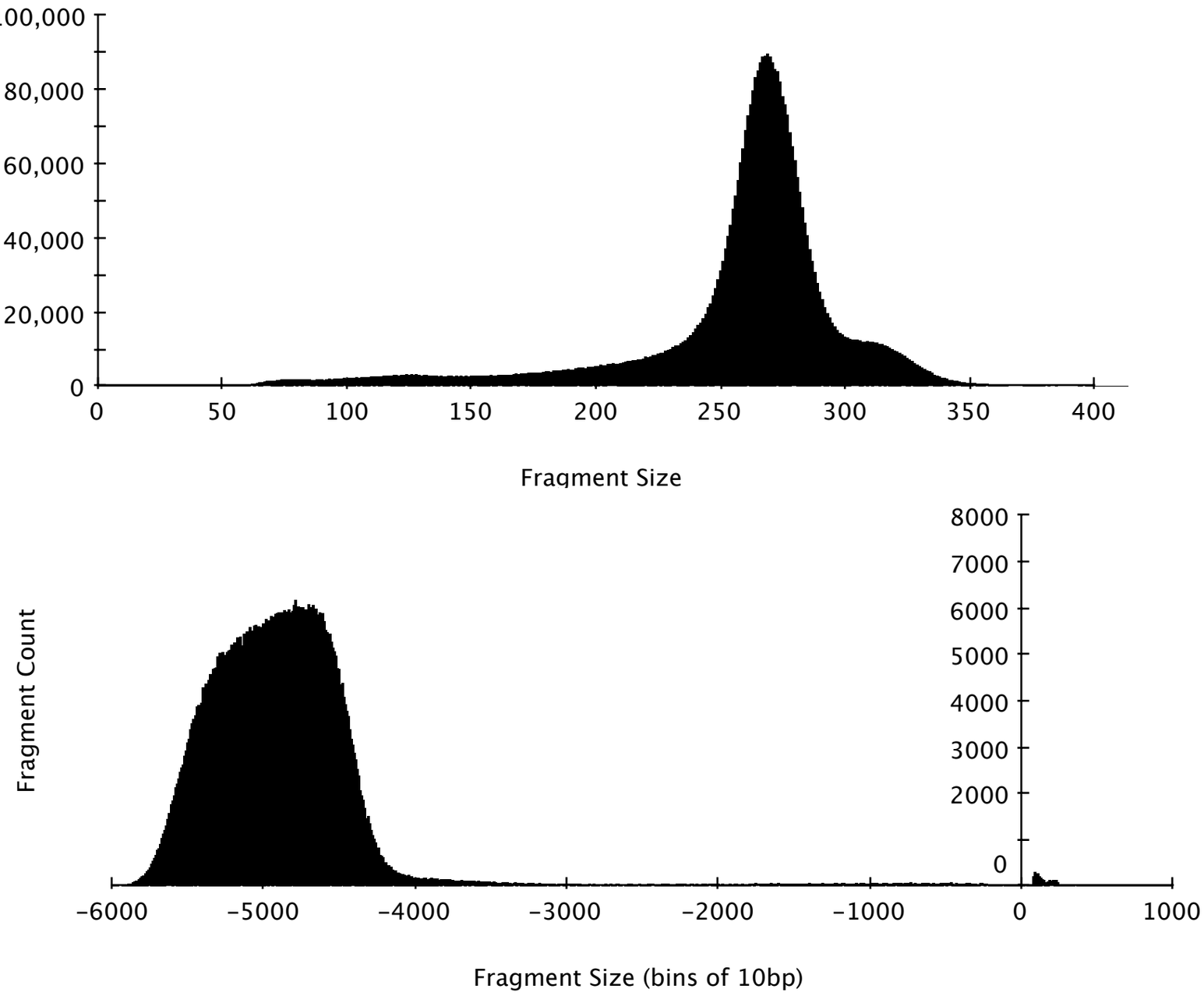


The final scaffold is calculated by resolving conflicting paths →



TGAC

Fragment Sizes



TGAC

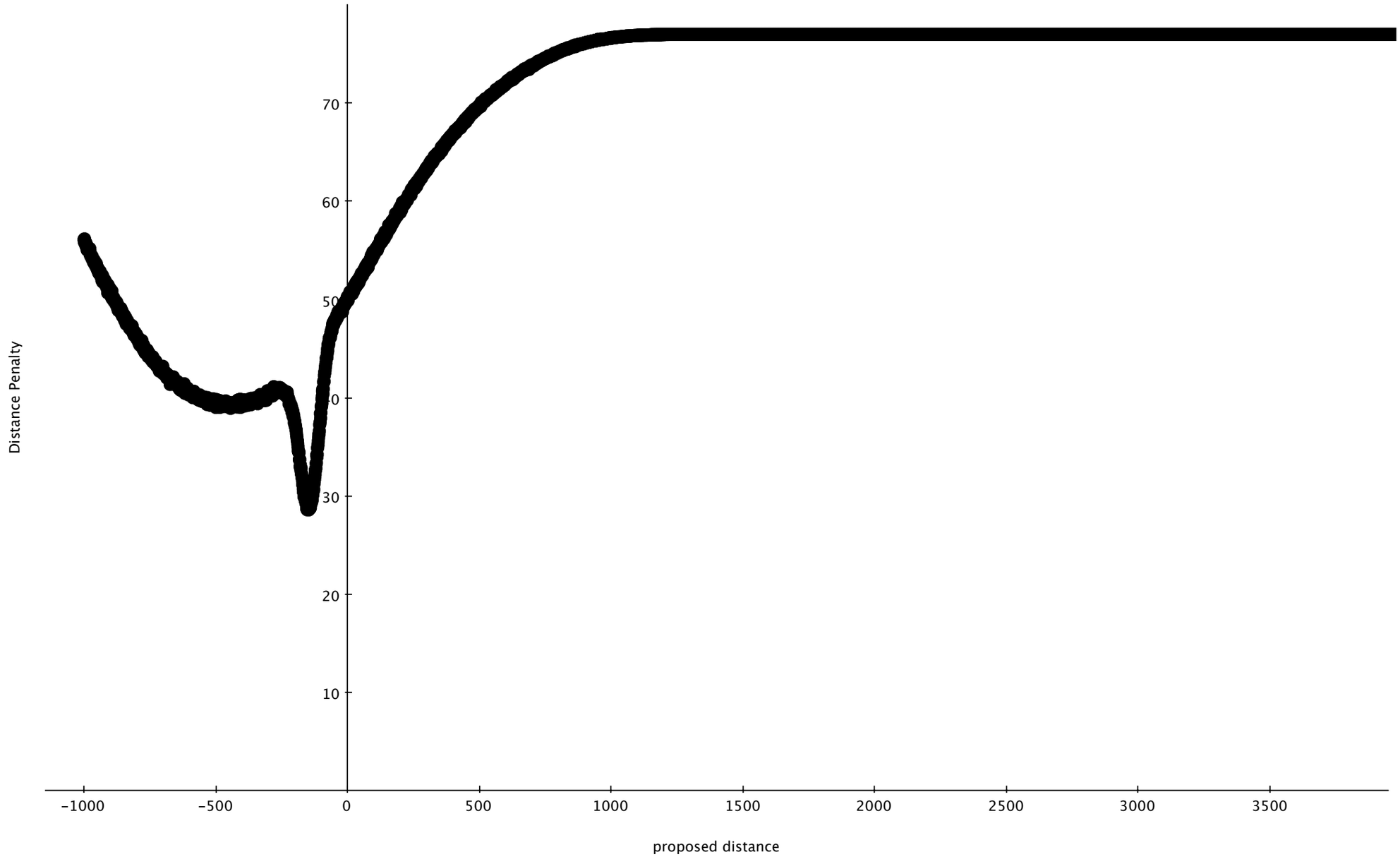
Read mapping stats

```
abyss-map  -j150 -l61   /scratch/clavijob/yeast_tests/diploid/s_3_1_seq
uence.txt /scratch/clavijob/yeast_tests/diploid/s_3_2_sequence.txt nyc3
574_k61-3.fa \
               |abyss-fixmate   -h pe1-3.hist \
               |sort -snk3 -k4 \
               |DistanceEst  -j150 -k61 -l61 -s200 -n10   -o pe1-3.dis
t pe1-3.hist
Building the suffix array...
Building the Burrows-Wheeler transform...
Building the character occurrence table...
Mateless         0
Unaligned    71619  1.14%
Singleton   516328  8.19%
FR         4018144  63.8%
RF              28  0.000444%
FF            8285  0.131%
Different  1686337  26.8%
Total      6300741
```

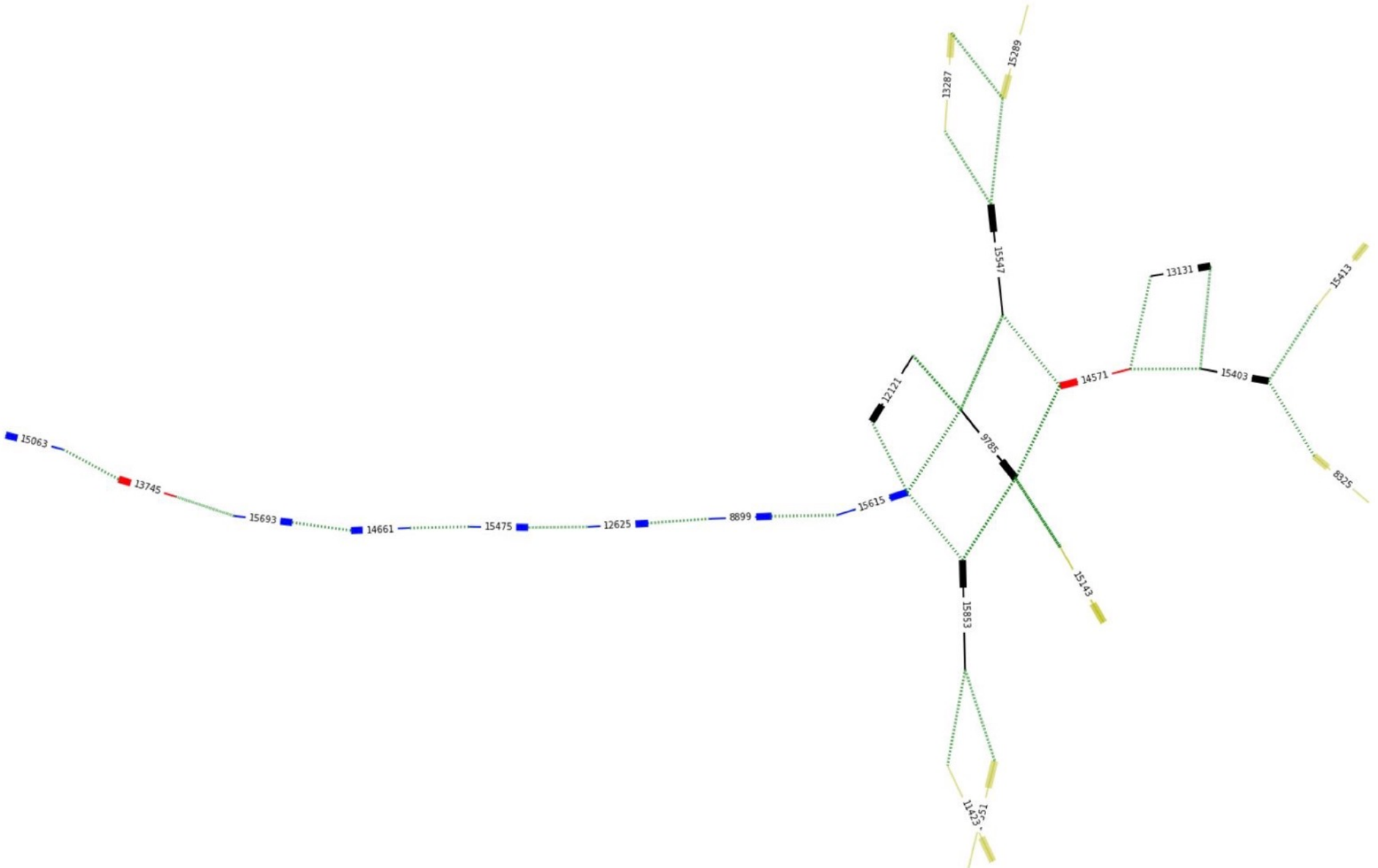
Read mapping stats

```
abyss-map  -j150 -l61   /scratch/clavijob/yeast_tests/diploid/LIB3796_c
lipped_A_R1.fastq /scratch/clavijob/yeast_tests/diploid/LIB3796_clipped
_A_R2.fastq nyc3574_k61-6.fa \
                |abyss-fixmate   -h lmp1-6.hist \
                |sort -snk3 -k4 \
                |DistanceEst --dot  -j150 -k61 -l61 -s200 -n10   -o lmp
1-6.dist.dot lmp1-6.hist
Building the suffix array...
Building the Burrows-Wheeler transform...
Building the character occurrence table...
Mateless         0
Unaligned   127754  6.8%
Singleton   828893  44.1%
FR            3191  0.17%
RF          668696  35.6%
FF           20536  1.09%
Different   230815  12.3%
Total      1879885
```

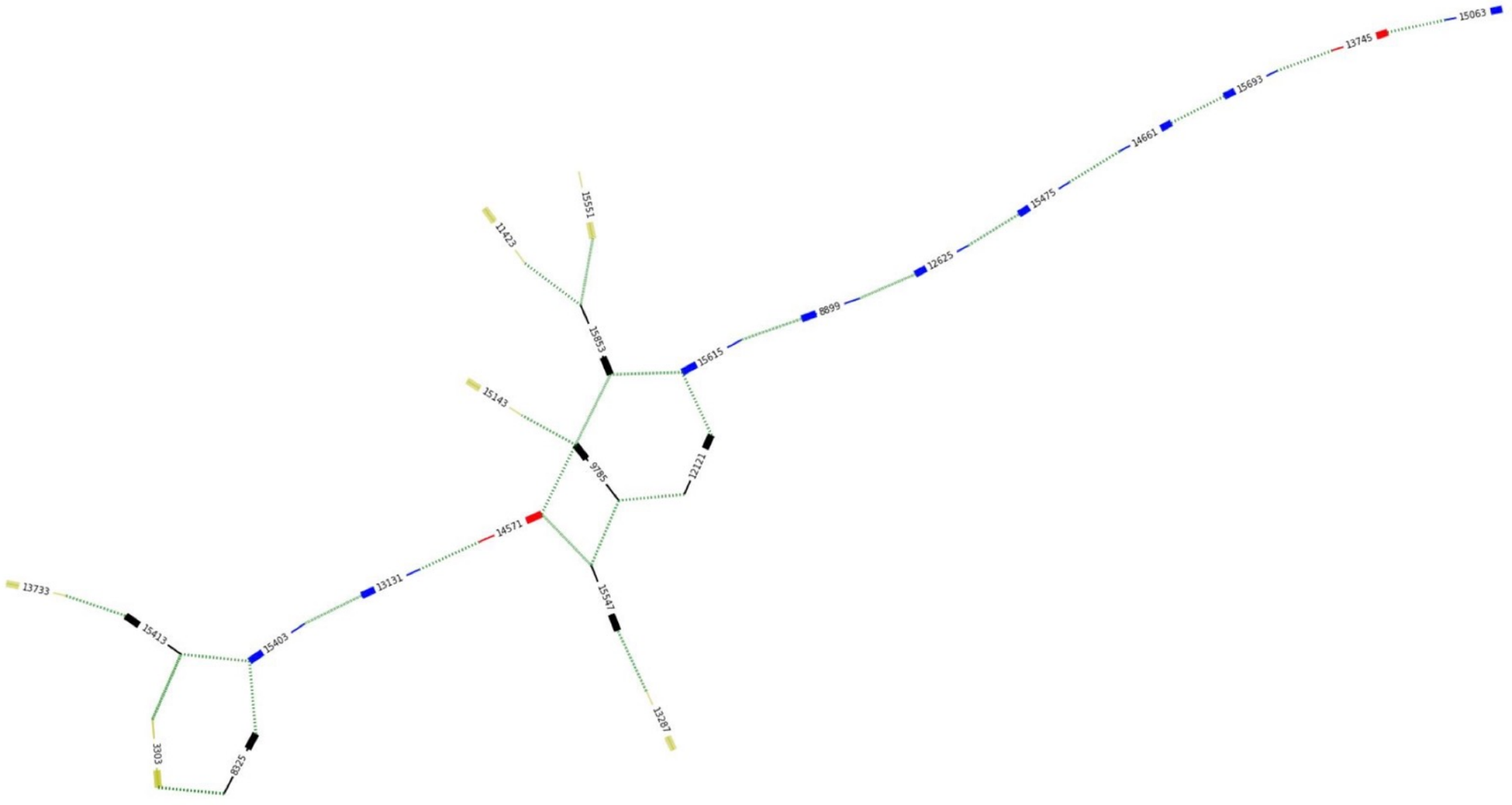
A distance penalty function (from A-scaff)



Repetition expansion and re-linking (from A-scaff)

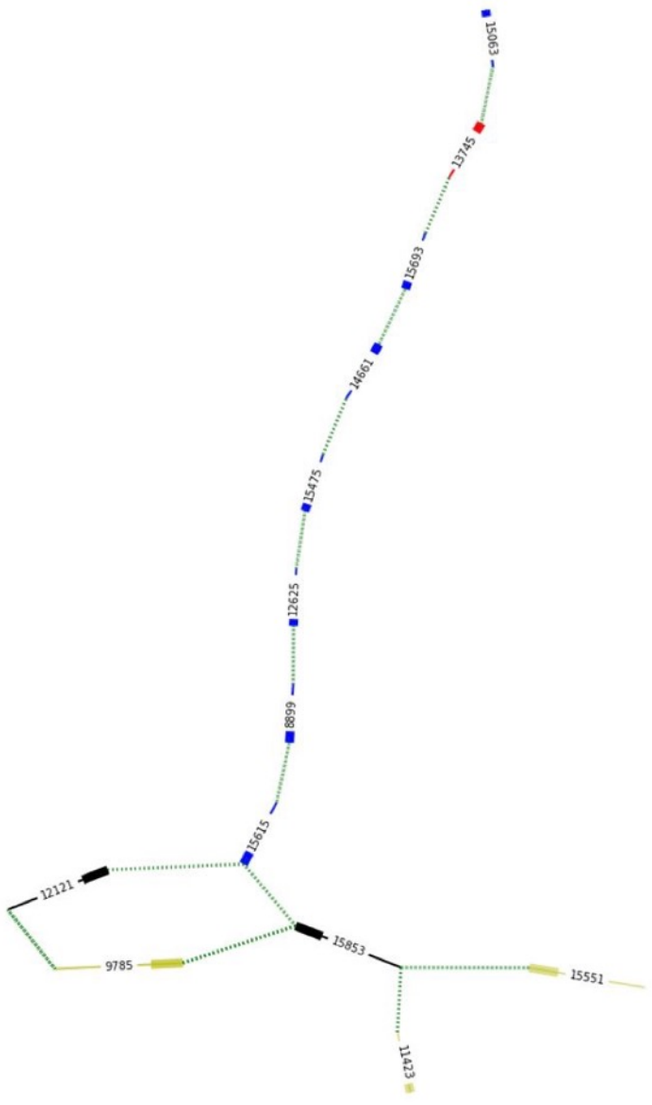


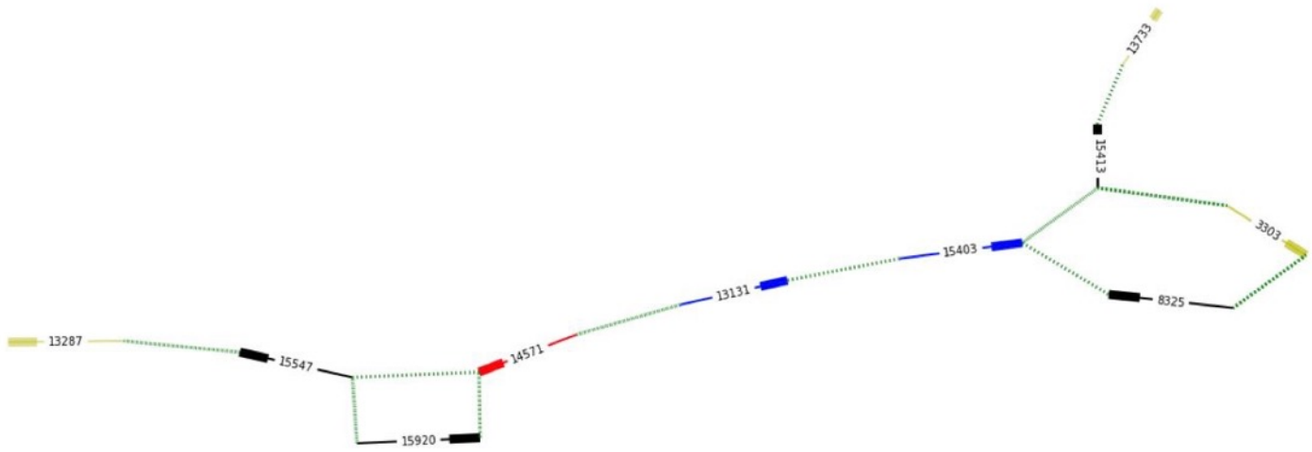
Repetition expansion and re-linking (from A-scaff)



TGAC

Repetition expansion and re-linking (from A-scaff)

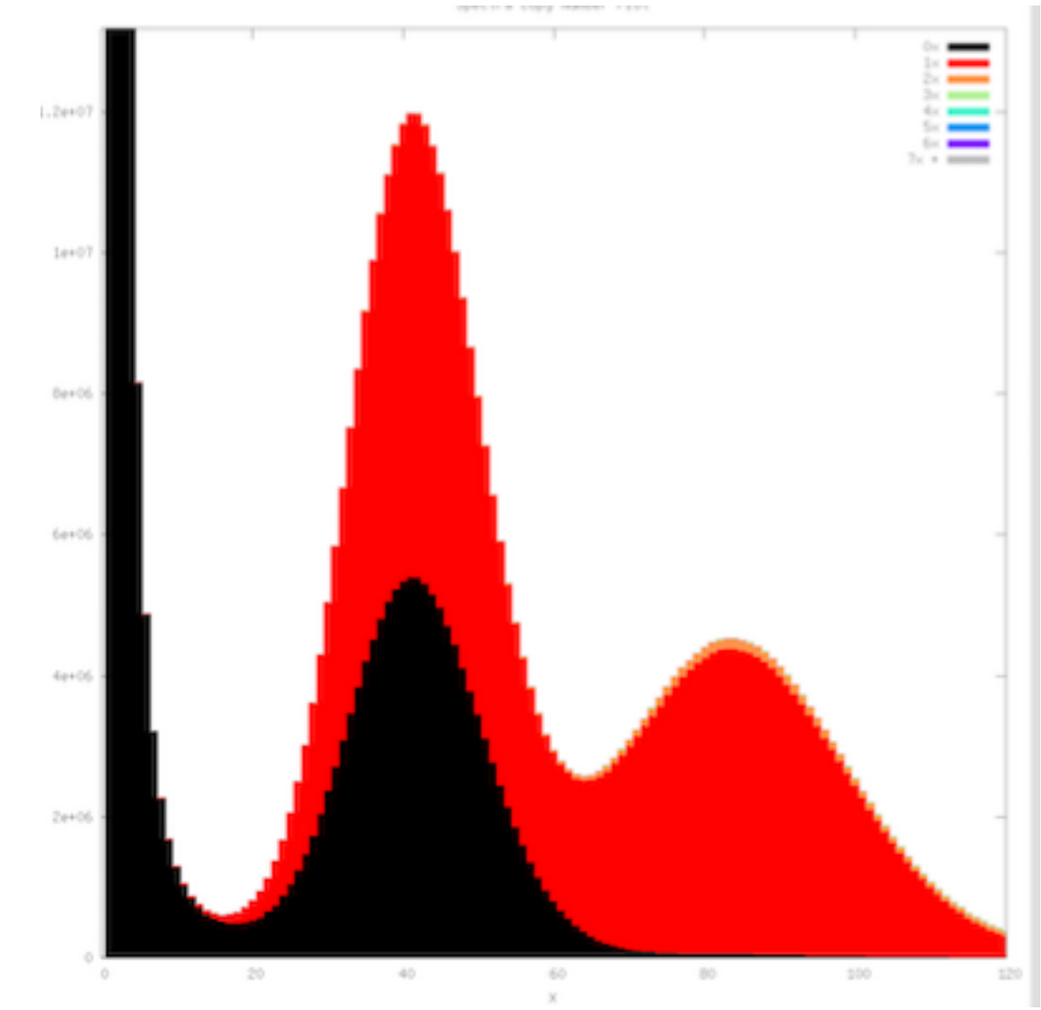




TGAC

Haplotype collapsing and re-scaffolding (from A-scaff)

n	n:500	L50	min	N80	N50	N20	E-size	max	sum	name
413248	53397	4046	500	9803	26774	55122	34956	293176	379.4e6	a.lines.f20.prep.contig
12527	12527	963	1001	37912	104196	217311	138250	771211	360.2e6	DAS_09-93_13_m1000-1.fa



TGAC

source: Genome1
destination: Genome2

lcl|scaffold66
lcl|scaffold229
lcl|scaffold6165
lcl|scaffold69
lcl|scaffold43
lcl|scaffold711
lcl|scaffold61
lcl|scaffold34
lcl|scaffold737
lcl|scaffold95
lcl|scaffold41
lcl|scaffold91
lcl|scaffold82
lcl|scaffold36
lcl|scaffold22
lcl|scaffold64
lcl|scaffold4880960
lcl|scaffold40
lcl|scaffold53
lcl|scaffold115
lcl|scaffold27
lcl|scaffold28

traes3bPseudomoleculeV1

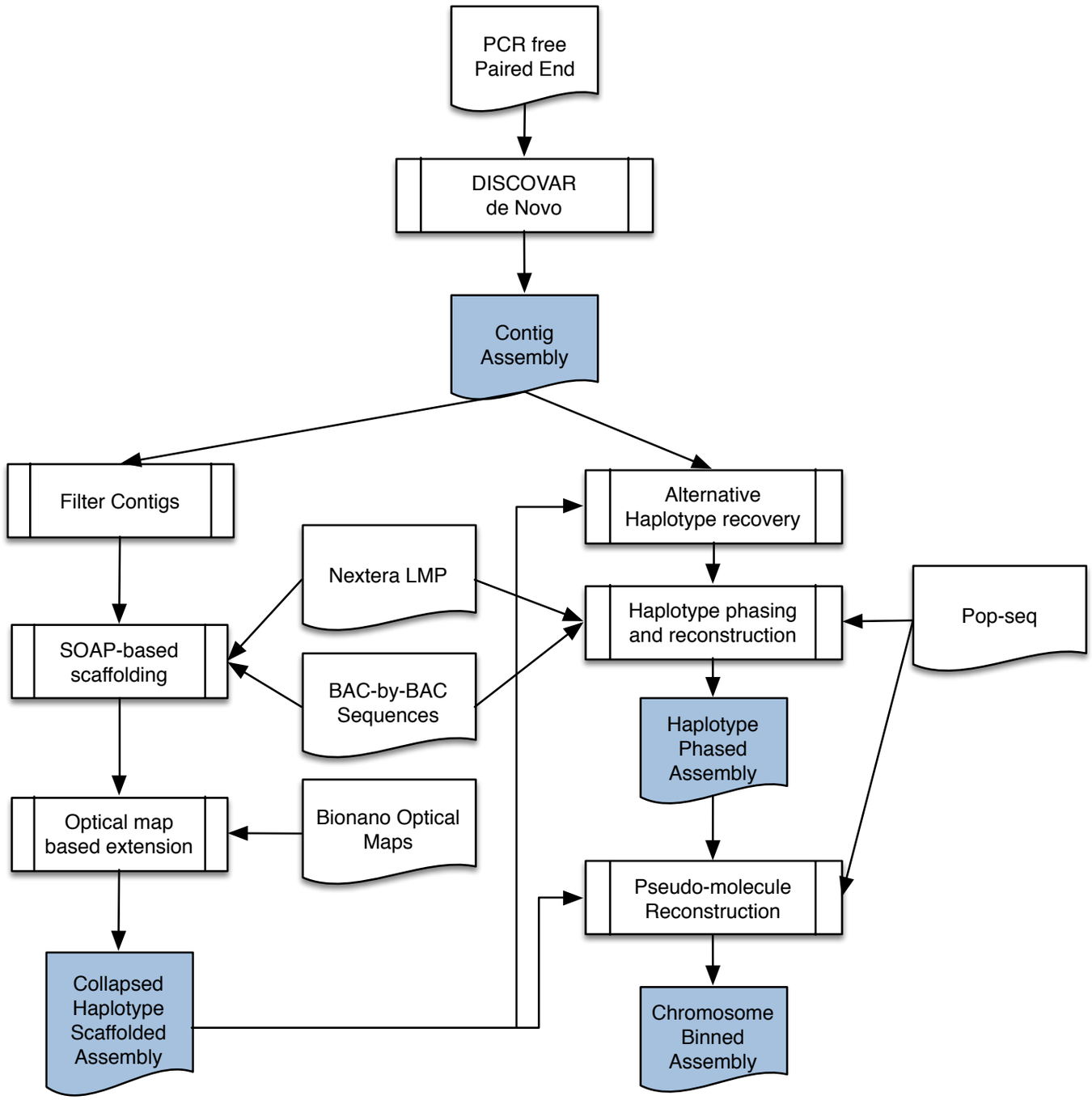
traes3bPseudomoleculeV1

line saturation
- +

About gap closing

- BEWARES:
  - Heuristics are too greedy
  - If there was a gap… When did we lose that information and why?
  - “Walking” is not the same as “bridging”
    - You can be masking problems.
- If you need to:
  - Last step
  - Check QC, metrics and stats before and after, eye-ball typical cases
  - Be conscious it IS a patch

A full assembly pipeline example



TGAC

Integrating non-NGS data, and more…

- Keep in mind the different biases
- Do not expect perfect integration
- You'll need to know your data really well
- Optical mapping and Hi-C (technically NGS) are becoming more typical
- Check internal coherence and significance of every data point you use

Questions?