

# 5 - Data QC, Preparation and First Pass Assembly

Wednesday afternoon

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



# Living on a biased environment

## Review Article

### Library preparation methods for next-generation sequencing: Tone down the bias

Erwin L. van Dijk<sup>a,\*</sup>, Yan Jaszczyzyn<sup>b</sup>, Claude Thermes<sup>a</sup>

<sup>a</sup>Centre de Génétique Moléculaire – CNRS, Avenue de la Terrasse, 91198 Gif sur Yvette, France

<sup>b</sup>Plateforme Intégrée IMAGIF – CNRS, Avenue de la Terrasse, 91198 Gif sur Yvette, France

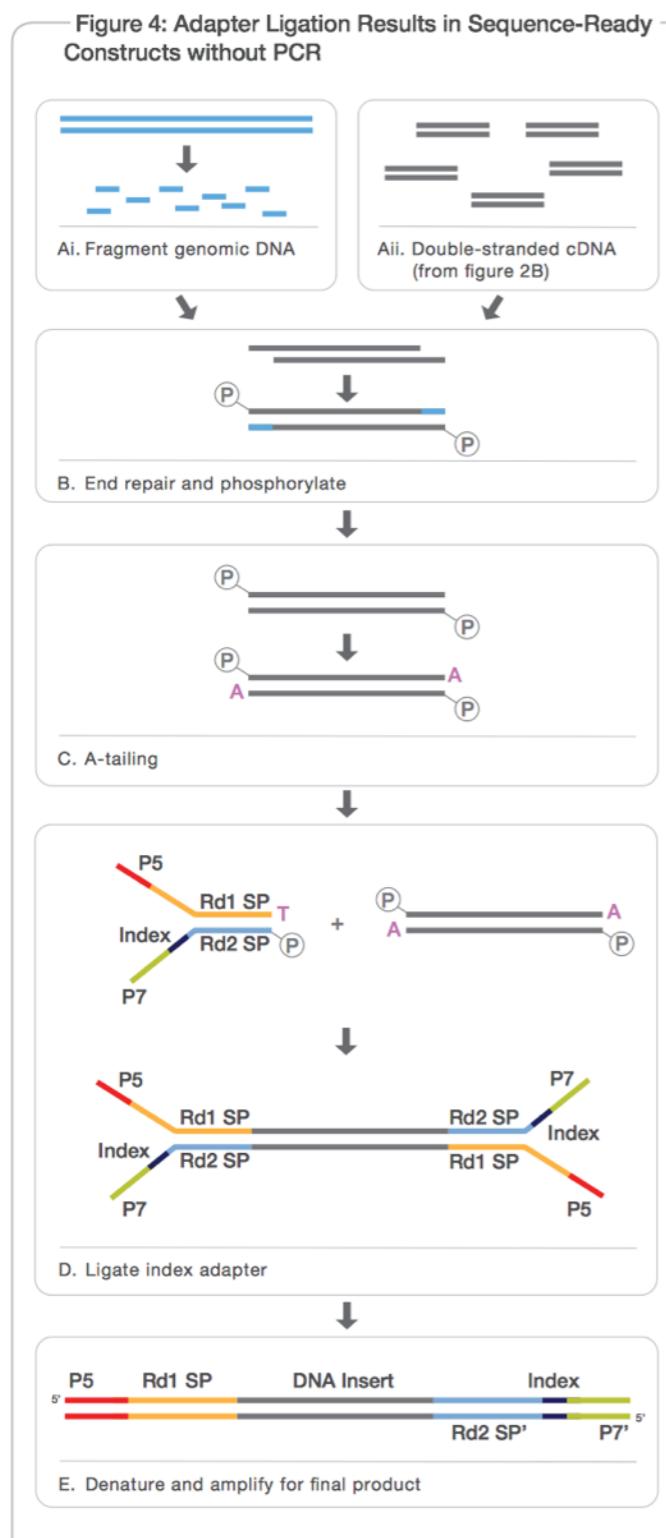
**Table 2 – Sources of bias in RNA-seq library prep and suggestions for improvement.**

Description	Suggestion for improvement	Reference
RNA extraction using Trizol: selective loss of GC poor or highly structured small RNAs at low RNA concentrations	Use similar RNA concentrations for samples that are to be compared or avoid Trizol extraction altogether. Use alternative protocols such as the MirVana miRNA isolation kit.	Kim et al., 2012 [43]
Ribosomal RNA (rRNA) depletion/ mRNA enrichment: bias due to exonuclease targeting partially degraded mRNAs	Use subtractive hybridization rather than exonuclease treatment to deplete rRNAs	He et al., 2010 [26]
RNA fragmentation by RNase III: not completely random, leading to reduced complexity	Use chemical treatment (e.g. zinc) rather than RNase III for RNA fragmentation	Wery et al., 2013 [28]
Random hexamer priming bias: not completely random	A read count reweighing scheme was proposed that adjusts for the bias and makes the distribution of reads more uniform	Hansen et al., 2010 [29]
Reverse transcription: antisense artefacts (especially in the SMART and the NSR methods)	Add actinomycin D to the reaction (not possible for the SMART method)	Perocchi et al., 2007 [38]
Adapter ligation bias due to substrate preferences of T4 RNA ligases	Use adapters with random nucleotides at the extremities to be ligated.	Levin et al., 2010 [25]
		Jayaprakash et al., 2011 [34]
		Sorefan et al., 2013 [31]
		Sun et al., 2011 [35]
		Zhuang et al., 2012 [33]
		Munafo and Robb, 2010 [46]
Reduced ligation efficiency due to RNA modifications	If necessary, perform the appropriate enzymatic treatments to generate a 5' monophosphate and a 3' OH. For RNAs with a 2'-O-methyl modification use adjusted ligation reaction conditions	

**Table 1 – Sources of bias in DNA-seq library preparation and suggestions for improvement.**

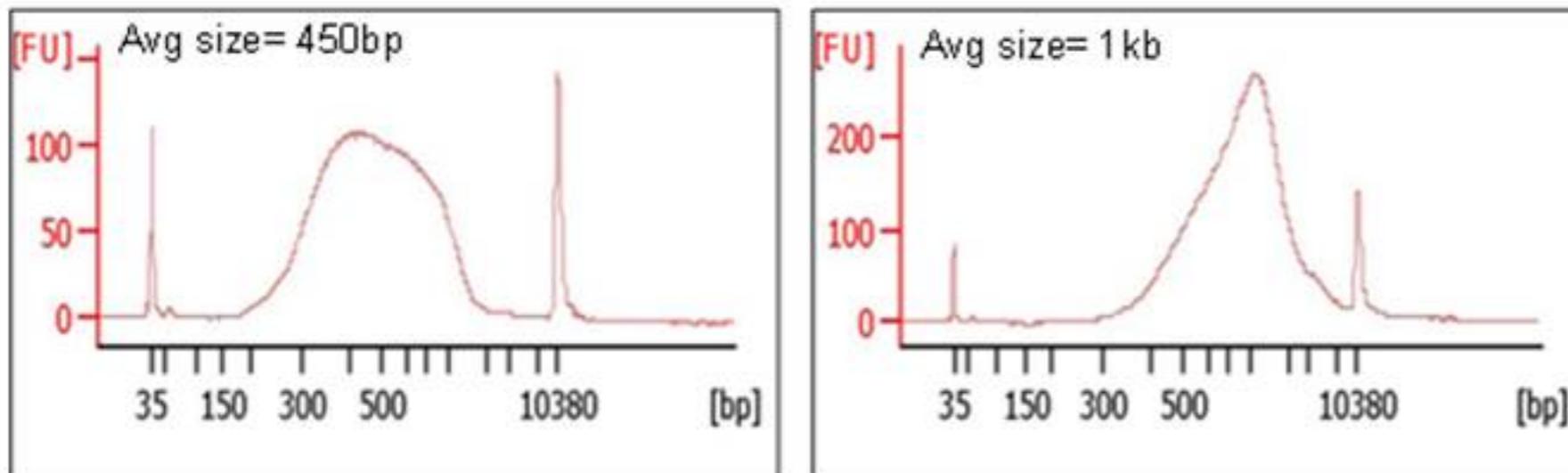
Description	Suggestions for improvement	Reference
<b>Fragmentation of chromatin</b> Bias due to more efficient sonication of euchromatin than heterochromatin	Second fragmentation after IP to concentrate fragments in the optimal size range	Mokry et al 2010 [5]
<b>Size selection</b> Bias due to heating agarose gel slices in chaotropic salt buffer	Melt agarose gel slices at room temperature	Quail et al 2008 [6]
<b>PCR</b> Bias due to preferential amplification of templates with neutral GC%	1. Use Kapa HiFi rather than Phusion polymerase. 2. For AT-rich genomes, use lower extension temperatures and/or use the PCR additive TMAC 3. For GC-rich genomes, use extended denaturation times and/or use the additive betaine 4. For the amplification of minute quantities of genomic DNA (single cell), use MDA rather than PCR	Quail et al 2012 [12] Oyola et al 2012 [13] Aird et al 2011[3] Dabney et al 2012 [14]

# Sample and library preparation: a source of bias



- DNA/RNA extraction techniques have bias:
  - And sample quality limit sequencing!
- Samples are never pure.
- PCR generates further bias.
- No chemical reaction is perfect, nor complete.
- You can learn what your typical biases are:
  - Assess them.
  - Take their impact into account.
  - Try to get better data produced.

# Do not neglect the QC data from the lab



- Concentrations.
- Sample contamination.
- Fragment sizes!

Re: ENQ-1004

Bernardo Clavijo (TGAC)

Sent: Monday, 11 May 2015 14:21

To: [Jens Maier](#) (JIC); [Aaron Heaven](#) (TGAC); [Diana Saunders](#) (TGAC)

Great guys! As you know I love to offload my work to the lab, so basically I make you generate fantastic data so I get to just press a button and deliver a nice result.

Much appreciated,

bj

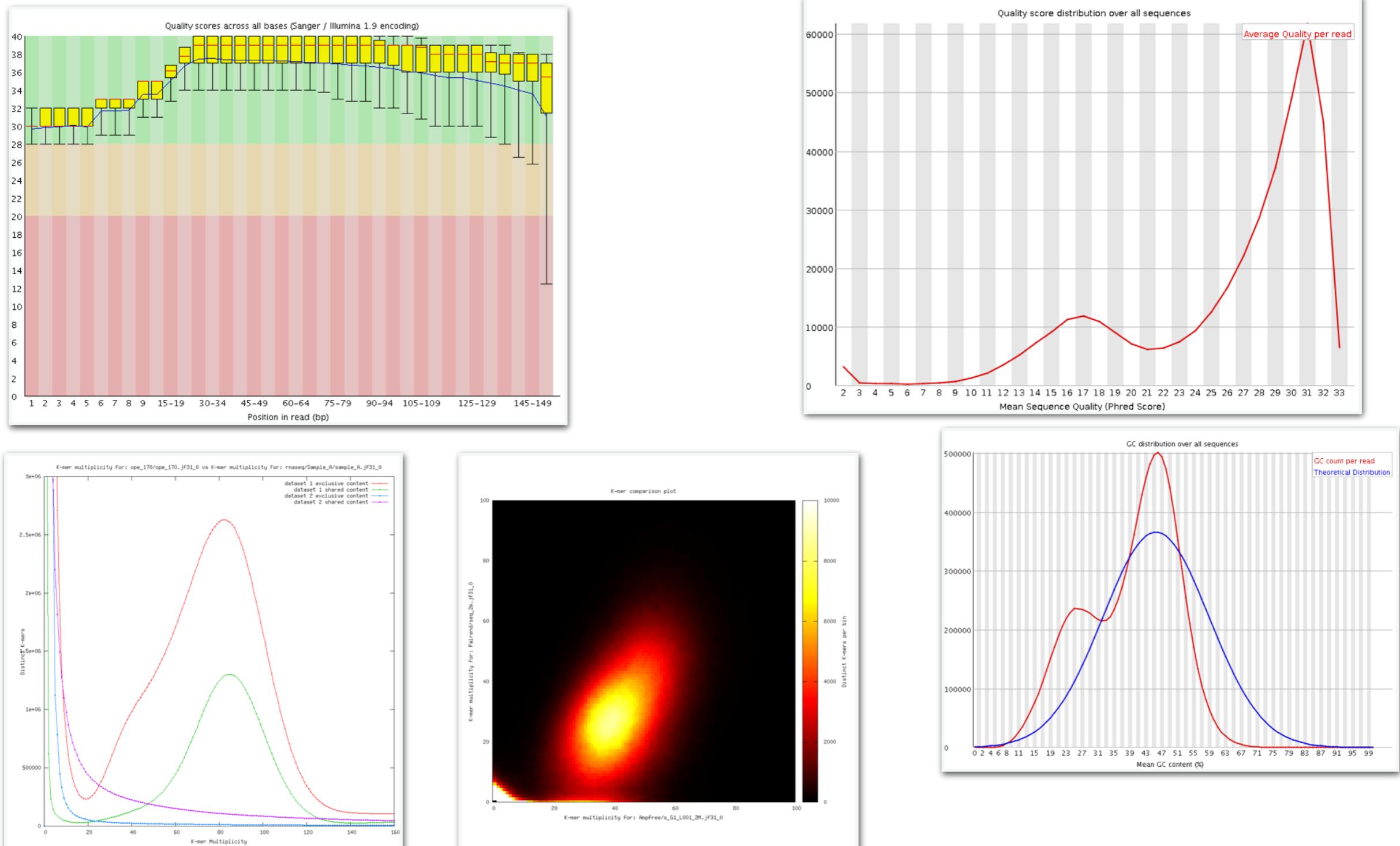
---

**From:** "[Jens Maier](#) (JIC)" <[jens.maier@jic.ac.uk](mailto:jens.maier@jic.ac.uk)>  
**Date:** Monday, 11 May 2015 14:19  
**To:** "[Aaron Heaven](#) (TGAC)" <[aaron.heaven@tgac.ac.uk](mailto:aaron.heaven@tgac.ac.uk)>, Bernardo Clavijo <[bernardo.clavijo@tgac.ac.uk](mailto:bernardo.clavijo@tgac.ac.uk)>, "[Diana Saunders](#) (TGAC)" <[diana.saunders@tgac.ac.uk](mailto:diana.saunders@tgac.ac.uk)>  
**Subject:** RE: ENQ-1004

Mine too! Will make more DNA this week, we still have [spores](#).

Cheers  
Jens

# Do QC before performing the analysis



# PreQC

---

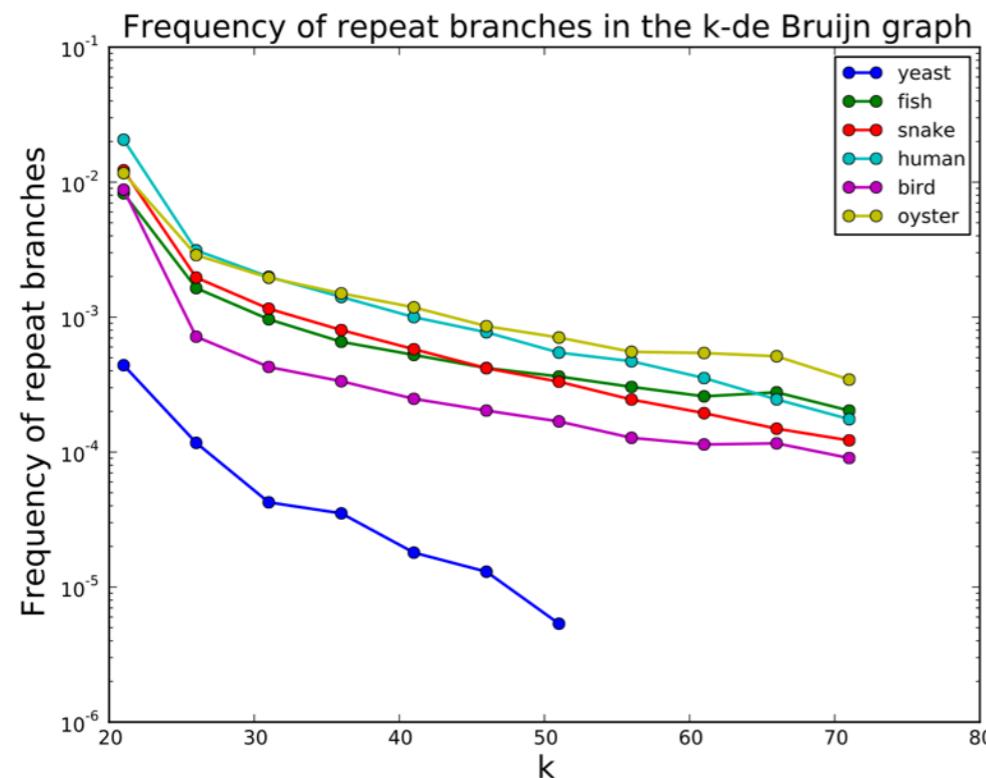


Figure 2: The estimated repeat branch rate for each genome as a function of  $k$ . The yeast data stops at  $k=51$  as the number of repeat branches found falls below the minimum threshold for emitting an estimate.

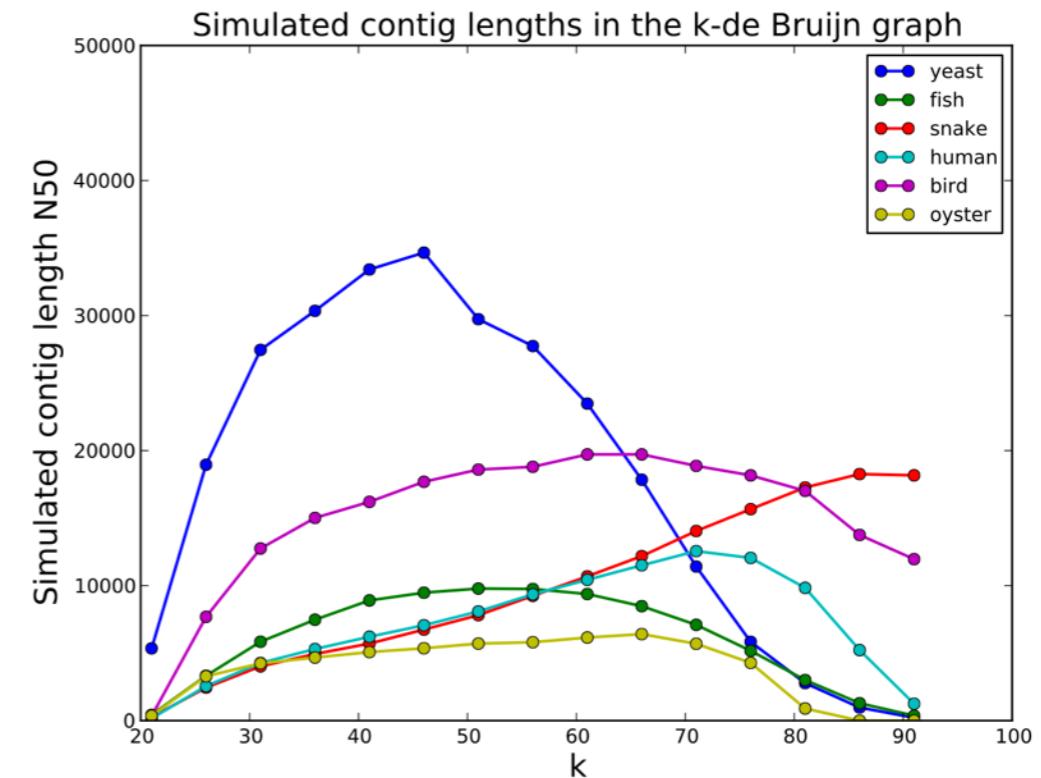
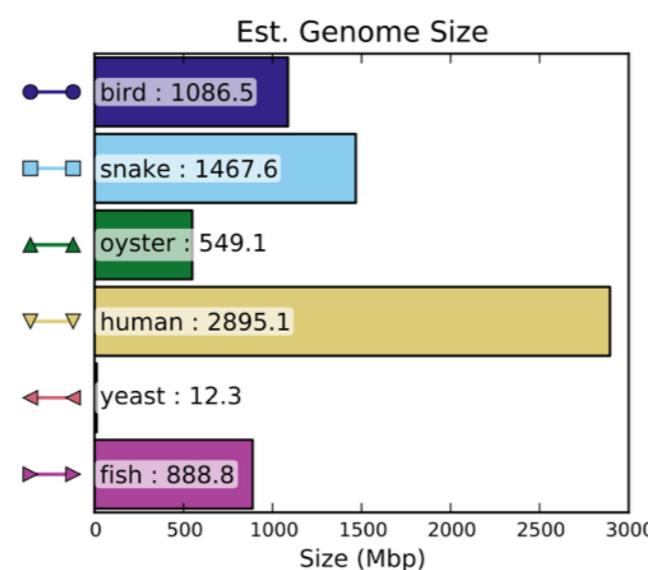
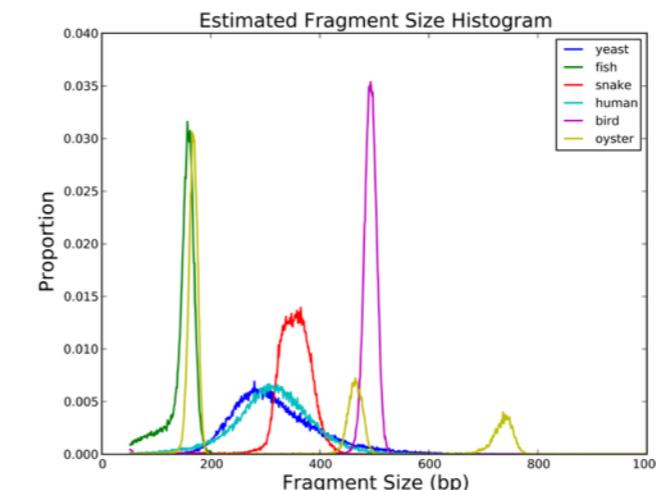


Figure 5: The N50 length of simulated contigs for  $k$  from 21 to 91, in increments of 5

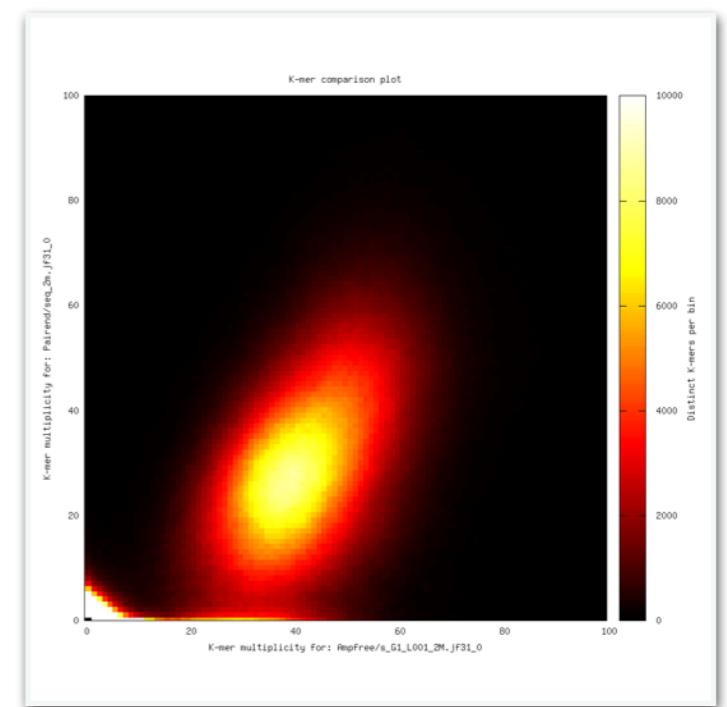
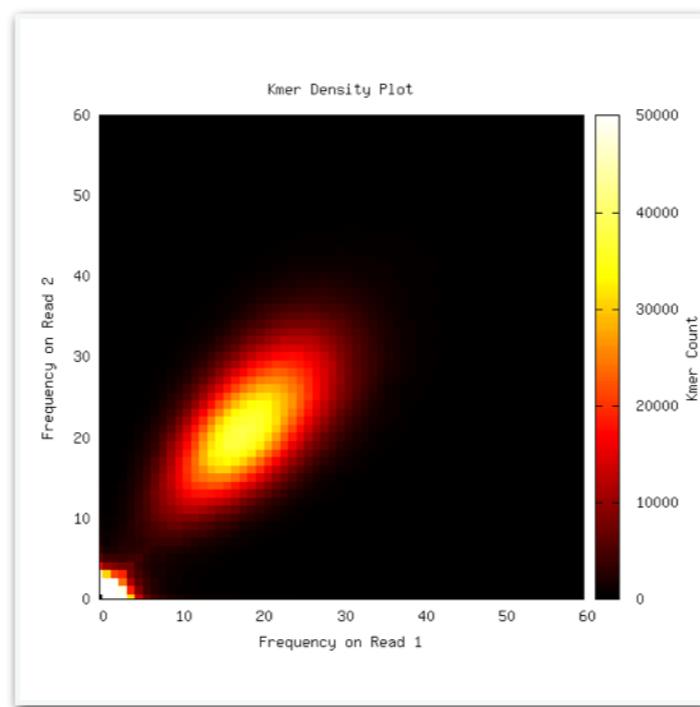
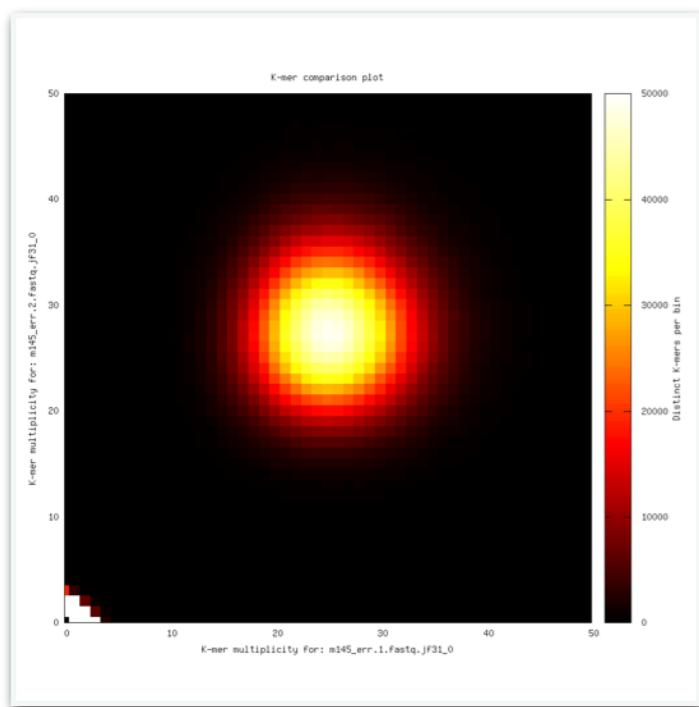
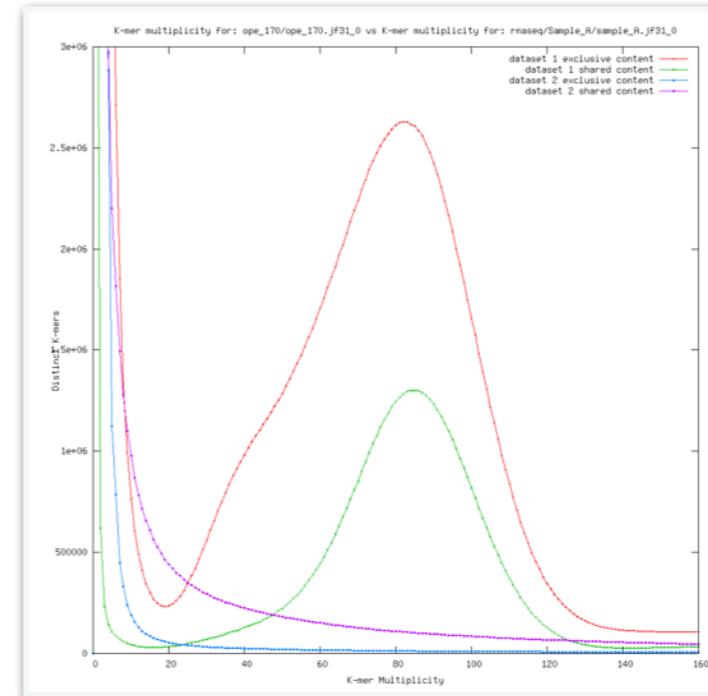
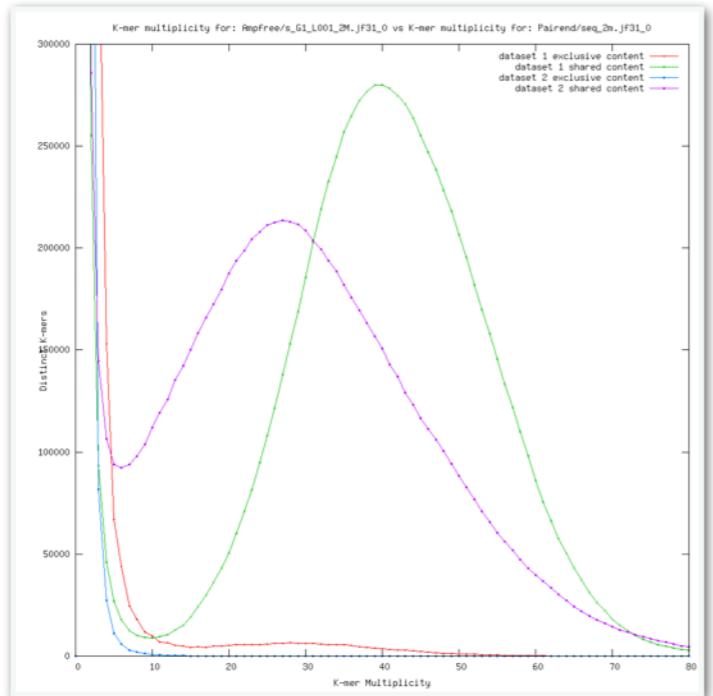


6

Figure 8: The estimated paired-end fragment size for each data set. The oyster data set is a mixture of three libraries.



# Internal and external library coherence (KAT)



# Read preparation:

---

- **Adaptor trimming:** if you have lots of adaptor sequence.
  - But **SPECIALLY** if you have linkers from LMP (check Nextclip).
- **Pair joining:** allows higher k on overlapping reads. Might loose longer frags.
- **Quality trimming:** only if your data is terrible and you are short of memory.
- **Error correction:** once it miscorrects, all subsequent processing is tainted.
  - Your approach should be able to cope with errors, EC is just one option.
  - Pacbio reads are a special case, more about that later.
- **Deduplication:** hard to do right, sometimes needed, scaffolders handle it.
- **Digital normalisation:** rna-\* / meta-\*, and if you understand what it does.
- **IN GENERAL:** illumina is better than it used to be. Keep it in mind.

# Running abyss as a first pass assembler

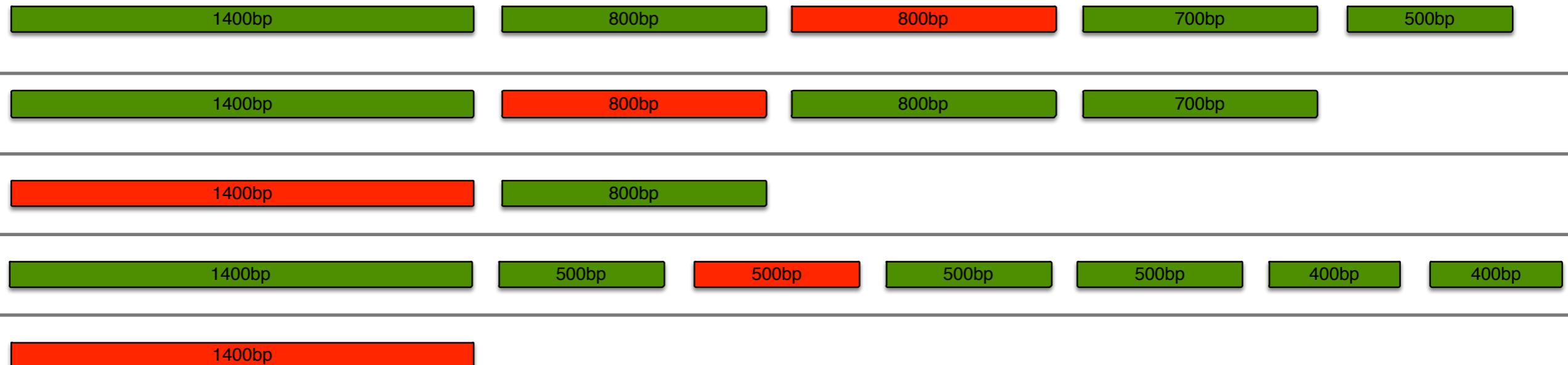
- It runs easily and can use both single and multi-host multiprocessing.
- Creates a ton of useful output, and a nice log.

runabyss.sh	nyc3574_k61-5.fa
Kmer spectra	coverage.hist
	nyc3574_k61-1.fa
	nyc3574_k61-bubbles.fa
	nyc3574_k61-1.adj
	nyc3574_k61-2.adj
	nyc3574_k61-1.path
	nyc3574_k61-3.adj
	nyc3574_k61-2.path
	nyc3574_k61-3.fa
	nyc3574_k61-indel.fa
PE fragment sizes histogram (mapped to unitigs)	nyc3574_k61-unitigs.fa
	pe1-3.hist
	pe1-3.dist
	nyc3574_k61-3.dist
	nyc3574_k61-4.fa
	nyc3574_k61-4.adj
	nyc3574_k61-4.path1
	nyc3574_k61-4.path2
	nyc3574_k61-4.path3
	nyc3574_k61-contigs.fa
	nyc3574_k61-contigs.dot
	lmp1-6.hist
	lmp1-6.dist.dot
	lmp2-6.hist
	lmp2-6.dist.dot
	nyc3574_k61-6.path1.dot
	nyc3574_k61-6.path1
	nyc3574_k61-6.path2
	nyc3574_k61-7.fa
	nyc3574_k61-7.dot
	nyc3574_k61-scaffolds.fa
	nyc3574_k61-scaffolds.dot
	nyc3574_k61-stats
	nyc3574_k61.log

# Beware of N50

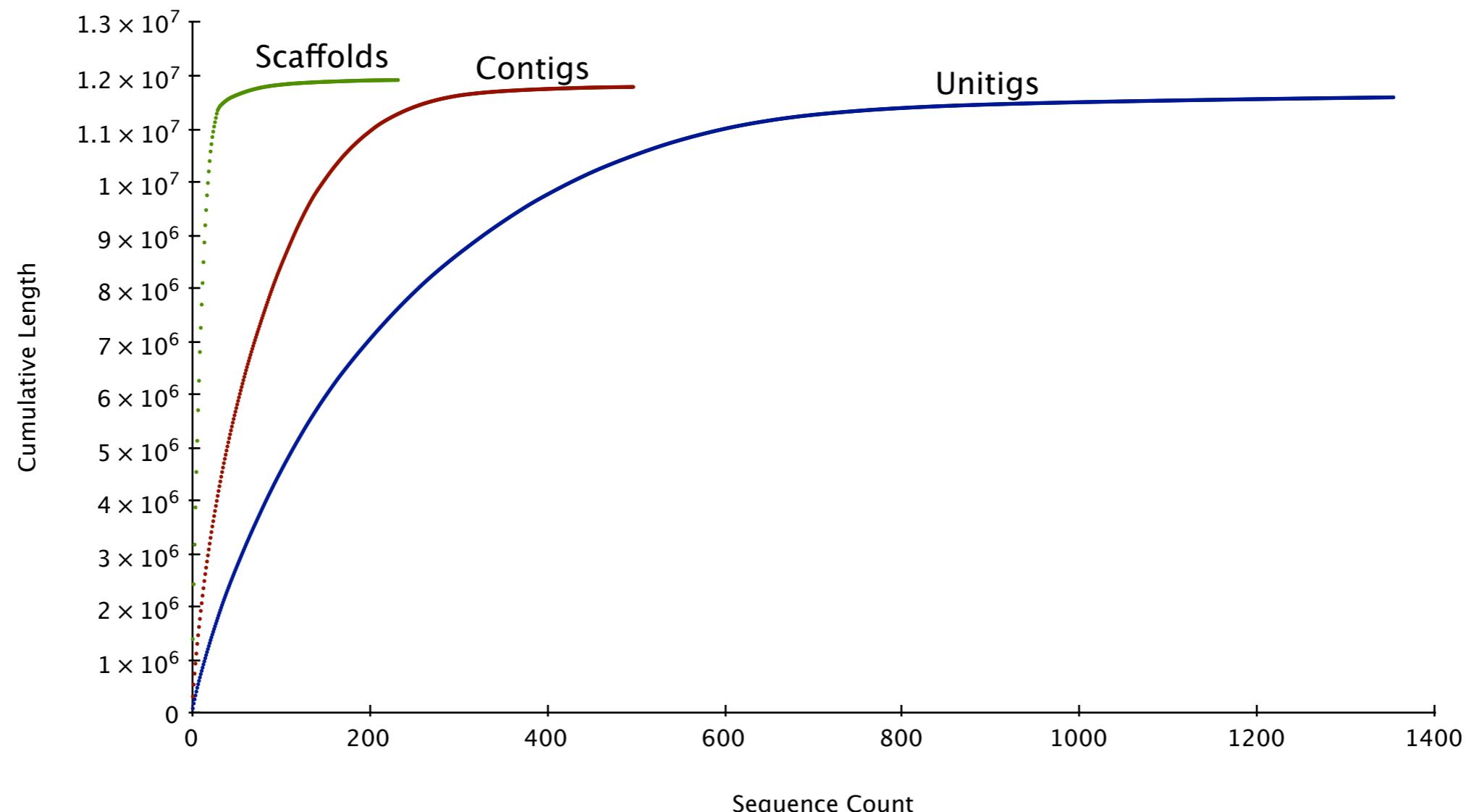
---

- N50 is the most used metric in assembly world... and it should not be:
  - Using contiguity as primary goal reward “risky joining”.
  - N50 is affected by filtering, and not very sensitive!



# Contiguity stats

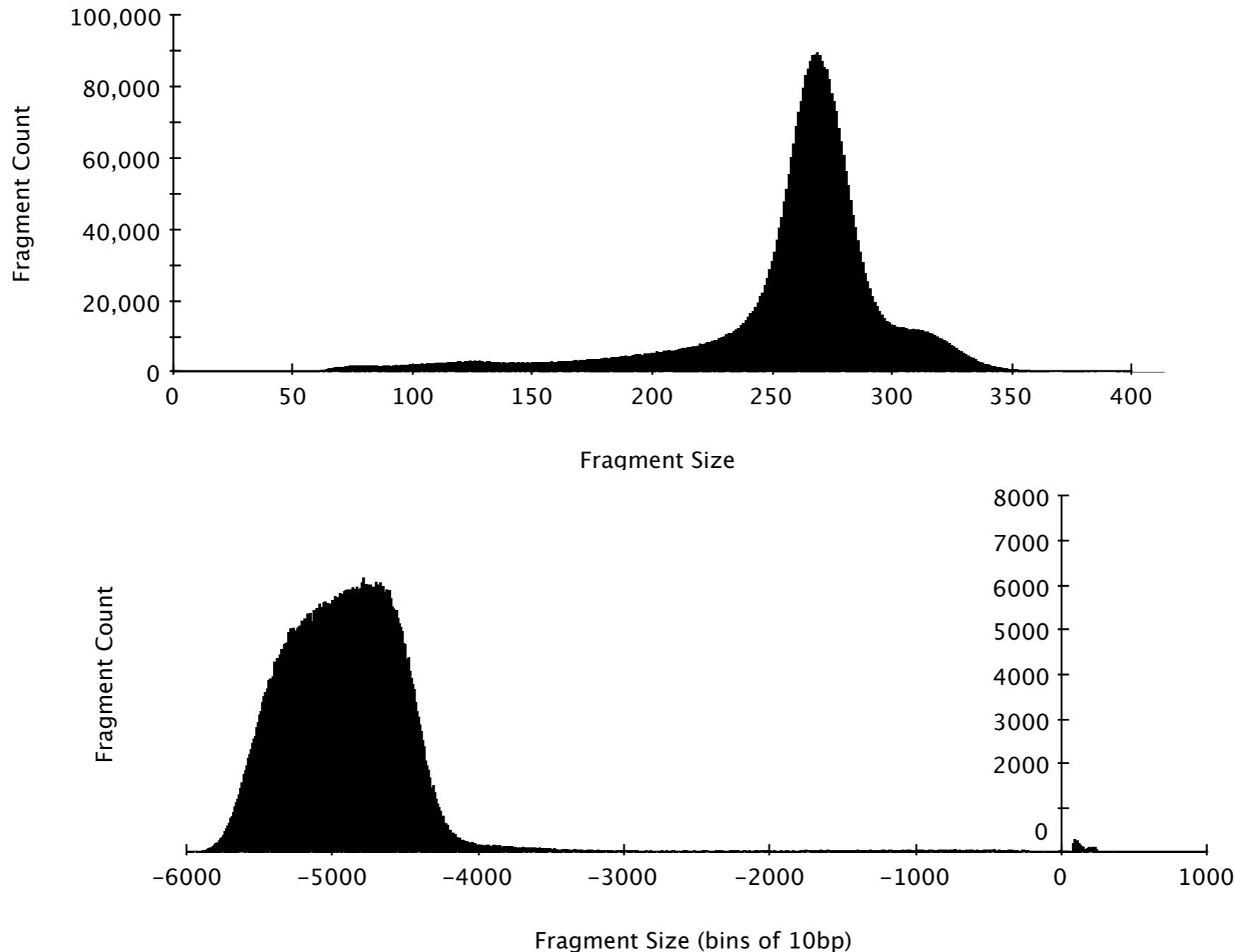
n	n:200	n:N50	min	N80	N50	N20	max	sum	
10773	1353	144	200	11170	25592	44031	96106	11.6e6	nyc3574_k61-unitigs.fa
8880	497	53	200	32554	66307	139116	322315	11.79e6	nyc3574_k61-contigs.fa
8615	232	8	200	269923	551245	1029531	1372216	11.74e6	nyc3574_k61-scaffolds.fa



- Don't forget to check your “Ns” !!!

# Fragment Sizes

---



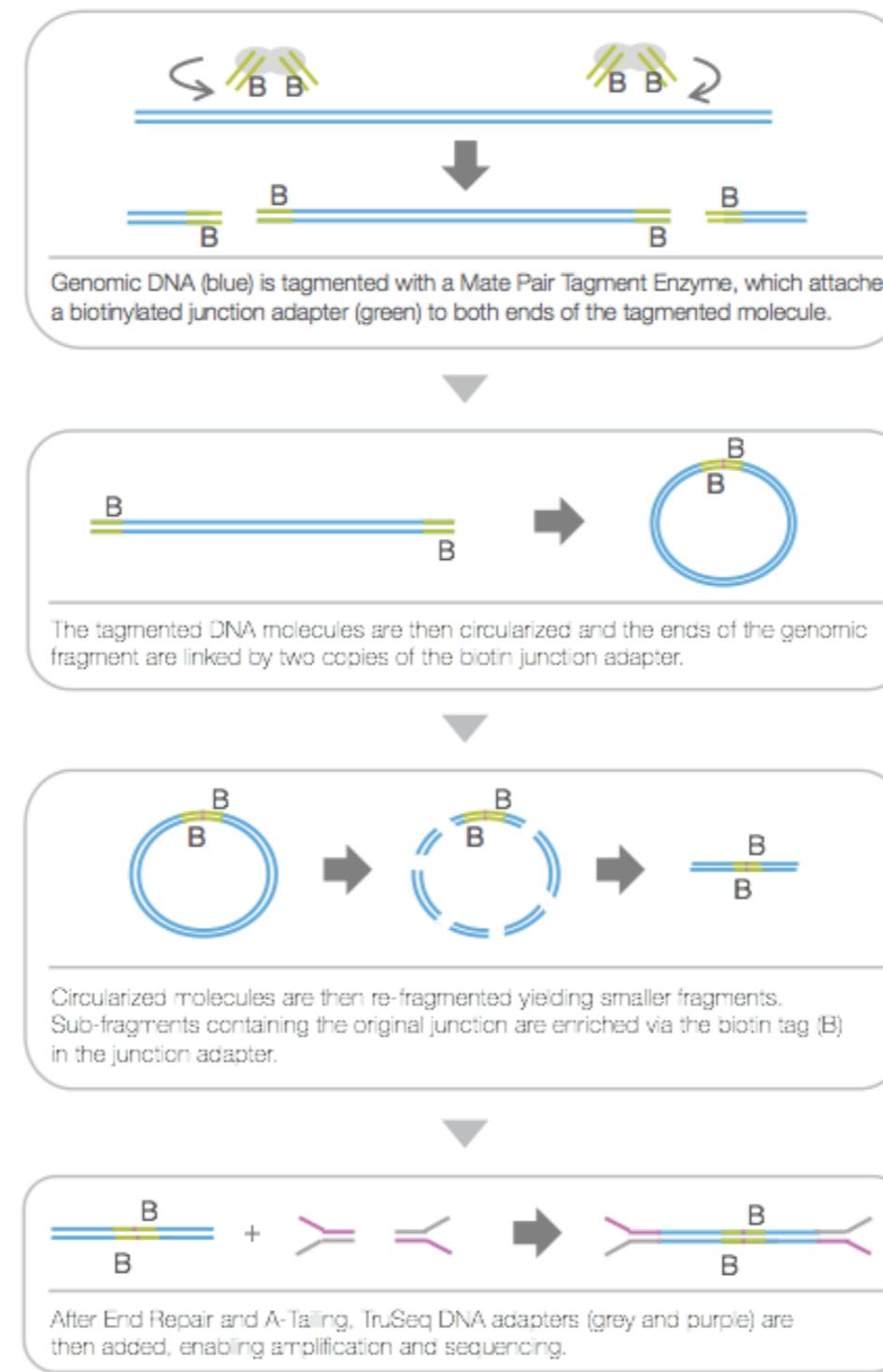
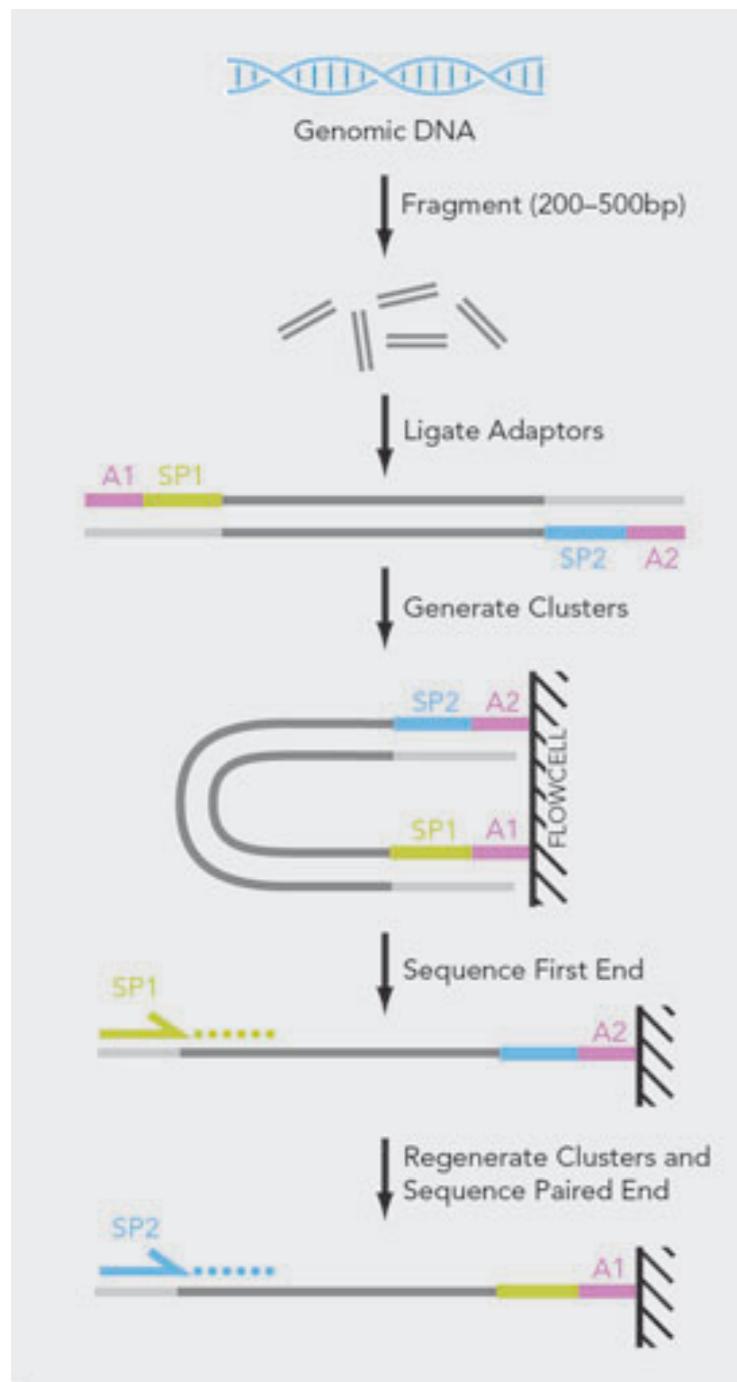
# Questions?



# Preparing LMP data...

... an exceptionally successful example.

# Paired End & Long Mate Paired Libraries



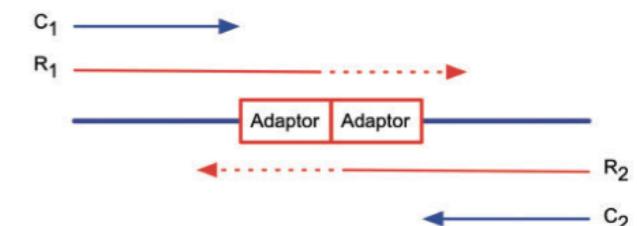
BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 4 2014, pages 566–568 doi:10.1093/bioinformatics/btt702

Sequence analysis Advance Access publication December 2, 2013

## NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries

Richard M. Leggett\*, Bernardo J. Clavijo, Leah Clissold, Matthew D. Clark and Mario Caccamo

The Genome Analysis Centre (TGAC), Norwich Research Park, Norwich NR4 7UH, UK  
Associate Editor: Michael Bruno



**Fig. 1.** Nextera mate pair fragments are formed by the joining of two junction adaptors. Reads R<sub>1</sub> and R<sub>2</sub> are produced from both ends and are clipped at the adaptor to produce C<sub>1</sub> and C<sub>2</sub>

**Table 1.** ABYSS *A. thaliana* assembly with and without NextClip clipping.

Reads used for assembly	Contig N50	Scaffold N50
Paired End only	15,627	21,939
PE and all raw LMP	15,627	15,628
PE and NextClip processed A,B,C cat.	15,627	245,226

# Yeast Assemblies

---

n	n:500	n:N50	min	N80	N50	N20	max	sum	
17783	7370	2171	500	744	1205	2024	6372	8031482	yeast_pe_k91-unitigs.fa
9030	4941	1079	500	1306	2708	5235	17385	9789489	yeast_pe_k91-contigs.fa
8810	4842	1037	500	1336	2753	5478	82405	9811356	yeast_pe_k91-scaffolds.fa
n	n:500	n:N50	min	N80	N50	N20	max	sum	
17783	7370	2171	500	744	1205	2024	6372	8031482	yeast_pe_raw_lmp1_k91-unitigs.fa
9030	4941	1079	500	1306	2708	5235	17385	9789489	yeast_pe_raw_lmp1_k91-contigs.fa
8571	4662	1024	500	1434	2898	5409	17385	9850977	yeast_pe_raw_lmp1_k91-scaffolds.fa
n	n:500	n:N50	min	N80	N50	N20	max	sum	
17783	7370	2171	500	744	1205	2024	6372	8031482	yeast_pe_clipped_lmp1_abc_k91-unitigs.fa
9030	4942	1079	500	1306	2708	5235	17385	9789862	yeast_pe_clipped_lmp1_abc_k91-contigs.fa
6907	2988	232	500	1967	10653	25702	74395	9843053	yeast_pe_clipped_lmp1_abc_k91-scaffolds.fa

# What if... I assembly the LMP library **alone**?

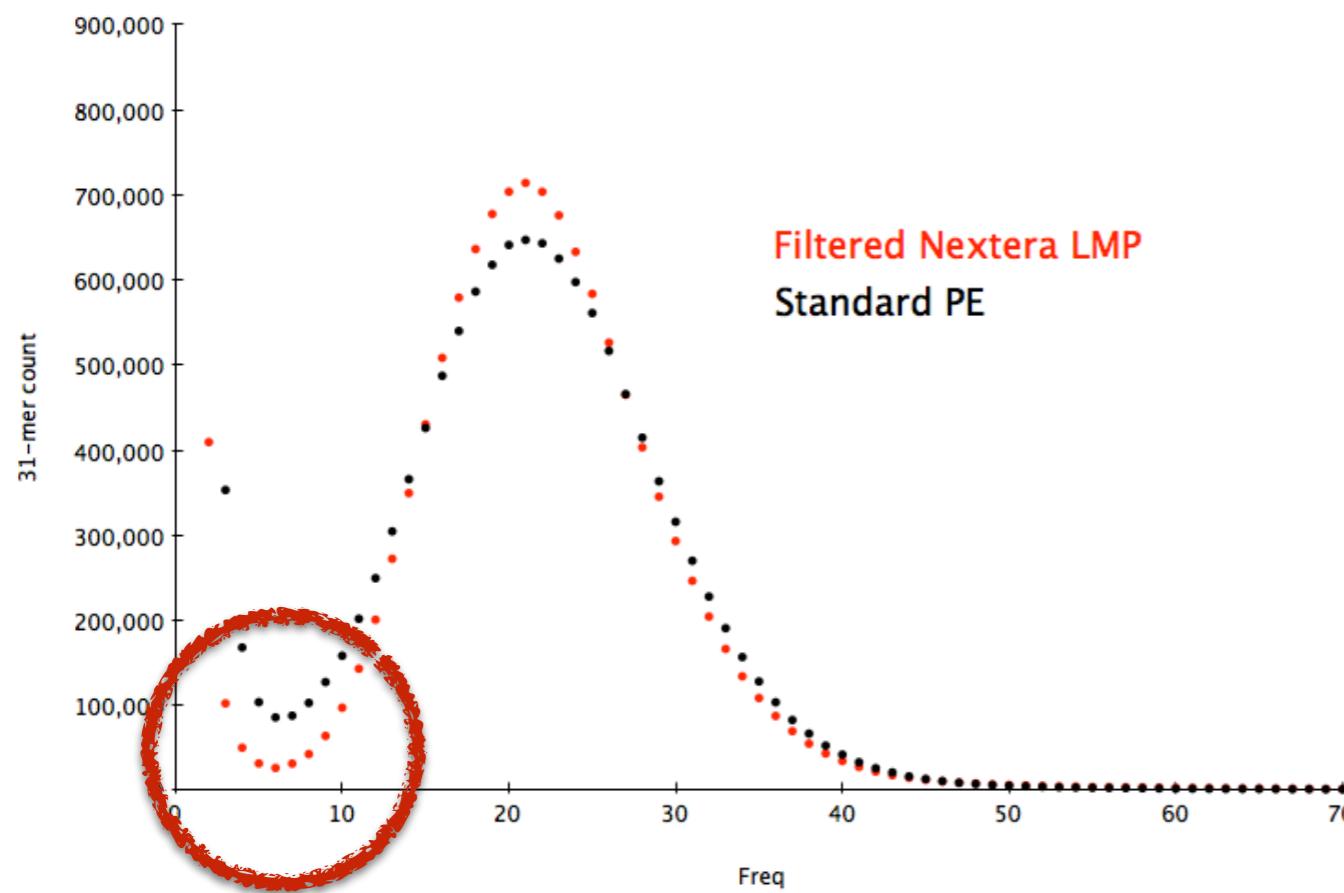
---

n	n:500	n:N50	min	N80	N50	N20	max	sum	
17783	7370	2171	500	744	1205	2024	6372	8031482	yeast_pe_k91-unitigs.fa
9030	4941	1079	500	1306	2708	5235	17385	9789489	yeast_pe_k91-contigs.fa
8810	4842	1037	500	1336	2753	5478	82405	9811356	yeast_pe_k91-scaffolds.fa
n	n:500	n:N50	min	N80	N50	N20	max	sum	
17783	7370	2171	500	744	1205	2024	6372	8031482	yeast_pe_raw_lmp1_k91-unitigs.fa
9030	4941	1079	500	1306	2708	5235	17385	9789489	yeast_pe_raw_lmp1_k91-contigs.fa
8571	4662	1024	500	1434	2898	5409	17385	9850977	yeast_pe_raw_lmp1_k91-scaffolds.fa
n	n:500	n:N50	min	N80	N50	N20	max	sum	
17783	7370	2171	500	744	1205	2024	6372	8031482	yeast_pe_clipped_lmp1_abc_k91-unitigs.fa
9030	4942	1079	500	1306	2708	5235	17385	9789862	yeast_pe_clipped_lmp1_abc_k91-contigs.fa
6907	2988	232	500	1967	10653	25702	74395	9843053	yeast_pe_clipped_lmp1_abc_k91-scaffolds.fa
n	n:500	n:N50	min	N80	N50	N20	max	sum	
2740	644	92	502	17019	36185	75693	144003	11.5e6	yeast_clipped_lmp1_abc_k91-unitigs.fa
1831	158	20	502	78404	188426	377063	431579	11.91e6	yeast_clipped_lmp1_abc_k91-contigs.fa
1759	104	12	502	135419	328003	648583	1026799	11.91e6	yeast_clipped_lmp1_abc_k91-scaffolds.fa

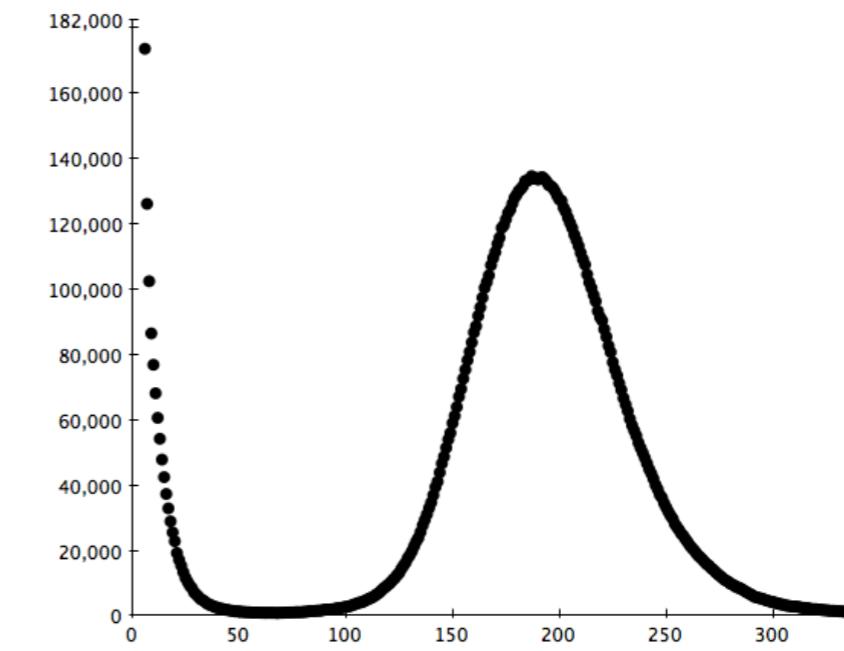
# What? Why?

---

- Unitigs are longer: better coverage\*, fewer errors (!)
- Contigs are longer: unitigs are longer, reads map better (!!)
- Scaffolds are longer: contigs are longer, reads map better (!!!)



Filtered Nextera LMP  
Standard PE



# Getting even more from **our** Nextera LMP

---

- R1 and R2 usually overlap:
  - Because we need to maximise chances of finding the linker
  - We can join the overlapping pairs but then we need to
- PCR-free, and very few errors (more later).
  - Kmer spectra is clean (after Nextclip!).
  - Error correction dream dataset!!!

# Fewer errors on Nextclip processed pairs?

Library	Strict			Standard				Relaxed			
	Total	Errors	Error %	Total	Errors	Error %	Error % (added)	Total	Errors	Error %	Error % (added)
LIB6470	2920076520	293879473	10.06	2960504161	304398051	10.28	26.02	2986028154	313188860	10.49	34.44
LIB6471	2975748804	198581925	6.67	3015112296	208350937	6.91	24.82	3040585503	216702885	7.13	32.79
LIB4659	2792344679	83943676	3.01	2814342003	88507030	3.14	20.75	2830348648	92654651	3.27	25.91
LIB5214	975451317	18809284	1.93	980158466	19687701	2.01	18.66	983874173	20564616	2.09	23.60
LIB5214b	2645386158	91401099	3.46	2677814948	97861650	3.65	19.92	2697174451	103550645	3.84	29.39



- Better linker match -> fewer errors.
- The worse reads won't match a linker.
- We can control strictness on matching.

# How would you ideally process a PAIR?

---

TGACTTGGAAAG**C**LLL**L**MM**M**TTTGGGAAACCC

CCCCAAAAAGGGGTTTCCCCAAAAGGGTT

ACCCTTGGGAACACCCCTTTGGGG

# How would you ideally process a PAIR?

---

TGACTTGGAAAGC~~LLL~~MMMTTTGGGAAACCC

CCCCCAAAAAGGGGTTTCCCCAAAAGGGTT

AACCCTTTGGGGAAAACCCCTTTGGGG  
TGACTTGGAGC~~LLL~~MMMTTTGGGAAACCC

# How would you ideally process a PAIR?

---

TGACTTGGAA**G**C**L****L****L****M****M**TTTGGGAAACCC

CCCCAAAAAGGGGTTTCCCCAAAAGGGTT

**TGACTTGGAGC****L****L****L****M****M**TTTGGGAAACCCTTTGGGGAAAACCCCTTTGGGG

# How would you ideally process a PAIR?

---

TGACTTGGAAAGC~~LLL~~MMMTTTGGGAAACCC

CCCCAAAAAGGGGTTTCCCCAAAAGGGTT

TGACTTGGAGC~~LLL~~MMMTTTGGGAAACCCTTTGGGGAAAACCCCTTTGGGG

TGACTTGGAGCC

TTTGGGAAACCCTTTGGGGAAAACCCCTTTGGGG  
CCCAAAAGGGTTTCCCAAGGGTTCCCAA

# How would you ideally process a PAIR?

---

TGACTTGGAA**G**C**LLL**M**MM**TTTGGGAAACCC

CCCCAAAAAGGGGTTTCCC**AAA**AGGGTT

TGACTTGG**A****G**C**LLL**M**MM**TTTGGGAAACC**CTT**TGGGGAAAACC**CTT**TTGGGG

TGACTTGG**A****G**CC

TTTGGGAAACC**CTT**TGGGGAAAACC**CTT**TTGGGG

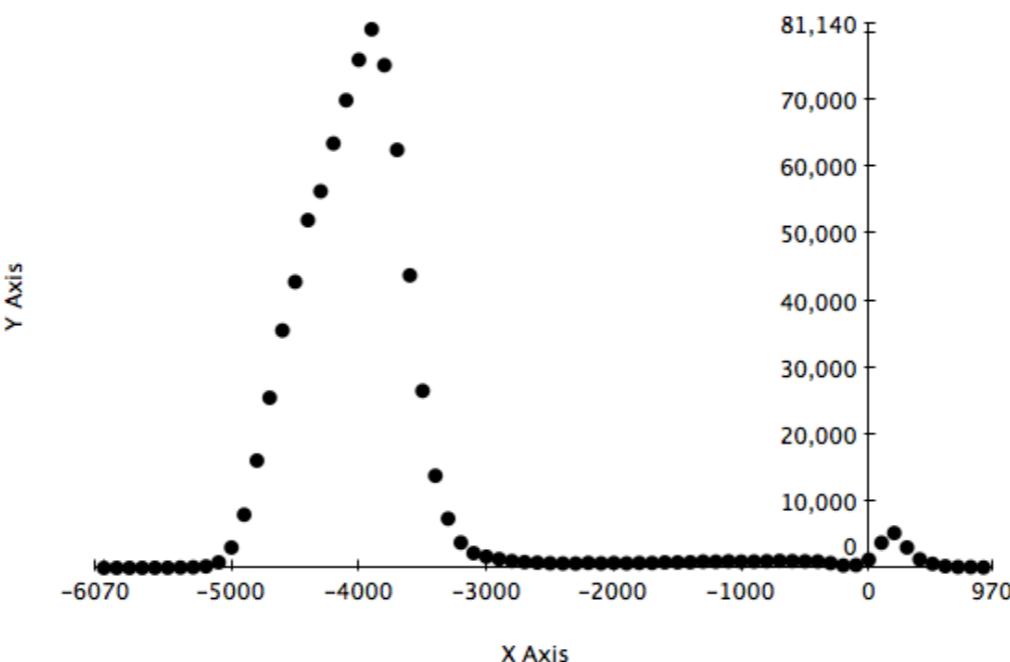
TGACTTGG**A****G**CC

CCCCAAAAAGGGGTTTCCC**AAA**AGGGTTCCC**AAA**

TGACTTGG**A****CC**

CCCCAAAAAGGGGTTTCCC**AAA**AGGGTTCCC**AAA**

# *A. thaliana* col-0 assembly



```
[clavijob@UVI 04-assembly]$ abyss-fac */*.scafSeq
n      n:500    n:N50    min     N80     N50     N20     max     sum
91954  3820     56     500   145714  504481  1225625  3219505  109.2e6 soap_4659/ath_4659.k71.scafSeq
88231  2206     26     500   232953  1090267  2540849  6126667  110.8e6 soap_5214b/ath_5214b.k71-101.scafSeq
110330  2189     28     500   233256  867462   2716797  6127137  110.8e6 soap_5214b/ath_5214b.k71.scafSeq
110260  2575     24     500   291558  1256637  3008384  6210307  112.7e6 soap_both/ath_both.k71-101.scafSeq
134053  2653     24     500   290498  1255489  3008402  6210370  112.6e6 soap_both/ath_both.k71.scafSeq
```

# Was it expected to work so well?

---

## De novo fragment assembly with short mate-paired reads: Does the read length matter?

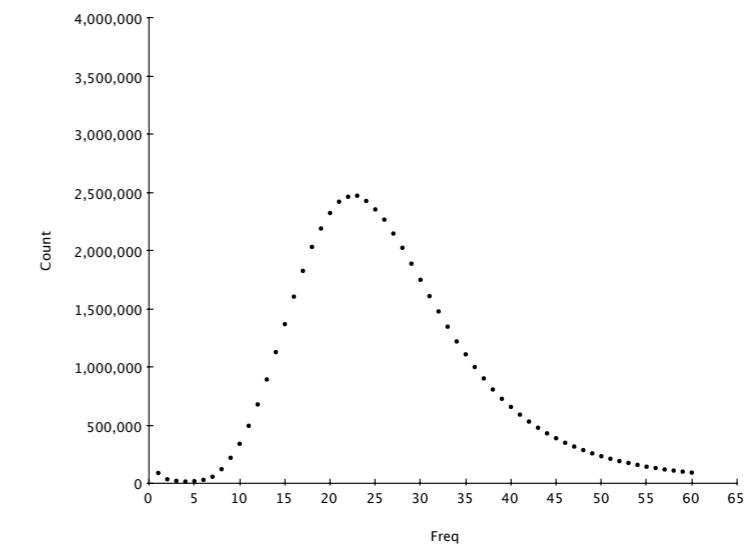
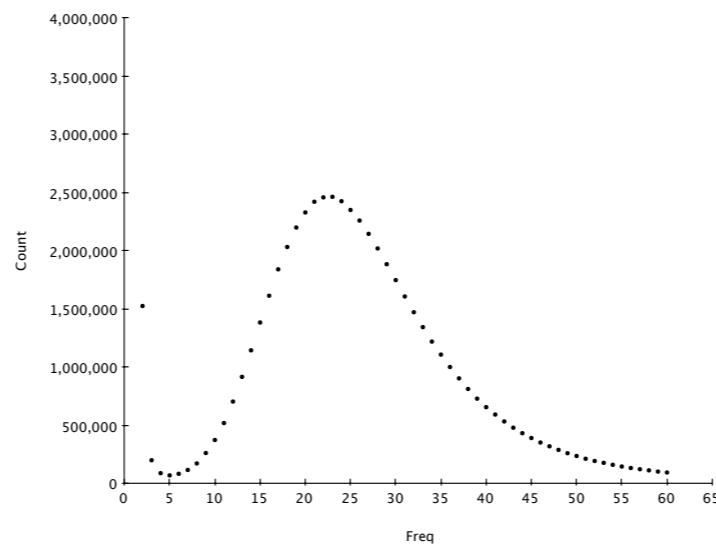
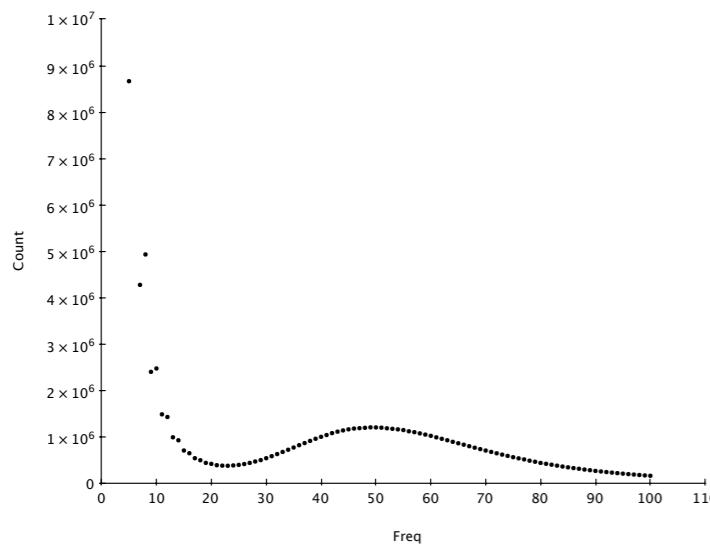
Mark J. Chaisson,<sup>1,3</sup> Dumitru Brinza,<sup>2</sup> and Pavel A. Pevzner<sup>2</sup>

<sup>1</sup>*Bioinformatics Program, University of California San Diego, La Jolla, California 92093, USA;* <sup>2</sup>*Department of Computer Science and Engineering, University of California San Diego, La Jolla, California 92093, USA*

Increasing read length is currently viewed as the crucial condition for fragment assembly with next-generation sequencing technologies. However, introducing mate-paired reads (separated by a gap of length,  $\text{GapLength}$ ) opens a possibility to transform short mate-pairs into long mate-reads of length  $\approx \text{GapLength}$ , and thus raises the question as to whether the read length (as opposed to  $\text{GapLength}$ ) even matters. We describe a new tool, EULER-USR, for assembling mate-paired short reads and use it to analyze the question of whether the read length matters. We further complement the ongoing experimental efforts to maximize read length by a new computational approach for increasing the effective read length. While the common practice is to trim the error-prone tails of the reads, we present an approach that substitutes trimming with error correction using repeat graphs. An important and counterintuitive implication of this result is that one may extend sequencing reactions that degrade with length “past their prime” to where the error rate grows above what is normally acceptable for fragment assembly.

# *C. fraxinea* assembly

n	n:200	n:N50	min	N80	N50	N20	max
5418	2583	306	200	27049	61813	115341	363023



```
[clavijob@UV1 soap_k111r_nok]$ abyss-fac cha_soap_ope_lmp_k111r.contig
n      n:500    n:N50   min      N80      N50      N20      max      sum
18093   6620    781     500     8359    22816    48113    130784   62.84e6 cha_soap_ope_lmp_k111r.contig
[clavijob@UV1 soap_k111r_nok]$ abyss-fac -s 10000 cha_soap_ope_lmp_k111r.scafSeq
n      n:10000  n:N50   min      N80      N50      N20      max      sum
11058    23      9     189181  2611331  2982393  4115367  4495304  62.08e6 cha_soap_ope_lmp_k111r.scafSeq
```

# Questions?

