

Session 1: k-mer histograms and genome profiling

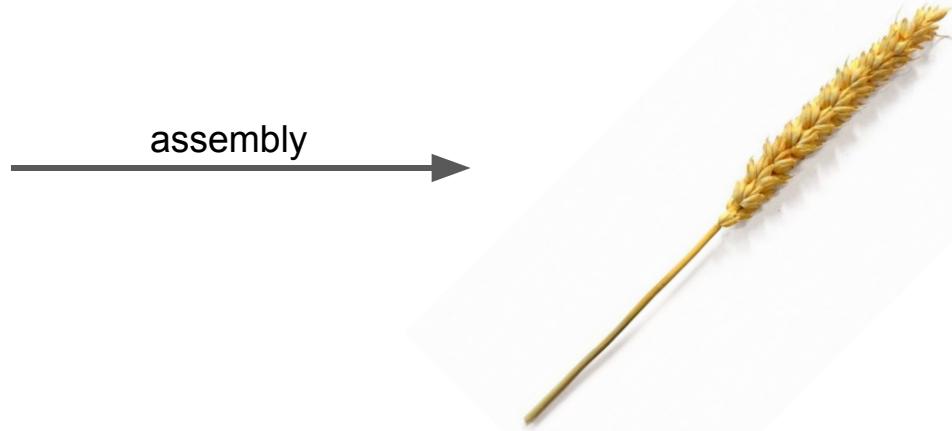


Kamil S. Jaron

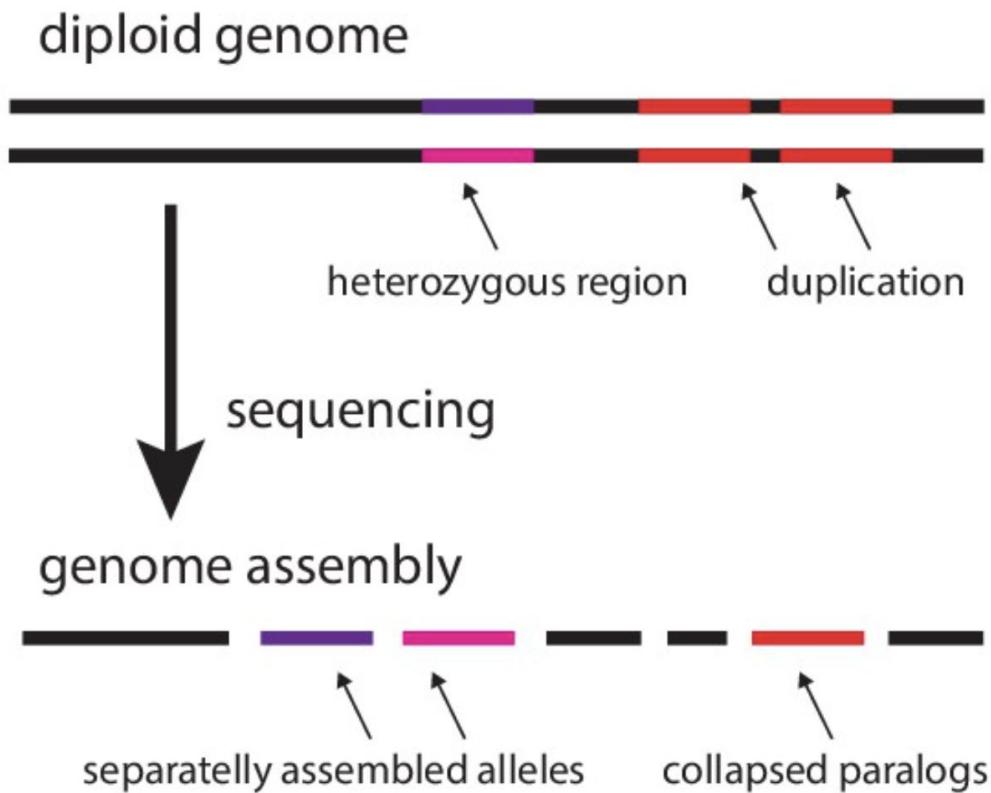
2nd June 2025

K-mer workshop for biodiversity genomics						
Time	Sunday	Monday	Tuesday	Wednesday	Thursday	
08:45 - 09:00						08:45 - 09:00
09:00 - 10:00						09:00 - 10:00
10:00 - 10:15						10:00 - 10:15
10:15 - 11:15		Session 1: Overview & Motivation of k-mer histograms and genome profiling		Hands-on for k-mer histograms and genome profiling		Work on k-mer histograms and genome profiling
11:15 - 11:30						11:15 - 11:30
11:30 - 12:30						11:30 - 12:30
12:30 - 12:45						12:30 - 12:45
12:45 - 13:45						12:45 - 13:45
13:45 - 14:00						13:45 - 14:00
14:00 - 15:00						14:00 - 15:00
15:00 - 15:15						15:00 - 15:15
15:15 - 16:15						15:15 - 16:15
16:15 - 16:30						16:15 - 16:30
16:30 - 17:30						16:30 - 17:30
17:30 - 17:45						17:30 - 17:45
17:45 - 18:45						17:45 - 18:45
18:45 - 18:55						18:45 - 18:55
18:55 - 19:00						18:55 - 19:00
19:00 - 19:15						19:00 - 19:15
19:15 - 19:30						19:15 - 19:30
19:30 - 19:45						19:30 - 19:45
19:45 - 20:00						19:45 - 20:00
20:00 - 20:15						20:00 - 20:15
20:15 - 21:15						20:15 - 21:15
21:15 - 21:30						21:15 - 21:30
21:30 - 21:45						21:30 - 21:45
21:45 - 22:00						21:45 - 22:00
22:00 - 22:15						22:00 - 22:15
22:15 - 22:30						22:15 - 22:30
22:30 - 22:45						22:30 - 22:45
22:45 - 22:55						22:45 - 22:55
22:55 - 23:00						22:55 - 23:00
23:00 - 23:15						23:00 - 23:15
23:15 - 23:30						23:15 - 23:30
23:30 - 23:45						23:30 - 23:45
23:45 - 23:55						23:45 - 23:55
23:55 - 24:00						23:55 - 24:00
24:00 - 24:15						24:00 - 24:15
24:15 - 24:30						24:15 - 24:30
24:30 - 24:45						24:30 - 24:45
24:45 - 24:55						24:45 - 24:55
24:55 - 25:00						24:55 - 25:00
25:00 - 25:15						25:00 - 25:15
25:15 - 25:30						25:15 - 25:30
25:30 - 25:45						25:30 - 25:45
25:45 - 25:55						25:45 - 25:55
25:55 - 26:00						25:55 - 26:00
26:00 - 26:15						26:00 - 26:15
26:15 - 26:30						26:15 - 26:30
26:30 - 26:45						26:30 - 26:45
26:45 - 26:55						26:45 - 26:55
26:55 - 27:00						26:55 - 27:00
27:00 - 27:15						27:00 - 27:15
27:15 - 27:30						27:15 - 27:30
27:30 - 27:45						27:30 - 27:45
27:45 - 27:55						27:45 - 27:55
27:55 - 28:00						27:55 - 28:00
28:00 - 28:15						28:00 - 28:15
28:15 - 28:30						28:15 - 28:30
28:30 - 28:45						28:30 - 28:45
28:45 - 28:55						28:45 - 28:55
28:55 - 29:00						28:55 - 29:00
29:00 - 29:15						29:00 - 29:15
29:15 - 29:30						29:15 - 29:30
29:30 - 29:45						29:30 - 29:45
29:45 - 29:55						29:45 - 29:55
29:55 - 30:00						29:55 - 30:00
30:00 - 30:15						30:00 - 30:15
30:15 - 30:30						30:15 - 30:30
30:30 - 30:45						30:30 - 30:45
30:45 - 30:55						30:45 - 30:55
30:55 - 31:00						30:55 - 31:00
31:00 - 31:15						31:00 - 31:15
31:15 - 31:30						31:15 - 31:30
31:30 - 31:45						31:30 - 31:45
31:45 - 31:55						31:45 - 31:55
31:55 - 32:00						31:55 - 32:00
32:00 - 32:15						32:00 - 32:15
32:15 - 32:30						32:15 - 32:30
32:30 - 32:45						32:30 - 32:45
32:45 - 32:55						32:45 - 32:55
32:55 - 33:00						32:55 - 33:00
33:00 - 33:15						33:00 - 33:15
33:15 - 33:30						33:15 - 33:30
33:30 - 33:45						33:30 - 33:45
33:45 - 33:55						33:45 - 33:55
33:55 - 34:00						33:55 - 34:00
34:00 - 34:15						34:00 - 34:15
34:15 - 34:30						34:15 - 34:30
34:30 - 34:45						34:30 - 34:45
34:45 - 34:55						34:45 - 34:55
34:55 - 35:00						34:55 - 35:00
35:00 - 35:15						35:00 - 35:15
35:15 - 35:30						35:15 - 35:30
35:30 - 35:45						35:30 - 35:45
35:45 - 35:55						35:45 - 35:55
35:55 - 36:00						35:55 - 36:00
36:00 - 36:15						36:00 - 36:15
36:15 - 36:30						36:15 - 36:30
36:30 - 36:45						36:30 - 36:45
36:45 - 36:55						36:45 - 36:55
36:55 - 37:00						36:55 - 37:00
37:00 - 37:15						37:00 - 37:15
37:15 - 37:30						37:15 - 37:30
37:30 - 37:45						37:30 - 37:45
37:45 - 37:55						37:45 - 37:55
37:55 - 38:00						37:55 - 38:00
38:00 - 38:15						38:00 - 38:15
38:15 - 38:30						38:15 - 38:30
38:30 - 38:45						38:30 - 38:45
38:45 - 38:55						38:45 - 38:55
38:55 - 39:00						38:55 - 39:00
39:00 - 39:15						39:00 - 39:15
39:15 - 39:30						39:15 - 39:30
39:30 - 39:45						39:30 - 39:45
39:45 - 39:55						39:45 - 39:55
39:55 - 40:00						39:55 - 40:00
40:00 - 40:15						40:00 - 40:15
40:15 - 40:30						40:15 - 40:30
40:30 - 40:45						40:30 - 40:45
40:45 - 40:55						40:45 - 40:55
40:55 - 41:00						40:55 - 41:00
41:00 - 41:15						41:00 - 41:15
41:15 - 41:30						41:15 - 41:30
41:30 - 41:45						41:30 - 41:45
41:45 - 41:55						41:45 - 41:55
41:55 - 42:00						41:55 - 42:00
42:00 - 42:15						42:00 - 42:15
42:15 - 42:30						42:15 - 42:30
42:30 - 42:45						42:30 - 42:45
42:45 - 42:55						42:45 - 42:55
42:55 - 43:00						42:55 - 43:00
43:00 - 43:15						43:00 - 43:15
43:15 - 43:30						43:15 - 43:30
43:30 - 43:45						43:30 - 43:45
43:45 - 43:55						43:45 - 43:55
43:55 - 44:00						43:55 - 44:00
44:00 - 44:15						44:00 - 44:15
44:15 - 44:30						44:15 - 44:30
44:30 - 44:45						44:30 - 44:45
44:45 - 44:55						44:45 - 44:55
44:55 - 45:00						44:55 - 45:00
45:00 - 45:15						45:00 - 45:15
45:15 - 45:30						45:15 - 45:30
45:30 - 45:45						45:30 - 45:45
45:45 - 45:55						45:45 - 45:55
45:55 - 46:00						45:55 - 46:00
46:00 - 46:15						46:00 - 46:15
46:15 - 46:30						46:15 - 46:30
46:30 - 46:45						46:30 - 46:45
46:45 - 46:55						46:45 - 46:55
46:55 - 47:00						46:55 - 47:00
47:00 - 47:15						47:00 - 47:15
47:15 - 47:30						47:15 - 47:30
47:30 - 47:45						47:30 - 47:45
47:45 - 47:55						47:45 - 47:55
47:55 - 48:00						47:55 - 48:00
48:00 - 48:15						48:00 - 48:15
48:15 - 48:30						48:15 - 48:30
48:30 - 48:45						48:30 - 48:45
48:45 - 48:55						48:45 - 48:55
48:55 - 49:00						48:55 - 49:00
49:00 - 49:15						49:00 - 49:15
49:15 - 49:30						49:15 - 49:30
49:30 - 49:45						49:30 - 49:45
49:45 - 49:55						49:45 - 49:55
49:55 - 50:00						49:55 - 50:00
50:00 - 50:15						50:00 - 50:15
50:15 - 50:30						50:15 - 50:30
50:30 - 50:45						50:30 - 50:45
50:45 - 50:55						50:45 - 50:55
50:55 - 51:00						50:55 - 51:00
51:00 - 51:15						51:00 - 51:15
51:15 - 51:30						51:15 - 51:30
51:30 - 51:45						51:30 - 51:45
51:45 - 51:55						51:45 - 51:55
51:55 - 52:00						51:55 - 52:00
52:00 - 52:15						52:00 - 52:15
52:15 - 52:30						52:15 - 52:30
52:30 - 52:45						52:30 - 52:45
52:45 - 52:55						52:45 - 52:55
52:55 - 53:00						52:55 - 53:00
53:00 - 53:15						53:00 - 53:15
53:15 - 53:30						53:15 - 53:30
53:30 - 53:45						53:30 - 53:45
53:45 - 53:55						53:45 - 53:55
53:55 - 54:00						53:55 - 54:00
54:00 - 54:15						54:00 - 54:15
54:15 - 54:30						54:15 - 54:30
54:30 - 54:45						54:30 - 54:45
54:45 - 54:55						54:45 - 54:55
54:55 - 55:00						54:55 - 55:00
55:00 - 55:15						55:00 - 55:15
55:15 - 55:30						55:15 - 55:30
55:30 - 55:45						55:30 - 55:45
55:45 - 55:55						55:45 - 55:55
55:55 - 56:00						55:55 - 56:00
56:00 - 56:15						56:00 - 56:15
56:15 - 56:30						56:15 - 56:30
56:30 - 56:45						56:30 - 56:45
56:45 - 56:55						56:45 - 56:55
56:55 - 57:00						

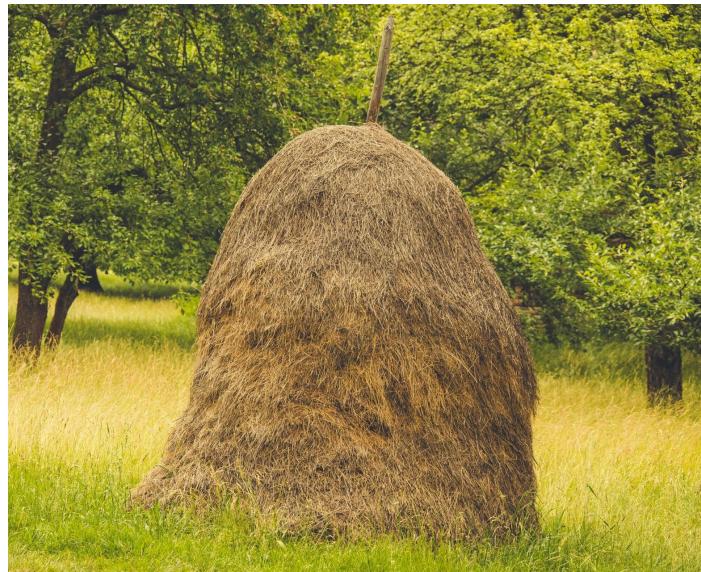
Genome sequencing



Assembly problems



Genome sequencing



assembly →

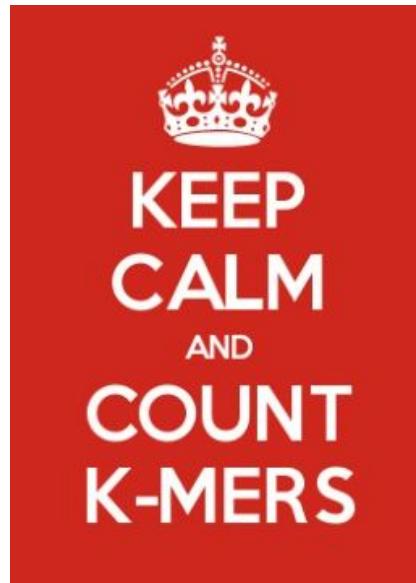


genome profiling →



13e6 individual plants in the harvest;
each weighted 843g and had 234 kernels on average

Part 1: Understanding k -mer histograms





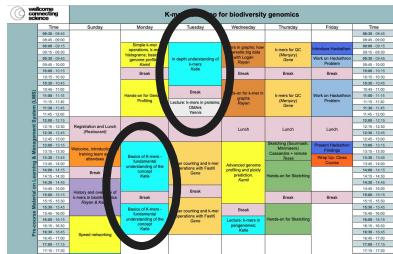
for all reads in the bag...

ATCTAAACGATCGATCGATCGA
ATCTAAA
TCTAAC
CTAAACG
TAAACGA

read

kmers
($k = 7$)

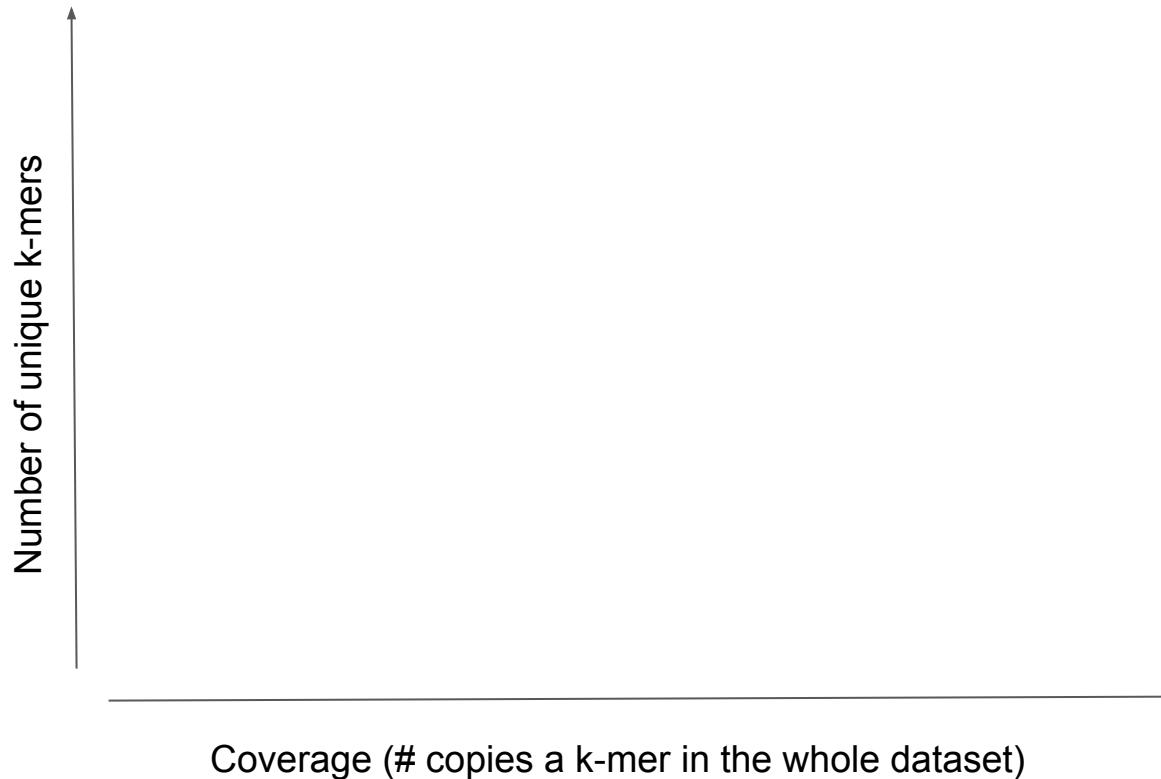
...



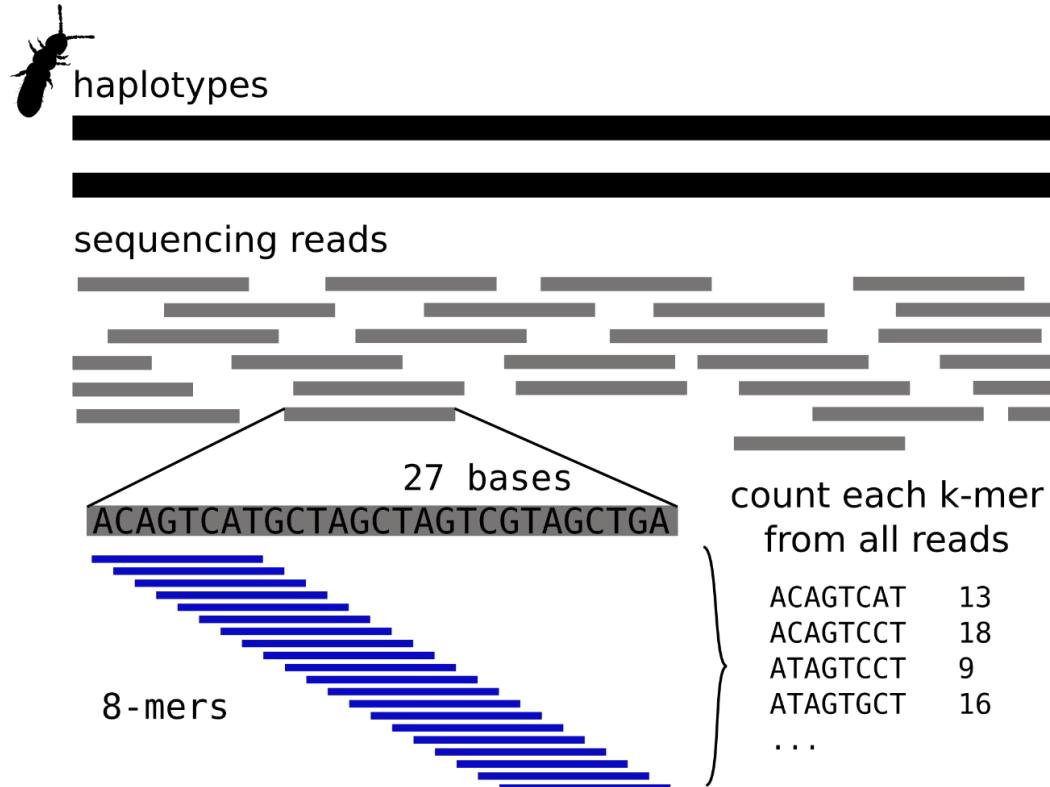
Counting frequency of all k-mers in all sequencing reads

kmer	coverage
AAAAAAAAAAACAAACGT	28
AAAAAATAAACACAAACGT	31
AAAAAATAAACACAAACGG	1
AAAAAATAACGCAACGT	47
AAAATTTACGCAACGT	2
AAAATTTACGCAACGA	17
...	...

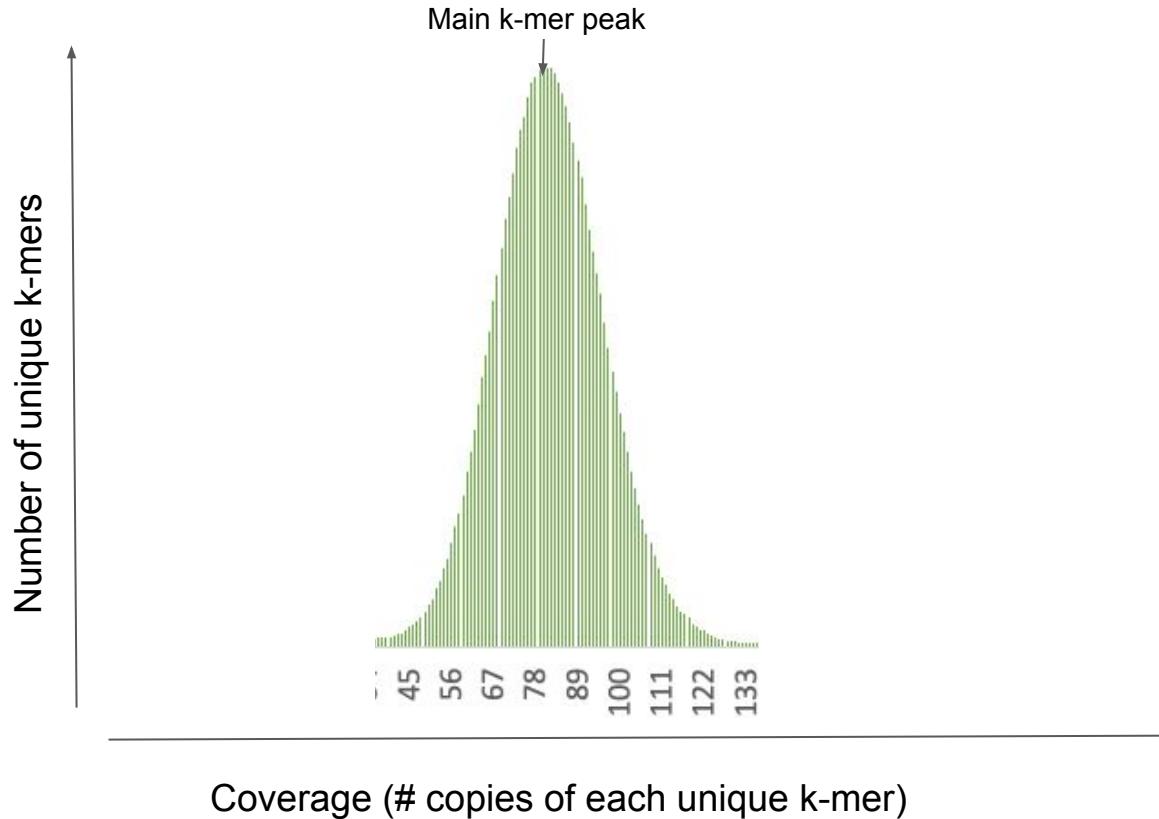
Plotting a k-mer histogram



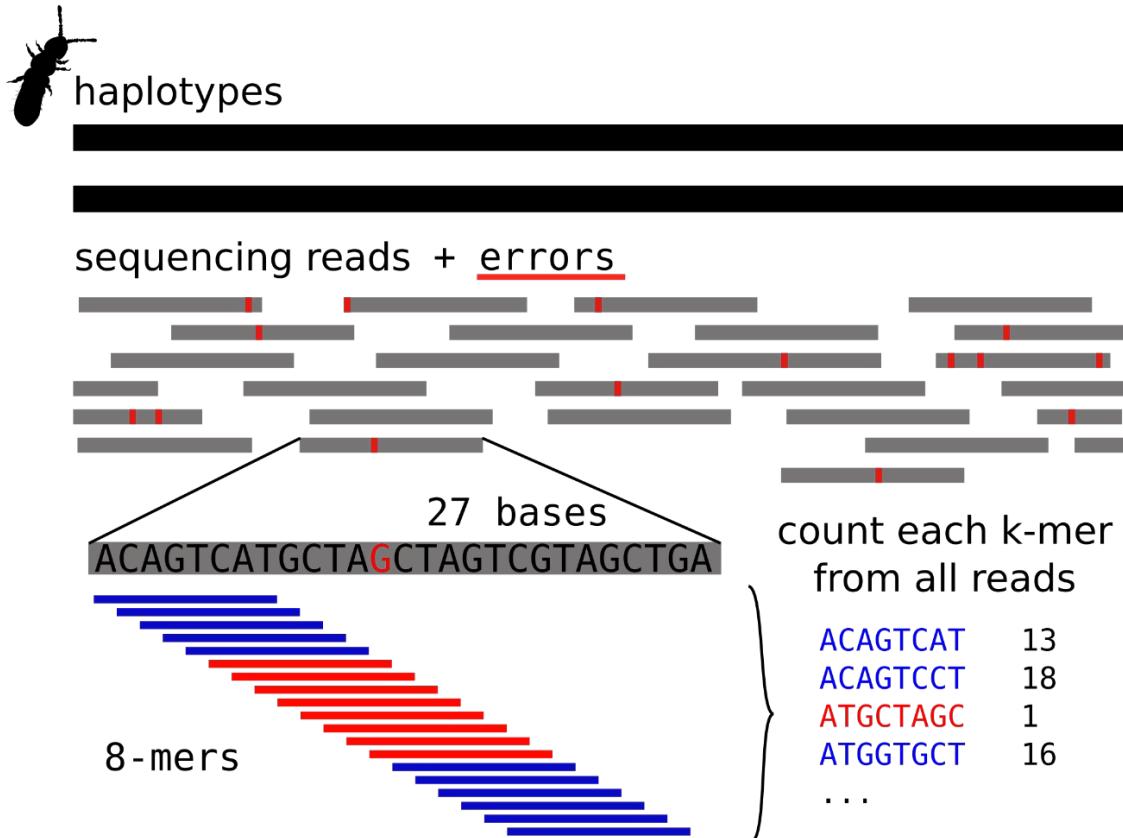
Plotting a perfect k-mer histogram



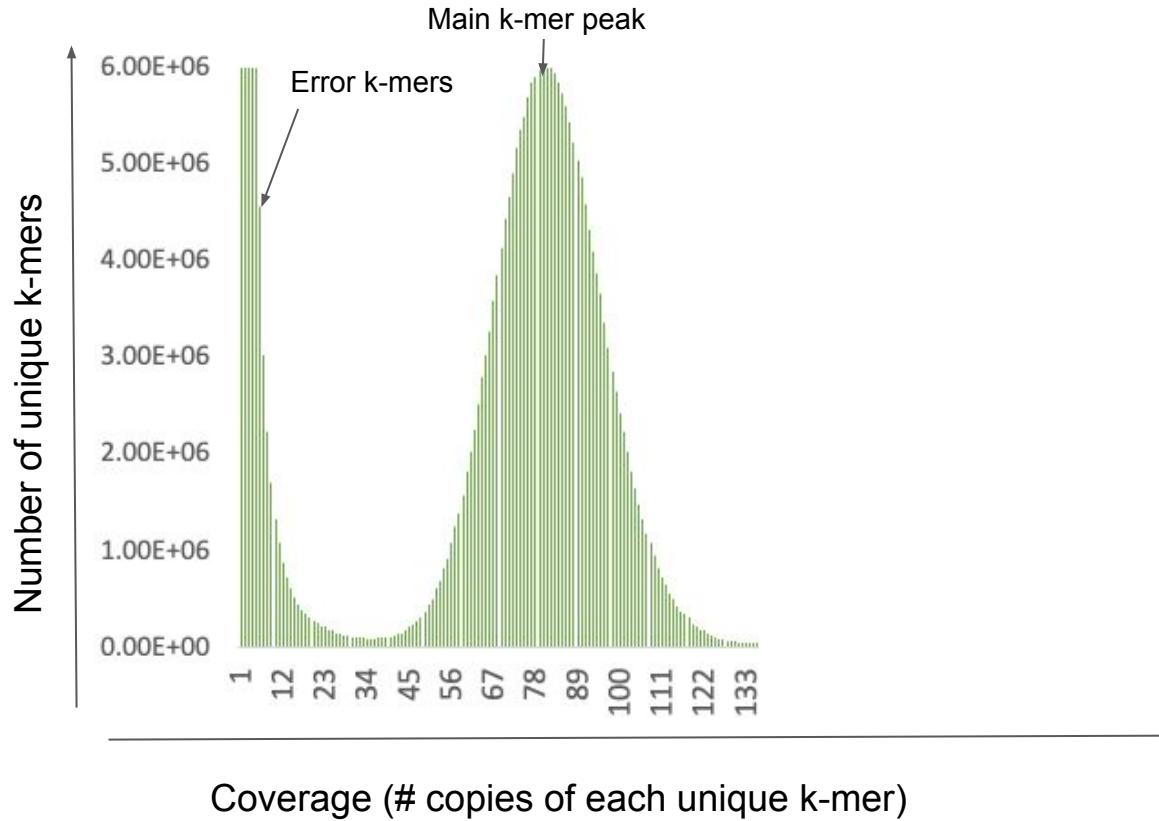
The perfect k-mer histogram



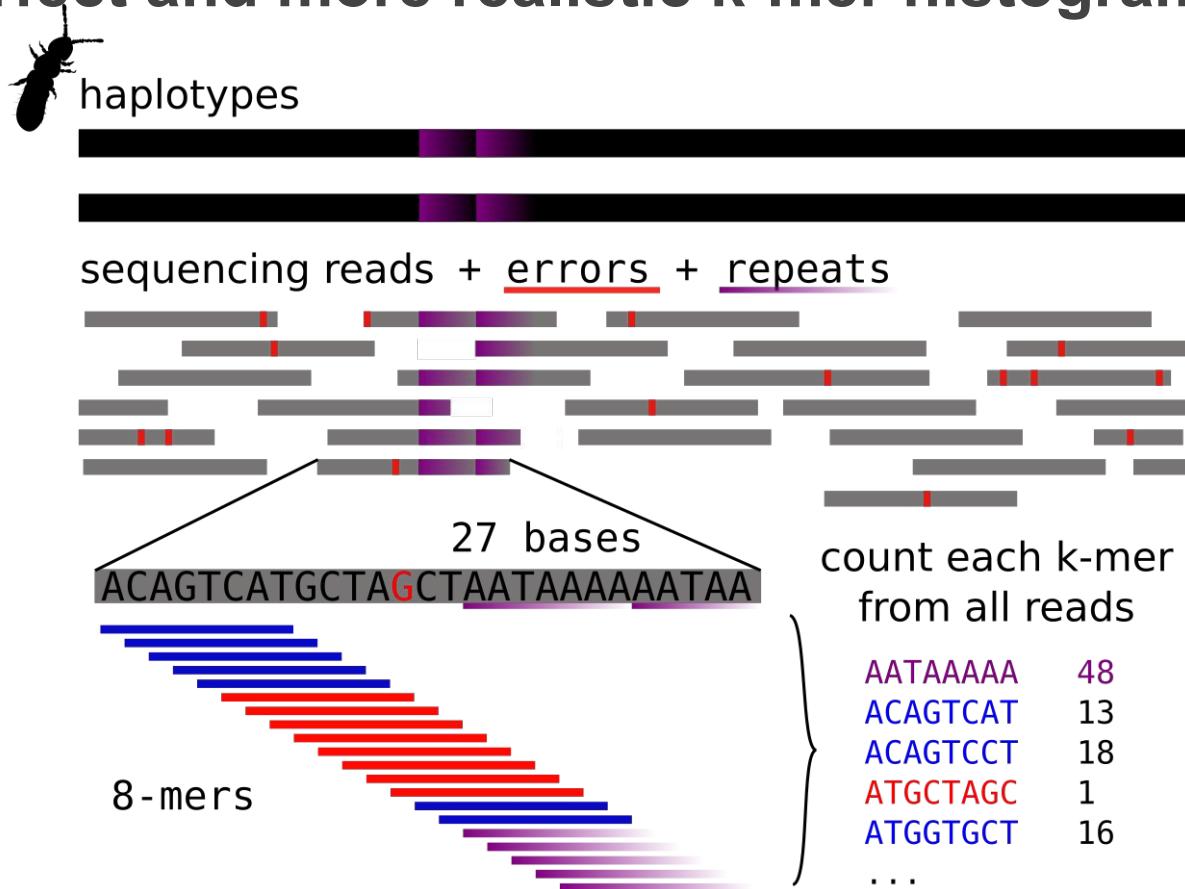
Creating an imperfect k-mer histogram



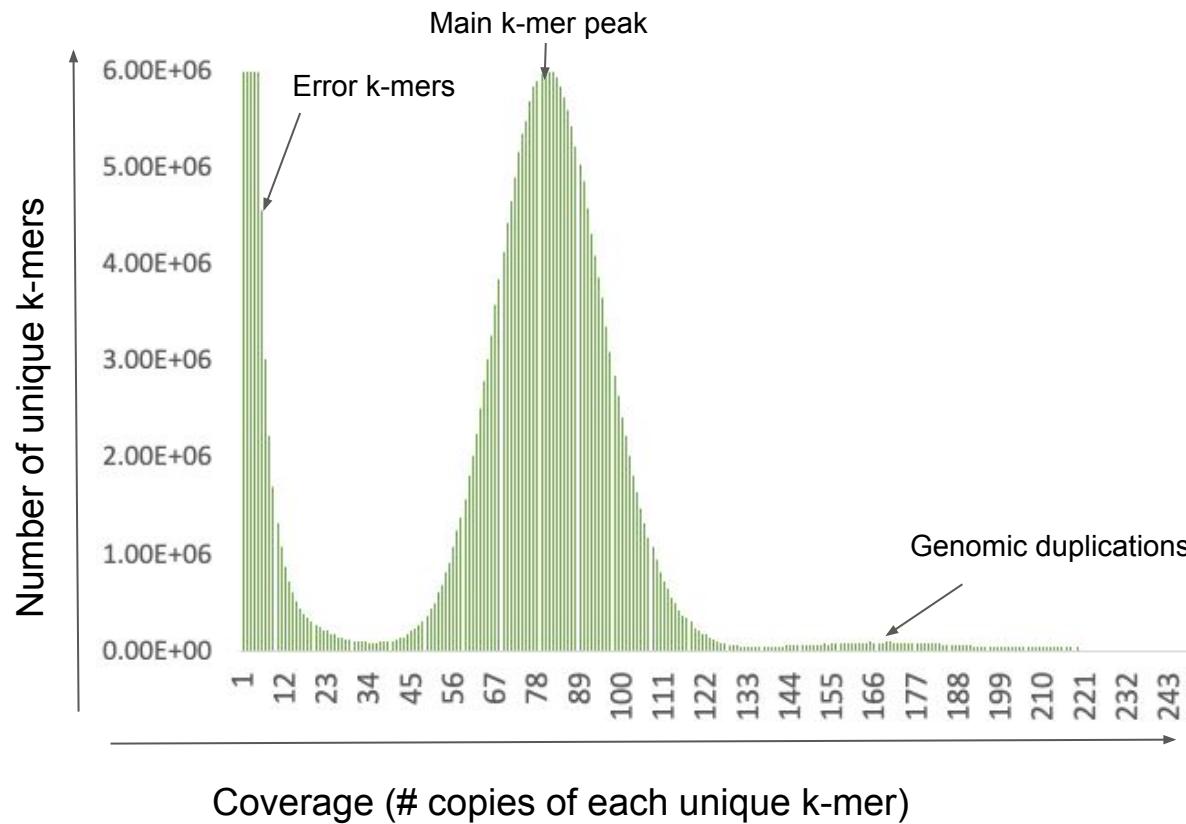
An imperfect histogram



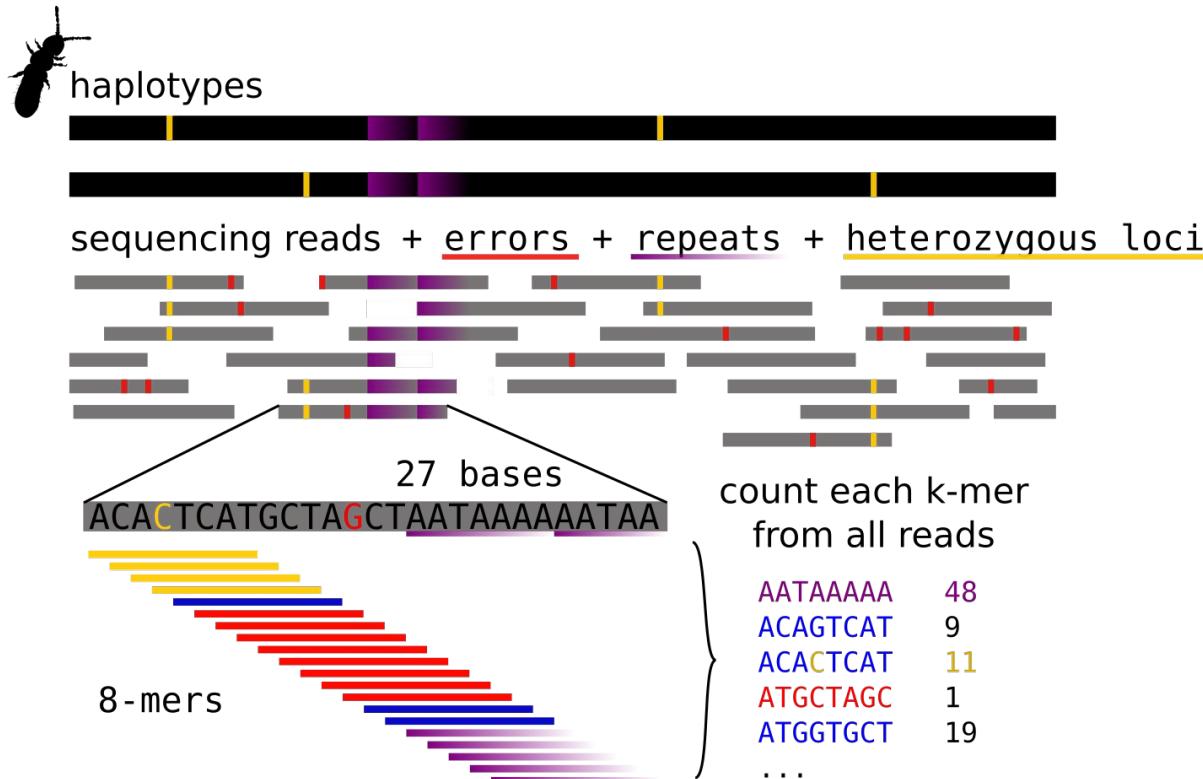
An imperfect and more realistic k-mer histogram



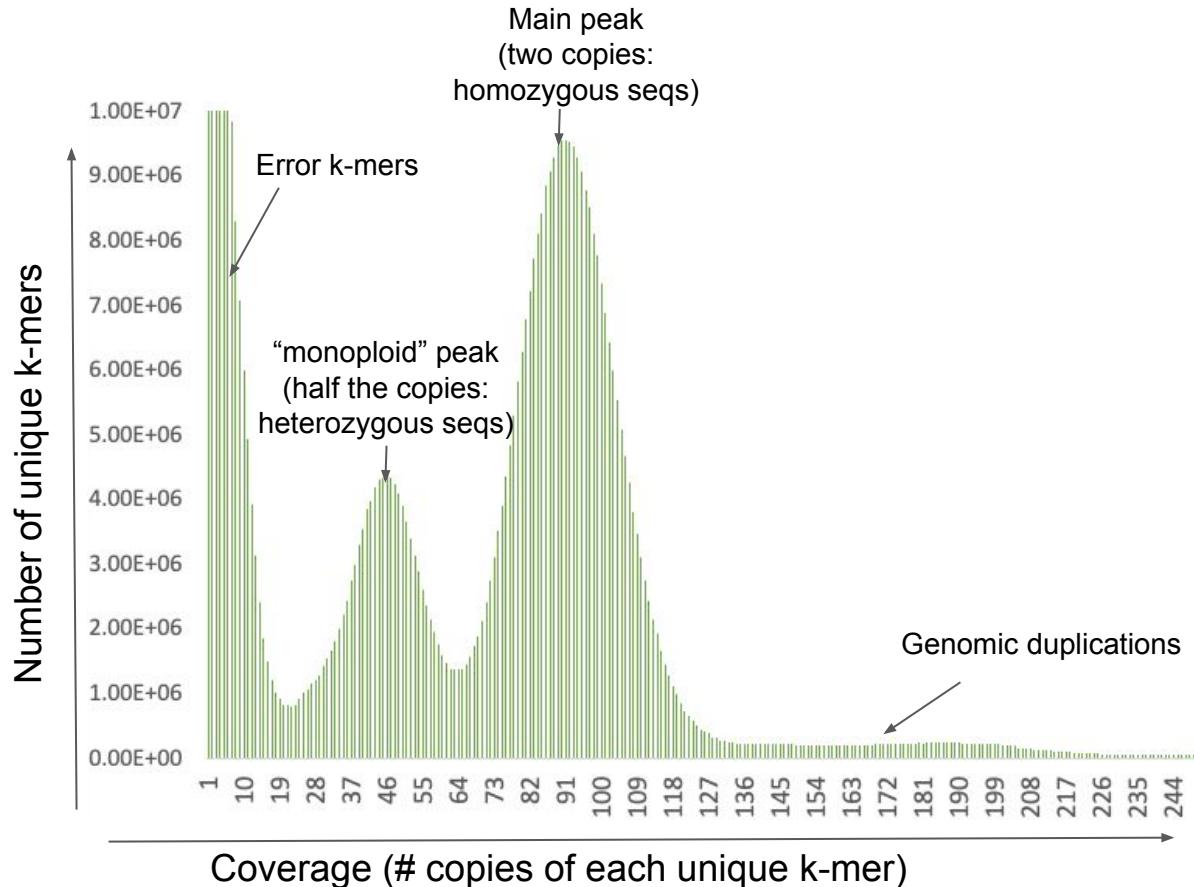
An imperfect and more realistic k-mer histogram



Counting k-mers...

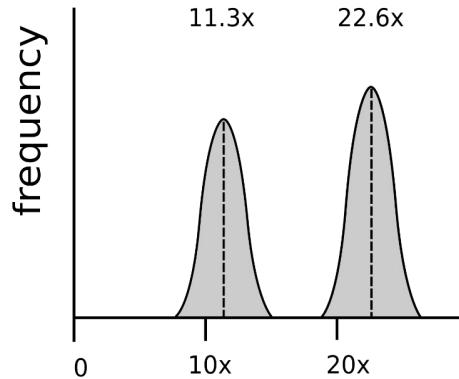


A realistic imperfect and non-inbred k-mer histogram

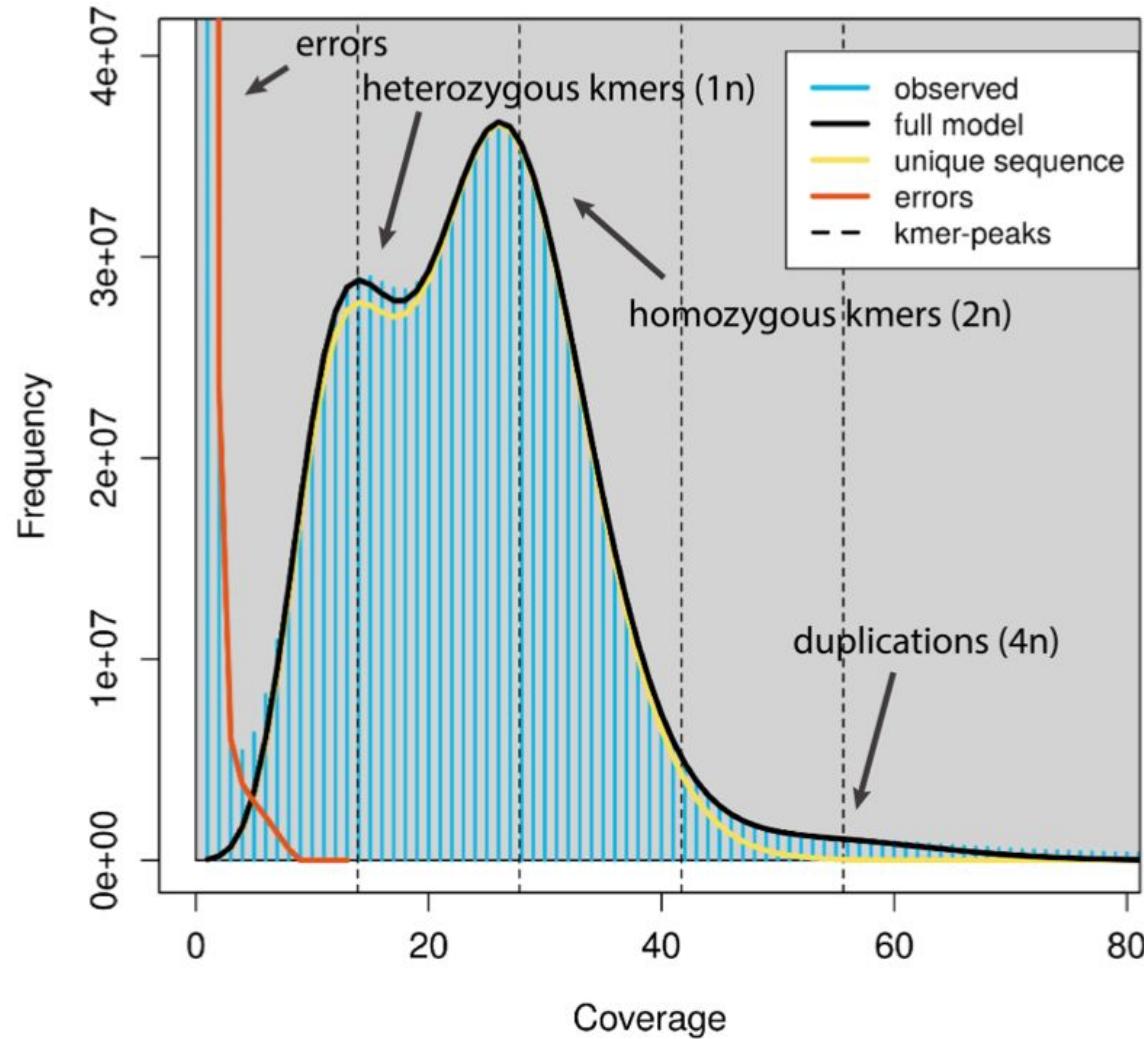


The principle of stoichiometry

The k -mers that occur in a single copy in the genome (e.g. heterozygous in diploid) will have on average $\frac{1}{2}$ coverage of the the k -mers that are present in two copies (homozygous in diploid)



There are millions of k -mers in each datasets, the deviations are not statistical but technical or biological!!!



Common bream

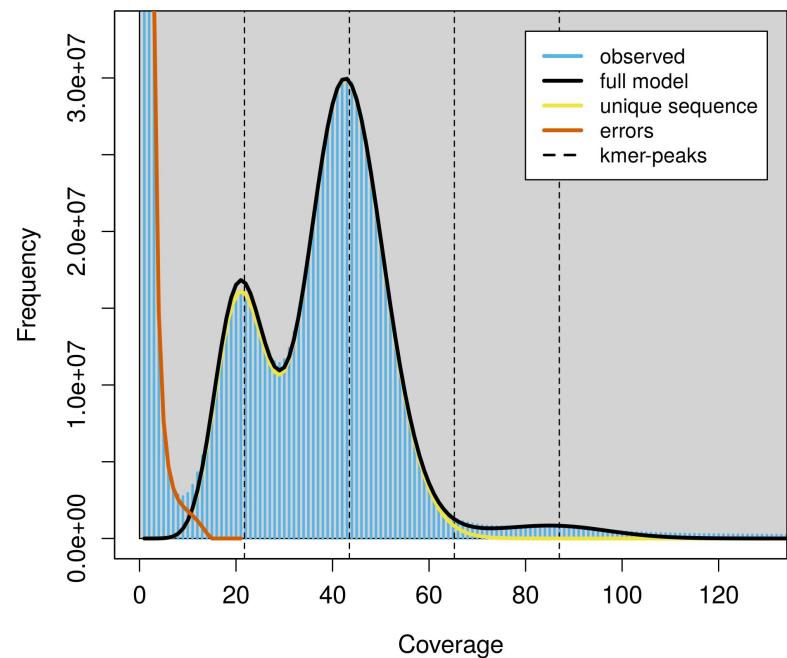
Do we have enough coverage?
Did sequencing run go well?

Rule of thumb - do you see clear genomic peak(s)?

Yes: Probably

Kind of: Possibly

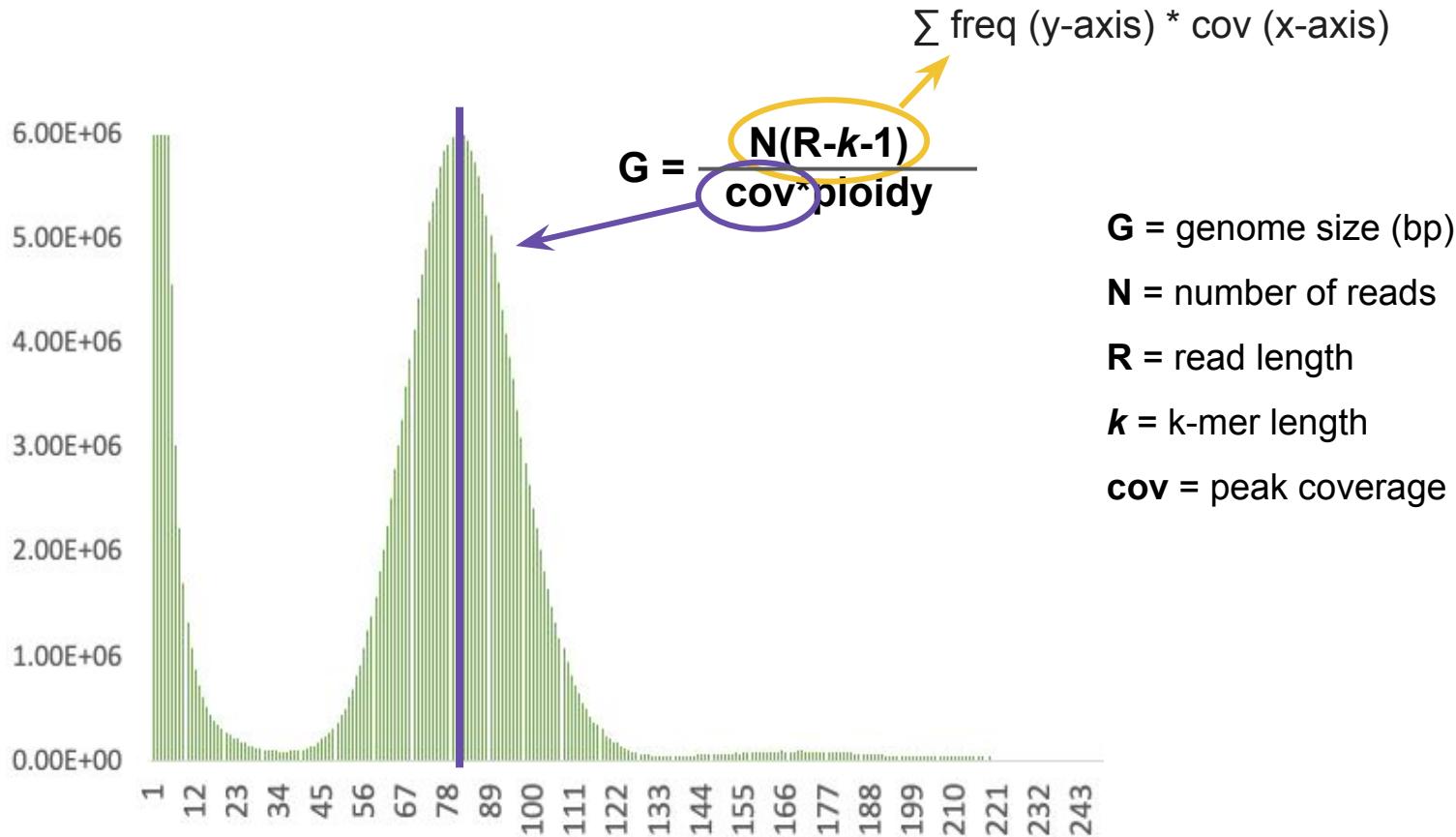
No: Most likely no!



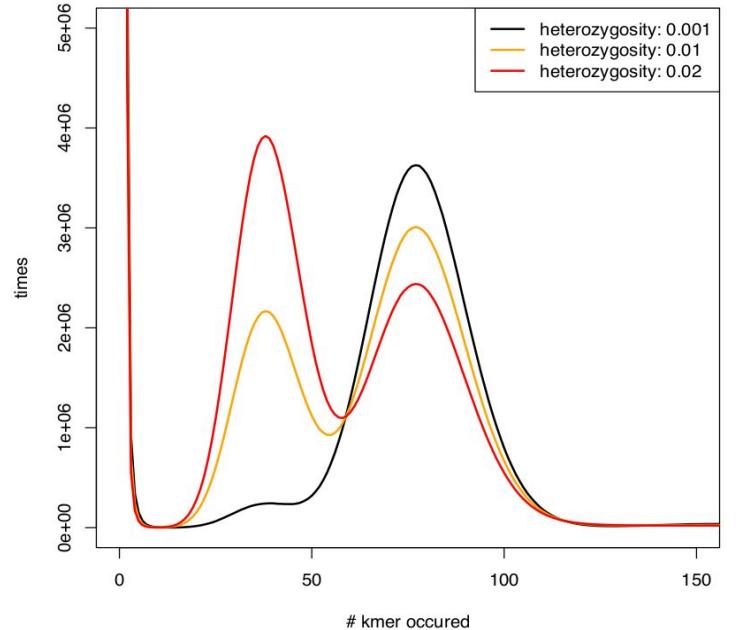
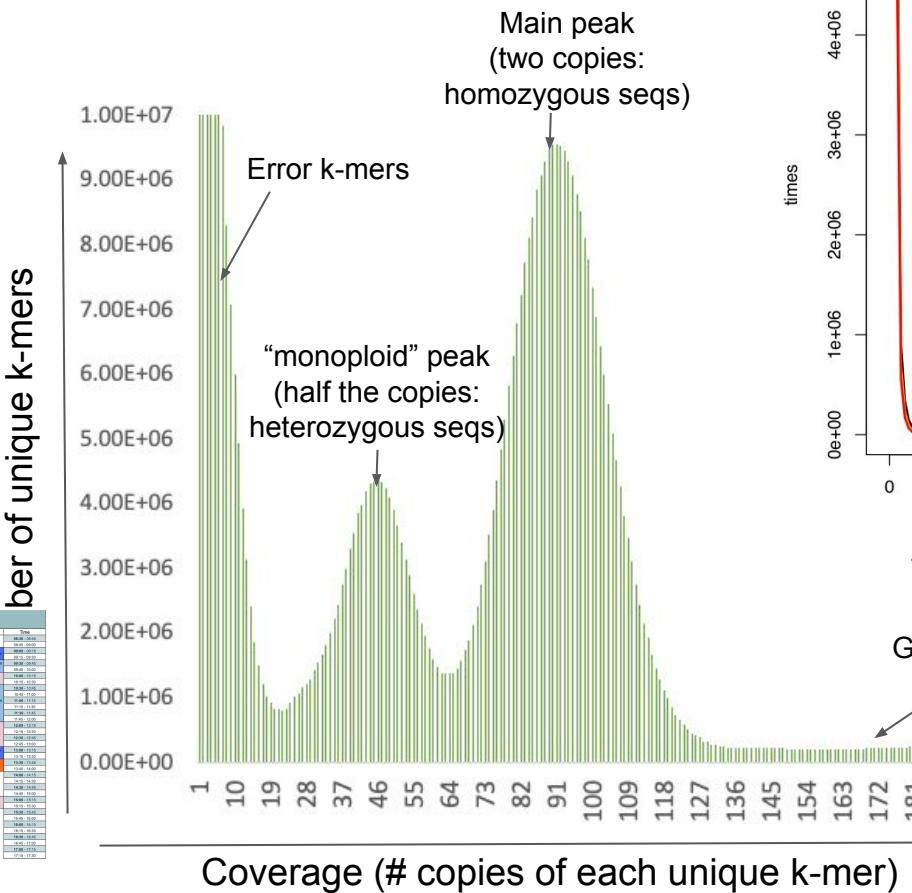
Part 2: Genome profiling



Modeling genome size



Modeling heterozygosity



from Supplementary materials of Vulture et al. 2017

Genomic duplications

For sane high coverage genomes spectra look good!

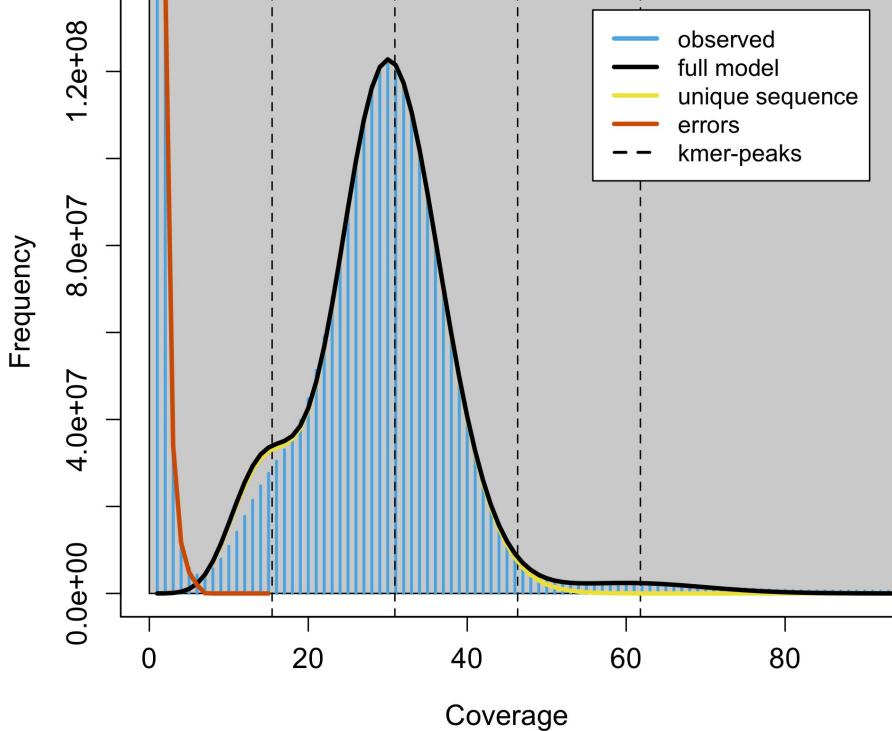


GenomeScope Profile

len:2,873,667,792bp uniq:72.3%

aa:99.6% ab:0.392%

kcov:15.4 err:0.112% dup:0.271 k:21 p:2

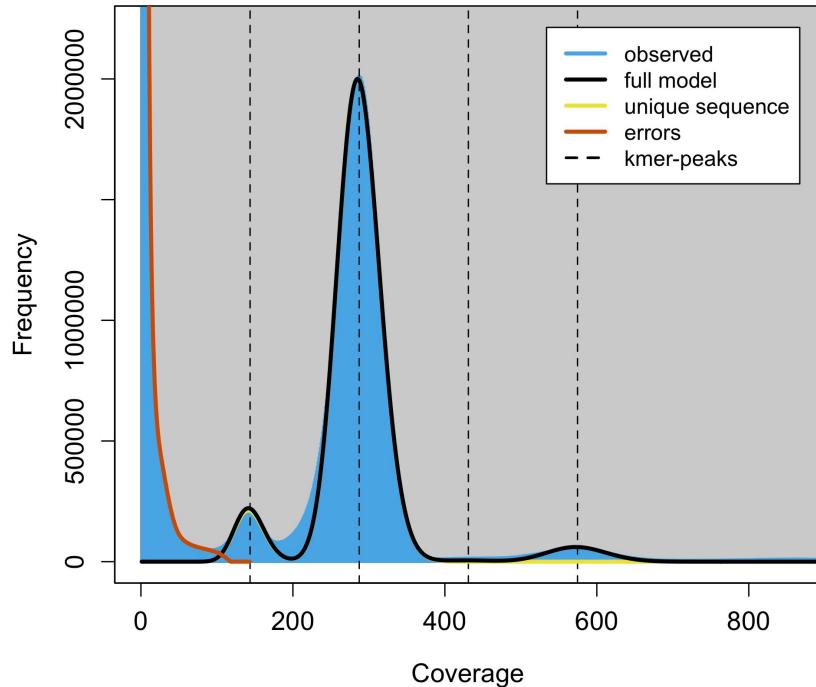


GenomeScope Profile

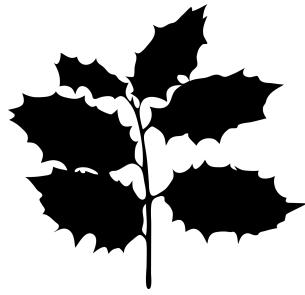
len:211,383,532bp uniq:68.5%

aa:99.8% ab:0.175%

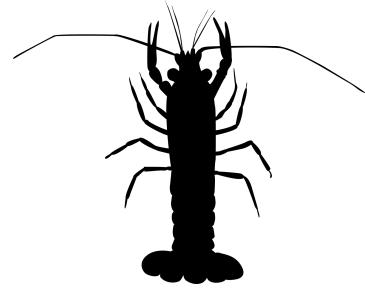
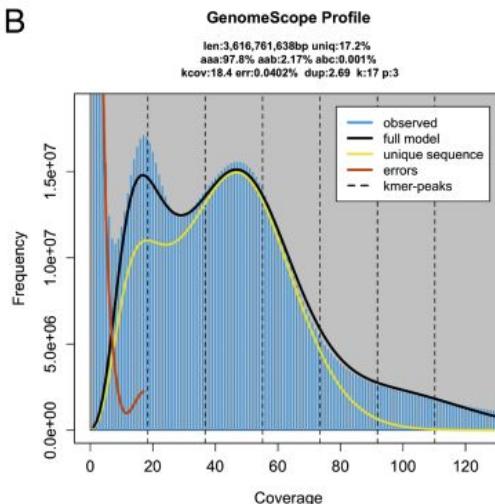
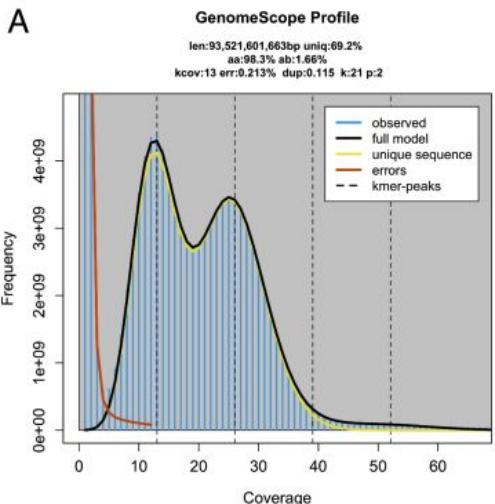
kcov:144 err:0.364% dup:1.71 k:21 p:2



Coverage patterns



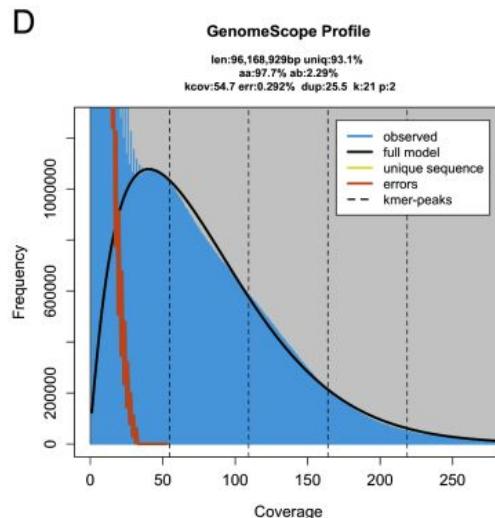
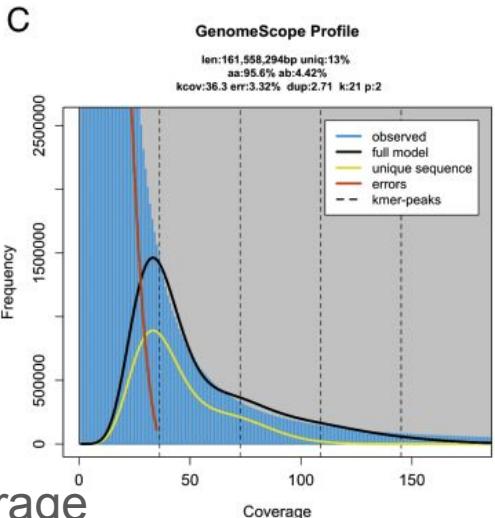
Great



Perhaps ok?



Not enough coverage



Contamination

Contamination is one and the only exception of possible good run without a clear coverage peak

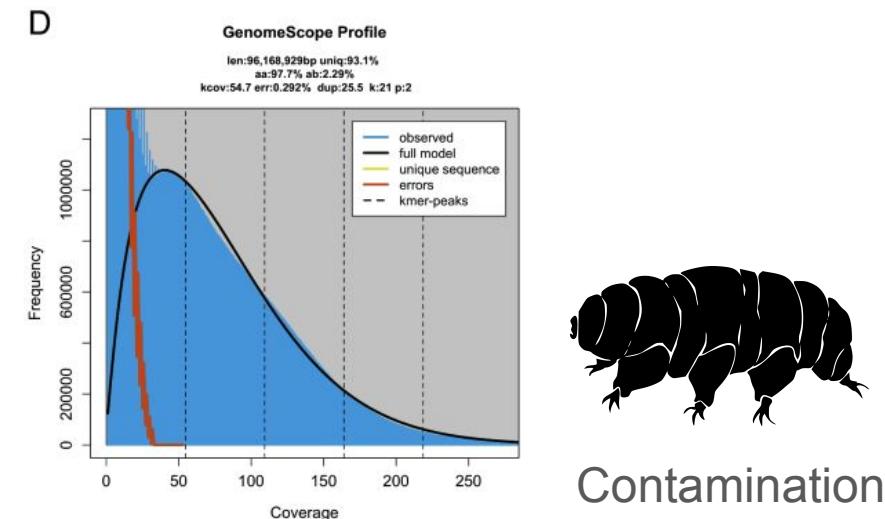
Would apply only for samples where very high levels of contamination can be expected... sequencing guts, environmental or otherwise messy samples...

Rule of thumb - do you see clear genomic peak(s)?

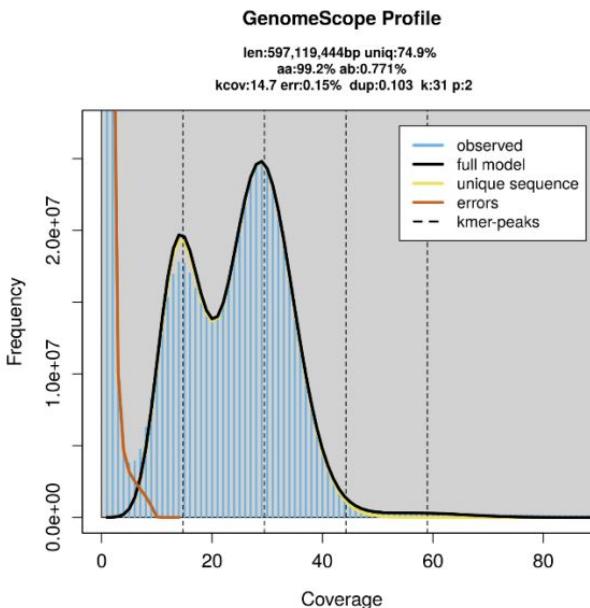
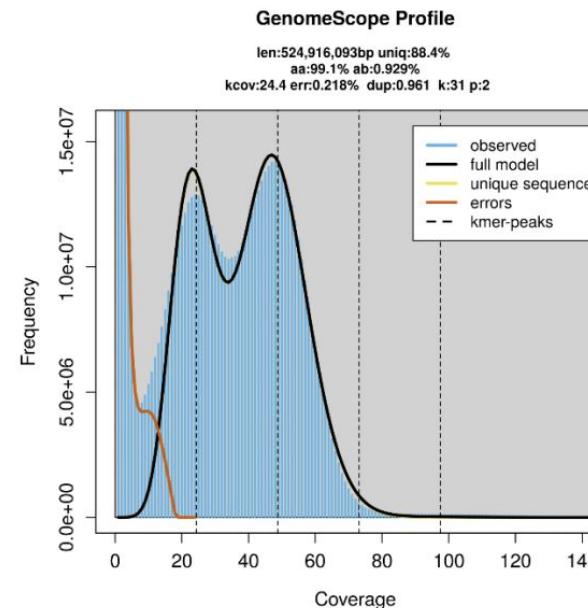
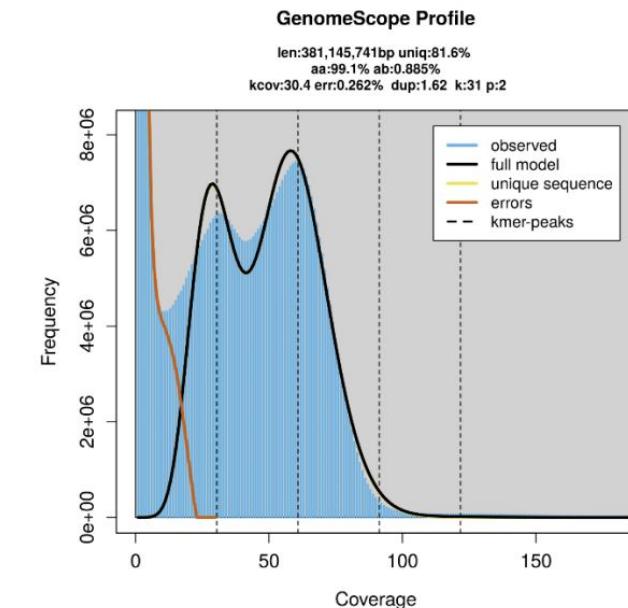
No? Is it a tiny messy sample, is the genome size in the right ballpark?

Yes - Might be fine

No - Low coverage!

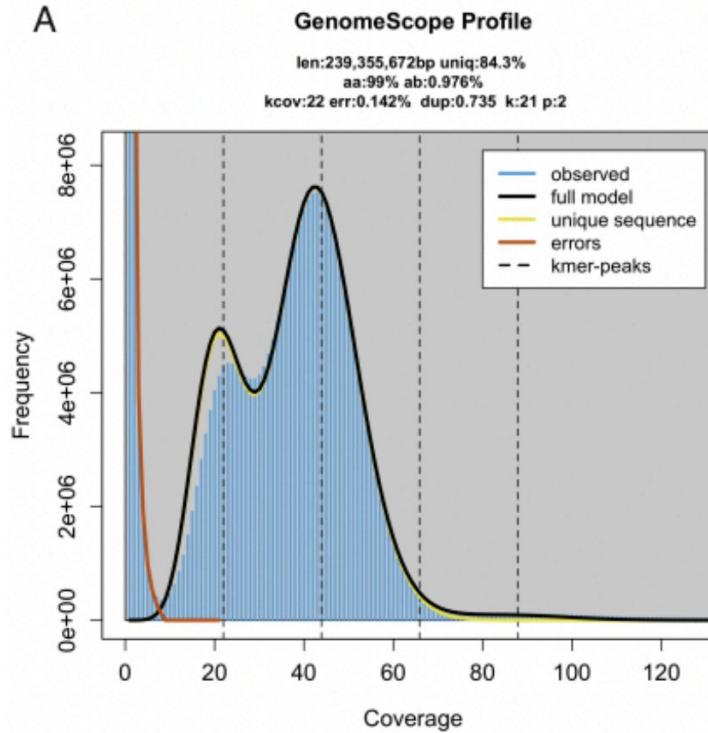


Sequencing biases can make “coverage bridge” (GA-rich regions)

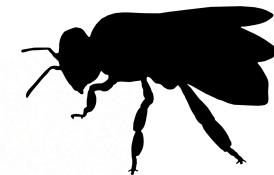
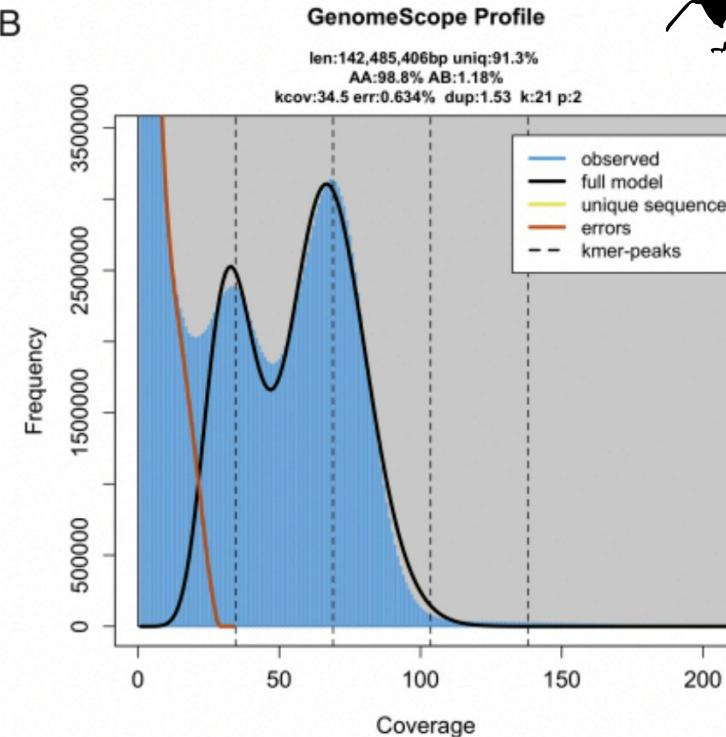
**A. stone loach****B. European flounder****C. nine-spined stickleback**

“Coverage bridge” can be due to poor sample handling

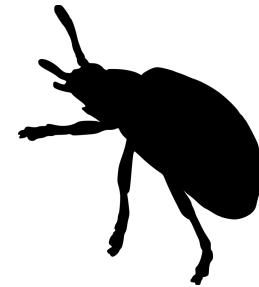
A



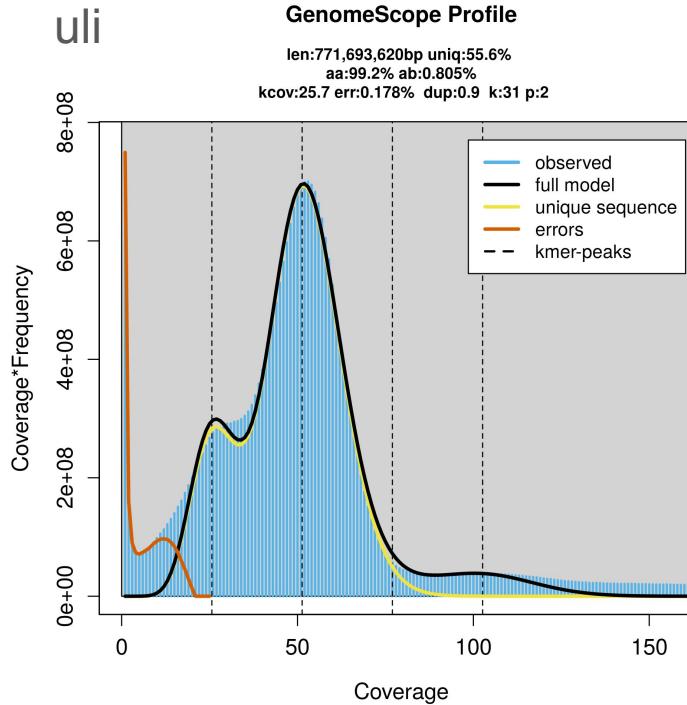
B



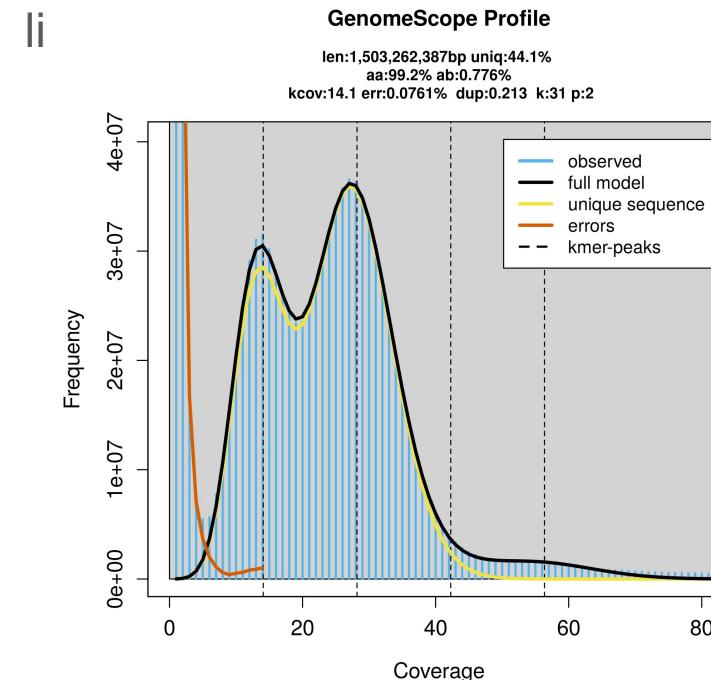
“Coverage bridge” can be due amplification



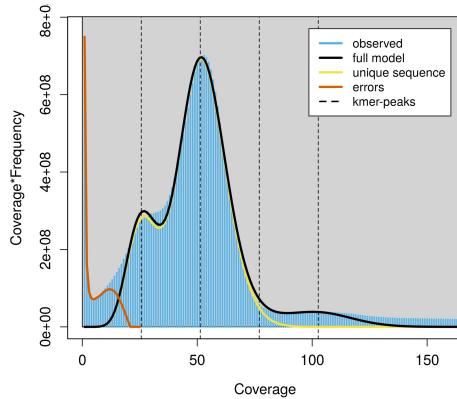
Chrysolina herbacea



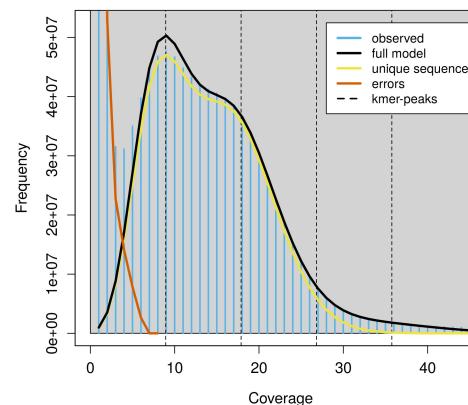
Chrysolina graminis



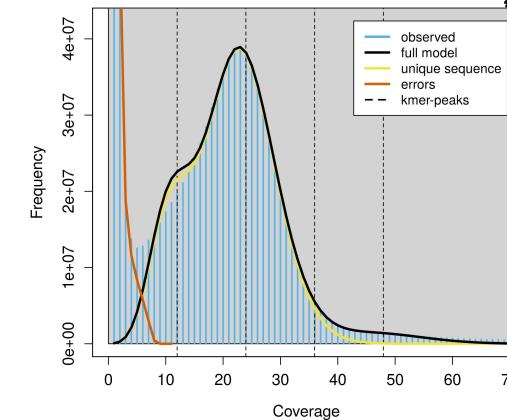
Example by Ksenia Krasheninnikova

Chrysolina herbacea

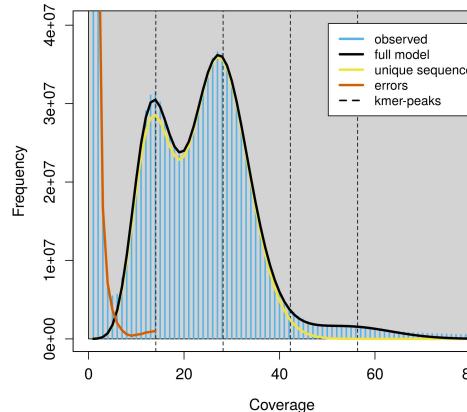
uli

Chrysolina graminis

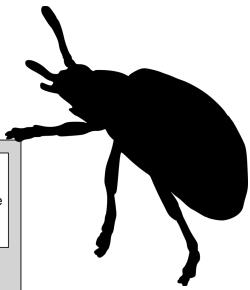
li

Chrysolina oricalcia

All bridges



Ok



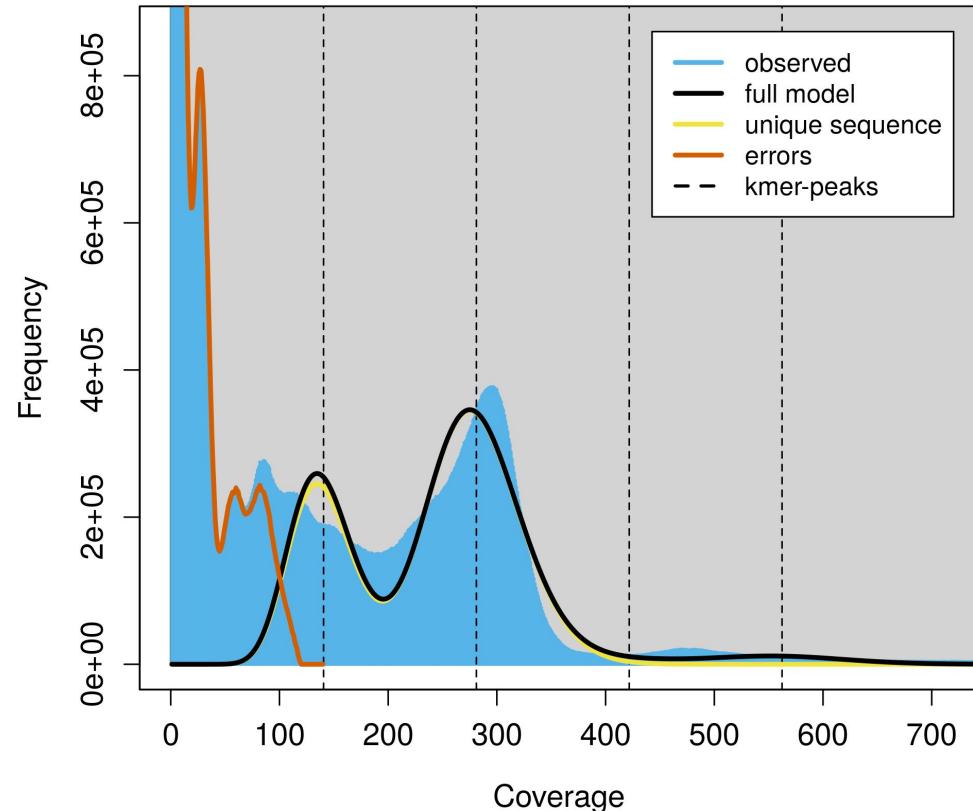
Example by Ksenia Krasheninnikova

“Pooled” sequencing also make bridges



Stylops aterrimus

- One clear genomic peak
- Lots of minor peaks that do not seem to have clear stochasticity to the main peak
 - > looks like pool of genotypes, possibly just because the sequenced female was mated?



Running GenomeScope on a server

<http://genomescope.org/genomescope2.0/>

GenomeScope 2.0

Estimate genome heterozygosity, repeat content, and size from sequencing reads using a kmer-based statistical approach.

Run GenomeScope

Click or drop .histo file here to upload

Description my sample

K-mer length 21

Ploidy 2

Max k-mer coverage -1

Average k-mer coverage for polyploid genome -1

Submit

Instructions

Upload results from running Jellyfish or KMC. Example: `inputk21.hist`

Instructions for running Jellyfish:

1. Download and install Jellyfish from: <http://www.genome.umd.edu/jellyfish.html#Release>
2. Count k-mers using Jellyfish:

```
$ jellyfish count -C -m 21 -s 1000000000 -t 10 *.fastq -o reads.jf
```

Note you should adjust the memory (-s) and threads (-t) parameters according to your server. This example will use 10 threads and 1GB of RAM. The k-mer length (-m) may need to be scaled if you have low coverage or a high error rate. You should always use "canonical k-mers" (-C).

3. Export the k-mer count histogram

```
$ jellyfish histo -t 10 reads.jf > reads.histo
```

Again the thread count (-t) should be scaled according to your server.

4. Upload `reads.histo` to GenomeScope

Instructions for running KMC:

1. Download and install KMC from: <http://sun.aei.polsl.pl/REFRESH/index.php?page=projects&project=kmc&subpage=download>
2. Count k-mers using KMC:

```
$ mkdir tmp
```



```
$ ls *.fastq > FILES
```



```
$ kmc -k21 -t10 -m64 -ci1 -cs10000 @FILES reads tmp/
```

Note you should adjust the memory (-m) and threads (-t) parameters according to your server. This example will use 10 threads and 64GB of RAM. The k-mer length (-k) may need to be scaled if you have low coverage or a high error rate. The lower (-ci) and upper (-cs) bounds exclude k-mers with counts outside these boundaries. `FILES` is a file with a list of input files.

3. Export the k-mer count histogram

k-mer histogram

1	1264991662
2	128715689
3	35786231
4	14698890
5	7658823
6	475066
7	3422518
8	2858068
9	2737239
10	2944294
11	3473705
12	4288452
13	5379104
14	6746789
15	8374781
16	10152558
17	11869712
18	1320337
19	14663647
20	15457083
21	15833391
22	15731512
23	15224706
24	14487391
25	13626577



Mike Schatz

Running GenomeScope in terminal

```
[kj11@ib117963s:~/Projects/kmer_stuff/genomescope2.0]$ genomescope.R --help
usage: /Users/kj11/bin/genomescope.R [-h] [-v] [-i INPUT] [-o OUTPUT]
                                     [-p PLOIDY] [-k KMER_LENGTH]
                                     [-n NAME_PREFIX] [-l LAMBDA]
                                     [-m MAX_KMERCOV] [--verbose]
                                     [--no_unique_sequence] [-t TOPOLOGY]
                                     [--initial_repetitiveness INITIAL_REPEATIVENESS]
                                     [--initial_heterozygosities INITIAL_HETEROZYGOSITIES]
                                     [--transform_exp TRANSFORM_EXP]
                                     [--testing] [--true_params TRUE_PARAMS]
                                     [--trace_flag] [--num_rounds NUM_ROUNDS]
                                     [--fitted_hist]
                                     [--start_shift START_SHIFT]
                                     [--typical_error TYPICAL_ERROR]
```

The workflow for this morning



FastK

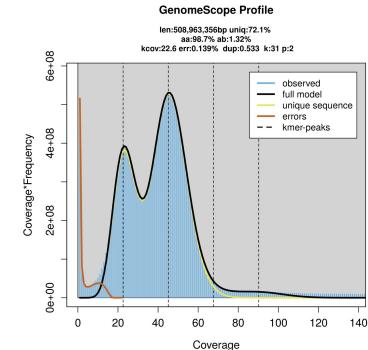
creates a *k*-mer database and

kmer	coverage
AAAAAAAAAAACACGT	28
AAAAATAACACAACGT	31
AAAATAACACAACGG	1
AAAAATAACGCAACGT	47
AAAATTACGCAACGT	2
AAAATTACGCAACGA	17
...	...

k-mer histogram

1 1264991662
2 128715689
3 35706231
4 14690890
5 7658823
6 4750866
7 3423518
8 2858068
9 2737239
10 2944294
11 3473705
12 4288452
13 5379194
14 6746789
15 8374781
16 10152558
17 11869712
18 13391733
19 14663647
20 15457083
21 15833391
22 15731512
23 15224796
24 14487391
25 13626577

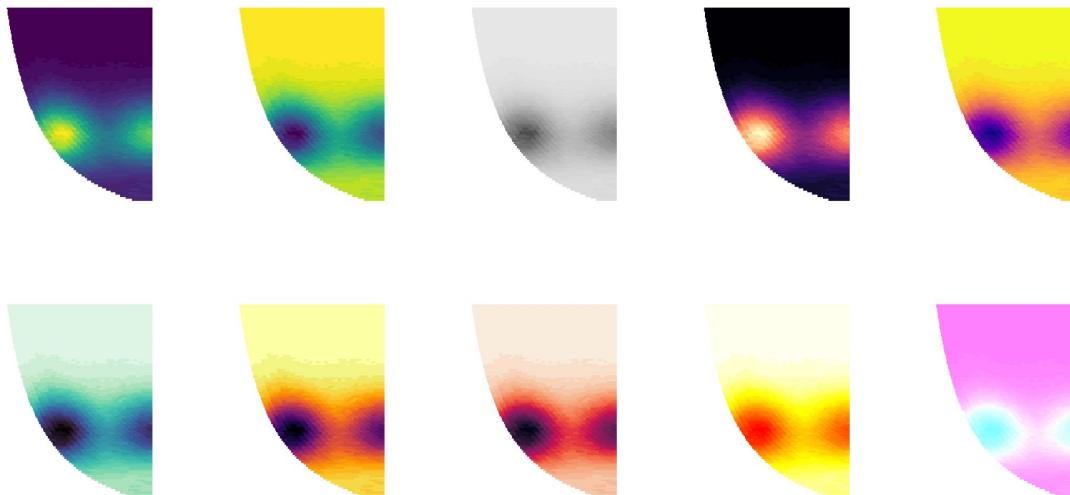
GenomeScope



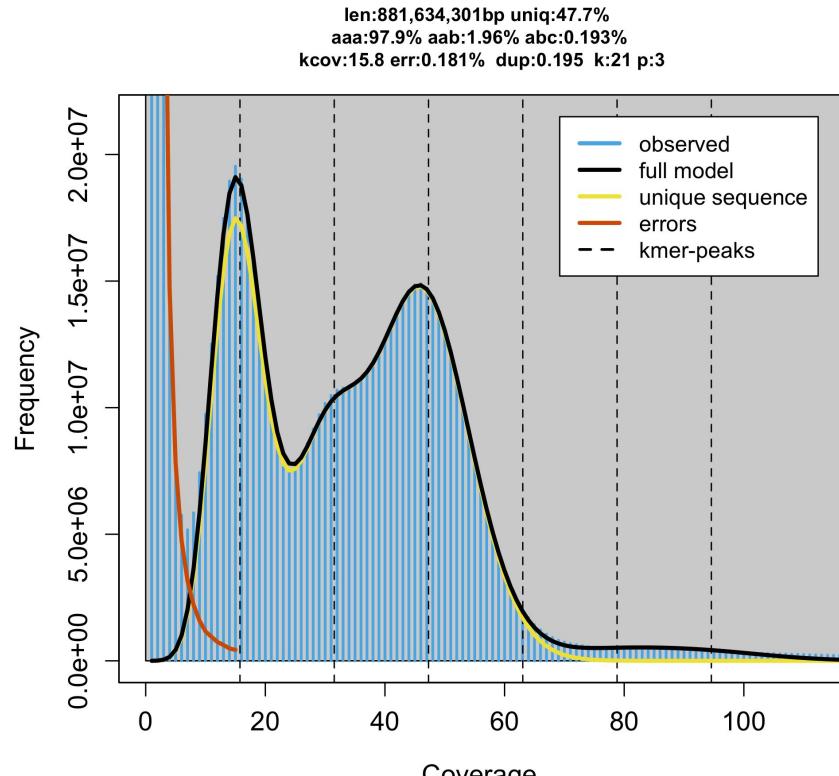
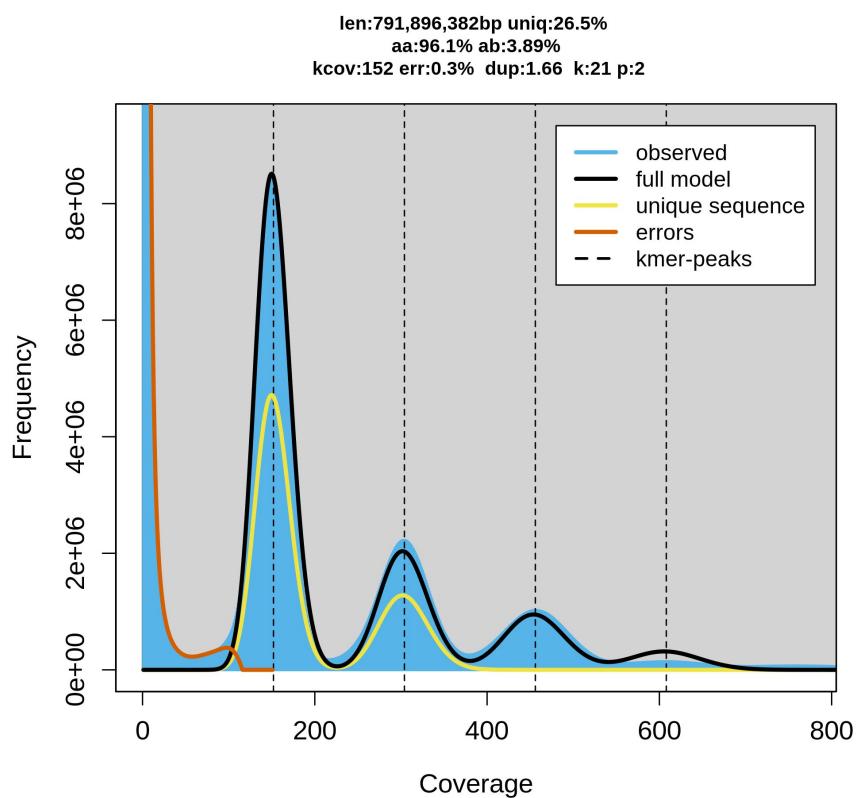
Your turn...

- Build *k*-mer spectrum
- Fit example GenomeScope model
- Fit many more
 - discuss what do you see / think !!!

Part 3: Genome profiling of polyploids



Spectra with less obvious interpretation



Find all “hetmers” - unique k-mers distant one nt

AACTCA
AACTAA
CCCTTA
GGCTCA
GGCTAA
GGCTTA

Yaay, a pair
Not a pair
Not unique

b

Structure	CovA + CovB	$\frac{\text{CovB}}{\text{CovA} + \text{CovB}}$
AB	2λ	0.5
AAB	3λ	0.333
AAAB	4λ	0.25
AABB	4λ	0.5
AAAAAB	5λ	0.2
AAABB	5λ	0.4
AAAAAAB	6λ	0.166
AAAABB	6λ	0.333
AAABBB	6λ	0.5
...



Find all “hetmers” - unique k-mers **distant one nt**

```
smudgeplot hetmers <FastK_kmer_db>
```

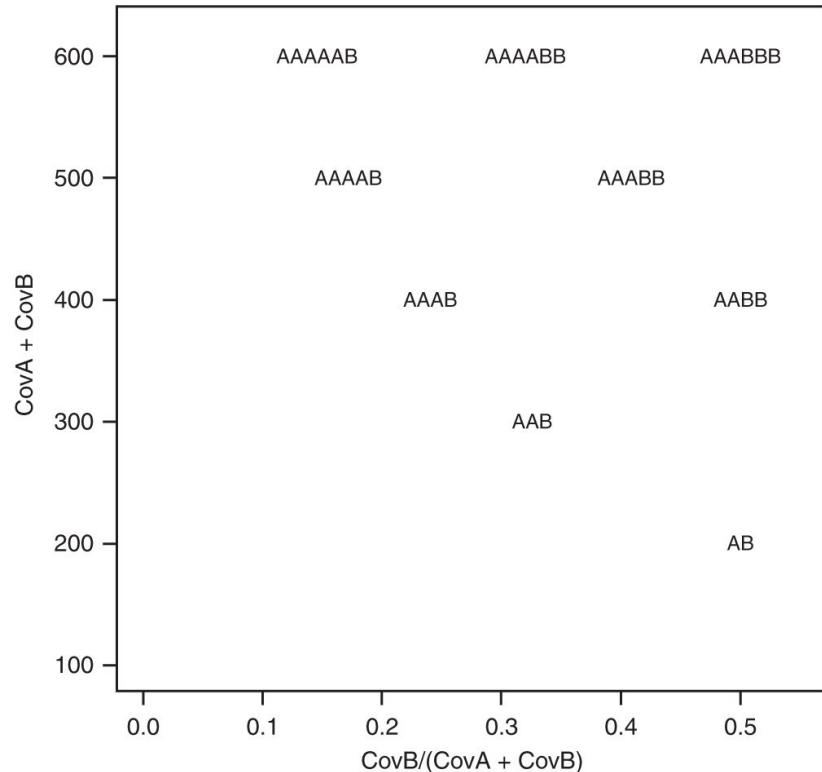
Fast implementation by Gene Myers

```
[kj11@nib117963s:~/Projects/kmer_stuff$ head ./smudgeplot/data/Fara2/kmerpairs_text.smu
26      26      6560
26      27      12888
26      28      12584
27      27      6512
26      29      12532
27      28      12764
26      30      11990
27      29      12624
28      28      6110
26      31      11110
```

covB covA freq

Plotting the two creates “smudgeplot”

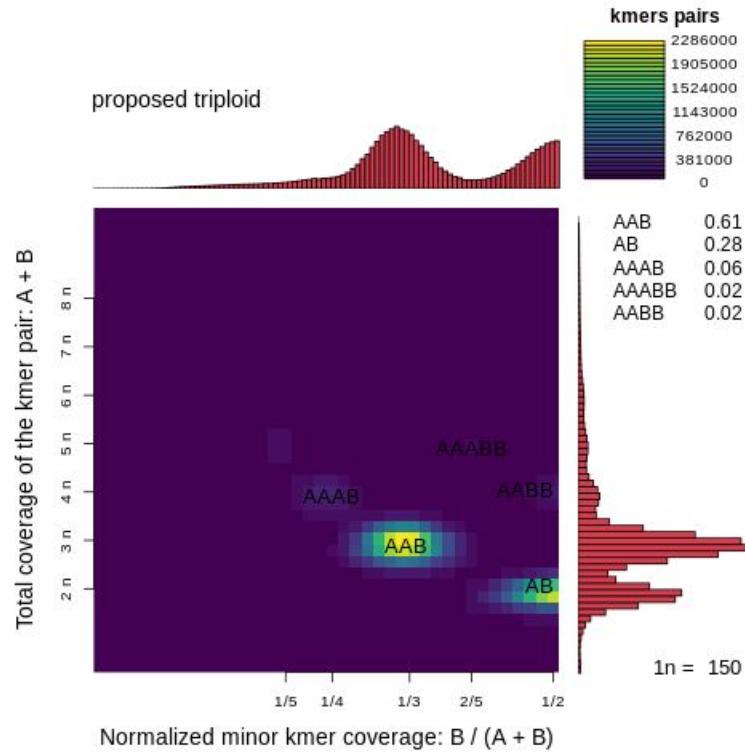
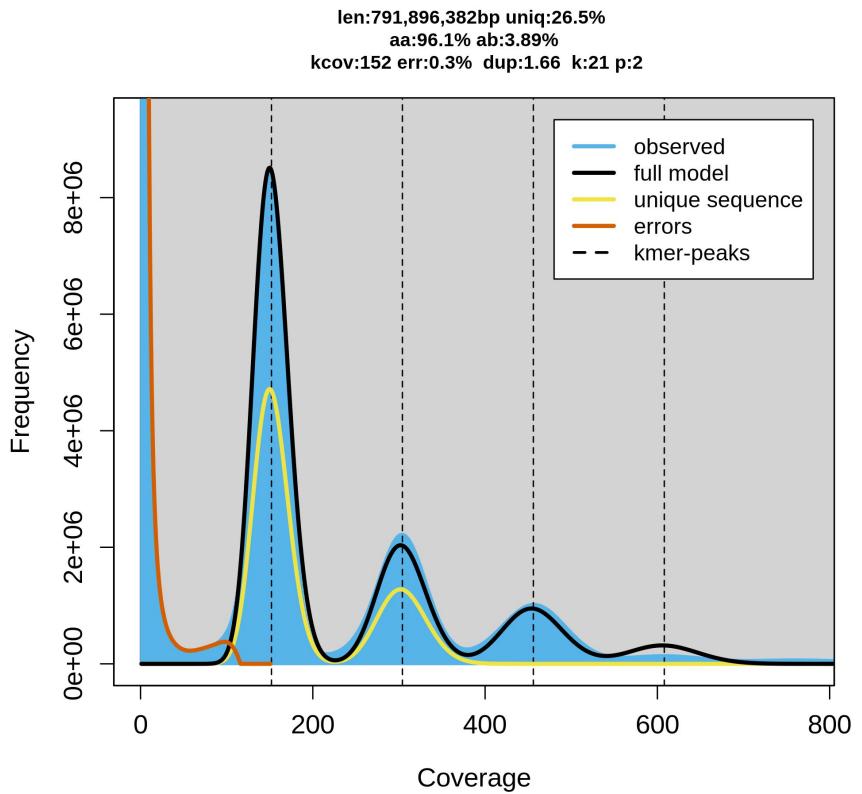
a



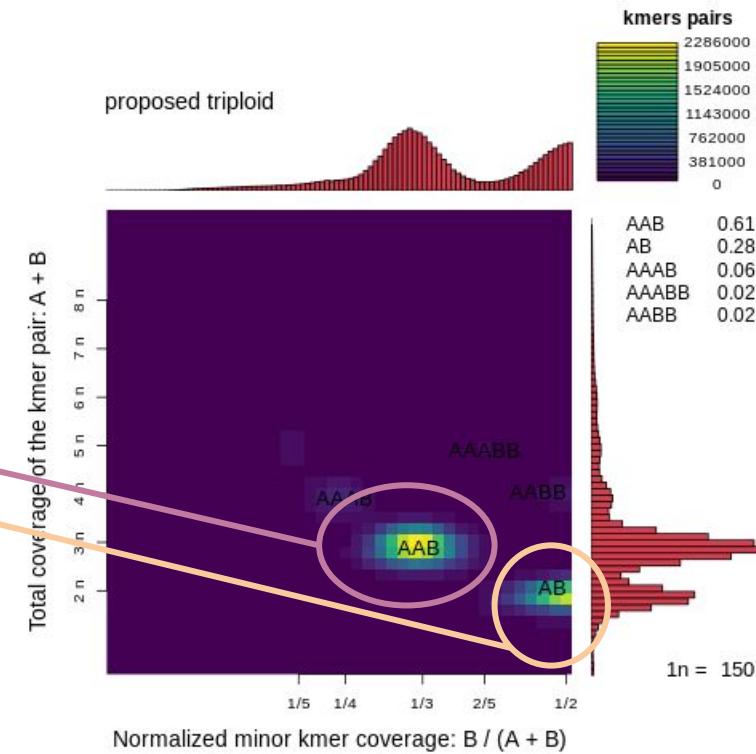
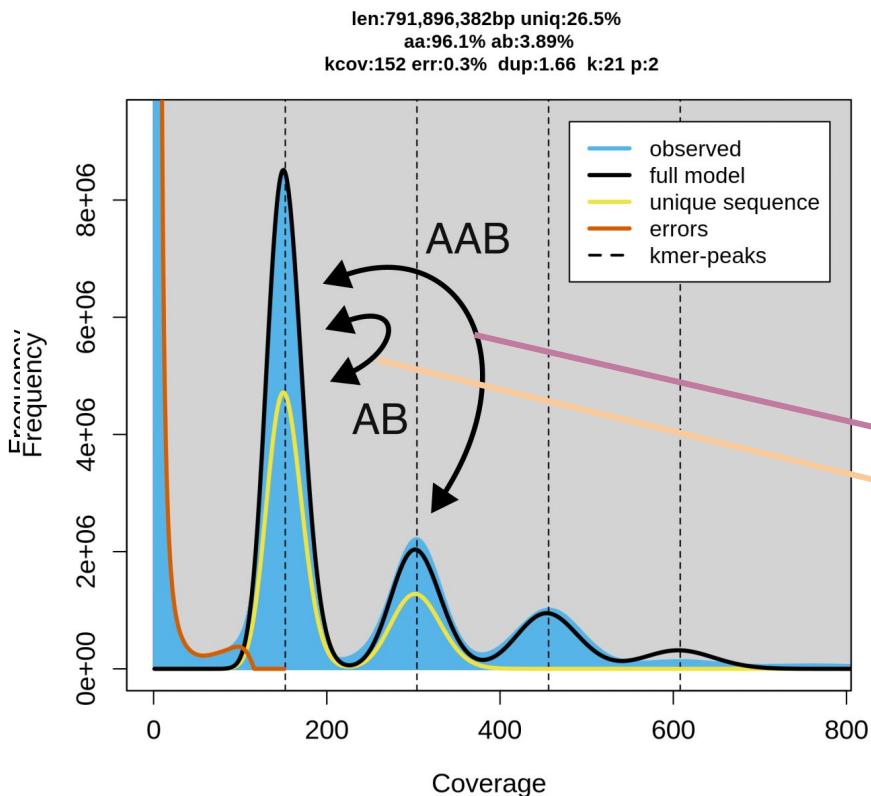
b

Structure	CovA + CovB	$\frac{\text{CovB}}{\text{CovA} + \text{CovB}}$
AB	2λ	0.5
AAB	3λ	0.333
AAAB	4λ	0.25
AABB	4λ	0.5
AAAAAB	5λ	0.2
AAABB	5λ	0.4
AAAAAAB	6λ	0.166
AAAABB	6λ	0.333
AAABBB	6λ	0.5
...

Smudgeplot indicates triploidy

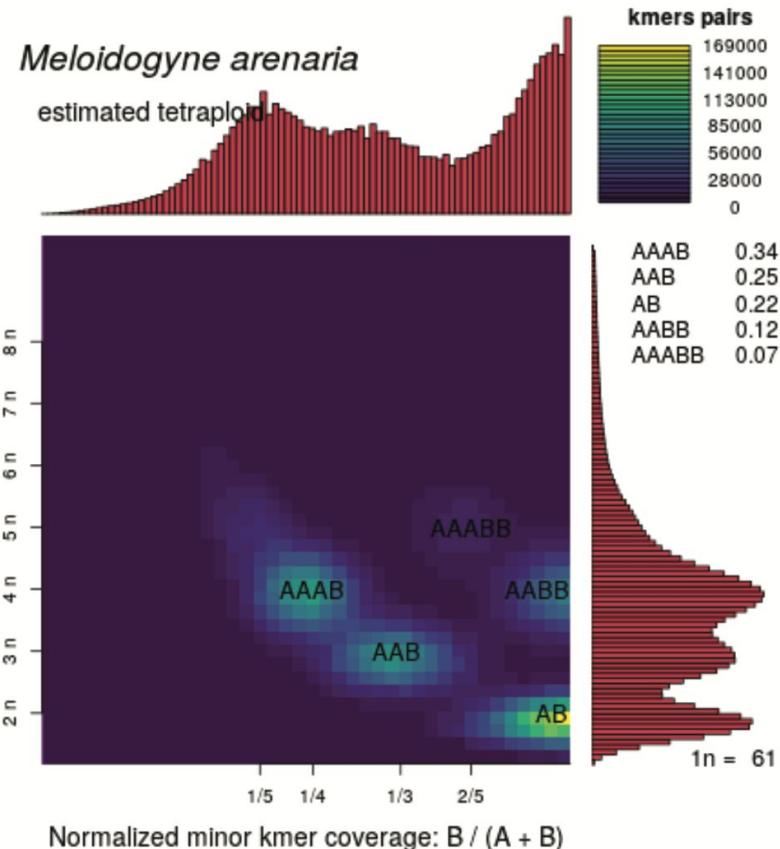
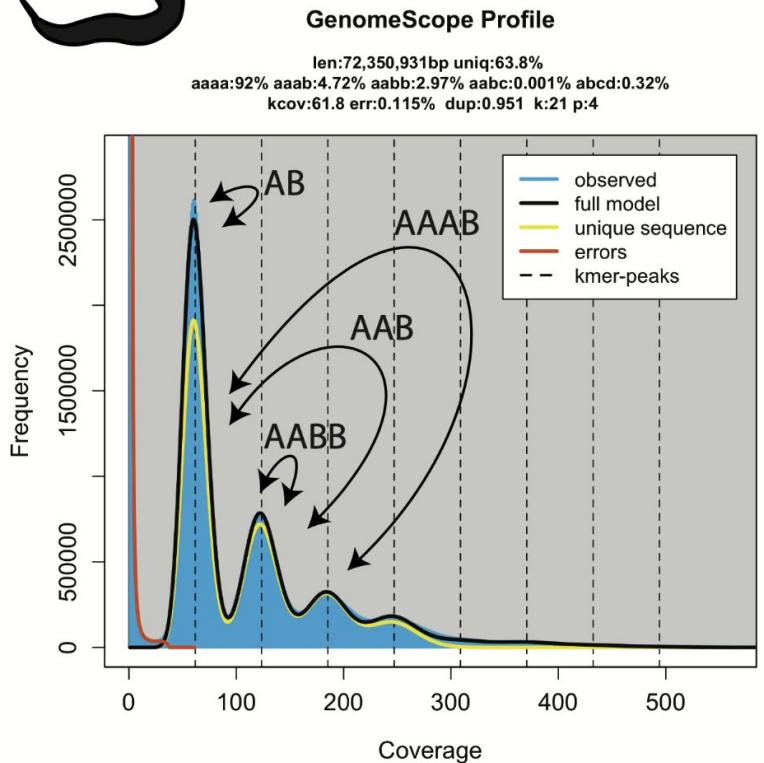


Joint interpretation of smudgeplot and genomescope

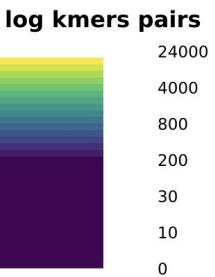


Posted by João G. Ferreira to smudgeplot Issues

Works for tetraploids too!

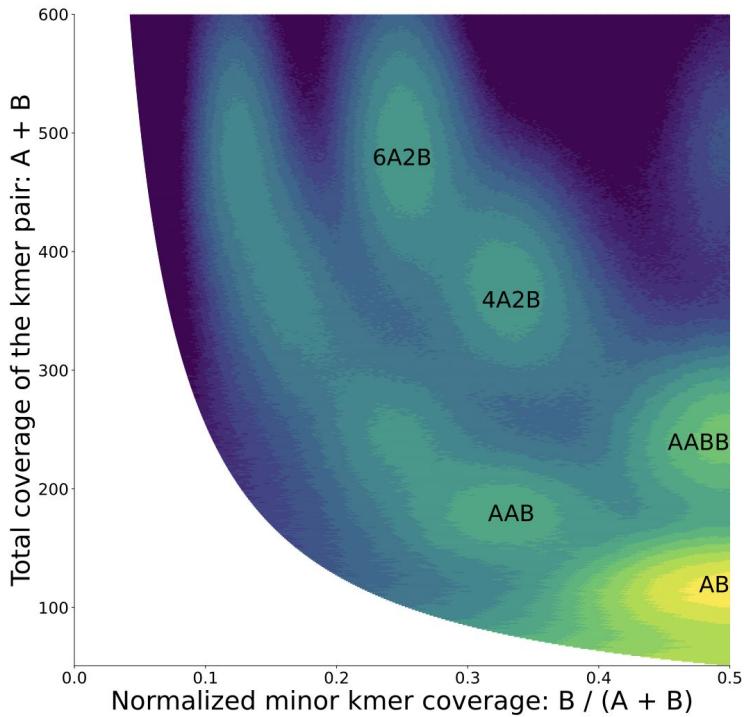


1n = 60.0
err = 0.007%

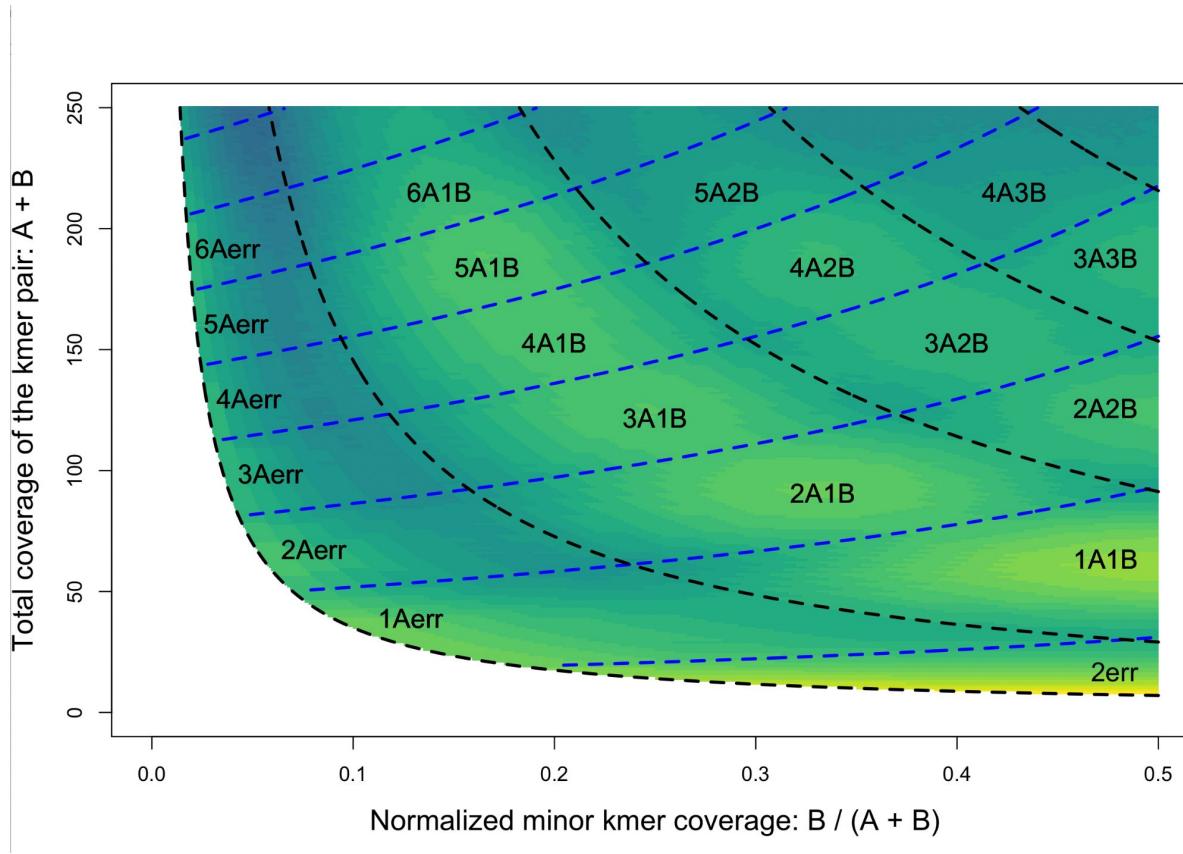


Smudgeplot works for high ploidies

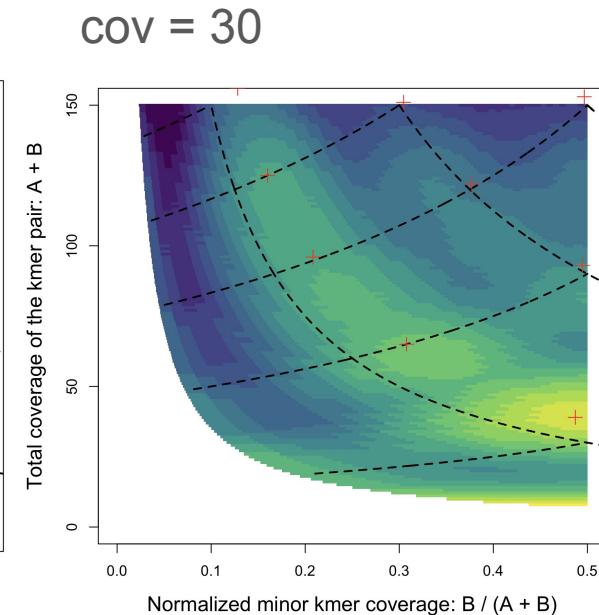
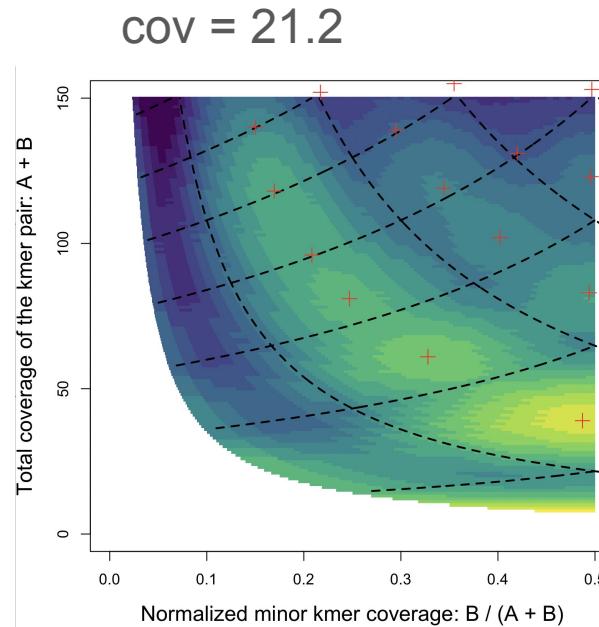
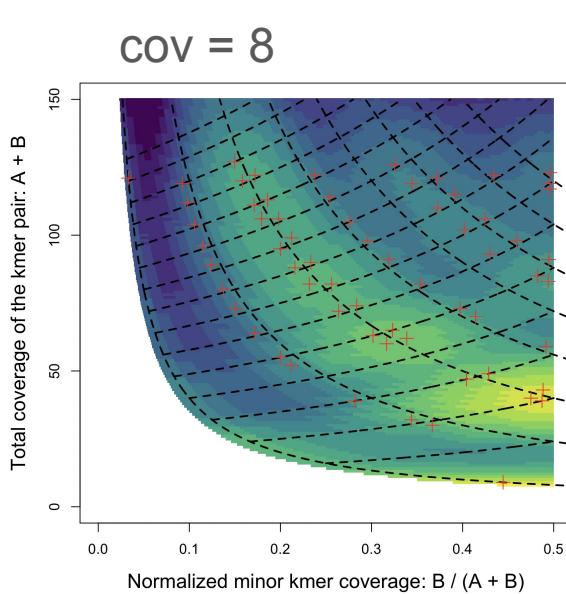
octoploid strawberry



Coverage inference by “fishnet”



We can test each coverage

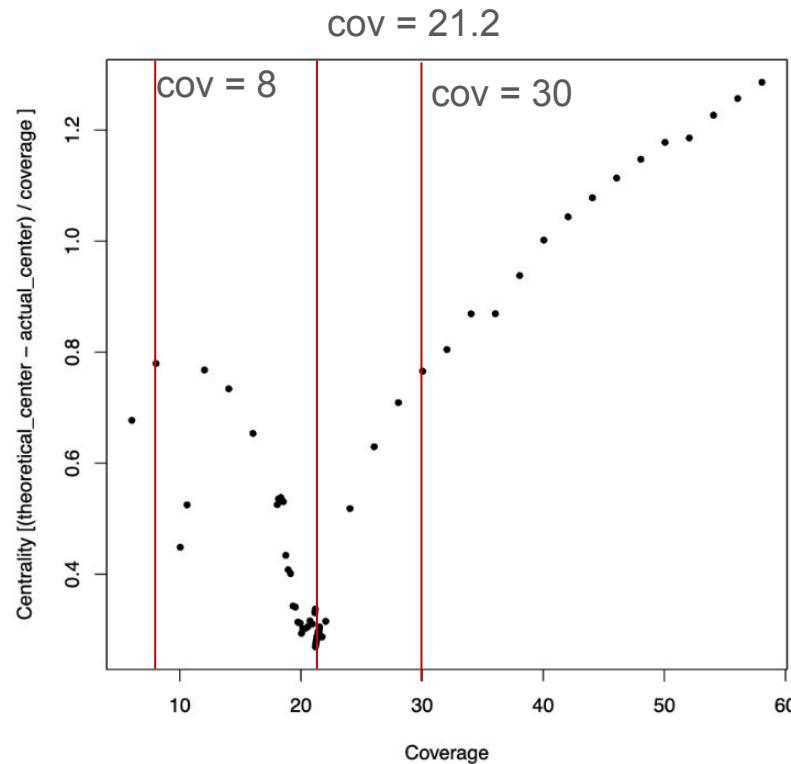


centrality: 0.8

centrality: 0.31

centrality: 0.78

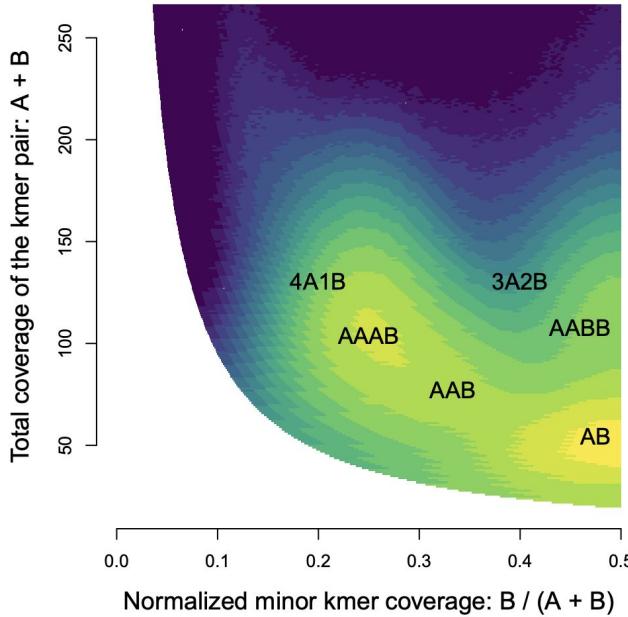
For each coverage, “cast the fishnet” and calculate “centrality”



willow

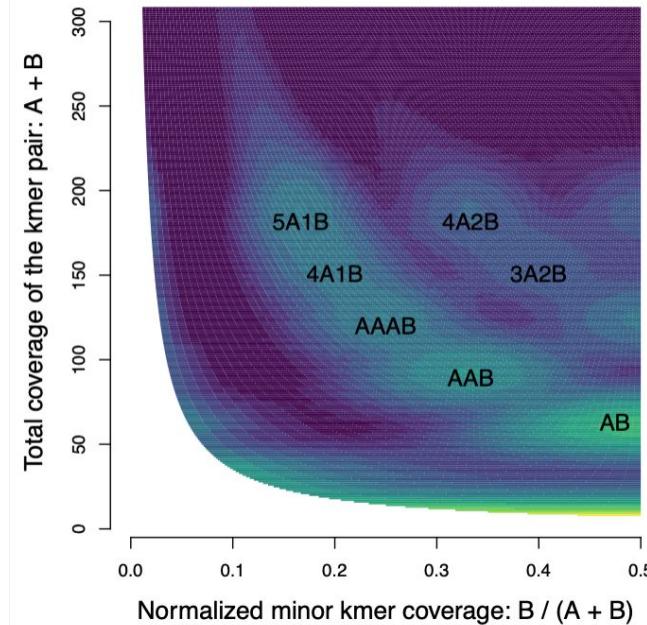
26M1

autotetraploid



sea wormwood

hexaploid

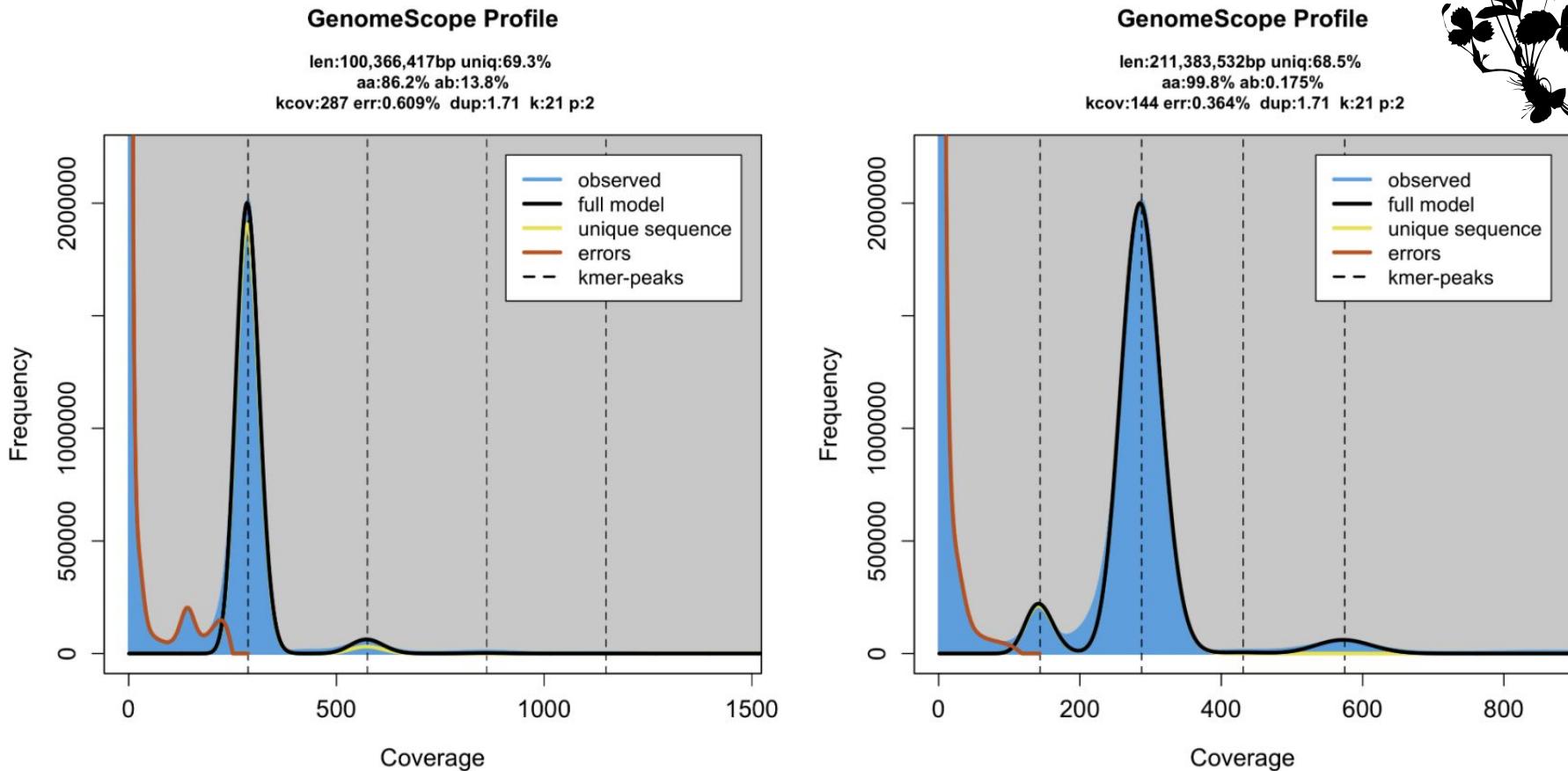


Smudgeplot by Vinciane Mossion

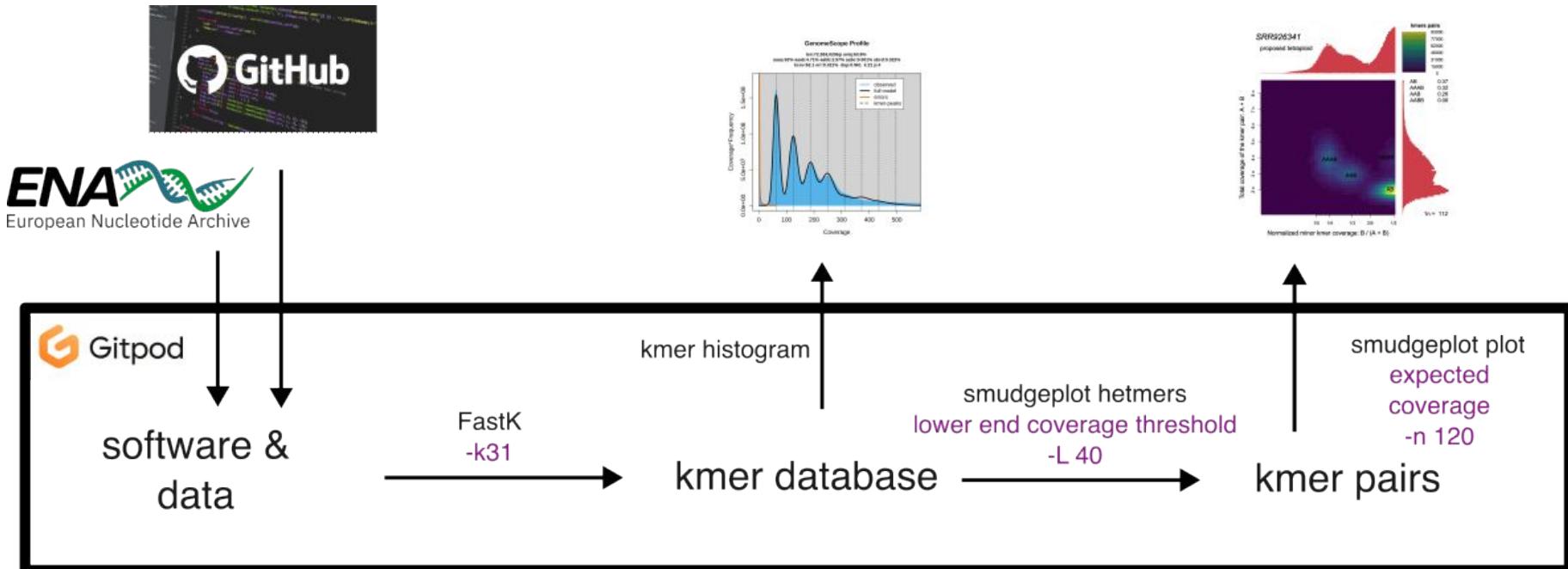
What to pay attention when genome profiling

- Find all possible “genome models”
 - Combinations of genomic features that would generate patterns observed by GenomeScope and Smudgeplot
- Make sure your model fits the data
- Make sure that all evidence makes sense together
 - E.g. Is the 1n coverage in smudgeplot and genomescope consistent?
- Make sure that your interpretation is somewhat compatible with known biology of the species
- Confront your model to the genome assembly
 - Est genome size vs assembled
 - Is the assembly with collapsed or uncollapsed haplotypes? Would you expect it given estimated heterozygosity?
 - Is the assembly ploidy consistent with smudgeplot?

Convergence problems - strawberry



The whole genome profiling pipeline



There is a place for all the freaks... all data are available on ToLQC

<https://tolqc.cog.sanger.ac.uk/>

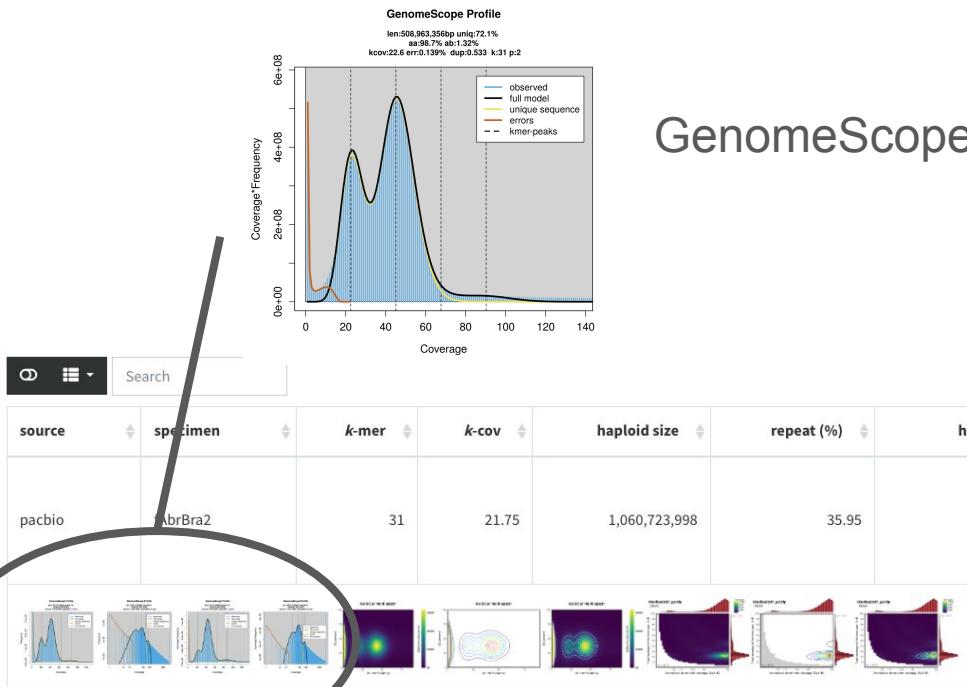


Shane McCarthy

k-mer
histogram

1	1264991662
2	128715689
3	35786231
4	146980890
5	7658823
6	475006
7	3422518
8	2858068
9	2737239
10	2944294
11	3473705
12	4288452
13	5379104
14	6746789
15	8374781
16	10152558
17	11869712
18	1300037
19	14663647
20	15457083
21	15833391
22	15731512
23	15224706
24	14487391
25	13626577

GenomeScope plots



Questions?

