4 - Genome Assembly and Validation (Concepts)

Wednesday afternoon

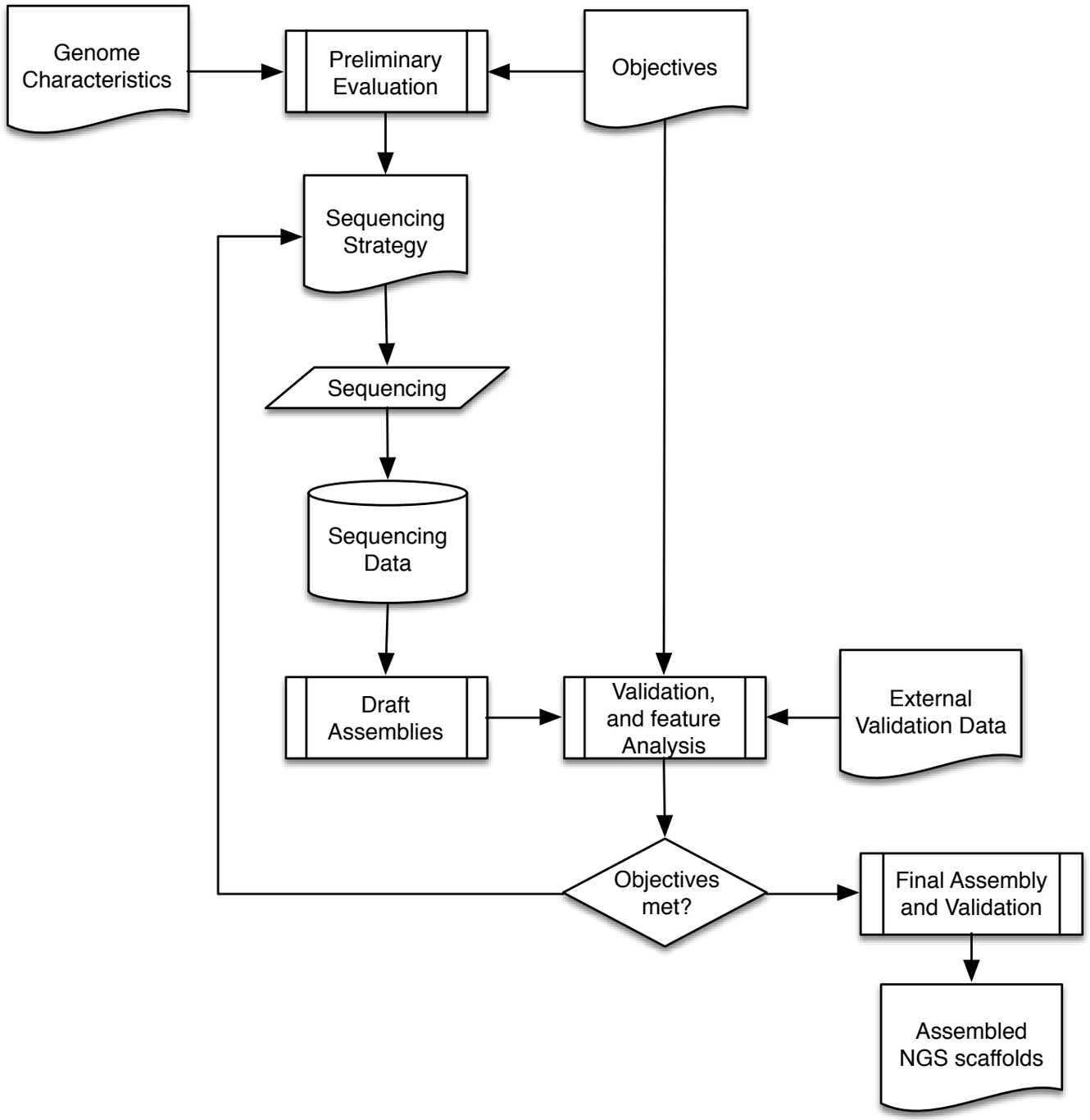
Bernardo J. Clavijo
Richard Smith-Unna
Gonzalo Garcia / Jon Wright





TGAC

Assembly project workflow | Prior Knowledge



- Kariotype: Genome size, Ploidy
- Heterozygocity
- GC content
- Contaminants / Symbionts
- Data Sets:
  - Close relatives
  - Genes / ESTs / RNAseq / Markers
- Mithocondria
- Chloroplast

TGAC

Experiment design (you choose the data!)

- **Know your biological question.**

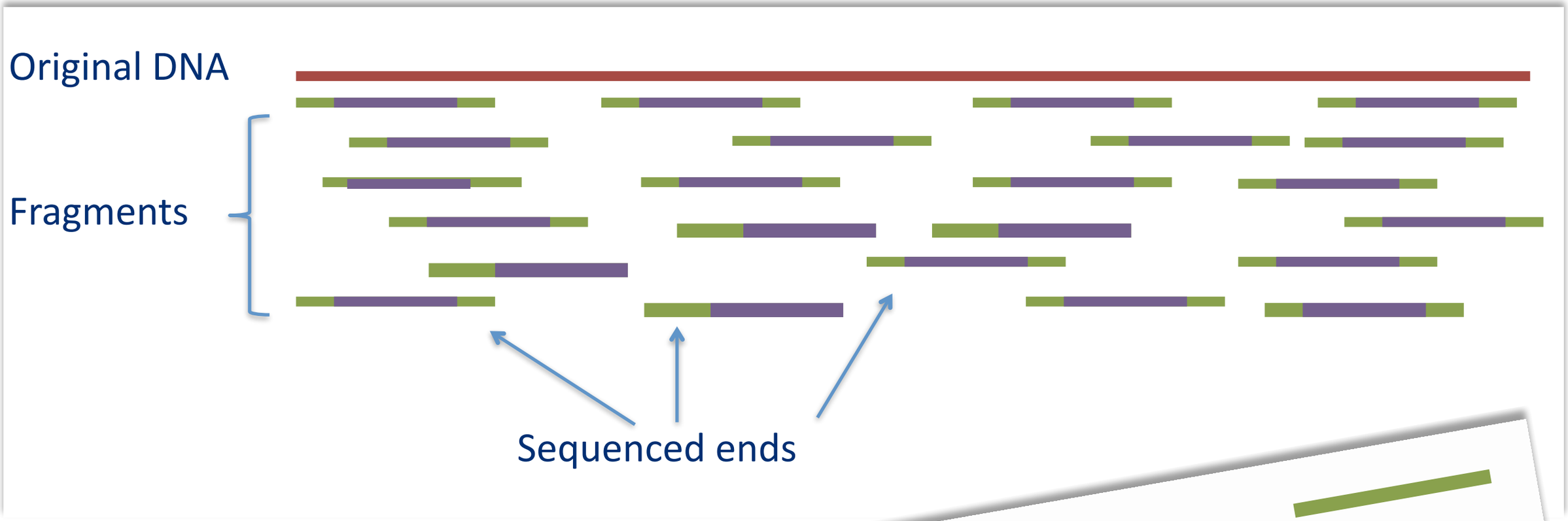
- Plan your data processing (from an information perspective).

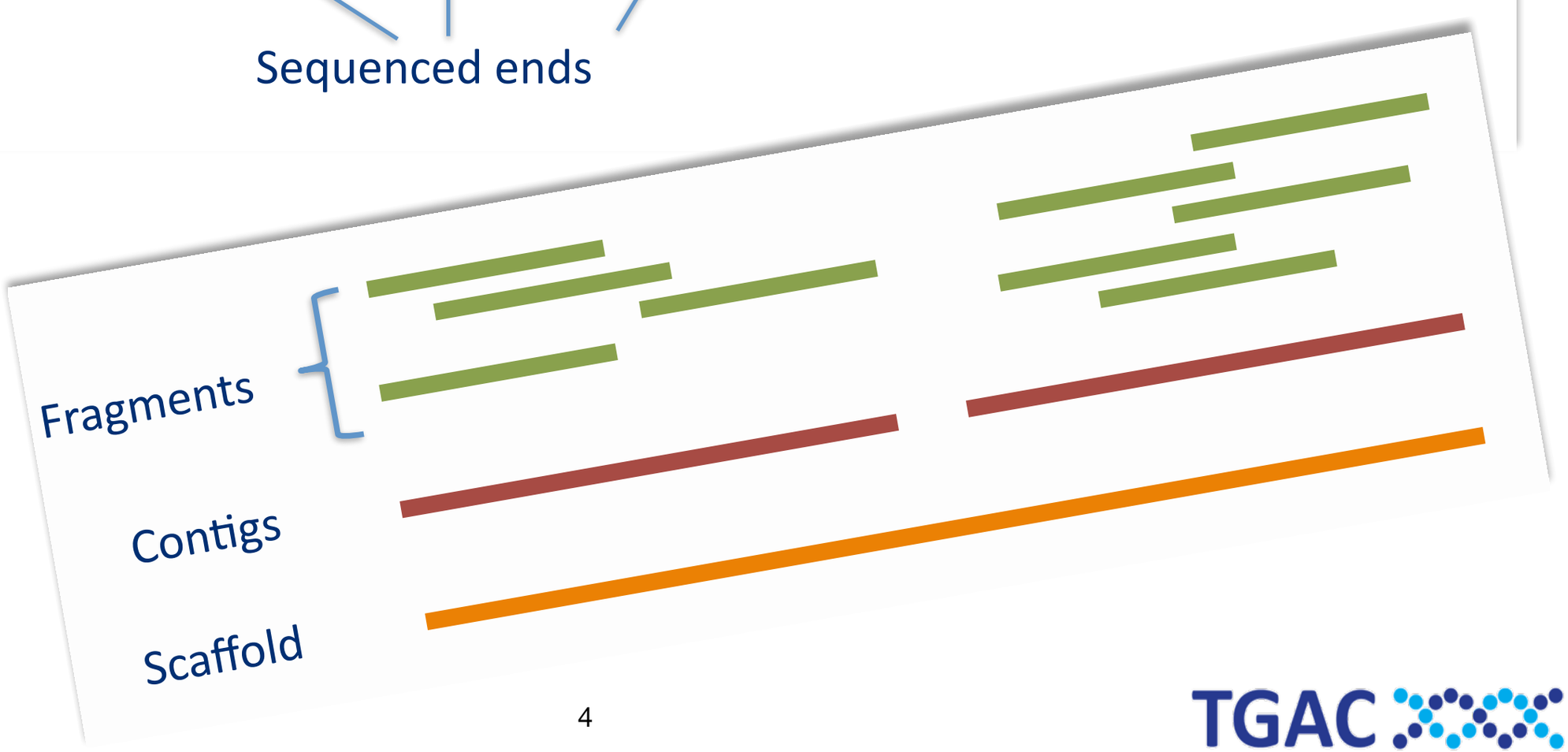
- Decide on conditions and biological/technical replicas.

- Decide on technologies and coverages:
  - How will the typical bias affect your experiment?
  - Is the coverage enough? Significant results?

TGAC

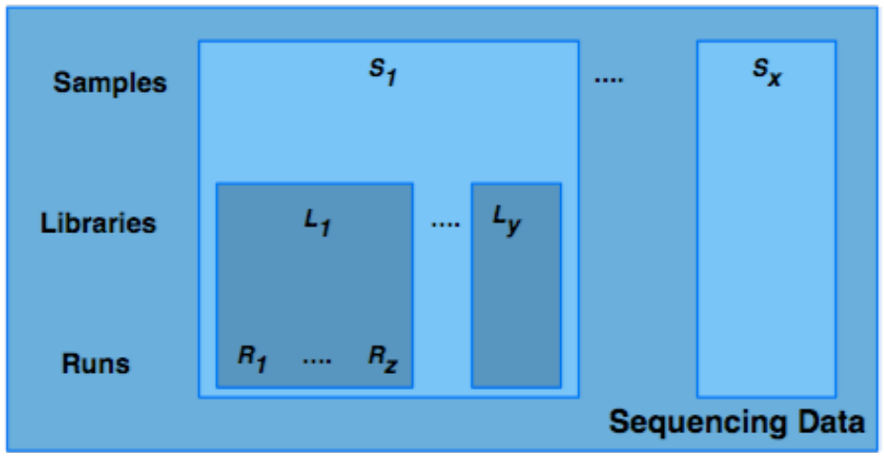
The genome assembly problem (WGS)





TGAC

Planning and "informed guesses"



Whole genome information (unknown?)
+Bias Information (unknown)

"known error profile"

short sequences ($10^{\{2\text{-}3\}}$)

large data (10^{X})

$Y << X$

small data (10^{Y})

long sequences (10^{10})

"unknown error profile"

Whole genome information (known?)
+Bias Information (unknown, reduced?)

Assemble and Scaffold

Scaffolds & Contigs

Validate and release

TGAC

The assembly is just a probabilistic model of a genome, condensing the information from the experimental evidence.

All the information is already present in the experimental results.

TGAC

A correct assembly has:

The right *motifs,*

the correct number of times,

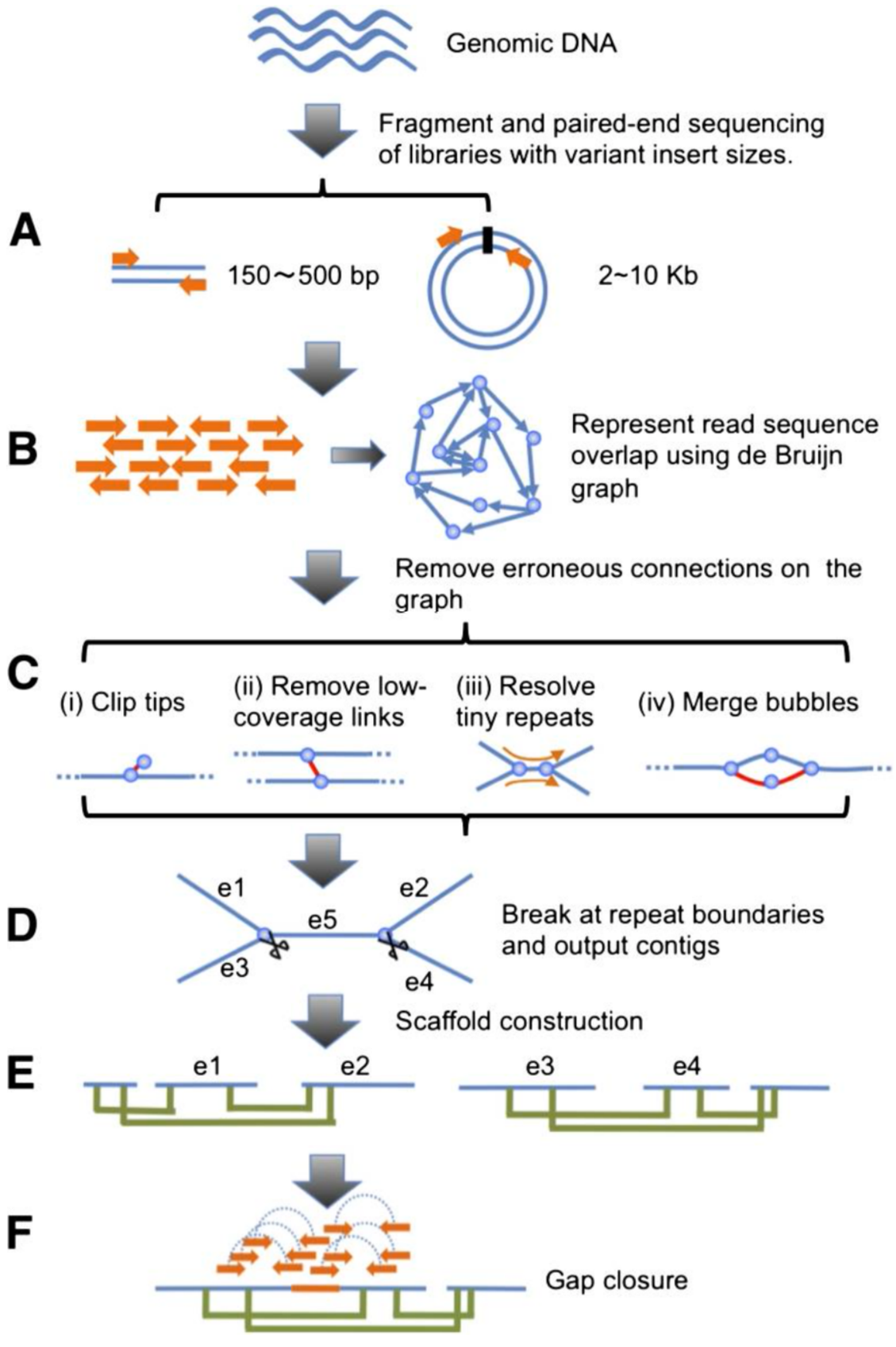
in correct order and position.

None of which is assessed by length stats.

TGAC

A modern assembler

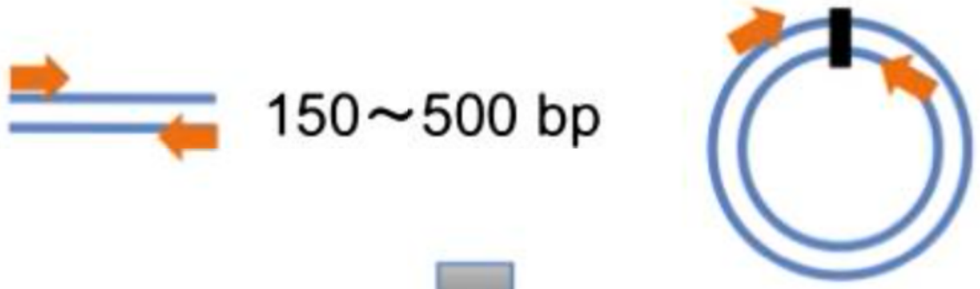
Using SOAPdenovo2 as an example


Genomic DNA
Fragment and paired-end sequencing of libraries with variant insert sizes.
A
150~500 bp
2~10 Kb
B
Represent read sequence overlap using de Bruijn graph
Remove erroneous connections on the graph
C
(i) Clip tips
(ii) Remove low-coverage links
(iii) Resolve tiny repeats
(iv) Merge bubbles
D
e1
e2
e5
e3
e4
Break at repeat boundaries and output contigs
Scaffold construction
E
e1
e2
e3
e4
F
Gap closure

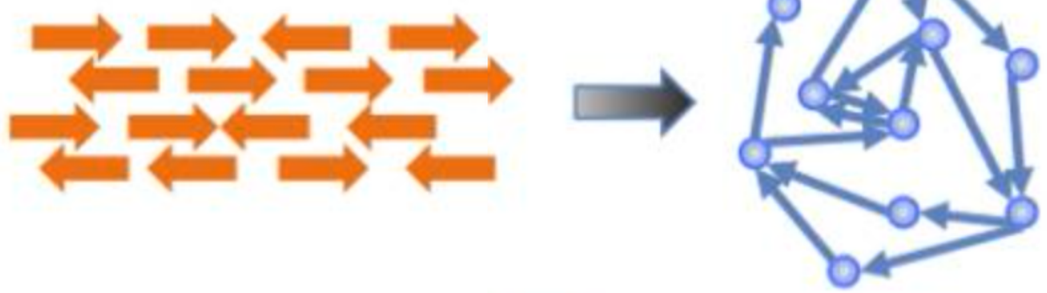
TGAC

of libraries with variant insert sizes.

A



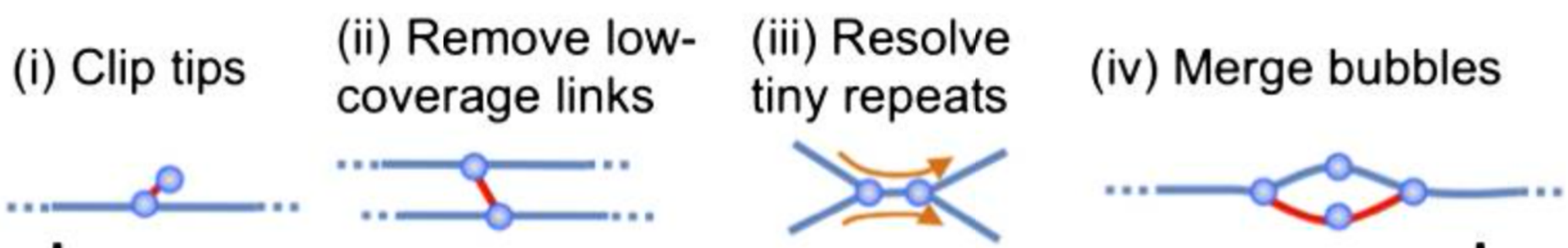
B



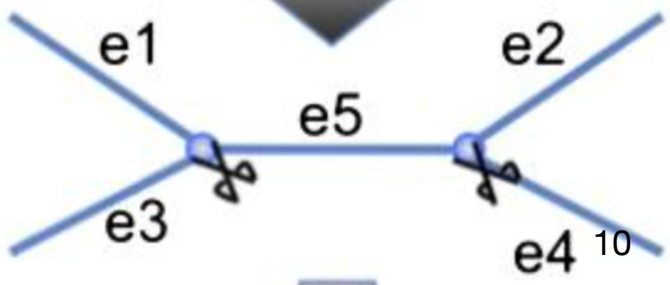
Represent read sequence overlap using de Bruijn graph

Remove erroneous connections on the graph

C



D



Break at repeat boundaries and output contigs

Remove erroneous connections on the graph

C

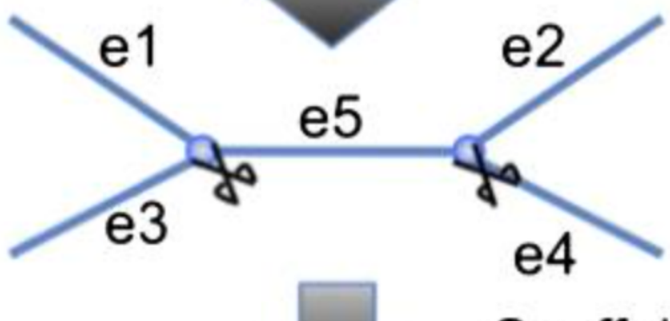
(i) Clip tips

(ii) Remove low-coverage links

(iii) Resolve tiny repeats

(iv) Merge bubbles

D

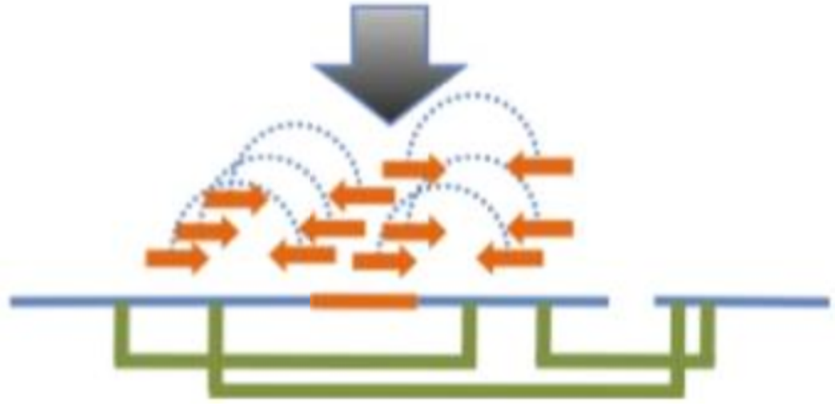


Break at repeat boundaries and output contigs

Scaffold construction

E

F



Gap closure

TGAC

Assembly validation

The power of the kmer spectra…

The right *motifs*,

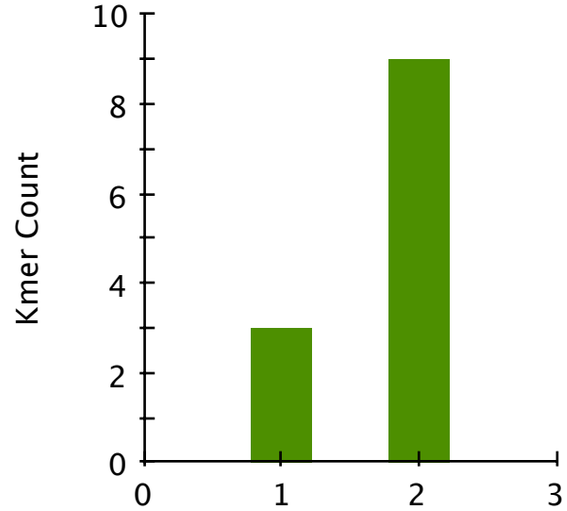
the correct number of times,

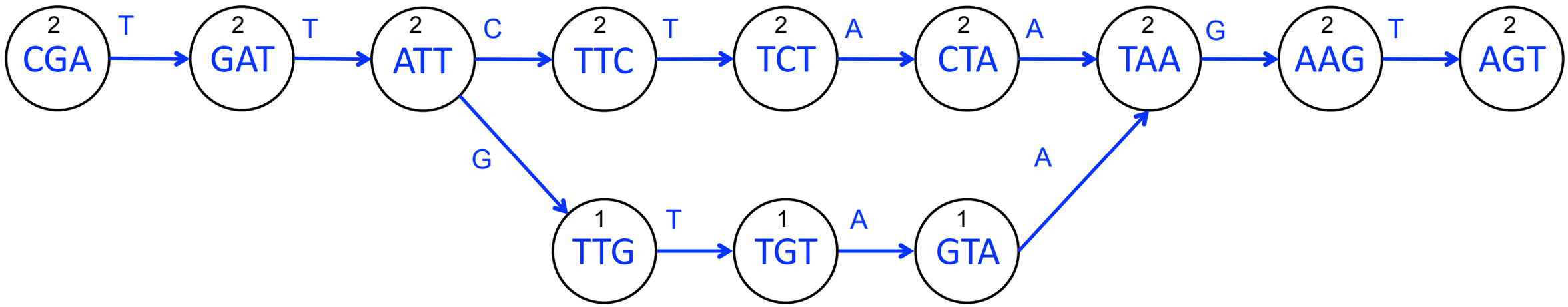
in correct order and position.

Counting kmers

```
>seq1
TTCTAAGT
>seq2
CGATTCTA
```

```
>seq3
CGATTGTAAGT
```





TGAC

The kmer spectra

---

spectra | ˈspɛktrə |
plural form of **SPECTRUM**.

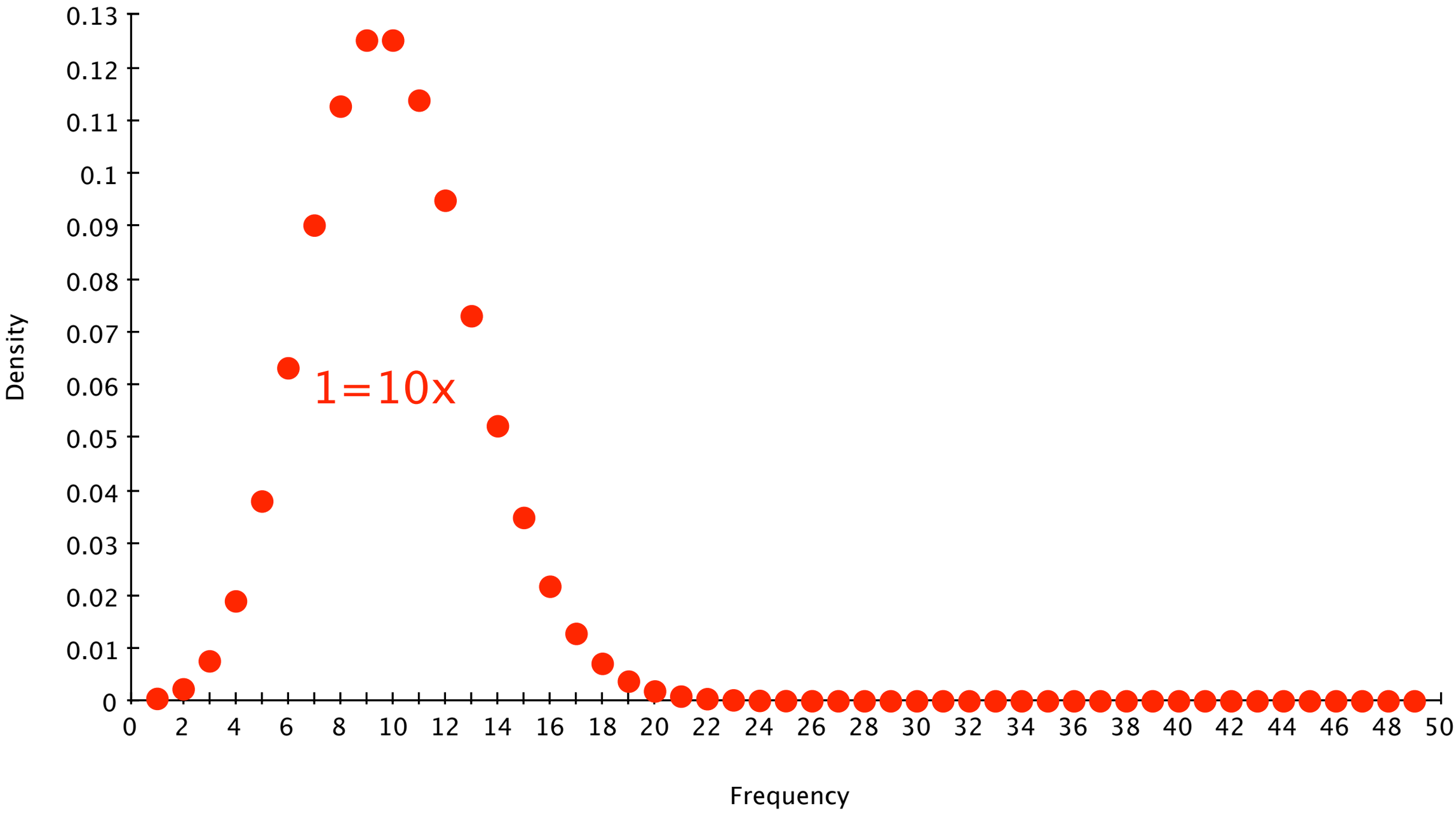
spectrum | ˈspɛktrəm |
noun (pl. **spectra** | -trə | or **spectrums**)

1 a band of colours, as seen in a rainbow, produced by separation of the components of light by their different degrees of refraction according to wavelength.
• (**the spectrum**) the entire range of wavelengths of electromagnetic radiation.
• a characteristic series of frequencies of electromagnetic radiation emitted or absorbed by a substance.
• the components of a sound or other phenomenon arranged according to such characteristics as frequency, charge, and energy.

2 used to classify something in terms of its position on a scale between two extreme points: *the left or the right of the political spectrum*.
• a wide range: *self-help books are covering a broader and broader spectrum*.

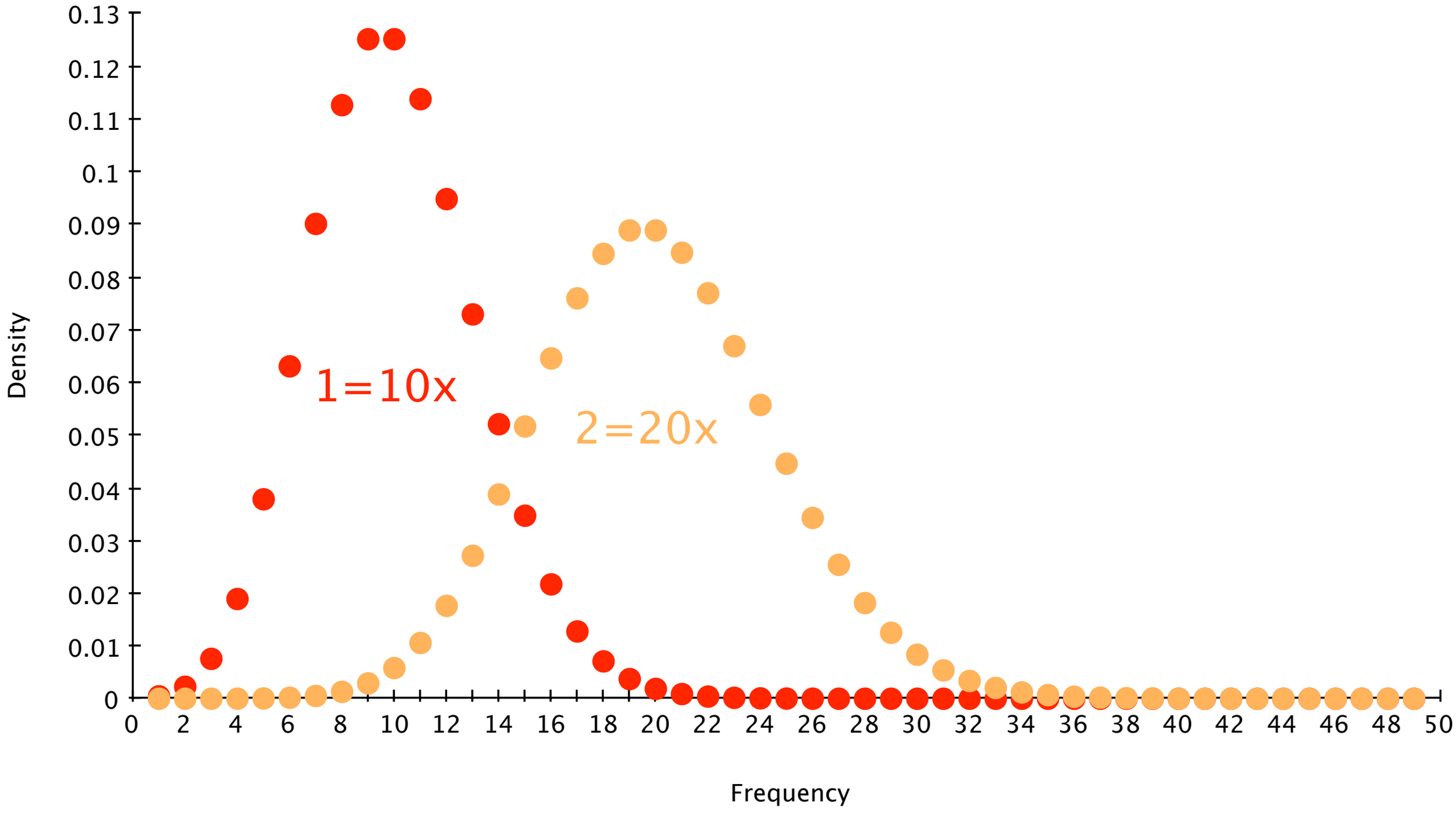
ORIGIN early 17th cent. (in the sense 'spectre'): from Latin, literally ***'image, apparition'***, from ***specere 'to look'***.

The kmer spectra's components



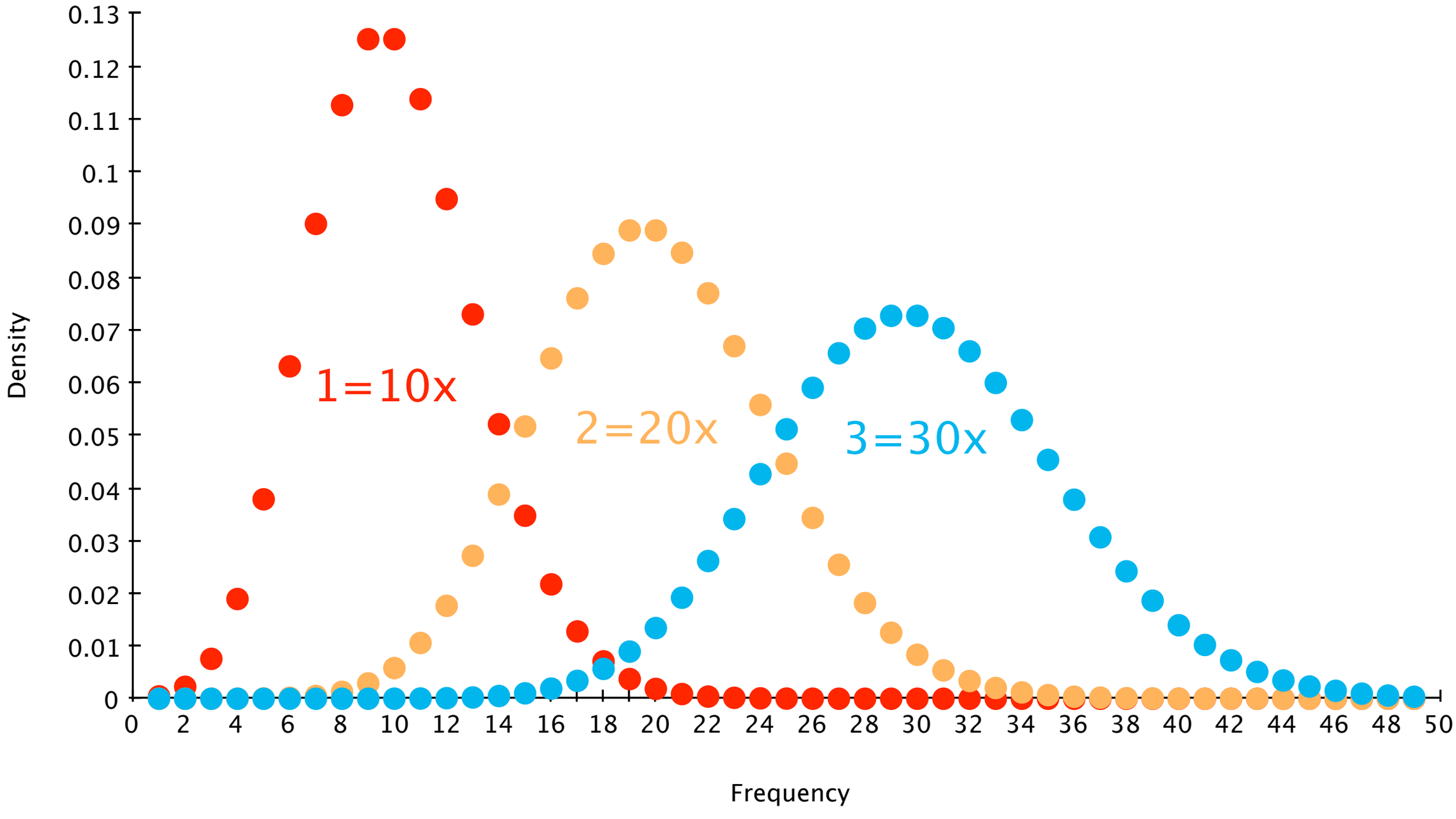
TGAC

The kmer spectra's components



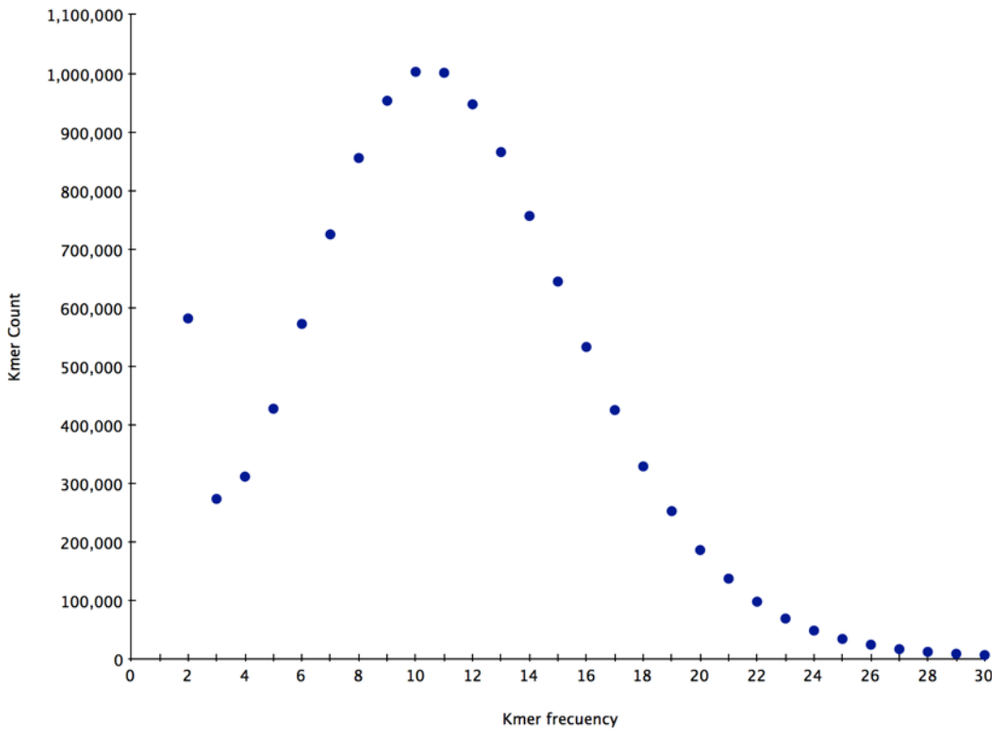
TGAC

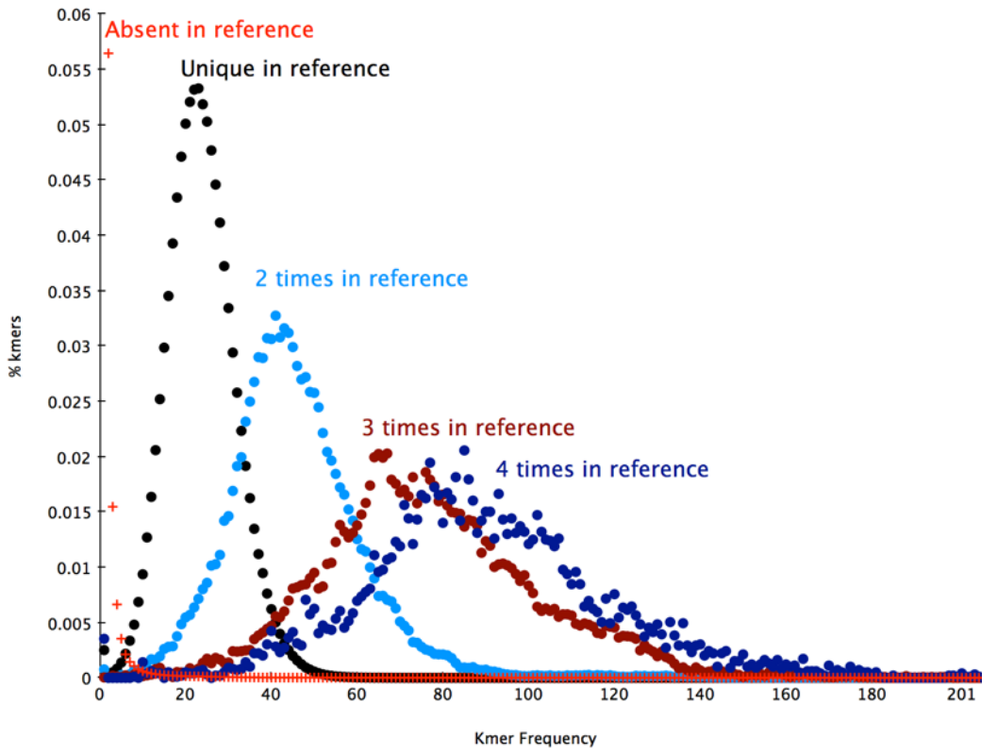
The kmer spectra's components



TGAC

The kmer spectrum… and its dissection.





- We typically use KAT to kmer-count.
- You can “read”:
  - Kmer coverage.
  - Genome size.
  - Errors vs. Good kmers.
- Comparing different spectrum (KAT):
  - Is a reference free library assessment.
  - Runs fast.
  - Gives at least a better vs. worse result.

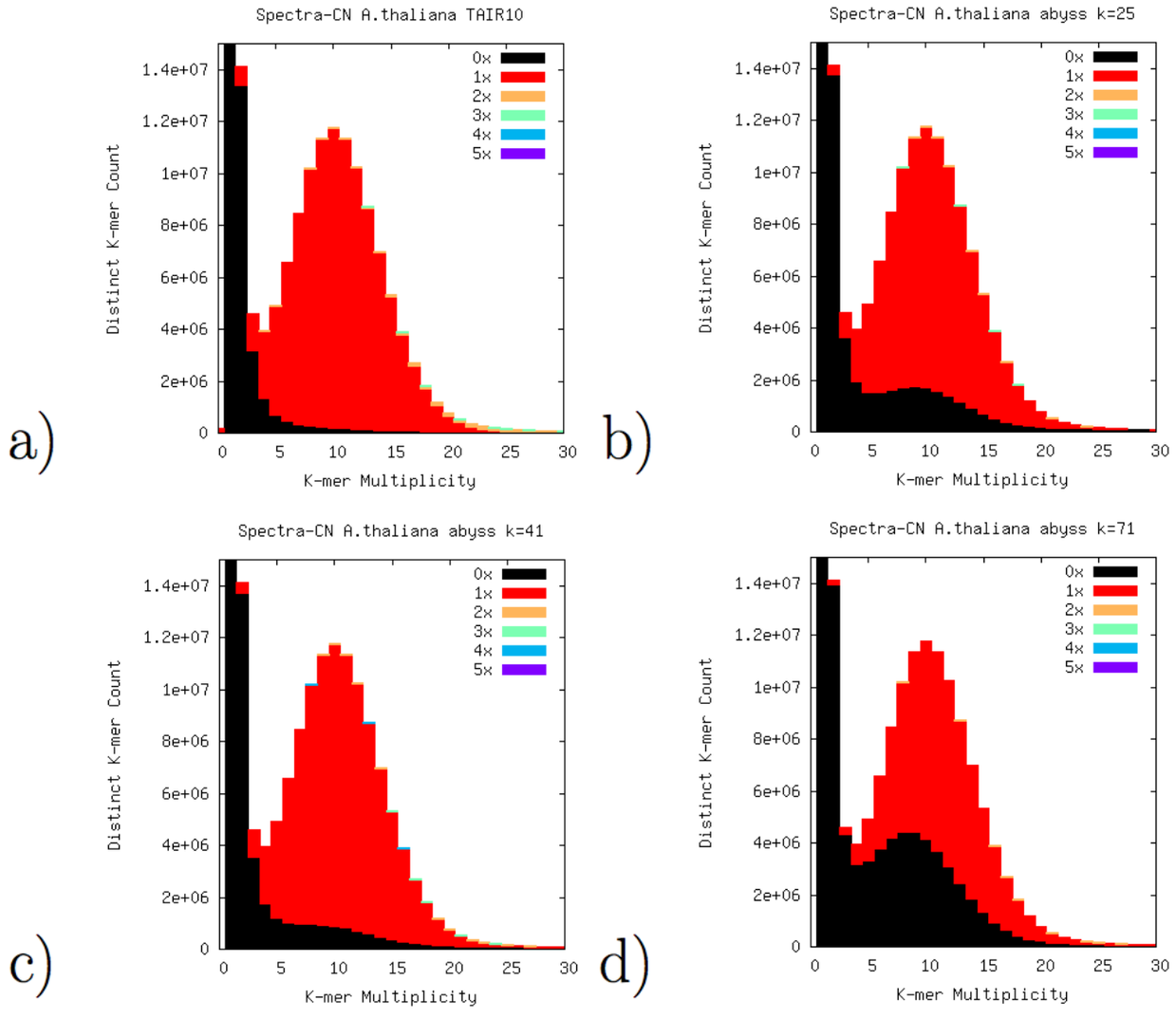
TGAC

If an assembly is correct, then the original reads should be a plausible sequencing set for the resulting genome model.

TGAC

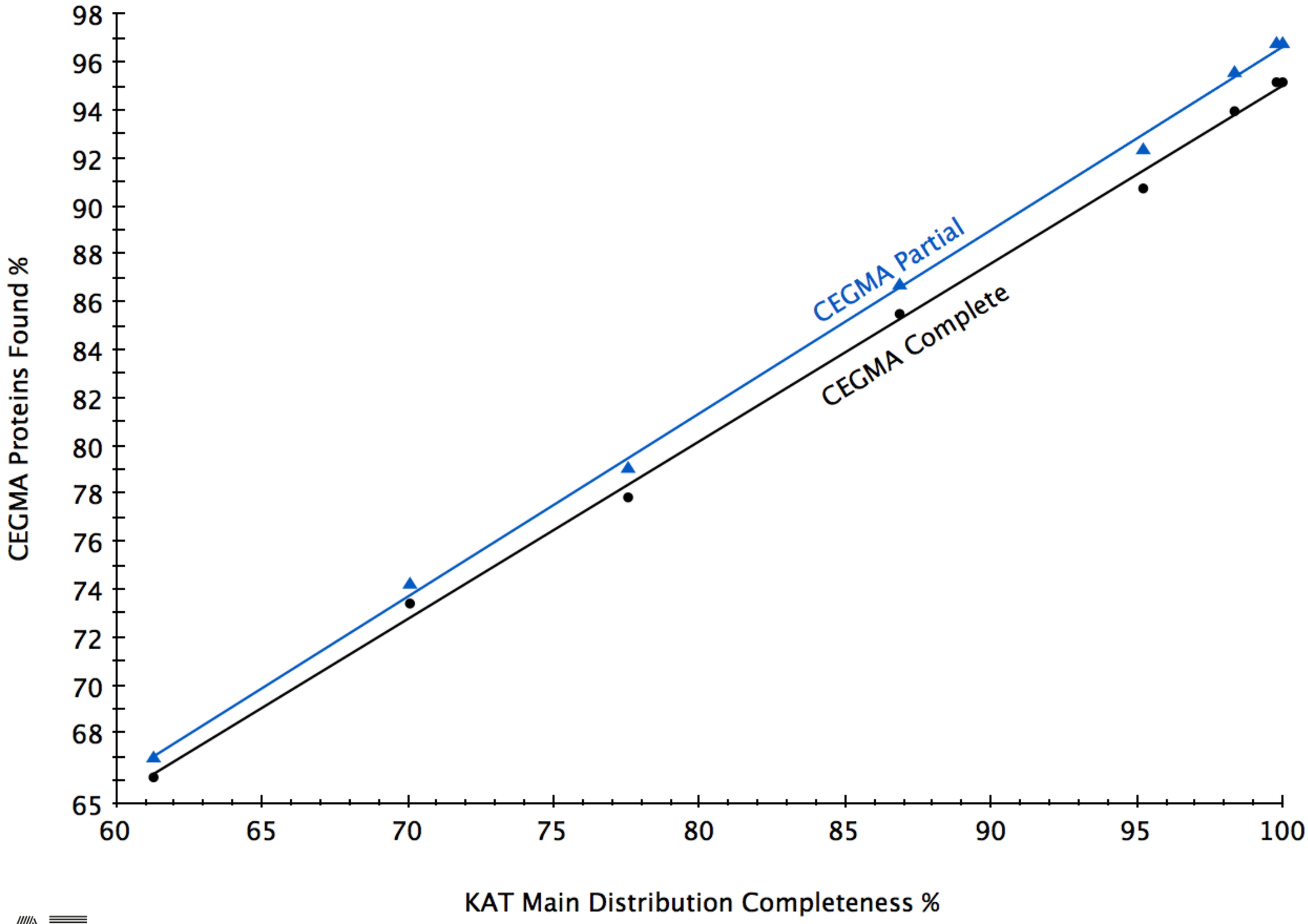
Checking content inclusion using KAT

- Just compare the frequency of kmers in the assembly to the reads spectrum.

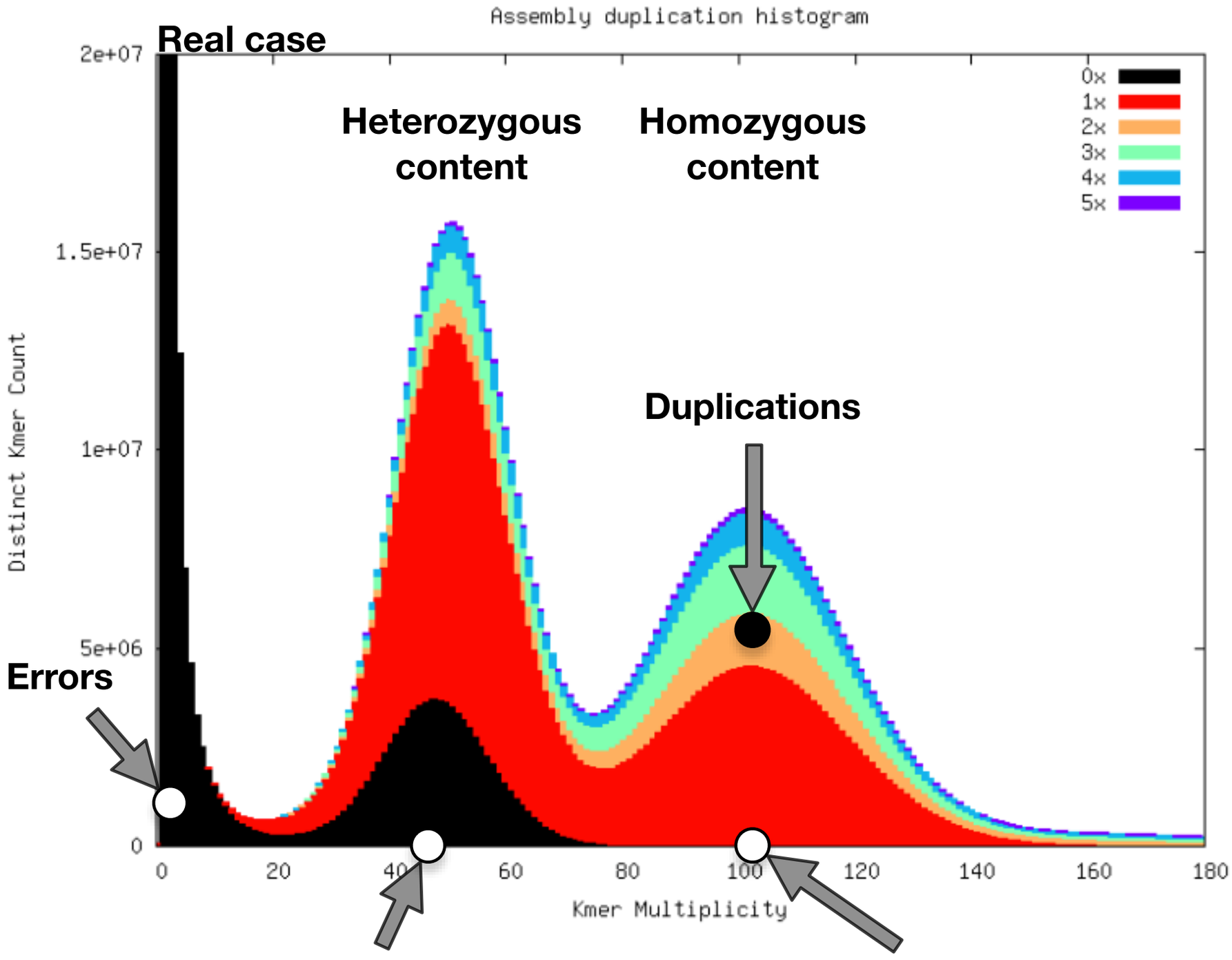


TGAC

KAT vs. CEGMA



C. fraxinea - “Ash Dieback” - NORNEX


Real case
Assembly duplication histogram
Heterozygous content
Homozygous content
Duplications
Errors
0x
1x
2x
3x
4x
5x
2e+07
1.5e+07
1e+07
5e+06
0
Distinct Kmer Count
0
20
40
60
80
100
120
140
160
180
Kmer Multiplicity
Discarded heterozygous content
Discarded homozygous content

Assembly validation

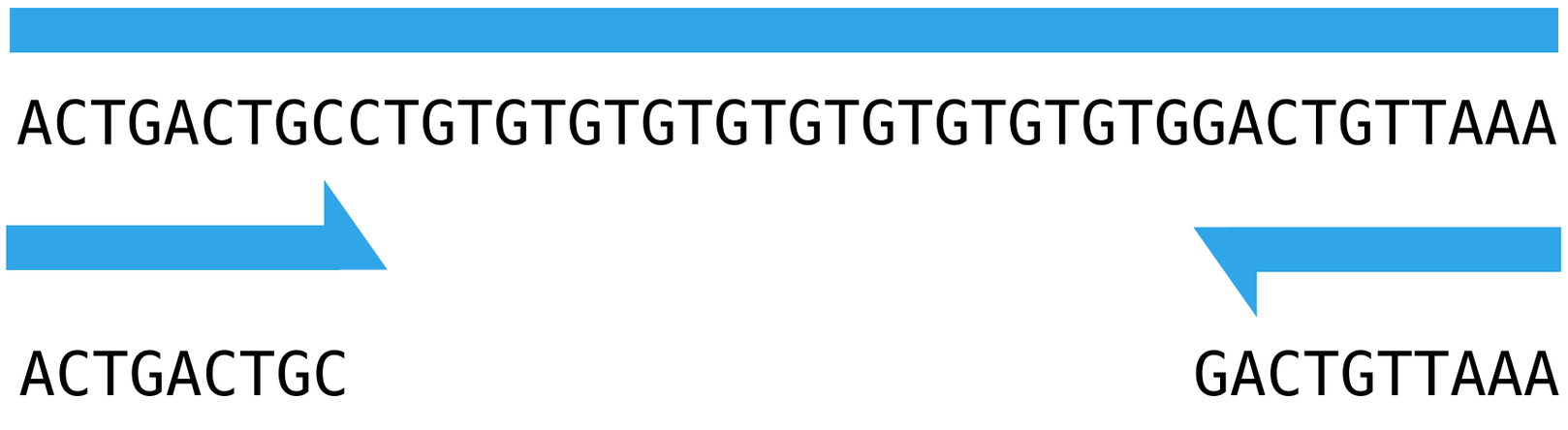
Using biological knowledge to figure out what are…

The right *motifs*,

the correct number of times,

in correct order and position.

Direct experimental evidence: the reads



The right *motifs*,

the correct number of times,

in correct order and position.

Error type	Transcripts	Assembly	Read evidence
Family collapse	geneAA geneAB geneAC n=3	n=1	bases in reads agreement 2 1 0
Chimerism	geneC geneB n=2	n=1	coverage
Unsupported insertion	n=1	n=1	no reads align to insertion
ncompleteness	n=1	n=1	read pairs align off end of conti
Fragmentation	n=1	n=4	bridging read pairs
Local misassembly	n=1	n=1	read pairs in wrong orientation
Redundancy	n=1	n=3	all reads assign to best contig

Direct experimental evidence: other evidence

- Genome size, ploidy
- GC content
- Symbionts
- Plastids
- ESTs, cDNAs, peptides, genome walking

The right *motifs,*

the correct number of times,

in correct order and position.

Indirect experimental evidence: genomes in general

- Genes! They have structure
- Repeats
- Chromosome macrostructure
  - (circular?, number, telomeres, …)

The right *motifs*,

the correct number of times,

in correct order and position.

Indirect experimental evidence: other species

- Close relatives: proteins, transcripts, genomes
- Distant relatives: single-copy genes,
  phylogeny, HGT

The right *motifs*,

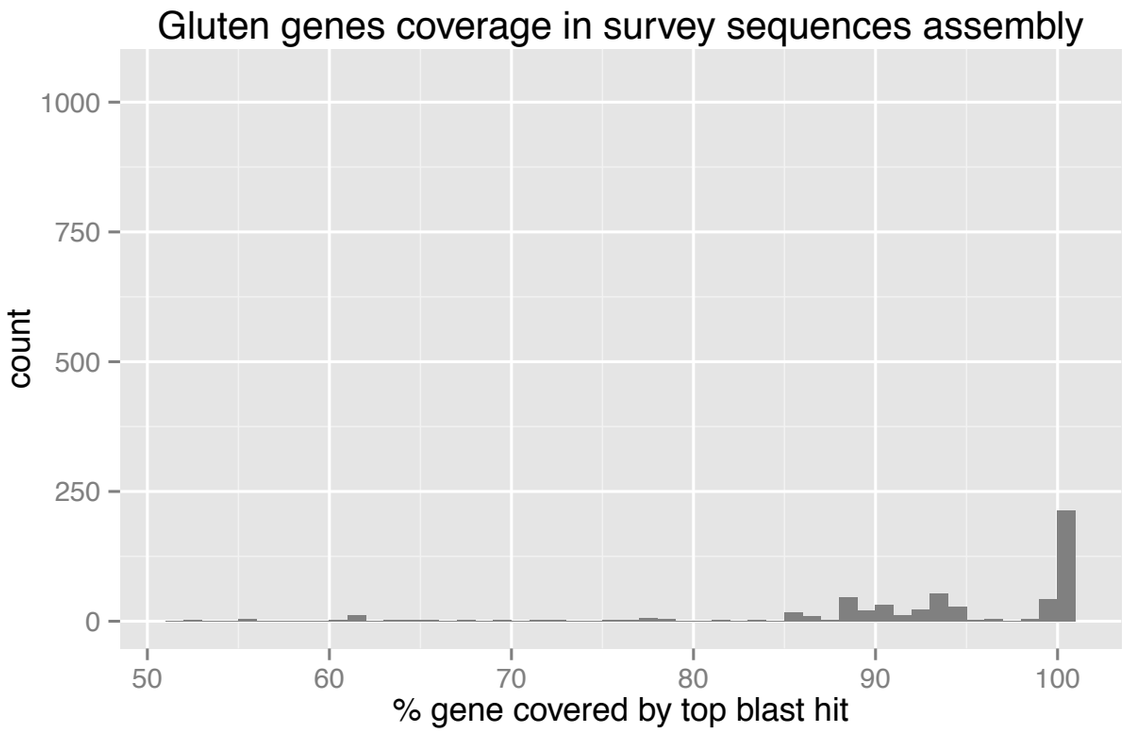
the correct number of times,

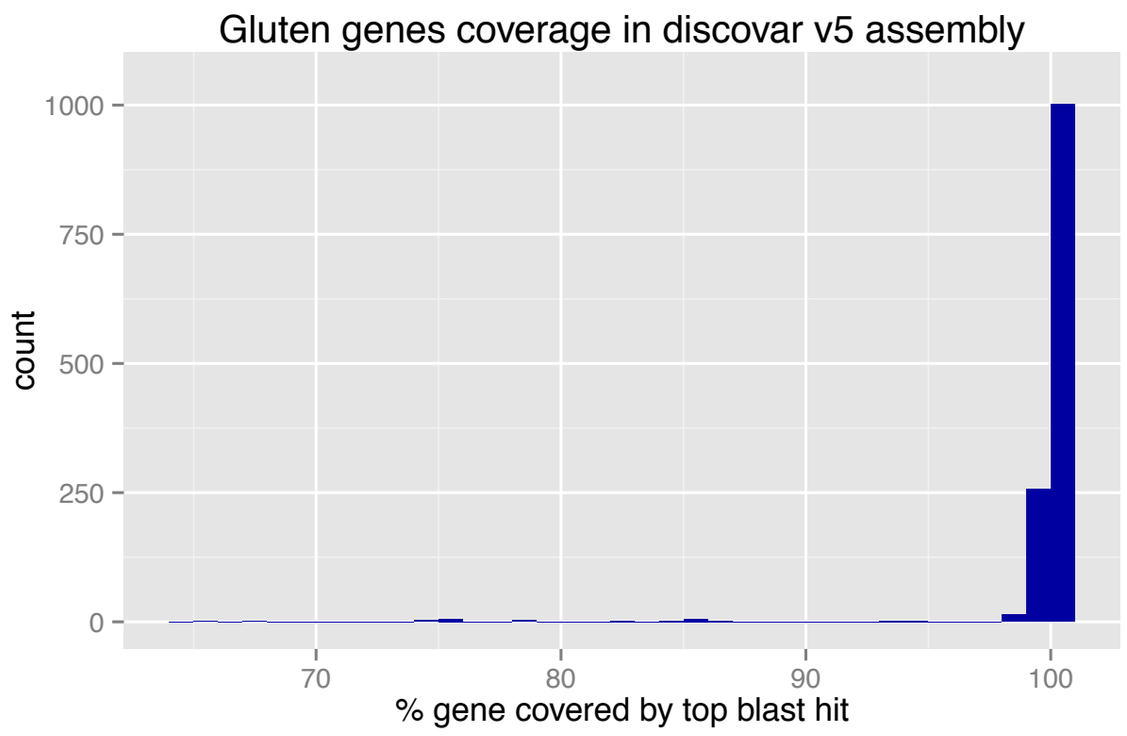
in correct order and position.

Examples from the wheat genome

```
>gi|146261042|gb|AB014771.1| HMW glutenin subunit 1By16 [Triticum aestivum]
MAKRLVLFATVVITLVALTAAEGEASRQLQCEREL QESSLEACR QVVD QQLAGRLPWSTGL QMRCC QQLR
DVSAKCRPVAVS QVVR QYE QIVVPPKGGSFYPGETTPL QQL QQVIFWGTSS QTV QGYYPSVSSP QQGPYY
PG QASP QQPG QG QQPGKW QELG QG QQGYYPTSLH QSG QG QQGYYPSSL QQPG QG QQTG QG QQGYYPTSL Q
QPG QG QQIG QG QQGYYPTSP QHPG QR QQPG QG QQIG QE QQLG QGR QIG QG QQSG QG QQGYYPTSP QQLG Q
G QQPG QW QQSG QG QQGYYPTS QQQPG QG QQG QYPAS QQQPG QG QQG QYPAS QQQPG QG QQG QYPAS QQQP
A QG QQG QYPAS QQQPG QG QQGHYLAS QQQPG QG QQRHYPASL QQPG QG QQGHYTASL QQPG QG QQGHYPA
SL QQVG QG QQIG QLG QR QQPG QG QQTR QG QQLE QG QQPG QG QQTR QG QQLE QG QQLE Q
G QQPG QG QQGYYPTSP QQSG QG QQPG QS QQPG QG QQGYYSTSL QQPG QG QQGHYPASL QQPG QGHPG QR Q
QPG QG QQPK QG QQPG QG QQGYYPTSS QQPG QGK QLG QG QQGYYPTSP QQPG QG QQPG QG QQGHCPTSP QQ
TG QA QQPG QG QQIG QV QEPG QG QQGYYPISL QQSG QG QQSG QG QQSG QGH QLG QG QQSG QE QQGYDNPYH
VNTE QQTASPKVAKV QQPAT QLPIMCRMEGGDALSTR Q
```

~1000 gluten gene classes are represented:
a-,b-, g- gliadins and high
and low molecular weights glutenins





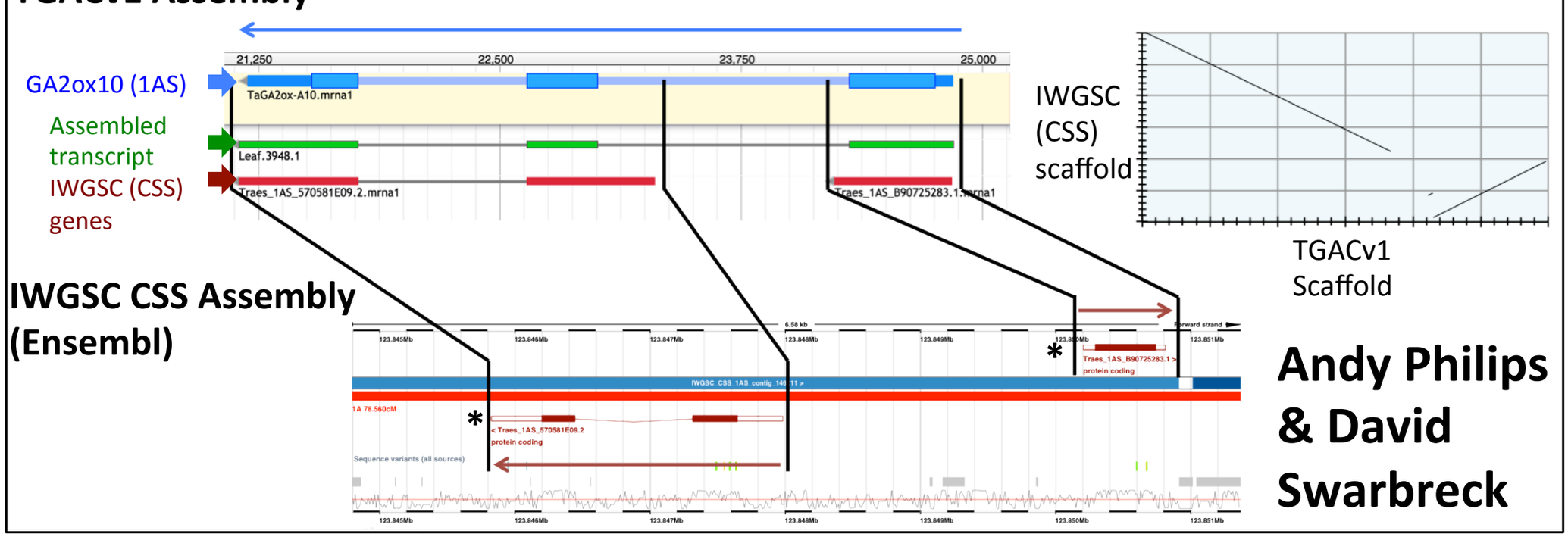
Ksenia Krasileva

Examples from the wheat genome (II)

- Gibberellin (GA) pathway plays a central role in plant development.
- 84 genes in the pathway (including homoeologues).
  - **25 genes (30%)** are found as full-length sequences in **CSS**
  - **79 of them (94%)** are full-length in **TGACv1**

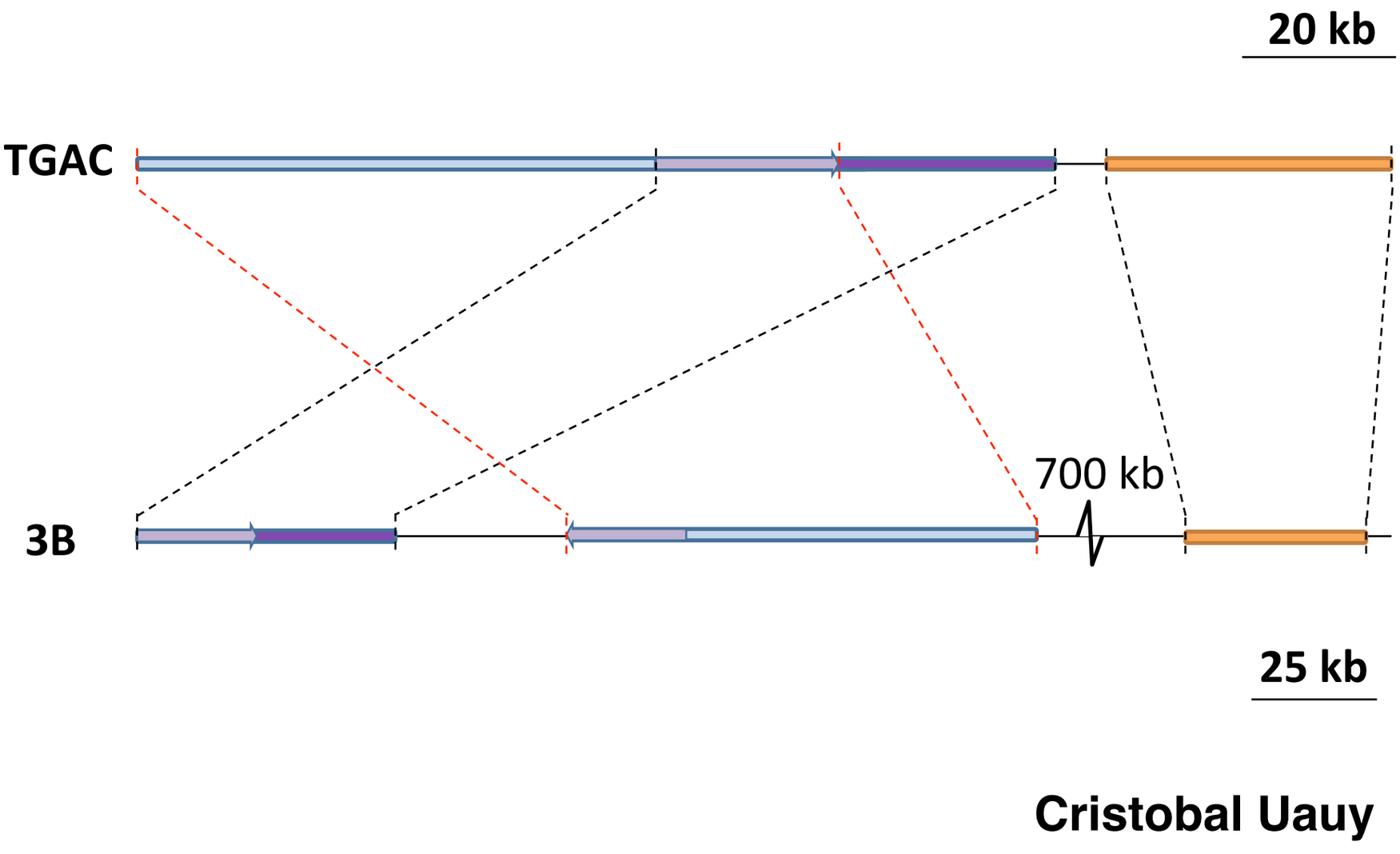
A) GA2ox10 (1AS) is incorrectly represented as two genes (*) on opposite strands in the IWGSC CSS assembly

B) IWGSC (CSS) assembly shows a inversion relative to the TGACv1 assembly in the region covering exon 1 of GA2ox10 (1AS)



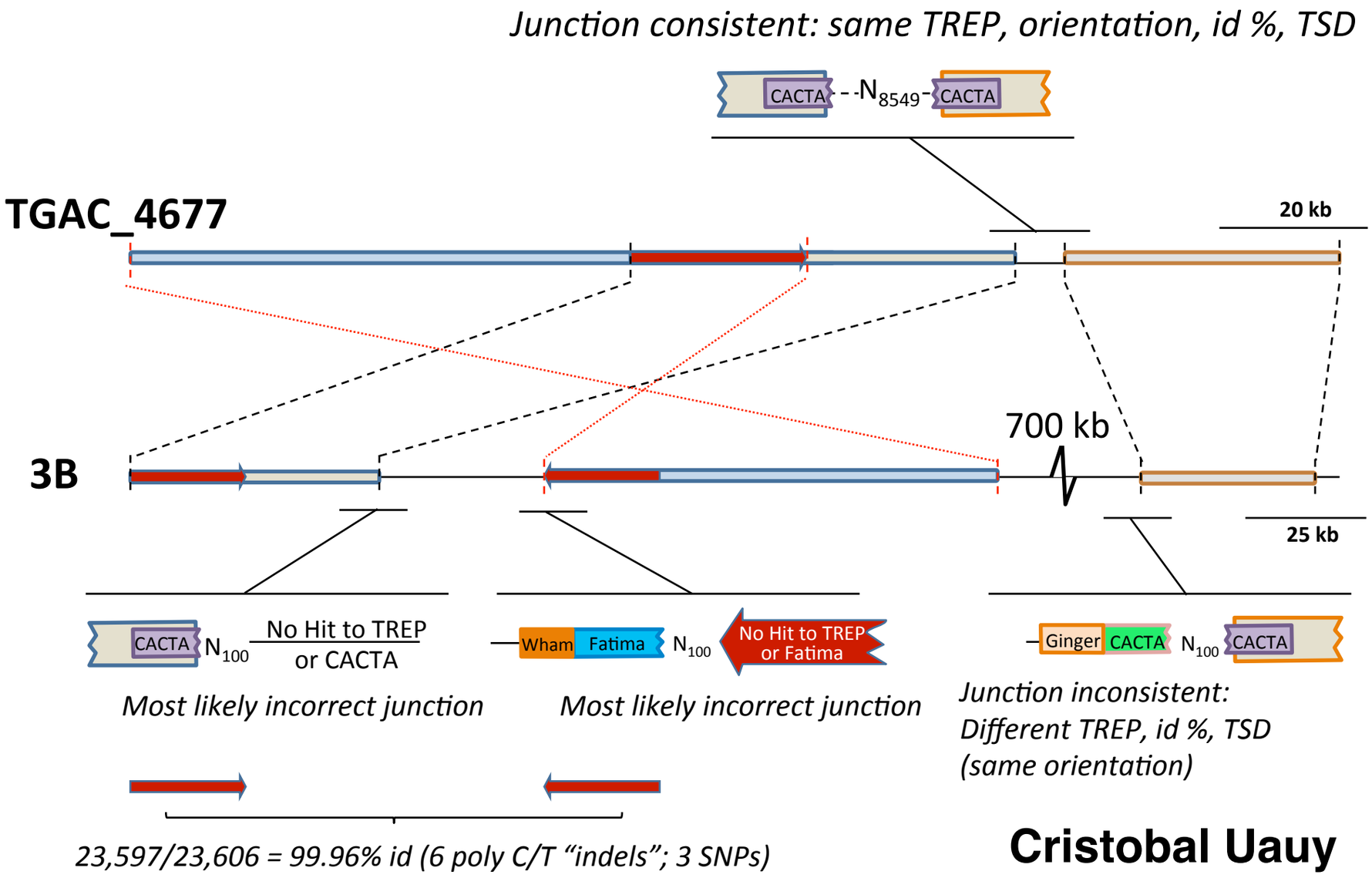
Andy Philips & David Swarbreck

Examples from the wheat genome (III)



Cristobal Uauy

Examples from the wheat genome (III)



Questions?