Experimental design for long read sequencing

If you have data, or will have data soon, please go to **"Experimental design"** in the wiki and fill out the spreadsheet there. We will take a look later.

The trade-off

Genome size

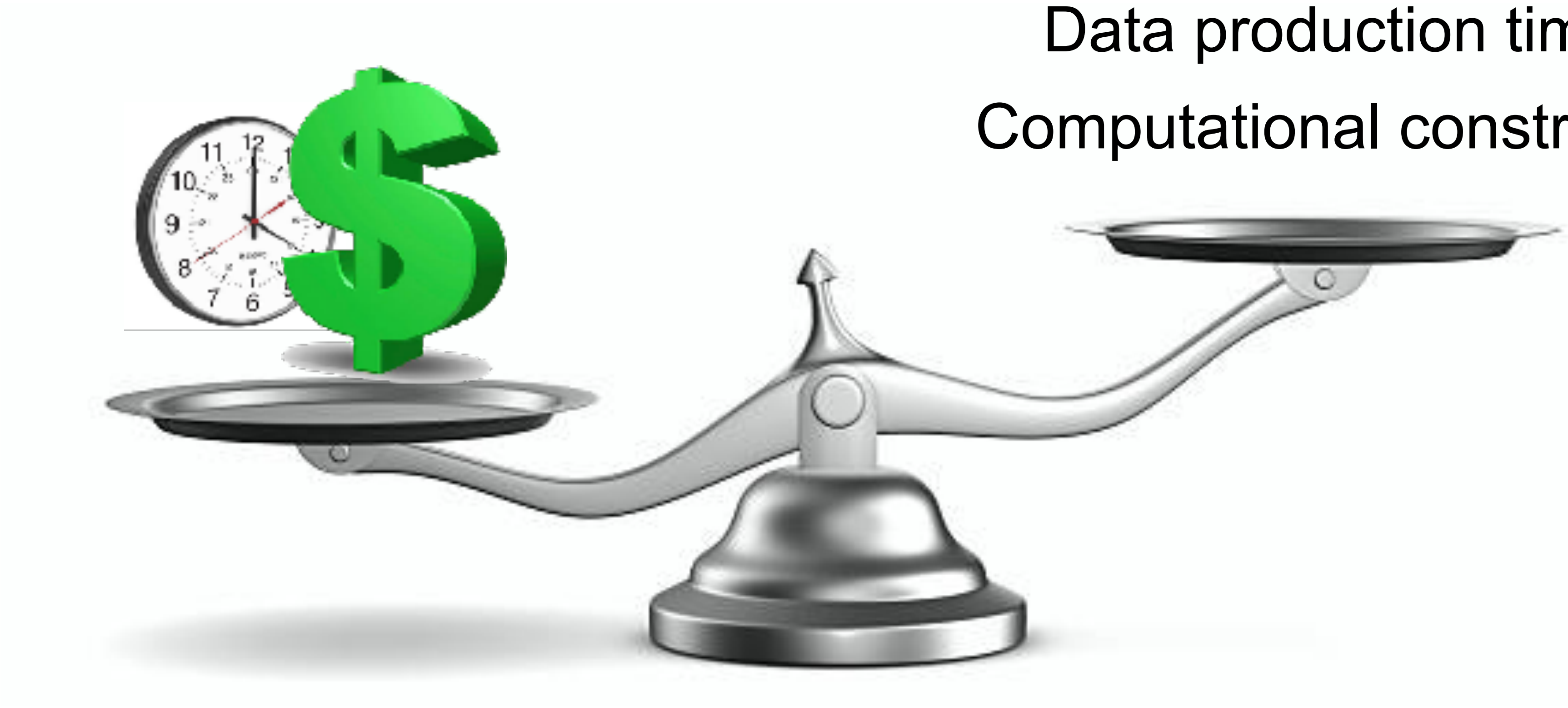
Heterozygosity

Repeat content

Material availability

Data production time

Computational constraints



Genome size

↑ Genome = ↑ Sequencing

(for same coverage)

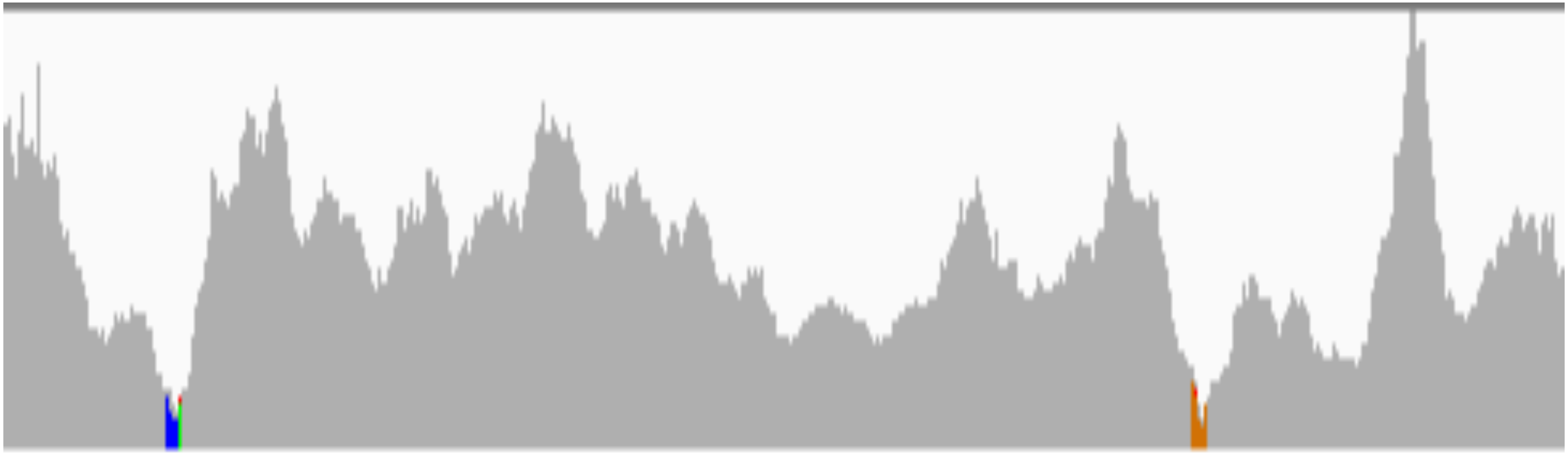
http://www.genomesize.com/

ANIMAL GENOME SIZE DATABASE

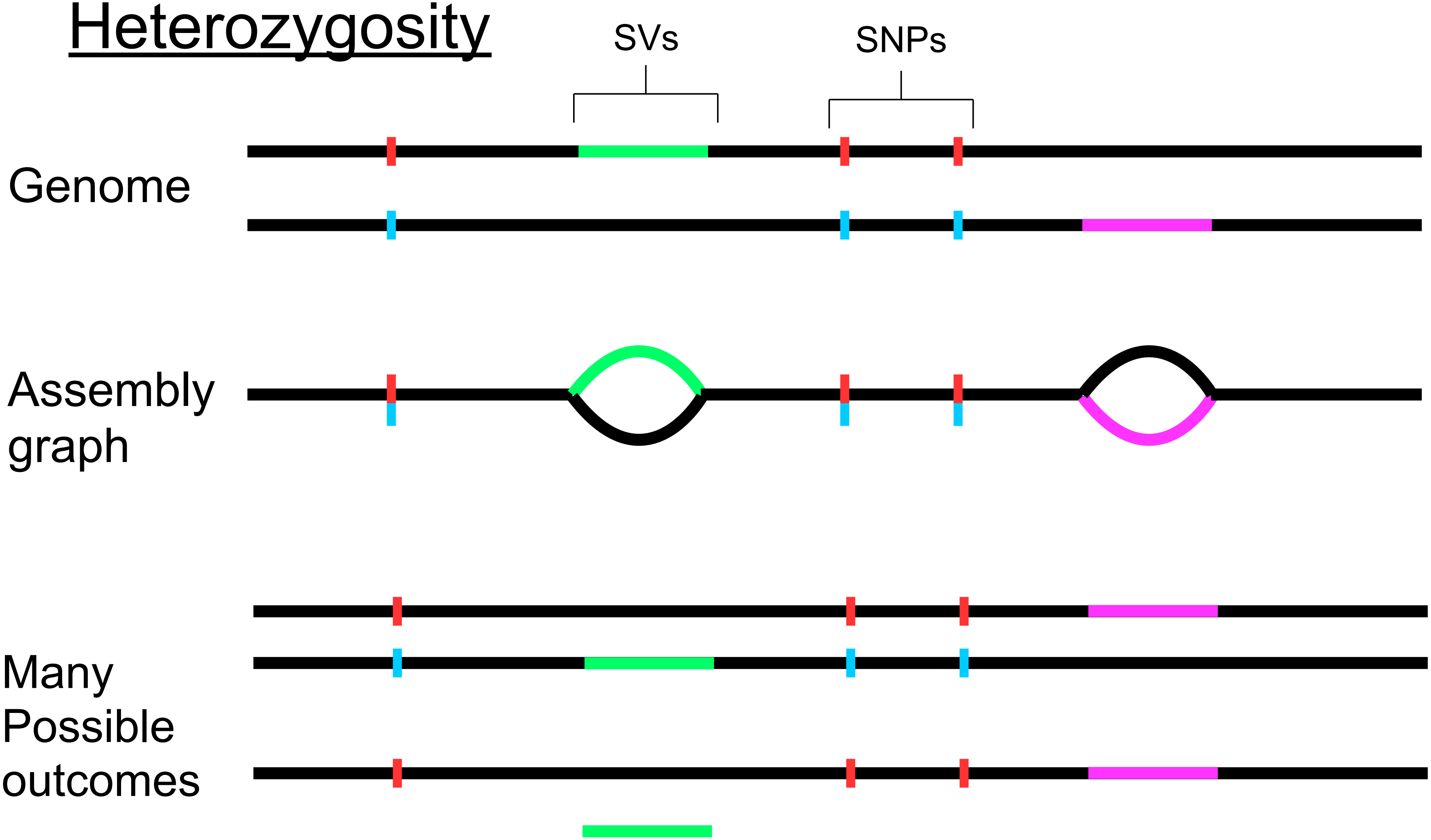
HOME

[illegible]

Remember . . . . coverage is anything but uniform



Heterozygosity



SVs

SNPs

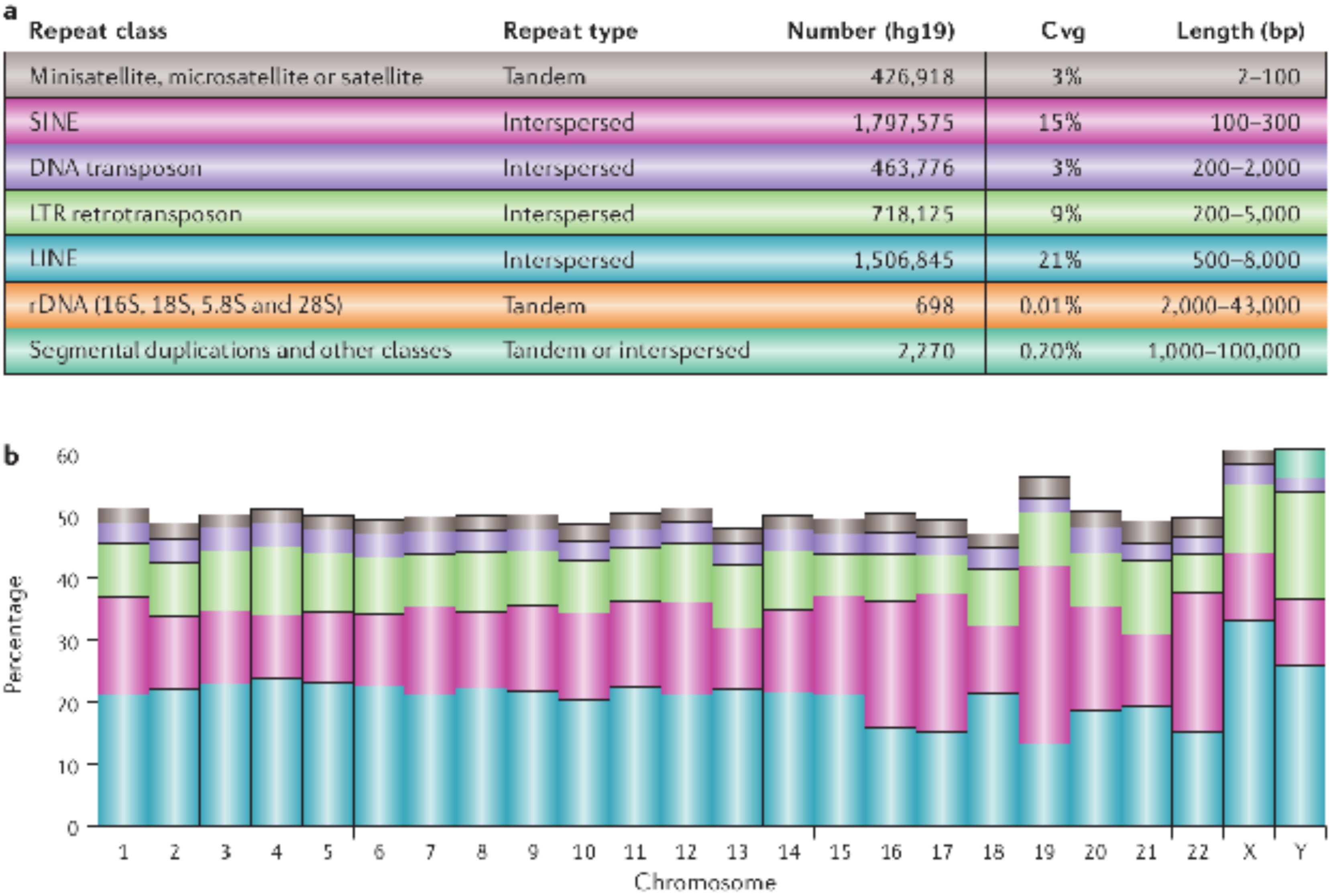
Genome

Assembly graph

Many Possible outcomes

Retrieving true haplotypes requires adequate sequencing depth for BOTH haplotypes!

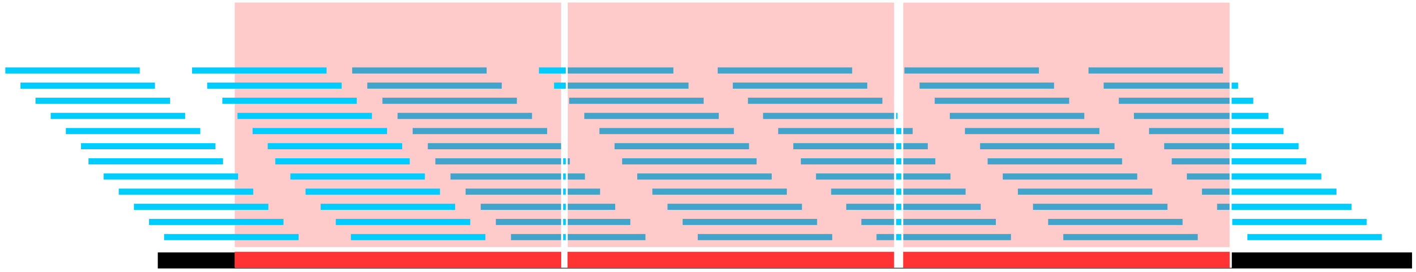
Repeat content



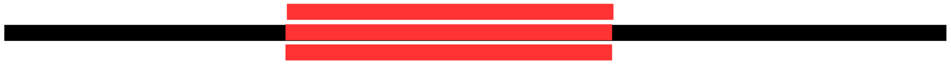
Repeat class	Repeat type	Number (hg19)	Cvg	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426,918	3%	2–100
SINE	Interspersed	1,797,575	15%	100–300
DNA transposon	Interspersed	463,776	3%	200–2,000
LTR retrotransposon	Interspersed	718,125	9%	200–5,000
LINE	Interspersed	1,506,845	21%	500–8,000
rDNA (16S, 18S, 5.8S and 28S)	Tandem	698	0.01%	2,000–43,000
Segmental duplications and other classes	Tandem or interspersed	2,270	0.20%	1,000–100,000

Treangen & Salzberg (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet.

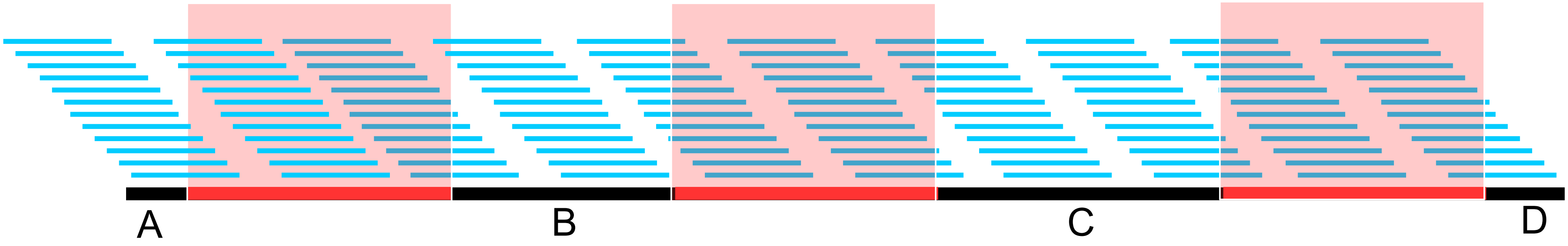
Tandem repeats



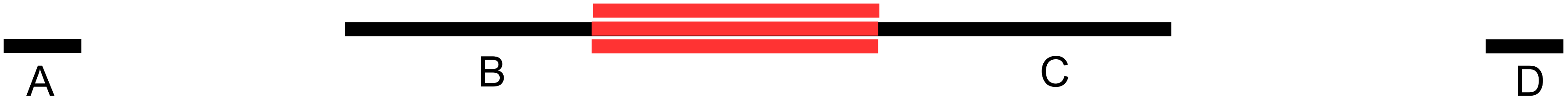
Possible (wrong) assembly - collapse repeats, split unique sequence into different contigs.



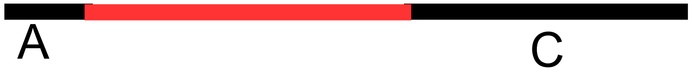
Interspersed repeats



Possible (wrong) assembly - collapse repeats, split unique sequence into different contigs.

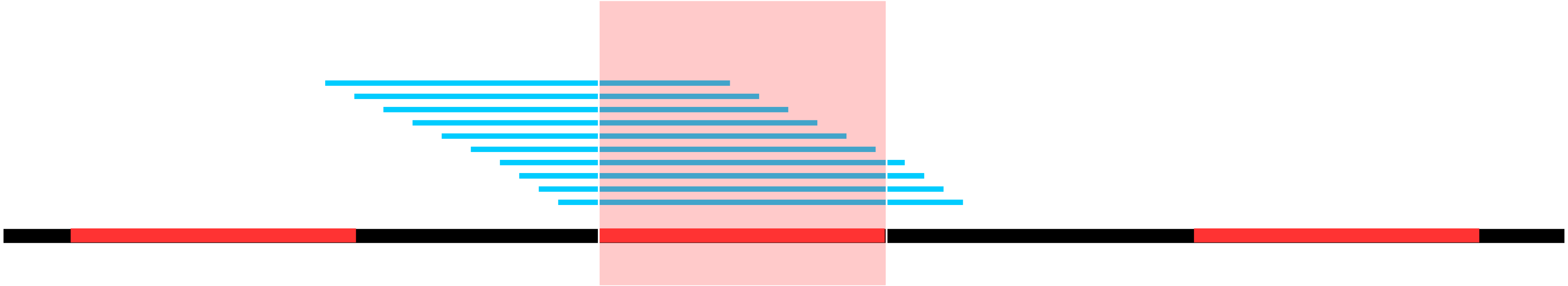


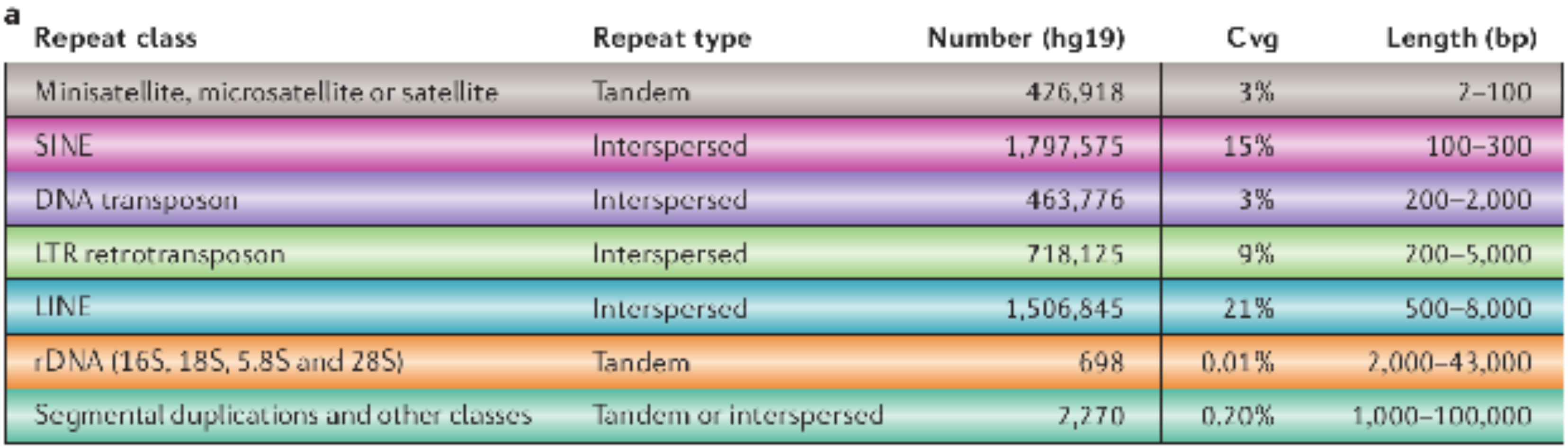
Possible (wrong) assembly - Chimeric contigs



Long reads to the rescue

The longer your reads, the better chance of resolving these regions



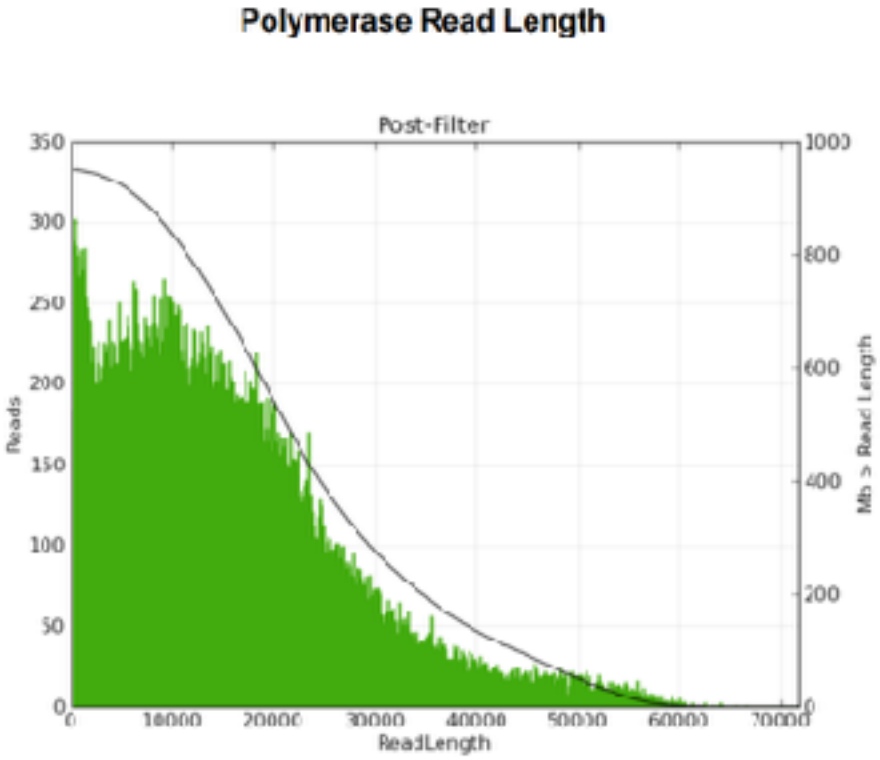


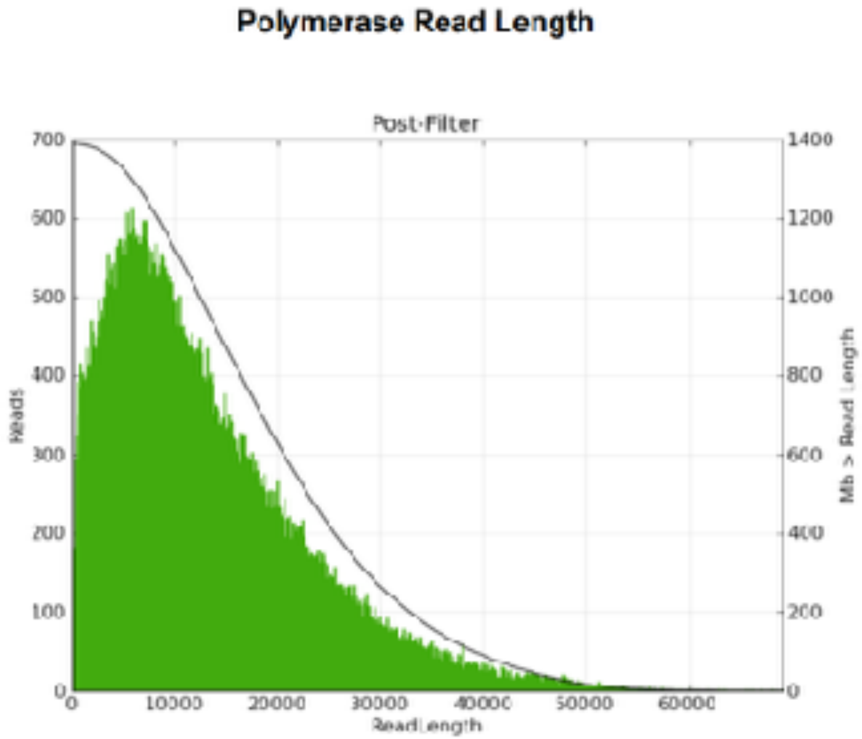
a

Repeat class	Repeat type	Number (hg19)	Cvg	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426,918	3%	2–100
SINE	Interspersed	1,797,575	15%	100–300
DNA transposon	Interspersed	463,776	3%	200–2,000
LTR retrotransposon	Interspersed	718,125	9%	200–5,000
LINE	Interspersed	1,506,845	21%	500–8,000
rDNA (16S, 18S, 5.8S and 28S)	Tandem	698	0.01%	2,000–43,000
Segmental duplications and other classes	Tandem or interspersed	2,270	0.20%	1,000–100,000

Material availability

Lots of high molecular weight DNA required!





———- > Frozen once ———->

13.9 kb mean subread length

12 kb mean subread length

Data production time . . .

Computational constraints

UNIL has 6000 cores, 6 petabytes of storage

Yet my 4.5 Gb genome with x50 coverage would have taken >4 months to assemble

So I paid some people lots of money to do it

DNAnexus