NORWEGIAN SEQUENCING CENTRE

What does it mean to do bioinformatics?

Lex Nederbragt, NSC and CEES

lex.nederbragt@ibv.uio.no

@lexnederbragt

Oslo universitetssykehus

The Research Council of Norway

CEES Centre for Ecological and Evolutionary Synthesis

HELSE SØR-ØST

UNIVERSITAS OSLOENSIS MDCCCXI

Who am I

Researcher at
the Centre for Ecological and Evolutionary Synthesis

CEES

Centre for Ecological and
Evolutionary Synthesis

Who am I

20% teaching position at
the Institute for Informatics

Who am I

Bioinformatician at the Norwegian Sequencing Centre

This presentation

Part 1: Defining bioinformatician

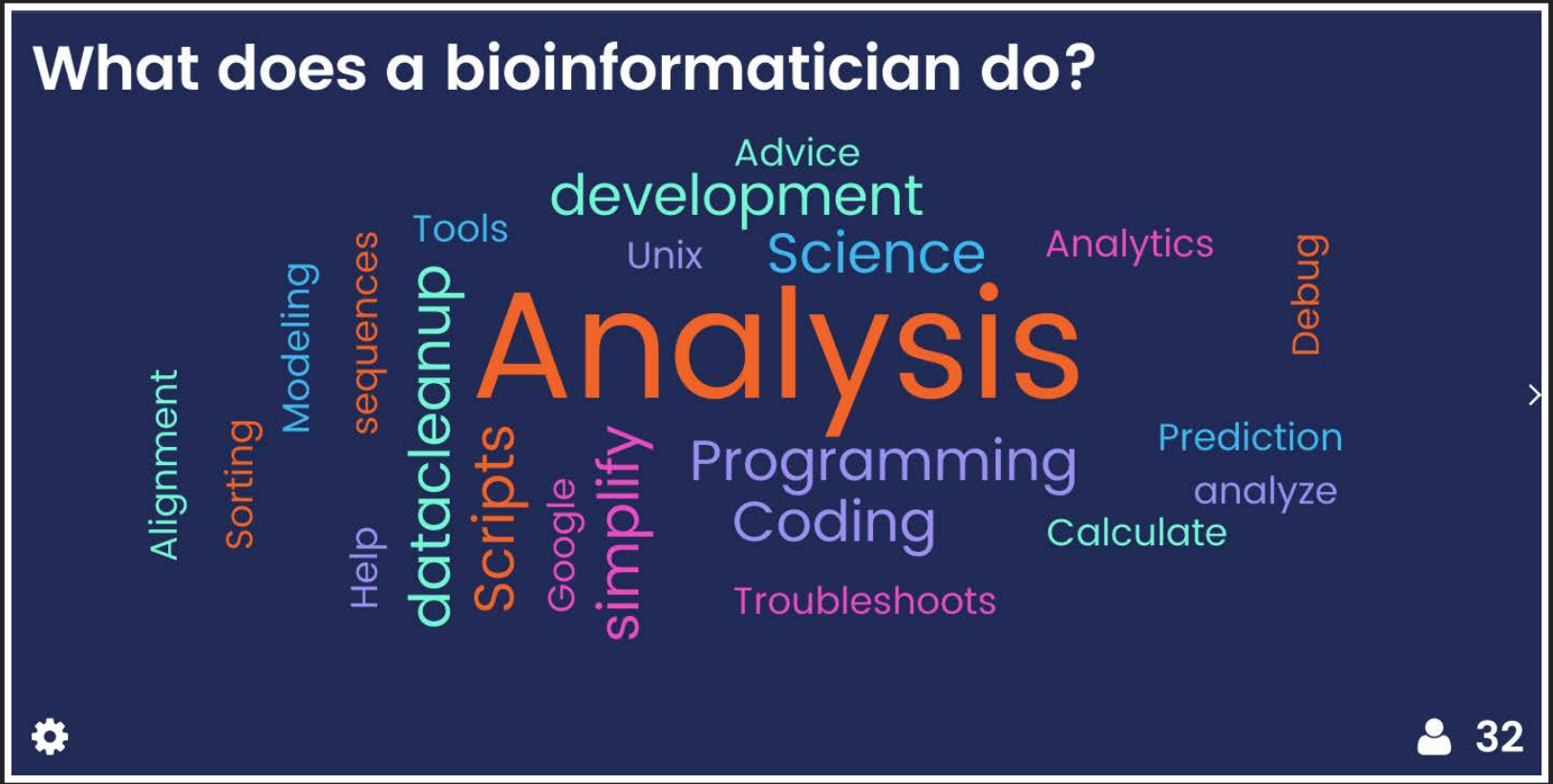
Part 2: How I became a bioinformatician

Part 3: What I think of bioinformatics

Part 4: How to become an efficient bioinformatician

Part 1: Defining bioinformatician

What does a bioinformatician do



What does a bioinformatician do

Biological data

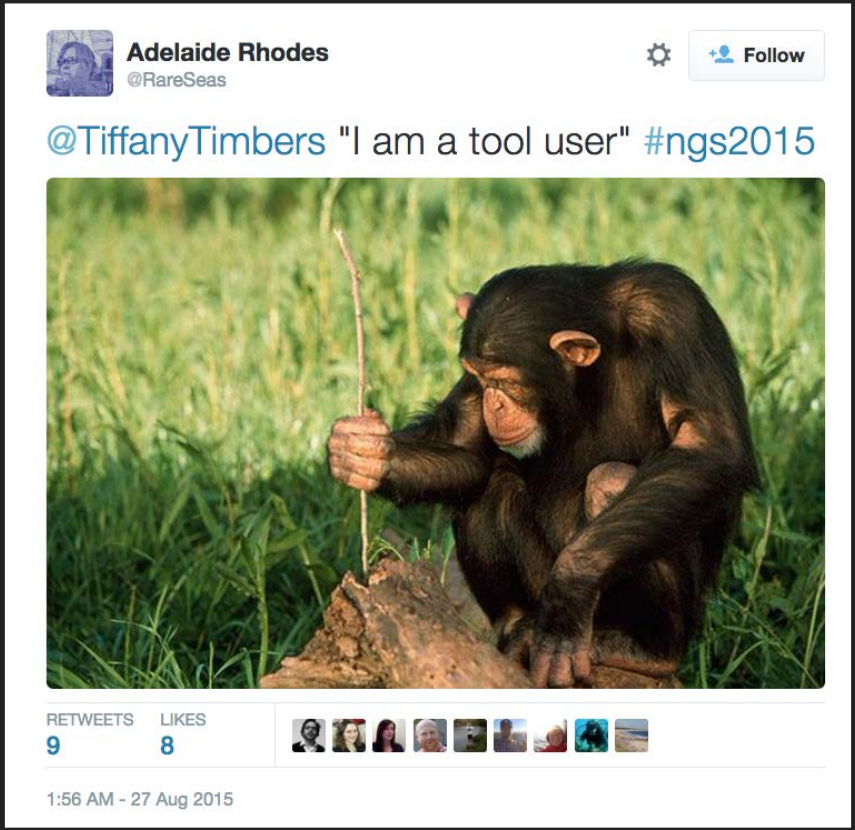
+

Software tools/statistics

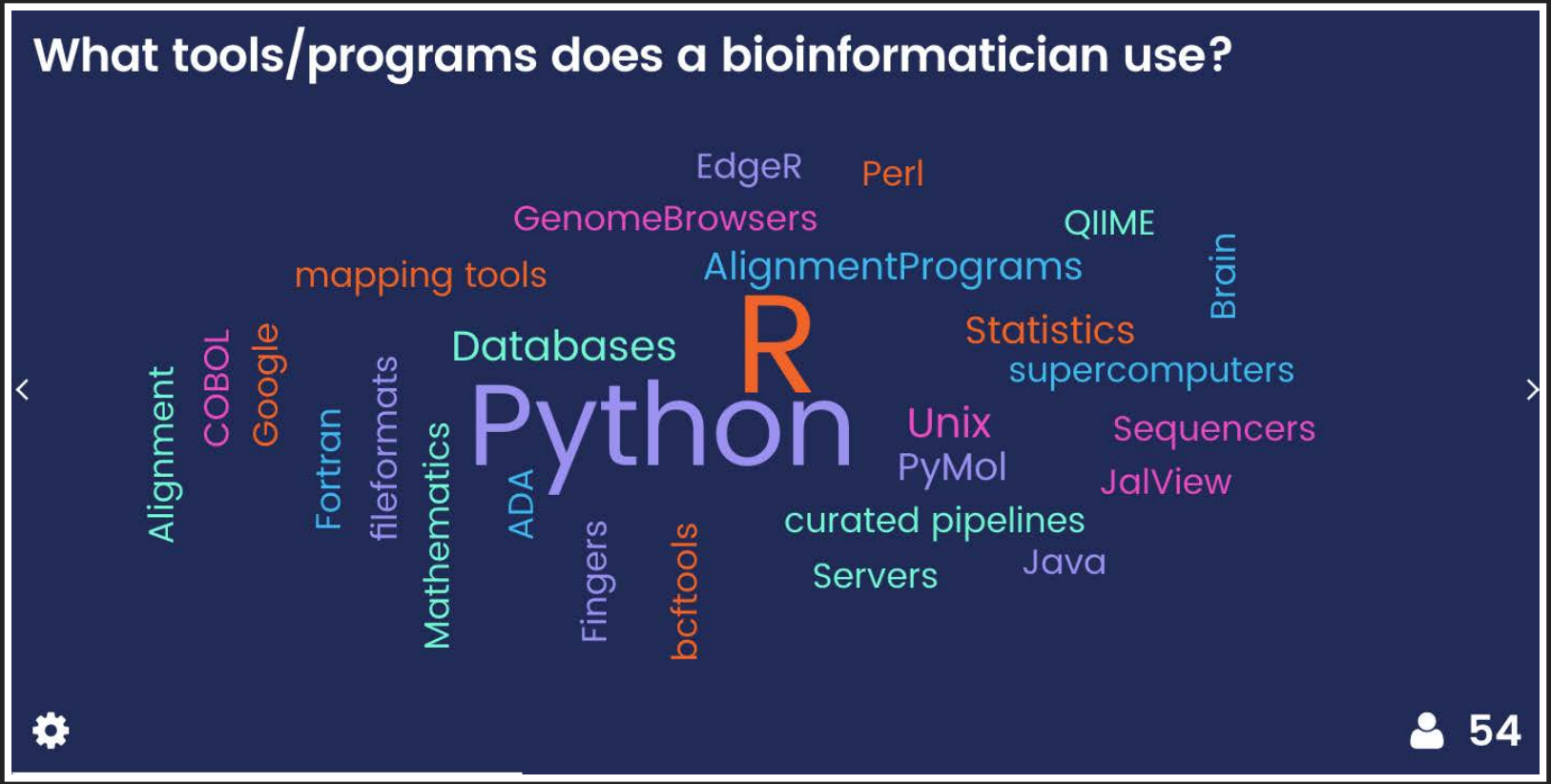
=

Biological knowledge

Tool developer versus tool user



What tools/programs does a bioinformatician use



Part 2: How I became a bioinformatician

2007: a grant

GS FLX from Roche/454

?

Genome Analyzer from Solexa/Illumina

Let’s try them out!

Specimen

Planktothrix rubescens NIVA CYA 98

Cyanobacteria

(blue-green algae)

Experiment

Half a million reads
Average length 250 nt

Assembly?

10 million reads
33 nucleotides each

Assembly?

Why is genome assembly so hard?

What is a genome assembly?

A hierarchical data structure

that maps the sequence data

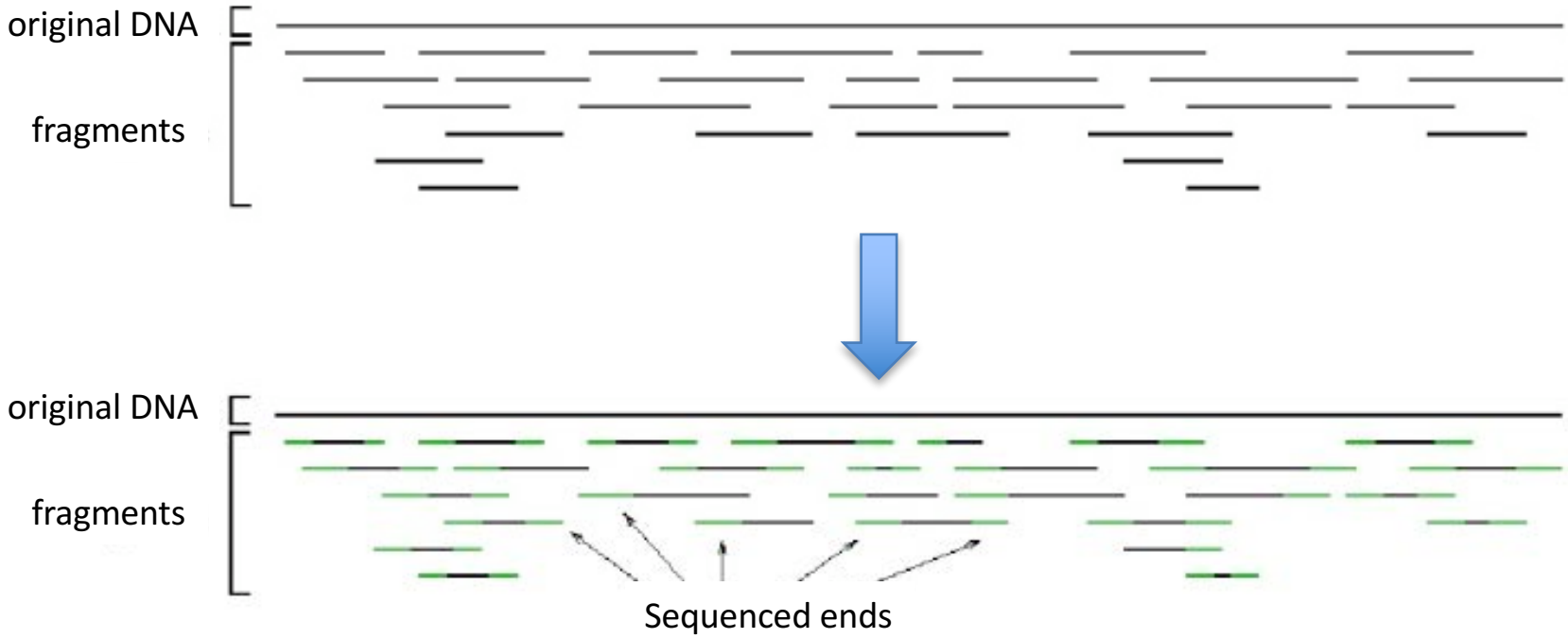
to a putative reconstruction of the target

Miller et al 2010, Genomics 95 (6): 315-327

Hierarchical structure

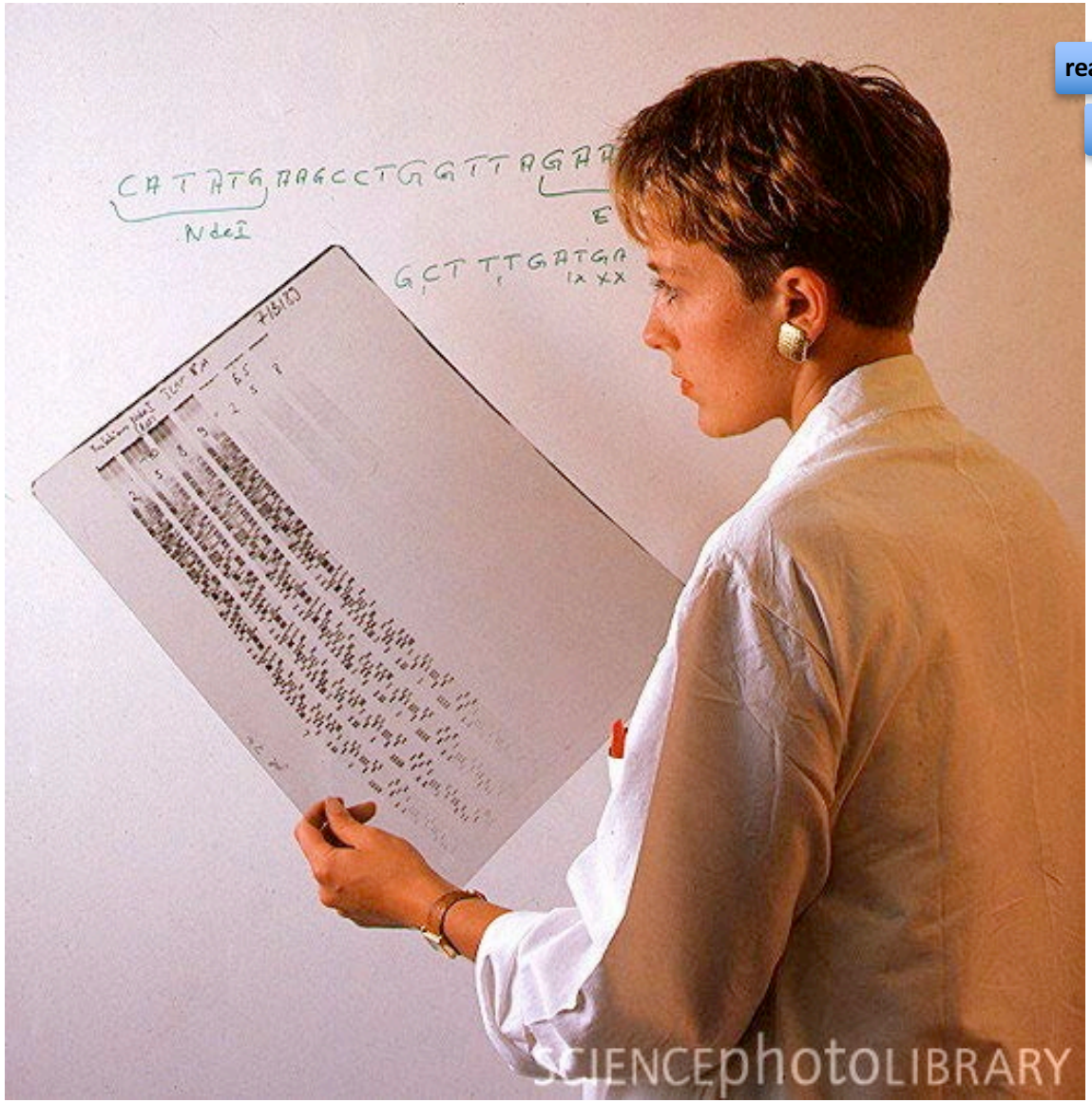
Sequence data

Reads



http://www.cbcb.umd.edu/research/assembly_primer.shtml

Reads!



reads

contigs

scaffolds

http://www.sciencephoto.com/media/210915/enlarge

Contigs

Building contigs

Aligned reads

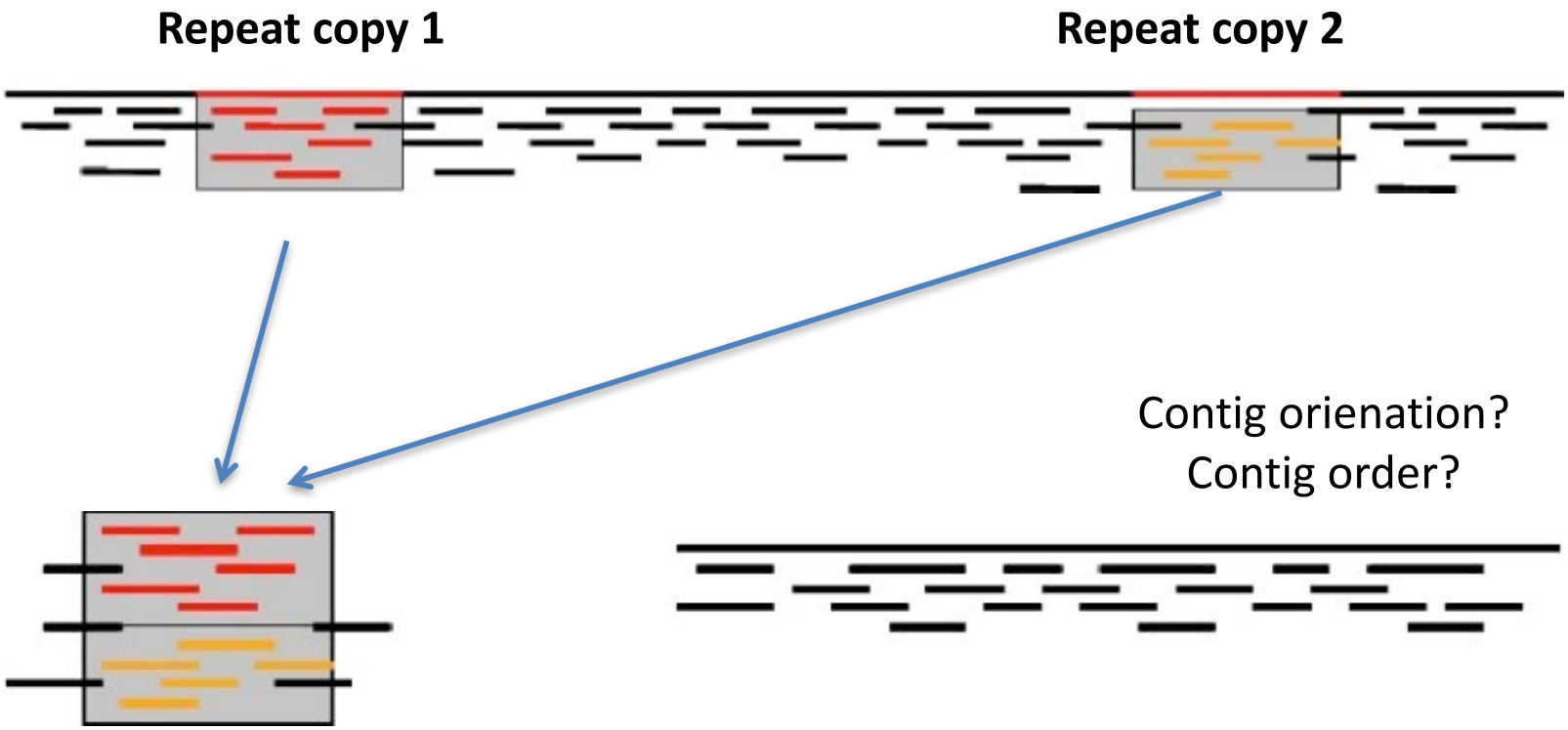
```
ACGCGATTCAGGTTACCACG
  GCGATTCAGGTTACCACGCG
    GATTCAGGTTACCACGCGTA
      TTCAGGTTACCACGCGTAGC
        CAGGTTACCACGCGTAGCGC
          GGTTACCACGCGTAGCGCAT
            TTACCACGCGTAGCGCATTA
              ACCACGCGTAGCGCATTACA
                CACGCGTAGCGCATTACACA
                  CGCGTAGCGCATTACACAGA
                    CGTAGCGCATTACACAGATT
                      TAGCGCATTACACAGATTAG
```

Consensus contig

```
ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG
```

Contigs

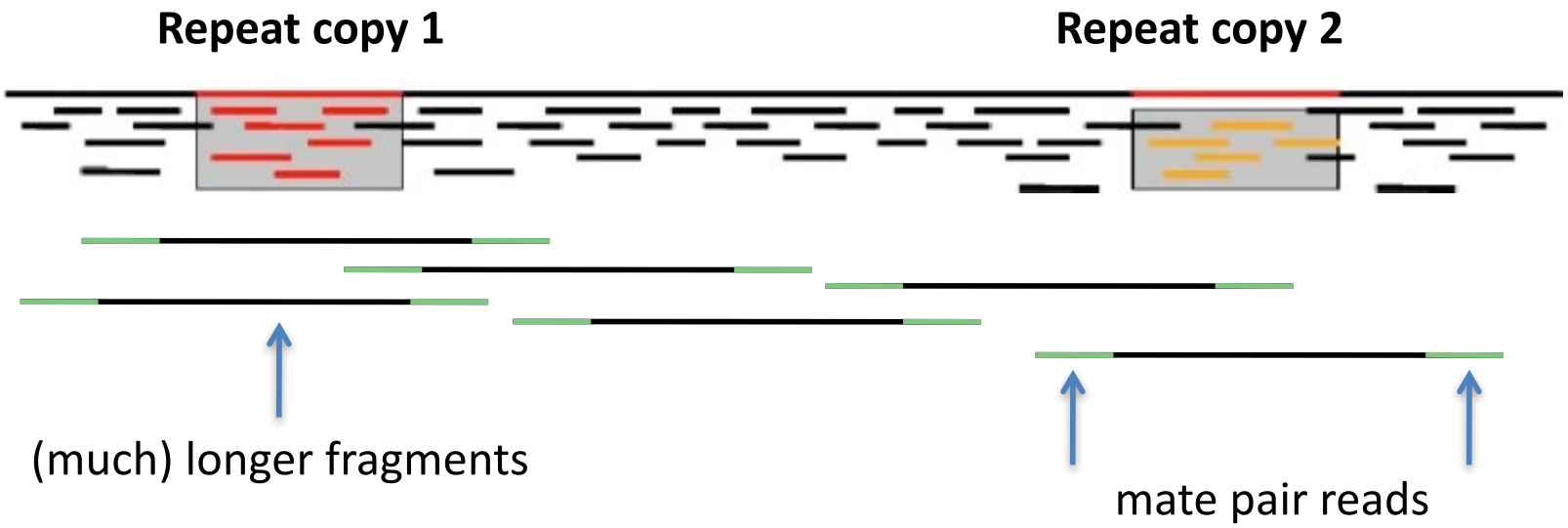
Building contigs



http://www.cbcb.umd.edu/research/assembly_primer.shtml

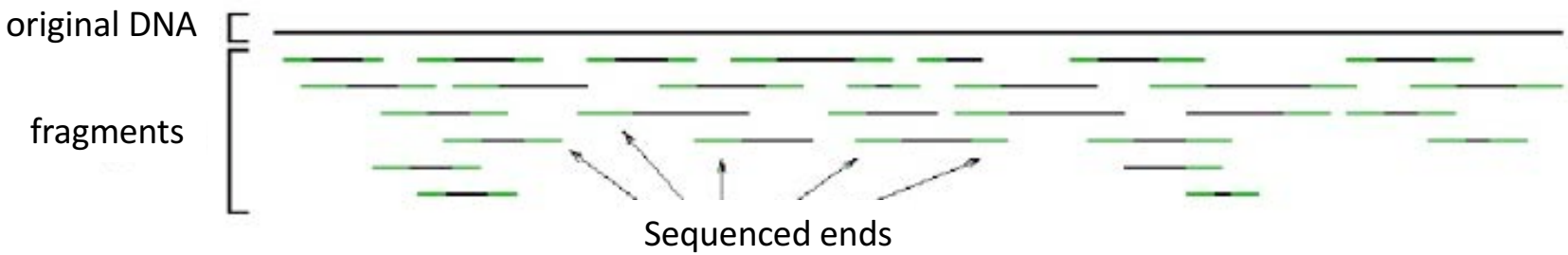
Mate pairs

Other read type

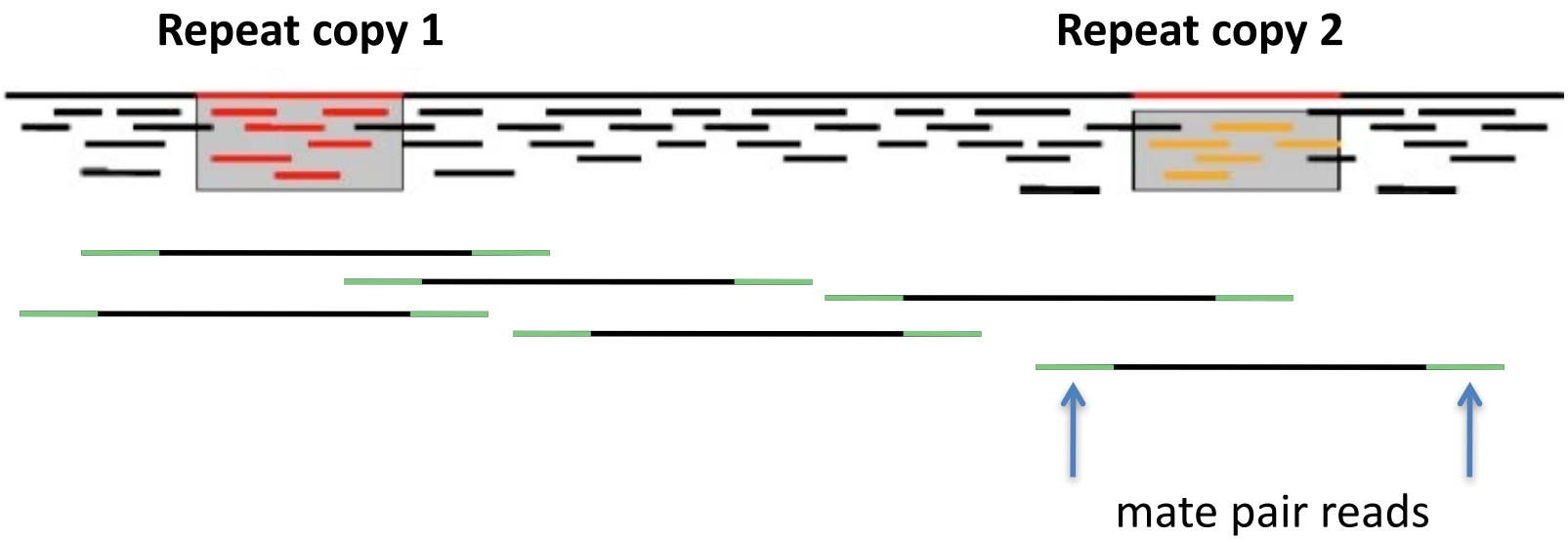


Mate pairs

Paired end reads → 100-500 bp insert

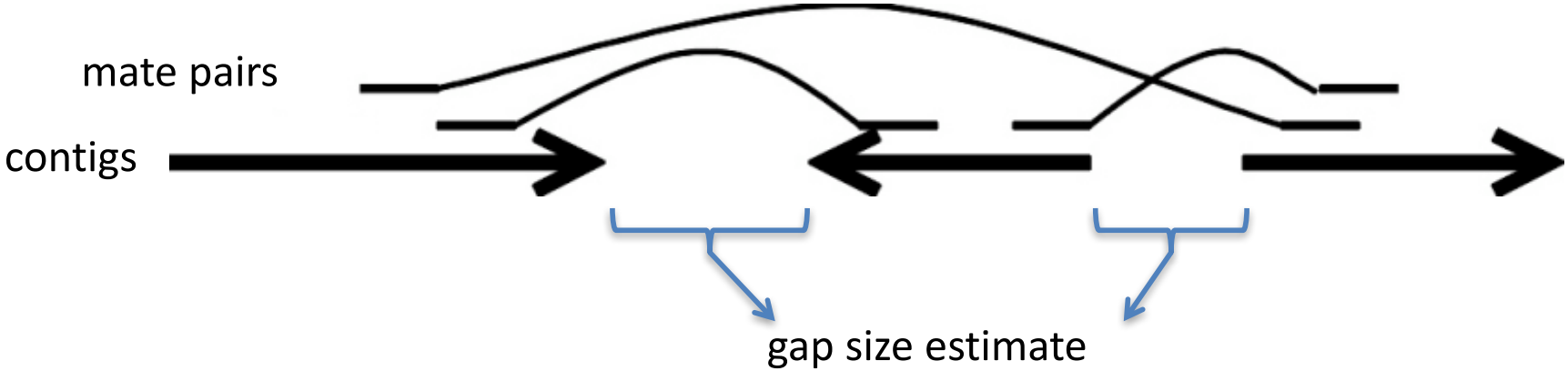


Mate pairs → 2-20 kb insert

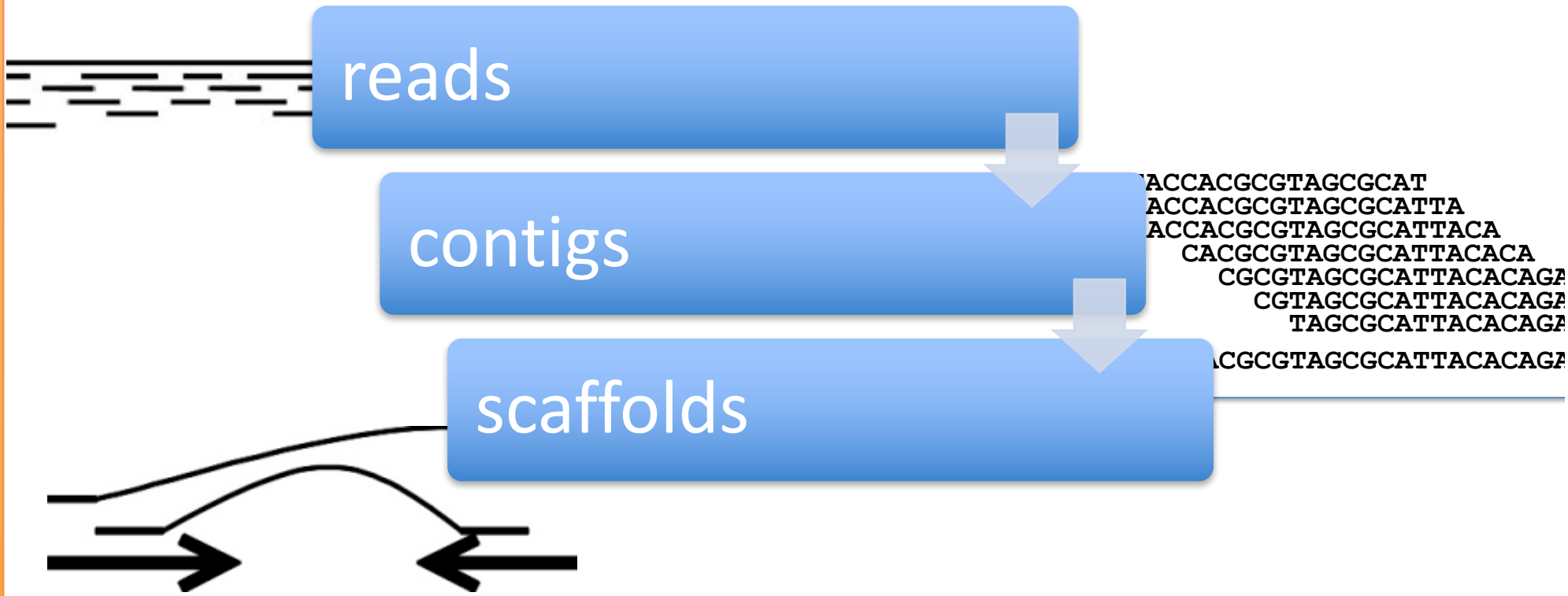


Scaffolds

Ordered, oriented contigs



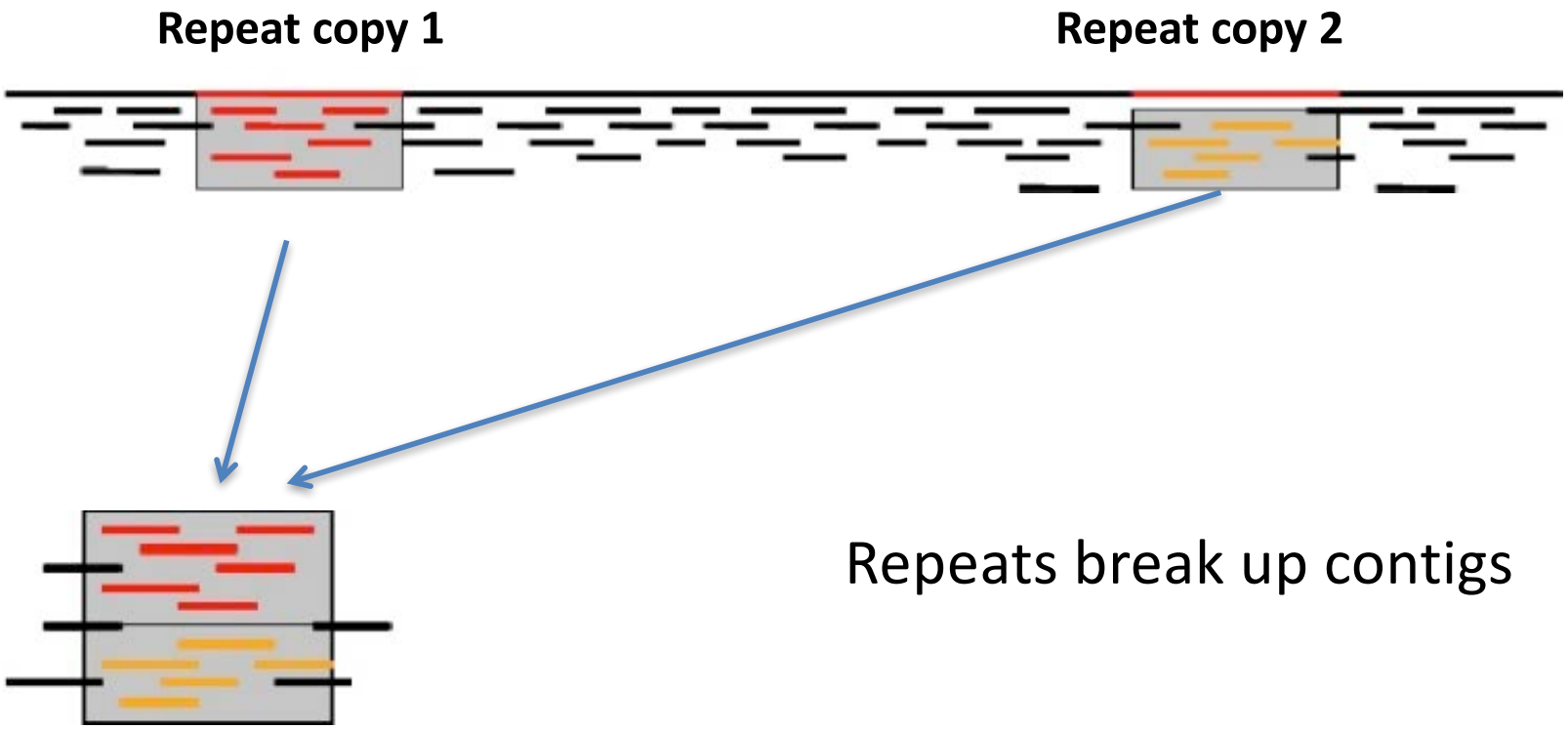
Hierarchical structure



Genome assembly

So, what's so hard about it?

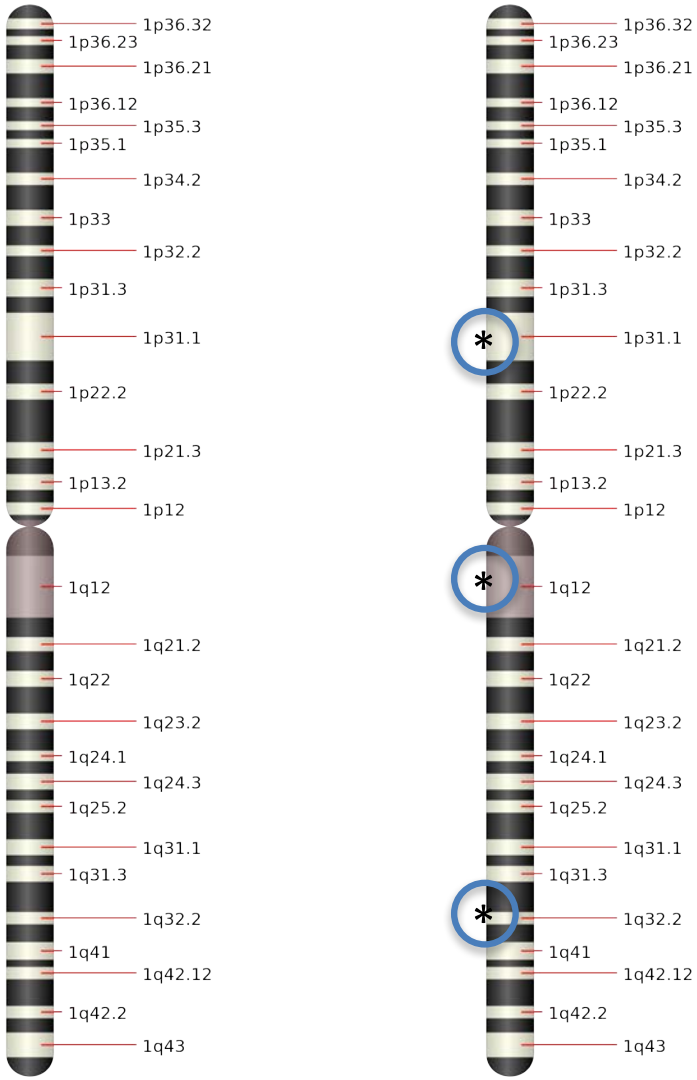
1) Repeats



Repeats break up contigs

http://www.cbcb.umd.edu/research/assembly_primer.shtml

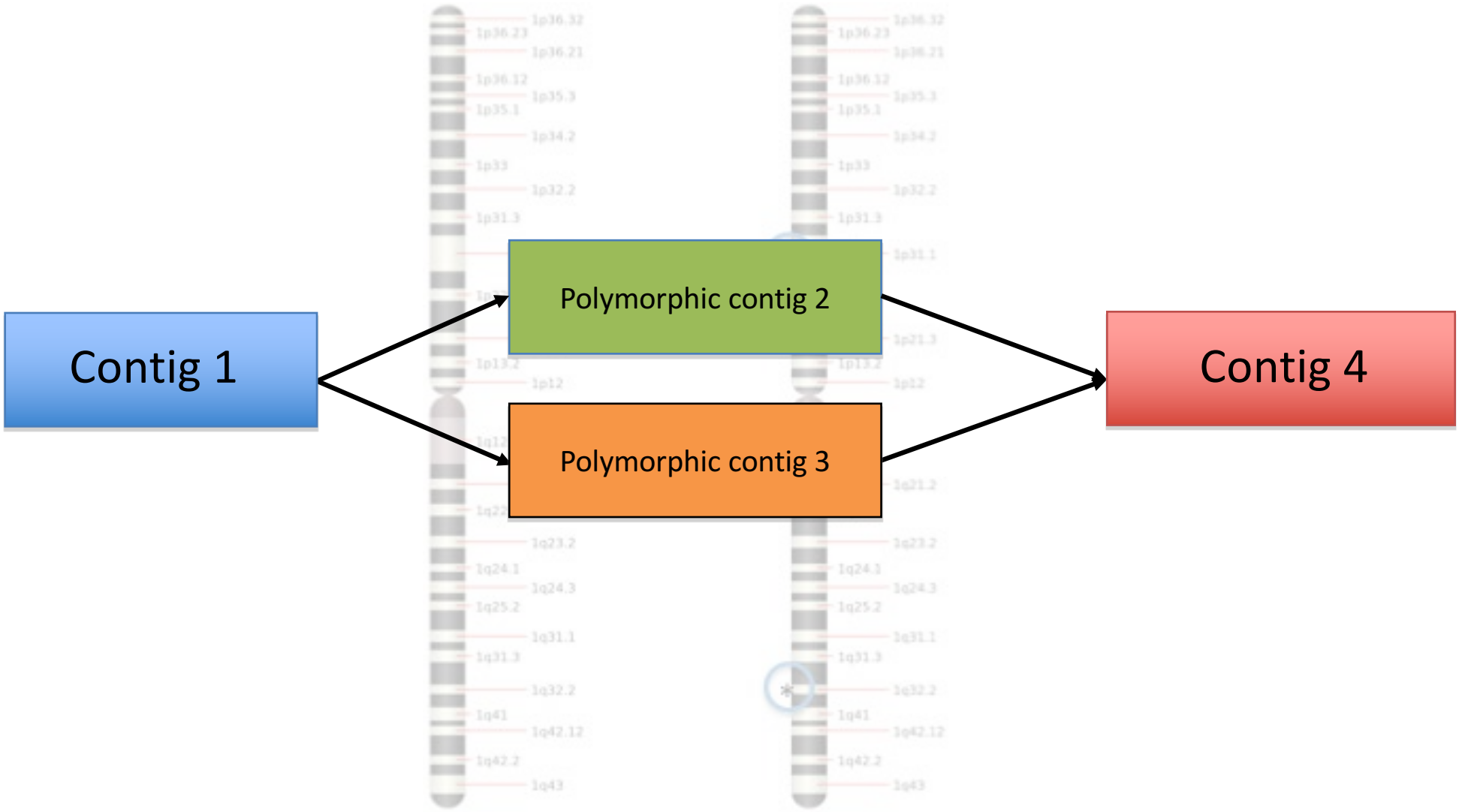
2) Heterozygosity



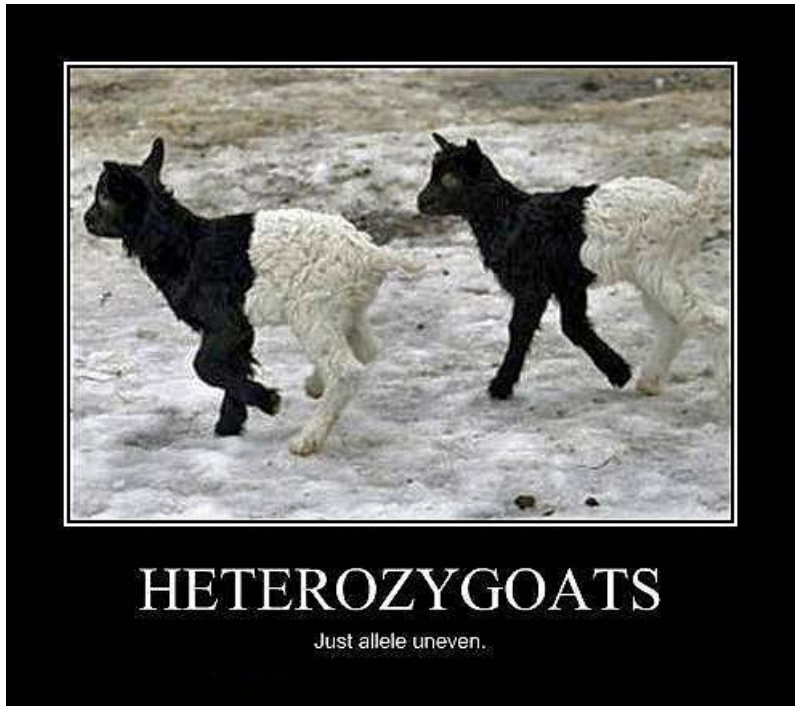
Differences between sister chromosomes

http://commons.wikimedia.org/wiki/File:Chromosome_1.svg

2) Heterozygosity

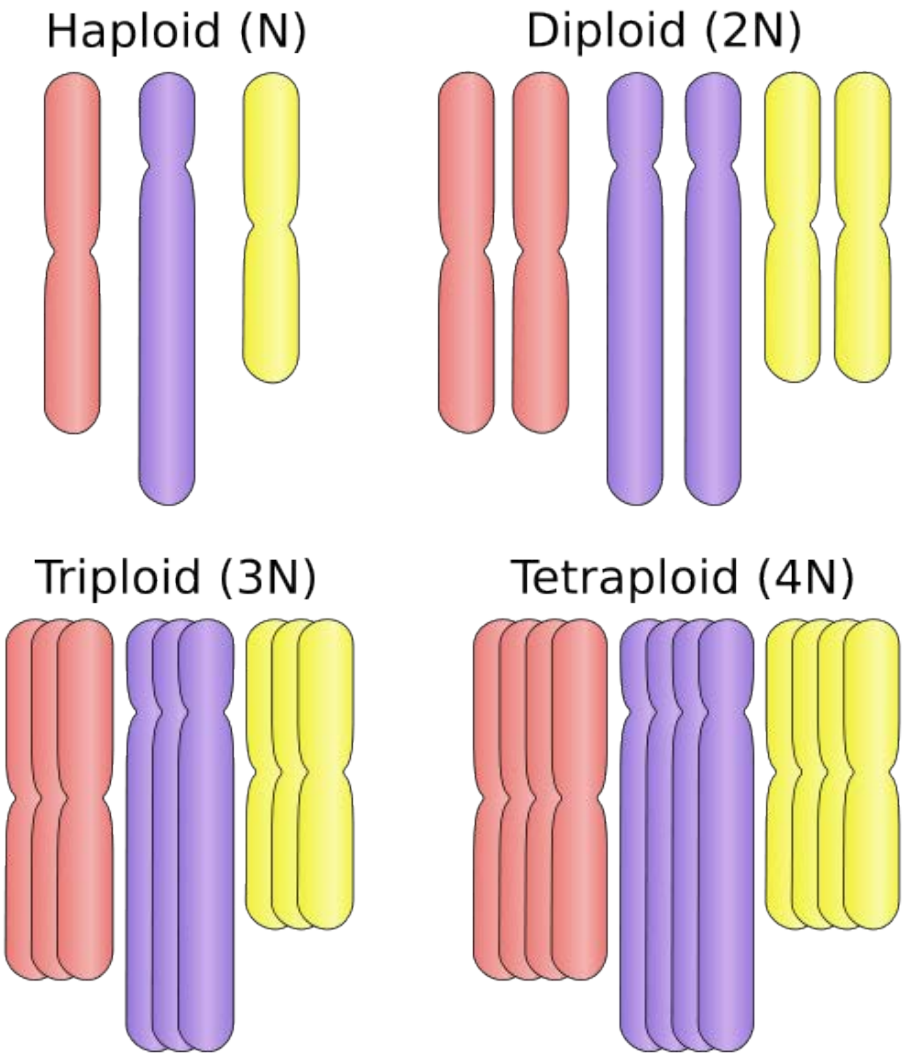


2) Heterozygosity



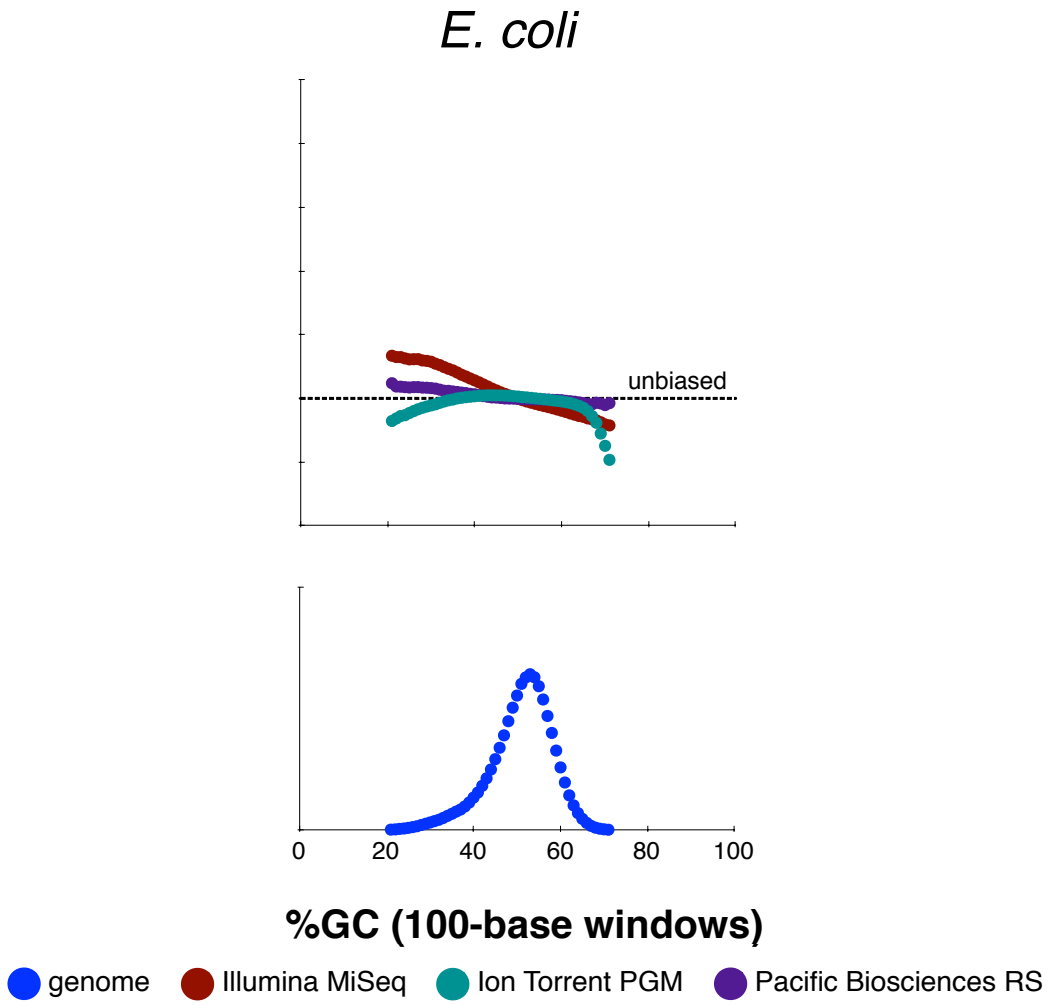
http://www.astraean.com/borderwars/wp-content/uploads/2012/04/heterozygoats.jpg
and many other sites

3) Polyploidy



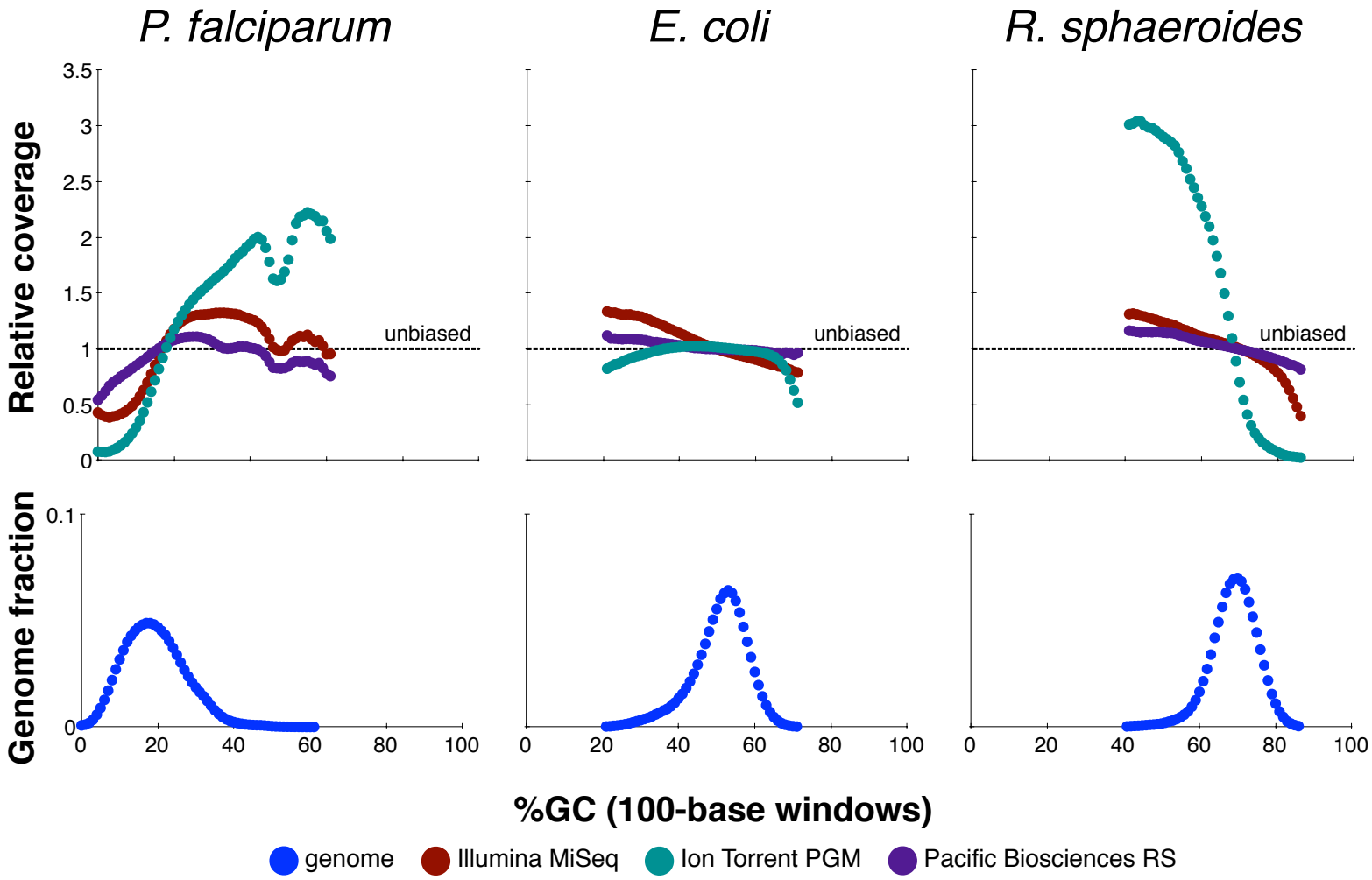
http://en.wikipedia.org/wiki/Polyploidy

4) Sequencing bias



Ross et al. (2013) Characterizing and measuring bias in sequence data.
doi:10.1186/gb-2013-14-5-r51

4) Sequencing bias



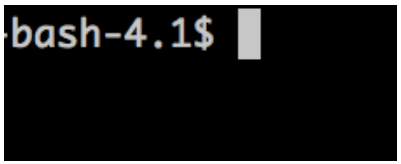
Ross et al. (2013) Characterizing and measuring bias in sequence data.
doi:10.1186/gb-2013-14-5-r51

Back to the how I became a bioinformatician

Planktothrix

Half a million reads
Average length 260 nt

10 million reads
33 nucleotides each



Perl

Planktothrix

Half a million reads
Average length 260 nt

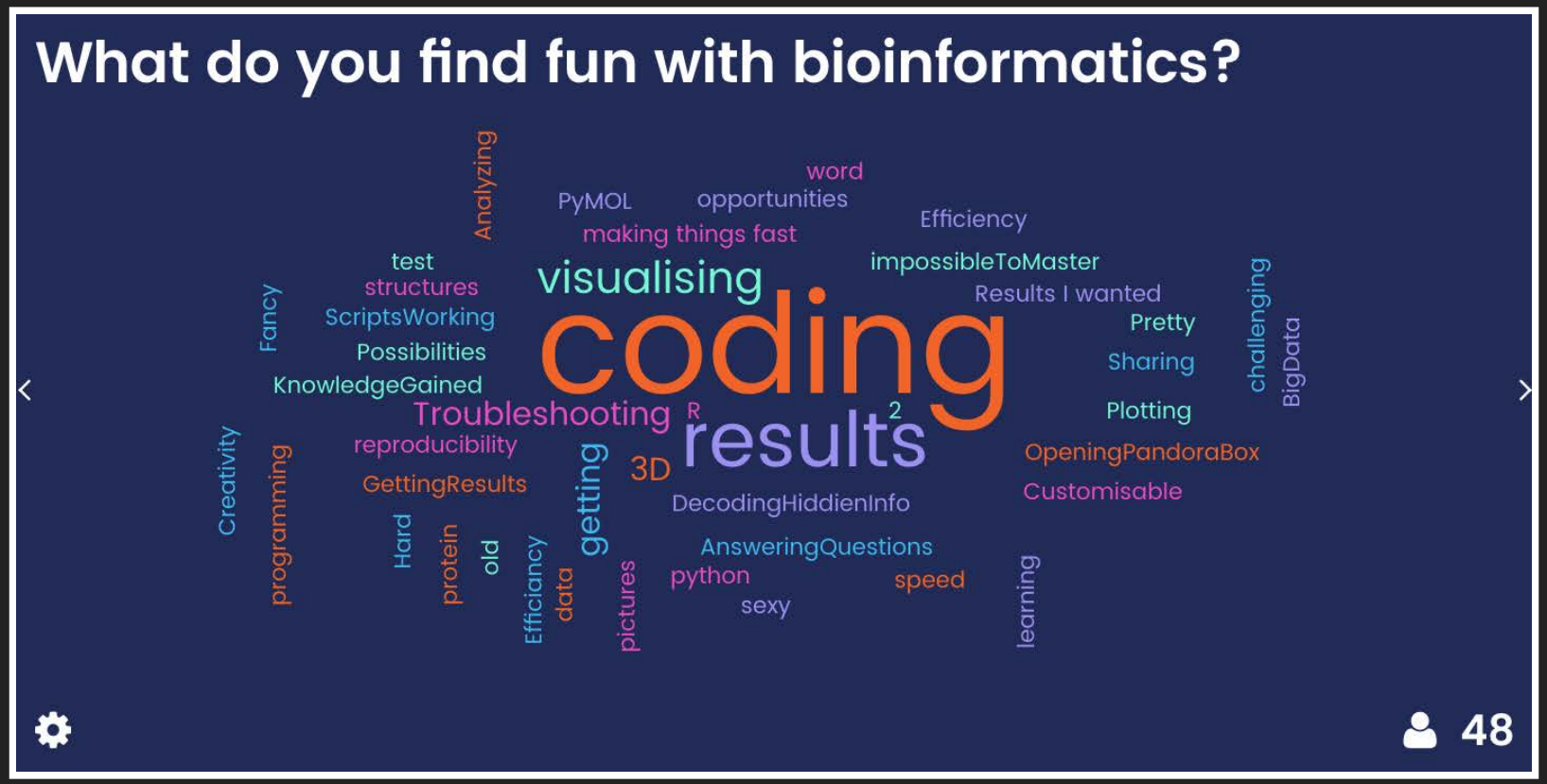
newbler

10 million reads
33 nucleotides each

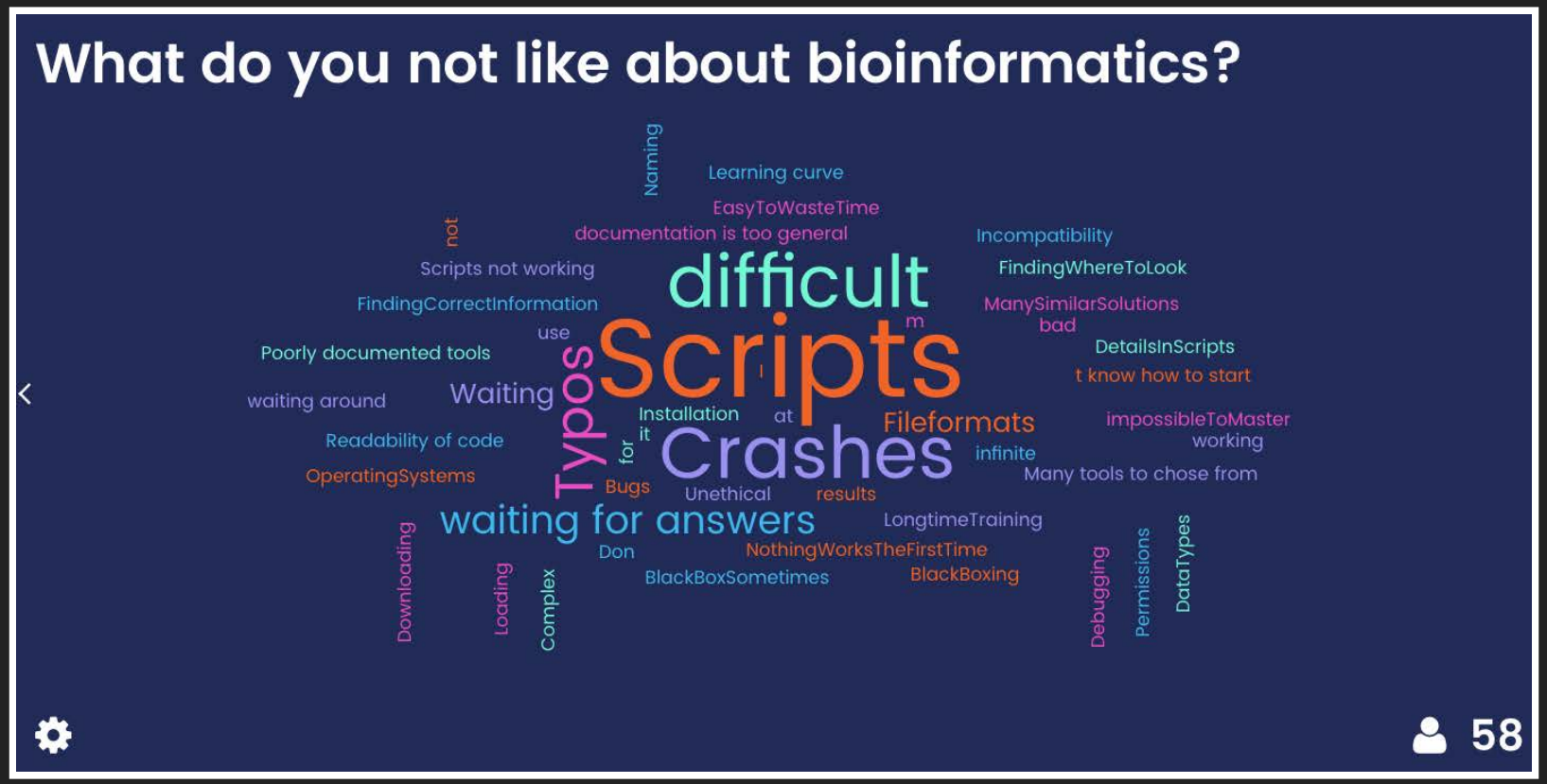
SHARCGS

Assembly

What do you find fun with bioinformatics?



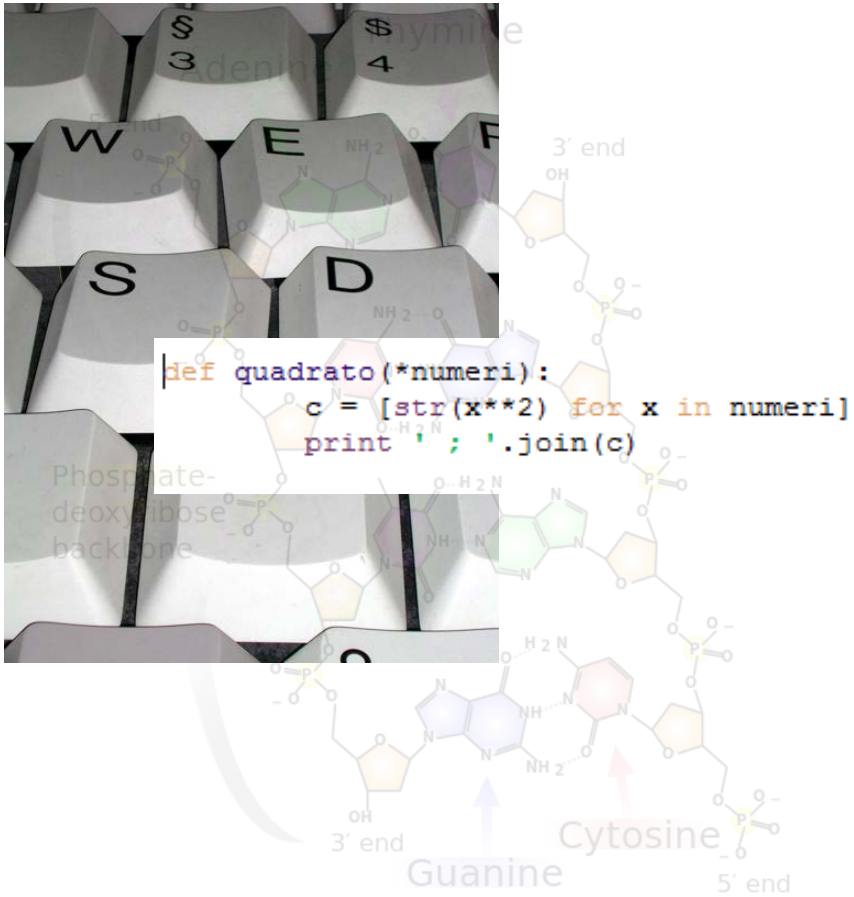
What do you not like about bioinformatics?



Part 3: What I think of bioinformatics

The fun parts

Coding



Mathias Bigge, Ricordisamoa, others (wikimedia commons)

The fun parts

Big computers

The fun parts

Learning something new all the time

www.gotcredit.com Via flickr https://flic.kr/p/rECgUF

The fun parts

Teaching it

The not-so-fun parts

Constant stream of new software

	Bioinformatics method	Biological technology	Operating system	Language	Maintained	Licence
4peaks	Sequence analysis	Sanger	Mac OS X		Yes	Freeware
AB Large Indel Tool	Mapping	ABI SOLiD	Linux 64	Perl	No	GPL
AB Small Indel Tool	Mapping Alignment	ABI SOLiD	Linux 64	Perl C++	Maybe	GPL
ABBA	Assembly Scaffolding		Linux		Maybe	Artistic License
ABMapper	Mapping Alignment	Illumina	Linux	C++ Perl	Yes	GPLv3
ABySS	Assembly De Bruijn graph	Illumina 454 ABI SOLiD Sanger	POSIX Linux Mac OS X	C++	Yes	Free for academic use
Adapter Removal	Adapter Removal	Illumina 454	Linux 64 Windows Mac OS X	Java	Yes	Custom Licence
AGE	Alignment Gap extension	Illumina			Maybe	Creative Commons license (Attribution-NonCommerical).
AGILE	Mapping	454		C	Yes	
Agp2amos	Format conversion		Windows Linux		Maybe	
Alcovna				Java	Maybe	
ALEXA-Seq				Perl	Maybe	GPLv3
ALLPATHS	Assembly De Bruijn graph				Yes	

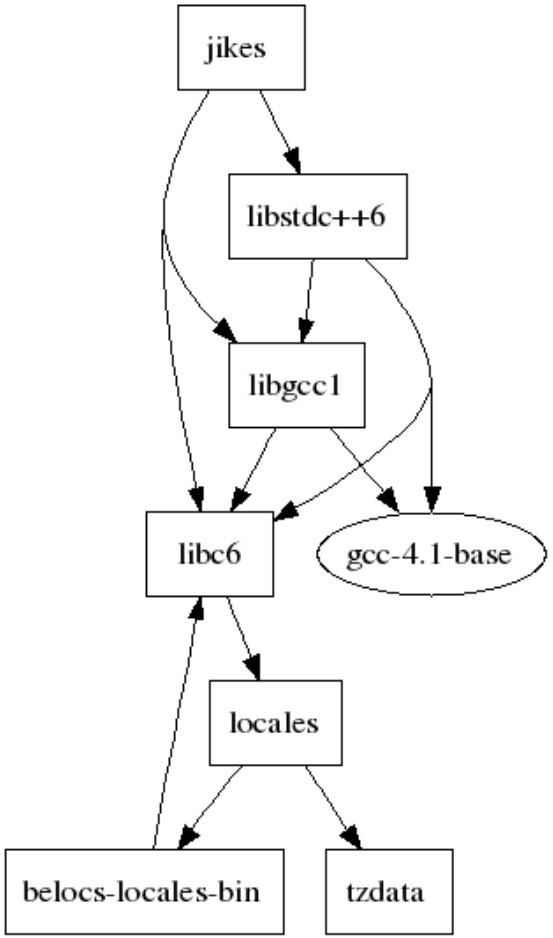
http://seqanswers.com/wiki/Software

The not-so-fun parts

Constant stream of new software

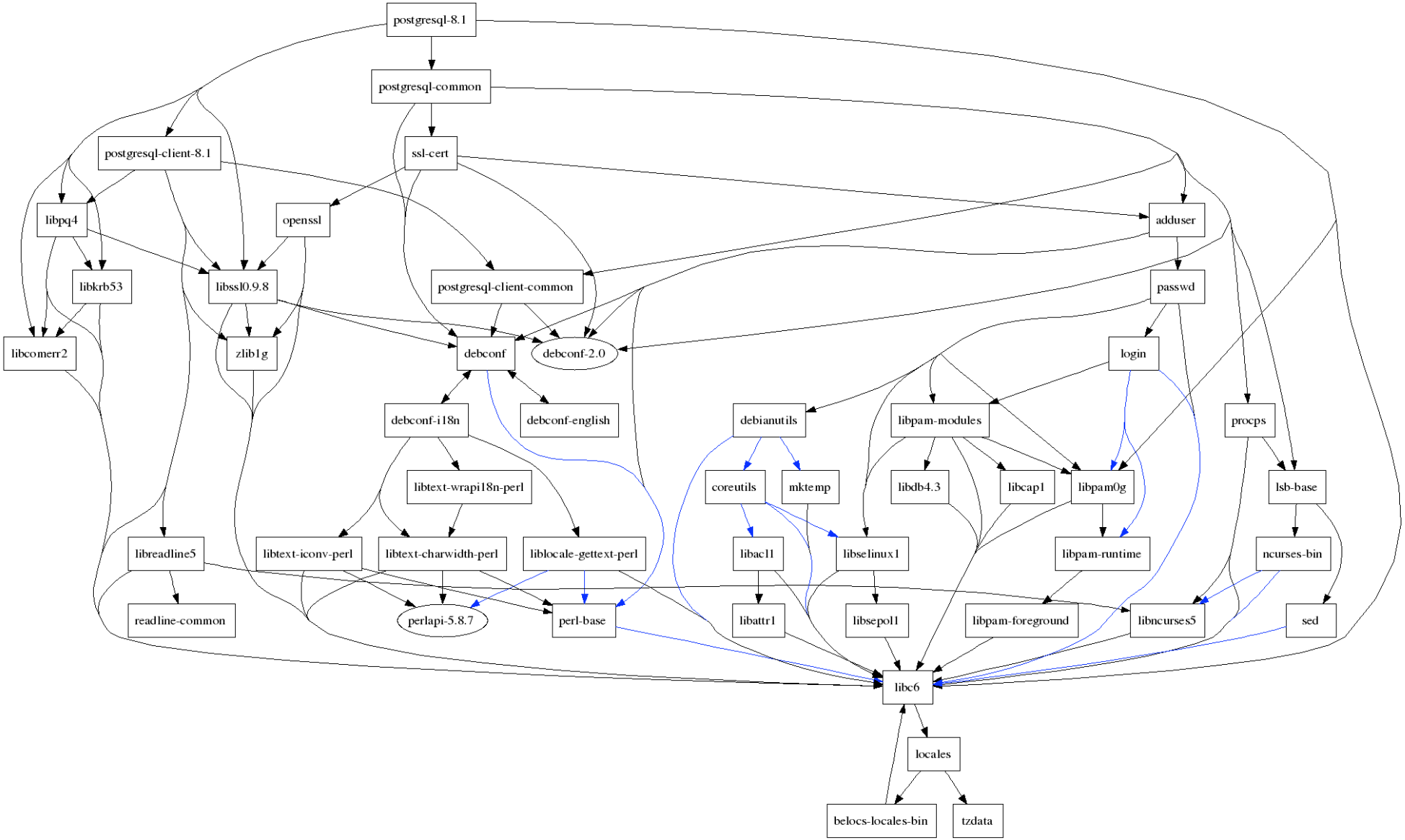
→hard to judge if programs are any good

→sometimes a challenge to install a program and get it working



http://neidetcher.com/ubuntu_package_dependency.html

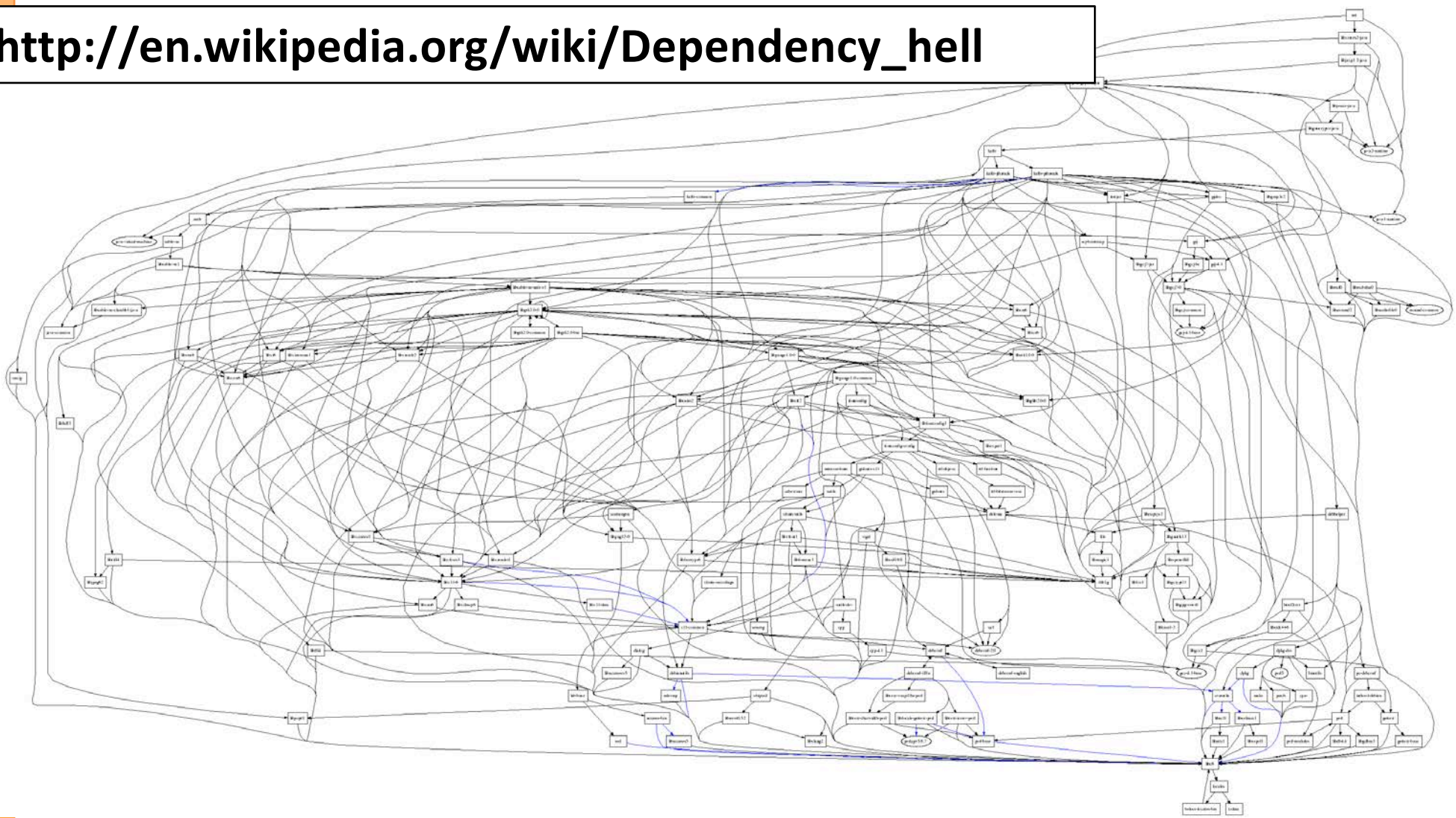
The not-so-fun parts



http://neidetcher.com/ubuntu_package_dependency.html

The not-so-fun parts

http://en.wikipedia.org/wiki/Dependency_hell



http://neidetcher.com/ubuntu_package_dependency.html

The not-so-fun parts

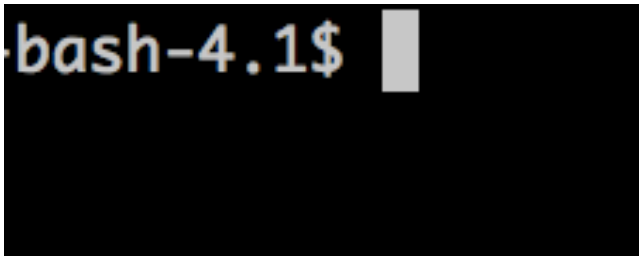
File formats

```
.csv
.txt
.tsv
.sam
.bam
.cram
.vcf
.bed
.gff
.fasta
.fastq
.fastg
.qual
...
```

Part 4: How to become an efficient bioinformatician

Learn

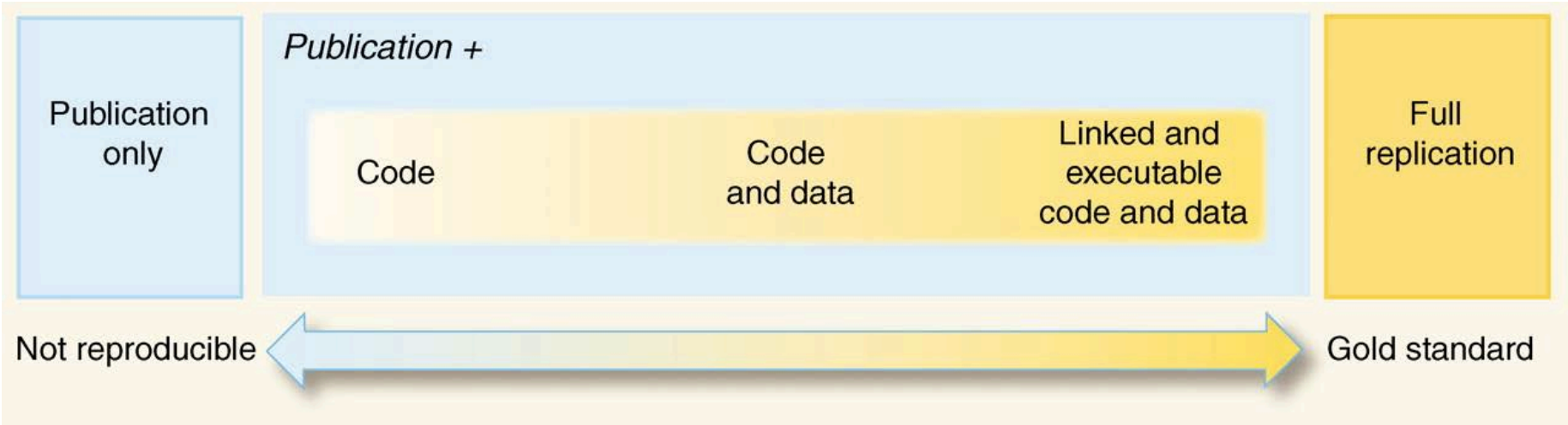
The command line



Learn

python™ A programming language

Make your work reproducible



Peng (2011) doi: 10.1126/science.1213847

Ask

Your neighbor

https://flic.kr/p/rJS6xM Flickr (user Spencer Means)

Ask

The internet

Learn

COMMENTARY

_computational BIOLOGY

So you want to be a computational biologist?

Nick Loman & Mick Watson

Two computational biologists give advice when starting out on computational projects.

http://www.nature.com/nbt/journal/v31/n11/full/nbt.2740.html
doi:10.1038/nbt.2740

Learn

Good Enough Practices in Scientific Computing

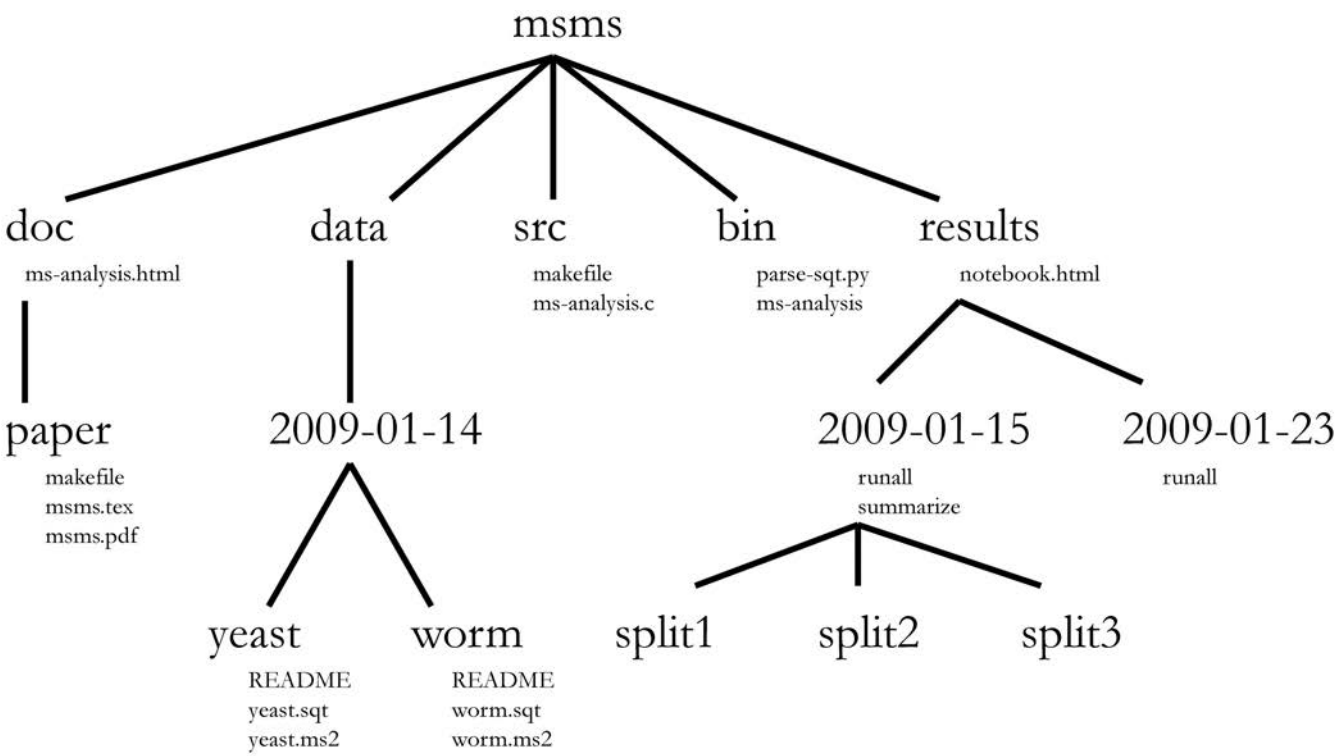
Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, Tracy K. Teal

(Submitted on 31 Aug 2016 (v1), last revised 14 Oct 2016 (this version, v2))

We present a set of computing tools and techniques that every researcher can and should adopt. These recommendations synthesize inspiration from our own work, from the experiences of the thousands of people who have taken part in Software Carpentry and Data Carpentry workshops over the past six years, and from a variety of other guides. Unlike some other guides, our recommendations are aimed specifically at people who are new to research computing.

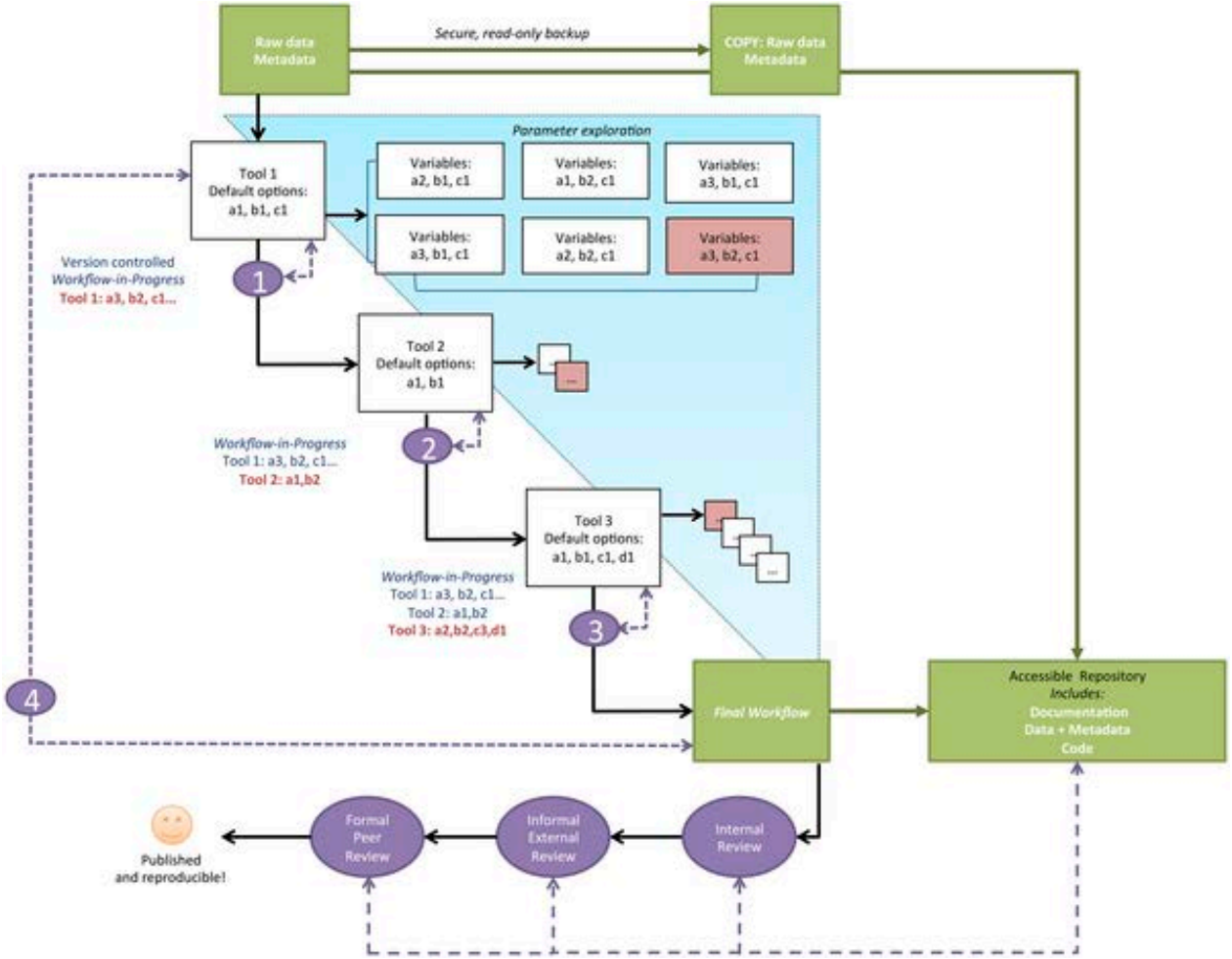
https://arxiv.org/abs/1609.00037

Learn



Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424. doi:10.1371/journal.pcbi.1000424

Learn



Shade A, Teal TK (2015) Computing Workflows for Biologists: A Roadmap. PLoS Biol 13(11): e1002303. doi:10.1371/journal.pbio.1002303

Learn

Attend a Software or Data Carpentry workshop

http://software-carpentry.org/
http://datacarpentry.org

uio-carpentry.github.io
(soon uio.no/carpentry)