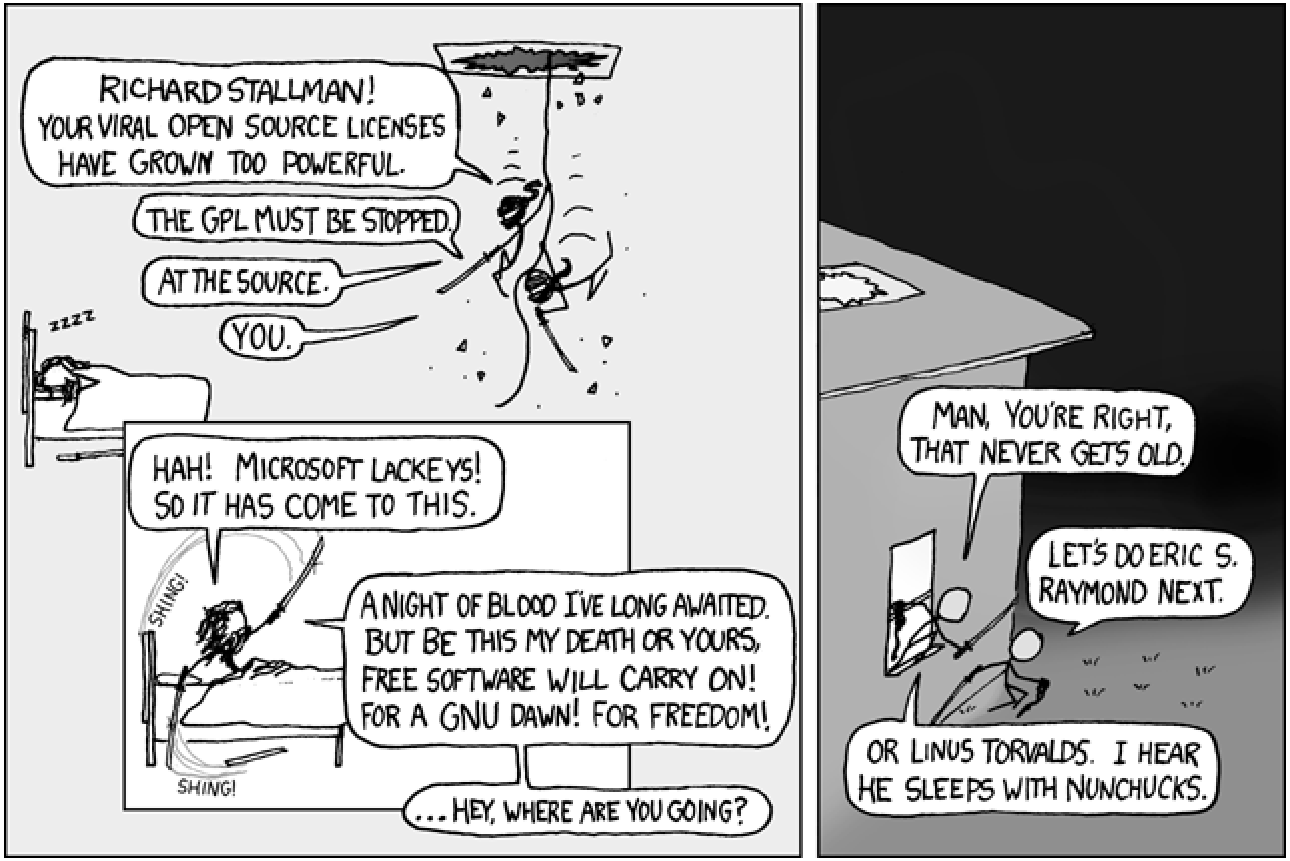
MIMM4750G

NGS QC and open source


RICHARD STALLMAN! YOUR VIRAL OPEN SOURCE LICENSES HAVE GROWN TOO POWERFUL.
THE GPL MUST BE STOPPED.
AT THE SOURCE.
YOU.
ZZZZ
HAH! MICROSOFT LACKEYS! SO IT HAS COME TO THIS.
SHING!
SHING!
A NIGHT OF BLOOD I'VE LONG AWAITED. BUT BE THIS MY DEATH OR YOURS, FREE SOFTWARE WILL CARRY ON! FOR A GNU DAWN! FOR FREEDOM!
...HEY, WHERE ARE YOU GOING?
MAN, YOU'RE RIGHT, THAT NEVER GETS OLD.
LET'S DO ERIC S. RAYMOND NEXT.
OR LINUS TORVALDS. I HEAR HE SLEEPS WITH NUNCHUCKS.

Now you have data

- Congratulations! You have several hundred gigabytes of data.
- Before you start to learn how to *analyze* the data, you need to check if it is any good.
- Quality control is a necessary and tedious step of NGS analysis.
- We will focus on Illumina sequencing - there are many other platforms but right now Illumina is fairly popular.

Demultiplexing

- One of the first steps in processing raw NGS outputs
- Generates FASTQ from base call `.bcl` files
- This conversion used to be performed with a Perl script `bcl2fastq.pl`
- Has now been re-implemented as a C++ program `bcl2fastq2`
- Separates reads labelled with different index tags into different FASTQ files.

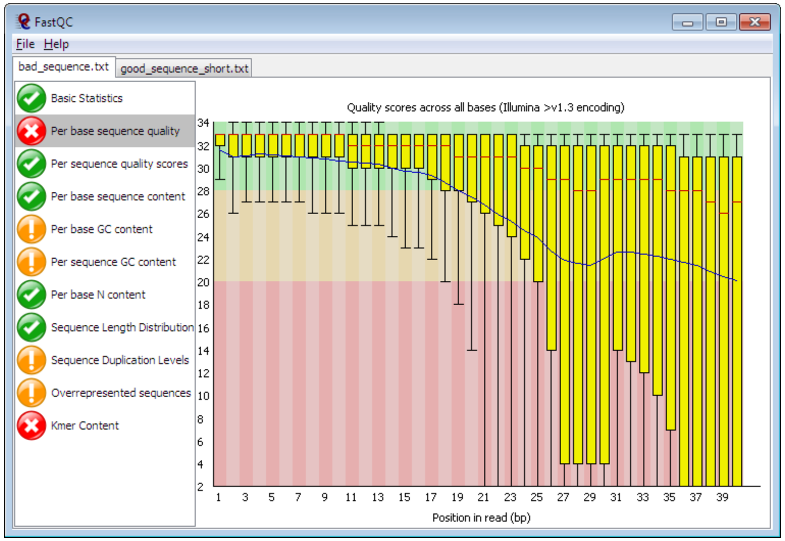
Quality control

- **Number of reads:** a sample may have a small number of reads, possibly due to inaccurate DNA quantification
- **Quality scores:** read quality tends to fall off over cycles.
- **Nucleotide frequencies:** skewed frequencies can reflect poor quality.

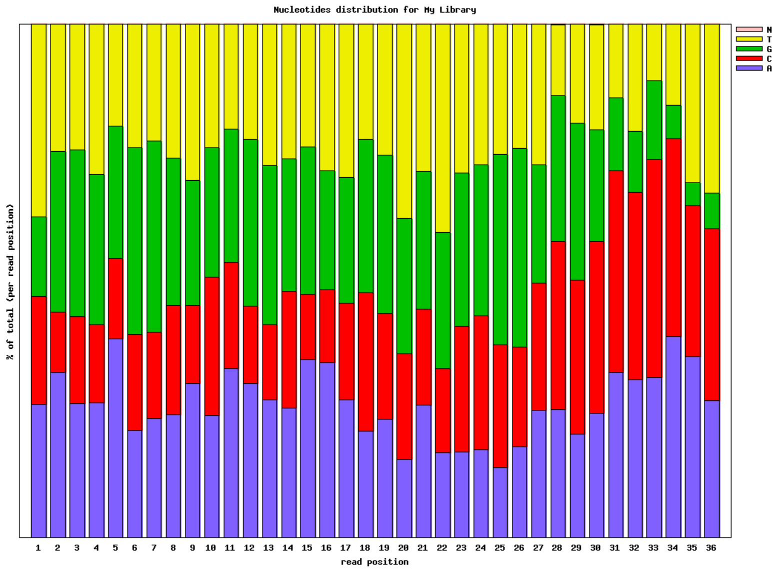
QC Software

- Several different software packages are available for screening FASTQ data for quality; for example:

FastQC



fastx_toolkit



Trimming adapters

- Illumina adapters are short nucleotide sequences (oligos) that are used in the construction of the sequencing library
- If the DNA fragment is shorter than the read, then the sequence may "read-through" to the adapter on the other end.
- Adapter contamination: many genomes in Genbank are contaminated with adapter sequences that were not removed by the authors.
- *e.g.*, the carp genome

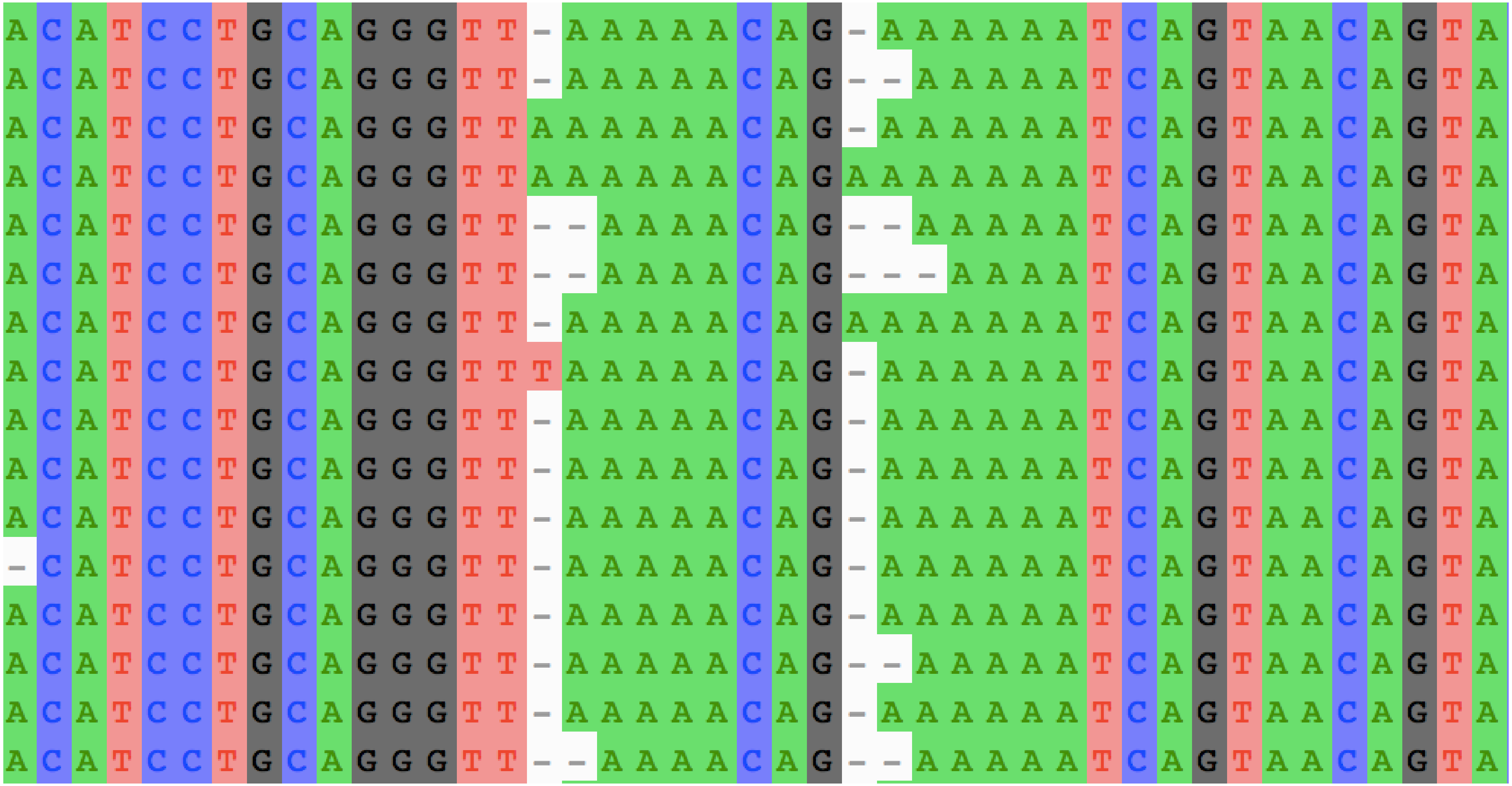
Software for trimming

- Trimmomatic - Java program for trimming Illumina data
- cutadapt - A Python module for removing adapter sequences and other artefacts
- AfterQC - Another Python module for trimming, discarding low quality bases and reconciling paired-end reads.

Sources of error

- PCR error
- Homopolymer errors (454, Ion Torrent)
- Cross-talk between clusters
- Sample cross-contamination
- Inter-run contamination

Homopolymer error



```
A C A T C C T G C A G G G T T - A A A A A C A G - A A A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T - A A A A A C A G - - A A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T A A A A A A C A G - A A A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T A A A A A A C A G A A A A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T - - A A A A C A G - - A A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T - - A A A A C A G - - - A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T - A A A A A C A G A A A A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T T A A A A A C A G - A A A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T - A A A A A C A G - A A A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T - A A A A A C A G - A A A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T - A A A A A C A G - A A A A A A T C A G T A A C A G T A
- C A T C C T G C A G G G T T - A A A A A C A G - A A A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T - A A A A A C A G - A A A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T - A A A A A C A G - - A A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T - A A A A A C A G - A A A A A A T C A G T A A C A G T A
A C A T C C T G C A G G G T T - - A A A A C A G - - A A A A A T C A G T A A C A G T A
```

φX174 control

- Bacteriophage genomic DNA (φX174) is commonly used as a positive control ("spike-in") for Illumina runs
- Reads mapping to φX714 are sorted by the vendor software and used to measure run-specific error rates
- Sometimes the software fails to remove all φX174 reads from the data! About 10% of genomes published by 2015 contaminated.
- These results are stored in one of the binary `InterOp` files.

InterOp files

- About 11 binary (`*.bin`) InterOp files produces for every Illumina run

- Binary means that these files are not plain-text:

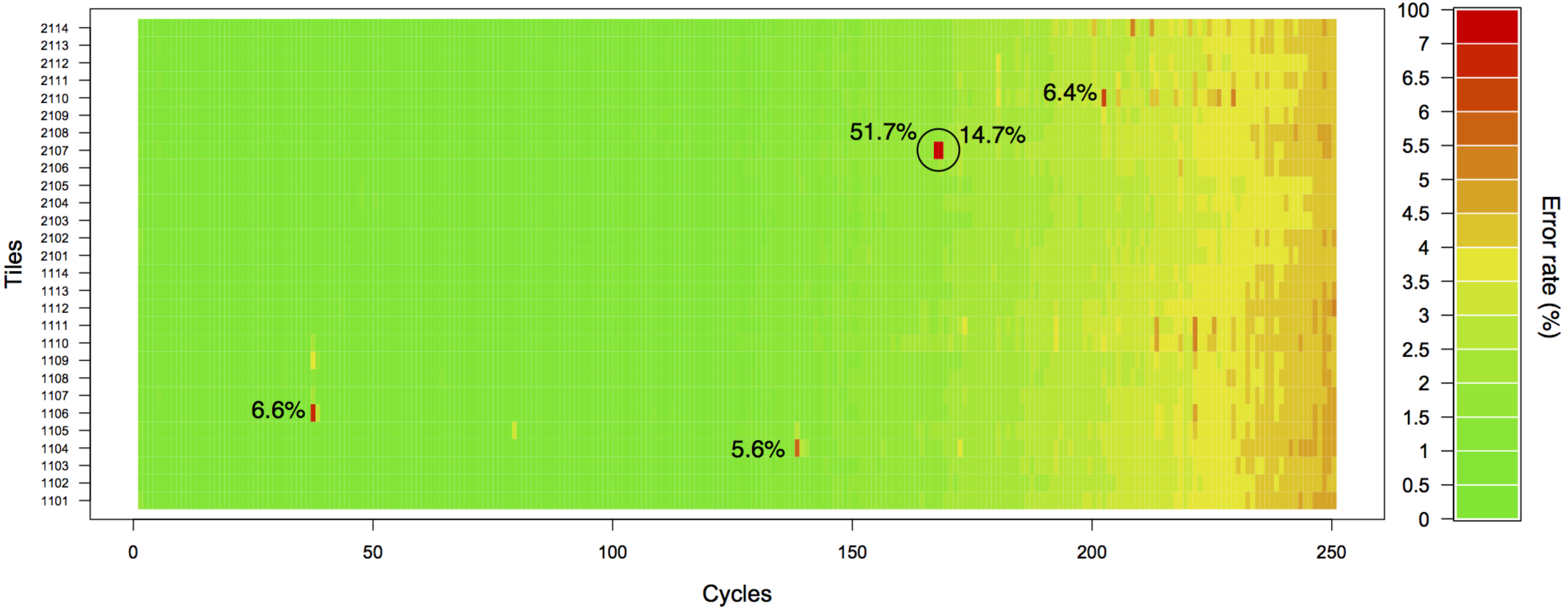
```
7??=??&M7???=??`M7???=???M?>7??M???>a??MW??=&??M?>Ýž=M????>???M?
?=D??M?j">??M???=È·Mw??=??M{?=h?N?a?>???N?Z>???N?+?=?N??=?8N?
U?=??gN??H>[??N?n>?:@>ÅŽYN    9iB>l??N?u>#??N
```

- Store metrics about that run, *e.g.*, tile-/cycle-mean quality scores

- `ErrorMetricsOut.bin` stores the φX174 error rates.

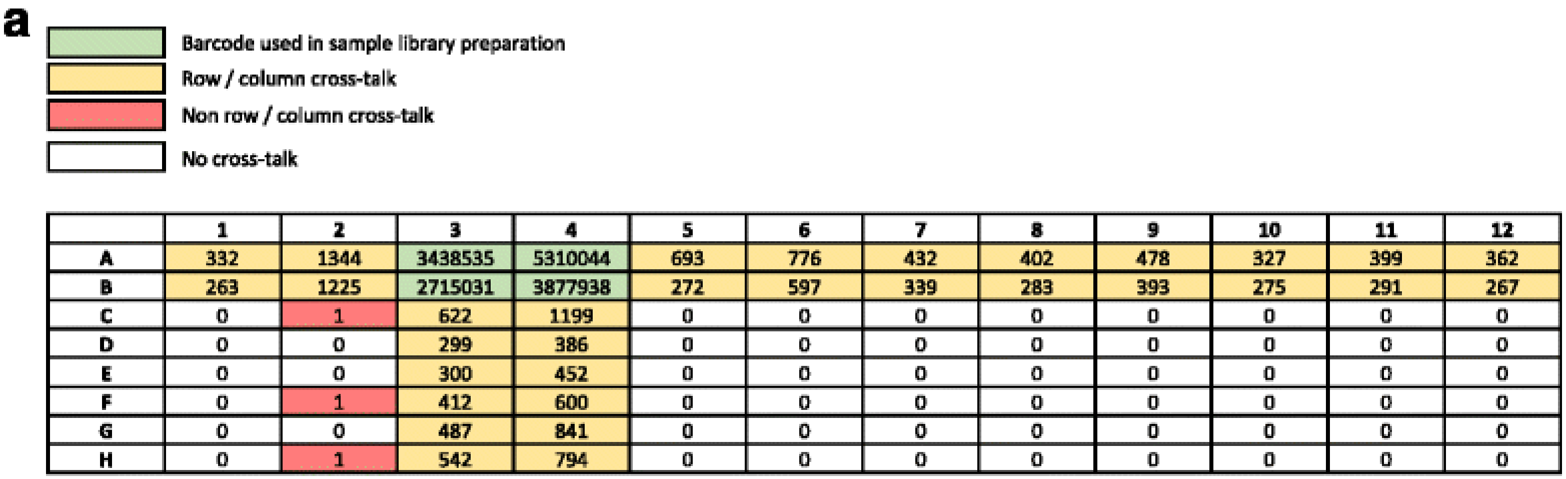
Bad tile-cycle combos

Extracted error rates from the InterOp file of a run where every HIV-1 patient was diagnosed with the same drug resistance mutation (K103N)



Cross-contamination

- Reads are incorrectly demultiplexed into other samples.
- "Index-hopping" - the barcode from one end recombines with the barcode from another template.
- Roughly 0.1% to 1% of reads assigned to a sample may come from other samples.



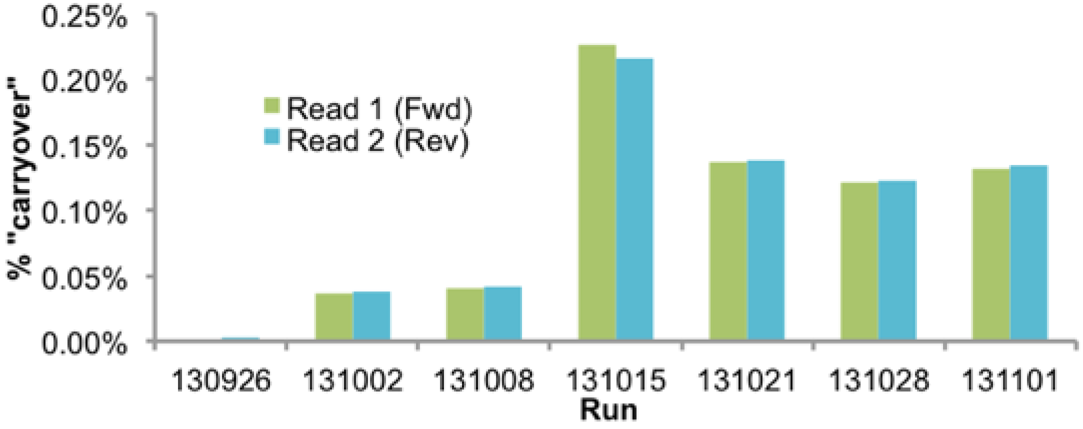
a

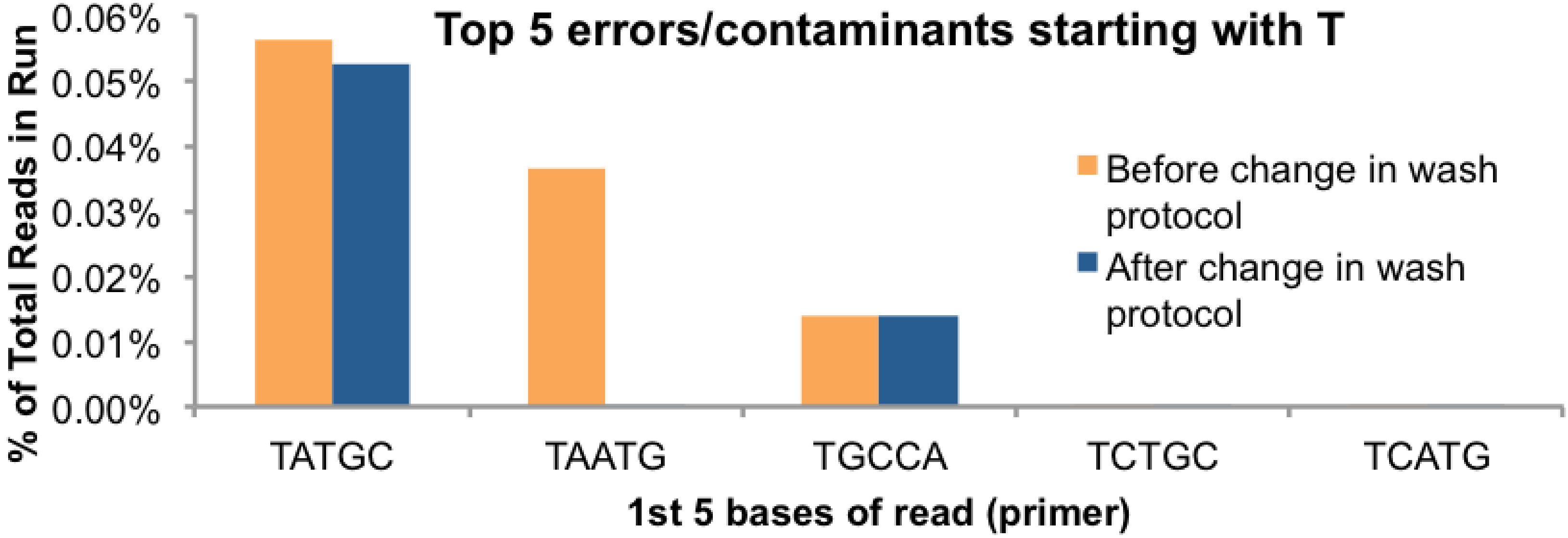
🟩 Barcode used in sample library preparation
🟨 Row / column cross-talk
🟥 Non row / column cross-talk
⬜ No cross-talk

	1	2	3	4	5	6	7	8	9	10	11	12
A	332	1344	3438535	5310044	693	776	432	402	478	327	399	362
B	263	1225	2715031	3877938	272	597	339	283	393	275	291	267
C	0	1	622	1199	0	0	0	0	0	0	0	0
D	0	0	299	386	0	0	0	0	0	0	0	0
E	0	0	300	452	0	0	0	0	0	0	0	0
F	0	1	412	600	0	0	0	0	0	0	0	0
G	0	0	487	841	0	0	0	0	0	0	0	0
H	0	1	542	794	0	0	0	0	0	0	0	0

Image from LE MacConaill *et al*. 2018 BMC Genomics 19:30

Carryover contamination





Open-source software (OSS)

- Source code is released under a license that grants users the freedom to use, modify and re-distribute it
- Some licenses require developers using the source code to release *their* code under the same license ("viral" GPL).
- The course materials at http://artpoon.github.io/BioID are released under the Creative Commons Attribution-ShareAlike license (same as Wikipedia).

Pros and cons of open-source

- OSS is often *free*, which promotes widespread use, *e.g.*, R.
- Transparency to be inspected by the developer community can make OSS more secure, reliable.
- The majority of bioinformatics software developed by researchers is released as OSS.
- OSS can only as good as it is actively maintained - there are many abandoned projects.
- Proprietary software may be more consistently maintained because developers are paid.

Compiling from source

- If a program is only distributed as source code or you need to customize it, you have to compile the code.
- A compiler (`gcc`, `javac`, `gfortran`) converts instructions that can be read by people into those that can be read by your computer.
- Many programs can be compiled by using the combination `./configure`, `make`, and `sudo make install`.

Package managers

- A public repository stores binaries (executable files compiled from source) that work for many different systems.
- A user can run a package manager that downloads the required binaries from a repository.
- Much easier than compiling from source, but can lead to unexpected problems.
- *e.g.*, NCBI `sra-toolkit` package was broken for a long time.