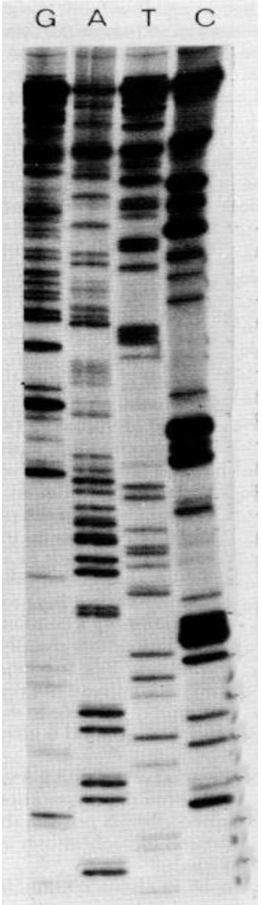
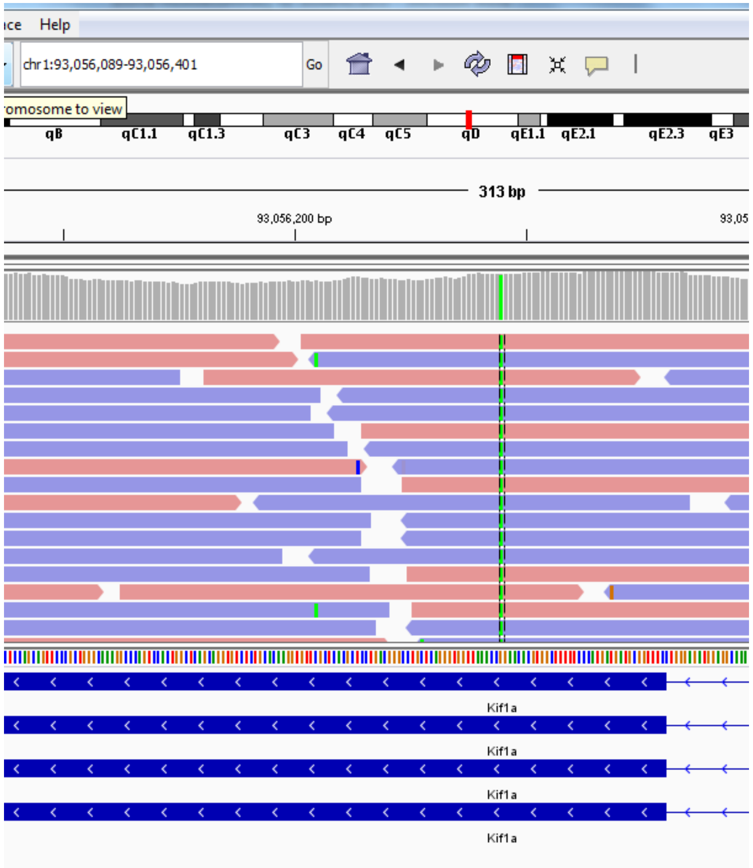
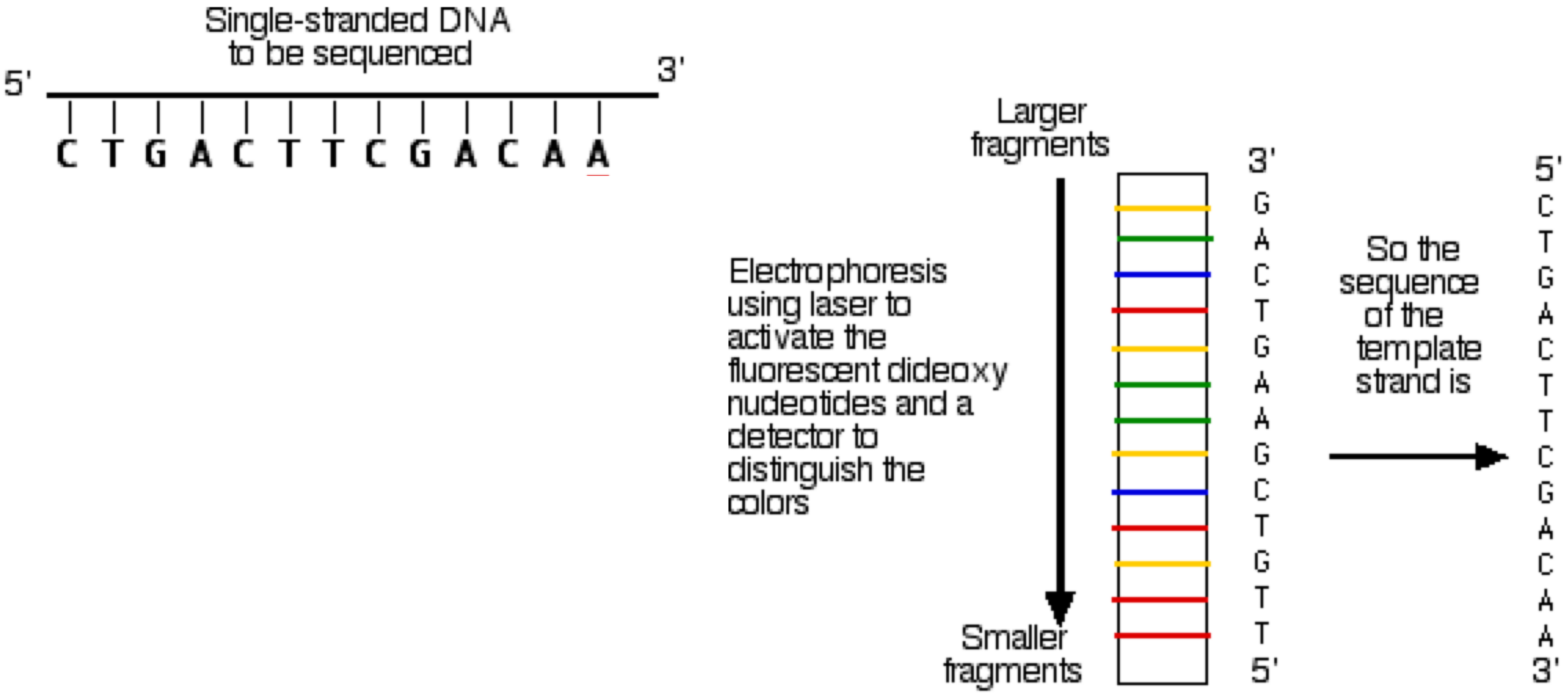
High Throughput Sequence & Sequence Analysis: A General Review




chr1:93,056,089-93,056,401

Sequencing - Methods

Chain termination (Sanger sequencing)


Single-stranded DNA to be sequenced
5'
3'
C T G A C T T C G A C A A
Larger fragments
Smaller fragments
3'
G A C T G A A G C T G T T
5'
5'
C T G A C T T C G A C A A
3'

Electrophoresis using laser to activate the fluorescent dideoxy nucleotides and a detector to distinguish the colors

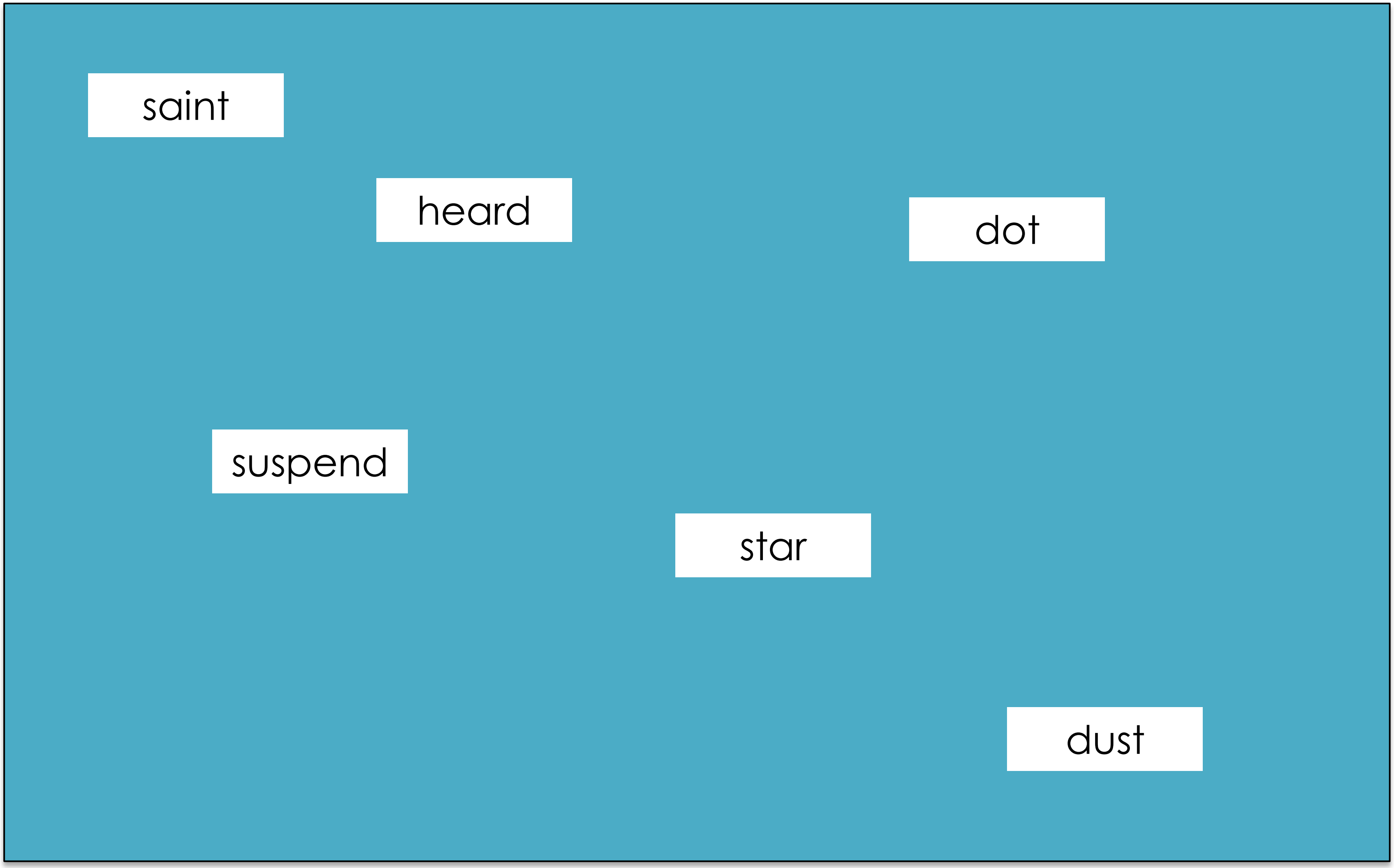
So the sequence of the template strand is

Targeted sequencing limitations

	Sanger	***Pyro***
Max. Length	800-1000 bp	350-500 bp
Error rate	0.001% to 1%	> 1%

✧ **TARGETED!**

How to build a genome by sequencing



saint

heard

dot

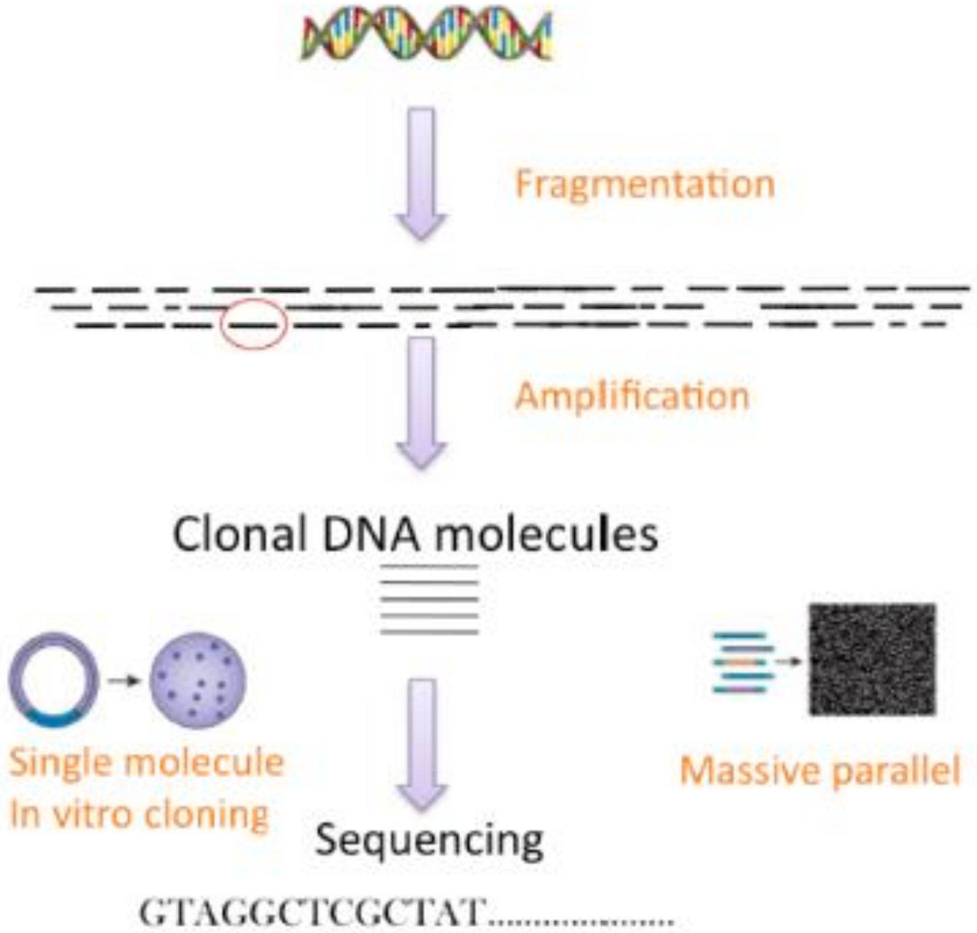
suspend

star

dust

✧ Where do the targeted sequences go in the genome?

How does next-gen sequencing work?



Fragmentation

Amplification

Clonal DNA molecules

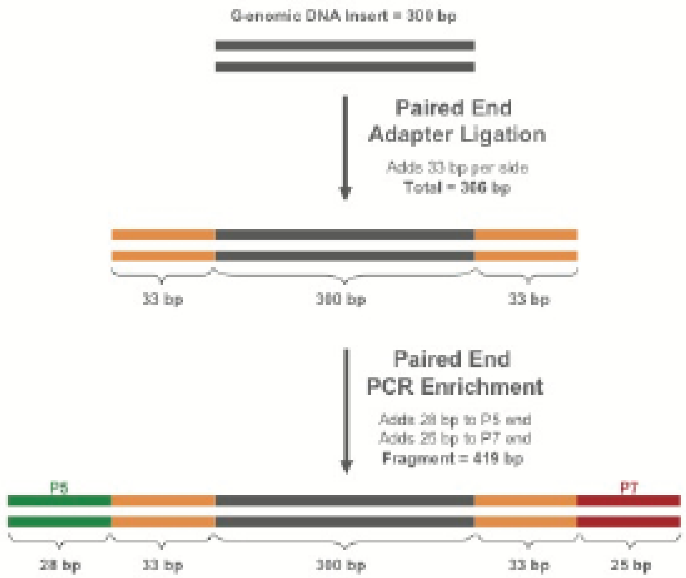
Single molecule
In vitro cloning

Massive parallel

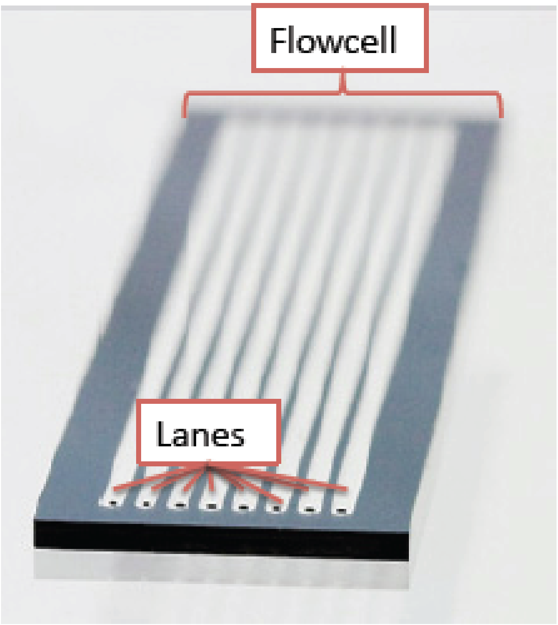
Sequencing

GTAGGCTCGCTAT....................

Terminology: libraries, lanes, and flowcells



Each reaction produces a unique **library** of DNA fragments for sequencing.

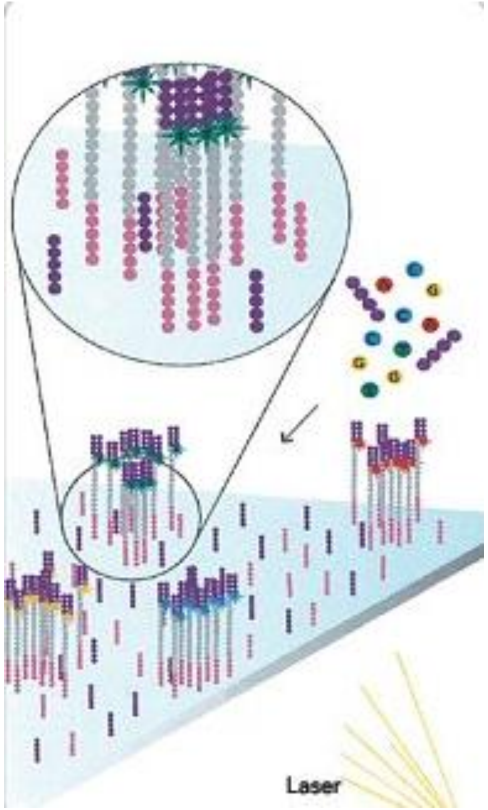


Each NGS machine processes a single **flowcell** containing several independent **lanes** during a single sequencing run

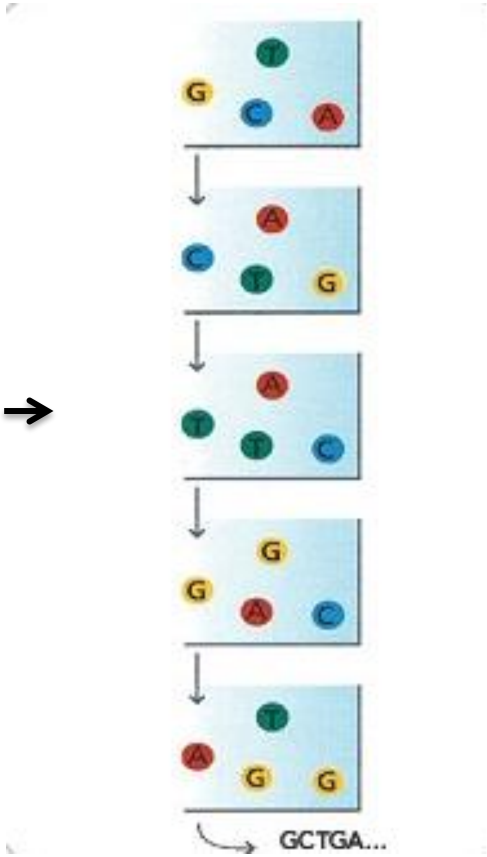
http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html

High throughput sequencing

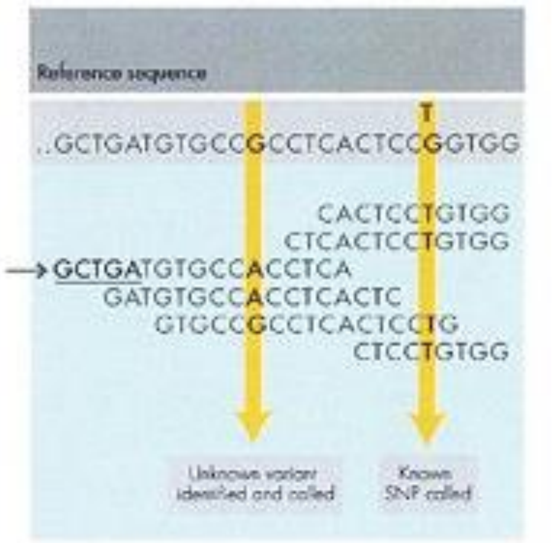
Illumina sequencing – chain termination



→



→



2nd generation sequencing output formats

Illumina ↓ FASTQ (various flavours)

SoLID/ABI-Life ↓ Color-space FASTA

Roche 454 ↓ SFF

Ion Torrent ↓ SFF or FASTQ

Platforms have errors and artifacts

Illumina

SoLID/ABI-Life

Roche 454

Ion Torrent

Removal of low quality bases
Removal of adaptor sequences
Platform specific artifacts (e.g homopolymers)

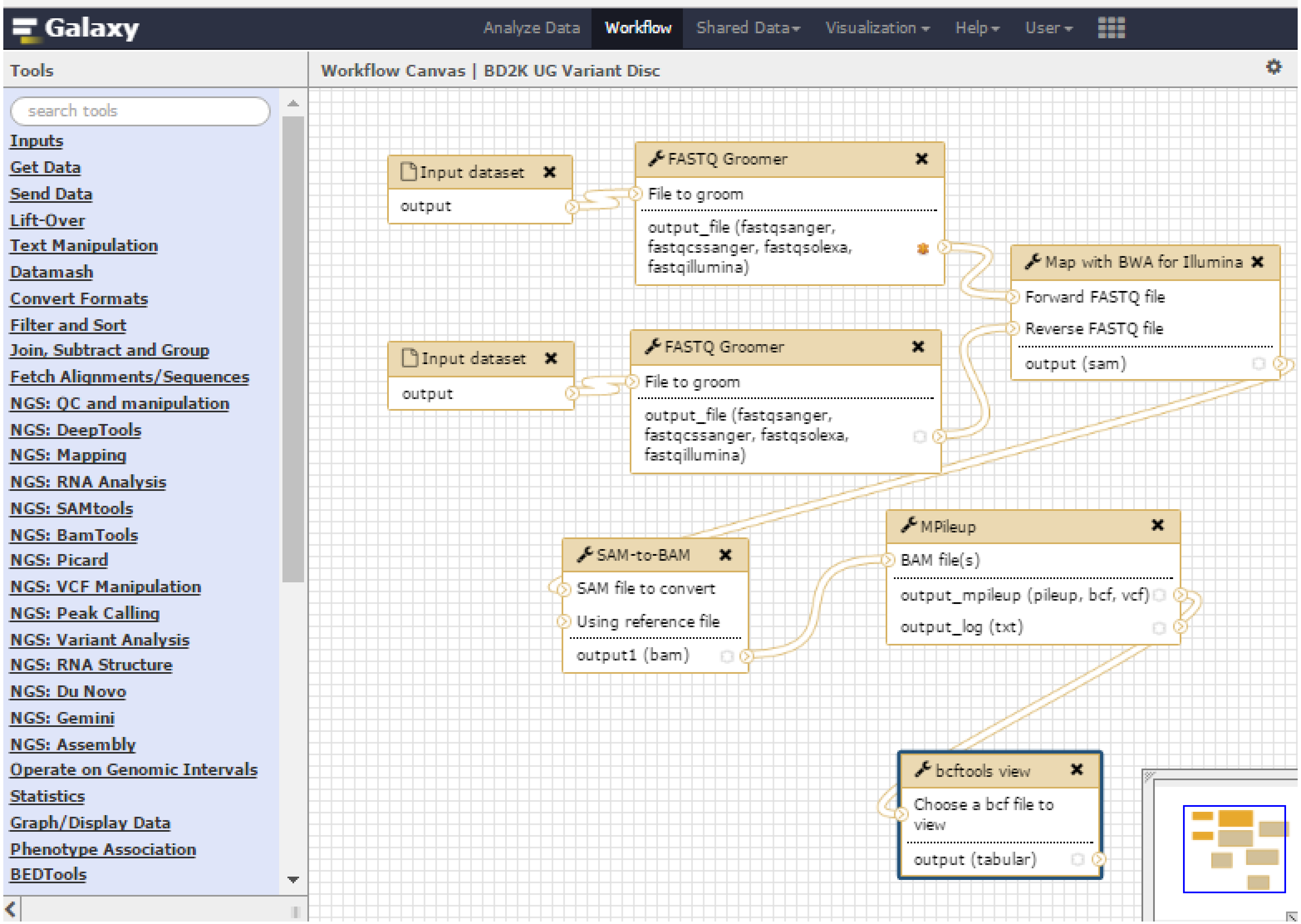
Computational choices in genomic data analysis

- Download and run at the command-line
- Public web-based server
- Private instance of a web-based server

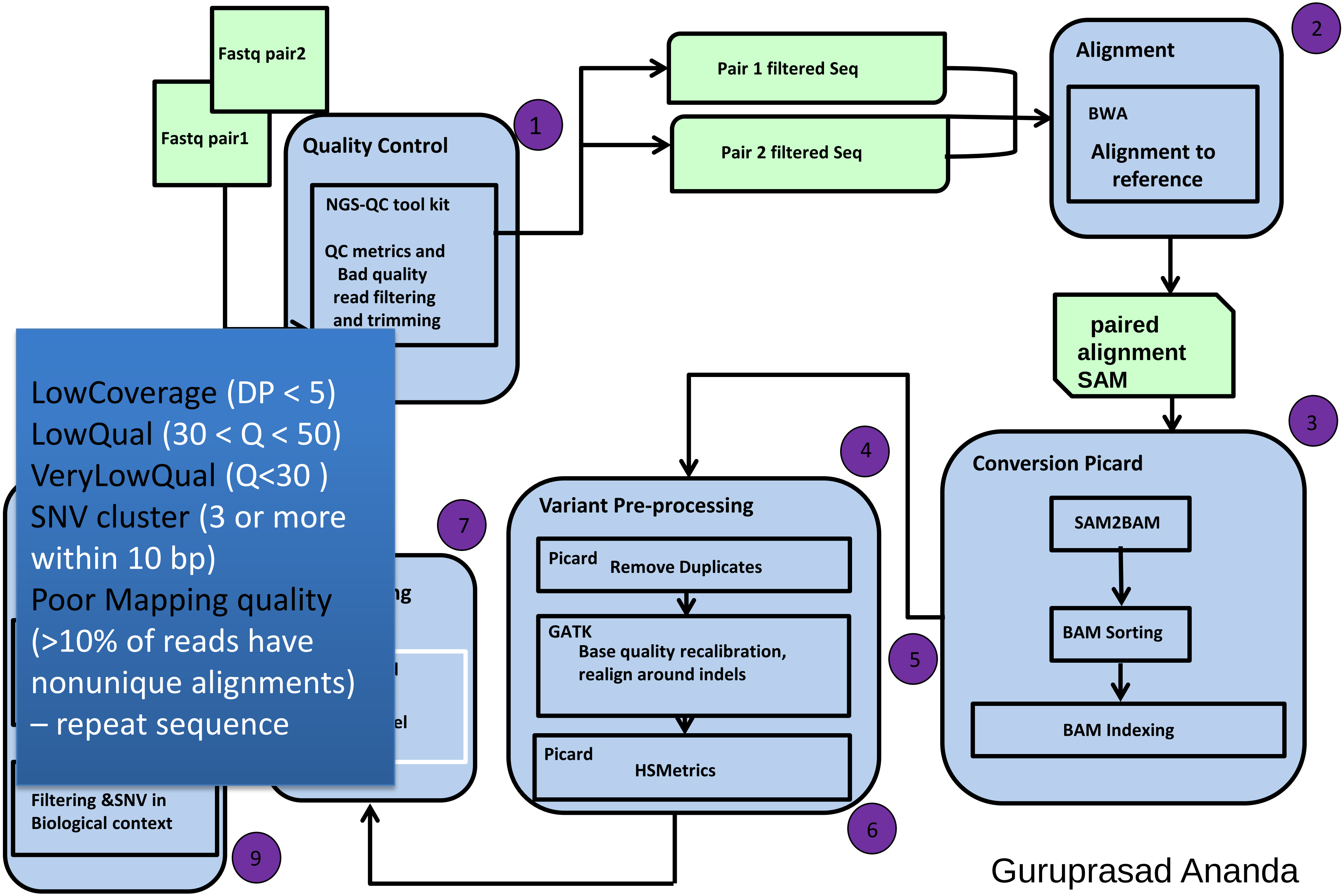
Things to remember about server-based analysis

- You are using someone else's compute resources
  - Good- your investment is small, someone else
  - Bad- server load and wait times can be unpredictable
- The majority of the tools are "wrapped" presenting a reduced set of options/functions
  - Full functionality is accessed from the command-line
- File transfer times to and from the server will be dependent upon network capacity.

Designing Data Analysis Workflows



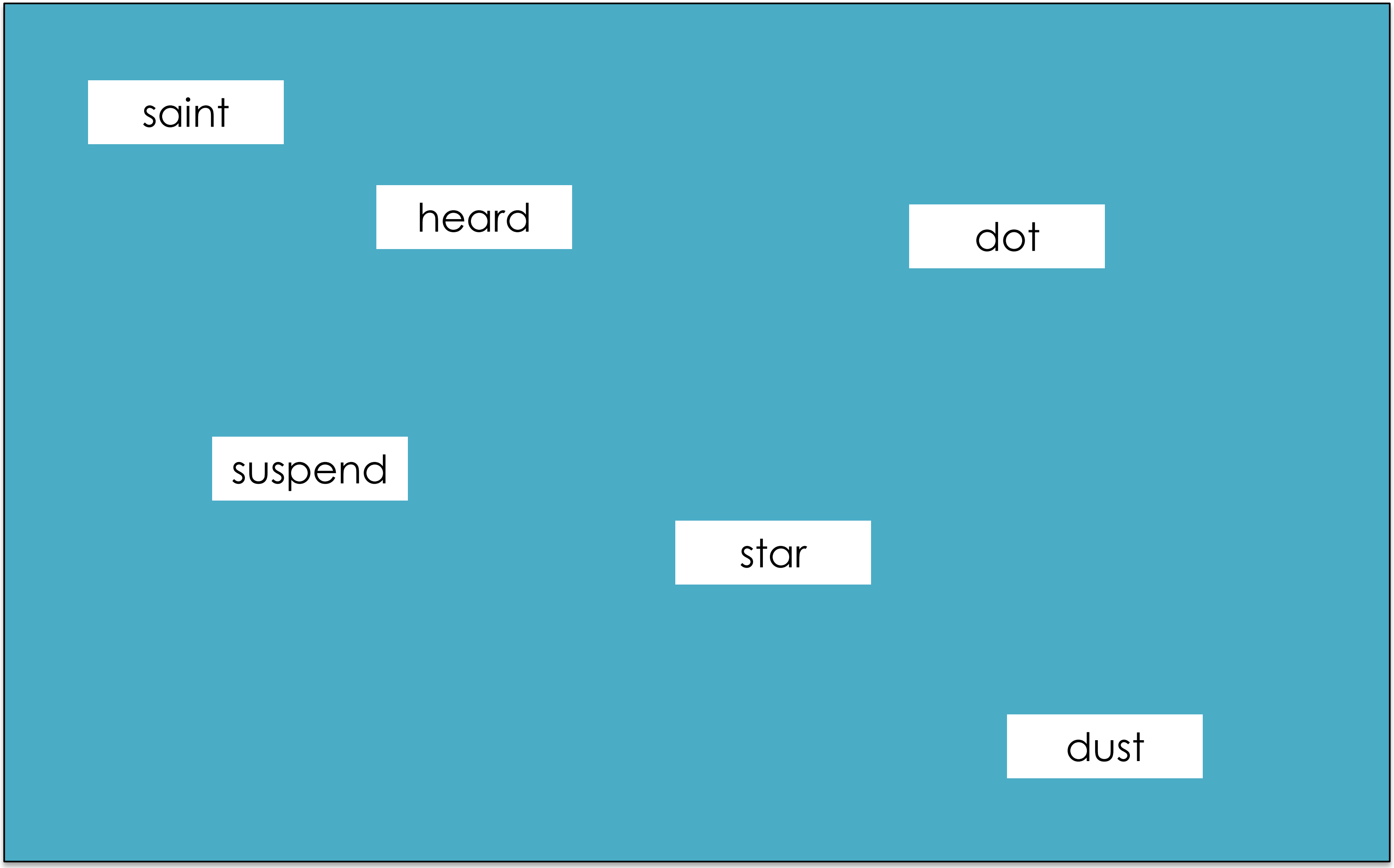
Mouse exome CIVET pipeline



LowCoverage (DP < 5)
LowQual (30 < Q < 50)
VeryLowQual (Q<30)
SNV cluster (3 or more within 10 bp)
Poor Mapping quality (>10% of reads have nonunique alignments) – repeat sequence

Guruprasad Ananda

How to build a genome by sequencing



saint

heard

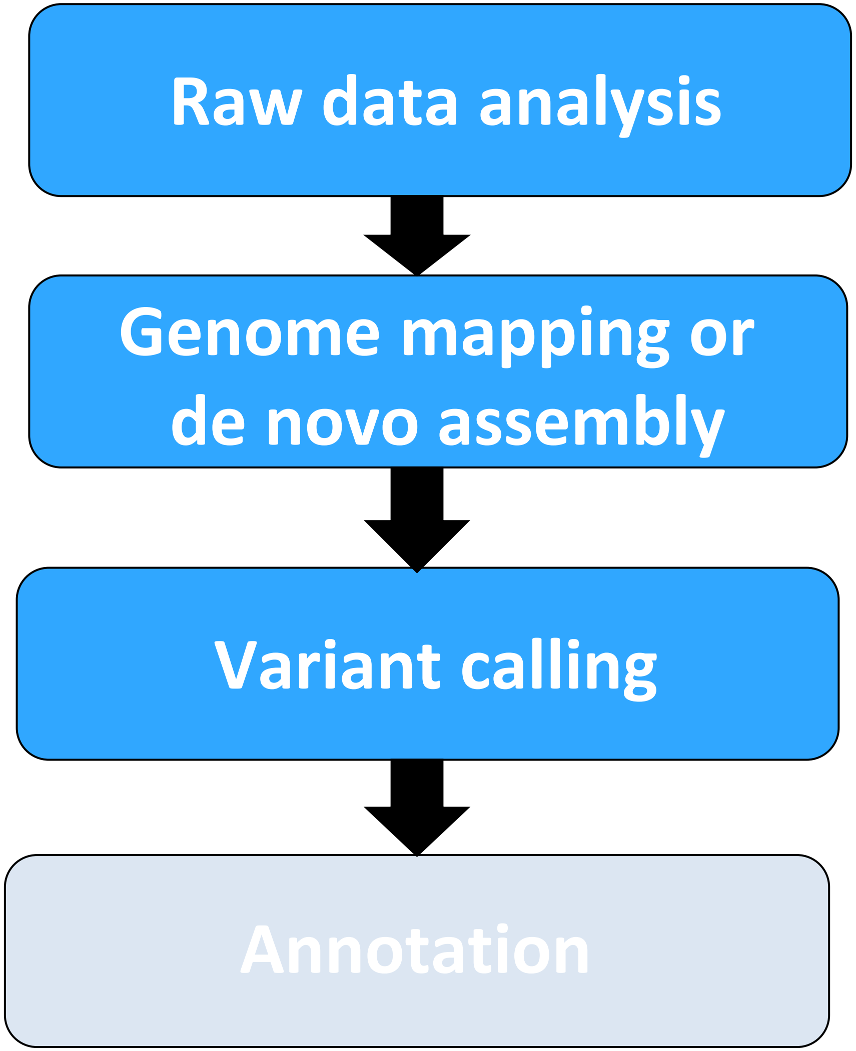
dot

suspend

star

dust

Basic NGS data analysis workflow



Raw data analysis

Image analysis, base calling and data pre-processing

Genome mapping or de novo assembly

Alignment to reference or sequence assembly

Variant calling

Detection of genetic variation (SNP/SNV, SV, CNV)

Annotation

Linking variants to biological information

Fastq format

```
@HISEQ2000:128:D230MACXX:3:1101:1083:2161 1:Y:0:TAGCTT
CCATAGAAAGACTGGTTTGNNNANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+
;;5=;?>?>@><>??>>@?###3##############################################
```

1.	@HISEQ2000	the unique instrument name
2.	128:D230MACXX	the run id, the flowcell id
3.	3	flowcell lane
4.	1101	tile number within the flowcell lane
5.	1083:2161	'x'-coordinate of the cluster within the tile:'y'- coordinate of the cluster within the tile
6.	1	the member of a pair, 1 or 2 (paired-end or mate-pair reads only)
7.	Y	Y if the read fails filter (read is bad), N otherwise
8.	0	0 when none of the control bits are on, otherwise it is an even number
9.	TAGCTT	index sequence

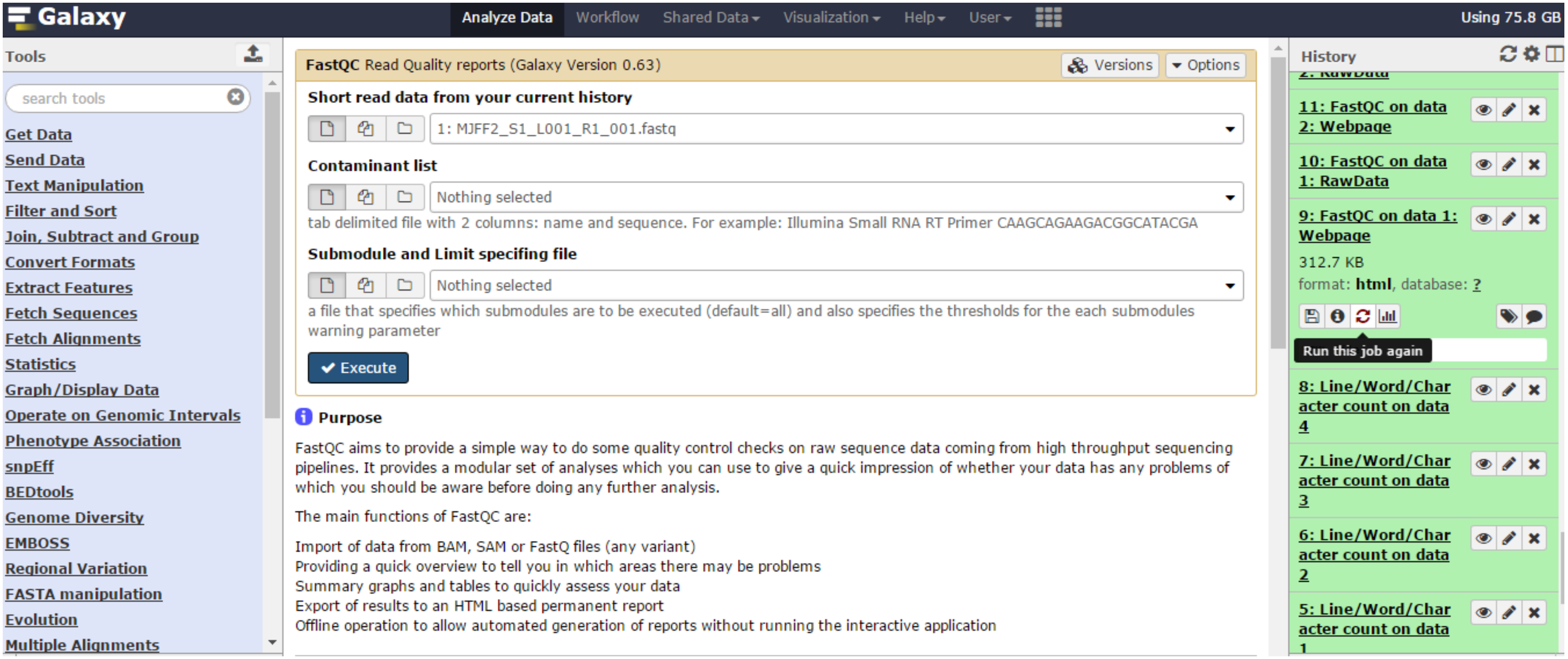
Quality value interpretation

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

$$\boldsymbol{Q = -10\ \log_{10} P}$$

Q = Phred Quality Scores
P = Base-calling error probabilities

FASTQC



FastQC Read Quality reports (Galaxy Version 0.63)

Short read data from your current history

1: MJFF2_S1_L001_R1_001.fastq

Contaminant list

Nothing selected

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifing file

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Purpose

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

Import of data from BAM, SAM or FastQ files (any variant)
Providing a quick overview to tell you in which areas there may be problems
Summary graphs and tables to quickly assess your data
Export of results to an HTML based permanent report
Offline operation to allow automated generation of reports without running the interactive application

2: RawData

11: FastQC on data 2: Webpage

10: FastQC on data 1: RawData

9: FastQC on data 1: Webpage

312.7 KB

format: **html**, database: ?

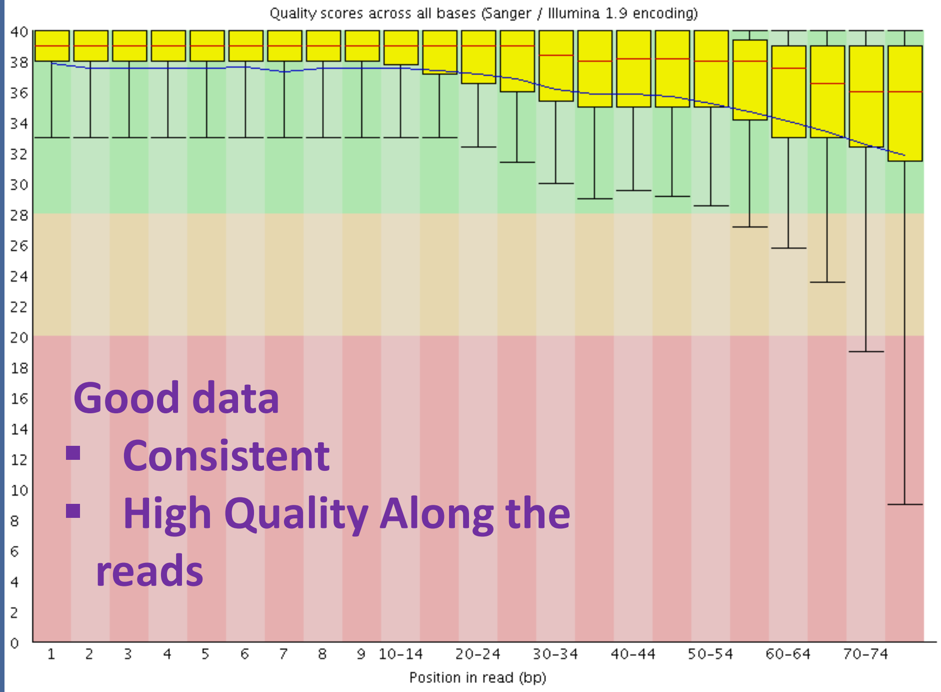
8: Line/Word/Character count on data 4

7: Line/Word/Character count on data 3

6: Line/Word/Character count on data 2

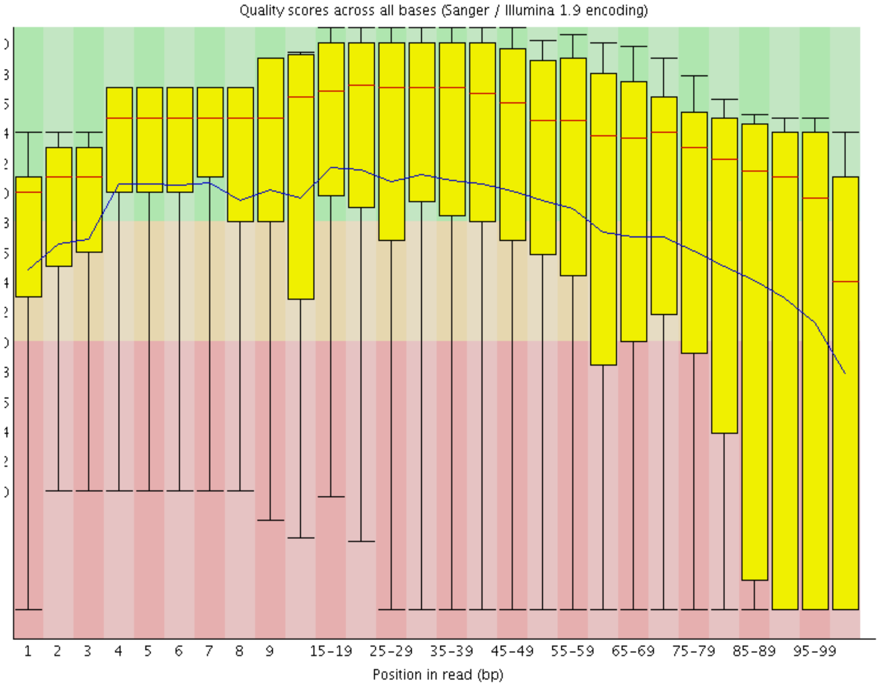
5: Line/Word/Character count on data 1

Sequence quality per base position



Good data

- **Consistent**
- **High Quality Along the reads**

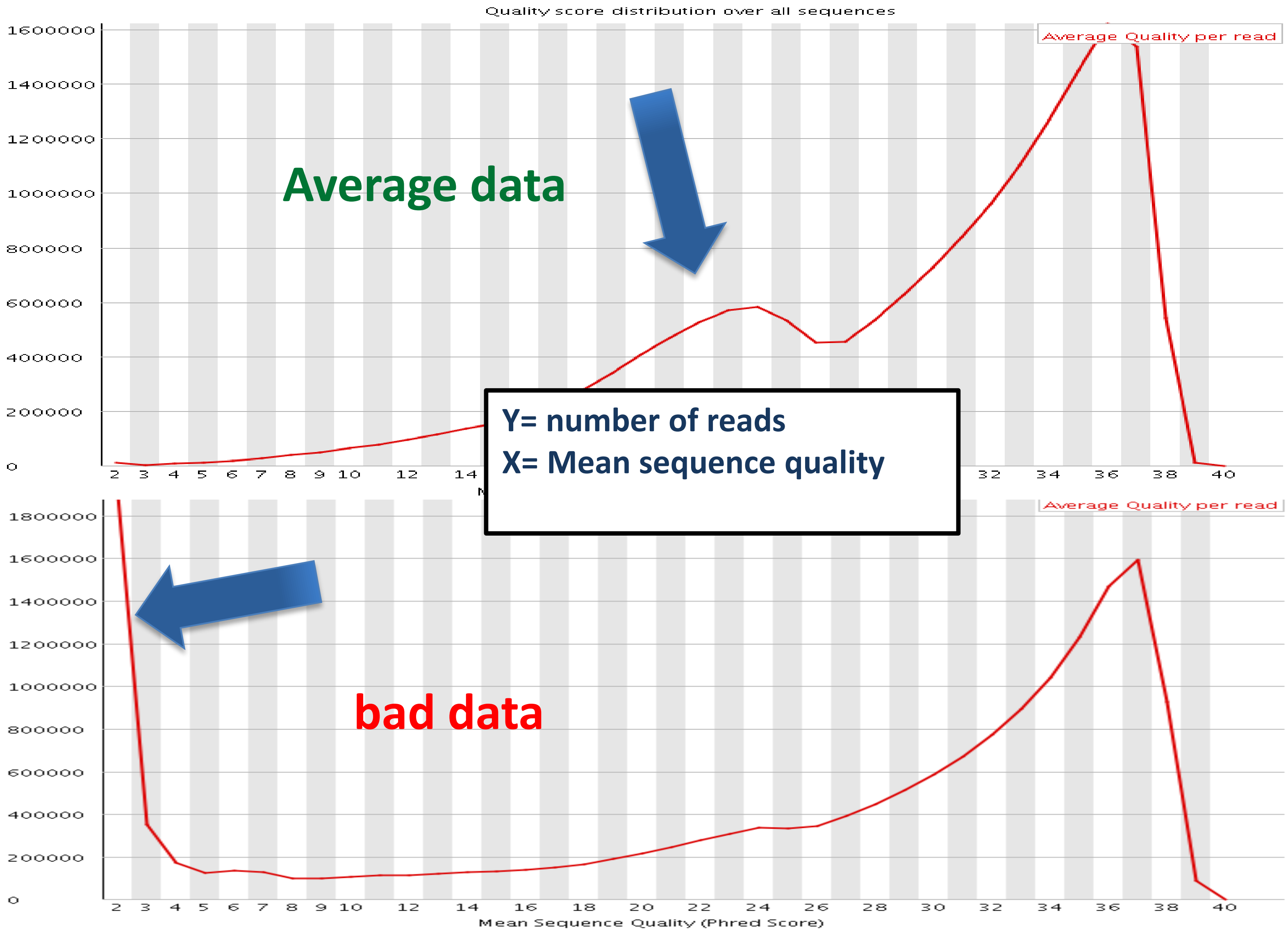


Bad data

- **High Variance**
- **Quality Decrease with Length**

- **The central red line is the median value**
- **The yellow box represents the inter-quartile range (25-75%)**
- **The upper and lower whiskers represent the 10% and 90% points**
- **The blue line represents the mean quality**

Per sequence quality distribution



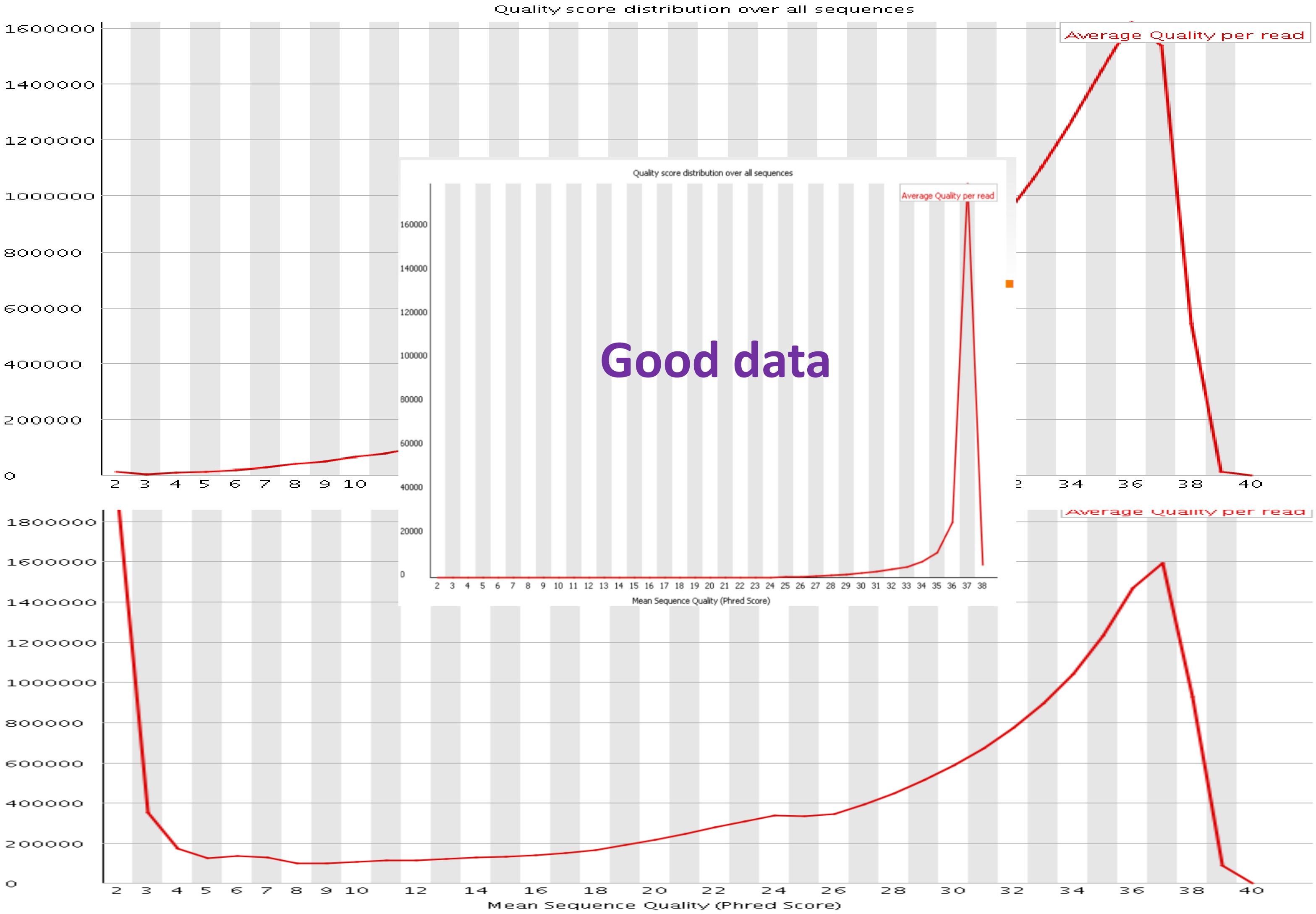
Average data

bad data

Y= number of reads

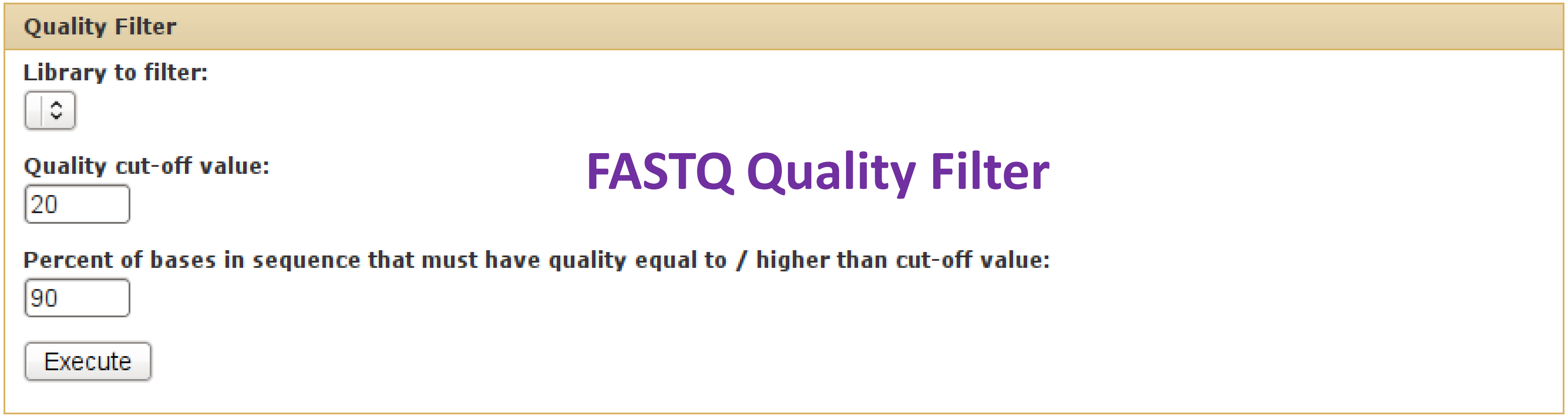
X= Mean sequence quality

Per sequence quality distribution



Good data

Galaxy tool for filtering and trimming



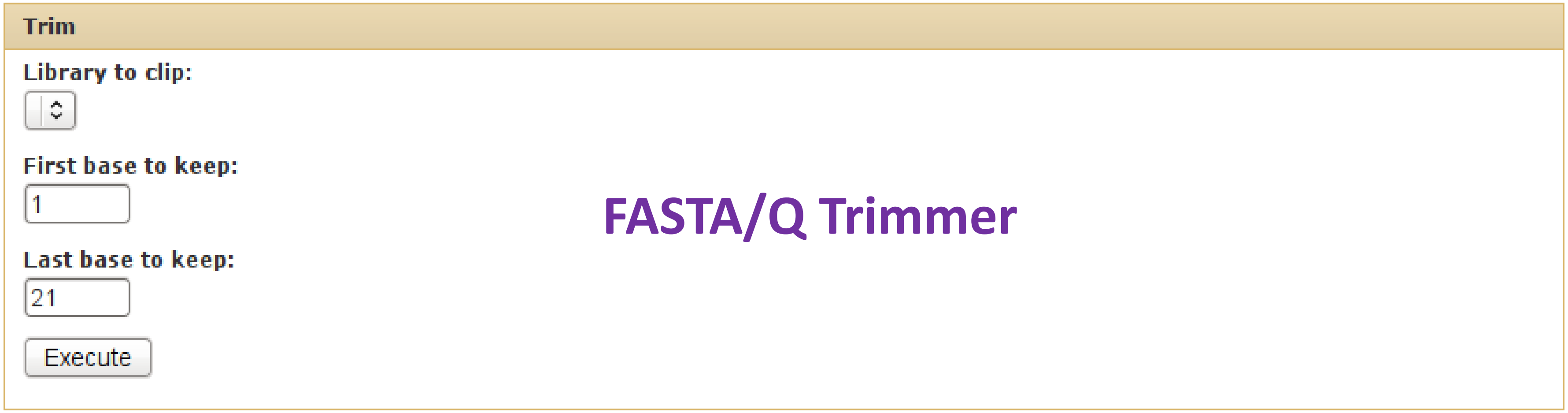
Quality Filter

Library to filter:

Quality cut-off value:
20

Percent of bases in sequence that must have quality equal to / higher than cut-off value:
90

FASTQ Quality Filter



Trim

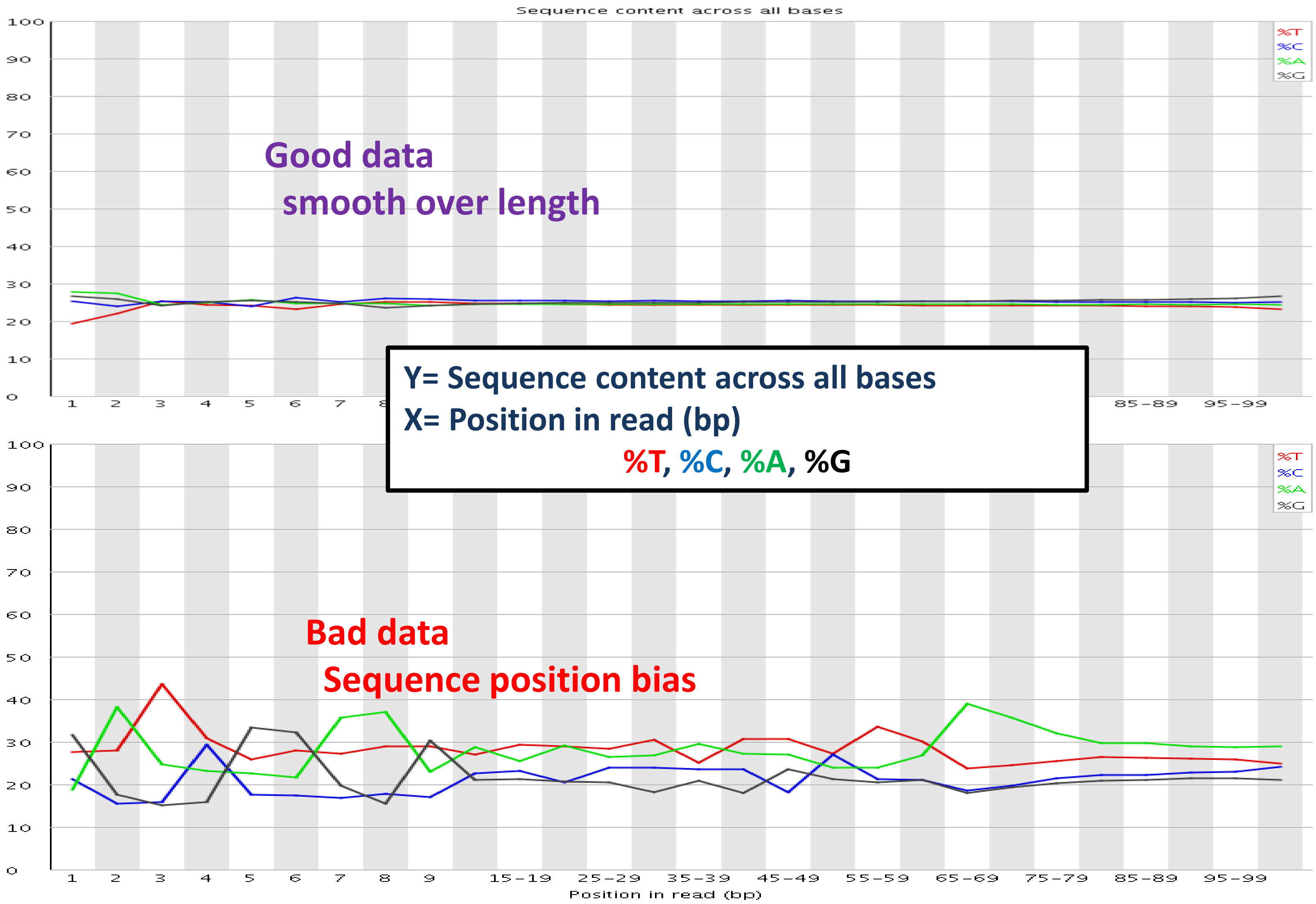
Library to clip:

First base to keep:
1

Last base to keep:
21

FASTA/Q Trimmer

Nucleotide content per position

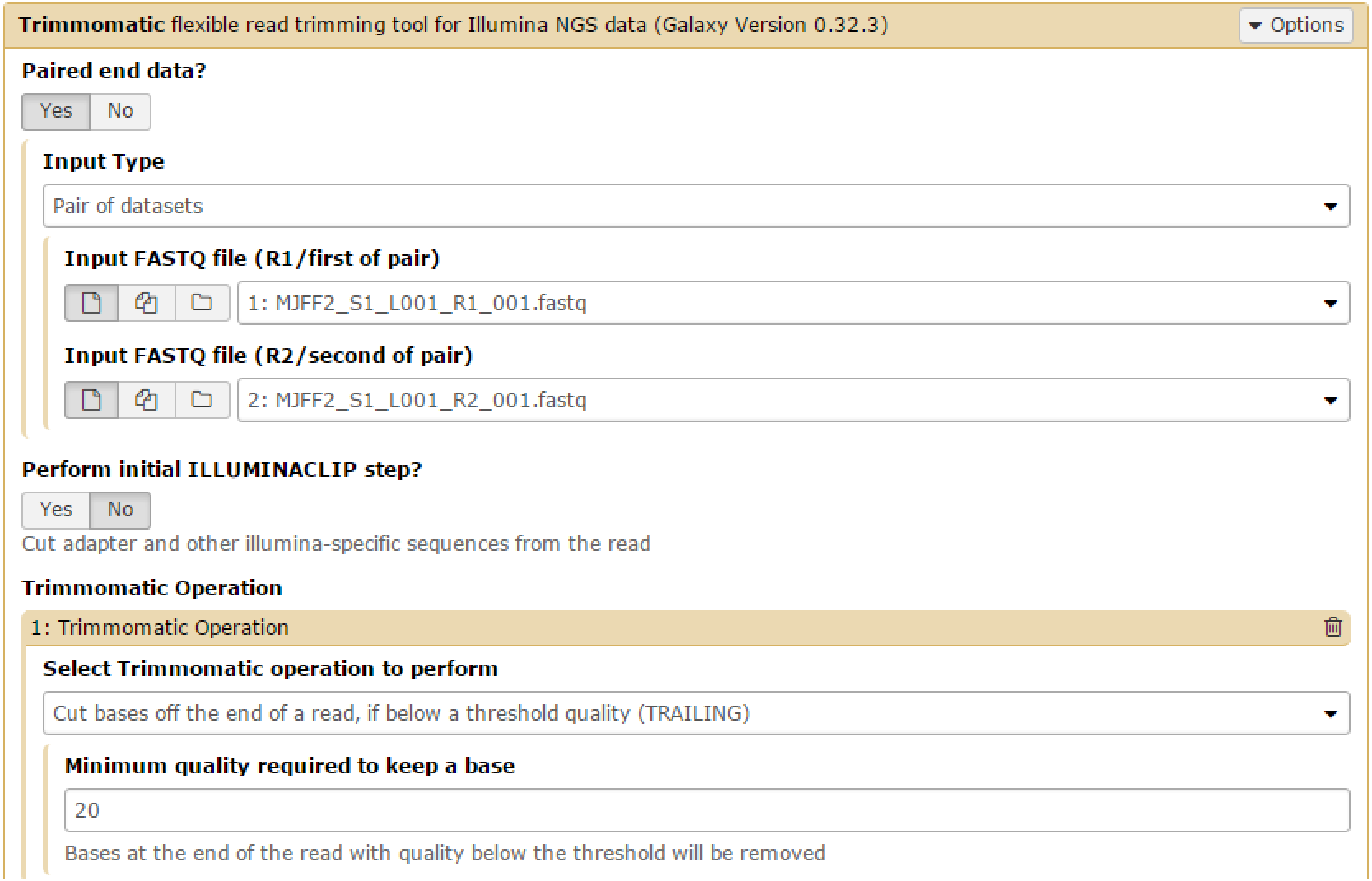


Good data
smooth over length

Y= Sequence content across all bases
X= Position in read (bp)
%T, %C, %A, %G

Bad data
Sequence position bias

Trimmomatic



Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.32.3)

Paired end data?

[✓ Yes] No

Input Type

Pair of datasets

Input FASTQ file (R1/first of pair)

1: MJFF2_S1_L001_R1_001.fastq

Input FASTQ file (R2/second of pair)

2: MJFF2_S1_L001_R2_001.fastq

Perform initial ILLUMINACLIP step?

Yes [✓ No]

Cut adapter and other illumina-specific sequences from the read

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

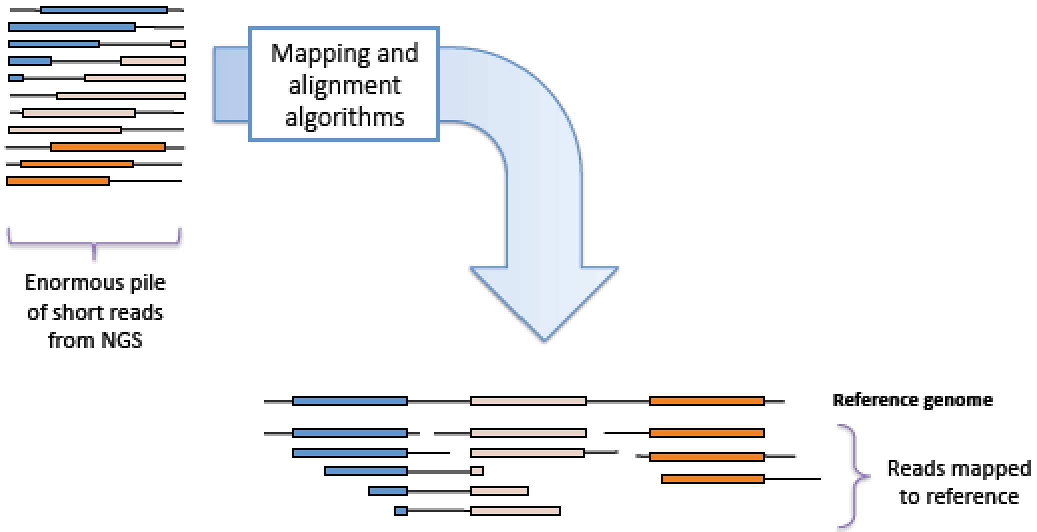
Cut bases off the end of a read, if below a threshold quality (TRAILING)

Minimum quality required to keep a base

20

Bases at the end of the read with quality below the threshold will be removed

Instruments generate short reads that need to be mapped to the reference


Mapping and alignment algorithms
Enormous pile of short reads from NGS
Reference genome
Reads mapped to reference

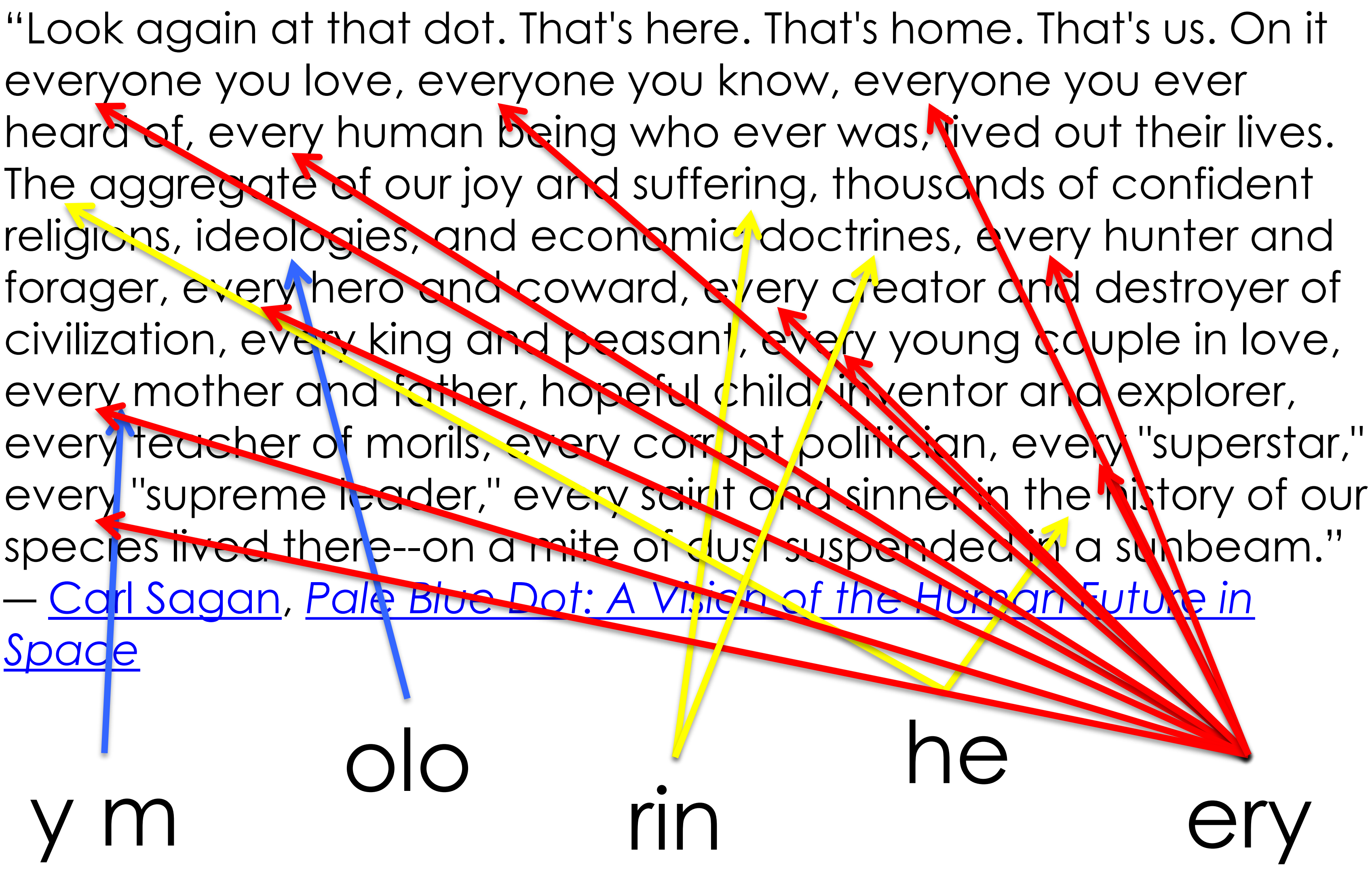
Aligning (mapping) short sequencing reads

1. Drop of water in the sea problem

2. Identical (or near identical) drops

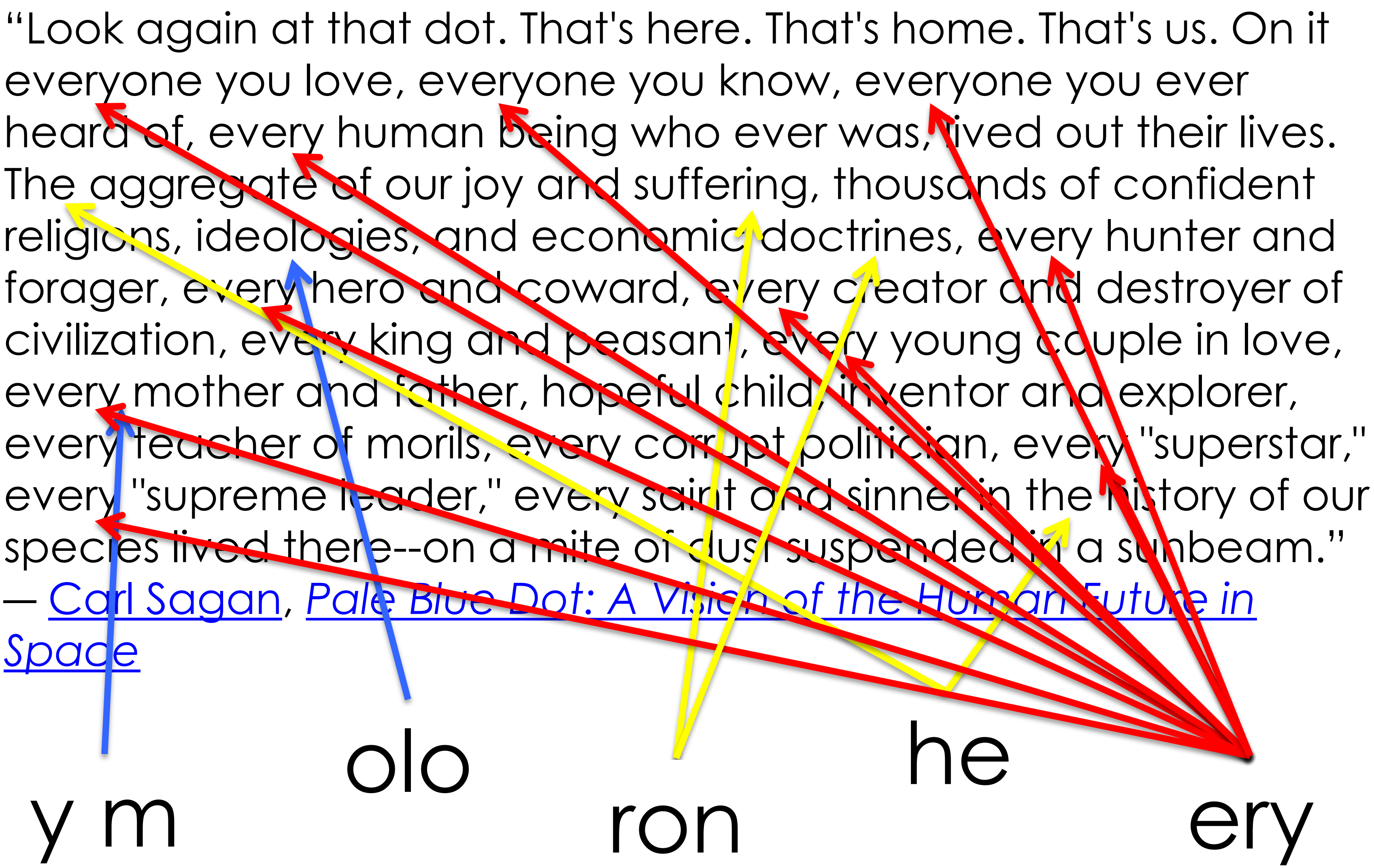
3. Drops of polluted water

Aligning (mapping) short sequencing reads



“Look again at that dot. That's here. That's home. That's us. On it everyone you love, everyone you know, everyone you ever heard of, every human being who ever was, lived out their lives. The aggregate of our joy and suffering, thousands of confident religions, ideologies, and economic doctrines, every hunter and forager, every hero and coward, every creator and destroyer of civilization, every king and peasant, every young couple in love, every mother and father, hopeful child, inventor and explorer, every teacher of morils, every corrupt politician, every "superstar," every "supreme leader," every saint and sinner in the history of our species lived there--on a mite of dust suspended in a sunbeam.”
— Carl Sagan, *Pale Blue Dot: A Vision of the Human Future in Space*

Aligning (mapping) short sequencing reads



"Look again at that dot. That's here. That's home. That's us. On it everyone you love, everyone you know, everyone you ever heard of, every human being who ever was, lived out their lives. The aggregate of our joy and suffering, thousands of confident religions, ideologies, and economic doctrines, every hunter and forager, every hero and coward, every creator and destroyer of civilization, every king and peasant, every young couple in love, every mother and father, hopeful child, inventor and explorer, every teacher of morils, every corrupt politician, every "superstar," every "supreme leader," every saint and sinner in the history of our species lived there--on a mite of dust suspended in a sunbeam."
― Carl Sagan, *Pale Blue Dot: A Vision of the Human Future in Space*

y m

olo

ron

he

ery

DNA Sequence alignment

- sequence analysis → sequence alignment
- what:

```
  aca--gacgcagtactttg-g-gc-caga-ac-cgt
       |||  ||| ||||  | || |||| || |||
cgacacagacgcagt-ctttgtgtgctcacacacacgtgct
```

- why:
  - similar sequence
  - infer homology
  - infer function

sequence → structure → function

Global vs. Local

```
SPQ-RTGKCCWIAGPGILHRMSL
|   |      ||||   | ||
SGALRCSWND-IAGPCAQH-MSA
```

Global: Needleman-Wunsch; similar length, highly similar sequences

Local: Smith-Waterman; finds region(s) of highest similarity and build outward

BLAST

Basic **L**ocal **A**lignment **S**earch **T**ool

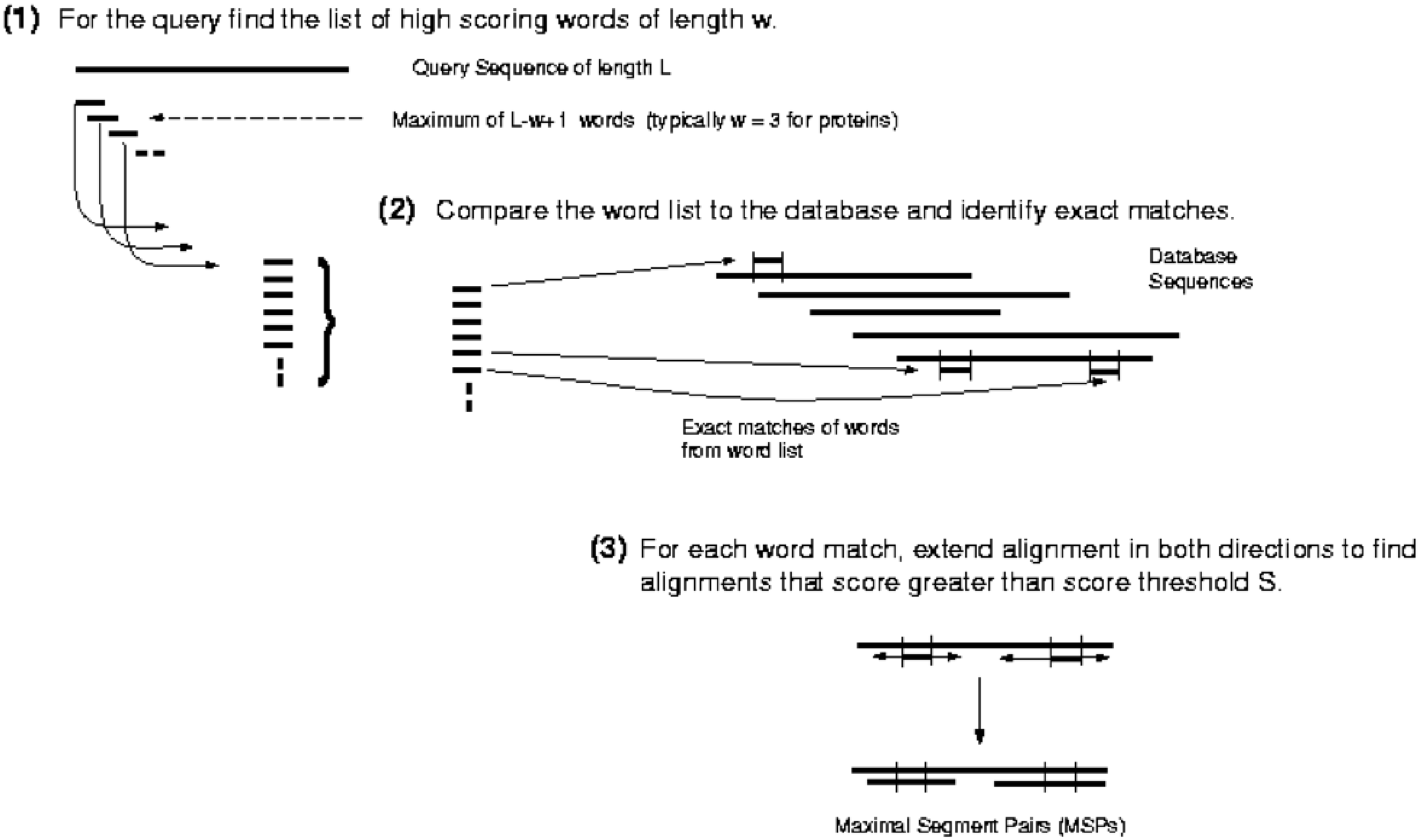
- **idea**: find high scoring local alignments between query sequence and target database

- **assumption**: true match alignments very likely to contain *within them* very high scoring matches

  heuristics theme: search quickly for homologous regions and then do slow/exact alignments

Traditional alignment methods

BLAST



(1) For the query find the list of high scoring words of length w.

Query Sequence of length L
Maximum of L-w+1 words (typically w = 3 for proteins)

(2) Compare the word list to the database and identify exact matches.

Database Sequences
Exact matches of words from word list

(3) For each word match, extend alignment in both directions to find alignments that score greater than score threshold S.

Maximal Segment Pairs (MSPs)

BLAST Steps

1. Seeding

Query word (*W* = 3)

Query: GSDFWQETRASFGCSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFGCATSWPI

Neighborhood words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
…	

Neighborhood score threshold (*T* = 13)

Hit

Query: SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA

Subject: TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA

Determine the locations of all common “words” between the query and the database (“word hits”).

(protein W = 3, DNA W = 11)

BLAST Steps

2. Extension

use dynamic programming to extend hits until the score drops a value of *X* – *expensive!!* -- *90% of time*

```
ABCDEFGHIJKLMNOPQRST
||||||  |||||
ABCDEFZYIJKLMXWVUTAB
12345654567898765654  -> Score
00000012100001234345  -> Drop off score
```

Match = 1
Mismatch = -1
X = 5

Scoring Matrices

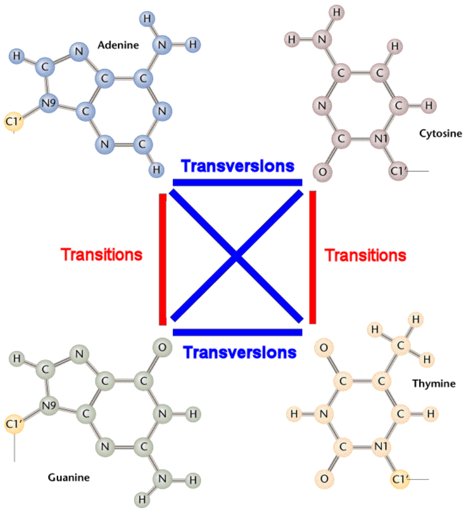
A simple matrix for DNA

	C	T	A	G
C	1	-1	-1	-1
T	-1	1	-1	-1
A	-1	-1	1	-1
G	-1	-1	-1	1

```
ATGGCCATG
|  |||  |
A-CGCCTCG
```

Score = 1

A more sophisticated matrix for DNA



	C	T	A	G
C	2	1	-1	-1
T	1	2	-1	-1
A	-1	-1	2	1
G	-1	-1	1	2

```
ATGGCCATG
|: |||  |
AC-GCCTCG
```

Score = 8

Alignment Overview

- **Reads: short DNA sequences usually up to 100-200 base pairs produced by a sequencing machine**
- **Reference: Genome sequence of organism of interest**
- **Aligner: Short-read aligner (BWA, bowtie, SOAP, MAQ etc.)**
- **Distances:**

  ❑ **Hamming Distance:**

  The hamming distance is defined only for strings of the same length. For two strings, it is the number of places in which the two string differ.

  ❑ **Edit distance:**

  The edit distance between two strings is the minimum number of insertions, deletions and substitutions needed to transform the first string into the second one.

Overview

MICHAEL
MICHELE

A = E

E = L

L = E

Hamming distance = 3

KOBY
BOBBY

K=B
Y=B
Y=inserted

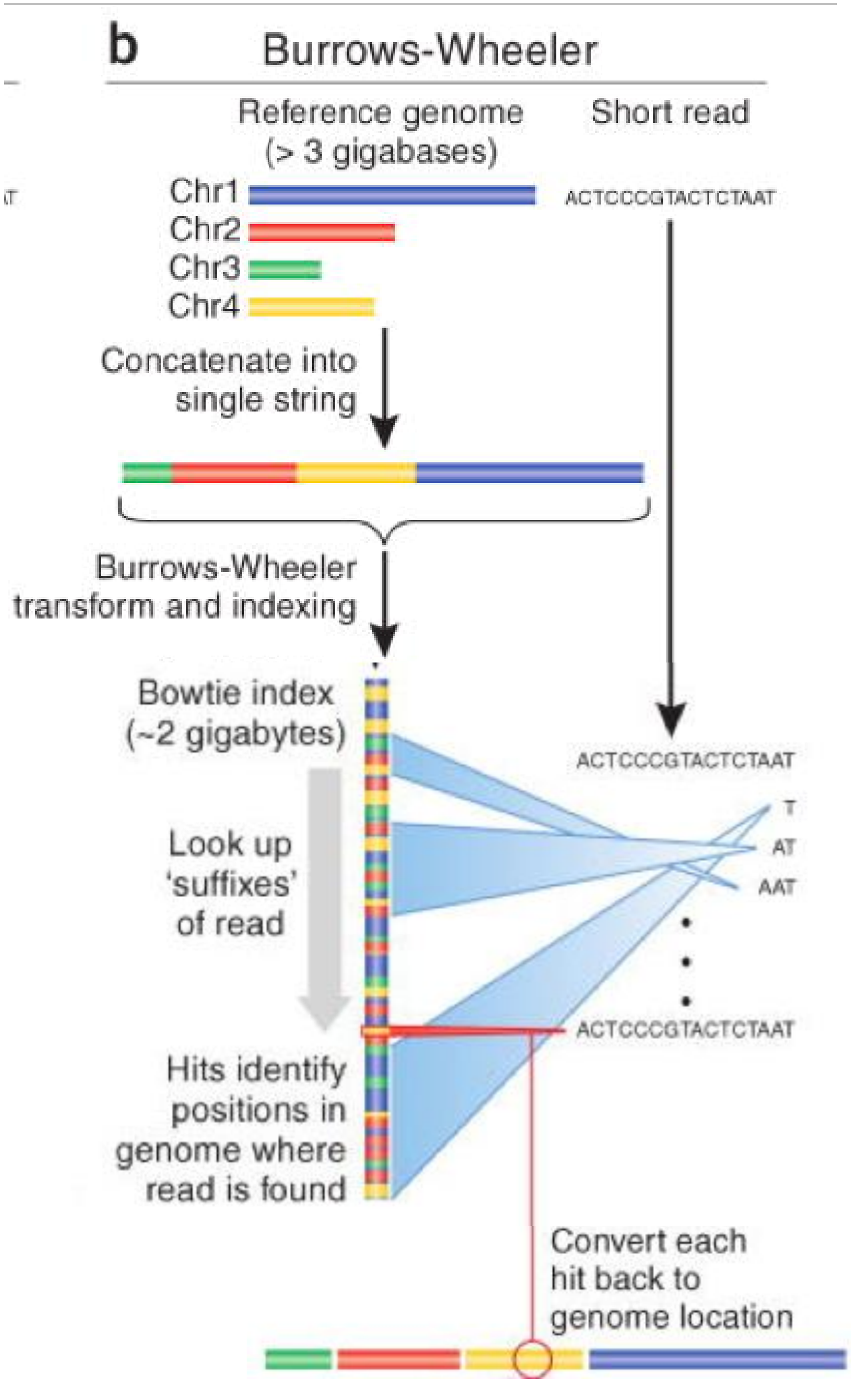
Edit distance = 3

The Burroughs Wheeler Alignment BWA tools

- BWA is used to map low divergent sequence reads to a reference genome
- It assumes that your ‘reads’ are from the genome you are aligning to
- There are three algorithms; the most common being BWA-MEM
- BWA outputs alignments in a SAM format
- Downstream you can use SAM tools or GATK to call variants or whatever

Alignment methods for NGS

Burrows-Wheeler transformation


b
Burrows-Wheeler
Reference genome (> 3 gigabases)
Short read
Chr1
Chr2
Chr3
Chr4
ACTCCCGTACTCTAAT
Concatenate into single string
Burrows-Wheeler transform and indexing
Bowtie index (~2 gigabytes)
Look up 'suffixes' of read
Hits identify positions in genome where read is found
ACTCCCGTACTCTAAT
T
AT
AAT
ACTCCCGTACTCTAAT
Convert each hit back to genome location

Nat Biotechnol. May 2009; 27(5): 455–457.

Tools for NGS aligning

BWA

“There is no such thing (yet) as an automated gearshift in short read mapping. It is all like stick-shift driving in San Francisco. In other words running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to understand the parameters by carefully reading the documentation and experimenting. Fortunately, Galaxy makes experimenting easy.”
- Galaxy Team

Parameters to test

Popular methods for alignment

BWA

BWA-aln:

- ❑ **Short reads up to 200 bp with errors <5%**
- ❑ **Gapped alignment**
- ❑ **global alignment**
- ❑ **Can do paired-end**
- ❑ **Report ambiguous hits**

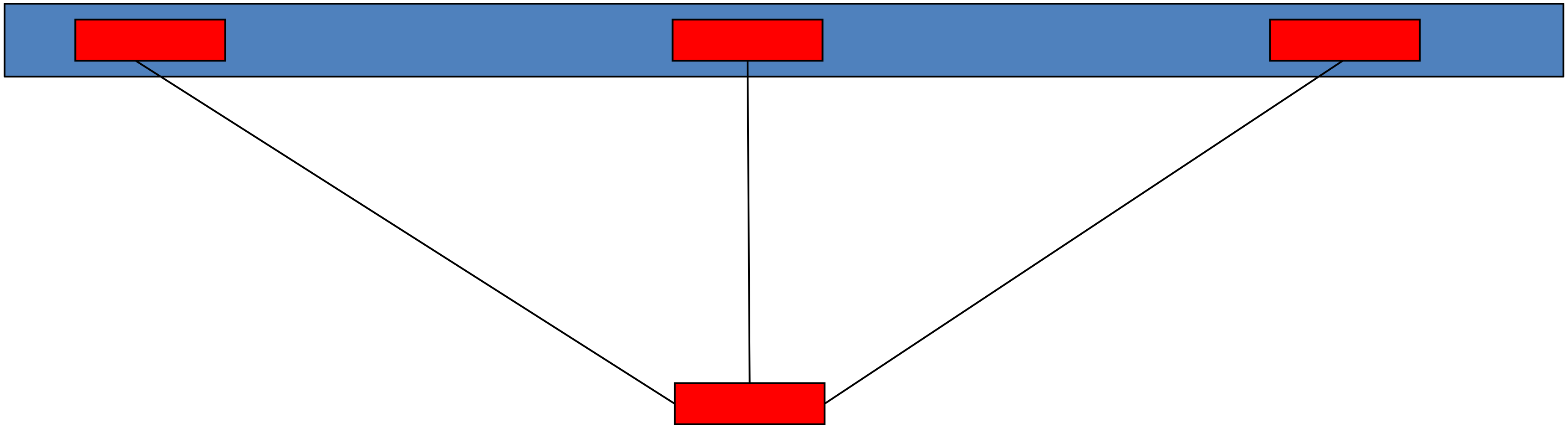
BWA-SW:

- ❑ **Can align longer reads (upto 1mbp)**
- ❑ **Local alignment**
- ❑ **The paired-end mode only works for Illumina short-insert libraries.**

BWA-MEM is the latest BWA algorithm – much faster

BWA outputs the final alignment in the SAM (Sequence Alignment/Map) format

Mapping Quality



- Reads can occur more than once in the reference genome
- One can restrict the analysis to exclude the reads which occur more than n times
- As n gets larger, one gets more data, but also more noise in the data

Mapping Quality

Understanding mapping qualities:

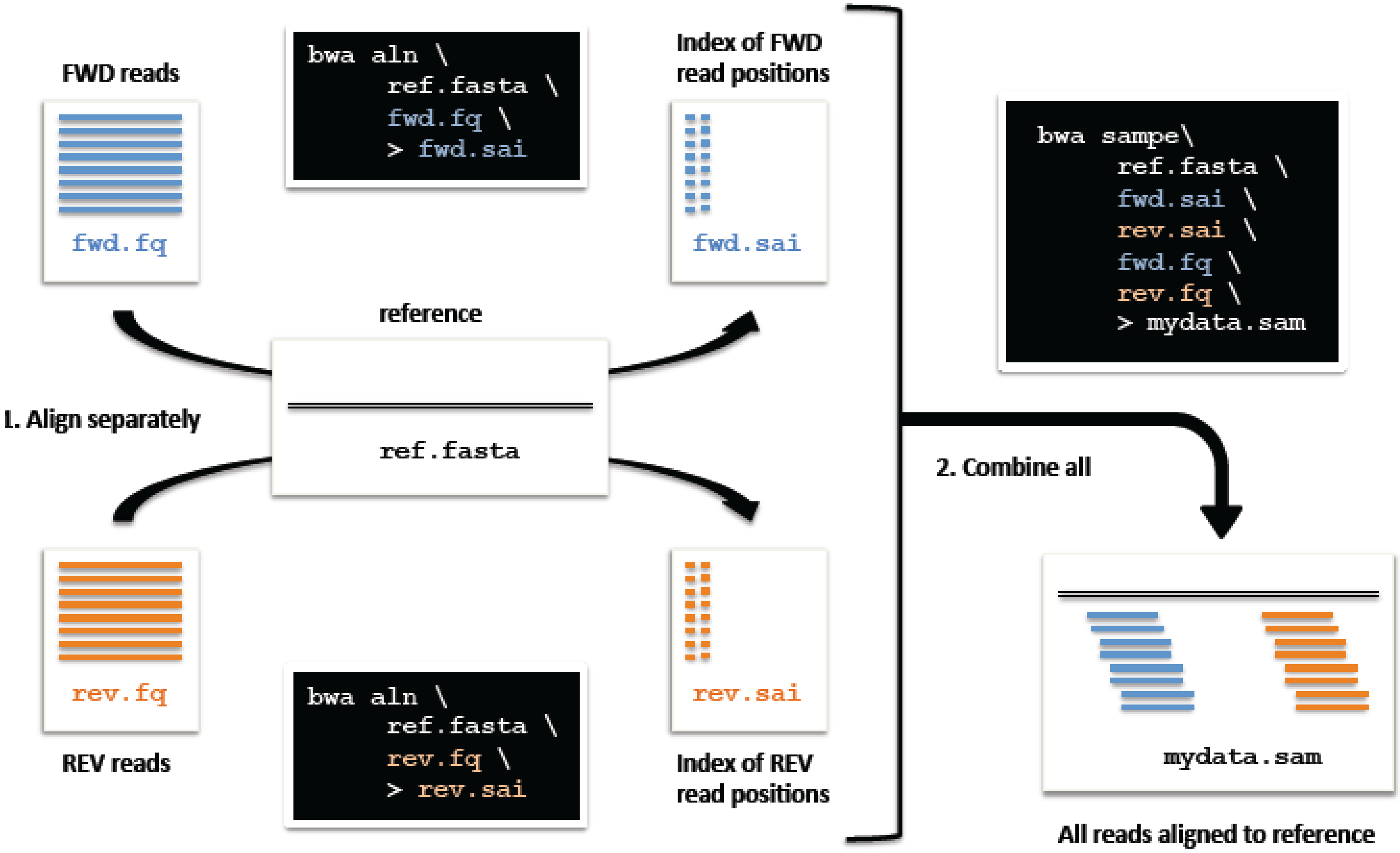
Mapping quality calculation consider all these factors:

- ❑ **Reference genome repeat structure**
- ❑ **Base qualities of read**
- ❑ **Paired end or not**

***Qs* = 30 implies there is a 1 in 1000 probability that the read is incorrectly mapped.**

- ❑ **The overall base quality of the read is good**
- ❑ **The best alignment has fewer mismatches**
- ❑ **The read has not matched to many places in genome**

Typical workflow using BWA to map paired-end data

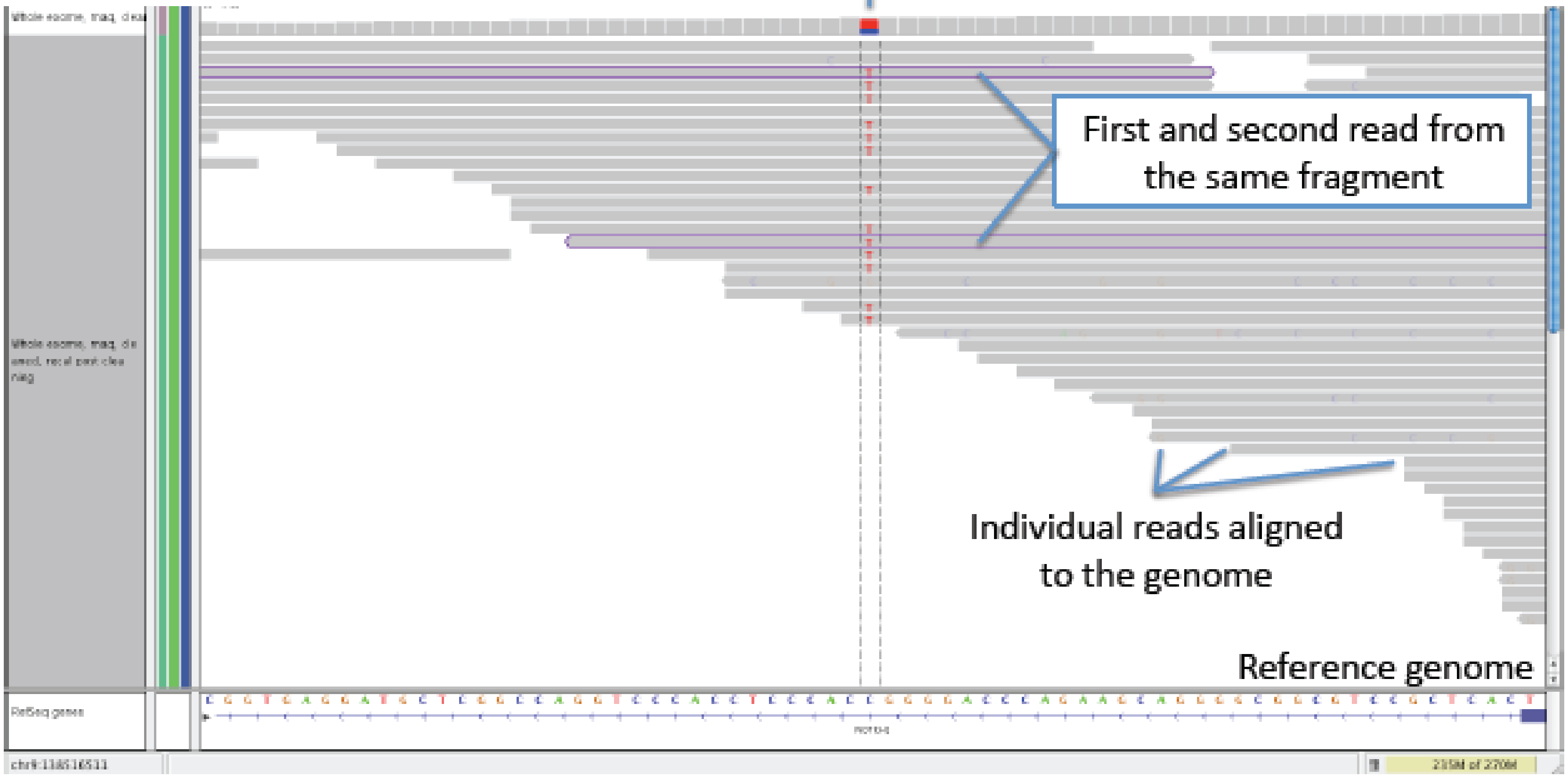


```
bwa aln \
     ref.fasta \
     fwd.fq \
     > fwd.sai
```

```
bwa aln \
     ref.fasta \
     rev.fq \
     > rev.sai
```

```
bwa sampe\
     ref.fasta \
     fwd.sai \
     rev.sai \
     fwd.fq \
     rev.fq \
     > mydata.sam
```

Prototypical IGV screenshot representing aligned NGS reads



Non-reference bases are colored; reference bases are grey

Clean C/T heterozygote

Depth of coverage

First and second read from the same fragment

Individual reads aligned to the genome

Reference genome

SAM format

SAM stands for Sequence Alignment/Map format

- ❑ **TAB-delimited text format consisting of optional header section and an alignment section**
- ❑ **Header lines start with “@” symbol; alignment lines don’t**
- ❑ **Each alignment line has 11 mandatory fields for essential alignment information**
- ❑ **Variable number of optional fields for flexible or aligner specific information.**

It’s compact version is BAM format (Binary alignment MAP)

SAM format

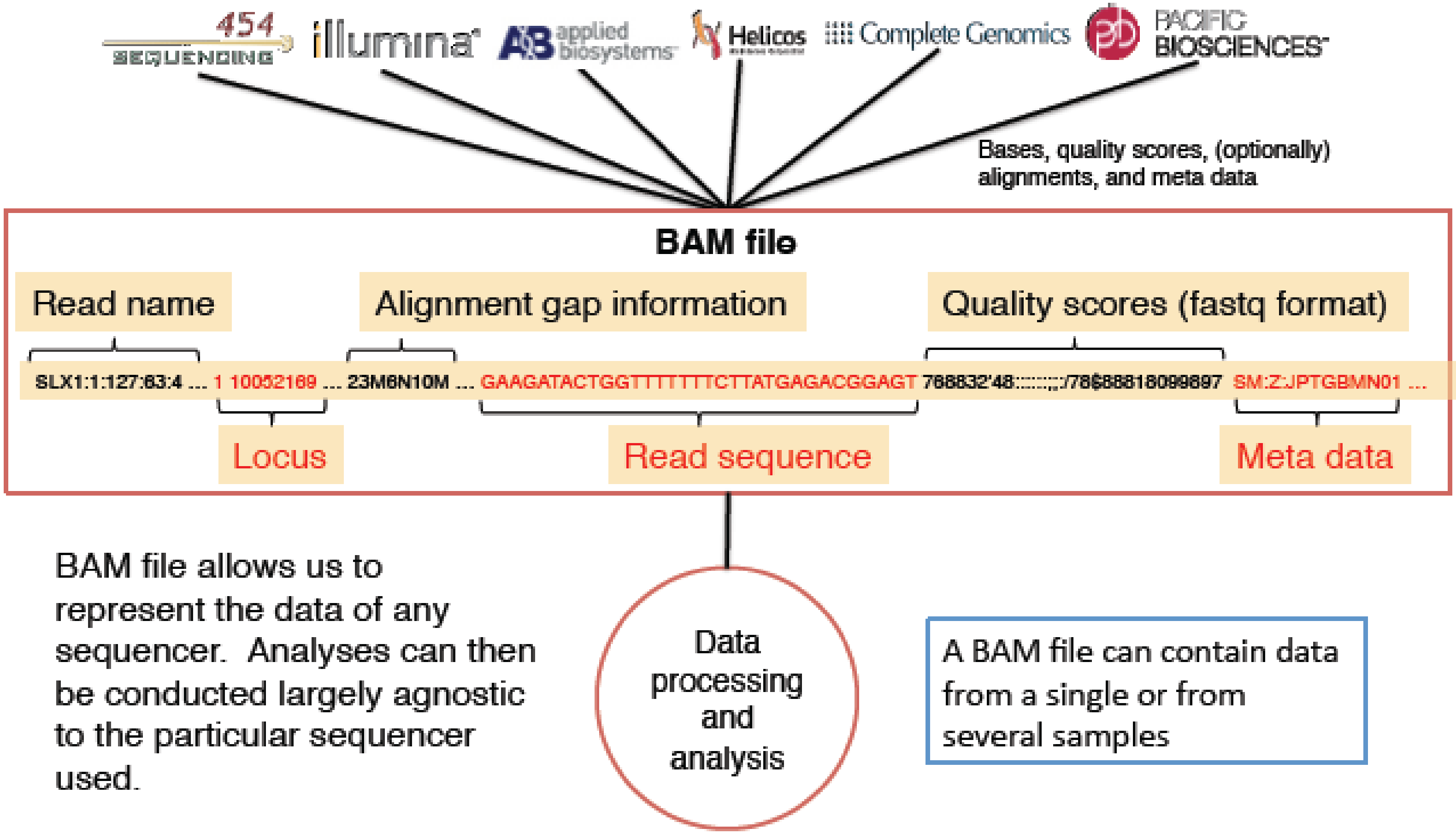
- ❑ **Alignment file SAM is converted to BAM format for efficient storage and access to alignment information**

- ❑ **BAM is also indexed to allow access to portions of information without loading the whole file**

- ❑ **BAM are re-ordered mostly by chromosomal coordinates**

- ❑ **Alignment paired or unpaired or different samples could be merged by samtools**

The BAM format stores aligned reads and is technology independent

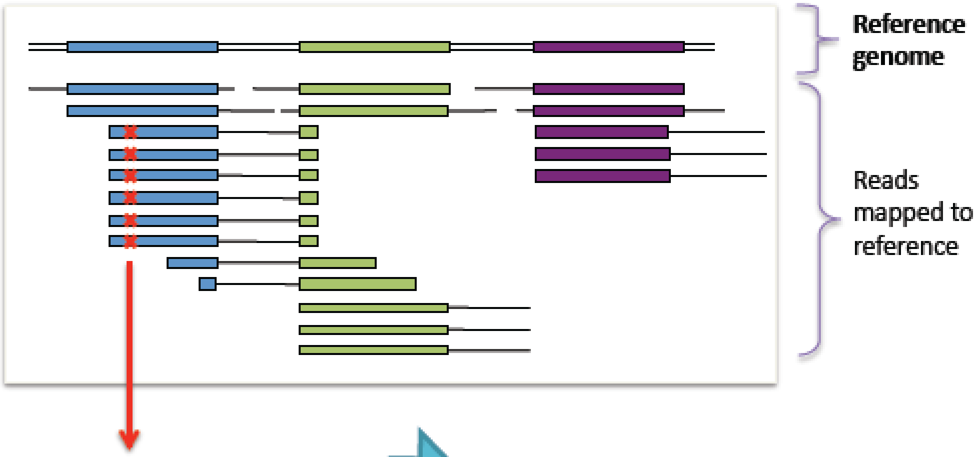

454 SEQUENCING
illumina
AB applied biosystems
Helicos
Complete Genomics
PACIFIC BIOSCIENCES
Bases, quality scores, (optionally) alignments, and meta data
BAM file
Read name
Alignment gap information
Quality scores (fastq format)
SLX1:1:127:63:4 ... 1 10052169 ... 23M6N10M ... GAAGATACTGGTTTTTTTCTTATGAGACGGAGT 768832'48::::::;;;/78$88818099897 SM:Z:JPTGBMN01 ...
Locus
Read sequence
Meta data
Data processing and analysis

BAM file allows us to represent the data of any sequencer. Analyses can then be conducted largely agnostic to the particular sequencer used.

A BAM file can contain data from a single or from several samples

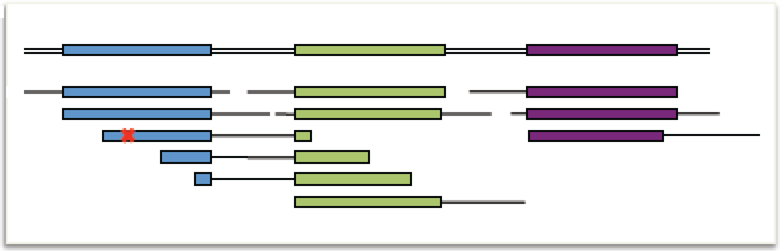
Cleaning up BAM Alignments
Avoiding bad data

✖ = sequencing error propagated in duplicates



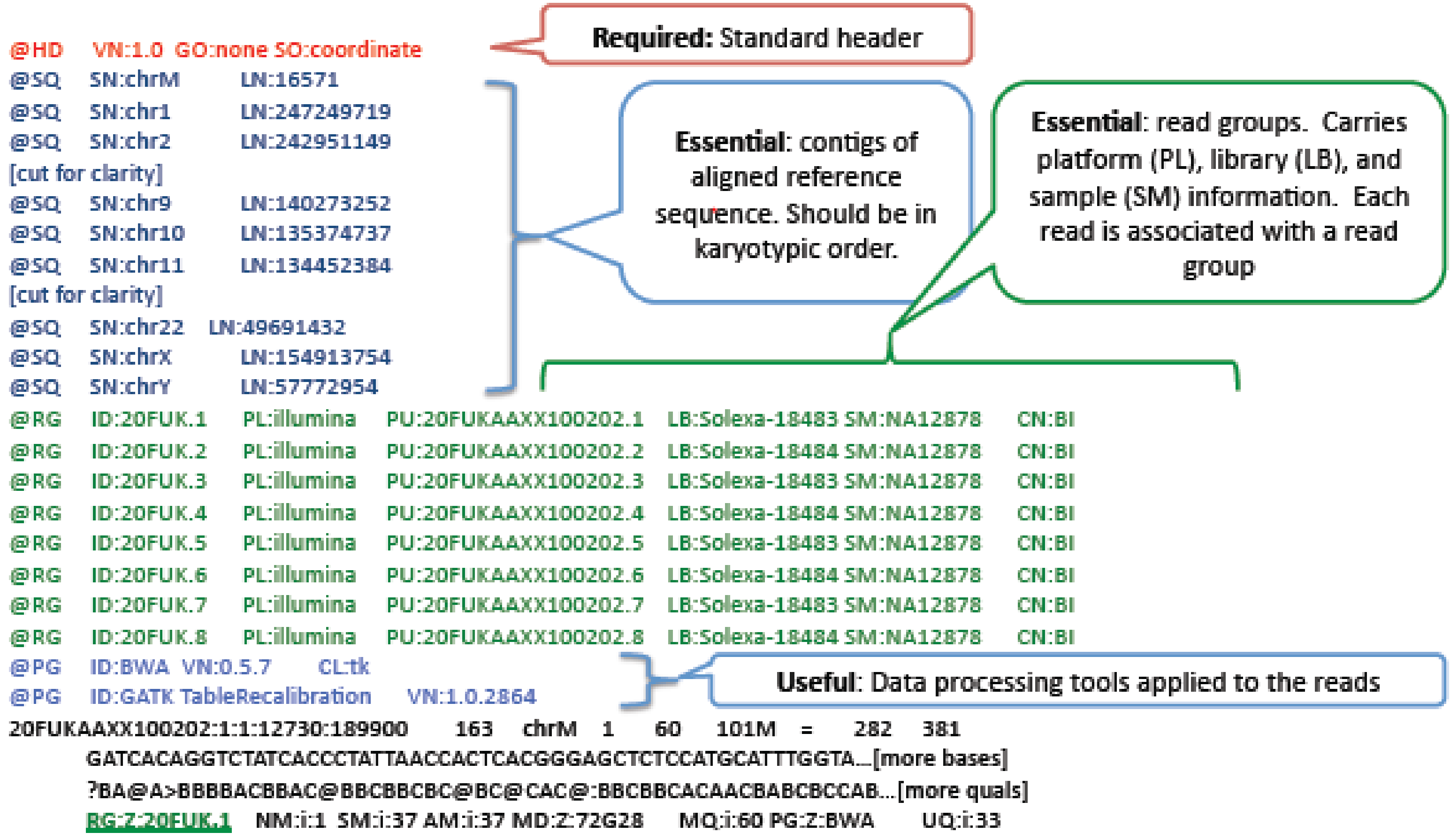
FP variant call (bad)

After marking duplicates, the GATK will only see :



… and thus be more likely to make the right call

BAM headers: an essential part of a BAM file



```
@HD    VN:1.0  GO:none SO:coordinate
@SQ    SN:chrM        LN:16571
@SQ    SN:chr1        LN:247249719
@SQ    SN:chr2        LN:242951149
[cut for clarity]
@SQ    SN:chr9        LN:140273252
@SQ    SN:chr10       LN:135374737
@SQ    SN:chr11       LN:134452384
[cut for clarity]
@SQ    SN:chr22   LN:49691432
@SQ    SN:chrX        LN:154913754
@SQ    SN:chrY        LN:57772954
@RG    ID:20FUK.1    PL:illumina    PU:20FUKAAXX100202.1    LB:Solexa-18483 SM:NA12878    CN:BI
@RG    ID:20FUK.2    PL:illumina    PU:20FUKAAXX100202.2    LB:Solexa-18484 SM:NA12878    CN:BI
@RG    ID:20FUK.3    PL:illumina    PU:20FUKAAXX100202.3    LB:Solexa-18483 SM:NA12878    CN:BI
@RG    ID:20FUK.4    PL:illumina    PU:20FUKAAXX100202.4    LB:Solexa-18484 SM:NA12878    CN:BI
@RG    ID:20FUK.5    PL:illumina    PU:20FUKAAXX100202.5    LB:Solexa-18483 SM:NA12878    CN:BI
@RG    ID:20FUK.6    PL:illumina    PU:20FUKAAXX100202.6    LB:Solexa-18484 SM:NA12878    CN:BI
@RG    ID:20FUK.7    PL:illumina    PU:20FUKAAXX100202.7    LB:Solexa-18483 SM:NA12878    CN:BI
@RG    ID:20FUK.8    PL:illumina    PU:20FUKAAXX100202.8    LB:Solexa-18484 SM:NA12878    CN:BI
@PG    ID:BWA  VN:0.5.7      CL:tk
@PG    ID:GATK TableRecalibration    VN:1.0.2864
20FUKAAXX100202:1:1:12730:189900     163    chrM    1     60     101M    =     282    381
        GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTA…[more bases]
        ?BA@A>BBBBACBBAC@BBCBBCBC@BC@CAC@:BBCBBCACAACBABCBCCAB…[more quals]
        RG:Z:20FUK.1   NM:i:1  SM:i:37 AM:i:37 MD:Z:72G28     MQ:i:60 PG:Z:BWA      UQ:i:33
```

Required: Standard header

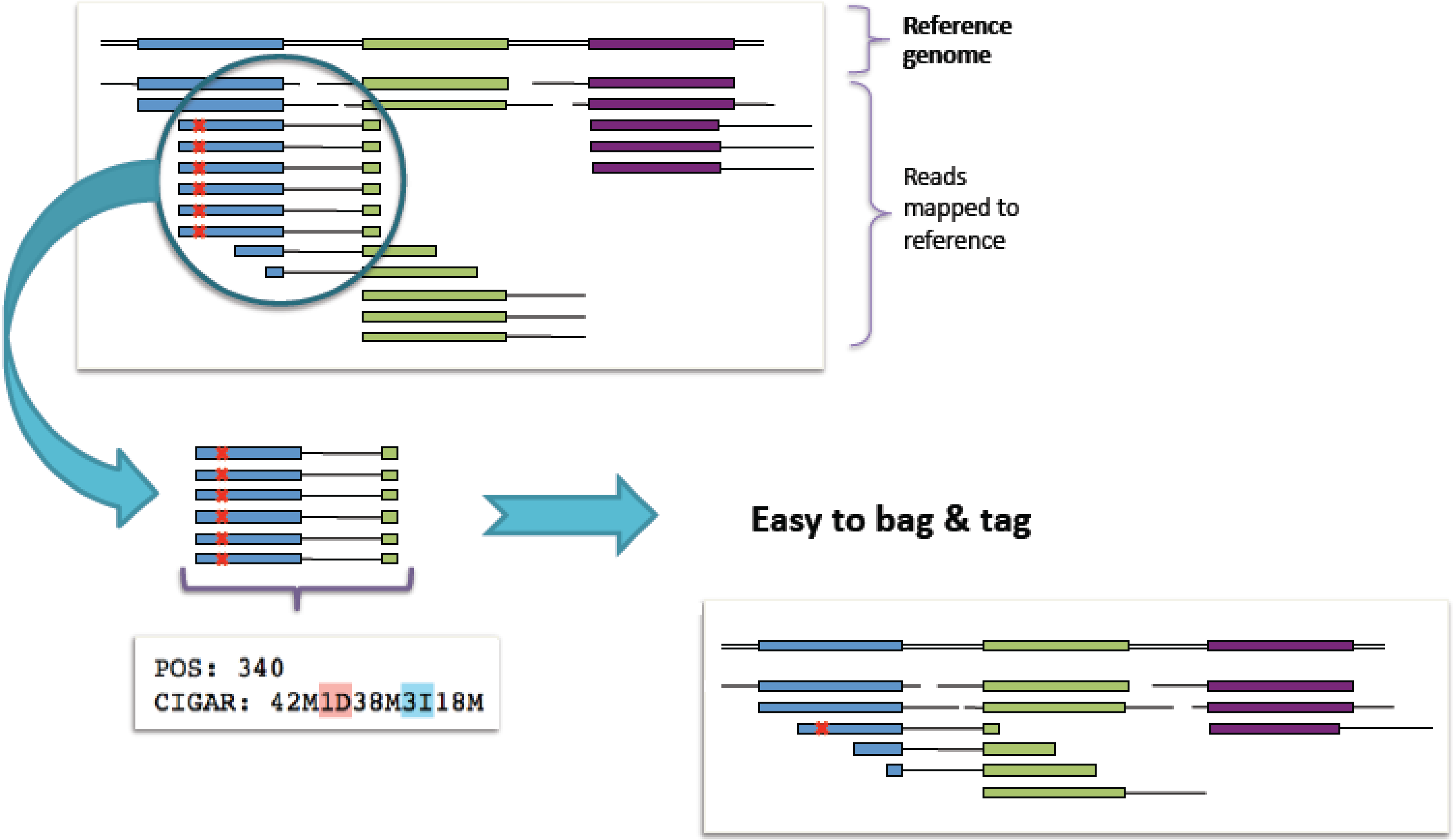
Essential: contigs of aligned reference sequence. Should be in karyotypic order.

Essential: read groups. Carries platform (PL), library (LB), and sample (SM) information. Each read is associated with a read group

Useful: Data processing tools applied to the reads

Official specification in http://samtools.sourceforge.net/SAM1.pdf

Identifying duplicates



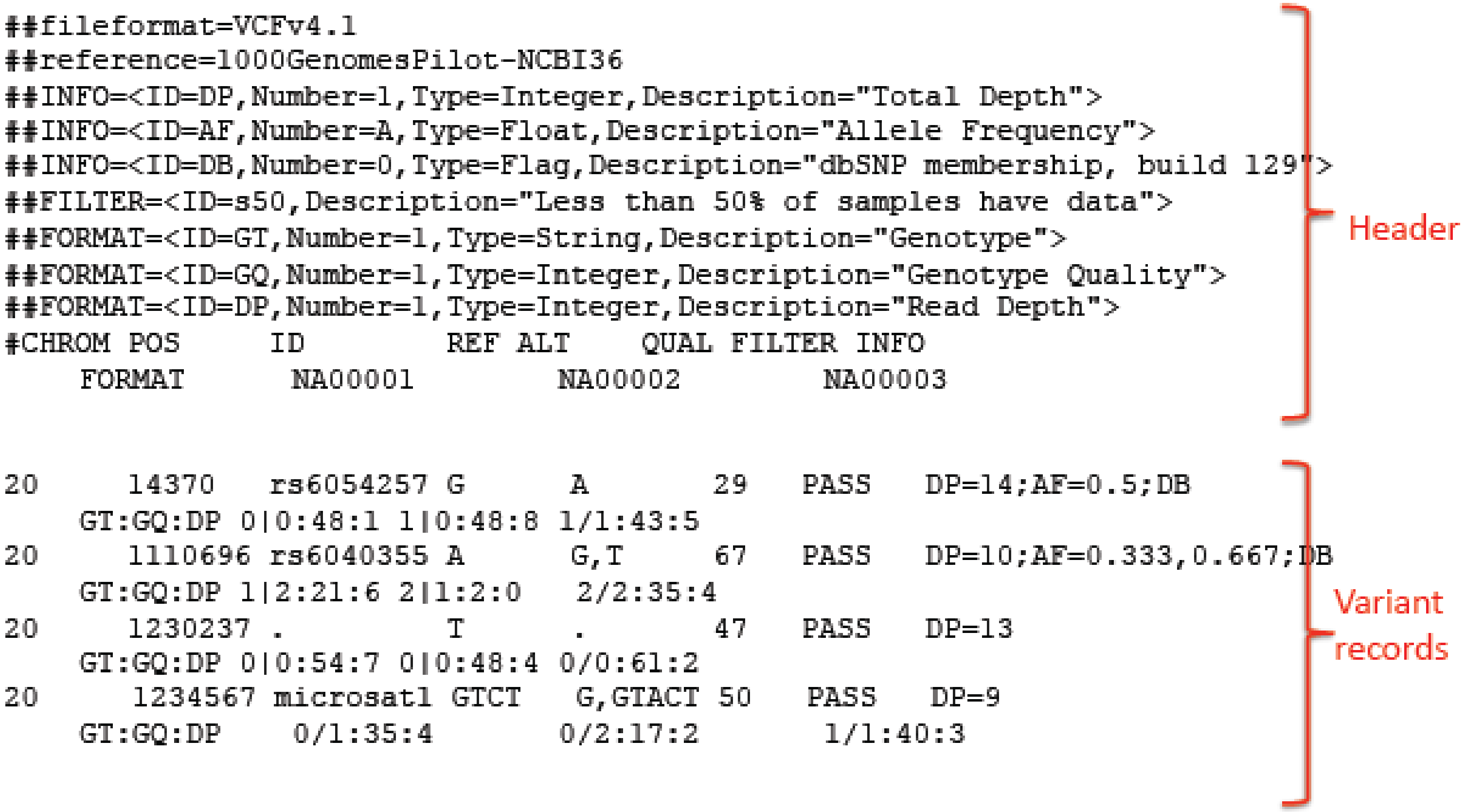
Reference genome

Reads mapped to reference

```
POS: 340
CIGAR: 42M1D38M3I18M
```

Easy to bag & tag

VCF Files store variant information

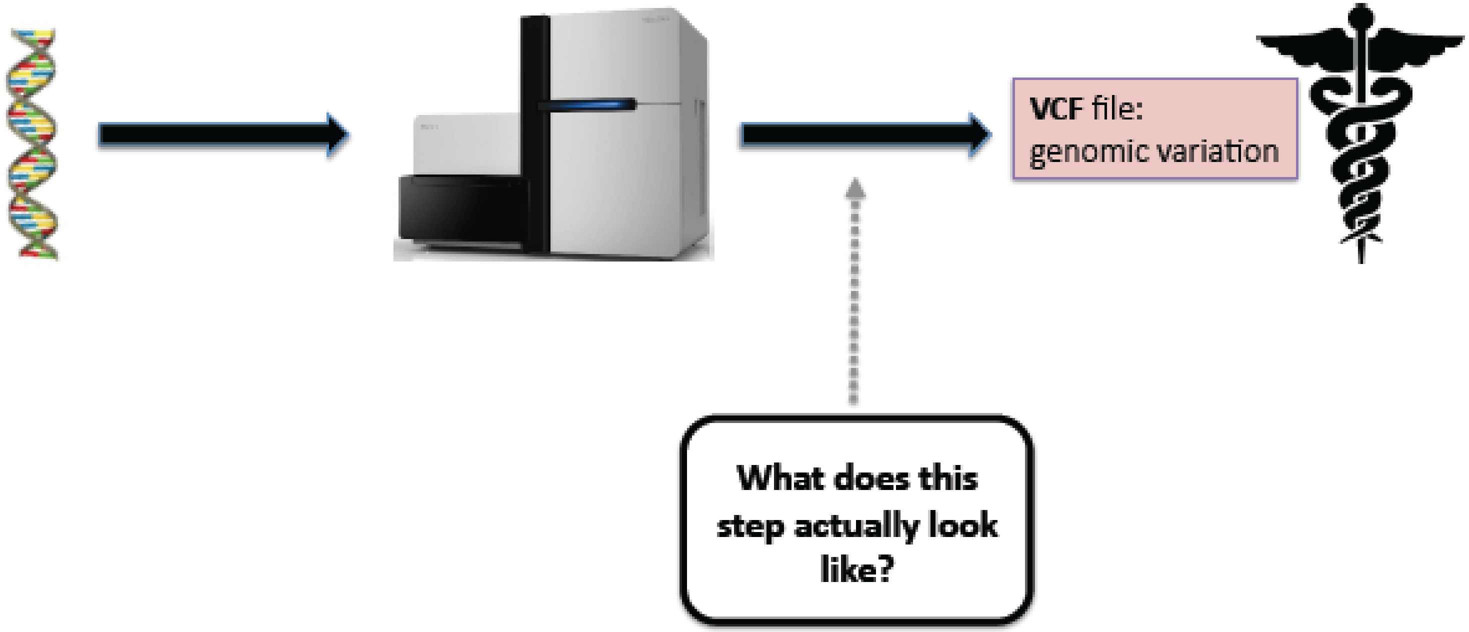


```
##fileformat=VCFv4.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO
    FORMAT      NA00001        NA00002        NA00003


20     14370   rs6054257 G      A       29   PASS   DP=14;AF=0.5;DB
    GT:GQ:DP 0|0:48:1 1|0:48:8 1/1:43:5
20     1110696 rs6040355 A      G,T     67   PASS   DP=10;AF=0.333,0.667;DB
    GT:GQ:DP 1|2:21:6 2|1:2:0   2/2:35:4
20     1230237 .         T      .       47   PASS   DP=13
    GT:GQ:DP 0|0:54:7 0|0:48:4 0/0:61:2
20      1234567 microsat1 GTCT   G,GTACT 50   PASS   DP=9
    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Official specification in
www.1000genomes.org/wiki/Analysis/Variant Call Format/vcf-variant-call-format-version-41

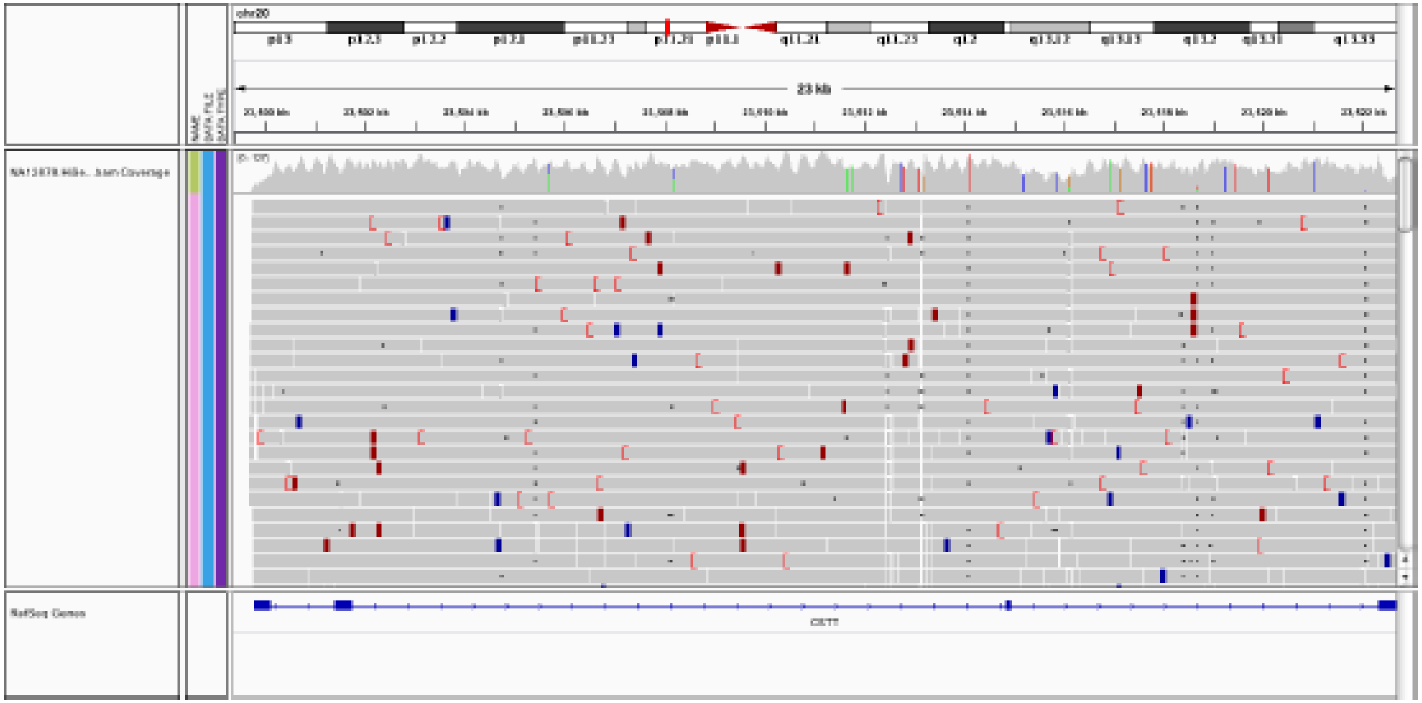
Is processing/analysis of NGS data really that easy?



VCF file:
genomic variation

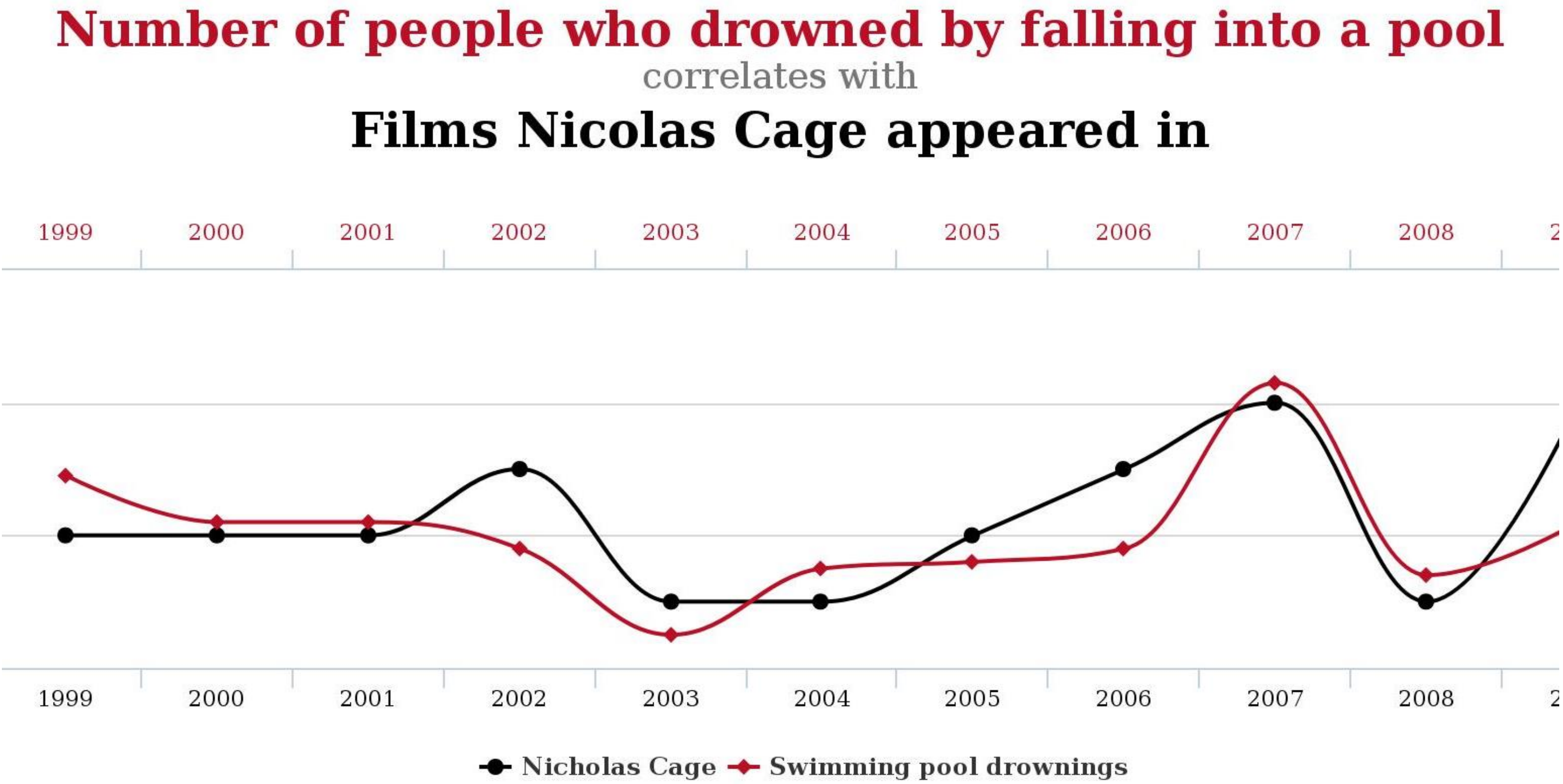
What does this step actually look like?

It’s going to involve dealing with messy situations like this:



How can we tell which mismatches represent real mutations and which are just noise?

If you look at a lot of data, you’re bound to find something



http://www.tylervigen.com/spurious-correlations