Whole-Genome Sequencing (WGS)

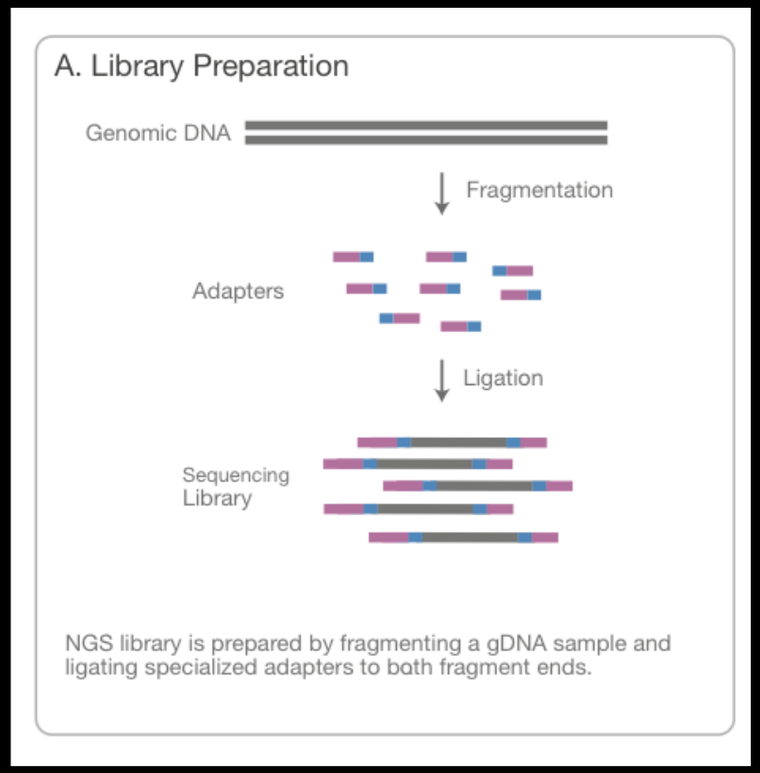
Aakrosh Ratan

PHS5500: Special Topics in Public Health - Public Health Genomics

14th March, 2016

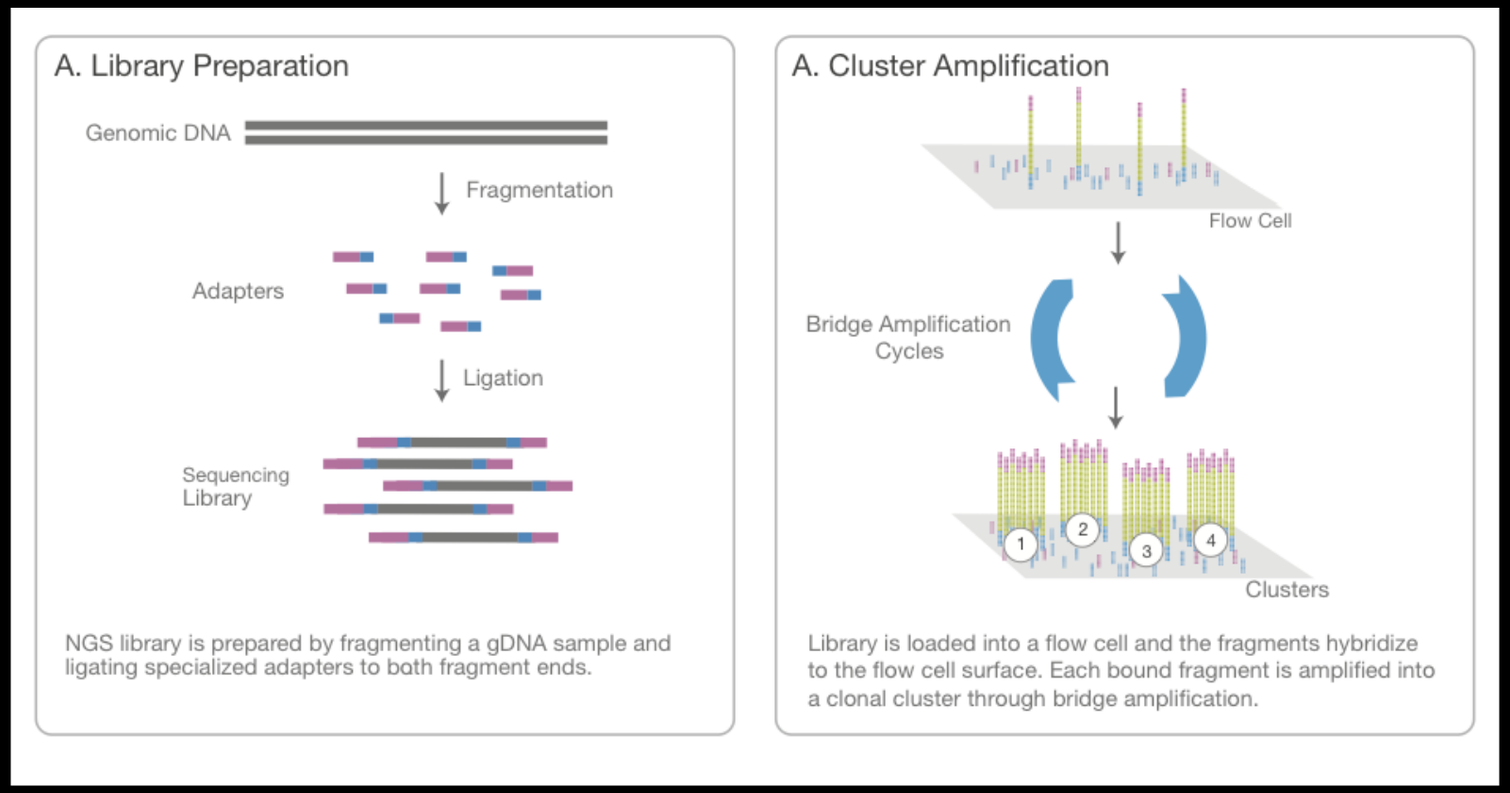
http://bims.virginia.edu/faculty/aakrosh-ratan

ratan@virginia.edu

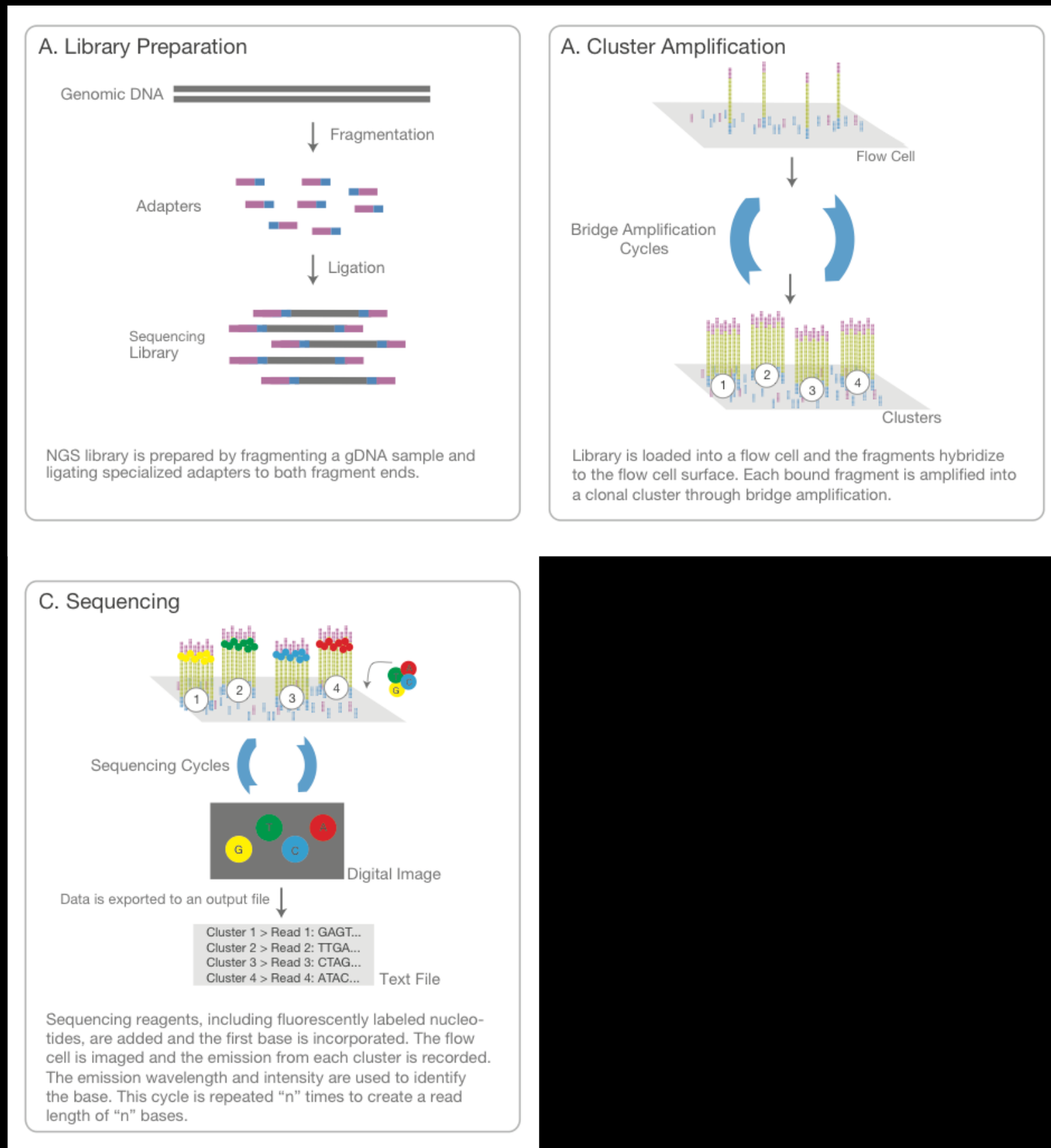


Source :http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

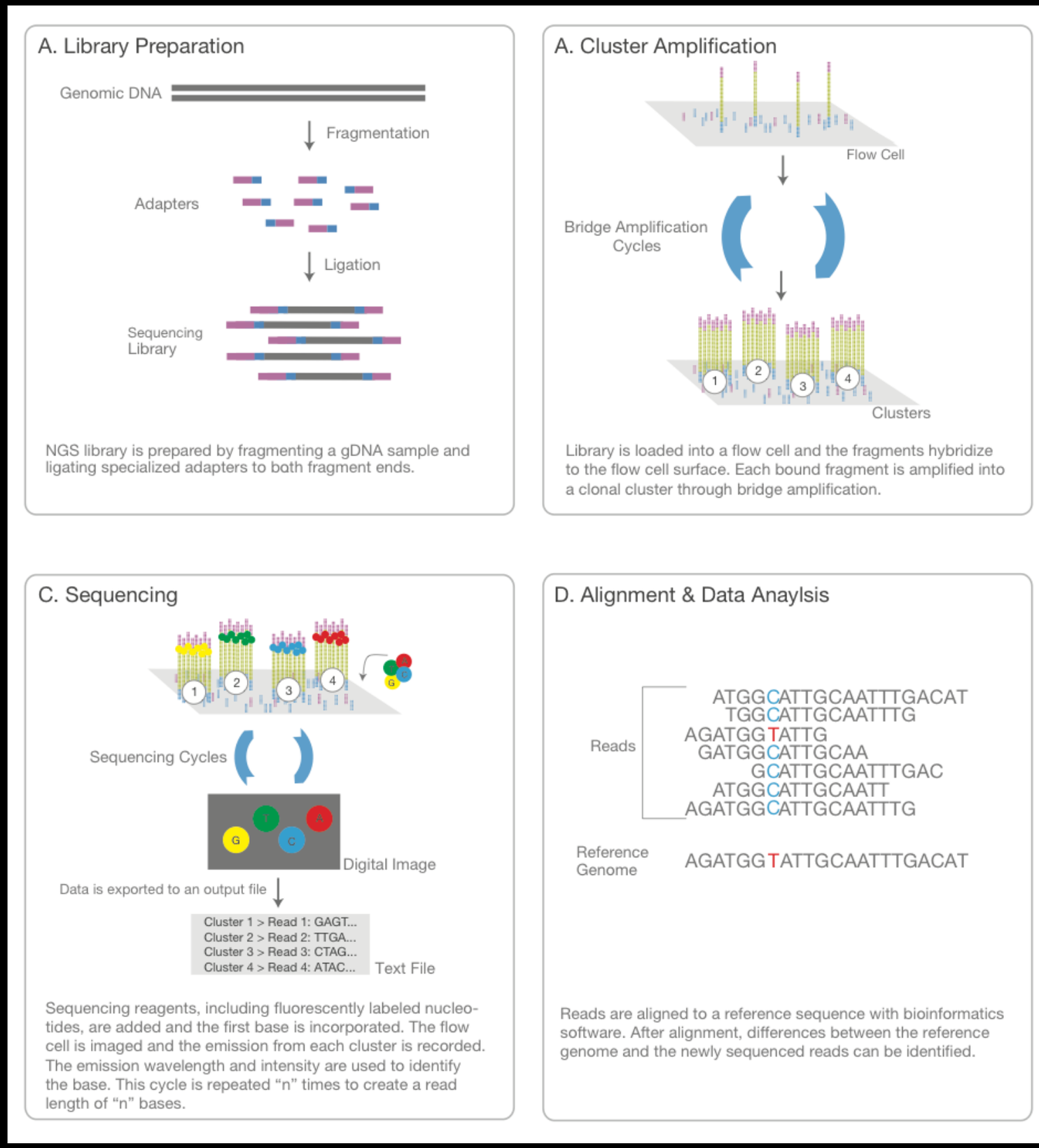
This should be called panel B.



Source :http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf



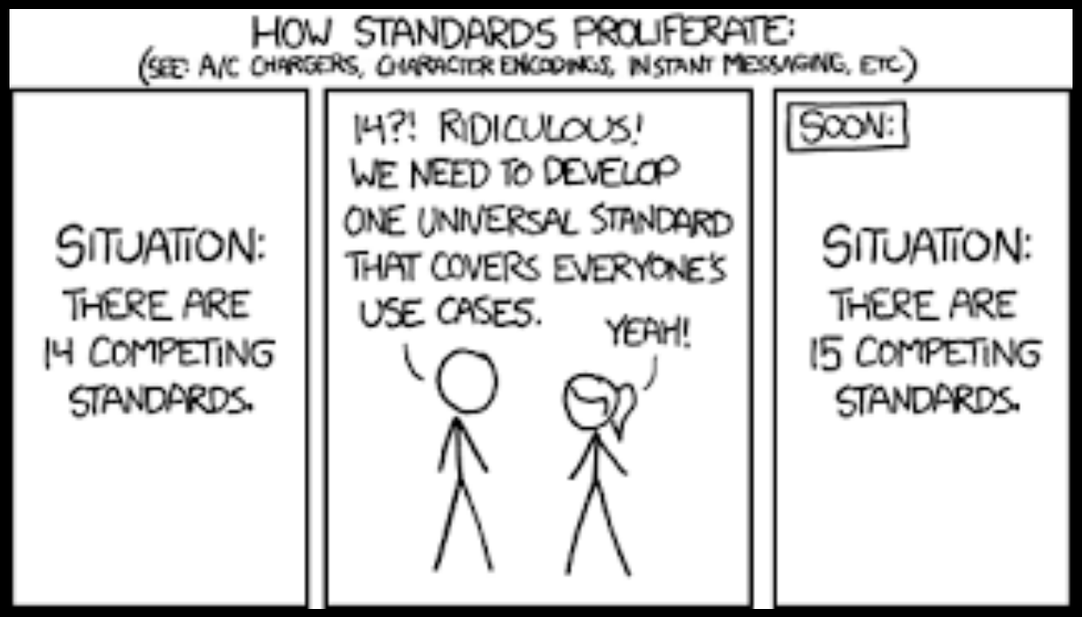
Source :http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf



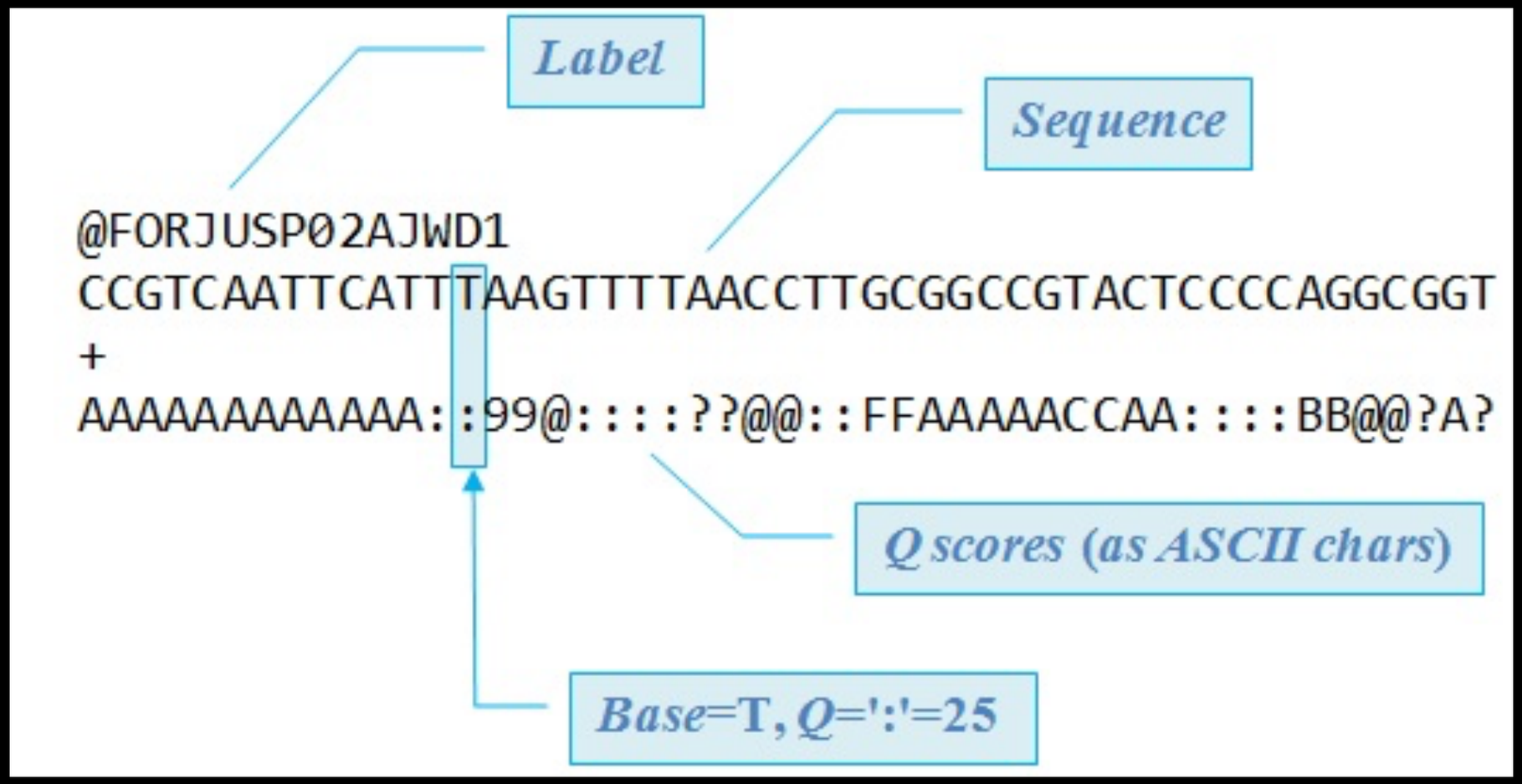
Source :http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

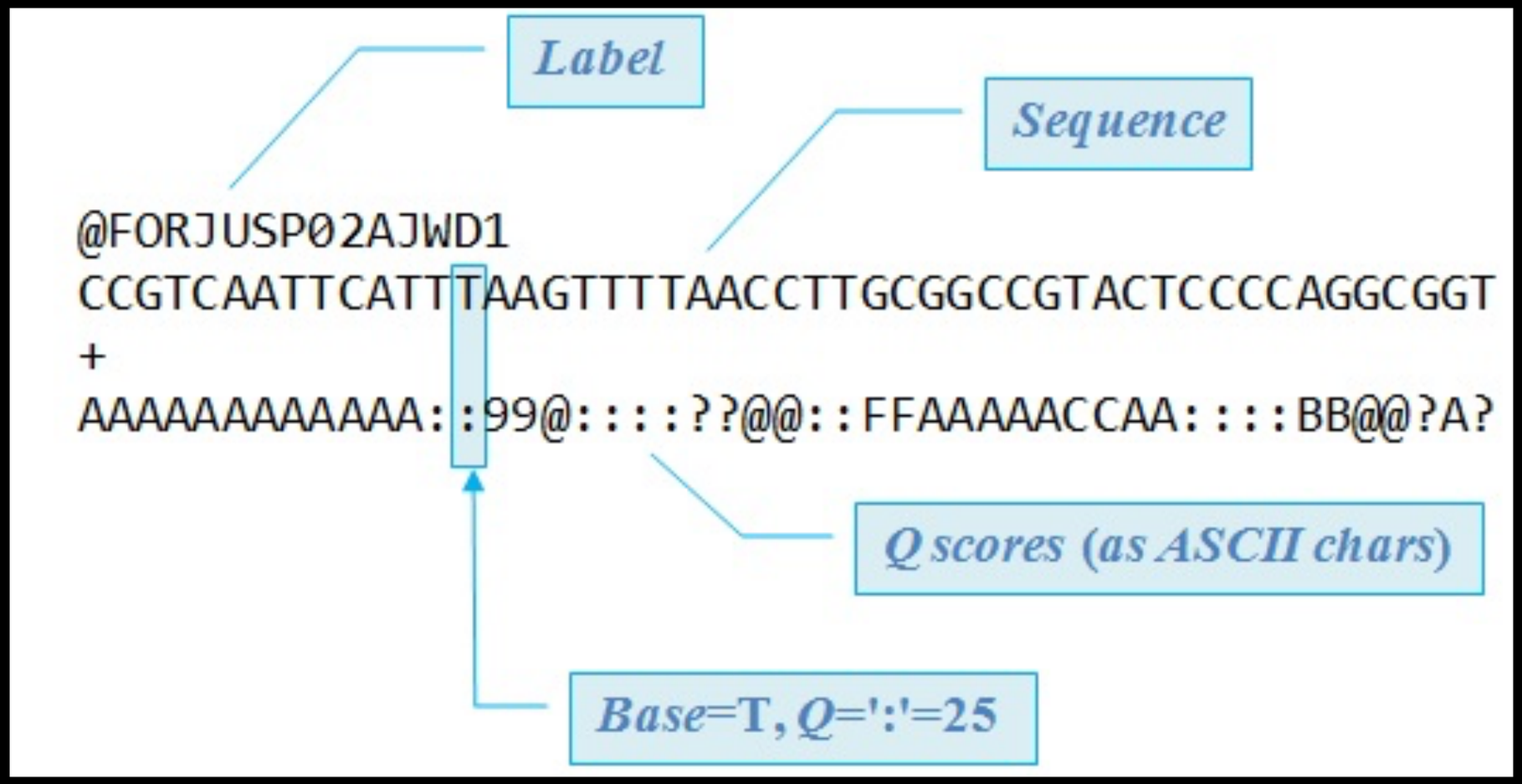
Paired-End Sequencing



Source: https://xkcd.com/927/



Source: http://drive5.com/usearch/manual/fastq_files.html



Source: http://drive5.com/usearch/manual/fastq_files.html

$$A = 1 - E \quad = 1 - 10^{-\left(\frac{Q}{10}\right)}$$

phred Q25 ~ 0.9968

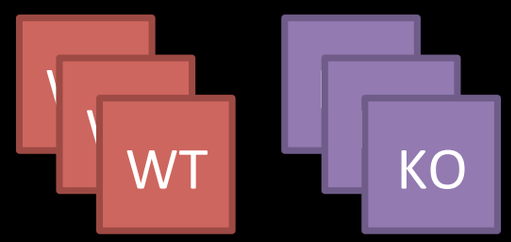
```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a``^\__`_```^a``a`^a_^__]a_]\]`a______`_^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaababbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^a`]a_abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_]_[^_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACA
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\^\\`aa]ba__bba[a_O_a`aa`aa`a]^V]X_a^YS\R_\H_[]\ZTDUZZUSOPX]]POP\GS\WSHHD
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAATACATCTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
```

**1X coverage of the human genome**

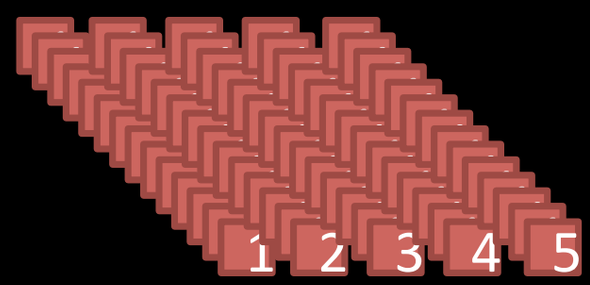
Assume read length of 100 bases

Assume read names take ~ 50 character

~7.5 GB
```



Small experiment; we can examine things in detail



Large experiment; we can't truly look at all details

Quality Control

QC should tell you what to look at more closely. It should NOT be used as an automated filter.

(Baby) Steps

- Decide what is “normal”
- Calculate the same metric in your datasets
- Check from deviance from normal
- Trigger an alarm (visual alarm, email…) to notify user to look at the data more closely
- Summarize, Visualize and Flag

Raw sequence:

FastQC

PRINSEQ

FastQ screen

Mapped sequence:

BamQC

QualiMap

Bismark (specialised)

Application specific:

- **RNA-Seq**
  SeqMonk RNA-Seq QC
  RNASeQC
- **Small RNA**
  SeqMonk QC
- **ChIP**
  ChIPQC package
- **Bi-Sulphite Sequencing**
  Bismark
- **Hi-C**
  HiCUP

Data Visualisation:

SeqMonk

Intergrated Genome Viewer (IGV)

Compile many QC analyses into a single report:

MultiQC

Few Existing Tools for QC

Raw sequence:

FastQC

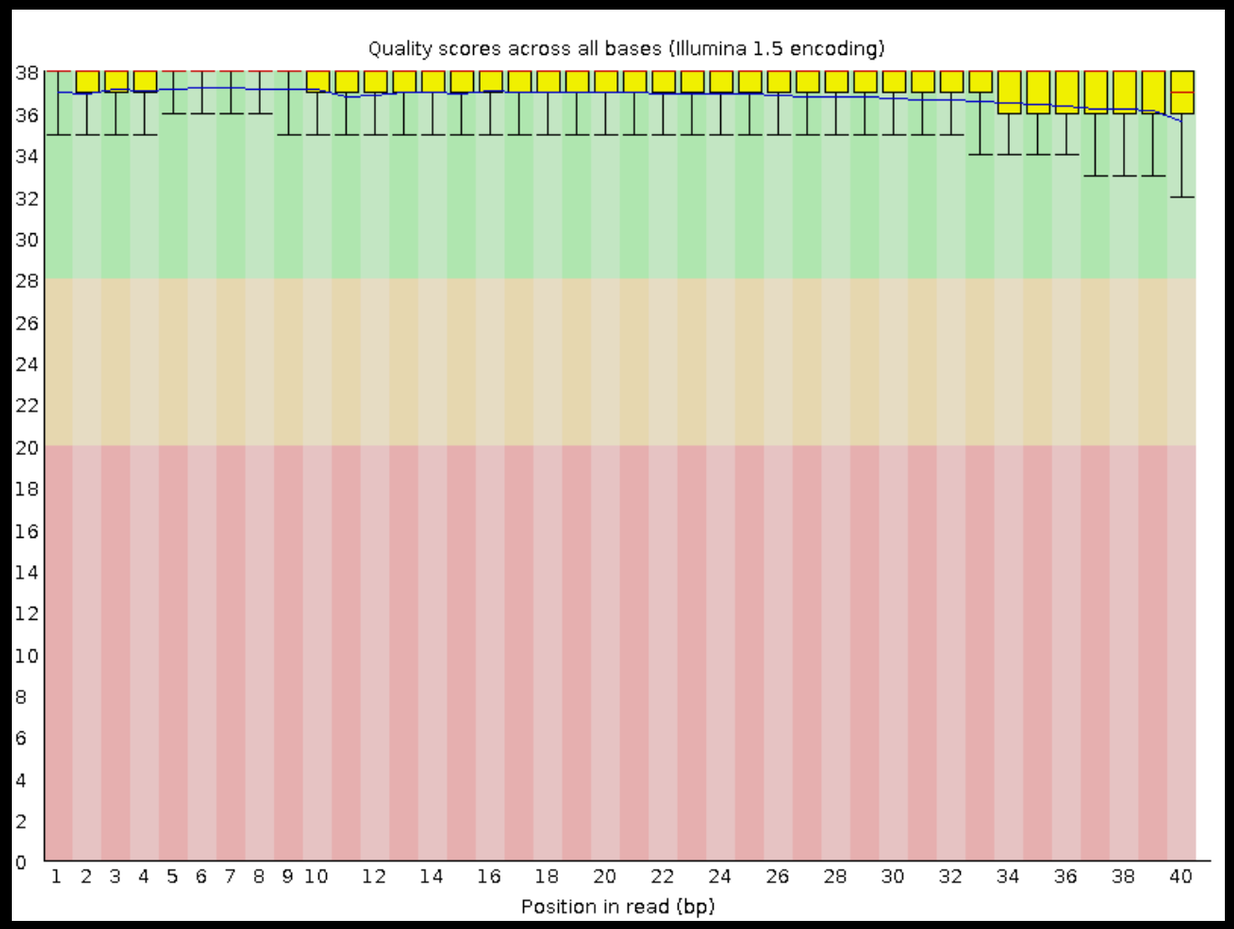
PRINSEQ

FastQ screen

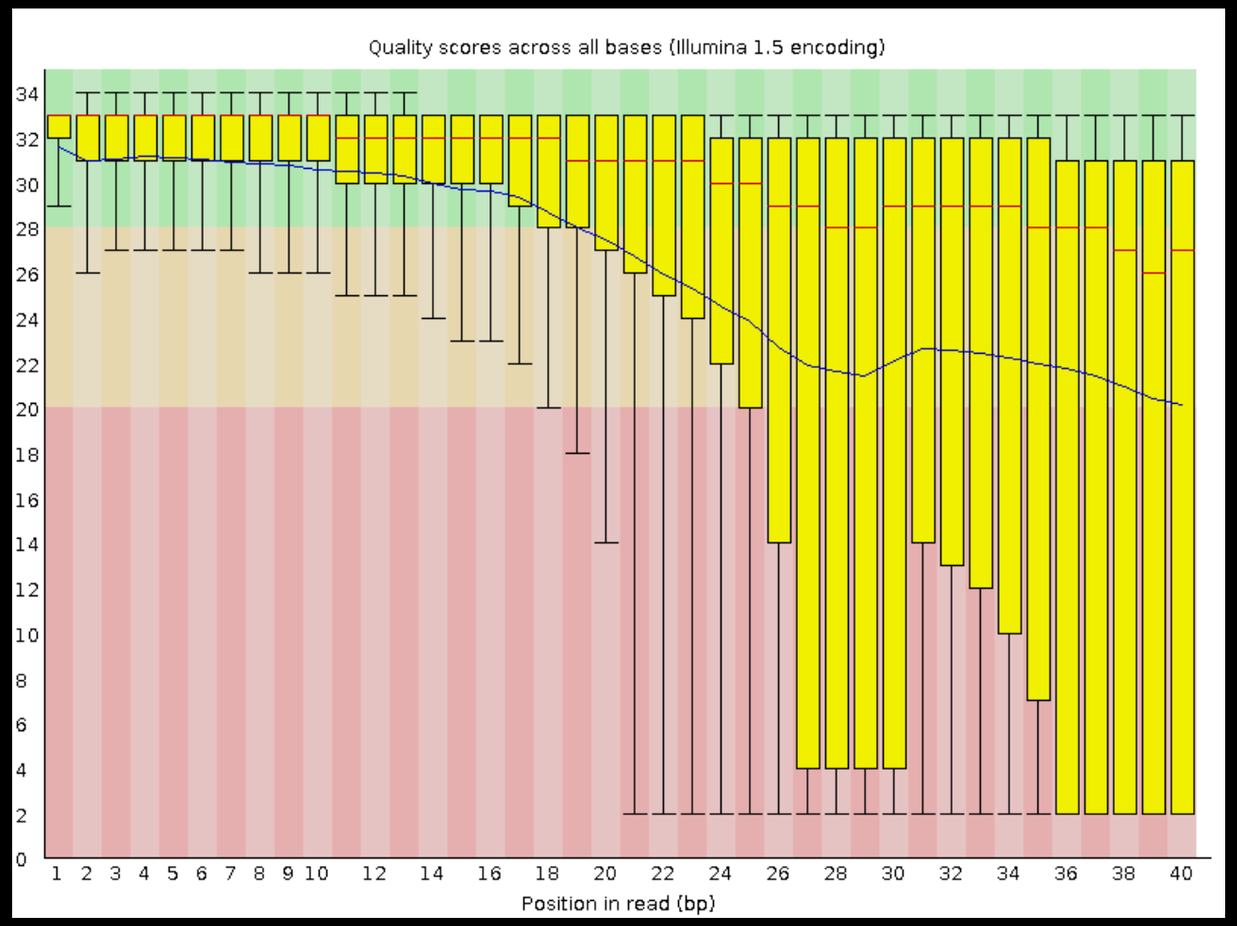
Few Existing Tools for QC

Per base sequence quality

Good



Bad

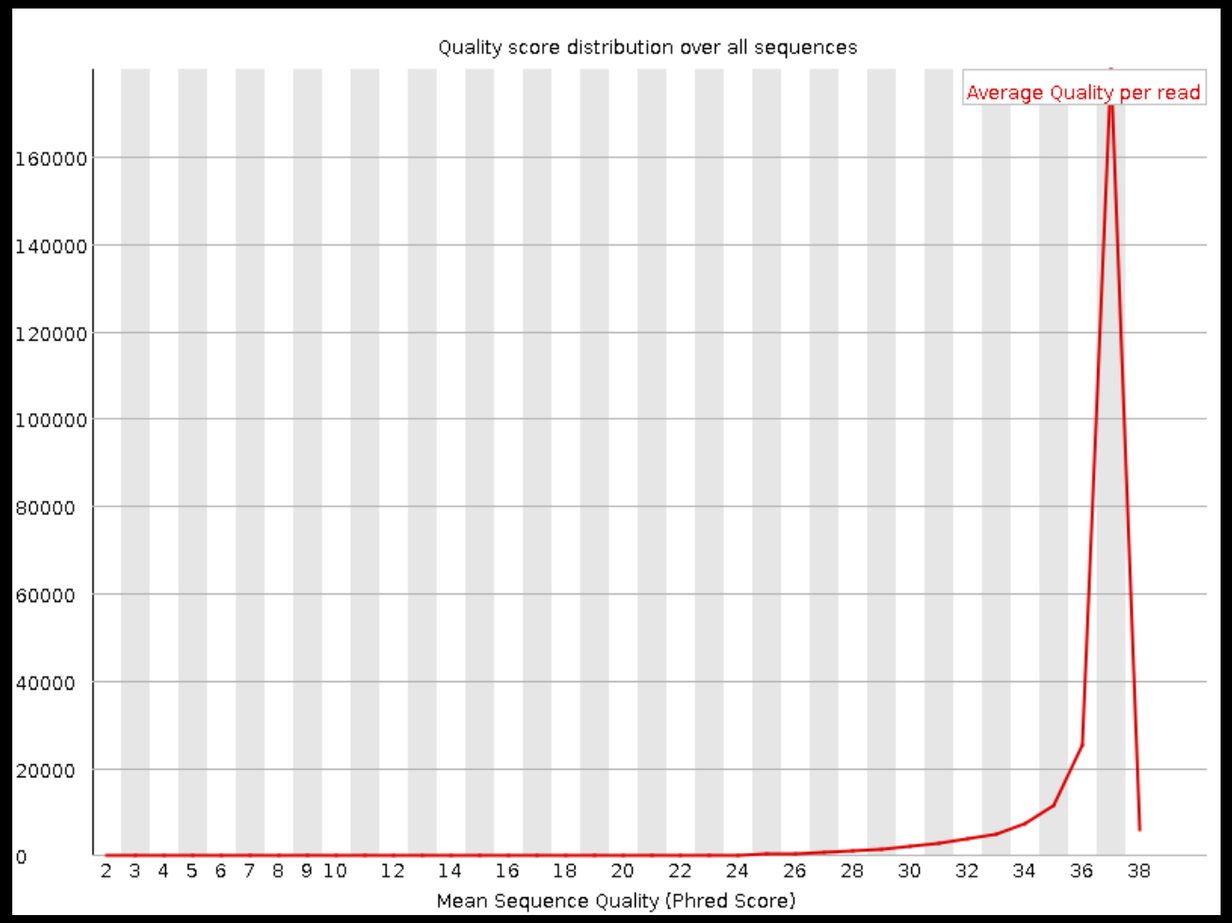


FASTQC

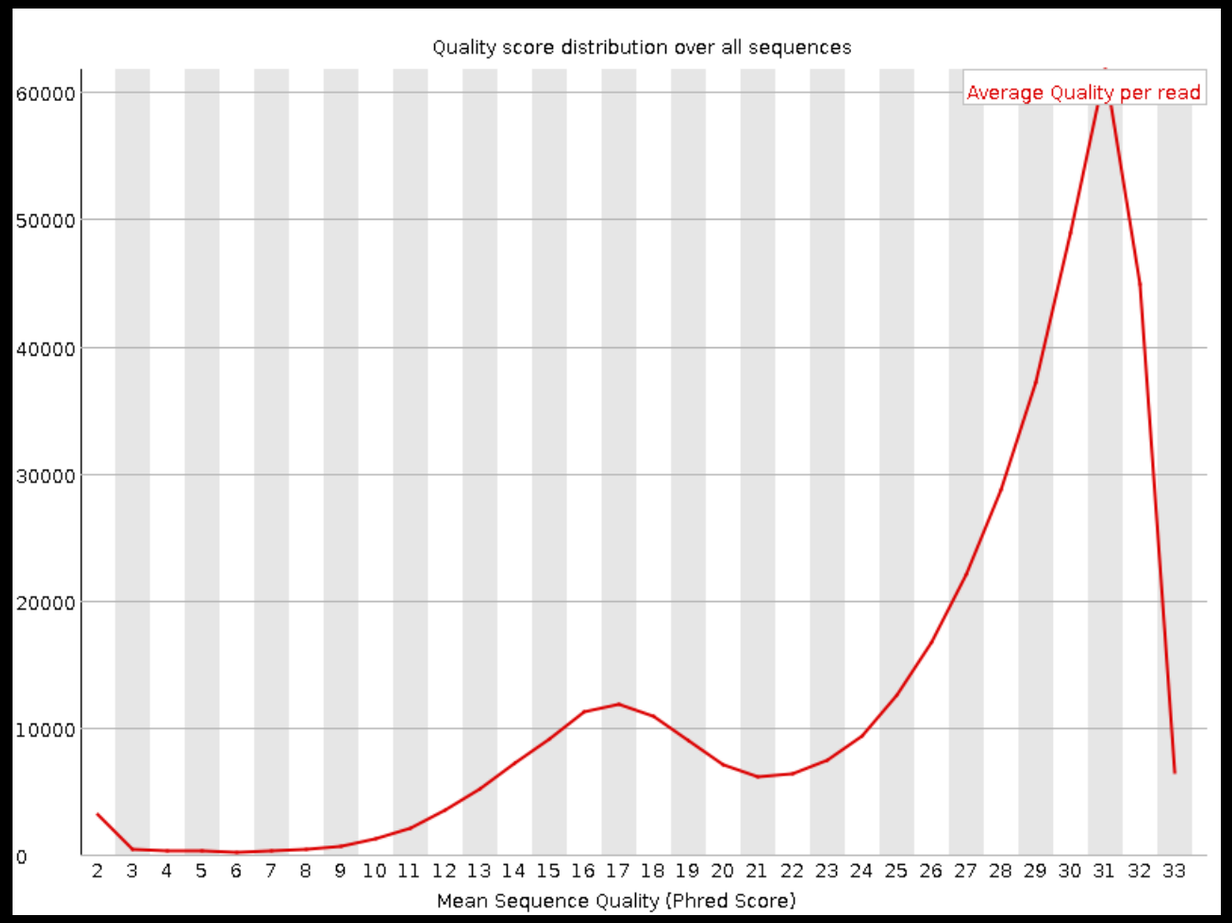
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Per sequence quality scores

Good



Bad

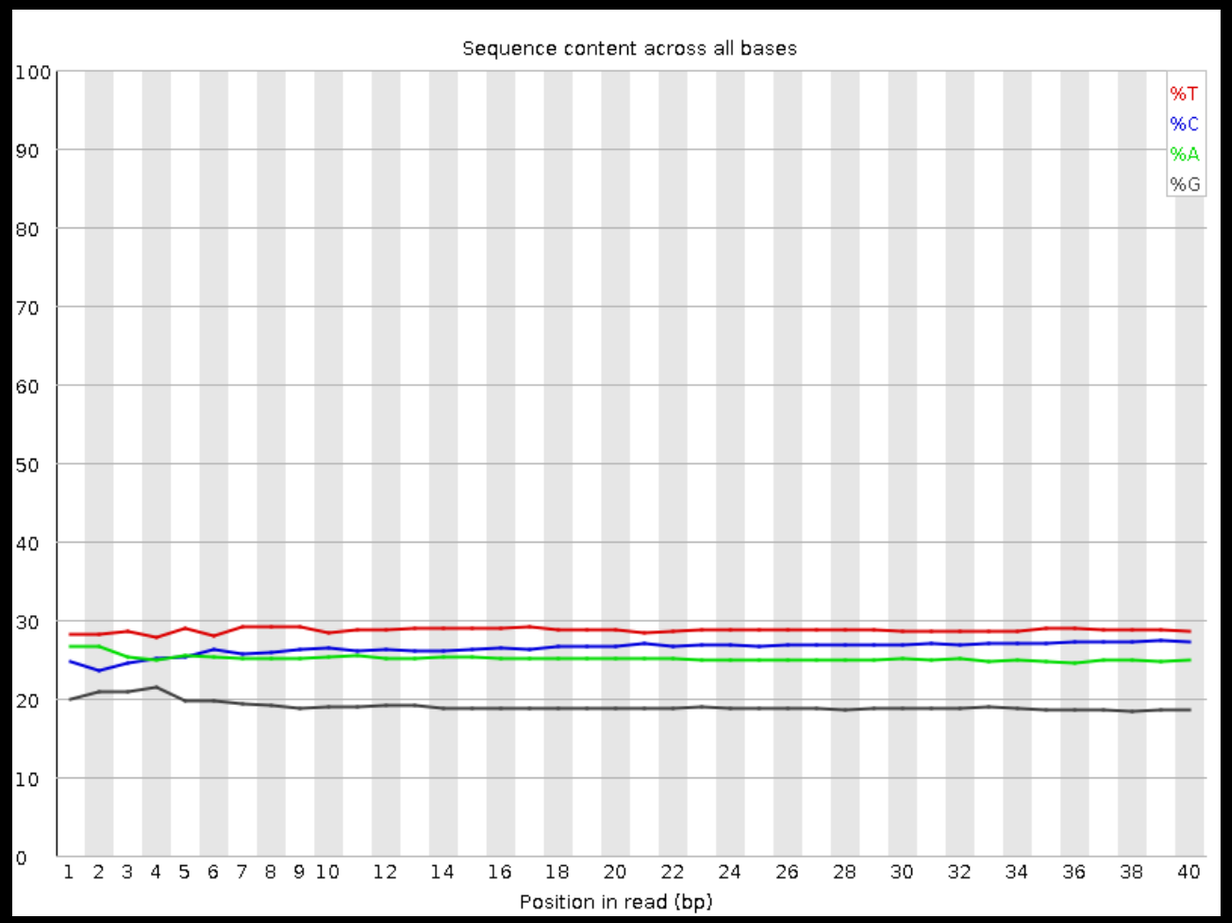


FASTQC

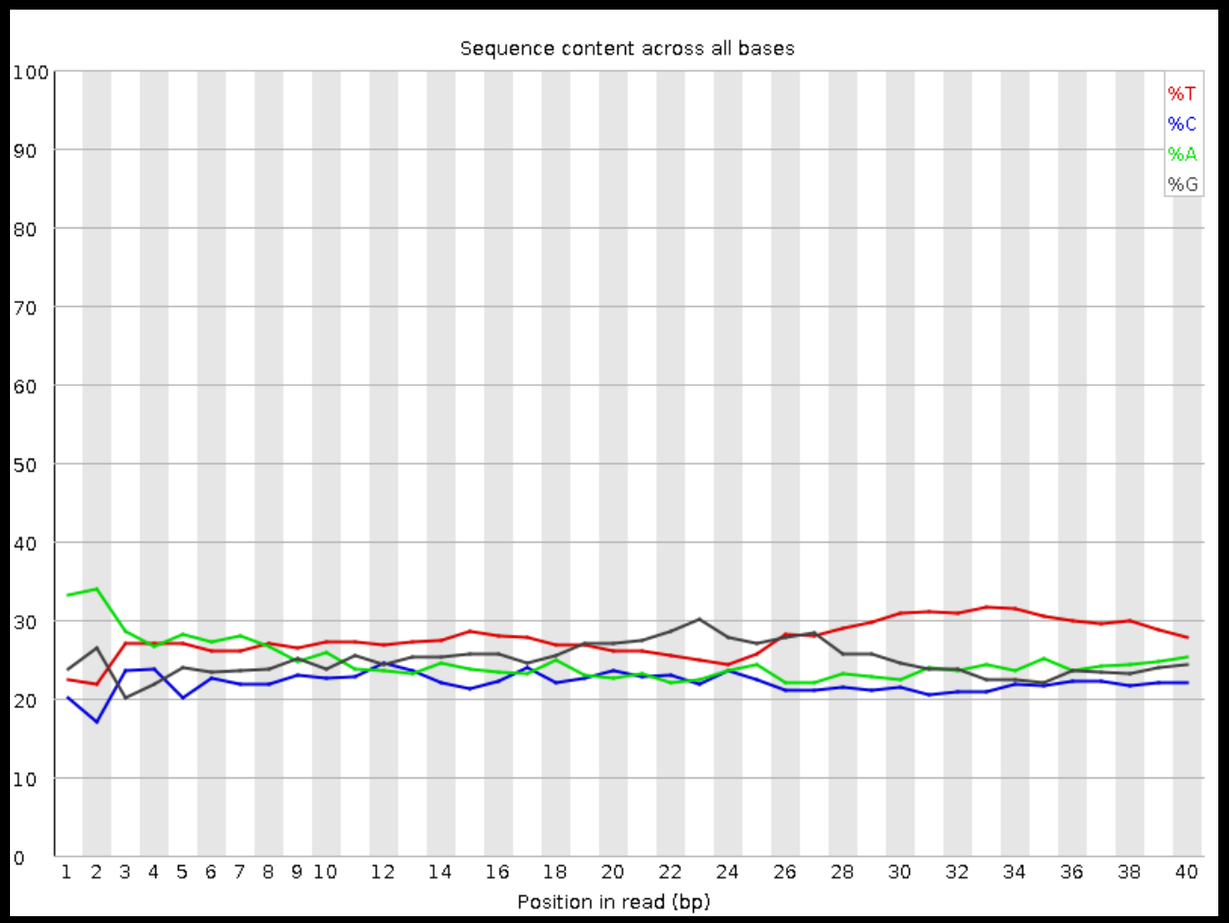
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Per base sequence quality

Good



Bad

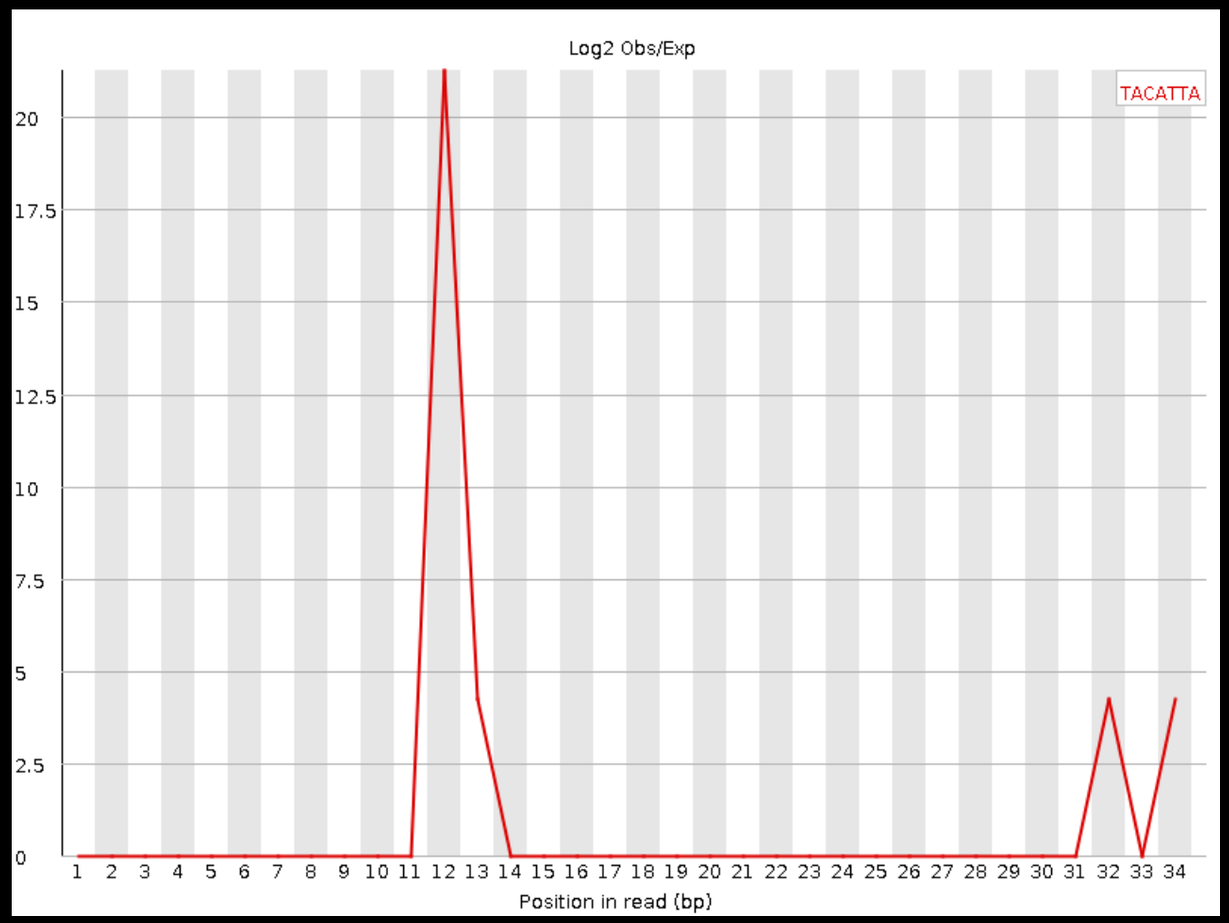


FASTQC

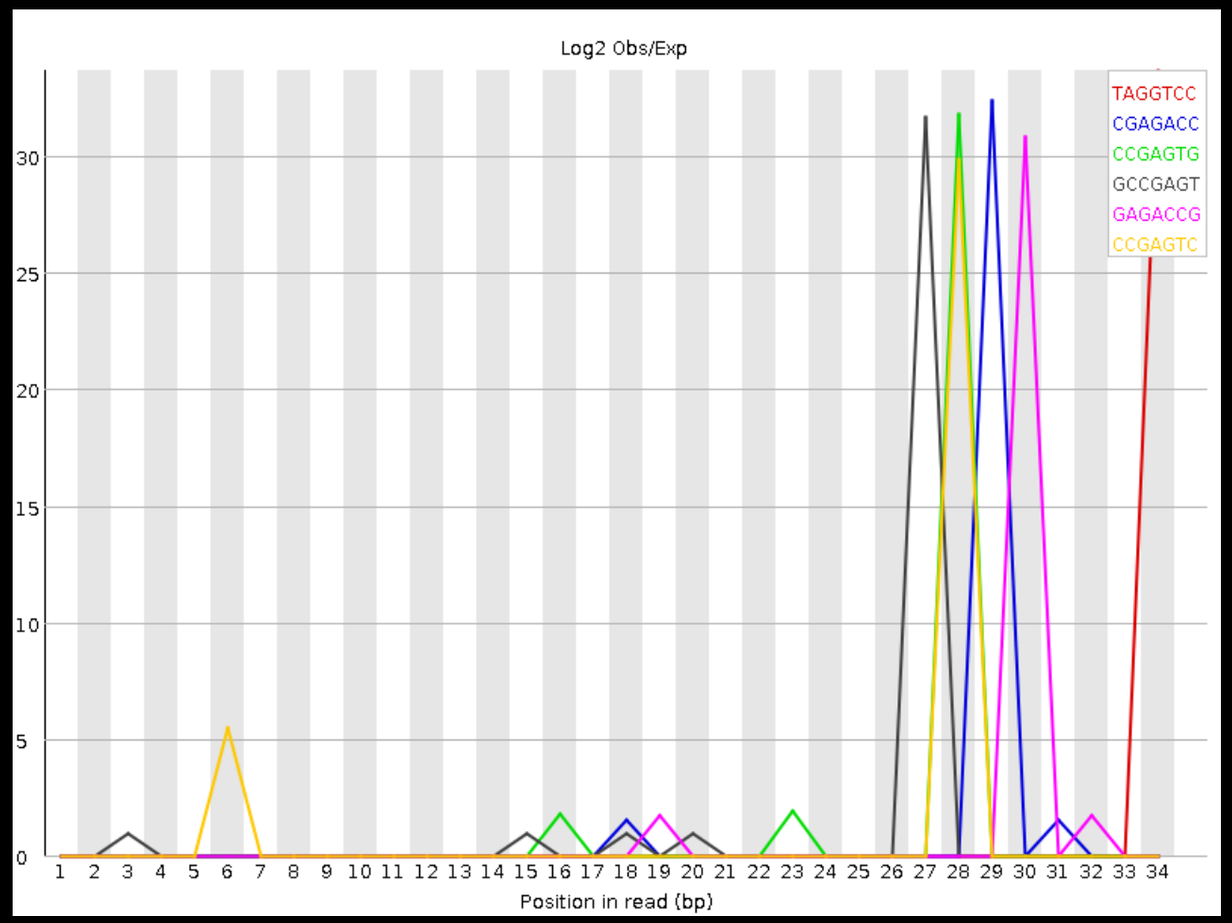
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

K-mer Content

Good (Not that good!!!)



Bad



FASTQC

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Several other metrics can be checked (based on application)

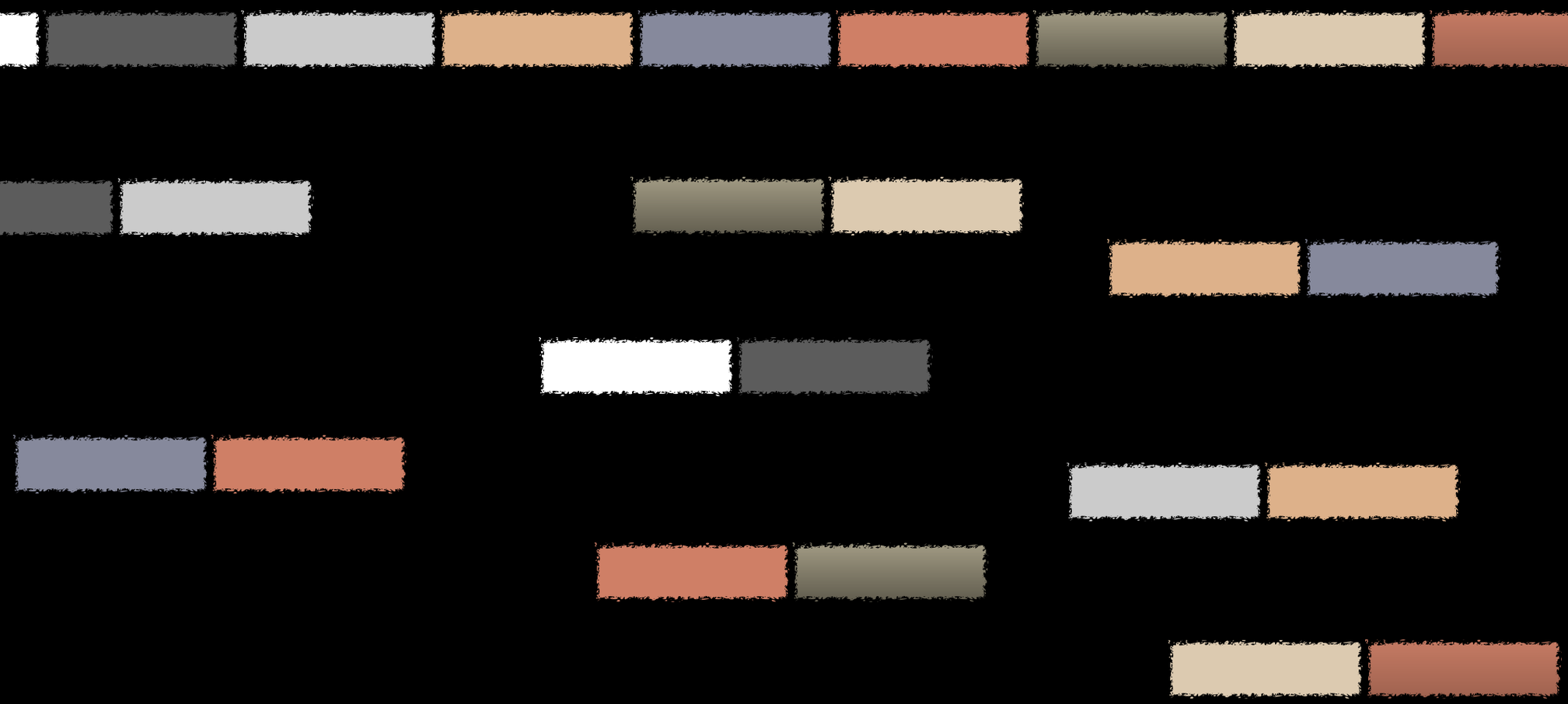
- Per tile sequence quality (if some specific region on the sequencer was enriched for bad sequences)
- Per base N content (if some specific base position was enriched for ambiguous base-calls)
- Sequence length distribution (Adapter contamination, could also signal degraded DNA)

Several other metrics can be checked (based on application)

- Per tile sequence quality (if some specific region on the sequencer was enriched for bad sequences)
- Per base N content (if some specific base position was enriched for ambiguous base-calls)
- Sequence length distribution (Adapter contamination, could also signal degraded DNA)

And of course, there is additional QC after every step in the pipeline

So let us sequence Mr. H.

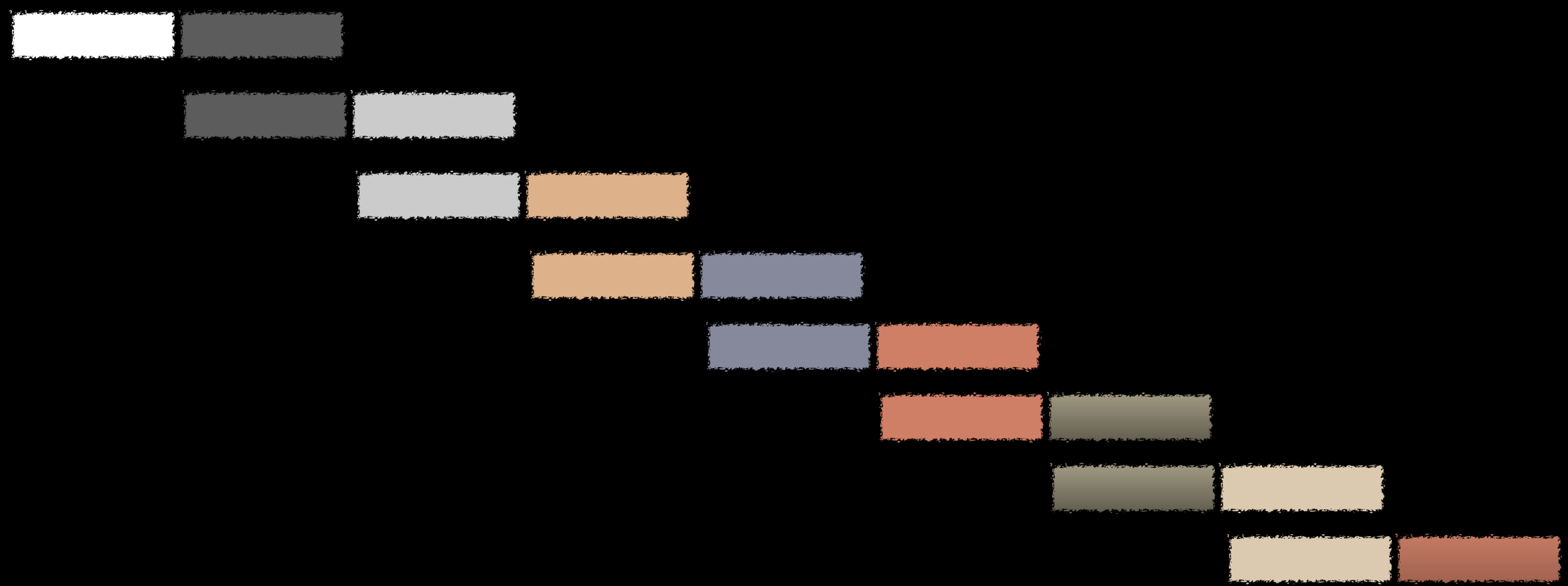


Using reads that are 2 bps long.

We actually do not know his genome sequence!!!

De Novo Assembly

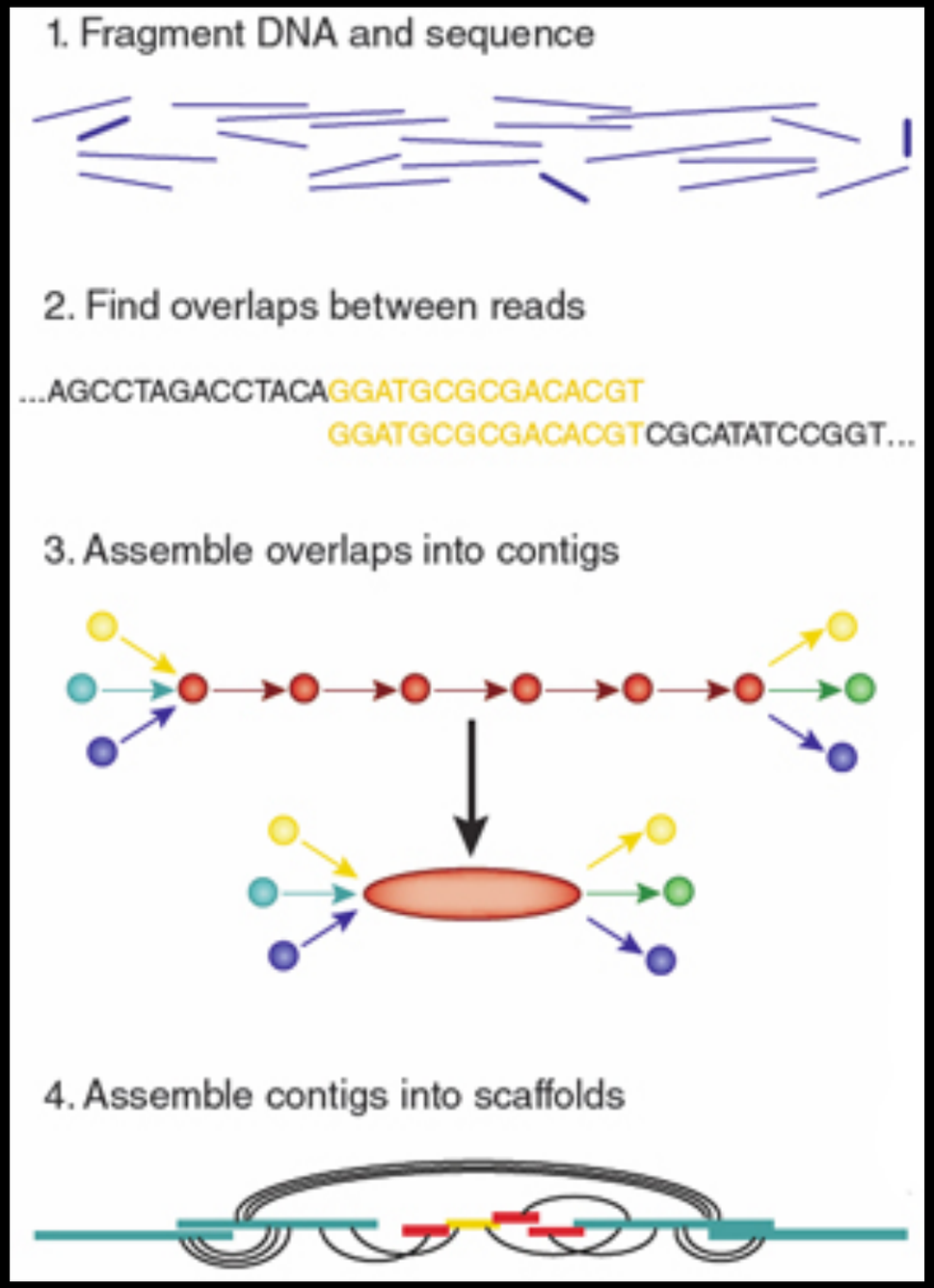
Overlapping Fragments



De Novo Assembly

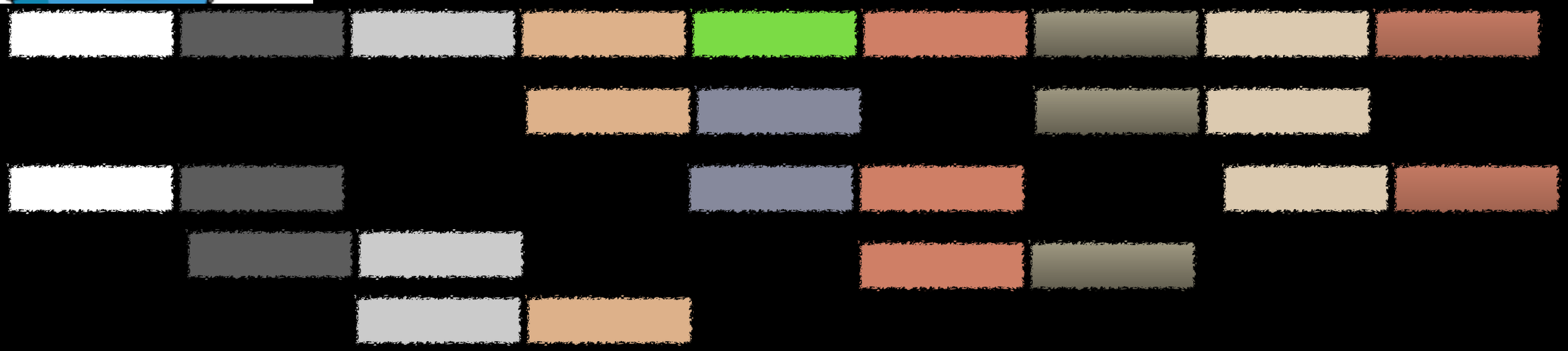
De Novo Assembly

- Repeats complicate assemblies.
- Typically require large amounts of memory for mammalian sized genomes
- Several approaches:
  - Overlap graphs
  - De Bruijn graphs
- Some de novo assemblers for short-reads
  - Velvet, ABySS, Forge, SOAPdenovo…

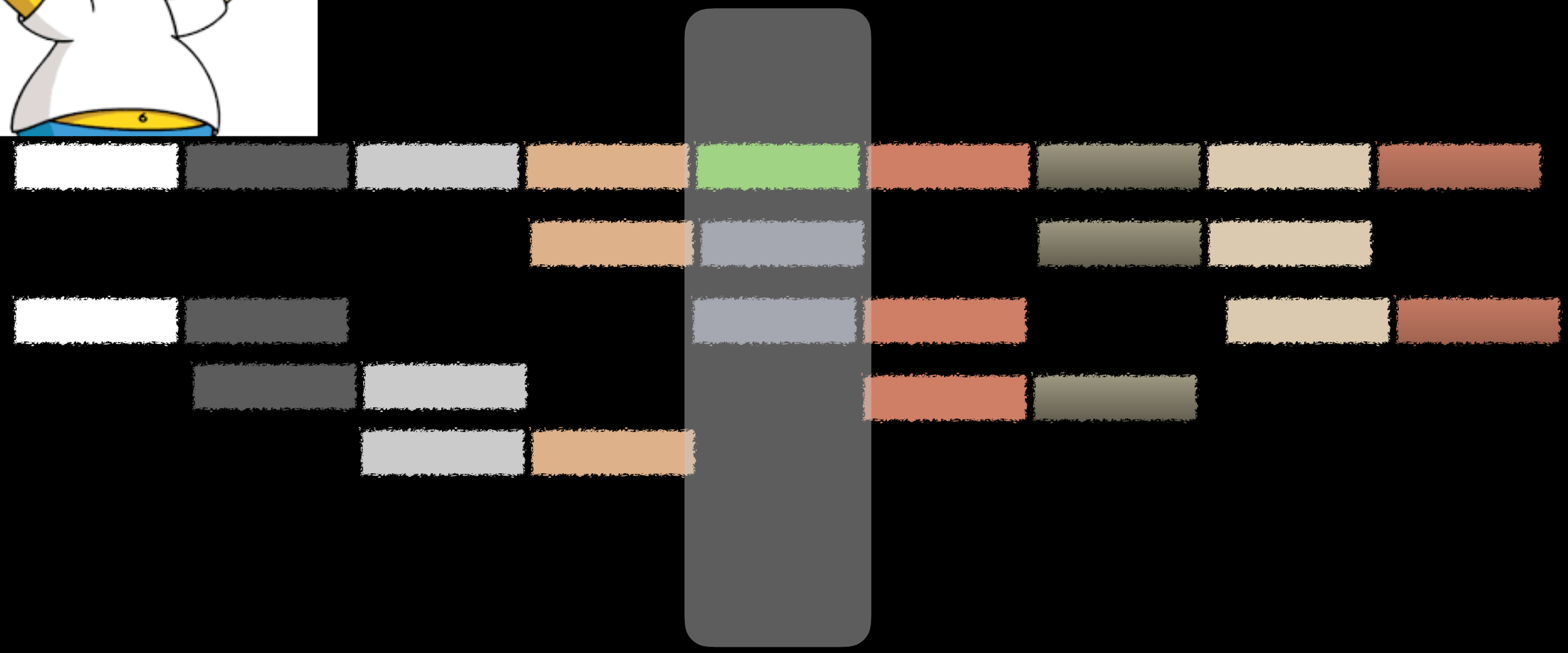


But wait !!! We have a reference genome from Mr. T.

So now we need to find the best placement for each sequence from Mr. H.



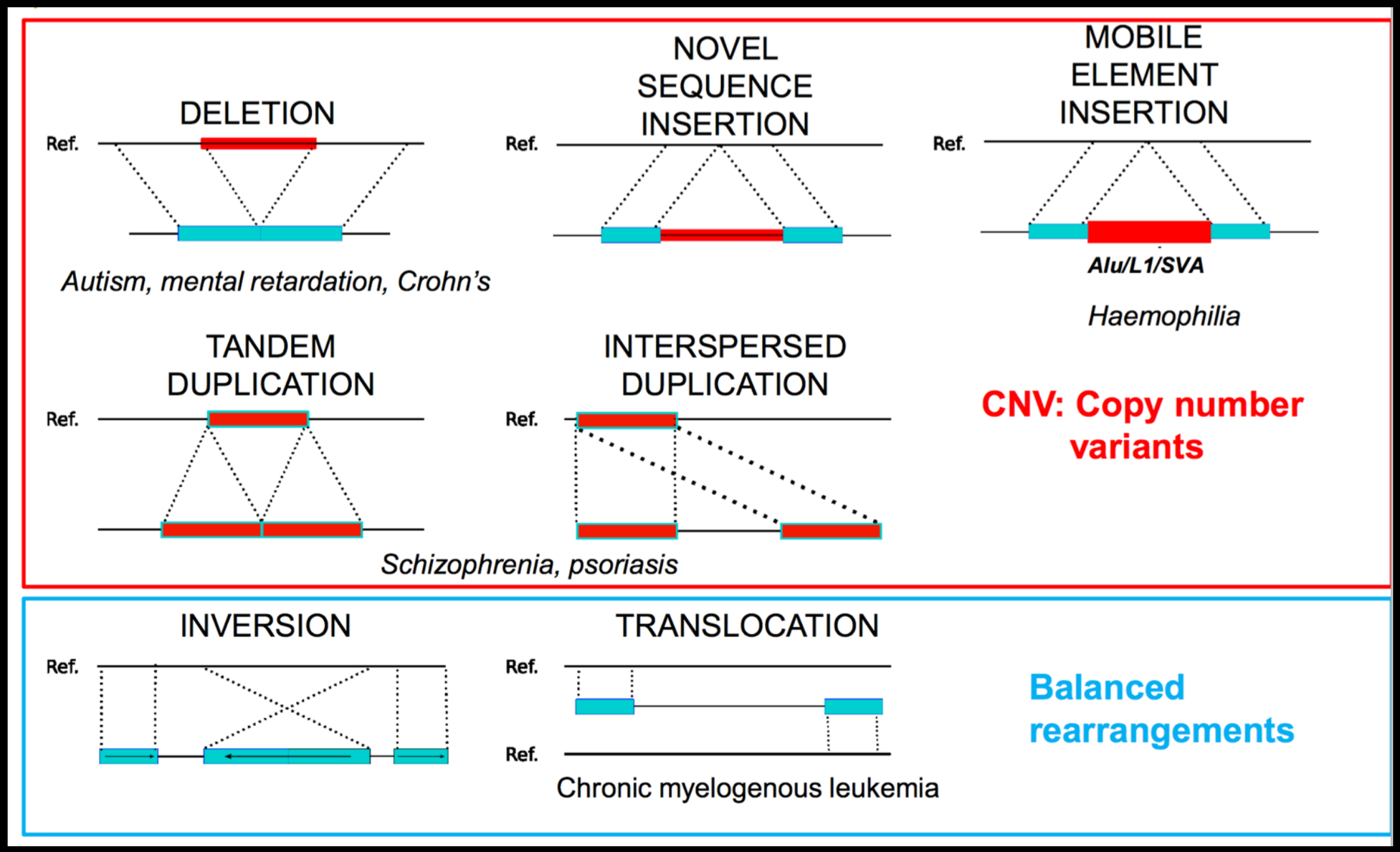
So now we need to find the best placement for each sequence from Mr. H.



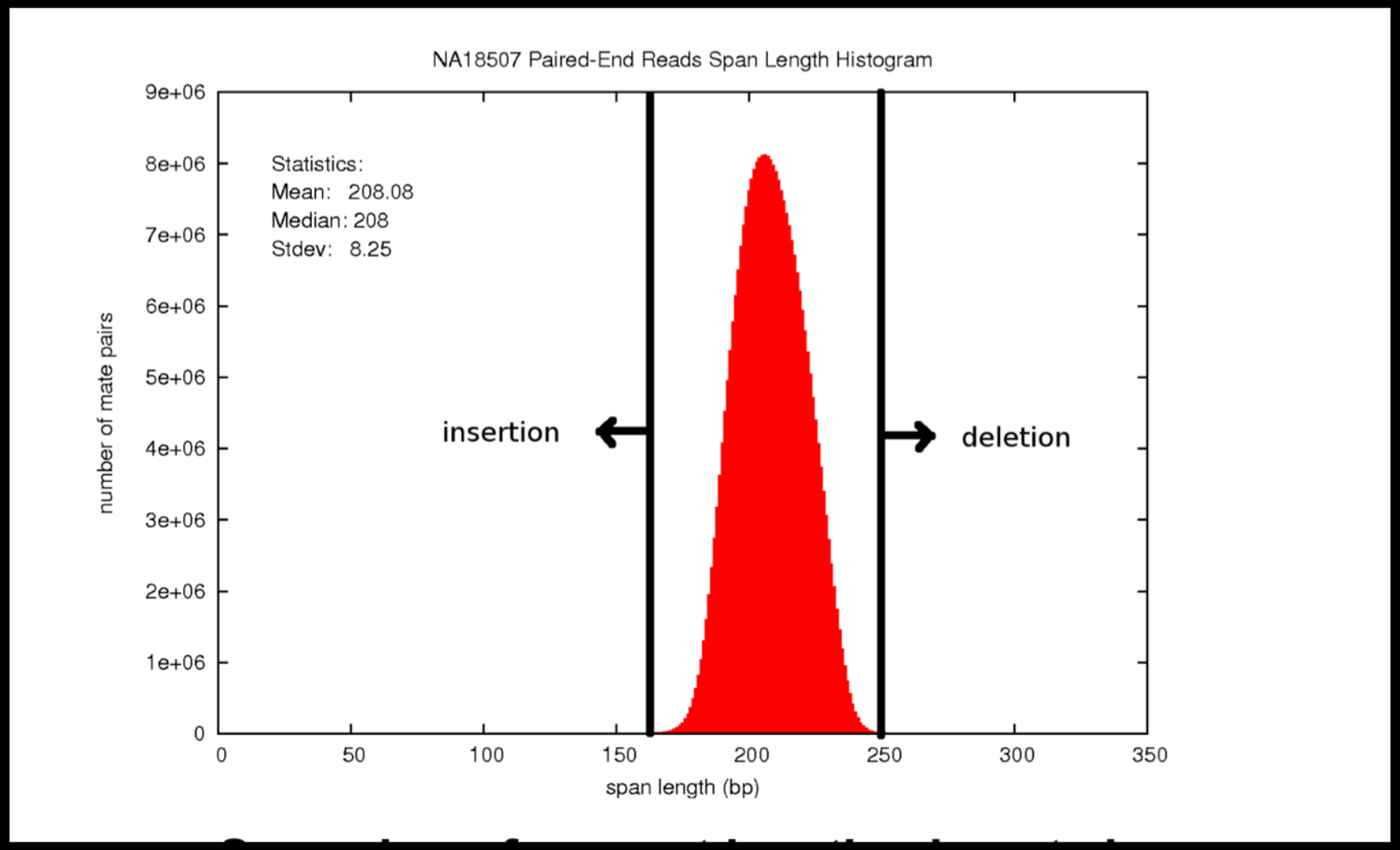
So now we need to find the best placement for each sequence from Mr. H.

Aligning to a reference

- Typically faster and requires less resources
- SNPs and other variations are more easily placed and identified
- Large fraction of sequence that does not align is either really divergent or not present in the reference
- Several approaches
  - Seed and extend
  - BWT …
- Some alignment tools:
  - BWA, Bowtie, LASTZ, LAST, BLAST …



Classes of other variants

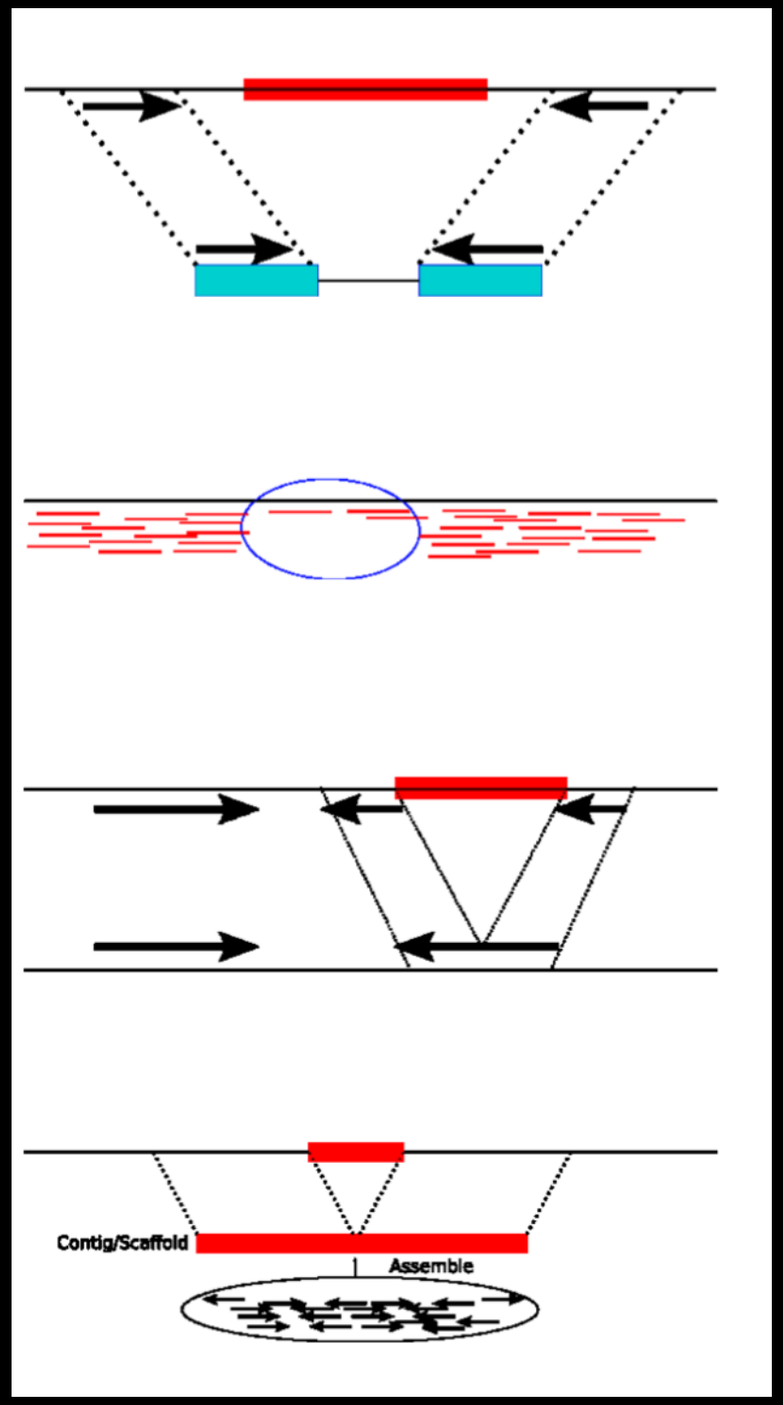


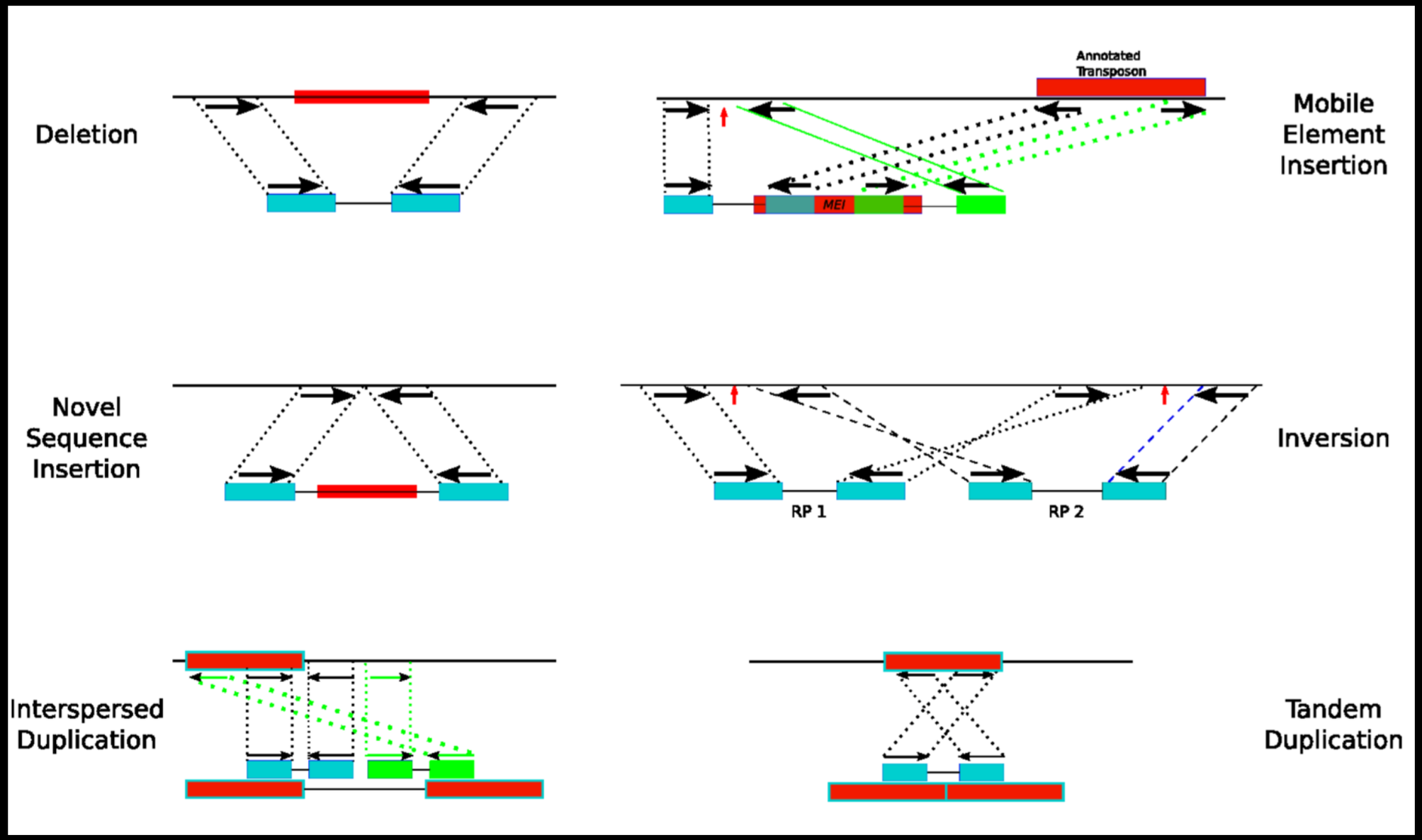
Concordant = read pairs that map in expected orientation & size
Discordant = read pairs that map different than what is expected

Insert Length Distribution

Sequence signatures of structural variation

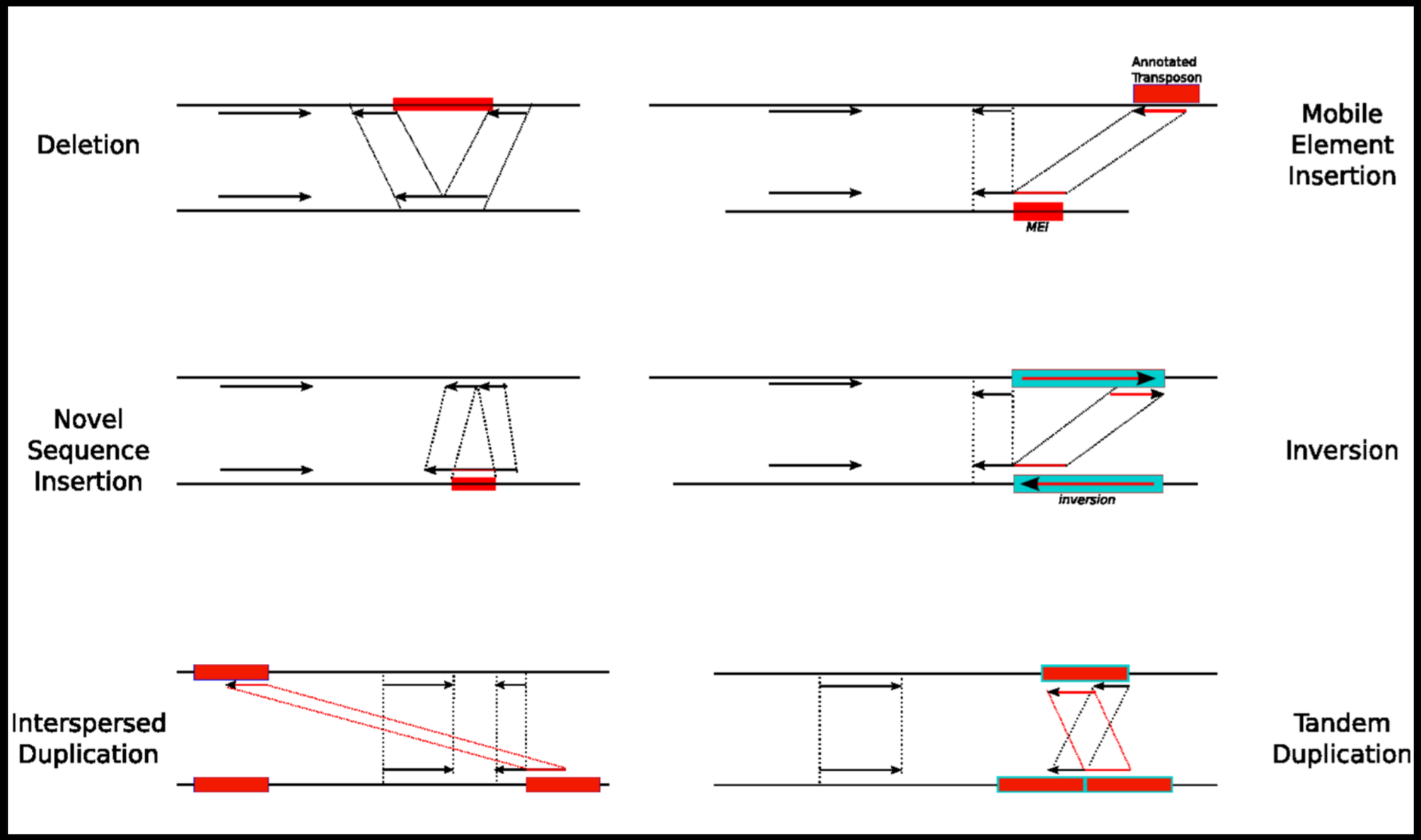
- Read pair analysis
  - Deletions,small novel insertions,inversions, transposons
  - Size and breakpoint resolution dependent to insert size
- Read depth analysis
  - Deletions and duplications only
  - Relatively poor breakpoint resolution
- Split read analysis
  - Small novel insertions/deletions, and mobile element insertions
  - 1bp breakpoint resolution
- Local and de novo assembly
  - SV in unique segments
  - 1bp breakpoint resolution





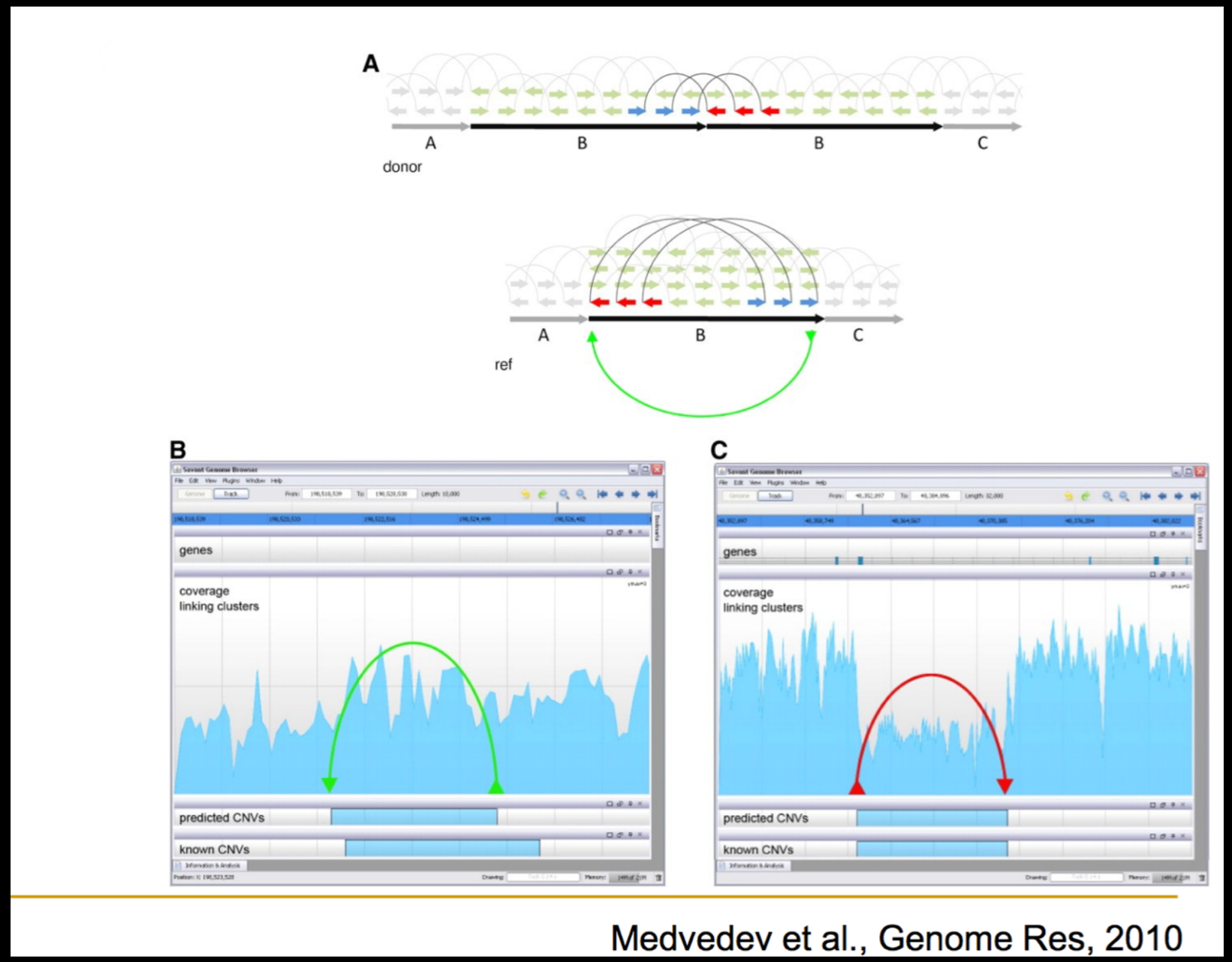
Patterns of SVs from Paired-End reads

BreakDancer, GenomeSTRiP, VariationHunter, HYDRA



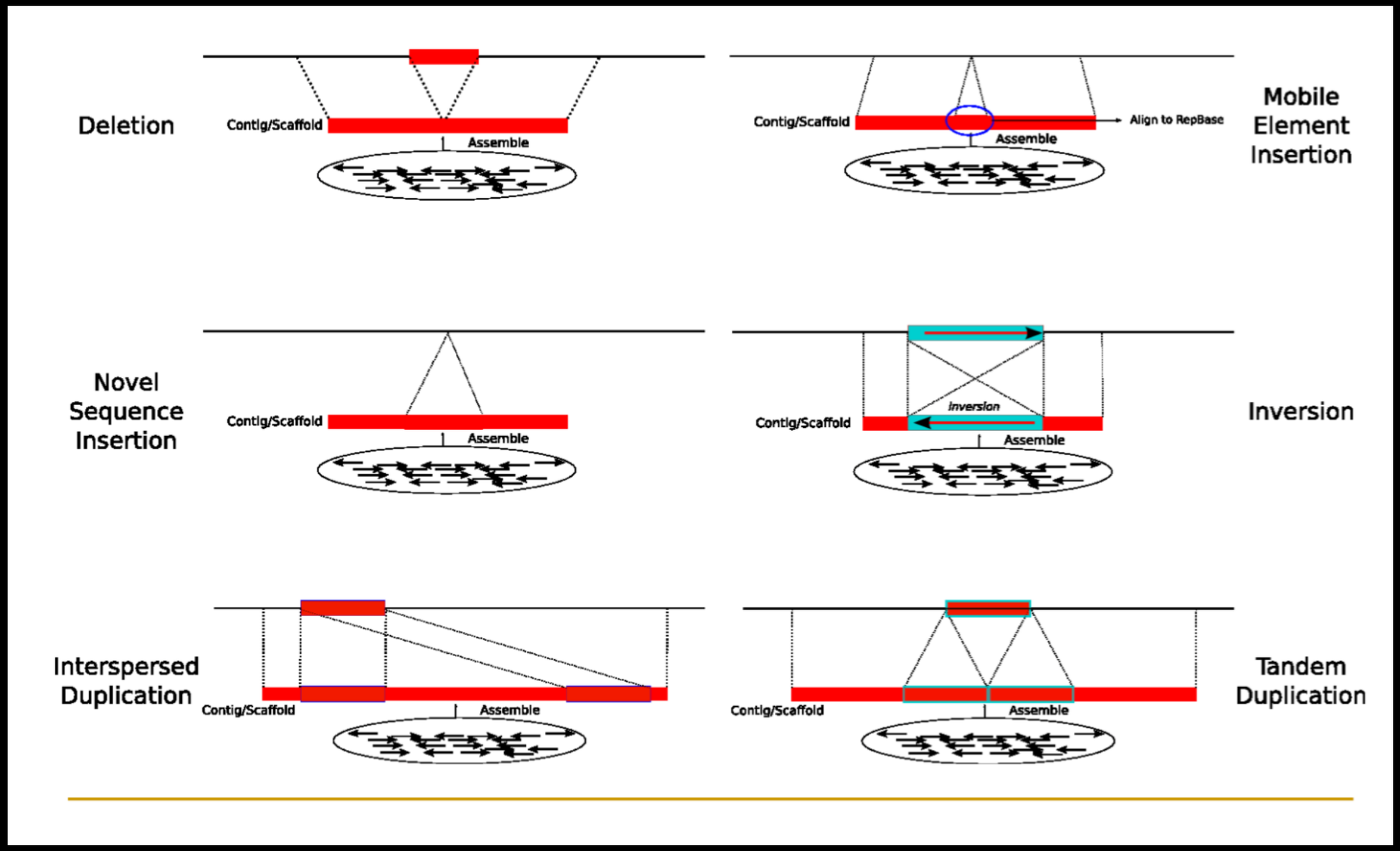
Patterns of SV from split-reads

PINDEL, SPLITREAD, indelMINER



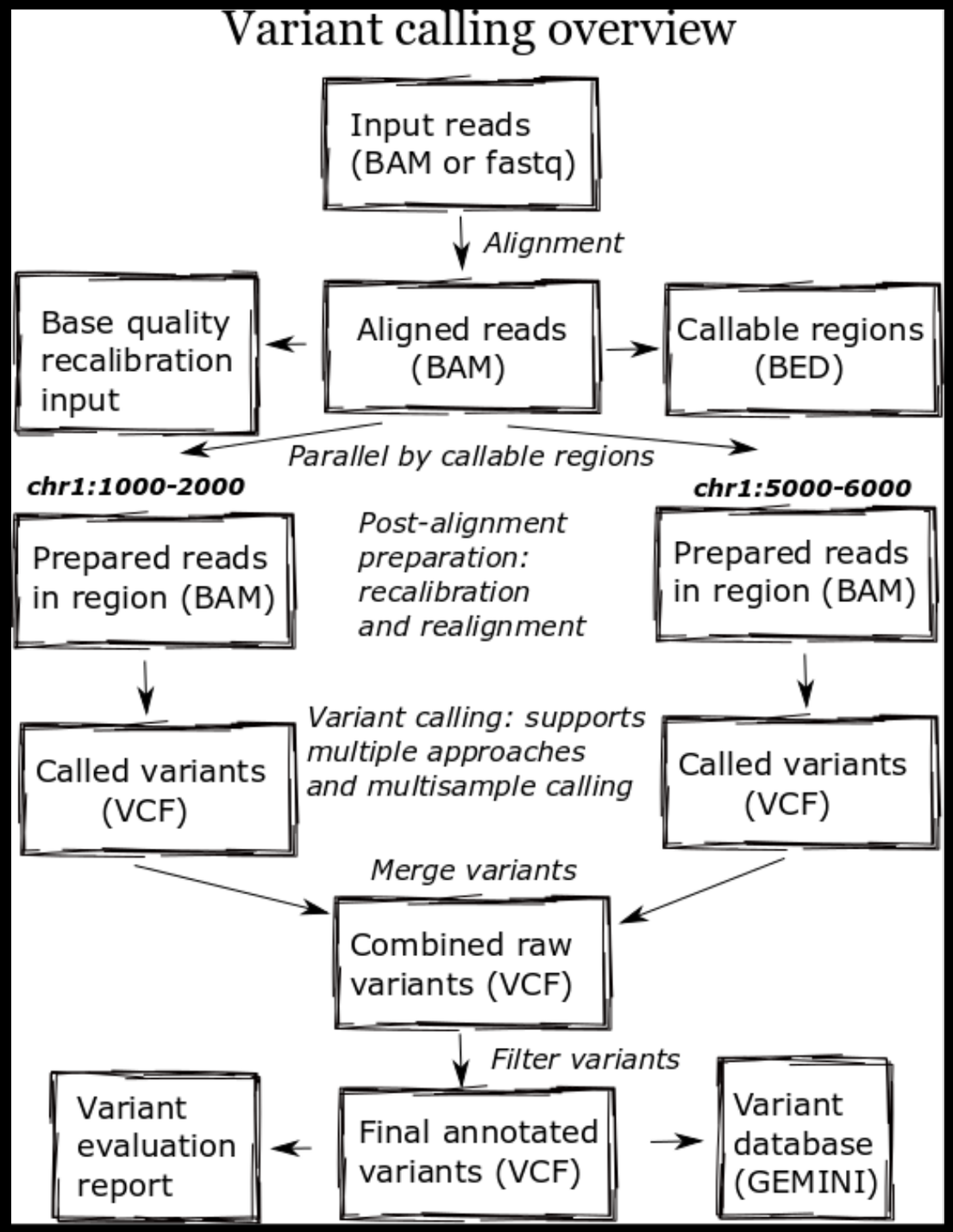
Using multiple signals

CNVer, LUMPY



Using Genome Assembly

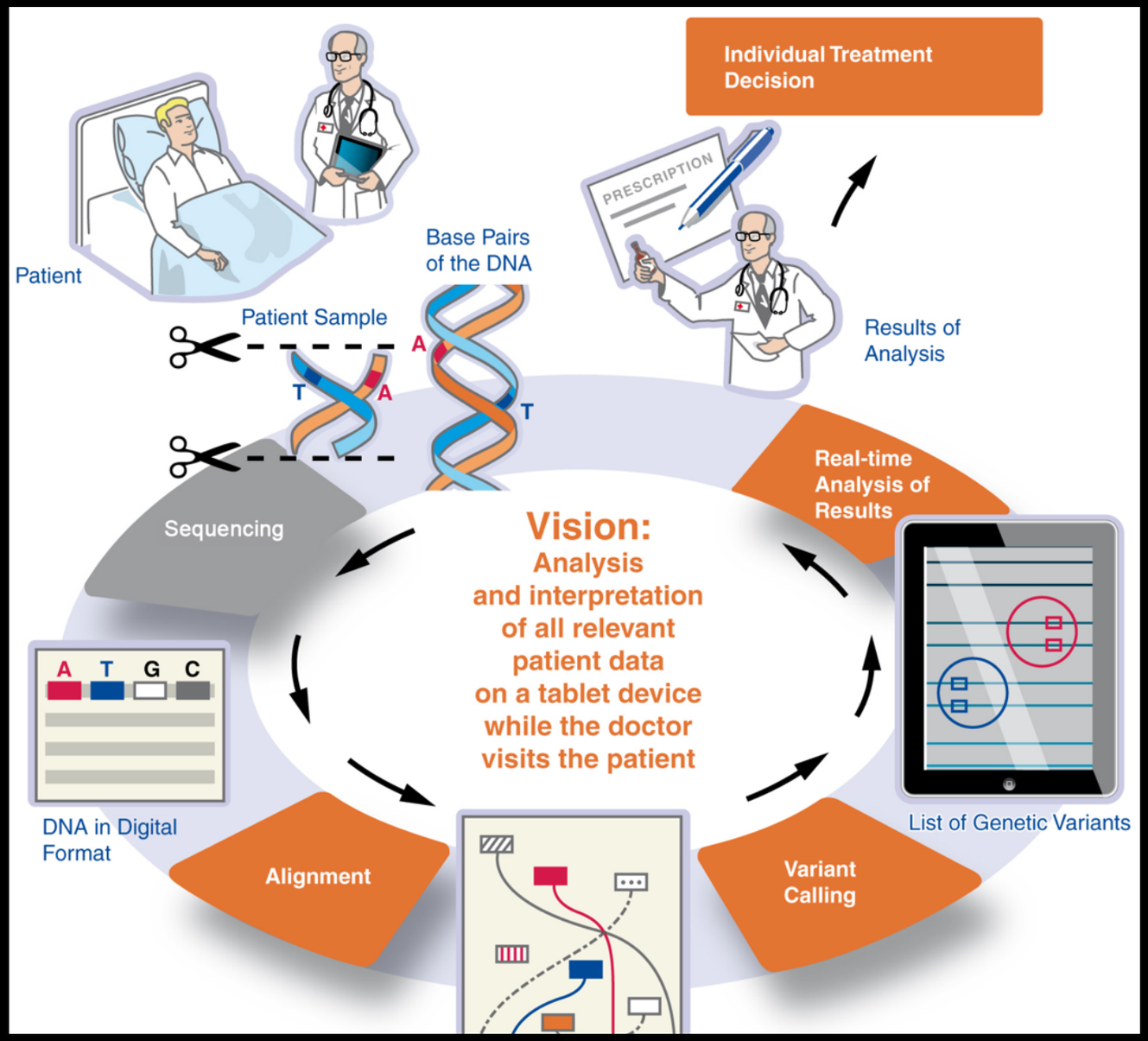
NovelSeq

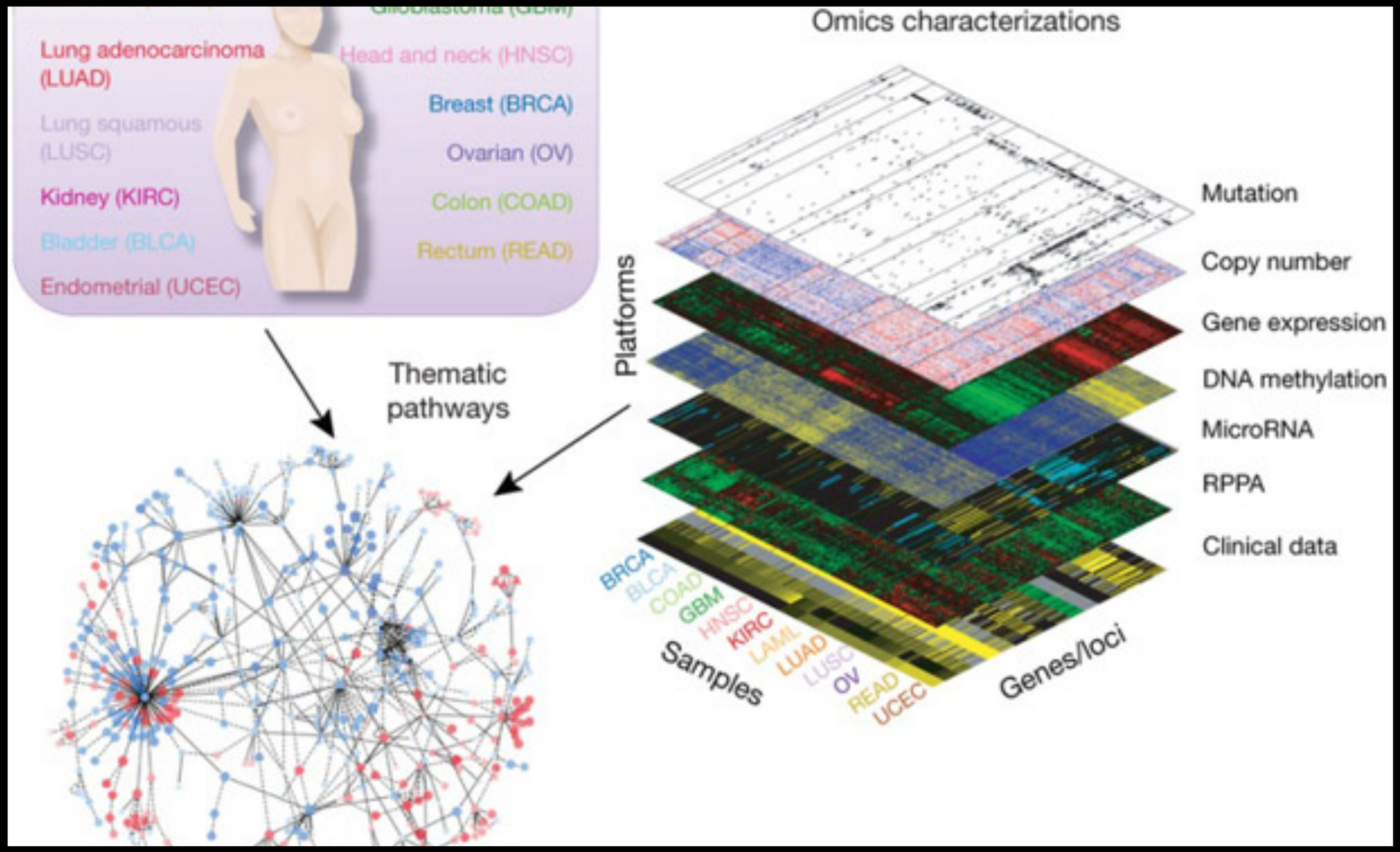


Typical Analysis Pipeline in human genome sequencing

Again, why sequence genomes?

- From genome sequences to know genome variation between individuals and study
  - Disease
  - Drug response
  - Biomes
  - Energy
  - Agriculture …





Combining Datatype to enable precision medicine, improvements in agriculture, energy …


π
3.14
HAPPY PI DAY