Practical Bioinformatics

Basic NGS

Part 1

Stefan Wyder
stefan.wyder@uzh.ch
URPP Evolution
www.evolution.uzh.ch

Universität Zürich[UZH]

Goals for today

- to know the important file formats
- to be able to use samtools and know its functions
- to know quality control measures

- to be able to use an interactive genome browser (IGV)
- to know the frequency of sequencing errors and their distribution across the read
- to know how to inspect sequence variation

NGS Applications

DNA - Genetic variability (SNPs, CNV, Indels)

Amplicon sequencing

exome re-sequencing

whole genome re-sequencing

de novo whole genome sequencing

RNA – Expression Levels and Alternative Splicing

RNA-seq (transcriptome sequencing)

small RNA (miRNA,long ncRNA,...)

Epigenetics

ChIP-seq (Chromatin immunoprecipitation)

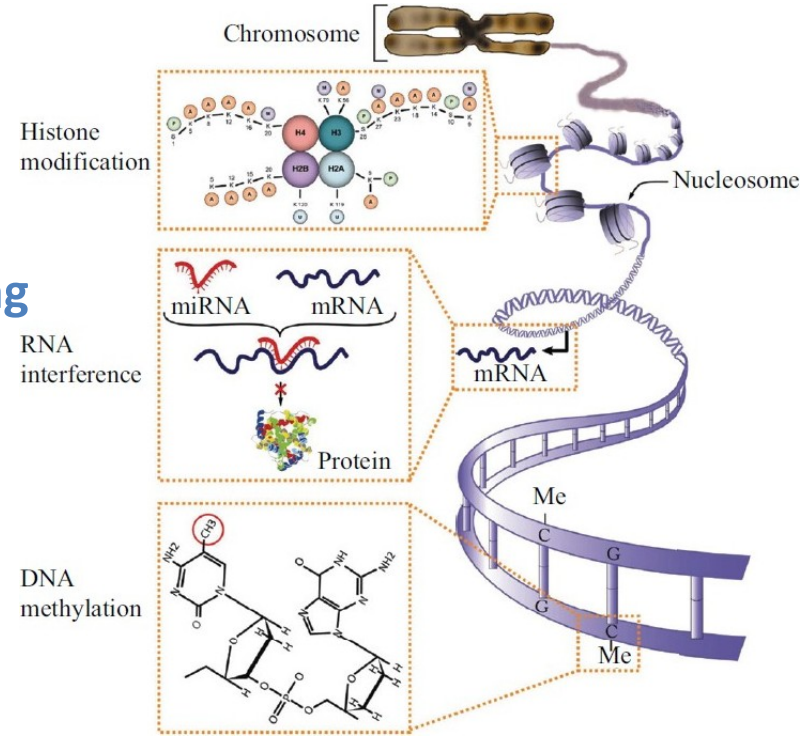
DNA methylation

Others

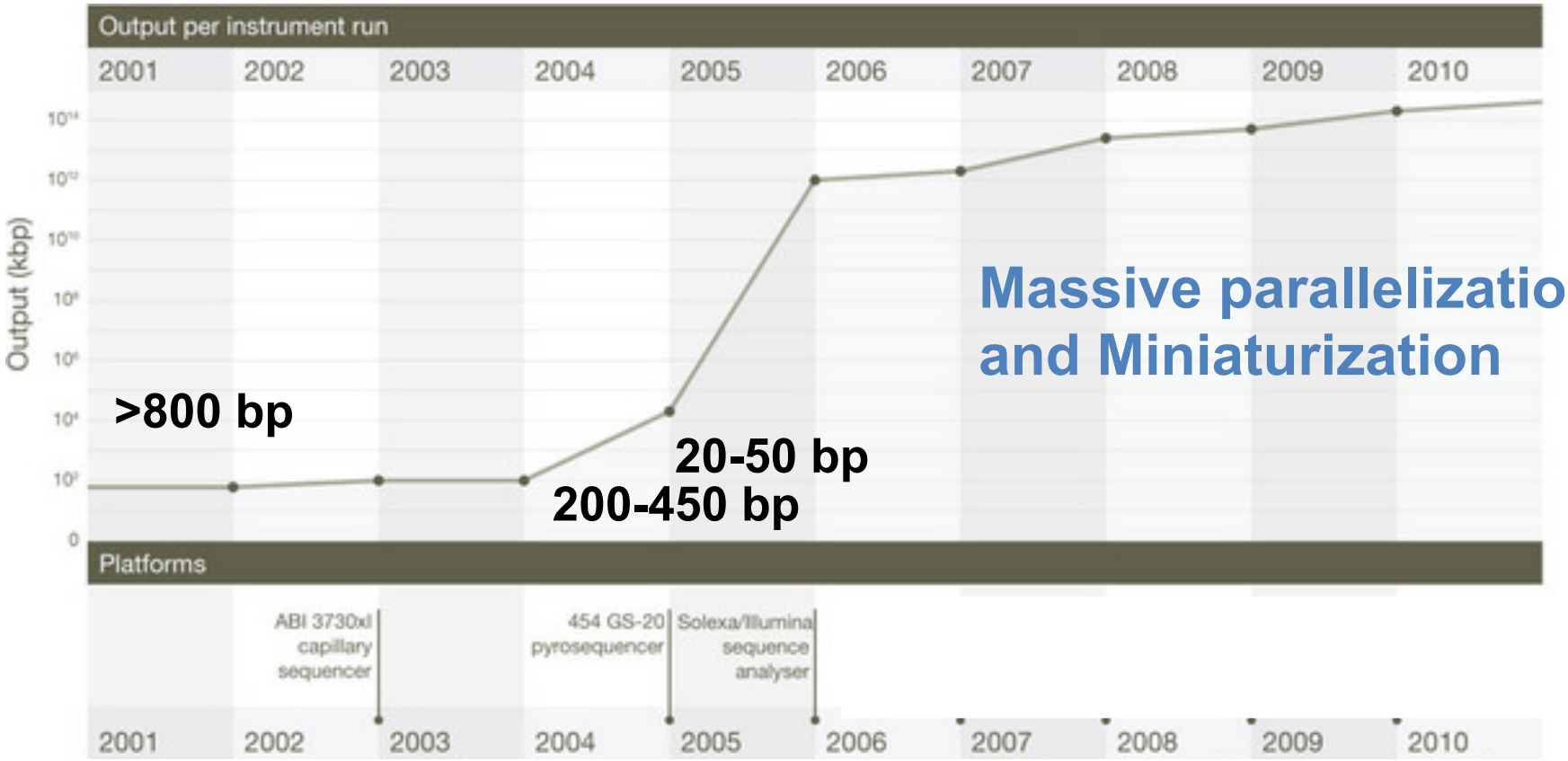
Metagenomics (16S rRNA Sequencing)

RIP-seq

....



Revolution in genomics



Massive parallelization and Miniaturization

Drawback: short reads
9x100 bp reads ≠ 1x900read

Available sequencing technology

Next Generation Sequencing

- Illumina HiSeq 2000
- Ion Torrent Proton
- Complete Genomics

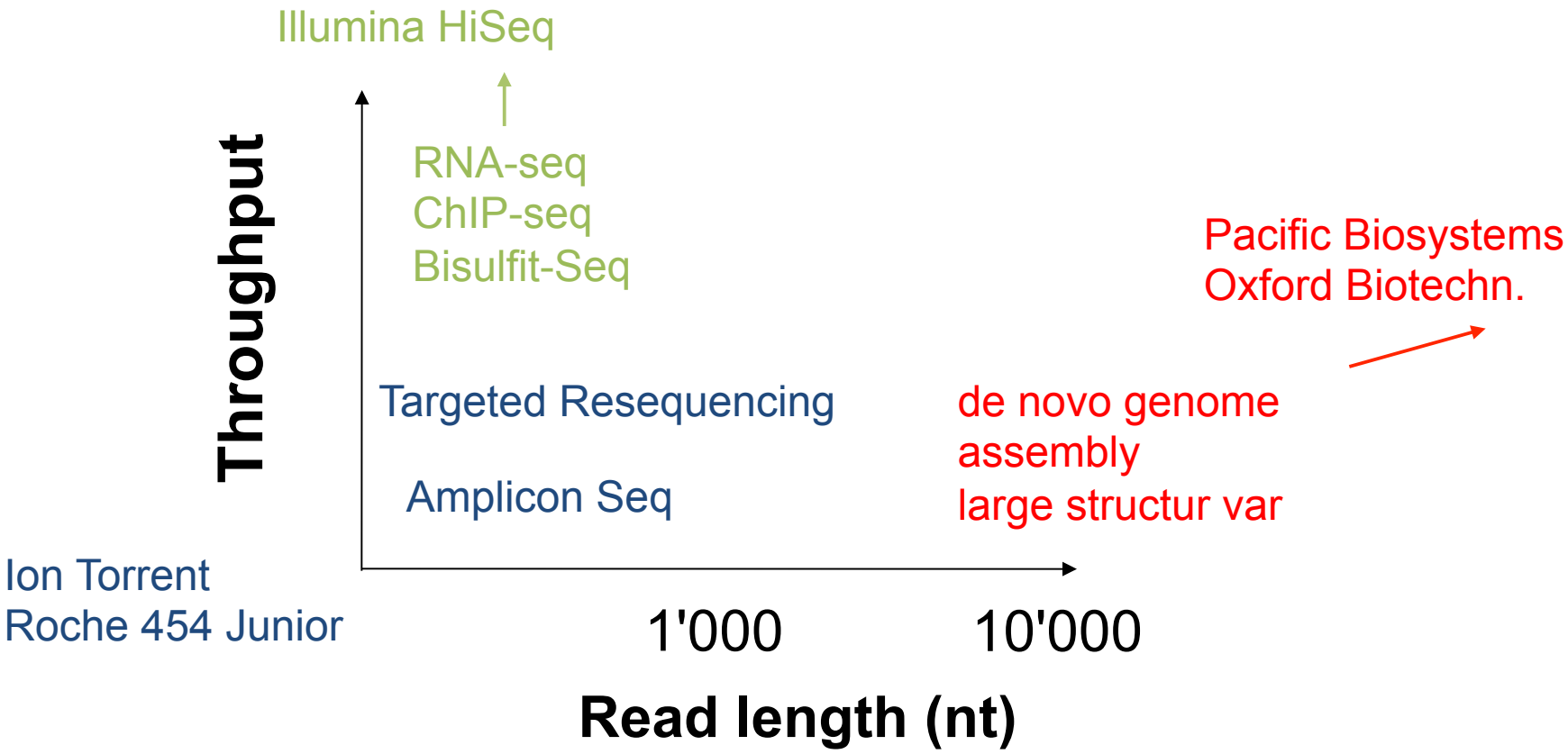
Next Next Generation Sequencing

- Pacific Biosciences
- Oxford Nanotechnologies (soon)
- .....

Benchtop sequencers

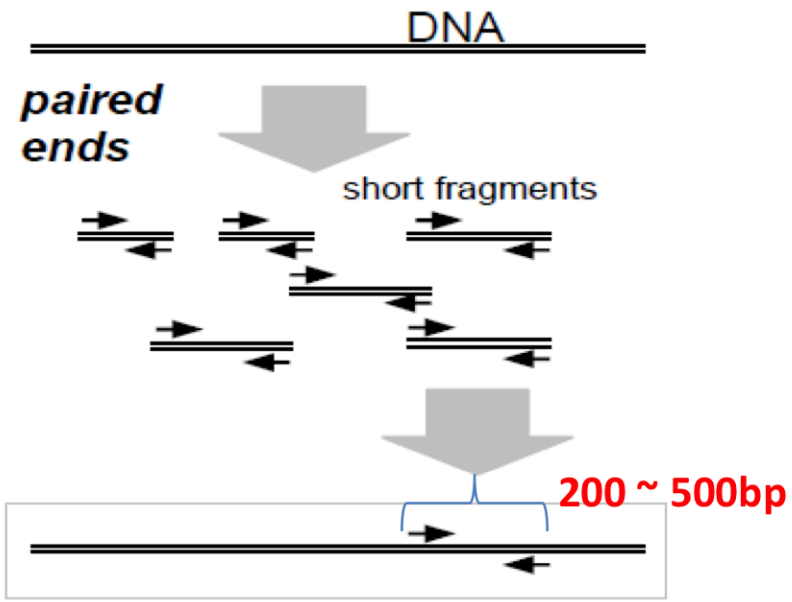
- Illumina MiSeq
- Ion Torrent PGM
- GnuBIO
- QIAGEN GeneReader (soon)
- ....

Different niches



Single-end, Paired-end, Mate-pair?

paired-ends (PE)

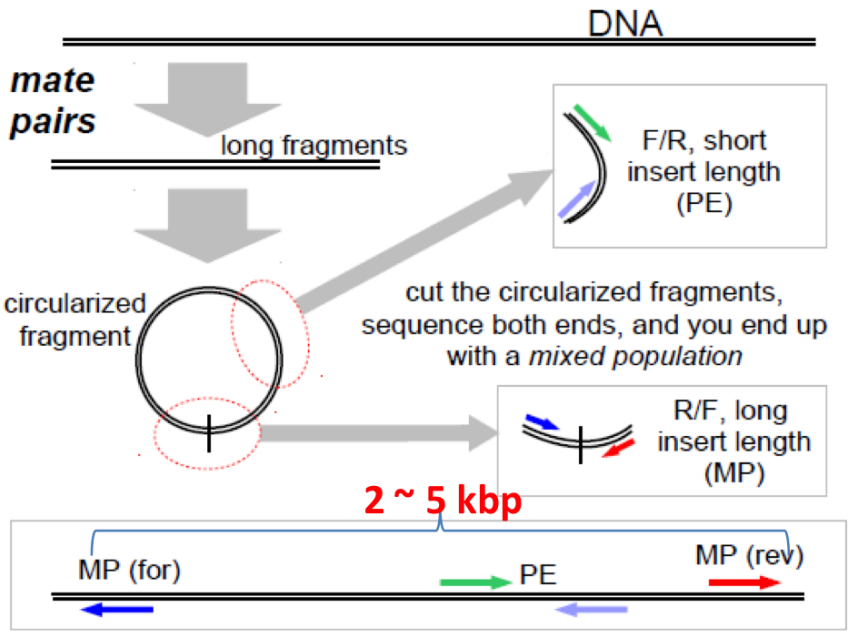


increases the mapping accuracy

RNA-seq

genome resequencing

mate-pair (MP)



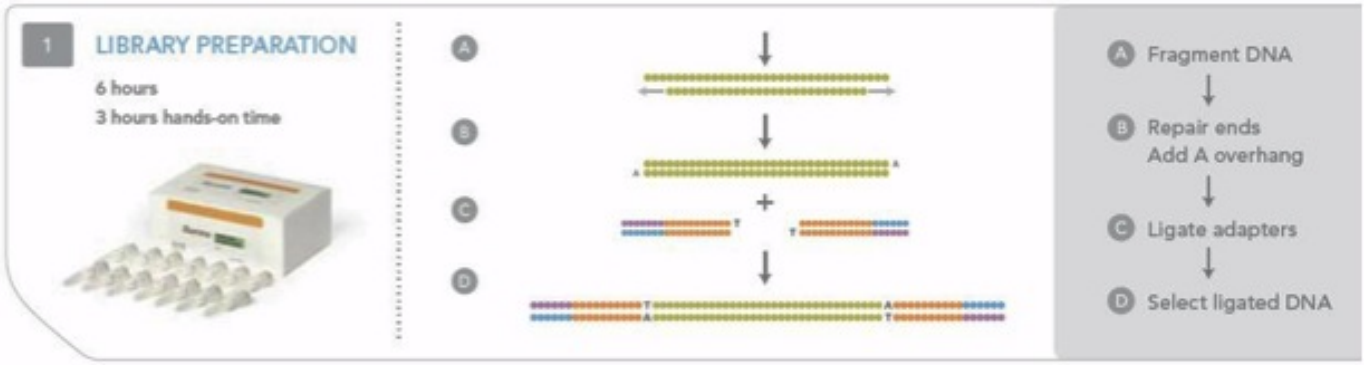
de novo genome sequencing

GDC, ETHZ

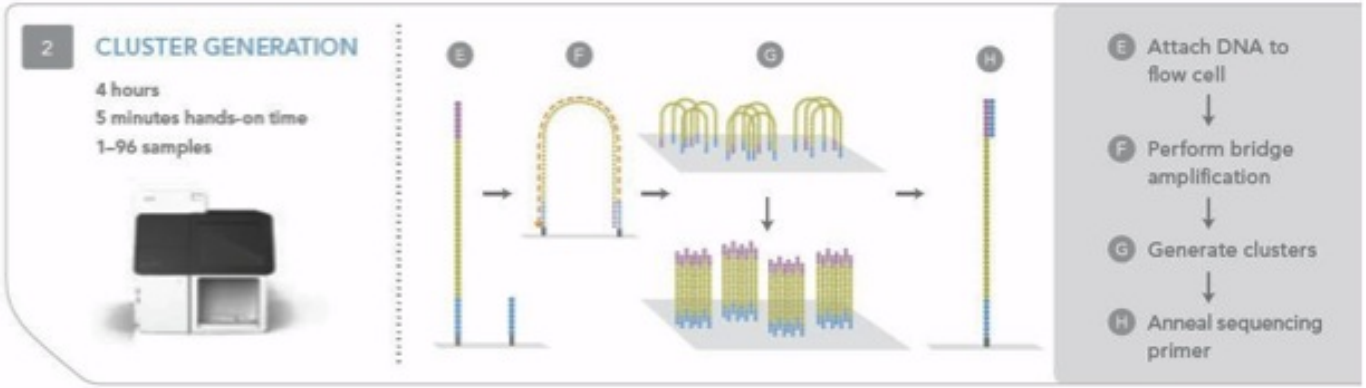
Illumina sequencing

- Illumina HiSeq 2000 (2x100 bp, >400 Gb)
- Benchtop sequencer: MiSeq (2x250 bp, max. 8 Gb)

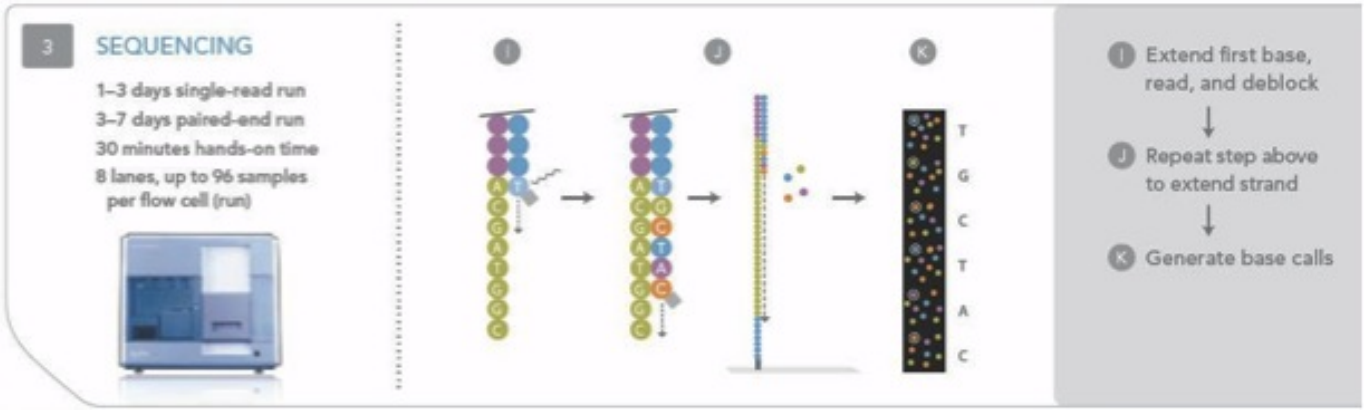
Illumina Sequencing



PCR template



Slide with lawn of primers



Sequencing-by-synthesis using 3'-blocked labeled nucleotides

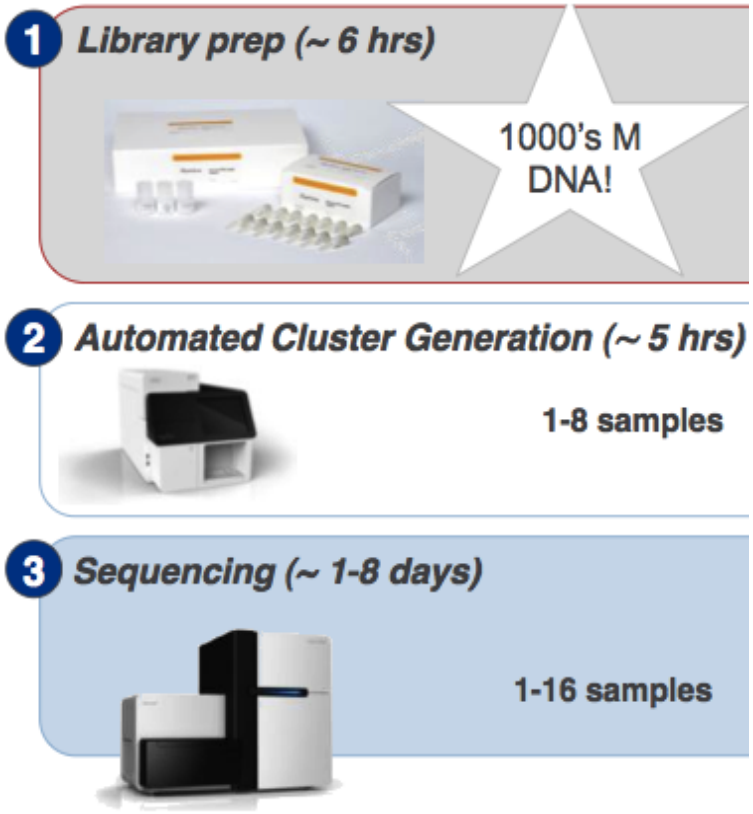
Illumina Sequencing

Animation

http://www.wellcome.ac.uk/Education-resources/Education-and-learning/Resources/Animation/

NGS has some drawbacks

- Work/Equipment for Library prep
- Amplification Bias (bridge PCR)
- Short reads (100 <<1000nt) inherently limited by your ability to keep all the nascent strands in sync



Biases

Simple counting - leaving behind all the problems with microarrays?

Published online 26 July 2008

Nucleic Acids Research, 2008, Vol. 36, No. 16 ***e105***
doi:10.1093/nar/gkn425

Substantial biases in ultra-short read data sets from high-throughput DNA sequencing

Juliane C. Dohm[1], Claudio Lottaz[2], Tatiana Borodina[1] and Heinz Himmelbauer[1,*]

Published online 14 April 2010

Nucleic Acids Research, 2010, Vol. 38, No. 12 ***e131***
doi:10.1093/nar/gkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen[1,*], Steven E. Brenner[2] and Sandrine Dudoit[1,3]

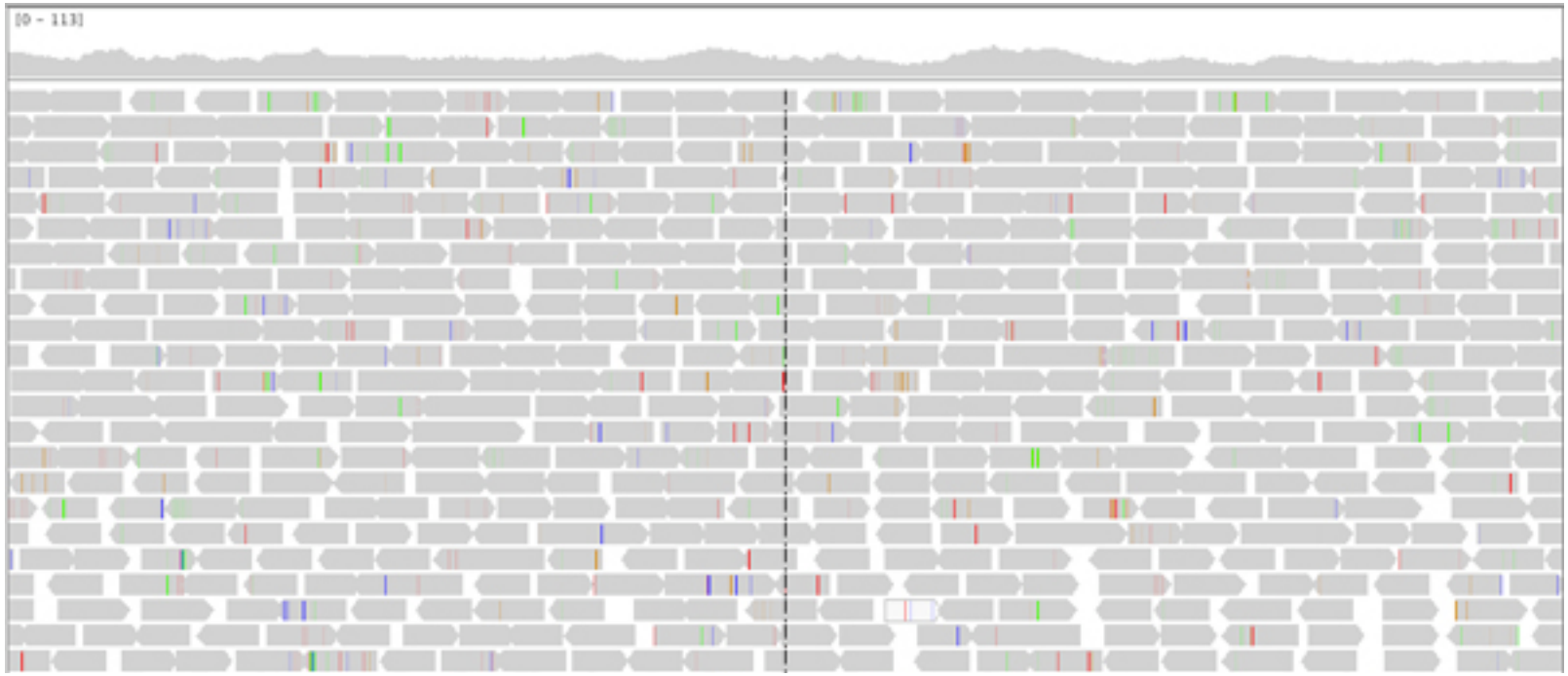
Published online 16 May 2011

Nucleic Acids Research, 2011, Vol. 39, No. 13 ***e90***
doi:10.1093/nar/gkr344

Sequence-specific error profile of Illumina sequencers

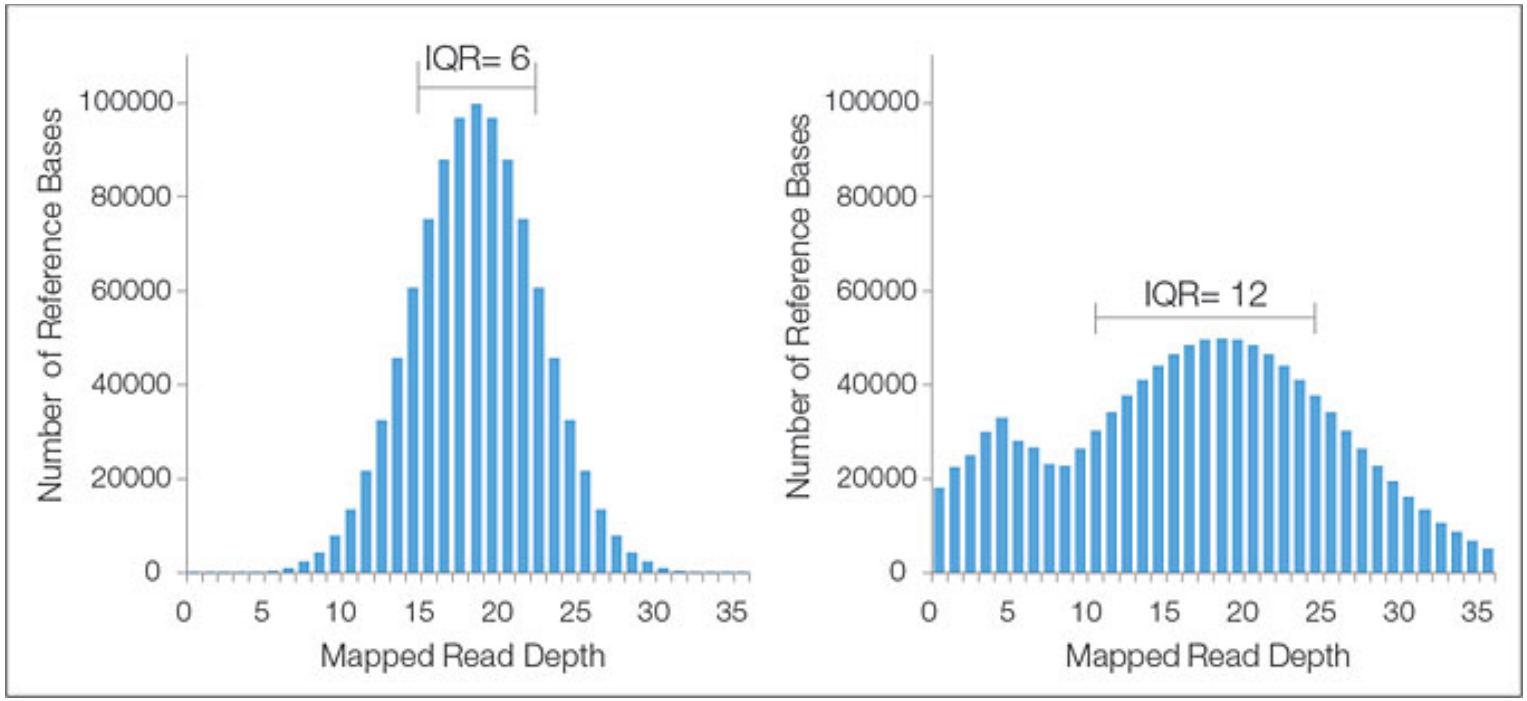
Kensuke Nakamura[1,*], Taku Oshima[2], Takuya Morimoto[2,3], Shun Ikeda[1], Hirofumi Yoshikawa[4,5], Yuh Shiwa[5], Shu Ishikawa[2], Margaret C. Linak[6], Aki Hirai[1], Hiroki Takahashi[1], Md. Altaf-Ul-Amin[1], Naotake Ogasawara[2] and Shigehiko Kanaya[1]

Unequal read depth



dependent on GC-content, library protocol, ...

Read depth histograms



Good

Bad

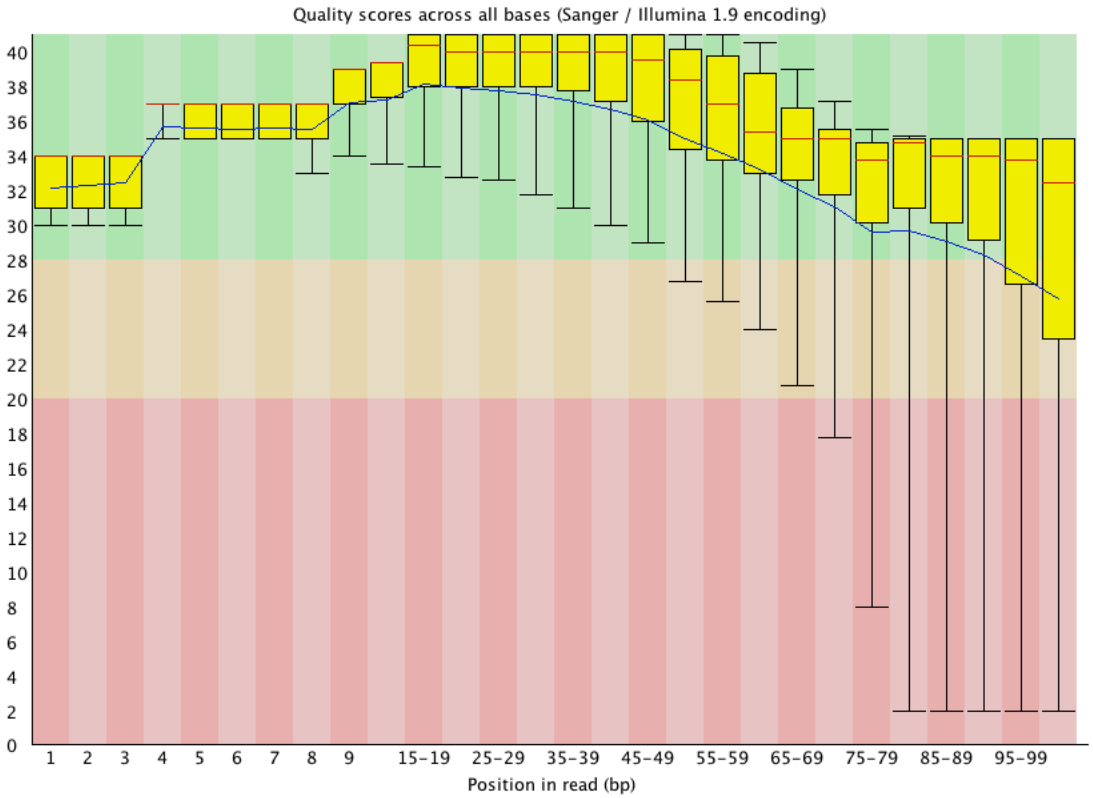
Illumina.com

Sequencing errors

Error rate and error profile are technology-specific

Illumina Sequencing

- Error Rate: > 0.1% (i.e. > 1 in 1000)
- mainly substitutions errors
- errors mostly at read's start or end



FASTQ format & base qualities

```
@read1
TTGTGTTCAAAATATATAATTTATTTATAAGCTATAATCTTATGNNNNNNNCTCCTTCTTAGCTT
+
@C@DDDDDFHHHHJJDHIIIJI@HHGGIDGEBDEIEIIIIJJII#######008BGGGHIIGGH>
```

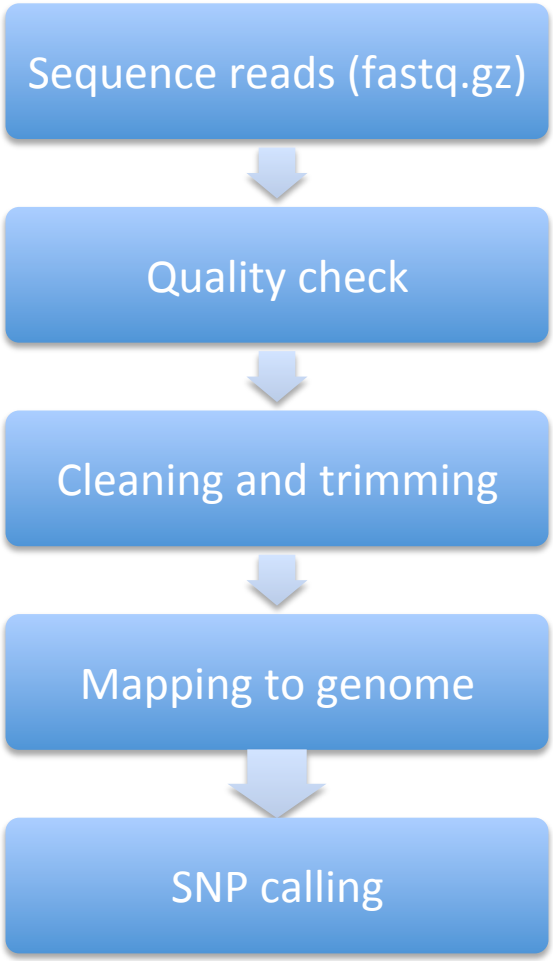
@ = ASCII code 64
BQ = ASCII code – 33 = **31**

Base Quality: Phred Score Q_{phred}

$Q_{phred} = -10 * \log_{10} (P_{error})$

Base Quality	P_{error}
3	50 %
5	32 %
10	10 %
20	1 %
30	0.1 %
40	0.01 %

Workflow



fastQC

(cutadapt)

bowtie2

samtools / GATK

Mapping Reads

Mapping / Alignment to genome or transcriptome:
Find the genomic location the read originates from
(by taking into account base call qualities)

How to deal with non-unique hits

- skip them
- place them randomly

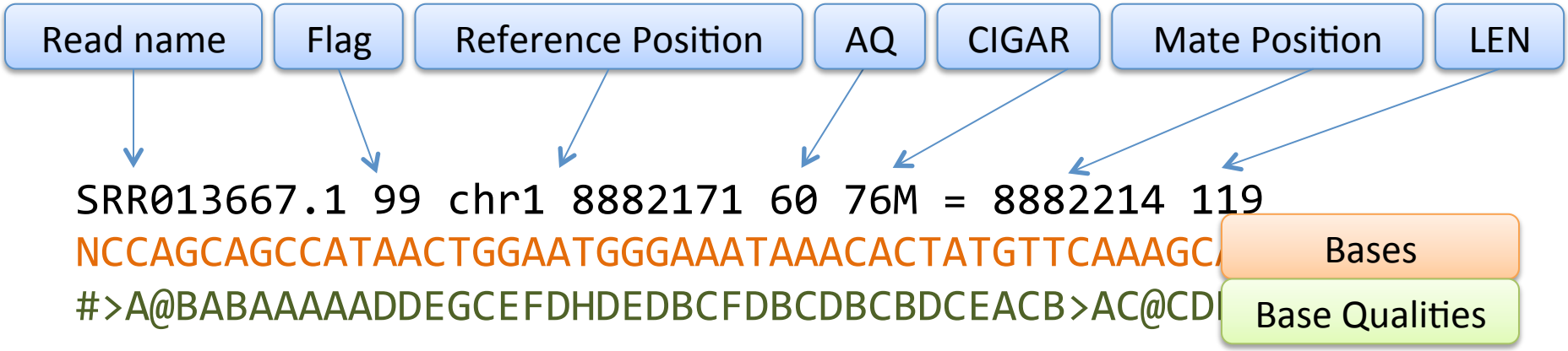
Mapping programs use heuristics

Many alignment software

- **Speed**
  - SNAP (http://snap.cs.berkeley.edu/)
  - iSAAC (http://www.illumina.com/)
- **Accuracy**
  - NovoAlign (http://www.novocraft.com)
  - Razers3 (http://www.seqan.de/projects/razers/)
- **All-round**
  - bwa & bwa-mem (http://bio-bwa.sourceforge.net/)
  - Bowtie (http://bowtie-bio.sourceforge.net)
- **Functionality** (e.g. de novo splice aligners)
  - STAR (https://code.google.com/p/rna-star/)
  - TopHat (http://tophat.cbcb.umd.edu/)

Output Formats: SAM & BAM

- SAM http://samtools.sourceforge.net/SAMv1.pdf



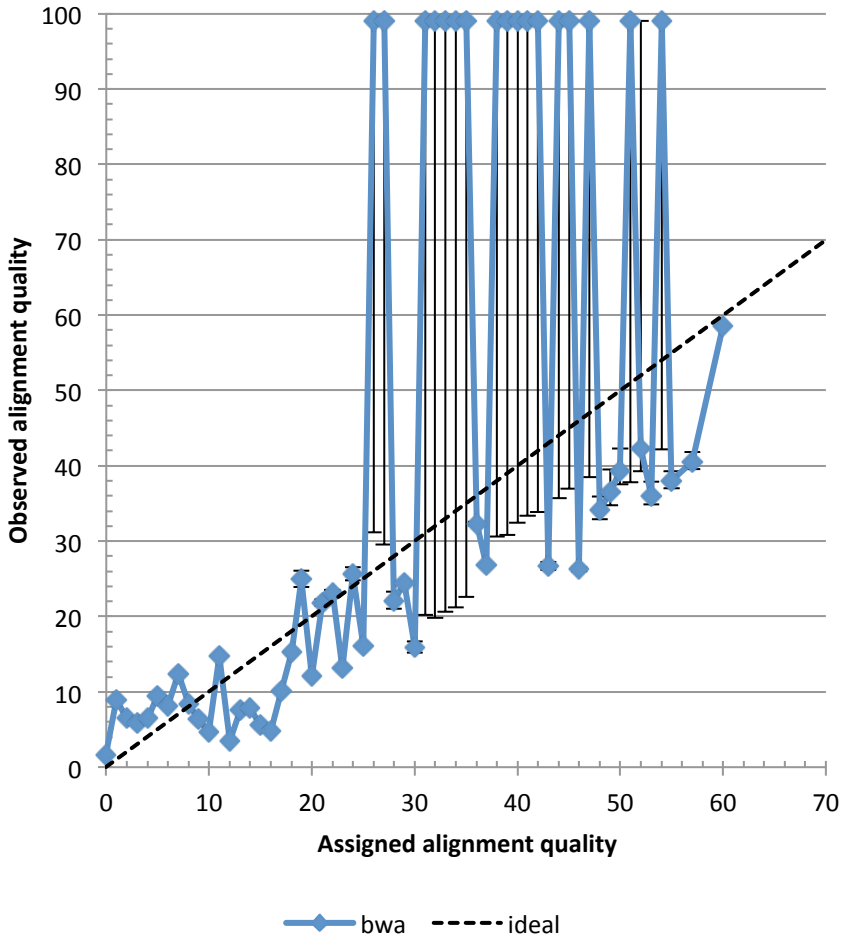
- BAM
  - binary version of SAM

Alignment Quality

The probability that an alignment has been misaligned:

$$AQ = -10 * \log_{10}(P_{misaligned})$$

AQ=0 means multiple hits to genome



Alignment postprocessing

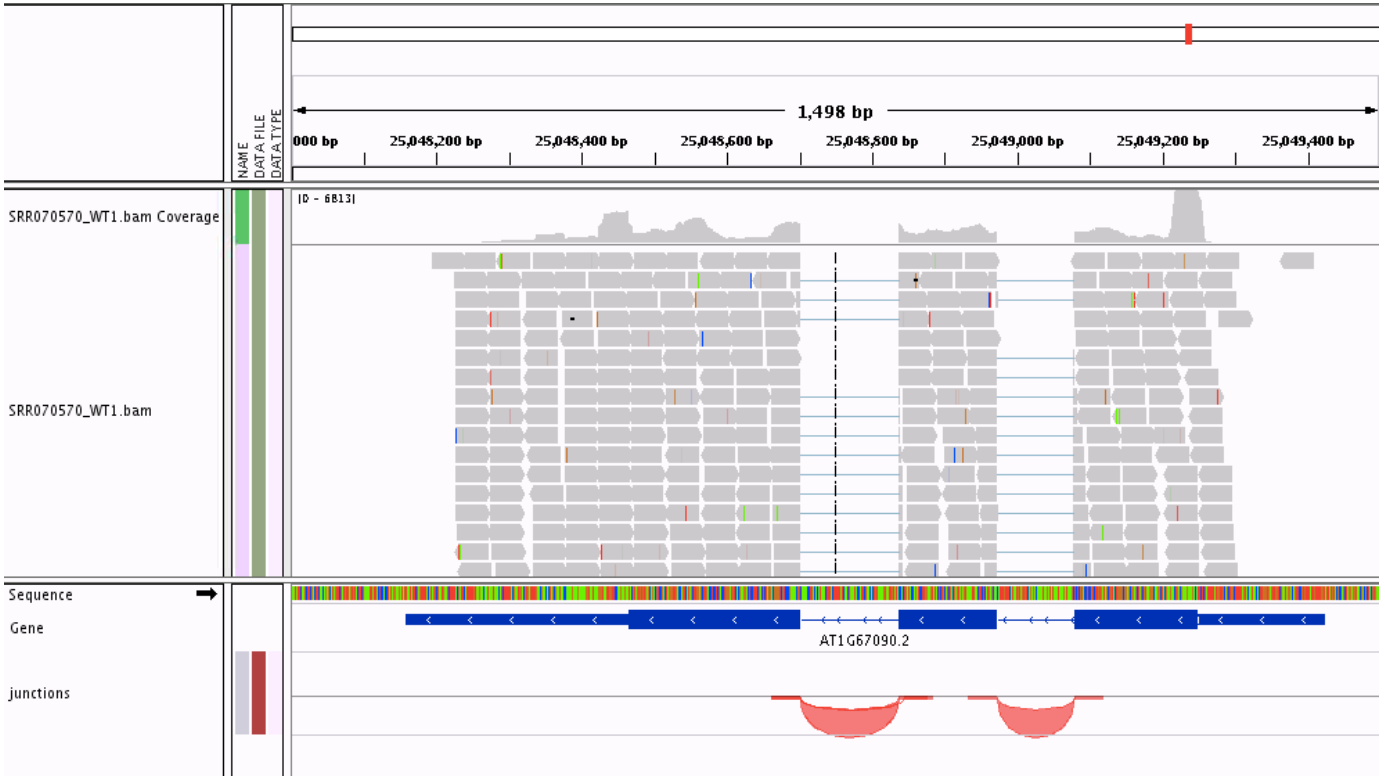
Depending on the application:

- Duplicate Removal
  (use samtools rmdup or Picard)
- Indel Cleaning / Realignment
  (GATK)
- Filtering reads based on flags
  (use samtools or Picard)

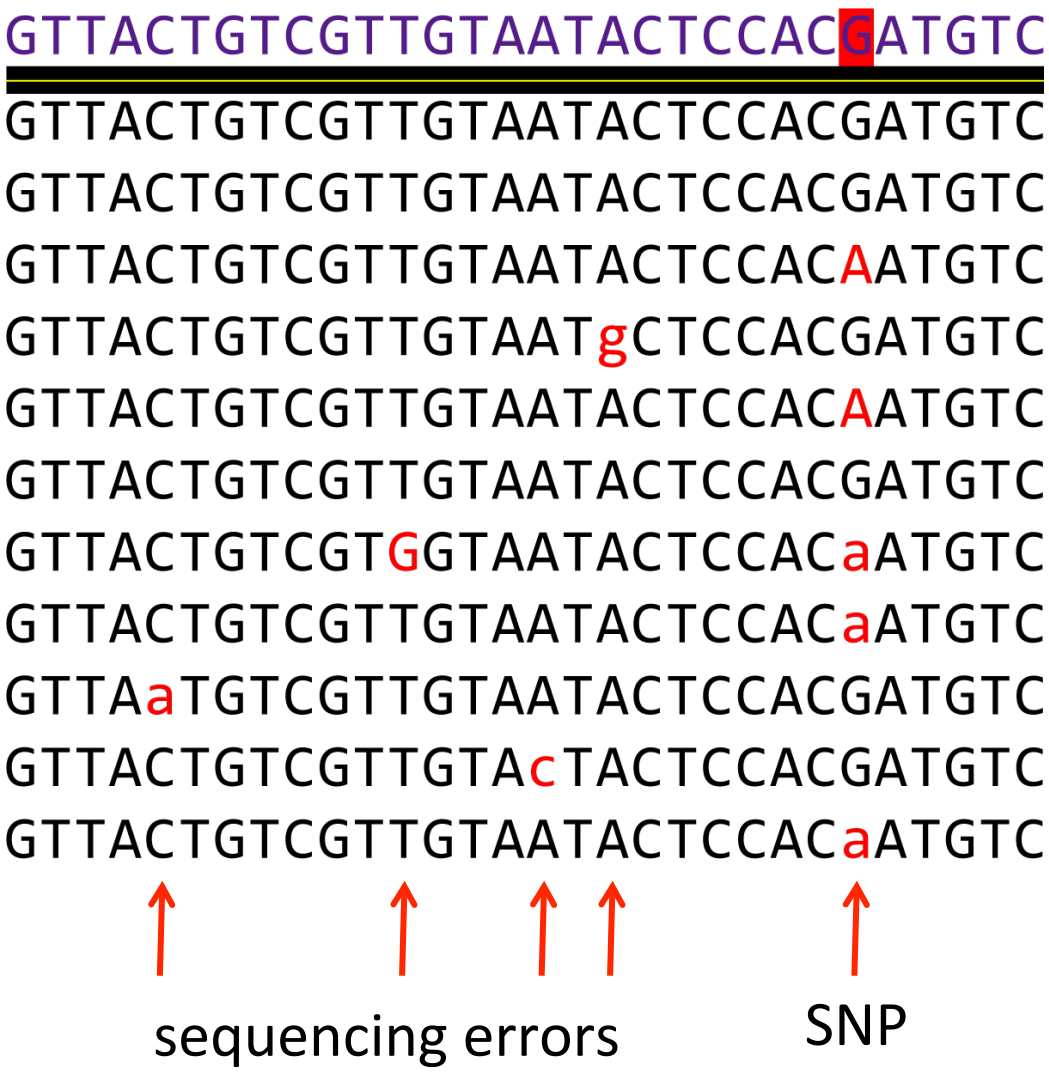
INDEL Cleaning

Alignment visualization

- Many Genome Browsers are available with different strength
- today: IGV



SNP Discovery



A word of caution

- Good experiments start with good planning

  Quality / quantity of DNA/RNA
  Choice of technology / protocol
  Enrichment / Capturing?

Experimental Design

- Statistical rules still apply!

  Sufficient number of biological replicates

  Sufficient coverage

Sources & Links

Article Collections

- Review Articles from Nature Reviews Genetics
- PLoS Computational Biology: Education

Material

- SEQanswers NGS forum http://seqanswers.com/
- Biostar http://biostars.org/
- List of Applications http://seqanswers.com/wiki/Special:BrowseData/