# Algorithms for Bioinformatics

## 2018/2019

## *High-throughput sequencing applications*

Pedro G. Ferreira
[d**CC**] @ Faculty of Sciences University of Porto

# Sequencing technologies



- High-throughput sequencing technologies
- DNA variant calling
- DNA Assembly
- Large scale sequencing projects
- Expression analysis
- Exercises

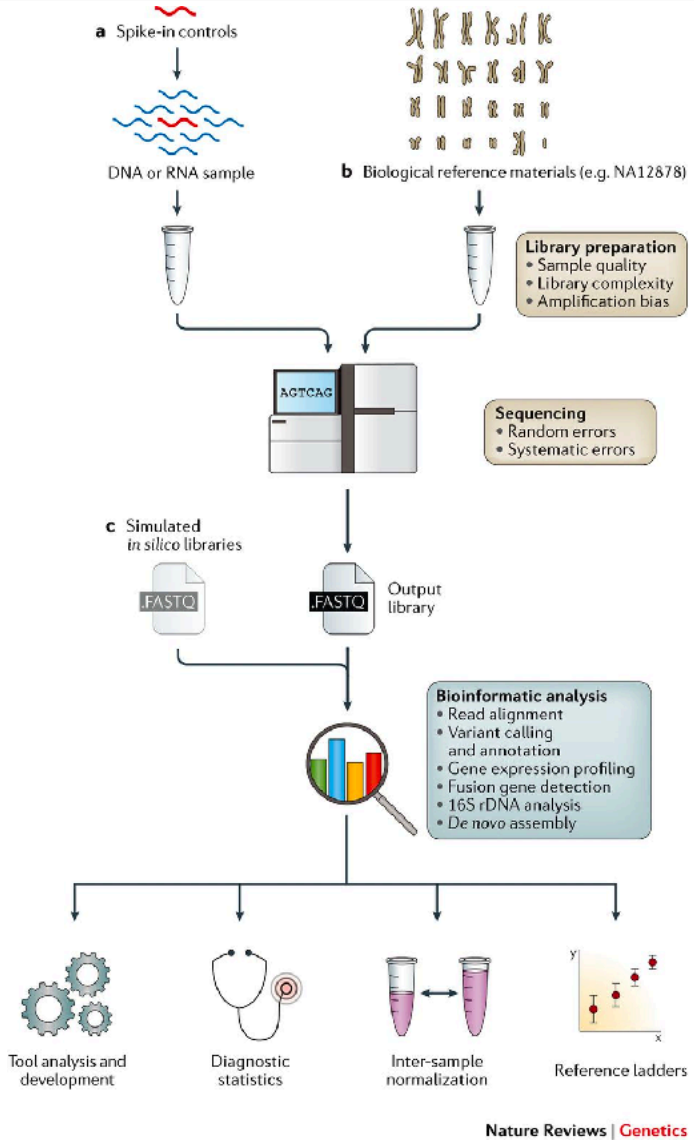
# Sequencing Technologies

Typical applications of massive (also called next-generation or high-throughput) sequencing technologies:

**DNA Sequencing**

- De novo Genome Assembly
- Single Nucleotide Variation discovery
- Copy Number Variation detection
- Structural Rearrangements

# Sequencing Technologies




Reference standards for next-generation sequencing
Hardwick et al. Nature Reviews Genetics volume 18, pages 473–484 (2017)

# *de novo* Genome Assembly



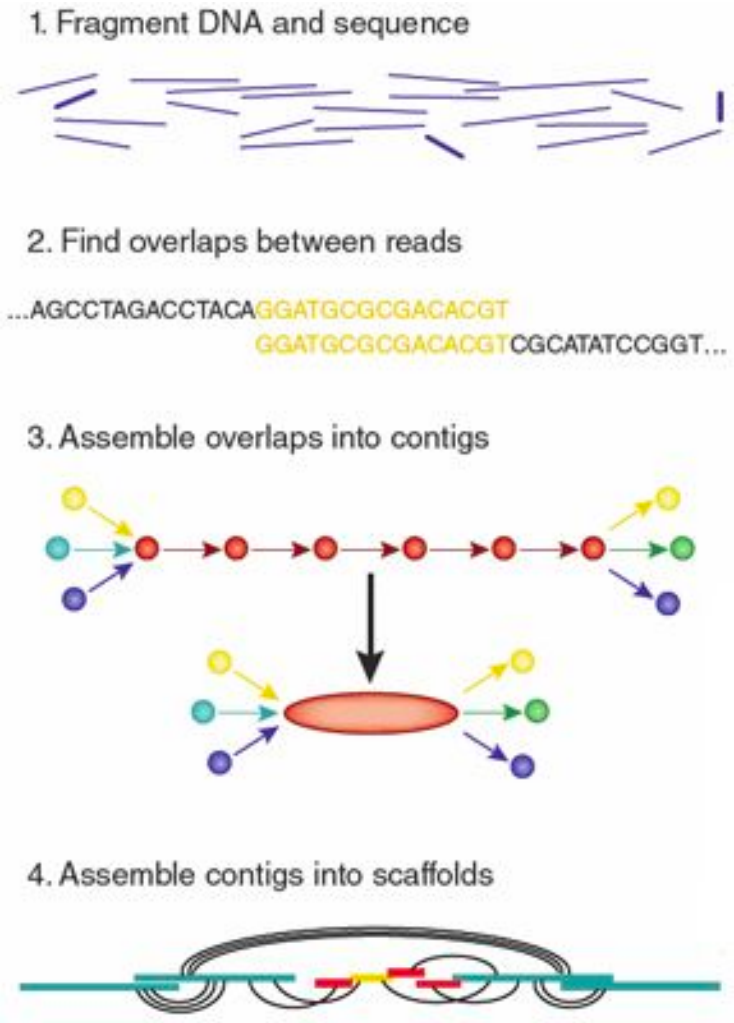
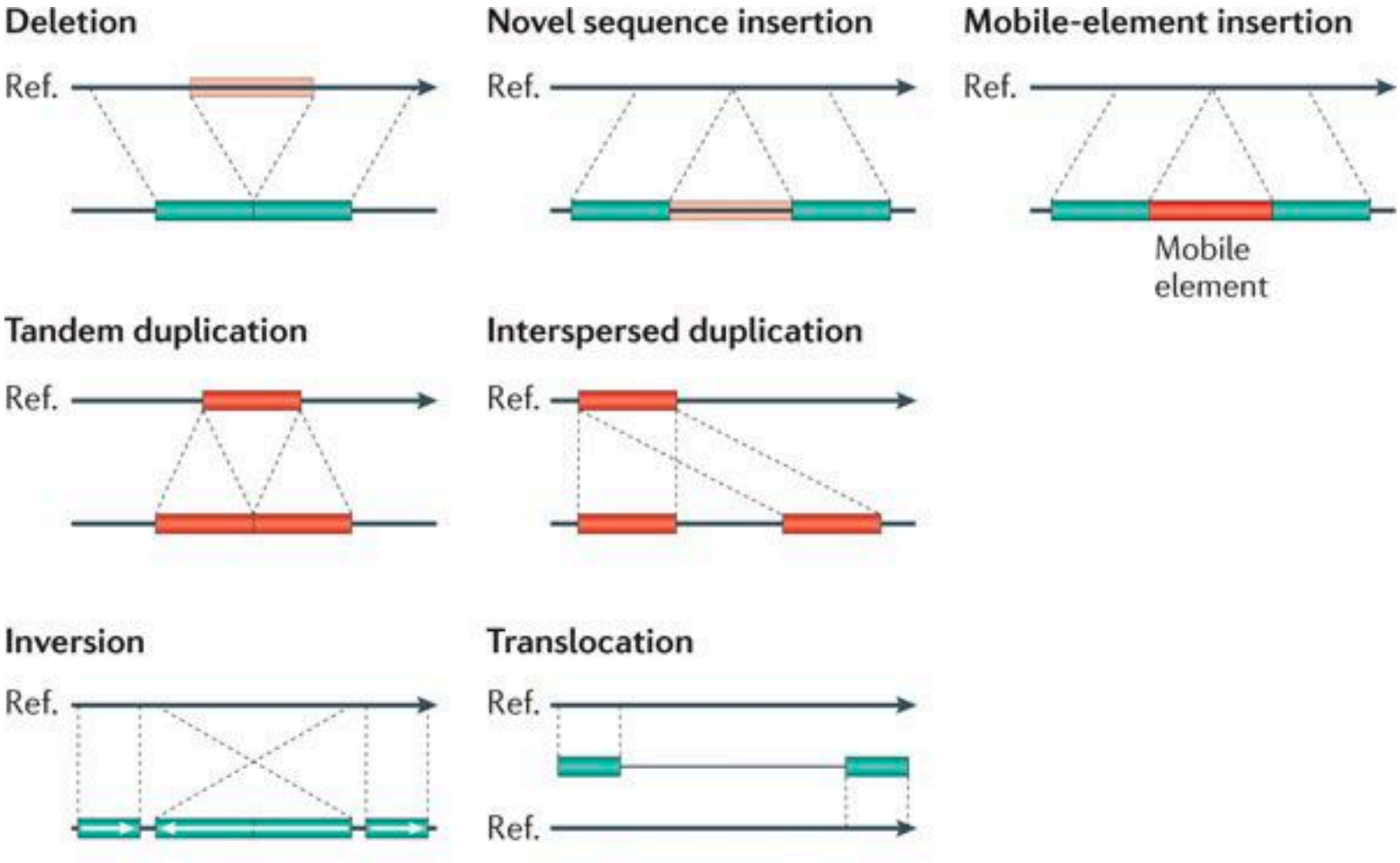



Image: Michael Schatz, Cold Spring Harbor
De novo genome assembly: what every biologist should know
Monya Baker Nature Methods volume 9, pages 333–337 (2012)

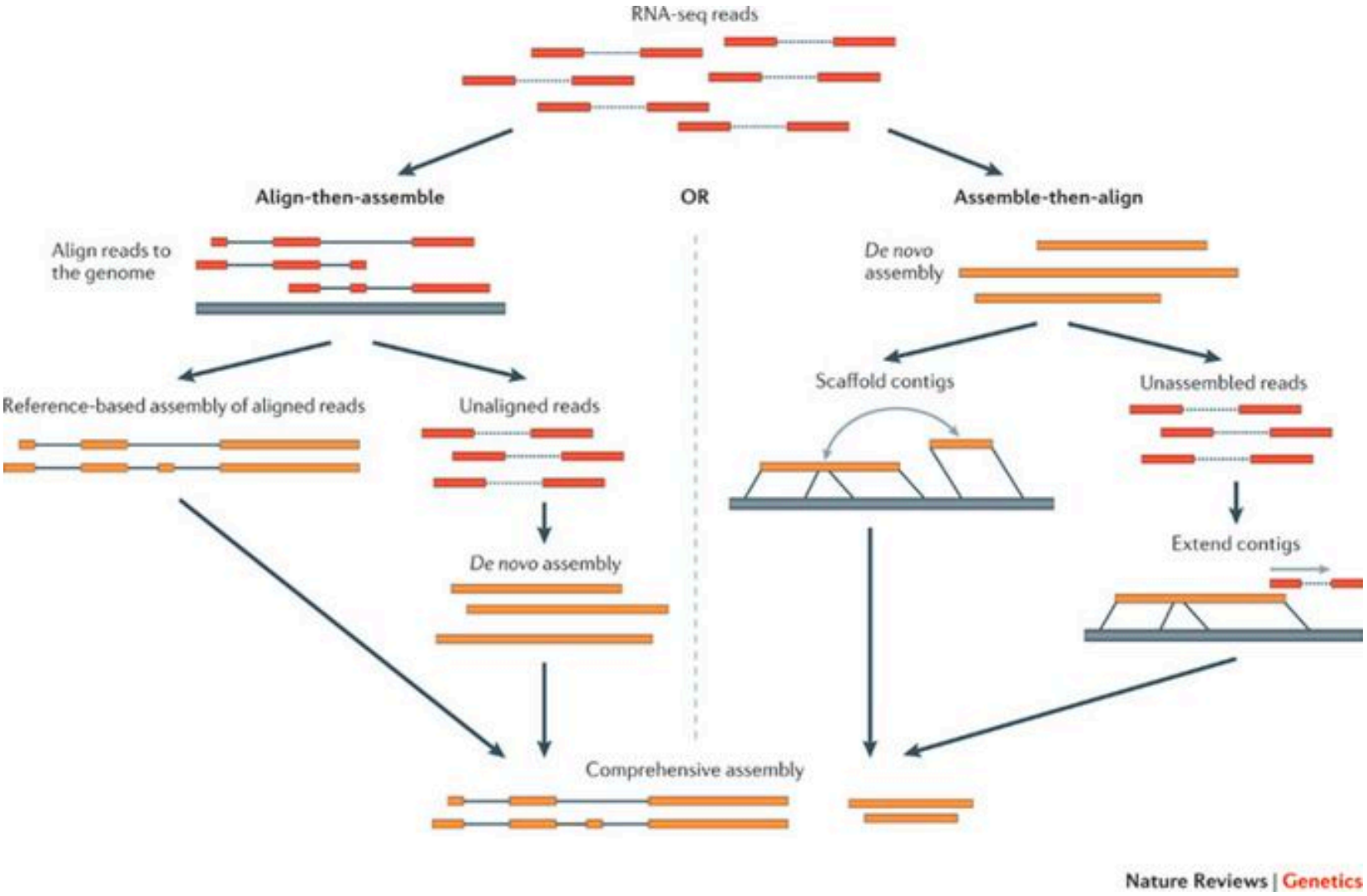
# Genome Structural Variation






Genome structural variation discovery and genotyping
Alkan et al. Nature Reviews Genetics volume 12, pages 363–376 (2011)

# Transcriptome Assembly



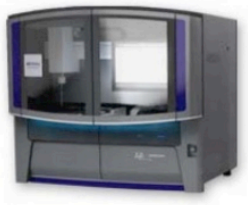

Next-generation transcriptome assembly
Martin & Wang. Nature Reviews Genetics volume 12, pages 671–682 (2011)

# Sequencing Technologies

- **454**, 800bp, 1 day, 0.7Gb
- **Illumina**, 2*100bp reads, 2-5 days, 120Gb
- **Solid**, 85bp, 8 days, 150Gb
- **PacBio**, 3kb-15kb, 20mins, 3Gb

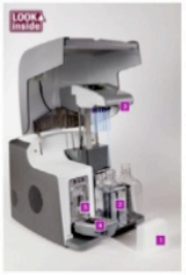
GS FLX 454
(ROCHE)

HiSeq 2000
(ILLUMINA)



5500xl SOLiD
(ABI)

Ion TORRENT



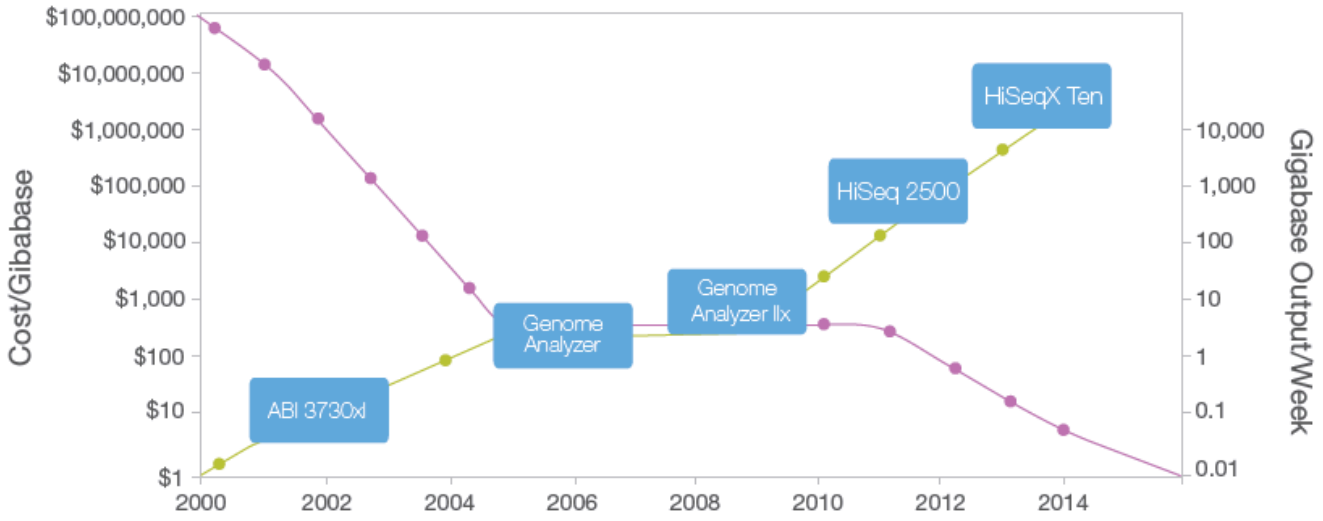
GS Junior

| |  MiniSeq System |  MiSeq Series |  NextSeq Series |  HiSeq Series |  HiSeq X Series* |
|---|---|---|---|---|---|
| **Key Methods** | Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing. | Small genome, amplicon, and targeted gene panel sequencing. | Everyday exome, transcriptome, and targeted resequencing. | Production-scale genome, exome, transcriptome sequencing, and more. | Population- and production-scale whole-genome sequencing. |
| **Maximum Output** | 7.5 Gb | 15 Gb | 120 Gb | 1500 Gb | 1800 Gb |
| **Maximum Reads per Run** | 25 million | 25 million† | 400 million | 5 billion | 6 billion |
| **Maximum Read Length** | 2 × 150 bp | 2 × 300 bp | 2 × 150 bp | 2 × 150 bp | 2 × 150 bp |

Adapted from Illumina.com

# Sequencing Cost and Data Output

## Cost and Ouput of Ilumina sequencing machines across the years




**MiSeq Series**
Small genome, amplicon and targeted gene panel sequencing.

**NextSeq Series**
Everyday genome, exome transcriptome sequencing, and more.

**HiSeq Series**
Production-scale genome, exome, transcriptome sequencing and more.

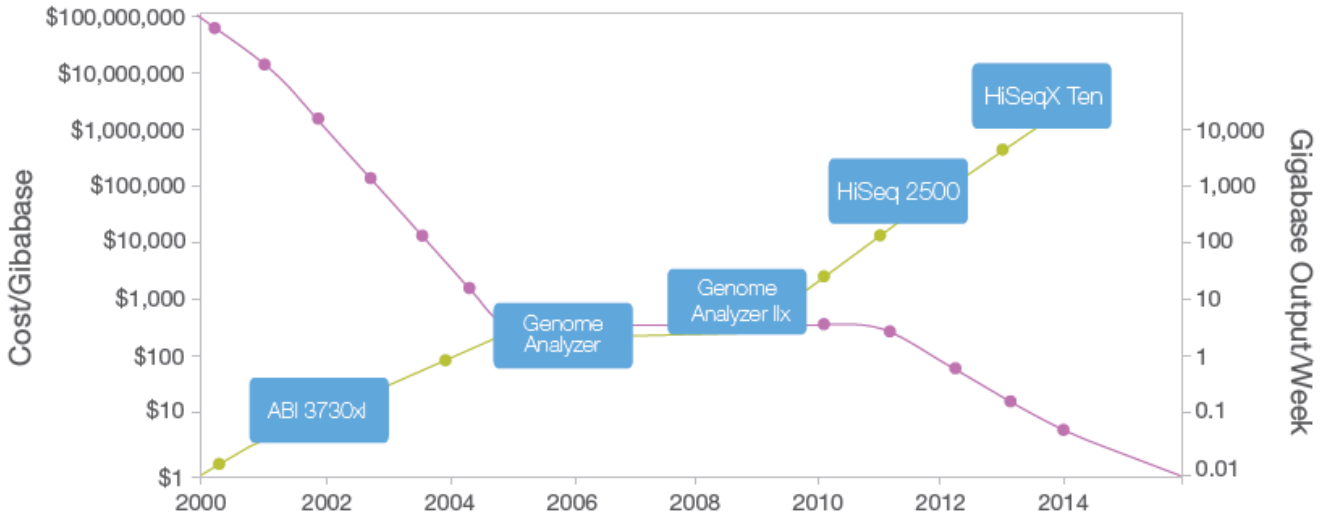
**HiSeq X Series**
Population- and production-scale human whole-genome sequencing.

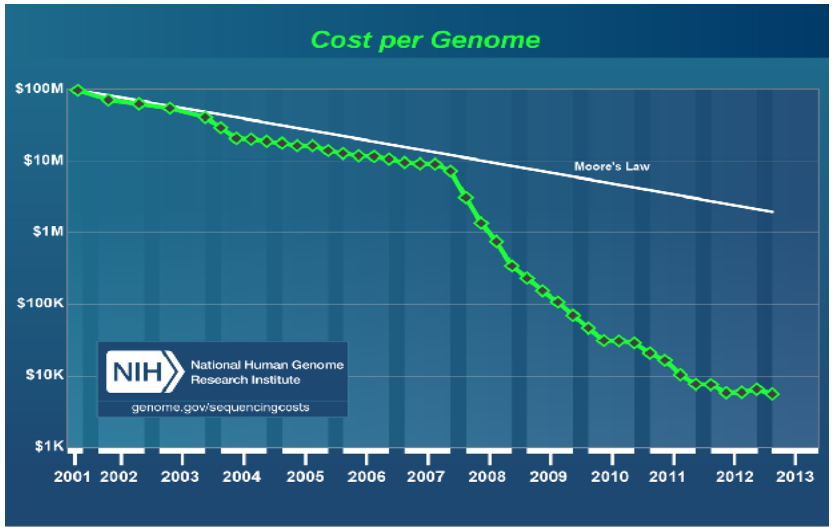
From Illumina.com

Adapted from http://www.illumina.com/technology/next-generation-

# Sequencing Cost and Data Output

## Cost and Ouput of Ilumina sequencing machines across the years




Adapted from http://www.illumina.com/technology/next-generation-sequencing.html

# Applications and Pipeline of DNA Sequencing



- Resequencing
  sequencing the genome of an organism with a known genome
- Exome sequencing / Targeted sequencing
  sequencing only selected regions from the genome
- de-novo sequencing
  sequencing the genome of an organism with a unknown genome (no reference sequence available for alignment).

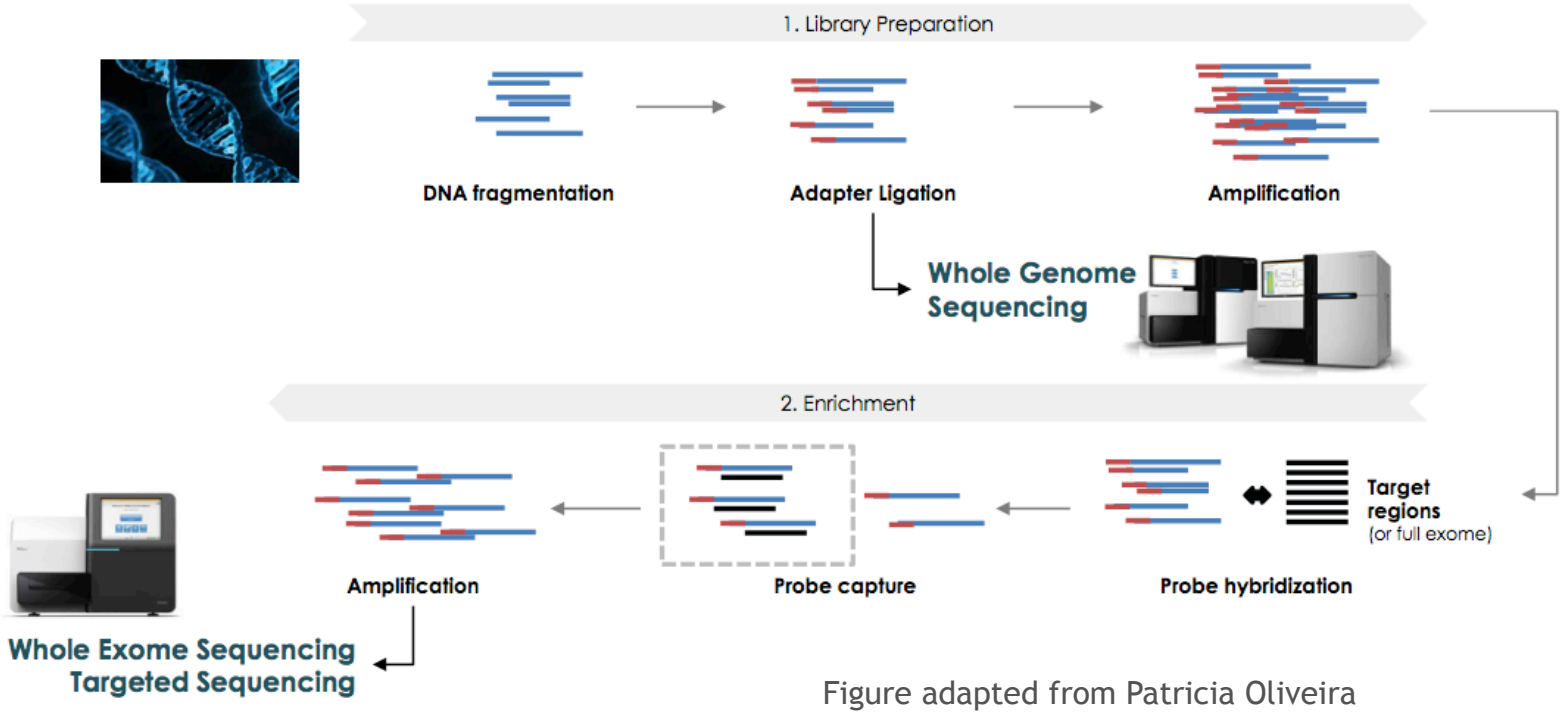
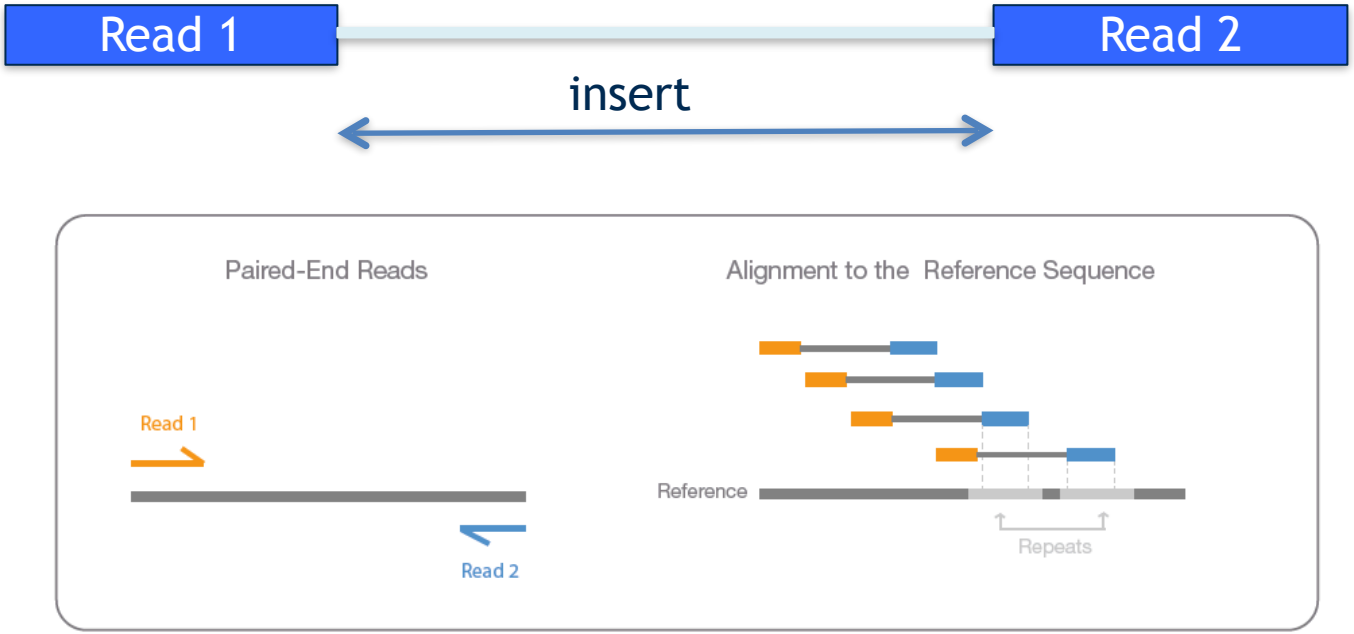



Figure adapted from Patricia Oliveira

# Sequence Reads (Concepts)

- **Insert** – the DNA fragment that is used for sequencing.
- **Read** – the part of the insert that is sequenced.
- **Single Read (SR)** – a sequencing procedure by which the insert is sequenced from one end only.
- **Paired End (PE)** – a sequencing procedure by which the insert is sequenced from both ends.
- **Insert size** - average distance between read pair mates

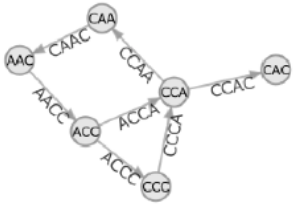



Adapted from Illumina.com

# *de novo* Assembly Analysis Pipeline



Tissue collection → Sample preparation → Sequencing → Assembly




~ 1 day

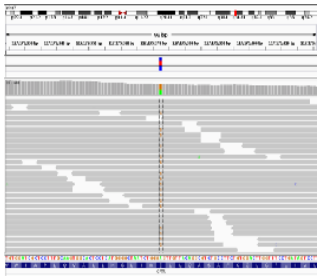
~ 1/2 days

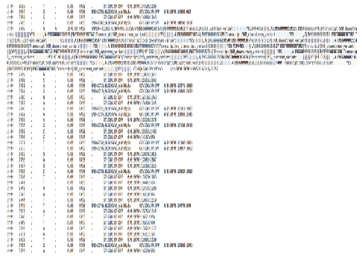
~ 1 week

# Variant Calling Analysis Pipeline

Bioinformatics
Pedro G. Ferreira

Tissue collection → Sample preparation → Sequencing → Mapping → Variant Calling





~ 1 day

~ 1/2 days

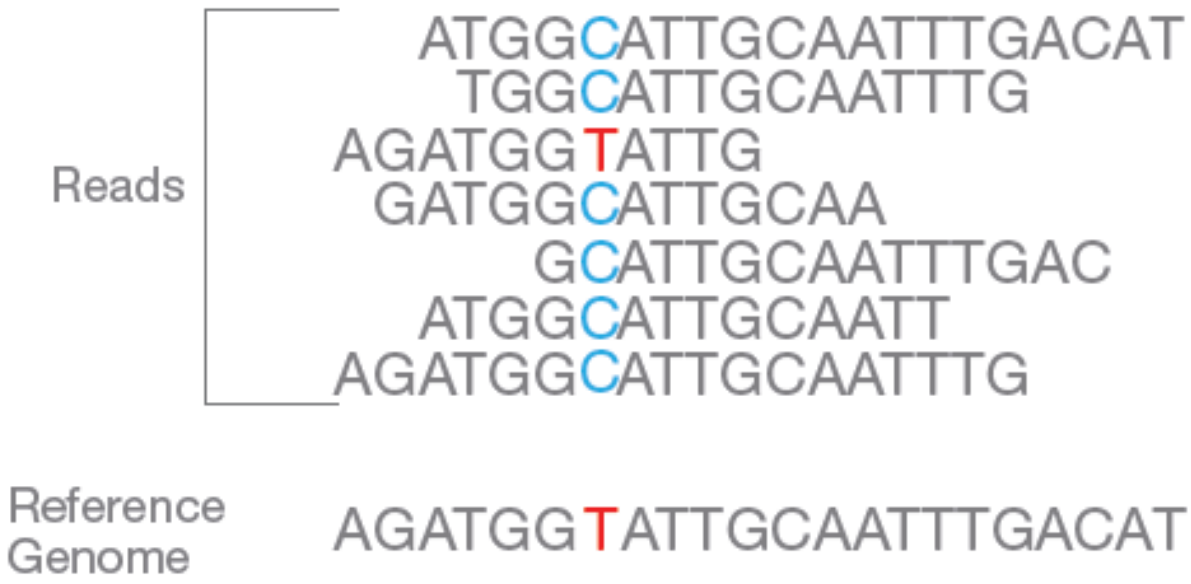
~ 2/5 days

~ 2 days

Several weeks

# Single Nucleotide Variant detection

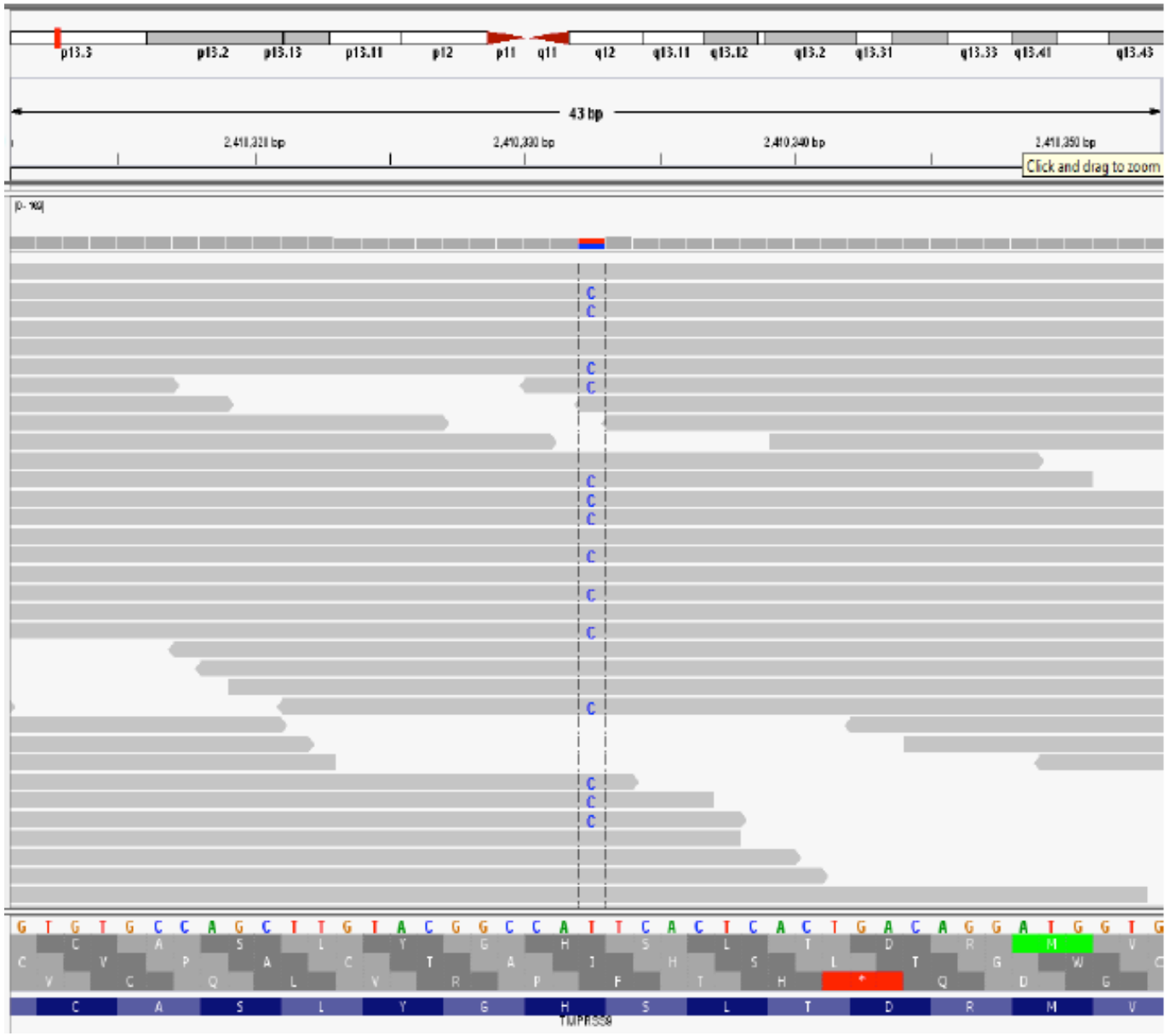




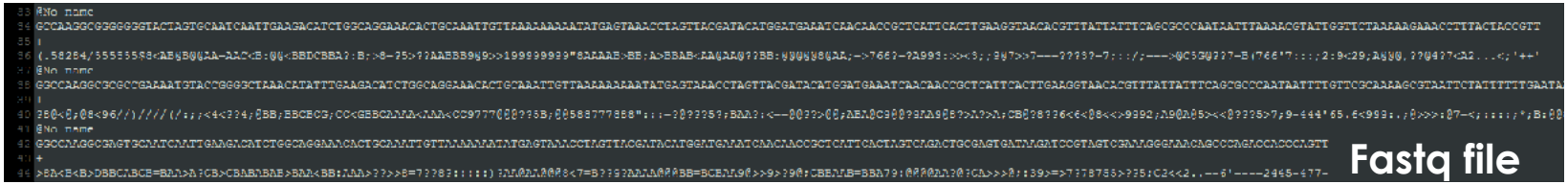

Adapted from Illumina.com

# Single Nucleotide Variant detection





# Alignment and Variant Calling (Bioinformatics Analysis)



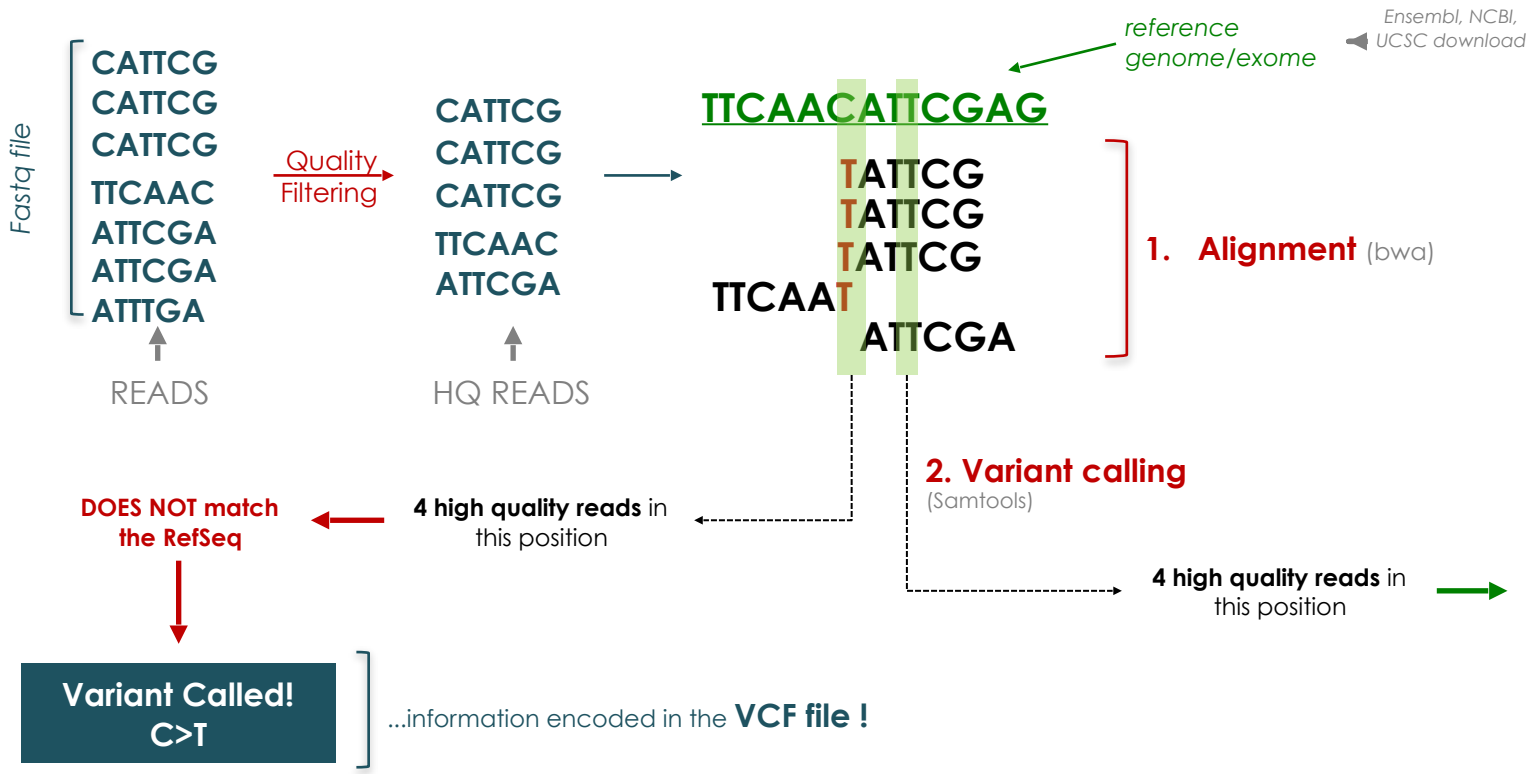
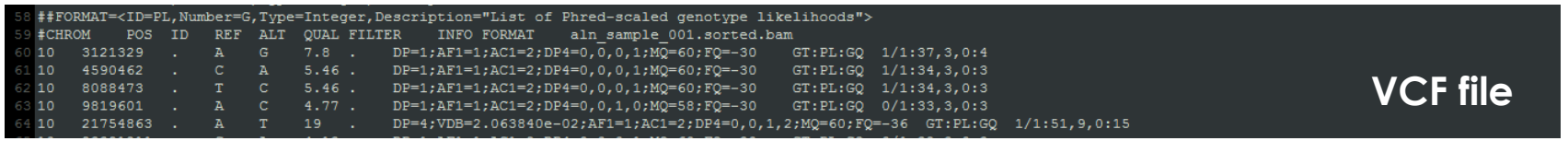




Figure adapted from Patricia Oliveira



# Variant Filtering based on Calling Quality

*VCF file version 4.1*

```
59 #CHROM    POS   ID   REF  ALT  QUAL  FILTER  INFO FORMAT    aln_sample_001.sorted.bam
60 10   3121329   .    A    G    7.8   .    DP=1;AF1=1;AC1=2;DP4=0,0,0,1;MQ=60;FQ=-30  GT:PL:GQ  1/1:37,3,0:4
```

## QUAL = Quality

**Definition:** Phred-scaled quality score for the assertion made for the alternative allele i.e. the probability that the call is wrong. Ranges from 1 to 225.

**Example:** QUAL=225, probability of error is $10^{-22.5}$

**High quality scores indicate high confidence calls**

| Phred Quality Score | Probability of incorrect base call | Base call Accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 30 | 1 in 1000 | 99.90% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 100.00% |

## INFO = Information

**Various information incluing depth, allele counts, etc**

*DP=1;AF1=1;AC1=2;DP4=0,0,0,1; MQ=60;FQ=-30*

## Genotype

*GT:PL:GQ* *1/1:37,3,0:4*

**GT:** Genotype (0/1, 1/1)

**PL:** Phred-scaled likelihood of genotype
0 is the most likely genotype

**GQ:** Phred-scaled quality of genotype
Ranges from 1-99

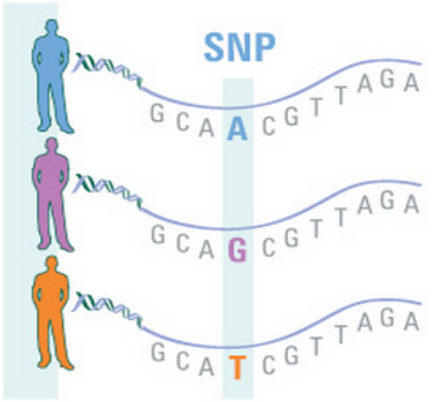
**High genotype quality scores indicate high confidence genotype calls**

**High confidence variant calls selected!**

Adapted from Patricia Oliveira

# DNA Sequencing Projects

Several and large scale projects are sequencing the DNA to **improve personalized medicine** and understand **population differences**:

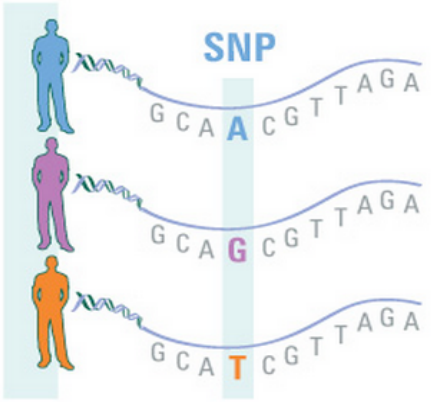



Adapted from www.broadinstitute.org

- The goal of the **1000 Genomes Project** is to find most genetic variants that have frequencies of at least 1% in the populations studied (2008).
- **UK10K**, Rare Genetic Variants in Health and Disease by studying and comparing the DNA of **4,000** people whose physical characteristics are well documented, and **6,000** people with extreme health problems (£10.5 million funding award, 2010-2013).
- Genomics England, a company wholly owned and funded by the Department of Health, will sequence **100,000** whole genomes from NHS patients by 2017.

# DNA Sequencing Projects

Several and large scale projects are sequencing the DNA to **improve personalized medicine** and understand **population differences**:




Adapted from www.broadinstitute.org

- The goal of the **1000 Genomes Project** is to find most genetic variants that have frequencies of at least 1% in the populations studied (2008).
- UK10K, Rare Genetic Variants in Health and Disease by studying and comparing the DNA of **4,000** people whose physical characteristics are well documented, and **6,000** people with extreme health problems (£10.5 million funding award, 2010-2013).
- **Genomics England**, a company wholly owned and funded by the Department of Health, will sequence **100,000** whole genomes from NHS patients by 2017.

# DNA Sequencing Projects

- **Saudi Human Genome Project**, read **~100,000** Saudi genomes representing normal and disease conditions.
- Obama Announces **$215m Precision Medicine Investment** for NIH, FDA (January 30, 2015).
- **deCODE** conducted whole genome sequencing on **2,636 individuals**—a little less than 1% of Iceland's population (March 25, 2015).
- The **Exome Aggregation Consortium (ExAC)** is seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects (**60,706** unrelated individuals).

# DNA Sequencing Projects

- **Saudi Human Genome Project**, read ~**100,000** Saudi genomes representing normal and disease conditions.
- Obama Announces **$215m Precision Medicine Investment** for NIH, FDA (January 30, 2015).
- **deCODE** conducted whole genome sequencing on **2,636 individuals**—a little less than 1% of Iceland's population (March 25, 2015).
- The **Exome Aggregation Consortium (ExAC)** is seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects (**60,706** unrelated individuals).

# Transcriptome Sequencing

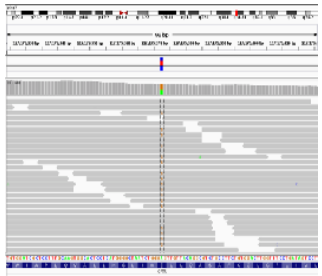
## Typical applications of transcriptome sequencing (RNA-seq) technologies:

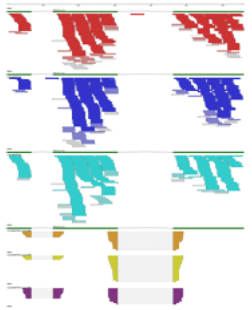
- Gene abundance quantification
  Compare 2 different experiments:
  - Drug treated vs untreated cell line;
  - Tumor vs Normal tissue;
- De novo transcript sequence reconstruction
  Species that lack a genome or have not been annotated yet/have poor annotation of gene models.
- Detection of alternative splicing events
  Each gene has multiple isoforms; Tissue type have alternative splicing, some isoforms are specific of a given tissue.
- Fusion transcripts detection
  Cancer samples may have aberrant gene expression with 2 genes fused that may not function properly.
- Allele Specific Expression
  If a gene expression and has a mutation, what is the isoform that has higher expression the one with the mutant allele or the wild type allele.

# Transcriptome Analysis Pipeline

Bioinformatics
Pedro G. Ferreira

Tissue collection → Sample preparation → Sequencing → Mapping → Expression And Splicing Quantification





~ 1 day

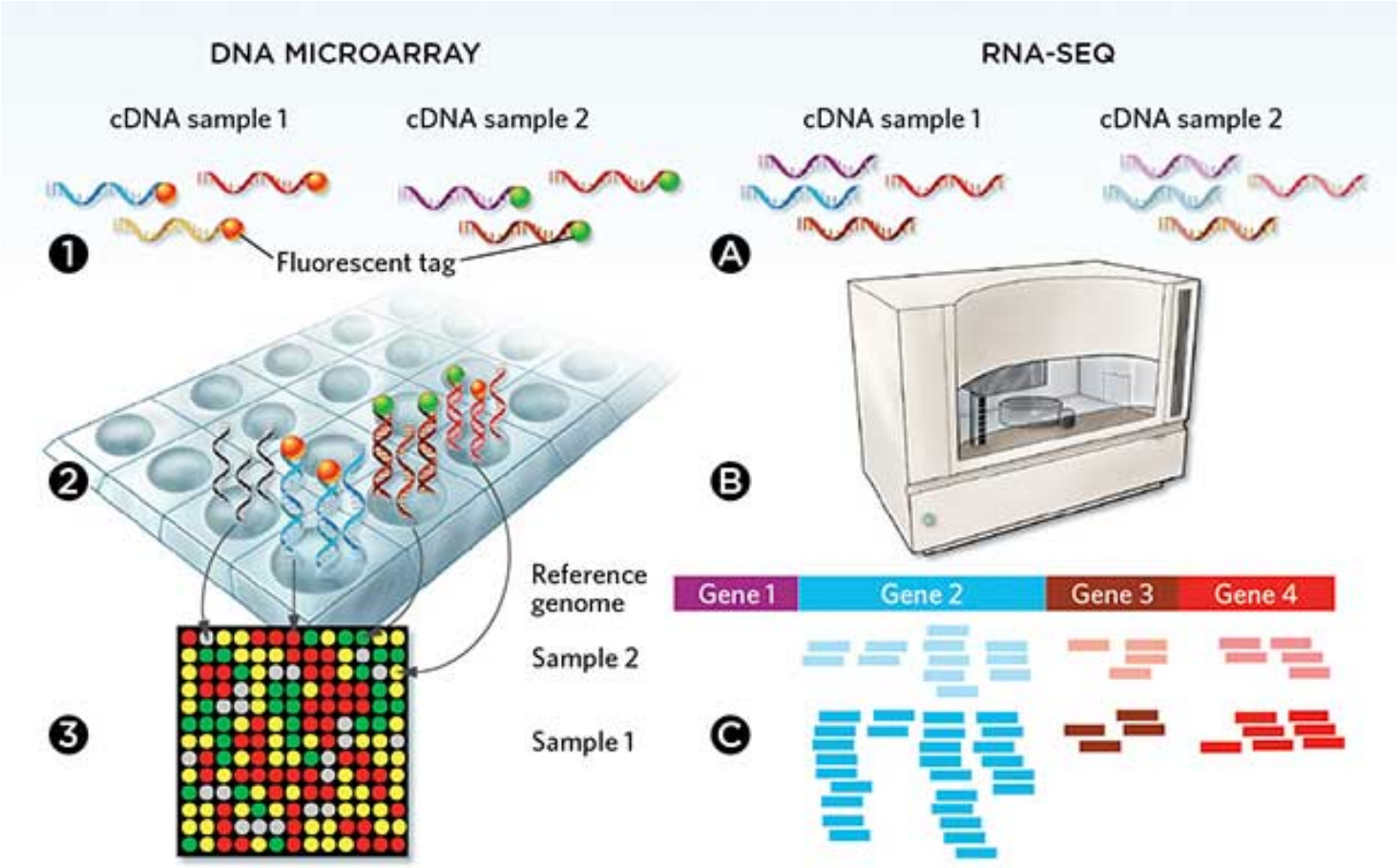
~ 1/2 days

~ 2/5 days

~ 2/4 days

Several weeks

# Hybridization VS Sequencing




*Adapted from: Kate Yandell. An Array of Options, A guide for how and when to transition from the microarray to RNA-seq. The Scientist, 2015*

# Hybridization VS Sequencing



## Analog signal *(microarrays)*

Continuous signal

Signal loss at low end and signal saturation at the high end

Requires *a priori* knowledge of the sequence of the sample

Less expensive

Methods are fully mature and straightforward

## Digital signal *(sequencing)*

Read Counts: discrete values

Weak background noise

Unbiased detection of novel transcripts

Broader dynamic range

Increased specificity and sensitivity

Easier detection of rare and low-abundance transcripts

Does not require *a priori* knowledge of the sequence of the sample *(non-model species)*

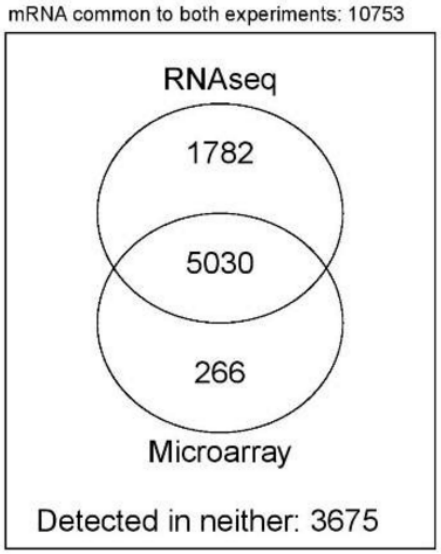
More Expensive

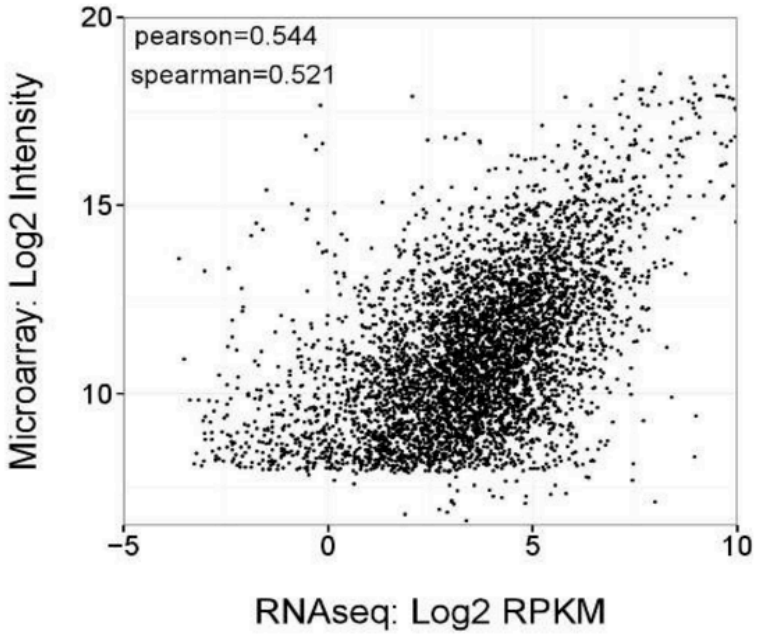
Bioinformatics pipelines not fully established yet *(no consensus)*, although quite ready advanced

# Hybridization VS Sequencing

## Comparison of Gencode 15 LncRNA microarray with RNAseq gene quantification

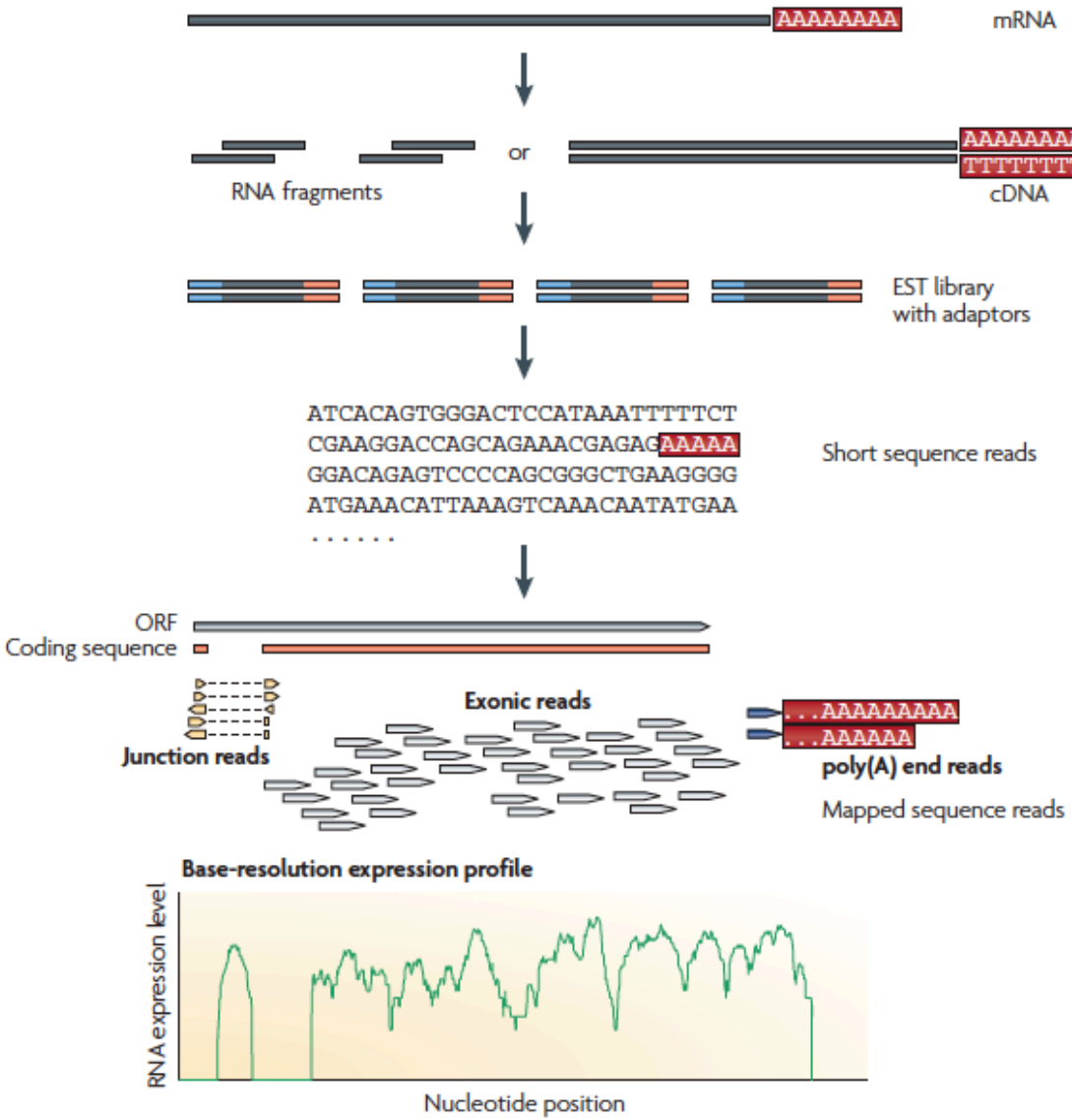
*For RNAseq data, genes or transcripts are considered present if RPKM mean > 0 and IDR < 0.1*



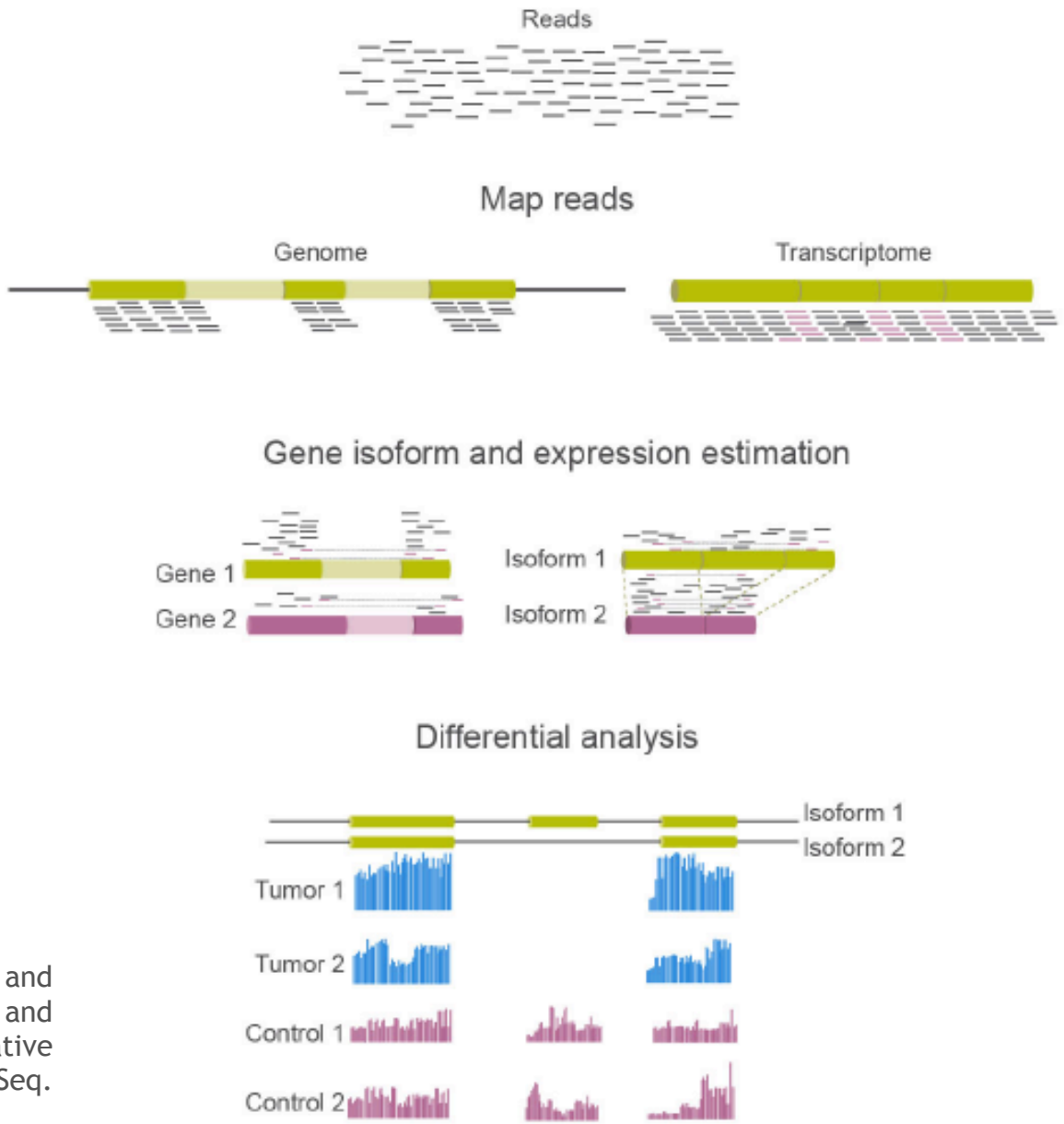

*Adapted from: GENCODE Custom lncRNA Expression Microarray Design, http://www.gencodegenes.org/custom_array.html*
*Derrien et al., The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012*

# RNA-Sequencing Experiment



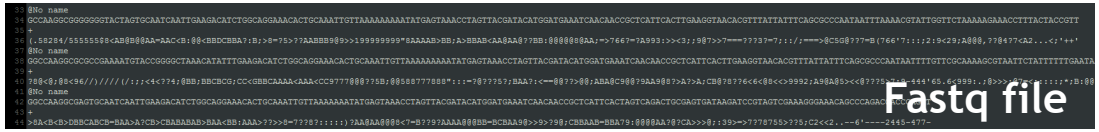

*Adapted from: Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009*

# Differential Expression Analysis




Adapted from Feng H., Qin Z. and Zhang X. (2012) Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. Cancer Letters

# Normalization of Expression values




Fastq file

CATTCG
CATTCG
CATTCG
TTCAAC
ATTCGA
ATTCGA
ATTTGA

READS

Quality Filtering

CATTCG
CATTCG
CATTCG
TTCAAC
ATTCGA

HQ READS

TTCAACATTCGAG

CATTCG
CATTCG
CATTCG
TTCAAC
ATTCGA

*reference transcriptome*

*Ensembl, NCBI, UCSC download*

**1. ALIGNMENT**

**2. COUNT READS**

**5 high quality reads in this transcribed region**

**3. NORMALIZATION**

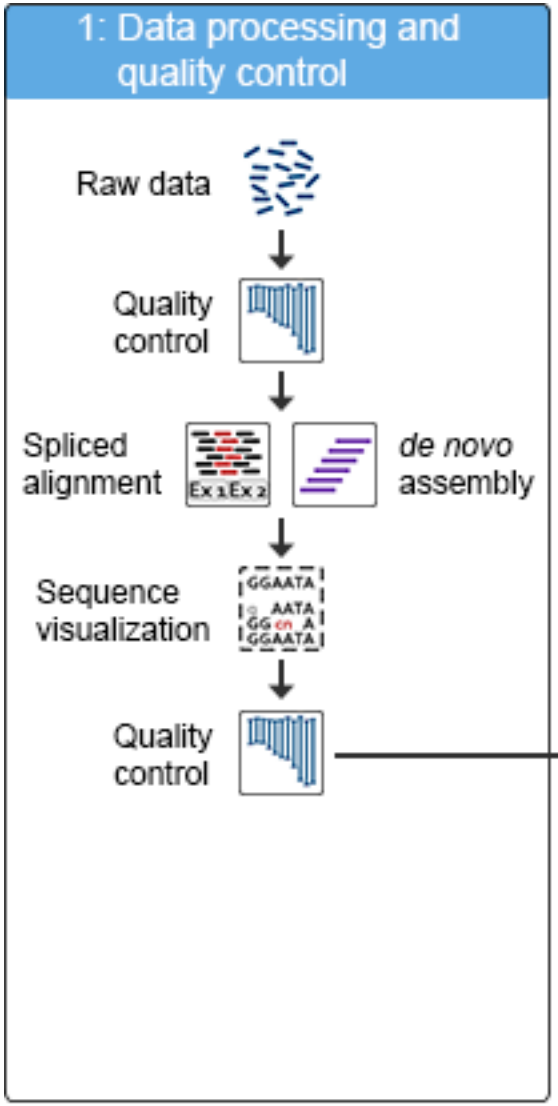
**length of transcript** *and/or* **length of gene**

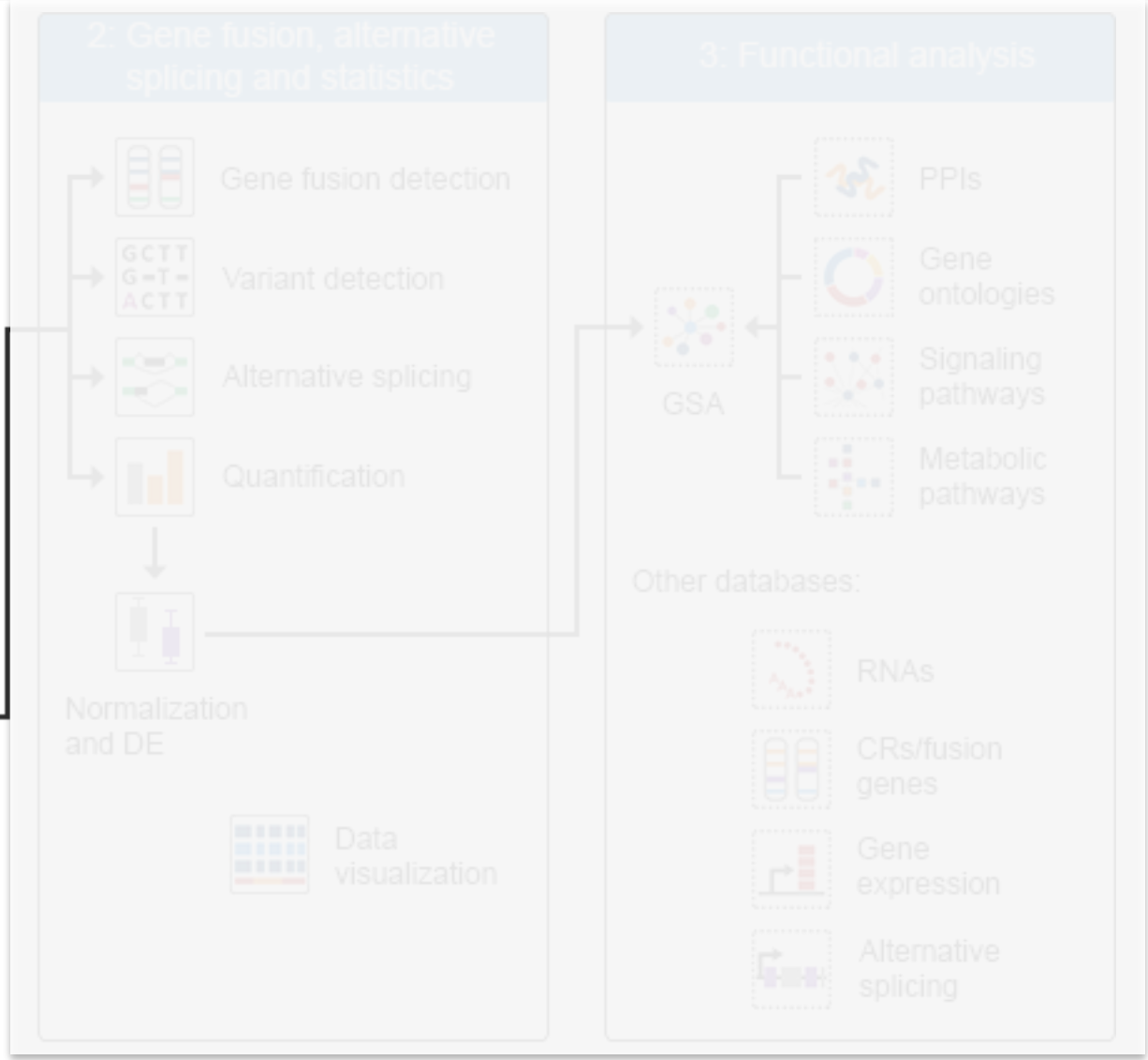
**total number of reads per sample** *and/or* **total number of reads across samples**

$$\text{RPKM or FPKM} = \frac{10^9 \times \text{Reads in the Transcript}}{\text{Total number of Reads in the Sample} \times \text{Number of bp in the Transcript}}$$

Adapted from Patricia

# RNA-Sequencing Experiment

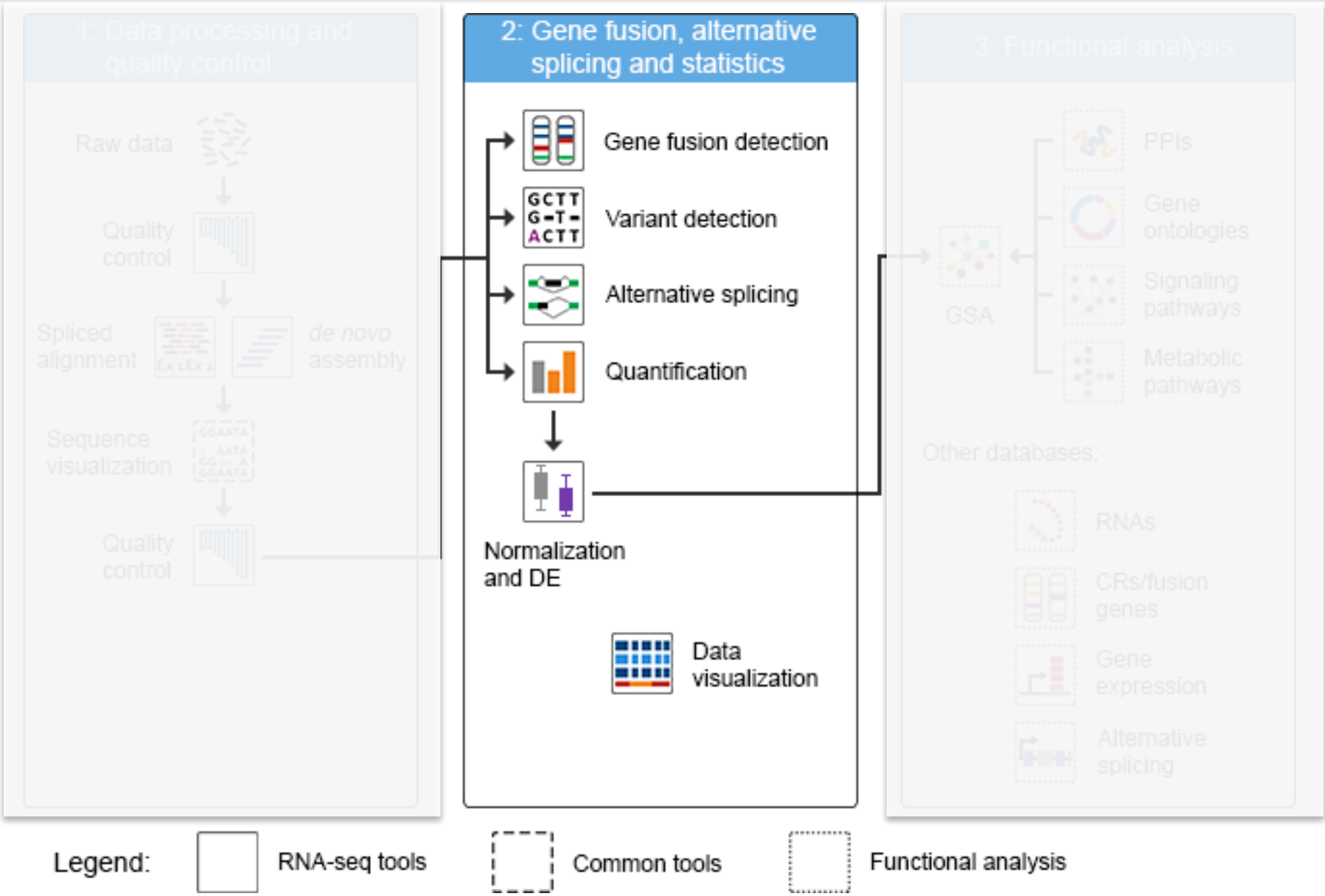



Legend: RNA-seq tools | Common tools | Functional analysis

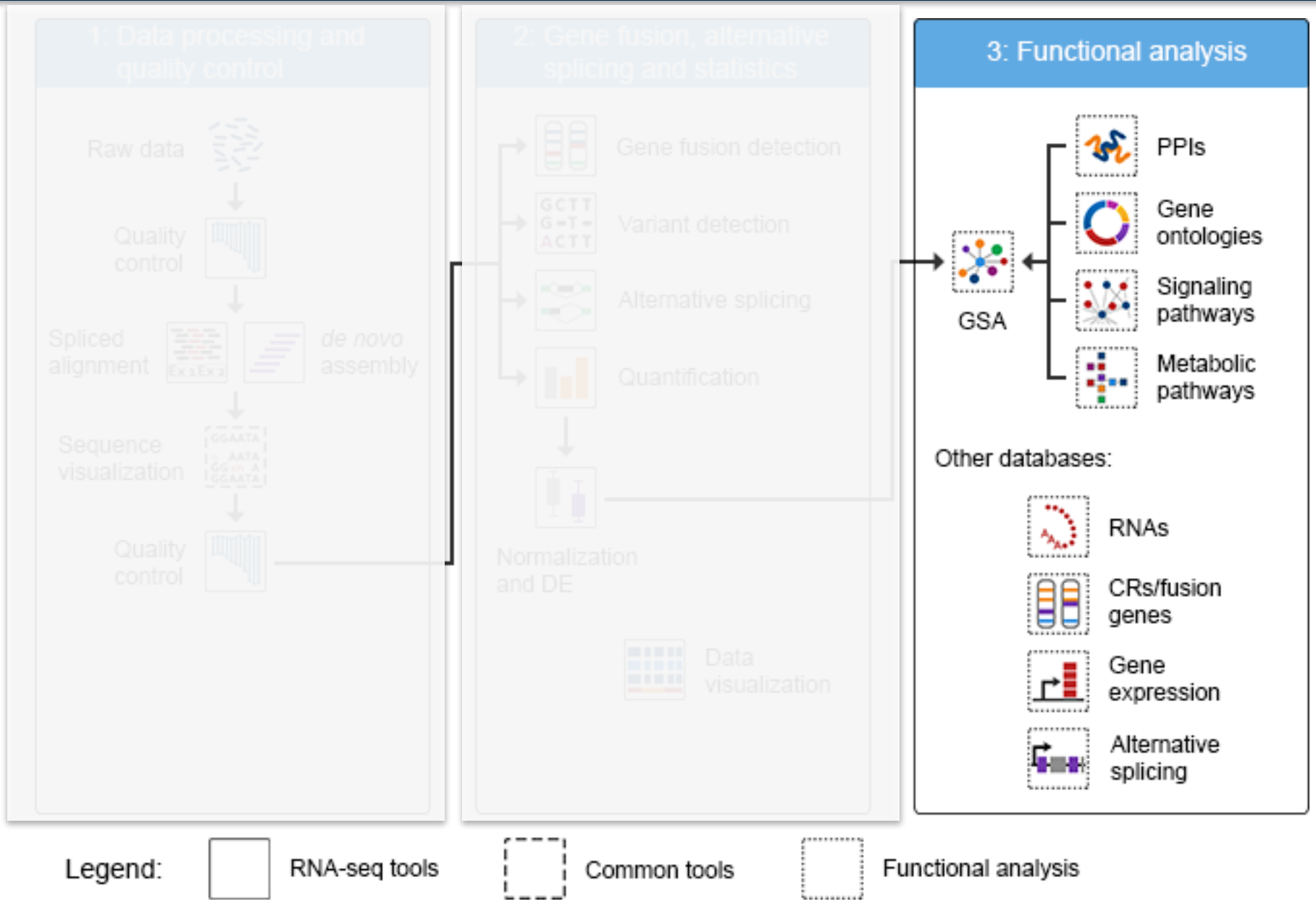
*Adapted from: omictools.com*

# RNA-Sequencing Experiment




*Adapted from: omictools.com*

# RNA-Sequencing Experiment




*Adapted from: omictools.com*

# Transcriptome Sequencing



*Transcriptome and genome sequencing uncovers functional variation in humans* by Lappalainen et. al., Nature 2013.

- Large scale RNA-seq and microRNA-seq of about 462 individuals (part of the 1000 Genomes project);
- Highly densed genotype (> 5Million SNPs);
- The 462 individuals covering five populations: the CEPH, Finns, British, Toscani and Yoruba have both genome and RNA sequencing data;

*Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals*, by Battle et. al. Genome Research 2013

- 922 individuals are of European descent: the largest transcriptomic study using RNA-seq;
- Analyzed how genetic variations influence gene expression variations in both Cis and Trans;

*Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics*. Khurana, Fu et al Science 2013

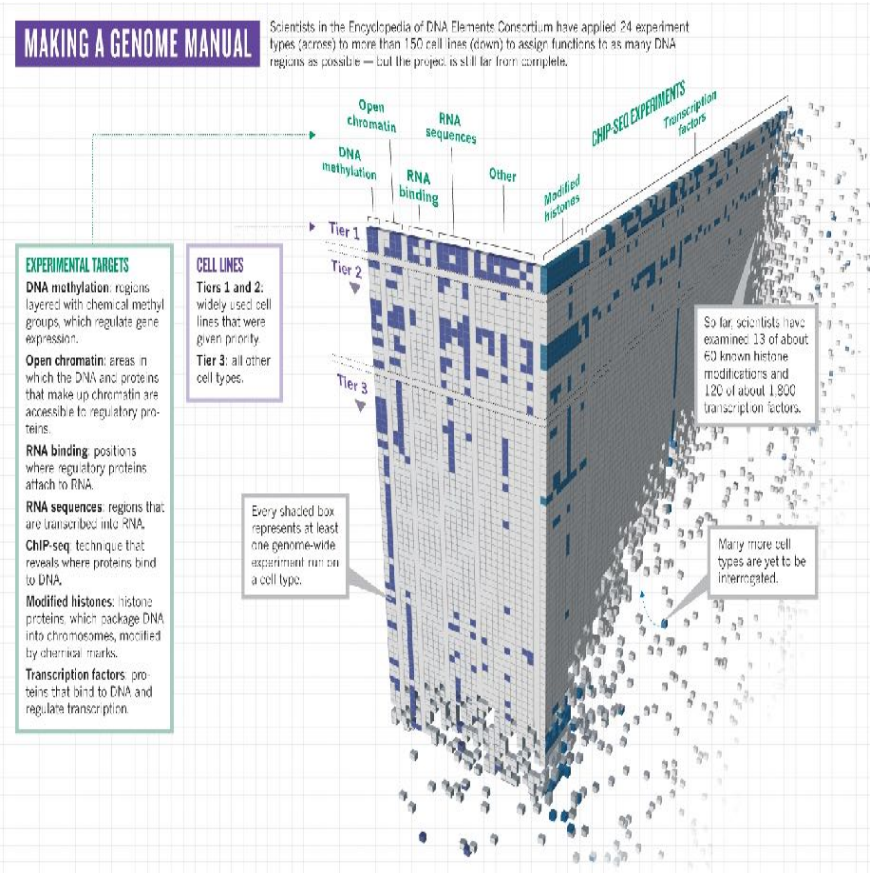
- 1092 individuals in the 1000 Genomes Project (Phase 1), genetically characterized: SNVs, indels, SVs;
- Patterns of selection in DNA elements from the ENCODE project;

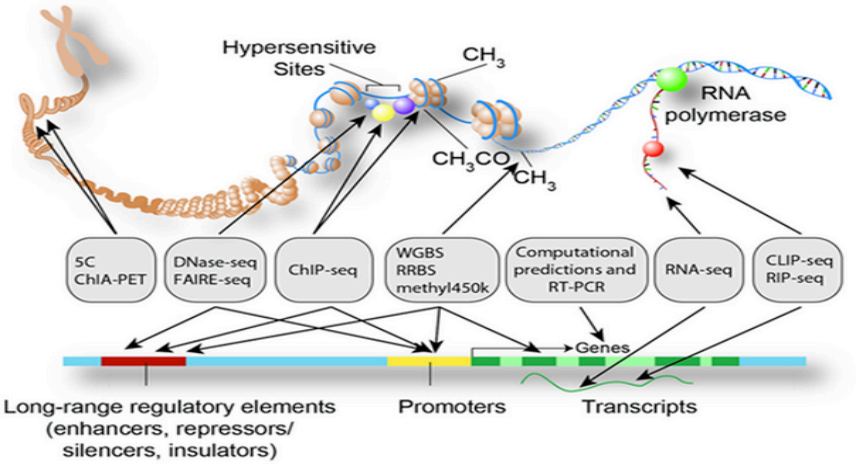
*The Genotype-Tissue Expression (GTEx) project*. Londsdale et al, Nature Genetics 2013

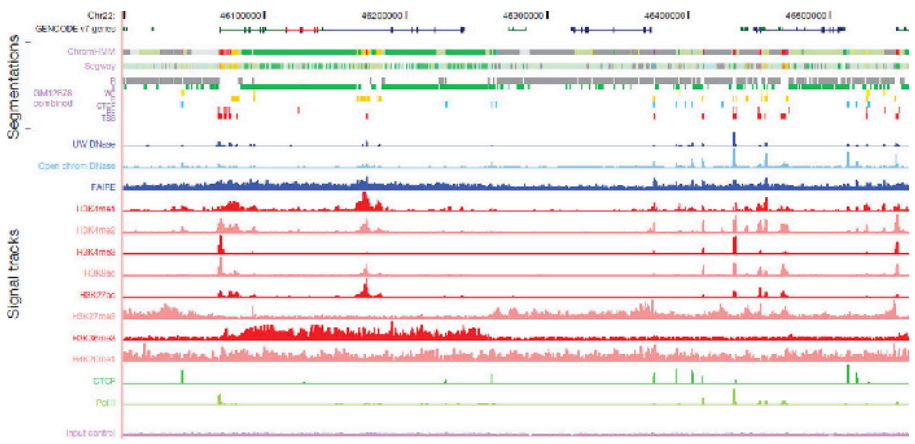
- Possibly the largest resource of genotype and gene expression: ~ 50 tissues in the human body, including the brain, muscle, heart, liver, kidney;
- ~ 1000 post-mortem donors;
- ~ 20K samples RNA-sequenced;

# Gene Regulation Analysis

- The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the NHGRI.
- The goal of ENCODE is to build a comprehensive list of functional elements in the human genome.




Adapted from www.nature.com/encode/

# Large Scale Cancer Projects



The Cancer Genome Atlas (**TCGA**), a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), aims to generate comprehensive, multi-dimensional maps of the key genomic changes in major types and subtypes of cancer.

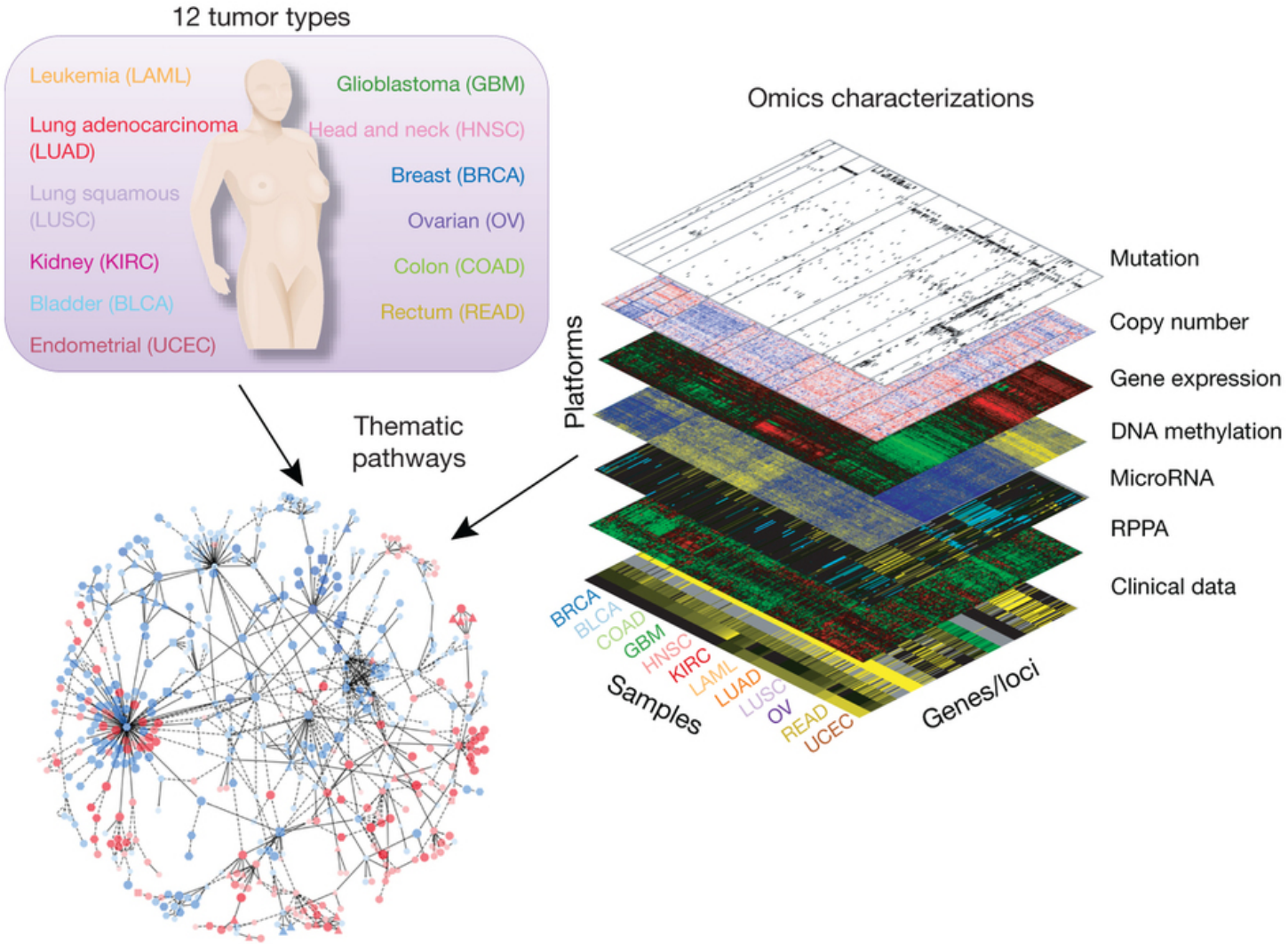
Projects with publicly available data: *Breast*, *Central Nervous System*, *Endocrine*, *Gastrointestinal*, *Gynecologic*, *Head and Neck*, *Hematologic*, *Skin* ,*Soft Tissue*, *Thoracic*, *Urologic*

**ICGC** Goal: To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.

Projects that are currently funded are examining tumors affecting: *the biliary tract*, *bladder*, *blood*, *bone*, *brain*, *breast*, *cervix*, *colon*, *eye*, *head and neck*, *kidney*, *liver*, *lung*, *nasopharynx*, *oral cavity*, *ovary*, *pancreas*, *prostate*, *rectum*, *skin*, *soft tissues*, *stomach*, *thyroid and uterus*.

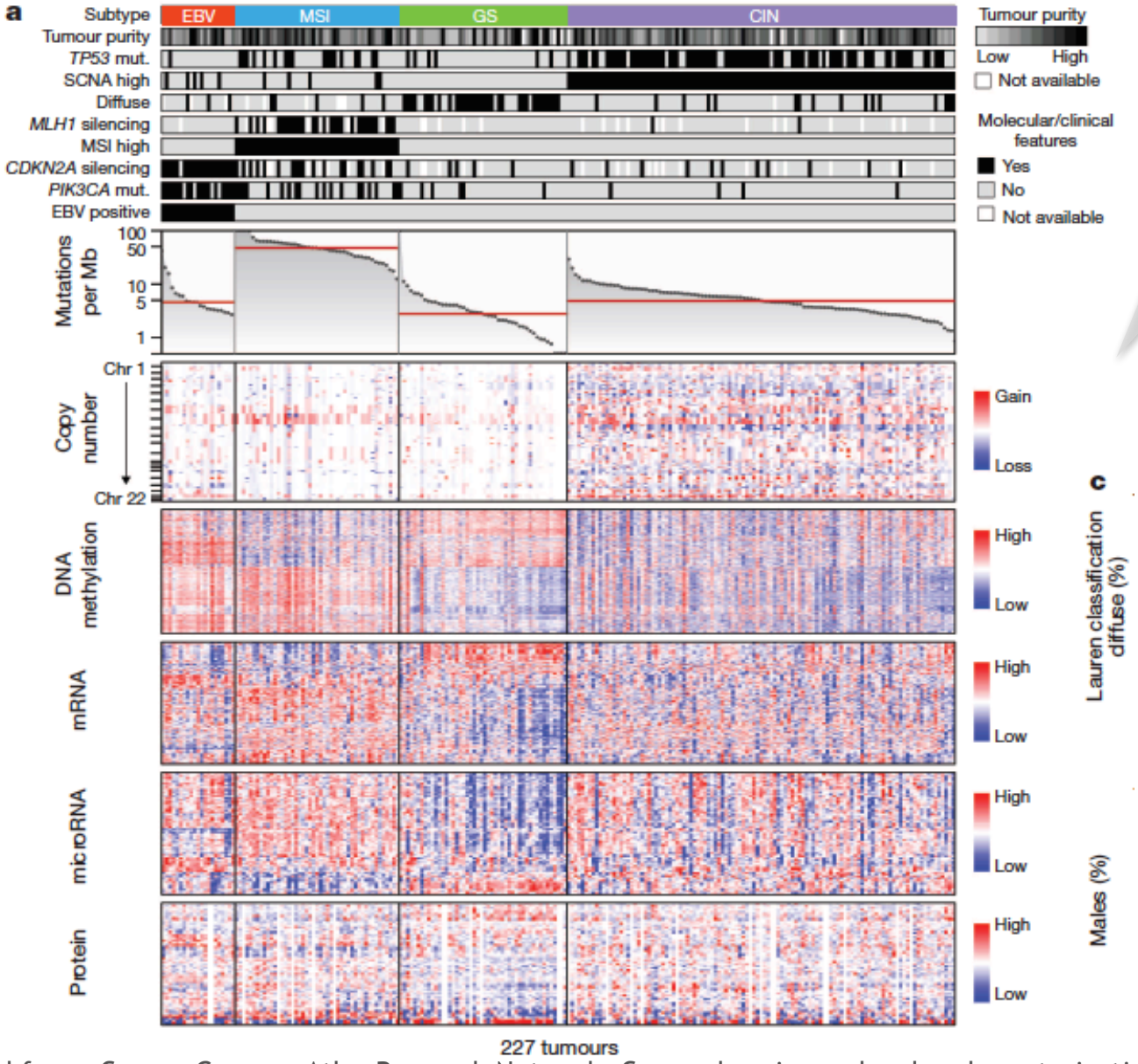
References:
http://cancergenome.nih.gov/abouttcga/overview
https://icgc.org/

# The Cancer Genome Atlas Multi-Dimensional data




"The Cancer Genome Atlas Pan-Cancer analysis project", Nature Genetics 45, 2013.

# Challenge: Integrate all this data

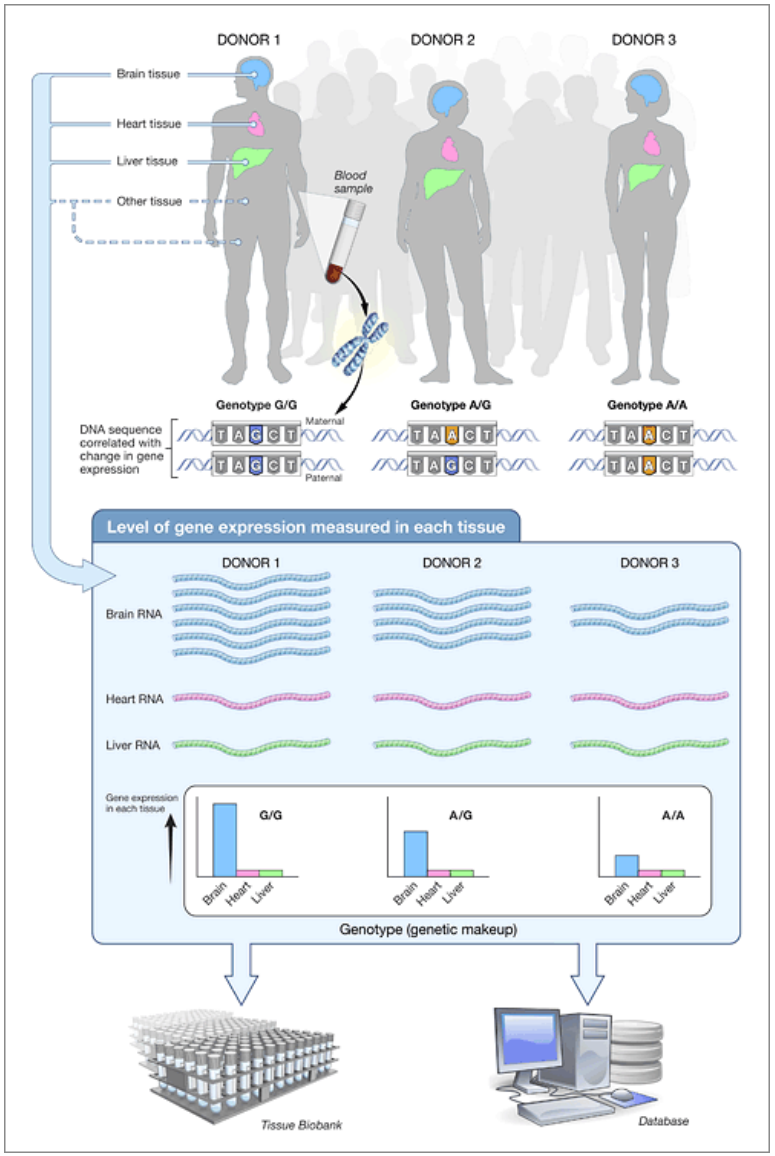



- Molecular subtypes of gastric cancer

Adapted from: Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014

# GTEx: Genotype-Tissue Expression






Adapted from https://commonfund.nih.gov/gtex/data

- Collect and analyze multiple human tissues from donors who are also densely genotyped.
- Create a resource to study expression and regulation and its relationship to genetic variation and other molecular and medical phenotypes.
- Pilot phase (2015): RNA-seq from 1641 samples
  - 175 individuals
  - 43 body sites

  29 solid organ tissues, 11 brain sub-regions, whole blood, and 2 cell lines (LCL, Fibroblasts).
- Middle phase (2016): 8555 RNA-seq samples
  - 550 individuals.
- Final phase (2017): ~20K RNA-seq samples
  - ~900 individuals.

# NGS has very high computational demands



## NGS pushes bioinformatics limits

- Very large text files are generated per experiment. With the current technology these files easily reach 50 million lines.
  - Familiar tools in python and perl will not handle the processing of these files;
  - Impossible to fit in main memory and high demand of execution time;

## NGS projects require supercomputing infrastructure

- Some genomic studies are very data intensive.
- Scenario: whole genome sequencing (WGS) of Tumor and Normal pairs.
  - 30X coverage of a genome pair (T/N) ~ 500 Gb;
  - 50 genome pairs ~25 Tb;

# NGS has very high computational demands



## NGS pushes bioinformatics limits

- Very large text files are generated per experiment. With the current technology these files easily reach 50 million lines.
  - Familiar tools in python and perl will not handle the processing of these files;
  - Impossible to fit in main memory and high demand of execution time;

## NGS projects require supercomputing infrastructure

- Some genomic studies are very data intensive.
- Scenario: whole genome sequencing (WGS) of Tumor and Normal pairs.
  - 30X coverage of a genome pair (T/N) ~ 500 Gb;
  - 50 genome pairs ~25 Tb;

# Exercises



| gene | Len | Reads | RPKM |
|---|---|---|---|
| A | 2 | 20 | |
| B | 3 | 10 | |
| C | 5 | 50 | |
| D | 10 | 100 | |

- Calculate RPKM for each gene.

**Reads Per KB per Million Mapped Reads**

**$RPKM(X) = 10^9 . C / N . L$**

Check out the video:

https://statquest.org/2015/07/09/rpkm-fpkm-and-tpm-clearly-explained/

# Exercises



- Develop a script that reads as input the file gene_exp.txt and calculates the RPKM values for each gene on the two conditions. Choose a constant value so the RPKM values are not very low.
- How many genes have a normalised expression greater in condition 1 than condition 2?

# Exercises



```
CGACGACGACGACGAATGATGTATTATCGAGCGAGCGGCAGATGCTA
----------------------------------------------
CGACGACGACGACGAATGAT    TATCGAGCGCGCGGCAGATG
 GACGACGACGACGAATGATG       CGAGCGCGCGGCAGATGCTA
 GACGACGACGACGAACGATG       CGAGCGCGCGGCAGATGCTA
         CGACGAACGATGTATTATCG
            CGAACGATGTATTATCGAGC
                  TGTATTATCGAGCGCGCGGC
```

- Identify the candidate mutations in this sequencing read pile-up. Report how many reads support each of the alleles (alternative and reference).
- Develop a script that given the input file identifies the mutations.