Next-Generation Sequencing in practice

Bioinformatics analysis techniques and some medical applications

Salvatore Alaimo, MSc.

Email: alaimos@dmi.unict.it

Overview

- Next Generation Sequencing:
  - how it works and why we need clever bioinformatics techniques
- NGS bioinformatics pipelines:
  - Pre-processing NGS data
  - Genomic Variants Extraction
  - RNA Expression Extraction
  - Identifying RNA Editing Events
- NGS for Cancer Biomarker Discovery
- The MedSeq Project

Next Generation Sequencing

How it works and Why we need clever bioinformatics techniques

Introduction

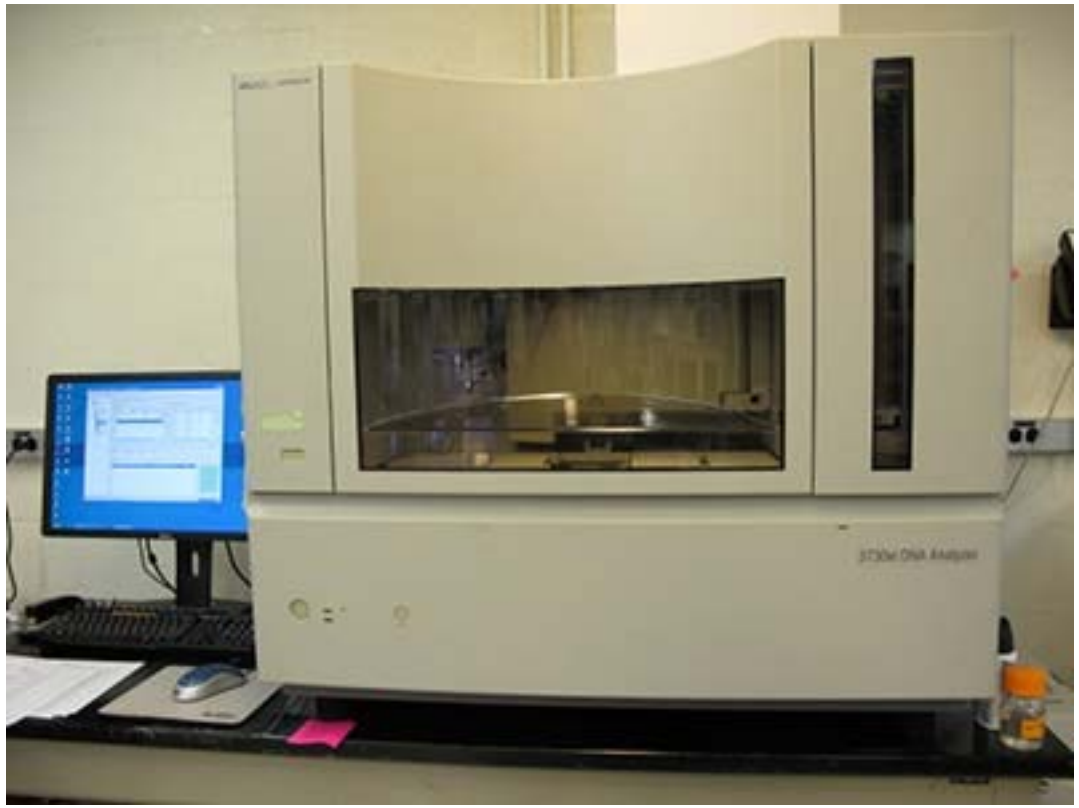
- High-throughput sequencing technologies have revolutionized our approach to omics studies.
- Example:
  - Conventional Microarrays:
    - requires prior knowledge of DNA sequence (reference genome sequence essential to build probes)
    - limited to genomic regions which are targeted by the probes.

    ➢conventional gene expression studies: no gene expression levels unless probes are available

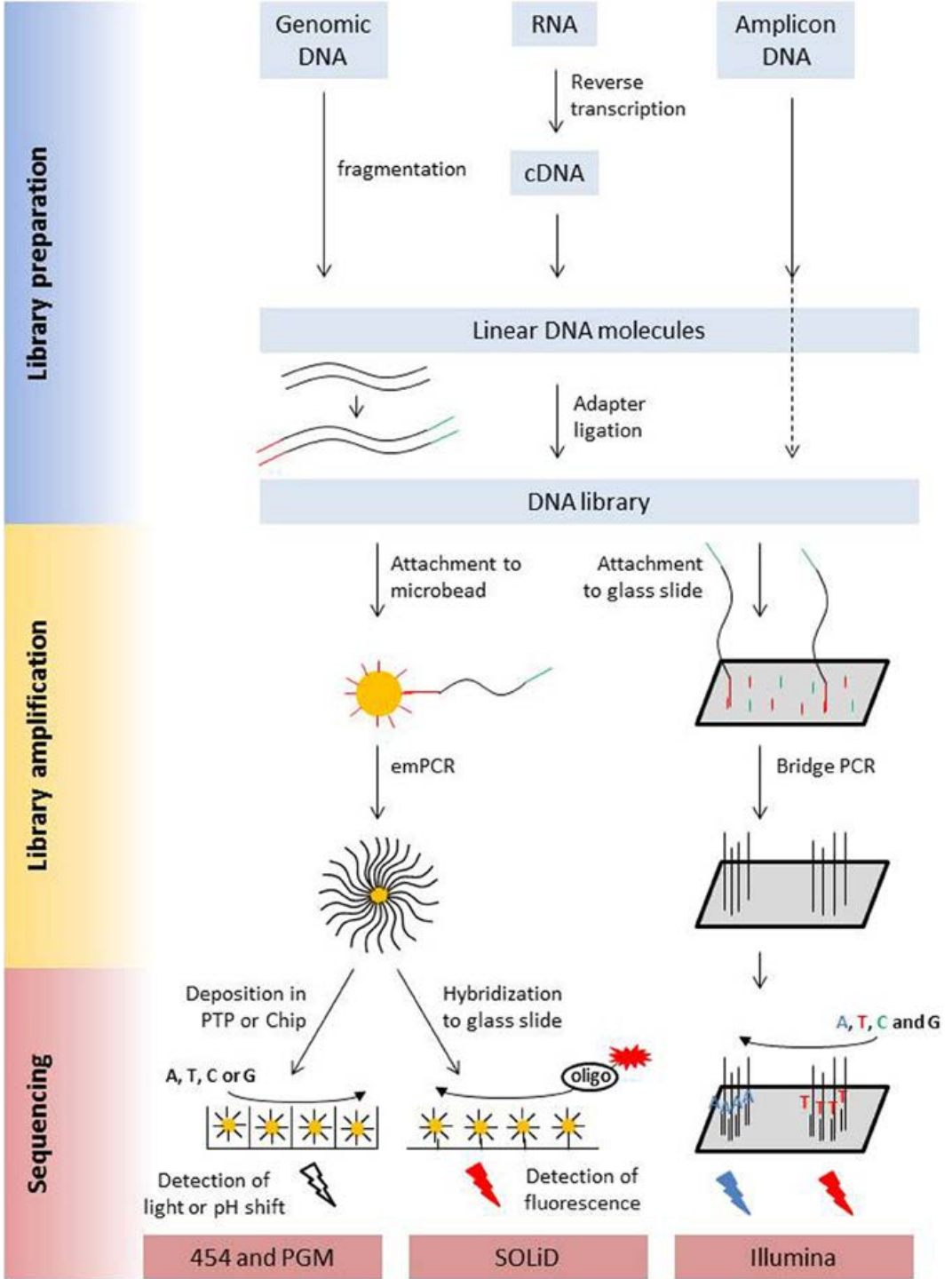
    ➢unknown transcript cannot be assessed.
- High-throughput sequencing approaches are able to capture all the DNA fragments and all the transcripts (coding and noncoding)
  - including low abundance ones…if the sequencing depth is sufficient.

From Sanger Sequencing to NGS Technologies

- Sanger sequencing
  - most widely used over the last three decades
  - invented in the late 1970s
  - used for various applications
  - still considered gold standard
    - commonly used as validation
  - sequence length of up to 1 kb.
- Major limitations:
  - ➢ low throughput (<100 kb)
  - ➢ high cost
    - ❖ Human Genome Project
      13 years and **$3 billion dollars**



Next-Generation Sequencing Technology

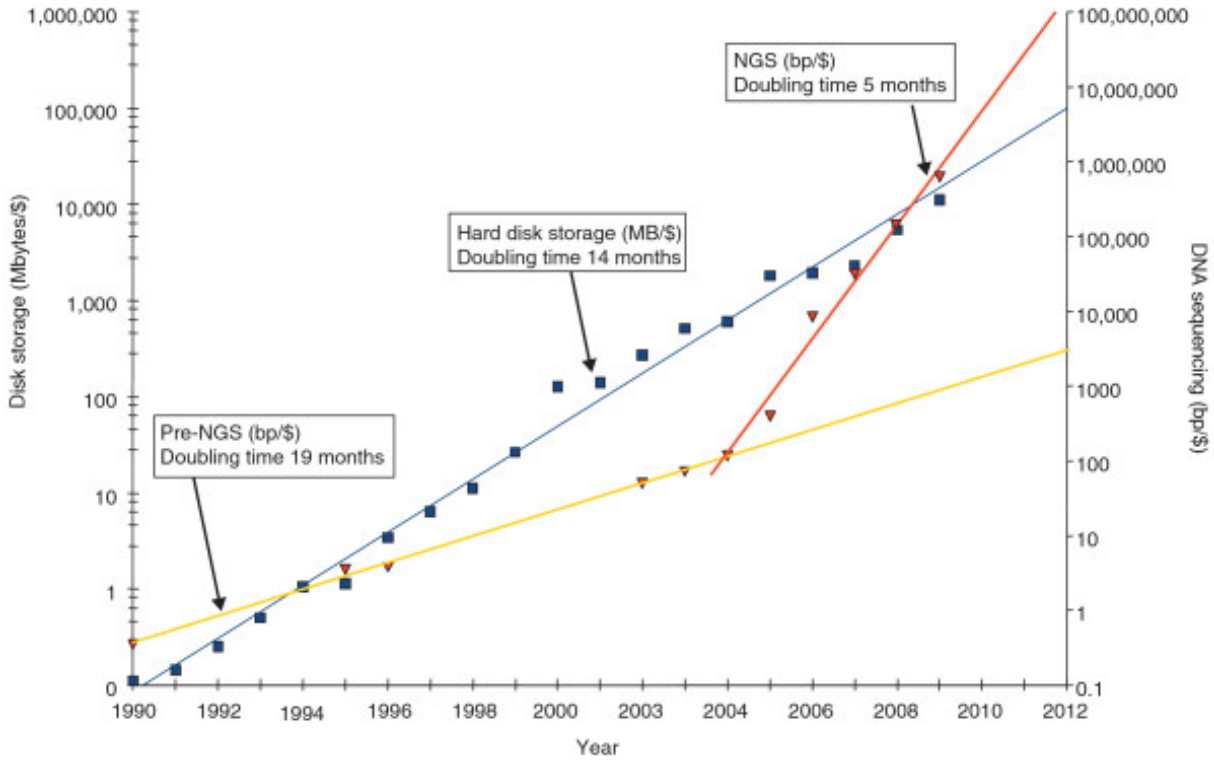


Knief C (2014) **Analysis of plant microbe interactions in the era of next generation sequencing technologies**. *Front. Plant Sci.* 5:216. doi: 10.3389/fpls.2014.00216

Future of NGS: 1000$ genome

- Storage:
  - Mb per $
  - ❖exponential curve
  - ❖doubling time ~1.5 years

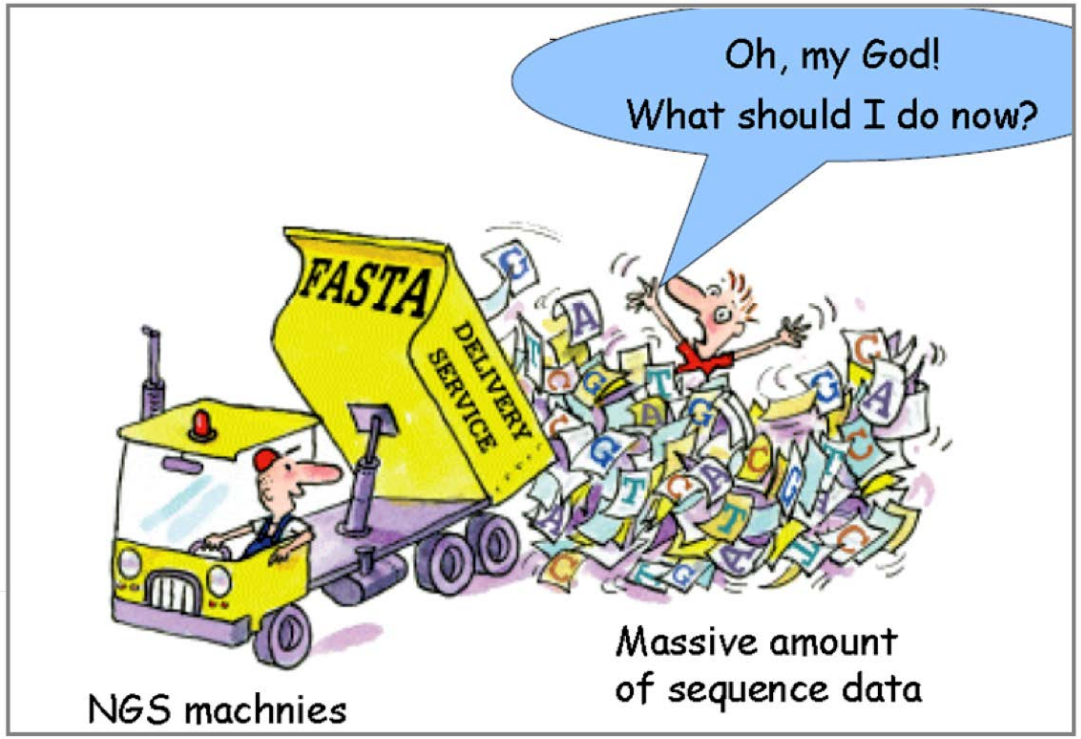
- DNA sequencing:
  - Base Pair per $
  - ❖exponential curve
  - ➢slow growth before 2004 (yellow line)
  - ➢with NGS doubling time of less than 6 months (red line)



Baker, Monya. "Next-generation sequencing: adjusting to data overload." *nature methods* 7.7 (2010): 495-499.

Bioinformatics Challenges

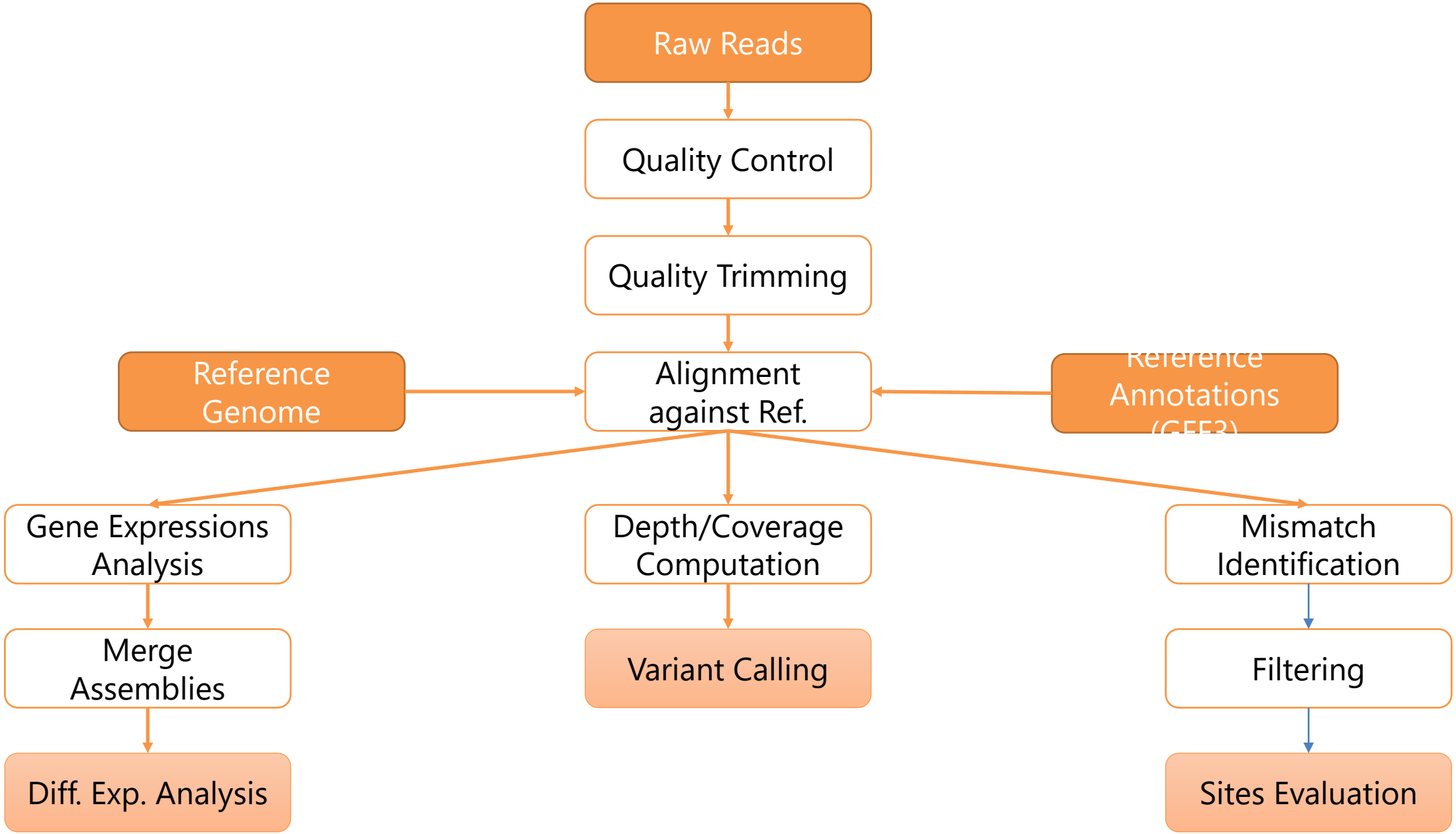
- NGS pushes bioinformatics needs up
  - Need for large amount of CPU power
    - Large compute clusters, Parallelization, …
  - VERY large text files (~10 million lines long)
    - New Tools, Memory Challenges, Browsing, …
  - Need sequence Quality filtering
  - Raw data are large
    - ➢…processed data are manageable for most people
    - ➢sometimes processing = loosing information
- In NGS we have to process really big amounts of data, which is not trivial in computing terms.
- Big NGS projects require supercomputing infrastructures



Bioinformatics Pipelines

Pre-processing of NGS Data

NGS Analysis Workflow



The FASTQ Format

- FASTQ format is a text-based format for storing both a biological sequence and its corresponding quality scores.

  ➢Both the sequence letter and quality score are each encoded with a single ASCII character for brevity.

  ➢Originally developed at the Wellcome Trust Sanger Institute.

  ✓Now the de facto standard for storing the output of high-throughput sequencing instruments

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

The FASTQ Format

- A FASTQ file normally uses four lines per sequence:
  - **Line 1** begins with a '@' character and is followed by a sequence identifier and an *optional* description.
  - **Line 2** is the raw sequence letters.
  - **Line 3** begins with a '+' character and is *optionally* followed by the same sequence identifier.
  - **Line 4** encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.
    - The character '!' represents the lowest quality while '~' is the highest.
    - The characters are ordered following ASCII codes.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

The FASTQ Format – Quality Score

- There are two methods to quantify Q.
  - Phred33 and Phred64 which consist of assigning a numerical value based on the ascii code to which must be added 33 (or 64).

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%

The GFF3 Format

- GFF3 (General Feature Format 3) is a simple tab delimited format for describing genomic features.

  - allows multi-level grouping and multi-level descriptive attributes.

  - is both more powerful and more restrictive than other GFF formats.

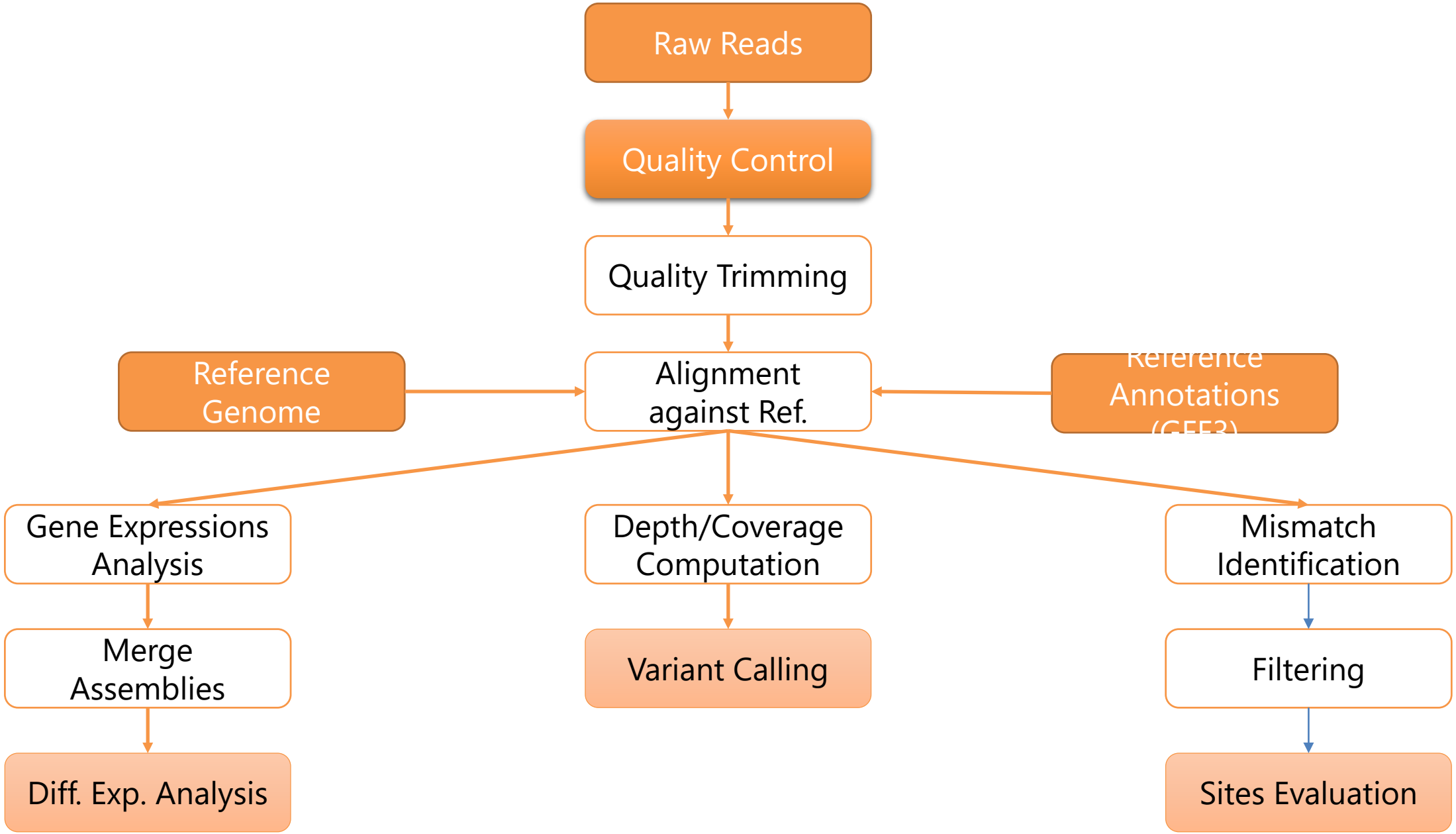
The GFF3 Format

- A GFF file consists of one or more records, each of which represents a simple start to stop feature.
  - One record per row.
  - Each record has 9 fields:
    - **seqname** - the name of the sequence.
    - **source** - The program that generated this feature.
    - **feature** - The name of this type of feature.
    - **start** - The starting position
    - **end** - The ending position.
    - **score** - A score between 0 and 1000.
    - **strand** - Valid entries include '+', '-', or '.'.
    - **frame** - If the feature is a coding exon, represents the reading frame of the first base.
    - **grouping** - used for identification and grouping of multiple records into a single composite record.

The GFF3 Format

```
human15.1 . gene            214301  215772 . + . ID=HsG8283
human15.1 . mRNA            214360  215771 . + . Comments=fixed+one+splice+junction;Parent=HsG8283;Evidence=70000000697438
human15.1 . CDS             214360  214441 . + . Parent=HsT20206
human15.1 . CDS             215299  215444 . + . Parent=HsT20206
human15.1 . CDS             215641  215766 . + . Parent=HsT20206
human15.1 . three_prime_UT  215767  215771 . + . Parent=HsT20206
human15.1 . mRNA            214590  215772 . + . Comments=fixed+one+splice+site%0A;Parent=HsG8283;Evidence=700000006960084
human15.1 . five_prime_UTR  214590  214590 . + . Parent=HsT20207
human15.1 . CDS             214591  214660 . + . Parent=HsT20207
human15.1 . CDS             215299  215444 . + . Parent=HsT20207
human15.1 . CDS             215641  215769 . + . Parent=HsT20207
human15.1 . three_prime_UT  215770  215772 . + . Parent=HsT20207
human15.1 . mRNA            214301  215769 . + . Parent=HsG8283;Evidence=7000000069974357;Transcript_type=Candidates+for+D
human15.1 . five_prime_UTR  214301  214302 . + . Parent=HsT16028
human15.1 . CDS             214303  214460 . + . Parent=HsT16028
human15.1 . CDS             215299  215467 . + . Parent=HsT16028
human15.1 . three_prime_UT  215468  215769 . + . Parent=HsT16028
human15.1 . mRNA            215218  215772 . + . Parent=HsG8283;Evidence=7000000069512231;Transcript_type=Novel_Transcript
human15.1 . five_prime_UTR  215218  215233 . + . Parent=HsT16029
human15.1 . CDS             215234  215444 . + . Parent=HsT16029
human15.1 . CDS             215641  215735 . + . Parent=HsT16029
human15.1 . three_prime_UT  215736  215772 . + . Parent=HsT16029
```

1st step: Quality control with FASTQC



1st step: Quality control with FASTQC

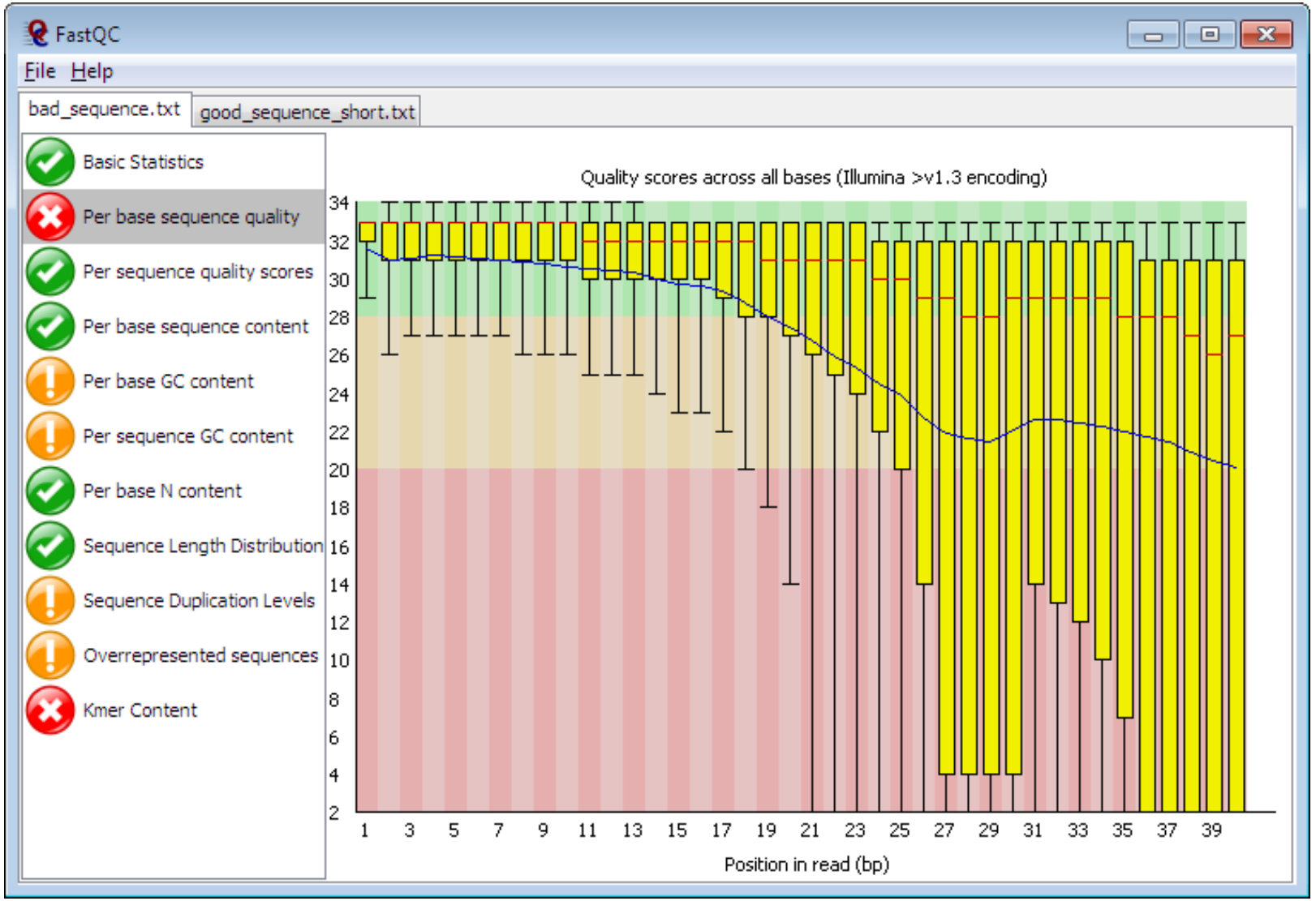
- FastQC provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines.

- 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.

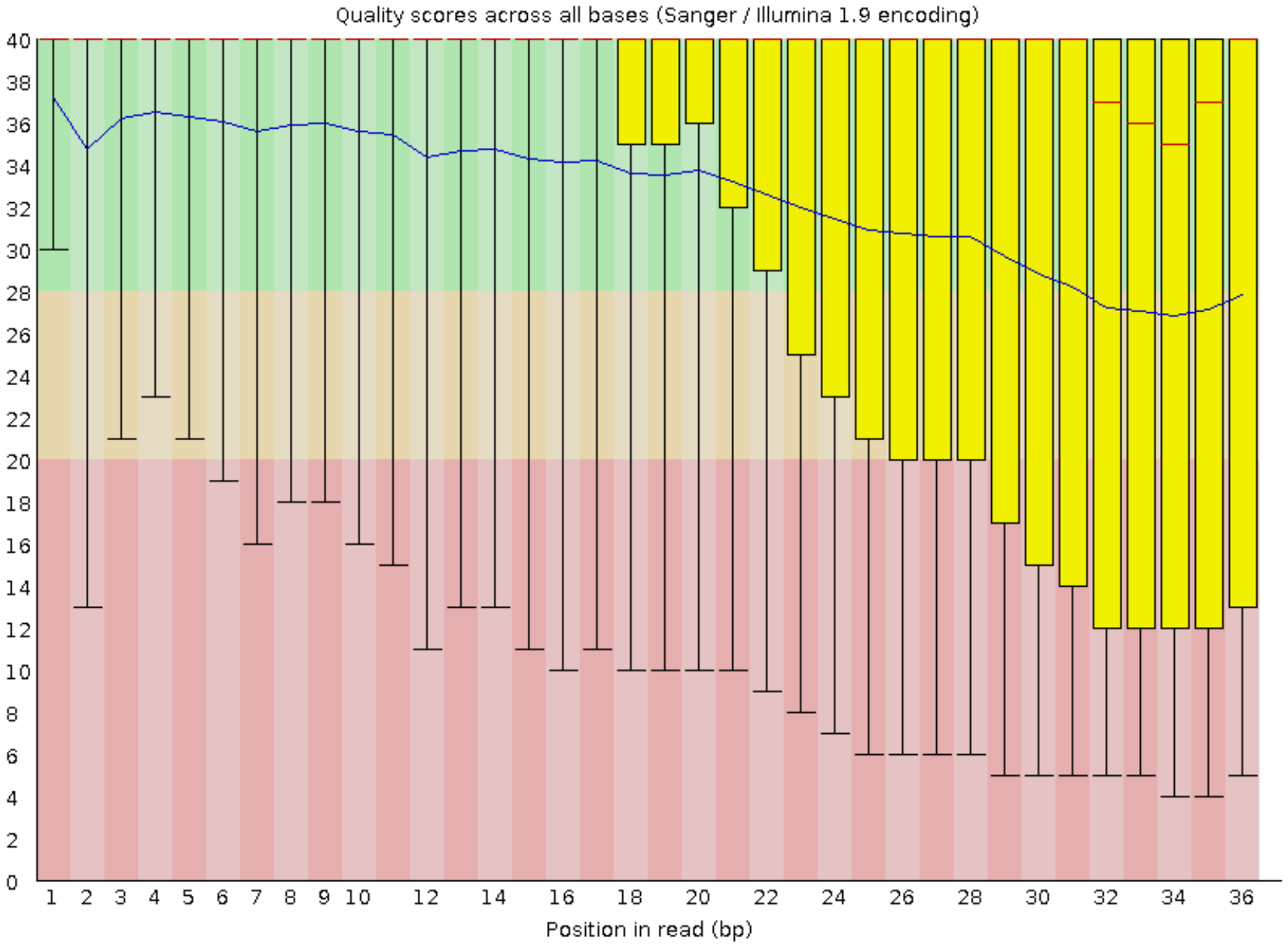
- Main functions:
  - ❖Import of data from BAM, SAM or FastQ files (any variant)
  - ❖Providing a quick overview to tell you in which areas there may be problems
  - ❖Summary graphs and tables to quickly assess your data
  - ❖Export of results

1st step: Quality control with FASTQC

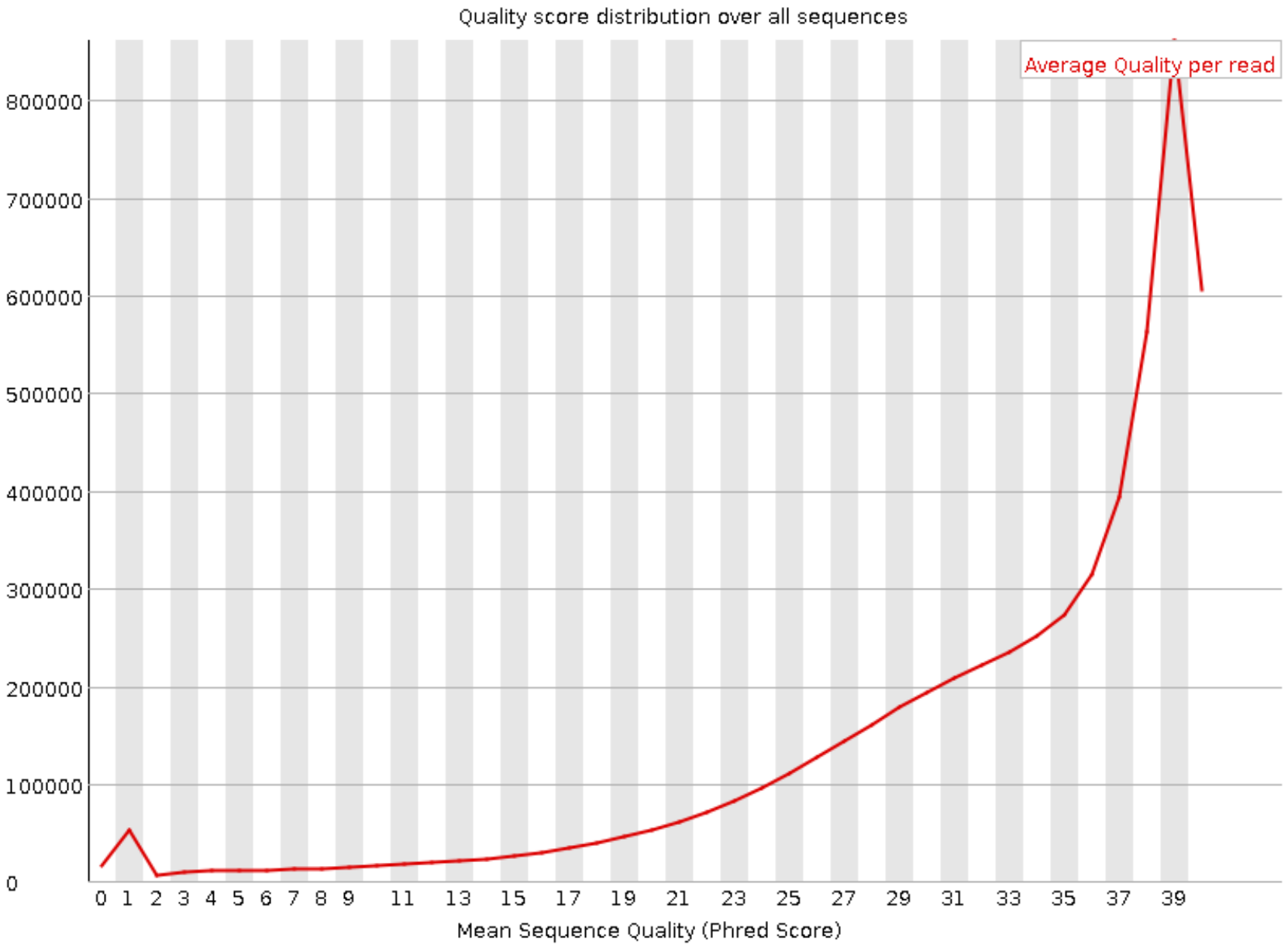
- How to use FASTQC:



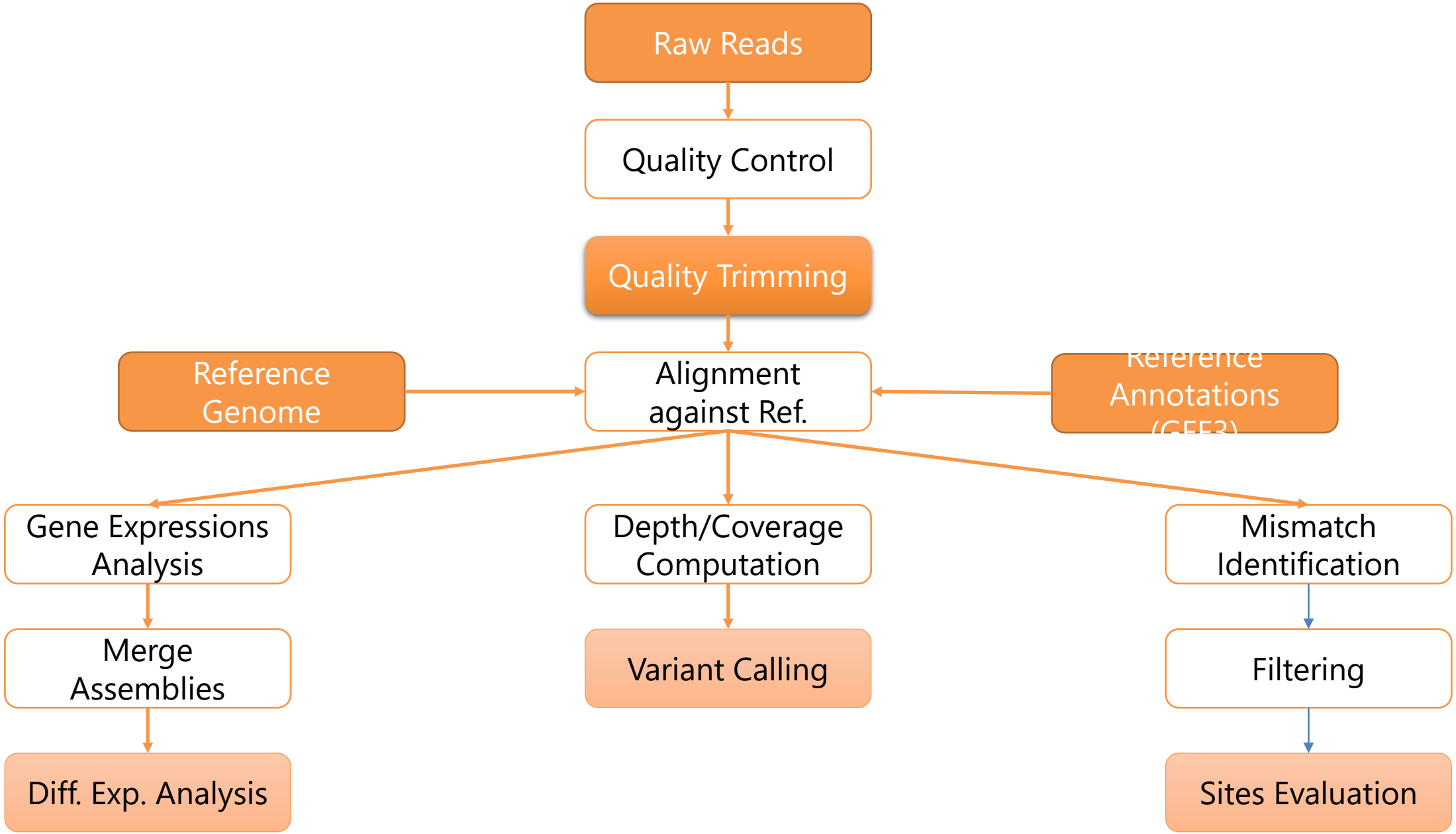
1st step: Quality control with FASTQC



1st step: Quality control with FASTQC



2nd step: Quality Trimming with erne-filter



2nd step: Quality Trimming with erne-filter

- Read trimming is the process, which aims at removing low quality portions while preserving the longest high quality part of a NGS read

- Trimming is shown to increase the quality and reliability of the analysis, with concurrent gains in performances.

- ERNE is a suite of tools, developed by IGA (Udine, Italy), whose goal is to provide an all-inclusive set of tools to handle short (NGS-like) reads.

2nd step: Quality Trimming with erne-filter

- ERNE-filter is a command line tool, which handles quality trimming and contamination filtering of reads.

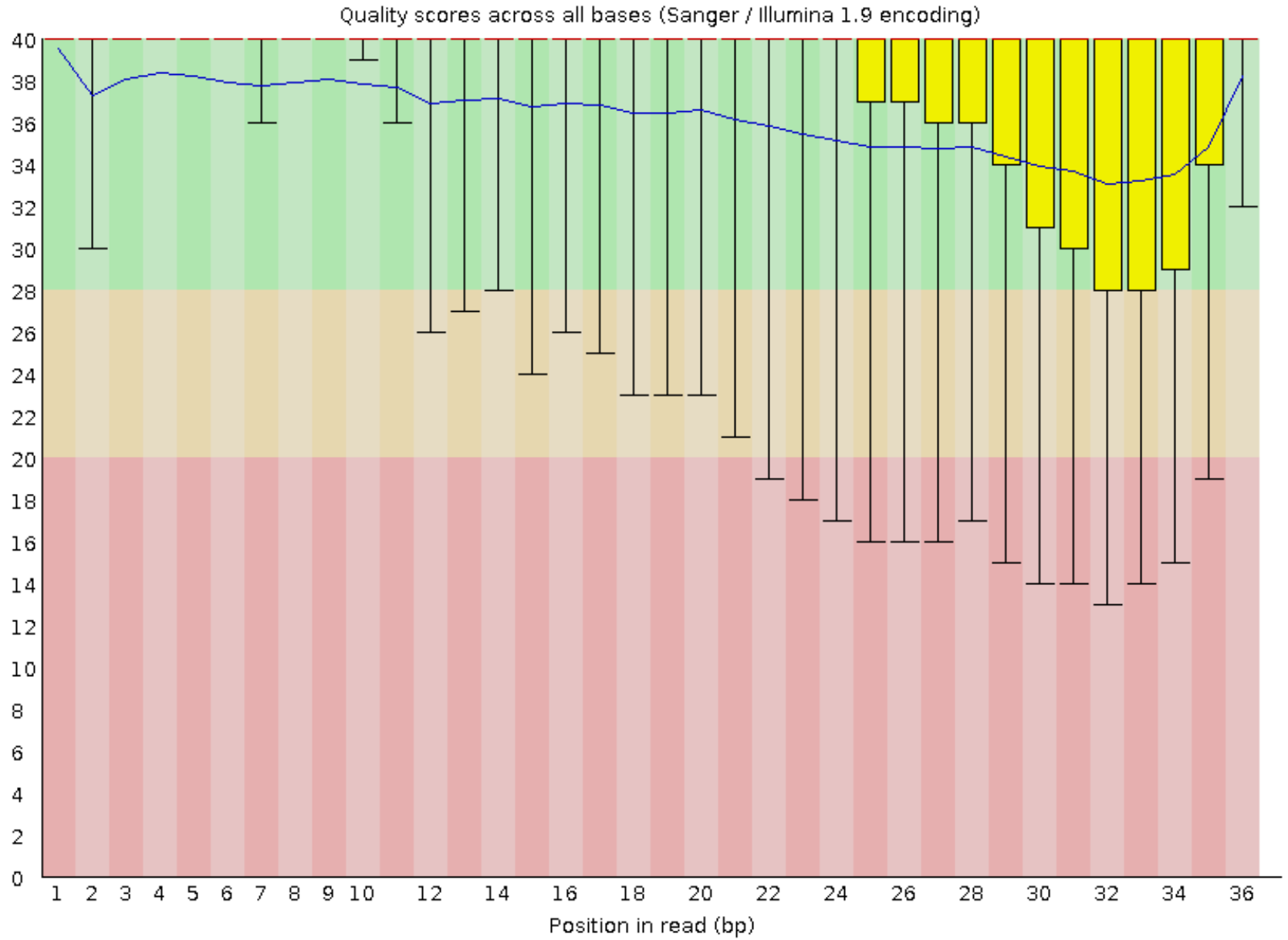
  - Given a threshold value Q, the algorithm works in two steps.
    - In the first step, it computes the first index where the quality is greater than Q.
    - In the second step, it computes the part of the sequence that can be kept without lowering overall quality.
    - Everything before and after is trimmed out.
    - After that, if the good region length is lower than a threshold or if the mean quality in the good region is lower than a threshold, then the read is discarded.

2nd step: Quality Trimming with erne-filter

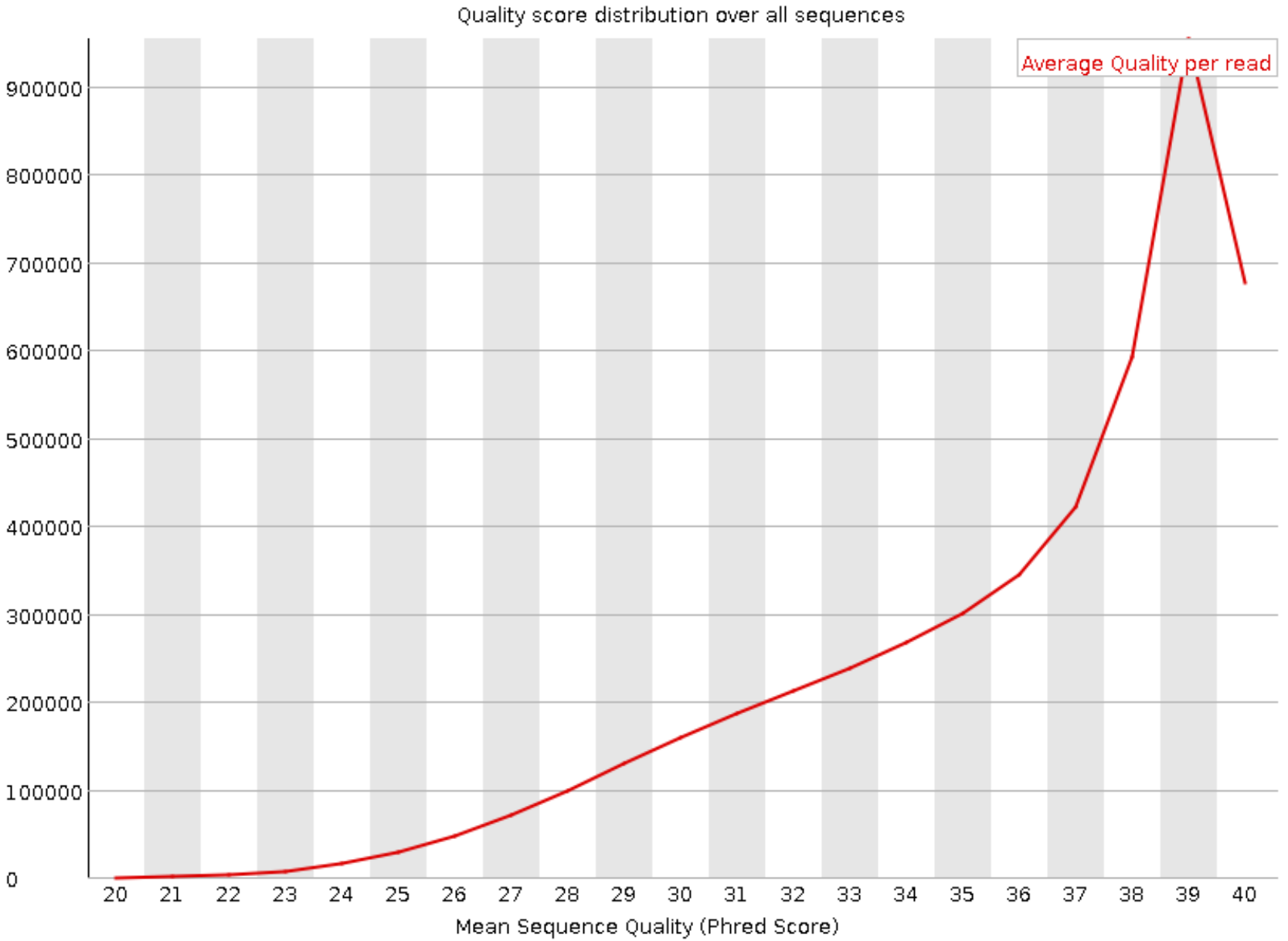
- How to use erne-filter:

```
erne-filter --gzip --threads n --query1 infile \
--query2 infile --output outbase
```

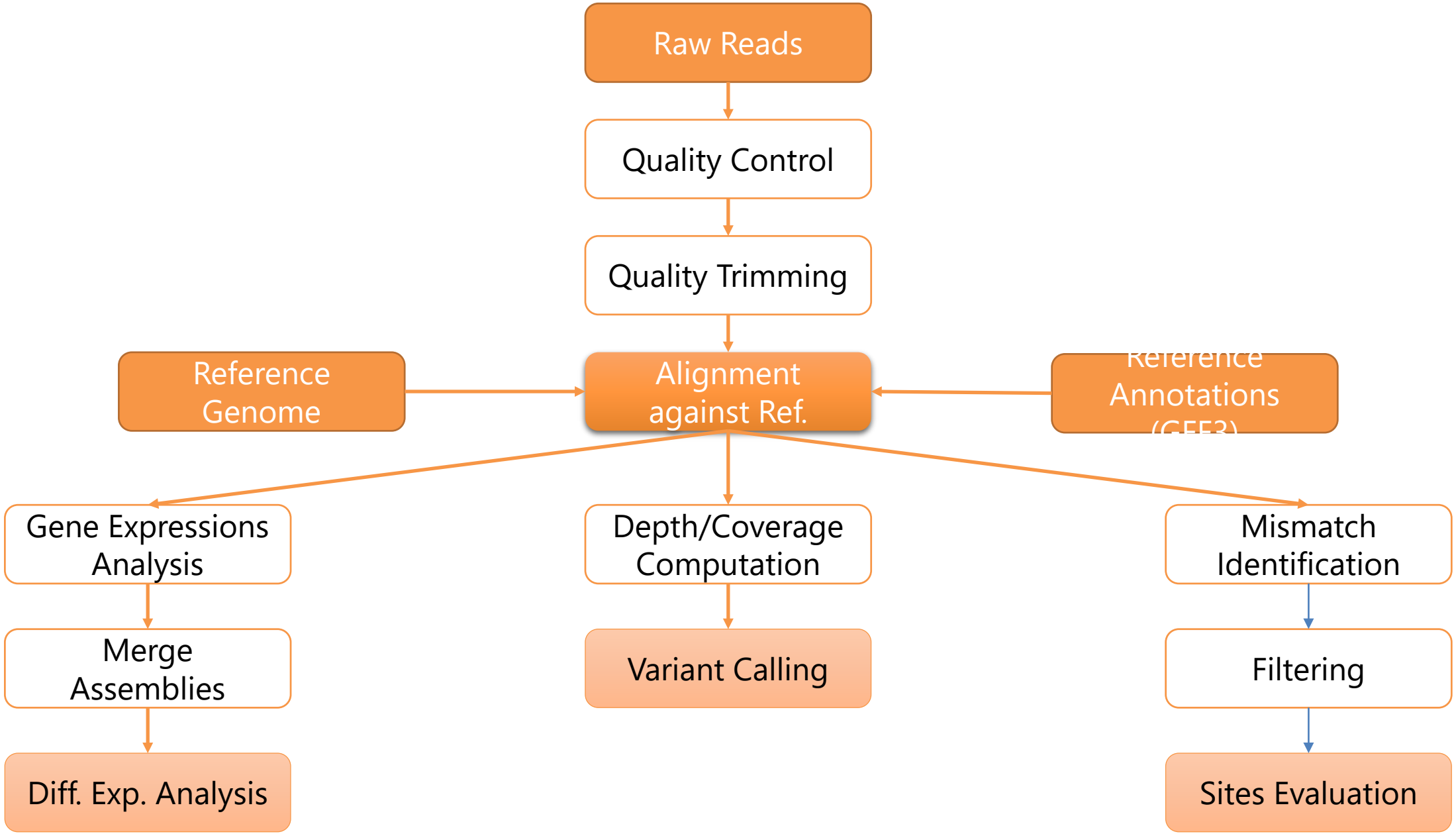
2nd step: Quality Trimming with erne-filter



2nd step: Quality Trimming with erne-filter



3rd step: Alignment against reference genome



3rd step: Alignment against reference genome

- For each read we are interested in:
  - Position and strand in the reference genome;
  - Position of mismatches against reference.

➢We need a fast algorithm!!

- Generally the alignment phase is preceded by a preprocessing of the genome in order to speed up the alignment itself.

➢Here we will introduce Bowtie 2 for DNA-Seq and TopHat 2 for RNA-Seq.

3rd step: Alignment against reference genome

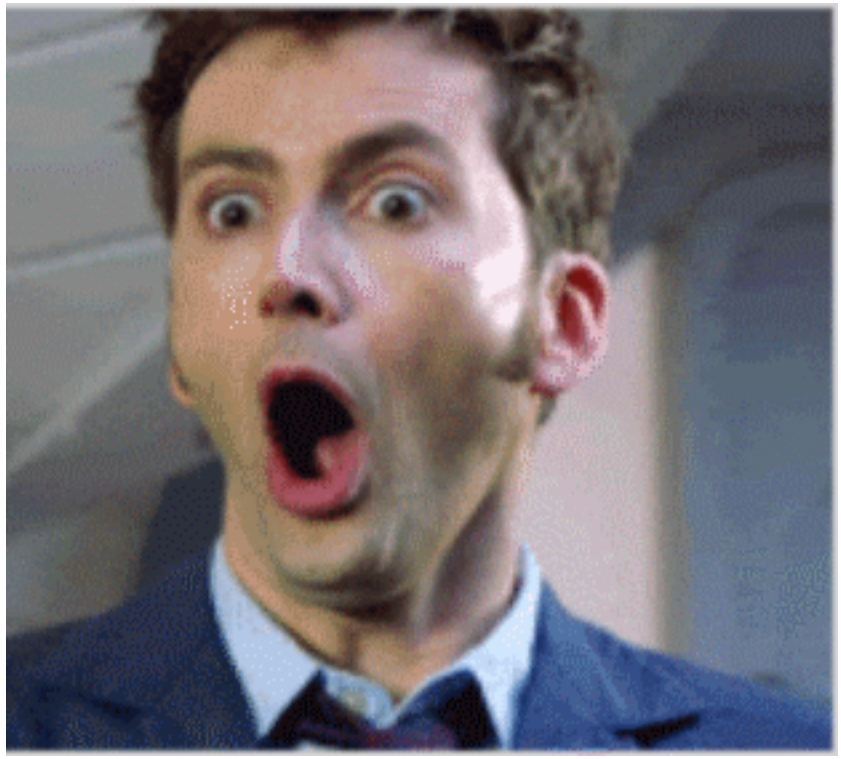
- How much computational resources we need?

➢Minimum requirements for a fast processing:

✓8 Cores Processor

✓32Gb of RAM (higher is better)

✓2Tb of Disk Storage



3rd step: Alignment against reference genome

- INPUT:
  - A sequence of known reference genome
  - A set (million) reads from the NGS experiment
- OUPUT:
  - Positioning of reads compared to the reference genome
    - A number of (small) mismatch is allowed.
- Generally the alignment phase is preceded by a preprocessing (INDEXING) for the purpose of speeding the alignment itself.

3rd step: Alignment against reference genome – Hash Table based methods

SEED (K=5)	Values
AAAAA	34,100000,…
AAAAC	3120,440,…
AAAAG	….
AAAAT	….
AAACA	….
…..	…
…..	…
…..	…
TTTTT	4591923,12…

- We build a dictionary of the reference genome to speed-up the lookup phase of each read.
- First consider the first K bases for a read (or seed K-mer) and search for the position in the reference genome.
- Once found verify that the read aligns with the reference by using an alignment algorithm (Smith-Waterman, seed-and-extends).
- Sometimes spaced-seeds are used.
  - Segments of length L where a match is required only for K bases (K < L).

- Pay attention to K!!!

- First generation aligners: MAQ, ELAND

3rd step: Alignment against reference genome – Suffix Array

- They store in an index all the suffixes of the genome.
  - Space and time requirement $O(|G|)$
- Suffixes are ordered and a simple binary search is used to find the alignment of a read.
  - Time requirement $O(|r| \cdot log|G|)$
- Second generation aligners: SOAP

i	*A*[*i*]
0	acaaacatat$
1	caaacatat$
2	aaacatat$
3	aacatat$
4	acatat$
5	catat$
6	atat$
7	tat$
8	at$
9	t$
10	$

i	SA	*A*[*SA*[*i*]]
0	2	aaacatat$
1	3	aacatat$
2	0	acaaacatat$
3	4	acatat$
4	6	atat$
5	8	at$
6	1	caaacatat$
7	5	catat$
8	7	tat$
9	9	t$
10	10	$

3rd step: Alignment against reference genome – Burrows Wheeler Transform

- Suppose we have a string:

$$G = BANANA\$$$

- Consider all possible rotations and order the results (the $ symbol can be seen as either the first or last)

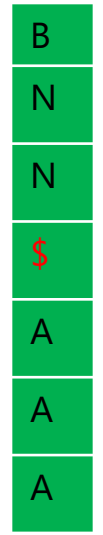
0	B	A	N	A	N	A	$
1	A	N	A	N	A	$	B
2	N	A	N	A	$	B	A
3	A	N	A	$	B	A	N
4	N	A	$	B	A	N	A
5	A	$	B	A	N	A	N
6	$	B	A	N	A	N	A

1	A	N	A	N	A	$	B
3	A	N	A	$	B	A	N
5	A	$	B	A	N	A	N
0	B	A	N	A	N	A	$
2	N	A	N	A	$	B	A
4	N	A	$	B	A	N	A
6	$	B	A	N	A	N	A

SA BWT

3rd step: Alignment against reference genome – Burrows Wheeler Transform

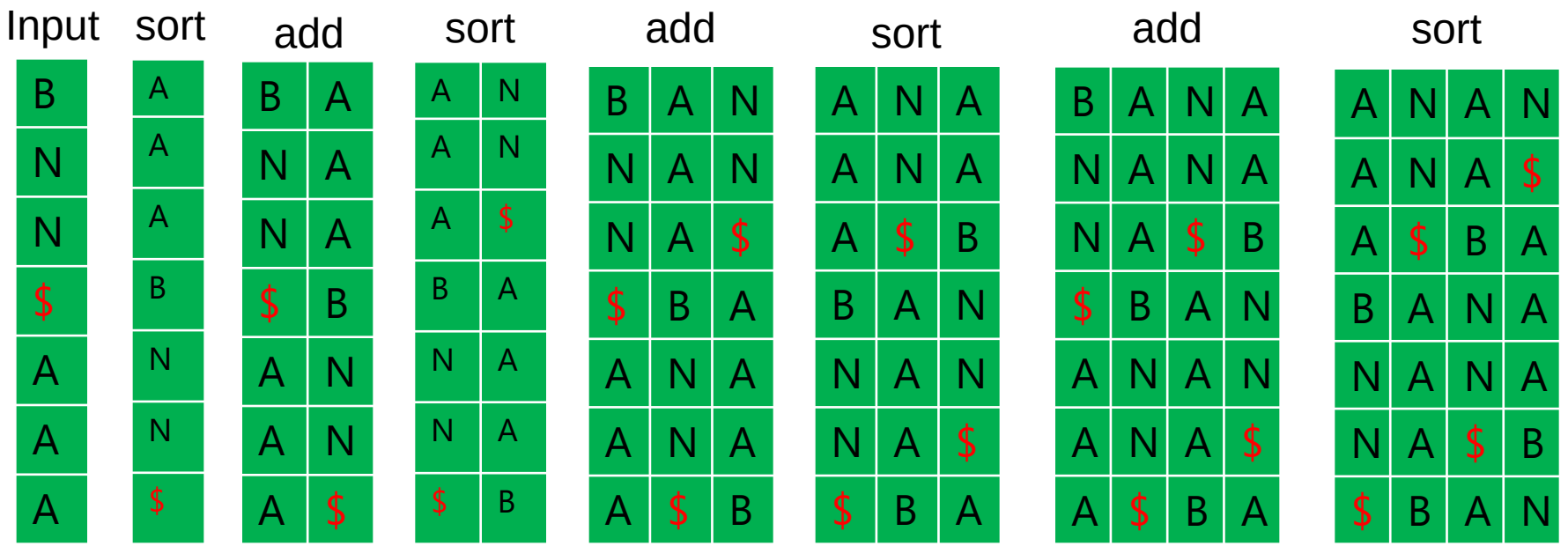
- BW transform is reversible!!



BWT

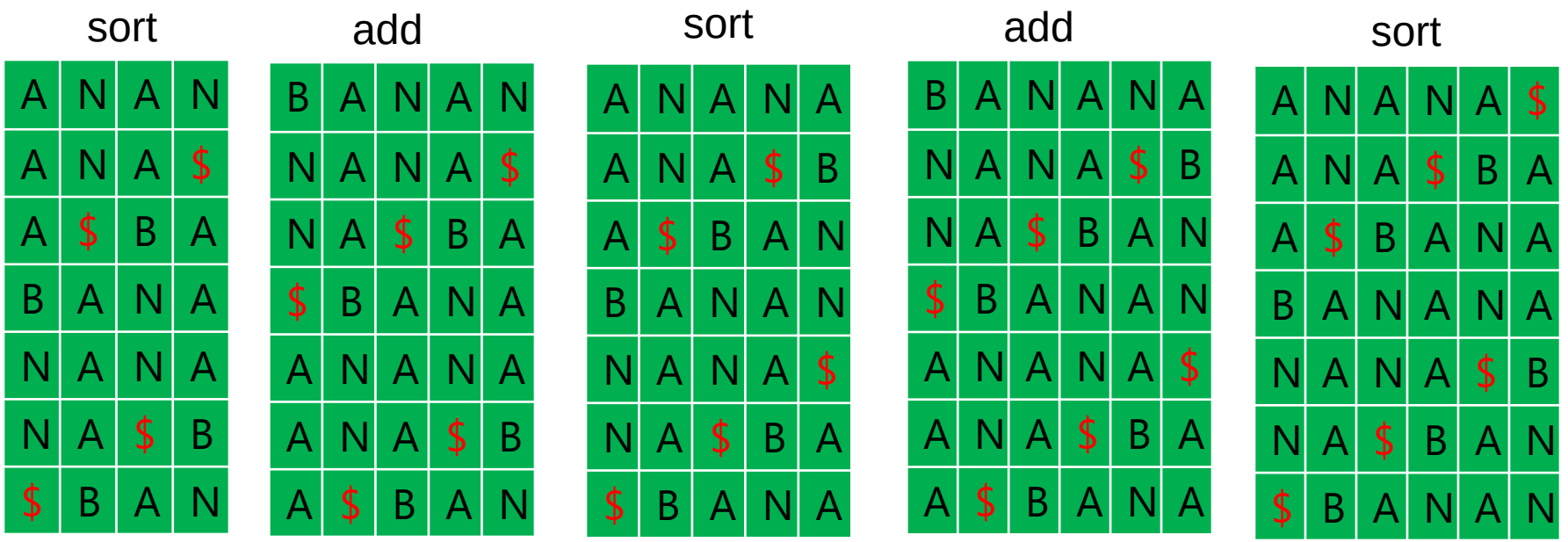
3rd step: Alignment against reference genome – Burrows Wheeler Transform

- Reversing BW transform:



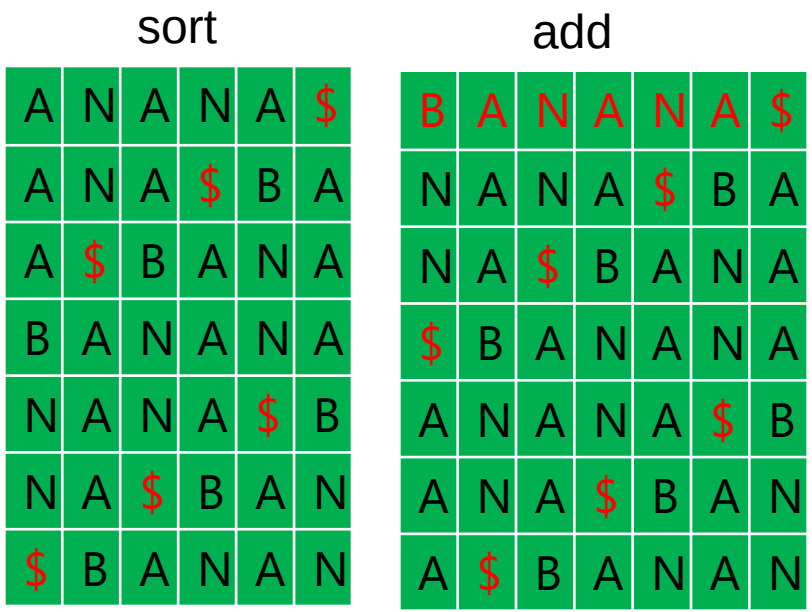
3rd step: Alignment against reference genome – Burrows Wheeler Transform

- Reversing BW transform:



3rd step: Alignment against reference genome – Burrows Wheeler Transform

- Reversing BW transform:



3rd step: Alignment against reference genome

- IMPORTANT: NGS transform can be compressed
  - even though the starting string could not

- BW-based (FM-index) methods allow search directly in the compressed space.

- Third generation aligners: BWA, Bowtie, SOAP2

3rd step: Alignment against reference genome – Bowtie 2

- Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.

- It works with reads up to 1000 characters.

- Particularly suited for aligning long genomes (e.g. mammalian)
- It uses an indexing algorithm to keep its memory footprint small:
  - human genome ~3.2 GB

3rd step: Alignment against reference genome – Bowtie 2

- How to use Bowtie 2:

I. Build an index of the reference genome with bowtie-build utility:

```
bowtie2-build reference_in index_base
```

II. Align the reads with the bowtie2 command:

```
bowtie2 –x bt2-idx -1 m1 -2 m2 -U r -S output.sam
```

3rd step: Alignment against reference genome – TopHat 2

- TopHat is a fast splice junction mapper for RNA-Seq reads.

  - It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie.

  - It analyzes the mapping results to identify splice junctions between exons.

3rd step: Alignment against reference genome – TopHat 2

- How to use TopHat 2:

  I. Build an index of the reference genome with bowtie-build utility:

  ```
  bowtie2-build reference_in index_base
  ```

  II. Align the reads with the tophat2 command:

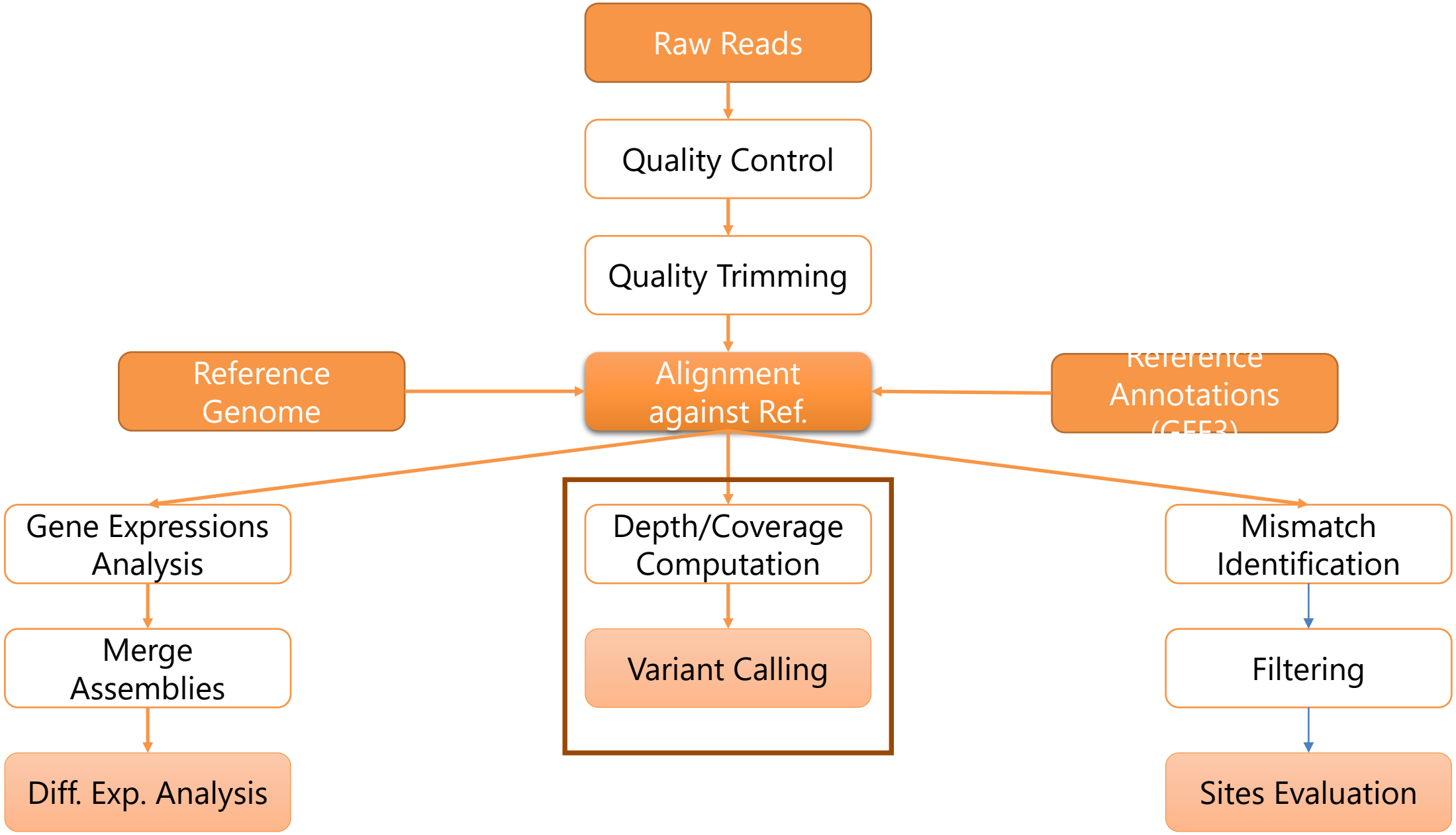
  ```
  tophat2 –G reference_annot bt2-idx \
      sample1,sample2,…
  ```

- TopHat produces a **tophat_out** directory.
- The file **accepted_hits.bam** inside the output directory contains the results of the alignment.

Bioinformatics Pipelines

Extracting Genomic Variants

Extracting Genomic Variants



SAMtools and BCFtools

- SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments.
- SAM aims to be a format that:
  - Is flexible enough to store all the alignment information generated by various alignment programs;
  - Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;
  - Is compact in file size;
  - Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory;
  - Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

➢**SAM Tools** provide various utilities for manipulating alignments

SAMtools and BCFtools

- Variant Call Format (VCF) is a text file format for storing marker and genotype data.
- Every VCF file has three parts in the following order:
  - Meta-information lines.
  - One header line.
  - Data lines contain marker and genotype data (one variant per line).

    A data line is called a VCF record.

➢ BCFtools is a set of utilities that manipulate variant calls in the Variant Call Format (VCF) and its binary counterpart BCF.

How to extract genomic variants

I. Convert SAM to BAM format and fix errors (needed for bowtie2):

```
samtools fixmate -O bam alignment_result.sam \
    fixed_result.bam
```

II. Sort BAM file:

```
samtools sort -O bam -o sorted_alignment.bam \
    -T temp_directory alignment_result.bam
```

How to extract genomic variants

III. Compute all possible genomic variants:

```
samtools mpileup –ug -o all_variant_sites.bcf \
    -f reference.fa sorted_alignment.bam
```

IV. Run variant caller:

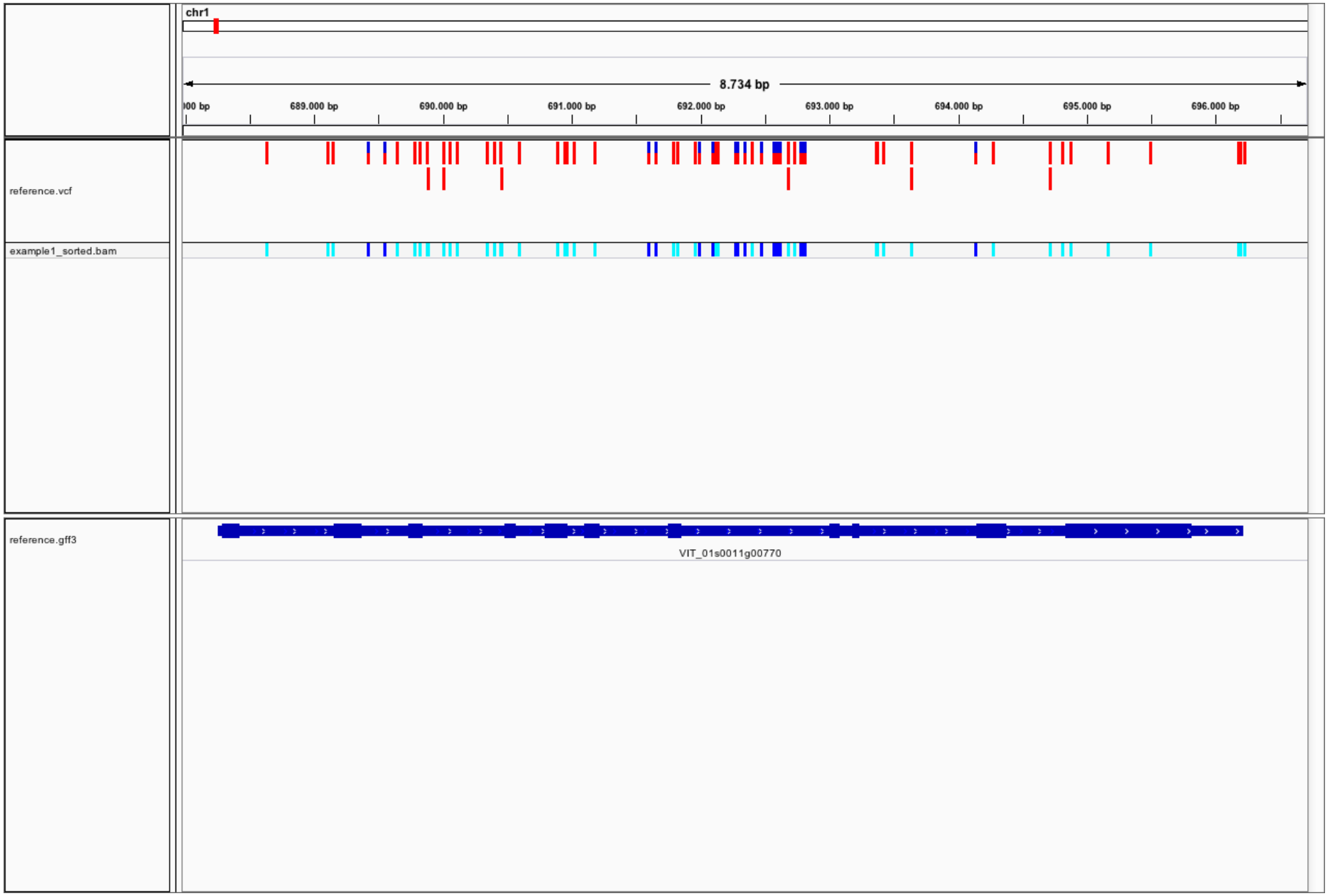
```
bcftools call –vm -O z -o called_sites.vcf.gz \
    all_variant_sites.bcf
```

- Results of variant caller can be viewed with tools like Broad IGV or UCSC browser

VCF Output Example

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr1	10749	.	T	A	156	.	DP=19;VDB=0.00181778;SGB=-0.690438;RPB=
chr1	10802	.	T	C	158	.	DP=12;VDB=0.0601435;SGB=-0.670168;MQ0F=
chr1	10815	.	A	G	109	.	DP=6;VDB=0.928942;SGB=-0.590765;MQ0F=0;
chr1	10836	.	C	A	103	.	DP=9;VDB=0.0697235;SGB=-0.616816;RPB=0.
chr1	10871	.	T	C	59	.	DP=14;VDB=0.446842;SGB=-0.556411;RPB=0.
chr1	11268	.	A	G	197	.	DP=8;VDB=0.0760875;SGB=-0.651104;MQSB=
chr1	11333	.	T	C	57	.	DP=3;VDB=0.421281;SGB=-0.511536;MQSB=1
chr1	11355	.	C	A	101	.	DP=7;VDB=0.329692;SGB=-0.616816;MQSB=1
chr1	11674	.	A	T	114.963	.	DP=1;SGB=-0.379885;MQ0F=0;ICB=1;HOB=0.5
chr1	11713	.	G	A	228	.	DP=12;VDB=0.593729;SGB=-0.680642;MQSB=
chr1	11890	.	A	C	119.226	.	DP=2;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP
chr1	11922	.	A	T	144.918	.	DP=2;VDB=0.42;SGB=-0.453602;MQSB=1;MQ0
chr1	11952	.	C	A	96.729	.	DP=1;SGB=-0.379885;MQ0F=0;ICB=1;HOB=0.5
chr1	12077	.	AA	AATA	53	.	INDEL;IDV=2;IMF=0.117647;DP=12;VDB=0.939
chr1	12078	.	A	AT	470.915	.	INDEL;IDV=6;IMF=0.352941;DP=12;VDB=0.901

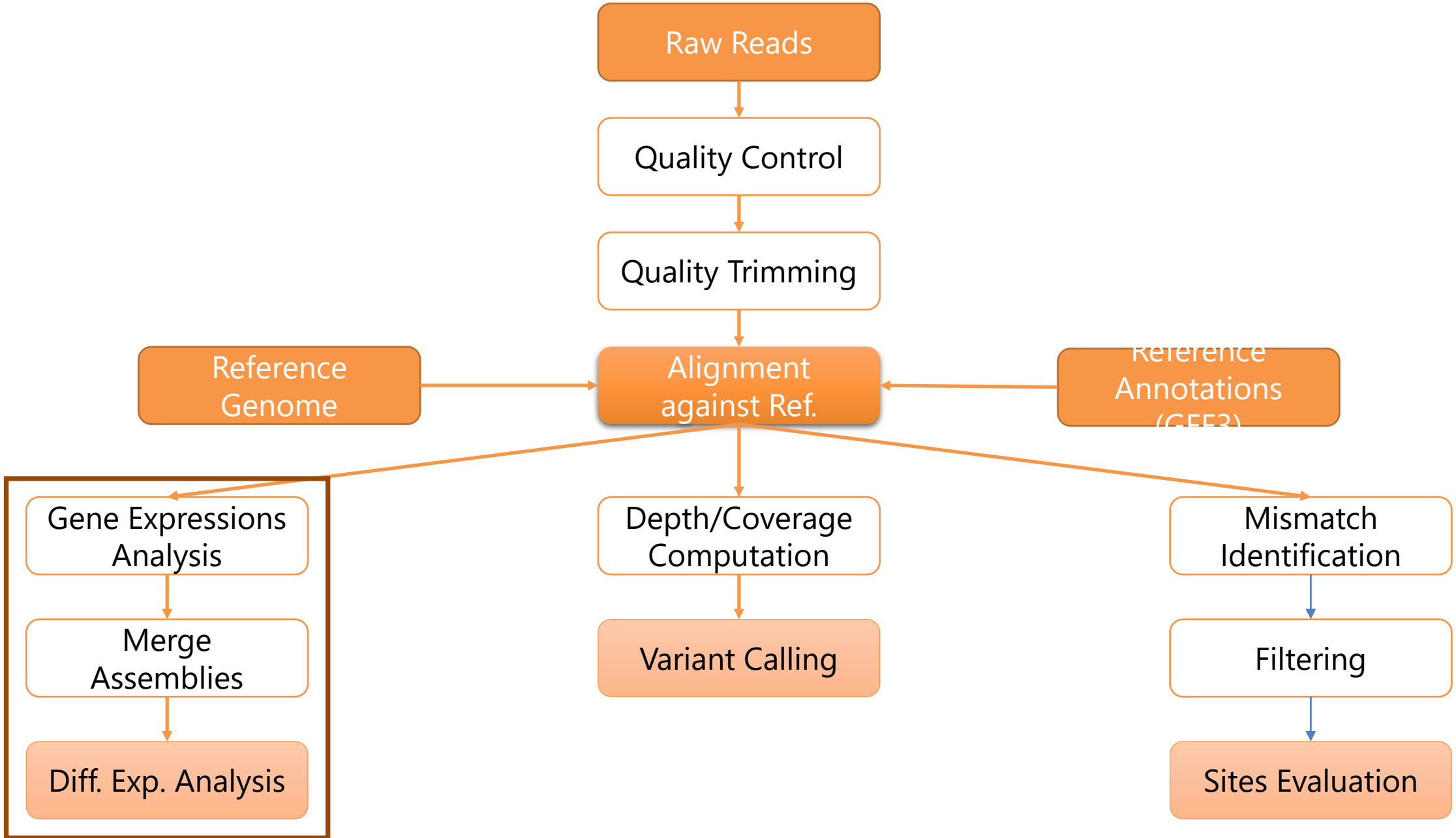
VCF Output Example



Bioinformatics Pipelines

Extracting RNA Expressions

Extracting RNA Expressions

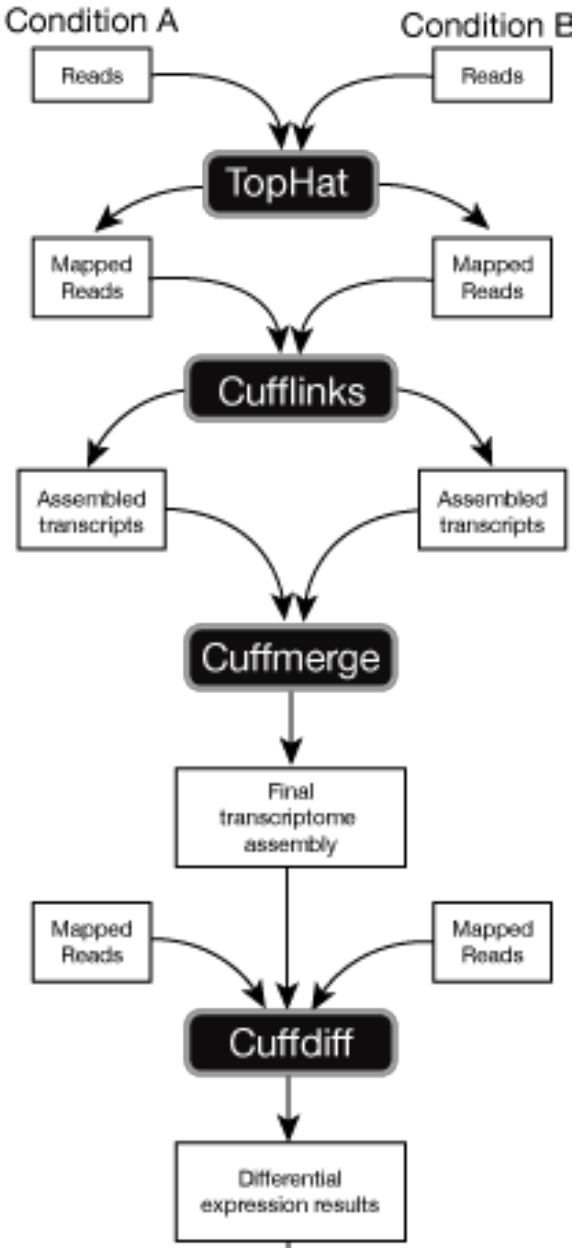


The Cufflinks suite

- The Cufflinks suite assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples:
  - ❖accepts aligned RNA-Seq reads
  - ❖assembles the alignments into a set of transcripts
  - ❖estimates the relative abundances of these transcripts (taking into account biases in library preparation protocols)

- The Cufflinks suite includes a number of different programs that work together to perform these analyses.

Cufflinks Workflow



The programs in the Cufflinks suite

- **Cufflinks**: assembles transcriptomes from RNA-Seq data and quantifies their expression (RPKM).

- **Cuffmerge**: for multiple RNA-Seq libraries, it merges the assemblies into a master transcriptome.

- **Cuffdiff**: compares expression levels of genes and transcripts, determining not only which genes are up- or down-regulated between two or more conditions, but also which genes are differentially spliced or are undergoing other types of isoform-level regulation.

Cufflinks output

- In each output directory, the application will create several files.

➢We are interested in:

✓`transcripts.gtf`

- This file contains the assembled transcripts.
- It can be viewed with tools like Broad IGV or UCSC browser

✓`genes.fpkm_tracking`

- It quantifies the expression of genes, specified in the annotation file.

✓`isoforms.fpkm_tracking`

- It quantifies the expression of transcripts, specified in the annotation file.

Cufflinks output

Cuffmerge output

- In each output directory, the application will create several files.

✓We are interested in `merged.gtf`

- This file contains the set of merged transcripts.

- Each line contains an annotation field that describes the nature of the overlap of this transcript with transcripts from the reference genome.

Cuffdiff output

- In each output directory, the application will create several files.

✓We are interested in `gene_exp.diff` and `isoform_exp.diff`

- These files contains the list of genes (isoforms) that are found differentially expressed.

➢For further description of the format and contents of this file, see http://cole-trapnell-lab.github.io/cufflinks/.

How to extract RNA expressions

I. Run cufflinks for each RNA sample:

```
cufflinks -g reference.fa -L label \
    -o output_dir tophat_out/accepted_hits.bam
```

II. Create a text file where each line corresponds to a "transcripts.gtf" file

III. Run cuffmerge:

```
cuffmerge -o output_dir -g reference.fa \
    list_of_cufflinks_results.txt
```

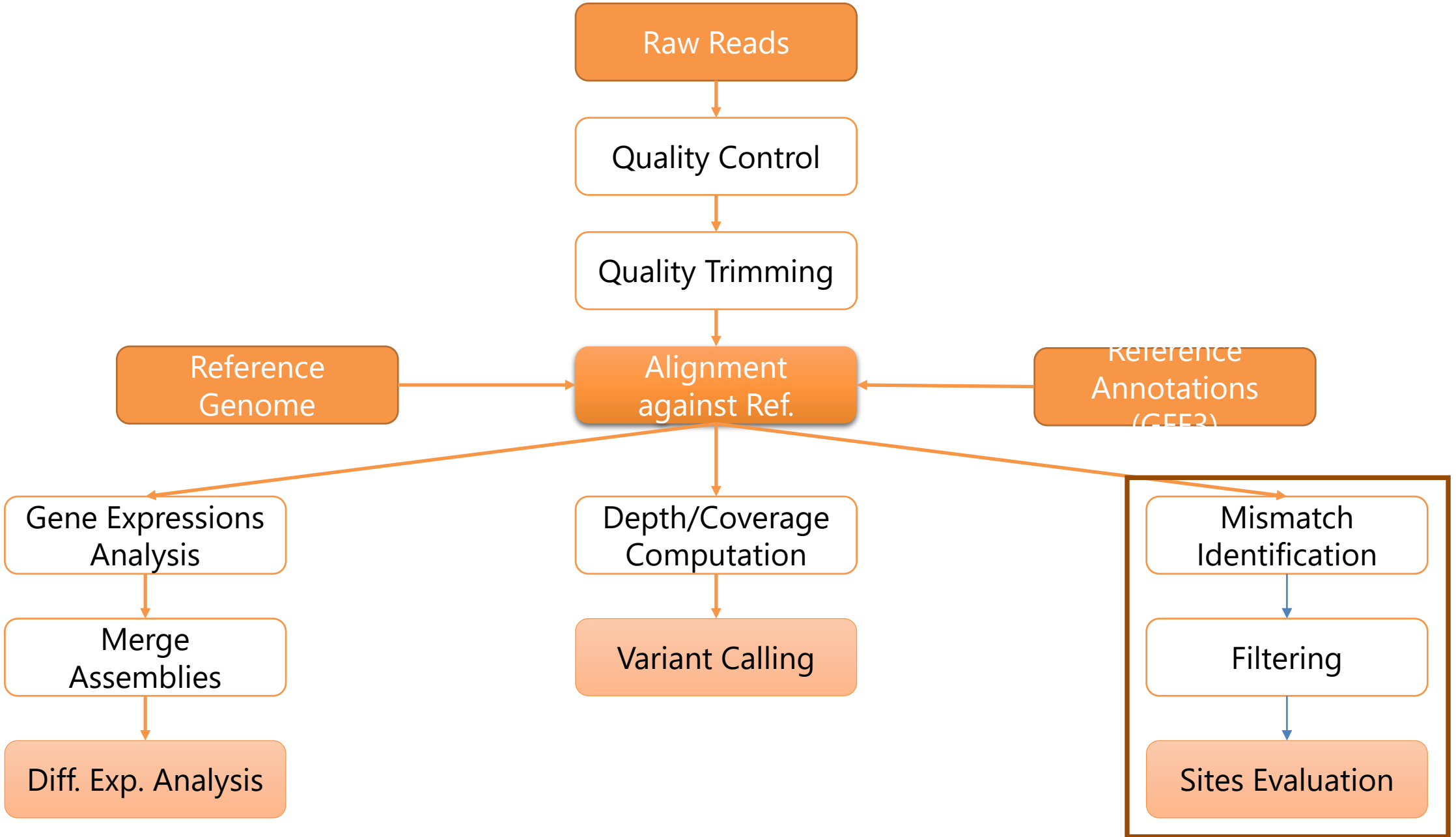
IV. Run cuffdiff:

```
cuffdiff -o output_dir -L label1,label2,… \
    cuffmerge_out/merged.gtf samples1,… samples2,… …
```

Bioinformatics Pipelines

Identifying RNA Editing Events

Identifying RNA Editing Events



What is editing?

- RNA editing is a post-transcriptional phenomenon involving the insertion/deletion or substitution of specific bases in precise RNA localizations.

- In human, RNA editing occurs mostly by the **adenosine to inosine** (A-to-I) conversion through ADAR enzymes.
  - ❖A-to-I substitutions may have profound functional consequences
  - ❖A-to-I substitutions have been linked to a variety of human diseases:
    - ➢neurological, neurodegenerative disorders, or cancer.

- Next generation sequencing technologies offer the unique opportunity to investigate in depth RNA editing.

REDItools

- REDItools are simple python scripts conceived to facilitate the investigation of RNA editing at large-scale and devoted to research groups that would to explore such phenomenon in own data but don't have sufficient bioinformatics skills.

  - They work on main operating systems
  - They can handle reads from whatever platform in the standard BAM format and implement a variety of filters.

How to identify putative RNA editing sites

I. If starting from SAM format, convert SAM to BAM:

```
samtools view -b -T reference.fa aligned.sam > aligned.bam
```

II. Sort your BAM file:

```
samtools sort aligned.bam aligned.sorted
```

III. Index sorted BAM file:

```
samtools index myfile.sorted.bam
```

IV. Download and pre-process GTF annotations from UCSC website

More information at: http://srv00.ibbe.cnr.it/reditools/#gtf

How to identify putative RNA editing sites

V. Run REDItoolDnaRNA to find putative editing sites:

```
REDItoolDnaRna.py -i rna.bam -j dna.bam -f reference.fa \
    -o out/
```

VI. Filter candidates:

```
selectPositions.py -i out/outfile -e -u –o candidates.txt
```

VII. Annotate candidates with informations in the RepeatMask database to look at Alu sites:

```
AnnotateTable.py -a rmsk.gtf.gz -i candidates.txt -u \
    -c 1,2,3 -n RepMask -o candidates.rmsk.txt
```

VIII. Add gene annotations using RefSeq database:

```
AnnotateTable.py –a refGene.gtf.gz -i candidates.rmsk.txt -u \
-c 1,2,3 -n RefSeq -o candidates.rmsk.alu.ann.txt
```

Output example

Region	Position	Reference	Strand	Coverage-q25	MeanQ	BaseCount[A,C,G,T]	AllSubs	Frequency	gCoverage-q25	gMeanQ	gBaseCount[A,C,G,T]	gAllSubs	gFrequency	RepMask_feat	RepMask_gid	RepMask_tid	RefSeq_feat	RefSeq_gid
chr21	47739578	A	0	14	38.50	[11, 0, 3, 0]	AG	0.21	26	30.27	[26, 0, 0, 0]	-	0.00	SINE	AluSq4	Alu-SINE	intron	C21orf58
chr21	47739644	A	0	18	36.61	[13, 0, 5, 0]	AG	0.28	23	30.22	[23, 0, 0, 0]	-	0.00	SINE	AluSq4	Alu-SINE	intron	C21orf58
chr21	47739647	A	0	16	36.12	[7, 0, 9, 0]	AG	0.56	18	30.67	[18, 0, 0, 0]	-	0.00	SINE	AluSq4	Alu-SINE	intron	C21orf58
chr21	47739724	A	0	8	37.00	[6, 0, 2, 0]	AG	0.25	30	30.10	[30, 0, 0, 0]	-	0.00	SINE	AluSq4	Alu-SINE	intron	C21orf58
chr21	47739725	A	0	8	37.75	[5, 0, 3, 0]	AG	0.38	30	29.93	[30, 0, 0, 0]	-	0.00	SINE	AluSq4	Alu-SINE	intron	C21orf58
chr21	47739764	A	0	6	37.33	[4, 0, 2, 0]	AG	0.33	19	30.37	[19, 0, 0, 0]	-	0.00	SINE	AluSq4	Alu-SINE	intron	C21orf58
chr21	47740295	A	0	10	34.00	[6, 0, 4, 0]	AG	0.40	30	29.77	[30, 0, 0, 0]	-	0.00	SINE	AluSz	Alu-SINE	intron	C21orf58
chr21	47741150	A	0	28	36.57	[25, 0, 3, 0]	AG	0.11	24	31.54	[24, 0, 0, 0]	-	0.00	SINE	AluSx4	Alu-SINE	intron	C21orf58
chr21	47741221	A	0	49	36.33	[44, 0, 5, 0]	AG	0.10	40	29.50	[40, 0, 0, 0]	-	0.00	SINE	AluSx4	Alu-SINE	intron	C21orf58

Next Generation Sequencing

for Cancer Biomarker Discovery

Biomarkers

- Biomarkers are key molecular, chemical or cellular characteristics that
  - ❖can be objectively measured
  - ❖can be used to describe biological processes, pathogenic state, and response to therapy.

- Three main type of biomarkers
  - ➢**Diagnostic** – used to establish the disease state
  - ➢**Prognostic** – provide information regarding potential clinical outcome irrespective of treatment
  - ➢**Predictive** – provide information regarding potential clinical outcome in response to treatment

Biomarkers

- For a biomarker to become accepted for clinical application, it should have the following characteristics:
  - ✓Readily and consistently detectable in biological fluids, tissues, or other biological specimens
  - ✓Rapidly detectable and stable
  - ✓High sensitivity and specificity
  - ✓Strong correlation with the phenotype or outcome of interest
  - ✓Better if detectable via noninvasive, and cost-effective tests
  - ✓Consistent across genders.

➢In cancer a biomarker should also be specific to the cancer subtype (metastatic potential, detectable in archived samples).

Limitation of traditional biomarkers

- Previous examples are very useful in establishing diagnosis in many cases. However
  - They suffer from low specificity and sensitivity.
  - It is becoming clear that blood biomarkers and imaging, may be useful in cancer diagnosis, but they are not sufficient to provide enough information for the best course of therapeutic outcome.

- CRITICAL: Targeted therapy needs reliable, precise, and clinical relevant biomarker

- Examples:
  - CML with BCR-ABL translocation and Imatinib
  - HER2/neu positive breast cancer and Trastuzumab

Why NGS?

✓NGS enable performing millions of sequencing reactions in parallels in a short time!!

- ❖unprecedented scale and rapidly declining cost
- ❖With NGS, we can sequence entire cancer genomes
  - ❖Powerful tool to get an unbiased view of the genome
  - ❖Significantly improves chances of identifying actionable genetic aberrations.
- ❖NGS can be also employed to study epigenetic factors (like methylation)

➢Commercially, there are NGS-based cancer panels already used in clinical practice to guide patients to most appropriate treatment.

How can NGS be employed?

- ✓Discovery of genetic variants
- ✓Gene expression profiles
- ✓Epigenetic modifications
- ✓MicroRNA

How can NGS be employed?

✓Discovery of genetic variants

❖Genetic variants for Breast Cancer

- ➢Most common cancer in woman (USA 250k cases and 40k deaths)
- ➢Arise from genetic mutations (~20% family history)

Gene Name	Mutation Frequency	NGS panel cancer test	Potential therapy
EGFR	28%	Ion AmpliSeq Cancer Hotspot v2 FoundationOneTM TruSeq Amplicon Cancer Panel	Erlotinib, Gefitinib, Vandetanib, Afatinib, Icotinib, Canertinib, Epitinib
KRAS	16%	Ion AmpliSeq Cancer Hotspot v2 FoundationOneTM TruSeq Amplicon Cancer Panel	
TP53	34%	Ambry Genetics CancerNext FoundationOneTM Ion AmpliSeq Cancer Hotspot v2 TruSeq Amplicon Cancer Panel	Ad5CMV-p53 gene

How can NGS be employed?

✓Gene expression profiles

❖Many Microarray based techniques are available for common cancer types

- (Mammaprint, BluePrint, TargetPrint, …)

❖These technology, however, report expression at "**gene level**"

❖In reality:

➢many **isoforms** exist!

➢Numerous studies have shown **altered expression of specific gene isoforms in cancer**

✓For Example, in colon cancer, a specific fusion of RSPO2 and RSPO3 were found in ~10% of tumors.

How can NGS be employed?

✓Epigenetic modifications

❖Epigenetic modifications are changes in DNA independent of variations in its sequence

- ❖DNA methylation, histone acetylation and methylation, …

❖They have profound impact on gene expression.

❖Can act as **biomarkers** for cancer detection:

- ✓Hypermethylation of GSTP1 in prostate cancer
- ✓Hypermethylation of DAPK and RASSF1A in bladder cancer

- ✓A specific pattern composed of more than 500 differentially methylated genes found in hepatocellular carcinoma

(Mah et al., Methylation profiles reveal distinct subgroup of hepatocellular carcinoma patients with poor prognosis. PLoS One. 2014)

How can NGS be employed?

✓MicroRNA

❖Small single stranded RNA molecules that play critical role in gene expression regulation

❖Fundamental in various process (like cell growth)

❖Hence, they have a key role in carcinogenesis

Cancer Type	Sample Type	RNA found	Sensitivity	Specificity	AUC
Breast Cancer	Serum	miR-16, miR-25, miR-222, miR-324-3p	91.7%	89.6%	0.954
Astrocytoma	Serum	miR-15b*, miR-23a, miR-133a, miR-150*, miR-197, miR-497, miR-548b-5p	88%	97.9%	0.972
Gastric Cancer	Serum	miR-1, miR-20, miR-27a, miR-34, miR-423-5p	80%	81%	0.879

Next Generation Sequencing

The MedSeq project

The MedSeq Project

- The **MedSeq** Project is the **FIRST randomized controlled trial** which is trying to introduce Whole Genome Sequencing in **everyday clinical practice**.

- The MedSeq Project is supported by the National Institutes of Health (NIH) National Human Genome Research Institute

- Institutional Review Board approval in August 2012.

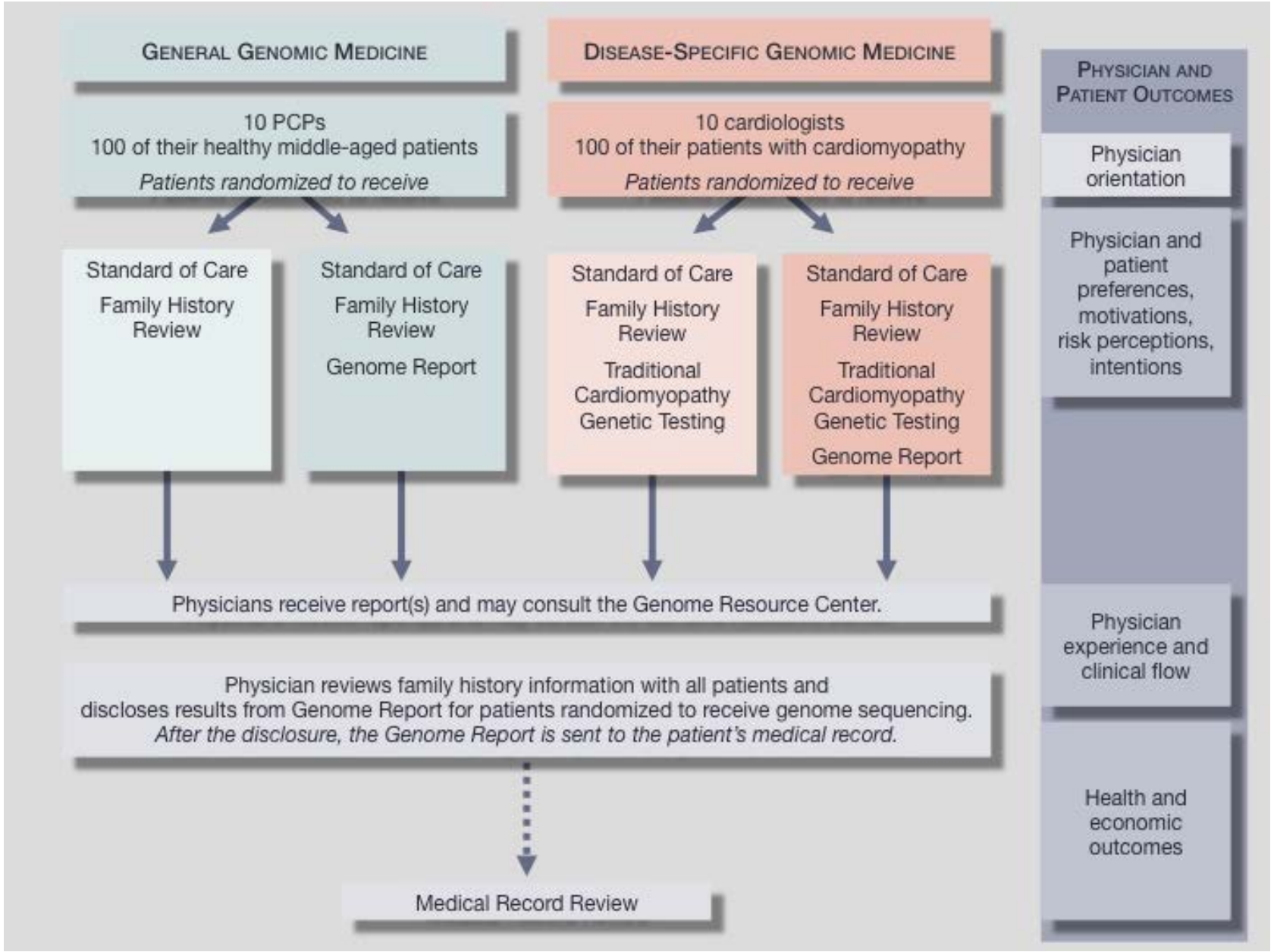
- To date:
  - enrolled 10 PCPs,
  - 9 cardiologists, and 200 of their patients.
  - 33 Genome Reports have been delivered to enrolled physicians and their patients.

The MedSeq Project

- Three main questions to answer

1. How should a clinical molecular genetics laboratory process and report WGS results to physicans and their patients?
2. With education and appropriate support, will non-geneticist physicians feel prepared to discuss and manage WGS results with their patients?
3. How will the delivery of WGS results impact the actions, attitudes and outcomes of patients and their physicians?

Study Schema



Patients Selection Creteria

	Both trials	Primary care trial	Cardiology trial
Inclusion criteria	Patients receiving care from MedSeq Project physician participants	Age 40-65 years	Age 18-90 years
		Generally healthy, in the judgment of the patient's participating physician	Diagnosis of hypertrophic cardiomyopathy (HCM) or dilated cardiomyopathy (DCM)
		No indication for a genetic test	Prior or concurrent targeted genetic testing for HCM or DCM
Exclusion criteria	Clinically significant anxiety (Hospital Anxiety and Depression Scale [HADS] anxiety subscale >14) or depression (HADS depression subscale >16) at baseline assessment	Presence of cardiovascular disease or diabetes	
	Reported current pregnancy or intention for future conception in the next year of participant or spouse/partner		

Vassy *et al.*

Report example

Name:DOE, JONATHAN
DOB: 12/34/5678
Sex: Male
Race: Caucasian
MRN:123456789
Specimen: Blood, Peripheral
Received: 12/34/5678

Accession ID: PMXX-12345
Family #: F12345
Referring physician: MedSeq
Referring facility: MedSeq

GENERAL GENOME REPORT

RESULT SUMMARY

Sequencing of this individual's genome was performed and covered 95.7% of all positions at 8X coverage or higher, resulting in over 5.2 million variants compared to reference genome. These data were analyzed to identify previously reported variants of potential clinical relevance as well as novel variants that could reasonably be assumed to cause disease (see methodology below). All results are summarized on page 1 with further details on subsequent pages.

MONOGENIC DISEASE RISK: 1 VARIANT IDENTIFIED

This test identified 1 genetic variant that may be responsible for existing disease or the development of disease in this individual's lifetime.

Disease (Inheritance)	Phenotype	Gene (Variant)	Classification
X-linked recessive chondrodysplasia punctata (X-linked)	Abnormal bone and cartilage development	ARSE (c.410G>C p.Gly137Ala)	Uncertain significance: Favor pathogenic

CARRIER RISK: 2 VARIANTS IDENTIFIED

This test identified carrier status for 2 autosomal recessive disorders.

Disease (Inheritance)	Phenotype	Gene (Variant)	Classification	Carrier Phenotype*
Cystic fibrosis (Autosomal recessive)	Chronic lung and digestive disease	CFTR (c.3846G>A p.Trp1282X)	Pathogenic	None Reported
Glycogen storage disease 7 (Autosomal recessive)	Severe exercise intolerance	PFKM (c.237+1G>A)	Pathogenic	None Reported

As a carrier for recessive genetic variants, this individual is at higher risk for having a child with one or more of these highly penetrant disorders. To determine the risk for this individual's future children to be affected, the partner of this individual would also need to be tested for these variants. Other biologically related family members may also be carriers of these variants. *Carriers for some recessive disorders may be at risk for certain phenotypes. Please see variant descriptions for more information.

PHARMACOGENOMIC ASSOCIATIONS

This test identified the following pharmacogenomic associations. Additional pharmacogenomic results may be requested, but will require additional molecular confirmation prior to disclosure.

Drug	Risk and Dosing Information
Warfarin	Increased dose requirement
Clopidogrel	Typical response to clopidogrel
Digoxin	Intermediate metabolism and serum concentration of digoxin
Metformin	Decreased glycemic response to metformin
Simvastatin	Typical risk of simvastatin-related myopathy

BLOOD GROUPS

This test identified the ABO Rh Blood Type as AB Negative. Based on their results, this person is a very desirable universally compatible platelet donor. Additional RBC and platelet antigen information is available at the end of the report.

It should be noted that the disease risk section of this report is limited only to variants with strong evidence for causing highly penetrant disease, or contributing to highly penetrant disease in a recessive manner. Not all variants identified have been analyzed, and not all regions of the genome have been adequately sequenced. These results should be interpreted in the context of the patient's medical evaluation, family history, and racial/ethnic background. Please note that variant classification and/or interpretation may change over time if more information becomes available. For questions about this report, please contact the Genome Resource Center at GRC@partners.org.

DETAILED VARIANT INFORMATION

MONOGENIC DISEASE RISK

Disease (Inheritance)	Gene (Transcript)	Variant (Classification)	Variant Frequency	Disease Prevalence	References
X-linked recessive chondrodysplasia punctata (X-linked)	ARSE (NM_000047.2)	c.410G>C p.Gly137Ala hemizygous (Uncertain significance: Favor pathogenic)	1/6728 (0.01%) European American	1:500,000	Sheffield 1998, Nino 2008, Franco 1995, Matos-Miranda 2013

VARIANT INTERPRETATION: The Gly137Ala variant in ARSE has been previously identified in 2 males with chondrodysplasia punctata; however, this variant was also identified in one unaffected male family member (Sheffield 1998, Nino 2008). Variants in a paralogous gene (ARSB) at the same position have also been identified in an individual with Maroteux-Lamy syndrome, which also features skeletal abnormalities (Franco 1995). Functional studies indicate that the Gly137Ala variant leads to reduced ARSE activity (Matos-Miranda 2013). In summary, although some data support a disease-causing role, there is currently insufficient evidence for pathogenicity leading to a current classification of uncertain significance.

DISEASE INFORMATION: X-linked chondrodysplasia punctata 1 (CDPX1), a congenital disorder of bone and cartilage development, is caused by a deficiency of the Golgi enzyme arylsulfatase E (ARSE). It is characterized by chondrodysplasia punctata (stippled epiphyses), brachytelephalangy (shortening of the distal phalanges), and nasomaxillary hypoplasia. Although most affected males have minimal morbidity and skeletal findings that improve by adulthood, some have significant medical problems including respiratory compromise, cervical spine stenosis and instability, mixed conductive and sensorineural hearing loss, and abnormal cognitive development. From GeneReviews abstract: http://www.ncbi.nlm.nih.gov/books/NBK1544/

FAMILIAL RISK: X-Linked chondrodysplasia punctata is inherited in an X-linked recessive manner, with primarily males being affected. Each child is at a 50% (or 1 in 2) chance of inheriting the variant from a carrier female, while all daughters will inherit the variant from an affected male.

CARRIER RISK

Disease (Inheritance)	Gene (Transcript)	Variant (Classification)	Variant Frequency	Disease Prevalence (Carrier Freq.)	References	Carrier Phenotype
Cystic fibosis (Autosomal recessive)	CFTR (NM_000492.3)	c.3846G>A p.Trp1282X heterozygous (Pathogenic)	6/8600 (0.07%) European American	1/3200 European American (1/25)	Hamosh 1991, Kerem 1990, Shoshani 1992, Vidaud 1990	None Reported

VARIANT INTERPRETATION: The Trp1282X variant in CFTR has been identified in numerous patients with cystic fibrosis (Vidaud 1990, Kerem 1990, Hamosh 1991, Shoshani 1992). This variant is present on the American Board of Medical Genetics CFTR mutation panel (http://www.acmg.net/Pages/ACMG_Activities/stds-2002/cf.htm). This nonsense variant leads to a premature termination codon at position 1282, which is predicted to lead to a truncated or absent protein. In summary, this variant meets our criteria for pathogenicity.

DISEASE INFORMATION: Cystic fibrosis effects the epithelia of the respiratory tract, exocrine pancreas, intestine, male genital tract, hepatobiliary system, and exocrine sweat glands, resulting in a complex multisystem disease. Pulmonary disease is the major cause of morbidity and mortality in CF. Affected individuals have lower airway inflammation and chronic endobronchial infection, progressing to end-stage lung disease characterized by extensive airway damage (bronchiectasis, cysts, and abscesses) and fibrosis of lung parenchyma. Meconium ileus occurs at birth in 15%-20% of newborns with CF. Pancreatic insufficiency with malabsorption occurs in the great majority of individuals with CF. More than 95% of males with CF are infertile as a result of azoospermia caused by absent, atrophic, or fibrotic Wolffian duct structures. Adapted from GeneReviews abstract: http://www.ncbi.nlm.nih.gov/books/NBK1250/

FAMILIAL RISK: Cystic fibrosis is inherited in an autosomal recessive manner. A carrier of cystic fibrosis has a 50% chance of passing on the CFTR variant to any children. The risk of this patient's child having cystic fibrosis is dependent on the CFTR carrier status of the patient's partner. This patient likely inherited the CFTR variant from one of his parents. Other biologically related family members may also be carriers of this variant.

Next Generation Sequencing

Concluding

Concluding

- Next generation sequencing poses **new opportunities and challenges** for clinical medicine.

- NGS-based studies have made immense contribution towards discovering many **novel and clinical relevant biomarkers**

- Most important contribution: the capability to decipher individual cancer genomes and truly unleash the potential of **high-precision medicine**

- However, there is a need for specific training, and close collaboration with **bioinformaticians**, which are able to correctly use computational means.

Concluding

- In the near future, NGS technologies will play crucial role in cancer diagnosis, prognosis and disease management.

- **Costs are continuously decreasing** and NGS technologies are becoming more precise end less error prone

- As training increase, NGS will be used continuously in common medical practice, and will become a typical analysis performed to all patients as a deep screening methodology.

Future Directions

- Longer READS (Next-Next Generation Sequencing)

- Single molecule Sequencing (Fourth Generation Sequencing)


Any
Questions?


The End