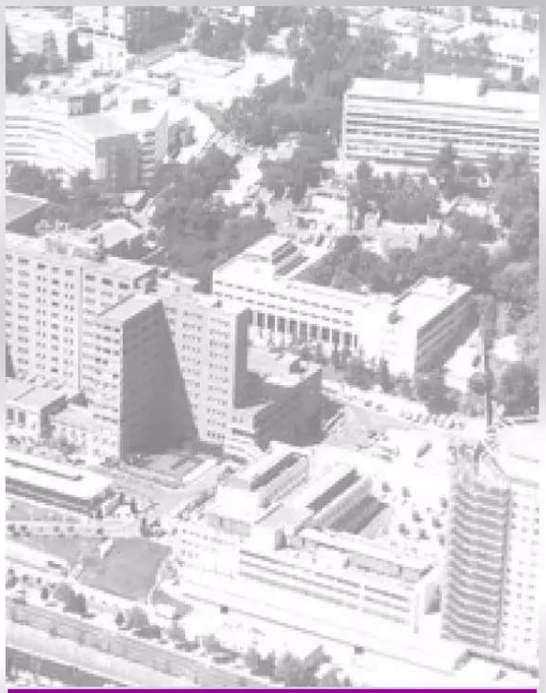
SUMMER TRAINING PROGRAMME (STP) ON INTRODUCTION TO COMPUTATIONAL BIOLOGY

INTRODUCTION TO NGS VARIANT CALLING ANALYSIS

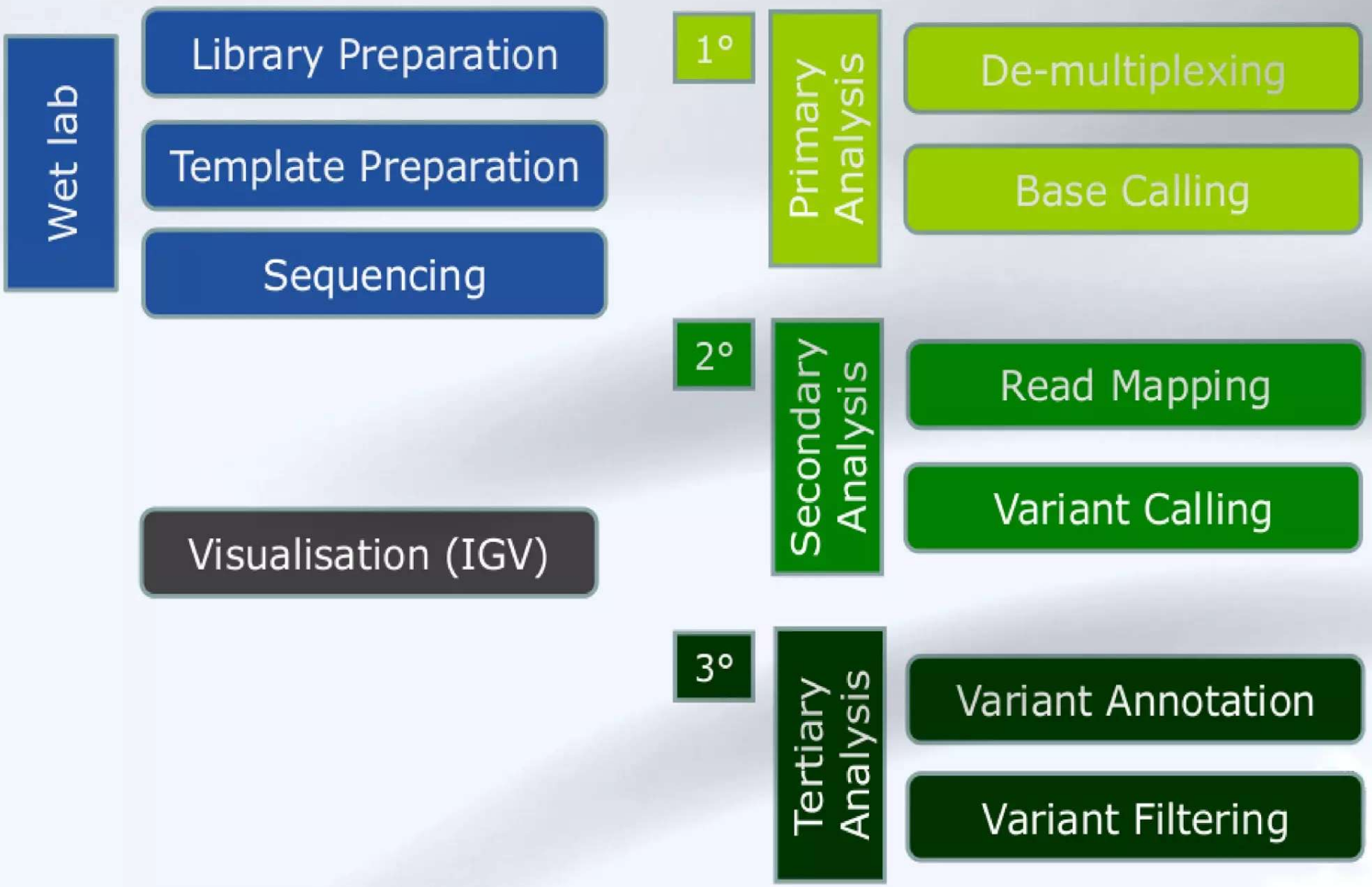
Instructor
Dr. Manisha Aswal
Bioinformatics Scientist-II
Elucidata, New Delhi
India

PRESENTATION OUTLINE

1. **NGS WORKFLOW OVERVIEW**
2. **WET LAB STEPS**
3. **IMPORTANT SEQUENCING CONCEPTS**
4. **NGS ANALYSIS WORKFLOW**
    1. Primary analysis: de-multiplexing, QC
    2. Secondary analysis: read mapping and variant calling
    3. Tertiary analysis: annotation, filtering...
5. **VISUALIZATION**
6. **COMMON PIPELINES AND FORMATS**
7. **CONCLUSIONS**



1 NGS WORKFLOW OVERVIEW



2 LIBRARY PREPARATION

Select target

Hybridization-based cature or PCR

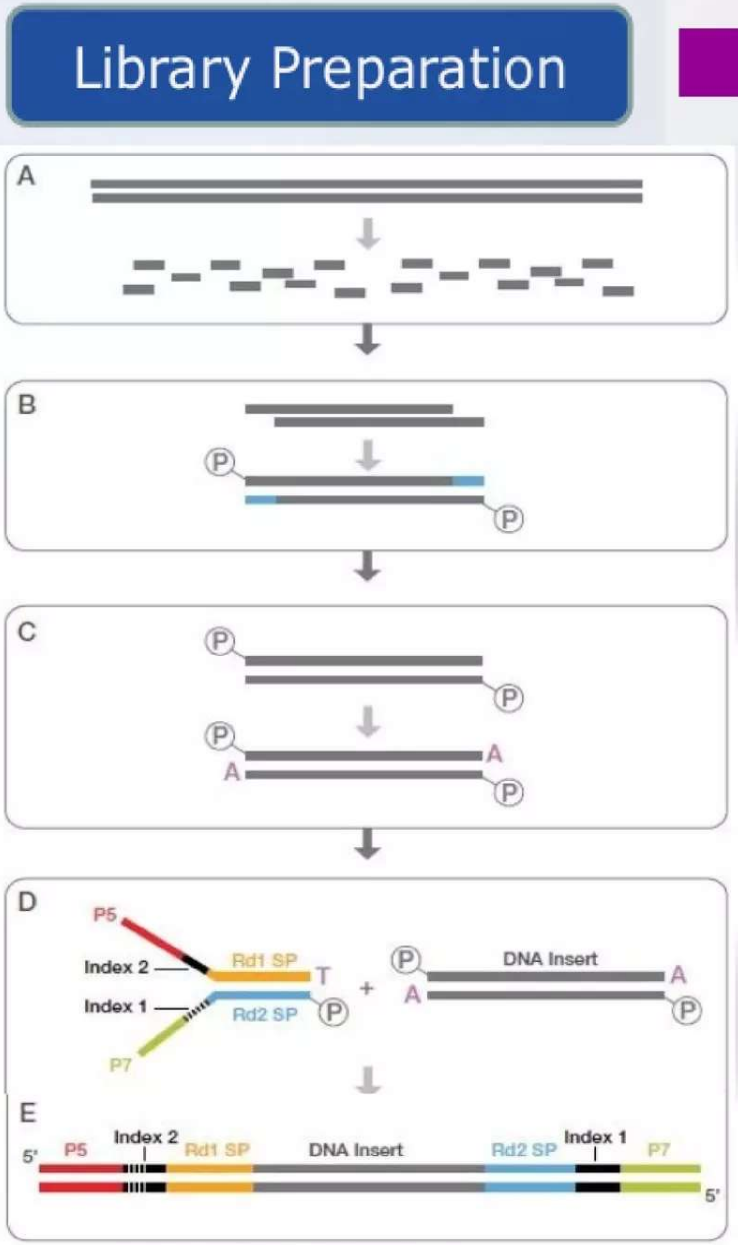
Add adapters

Contain binding sequences
Barcodes
Primer sequences

Amplify material

2 LIBRARY PREPARATION

Wet lab



Select target

Hybridization-based cature or PCR

Add adapters

Contain binding sequences
Barcodes
Primer sequences

Amplify material

A) Fragment DNA

B) End-repair

C) A-tailing, adapter ligation and PCR

D) Final library contains

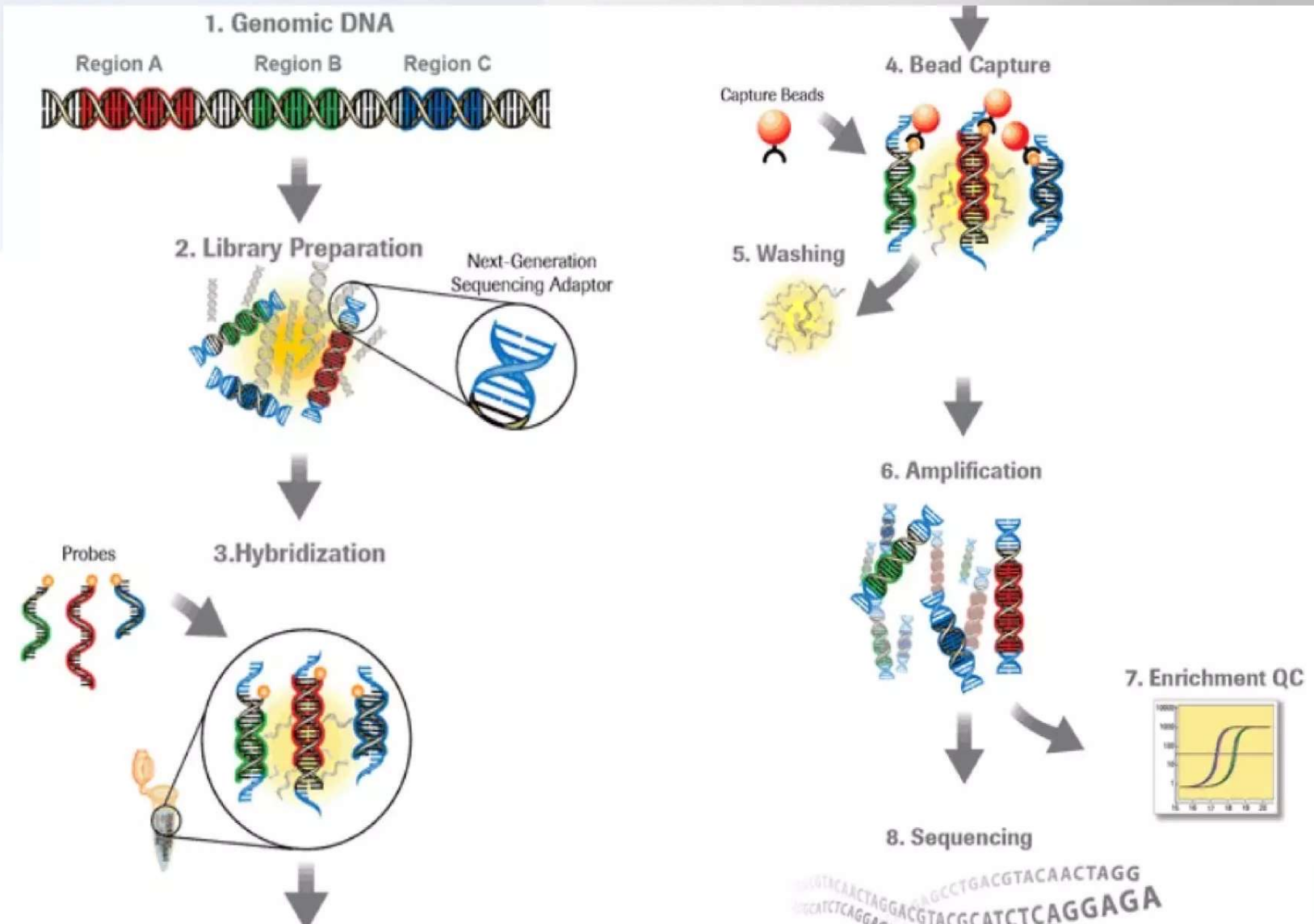
- sample insert
- indices (barcodes)
- flowcell binding sequences
- primer binding sequences

2 LIBRARY PREPARATION

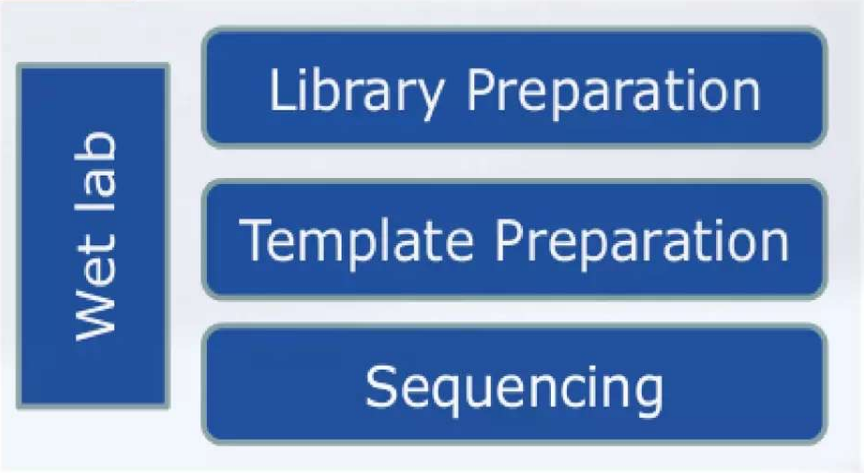
Wet lab

Library Preparation

Select target
Hybridization-based cature or PCR



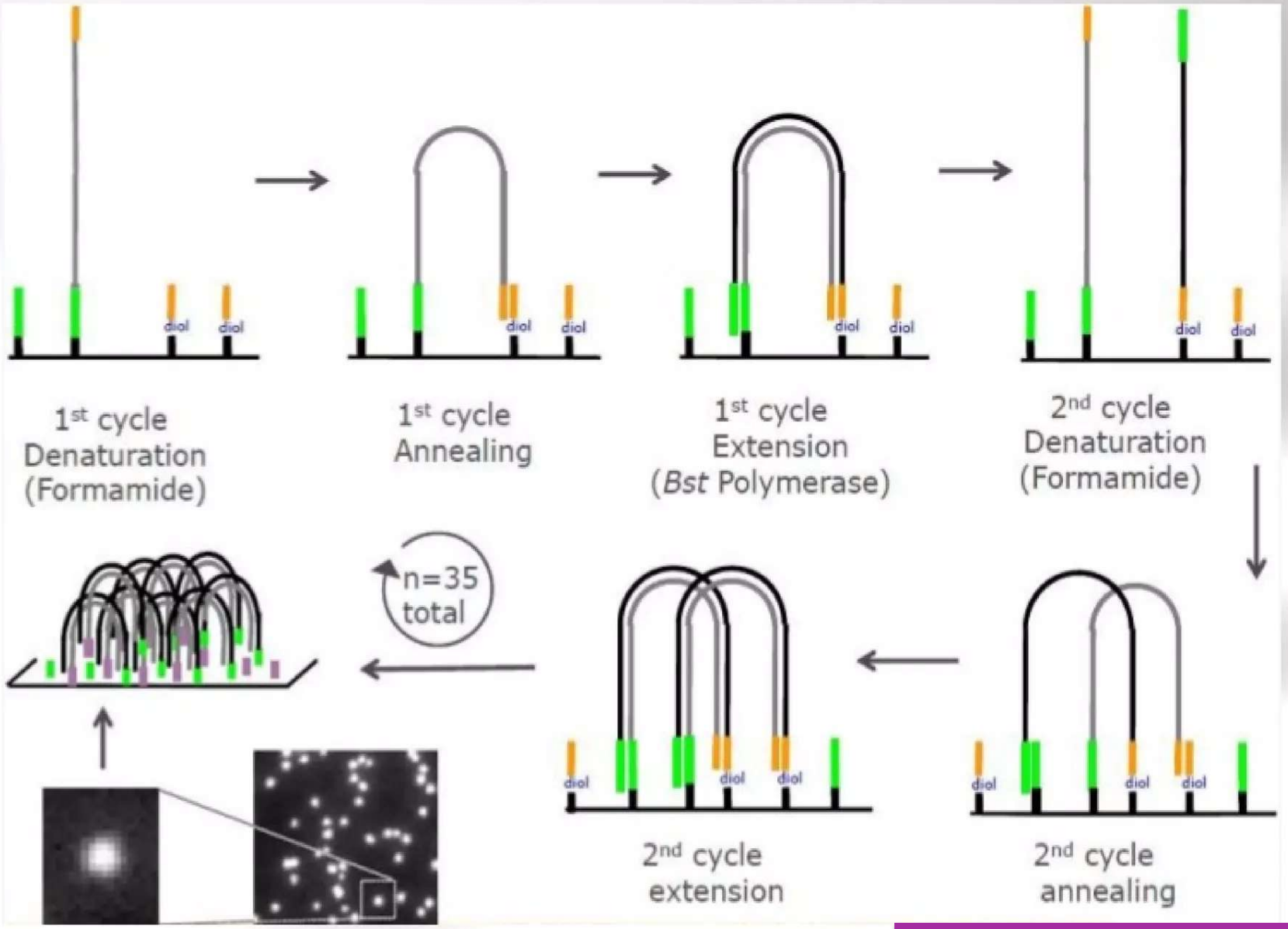
2 TEMPLATE PREPARATION



Attachment of library
e.g. To Illumina Flowcell

Amplification of library molecules
e.g. Brigde amplification

2 BRIDGE AMPLIFICATION



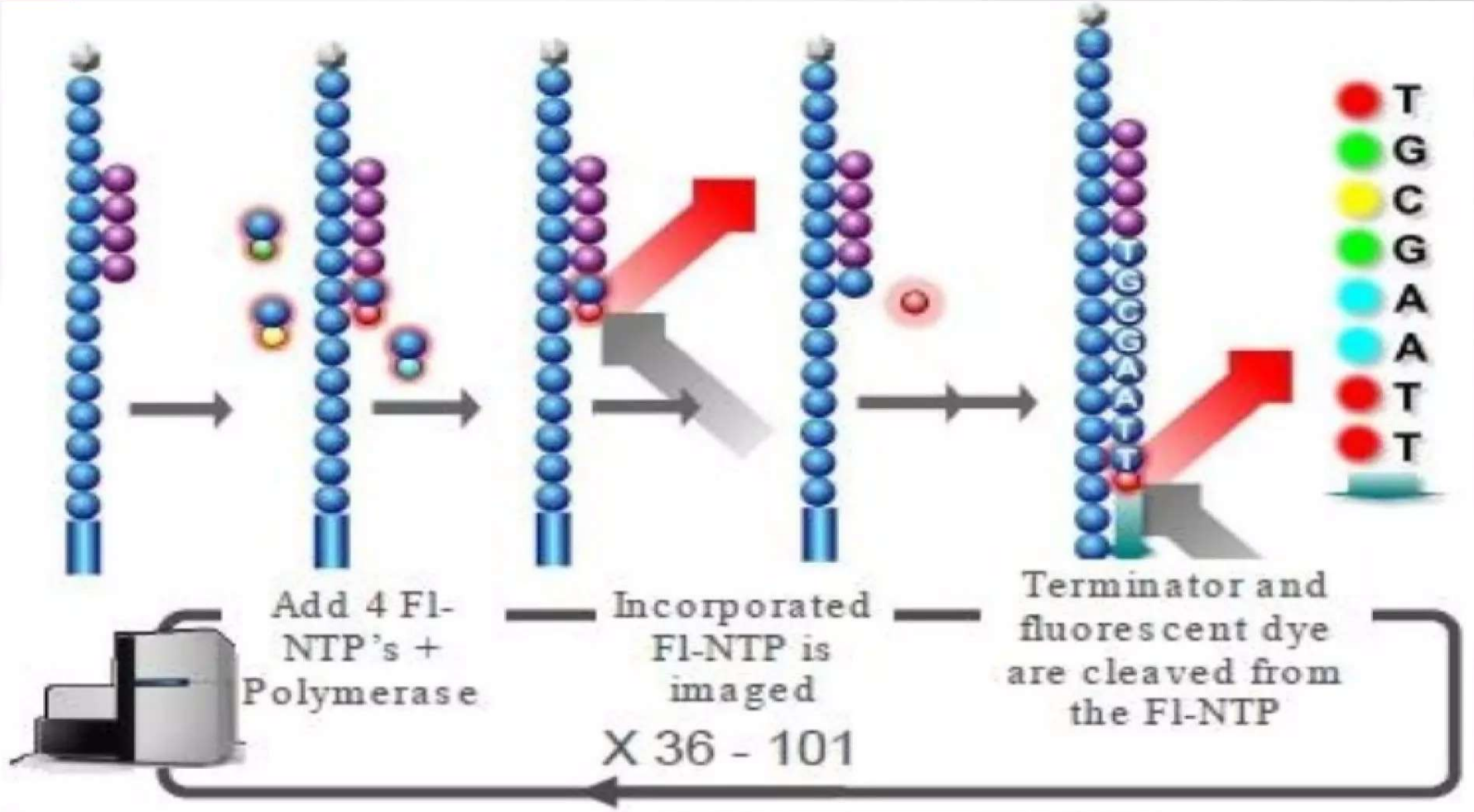
2 SEQUENCING

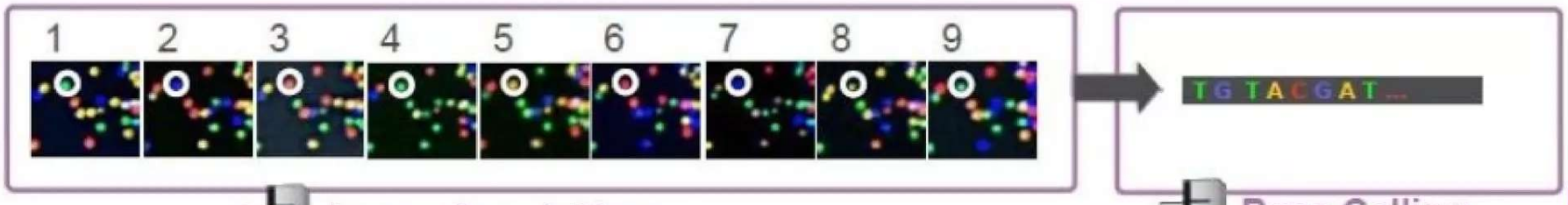
Sequencing-by-Synthesis

Detection by:

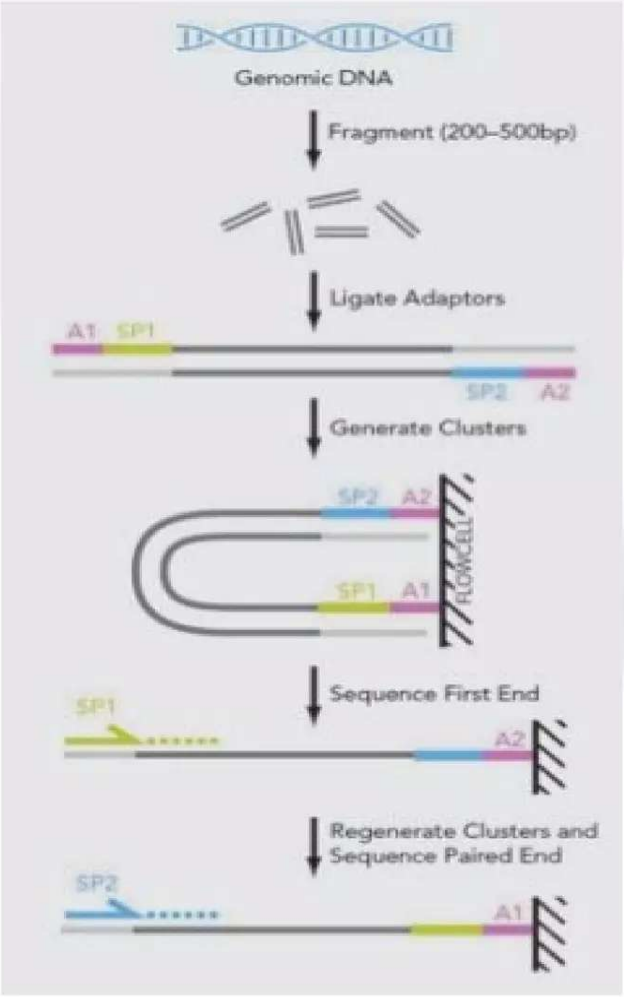
- Illumina – fluorescence
- Ion Torrent – pH
- ROCHE 454 – PO_4 and light

2 SEQUENCING-BY-SYNTHESIS (ILLUMINA)





3 IMPORTANT SEQUENCING CONCEPTS



Barcoding/Indexing:
allows multiplexing of different samples

Single-end vs paired-end sequencing

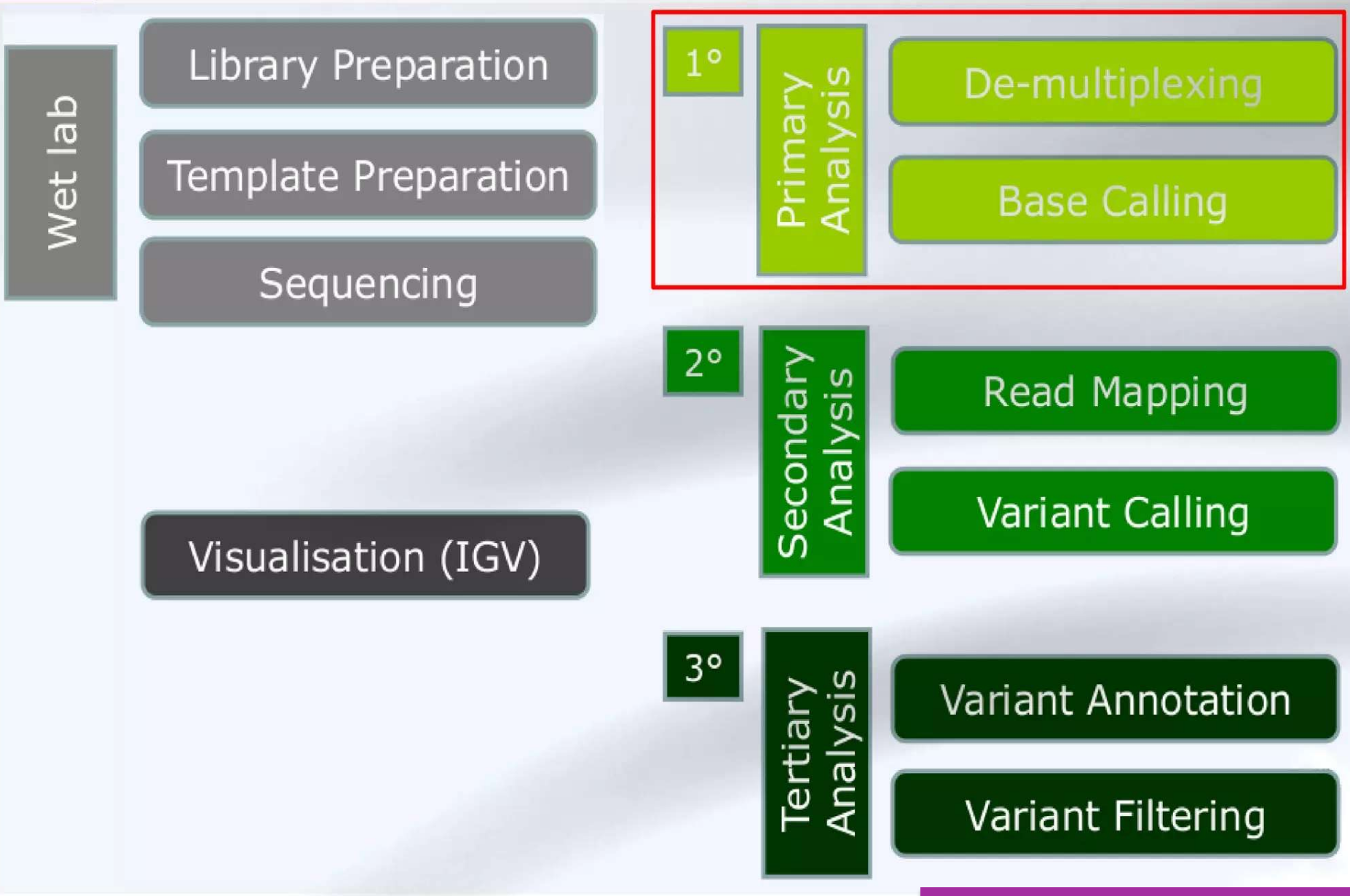
Coverage: avg. number reads per target

Quality scores (Qscore): log-scales!

Quality Score	Probability of a wrong base call	Accuracy of a base call
Q 10	1 in 10	90%
Q 20	1 in 100	99%
Q 30	**1 in 1000**	**99.90%**
Q 40	1 in 10000	99.99%
Q 50	1 in 100000	100.00%

```
fragment            ====================
Single read         ------->
Paired-end reads    R1------->          <------R2
```

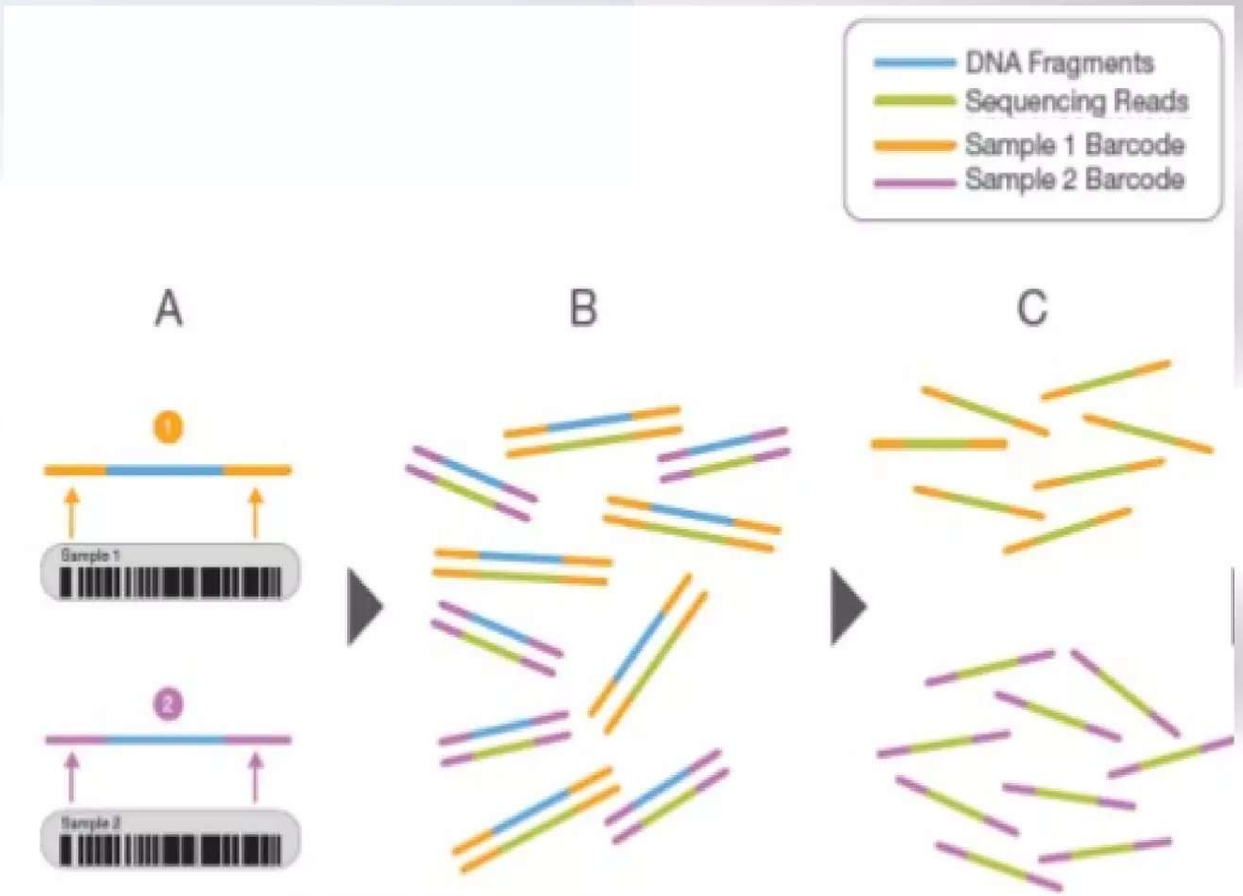
4 NGS DATA ANALYSIS WORKFLOW



4 DE-MULTIPLEXING (BARCODE SPLITTING)

1° Primary Analysis

De-multiplexing



4 FASTQ FORMAT

Read Identifier	@D3NZ4HQ1:111:D2DM2ACXX:1:1101:1243:2110 2:N:0:TGACCA
Sequence	GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTC
+	+
Error probability (quality)	!''*(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CC

!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ

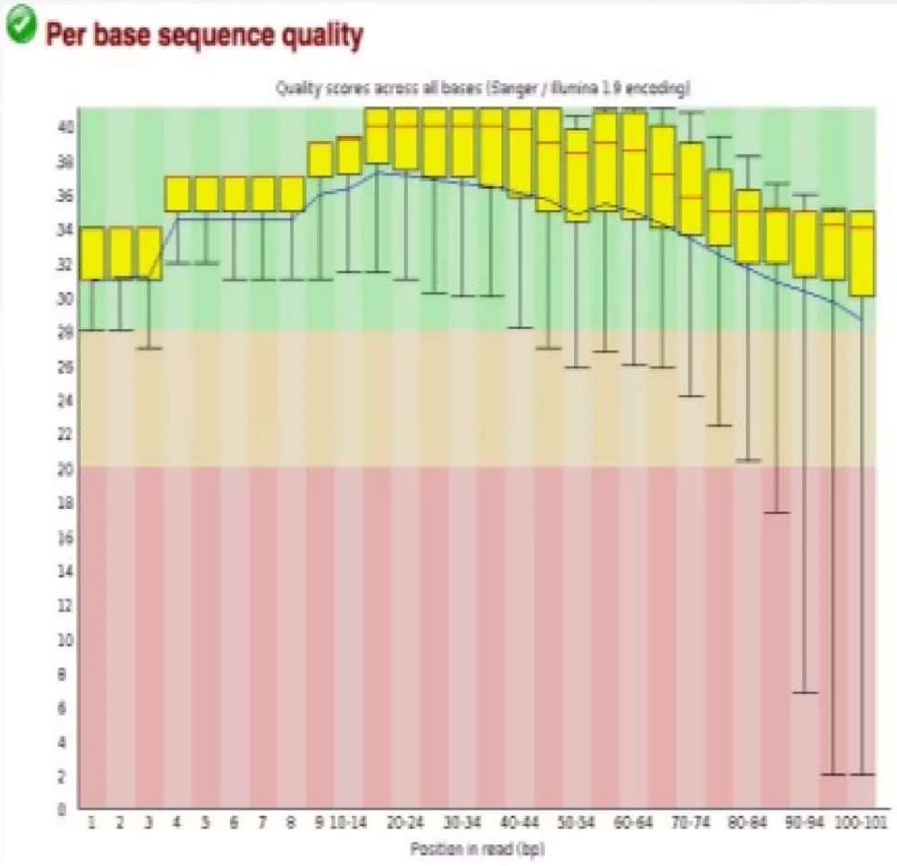
Phred score 0.....41

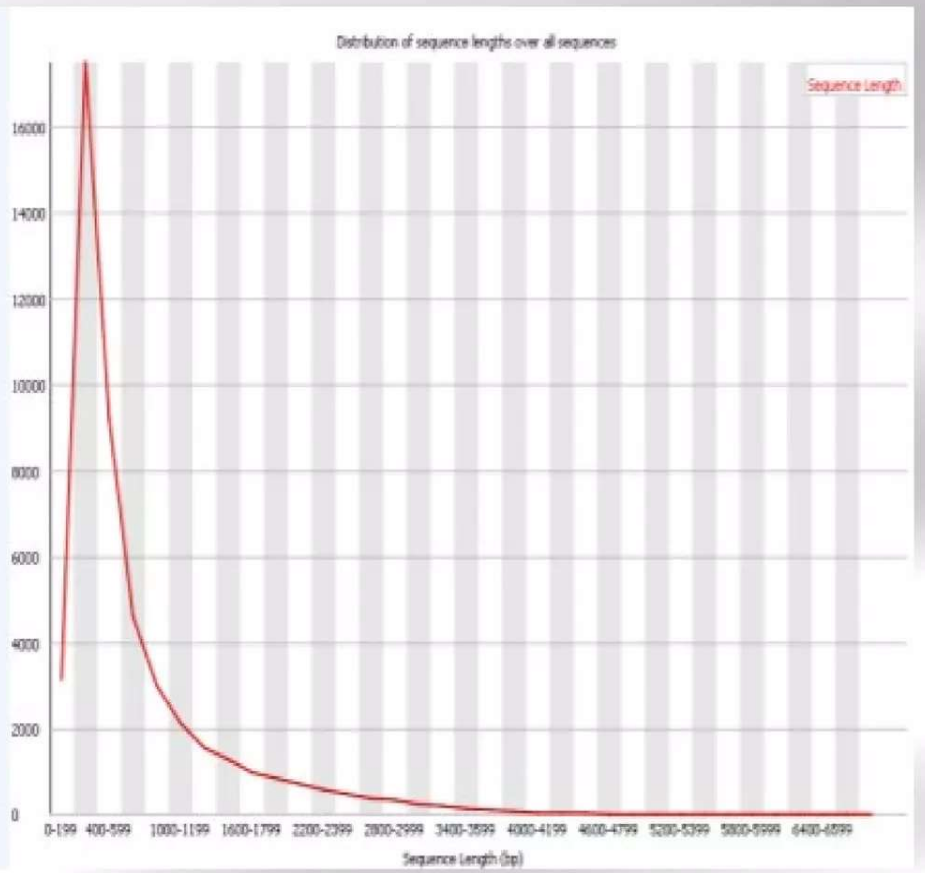
Probability 1.....0.0001

Phred score = $-10 \log_{10} P$

see en.wikipedia.org/wiki/FASTQ_format

4 SEQUENCE QUALITY: fastQC



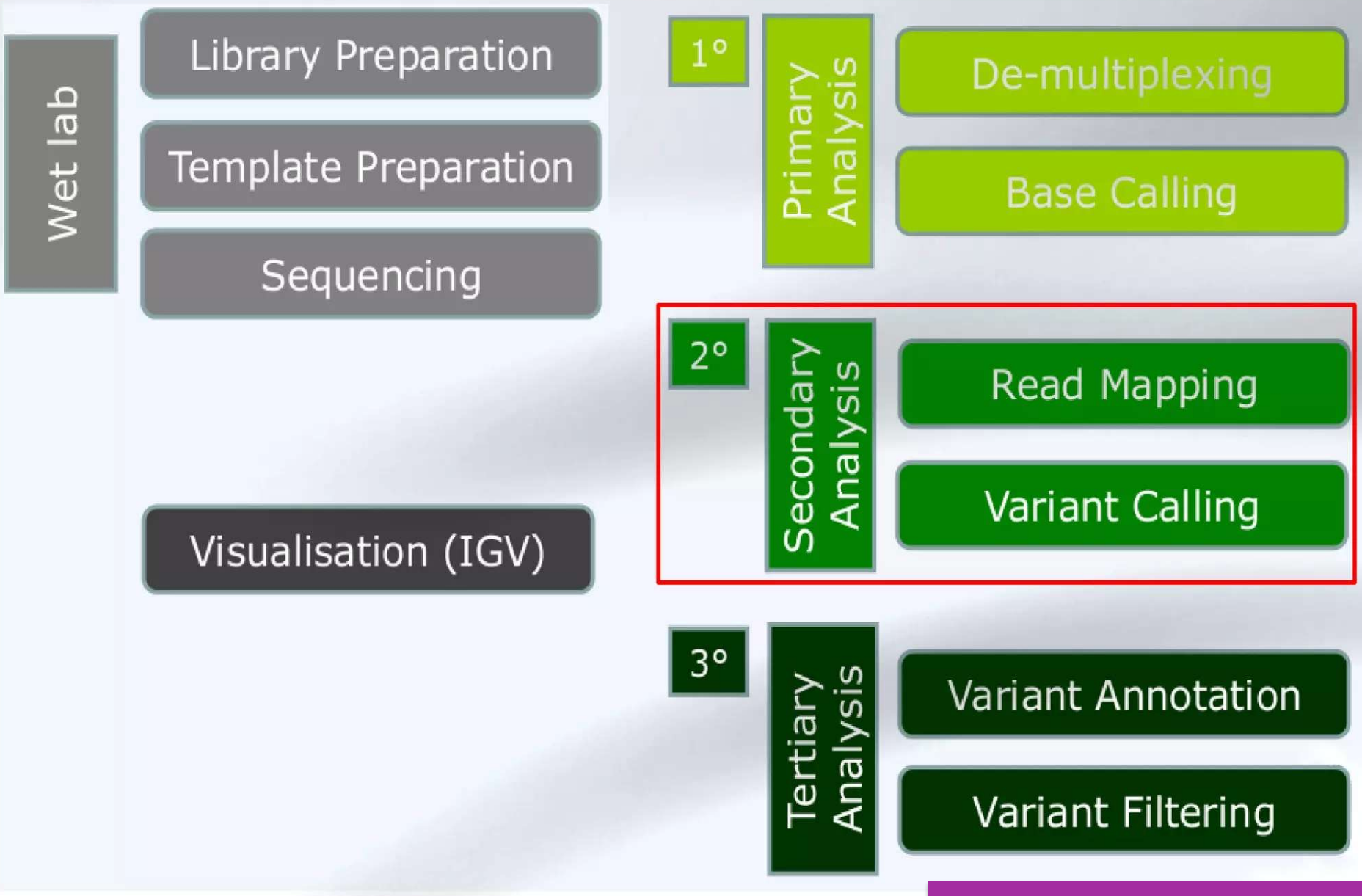


Babraham Bioinformatics

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

Details of the output https://docs.google.com/document/pub?id=16GwPmwYW7o_r-ZUgCu8-oSBBY1gC97TfTTinGDk98Ws

4 NGS DATA ANALYSIS WORKFLOW



4 READ MAPPING (BASIC ALIGNMENT)

Comparison against reference genome
(! not assembly !)

Many aligners
(short reads, longer reads, RNAseq...)
Examples: BWA, Bowtie

SAM/BAM files

4 SAM/BAM FILES

Suppose we have the following alignment with bases in lower cases clipped from the alignment. Read r001/1 and r001/2 constitute a read pair; r003 is a chimeric read; r004 represents a split alignment.

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                      ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                         CAGCGGCAT
```

The corresponding SAM format is:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001  163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001   83 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

see http://samtools.sourceforge.net/SAMv1.pdf

4 SAM/BAM FILES

@ Header (information regarding reference genome, alignment method...)

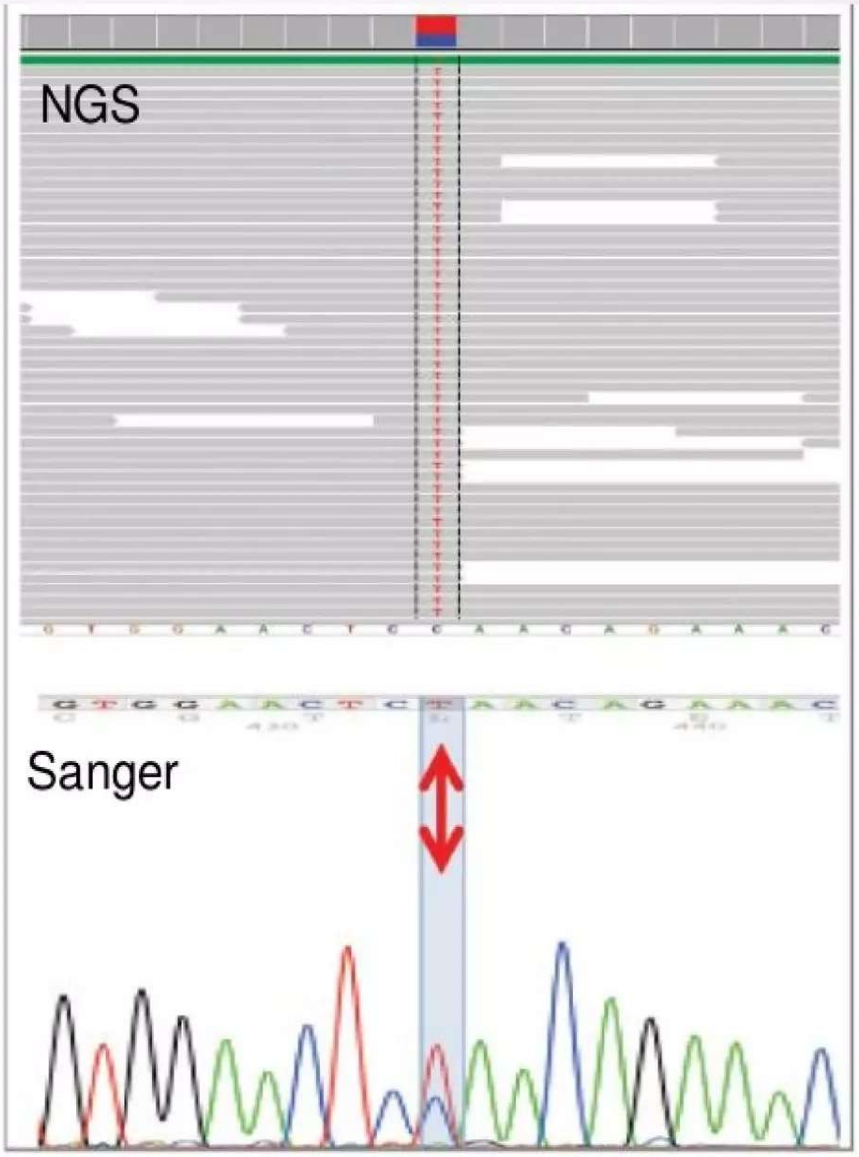
1) Read ID (QNAME)
2) Bitwise FLAG (first/second read in pair, both reads mapped...)
3) ReferenceSequence Name (RNAME)
4) Position (POS, coordinate)
5) MapQuality (MAPQ = -10log10P[wrong mapping position])
6) CIGAR (describes alignment – matches, skipped regions, insertions..)
7) ReferenceSequence (RNEXT, Ref seq of the pair)
8) Position of the pair (PNEXT)
9) TemplateLength (TLEN)
10) ReadSequence
11) QUAL (in Fastq format, '*' if NA)
...

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001  163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001   83 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

4 VARIANT CALLING

Identify sequence variants
Distinguish signal vs noise
VCF files
Examples: SAMtools, SNVmix

4 SEQUENCE VARIANTS

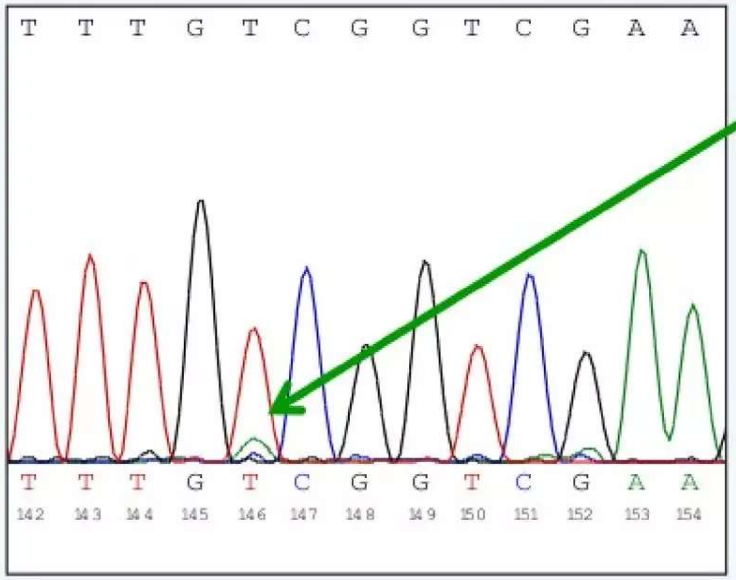


Differences to the reference

Reference: C

Sample: C/T

4 SEQUENCE VARIANTS



Sanger: is it real??

Total count: 204
A : 185 (91%, 92+, 93-)
C : 0
G : 1 (0%, 0+, 1-)
T : 18 (9%, 12+, 6-)
N : 0

NGS: read count

Provides confidence (statistics!)

Sensitivity tune-able parameter
(dependent on coverage)

4 VARIANT CALLING: GATK

Genome Analysis Toolkit (BROAD Institute)

- **Initially developed for 1000 Genomes Project**
- **Single or multiple sample analysis (cohort)**
- **Popular tool for germline variant calling**
- **Evaluates probability of genotype given read data**

see http://www.broadinstitute.org/gatk/
and McKenna et al. Genome Research 2010

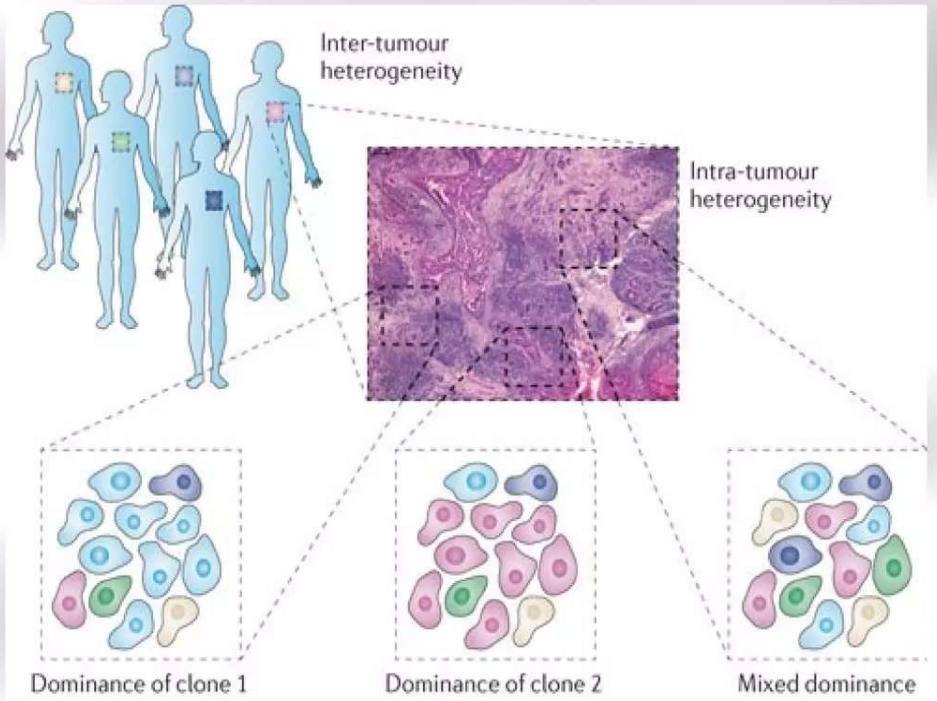
4 SOMATIC VARIANT CALLING

Somatic mutations can occur at low freq. (<10%) due to:

- **Tumor heterogeneity (multiple clones)**
- **Low tumor purity (% normal cells in tumor sample)**

Requires different thresholds than germline variant calling when evaluating signal vs noise

Trade-off between sensitivity (ability to detect mutation) and specificity (rate of false positives)



Nature Reviews Cancer 12, 323-334 (May 2012)

4 INDELS DETECTION

Small insertions/ deletions

The trouble with mapping approaches

coor	12345678901234	5678901234567890123456
ref	aggttttataaaac	----aattaagtctacagagcaacta
sample	aggttttataaaac	AAATaattaagtctacagagcaacta

Insertion AAAT in our sample!!!

4 INDELS DETECTION

Small insertions/ deletions

The trouble with mapping approaches

```
coor      12345678901234     5678901234567890123456
ref       aggttttataaaac----aattaagtctacagagcaacta
sample    aggttttataaaacAAATaattaagtctacagagcaacta
read1     aggttttataaaac    aaAtaa
read2      ggttttataaaac    aaAtaaTt
read3          ttataaaacAAATaattaagtctaca
read4             CaaaT     aattaagtctacagagc
read5              aaT      aattaagtctacagagc
read6               T       aattaagtctacagagc
```

Insertion AAAT in our

By default, aligners prefer placing reads w/ a mismatch than with an insertion, esp. at ends of read!!

4 INDELS DETECTION

Small insertions/ deletions

The trouble with mapping approaches

Insertion AAAT in our

By default, aligners prefer placing reads w/ a mismatch than with

Information from other reads can be used to improve alignment; After local realignment the insertion has been correctly placed!!

4 EVALUATING VARIANT QUALITY

TAKING INTO ACCOUNT:

- **Coverage at position**
- **Number independent reads supporting variant**
- **Observed allele fraction vs expected (somatic / germline)**
- **Strand bias**
- **Base qualities at variant position**
- **Mapping qualities of reads supporting variant**
- **Variant position within reads (near ends or at centre)**

4 VCF FILES

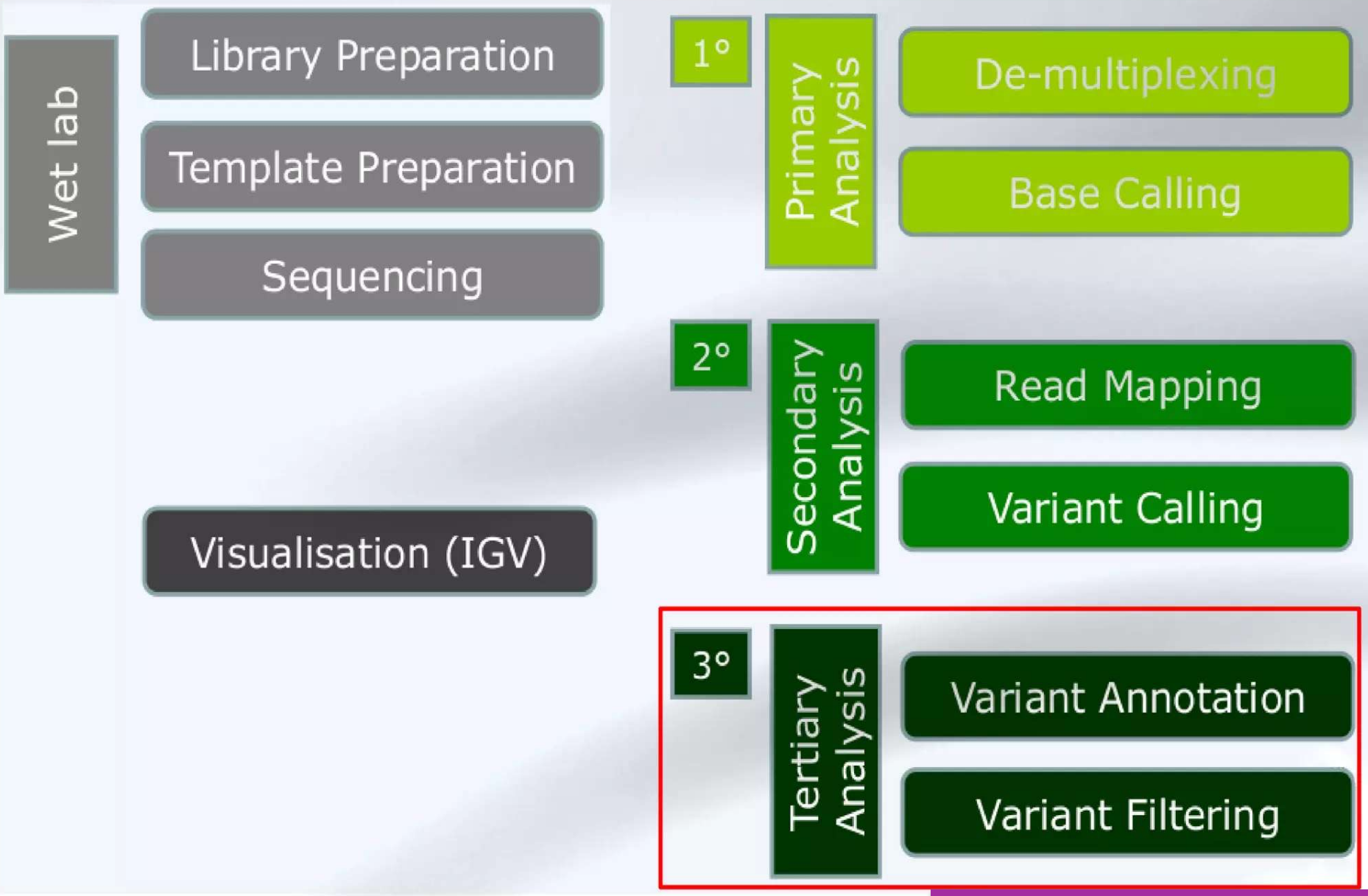
```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myProgramV3
##reference=file:///seq/NCBI36.fasta
...
#CHROM  POS     ID          REF  ALT  QUAL  FILTER
20      14370   rs6054257   G    A    29    PASS
20      17330   .           T    A    3     q10

INFO                              FORMAT      SAMPLE1   ...
NS=2;DP=14;AF=0.5;DB;H2           GT:GQ:DP    1|0:48:8
NS=2;DP=11;AF=0.017               GT:GQ:DP    0|0:49:3
```

Header lines (marked by ##): Metadata of analysis

Data lines: Individual variant calls

GT: genotype: 1|0 het, 0|0 hom
DP: read depth

4 NGS DATA ANALYSIS WORKFLOW

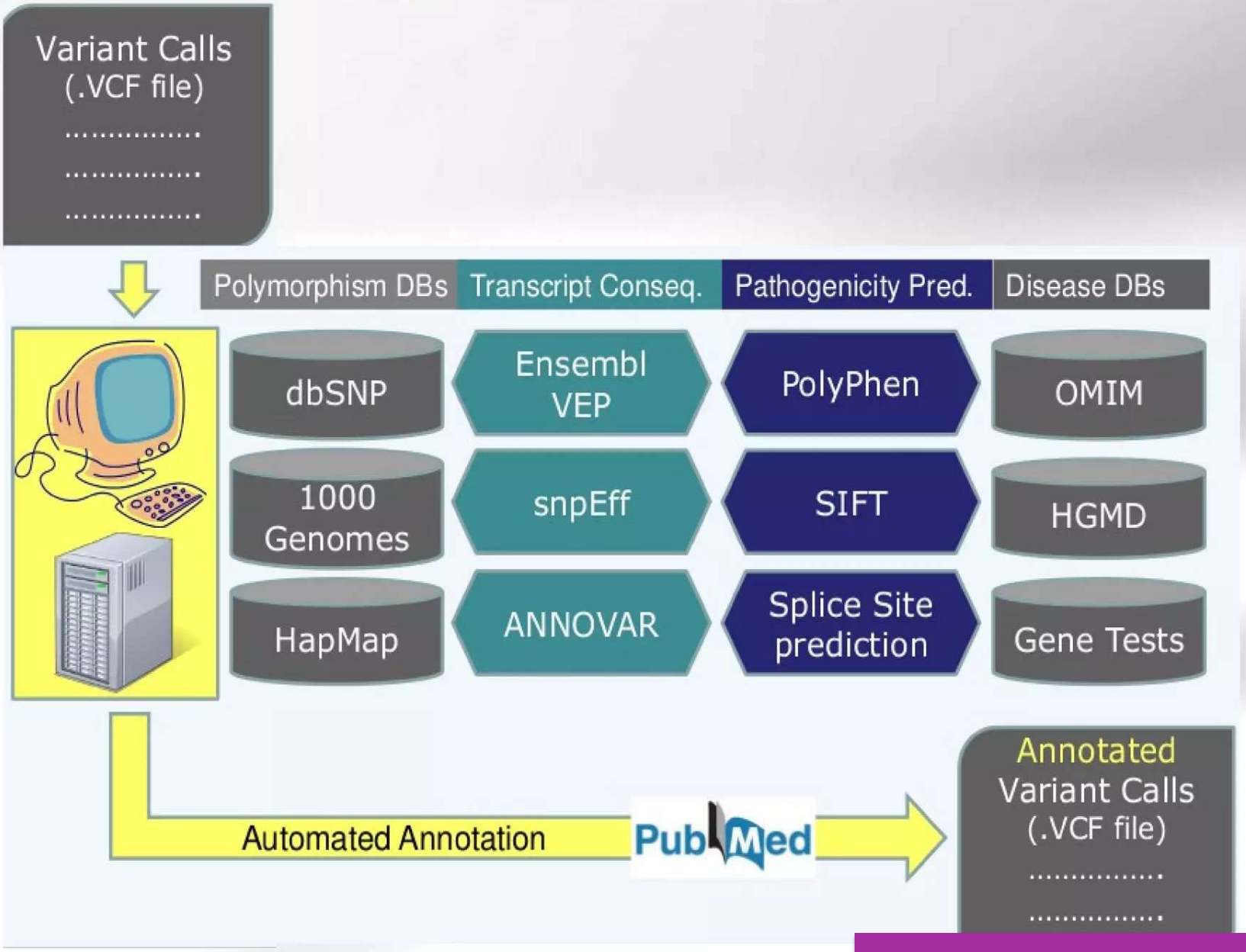


4 VARIANT ANNOTATION

Provide biological & clinical context

Identify disease-causing mutations (among 1000s of variants)

4 ANNOTATION OVERVIEW



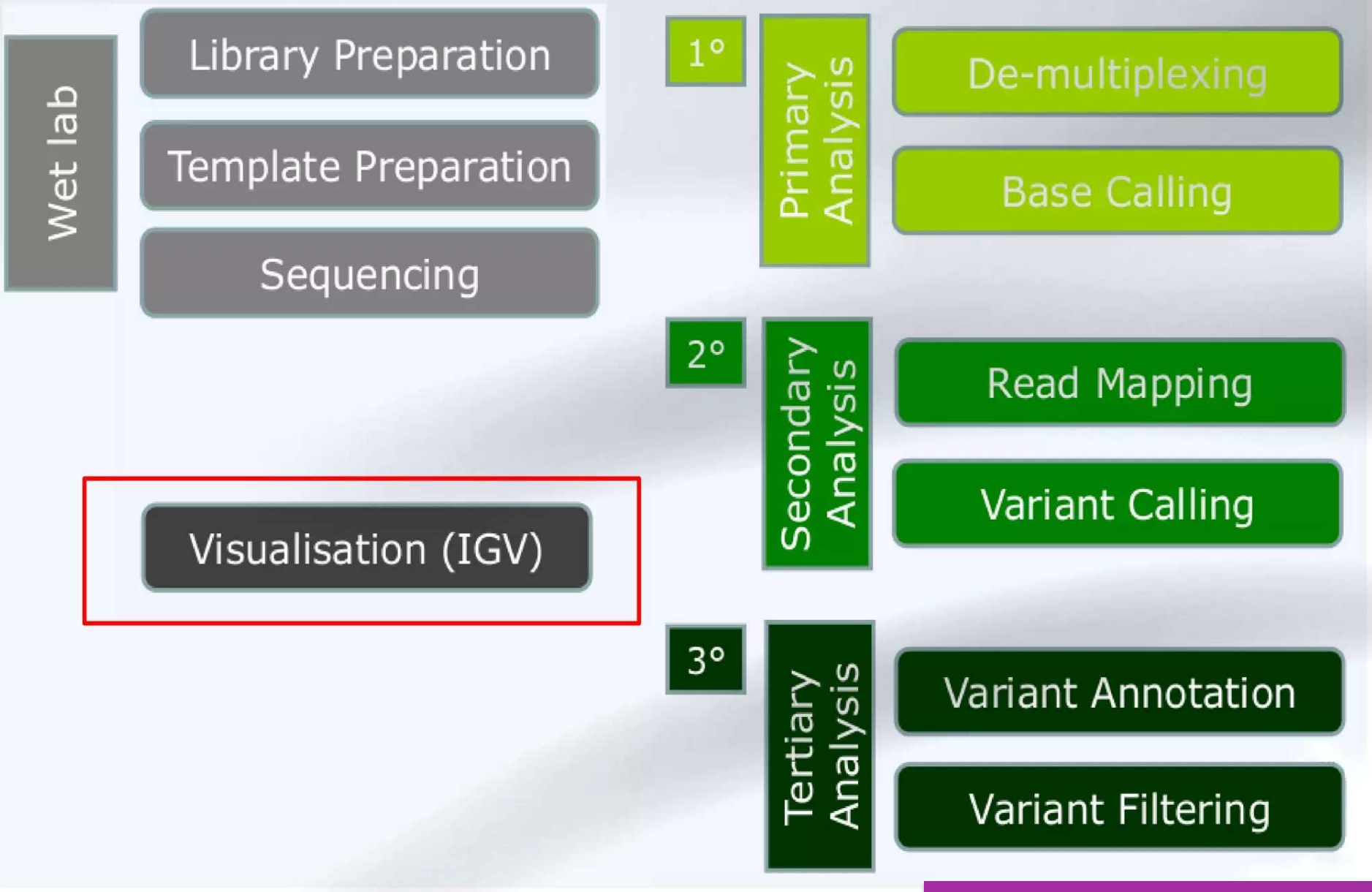
4 VARIANT FILTERING AND PRIORIZATION

PURPOSE:
Identify pathogenic or disease-associated mutation(s)
Reduce candidate variants to reportable set

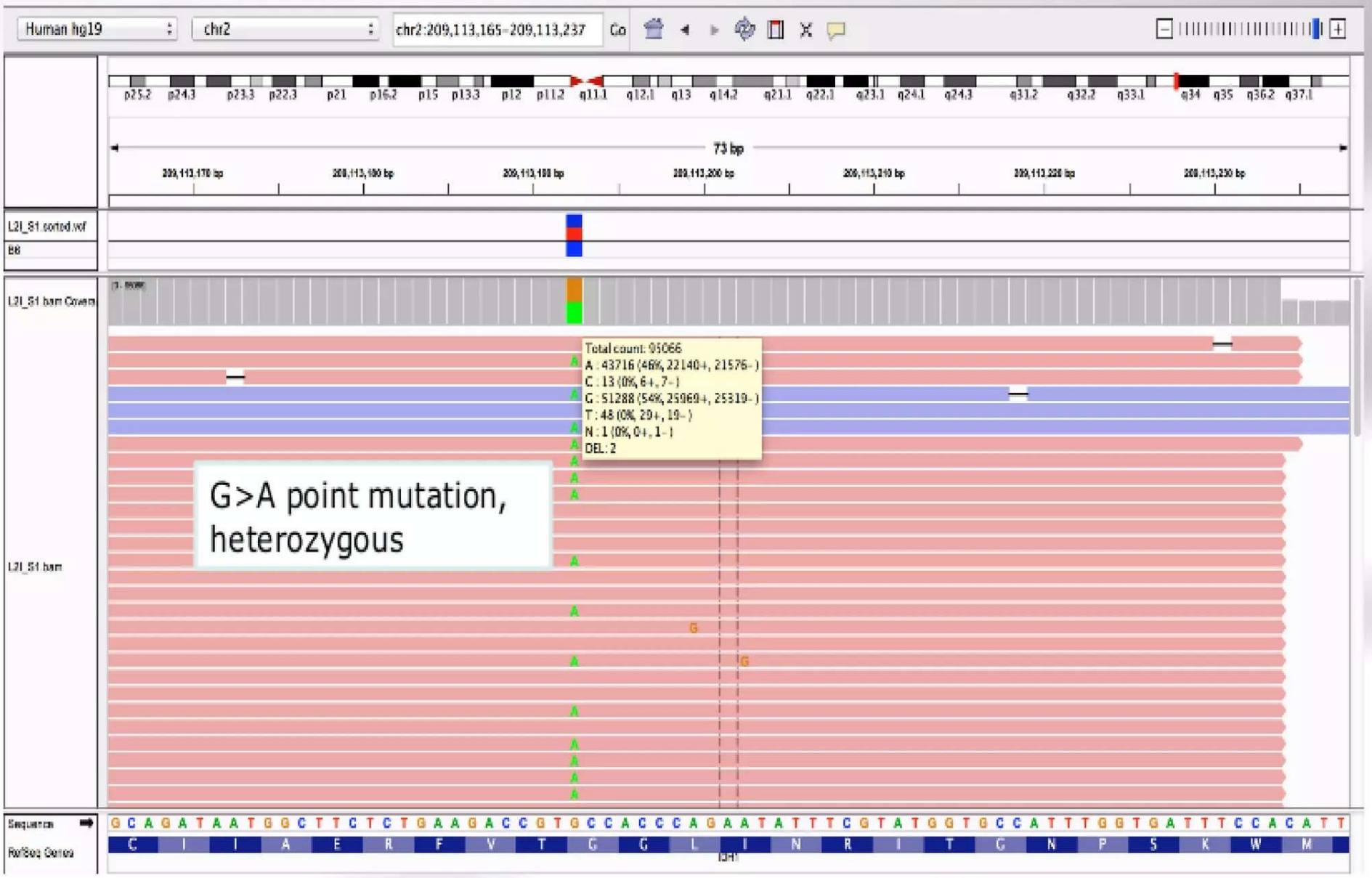
COMMON STEPS:

- **Remove poor quality variant calls**
- **Remove common polymorphisms**
- **Prioritize variants with high functional impact**
- **Compare against known disease genes**
- **Consider mode of inheritance (autosomal recessive, X-linked...)**
- **Consider segregation in family (where multiple samples available)**

5 NGS DATA ANALYSIS WORKFLOW



5 VISUALIZATION – IGV (or Genome Browser, Circos...)



6 COMMON PIPELINE

bcl2fastq (Illumina)
FastQC (open-source)

Exomes (HiSeq):
BWA(open-source), GATK (Broad)

Gene panels (MiSeq, PGM):
MiSeq Reporter (Illumina)
Torrent Suite (Ion Torrent)

Custom scripts and third party tools
(Annovar, snpEff, PolyPhen, SIFT...)

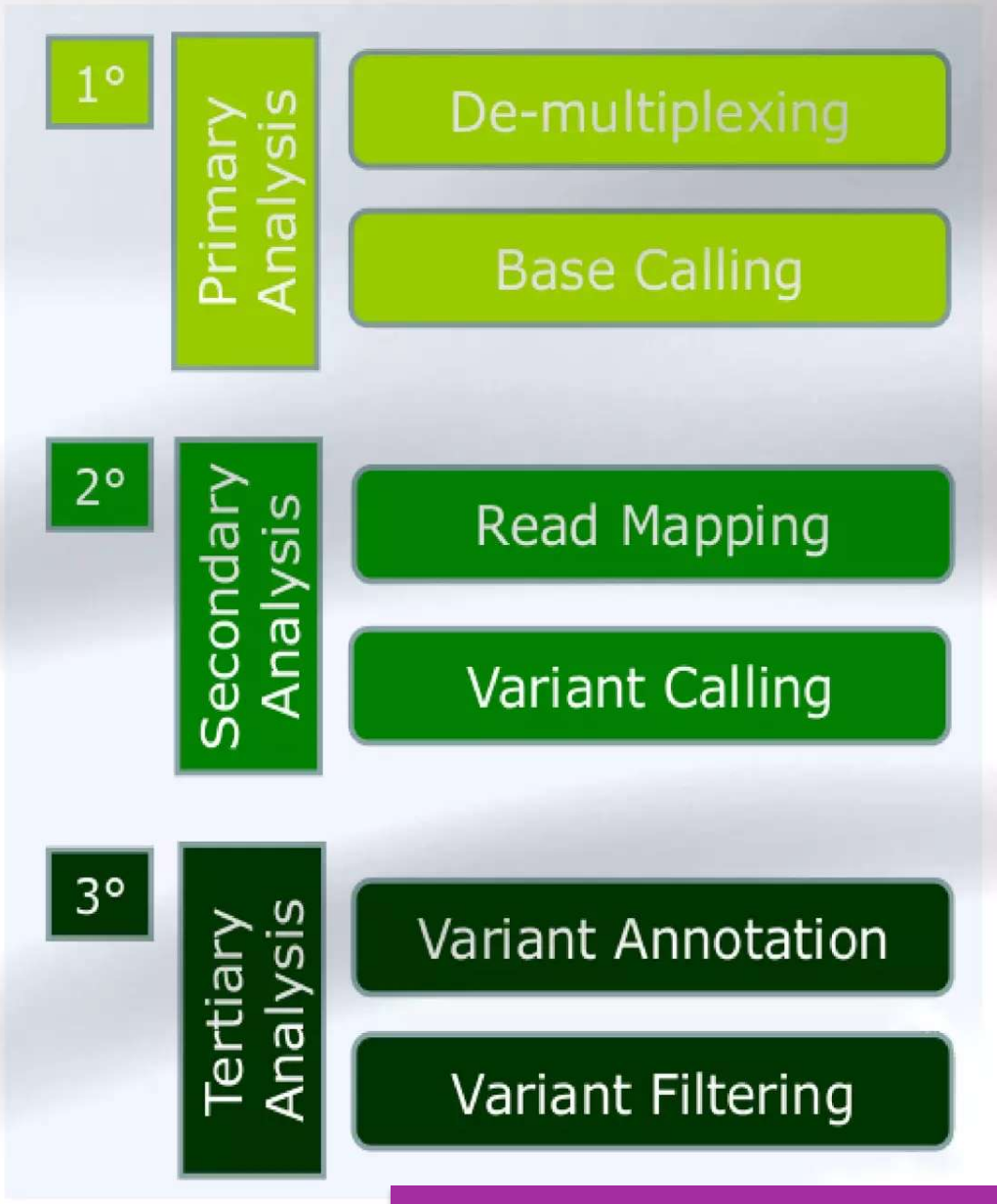
Commercial annotation software
(GeneticistAssistant, VariantStudio...)

6 COMMON DATA FORMATS

.bcl
.fastq

.BAM
.VCF

.csv
.txt
.xls
.html
...



7 CONCLUSIONS

NGS data - the new currency of (molecular) biology

Broad applications (ecology, evolution, ag sciences, medical research and clinical diagnostics...).

Rapidly evolving (sequencing technologies, library preparation methods, analysis approaches, software).

Different tools/pipelines/parametrization gives different results, (more standards needed).

Bioinformatics pipelines typically combine vendor software, third-party tools and custom scripts.

Requires skills in scripting, Linux/Unix, HPC.

Requires advanced hardware (not always available).

Understanding of data (SE, PE, RNA-Seq) important for successful analysis.