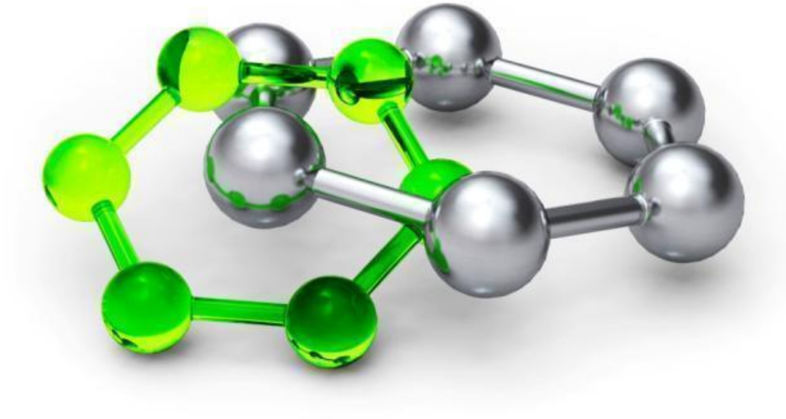
CEITEC

Central European Institute of Technology
BRNO | CZECH REPUBLIC

Moderní metody analýzy genomu - analýza

Mgr. Nikola Tom

Brno, 20.11.2015



EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND
INVESTING IN YOUR FUTURE

2007–13
OP Research and
Development for Innovation

Before we start analysis

We have to know what we are dealing with… and what we want to find out…

Concept of the project
DNA/RNA/methylation/…

DNA
Targeted sequencing (amplicons, gene panels, exomes)
Whole genome sequencing

- Finding differences to known reference genome = re-sequencing

***De novo* assembly**
- Genome construction

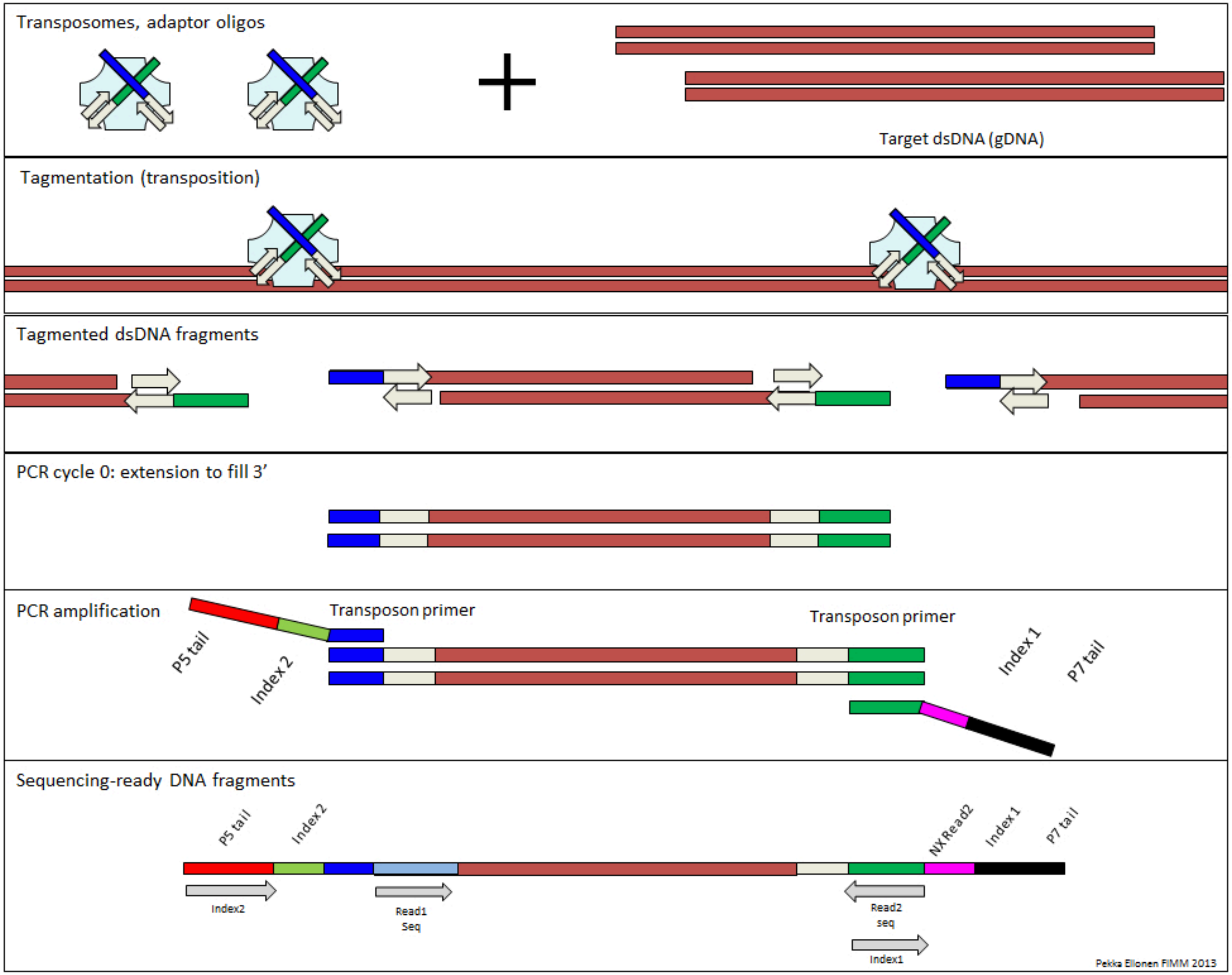
Before we start analysis…

RNA
- Gene expression, alternative splicing

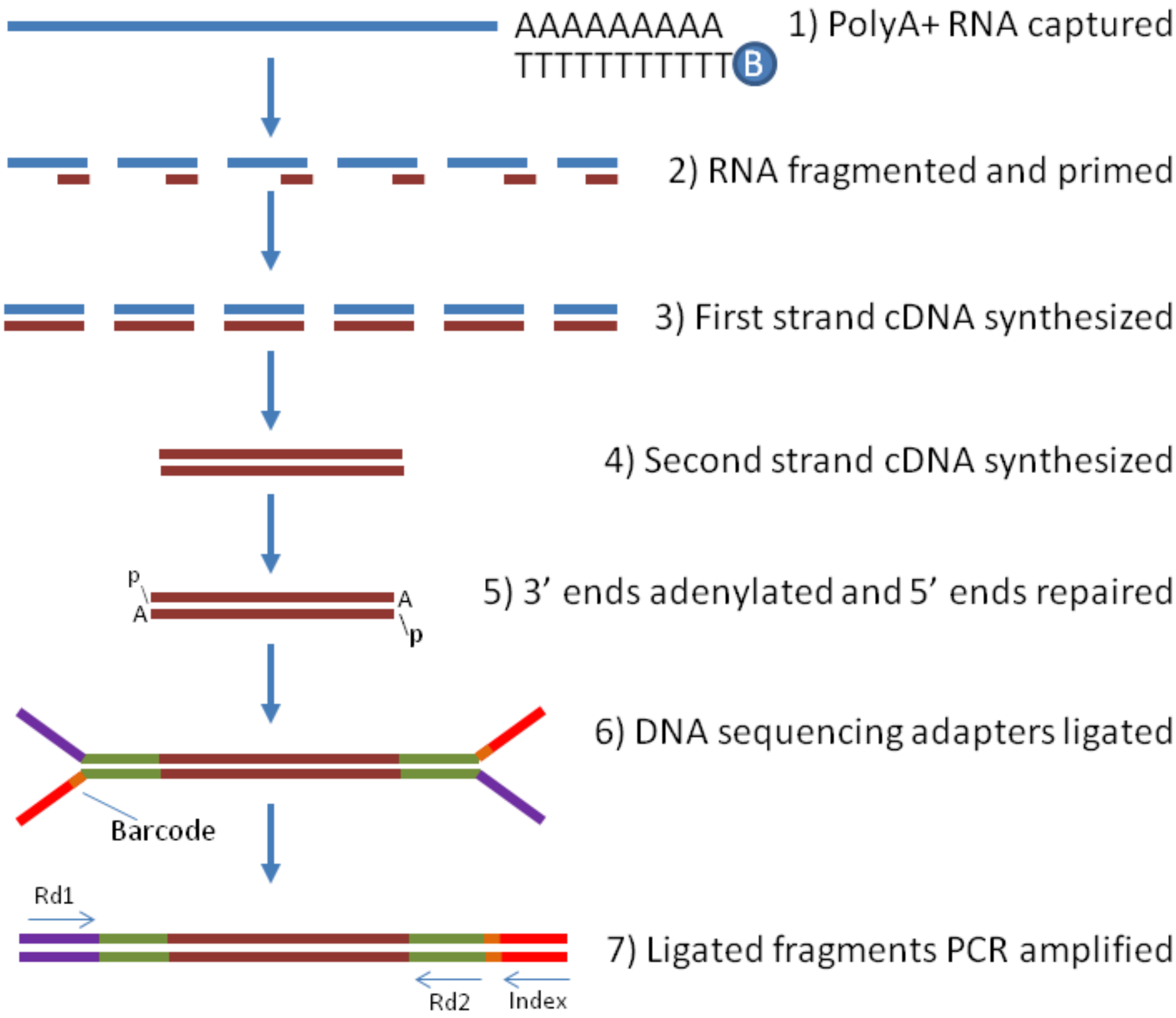
Metagenomics (bacteria, viruses)
- Their composition, variants

ChIP sequrcing (DNA-protein interactions)

Library preparation – example of DNA library



Library preparation – example of mRNA library



1) PolyA+ RNA captured

2) RNA fragmented and primed

3) First strand cDNA synthesized

4) Second strand cDNA synthesized

5) 3’ ends adenylated and 5’ ends repaired

6) DNA sequencing adapters ligated

7) Ligated fragments PCR amplified

Bioinformatics

Bioinformatics is a quite new field… (first NGS in 2005)

How to analyse data defived from NGS = bottleneck of NGS

A lot of tools/software for NGS data analysis…

Most of the tools are command-line based

No tool is working perfectly…☹

Each tools solves only a peace of the cake…
NO tool, that is able to perform analysis from the very beginning
to the end => Need for setup the **pipeline**

Bioinformatics

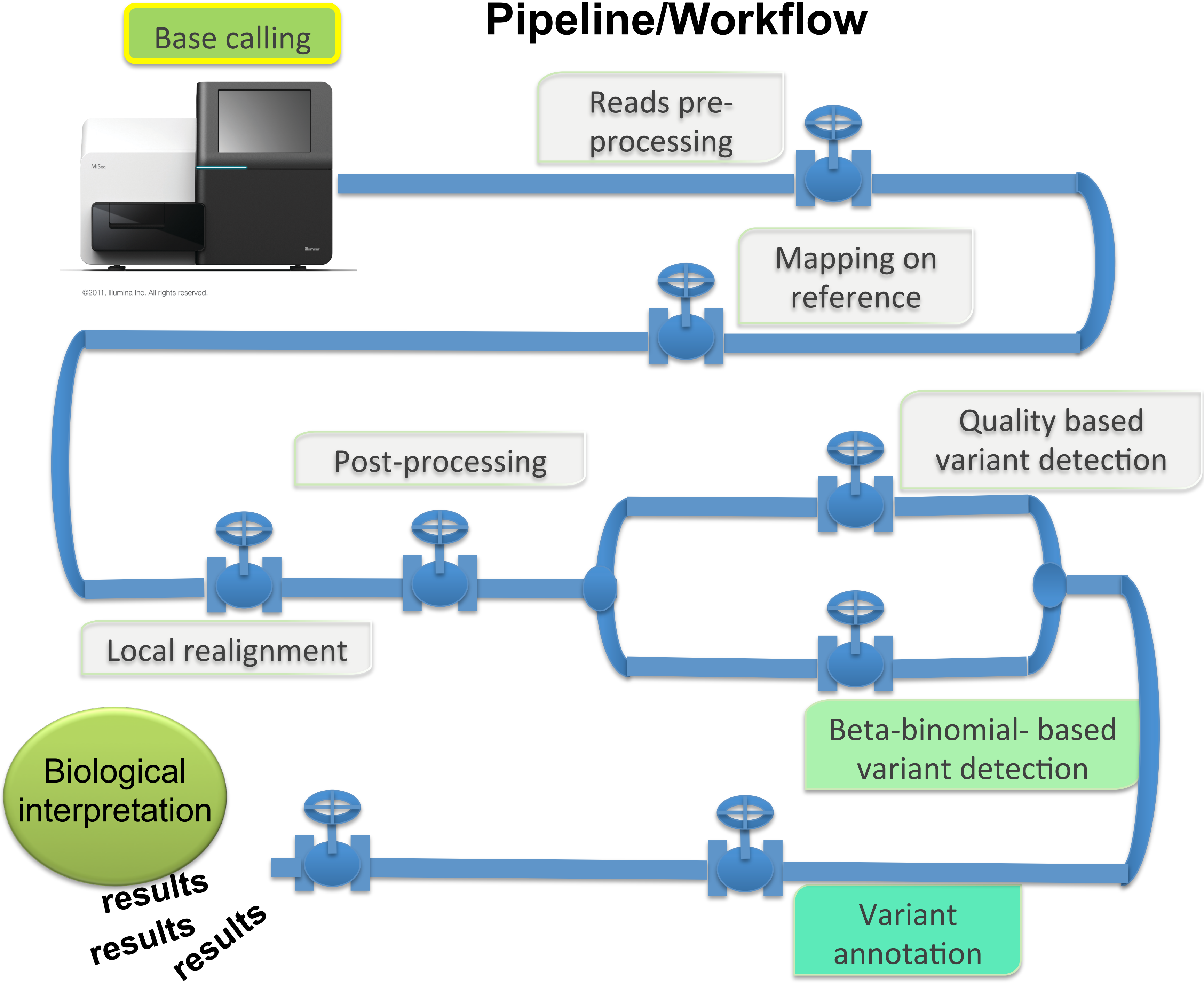
Exception: commercial software and ready to use pipelines **BUT** they have usually not-transparent settings and/or not enough of options

Heavily depends on type of experiment, library preparation and project

Laptop or PC are usually not enough… need for cluster

×

Pipeline/Workflow



©2011, Illumina Inc. All rights reserved.

Base calling

Reads pre-processing

Mapping on reference

Local realignment

Post-processing

Quality based variant detection

Beta-binomial- based variant detection

Variant annotation

Biological interpretation

results

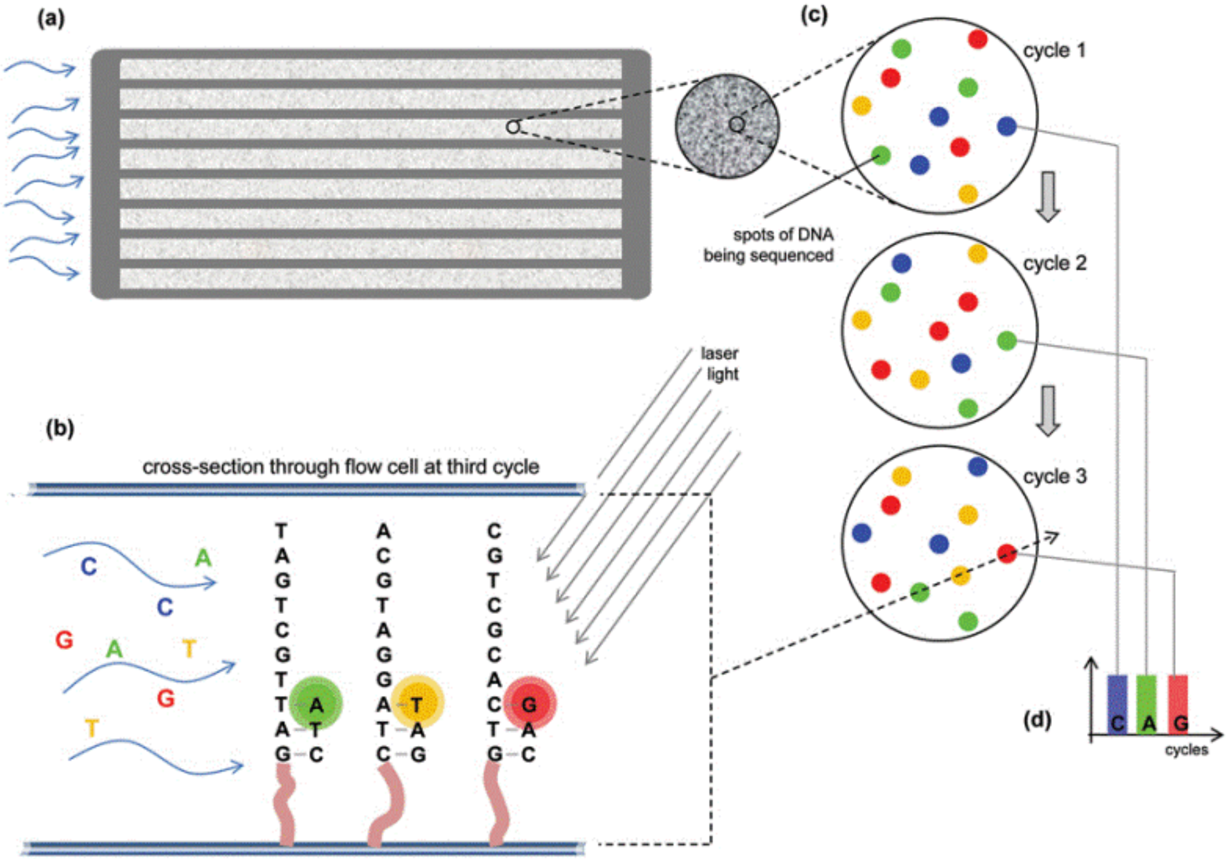
results

results

Base calling

Signal to sequence conversion and assigning base quality scores (fastq file)

Phred score – probability of arising an error (log based)

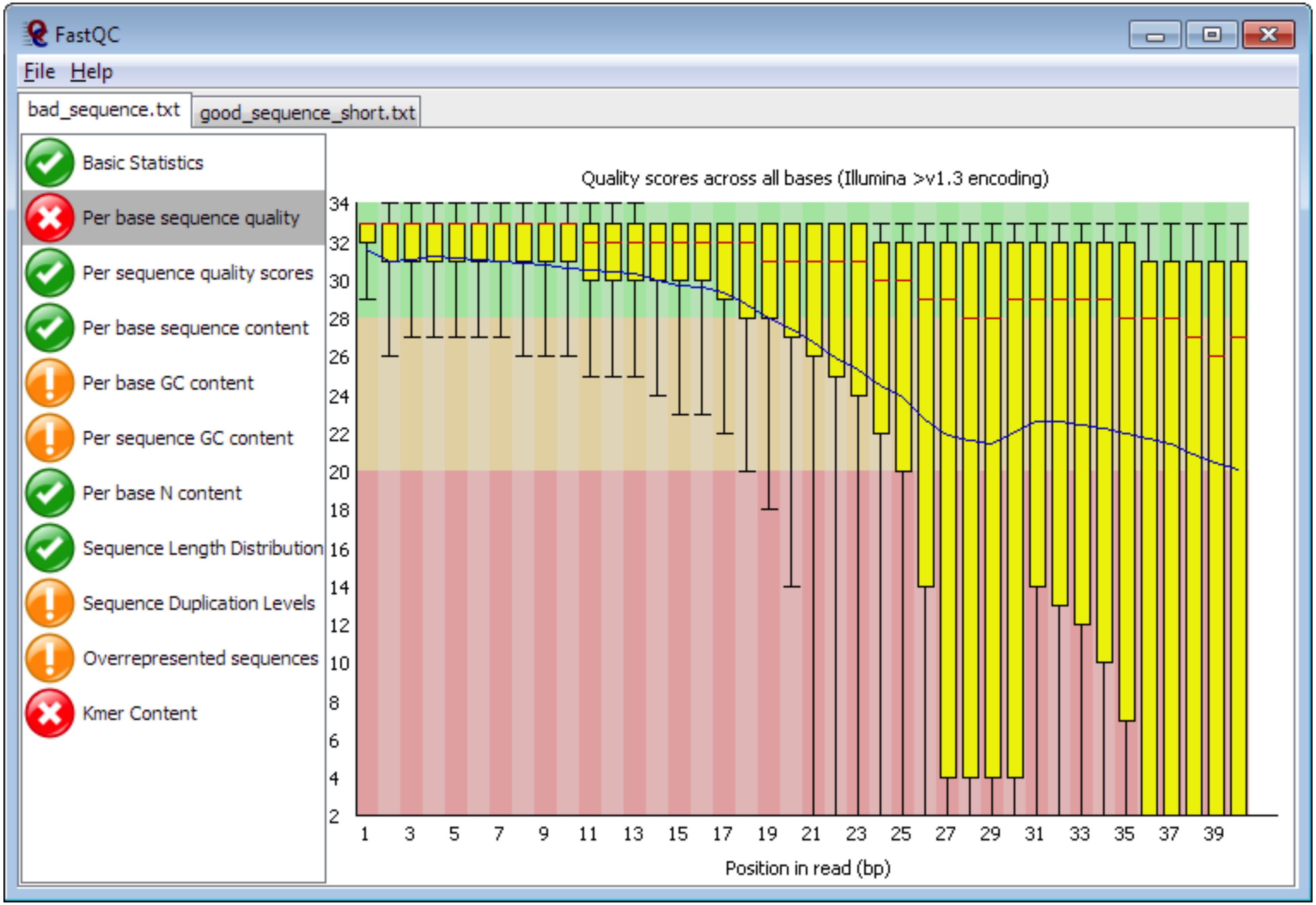


fastq

- Consists of reads - biological sequences
(each read represents 1 input molecule sequenced on flowcell)
- Corresponding quality score for each base
- ASCII character
- (fasta+ qual, csfasta + csqual, sff)
- Pair-end sequencing – 2 fastq files

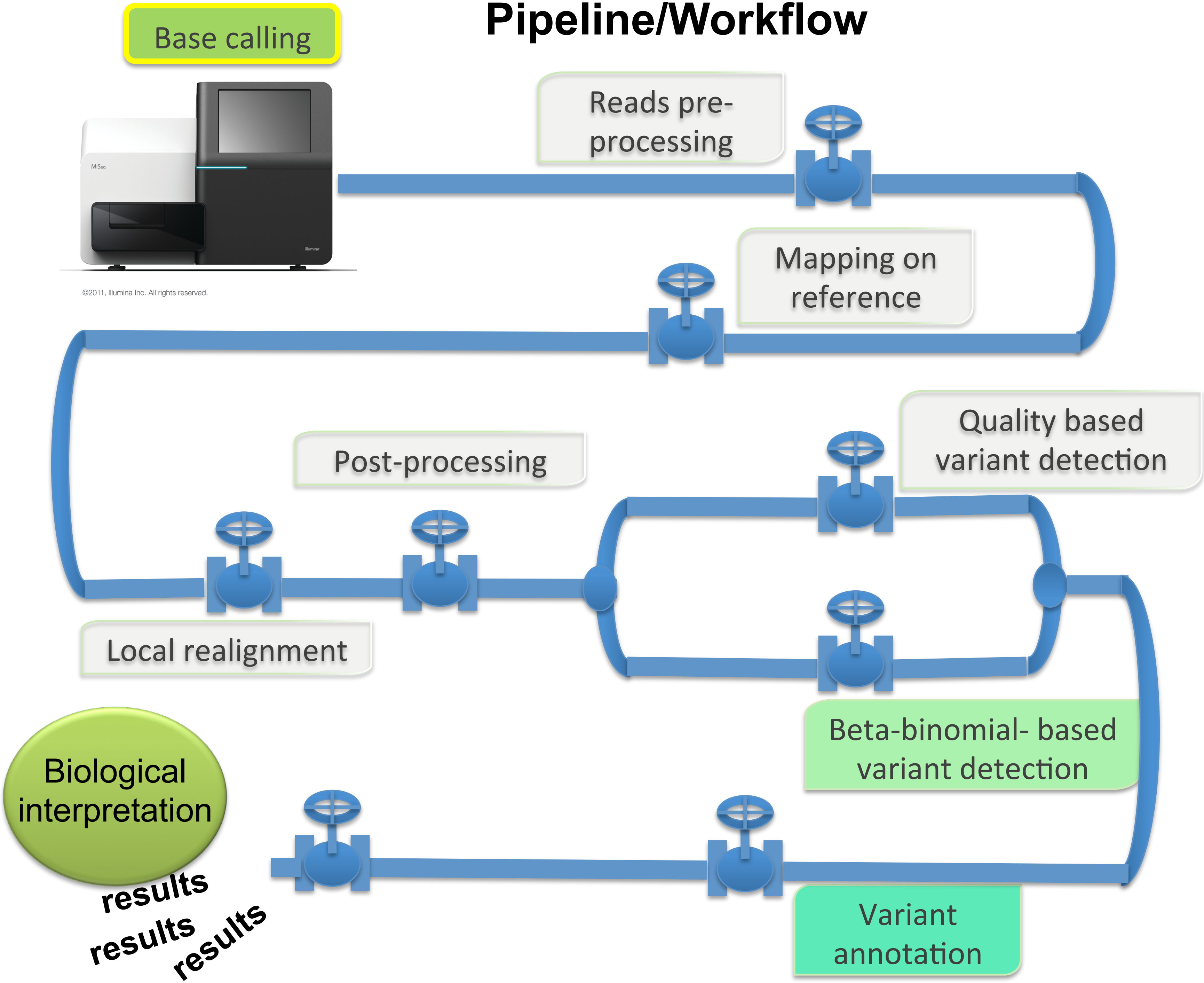
```
@
SEQ_ID GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
 !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Quality control (FastQC)



FastQC
File
Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
Position in read (bp)

Pipeline/Workflow



Base calling

©2011, Illumina Inc. All rights reserved.

Reads pre-processing

Mapping on reference

Post-processing

Local realignment

Quality based variant detection

Beta-binomial- based variant detection

Variant annotation

Biological interpretation

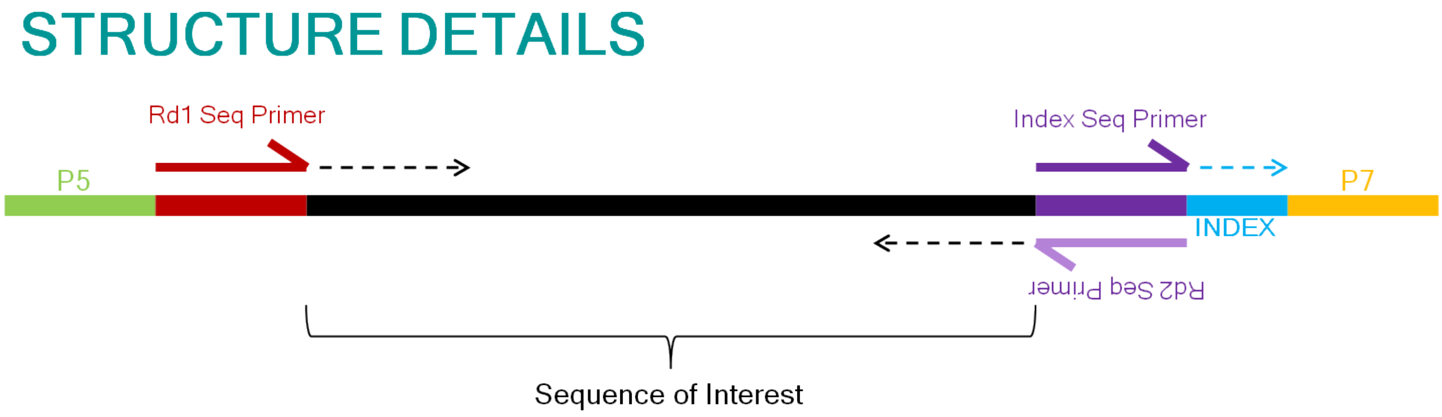
results

results

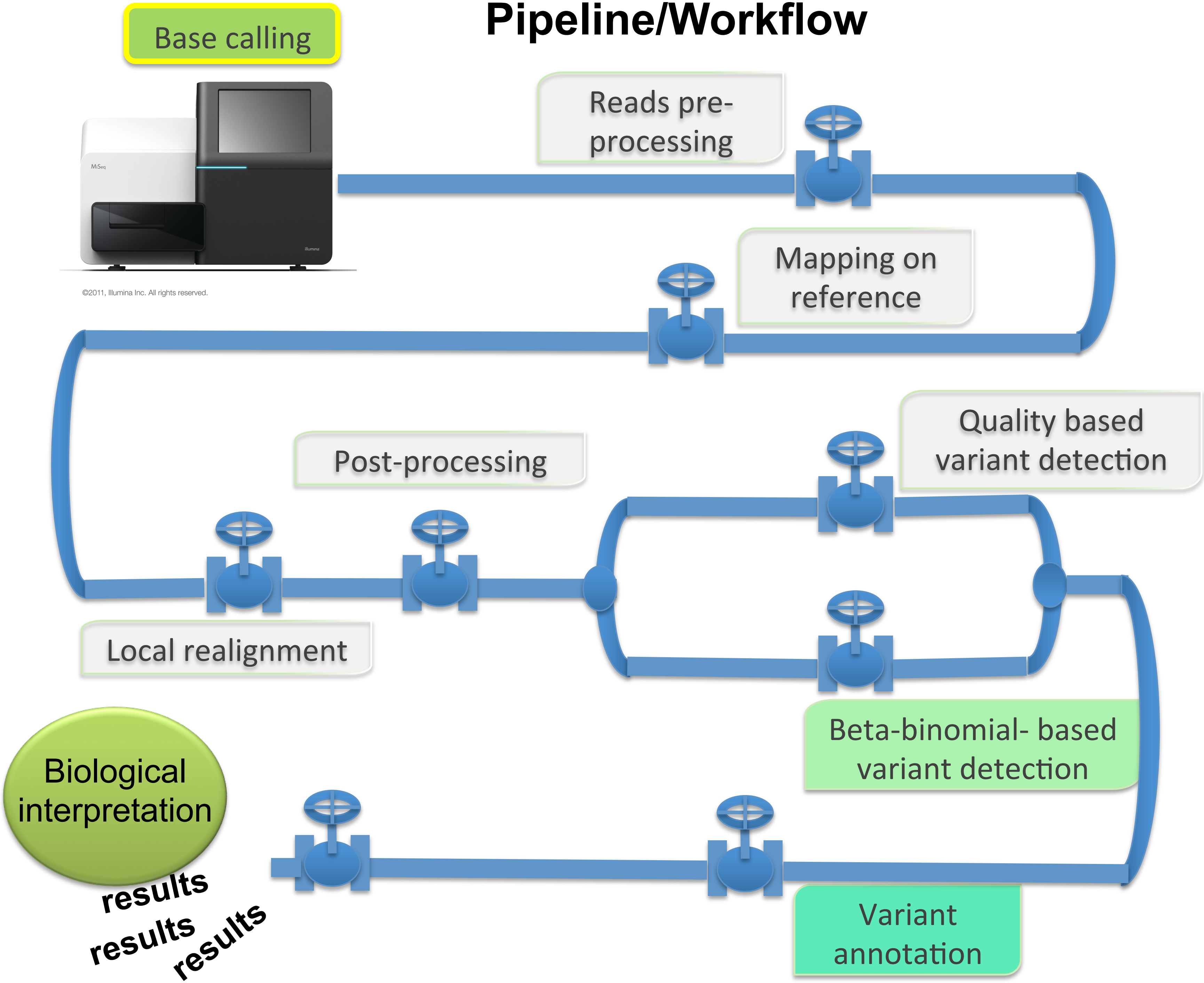
results

Cleaning reads (Cutadapt)

- Adaptor trimming (miRNA)
- Quality trimming
- Length filtering



Pipeline/Workflow



Base calling

©2011, Illumina Inc. All rights reserved.

Reads pre-processing

Mapping on reference

Post-processing

Quality based variant detection

Local realignment

Beta-binomial- based variant detection

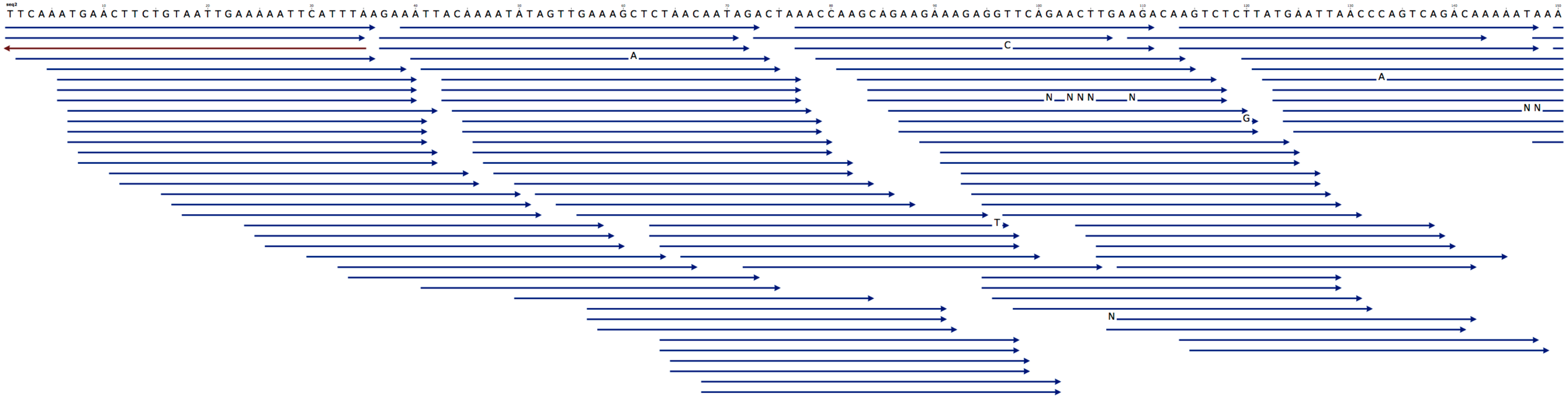
Variant annotation

Biological interpretation

Read mapping (alignment)

- Usually mapping reads on reference sequence (DNA/cDNA/16S/ other seq) to find corresponding location & differences

- Problem with too many sequences and billions bp long references – need for special algorithms (Burrows-Wheeler transform, hash table indexing)


seq2
TTCAAATGAACTTCTGTAATTGAAAAATTCATTTAAGAAATTACAAAATATAGTTGAAAGCTCTAACAATAGACTAAACCAAGCAGAAGAAAGAGGTTCAGAACTTGAAGACAAGTCTCTTATGAATTAACCCAGTCAGACAAAAATAAA
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
A
C
A
N
N N N
N
N N
G
T
N

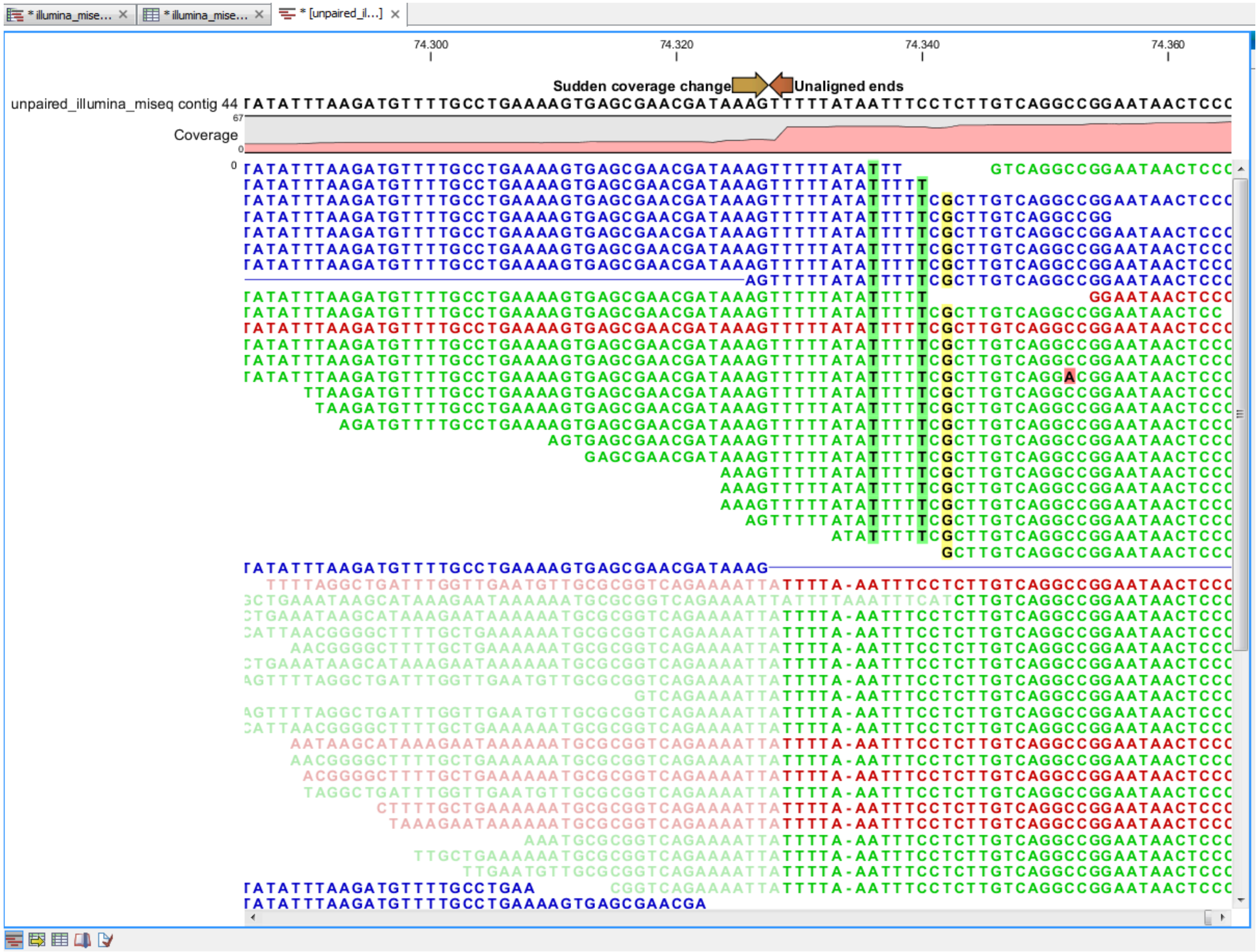
Mapping of DNA reads

•On Existing DNA reference sequence
(ready for many organisms)

•To find substitutions, insertions, deletions, inversions, etc…
Precisely!

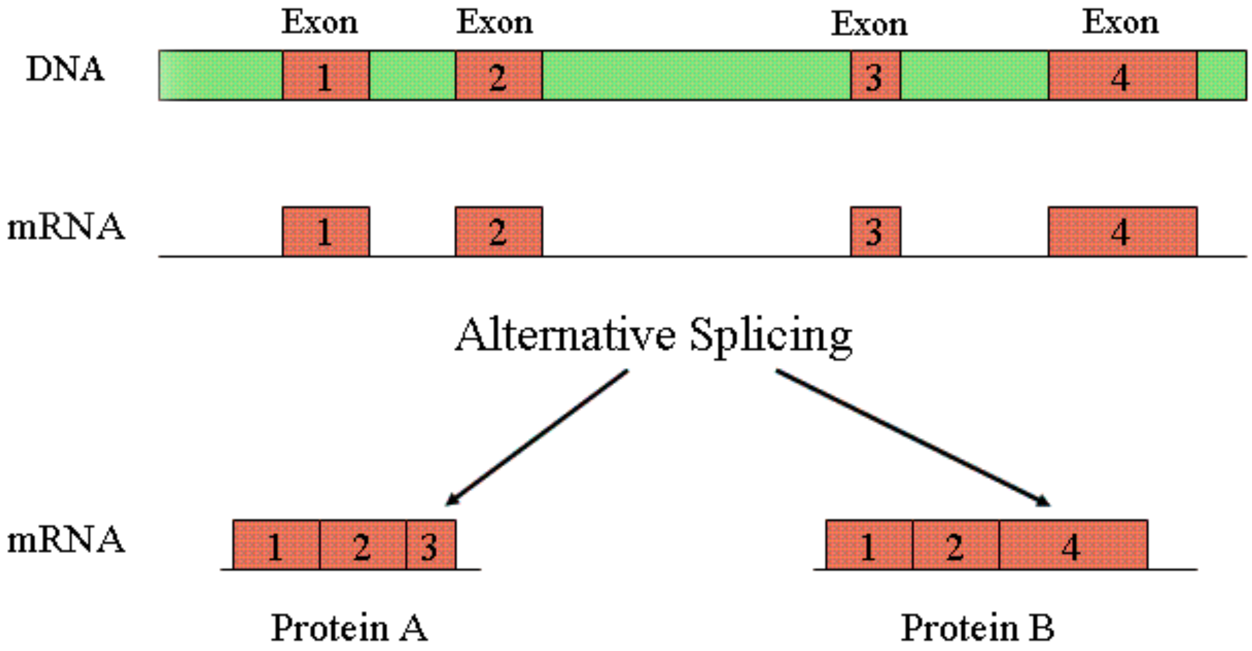
–BWA, Bowtie, Bfast, SHRiMP

Example of DNA re-sequencing



Mapping of RNA reads – alternative splicing

Reads can span exon junctions
- mRNA splicing



Mapping of RNA reads

•To measure gene expression OR alternative splicing

•On existing **DNA** reference sequence

•To find **alternative splicing**

•More tricky, complicated, slower

- TopHat (*de novo* splice aligner)

On **transcriptome** reference sequences

Reads can map to multiple transcripts (shared exons)

Easier, faster, no need for special aligners

- BWA

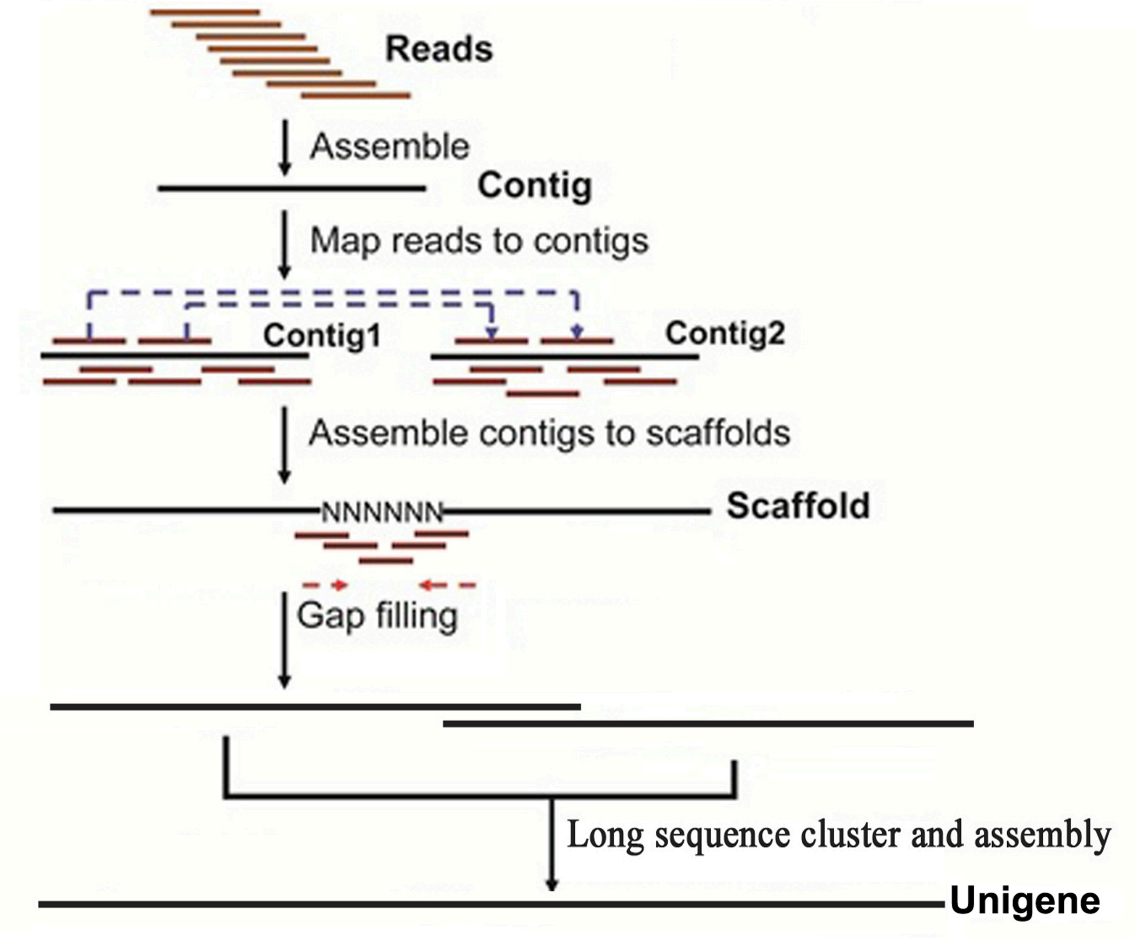
•On **miRNA** sequences - miRBase

–Grouping and annotate against mirBase

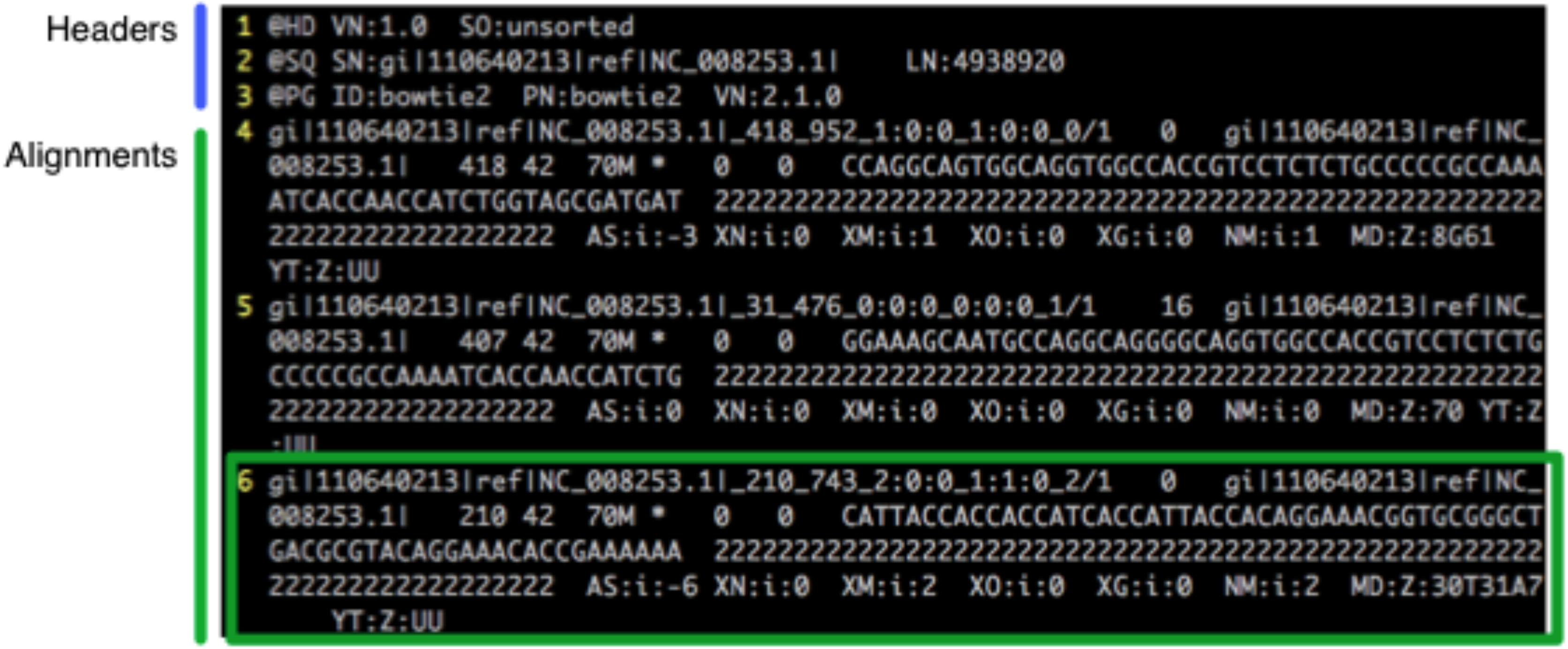
De novo assembly

– to uncover unknown genomes/transcriptomes

- To detect large structural variants



SAM/BAM

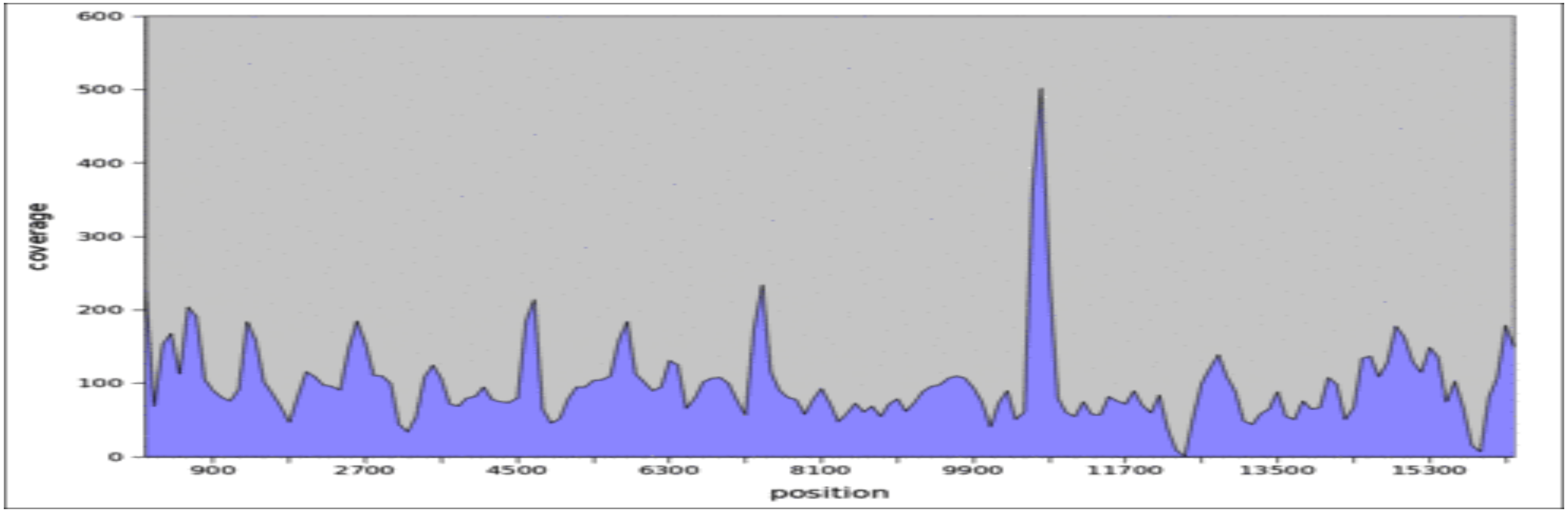


```
@HD VN:1.0  SO:unsorted
@SQ SN:gi|110640213|ref|NC_008253.1|    LN:4938920
@PG ID:bowtie2  PN:bowtie2  VN:2.1.0
gi|110640213|ref|NC_008253.1|_418_952_1:0:0_1:0:0_0/1   0   gi|110640213|ref|NC_
008253.1|   418 42  70M *   0   0   CCAGGCAGTGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAA
ATCACCAACCATCTGGTAGCGATGAT  2222222222222222222222222222222222222222222222222222
222222222222222222  AS:i:-3 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:8G61
YT:Z:UU
gi|110640213|ref|NC_008253.1|_31_476_0:0:0_0:0:0_1/1    16  gi|110640213|ref|NC_
008253.1|   407 42  70M *   0   0   GGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTG
CCCCCGCCAAAATCACCAACCATCTG  2222222222222222222222222222222222222222222222222222
222222222222222222  AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:70 YT:Z
:UU
gi|110640213|ref|NC_008253.1|_210_743_2:0:0_1:1:0_2/1   0   gi|110640213|ref|NC_
008253.1|   210 42  70M *   0   0   CATTACCACCACCATCACCATTACCACAGGAAACGGTGCGGGCT
GACGCGTACAGGAAACACCGAAAAAA  2222222222222222222222222222222222222222222222222222
222222222222222222  AS:i:-6 XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:30T31A7
    YT:Z:UU
```

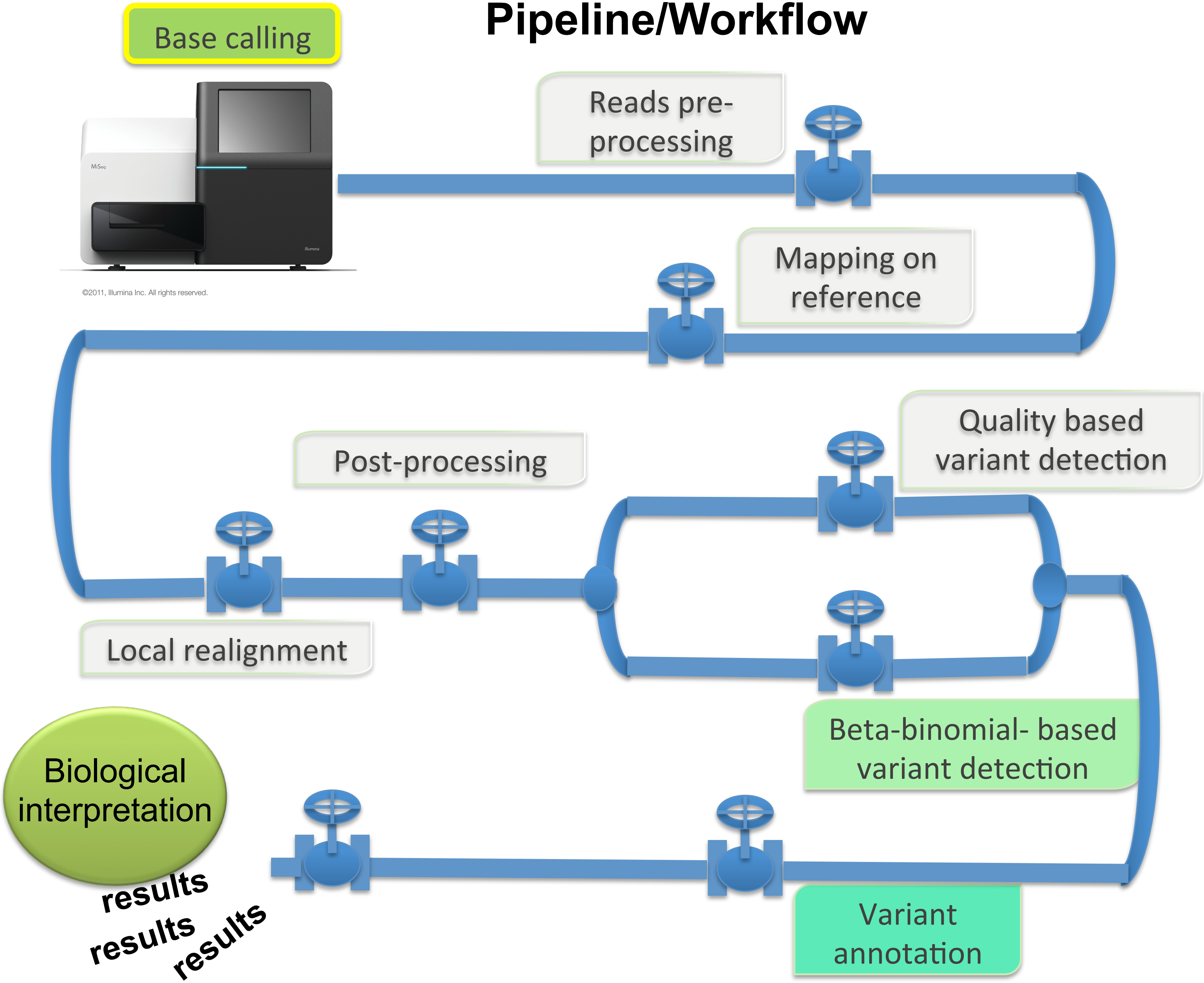
Each row describes a single alignment of a raw read against the reference genome. Each alignment has 11 mandatory fields, followed by any number of optional fields.

Mapping, Coverage reports

- Repeat alignment/other steps with different criteria?
- Important checkout for lab protocol
- Specificity of PCR
- Settings of variant calling threshold, CNV



Pipeline/Workflow



Base calling

©2011, Illumina Inc. All rights reserved.

Reads pre-processing

Mapping on reference

Local realignment

Post-processing

Quality based variant detection

Beta-binomial- based variant detection

Variant annotation

Biological interpretation

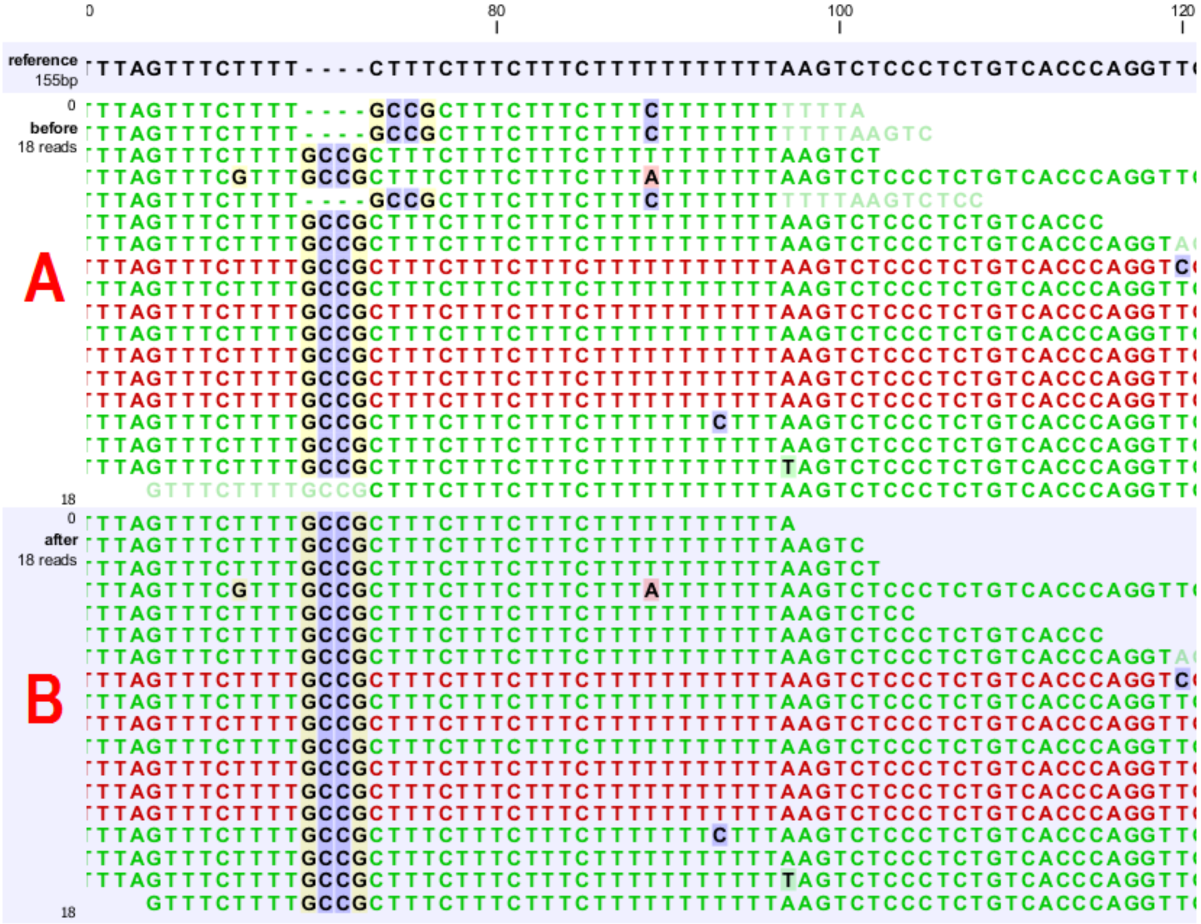
results

results

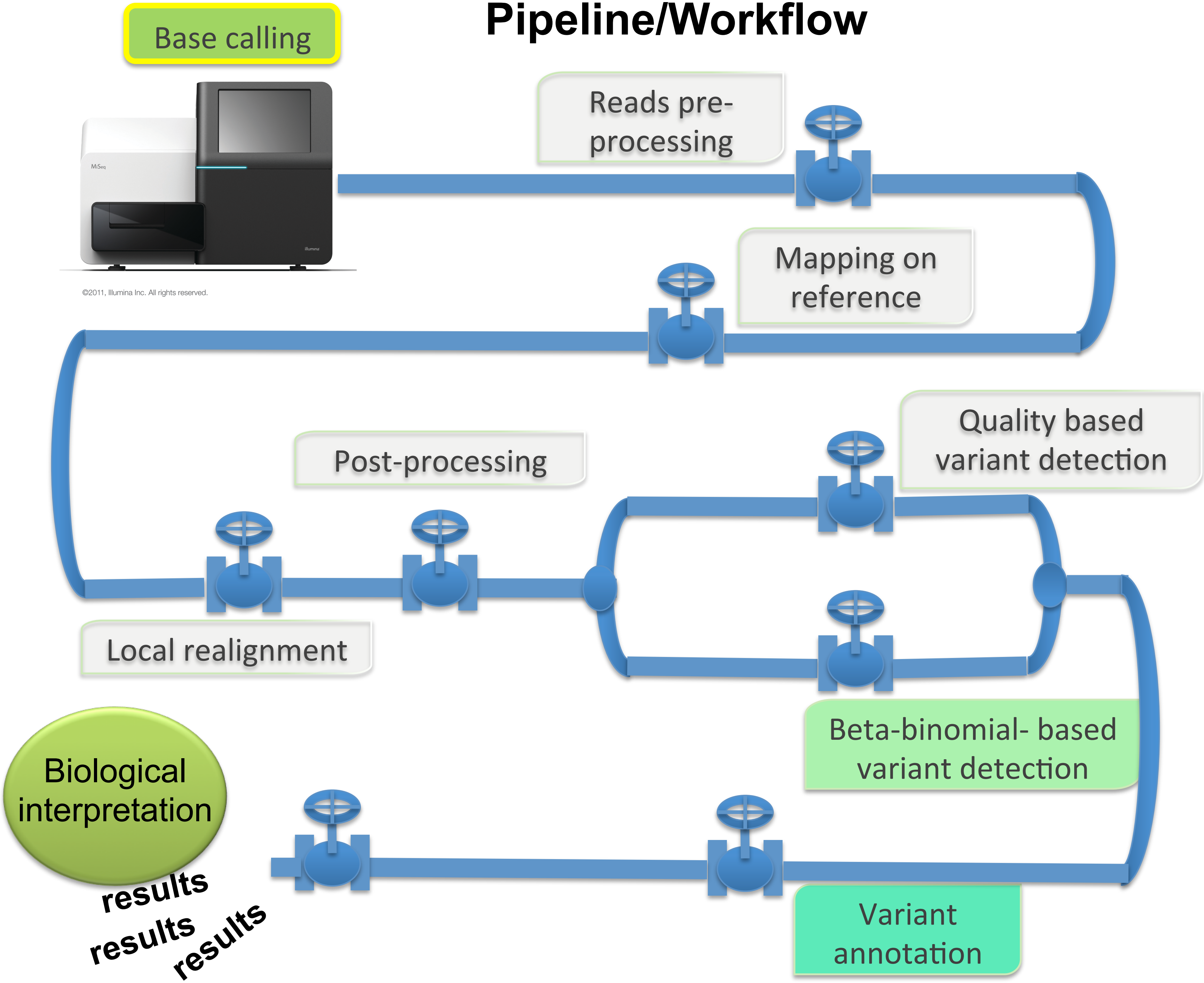
results

Indel realignment

Usually alignment is not perfect – false positive indels & Substitutions => Need for local indel realignment



Pipeline/Workflow



Base calling

©2011, Illumina Inc. All rights reserved.

Reads pre-processing

Mapping on reference

Post-processing

Quality based variant detection

Local realignment

Beta-binomial- based variant detection

Variant annotation

Biological interpretation

Remove PCR duplicates

Each read represents 1 input molecule

THEORY:
E.g. in case of DNA re-sequencing, 1 diploid cell is represented by 2 reads because of 2 chromosomes

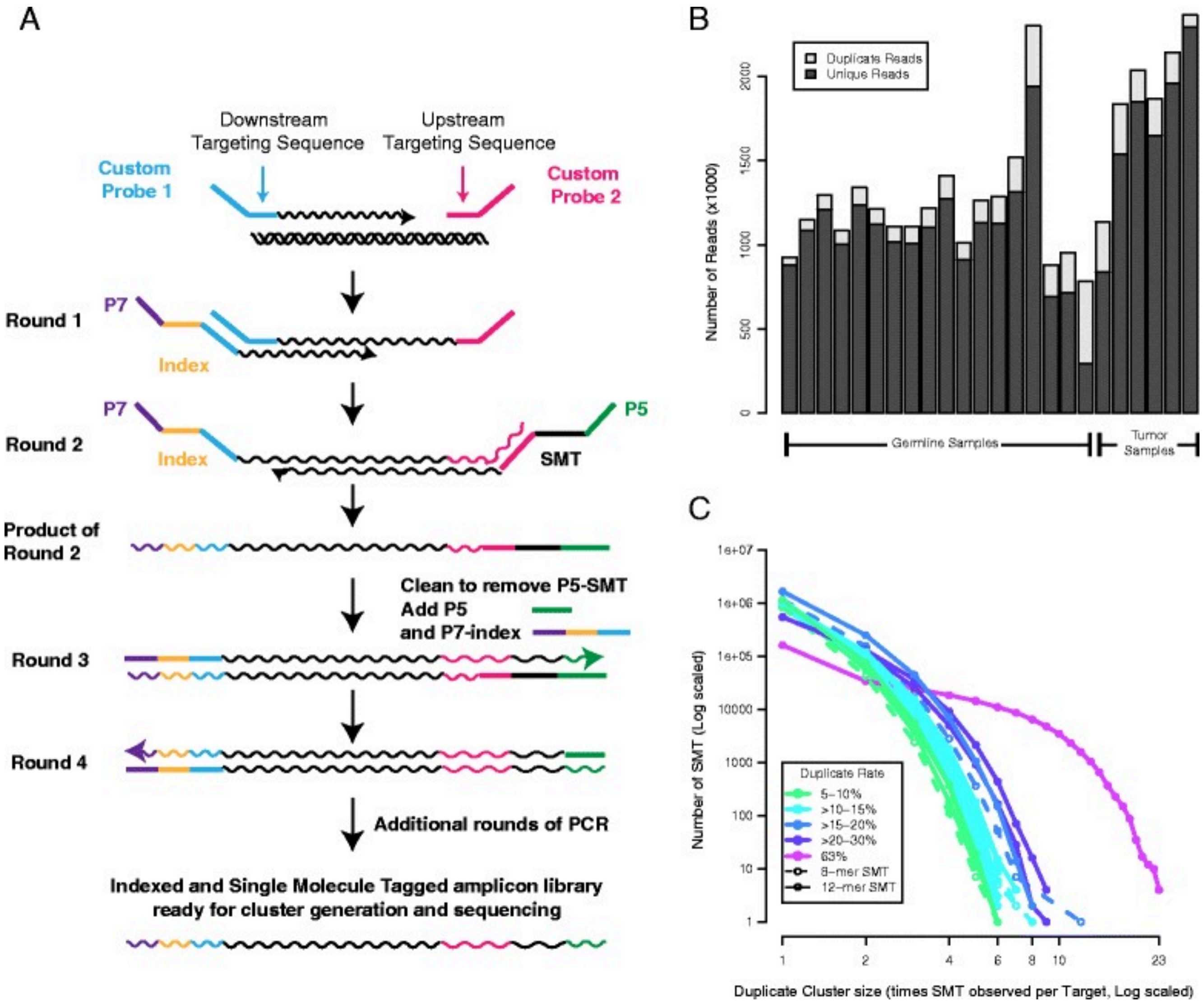
BUT

there is a PCR to amplify genetic material to be analyzable =>
1 input molecule from 1 cell could be after PCR represented by more reads => Biased variant allele frequency

How to solve it?

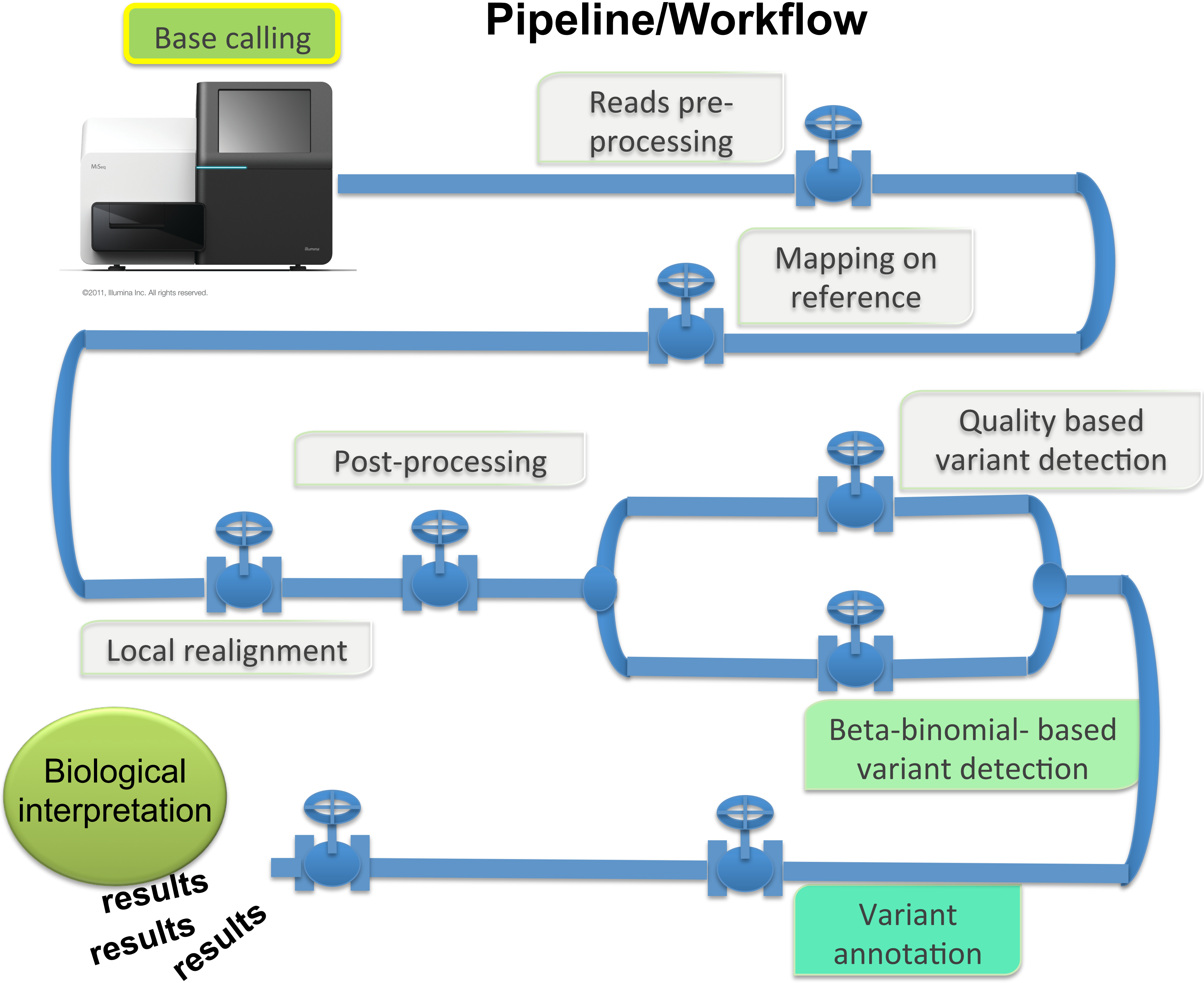
1) Molecular barcodes (very new method)
2) Identity of start-end positions of read pair

Molecular barcodes



Smith et al. 2014

Pipeline/Workflow



DNA Seq - variant calling

- To detect differences from reference sequence

- Single/multi-nucleotide
- Substitutions
- Insertions
- Deletions

- Inversions
- Large structural variations (translocations, indels)
- Copy number variations

DNA Seq variant calling

based on many criteria like:

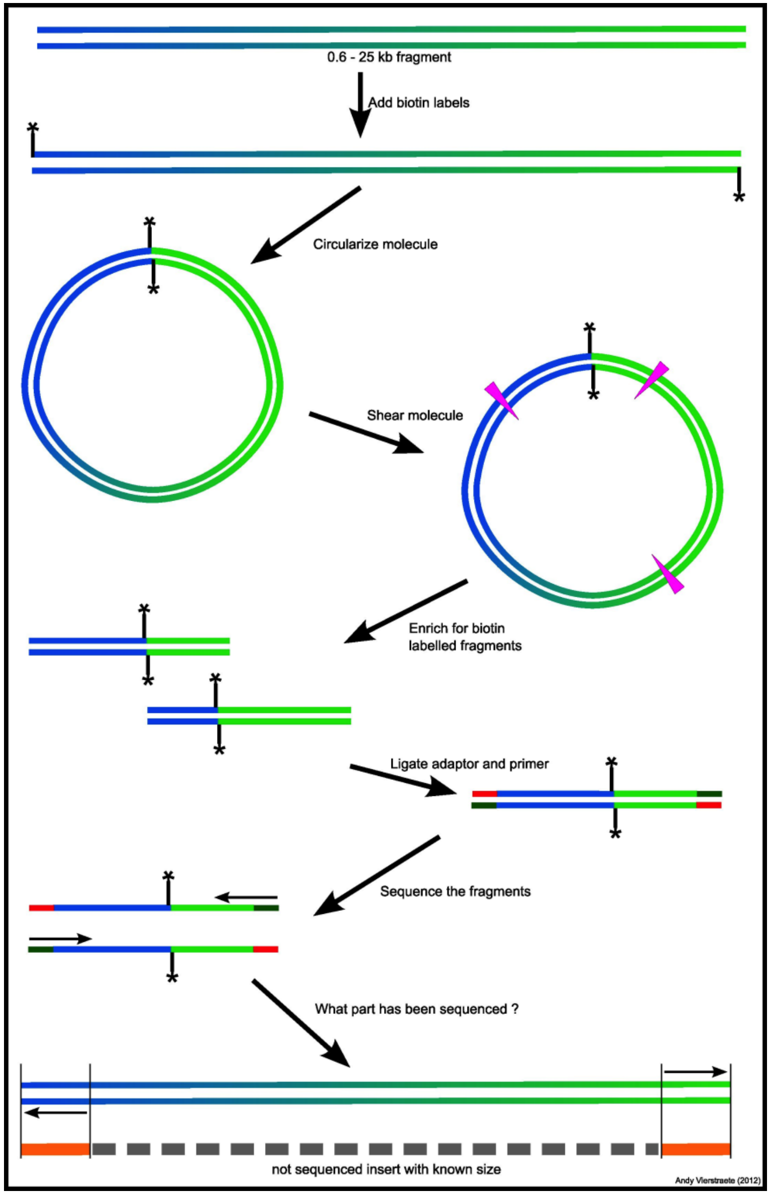
- Coverage
- Variant alelle frequency
- Base quality

- Depends also on:
- Genomic context (homopolymers)
- Nucleotide type
- Position in read (errors at the read end)
- Alignment errors (importance of realignment)
- Presence in both forward and reverse reads

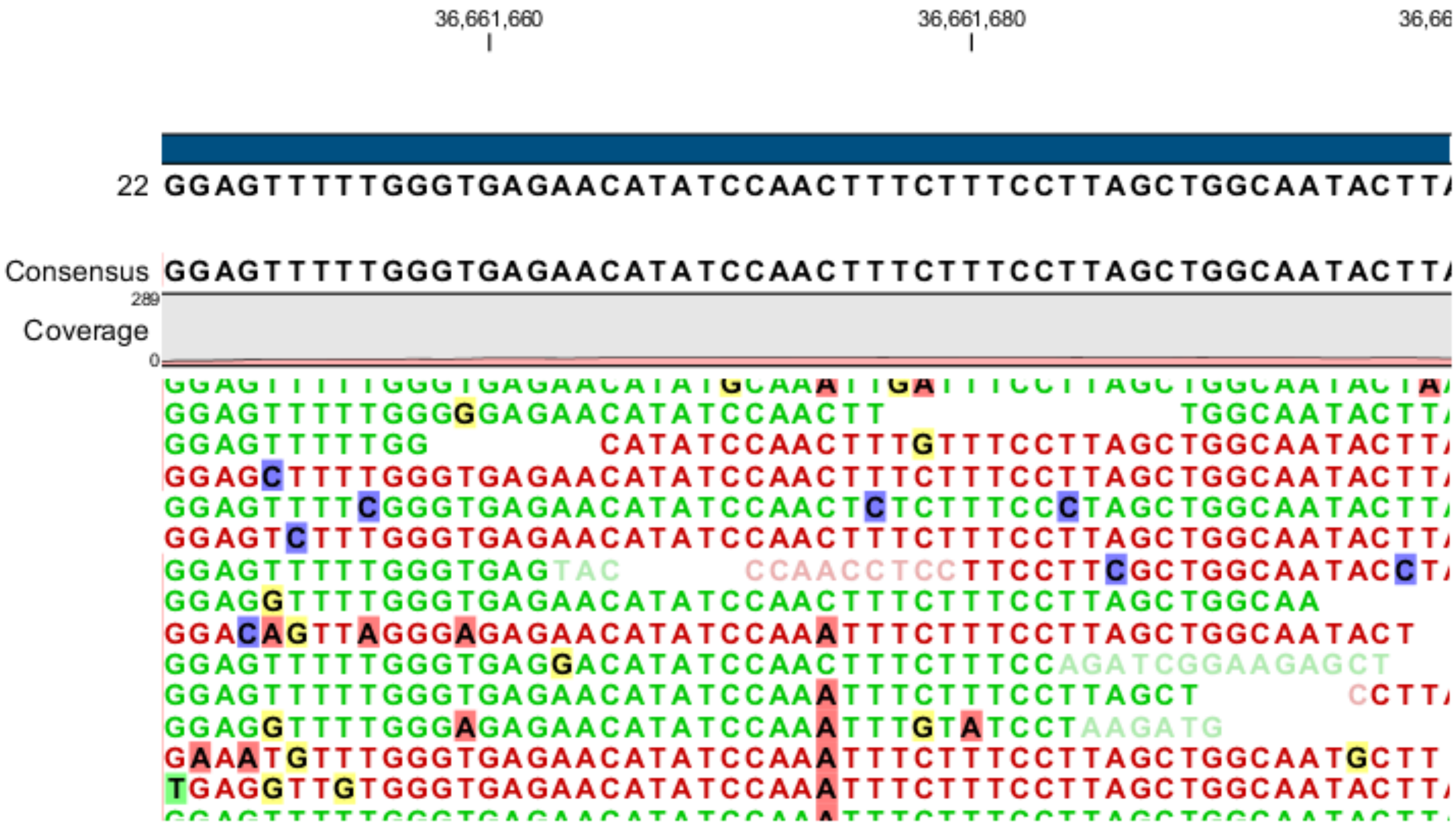
Necessary to take into account type of library preparation (single end; pair end; mate pair)

DNA Seq variant calling

- Mate-pair library
- Detection of large indels
& translocations


0.6 - 25 kb fragment
Add biotin labels
Circularize molecule
Shear molecule
Enrich for biotin labelled fragments
Ligate adaptor and primer
Sequence the fragments
What part has been sequenced ?
not sequenced insert with known size
Andy Vierstraete (2012)

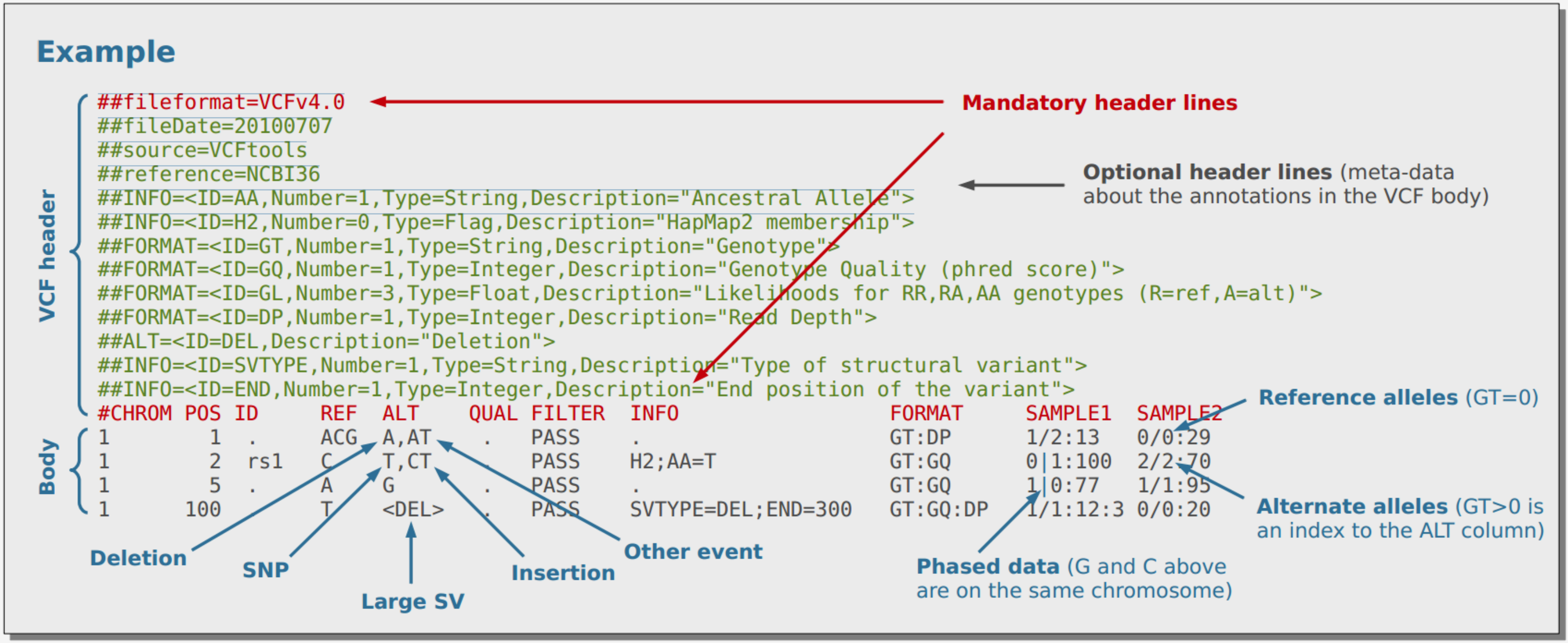
DNA Seq variant calling



```
                 36,661,660                 36,661,680                 36,66

22        GGAGTTTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAATACTT
Consensus GGAGTTTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAATACTT
289
Coverage
0
          GGAGTTTTGGGTGAGAACATATGCAAATTGATTTCCTTAGCTGGCAATACTA
          GGAGTTTTGGGGGAGAACATATCCAACTT            TGGCAATACTT
          GGAGTTTTGG       CATATCCAACTTTGTTTCCTTAGCTGGCAATACTT
          GGAGCTTTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAATACTT
          GGAGTTTCGGGTGAGAACATATCCAACTCTCTTTCCCTAGCTGGCAATACTT
          GGAGTCTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAATACTT
          GGAGTTTTGGGTGAGTAC      CCAACCTCCTTCCTTCGCTGGCAATACCT
          GGAGGTTTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAA
          GGACAGTTAGGGAGAGAACATATCCAAATTTCTTTCCTTAGCTGGCAATACT
          GGAGTTTTGGGTGAGGACATATCCAACTTTCTTTCCAGATCGGAAGAGCT
          GGAGTTTTGGGTGAGAACATATCCAAATTTCTTTCCTTAGCT       CCTT
          GGAGGTTTTGGGAGAGAACATATCCAAATTTGTATCCTAAGATG
          GAAATGTTTGGGTGAGAACATATCCAAATTTCTTTCCTTAGCTGGCAATGCTT
          TGAGGTTGTGGGTGAGAACATATCCAAATTTCTTTCCTTAGCTGGCAATACTT
          GGAGTTTTTGGGTGAGAACATATCCAAATTTCTTTCCTTAGCTGGCAATACTT
```

vcf file



Example

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID    REF  ALT    QUAL FILTER  INFO                 FORMAT     SAMPLE1   SAMPLE2
1        1 .     ACG  A,AT    .   PASS    .                    GT:DP      1/2:13    0/0:29
1        2 rs1   C    T,CT    .   PASS    H2;AA=T              GT:GQ      0|1:100   2/2:70
1        5 .     A    G       .   PASS    .                    GT:GQ      1|0:77    1/1:95
1      100 .     T    <DEL>   .   PASS    SVTYPE=DEL;END=300   GT:GQ:DP   1/1:12:3  0/0:20
```

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Phased data (G and C above are on the same chromosome)

DNA Seq variant calling

- Tumor only (amplicon sequencing & diagnostics)
- Tumor & normal (exome sequencing)

-to do variant calling and genotyping more precisely (somatic, germinal mutations)

- Option is also to analyze tumor vs. group of tumors

Application of many statistical tests:

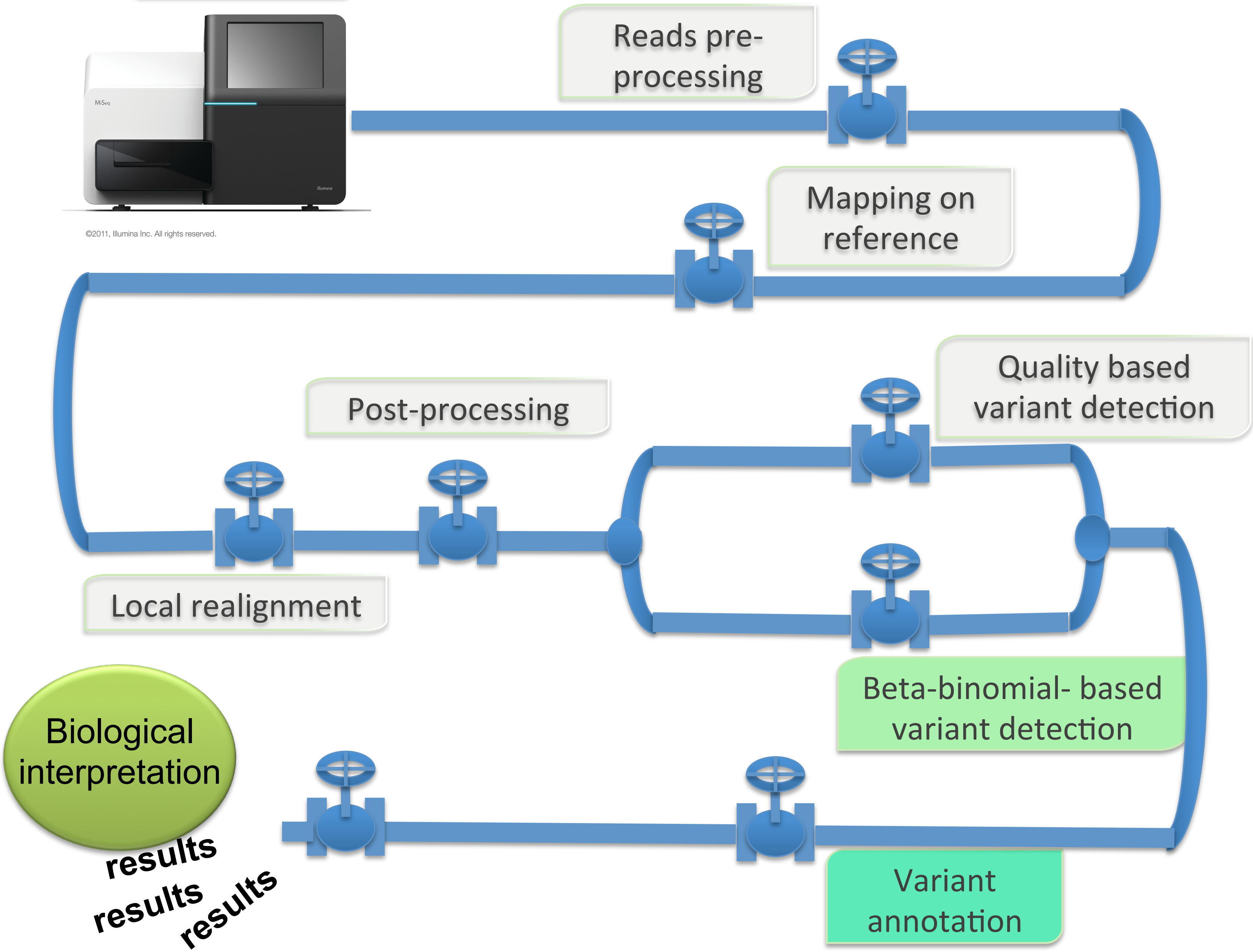
- negative beta-binomial test
- Bayesian statistics
- Fisher exact test

As higher coverage as higher sensitivity and specificity (but limited)

More about statistics and RNA sequencing in the next courses

Pipeline/Workflow

Base calling



Reads pre-processing

Mapping on reference

Local realignment

Post-processing

Quality based variant detection

Beta-binomial- based variant detection

Variant annotation

Biological interpretation

results

results

results

Annotating and filtering of detected variants

- Gene
- Transcript
- dbSNP
- Regulation
- Comparative genomics
- Repeats
- Functional
- Gene ontology
- Etc.