CEITEC

Central European Institute of Technology

BRNO | CZECH REPUBLIC

Moderní metody analýzy genomu - analýza

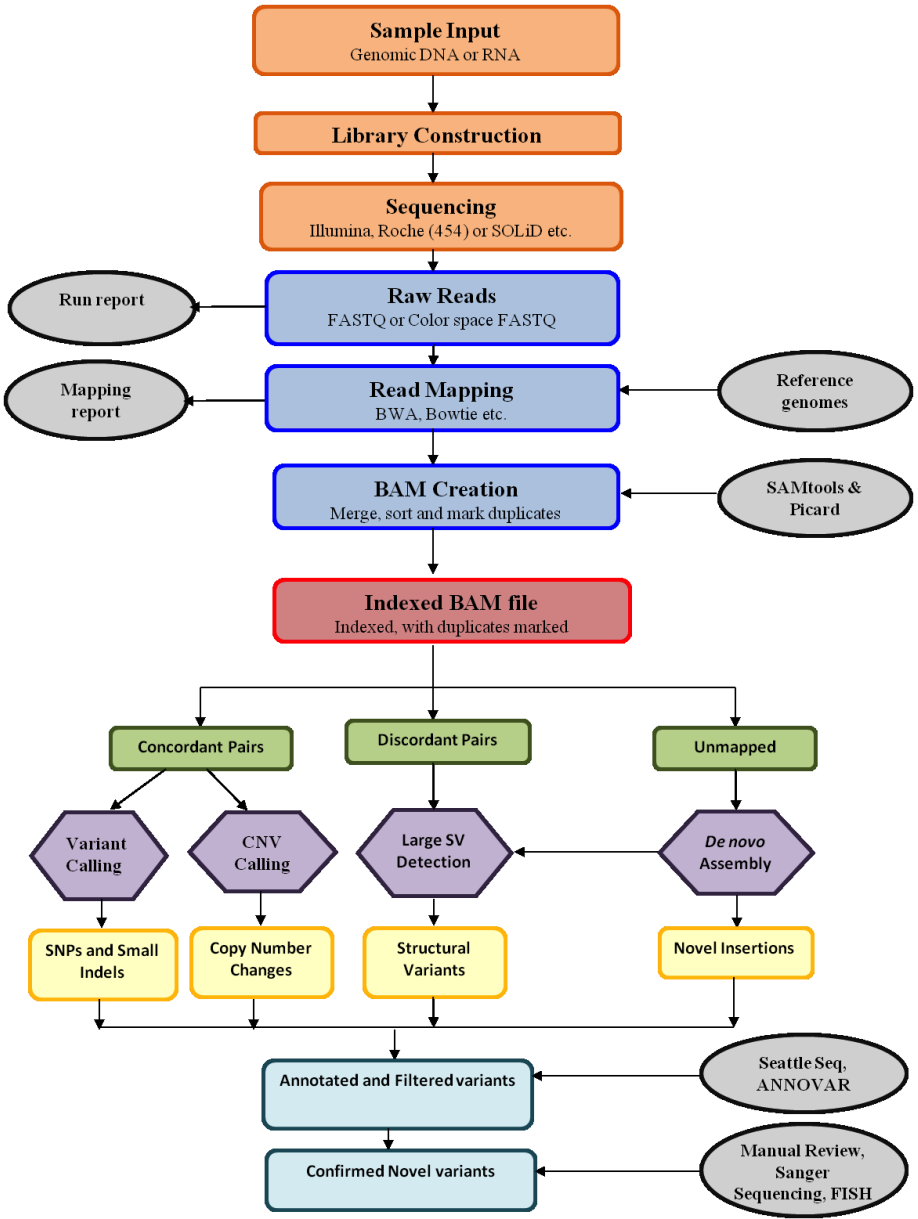
Mgr. Nikola Tom

Brno, 2.4.2014

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND
INVESTING IN YOUR FUTURE

2007–13
OP Research and Development for Innovation

Workflow

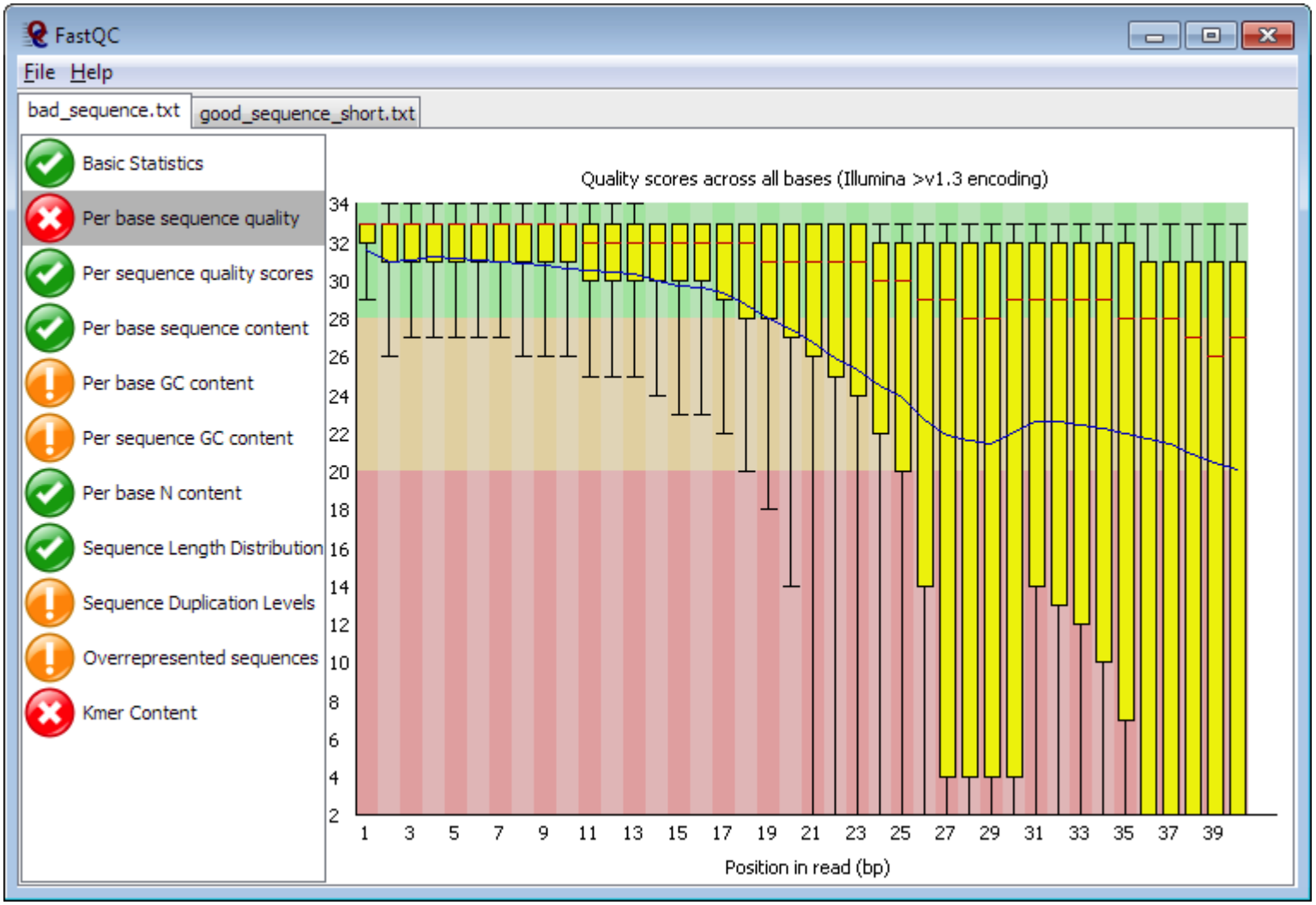

Sample Input
Genomic DNA or RNA
Library Construction
Sequencing
Illumina, Roche (454) or SOLiD etc.
Raw Reads
FASTQ or Color space FASTQ
Run report
Read Mapping
BWA, Bowtie etc.
Mapping report
Reference genomes
BAM Creation
Merge, sort and mark duplicates
SAMtools & Picard
Indexed BAM file
Indexed, with duplicates marked
Concordant Pairs
Discordant Pairs
Unmapped
Variant Calling
CNV Calling
Large SV Detection
De novo Assembly
SNPs and Small Indels
Copy Number Changes
Structural Variants
Novel Insertions
Annotated and Filtered variants
Seattle Seq, ANNOVAR
Confirmed Novel variants
Manual Review, Sanger Sequencing, FISH

Raw sequence = fastq

- Biological sequence
- Corresponding quality scores
- ASCII character
- (fasta+ qual, csfasta + csqual, sff)

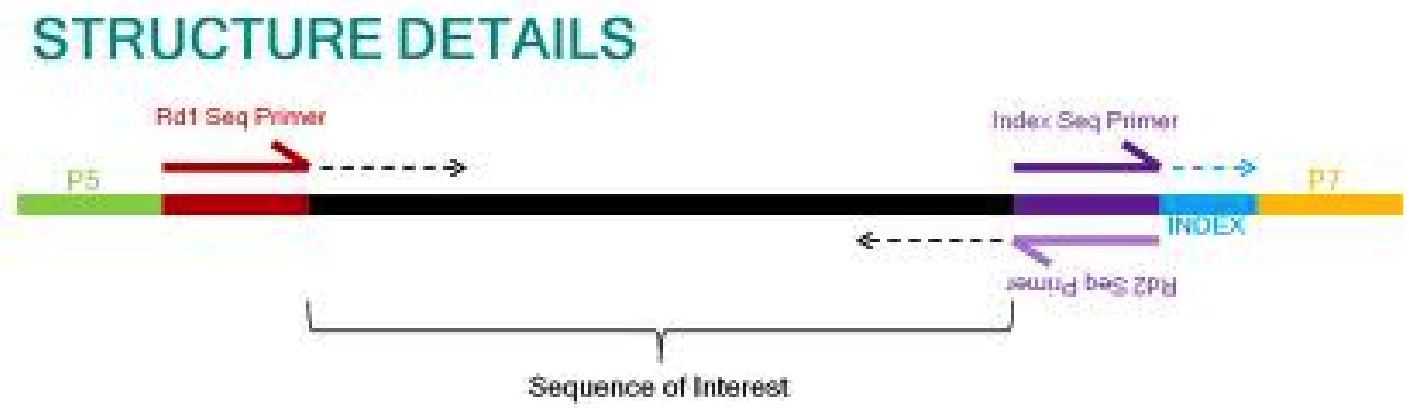
@
SEQ_ID GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

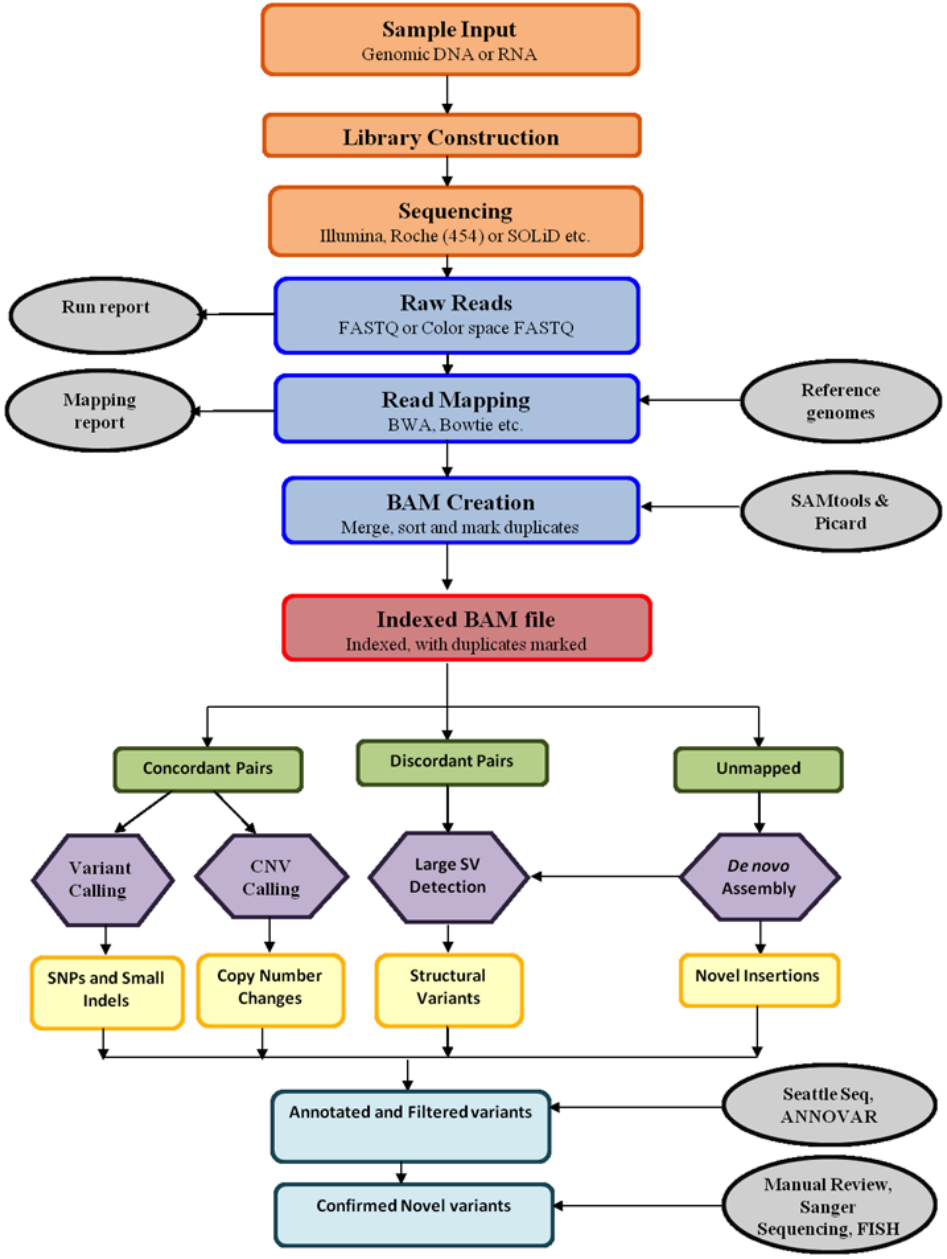
FastQC


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)

Cutadapt

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




Sample Input
Genomic DNA or RNA
Library Construction
Sequencing
Illumina, Roche (454) or SOLiD etc.
Raw Reads
FASTQ or Color space FASTQ
Run report
Read Mapping
BWA, Bowtie etc.
Mapping report
Reference genomes
BAM Creation
Merge, sort and mark duplicates
SAMtools & Picard
Indexed BAM file
Indexed, with duplicates marked
Concordant Pairs
Discordant Pairs
Unmapped
Variant Calling
CNV Calling
Large SV Detection
De novo Assembly
SNPs and Small Indels
Copy Number Changes
Structural Variants
Novel Insertions
Annotated and Filtered variants
Seattle Seq, ANNOVAR
Confirmed Novel variants
Manual Review, Sanger Sequencing, FISH

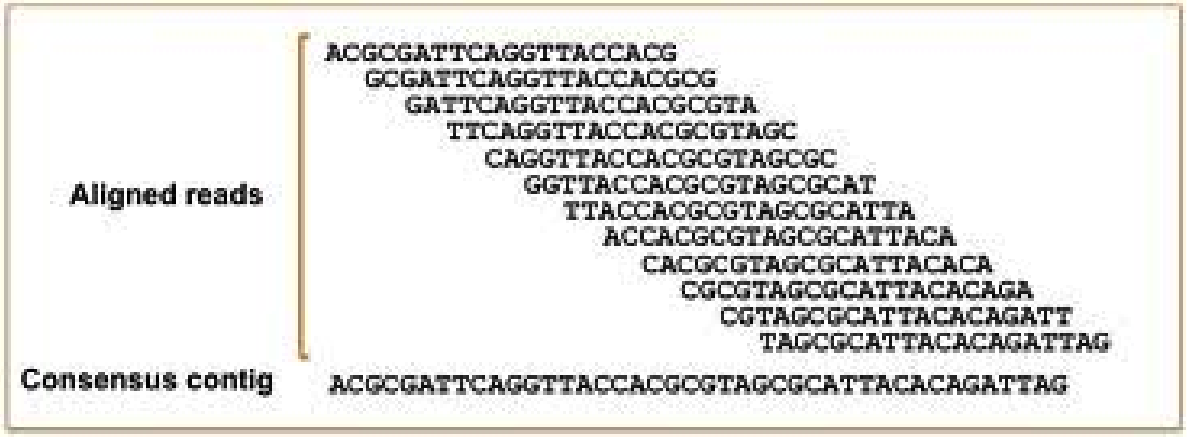
Read mapping => SAM, BAM

- Usually mapping reads on reference
- miRNA - special case
  - Grouping and annotate against mirBase
- **DNA**
  - BWA, Bowtie, Bfast, SHRiMPclc
- **RNA**
  - TopHat (*de novo* splice aligner)
- **Commercial**
  - CLC Genomics Workbench

SAM

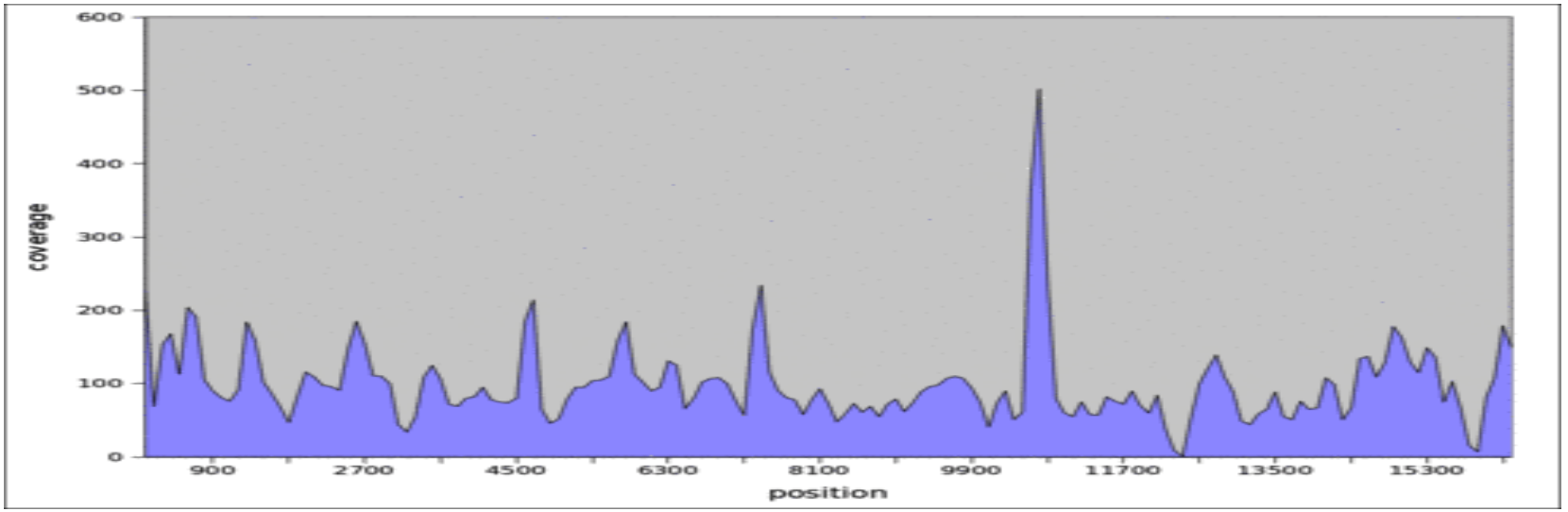
Each row describes a single alignment of a raw read against the reference genome.
Each alignment has 11 mandatory fields, followed by any number of optional fields.

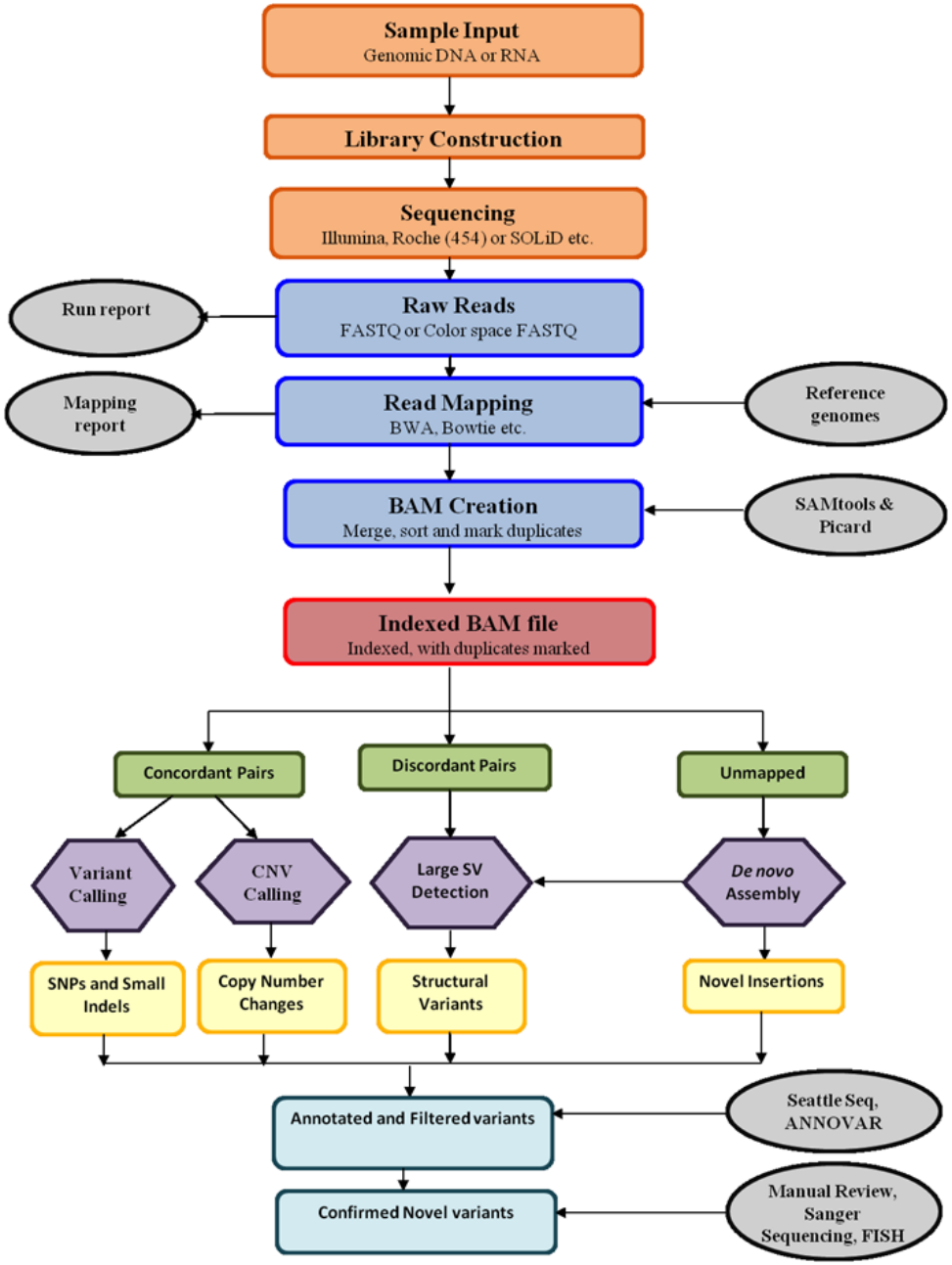
Alignment


Aligned reads
ACGCGATTCAGGTTACCACG
GCGATTCAGGTTACCACGCG
GATTCAGGTTACCACGCGTA
TTCAGGTTACCACGCGTAGC
CAGGTTACCACGCGTAGCGC
GGTTACCACGCGTAGCGCAT
TTACCACGCGTAGCGCATTA
ACCACGCGTAGCGCATTACA
CACGCGTAGCGCATTACACA
CGCGTAGCGCATTACACAGA
CGTAGCGCATTACACAGATT
TAGCGCATTACACAGATTAG
Consensus contig
ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG

Mapping, Coverage reports

- Important checkout for lab protocol
- Specificity of PCR
- Normalization
- Settings of variant calling threshold, CNV




Sample Input
Genomic DNA or RNA
Library Construction
Sequencing
Illumina, Roche (454) or SOLiD etc.
Raw Reads
FASTQ or Color space FASTQ
Run report
Read Mapping
BWA, Bowtie etc.
Mapping report
Reference genomes
BAM Creation
Merge, sort and mark duplicates
SAMtools & Picard
Indexed BAM file
Indexed, with duplicates marked
Concordant Pairs
Discordant Pairs
Unmapped
Variant Calling
CNV Calling
Large SV Detection
De novo Assembly
SNPs and Small Indels
Copy Number Changes
Structural Variants
Novel Insertions
Annotated and Filtered variants
Seattle Seq, ANNOVAR
Confirmed Novel variants
Manual Review, Sanger Sequencing, FISH

SNV and small InDel Calling

- Coverage
- Frequency
- Base quality

- !!!
- Genomic context (homopolymers)
- Nucleotide type
- Position in read (errors at the read end)
- Alignment errors

CNV variations

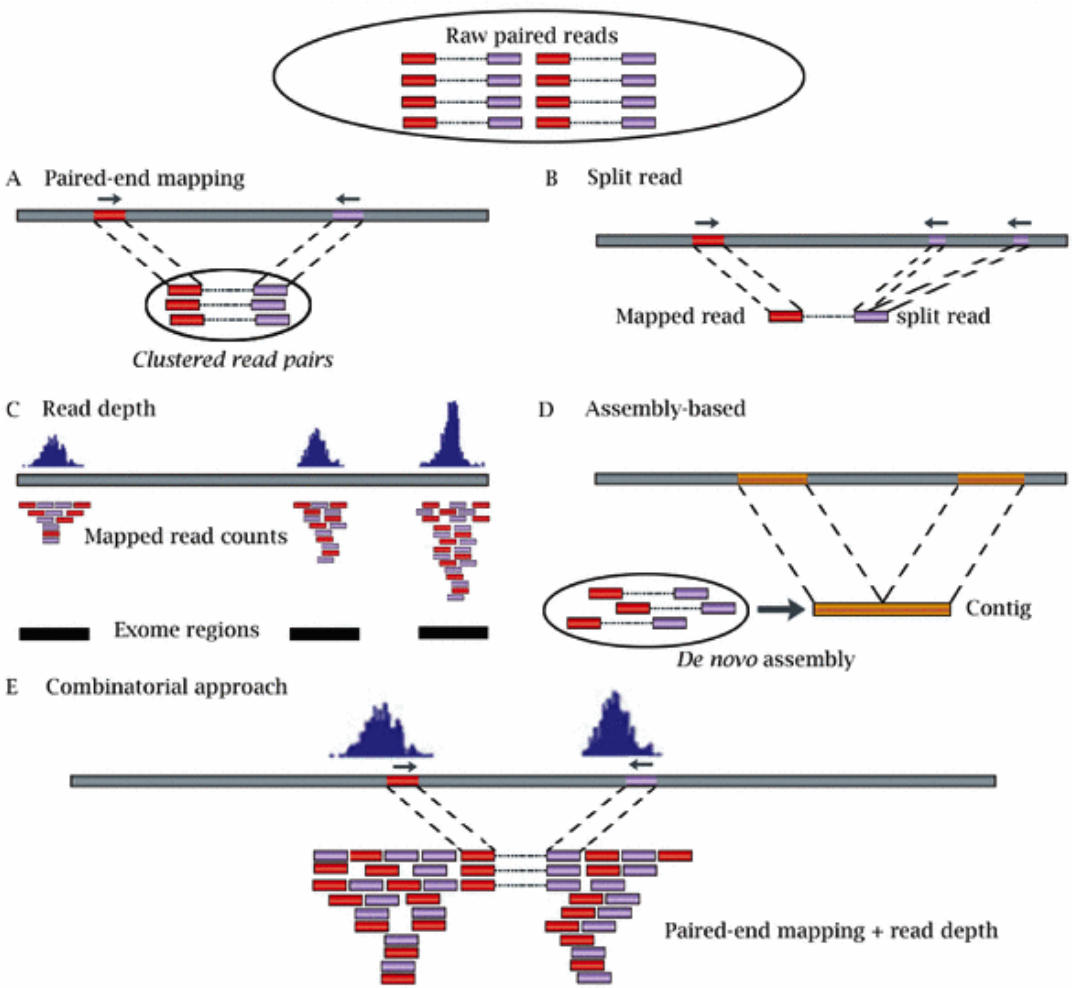
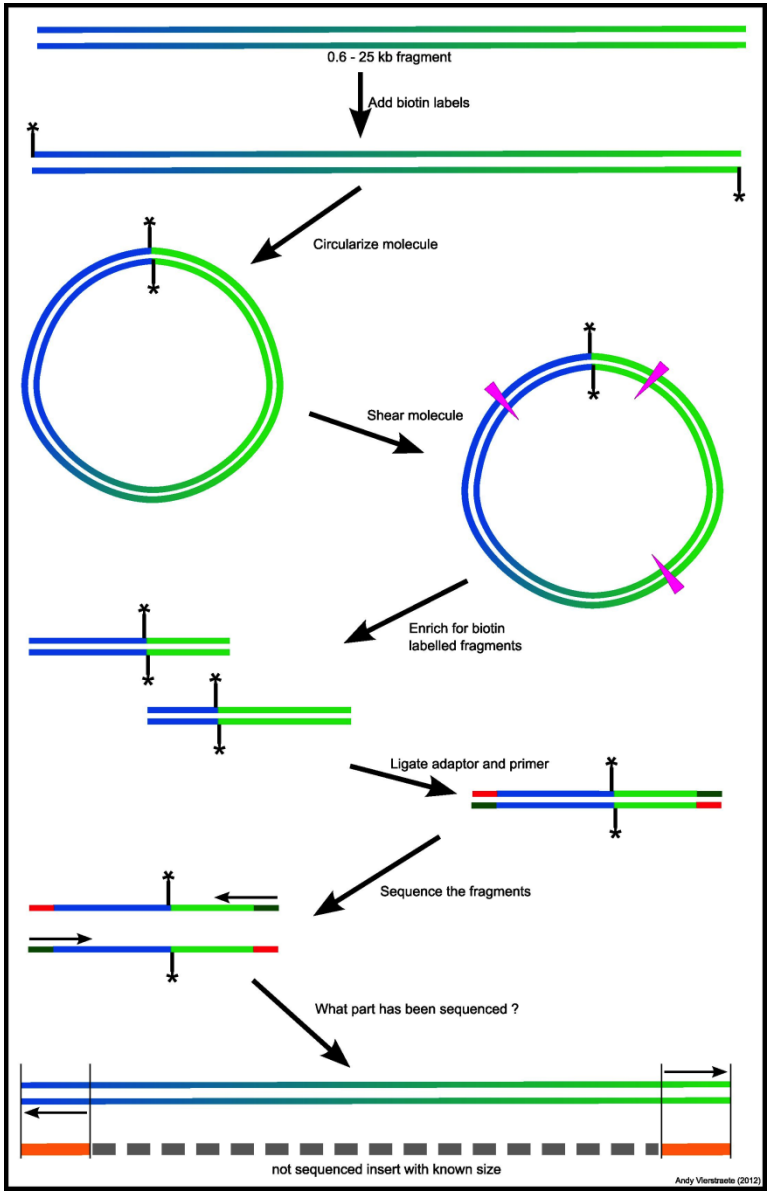
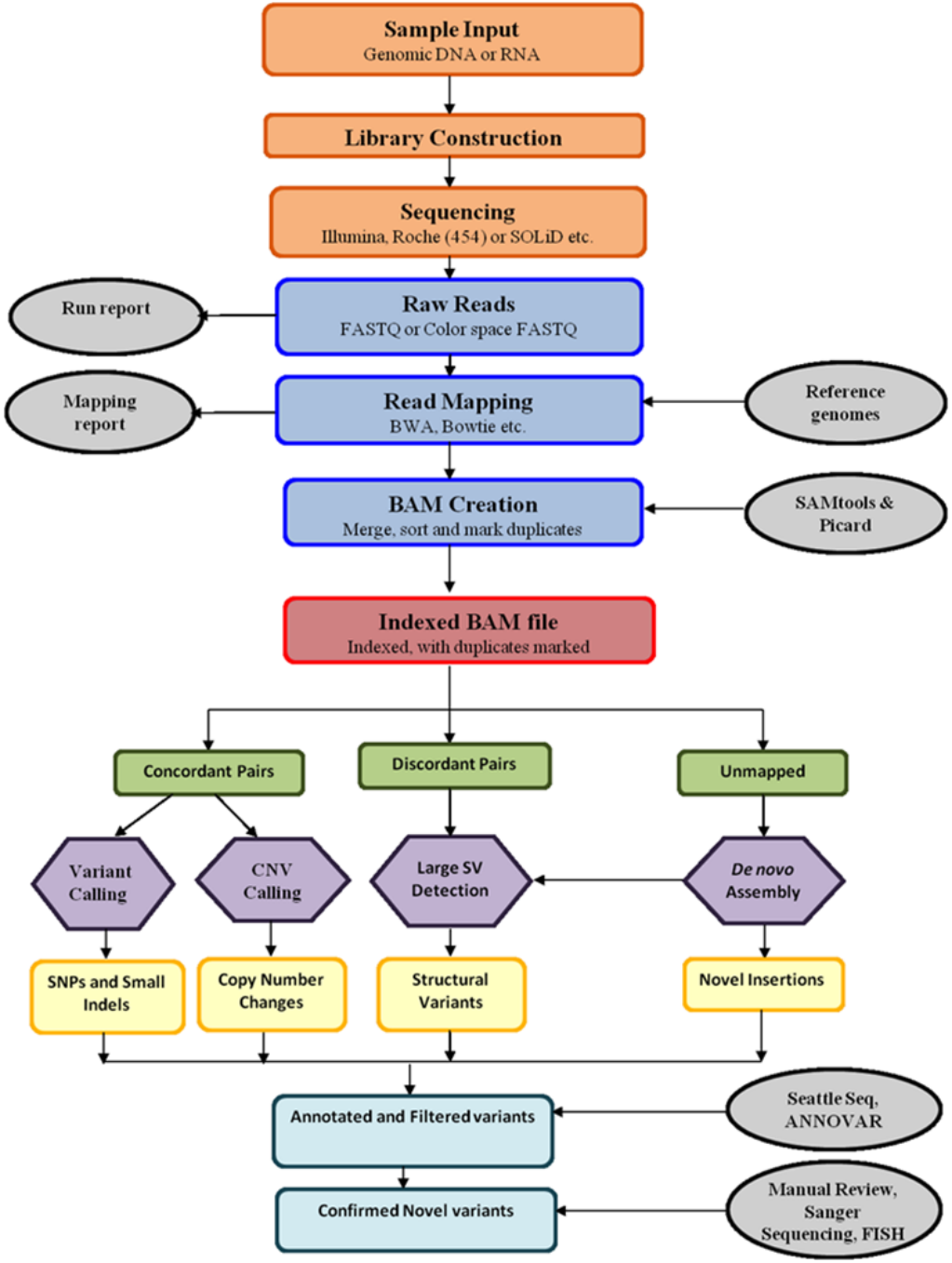


Figure 1 Five approaches to detect CNVs from NGS short reads. A. Paired-end mapping (PEM) strategy detects CNVs through discordantly mapped reads. A discordant mapping is produced if the distance between two ends of a read pair is significantly different from the average insert size. B. Split read (SR)-based methods use incompletely mapped read from each read pair to identify small CNVs. C. Read depth (RD)-based approach detects CNV by counting the number of reads mapped to each genomic region. In the figure, reads are mapped to three exome regions. D. Assembly (AS)-based approach detects CNVs by mapping contigs to the reference genome. E. Combinatorial approach combines RD and PEM information to detect CNVs.

Structural variations

- Mate-pair library
- Long InDel
- Translocation




Sample Input
Genomic DNA or RNA
Library Construction
Sequencing
Illumina, Roche (454) or SOLiD etc.
Raw Reads
FASTQ or Color space FASTQ
Run report
Read Mapping
BWA, Bowtie etc.
Mapping report
Reference genomes
BAM Creation
Merge, sort and mark duplicates
SAMtools & Picard
Indexed BAM file
Indexed, with duplicates marked
Concordant Pairs
Discordant Pairs
Unmapped
Variant Calling
CNV Calling
Large SV Detection
De novo Assembly
SNPs and Small Indels
Copy Number Changes
Structural Variants
Novel Insertions
Annotated and Filtered variants
Seattle Seq, ANNOVAR
Confirmed Novel variants
Manual Review, Sanger Sequencing, FISH

Annotating and filtering

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