UPPSALA UNIVERSITET

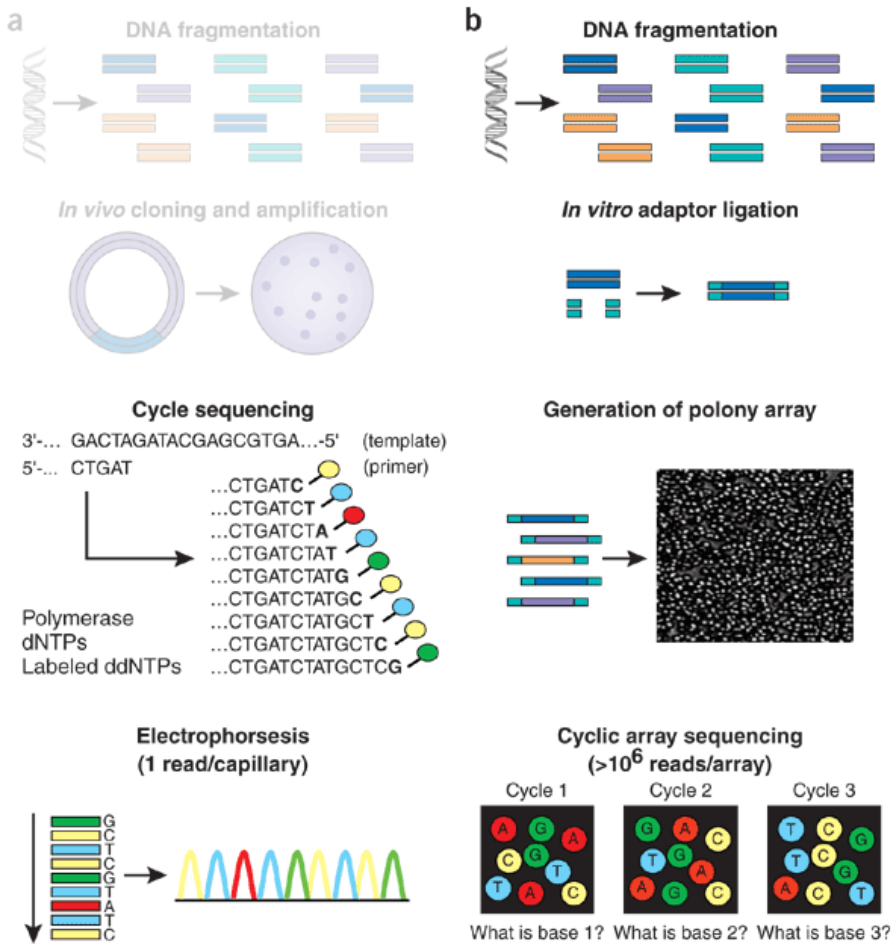
Basic bioinformatics - from fastq to variants

2nd ERIC workshop on TP53 analysis in Chronic Lymphocytic Leukemia

7/11 - 2017

Viktor Ljungström
Department of Immunology, Genetics and Pathology
Uppsala University

Sanger vs Next-generation sequencing

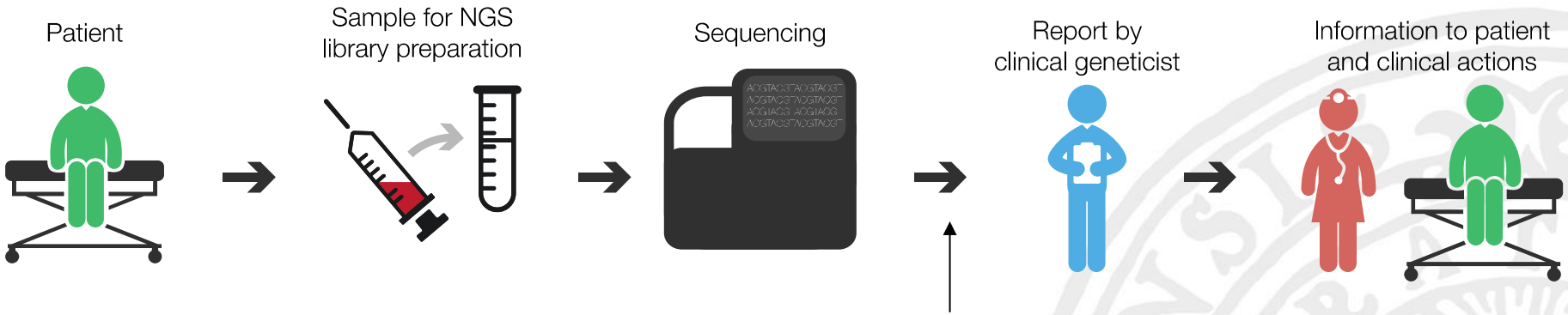


Shender et al., Nature Biotech 2008

Sanger sequencing
- One region in one patient
- Robust
- *Manual* analysis possible

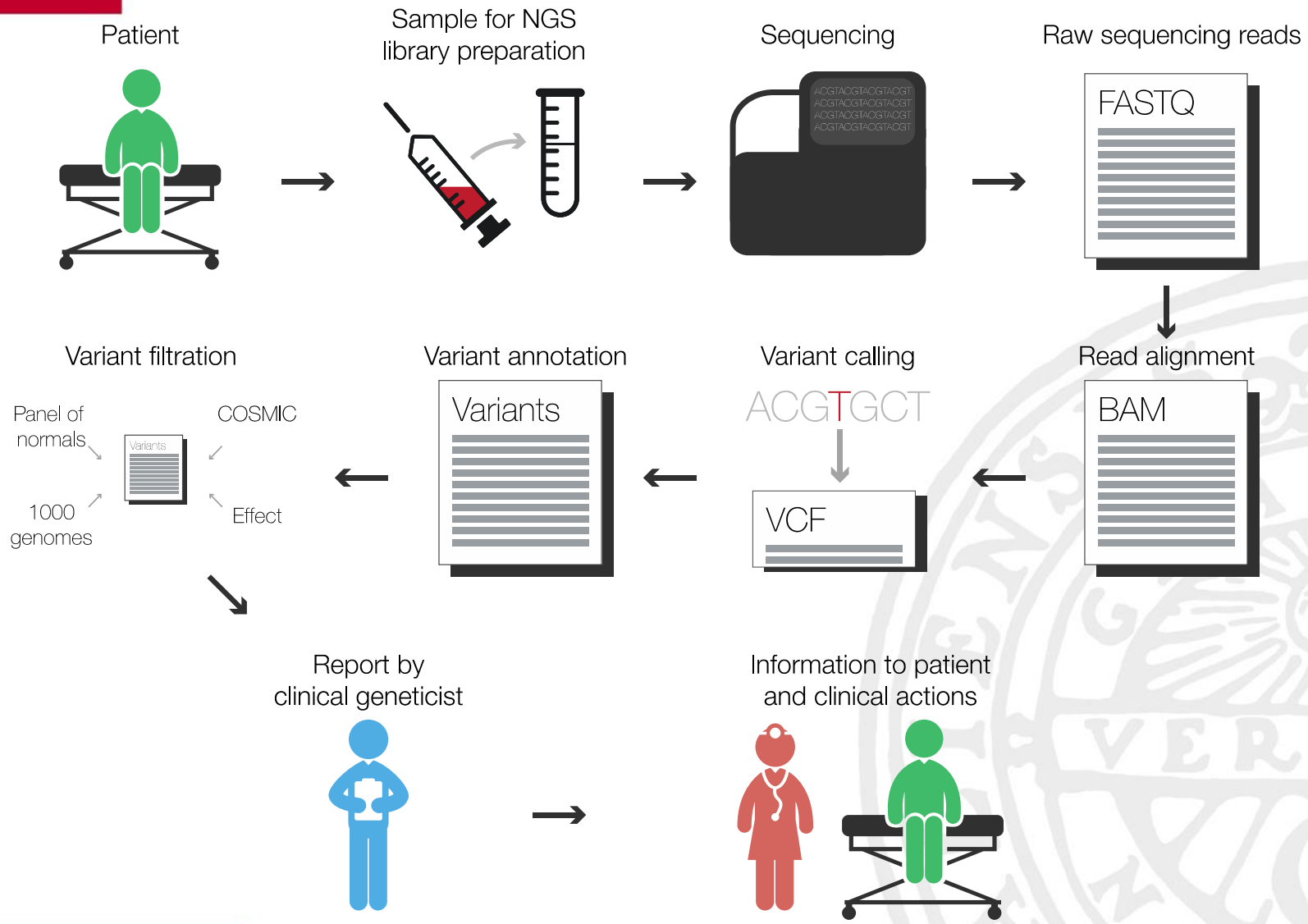
NGS
- Multiplexing regions and patients
- Sensitive
- Need of computational analysis

NGS in the precision medicine workflow



Computational analysis!

NGS in the precision medicine workflow



What is bioinformatics?

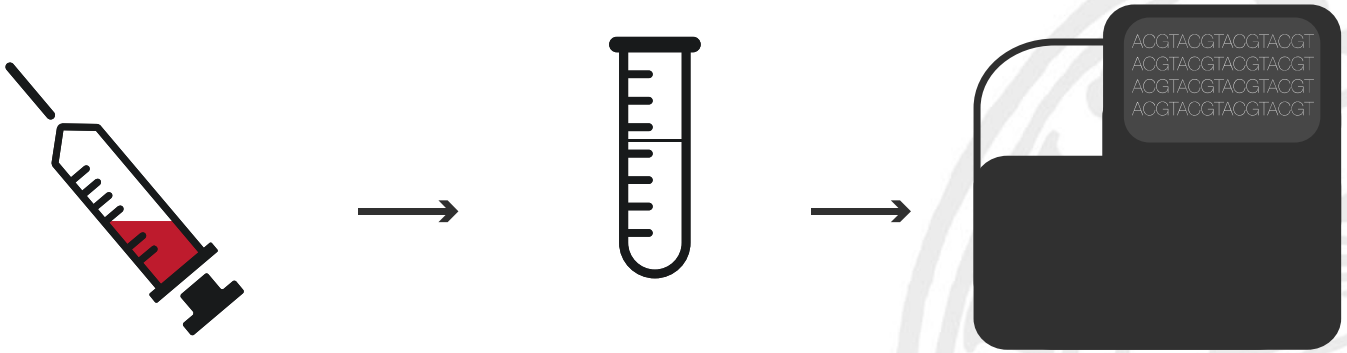
- Broad term
  - From AI to biostatistics
- Here: Computational analysis of NGS data
- From the sequencing machine output to a list of variants that makes sense to the geneticist

Several NGS applications today

Different applications and different platforms

Today: Focus on targeted deep sequencing with Illumina technique

Sample | Library preparation | Sequencing



DNA
RNA

TruSight
HaloPlex
Multiplicom

Illumina
Ion torrent
PacBio
Nanopore

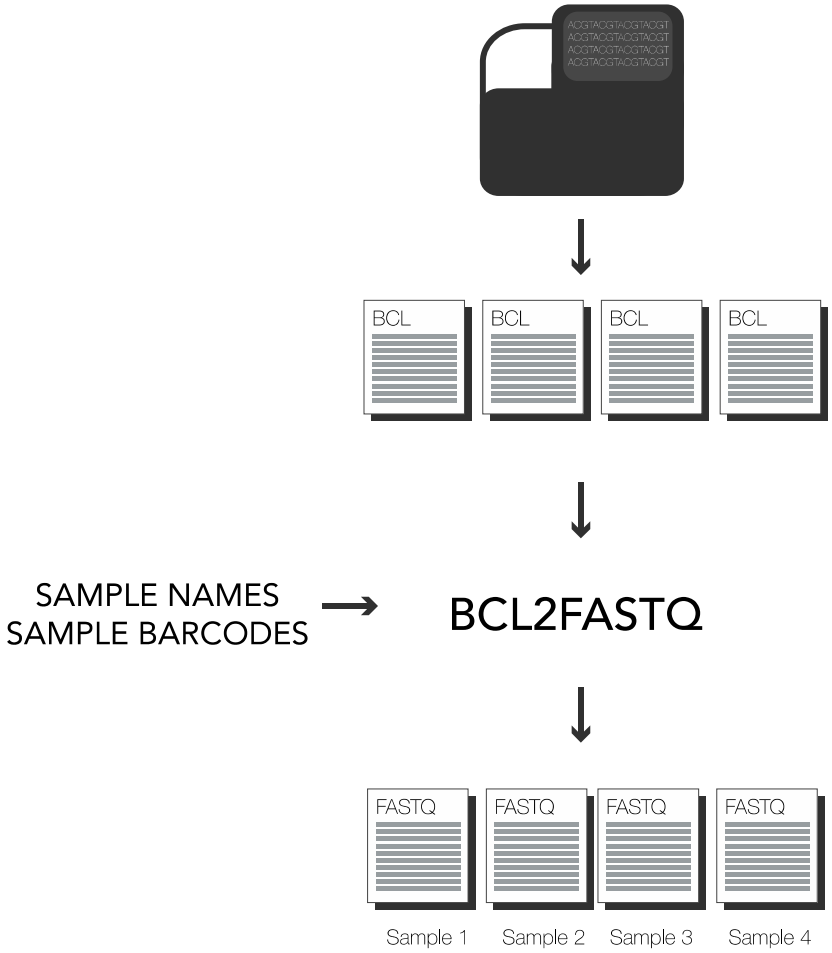
UPPSALA UNIVERSITET

The analysis workflow

UPPSALA UNIVERSITET

The analysis workflow

1. BCL to FASTQ conversion and demultiplexing



- BCL – raw sequencing data
- Convert to FASTQ and split into sample files
- Sample sheet information, DNA barcodes
- Usually automated on the sequencer

The FASTQ format

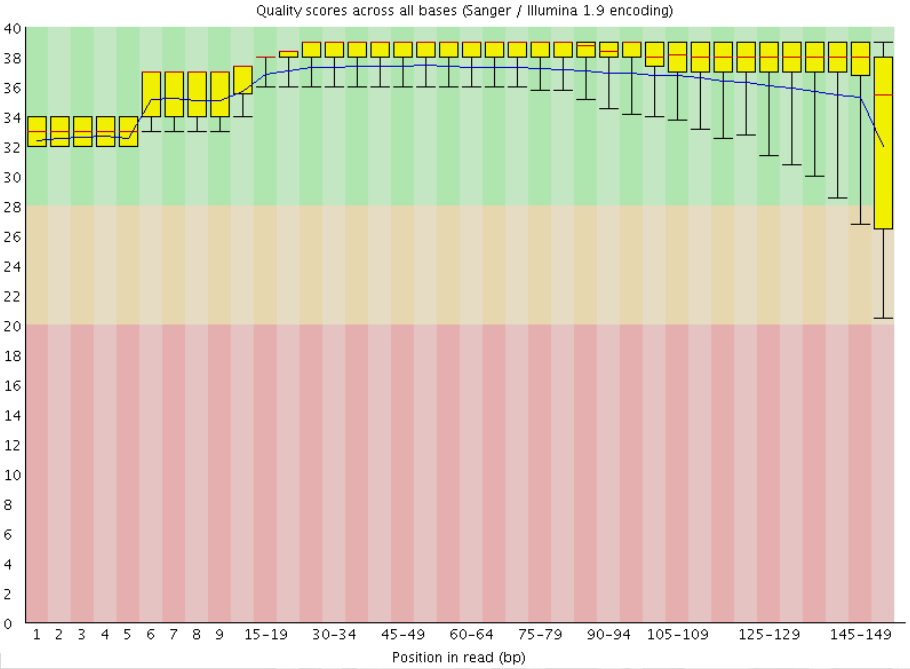
- FASTQ = FASTA + Quality

1. Sequence identifier
2. Nucleotide sequence (the *read*)
3. Phred quality information per base (ASCI encoded)

```
@HISEQ2000-02:420:C2E47ACXX:7:2214:18015:39495/1
CACTCCAGCCTGGGTGACAGAGCGAGATTCCGTCTCAAAAAGTAAAATAAAATAAA
+
EAD@@@?@A@?>>??@@?A?@??>@>ACCAA@A@@@AABAAA?AAAAAAAAAA
```

1. BCL to FASTQ conversion and demultiplexing

- First quality control by eye
  - Are all files present?
  - Are the files of expected size?
- Other quality controls
  - Qscore distribution, GC content, sequence enrichment
- FASTQC

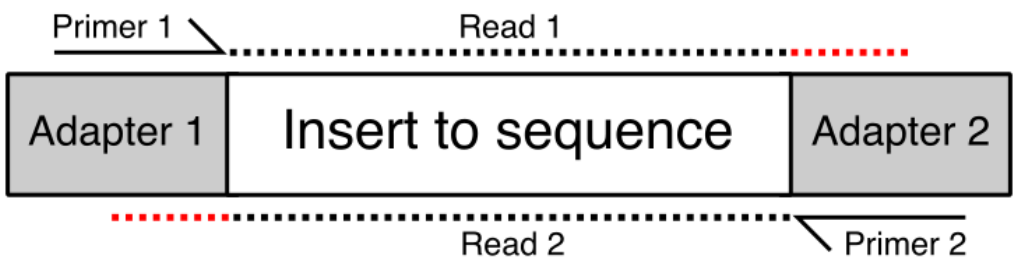


UPPSALA UNIVERSITET

The analysis workflow

2. Read trimming

- Adapter read through
  - Insert shorter than read length
- Low quality bases
- Enzyme footprints (Agilent Haloplex)
- Necessary?



https://sequencing.qcfail.com/articles/read-through-adapters-can-appear-at-the-ends-of-sequencing-reads/

Tool examples:
Cutadapt
Trim Galore!
Agilent Agent

UPPSALA UNIVERSITET

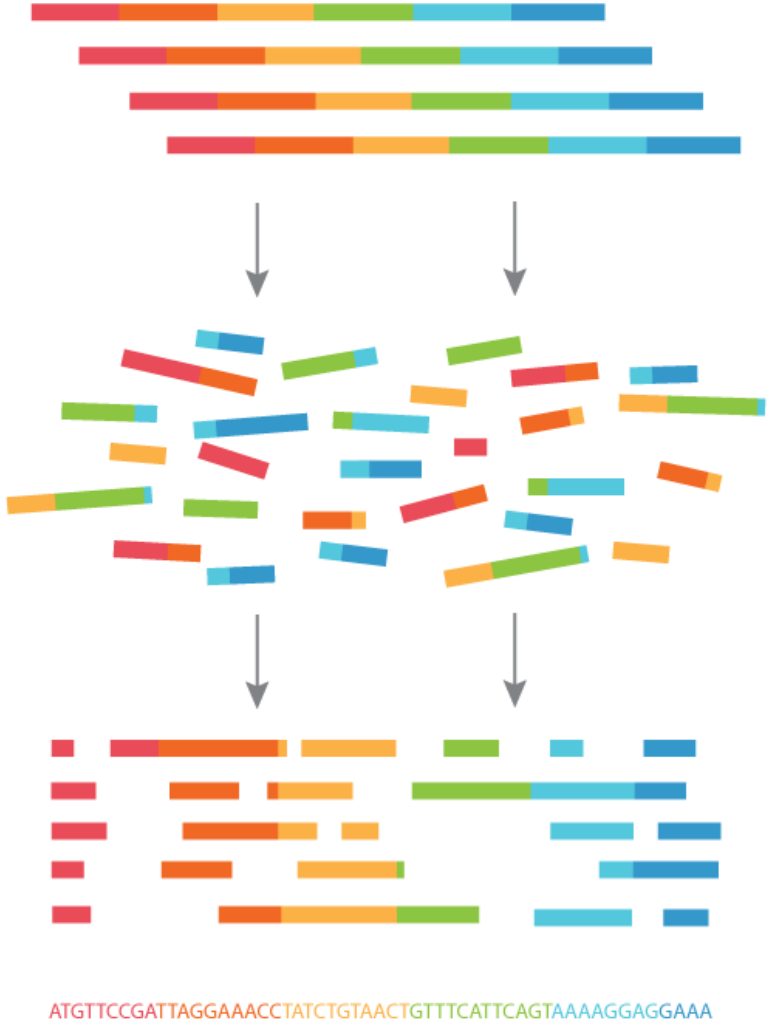
The analysis workflow

3. Read alignment

- Which loci do the read originate from?
- Compare to reference genome
- Technical and biological challenges:
  - The reference is large
  - Somatic and inherited variants?
  Pseudogenes?
- Input: FASTQ files
- Output: SAM/BAM files

Tool examples:
BWA-mem
Novoalign
Bowtie
MOSAIK

The SAM/BAM format



Template DNA

Short reads from Sequencer (FASTQ)

Mapped reads (SAM/BAM-file)

https://www.abmgood.com/marketing/knowledge_base/next_generation_sequencing_data_analysis.php

The SAM/BAM format

- Sequence Alignment/Map format
- Similar to FASTQ but added information

@HISEQ2000-02:420:C2E47ACXX:7:2214:18015:39495. 99. chr1 17644 37 37M = 17919 314
CACTCCAGCCTGGGTGACAGAGCGAGATTCCGTCTCAAAAAGTAAAATAAAATAAAATAAAAAATAAAAGTTTG
EAD@@@?@A@?>>??@@?A?@??>@>ACCAA@A@@@AABAAA?AAAAAAAAAACCCBBBBBAAABA@
RG:Z:UM0098:1. XT:A:R. NM:i:0 SM:i:0. AM:i:0. X0:i:4. X1:i:0. XM:i:0. XO:i:0. XG:i:0. MD:Z:37

Field	
QNAME	@HISEQ2000-02:420:C2E47ACXX:7:2214:18015:39495
FLAG	99
RNAME	chr1
POS	17644
MAPQ	37
CIGAR	37M
MRNM/RNEXT	=
MPOS/PNEXT	17919
ISIZE/TLEN	314
SEQ	CACTCCAGCCTGGGTGACAGAGCG…
QUAL	EAD@@@?@A@?>>??@@?A?@...
TAGs	RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37

UPPSALA UNIVERSITET

The analysis workflow

4. Variant calling

- Is there a variation in the tumor sequence compared to the reference?
- Small variants:
  - Single nucleotide variants (SNVs)
  - Insertions and deletions < ~20bp (InDels)
- Input file: BAM-file
- Output file: VCF-file

4. Variant calling

Aligned reads

C>G mutation

```
GTCTGGCCCCTCGTCAGCATCTTATCC
  CTGGCCCCTCCTCAGCATCTTATCC
 TCTGGCCCCTCGTCAGCATCTTATCCG
  CTGGCCCCTCCTCAGCATCTTATCC
  CTGGCCCCTCGTCAGCATCTTATCCG
  CTGGCCCCTCCTCAGCATCTTATCCG
    GCCCCTCGCTCAGCATCTTATCCG
  CTGGCCCCTCGTCAGCATCTTATCCG
  CTGGCCCCTCCTCAGCATCTTATCCG
  CTGGCCCCTCCTCAGCATCTTATCCG
  CTGGCCCCTCGTCAGCATCTTATCCG
```

GT deletion

```
GAAATTTGCGTGTGGAGTATTTGGATGACAG
GAAATTTGC   GTGGAGTATTTGGATGACAGA
GAAATTTGCGTGTGGAGTATTTGGATGACAGA
GAAATTTGCGTGTGGAGTATTTGGATGACAG
GAAATTTGC   GTGGAGTATTTGGATGACAGA
 AAATTTGC   GTGGAGTATTTGGATGACAGA
 AAATTTGC   GTGGAGTATTTGGATGACAGA
 AAATTTGCGTGTGGAGTATTTGGATGACAG
 AAATTTGC   GTGGAGTATTTGGATGACAG
 AAATTTGC   GTGGAGTATTTGGATGACAG
 AAATTTGCGTGTGGAGTATTTGGATGAC
```

GTCTGGCCCCTCCTCAGCATCTTATCCGAGTGGAAGGAAATTTGCGTGTGGAGTATTTGGATGACAGAAACACTTTTCGACATAGT

Reference sequence

Captured region 1

Captured region 2

4. Variant calling

Tool examples:
VarScan2 (U+P)
Mutect2 (P)
Strelka (U+P)
GATK (U)

- Reports all detectable variation
  - Unaware of effects and gene borders
  - Biological and technical variation
- Paired vs unpaired (somatic / germline)
  - Unpaired: Direct comparison to reference genome
  - Paired: Filter against matched normal sample – germline and noise removal
  - True germline callers may not be best suited for cancer samples

The variant call format - VCF

- Raw output from the variant caller
- Variant and its position + technical data
  - Read depth (11x)
  - VAR (5/11 ≈ 45%)
  - Quality score
- No gene information

```
GTCTGGCCCCTCGTCAGCATCTTATCC
  CTGGCCCCTCCTCAGCATCTTATCC
 TCTGGCCCCTCGTCAGCATCTTATCCG
  CTGGCCCCTCCTCAGCATCTTATCC
  CTGGCCCCTCGTCAGCATCTTATCCG
  CTGGCCCCTCCTCAGCATCTTATCCG
    GCCCCTCGCTCAGCATCTTATCCG
  CTGGCCCCTCGTCAGCATCTTATCCG
  CTGGCCCCTCCTCAGCATCTTATCCG
  CTGGCCCCTCCTCAGCATCTTATCCG
  CTGGCCCCTCGTCAGCATCTTATCCG

GTCTGGCCCCTCCTCAGCATCTTATCCGAGT
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr1	211545955	.	G	A	.	.	AO=34296;DF
chr11	108137985	.	C	T	.	.	';AO=1100;DF
chr17	7578266	.	T	A	.	.	AO=13131;DF

UPPSALA UNIVERSITET

The analysis workflow

5. Variant annotation

- Information from genomic databases
- Add information to each variant
  - Gene name
  - Transcript
  - Amino acid consequence
  - dbSNP / 1000 genomes
  - COSMIC

Tool examples:
Annovar
Oncotator
Nirvana
SeattleSeq Annotation

5. Variant filtration - biological

- Clinical setting – usually no matched normal
  - Remove *unimportant* variants
- Remove known germline variants in population
  - Improving databases (e.g. dbSNP -> 1000 genomes -> 1000 genomes Europe -> SweGen)
  - Careful with patient samples of other genetic background
- Remove non-coding and synonymous variants
  - UTR3' and 5'?
  - Splice variants?

5. Variant filtration - technical

- Clinical setting – usually no matched normal
  - Remove technical errors/noise
- Technical quality of variants
  - VAR cutoff
  - Read depth cutoff
  - Variant quality score cutoff (?)
- Panel of normals / negative controls?
  - Potentially efficient for recurrent panel errors
  - How many samples?

UPPSALA UNIVERSITET

The analysis workflow

6. Quality control

- General quality of the sequencing run
  - Base qualities
  - Sequencing yield
  - Over/under clustering
  - Percent on target reads
- Sample specific QC
  - Depth of coverage
  - MAPQ
  - % reads mapped

6. Quality control

- No consensus yet

Quality Measure	ACGS	ACMG	CAP	EuroGentest	RCPA	CDC
Base quality	X		X	X	X	X
GC bias	X			X	X	X
Distinct reads		X		X	X	X
Reads mapped (%)	X			X		
Percentage of reads on target region		X	X	X		
Mapping quality	X		X	X	X	X
Average depth on target region	X			X	X	X
Uniformity		X		X	X	X
Off Target						X
Percent of target captured	X	X	X			

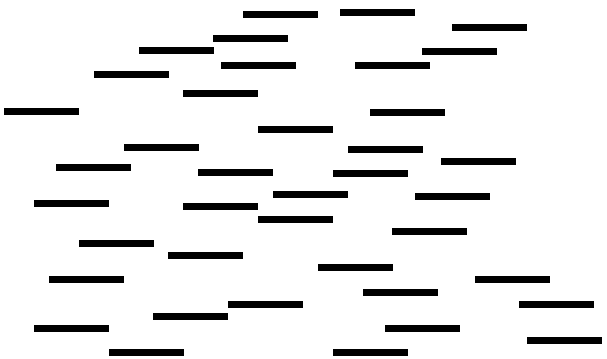
http://euformatics.com/evolving-standards-in-clinical-ngs/

Depth of coverage

- The number of times a base-pair is covered by aligned reads
- Targeted deep sequencing: Mean coverage within the target regions

Depth of coverage

Unaligned reads



X gigabases of sequencing
- Per sample?
- Per region/base?

Depth of coverage

- The number of times a base-pair is covered by aligned reads
- Targeted deep sequencing: Mean coverage within the target regions
- Best cutoff metric
  - Mean coverage?
  - Percent bases covered 100x/1000x?
  - Target specific?

Tool examples:
Sambamba
Samtools
Bedtools

6. Quality control and inspection

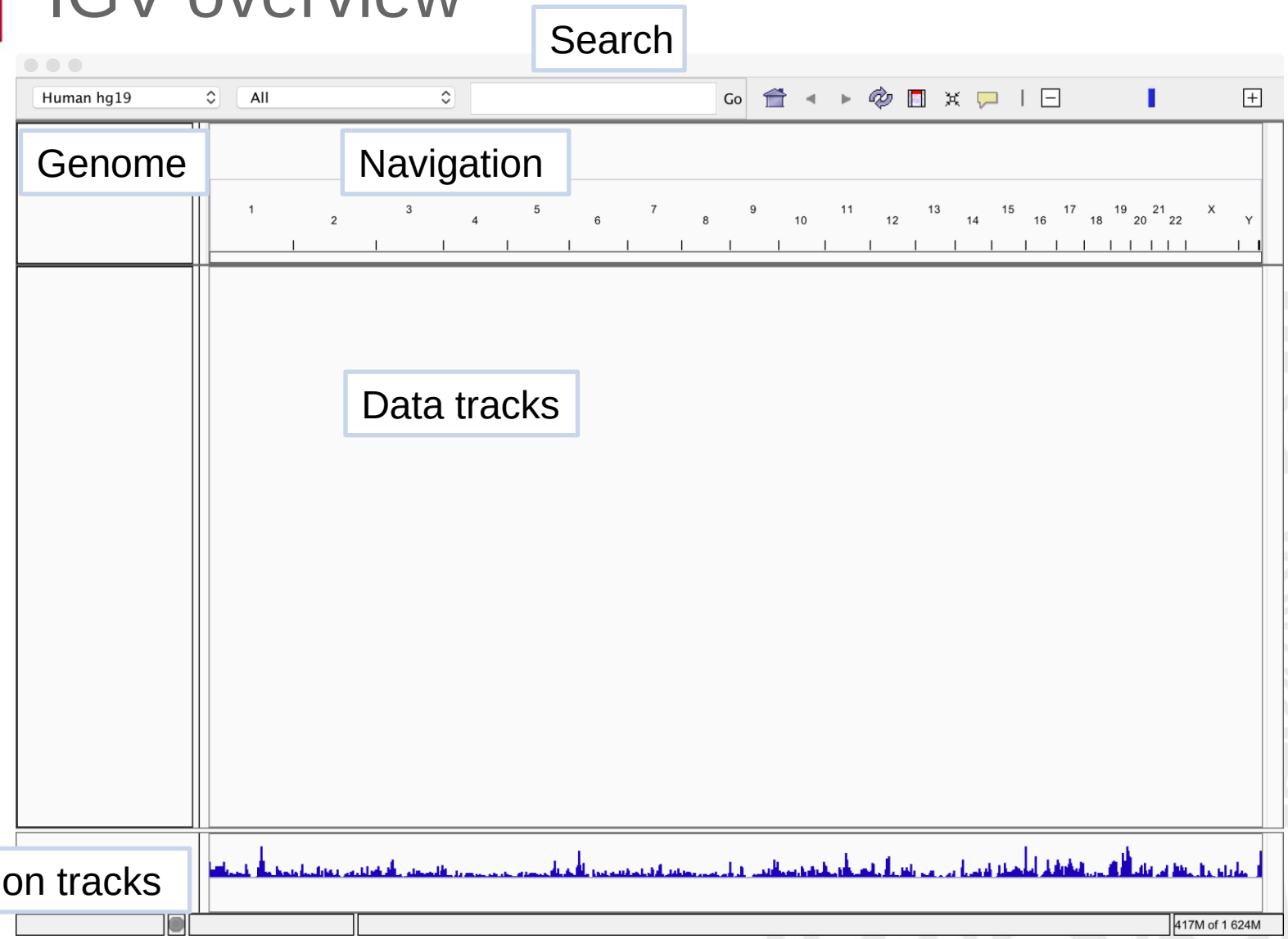
- Variant lists good for big data quantities
- Information about a specific variant?
- Inspect problematic regions and alignment results

- IGV

What is IGV?

- **I**ntegrative **G**enomics **V**iewer
- Desktop genome browser - "visualization tool for interactive exploration of large, integrated genomic datasets”
- Display reads and variants
- Runs locally

IGV overview



IGV input file formats

- **BAM-file**
  - coordinate sorted
  - indexed
- **BED-files**
- **VCF-files**
- **Many others**

What can we do in IGV?

1. **Inspect alignments and coverage**
   - File > Load from file > Select BAM file
   - Reset: File > New session

BAM file overview

Zoom

Coverage track

Alignments

Double click to zoom
Drag to move
Zoom in to show variants
Right click: Collapsed/Expanded

Annotation tracks

What can we do in IGV?

1. **Inspect alignments and coverage**
2. **Inspect SNVs**

Chr	Start	End	Reference_base	Variant_base	Gene	Type	Exonic_type	Variant_allele_ratio%	#reference_alleles	#variant_alleles	Read_depth
chr17	7578466	7578466	G	A	TP53	exonic	nonsynonymous SNV	66,88	52	105	157

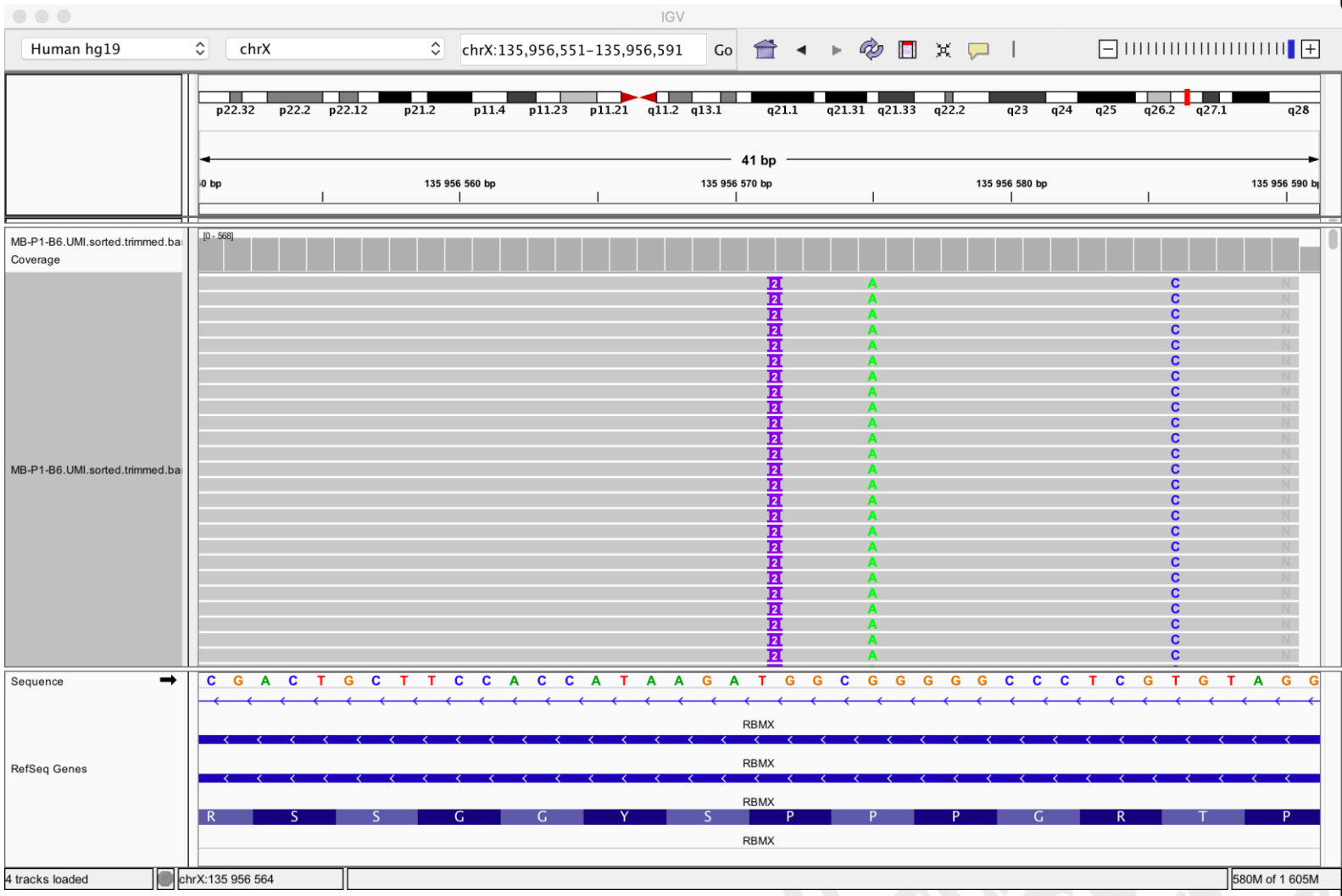
Variant inspection (SNVs)

Clean reads?
Surrounding reads?
Surrounding indels?

What can we do in IGV?

1. **Inspect alignments and coverage**
2. **Inspect SNVs**
3. **Inspect InDels**

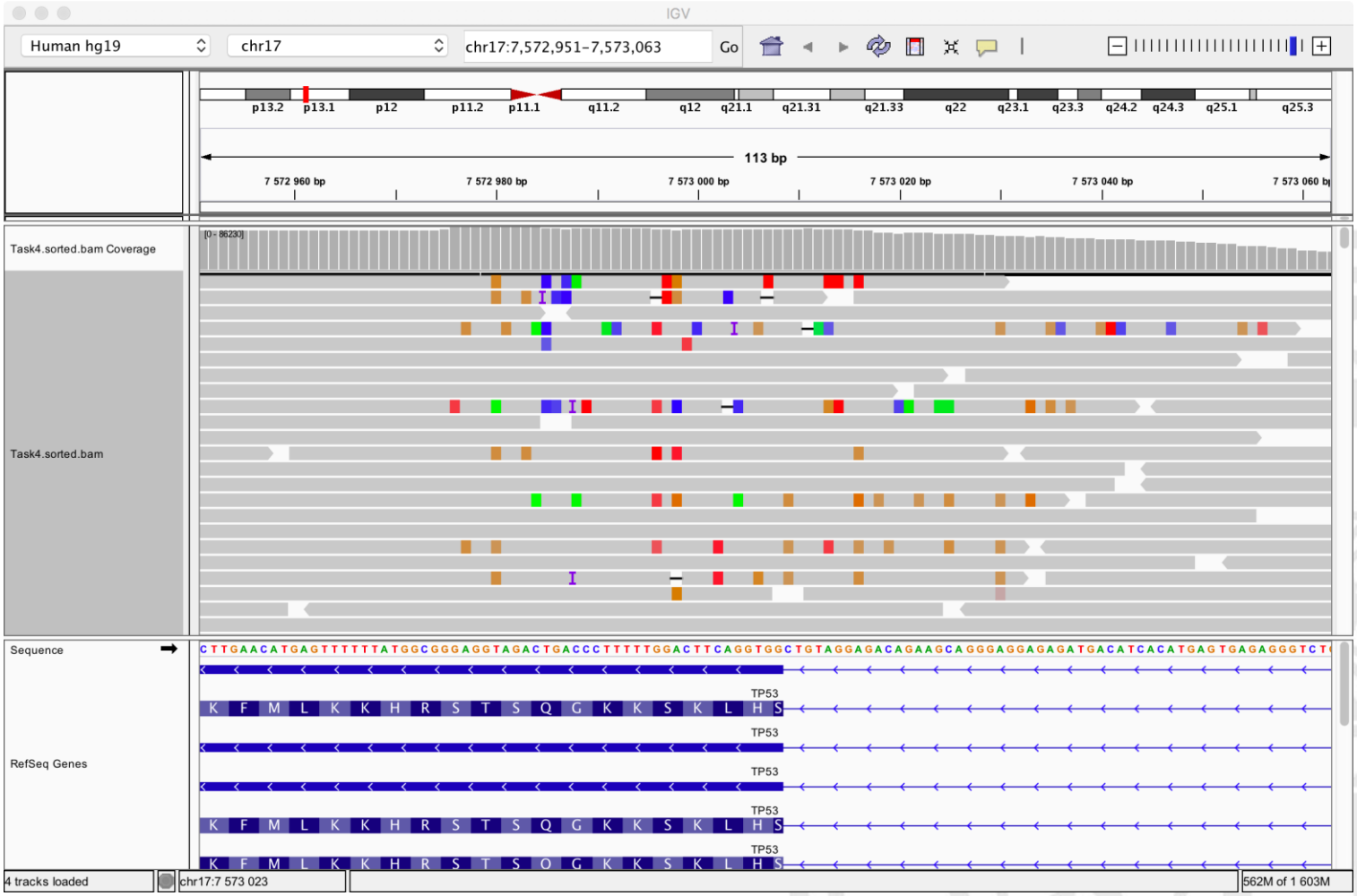
Variant inspection (insertion)



What can we do in IGV?

1. **Inspect alignments and coverage**
2. **Inspect SNVs**
3. **Inspect InDels**
4. **Inspect low quality variants**

Variant inspection (low quality SNV)

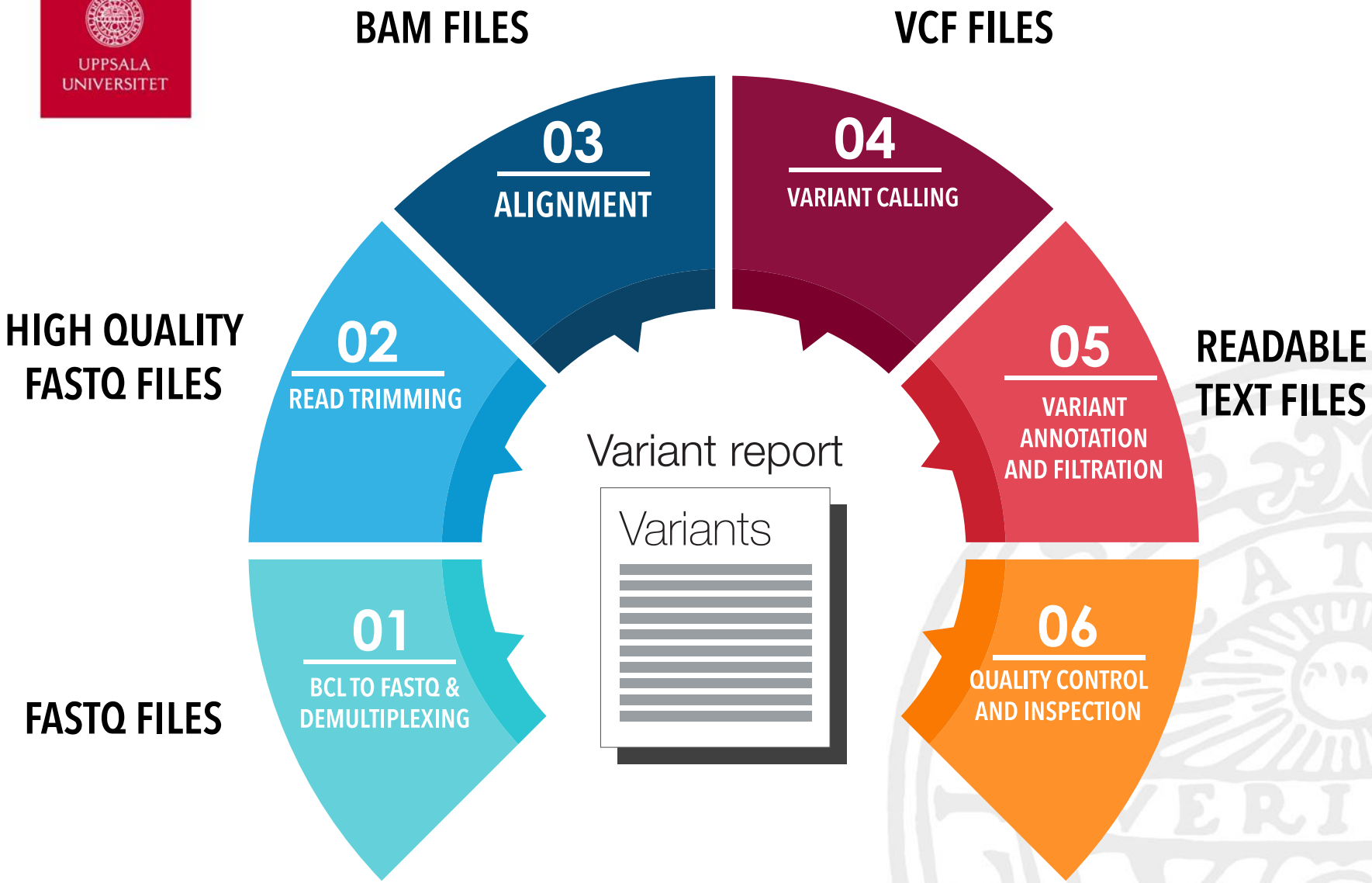


UPPSALA UNIVERSITET

More IGV in the hands-on workshop tomorrow

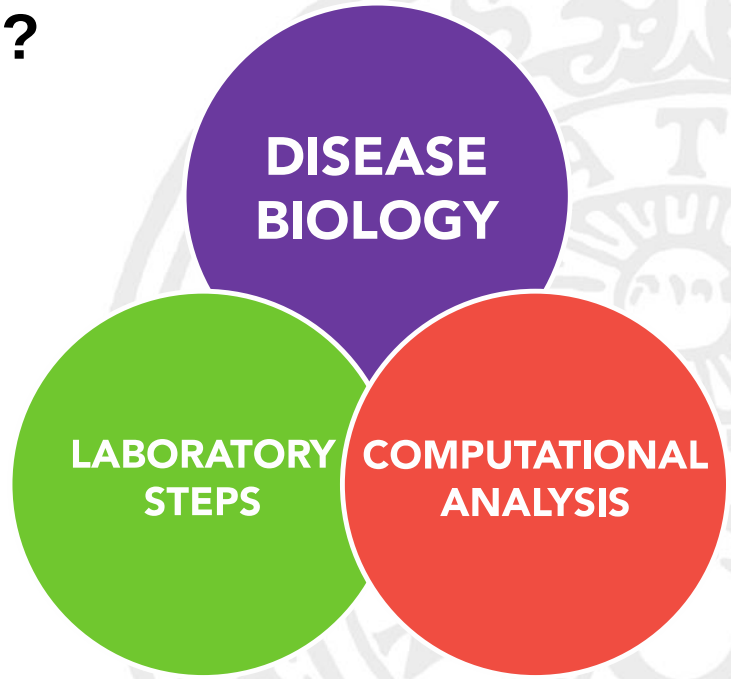
Read the email and download IGV tonight

UPPSALA UNIVERSITET



Final remarks

- **Which tools to use**
  - Open source vs proprietary software
  - Still no best practice on the somatic side
- **Bioinformatics pipelines**
  - Feeding from one tool to another
  - Can we agree on one?
- **Cloud solutions**
- **Bioinformatics**
  - Part of the puzzle
- **Future**
  - UMI analysis
  - CNV analysis



Acknowledgements

CERTH, Thessaloniki

Stavroula Ntoufa

Kostas Stamatopoulos

CEITEC, Brno

Karla Plevova

Jana Kotaskova

Sarka Pospisilova

Lund University

Gunnar Juliusson

NIHR, Oxford

Ruth Clifford

Anna Schuh

G. Papanicolaou Hospital, Thessaloniki

Niki Stavroyianni

IRCCS San Raffaele, Milan

Andreas Agathangelidis

Paolo Ghia

University of Southampton

Stuart Blakemore

Jonathan C. Strefford

Karolinska Institutet, Stockholm

Karin E. Smedby

University Hospital, Kiel

Christiane Pott

Feinstein Institute, NY

Nicholas Chiorazzi

Nikea Hospital, Athens

Chrysoula Belessi

Hopital Pitie-Salpetriere, Paris

Frederic Davi

Padua University

Livio Trentin

Royal Bournemouth Hospital

David Oscier

University of Athens

Panagiotis Panagiotidis

Erasmus MC, Rotterdam

Anton W. Langerak

University of Eastern Piedmont Novara

Davide Rossi

Gianluca Gaidano

Acknowledgements

Richard Rosenquist
Tobias Sjöblom
Larry Mansouri
Mats Nilsson

Panagiotis Baliakas
Sujata Bhoi
Diego Cortese
Karin Larsson
Mattias Mattson
Aron Skaftason
Lesley-Ann Sutton
Emma Young

Tom Adlerteg
Karin Hartman
Snehangshu Kundu
Chatarina Larsson
Lucy Mathot
Verónica Rendo
Ivaylo Stoimenov

Lucia Cavalier
Claes Ladenwall
Malin Melin
Lotte Moens
Tatjana Pandzic
Johan Rung
Patrik Smeds

UPPSALA
UNIVERSITET

Thank you!