GÖTEBORGS UNIVERSITET

Bioinformatics
Core Facility

IDENTIFYING A DISEASE CAUSING MUTATION

MARCELA DAVILA

2/03/2017

GÖTEBORGS UNIVERSITET

Bioinformatics
Core Facility

Core Facilities at Sahlgrenska Academy

The individual centres

- Bioinformatics
- Centre for Cellular Imaging (CCI)
- Mammalian Protein Expression (MPE)
- Proteomics
- Genomics

www.cf.gu.se

GÖTEBORGS UNIVERSITET

Bioinformatics
Core Facility

Bioinformatics Core Facility

- 5 statisticians, 3 bioinformaticians
- Consultation
- 7-8 Courses / year

Contact information

Visiting address:
Medicinaregatan 3B, F1000-2000

bioinformatics@gu.se

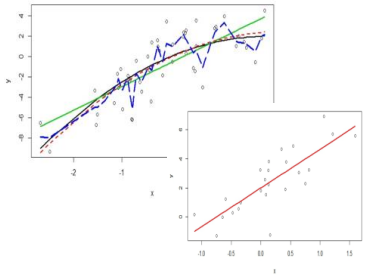
www.cf.gu.se/english/Bioinformatics/

GÖTEBORGS UNIVERSITET

Bioinformatics
Core Facility

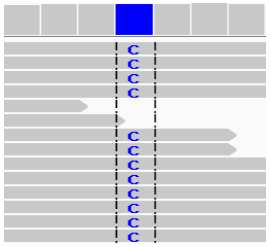
Projects

Clustering



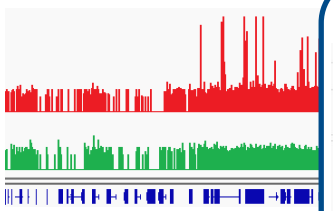
Statistical modeling

Exome-seq



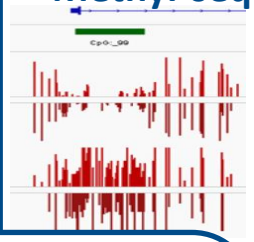
SNV detection

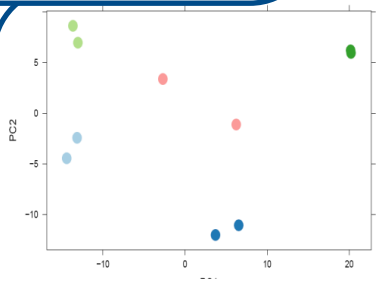
ChIP-Seq



Peak detection

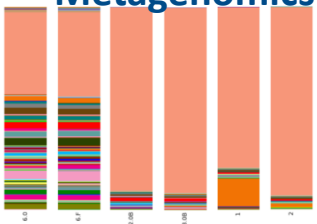
Methyl-seq





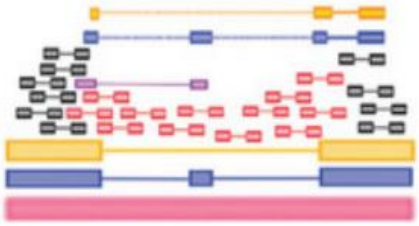
Variability predictors

Metagenomics

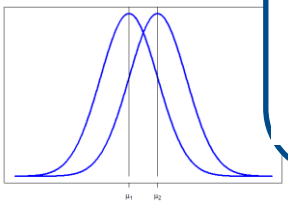


Taxa summary

RNA-seq

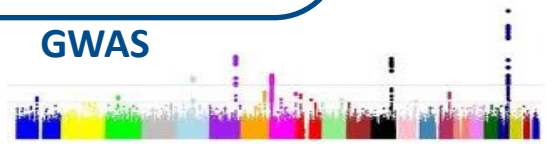


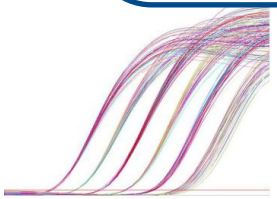
Transcript estimation

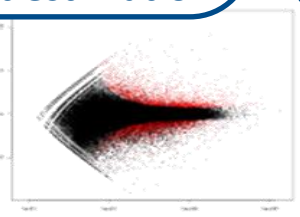


Statistical tests

GWAS







Expression analysis

GÖTEBORGS UNIVERSITET

Bioinformatics Core Facility

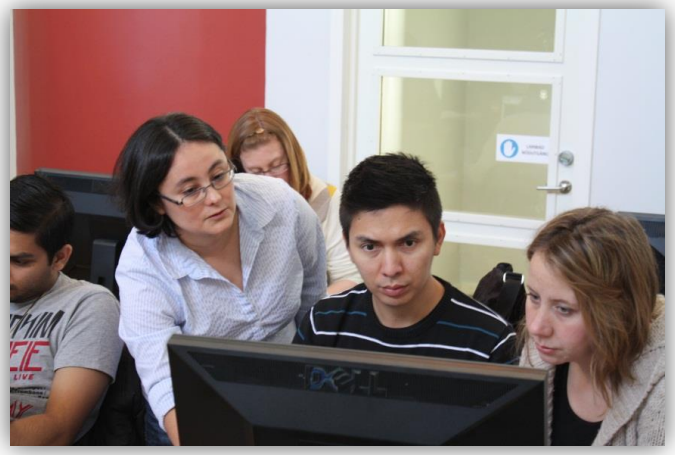
Increasing statistical and bioinformatics knowledge

2017

Week	Course
4	Unix
5	
6	
7	
8	
9	
10	Bioinformatics in Genomics
11	
12	
13	
14	
15	
16	
17	
18	Python
19	
20	
21	
22	Multivarate Analysis
23	
24	

Week	Course	
36	Unix	Microarray analysis using R
37		
38		
39		
40		
41		
42		
43	Advanced NGS	
44		
45		
46		
47	Python	
48		
49		
50		
51		

- Courses
- Seminars and workshops
- Personalized training



GÖTEBORGS UNIVERSITET

Bioinformatics
Core Facility

Supporting local bioinformaticians

Master's thesis projects

Currently available projects

Analysis of the Ig heavy chain repertoire i the absence of SL chain (project plan)
Contact: Lill Mårtensson-Bopp, Inst. of Medicine

In search for the cell of origin in sarcoma. Transcriptome and DNAmethylome analysis of local and public databases combined with wet experiment data (project plan)
Contact: Pierre Åman (phone: 0706-846085), Sahlgrenska Cancer Center, Dept. of Pathology

Estimating minimum host population size for Varicella zoster virus given different assumptions of reinfections (project plan)
Contact: Peter Norberg (phone: 0735-316166), Dept. of Infectious Medicine

Continuous Vector Space Models for Medical Terms (project plan)
Contact: Devdatt Dubhashi, Department of Computer Science and Engineering, Chalmers University of Technology

Latent Topic Models for Medical Documents (project plan)
Contact: Devdatt Dubhashi, Department of Computer Science and Engineering, Chalmers University of Technology

Acute myeloid leukemia analyzed with exome sequencing (project plan)
Contact: Linda Fogelstrand (phone: 46 31 342 9296), Department of Clinical Chemistry and Transfusion Medicine

http://cf.gu.se/english/Bioinformatics/education_and_training/master-s-thesis-projects

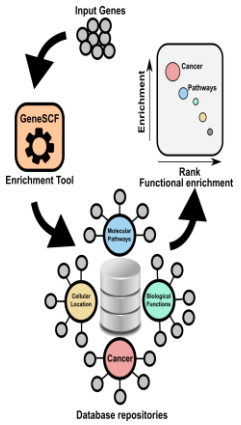
Supporting local bioinformaticians

GOTBIN
GOTHENBURG BIOINFORMATICS NETWORK

https://groups.google.com/forum/#!forum/gotbin

1) …to maintain a **voluntary registry**
2) …to maintain a **list of hardware and software resources**
3) …to **promote and forward bioinformatics in Gothenburg**

GeneSCF



Hirbin

Metaxa2

piano
For R/Bioconductor
PLATFORM FOR INTEGRATIVE ANALYSIS OF OMICS DATA


GÖTEBORGS
UNIVERSITET

Identifying a disease causing mutation

	Alpers disease	Polyglucosan Body Myopathy
Clinical findings		
	Psychomotor regression Feeding difficulties	Leg weakness accumulation of polyglucosan
Prior knowledge		
	POLG1	???

Identifying a disease causing mutation

Alpers disease

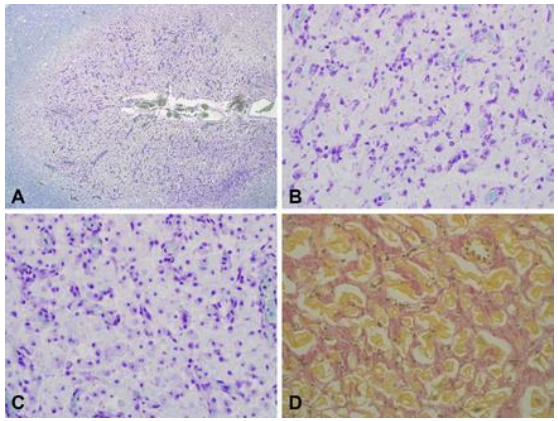
Polyglucosan Body Myopathy

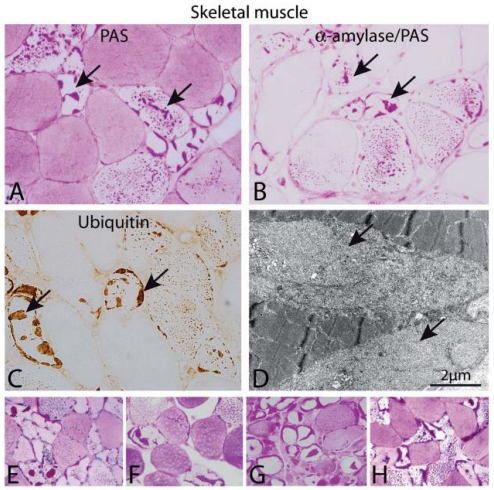
Clinical findings

Psychomotor regression
Feeding difficulties

Leg weakness
accumulation of polyglucosan

Morphology





Identifying a disease causing mutation

Alpers disease

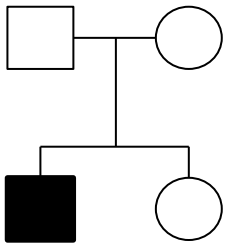
Polyglucosan Body Myopathy

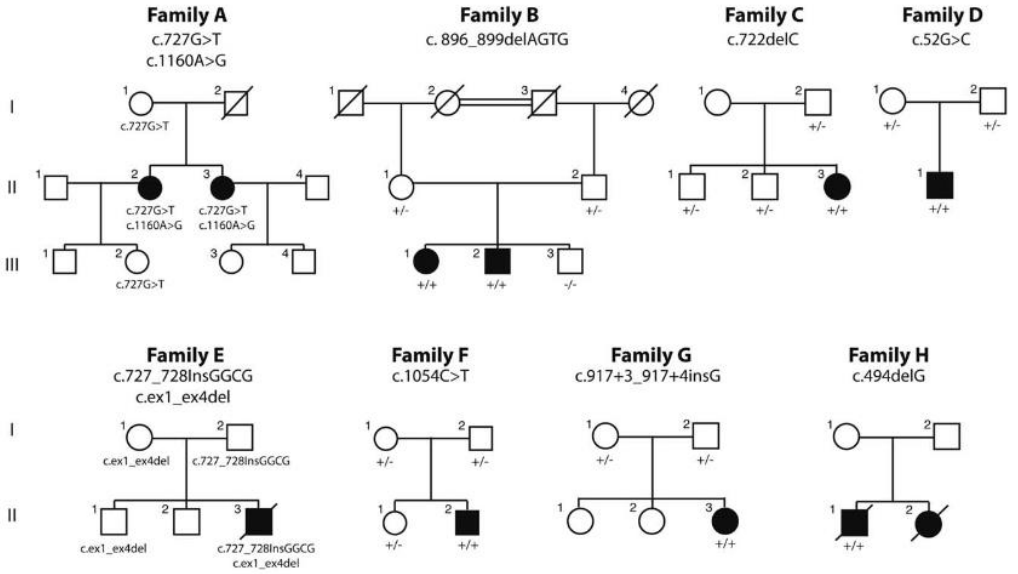
Prior knowledge

POLG1

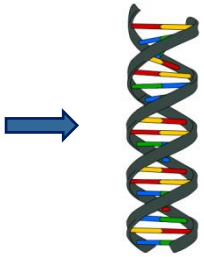
???

Cases

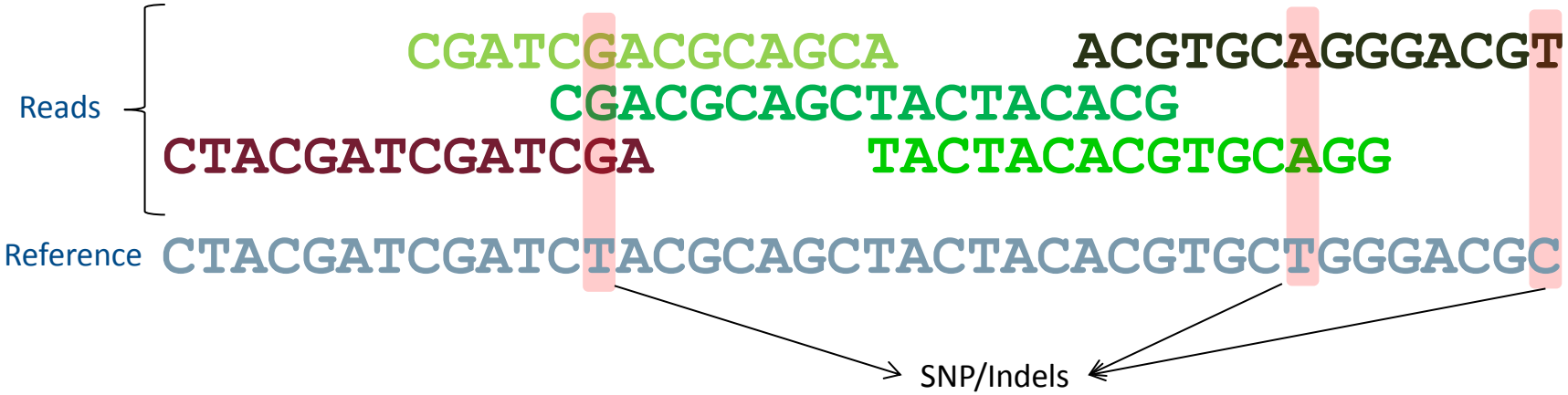




Targeted resequencing



Sample preparation
sequencing



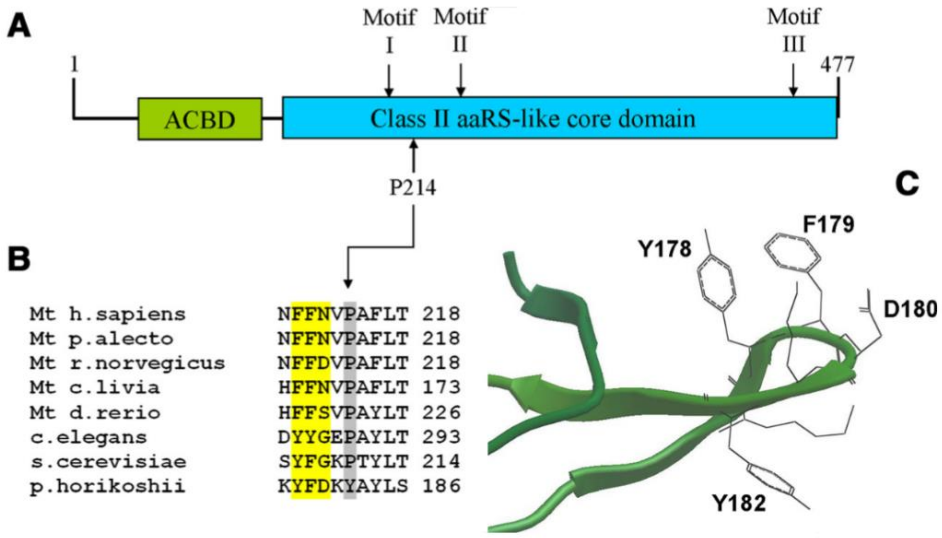
Identifying a disease causing mutation

Alpers disease

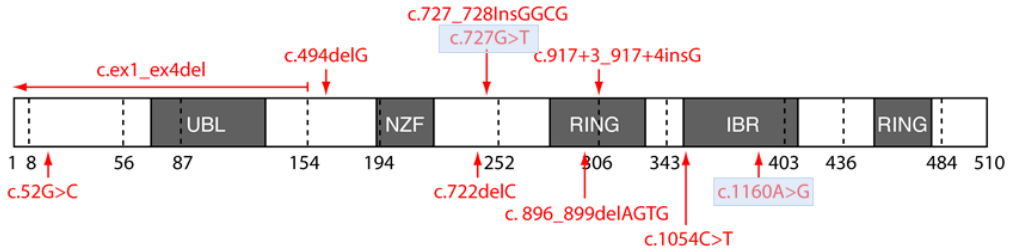
Polyglucosan Body Myopathy

Functional analysis

NARS2



RBCK1




GÖTEBORGS
UNIVERSITET

Illumina's sequencers

MiSeq
Focused power. Speed and simplicity for targeted and small genome sequencing.

NextSeq 500
Flexible power. Speed and simplicity for everyday genomics.

HiSeq 2500
Production power. Power and efficiency for large-scale genomics.

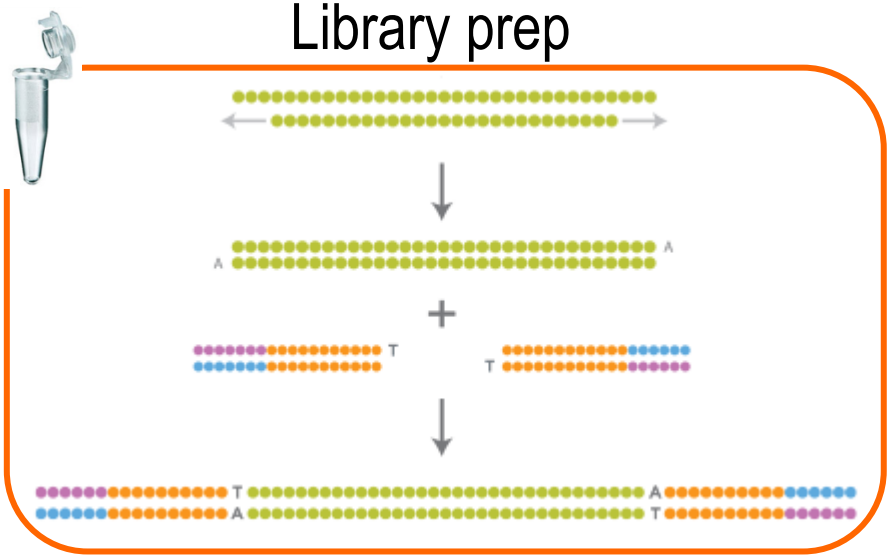
HiSeq X*
Population power. $1,000 human genome and extreme throughput for population-scale sequencing.

Key applications	Small genome, amplicon, and targeted gene panel sequencing.	Everyday genome, exome, transcriptome sequencing, and more.		Production-scale genome, exome, transcriptome sequencing, and more.		Population-scale human whole-genome sequencing.
Run mode	N/A	Mid-Output	High-Output	Rapid Run	High-Output	N/A
Flow cells processed per run	1	1	1	1 or 2	1 or 2	1 or 2
Output range	0.3-15 Gb	20-39 Gb	30-120 Gb	10-180 Gb	50-1000 Gb	1.6-1.8 Tb
Run time	5-65 hours	15-26 hours	12-30 hours	7-40 hours	< 1 day - 6 days	< 3 days
Reads per flow cell†	25 Million‡	130 Million	400 Million	300 Million	2 Billion	3 Billion
Maximum read length	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 125 bp	2 × 150 bp

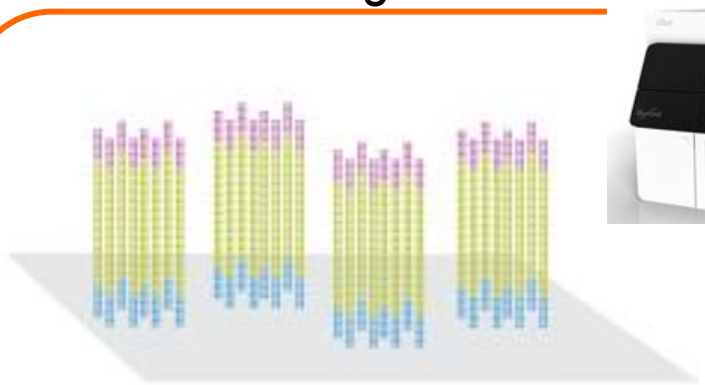
GenCore at SU: genomics@gu.se

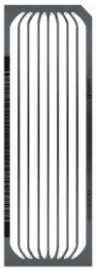
Illumina workflow

Library prep

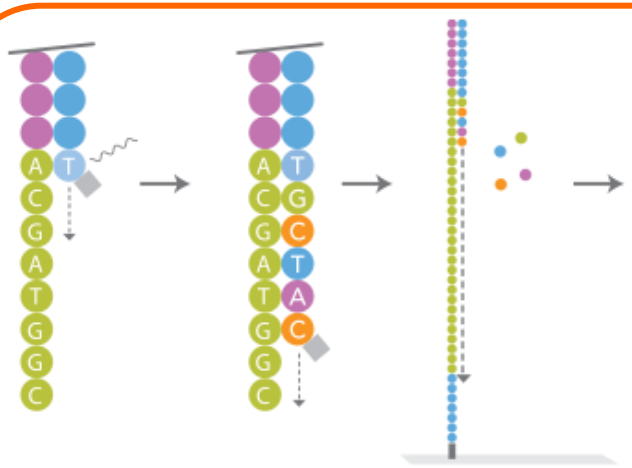


Cluster generation





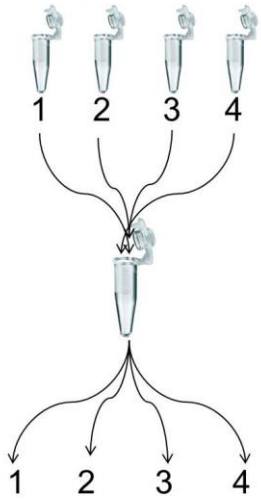
Sequencing, imaging and base calling



Fastq files

Sequencing run Quality

Demultiplexing



Total Reads	PF Reads	% Reads Identified (PF)	CV	Min	Max
116344024	100675880	96.5715	0.0514	22.6164	25.5666

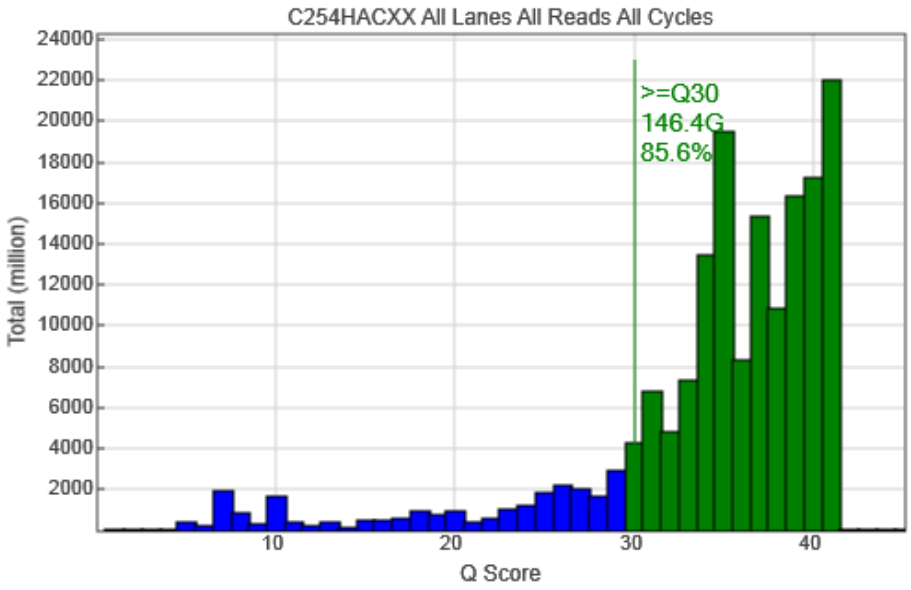
Index Number	Sample Id	Project	Index 1 (I7)	Index 2 (I5)	% Reads Identified (PF)
1	S1		CGATGT		23.8324
2	S2		TTAGGC		25.5666
3	S3		TGACCA		22.6164
4	S4		AAACAT		24.5561

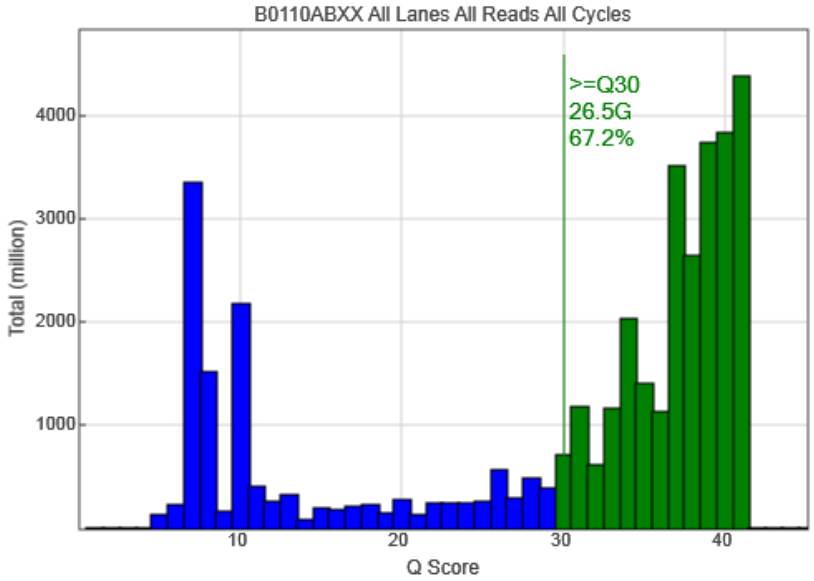
Total Reads	PF Reads	% Reads Identified (PF)	CV	Min	Max
29906232	28449264	98.0977	0.2024	11.7508	21.0338

Index Number	Sample Id	Project	Index 1 (I7)	Index 2 (I5)	% Reads Identified (PF)
1	S1		CGATGT		14.2264
2	S2		TGACCA		15.0889
3	S3		ACAGTG		7.75
4	S4		GCCAAT		18.2478
5	S5		CAGATC		11.7508
6	S6		CTTGTA		21.0338

Sequencing run Quality

QScore Distribution





A succesful run should have 80% >= Q30

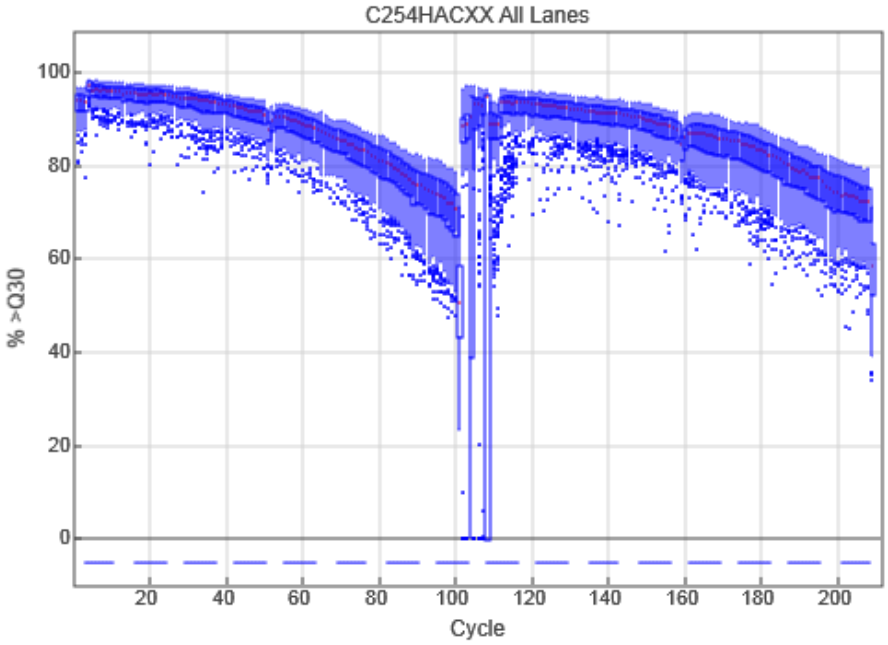
Illumina Sequencing Analysis Viewer

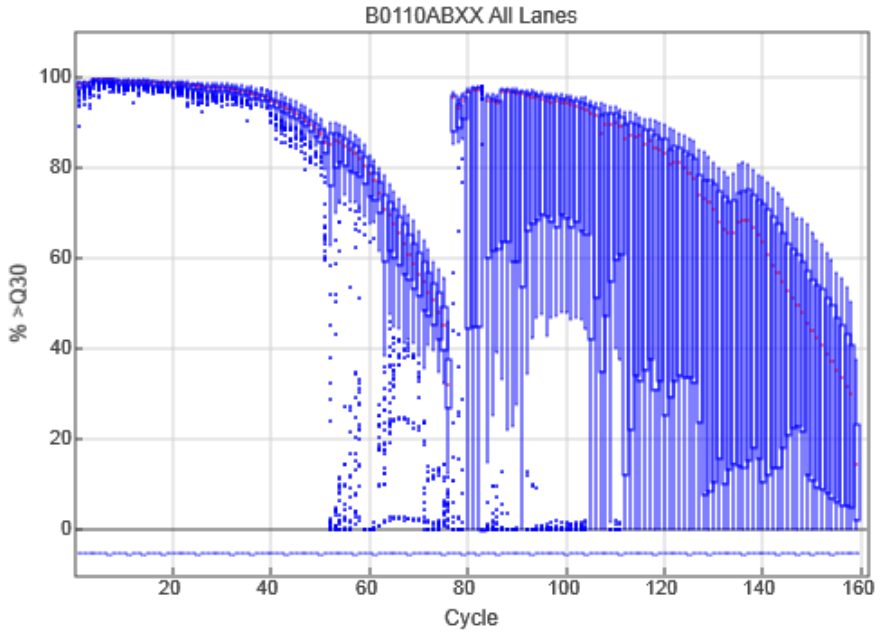
GÖTEBORGS UNIVERSITET

Bioinformatics
Core Facility

Sequencing run Quality

Data by Cycle





Illumina Sequencing Analysis Viewer

Fastq format

1) @SEQ_ID
   instrument:run:flowcell:lane:tile:x:y pair:fail:control:index
2) sequence
3) marker
4) quality

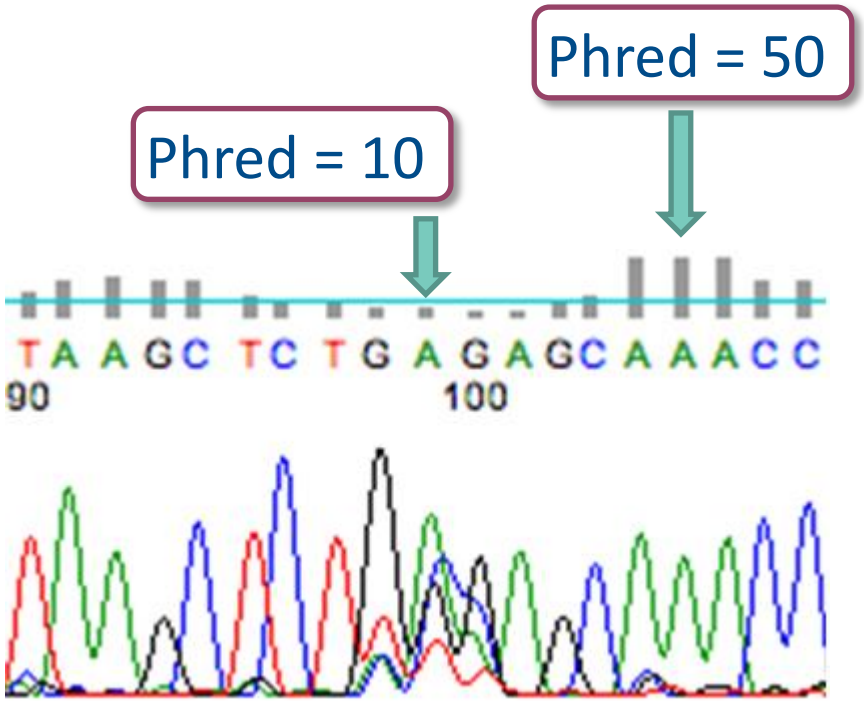
```
1)   @HWI-H200:53:D08U2ACXX:5:1101:1231:2012 1:N:0:
2)   GCATTTTAGTAGAACCAGNCATTTCCCCCNACNTCNNTNCGNNANNNNTAA
3)   +
4)   @CCFFFFFHFFHHJJJJJ#3<FGIJJJJJ#1?#####################
```

31 37 39 18 16 2

Phred score

Probability that the base has been erroneously called

Phred score	P(called wrong)	Accuracy base call
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%




GÖTEBORGS
UNIVERSITET

GÖTEBORGS UNIVERSITET

Bioinformatics
Core Facility

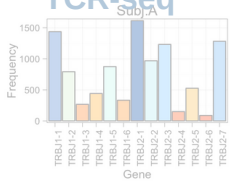
Data handling workflow

Quality Check

Quality Filter

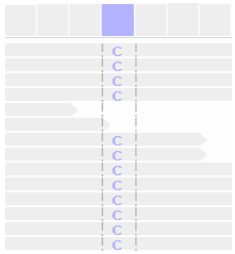
Mapping to ref genome
De novo assembly

TCR-seq



TCR profiling

Exome-seq

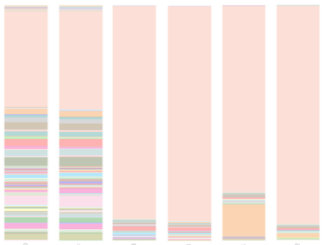


SNV detection

Methyl-seq

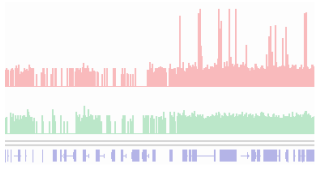
CpG patterns

Metagenomics



Taxa summary

ChIP-Seq



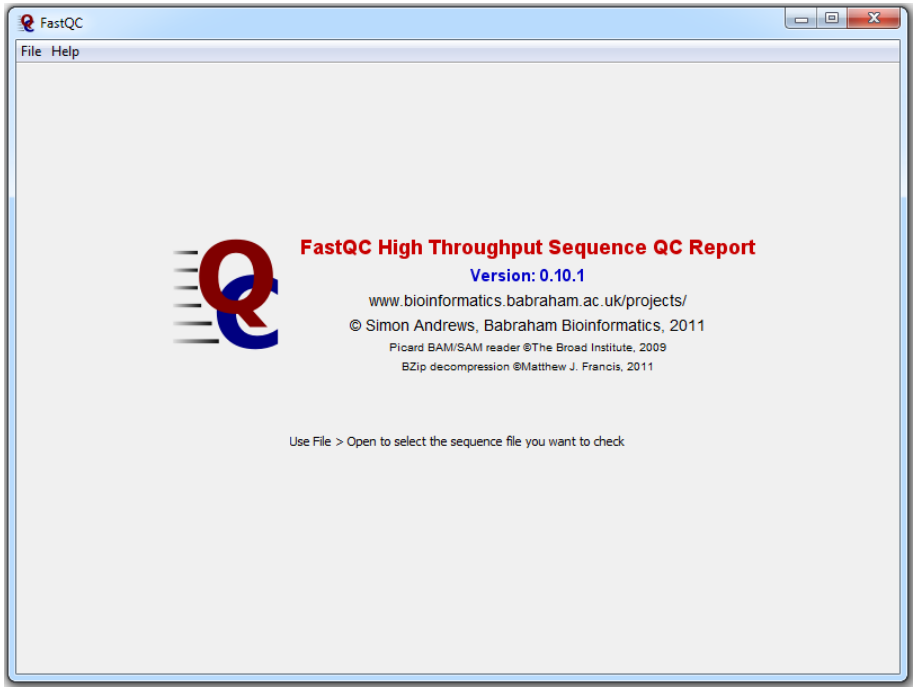
Peak detection

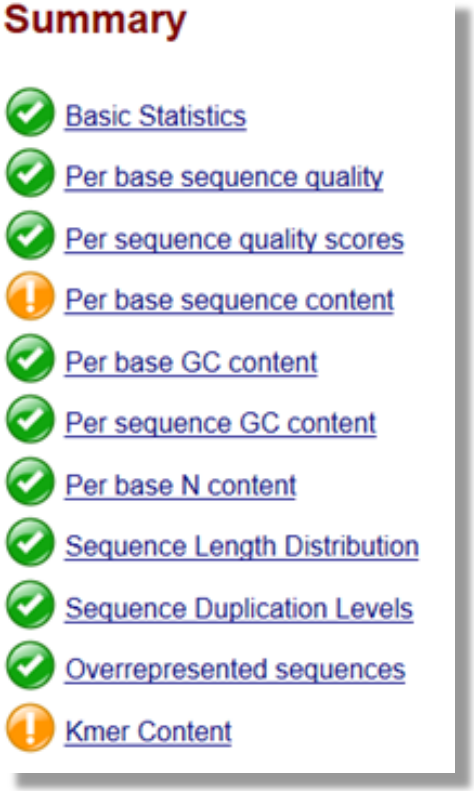
RNA-seq

Transcript estimation

Quality check

Set of quality checks to produce a report which allows you to quickly assess the overall quality of your run



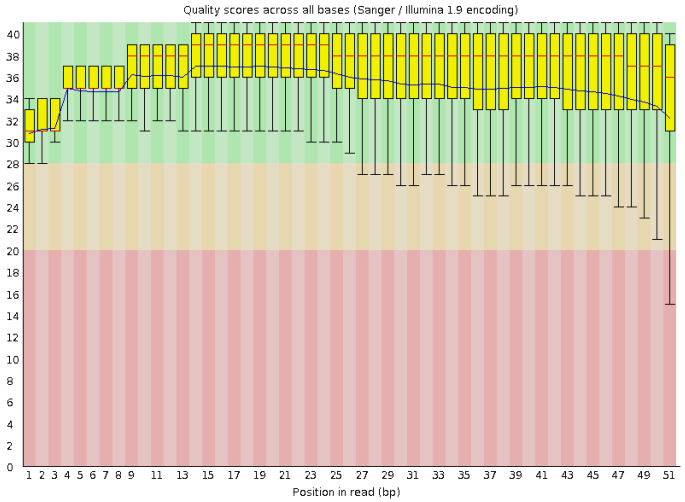


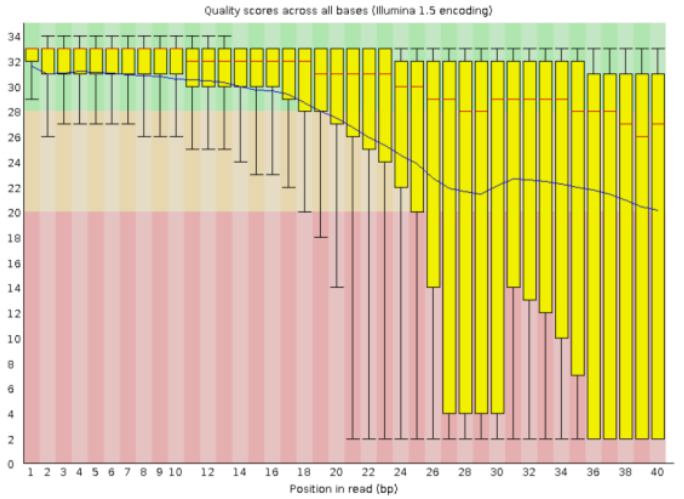
FastQC, PrinSeq

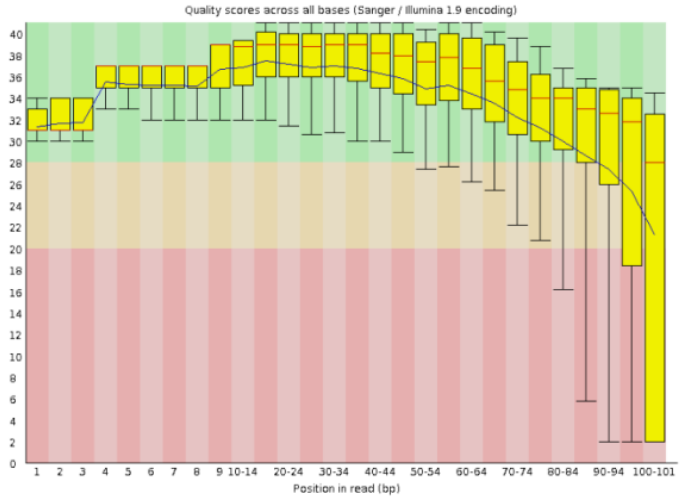
GÖTEBORGS UNIVERSITET

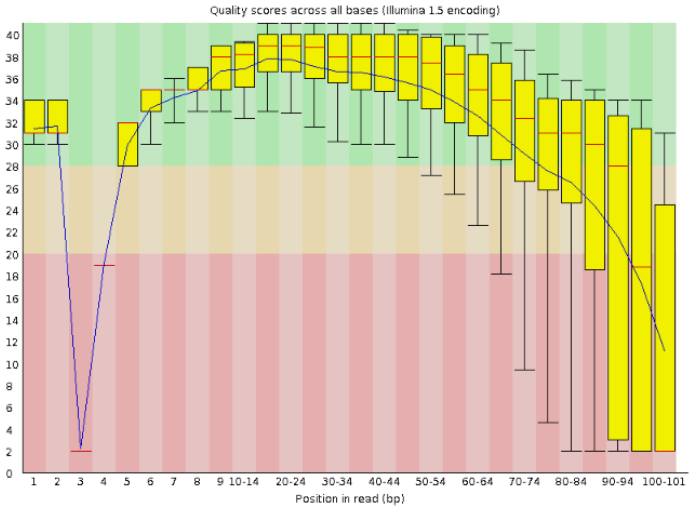
Bioinformatics
Core Facility

Per base sequence quality

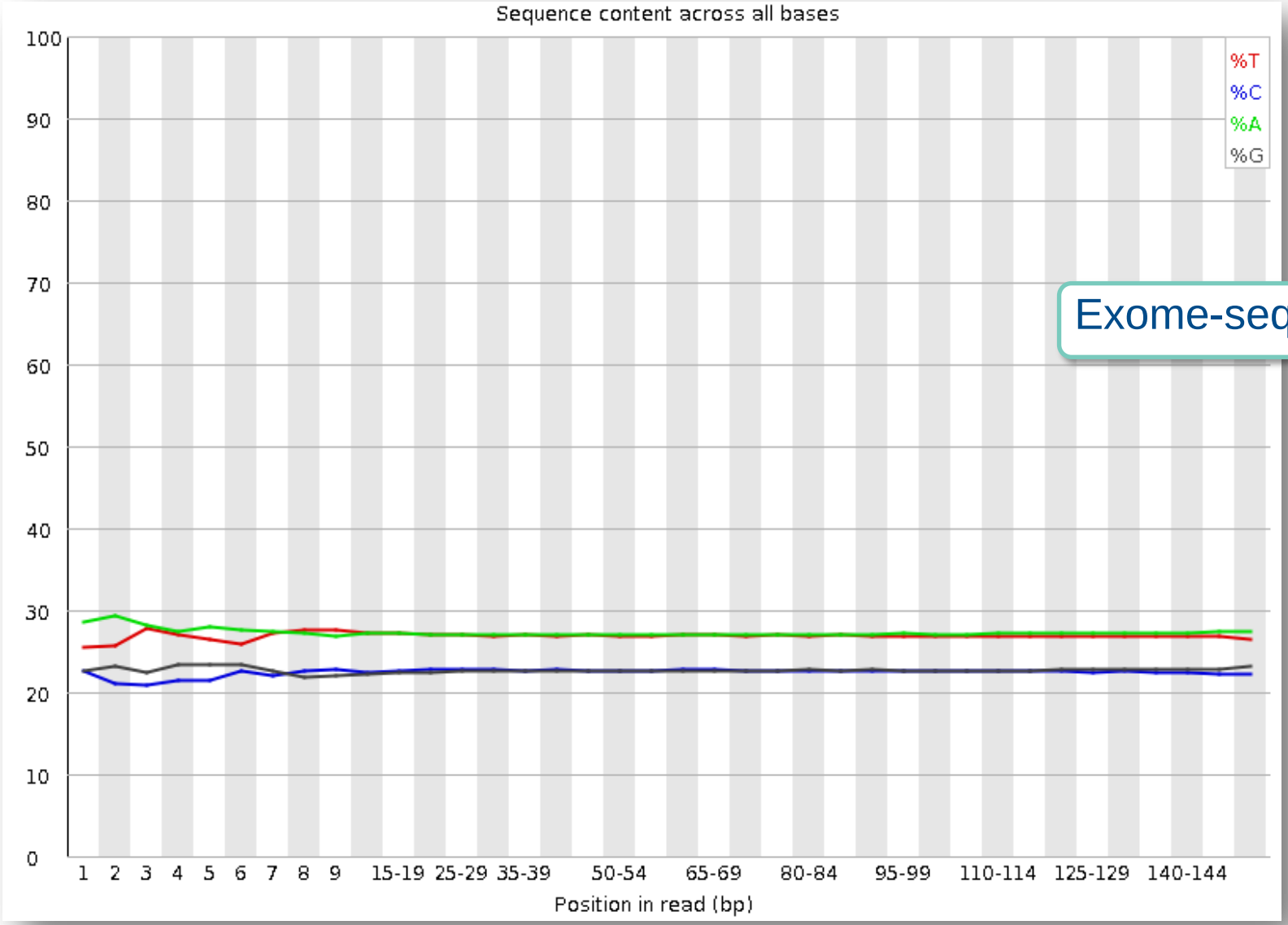








Per base sequence content



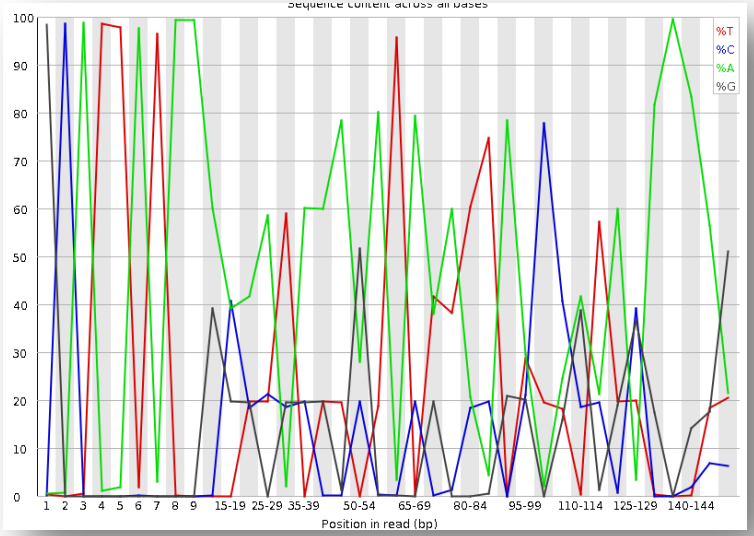
Exome-seq

GÖTEBORGS UNIVERSITET

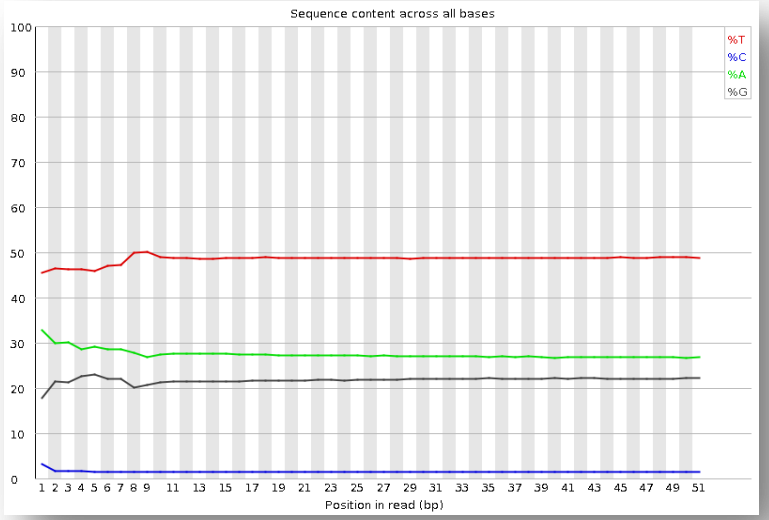
Bioinformatics
Core Facility

Per base sequence content

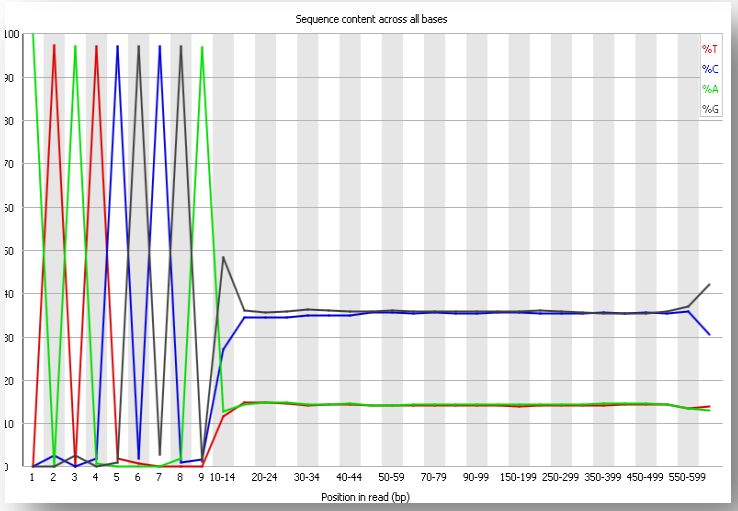
A



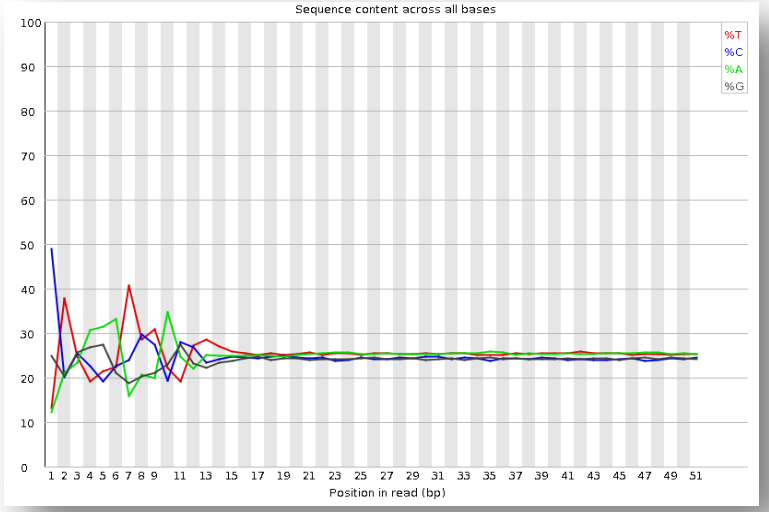
B



C



D

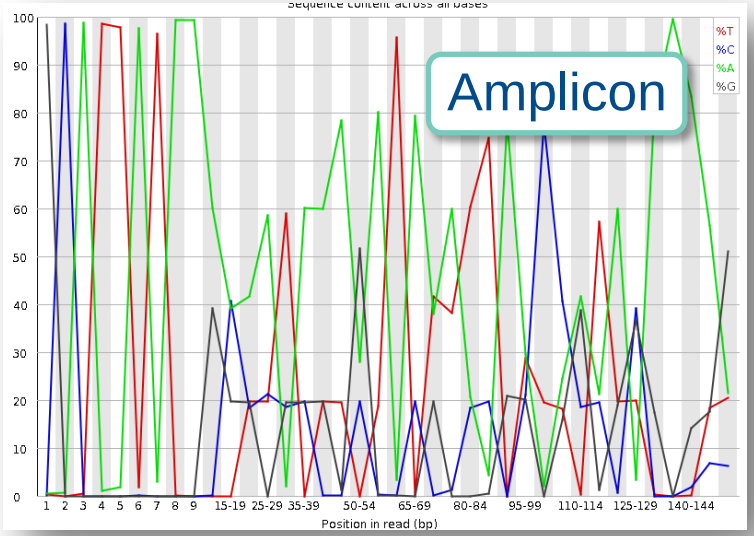


GÖTEBORGS UNIVERSITET

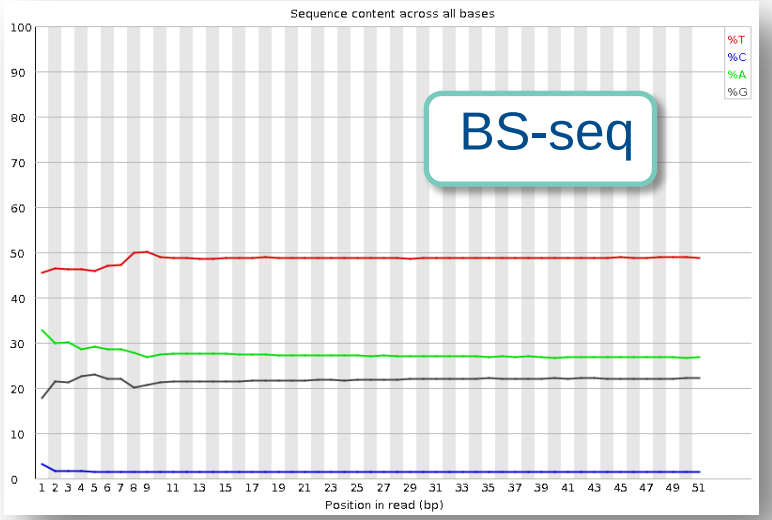
Bioinformatics
Core Facility

Per base sequence content

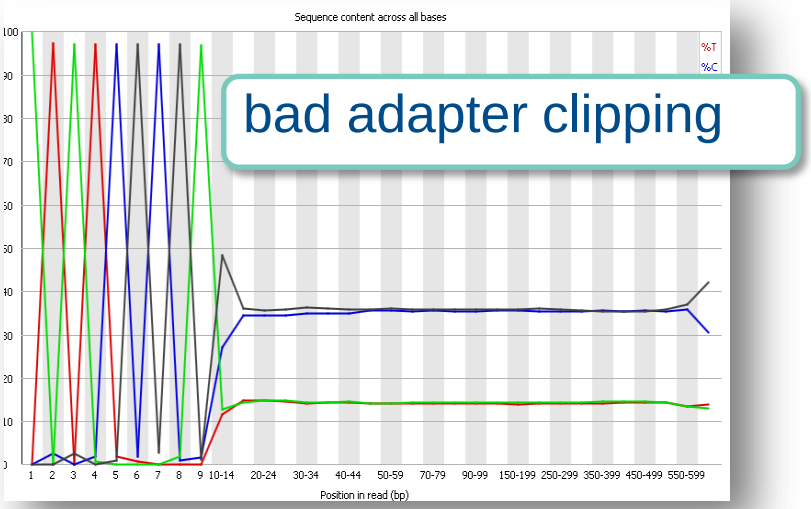
A



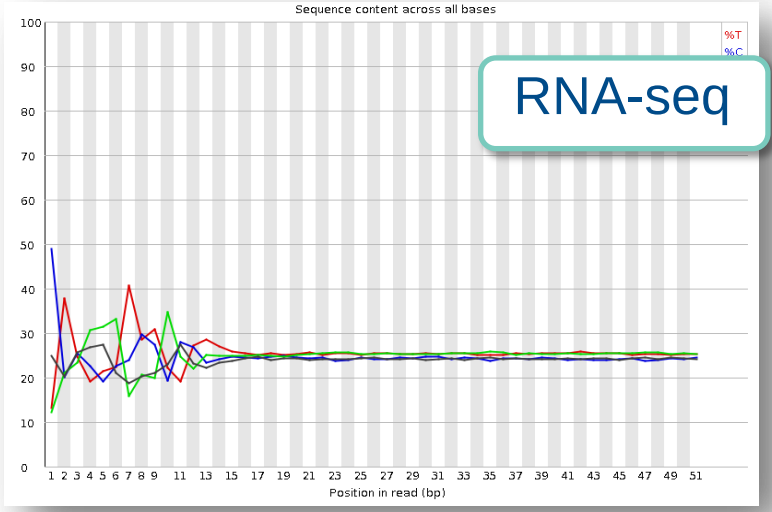
B



C



D



GÖTEBORGS UNIVERSITET

Bioinformatics
Core Facility

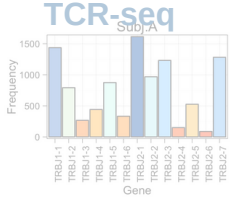
Data handling workflow

Quality Check

Quality Filter

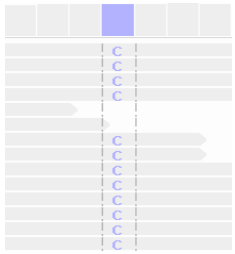
Mapping to ref genome
De novo assembly

TCR-seq



TCR profiling

Exome-seq

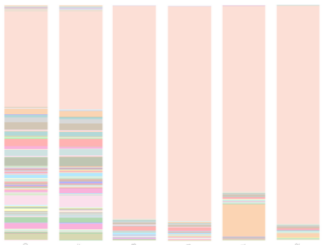


SNV detection

Methyl-seq

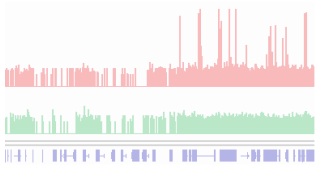
CpG patterns

Metagenomics



Taxa summary

ChIP-Seq



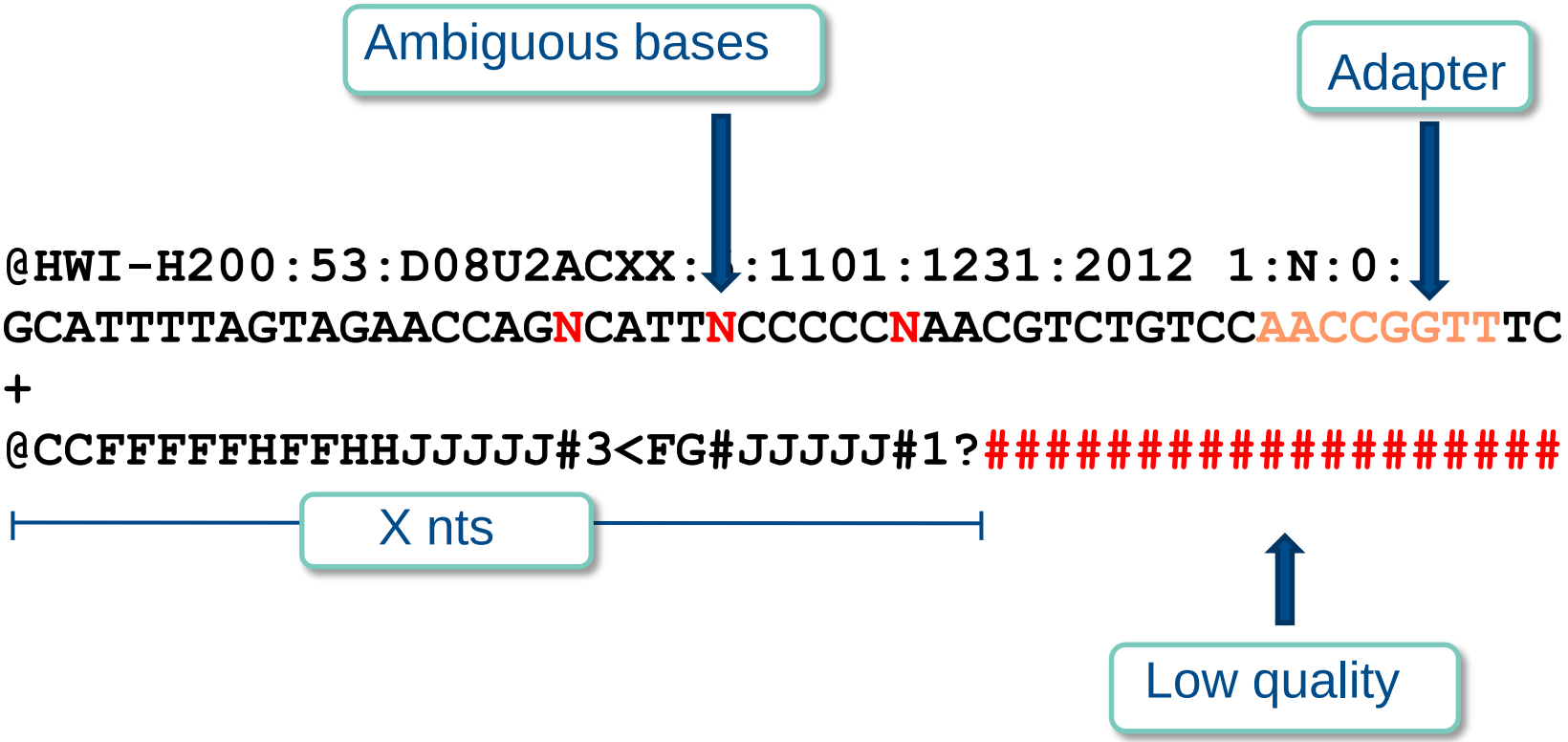
Peak detection

RNA-seq

Transcript estimation

Quality filter

A collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.



FastX, PRINSeq, Cutadapt, Trim Galore!

GÖTEBORGS UNIVERSITET

Bioinformatics
Core Facility

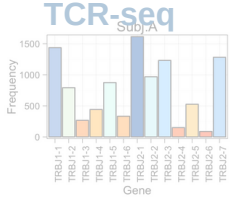
Data handling workflow

Quality Check

Quality Filter

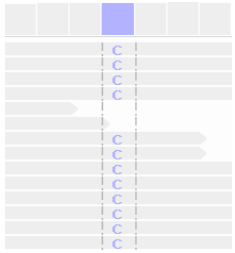
Mapping to ref genome
De novo assembly

TCR-seq



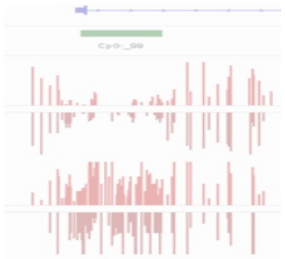
TCR profiling

Exome-seq



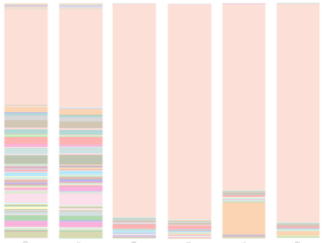
SNV detection

Methyl-seq



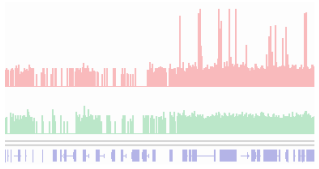
CpG patterns

Metagenomics



Taxa summary

ChIP-Seq

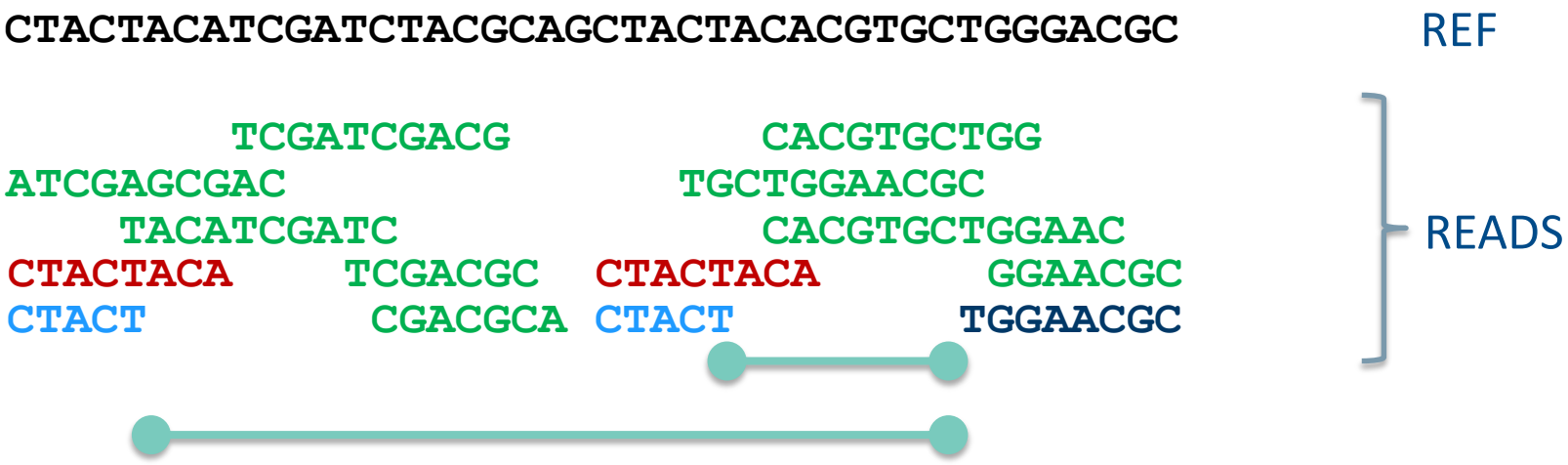


Peak detection

RNA-seq

Transcript estimation

Mapping



WHERE to place the reads?

a) Unique reads
b) Everywhere possible
c) Choose one randomly
d) Use pair-end data

fragment size: 40

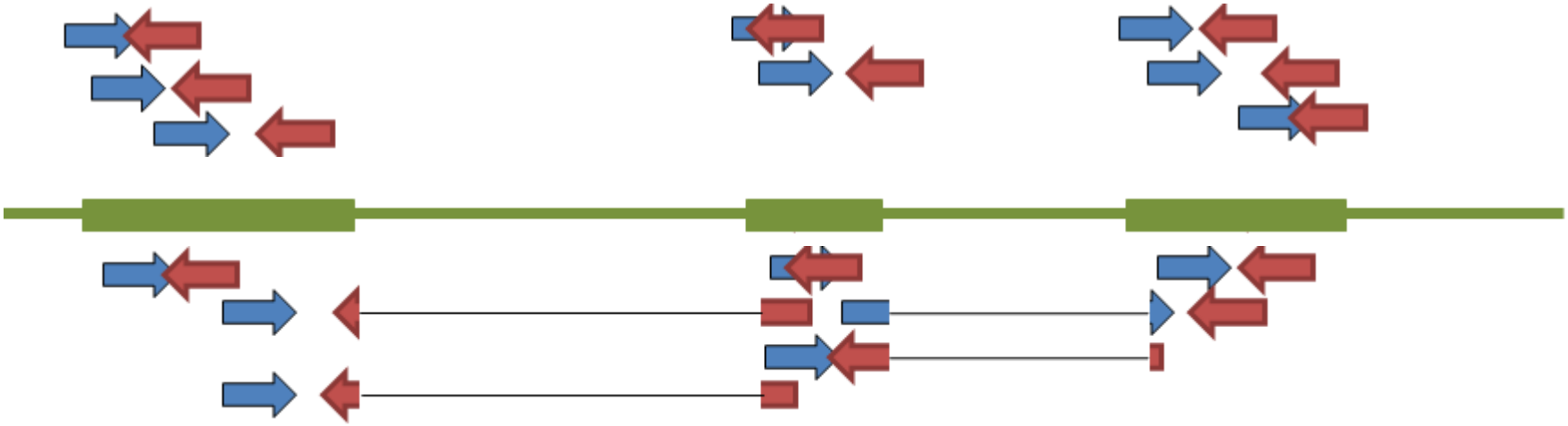
Mapping

```
CTACTACATCGATCTACGCAGCTACTACACGTGCTGGGACGC                    REF

          TCGATCGACG              CACGTGCTGG
ATCGAGCGAC                  TGCTGGAACGC
     TACATCGATC                   CACGTGCTGGAAC
CTACTACA      TCGACGC   CTACTACA         GGAACGC
CTACT           CGACGCA CTACT              TGGAACGC
```

READS

HOW to place the reads? Ungapped, Gapped



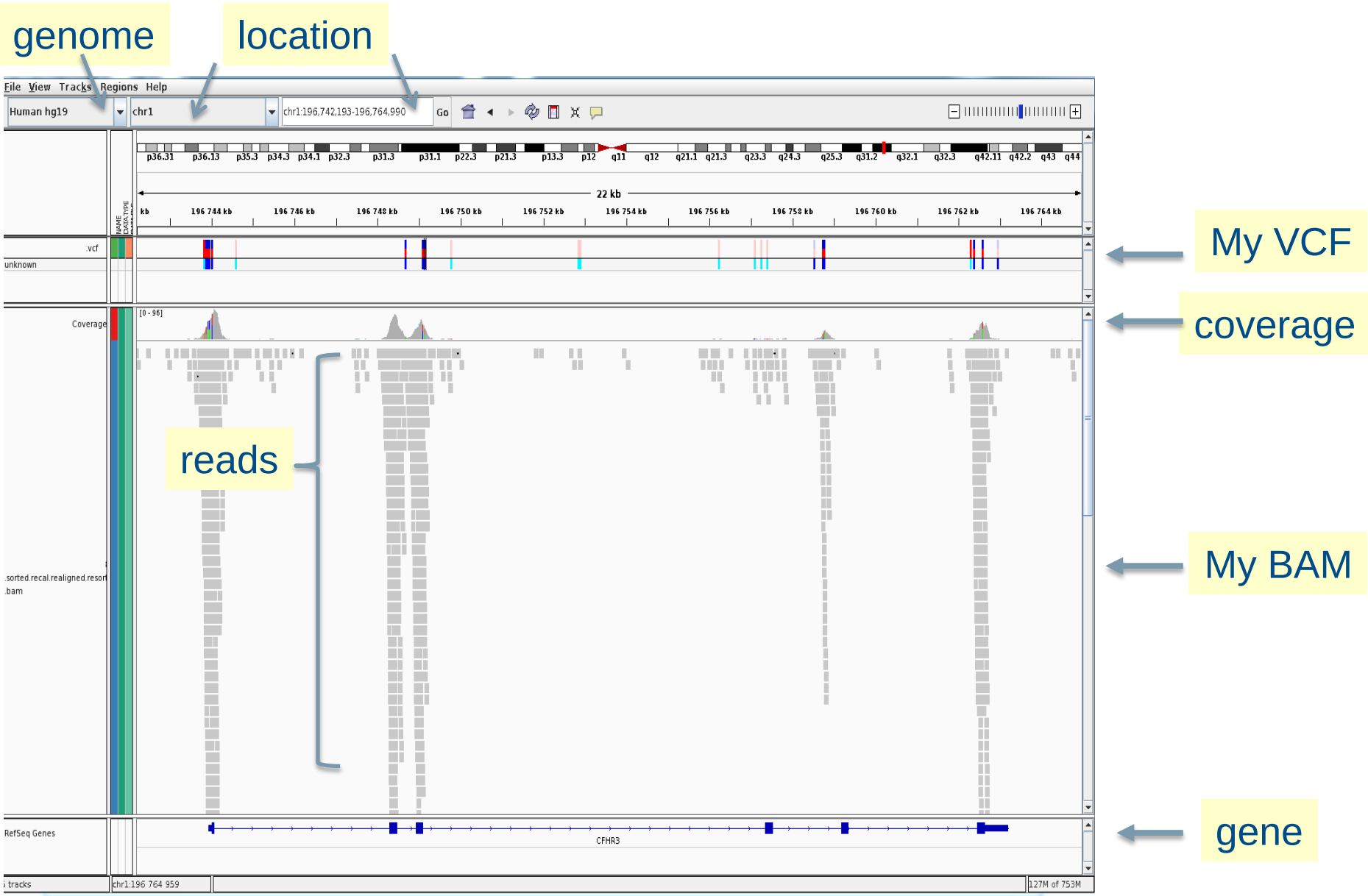
SAM/BAM format

SAM (Sequence Alignment/Map)
BAM (Binary Alignment/Map)

```
HWI-H200:53:D08U2ACXX:6:1108:18555:16623        99      chr1    10001   60      45M6S   =       10174   224
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT
AACCCTAAAGATCG  @?@DDDBDAH??FHDGFFFHIIIGDGEHHI<ABHICHIEHCDD3BDEDGEC     MD:Z:45 RG:Z:1  XG:i:0  AM:i:0  NM:i:0
SM:i:0  XM:i:0  XO:i:0  XT:A:M
```

Query name	HWI-H200:53:D08U2ACXX:6:1108:18555:16623
Flag	99
Reference name	chr1
Leftmost position	10001
Mapping quality	60
CIGAR string	45M6S
Mate reference	=
Mate position	10174
Insert size	224
Query sequence	TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT..
Quality	@?@DDDBDAH??FHDGFFFHIIIGDGEHHI<ABHICH..
Optional fields	MD:Z:45 RG:Z:1 XG:i:0 AM:i:0 NM:i:0 SM:i:0 XM:i:0 XO:i:0 XT:A:M

IGV – integrative genome viewer

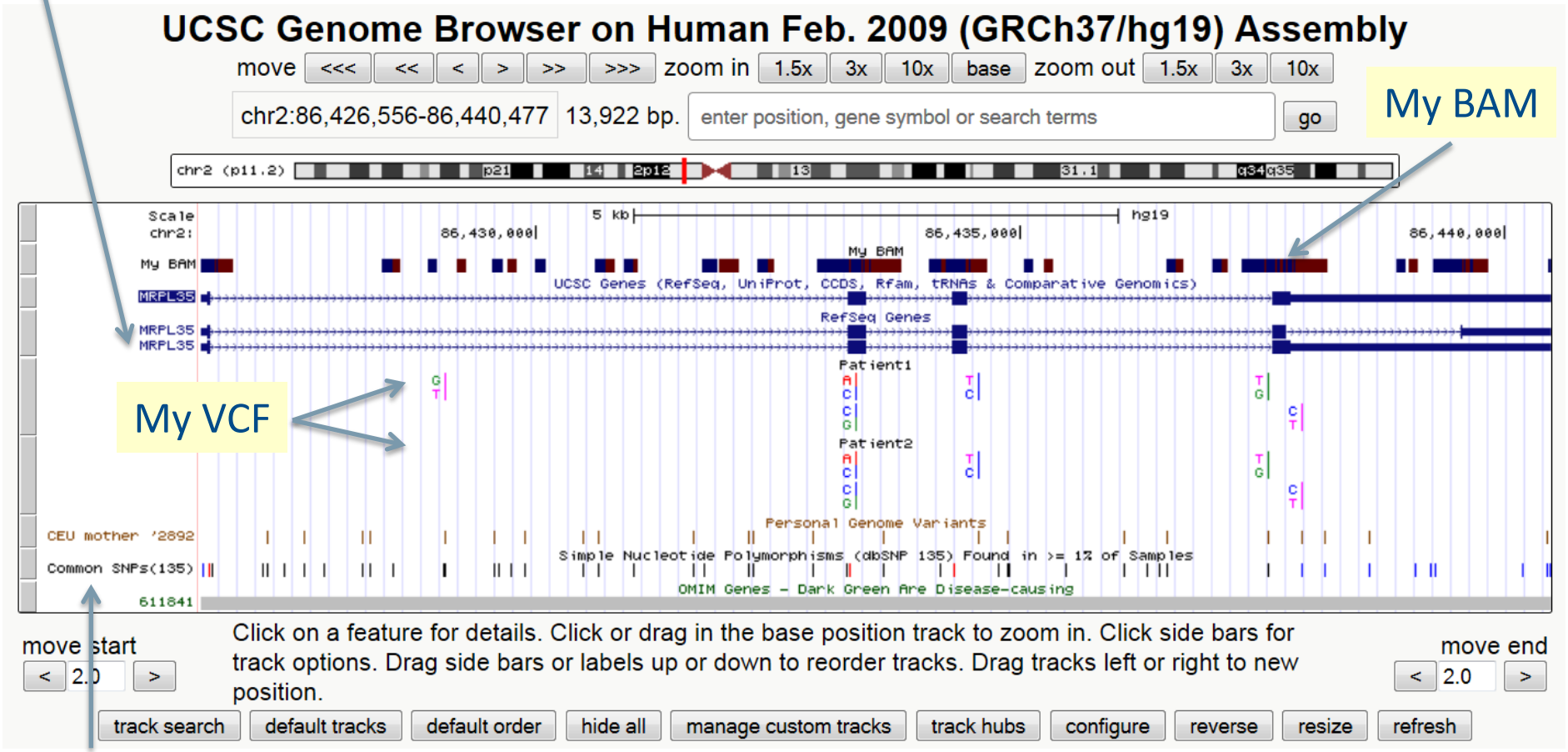


GÖTEBORGS UNIVERSITET

Bioinformatics
Core Facility

UCSC browser

gene variants



Variation track

GÖTEBORGS UNIVERSITET

Bioinformatics
Core Facility

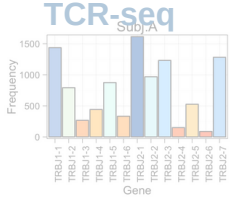
Data handling workflow

Quality Check

Quality Filter

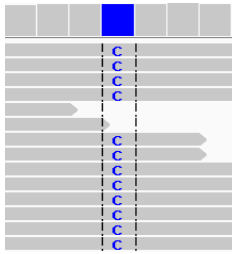
Mapping to ref genome
De novo assembly

TCR-seq



TCR profiling

Exome-seq

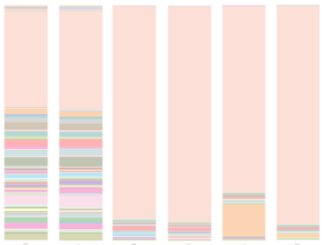


SNV detection

Methyl-seq

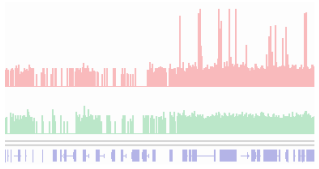
CpG patterns

Metagenomics



Taxa summary

ChIP-Seq



Peak detection

RNA-seq

Transcript estimation

Single Nucleotide Polymorphism

3 million differences
(0.1%)

SNP	AAGC-TA AAGC**T**TA	AAGC**C**TA AAGC**T**TA	AAGC**C**TA AAGC-TA

- Disease risk
- Drug efficacy
- Heritable phenotypes

rs4988235 lactose intolerance
rs1333049 coronary heart dis
rs9939609 triggers obesity

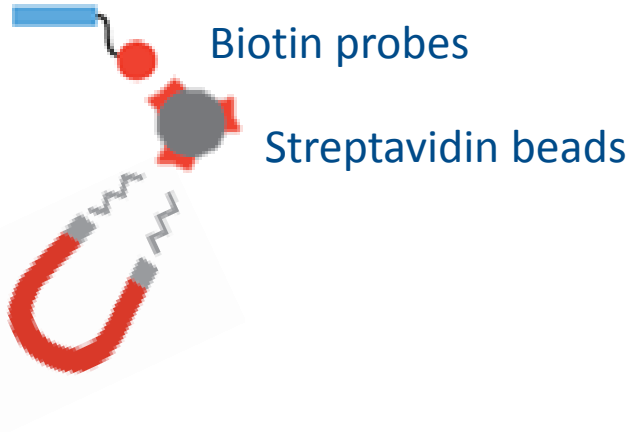
Monogenic diseases

Modifications of a single gene over 10,000 of human diseases (½ have a gene associated)

DISEASE	GENE	MUTATION
Thalassaemia	HBB	Δ → frameshift
Sickle cell anemia	HBB	G6V
Cystic Fibrosis	CFTR	G542X …
Fragile X syndrome	FMR1	CGG expansion
Huntington's	HTT	CAG +36 repeats
Tay sachs	HEXA	65 single base mutations 14 splice site lesions 10 deletions 2 insertions

Exome sequencing (targeted)

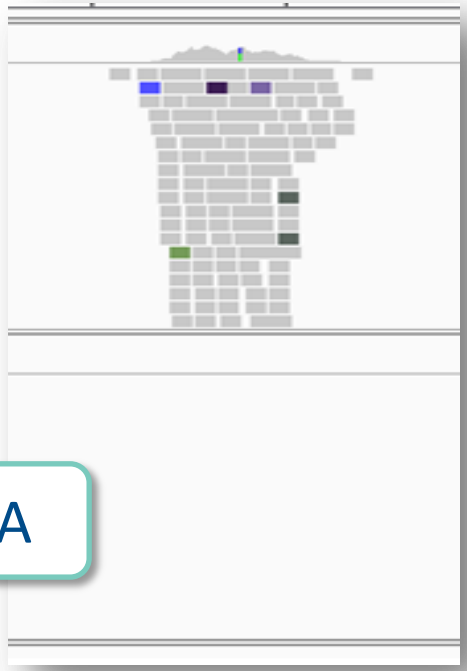
85% of the disease-causing mutations are located in protein coding regions



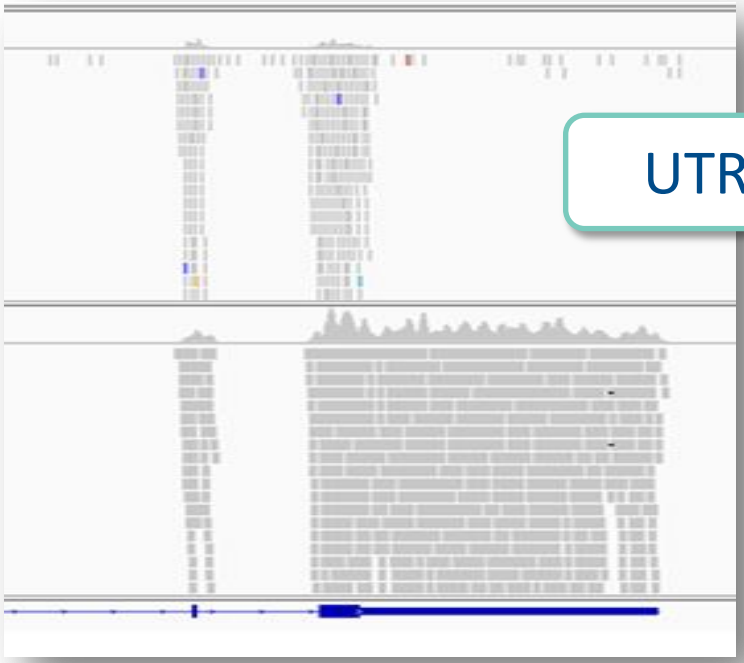
Genome (3 GB) vs exome (50 MB)

Enrichment kits

	NimbleGen v3	Agilent	TruSeq
Total	64,190,759	51,542,882	61,884,224
RefSeq (coding)	33,491,892	32,326,914	31,817,166
RefSeq (UTR)	NA	3,920,825	31,642,004
Ensembl (CDS)	31,690,383	33,472,589	31,918,846
Ensembl (all exons)	33,731,215	38,123,201	59,275,652
miRBase	59,996	55,249	27,963



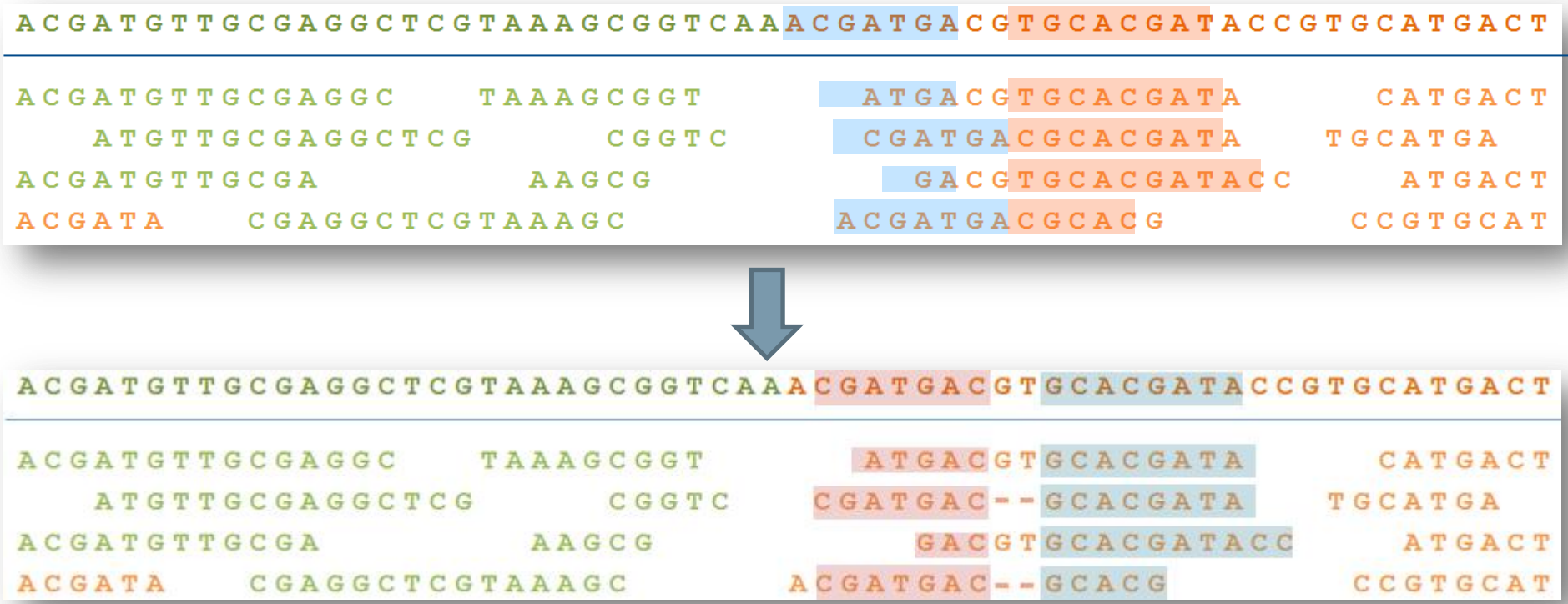
miRNA



UTR’s

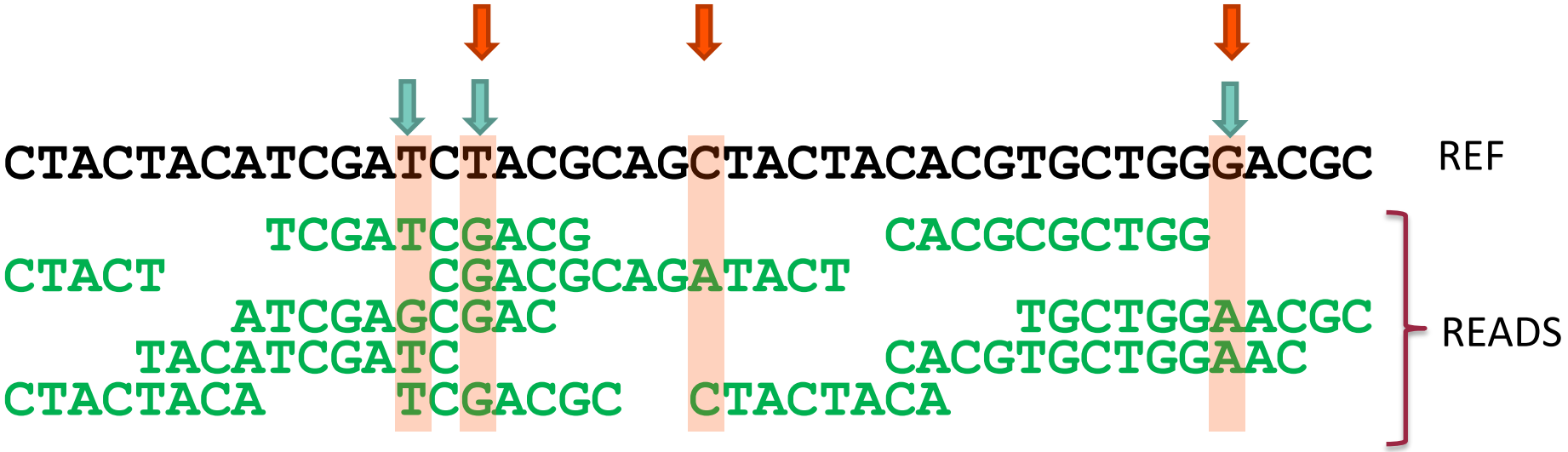
Realignment and recalibration

Correct alignments due to the presence of indels
Differenciate between polymorphisms and sequencing errors



GATK

Variant calling



Is it a variant allele?

What is the most likely genotype?

SOAP2, samtools, GATK, Beagle, CRISP, Dindel, FreeBayes, SeqEM, VarScan, Mutect

VCF format

Variant call format http://www.1000genomes.org/node/101

HEADER

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

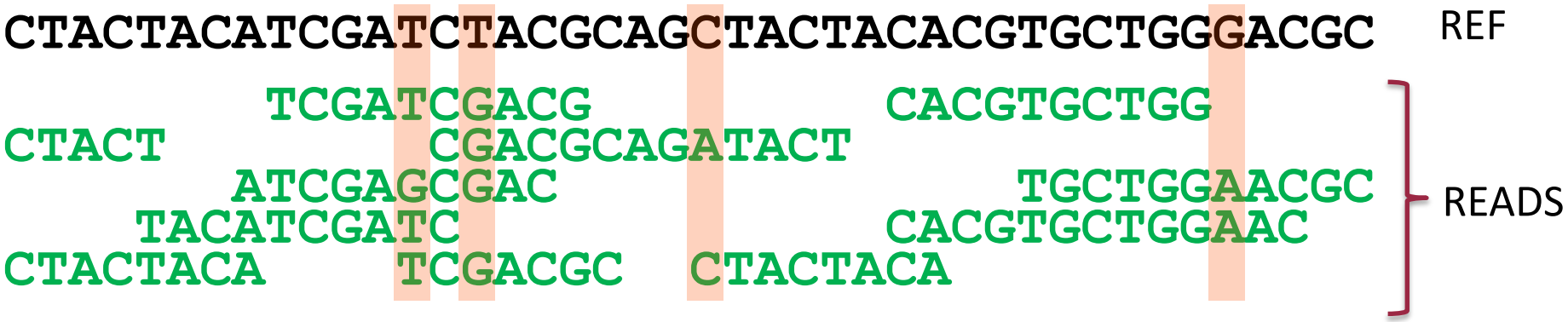
BODY

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTCT	G,GTACT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

GÖTEBORGS UNIVERSITET

Bioinformatics
Core Facility

Variant annotation



In which gene is it located?
Name, Description,
OMIM, Pathway, GO,
Expression profiles . . .

Where in the gene is it located?
Intron, exon, UTR,
intergenic region, splice site

Is there any AA change?

```
GAA -> GAG = E->E
GTT -> CTT = V->L
TGG -> TGA = W->X
TGA -> CGA = X->R
```

Is it a known SNP?

What impact does the AA change have?
Damaging, benign

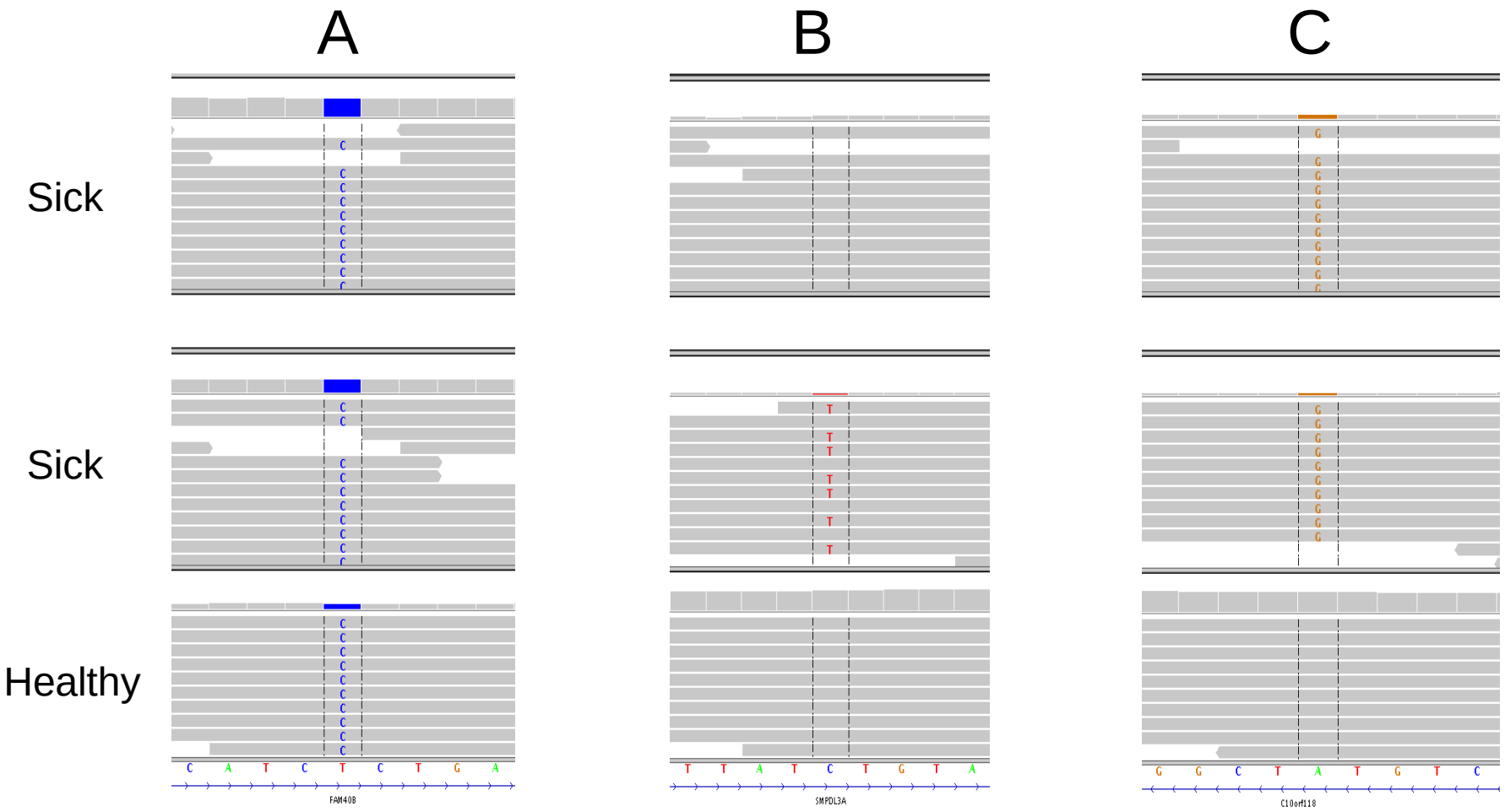
Annovar,
SIFT, PP2,
dbSNP,
GO,
KEGG,
OMIM
1000G

Variant list

CHR	POS	REF	OBS	ALLELE	GENE	DESCRIPTION	VARIANT_FUNCTION	EXONIC_FUNCTION
chr1	780785	T	A	homozygous	LOC643837	-	ncRNA_intronic	-
chr1	802496	C	T	heterozygous	FAM41C	-	downstream	-
chr1	887801	A	G	homozygous	NOC2L	Nucleolar complex protein 2 homolog	exonic	Synonymous
chr1	1265154	T	C	homozygous	GLTPD1	Glycolipid transfer protein domain-containing protein 1	downstream	-
chr1	151733327	T	C	heterozygous	MRPL9	39S ribosomal protein L9, mitochondrial	ncRNA_exonic	nonsynonymous
chr1	151733335	T	G	homozygous	MRPL9	39S ribosomal protein L9, mitochondrial	ncRNA_exonic	nonsynonymous
chr1	52306064	TCT	-	heterozygous	NRD1	Nardilysin	ncRNA_exonic	frameshift deletion
chr1	54605319	G	GC	homozygous	CDCP2	CUB domain-containing protein 2	exonic	frameshift substitution
chr3	189507518	C	CAGA	homozygous	TP63	Tumor protein 63	UTR5	-

AA_CHANGE_POS	AA_CHANGE	dbSNP	BUILD	SIFT	PP2	LRT	OMIM	CONSERVED
-		rs2977612	101					
-		rs10157494	119					conserved
NOC2L:uc001abz.3:exon10:c.T1182C:p.T394T	T => T	rs3828047	107					
-		rs307355	79					conserved
MRPL9:uc001eyv.2:exon6:c.A637G:p.I213V,	I => V	rs74228558	130	tolerated	bening	deleterius	611824	conserved
MRPL9:uc001eyv.2:exon6:c.A629C:p.E210A	K => Q	rs8480	52	damaging	damaging	neutral	611824	
NRD1:uc010ong.1:exon2:c.208_0del:p.70_0del,		rs145326984	134					
CDCP2:uc001cwv.1:exon4:c.1224_1224delinsGC,		rs66537746	130					
-		rs34201045	126					conserved

Variant filtering

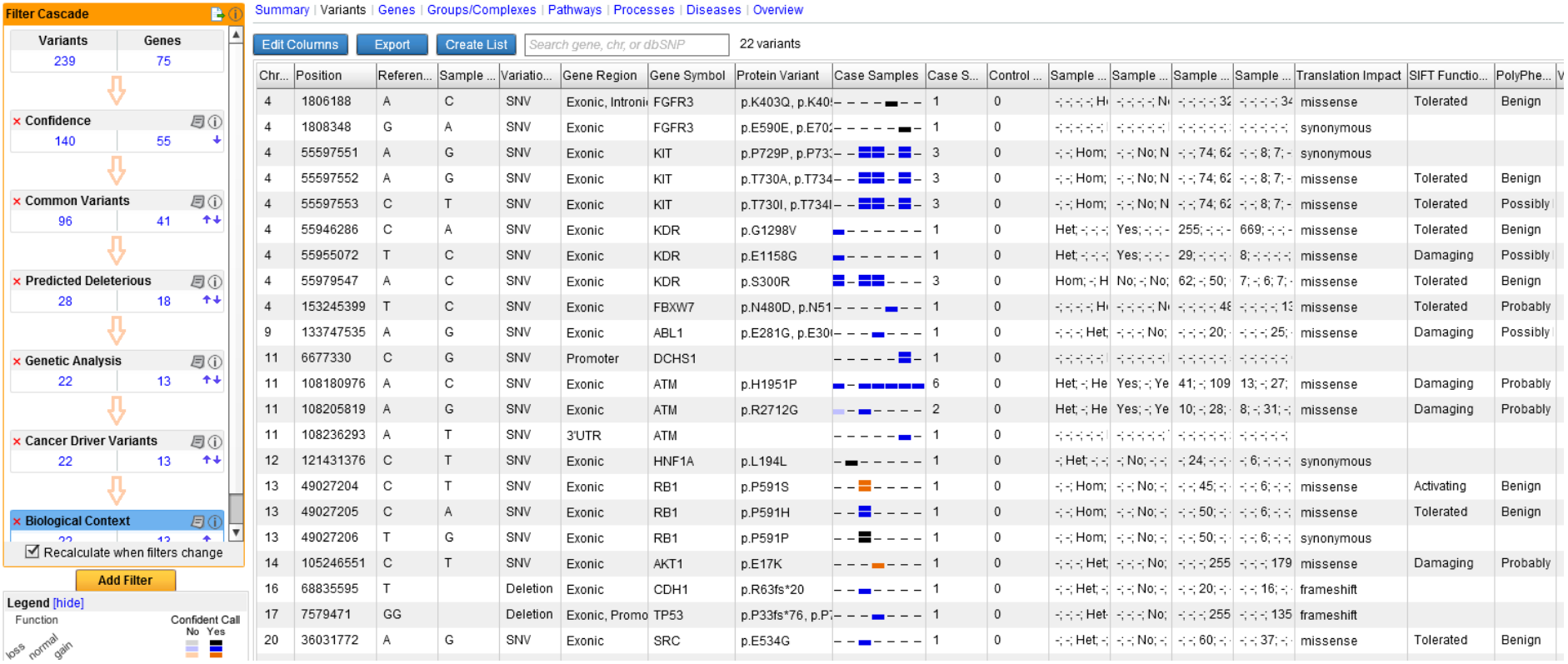


GÖTEBORGS UNIVERSITET

Variant Analysis

Bioinformatics
Core Facility

Ingenuity Variant Analysis (IVA)



Making sense of the data

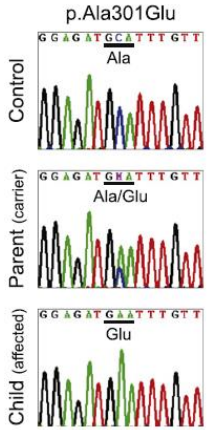
Exome sequencing cases → Coding variants → Disease model / Disease knowledge → **Candidate genes**

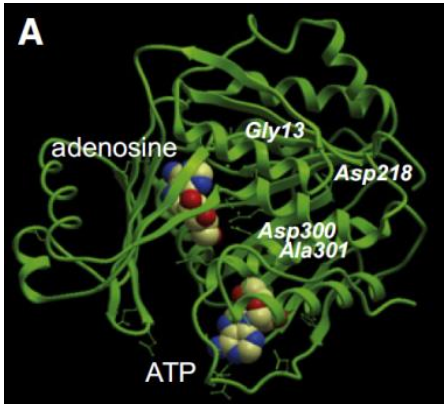
Family filters

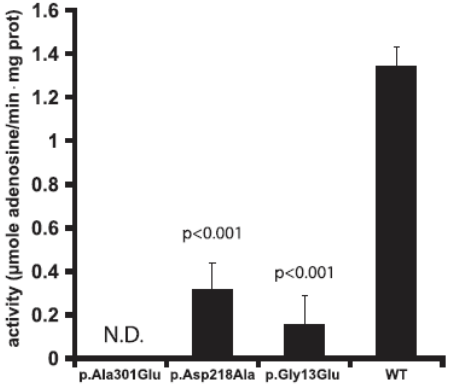
Controls

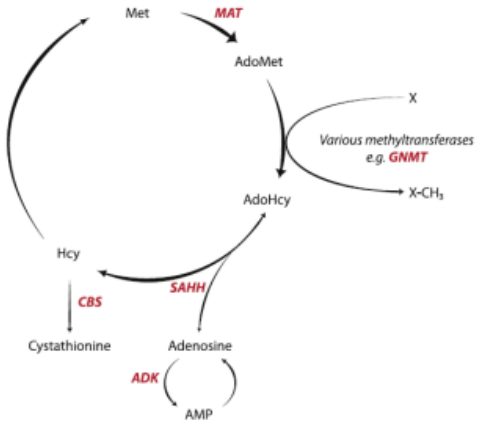
Genetic variation DBs

Your real work begins…










GÖTEBORGS
UNIVERSITET