IDENTIFYING A DISEASE CAUSING MUTATION

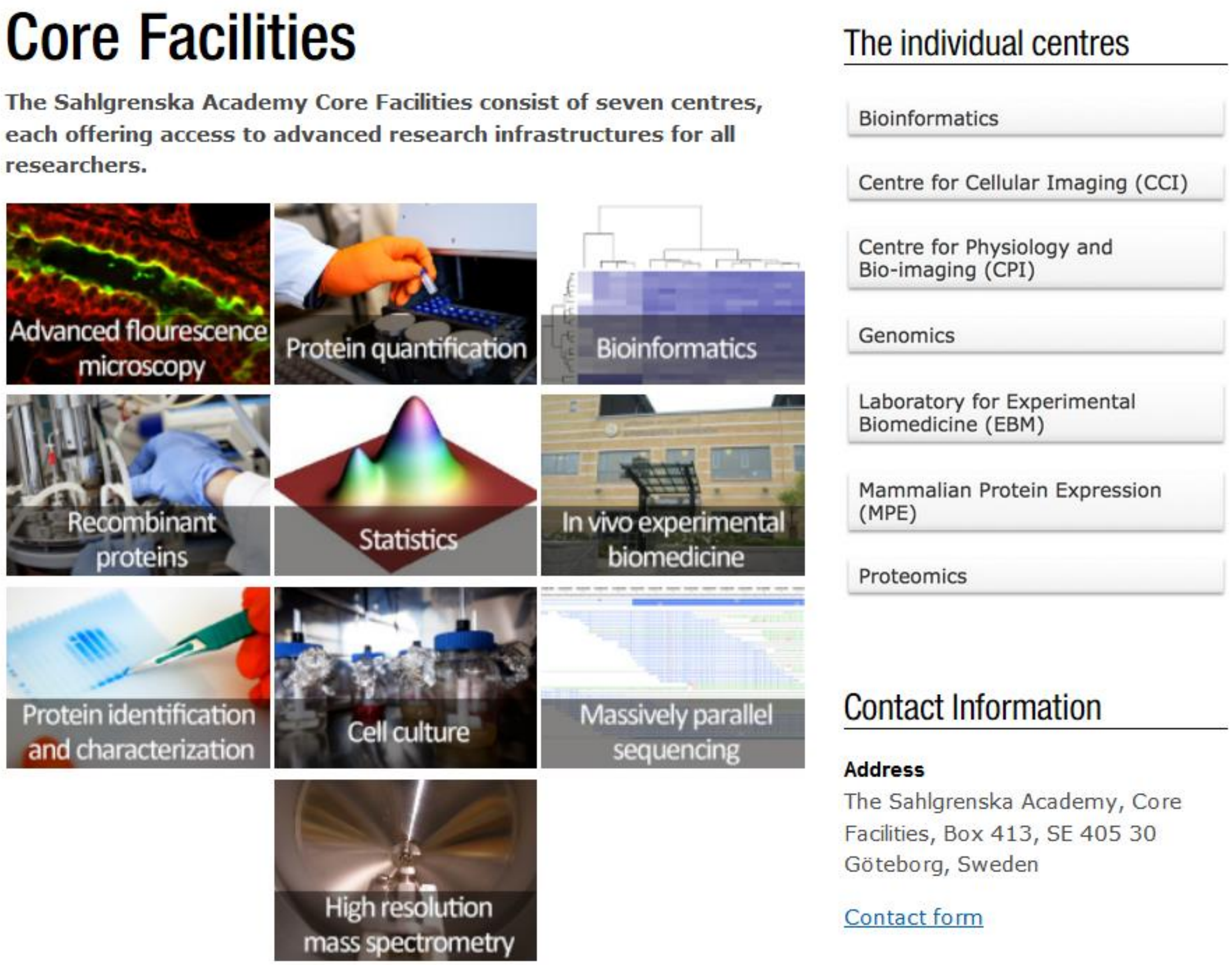
Targeted resequencing

MARCELA DAVILA

3/MZO/2016

Bioinformatics
Core Facility

Core Facilities at Sahlgrenska Academy


Advanced flourescence microscopy
Protein quantification
Bioinformatics
Recombinant proteins
Statistics
In vivo experimental biomedicine
Protein identification and characterization
Cell culture
Massively parallel sequencing
High resolution mass spectrometry

Core Facilities

The Sahlgrenska Academy Core Facilities consist of seven centres, each offering access to advanced research infrastructures for all researchers.

The individual centres

- Bioinformatics
- Centre for Cellular Imaging (CCI)
- Centre for Physiology and Bio-imaging (CPI)
- Genomics
- Laboratory for Experimental Biomedicine (EBM)
- Mammalian Protein Expression (MPE)
- Proteomics

Contact Information

Address

The Sahlgrenska Academy, Core
Facilities, Box 413, SE 405 30
Göteborg, Sweden

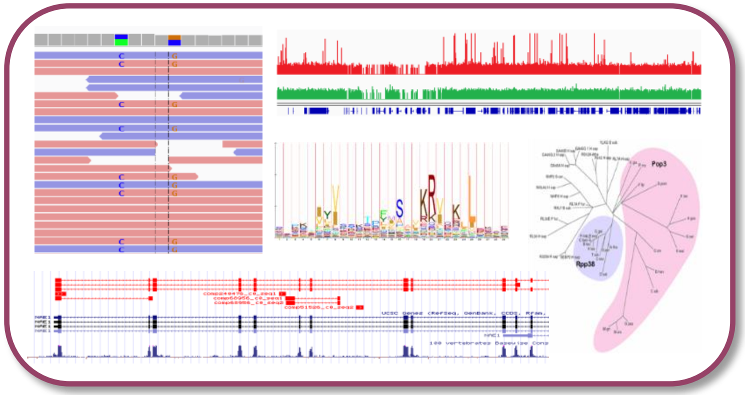
Contact form

GÖTEBORGS
UNIVERSITET

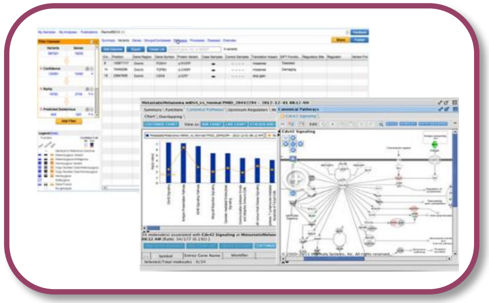
Bioinformatics
Core Facility

Bioinformatics Core Facility

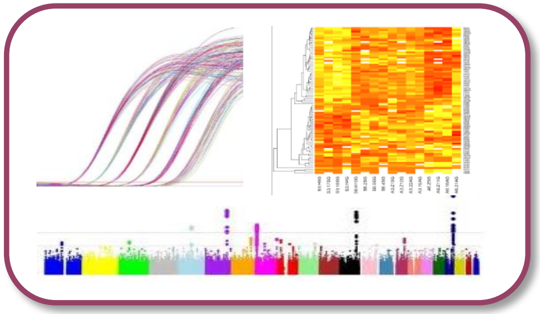
Bioinformatics



Software



Statistics

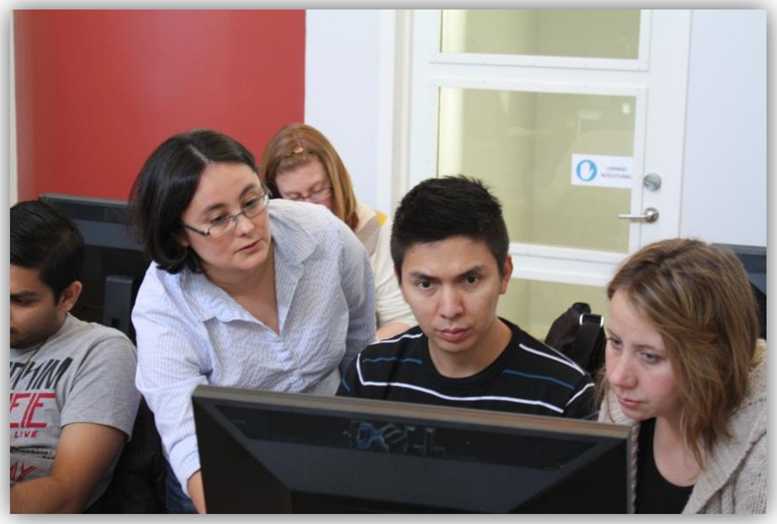


bioinformatics@gu.se

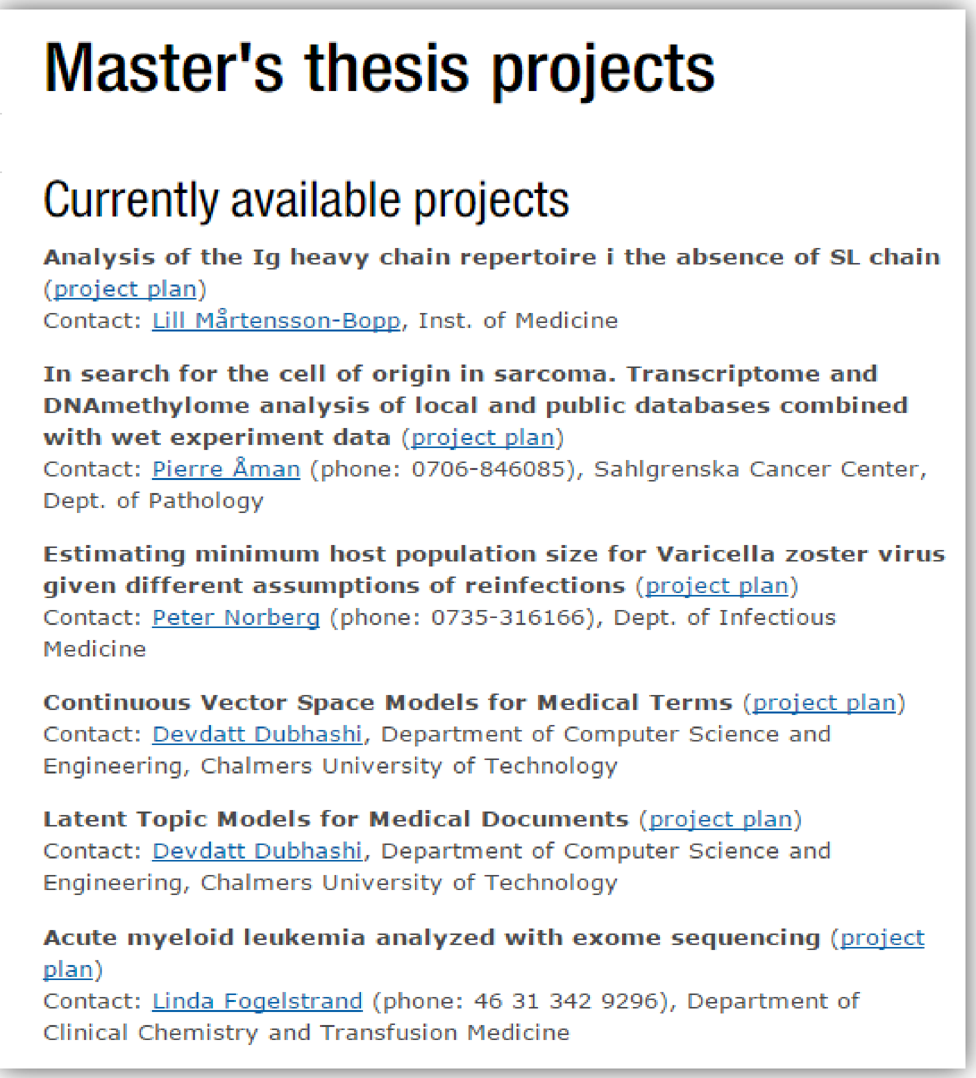
www.cf.gu.se/english/Bioinformatics/

Increasing statistical and bioinformatics knowledge

- Personalized training (software/programming)
- Courses
  - ❖ Genomics and Bioinformatics
  - ❖ Advanced NGS data analysis
  - ❖ Perl for life science researchers
  - ❖ Unix with applications to NGS data
- Seminars and workshops



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

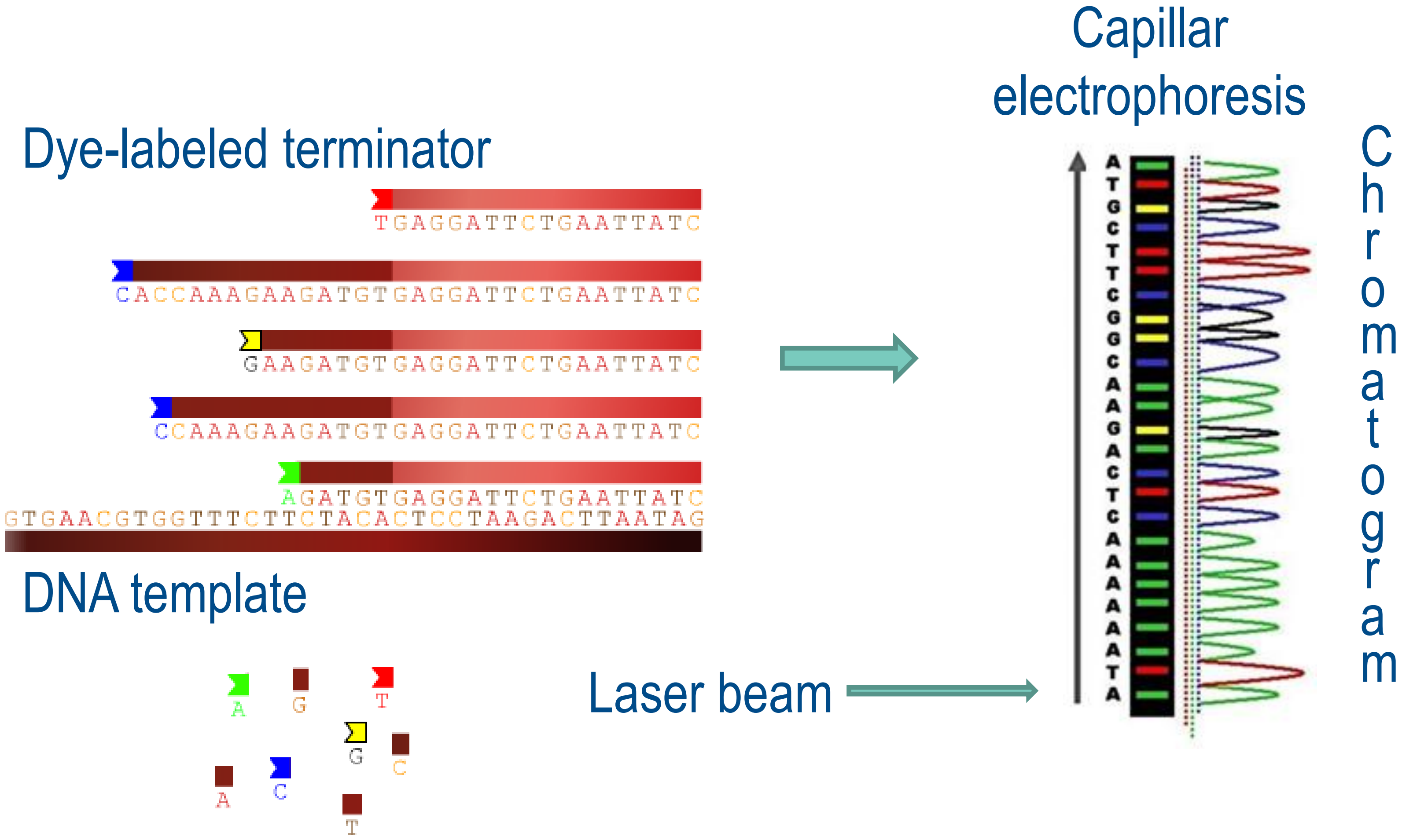
Continuous Vector Space Models for Medical Terms (project plan)
Contact: Devdatt Dubhashi, Department of Computer Science and Engineering, Chalmers University of Technology

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

Sanger sequencing



Dye-labeled terminator

DNA template

Capillar electrophoresis

Laser beam

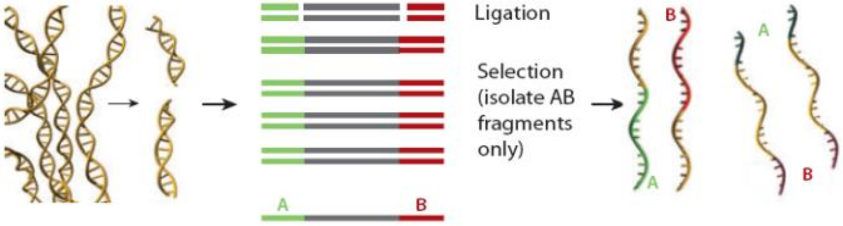
Chromatogram

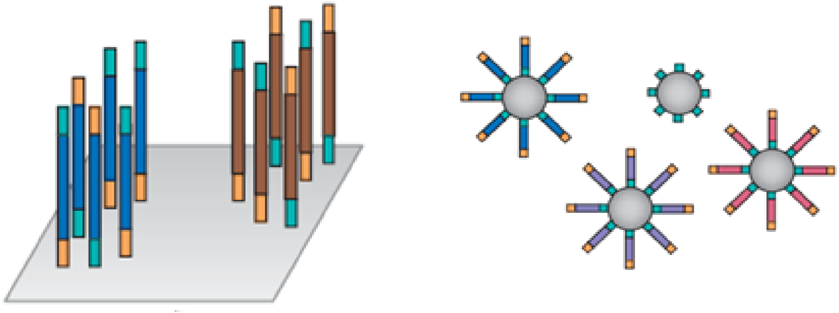
TGAGGATTCTGAATTATC
CACCAAAGAAGATGTGAGGATTCTGAATTATC
GAAGATGTGAGGATTCTGAATTATC
CCAAAGAAGATGTGAGGATTCTGAATTATC
AGATGTGAGGATTCTGAATTATC
GTGAACGTGGTTTCTTCTACACTCCTAAGACTTAATAG
A
G
T
G
C
A
C
T
A T G C T T C G G C A A G A C T C A A A A A A T A

Next Generation Sequencing

Roche 454
Solexa Illumina
SOLiD Life Technologies
Ion Torrent Life Technologies
Qiagen

- ❖ DNA library preparation
- ❖ Amplification (ePCR, bridge PCR)
- ❖ Sequencing reaction
- ❖ Imaging
- ❖ Decoding



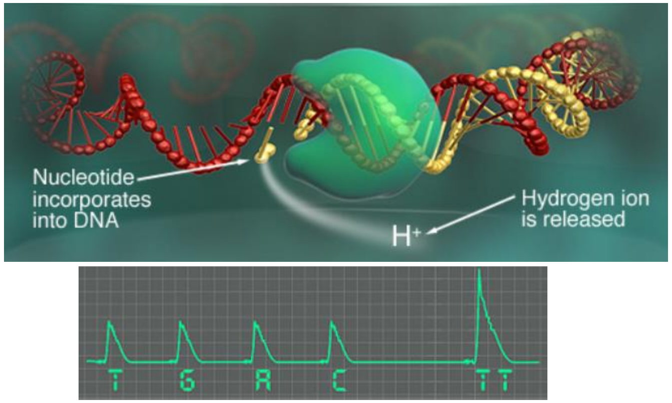


Third Generation Sequencing

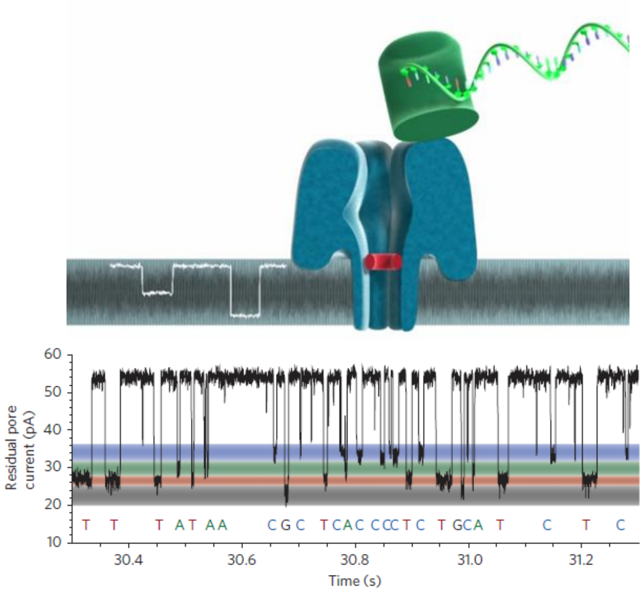
Single molecule- real time

No optics

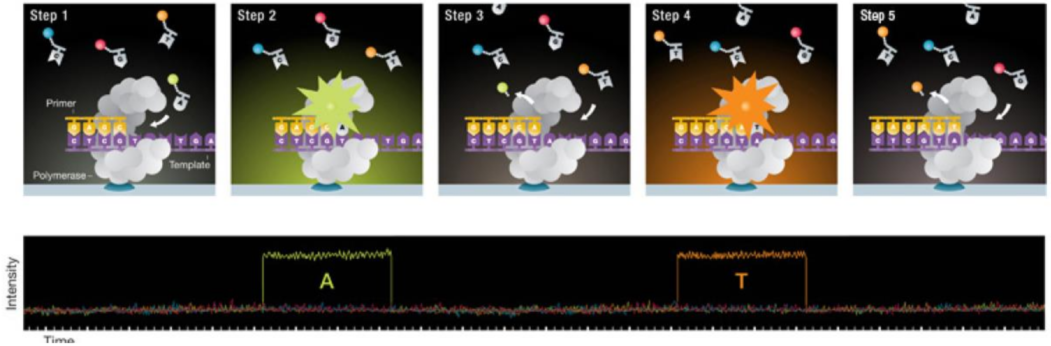
Increased sequencing speed



Ion Torrent

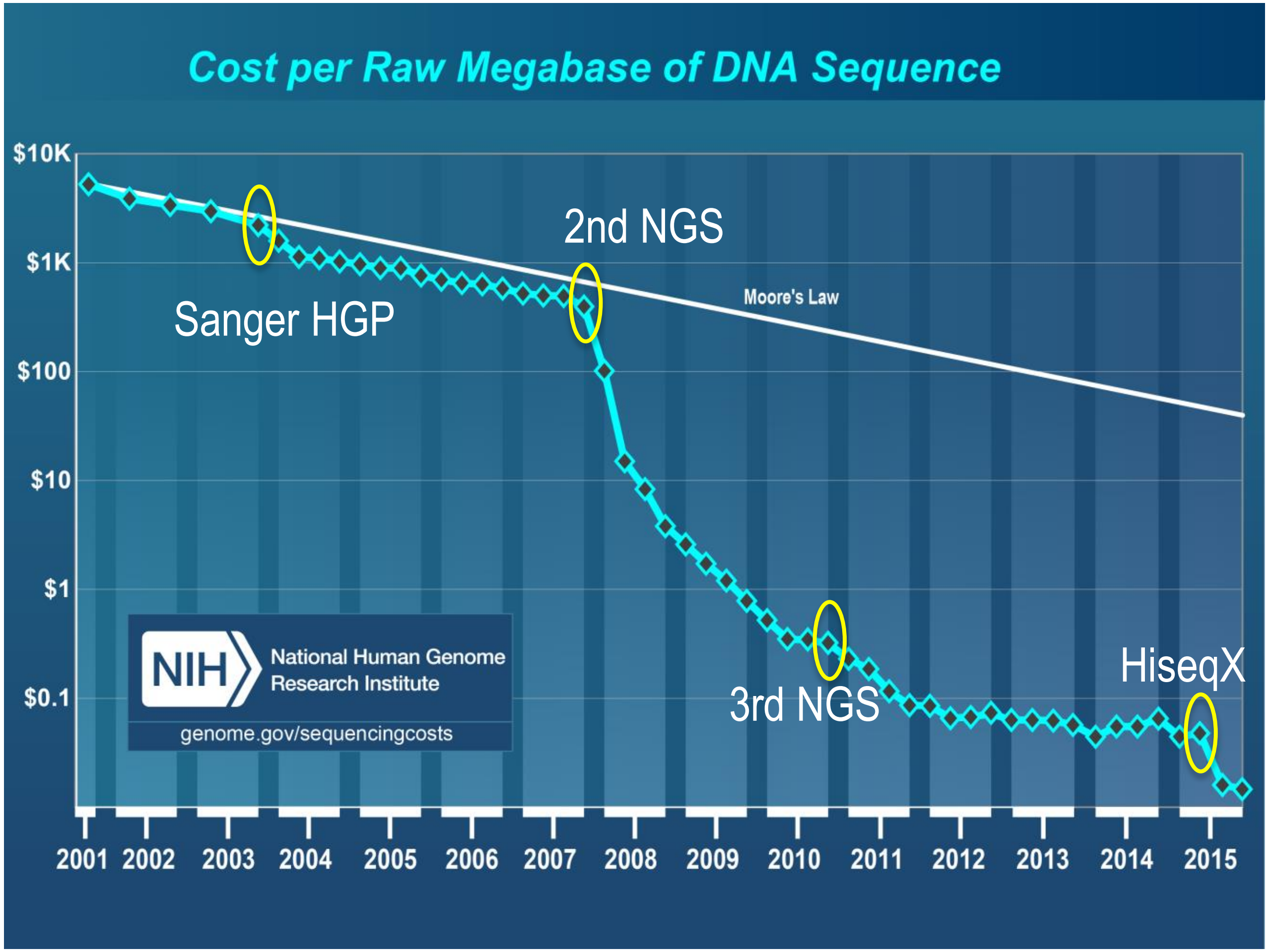

Residual pore current (pA)
60
50
40
30
20
10
T T T A T A A C G C T C A C C C C T C T G C A T C T C
30.4
30.6
30.8
31.0
31.2
Time (s)

Nanopore



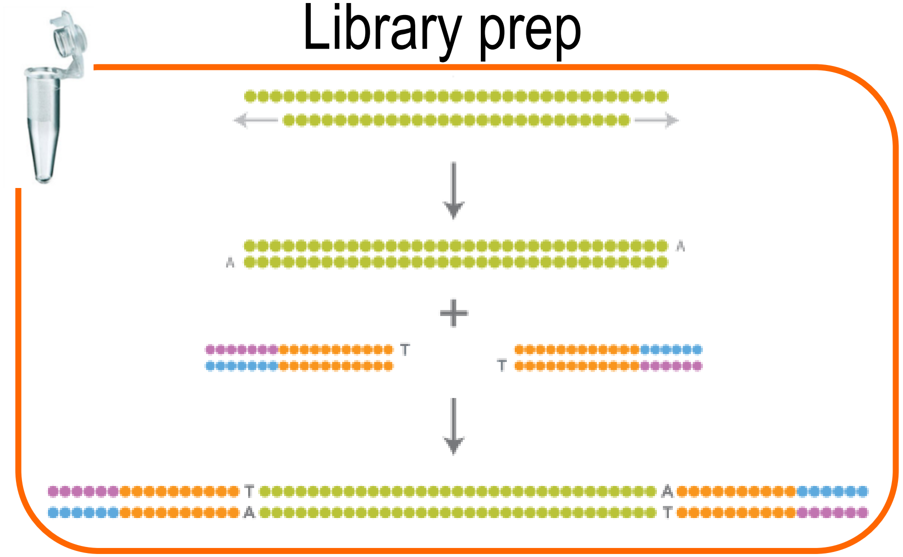
SMRT

Sequencing Costs

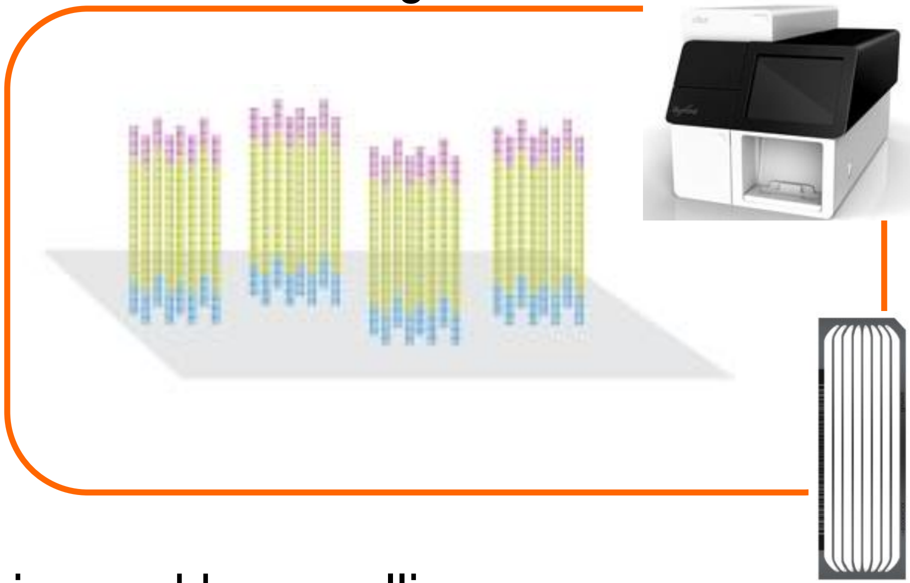


Illumina workflow

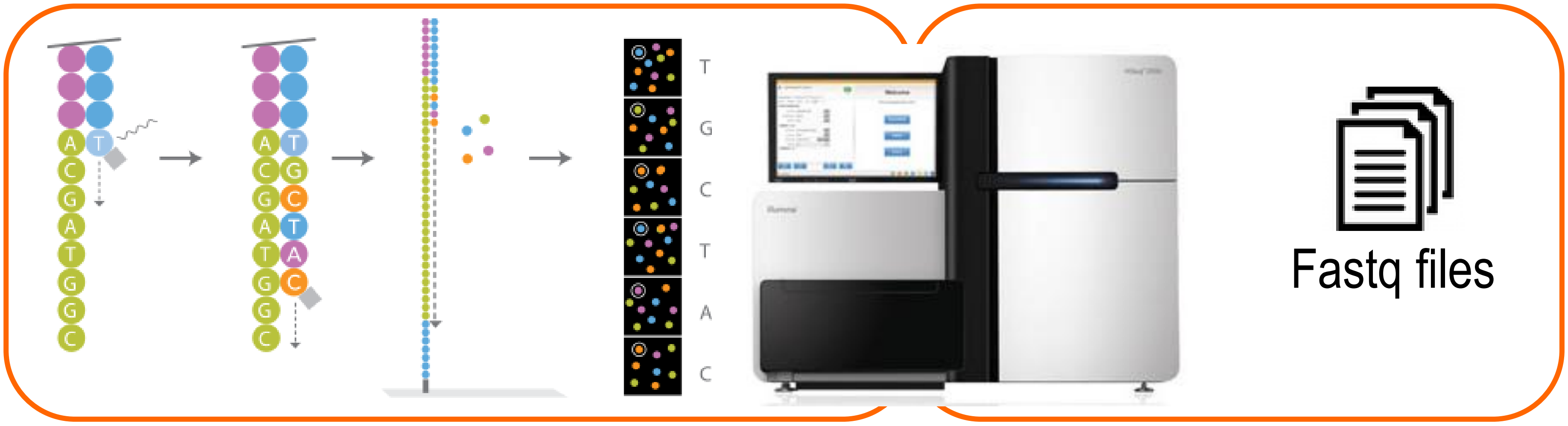
Library prep



Cluster generation



Sequencing, imaging and base calling



Fastq files

Fastq format

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

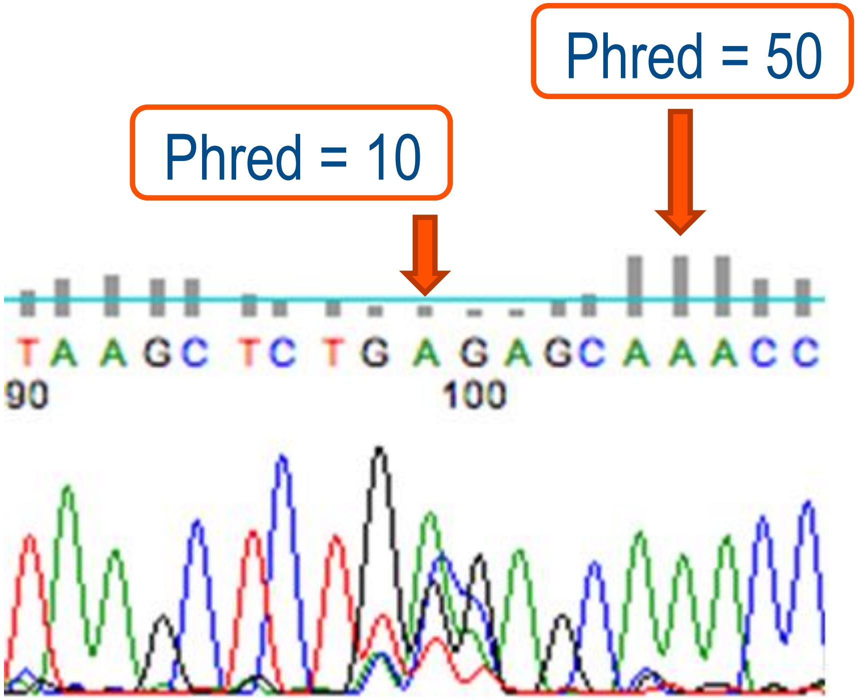
```
1) @HWI-H200:53:D08U2ACXX:5:1101:1231:2012 1:N:0:TACAGC
2) GCATTTTAGTAGAACCAGNCATTTCCCCCNACNTCNNTNCGNNANNNNTAA
3) +
4) @CCFFFFFHFFHHJJJJJ#3<FGIJJJJJ#1?###################
```

⬇ 31 ⬇ 37 ⬇ 39 ⬇ 18 ⬇ 16 ⬇ 2

Phred quality 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%

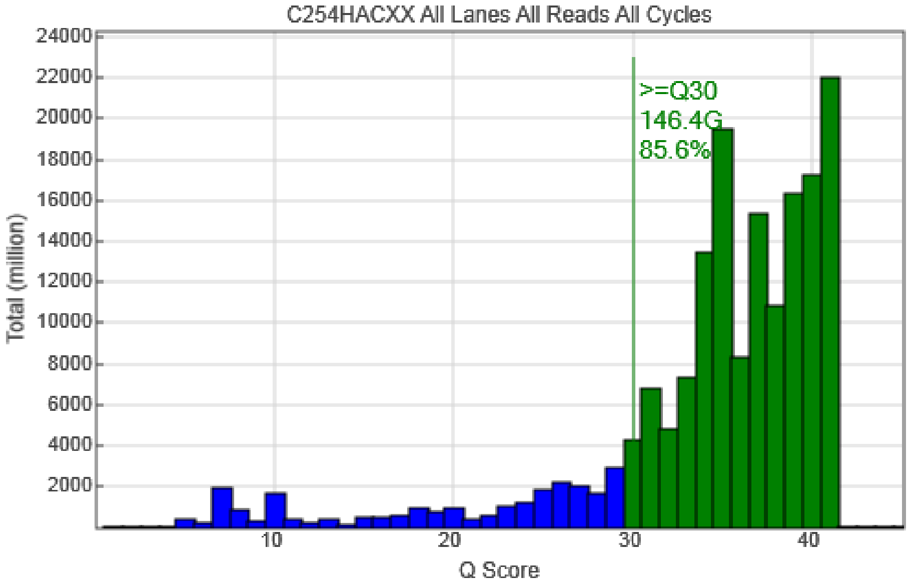


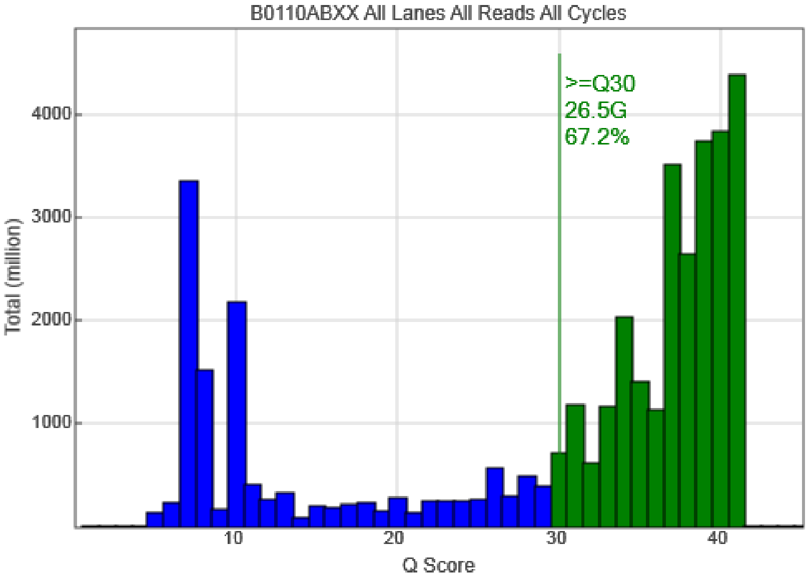
Phred = 10

Phred = 50

Sequencing run Quality

QScore Distribution





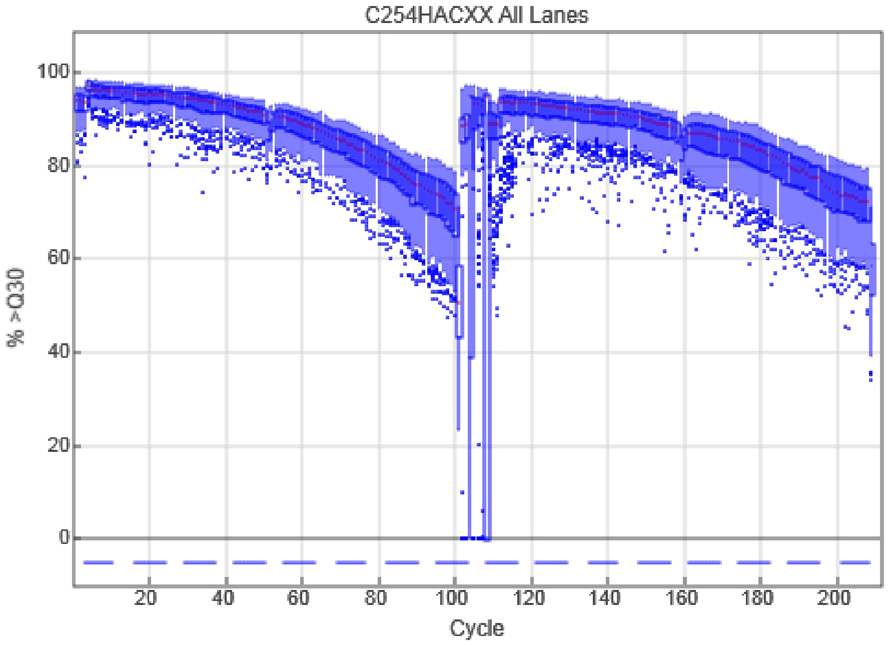
A succesful run should have 80% >= Q30

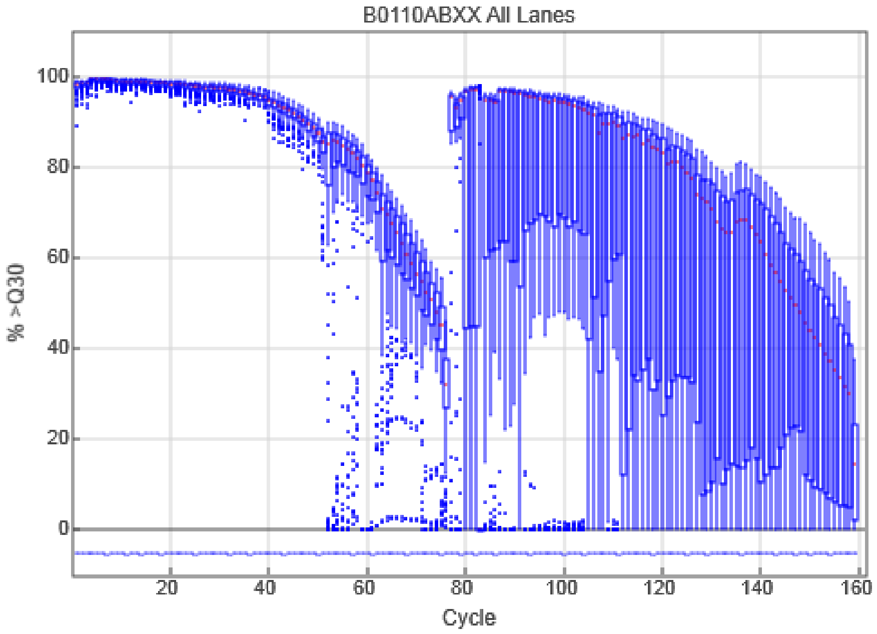
Illumina Sequencing Analysis Viewer

Bioinformatics
Core Facility

Sequencing run Quality

Data by Cycle

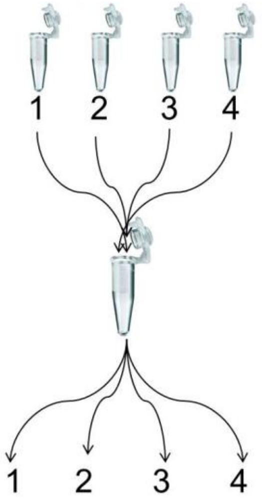


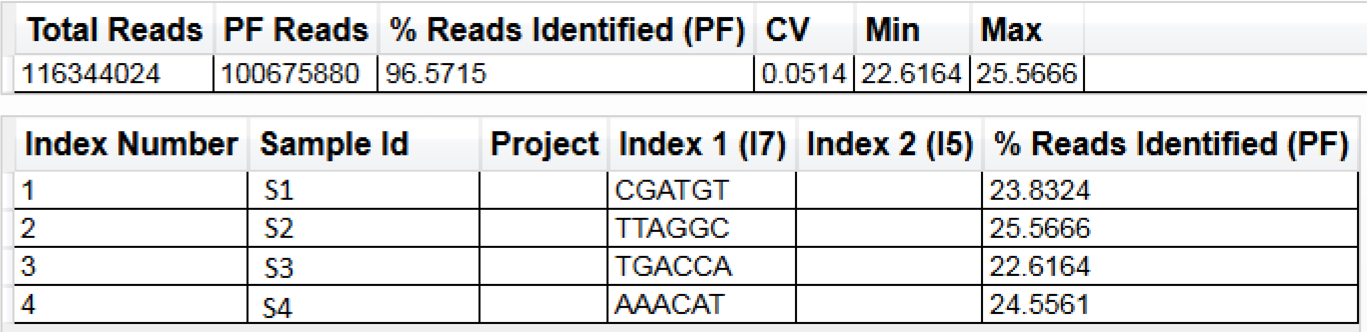


Illumina Sequencing Analysis Viewer

Sequencing run Quality

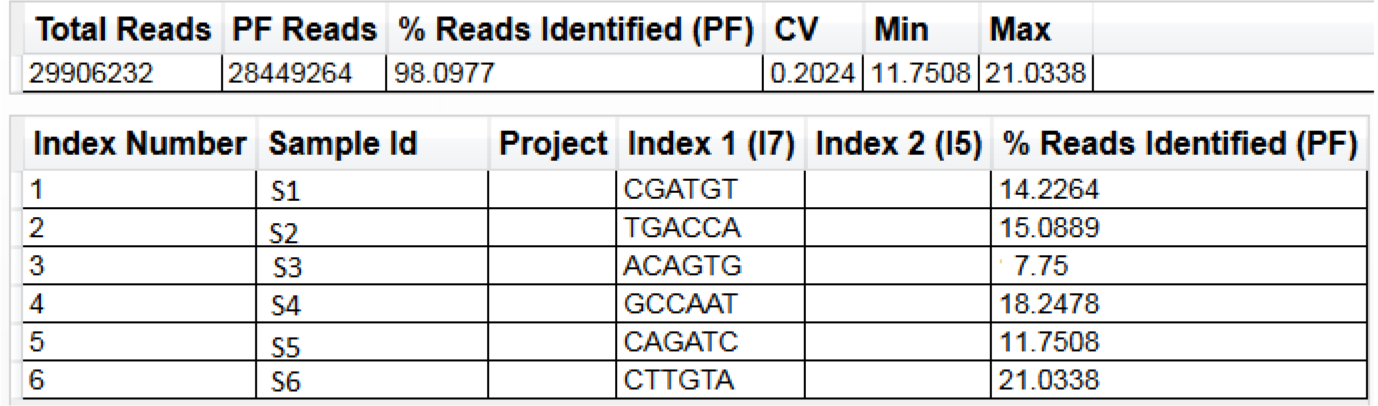
Demultiplexing





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



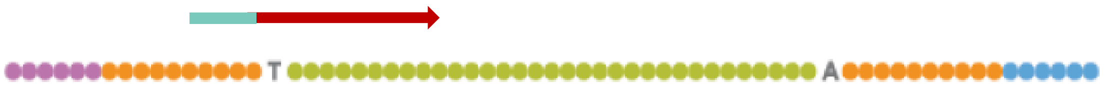
Total Reads	PF Reads	% Reads Identified (PF)	CV	Min	Max
29906232	28449264	98.0977	0.2024	11.7508	21.0338

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

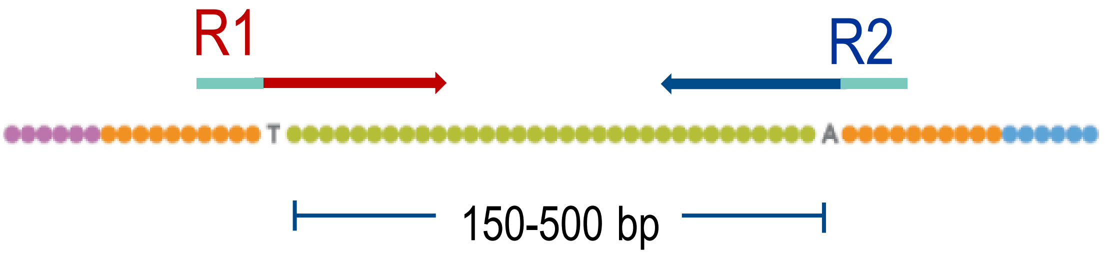
Illumina Sequencing Analysis Viewer - CASAVA

Different recepies

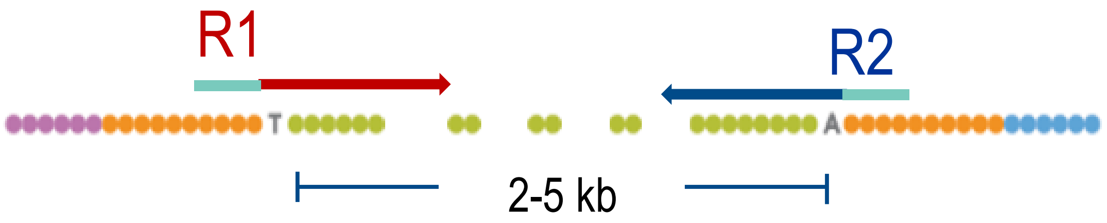
Single end (SE)



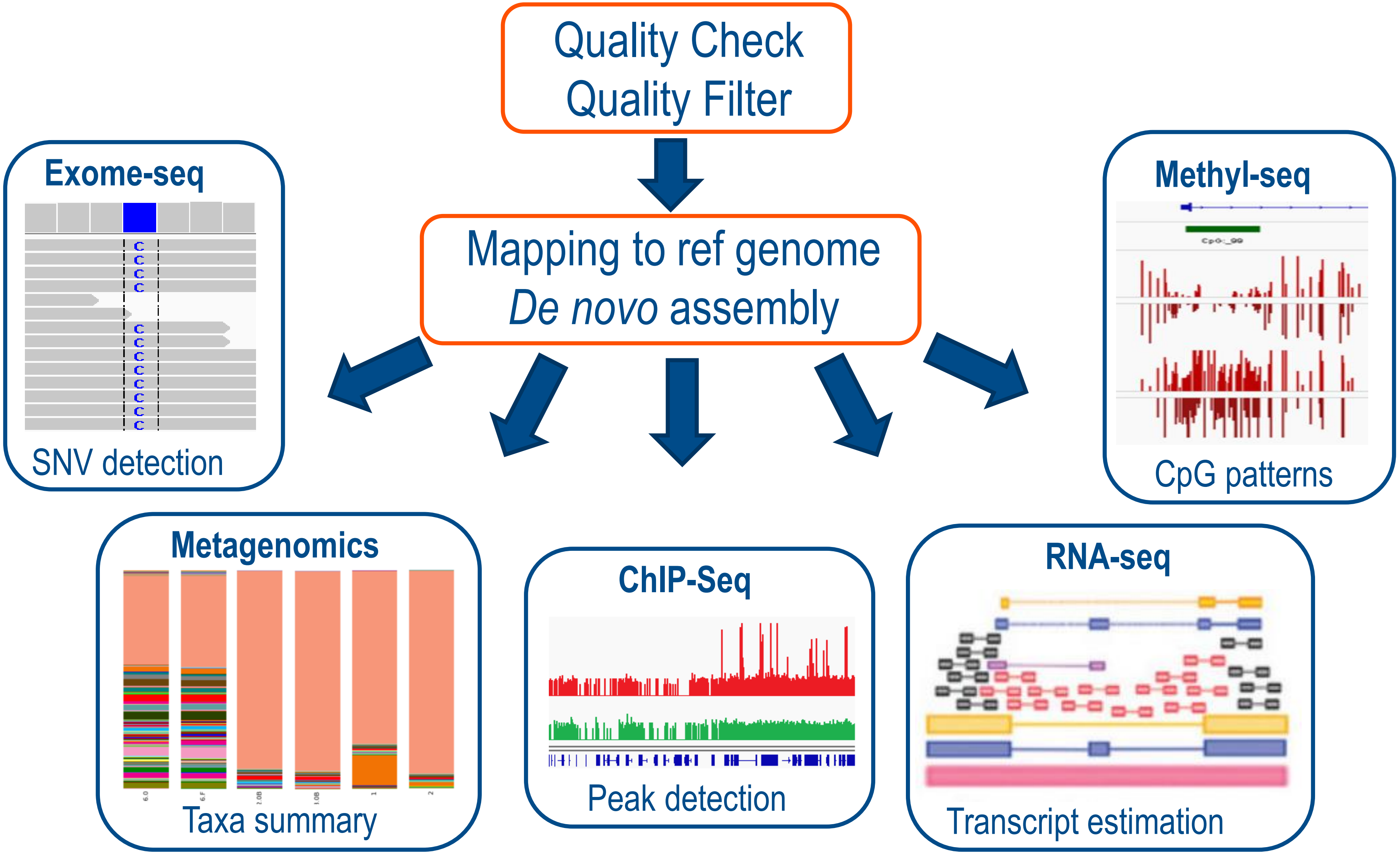
Paired-end (PE)



Mate-pair (MP)



Data handling workflow



Quality Check
Quality Filter

Mapping to ref genome
De novo assembly

Exome-seq

SNV detection

Methyl-seq

CpG patterns

Metagenomics

Taxa summary

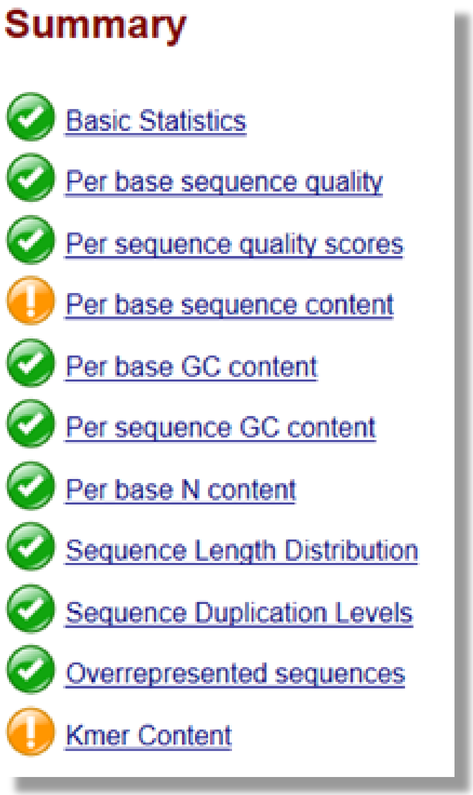
ChIP-Seq

Peak detection

RNA-seq

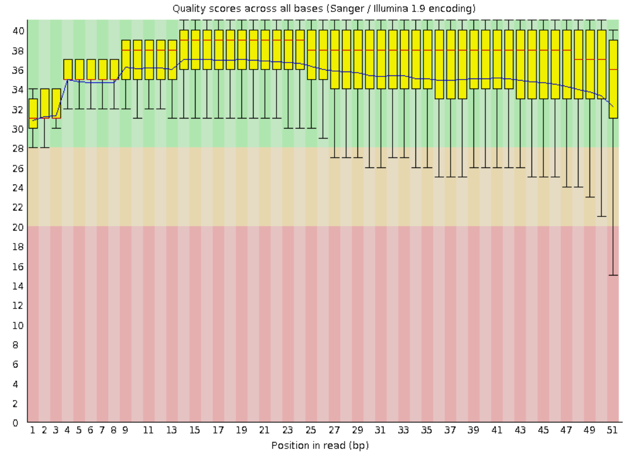
Transcript estimation

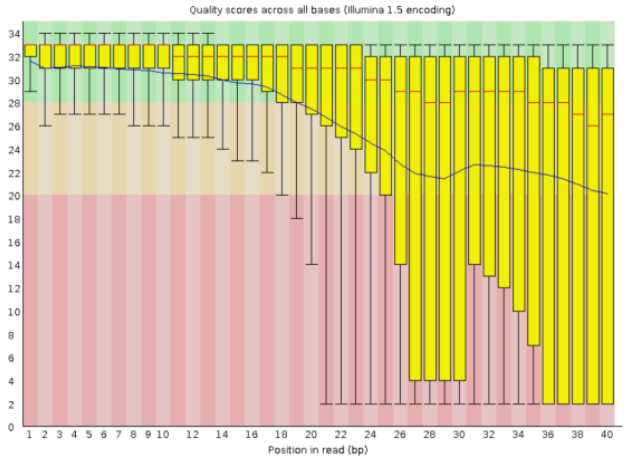
Quality check with FastQC

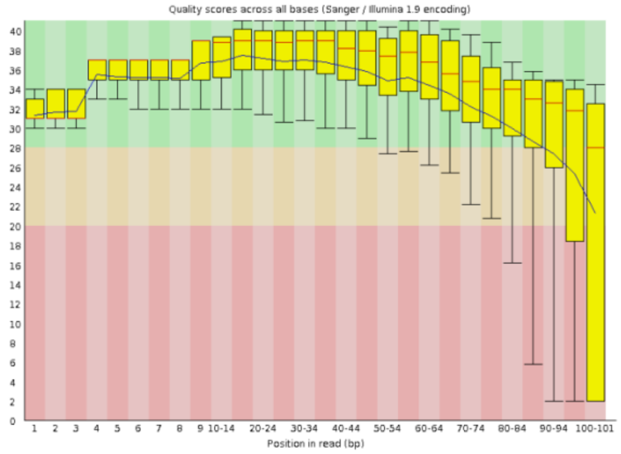


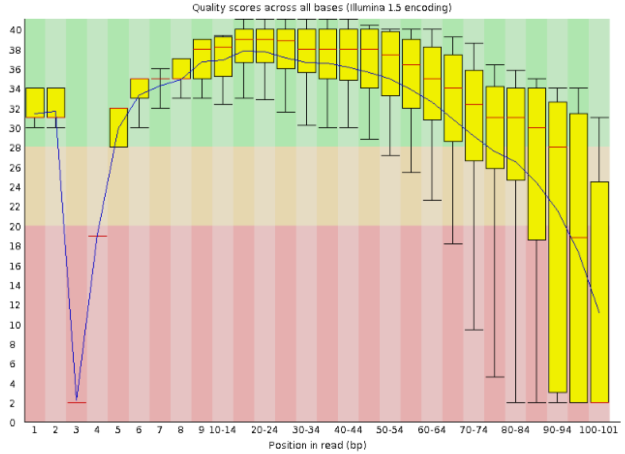
Summary

- ✓ 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



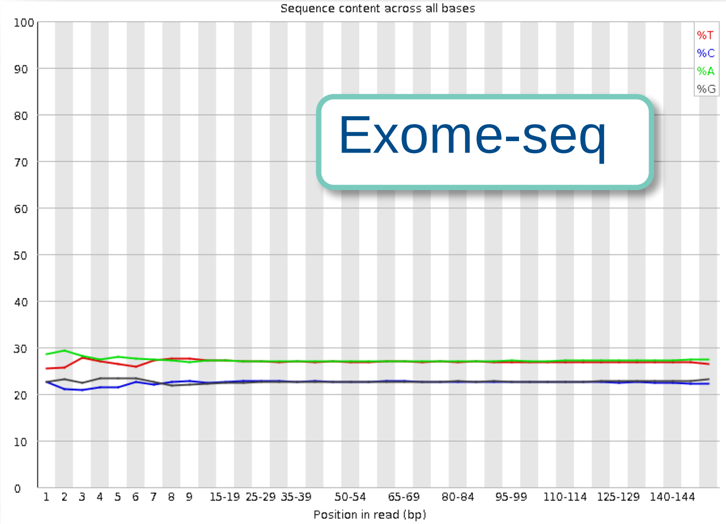






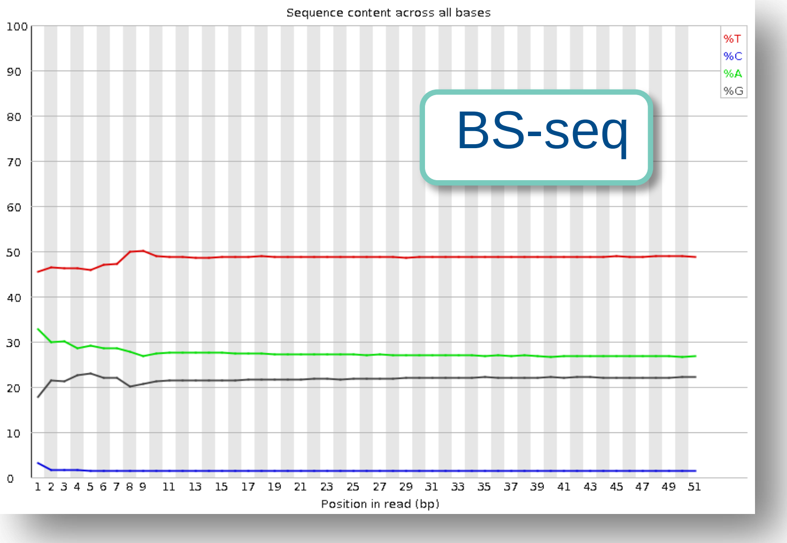
Per base sequence content

A



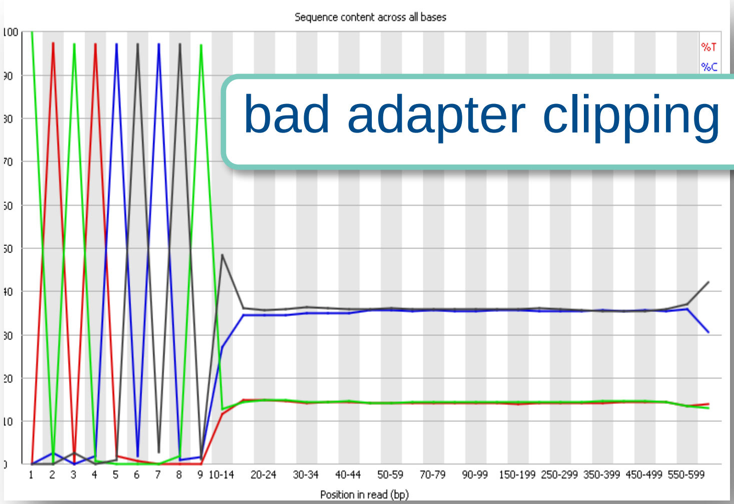
Exome-seq

C



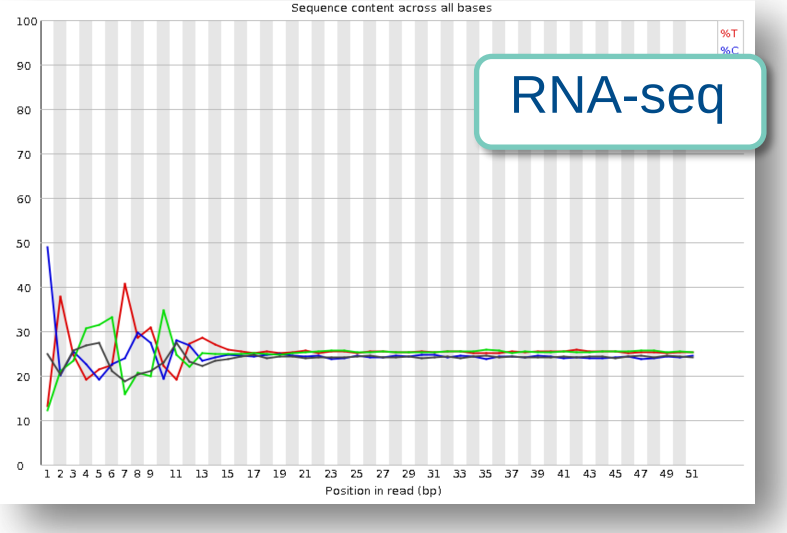
BS-seq

B



bad adapter clipping

D

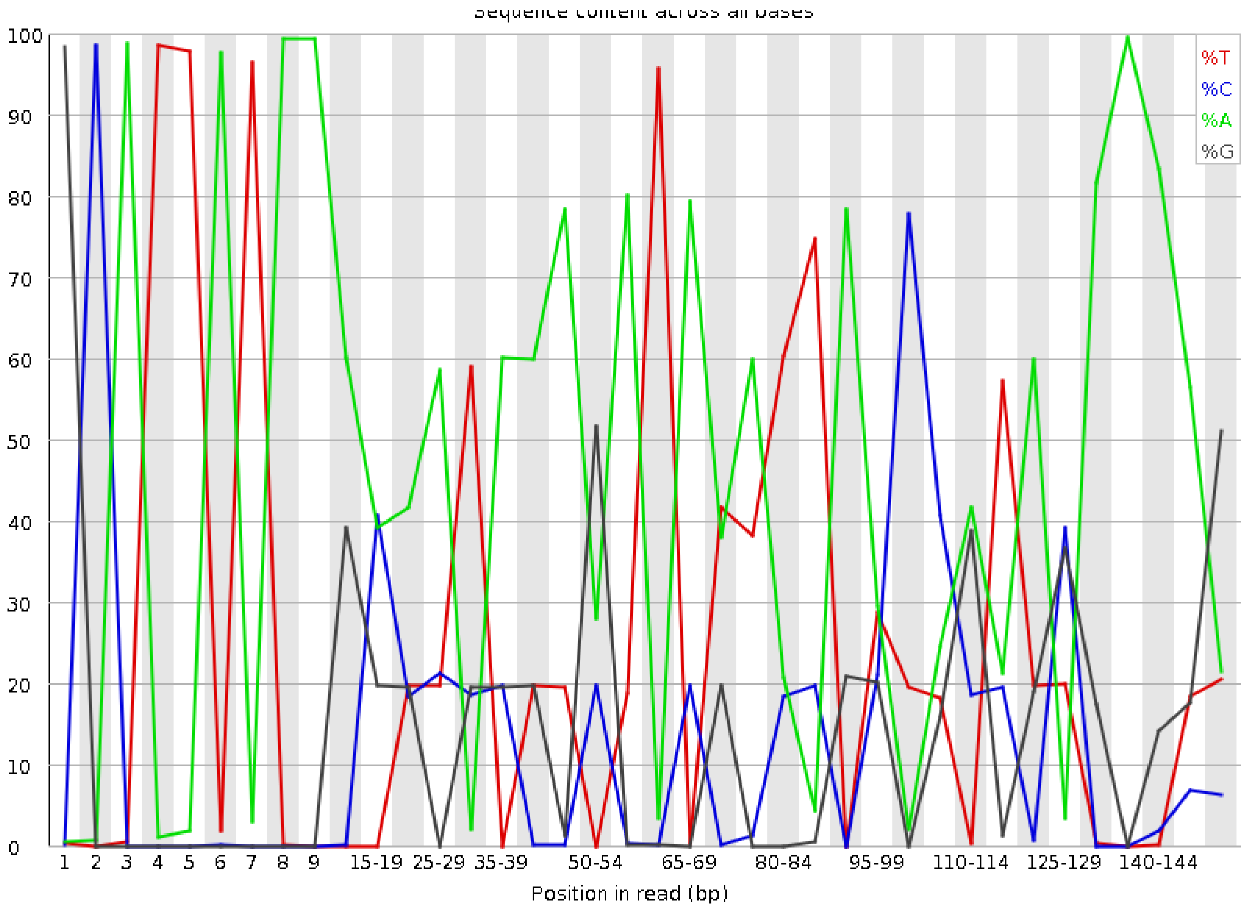


RNA-seq

GÖTEBORGS UNIVERSITET

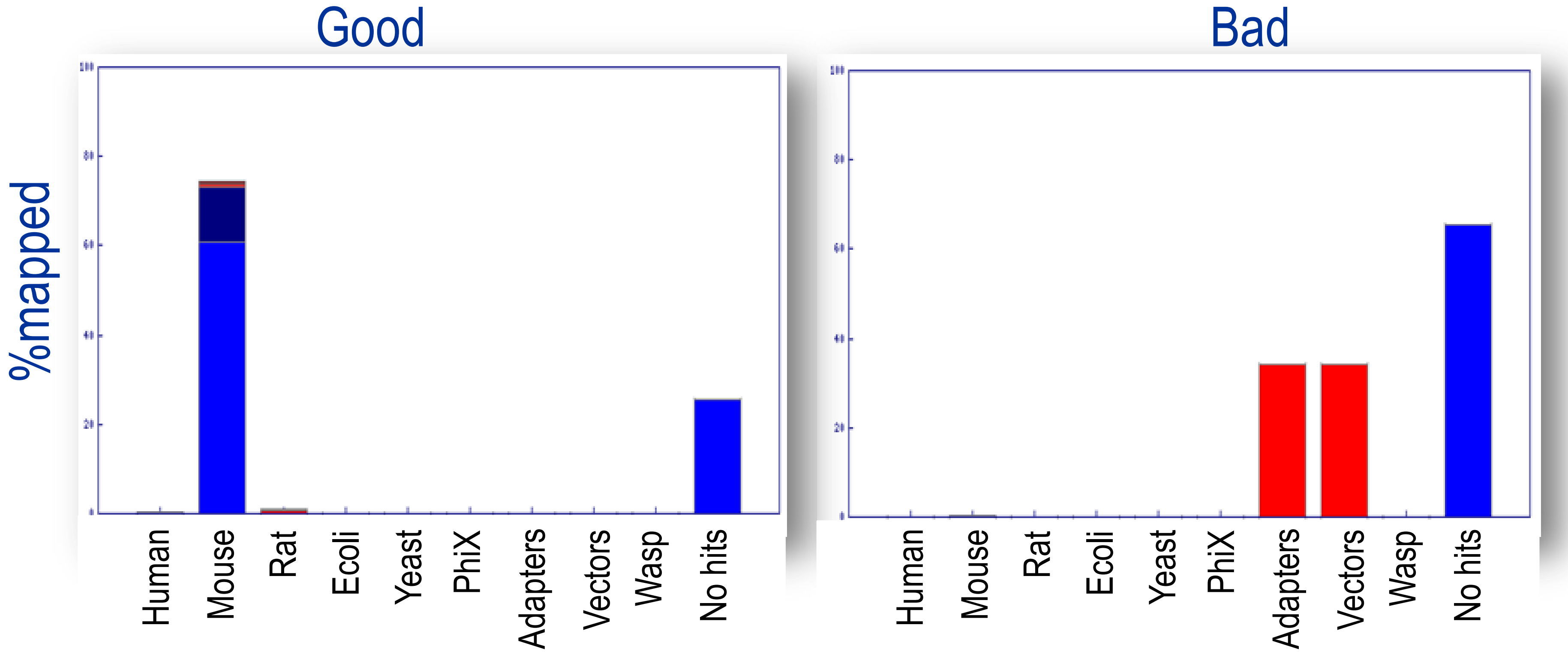
Bioinformatics
Core Facility

Per base sequence content

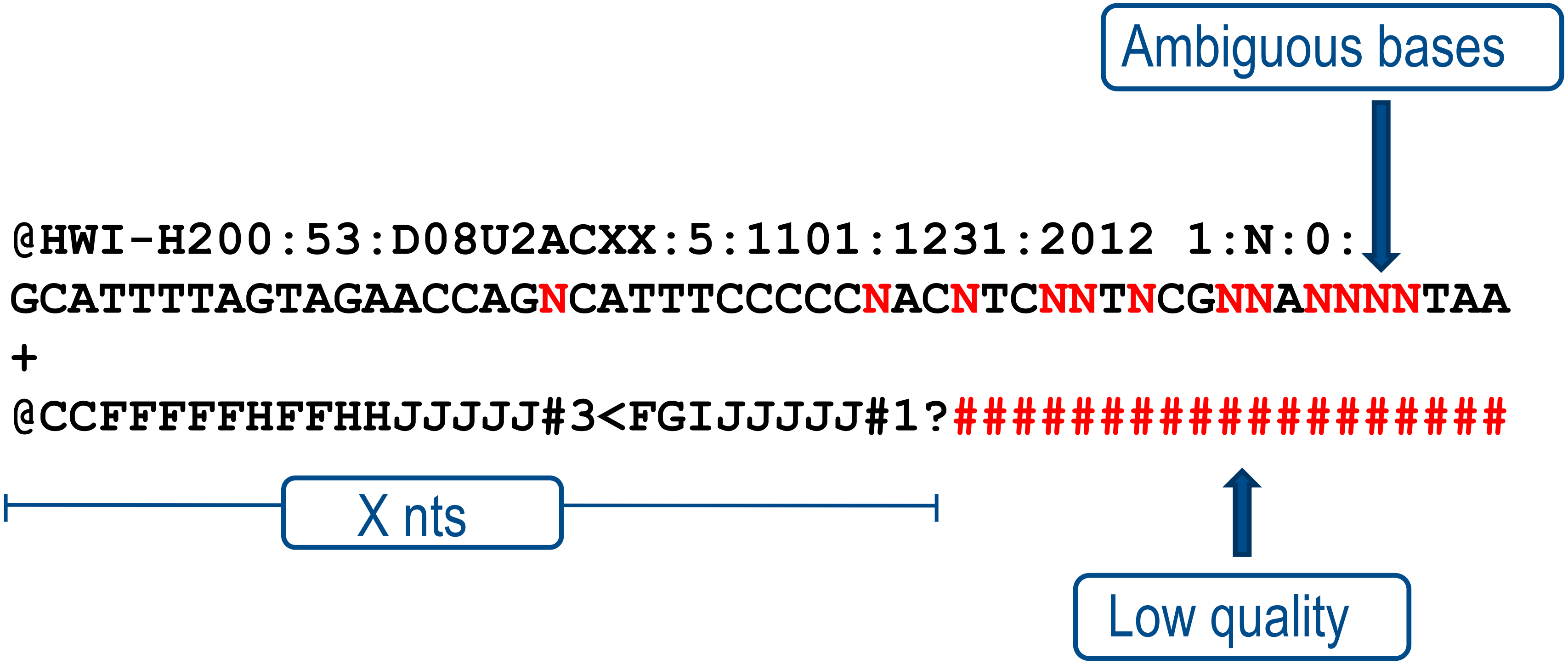


Amplicon seq

Contamination check with FastQScreen



Quality Filter with PRINSeq



Ambiguous bases

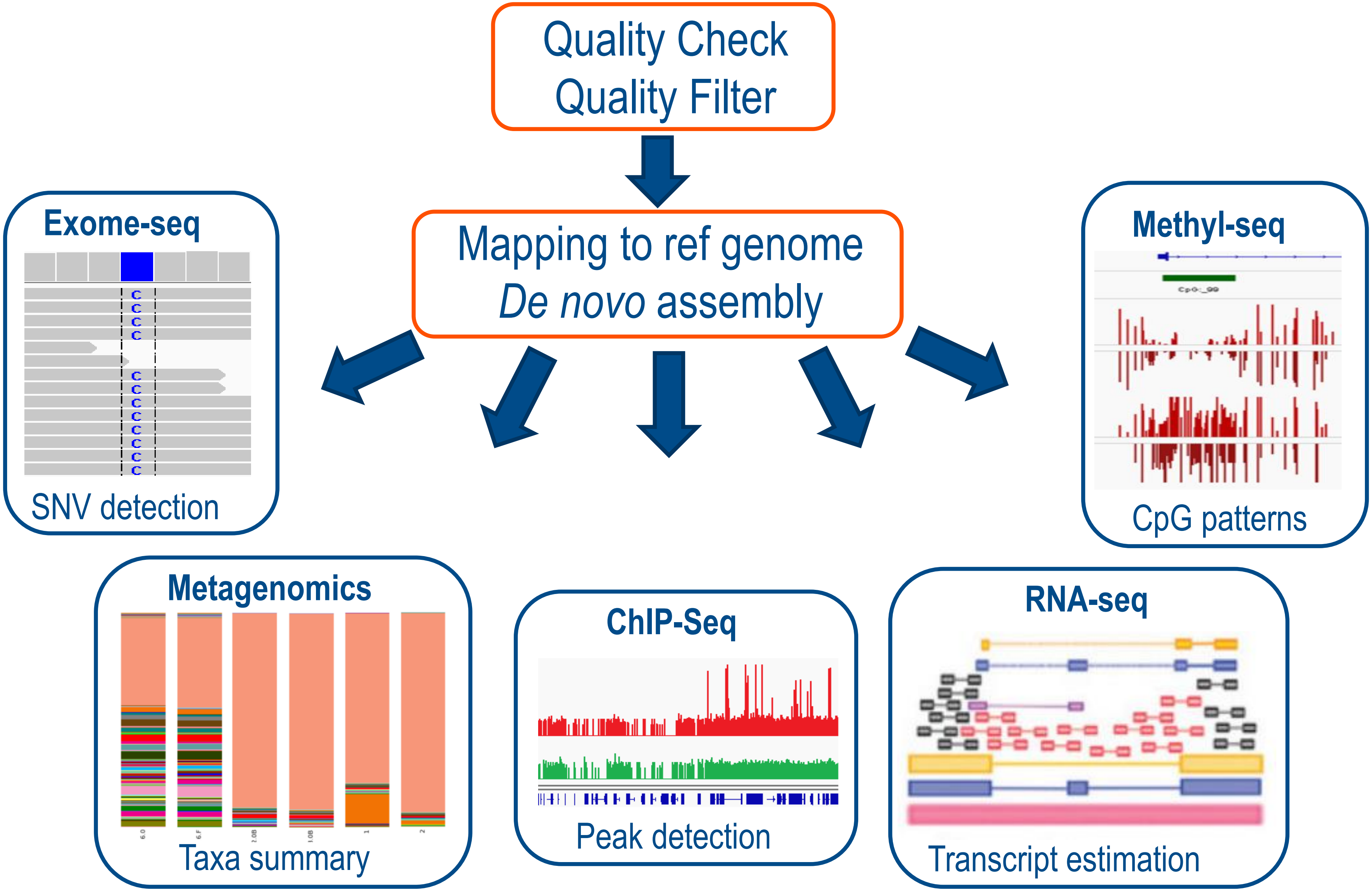
```
@HWI-H200:53:D08U2ACXX:5:1101:1231:2012 1:N:0:
GCATTTTAGTAGAACCAGNCATTTCCCCCNACNTCNNTNCGNNANNNNTAA
+
@CCFFFFFHFFHHJJJJJ#3<FGIJJJJJ#1?###################
```

X nts

Low quality

TRIM-galore

Data handling workflow



Quality Check
Quality Filter

Mapping to ref genome
De novo assembly

Exome-seq
SNV detection

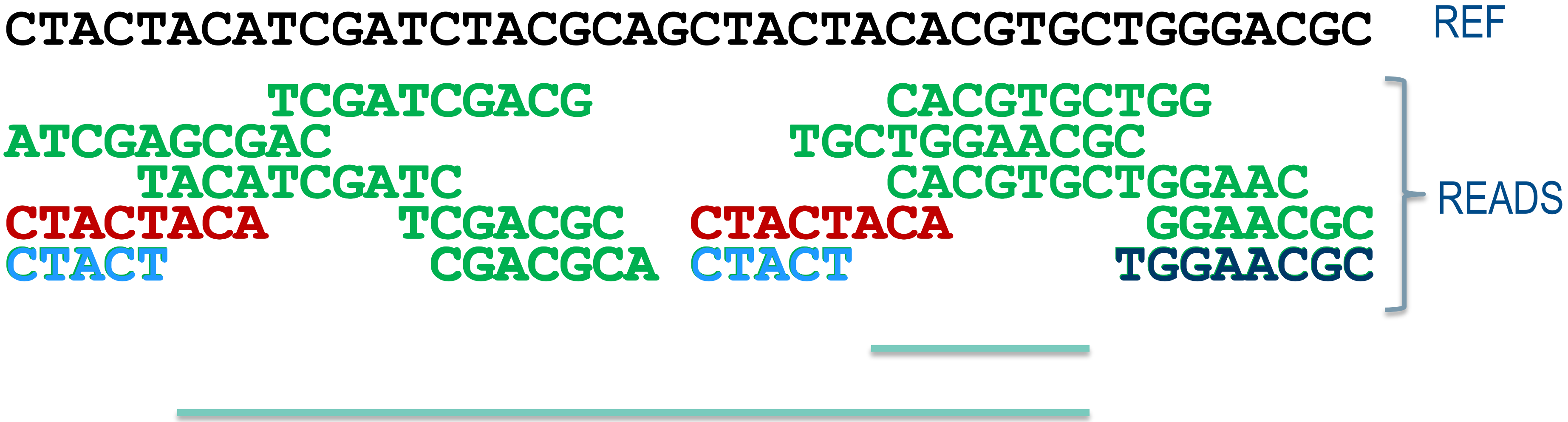
Methyl-seq
CpG patterns

Metagenomics
Taxa summary

ChIP-Seq
Peak detection

RNA-seq
Transcript estimation

Mapping



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

WHERE to place the reads?

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

mean DNA fragment size: 40

Bfast, BioScope, **Bowtie**, **BWA,** CLC bio, CloudBurst, Eland/Eland2, GenomeMapper, GnuMap, Karma, **MAQ**, MOM, **Mosaik**, MrFAST/MrsFAST, NovoAlign, PASS, PerM, RazerS, RMAP, SSAHA2, Segemehl, …

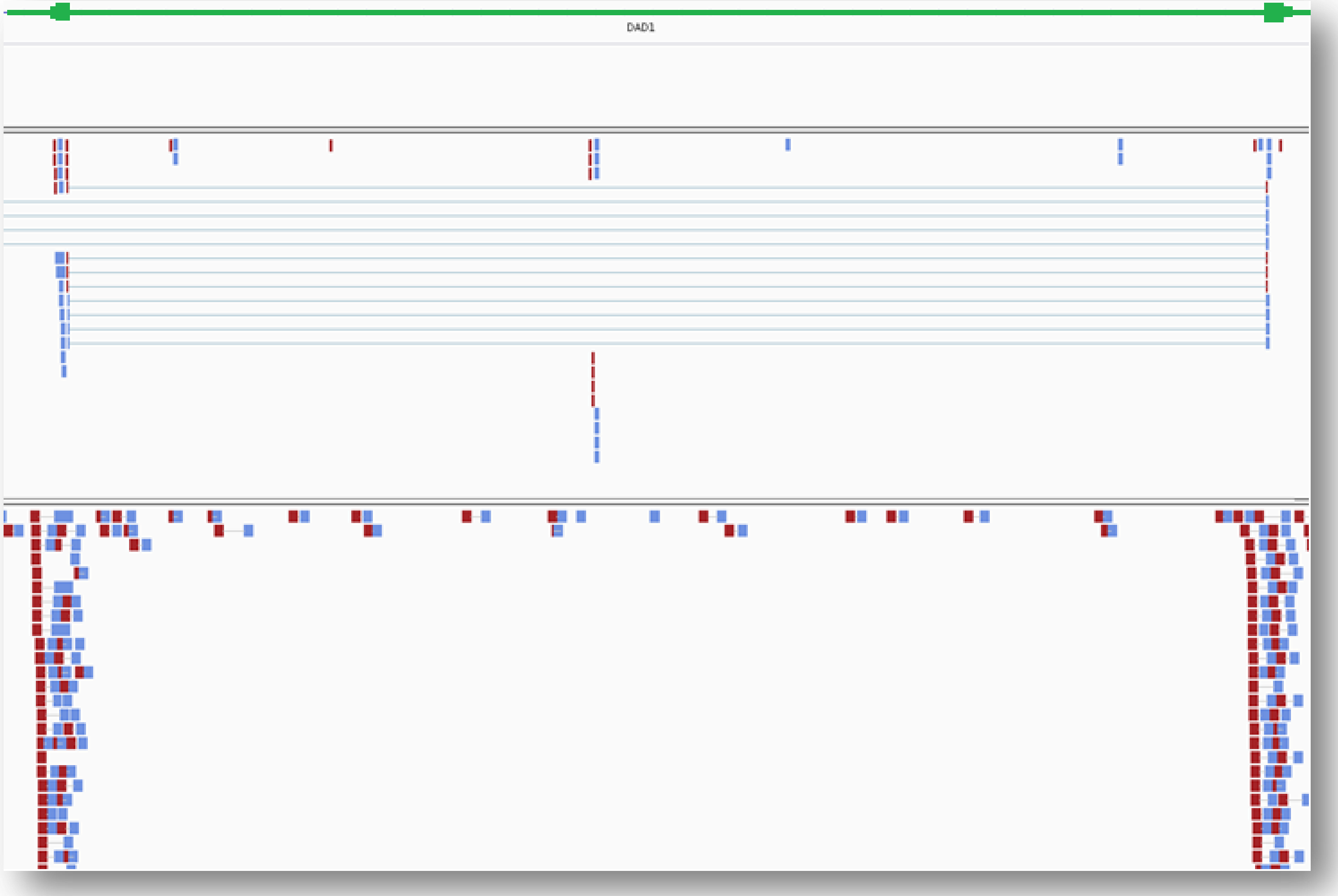
BAM files

Mapping

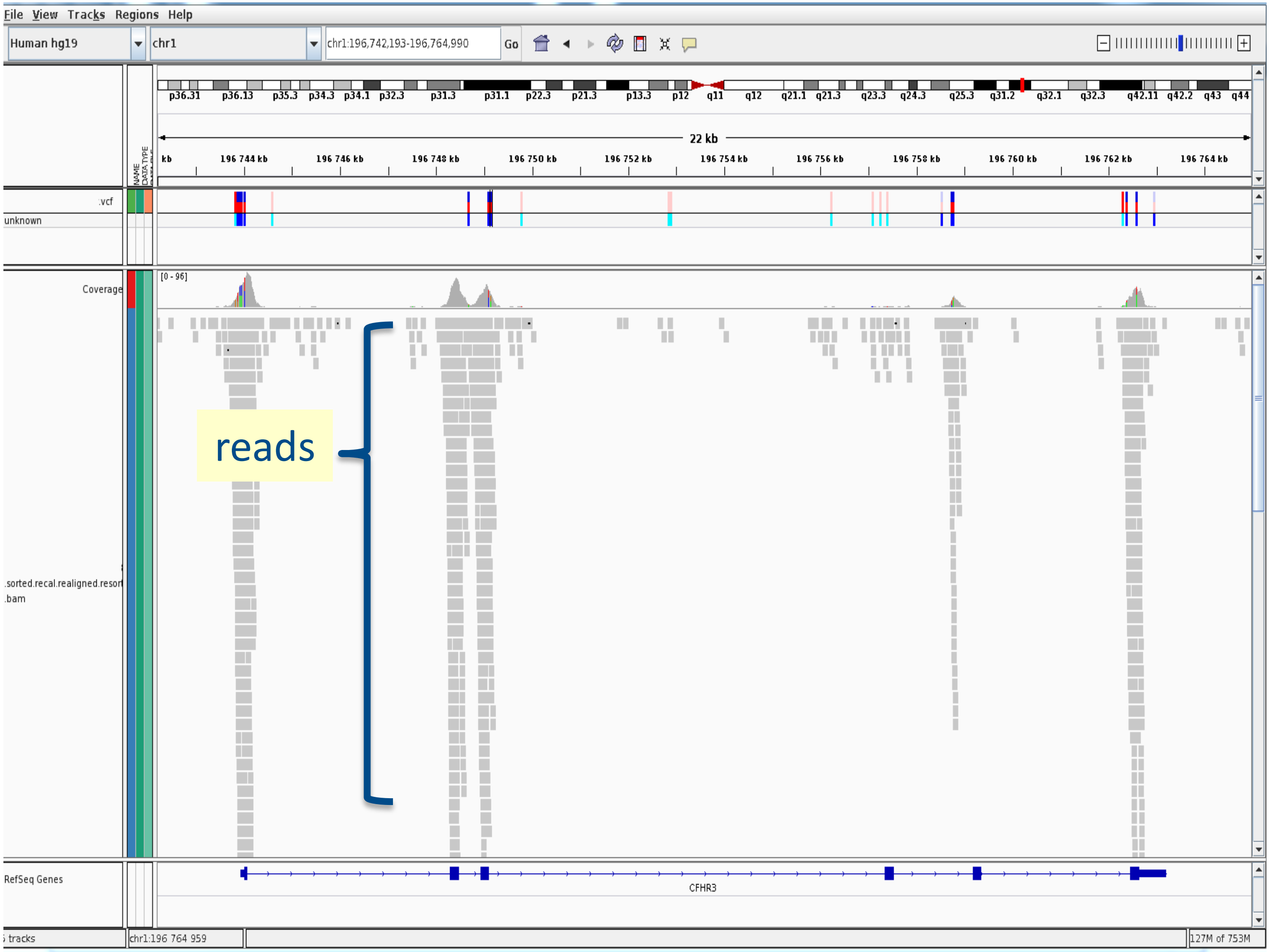
HOW to place the reads? Ungapped, Gapped

RNA-seq

Exome



IGV – Integrative Genome Viewer


File View Tracks Regions Help
Go
p36.31 p36.13 p35.3 p34.3 p34.1 p32.3 p31.3 p31.1 p22.3 p21.3 p13.3 p12 q11 q12 q21.1 q21.3 q23.3 q24.3 q25.3 q31.2 q32.1 q32.3 q42.11 q42.2 q43 q44
22 kb
kb 196 744 kb 196 746 kb 196 748 kb 196 750 kb 196 752 kb 196 754 kb 196 756 kb 196 758 kb 196 760 kb 196 762 kb 196 764 kb
NAME
DATA TYPE
.vcf
unknown
Coverage
[0 - 96]
.sorted.recal.realigned.resort
.bam
RefSeq Genes
CFHR3
tracks
127M of 753M

Human hg19
chr1
chr1:196,742,193-196,764,990
chr1:196 764 959

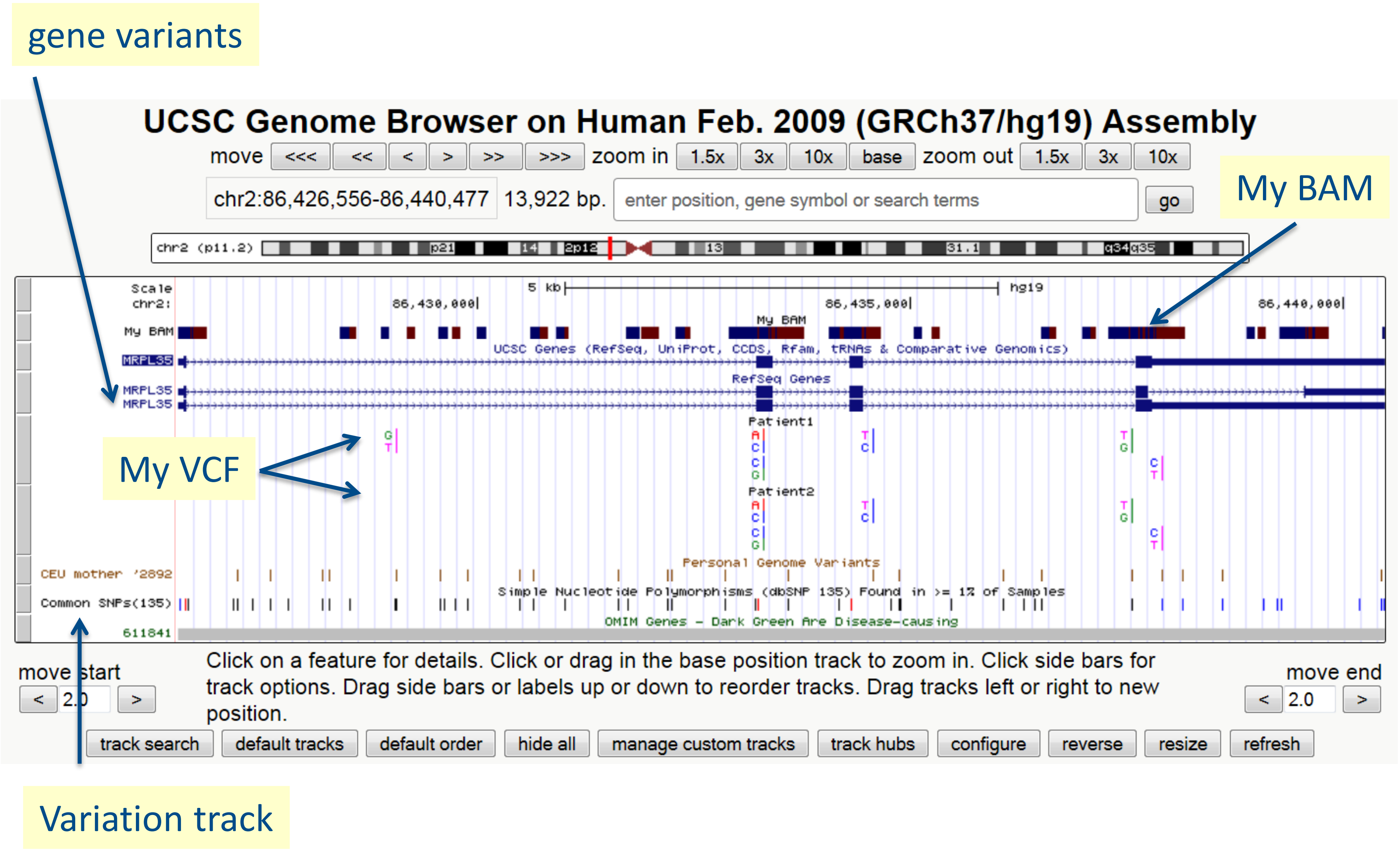
reads

coverage

My BAM

gene

UCSC browser



gene variants

My BAM

My VCF

Variation track

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr2:86,426,556-86,440,477 13,922 bp.

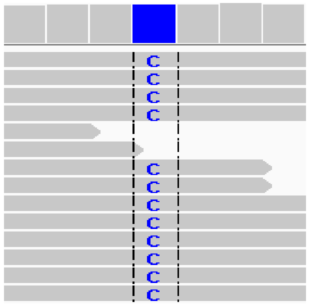
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Data handling workflow

Quality Check
Quality Filter

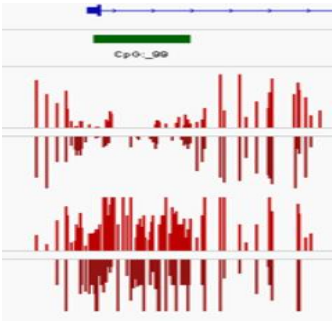
Mapping to ref genome
De novo assembly

Exome-seq



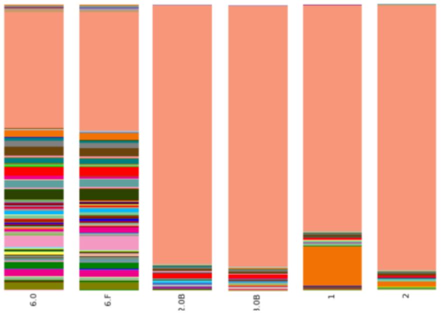
SNV detection

Methyl-seq



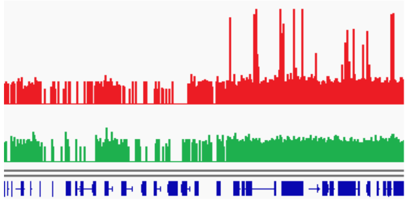
CpG patterns

Metagenomics



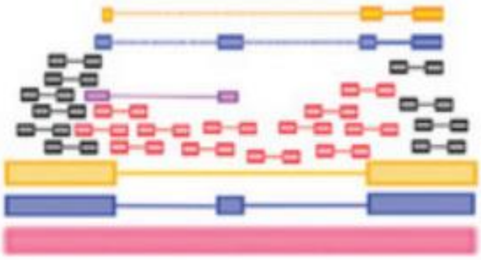
Taxa summary

ChIP-Seq



Peak detection

RNA-seq



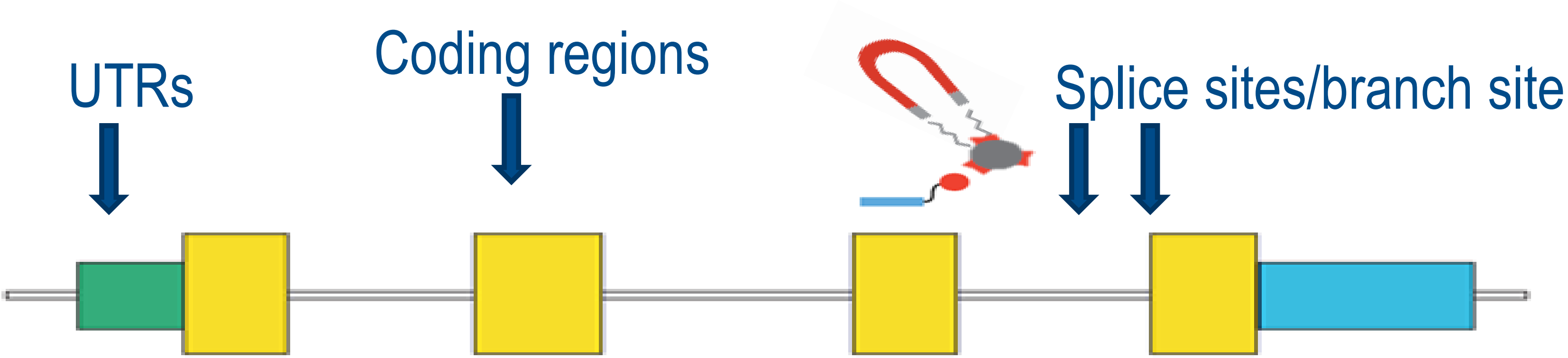
Transcript estimation

Monogenic diseases

http://www.who.int/

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

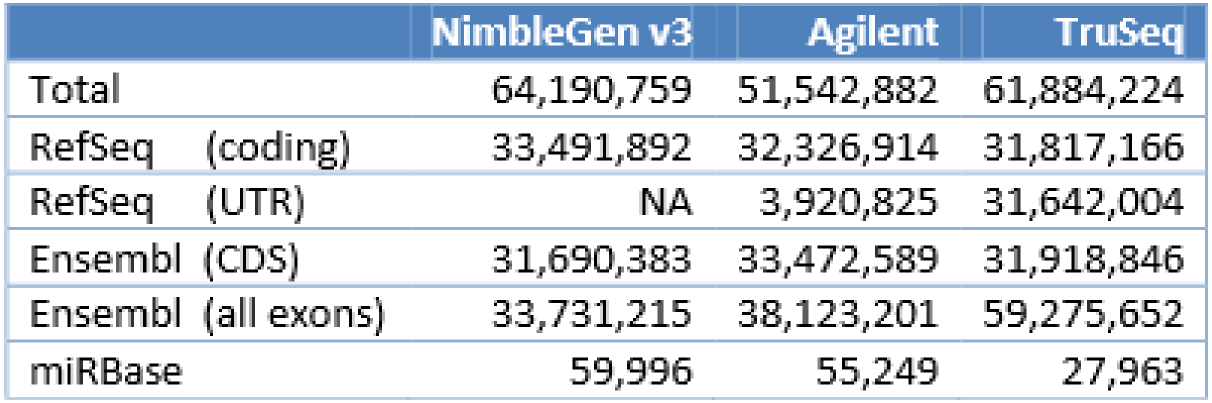


UTRs

Coding regions

Splice sites/branch site

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

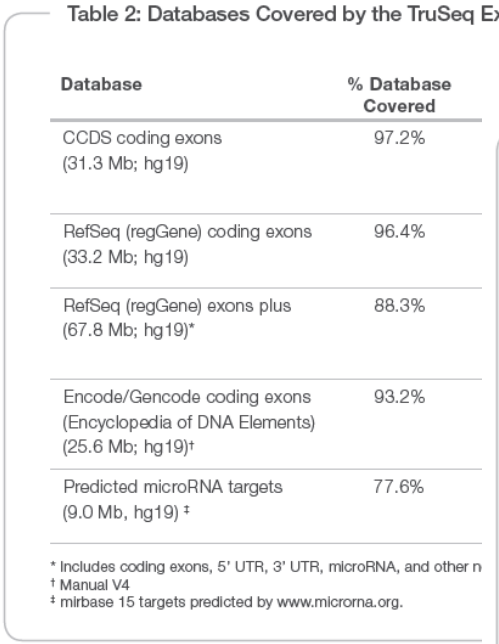


Table 2: Databases Covered by the TruSeq Exome Enrichment Kit

Database	% Database Covered	Description
CCDS coding exons (31.3 Mb; hg19)	97.2%	
RefSeq (regGene) coding exons (33.2 Mb; hg19)	96.4%	
RefSeq (regGene) exons plus (67.8 Mb; hg19)*	88.3%	
Encode/Gencode coding exons (Encyclopedia of DNA Elements) (25.6 Mb; hg19)†	93.2%	
Predicted microRNA targets (9.0 Mb, hg19) ‡	77.6%	

* Includes coding exons, 5' UTR, 3' UTR, microRNA, and other n
† Manual V4
‡ mirbase 15 targets predicted by www.microrna.org.

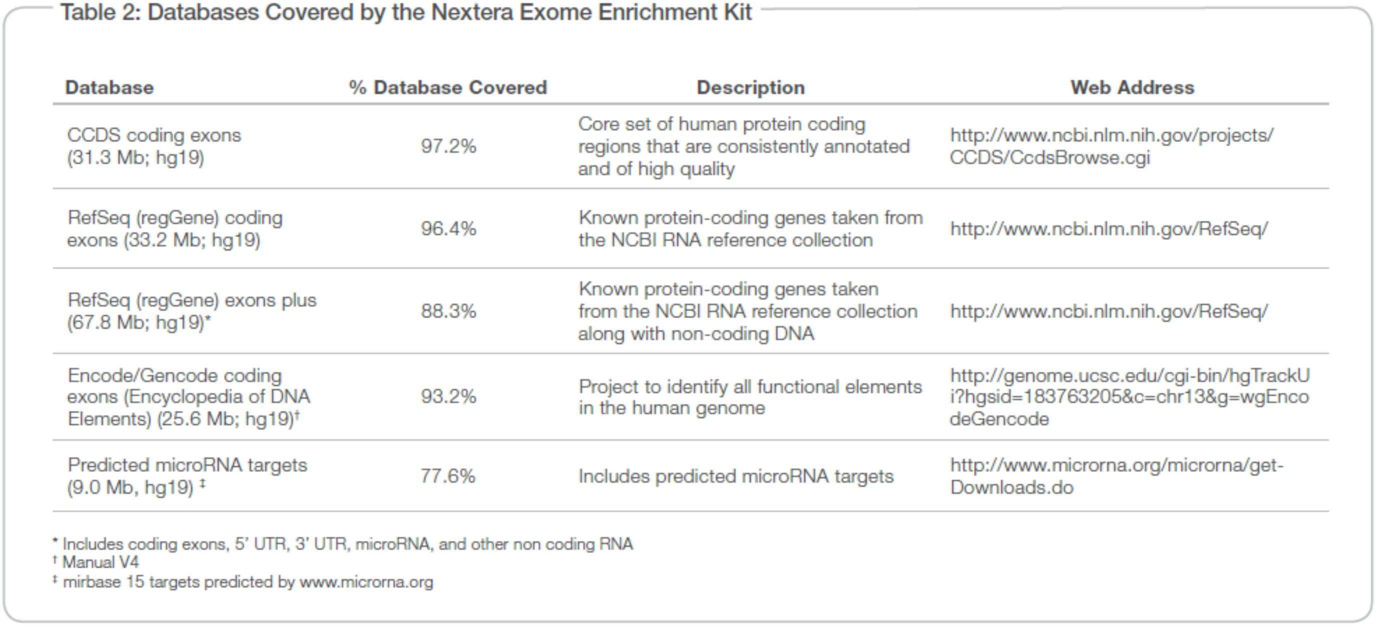


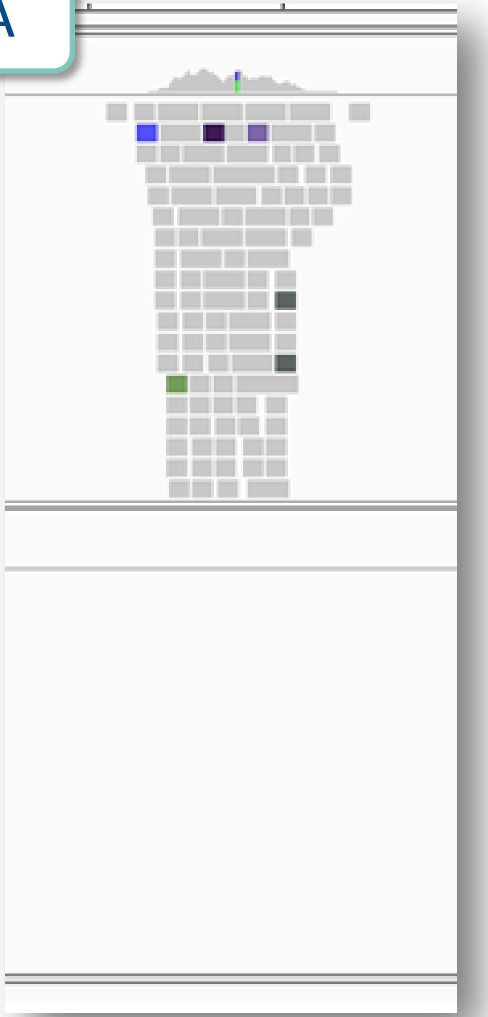
Table 2: Databases Covered by the Nextera Exome Enrichment Kit

Database	% Database Covered	Description	Web Address
CCDS coding exons (31.3 Mb; hg19)	97.2%	Core set of human protein coding regions that are consistently annotated and of high quality	http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi
RefSeq (regGene) coding exons (33.2 Mb; hg19)	96.4%	Known protein-coding genes taken from the NCBI RNA reference collection	http://www.ncbi.nlm.nih.gov/RefSeq/
RefSeq (regGene) exons plus (67.8 Mb; hg19)*	88.3%	Known protein-coding genes taken from the NCBI RNA reference collection along with non-coding DNA	http://www.ncbi.nlm.nih.gov/RefSeq/
Encode/Gencode coding exons (Encyclopedia of DNA Elements) (25.6 Mb; hg19)†	93.2%	Project to identify all functional elements in the human genome	http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=183763205&c=chr13&g=wgEncodeGencode
Predicted microRNA targets (9.0 Mb, hg19) ‡	77.6%	Includes predicted microRNA targets	http://www.microrna.org/microrna/get-Downloads.do

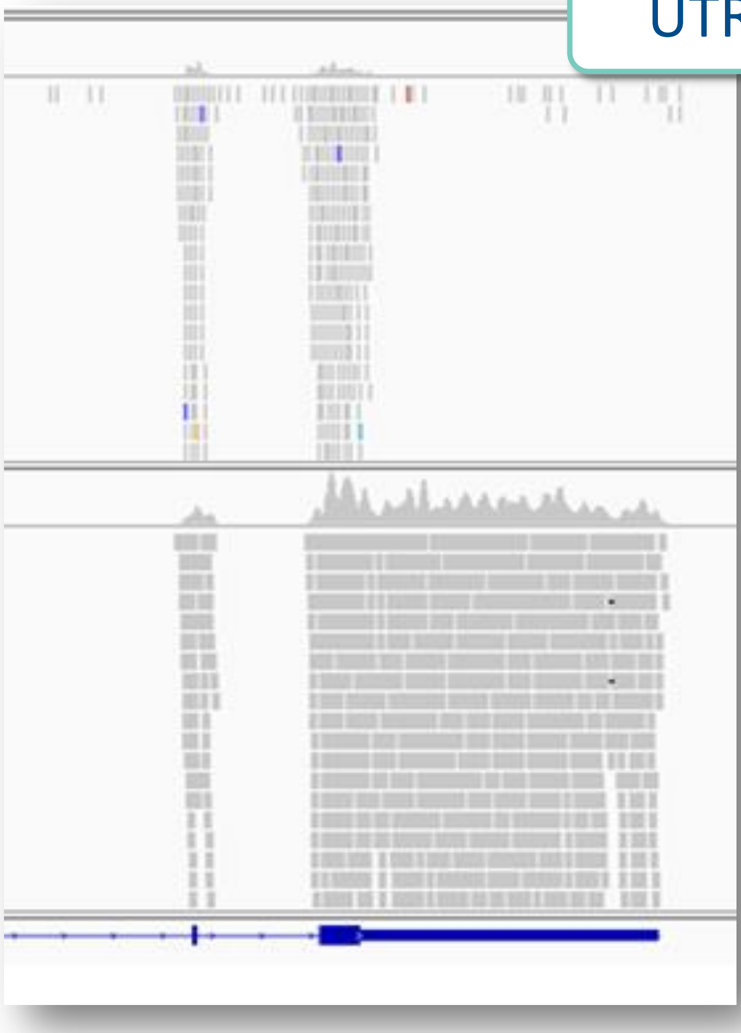
* Includes coding exons, 5' UTR, 3' UTR, microRNA, and other non coding RNA
† Manual V4
‡ mirbase 15 targets predicted by www.microrna.org

Enrichment kits

miRNA



UTR's

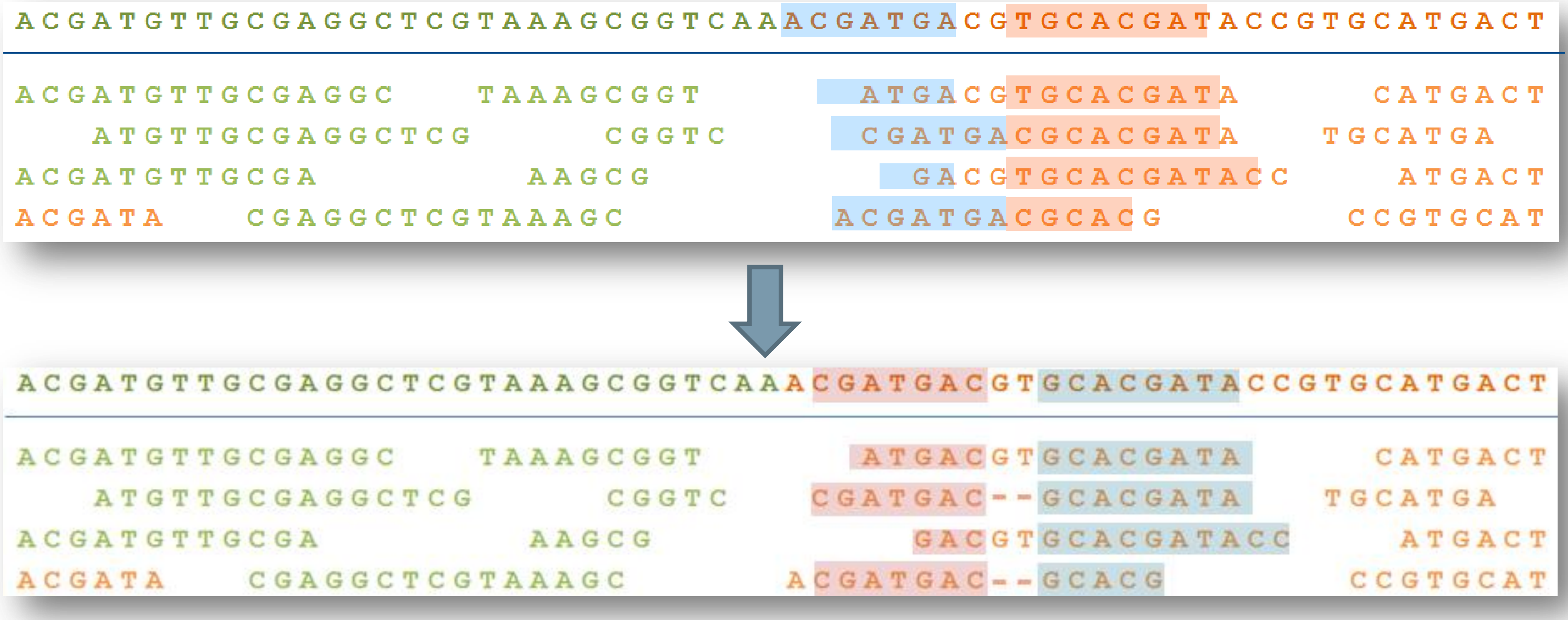


GÖTEBORGS UNIVERSITET

Bioinformatics
Core Facility

Realignment and recalibration

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

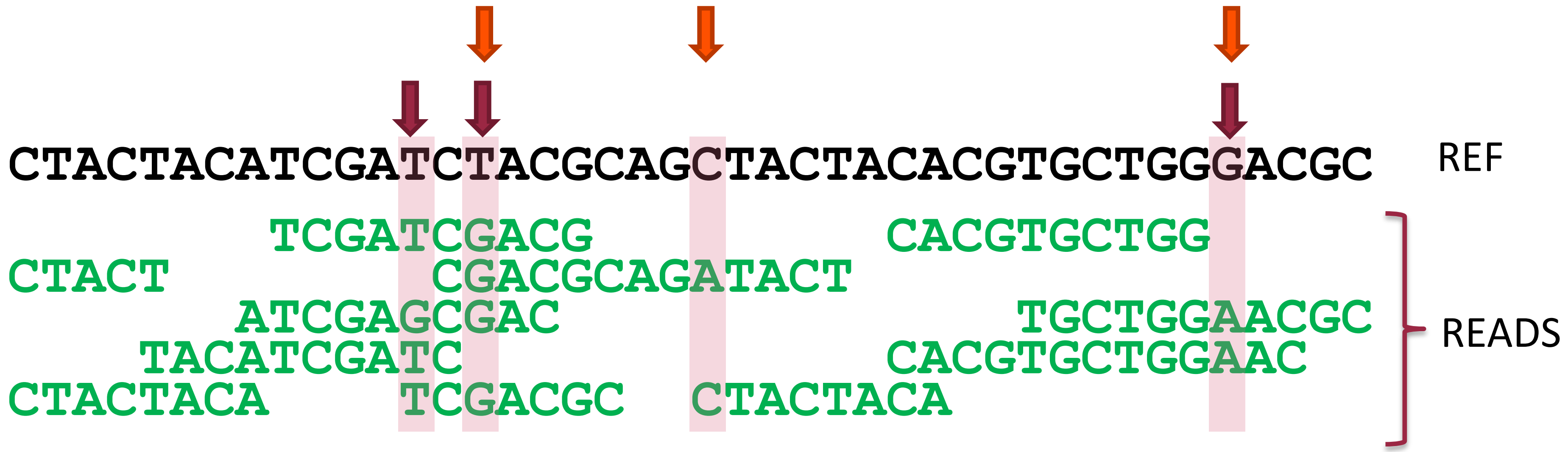


```
ACGATGTTGCGAGGCTCGTAAAGCGGTCAAACGATGACGTGCACGATACCGTGCATGACT

ACGATGTTGCGAGGC   TAAAGCGGT      ATGACGTGCACGATA     CATGACT
   ATGTTGCGAGGCTCG     CGGTC     CGATGACGCACGATA   TGCATGA
ACGATGTTGCGA        AAGCG          GACGTGCACGATACC    ATGACT
ACGATA   CGAGGCTCGTAAAGC        ACGATGACGCACG       CCGTGCAT
```

```
ACGATGTTGCGAGGCTCGTAAAGCGGTCAAACGATGACGTGCACGATACCGTGCATGACT

ACGATGTTGCGAGGC   TAAAGCGGT      ATGACGTGCACGATA     CATGACT
   ATGTTGCGAGGCTCG     CGGTC   CGATGAC--GCACGATA   TGCATGA
ACGATGTTGCGA        AAGCG          GACGTGCACGATACC    ATGACT
ACGATA   CGAGGCTCGTAAAGC      ACGATGAC--GCACG       CCGTGCAT
```

GATK

Variant calling



```
CTACTACATCGATCTACGCAGCTACTACACGTGCTGGGACGC    REF

        TCGATCGACG         CACGTGCTGG
CTACT        CGACGCAGATACT
       ATCGAGCGAC              TGCTGGAACGC    READS
    TACATCGATC             CACGTGCTGGAAC
CTACTACA    TCGACGC  CTACTACA
```

Is it a variant?

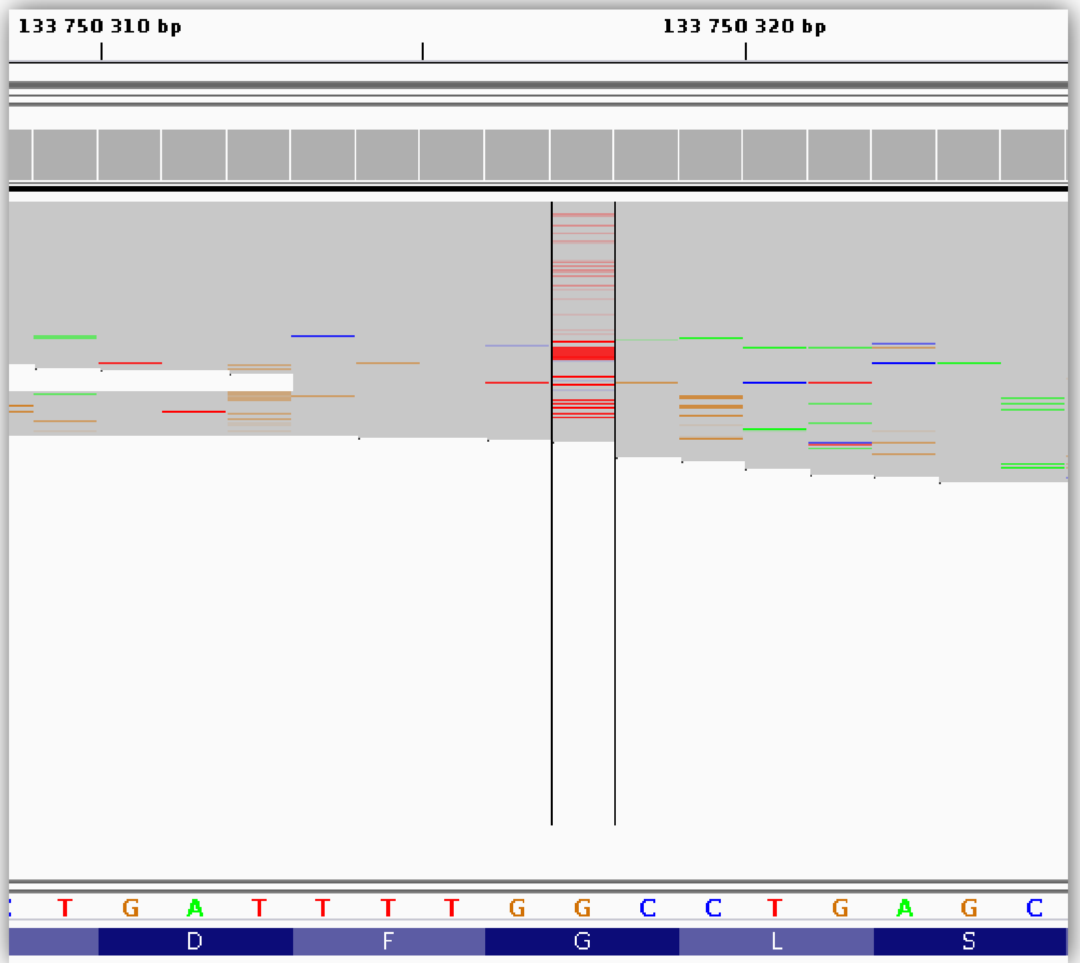
What is the most likely genotype?

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

VCF files

Variant calling

Amplicon, quite noisy

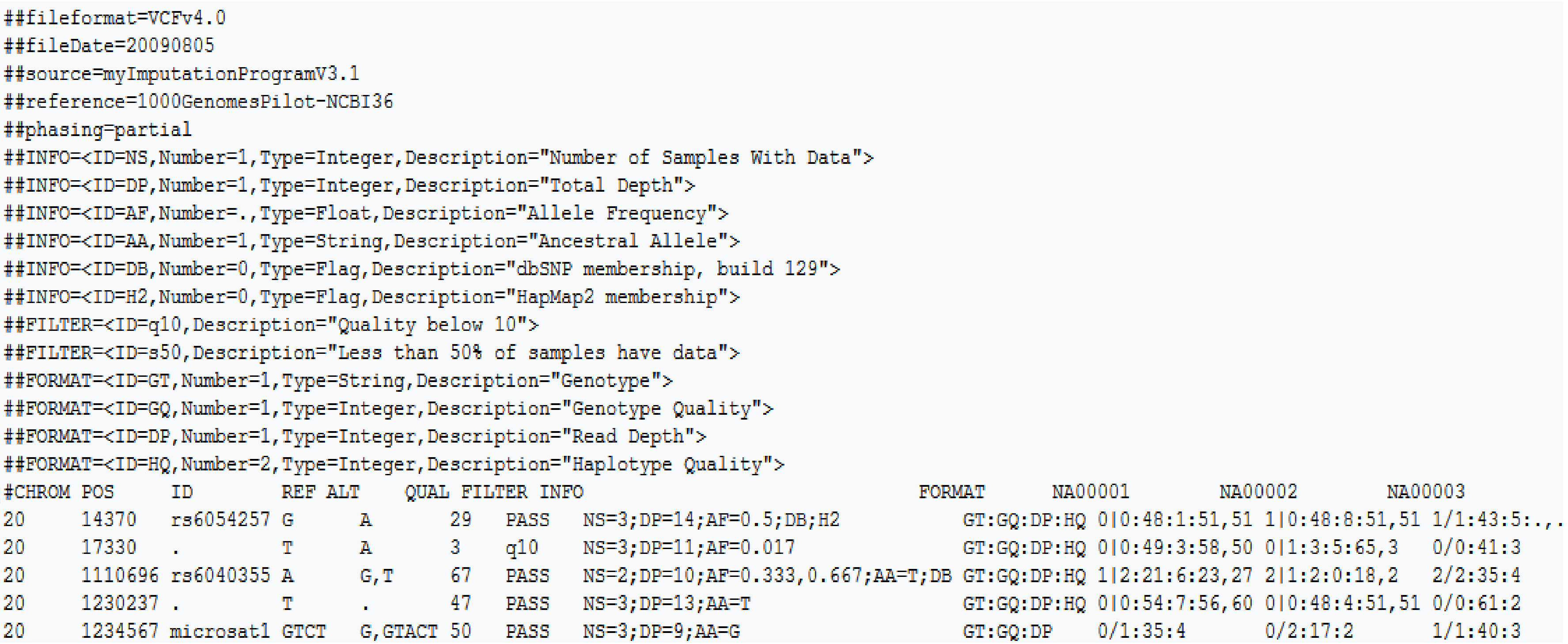


VCF format

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

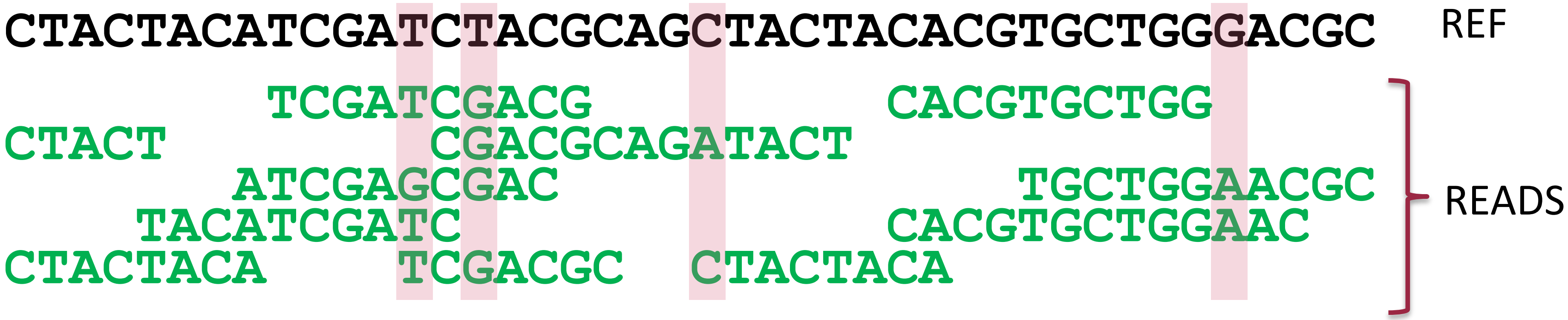
HEADER

BODY



```
##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">
#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
```

Variant annotation



```
CTACTACATCGATCTACGCAGCTACTACACGTGCTGGGACGC   REF
        TCGATCGACG         CACGTGCTGG
CTACT        CGACGCAGATACT
       ATCGAGCGAC              TGCTGGAACGC
    TACATCGATC             CACGTGCTGGAAC
CTACTACA    TCGACGC  CTACTACA
```

READS

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

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?

GA**A** -> GA**G** = E->E
GTT -> **C**TT = V->L
TG**G** -> TG**A** = W->X
TGA -> **C**GA = X->R

Is it a known SNP?

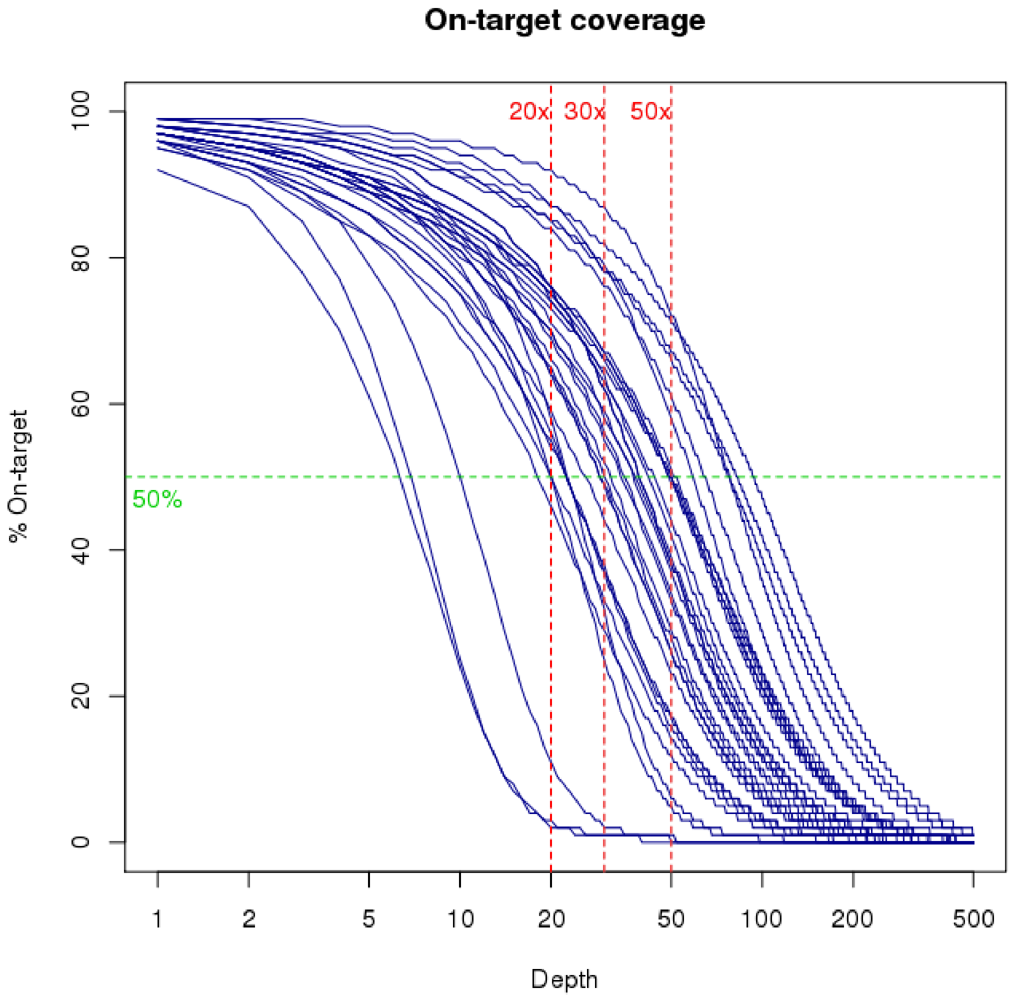
What impact does the AA change have?

Damaging, benign

CSV files

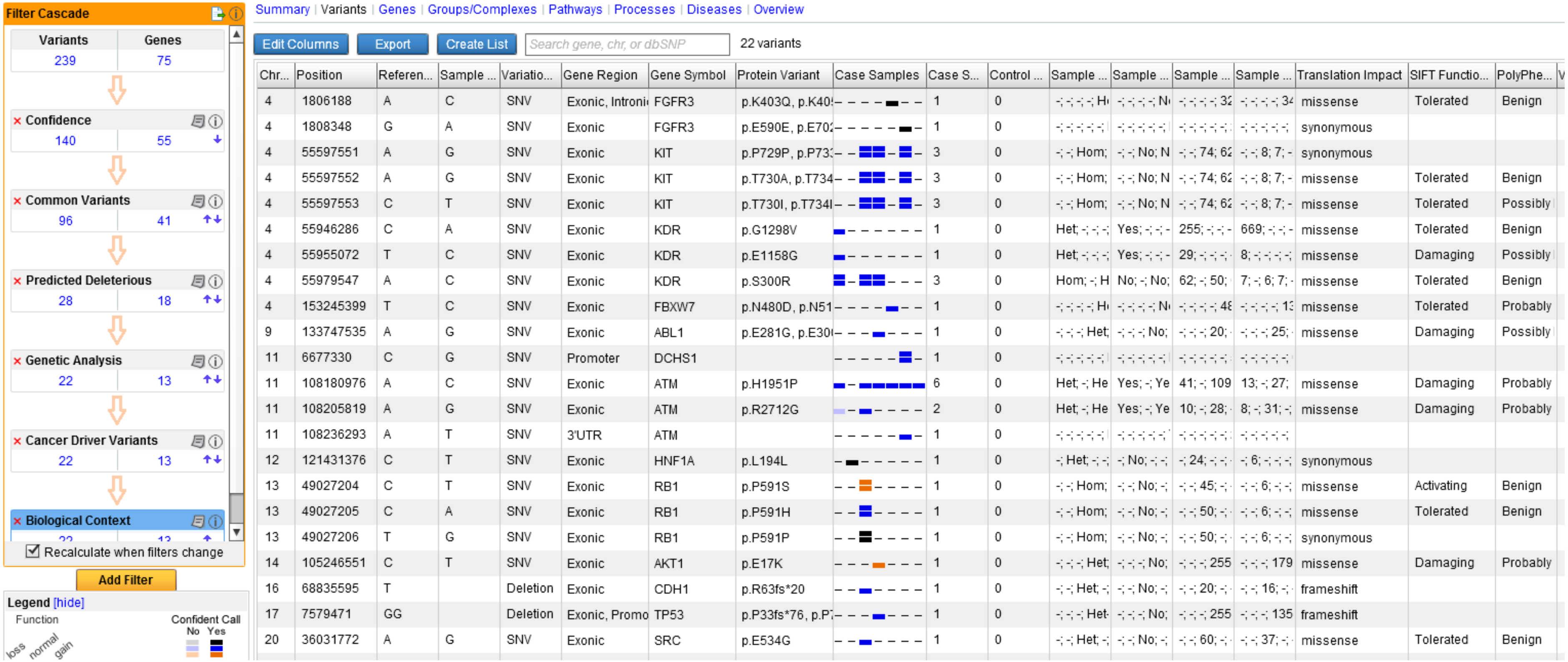
Coverage

	S1	S2	S3
Gene1	100	200	50
Gene2	50	0	50
Gene3	50	0	55
Gene4	10	10	55
Coverage	52.5X	52.5X	52.5X



Variant Analysis

Ingenutity Variant Analysis (IVA)



Filter Cascade

	Variants	Genes
	239	75
× Confidence	140	55 ↓
× Common Variants	96	41 ↑↓
× Predicted Deleterious	28	18 ↑↓
× Genetic Analysis	22	13 ↑↓
× Cancer Driver Variants	22	13 ↑↓
× Biological Context	22	13 ↑

☑ Recalculate when filters change

Add Filter

Legend [hide]

Function: loss, normal, gain

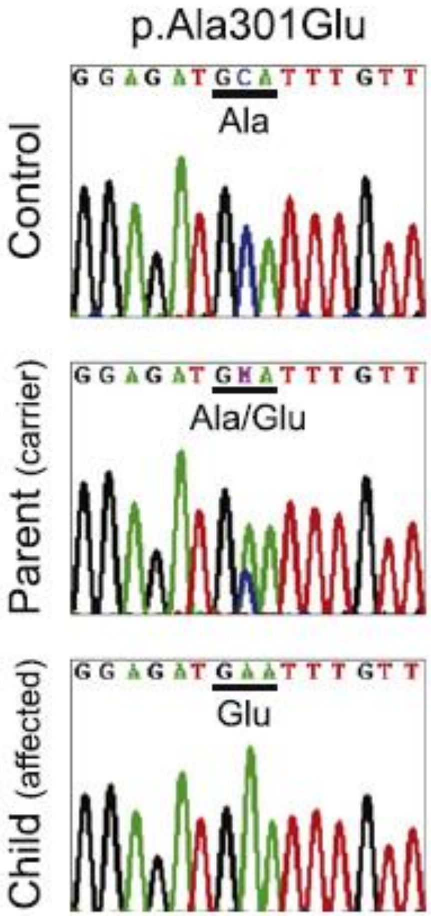
Confident Call	No	Yes
loss	▒	⬛
normal	▒	🟦
gain	▒	🟧

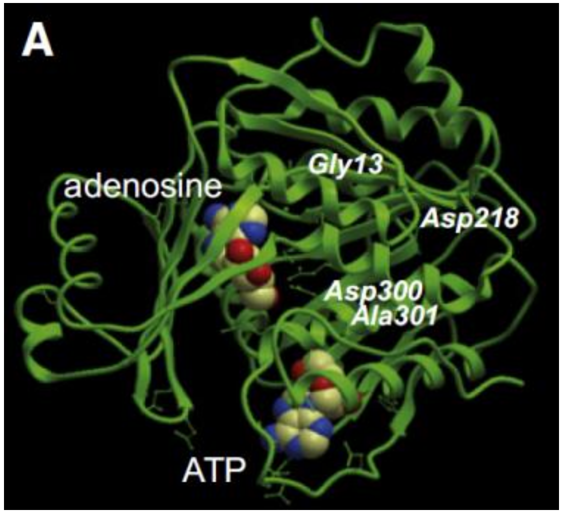
Summary | Variants | Genes | Groups/Complexes | Pathways | Processes | Diseases | Overview

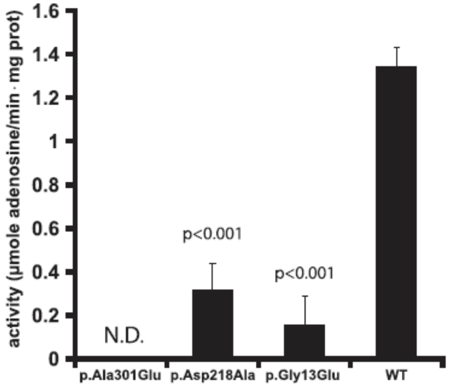
Edit Columns | Export | Create List | Search gene, chr, or dbSNP | 22 variants

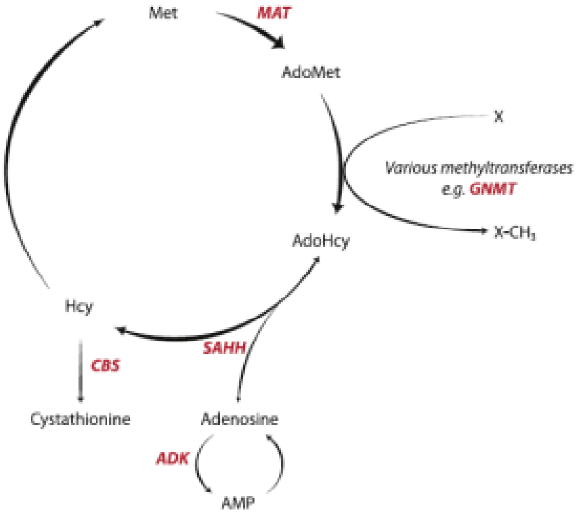
Chr...	Position	Referen...	Sample ...	Variatio...	Gene Region	Gene Symbol	Protein Variant	Case Samples	Case S...	Control ...	Sample ...	Sample ...	Sample ...	Sample ...	Translation Impact	SIFT Functio...	PolyPhe...	V
4	1806188	A	C	SNV	Exonic, Introni	FGFR3	p.K403Q, p.K40	– – – – ⬛ – –	1	0	-; -; -; -; H	-; -; -; -; N	-; -; -; -; 32	-; -; -; -; 34	missense	Tolerated	Benign	
4	1808348	G	A	SNV	Exonic	FGFR3	p.E590E, p.E70	– – – – – ⬛ –	1	0	-; -; -; -; -;	-; -; -; -; -;	-; -; -; -; -;	-; -; -; -; -;	synonymous			
4	55597551	A	G	SNV	Exonic	KIT	p.P729P, p.P73	– – 🟦🟦 – 🟦 –	3	0	-; -; Hom;	-; -; No; N	-; -; 74; 62	-; -; 8; 7; -	synonymous			
4	55597552	A	G	SNV	Exonic	KIT	p.T730A, p.T734	– – 🟦🟦 – 🟦 –	3	0	-; -; Hom;	-; -; No; N	-; -; 74; 62	-; -; 8; 7; -	missense	Tolerated	Benign	
4	55597553	C	T	SNV	Exonic	KIT	p.T730I, p.T734I	– – 🟦🟦 – 🟦 –	3	0	-; -; Hom;	-; -; No; N	-; -; 74; 62	-; -; 8; 7; -	missense	Tolerated	Possibly	
4	55946286	C	A	SNV	Exonic	KDR	p.G1298V	🟦 – – – – – – –	1	0	Het; -; -; -;	Yes; -; -; -	255; -; -; -	669; -; -; -	missense	Tolerated	Benign	
4	55955072	T	C	SNV	Exonic	KDR	p.E1158G	🟦 – – – – – – –	1	0	Het; -; -; -;	Yes; -; -; -	29; -; -; -;	8; -; -; -; -;	missense	Damaging	Possibly	
4	55979547	A	C	SNV	Exonic	KDR	p.S300R	🟦 – 🟦🟦 – – –	3	0	Hom; -; H	No; -; No;	62; -; 50;	7; -; 6; 7; -	missense	Tolerated	Benign	
4	153245399	T	C	SNV	Exonic	FBXW7	p.N480D, p.N51	– – – – 🟦 – –	1	0	-; -; -; -; H	-; -; -; -; N	-; -; -; -; 48	-; -; -; -; 13	missense	Tolerated	Probably	
9	133747535	A	G	SNV	Exonic	ABL1	p.E281G, p.E30	– – – 🟦 – – –	1	0	-; -; -; Het;	-; -; -; No;	-; -; -; 20;	-; -; -; 25;	missense	Damaging	Possibly	
11	6677330	C	G	SNV	Promoter	DCHS1		– – – – – 🟦 –	1	0	-; -; -; -; -;	-; -; -; -; -;	-; -; -; -; -;	-; -; -; -; -;				
11	108180976	A	C	SNV	Exonic	ATM	p.H1951P	🟦 – 🟦🟦🟦🟦🟦	6	0	Het; -; He	Yes; -; Ye	41; -; 109	13; -; 27;	missense	Damaging	Probably	
11	108205819	A	G	SNV	Exonic	ATM	p.R2712G	▒ – 🟦 – – – –	2	0	Het; -; He	Yes; -; Ye	10; -; 28;	8; -; 31; -;	missense	Damaging	Probably	
11	108236293	A	T	SNV	3'UTR	ATM		– – – – – 🟦 –	1	0	-; -; -; -; -;	-; -; -; -; -;	-; -; -; -; -;	-; -; -; -; -;				
12	121431376	C	T	SNV	Exonic	HNF1A	p.L194L	– ⬛ – – – – –	1	0	-; Het; -; -;	-; No; -; -;	-; 24; -; -;	-; 6; -; -; -;	synonymous			
13	49027204	C	T	SNV	Exonic	RB1	p.P591S	– – 🟧 – – – –	1	0	-; -; Hom;	-; -; No; -;	-; -; 45; -;	-; -; 6; -; -;	missense	Activating	Benign	
13	49027205	C	A	SNV	Exonic	RB1	p.P591H	– – 🟦 – – – –	1	0	-; -; Hom;	-; -; No; -;	-; -; 50; -;	-; -; 6; -; -;	missense	Tolerated	Benign	
13	49027206	T	G	SNV	Exonic	RB1	p.P591P	– – ⬛ – – – –	1	0	-; -; Hom;	-; -; No; -;	-; -; 50; -;	-; -; 6; -; -;	synonymous			
14	105246551	C	T	SNV	Exonic	AKT1	p.E17K	– – – 🟧 – – –	1	0	-; -; -; Het;	-; -; -; No;	-; -; -; 255	-; -; -; 179	missense	Damaging	Probably	
16	68835595	T		Deletion	Exonic	CDH1	p.R63fs*20	– – 🟦 – – – –	1	0	-; -; Het; -;	-; -; No; -;	-; -; 20; -;	-; -; 16; -;	frameshift			
17	7579471	GG		Deletion	Exonic, Promo	TP53	p.P33fs*76, p.P7	– – – 🟦 – – –	1	0	-; -; -; Het	-; -; -; No;	-; -; -; 255	-; -; -; 135	frameshift			
20	36031772	A	G	SNV	Exonic	SRC	p.E534G	– – 🟦 – – – –	1	0	-; -; Het; -;	-; -; No; -;	-; -; 60; -;	-; -; 37; -;	missense	Tolerated	Benign	

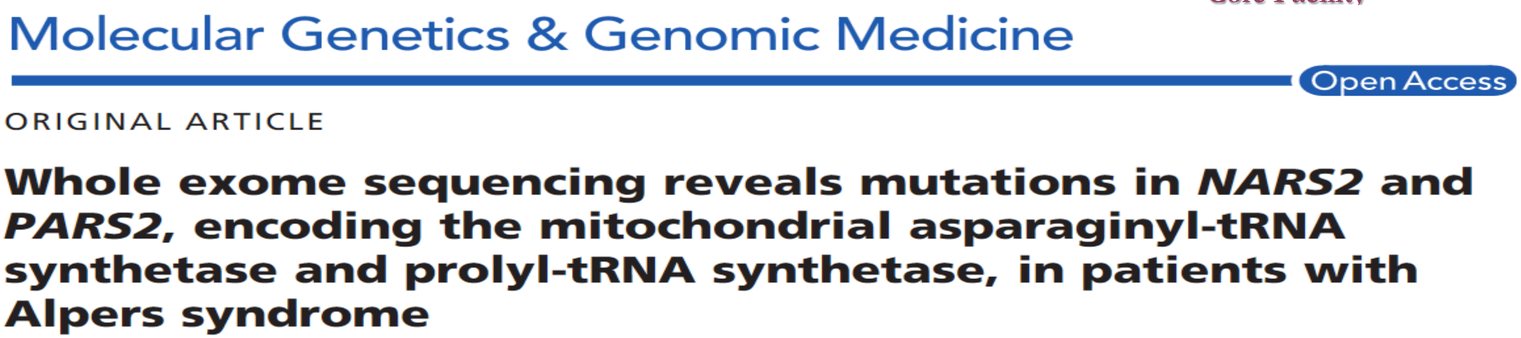
Identification of disease causing mutation











Molecular Genetics & Genomic Medicine

Open Access

ORIGINAL ARTICLE

Whole exome sequencing reveals mutations in *NARS2* and *PARS2*, encoding the mitochondrial asparaginyl-tRNA synthetase and prolyl-tRNA synthetase, in patients with Alpers syndrome

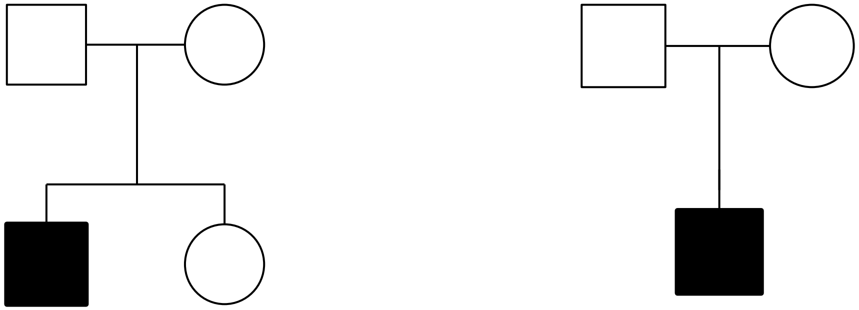
Alpers syndrome: progressive neurodegenerative dissorder

POLG1 – Alpers Huttenlocher

FARS2 – encoding enzyme to charge mt tRNA with Phe

19 patients: 6 had POLG1 mutations

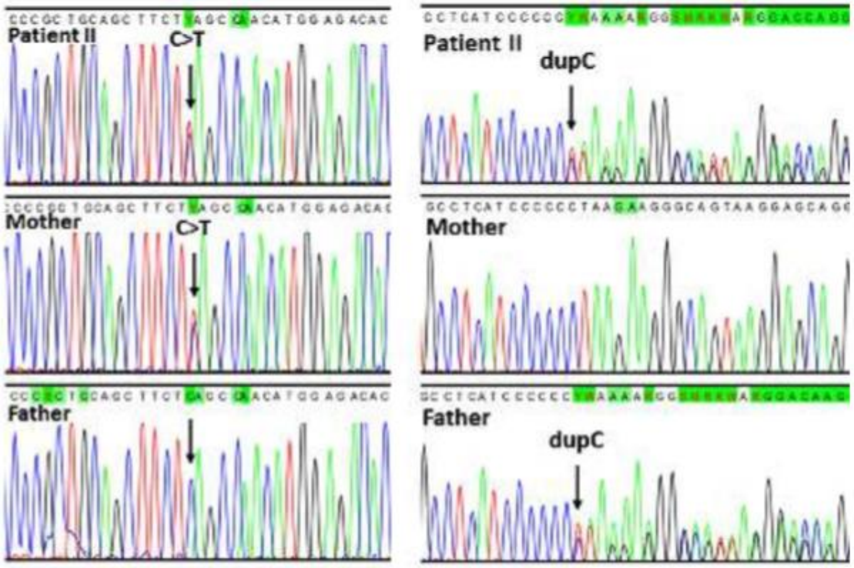
For this study:



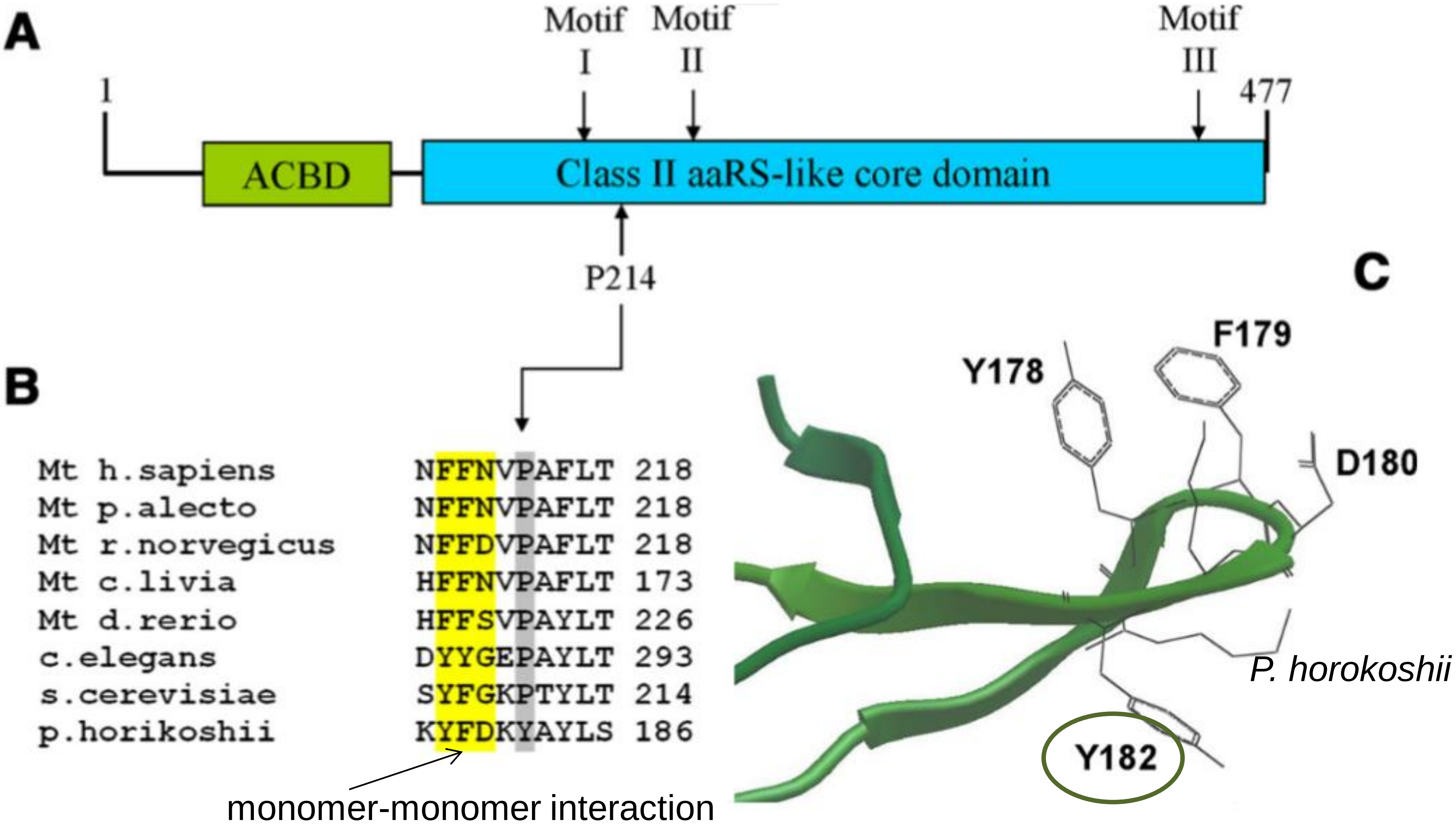
Exome sequencing

	Patient I		Patient II	
	Variants	Genes	Variants	Genes
Total	124,631	15,978	129,098	16,015
Genes encoding mitochondrial protein	1698	671	1882	681
Allele frequency <3%	98	94	100	86
Predicted deleterious	32	27	18	18
Recessive pattern of inheritance	1	1	2	1

Mutations in *PARS2* (Pro) and *NARS2* (Asn)



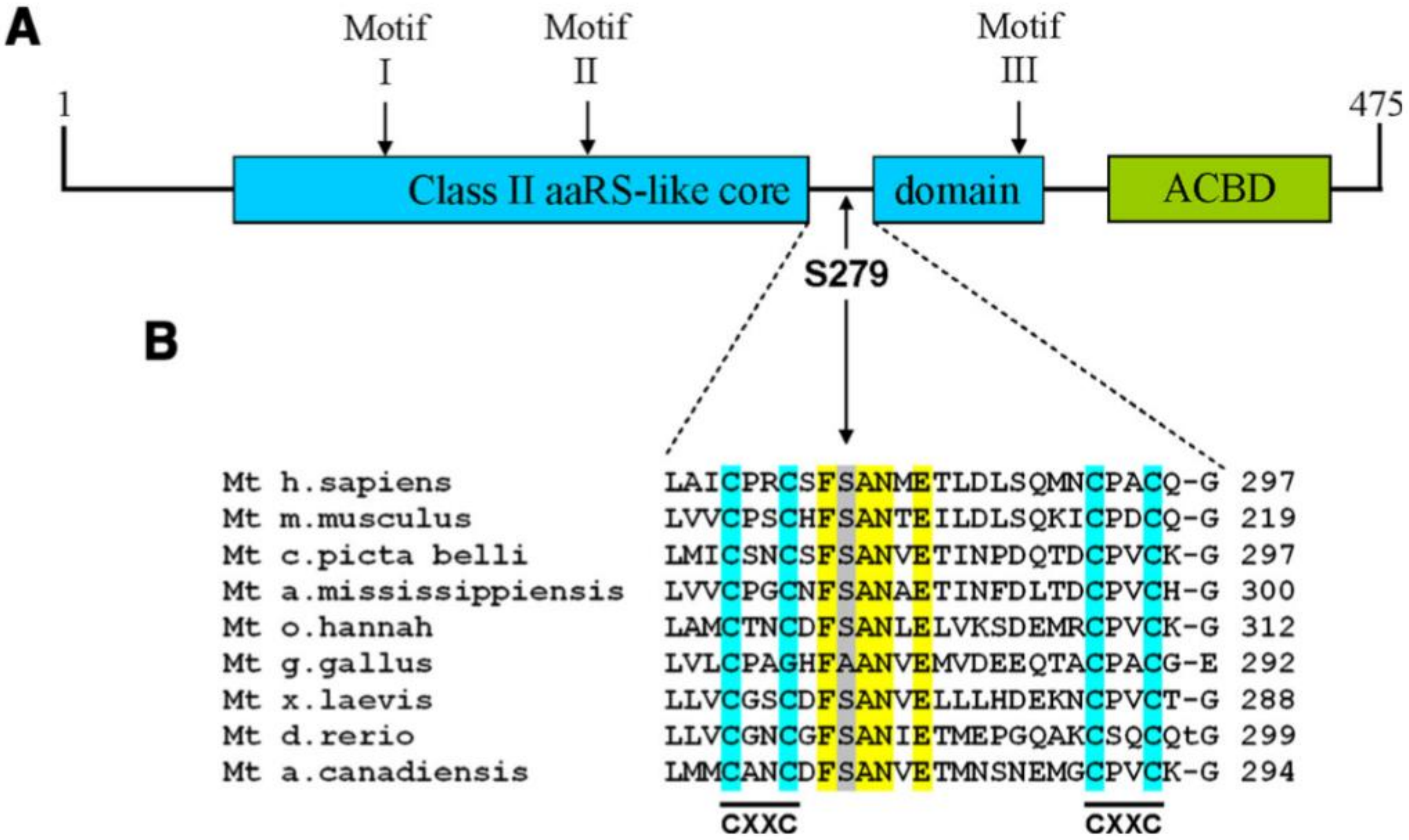
NARS


A
Motif I
Motif II
Motif III
1
477
ACBD
Class II aaRS-like core domain
P214
B
Mt h.sapiens NFFNVPAFLT 218
Mt p.alecto NFFNVPAFLT 218
Mt r.norvegicus NFFDVPAFLT 218
Mt c.livia HFFNVPAFLT 173
Mt d.rerio HFFSVPAYLT 226
c.elegans DYYGEPAYLT 293
s.cerevisiae SYFGKPTYLT 214
p.horikoshii KYFDKYAYLS 186
C
Y178
F179
D180
Y182

P. horokoshii

monomer-monomer interaction

PARS



```
Mt h.sapiens                LAICPRCSFSANMETLDLSQMNCPACQ-G 297
Mt m.musculus               LVVCPSCHFSANTEILDLSQKICPDCQ-G 219
Mt c.picta belli            LMICSNCSFSANVETINPDQTDCPVCK-G 297
Mt a.mississippiensis       LVVCPGCNFSANAETINFDLTDCPVCH-G 300
Mt o.hannah                 LAMCTNCDFSANLELVKSDEMRCPVCK-G 312
Mt g.gallus                 LVLCPAGHFAANVEMVDEEQTACPACG-E 292
Mt x.laevis                 LLVCGSCDFSANVELLLHDEKNCPVCT-G 288
Mt d.rerio                  LLVCGNCGFSANIETMEPGQAKCSQCQtG 299
Mt a.canadiensis            LMMCANCDFSANVETMNSNEMGCPVCK-G 294
```