NATIONAL GENOMICS INFRASTRUCTURE

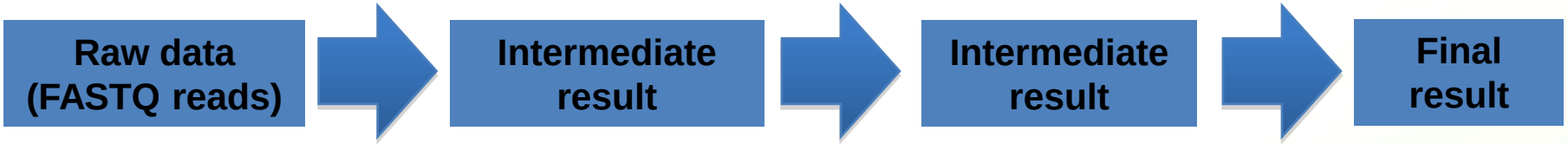
Next Generation Sequencing and Bioinformatics Analysis Pipelines

Adam Ameur
National Genomics Infrastructure
SciLifeLab Uppsala
adam.ameur@igp.uu.se

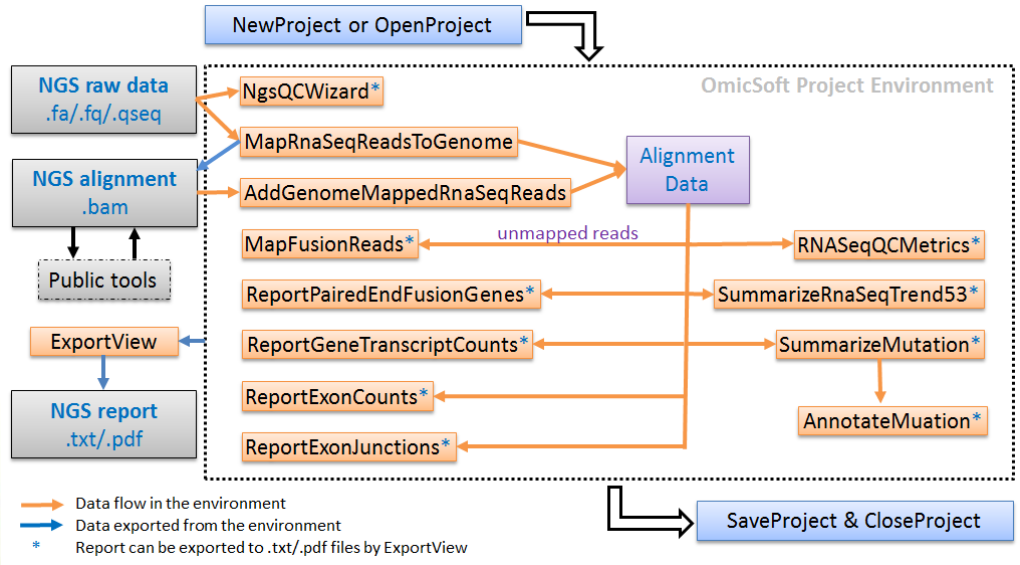
Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet · SciLifeLab

What is an analysis pipeline?

- Basically just a number of steps to analyze data

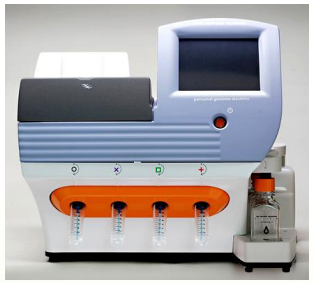


- Pipelines can be simple or very complex…



Today's lecture

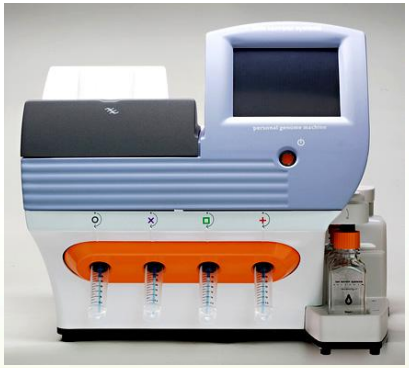
- Sequencing instruments and 'standard' pipelines
  - IonTorrent/PacificBiosciences
- In-house bioinformatics pipelines, some examples
- News and future plans



Ion Torrent - PGM/Proton

- The Ion Torrent System
  - 6 instruments available in Uppsala, early access users
  - Two instruments: PGM and Proton
  - For small scale (PGM) and large scale sequencing (Proton)
  - Rapid sequencing (run time ~ 2-4 hours)
  - Measures changes in pH
  - Sequencing on a chip

Personal Genome Machine (PGM)



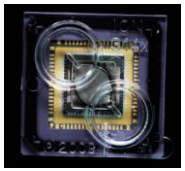
Ion Proton

Ion Torrent output

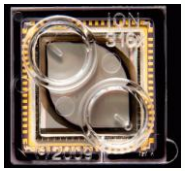
- Ion Torrent throughput
  ~ from 10Mb to >10Gb, depending on the chip

2 human exomes (PI chip)
2 human transcriptomes
1 human genome = 6 PI chips

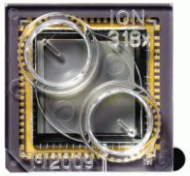
314 (PGM)



316 (PGM)



318 (PGM)



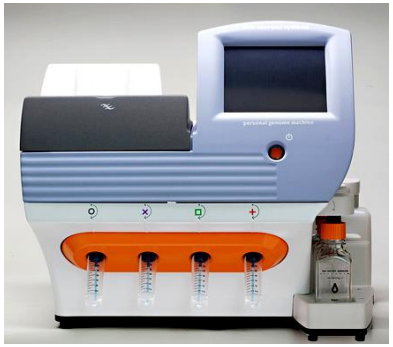
PI (Proton)

- Read lengths: 400bp (PGM), 200bp (Proton)

- Output file format: FASTQ

- What can we use Ion Torrent for?
  – Anything, except perhaps very large genomes

Ion Torrent analysis workflow



Torrent Server

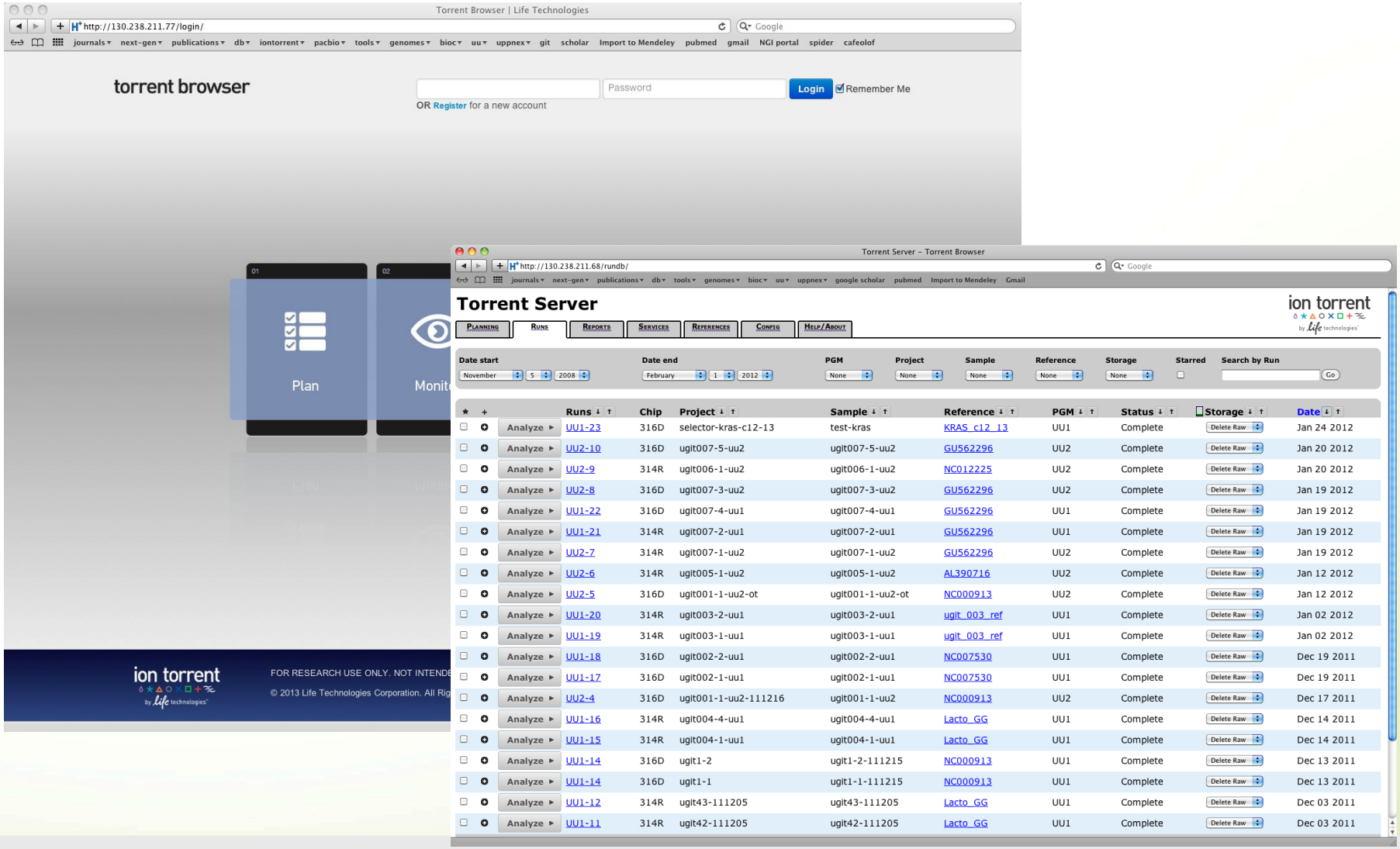
life technologies | Semiconductor Sequencing for Life™ ion torrent

.fastq
.bam
.fasta

UPPNEX

Downstream analysis

Torrent Suite Software



Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet · SciLifeLab

Torrent Suite Software Analysis

- Plug-ins within the Torrent Suite Software
  - Alignment
    - TMAP: Specifically developed for Ion Torrent data
  - Variant Caller
    - SNP/Indel detection
  - Assembler
    - MIRA
  - AmpliSeq analysis (Human Exomes and Transcriptomes)
    - SNP/Indel detection in amplicon-seq data
    - Expression analysis by AmpliSeq
  - …

- Analyses are started automatically when run is complete

Pacific Biosciences

- Pacific Biosciences
  - Installed summer 2013
  - Single molecule sequencing
  - Very long read lengths (up to 40 kb)
  - Rapid sequencing
  - Can detect base modifications (i.e. methylation)
  - Relatively low throughput

PacBio output

- PacBio throughput
  ~ 1 Gb/SMRT cell

~1 bacterial genome
~1 bacterial transcriptome
1 human genome = 100 SMRT cells?

- PacBio read lengths: 500bp-40kb

- Output file format: FASTQ

- What can we use PacBio for?
  – Anything, except really large genomes

PacBio analysis workflow

In-house PacBio cluster

.fastq
.bam
.fasta

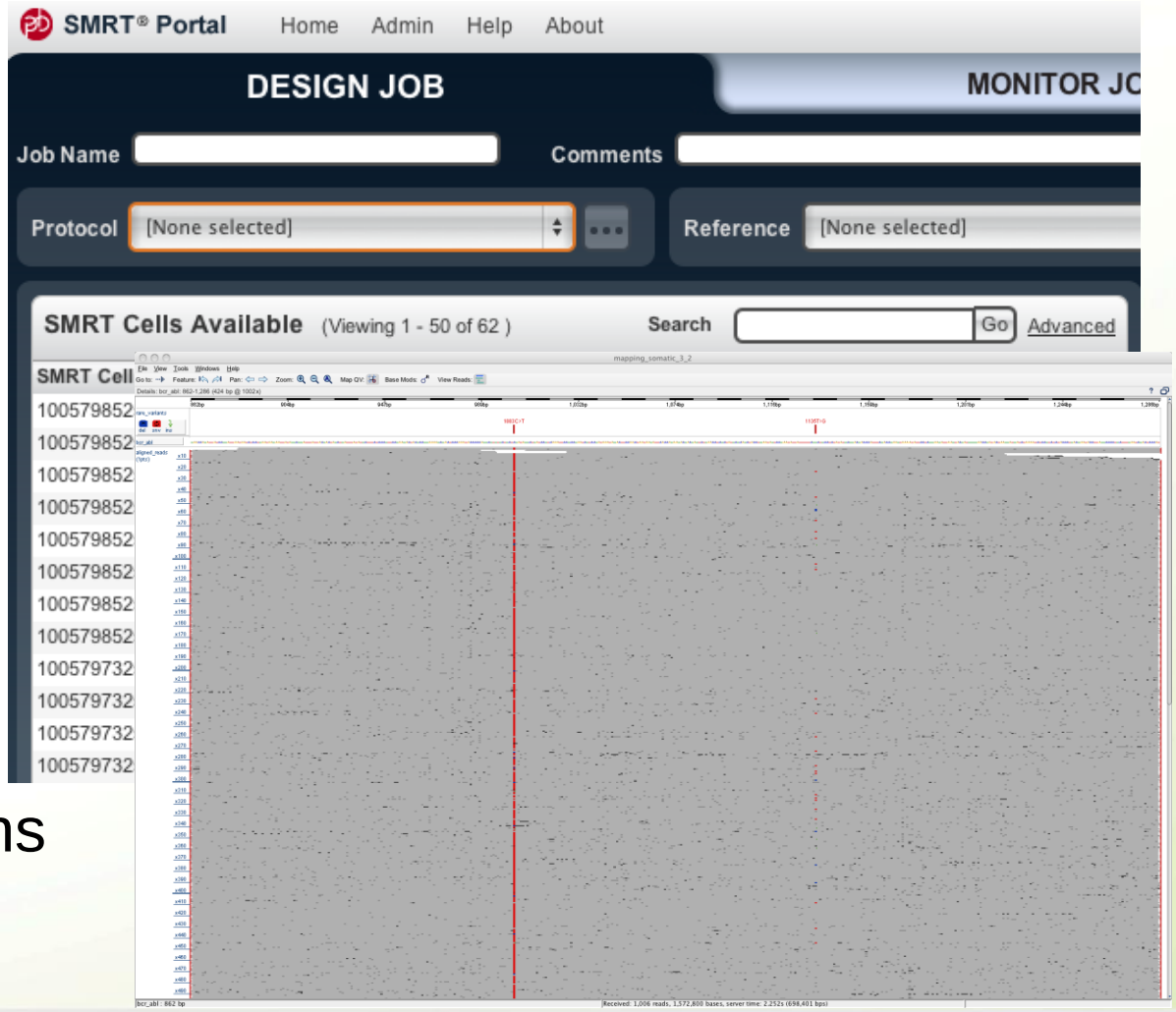
UPPNEX

Downstream analysis

SMRT analysis portal

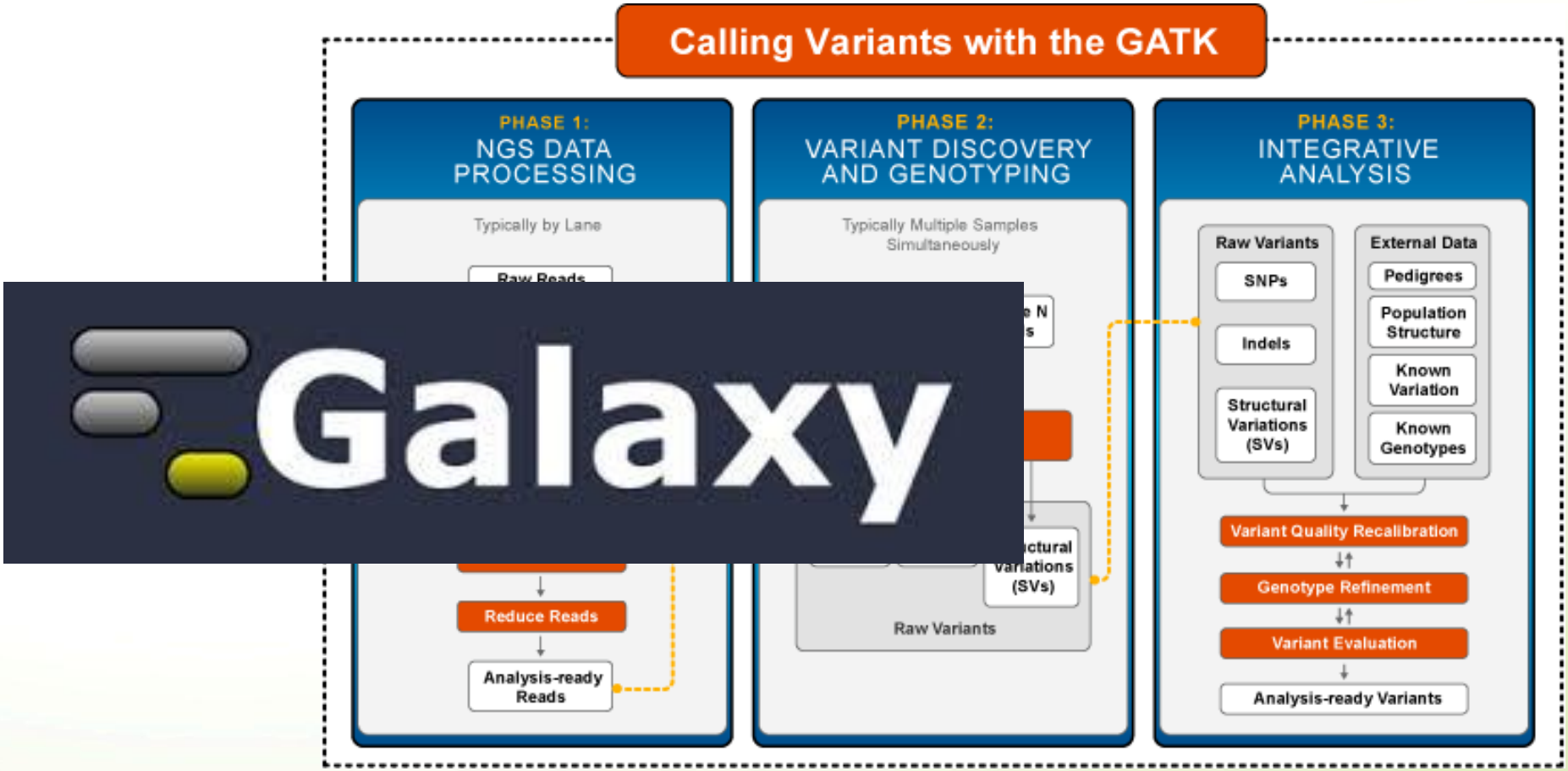
SMRT analysis pipelines

- Mapping
- Variant calling
- Assembly
- Scaffolding
- Base modifications



What about Illumina?

- There are many different pipelines for Illumina…



In-house development of pipelines

- The standard analysis pipelines are nice…

  … but sometimes we need to do own developments

  … or adapt the pipelines to our specific applications

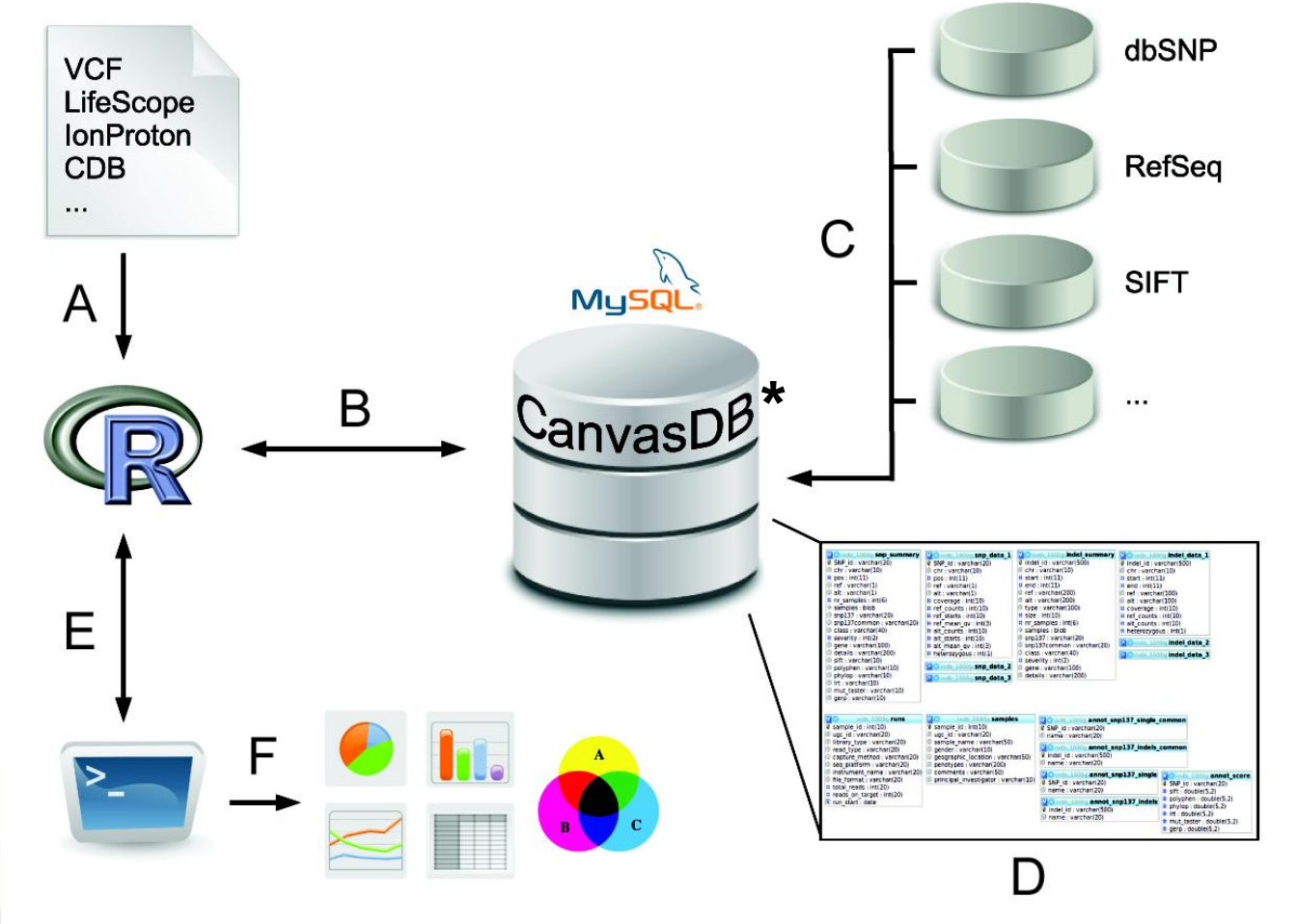
- Some examples of in-house developments:

I. Building a local variant database (WES/WGS)

II. Assembly of genomes using long reads

III. Clinical sequencing – Leukemia Diagnostics

Example I:
Computational infrastructure for exome-seq data



Background: exome-seq

- Main application of exome-seq
  - Find disease causing mutations in humans

- Advantages
  - Allows investigate all protein coding sequences
  - Possible to detect both SNPs and small indels
  - Low cost (compared to WGS)
  - Possible to multiplex several exomes in one run
  - Standardized work flow for data analysis

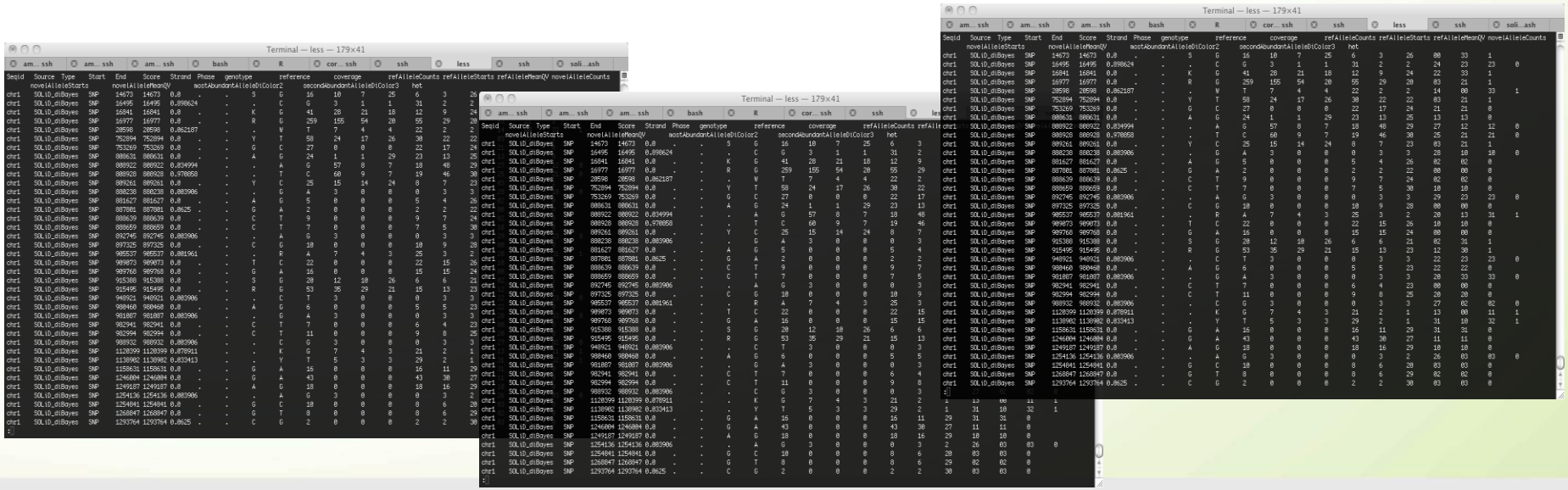
- Disadvantage
  - All genetic variants outside of exons are missed (~98%)

Exome-seq throughput

- We are producing a lot of exome-seq data
  - 4-6 exomes/day on Ion Proton
  - In each exome we detect
    - Over 50,000 SNPs
    - About 2000 small indels

=> Over 1 million variants/run!

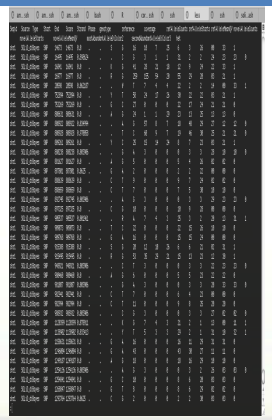
- In plain text files

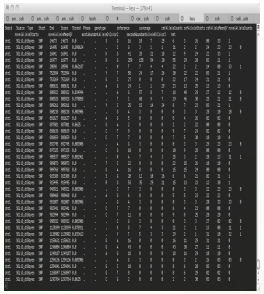


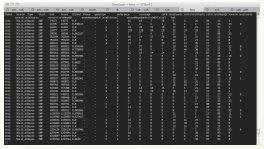
How to analyze this?

- Traditional analysis - A lot of filtering!
  - Typical filters
    - Focus on rare SNPs (not present in dbSNP)
    - Remove FPs (by filtering against other exomes)
    - Effect on protein: non-synonymous, stop-gain etc
    - Heterozygous/homozygous
  - This analysis can be automated (more or less)

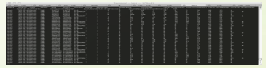
Start:
All identified SNPs







Result:
A few candidate causative SNP(s)!

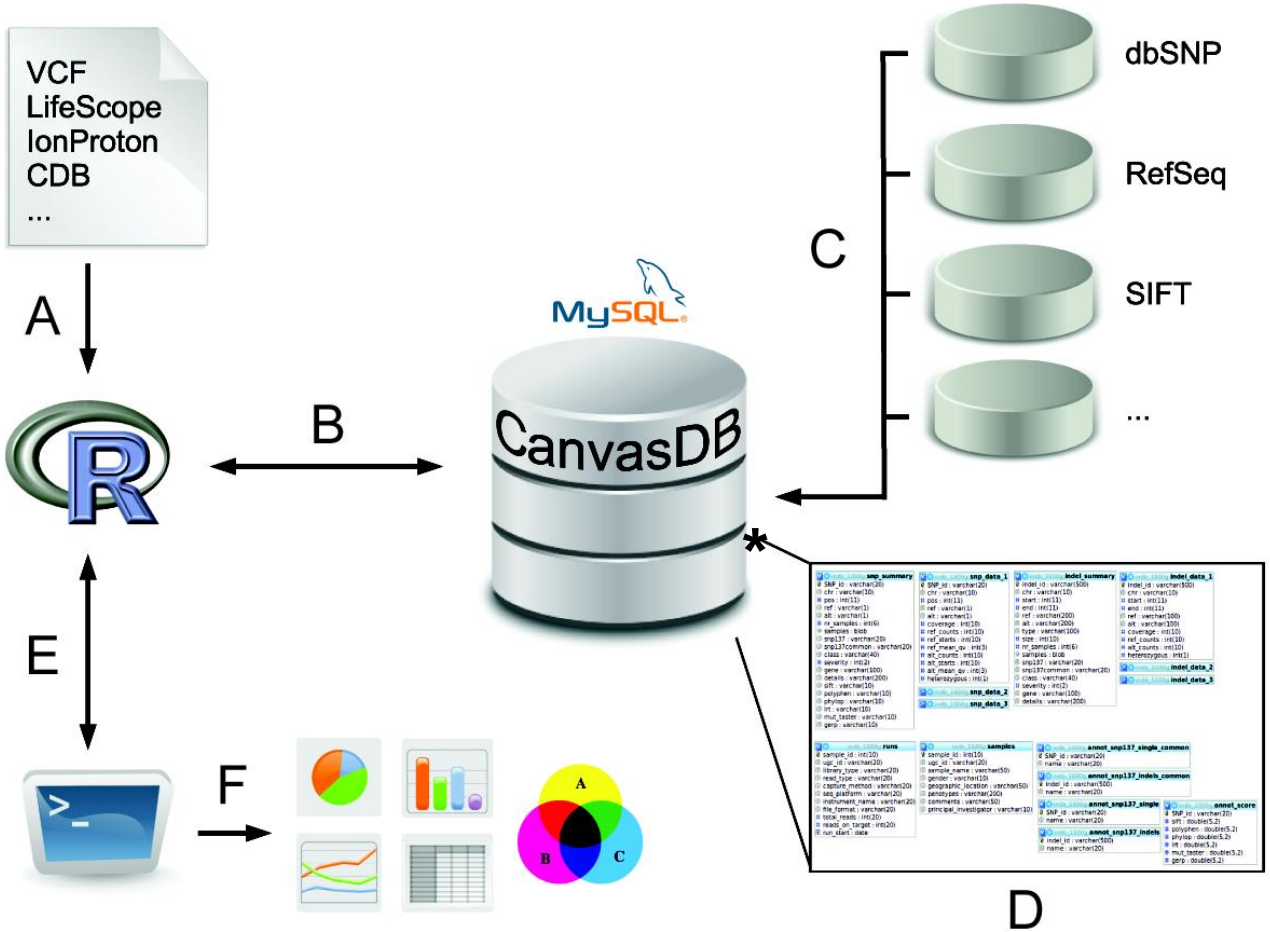


Why is this not optimal?

- Drawbacks
  - Work on one sample at time
    - Difficult to compare between samples
  - Takes time to re-run analysis
    - When using different parameters
  - No standardized storage of detected SNPs/indels
    - Difficult to handle 100s of samples

- Better solution
  - A database oriented system
    - Both for data storage and filtering analyses

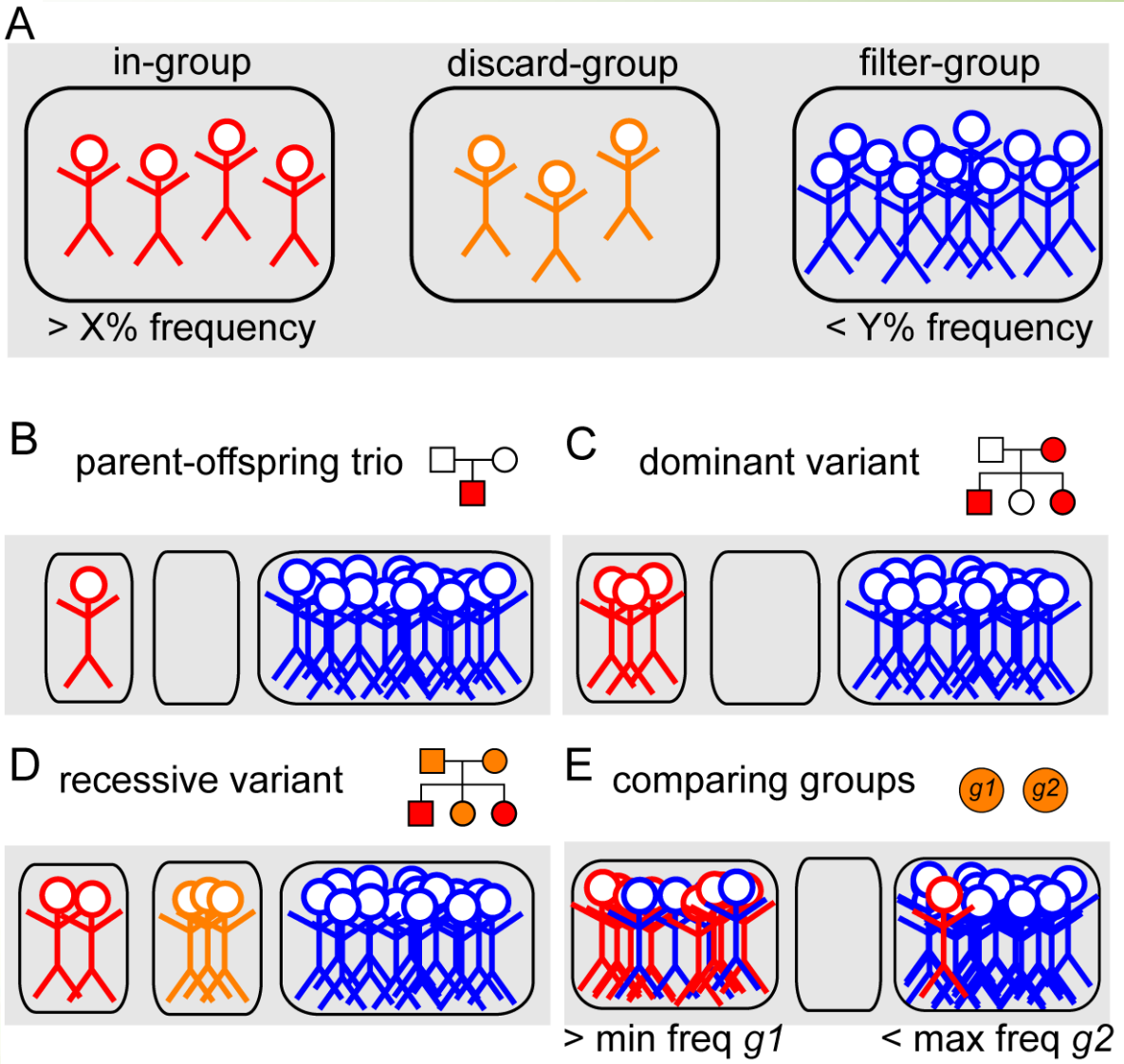
Analysis: In-house variant database



***CANdidate Variant Analysis System and Data Base**

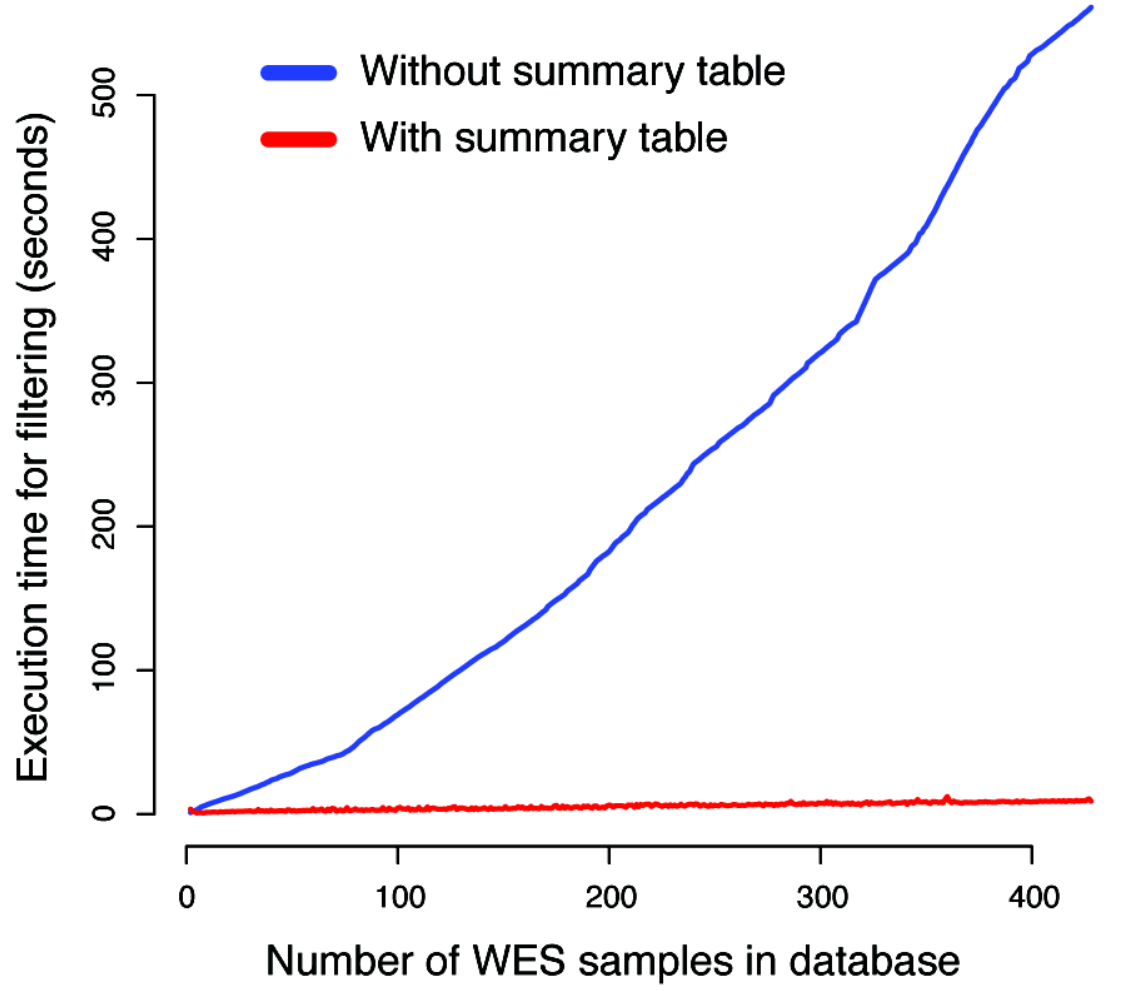
Ameur et al., Database Journal, 2014

CanvasDB - Filtering



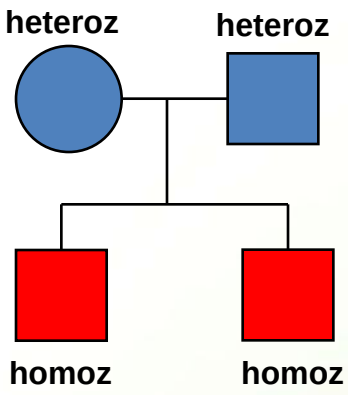
CanvasDB - Filtering speed

- Rapid variant filtering, also for large databases



A recent exome-seq project

- Hearing loss: 2 affected brothers
  - Likely a rare, recessive disease
  - => Shared homozygous SNPs/indels

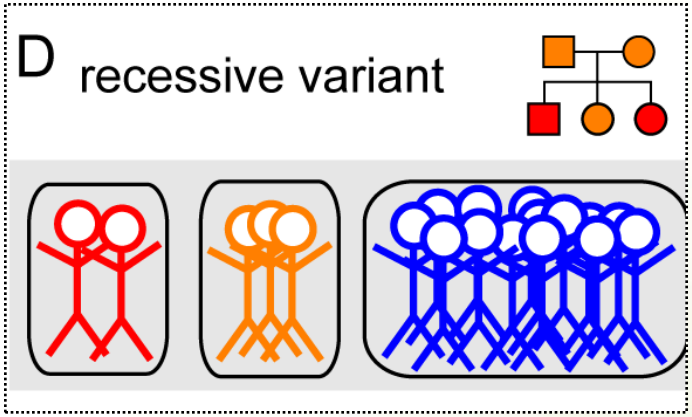


- Sequencing strategy
  - TargetSeq exome capture
  - One sample per PI chip

nr reads	(% mapped)	76M-89M (97%)
mapped reads	(% on target)	73M-88M (83%)
SNPs	(% in dbSNP)	85k-93k (93%)
Indels	(% in dbSNP)	5k-6k (48%)

Filtering analysis

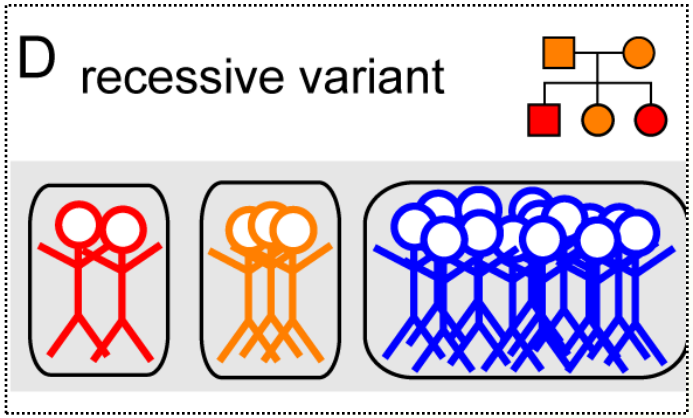
- *CanvasDB* filtering for a variant that is…
  - rare
    - at most in 1% of ~700 exomes
  - shared
    - found in both brothers
  - homozygous
    - in brothers, but in no other samples
  - deleterious
    - non-synonymous, frameshift, stop-gain, splicing, etc..




```
> cand <- filterRecessive(c("up_001_1","up_001_2"), outfile="cand.txt")
Total time for filtering: 27.012s
```

Filtering results

- Homozygous candidates
  - 2 SNPs
    - stop-gain in *STRC*
    - non-synonymous in *PCNT*
  - 0 indels
- Compound heterozygous candidates (lower priority)
  - in 15 genes



sample_name	class	chr	pos	ref	alt	snp137	gene	ref_counts	alt_counts
up_001_1	stopgain	chr15	43896948	G	A	rs144948296	STRC	3	58
up_001_2	stopgain	chr15	43896948	G	A	rs144948296	STRC	5	55
up_001_1	nonsynonymous	chr21	47808772	G	A	rs35044802	PCNT	0	21
up_001_2	nonsynonymous	chr21	47808772	G	A	rs35044802	PCNT	1	14

=> Filtering is fast and gives a short candidate list!

STRC - a candidate gene

STRC

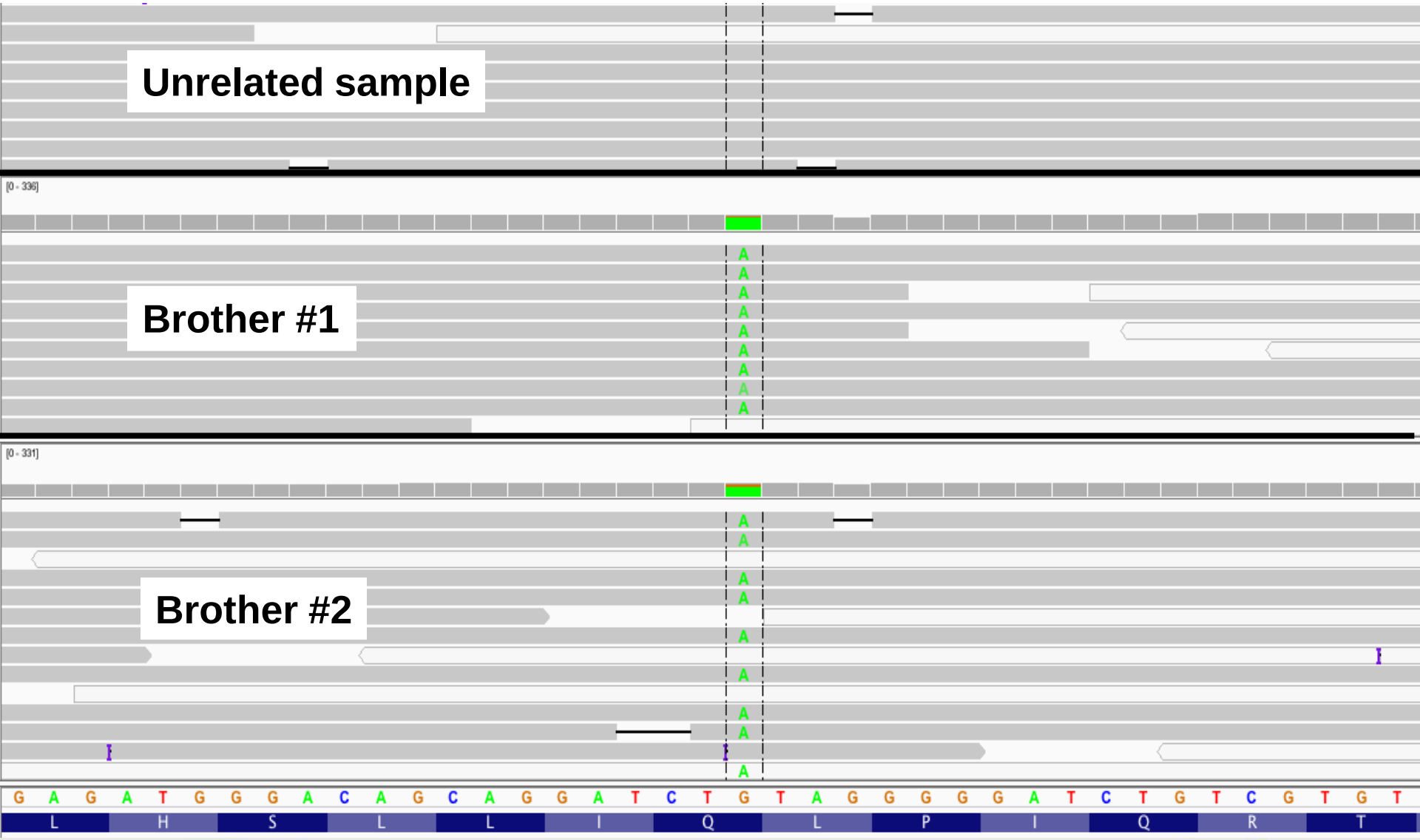
From Wikipedia, the free encyclopedia

Stereocilin is a protein that in humans is encoded by the *STRC* gene.[1][2][3]

This gene encodes a protein that is associated with the hair bundle of the sensory hair cells in the inner ear. The hair bundle is composed of stiff microvilli called stereocilia and is involved with mechanoreception of sound waves. This gene is part of a tandem duplication on chromosome 15; the second copy is a pseudogene. Mutations in this gene cause autosomal recessive non-syndromic deafness.[3]

=> Stop-gain in STRC is likely to cause hearing loss!

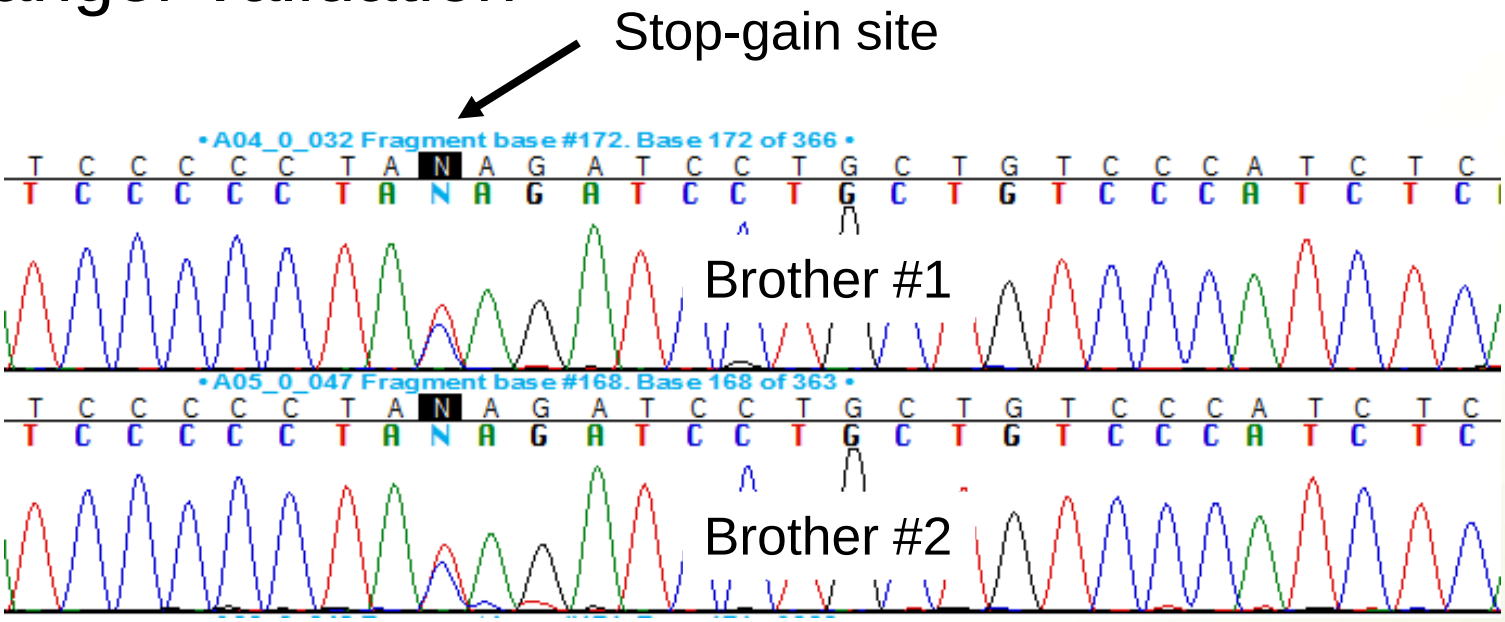
IGV visualization: Stop gain in STRC



Karolinska Institutet
Stockholms universitet
UPPSALA UNIVERSITET
SciLifeLab

STRC, validation by Sanger

- Sanger validation



- Does not seem to be homozygous..
  - Explanation: difficult to sequence STRC by Sanger
    - Pseudo-gene with very high similarity
- New validation showed mutation is homozygous!!

CanvasDB – some success stories

Solved cases, exome-seq - Niklas Dahl/Joakim Klar

Neuromuscular disorder	*NMD11*
Artrogryfosis	*SKD36*
Lipodystrophy	*ACR1*
Achondroplasia	*ACD2*
Ectodermal dysplasia	*ED21*
Achondroplasia	*ACD9*
Ectodermal dysplasia	*ED1*
Arythroderma	*AV1*
Ichthyosis	*SD12*
Muscular dystrophy	*DMD7*
Neuromuscular disorder	*NMD8*
Welanders myopathy (D)	*W*
Skeletal dysplasia	*SKD21*
Visceral myopathy (D)	*D:5156*
Ataxia telangiectasia	*MR67*
Exostosis	*SKD13*
Alopecia	*AP43*
Epidermolysis bullosa	*SD14*
Hearing loss	*D:9652*

Success rate >80% for recent Proton projects!

CanvasDB - Availability

- CanvasDB system now freely available on GitHub!

Installation of the CanvasDB system

This section describes how to download and install CanvasDB on your local computer. Make sure that MySQL, R and ANNOVAR are running on your computer before starting the installation.

Step 1. Download code from github

```
$ git clone https://github.com/UppsalaGenomeCenter/CanvasDB.git
$ cd CanvasDB
```

Step 2. Set the current path to 'rootDir' in canvasDB.R

Next Step: Whole Genome Sequencing

- New instruments at SciLifeLab for human WGS…

Capacity of HiSeq Xten: 320 whole human genomes/week!!!

- More work on pipelines and databases needed!!!

Example II:

Assembly of genomes using Pacific Biosciences

Genome assembly using NGS

- Short-read *de novo* assembly by NGS
  - Requires mate-pair sequences
    - Ideally with different insert sizes
  - Complicated analysis
    - Assembly, scaffolding, finishing
    - Maybe even some manual steps

  => Rather expensive and time consuming

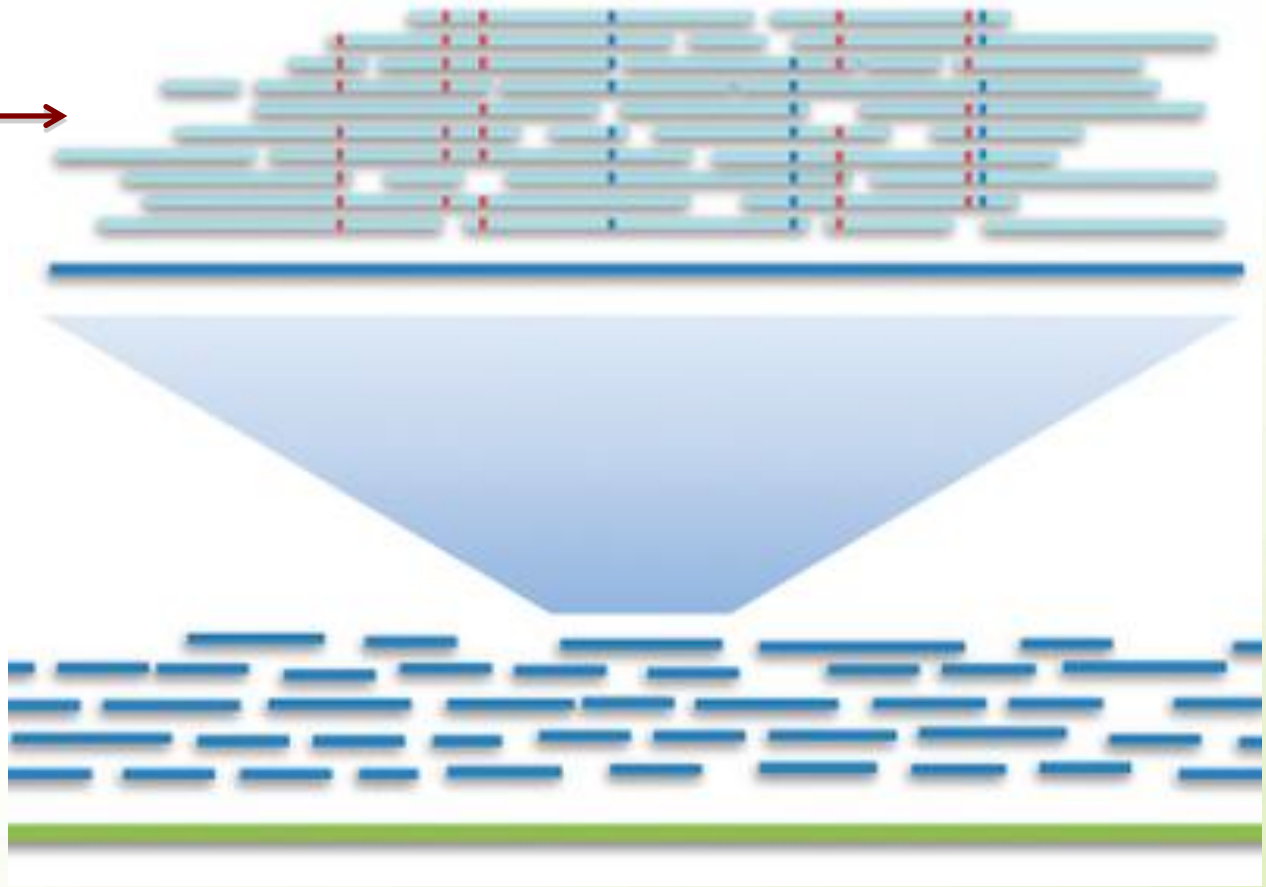
- Long reads really makes a difference!!
  - We can assemble genomes using PacBio data only!

HGAP *de novo* assembly

- HGAP uses both long and shorter reads

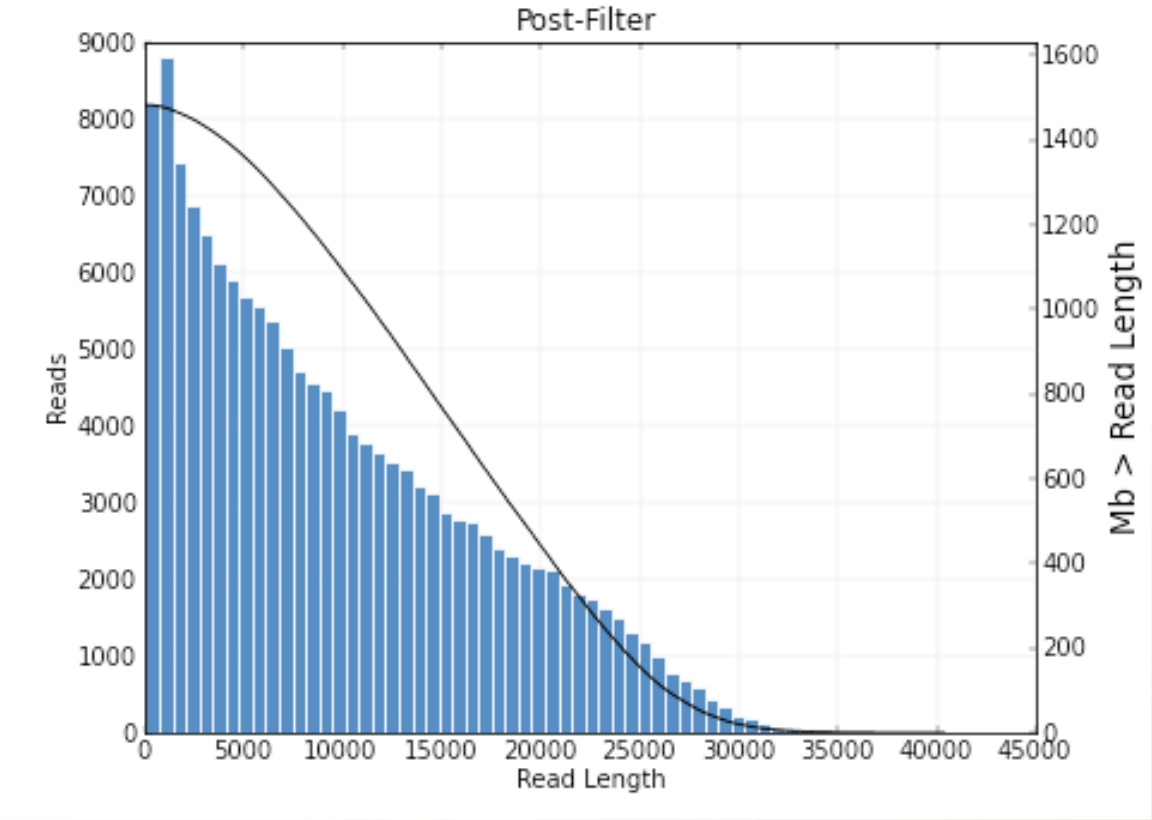
Short reads

Long reads (seeds)



PacBio – Current throughput & read lengths

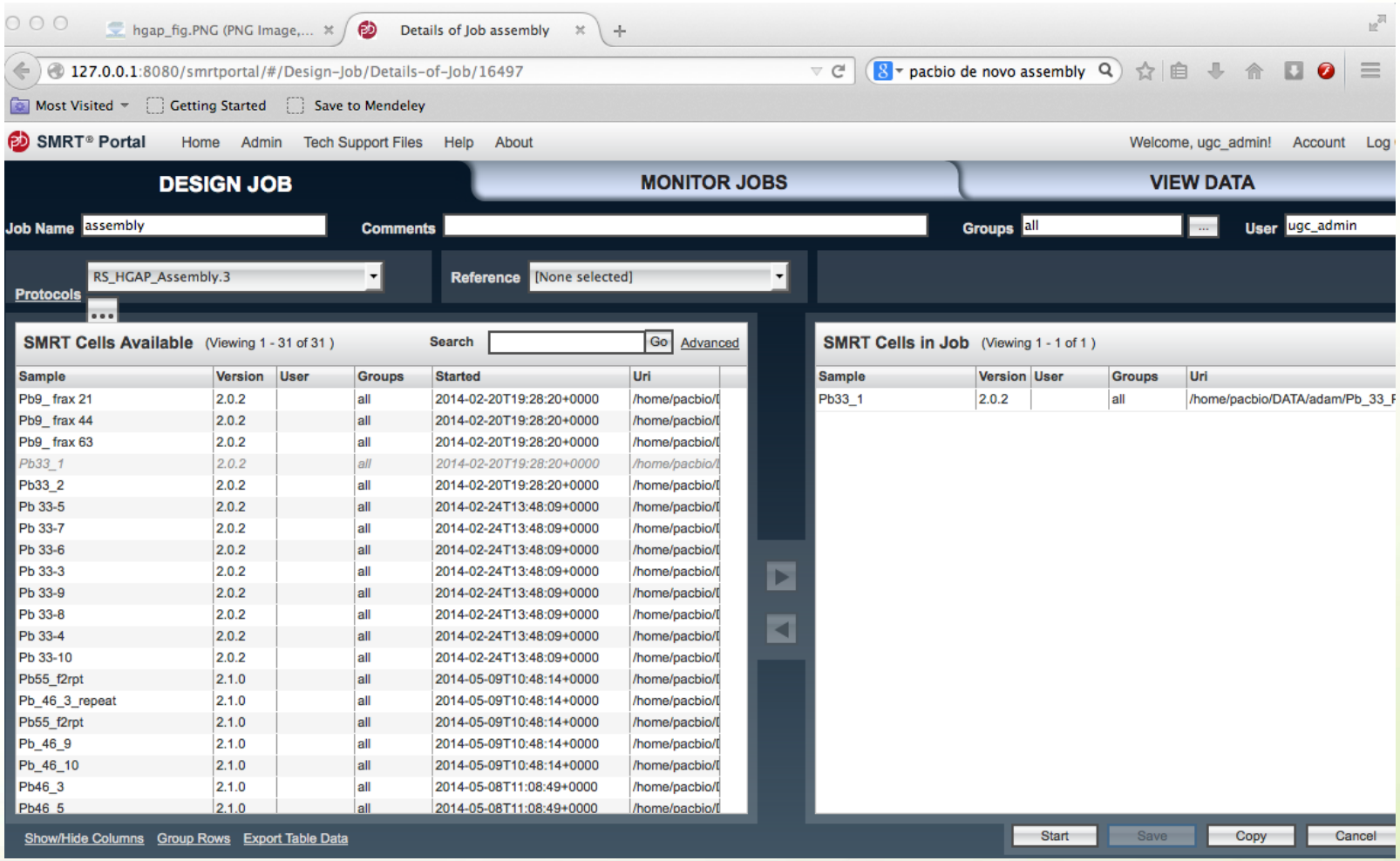
- >10kb average read lengths! (run from April 2014)



- ~ 1 Gb of sequence from one PacBio SMRT cell

PacBio assembly analysis

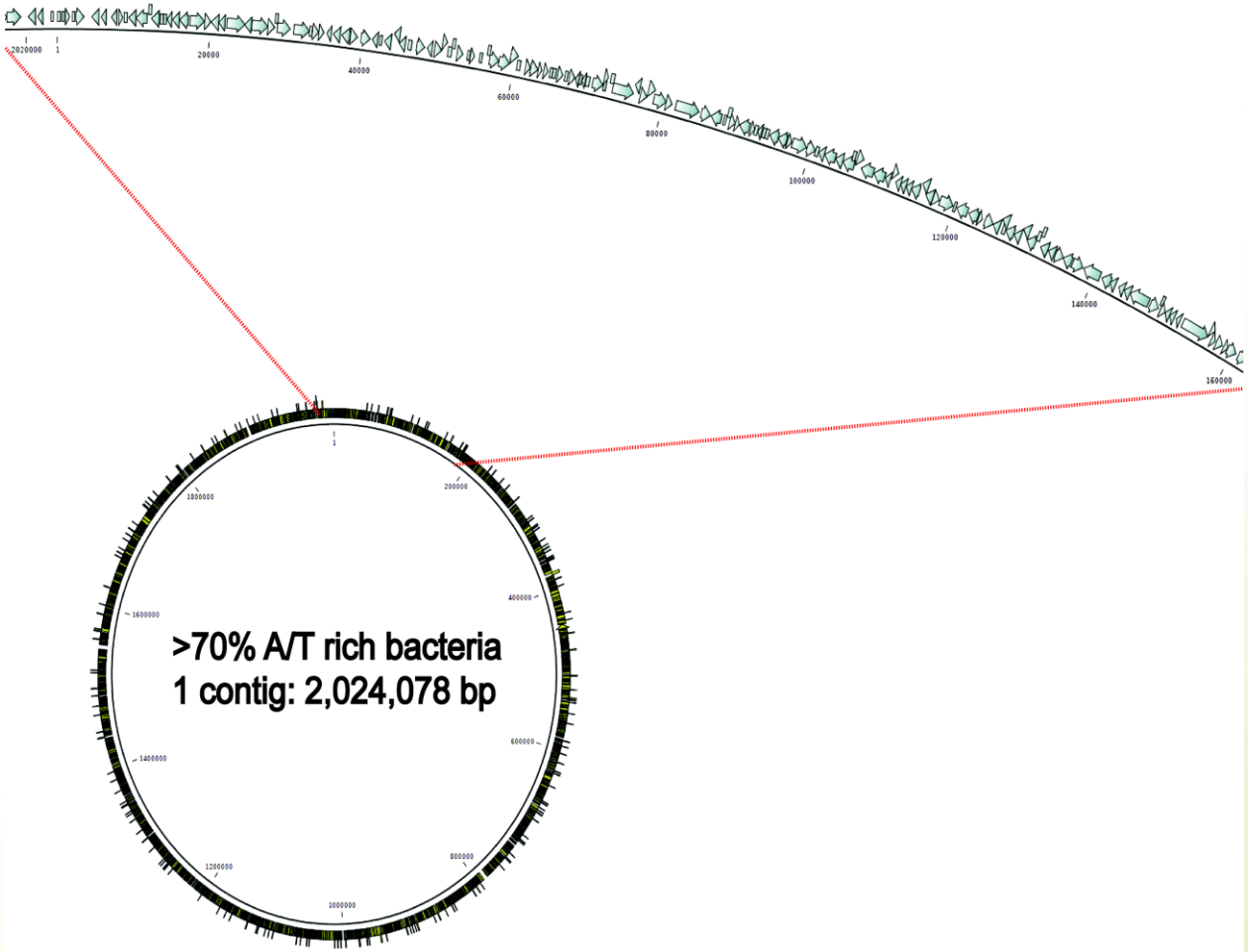
- Simple -- just click a button!!



Karolinska Institutet
Stockholms universitet
UPPSALA UNIVERSITET
SciLifeLab

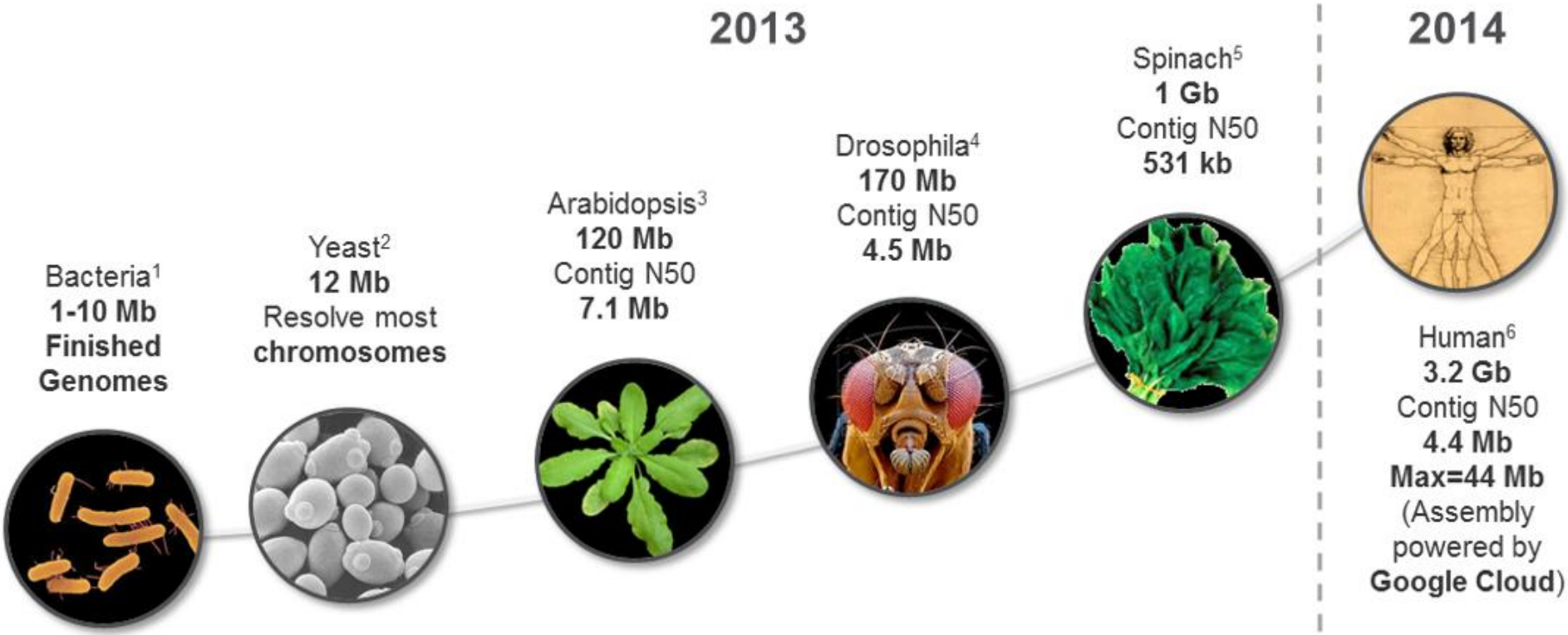
PacBio assembly, example result

- Example: Complete assembly of a bacterial genome



PacBio assembly – recent developments

- Also larger genomes can be assembled by PacBio..



Next step: assembly of large genomes

- A computational challenge!!

WEDNESDAY, FEBRUARY 12, 2014

Data Release: ~54x Long-Read Coverage for PacBio-only De Novo Human Genome Assembly

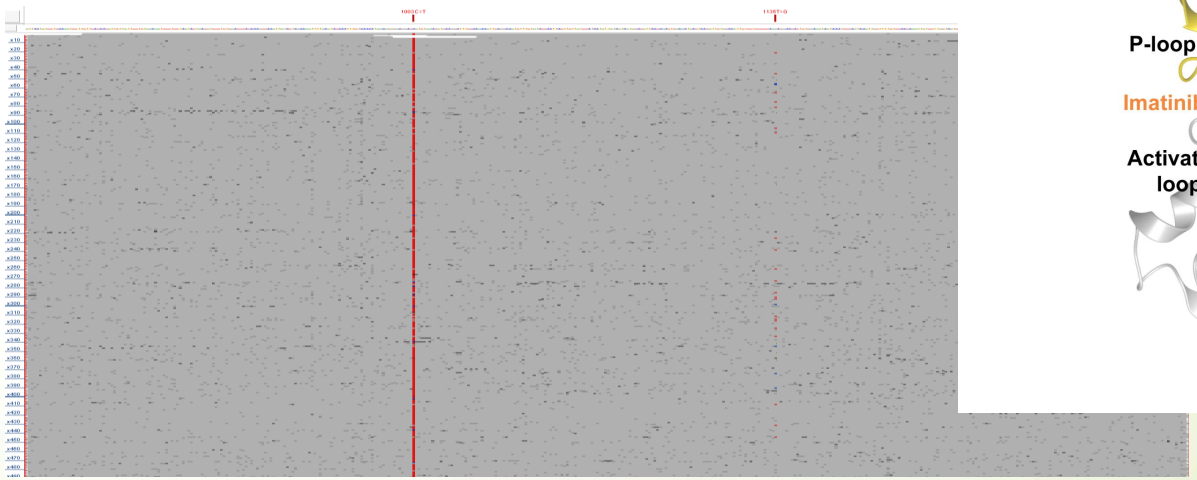
We are pleased to make publicly available a new shotgun sequence dataset of long PacBio® reads from a human DNA sample. We previously released sequence data using Single Molecule, Real-Time (SMRT®) Sequencing of ~10x coverage of this sample, sufficient for reference-based detection of structural variation. Today we expand on that release with additional data that increases the total sequencing coverage to ~54x. This long-read data has enabled the generation of the first *de novo* human genome assembly from PacBio-only sequence reads.
Download the 54x long-read coverage dataset.

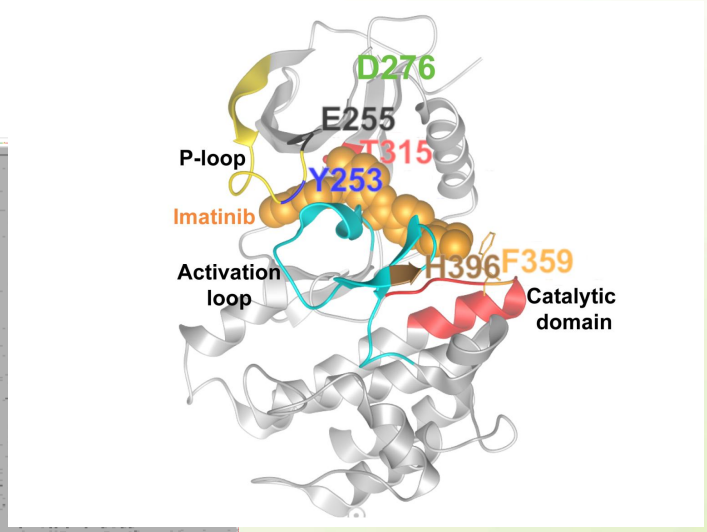
405,000 CPUh used on Google Cloud!

- We need to install such pipelines at UPPNEX!!

Example III:

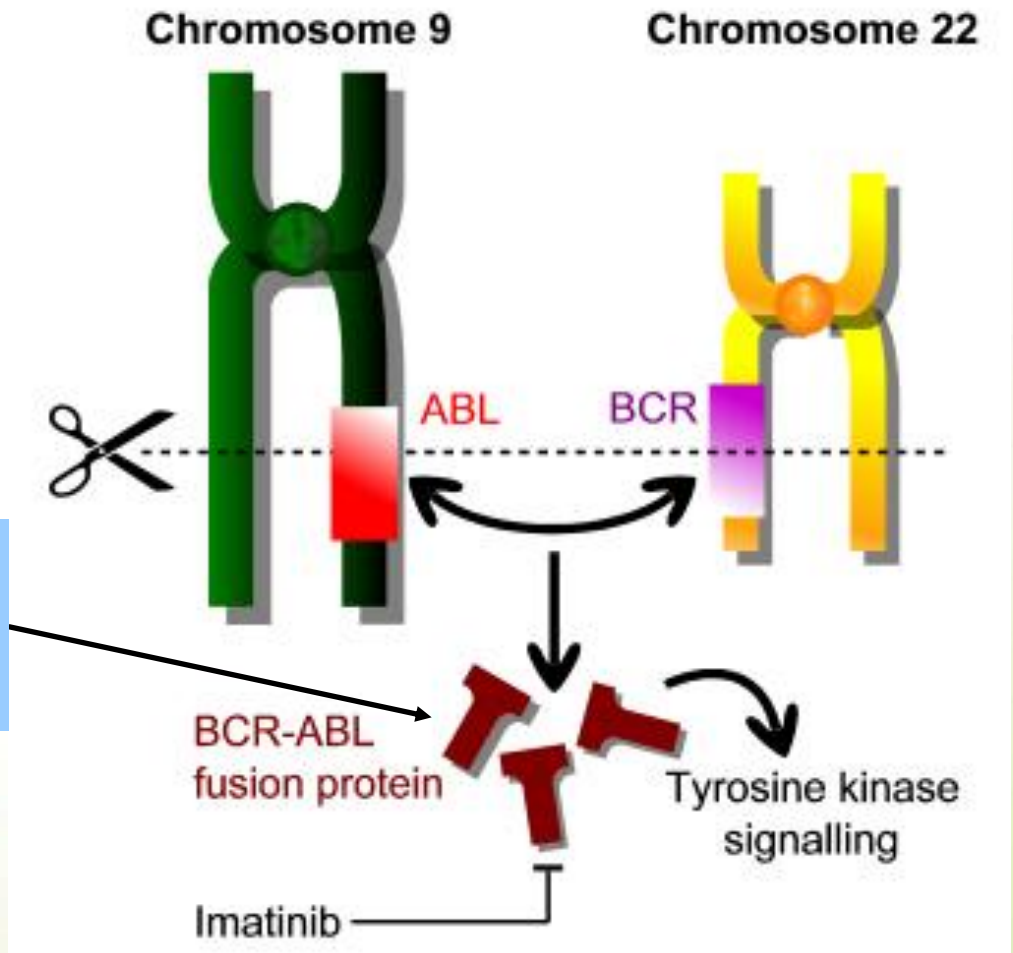
Clinical sequencing for Leukemia Treatment





Chronic Myeloid Leukemia

- BCR-ABL1 fusion protein – a CML drug target



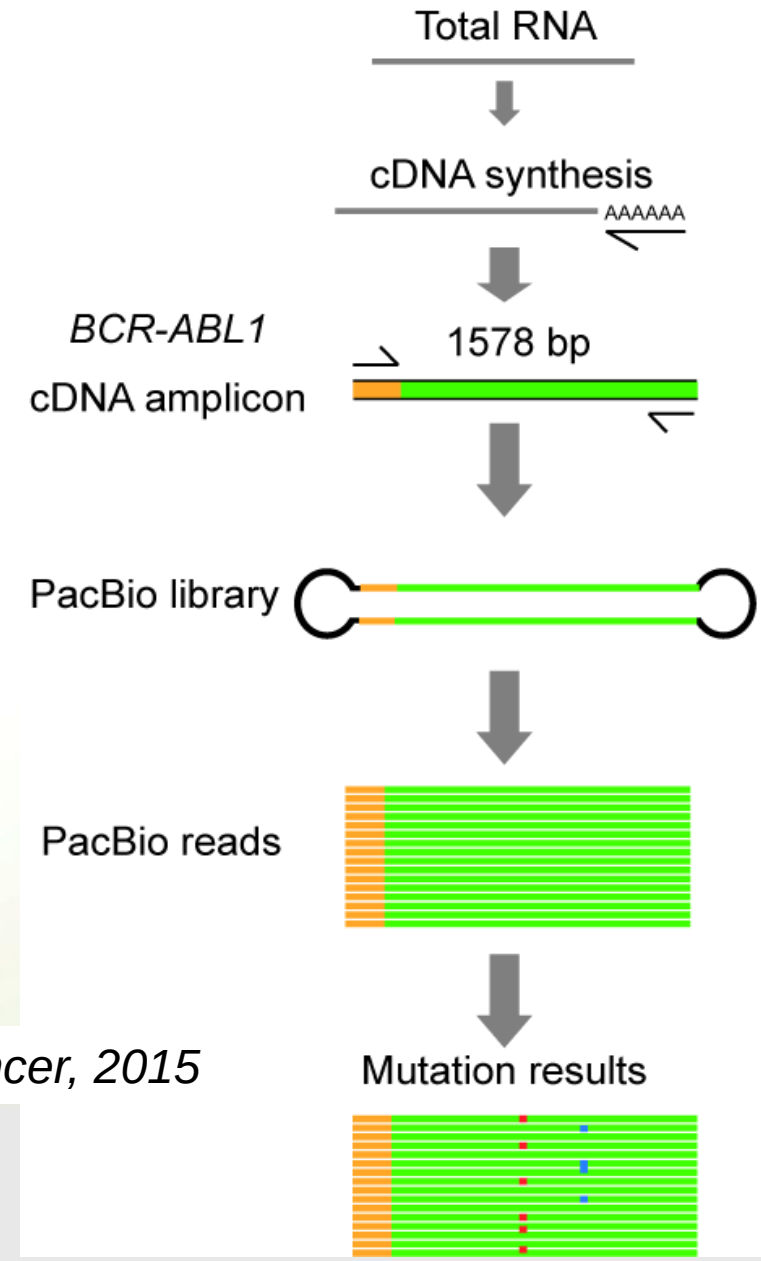
The BCR-ABL1 fusion protein can acquire resistance mutations following drug treatment

www.cambridgemedicine.org/article/doi/10.7244/cmj-1355057881

BCR-ABL1 workflow – PacBio Sequencing

From sample to results: < 1 week

1 sample/SMRT cell

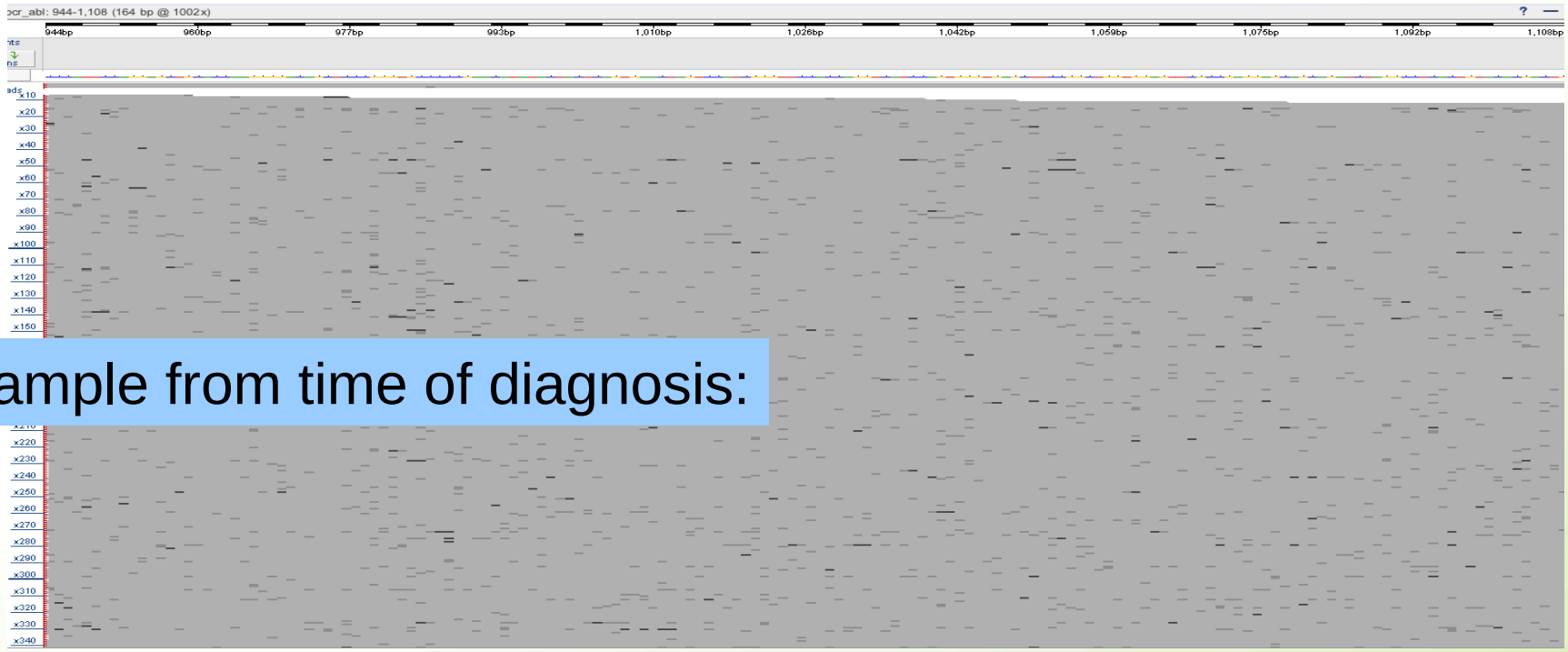


Cavelier et al., BMC Cancer, 2015

BCR-ABL1 mutations at diagnosis

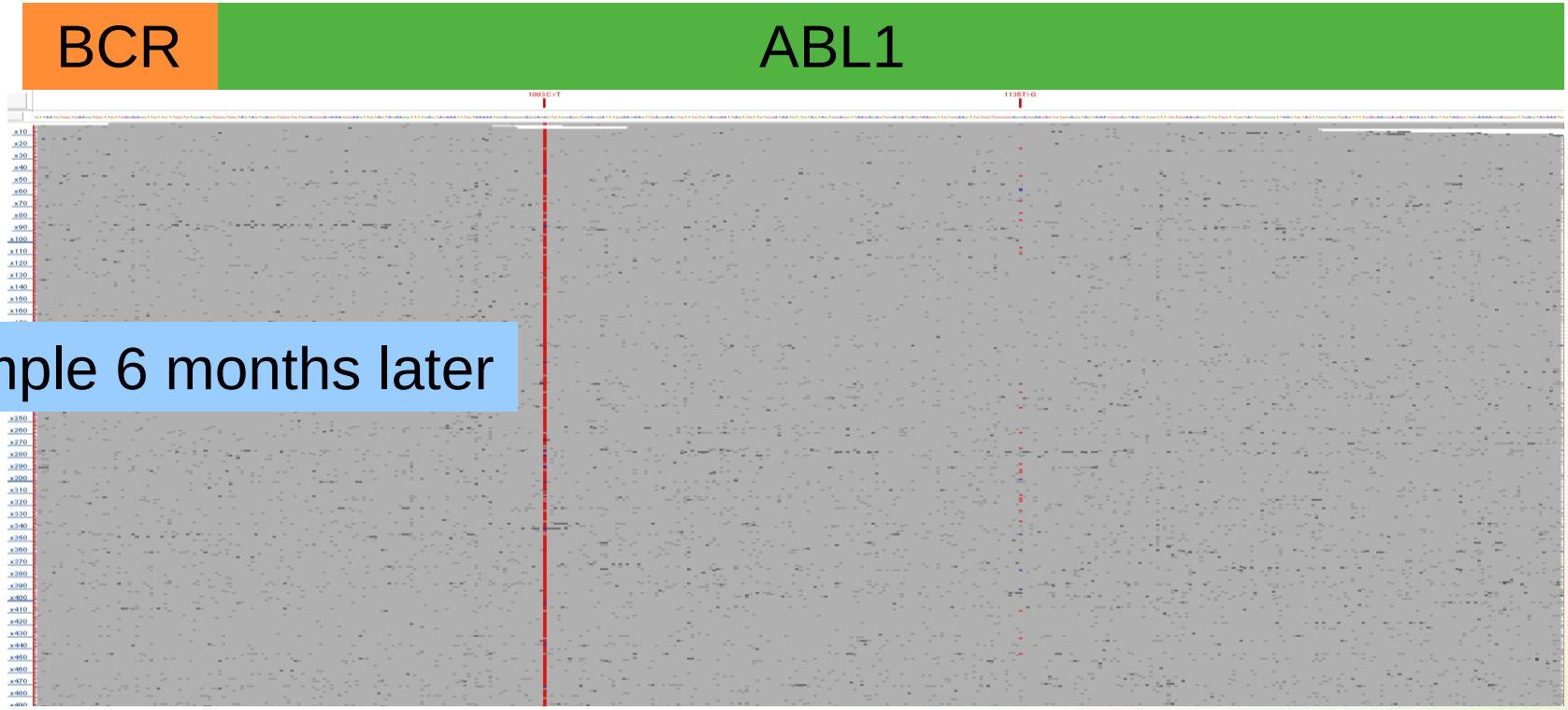
PacBio sequencing generates ~10 000X coverage!

BCR | ABL1



Sample from time of diagnosis:

BCR-ABL1 mutations in follow-up sample

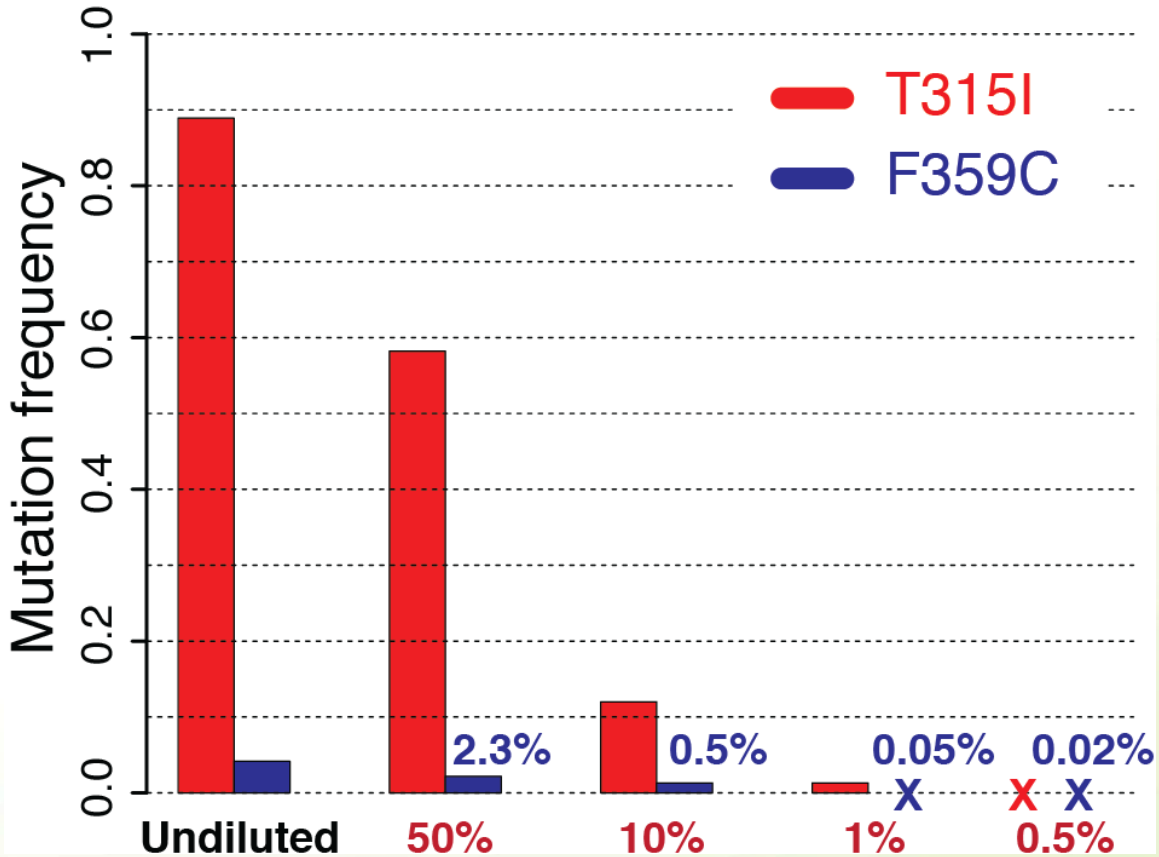


Sample 6 months later

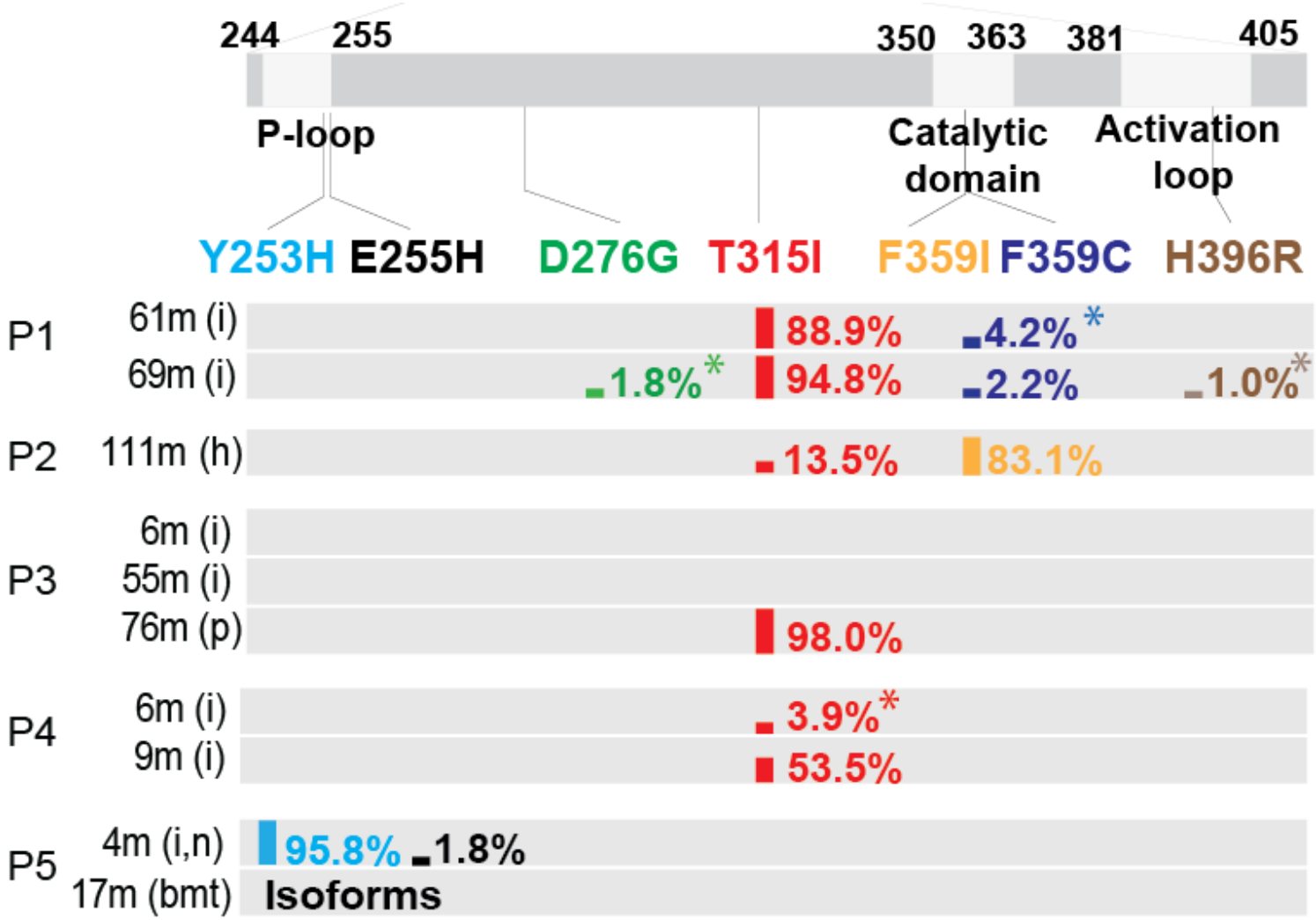
Mutations acquired in fusion transcript.
Might require treatment with alternative drug.

BCR-ABL1 dilution series results

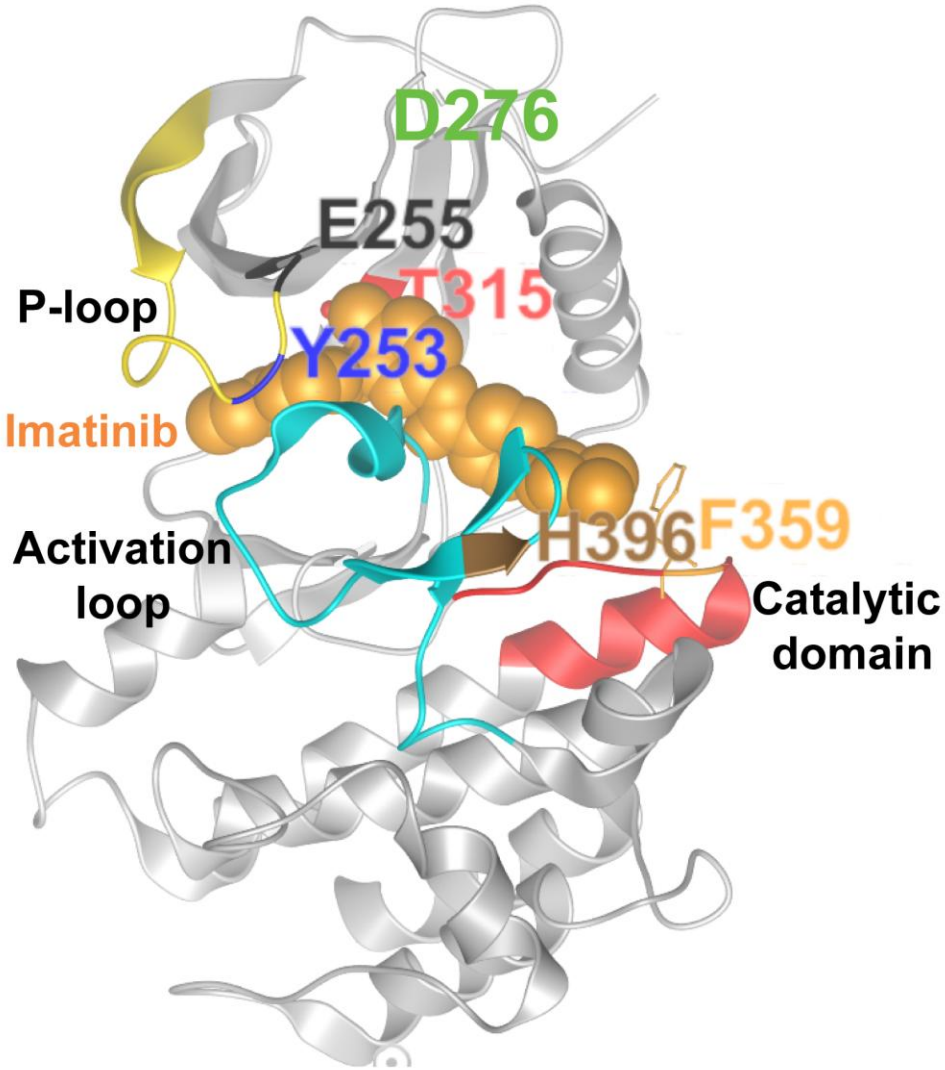
- Mutations down to 1% detected!



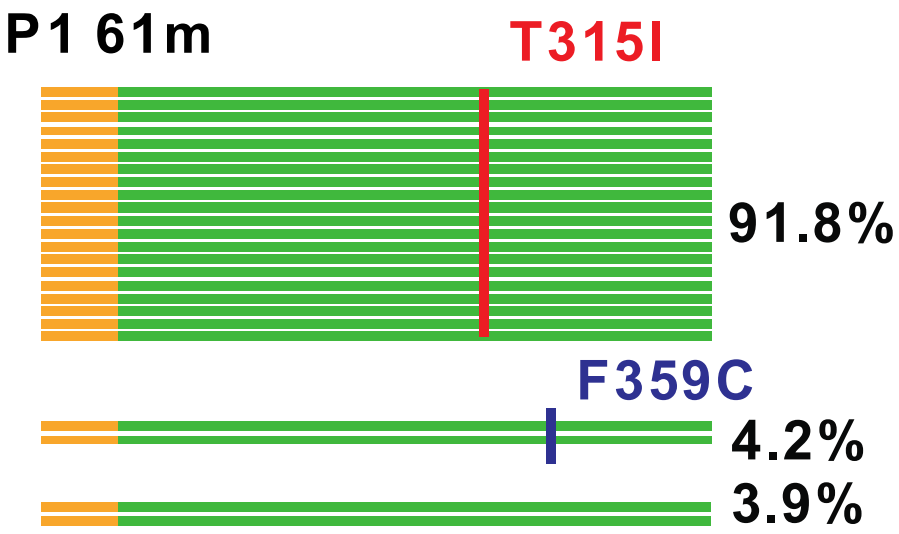
Summary of mutations in 5 CML patients

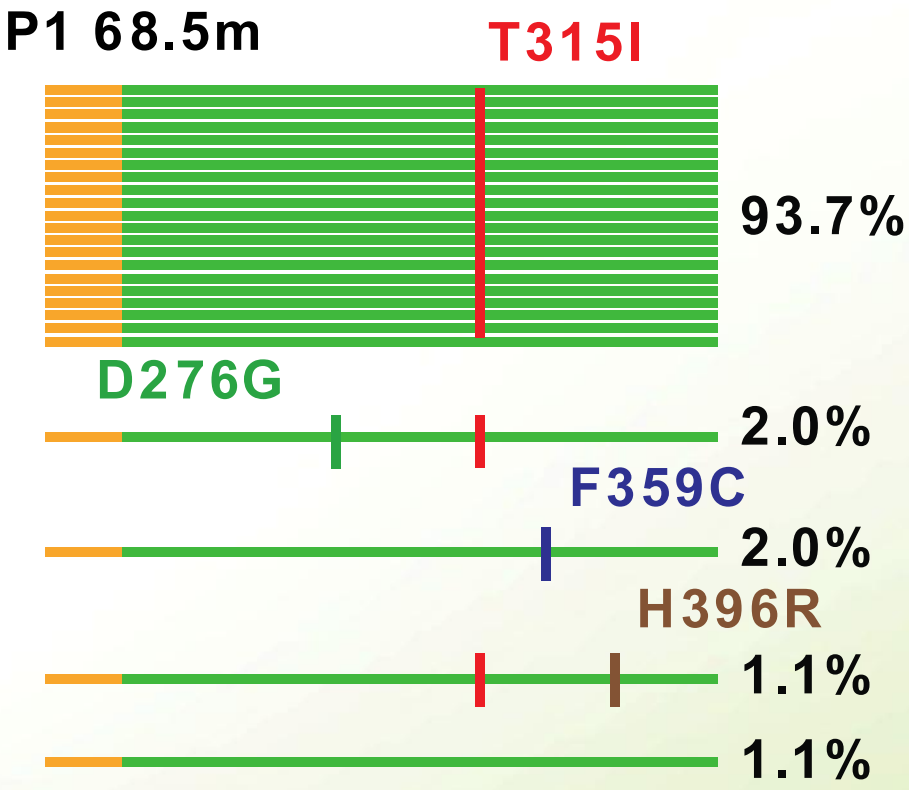


Mutations mapped to protein structure

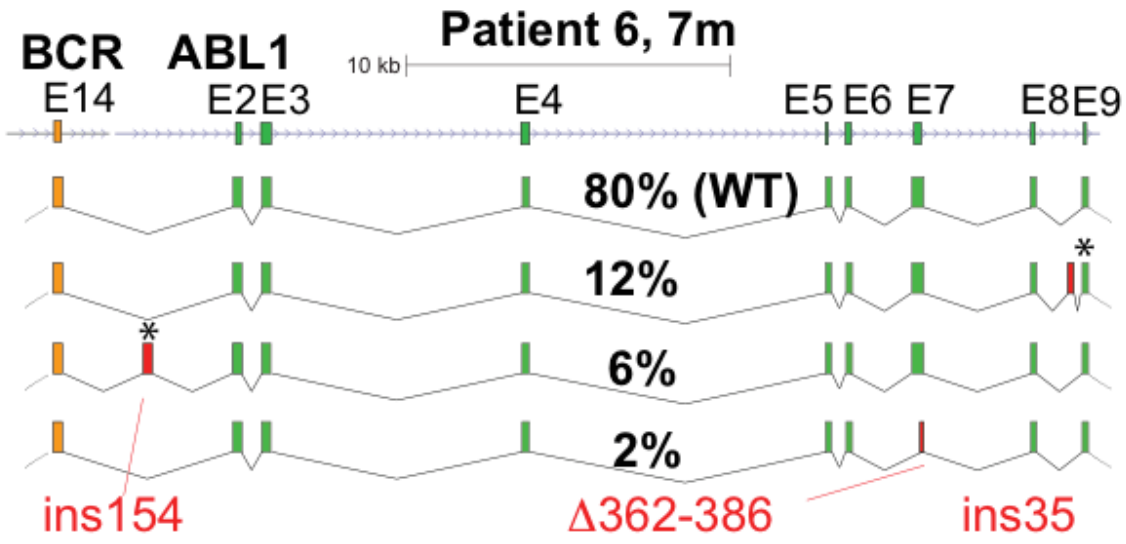


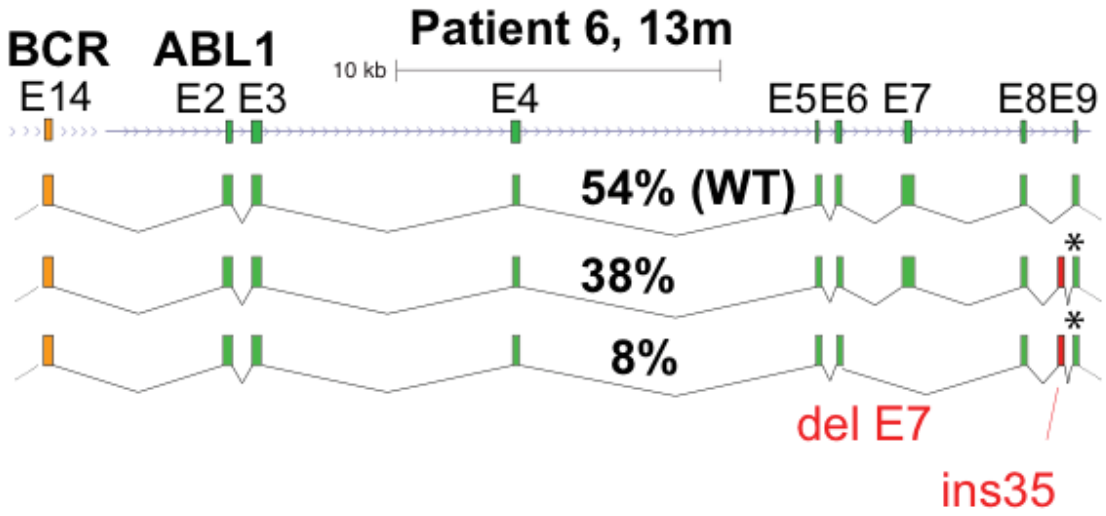
BCR-ABL1 - Compound mutations



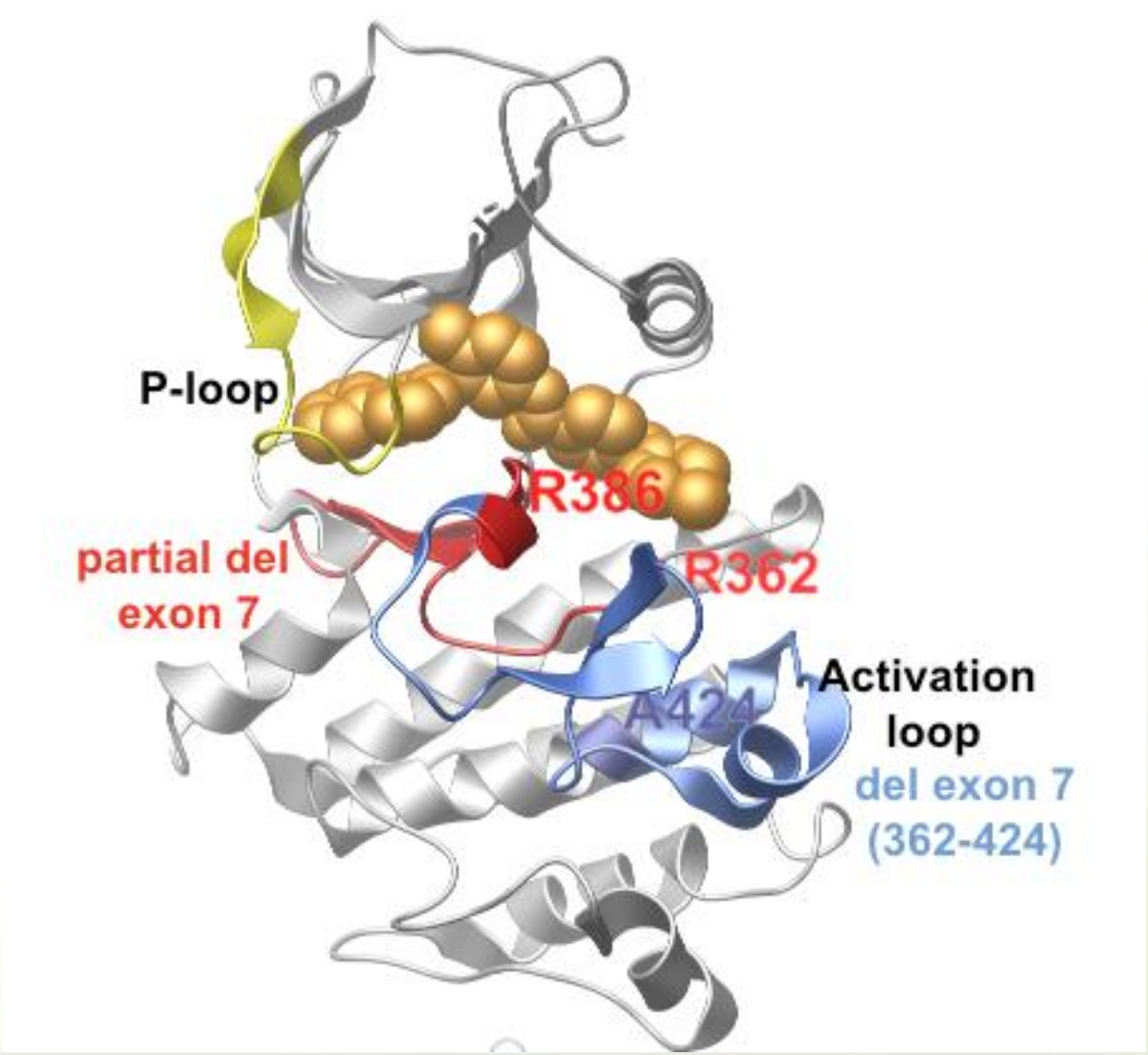


BCR-ABL1 - Multiple isoforms in one individual!



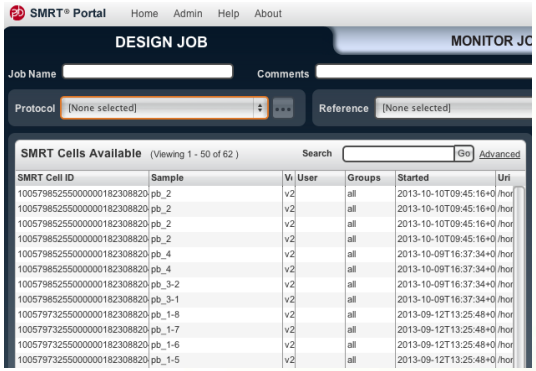


BCR-ABL1 – Isoforms and protein structure



Next step: A clinical diagnostics pipeline!

Step1. Create CCS reads



Step2. Run mutation analysis

E255K	CCAGTACGGG	[G/A]	AGGTGTACGA	7883	7143	0.475	8516	7669	0.474	16399	14812	0.475	positive
F359V	GAAGAAAAAC	[T/G]	TCATCCACAG	11646	3794	0.246	12231	3968	0.245	23877	7762	0.245	positive
L384M	TGATTTTGGC	[C/A]	TGAGCAGGTT	12704	1679	0.117	13545	1734	0.113	26249	3413	0.115	positive
M244V	GGACATCACC	[A/G]	TGAAGCACAA	14209	1550	0.098	15194	1695	0.1	29403	3245	0.099	positive
T315I	TATATCATCA	[C/T]	TGAGTTCATG	15392	793	0.049	16291	854	0.05	31683	1647	0.049	positive
L387M	CCTGAGCAGG	[T/A]	TGATGACAGG	13069	321	0.024	13977	403	0.028	27046	724	0.026	positive
K247R	ATGAAGCACA	[A/G]	GCTGGGCGGG	13901	14	0.001	14805	8	0.001	28706	22	0.001	negative
L248V	GAAGCACAAG	[C/G]	TGGGCCGGGG	13708	0	0	14823	0	0	28531	0	0	negative
G250E	AAGCTGGGCG	[G/A]	GGGCCAGTAC	13330	2	0	14453	4	0	27783	6	0	negative
Q252H	GCGGGGGCCA	[G/T]	TACGGGGAGG	6895	0	0	7489	2	0	14384	2	0	negative
Y253H	CGGGGGCCAG	[T/C]	ACGGGGAGGT	6877	1	0	7439	1	0	14316	2	0	negative
Y253F	GGGGGCCAGT	[A/T]	CGGGGGAGGTG	7146	0	0	7721	0	0	14867	0	0	negative
E255V	CAGTACGGGG	[A/T]	GGTGTACGAG	7932	0	0	8548	0	0	16480	0	0	negative
L273M	CGTGAAGACC	[T/A]	TGAAGGAGGA	15642	0	0	16694	0	0	32336	0	0	negative
D276N	CTTGAAGGAG	[G/A]	ACACCATGGA	15772	5	0	16786	9	0.001	32558	14	0	negative
D276G	TTGAAGGAGG	[A/G]	CACCATGGAG	15840	40	0.003	16855	37	0.002	32695	77	0.002	negative
T277P	GAAGGAGGAC	[A/G]	CCATGGAGGT	15786	10	0.001	16815	26	0.002	32601	36	0.001	negative
T277S	GAAGGAGGAC	[A/T]	CCATGGAGGT	15786	1	0	16815	0	0	32601	1	0	negative
T277N	AAGGAGGACA	[C/A]	CATGGAGGTG	15899	2	0	16939	2	0	32838	4	0	negative

Step3. Upload to result server

Karolinska Institutet
KTH
Stockholms universitet
UPPSALA UNIVERSITET
SciLifeLab

Reporting system for mutation results

Sample ID

Run ID

Date (yyyy-mm-dd)

Mutation

Sequenced

Search

Reset

Details	Sample ID	Run ID	M244V	Y253H	Y253H[E255V]	E255K	E255V	D276G	T315I	F359C	F359V	F359I	L384M	L387M	H396R	rs222798	Seq	Date
1	R3740	pb_003_1	9.9			47.5			4.9		24.5		11.5	2.6			1	2014-12-0
2	R7394	pb_003_2							91.7	4.7								2014-12-0
3	R7840	pb_014_3						2.0	96.8	2.1					1.2	39.7		2014-12-0
4	R9171	pb_014_4	100.0															2014-12-0
5	R8484	pb_014_5	100.0						14.4			85.4						2014-12-0
6	R4419	pb_015_1														69.3		2014-12-0
7	R4765	pb_015_2																2014-12-0
8	R7715	pb_015_3	0.6						99.9									2014-12-0
9	R9452	pb_015_4							99.9									2014-12-0
10	R5208	pb_033_1		99.8														2014-12-0

Collaboration with Wesley Schaal & Ola Spjuth, UPPNEX/Uppsala Univ

Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet · SciLifeLab

Ion Torrent – News and updates

- AmpliSeq Human Whole Transcriptome panel
  - Expression levels for ~20.000 human genes
  - 10-100 ng of input is enough!
  - Works on FFPE samples!!
  - Cheaper than conventional RNA-seq
  - Simple bioinformatics

- HiQ chemistry
  - Improves accuracy in sequencing
  - Reduces indel error rates

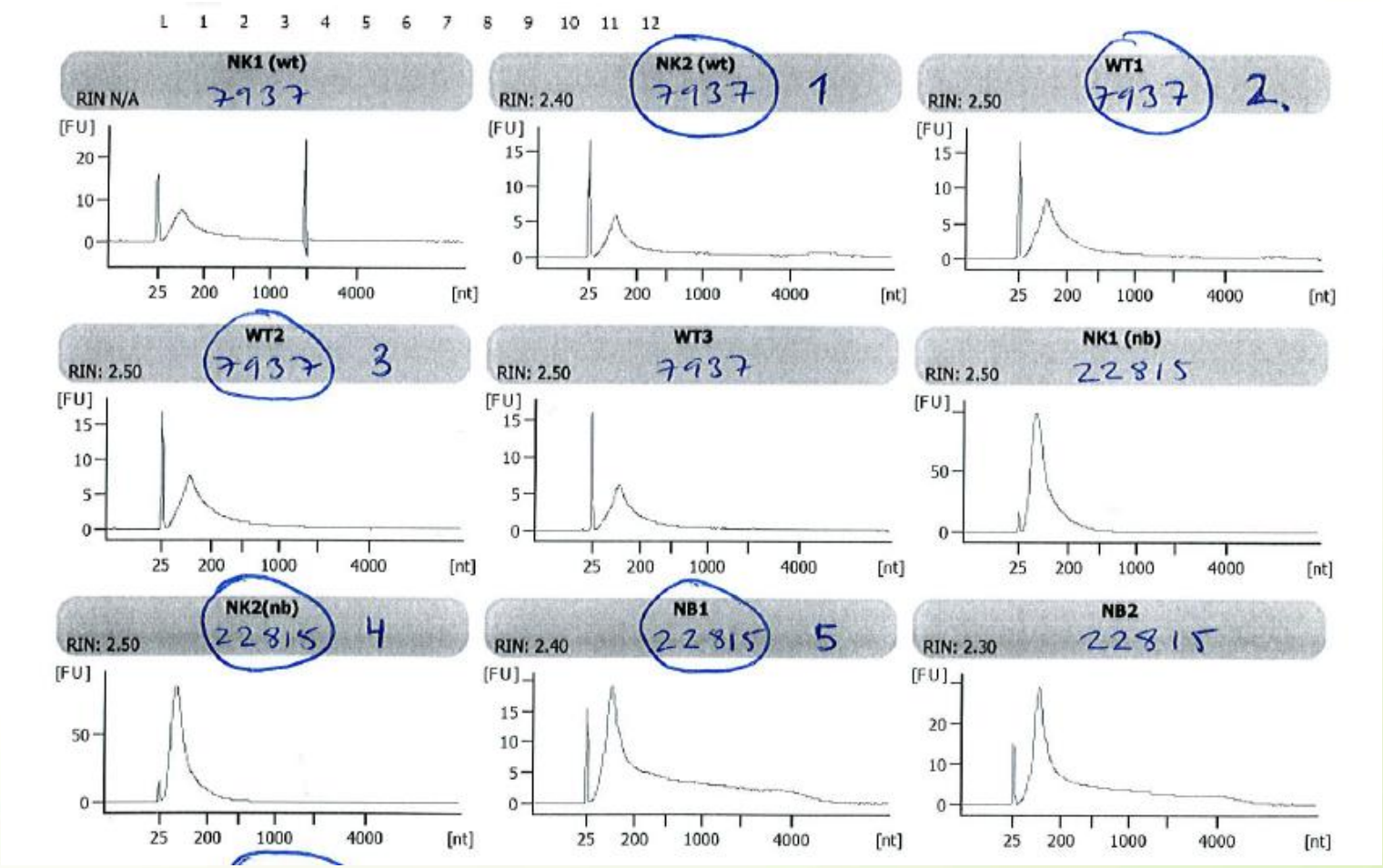
R_2014_10_07_13_31_43_user_PR5-44-up_109_111_A...

	A	B	C	D	E
1	Gene	Target	IonXpress_001	IonXpress_002	IonXpress_010
2	SEC24B-AS1	AMPL377418	0.96	1.568	1.369
3	A1BG	AMPL174256	0.107	0	0.152
4	A1CF	AMPL365934	0	0	0
5	GGACT	AMPL173676	0.213	0.922	1.065
6	A2M	AMPL1384	77.965	110.653	0
7	A2ML1	AMPL359429	132.679	160.723	0
8	A2MP1	AMPL376317	0.213	0.184	0.076
9	A4GALT	AMPL318887	12.052	26.188	0.076
10	A4GNT	AMPL323789	0	0	0
11	AAAS	AMPL336793	11.412	11.987	5.324
12	AACS	AMPL369958	48.635	71.74	5.857
13	AADAC	AMPL582558	31.89	28.216	0
14	AADACL2	AMPL223111	100.683	163.858	0
15	AADACL3	AMPL444945	0	0	0
16	AADACL4	AMPL612401	0	0	0
17	AADAT	AMPL326139	1.92	2.305	0
18	AAGAB	AMPL144320	51.088	38.175	14.984
19	AAK1	AMPL259042	1.6	1.014	7.835
20	AAMP	AMPL346680	79.885	69.066	41.454
21	AANAT	AMPL327561	0	0	0
22	AARS	AMPL107840	45.755	46.474	8.595
23	AARS2	AMPL314692	7.786	10.696	10.192
24	PTGES3L-AAR	AMPL466342	0.213	0	0
25	AASDH	AMPL214471	1.387	4.242	0.532
26	AASDHPPT	AMPL250926	31.037	33.565	8.747
27	AASS	AMPL293263	2.666	6.363	0.456
28	AATF	AMPL125583	75.405	63.349	111.28
29	AATK	AMPL554291	1.6	3.227	37.423
30	ABAT	AMPL338537	2.666	5.533	10.192
31	ABCA1	AMPL283855	40.742	46.659	31.87
32	ABCA10	AMPL185495	9.279	23.79	0.685
33	ABCA12	AMPL158582	306.207	231.172	0
34	ABCA13	AMPL344817	1.173	1.568	1.217
35	ABCA17P	AMPL198774	0	0.184	0
36	ABCA2	AMPL809904	29.437	43.8	45.562
37	ABCA3	AMPL507627	6.079	16.782	7.378

R_2014_10_07_13_31_43_user_PR5-

Ion Torrent – RNA-Seq on FFPE

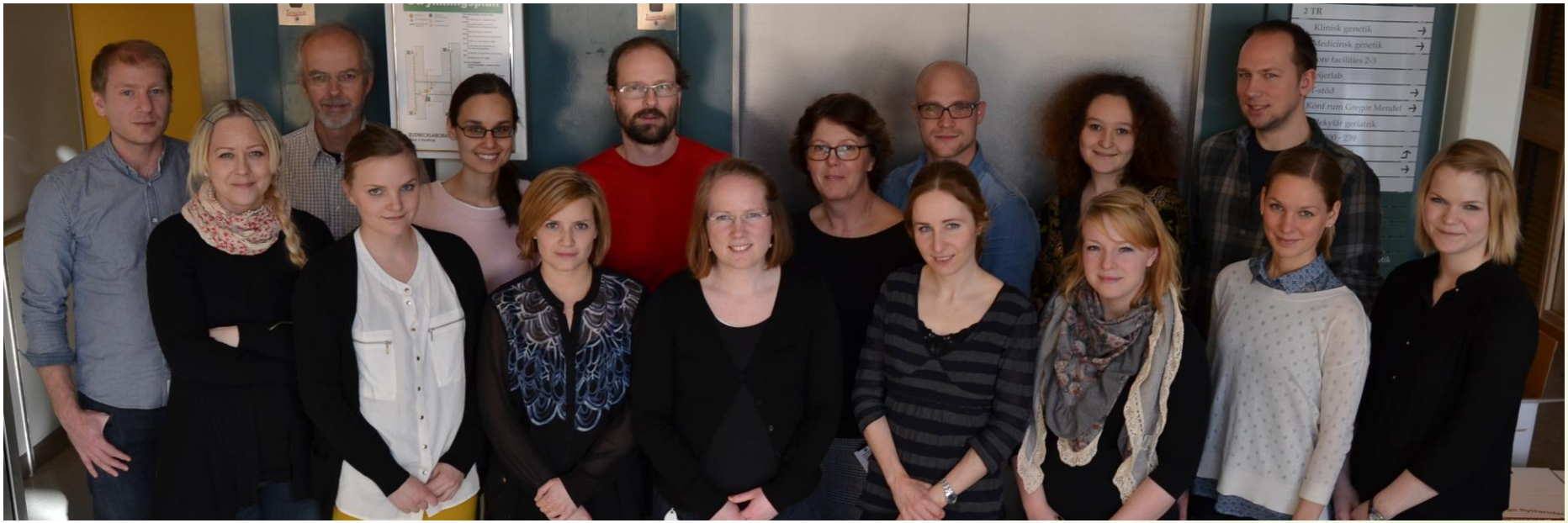
- Good results obtained for most of these samples!



PacBio – News and updates

- HLA typing
  - Full length sequencing of HLA genes
  - Multiplexing of several individuals in one run
- Fast track clinical samples
  - Preparing workflows for rapid sequencing
  - Organ transplantation, diagnostics, outbreaks, ...
- New chemistry and active loading of SMRT cells
  - Improved quality, longer reads
  - Increased throughput (early 2015)

Thank you!



NATIONAL GENOMICS INFRASTRUCTURE

SciLifeLab

Karolinska Institutet

KTH

Stockholms universitet

UPPSALA UNIVERSITET

SciLifeLab