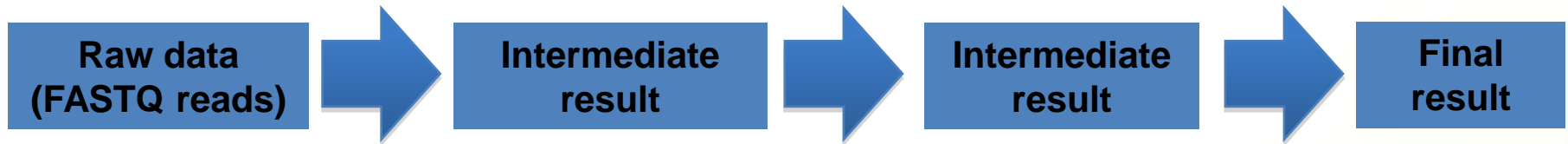


Next Generation Sequencing and Bioinformatics Analysis Pipelines

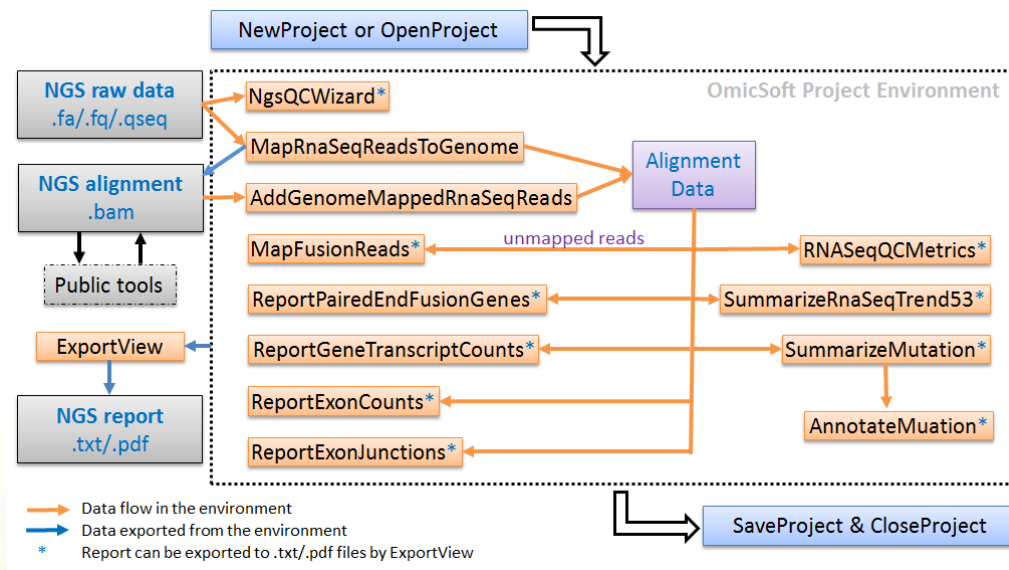
Adam Ameer
National Genomics Infrastructure
SciLifeLab Uppsala
adam.ameur@igp.uu.se

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)



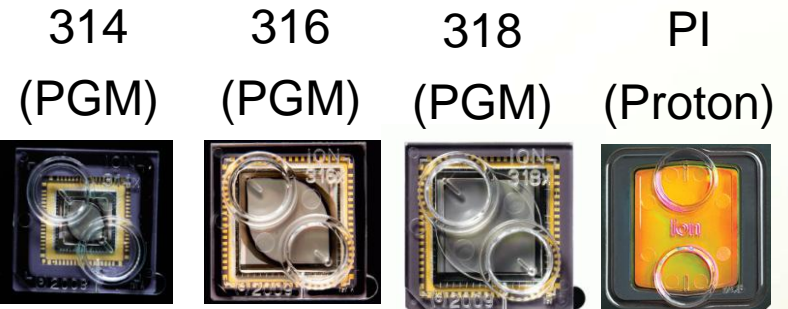
Ion Proton



Ion Torrent output

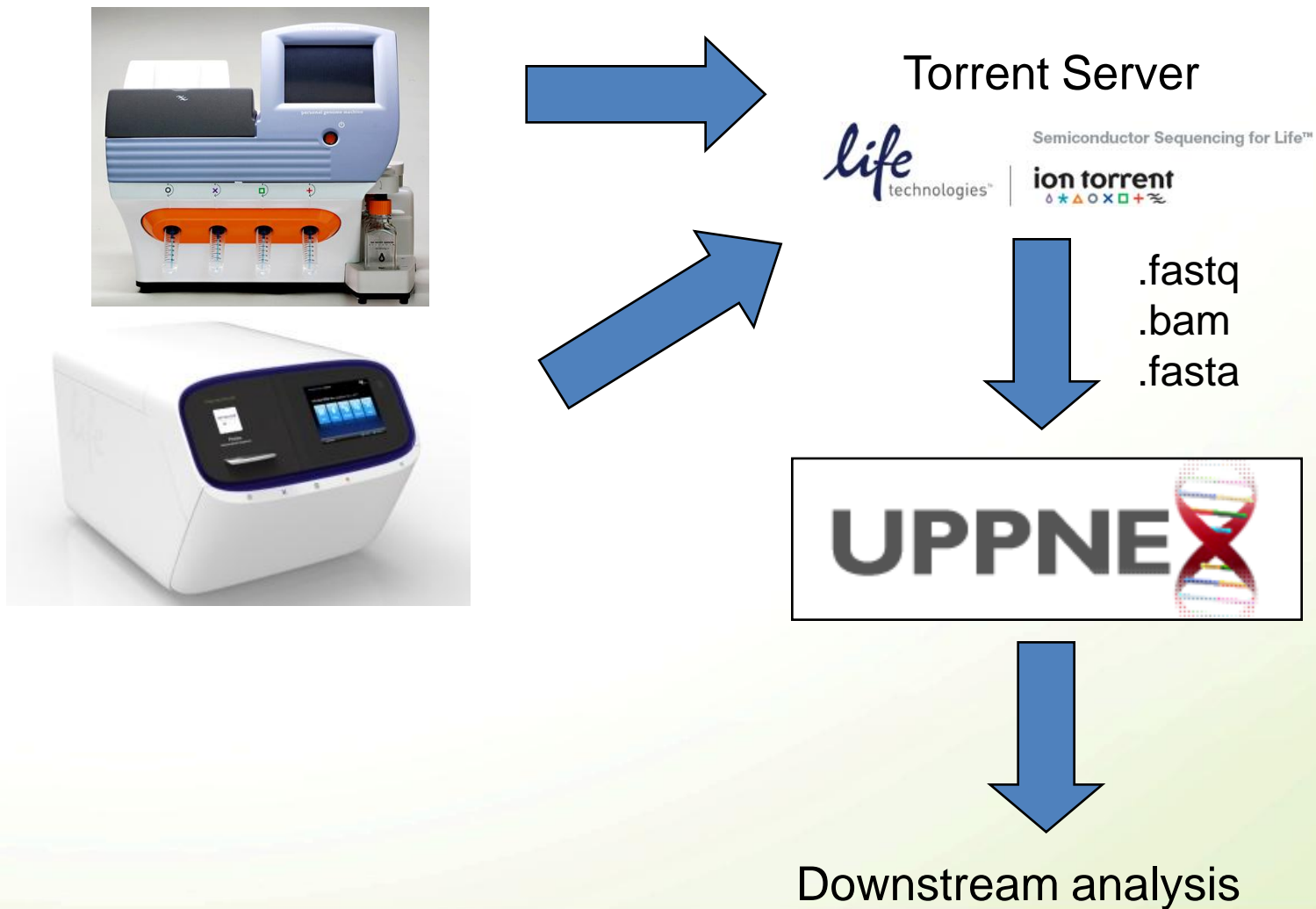
- Ion Torrent throughput
 - ~ from 10Mb to >10Gb, depending on the chip

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

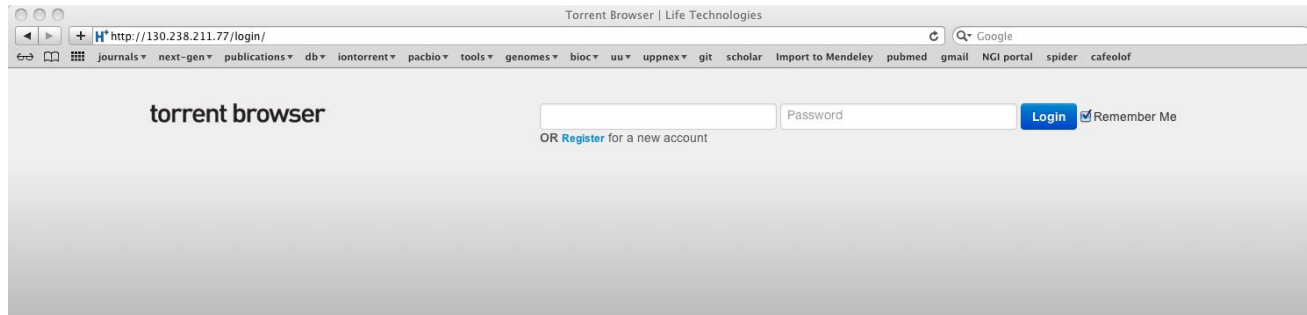


- Read lengths: 400bp (PGM), 200bp (Proton)
- Output file format: FASTQ
- What can we use Ion Torrent for?
 - Anything, except perhaps very large genomes

Ion Torrent analysis workflow



Torrent Suite Software



A screenshot of the 'Torrent Server' interface. The page title is 'Torrent Server - Torrent Browser'. The interface includes a navigation menu with options like 'PLANNING', 'RUNS', 'REPORTS', 'SERVICES', 'REFERENCES', 'CONFIG', and 'HELP/ABOUT'. Below the navigation, there are filters for 'Date start', 'Date end', 'PGM', 'Project', 'Sample', 'Reference', 'Storage', 'Starred', and 'Search by Run'. The main content is a table with columns for 'Runs', 'Chip', 'Project', 'Sample', 'Reference', 'PGM', 'Status', 'Storage', and 'Date'. The table lists various runs with their respective details.

Runs	Chip	Project	Sample	Reference	PGM	Status	Storage	Date
UU1-23	316D	selector-kras-c12-13	test-kras	KRAS_c12_13	UU1	Complete	Delete Raw	Jan 24 2012
UU2-10	316D	ugit007-5-uu2	ugit007-5-uu2	GU562296	UU2	Complete	Delete Raw	Jan 20 2012
UU2-9	314R	ugit006-1-uu2	ugit006-1-uu2	NC012225	UU2	Complete	Delete Raw	Jan 20 2012
UU2-8	316D	ugit007-3-uu2	ugit007-3-uu2	GU562296	UU2	Complete	Delete Raw	Jan 19 2012
UU1-22	316D	ugit007-4-uu1	ugit007-4-uu1	GU562296	UU1	Complete	Delete Raw	Jan 19 2012
UU1-21	314R	ugit007-2-uu1	ugit007-2-uu1	GU562296	UU1	Complete	Delete Raw	Jan 19 2012
UU2-7	314R	ugit007-1-uu2	ugit007-1-uu2	GU562296	UU2	Complete	Delete Raw	Jan 19 2012
UU2-6	314R	ugit005-1-uu2	ugit005-1-uu2	AL390716	UU2	Complete	Delete Raw	Jan 12 2012
UU2-5	316D	ugit001-1-uu2-ot	ugit001-1-uu2-ot	NC000913	UU2	Complete	Delete Raw	Jan 12 2012
UU1-20	314R	ugit003-2-uu1	ugit003-2-uu1	ugit_003_ref	UU1	Complete	Delete Raw	Jan 02 2012
UU1-19	314R	ugit003-1-uu1	ugit003-1-uu1	ugit_003_ref	UU1	Complete	Delete Raw	Jan 02 2012
UU1-18	316D	ugit002-2-uu1	ugit002-2-uu1	NC007530	UU1	Complete	Delete Raw	Dec 19 2011
UU1-17	316D	ugit002-1-uu1	ugit002-1-uu1	NC007530	UU1	Complete	Delete Raw	Dec 19 2011
UU2-4	316D	ugit001-1-uu2-111216	ugit001-1-uu2	NC000913	UU2	Complete	Delete Raw	Dec 17 2011
UU1-16	314R	ugit004-4-uu1	ugit004-4-uu1	Lacto_GG	UU1	Complete	Delete Raw	Dec 14 2011
UU1-15	314R	ugit004-1-uu1	ugit004-1-uu1	Lacto_GG	UU1	Complete	Delete Raw	Dec 14 2011
UU1-14	316D	ugit1-2	ugit1-2-111215	NC000913	UU1	Complete	Delete Raw	Dec 13 2011
UU1-14	316D	ugit1-1	ugit1-1-111215	NC000913	UU1	Complete	Delete Raw	Dec 13 2011
UU1-12	314R	ugit43-111205	ugit43-111205	Lacto_GG	UU1	Complete	Delete Raw	Dec 03 2011
UU1-11	314R	ugit42-111205	ugit42-111205	Lacto_GG	UU1	Complete	Delete Raw	Dec 03 2011

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?

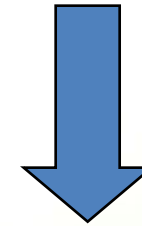


- PacBio read lengths: 500bp-40kb
- Output file format: FASTQ
- What can we use PacBio for?
 - Anything, except really large genomes

PacBio analysis workflow



In-house PacBio cluster



.fastq
.bam
.fasta



Downstream analysis

SMRT analysis portal

SMRT® Portal Home Admin Help About Welcome, ugc_admin! Account Log Off

DESIGN JOB MONITOR JOBS VIEW DATA

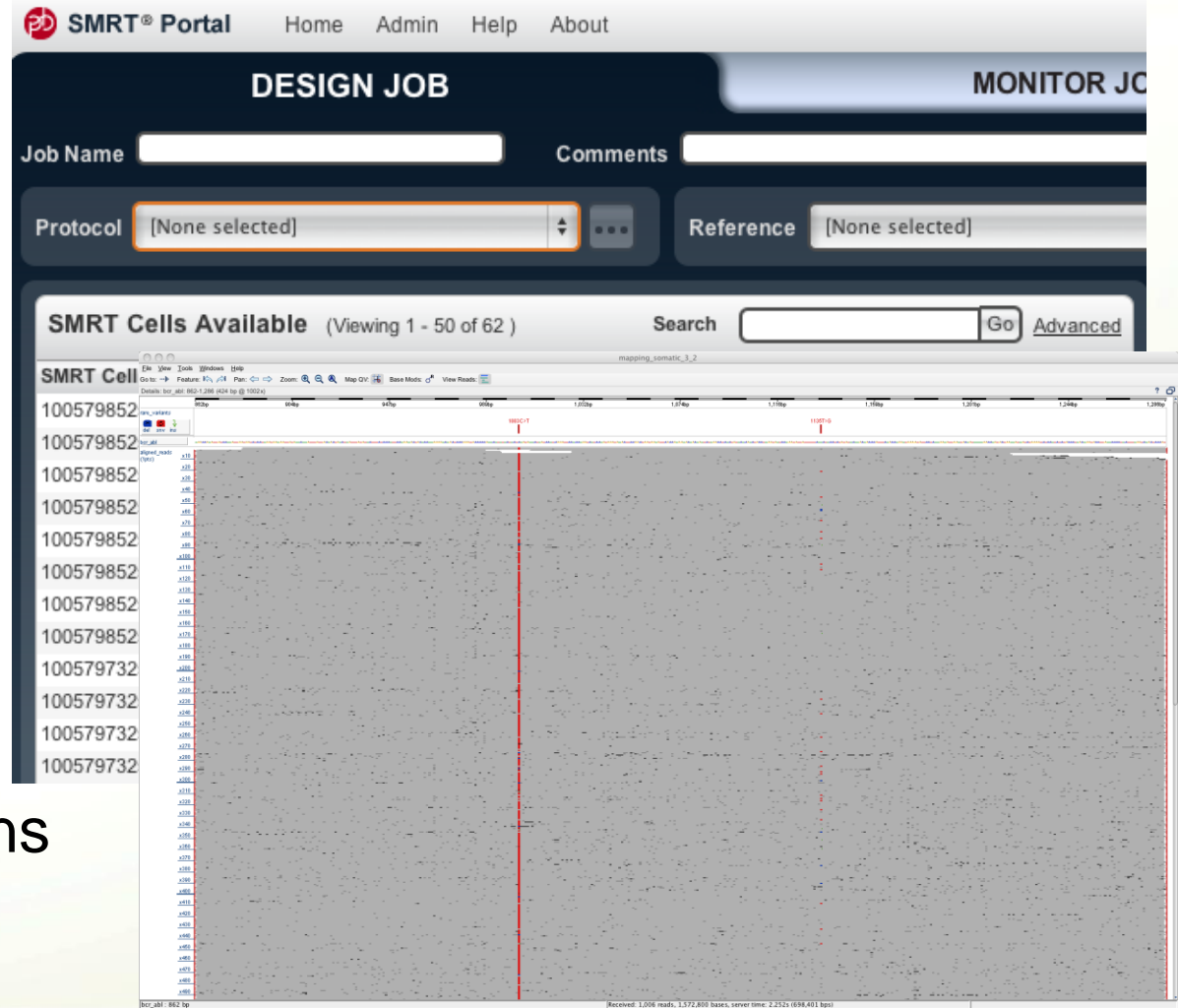
Open Existing Create New Import and Manage

RECENT JOBS

Job Name	Protocol	Reference Sequence	Started	Status	User
mapping_sample_S20	RS_Resequencing.1	rcatest	2013-09-13T14:21:	Completed	ugc_admin
testaaa	RS_HGAP_Assembly.1		2013-09-13T13:29:	Completed	ugc_admin
mapping_sample_S19	RS_Resequencing.1	rcatest	2013-09-13T12:55:	Completed	ugc_admin
mapping_sample_S18	RS_Resequencing.1	rcatest	2013-09-13T11:39:	Completed	ugc_admin
sample_S3_mapping	RS_Resequencing.1	rcatest	2013-09-13T11:35:	Completed	ugc_admin

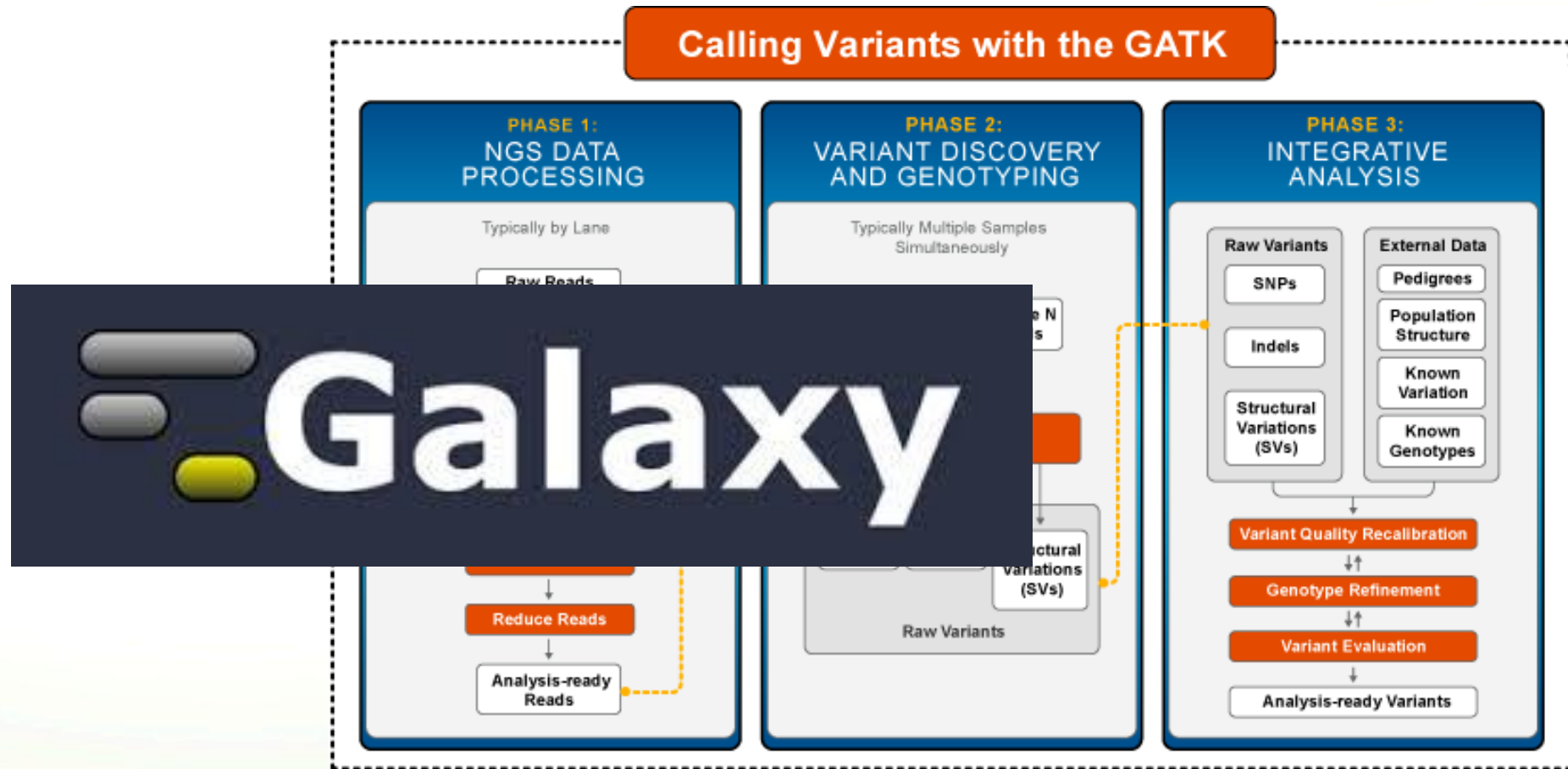
SMRT analysis pipelines

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



What about Illumina?

- There are many different pipelines for Illumina...

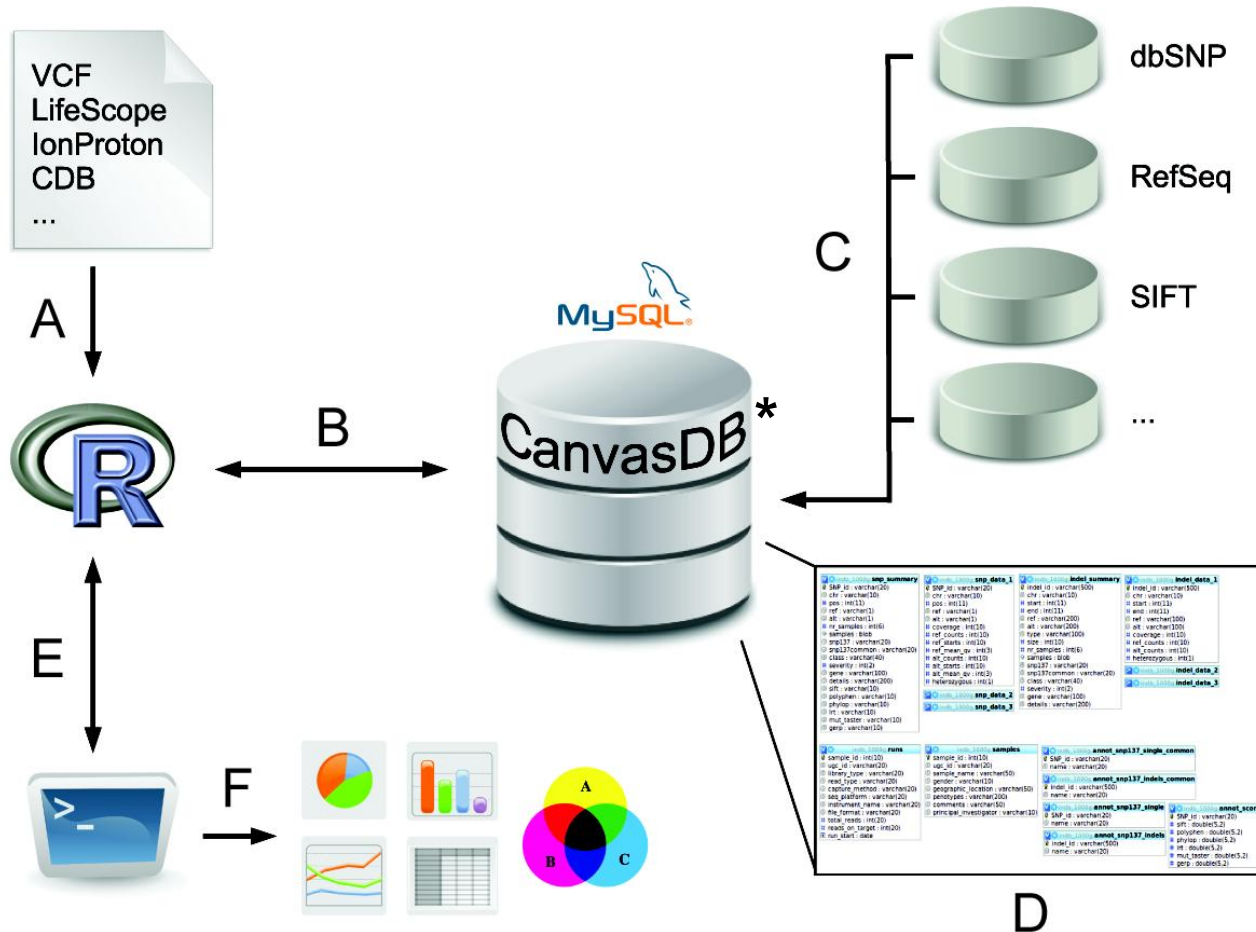


In-house development of pipelines

- 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

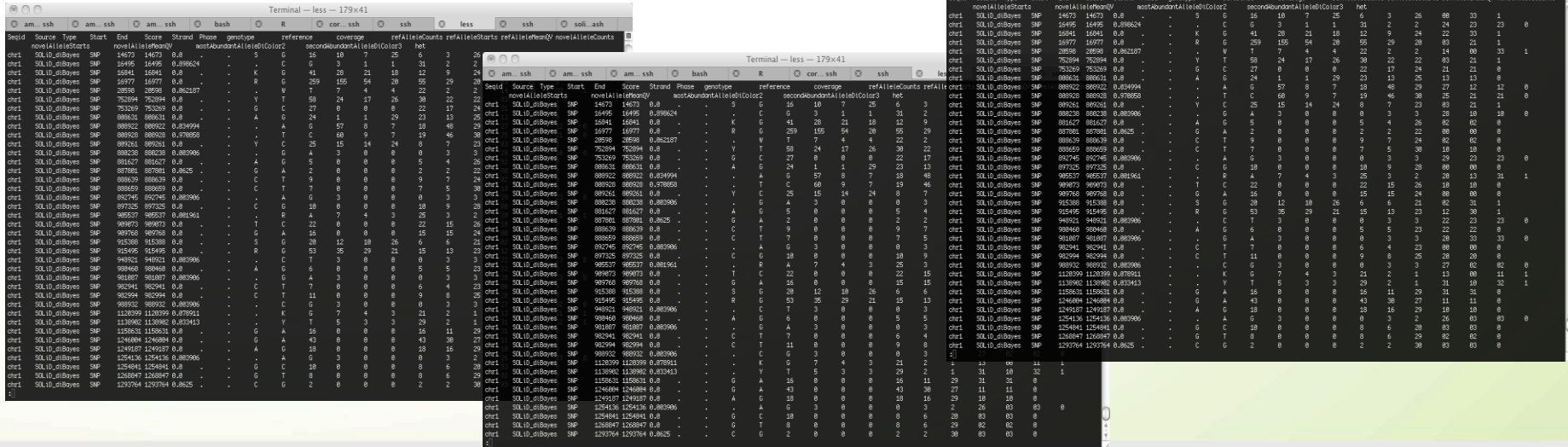


Background: exome-seq

- 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

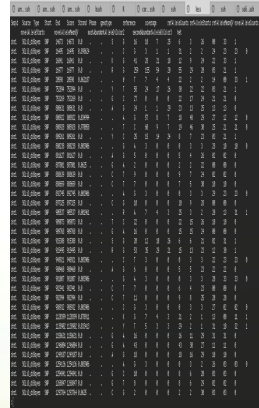


```
Terminal -- less - 179x41
Seid Source Type Start End Score Strand Phase genotype reference coverage refAlleleCounts refAlleleStarts refAlleleEnds refAlleleCounts
chr1 SOLiD_dBases SNP 14573 14573 0.8 - - - S G 16 10 7 25 6 3 26 00 33 1
chr1 SOLiD_dBases SNP 14595 14595 0.808624 - - - C G 3 1 1 31 2 2 24 23 23 0
chr1 SOLiD_dBases SNP 15591 15591 0.8 - - - K G 41 20 21 18 12 9 24 22 33 1
chr1 SOLiD_dBases SNP 15977 15977 0.8 - - - R G 259 155 54 20 95 29 43 24 21 83 2
chr1 SOLiD_dBases SNP 20598 20598 0.802107 - - - V T 7 4 4 22 2 2 14 00 33 1
chr1 SOLiD_dBases SNP 752094 752094 0.8 - - - Y T 58 24 17 26 39 22 22 83 21 1
chr1 SOLiD_dBases SNP 752095 752095 0.8 - - - R G 27 0 0 8 22 17 24 21 22 0
chr1 SOLiD_dBases SNP 888631 888631 0.8 - - - A G 24 1 1 29 23 13 25 13 13 0
chr1 SOLiD_dBases SNP 888632 888632 0.804994 - - - A G 57 0 7 19 46 39 22 83 21 1
chr1 SOLiD_dBases SNP 888633 888633 0.807658 - - - T C 60 9 7 19 46 38 22 83 21 1
chr1 SOLiD_dBases SNP 889261 889261 0.8 - - - Y C 25 15 14 24 8 7 23 23 13 0
chr1 SOLiD_dBases SNP 889262 889262 0.803966 - - - A G 3 0 0 8 3 3 24 23 13 0
chr1 SOLiD_dBases SNP 881627 881627 0.8 - - - A G 5 0 0 8 5 4 26 22 17 0
chr1 SOLiD_dBases SNP 887081 887081 0.8025 - - - G A 2 0 0 8 2 2 22 22 0 0
chr1 SOLiD_dBases SNP 888859 888859 0.8 - - - C T 9 0 0 8 9 7 24 23 13 0
chr1 SOLiD_dBases SNP 888859 888859 0.8 - - - C T 7 0 0 8 7 5 30 24 18 0
chr1 SOLiD_dBases SNP 888859 888859 0.803966 - - - A G 3 0 0 8 3 3 30 24 18 0
chr1 SOLiD_dBases SNP 888859 888859 0.803966 - - - G A 15 0 0 8 15 9 28 23 13 0
chr1 SOLiD_dBases SNP 892537 892537 0.801941 - - - P A 7 4 3 25 3 2 24 23 13 0
chr1 SOLiD_dBases SNP 899073 899073 0.8 - - - T C 22 0 0 8 22 15 26 26 6 21 0
chr1 SOLiD_dBases SNP 899760 899760 0.8 - - - G A 16 0 0 8 15 15 24 23 13 0
chr1 SOLiD_dBases SNP 915388 915388 0.8 - - - S G 28 12 18 26 6 21 21 15 13 0
chr1 SOLiD_dBases SNP 915495 915495 0.8 - - - R G 53 35 29 21 15 13 23 22 22 0
chr1 SOLiD_dBases SNP 915495 915495 0.8 - - - R G 53 35 29 21 15 13 23 22 22 0
chr1 SOLiD_dBases SNP 915495 915495 0.803966 - - - A G 6 0 0 8 5 5 23 22 22 0
chr1 SOLiD_dBases SNP 918167 918167 0.803966 - - - A G 6 0 0 8 5 5 23 22 22 0
chr1 SOLiD_dBases SNP 925941 925941 0.8 - - - C T 3 0 0 8 3 3 25 24 18 0
chr1 SOLiD_dBases SNP 925941 925941 0.8 - - - C T 11 0 0 8 9 8 25 24 18 0
chr1 SOLiD_dBases SNP 989926 989926 0.803966 - - - C T 6 7 0 8 6 7 21 21 21 0
chr1 SOLiD_dBases SNP 1120399 1120399 0.807658 - - - C T 3 0 0 8 3 3 29 22 2 1
chr1 SOLiD_dBases SNP 1139802 1139802 0.803413 - - - Y T 5 3 3 29 2 1 29 22 2 1
chr1 SOLiD_dBases SNP 1152036 1152036 0.8 - - - G A 43 0 0 8 43 38 27 11 13 0
chr1 SOLiD_dBases SNP 1246094 1246094 0.8 - - - G A 43 0 0 8 43 38 27 11 13 0
chr1 SOLiD_dBases SNP 1249287 1249287 0.8 - - - A G 18 0 0 8 18 16 29 18 10 0
chr1 SOLiD_dBases SNP 1254941 1254941 0.803966 - - - G A 3 0 0 8 3 3 29 18 10 0
chr1 SOLiD_dBases SNP 1254941 1254941 0.803966 - - - G C 18 0 0 8 6 6 29 18 10 0
chr1 SOLiD_dBases SNP 1258471 1258471 0.8025 - - - G T 8 0 0 8 6 6 29 83 03 0
chr1 SOLiD_dBases SNP 1259764 1259764 0.8025 - - - C G 2 0 0 8 2 2 30 83 03 0
chr1 SOLiD_dBases SNP 1259764 1259764 0.8025 - - - C G 2 0 0 8 2 2 30 83 03 0
```

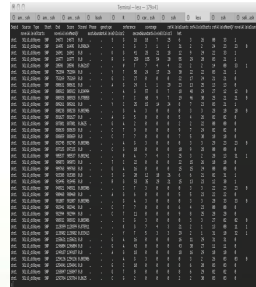
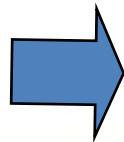
How to analyze this?

- 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)

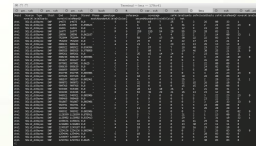
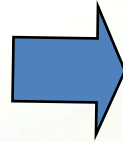
Start:
All identified SNPs



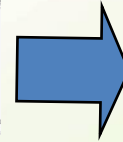
A screenshot of a text file containing a list of identified SNPs. The file is titled 'SNPs.txt' and contains a large number of lines, each representing a SNP. The columns include SNP ID, chromosome, position, and other genomic information.



A screenshot of a text file showing the first result of the analysis. The file is titled 'SNPs.txt' and contains a list of SNPs that have been filtered based on the criteria mentioned in the text.



A screenshot of a text file showing the second result of the analysis. The file is titled 'SNPs.txt' and contains a list of SNPs that have been further filtered based on the criteria mentioned in the text.



Result:
A few candidate
causative
SNP(s)!

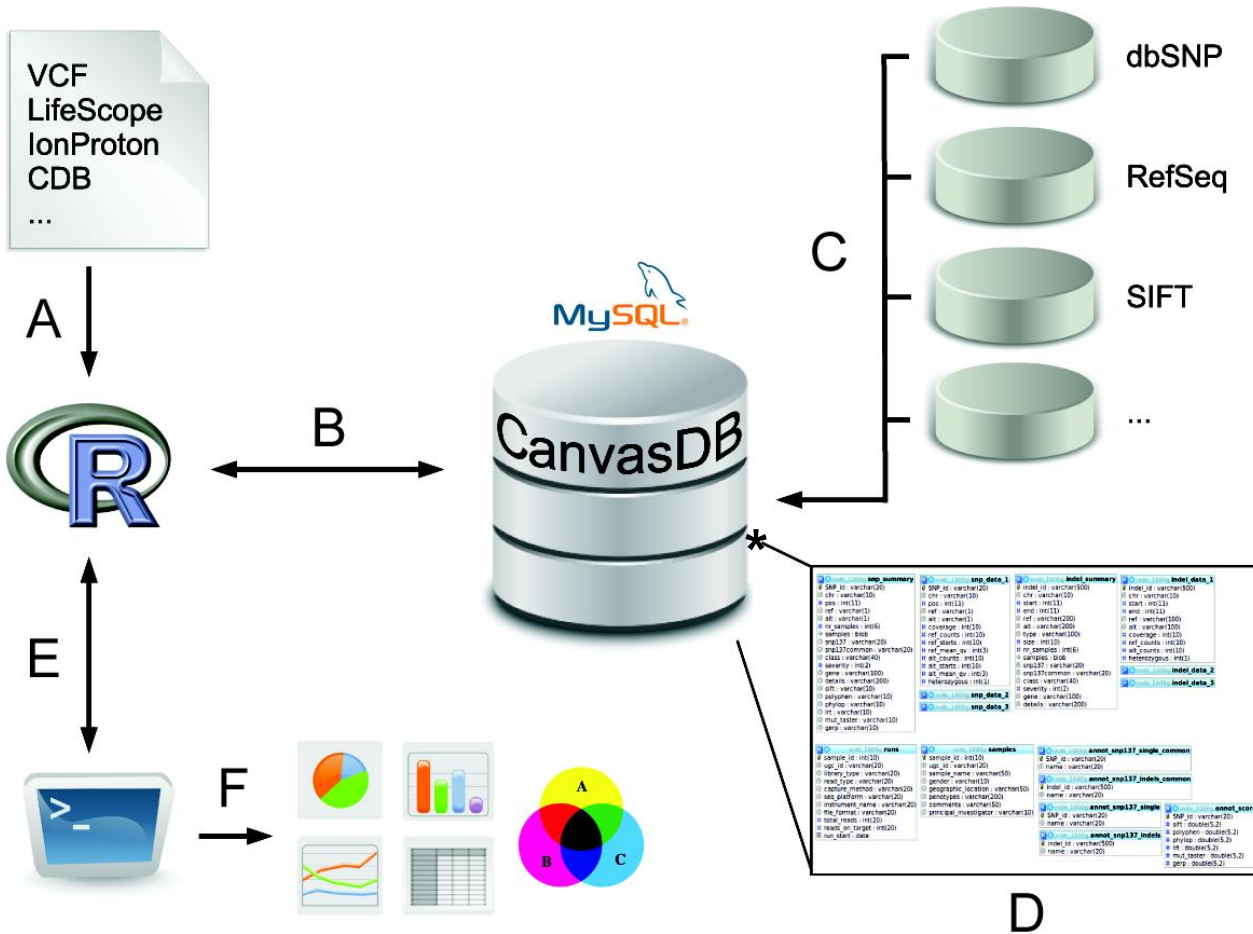


A screenshot of a text file showing the final result of the analysis. The file is titled 'SNPs.txt' and contains a list of SNPs that are considered candidate causative SNPs.

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

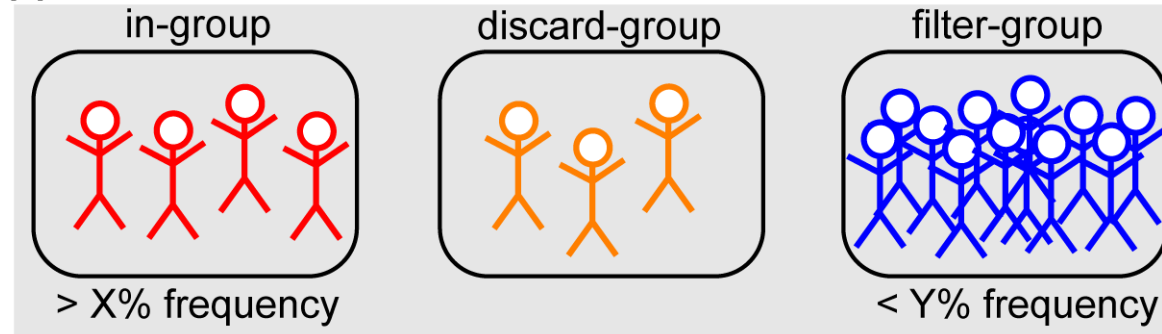


***CANdicate Variant Analysis System and Data Base**

Ameur et al., Database Journal, 2014

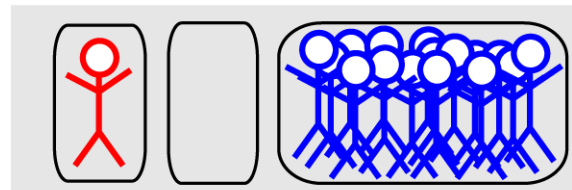
CanvasDB - Filtering

A

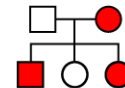


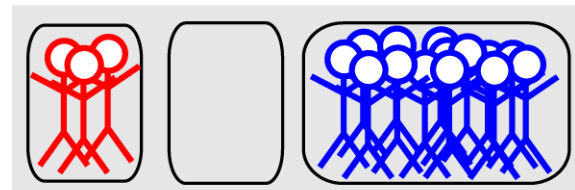
B

parent-offspring trio 



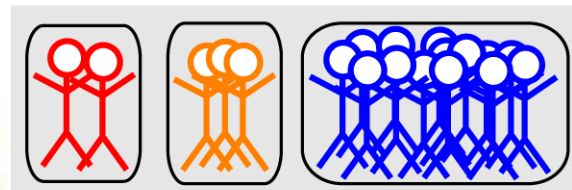
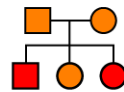
C

dominant variant 



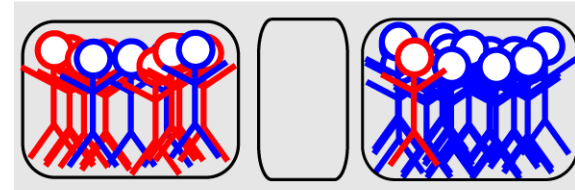
D

recessive variant



E

comparing groups

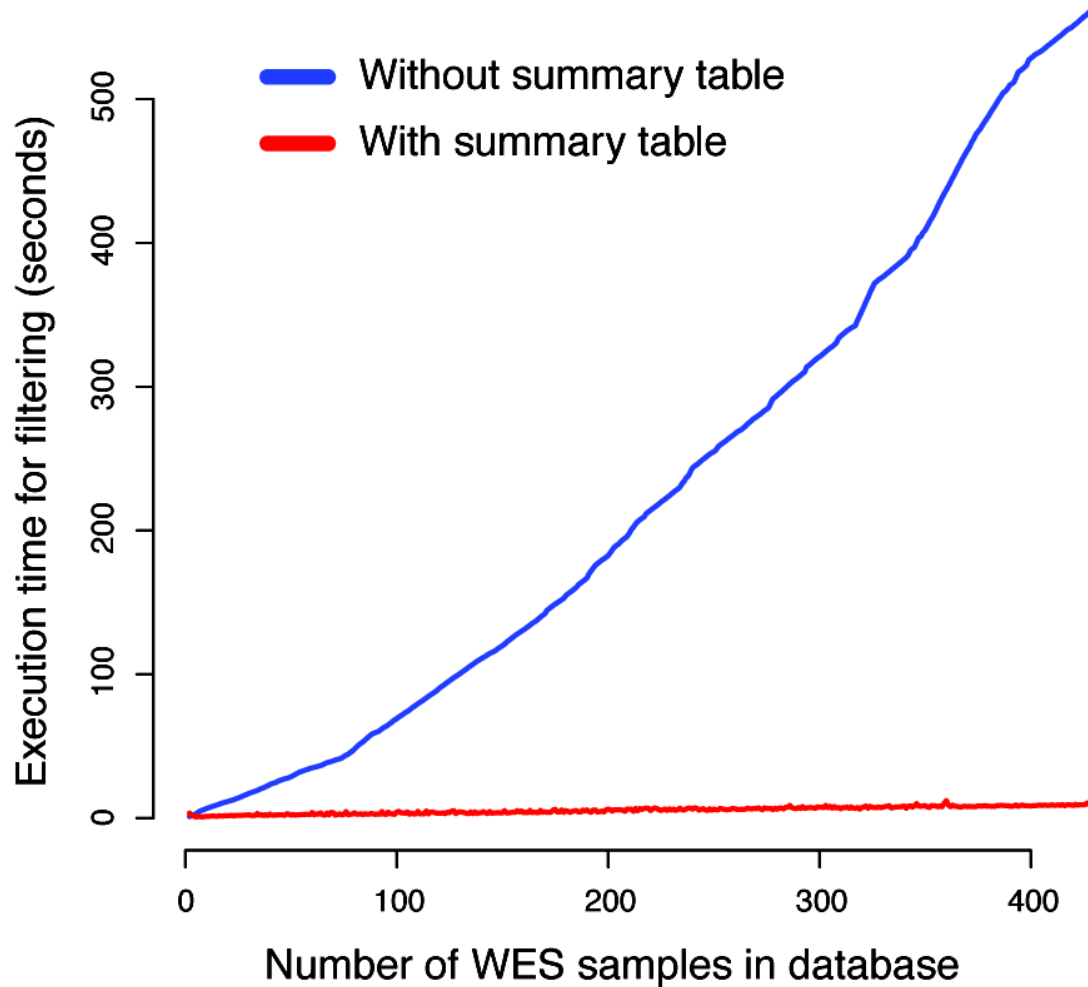


> min freq g1

< max freq g2

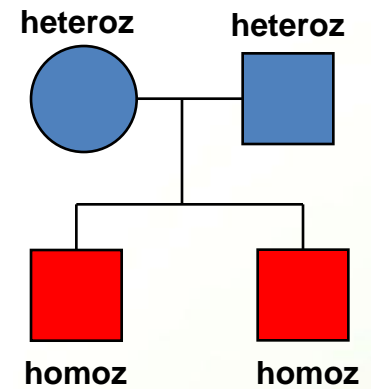
CanvasDB - Filtering speed

- Rapid variant filtering, also for large databases



A recent exome-seq project

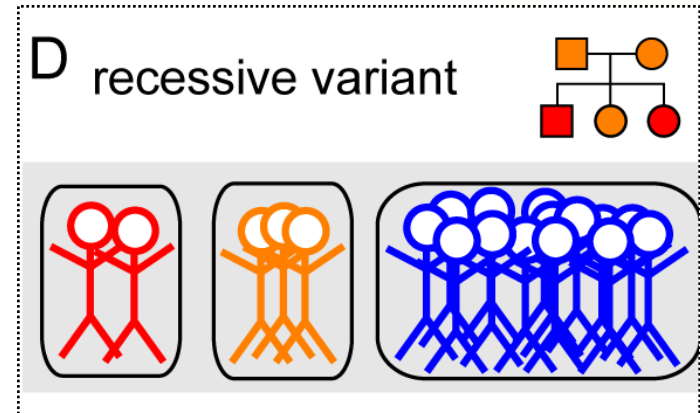
- Hearing loss: 2 affected brothers
 - Likely a rare, recessive disease
 - => Shared homozygous SNPs/indels
- Sequencing strategy
 - TargetSeq exome capture
 - One sample per PI chip



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

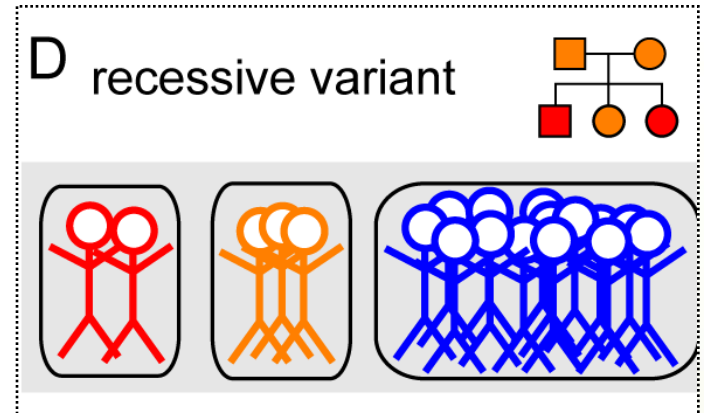
- *CanvasDB* filtering for a variant that is...
 - rare
 - at most in 1% of ~700 exomes
 - shared
 - found in both brothers
 - homozygous
 - in brothers, but in no other samples
 - deleterious
 - non-synonymous, frameshift, stop-gain, splicing, etc..



```
> 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

Unrelated sample

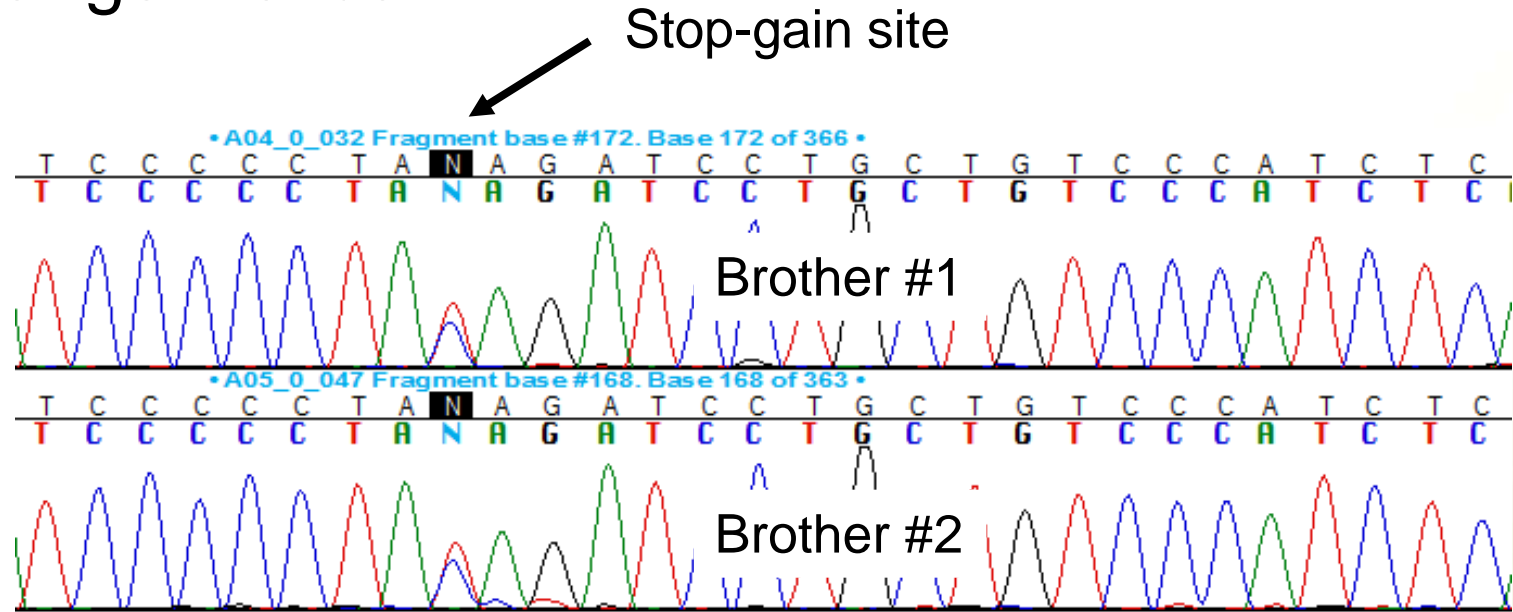
Brother #1

Brother #2

G A G A T G G G A C A G C A G G A T C T G T A G G G G A T C T G T C G T G T
L H S L L I Q L P I Q R T

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

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

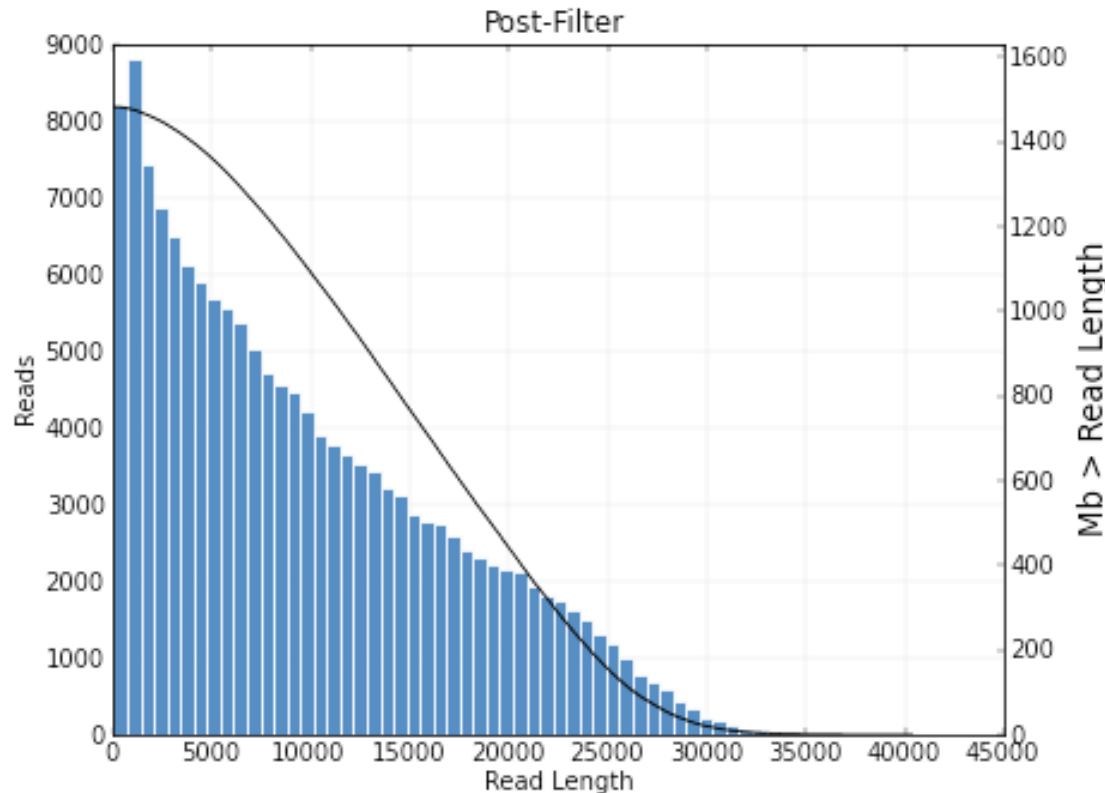


Long reads (seeds) →



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!!

The screenshot shows the PacBio SMRT Portal interface. The browser address bar displays the URL `127.0.0.1:8080/smrtportal/#/Design-Job/Details-of-Job/16497`. The page title is "Details of Job assembly". The navigation bar includes "SMRT® Portal", "Home", "Admin", "Tech Support Files", "Help", and "About". The user is logged in as "ugc_admin".

The main content area is divided into three tabs: "DESIGN JOB", "MONITOR JOBS", and "VIEW DATA". The "DESIGN JOB" tab is active, showing the following details:

- Job Name: assembly
- Comments: [Empty field]
- Groups: all
- User: ugc_admin
- Protocols: RS_HGAP_Assembly.3
- Reference: [None selected]

Below these details are two tables:

SMRT Cells Available (Viewing 1 - 31 of 31)

Sample	Version	User	Groups	Started	Uri
Pb9_frax 21	2.0.2		all	2014-02-20T19:28:20+0000	/home/pacbio/...
Pb9_frax 44	2.0.2		all	2014-02-20T19:28:20+0000	/home/pacbio/...
Pb9_frax 63	2.0.2		all	2014-02-20T19:28:20+0000	/home/pacbio/...
Pb33_1	2.0.2		all	2014-02-20T19:28:20+0000	/home/pacbio/...
Pb33_2	2.0.2		all	2014-02-20T19:28:20+0000	/home/pacbio/...
Pb 33-5	2.0.2		all	2014-02-24T13:48:09+0000	/home/pacbio/...
Pb 33-7	2.0.2		all	2014-02-24T13:48:09+0000	/home/pacbio/...
Pb 33-6	2.0.2		all	2014-02-24T13:48:09+0000	/home/pacbio/...
Pb 33-3	2.0.2		all	2014-02-24T13:48:09+0000	/home/pacbio/...
Pb 33-9	2.0.2		all	2014-02-24T13:48:09+0000	/home/pacbio/...
Pb 33-8	2.0.2		all	2014-02-24T13:48:09+0000	/home/pacbio/...
Pb 33-4	2.0.2		all	2014-02-24T13:48:09+0000	/home/pacbio/...
Pb 33-10	2.0.2		all	2014-02-24T13:48:09+0000	/home/pacbio/...
Pb55_f2rpt	2.1.0		all	2014-05-09T10:48:14+0000	/home/pacbio/...
Pb_46_3_repeat	2.1.0		all	2014-05-09T10:48:14+0000	/home/pacbio/...
Pb55_f2rpt	2.1.0		all	2014-05-09T10:48:14+0000	/home/pacbio/...
Pb_46_9	2.1.0		all	2014-05-09T10:48:14+0000	/home/pacbio/...
Pb_46_10	2.1.0		all	2014-05-09T10:48:14+0000	/home/pacbio/...
Pb46_3	2.1.0		all	2014-05-08T11:08:49+0000	/home/pacbio/...
Pb46_5	2.1.0		all	2014-05-08T11:08:49+0000	/home/pacbio/...

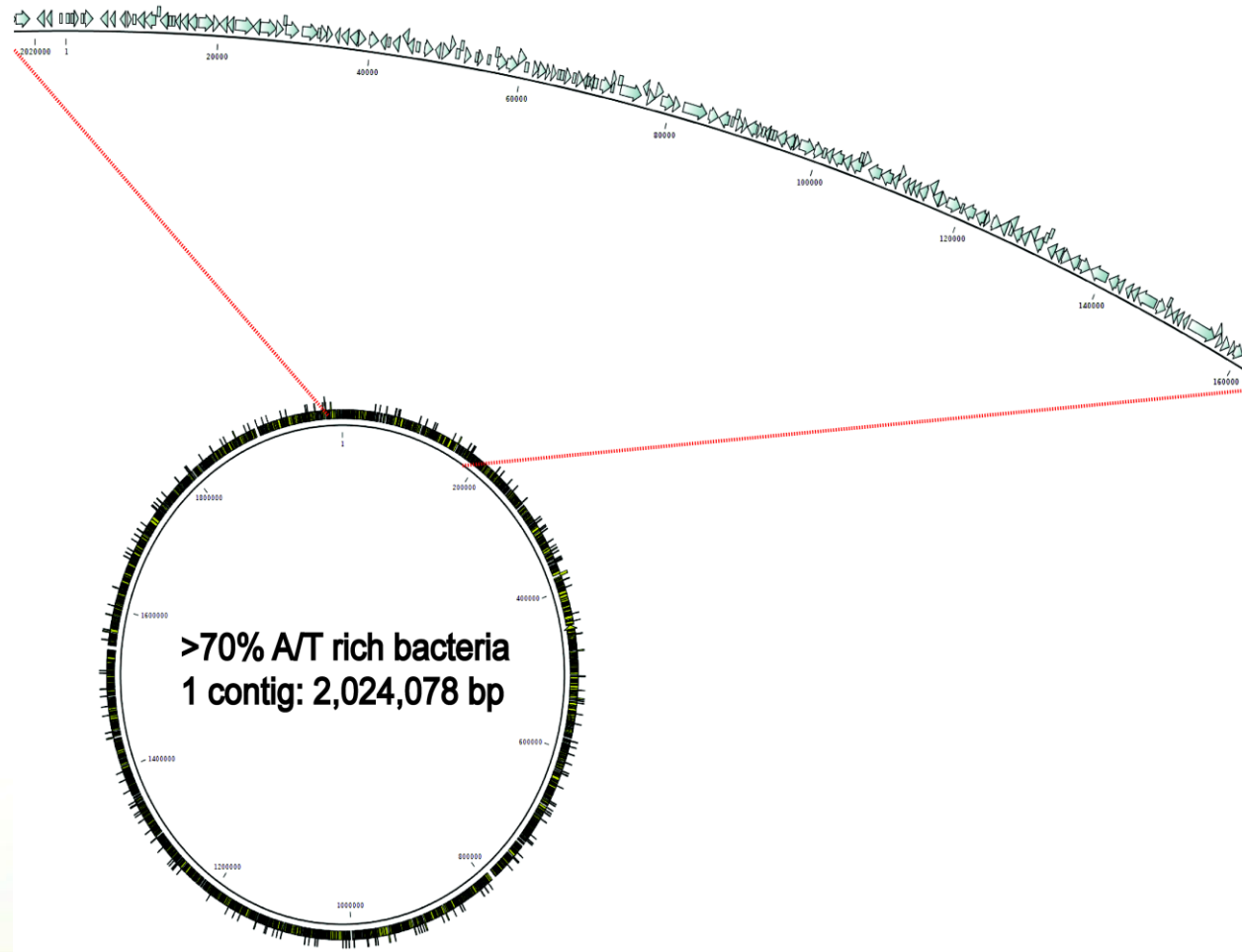
SMRT Cells in Job (Viewing 1 - 1 of 1)

Sample	Version	User	Groups	Uri
Pb33_1	2.0.2		all	/home/pacbio/DATA/adam/Pb_33_F...

At the bottom of the interface, there are buttons for "Start", "Save", "Copy", and "Cancel".

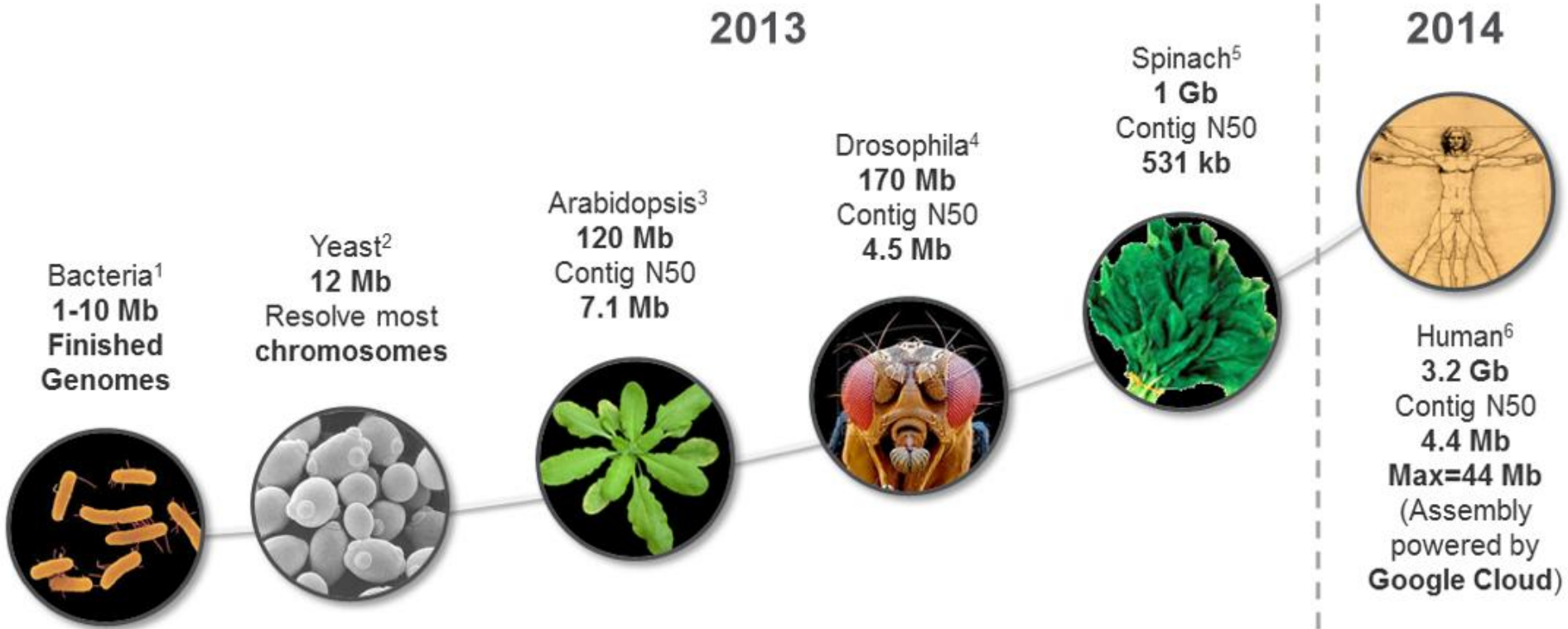
PacBio assembly, example result

- 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

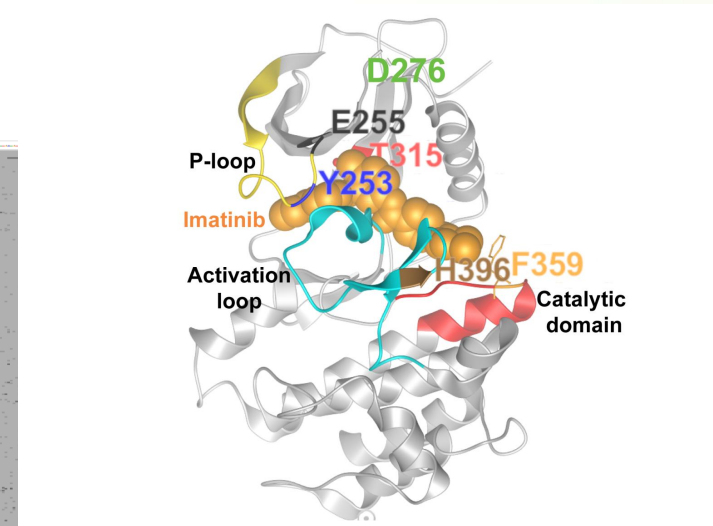
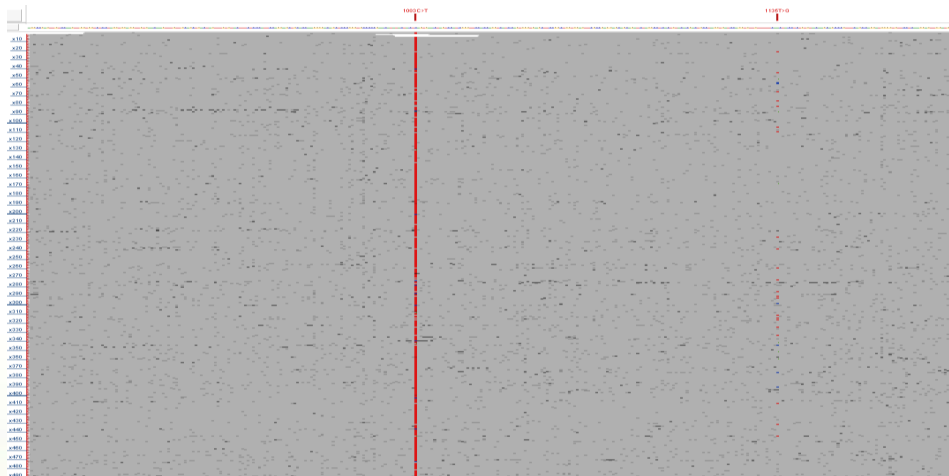
We are pleased to make publicly available a new shotgun sequence dataset of long PacBio® reads from a human DNA sample. We previously released sequence data using Single Molecule, Real-Time (SMRT®) Sequencing of ~10x coverage of this sample, sufficient for reference-based detection of structural variation. 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!

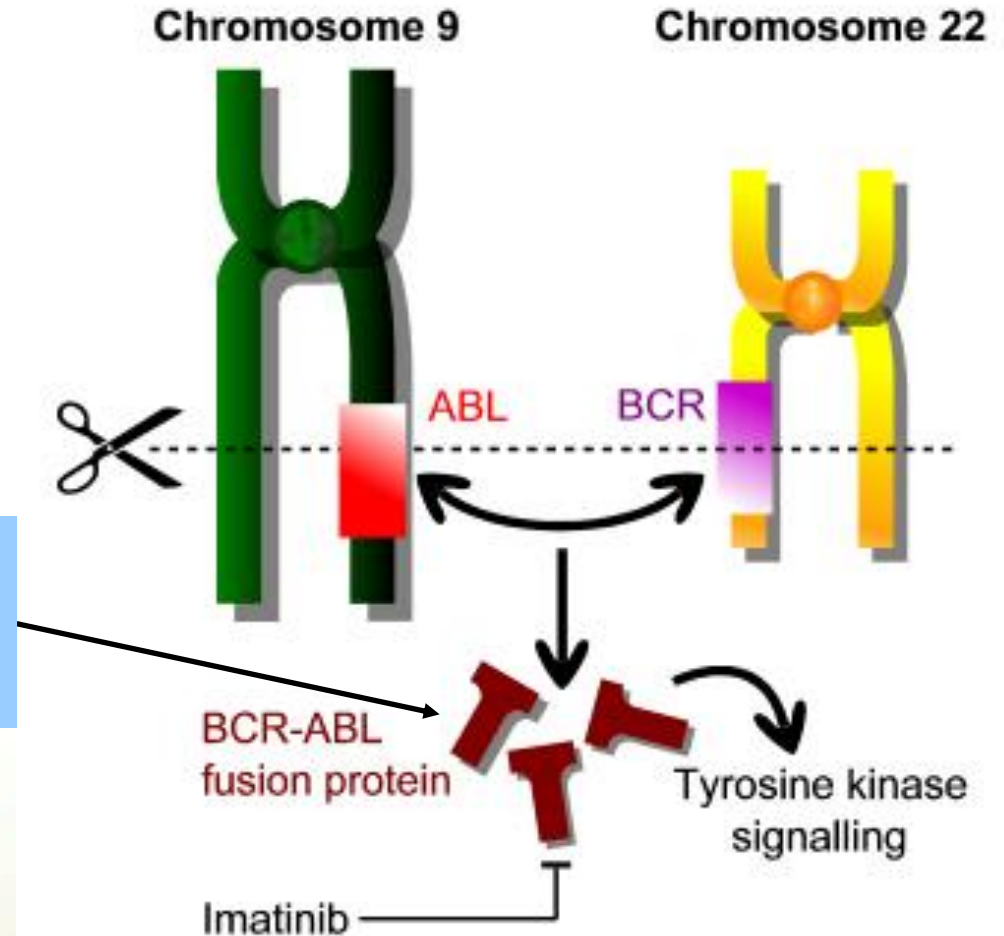
- We need to install such pipelines at UPPNEX!!

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

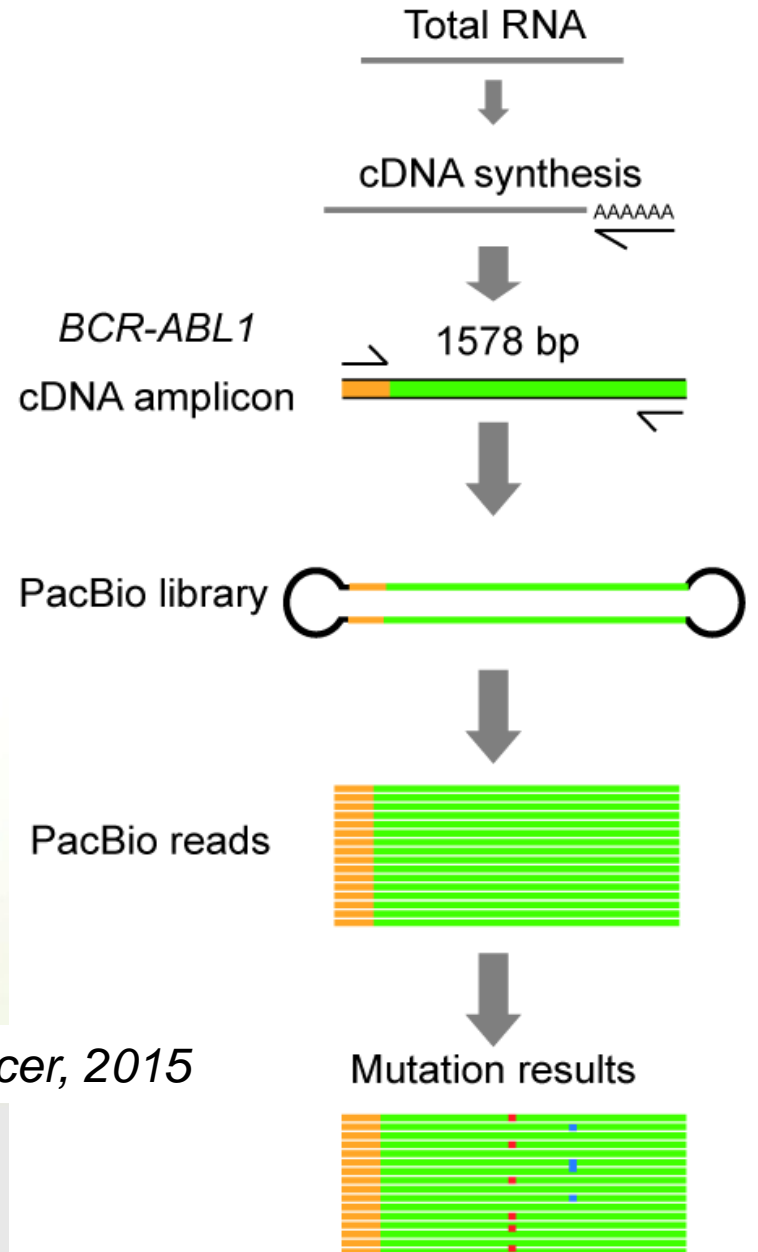
www.cambridgemedicine.org/article/doi/10.7244/cmj-1355057881

BCR-ABL1 workflow – PacBio Sequencing

From sample to results: < 1 week



1 sample/SMRT cell



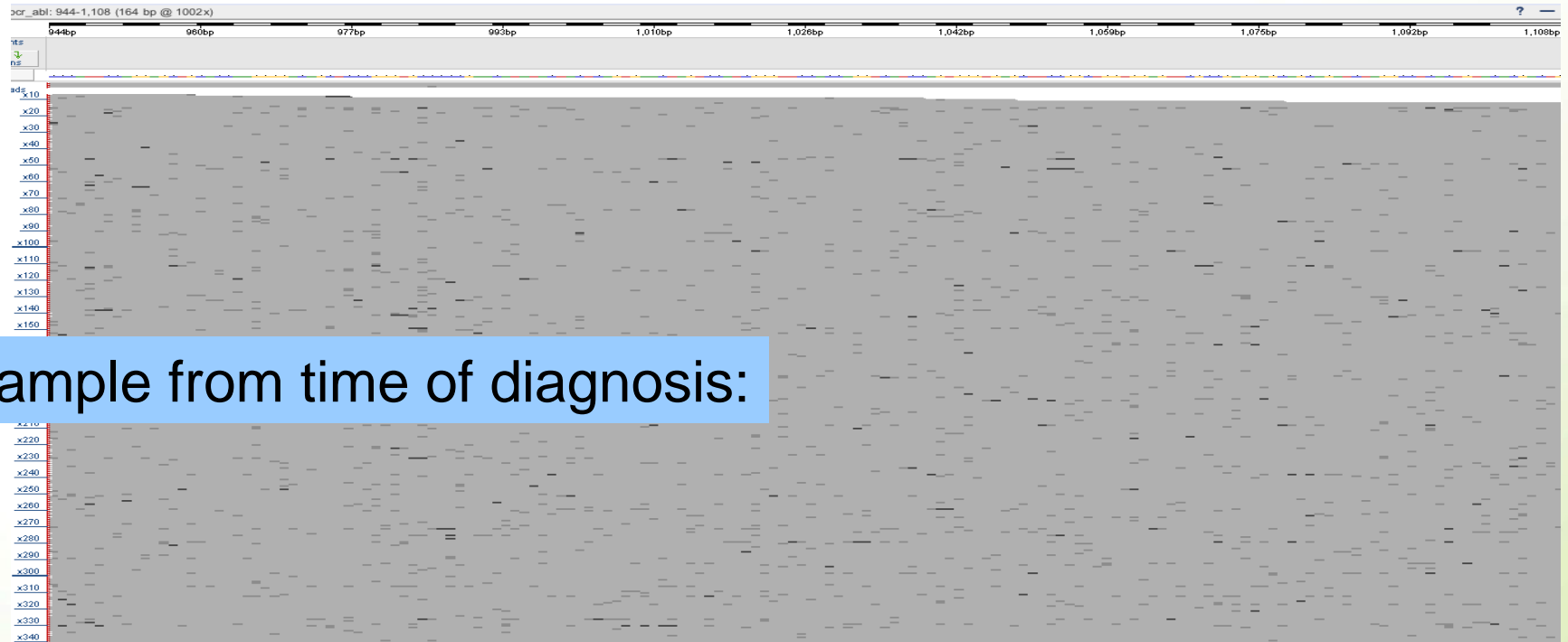
Cavelier et al., BMC Cancer, 2015

BCR-ABL1 mutations at diagnosis

PacBio sequencing generates ~10 000X coverage!

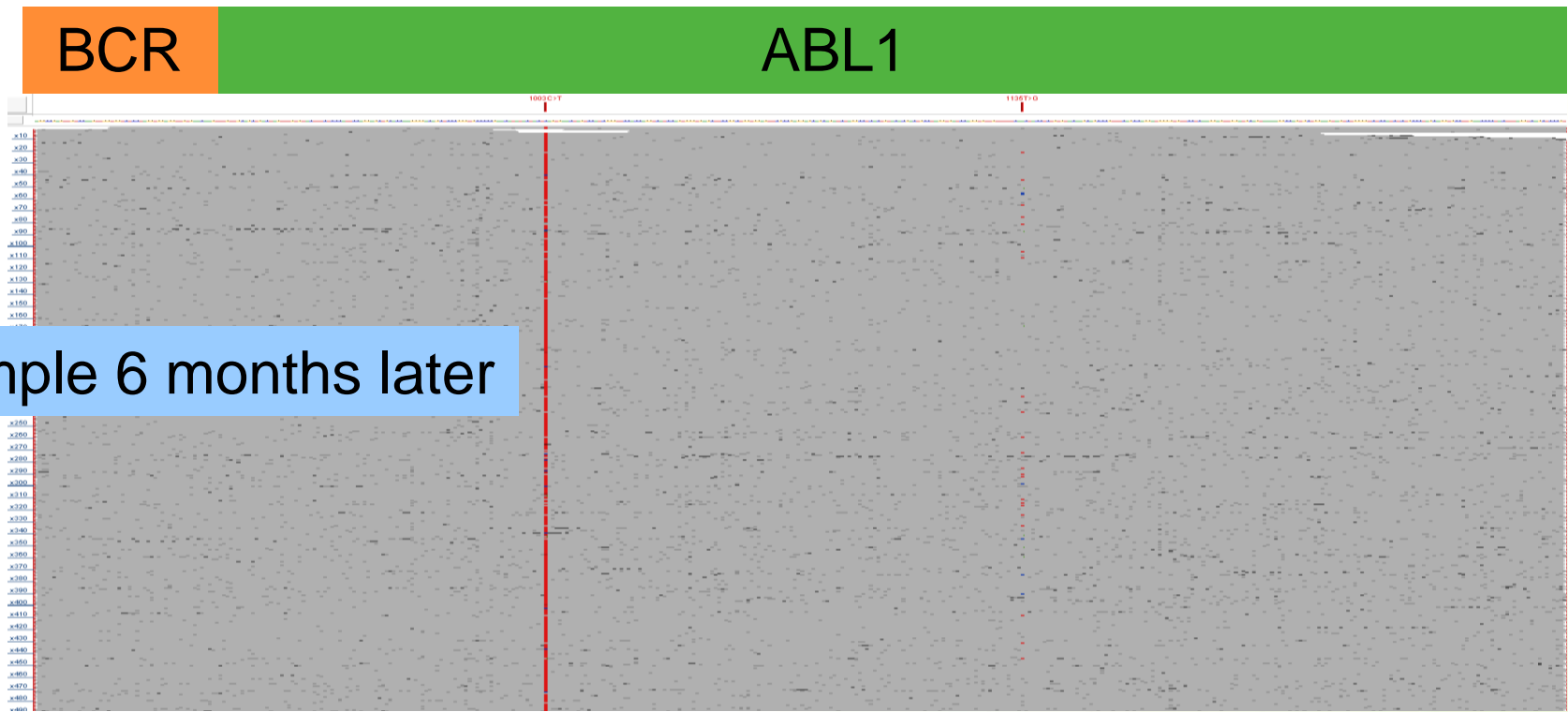
BCR

ABL1



Sample from time of diagnosis:

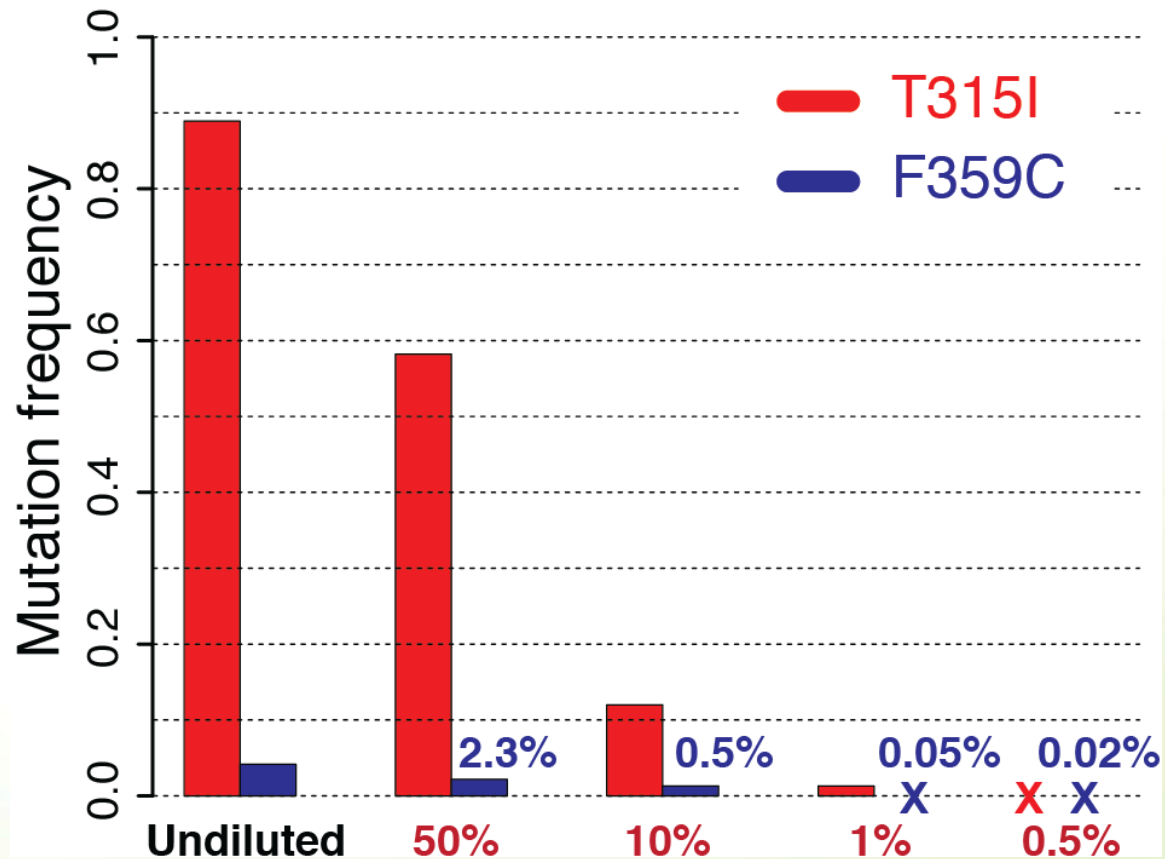
BCR-ABL1 mutations in follow-up sample



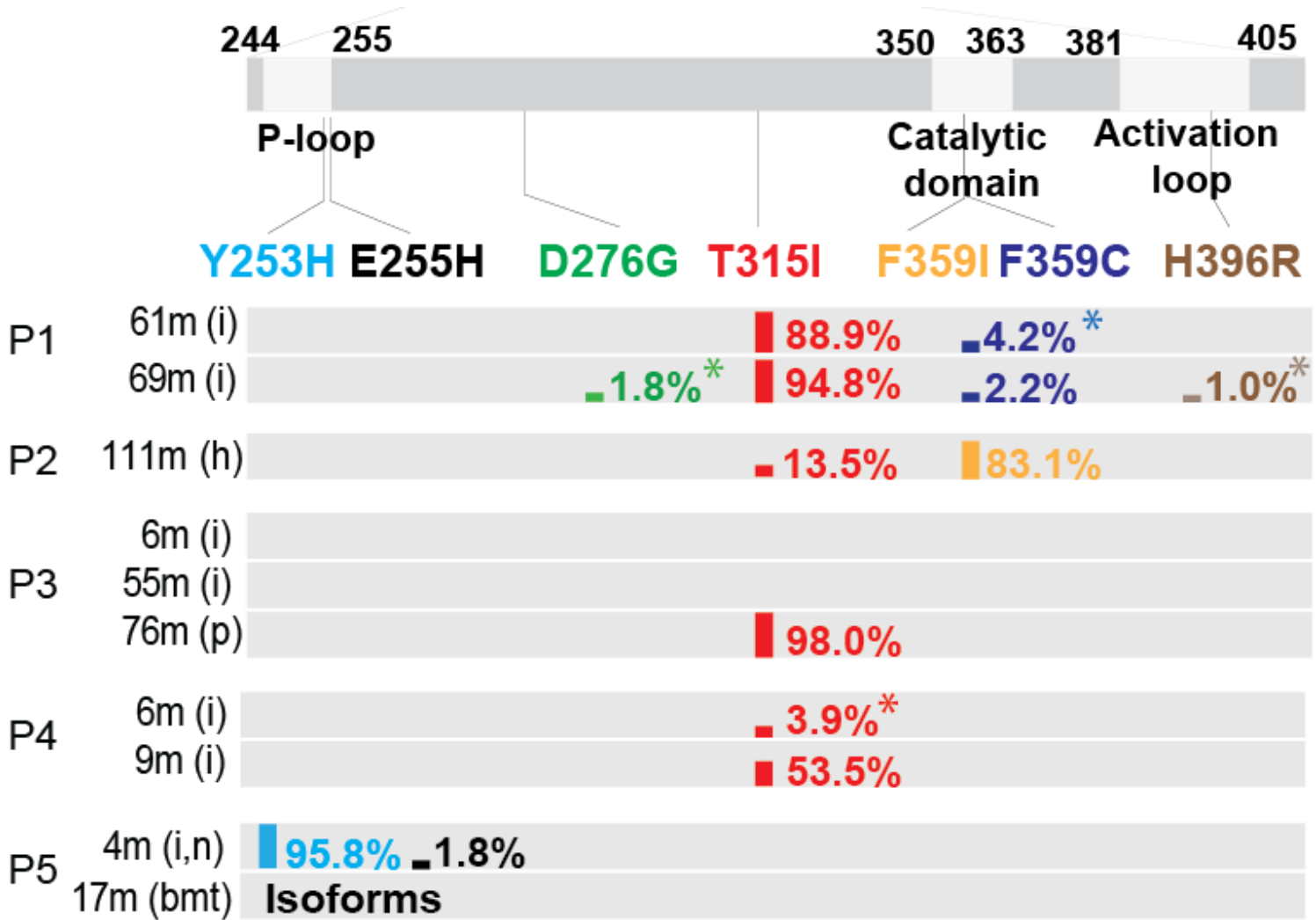
Mutations acquired in fusion transcript.
Might require treatment with alternative drug.

BCR-ABL1 dilution series results

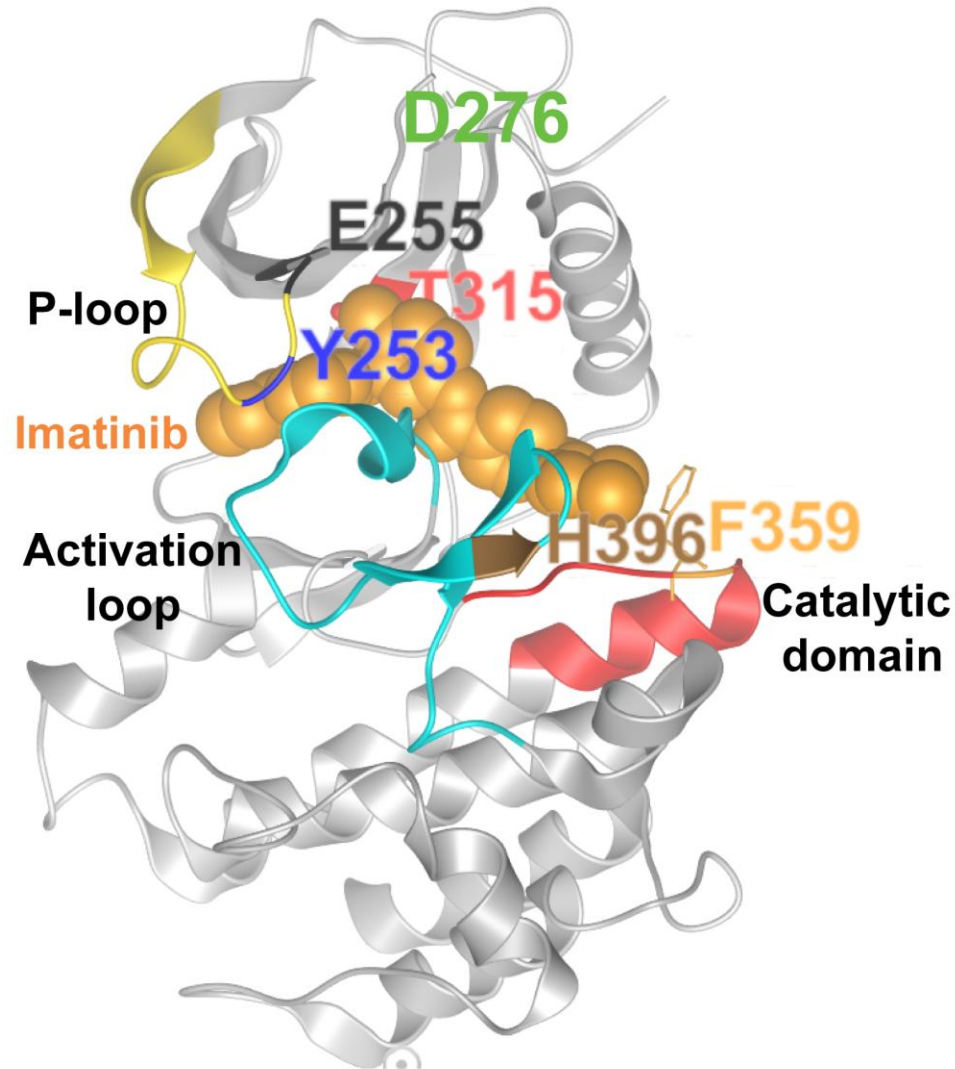
- Mutations down to 1% detected!



Summary of mutations in 5 CML patients



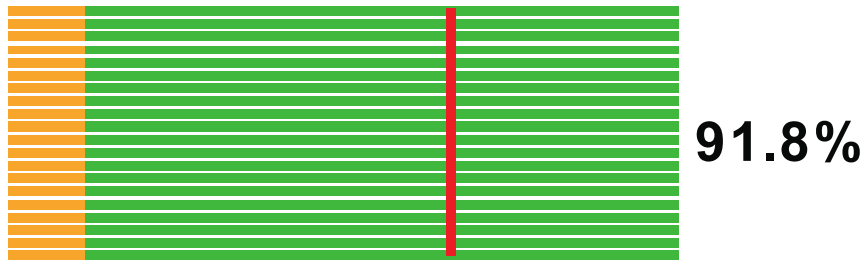
Mutations mapped to protein structure



BCR-ABL1 - Compound mutations

P1 61m

T315I

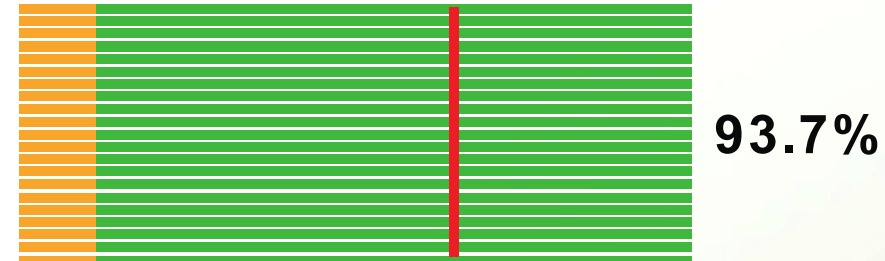


F359C



P1 68.5m

T315I



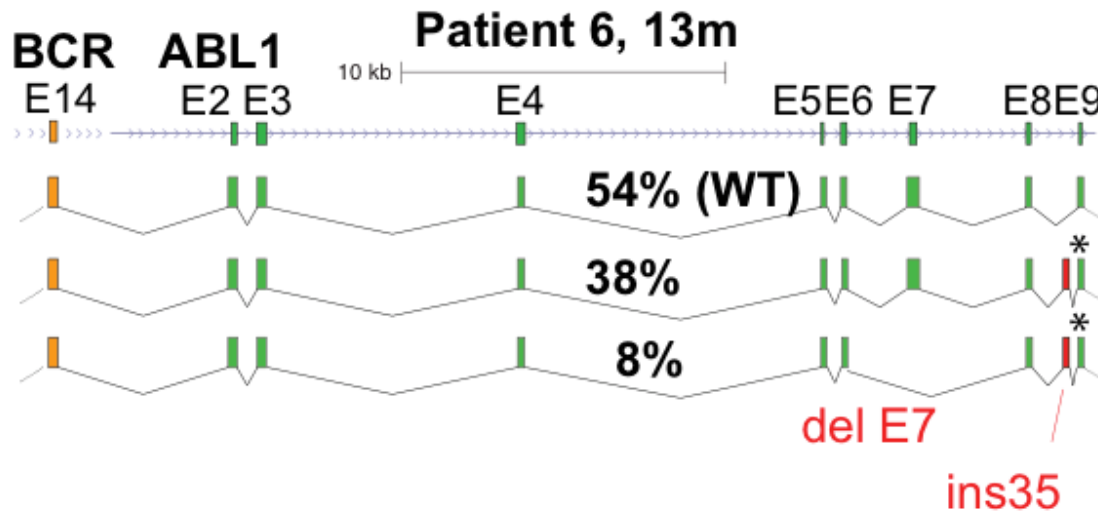
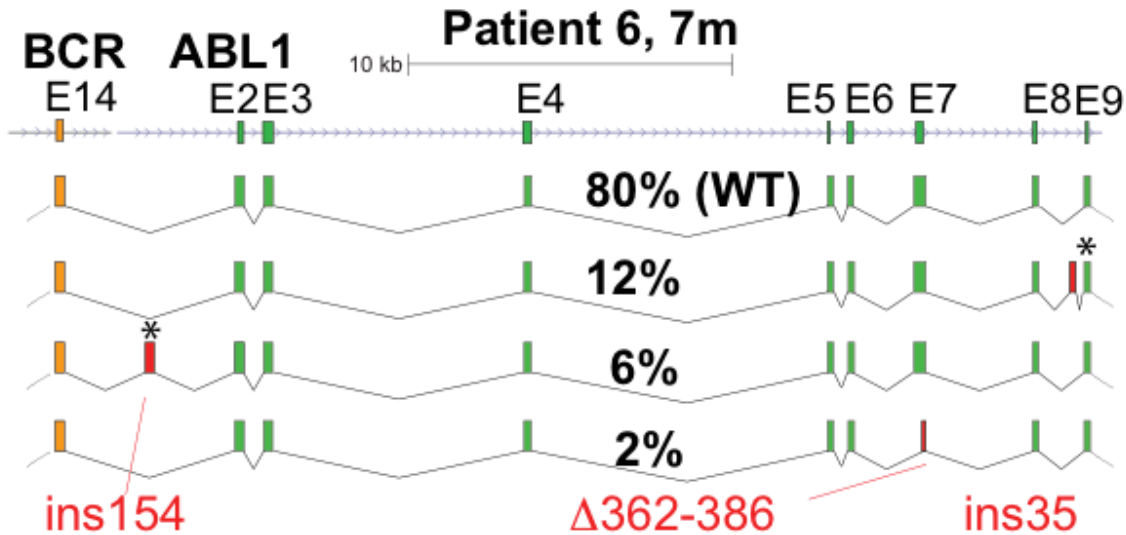
D276G



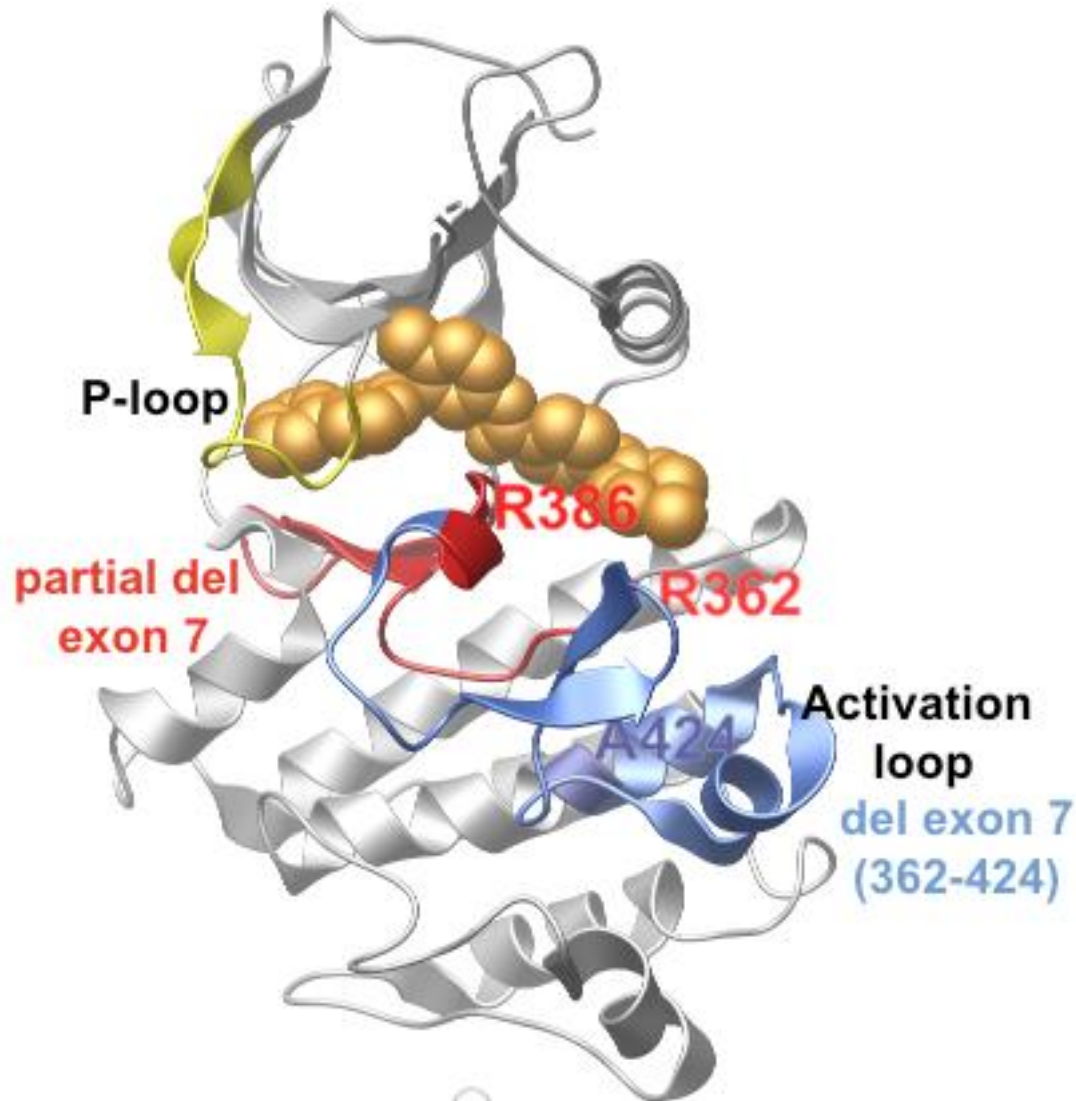
H396R



BCR-ABL1 - Multiple isoforms in one individual!

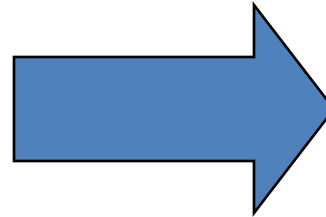


BCR-ABL1 – Isoforms and protein structure



Next step: A clinical diagnostics pipeline!

Step1. Create CCS reads



SMRT® Portal Home Admin Help About

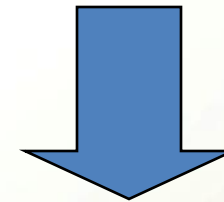
DESIGN JOB MONITOR JOB

Job Name: _____ Comments: _____

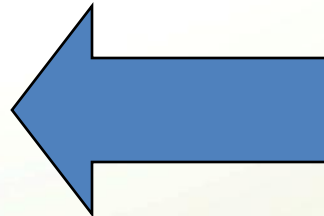
Protocol: [None selected] Reference: [None selected]

SMRT Cells Available (Viewing 1 - 50 of 62)

SMRT Cell ID	Sample	V	User	Groups	Started	Uri
1005798525500000182308820	pb_2	v2	all	all	2013-10-10T09:45:16+0	hor
1005798525500000182308820	pb_2	v2	all	all	2013-10-10T09:45:16+0	hor
1005798525500000182308820	pb_2	v2	all	all	2013-10-10T09:45:16+0	hor
1005798525500000182308820	pb_2	v2	all	all	2013-10-10T09:45:16+0	hor
1005798525500000182308820	pb_4	v2	all	all	2013-10-09T16:37:34+0	hor
1005798525500000182308820	pb_4	v2	all	all	2013-10-09T16:37:34+0	hor
1005798525500000182308820	pb_3-2	v2	all	all	2013-10-09T16:37:34+0	hor
1005798525500000182308820	pb_3-1	v2	all	all	2013-10-09T16:37:34+0	hor
1005797325500000182308820	pb_1-8	v2	all	all	2013-09-12T13:25:48+0	hor
1005797325500000182308820	pb_1-7	v2	all	all	2013-09-12T13:25:48+0	hor
1005797325500000182308820	pb_1-6	v2	all	all	2013-09-12T13:25:48+0	hor
1005797325500000182308820	pb_1-5	v2	all	all	2013-09-12T13:25:48+0	hor



Step3. Upload to result server



Step2. Run mutation analysis

E25K	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	12784	1679	0.117	13545	1734	0.113	26249	3413	0.115	positive
M244V	GGACATCACC[A/G]TGAAGCACAA	14209	1558	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]TGTATGACAGG	13869	321	0.024	13977	403	0.028	27846	724	0.026	positive
K247R	ATGAAGCAC[A/G]GCTGGCGGG	13901	14	0.001	14805	8	0.001	28786	22	0.001	negative
L248V	GAAGCACAG[C/G]TGGCGGGGG	13708	0	0	14823	0	0	28531	0	0	negative
G258E	AAGCTGGGCG[G/A]GGCGGAGT	13338	2	0	14453	4	0	27783	6	0	negative
Q252H	CGGGGGCCCA[G/T]TACGGGGAGG	6895	0	0	7489	2	0	14384	2	0	negative
Y253F	CGGGGGCCAG[T/C]ACGGGGAGGT	6877	1	0	7439	1	0	14316	2	0	negative
Y253F	GGGGGCCAGT[A/T]JCGGGGAGGTG	7146	0	0	7721	0	0	14867	0	0	negative
E255V	CAGTACGGG[A/T]JGGTACGAG	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	CTTGAAGGAC[G/A]ACACATGGA	15772	5	0	16786	9	0.001	32558	14	0	negative
D276G	TTGAAGGAGG[A/G]CCACCATGG	15840	40	0.003	16855	37	0.002	32695	77	0.002	negative
T277P	GAAGGAGGAC[A/G]CCATGAGGT	15786	10	0.001	16815	26	0.002	32601	36	0.001	negative
T277S	GAAGGAGGAC[A/T]CCATGAGGT	15786	1	0	16815	0	0	32601	1	0	negative
T277N	AAGGAGGACA[C/A]TCATGGAGGTG	15899	2	0	16939	2	0	32838	4	0	negative

Reporting system for mutation results

Sample ID <input type="text"/>	Run ID <input type="text"/>	Date (yyyy-mm-dd) <input type="text"/>
Mutation <input type="text"/>	Sequenced <input type="text"/>	

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

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

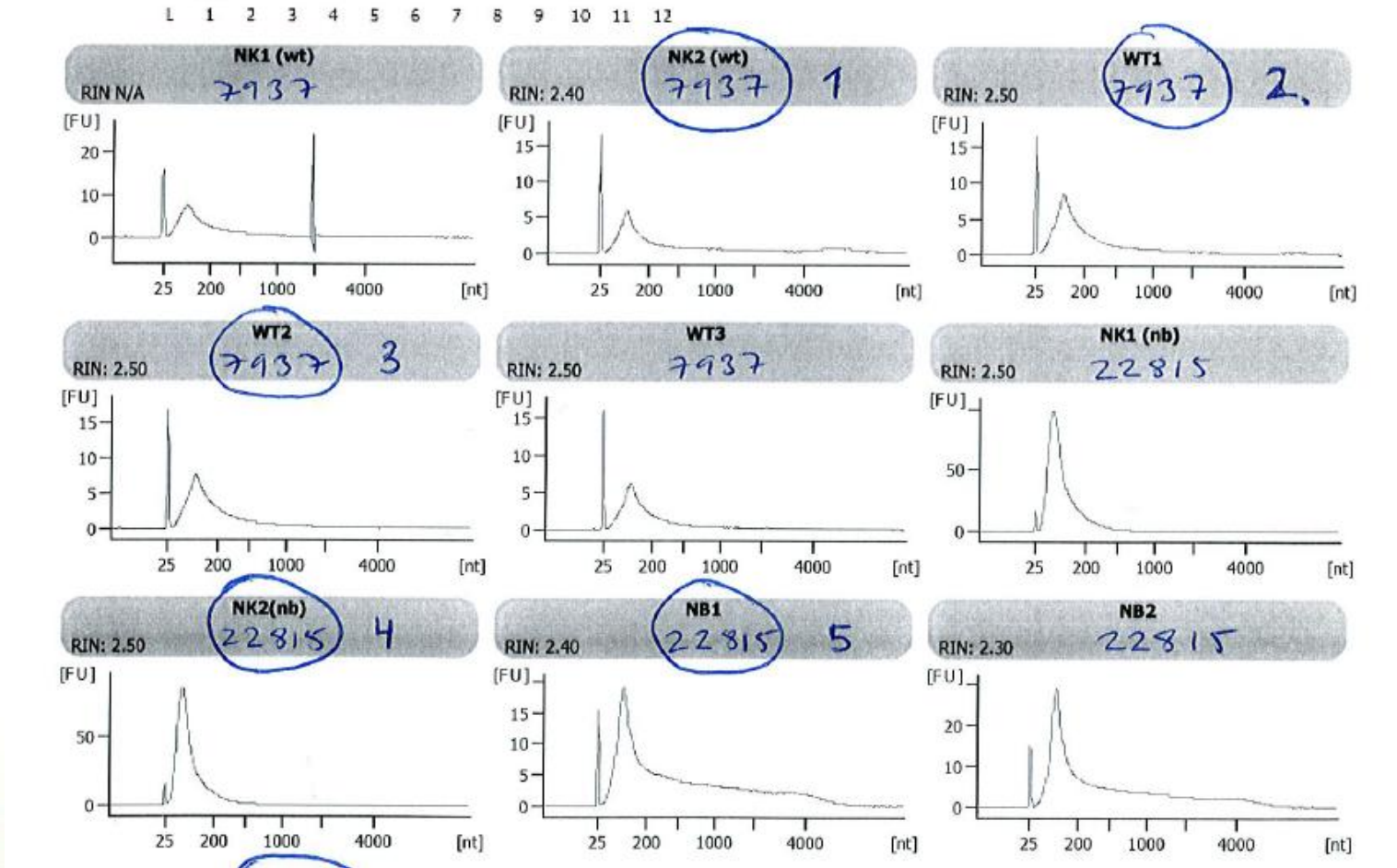
	A	B	C	D	E
	Gene	Target	IonXpress_001	IonXpress_002	IonXpress_010
1	SEC24B-AS1	AMPL377418	0.96	1.568	1.369
2	A1BG	AMPL174256	0.107	0	0.152
3	A1CF	AMPL365934	0	0	0
4	GGACT	AMPL173676	0.213	0.922	1.065
5	A2M	AMPL1384	77.965	110.653	0
6	A2ML1	AMPL359429	132.679	160.723	0
7	A2MP1	AMPL376317	0.213	0.184	0.076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

- HiQ chemistry

- Improves accuracy in sequencing
- Reduces indel error rates

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!



INTERNATIONAL CTAC
AT CAGENOMIC SGT
INFRASTRUCTURE

SciLifeLab