

Next Generation Sequencing and Bioinformatics Analysis Pipelines

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

Today's lecture

- Management of NGS data at NGI/SciLifeLab
- Examples of analysis pipelines:
 - Human exome & whole genome sequencing
 - Assembly using long reads
 - Clinical routine sequencing

illumina®



ThermoFisher Scientific
life ion torrent
life technologies

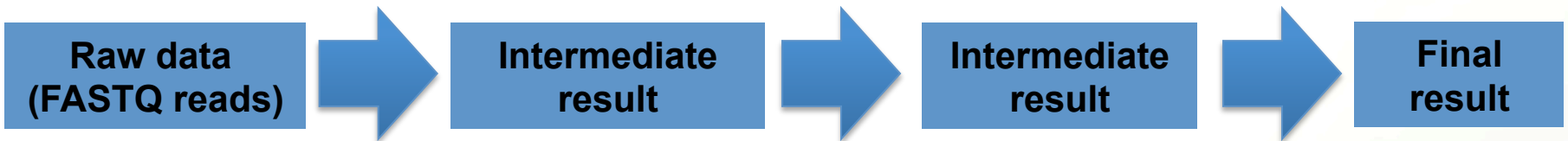


PACIFIC BIOSCIENCES®

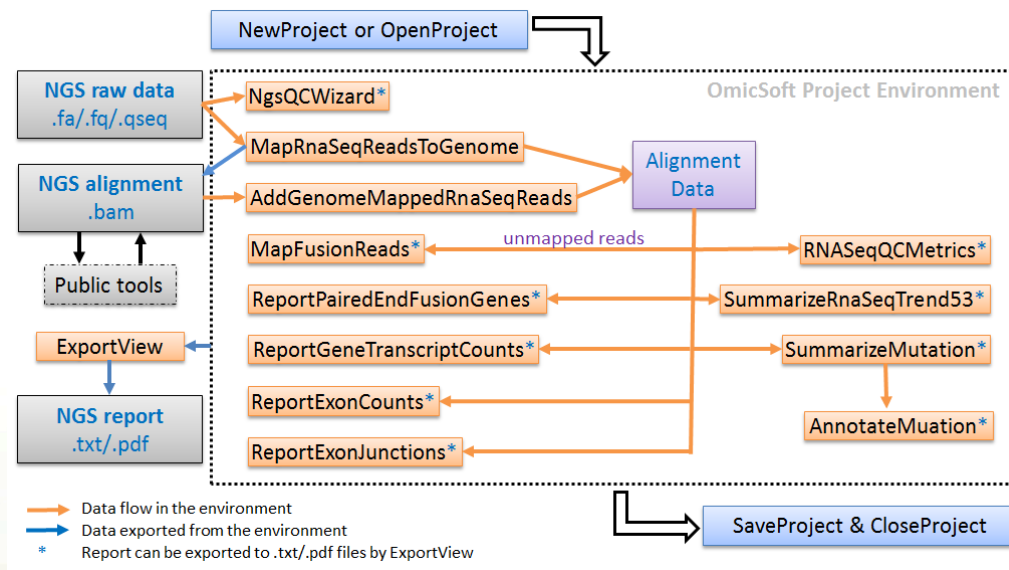


What is an analysis pipeline?

- Basically just a number of steps to analyze data

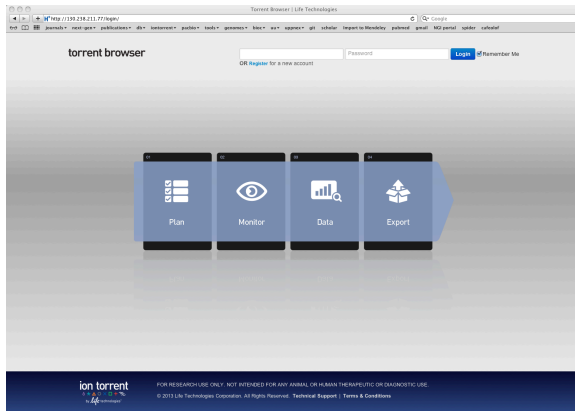


- Pipelines can be simple or very complex...

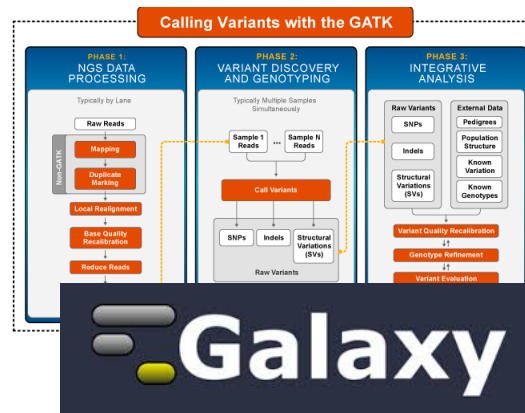


Some analysis pipelines for NGS data

Ion Torrent Torrent Suite Software



Illumina GATK, Galaxy,...



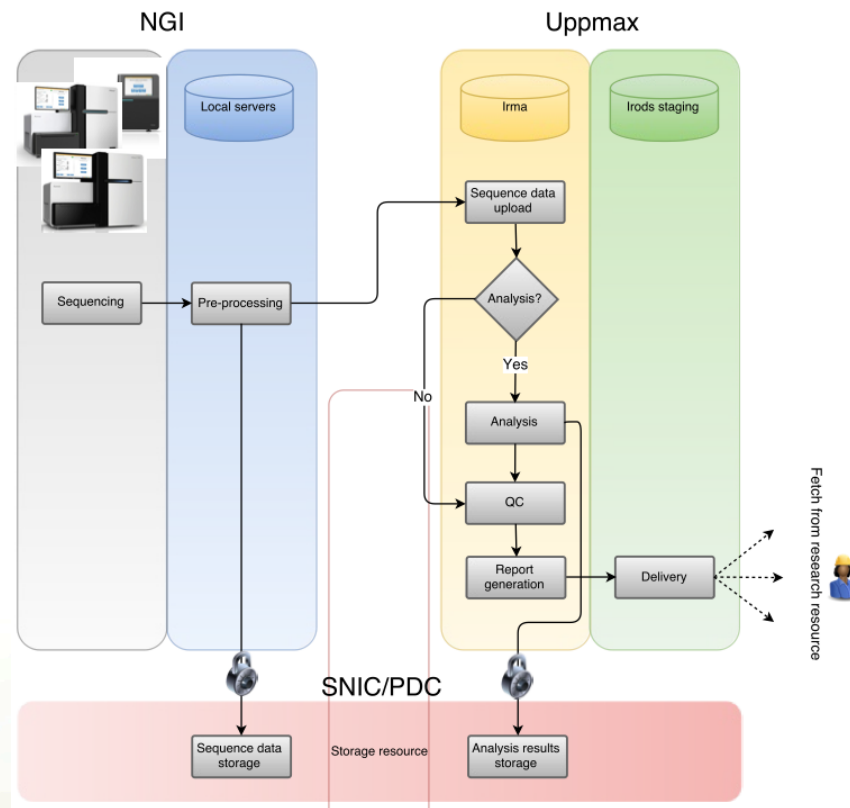
PacBio SMRT analysis portal



- Enables variant calling, de novo assembly, RNA expression analyses, ...
- Many other tools exist, also from commercial vendors

Data processing at NGI

- Raw data from is processed in automated pipelines
- Delivered to user accounts at UPPNEX



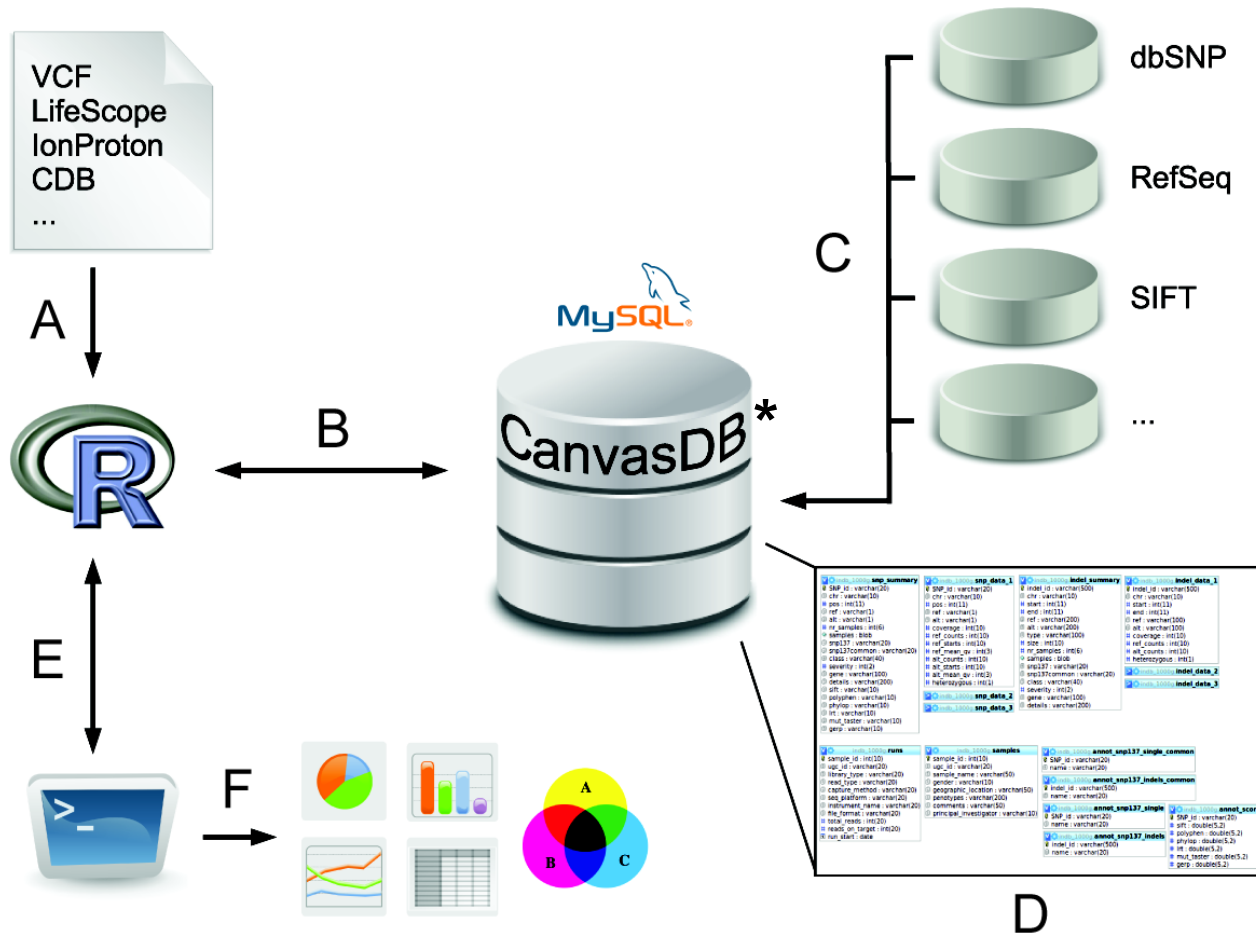
In-house development of pipelines

- In some cases NGI develops own pipelines
- But only when we see a need for a specific analysis

Some examples follows:

- I. Building a local variant databases (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
am_ssh am_ssh am_ssh bash cor_ssh ssh less ssh soli_ssh
SeqID Source Type Start End Score Strand Phase genotype reference coverage
novelAllelStart novelAllelEndNovelAbundanceAllelColor2 secondAbundanceAllelColor3 het
chr1 SOLiD_dlibases SNP 14973 14973 0.0 - - - S 6 16 10 7 25 6 3 26 00 33 1
chr1 SOLiD_dlibases SNP 14995 14995 0.99824 - - - C 6 3 1 1 31 2 2 2 2 24 23 0
chr1 SOLiD_dlibases SNP 15891 15891 0.0 - - - K 6 41 28 21 18 12 9 24 22 33 1
chr1 SOLiD_dlibases SNP 15977 15977 0.0 - - - R 6 259 155 54 20 55 29 29 29 29 29 0
chr1 SOLiD_dlibases SNP 20598 20598 0.002187 - - - W T 7 4 4 22 2 2 2 2 2 14 00 33 1
chr1 SOLiD_dlibases SNP 25204 25204 0.0 - - - Y T 58 24 17 26 39 22 22 83 21 1
chr1 SOLiD_dlibases SNP 75269 75269 0.0 - - - G C 27 0 0 0 22 17 24 21 0
chr1 SOLiD_dlibases SNP 88631 88631 0.0 - - - A G 24 1 1 29 23 13 25 83 21 1
chr1 SOLiD_dlibases SNP 88632 88632 0.003994 - - - A G 24 1 1 29 23 13 25 83 21 1
chr1 SOLiD_dlibases SNP 88676 88676 0.97058 - - - T C 60 9 7 19 46 38 38 38 38 38 0
chr1 SOLiD_dlibases SNP 89261 89261 0.0 - - - Y C 25 15 14 24 8 7 23 38 38 38 0
chr1 SOLiD_dlibases SNP 89268 89268 0.00396 - - - A G 5 0 0 0 8 3 3 38 38 38 0
chr1 SOLiD_dlibases SNP 88182 88182 0.0 - - - A G 5 0 0 0 5 4 26 38 38 38 0
chr1 SOLiD_dlibases SNP 88781 88781 0.0025 - - - G A 2 2 0 0 2 2 22 17 24 21 0
chr1 SOLiD_dlibases SNP 88859 88859 0.0 - - - C T 7 0 0 0 9 7 7 19 46 38 38 0
chr1 SOLiD_dlibases SNP 88859 88859 0.0 - - - C T 7 0 0 0 7 5 38 38 38 38 0
chr1 SOLiD_dlibases SNP 89276 89276 0.00396 - - - A G 6 3 0 0 8 3 3 38 38 38 0
chr1 SOLiD_dlibases SNP 89735 89735 0.0 - - - G 10 0 0 0 10 9 28 38 38 38 0
chr1 SOLiD_dlibases SNP 98537 98537 0.001941 - - - R A 7 4 3 25 3 2 2 15 26 6 21 15 13 23 0
chr1 SOLiD_dlibases SNP 98973 98973 0.0 - - - T C 22 0 0 0 22 15 26 6 21 15 13 23 0
chr1 SOLiD_dlibases SNP 98976 98976 0.0 - - - G A 16 0 0 0 15 15 24 23 23 23 0
chr1 SOLiD_dlibases SNP 91538 91538 0.0 - - - S G 28 12 10 26 6 4 21 38 38 38 0
chr1 SOLiD_dlibases SNP 91549 91549 0.0 - - - R G 53 35 29 21 15 13 23 23 23 0
chr1 SOLiD_dlibases SNP 94651 94651 0.00396 - - - T C 7 0 0 0 7 3 3 22 23 23 0
chr1 SOLiD_dlibases SNP 98468 98468 0.0 - - - A G 6 0 0 0 5 5 23 38 38 38 0
chr1 SOLiD_dlibases SNP 98187 98187 0.00396 - - - G A 3 0 0 0 3 3 23 38 38 38 0
chr1 SOLiD_dlibases SNP 98294 98294 0.0 - - - C T 11 0 0 0 9 8 25 38 38 38 0
chr1 SOLiD_dlibases SNP 98626 98626 0.00396 - - - C G 7 0 0 0 7 3 3 22 23 23 0
chr1 SOLiD_dlibases SNP 112839 112839 0.002914 - - - C T 2 0 0 0 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1
chr1 SOLiD_dlibases SNP 113982 113982 0.002413 - - - Y T 5 3 3 29 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1
chr1 SOLiD_dlibases SNP 115261 115261 0.0 - - - G A 43 0 0 0 43 38 27 38 38 38 0
chr1 SOLiD_dlibases SNP 124004 124004 0.0 - - - G A 43 0 0 0 43 38 27 38 38 38 0
chr1 SOLiD_dlibases SNP 124917 124917 0.0 - - - A G 18 0 0 0 18 15 29 38 38 38 0
chr1 SOLiD_dlibases SNP 125418 125418 0.00396 - - - G A 3 0 0 0 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
chr1 SOLiD_dlibases SNP 125481 125481 0.0 - - - C T 18 0 0 0 8 6 28 38 38 38 0
chr1 SOLiD_dlibases SNP 126847 126847 0.0 - - - G T 8 0 0 0 8 6 29 38 38 38 0
chr1 SOLiD_dlibases SNP 129704 129704 0.0025 - - - C G 2 0 0 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
    
```

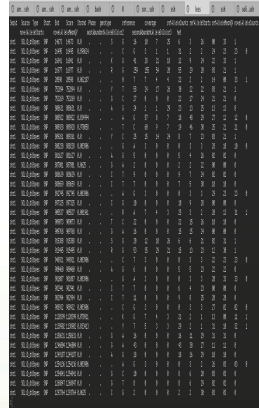
```

Terminal -- less -- 179x41
am_ssh am_ssh am_ssh bash cor_ssh ssh less ssh soli_ssh
SeqID Source Type Start End Score Strand Phase genotype reference coverage
novelAllelStart novelAllelEndNovelAbundanceAllelColor2 secondAbundanceAllelColor3 het
chr1 SOLiD_dlibases SNP 14973 14973 0.0 - - - S 6 16 10 7 25 6 3 26 00 33 1
chr1 SOLiD_dlibases SNP 14995 14995 0.99824 - - - C 6 3 1 1 31 2 2 2 2 24 23 0
chr1 SOLiD_dlibases SNP 15891 15891 0.0 - - - K 6 41 28 21 18 12 9 24 22 33 1
chr1 SOLiD_dlibases SNP 15977 15977 0.0 - - - R 6 259 155 54 20 55 29 29 29 29 29 0
chr1 SOLiD_dlibases SNP 20598 20598 0.002187 - - - W T 7 4 4 22 2 2 2 2 2 14 00 33 1
chr1 SOLiD_dlibases SNP 25204 25204 0.0 - - - Y T 58 24 17 26 39 22 22 83 21 1
chr1 SOLiD_dlibases SNP 75269 75269 0.0 - - - G C 27 0 0 0 22 17 24 21 0
chr1 SOLiD_dlibases SNP 88631 88631 0.0 - - - A G 24 1 1 29 23 13 25 83 21 1
chr1 SOLiD_dlibases SNP 88632 88632 0.003994 - - - A G 24 1 1 29 23 13 25 83 21 1
chr1 SOLiD_dlibases SNP 88676 88676 0.97058 - - - T C 60 9 7 19 46 38 38 38 38 38 0
chr1 SOLiD_dlibases SNP 89261 89261 0.0 - - - Y C 25 15 14 24 8 7 23 38 38 38 0
chr1 SOLiD_dlibases SNP 89268 89268 0.00396 - - - A G 5 0 0 0 8 3 3 38 38 38 0
chr1 SOLiD_dlibases SNP 88182 88182 0.0 - - - A G 5 0 0 0 5 4 26 38 38 38 0
chr1 SOLiD_dlibases SNP 88781 88781 0.0025 - - - G A 2 2 0 0 2 2 22 17 24 21 0
chr1 SOLiD_dlibases SNP 88859 88859 0.0 - - - C T 7 0 0 0 9 7 7 19 46 38 38 0
chr1 SOLiD_dlibases SNP 88859 88859 0.0 - - - C T 7 0 0 0 7 5 38 38 38 38 0
chr1 SOLiD_dlibases SNP 89276 89276 0.00396 - - - A G 6 3 0 0 8 3 3 38 38 38 0
chr1 SOLiD_dlibases SNP 89735 89735 0.0 - - - G 10 0 0 0 10 9 28 38 38 38 0
chr1 SOLiD_dlibases SNP 98537 98537 0.001941 - - - R A 7 4 3 25 3 2 2 15 26 6 21 15 13 23 0
chr1 SOLiD_dlibases SNP 98973 98973 0.0 - - - T C 22 0 0 0 22 15 26 6 21 15 13 23 0
chr1 SOLiD_dlibases SNP 98976 98976 0.0 - - - G A 16 0 0 0 15 15 24 23 23 23 0
chr1 SOLiD_dlibases SNP 91538 91538 0.0 - - - S G 28 12 10 26 6 4 21 38 38 38 0
chr1 SOLiD_dlibases SNP 91549 91549 0.0 - - - R G 53 35 29 21 15 13 23 23 23 0
chr1 SOLiD_dlibases SNP 94651 94651 0.00396 - - - T C 7 0 0 0 7 3 3 22 23 23 0
chr1 SOLiD_dlibases SNP 98468 98468 0.0 - - - A G 6 0 0 0 5 5 23 38 38 38 0
chr1 SOLiD_dlibases SNP 98187 98187 0.00396 - - - G A 3 0 0 0 3 3 23 38 38 38 0
chr1 SOLiD_dlibases SNP 98294 98294 0.0 - - - C T 11 0 0 0 9 8 25 38 38 38 0
chr1 SOLiD_dlibases SNP 98626 98626 0.00396 - - - C G 7 0 0 0 7 3 3 22 23 23 0
chr1 SOLiD_dlibases SNP 112839 112839 0.002914 - - - C T 2 0 0 0 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1
chr1 SOLiD_dlibases SNP 113982 113982 0.002413 - - - Y T 5 3 3 29 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1
chr1 SOLiD_dlibases SNP 115261 115261 0.0 - - - G A 43 0 0 0 43 38 27 38 38 38 0
chr1 SOLiD_dlibases SNP 124004 124004 0.0 - - - G A 43 0 0 0 43 38 27 38 38 38 0
chr1 SOLiD_dlibases SNP 124917 124917 0.0 - - - A G 18 0 0 0 18 15 29 38 38 38 0
chr1 SOLiD_dlibases SNP 125418 125418 0.00396 - - - G A 3 0 0 0 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
chr1 SOLiD_dlibases SNP 125481 125481 0.0 - - - C T 18 0 0 0 8 6 28 38 38 38 0
chr1 SOLiD_dlibases SNP 126847 126847 0.0 - - - G T 8 0 0 0 8 6 29 38 38 38 0
chr1 SOLiD_dlibases SNP 129704 129704 0.0025 - - - C G 2 0 0 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
    
```

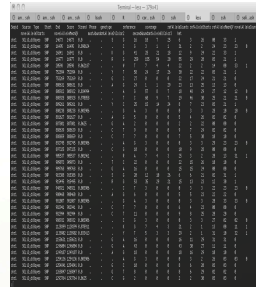
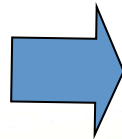
How to analyze this?

- 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)

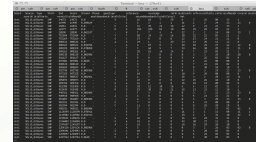
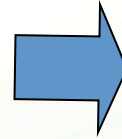
Start:
All identified SNPs



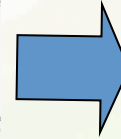
A screenshot of a text-based list of identified SNPs. The list is organized in columns, with the first column containing SNP IDs (e.g., rs1044544, rs1044545, etc.) and subsequent columns containing various data points such as genomic coordinates and allele frequencies. The text is dense and spans many lines.



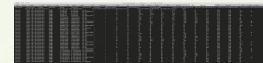
A screenshot of a filtered list of SNPs. The list is shorter than the 'Start' box, indicating that many SNPs have been removed through the filtering process. The format is similar to the 'Start' box, with columns for SNP IDs and associated data.



A screenshot of a further filtered list of SNPs. This list is significantly shorter than the previous ones, representing the 'A few candidate causative SNP(s)'. The format remains consistent with the previous boxes.



Result:
A few candidate
causative
SNP(s)!

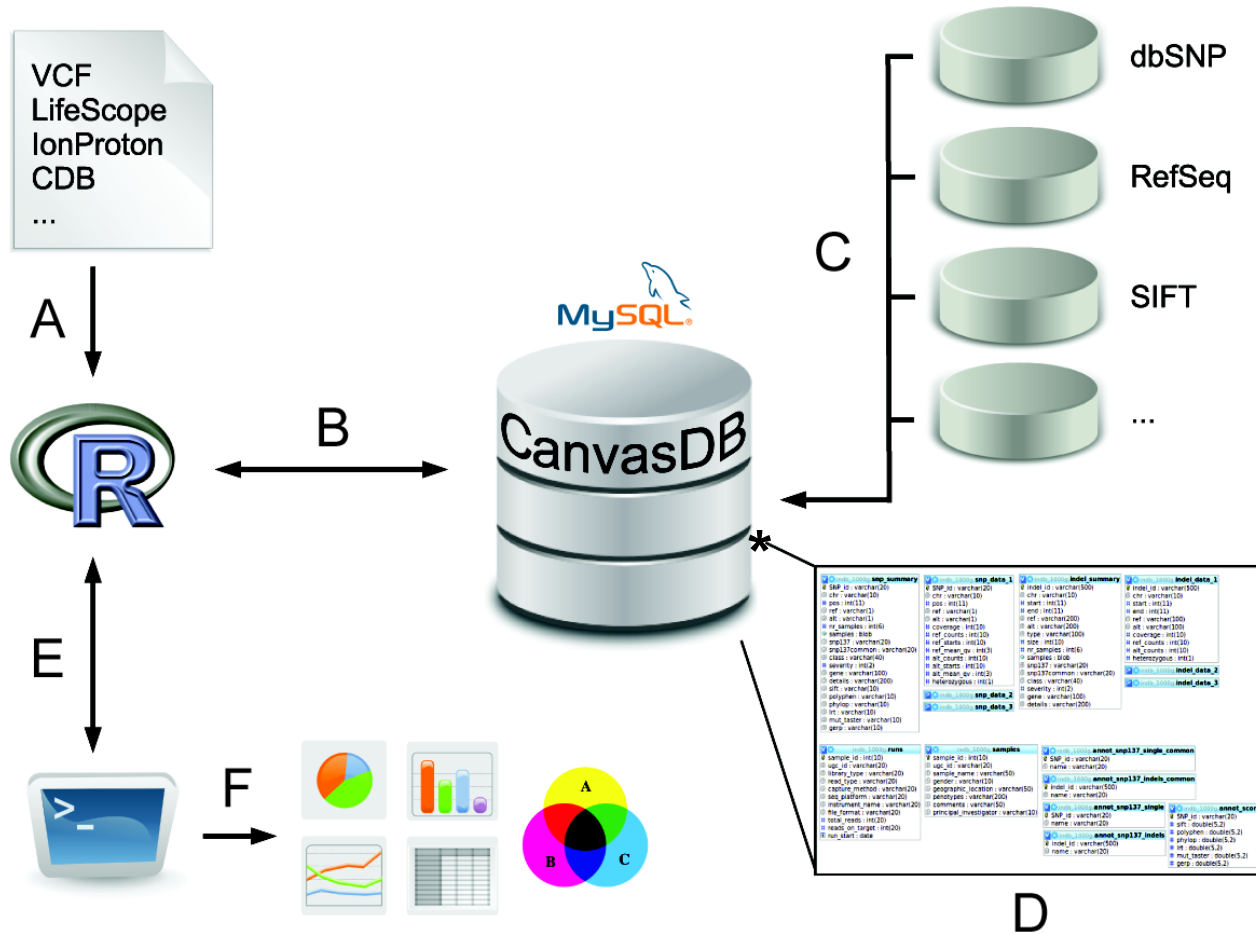


A screenshot of the final candidate SNPs. This is the smallest list shown, containing only a few SNPs that are considered as potential causative variants. The format is consistent with the other boxes.

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

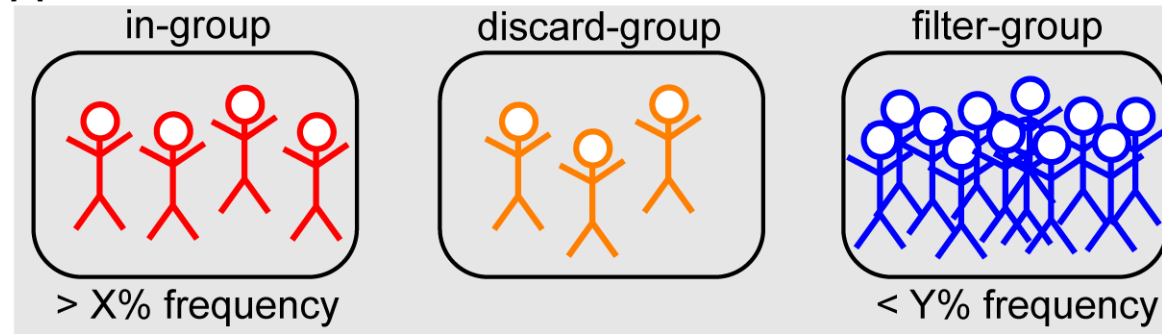


***CANdicate Variant Analysis System and Data Base**


Ameur et al., Database Journal, 2014

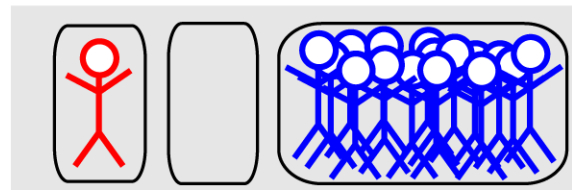
CanvasDB - Filtering

A

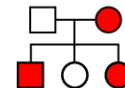


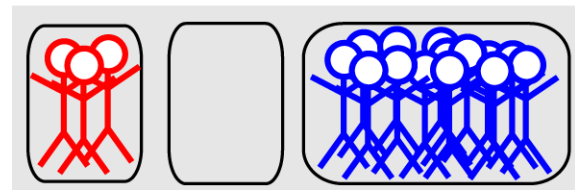
B

parent-offspring trio 



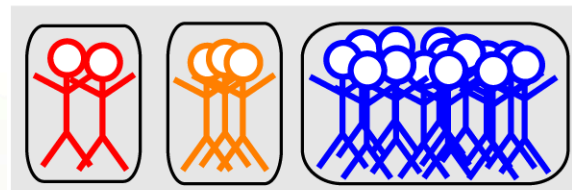
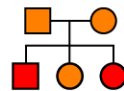
C

dominant variant 



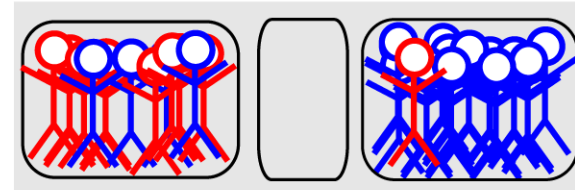
D

recessive variant



E

comparing groups

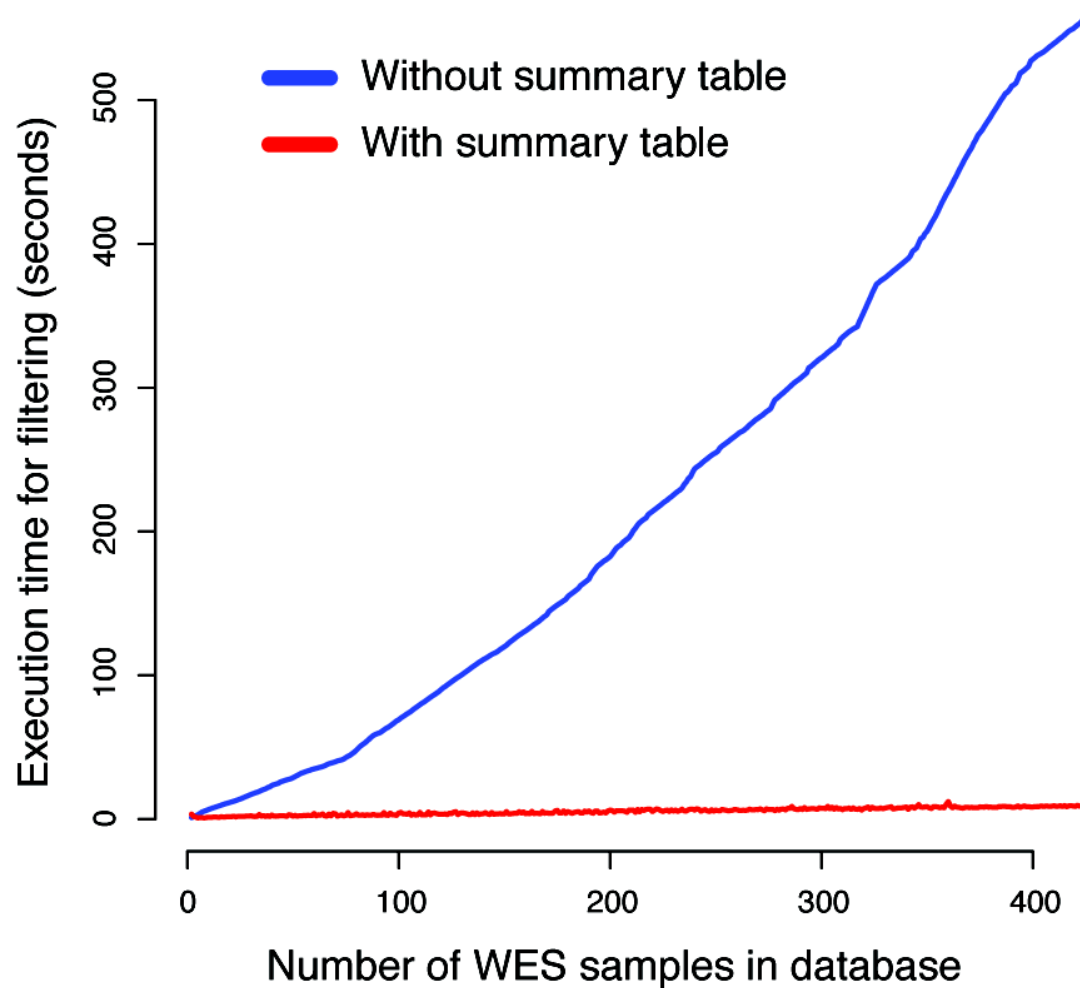


> min freq g_1

< max freq g_2

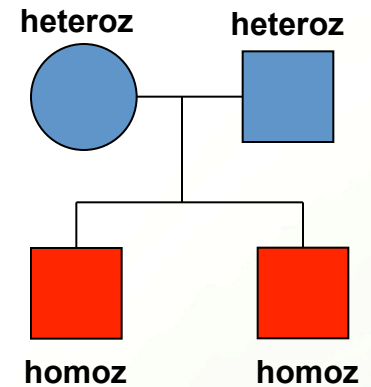
CanvasDB - Filtering speed

- Rapid variant filtering, also for large databases



An exome-seq project

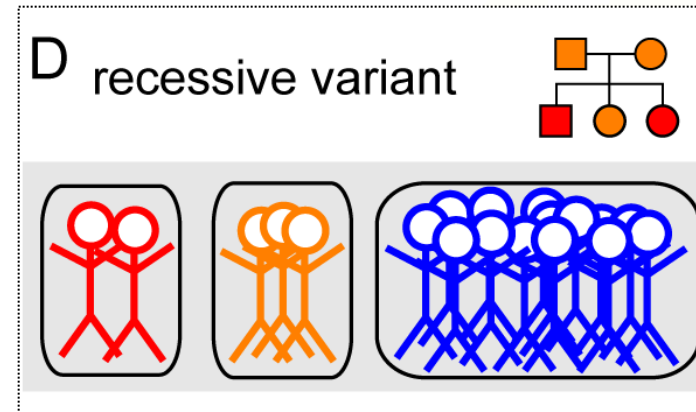
- Hearing loss: 2 affected brothers
 - Likely a rare, recessive disease
 - => Shared homozygous SNPs/indels
- Sequencing strategy
 - TargetSeq exome capture
 - One sample per PI chip



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

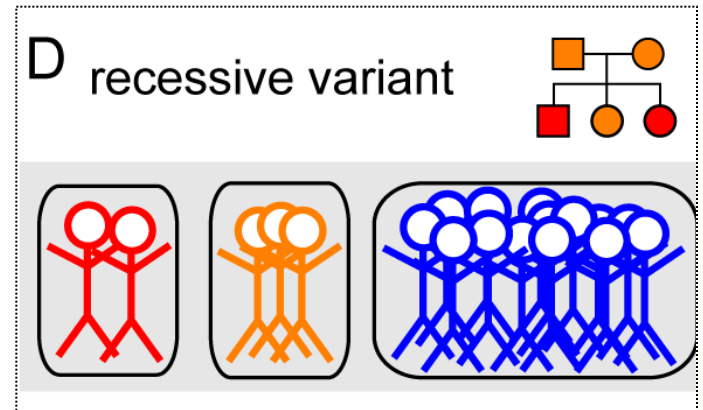
- 2 SNPs

- 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

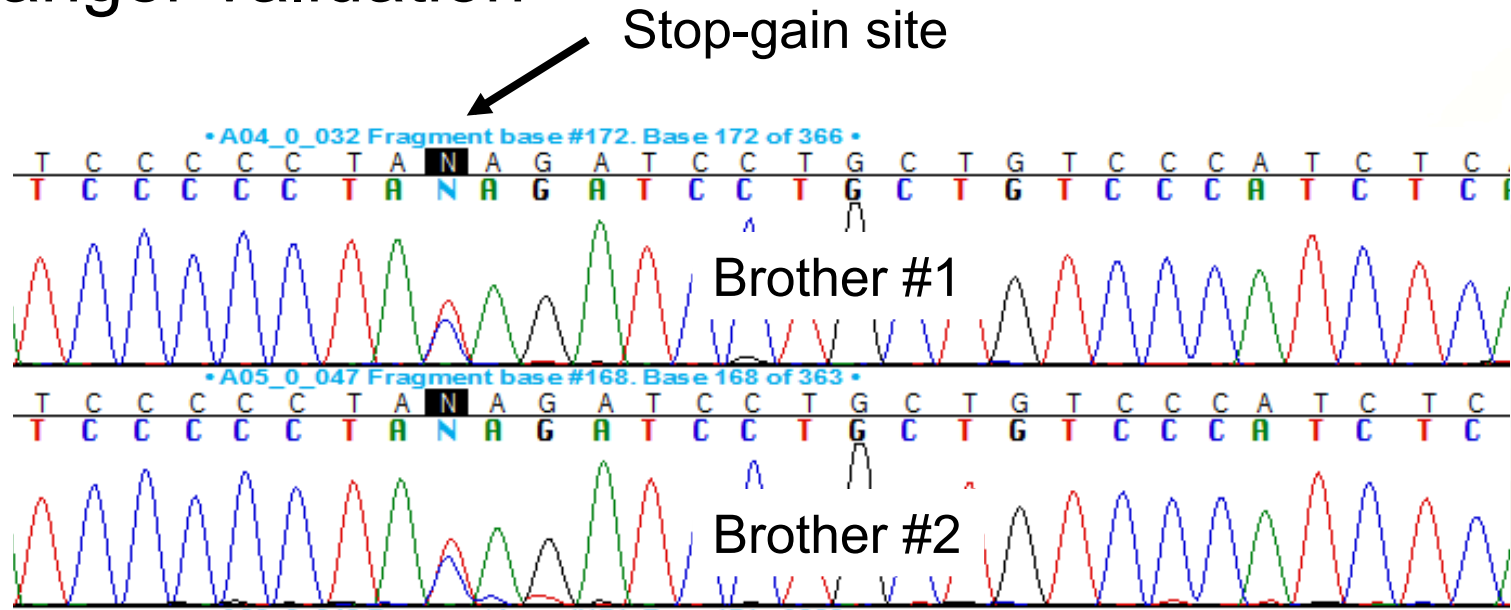
Brother #1

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 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



Capacity of HiSeq X Ten: 320 whole human genomes/week!!!

⇒ More work on pipelines and databases needed!

Whole Genome Sequencing Projects

Business & Financial News, Breaking US & International News | Reuters.com

 REUTERS

» Print

This copy is for your personal, non-commercial use only. To order presentation-ready copies for distribution to colleagues, clients or customers, use the Reprints tool at the top of any article or visit: www.reutersreprints.com.

UK to launch 100,000 genomes project as Obama backs DNA drive

Fri, Jan 30 2015

* Genomes project starts recruitment in England on Feb. 2

* U.S. to analyse genetic data from 1 million Americans

By [Ben Hirschler](#)

LONDON, Jan 30 (Reuters) - Gene research is getting a boost on both sides of the Atlantic, with scientists in England set to launch a project on Feb. 2 to analyse 100,000 entire human genomes and U.S. President Barack Obama backing a big new DNA data drive.

The twin projects show the accelerating work by researchers to understand the underlying basis of diseases and develop medicines targeted to the genetic profile of individual patients.

Obama will announce the U.S. plan to analyse genetic information from more than 1 million American volunteers on Friday as a central part of an initiative to promote so-called precision medicine, officials said.

The 100,000 genomes project in England, meanwhile, was first unveiled by the British government two years ago -- but the 11 centres charged with collecting samples will only begin full-scale recruitment from next week. The aim is to complete the programme by the end of 2017.

Such large-scale genomic research has become possible because the cost of genome sequencing has plummeted in recent years to around \$1,000 per genome. That is a far cry from 15 years ago when it cost some \$3 billion to get the first human genome.

In the case of the British project, all the sequencing will be carried out by U.S. biotech company Illumina, which has pioneered fast and cheap technology to read genetic code.

The 100,000 genomes project is focusing on patients with rare diseases, and their families, as well as people with common cancers. The idea is to tease out the common drivers of disease to help develop better drugs and diagnostic tests.

In addition to helping doctors understand more about disease, the government also hopes the scheme will make the state-run National Health Service a world leader in science and boost Britain's life sciences industry.

The project will actually recruit around 75,000 participants, rather than 100,000, since people with cancer will provide two genomes -- one derived from the healthy cells in their body and one from their tumour. By comparing the two, experts hope to find the exact genetic changes causing cancer. (Editing by Susan Thomas)

© Thomson Reuters 2015. All rights reserved. Users may download and print extracts of content from this website for their own personal and non-commercial use only. Republication or redistribution of Thomson Reuters content, including by framing or similar means, is expressly prohibited without the prior written consent of Thomson Reuters. Thomson Reuters and its logo are registered trademarks or trademarks of the Thomson Reuters group of companies around the world.

Thomson Reuters journalists are subject to an Editorial Handbook which requires fair presentation and disclosure of relevant interests.

This copy is for your personal, non-commercial use only. To order presentation-ready copies for distribution to colleagues, clients or customers, use the Reprints tool at the top of

PERSPECTIVE

Big Data: Astronomical or Genomical?

Table 1. Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Analysis of WGS data @ SciLifeLab

We have a working group for WGS at SciLifeLab!

wgs-toolbox@scilifelab.se

Contacts with Genomics England initiated for analyses

Genomics
england



The SciLifeLab Human WGS Initiative

- WGS of patient cohorts (n=10,000 ind/year)
- Genetic Variant Database for the Swedish Population (n=1000)



The Swedish Genetic Variant Project

- A. Identify a cohort that reflects the genetic structure of the Swedish population
- B. Generate WGS data using short- and long-read MPS technologies
- C. Establish a user-friendly database to make information available to the research community (association analyses) and clinical genetics laboratories.

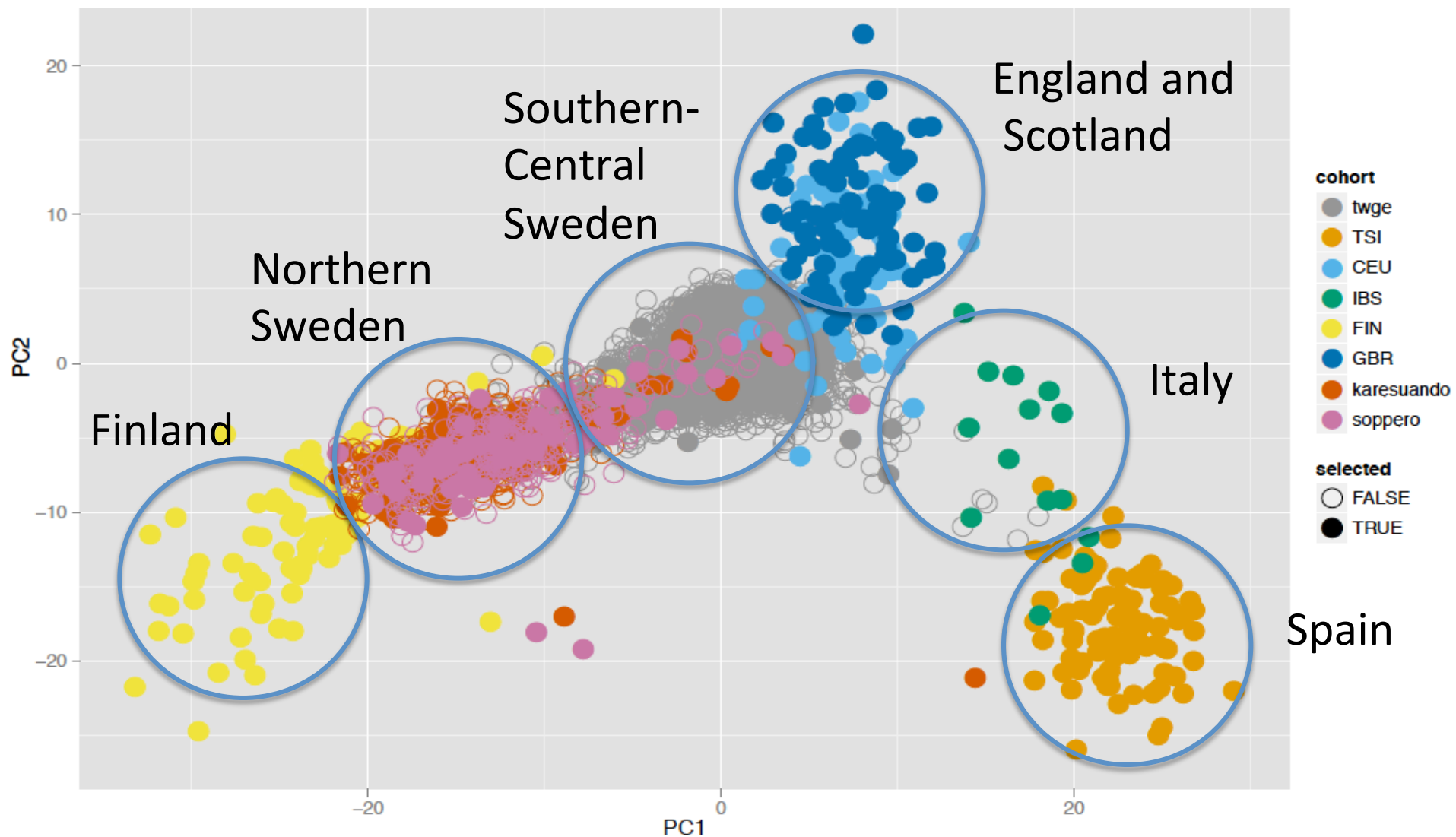
Twin Registry samples used as control cohort

- Inclusion based on twinning
- Distribution like population density
- General population-prevalence of disease
- 10,000 individuals have been analysed with SNP arrays



Identify 1,000 individuals based on genetic structure and diversity across Sweden

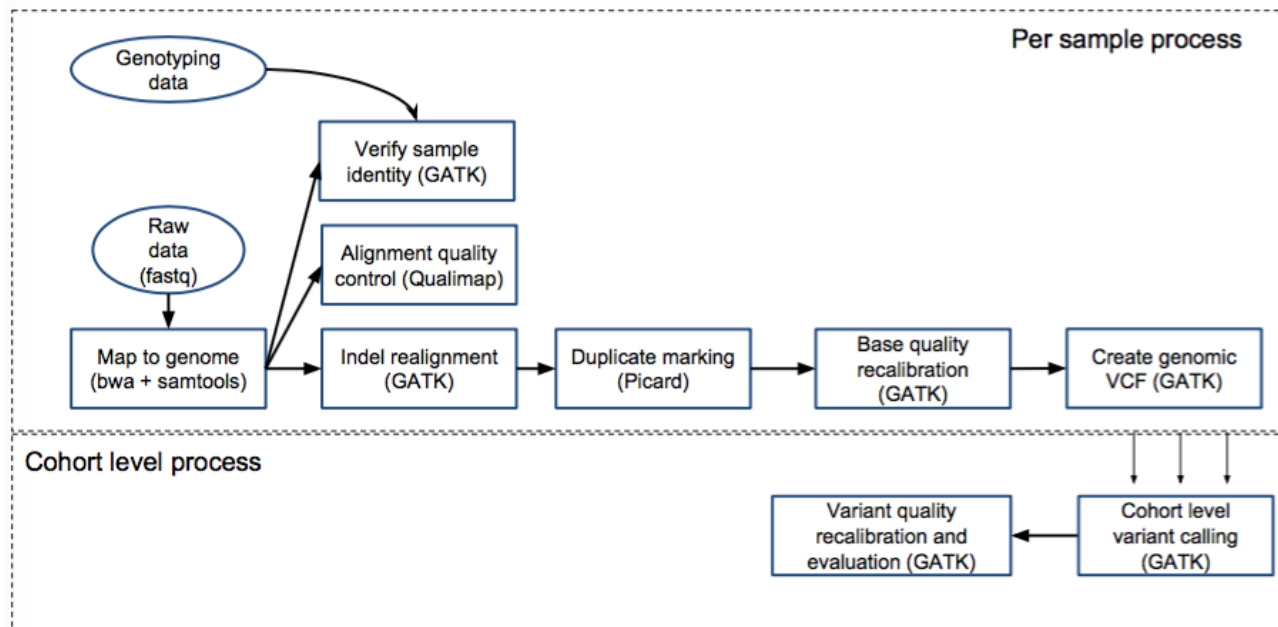
Principal components of European samples from 1,000 genomes project and 10,000 Swedish samples



Illumina WGS of Swedish control cohort

Step 1: 30X Illumina data of the 1,000 individuals

Step 2: Mapping and variant calling



Step 3: Making genotype frequencies available for download

Making frequency data available

SweFreq

Data Beacon

ExAC Browser

Login

SweGen Variant Frequency Database

This server hosts whole-genome variant frequencies for 1000 Swedish individuals generated within the SweGen project. The frequency data is intended to be used as a resource for the research community and clinical genetics laboratories. Individual positions in the genome can be viewed using the Data Beacon or ExAC Browser by clicking the links above. To access the variant frequency file you need to register.

Please note that the 1000 individuals included in the SweGen project represent a cross-section of the Swedish population and that no disease information has been used for the selection. The frequency data may therefore include genetic variants that are associated with, or causative of, disease.

We request that any use of data from the SweGen project cite this preprint on bioRxiv.



SciLifeLab

NATIONAL CTAC
ATLAS GENOMICS
INFRASTRUCTURE

NBS

elixir
SWEDEN

Aggregated frequencies now available from ***swefreq.nbis.se!***

Example II:

Assembly of genomes using Pacific Biosciences



Genome assembly using NGS

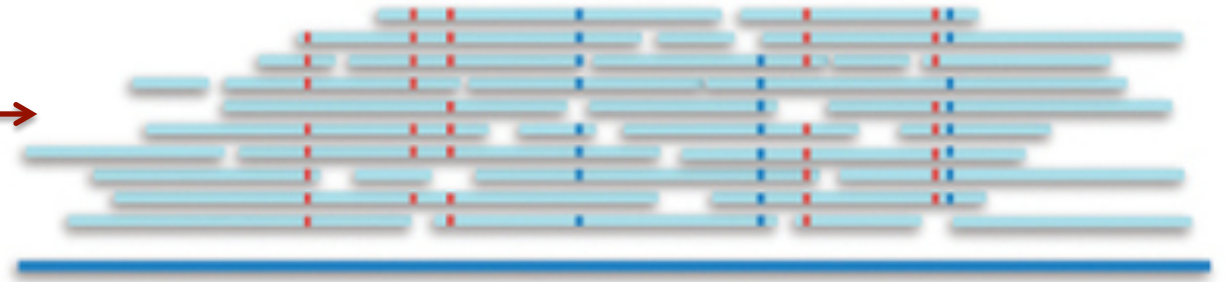
- Short-read *de novo* assembly by NGS
 - Requires mate-pair sequences
 - Ideally with different insert sizes
 - Complicated analysis
 - Assembly, scaffolding, finishing
 - Maybe even some manual steps

=> 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 assembly analysis

- Simple -- just click a button!!

The screenshot shows the PacBio SMRT Portal interface. The browser address bar displays `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/...

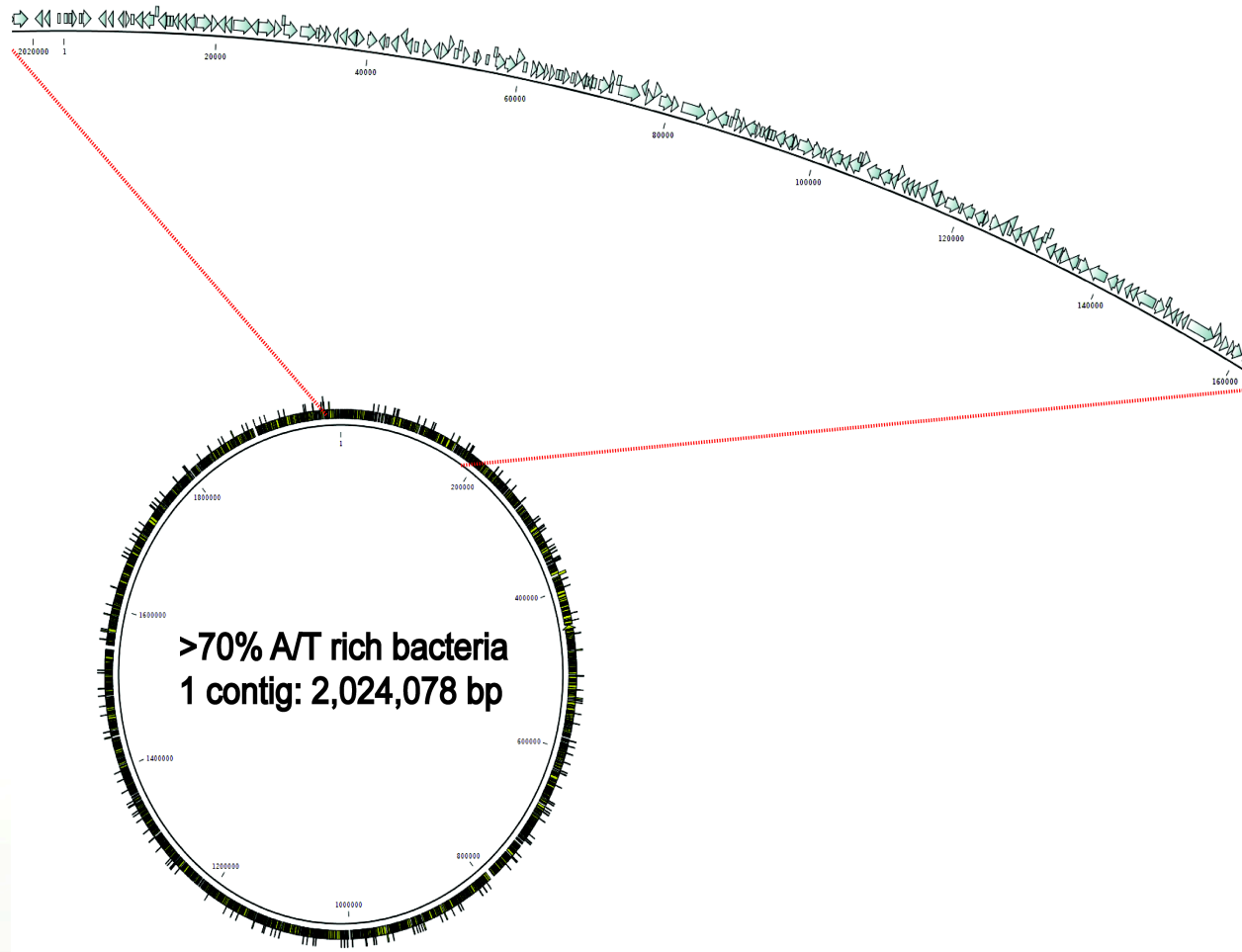
SMRT Cells in Job (Viewing 1 - 1 of 1)

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".

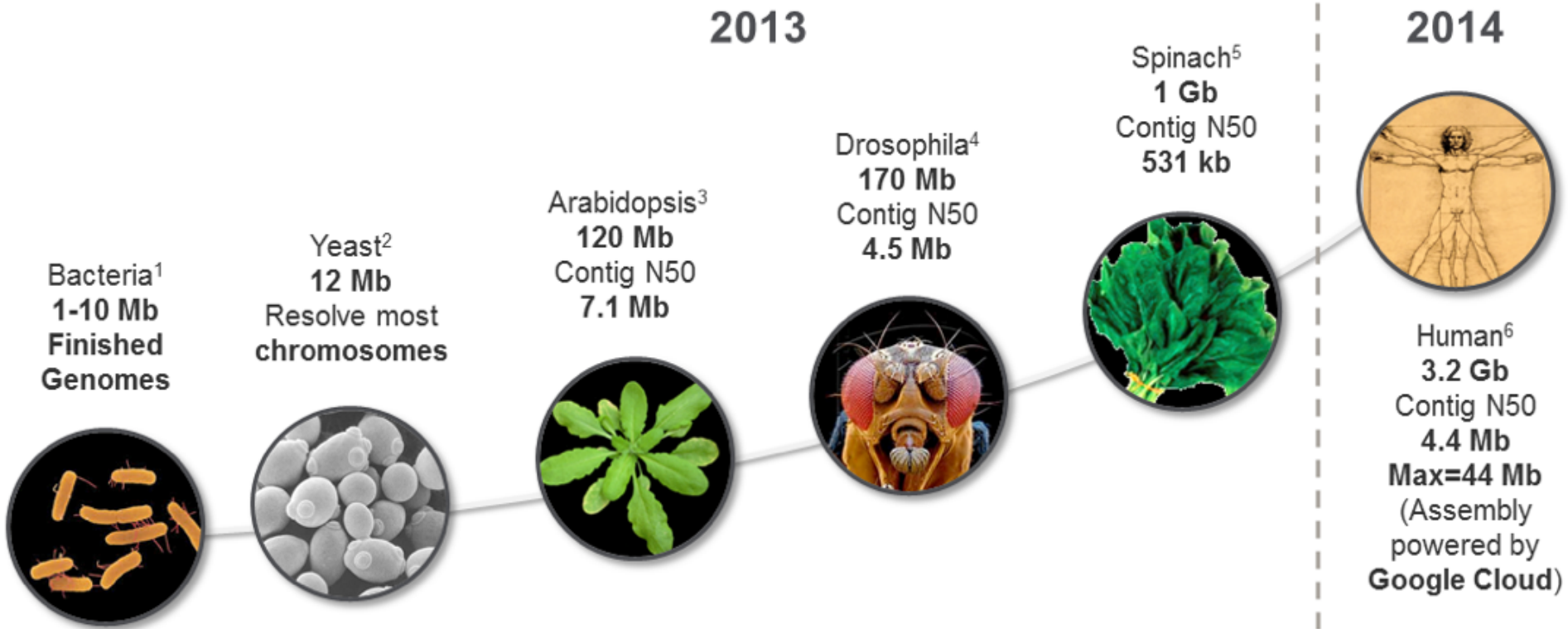
PacBio assembly, example result

- Example: Complete assembly of a bacterial genome



PacBio assembly – recent developments

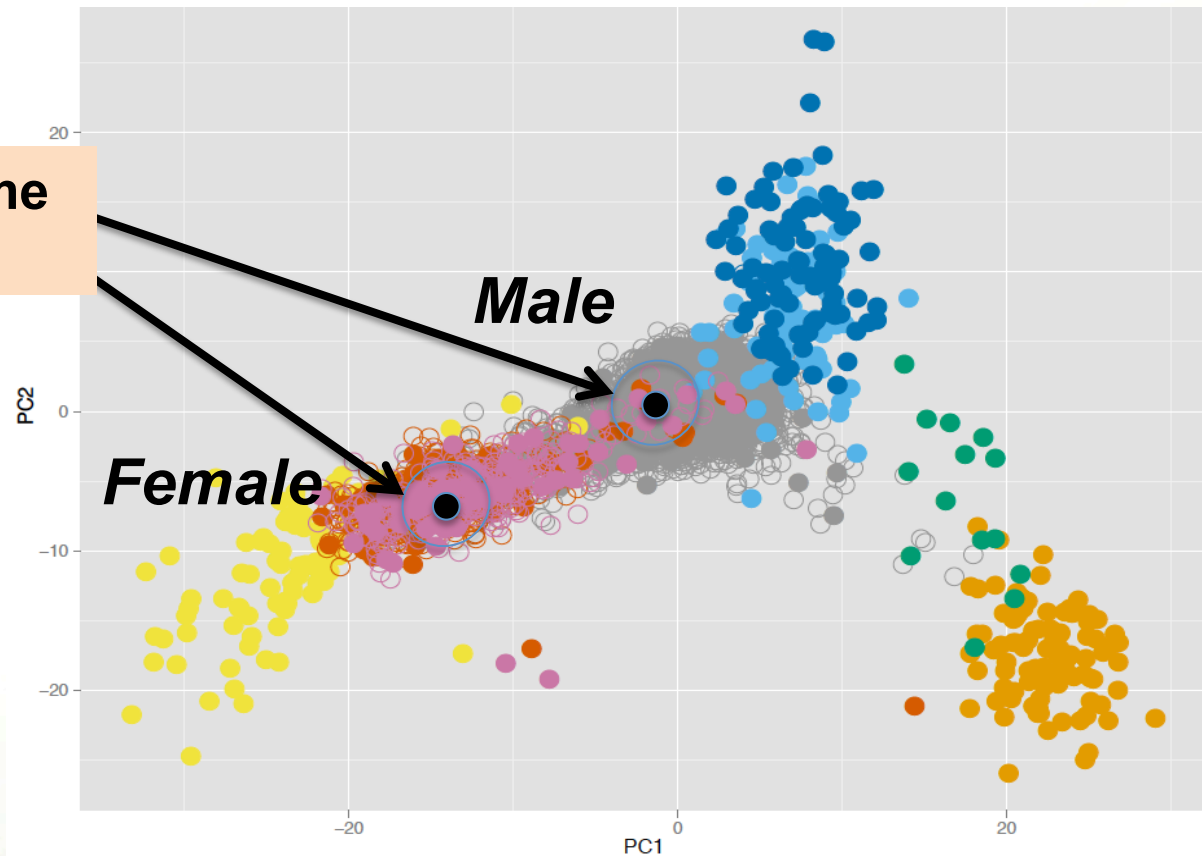
- Also larger genomes can be assembled by PacBio..



De novo WGS of Swedish cohort

Establish Swedish reference genome sequences by *de novo* assembly of long-reads: ***PacBio+BioNano+10X Genomics***

Reference genome individuals

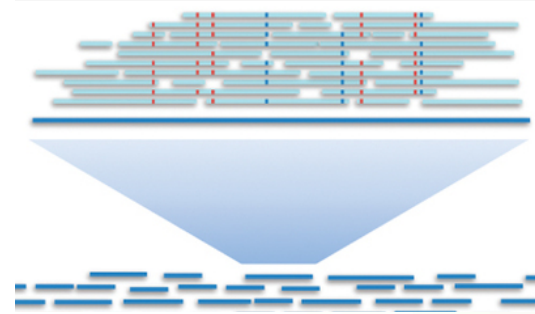


De novo assembly of 75X PacBio data

Assembly (FALCON)



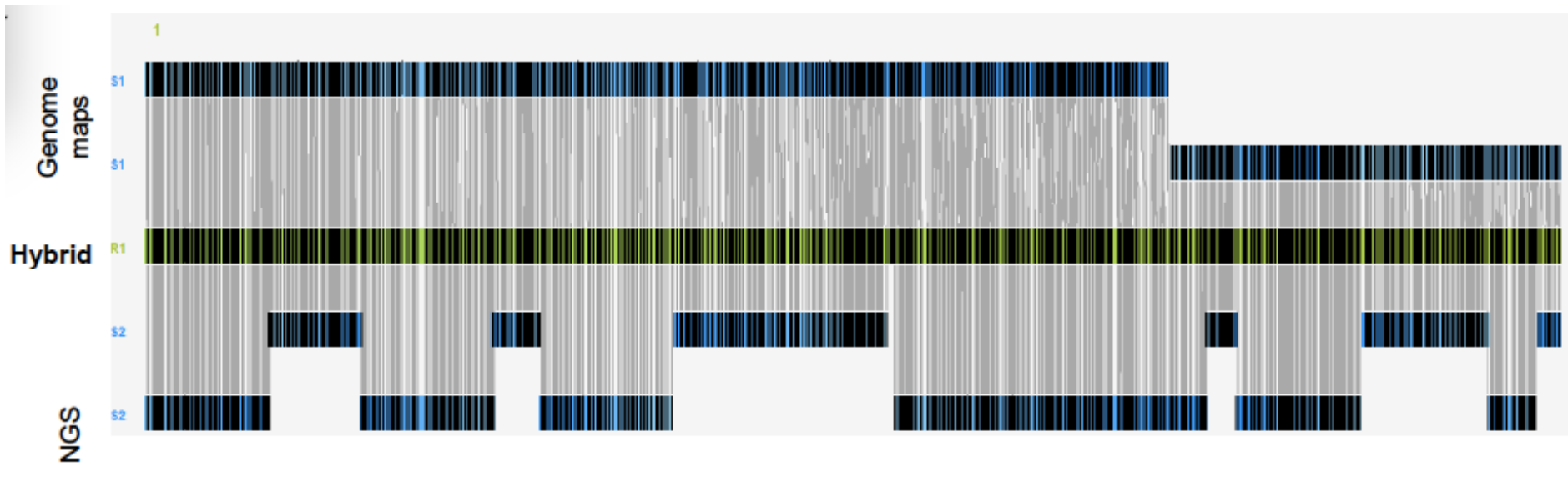
Error correction (2 x Quiver)



Analysis time: 1 month/genome

	Individual 1	Individual 2
Assembly size	3,039,619,582	3,024,752,299
Nr contigs	11,249	11,601
Longest contig	36,8 Mb	54,1 Mb
N50	8,9 Mb	8,3 Mb

Hybrid scaffolding, PacBio + BioNano



Hybrid scaffolding with two labellings resulted in

- **3,1 Gb** assembly, **51 Mb** N50 (for individual #1)
- **3,1 Gb** assembly, **46 Mb** N50 (for individual #2)

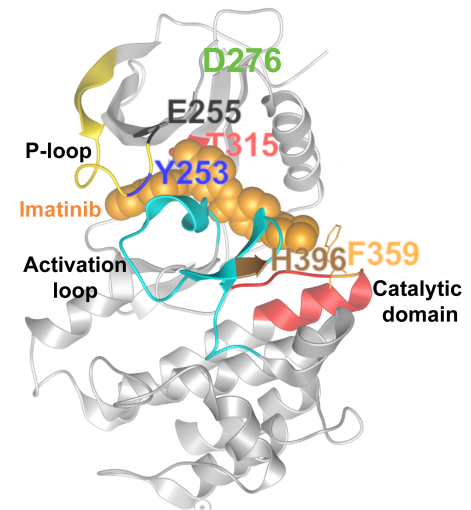
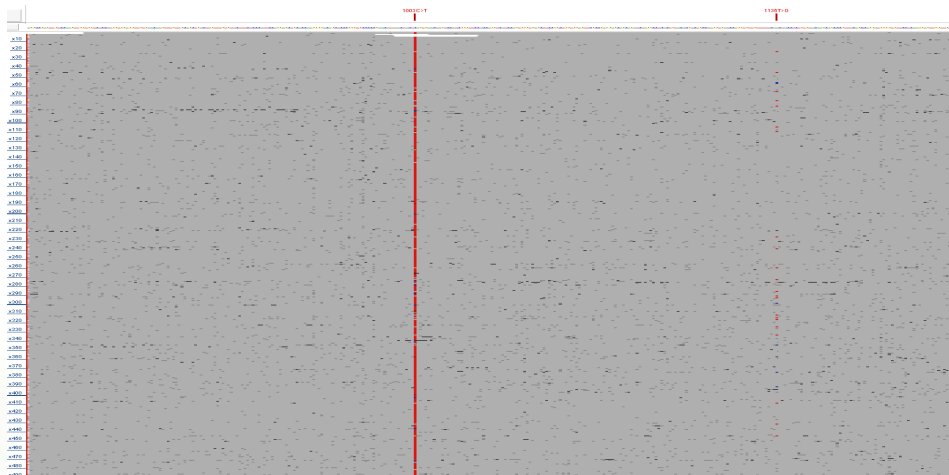
Aligning contigs to human reference

> 99% of bases can be aligned to human reference (hg38)



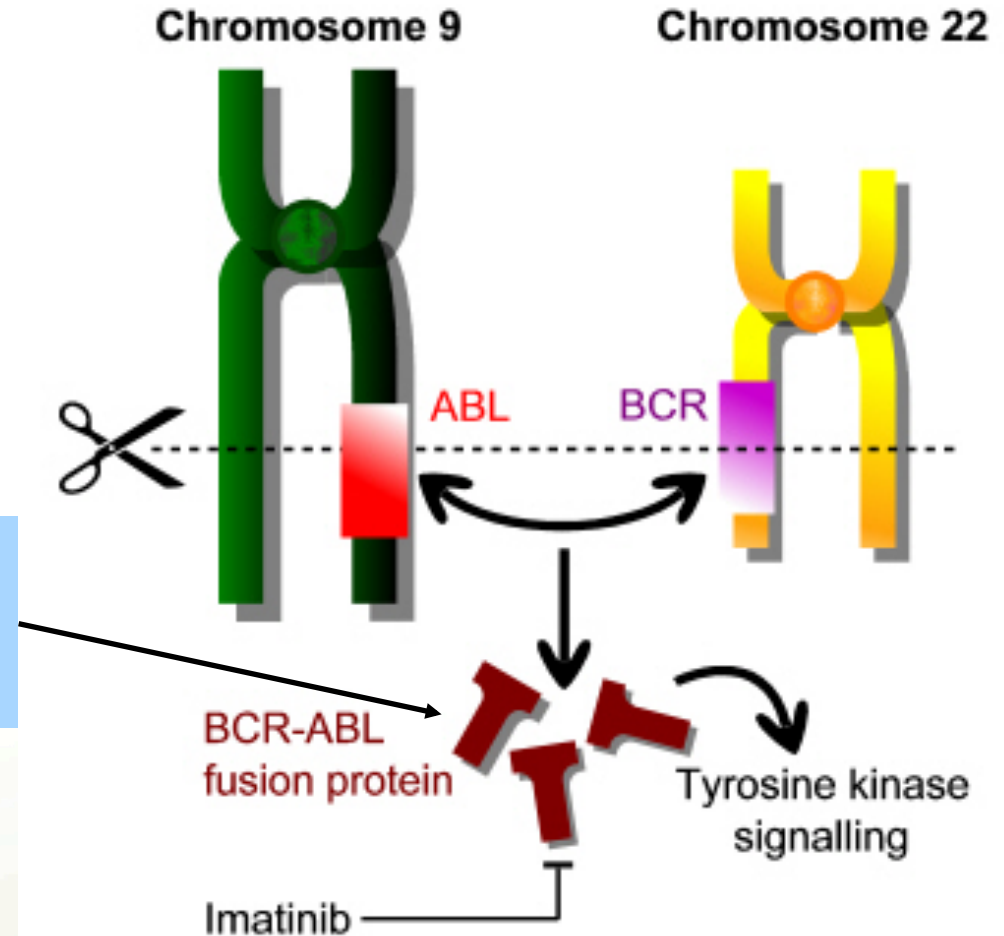
Example III:

Clinical sequencing for Leukemia Treatment



Chronic Myeloid Leukemia

- BCR-ABL1 fusion protein – a CML drug target

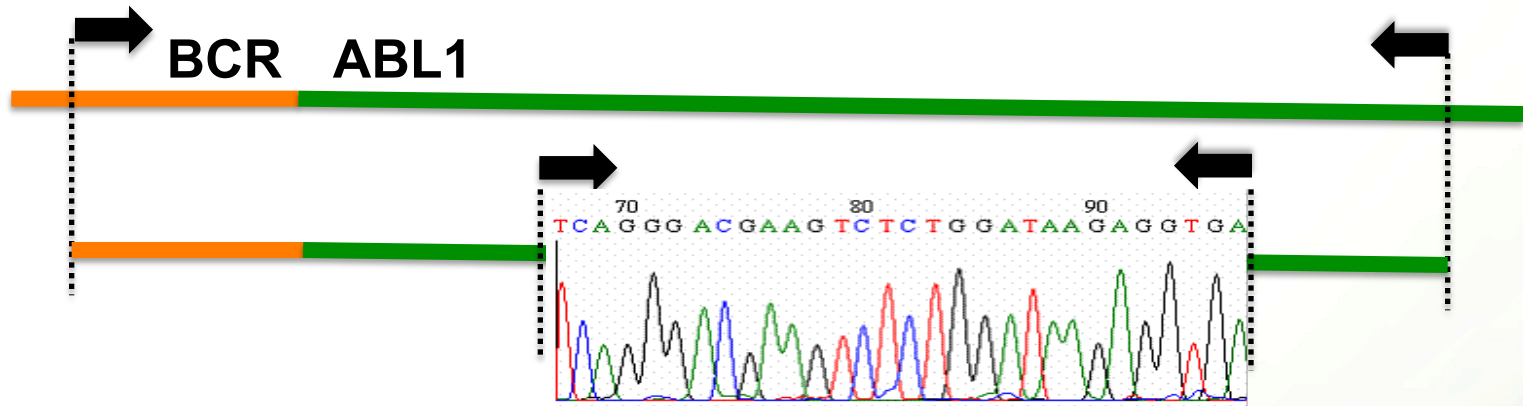


The BCR-ABL1 fusion protein can acquire resistance mutations following drug treatment

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

Traditional mutation screening in BCR-ABL1

Nested PCR and Sanger sequencing:



Limitations:

- Mutations at frequencies below 10-20% not seen
- Biases may be introduced by nested PCR
- Whole BCR-ABL1 fusion transcript not sequenced
- Clonal composition of mutations not determined

Our clinical diagnostics pipeline for BCR-ABL1

Cavelier et al. BMC Cancer (2015) 15:45
DOI 10.1186/s12885-015-1046-y



RESEARCH ARTICLE

Open Access

Clonal distribution of *BCR-ABL1* mutations and splice isoforms by single-molecule long-read RNA sequencing

Lucia Cavelier^{1*}, Adam Ameur^{1†}, Susana Häggqvist¹, Ida Höjjer¹, Nicola Cahill¹, Ulla Olsson-Strömberg² and Monica Hermanson¹

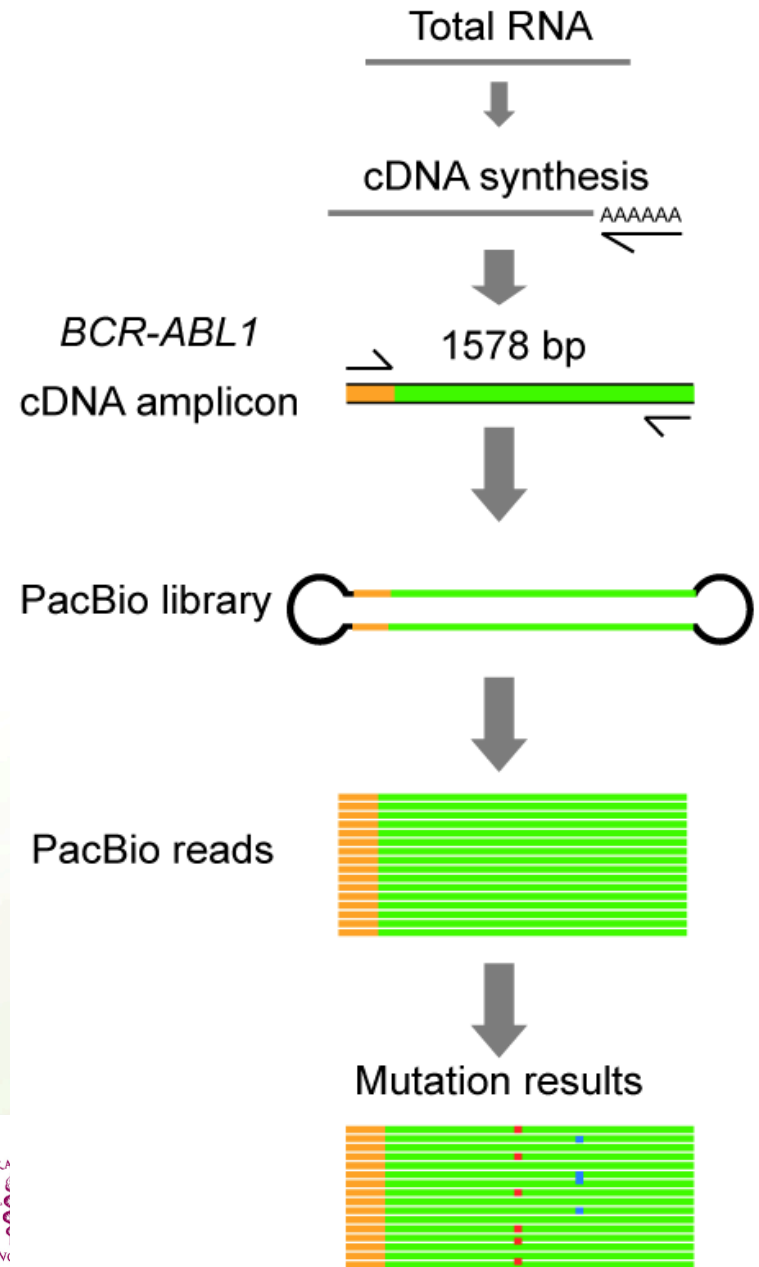
Abstract

Background: The evolution of mutations in the *BCR-ABL1* fusion gene transcript renders CML patients resistant to tyrosine kinase inhibitor (TKI) based therapy. Thus screening for *BCR-ABL1* mutations is recommended particularly in patients experiencing poor response to treatment. Herein we describe a novel approach for the detection and surveillance of *BCR-ABL1* mutations in CML patients.

Methods: To detect mutations in the *BCR-ABL1* transcript we developed an assay based on the Pacific Biosciences (PacBio) sequencing technology, which allows for single-molecule long-read sequencing of *BCR-ABL1* fusion transcript molecules. Samples from six patients with poor response to therapy were analyzed both at diagnosis and follow-up. cDNA was generated from total RNA and a 1,6 kb fragment encompassing the *BCR-ABL1* transcript was amplified using long range PCR. To estimate the sensitivity of the assay, a serial dilution experiment was performed.

Results: Over 10,000 full-length *BCR-ABL1* sequences were obtained for all samples studied. Through the serial dilution analysis, mutations in CML patient samples could be detected down to a level of at least 1%. Notably, the assay was determined to be sufficiently sensitive even in patients harboring a low abundance of *BCR-ABL1* levels. The PacBio sequencing successfully identified all mutations seen by standard methods. Importantly, we identified several mutations that escaped detection by the clinical routine analysis. Resistance mutations were found in all but one of the patients. Due to the long reads afforded by PacBio sequencing, compound mutations present in the same molecule were readily distinguished from independent alterations arising in different molecules. Moreover, several transcript isoforms of the *BCR-ABL1* transcript were identified in two of the CML patients. Finally, our assay allowed for a quick turn around time allowing samples to be reported upon within 2 days.

Conclusions: In summary the PacBio sequencing assay can be applied to detect *BCR-ABL1* resistance mutations in both diagnostic and follow-up CML patient samples using a simple protocol applicable to routine diagnosis. The method besides its sensitivity, gives a complete view of the clonal distribution of mutations, which is of importance when making therapy decisions.

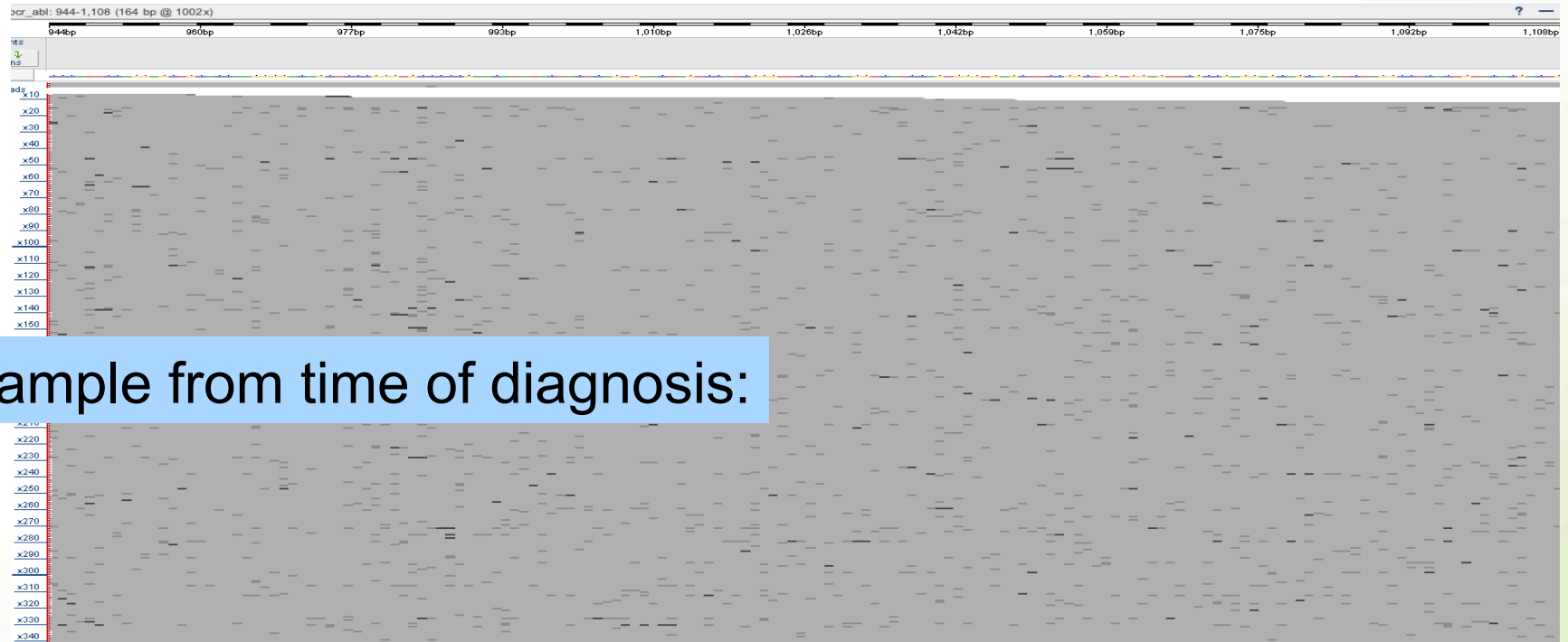


BCR-ABL1 mutations at diagnosis

PacBio sequencing generates ~10 000X coverage!

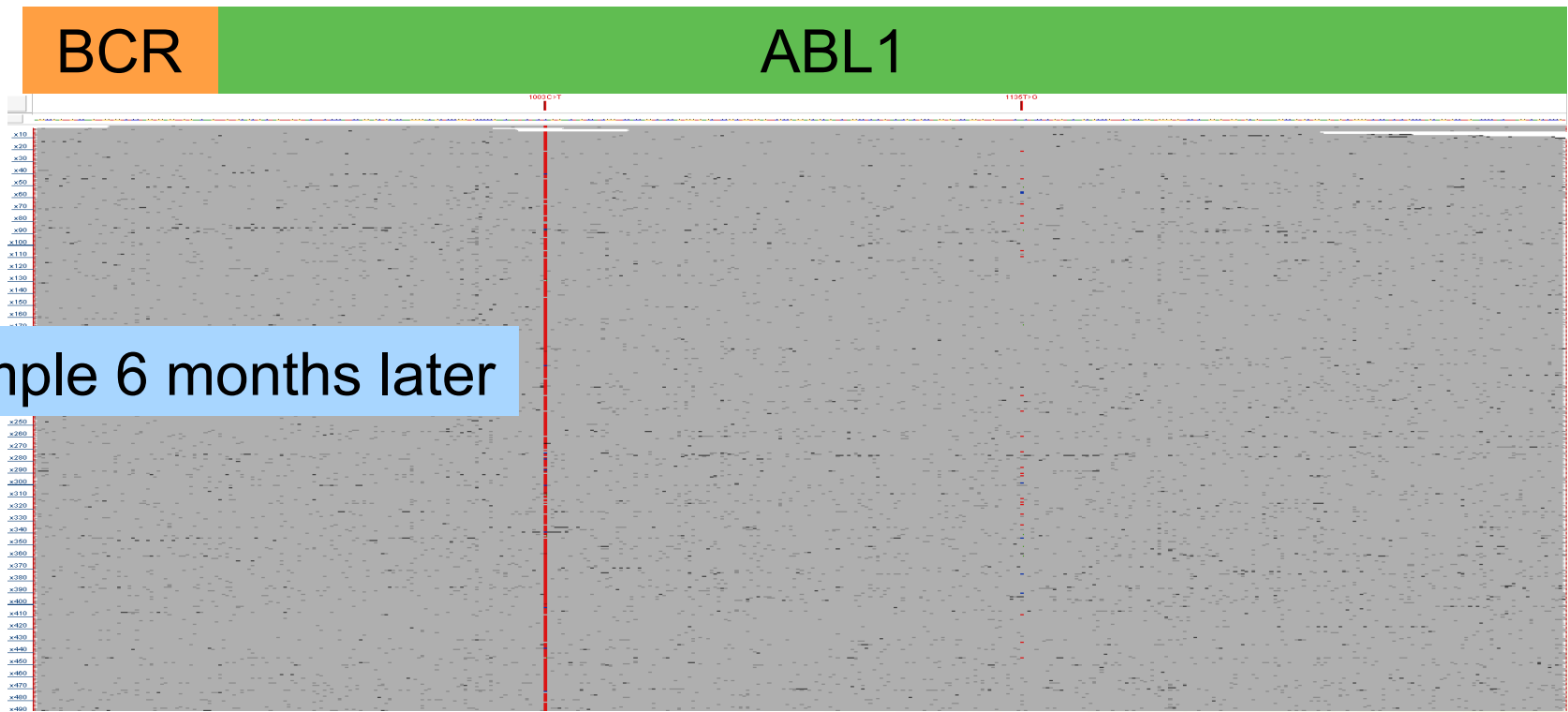
BCR

ABL1



Sample from time of diagnosis:

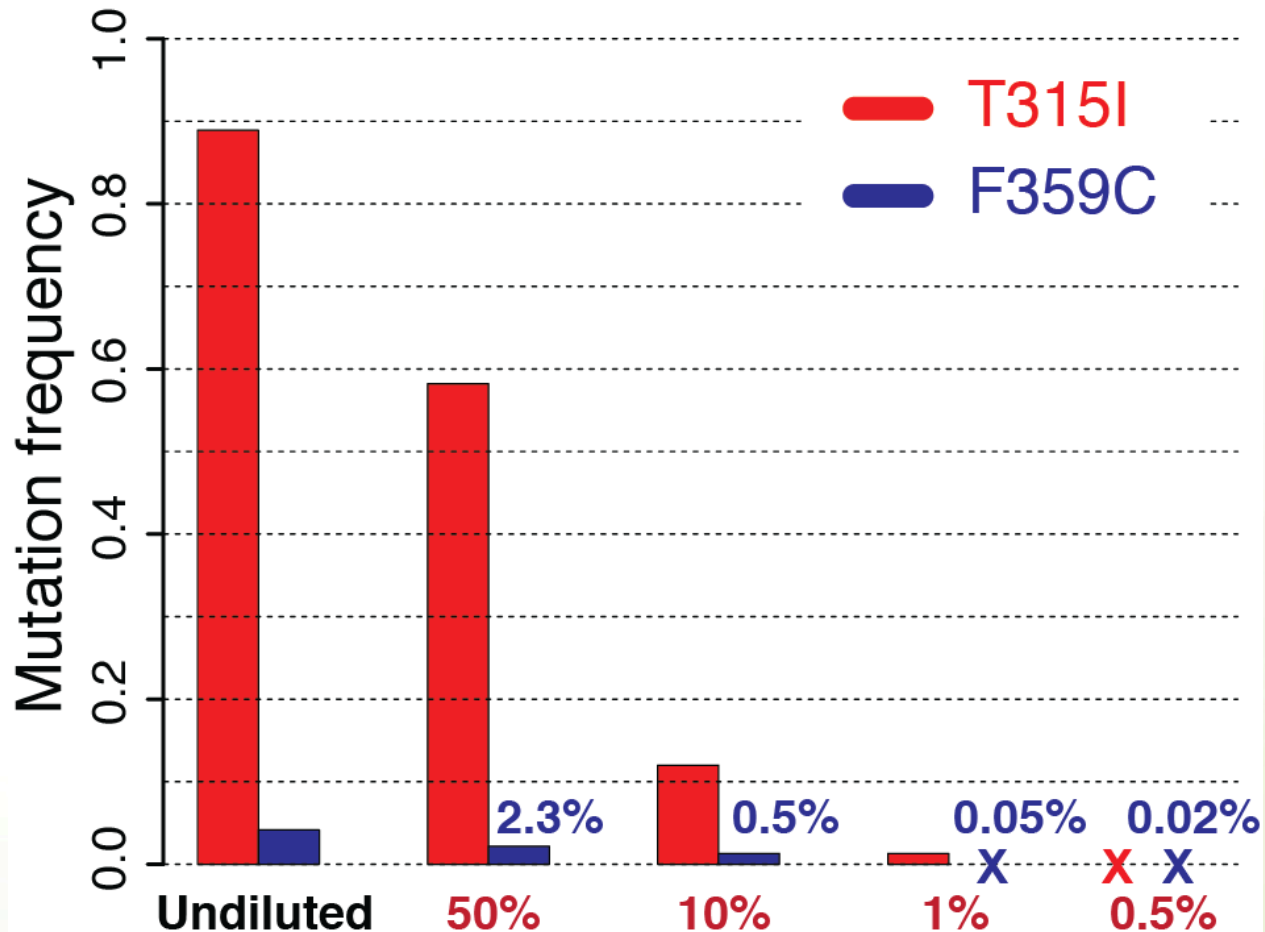
BCR-ABL1 mutations in follow-up sample



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

BCR-ABL1 dilution series results

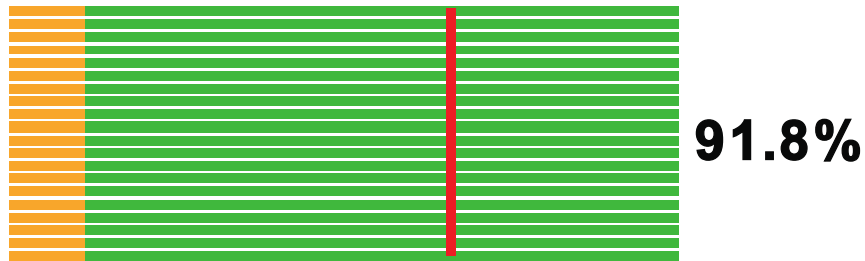
- Mutations down to 1% detected!



BCR-ABL1 - Compound mutations

49 months

T315I



F359C

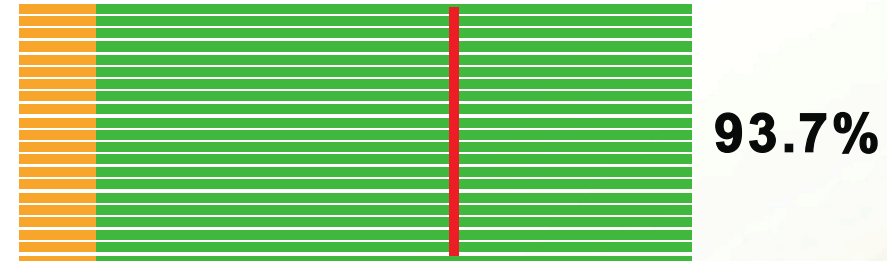


3.9%

A horizontal bar chart with 15 green bars. The first 3 bars on the left are orange, and the remaining 12 bars are green.

55 months

T315I



D276G



F359C



H396R



Analysis method for BCR-ABL1 mutations

- Create CCS reads and screen all known resistance mutations
- CAVA analysis - count number of WT and MUT sequences

WT sequence: **TATATCATCACTGAGTTCATG**

MUT sequence: **TATATCATCA**T**GAGTTCATG**



BCR-ABL1 resistance mutation

- Classify each mutation
 - Less than 500X coverage => **Unresolved**
 - At least 0.5% mutation frequency => **Positive**
 - Otherwise => **Negative**

Web system for sharing results

Details	Sample ID	Run ID	Unresolved (count)	Unknown (count)	M244V	Q252H	Y253H	E255K	E255V	K262N	D276G	T277A	L298V	T315I	T315A	M351T	F359V	L387M	E450G	E453G	E459G	M472I	E499E	Date
---------	-----------	--------	--------------------	-----------------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	------

101 **Sample 102** 103 [New Search](#)

Sample ID
R12095

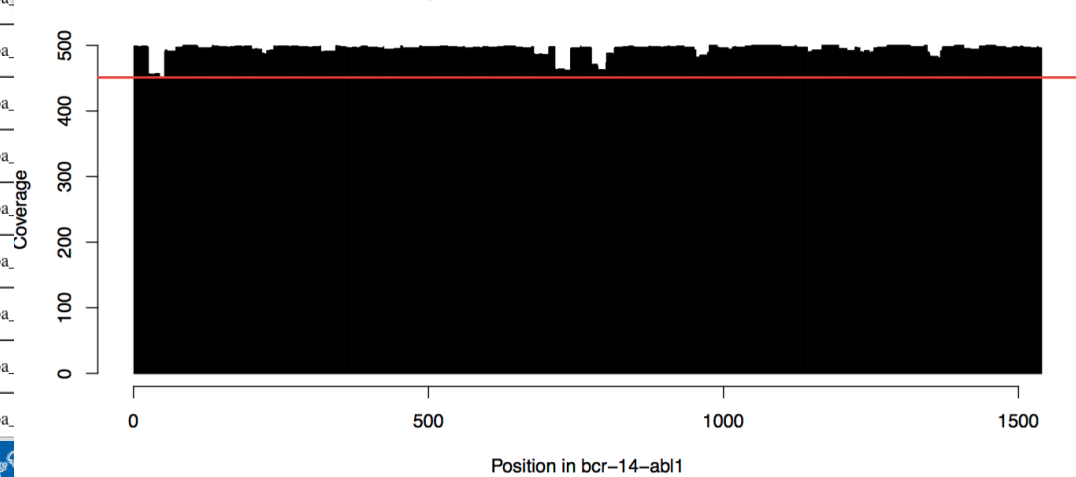
Run ID
cba_012_4

Date
2015-09-17

[Results](#) [Sequence](#) **Downloads:**
[Coverage](#) [Clonal txt](#) [Clonal pdf](#) [Log](#)

mutation	sequence	wt_reads	mut_reads	other_reads	freq	detection
M351T	CACTCAGATCTCGTCAGCCA[T/C]GGAGTACCTGGAGAAGAAAA	16134	19065	3	0.542	positive
Q252H	CACAAGCTGGGCGGGGGCCA[G/C]TACGGGGAGGTGTACCAGGG	12052	9920	8	0.451	positive
K262N	GTGTACGAGGGCGTGTGGAA[G/T]AAATACAGCCTGACGGTGGC	25597	6996	16	0.215	positive
M244V	TGGAACGCACGGACATCACC[A/G]TGAAGCACAAGCTGGGCGGG	32779	32	2	0.001	negative
K247R	GGACATCACCATGAAGCACA[A/G]GCTGGGCGGGGGCCAGTACG	27076	32	9	0.001	negative

Coverage of bcr-14-abl1, based on 500 reads



Frequency	Reads	detection
49.9 %	9268	negative
23.8 %	4418	negative
17.4 %	3245	negative
8.69 %	1613	negative

Clinical Diagnosis of BCR-ABL1 mutations

Clinical Genetics



- Collection of samples
- Seq library preparation

Sequencing Facility



- SMRT sequencing
- CAVA analysis

IT developers

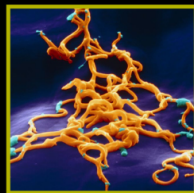
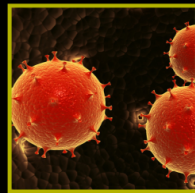
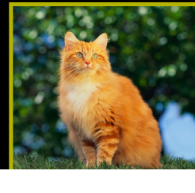
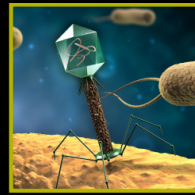
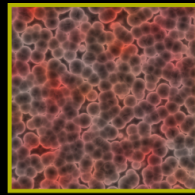


- Web server for results

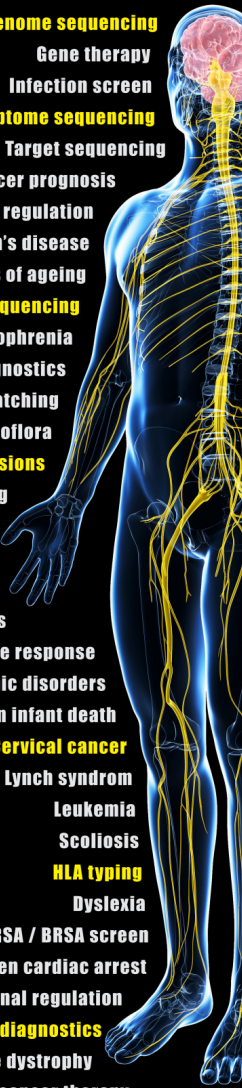
- Ongoing routine service, 0-4 samples/week.
- Over 150 patient samples run so far
- 100% consistency with Sanger results

What we sequence at NGI /

SciLifeLab



THANK YOU

- 
- Diabetes
 - Alzheimer's disease
 - Whole-genome sequencing**
 - Gene therapy
 - Infection screen
 - Whole-transcriptome sequencing**
 - Target sequencing
 - Cancer prognosis
 - Gene regulation
 - Crohn's disease
 - Genomics of ageing
 - Exome sequencing**
 - Schizophrenia
 - Cancer diagnostics
 - Organ donor matching
 - Gut microflora
 - Gene fusions**
 - RNA editing
 - HIV
 - HPV**
 - HCV
 - Scoliosis
 - Immune response
 - Monogenic disorders
 - Sudden infant death
 - Cervical cancer**
 - Lynch syndrome
 - Leukemia
 - Scoliosis
 - HLA typing**
 - Dyslexia
 - MRSA / BRSa screen
 - Sudden cardiac arrest
 - Transcriptional regulation
 - Prenatal diagnostics**
 - Muscle dystrophy
 - Individualised cancer therapy
 - and much more...