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

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

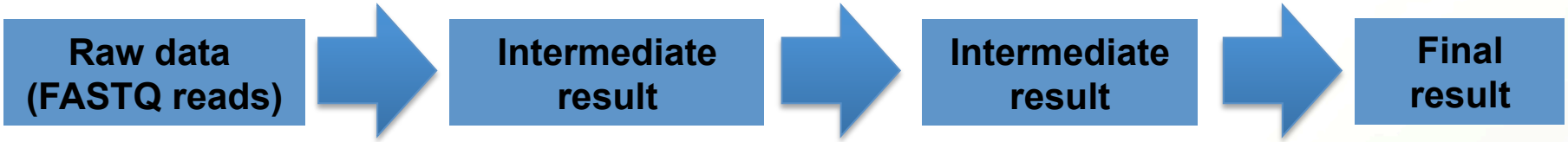
ion torrent

PACIFIC BIOSCIENCES

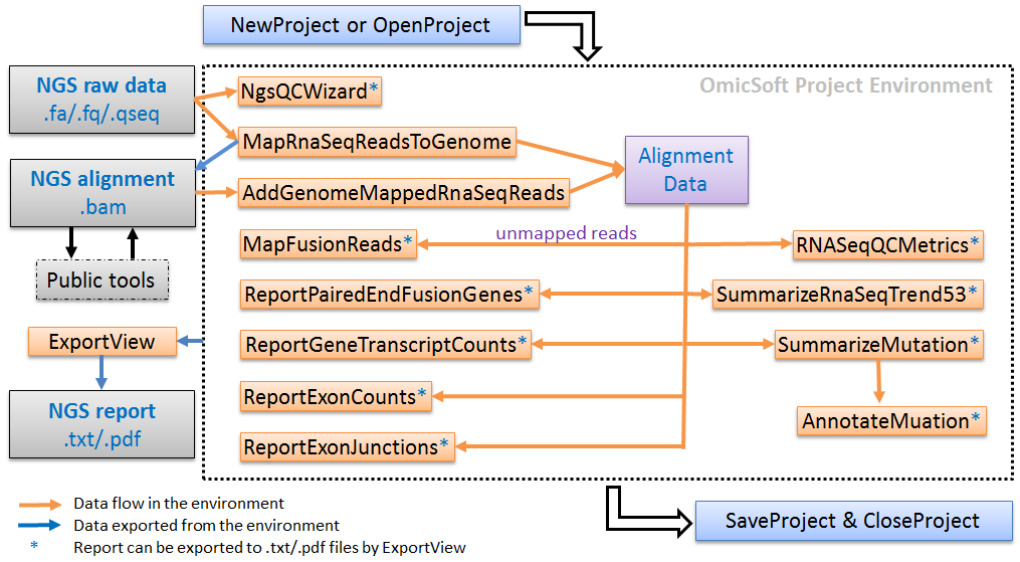
Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet · SciLifeLab

What is an analysis pipeline?

- Basically just a number of steps to analyze data



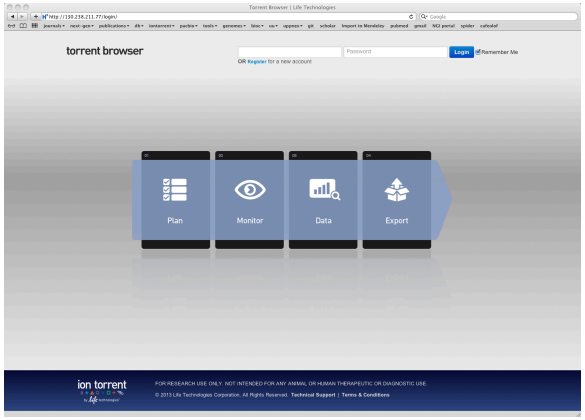
- Pipelines can be simple or very complex…



Some analysis pipelines for NGS data

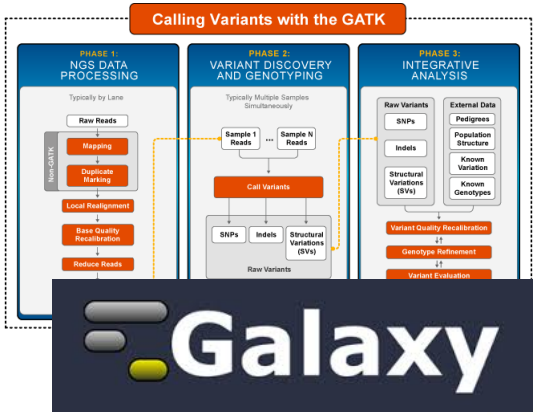
Ion Torrent

Torrent Suite Software



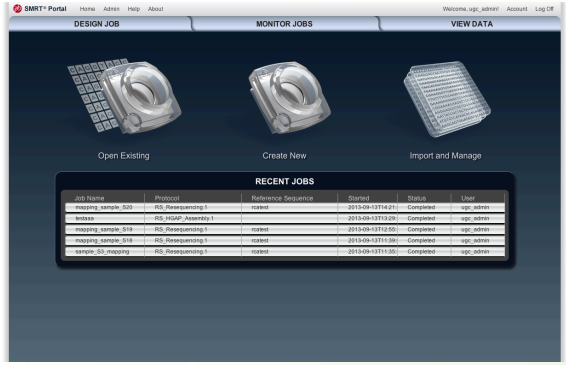
Illumina

GATK, Galaxy,…



PacBio

SMRT analysis portal

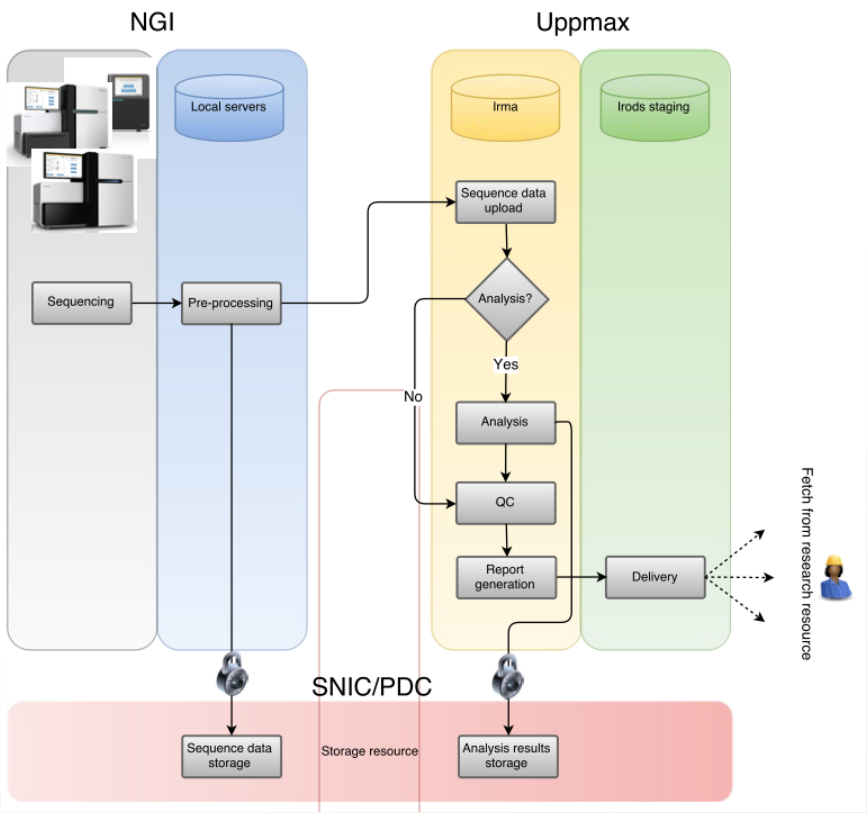


- Enables variant calling, de novo assembly, RNA expression analyses, …
- Many other tools exists, also from commercial vendors

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

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

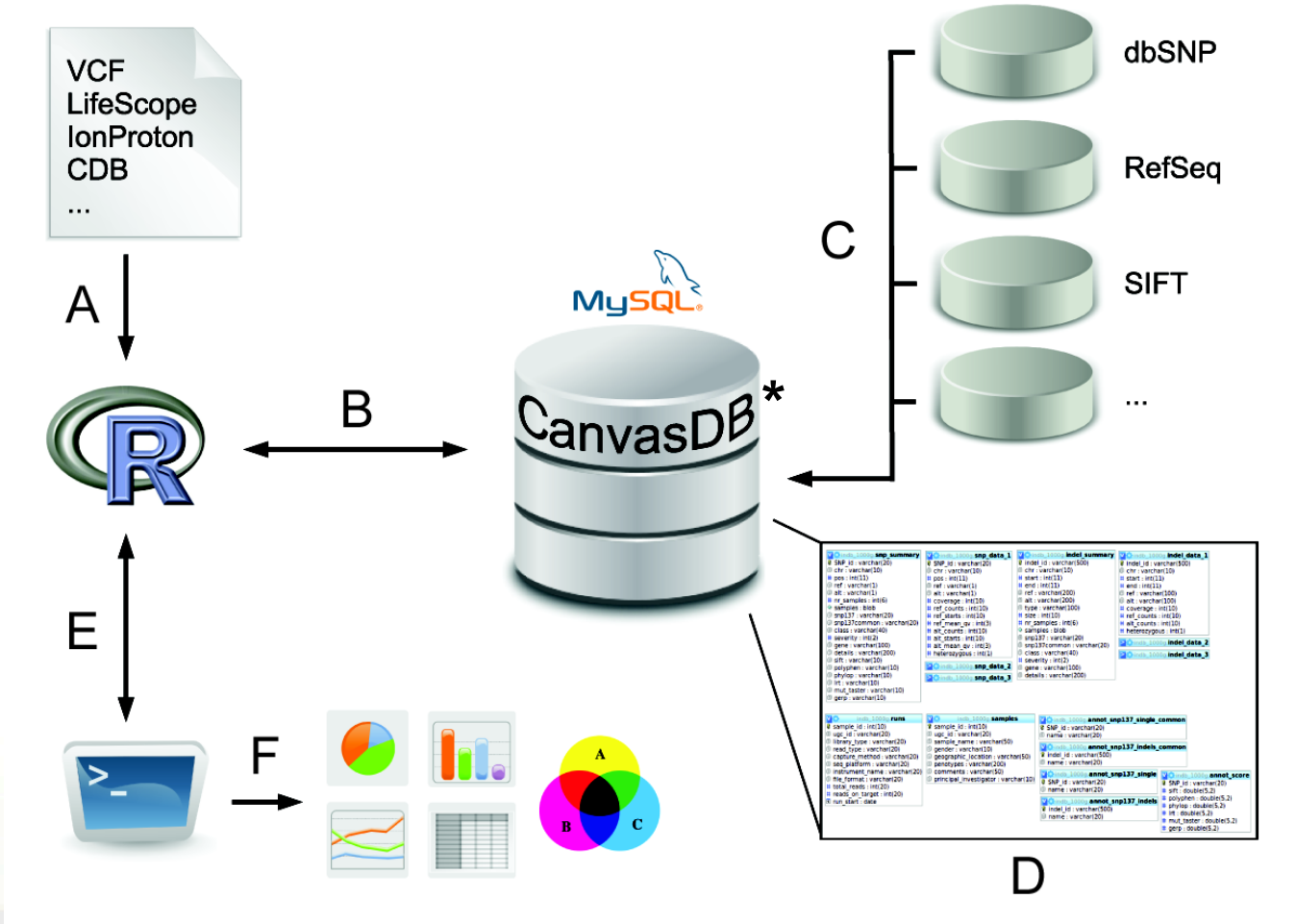
Some examples follows:

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

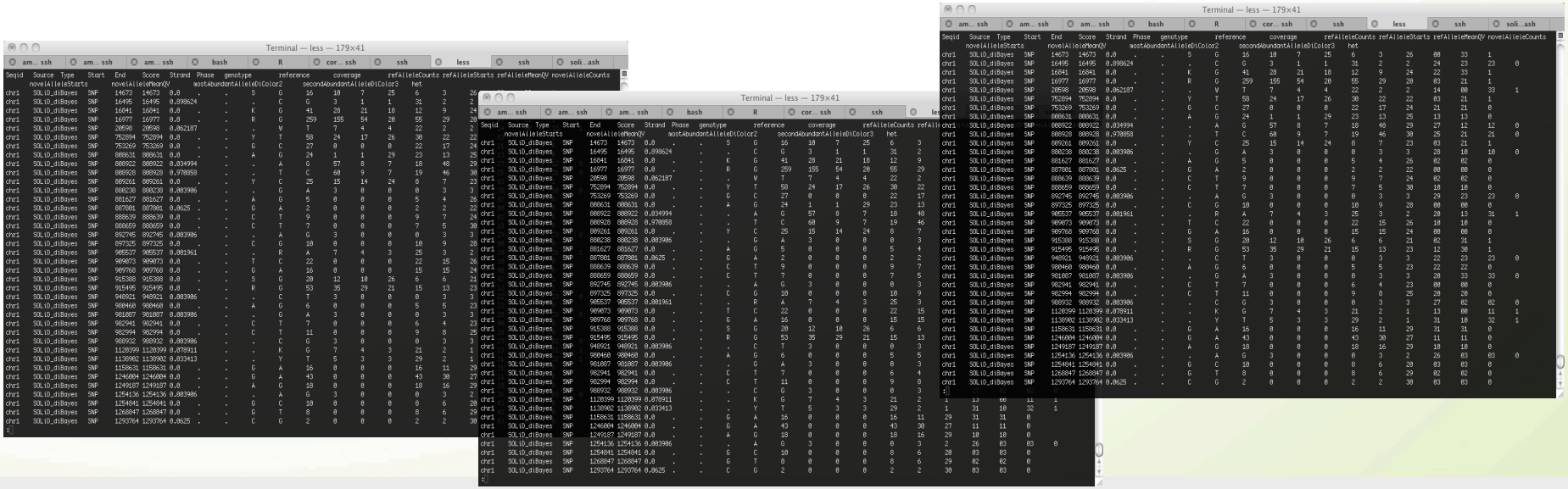
- Advantages
  - Allows investigate all protein coding sequences
  - Possible to detect both SNPs and small indels
  - Low cost (compared to WGS)
  - Possible to multiplex several exomes in one run
  - Standardized work flow for data analysis

- 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

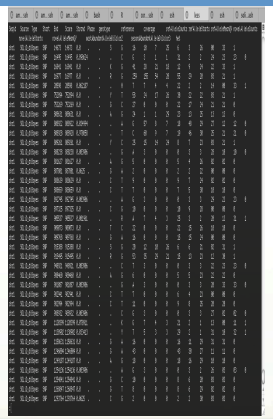


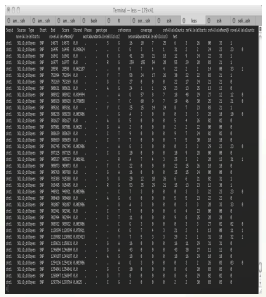
Karolinska Institutet
KTH Vetenskap och Konst
Stockholms universitet
Uppsala Universitet
SciLifeLab

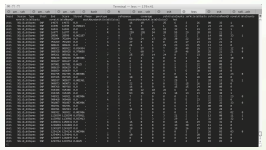
How to analyze this?

- 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)

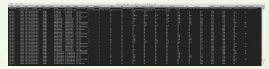
Start:
All identified SNPs







Result:
A few candidate causative SNP(s)!

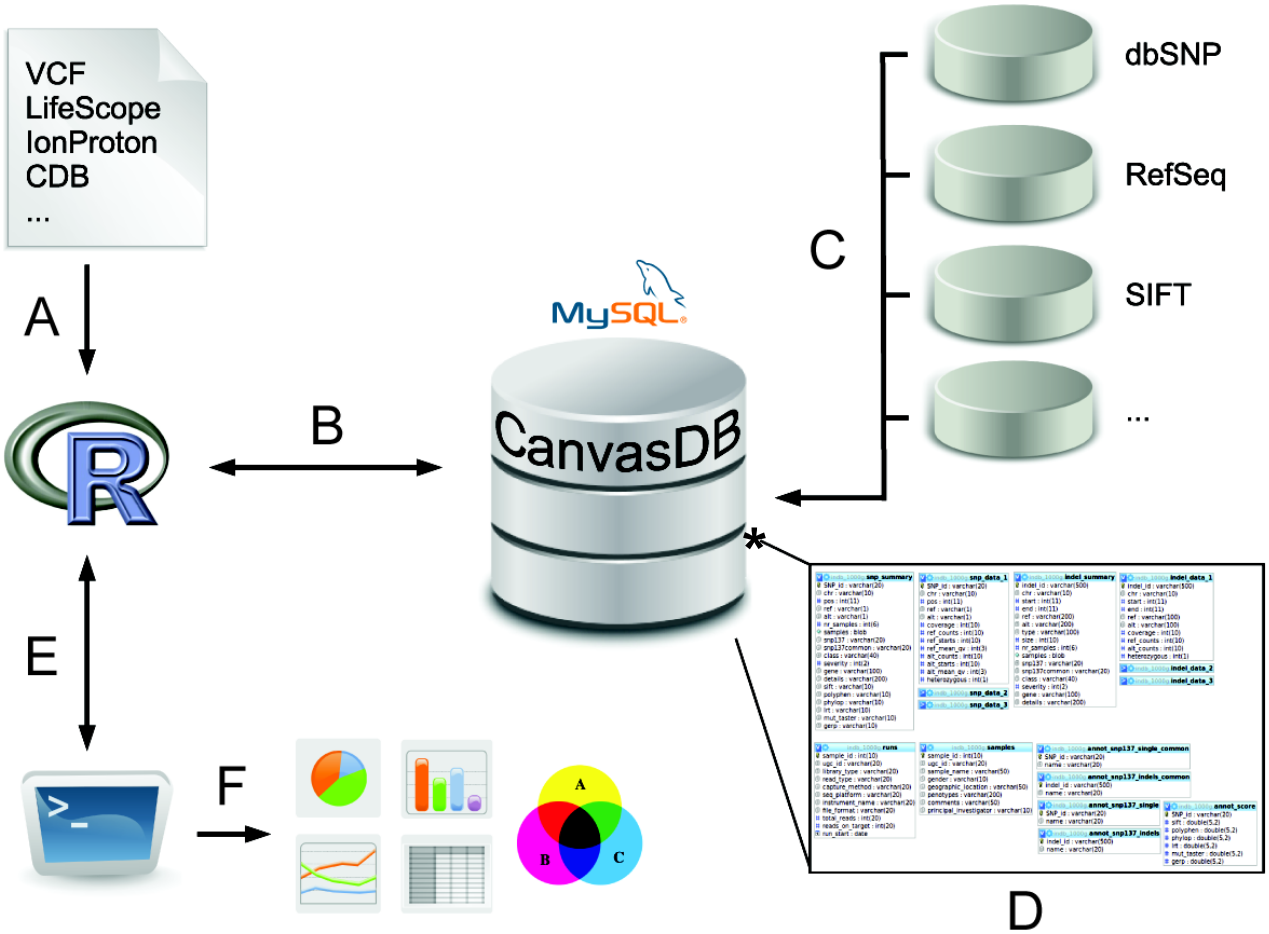


Why is this not optimal?

- 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

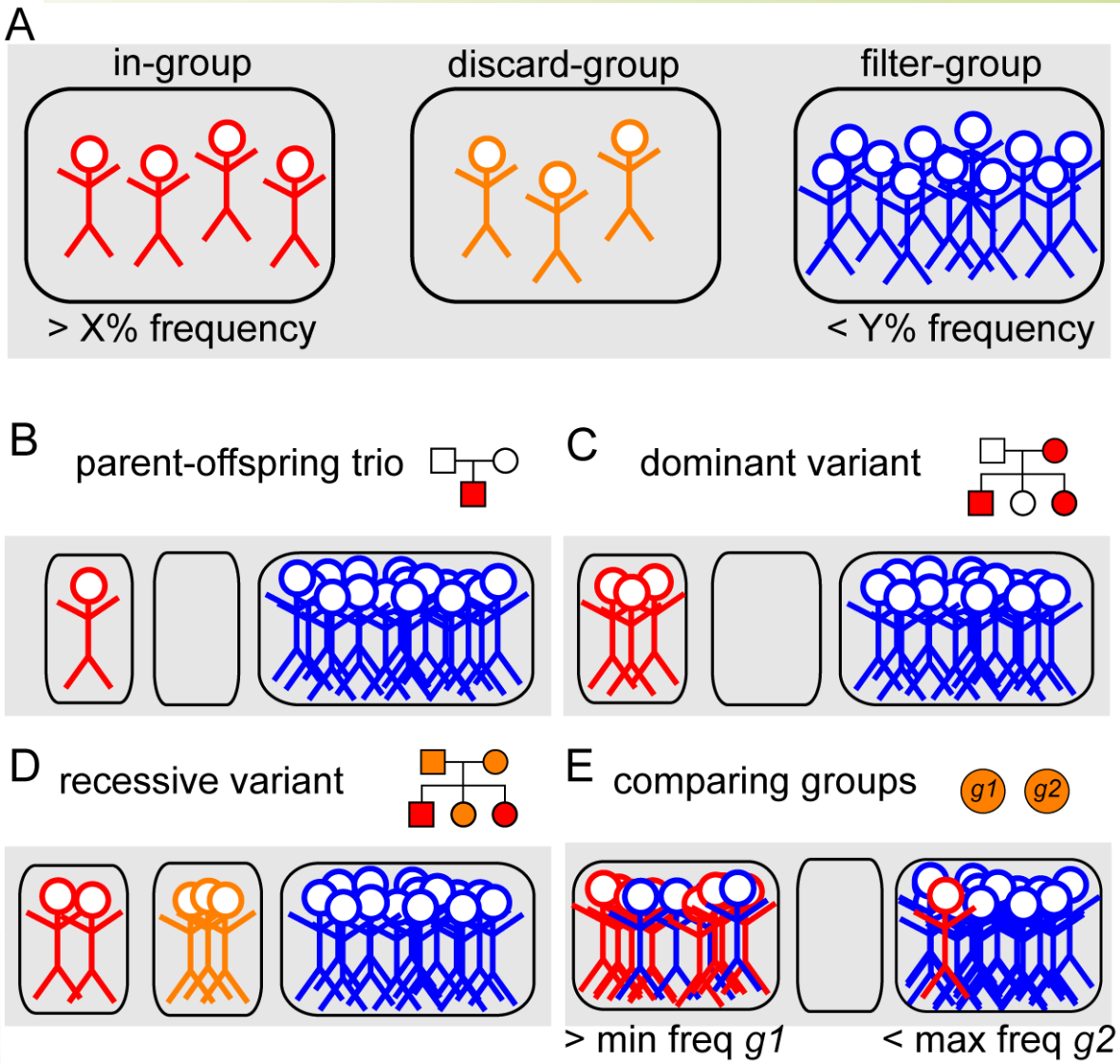


***CANdidate Variant Analysis System and Data Base**

Ameur et al., Database Journal, 2014

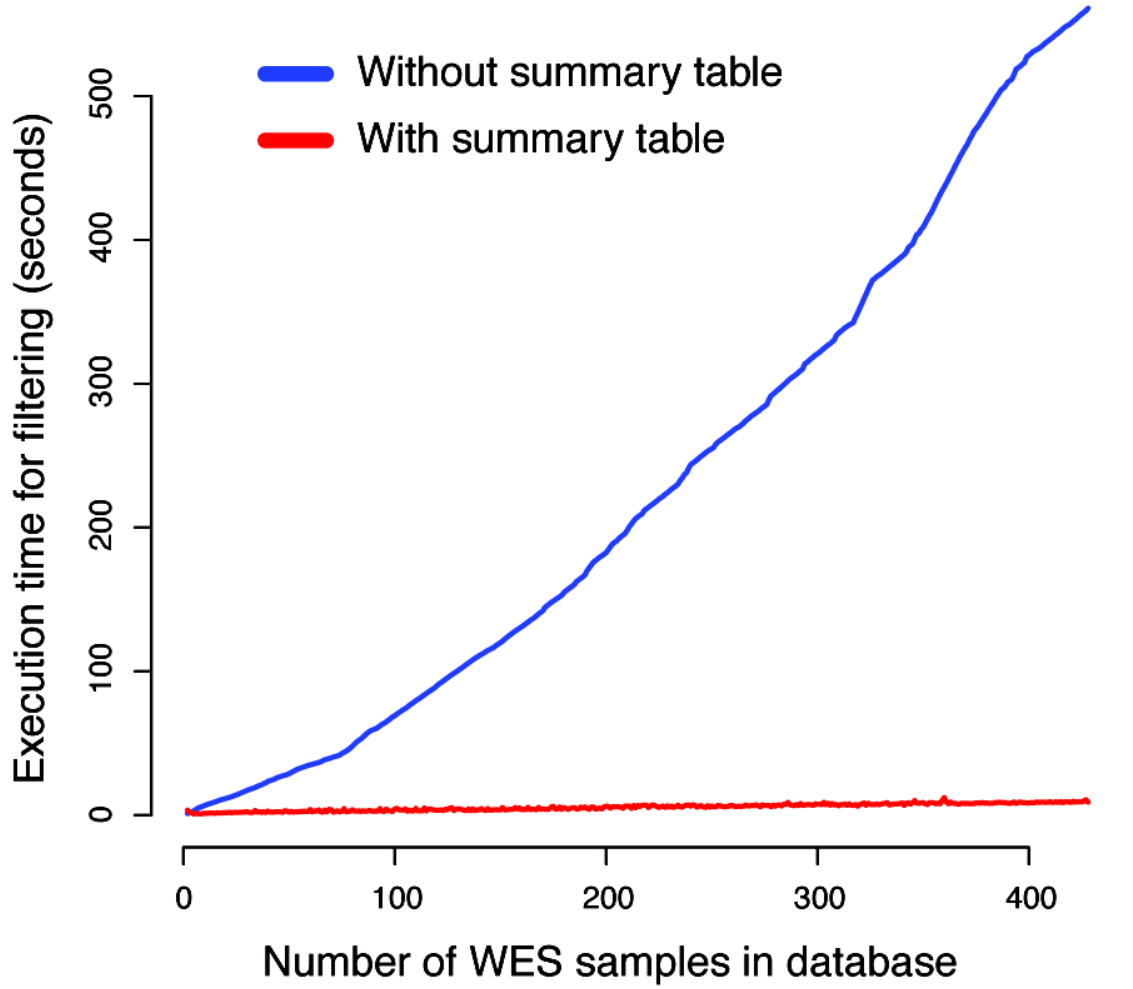
Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet · SciLifeLab

CanvasDB - Filtering



CanvasDB - Filtering speed

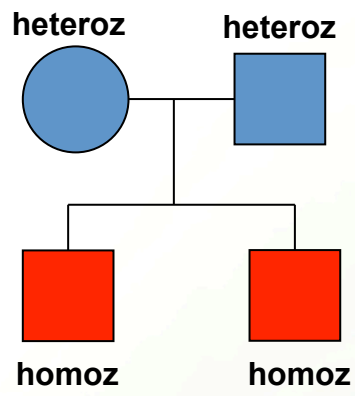
- Rapid variant filtering, also for large databases



Karolinska Institutet
KTH
Stockholms universitet
UPPSALA UNIVERSITET
SciLifeLab

A recent exome-seq project

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

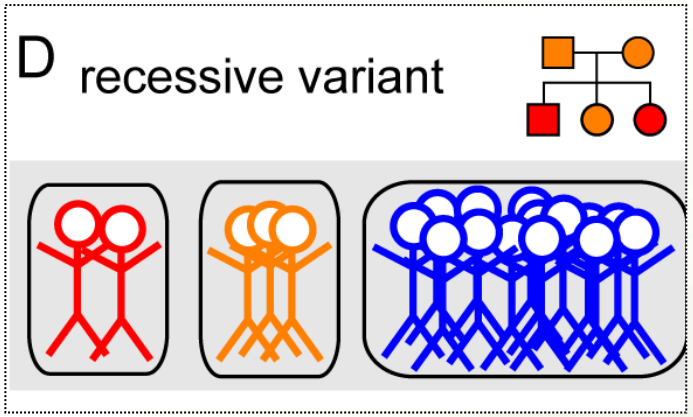


- Sequencing strategy
  - TargetSeq exome capture
  - One sample per PI chip

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

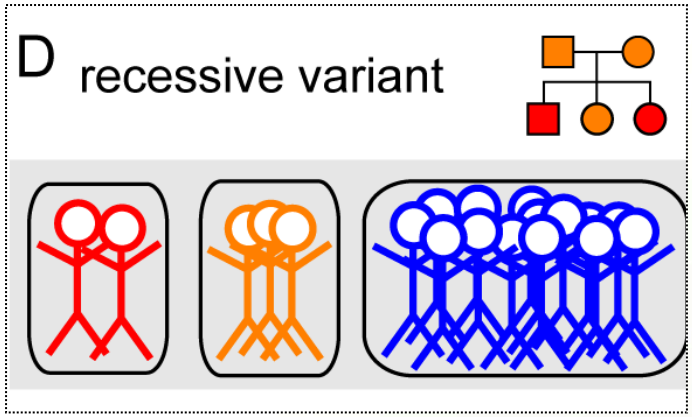
- *CanvasDB* filtering for a variant that is…
  - rare
    - at most in 1% of ~700 exomes
  - shared
    - found in both brothers
  - homozygous
    - in brothers, but in no other samples
  - deleterious
    - non-synonymous, frameshift, stop-gain, splicing, etc..




```
> 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

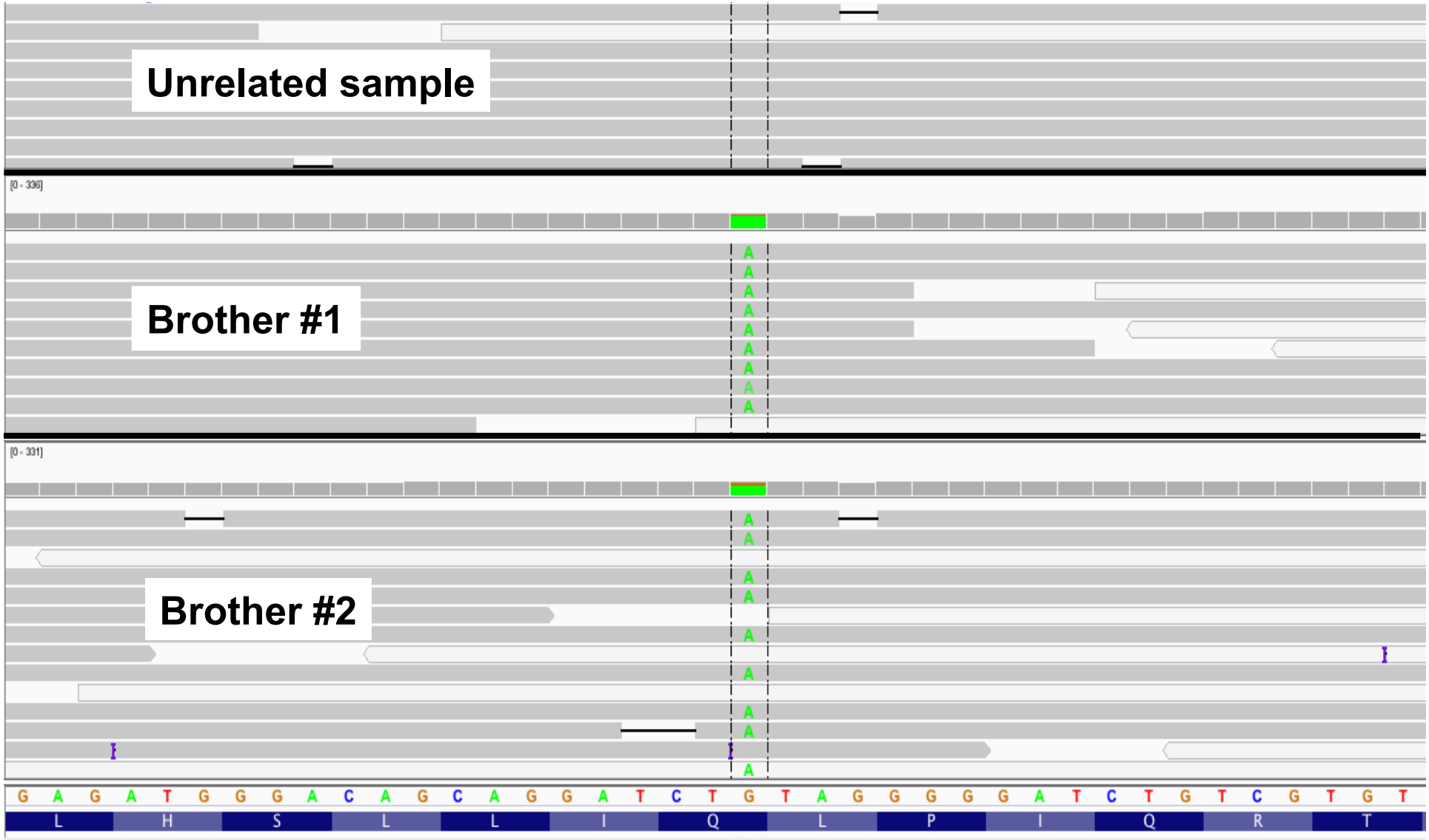
From Wikipedia, the free encyclopedia

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!

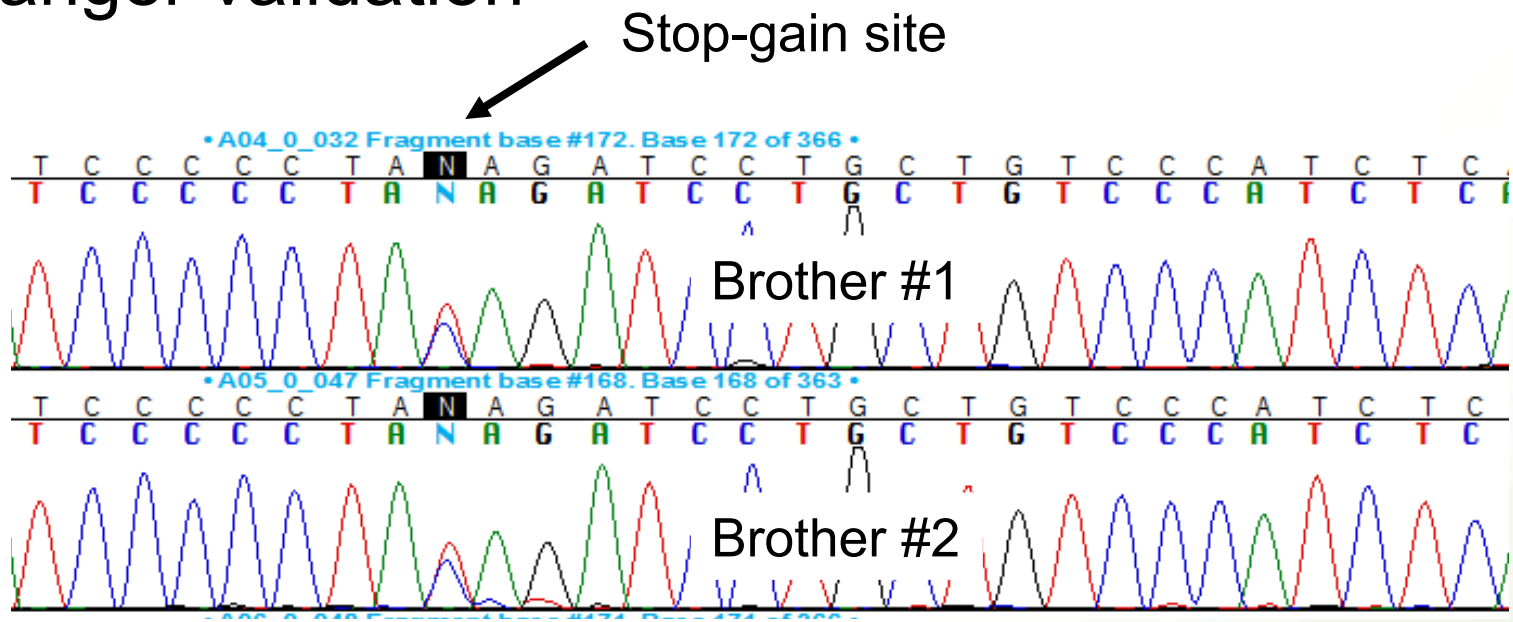
IGV visualization: Stop gain in STRC



Karolinska Institutet
KTH
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 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!

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

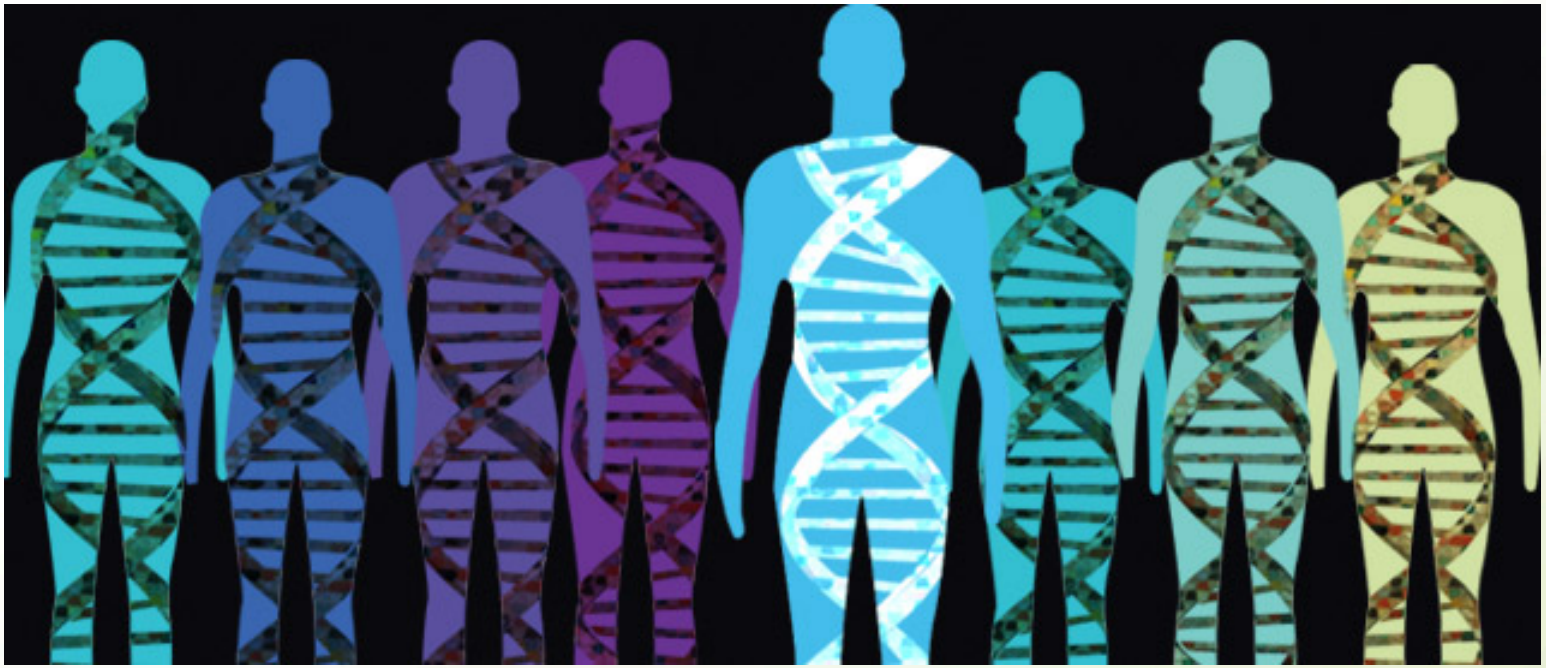
Genomics
england

SciLifeLab Uppsala Genome Center SNP&SEQ Technology Platform National Genomics Infrastructure Karolinska Institutet KTH Stockholms universitet Uppsala Universitet

The SciLifeLab Human WGS Initiative

- WGS of patient cohorts (n=10,000 ind/year)

- Genetic Variant Database for the Swedish Population (n=1000)



SciLifeLab Uppsala Genome Center SNP&SEQ Technology Platform National Genomics Infrastructure Karolinska Institutet KTH Stockholms universitet Uppsala Universitet

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.

SciLifeLab Uppsala Genome Center SNP&SEQ Technology Platform National Genomics Infrastructure Karolinska Institutet KTH Stockholms universitet Uppsala Universitet

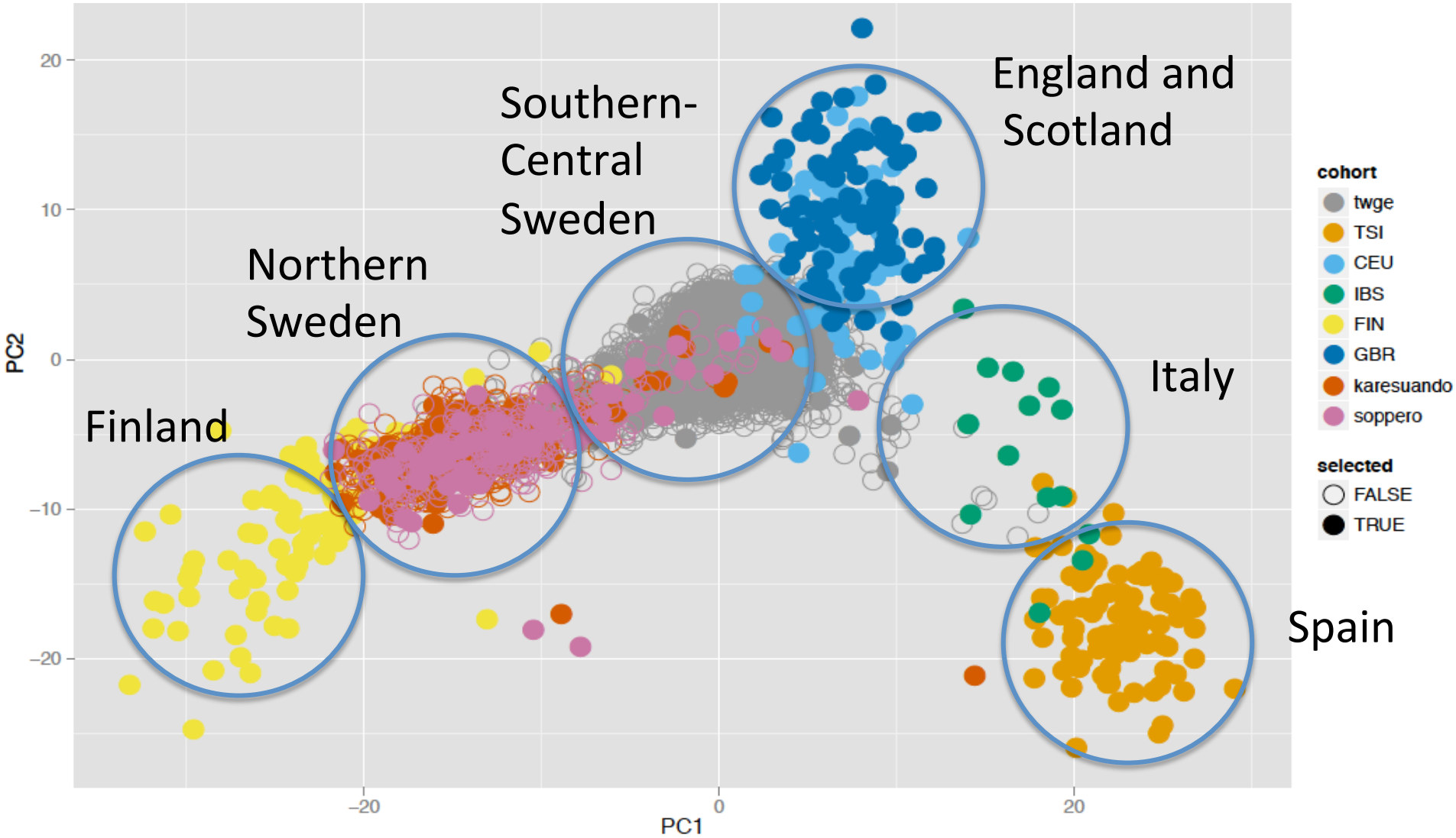
Twin Registry samples used as control cohort

- 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

SciLifeLab Uppsala Genome Center SNP&SEQ Technology Platform National Genomics Infrastructure Karolinska Institutet KTH Stockholms universitet Uppsala Universitet

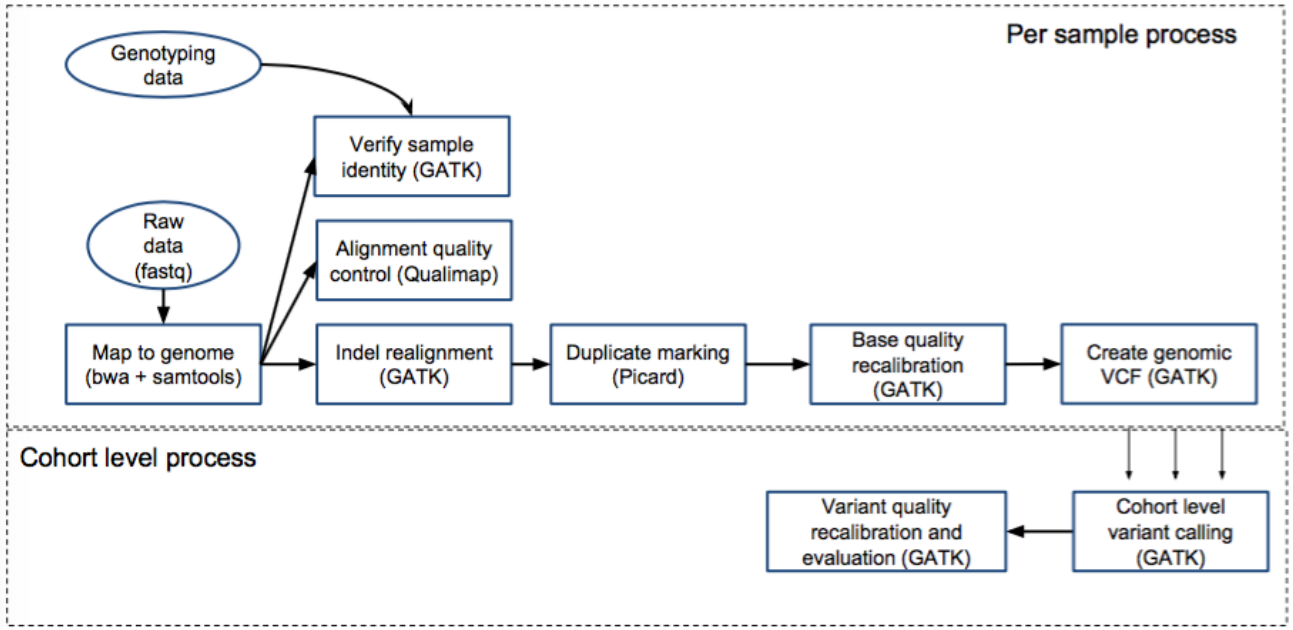
Principal components of European samples from 1,000 genomes project and 10,000 Swedish samples



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

A web server for ‘SweGen’ data

SweFreq | About | Terms of use | Data Beacon | ExAC Browser | Download Data | Admin | Adam Ameur Logout adam.ugc@gmail.com

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.

SweGen

Released on Oct 19th! Data available from: swefreq.nbis.se

Karolinska Institutet | KTH | Stockholms universitet | UPPSALA UNIVERSITET | SciLifeLab

Example II:

Assembly of genomes using Pacific Biosciences

Karolinska Institutet
KTH
Stockholms universitet
UPPSALA UNIVERSITET
SciLifeLab

Genome assembly using NGS

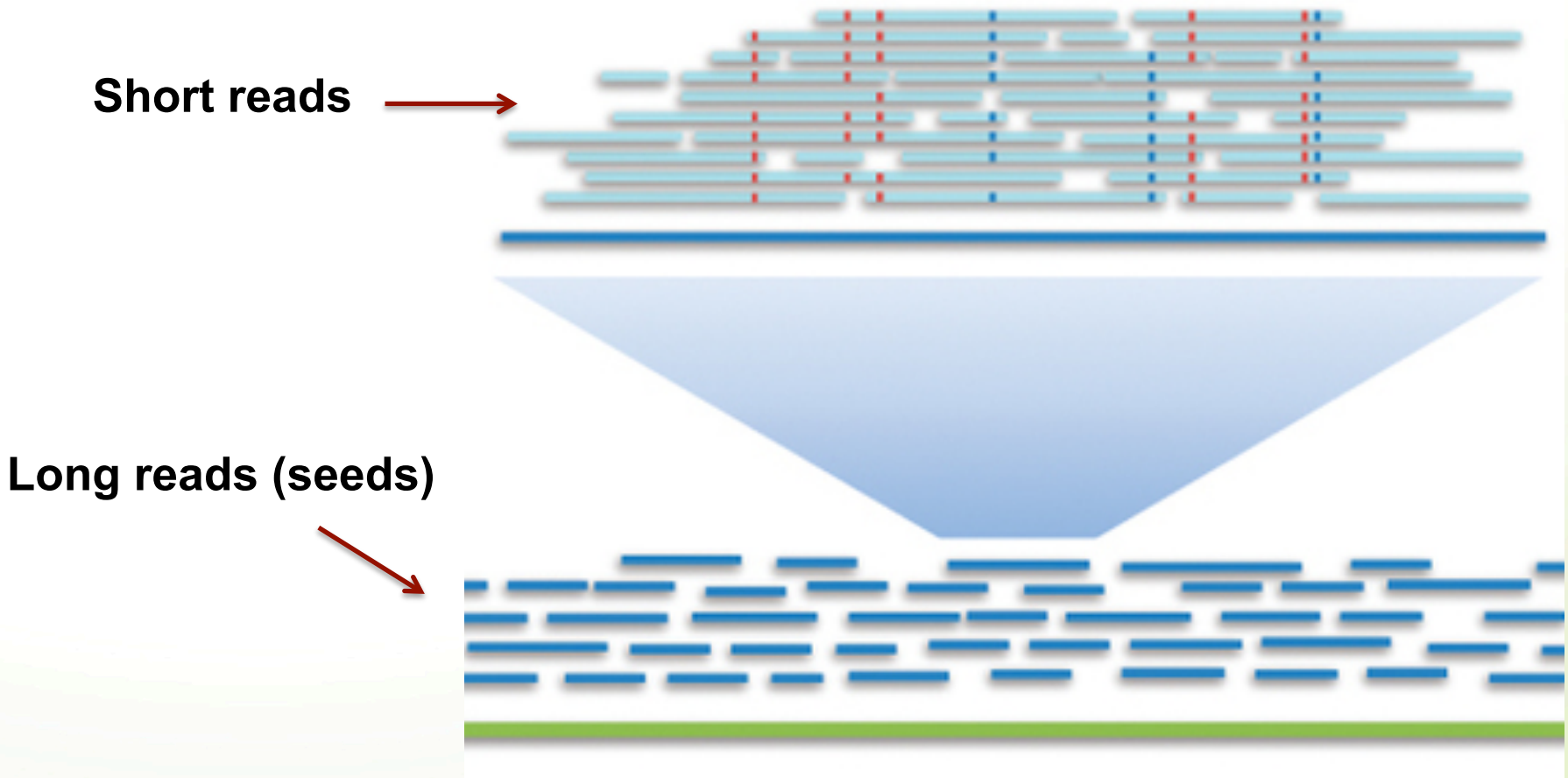
- Short-read *de novo* assembly by NGS
  - Requires mate-pair sequences
    - Ideally with different insert sizes
  - Complicated analysis
    - Assembly, scaffolding, finishing
    - Maybe even some manual steps

  => 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



PacBio assembly analysis

- Simple -- just click a button!!

hgap_fig.PNG (PNG Image,... | Details of Job assembly

127.0.0.1:8080/smrtportal/#/Design-Job/Details-of-Job/16497

pacbio de novo assembly

Most Visited | Getting Started | Save to Mendeley

SMRT® Portal | Home | Admin | Tech Support Files | Help | About | Welcome, ugc_admin! | Account | Log

DESIGN JOB | MONITOR JOBS | VIEW DATA

Job Name: assembly | Comments: | Groups: all | User: ugc_admin

Protocols: RS_HGAP_Assembly.3 | Reference: [None selected]

SMRT Cells Available (Viewing 1 - 31 of 31) | Search | Go | Advanced

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)

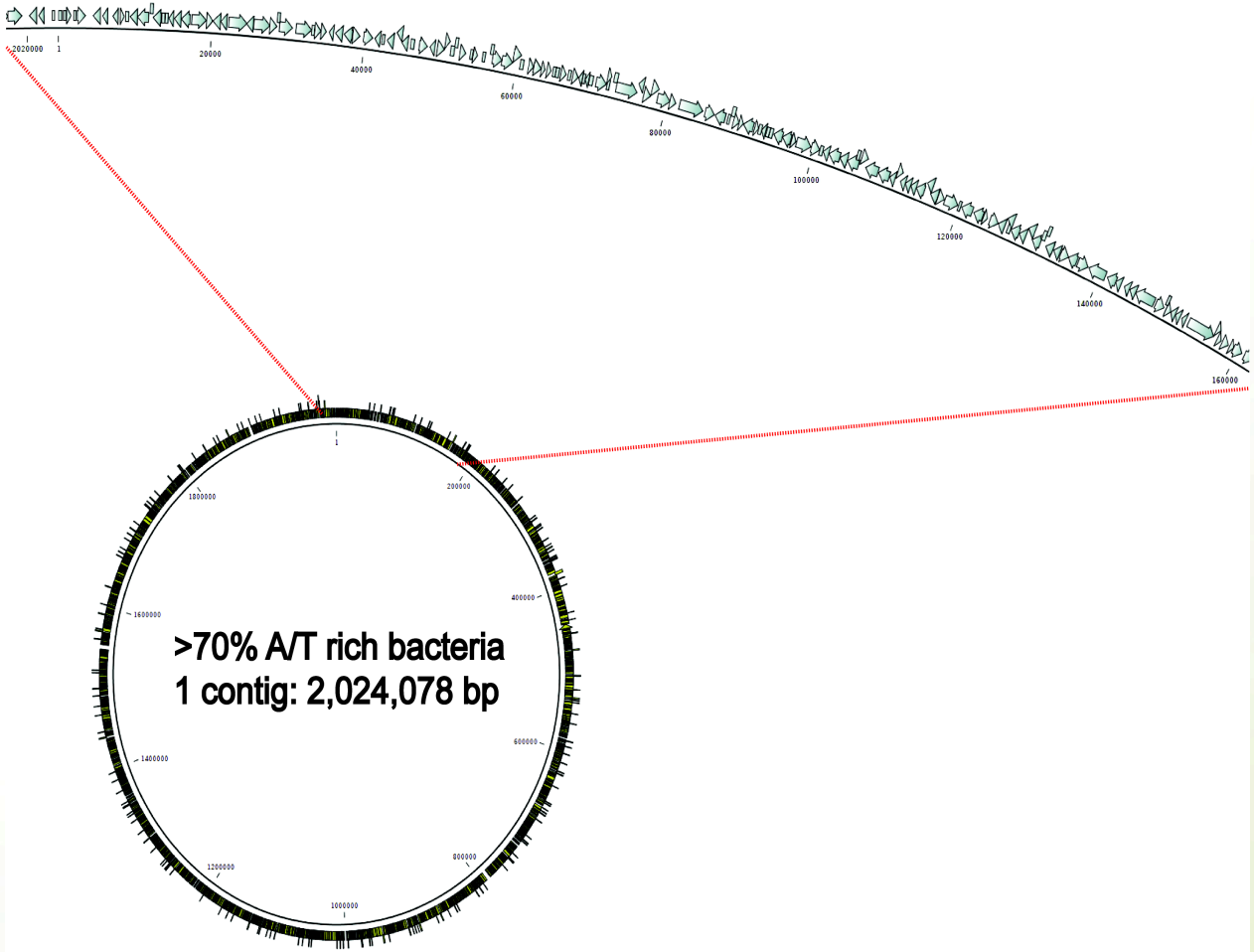
Sample	Version	User	Groups	Uri
Pb33_1	2.0.2		all	/home/pacbio/DATA/adam/Pb_33_F

Show/Hide Columns | Group Rows | Export Table Data

Start | Save | Copy | Cancel

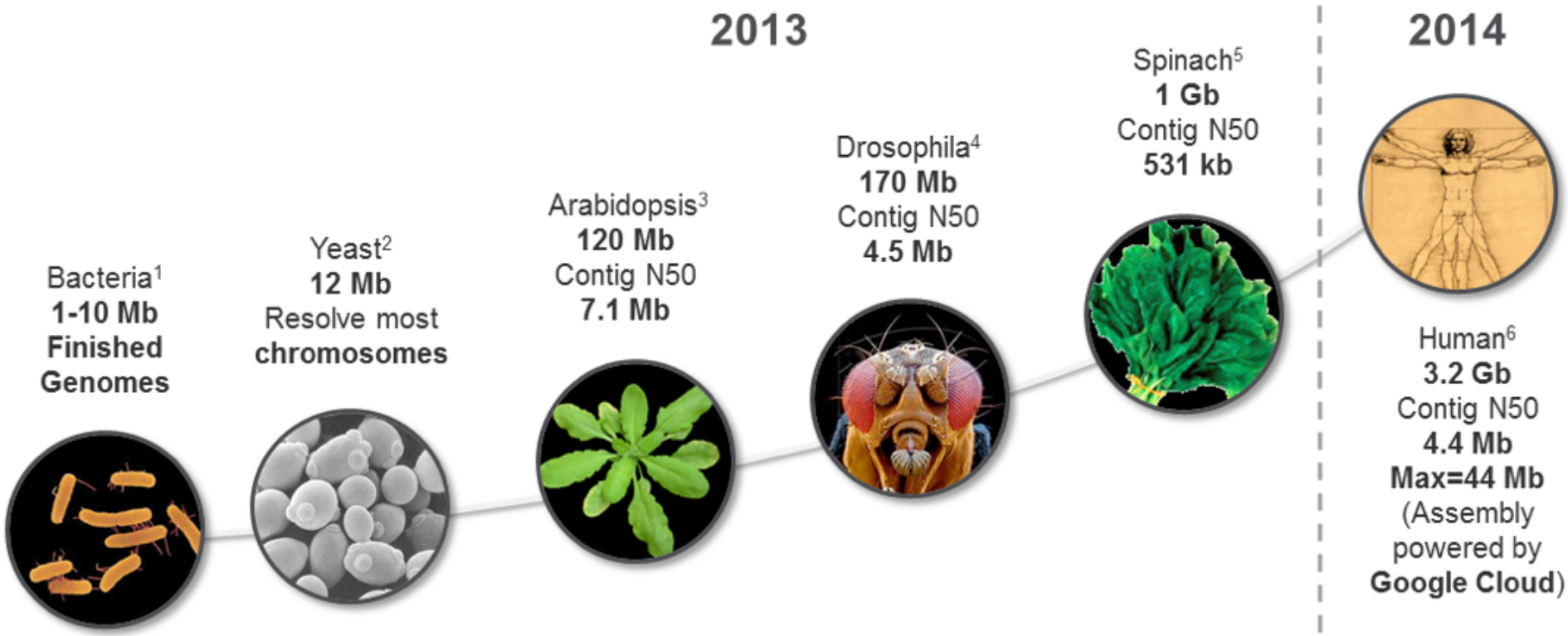
PacBio assembly, example result

- Example: Complete assembly of a bacterial genome



PacBio assembly – recent developments

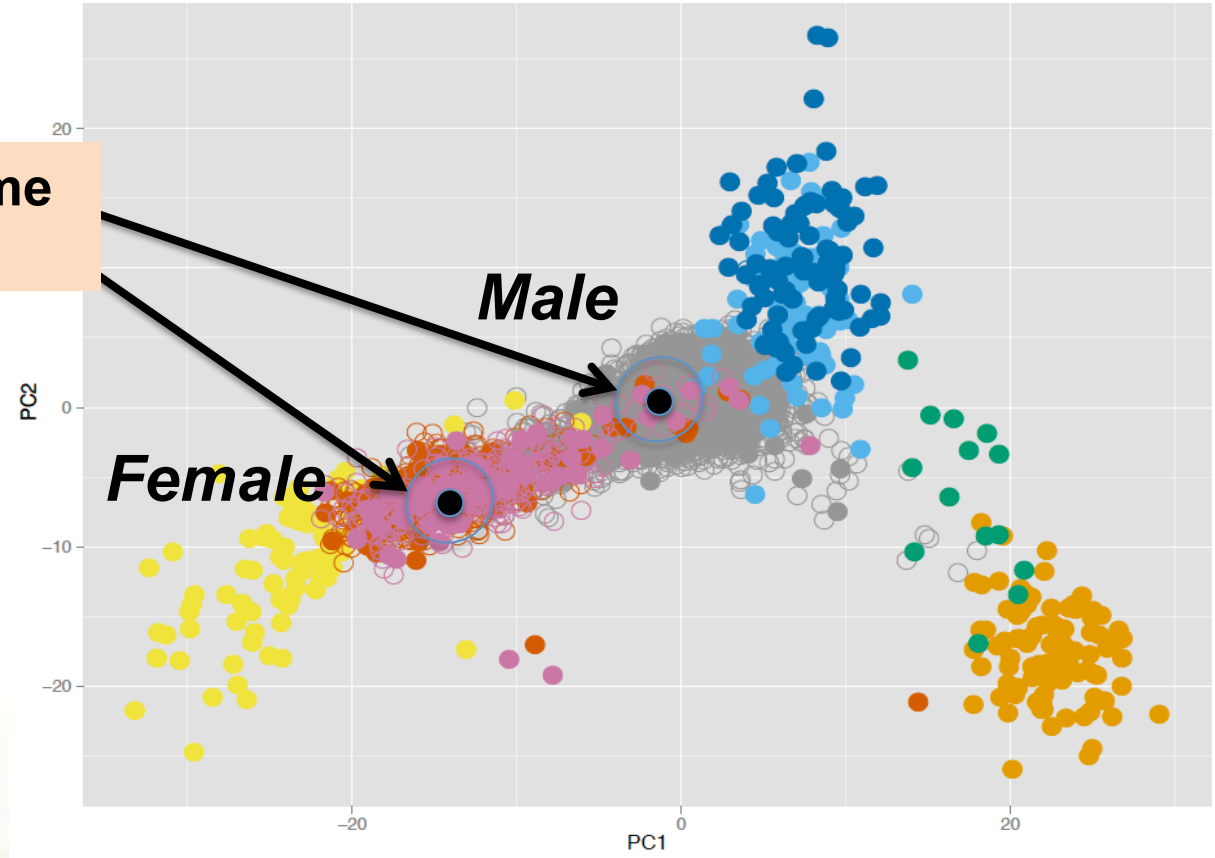
- Also larger genomes can be assembled by PacBio..



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

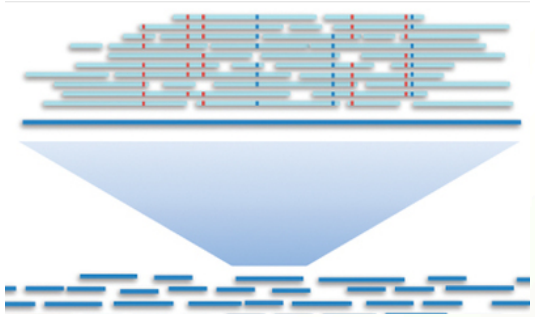


Karolinska Institutet
KTH
Stockholms universitet
UPPSALA UNIVERSITET
SciLifeLab

De novo assembly of 75X PacBio data

Assembly (FALCON)

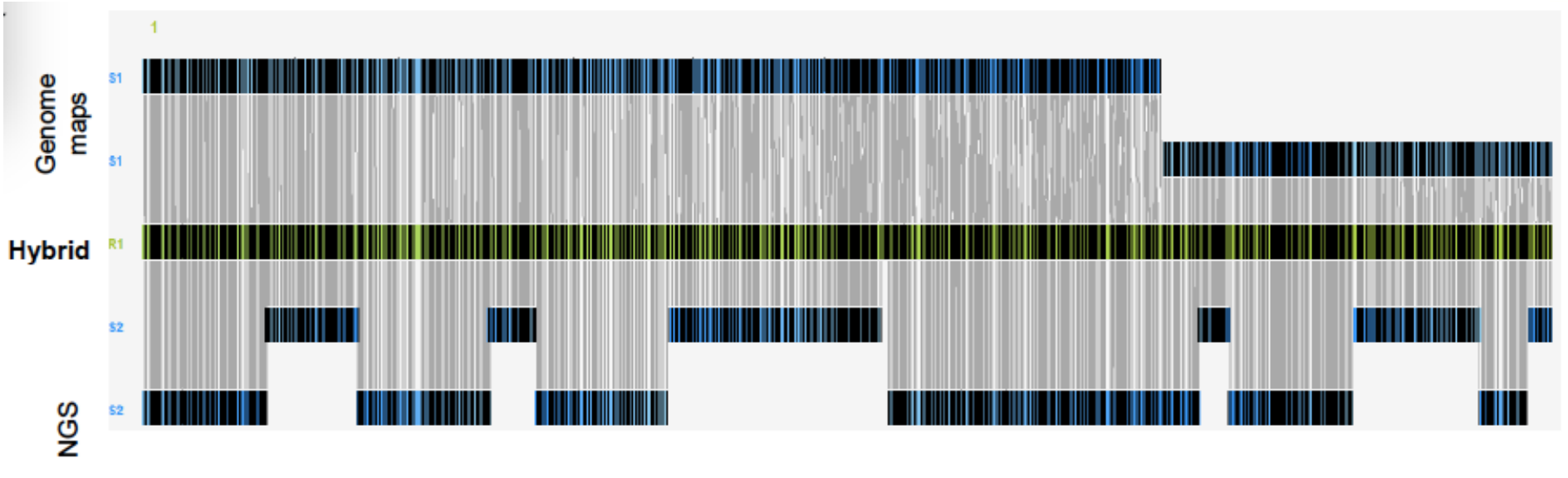
Error correction (2 x Quiver)



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

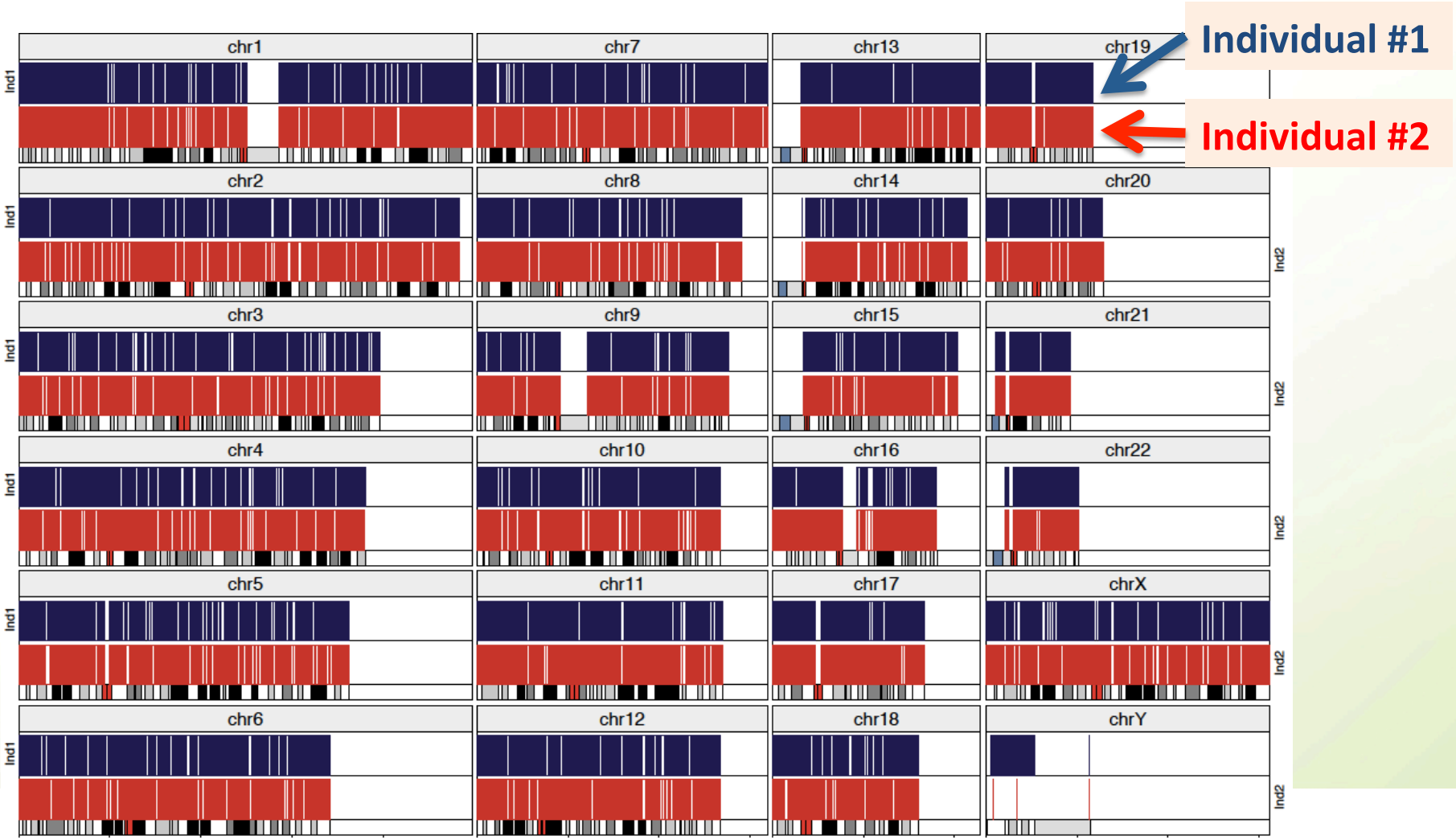


Hybrid scaffolding with two labellings resulted in

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

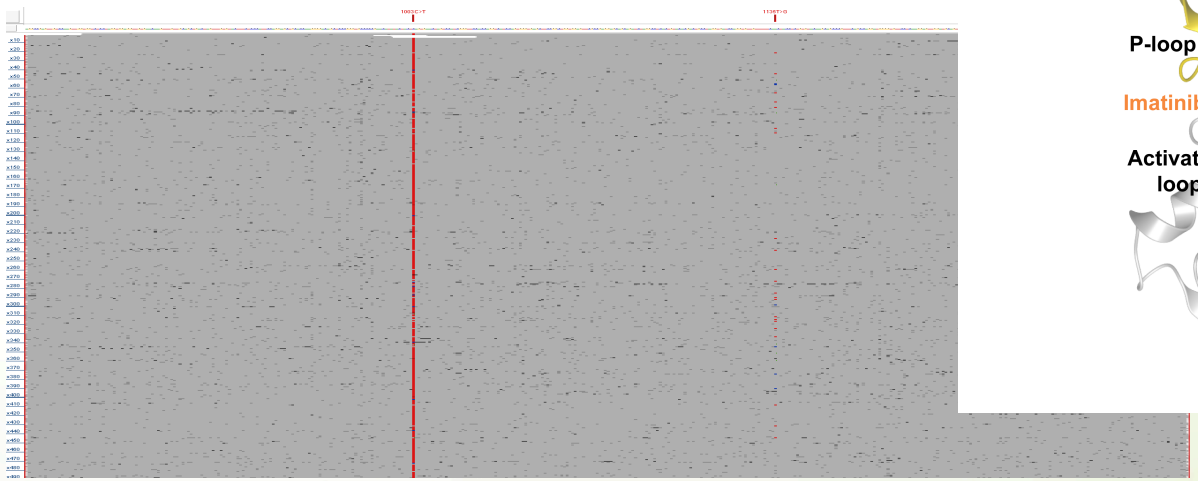
Aligning contigs to human reference

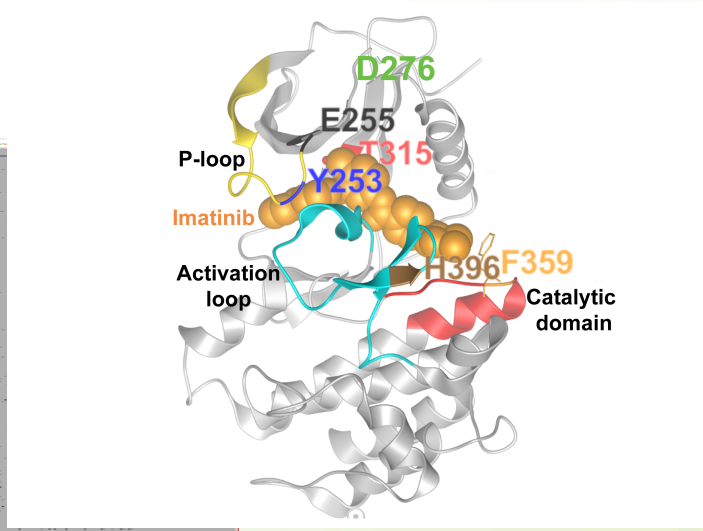
> 99% of bases can be aligned to human reference (hg38)



Example III:

Clinical sequencing for Leukemia Treatment

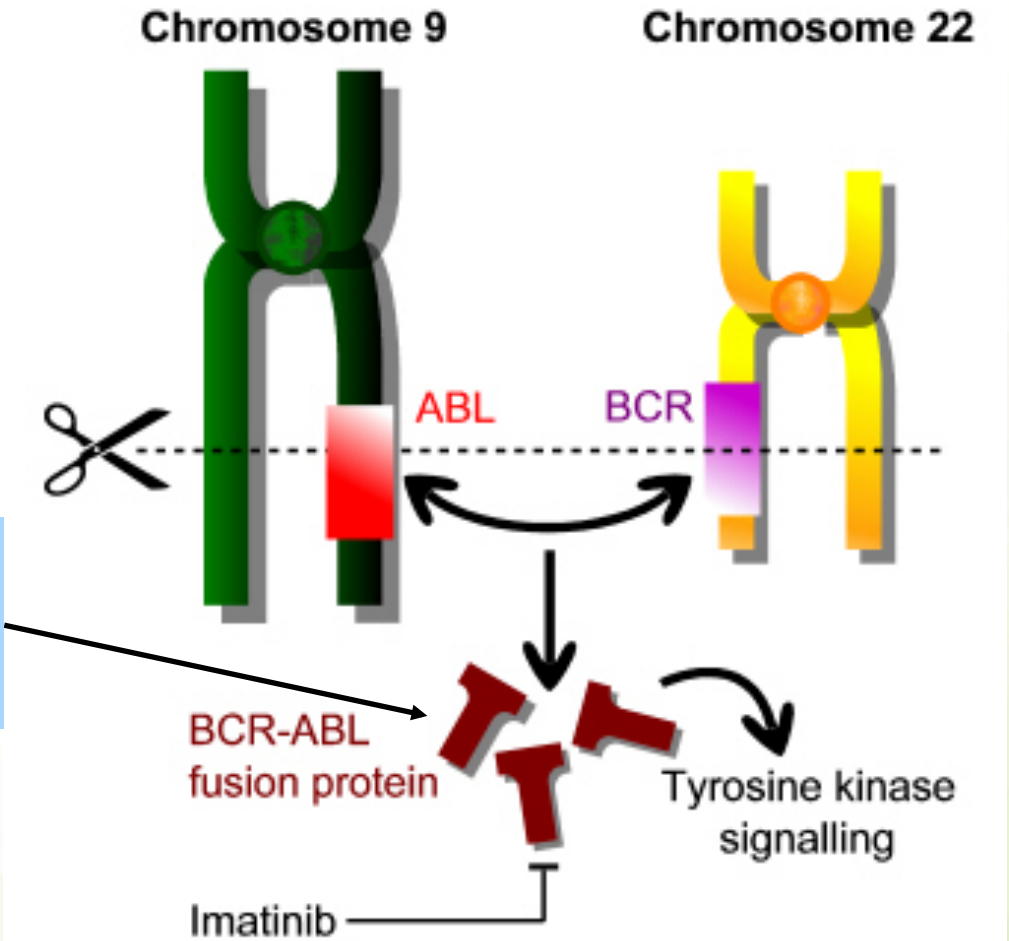




Karolinska Institutet
KTH
Stockholms universitet
UPPSALA UNIVERSITET
SciLifeLab

Chronic Myeloid Leukemia

- BCR-ABL1 fusion protein – a CML drug target

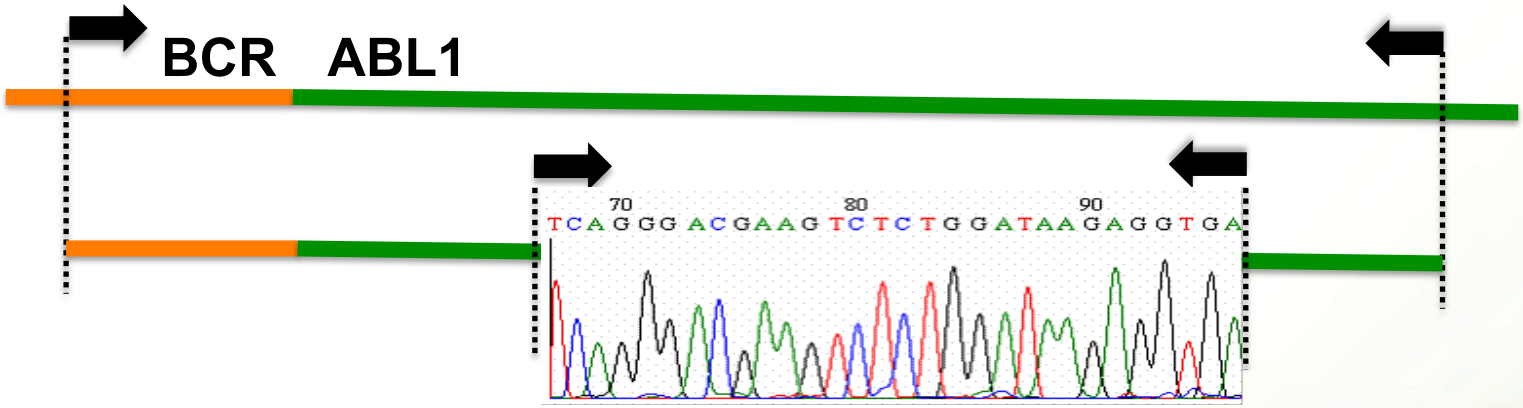


The BCR-ABL1 fusion protein can acquire resistance mutations following drug treatment

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

BMC Cancer

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 Häggqvist[1], Ida Höijer[1], Nicola Cahill[1], Ulla Olsson-Strömberg[2] and Monica Hermanson[1]

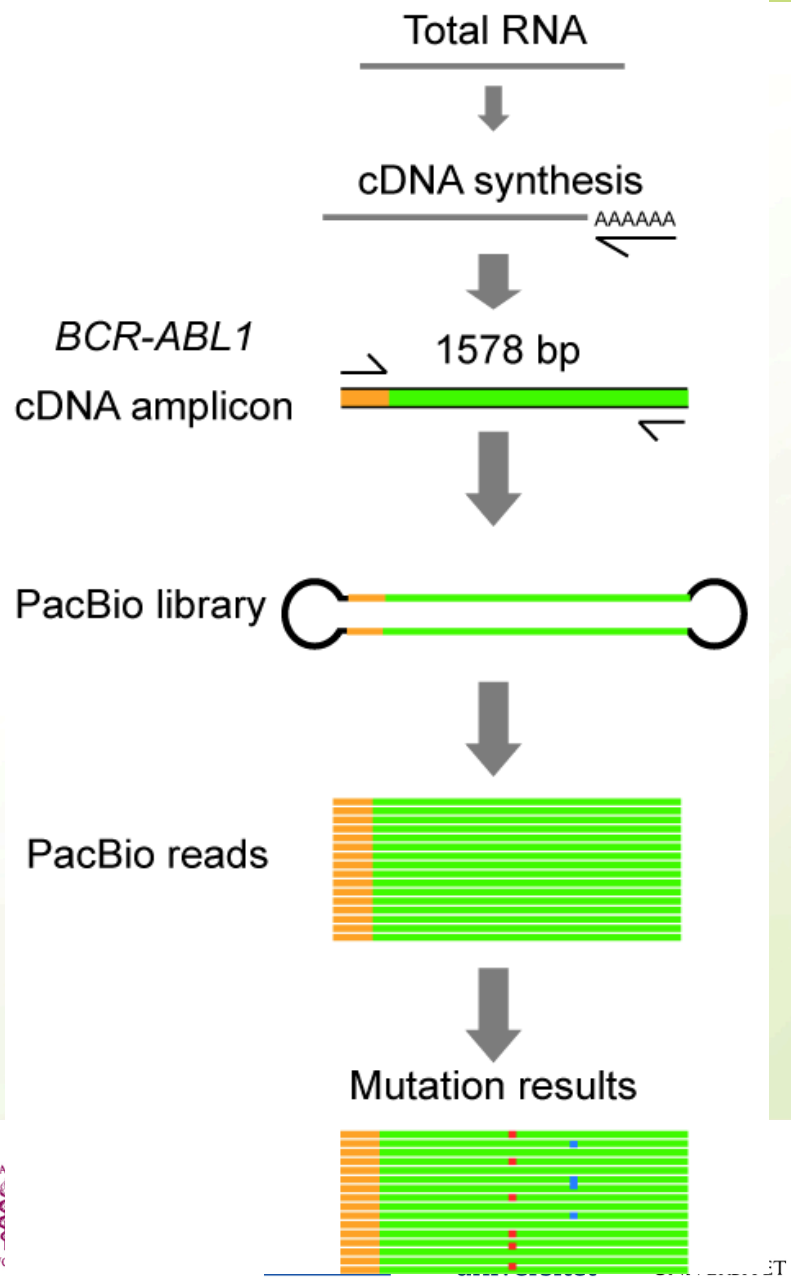
Abstract

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.

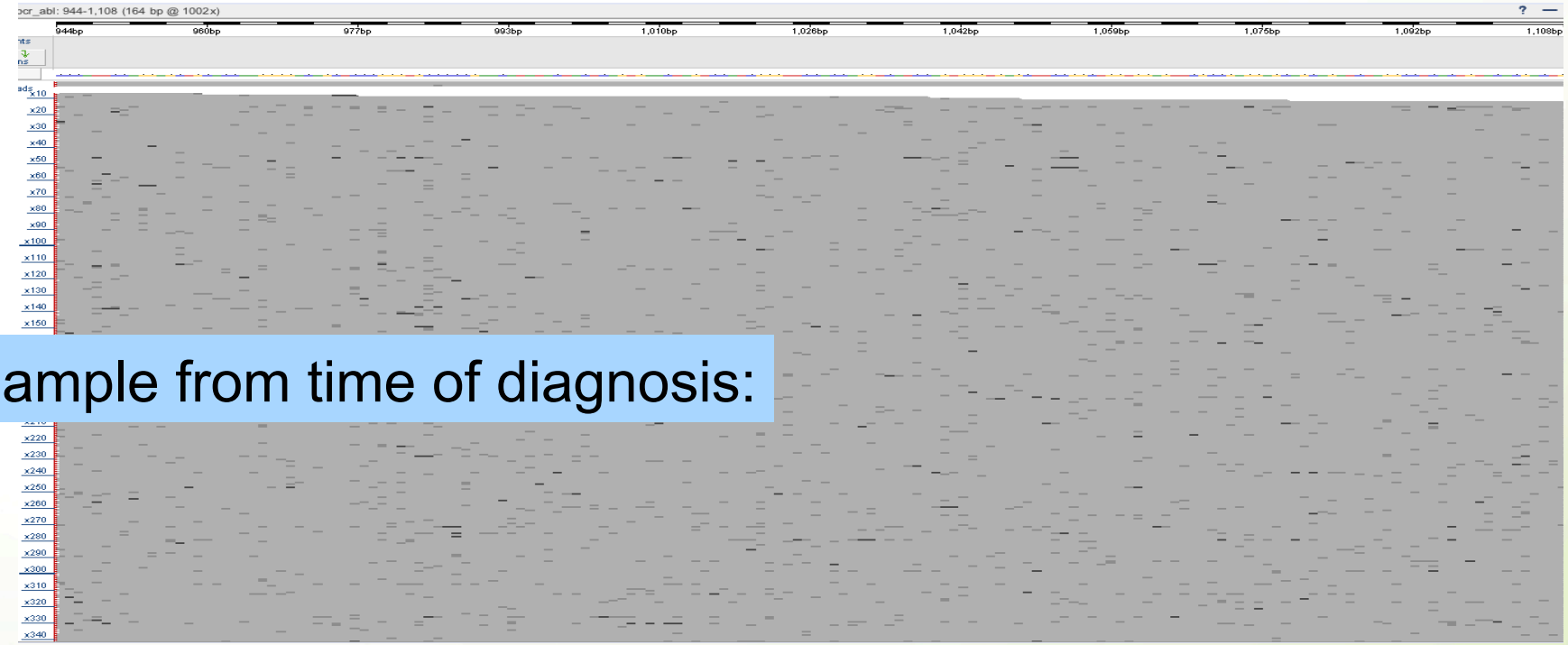


SciLifeLab Uppsala Genome Center SNP&SEQ Technology Platform NATIONAL GENOMICS INFRASTRUCTURE

BCR-ABL1 mutations at diagnosis

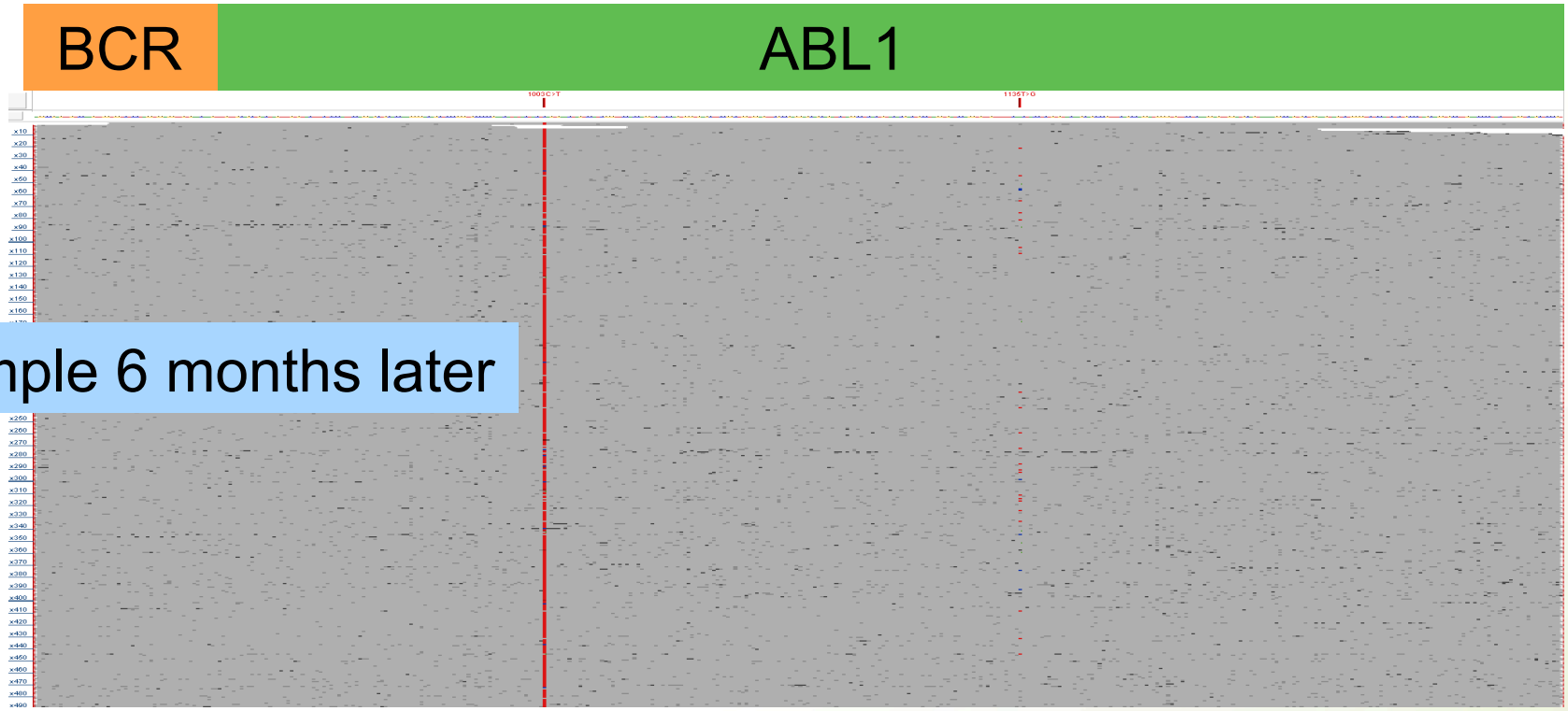
PacBio sequencing generates ~10 000X coverage!

BCR | ABL1



Sample from time of diagnosis:

BCR-ABL1 mutations in follow-up sample

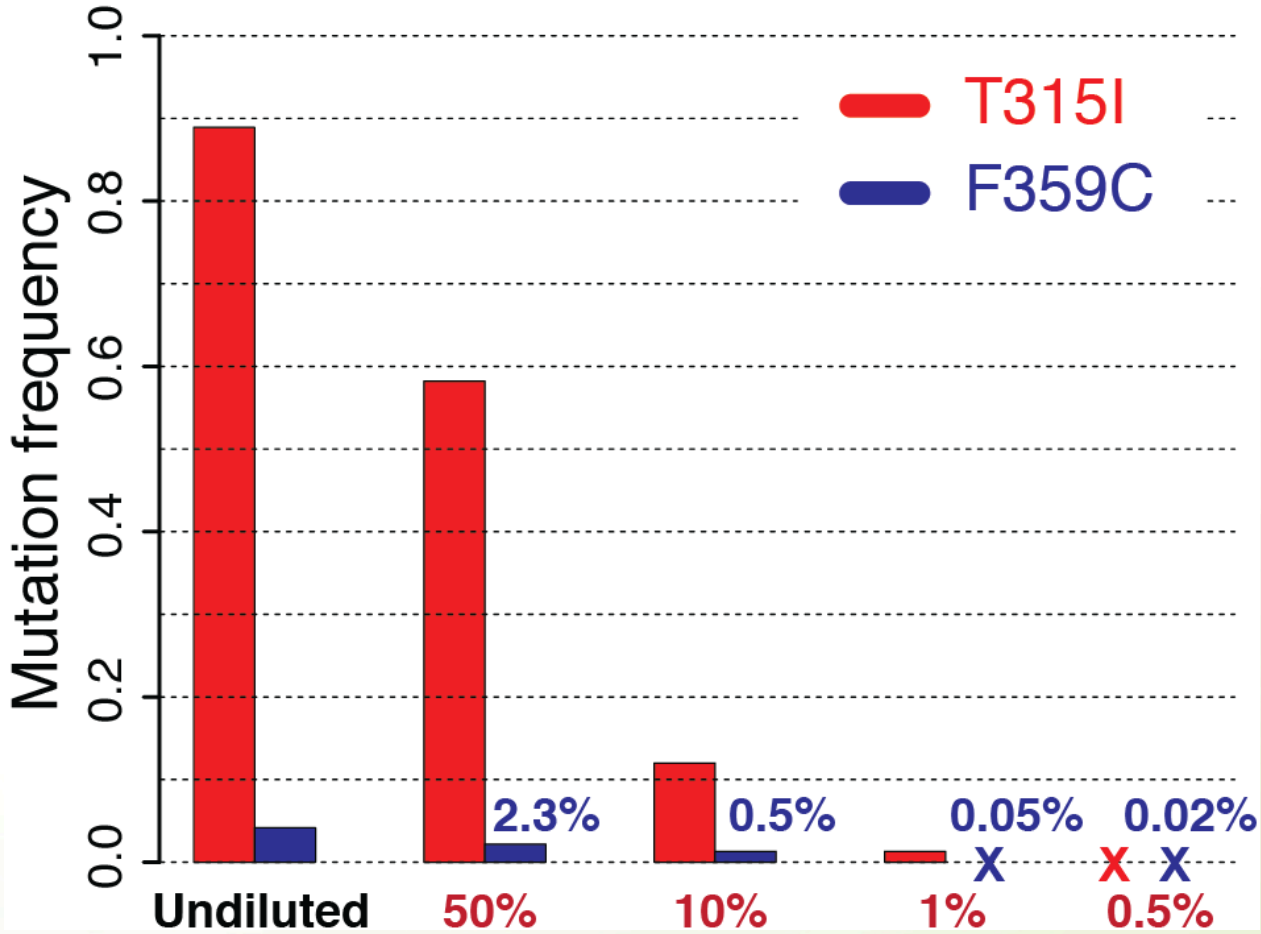


Sample 6 months later

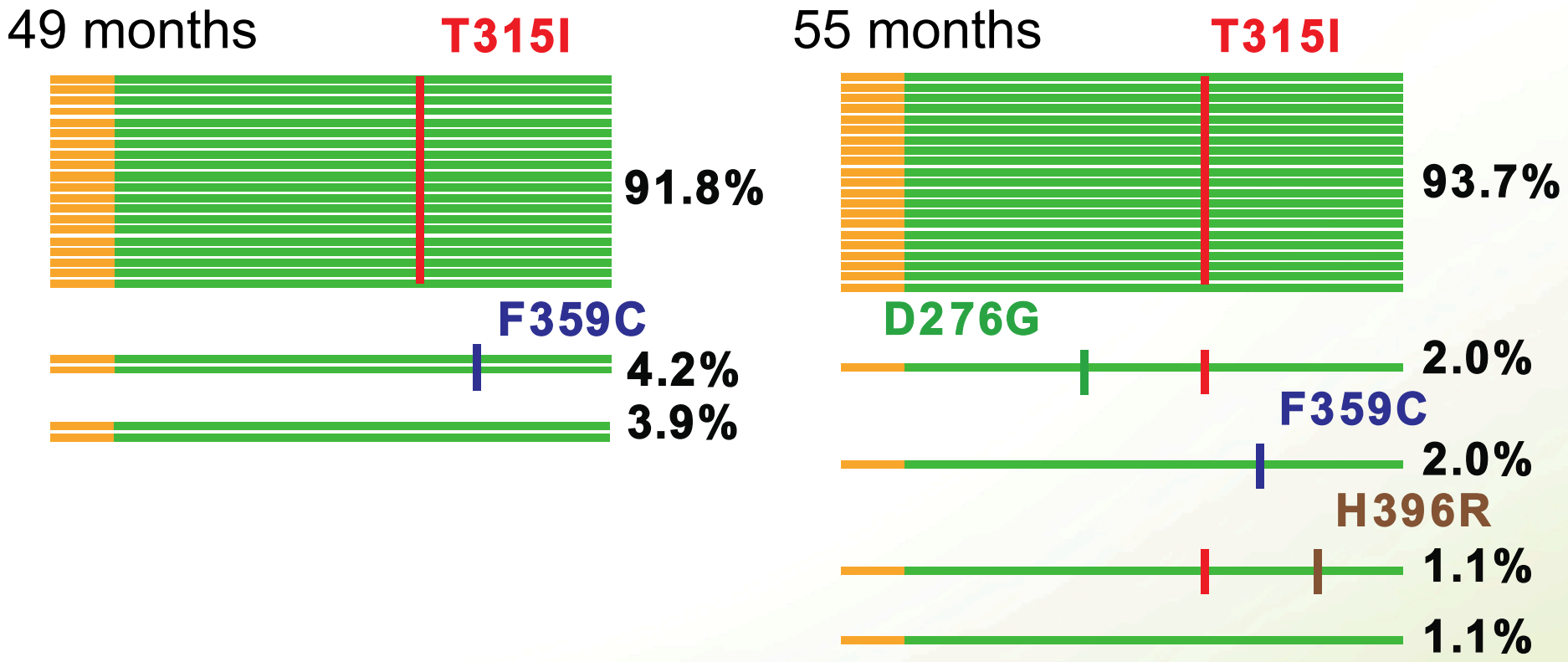
Mutations acquired in fusion transcript.
Might require treatment with alternative drug.

BCR-ABL1 dilution series results

- Mutations down to 1% detected!



BCR-ABL1 - Compound mutations



SciLifeLab Uppsala Genome Center SNP&SEQ Technology Platform National Genomics Infrastructure Karolinska Institutet KTH Stockholms universitet Uppsala Universitet

Analysis method for BCR-ABL1 mutations

- Create CCS reads and screen all known resistance mutations

- CAVA analysis - count number of WT and MUT sequences

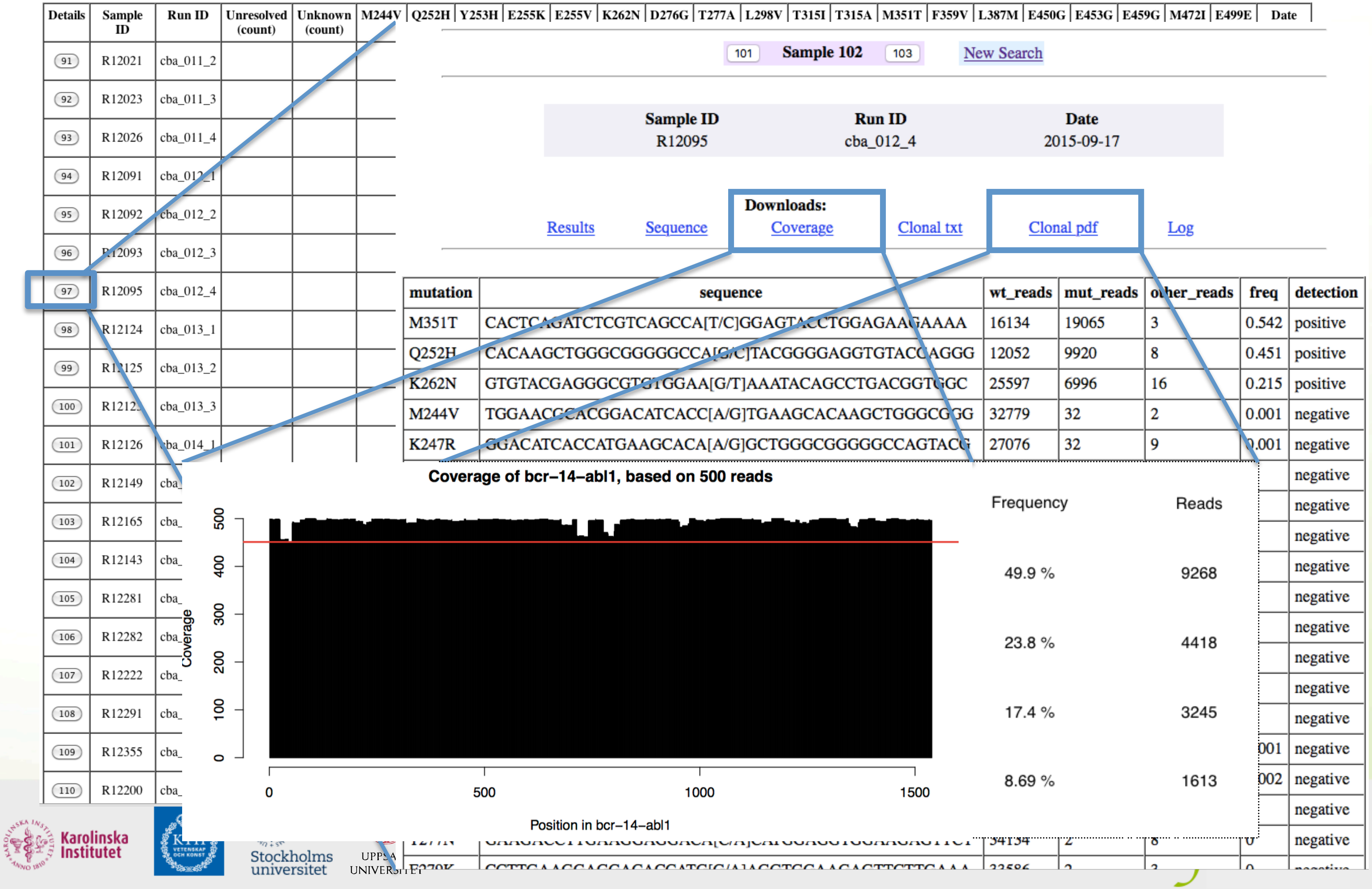
WT sequence: **TATATCATCACTGAGTTCATG**

MUT sequence: **TATATCATCATTGAGTTCATG**

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
91	R12021	cba_011_2																						
92	R12023	cba_011_3																						
93	R12026	cba_011_4																						
94	R12091	cba_012_1																						
95	R12092	cba_012_2																						
96	R12093	cba_012_3																						
97	R12095	cba_012_4																						
98	R12124	cba_013_1																						
99	R12125	cba_013_2																						
100	R1212	cba_013_3																						
101	R12126	cba_014_1																						
102	R12149	cba_																						
103	R12165	cba_																						
104	R12143	cba_																						
105	R12281	cba_																						
106	R12282	cba_																						
107	R12222	cba_																						
108	R12291	cba_																						
109	R12355	cba_																						
110	R12200	cba_																						

Sample ID	Run ID	Date
R12095	cba_012_4	2015-09-17

Downloads:

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]TACGGGGAGGTGTACGAGGG	12052	9920	8	0.451	positive
K262N	GTGTACGAGGGCGTGTGGAA[G/T]AAATACAGCCTGACGGTGGC	25597	6996	16	0.215	positive
M244V	TGGAACGGCACGGACATCACC[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
49.9 %	9268
23.8 %	4418
17.4 %	3245
8.69 %	1613

Karolinska Institutet

Stockholms universitet

UPPSALA UNIVERSITET

Clinical Diagnosis of BCR-ABL1 mutations

Clinical Genetics

- Collection of samples
- Seq library preparation

Sequencing Facility

- SMRT sequencing
- CAVA analysis

IT developers

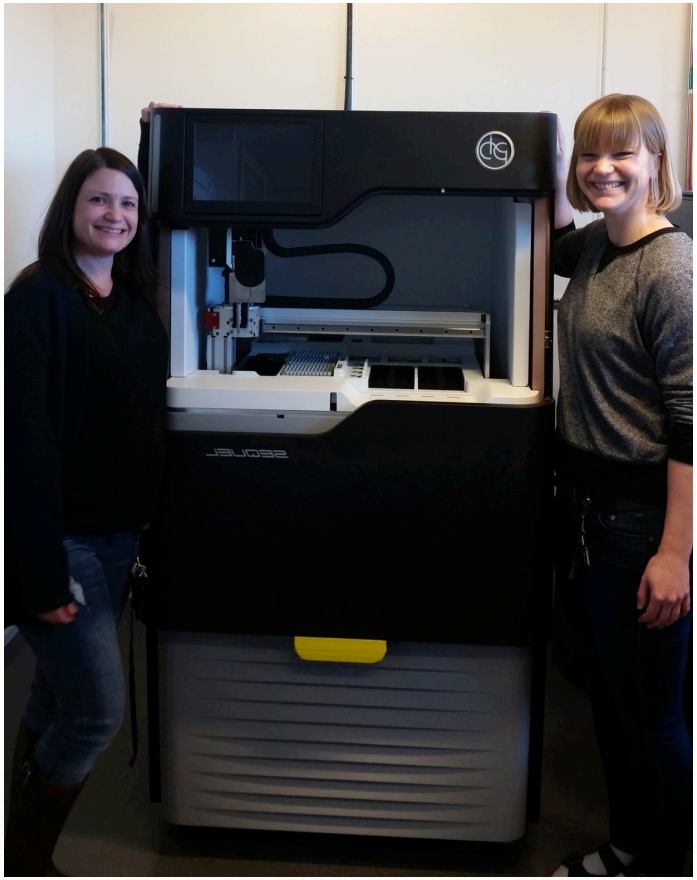
- Web server for results

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

SciLifeLab · Uppsala Genome Center · SNP&SEQ Technology Platform · National Genomics Infrastructure · Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet

PacBio - Ongoing developments

Sequel - New instrument with higher throughput!



7x more data per SMRT cell!

Installation at NGI during 2016

Who does the sequencing?

Ulf Gyllensten
Platform director

Inger Jonasson
Facility manager

Olga Vinnere Pettersson
Project coordinator

Adam Ameur
Bioinformatician, NGS

Ignas Bunikis
Bioinformatician, NGS

Christian Tellgren-Roth
Bioinformatician, NGS

Susana Häggqvist
Research engineer
NGS

Ida Höijer
Research engineer
NGS

Cecilia Lindau
Research engineer
NGS

Maria Schenström
Research engineer
NGS

Magdalena Andersson
Research engineer
NGS

Ulrika Broström
Research engineer
NGS

Nina Williams
Research engineer
NGS

Carolina Ilbäck
Research engineer
NGS

Anna Petri
Research engineer
Sequencing Service

Anne-Christine Lindström
Research engineer
Sequencing Service

What we sequence at NGI /

SciLifeLab

THANK YOU