Next Generation Sequencing

and

Bioinformatics Analysis Pipelines

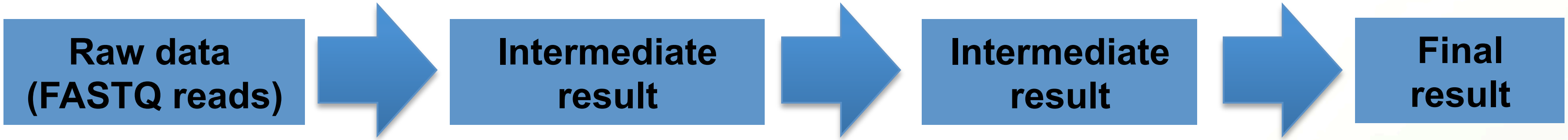
Adam Ameur
National Genomics Infrastructure
SciLifeLab Uppsala
adam.ameur@igp.uu.se

Today's lecture

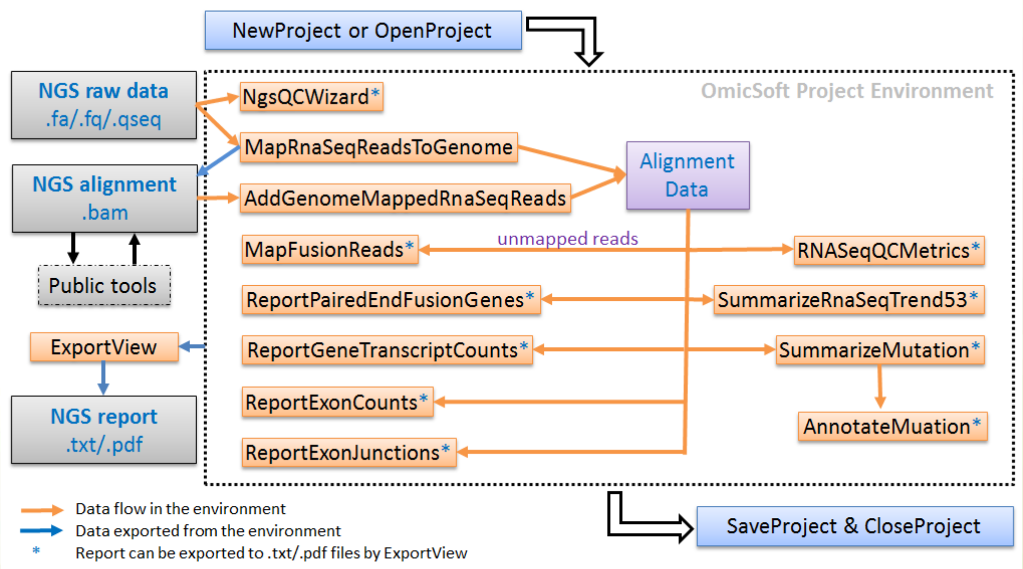
- Management of NGS data at NGI/SciLifeLab
- Examples of analysis pipelines:
  - Human exome/genome sequencing
  - Assembly using long reads
  - Clinical routine sequencing

What is an analysis pipeline?

- Basically just a number of steps to analyze data



- Pipelines can be simple or very complex…

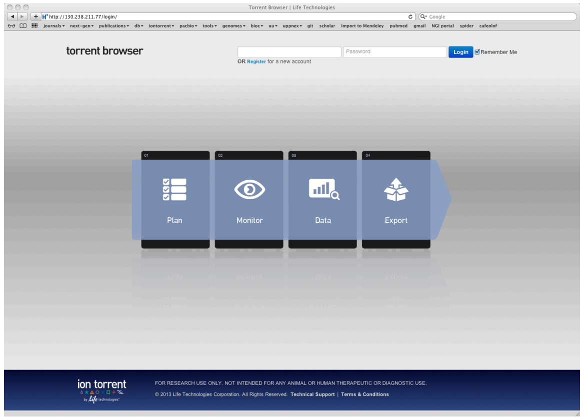


Data flow in the environment
Data exported from the environment
* Report can be exported to .txt/.pdf files by ExportView

Some analysis pipelines for NGS data

Ion Torrent

Torrent Suite Software

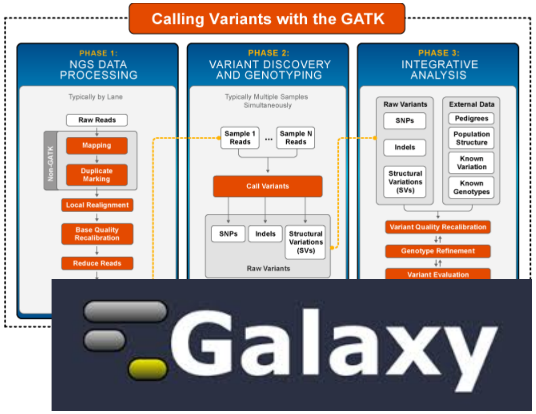


FOR RESEARCH USE ONLY. NOT INTENDED FOR ANY ANIMAL OR HUMAN THERAPEUTIC OR DIAGNOSTIC USE.

© 2013 Life Technologies Corporation. All Rights Reserved. Technical Support | Terms & Conditions

Illumina

GATK, Galaxy,…



Calling Variants with the GATK

PHASE 1: NGS DATA PROCESSING

Typically by Lane

Raw Reads

Non-GATK

Mapping

Duplicate Marking

Local Realignment

Base Quality Recalibration

Reduce Reads

PHASE 2: VARIANT DISCOVERY AND GENOTYPING

Typically Multiple Samples Simultaneously

Sample 1 Reads … Sample N Reads

Call Variants

SNPs

Indels

Structural Variations (SVs)

Raw Variants

PHASE 3: INTEGRATIVE ANALYSIS

Raw Variants

SNPs

Indels

Structural Variations (SVs)

External Data

Pedigrees

Population Structure

Known Variation

Known Genotypes

Variant Quality Recalibration

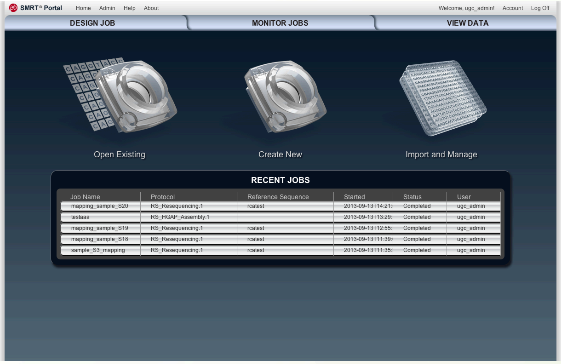
Genotype Refinement

Variant Evaluation

Galaxy

PacBio

SMRT analysis portal



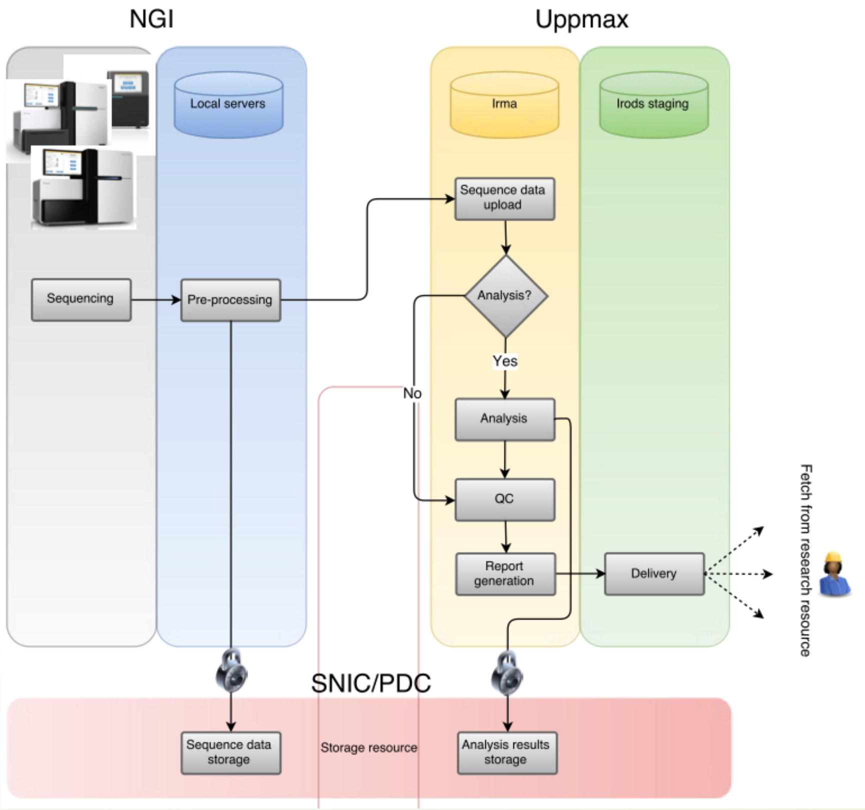
RECENT JOBS

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

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

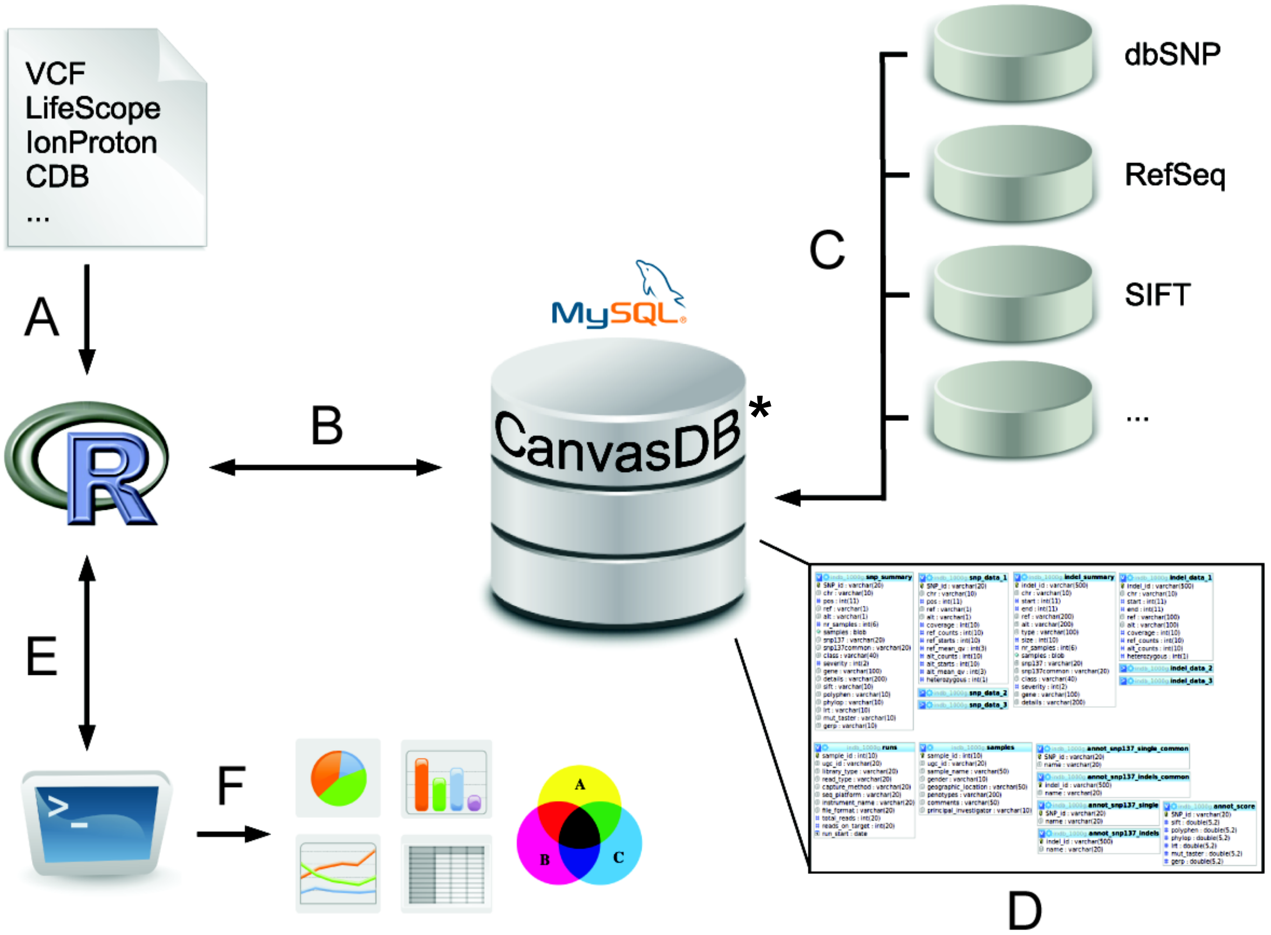
Some examples follows:

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

II. Assembly of genomes using long reads

III. Clinical sequencing – Leukemia Diagnostics

Example I:
Computational infrastructure for exome-seq data



```
indb_1000g.snp_summary
SNP_id : varchar(20)
chr : varchar(10)
pos : int(11)
ref : varchar(1)
alt : varchar(1)
nr_samples : int(6)
samples : blob
snp137 : varchar(20)
snp137common : varchar(20)
class : varchar(40)
severity : int(2)
gene : varchar(100)
details : varchar(200)
sift : varchar(10)
polyphen : varchar(10)
phylop : varchar(10)
lrt : varchar(10)
mut_taster : varchar(10)
gerp : varchar(10)

indb_1000g.snp_data_1
SNP_id : varchar(20)
chr : varchar(10)
pos : int(11)
ref : varchar(1)
alt : varchar(1)
coverage : int(10)
ref_counts : int(10)
ref_starts : int(10)
ref_mean_qv : int(3)
alt_counts : int(10)
alt_starts : int(10)
alt_mean_qv : int(3)
heterozygous : int(1)

indb_1000g.snp_data_2
indb_1000g.snp_data_3

indb_1000g.indel_summary
indel_id : varchar(500)
chr : varchar(10)
start : int(11)
end : int(11)
ref : varchar(200)
alt : varchar(200)
type : varchar(100)
size : int(10)
nr_samples : int(6)
samples : blob
snp137 : varchar(20)
snp137common : varchar(20)
class : varchar(40)
severity : int(2)
gene : varchar(100)
details : varchar(200)

indb_1000g.indel_data_1
indel_id : varchar(500)
chr : varchar(10)
start : int(11)
end : int(11)
ref : varchar(100)
alt : varchar(100)
coverage : int(10)
ref_counts : int(10)
alt_counts : int(10)
heterozygous : int(1)

indb_1000g.indel_data_2
indb_1000g.indel_data_3

indb_1000g.runs
sample_id : int(10)
ugc_id : varchar(20)
library_type : varchar(20)
read_type : varchar(20)
capture_method : varchar(20)
seq_platform : varchar(20)
instrument_name : varchar(20)
file_format : varchar(20)
total_reads : int(20)
reads_on_target : int(20)
run_start : date

indb_1000g.samples
sample_id : int(10)
ugc_id : varchar(20)
sample_name : varchar(50)
gender : varchar(10)
geographic_location : varchar(50)
penotypes : varchar(200)
comments : varchar(50)
principal_investigator : varchar(10)

indb_1000g.annot_snp137_single_common
SNP_id : varchar(20)
name : varchar(20)

indb_1000g.annot_snp137_indels_common
indel_id : varchar(500)
name : varchar(20)

indb_1000g.annot_snp137_single
SNP_id : varchar(20)
name : varchar(20)

indb_1000g.annot_snp137_indels
indel_id : varchar(500)
name : varchar(20)

indb_1000g.annot_score
SNP_id : varchar(20)
sift : double(5,2)
polyphen : double(5,2)
phylop : double(5,2)
lrt : double(5,2)
mut_taster : double(5,2)
gerp : double(5,2)
```

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

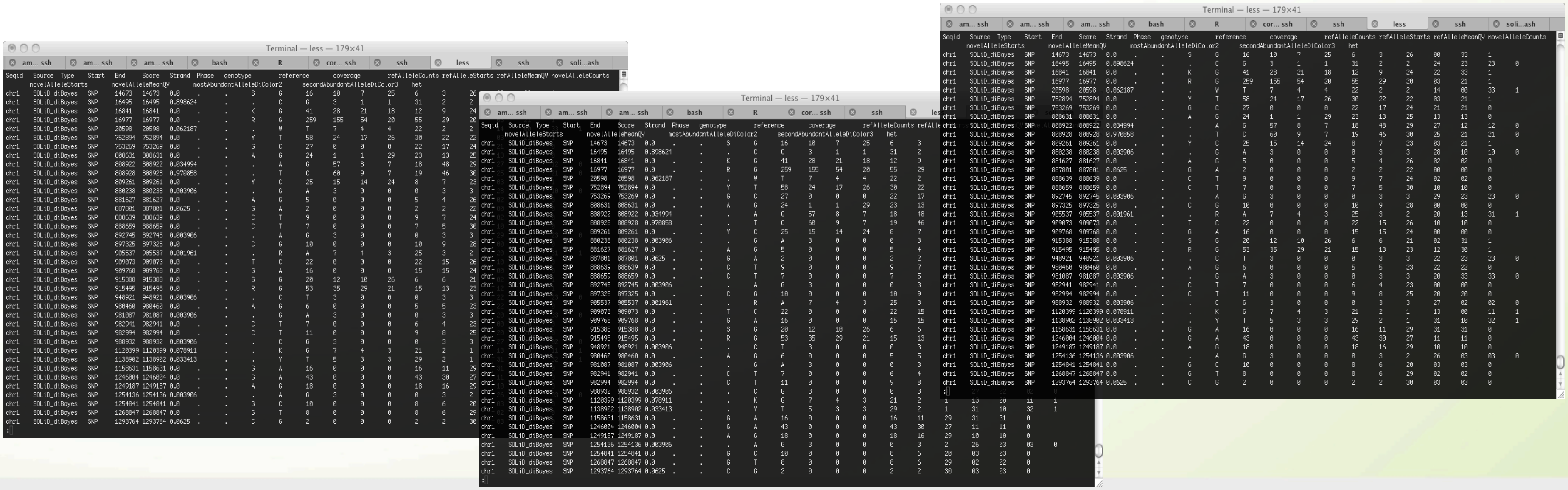
- Disadvantage
  - All genetic variants outside of exons are missed (~98%)

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

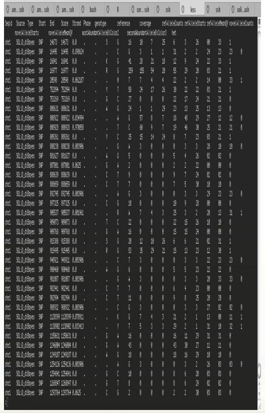


```
Seqid   Source  Type    Start   End     Score   Strand  Phase   genotype        reference       coverage        refAlleleCounts refAlleleStarts refAlleleMeanQV novelAlleleCounts
        novelAlleleStarts       novelAlleleMeanQV       mostAbundantAlleleDiColor2      secondAbundantAlleleDiColor3    het
chr1    SOLiD_diBayes   SNP     14673   14673   0.0     .       .       S       G       16      10      7       25      6       3       26      00      33      1
chr1    SOLiD_diBayes   SNP     16495   16495   0.898624        .       .       C       G       3       1       1       31      2       2       24      23      23      0
chr1    SOLiD_diBayes   SNP     16841   16841   0.0     .       .       K       G       41      28      21      18      12      9       24      22      33      1
chr1    SOLiD_diBayes   SNP     16977   16977   0.0     .       .       R       G       259     155     54      20      55      29      20      03      21      1
chr1    SOLiD_diBayes   SNP     20598   20598   0.062187        .       .       W       T       7       4       4       22      2       2       14      00      33      1
chr1    SOLiD_diBayes   SNP     752894  752894  0.0     .       .       Y       T       58      24      17      26      30      22      22      03      21      1
chr1    SOLiD_diBayes   SNP     753269  753269  0.0     .       .       G       C       27      0       0       0       22      17      24      21      21      0
chr1    SOLiD_diBayes   SNP     808631  808631  0.0     .       .       A       G       24      1       1       29      23      13      25      13      13      0
chr1    SOLiD_diBayes   SNP     808922  808922  0.034994        .       .       A       G       57      8       7       18      48      29      27      12      12      0
chr1    SOLiD_diBayes   SNP     808928  808928  0.970858        .       .       T       C       60      9       7       19      46      30      25      21      21      0
chr1    SOLiD_diBayes   SNP     809261  809261  0.0     .       .       Y       C       25      15      14      24      8       7       23      03      21      1
chr1    SOLiD_diBayes   SNP     880238  880238  0.003906        .       .       G       A       3       0       0       0       3       3       28      10      10      0
chr1    SOLiD_diBayes   SNP     881627  881627  0.0     .       .       A       G       5       0       0       0       5       4       26      02      02      0
chr1    SOLiD_diBayes   SNP     887801  887801  0.0625  .       .       G       A       2       0       0       0       2       2       22      00      00      0
chr1    SOLiD_diBayes   SNP     888639  888639  0.0     .       .       C       T       9       0       0       0       9       7       24      02      02      0
chr1    SOLiD_diBayes   SNP     888659  888659  0.0     .       .       C       T       7       0       0       0       7       5       30      10      10      0
chr1    SOLiD_diBayes   SNP     892745  892745  0.003906        .       .       A       G       3       0       0       0       3       3       29      23      23      0
chr1    SOLiD_diBayes   SNP     897325  897325  0.0     .       .       C       G       10      0       0       0       10      9       28      00      00      0
chr1    SOLiD_diBayes   SNP     905537  905537  0.001961        .       .       R       A       7       4       3       25      3       2       20      13      31      1
chr1    SOLiD_diBayes   SNP     909073  909073  0.0     .       .       T       C       22      0       0       0       22      15      26      10      10      0
chr1    SOLiD_diBayes   SNP     909768  909768  0.0     .       .       G       A       16      0       0       0       15      15      24      00      00      0
chr1    SOLiD_diBayes   SNP     915388  915388  0.0     .       .       S       G       20      12      10      26      6       6       21      02      31      1
chr1    SOLiD_diBayes   SNP     915495  915495  0.0     .       .       R       G       53      35      29      21      15      13      23      12      30      1
chr1    SOLiD_diBayes   SNP     948921  948921  0.003906        .       .       C       T       3       0       0       0       3       3       22      23      23      0
chr1    SOLiD_diBayes   SNP     980460  980460  0.0     .       .       A       G       6       0       0       0       5       5       23      22      22      0
chr1    SOLiD_diBayes   SNP     981087  981087  0.003906        .       .       G       A       3       0       0       0       3       3       20      33      33      0
chr1    SOLiD_diBayes   SNP     982941  982941  0.0     .       .       C       T       7       0       0       0       6       4       23      00      00      0
chr1    SOLiD_diBayes   SNP     982994  982994  0.0     .       .       C       T       11      0       0       0       9       8       25      20      20      0
chr1    SOLiD_diBayes   SNP     988932  988932  0.003906        .       .       C       G       3       0       0       0       3       3       27      02      02      0
chr1    SOLiD_diBayes   SNP     1120399 1120399 0.078911        .       .       K       G       7       4       3       21      2       1       13      00      11      1
chr1    SOLiD_diBayes   SNP     1138902 1138902 0.033413        .       .       Y       T       5       3       3       29      2       1       31      10      32      1
chr1    SOLiD_diBayes   SNP     1158631 1158631 0.0     .       .       G       A       16      0       0       0       16      11      29      31      31      0
chr1    SOLiD_diBayes   SNP     1246004 1246004 0.0     .       .       G       A       43      0       0       0       43      30      27      11      11      0
chr1    SOLiD_diBayes   SNP     1249187 1249187 0.0     .       .       A       G       18      0       0       0       18      16      29      10      10      0
chr1    SOLiD_diBayes   SNP     1254136 1254136 0.003906        .       .       A       G       3       0       0       0       3       2       26      03      03      0
chr1    SOLiD_diBayes   SNP     1254841 1254841 0.0     .       .       G       C       10      0       0       0       8       6       20      03      03      0
chr1    SOLiD_diBayes   SNP     1268847 1268847 0.0     .       .       G       T       8       0       0       0       8       6       29      02      02      0
chr1    SOLiD_diBayes   SNP     1293764 1293764 0.0625  .       .       C       G       2       0       0       0       2       2       30      03      03      0
:
```

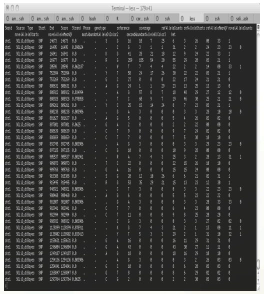
How to analyze this?

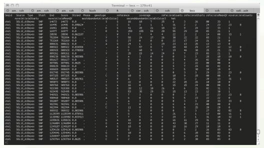
- Traditional analysis - A lot of filtering!
  - Typical filters
    - Focus on rare SNPs (not present in dbSNP)
    - Remove FPs (by filtering against other exomes)
    - Effect on protein: non-synonymous, stop-gain etc
    - Heterozygous/homozygous
  - This analysis can be automated (more or less)

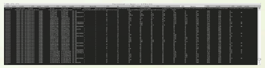
Start:
All identified SNPs



```
Seqid  Source  Type  Start  End  Score  Strand  Phase  genotype  reference  coverage  refAlleleCounts refAlleleStarts refAlleleMeanQV novelAlleleCounts
      novelAlleleStarts  novelAlleleMeanQV  mostAbundantAlleleDiColor2  secondAbundantAlleleDiColor3  het
chr1  SOLiD_diBayes  SNP  14673  14673  0.0  .  .  S  G  16  10  7  25  6  3  26  00  33  1
chr1  SOLiD_diBayes  SNP  16495  16495  0.898624  .  .  C  G  3  1  1  31  2  2  24  23  23  0
chr1  SOLiD_diBayes  SNP  16841  16841  0.0  .  .  K  G  41  28  21  18  12  9  24  22  33  1
chr1  SOLiD_diBayes  SNP  16977  16977  0.0  .  .  R  G  259  155  54  20  55  29  20  03  21  1
chr1  SOLiD_diBayes  SNP  28598  28598  0.062187  .  .  W  T  7  4  4  22  2  2  14  00  33  1
chr1  SOLiD_diBayes  SNP  752894  752894  0.0  .  .  Y  T  58  24  17  26  30  22  22  03  21  1
chr1  SOLiD_diBayes  SNP  753269  753269  0.0  .  .  G  C  27  0  0  0  22  17  24  21  21  0
chr1  SOLiD_diBayes  SNP  808631  808631  0.0  .  .  A  G  24  1  1  29  23  13  25  13  13  0
chr1  SOLiD_diBayes  SNP  808922  808922  0.034994  .  .  A  G  57  8  7  18  48  29  27  12  12  0
chr1  SOLiD_diBayes  SNP  808928  808928  0.970858  .  .  T  C  60  9  7  19  46  30  25  21  21  0
chr1  SOLiD_diBayes  SNP  809261  809261  0.0  .  .  Y  C  25  15  14  24  8  7  23  03  21  1
chr1  SOLiD_diBayes  SNP  880238  880238  0.003906  .  .  G  A  3  0  0  0  3  3  28  10  10  0
chr1  SOLiD_diBayes  SNP  881627  881627  0.0  .  .  A  G  5  0  0  0  5  4  26  02  02  0
chr1  SOLiD_diBayes  SNP  887801  887801  0.0625  .  .  G  A  2  0  0  0  2  2  22  00  00  0
chr1  SOLiD_diBayes  SNP  888639  888639  0.0  .  .  C  T  9  0  0  0  9  7  24  02  02  0
chr1  SOLiD_diBayes  SNP  888659  888659  0.0  .  .  C  T  7  0  0  0  7  5  30  10  10  0
chr1  SOLiD_diBayes  SNP  892745  892745  0.003906  .  .  A  G  3  0  0  0  3  3  29  23  23  0
chr1  SOLiD_diBayes  SNP  897325  897325  0.0  .  .  C  G  10  0  0  0  10  9  28  00  00  0
chr1  SOLiD_diBayes  SNP  905537  905537  0.001961  .  .  R  A  7  4  3  25  3  2  20  13  31  1
chr1  SOLiD_diBayes  SNP  909073  909073  0.0  .  .  T  C  22  0  0  0  22  15  26  10  10  0
chr1  SOLiD_diBayes  SNP  909768  909768  0.0  .  .  G  A  16  0  0  0  15  15  24  00  00  0
chr1  SOLiD_diBayes  SNP  915388  915388  0.0  .  .  S  G  20  12  10  26  6  6  21  02  31  1
chr1  SOLiD_diBayes  SNP  915495  915495  0.0  .  .  R  G  53  35  29  21  15  13  23  12  30  1
```







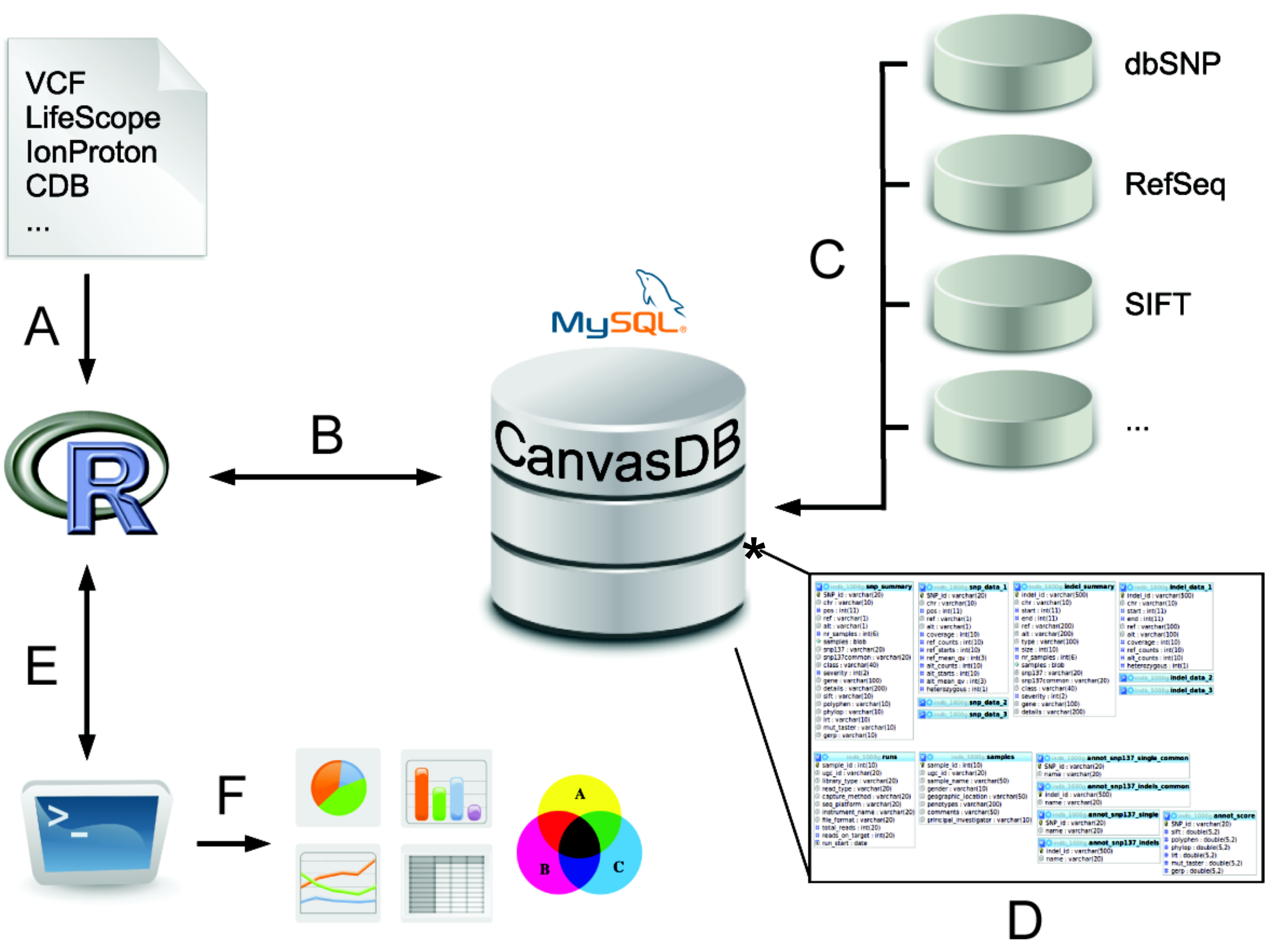
Result:
A few candidate causative SNP(s)!

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

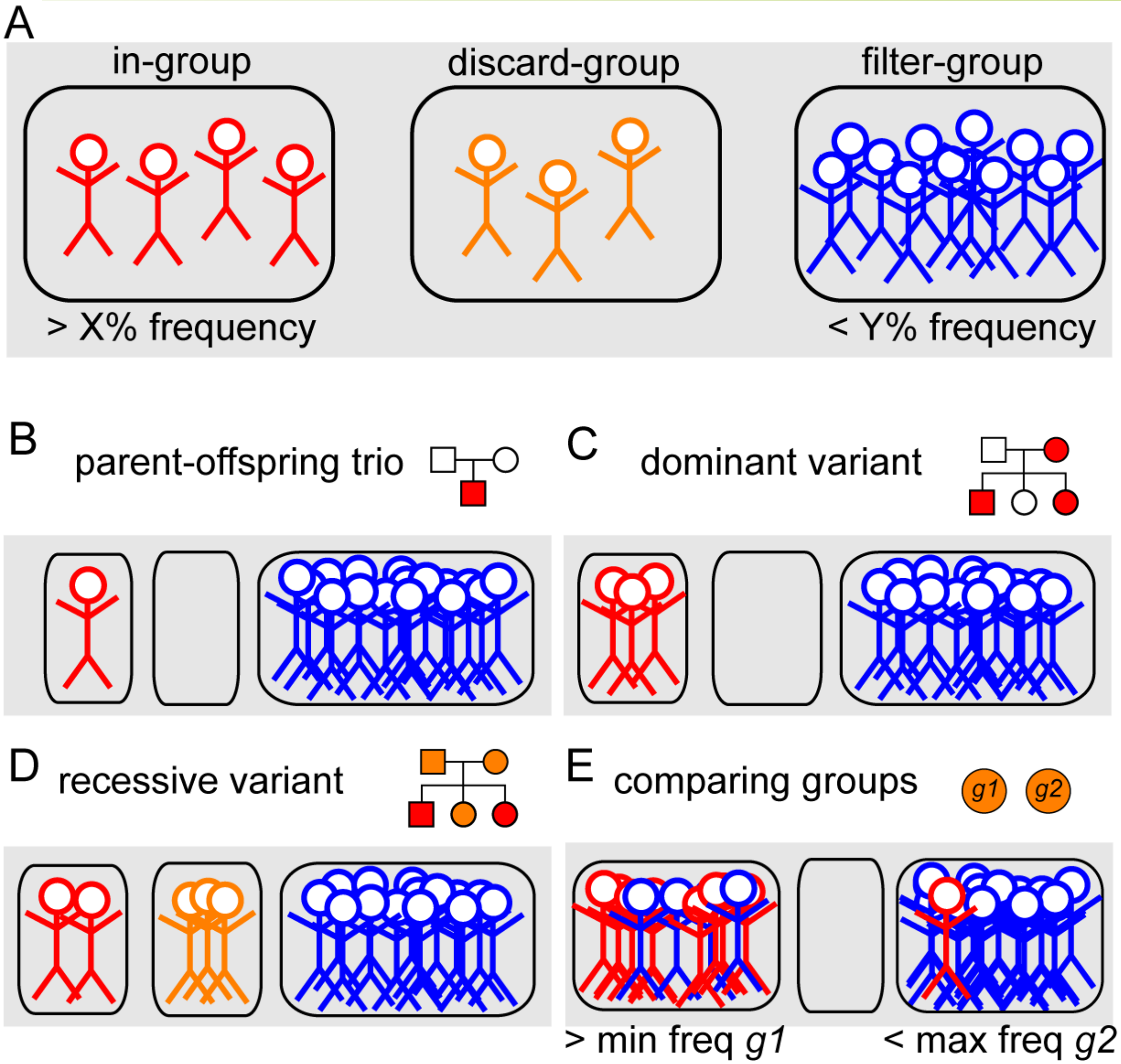


```
indb_1000g snp_summary
SNP_id : varchar(20)
chr : varchar(10)
pos : int(11)
ref : varchar(1)
alt : varchar(1)
nr_samples : int(6)
samples : blob
snp137 : varchar(20)
snp137common : varchar(20)
class : varchar(40)
severity : int(2)
gene : varchar(100)
details : varchar(200)
sift : varchar(10)
polyphen : varchar(10)
phylop : varchar(10)
lrt : varchar(10)
mut_taster : varchar(10)
gerp : varchar(10)

indb_1000g snp_data_1
SNP_id : varchar(20)
chr : varchar(10)
pos : int(11)
ref : varchar(1)
alt : varchar(1)
coverage : int(10)
ref_counts : int(10)
ref_starts : int(10)
ref_mean_qv : int(3)
alt_counts : int(10)
alt_starts : int(10)
alt_mean_qv : int(3)
heterozygous : int(1)

indb_1000g snp_data_2

indb_1000g snp_data_3

indb_1000g indel_summary
indel_id : varchar(500)
chr : varchar(10)
start : int(11)
end : int(11)
ref : varchar(200)
alt : varchar(200)
type : varchar(100)
size : int(10)
nr_samples : int(6)
samples : blob
snp137 : varchar(20)
snp137common : varchar(20)
class : varchar(40)
severity : int(2)
gene : varchar(100)
details : varchar(200)

indb_1000g indel_data_1
indel_id : varchar(500)
chr : varchar(10)
start : int(11)
end : int(11)
ref : varchar(100)
alt : varchar(100)
coverage : int(10)
ref_counts : int(10)
alt_counts : int(10)
heterozygous : int(1)

indb_1000g indel_data_2

indb_1000g indel_data_3

indb_1000g runs
sample_id : int(10)
ugc_id : varchar(20)
library_type : varchar(20)
read_type : varchar(20)
capture_method : varchar(20)
seq_platform : varchar(20)
instrument_name : varchar(20)
file_format : varchar(20)
total_reads : int(20)
reads_on_target : int(20)
run_start : date

indb_1000g samples
sample_id : int(10)
ugc_id : varchar(20)
sample_name : varchar(50)
gender : varchar(10)
geographic_location : varchar(50)
penotypes : varchar(200)
comments : varchar(50)
principal_investigator : varchar(10)

indb_1000g annot_snp137_single_common
SNP_id : varchar(20)
name : varchar(20)

indb_1000g annot_snp137_indels_common
indel_id : varchar(500)
name : varchar(20)

indb_1000g annot_snp137_single
SNP_id : varchar(20)
name : varchar(20)

indb_1000g annot_snp137_indels
indel_id : varchar(500)
name : varchar(20)

indb_1000g annot_score
SNP_id : varchar(20)
sift : double(5,2)
polyphen : double(5,2)
phylop : double(5,2)
lrt : double(5,2)
mut_taster : double(5,2)
gerp : double(5,2)
```

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

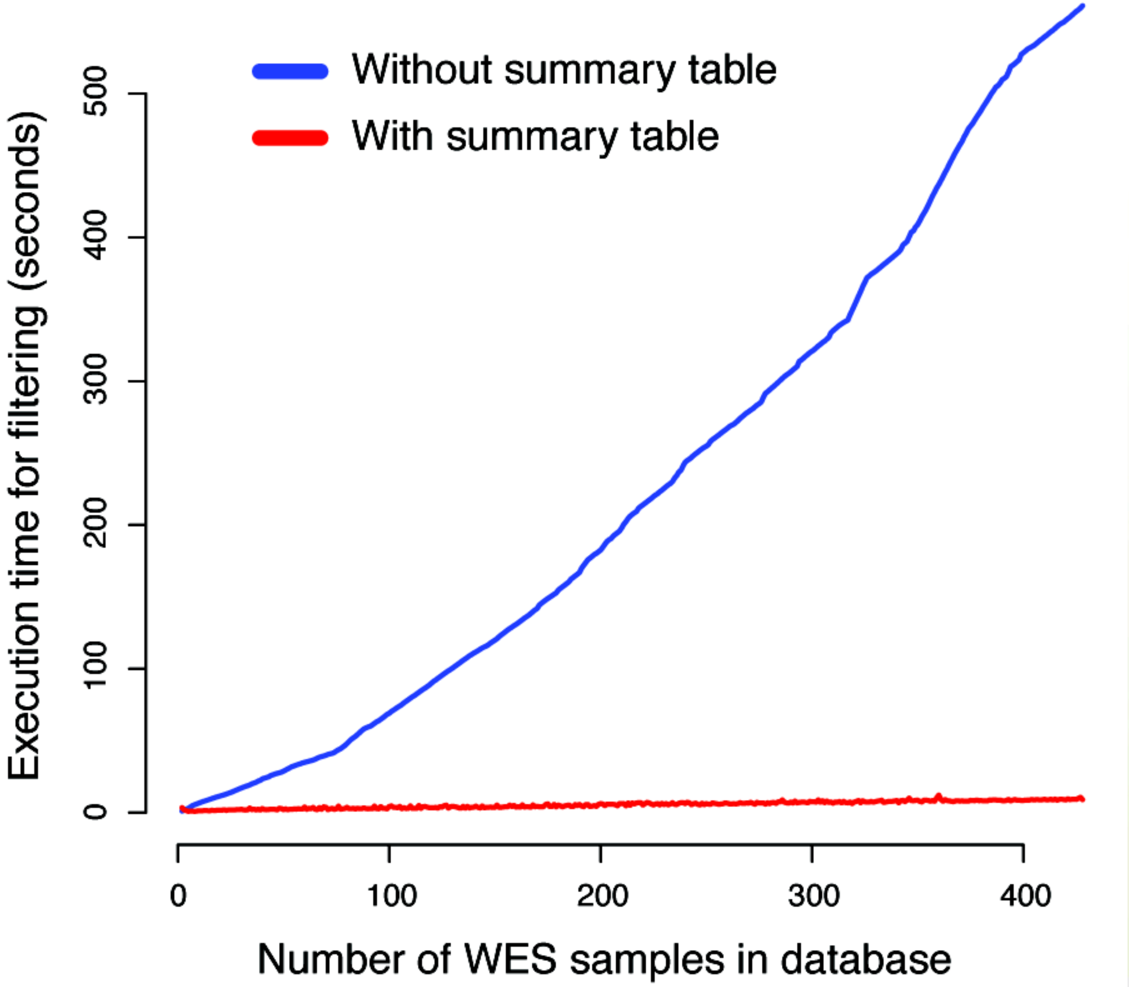
Ameur et al., Database Journal, 2014

CanvasDB - Filtering



CanvasDB - Filtering speed

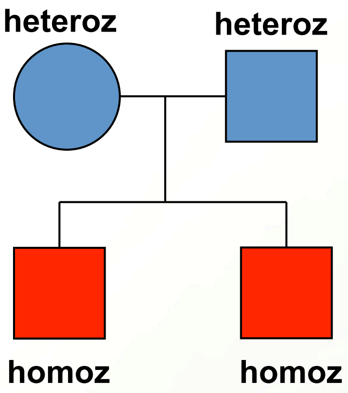
- Rapid variant filtering, also for large databases



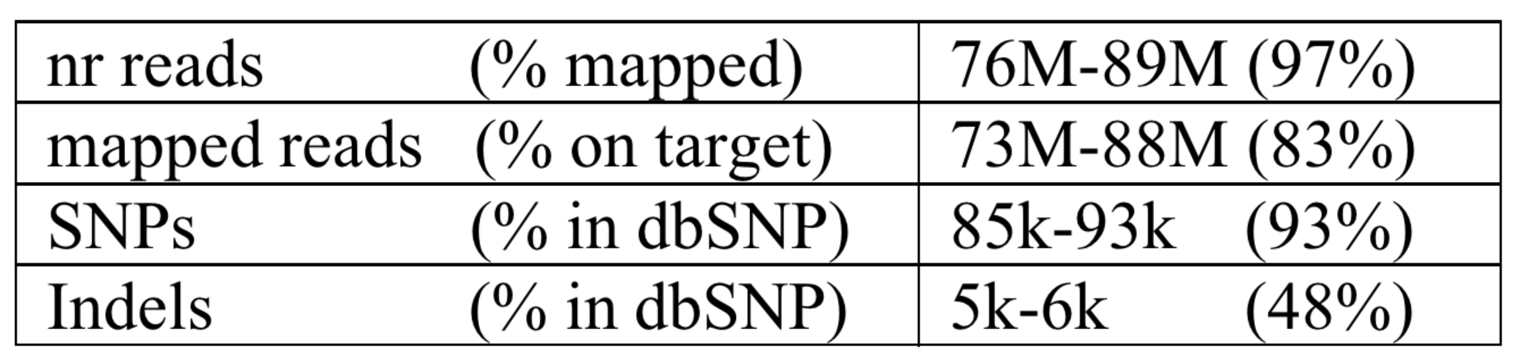
A recent exome-seq project

- Hearing loss: 2 affected brothers
  - Likely a rare, recessive disease

  => Shared homozygous SNPs/indels



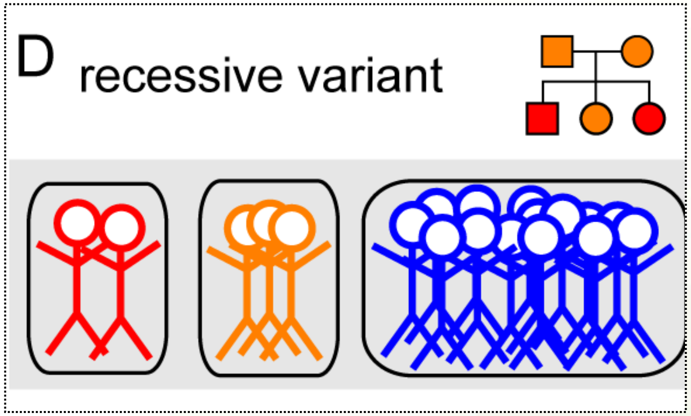
- Sequencing strategy
  - TargetSeq exome capture
  - One sample per PI chip

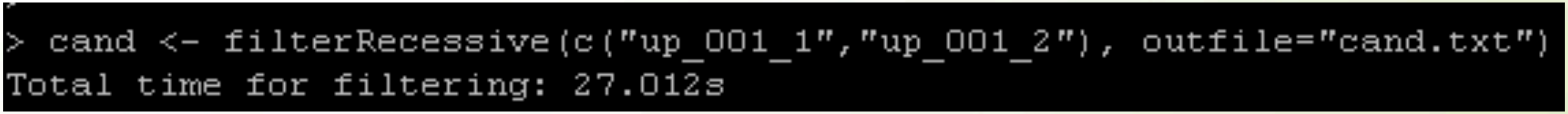


nr reads (% mapped)	76M-89M (97%)
mapped reads (% on target)	73M-88M (83%)
SNPs (% in dbSNP)	85k-93k (93%)
Indels (% in dbSNP)	5k-6k (48%)

Filtering analysis

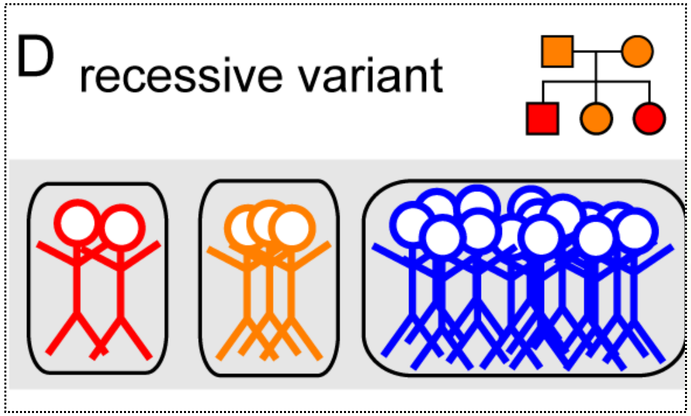
- *CanvasDB* filtering for a variant that is…
  - rare
    - at most in 1% of ~700 exomes
  - shared
    - found in both brothers
  - homozygous
    - in brothers, but in no other samples
  - deleterious
    - non-synonymous, frameshift, stop-gain, splicing, etc..



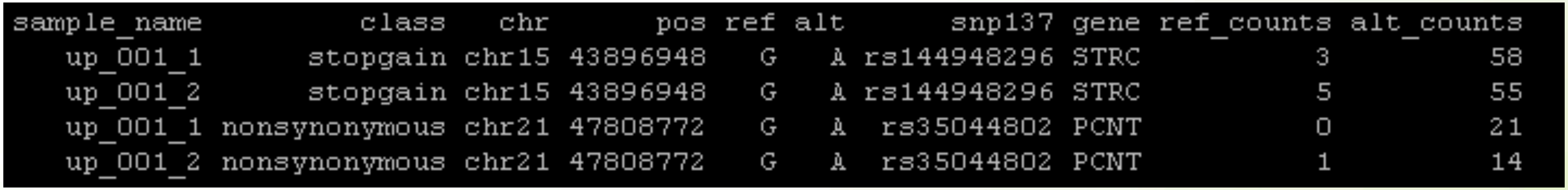


```
> cand <- filterRecessive(c("up_001_1","up_001_2"), outfile="cand.txt")
Total time for filtering: 27.012s
```

Filtering results



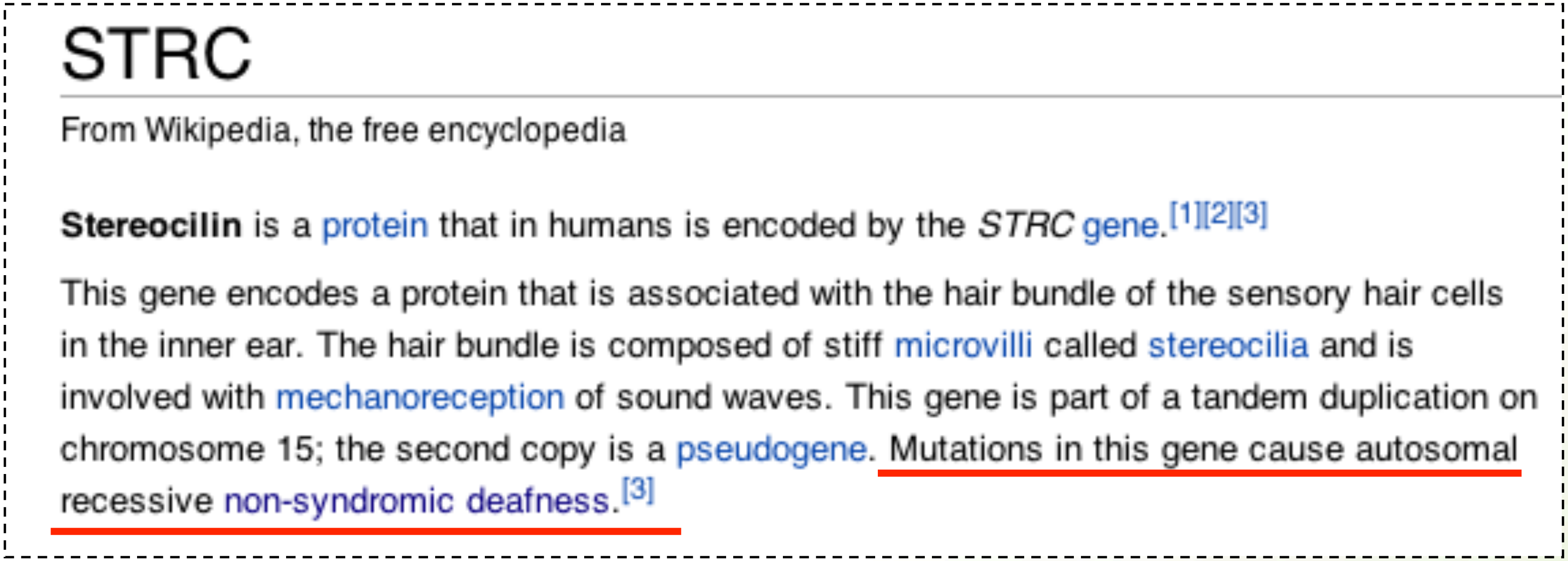
- Homozygous candidates
  - 2 SNPs
    - stop-gain in *STRC*
    - non-synonymous in *PCNT*
  - 0 indels
- Compound heterozygous candidates (lower priority)
  - in 15 genes



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

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

STRC - a candidate gene



STRC

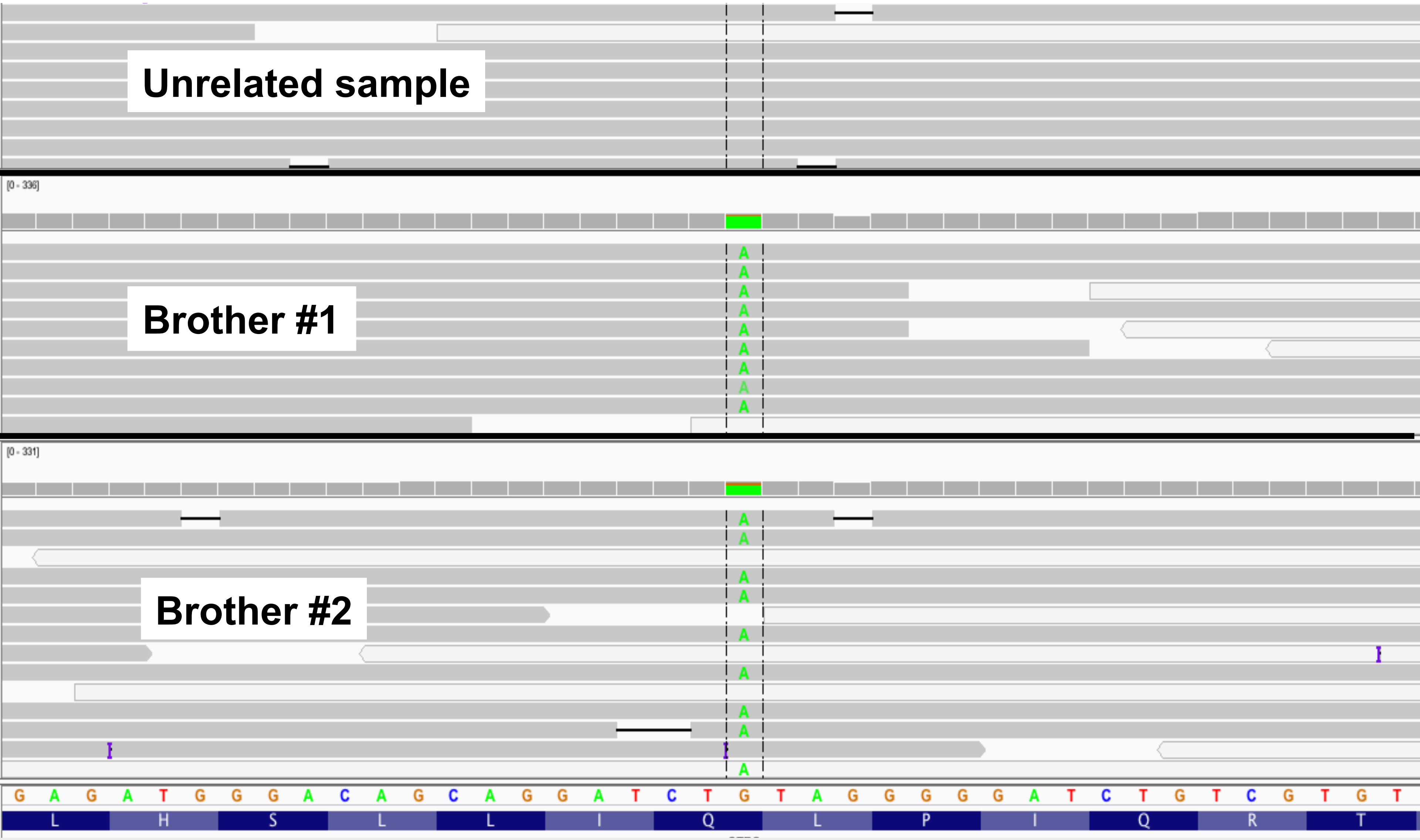
From Wikipedia, the free encyclopedia

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

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

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

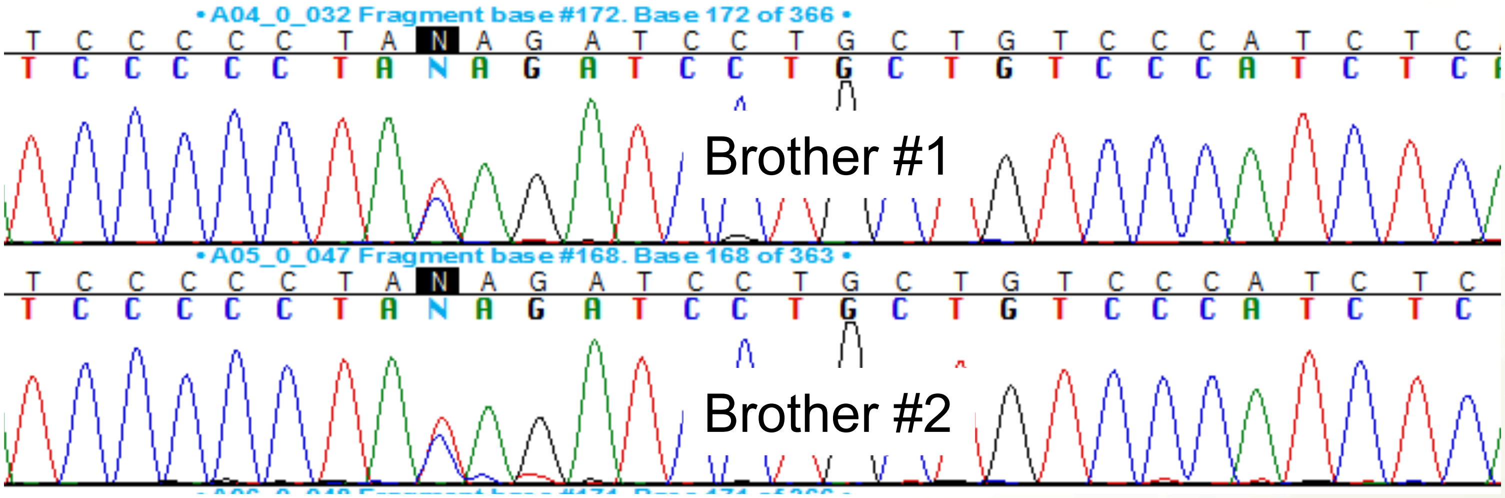
IGV visualization: Stop gain in STRC


Unrelated sample
[0 - 336]
Brother #1
A A A A A A A A A
[0 - 331]
Brother #2
A A A A A A A A A
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 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

Stop-gain site



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

CanvasDB – some success stories

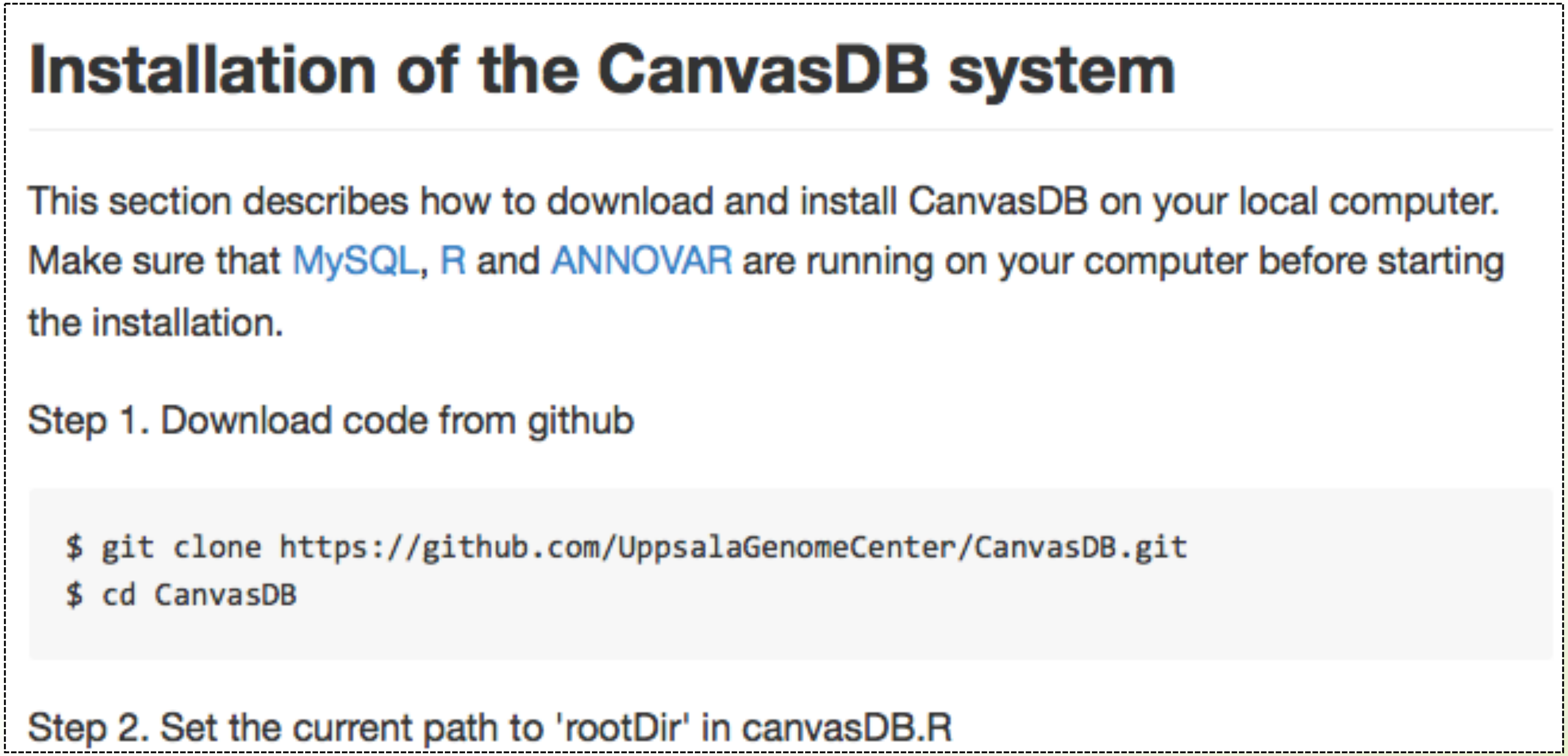
Solved cases, exome-seq - Niklas Dahl/Joakim Klar

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

Success rate >80% for recent Proton projects!

CanvasDB - Availability

- CanvasDB system now freely available on GitHub!



Installation of the CanvasDB system

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

Step 1. Download code from github

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

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

Next Step: Whole Genome Sequencing

- New instruments at SciLifeLab for human WGS…

Capacity of HiSeq 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

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.

SciLifeLab

Uppsala Genome Center

SNP&SEQ Technology Platform

NATIONAL GENOMICS INFRASTRUCTURE

Karolinska Institutet

KTH VETENSKAP OCH KONST

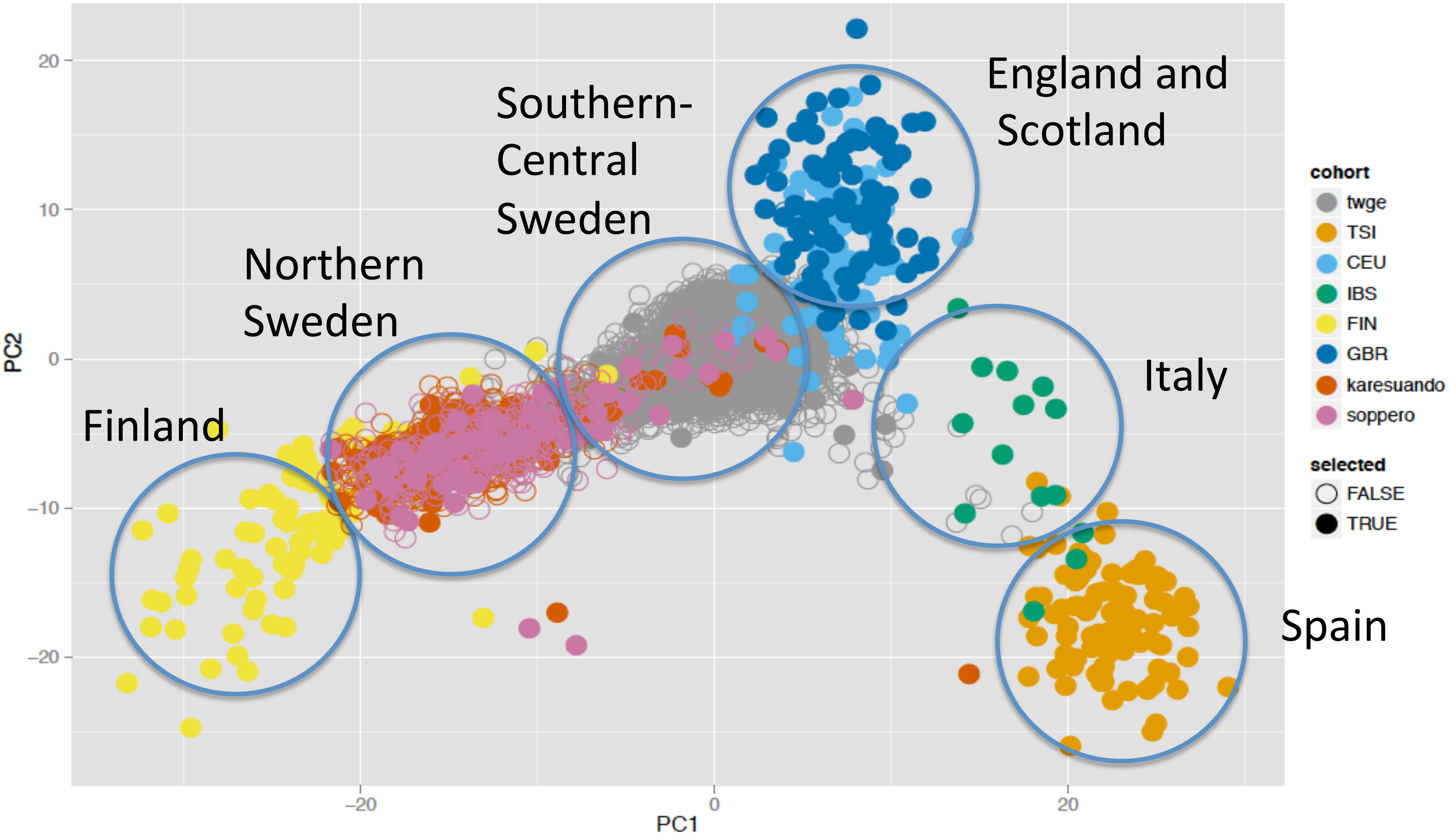
Stockholms universitet

UPPSALA UNIVERSITET

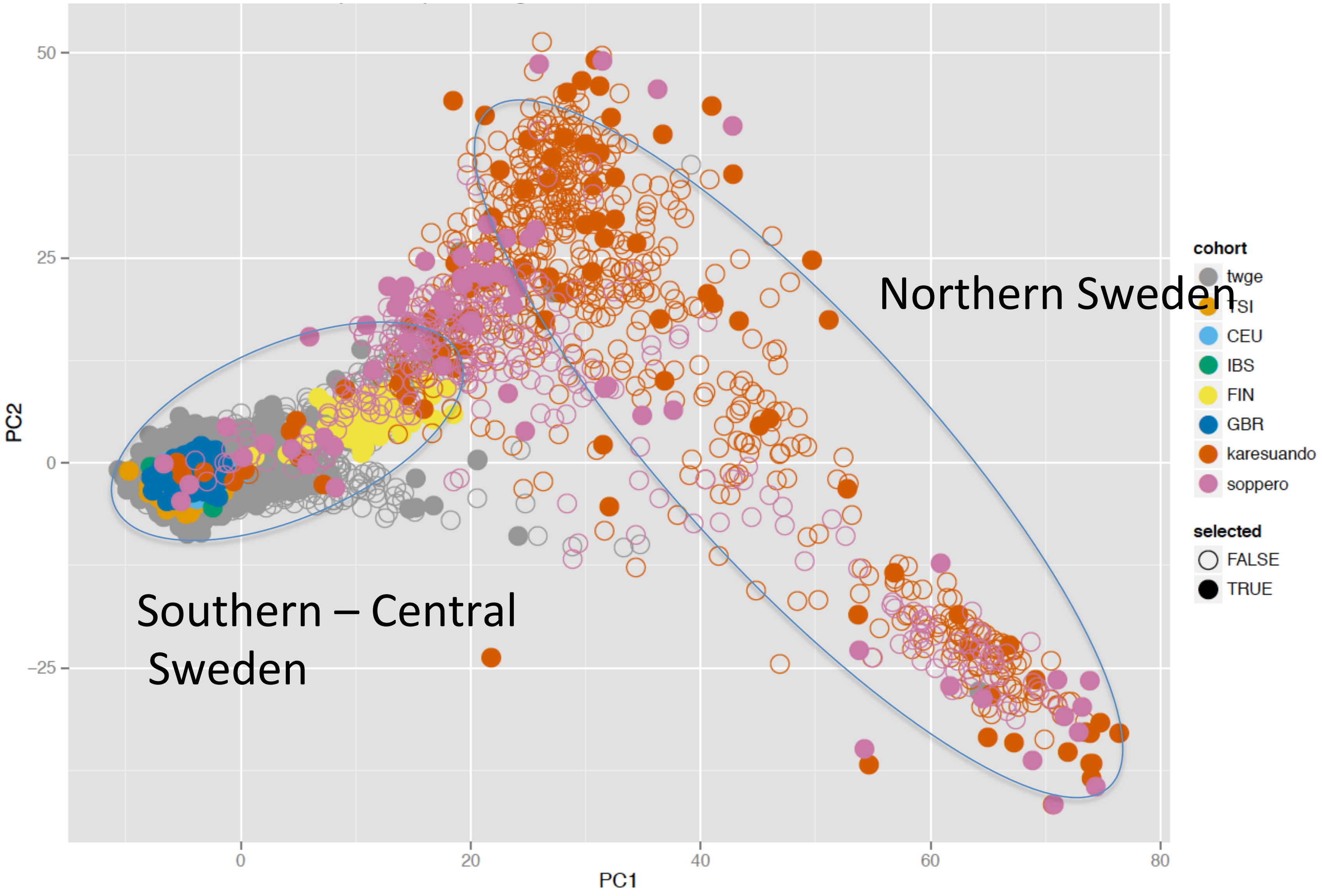
The Swedish Twin registry

- Inclusion based on twinning
- Distribution like population density
- General population-prevalence of any 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



European samples from 1,000 genomes project and 1,000 selected Swedish samples



Northern Sweden

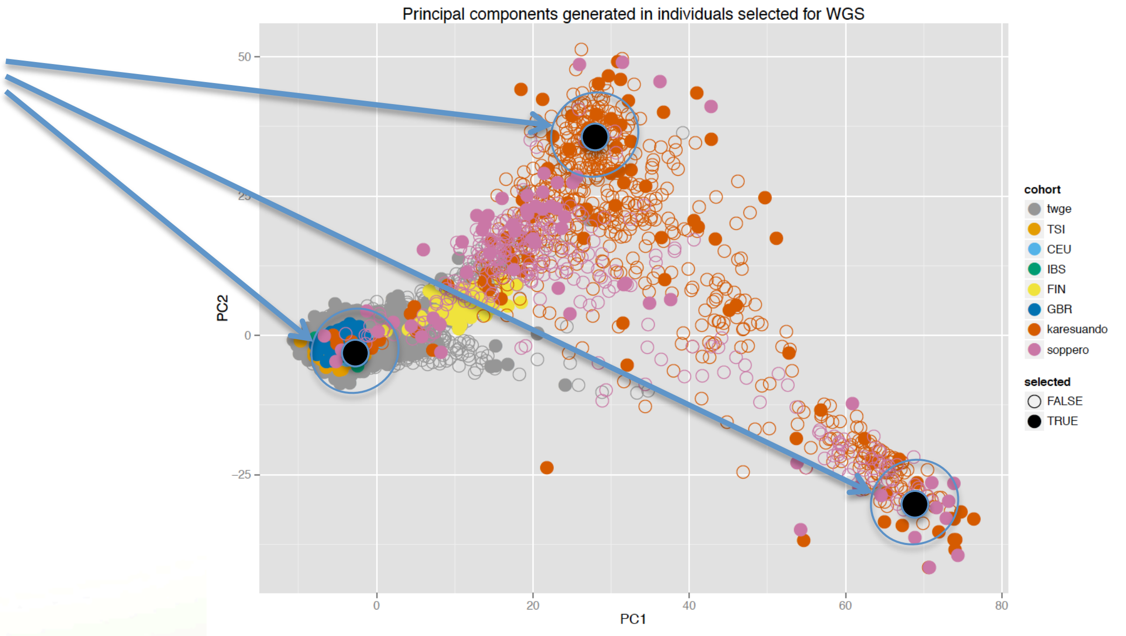
Southern – Central Sweden

WGS of Swedish control cohort

Step 1: 30X Illumina data of the 1,000 individuals (Q2 2016)

Step 2: PacBio *de novo* sequencing of 3 individuals (Q2 2016)

Ref genome
individuals



Step 3: Sequencing of HLA and other clinically relevant loci

Example II:

Assembly of genomes using Pacific Biosciences

Genome assembly using NGS

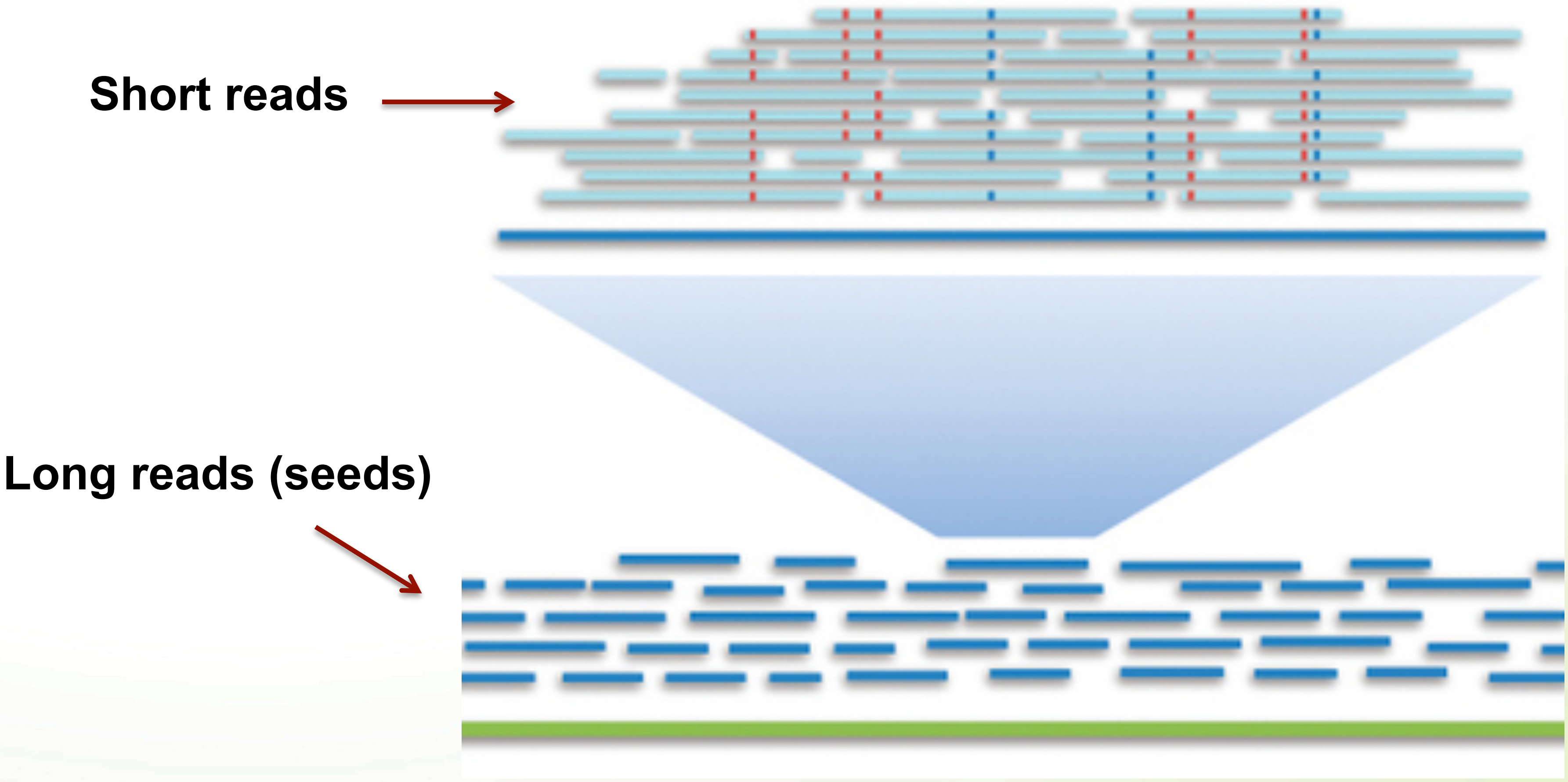
- Short-read *de novo* assembly by NGS
  - Requires mate-pair sequences
    - Ideally with different insert sizes
  - Complicated analysis
    - Assembly, scaffolding, finishing
    - Maybe even some manual steps

  => Rather expensive and time consuming

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

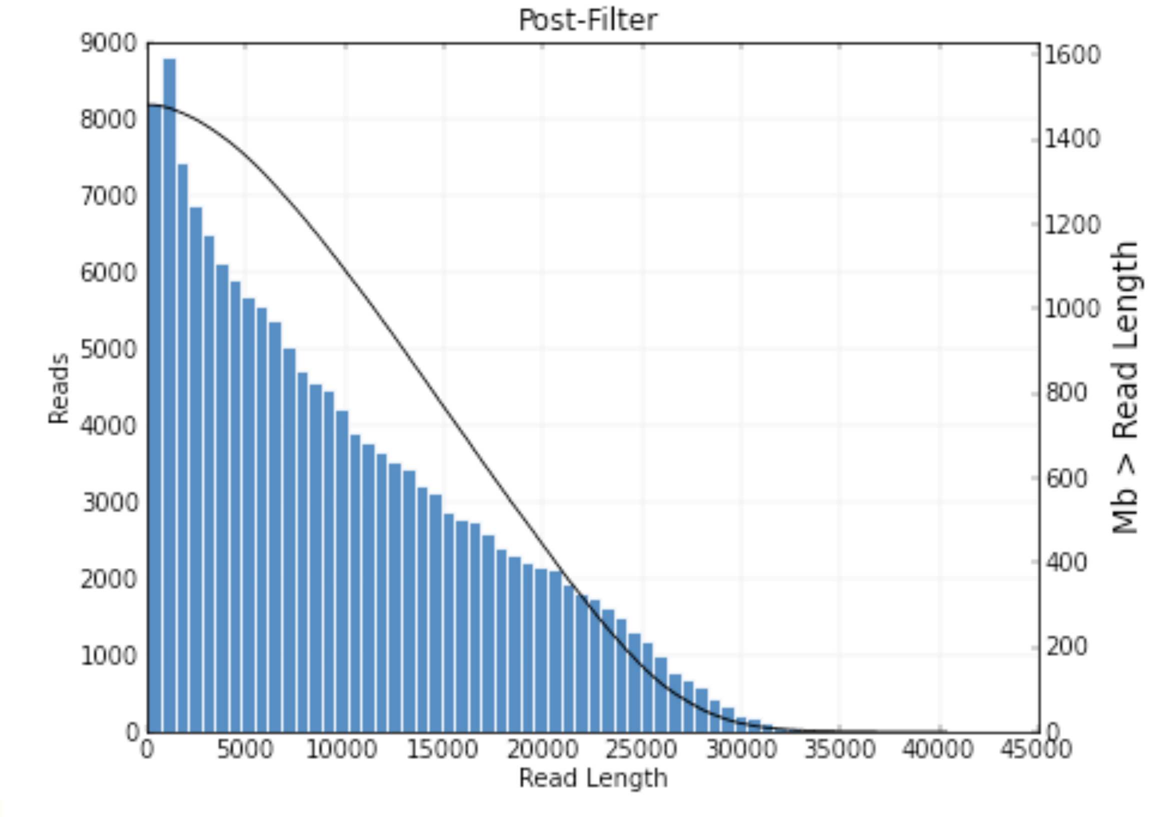
HGAP *de novo* assembly

- HGAP uses both long and shorter reads



PacBio – Throughput & read lengths

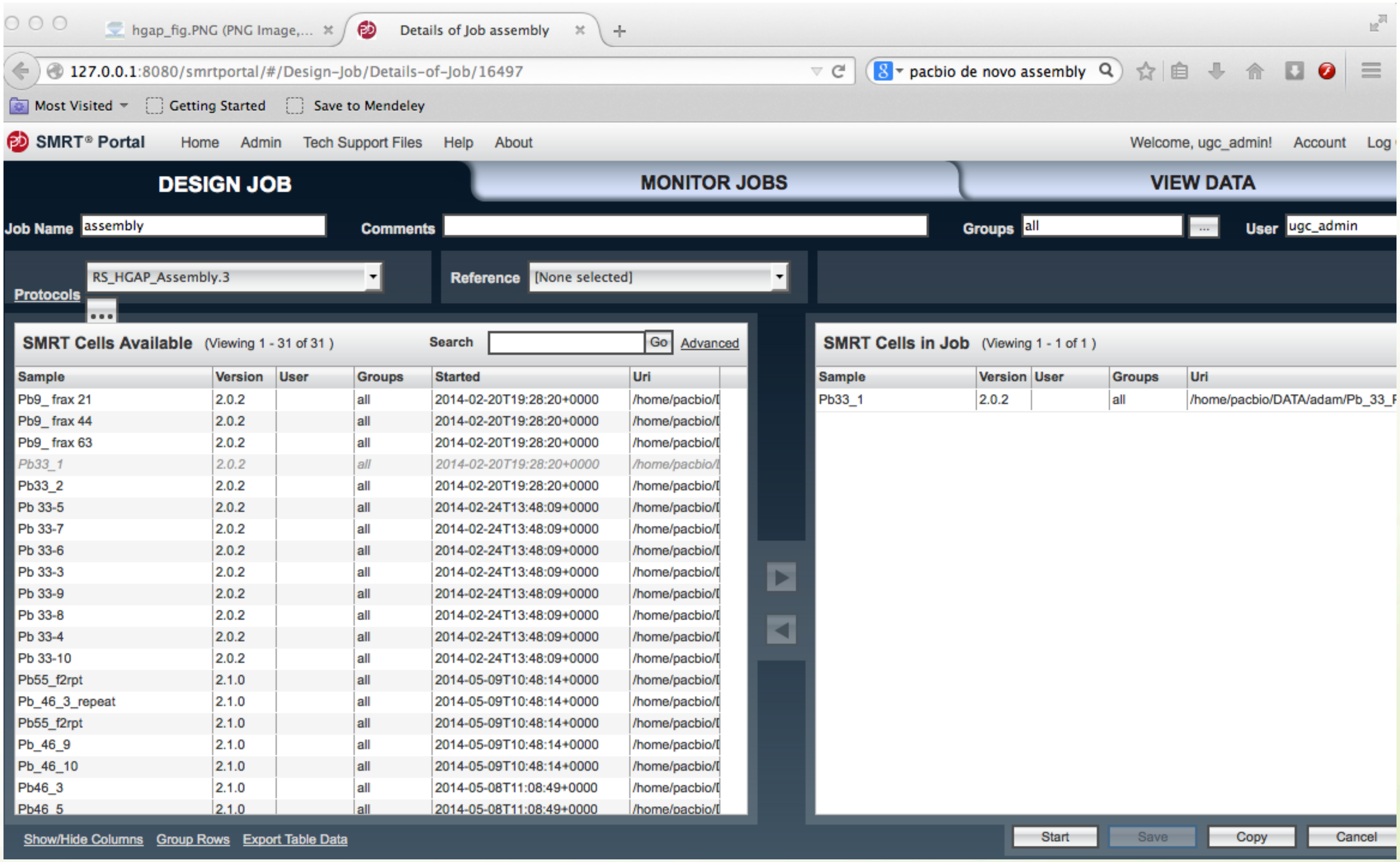
- >10kb average read lengths! (run from April 2014)



- ~ 1 Gb of sequence from one PacBio SMRT cell

PacBio assembly analysis

- Simple -- just click a button!!



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:

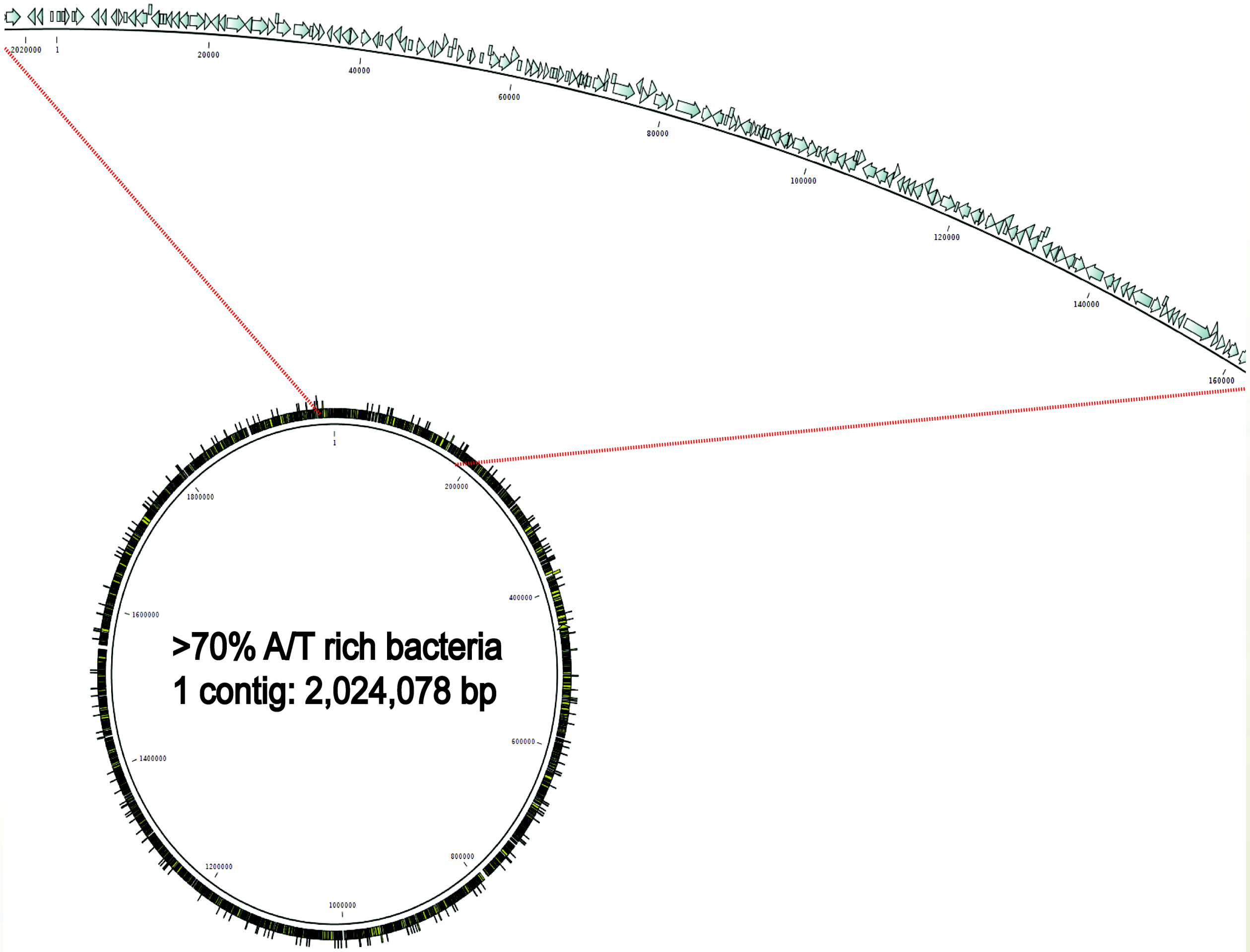
Sample	Version	User	Groups	Started	Uri
Pb9_ frax 21	2.0.2		all	2014-02-20T19:28:20+0000	/home/pacbio/
Pb9_ frax 44	2.0.2		all	2014-02-20T19:28:20+0000	/home/pacbio/
Pb9_ frax 63	2.0.2		all	2014-02-20T19:28:20+0000	/home/pacbio/
Pb33_1	*2.0.2*		*all*	*2014-02-20T19:28:20+0000*	*/home/pacbio/*
Pb33_2	2.0.2		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

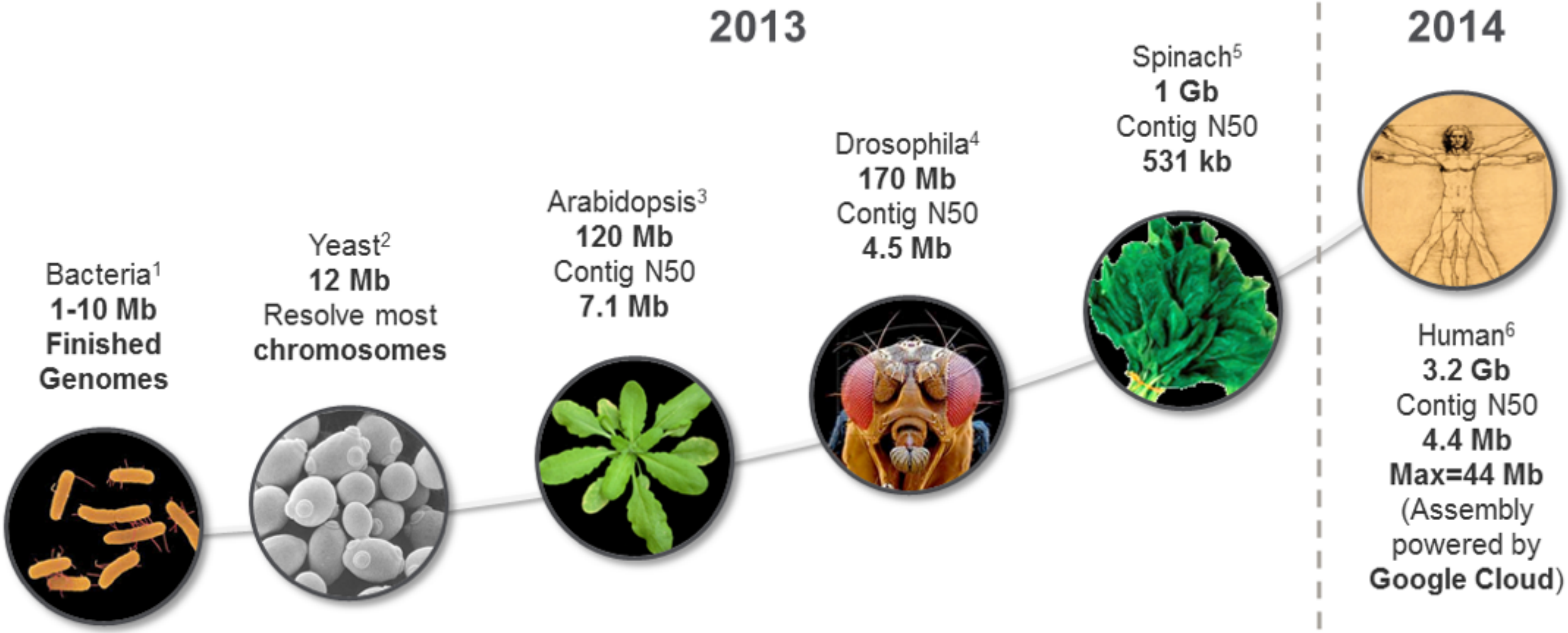
PacBio assembly, example result

- Example: Complete assembly of a bacterial genome



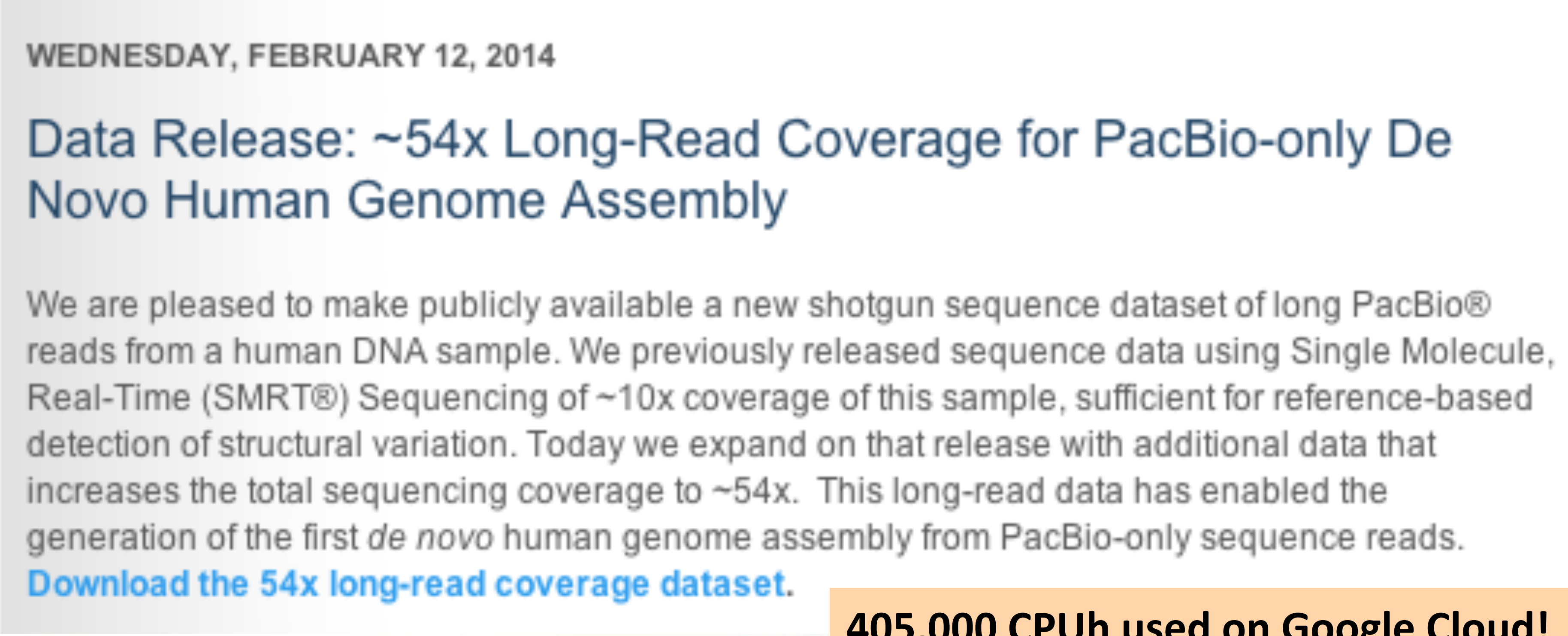
PacBio assembly – recent developments

- Also larger genomes can be assembled by PacBio..



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

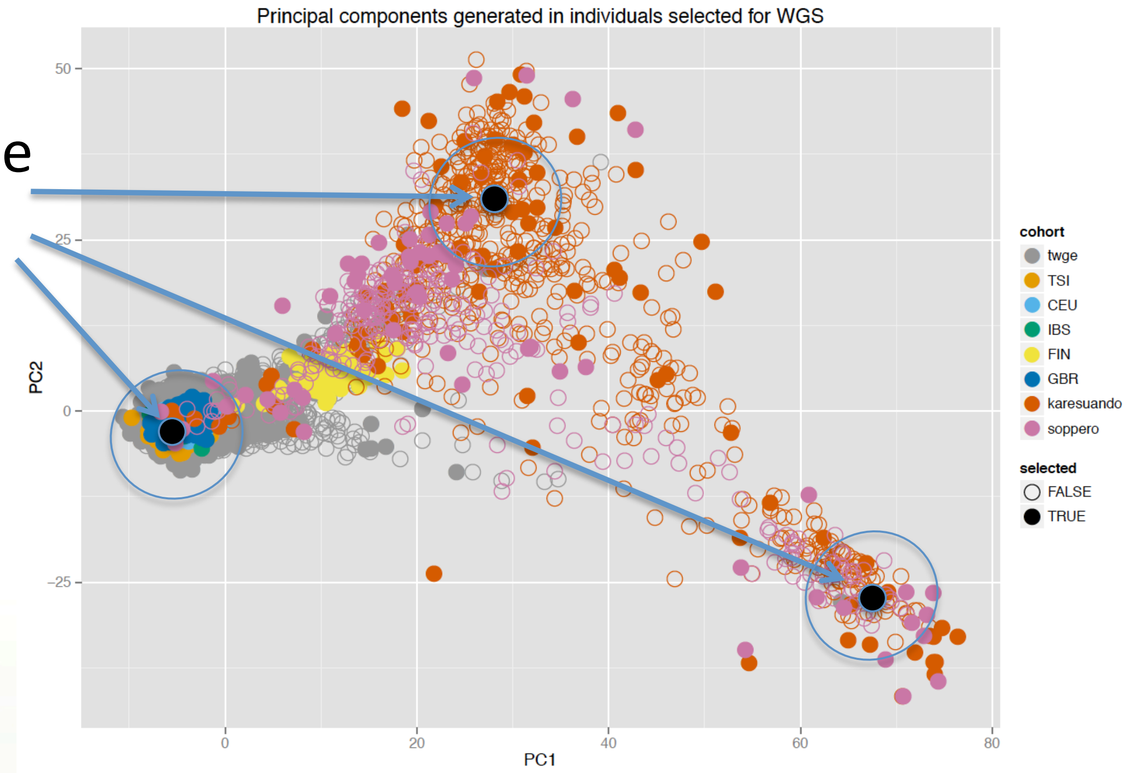
We are pleased to make publicly available a new shotgun sequence dataset of long PacBio® reads from a human DNA sample. We previously released sequence data using Single Molecule, Real-Time (SMRT®) Sequencing of ~10x coverage of this sample, sufficient for reference-based detection of structural variation.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

De novo WGS of Swedish cohort

Establish Swedish reference genome sequences by *de novo* assembly of long-read PacBio data (+10X Genomics?)

Ref genome individuals



First Swedish PacBio WGS

- 20 kb library
- 157 SMRT cells
- 140 Gb data (~45X)
- FALCON assembly

	First PacBio Assembly
# of contigs (>=0 bp)	7708
# of contigs (>=1000 bp)	7653
Total length (>=0 bp)	2844 Mb
Total length (>=1000 bp)	2844 Mb
No of contigs	7692
Largest contig	19.5 Mb
Total contig length	2844 Mb
N50	4.35 Mb
N75	1.97 Mb

Why clinical WGS using long reads?

Precision medicine requires high-quality genome sequences!

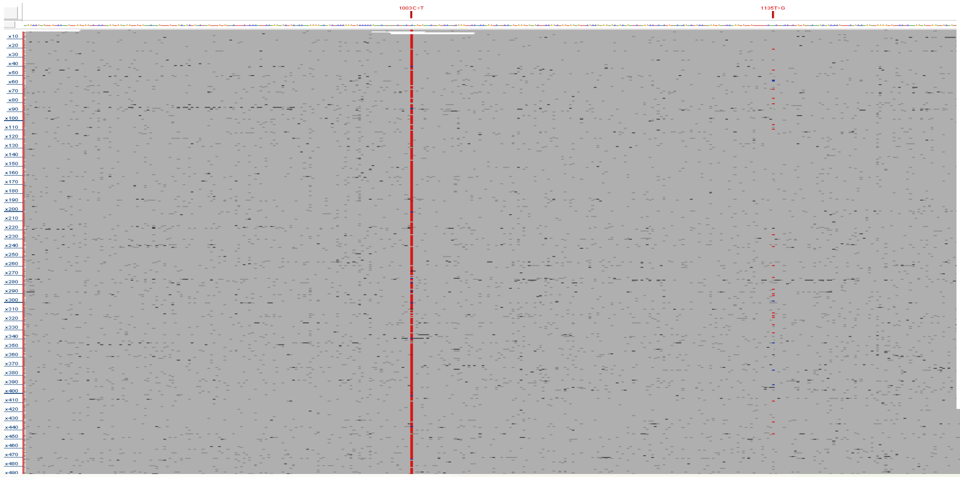
- Resolving repetitive and complex regions
- Annotation of unknown genomic regions
- Haplotype phasing
- …

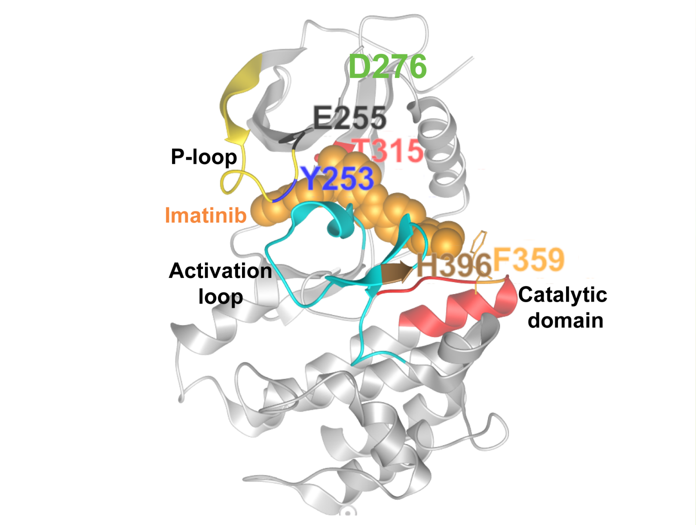
Jim Lupski: "The Goal Is De Novo Assembly in the Clinic"

Jim Lupski, Baylor

Example III:

Clinical sequencing for Leukemia Treatment

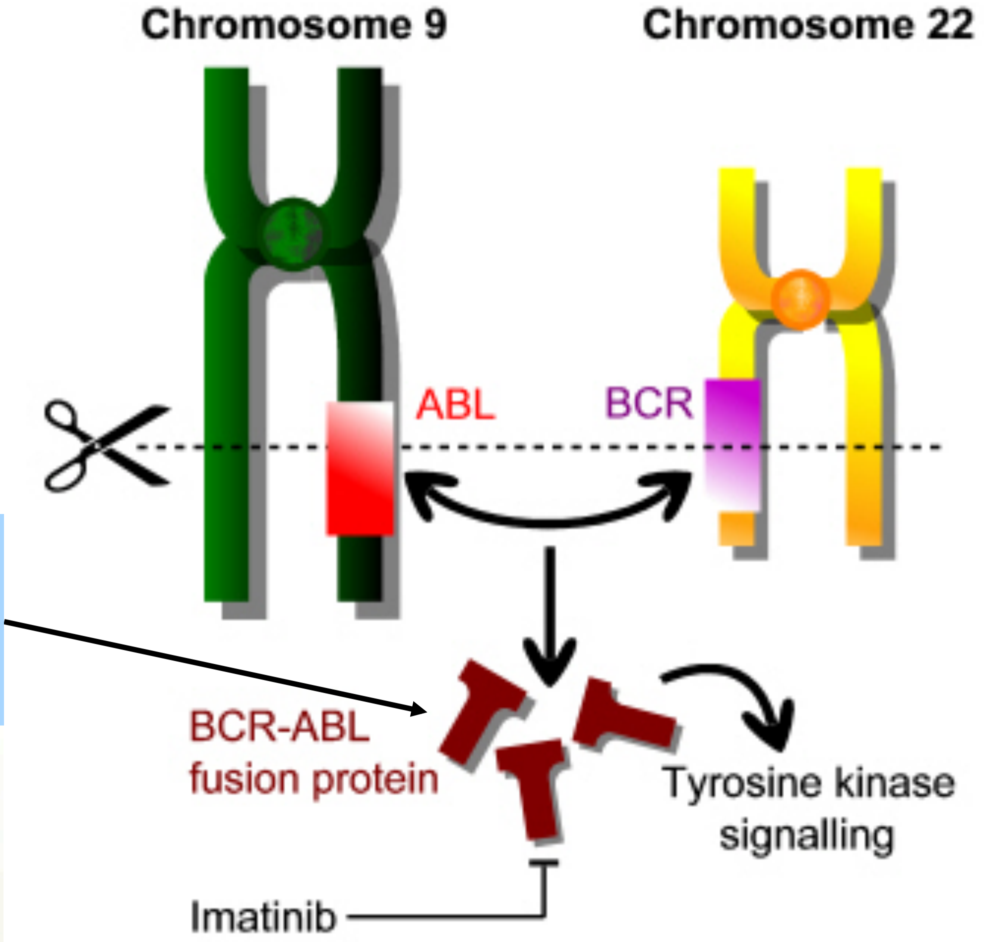




Chronic Myeloid Leukemia

- BCR-ABL1 fusion protein – a CML drug target

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

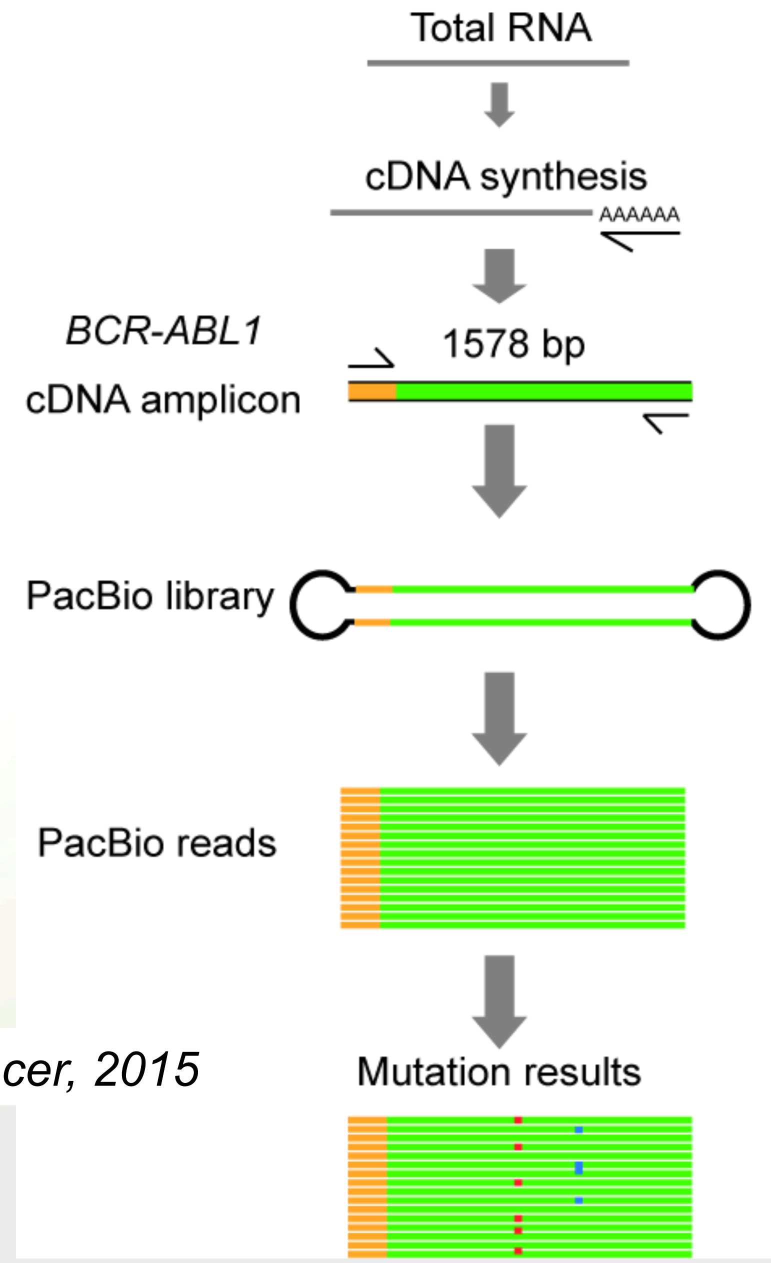


www.cambridgemedicine.org/article/doi/10.7244/cmj-1355057881

BCR-ABL1 workflow – PacBio Sequencing

From sample to results: < 1 week

1 sample/SMRT cell



Total RNA

cDNA synthesis

AAAAAA

BCR-ABL1

cDNA amplicon

1578 bp

PacBio library

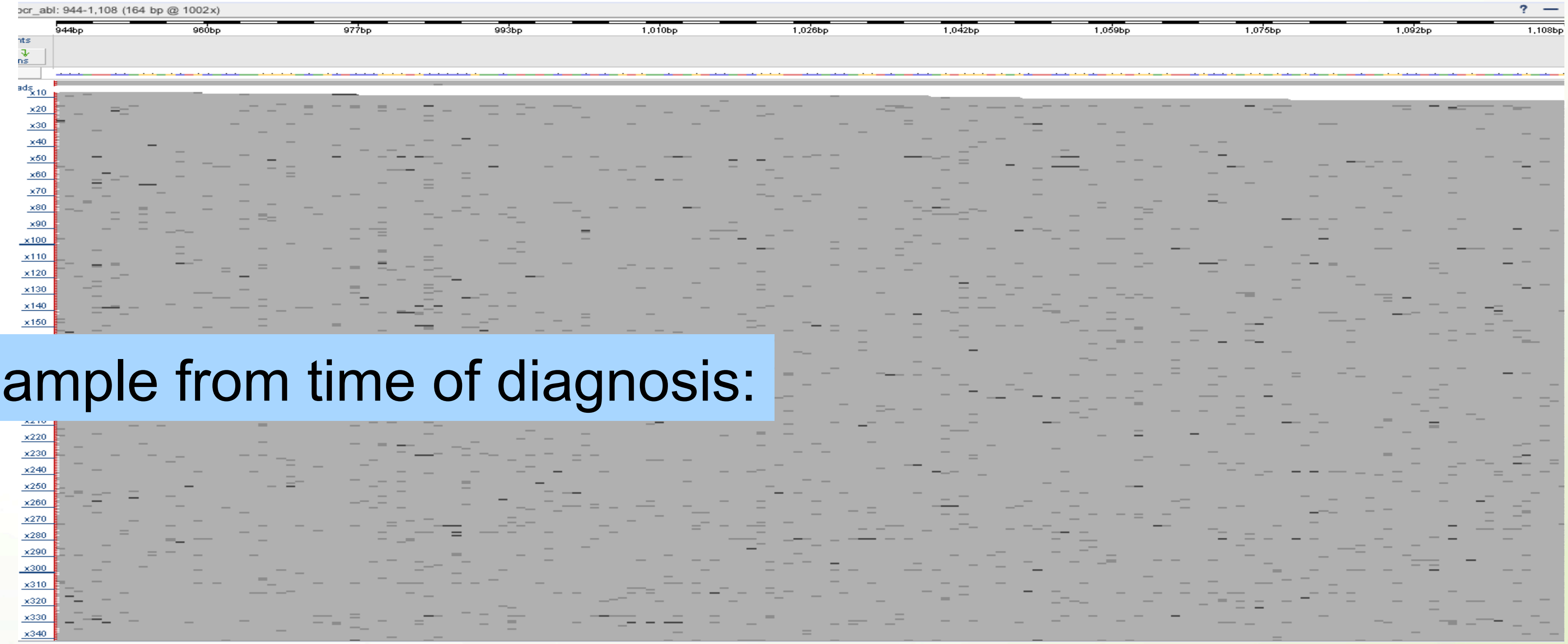
PacBio reads

Mutation results

Cavelier et al., BMC Cancer, 2015

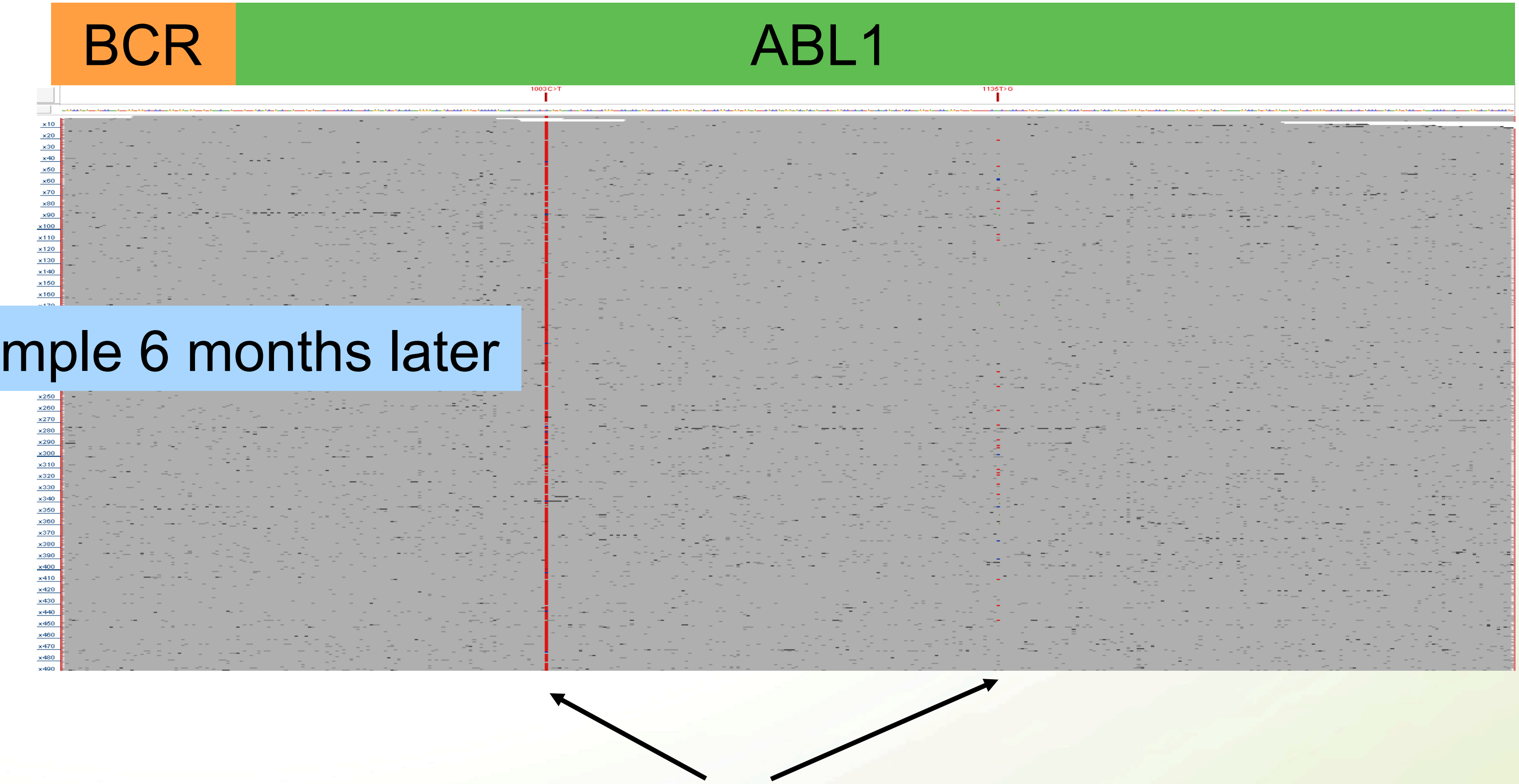
BCR-ABL1 mutations at diagnosis

PacBio sequencing generates ~10 000X coverage!



Sample from time of diagnosis:

BCR-ABL1 mutations in follow-up sample

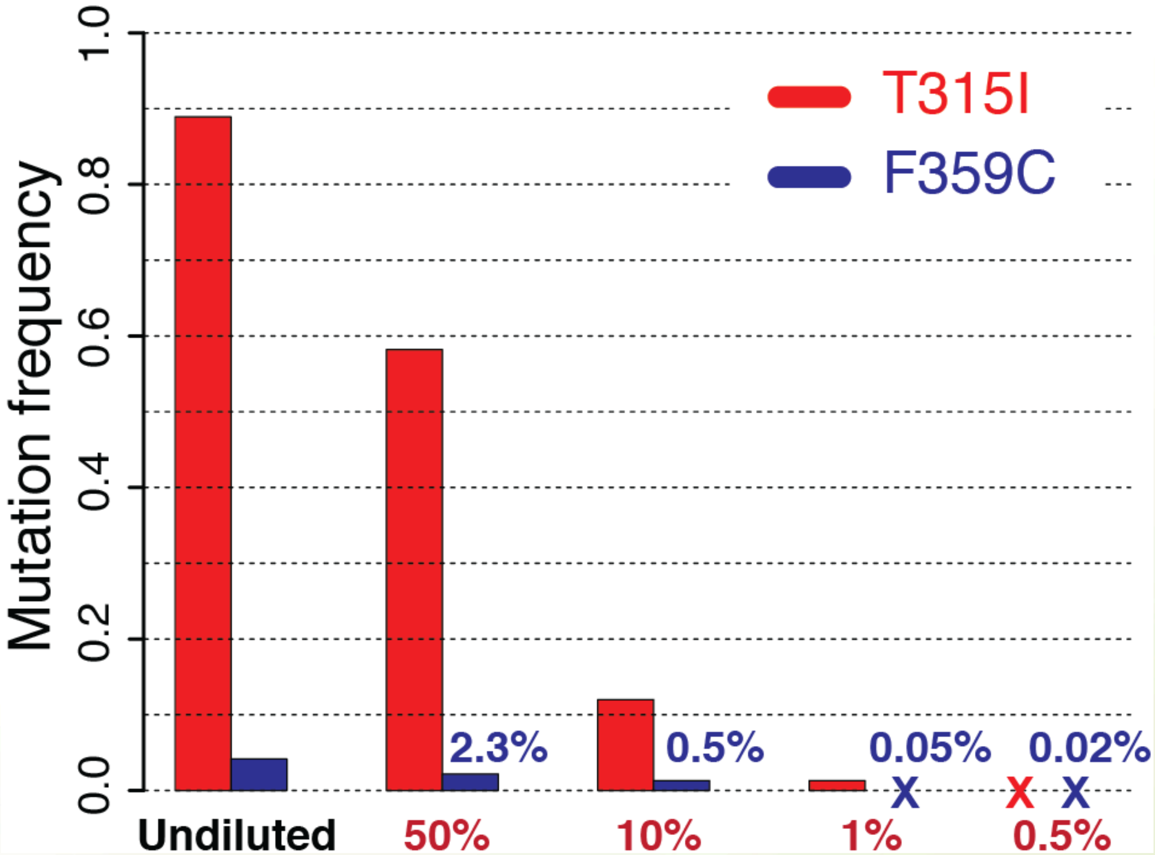


Sample 6 months later

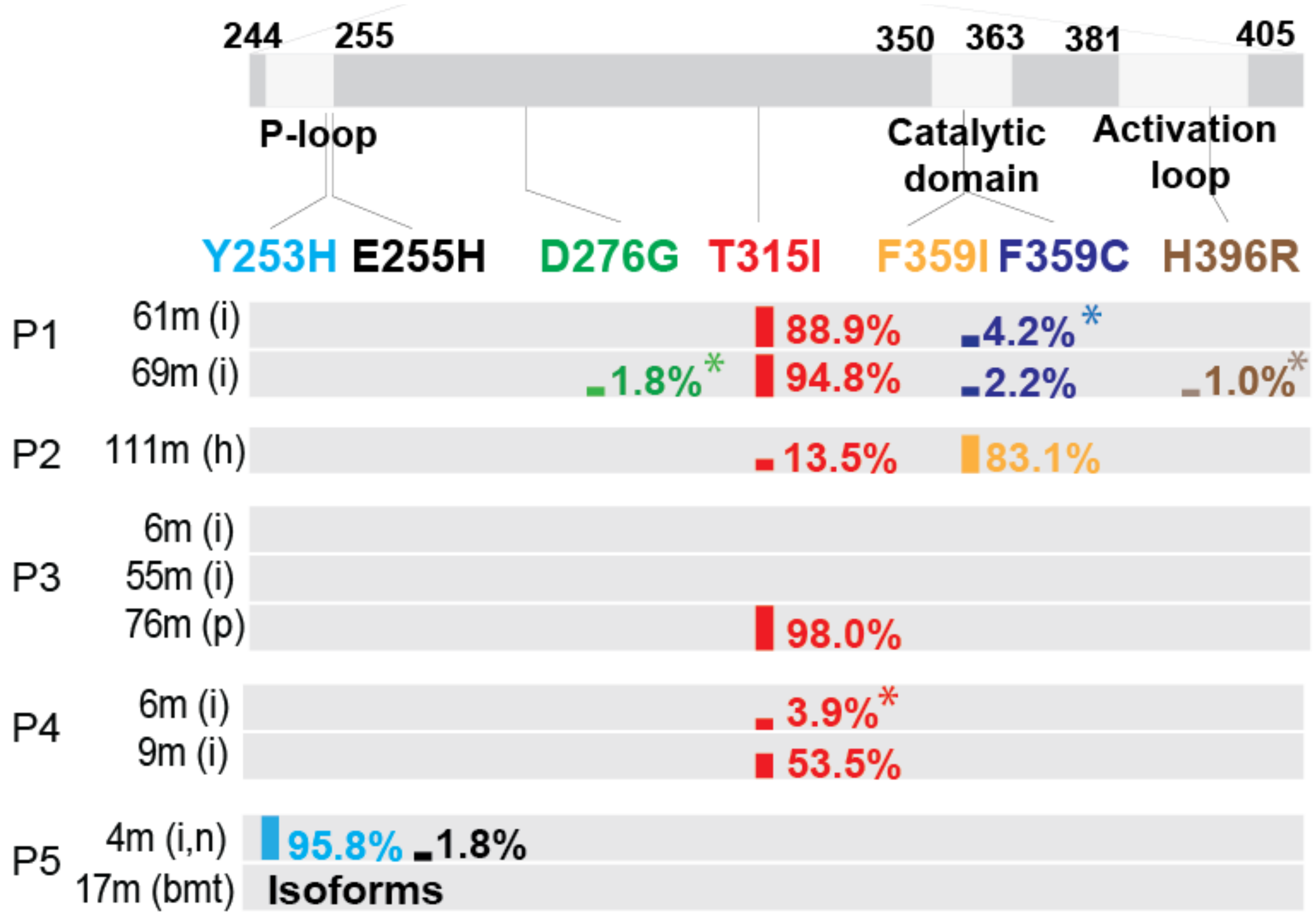
Mutations acquired in fusion transcript.
Might require treatment with alternative drug.

BCR-ABL1 dilution series results

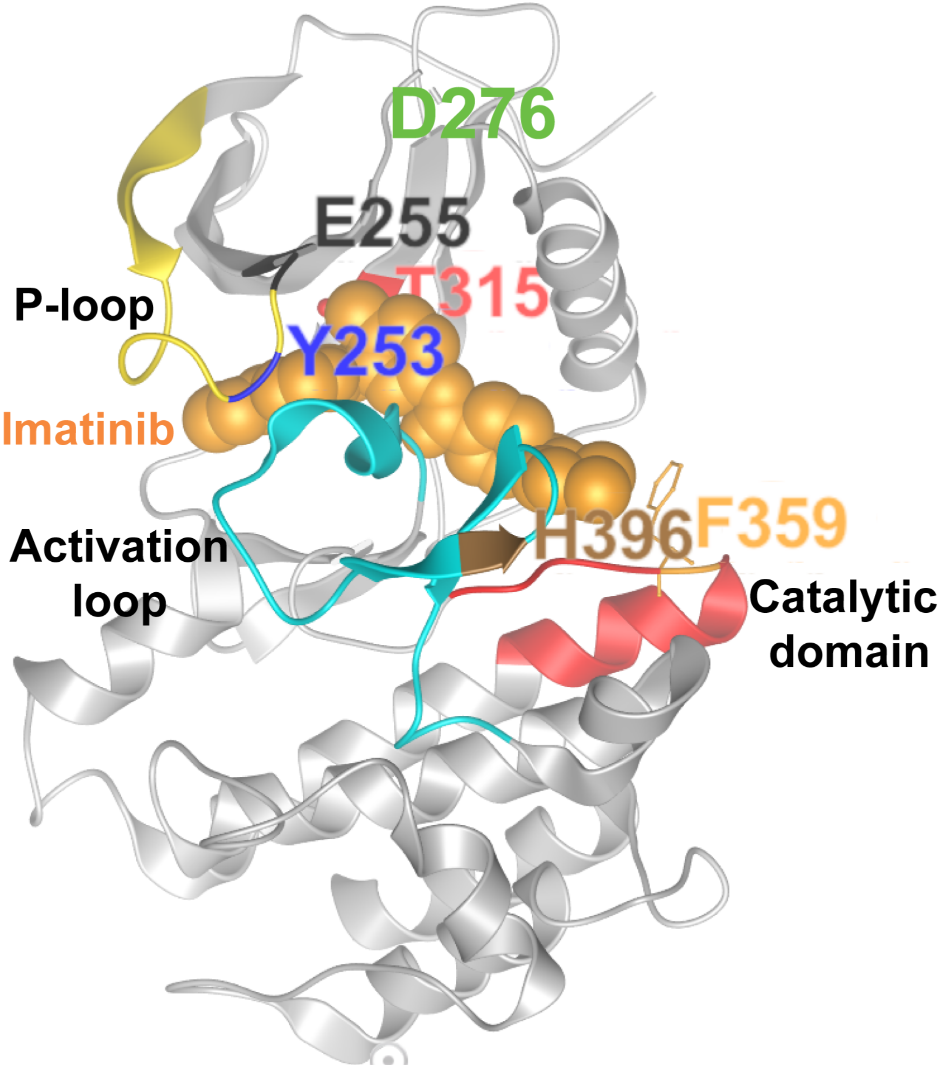
- Mutations down to 1% detected!



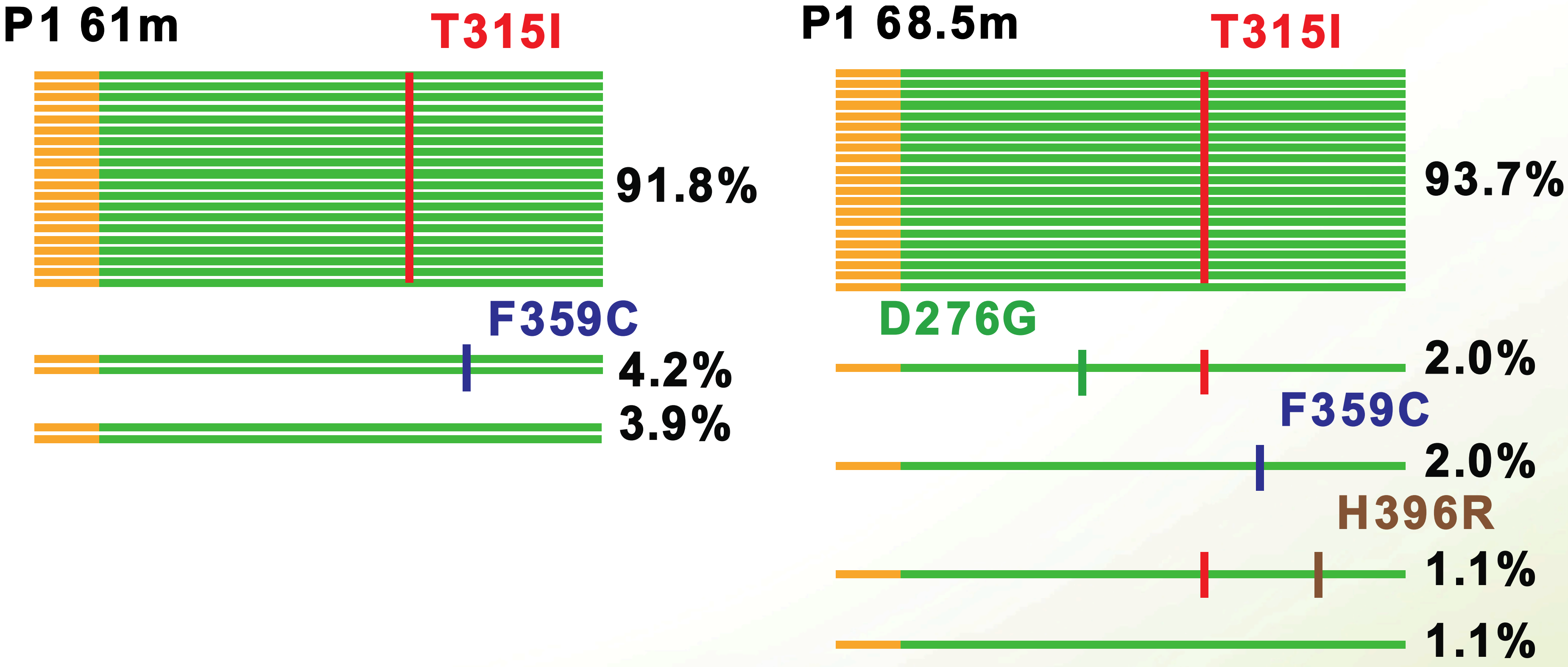
Summary of mutations in 5 CML patients



Mutations mapped to protein structure



BCR-ABL1 - Compound mutations



Karolinska Institutet

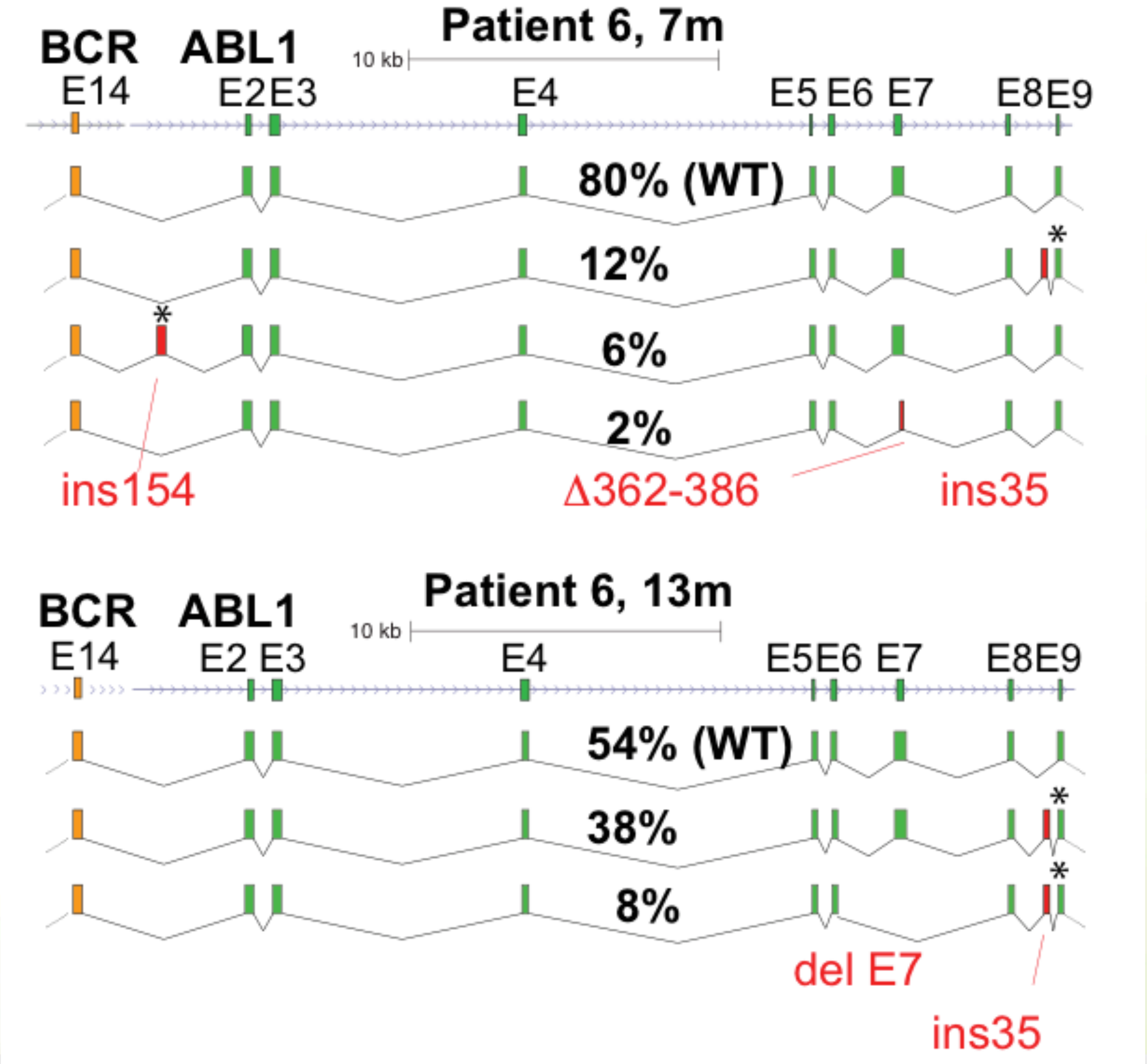
KTH

Stockholms universitet

UPPSALA UNIVERSITET

SciLifeLab

BCR-ABL1 - Multiple isoforms in one individual!



Karolinska Institutet

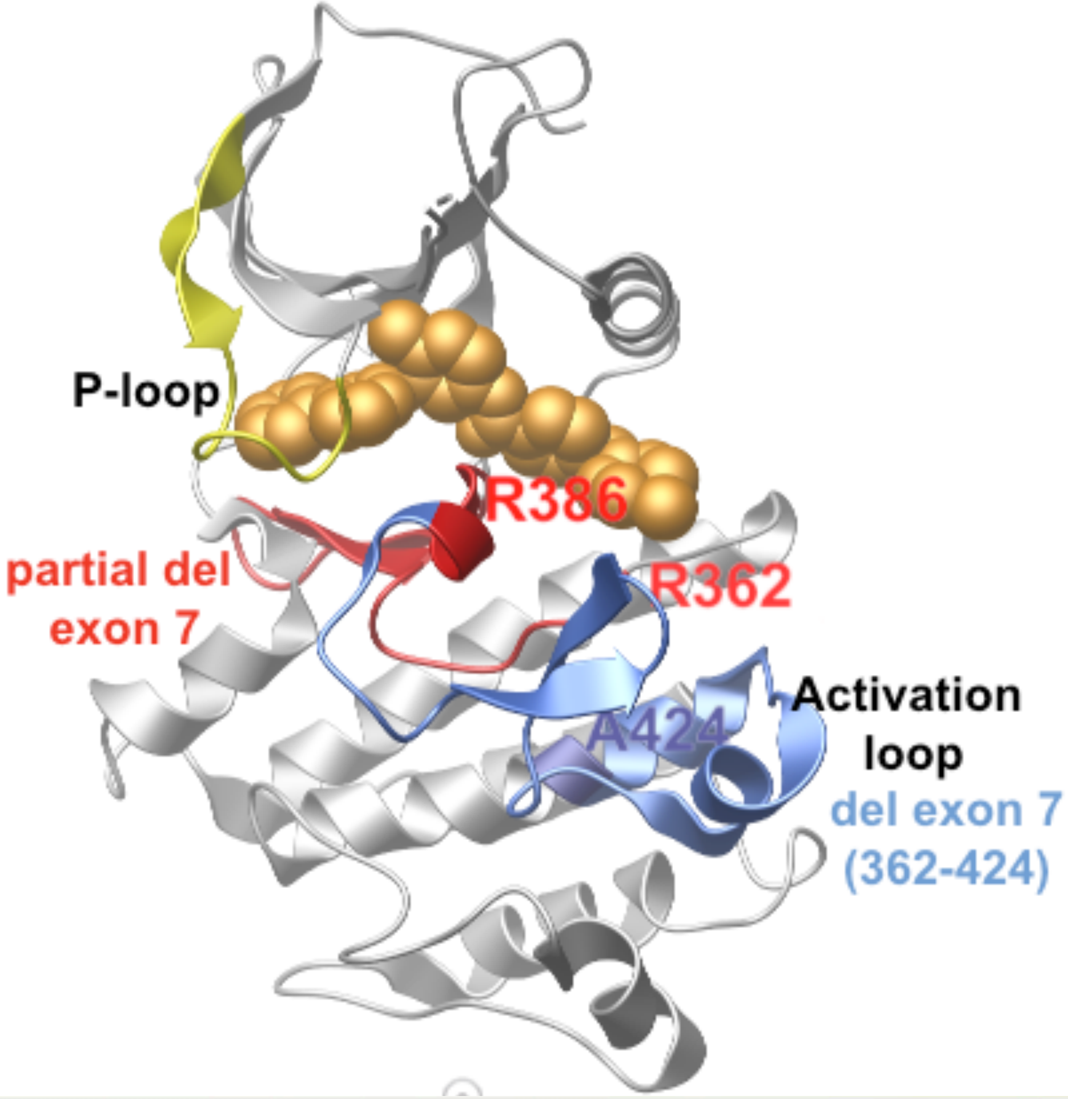
KTH

Stockholms universitet

UPPSALA UNIVERSITET

SciLifeLab

BCR-ABL1 – Isoforms and protein structure



Clinical Diagnosis of BCR-ABL1 mutations

Clinical Genetics

- Collection of samples
- Seq library preparation

Sequencing Facility

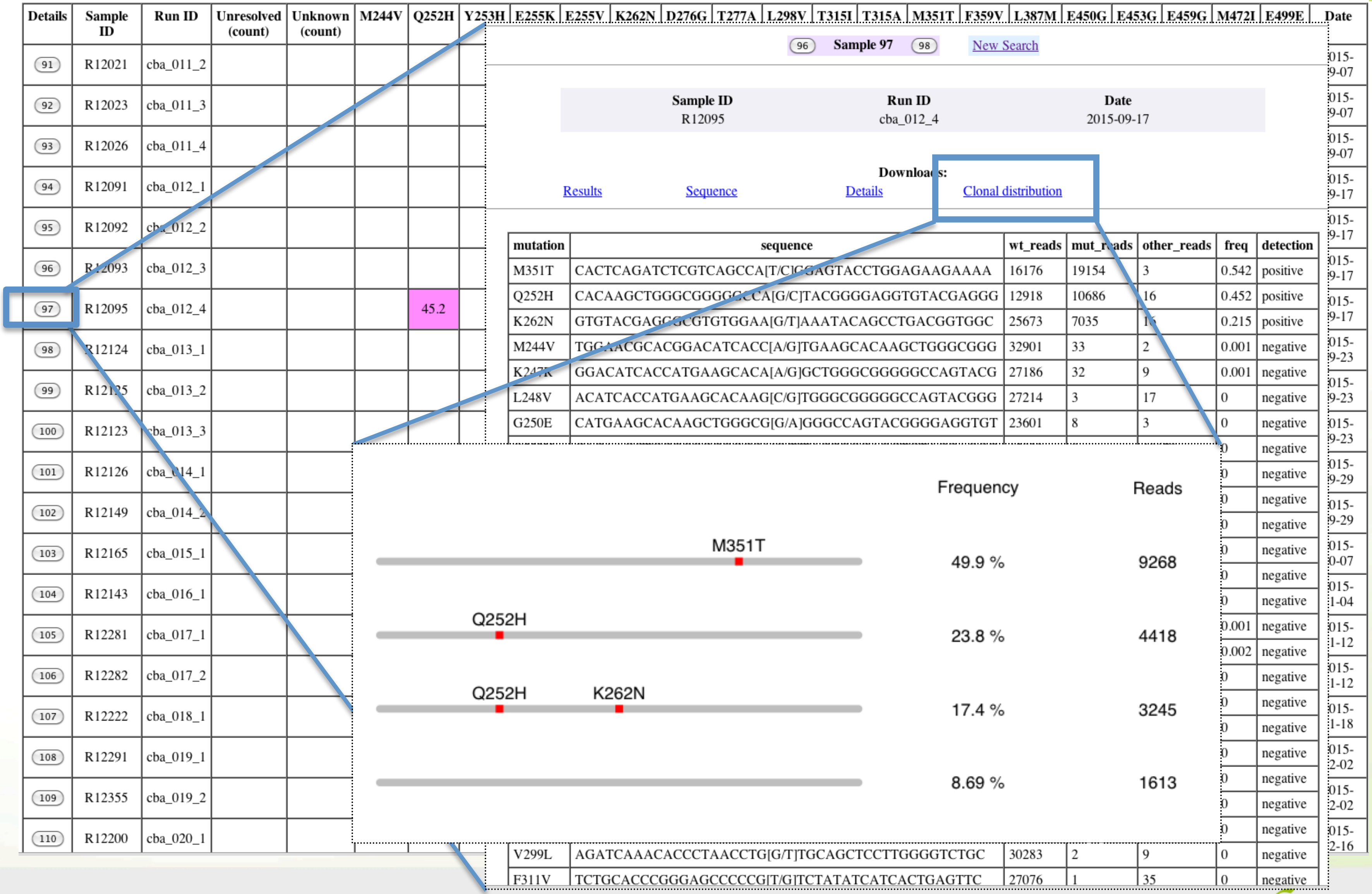
- SMRT sequencing
- Mutational analysis

IT developers

- Web server for results

- Ongoing routine service, 0-4 samples/week
- Over 120 patient samples run so far
- 100% of Sanger-positive mutations recovered
- Developments: Detect low frequency mutations down to 0.1%

Web system for result sharing



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																						015-9-07
92	R12023	cba_011_3																						015-9-07
93	R12026	cba_011_4																						015-9-07
94	R12091	cba_012_1																						015-9-17
95	R12092	cba_012_2																						015-9-17
96	R12093	cba_012_3																						015-9-17
97	R12095	cba_012_4				45.2																		015-9-17
98	R12124	cba_013_1																						015-9-23
99	R12125	cba_013_2																						015-9-23
100	R12123	cba_013_3																						015-9-23
101	R12126	cba_014_1																						015-9-29
102	R12149	cba_014_2																						015-9-29
103	R12165	cba_015_1																						015-0-07
104	R12143	cba_016_1																						015-1-04
105	R12281	cba_017_1																						015-1-12
106	R12282	cba_017_2																						015-1-12
107	R12222	cba_018_1																						015-1-18
108	R12291	cba_019_1																						015-2-02
109	R12355	cba_019_2																						015-2-02
110	R12200	cba_020_1																						015-2-16

96 Sample 97 98 New Search

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

Downloads:

Results Sequence Details Clonal distribution

mutation	sequence	wt_reads	mut_reads	other_reads	freq	detection
M351T	CACTCAGATCTCGTCAGCCA[T/C]GGAGTACCTGGAGAAGAAAA	16176	19154	3	0.542	positive
Q252H	CACAAGCTGGGCGGGGGCCA[G/C]TACGGGGAGGTGTACGAGGG	12918	10686	16	0.452	positive
K262N	GTGTACGAGGGCGTGTGGAA[G/T]AAATACAGCCTGACGGTGGC	25673	7035	16	0.215	positive
M244V	TGGAACGCACGGACATCACC[A/G]TGAAGCACAAGCTGGGCGGG	32901	33	2	0.001	negative
K247R	GGACATCACCATGAAGCACA[A/G]GCTGGGCGGGGGCCAGTACG	27186	32	9	0.001	negative
L248V	ACATCACCATGAAGCACAAG[C/G]TGGGCGGGGGCCAGTACGGG	27214	3	17	0	negative
G250E	CATGAAGCACAAGCTGGGCG[G/A]GGGCCAGTACGGGGAGGTGT	23601	8	3	0	negative
					0	negative
					0	negative
					0	negative
					0	negative
					0	negative
					0	negative
					0	negative
					0.001	negative
					0.002	negative
					0	negative
					0	negative
					0	negative
					0	negative
					0	negative
					0	negative
					0	negative
V299L	AGATCAAACACCCTAACCTG[G/T]TGCAGCTCCTTGGGGTCTGC	30283	2	9	0	negative
F311V	TCTGCACCCGGGAGCCCCCG[T/G]TCTATATCATCACTGAGTTC	27076	1	35	0	negative

	Frequency	Reads
M351T	49.9 %	9268
Q252H	23.8 %	4418
Q252H K262N	17.4 %	3245
	8.69 %	1613

Karolinska Institutet

KTH VETENSKAP OCH KONST

Stockholms universitet

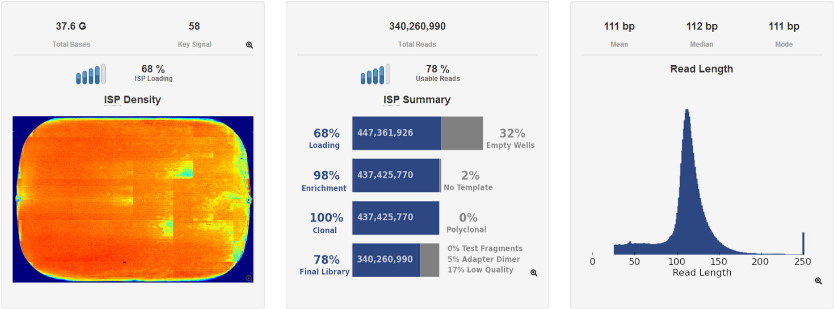
UPPSALA UNIVERSITET

SciLifeLab

Ion Torrent – Ongoing developments

Ion S5 XL system

Ion Proton PII chip (EA)



37.6 G
Total Bases

58
Key Signal

68 %
ISP Loading

ISP Density

340,260,990
Total Reads

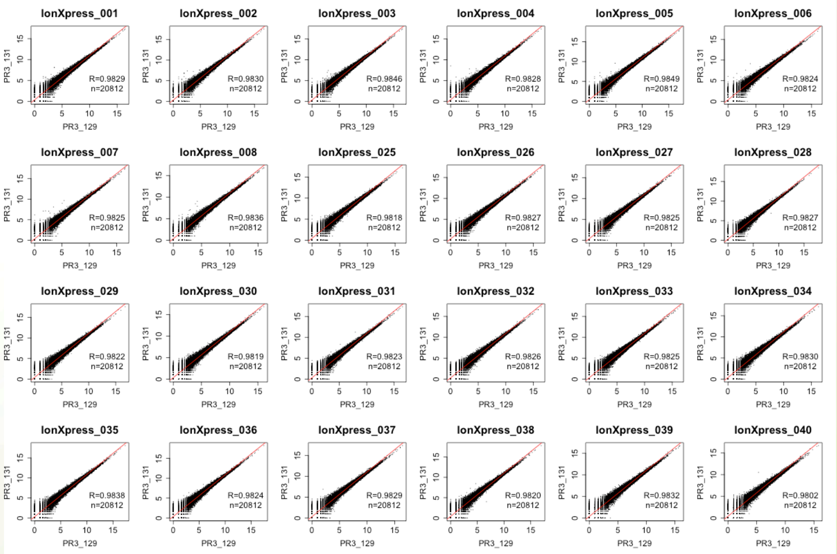
78 %
Usable Reads

ISP Summary

68% Loading	447,361,926	32% Empty Wells
98% Enrichment	437,425,770	2% No Template
100% Clonal	437,425,770	0% Polyclonal
78% Final Library	340,260,990	0% Test Fragments 5% Adapter Dimer 17% Low Quality

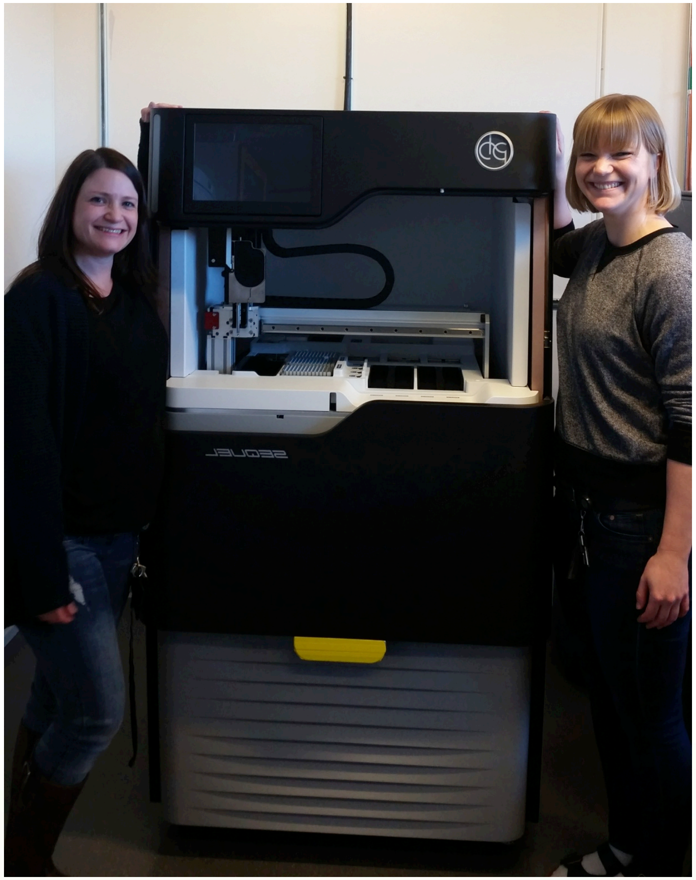
111 bp	112 bp	111 bp
Mean	Median	Mode

Read Length



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

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 syndrom
Leukemia
Scoliosis
HLA typing
Dyslexia
MRSA / BRSA screen
Sudden cardiac arrest
Transcriptional regulation
Prenatal diagnostics
Muscle dystrophy
Individualised cancer therapy
and much more...