

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 whole genome sequencing
 - Assembly using long reads
 - Clinical routine sequencing

illumina®



ThermoFisher
SCIENTIFIC
life technologies
ion torrent
◊ * △ ○ × □ + ≈



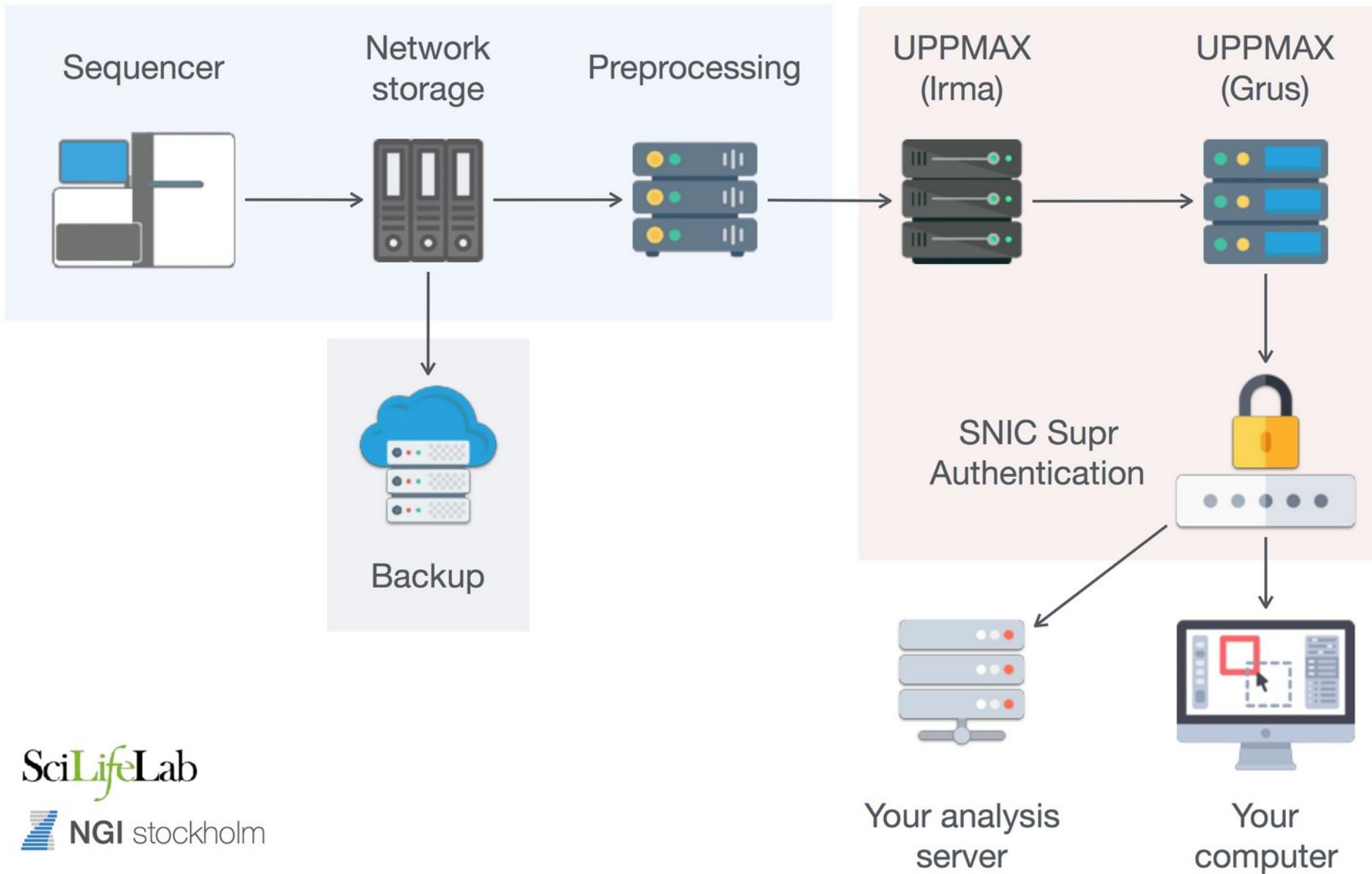
PB PACIFIC BIOSCIENCES®



Oxford
NANOPORE
Technologies



- NGI Data Handling



— Grus Deliveries

- UPPMAX tool for NGI data deliveries
 - NGI creates a SNIC Supr "delivery project" for each NGI sequencing project
 - Project PI and contact person given access, according to what was put on the order form
 - Email sent with project ID and instructions
- Grus is for secure short term storage only
 - Requires two-factor authentication



- Analysis Pipelines

- Initial data analysis for major protocols
- Internal QC and standardised starting point for users
- All software open source and on GitHub
 - <http://opensource.scilifelab.se/>
 - <http://github.com/SciLifeLab/>
- Accredited facility

SciLifeLab

 NGI stockholm



Ackred. nr 1850
Provning
ISO/IEC 17025

- Analysis Requirements



Automated



Reliable

nextflow



Easy for others to run



Reproducible results

Sarek



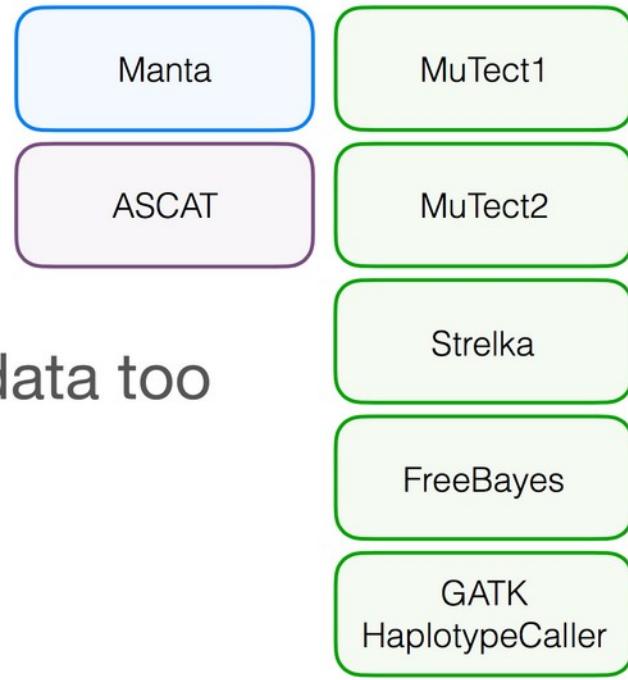
GitHub

<https://github.com/SciLifeLab/Sarek>

- Tumour/Normal pair WGS analysis based on GATK best practices
 - SNPs, SNVs and indels
 - Structural variants
 - Heterogeneity, ploidy and CNVs
- Works with regular WGS and Exome data too



Sarek



nf-core

- A community effort to collect a curated set of Nextflow analysis pipelines
 - GitHub organisation to collect pipelines in one place
 - No institute-specific branding
 - Strict set of guideline requirements
 - Automated testing for code style and function



SciLifeLab

 NGI stockholm

- nf-core

The screenshot shows the nf-core website homepage. At the top, there is a navigation bar with links for Home, Pipelines, Tasks, Docs, and About. Below the navigation is the nf-core logo, which includes a green stylized DNA helix icon. A subtext below the logo reads: "A community effort to collect a curated set of analysis pipelines built using Nextflow." A large green button labeled "VIEW PIPELINES" is centered. Below this, there are three main sections: "For facilities", "For users", and "For developers". Each section has a brief description and a corresponding icon. The "For facilities" section features a blue "Highly optimised pipelines with excellent reporting. Validated releases ensure reproducibility." icon. The "For users" section features a yellow "Portable, documented and easy to use workflows. Pipelines that you can trust." icon. The "For developers" section features a red "Companion templates and tools help to validate your code and simplify common tasks." icon.



- Easy to run pipelines
- Helpful community
- Super reproducible results

This section contains three cards. The first card, "Documentation", shows a clipboard icon and text about extensive documentation. The second card, "CI Testing", shows a Travis CI logo and text about continuous integration testing. The third card, "Stable Releases", shows a green checkmark icon and text about using GitHub releases to tag stable versions of the code.

This section contains three cards. The first card, "Docker", shows a Docker logo and text about software dependencies being available in a bundled Docker container. The second card, "Singularity", shows a Singularity logo and text about Singularity containers. The third card, "Bioconda", shows a Bioconda logo and text about pipelines coming with a Bioconda environment file.

Quality Control

- Every project has some level of quality control checks
 - Sequencing quality
 - FastQC, FastQ Screen
- Analysis pipelines give application-specific QC
 - Qualimap, RSeQC
- Reporting is done using MultiQC

P1234: Test_NGI_Project

General Stats

NGI-RNaseq

Sample Similarity

MDS Plot

STAR

Cutadapt

FastQC

Sequence Quality Histograms

Per Sequence Quality Scores

Per Base Sequence Content

Per Sequence GC Content

Per Base N Content

Sequence Length Distribution

Sequence Duplication Levels

Overrepresented sequences

Adapter Content



P1234: Test_NGI_Project

This is an example project. All identifying data has been removed.

Contact E-mail: phil.ewels@scilifelab.se
 Application Type: RNA-seq
 Sequencing Platform: HiSeq 2500 High Output V4
 Sequencing Setup: 2x125
 Reference Genome: hg19

Report generated on 2017-05-17, 18:43 based on data in:
/Users/philewels/GitHub/MultiQC_website/public_html/examples/ngi-rna/data

☰ NGI names 👤 User supplied names

General Statistics

📋 Copy table configure Configure Columns 📈 Plot Showing 22/22 rows and 5/5 columns.

Sample Name	% Aligned	M Aligned	% Trimmed	% Dups	% GC	M Seqs
P1234_1001	68.2%	22.8	10.3%	71.3%	49%	33.7
P1234_1002	67.9%	20.9	10.7%	70.1%	50%	31.1
P1234_1003	64.7%	21.7	11.0%	72.3%	50%	33.7
P1234_1004	55.2%	17.0	13.2%	73.4%	51%	31.2
P1234_1005	53.0%	17.7	15.9%	75.8%	52%	33.8
P1234_1006	52.7%	16.1	14.1%	73.8%	52%	30.8
P1234_1007	33.0%	7.0	32.0%	60.5%	52%	21.8
P1234_1008	27.5%	4.3	44.2%	79.1%	50%	16.7
P1234_1009	52.3%	10.5	20.9%	64.2%	48%	20.5

Toolbox

A

D

H

M

C

Getting MultiQC



ewels / MultiQC

BIOCONDA



<http://multiqc.info>



PyPI

The screenshot shows a web browser displaying the MultiQC homepage. The address bar contains "multiqc.info". The page features a large "MultiQC" logo with a magnifying glass icon. Below it, a tagline reads: "Aggregate results from bioinformatics analyses across many samples into a single report". A descriptive paragraph explains: "MultiQC searches a given directory for analysis logs and compiles a HTML report. It's a general use tool, perfect for summarising the output from numerous bioinformatics tools." On the right side, there is a navigation menu with links to "Home", "Docs", "Example Reports", and "GitHub". Below this, four blue buttons provide quick access: "Documentation", "View on PyPI", "View on GitHub", and "Quick Install". The "Quick Install" section includes a command-line example:

```
pip install multiqc    # Install  
multiqc .             # Run
```

. At the bottom, a note states: "Python or pip not installed? See the full installation instructions."

Human Whole Genome Sequencing



Crowd image via www.shutterstock.com

WGS projects all around the world

Genomics england

#100kThankYou Google Custom Search

About Us Understanding Genomics Information for Participants Research Industry Partnerships News & Events

Genomics england NHS

Th for h

HEALTH & MEDICINE

A Chinese Province Is Sequencing One Million of Its Residents' Genomes

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

Whole Genome Sequencing in Sweden

- Over 10k genomes/year can be produced in Sweden!
 - Research projects (disease cohorts)
 - Reference database (cross-section of population)
 - Clinical sequencing



The Swedish 1000 Genomes Project

A resource for researchers and clinical labs:



From SweGen release party on Oct 19th 2016!

Why 1000 Swedish whole genomes?

An important resource both for researchers and clinicians

Examples:

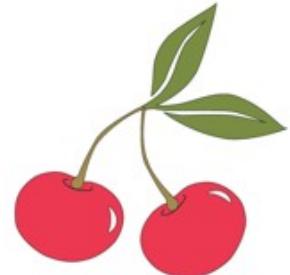
- Look up genetic variant frequencies
- Use as matched controls
- Study population genetics
- Study human evolutionary history

High demand for the data from many different groups:

→ Make the data available as **quickly** and **openly** as possible!

Deciding on a cohort to use for the project

The Swedish Twin Registry:

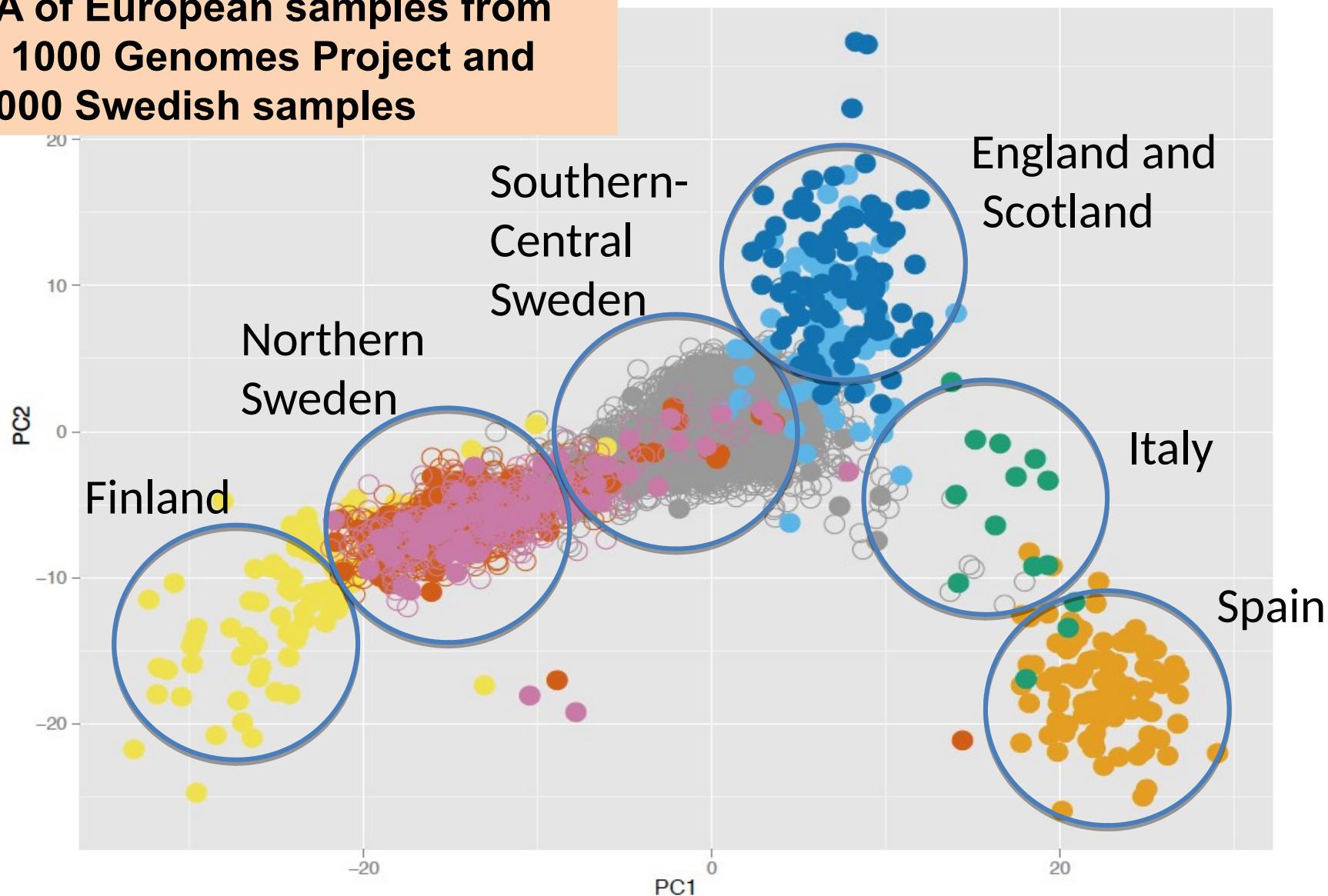


- Inclusion based on twinning
- Distribution like population density
- General population-prevalence of disease
- 10,000 individuals have been analysed with SNP arrays

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

Selecting SweGen samples based on PCA

PCA of European samples from
the 1000 Genomes Project and
10,000 Swedish samples



Step 1: Whole Genome Sequencing

30X Illumina WGS data generated for all 1,000 individuals

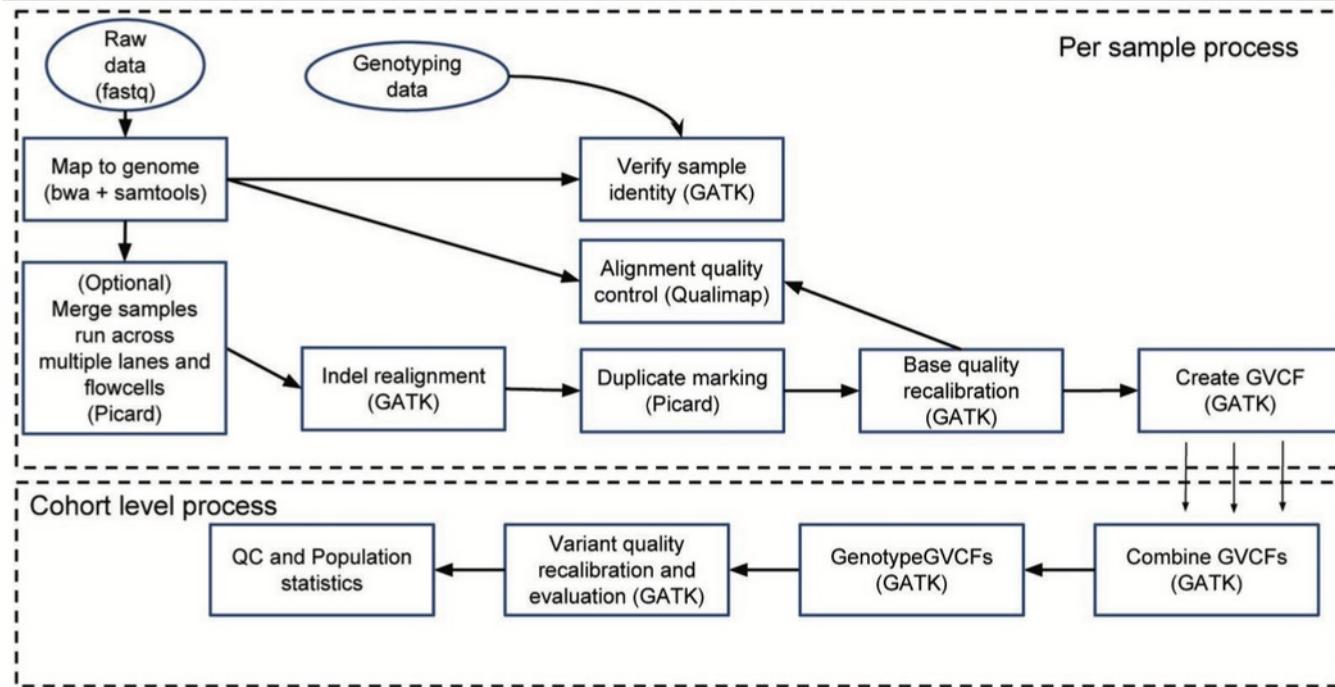


- 509 samples sequenced at NGI Stockholm
- 491 samples sequenced at NGI Uppsala

Sequencing of all 1000 samples completed in September 2016

Step 2: Data management and analysis

Analysis pipeline developed for mapping and variant calling



- About 100,000 Gb data generated within the project
- Over 2 million CPU hours used so far...

This pipeline will become standard for all WGS projects at NGI

Step 3: 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
GENOTYPE
GENOMICS
INFRASTRUCTURE

NBIS

elixir
SWEDEN

Aggregated frequencies available from: [swefreq.nbis.se!](http://swefreq.nbis.se/)

Tracking the usage of SweGen data

~ 600 monthly visits to web site at present



Rank	Location	Visits	Percentage
1.	Uppsala	1 220	32,73 %
2.	Solna	360	9,66 %
3.	Stockholm	346	9,28 %
4.	Gothenburg	166	4,45 %
5.	Huddinge	146	3,92 %
6.	Umeå	134	3,60 %
7.	Lund	102	2,74 %
8.	Linköping	82	2,20 %
9.	(not set)	64	1,72 %
10.	Orebro	49	1,31 %

SweGen – results at a glance

- **33 million** small genetic variants detected in the 1000 samples
 - 29.2 million SNPs, 3.8 million indels
 - **9.9 million** of these were novel (i.e. not in dbSNP147)
 - **26,635** of the novel variants are changing protein sequence
 - An average of **7,199** novel variants detected per individual
- The SweGen dataset contributes novel genetic variation

Genome assembly using long reads



PacBio assembly analysis

- Simple -- just click a button! (for small genomes)

The screenshot shows the SMRT Portal interface for a 'Design Job' assembly. The top navigation bar includes links for Home, Admin, Tech Support Files, Help, and About, along with user information: Welcome, ugc_admin!, Account, and Log.

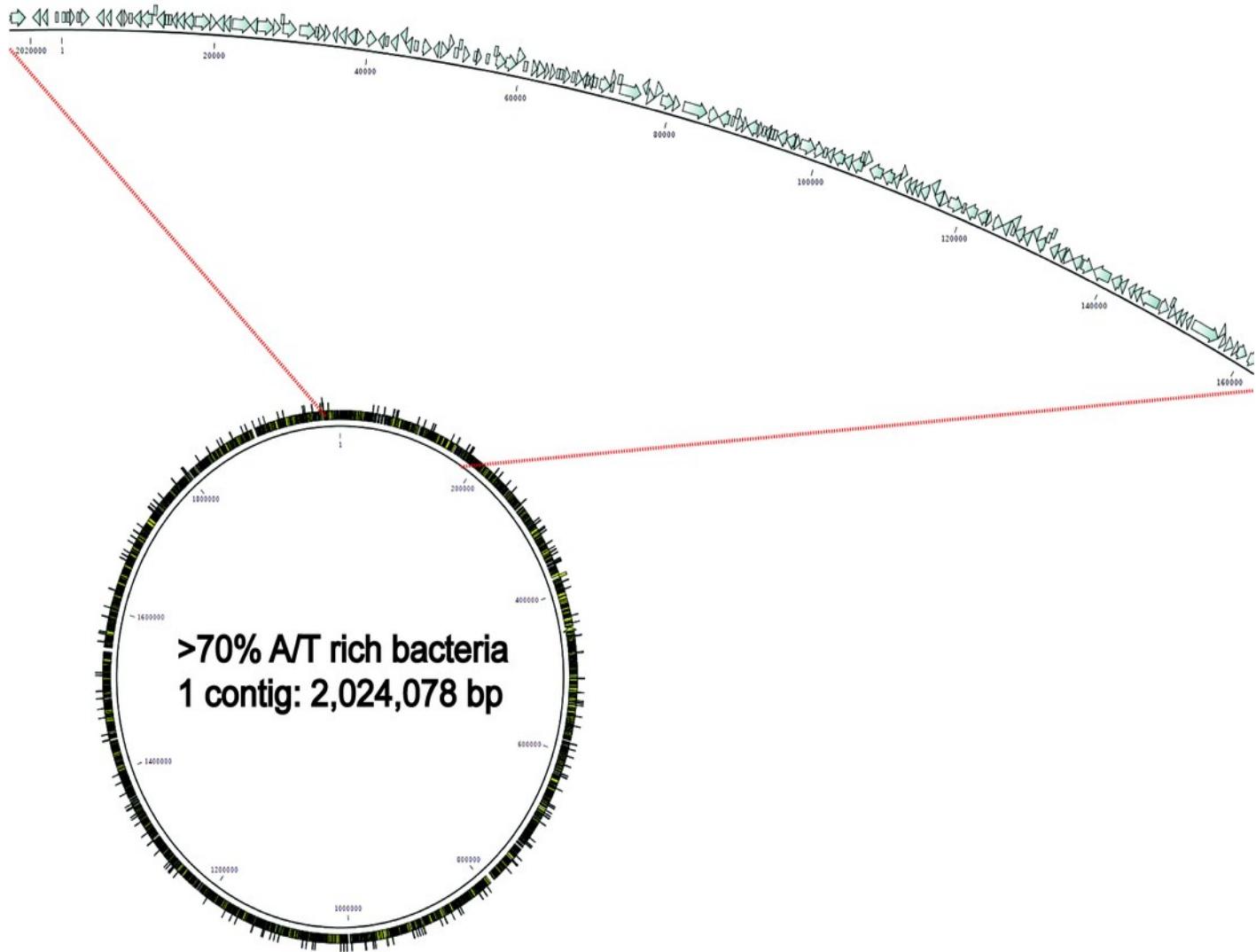
The main interface is divided into three sections: DESIGN JOB, MONITOR JOBS, and VIEW DATA. The DESIGN JOB section contains fields for Job Name (assembly), Comments, Protocols (set to RS_HGAP_Assembly.3), and Reference (None selected). The MONITOR JOBS section displays a table titled 'SMRT Cells Available' (Viewing 1 - 31 of 31) with columns for Sample, Version, User, Groups, Started, and Uri. The table lists numerous samples, mostly starting with 'Pb', with various versions (2.0.2, 2.1.0) and start times from 2014-02-20T19:28:20+0000 to 2014-05-08T11:08:49+0000. The VIEW DATA section displays a table titled 'SMRT Cells in Job' (Viewing 1 - 1 of 1) with a single entry for sample Pb33_1, version 2.0.2, with all groups and uri /home/pacbio/DATA/adam/Pb_33_F.

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_f2 rpt	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_f2 rpt	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/

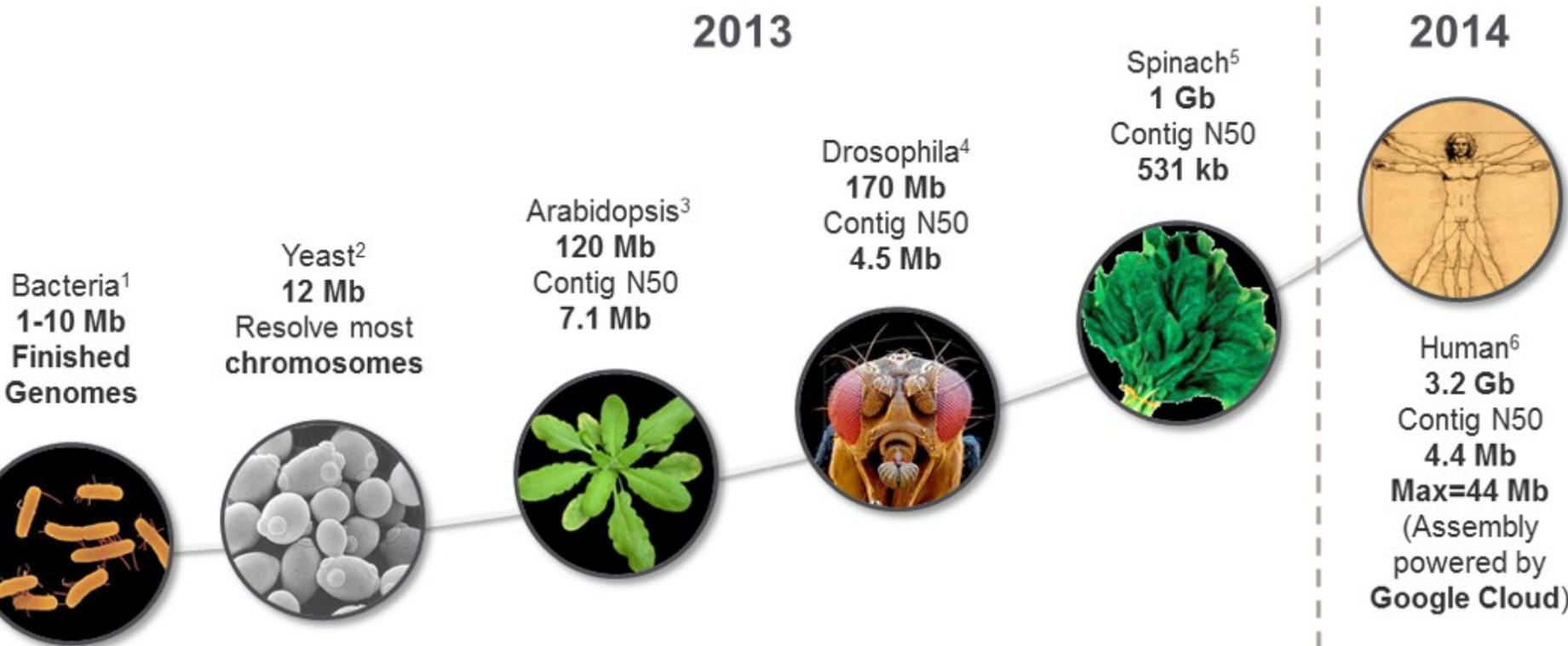
At the bottom, there are buttons for Show/Hide Columns, Group Rows, Export Table Data, Start, Save, Copy, and Cancel.

PacBio assembly, example result

- Example: Complete assembly of a bacterial genome

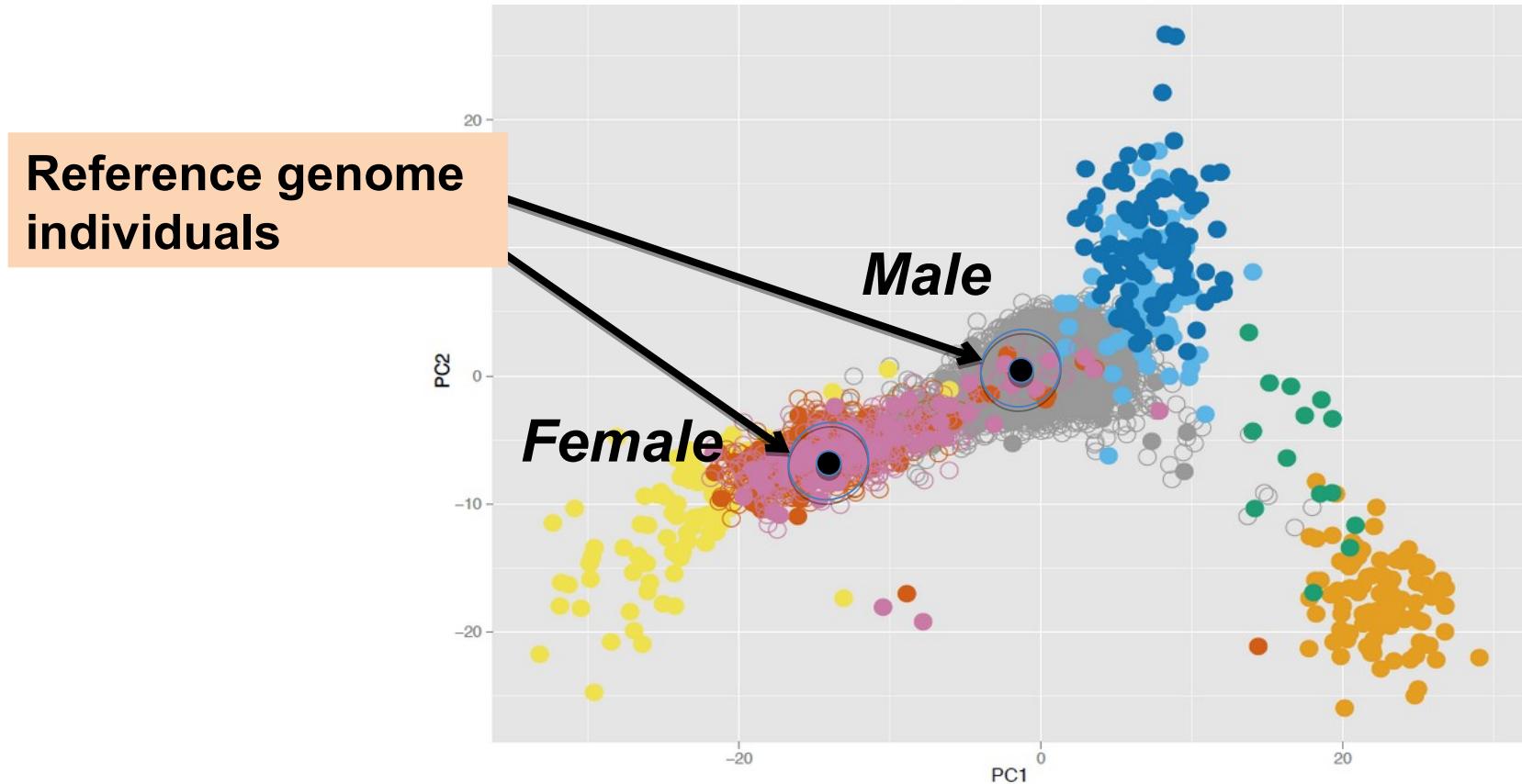


PacBio assembly – also for large genomes



De novo assembly of Swedish genomes

Long-read *de novo* sequencing of 2 individuals:



Available data for two reference individuals

Data type	Amount (per individual)
SMRT PacBio	75X coverage
BioNano	2 x 100X coverage
10X Chromium	50X coverage
Illumina WGS	30X coverage
Oxford Nanopore	30X coverage
MGI	??X coverage

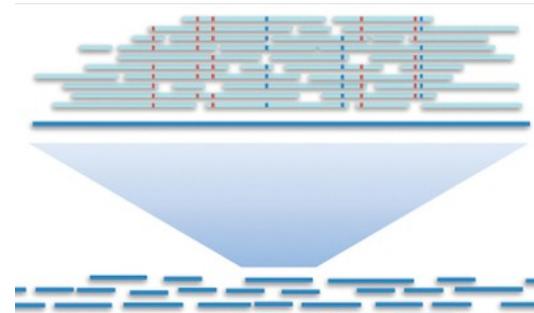
Aim: use all this data to create high-quality references!

De novo assembly of 75X PacBio data

Assembly (FALCON)



Error correction (2 x Quiver)

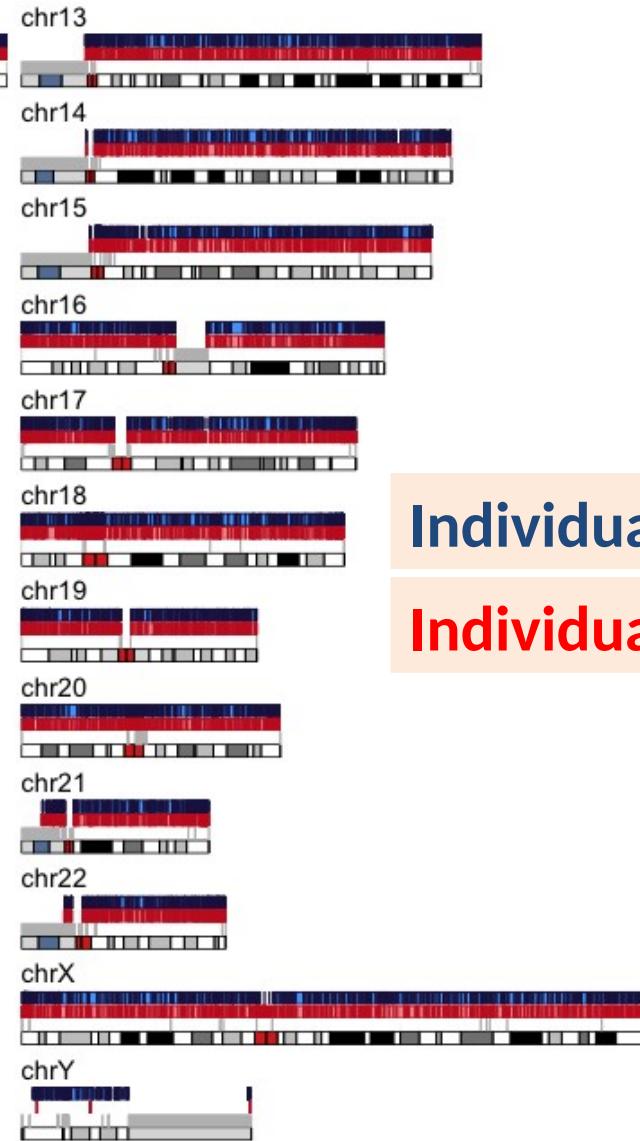


Analysis time: ~3 weeks/genome

	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

Aligning contigs to human reference

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



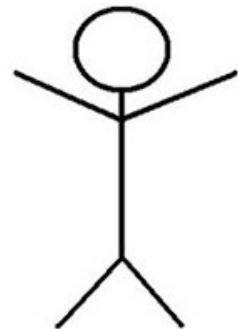
Individual #1

Individual #2

Discovery of ‘novel’ sequences

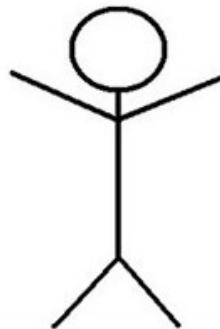
We find a lot of ‘novel’ sequence (NS), i.e. not matching hg38

Individual 1



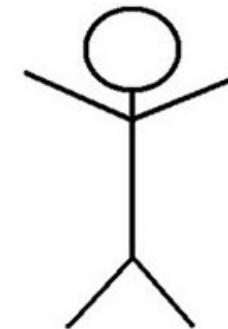
~14 Mb

Individual 2



~10 Mb

Chinese



~13 Mb



ARTICLE

Received 23 Nov 2015 | Accepted 26 May 2016 | Published 30 Jun 2016

DOI: 10.1038/ncomms13061 OPEN

Long-read sequencing and *de novo* assembly of a Chinese genome

Lingling Shi^{1,2,*}, Yunfei Guo^{4*}, Chengliang Dong⁴, John Huddleston⁵, Hui Yang⁴, Xiaolu Han⁶, Alisi Fu⁷, Quan Li⁸, Na Li⁷, Siyi Gong², Katherine E. Lintner⁸, Qiong Ding⁷, Zou Wang⁷, Jiang Hu⁹, Depeng Wang⁹, Feng Wang¹⁰, Lin Wang¹¹, Gholton J. Lyon¹², Yongtao Guan¹³, Yufeng Shen¹⁴, Oleg V. Egryakov¹³, James A. Knowles^{1,15}, Françoise Thibaud-Nissen¹⁶, Valerie Schneider¹⁶, Chack-Yung Yu¹⁷, Libing Zhou^{12,13}, Evan E. Eichler², Kwok-Fai So^{12,13,17,18} & Kai Wang¹³

Shi et al, 2016

Amount of novel sequences

~6 Mb of NS overlapping in all three genomes!

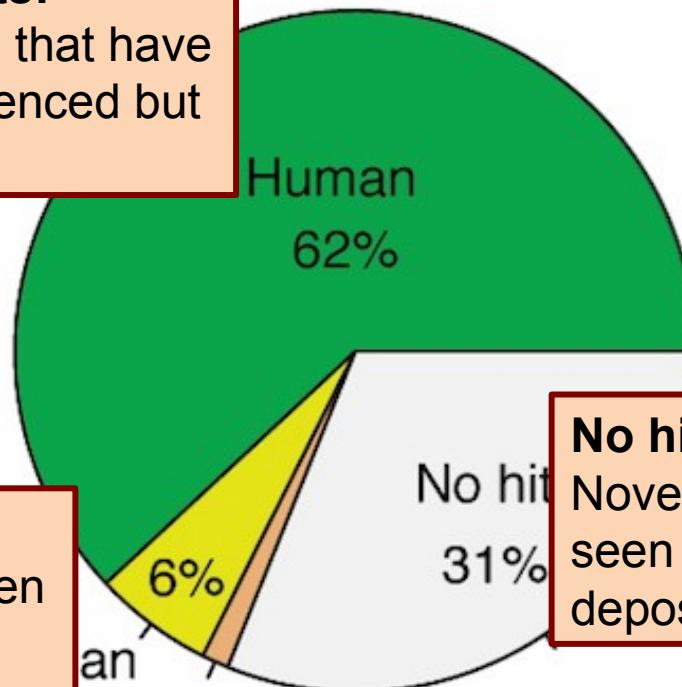
→ Likely missing sequence in hg38 reference

Have novel sequences been seen before?

Results of BLAST of NSs to NCBI database:

Human BLAST hits:

Mainly BAC clones that have been Sanger sequenced but not placed in hg38



Non-human primates:

Regions that have only been sequenced in primates before, but not in humans

No hit:

Novel sequences, never seen before (or at least not deposited in a database)

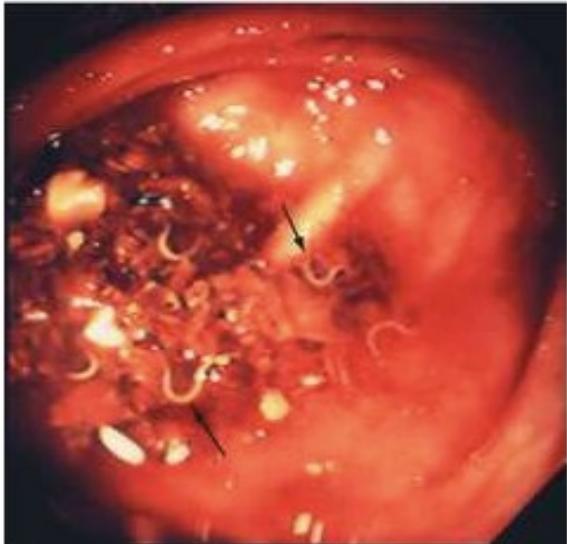
Other:

A few interesting other things, including a complete HPV35 genome. And some parasitic worms!

Why worm DNA in human genomes???

Worm DNA found in the two Swedish genomes and Chinese!

Enterobius vermicularis



Spirometra erinaceieuropaei



Dracunculus medinensis

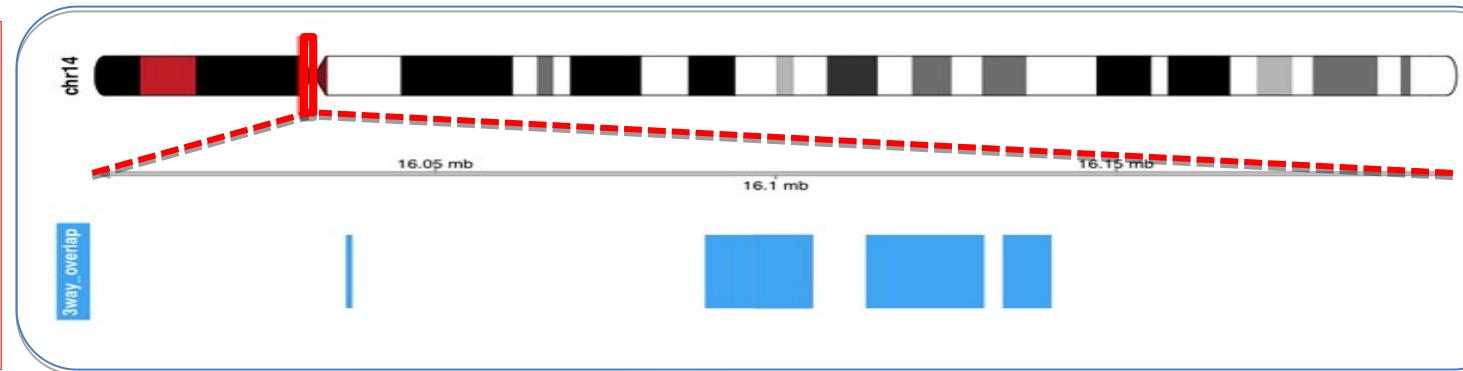


Our hypothesis: These “worm” sequences are in fact human, but erroneously annotated as worm DNA!

Novel sequences – examples

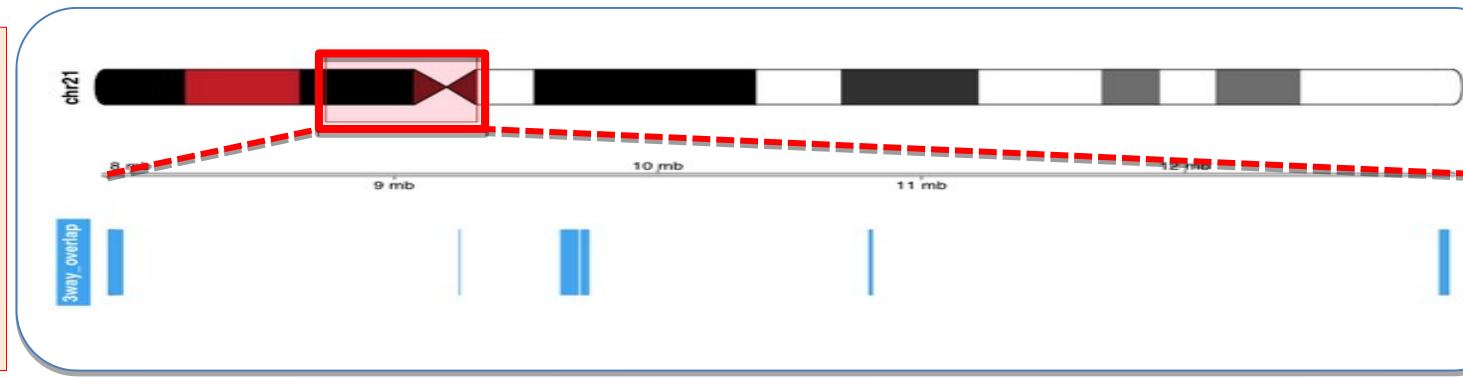
Ex 1: chr14

Missing in
reference
(3-way olap)



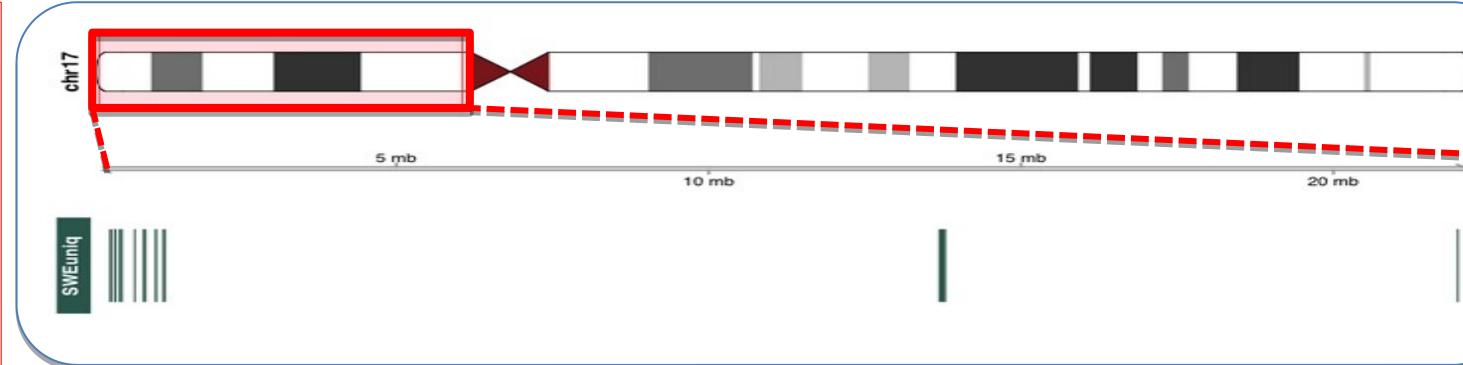
Ex 2: chr21

Missing in
reference
(3-way olap)



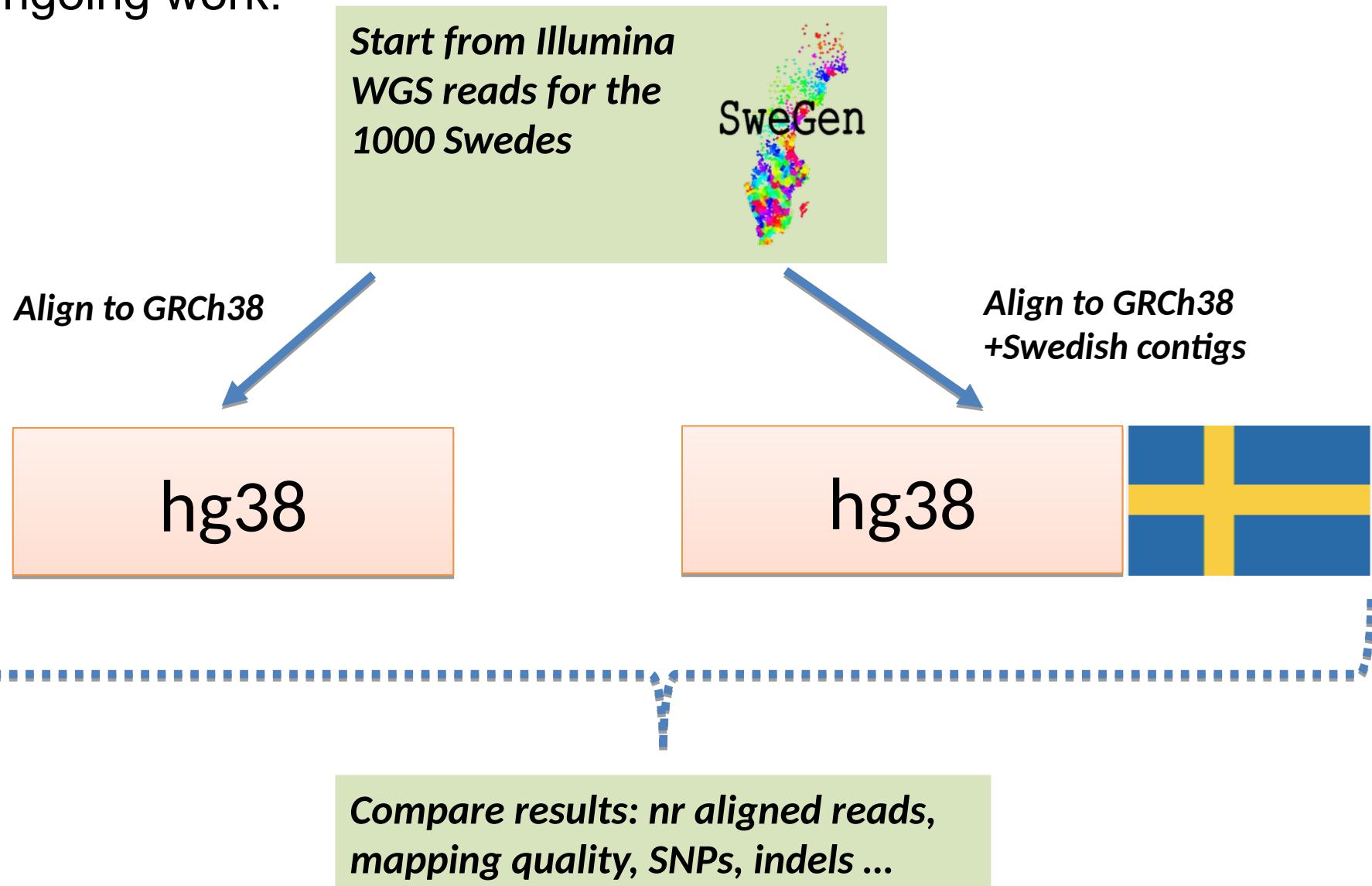
Ex 3: chr17

Population
specific?
(not in
Chinese)



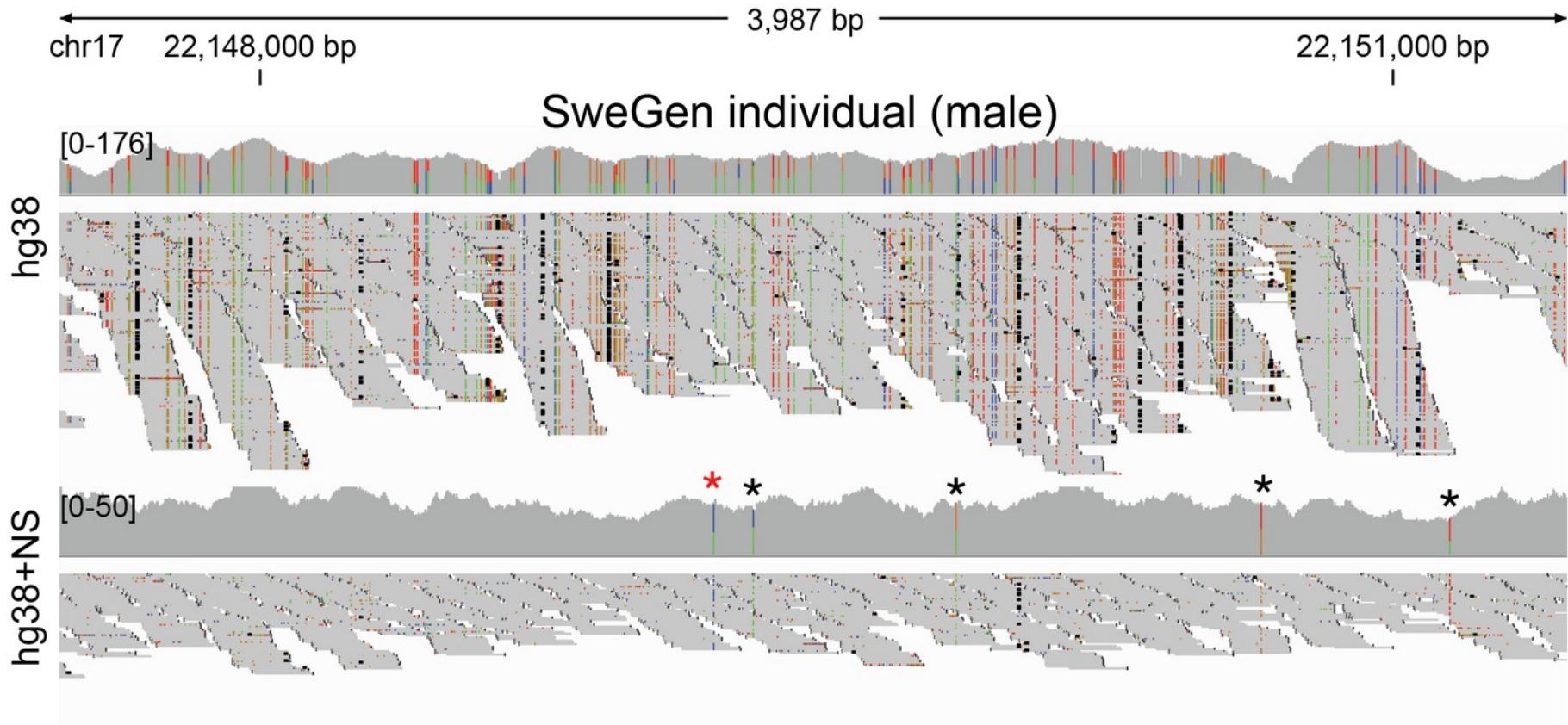
What is the benefit of a Swedish reference?

Ongoing work:



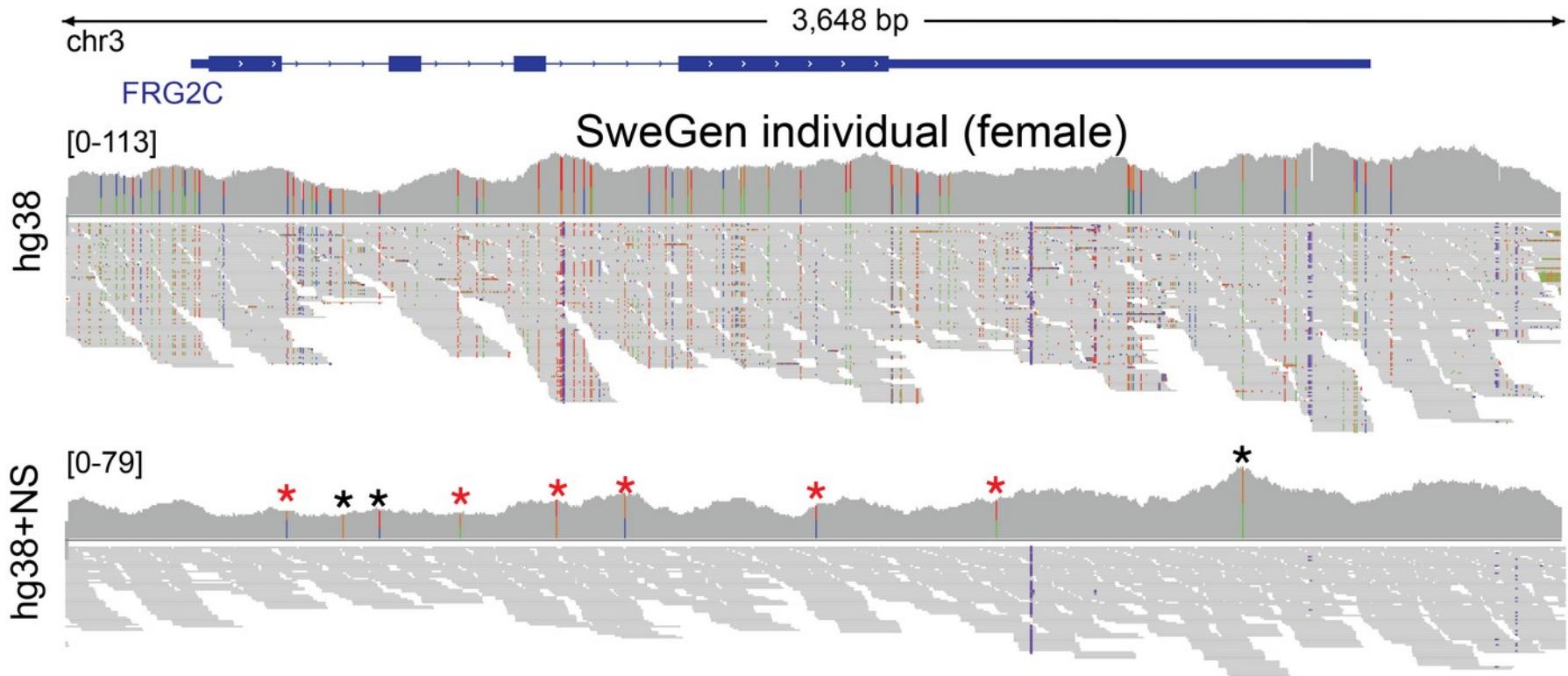
Cleaning up “ugly” regions in the genome

Some false positive SNPs seem to disappear!



Cleaning up “ugly” regions in the genome (2)

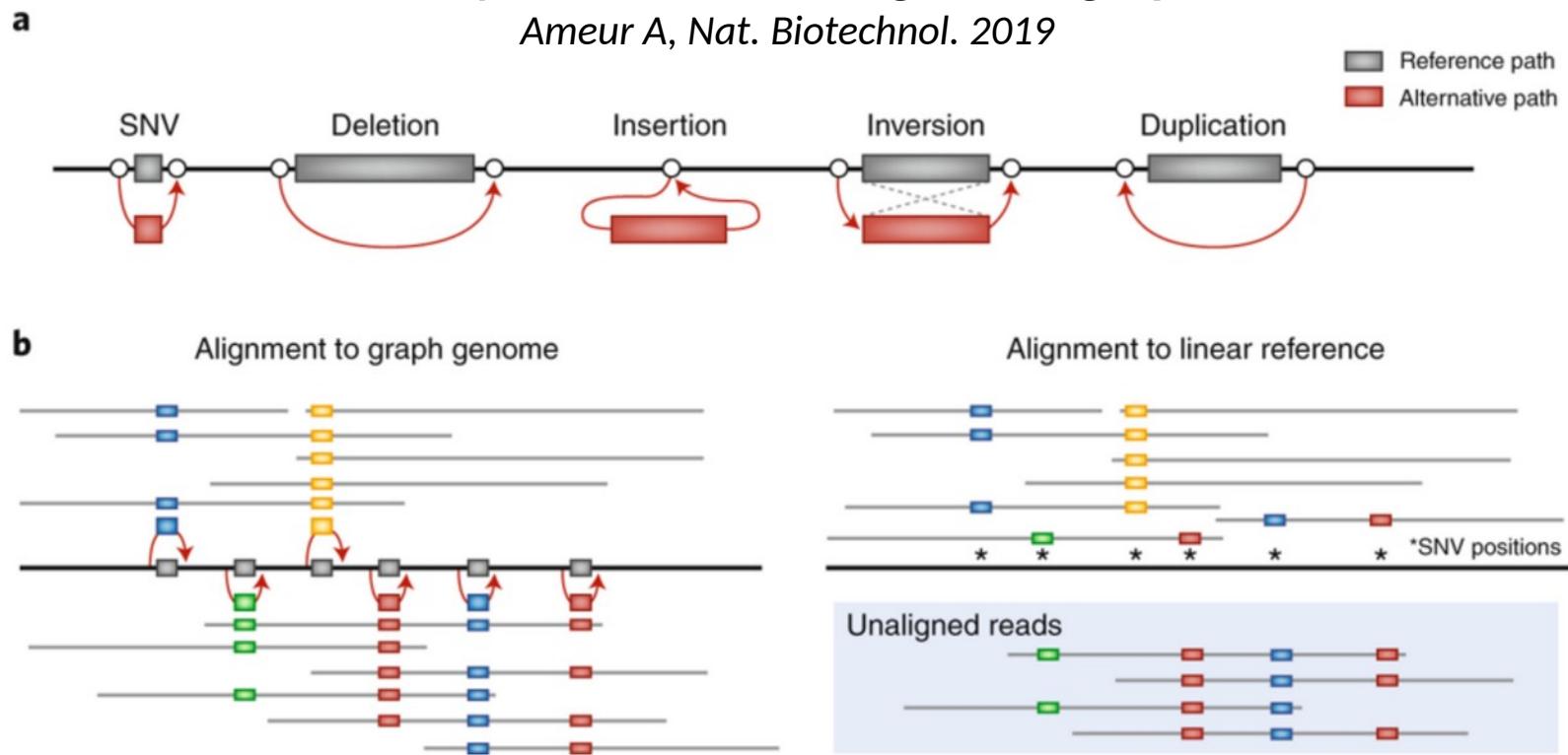
Another example, in protein coding region!



Can we represent the genome as a graph?

"Goodbye reference, hello genome graphs"

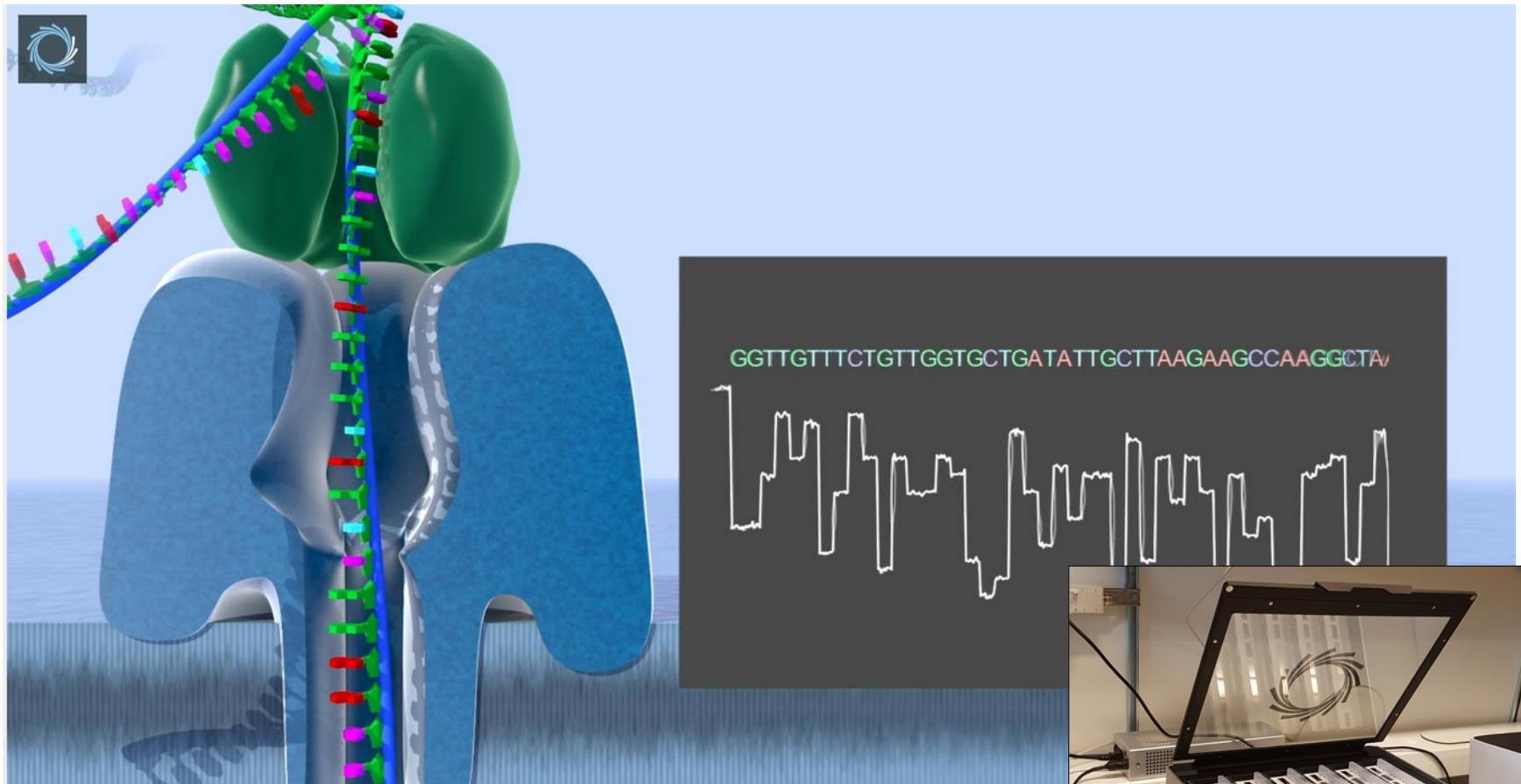
Ameur A, Nat. Biotechnol. 2019



Recent genome graph i:

- HISAT2/HISAT2-Genotype (Nat. Biotechnol. 2019)
- Vg (Nat. Biotechnol. 2018)
- Graph Genome Pipeline (Nat. Genet. 2019)

Human WGS on PromethION



- ~30X coverage from one flow cell
- Read lengths up to 1 Mega base!

SV calling - Nanopore vs PacBio

Sample	Insertion (INS)	Deletion (DEL)	Duplication (DUP)	Inversion (INV)
Nanopore, Swe1	8746	7769	185	133
PacBio, Swe1	12,441	9052	331	196
Nanopore, Swe2	8369	7820	147	128
PacBio, Swe2	12,218	9101	291	193

Long-Read sequencing in the clinic

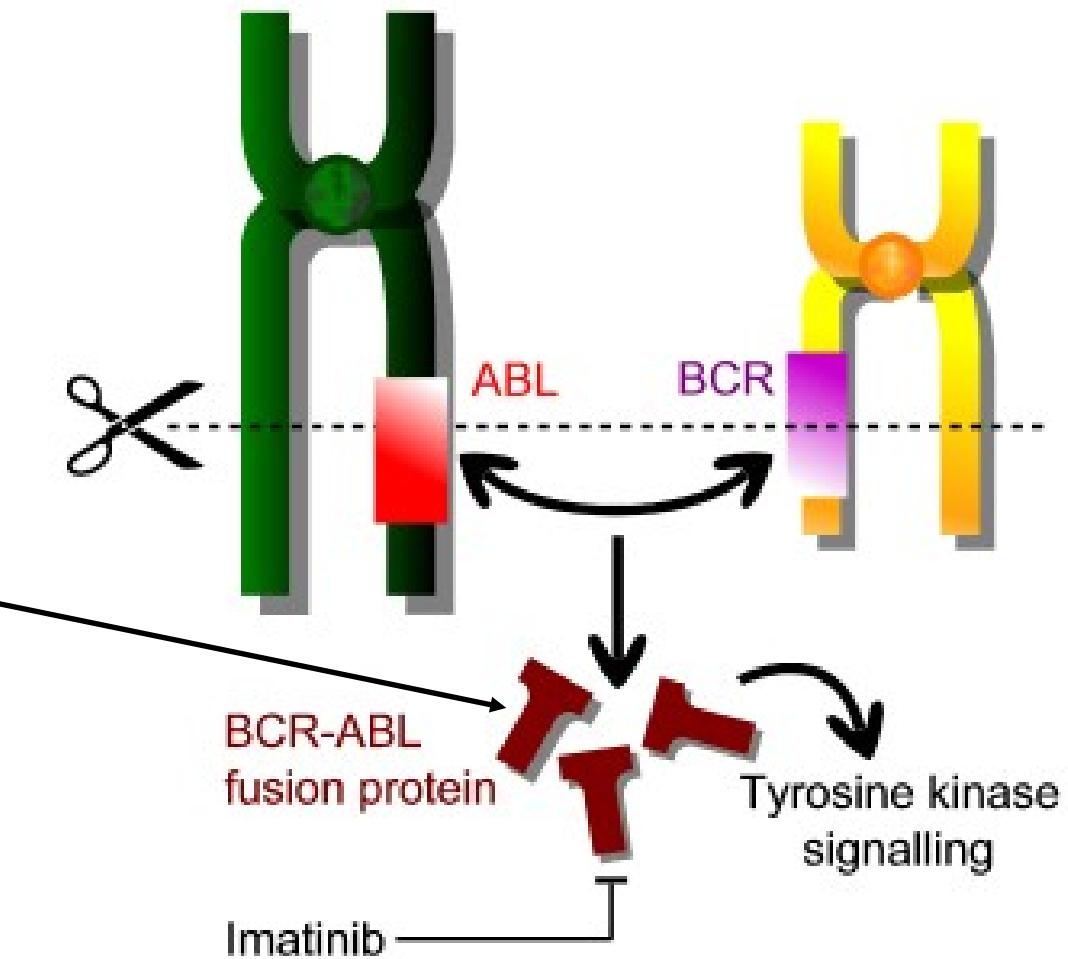


Project 1: Chronic Myeloid Leukemia

- BCR-ABL1 - a drug target

Chromosome 9

Chromosome 22



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

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 Höijer¹, Nicola Cahill¹, Ulla Olsson-Strömberg² and Monica Hermanson¹

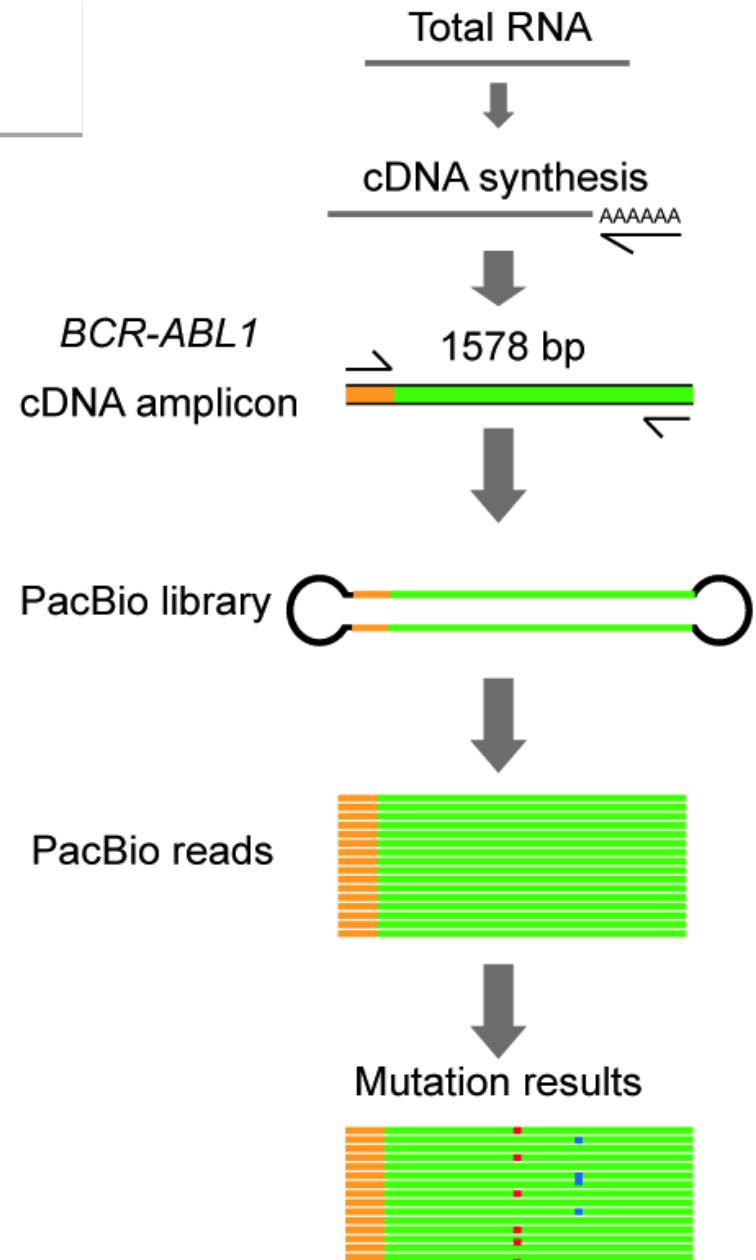
Abstract

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.

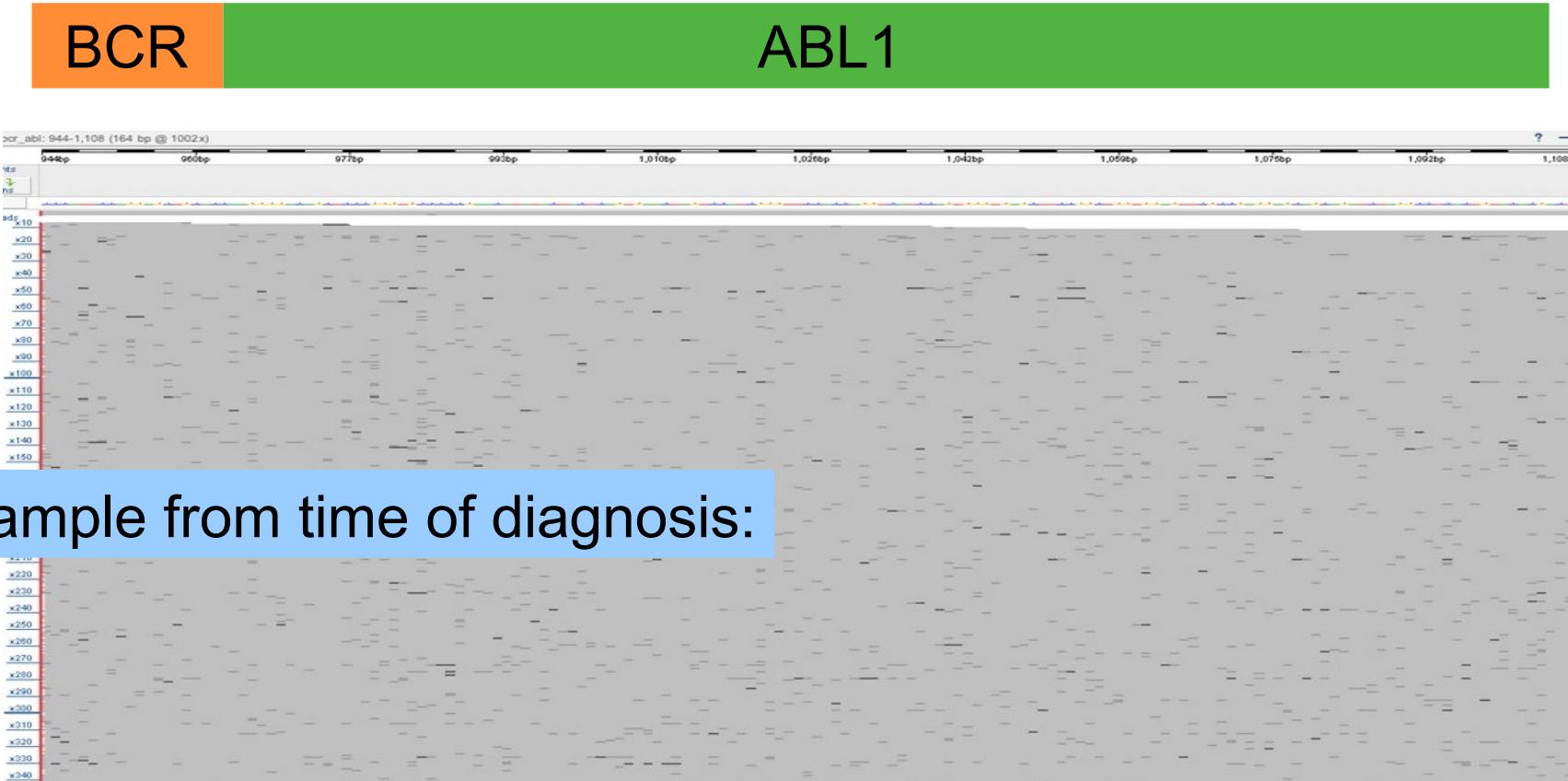
Results: Over 10,000 full-length *BCR-ABL1* sequences were obtained for all samples studied. Through the serial dilution analysis, mutations in CML patient samples could be detected down to a level of at least 1%. Notably, the assay was determined to be sufficiently sensitive even in patients harboring 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.

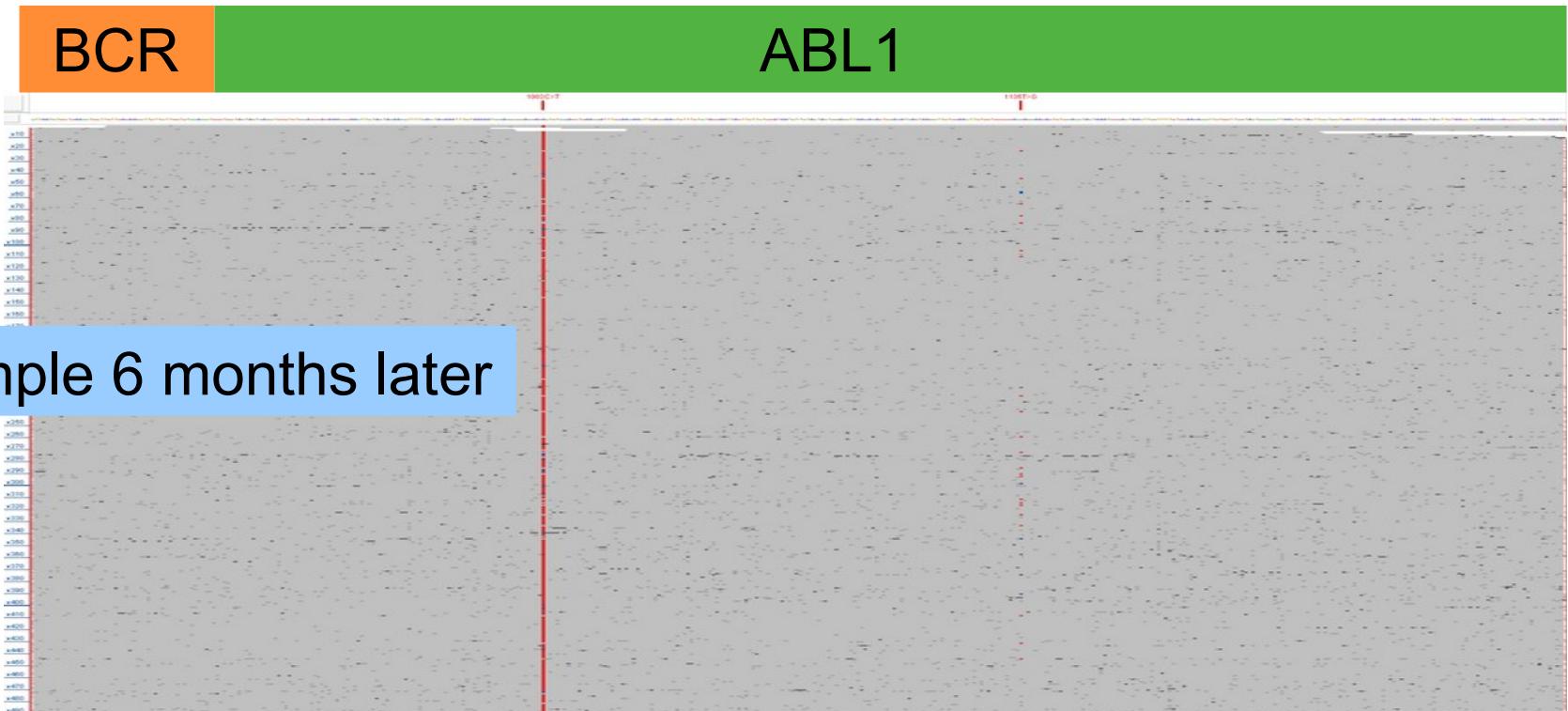


BCR-ABL1 mutations at diagnosis

PacBio sequencing generates ~10 000X coverage!

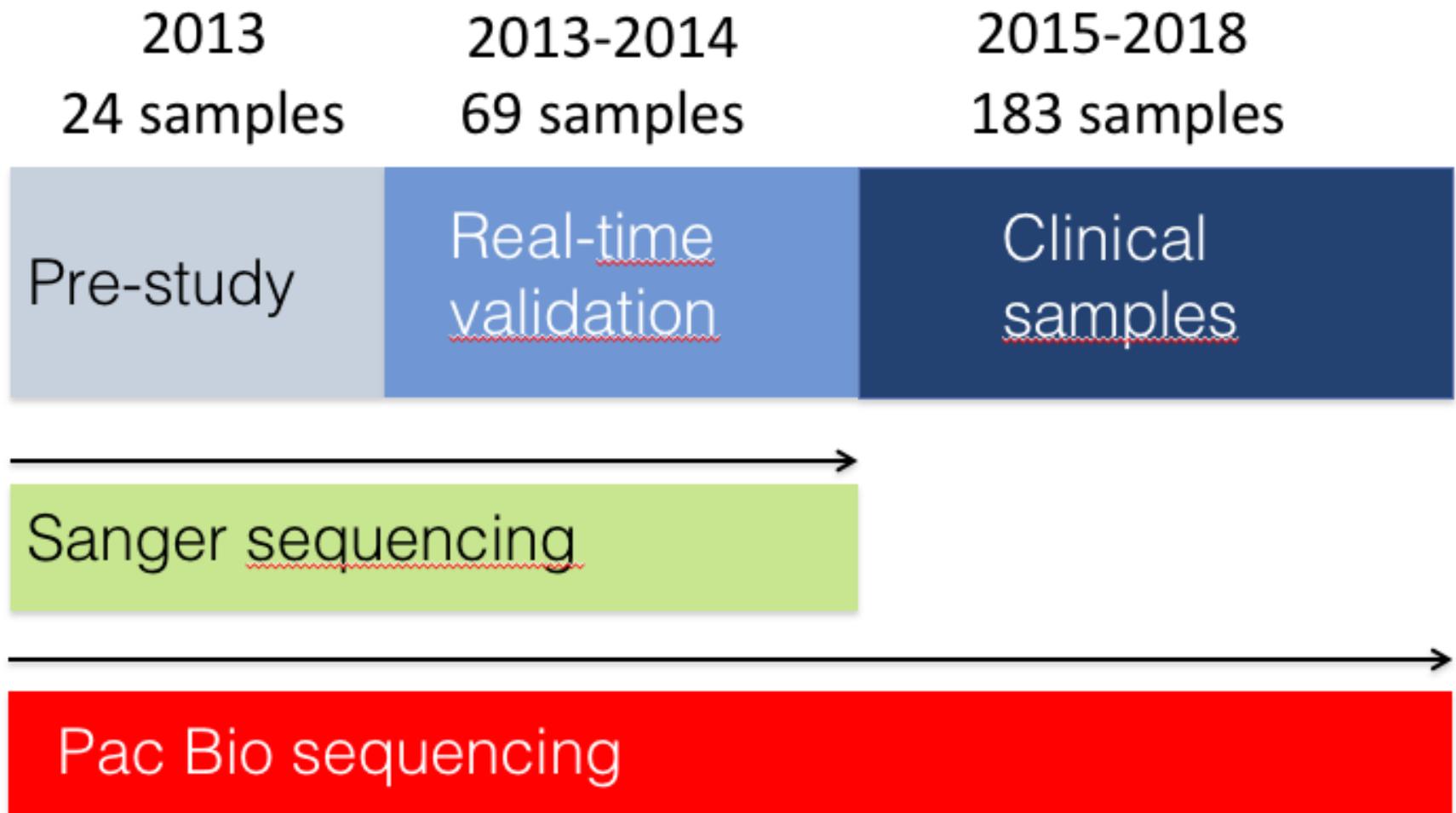


BCR-ABL1 mutations in follow-up sample



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

PacBio BCR-ABL1 timeline



Clinical Diagnosis of BCR-ABL1 mutations

Clinical Genetics



- Collection of samples
- Seq library preparation

Sequencing Facility



- SMRT sequencing
- Automated data analysis

IT developers



- Web server for results

- Ongoing routine service, up to 4 samples/week
- Over 250 samples run so far
- 100% consistency with Sanger results during pre-study and real time validation period

CML results viewed in our system

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)	R12126	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_																						

101 Sample 102 103 [New Search](#)

Sample ID	Run ID	Date
R12095	cba_012_4	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]GGAGTACCTGGAGAAAGAAAA	16134	19065	3	0.542	positive
Q252H	CACAAGCTGGGGGGGGCCAG[C/C]TACGGGGAGGTGTACCAAGGG	12052	9920	8	0.451	positive
K262N	GTGTACGAGGGCGTGAGAA[G/T]AAATACAGCCTGACGGTGCG	25597	6996	16	0.215	positive
M244V	TGGAACCCACGGACATCACC[A/G]TGAAGCACAAGCTGGCGGG	32779	32	2	0.001	negative
K247R	GGACATCACCATGAAGCACA[A/G]GCTGGCGGGGCCAGTACG	27076	32	9	0.001	negative

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

Position in bcr-14-abl1

Position	Frequency	Reads
0-500	49.9 %	9268
500-1000	23.8 %	4418
1000-1500	17.4 %	3245
1500-2000	8.69 %	1613

Frequency Reads

1000 1500 2000

Position in bcr-14-abl1

1000 1500 2000

Position in bcr-14-abl1

Example, patient with CML

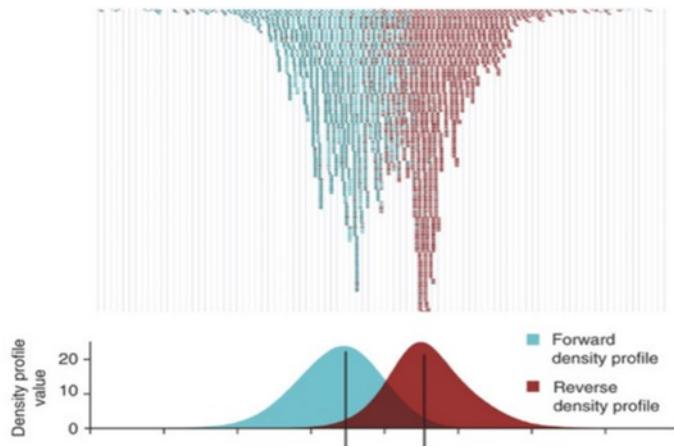
M244V detected at 2,1%



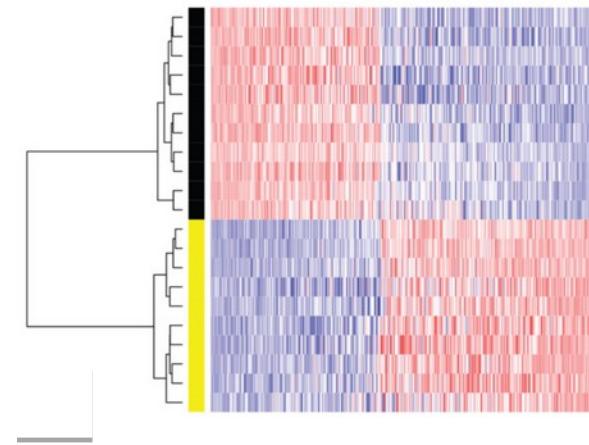
Imatinib → Nilotinib

Many topics I have not covered...

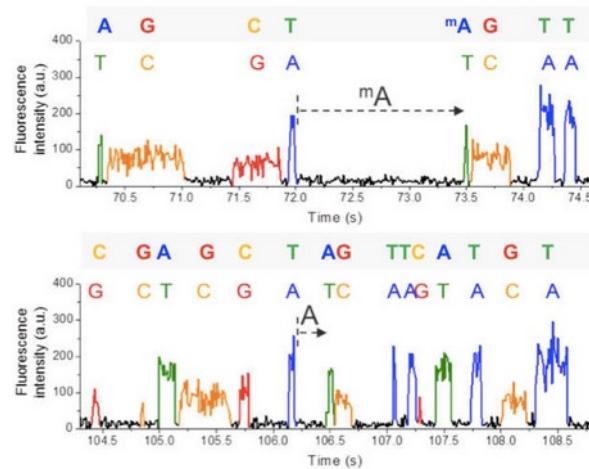
ChIP-seq analysis



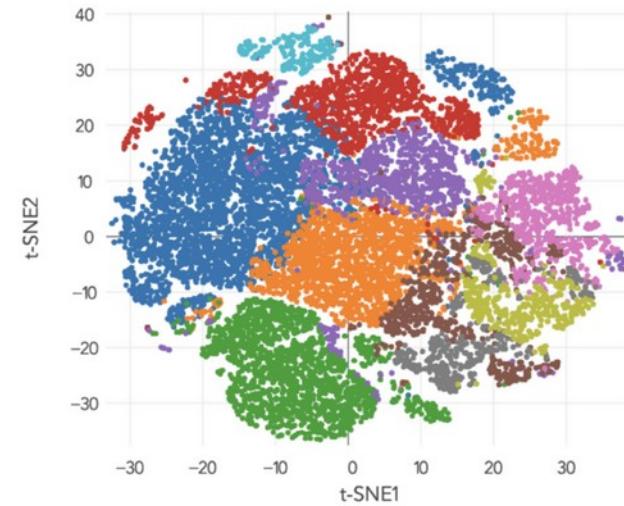
RNA-Seq



DNA base modifications



Single cell RNA-Seq



Simply too much to talk about in just one lecture...

Long-Read Sequencing

October 12-14, Uppsala

Adam Phillippy



Program Abst

er your research

Kathy Below



Practical inform

sequencing technolo

Karen Miga



In recent years, long-read DNA sequencing technologies has replaced standard sequencing methods. In addition to producing high quality *de novo* genome assembly, long-read sequencing can be used to study epigenetic signals, and much more. The adaptation of long-read sequencing is sweeping through several areas of the life sciences including agricultural, environmental, and medical research.

Join us in Uppsala in October to catch up with the latest developments in long-read sequencing technologies and their applications, get inspired by peers presenting their research, and enjoy discussions with leading experts and company representatives.

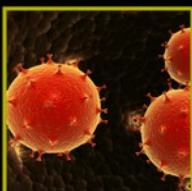
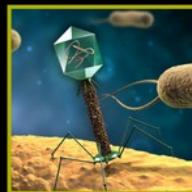
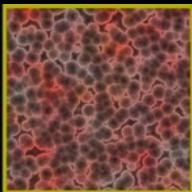
Registration will open in February 2020!



#LRUA20

What we sequence at NGI /

SciLifeLab



- 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...
- 