NATIONAL GENOMICS INFRASTRUCTURE

SMRT Leiden, May 3, 2017

Resolving challenging regions in the human genome

Adam Ameur, PhD

IGP, Science for Life Laboratory, Uppsala University

National Genomics Infrastructure, Sweden

Enabler for Life Sciences

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

National Genomics Infrastructure

NGI staff: 75 FTE, including Head of facility, lab research engineers, bioinformaticians, IT-experts, project coordinators

NGI production during 2016

- 1000 projects
- 34 000 samples

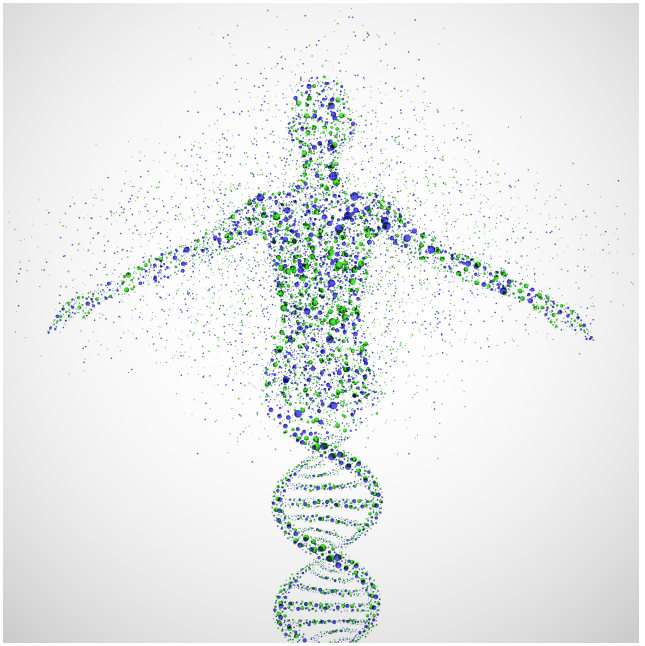
NGI Sequencing equipment

DNA sequencing at all scales

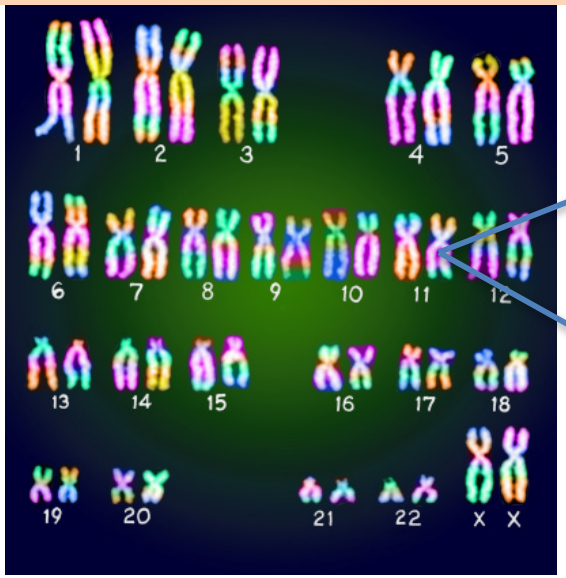
• 10 Illumina HiSeqX (Xten)	Whole genome sequencing & bisulfite seq
• 10 Illumina HiSeq2500/2000	Large projects – WGS, WES, Transcriptome
• 3 Illumina MiSeq	Microbial genomes, Targeted seq, Biodiversity
• 1 Illumina NextSeq	Technology development
• 1 5SXL (Thermo Fisher)	Microbial genome, Targeted seq, Biodiversity
• 5 IonProton (Thermo Fisher)	Large projects – WES, Transcriptome
• 2 Pacific Biosciences RSII	*De Novo* sequencing, long amplicon, Base modifications
• 1 Pacific Biosciences Sequel	*De Novo* sequencing, long amplicon, Iso-seq
• 1 Sanger ABI3730	Variant validation
• 1 Oxford Nanopore MinIon	Technology development
• 2 10X Genomics Chromium	Haplotype phasing, RNA-seq
• 1 BioNano Irys system	*De Novo* sequencing, structural variation

Why SMRT sequencing of human DNA?

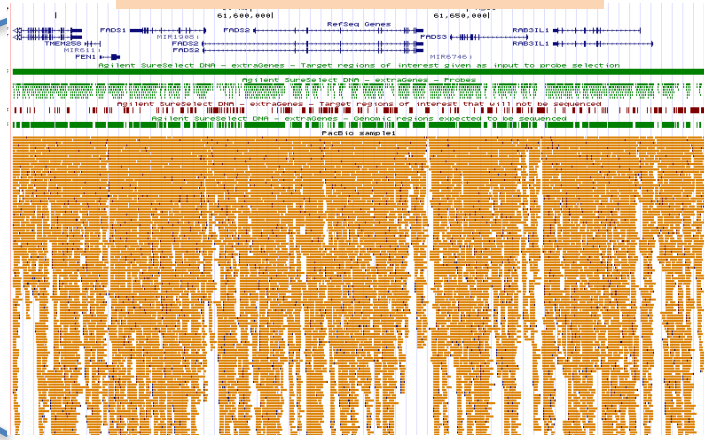
- Resolve complex regions (e.g. repeats)
- Detect low frequency mutations
- Get phasing information for variants
- Assemble complete human genomes
- Study patients in clinical setting



Whole Genome Sequencing



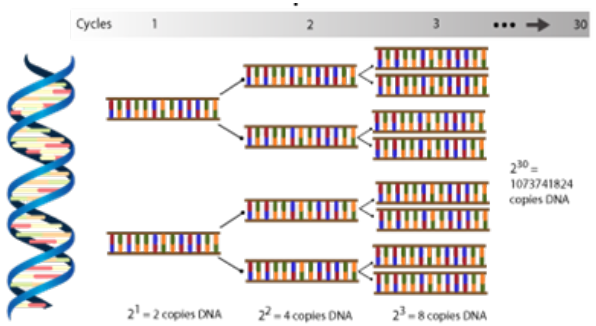
Targeted Sequencing



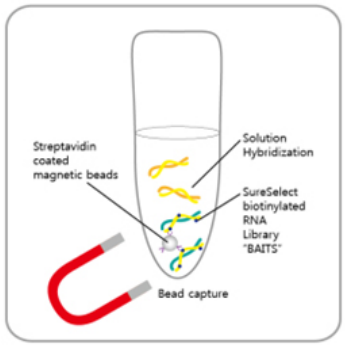
Targeted SMRT sequencing

Many options available… but no perfect solution

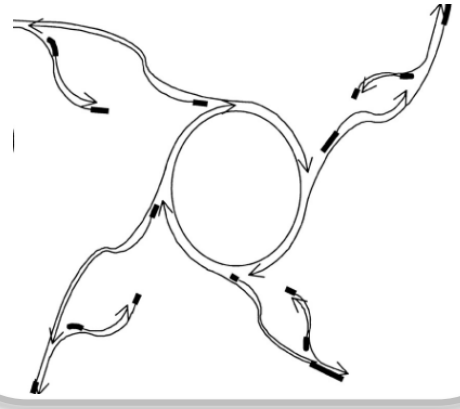
LR-PCR



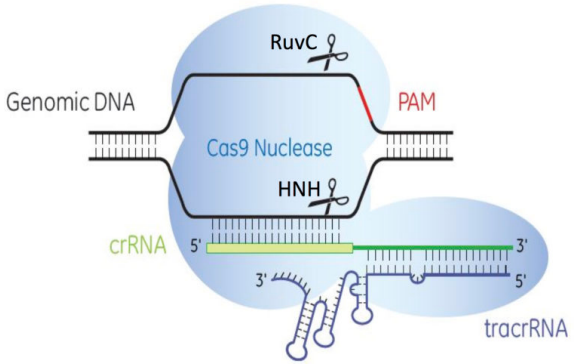
Hybridization



RCA

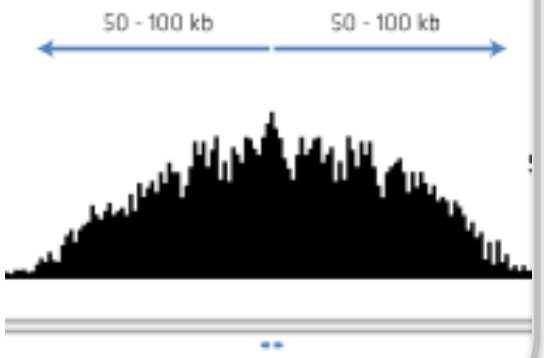


CRISPR/Cas9



Droplet amp.

TLA

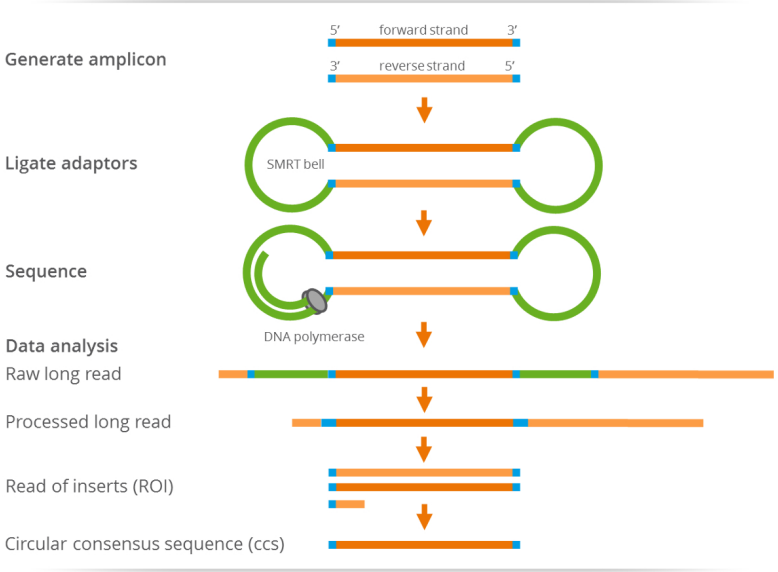


We use different methods depending on the project!

Long range PCR and SMRT sequencing

Simple and efficient method

1. Amplify region of interest (up to 15kb)
2. Do SMRT sequencing
3. Generate CCS reads (or subreads)
4. Data analysis



We use LR-PCR for many human sequencing projects:

- Both research & Clinical diagnostics

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]

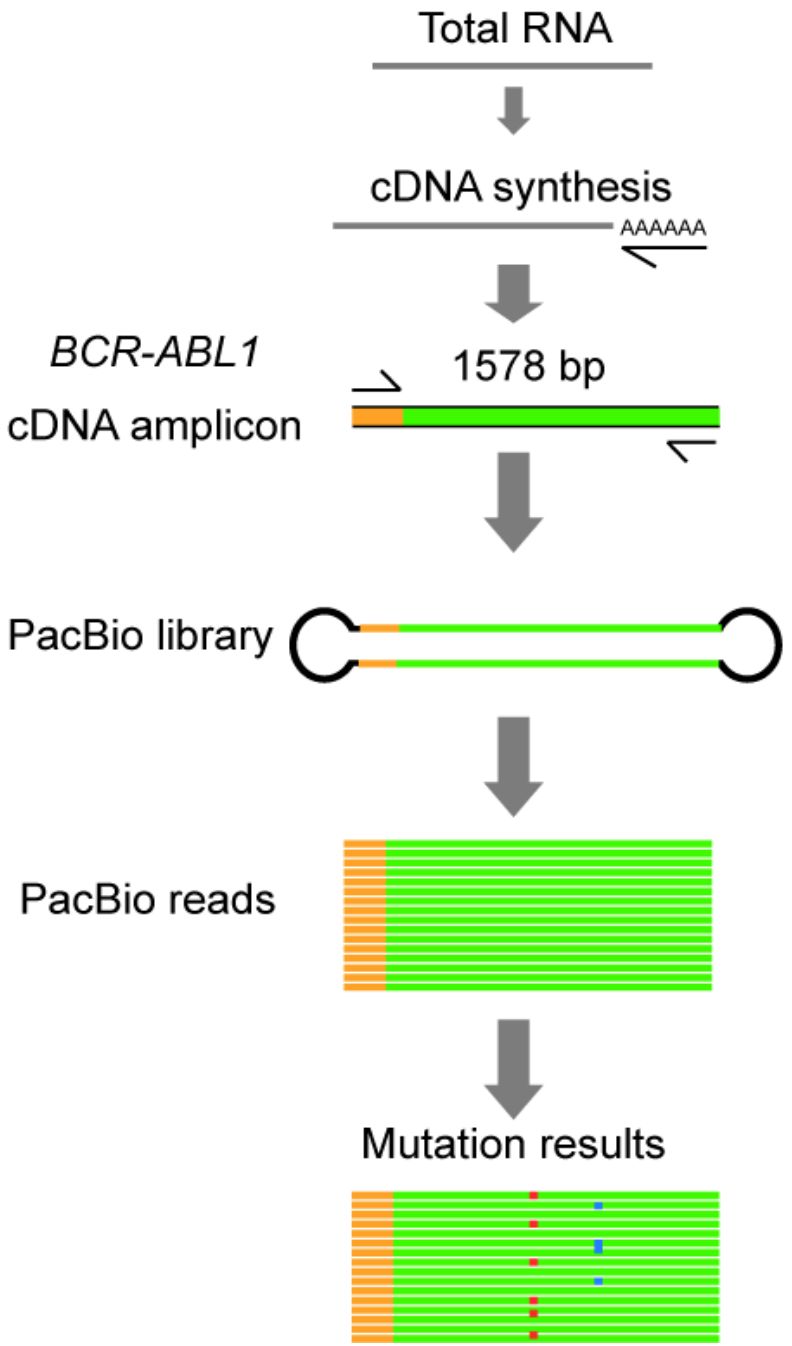
Abstract

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.

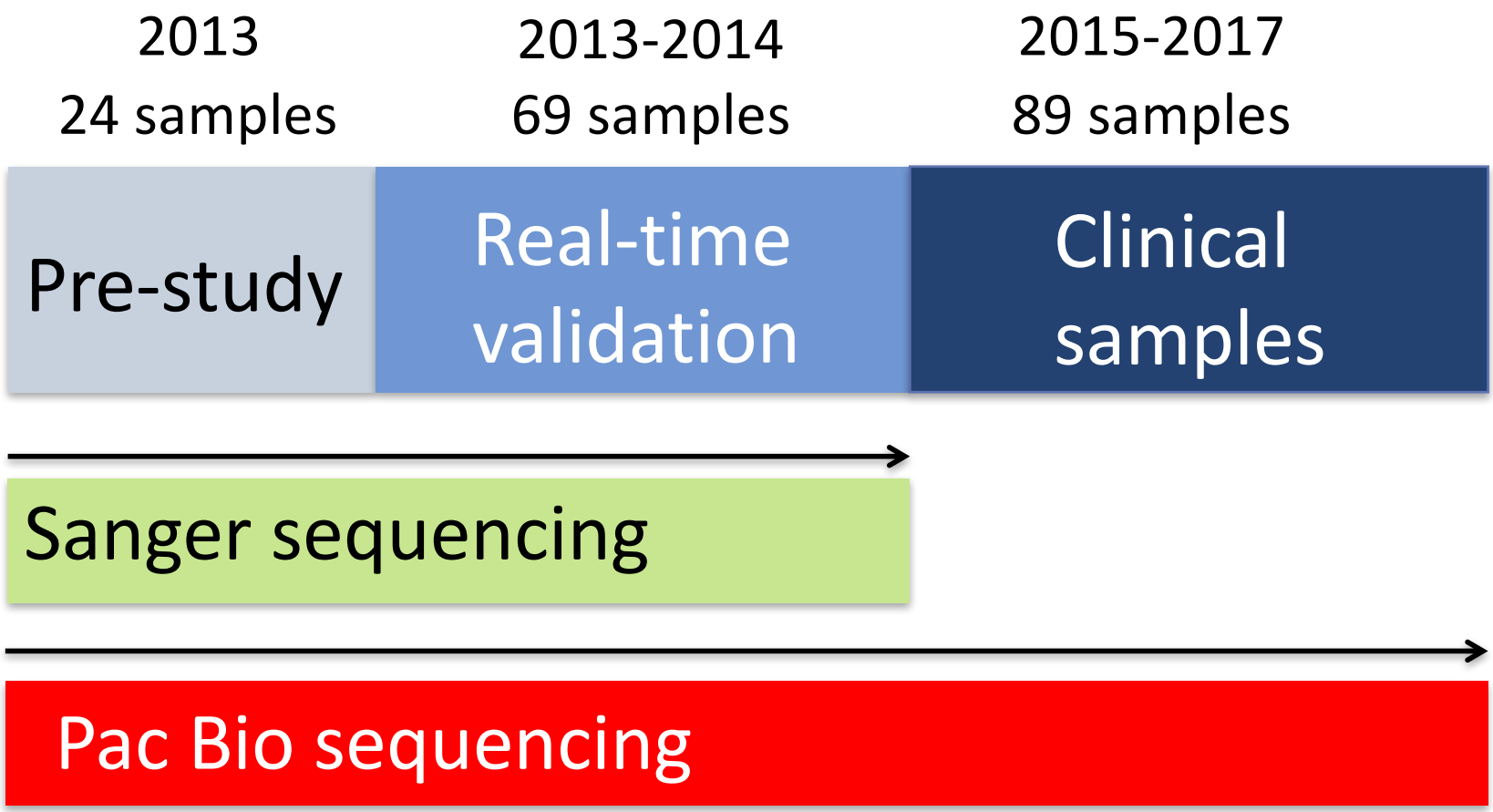
Methods: To detect mutations in the *BCR-ABL1* transcript we developed an assay based on the Pacific Biosciences (PacBio) sequencing technology, which allows for single-molecule long-read sequencing of *BCR-ABL1* fusion transcript molecules. Samples from six patients with poor response to therapy were 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.



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 180 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_																						

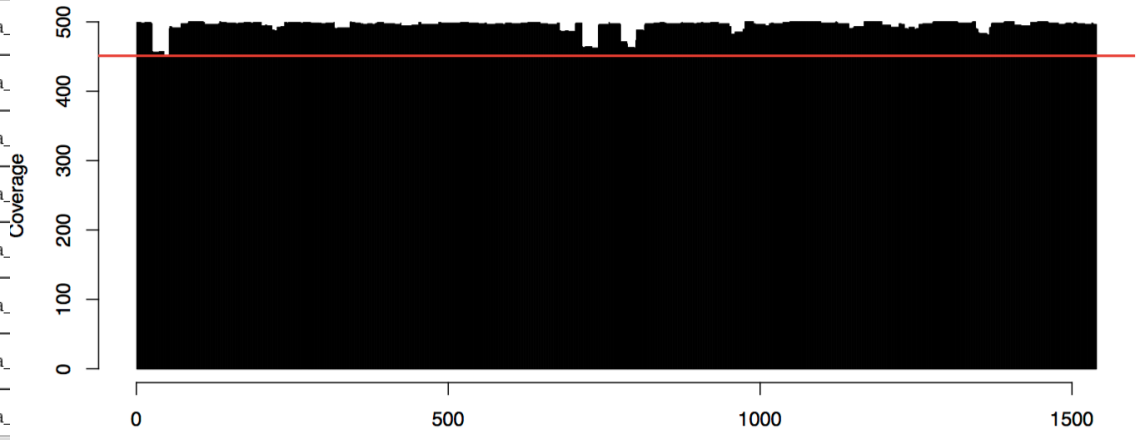
101 **Sample 102** 103 New Search

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

Results Sequence **Downloads:** Coverage Clonal txt Clonal pdf Log

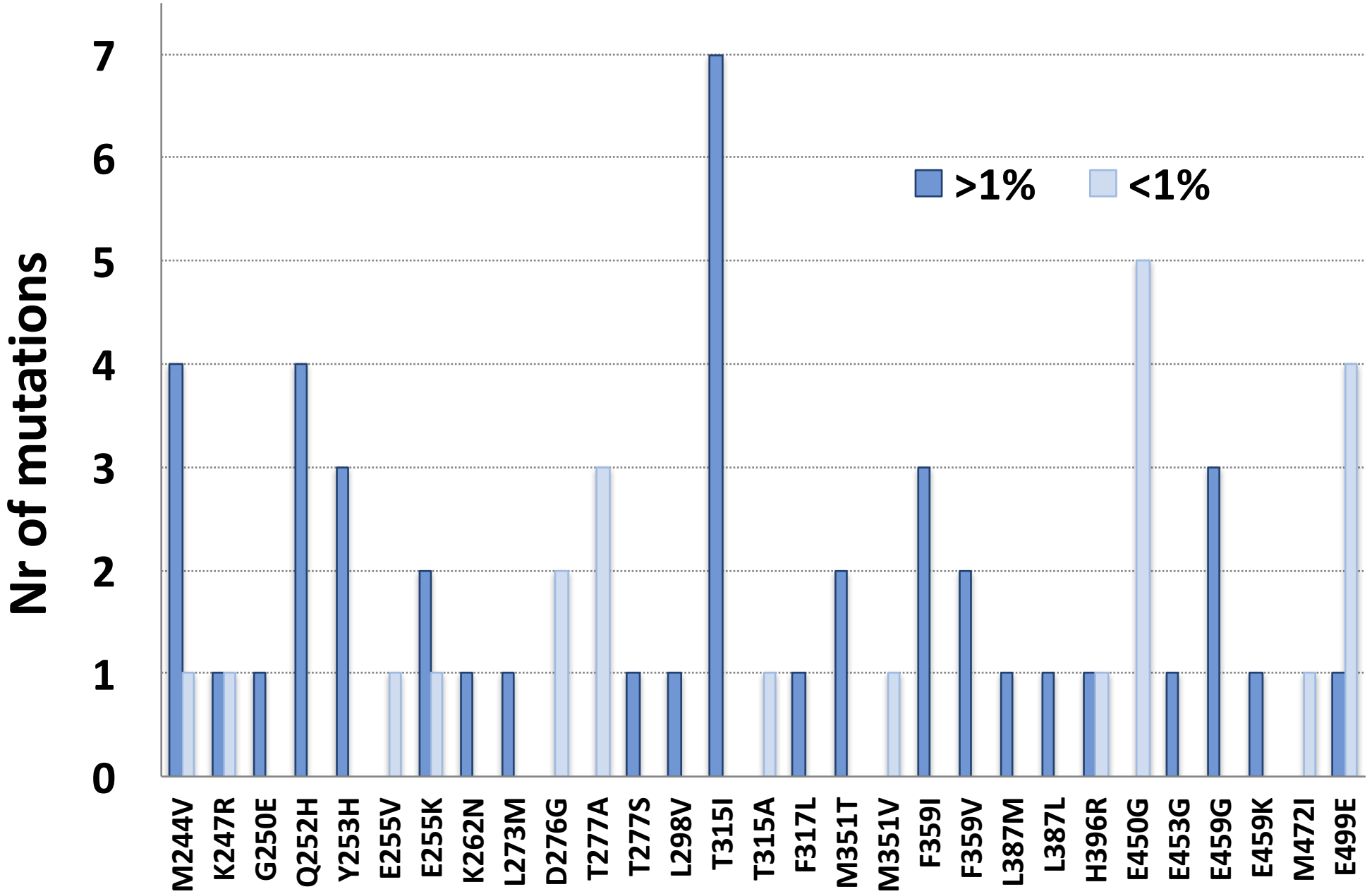
mutation	sequence	wt_reads	mut_reads	other_reads	freq	detection
M351T	CACTCAGATCTCGTCAGCCA[T/C]GGAGTACCTGGAGAAGAAAA	16134	19065	3	0.542	positive
Q252H	CACAAGCTGGGCGGGGGCCA[G/C]TACGGGGAGGTGTACCAGGG	12052	9920	8	0.451	positive
K262N	GTGTACGAGGGCGTGTGGAA[G/T]AAATACAGCCTGACGGTGGC	25597	6996	16	0.215	positive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

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



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

Mutations detected in 89 clinical samples



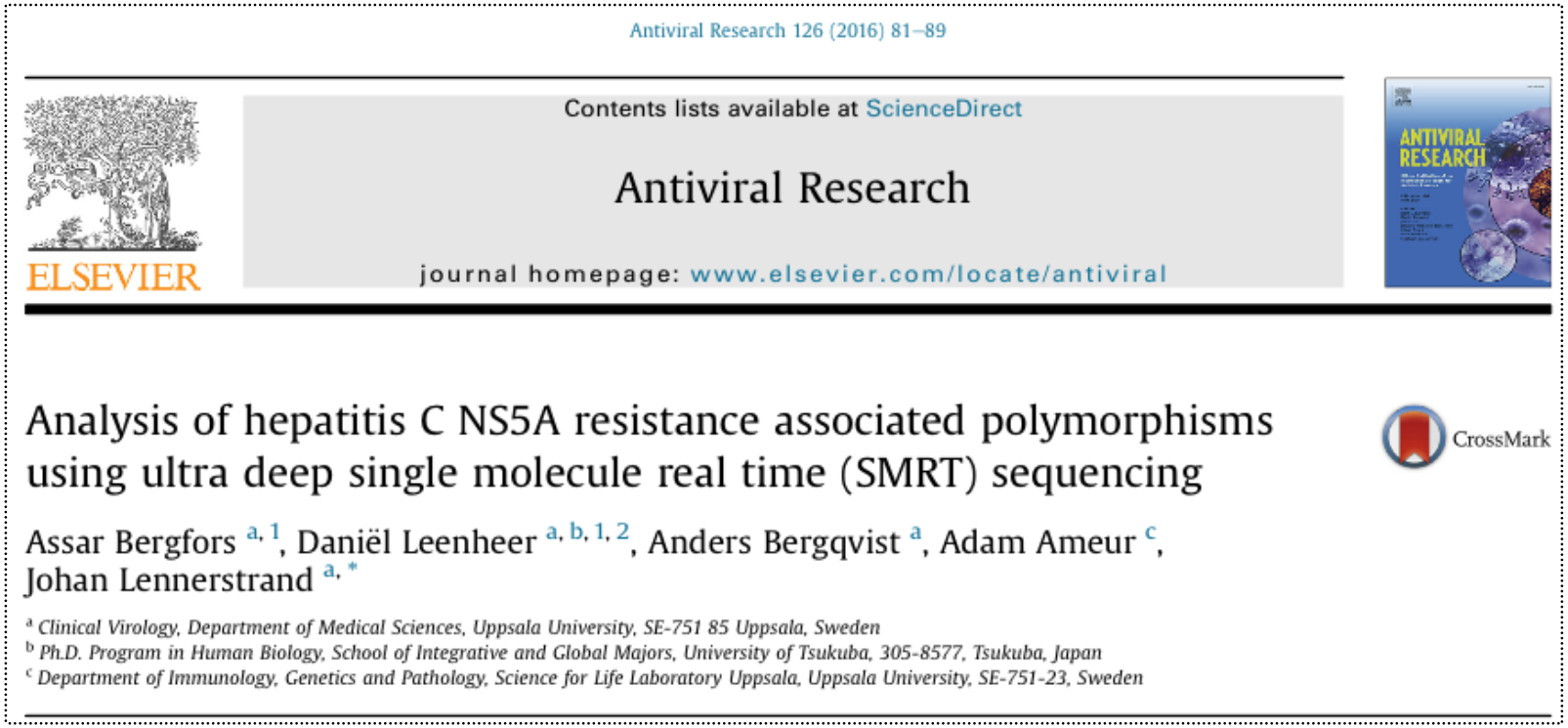
Advantages of SMRT sequencing for CML

- Simple and rapid workflow
- Sensitivity higher than Sanger
- No nested PCR bias: true mutation frequency estimates
- Detection of compound and independent mutations
- Detection of transcript isoforms
- Easy and sensitive bioinformatic analysis

Lucia Cavelier, PhD
Clinical Genetics,
Uppsala, Sweden

Examples of other clinical targets - HCV

Hepatitis C virus resistance monitoring during treatment:



Antiviral Research 126 (2016) 81–89

Contents lists available at ScienceDirect

Antiviral Research

journal homepage: www.elsevier.com/locate/antiviral

Analysis of hepatitis C NS5A resistance associated polymorphisms using ultra deep single molecule real time (SMRT) sequencing

Assar Bergfors [a, 1], Daniël Leenheer [a, b, 1, 2], Anders Bergqvist [a], Adam Ameur [c], Johan Lennerstrand [a, *]

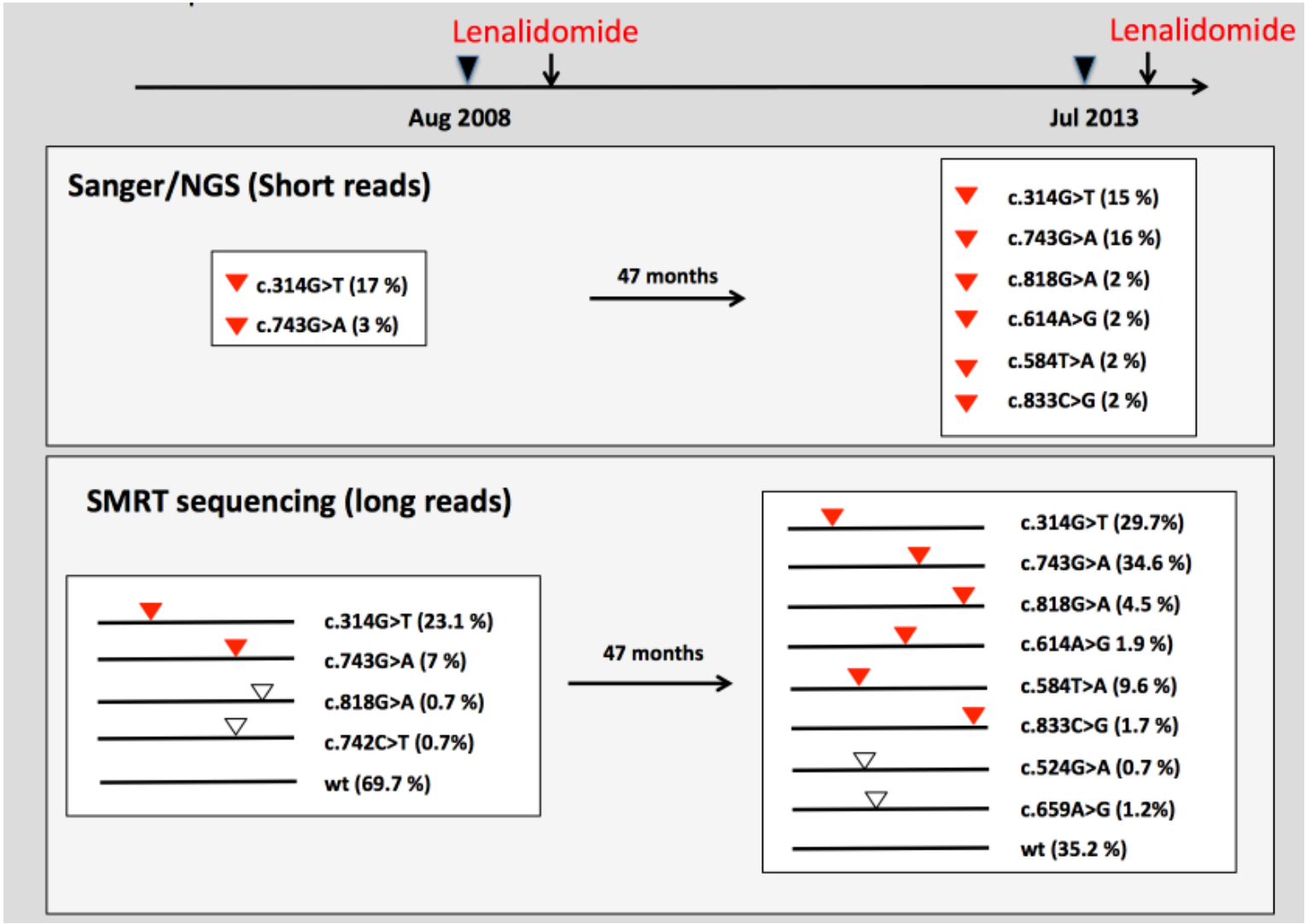
[a] Clinical Virology, Department of Medical Sciences, Uppsala University, SE-751 85 Uppsala, Sweden
[b] Ph.D. Program in Human Biology, School of Integrative and Global Majors, University of Tsukuba, 305-8577, Tsukuba, Japan
[c] Department of Immunology, Genetics and Pathology, Science for Life Laboratory Uppsala, Uppsala University, SE-751-23, Sweden

Proof-of-concept published in 2016

Validation study ongoing, 64 samples run on PacBio & Sanger

Examples of other clinical targets – TP53

Application of PacBio long reads to study TP53 mutations in leukemia



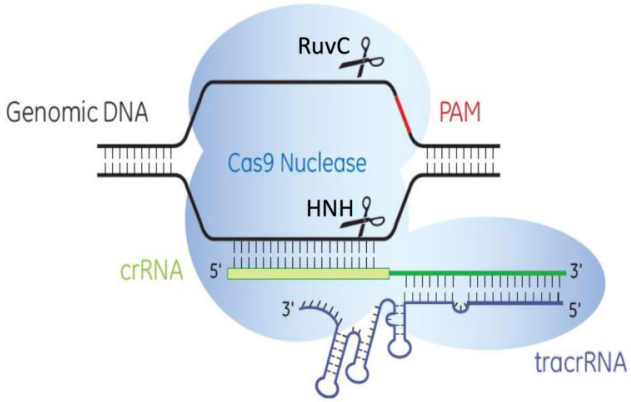
Thierry Soussi
Prof. Karolinska Institute

Manuscript in preparation

Amplification-free Targeted Enrichment

Why?

- Challenging regions for PCR amplification
  - Repeat expansions
  - Low complexity regions
- No PCR bias
- Preserves epigenetic modification signals



http://dharmacon.gelifesciences.com/applications/gene-editing/

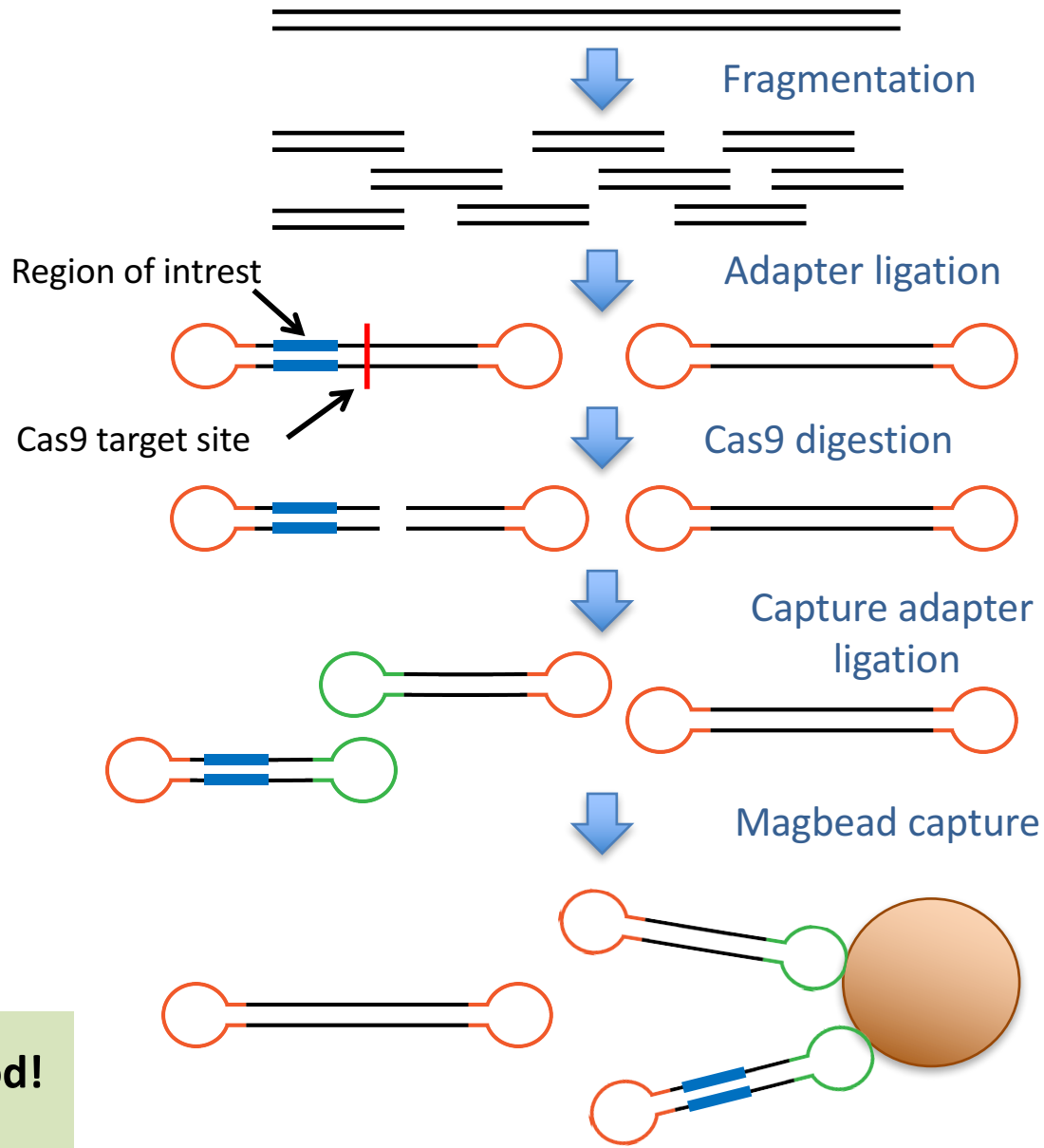
How?

- Novel method utilizing the CRISPR/Cas9 system
- Capture adapters/magnetic beads for enrichment
- Collaboration with PacBio & Mt Sinai

CRISPR/Cas9 enrichment protocol

Ida Höijer
NGI Uppsala

- RE fragmentation of gDNA
- Adapter ligation to form standard SMRTbell
- Digestion of SMRTbell using cas9
- Ligation of capture adapter to enable enrichment of SMRTbells containing region of intrerest
- Capture using magbeads
- Sequencing of enriched SMRTbells



+ Development of new analysis method!
(more about this tomorrow)

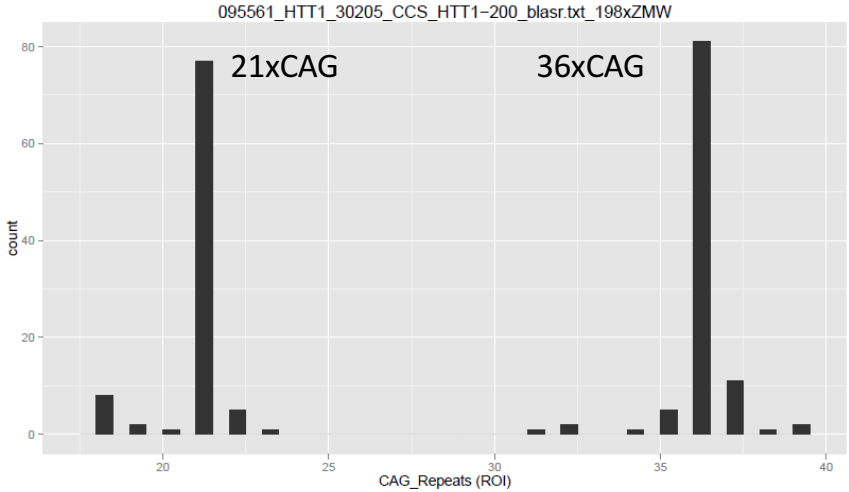
Results for 12 Huntington's patient samples

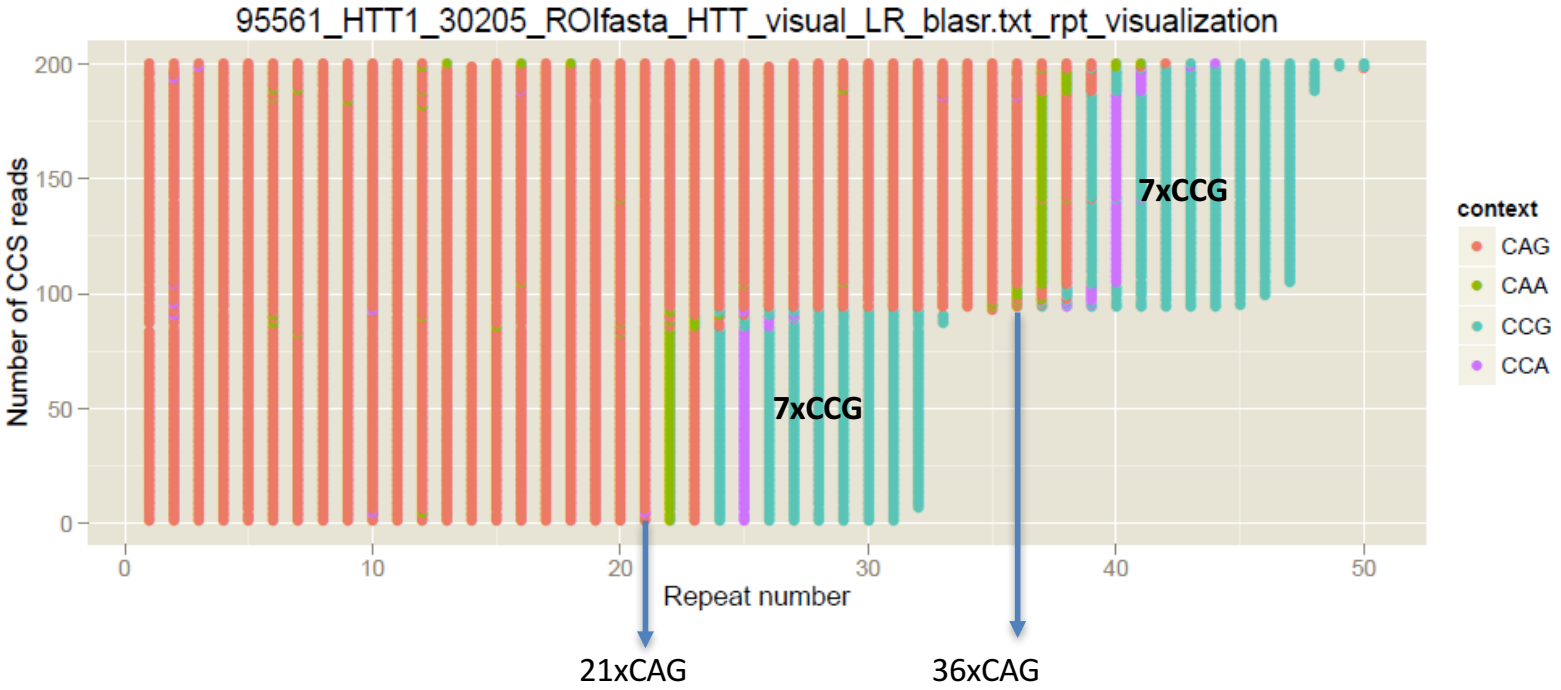
Number of CAG repeats in HTT gene:

	Repeat Count Fragment Analysis		Repeat Count CRISPR/Cas9 Method		
Sample	Allele 1	Allele 2	Allele 1	Allele 2	Reads on Target
1	15	54	15	54	77
2	17	51	17	52	73
3	17	41	17	41	90
4	17	39	17	39	135
5	15	38	15	38	120
6	21	36	21	36	200
7	17	33	17	35	274
8	21	29	21	29	193
9	20	27	20	27	164
10	17	25	17	25	86
11	17	23	17	23	304
12	17	18	17	18	131

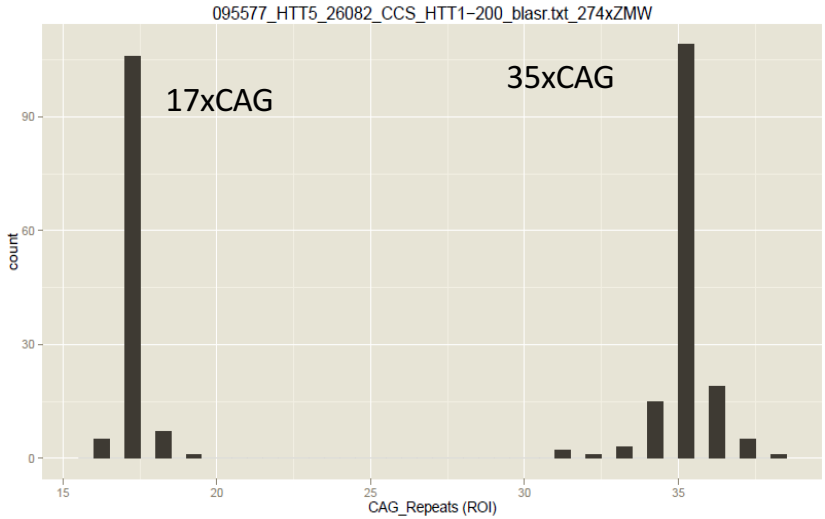
Identical results as compared to fragment analysis for 22 of 24 alleles

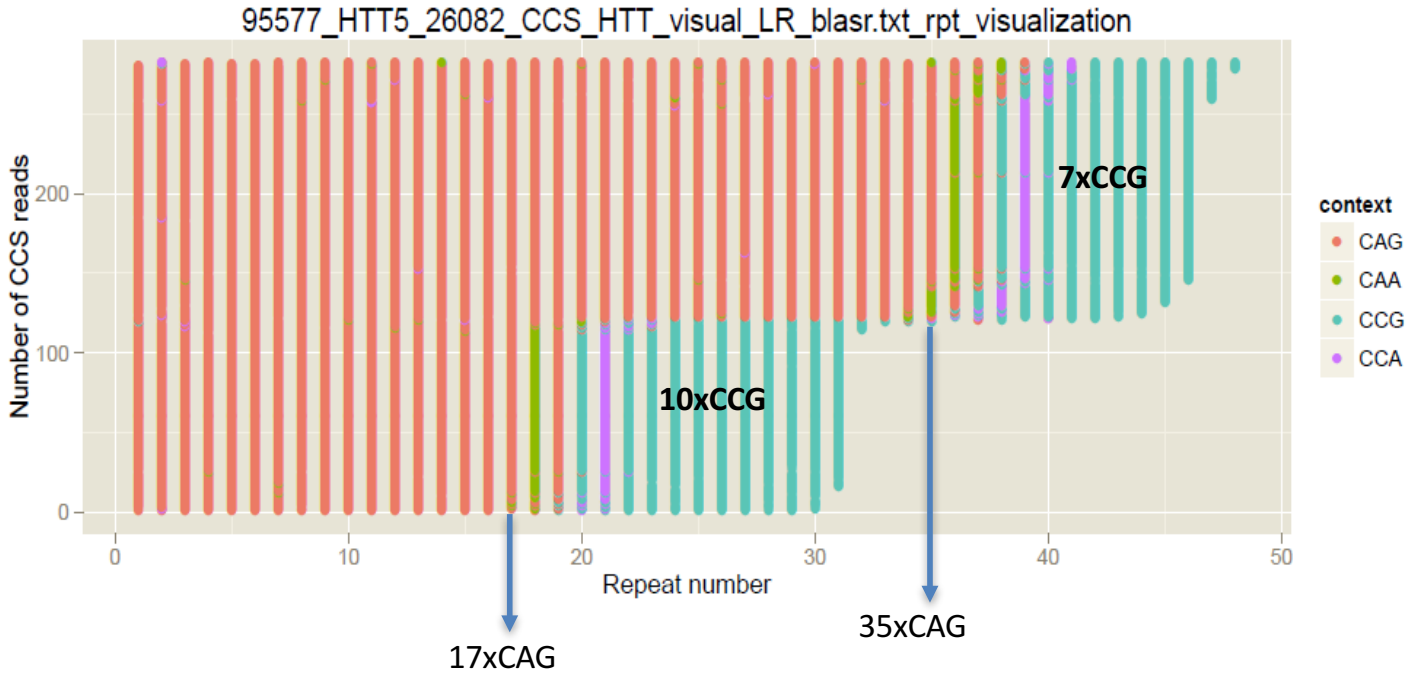
Results - Huntington's patient 1





Results - Huntington's patient 2





Pros and cons with current protocol

Advantages

- No PCR bias
- Multiplexing of targets possible
- Accurately sequence through long repetitive and low complexity regions
- Preserves epigenetic modification signals

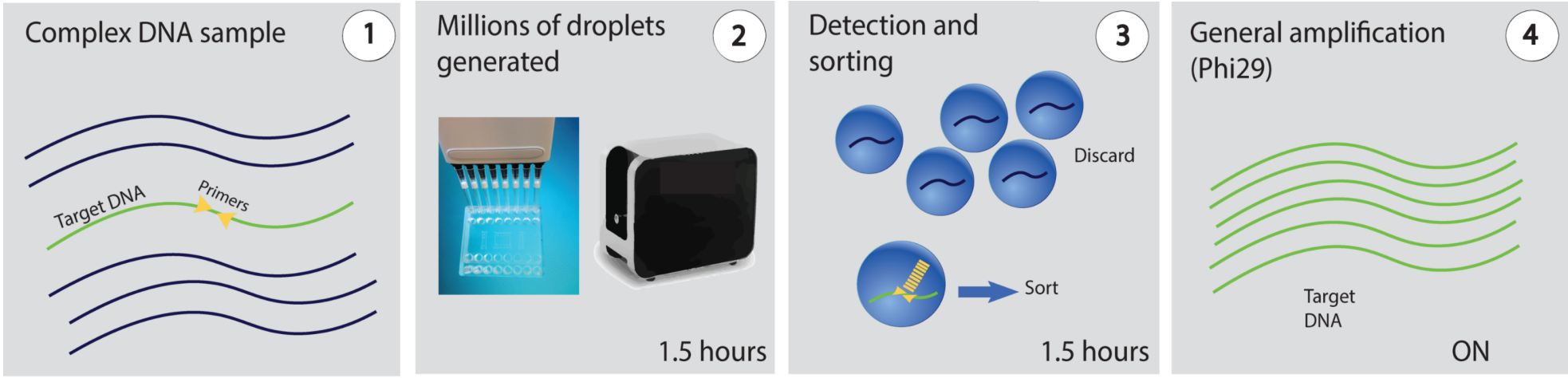
Disadvantages

- Requires a lot of input DNA
- Long turn-around time
- Complex protocol – requires a certain level of expertise
- Limited QC abilities
- Limited number of reads on target generated

PINS – Microfluidic droplet enrichment

- New technology for targeted enrichment of long fragments
- PINS is being commercialized by Samplix (DK)

Sampl/x

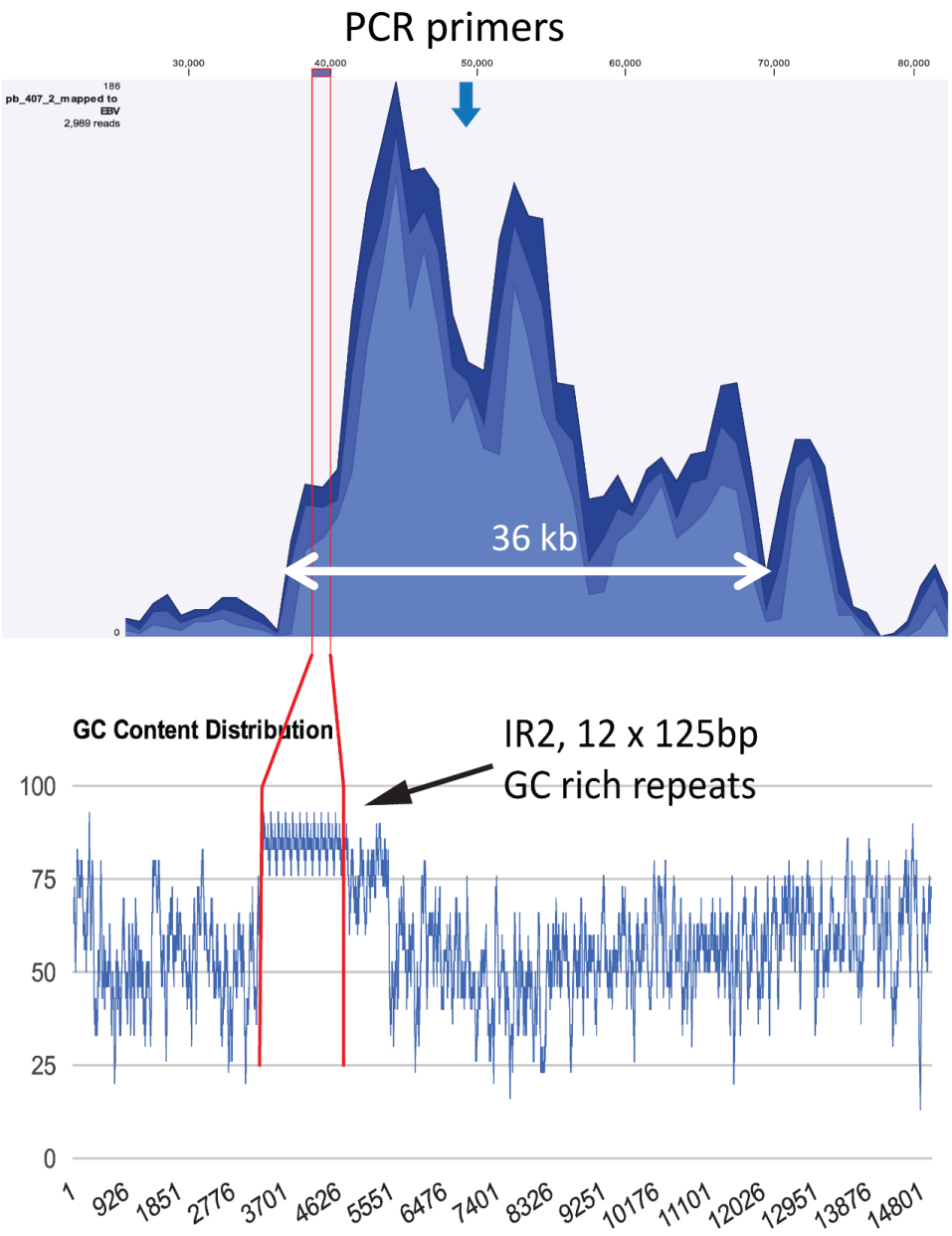


1) Mix long DNA fragments directly from DNA purification with reagents for specific PCR detection of the target fragment
2) Generate 10 million droplets, run PCR
3) Detect and sort droplets based on fluorescence
4) Amplify long target fragments from positive droplets using Phi29

Enrichment of EBV from mixed sample

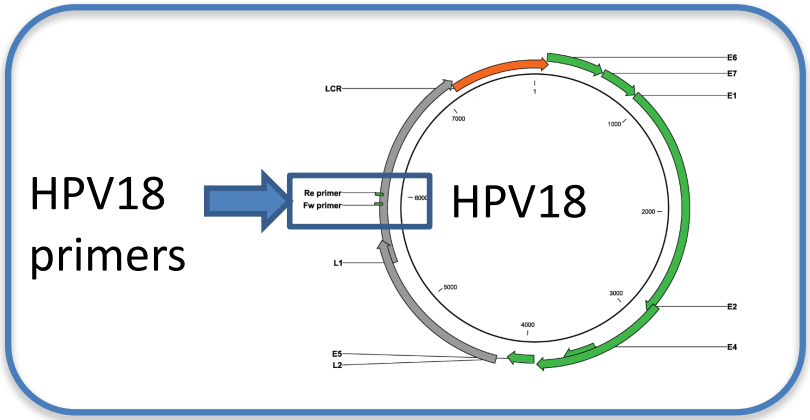
PacBio sequencing of Epstein Barr Virus in a background of human DNA

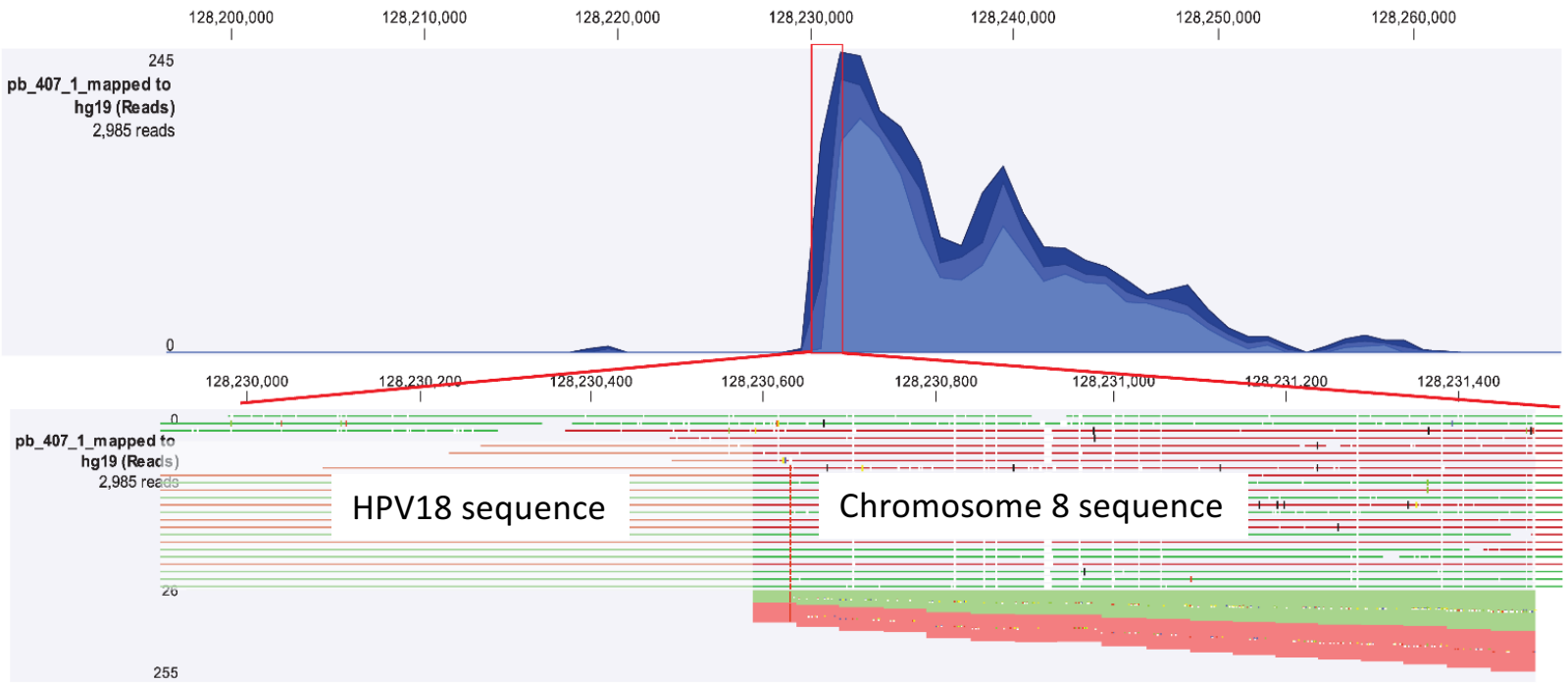
- 2 kb library prep for SMRT sequencing
- 36 kb EBV sequence obtained from one set of primers
- Includes 12x125 bp GC rich repeat!



Detection of integrated HPV 18

- PINS using detection primers in HPV18 L2
- 2 kb library prep for SMRT sequencing
- >20 kb sequence from the human side of the main breakpoint





Next steps for PINS droplet enrichment

We think PINS is a promising method!

- Sensitive detection of low-frequency molecules (similar to ddPCR)
- Low input protocol: >3ng human DNA required
- All amplification in droplets, no inter-allele chimeric reads
- Possible to multiplex samples and targets

EU project started (April 2017)

- Samplix will build a complete enrichment system
- NGI Uppsala will evaluate the technology using PacBio sequencing

Samplix

eurostars™

EUREKA

Human Whole Genome SMRT Sequencing

Crowd image via www.shutterstock.com

The Swedish 1000 Genomes Project

SweFreq | About | Terms of use | Data Beacon | SweGen Browser | Download Data | Admin | Logout

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 SweGen Browser by clicking the links above. To access the variant frequency file you need to register.

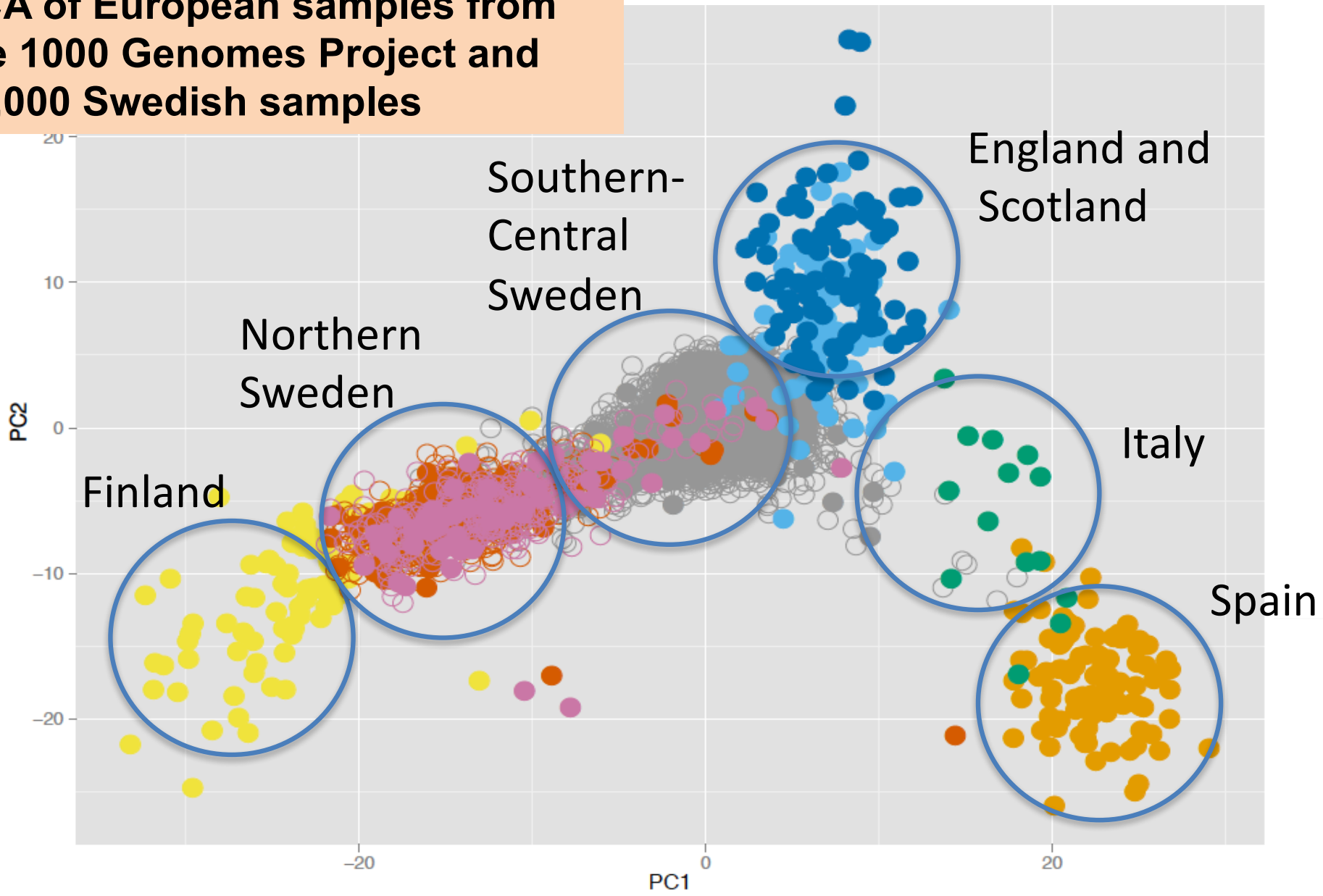
Please note that the 1000 individuals included in the SweGen project represent a cross-section of the Swedish population and that no disease information has been used for the selection. The frequency data may therefore include genetic variants that are associated with, or causative of, disease.

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

Ameur et al, submitted

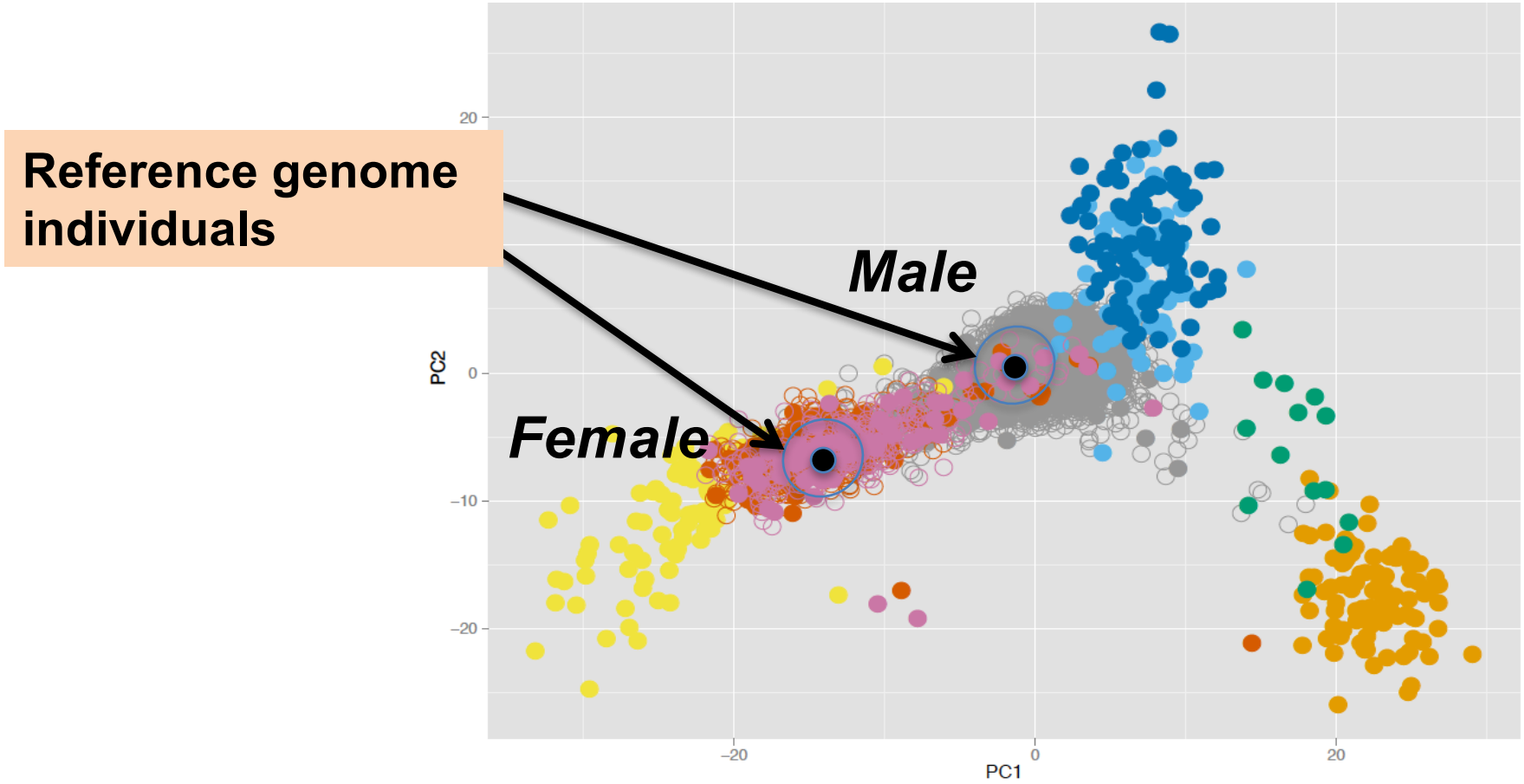
Selecting SweGen samples based on PCA

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



WGS of Swedish control cohort

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



Available data for two reference individuals

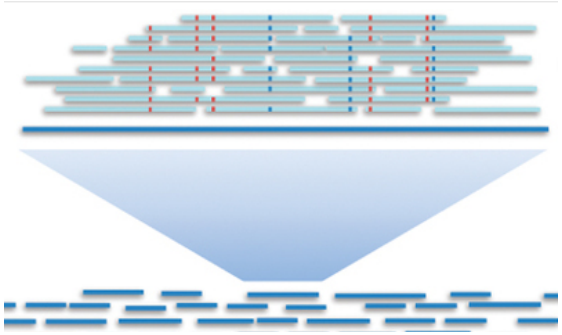
Data type	Amount (per individual)
SMRT PacBio	75X coverage
BioNano	2 x 100X coverage
10X Chromium	50X coverage
Illumina WGS	30X coverage
SNP array	700k SNPs

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

De novo assembly of 75X PacBio data

Assembly (FALCON)

Error correction (2 x Quiver)

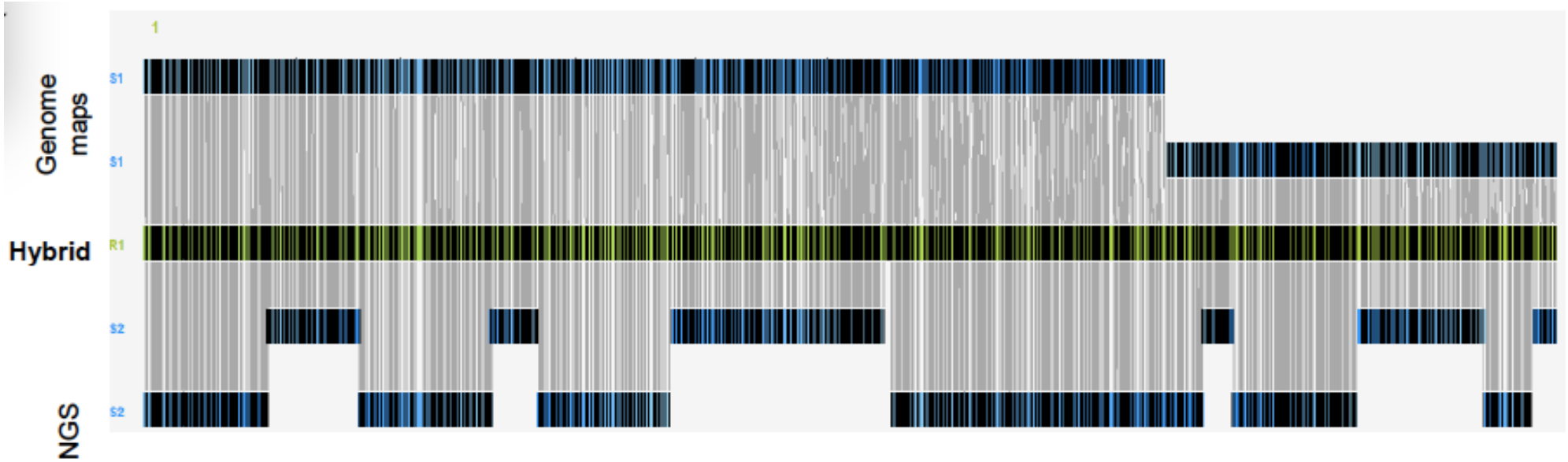


Analysis time: ~3 weeks/genome

	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

Hybrid scaffolding, PacBio + BioNano

Genome maps created, starting from frozen blood (!)

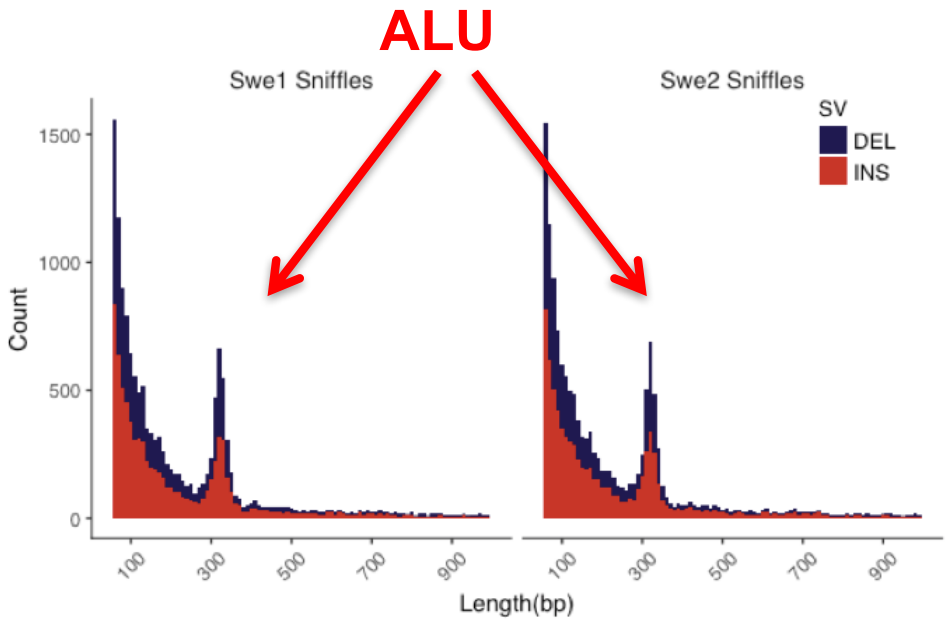


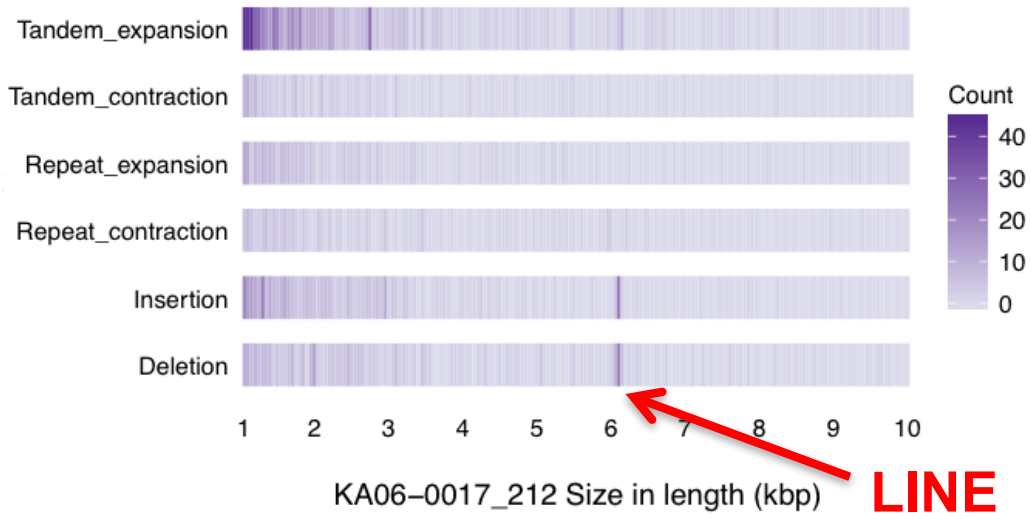
Hybrid scaffolding with two labelings resulted in

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

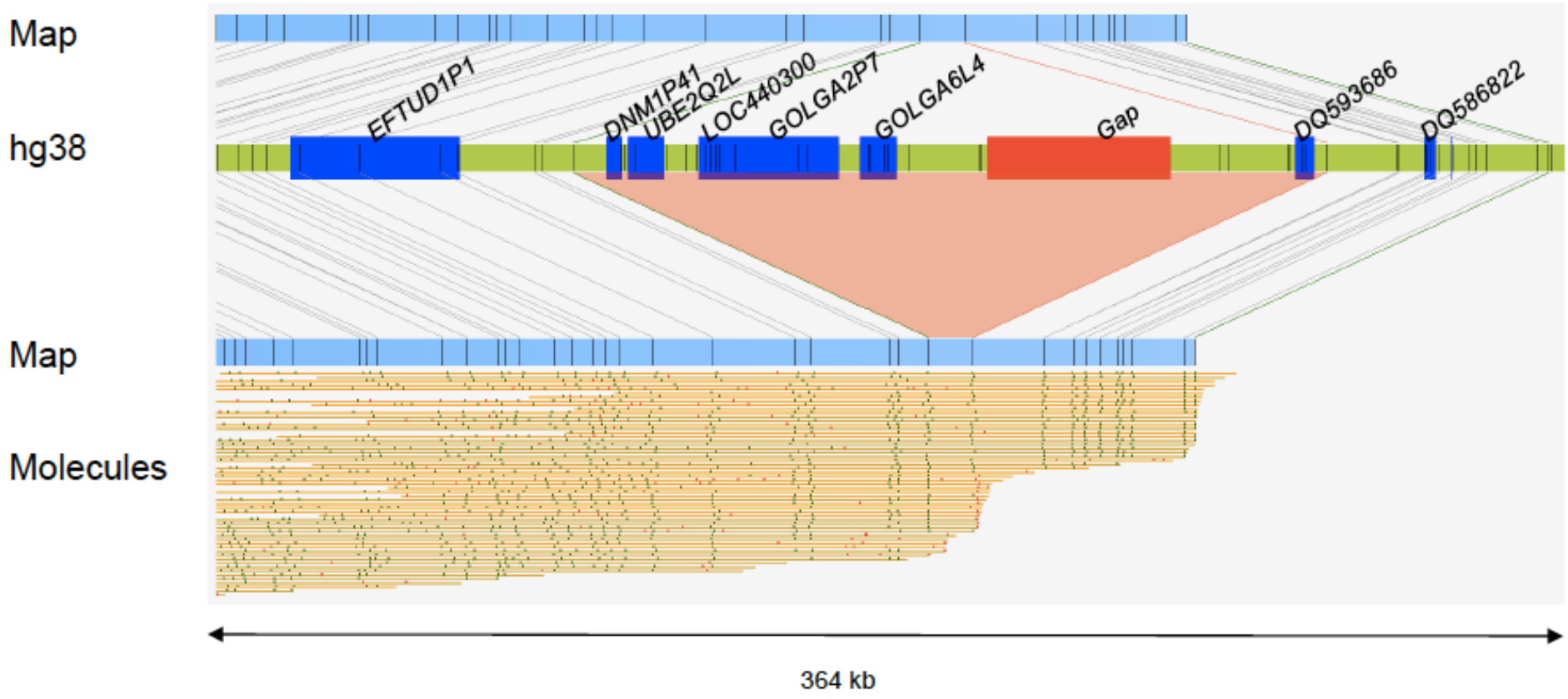
Overview of PacBio Structural Variation

Results for NGM-LR + Sniffles	Swe1	Swe2
INS	9290	9172
DEL	7673	7565
DUP	1385	1270
INV	278	295
DUP/INS	27	30
TRA	1926	1831
Total (in numbers)	**20579**	**20163**



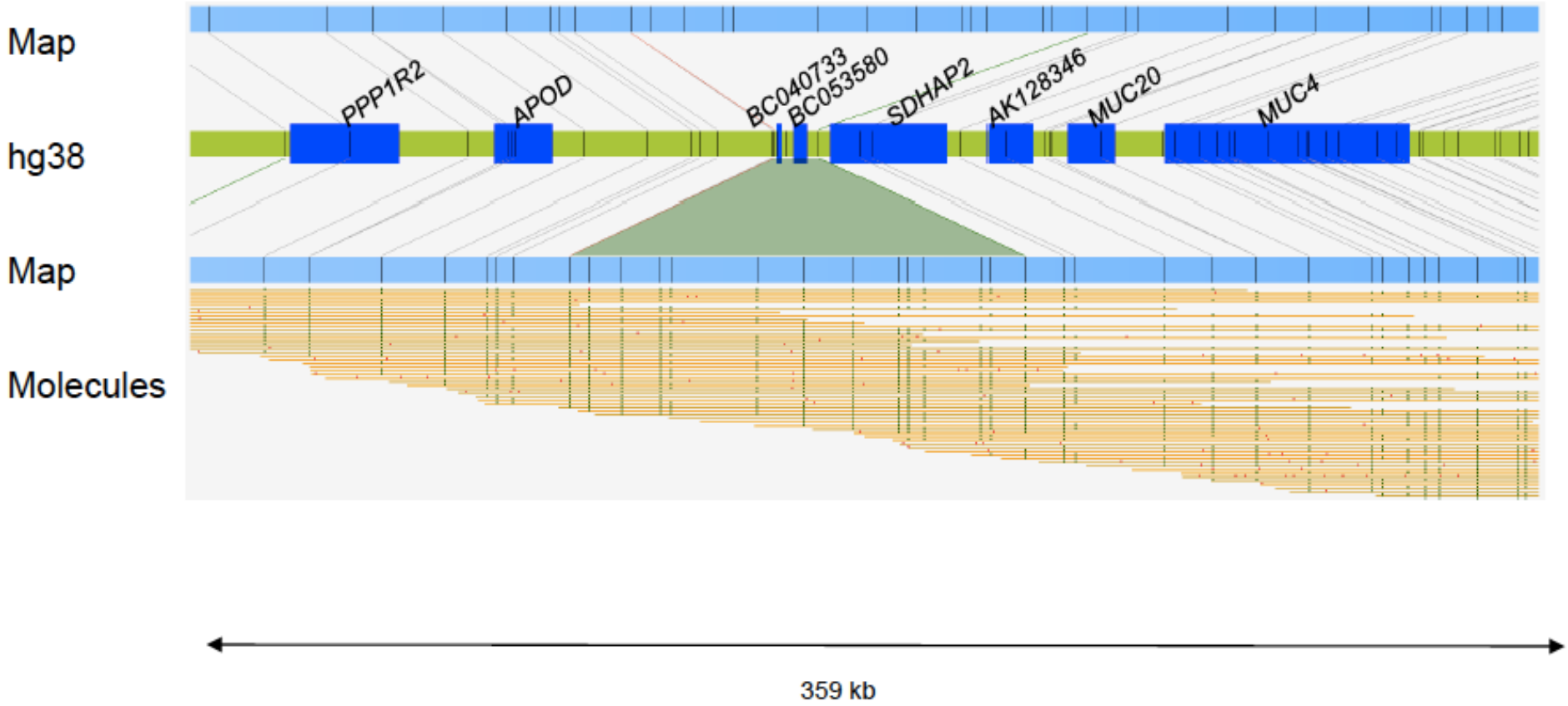


Example of a deletion in KA06-0017



Homozygous deletion of 193 kb at chr15:84,157,598-84,362,745.
Part of the deletion overlaps N-base gap in hg38.
Nt.BspQI data shown. Deletion was also detected in Nb.BssSI data.

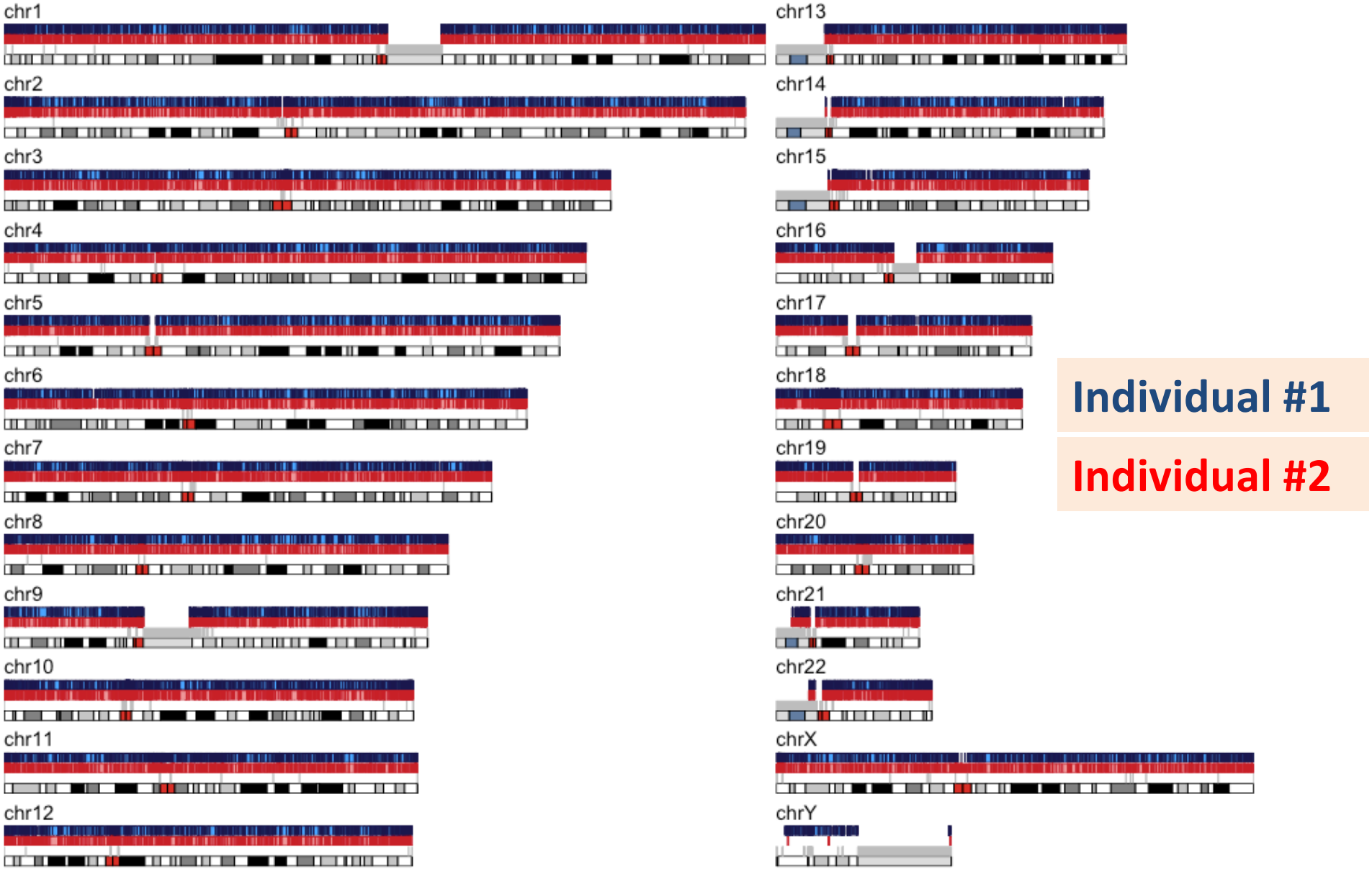
Example of a homozygous insertion in KA06-0031



Homozygous insertion of 109 kb at chr3:195,643,253-195,655,205.
Nt.BspQI data shown. Insertion was also detected in Nb.BssSI data.

Aligning contigs to human reference

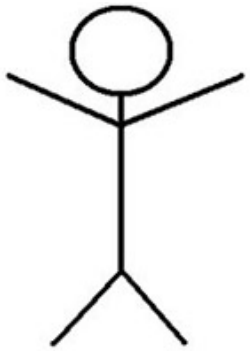
> 99% of bases can be aligned to human reference (hg38)



Discovery of ‘novel’ sequences

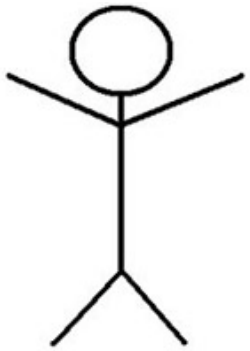
We find a lot of ‘novel’ sequence (NS), i.e. not matching hg38

Individual 1



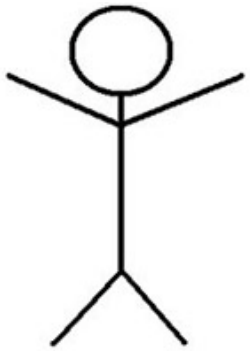
~14 Mb

Individual 2



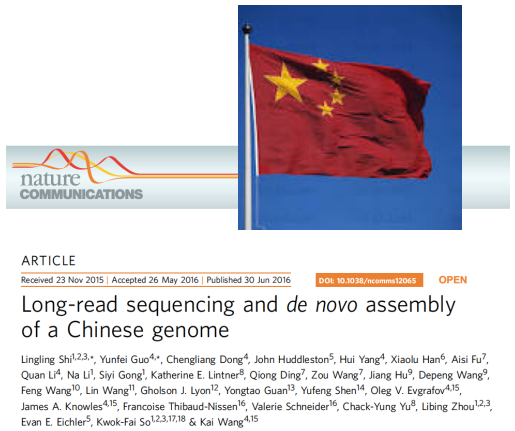
~10 Mb

Chinese



~13 Mb

Amount of novel sequences



Shi et al, 2016

~6 Mb of NS overlapping in all three genomes!

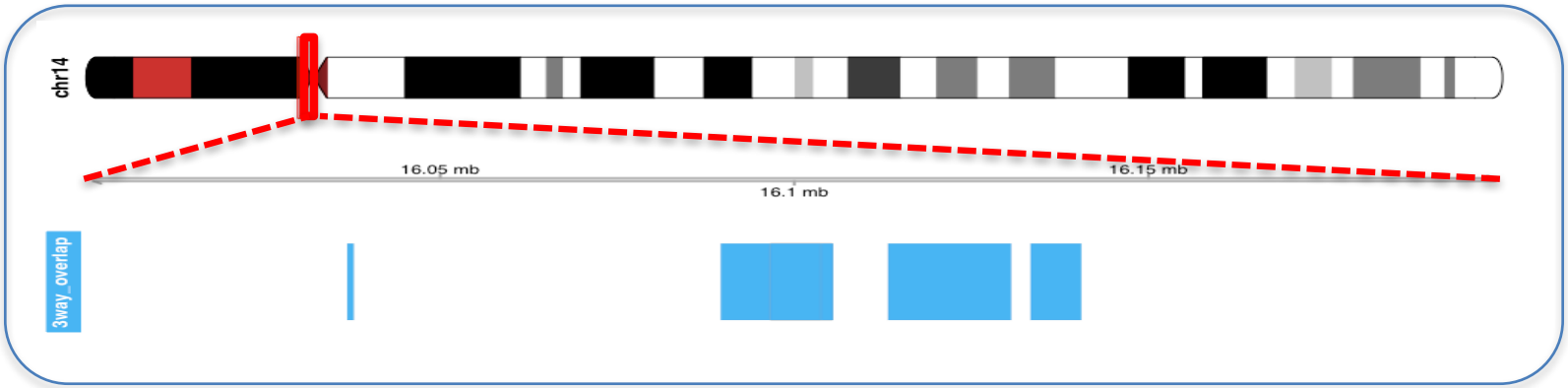
➔ Likely missing sequence in hg38 reference

Novel sequences – examples

Ex 1: chr14

Missing in reference

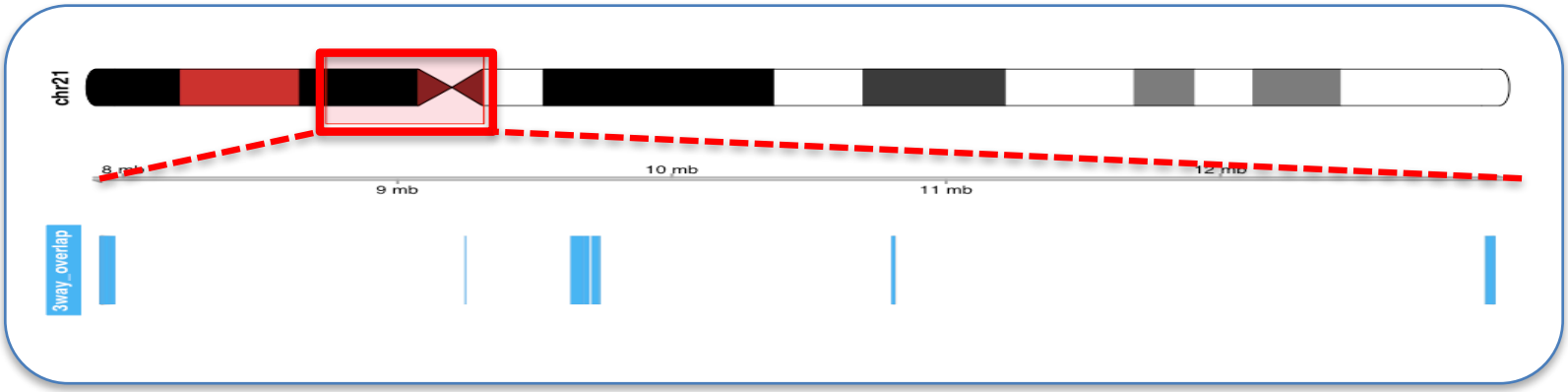
(3-way olap)



Ex 2: chr21

Missing in reference

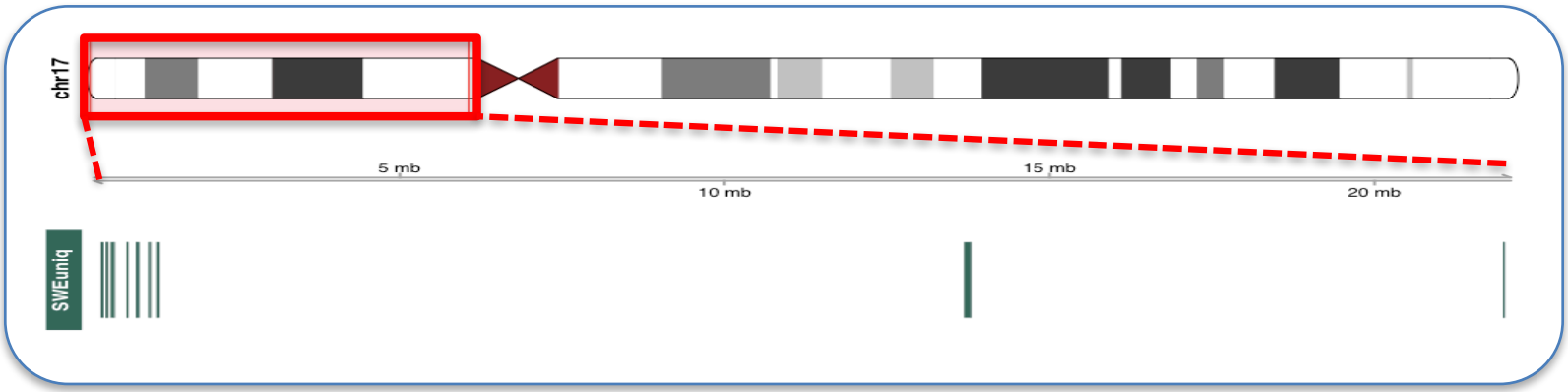
(3-way olap)



Ex 3: chr17

Population specific?

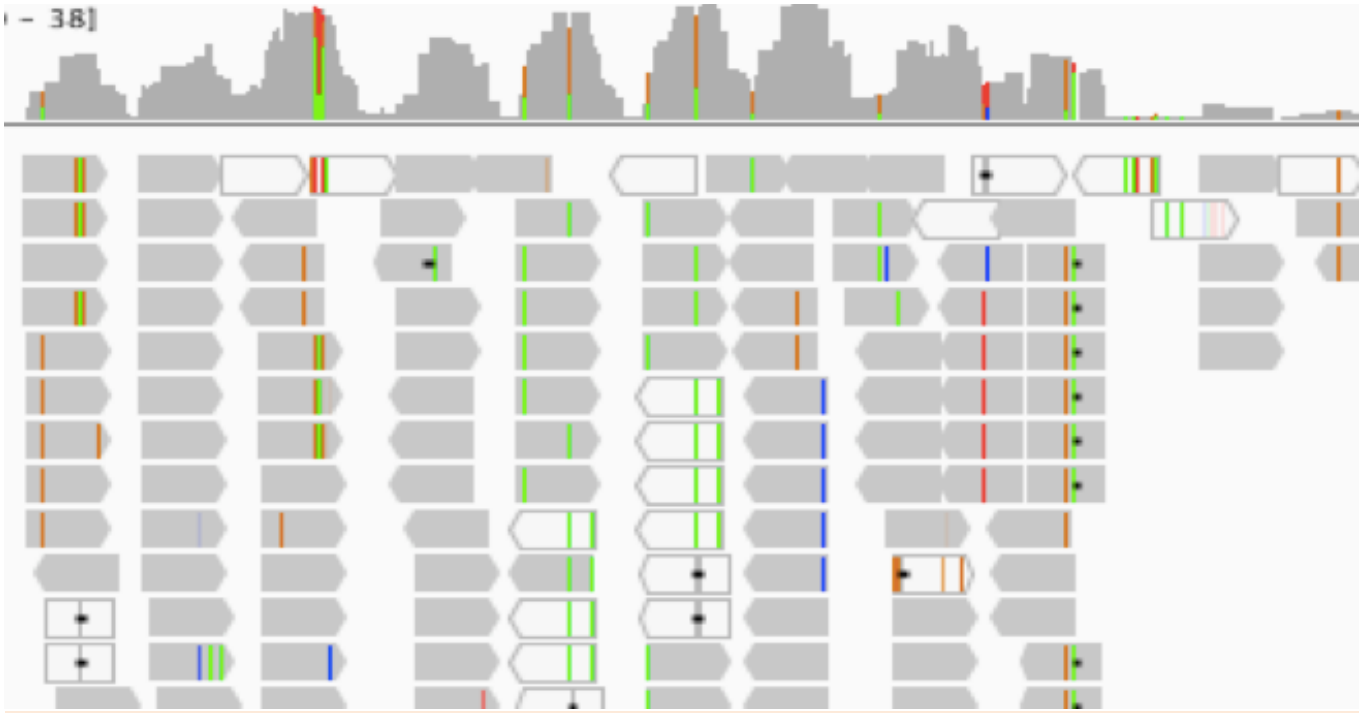
(not in Chinese)



Novel sequences – what are they?

Next steps to understand these novel sequences:

- Detect putative coding sequences (Maker)
- Analyse RNA-seq data (Illumina Body Map)
- Analyse epigenetic data (ChIP-seq, ENCODE)
- Analyse conservation
- ….



Example, RNA-seq data mapped to novel contig

What is the benefit of a Swedish reference?

Ongoing work:

Start from Illumina WGS reads for the 1000 Swedes

Align to GRCh38

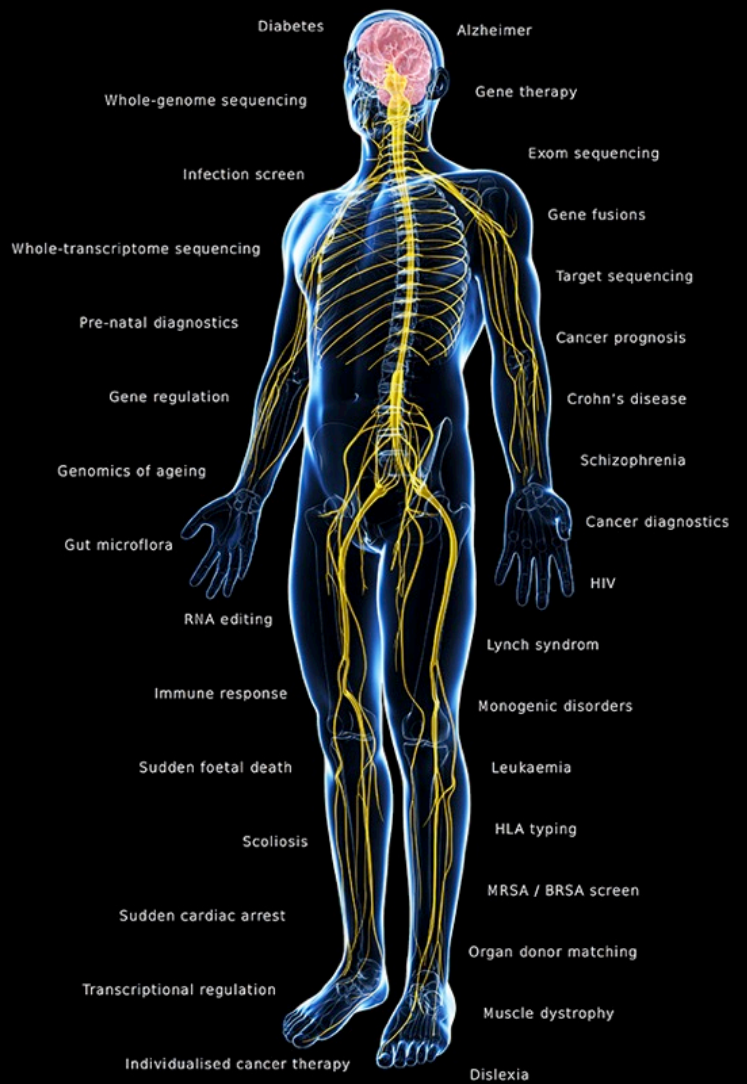
Align to GRCh38 +Swedish contigs

hg38

hg38

Compare results: nr aligned reads, mapping quality, SNPs, indels …

What we sequenced at SciLifeLab



16S amplicon, Acinetobacter baumanii, Acrasis kona, Acridotheres javanicus, Actinobacillus succinogenes, African swine fever virus, Agaricomycotina sp, **Alces alces**, **Alligator mississippiensis**, Amphiura filiformis, Apis mellifera, Aquila chrysaetus, Arabidopsis thaliana, Arabis alpina, Archaeorhizomycetes finlayi, Arctocephalus gazella, artificial sequences, Arvicola amphibius, Ascaridia galli, Aspergillus oryzae, Astrapia stephaniae, Atlantic herring, Atlantic salmon, **Avena sativa**, Baccharis breviseta, Baccharis dracunculifolia, Bacteriophages, Balaenoptera musculus, Balaenoptera physalus, Balanus improvisus, Baltic Sea microorganisms, Bathynomus sp, Bifidobactrium sp, **Borrelia burgdorferi**, Borrelia garinii, Bos taurus, Bovine viral diarrhea virus, Brachyspira suanatina, Brassica sp, Brettanomyces naardensis, Caenorhabditis elegans, Callosobruchus maculatus, Candida intermedia, Candida parapsilosis, Candidatus Neoehrlichia mikurensis, **Canis lupus**, Capreolus capreolus, Capsella bursa-pastoris, Capsella grandiflora, Capsella orientalis, Capsella rubella, Ceanothus thyrsiflorus, Cervus dama, Cervus elaphus, Childia submaculatum, Clonostachys rosea, Clostridium ultunense, Coelodonta antiquitatis, Colias crocea, Collinsia heterophylla, Coregonus lavaretus, Coronavirus, Corvus corone, Corvus monedula, Crassostrea gigas, Cricetulus griseus, Cryptococcus tephrensis, Cubanola dominguensis, Cytomegalovirus, Danio rerio, Datisca glomerata, Deformed wing virus, Dekkera bruxellensis, Dicerorhinus sumatrensis, Dictyostelium discoideum, Diopisthoporus gymnopharyngeus, Diopistoporus longitubus, Drosophila melanogaster, Drosophila paulistorum, Electrophorus electricus, Enterobacter cloacae, Enterococcus faecium, **Equus caballus**, Escherichia coli, Eumecynostomum macrobursalium, Euphorbia lathyris, Euphorbia peplus, Euplectes afer, Euplectes ardens, Euplectes aureus, Euplectes hordeaceus, Euplectes macrourus, Euplectes orix, **Felis catus**, Ficedula albicollis, Ficedula hypoleuca, **Fragaria ananassa**, Freshwater microbial communitites, Fucus radicans, Fucus vesiculosus, Fumaria sp, Galerucella, Gallus gallus, Geospiza magnirostris, Giardia muris, Globodera rostochiensis, Gnetum gnemon, Gnetum luofuense, Gnetum montanum, Gnetum parvifolium, Gnetum pendulum, Gonyostomum semen, Gonzalagunia, Gut microbiota, Hamelia marantha, Heterobasidion abietinum, Heterobasidion annosum, **Hippophae rhamnoides**, **Homo sapiens**, Human Immunodeficiency Virus, Huperzia selago, Hymenoscyphus albidus, Hymenoscyphus pseudoalbidus, Idotea baltica, **Influenza A virus**, Klebsiella pneumoniae, Laccaria bicolor, Lactobacillus, Lepidium campestre, **Leptidea sinapis**, Letharia rugosa, Letharia vulpina, Littorina saxatilis, Lycocorax pyrrhopterus, **Lynx lynx**, Malassezia sympodialis, **Malus domestica**, Malus sylvestris, **Mammuthus primigenius**, Marchantia polymopha, Marine bacteria whole community, **Meligethes aeneus**, Metagenomes, Methanoculleus sp, Metschnikowia andauensis, Metschnikowia hawaiiensis, Metschnikowia lopburiensis, Metschnikowia pulcherrima, Metschnikowia saccharicola, Mixornis gularis, Moorella thermoacetica, Mus musculus, Mycobacterium malmoense, Mycobacterium marinum, **Mytilus edulis**, Nemertoderma westbladi, Nesophontes sp, Neurospora crassa, Neurospora hispaniola, Neurospora intermedia, Neurospora metzenbergii, Neurospora perkinsii, Neurospora sitophila, Neurospora tetrasperma, Nora Virus, Nothoprocta ornata, Nothoprocta perdicaria, **Notophthalmus viridescens**, Nyctereutes procyonoides, Ogataea pini, **Oryctolagus cuniculus**, Rana arvalis, Oryzias latipes, **Pacifastacus leniusculus**, Paenibacillus polymyxa, **Panthera leo**, Panthera pardus, Paradisaea rubra, Parus major, Passer montanus, Paxillus involutus, Penicillium sp, **Perca fluviatilis**, Peridinium aciculiferum, Philomachus pugnax, Phoca sibirica, Phylloscopus collybita, Phylloscopus trochilus, Wolbachia persica, Physcomitrella patens, Phytophthora infestans, **Picea abies**, Pieris napi, Pieris rapae, Pinus pinaster, Pinus sylvestris, **Pisum sativum**, Planctomycetes sp, Plasmodiophora brassicae, Plasmodium falciparum, Podospora anserina, Polystachya paniculata, Pomatoschistus minutus, Populus maximowiczii, Populus tremula, Populus trichocarpa, Posoqueria sp, Pseudomonas aeruginosa, Pseudomonas brassicacearum, Pseudomonas chlororaphis, Pseudomonas putida, Pteridophora alberti, Ptiloris paradiseus, Puccinia striiformis, Pythium oligandrum, Quelea quelea, Rangifer tarandus, Rattus rattus, Rhizoctonia sp, Saccharomyces cerevisiae, Salix purpurea, Salix viminalis, Salmonella enterica, Salmonella typhimurium, Salmo salar, Salmo trutta, Schizophyllum commune, Schizosaccharomyces pombe, Scrippsiella hangoei, **Semibalanus balanoides**, Setaria digitata, Silene conoidea, Silene latifolia, Silene viscaria, Sindbis virus, Siphocampylus, Siphocampylus retrorsus, Siphoviridae phage, Skeletonema marinoi, **Solanum tuberosum**, Sorghum sp, Spironucleus barkahanus, Spironucleus salmonicida, Spironucleus vortens, Staphylococcus aureus, Staphylococcus pseudintermedius, Stemmadenia sp, Streptococcus pneumoniae, Streptococcus pyogenes, Streptomyces coelicolor, Struthio camelus, Sulfolobus acidocaldarius, **Sus scrofa**, Synthetic DNA, Syntrophaceticus schinkii, Taphrina betulina, Tepidanarobacter aceatoxydans, Thamnolia vermicularis, Theileria parva, Trypanosoma cruzi, Trypanosoma rangeli, **Ursus spelaeus**, Vitex agnus-castus, Yarrowia lipolytica, Zalophus californianus, Zalophus wollebaeki, Zygopetalum crinitum, Zygosaccharomyces bailii

Version 2.0 of the Spruce genome

Genome size 20Gbp

>95% repeat content

2013: First version of Spruce genome published

- Mainly based on Illumina sequencing

ARTICLE OPEN

doi:10.1038/nature12211

The Norway spruce genome sequence and conifer genome evolution

Nystedt et al, Nature, 2013

2017: Starting on an improved Spruce genome

- ~200 PacBio Sequel SMRT cells estimated
- Complemented with BioNano maps
- Largest PacBio project in Sweden!

Nathaniel Street, PhD,

Umeå Plant Science Center

Acknowledgements

SciLifeLab and Uppsala University
Ulf Gyllensten
Inger Jonasson
Huiwen Che
Ida Höijer
Ignas Bunikis
Marie-Louise Bondeson
Lucia Cavelier

Karolinska Institute
Thierry Soussi

BioNano Genomics
Alex Hastie

PacBio
Tyson Clark
Yu-Chih Tsai
Paul Kotturi
David Greenberg

Samplix
Marie Just Mikkelsen
Thomas Kvist
Esben Madsen

NATIONAL GENOMICS INFRASTRUCTURE

SciLifeLab

UPPSALA UNIVERSITET