

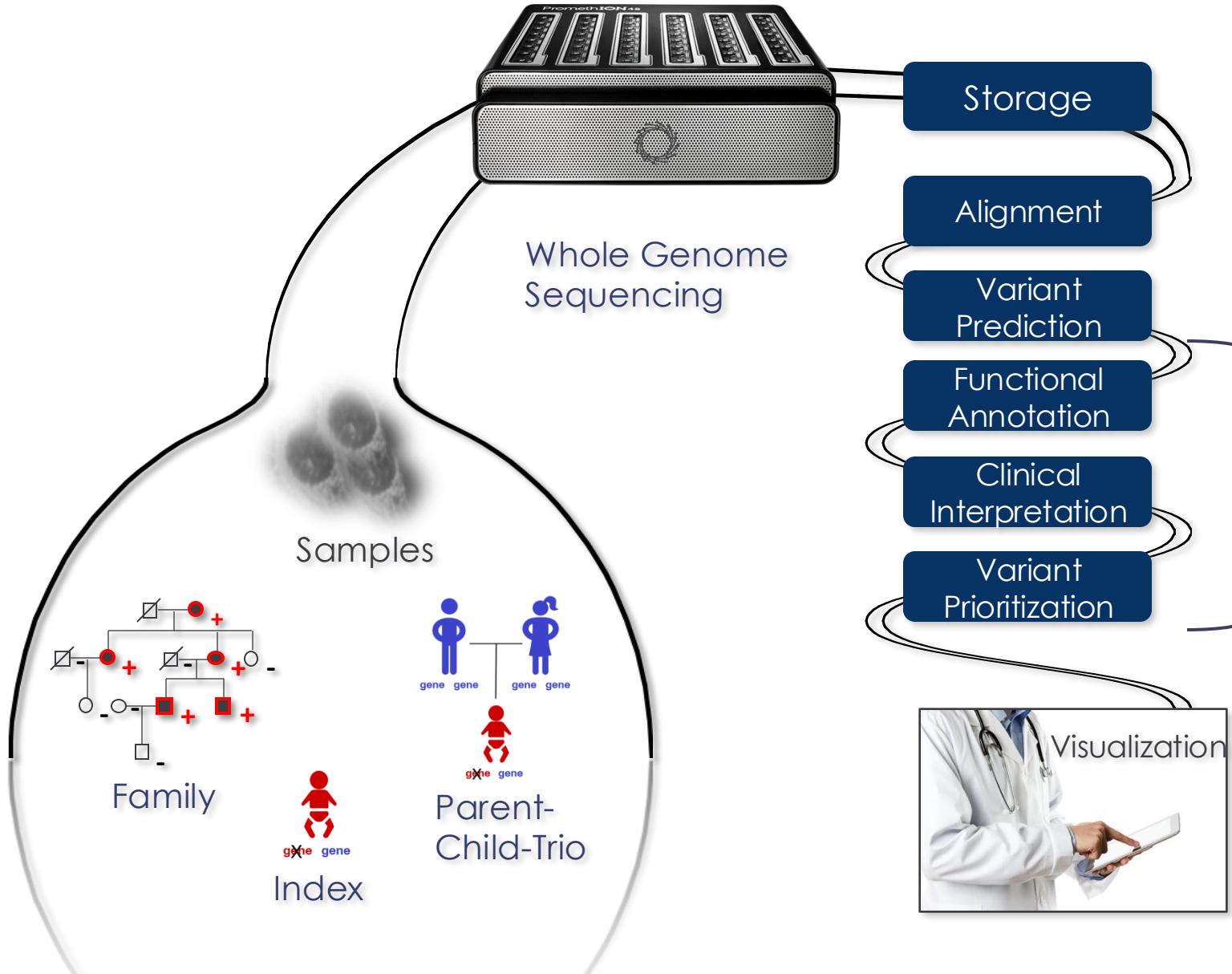
GHGA 2024

# Rare Disease Diagnostics with Long Leads

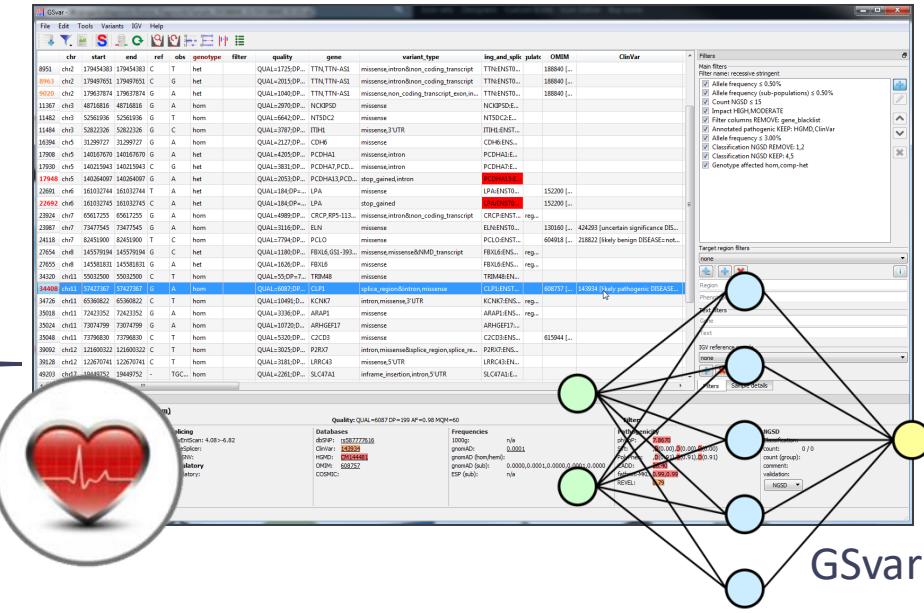




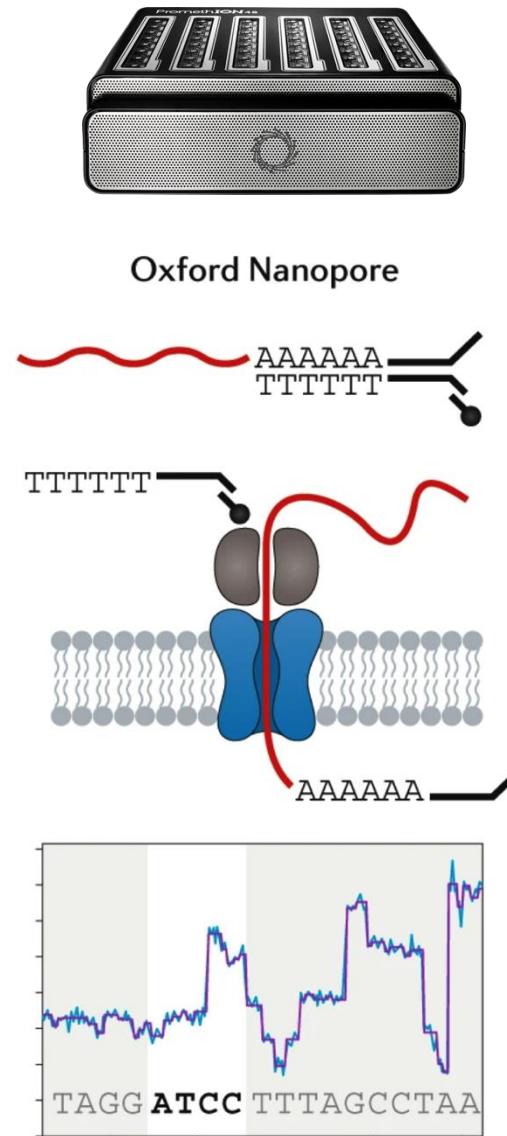
# Goal: Rare Disease Diagnostics with Nanopore



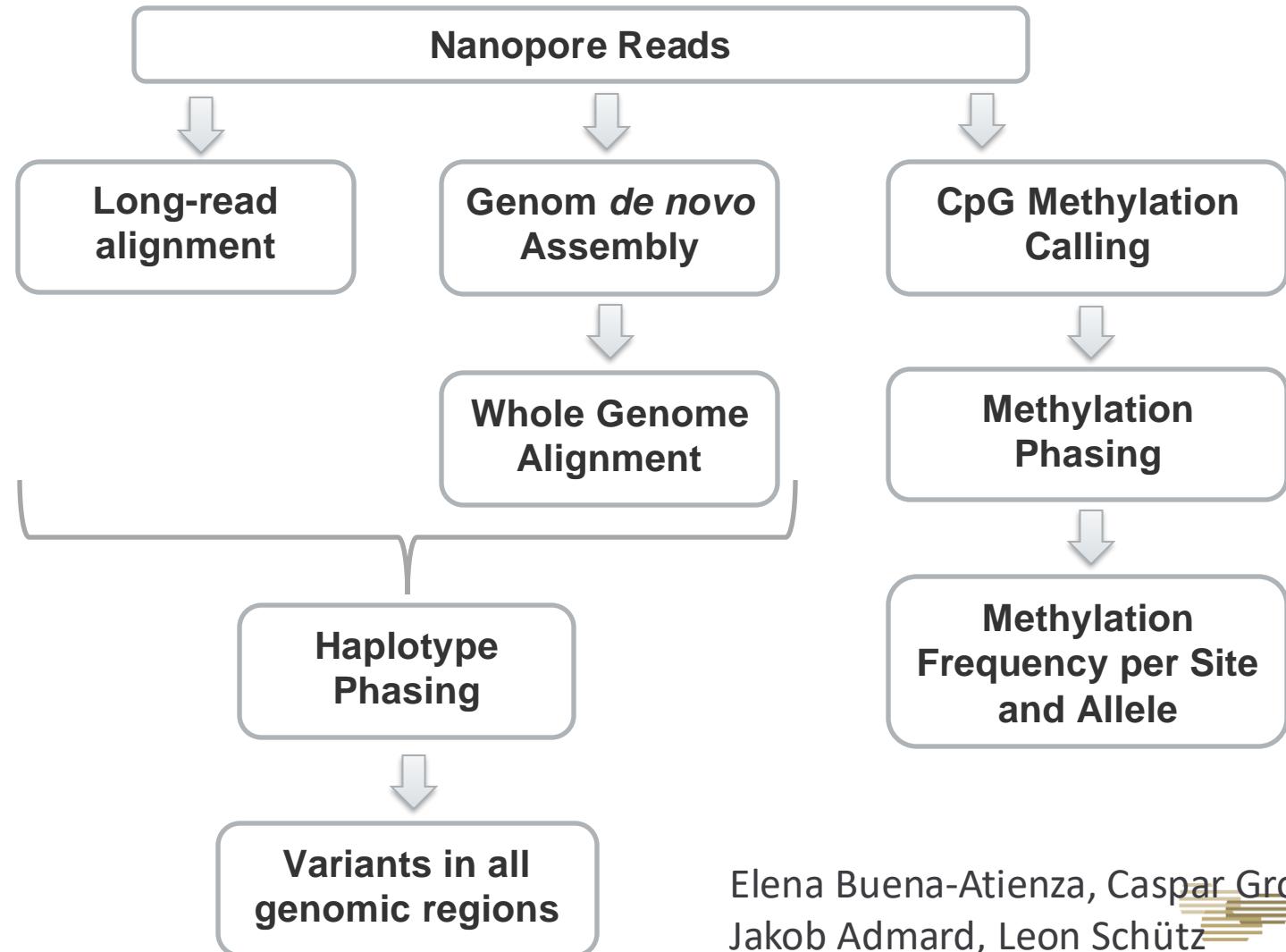
Clinical Decision Support Systems,  
Clinical Databases, EHR-Textmining, AI



# Genome & Methylome Diagnostics with Long Reads

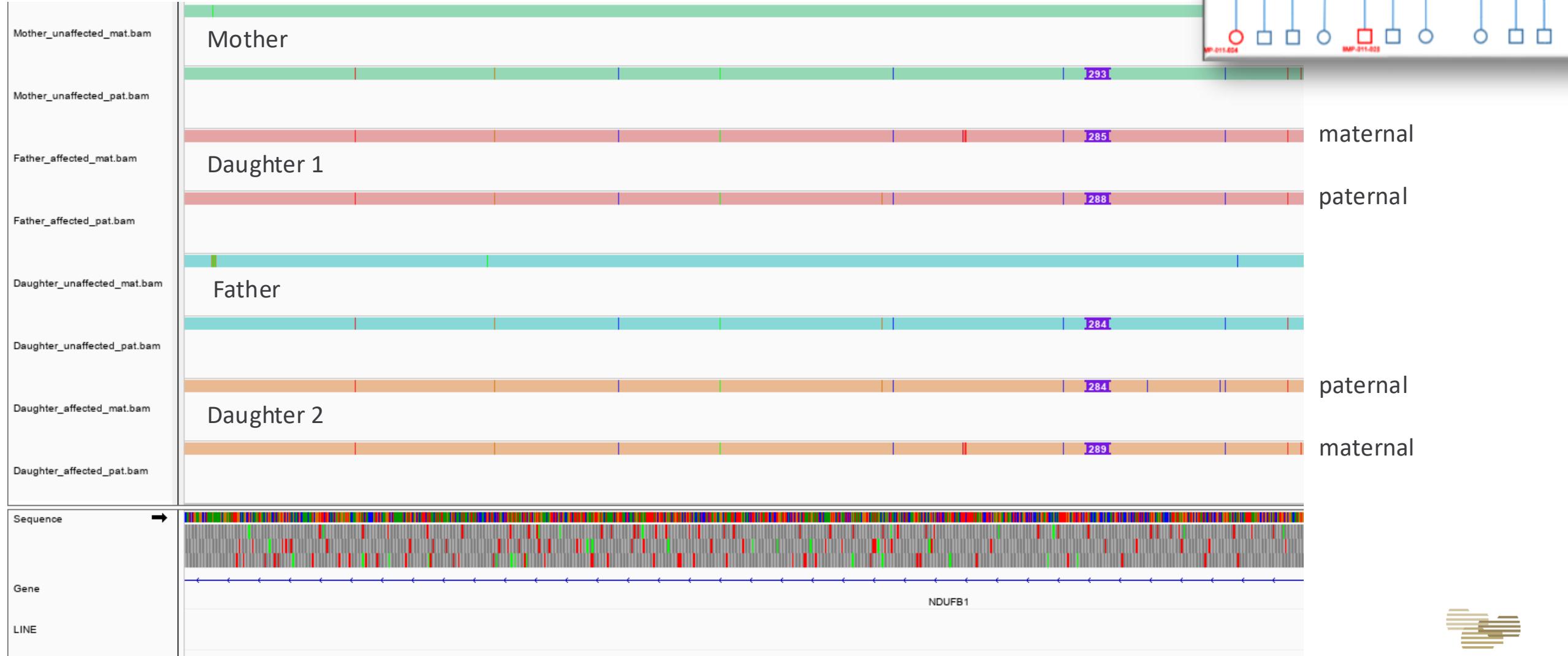


## New analysis options:



Elena Buena-Atienza, Caspar Gross,  
Jakob Admard, Leon Schütz

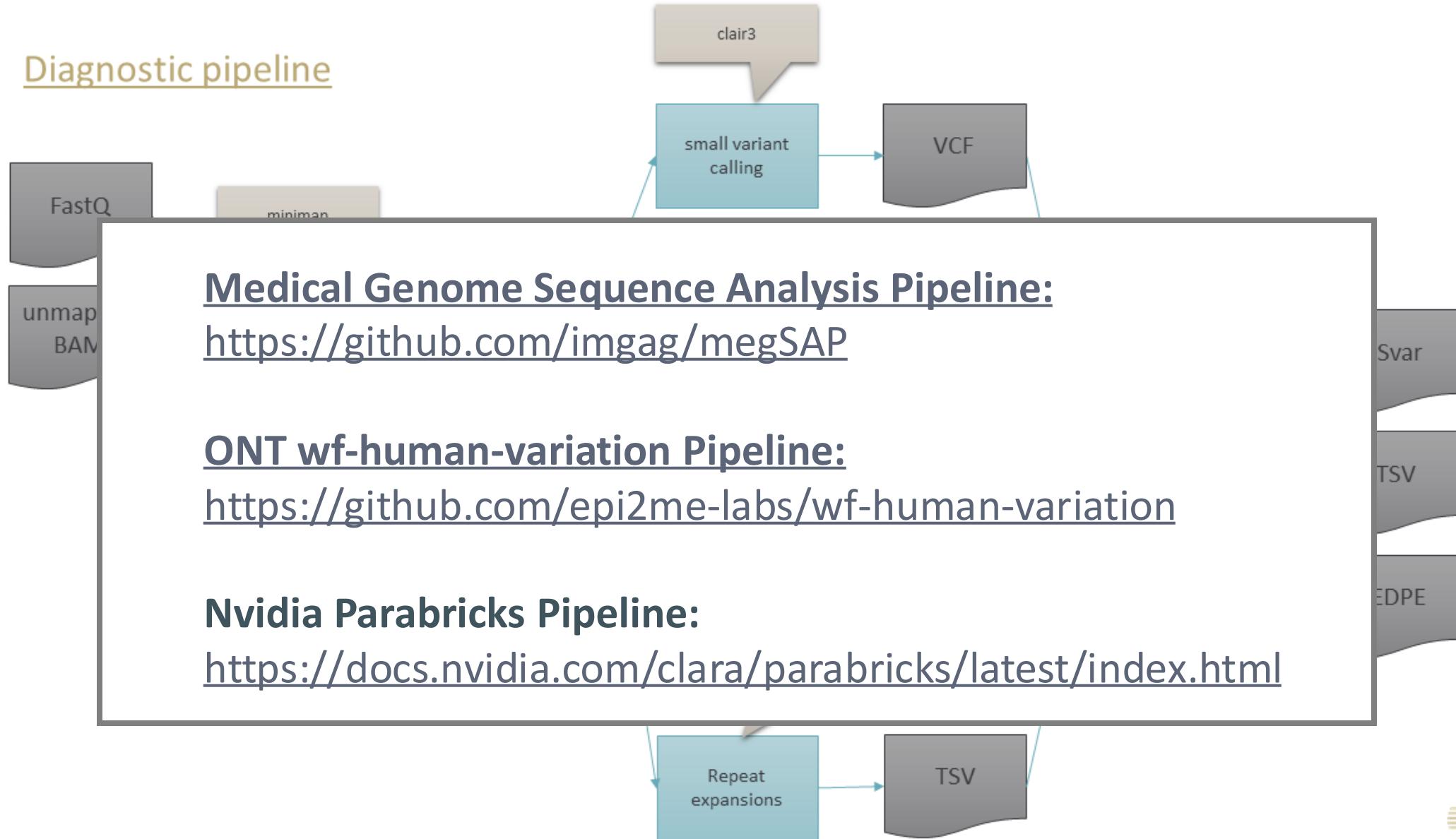
# Long-Reads give Haplotype-Phased Genomes (Example in IGV)





# Variant-Analysis Pipelines for Nanopore Data

## Diagnostic pipeline



# Clinical Decision Support System Gsvar Fully Supports Long Read Diagnostics (DNA, RNA, Methylation)

**Gsvar - single-sample analysis DNA2301278A1\_01**

File NGSD Tools Conversion IGV Help

Variants X

Repeat Expansions of single-sample analysis DX20242\_01

filter quality

QUAL=388;DP=42;QD=9.24;AF=0... STK

QUAL=1963;DP=198;QD=9.91;AF=0... AB

QUAL=2816;DP=242;QD=11.64;AF=0... PO

QUAL=673;DP=66;QD=10.20;AF=0... LG

QUAL=3626;DP=289;QD=12.55;AF=0... SLC

QUAL=1795;DP=160;QD=11.22;AF=0... PPI

QUAL=1191;DP=95;QD=12.54;AF=0... RFX

chr start end repeat\_id repeat\_unit repeats wt\_repeat repeat\_ci filter locus\_coverage reads\_flanking reads\_in\_repeat reads\_spanning

1 chr1 1493990802 1493990841 NOTCH2NL GGC 8/15 16 8-8/15 PASS 64.34 66/95 0/0 32/18

2 chr2 190888072 190889029 GLS GCA

3 chr3 63912684 63912714 ATXN7 GCA

4 chr3 63912714 63912785 ATXN7\_GCC GCG

5 chr3 129172576 129172656 CNBP CAGG

6 chr3 129172556 129172656 CNBP\_CAGA CAGA

7 chr3 129172696 129172723 CNBP\_CA CA

8 chr4 3074876 3074933 HTT CAG

9 chr4 3074939 3074966 HTT\_CCG CCG

10 chr4 39348424 39348479 FCF1 AARR

11 chr4 41745972 41746032 PHOX2B GCN

12 chr5 146878727 146878757 PPP2R8B GCT

13 chr6 16327723 16327723 ATXN1 TGC

14 chr7 170561906 170562017 TEP GCA

15 chr9 2757528 2757354 C9ORF72 GGCC

16 chr9 69037286 69037304 FXN\_A A

17 chr9 69037286 69037304 FXN GAA

chr11 119206289 119206322 CBL CGG

chr12 6936716 6936773 ATNI CAG

chr20 50505001 5050502 DIP2Q GGC

chr12 111598949 111599018 ATXN2 GCT

chr23 70139353 70139383 ATXNBOS\_CTA CTG

chr13 70139383 70139423 ATXNBOS CTG

chr14 23321472 23321490 PABPN1 GCG

chr14 92071009 92071042 ATXN3 GCT

chr15 22786677 22786701 NIPA1 GCG

chr17 87604287 87604329 JPH3 CTG

chr18 55586155 55586227 TCF4 CAG

chr19 13207858 13207897 CACNA1A CTG

chr20 4577024 4577024 DMPK CAG

chr20 2652733 2652757 NOP56 GGGC

chr20 2652757 2652775 NOP56\_CGCG CGGC

chr21 43776443 43776479 CSTB CGGC

chr22 45795334 45795424 ATXN10 ATTC

chr22 67545316 67545385 AR GCA

chr24 147912050 147912101 FMR1 CGG

chr24 148500631 148500691 AFF2 GCC

Gene(s): SLC17A5 (inh=AR, oe=AR)

**SLC17A5 ENST0000035577**

RefSeq: NM\_012434

Type: missense\_variant

Impact: MODERATE

Exon: 2/11

cDNA: c.115C>T

Protein: p.Arg39Cys

Domain:

361 of 128874 variants passed filters.

**Circos Plot of single-sample analysis DX197989\_01**

**Expression Data of RNA2206898A1\_02**

gene_id	gene_name	gene_biotype	raw	tpm	cohort_mean	log2fc	zscore	pval	hpa_tissue_tpm	hpa_tissue
1	ENSG00000175879	HOXD8	protein_coding	14	8.25	1.21	2.57	2.08	0.012	5.400
2	ENSG00000178568	ERBB4	protein_coding	4	0.48	0.10	0.43	2.394	0.017	0.000
3	ENSG00000154803	FLCN	protein_coding	142	22.49	10.68	1.09	2.389	0.017	11.100
4	ENSG00000187098	MITF	protein_coding	36	7.52	3.71	0.91	2.315	0.021	5.000
5	ENSG00000205755	CRLF2	protein_coding	4	3.13	0.63	1.52	2.300	0.021	0.400
6	ENSG00000164736	SOX17	protein_coding	3	1.98	0.50	1.08	2.227	0.026	0.900
7	ENSG00000166923	GREM1	protein_coding	381	39.94	8.33	2.78	2.156	0.031	29.400
8	ENSG00000136634	IL10	protein_coding	5	1.78	0.74	0.72	2.021	0.043	1.500
9	ENSG00000119139	TIP2	protein_coding	20	2.03	7.02	-1.28	-2.032	0.042	38.300
10	ENSG0000065057	NTHL1	protein_coding	1	0.64	5.18	-1.73	-2.183	0.029	6.000
11	ENSG00000167985	SDHAF2	protein_coding	5	4.57	12.17	-1.16	-2.202	0.028	31.700
12	ENSG00000163930	BAP1	protein_coding	25	8.21	18.00	-0.98	-2.203	0.028	26.200
13	ENSG00000133056	PIK3CB	protein_coding	139	21.18	40.35	-0.85	-2.228	0.026	14.000
14	ENSG00000146232	NFKBIE	protein_coding	5	3.00	9.04	-1.24	-2.335	0.020	6.400
15	ENSG00000172996	MYD88	protein_coding	51	18.54	33.69	-0.79	-2.360	0.018	34.500
16	ENSG00000136936	XPA	protein_coding	1	0.85	12.33	-2.57	-2.413	0.016	10.800
17	ENSG00000122729	ACO1	protein_coding	50	10.26	15.31	-0.52	-2.453	0.014	22.400

Region & Intron

Cohort determination

- germline (same tissue)
- germline (same tissue and project)
- somatic (HPO/ICD based)
- custom cohort

Quality filter: Remove samples with quality 'bad'

Cohort size: 10 show cohort samples

Filter

Gene(s):

min. abs. logFC: 0,00

min. abs. t-score: 2,00

min. TPM (sample): 0,00

min. TPM (cohort): 0,00

low expression (TPM): 0,10

Biotype:

- IG C gene
- IG C pseudogene
- IG D gene
- IG J gene
- IG J pseudogene
- IG V gene
- IG V pseudogene
- Ig pseudogene
- Mt rRNA
- Mt tRNA
- TEC
- TD C gene

Select HPO terms

Phenotype browser (double-click to select)

heart morph

Abnormal cardiac atrium morphology

Abnormal cardiac septum morphology

Abnormal heart morphology

Abnormal heart valve morphology

Abnormal left ventricle morphology

Calced amorphous tumor of the heart

Selected phenotypes (double-click to remove)

HP:0000278 - Retrognathia

HP:0001511 - Intrauterine growth retardation

HP:0001629 - Ventricular septal defect

HP:0006711 - Aplasia/Hypoplasia involving bones of the thorax

HP:0012165 - Oligodactyly

HP:0001627 - Abnormal heart morphology

**Main filters**

Default filter: CHINCH default

- CHI size > 10.00 kB
- CHI log likelihood > 12.00 located by region
- CHI q-value < 0.05
- CHI co-hybridization region < 0.95 (column overlap)
- CHI CMM genes
- CHI gene overlap complete, exon/splicing
- CHI compound-heterozygote
- CHI gene overlap incomplete

**Target region filters**

none

Region

Phenotypes

Text filters

**Classification NGSD KEP: 3,4,5,M**

# Generating Diagnostic Reports for Long-Read Diagnostics

**GSvar - Variant View**

File Edit Tools Variants IGV Help

chr start end ref obs genotype filter quality gene variant\_type ing\_and\_splic plate OMIM ClinVar

8951 chr2 179454383 179454383 C T het QUAL=175;DP=... TTN;TTN-AS1 missense,intron&non\_coding,transcript TTNENST0... 188840 [...]

8963 chr2 179497051 179497051 C G het QUAL=2015;DP=... TTN;TTN-AS1 missense,intron&non\_coding,transcript TTNENST0... 188840 [...]

9020 chr2 179637874 179637874 G A het QUAL=1040;DP=... TTN;TTN-AS1 missense,intron\_coding,transcript\_exon... TTNENST0... 188840 [...]

11367 chr3 48716816 48716816 G A hom QUAL=2970;DP=... NCKIPSD missense NCKIPSD...

11482 chr3 52561936 52561936 G T hom QUAL=6642;DP=... NTSDC2 missense NTSDC2...

11484 chr3 52822326 52822326 G C hom QUAL=3787;DP=... JTH1 missense,3'UTR JTH1-ENST...

16394 chr5 31299727 31299727 G A hom QUAL=2127;DP=... CDH6 missense CDH6-ENST...

17908 chr5 140167670 140167670 G A het QUAL=4205;DP=... PCDHA1 missense,intron PCDHA1-ENST...

17930 chr5 140215943 140215943 G C het QUAL=3831;DP=... PCDHA7;PCD... missense,intron PCDHA7-ENST...

**17948 chr5 140264097 140264097 G A het QUAL=2053;DP=... PCDHA13;PCD... stop\_gained,intron PCDHA13-ENST...**

22691 chr6 161032744 161032744 T A het QUAL=184;DP=... LPA missense LPA-ENST... 152200 [...]

**22692 chr6 161032745 161032745 C A het QUAL=184;DP=... LPA stop\_gained LPA-ENST... 152200 [...]**

23924 chr7 65617255 65617255 G A hom QUAL=4989;DP=... CRCP;RPS-11 missense,intron&non\_coding,transcript CRCP-ENST... reg...

23987 chr7 73477545 73477545 G A hom QUAL=31616;DP=... ELN missense ELN-ENST... 130160 [...] 424293 [Uncertain significance DIS...

24118 chr7 82451900 82451900 T C hom QUAL=7795;DP=... PCLO missense PCLO-ENST... 604918 [...] 218822 [likely benign DISEASE-not...

27054 chr8 145579194 145579194 G C het QUAL=1180;DP=... FBXL6;GSI-398... missense,misense&NMD\_transcript FBXL6-ENST... reg...

27655 chr8 145581831 145581831 G A het QUAL=1620;DP=... FBXL6 missense FBXL6-ENST... reg...

34320 chr11 55032500 55032500 C T hom QUAL=55;DP=... TRIM4 missense TRIM4-ENST...

**34408 chr11 57477387 57477387 G A hom QUAL=6087;DP=... CLP1 splice\_region&intron missense CLP1-ENST... 608757 [...] 143934 [likely pathogenic DISEASE...**

34726 chr11 63560822 63560822 C T hom QUAL=10491;DP=... KCNQ7 intron,missense,3'UTR KCNQ7-ENST... reg...

35018 chr11 72423352 72423352 G A hom QUAL=3336;DP=... ARHGEF17 missense ARHGEF17-ENST... reg...

35024 chr11 73074799 73074799 G A hom QUAL=10720;DP=... ARHGEF17 missense ARHGEF17-ENST...

35048 chr11 73796830 73796830 T C hom QUAL=5320;DP=... C2CD3 missense C2CD3-ENST... 615944 [...]

39092 chr12 121600322 121600322 C T hom QUAL=3025;DP=... P2RX7 intron,missense&splice\_region,splice\_re... P2RX7-ENST...

39128 chr12 122670741 122670741 C T hom QUAL=3181;DP=... LRRC43 missense,5'UTR LRRC43-ENST...

49203 chr17 19449752 19449752 - TGC... hom QUAL=2261;DP=... SLC47A1 inframe\_insertion,intron,5'UTR SLC47A1-ENST...

**Variant details**

**chr11:57477387-57477387 G>A (hom)**

Gen(s): CLP1 (In = 0.01, pl = 0.91)

Quality: QUAL=6087 DP=199 AF=0.98 MQM=60

**Filter:**

CLP1 ENST00000302231 (1/4) < Splicing

Type: Intron\_Region\_Variant,intron\_variant

Impact: LOW

Exon: 2/2

cDNA: c.414+5G>A

Protein: COSMIC: CSIGIC: Regulatory:

Regulatory: COSMIC: CSIGIC:

Frequencies

Databases

Allele: rs37777616 ClinVar: 143934

Gene/Symbol: CLP1 HGMD: OMIM: 608757

dbSNP: rs4444481

Phenotype: Phenotype: PolyPhen: 0.91(0.91)(0.91)(0.91)

gnomAD (hom): 0.0001 CADD: 26.40

gnomAD (sub): 0.0000,0.0001,0.0000,0.0001,0.0000

ESP (sub): n/a

fathmm-MKL: 0.99,0.99 REVEL: 0.79

ICGC: NSGO

Tübingen: 25 genetic analysts use GSvar to identify causal variants and to generate reports

**RNA-FFPE\_Auswertung | Befund-ID: 472630**

chromosome 7 p11.2

breakpoint1 chr7:5501985 6582

breakpoint2 chr7:55155830 28

Coverage

EGFR ENST00000275493

EGFR ENST00000275493

6 kbp intron not scale

**ANLAGEN**

Potentiel relevante somatische Veränderungen:

Punktmutationen (SNVs) und kleine Insertionen/Deletionen (INDELs) (RNA2204787A1\_01-DNA2204785A1\_01-DNA2203877A1\_01)

Gen	Veränderung	Typ	Anteil	Anteil	Tumorsequenz	Normalprobe	Normalsequenz	Veränderung (x-fach)
EGFR	c.787A>C:p.Thr263Pro	missense	0.78	0.71	1533	22	671	2.3
MSH6	c.808A>T;p.Lys270Ter	stop_gained	0.40	0.70	12	21	15	0.8
TERT	c.-124C>T	upstream_gene	0.85	n/a	3	0	2	-
TET2	c.3511A>T;p.Lys117Ter	stop_gained	0.41	0.00	56	5	35	1.6

SUPPORTING READ COUNT

Split reads at breakpoint1 = 0  
Split reads at breakpoint2 = 3  
Discordant mates = 0

**Kopienzahlveränderungen (CNVs)**

alprobe	Bewertung	Normalsequenz	Normalprobe	Normalsequenz	Veränderung
PM	TPM	MW TPM	x-fach		
16	1	34	1.4		
4	1	130	2.5		
22	1	671	2.3		
10	1	198	0.7		
6	1	22	1.3		
38	2	33	1.1		
101	2	20	1.1		
22	2	53	0.8		
3	3	3	-		
2	3	15	-		
3	3	8	-		
3	3	9	-		
10	3	17	-		
0	3	1	-		
3	3	8	-		

Universitätsklinikum Tübingen

Institut für Medizinische Genetik und Angewandte Genomik  
Ärztlicher Direktor Prof. Dr. med. Olaf Rieß  
Med. Versorgungszentrum des UKT Fachgebiet Medizinische Genetik  
Hoppe-Seyler-Straße 3 72076 Tübingen

MVZ Institut für Medizinische Genetik und Angewandte Genomik Calwerstraße 7 - 72076 Tübingen

Frau

72076 Tübingen

Seite 1 von 5

**Bericht zur somatischen Tumordiagnostik (NGS Transkriptom-Analyse)**

Patient: Berichtsdatum: 07.07.2022

Klinische Angaben: Tumorsequenz: Auftrags-ID: 47390

Auftrag vom: 13.05.2022 Probeneingang: 17.06.2022

Probenfreigabe: 17.06.2022 Auftragsfreigabe: 17.06.2022

Sehr geehrte Frau

wir bedanken uns für die Anforderung einer molekulargenetischen Diagnostik bei [REDACTED] In der nachfolgenden Übersicht finden sich Expressionsdaten zu den bereits identifizierten therapierelevanten somatischen Veränderungen in derselben Tumorsequenz (s. unsere Befund-ID 47263). Die Expression bestimmter Gene aus therapierelevanten Signalkaskaden ist im Anhang dargestellt. Weitere Expressionsdaten können auf Wunsch elektronisch zur Verfügung gestellt werden.

**ZUSAMMENFASSUNG**

- Die Expression der therapierelevanten somatischen Varianten in EGFR und MSH6 wurde bestätigt. Splice- oder trunkierende Varianten können zu RNA-mediated decay führen und werden oft im Transkriptom nicht sicher erfasst.
- Es zeigte sich ein Hinweis auf die onkogene Variante EGFRvIII. Eine Expression der unklaren Translokation zwischen MTHFR und ATP6V1E2 wurde nicht nachgewiesen.
- Eine differenzielle Expressionsanalyse im Bezug zu einer Normalprobe (Großhirnrinde) zeigt eine Überexpression von EGFR (70-fach), CDK6 (83-fach) und MDM4 (15-fach).
- Die Bewertung der Expression bestimmter Gene aus therapierelevanten Signalkaskaden deuten auf die Aktivierung des CDK4/6 Signalweges über CCND2 und CDK6.
- Das Expressionsprofil deutet auf eine Zugehörigkeit zur Gruppe der in-house Glioblastom-Patienten hin.

AP 2021\_12\_167-g25b96bc  
2022\_04\_165-g8f8be95  
3570 %  
X  
X  
next Ultra II Directional RNA Library Prep  
5000 als 2x75 bzw. 2x100 bp paired-end  
Die Aufbereitung der Daten erfolgte  
erte Genexpressio, angegeben in TPM  
n wurden mit der Software STAR-Fusion

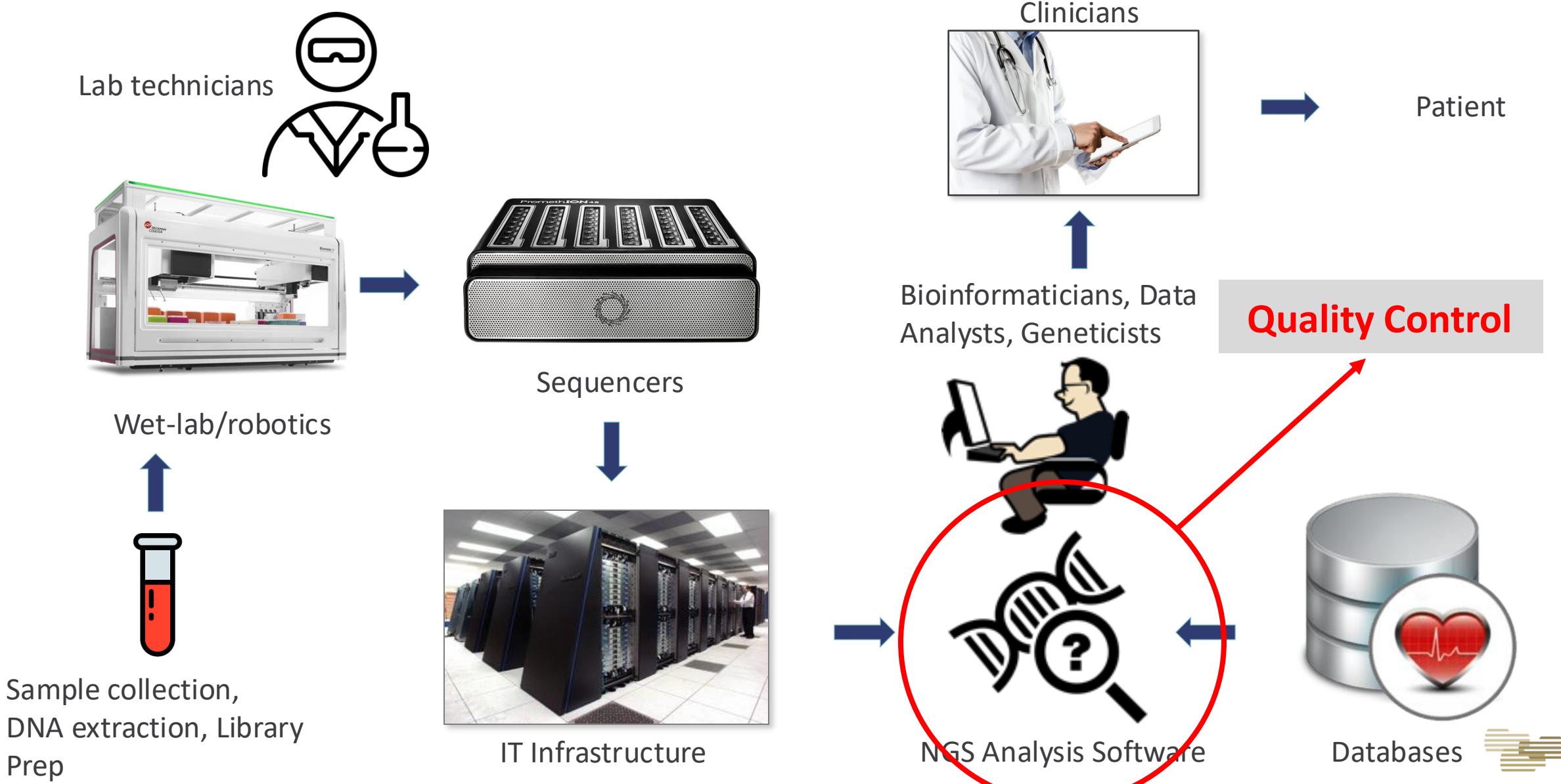
CNVs (Copy Number Variations) wurden bereits in  
erungen wurden mit der hauseligen  
er Varianten wurde nach den Variant  
/one\_path\_snp/).

kommt, wenn vorhanden, eine Angabe  
Human Protein Atlas. Vergleichende

Seite 5 von 5

Seite 3 von 5

# Accreditation of NGS Laboratory Developed Tests



# IonGER\* - Clinical Long-read Genome Initiative



- National German initiative to evaluate the clinical applications of Nanopore sequencing for rare disease
- Sequencing will be done at 4 university medical centers across Germany
- Relies on a clinically well characterized multi-center cohort of rare disease patients
- 1000 samples in a 2-year period + 80 Pilot samples
- Aims:
  - Demonstration of added diagnostic value
  - Proof of sustainability in clinical practice
  - Benchmarking and accreditation
  - Establishing „wet lab“ SOPs and bioinformatic pipeline for streamlined enrollment in other university medical centers



\*IonGER Study is supported by ONT



# IonGER Pilot Study: Overview



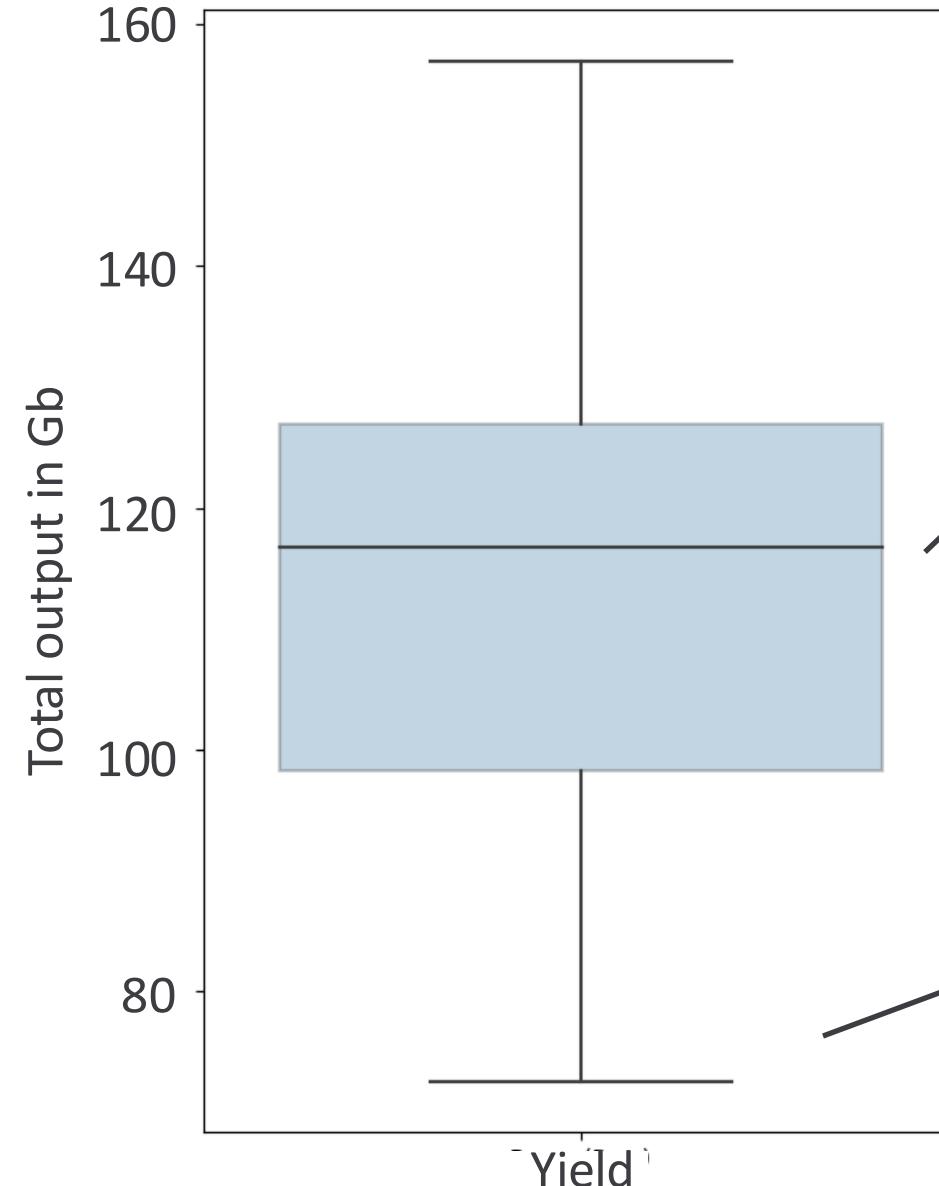
- 2 Genome in a Bottle (**GiaB**) cell lines sequenced at each site (total of 12 FC)
- 68 patient samples (1 FC per genome)
- R10.4.1 with V14 chemistry
- Clinical use cases
  - Repeat Expansions (e.g. FGF14, SCA03 ...)
  - Low complexity regions (e.g. RPGR, KANSL1)
  - Pseudogenes (e.g. SMN1/SMN2)
  - Pathogenic haplotypes (e.g. OPN1-Cluster)
  - Imprinting disorders (e.g. FMR1, SNRPN)





# Yield per Flowcell

40x Coverage  
30x Coverage

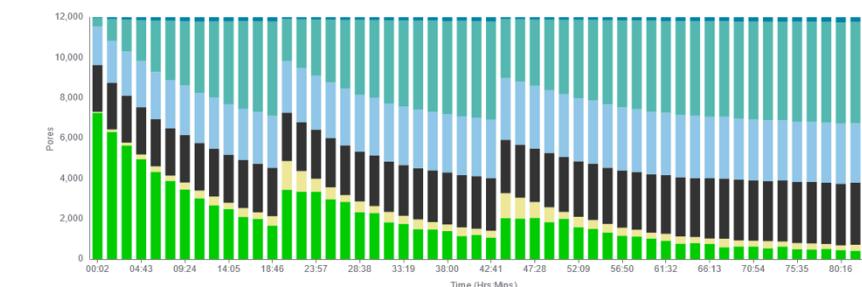
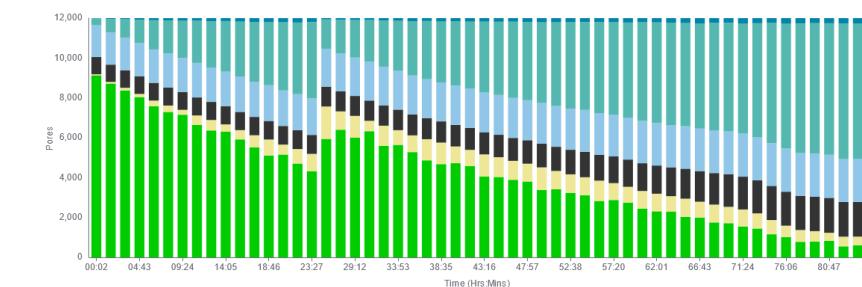


## PORE SCAN

A Pore scan is performed at configurable time intervals to determine the current status of pores within channels on a Flow Cell. For this run a Pore scan is performed every 1.5 hrs.

### Legend

- ▲ Single Pore Pore in channel available for sequencing
- Reserved Pore Pore in reserve, will return to available when required
- Saturated Possible contamination in the sample
- Zero No current is passing through this pore, possibly due to bubbles on the membrane
- Unavailable Pore inhibited from sequencing
- Inactive Pore no longer suitable for further sequencing





# Quality Control for Nanopore Sequencing

Sample	N50 – Pass (bp)	Bases – Pass (Gb)	avg. read depth	20x coverage (%)	Runtime: Analysis
TUEB_04	19603	121.5	37.24	99.80	30:57
TUEB_05	11907	111.1	33.89	98.93	15:39
TUEB_06	15948	91.4	27.91	90.81	24:50
TUEB_07	18110	113.9	34.82	99.37	31:22
TUEB_08	10611	146.7	44.95	99.99	30:35
TUEB_09	14586	98.2	30.03	96.42	21:49
TUEB_10	13570	98.2	29.87	96.11	24:09
TUEB_11	13767	98.1	29.98	95.98	17:37
TUEB_12	14302	131.0	40.24	99.92	29:57
TUEB_13	9931	126.9	38.30	99.79	25:21
TUEB_14	9943	72.6	22.27	73.41	21:07
TUEB_15	18946	157.0	48.20	99.99	32:52
TUEB_16	12949	85.9	26.23	87.03	19:56
TUEB_17	14697	121.4	37.16	99.58	29:19
TUEB_18	15428	123.7	38.00	99.60	24:50
TUEB_19	14479	150.1	45.92	99.98	31:27
TUEB_20	10428	116.7	35.59	99.46	29:48
Average	<b>13725.1</b>	<b>115.2</b>	<b>35.6</b>	<b>96.0</b>	<b>~26h</b>





# Genome in a Bottle Reference Data for Germline NGS

NIST

Search NIST



Menu

PROJECTS/PROGRAMS

## Genome in a Bottle

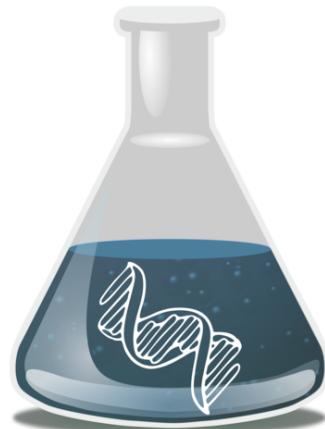
### Summary

Consortium hosted by NIST dedicated to authoritative characterization of benchmark human genomes. Sign up for [General GIAB](#) and [Analysis Team](#) email lists. [Public workshops](#) held annually - next workshop will be rescheduled after COVID-19. Interested in job opportunities with us? Contact Justin Zook at the email in the right panel.  
[Click here for the GIAB FAQ](#)

### DESCRIPTION

#### Consortium goals:

The Genome in a Bottle Consortium is a public-private-academic consortium hosted by NIST to develop the technical infrastructure (reference standards, reference methods, and reference data) to enable translation of whole human genome sequencing to clinical practice and innovations in technologies. The priority of GIAB is authoritative characterization of human genomes for use in benchmarking, including analytical validation and technology development, optimization, and demonstration.



#### Reference samples:

GIAB has currently characterized a pilot genome (NA12878/HG001) from the [HapMap project](#), and two son/father/mother trios of Ashkenazi Jewish and Han Chinese ancestry from the [Personal Genome Project](#) (selected because, unlike the pilot genome, they are consented for commercial redistribution). These samples and their IDs from [NIST](#), [CoreLife](#), and [PGP](#) are in [this table](#).

#### Benchmark (or "High-confidence") variant calls and regions:

We developed an integration pipeline to utilize sequencing data generated by multiple technologies to generate variant calls and regions for use in benchmarking and validating variant calling pipelines. Currently, benchmark VCF and BED files for small variants are available for GRCh37 and GRCh38 under each genome at <https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/>

GIAB's versions of GRCh37 and GRCh38 reference fasta files, including a new GRCh38 reference in collaboration with the GRC that masks false duplications in GRCh38, are at <https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/references/>

#### New benchmarks for difficult variants and regions:

Structural variants: [Currently available for HG002 on GRCh37](#) and in Challenging Medically Relevant Gene benchmark below

### ORGANIZATIONS

Material Measurement Laboratory  
Biosystems and Biomaterials Division  
Biomarker and Genomic Sciences Group

### NIST STAFF

Justin Zook  
Nathanael David Olson  
Justin Wagner  
Nathan Dwarshuis  
Megan Cleveland  
Lindsay Harris  
Sierra Miller

### FORMER STAFF

Marc Salit  
Jennifer McDaniel  
Lesley Chapman

### CONTACT

Justin Zook  
[justin.zook@nist.gov](mailto:justin.zook@nist.gov)  
(301) 975-4133

### DATES

Started: August 2012

### PROJECT STATUS

ONGOING

## Reference samples:

GIAB has currently characterized a pilot genome (NA12878/HG001) from the [HapMap project](#), and two son/father/mother trios of Ashkenazi Jewish and Han Chinese ancestry from the [Personal Genome Project](#)

## Benchmark (or "High-confidence") variant calls and regions:

We developed an integration pipeline to utilize sequencing data generated by multiple technologies to generate variant calls and regions for use in benchmarking and validating variant calling pipelines. Currently, benchmark VCF and BED files for small variants are available for GRCh37 and GRCh38 under each genome at <https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/>

## New benchmarks for difficult variants and regions:

Structural variants: [Currently available for HG002 on GRCh37](#) and in Challenging Medically Relevant Gene benchmark below

Small variants in more difficult regions: [v4.2.1 is available for all 7 GIAB samples on GRCh37 and GRCh38 \(manuscript\)](#).

MHC: Included in [v4.2.1 small variant benchmark for HG001-HG007 \(Manuscript describing MHC benchmark\)](#)

[273 Challenging Medically Relevant Genes small variant and SV benchmarks in HG002](#) and [Preliminary benchmark for T2T-CHM13v1.0](#)

# GiaB Technology

NA12878:HG001 <https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/>

<https://www.nist.gov/programs-projects/genome-bottle>

[Published: 08 July 2015](#)

## Genome in a bottle—a human DNA standard

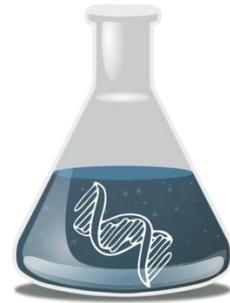
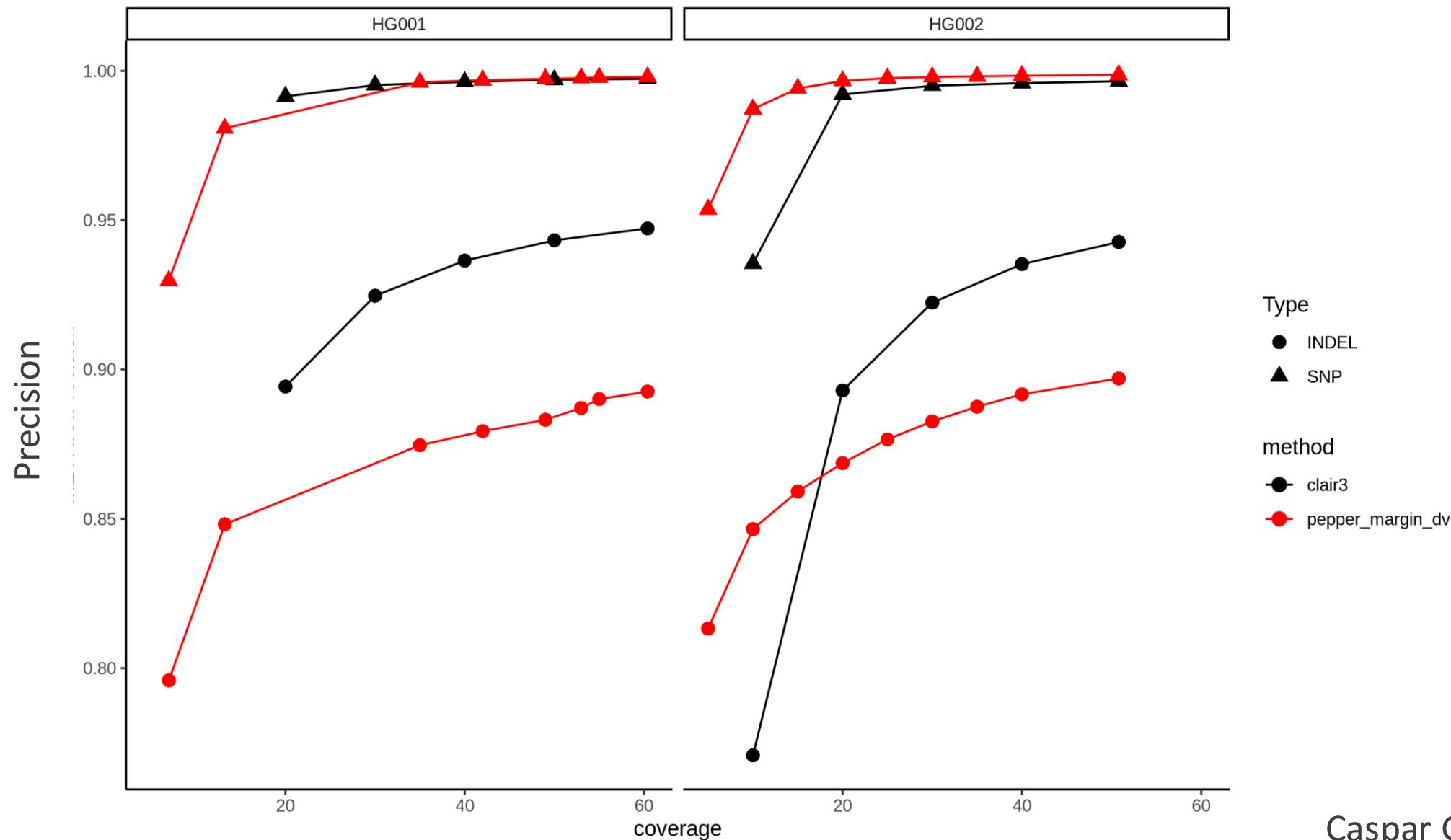
[Nature Biotechnology](#) 33, 675 (2015) | [Cite this article](#)

4171 Accesses | 3 Citations | 18 Altmetric | [Metrics](#)

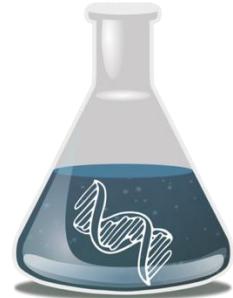
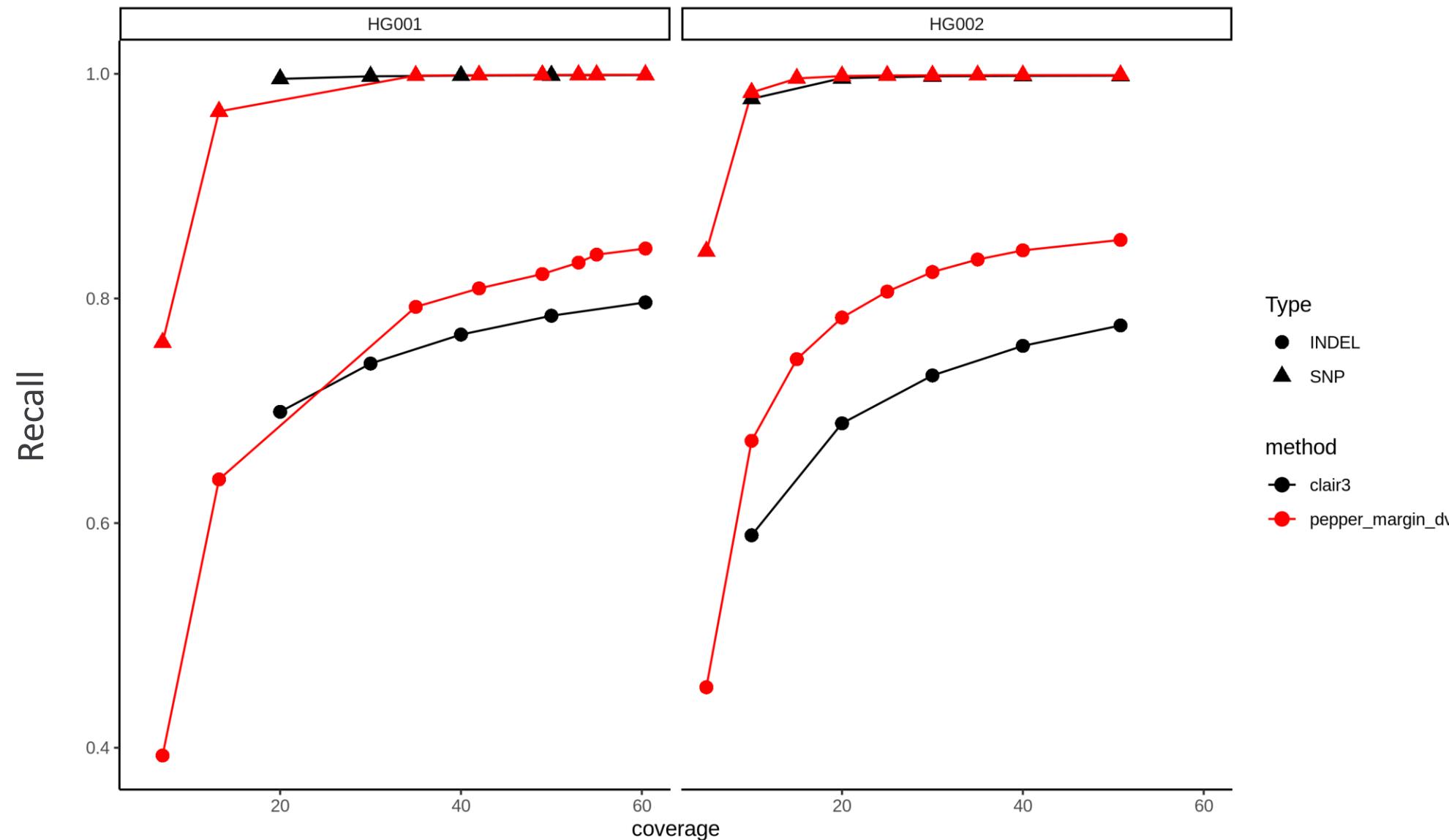
Sequencing Platform	Sequence	Alignment
Illumina WGS 2x150bp 300X	HG001	bwamem: <a href="#">HG001:hg19 (downsampled30x)</a> novoalign: <a href="#">HG001</a>
Illumina HiSeq Exome	HG001 <a href="#">HG001:trimmed_fastq</a>	bwamem: <a href="#">HG001:hg19</a>
Illumina TruSeq Exome		bwamem: <a href="#">HG001:hg19</a>
10X Genomics		bwamem: <a href="#">HG001:hg19</a> bwamem: <a href="#">HG001:hg19 (size_selected)</a>
10X Genomics ChromiumGenome		LongRanger2.0: <a href="#">HG001:hg19-hg38</a> LongRanger2.1: <a href="#">HG001:hg19-hg38</a>
CompleteGenomics		CGAtools: <a href="#">HG001:hg19</a>
Ion Proton 1000x Exome		TMAP: <a href="#">HG001:hg19</a>
NA12878 SOLiD5500W		LifeScope: <a href="#">HG001:hg19</a>
BGI BGISEQ500	sequence.index	alignment.index
BGI MGISEQ	sequence.index	alignment.index
BGI stLFR	sequence.index	alignment.index
PacBio 40x	<a href="#">HG001:hdf5</a>	
PacBio SequelII CCS 11kb	sequence.index	alignment.index
Ultralong_OxfordNanopore	-	minimap2: <a href="#">HG001</a>



# SNV and InDel Precision



# SNV and InDel Recall



Type

- INDEL
- ▲ SNP

method

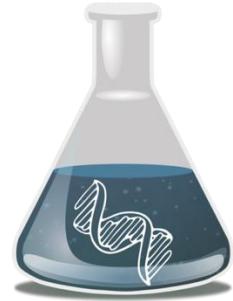
- clair3
- pepper\_margin\_dv





# Higher InDel Quality in Coding Regions

Genome-in-a-Bottle Reference NA12878 (HG001), 40x coverage



## SNV/INDEL validation (whole genome):

	recall/sensitivity	precision	genotyping accuracy
SNVs	0.999	0.999	0.999
INDELS	0.794	0.966	0.998

## SNV/INDEL validation (only protein-coding exons):

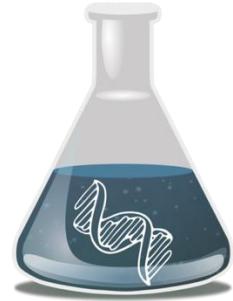
	recall/sensitivity	precision	genotyping accuracy
SNVs	1.000	0.999	1.000
INDELS	0.936	0.978	0.998





# How much Nanopore read coverage is needed?

Genome-in-a-Bottle Reference NA12878 (HG001)



SNV/INDEL validation (only protein-coding exons):

Coverage	SNV			INDEL		
	sensitivity	Precision	genotyping	sensitivity	Precision	genotyping
50x	0.999	0.996	1.000	0.940	0.967	0.998
40x	0.999	0.994	1.000	0.932	0.944	0.996
30x	0.999	0.990	0.999	0.921	0.935	0.998
20x	0.998	0.989	0.999	0.900	0.848	0.996
10x	0.986	0.888	0.994	0.812	0.494	0.998

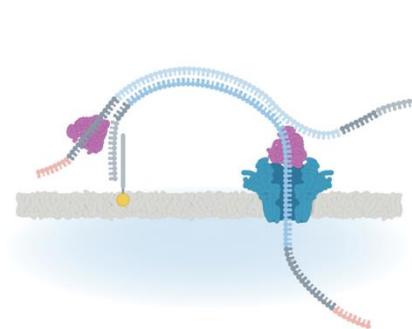
Can we further improve InDel calling: Duplex Reads!



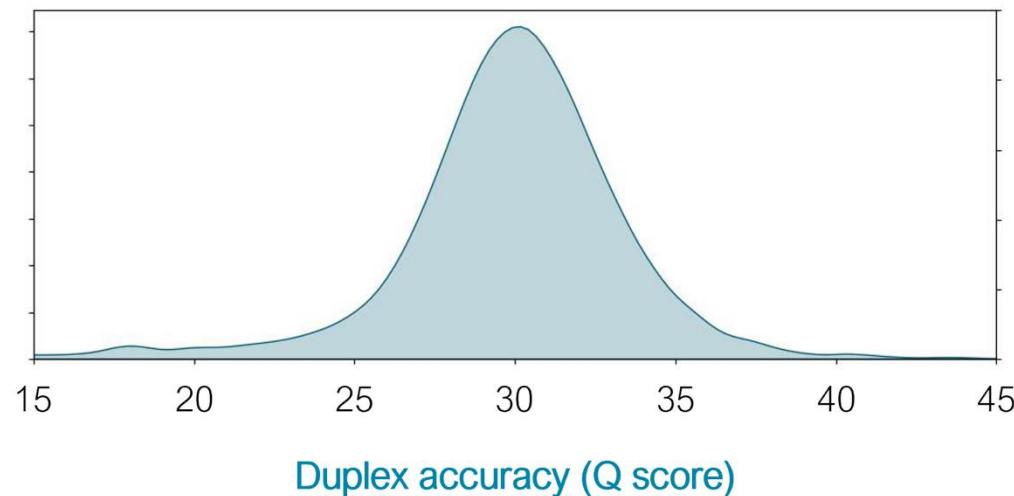
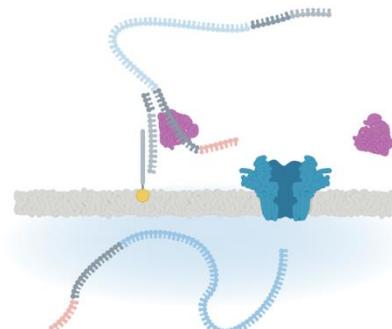


# Duplex Reads

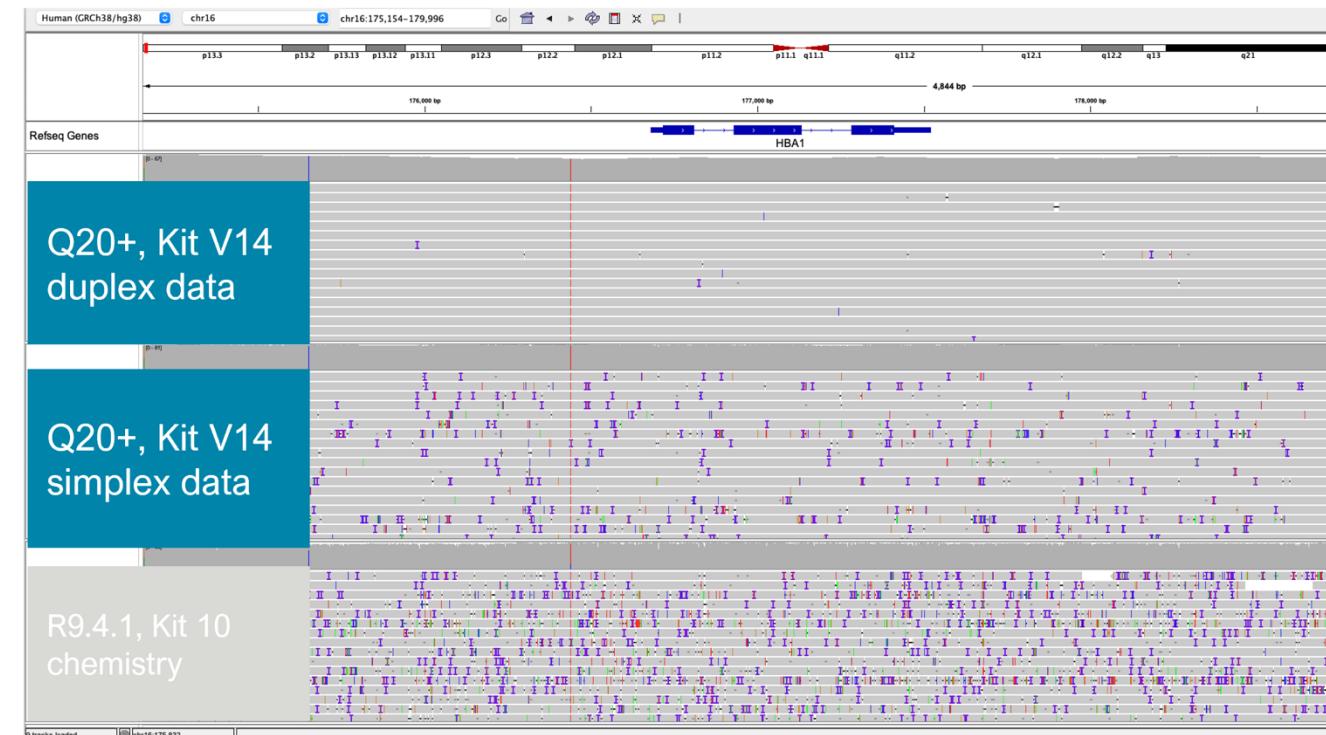
Linear dsDNA molecule adapted on both ends and first strand sequenced



Second strand captured and sequenced subsequently



## Alignments in IGV



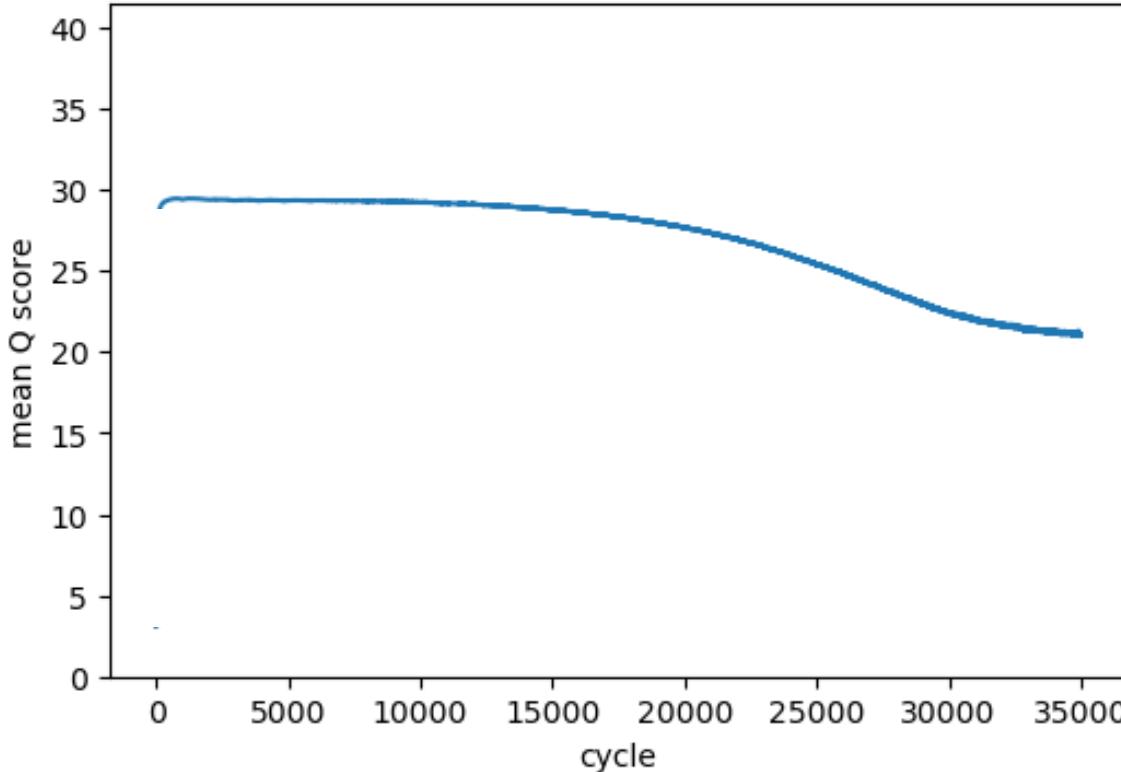
<https://nanoporetech.com/q20plus-chemistry>



# Duplex Read Quality of Sample HG001 (NA12878)

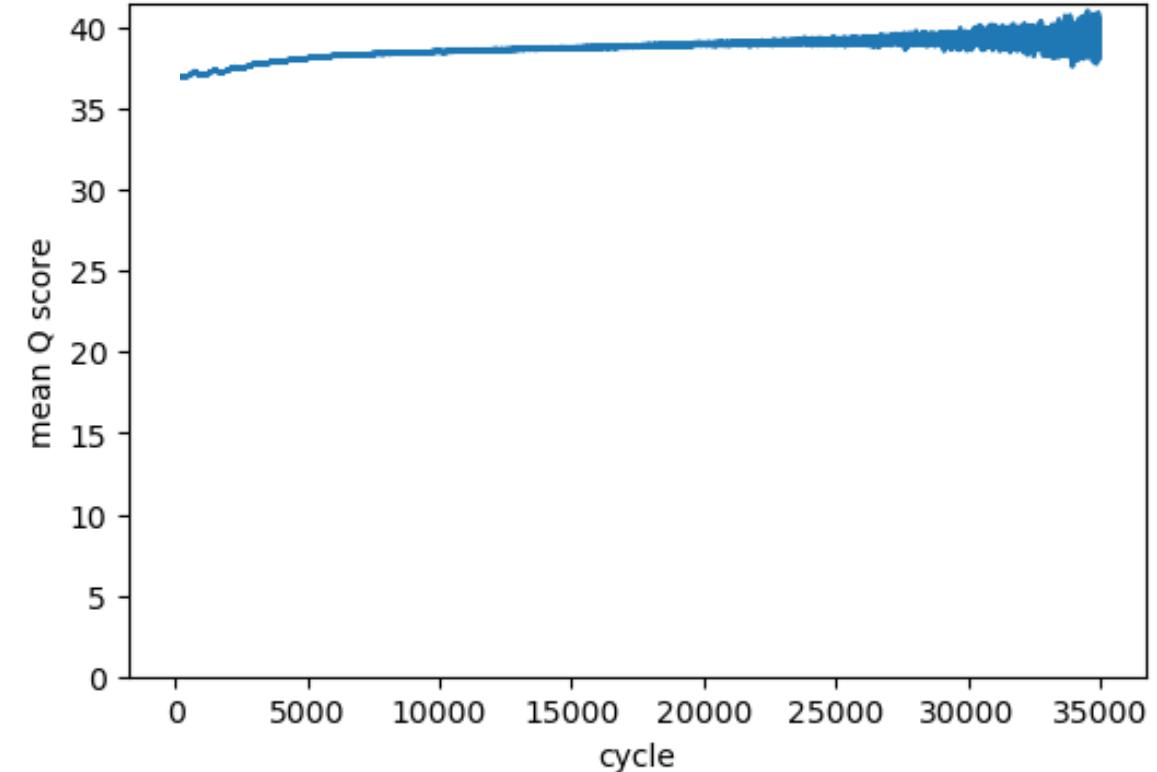


Simplex Reads



- Q20 read percentage: 97.09
- Q30 read percentage: 58.73

Duplex Reads



- Q20 read percentage: 99.96
- Q30 read percentage: **89.68**



# Duplex Reads Increase Quality



	Simplex	Duplex
Error Rate	2.03%	0.54%
Mismatch Rate	1.06%	0.32%
Insertion Rate	0.45%	0.12%
Deletion Rate	0.52%	0.10%
Soft-clip Rate	99.85%	12.20%

TODO: test improvement for indel calling by pooling duplex reads from multiple flowcells (HG001, HG002)





# Haplotype Aware Error Correction?

11:50 4G+ 🔋

Suche

**Mile Sikic · 2.**  
AI in Genomics Lab at Genome Institute ...  
20 Std.

+ Folgen

Welcome to the era of Q30+ (accuracy > 99.9%) long nanopore reads. Error correction method Herro, developed by [Dominik Stanojević](#) and thoroughly tested by [Dehui Lin](#) with support from [Sergey Nurk](#) and [Paola Florez de Sessions](#), increases the accuracy of R9.4.1 and R10.4.1 reads up to two orders of magnitude.

The image below shows the improvement in accuracy for different datasets (CHM13 - R9.4.1).

The link to GitHub is in the comment.

Before/after Concordance QV by dataset

Dataset	Before (Median)	After (Median)
HG002	~18	~38
I002C	~18	~38
CHM13	~12	~35
Arabidopsis	~22	~40
Zebrafish	~18	~30

Product Solutions Open Source Pricing Search or jump to

[lbcn-sci/herro](#) Public Notif

Code Issues 6 Pull requests 2 Actions Projects Security Insights

main ▾ 6 Branches 0 Tags Go to file Code ▾

dominikstanojevic Update README.md 649afcb · last week 187 Commits

resources	Added 2-bit sequence encoding	last year
scripts	refactoring & formating	last month
src	mak -c not required parameter	2 months ago
.gitignore	First commit	2 years ago
Cargo.toml	Prepared repo for public release	3 months ago

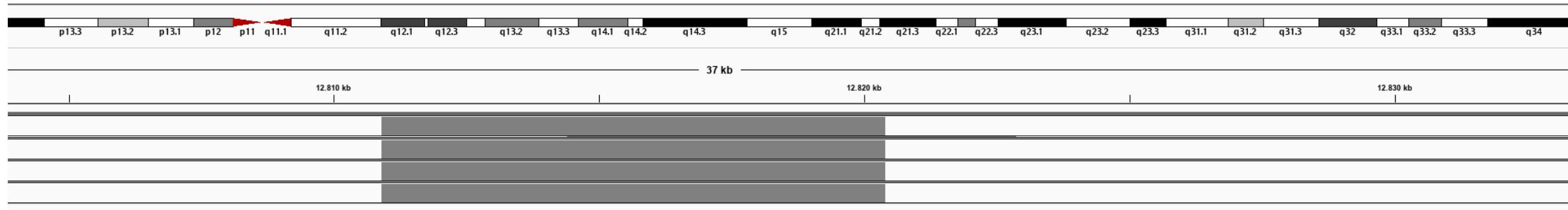
## HERRO

HERRO (Haplotype-aware ERRober cOrrection) is a highly accurate, haplotype-aware, deep-learning tool for error correction of Nanopore R10.4.1, Kit 14 reads (length of  $\geq 10000$ bp is recommended). An experimental model for R9.4.1 data is also provided for download.

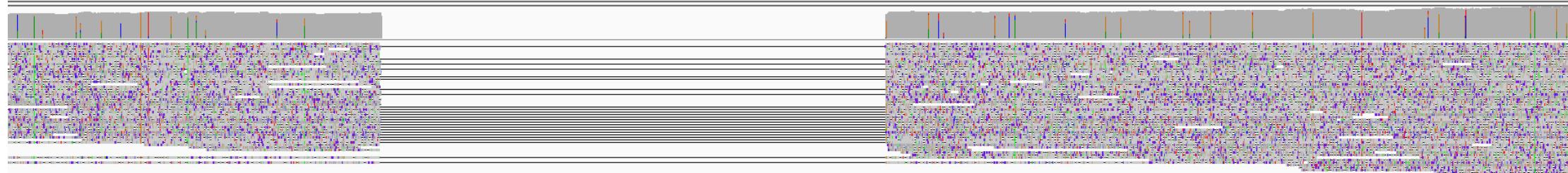


# Structural Variant Detection

BioNano\*



Nanopore



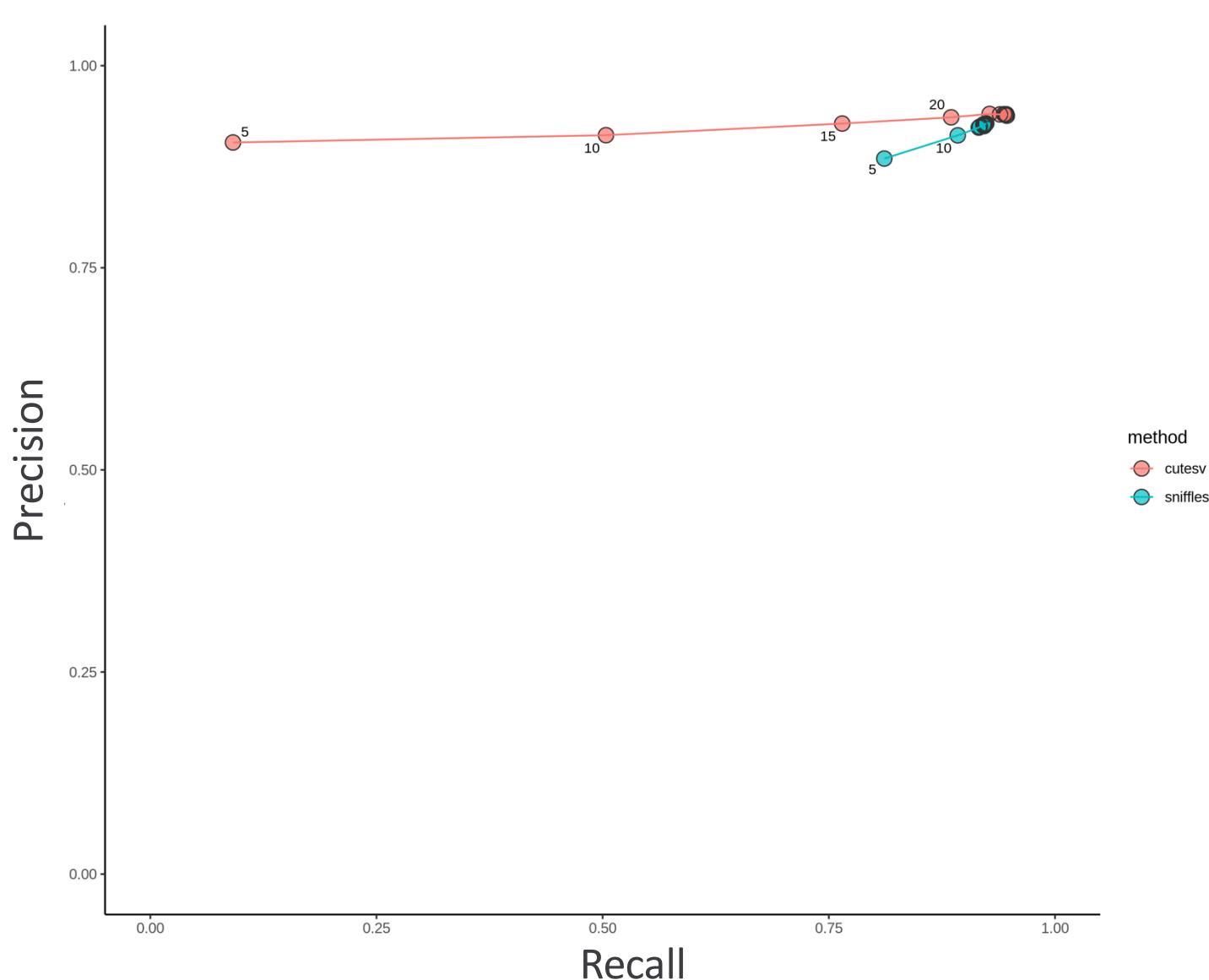
Short Read



\*Bionano: optical genome mapping

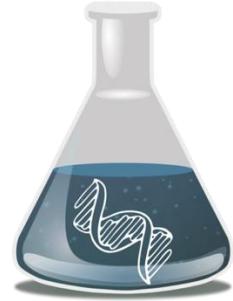


# Structural Variants: High Recall and Precision



Sniffles with 50x:

- Precision: 92.2
- Recall: 91.8



CuteSV with 50x:

- Precision: **93.8**
- Recall: **94.7**

Sniffles with 5x:

- Precision: 81.1
- Recall: 84.7

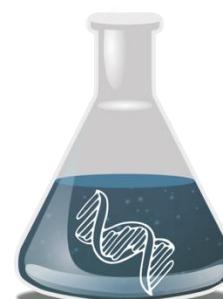


# Structural Variants with Sniffles



Coverage	True Pos	False Neg	False Pos	Precision	Recall
50x	17286	757	670	0.9219	0.9176
40x	17317	735	661	<b>0.9223</b>	<b>0.9187</b>
30x	17296	735	662	<b>0.9204</b>	<b>0.9169</b>
20x	17122	846	656	0.9162	0.9070
10x	16149	1474	676	0.8935	0.8557

- 30-40x coverage is sufficient for high quality SV calling with sniffles
- At > 40x coverage cuteSV outperforms sniffles
- With short reads, recall is less than 50%
- Genome-wide: ~10.000 SV with short and 23.000 SV with long reads



# Examples for Improved Clinical Diagnostics



1. **Causal Compound Heterozygotes** when sequencing only the index case (thanks to Haplotype Phasing)

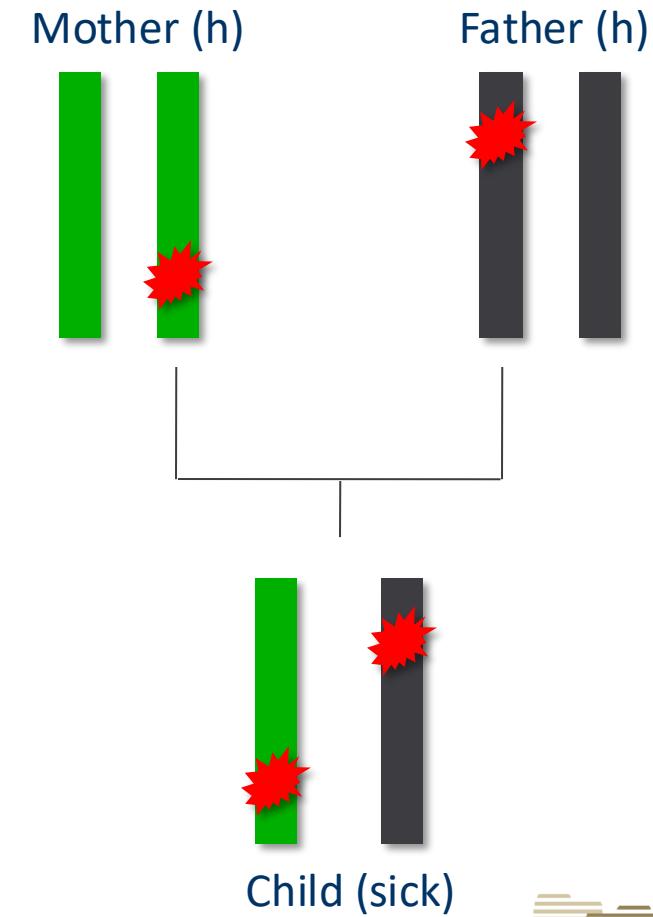
2. **Structural variants and Mobile elements:** doubling the recall for SVs

3. **Duplicate gene analysis** (or other ‘dark regions of the genome’)

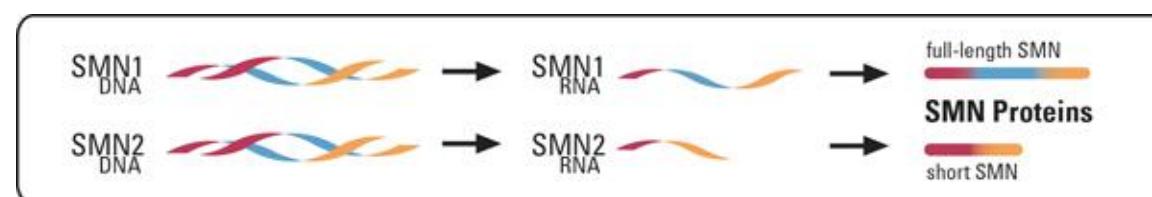
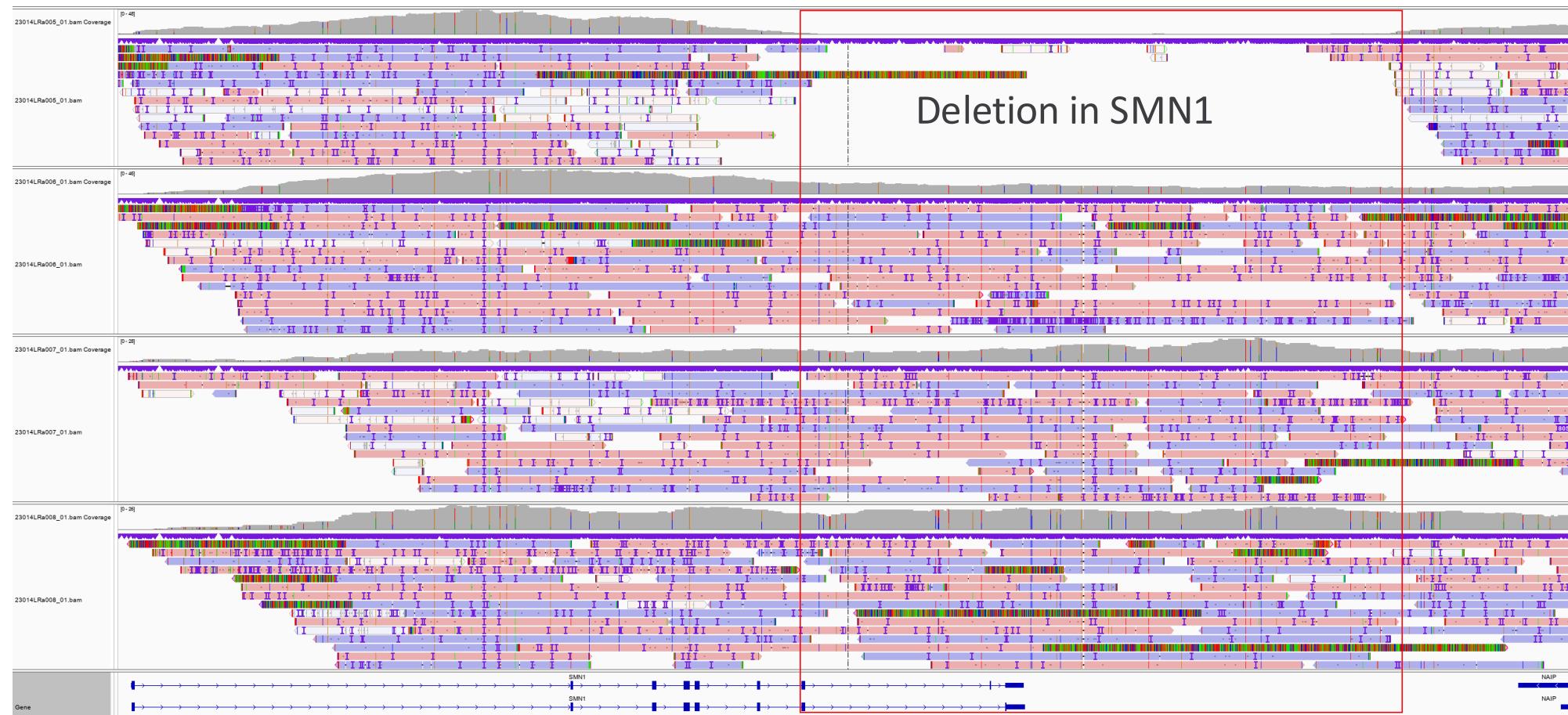
4. **Repeat Expansions**

5. **DNA Methylation, Imprinting**

Compound Heterozygotes  
in recessive diseases



# Duplicate Genes: SMN1 and SMN2 in Spinal Muscular Atrophy



<https://www.mda.org/disease/spinal-muscular-atrophy/causes-inheritance>

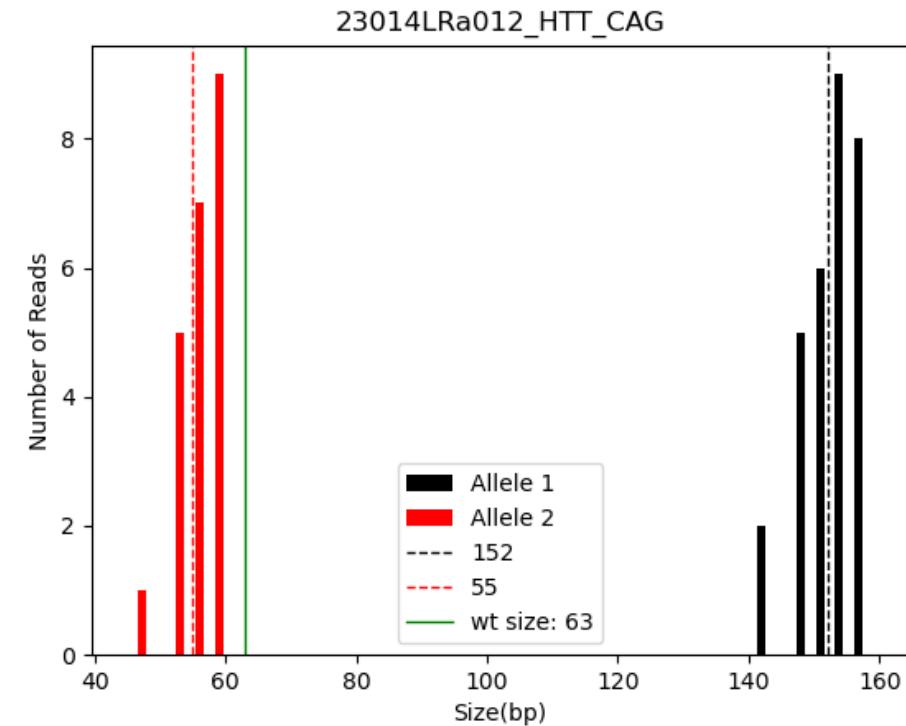
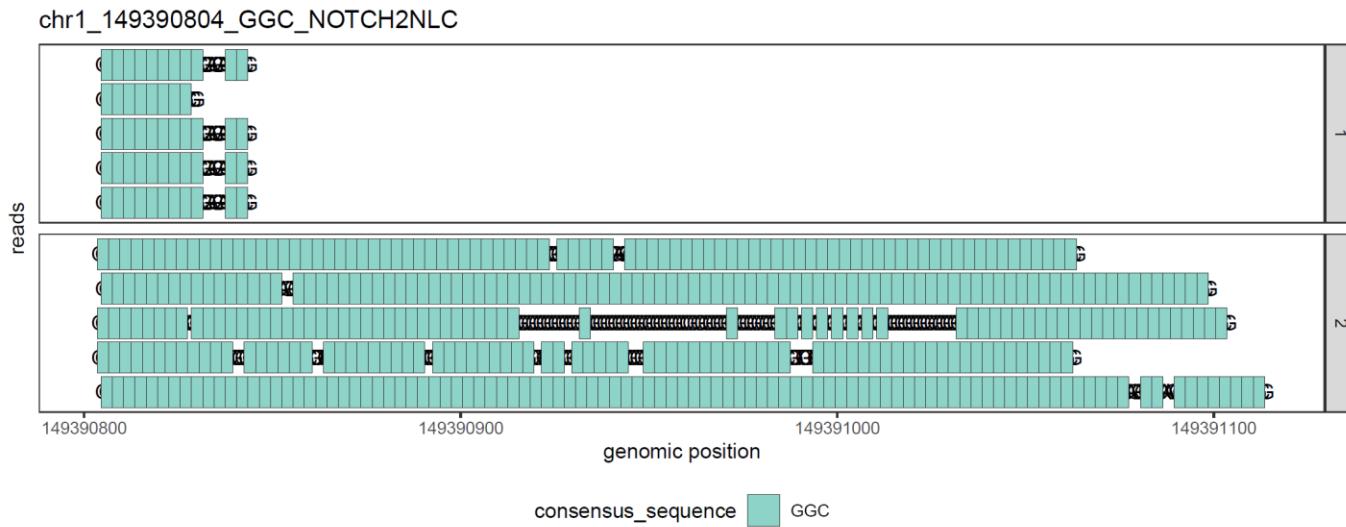


# Repeat Expansion Detection



## Algorithm:

- Phase repeat haplotypes of cases and controls
  - Generate consensus repeat-length for each haplotype
  - Measure minimum length causing disease
  - We currently use the Straglr tool with in-house visualization

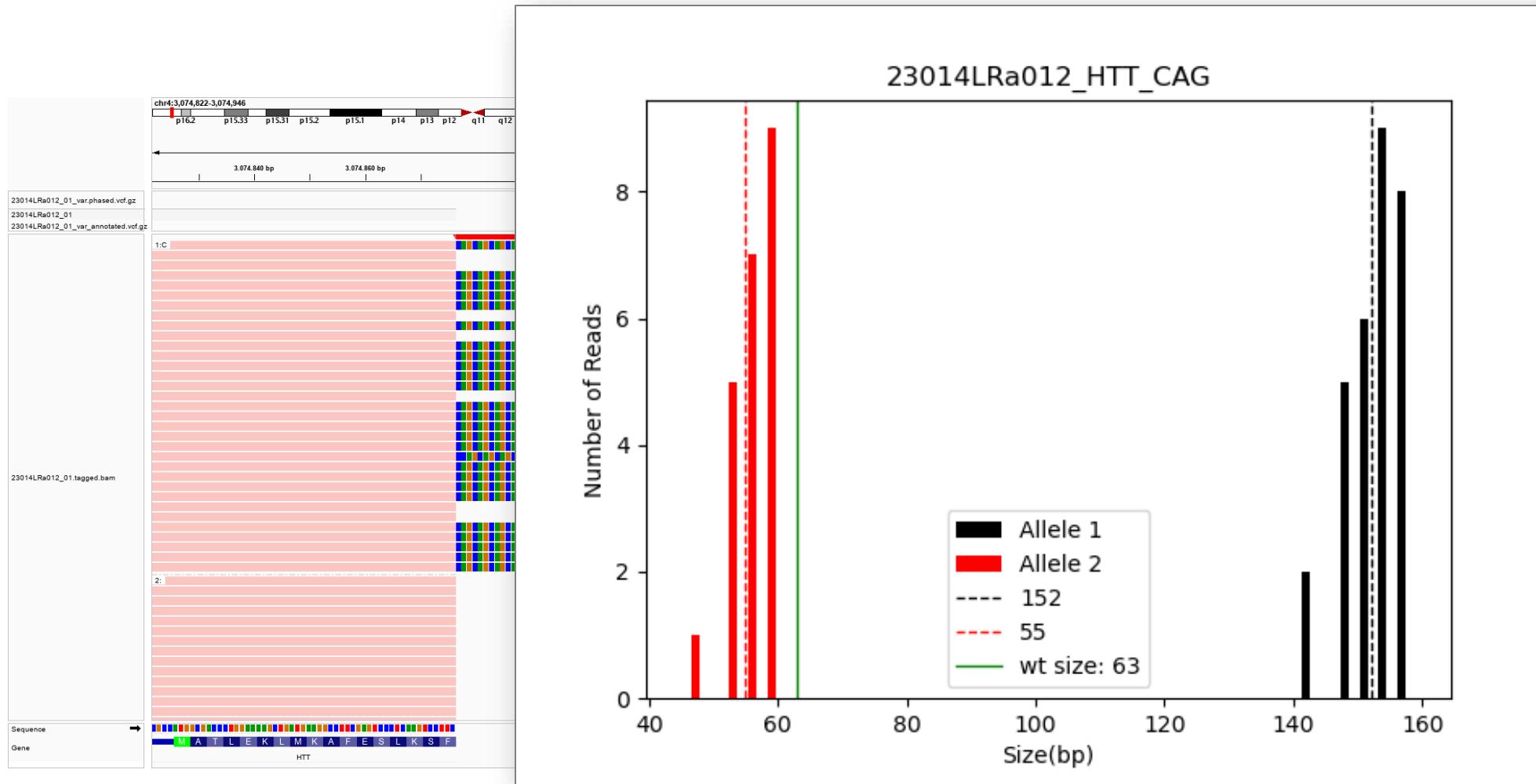


Caspar Gross, Thomas Braun  
Chia Ying Ko, Leon Schütz

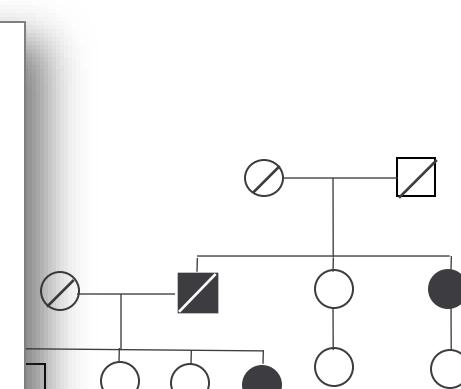




# Example: Huntington Disease (Tübingen Case 15)



Extended allele: 53 CAG Repeats in HTT



Huntington disease  
kinetic movement disorder  
sturbance  
ine motor coordination

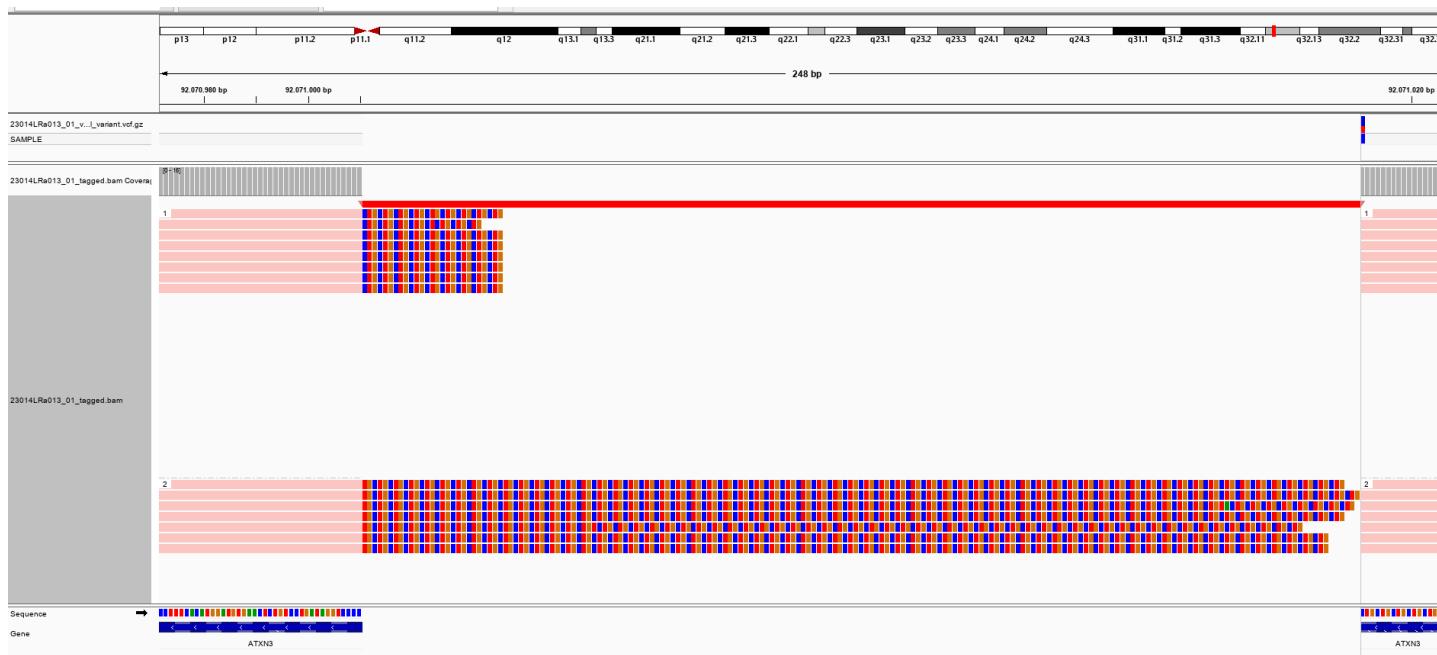
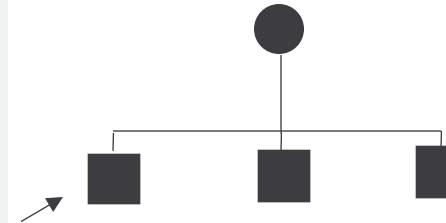




# Example: Ataxia (SCA3, Tübingen Case 16)

## Indication: SCA 3

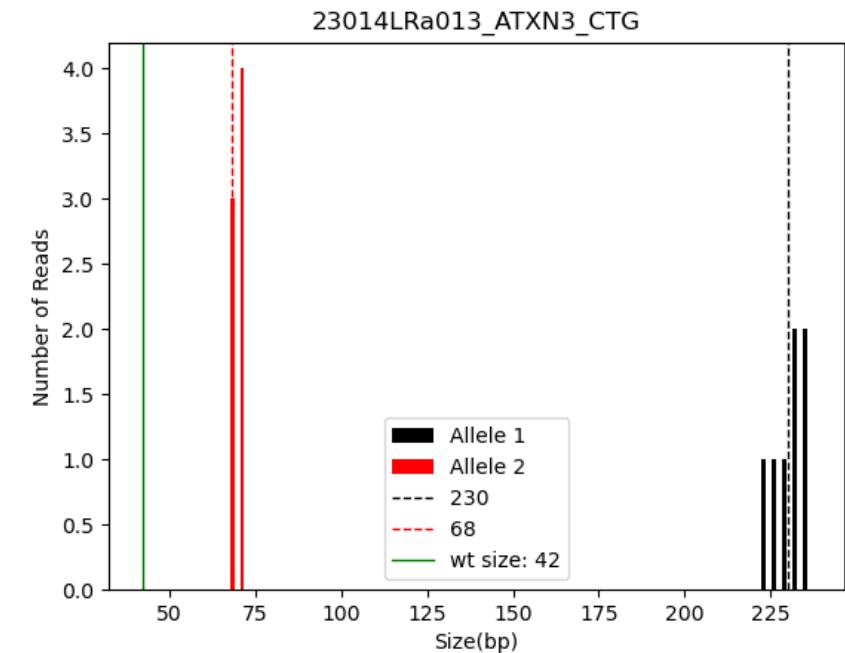
- ataxia
- dysarthria
- dysphagia
- cognitive impairment
- ophthalmoparesis
- parkinsonism



Extended allele: 77 CAG Repeats in ATXN3

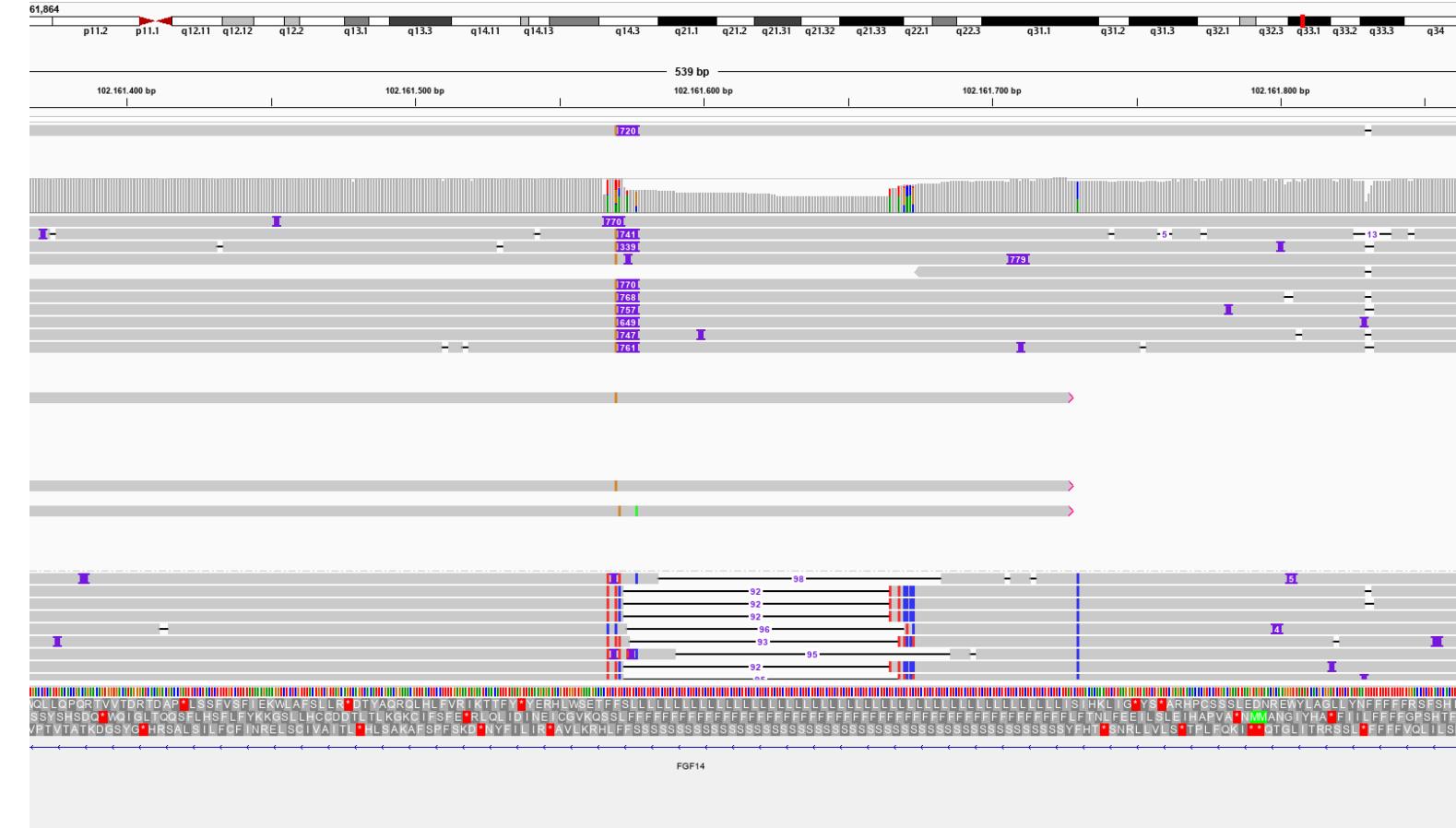
## ONT Long Read Analysis

- ATXN3:  $(CAG)_{23} / (CAG)_{77}$





# Ataxia with FGF14 Repeat Expansion (Tü Case 5)



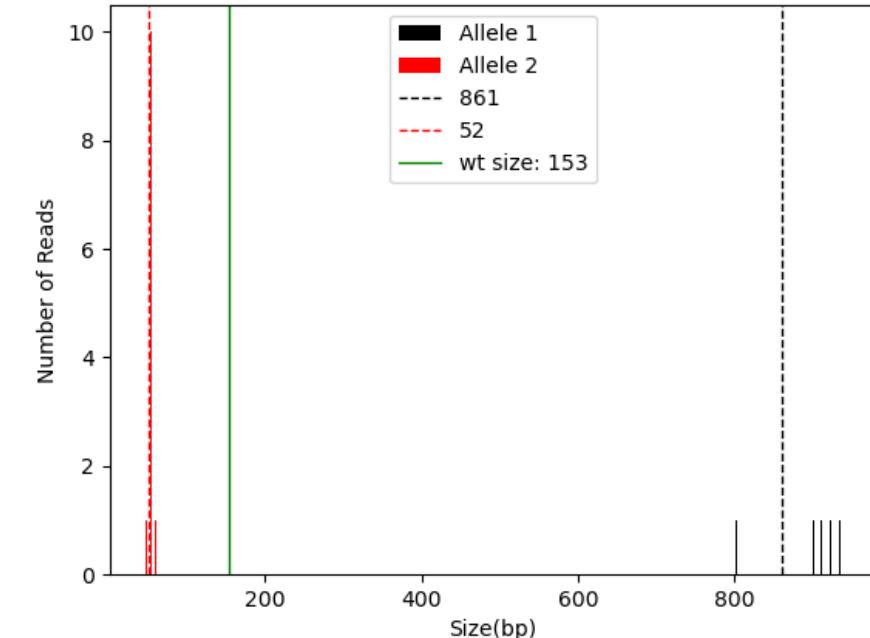
## GAA repeat in FGF14:

- allele 1: ~ 30 repeats,
- allele 2: >300 repeats

## ONT Long Read Analysis

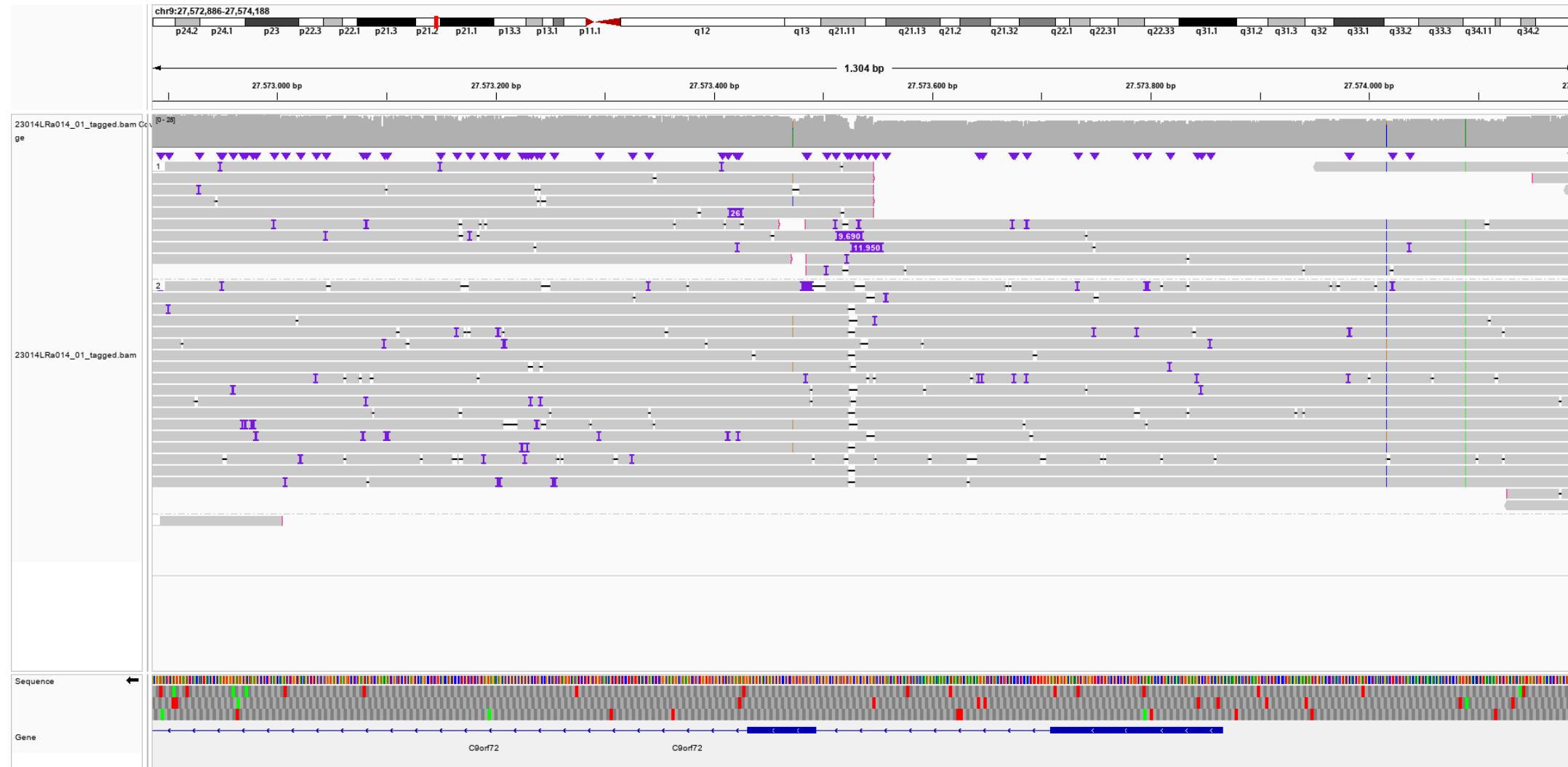
- FGF14: (GAA)<sub>309</sub>

23014LRa002\_FGF14\_AAG

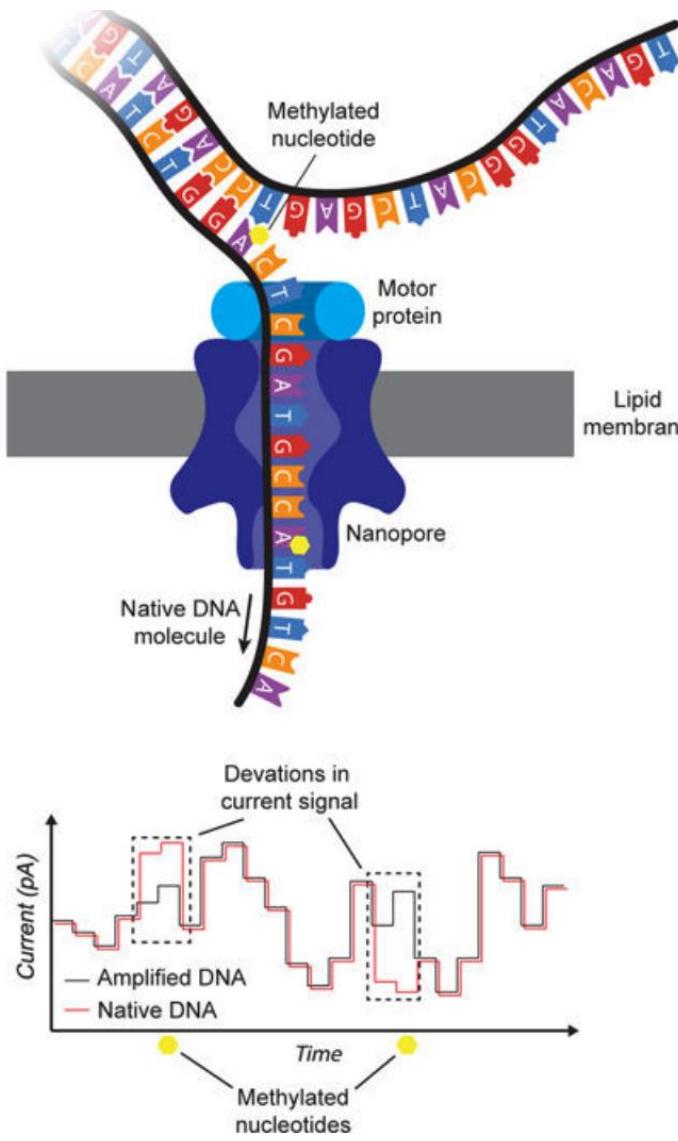




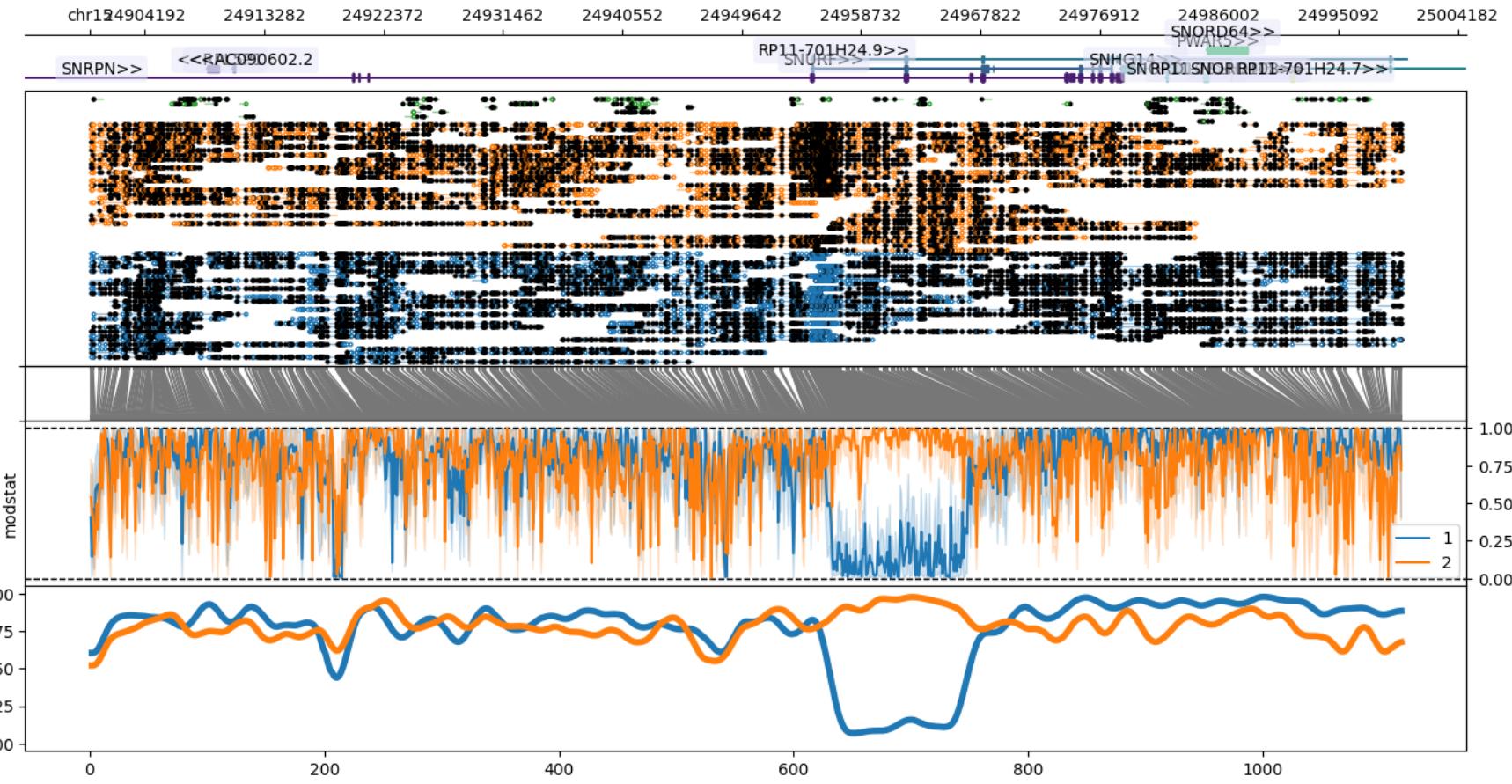
# Amyotrophic lateral sclerosis: *c9orf72* Repeat Expansion



# Haplotype-Phased DNA Methylation Calling



# Plot: MethylArtist

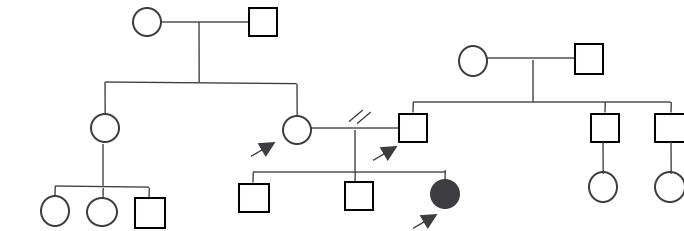
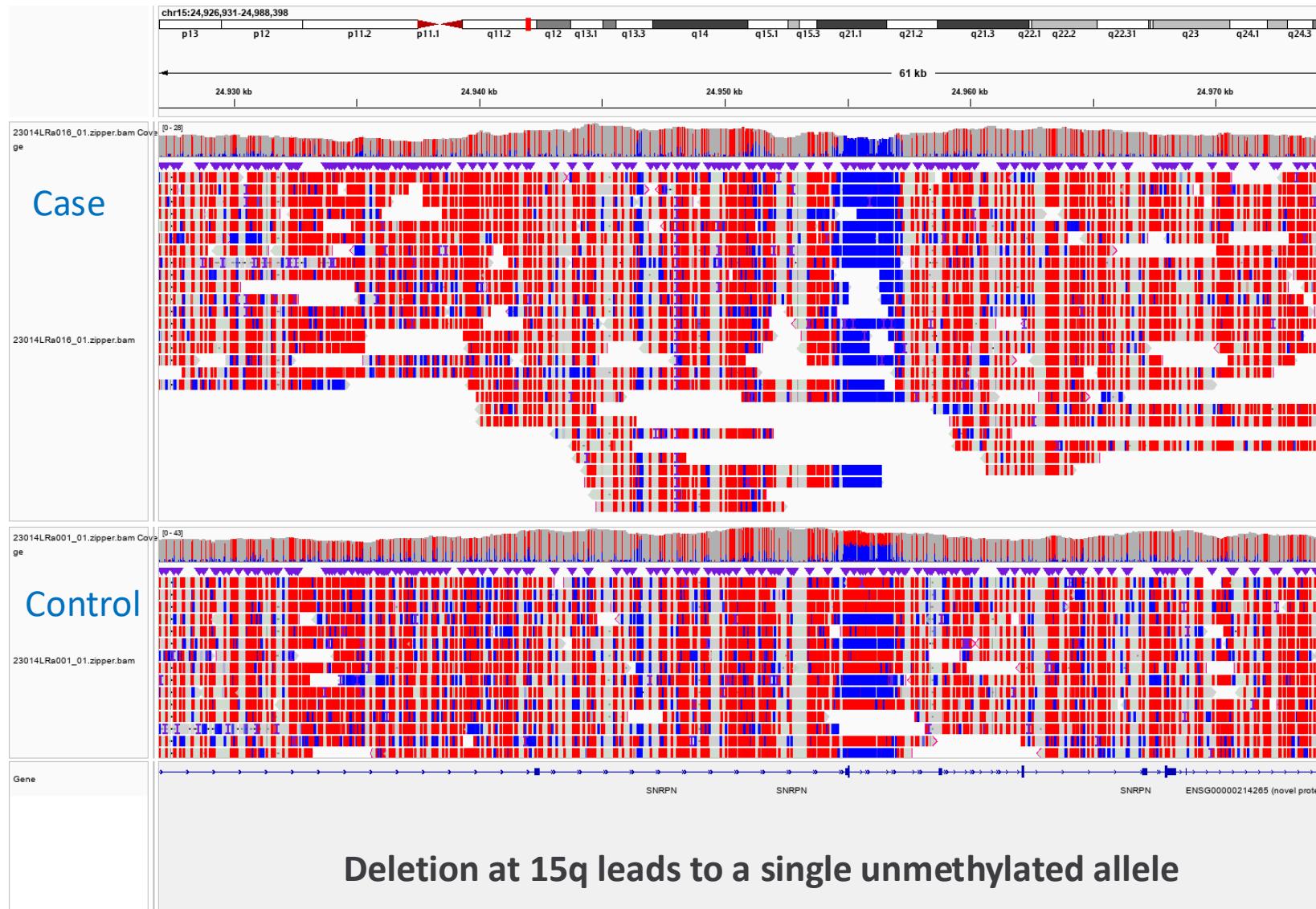


# Imprinting Locus

# Caspar Gross



# Autism Spectrum Disorder: Hypomethylation of SNRPN



## Phenotypes:

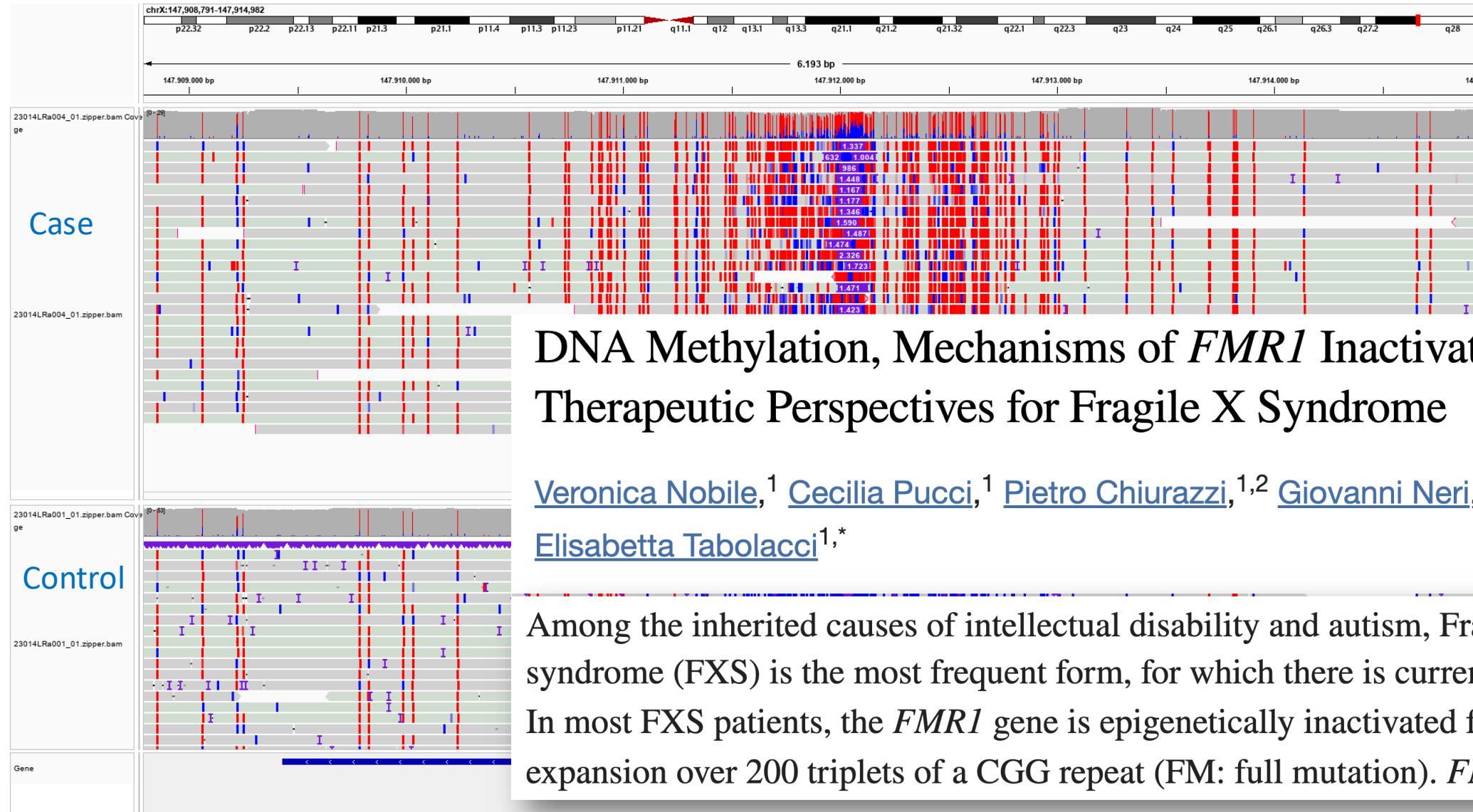
- developmental delay
- muscular hypotonia
- microcephaly
- behavioral abnormalities





# Fragile X Syndrome: FMR1 Inactivation by Methylation

CGG expansion & Hypermethylation in FMR1 (X chromosome, male patient),



## DNA Methylation, Mechanisms of *FMR1* Inactivation and Therapeutic Perspectives for Fragile X Syndrome

Veronica Nobile,<sup>1</sup> Cecilia Pucci,<sup>1</sup> Pietro Chiurazzi,<sup>1,2</sup> Giovanni Neri,<sup>1,3</sup> and Elisabetta Tabolacci<sup>1,\*</sup>

Among the inherited causes of intellectual disability and autism, Fragile X syndrome (FXS) is the most frequent form, for which there is currently no cure. In most FXS patients, the *FMR1* gene is epigenetically inactivated following the expansion over 200 triplets of a CGG repeat (FM: full mutation). *FMR1* encodes



# Discovery of a Novel Repeat Expansion Genes

nature genetics

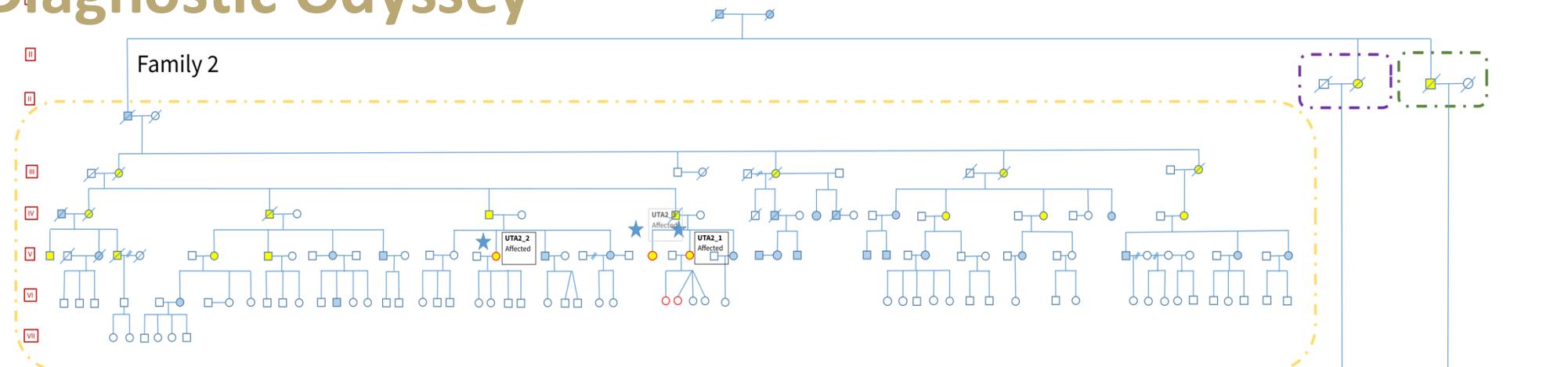
Letter | Published: 29 April 2024

## A GGC-repeat expansion in *ZFHX3* encoding polyglycine causes spinocerebellar ataxia type 4 and impairs autophagy

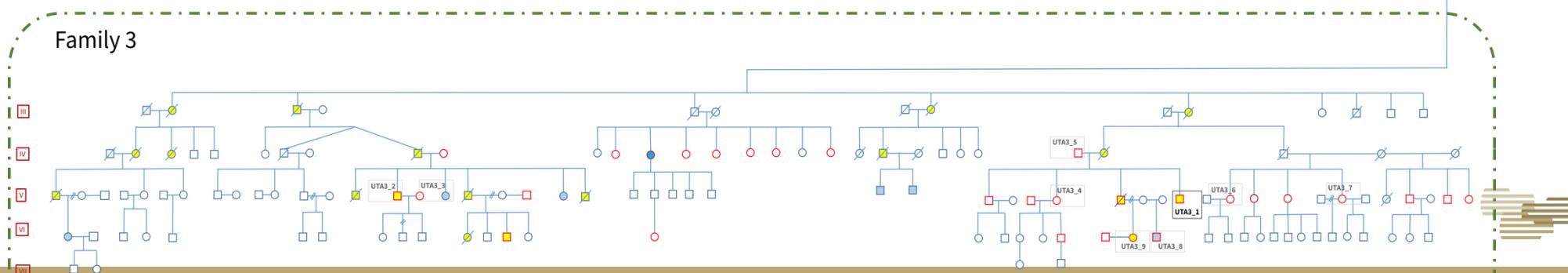
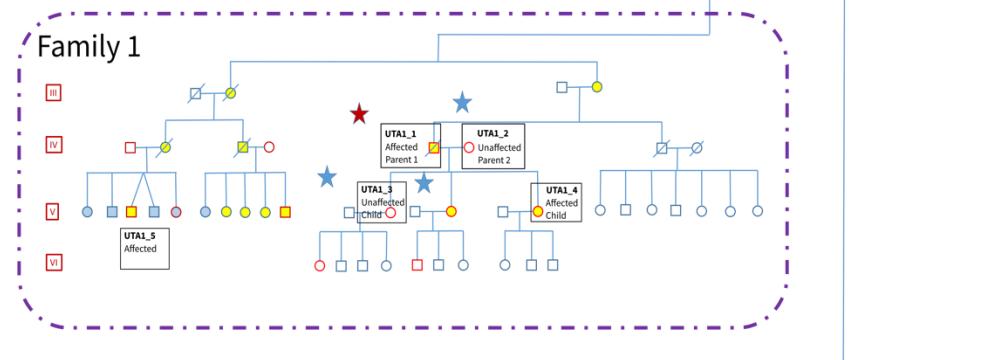
[Karla P. Figueroa](#), [Caspar Gross](#), [Elena Buena-Atienza](#), [Sharan Paul](#), [Mandi Gadelman](#),  
[Naseebullah Kakar](#), [Marc Sturm](#), [Nicolas Casadei](#), [Jakob Admard](#), [Joohyun Park](#), [Christine Zühlke](#), [Yorck Hellenbroich](#), [Jelena Pozojevic](#), [Saranya Balachandran](#), [Kristian Händler](#),  
[Simone Zittel](#), [Dagmar Timmann](#), [Friedrich Erdlenbruch](#), [Laura Herrmann](#), [Thomas Feindt](#),  
[Martin Zenker](#), [Thomas Klopstock](#), [Claudia Dufke](#), [Daniel R. Scoles](#), [Arnulf Koeppen](#), [Malte Spielmann](#), [Olaf Riess](#)✉, [Stephan Ossowski](#), [Tobias B. Haack](#) & [Stefan M. Pulst](#)✉



# Spinocerebellar Ataxia Type 4: Ending a 25-Year Diagnostic Odyssey



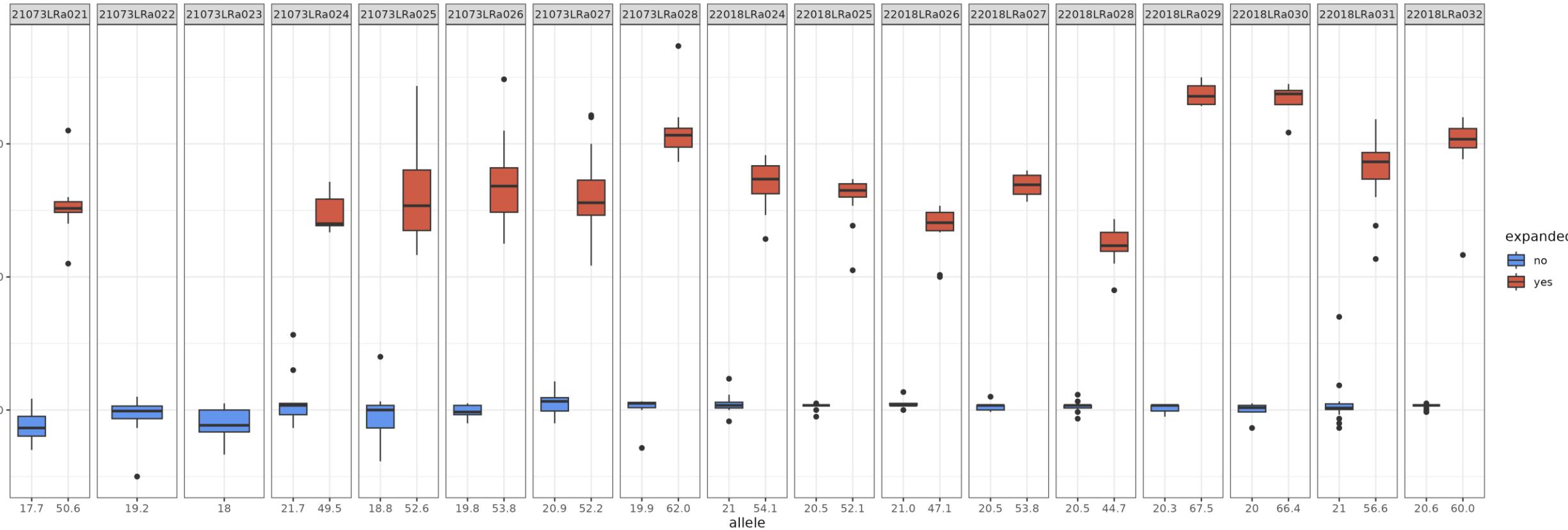
- Multi-Generation family from Utah (Swedish ancestry) with dominant SCA4 (many cases)
- Linkage on 16q, but causal gene unknown
- 6 cases + 2 Controls selected for long-read sequencing (PacBio-HiFi and ONT-Nanopore)



# Repeat Length Analysis using Nanopore Reads with Straglr



Minimap2 + Straglr + in-house visualization



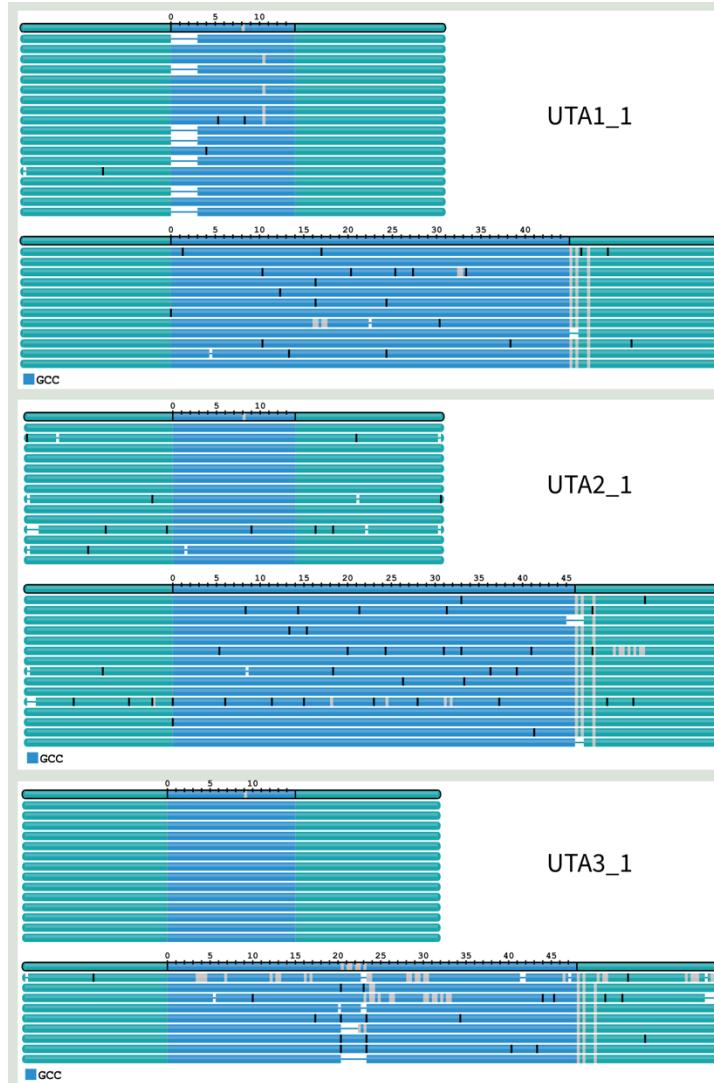
- We identified a heterozygous GGC repeat expansion in gene ZFHX3 in all cases
- Pathogenic repeat length is >42 repeats



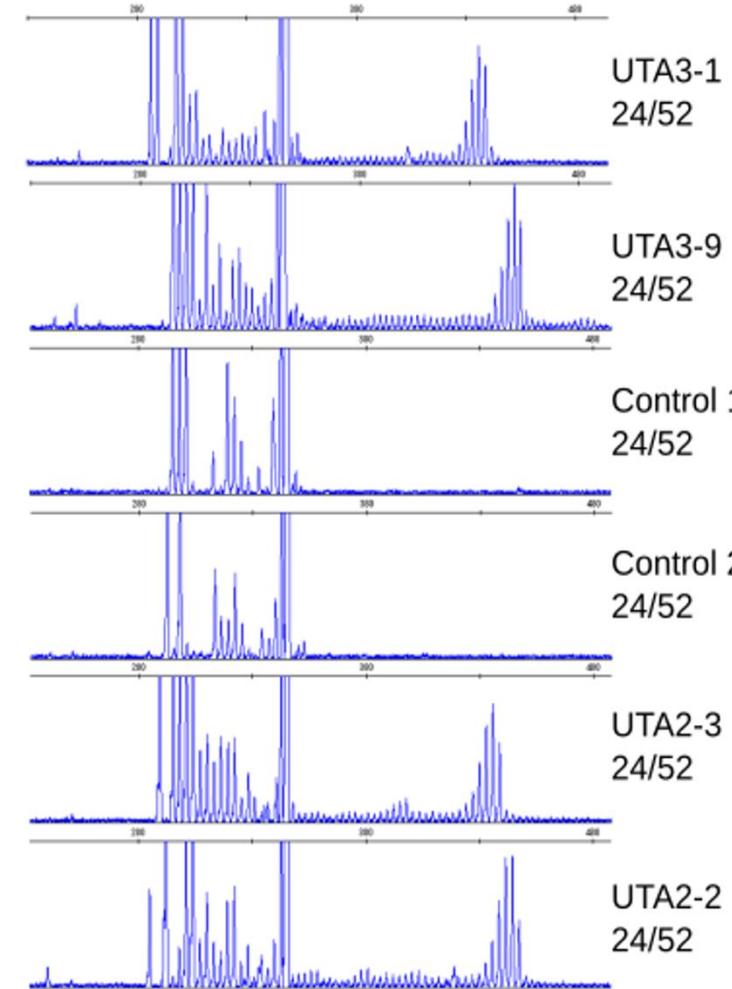


# SCA4: Pathogenic GGC Repeat Expansion in ZFHX3

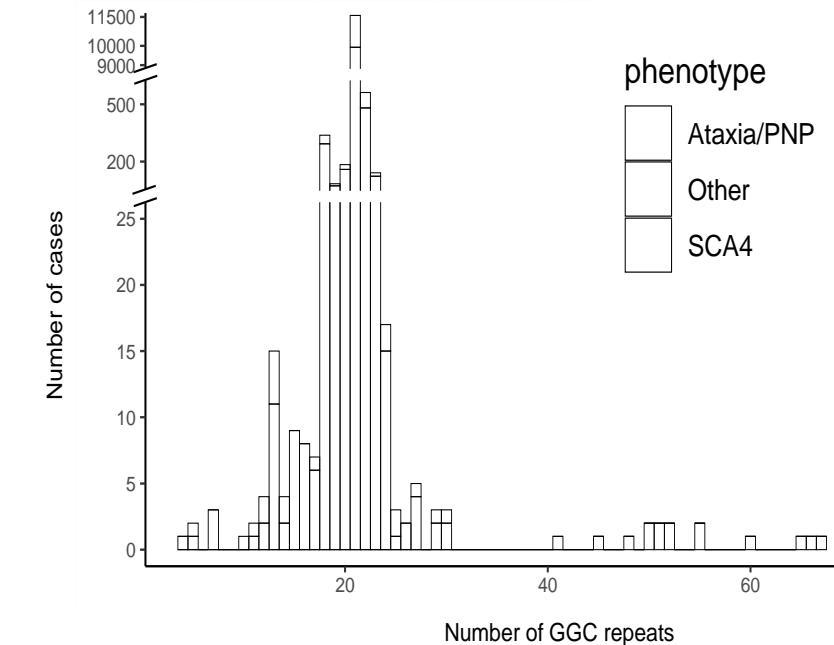
Discovery: HiFi + TRGT



Validation (PCR amplification)



Pathogenic repeat length





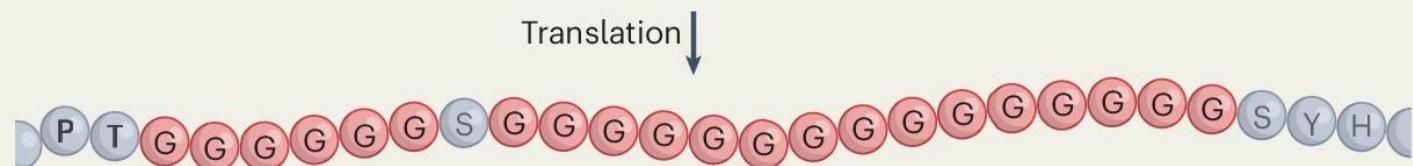
# An unexpected polyglycine route to spinocerebellar ataxia

Nicolas Charlet-Berguerand 

*Nature Genetics* (2024) | Cite this article

### ZFHX3 gene with a normal number of GGC repeats (~21 units)

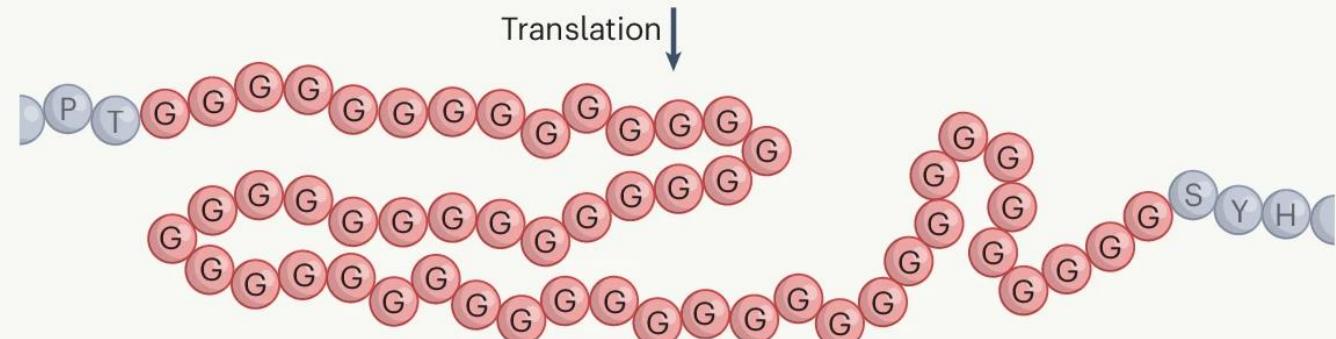
...CCCACCGGGCGGCGGGCGGGCGGTGGCAGTGGCGGGCGGGCGGGCGGGCGGGCGGGCGGGCGGGCGGGCTCGTACCAAC...



## ZFHX3/ATBF1 transcription factor

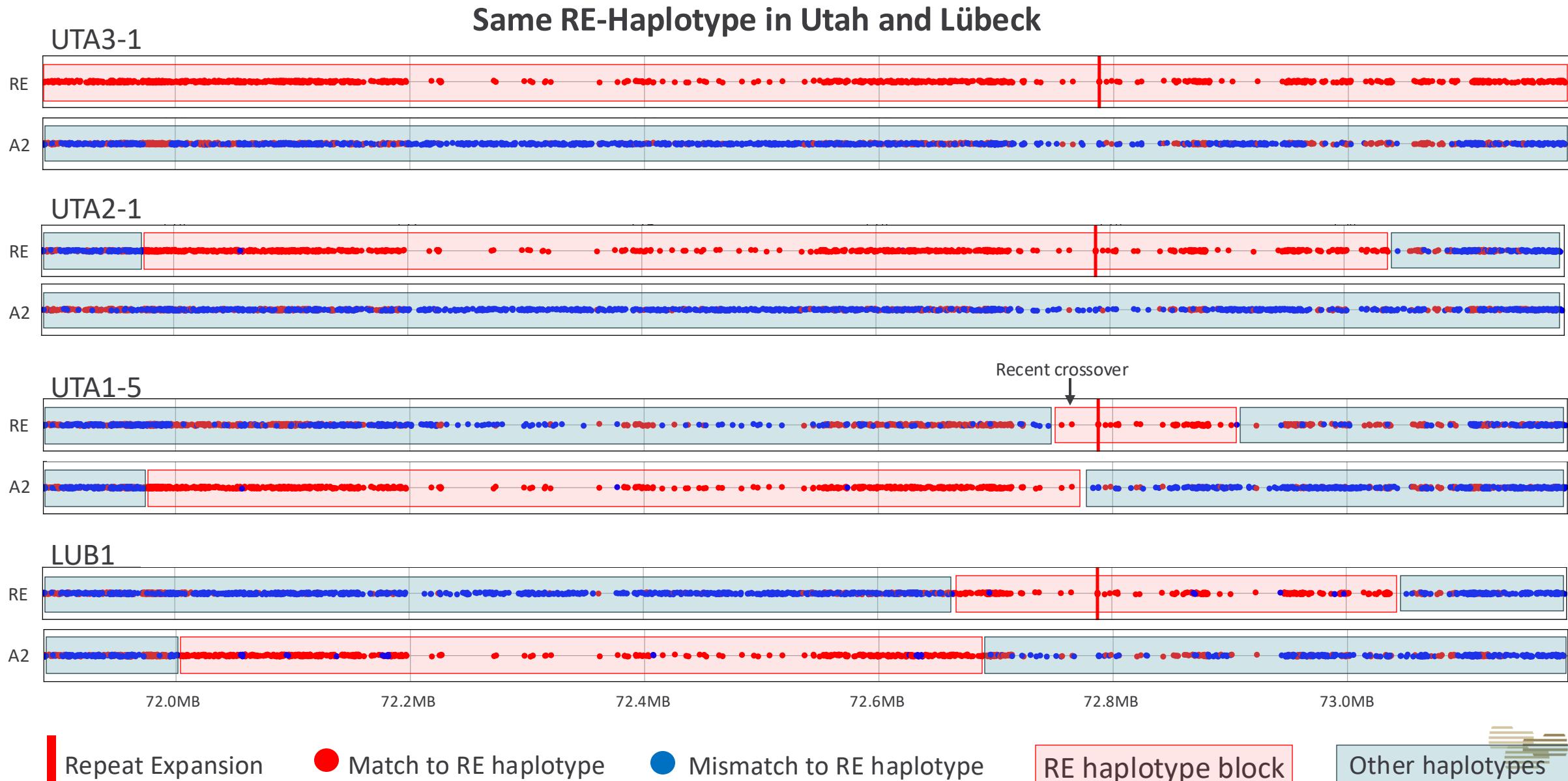
SCA4

## ZFHX3 gene with a GGC repeat expansion (>40 units)

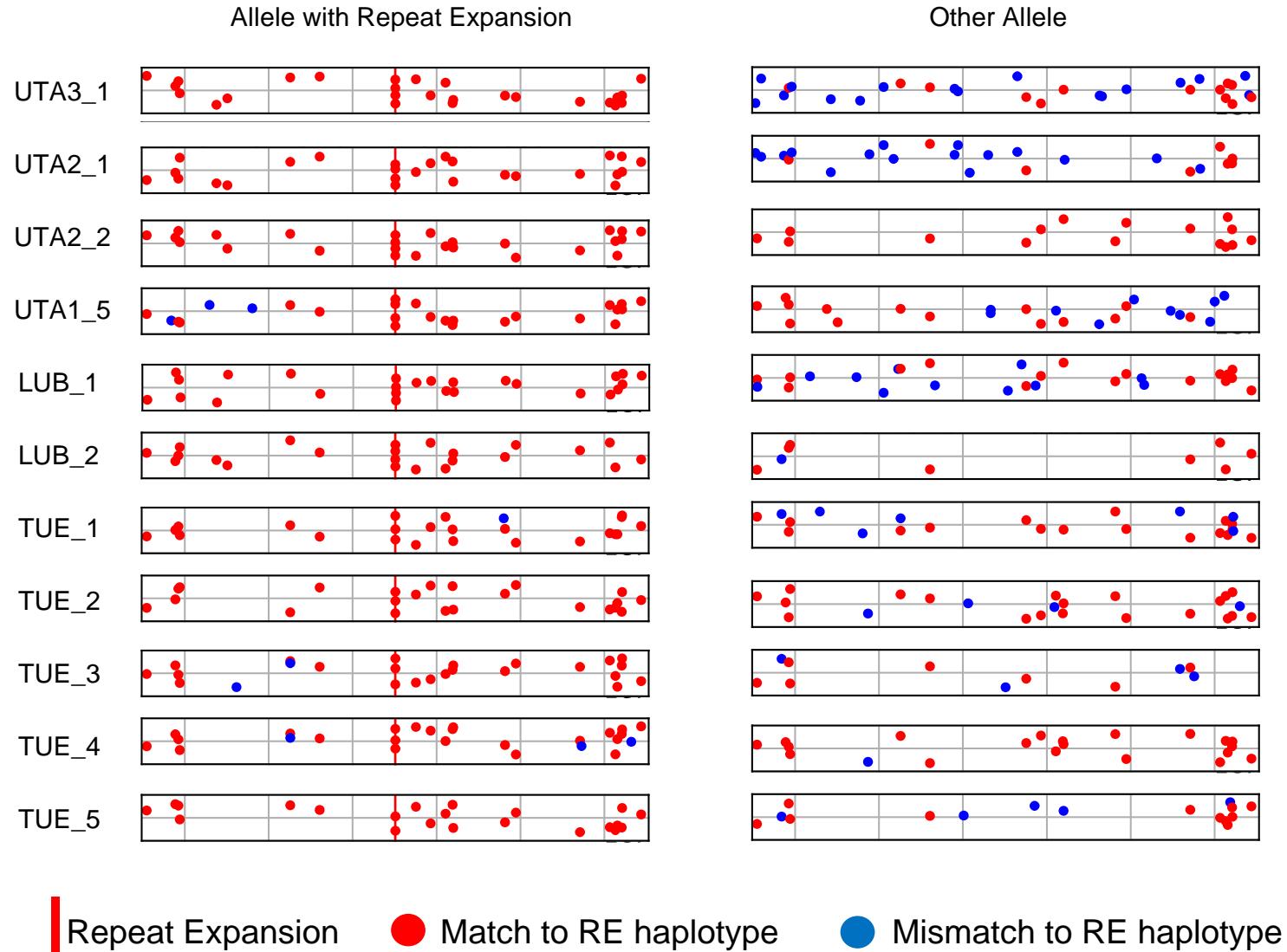


ZFHX3/ATBF1 protein with an extended polyglycine stretch  
Impaired transcriptional functions and/or toxic gain of function  
Intranuclear inclusions and neuronal cell death

# Independent Discovery in 2 Cases of Lübeck (Spielman Lab)



# Characteristics SNVs Only Found in RE-Linked Haplotype



6 Ultra-rare SNVs are specific to the Repeat Expansion Allele



# Screening in 6,495 srGS Datasets from U. Tübingen using Unique SNVs & ExpansionHunter: 7 more cases identified

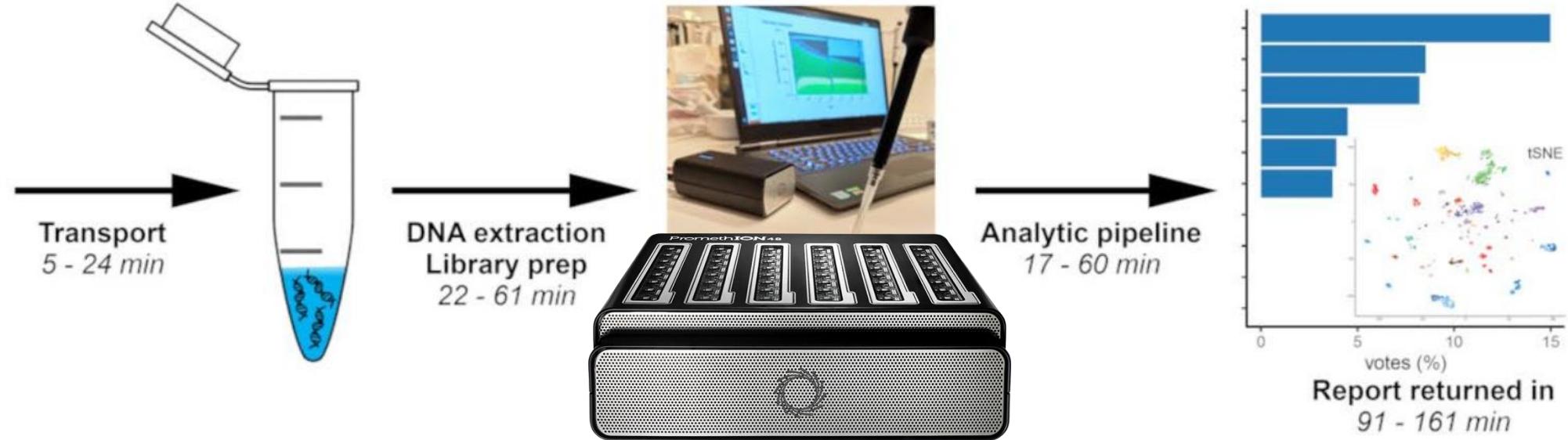
					Characteristic variants flanking repeat expansion					
Sample	Disease status	Repeat length	Seq-Tech	chr16:7273445 7 A>G	chr16:727377 03 T>C	chr16:727877 19 A>G	chr16:727877 37 A>G	chr16:727877 39 T>C	chr16:727877 43 A>G	
UTA2	Unaffected	normal	lr	no	no	no	no	no	no	
UTA3	Unaffected	normal	lr	no	no	no	no	no	no	
UTA1	Affected	expanded	lr	yes	yes	yes	yes	yes	yes	
UTA4	Affected	expanded	lr	yes	yes	yes	yes	yes	yes	
UTA5	Affected	expanded	lr	yes	yes	yes	yes	yes	yes	
UTA6	Affected	expanded	lr	yes	yes	yes	yes	yes	yes	
UTA7	Affected	expanded	lr	yes	yes	yes	yes	yes	yes	
UTA8	Affected	expanded	lr	yes	yes	yes	yes	yes	yes	
LUB1	Affected	expanded	lr	yes	yes	yes	yes	yes	yes	
LUB2	Affected	expanded	lr	yes	yes	yes	yes	yes	yes	
TUB1	Affected	expanded	sr	yes	no	yes	yes	yes	yes	
TUB2	Affected	expanded	sr	yes	no	yes	yes	yes	yes	
TUB3	Affected	expanded	sr	yes	no	yes	yes	yes	yes	
TUB4	Affected	expanded	sr	yes	no	yes	yes	yes	yes	
TUB5	Affected	expanded	sr	yes	no	yes	low quality	low quality	yes	
TUB6	Affected	expanded	sr	yes	no	yes	yes	yes	yes	
TUB7	Affected	expanded	sr	yes	no	yes	yes	yes	yes	



# LeOPARD: Intraoperative classification of tumors



Tumor biopsy  
early in surgery



Goal:

- Classification of tumor subtypes within 2 hours
- Adjustment of surgery strategy during surgery based on results

Coordination: Philipp Euskirchen, Charité Berlin



# nanoDX pipeline: methylation-based classification of tumors



<https://gitlab.com/pesk/nanoDx>

nanoDX:

- Uses low-pass Nanopore whole genome sequencing data
- Classifies tumor types based on methylation profile
- Two neural networks available: Brain-tumors (Capper et al.) and Pan-Cancer

medRxiv

THE PREPRINT SERVER FOR HEALTH SCIENCES



Cold  
Spring  
Harbor  
Laboratory

BMJ Yale

Follow this preprint

**crossNN: an explainable framework for cross-platform DNA methylation-based classification of cancer**

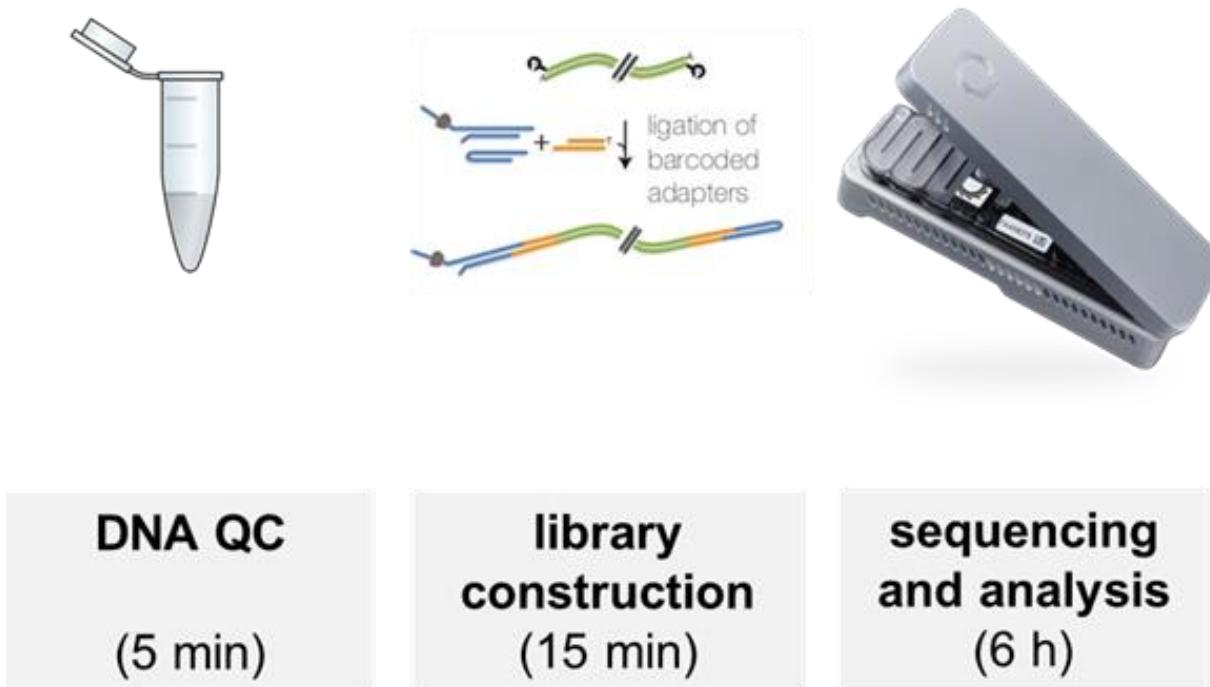
Dongsheng Yuan, Robin Jugas, Petra Pokorna, Jaroslav Sterba, Ondrej Slaby, Simone Schmid, Christin Siewert, Brendan Osberg, David Capper, Pia Zeiner, Katharina Weber, Patrick Harter, Nabil Jabareen, Sebastian Mackowiak, Naveed Ishaque, Roland Eils, Sören Lukassen, Philipp Euskirchen

**doi:** <https://doi.org/10.1101/2024.01.22.24301523>



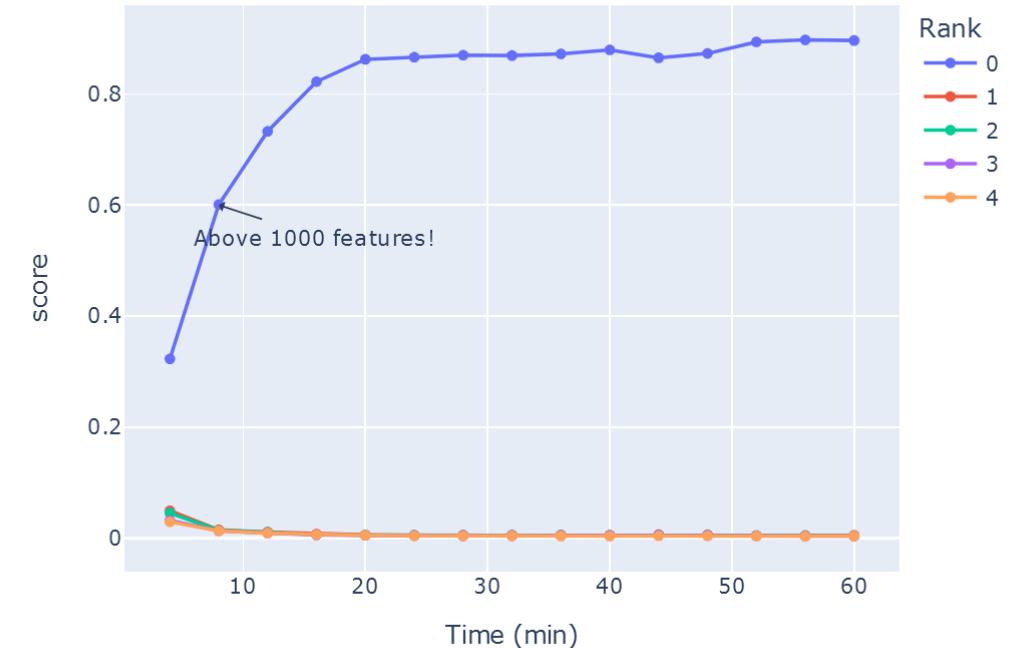


# nanoDX pipeline: round robin test



- Samples: DNA derived from brain tumor tissue
- Realtime analysis using nanoDx (with neural network for brain tumors)
- crossCNN updates results every minute

Capper et al. classifier. Top class (Rank 0) is MNG.





# nanoDX pipeline: Clinical Report

nanopore low-pass whole genome sequencing report

**Sample ID:** bams\_ALL

## Quality control metrics

**Barcode statistics** Demultiplexing statistics are shown for quality control only using a maximum of randomly subsampled 100,000 reads. All reads are considered for downstream analysis.

Adapters detected in 811 of 100000 reads

NBD104/NBD114	811:	0.81 %
none	86275:   #####	86.28 %

Barcodes detected in 811 of 100000 adapters

barcode01	810:	0.81 %
barcode16	1:	0.00 %
none	86275:   #####	86.28 %

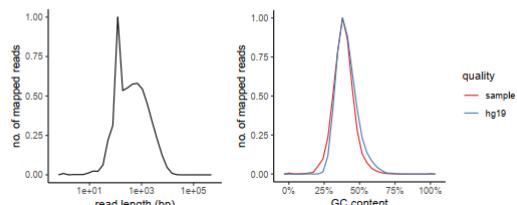
12914 reads were skipped due to the min. length filter.

Demultiplexing finished in 149.50s

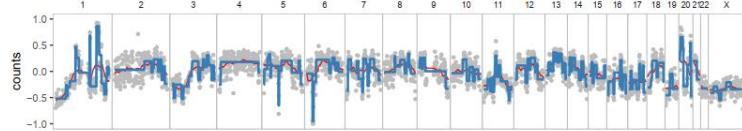
## Read statistics

General summary:  
 Average percent identity: 96.9  
 Fraction of bases aligned: 1.0  
 Mean read length: 1,191.0  
 Mean read quality: 14.9  
 Median percent identity: 97.9  
 Median read length: 573.0  
 Median read quality: 16.6  
 Number of reads: 1,225,509.0  
 Read length N50: 2,543.0  
 STDEV read length: 1,737.2  
 Total bases: 1,459,604,278.0  
 Total bases aligned: 1,435,112,548.0  
 Number, percentage and megabases of reads above quality cutoffs  
 >Q10: 1208064 (98.6%) 1443.3Mb

Mean genome coverage is 0.46X.



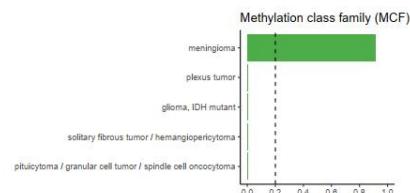
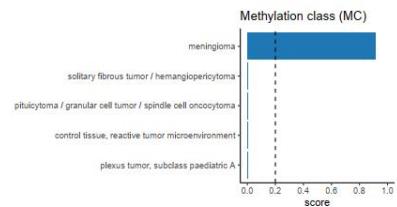
## Copy number profile



## Methylation-based classification

Methylation-based classification is based on 83061 CpG sites (overlap of sites covered in this sample and the model). At the methylation class (MC) level, the sample has been classified as **meningioma**. This prediction has a confidence score of **0.911**. At the methylation class family (MCF) level, the sample has been classified as **meningioma**. The MCF prediction has a confidence score of **0.911**.

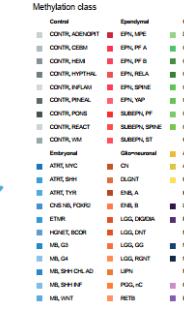
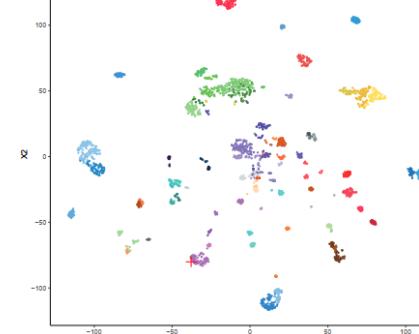
Scores for the Top 5 entities on MC and MCF level are given below. Vertical dashed lines indicate the recommended >0.2 cut-off for classification.



## Dimensionality reduction plot

Dimensionality reduction plots are only intended for visual quality control, not classification, and must be interpreted with caution.

t-SNE, no. of PCA dimensions = 94, perplexity = 30, max no. of iterations = 2500



**Disclaimer** Methylation-based classification using nanopore whole genome sequencing is a research tool currently under development. It has not been clinically validated in sufficiently large cohorts. Interpretation and implementation of the results in a clinical setting is in the sole responsibility of the treating physician.

(Report generated on 2024-10-02 09:22:41.916161. Pipeline version: v1.0rc3-9-gce225ba, ce225ba)



# Thanks!

## University Hospital Aachen

Ingo Kurth  
Florian Kraft  
Sebastian Gießelmann

## Charité – Universitätsmedizin Berlin

Nadja Ehmke

## Berlin Institute of Health

Janine Altmüller  
Manuel Holtgrefe  
Claudia Quedenau

## Medical School Hannover

Bernd Auber  
Gunnar Schmidt

## University of Tübingen

Elena Buena-Atienza  
Caspar Gross  
Leon Schütz  
Marc Sturm  
Jakob Admard  
Vladislav Lysenkov  
German Demidov  
Chia Ying Ko  
Thomas Braun  
Nicolas Casadei  
Tobias Haack  
Olaf Riess

## SCA4 Project

Malte Spielmann (U. of Lübeck and Kiel)  
Karla P. Figueroa (U. of Utah)  
Daniel R. Scoles (U. of Utah)  
Stefan M. Pulst (U. of Utah)



## Oxford Nanopore Technologies

Manuela Saathoff  
Tonya McSherry  
Gerald Goh  
Alexander Vogel  
Anthony Doran  
Alexander Rotmann

## NVIDIA

Uwe Samer  
Harry Clifford  
Lotfi Slim

# IonGER Kickoff-Meeting Frankfurt, April 2023



Elena Buena-Atienza



Caspar Gross



Jakob Admard

Posters:

P15.088.A

P21.060.C

Talk: Tobias Haack (tomorrow)

# Nanopore Experts in Tübingen



**Elena Buena-Atienza**

Nanopore Sequencing,  
wet-lab and analysis



**Marc Sturm**

Lead Diagnostic Bioinformatics



**Caspar Gross**

Long-read bioinformatics  
method development and analysis



**Leon Schütz**

Diagnostic Nanopore Sequencing  
megSAP / Gsvar Software



**Jakob Admard**

Long-read bioinformatics  
method development and analysis



**German Demidov**

ERDERA Rare-Disease Consortium  
Nanopore Workpackage

