Application of Long-Read Sequencing (LRS) in resolving complex regions and detecting diseases

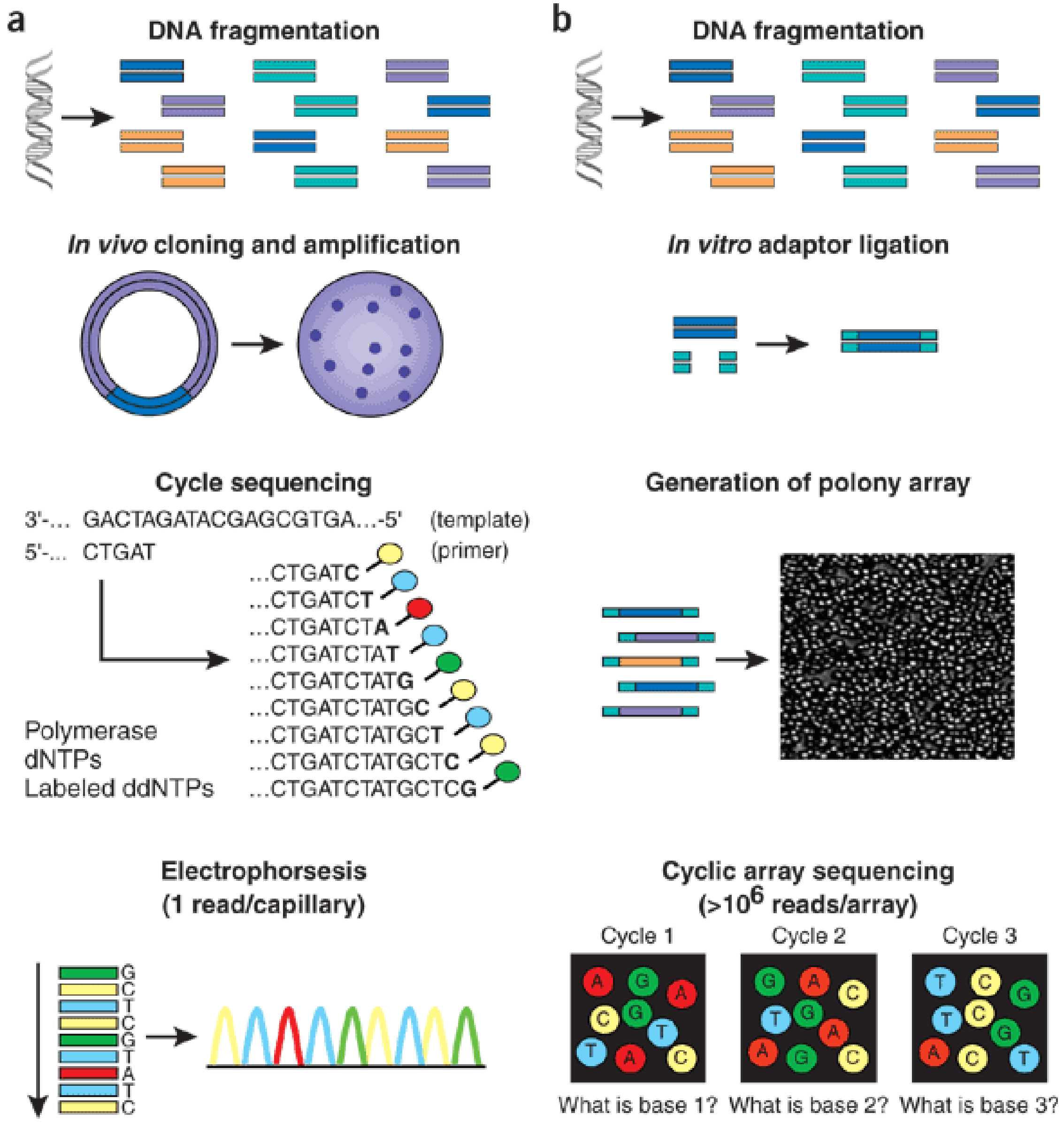
Long-Read Sequencing Is All You Need

Quanyu Chen

25.6.20

Short-Read Sequencing helps a lot

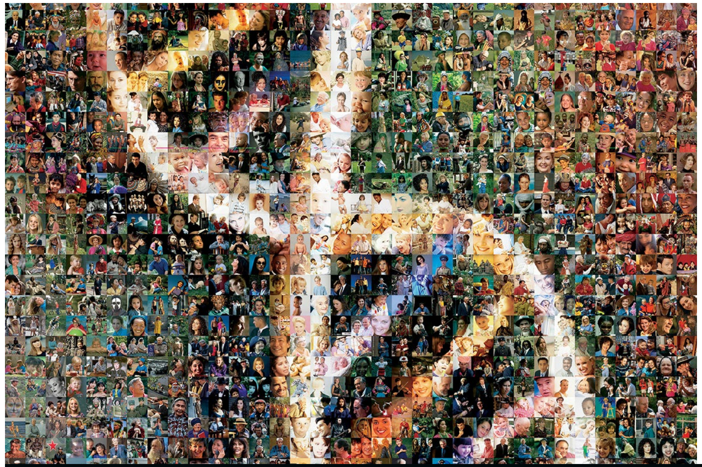
Sanger Sequencing **Short-gun Sequencing**

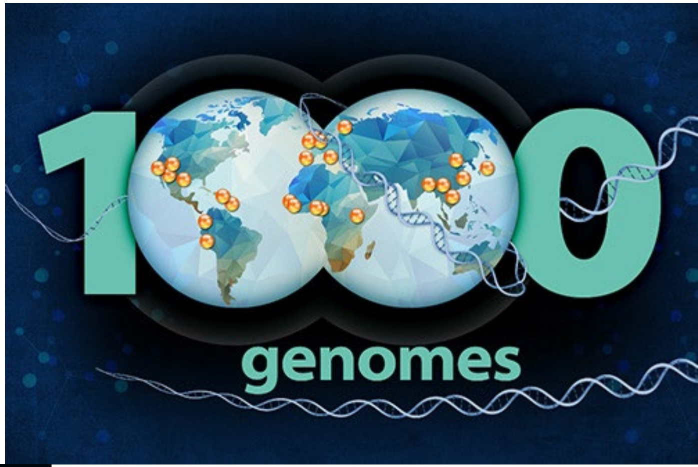


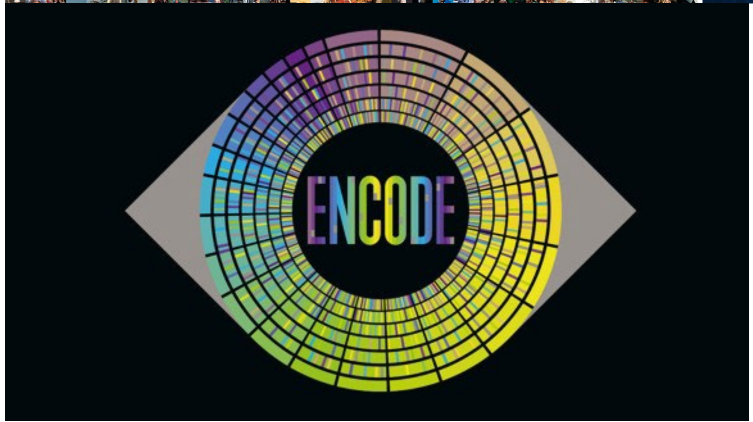
Advantages:

- **Ultra-high throughput, scalability, speed, cost effectiveness**
- **Accuracy:** detect genetic variants accurately
- **Versatility:** whole-genome sequencing, whole-exome, targeted regions associated with diseases

SRS was revolutionary!

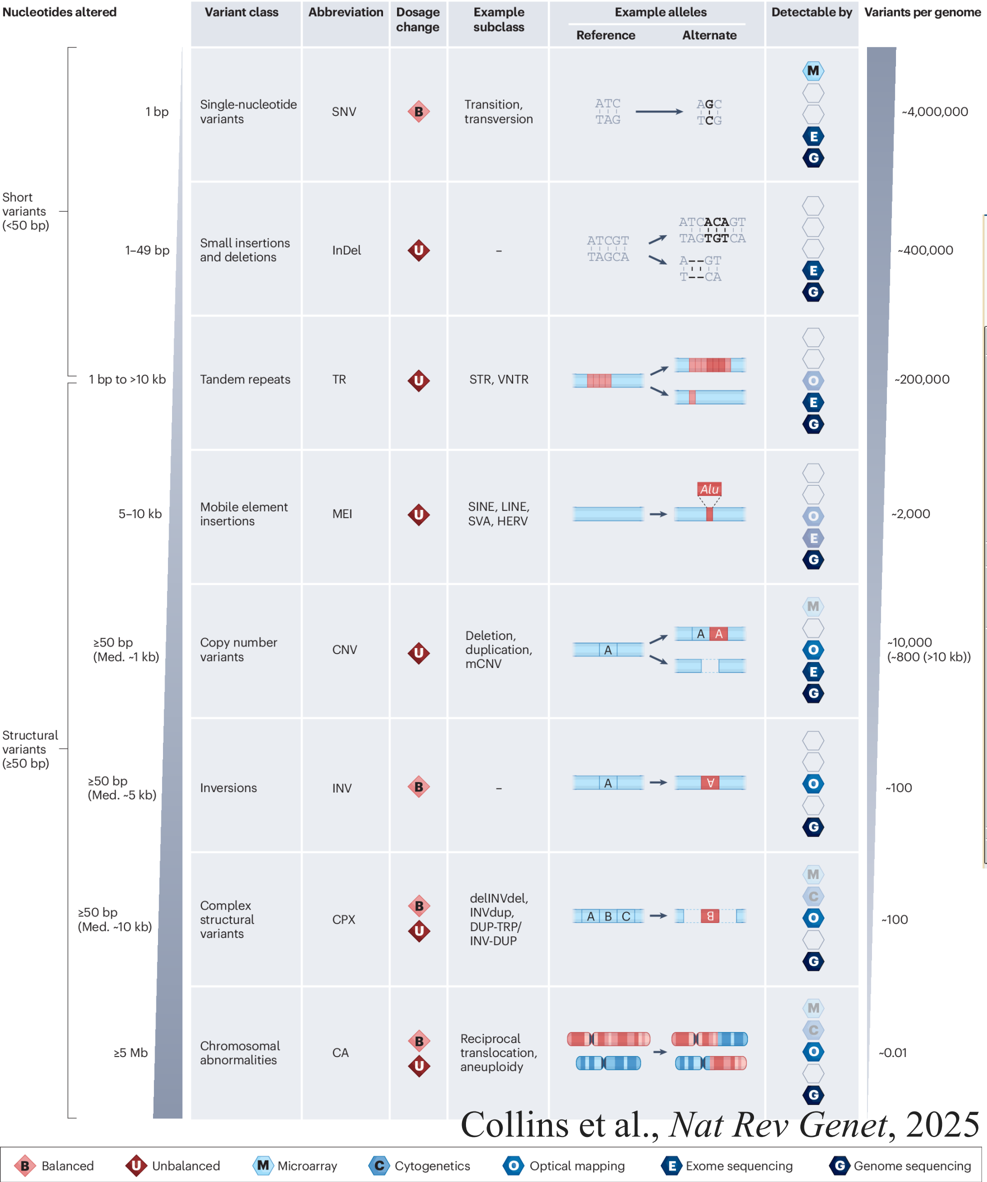






Short-Read Sequencing has a lot of limitations too

The root of all limitations: **Short!**

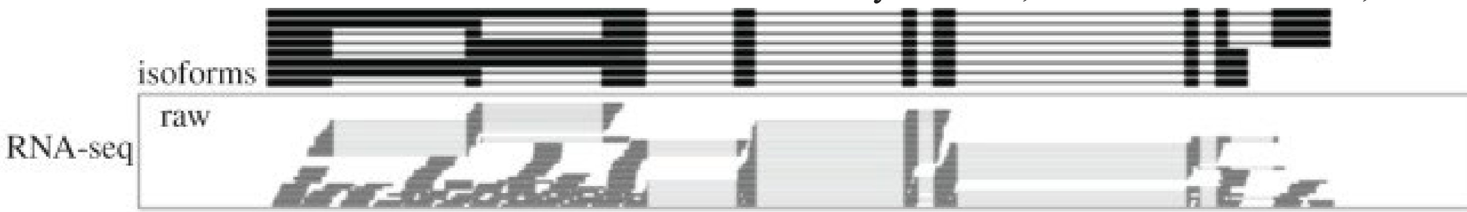


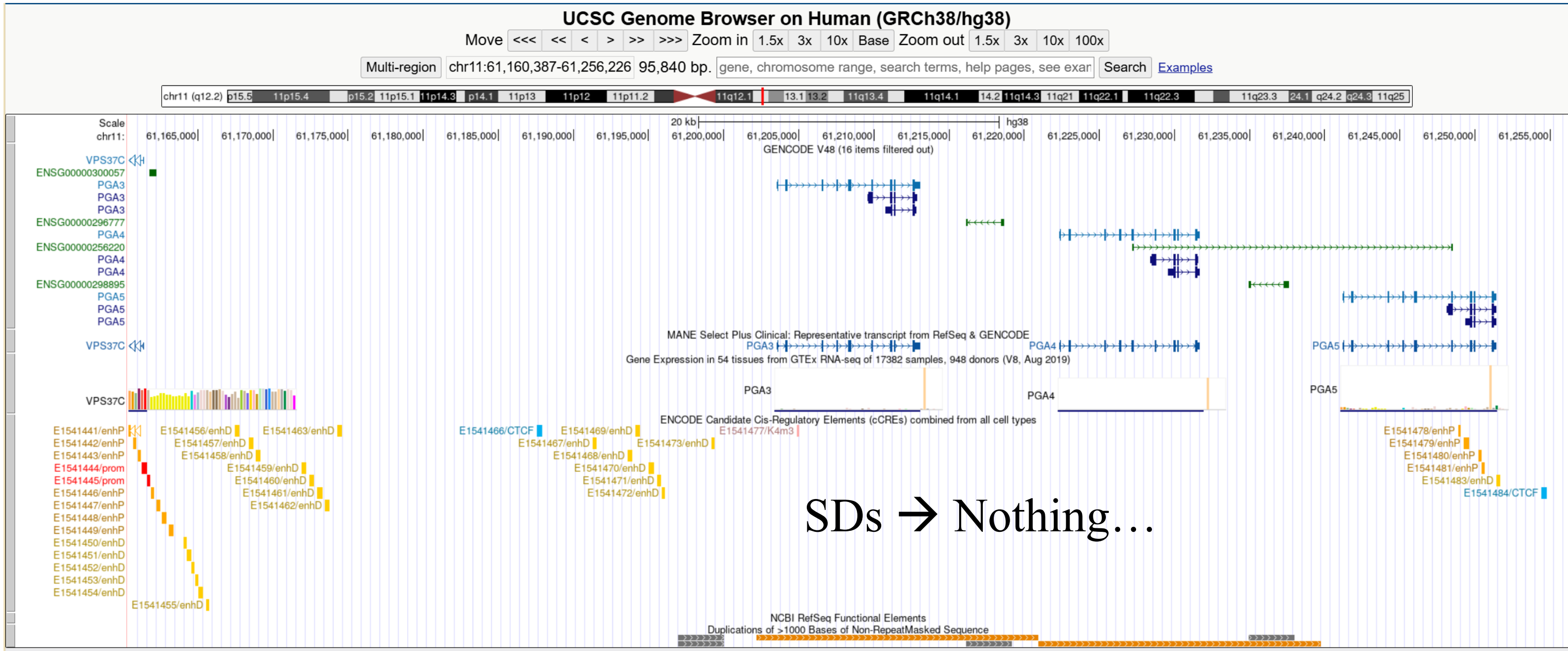
Nucleotides altered	Variant class	Abbreviation	Dosage change	Example subclass	Example alleles: Reference	Example alleles: Alternate	Detectable by	Variants per genome
Short variants (<50 bp): 1 bp	Single-nucleotide variants	SNV	B	Transition, transversion	ATC TAG	AGC TCG	M, E, G	~4,000,000
1–49 bp	Small insertions and deletions	InDel	U	–	ATCGT TAGCA	ATCACAGT TAGTGTCA; A--GT T--CA	E, G	~400,000
1 bp to >10 kb	Tandem repeats	TR	U	STR, VNTR			O, E, G	~200,000
Structural variants (≥50 bp): 5–10 kb	Mobile element insertions	MEI	U	SINE, LINE, SVA, HERV		Alu	O, E, G	~2,000
≥50 bp (Med. ~1 kb)	Copy number variants	CNV	U	Deletion, duplication, mCNV	A	A A	M, O, E, G	~10,000 (~800 (>10 kb))
≥50 bp (Med. ~5 kb)	Inversions	INV	B	–	A	∀	O, G	~100
≥50 bp (Med. ~10 kb)	Complex structural variants	CPX	B, U	delINVdel, INVdup, DUP-TRP/INV-DUP	A B C	B	M, C, O, G	~100
≥5 Mb	Chromosomal abnormalities	CA	B, U	Reciprocal translocation, aneuploidy			M, C, O, G	~0.01

B Balanced · U Unbalanced · M Microarray · C Cytogenetics · O Optical mapping · E Exome sequencing · G Genome sequencing

Collins et al., *Nat Rev Genet*, 2025

Byrne et al., *Phil. Trans. R. Soc. B*, 2019



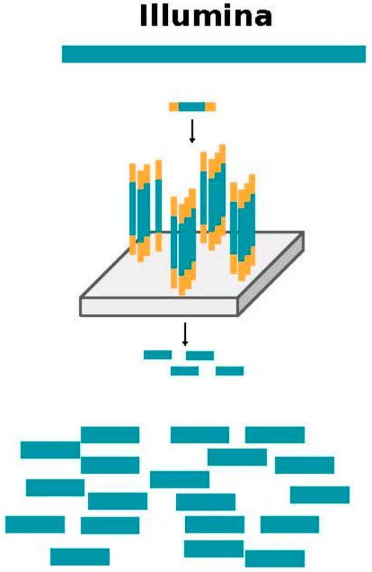


SDs → Nothing…

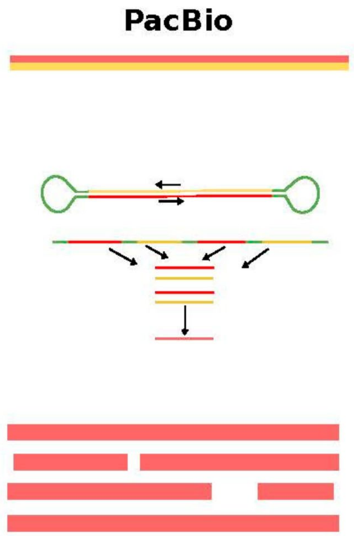
reads length from SRS is so short that:

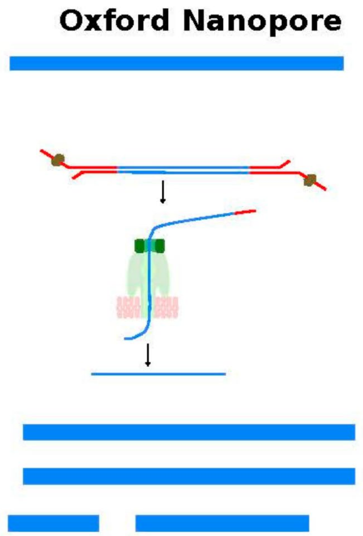
- can't span large structural variations (SVs)
- can't fully detect the isoforms
- challenging to reveal complex cis-acting chromatin states
- …

Long-Read Sequencing Is All You Need



read length is short? → let's get longer!





NGS sequencing → LRS sequencing

Genome

RNA-seq (bulk expression) → Iso-seq (transcripts)

Transcriptome

BS-seq → LRS-seq

Methylation Epigenome

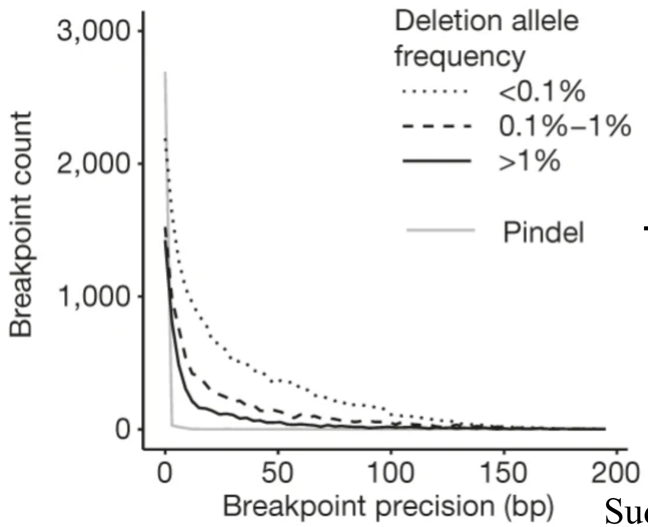
ATAC-seq & DNase-seq → Fiber-seq

Chromatin Epigenome

Long-Read Sequencing is promising

SRS limitations

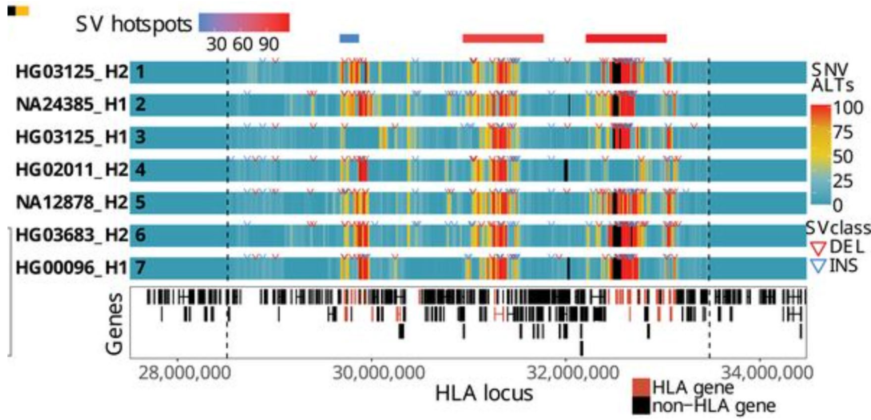
can’t span large SVs



break points

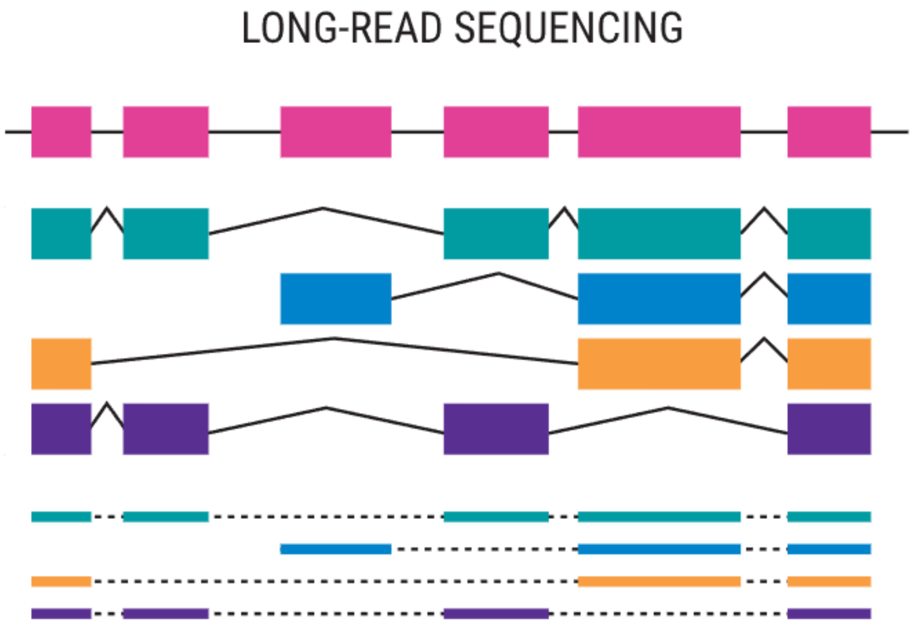
SVs haplotypes

Sudmant et al., *Nature*, 2015

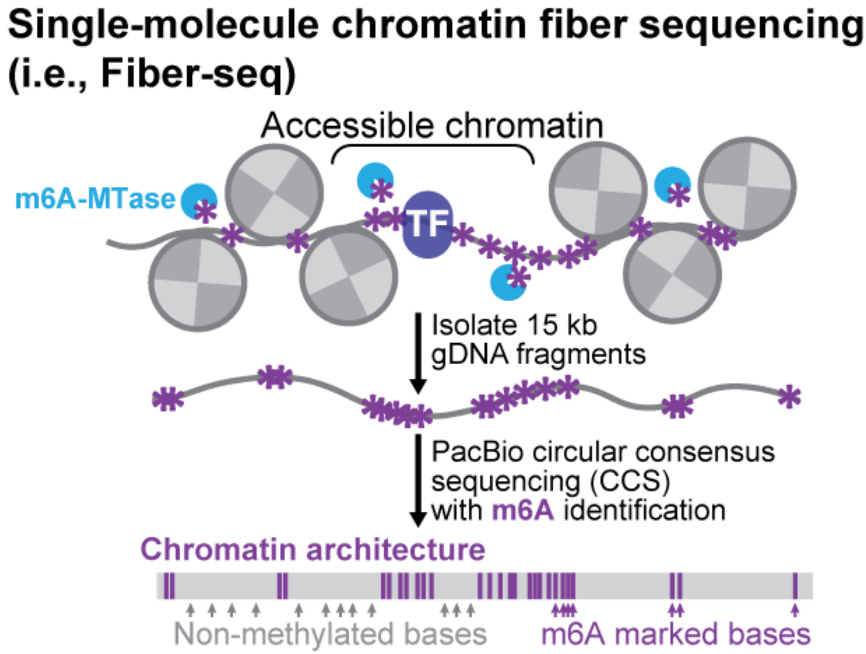


Ebert et al., *Science*, 2021

can’t fully detect the isoforms



challenging to reveal complex cis-acting chromatin states



Stergachis et al., *Science*, 2020

Resolving complex regions and detecting diseases

New Results

Follow this preprint

Genetic diversity and regulatory features of human-specific *NOTCH2NL* duplications

Taylor D. Real, Prajna Hebbar, DongAhn Yoo, Francesca Antonacci, Ivana Pačar, Mark Diekhans, Gregory J. Mikol, Oyeronke G. Popoola, Benjamin J. Mallory, Mitchell R. Vollger, Philip C. Dishuck, Xavi Guitart, Allison N. Rozanski, Katherine M. Munson, Kendra Hoekzema, Jane E. Ranchalis, Shane J. Neph, Adriana E. Sedeño-Cortes, Benedict Paten, Sofie R. Salama, Andrew B. Stergachis, Evan E. Eichler

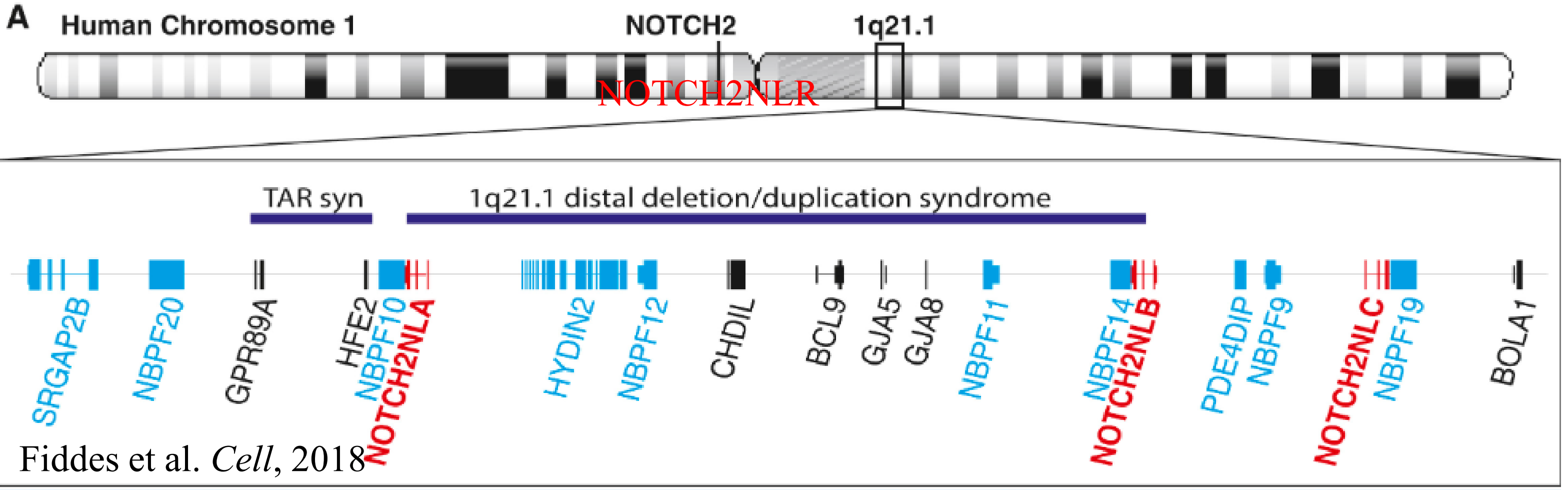
doi: https://doi.org/10.1101/2025.03.14.643395

nature genetics

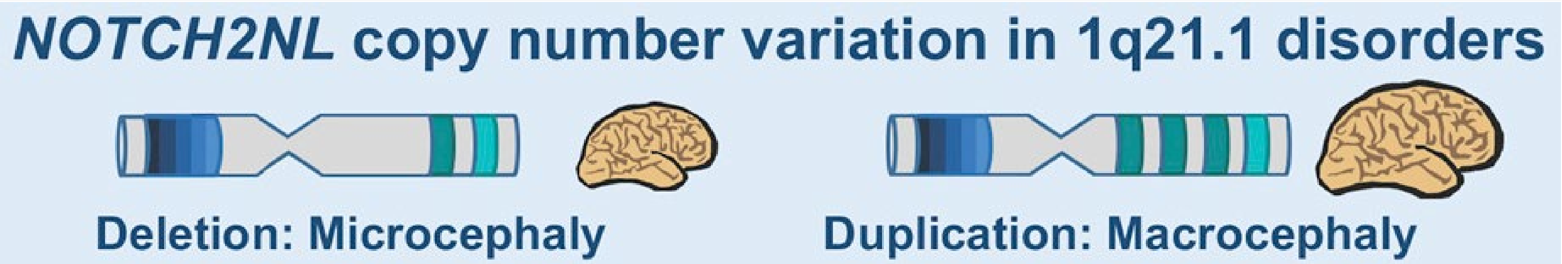
Technical Report https://doi.org/10.1038/s41588-024-02067-0

Synchronized long-read genome, methylome, epigenome and transcriptome profiling resolve a Mendelian condition

NOTCH2/NLs are associated with human brain evolution and a few genetic disorders



Fiddes et al. *Cell*, 2018



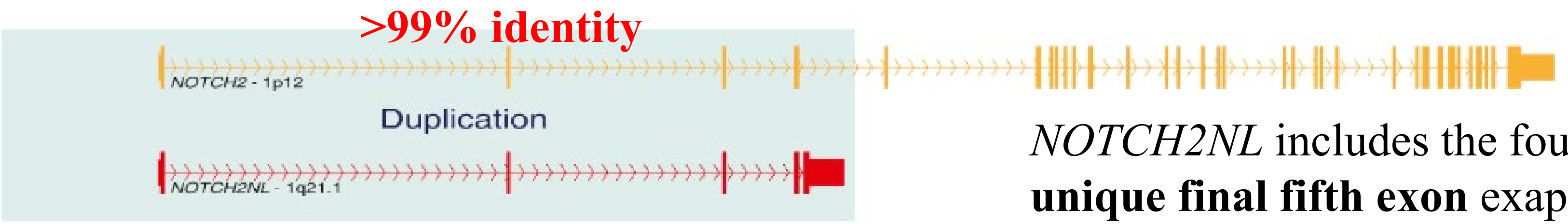
Function: increase neuronal mass during cortical neurogenesis

Disorders

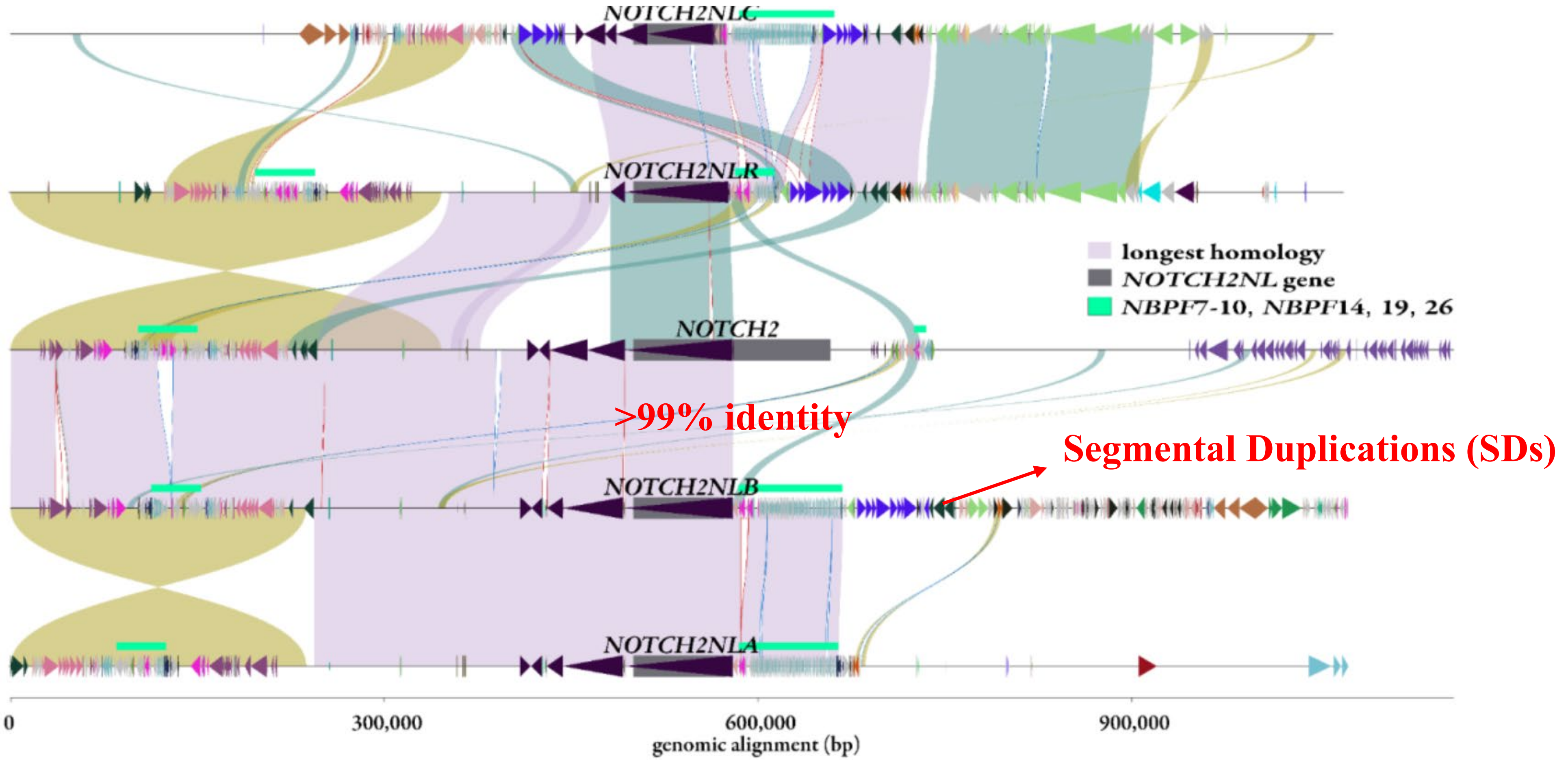
Alagille syndrome

TAR (thrombocytopenia-absent radius) syndrome

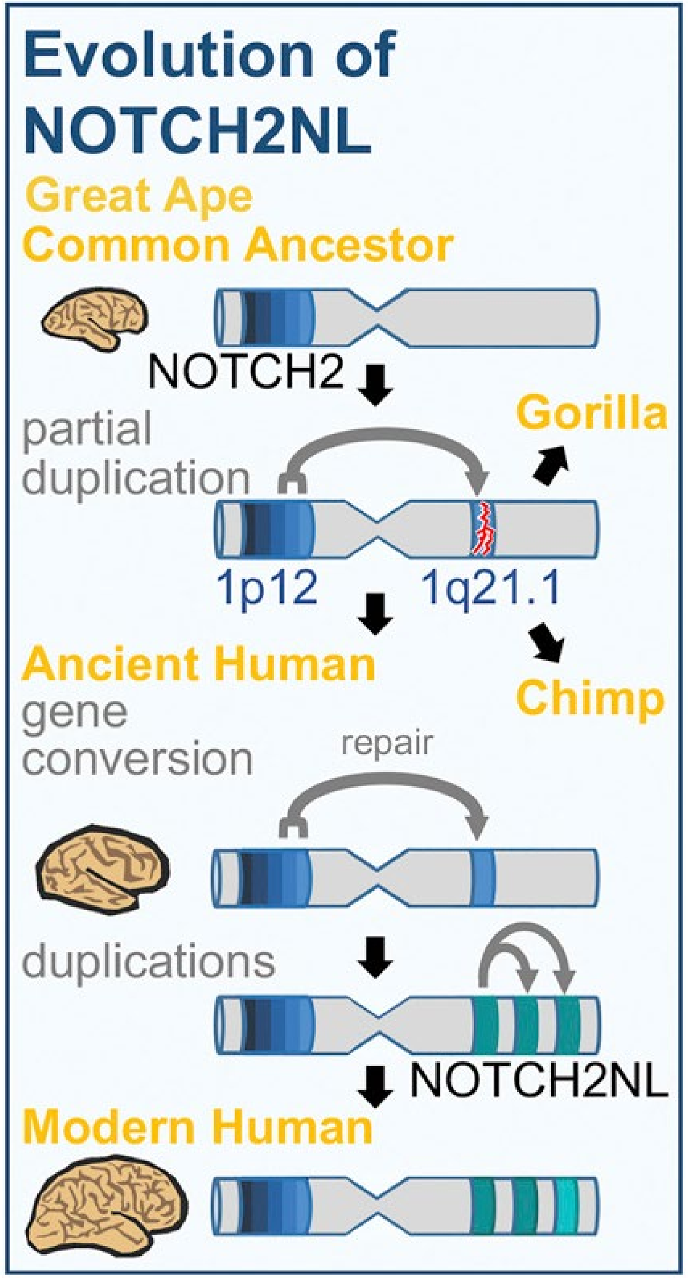
NOTCH2/NL locus is complex high-identity regions enriched in SDs and rearrangement



NOTCH2NL includes the four N-terminal exons of *NOTCH2* and includes a **unique final fifth exon** exapted from the fourth intron of *NOTCH2*



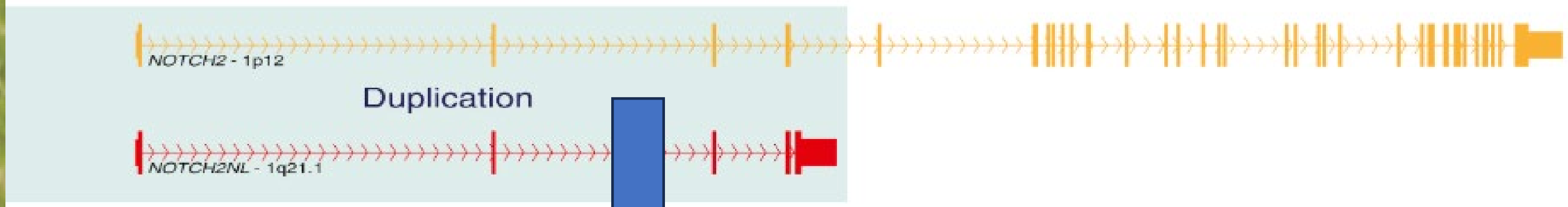
Independent *NOTCH2NL* duplications in apes



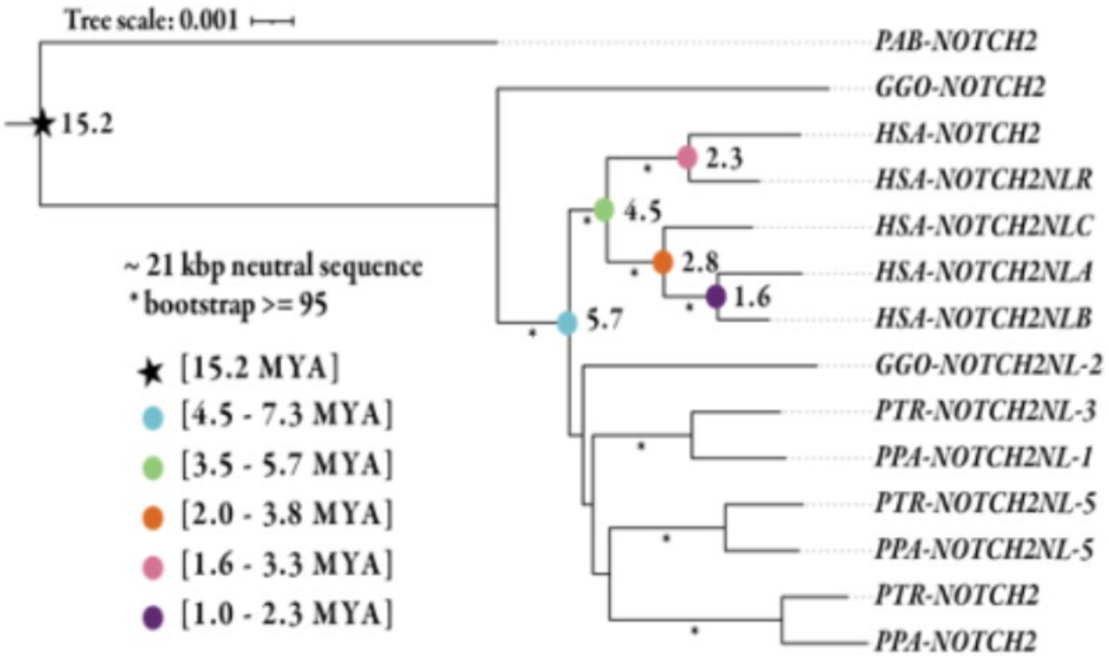
Duplications of *NOTHC2NL* had previously been noted in chimpanzee and gorilla, so why they can't have the function like humans?

Fiddes et al. *Cell*, 2018

Independent *NOTCH2NL* duplications in apes

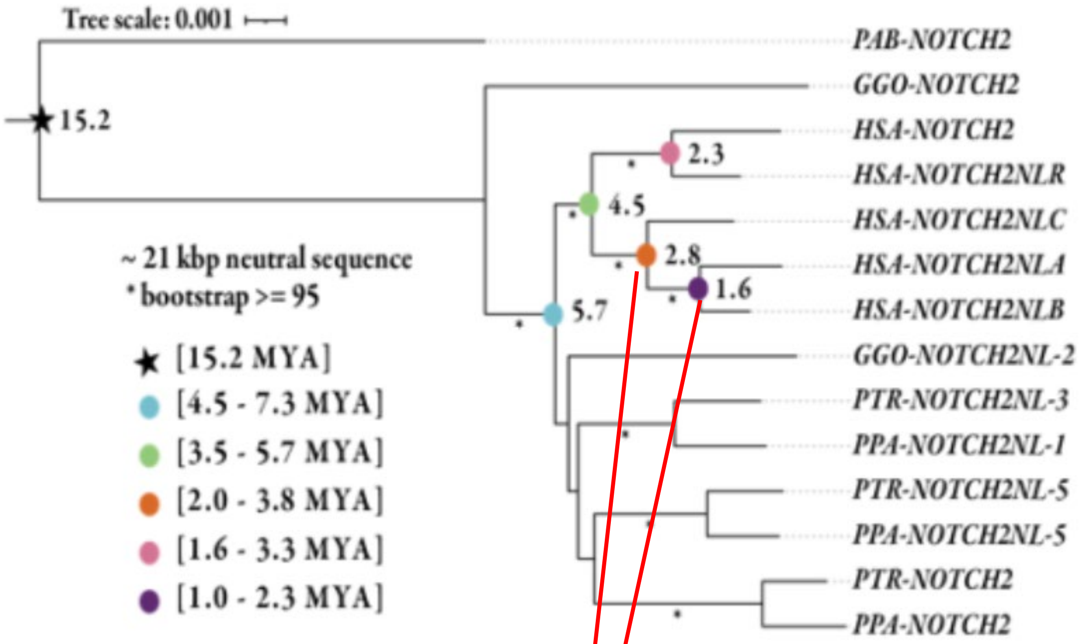


intron 2



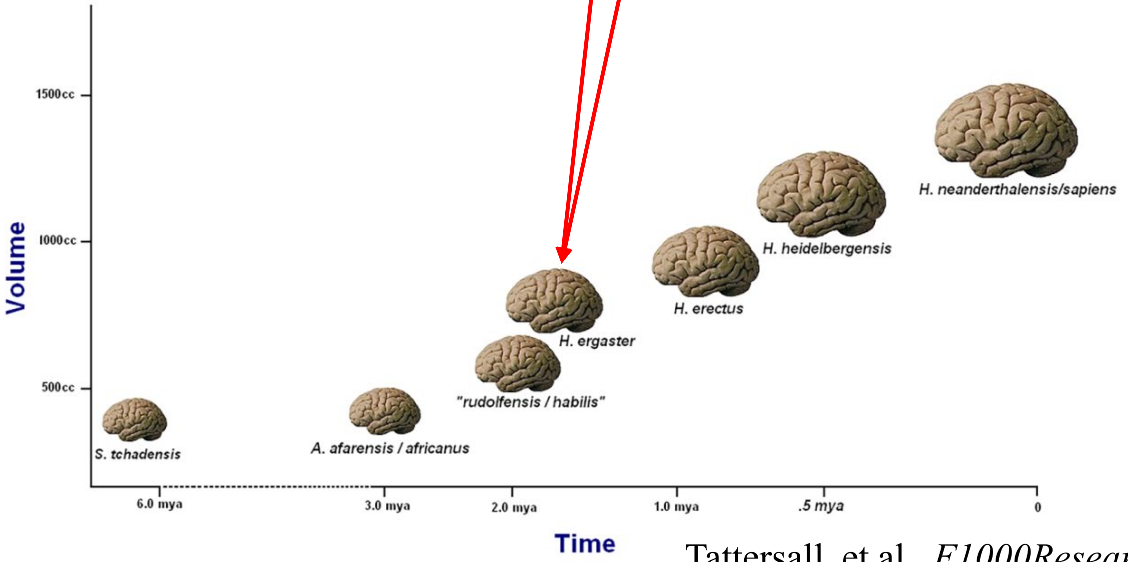
Recurrent expansions of *NOTCH2NL* in humans, the *Pan* genus, and gorillas

Independent *NOTCH2NL* duplications in apes



Using orangutan as an outgroup to estimate the timing of the duplications events in the human lineage:

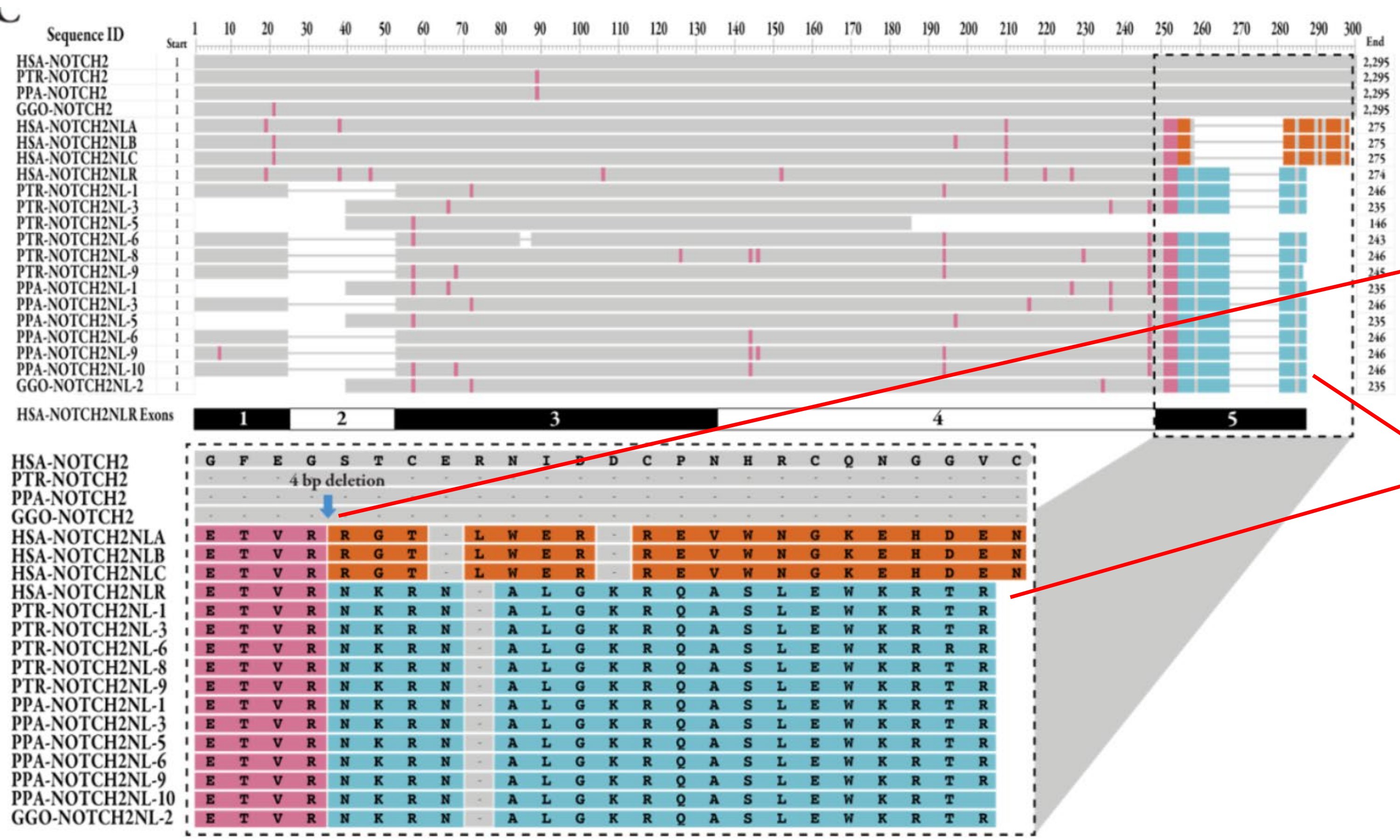
- About **2.8 MYA (2.0-3.8 MYA),** the human-specific copies began to diverge, distinguishing *NOTCH2NLC* from *NOTCH2NLA/B*.
- *NOTCH2NLA* and *NOTCH2NLB* appeared to have diverged around **1.6 MYA (1.0-2.3 MYA)**



The time of increasing in cranial volume taking place between **2.0-1.5 MYA** consistent with the diversification of NOTCH2NL genes in humans

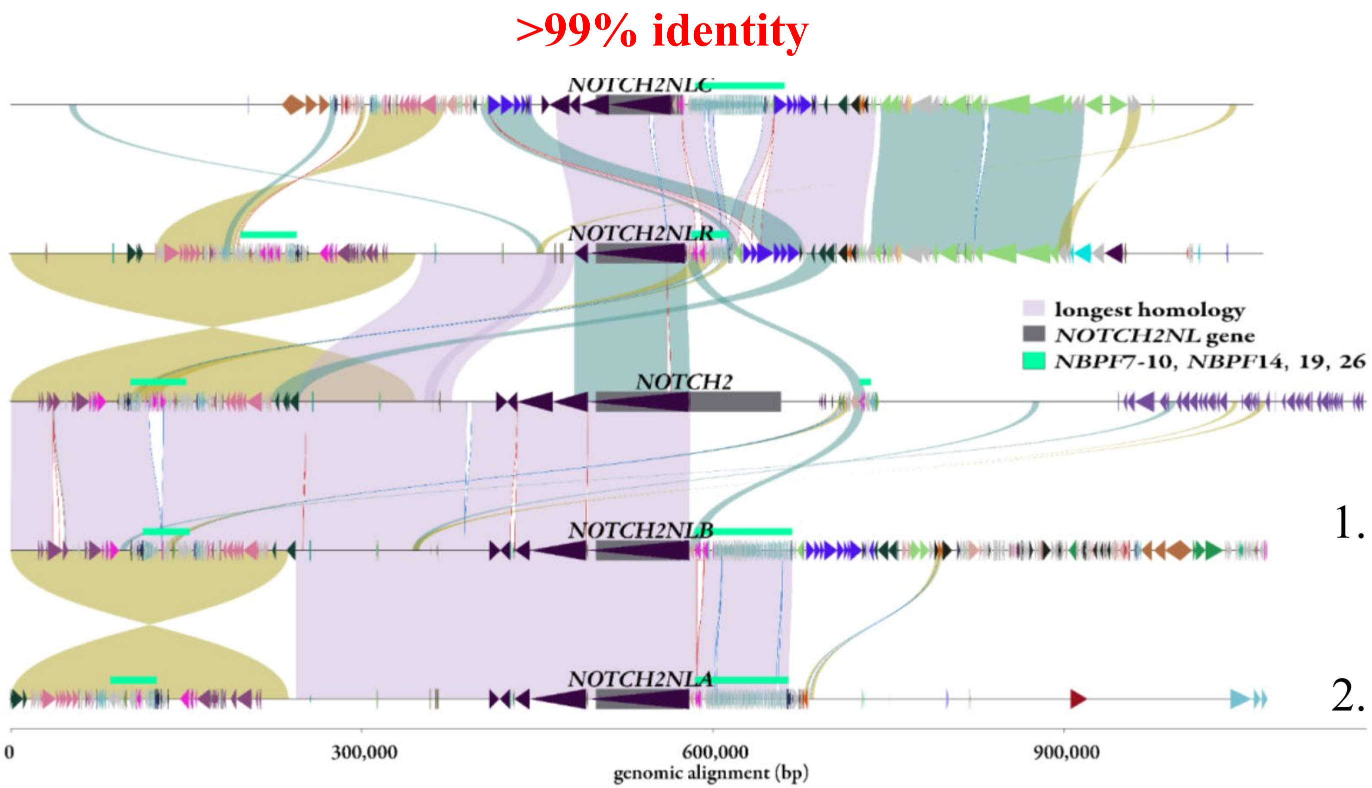
Tattersall, et al., *F1000Research*, 2023

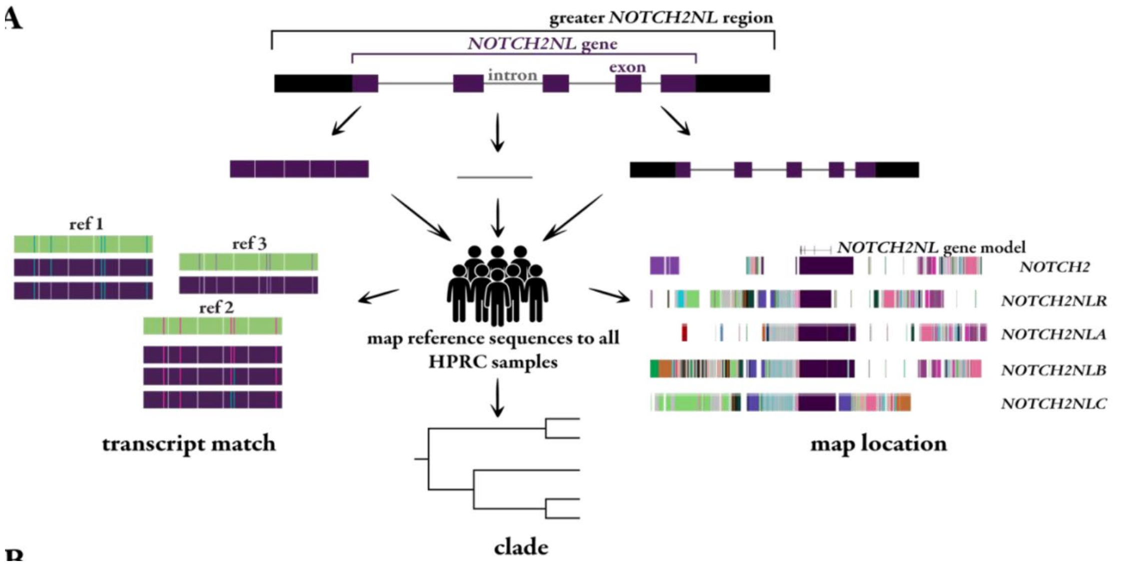
NHA (perhaps) have no functional *NOTCH2NL* transcripts



- **Humans appear to be the only species with *NOTCH2NL* transcripts that are predicted to make a stable protein**, likely because NHA copies lack the **4 bp deletion** that was found to be essential for *NOTCH2NLA/B/C* protein
- This 4 bp deletion modifies the final 19-20 AAs of the carboxy terminus in **not just a paralog-specific, but also human-specific**, fashion
- However, we cannot definitively comment on the functional role of NHA transcripts, most of which have strong Iso-Seq support and maintain reasonable ORFs

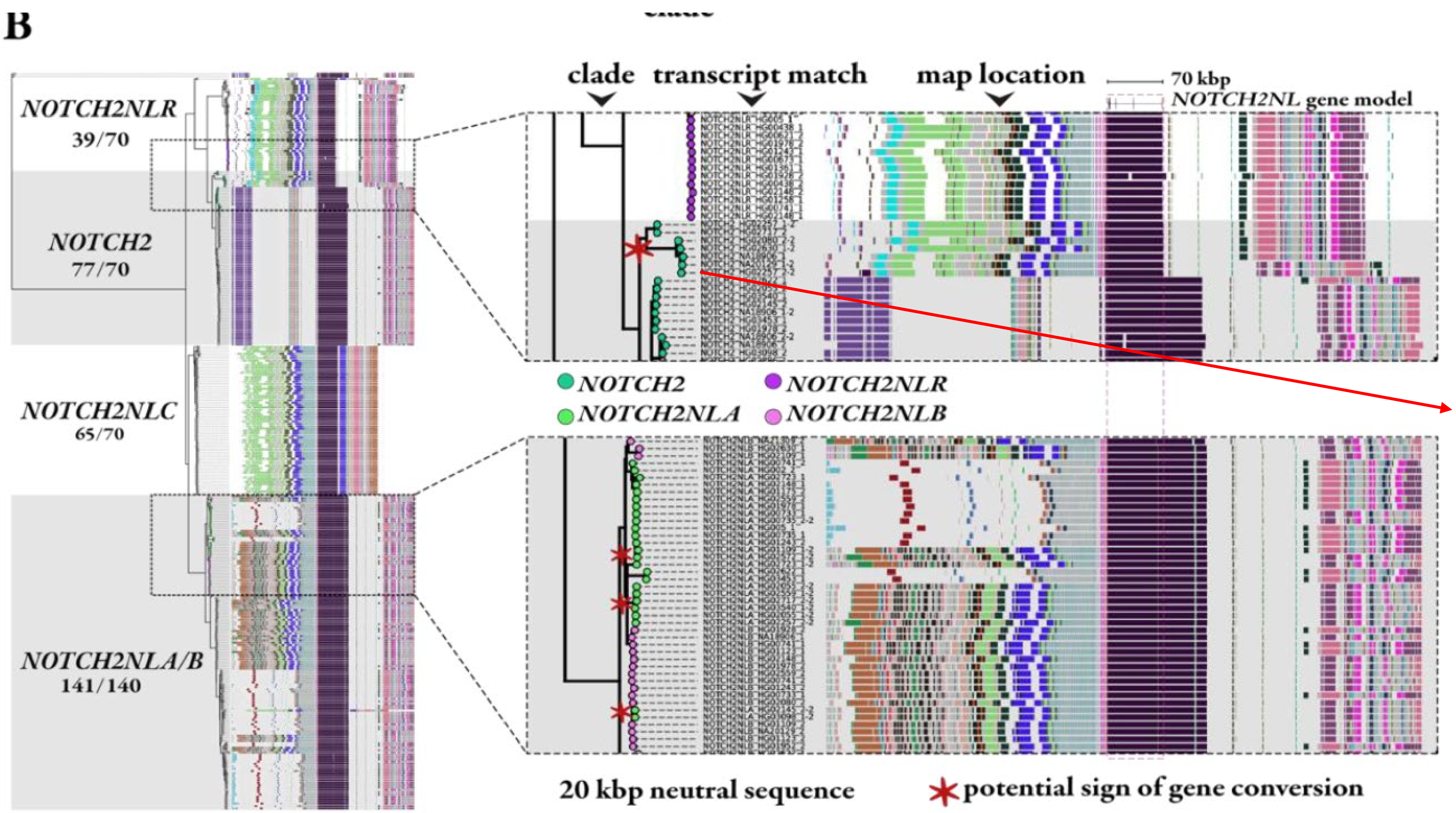
Identifying *NOTCH2/NL* structural haplotypes in human pangenome





1. First, we examined the best transcript match by identifying which NOTCH2NL coding sequence best matches NOTCH2NL copies assigned in the T2T-CHM13 reference
2. Second, we used NOTCH2NL intronic sequence to construct a tree identifying a phylogenetic framework **for each NOTCH2NL haplotype assigning different haplotypes to related clades**
3. Third, we used the **extended duplication organization as defined by the DupMasker barcode to examine the long-range organization of the region flanking NOTCH2NL.**

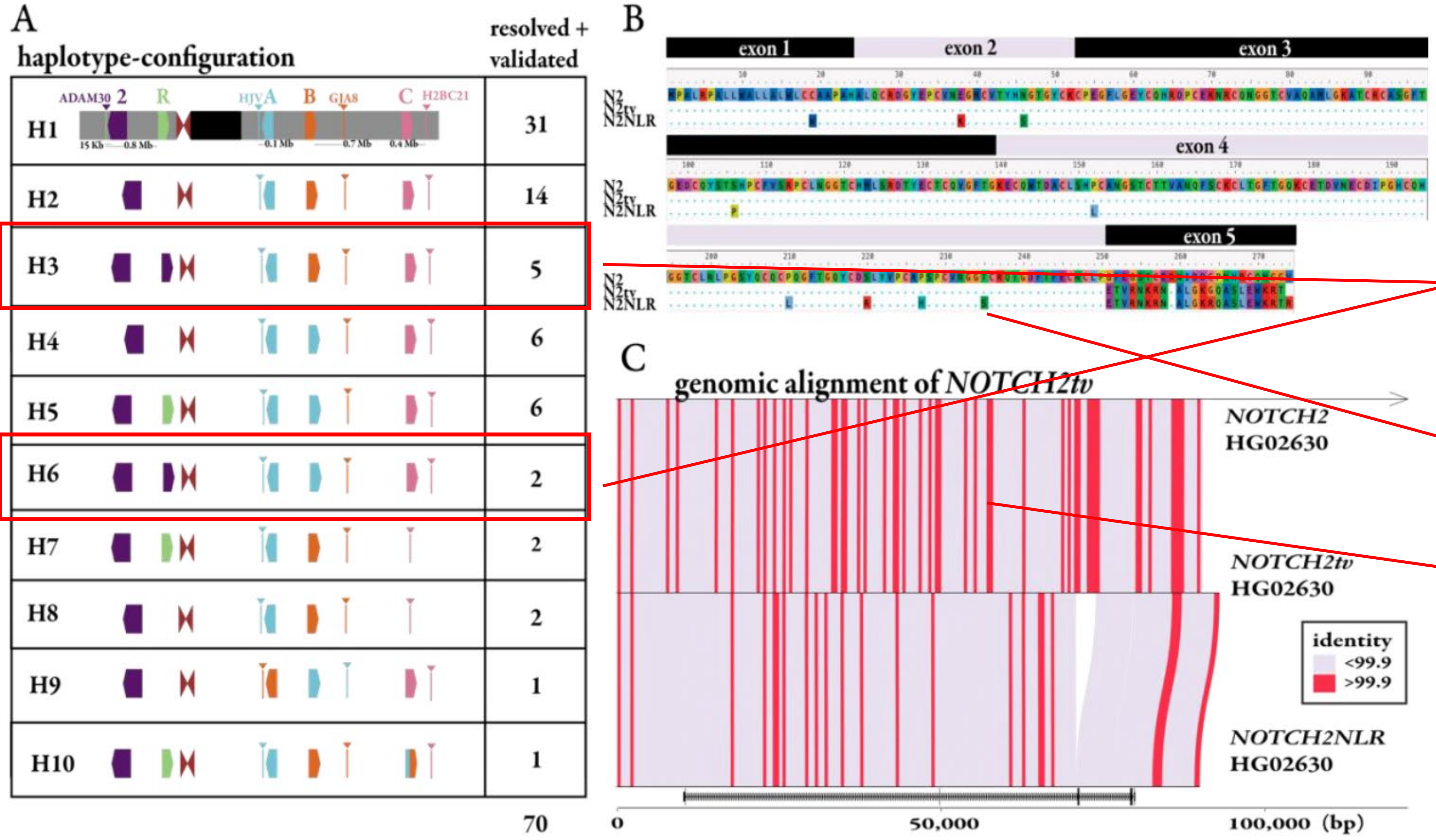
Identifying *NOTCH2/NL* structural haplotypes in human pangenome



NOTCH2 → *NOTCH2NLR*

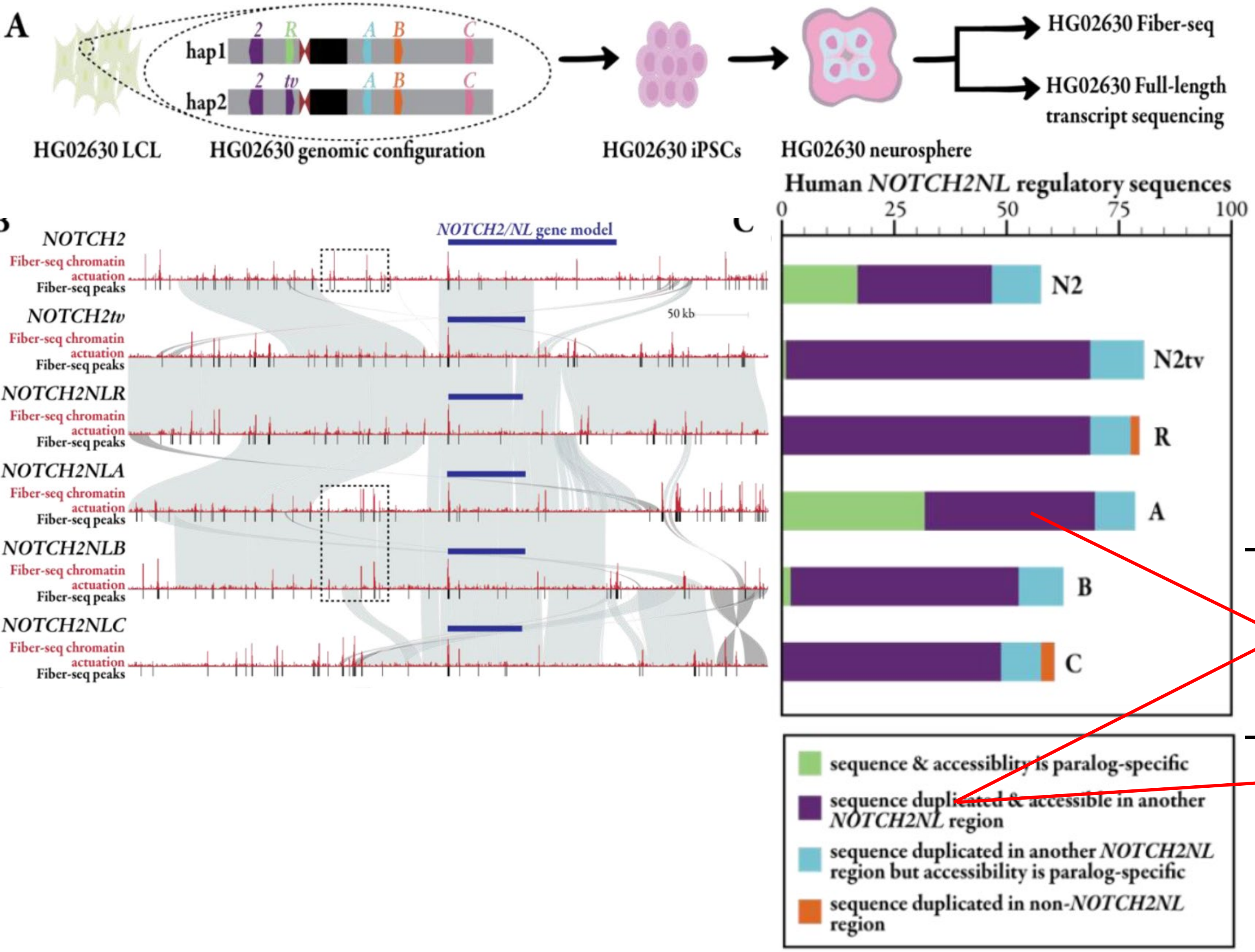
transcript matchs *NOTCH2* but flanking region mirrors *NOTCH2NLR*

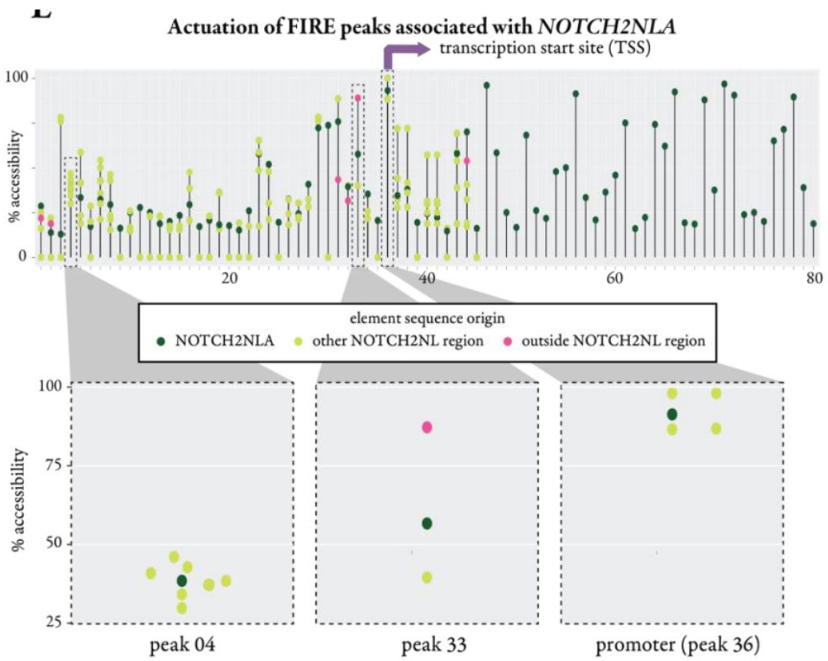
Identifying *NOTCH2/NL* structural haplotypes in human pangenome



- New gene conversion event between *NOTCH2* and *NOTCH2NLR* resulting a truncated version of *NOTCH2*, defined as ***NOTCH2tv***
- *NOTCH2* share the AA changes with *NOTCH2tv* in the first four exons
- More >99% identity bins between *NOTCH2* and *NOTCH2tv* than between the *NOTCH2tv* and *NOTCH2NLR*

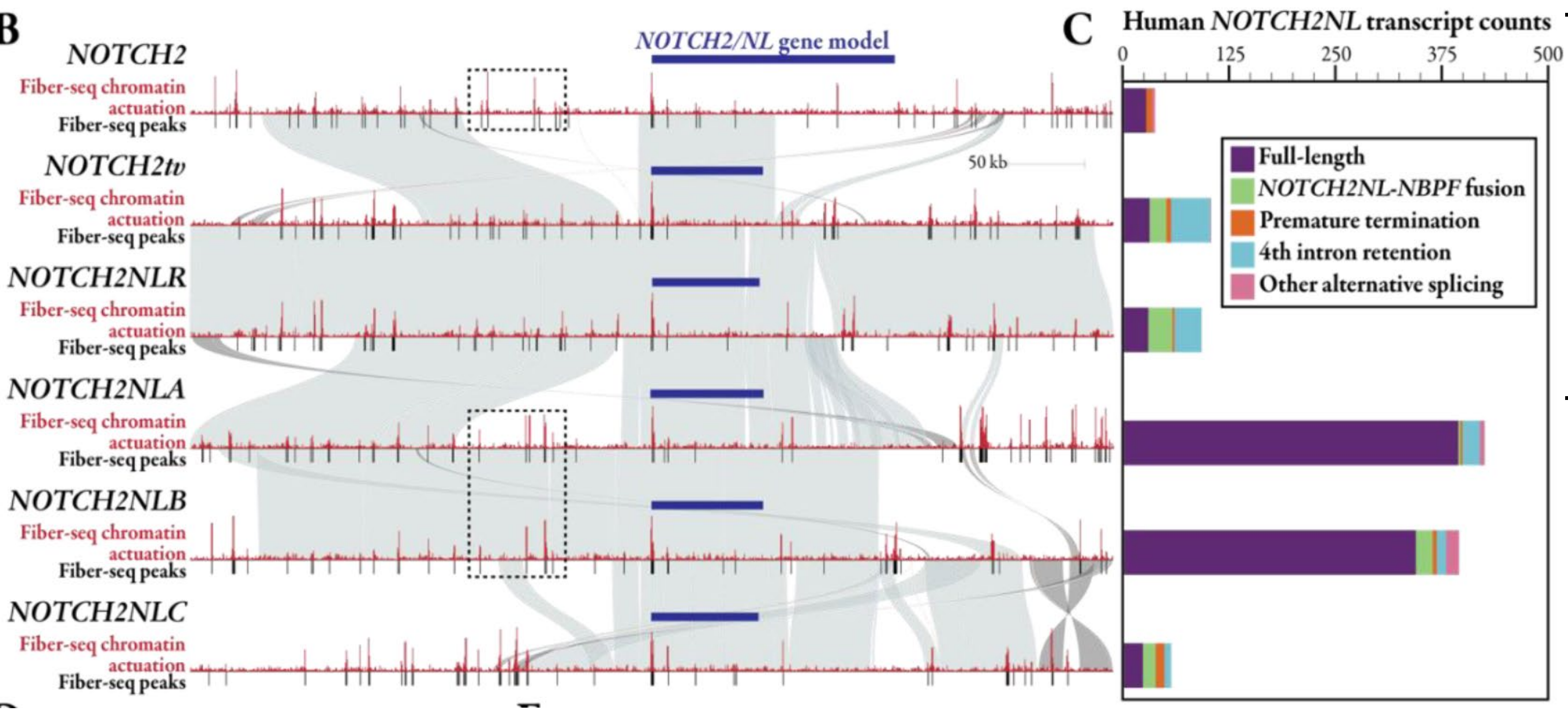
Regulatory landscape of *NOTCH2/NL* paralogs

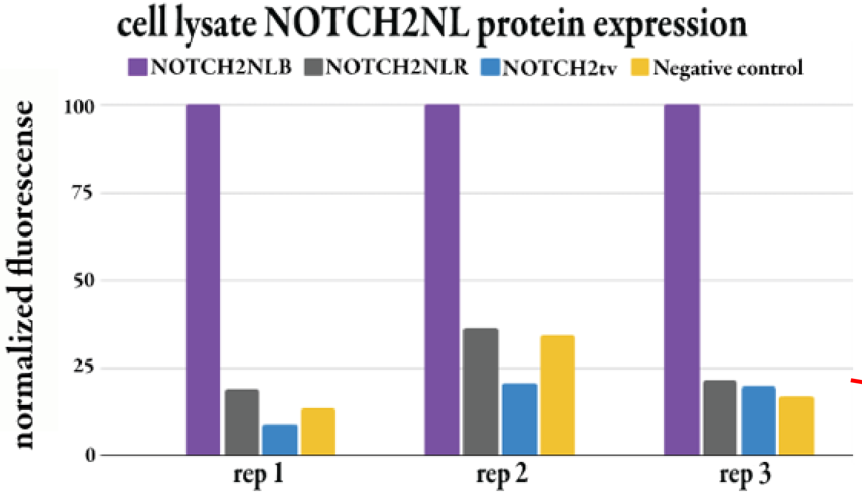




- The accessible chromatin landscape surrounding each *NOTCH2NL* paralog appears predominantly populated by these multi-paralog accessible chromatin elements
- The degree of chromatin accessibility can vary quite substantially between the two duplicates, suggesting the predominant effect was **quantitative differences in chromatin accessibility** as opposed to drastic changes to on/off actuation
- …

Expression of *NOTCH2/NL* paralogs





- Distinct differences within the transcript abundance of each of the *NOTCH2* paralogs were found, indicating that these paralog-specific accessible chromatin elements may be creating unique gene regulatory environments for each of the *NOTCH2* paralogs
- Although the promoter and transcript sequence of *NOTCH2tv* mirrors that of *NOTCH2*, **the transcript abundance and composition of *NOTCH2tv* appeared to mirror most closely that of *NOTCH2NLR***
- This indicates that despite the transcript identity of *NOTCH2tv* matching the first four exons of *NOTCH2*, **the surrounding gene regulatory architecture in fact mirrors that of *NOTCH2NLR***, potentially impacting the overall function of *NOTCH2tv*
- *NOTCH2tv* was similarly unable to form a stable protein product like *NOTCH2NLR*

Take home message:

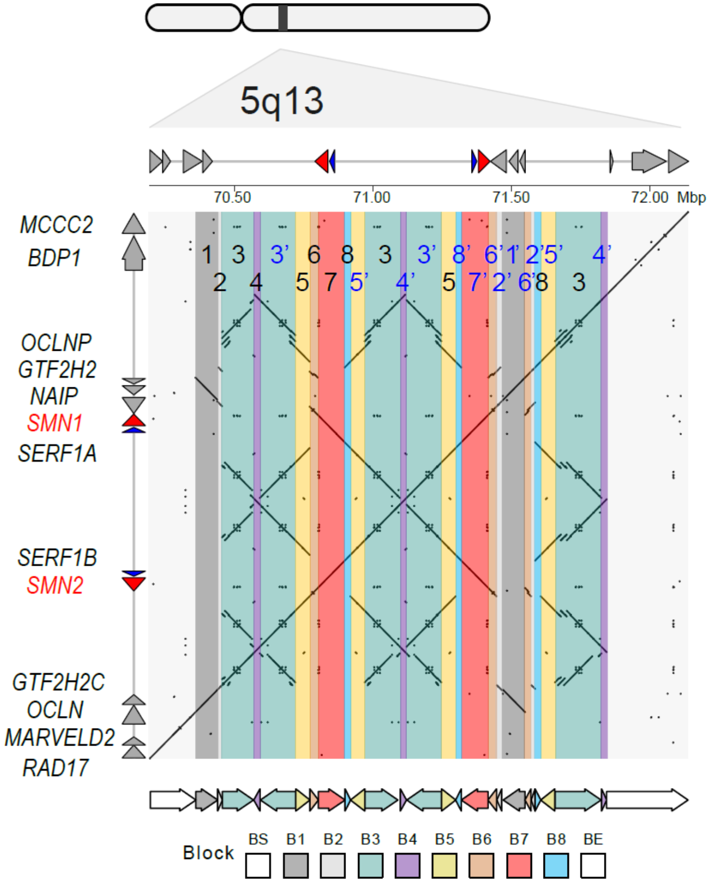
- *NOTCH2/NL* was associated with human brain evolution and a few genetic disorders
- *NOTCH2NL* had undergone independent duplications in great apes but there were functional copies only in humans
- **Utilizing gene sequence, phylogenic tree, and flanking regions might be helpful for distinguishing all the high-identity gene paralogs and gene conversion events**
- **The degree of chromatin accessibility can vary quite substantially**
- The surrounding gene regulatory architecture was important

Q.A.

Long-read sequencing enhances the detection and resolution for complex diseases

脊髓性肌萎缩症 （SMA） 是一种罕见的神经肌肉疾病，可导致运动神经元丧失和进行性肌肉萎缩。通常在婴儿期或儿童早期被诊断出来，如果不及时治疗，它是婴儿死亡的最常见遗传原因。它也可能在以后的生活中出现，然后病情较轻。共同特征是随意肌进行性无力，手臂、腿部和呼吸肌首先受到影响。 相关问题可能包括头部控制不佳、吞咽困难、脊柱侧凸和关节挛缩。

脊髓性肌萎缩症是由于 SMN1 基因的异常（突变），该基因编码 SMN，SMN 是运动神经元生存所必需的蛋白质。脊髓中这些神经元的丢失会阻止大脑和骨骼肌之间的信号传递。另一个基因 SMN2 被认为是疾病修饰基因，因为通常 SMN2 拷贝越多，病程越轻。SMA 的诊断基于症状并通过基因检测确认。



LRS could also be used for the detection and resolution for **complex diseases** because lots of pathogenic genes were located in **complex regions which are inaccessible to SRS**

Synchronized long-read multi-ome profiling

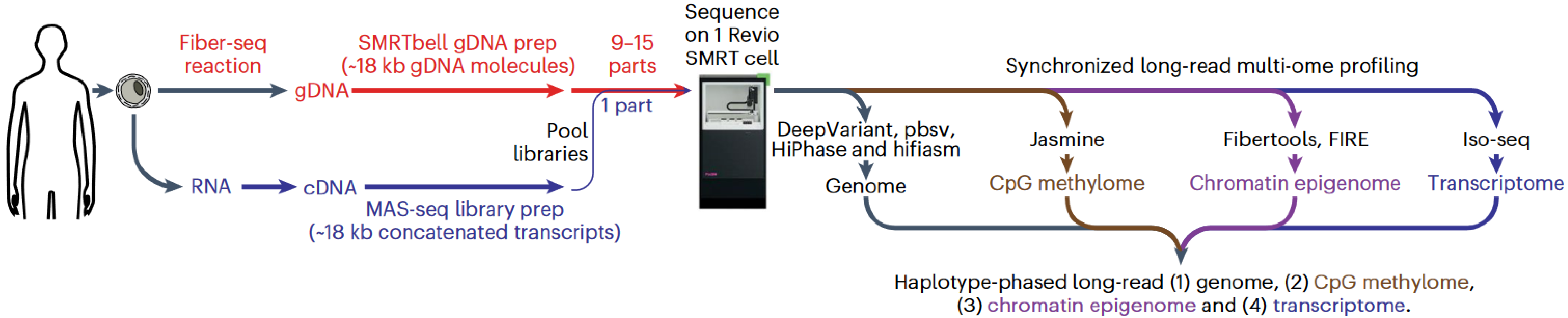
nature genetics

Technical Report

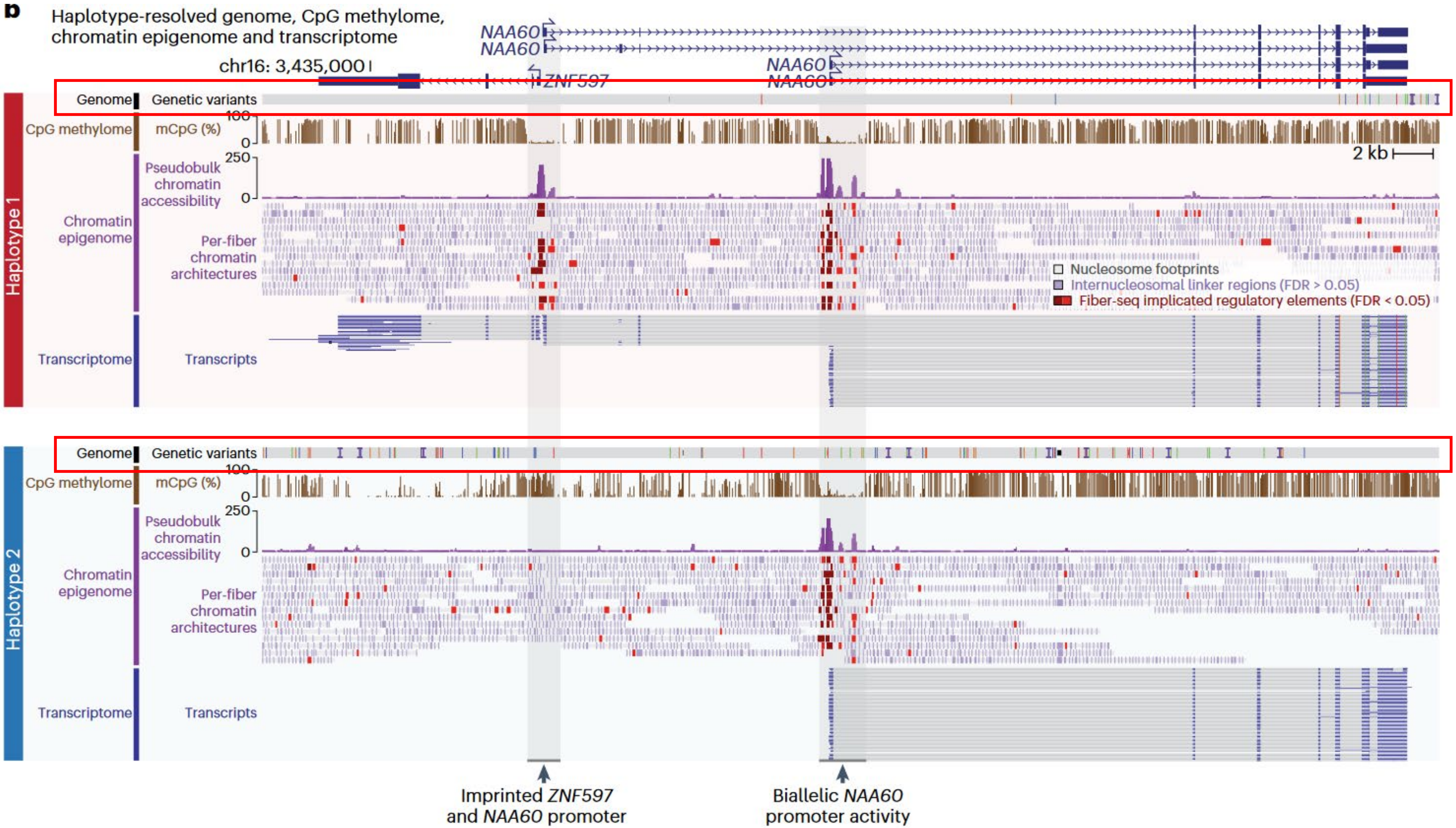
https://doi.org/10.1038/s41588-024-02067-0

Synchronized long-read genome, methylome, epigenome and transcriptome profiling resolve a Mendelian condition

- Cells are subjected to a Fiber-seq reaction followed by gDNA extraction and SMRTbell library preparation
- In parallel, cells are subjected to an RNA extraction followed by cDNA synthesis and MAS-seq (multiplexed arrays isoform sequencing) library preparation.
- The two libraries are then mixed together and sequenced simultaneously using a single sequencing run

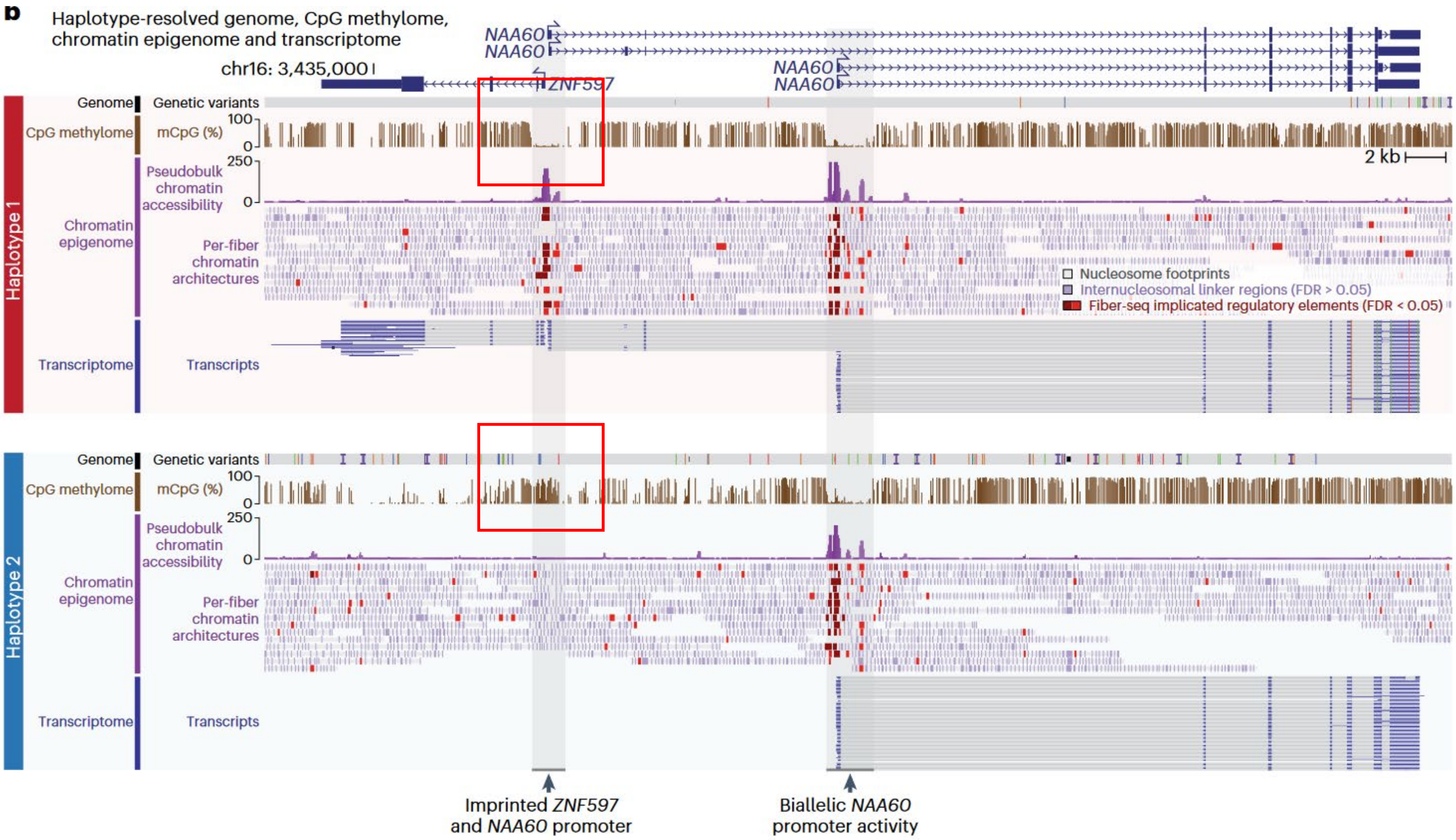


Synchronized long-read multi-ome profiling



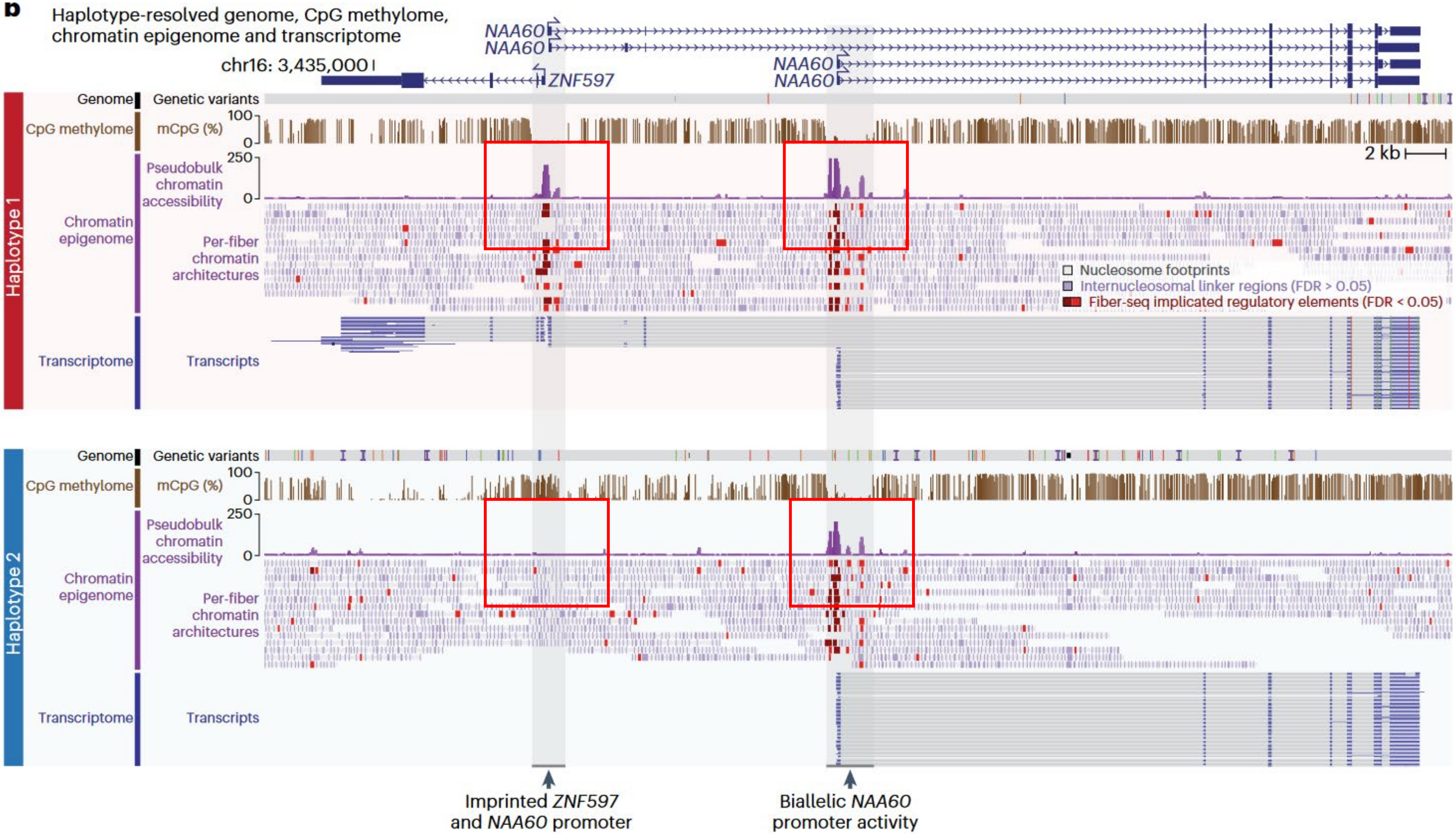
Haplotype-resolved genome implied by genetic variants

Synchronized long-read multi-ome profiling



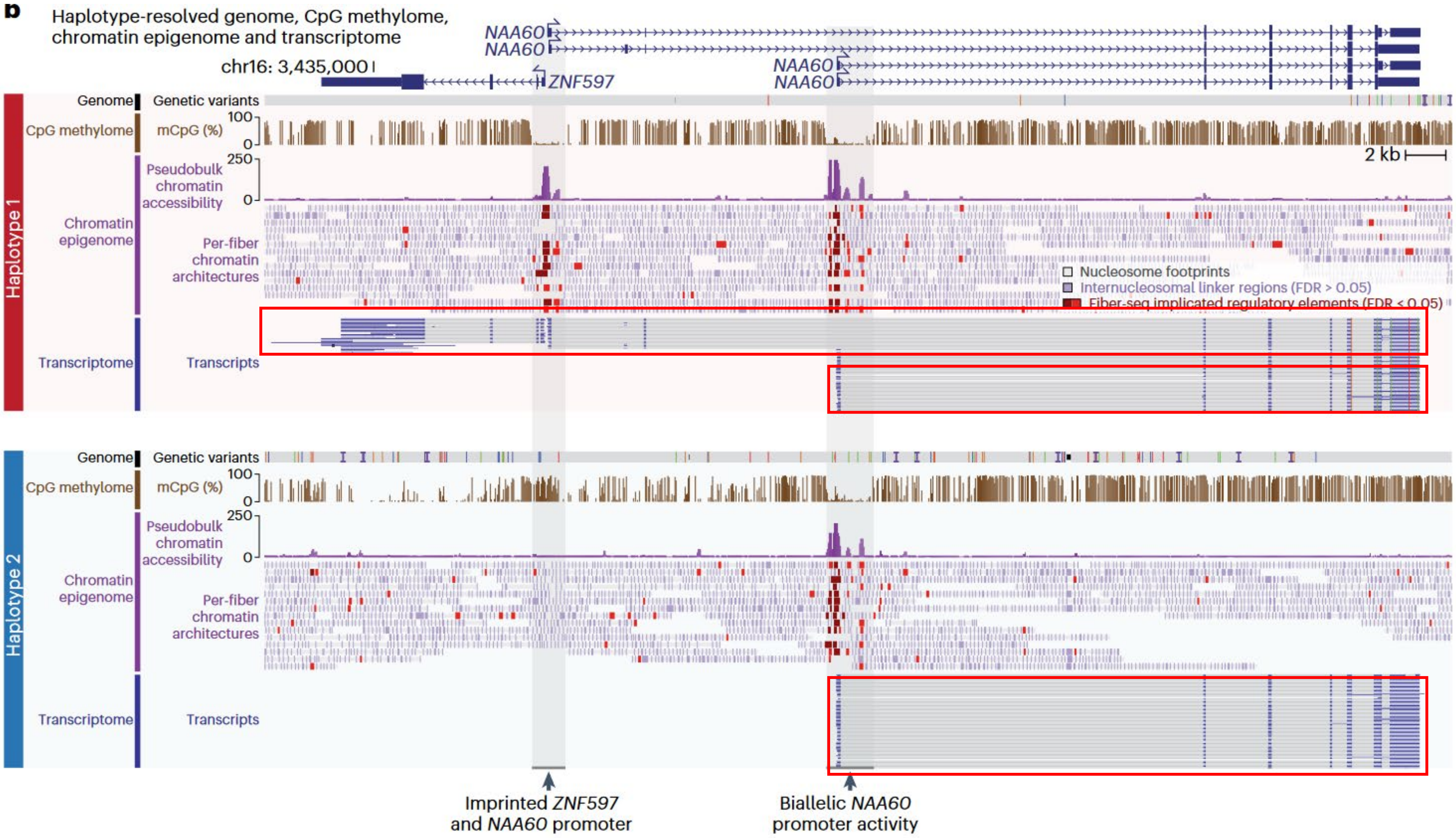
Haplotype-resolved CpG methylome implied by imprinted locus

Synchronized long-read multi-ome profiling



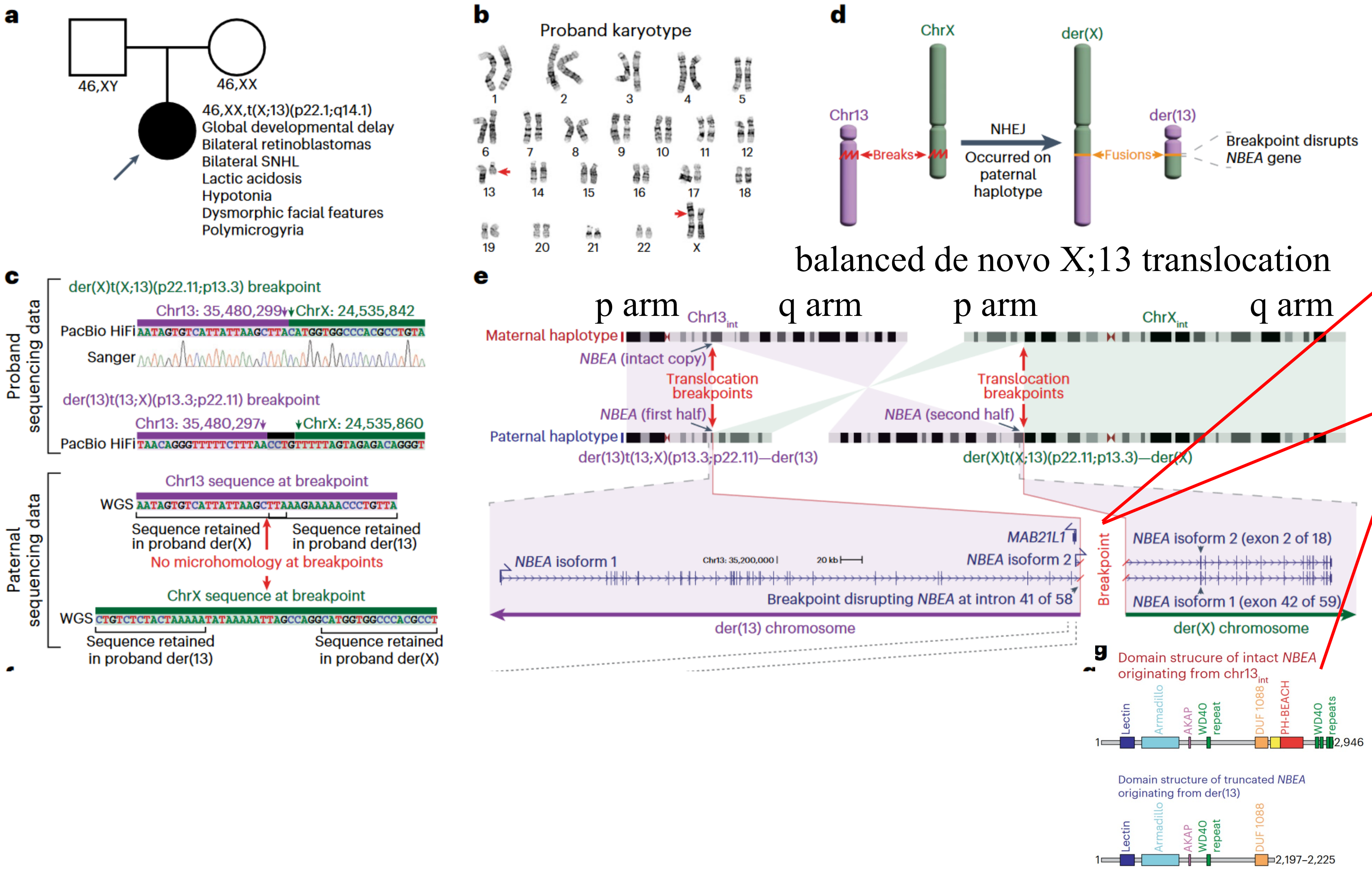
Haplotype-resolved chromatin epigenome implied by the degree of chromatin accessibility

Synchronized long-read multi-ome profiling



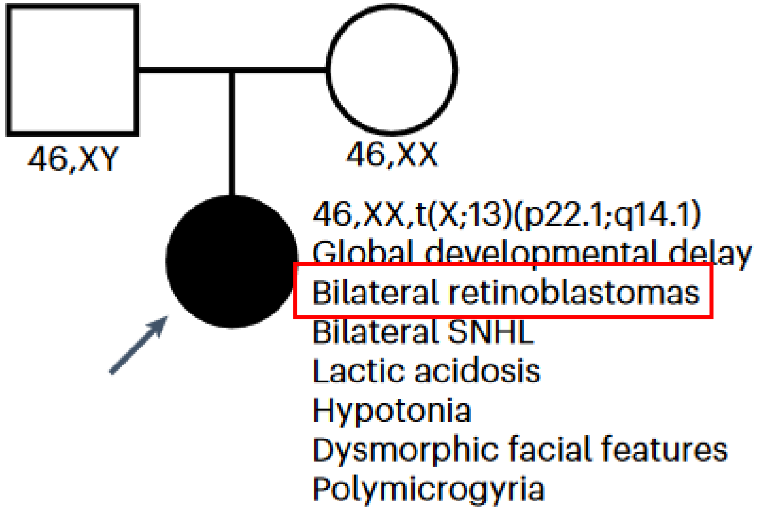
Haplotype-resolved transcriptome implied by full-length isoforms

Long-read genome exposes *NBEA* disruption



- Genome assembly of these sequencing reads delineated the precise translocation breakpoints
- The translocation breakpoint on 13q is located in intron 41 of the 58-exon gene *NBEA*
- This resulted in the formation of a **truncated *NBEA* protein product**
- Such a truncation could have unique impacts on dendrite (树突) formation

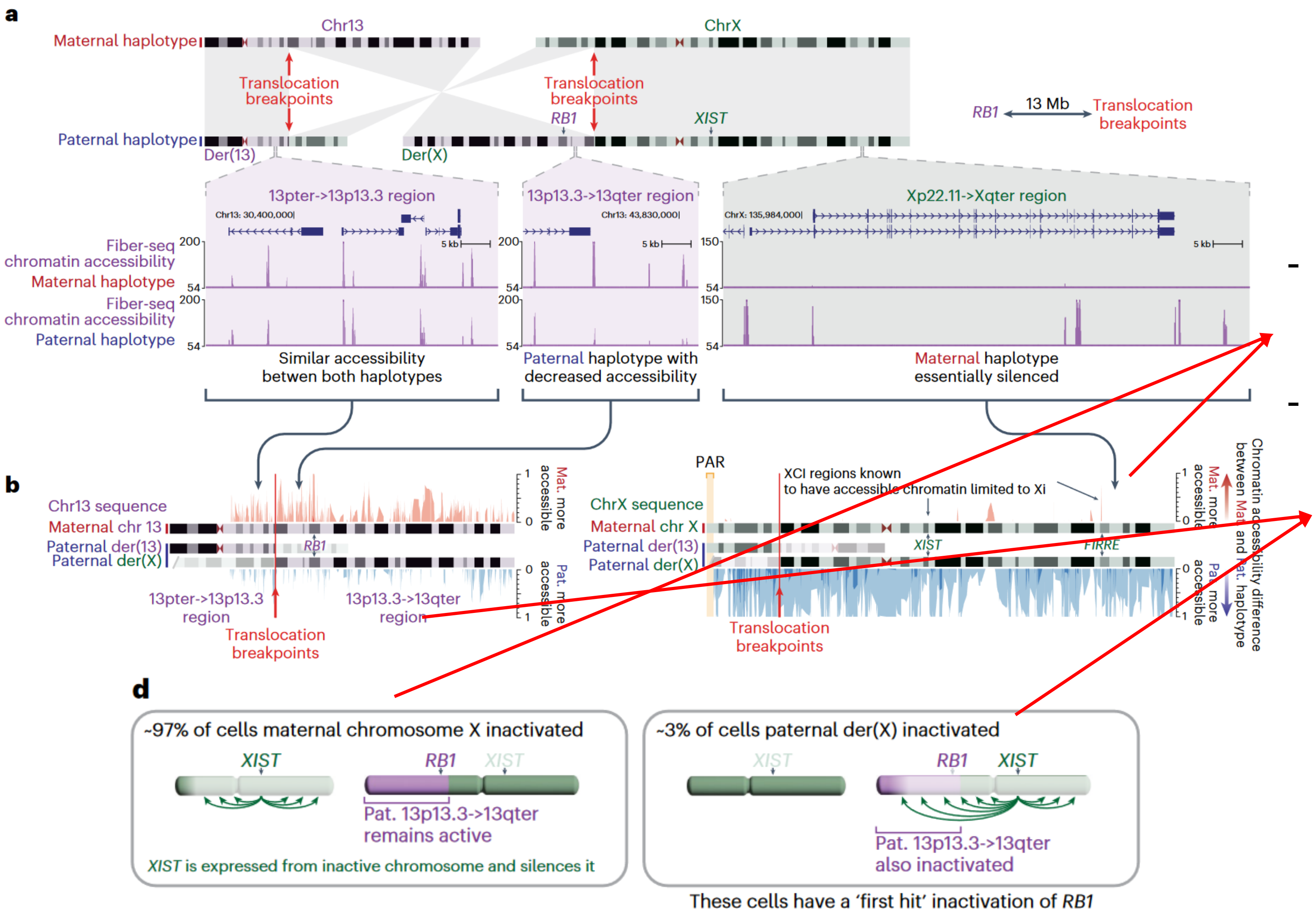
Long-read multi-ome exposes inappropriate XCI of autosomal genes



双侧视网膜母细胞瘤：视网膜母细胞瘤最常见和最明显的体征是通过瞳孔观察视网膜的异常外观，其医学术语是白瞳。其他体征和症状包括视力恶化、青光眼，眼睛发红和发炎，以及生长迟缓或发育迟缓。一些患有视网膜母细胞瘤的儿童可能会出现斜视，通常被称为"斗鸡眼"。

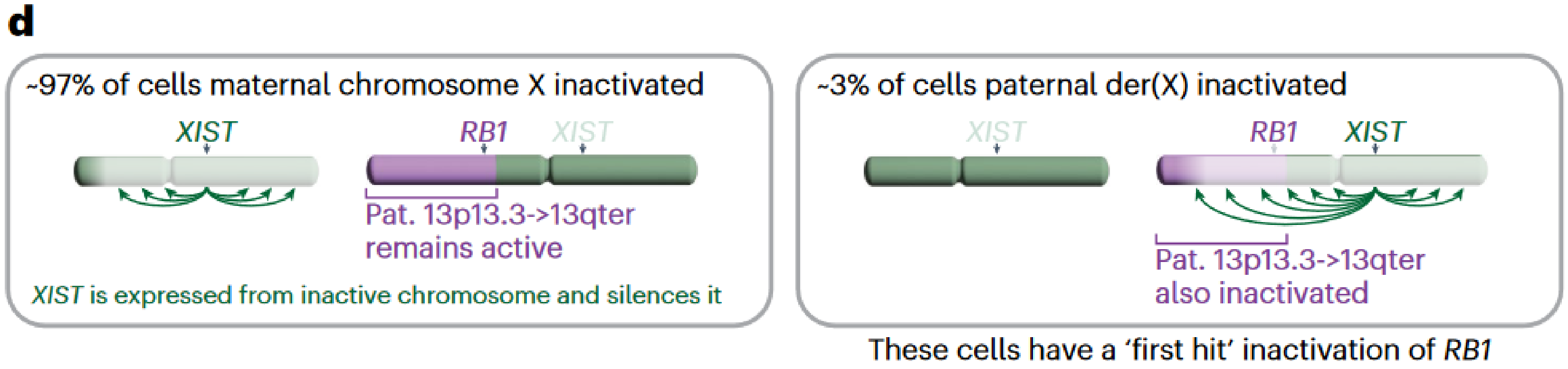
RB1 is the only gene associated with hereditary bilateral retinoblastomas, yet *RB1* is located 13.5 Mb away from the breakpoint along der(X)

Long-read multi-ome exposes inappropriate XCI of autosomal genes



- Intact chromosome X was preferentially subjected to (X-chromosome inactivation) XCI
- The 13p13.3 to 13qter region selectively and significantly exhibited an imbalance in allelic chromatin accessibility with this region on der(X) being silenced in 3–10% of cells

Long-read multi-ome exposes inappropriate XCI of autosomal genes



1. 理解RB1基因与视网膜母细胞瘤

- **RB1是抑癌基因：** *RB1*是一个经典的抑癌基因。它的正常功能是抑制细胞过度生长，防止肿瘤的发生。
- **两次打击假说（Two-Hit Hypothesis）：** 要使一个细胞癌变，通常需要其两个拷贝（一个来自父亲，一个来自母亲）的抑癌基因**都**失去功能。
  - **第一次打击（First Hit）：** 一个等位基因失活。
  - **第二次打击（Second Hit）：** 另一个等位基因失活。
  - 只有当"两次打击"都发生在同一个细胞中时，这个细胞才会失去抑癌基因的保护，从而走向癌变。

2. 区分遗传性与散发性视网膜母细胞瘤

- **散发性（通常是单侧）：** 患儿出生时，所有细胞的两个*RB1*等位基因都是完好的。需要在一个视网膜细胞中，偶然地、先后发生两次独立的体细胞突变（第一次和第二次打击），才会形成肿瘤。这个"双重偶然"事件的概率极低，因此通常只发生在一只眼睛，且发病年龄较晚。
- **遗传性（通常是双侧）：** 患儿出生时，全身的**每一个细胞**中都已经携带了一个失活的*RB1*等位基因。这就是遗传来的**"第一次打击"。**因此，他/她全身数百万个视网膜细胞都处于"一触即发"的状态。只需要在任何一个视网膜细胞中，那个唯一剩下的、功能完好的*RB1*等位基因再发生一次随机的体细胞突变（"第二次打击"），肿瘤就会形成。因为全身细胞都有这个"第一次打击"，所以"第二次打击"在双眼独立发生的概率非常高，导致患者常常在幼年就出现**双侧或多发性**的肿瘤。

Synchronized long-read multi-omic profiling mechanistically resolved complex phenotypes

Table 2 | Overview of molecular variants identified in UDN318336 via multi-ome long-read sequencing

Molecular event	Ome(s) required for identification	Associated clinical phenotypes in UDN318336	Proposed mechanism	Overlapping Mendelian condition
NBEA–chrX fusion transcripts	Genome; transcriptome		*NBEA* haploinsufficiency and/or production of truncated protein	OMIM 619157
PDK3–*MAB21L1* fusion kinase transcript	Genome; transcriptome	Polymicrogyria, SNHL, developmental delay, lactic acidosis and hypotonia.	Overexpression of *PDK3* in tissue that endogenously expresses *MAB21L1*. Potentially altered regulation of PDK3–MAB21L1 fusion	OMIM 300905; OMIM 312170
PDK3 adoption of *MAB21L1* enhancer and subsequent *PDK3* ectopic GOE	Chromatin epigenome			
XCI of *RB1* locus	Chromatin epigenome	Bilateral retinoblastomas	'First hit' in the development of biallelic *RB1* LOF	OMIM 180200
Transcriptional readthrough silencing of *MAB21L1*	CpG methylome; chromatin epigenome; transcriptome	No effect on patient phenotype as only one *MAB21L1* haplotype impacted, with other haplotypes demonstrating intact gene regulation	N/A	OMIM 618479

In short, LRS multi-omic profiling revealed that this translocation disrupted the functioning of four separate genes (*NBEA*, *PDK3*, *MAB21L1* and *RB1*) previously associated with single-gene Mendelian conditions. Notably, the function of each gene was disrupted via a distinct mechanism that required integration of the four 'omes' to resolve.

Long-read sequencing v2.0

Sequence more, learn more

Nurk, et al., *Science*, 2022

Liao, et al., *Nature*, 2023

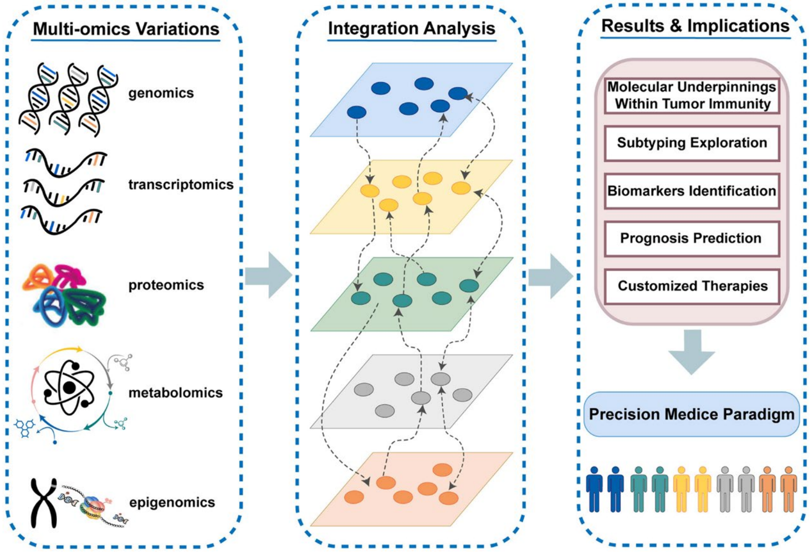
A head start

The Earth BioGenome Project could coordinate the efforts below and others that are already sequencing broad swaths of the planet's life.

PROJECT	YEAR STARTED	SEQUENCING GOAL	NUMBER SEQUENCED
G10K	2009	9478 vertebrate genera	100
i5K	2011	5000 arthropods	30
GIGA	2013	7000 marine invertebrates	60
GAGA	2016	All 300 ant genera	25
B10K	2016	All 10,500 bird species	300
AOCC	2013	101 African food crops	

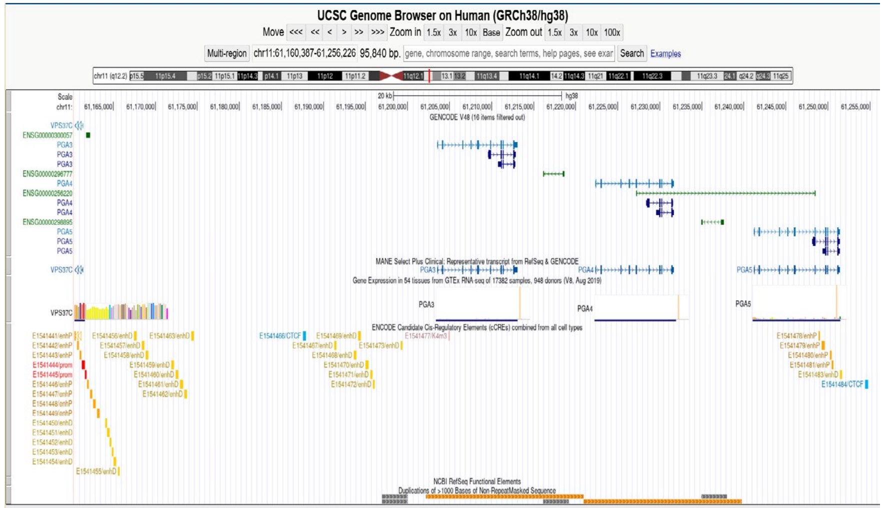
Pennisi, *Science*, 2017

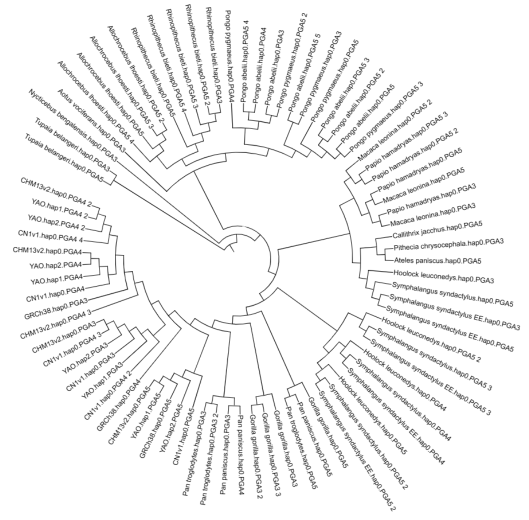
Multi-omics



Chen, et al., *Front. Immunol.*, 2022

Complex regions evolutionary history





Thanks and Q.A.