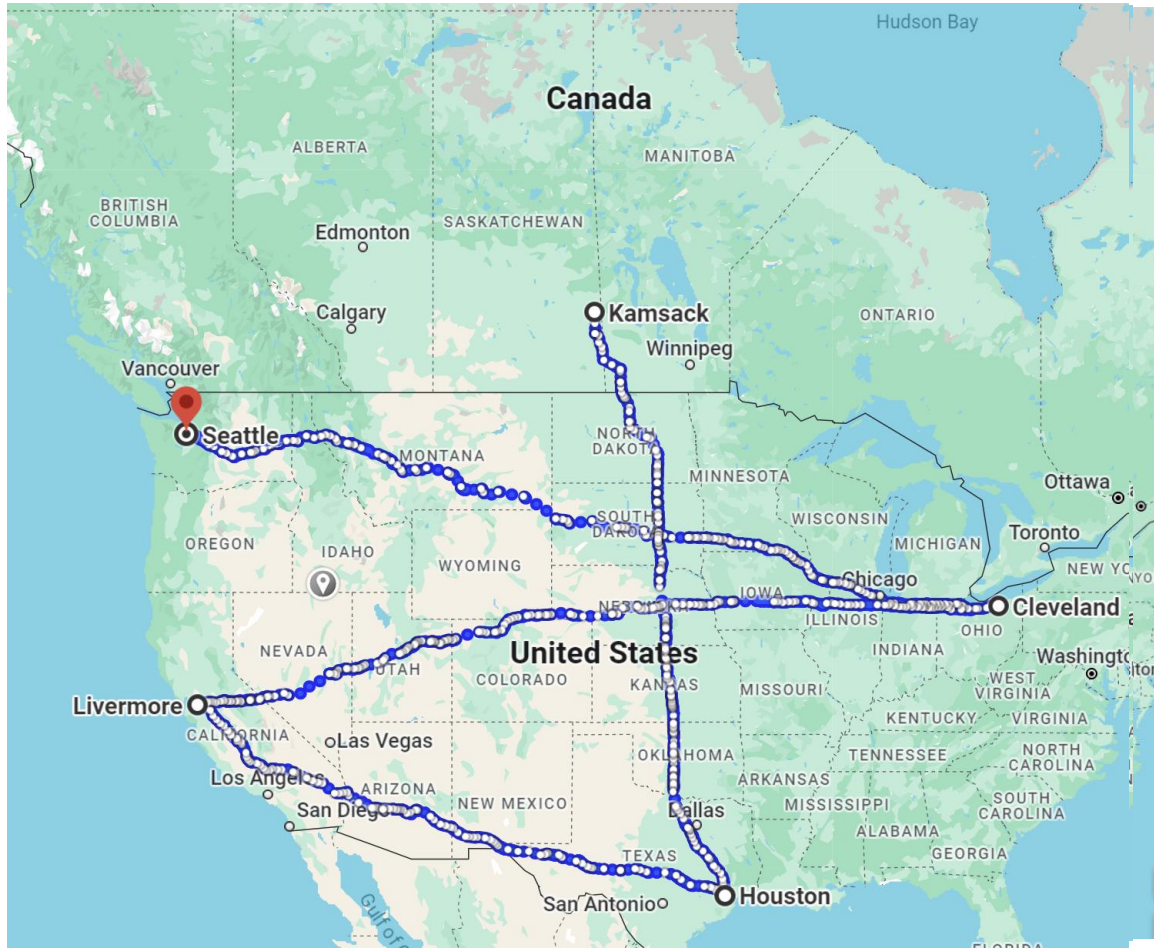
Genome Structural Variation

Evan Eichler
Howard Hughes Medical Institute
University of Washington

January 13th, 2025, Genomics Workshop, Český Krumlov

Who am I?

- Canadian and American
- 1991-1995--Ph.D. Baylor College of Medicine with David Nelson : triplet repeat instability and Fragile X
- 1995-1997- Postdoc –LLNL Human Genome Project
- 1997-2004 Assistant & Associate Professor Case Western Reserve Univ
- 2004-present Professor and HHMI investigator at University of Washington, Seattle
- Recently duplicated genes and dynamic regions of structural variation their role in human disease and evolution



Genetic Variation

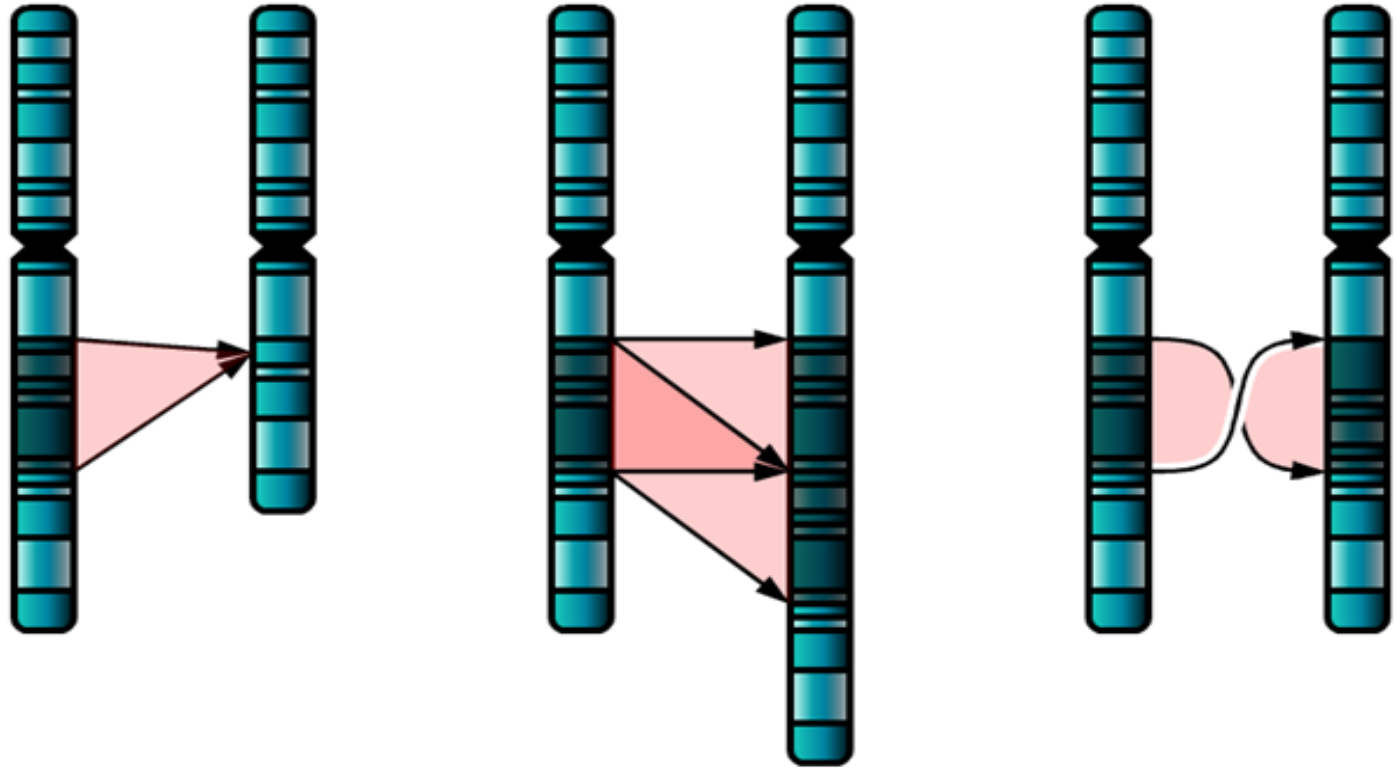
Types

Sequence

- Single base-pair changes – point mutations
- Small insertions/deletions– frameshift, microsatellite, minisatellite
- Mobile elements—retroelement insertions (300bp -10 kb in size)
- Large-scale genomic variation (>1 kb)
  - Large-scale Deletions, Inversion, translocations
  - Segmental Duplications
- Chromosomal variation—translocations, inversions, fusions.

Cytogenetics

Genome Structural Variation



Deletion **Duplication** **Inversion**

Introduction

- **Genome structural variation :** gains and losses of DNA (copy-number variation (CNV)) as well as balanced events such as inversions and translocations—operationally defined >=50 bp
- **Objectives**
  1. Genomic architecture and disease impact.
  2. Detection and characterization methods
  3. Primate genome evolution

Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans

Timothy J. Aitman[1], Rong Dong[1]*, Timothy J. Vyse[2]*, Penny J. Norsworthy[1]*, Michelle D. Johnson[1], Jennifer Smith[3], Jonathan Mangion[1], Cheri Roberton-Lowe[1,2], Amy J. Marshall[1], Enrico Petretto[1], Matthew D. Hodges[1], Gurjeet Bhangal[3], Sheetal G. Patel[1], Kelly Sheehan-Rooney[1], Mark Duda[1,3], Paul R. Cook[1,3], David J. Evans[3], Jan Domin[3], Jonathan Flint[4], Joseph J. Boyle[5], Charles D. Pusey[3] & H. Terence Cook[5]

Nature. 2006

The Influence of *CCL3L1* Gene–Containing Segmental Duplications on HIV-1/AIDS Susceptibility

Enrique Gonzalez,[1*] Hemant Kulkarni,[1*] Hector Bolivar,[1*†] Andrea Mangano,[2*] Racquel Sanchez,[1‡] Gabriel Catano,[1‡] Robert J. Nibbs,[3‡] Barry I. Freedman,[4‡] Marlon P. Quinones,[1‡] Michael J. Bamshad,[5] Krishna K. Murthy,[6] Brad H. Rovin,[7] William Bradley,[8,9] Robert A. Clark,[1] Stephanie A. Anderson,[8,9] Robert J. O'Connell,[9,10] Brian K. Agan,[9,10] Seema S. Ahuja,[1] Rosa Bologna,[11] Luisa Sen,[2] Matthew J. Dolan,[9,10,12]§ Sunil K. Ahuja[1]§

Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome

Andrew J Sharp[1], Sierra Hansen[1], Rebecca R Selzer[2], Ze Cheng[1], Regina Regan[3], Jane A Hurst[4], Helen Stewart[4], Sue M Price[4], Edward Blair[4], Raoul C Hennekam[5,6], Carrie A Fitzpatrick[7], Rick Segraves[8], Todd A Richmond[2], Cheryl Guiver[3], Donna G Albertson[8,9], Daniel Pinkel[8], Peggy S Eis[2], Stuart Schwartz[7], Samantha J L Knight[3] & Evan E Eichler[1] VOLUME 38 | NUMBER 9 | SEPTEMBER 2006 NATURE GENETICS

Schizophrenia risk from complex variation of complement component 4

Aswin Sekar, Allison R. Bialas, Heather de Rivera, Avery Davis, Timothy R. Hammond, Nolan Kamitaki, Katherine Tooley, Jessy Presumey, Matthew Baum, Vanessa Van Doren, Giulio Genovese, Samuel A. Rose, Robert E. Handsaker, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Mark J. Daly, Michael C. Carroll, Beth Stevens & Steven A. McCarroll ✉

Nature **530**, 177–183(2016) | Cite this article

Association between Microdeletion and Microduplication at 16p11.2 and Autism

Lauren A. Weiss, Ph.D., Yiping Shen, Ph.D., Joshua M. Korn, B.S., Dan E. Arking, Ph.D., David T. Miller, M.D., Ph.D., Ragnheidur Fossdal, B.Sc., Evald Saemundsen, B.A., Hreinn Stefansson, Ph.D., Manuel A.R. Ferreira, Ph.D., Todd Green, B.S., Orah S. Platt, M.D., Douglas M. Ruderfer, M.S., Christopher A. Walsh, M.D., Ph.D., David Altshuler, M.D., Ph.D., Aravinda Chakravarti, Ph.D., Rudolph E. Tanzi, Ph.D., Kari Stefansson, M.D., Ph.D., Susan L. Santangelo, Sc.D., James F. Gusella, Ph.D., Pamela Sklar, M.D., Ph.D., Bai-Lin Wu, M.Med., Ph.D., and Mark J. Daly, Ph.D., for the Autism Consort N Engl J Med 2008;358:667-75

Strong Association of De Novo Copy Number Mutations with Autism

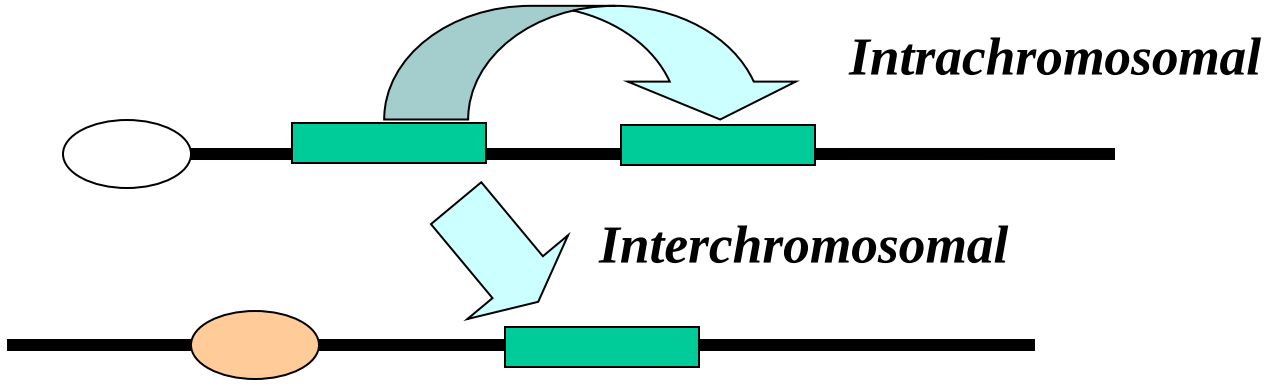
Jonathan Sebat,[1*] B. Lakshmi,[1] Dheeraj Malhotra,[1*] Jennifer Troge,[1*] Christa Lese-Martin,[2] Tom Walsh,[3] Boris Yamrom,[1] Seungtai Yoon,[1] Alex Krasnitz,[1] Jude Kendall,[1] Anthony Leotta,[1] Deepa Pai,[1] Ray Zhang,[1] Yoon-Ha Lee,[1] James Hicks,[1] Sarah J. Spence,[4] Annette T. Lee,[5] Kaija Puura,[6] Terho Lehtimäki,[7] David Ledbetter,[2] Peter K. Gregersen,[5] Joel Bregman,[8] James S. Sutcliffe,[9] Vaidehi Jobanputra,[10] Wendy Chung,[10] Dorothy Warburton,[10] Mary-Claire King,[3] David Skuse,[11] Daniel H. Geschwind,[12] T. Conrad Gilliam,[13] Kenny Ye,[14] Michael Wigler[1]†

NCE

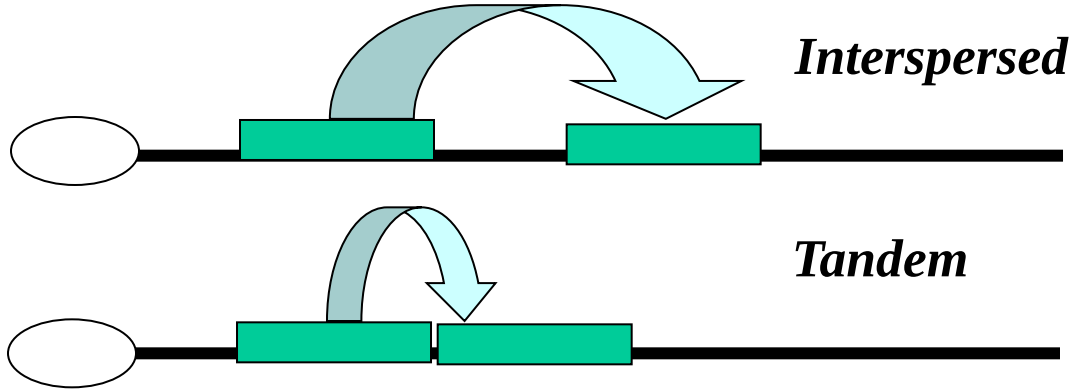
SCIENCE VOL 316 20 APRIL 2007

Perspective: Segmental Duplications (SD)

Definition: Continuous portion of genomic sequence represented more than once in the genome (>90% and > 1kb in length)=historical copy number variation

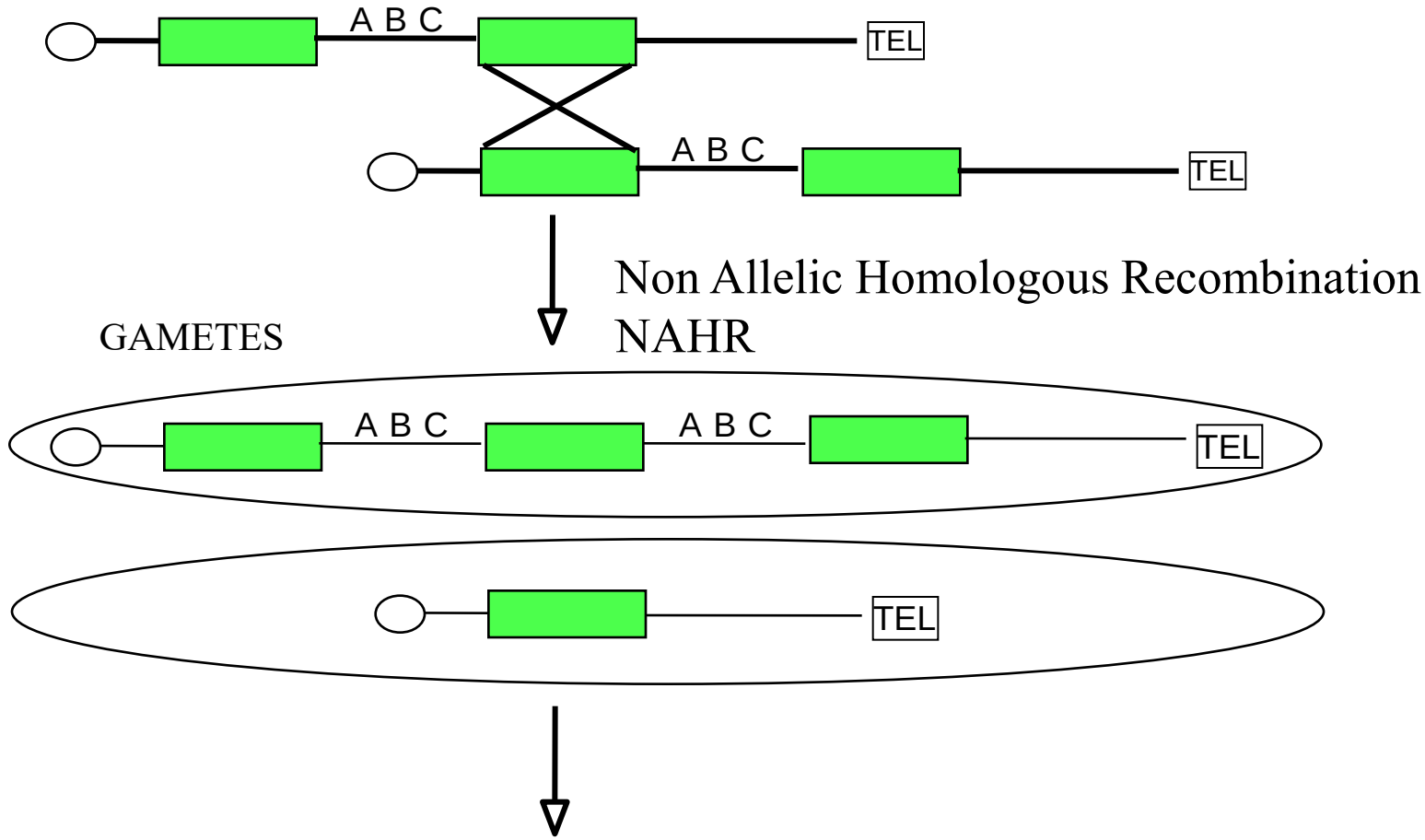


Distribution



Configuration

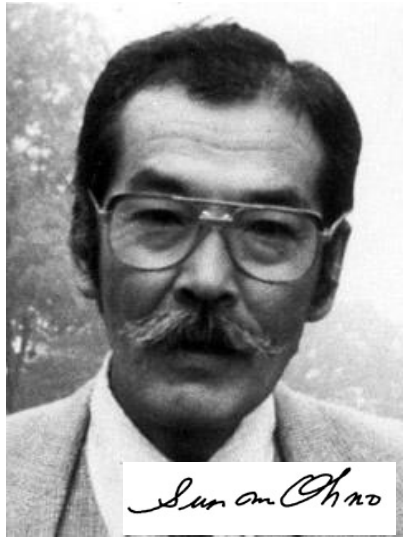
Importance:
SDs promote genome structural variation



Human Disease

Triplosensitive, Haploinsufficient and Imprinted Genes

Importance: Evolution of New Gene Function



GeneA

Duplication

GeneA'

Acquire New/
Modified Function

Mutation

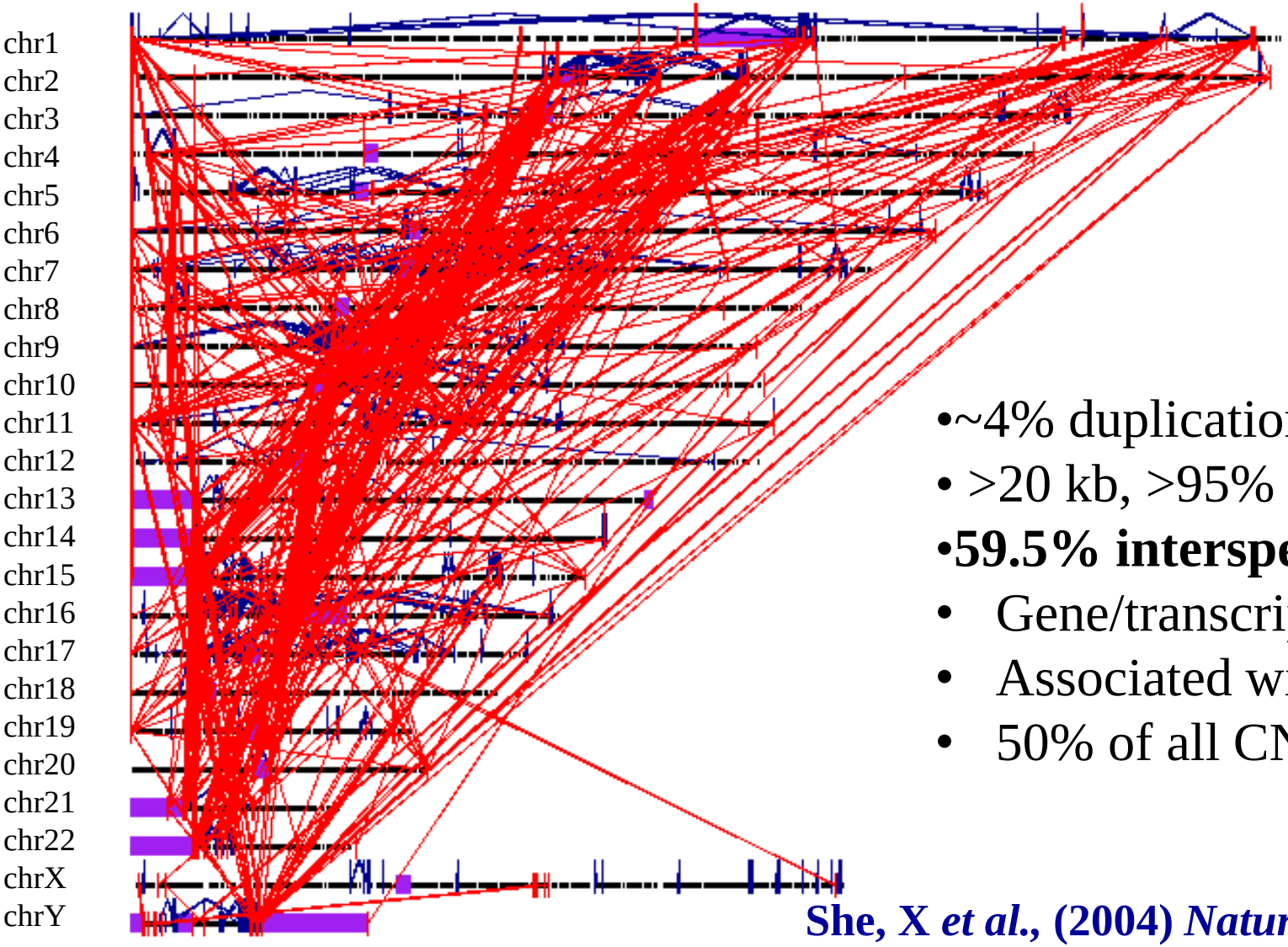
Mutation

Maintain old
Function

Mutation

Loss of Function

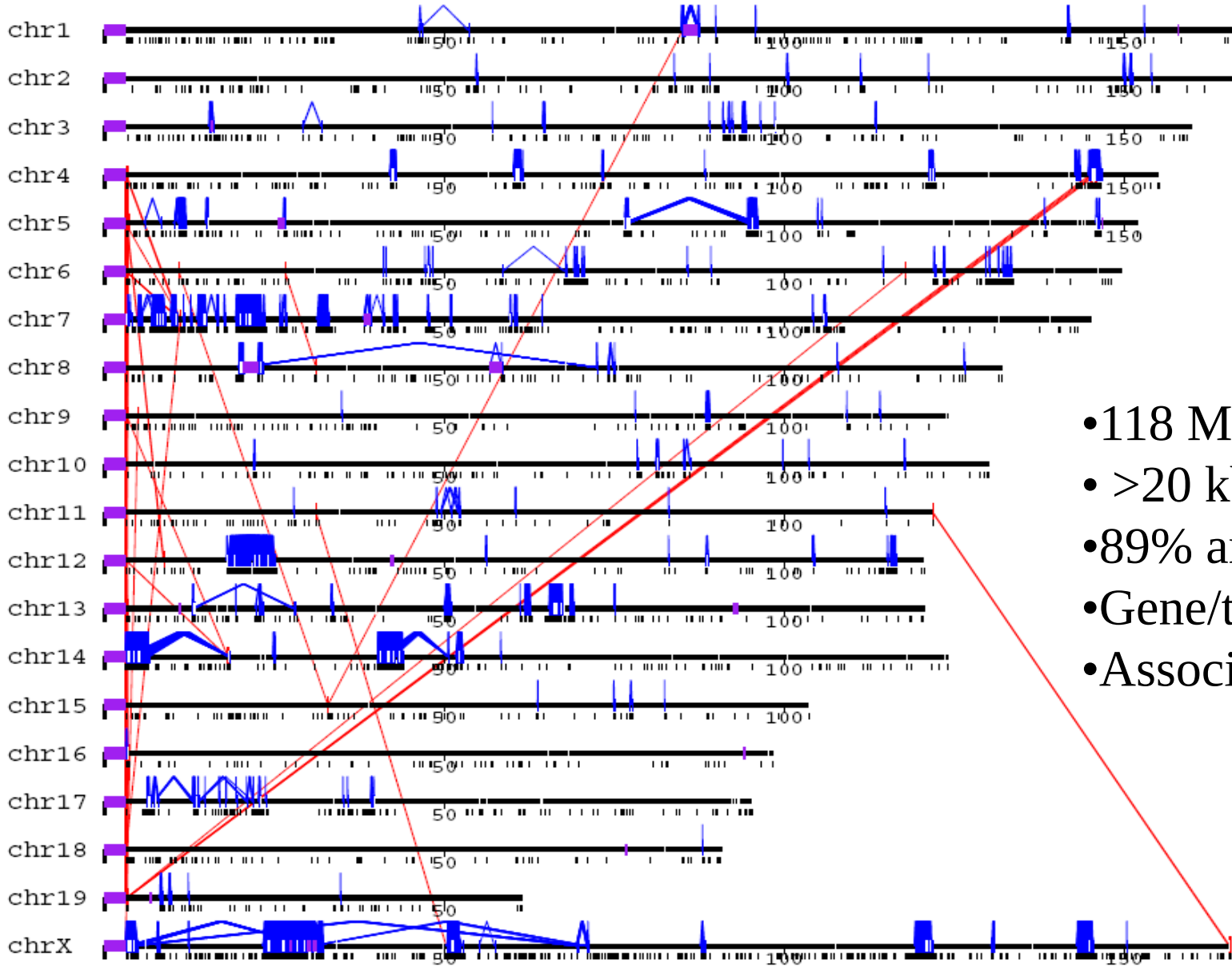
Human Genome Segmental Duplication Pattern



- ~4% duplication (125 Mb)
- >20 kb, >95%
- **59.5% interspersed**
- Gene/transcript rich
- Associated with Alu repeat
- 50% of all CNV > 1 kbp

She, X *et al.*, (2004) *Nature* 431:927-30

Mouse Segmental Duplication Pattern



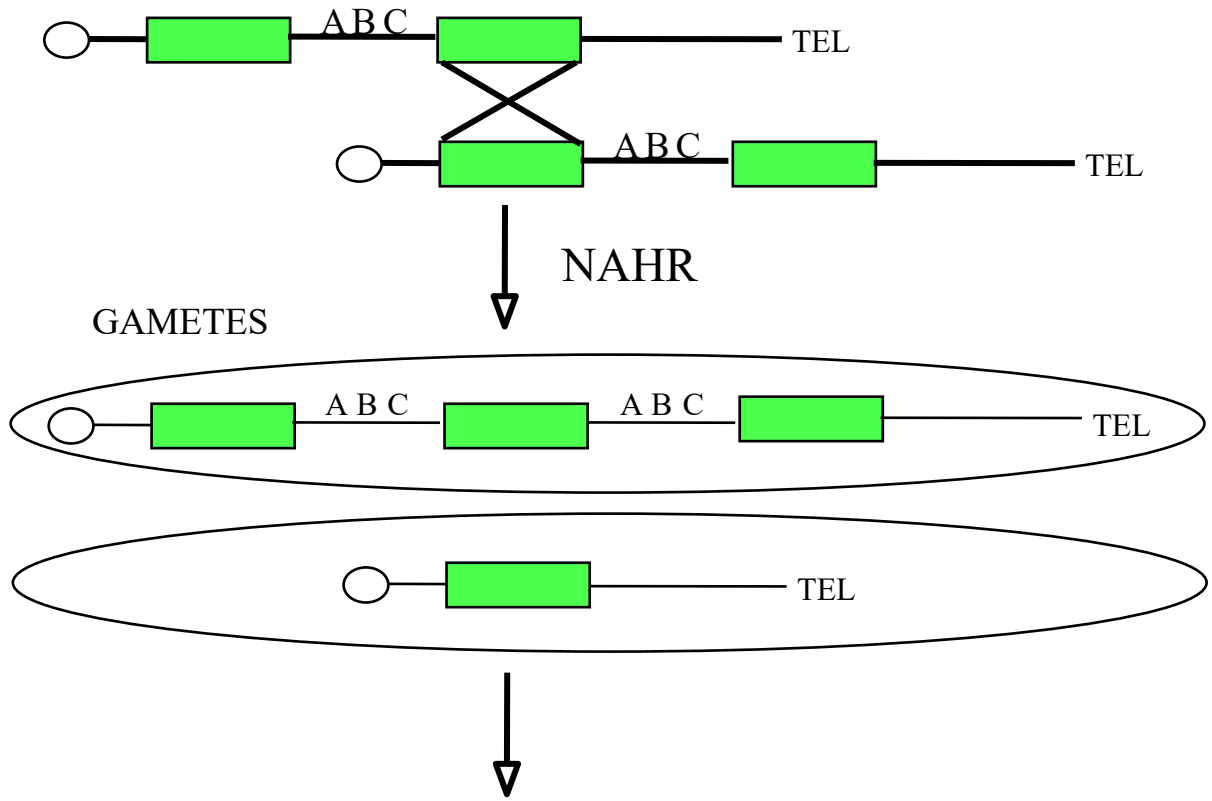
- 118 Mb or ~4% dup
- >20 kb, >95%
- 89% are tandem
- Gene/transcript poor
- Associated with LINEs

She, X *et al.*, (2008) *Nature Genetics*

Human Segmental Duplications Properties

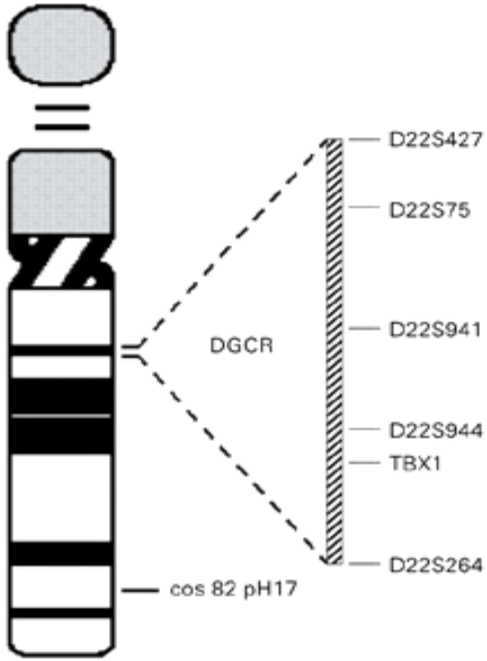
- Large (>10 kb)
- Recent (>95% identity)
- **Interspersed (60% are separated by more than 1 Mb)**
- Modular in organization
- Difficult to resolve

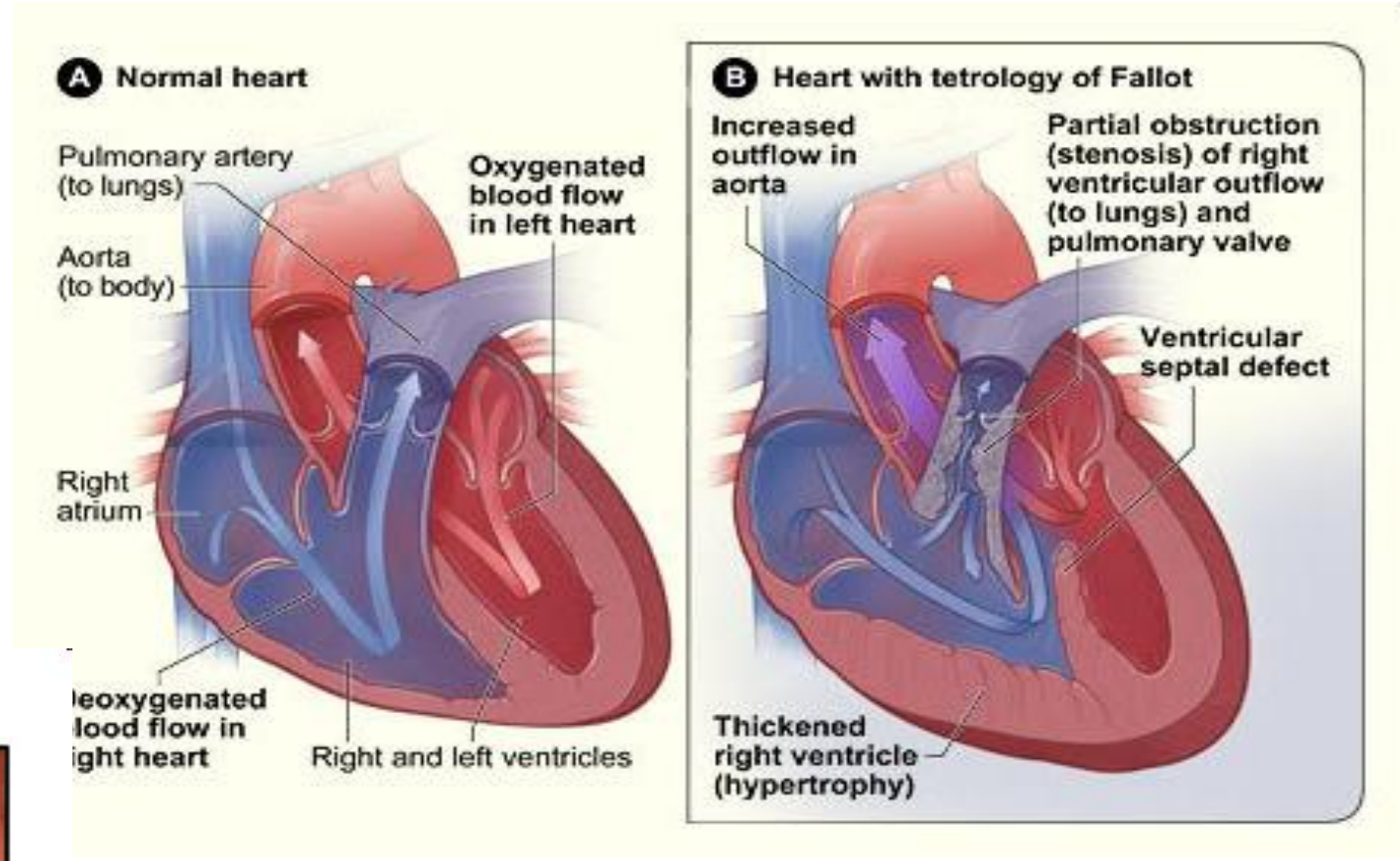
Rare Structural Variation & Disease



•**Genomic Disorders:** A group of diseases that results from genome rearrangement mediated mostly by non-allelic homologous recombination**. (***Inoue & Lupski , 2002***).**

DiGeorge/VCFS/22q11 Syndrome

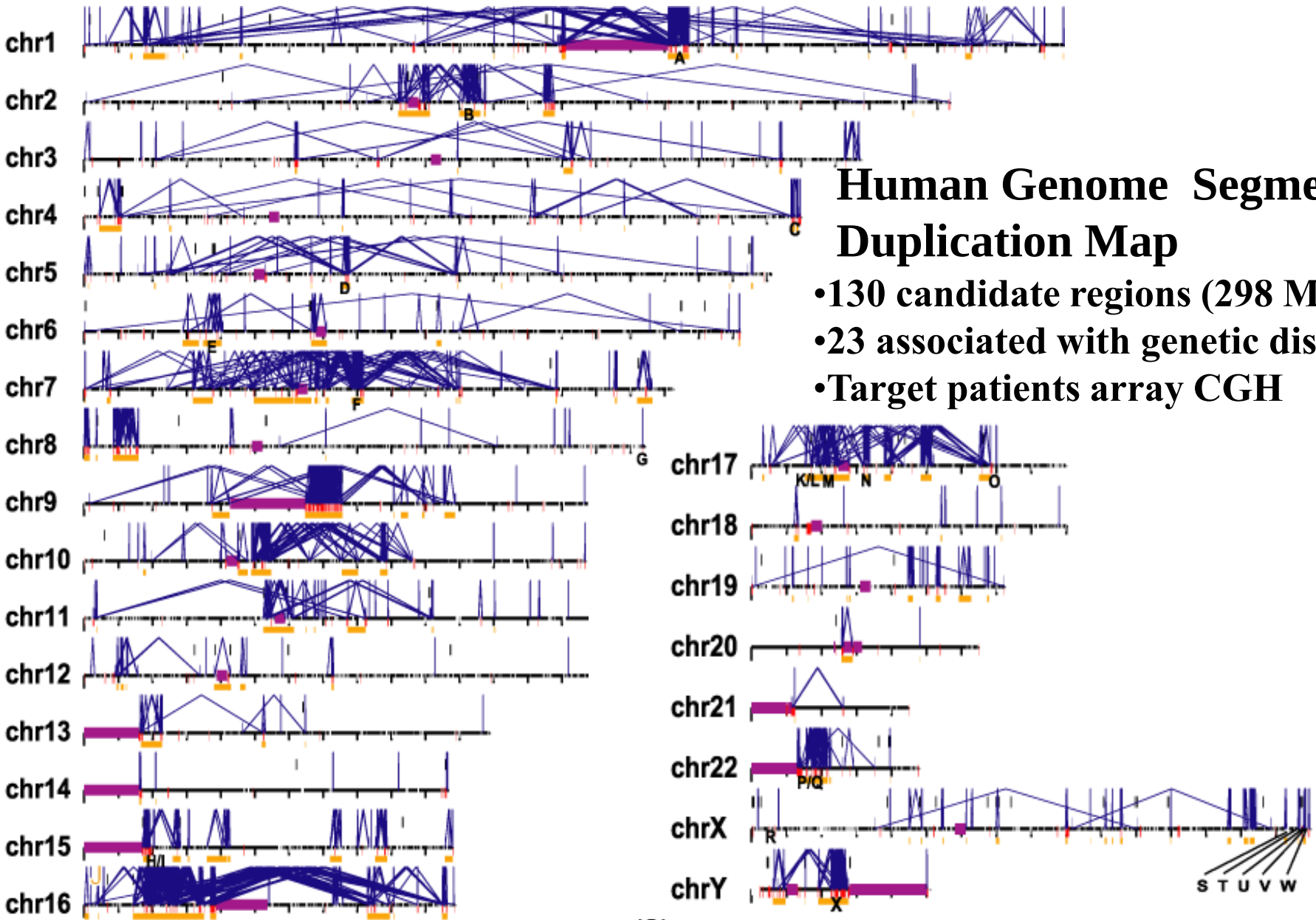




1/2000 live births

180 phenotypes

75-80% are sporadic (not inherited)

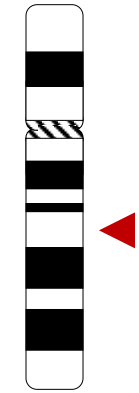


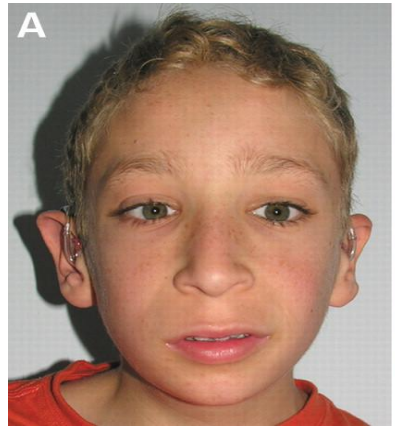
Human Genome Segmental Duplication Map

- 130 candidate regions (298 Mb)
- 23 associated with genetic disease
- Target patients array CGH

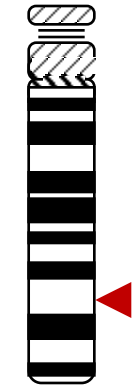
Bailey *et al.* (2002), *Science*

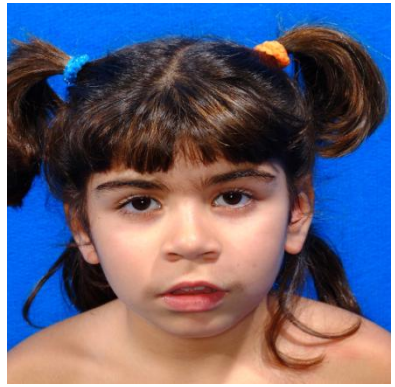
Chromosome 17





Chromosome 15

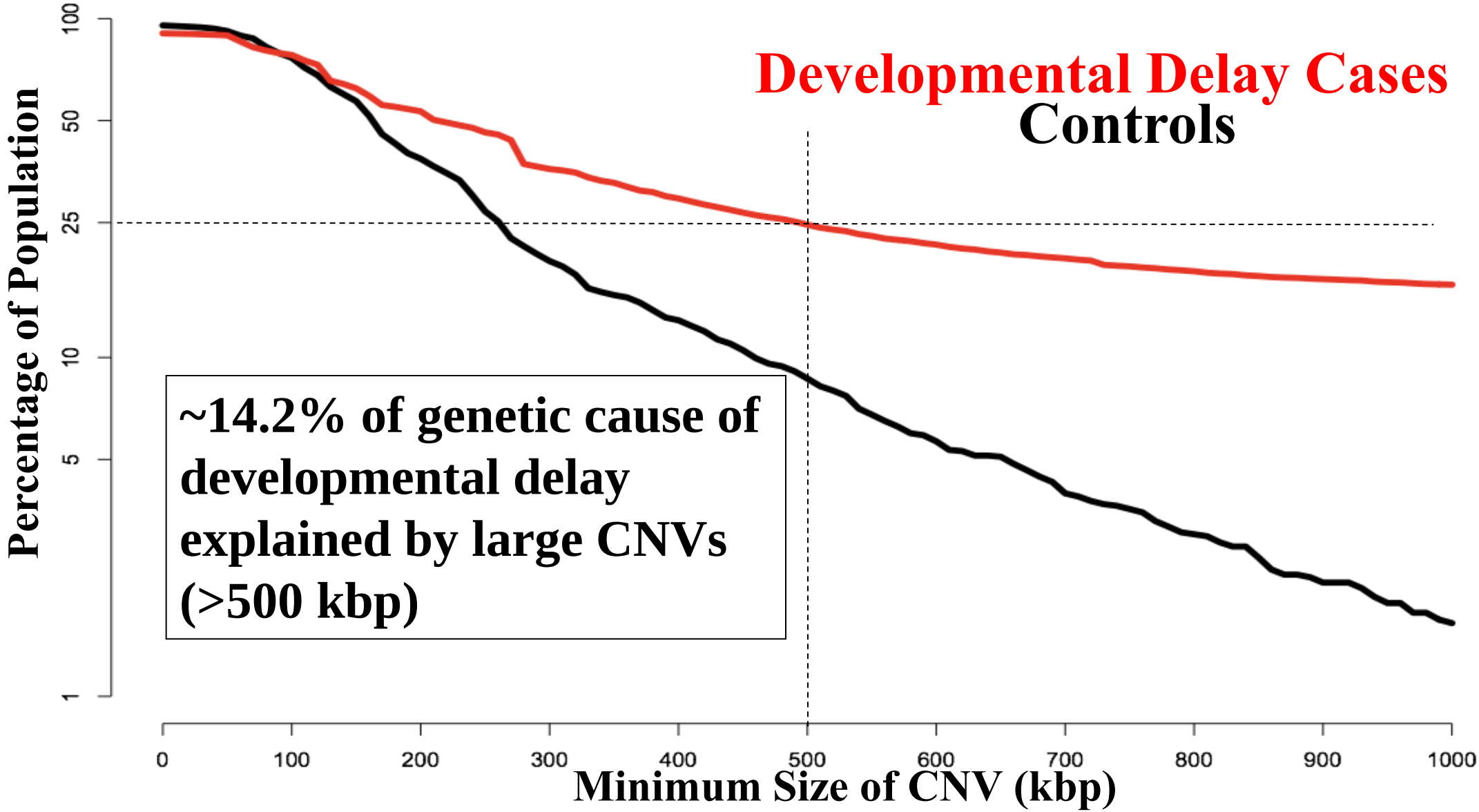




Chromosome 15

Genome Wide CNV Burden

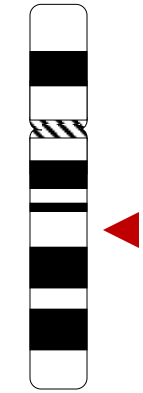
(15,767 cases of ID,DD,MCA vs. 8,328 controls)

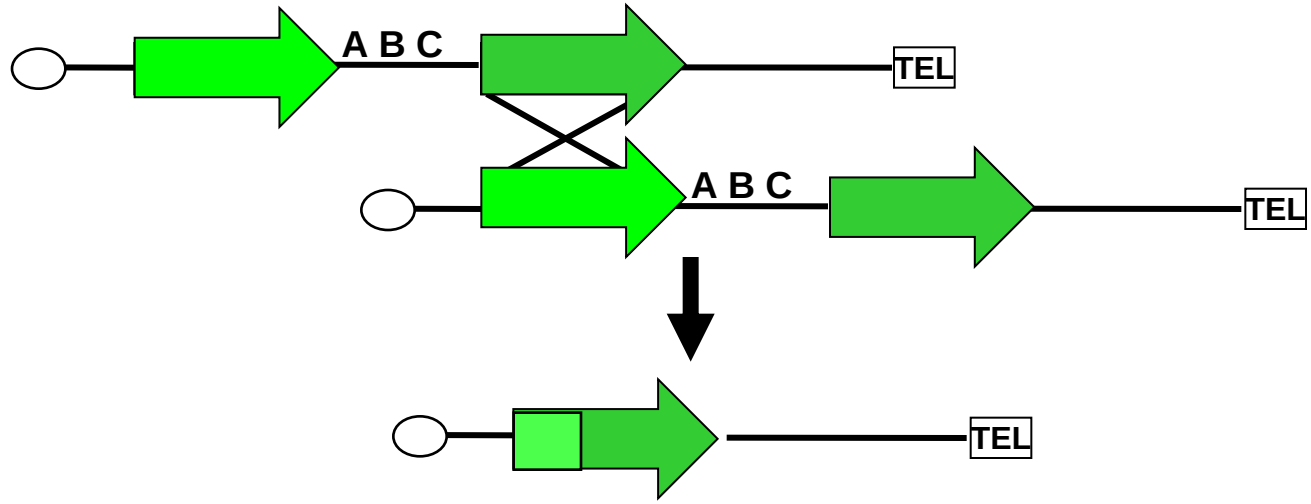


Cooper et al., *Nat. Genet*, 2011

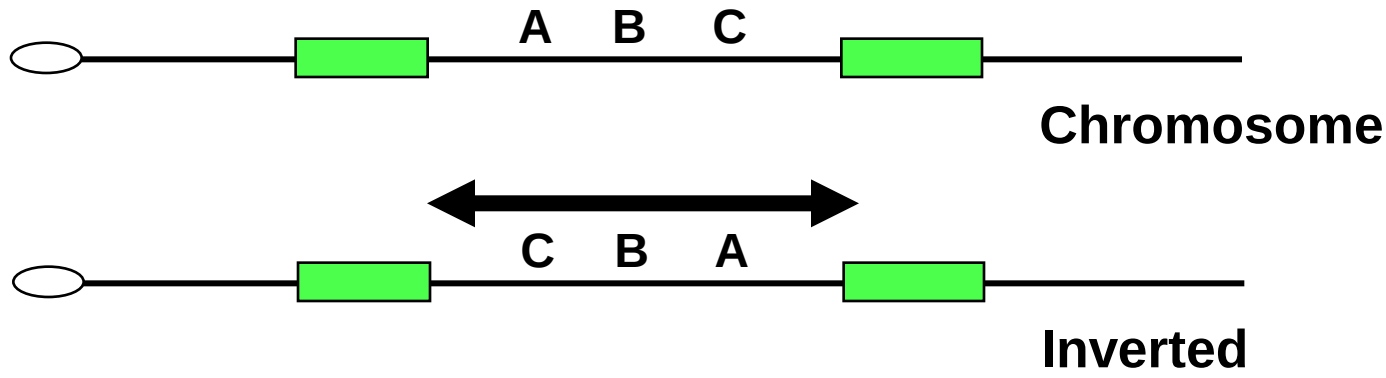
Common and rare structural variation are linked 17q21.31 deletion syndrome

Chromosome 17





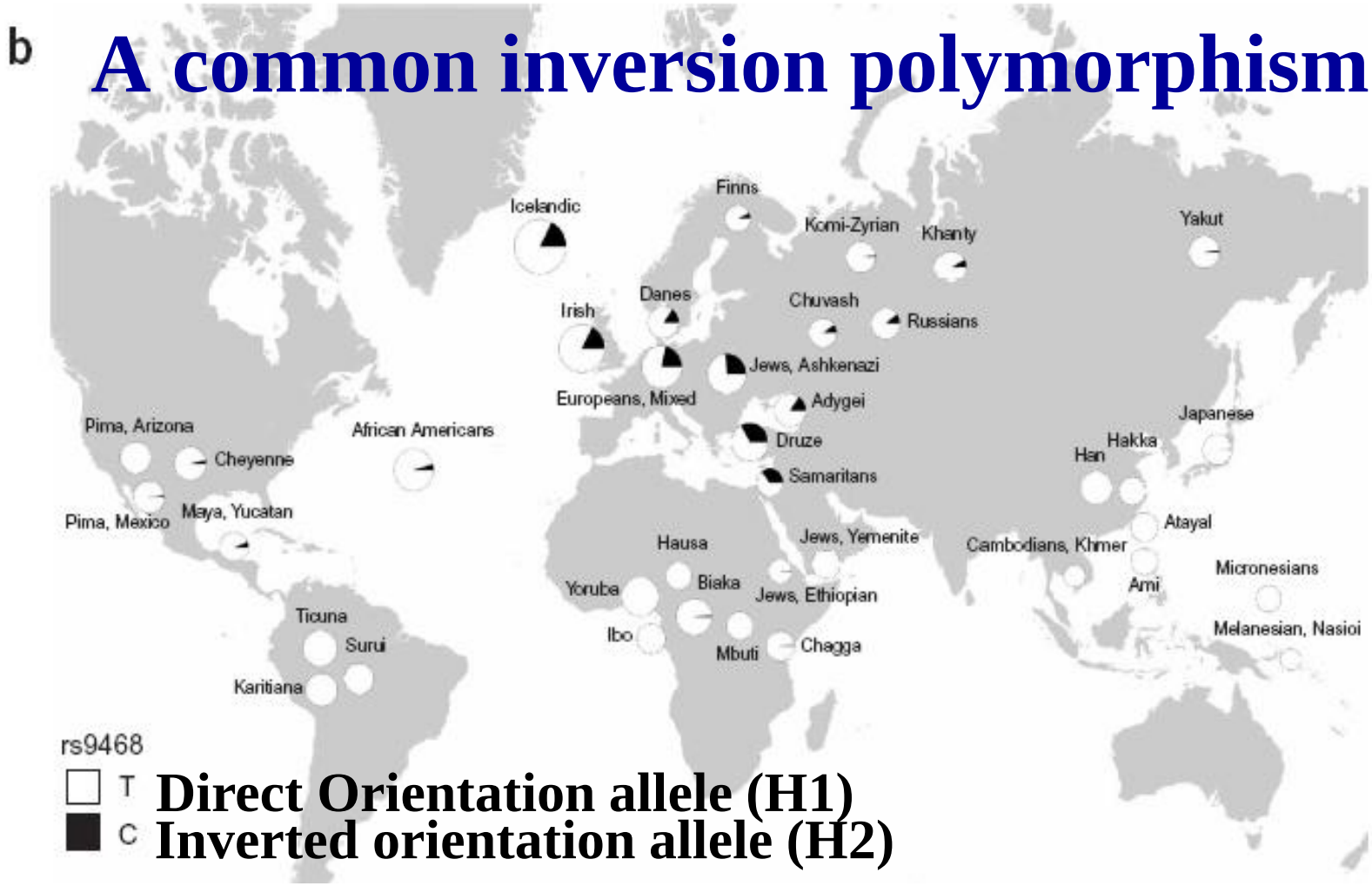
17q21.31 inversion



- Region of recurrent deletion is a site of common inversion polymorphism in the human population
- Inversion is largely restricted to Caucasian populations
  - 20% frequency in European and Mediterranean populations
- **Inversion is associated with increase in global recombination and increased fecundity**

Stefansson, K *et al.*, (2005) *Nature Genetics*

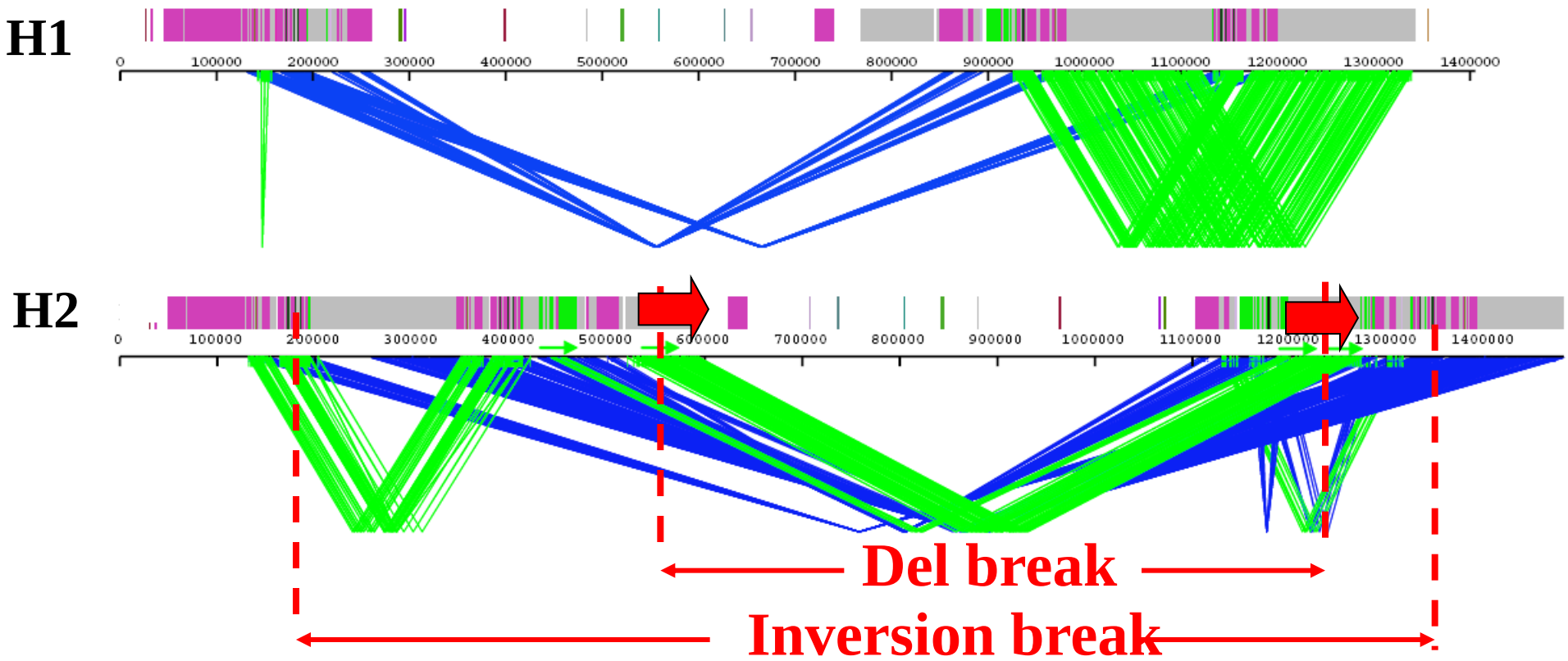
A common inversion polymorphism



Direct Orientation allele (H1)
Inverted orientation allele (H2)

- **Tested 17 parents of children with microdeletion and found that every parent within whose germline the deletion occurred carried an inversion**
- **Inversion polymorphism is a risk factor for the microdeletion event**

Duplication Architecture of 17q21.31 Inversion (H2) vs. Direct (H1) Haplotype



- Inversion occurred 2.3 million years ago and was mediated by the LRRC37A core duplicon
- H2 haplotype acquired human-specific duplications in direct orientation that mediate rearrangement and disrupts *KANSL1* gene

Zody *et al.*, *Nat. Genet.* 2008, Itsara et al., Am J. Human Genet 2012

Summary

- Human genome is enriched for segmental duplications which predisposes to recurrent large CNVs during germ-cell production
- 15% of neurodevelopmental disease in intellectual disabled children is “caused” by large CNVs—8% of normals carry large events
- Segmental duplications enriched >10 fold for structural variation.
- Increased complexity is beneficial and deleterious: Ancestral duplication predisposes to inversion polymorphism, inversion polymorphisms acquires duplication, haplotype becomes positively selected and now predisposes to microdeletion

II. Genome-wide SV Discovery Approaches

Hybridization-based

- Iafrate et al., 2004, Sebat et al., 2004
- SNP microarrays: McCarroll *et al*., 2008, Cooper *et al*., 2008, Itsara *et al*., 2009
- Array CGH: Redon *et al.* 2006, Conrad *et al*., 2010, Park *et al*., 2010, WTCCC, 2010

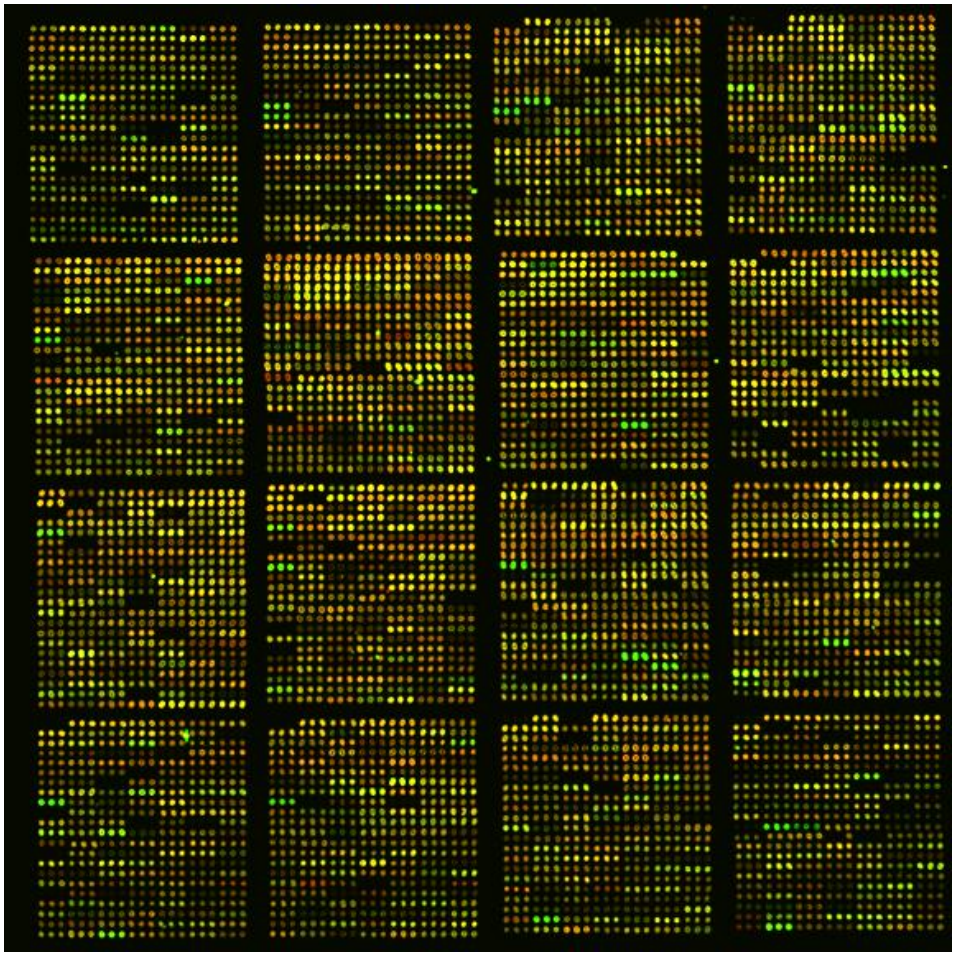
Single molecule mapping

- Optical mapping**:** Teague et al., 2010 e.g. Bionano Genomics: Levy-Sakin et al, 2019

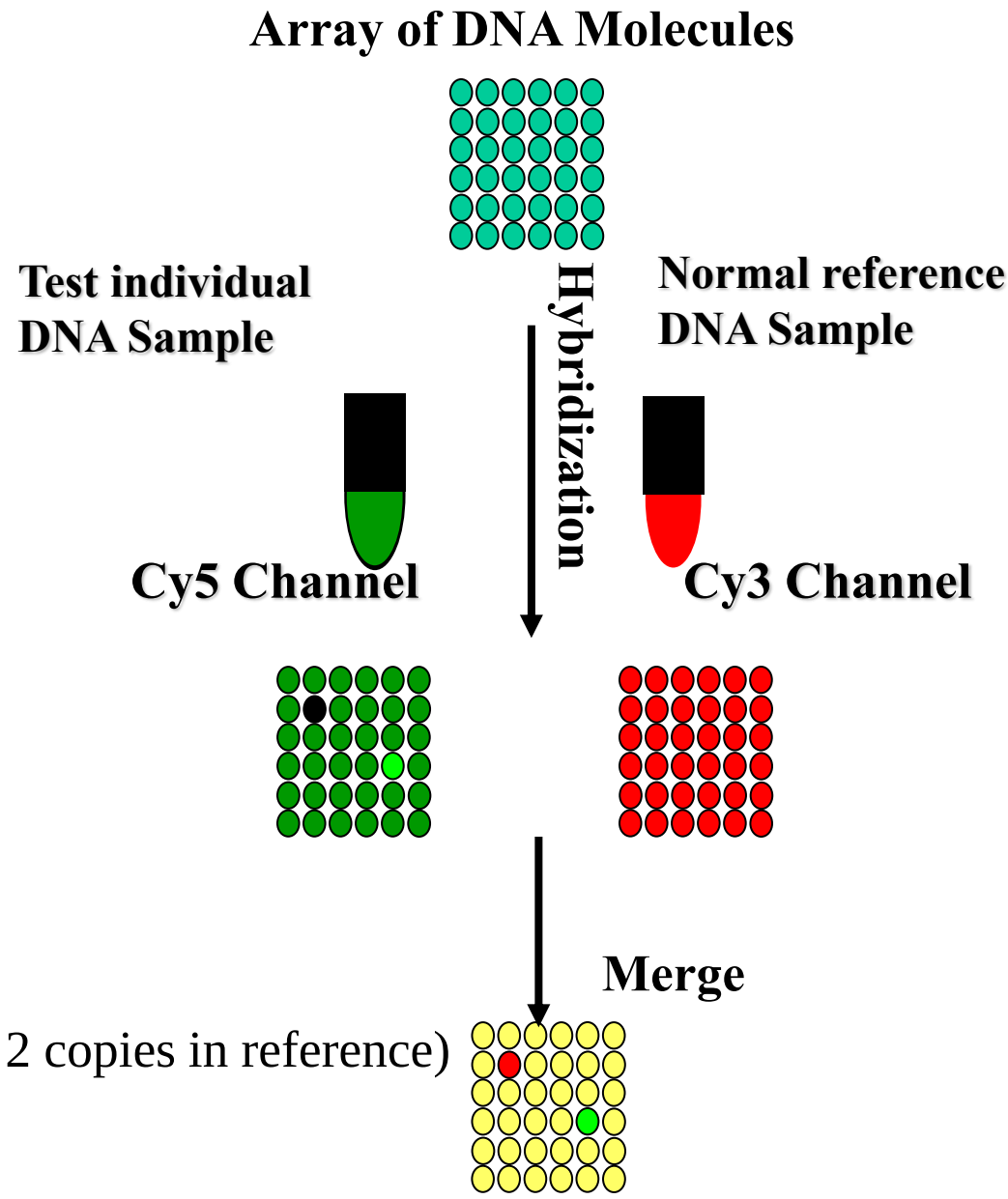
Sequencing-based

- Read-depth: Bailey et al, 2002
- Fosmid ESP: Tuzun *et al.* 2005, Kidd *et al.* 2008
- Next-gen sequencing: Korbel *et al.* 2007, Yoon *et al*., 2009, Alkan et al., 2009, Chen *et al.* 2009; Mills 1000 Genomes Project, 2011, Sudmant *et al.* 2015a,
- Long-read sequencing and assembly: Chaisson *et al*., 2015, 2019, Pendleton *et al*., 2015, Sedlazeck et al., 2018 Audano *et al*, 2019, Ebert et al,, 2021

Array Comparative Genomic Hybridization



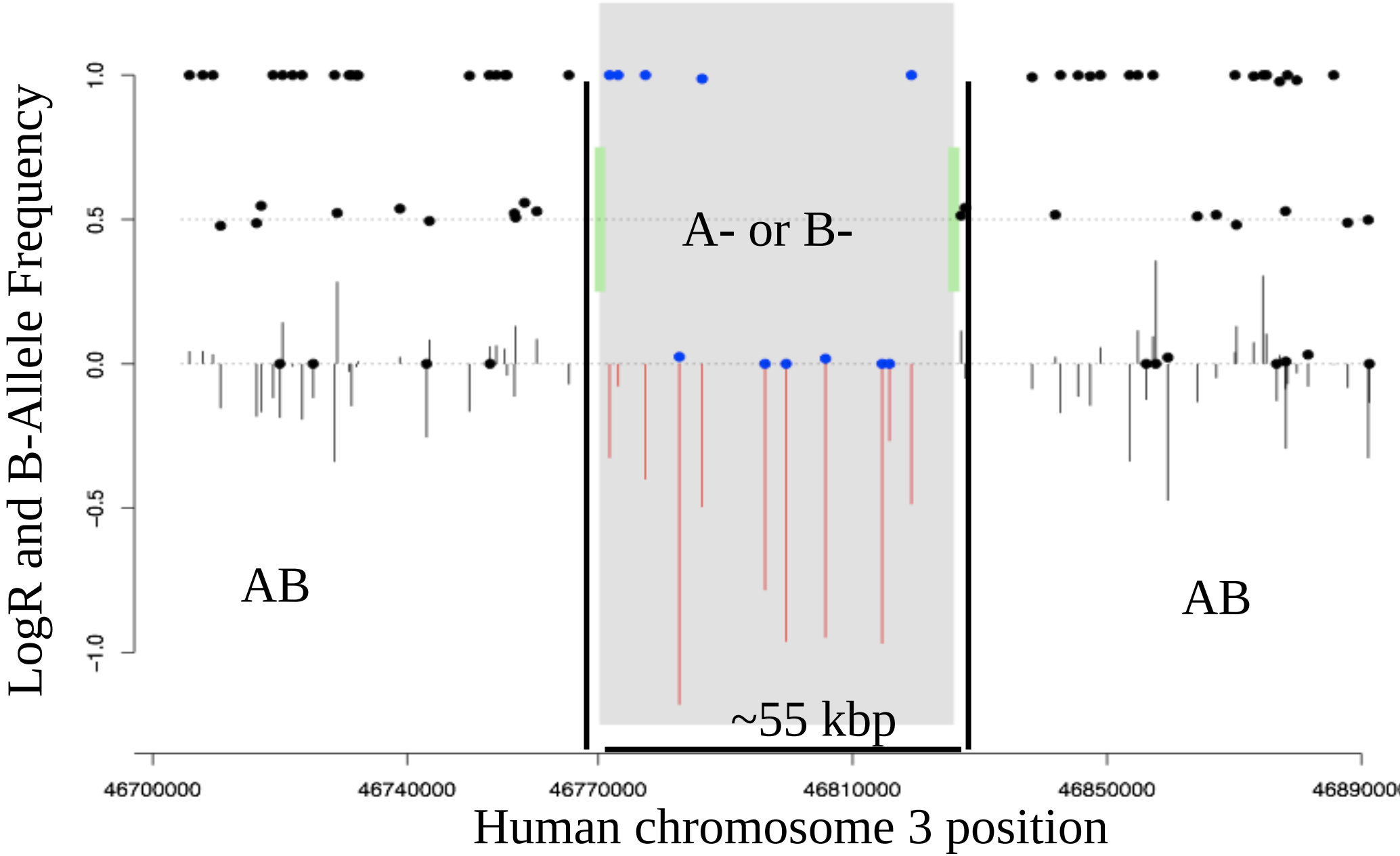
12 mm



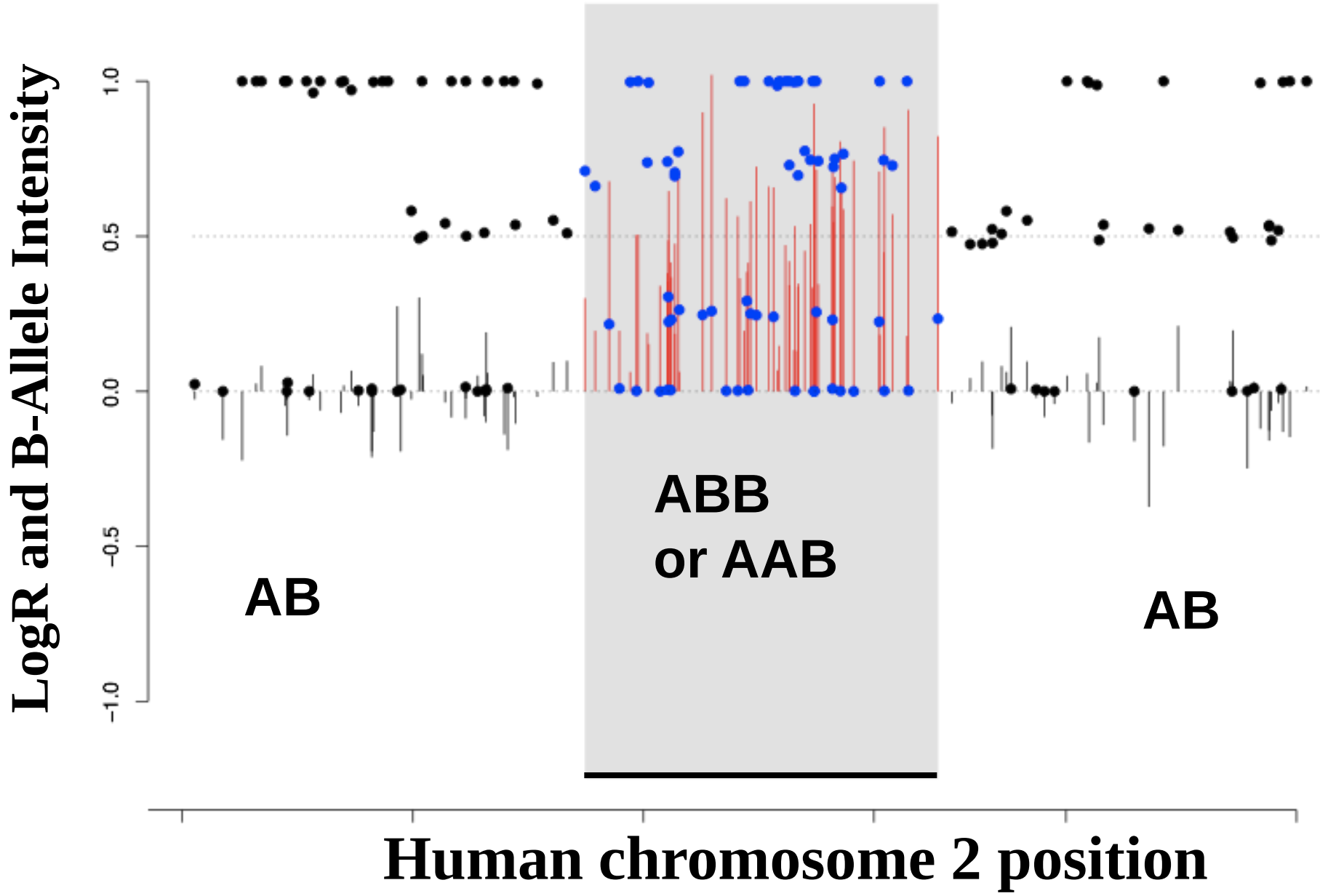
One copy gain = $\log_2(3/2) = 0.57$ (3 copies vs. 2 copies in reference)

One-copy loss = log2(1/2) = -1

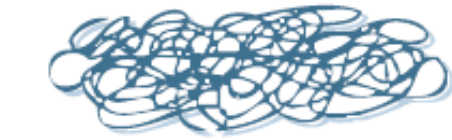
SNP Microarray detection of Deletion (Illumina)



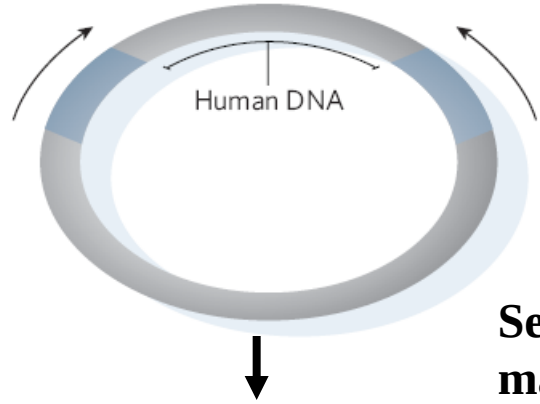
SNP Microarray detection of duplication



Using sequence read pairs to resolve structural variation

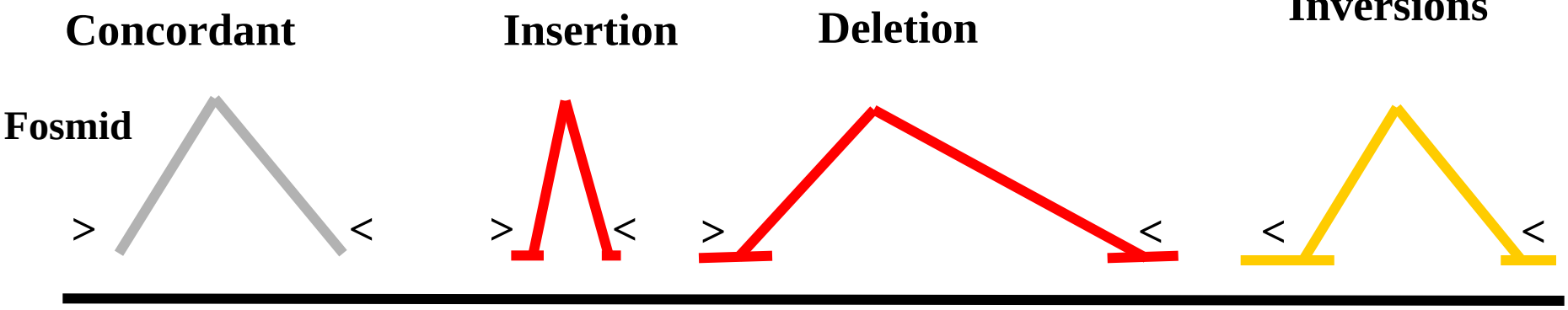


Human Genomic DNA



Genomic Library (1 million clones)

Sequence ends of genomic inserts & map paired-ends to human genome



Dataset: 1,122,408 fosmid pairs preprocessed (15.5X genome coverage)
639,204 fosmid pairs BEST pairs (8.8 X genome coverage)

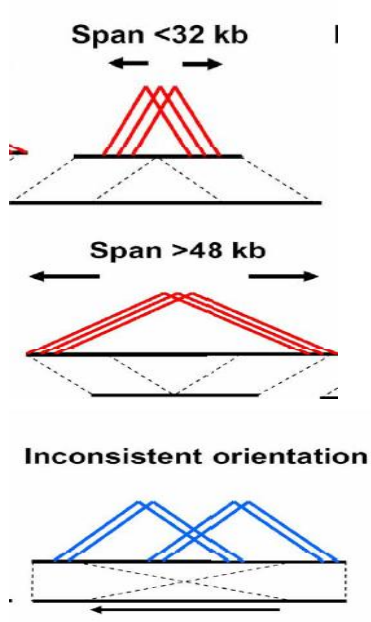
Genome-wide detection of structural variation (>8kb) by end-sequence pairs or "mate pairs"

a)

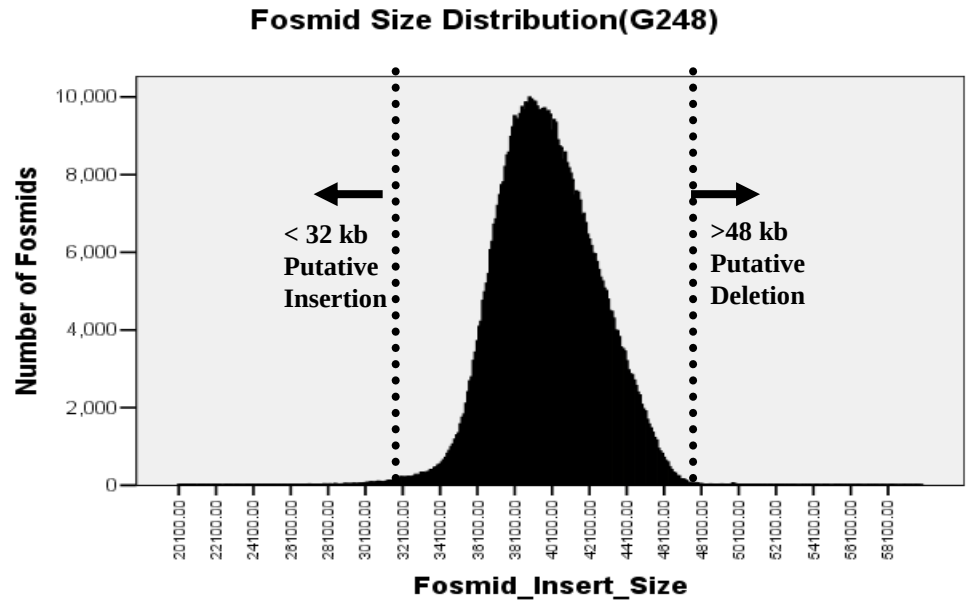
Insertion

Deletion

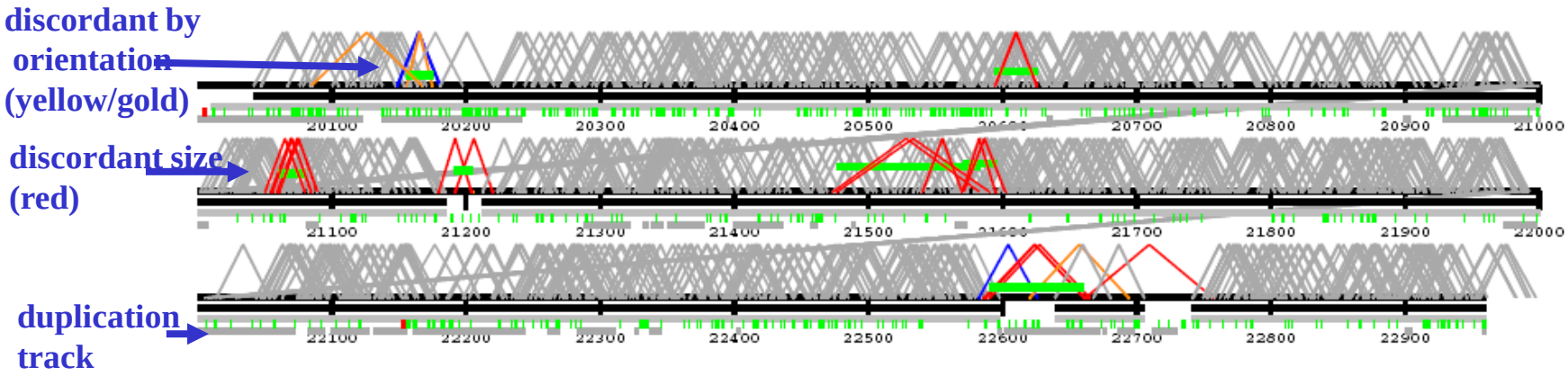
Inversion



b)



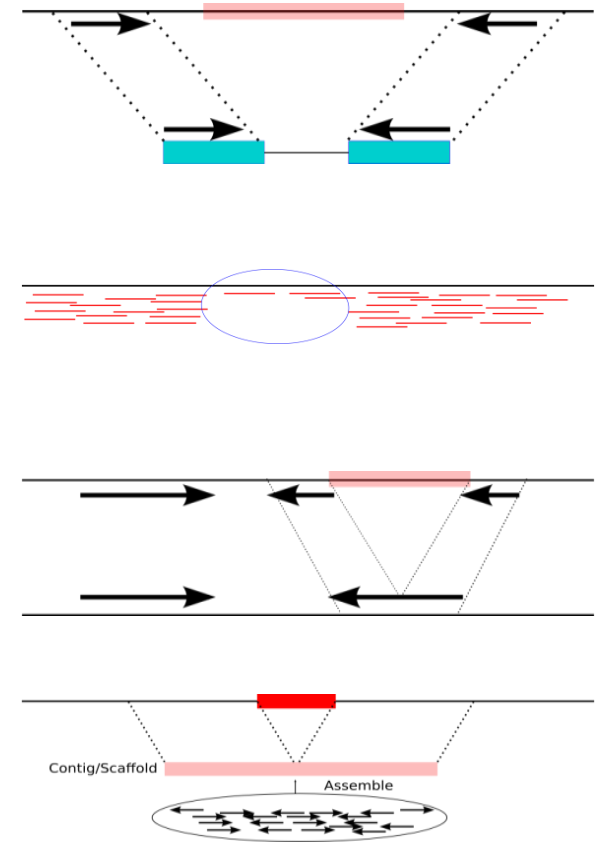
c)



Tuzun et al, *Nat. Genetics*, 2005; Kidd et al., *Nature*, 2008

Next-Generation Sequencing Methods

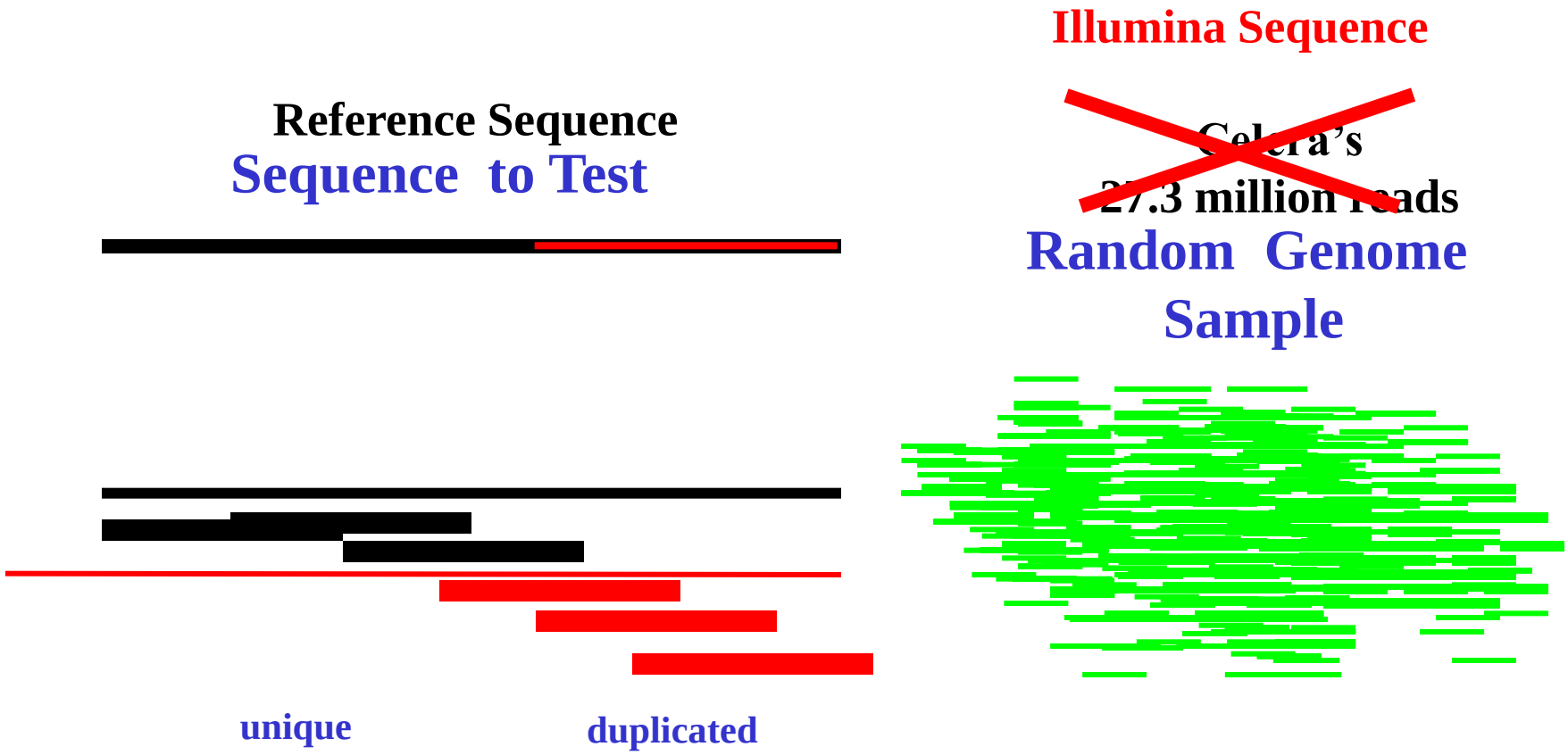
- **Read pair analysis**
  - Deletions, small novel insertions, inversions, transposons
  - Size and breakpoint resolution dependent to insert size
- **Read depth analysis**
  - Deletions and duplications (CNV) only
  - Relatively poor breakpoint resolution
- **Split read analysis**
  - Small novel insertions/deletions, and mobile element insertions
  - 1bp breakpoint resolution
- **Local and *de novo* assembly**
  - SV in unique segments
  - 1bp breakpoint resolution



Alkan et al., *Nat Rev Genet, 2011*

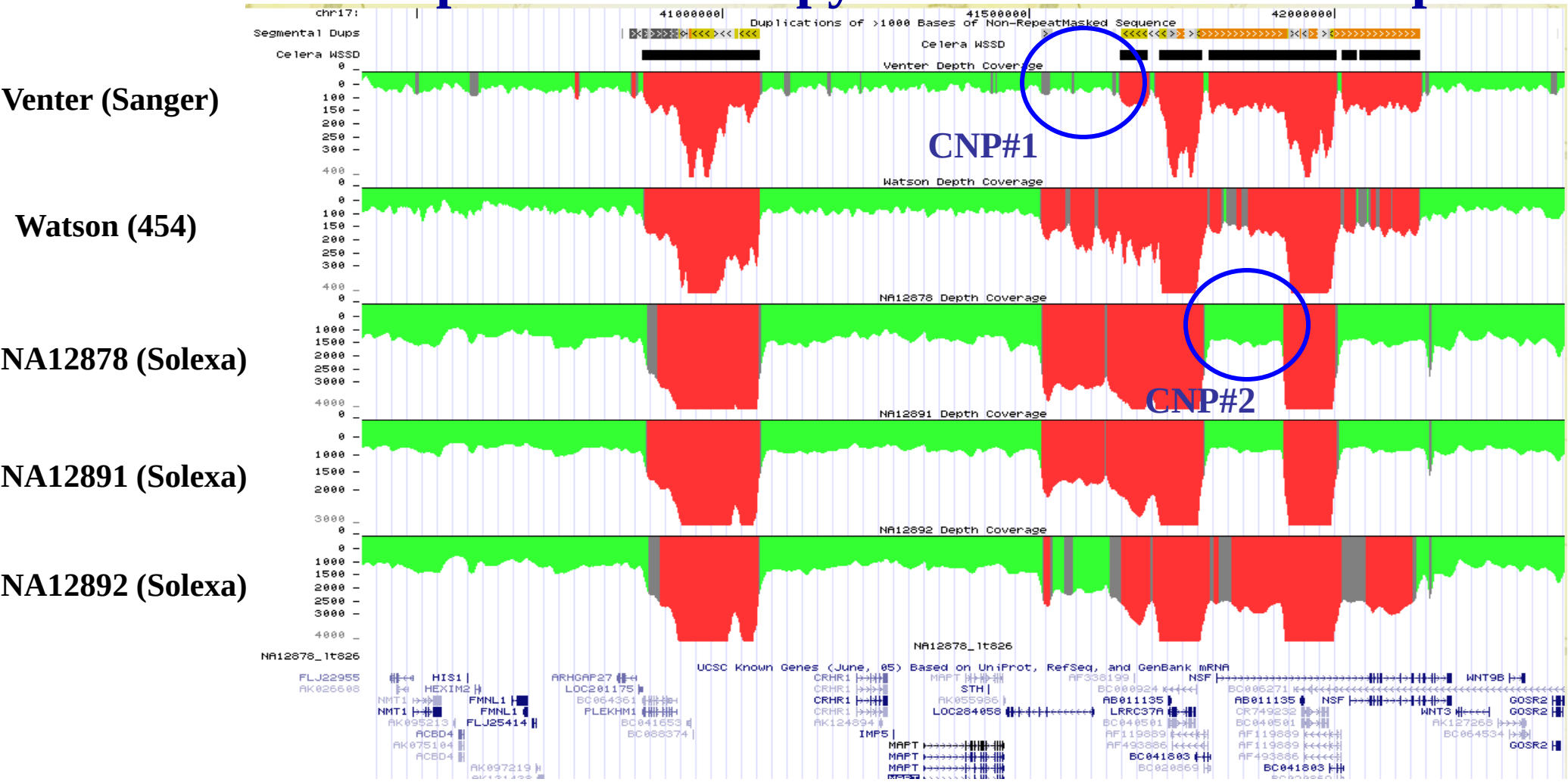
Using Sequence Read Depth

- Map whole genome sequence to reference genome
  - Variation in copy number correlates linearly with read-depth



Bailey et al., Science, 2002

Personalized Duplication or Copy-Number Variation Maps

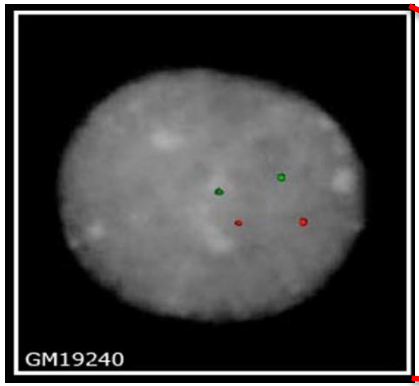


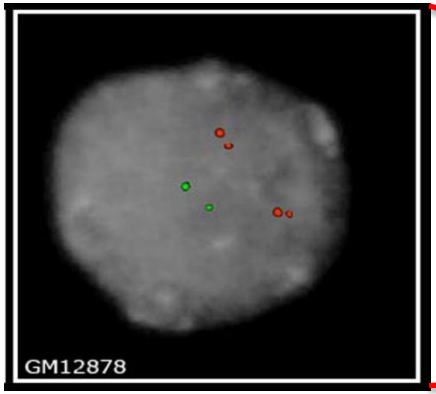
- **Two known ~70 kbp CNPs, CNP#1 duplication absent in Venter but predicted in Watson and NA12878, CNP#2 present mother but neither father or child**

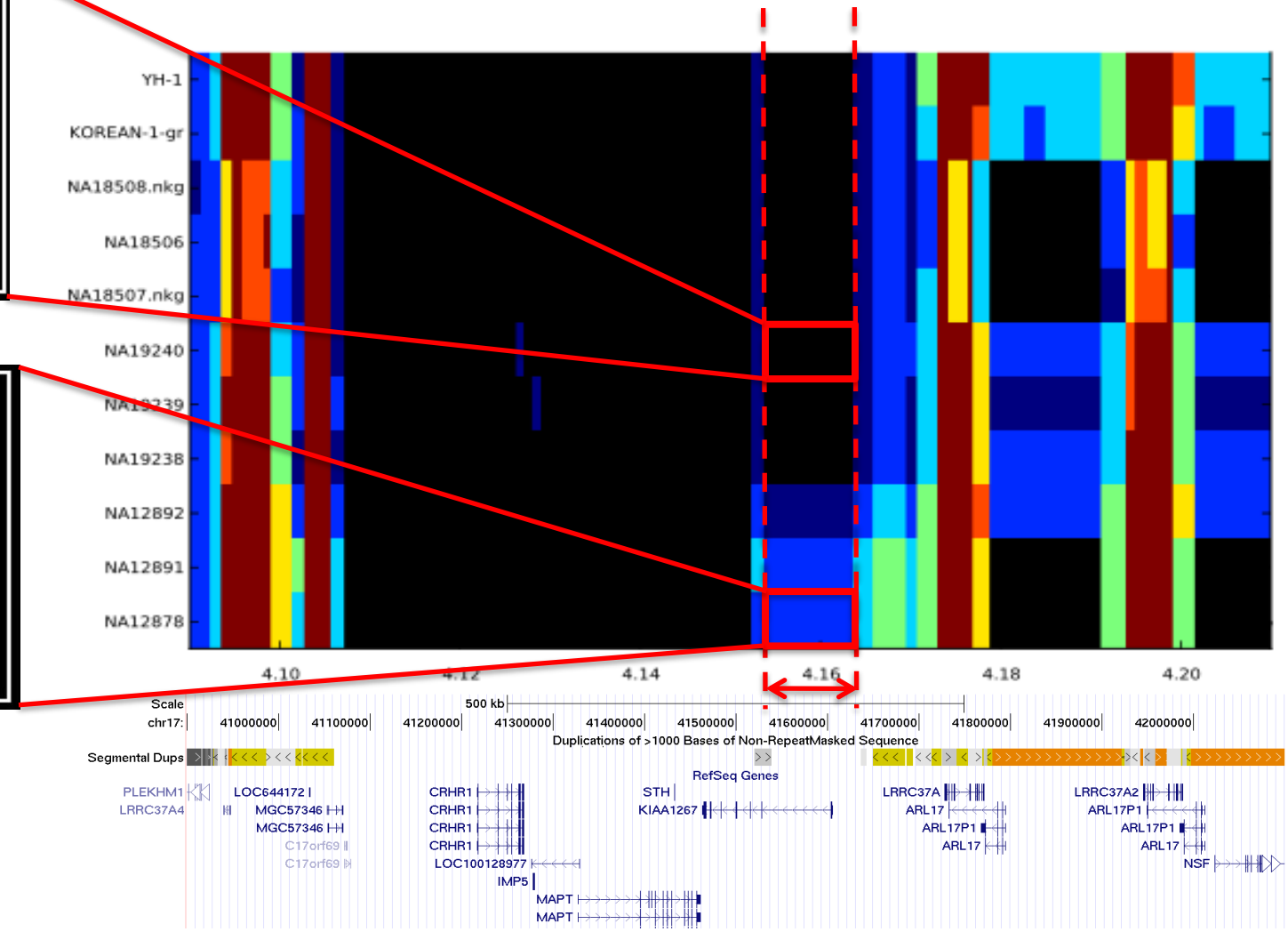
Alkan, Nat. Genet, 2009

Read-Depth CNV Heat Maps vs. FISH

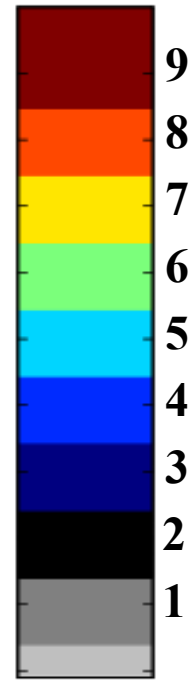
Interphase FISH



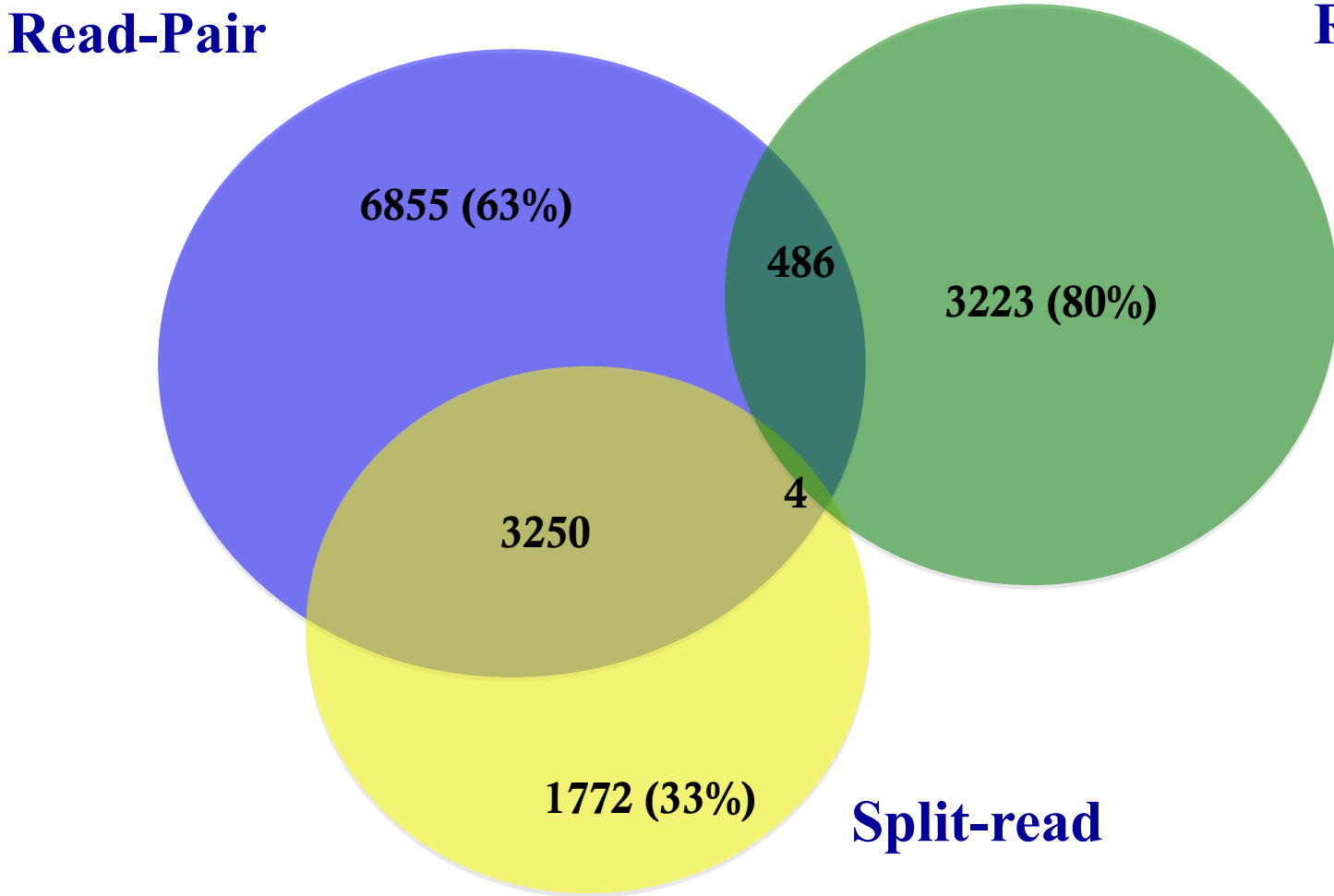




Copy Number



Indirect sequence-based approaches are incomplete
159 genomes (2-4X) (deletions only)

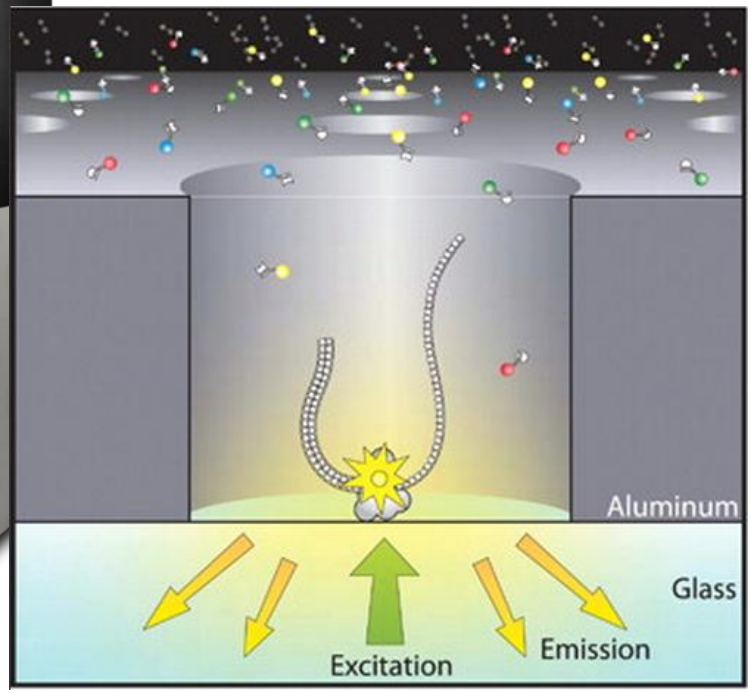


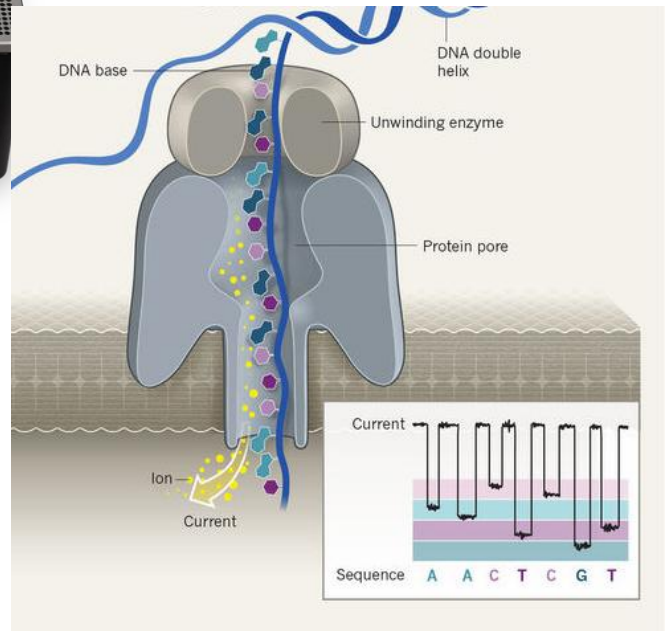
Mills *et al.*, Nature 2011

Challenges

- Size spectrum—>5 kbp discovery limit for most experimental platforms; NGS can detect much smaller but misses events mediated by repeats.
- Class bias: deletions>>>duplications>>>>balanced events (inversions)
- Multiallelic copy number states—incomplete references and the complexity of repetitive DNA
- False negatives.

Long read Genome Sequencing Revolution

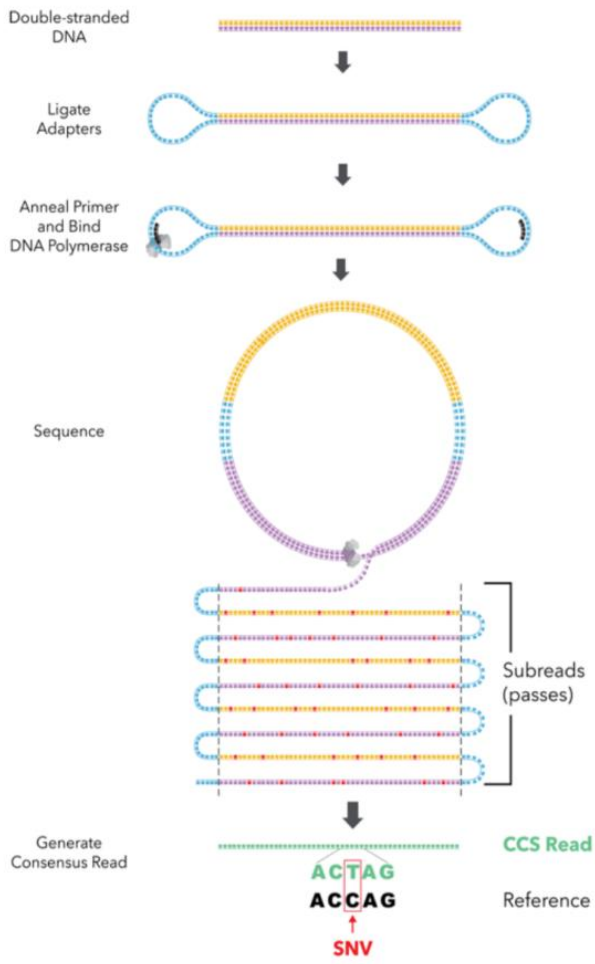




Pacific Biosciences (PacBio)—single-molecule real-time sequence (SMRT) data (15-50) kbp sequence reads
ONT (Oxford Nanopore Technology)—higher error rate but, portable, scalable native DNA sequencing of long-reads

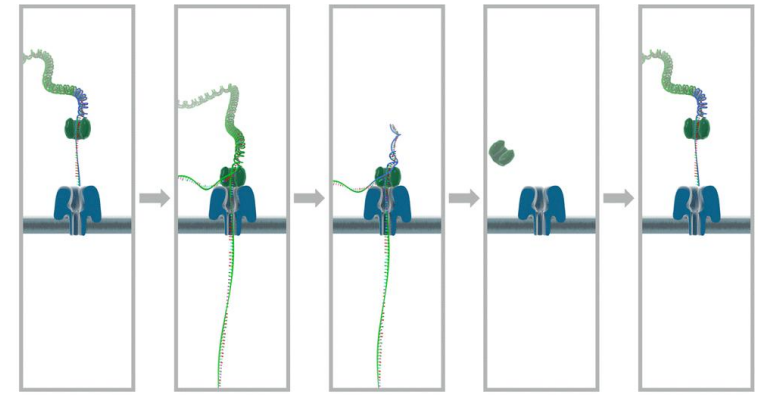
Advances in long-read sequencing

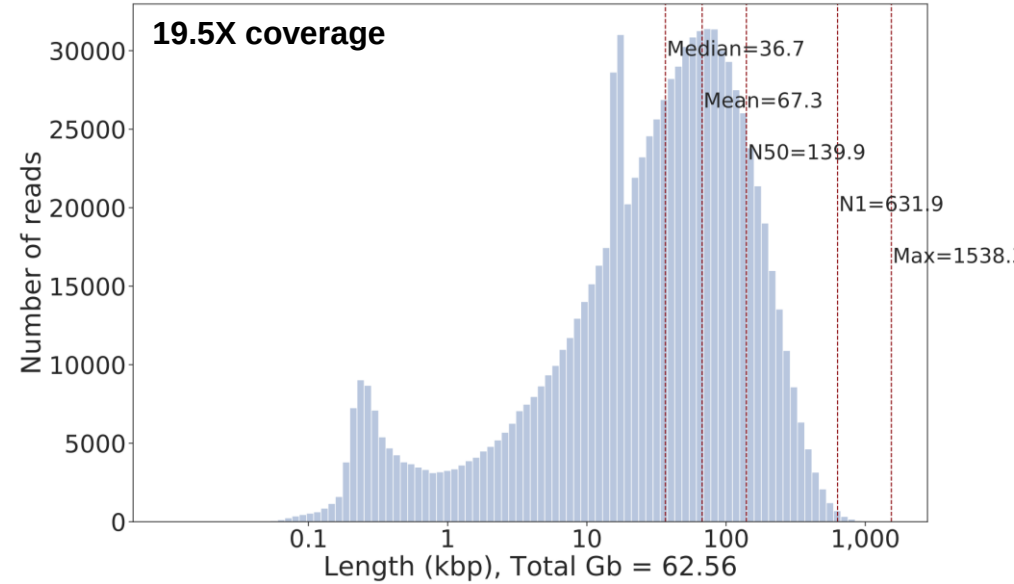
HiFi Pac Bio Sequencing



99.9% accurate 18-23 kbp reads

Ultra-long reads ONT





>100 kbp in length

Advantages of long read sequencing

Ultra-long Oxford Nanopore Technology (ONT)

~139 kbp

HiFi PacBio

~18-20 kbp

Illumina

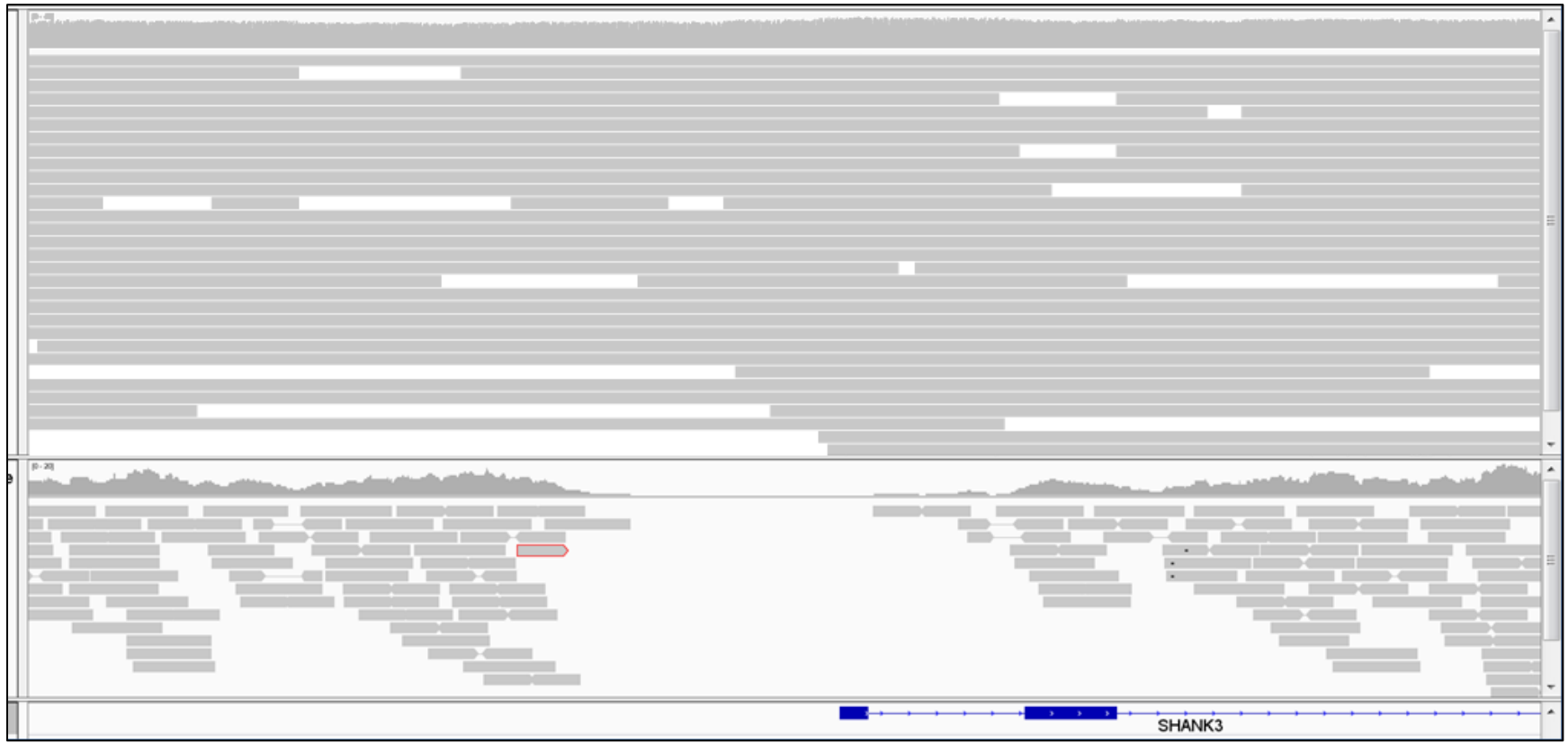
150-300 bp

More uniform coverage and sequencing of native DNA

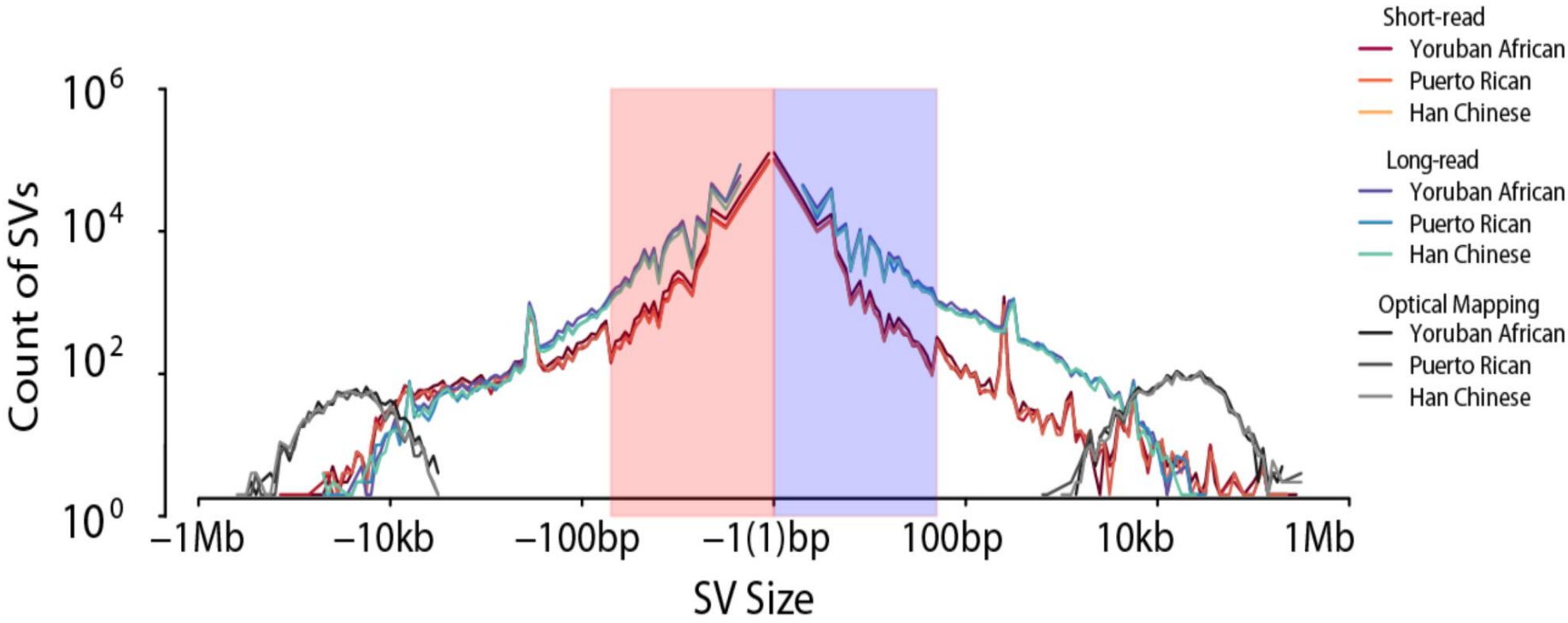
SHANK3

PacBio Sequence Coverage

Illumina Sequence Coverage



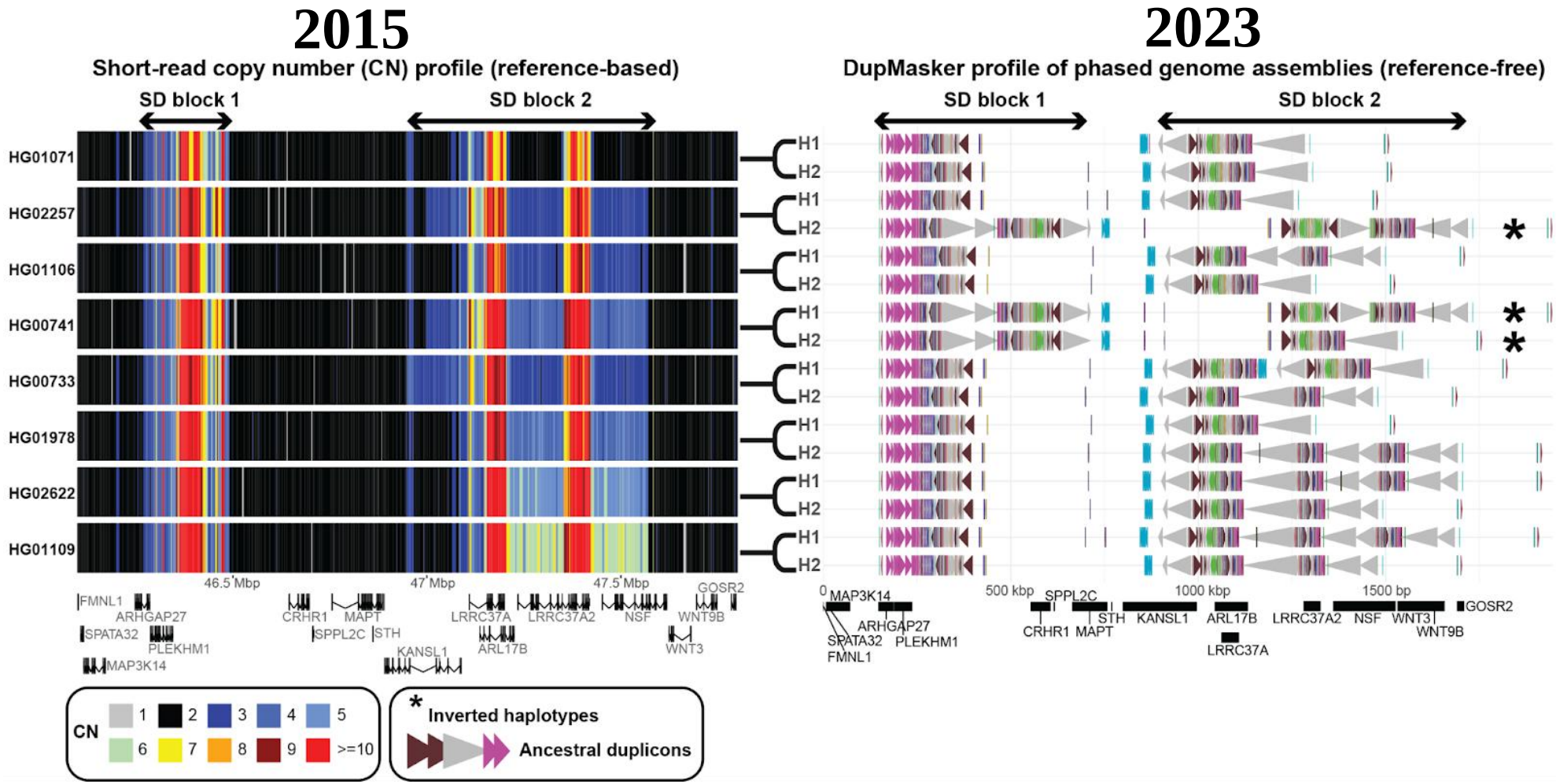
Increased sensitivity for structural variation (SV)



- ~25,000 PacBio SVs vs. 11,000 Illumina SVs >50 bp
- Eleven Illumina callers combined detect 49% of deletions and 11% of insertions in a human genome--**NGS misses 75% of SVs**

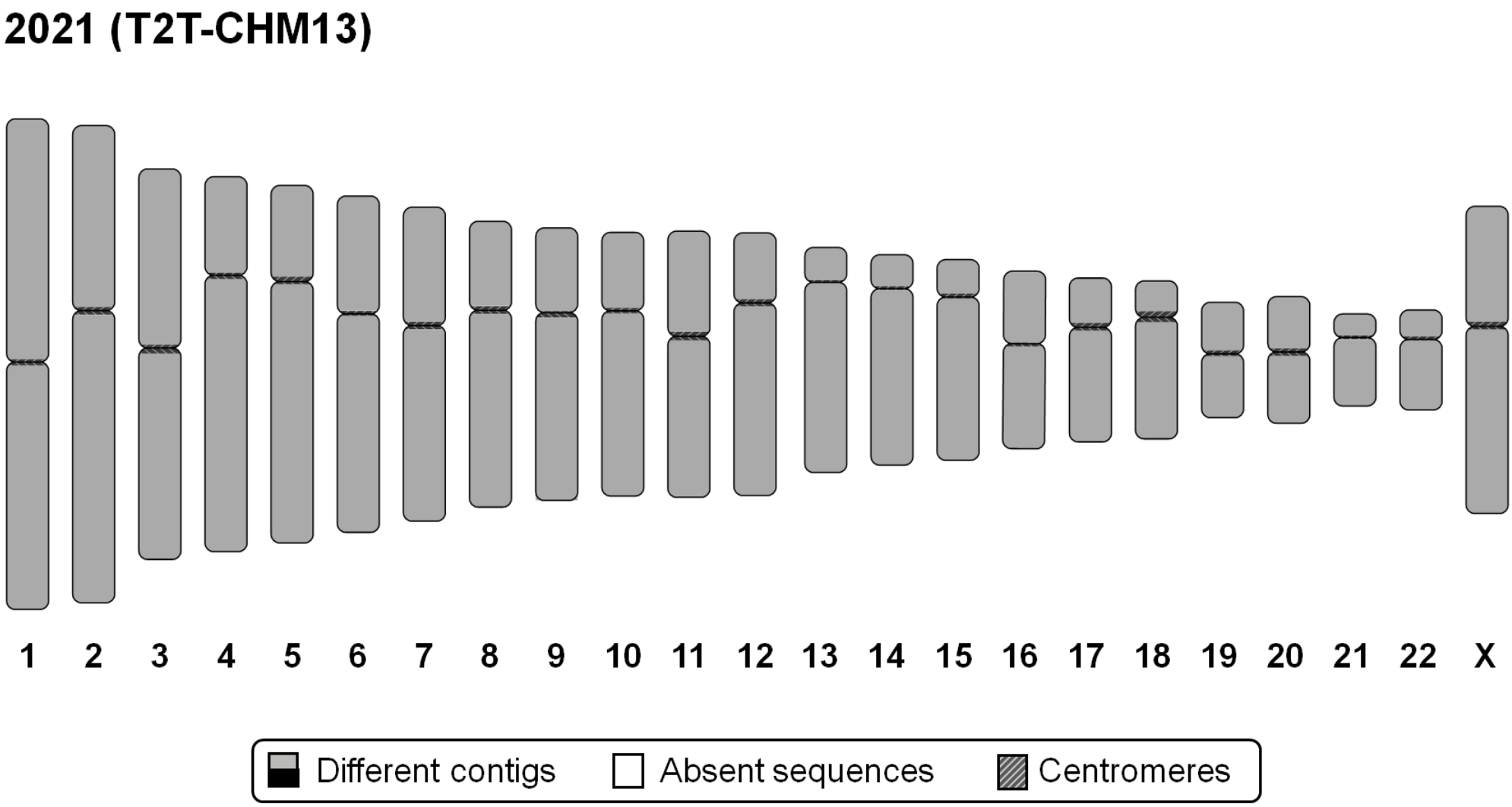
Chaisson et al, Nature, 2015; Chaisson et al., Nat Comm, 2019

LRS has transformed how we characterize copy number and structural variation



Sudmant et al, Science, 2015, Porubsky et al, Cell, 2024

Complete sequence of human genome

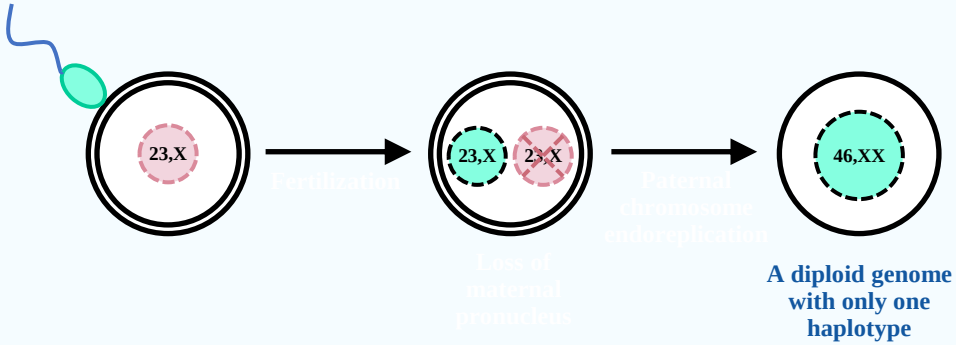


Nurk et al, bioRxiv, 2021, Science 2022

So how did we do it?

We used an *effectively haploid* human cell line known as CHM13

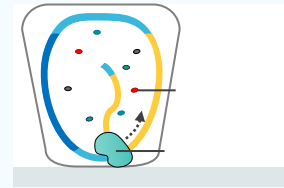
CHM13 is a complete hydatidiform mole



This greatly simplifies this problem because it allows us to assemble each chromosome without interference from a second set of chromosomes

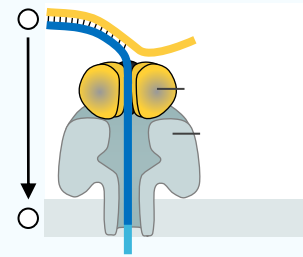
We used two long-read sequencing technologies with complementary strengths

1. Pacific Biosciences (PacBio) high-fidelity (HiFi)

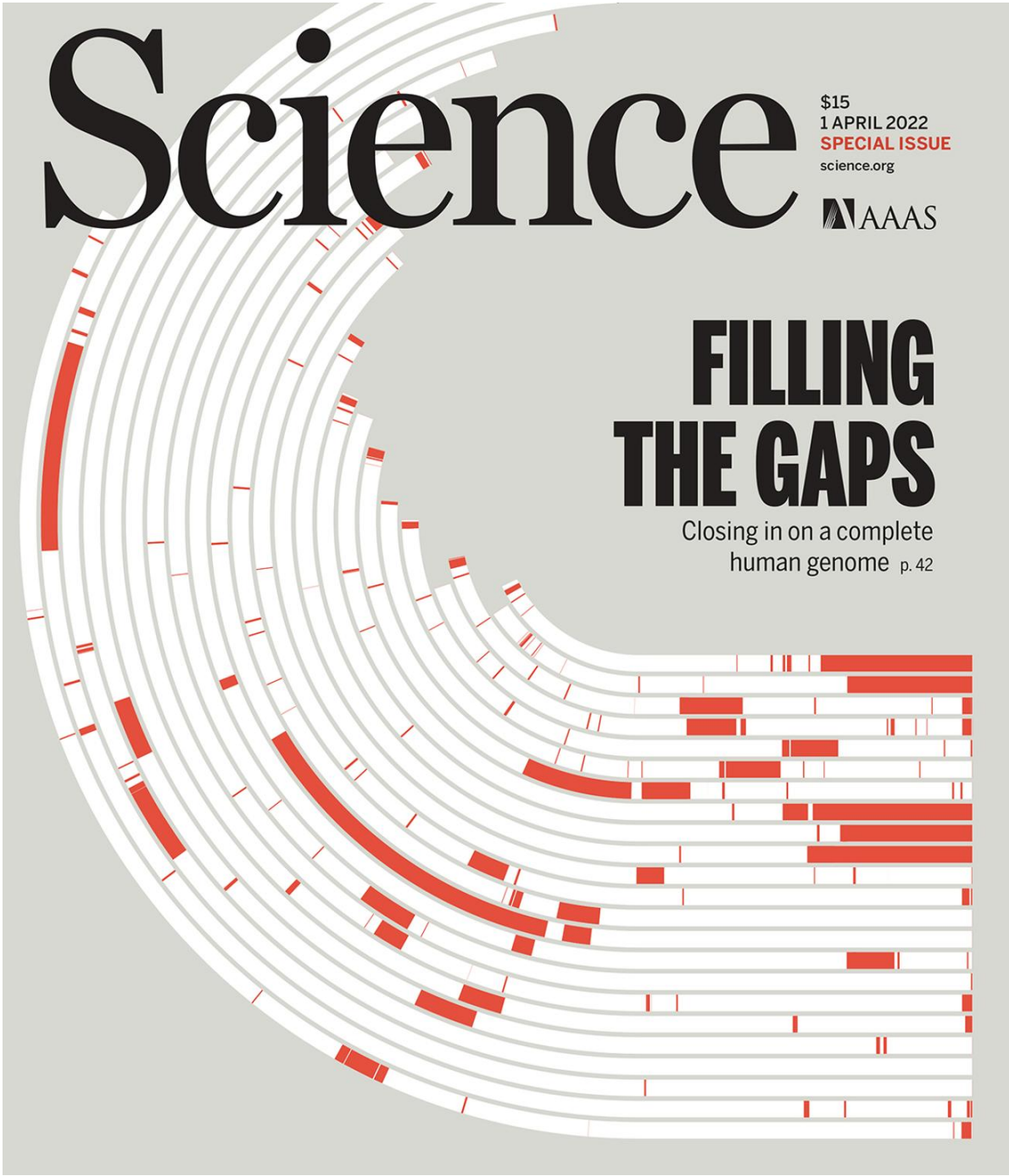


- 15-25 kbp long
- >99% accurate (similar to Illumina)
- Strength: Extremely accurate

2. Oxford Nanopore Technologies (ONT)



- No limit in read length!
- 93-99% accurate
- Strength: Extremely long

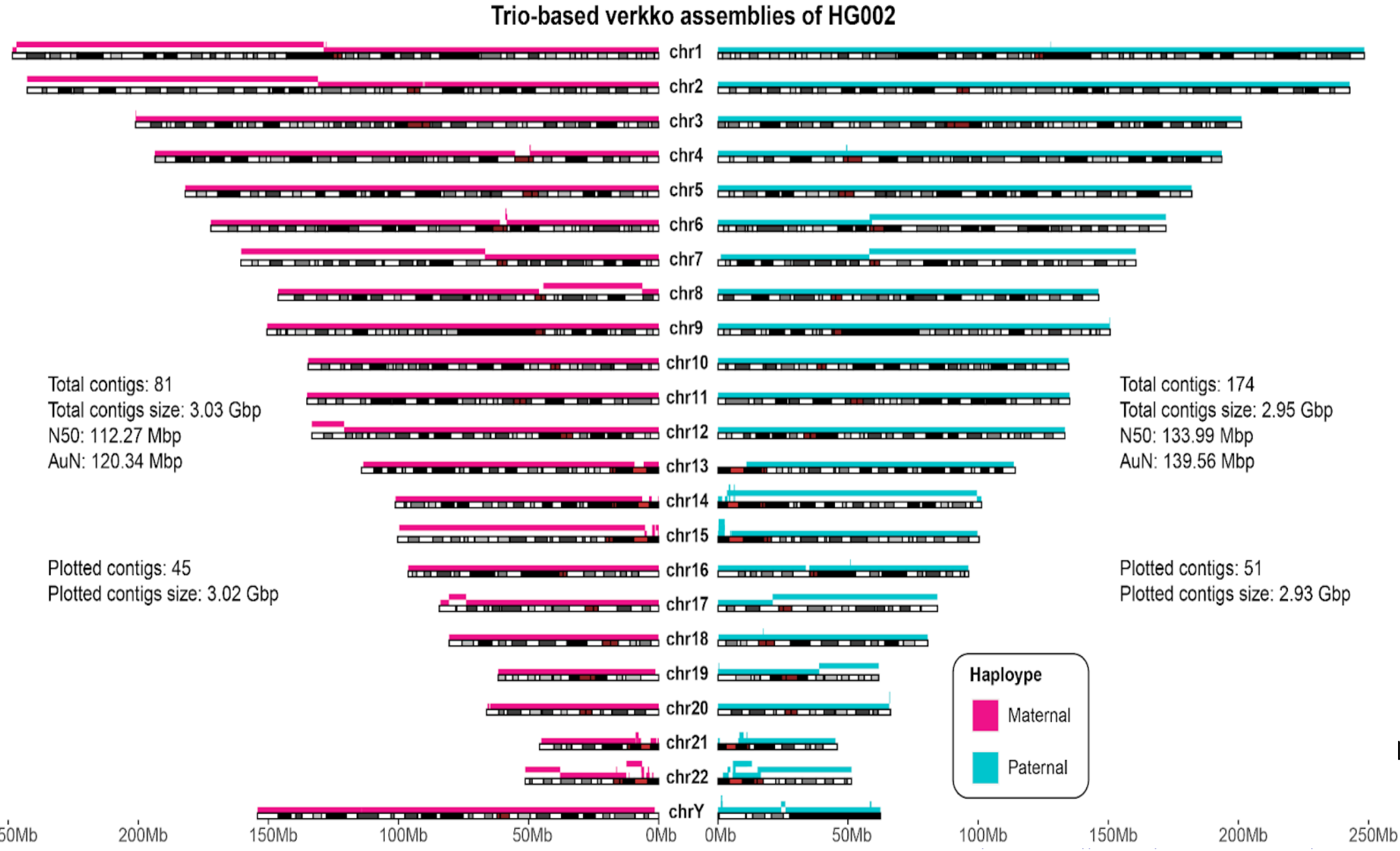


- 8% of missing genome sequence added (>200 Mbp)
- Complete sequence of centromeres, acrocentric and segmental duplications
- Adds 1956 gene predictions of which 130-190 are protein coding
- Framework for understanding the genetically most complex regions of our genome.

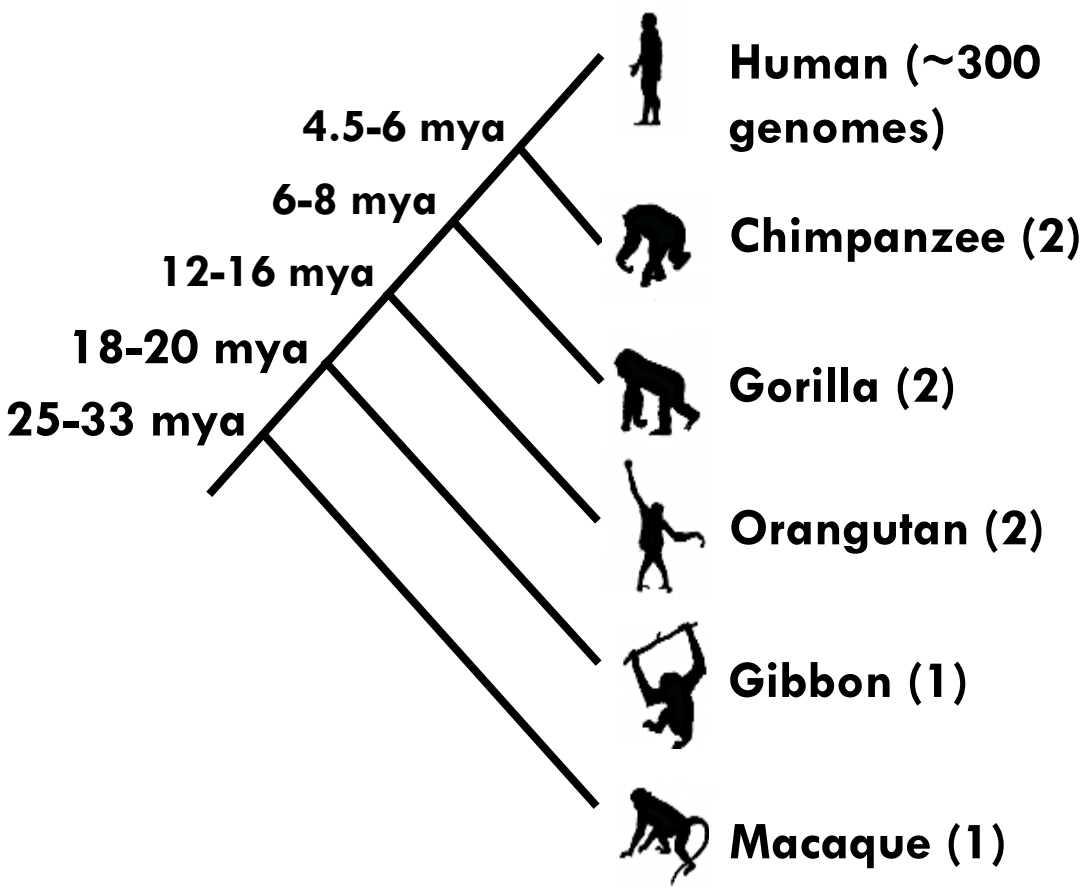
A 6 Gbp Human Genome Assembly

(contig N50=25-28 Mbp)

Porubsky et al, Nat. Biotech, 2020


Trio-based verkko assemblies of HG002
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrY
Total contigs: 81
Total contigs size: 3.03 Gbp
N50: 112.27 Mbp
AuN: 120.34 Mbp
Plotted contigs: 45
Plotted contigs size: 3.02 Gbp
Total contigs: 174
Total contigs size: 2.95 Gbp
N50: 133.99 Mbp
AuN: 139.56 Mbp
Plotted contigs: 51
Plotted contigs size: 2.93 Gbp
Haploype
Maternal
Paternal
250Mb
200Mb
150Mb
100Mb
50Mb
0Mb
0Mb
50Mb
100Mb
150Mb
200Mb
250Mb

Primate phased genome assembly efforts

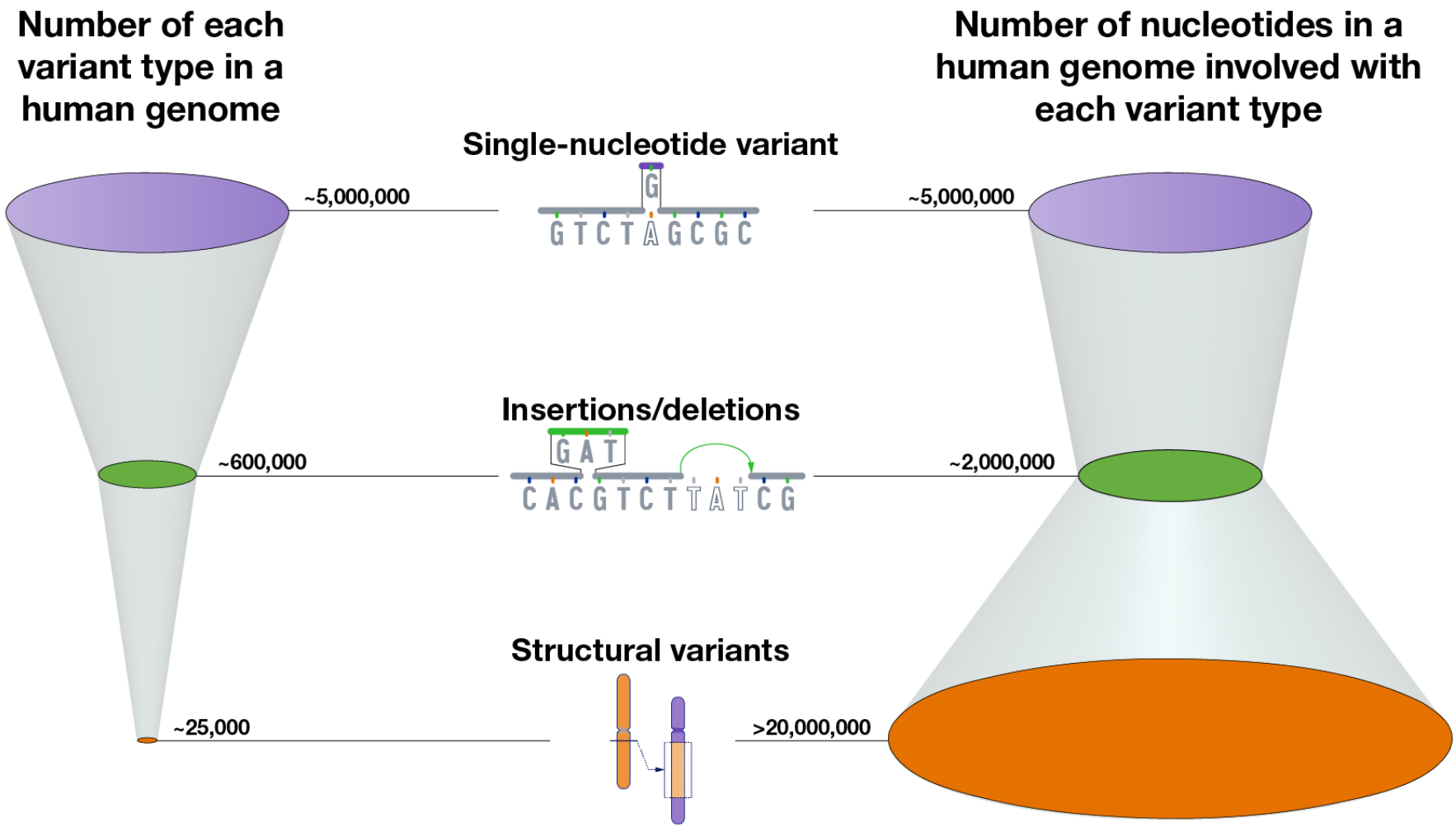


HGSVC

HUMAN PANGENOME

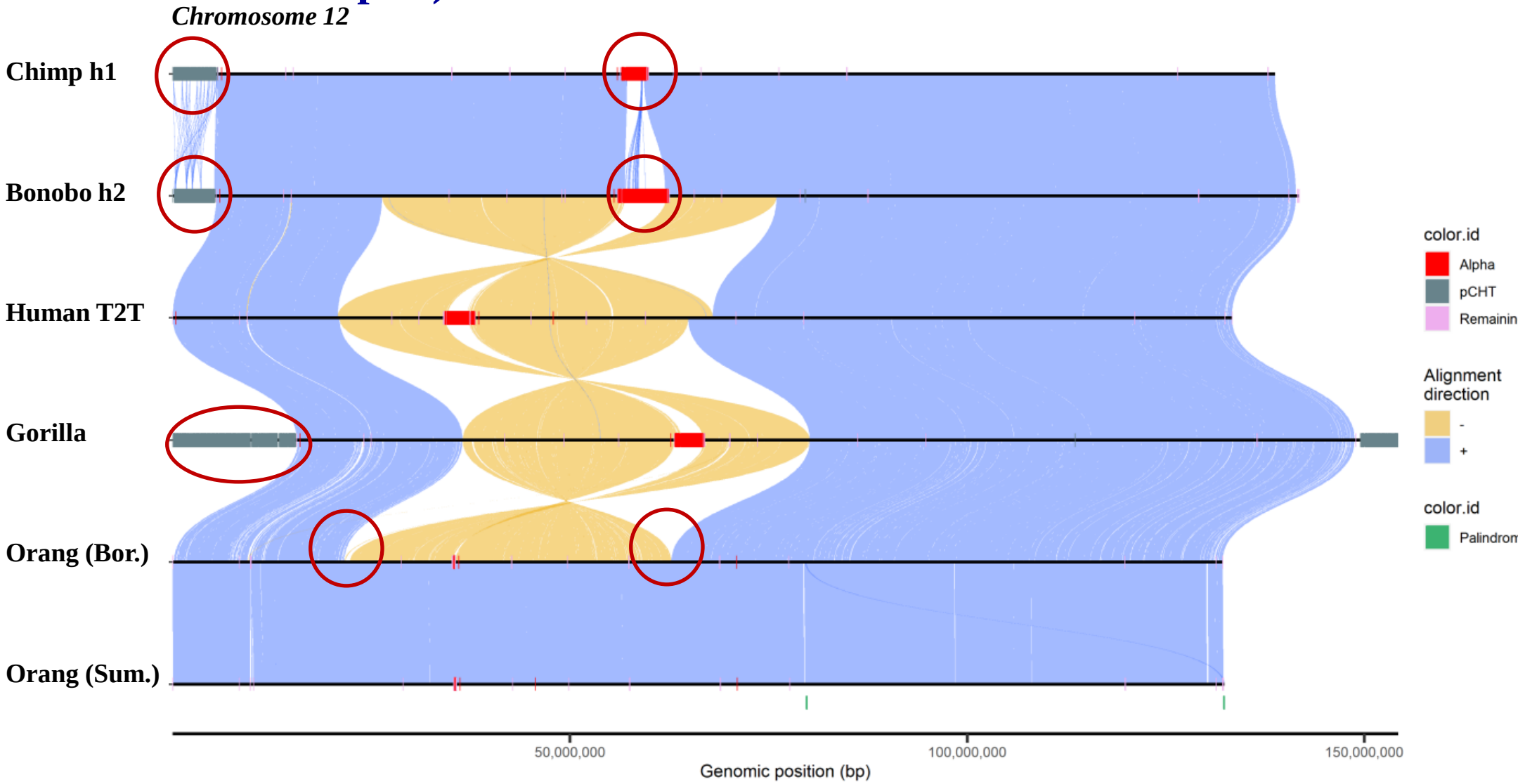
Kronenberg et al, *Science*, 2017; Ebert et al., Science, 2021; Liao/HPRC Nature, 2023; Yoo et al, bioRxiv, 2024

Complete spectrum of human genetic variation



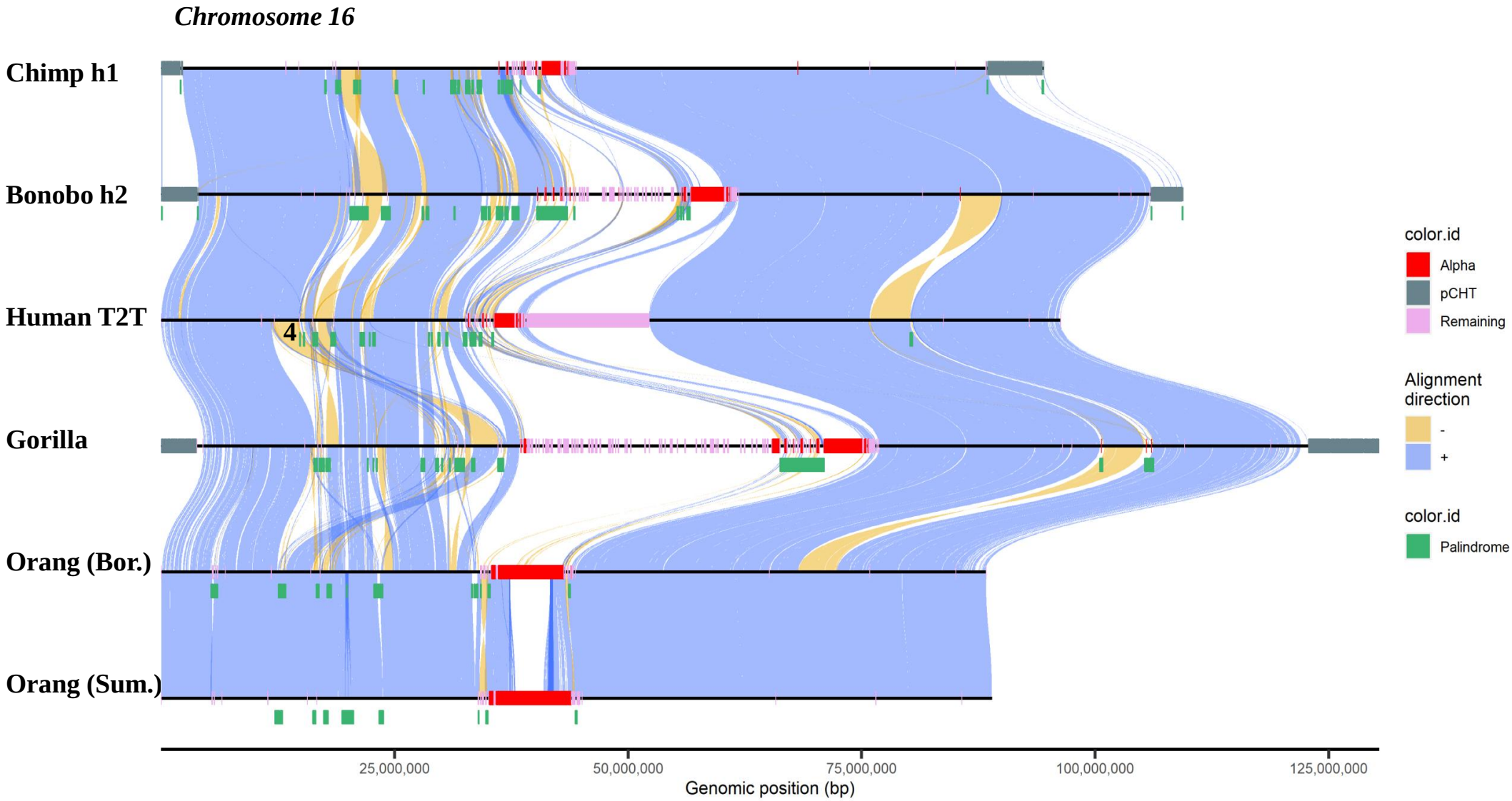
Porubsky & Eichler, *Cell*, 2024

Complete sequencing of ape chromosomes (SVbyEye "stacked" plot)

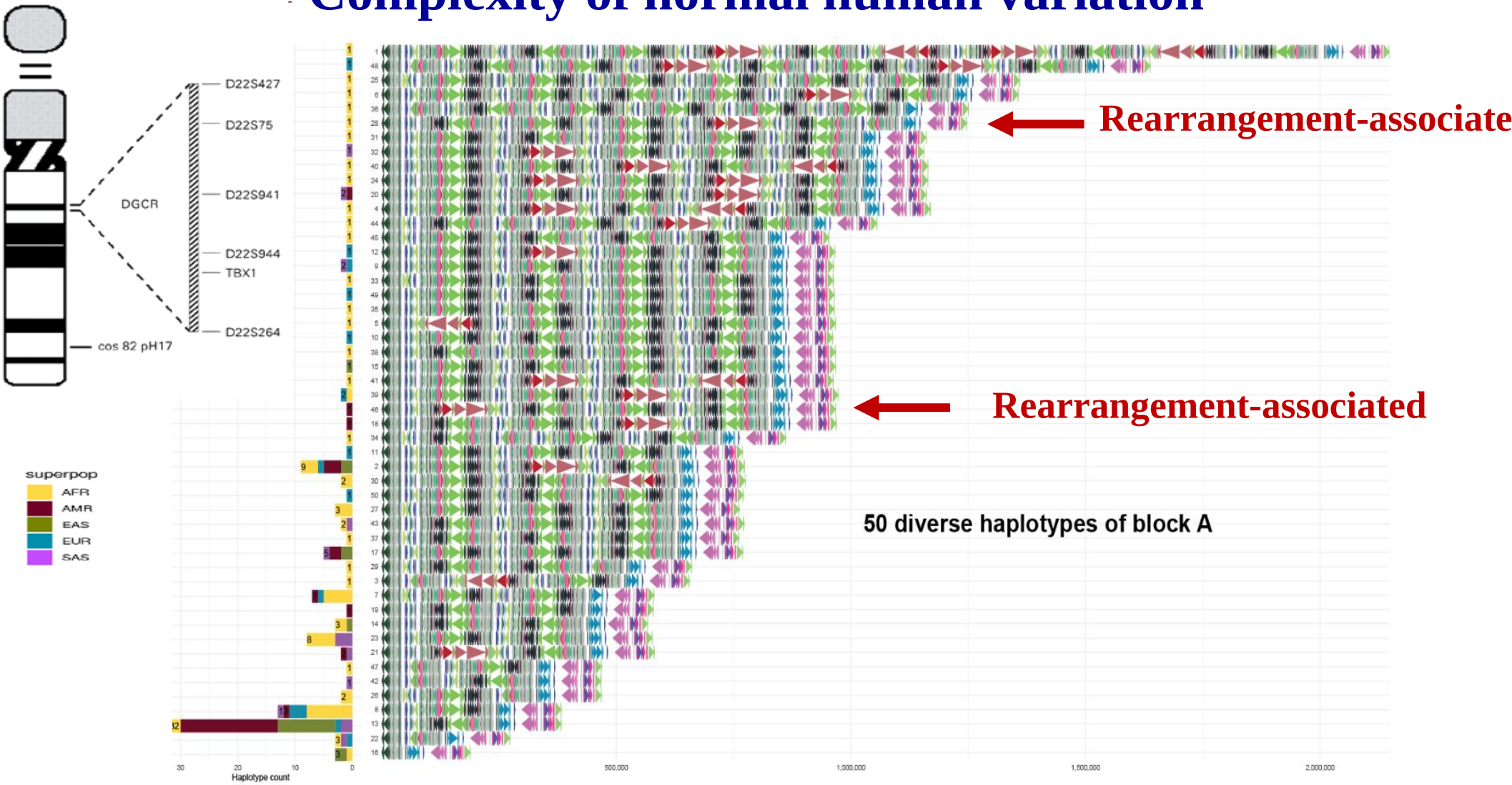


Yoo et al, bioRxiv, 2024

Complete sequencing of ape chromosomes

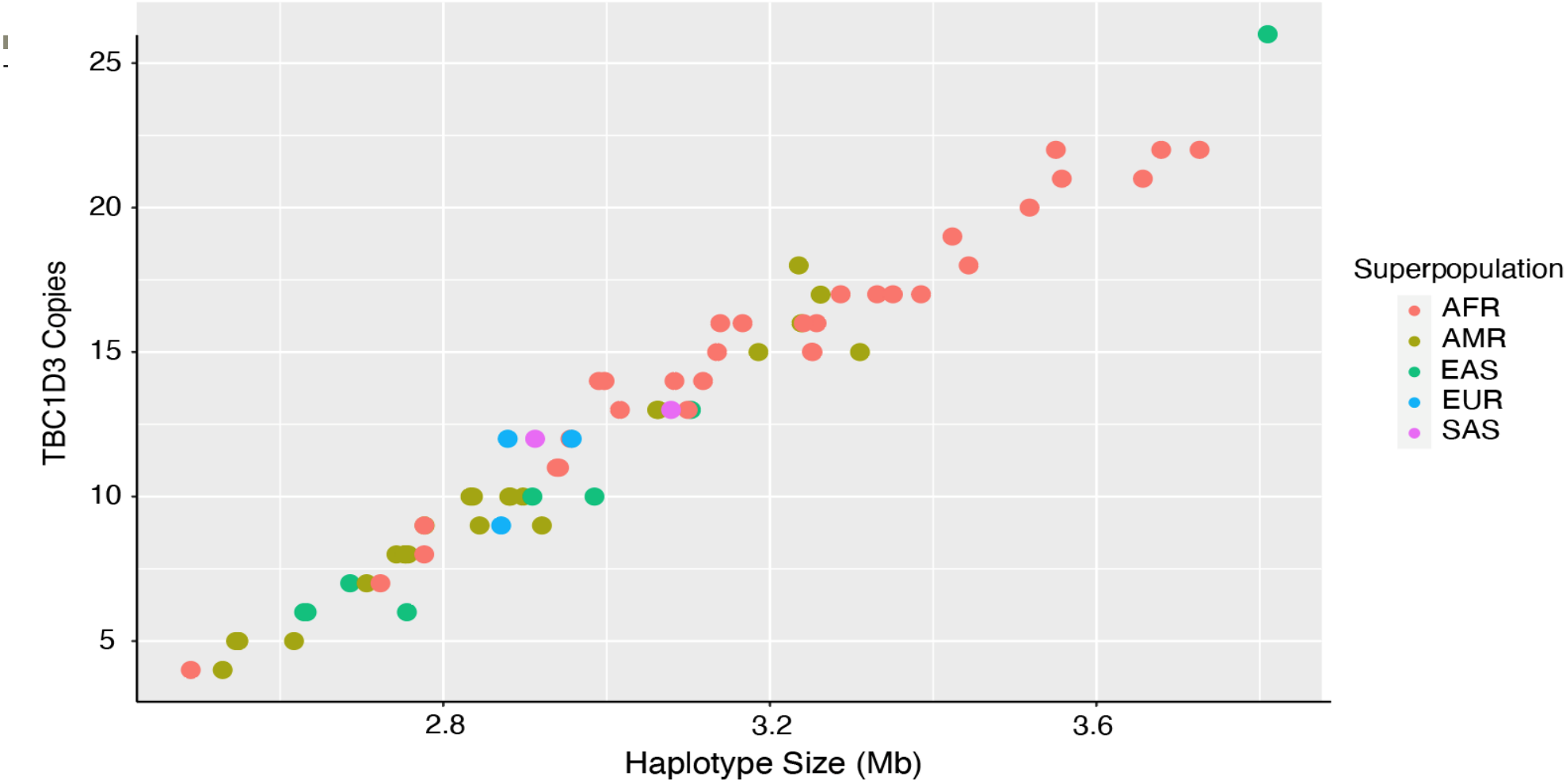


22q11 syndrome LCRA.
Complexity of normal human variation



Porubsky *unpublished*

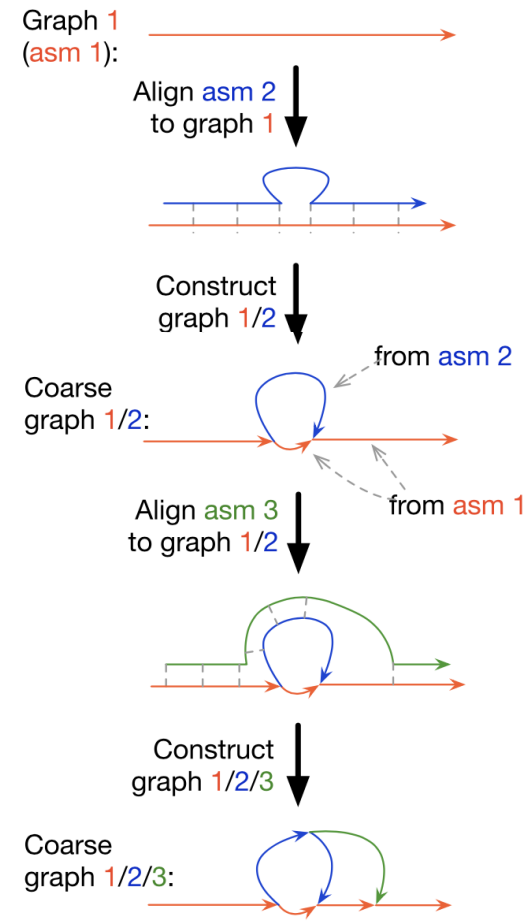
Copy number and structural variation of *TBC1D3*



Guitart et al, *Genome Res*, 2024

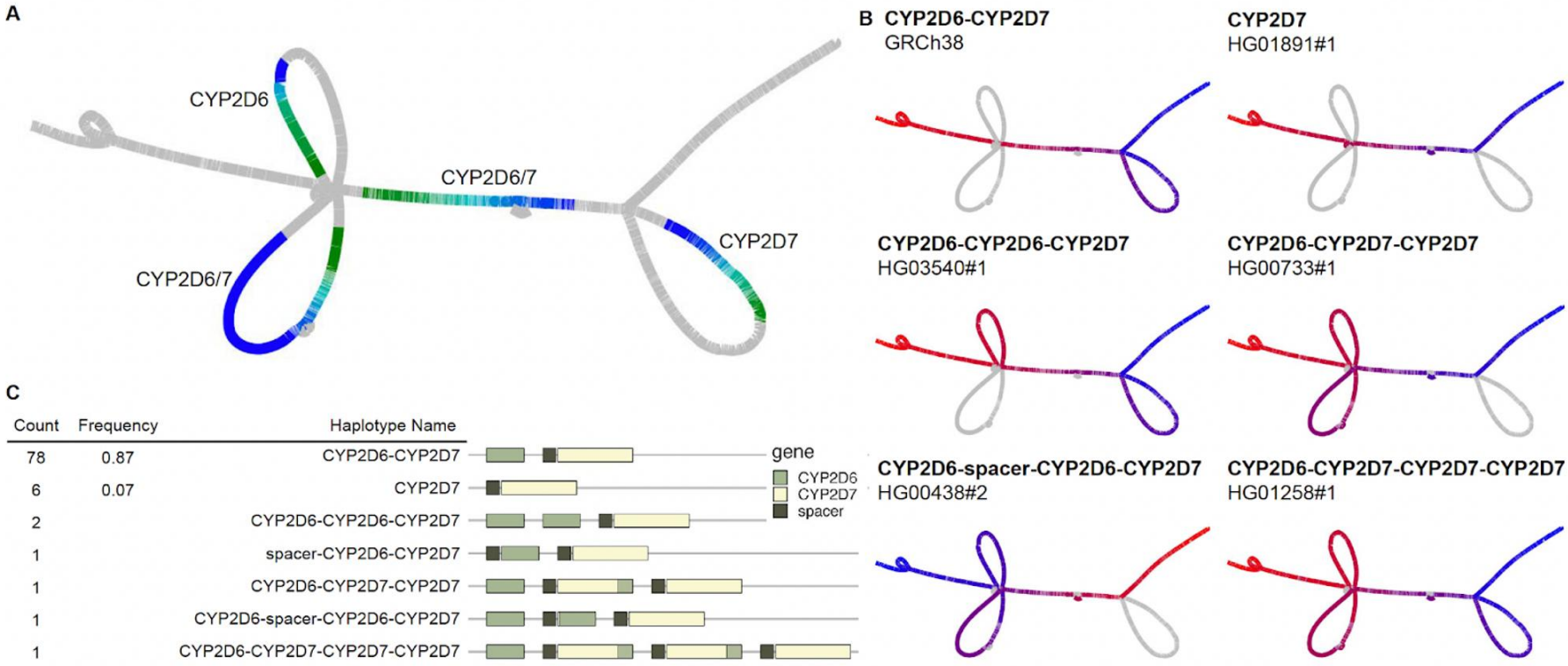
A graph can capture such variation e.g. Minigraph

1. **Generate phase genome assemblies**
2. **Iteratively introduce assembly sequence to a graph.**
3. **Distinguish query sequence already present in graph from novel sequence**
4. **Include novel sequence as new segments or edges between segments in graph.**
5. **Repeat with next assembly**



Guitart, X

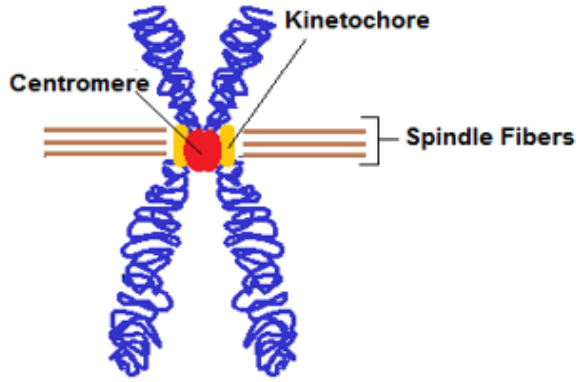
A graph-based representation of structural variation

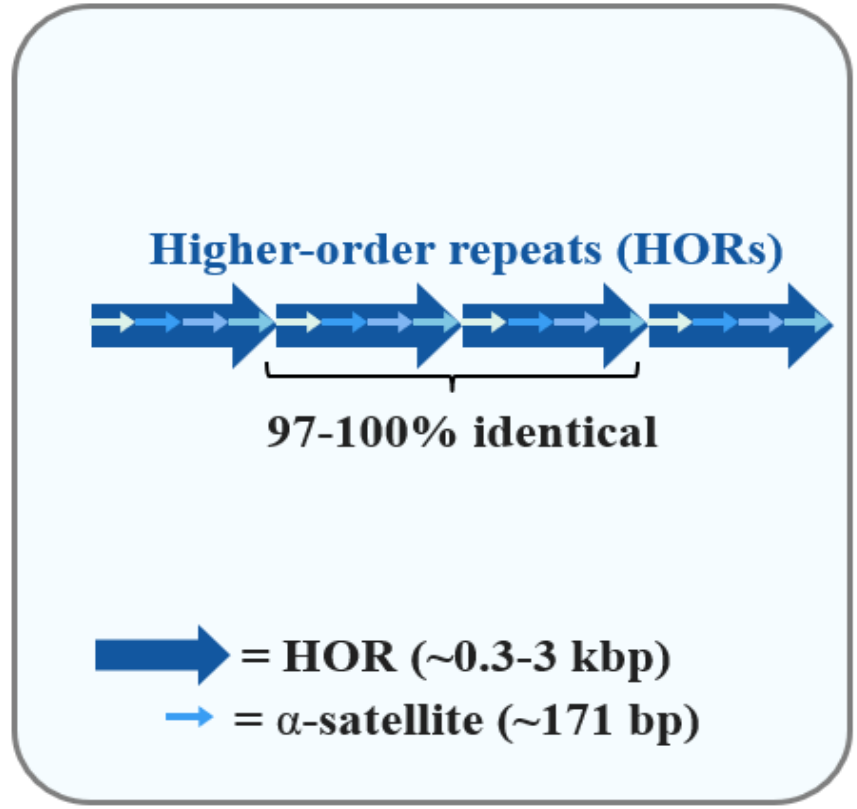


Liao et al., bioRxiv, 2022, Nature, 2023

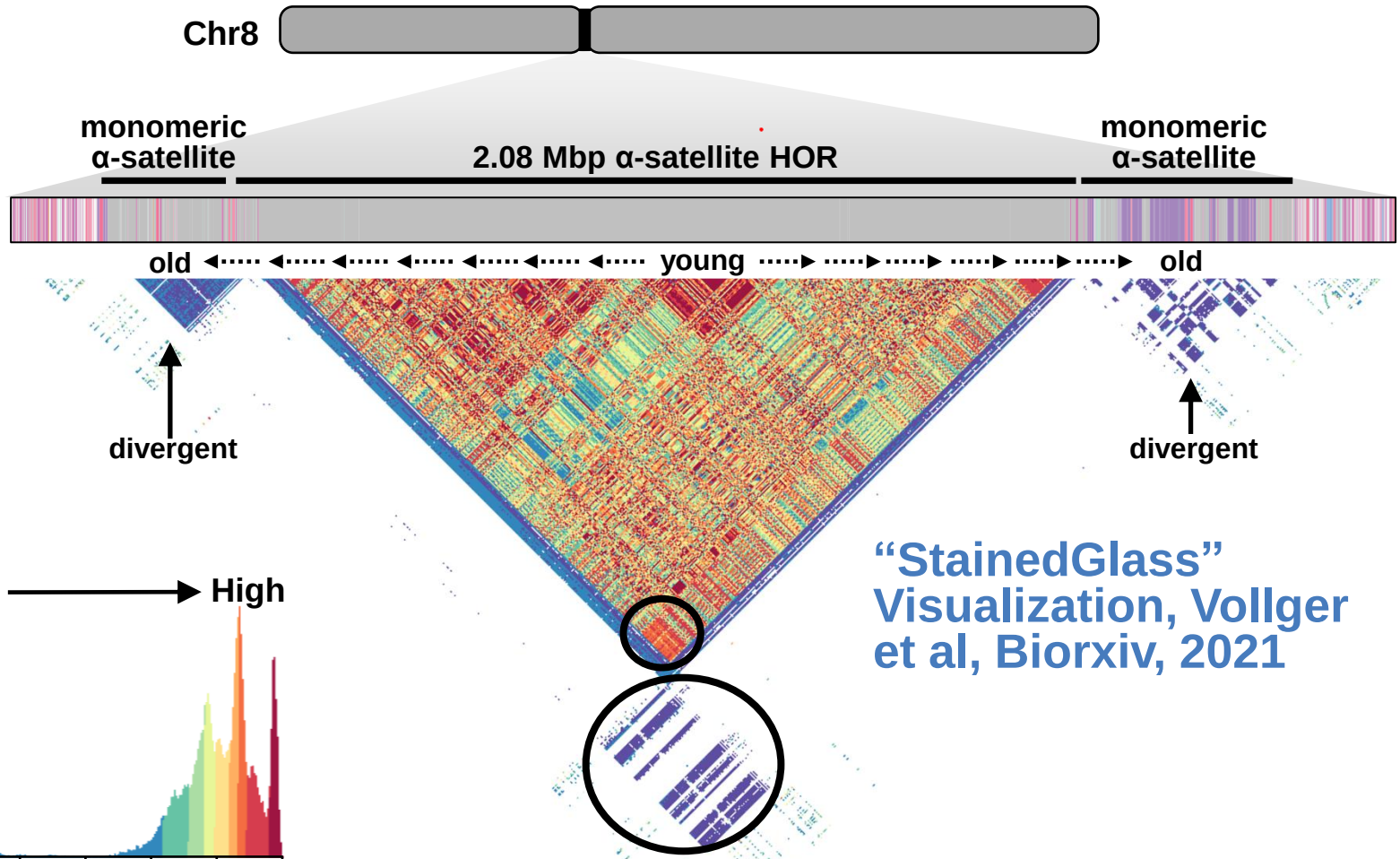
A graph-based representation of the entire human genome as a conceptual new reference.

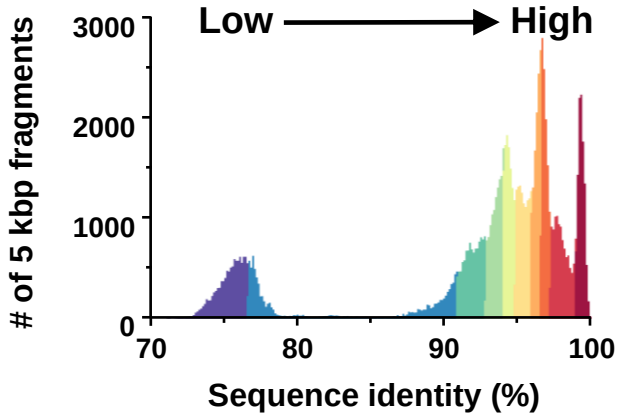
Access to previously inaccessible regions of human genome: Centromeres





Centromere organization



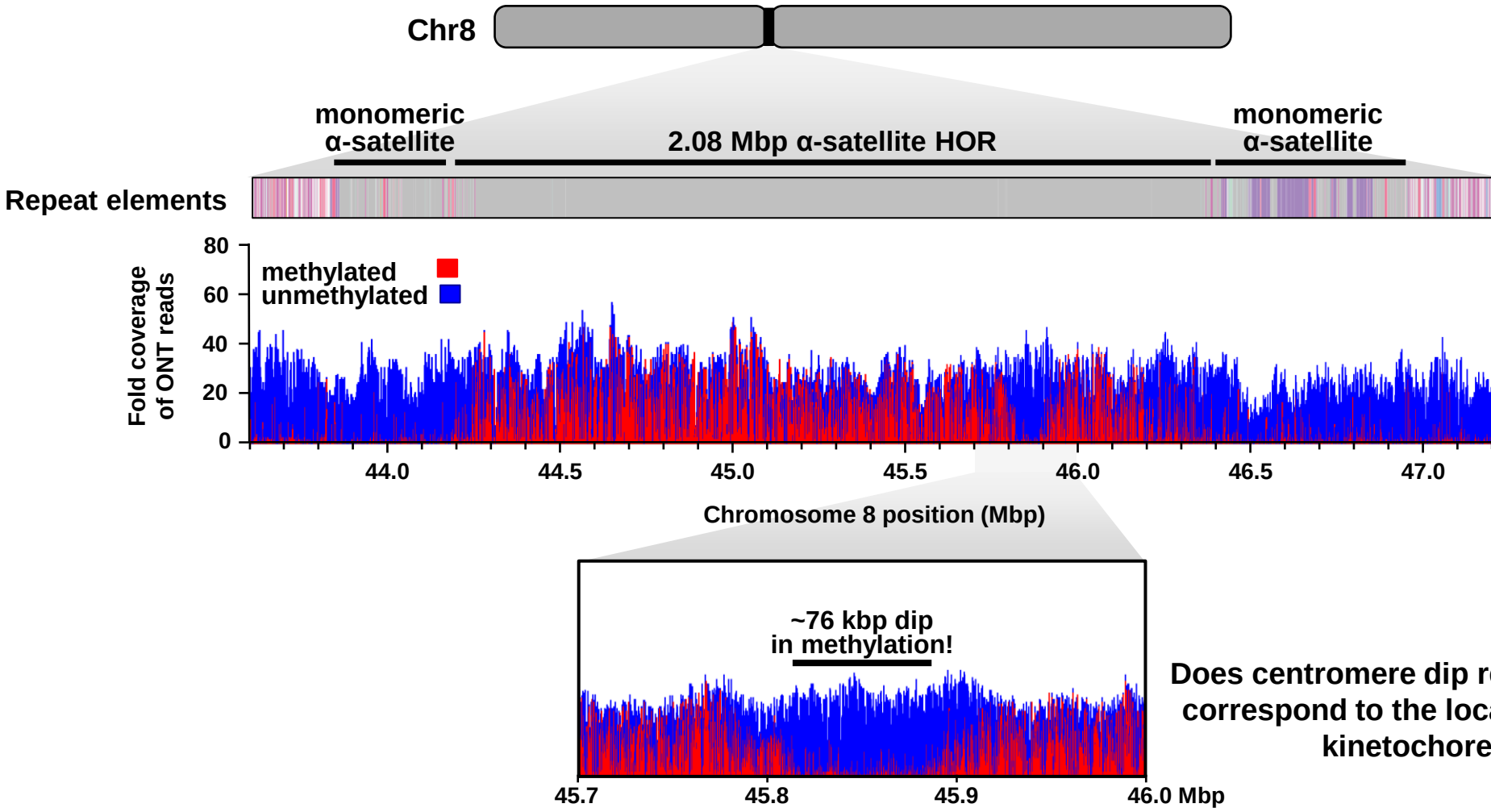


"StainedGlass" Visualization, Vollger et al, Biorxiv, 2021

1. Mirror symmetry
2. Layered nature

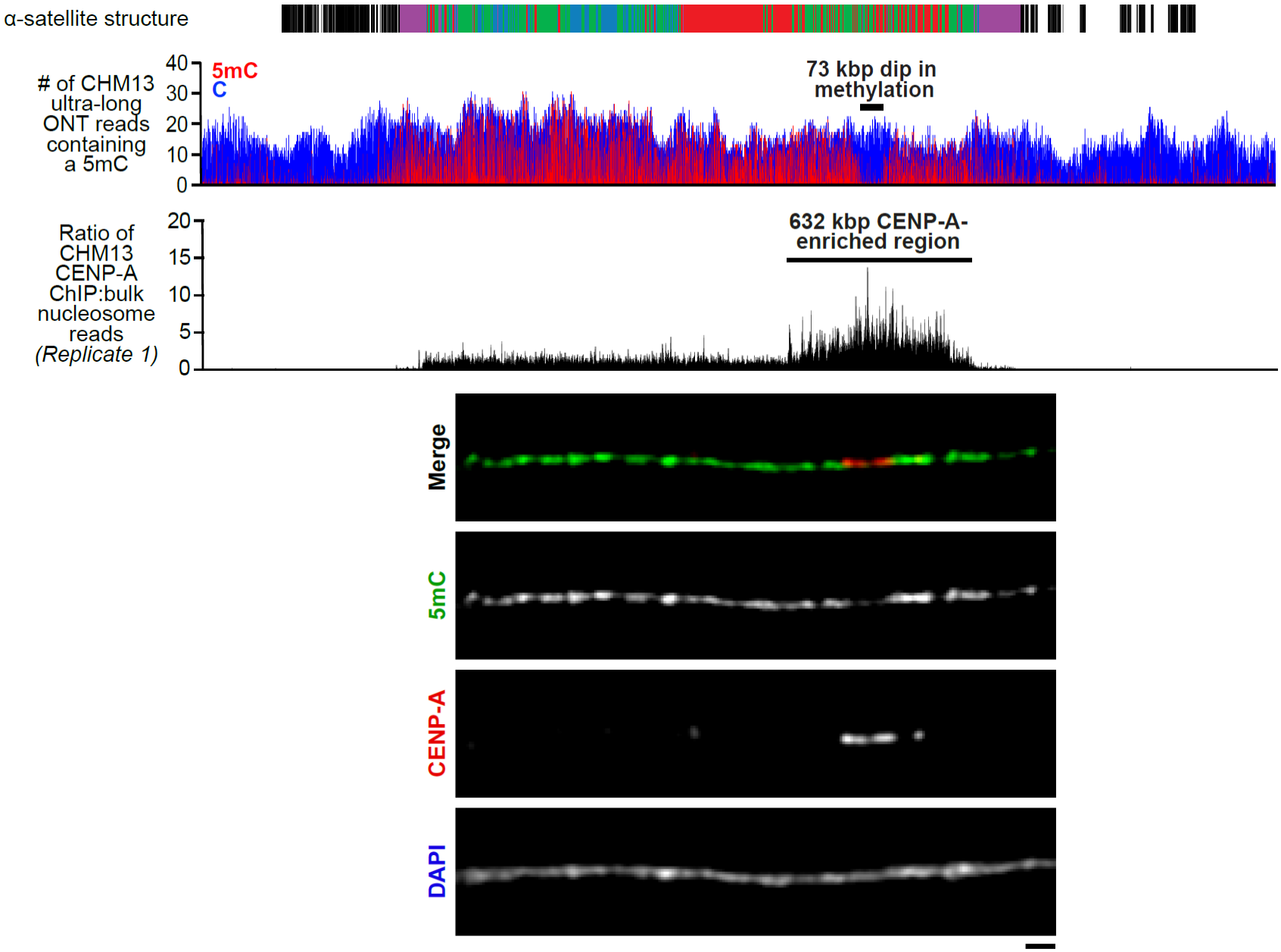
Logsdon et al., bioRxiv, 2020, Nature, 2021

Understanding centromere structure and function



Does centromere dip region (CDR) correspond to the location of the kinetochore?

Understanding centromere structure and function

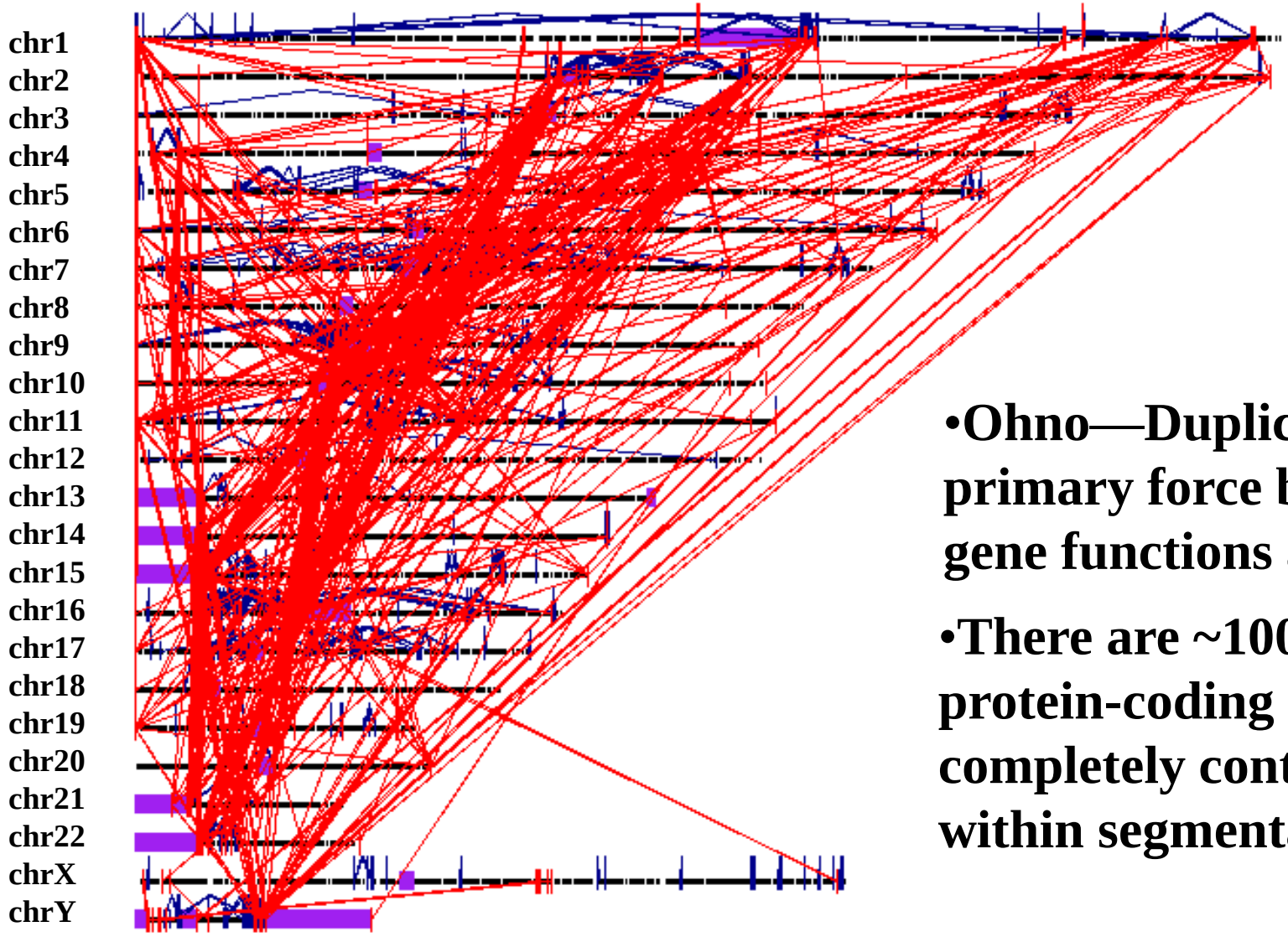


Logsdon et al., bioRxiv, 2020, Nature, 2021

Summary

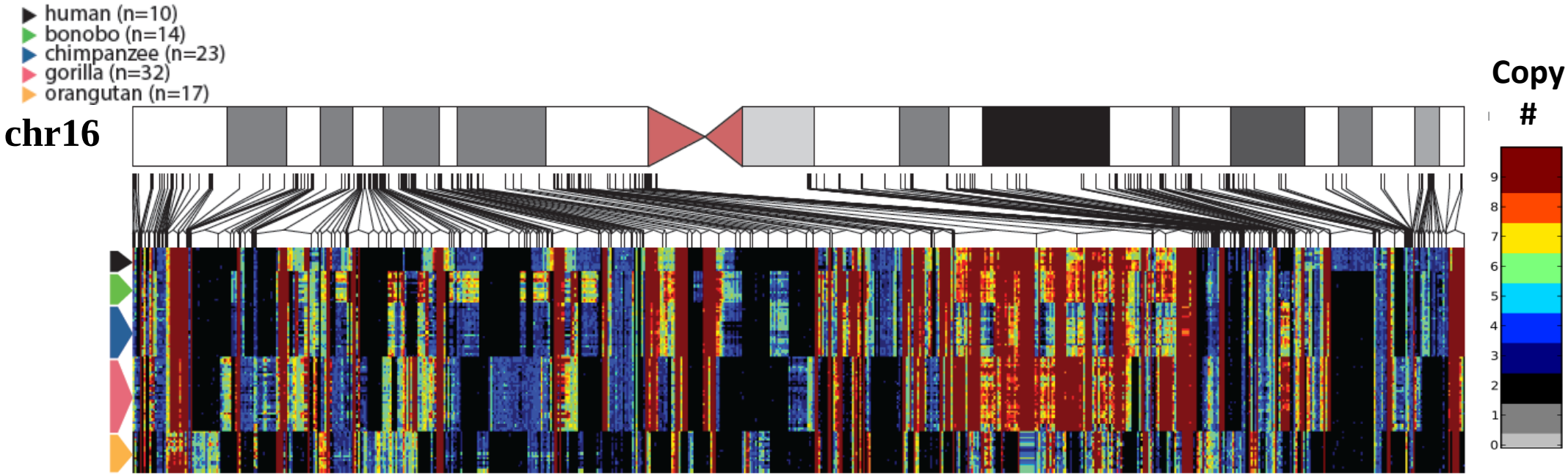
- Short read NGS approaches
  - Multiple methods are needed—readpair+read-depth+splitread often with orthogonal validation such as SNP microarray
  - ~75% of SVs are missed because SVs are non-randomly distributed to repetitive regions where mapping quality is low
  - Read-depth approaches allow CNV prediction but not structure
- Long-read sequencing methods provide complete SV but currently limited throughput
  - Read-based versus assembly-based approaches
  - Telomere-to-telomere assemblies of human genomes now possible or nearly so for diploid—**complete genetic information where all variants are phased.**
  - First human pangenomes now available—a new concept to eventually replace a singular reference.

III. Why?



- **Ohno—Duplication is the primary force by which new gene functions are created**
- **There are ~1000 annotated protein-coding genes completely contained within segmental duplications**

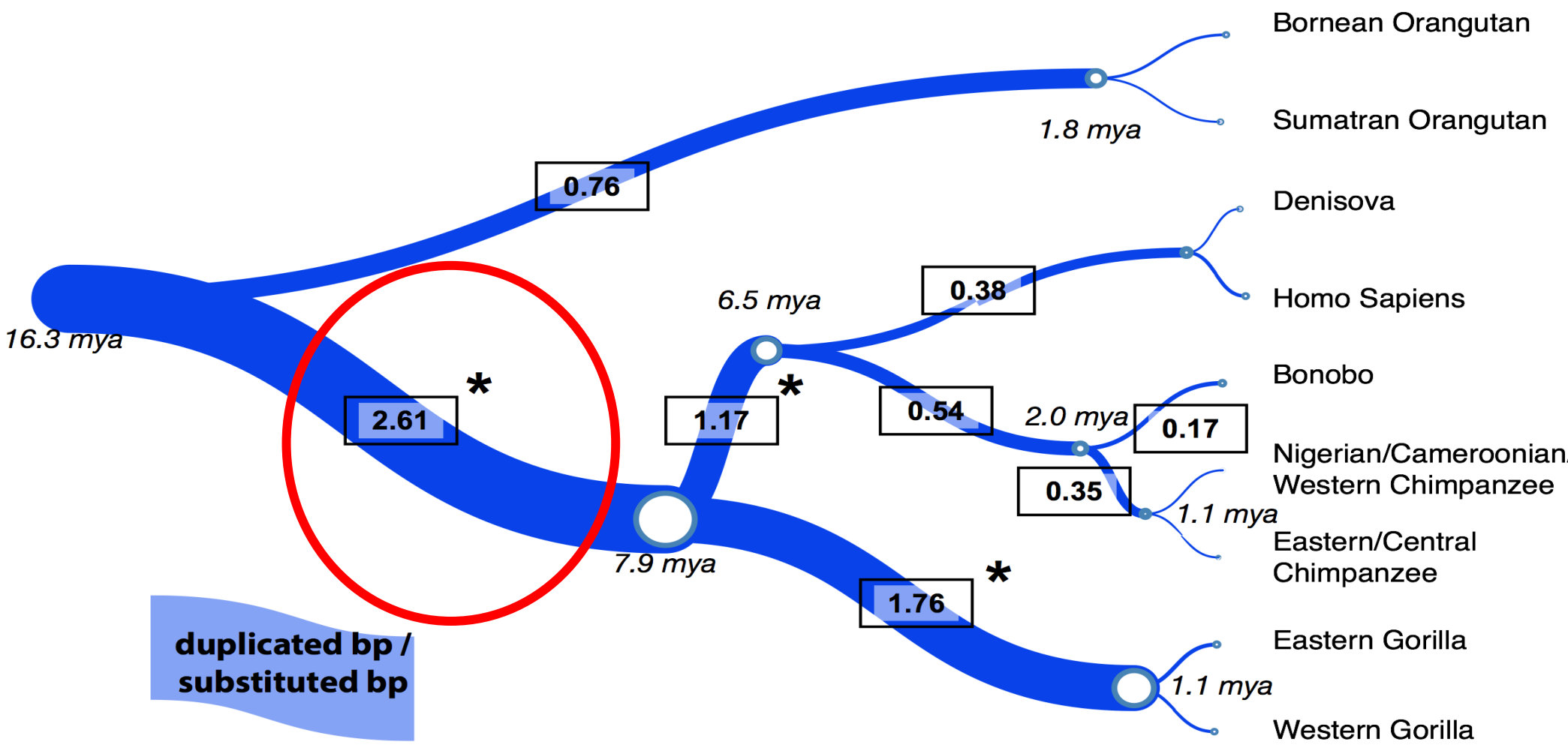
Dynamic Genetic Variation



- **Genomic copy number changes contributes more genetic difference between apes and humans than SNVs**
- **468 Mbp CNV vs. 167 Mbp SNVs (ration: 2.8)**

Sudmant *et al*., Genome Res., 2013, Sudmant *et al, Science, 2015*

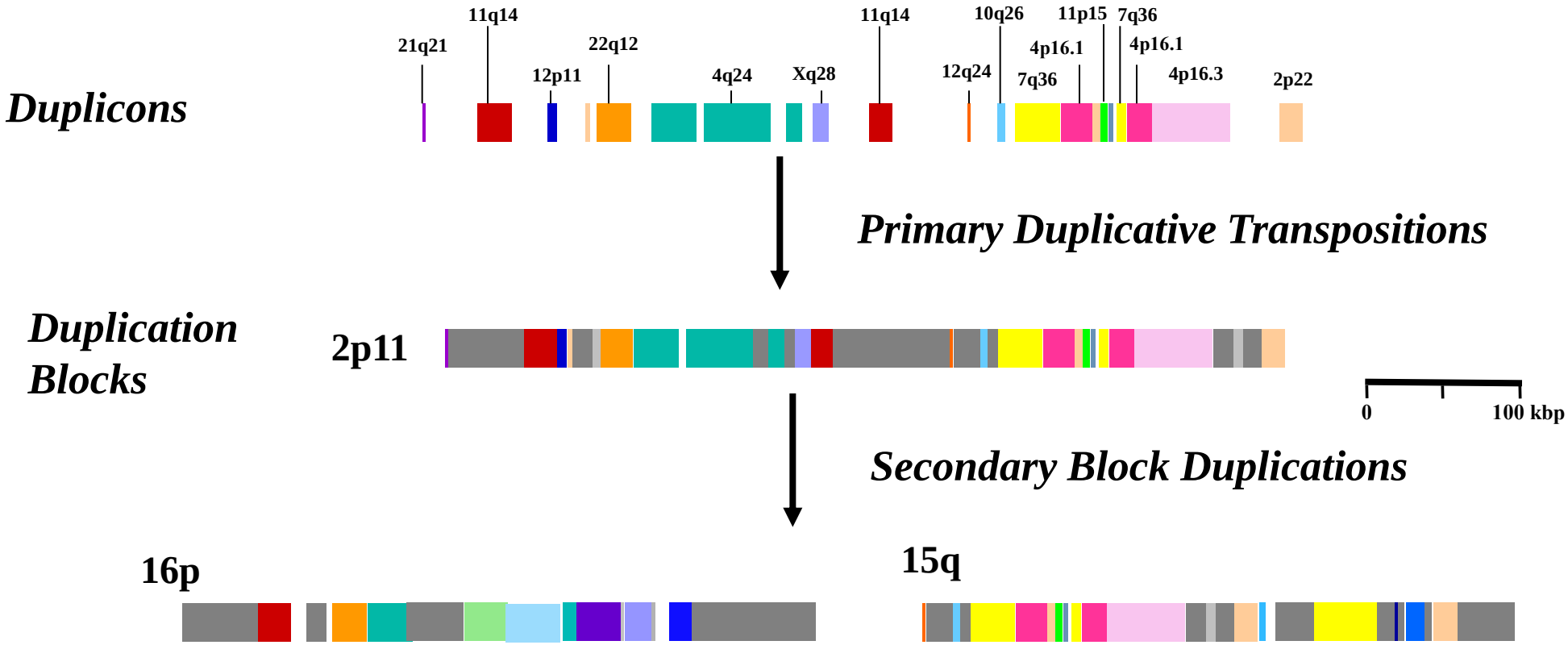
Rate of Duplication



$p=9.786 \times 10^{-12}$

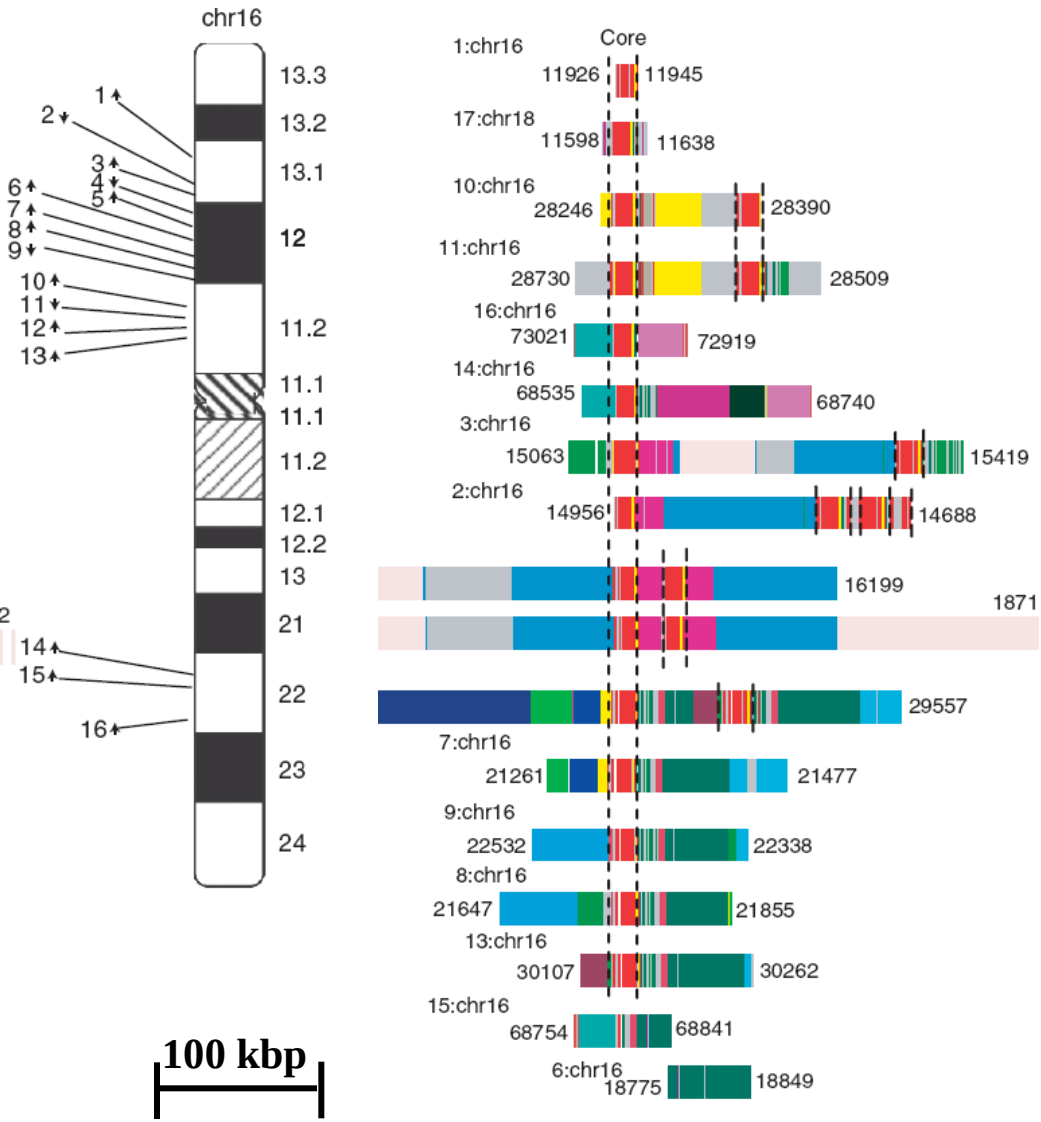
Sudmant PH et al. , *Genome Res.* 2013

Mosaic Architecture



- A mosaic of recently transposed duplications
- Duplications within duplications.
- Potentiates “exon shuffling”, regulatory innovation

Human Chromosome 16 Core Duplicon

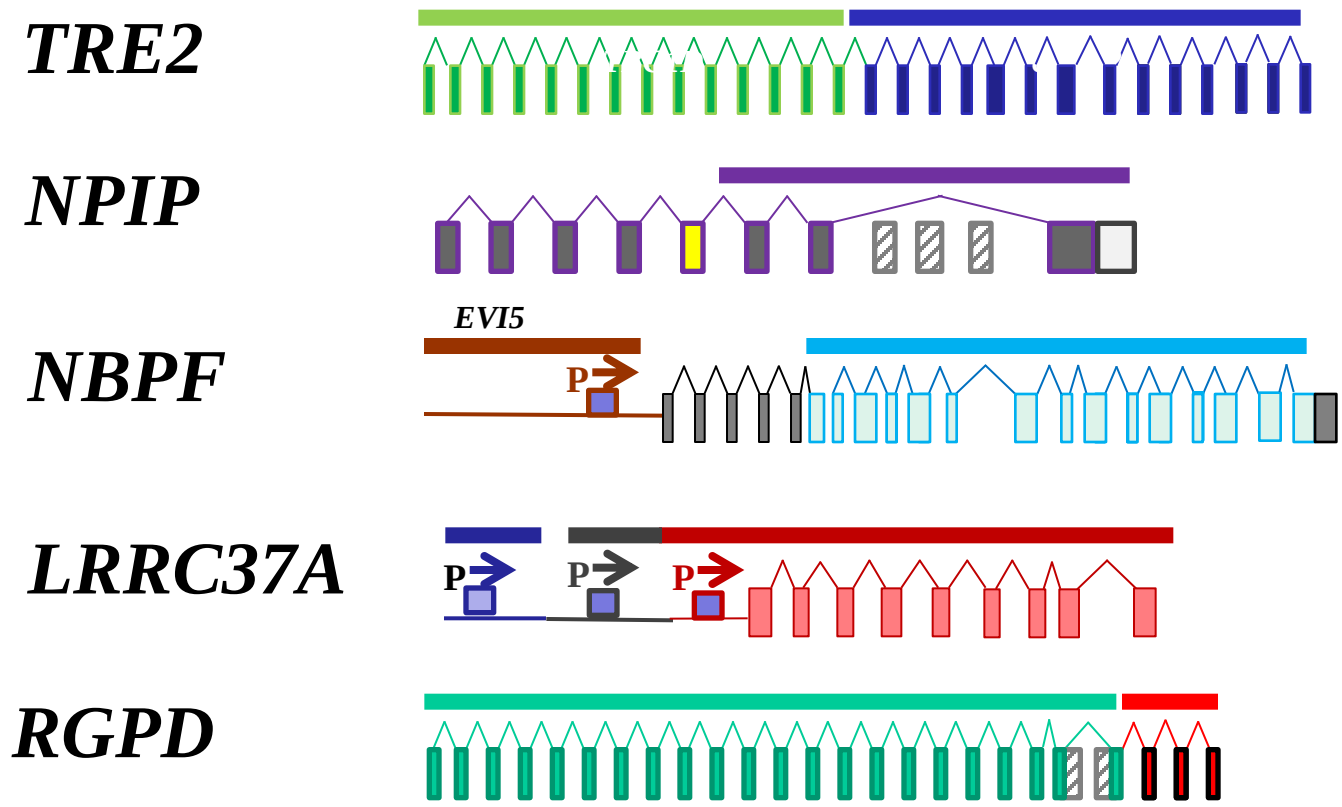


- **The burst of segmental duplications 8-12 mya corresponds to core-associated duplications which have occurred on six human chromosomes (chromosomes 1,2, 7, 15, 16, 17)**

- **Most of the recurrent genomic disorders associated with developmental delay, epilepsy, intellectual disability, etc. are mediated by duplication blocks centered on a core.**

Jiang et al, *Nat. Genet.*, 2007

Human/Great-ape "Core Duplicons" have led to the emergence of new genes



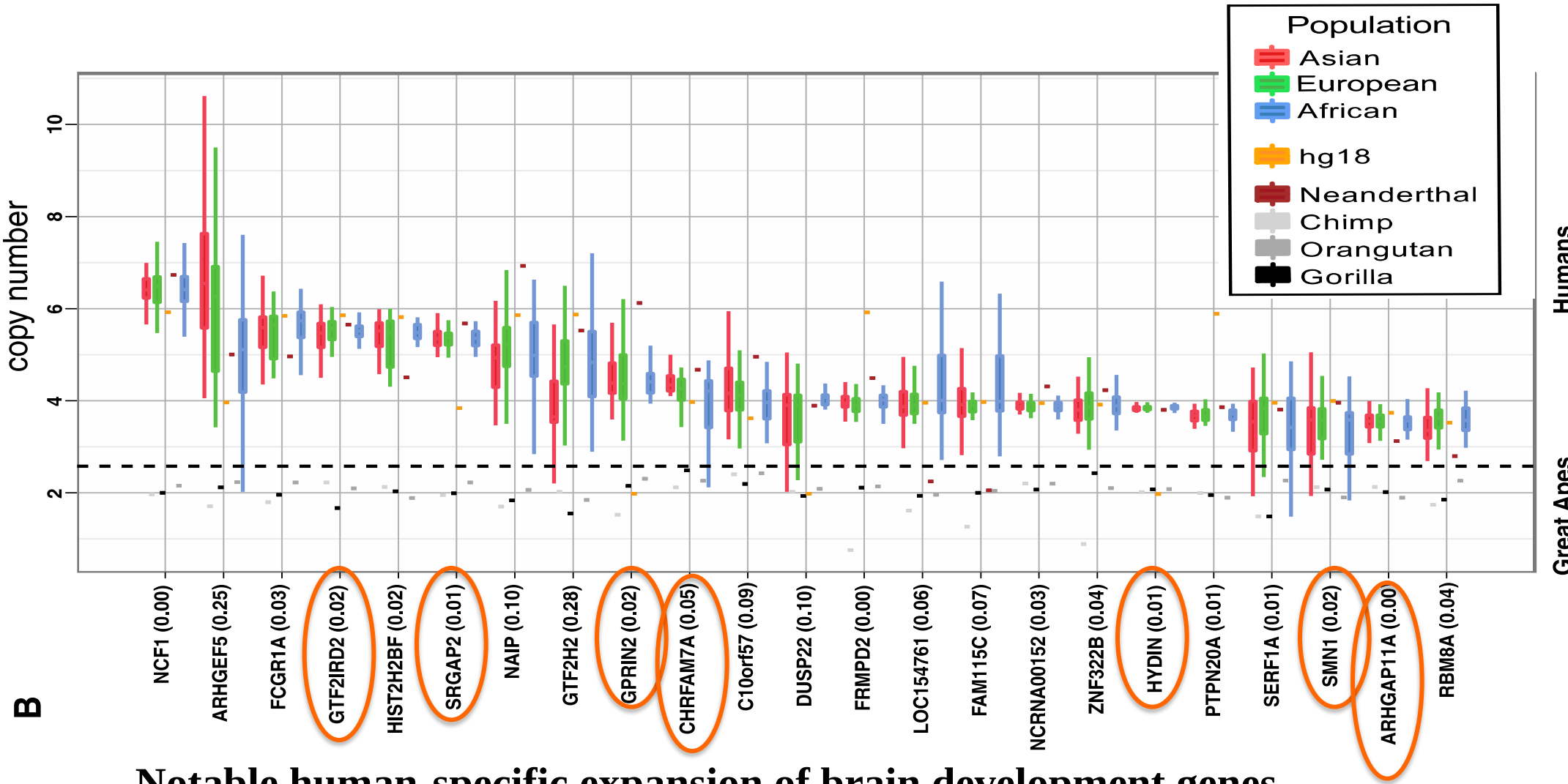
Features: No orthologs in mouse; multiple copies in chimp & human dramatic changes in expression profile; signatures of positive selection

Core Duplicon Hypothesis

The selective disadvantage of interspersed duplications is offset by the benefit of evolutionary plasticity and the emergence of new genes with new functions associated with core duplicons.

Marques-Bonet and Eichler, CSHL *Quant Biol*, 2008

Human-specific gene family expansions

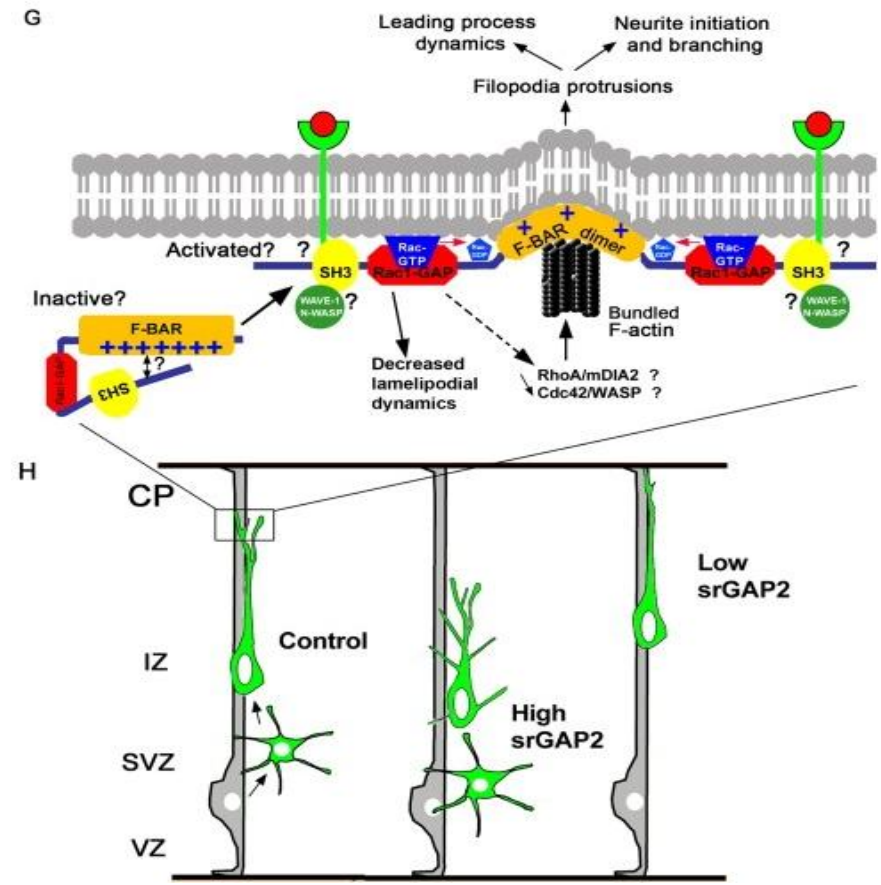


Notable human-specific expansion of brain development genes.
Neuronal cell death: p=5.7e-4; Neurological disease: p=4.6e-2

Sudmant et al., *Science*, 2010

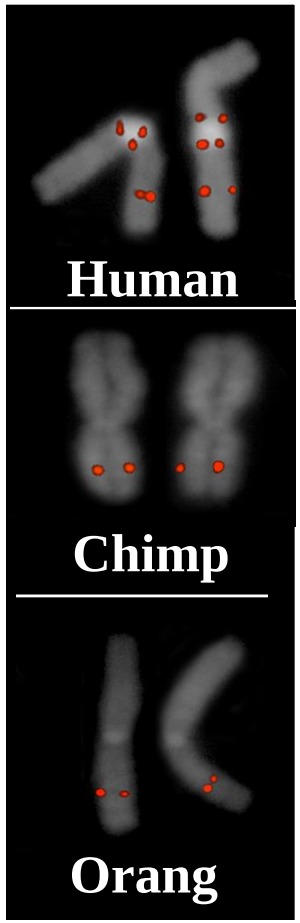
SRGAP2 function

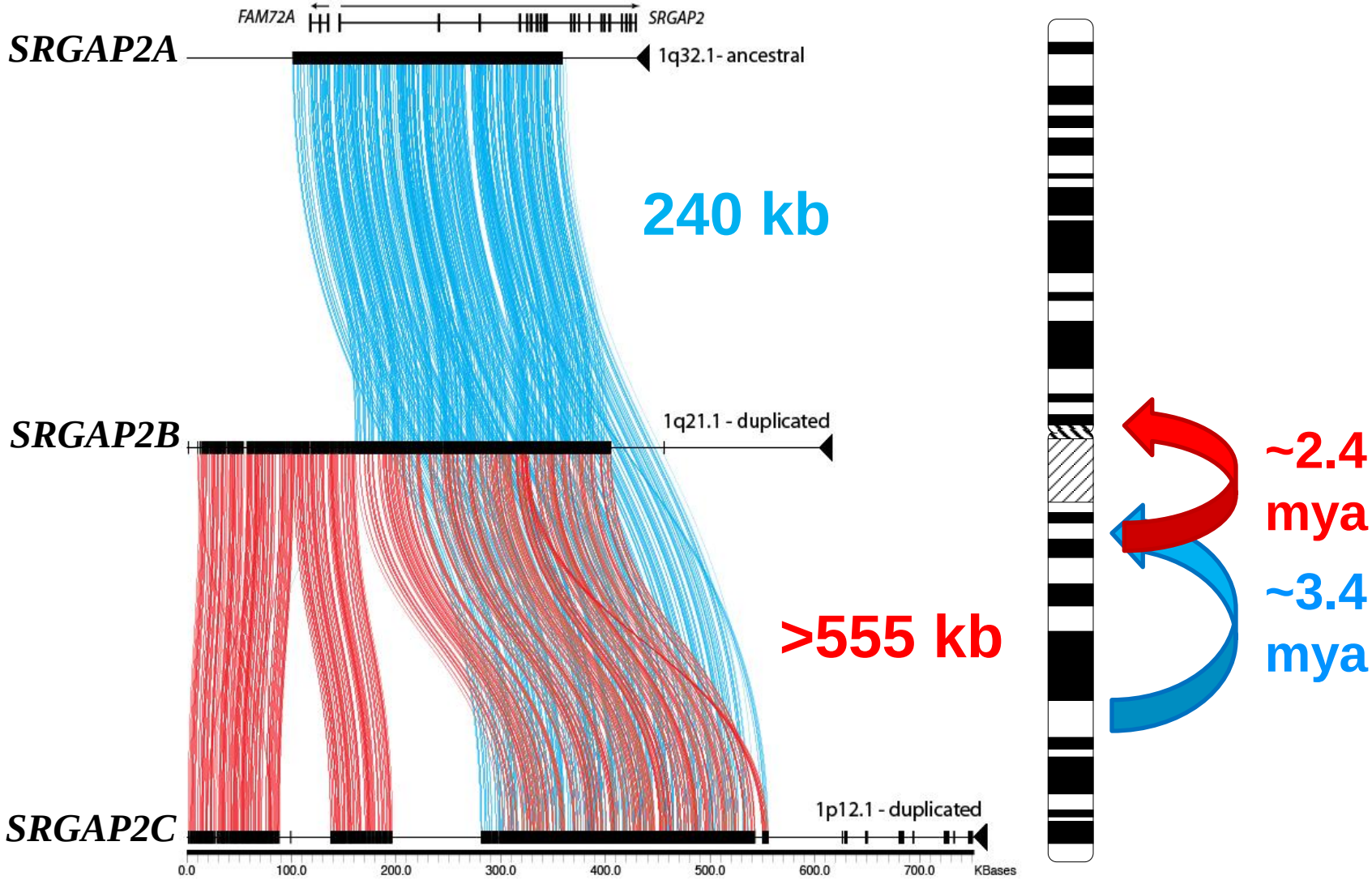
- *SRGAP2* (**SLIT**-ROBO Rho GTPase activating protein 2) functions to control migration of neurons and dendritic formation in the cortex
- Gene has been duplicated three times in human and no other mammalian lineage
- Duplicated loci not in human genome



Guerrier et al., *Cell*, 2009

SRGAP2 Human Specific Duplication

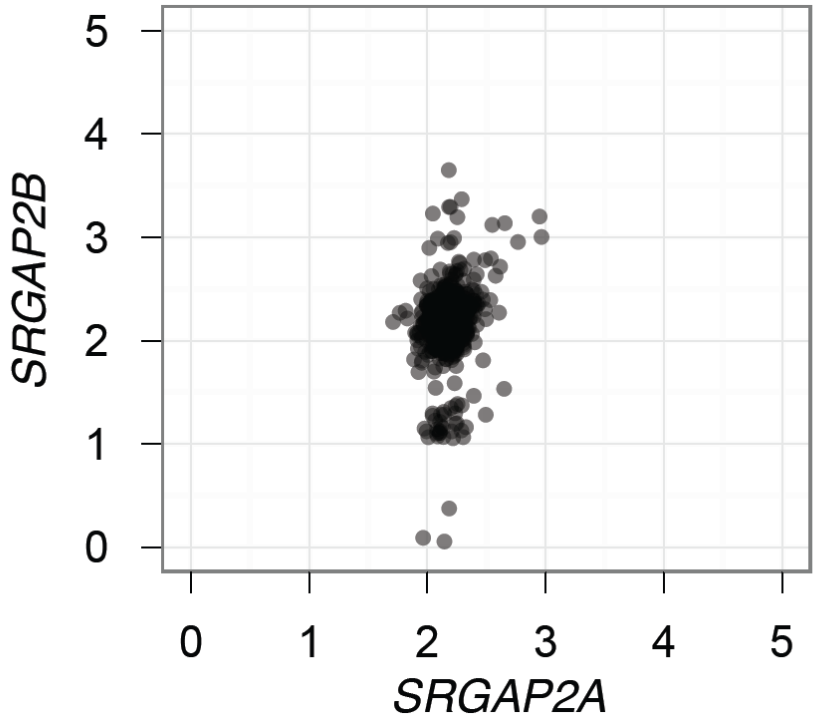


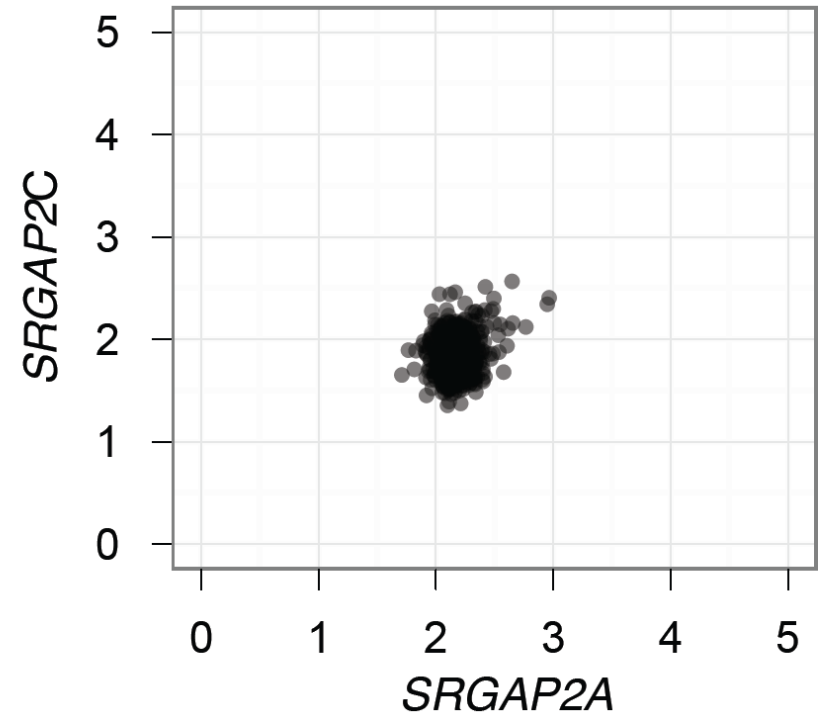


Dennis, Nuttle et al., *Cell*, 2012

SRGAP2C is fixed in humans

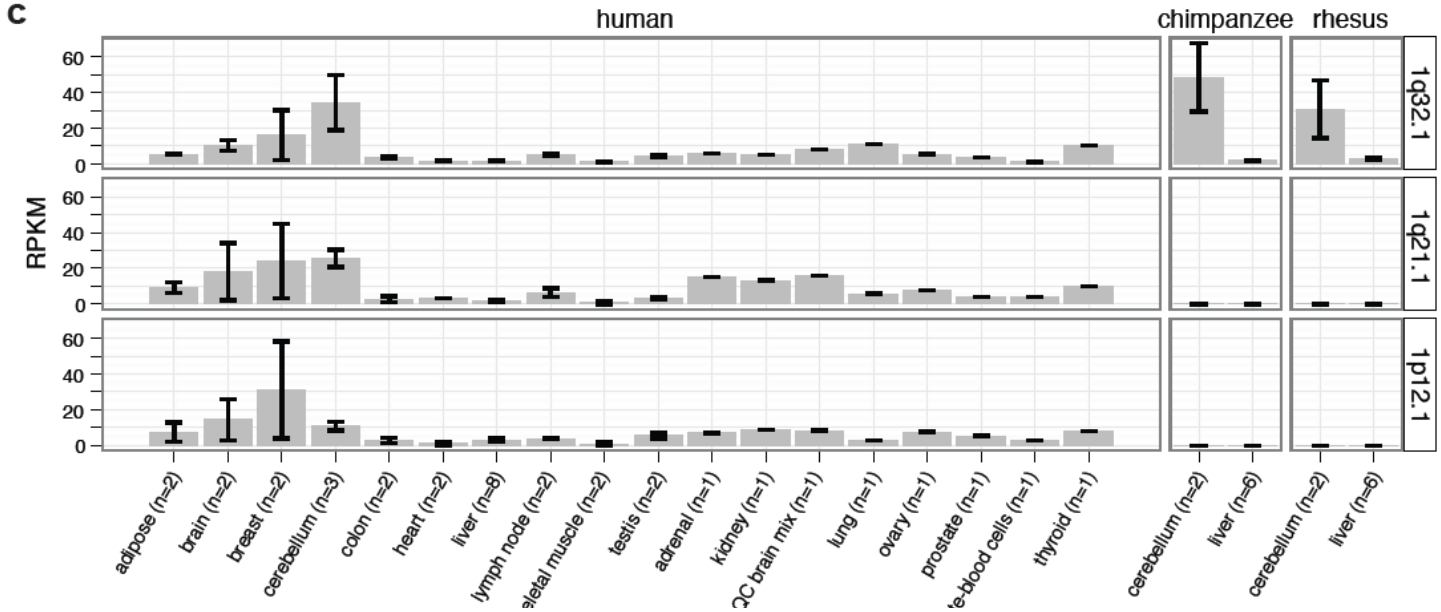
(n=661 individual genomes)



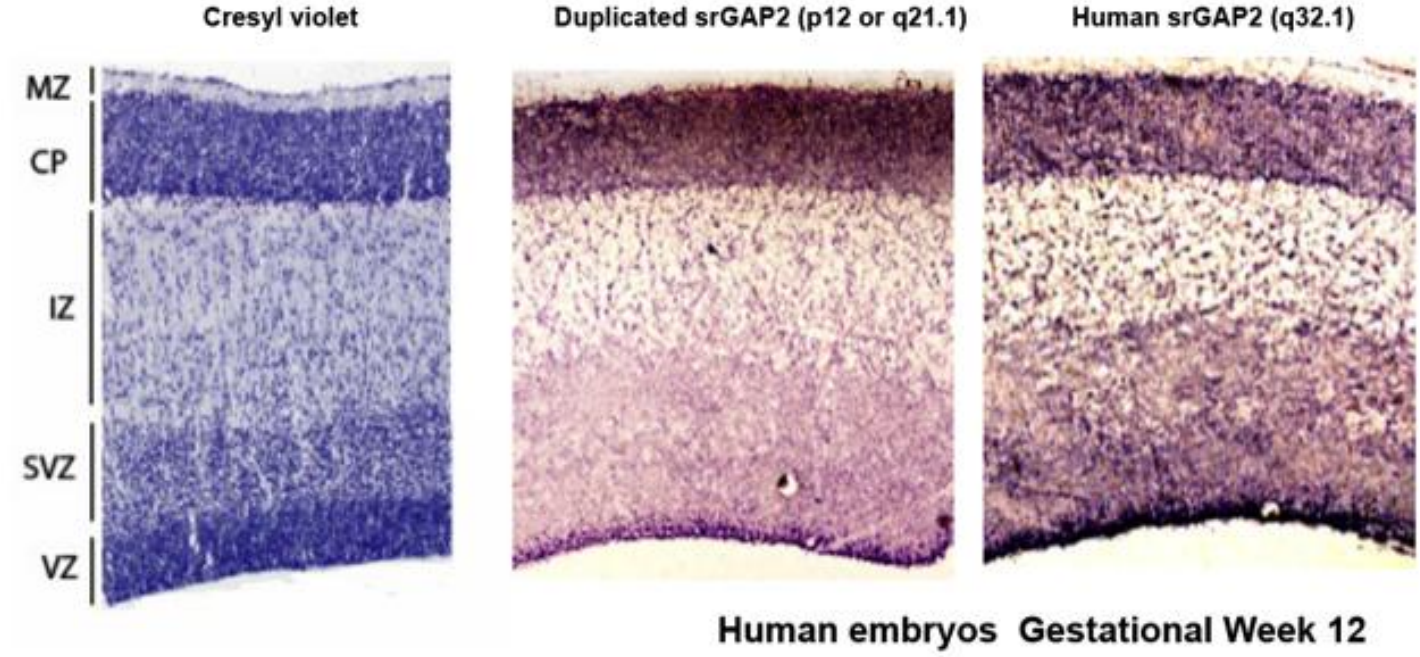


SRGAP2 duplicates are expressed

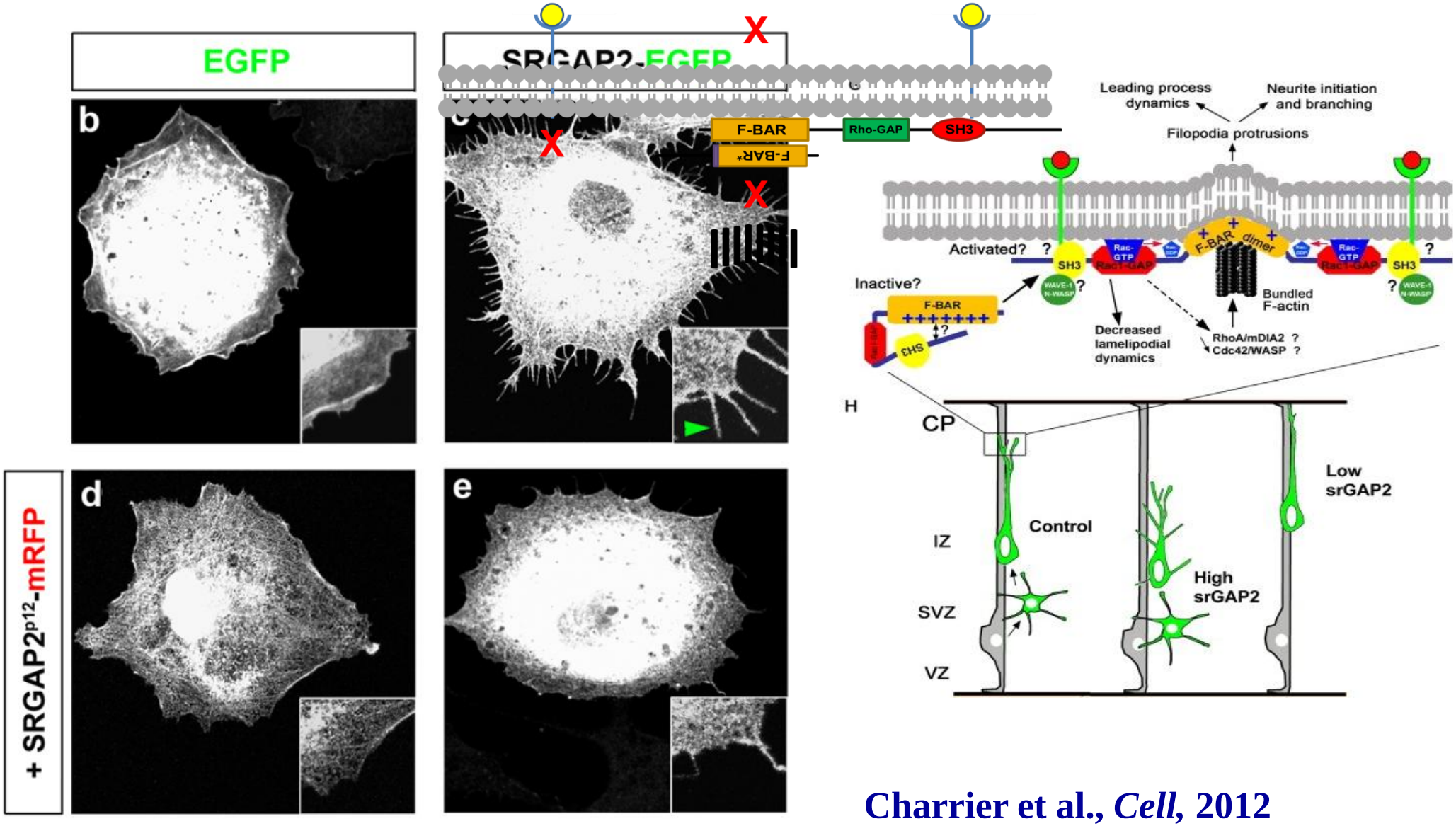
RNAseq



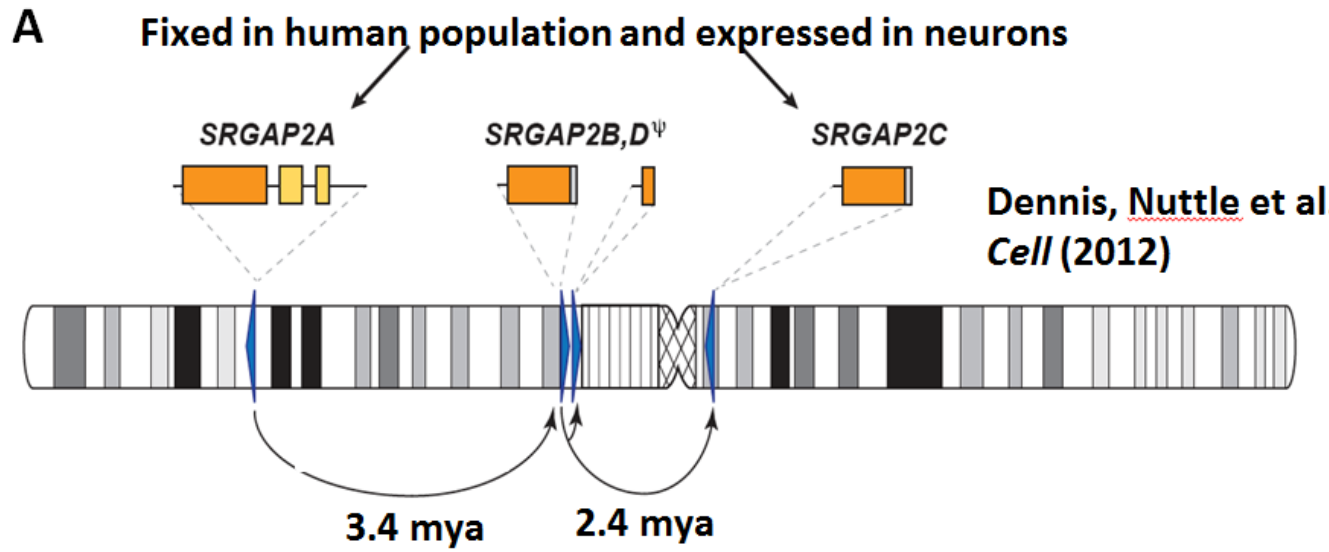
In situ



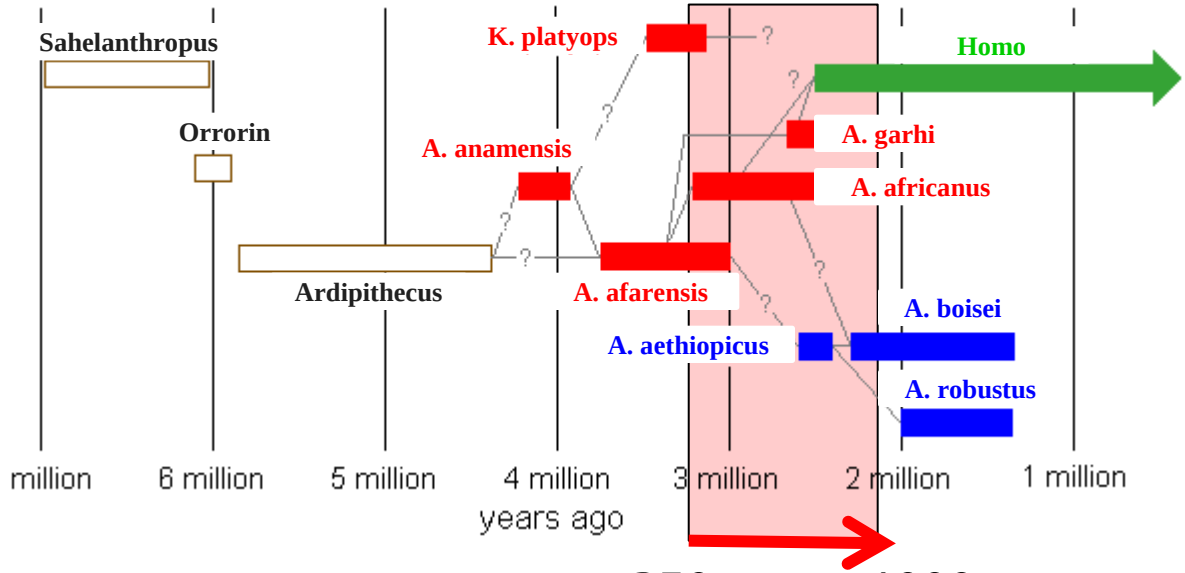
SRGAP2C duplicate antagonizes function



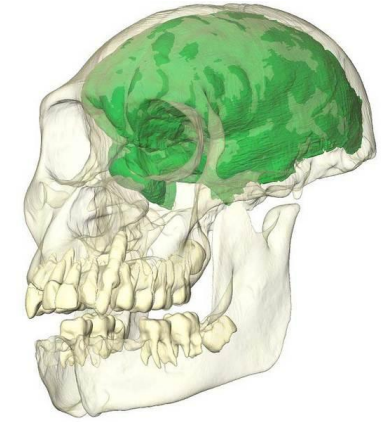
Charrier et al., *Cell*, 2012



Australopithecus



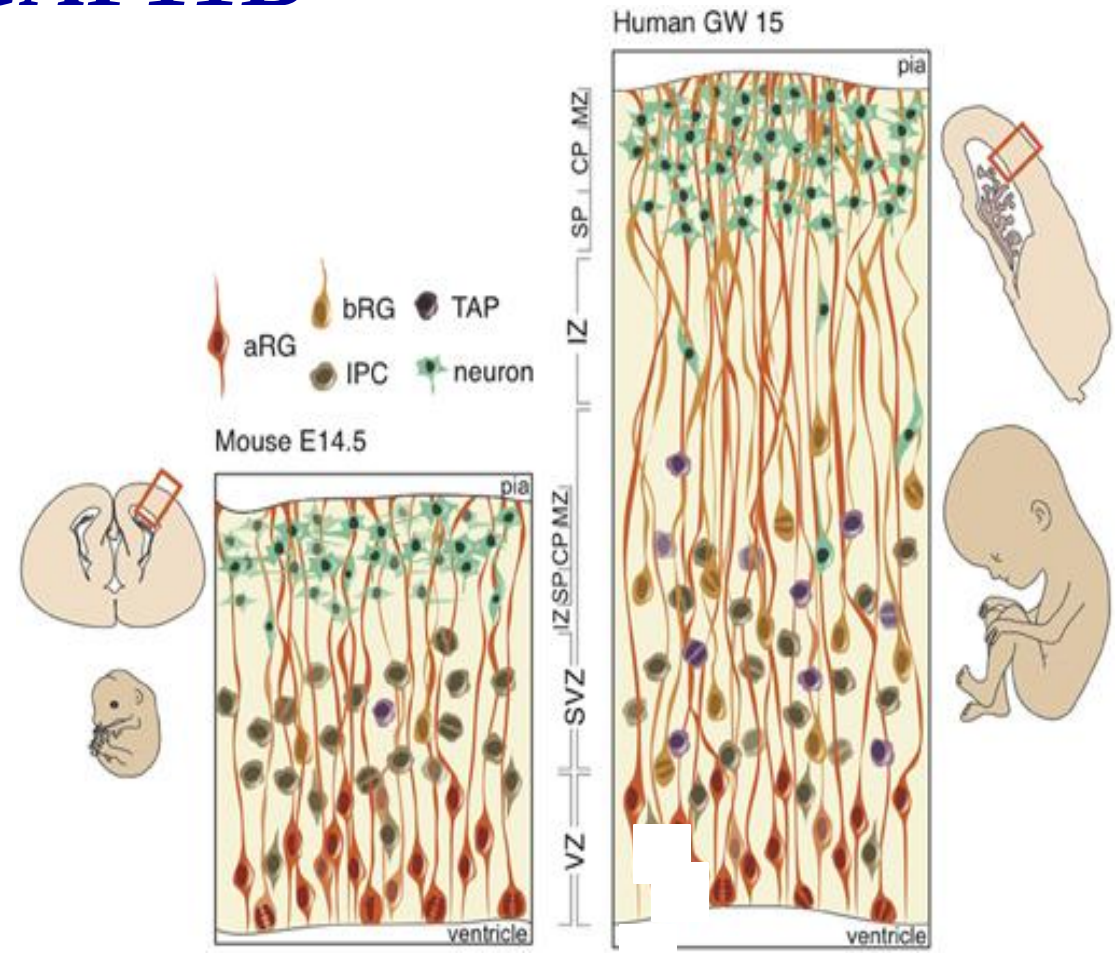
~350 cc ~1000 cc



Homo habilis

Example 2: Human-specific Duplication of *ARHGAP11B*

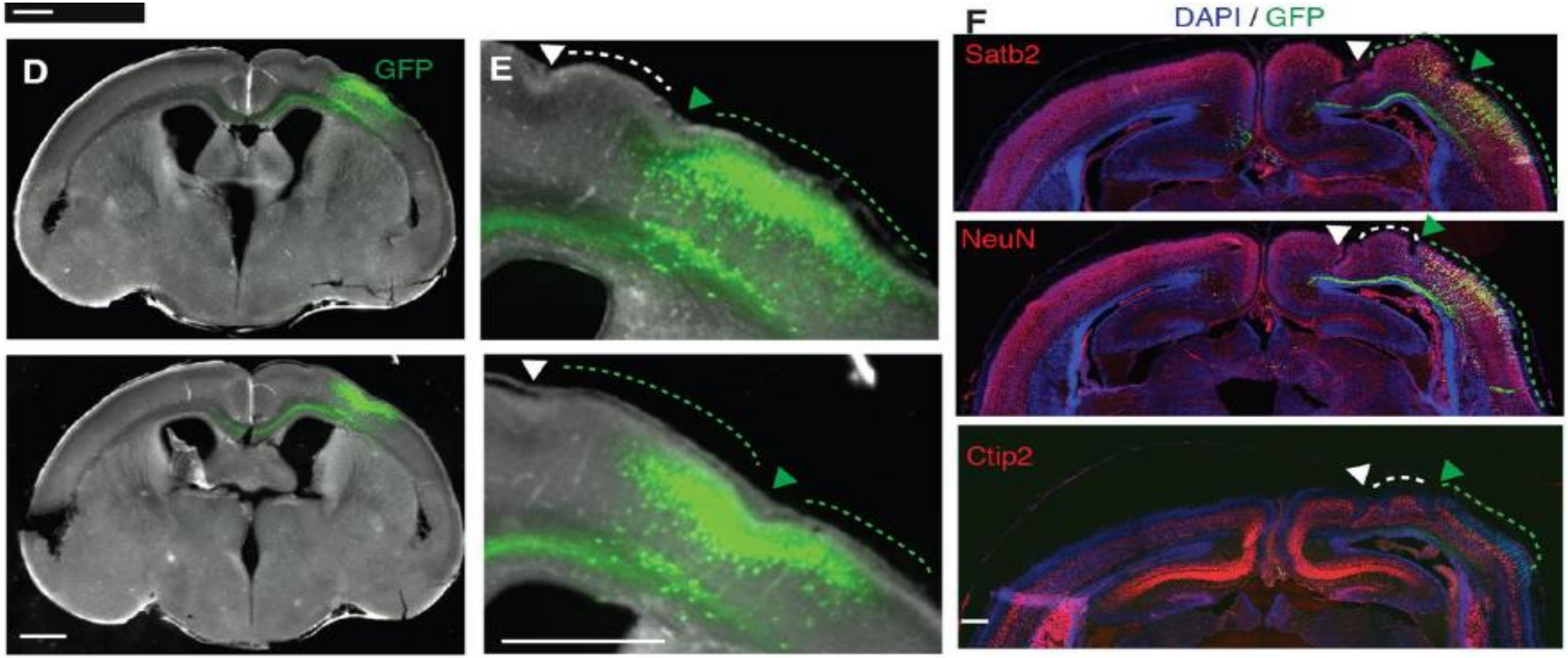
- Hypothesis: increase in number of basal radial glial cells or prolonged proliferation may lead to enlargement of the subventricular zone in humans
- Search for genes that are dramatically increased in concentration in basal radial glial cells as compared to neurons during development
- Only one gene of 56 not present in mouse *ARHGAP11B*



Florea *et al.*, *Science* 2015, Antonacci *et al.*, *Nat. Genet.*, 2014

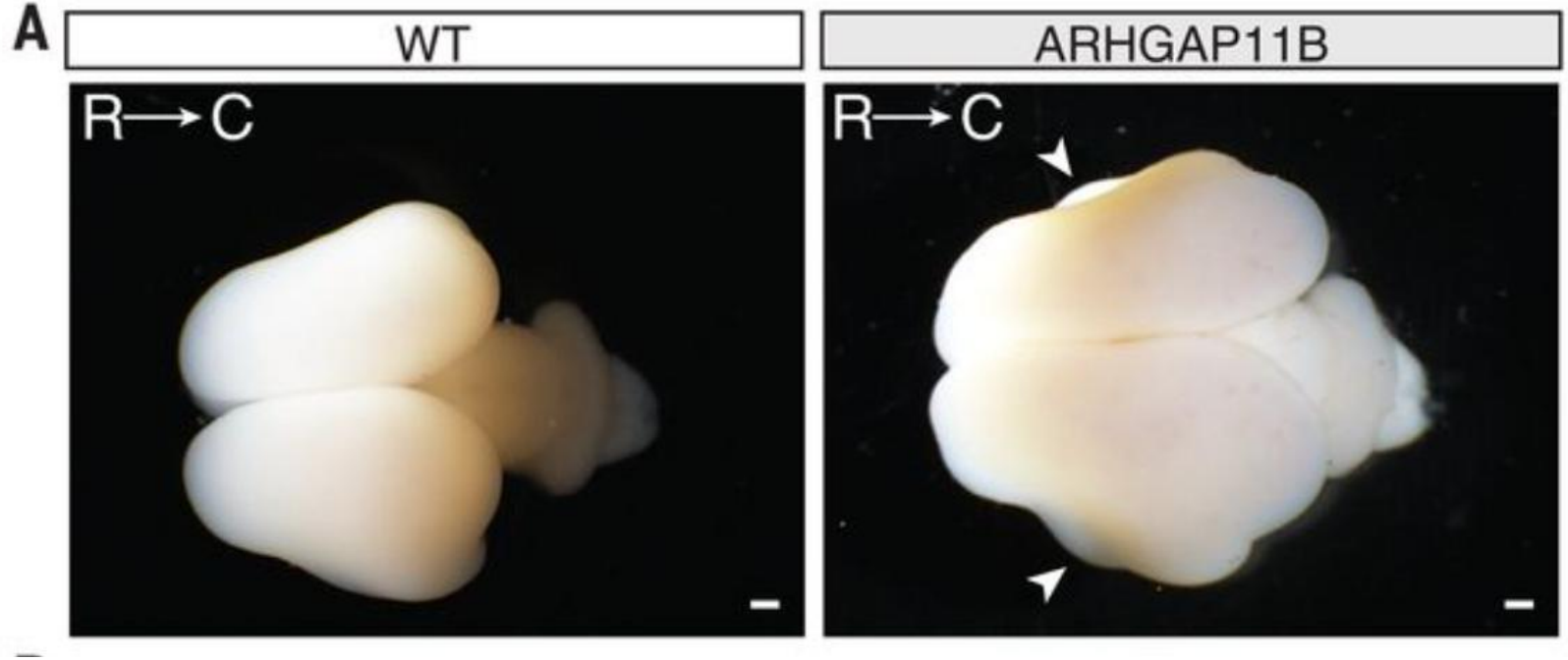
ARHGAP11B induced gyrification of mouse brain

- E13.5 microinjection of *ARHGAP11B* induced folding in the neocortex by E18.5 in ½ of the cases– a significant increase in cortical area.



Florea *et al.*, *Science* 2015

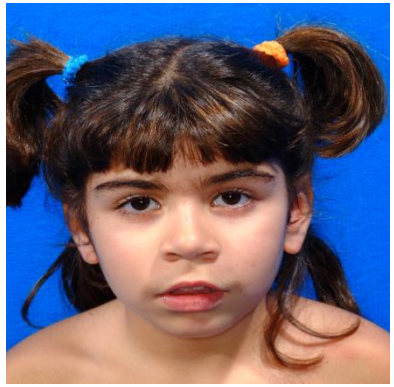
Transgenic human-specific duplicate *ARHGAP11B:* Marmoset fetal brain with human promoter



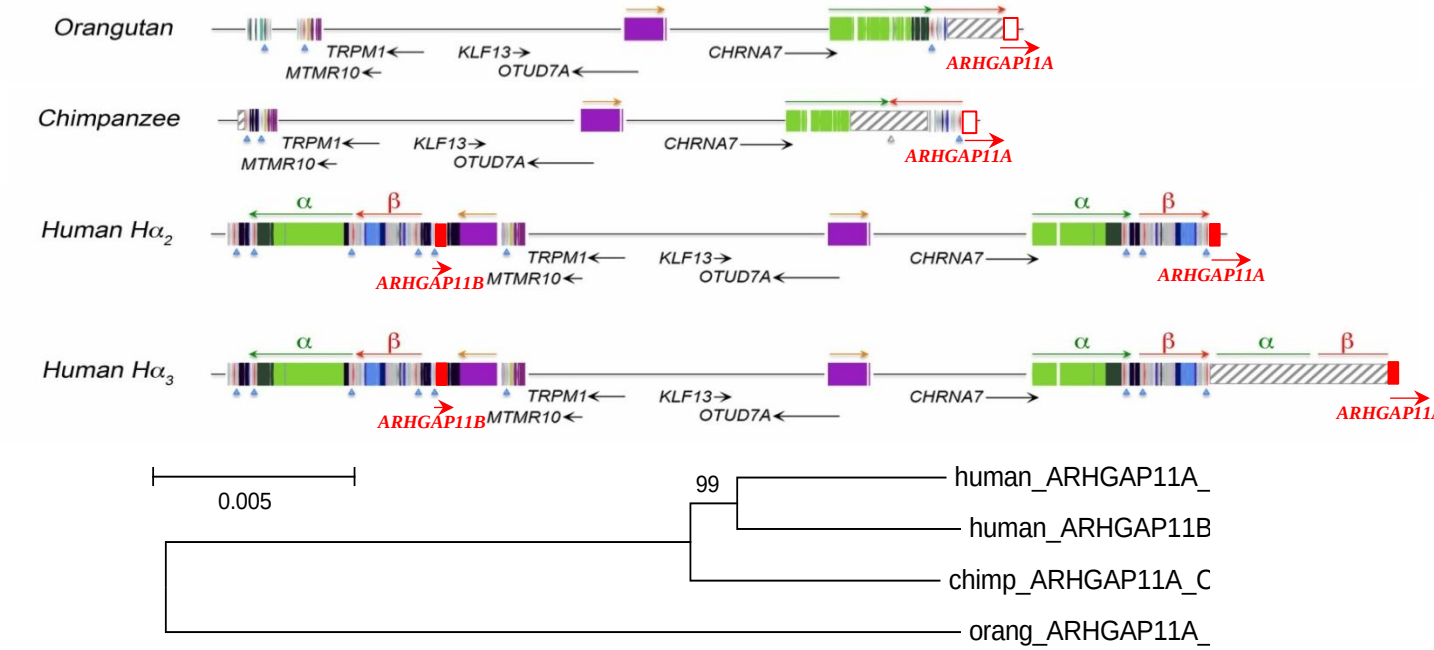
WT brain and brain expressing *ARHGAP11B* in neocortex (TG3). Arrowheads indicate cortical folds. R, rostral; C, caudal. Scale bars, 1 mm

- **Increased the numbers of basal radial glia progenitors in the marmoset outer subventricular zone, increased the numbers of upper-layer neurons, enlarged the neocortex, and induced its folding.**

Duplication of *ARHGAP11B* and 15q13.3 Syndrome



Chromosome 15

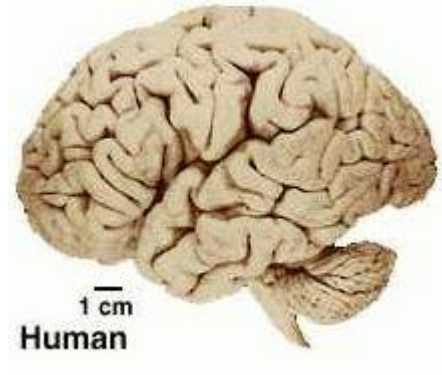


Duplication from*ARHGAP11A* to *ARHGAP11B* estimated to have occurred 5.3 +/- 0.5 million years ago.

Antonacci et al., *Nat Genet,* 2014,

Human-specific duplicated gene innovations and brain development

- *SRGAP2C*— 3.2 mya—produces a truncated protein that heterodimerizes with the parental product and alters neuronal migration, dendritic morphology and density of synapses (Dennis *et al*., *Cell*, 2012; Charrier *et al*., *Cell*, 2012).
- *ARHGAP11B*— truncated duplicate is expressed in basal radial glial cells appears to expand neuronal count and expand subventricular zone (Antonacci *et al*., *Nat Genet*, 2014: Florio *et al*., *Science*, 2015,).
- *BOLA2B*--- *(*256 kya) duplication of gene family specifically at root of Homo sapiens, rapid fixation and largest difference between Neandertals and human genomes and is important in iron homeostasis (Nuttle *et al*., *Nature*, 2016, Gianuzzi *et al*., *Am J Hum Genet* 2019).
- *NOTCH2NL*--- *(*<3 mya) partial duplication expressed in radial glial where interacts with NOTCH2 receptors and delays neuronal progenitor differentiation(Fiddes *et al*., *Cell*, 2018)
- Properties: Nearly fixed for copy number in the human population, predispose to disease instability and the duplications are incomplete with respect to gene structure. **NONE present in original human genome.**



Summary

- Interspersed duplication architecture sensitized our genome to copy-number variation increasing our species predisposition to disease—children with autism and intellectual disability
- Duplication architecture has evolved recently in a punctuated fashion around core duplicons which encode human great-ape specific gene innovations (eg. *NPIP*, *NBPF*, *LRRC37*, etc.).
- Cores have propagated in a stepwise fashion "transducing" flanking sequences---human-specific acquisitions flanks are associated with brain developmental genes.
- **Core Duplicon Hypothesis**: Selective disadvantage of these interspersed duplications offset by newly minted genes and new locations within our species. Eg. *SRGAP2C*

Overall Summary

- **I. Disease**: Role of CNVs in human disease—relationship of common and rare variants—biased toward interspersed SDs due to NAHR
- **II. Methods**: NGS Read-pair and read-depth methods to characterize SVs—long-read genomes can now be fully phased and assembled achieving complete telomere-to-telomere assembly & complete variation discovery.
- **III**: **Evolution**: Rapid evolution of complex human architecture that predisposes to disease also coupled to human-specific gene innovations that make us uniquely human


Disease
Evolution

Acknowledgements

Glossary

SV-structural variation

SD-segmental duplication

CNV- copy number variation

CNP—copy number polymorphism

NGS—next generation sequencing (eg. Illumina short read)

Indel-insertion/deletion event

SMRT-single-molecule real-time sequencing

CCS—circular consensus sequencing

HiFi-high fidelity long-read

CLR—continuous long-read sequencing

WGS—whole genome shotgun sequencing

ONT—Oxford Nanopore Technology

PacBio—Pacific Biosciences

ZMW-zero-mode wave guide

CDR—centromere dip region

NAHR—non-allelic homologous recombination

SV Software

- *PennCNV* (Kai Wang) and *CNVPartition*—calling CNVs from SNP microarray
- *Genomestrip*—Handsaker/McCarroll—combines read-depth and readpair data to identify potential sites of SV data from population genomic data; *dCGH*—Sudmant/Eichler—measure Illumina read-depth using multi-read sequence mapper (mrsFAST/mrFAST) ; *Delly*—EMBL Rausch/Korbel—uses split-read and readpair signatures; *Lumpy* --Quinlan/Hall—uses probabilistic framework to integrate multiple SV such as discordant paired-end alignments and split-read alignments; *GATK-SV*—Talkowski—integrates multiple short reads signatures; *Manta*—Illumina split and paried-end reads followed by assembly
- *Conifer /XHMM*— Krumm/Eichler & Frommer/Purcell-exome CNV calling
- *PBSV*—Aaron Wenger (PacificBiosciences software) signatures from pbmm2 alignments; *SNIFFLES2*—Sedlaczeck/Schatz– NGLMR mapping of PacBio or ONT data using split-read alignments, high-mismatch regions, and coverage
- *PAV*—Audano/Eichler & *SVIM-asm*—Heller/Vingron--assembly-to-assembly based discovery of SVs using minimap and LRSassembled genomes
- *Verkko*—Koren/ Philippy & *HiFiasm*- Cheng/Li—graph based approaches to generate near T2T assemblies using UL-ONT and HiFi sequencing data
- *Saffire-SV*, *StainedGlass* & *SVbyEye*- (Vollger/ Porubsky/Eichler)—visualization tools to characterize chromosomal level SV and centromeric satellite DNA

SD-Mediated Rearrangements