

Genome Structural Variation

Evan Eichler

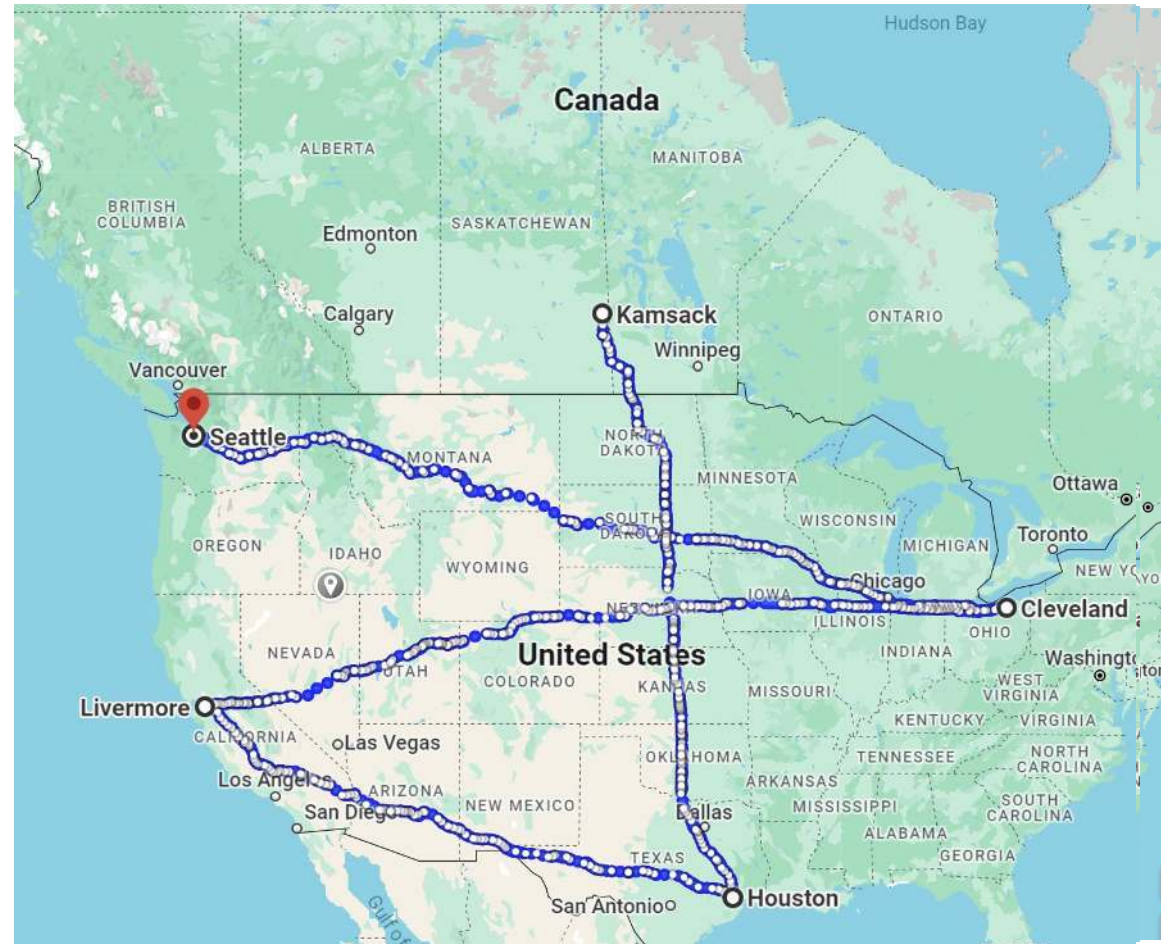
Howard Hughes Medical Institute

University of Washington

January 19th, 2026, 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

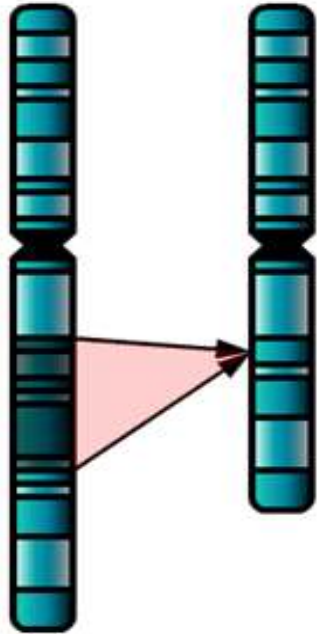
Types

- 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.

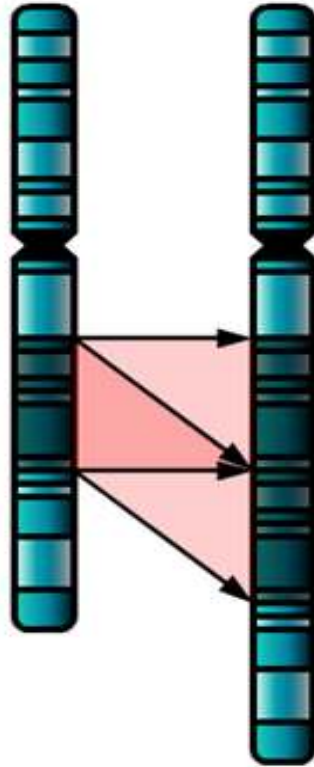
Sequence

Cytogenetics

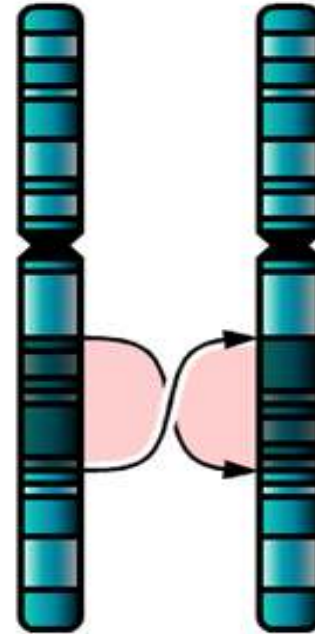
Genome Structural Variation



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. Aitman¹, Rong Dong^{1*}, Timothy J. Vyse^{2*}, Penny J. Norsworthy^{1*}, Michelle D. Johnson¹, Jennifer Smith³, Jonathan Mangion¹, Cheri Robertson-Lowe^{1,2}, Amy J. Marshall¹, Enrico Petretto¹, Matthew D. Hodges¹, Gurjeet Bhargal³, Sheetal G. Patel¹, Kelly Sheehan-Rooney¹, Mark Duda^{1,3}, Paul R. Cook^{1,3}, David J. Evans³, Jan Domin³, Jonathan Flint⁴, Joseph J. Boyle⁵, Charles D. Pusey³ & H. Terence Cook⁵ [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§}

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](#)

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

Andrew J Sharp¹, Sierra Hansen¹, Rebecca R Selzer², Ze Cheng¹, Regina Regan³, Jane A Hurst⁴, Helen Stewart⁴, Sue M Price⁴, Edward Blair⁴, Raoul C Hennekam^{5,6}, Carrie A Fitzpatrick⁷, Rick Segraves⁸, Todd A Richmond², Cheryl Guiver³, Donna G Albertson^{8,9}, Daniel Pinkel⁸, Peggy S Eis², Stuart Schwartz⁷, Samantha J L Knight³ & Evan E Eichler¹ [VOLUME 38 | NUMBER 9 | SEPTEMBER 2006 NATURE GENETICS](#)

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

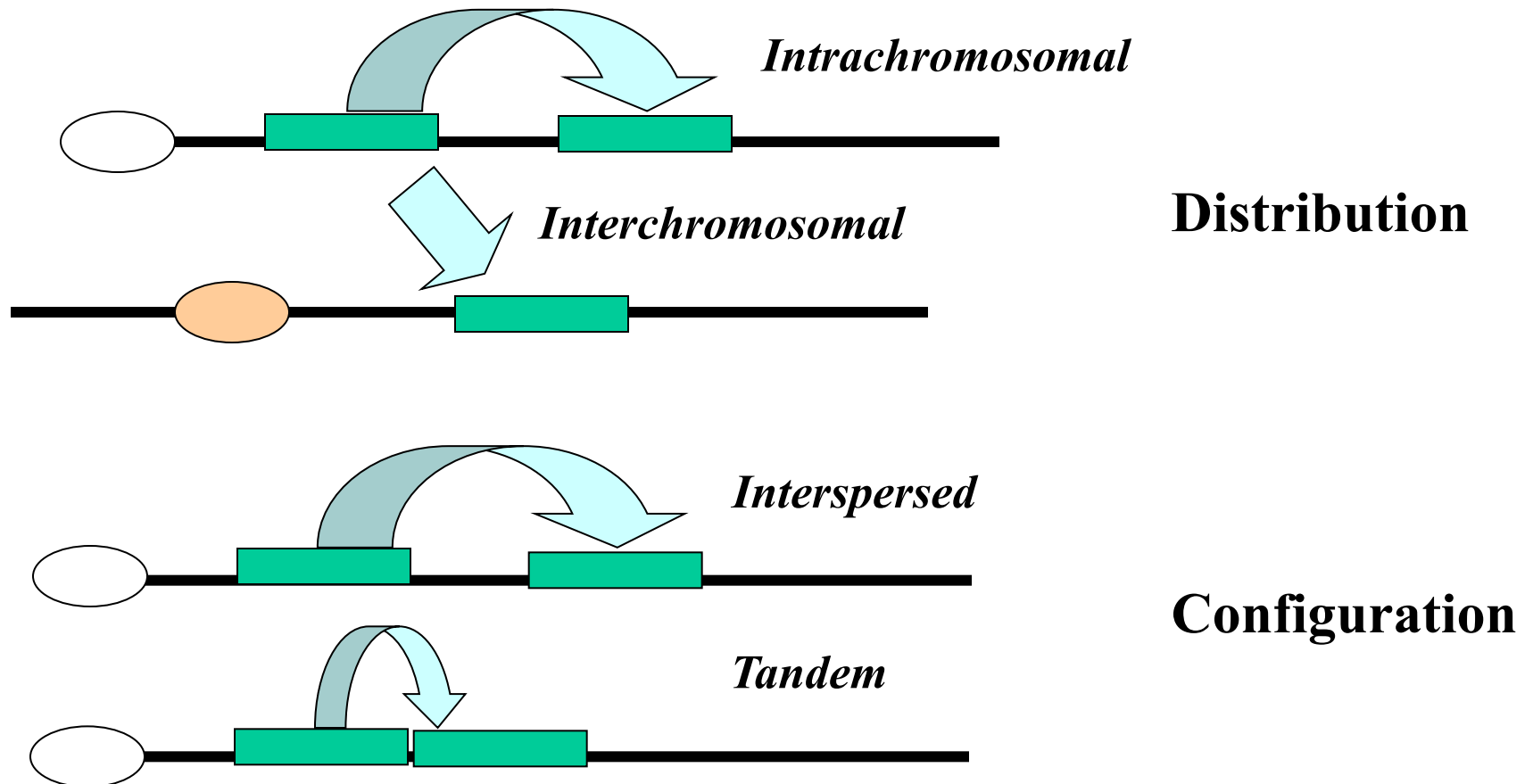
Lauren A. Weiss, Ph.D., Yiping Shen, Ph.D., Joshua M. Korn, B.S., Dan E. Arking, Ph.D., David T. Miller, M.D., Ph.D., Ragnheidur Fossdal, B.Sc., Evald Saemundsen, B.A., Hreinn Stefansson, Ph.D., Manuel A.R. Ferreira, Ph.D., Todd Green, B.S., Orah S. Platt, M.D., Douglas M. Ruderfer, M.S., Christopher 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 Consortium [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 Seungtae 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†} [SCIENCE VOL 316 20 APRIL 2007](#)

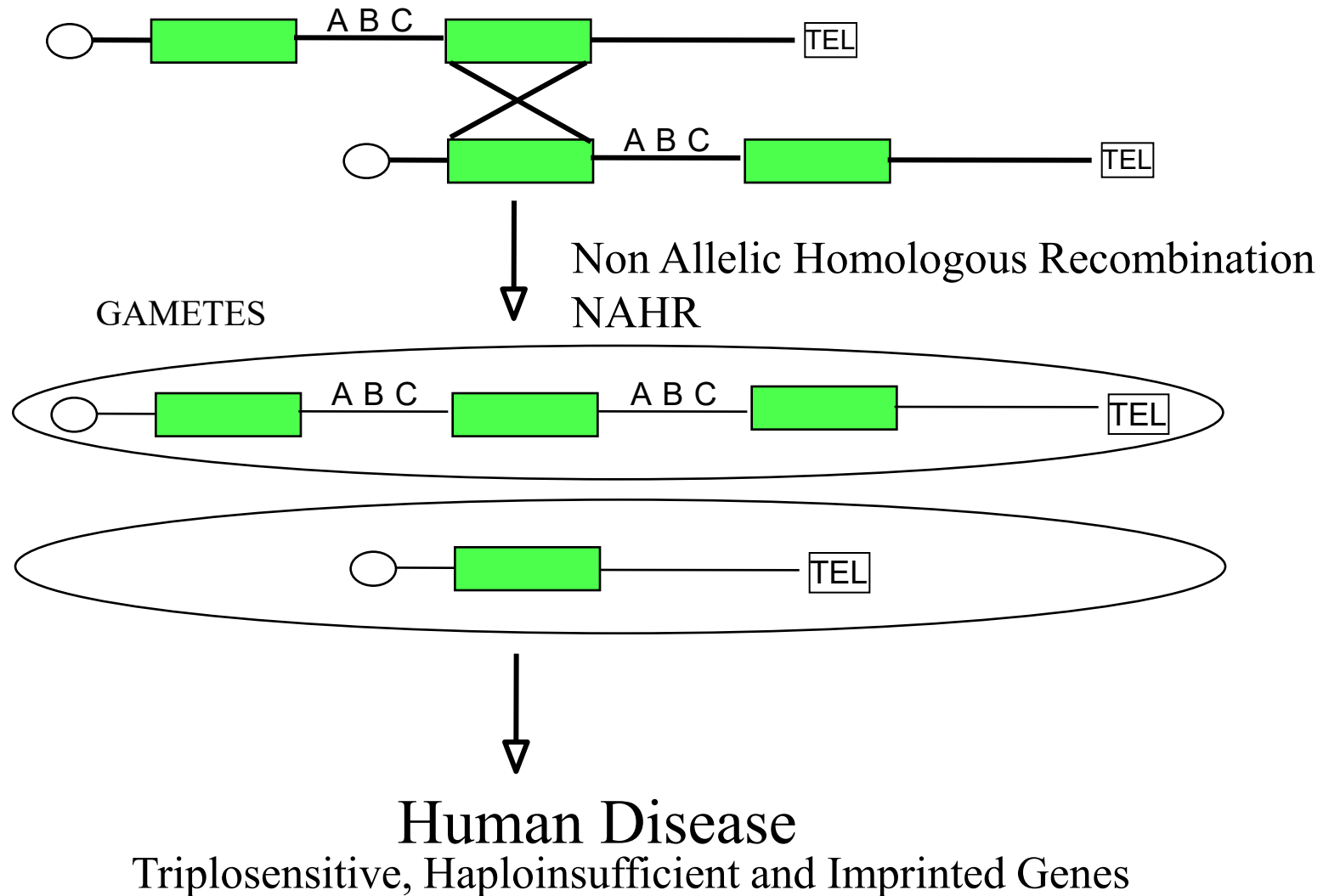
Perspective: Segmental Duplications (SD)

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

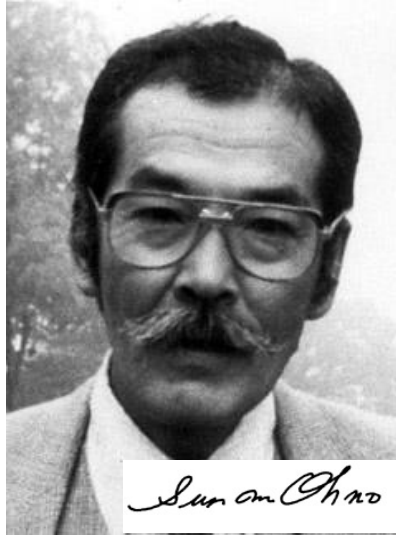


Importance:

SDs promote genome structural variation



Importance: Evolution of New Gene Function



GeneA

Mutation

Maintain old
Function

Duplication



Acquire New/
Modified Function

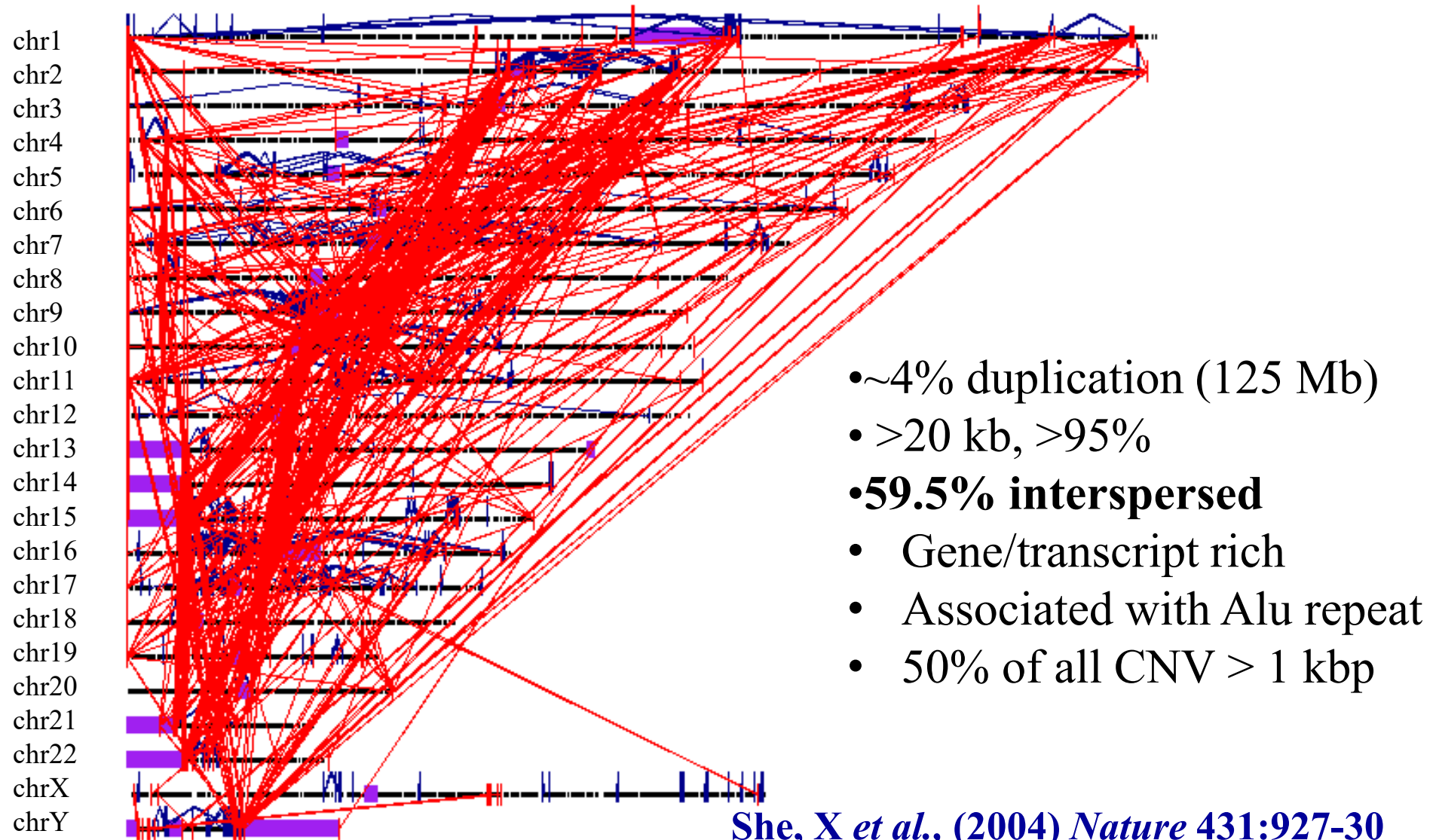
Mutation

GeneA'

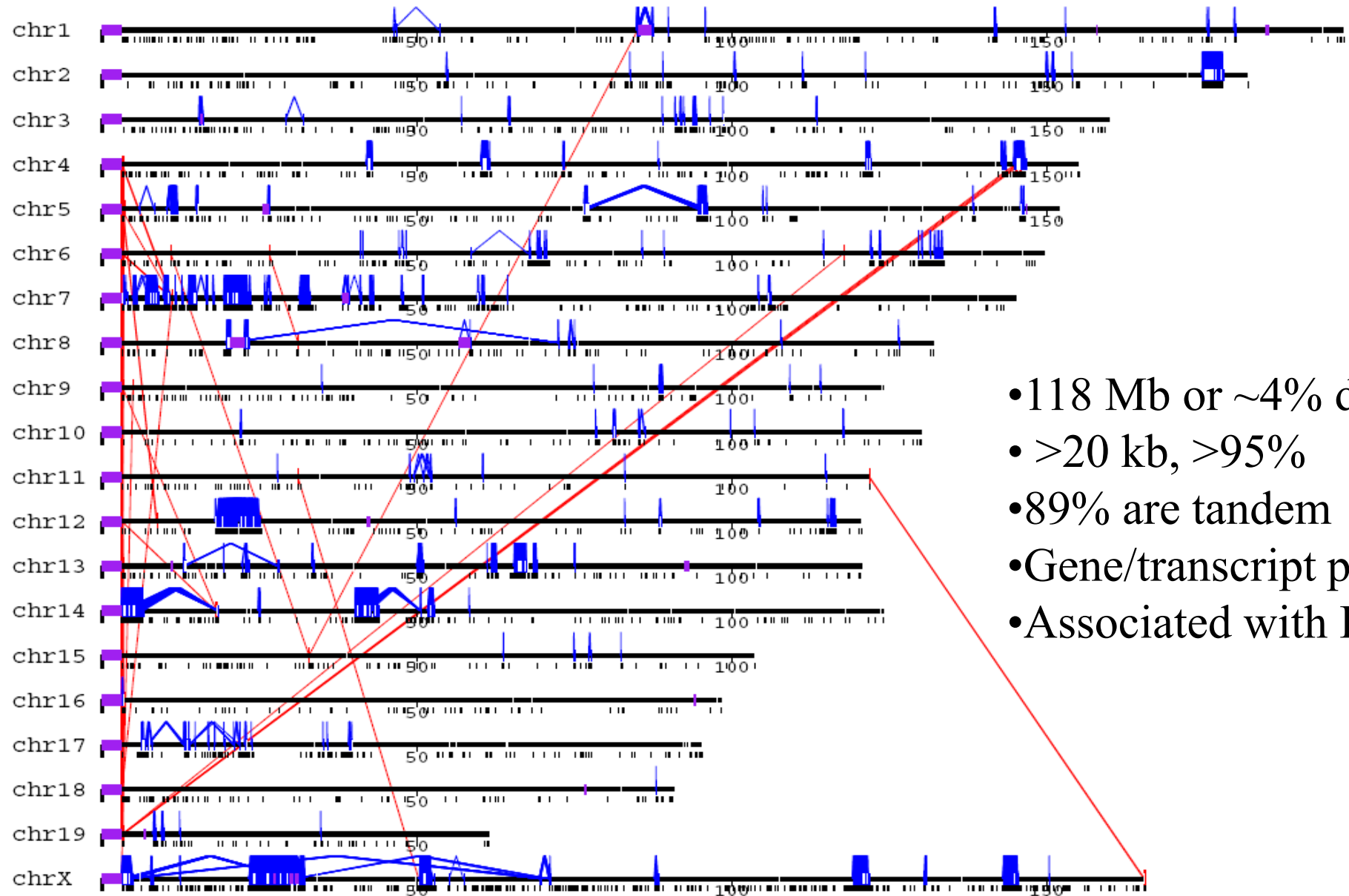
Mutation

Loss of Function

Human Genome Segmental Duplication Pattern



Mouse Segmental Duplication Pattern

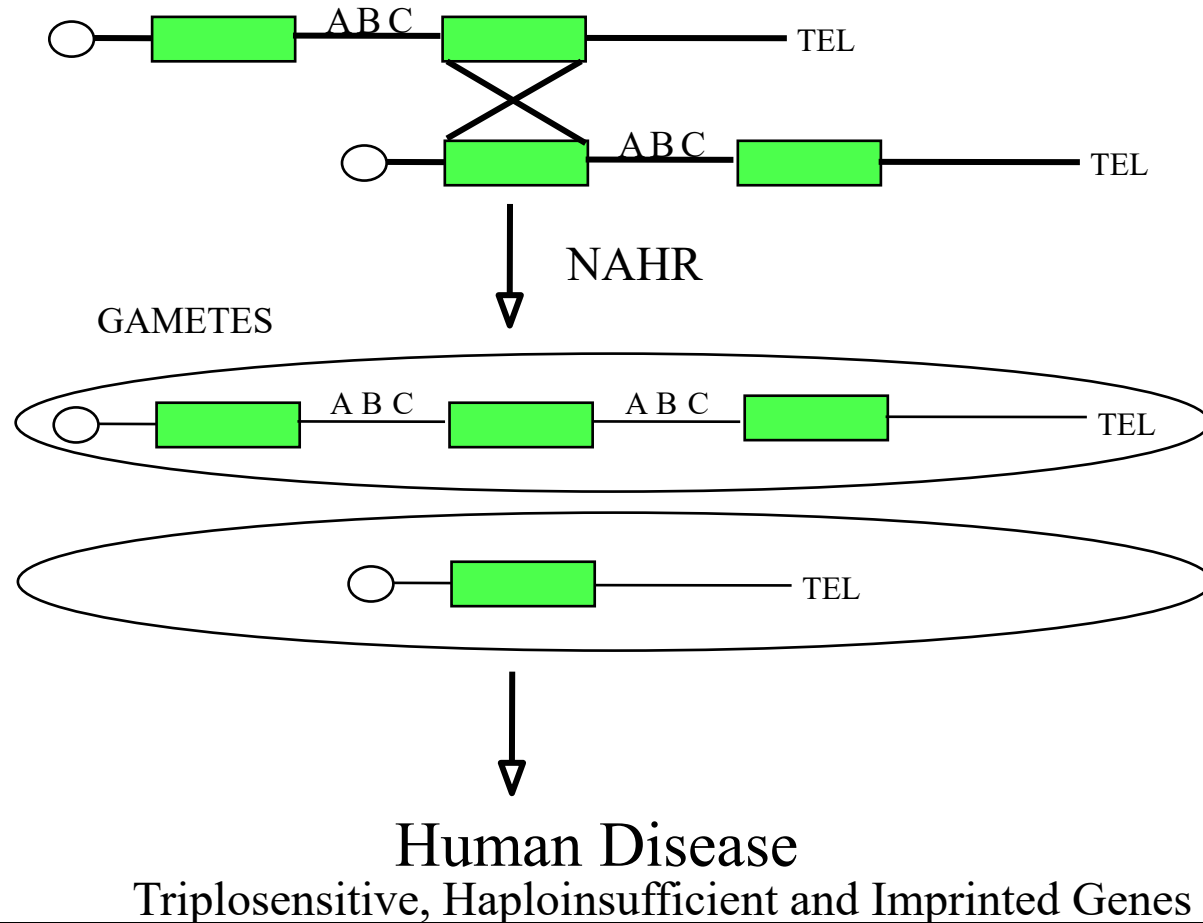


- 118 Mb or ~4% dup
- >20 kb, >95%
- 89% are tandem
- Gene/transcript poor
- Associated with LINEs

Human Segmental Duplications Properties

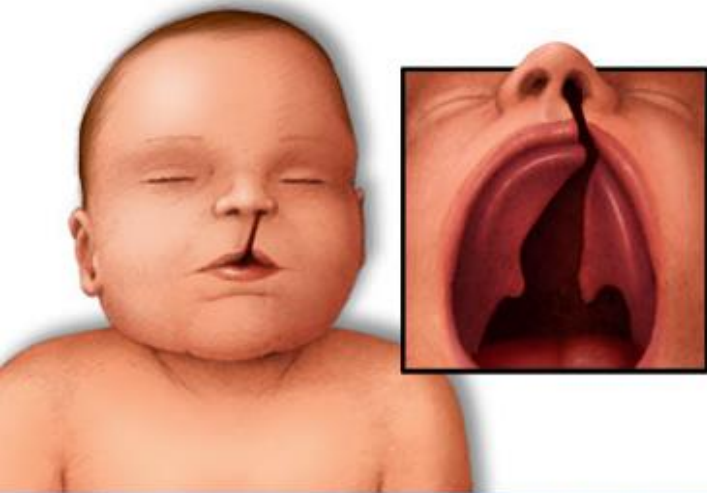
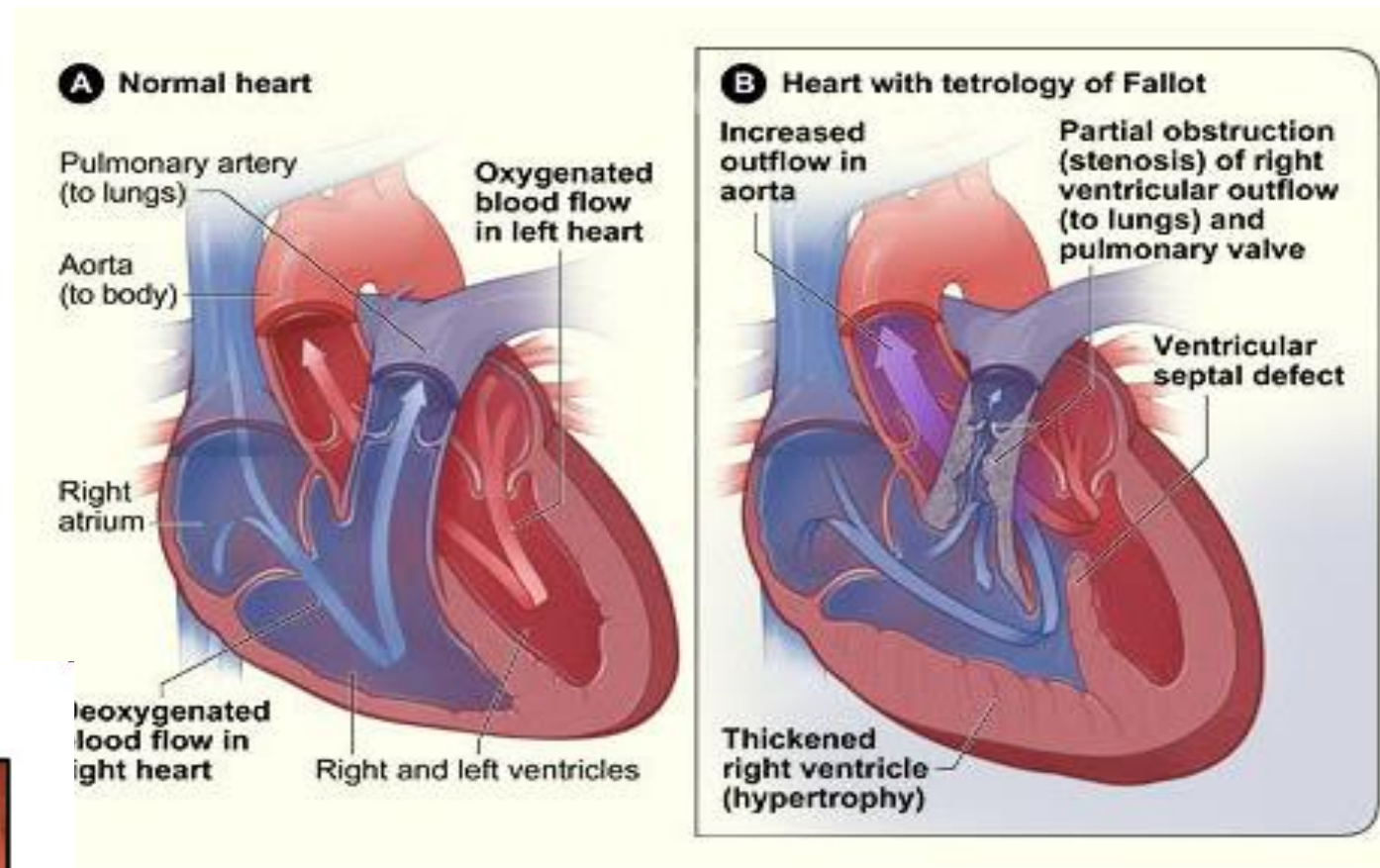
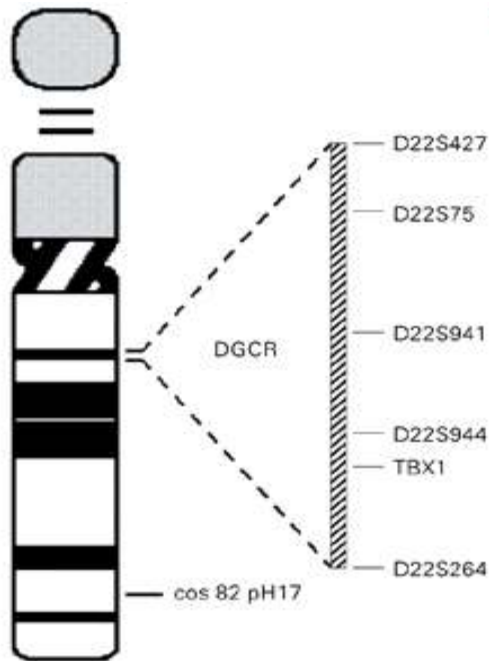
- Large (>10 kb)
- Recent ($>95\%$ identity)
- **Interspersed (60% are separated by more than 1 Mb)**
- Modular in organization
- Difficult to resolve

Rare Structural Variation & Disease

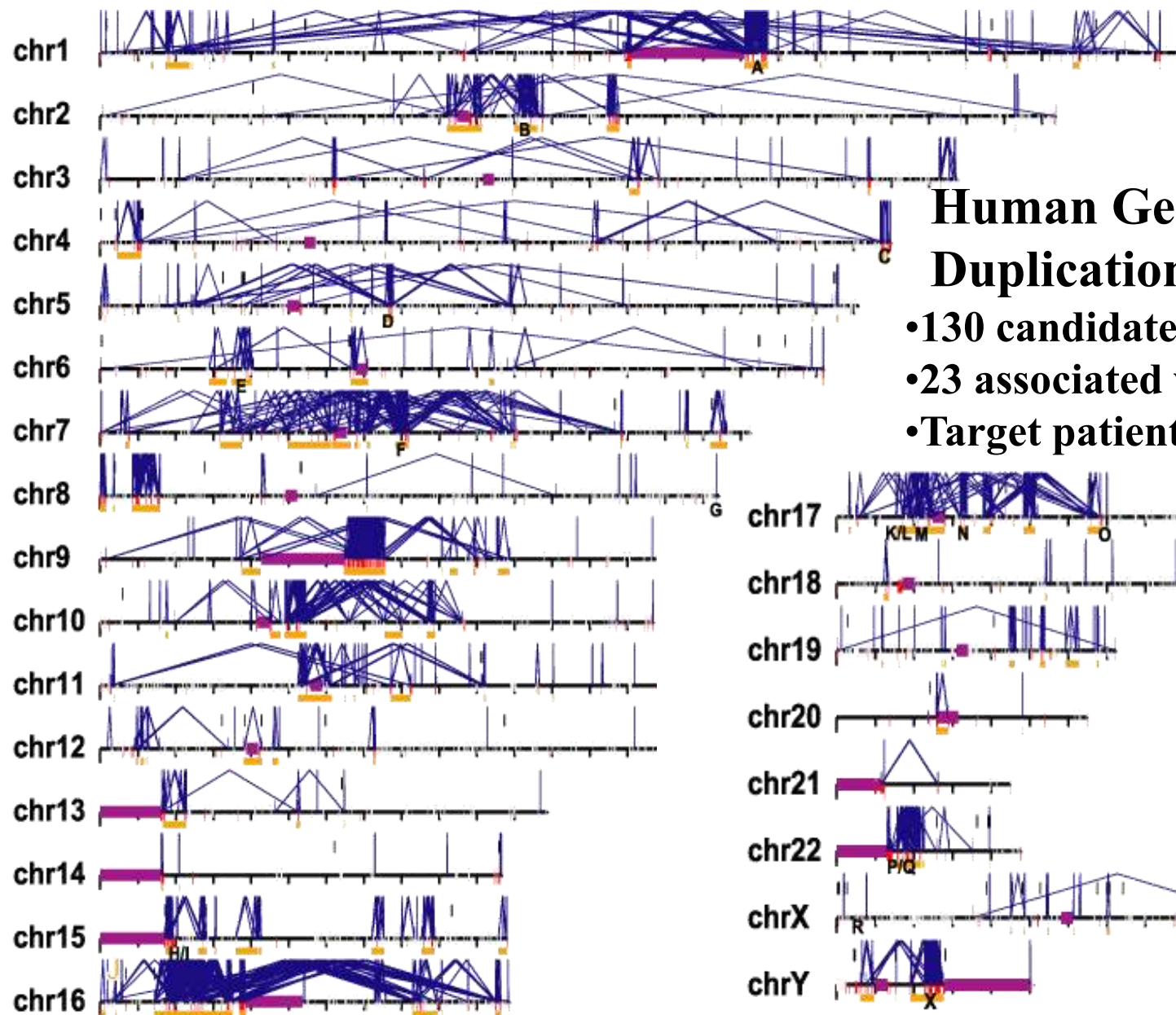


•**Genomic Disorders:** A group of diseases that results from genome rearrangement mediated mostly by non-allelic homologous recombination. (*Inoue & Lupski , 2002*).

DiGeorge/VCF/22q11 Syndrome

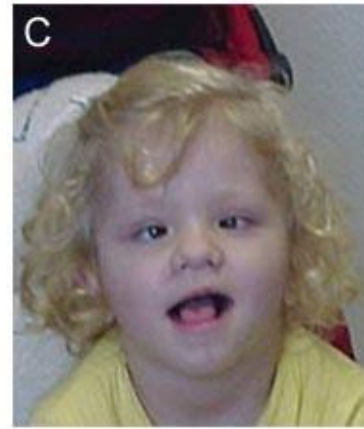


1/3000 live births
180 phenotypes
75-80% are sporadic (not inherited)

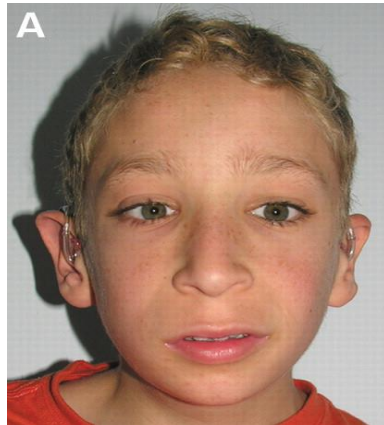
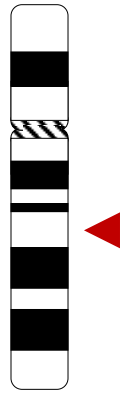


Human Genome Segmental Duplication Map

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



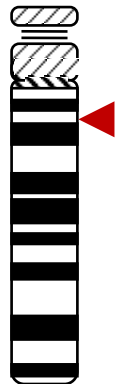
Chromosome 17



Chromosome 15

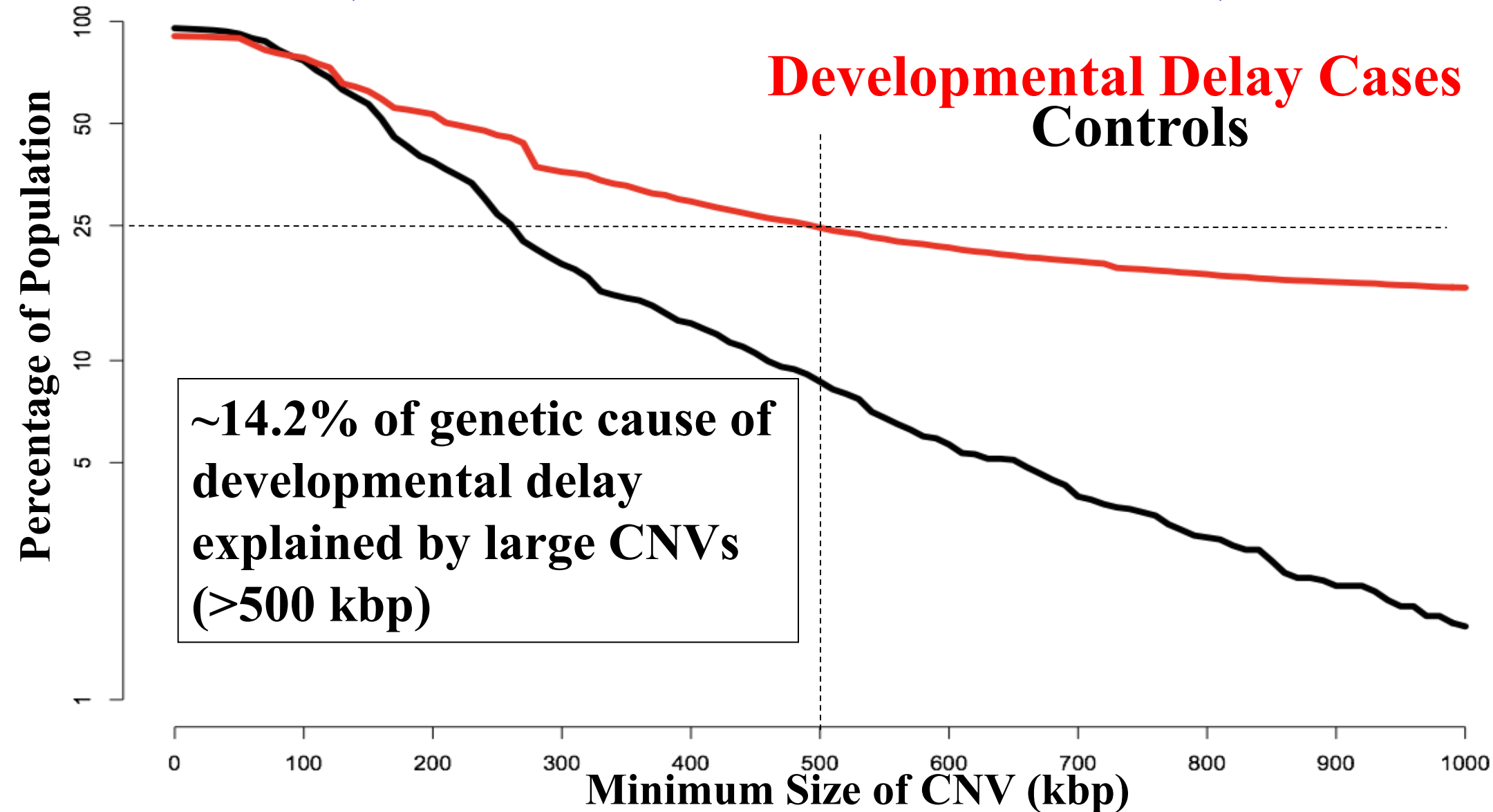


Chromosome 15

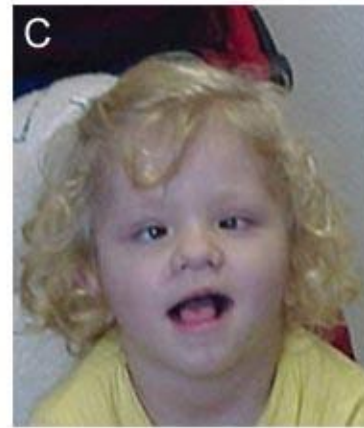
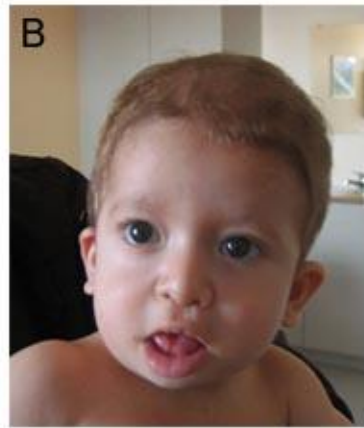


Genome Wide CNV Burden

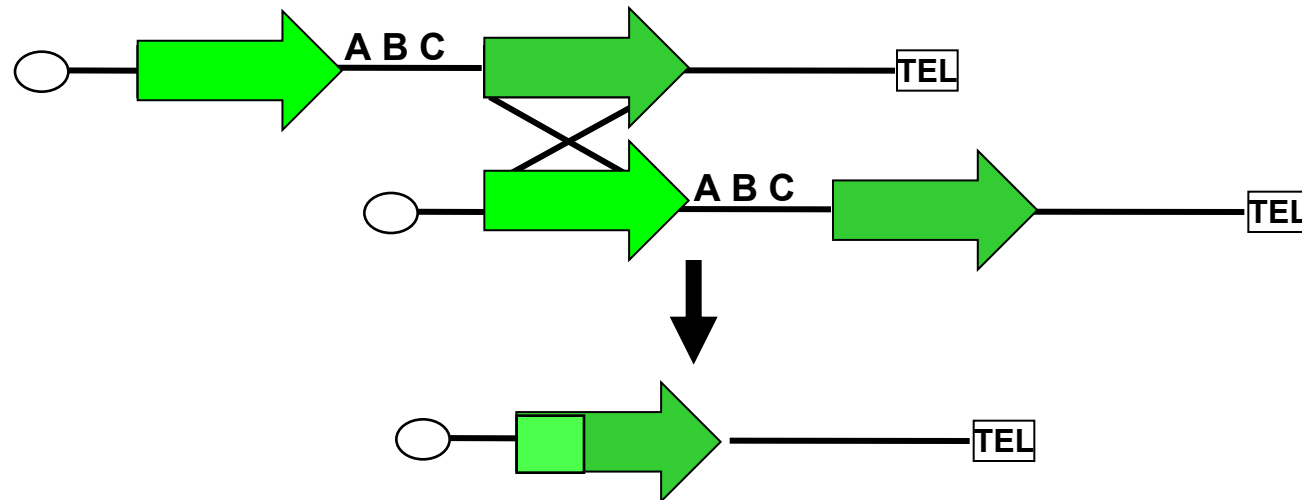
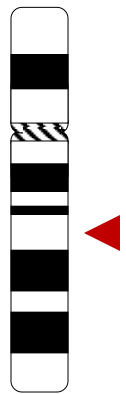
(15,767 cases of ID,DD,MCA vs. 8,328 controls)

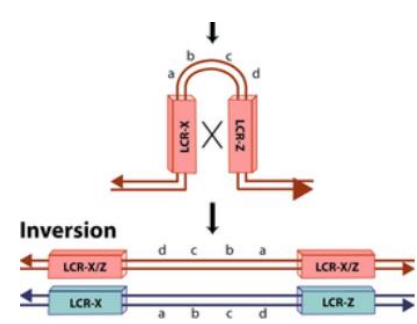


Common and rare structural variation are linked 17q21.31 deletion syndrome

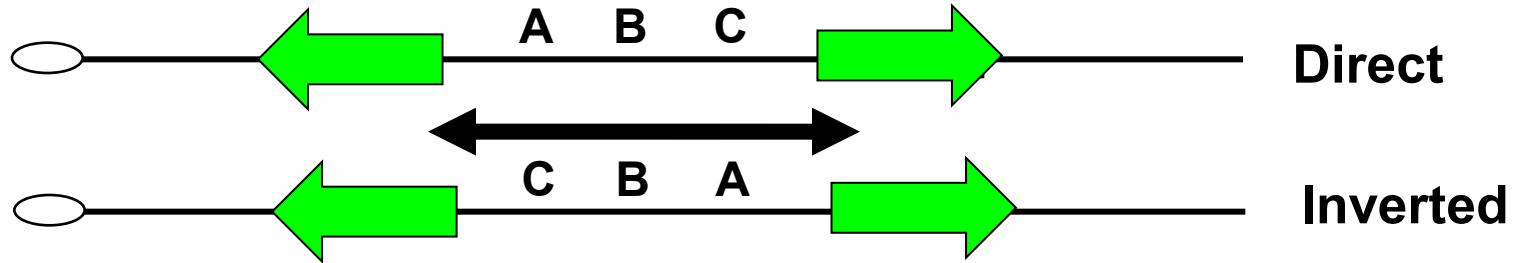


Chromosome 17



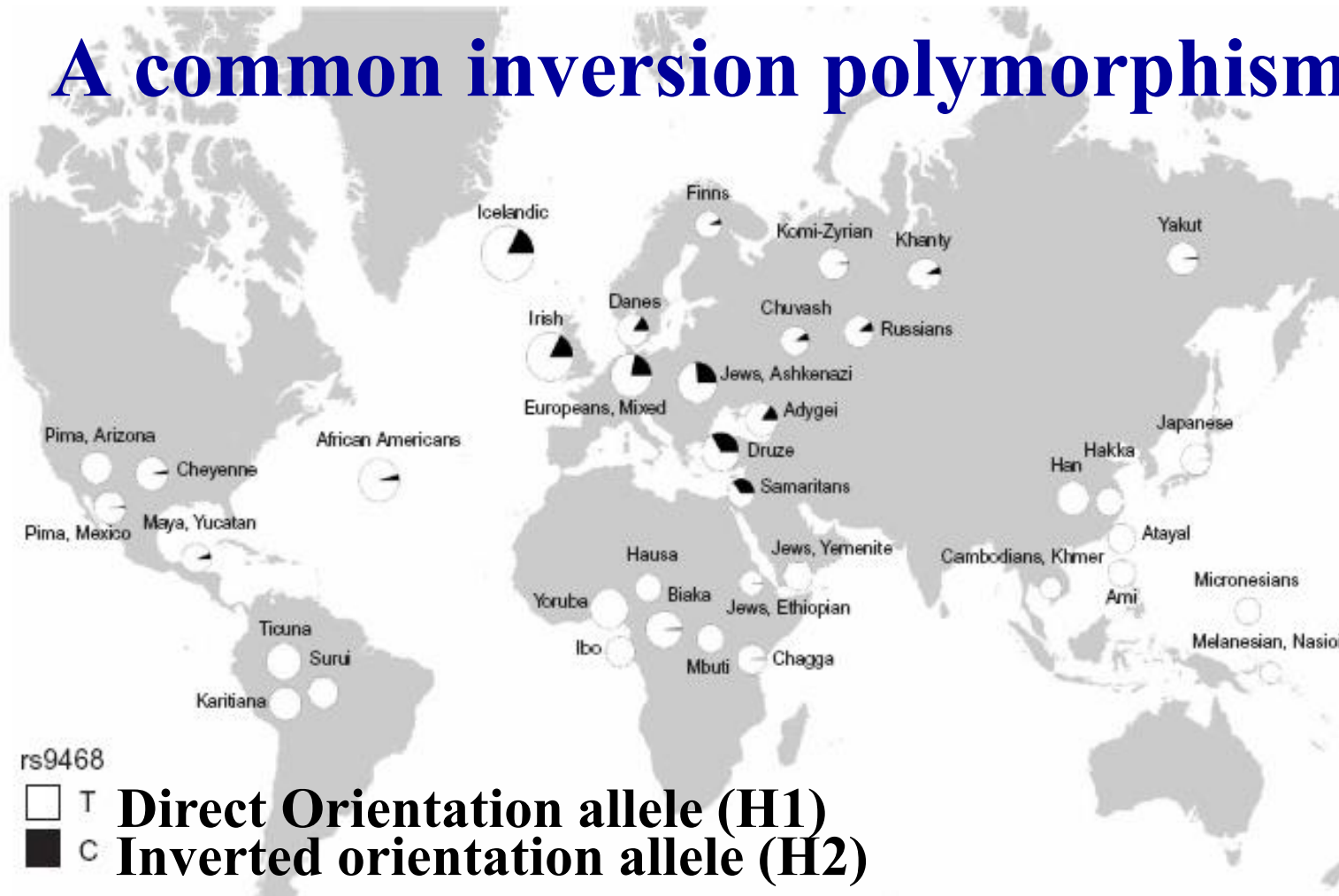


17q21.31 inversion



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

b A common inversion polymorphism



- 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

Summary

- Human genome is enriched for segmental duplications which predisposes to recurrent large CNVs during germ-cell production
- 14-15% of neurodevelopmental disease in intellectual disabled children is “caused” by large CNVs—8% of normal individuals carry large events
- Segmental duplications enriched >10 fold for structural variation because of increased probability of non-allelic homologous recombination
- Increased complexity is beneficial and deleterious: Ancestral duplication predisposes to inversion polymorphism, inversion polymorphisms acquires duplication, haplotype becomes positively selected and now predispose 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

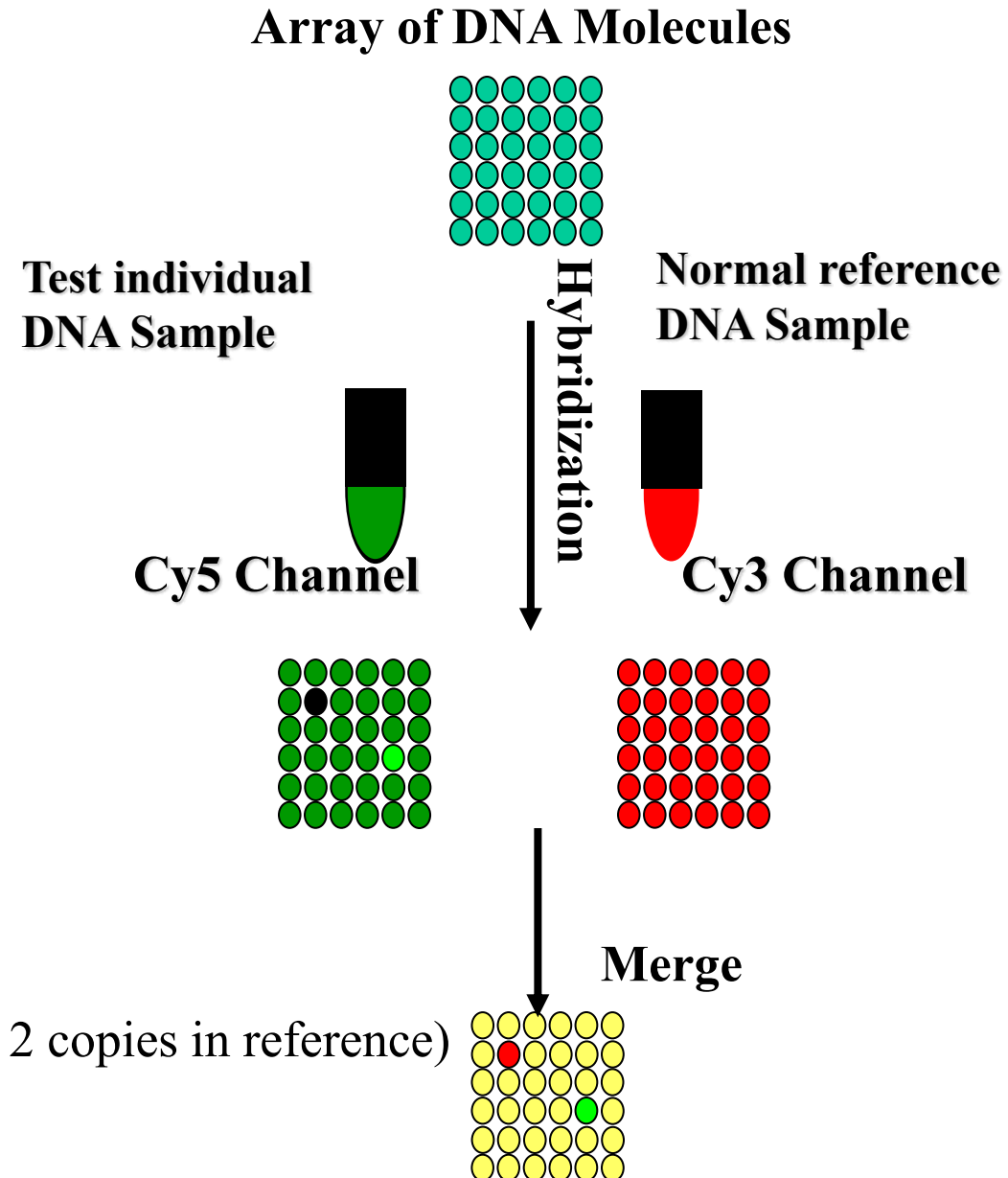
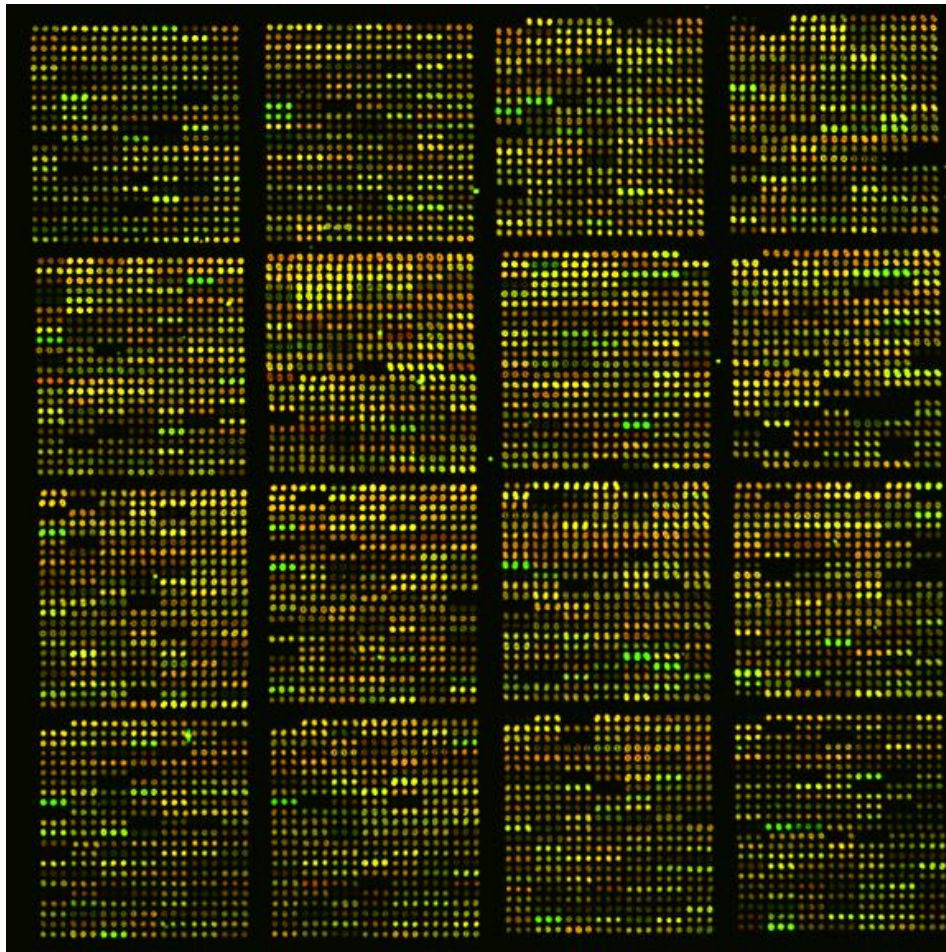
Single molecule mapping

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

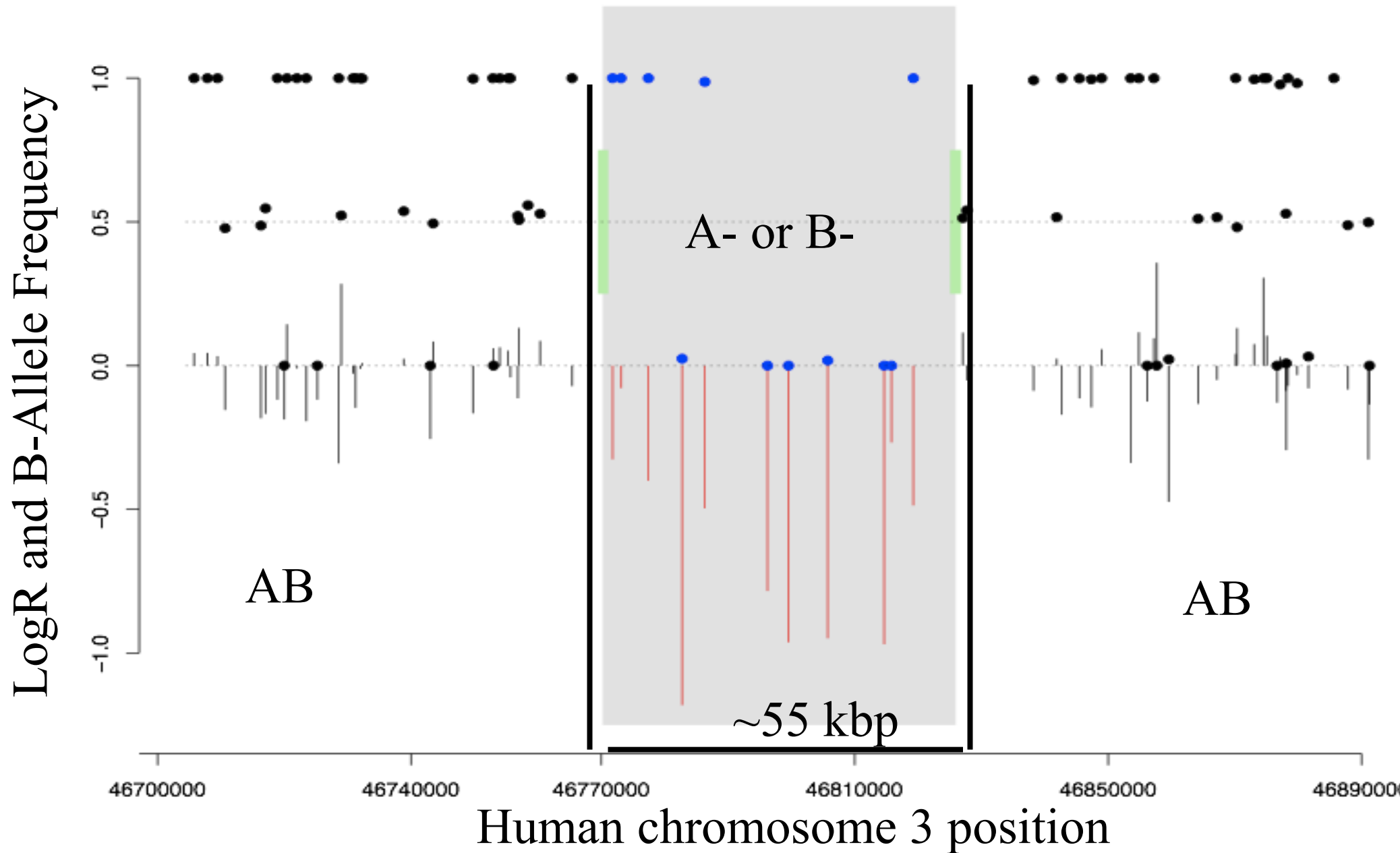
Sequencing-based

- 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

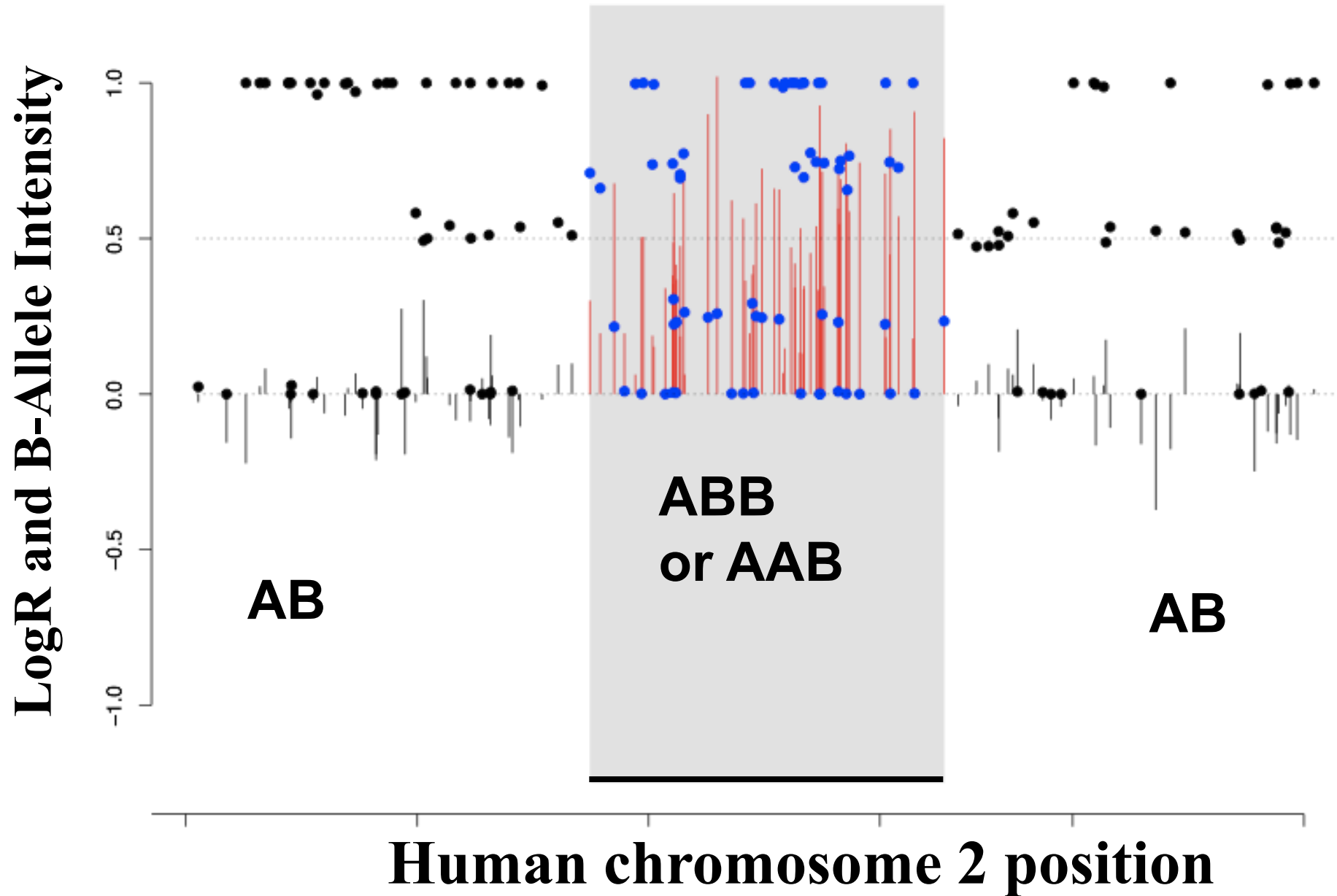
Array Comparative Genomic Hybridization



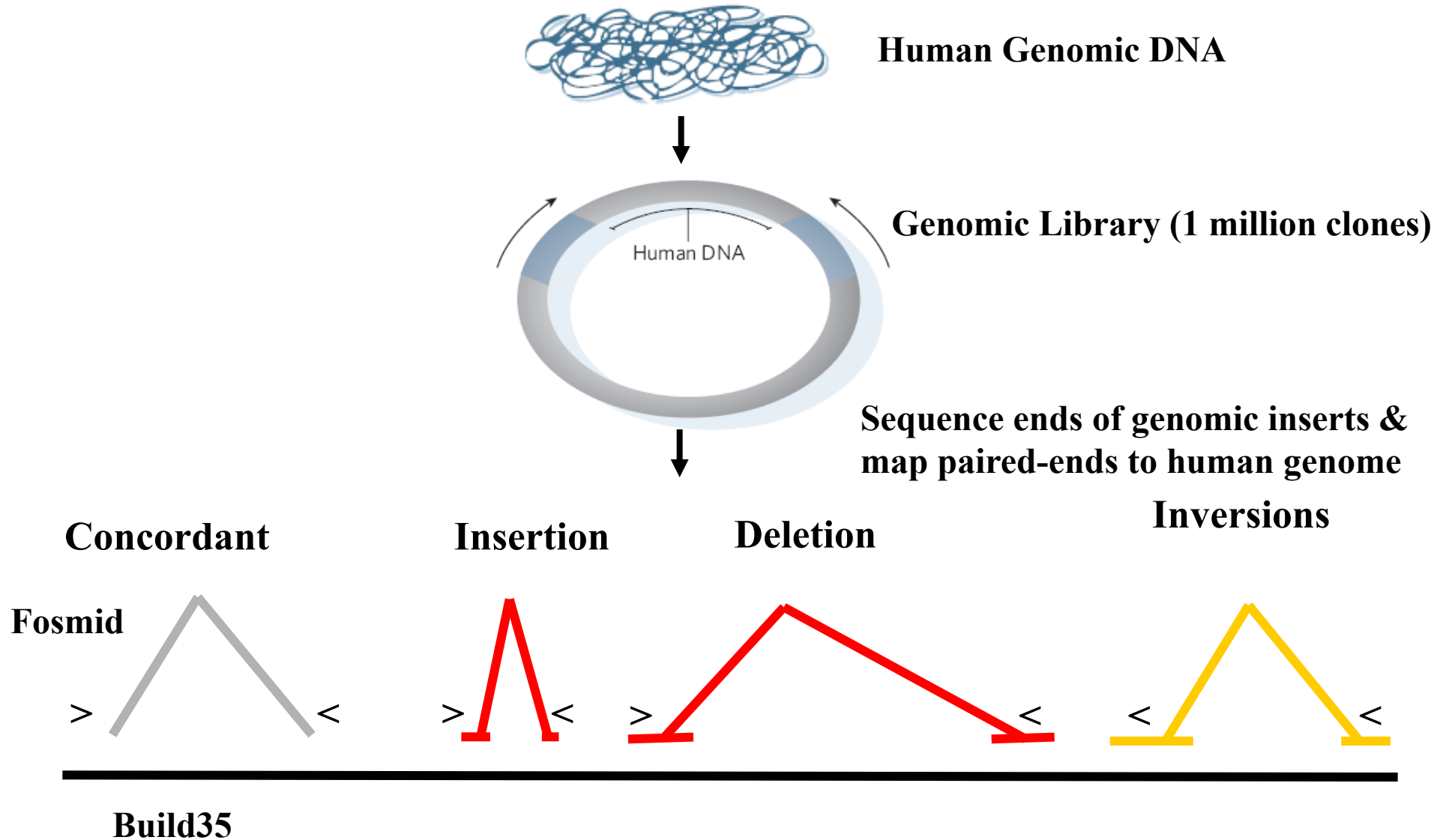
SNP Microarray detection of Deletion (Illumina)



SNP Microarray detection of duplication

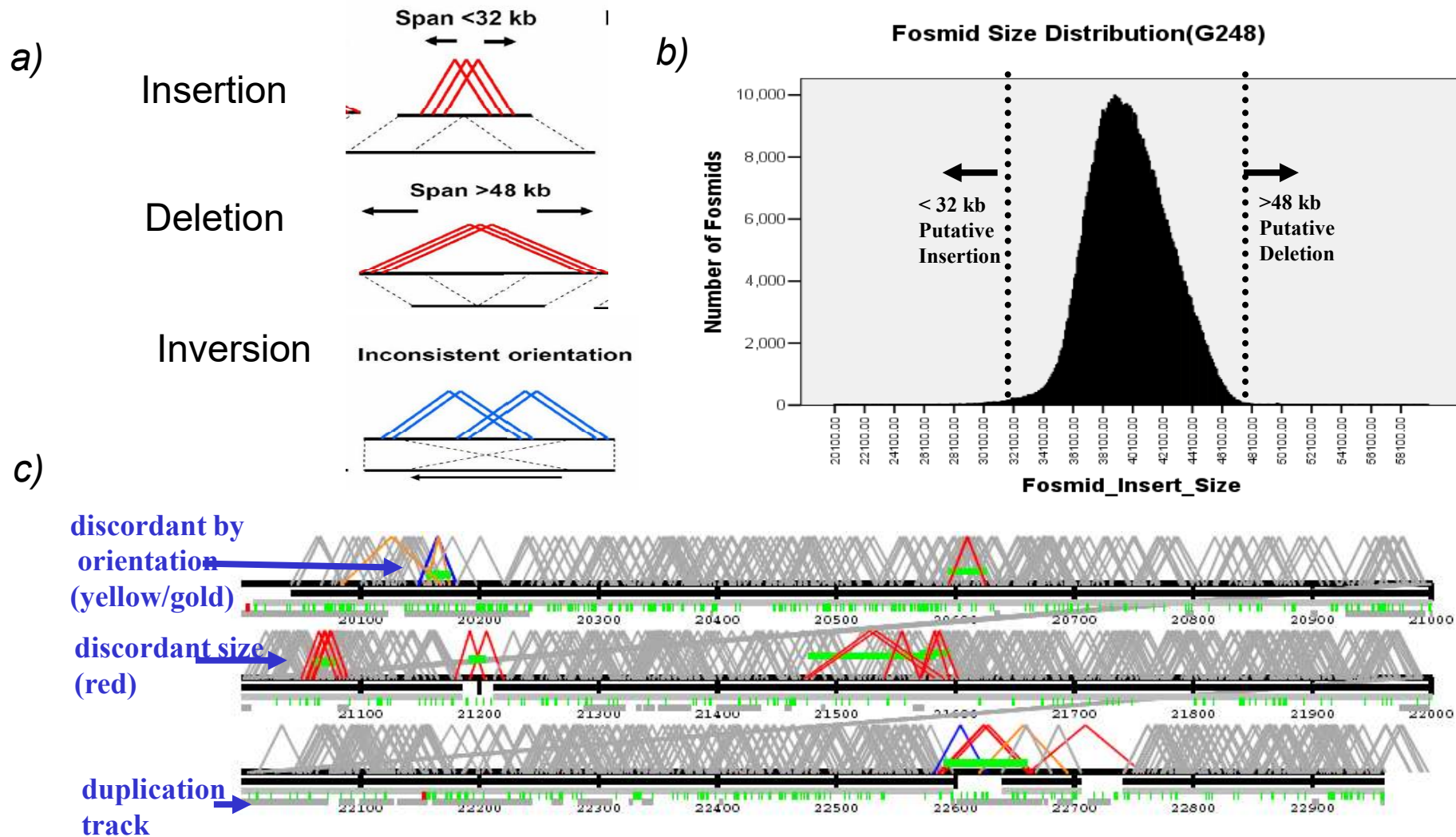


Using sequence read pairs to detect structural variation



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

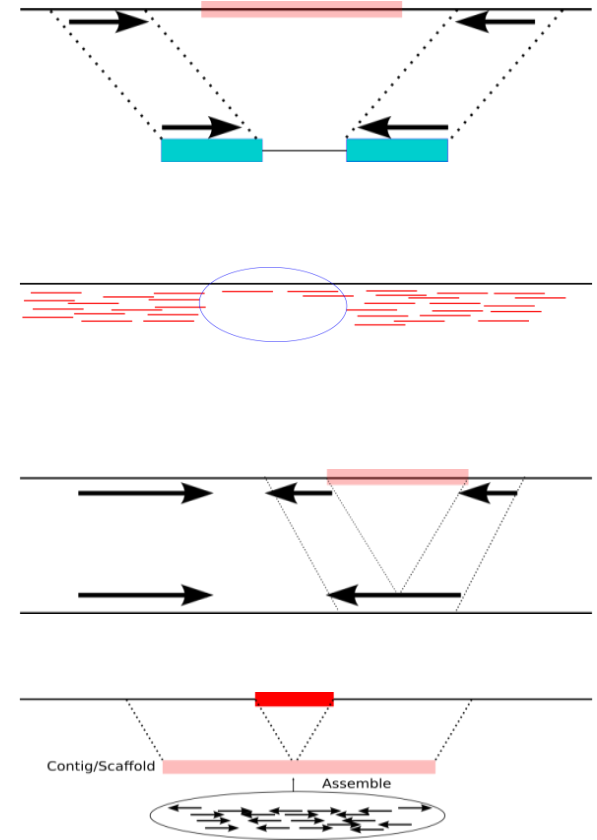
Genome-wide detection of structural variation (>8kb) by end-sequence pairs or “mate pairs”



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

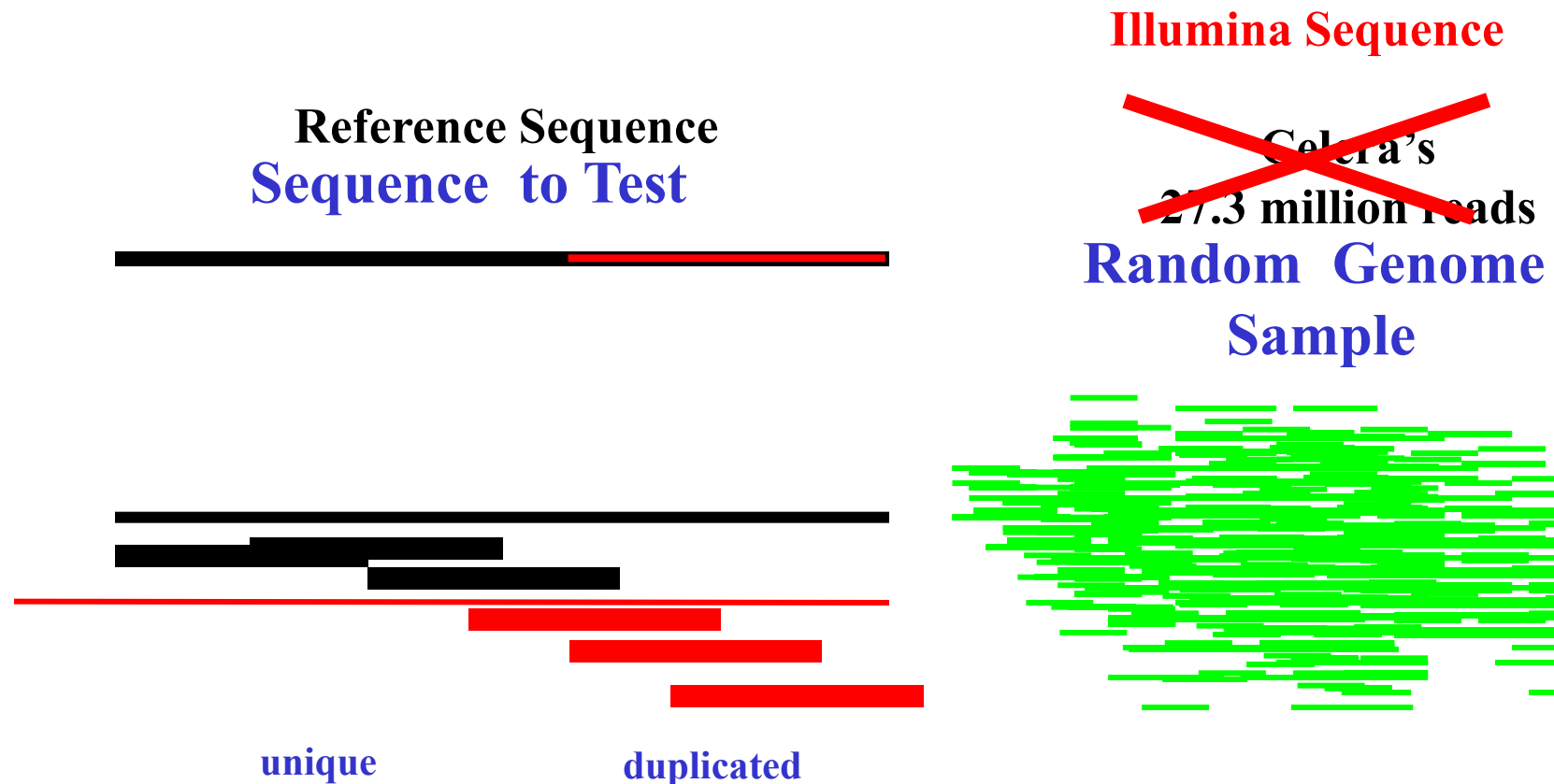
Next-Generation Sequencing Methods

- **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



Using Sequence Read Depth

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



Personalized Duplication or Copy-Number Variation Maps

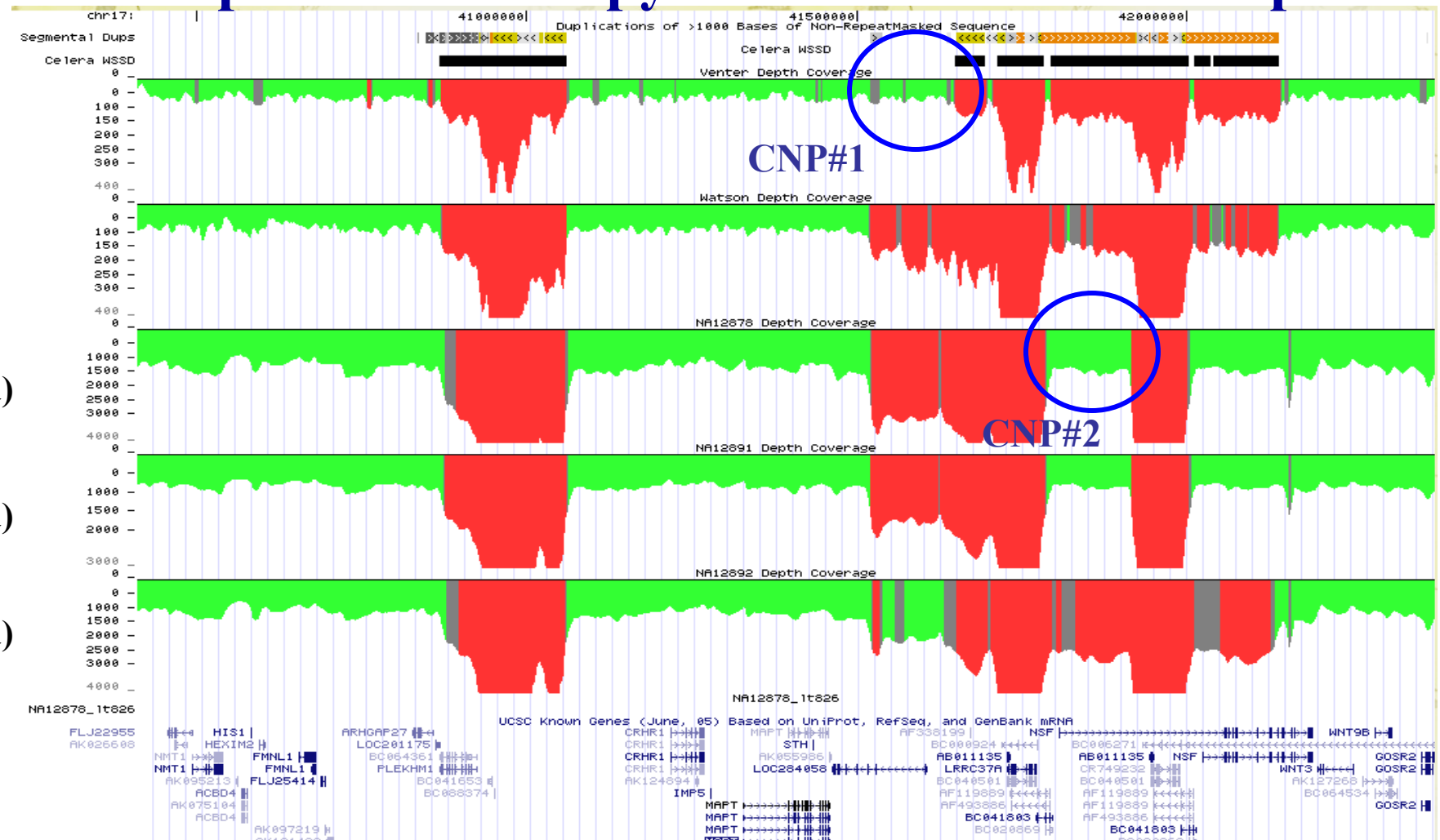
Venter (Sanger)

Watson (454)

NA12878 (Solexa)

NA12891 (Solexa)

NA12892 (Solexa)

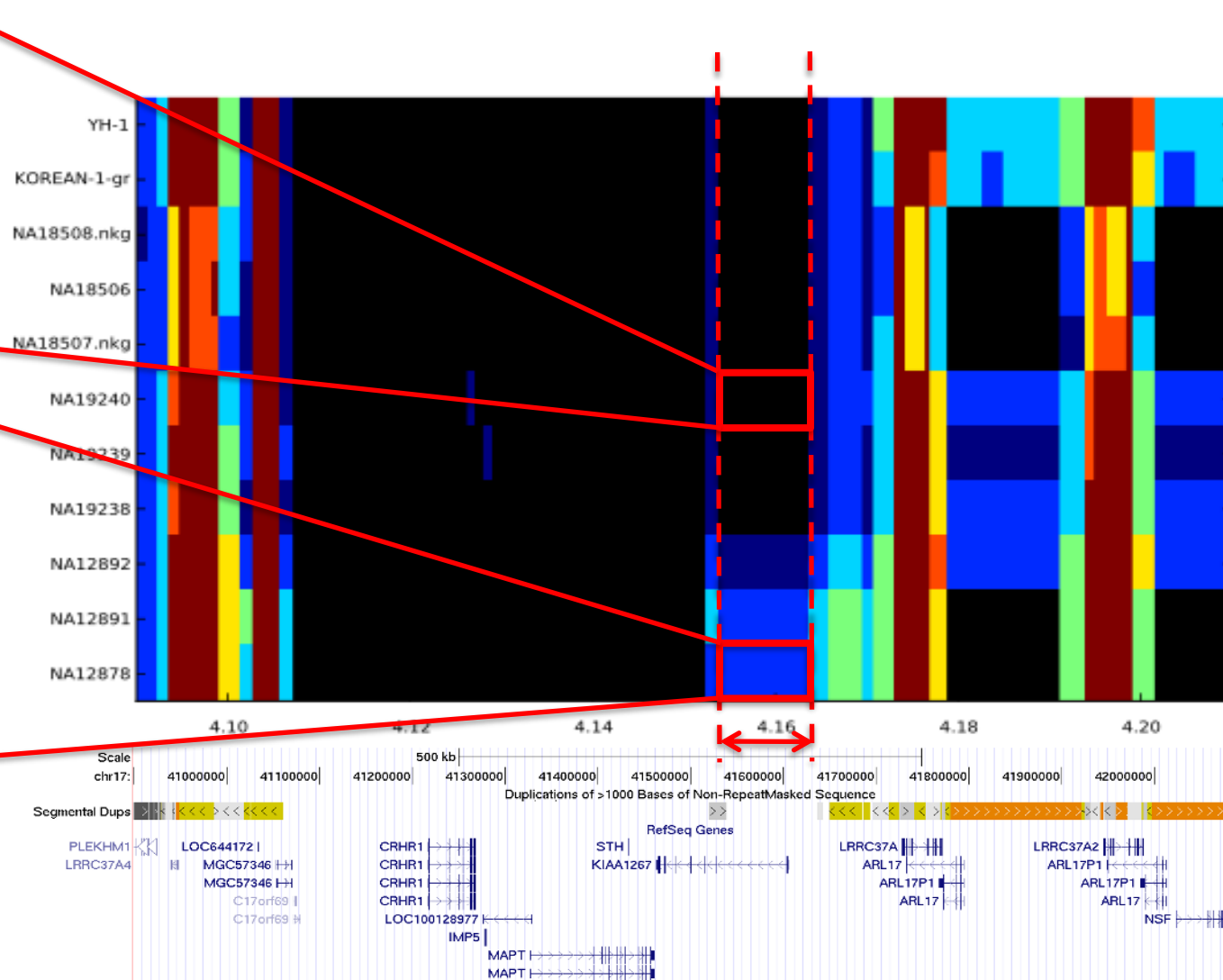
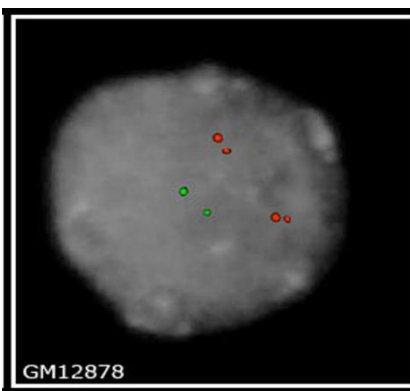
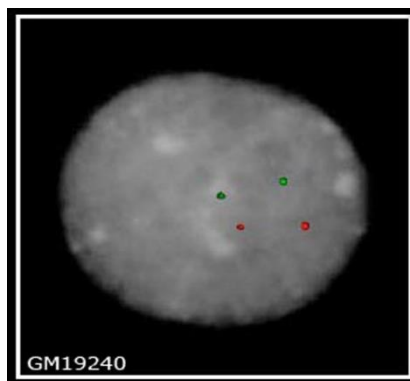


•Two known ~70 kbp CNPs, CNP#1 duplication absent in Venter but predicted in Watson and NA12878, CNP#2 present mother but neither father or child

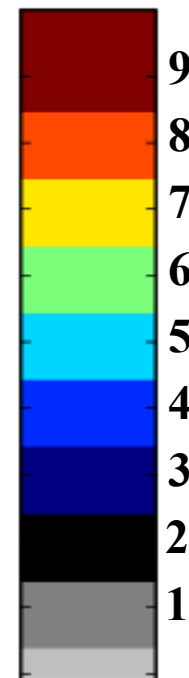
Alkan, Nat. Genet, 2009

Read-Depth CNV Heat Maps vs. FISH

Interphase FISH



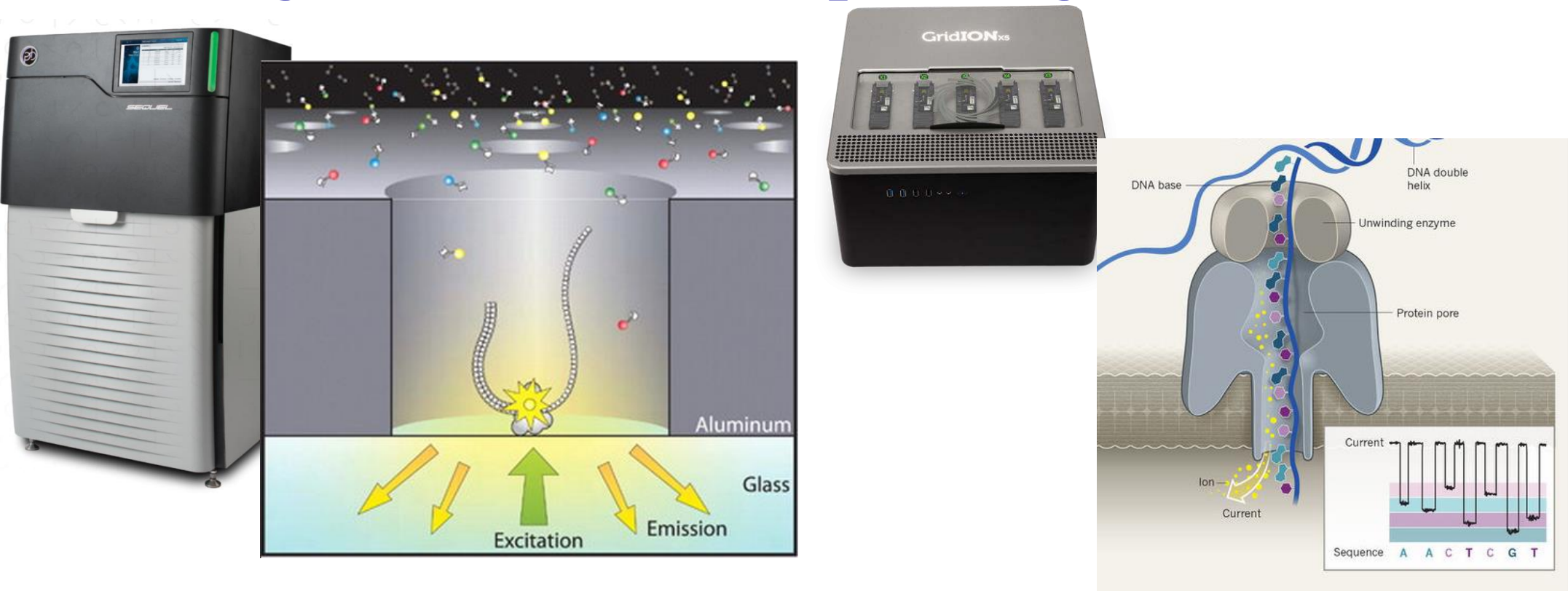
Copy Number



Challenges with indirect short-read approaches

- Repeat biases—a large fraction of events mediated by repeats are missed
- Size biases—read-pair methods limited by insert size while read-depth typically requires $> 5\text{kbp}$
- Class bias: deletions $\gg$ duplications $\gg$ balanced events (inversions)
- Multiallelic copy number states—incomplete references and the complexity of repetitive DNA
- False negatives.

Long read Genome Sequencing Revolution

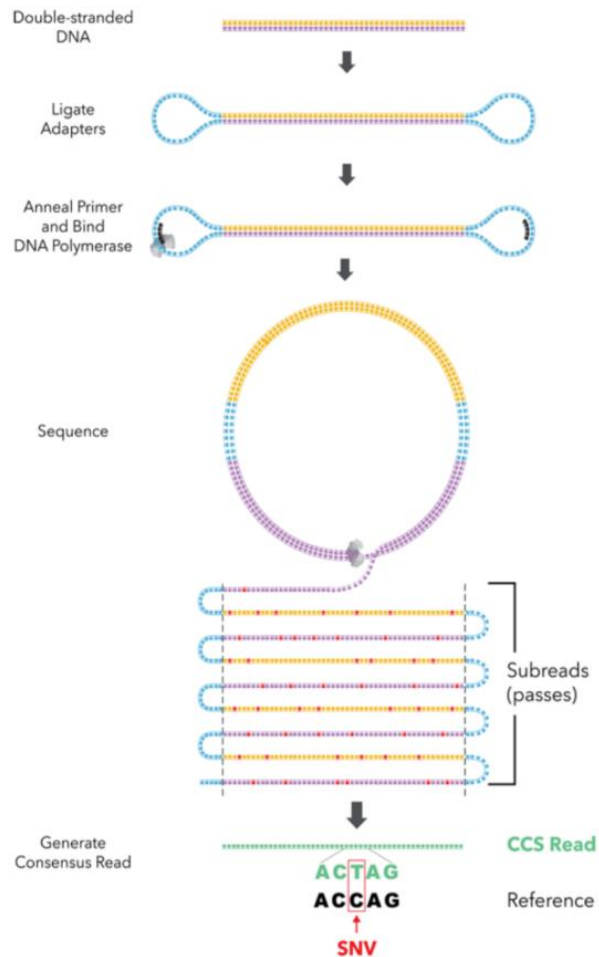


Pacific Biosciences (PacBio)—single-molecule real-time sequence (SMRT) data (15-50) kbp sequence reads

ONT (Oxford Nanopore Technology)—higher error rate but, portable, scalable native DNA sequencing of long-reads

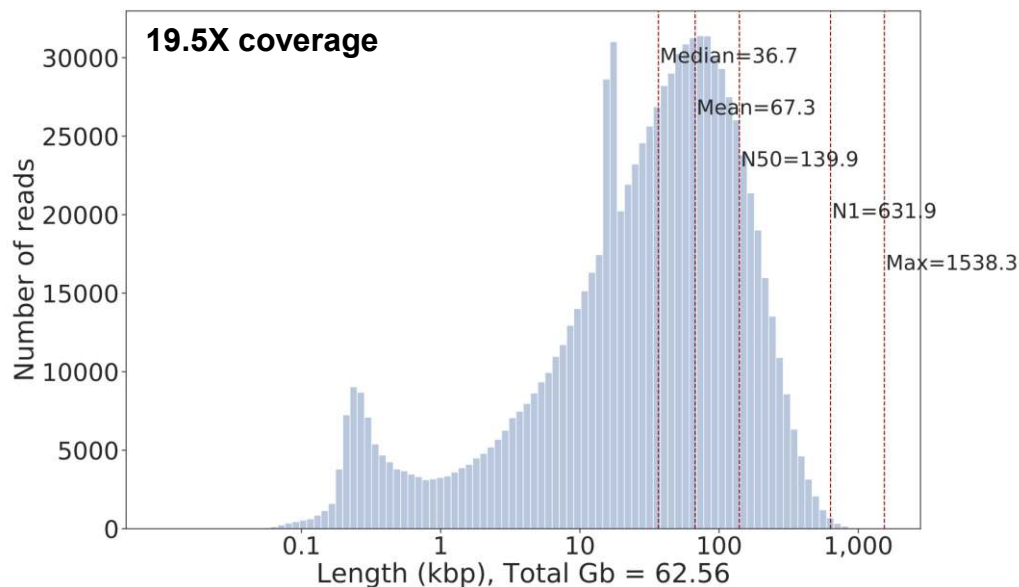
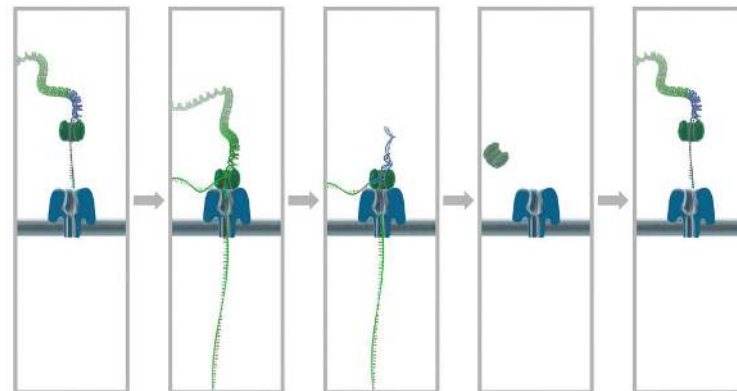
Advances in long-read sequencing

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



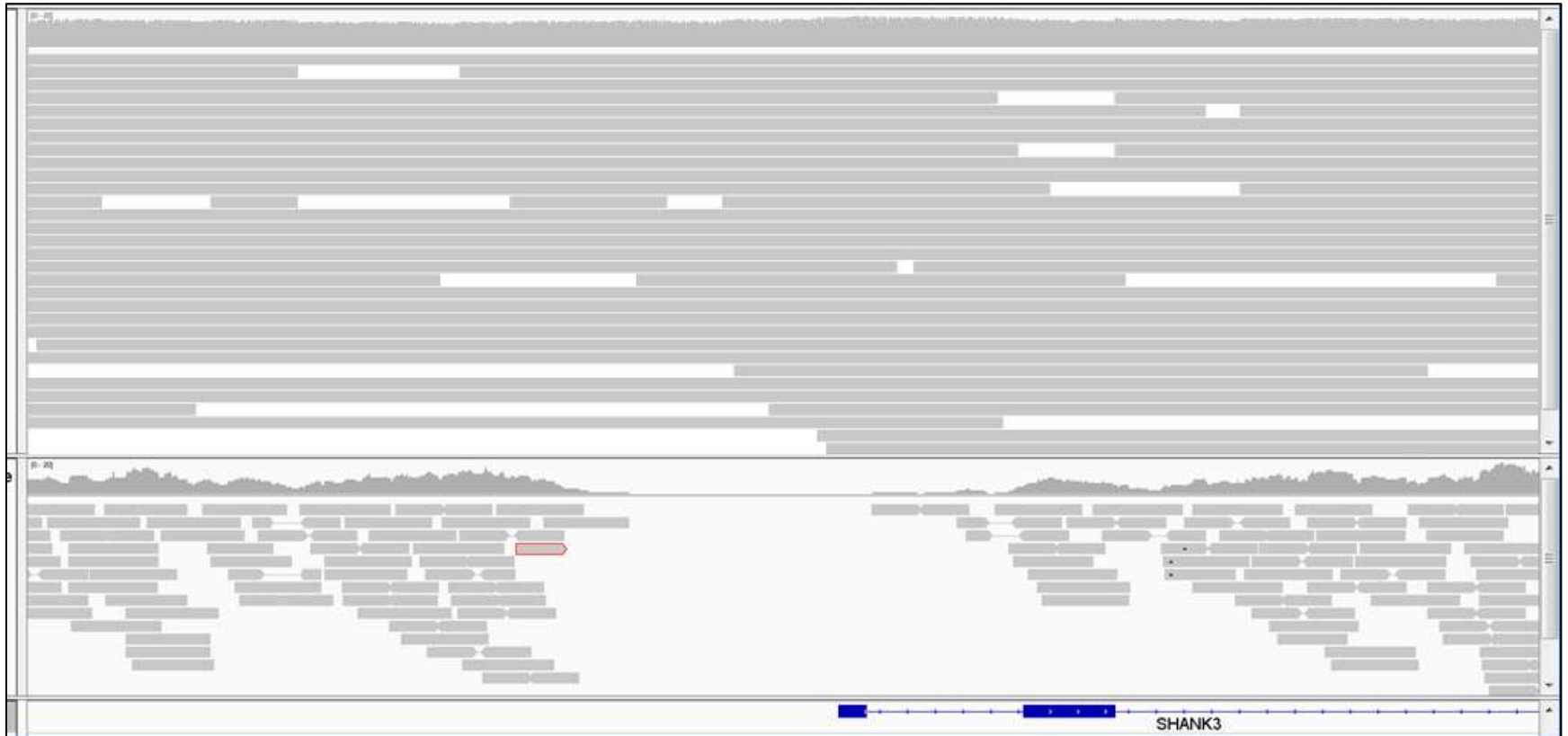
Illumina
-
150-300 bp

More uniform coverage and sequencing of native DNA

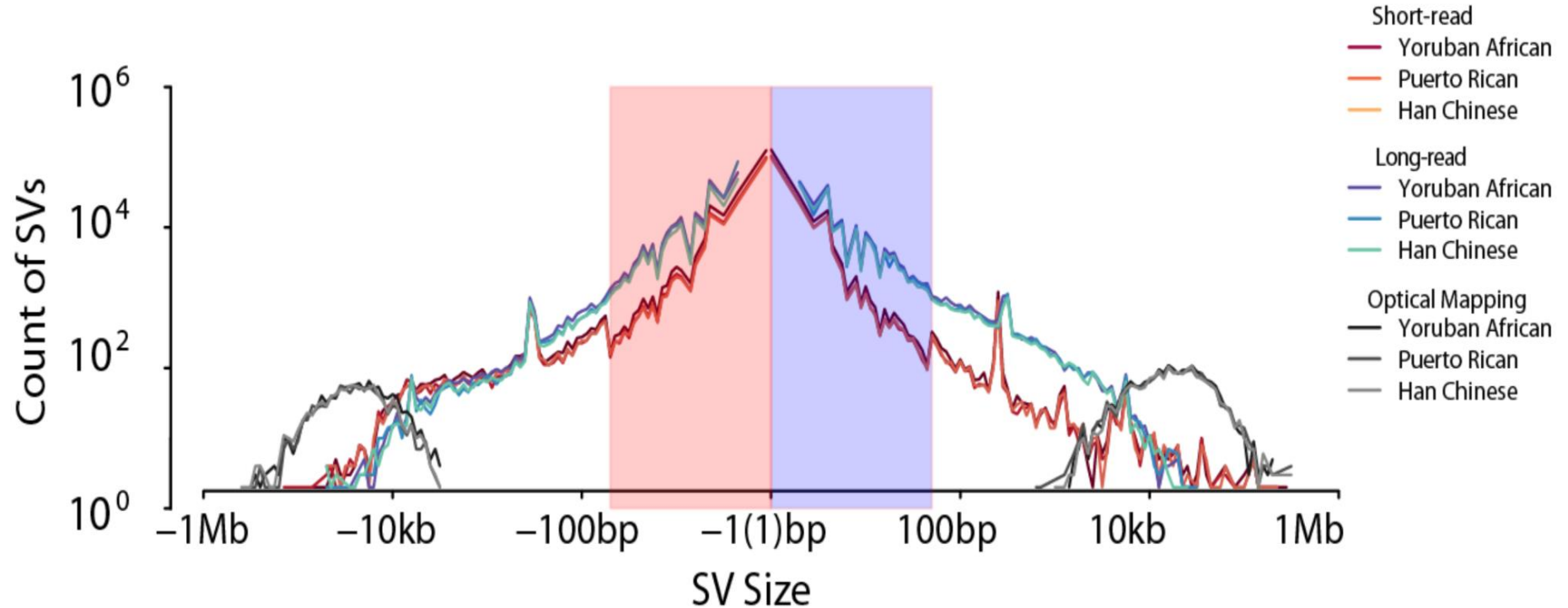
SHANK3

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**

Complete sequence of a human genome

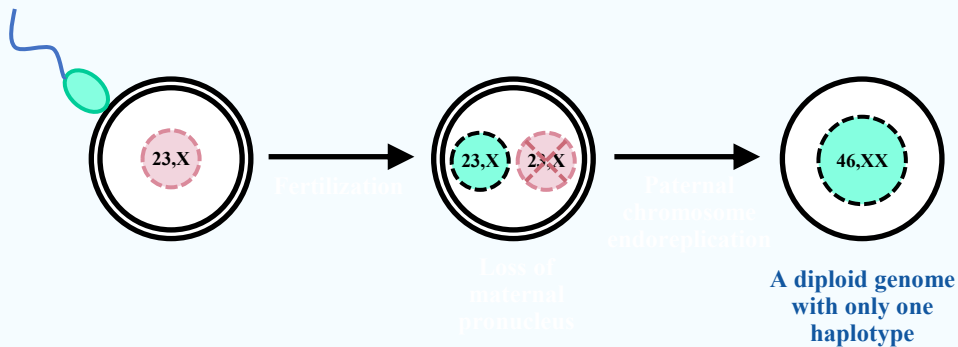


- 8% of missing genome sequence added (>200 Mbp)
- Complete sequence of centromeres, acrocentric and segmental duplications
- Adds 1956 gene predictions of which 130-190 are protein coding
- Framework for understanding the genetically most complex regions of our genome.

So how did we do it?

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

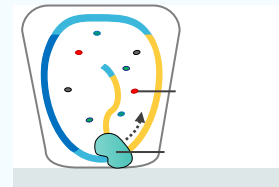
CHM13 is a complete hydatidiform mole



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

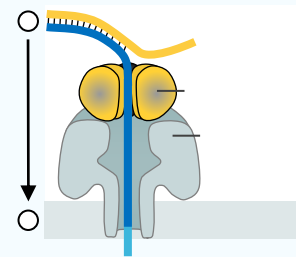
We used two long-read sequencing technologies with complementary strengths

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



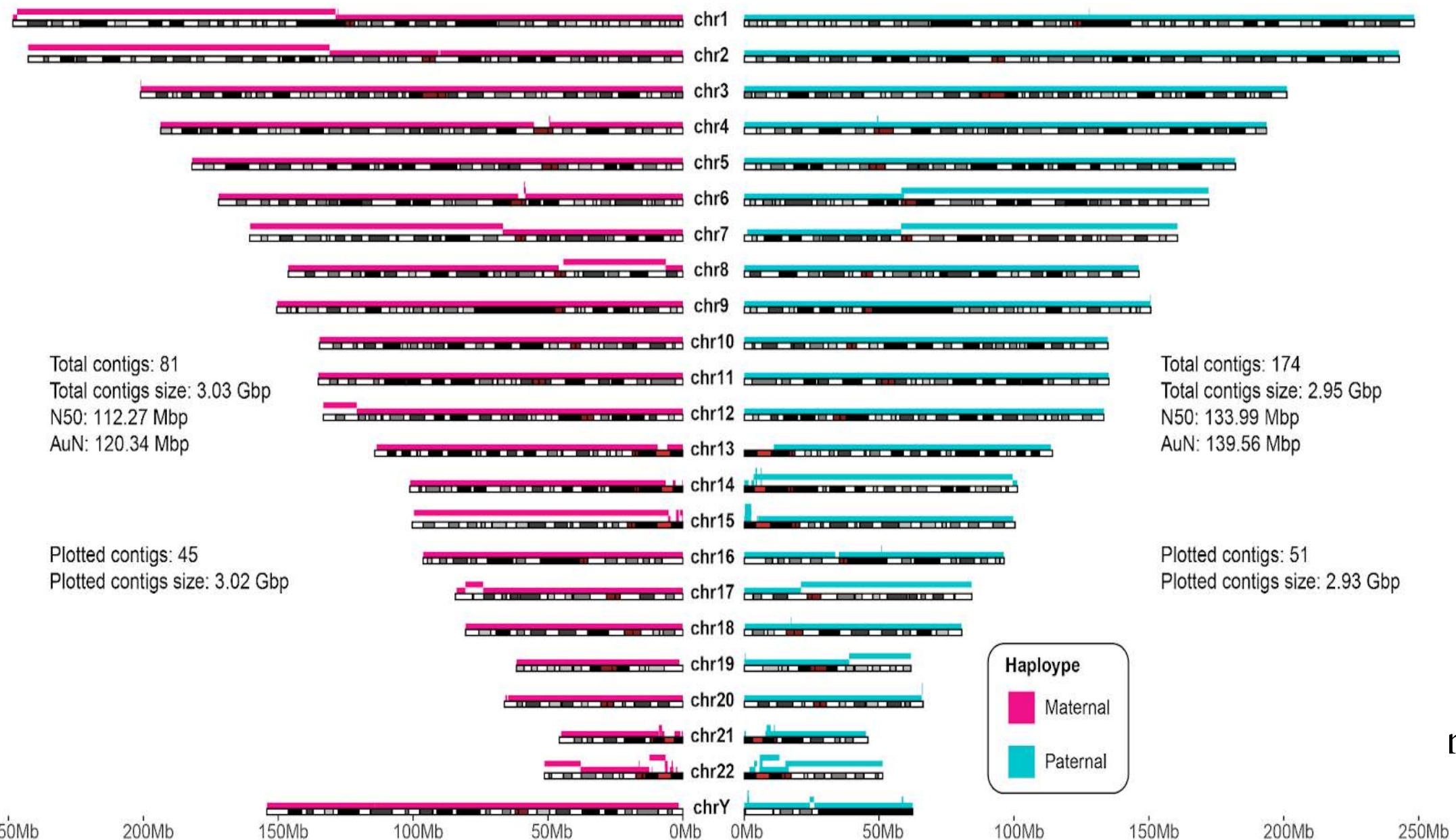
- 15-25 kbp long
- >99% accurate (similar to Illumina)
- Strength: Extremely accurate

2. Oxford Nanopore Technologies (ONT)



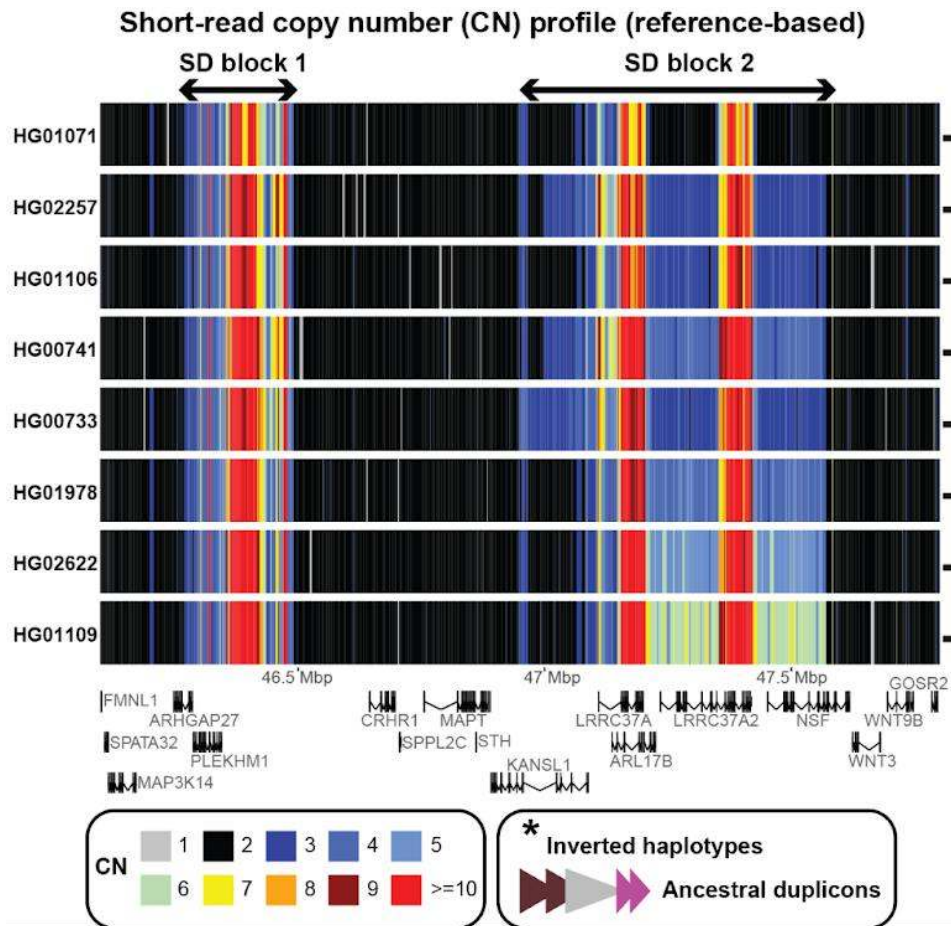
- No limit in read length!
- 93-99% accurate
- Strength: Extremely long

Trio-based verkko assemblies of HG002

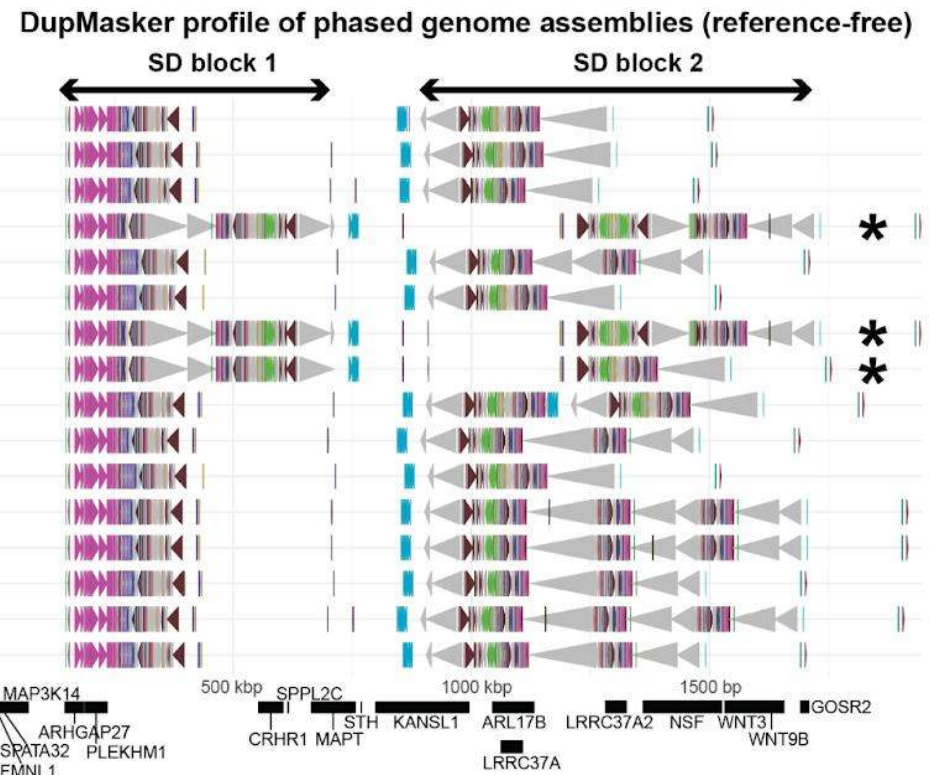


LRS has transformed how we characterize copy number and structural variation

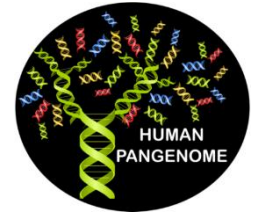
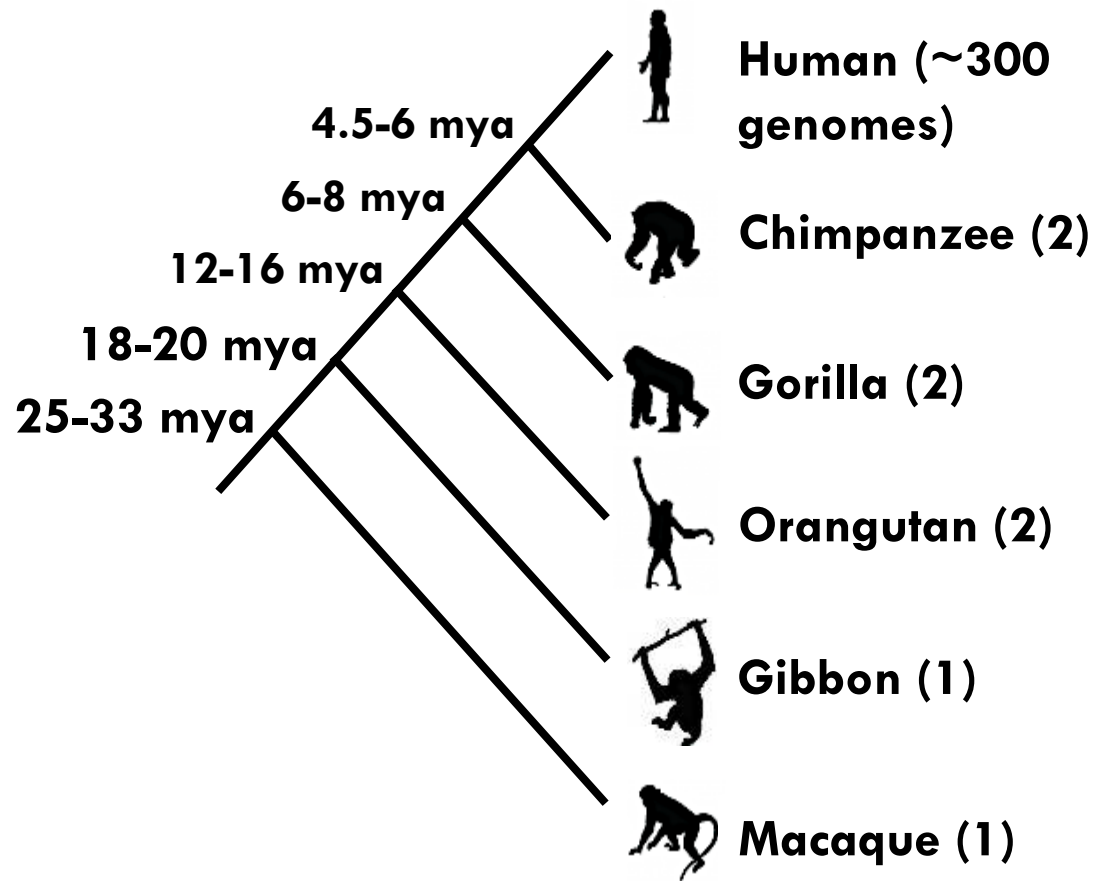
2015



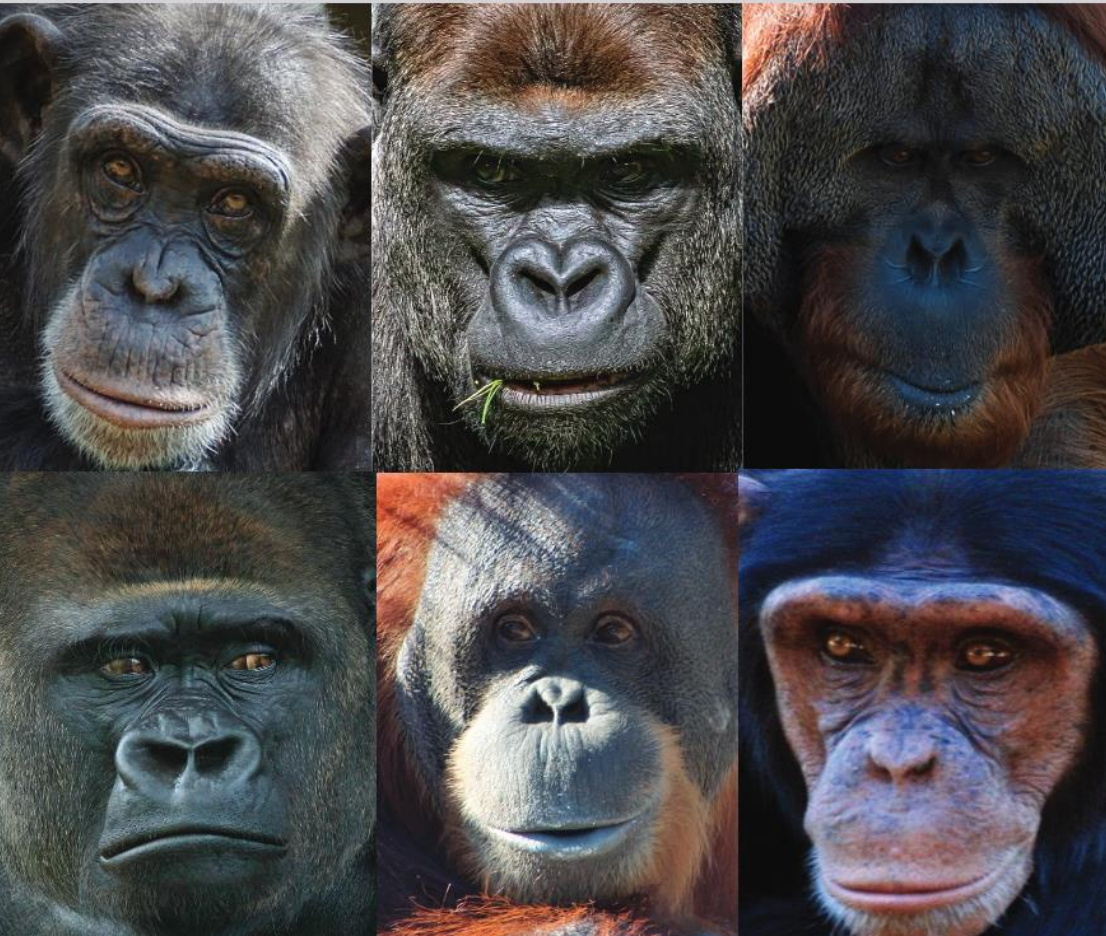
2023



Primate phased genome assembly efforts



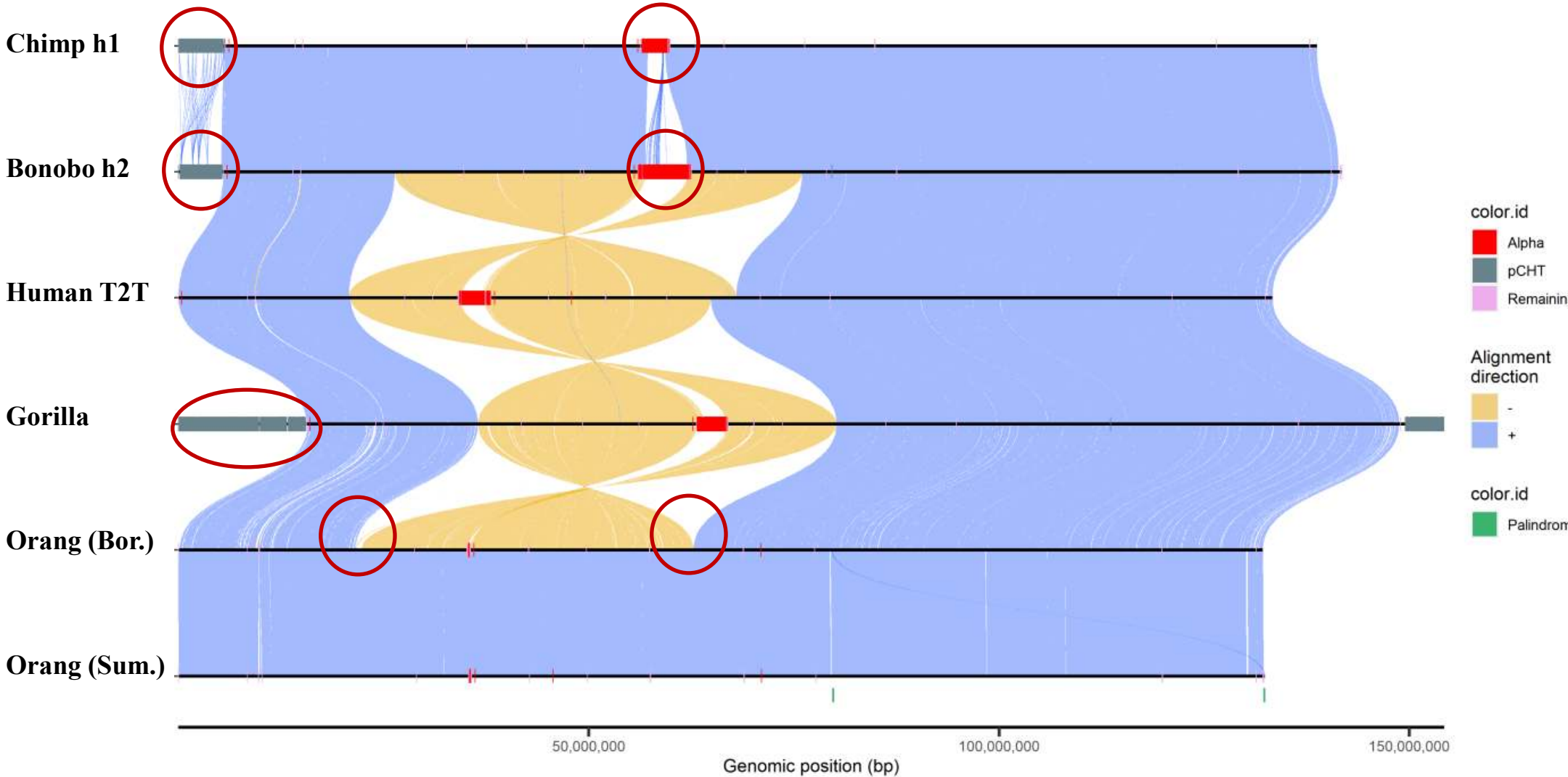
Ape telomere-to-telomere sequencing



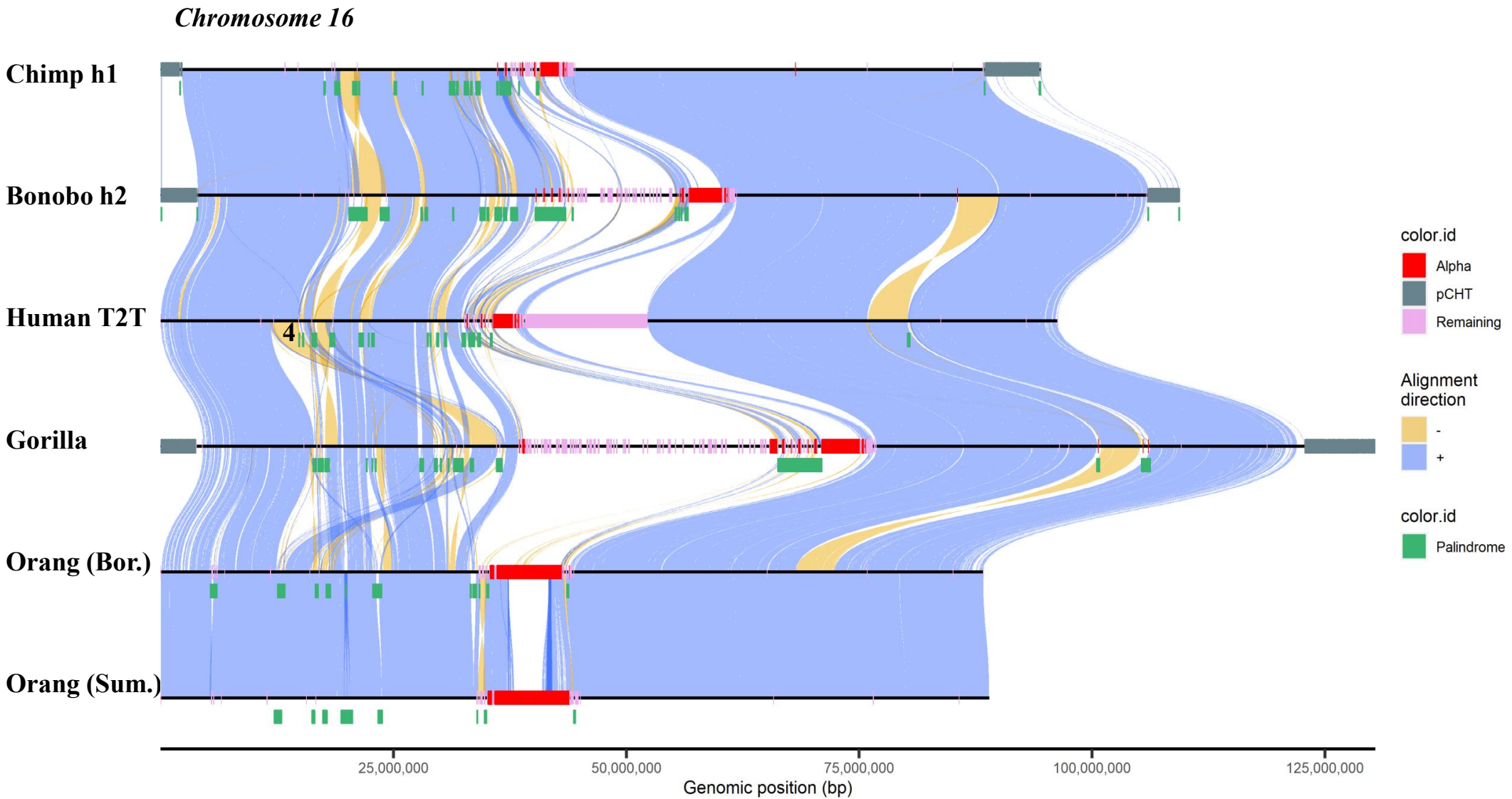
- Chimpanzee, bonobo, Western lowland gorilla, Bornean orangutan, Sumatran orangutan, Siamang
 - led by UW, Penn State, NIH (60 labs)
 - Assembled 12 genomes (99.98-99.99% complete)
- 94% non-rDNA chromosomes T2T
- QV of 61.7-66.4 -- less than 1 error per 2.7 million base pairs

Complete sequencing of ape chromosomes (SVbyEye “stacked” plot)

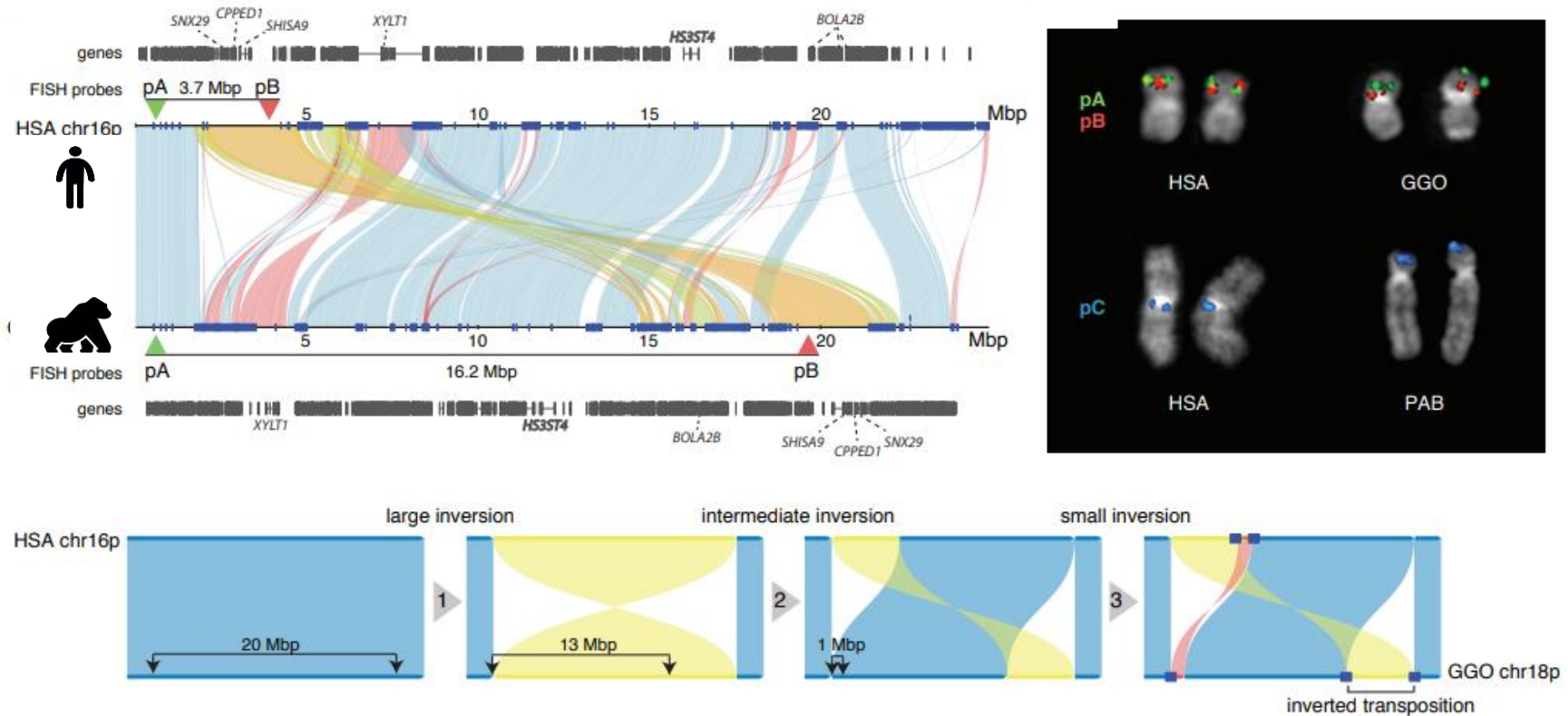
Chromosome 12



Complete sequencing of ape chromosomes



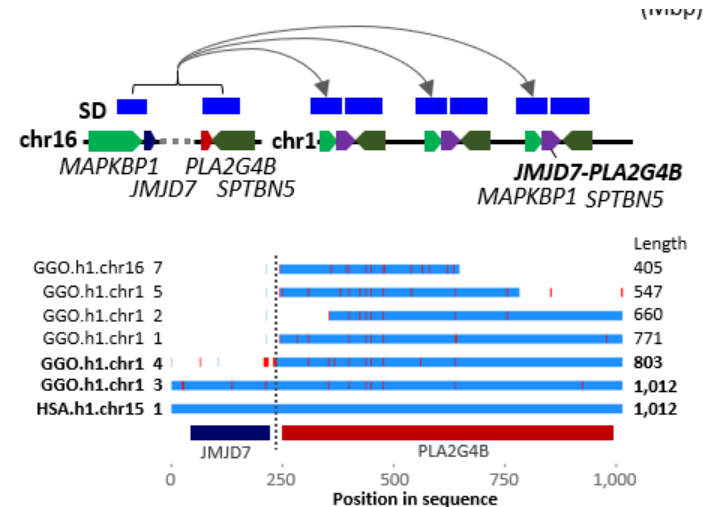
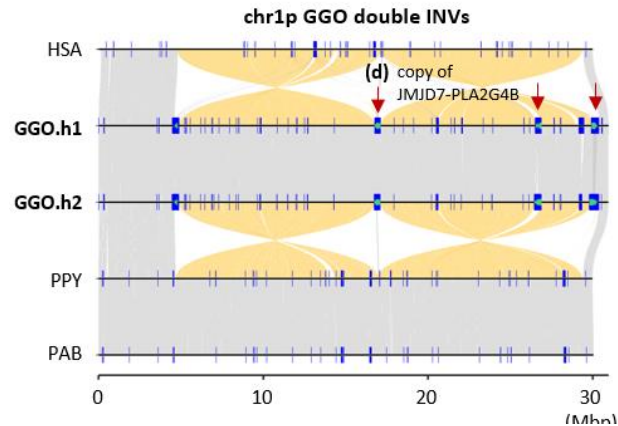
Inversions & SDs are the largest and most abundant form of structural variation affecting genes



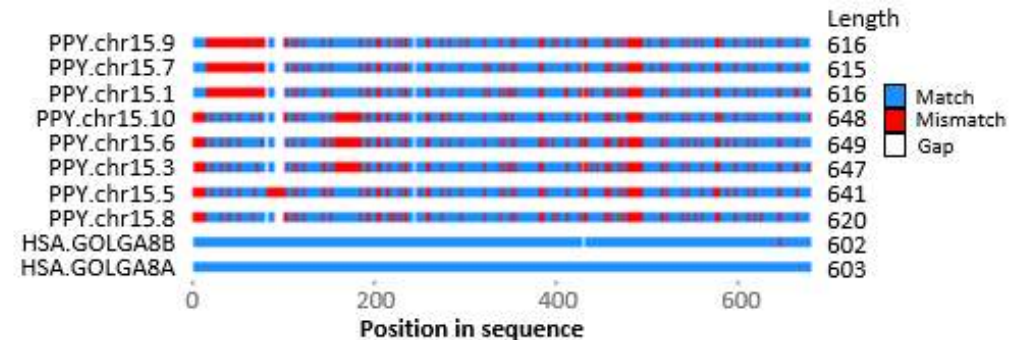
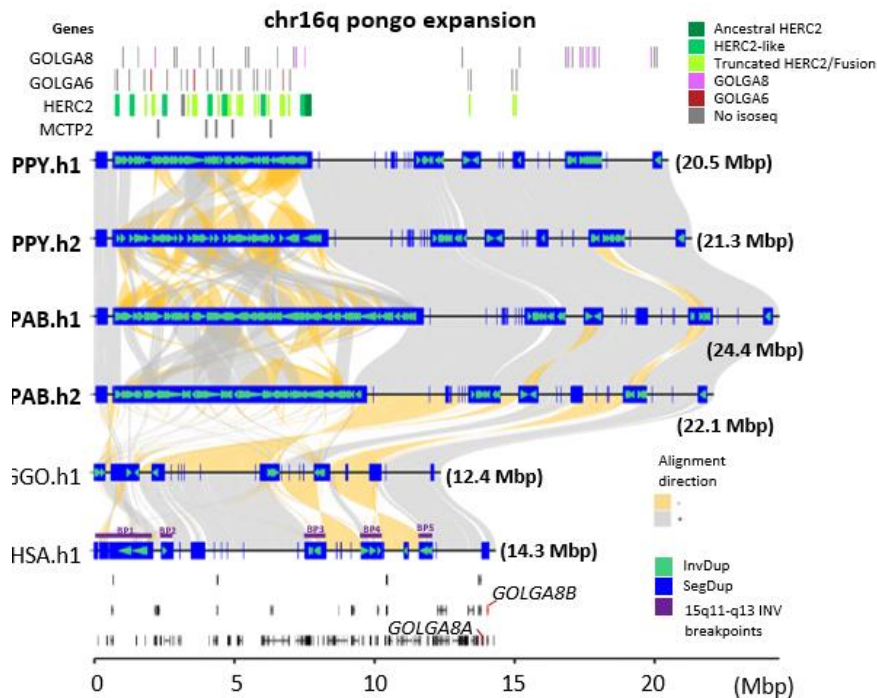
- Discovered a total of 1,140 inversions between ape species--522 are novel; 63.5% are flanked by large highly identical duplications

Lineage-specific duplicate genes emerge at the edges of large-scale inversions

Gorilla



Orang



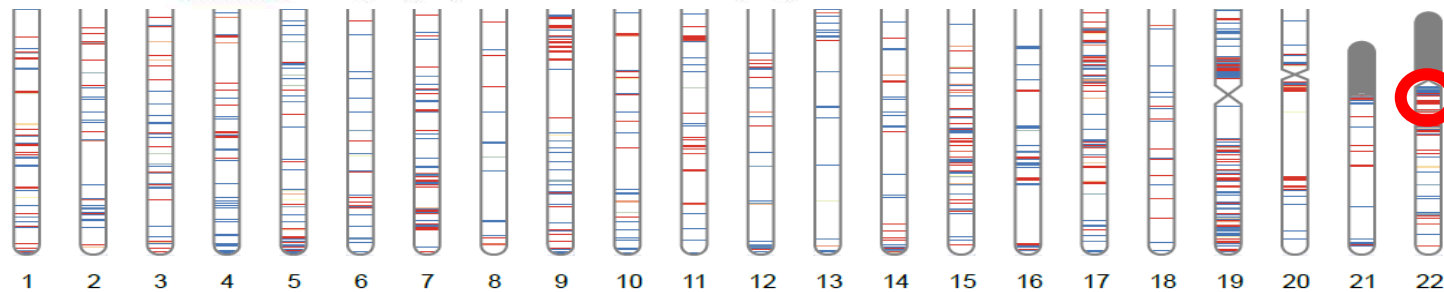
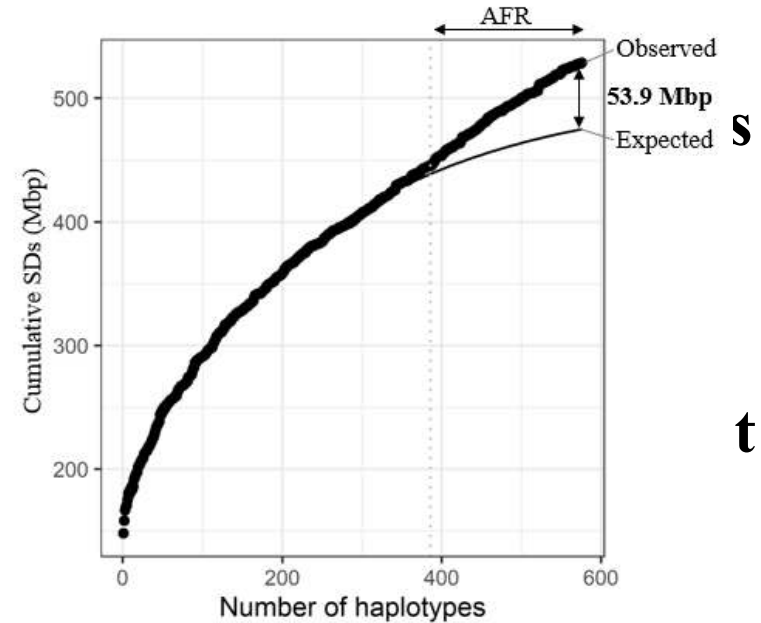
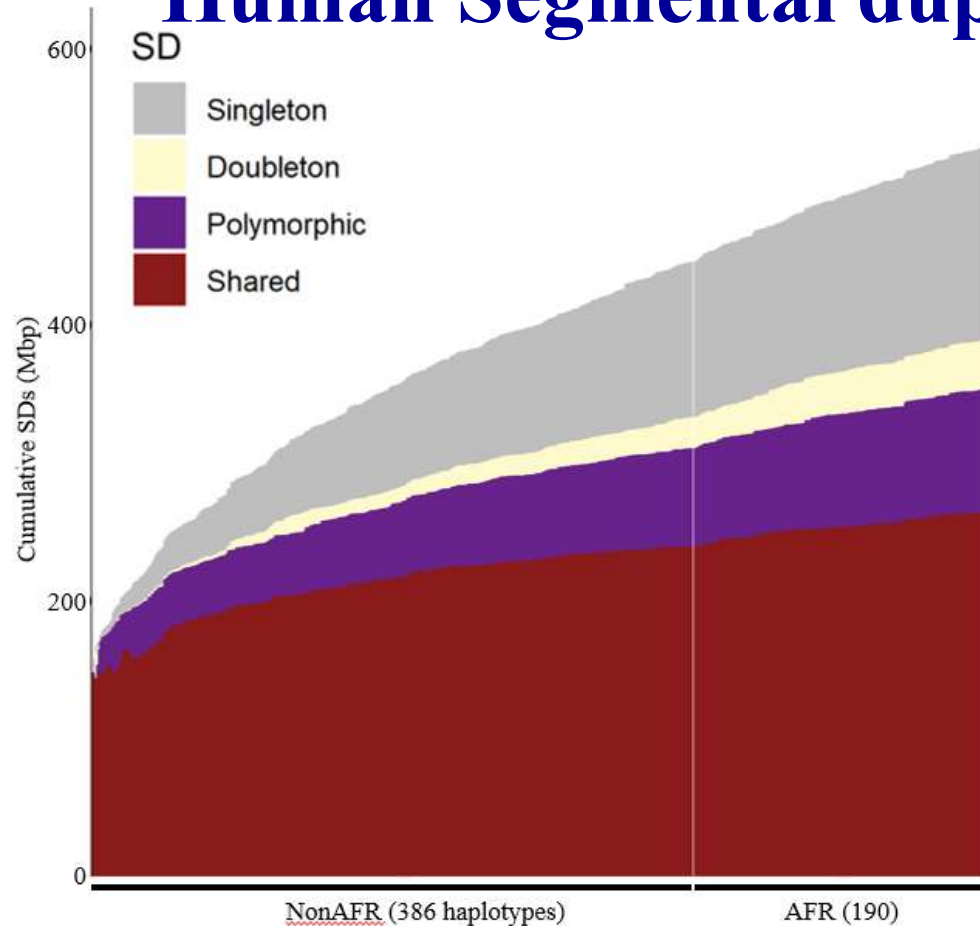
Lineage-specific ape protein-coding genes

Species	Human transcript s mapped	Orthologou s loci	Novel Loci with Isoseq support	Novel loci overlapping SDs	Novel isoforms Outside of novel loci
Bonobo (PPA)	99.82%	93.52%	1482 (6.48%)	1004	1708 (2.1%)
Chimp (PTR)	99.77%	94.82%	1167 (5.16%)	904	2560 (2.9%)
Gorilla (GGO)	99.80%	94.98%	1153 (5.12%)	812	2105 (2.6%)
S. Orang (PAB)	99.44%	94.01%	1328 (5.99%)	909	2363 (2.9%)
B. Orang (PPY)	99.40%	93.95%	1336 (6.05%)	906	2239 (2.8%)
Siamang (SSY)	99.27%	96.51%	770 (3.49%)	471	3638 (5.1%)

*Number of protein coding genes annotated: > 200 AA protein length filter, 50% protein homology percentage filter, IsoSeq filter applied

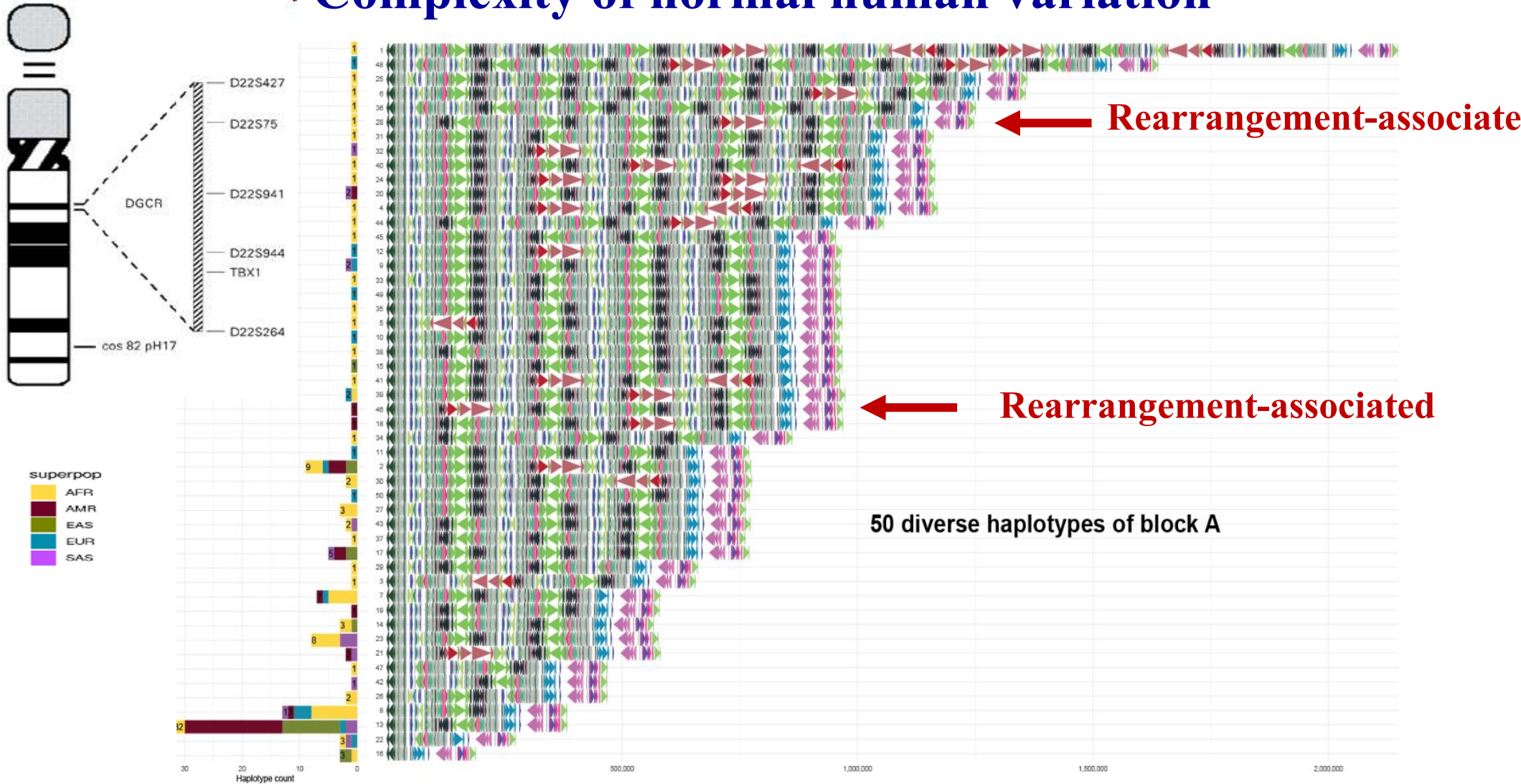
- **5-6% novel, multi-exonic protein coding per ape genes -70-80% driven by lineage-specific SDs** [Yoo et al, bioRxiv, 2024, Nature, 2025](#)

Human Segmental duplication polymorphism

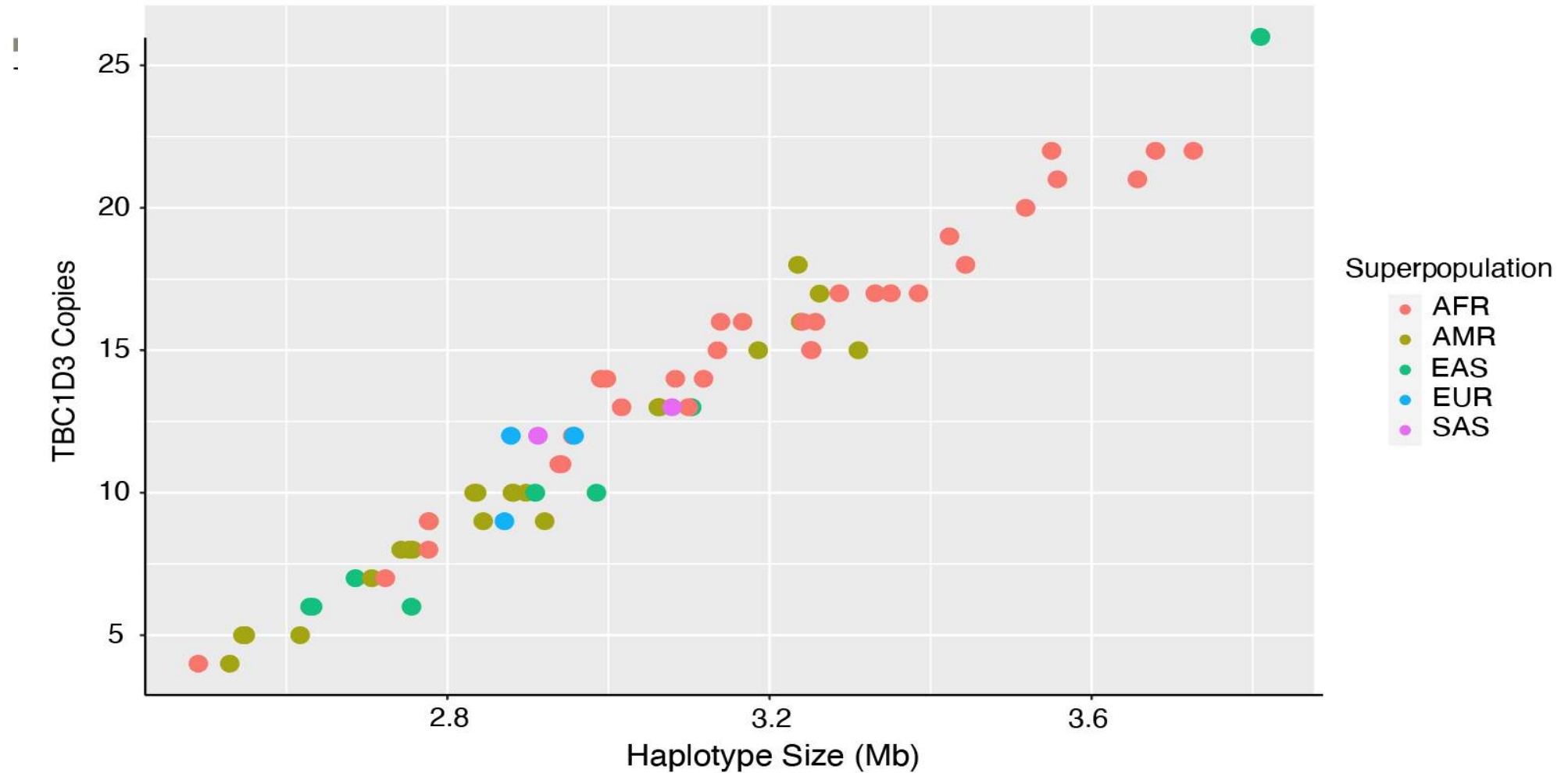


22q11 syndrome LCRA.

Complexity of normal human variation

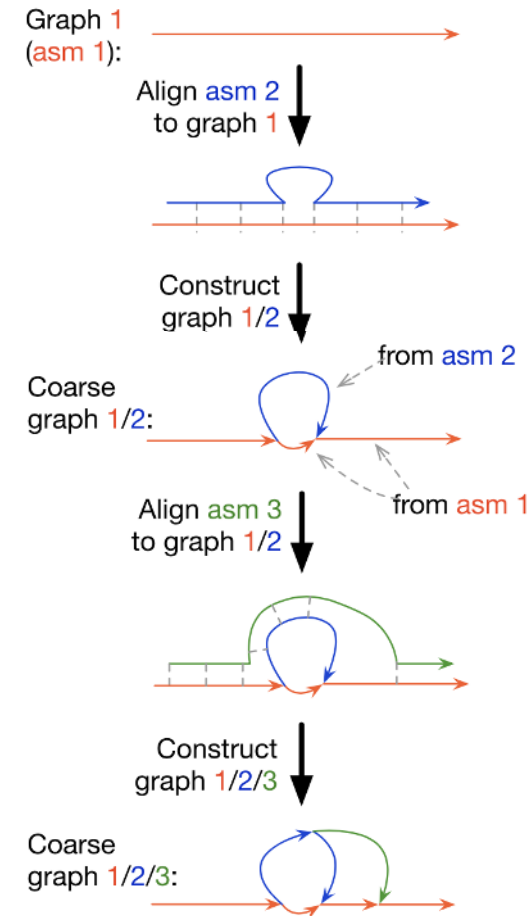


Copy number and structural variation of *TBC1D3*

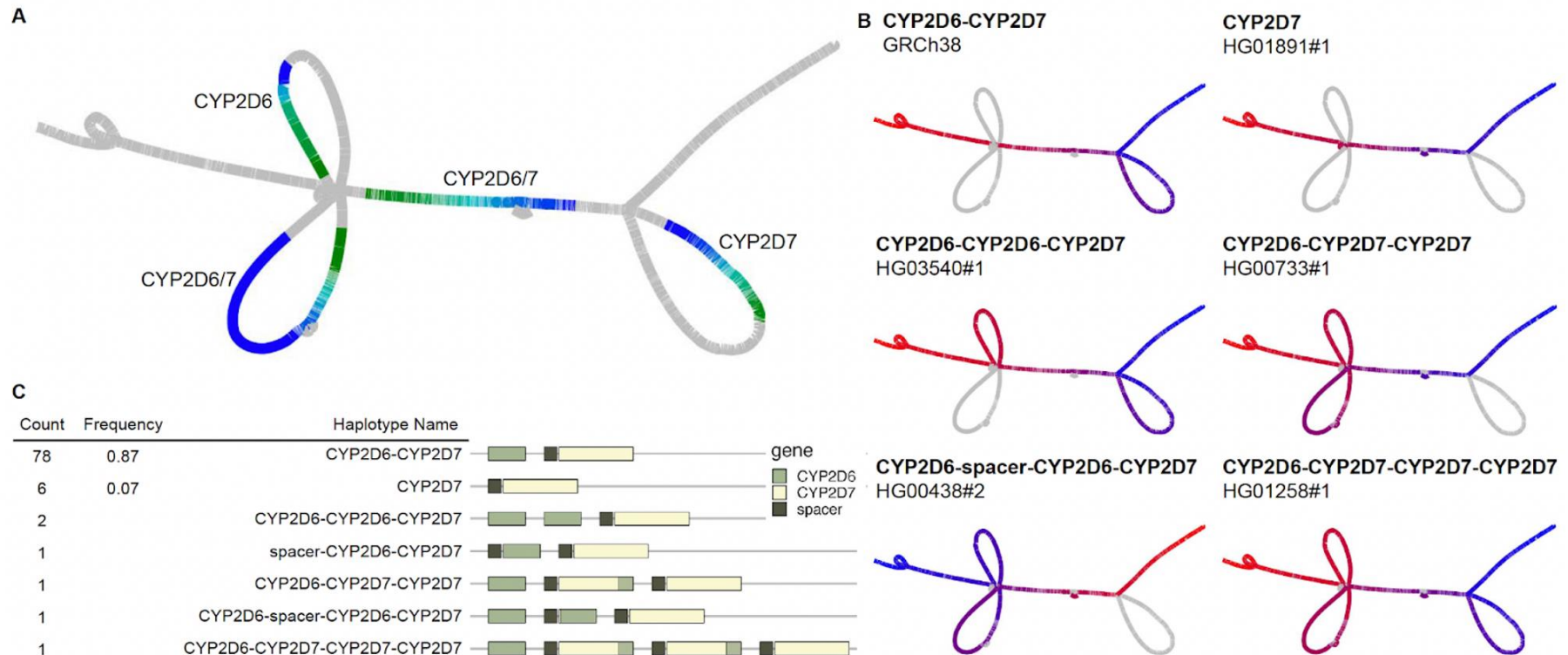


A graph can capture such variation e.g. Minigraph

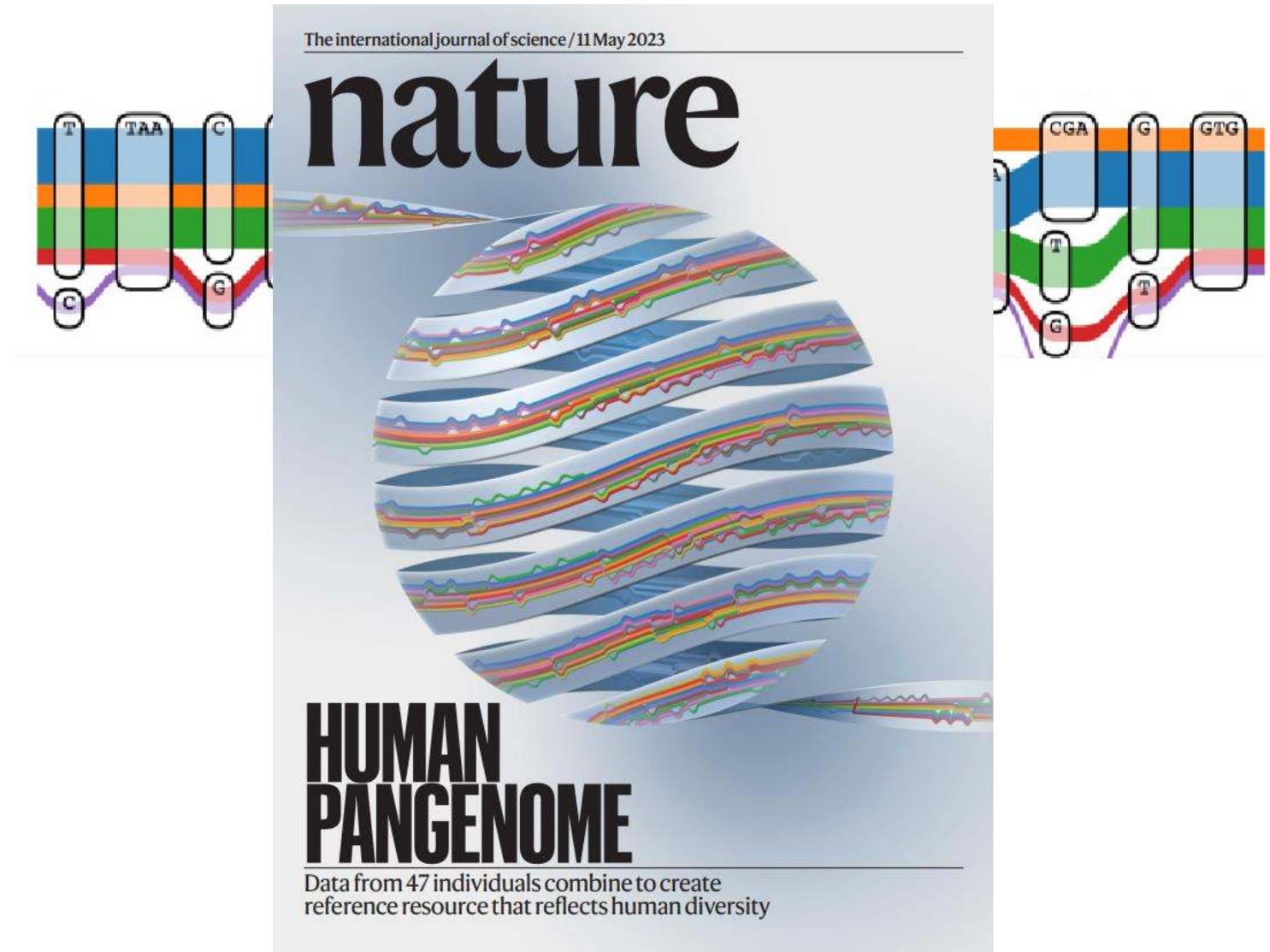
1. Generate phase genome assemblies
2. Iteratively introduce assembly sequence to a graph.
3. Distinguish query sequence already present in graph from novel sequence
4. Include novel sequence as new segments or edges between segments in graph.
5. Repeat with next assembly



A graph-based representation of structural variation



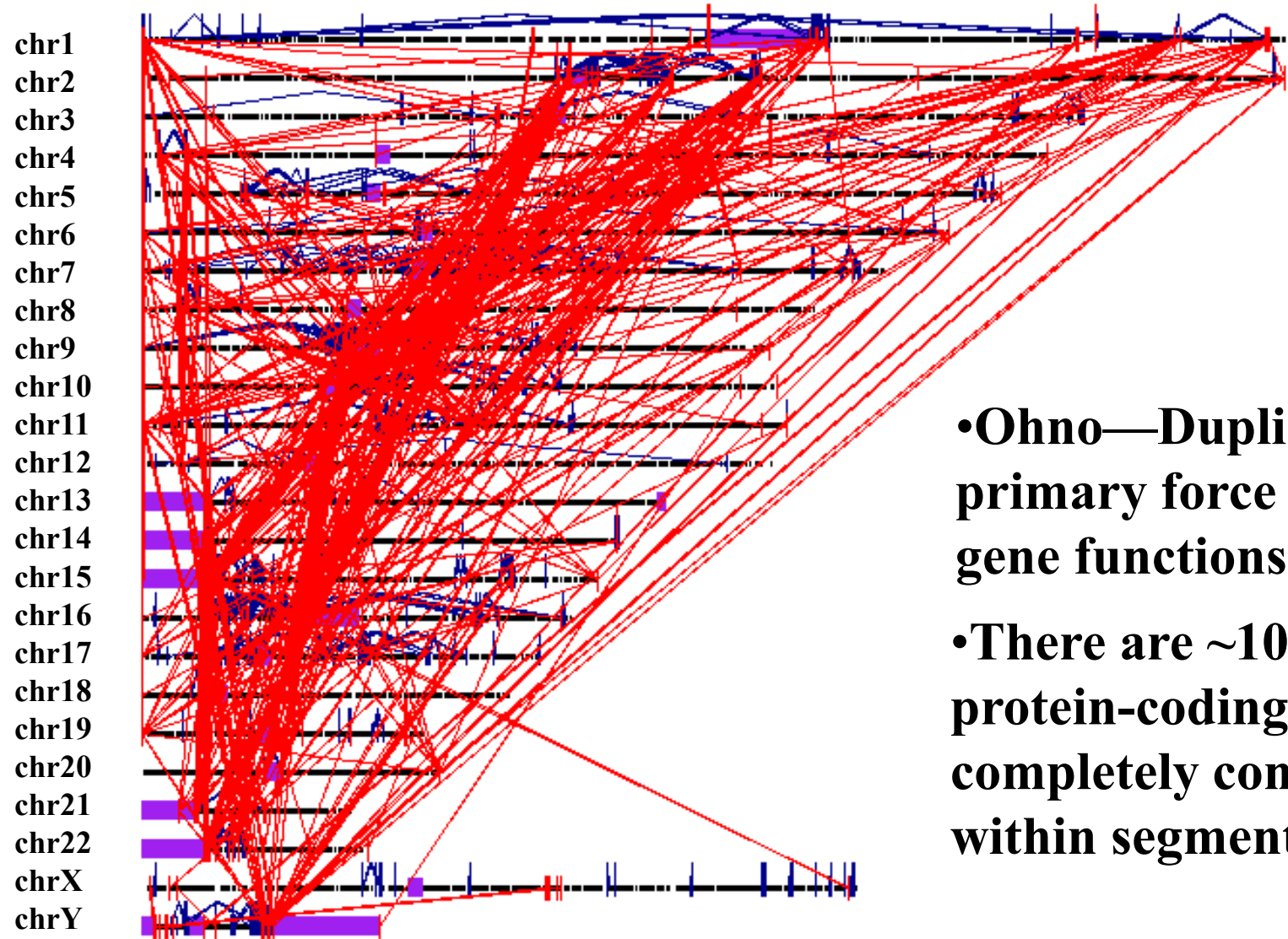
A graph-based representation of the entire human genome as a conceptual new reference.



Summary

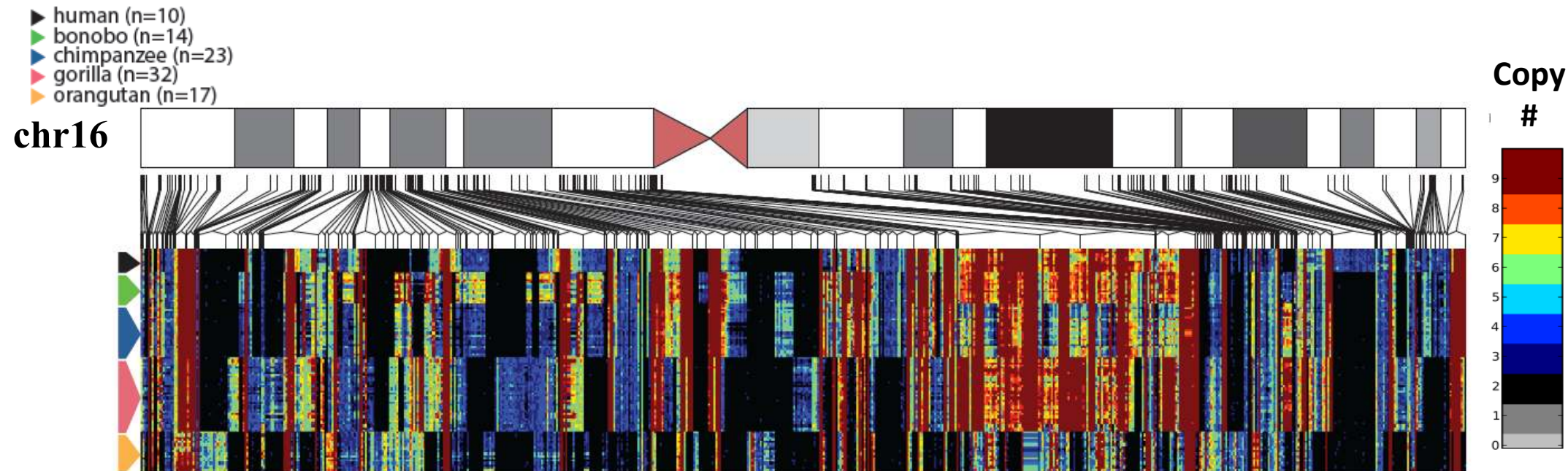
- Short read NGS approaches
 - Multiple methods are needed—readpair+read-depth+splitread often with orthogonal validation such as SNP microarray
 - ~75% of SVs are missed because SVs are non-randomly distributed to repetitive regions where mapping quality is low
 - Read-depth approaches allow CNV prediction 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?



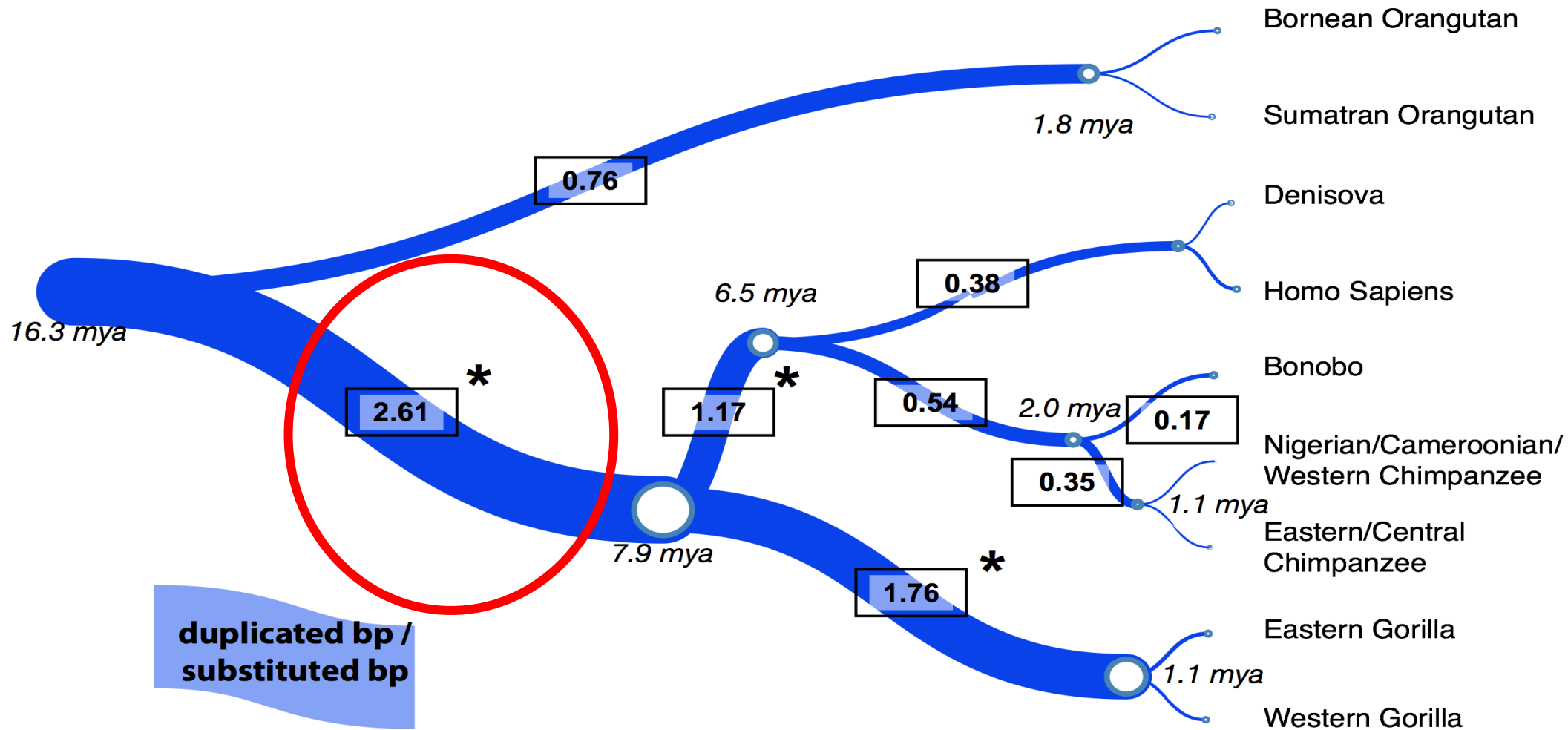
- **Ohno—Duplication is the primary force by which new gene functions are created**
- **There are ~1000 annotated protein-coding genes completely contained within segmental duplications**

Dynamic Genetic Variation



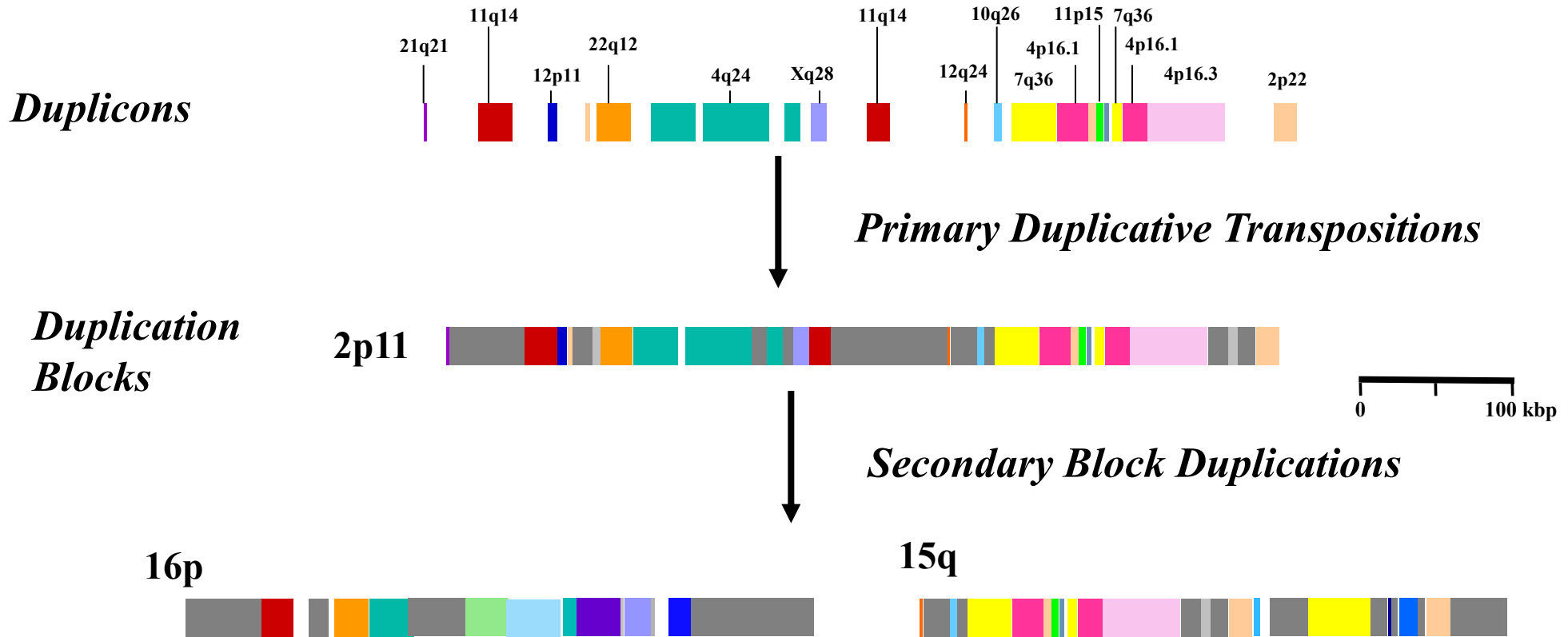
- Genomic copy number changes contributes more genetic difference between apes and humans than SNVs
- 468 Mbp CNV vs. 167 Mbp SNVs (ration: 2.8)

Rate of Duplication



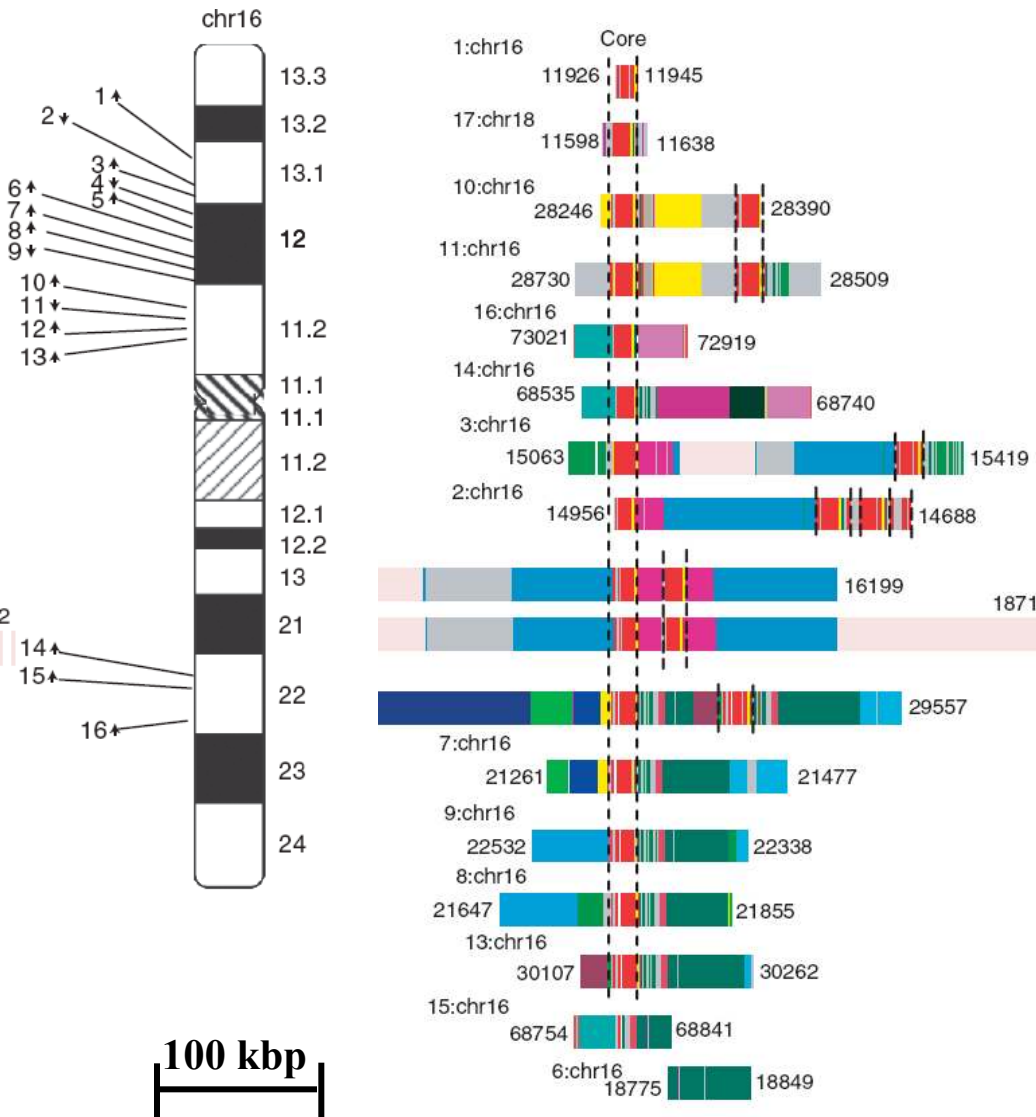
$p=9.786 \times 10^{-12}$

Mosaic Architecture



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

Human Chromosome 16 Core Duplicon

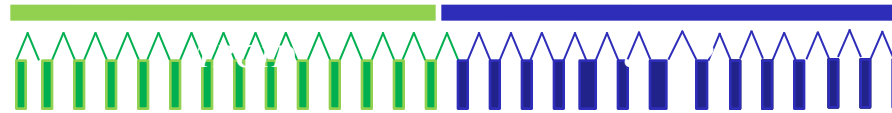


•The burst of segmental duplications 8-12 mya corresponds to core-associated duplications which have occurred on six human chromosomes (chromosomes 1,2, 7, 15, 16, 17)

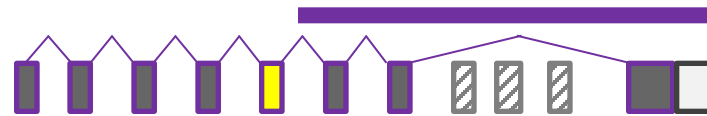
•Most of the recurrent genomic disorders associated with developmental delay, epilepsy, intellectual disability, etc. are mediated by duplication blocks centered on a core.

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

TRE2

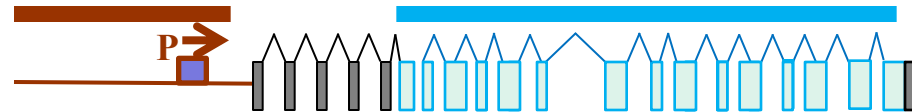


NPIP

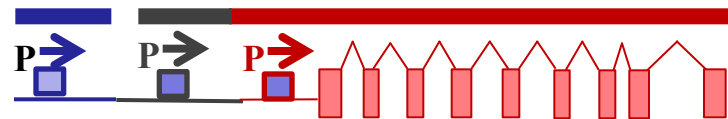


EVI5

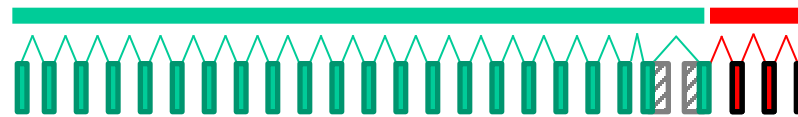
NBPF



LRRC37A



RGPD



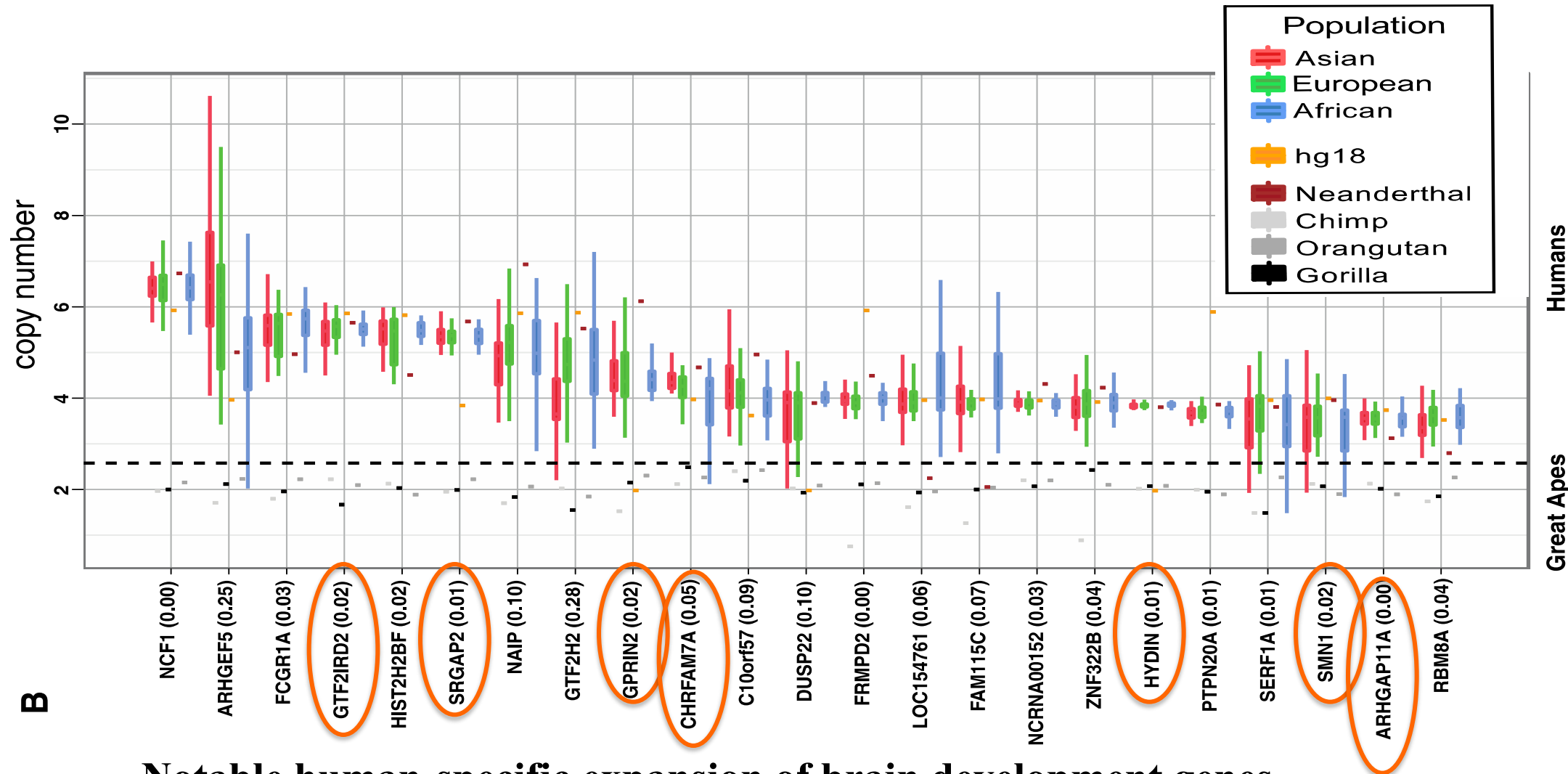
**Features: No orthologs in mouse; multiple copies in chimp & human
dramatic changes in expression profile; signatures of positive selection**

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



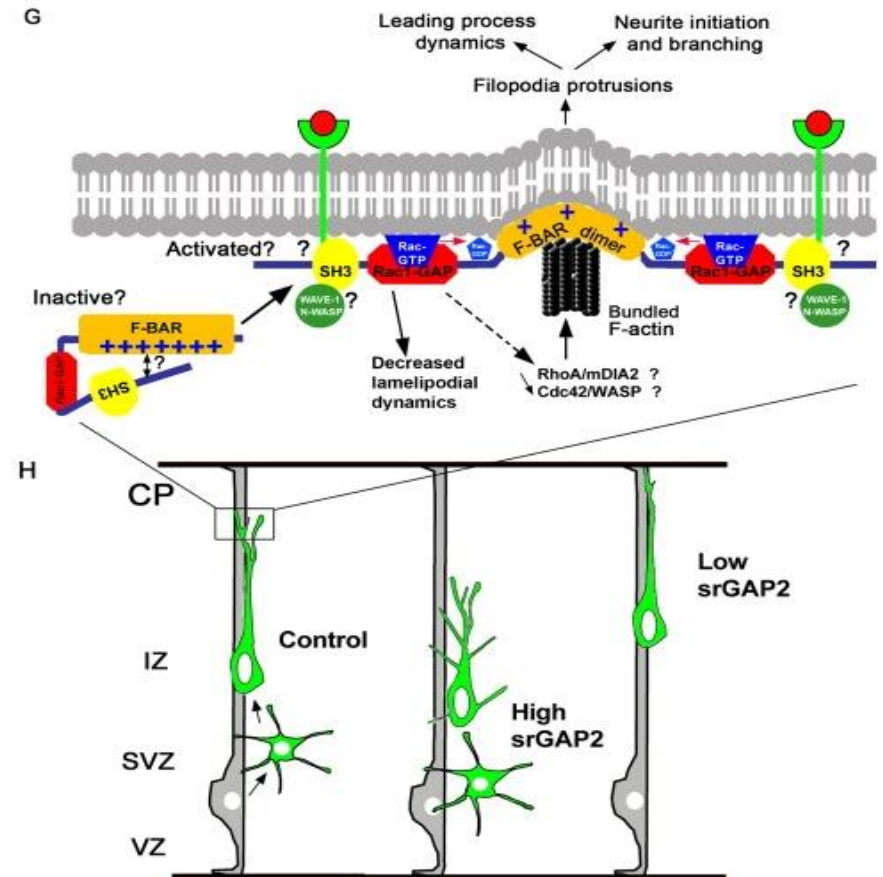
Notable human-specific expansion of brain development genes.

Neuronal cell death: $p=5.7e-4$; Neurological disease: $p=4.6e-2$

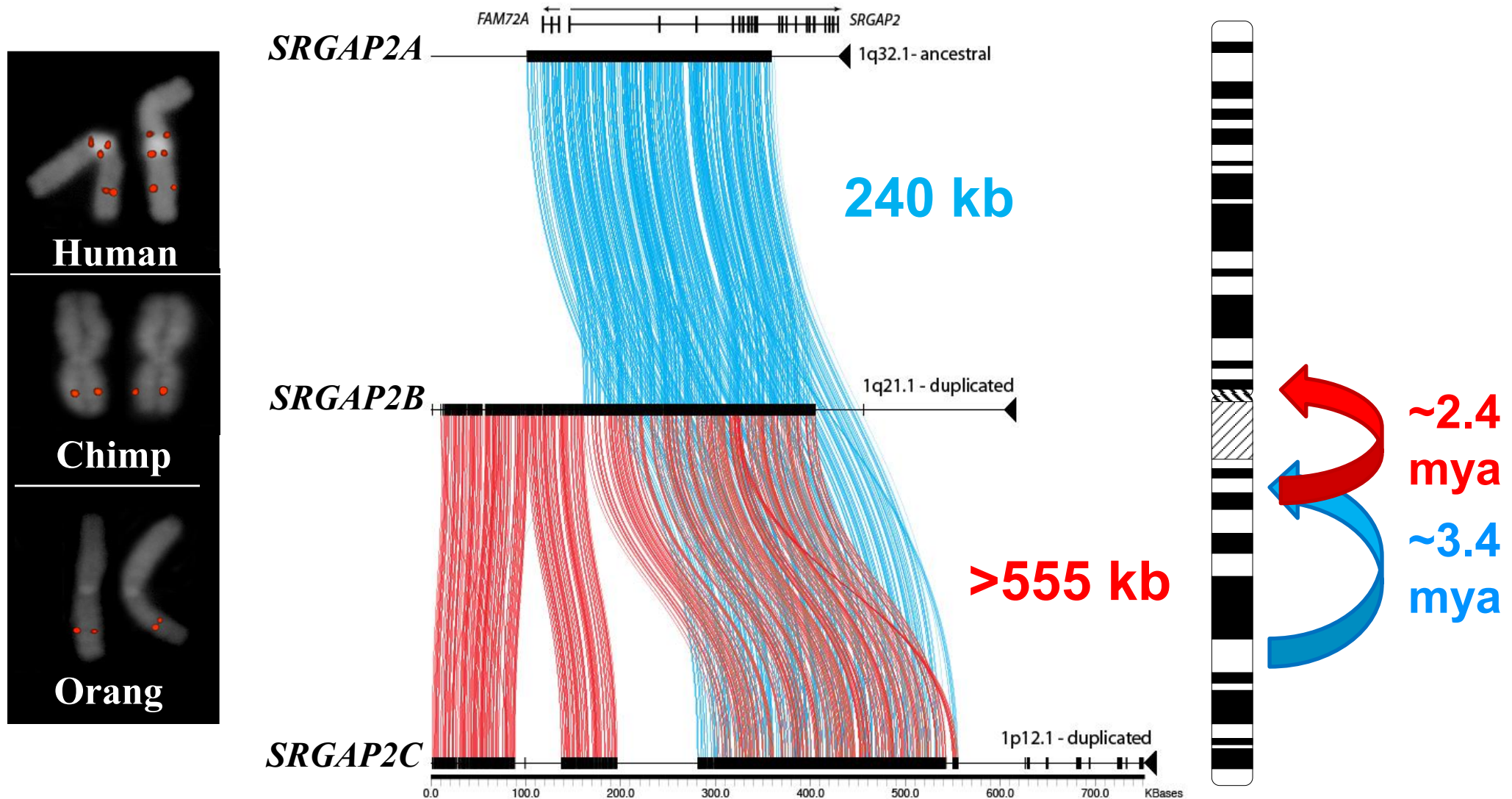
Sudmant et al., *Science*, 2010

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

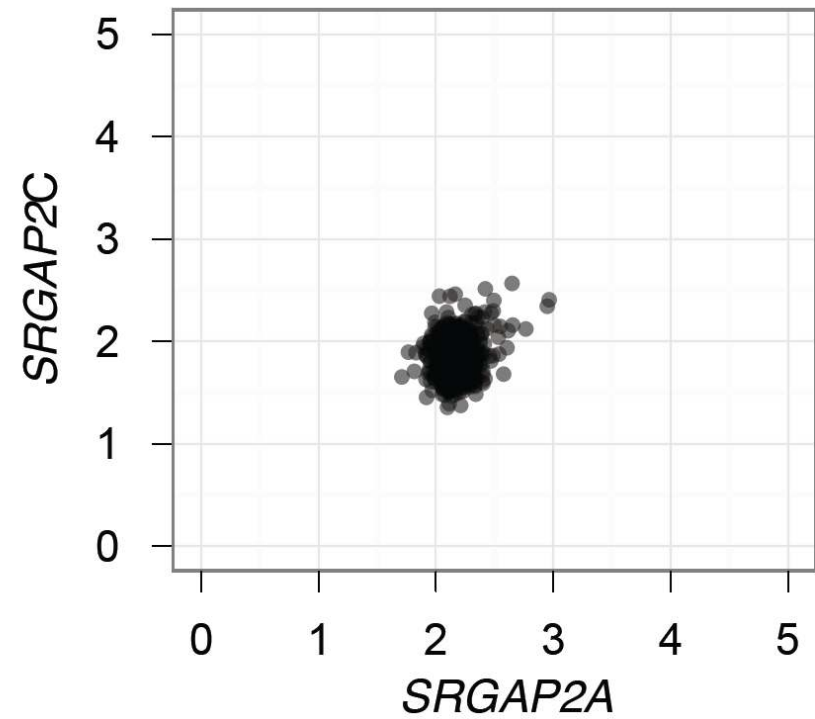
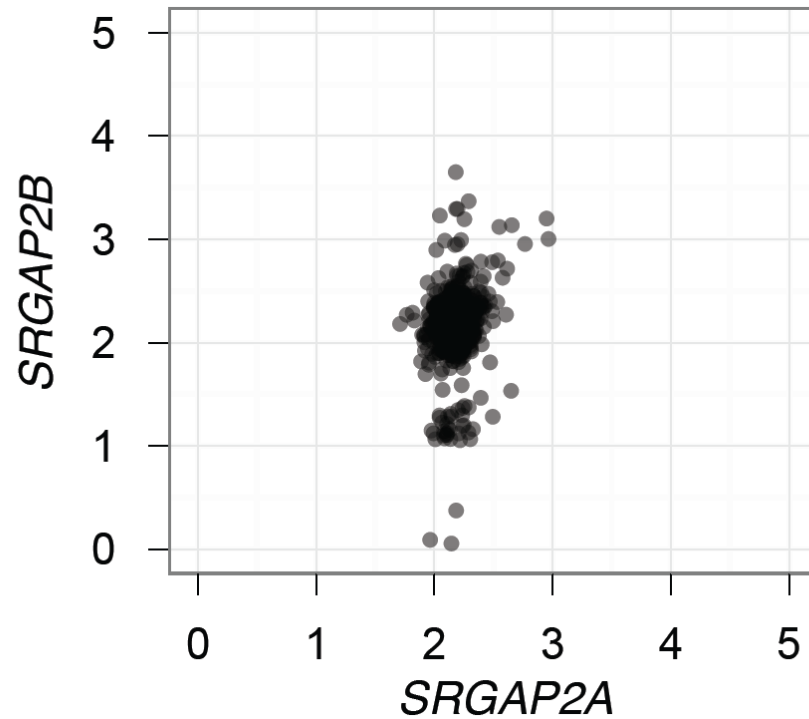


SRGAP2 Human Specific Duplication



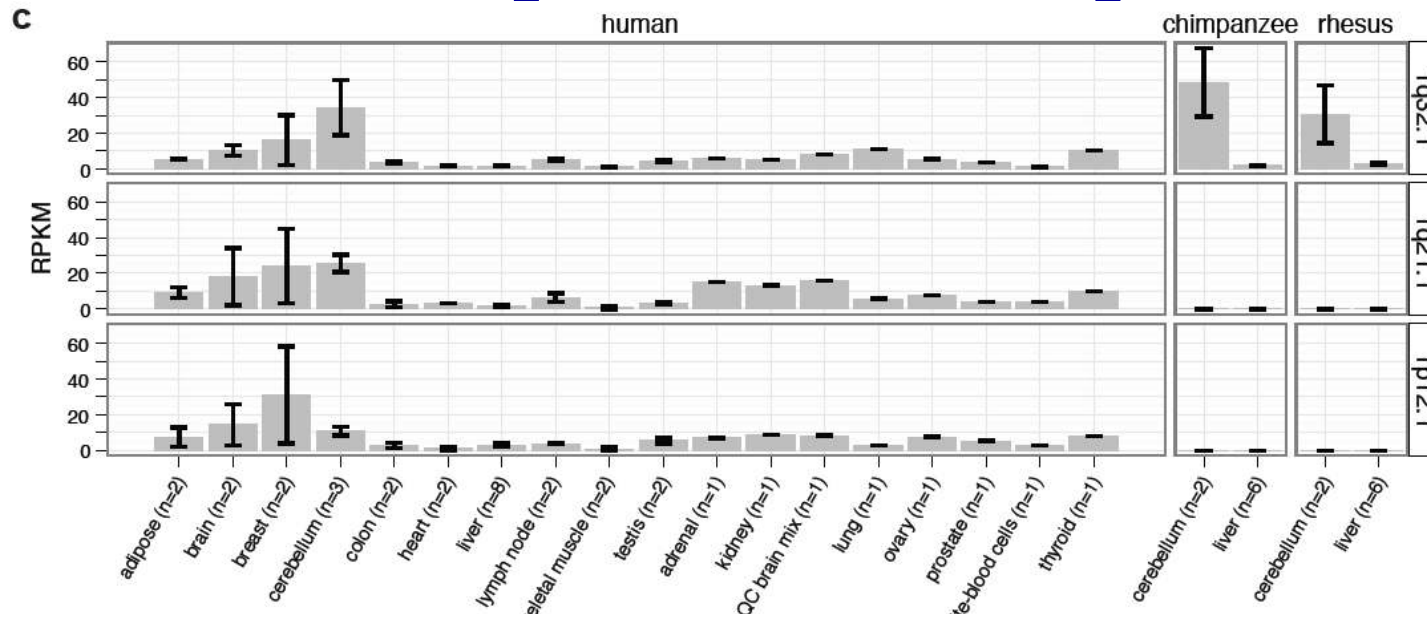
SRGAP2C is fixed in humans

(n=661 individual genomes)

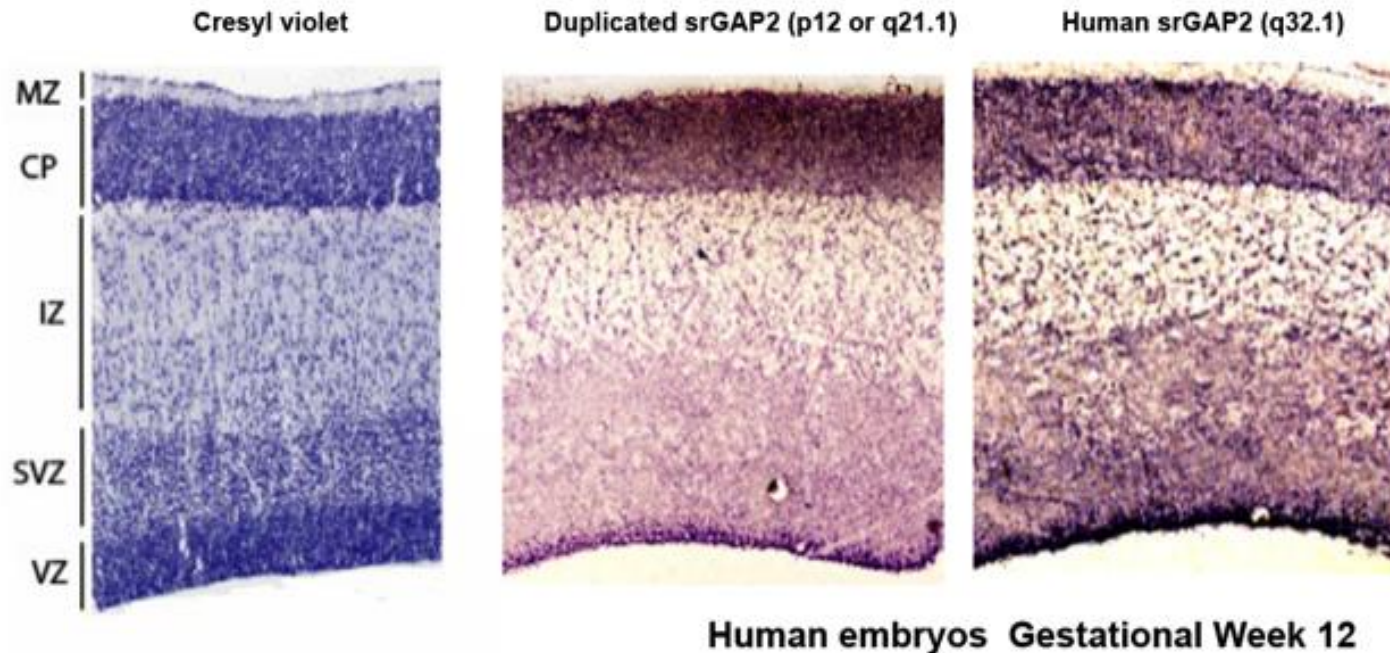


SRGAP2 duplicates are expressed

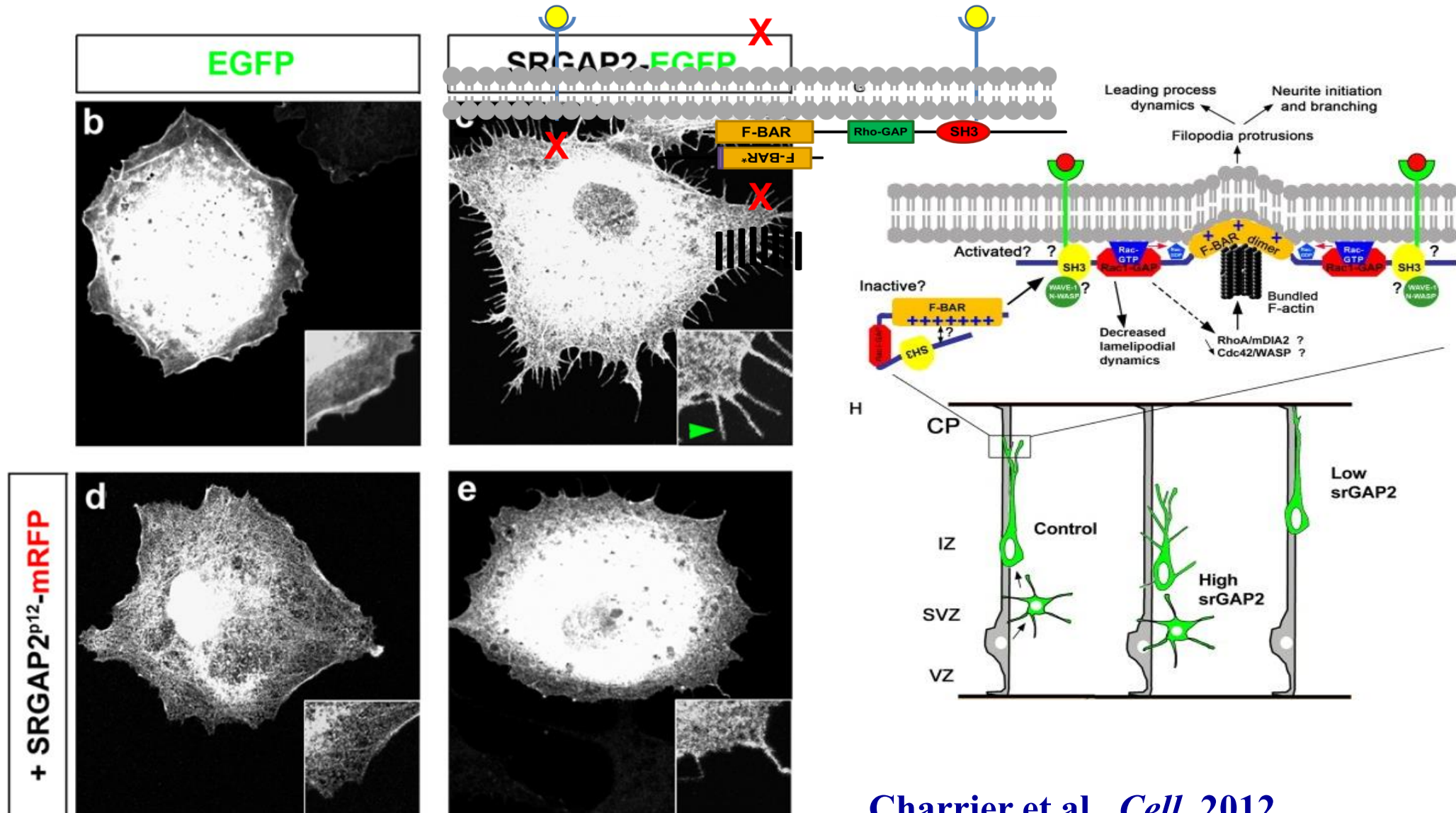
RNAseq

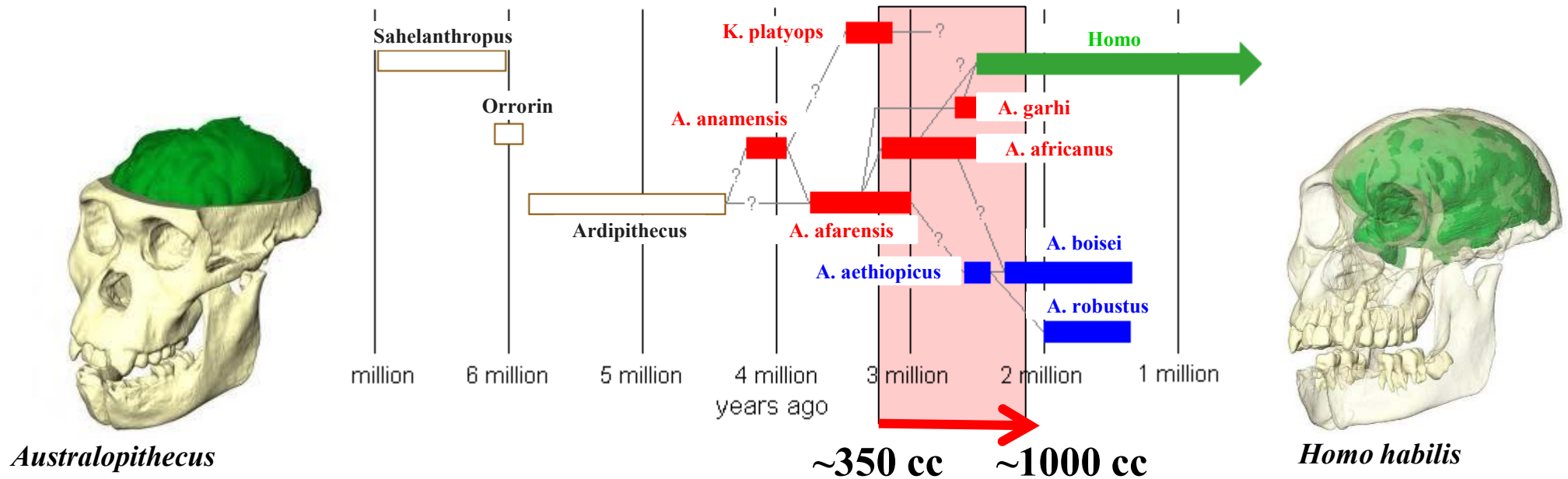
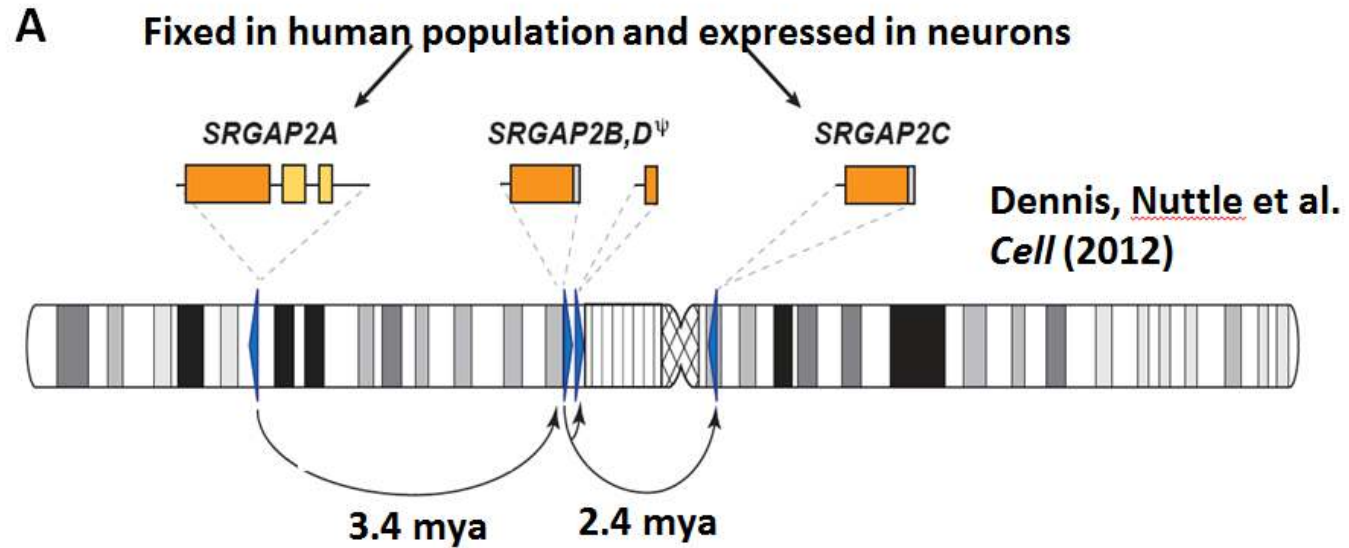


In situ



SRGAP2C duplicate antagonizes function

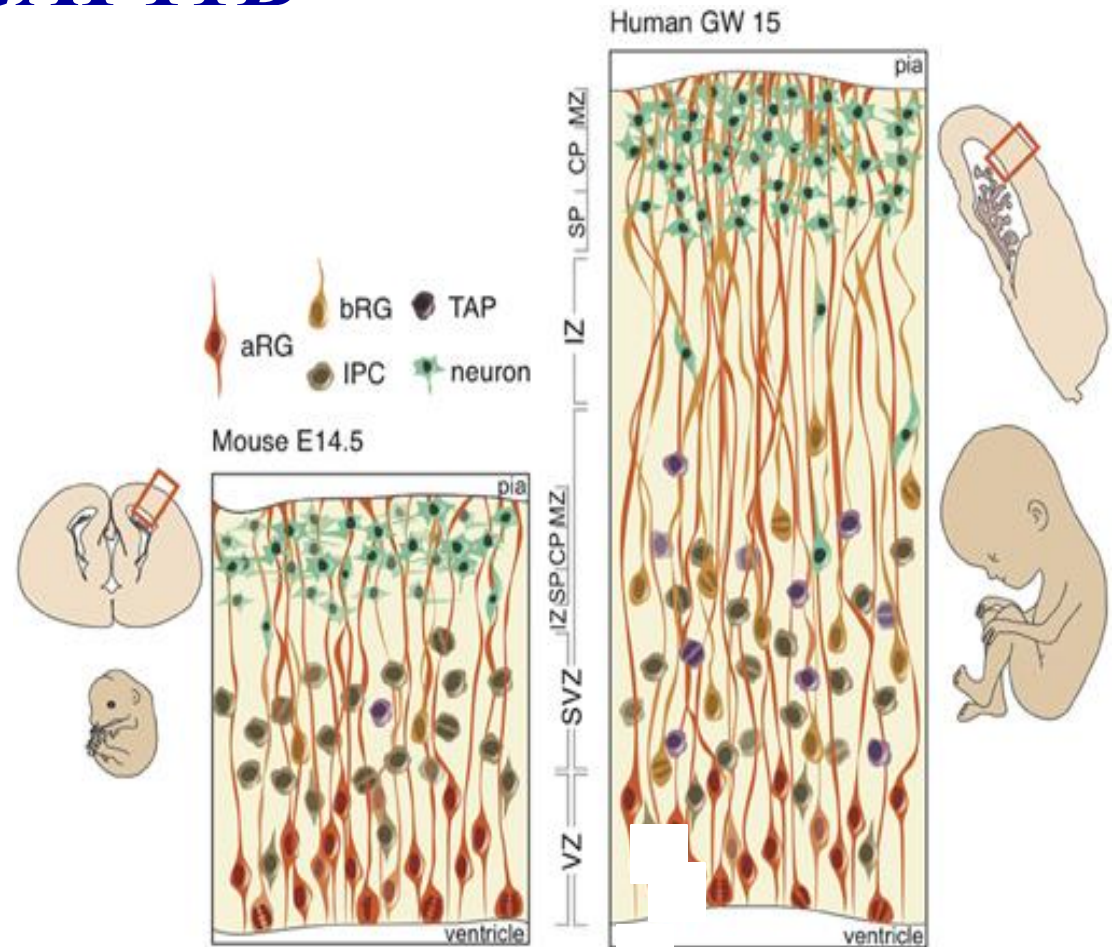




Example 2: Human-specific Duplication of *ARHGAP11B*

- 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

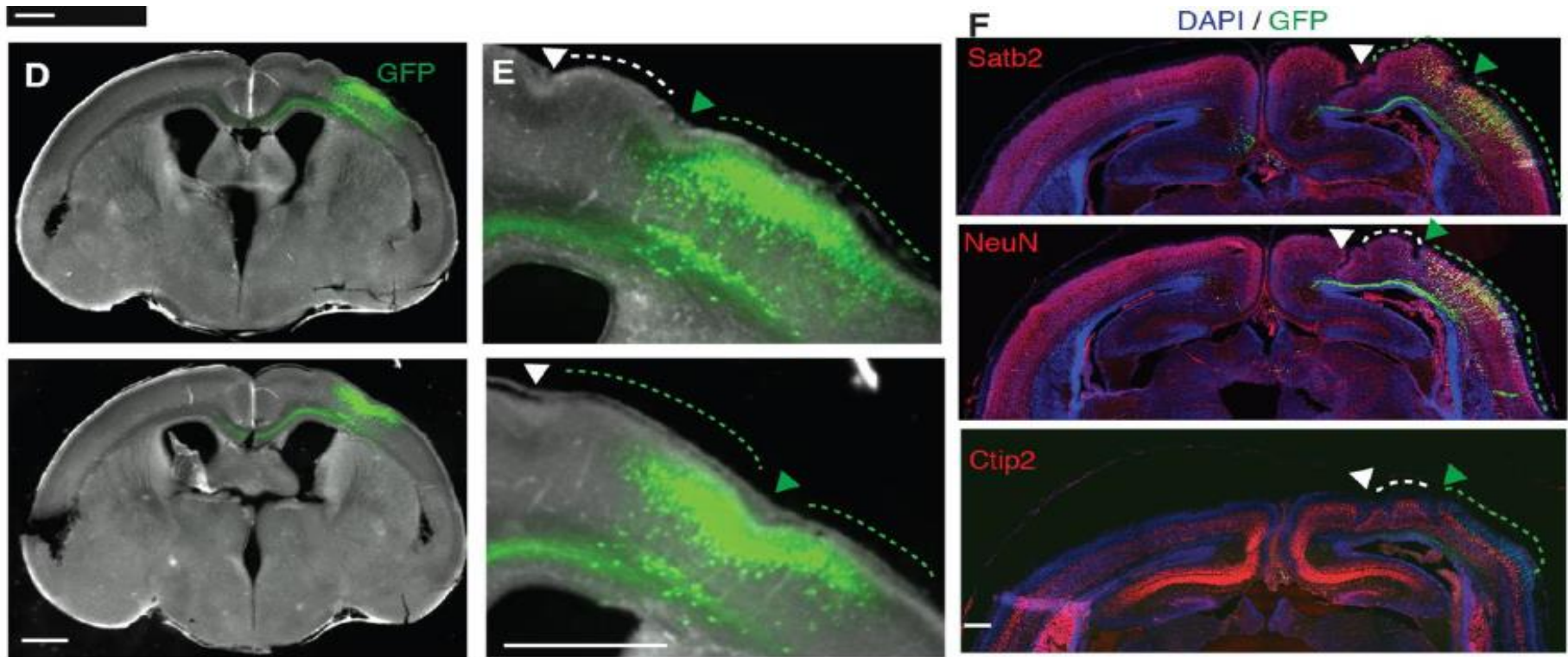
ARHGAP11B



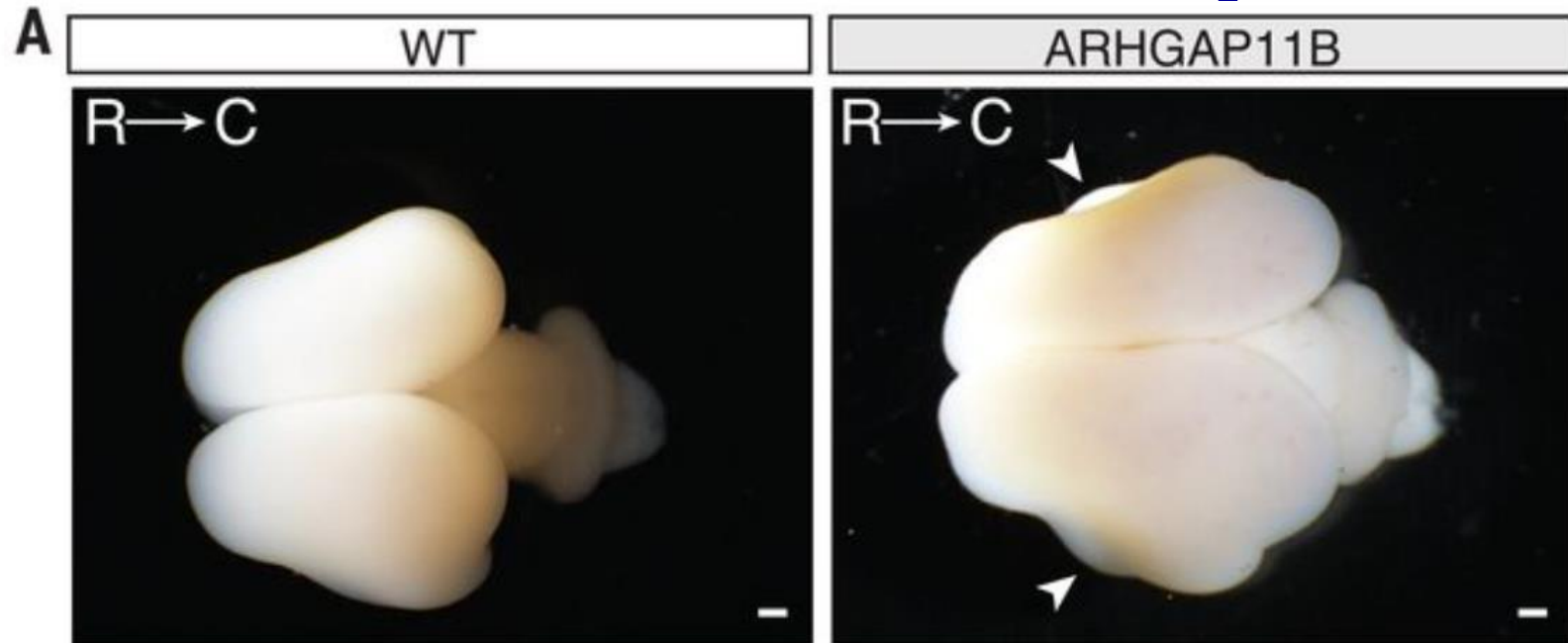
Florea *et al.*, *Science* 2015, Antonacci *et al.*, *Nat. Genet.*, 2014

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.



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

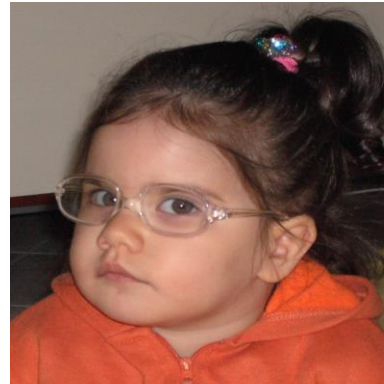
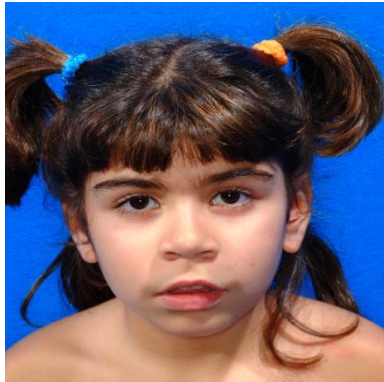


WT brain and brain expressing *ARHGAP11B* in neocortex (TG3). Arrowheads indicate cortical folds. R, rostral; C, caudal. Scale bars, 1 mm

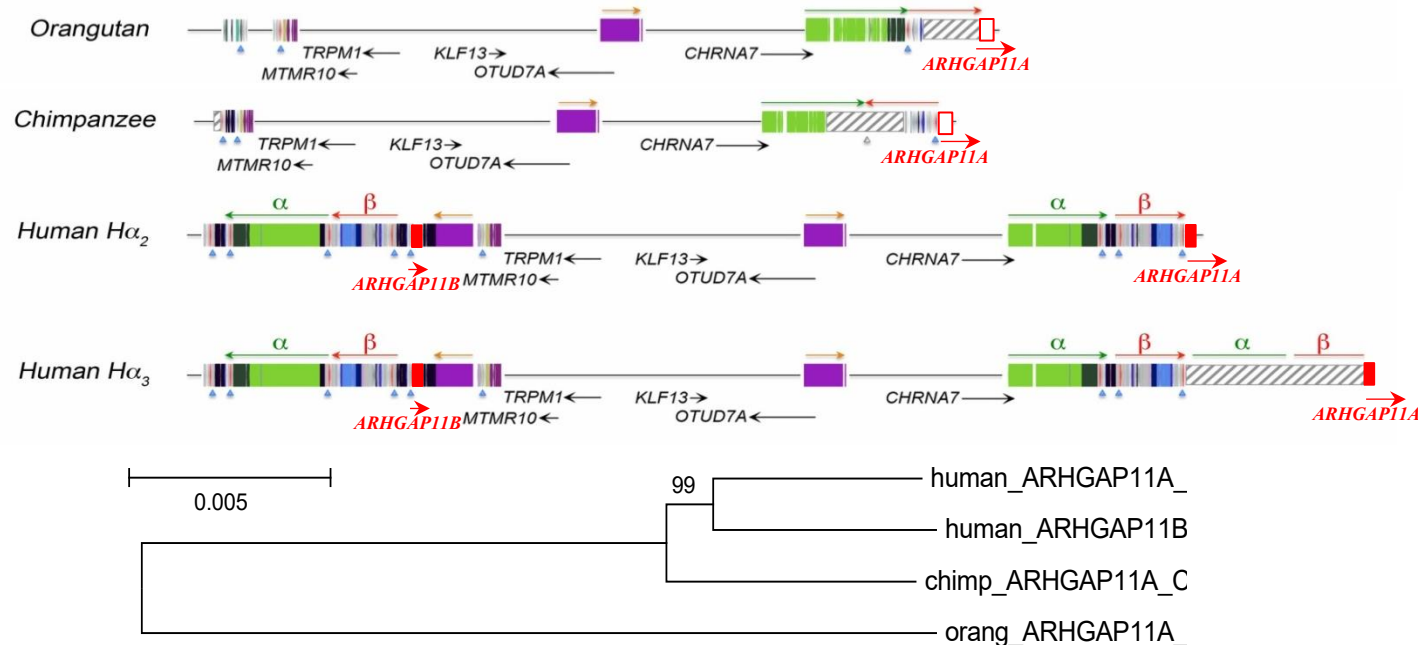
- Increased the numbers of basal radial glia progenitors in the marmoset outer subventricular zone, increased the numbers of upper-layer neurons, enlarged the neocortex, and induced its folding.

Heide et al., Science 2020

Duplication of *ARHGAP11B* and 15q13.3 Syndrome



Chromosome 15

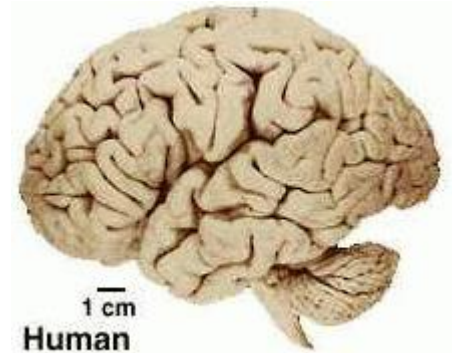


Duplication from *ARHGAP11A* to *ARHGAP11B* estimated to have occurred 5.3 +/- 0.5 million years ago.

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 ([Nuttall 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.**



Genetic differences that make us uniquely human



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; *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 paired-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

