Genome Structural Variation

Evan Eichler
Howard Hughes Medical Institute
University of Washington

January 11th, 2020, Genomics Workshop, Český Krumlov

Genetic Variation

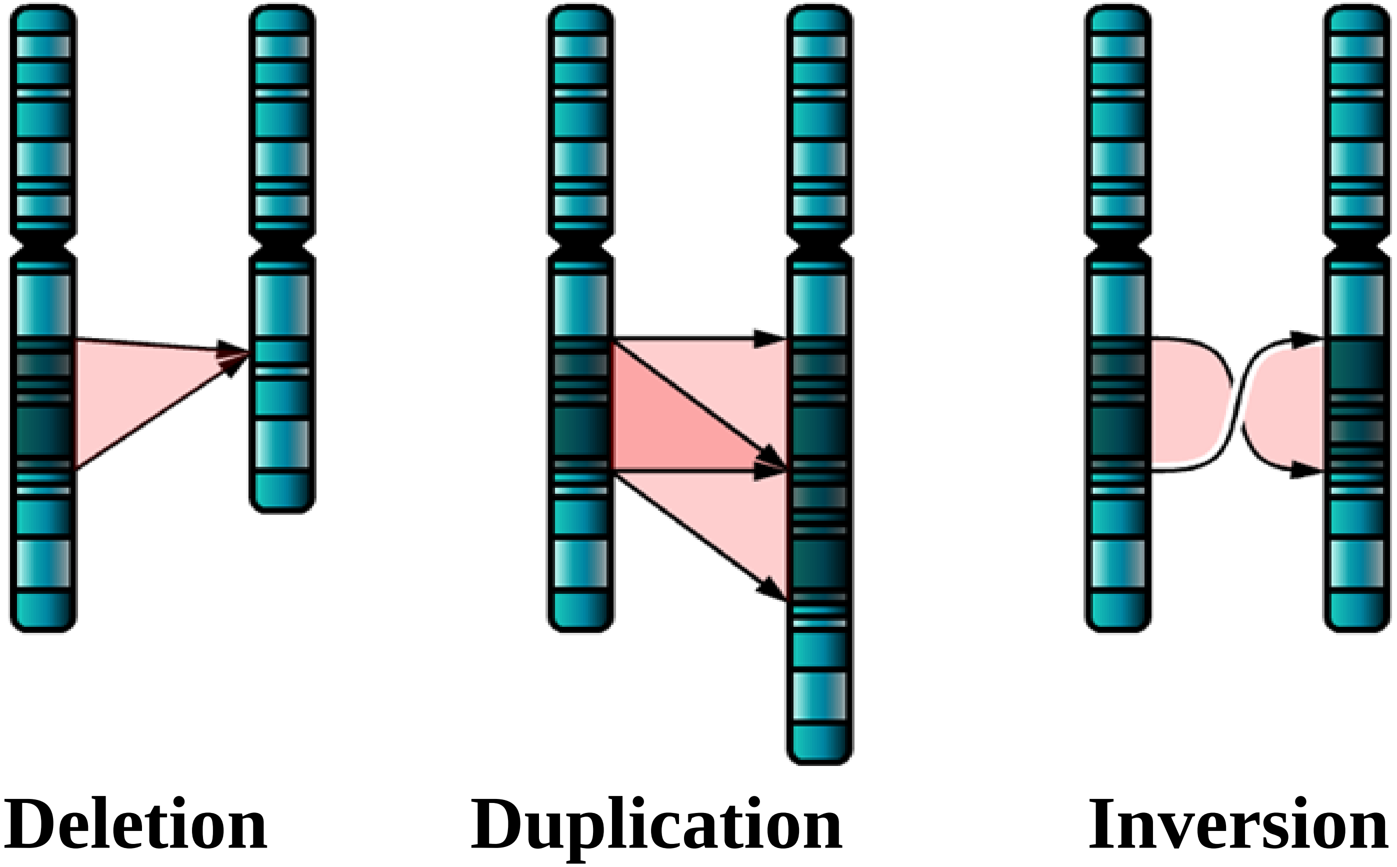
Types

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



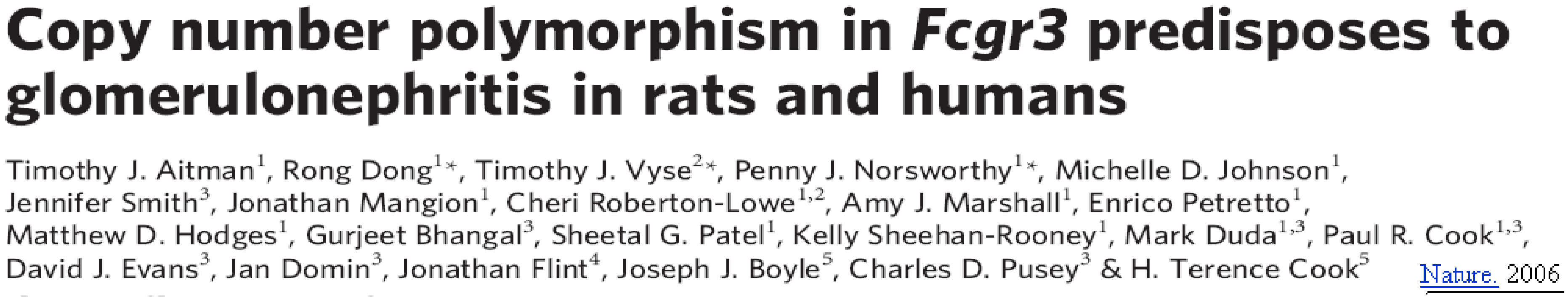
Deletion

Duplication

Inversion

Introduction

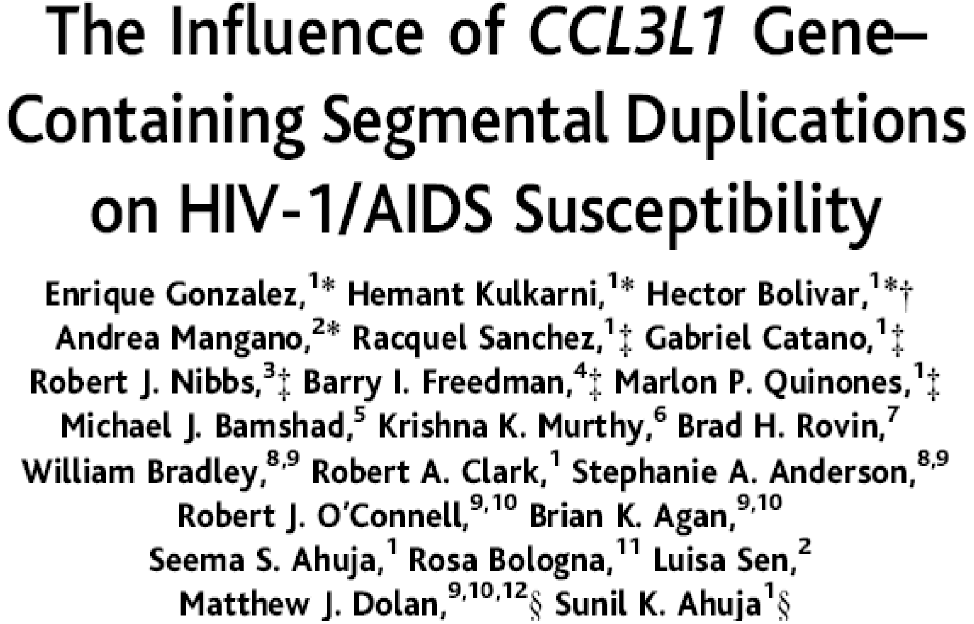
- **Genome structural variation :** gains and losses of DNA (copy-number variation (CNV)) as well as balanced events such as inversions and translocations—operationally defined >50 bp
- **Objectives**
  1. Genomic architecture and disease impact.
  2. Detection and characterization methods
  3. Primate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



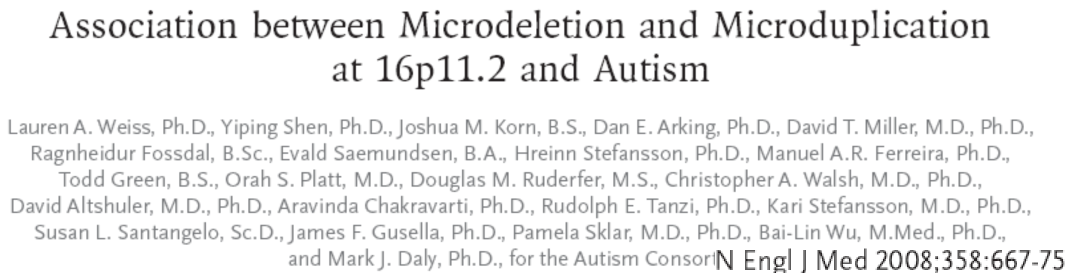
The Influence of *CCL3L1* Gene–Containing Segmental Duplications on HIV-1/AIDS Susceptibility

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



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

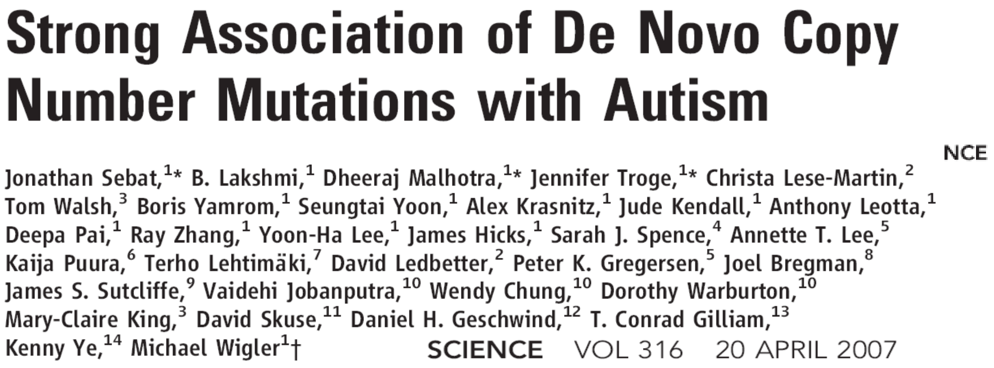
Lauren A. Weiss, Ph.D., Yiping Shen, Ph.D., Joshua M. Korn, B.S., Dan E. Arking, Ph.D., David T. Miller, M.D., Ph.D., Ragnheidur Fossdal, B.Sc., Evald Saemundsen, B.A., Hreinn Stefansson, Ph.D., Manuel A.R. Ferreira, Ph.D., Todd Green, B.S., Orah S. Platt, M.D., Douglas M. Ruderfer, M.S., Christopher 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

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



Strong Association of De Novo Copy Number Mutations with Autism

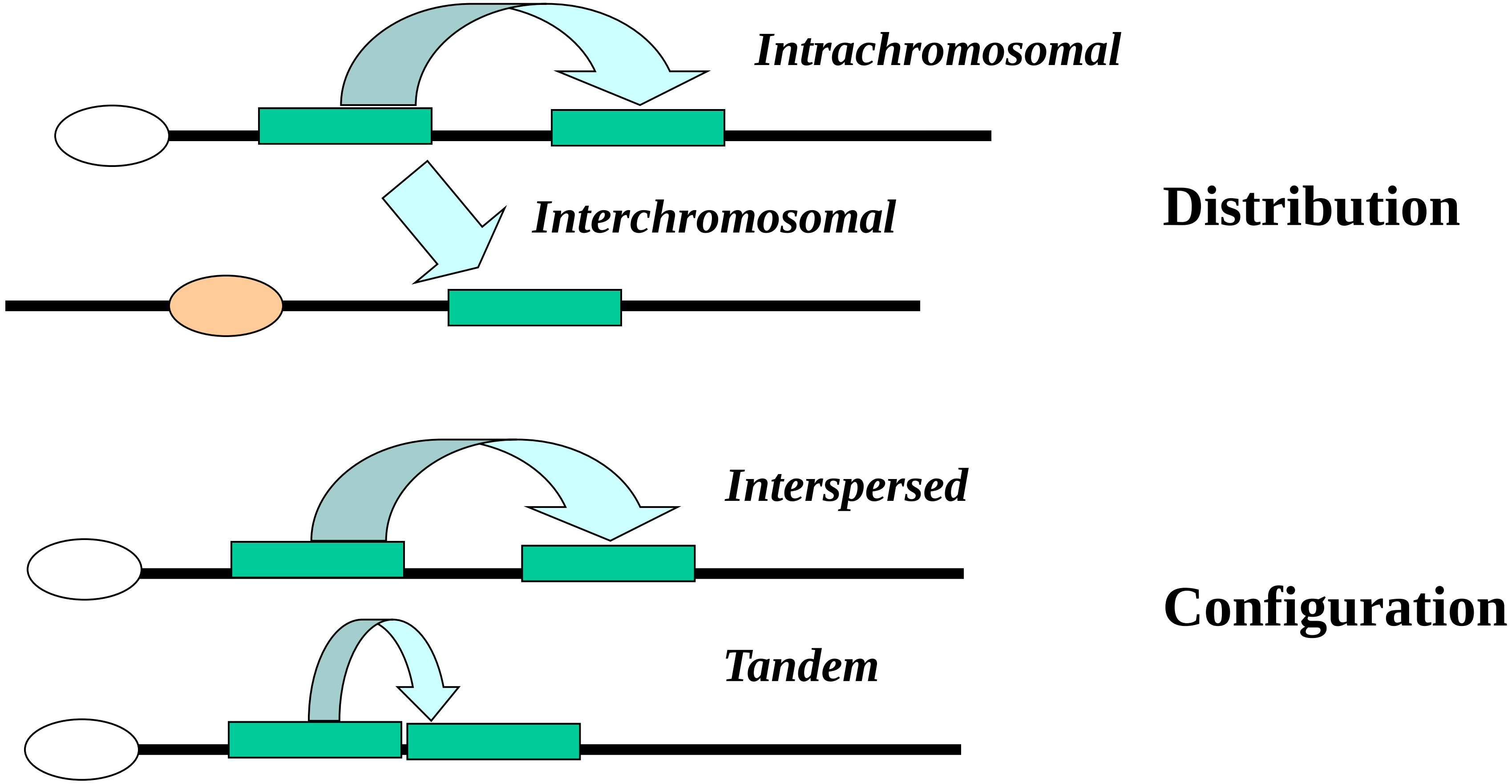
NCE

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†]

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



Intrachromosomal

Interchromosomal

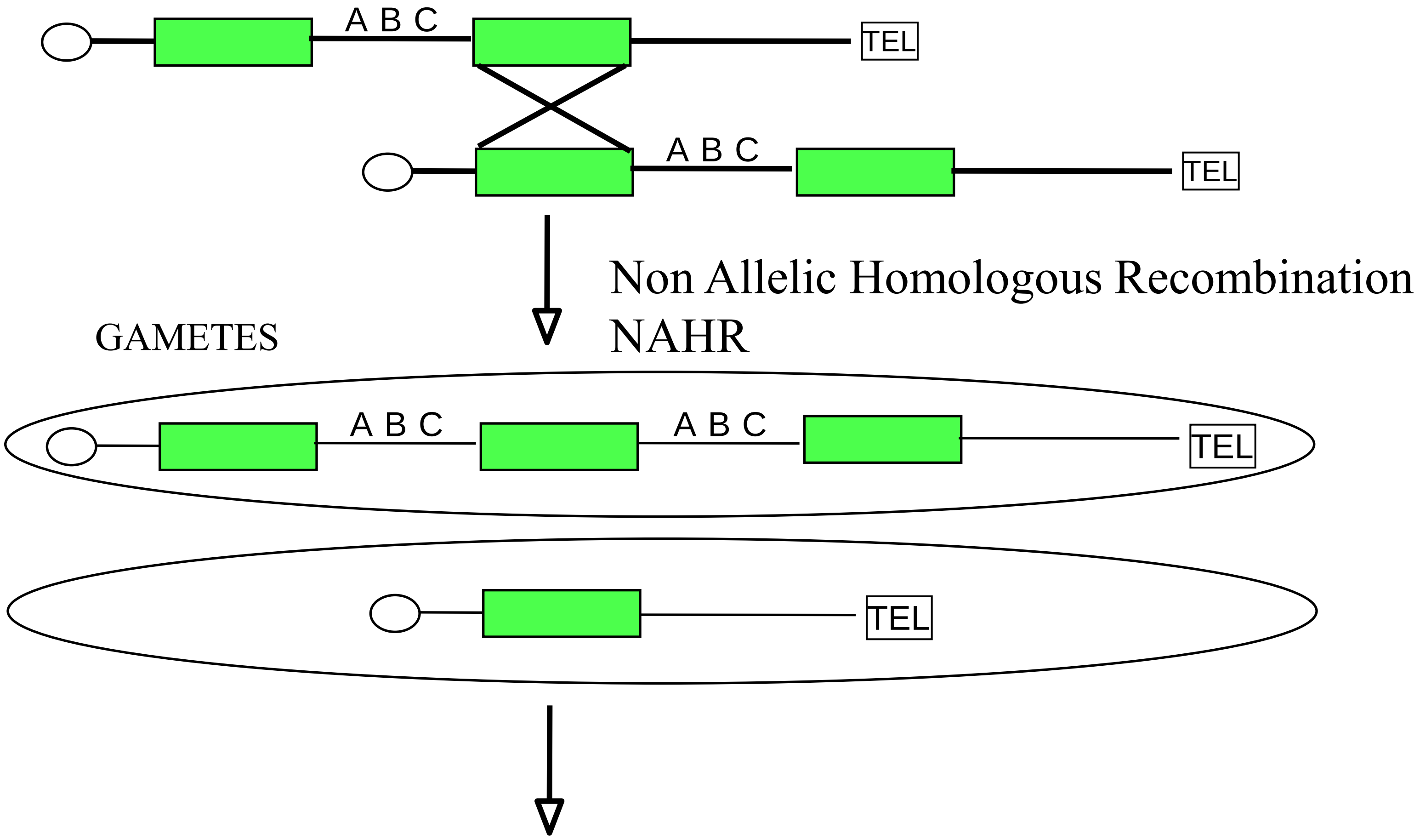
Distribution

Interspersed

Tandem

Configuration

Importance:
SDs promote Structural Variation



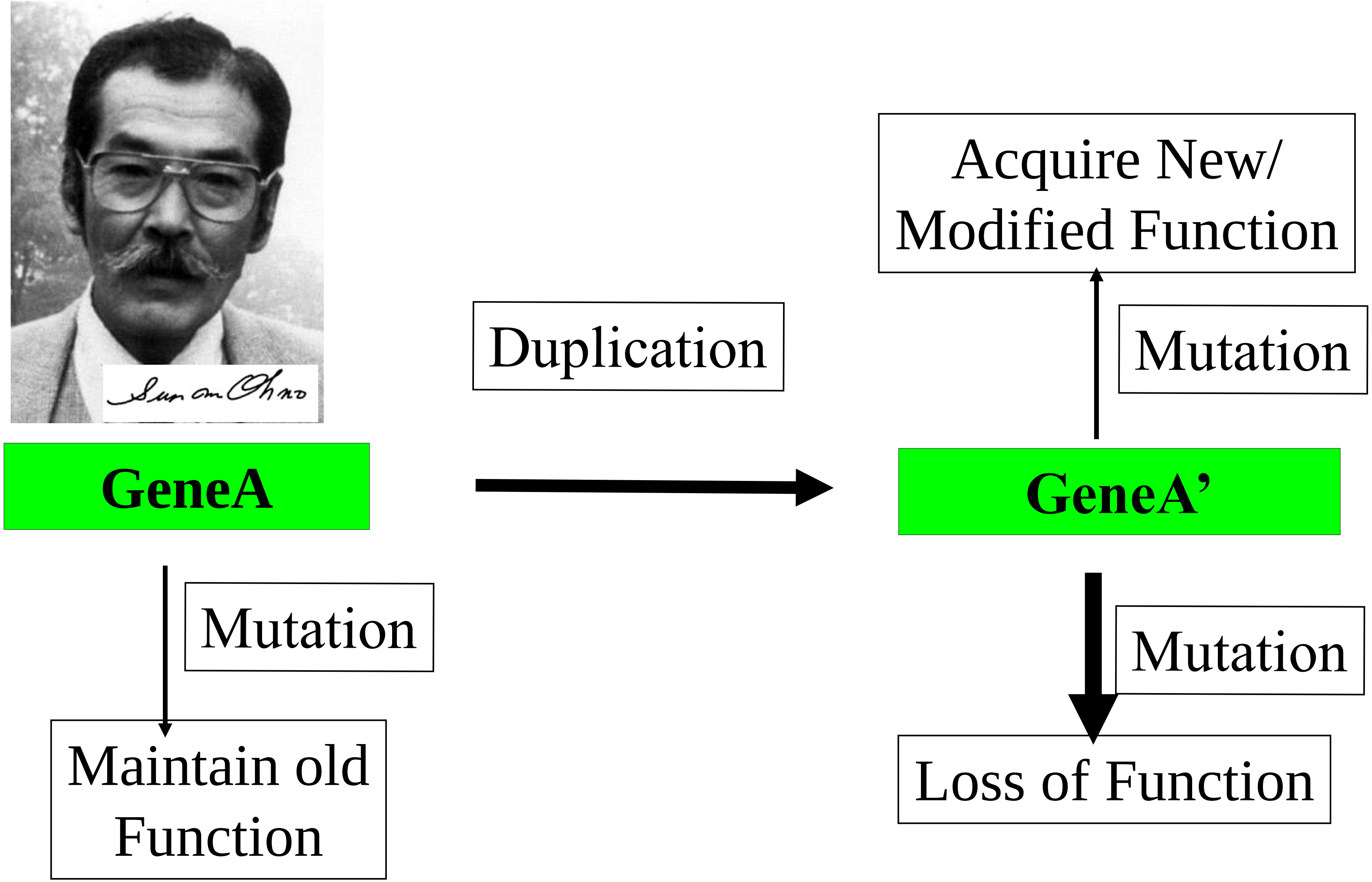
Non Allelic Homologous Recombination
NAHR

GAMETES

Human Disease

Triplosensitive, Haploinsufficient and Imprinted Genes

Importance: Evolution of New Gene Function



Duplication

GeneA

Mutation

Maintain old Function

GeneA’

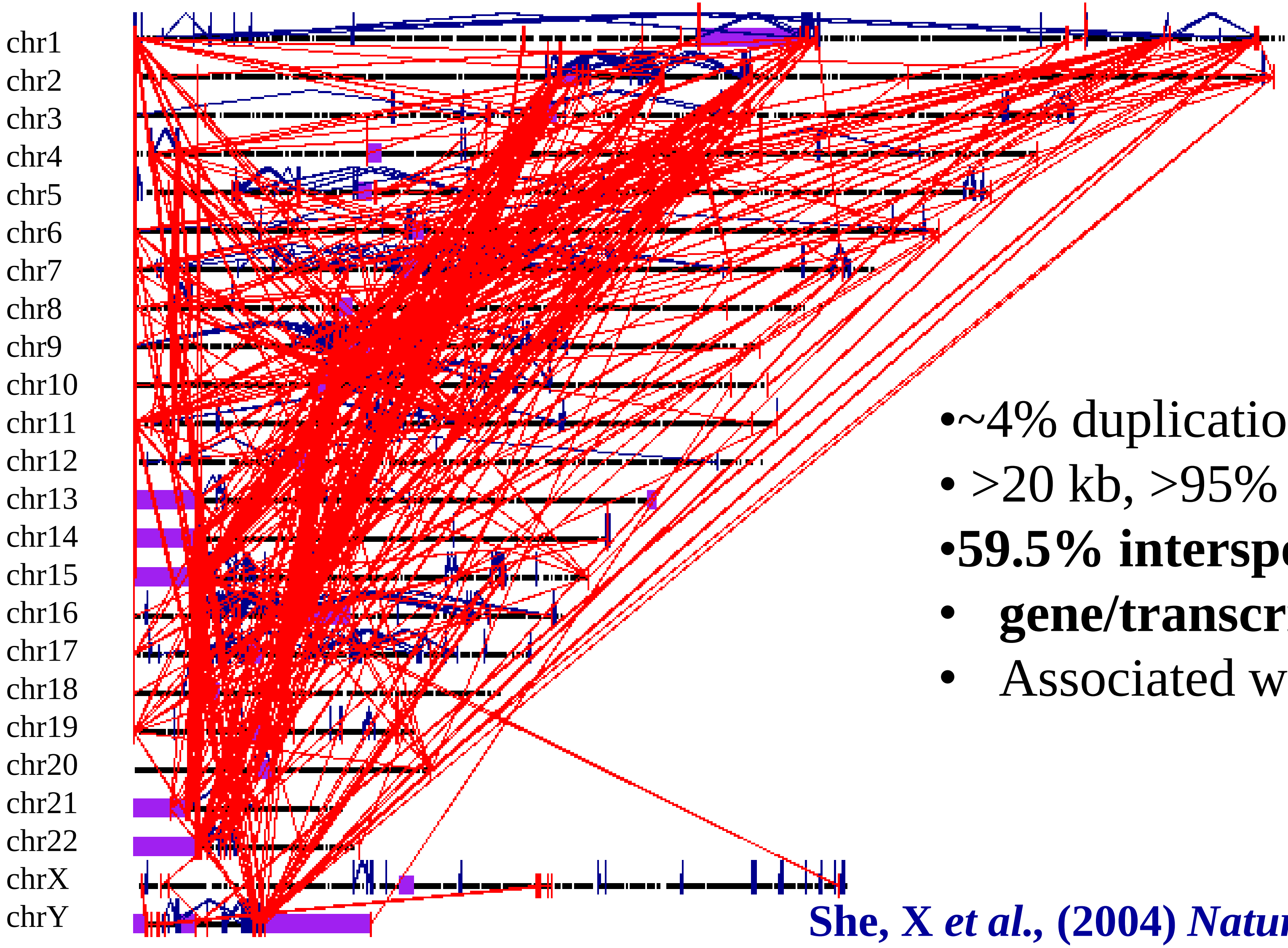
Mutation

Acquire New/ Modified Function

Mutation

Loss of Function

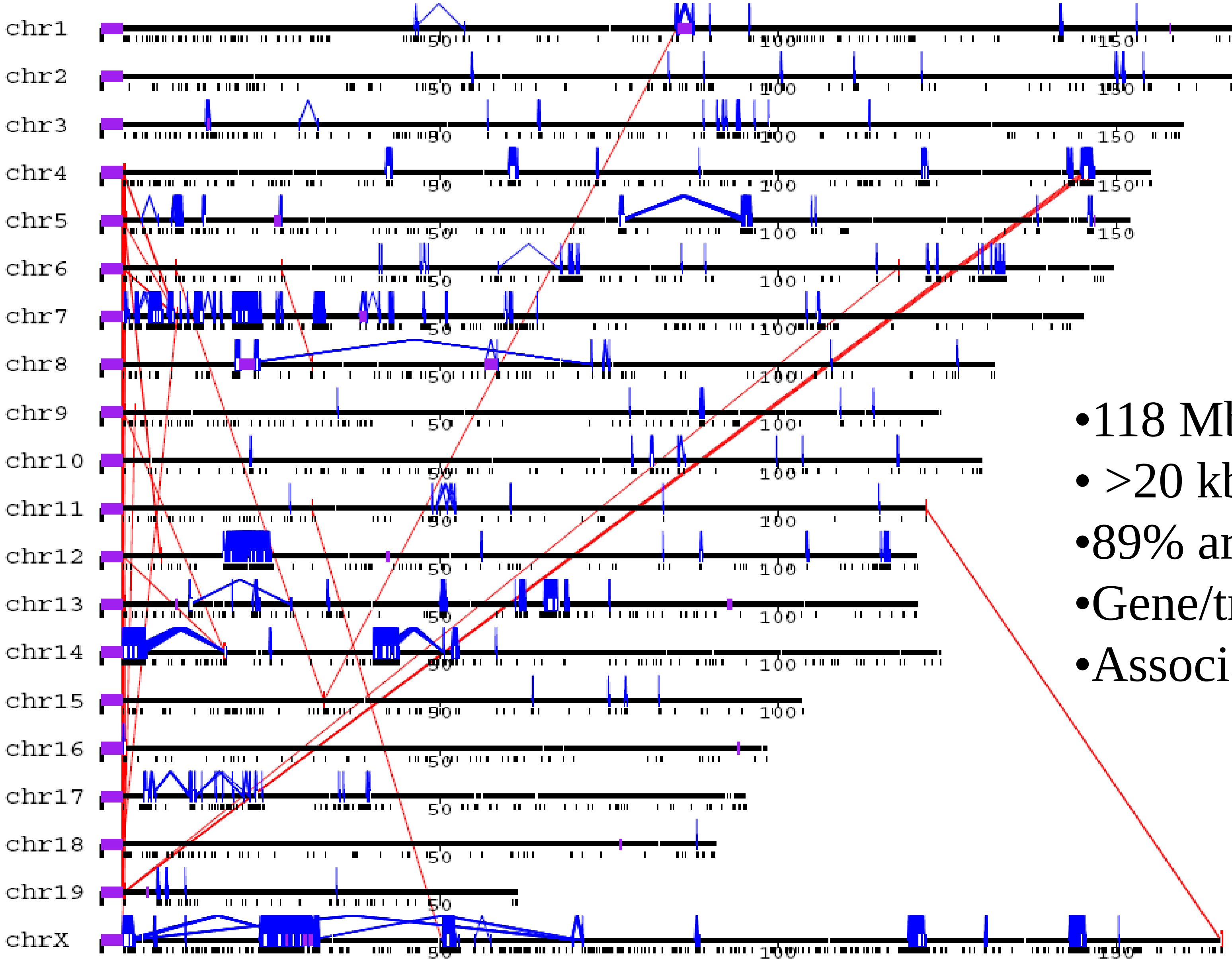
I. Human Genome Segmental Duplication Pattern



- •~4% duplication (125 Mb)
- >20 kb, >95%
- **•59.5% interspersed**
- **gene/transcript rich**
- Associated with Alu repeats

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

Mouse Segmental Duplication Pattern



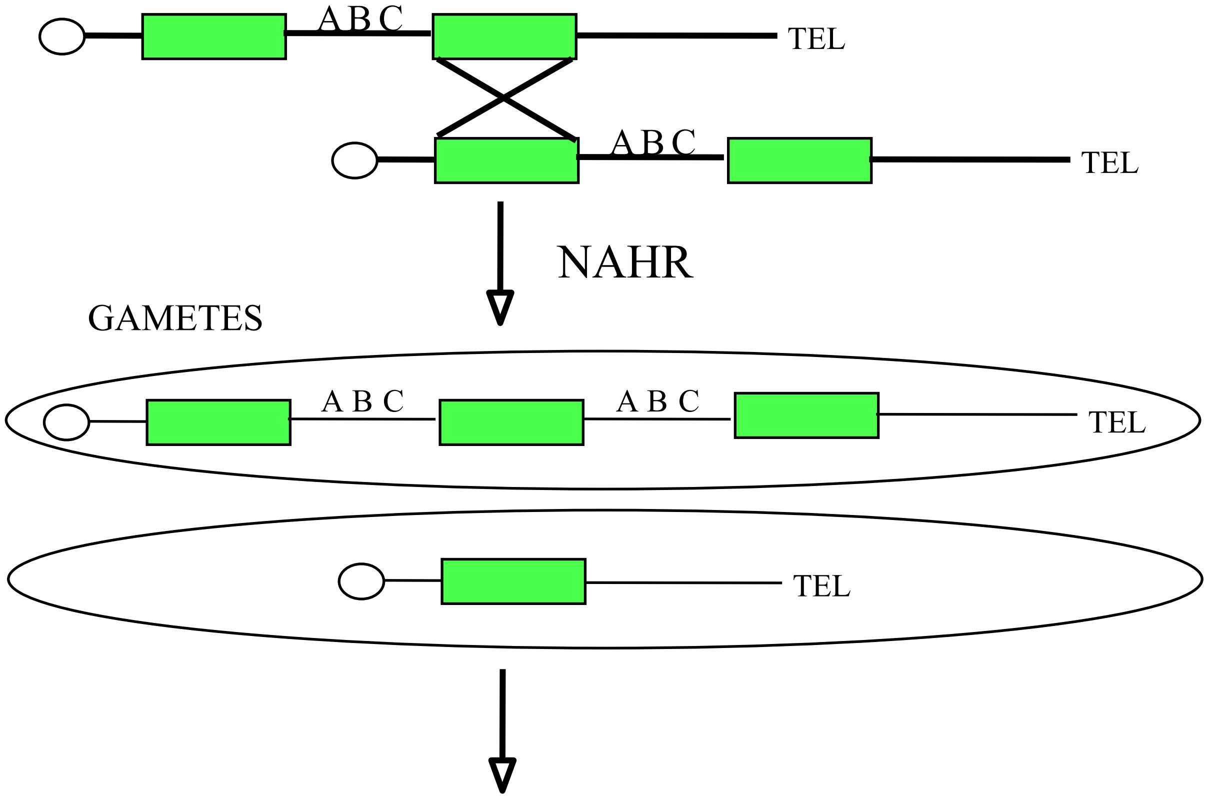
- 118 Mb or ~4% dup
- >20 kb, >95%
- 89% are tandem
- Gene/transcript poor
- Associated with LINEs

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

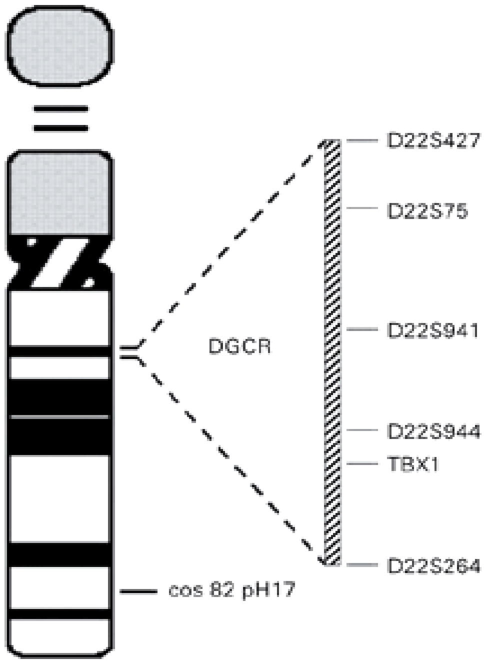


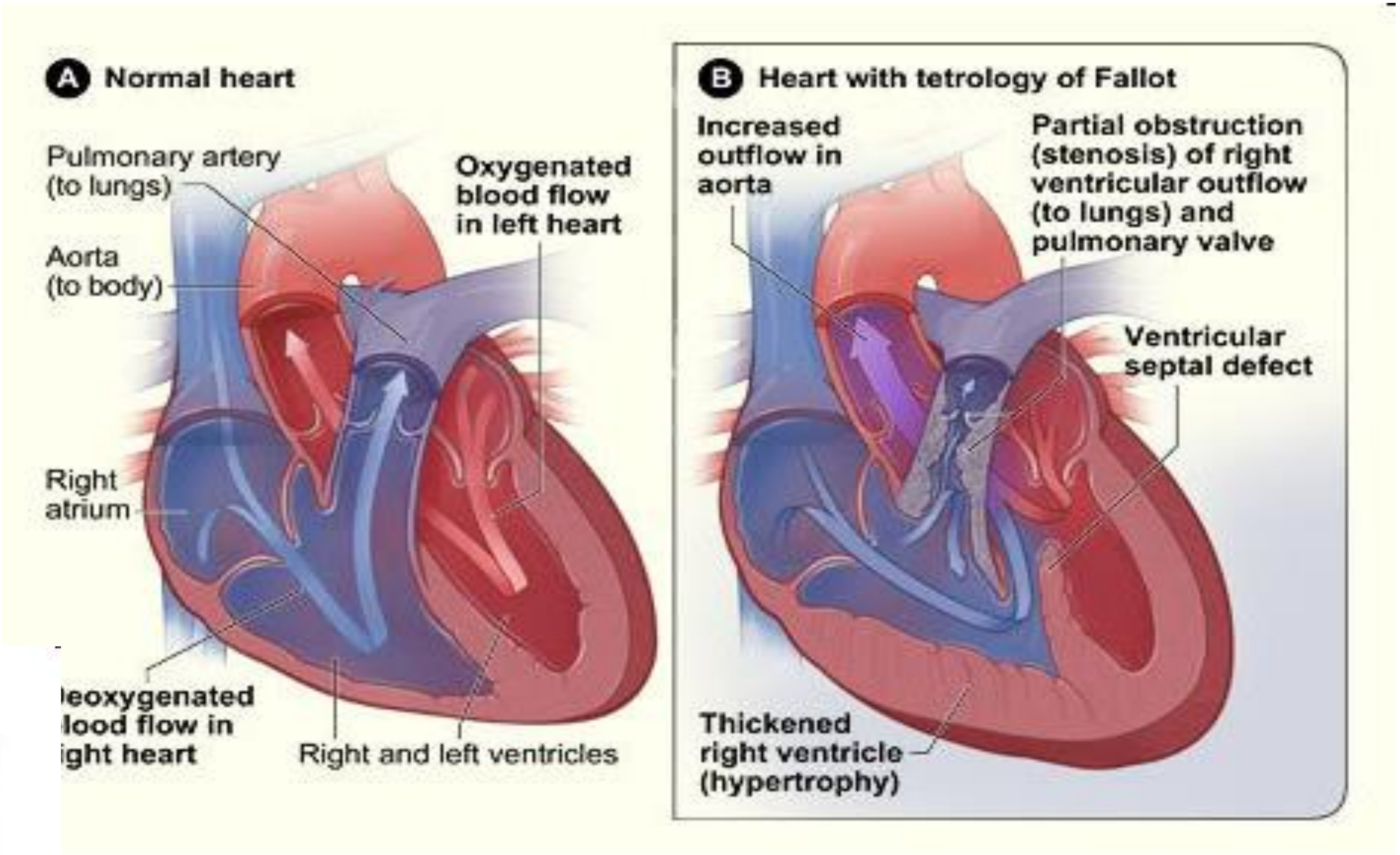
Human Disease

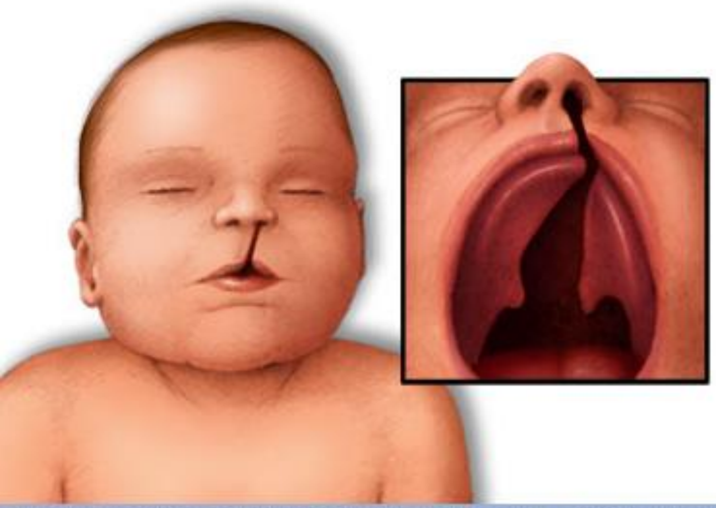
Triplosensitive, Haploinsufficient and Imprinted Genes

- **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



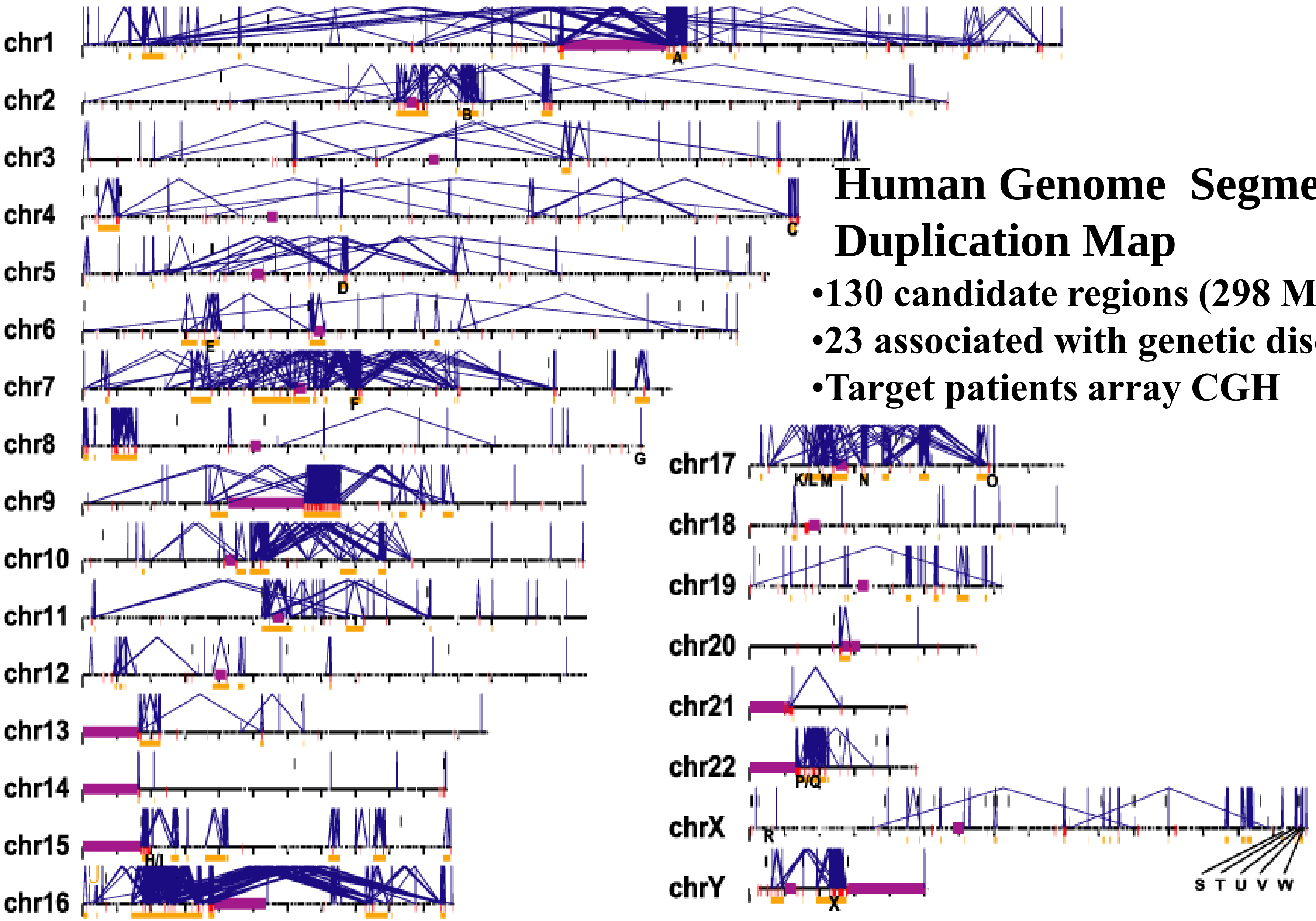

A Normal heart
Pulmonary artery (to lungs)
Aorta (to body)
Oxygenated blood flow in left heart
Right atrium
eoxygenated lood flow in ight heart
Right and left ventricles
B Heart with tetrology of Fallot
Increased outflow in aorta
Partial obstruction (stenosis) of right ventricular outflow (to lungs) and pulmonary valve
Ventricular septal defect
Thickened right ventricle (hypertrophy)



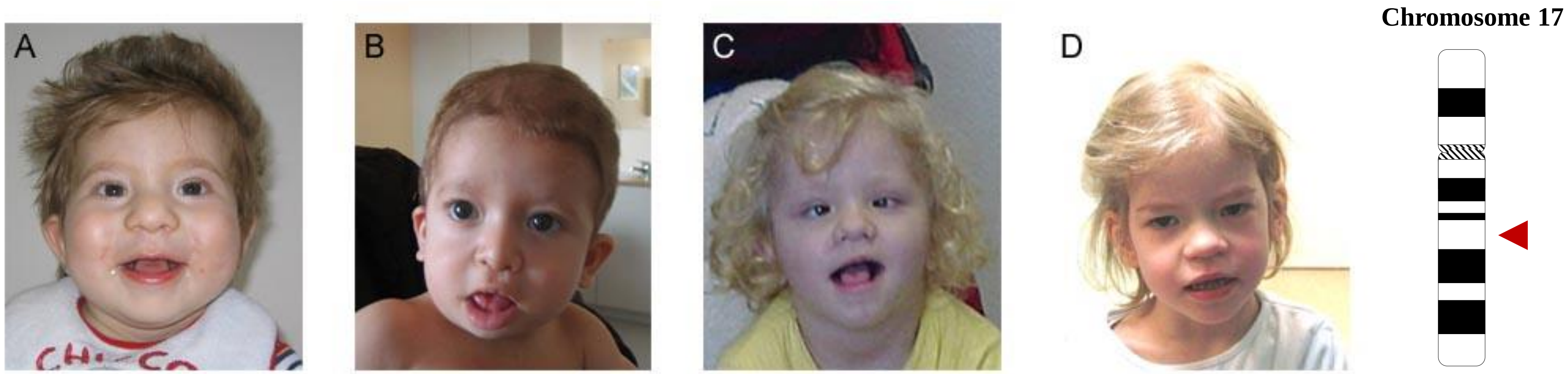
1/2000 live births
180 phenotypes
75-80% are sporadic (not inherited)

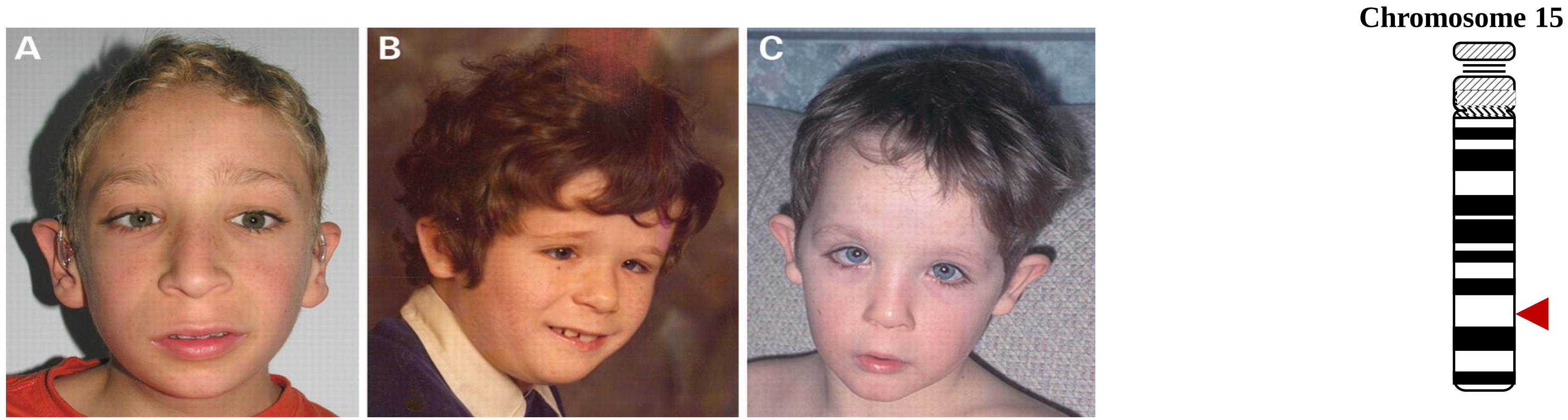
Human Genome Segmental Duplication Map

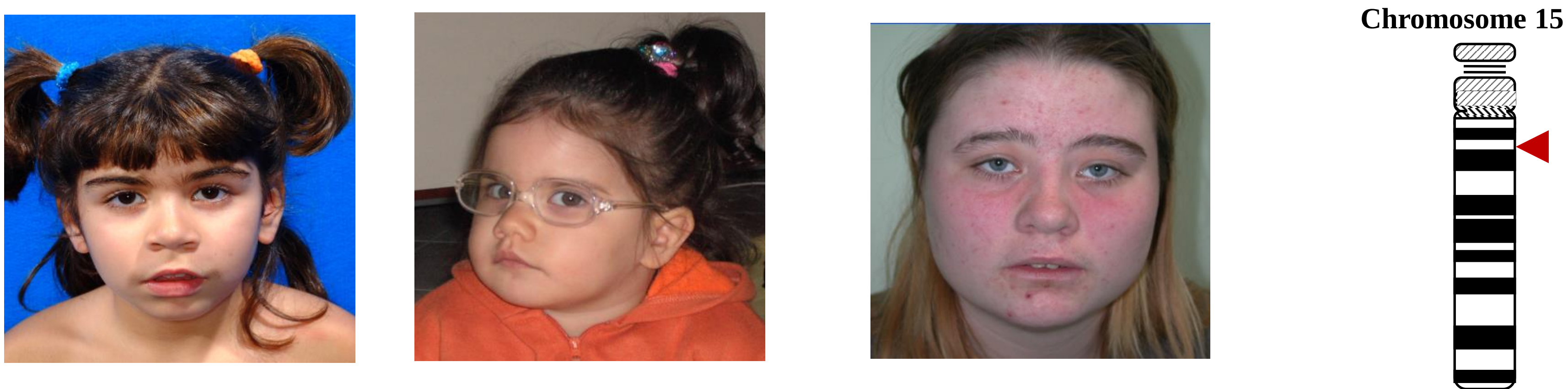
- 130 candidate regions (298 Mb)
- 23 associated with genetic disease
- Target patients array CGH



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

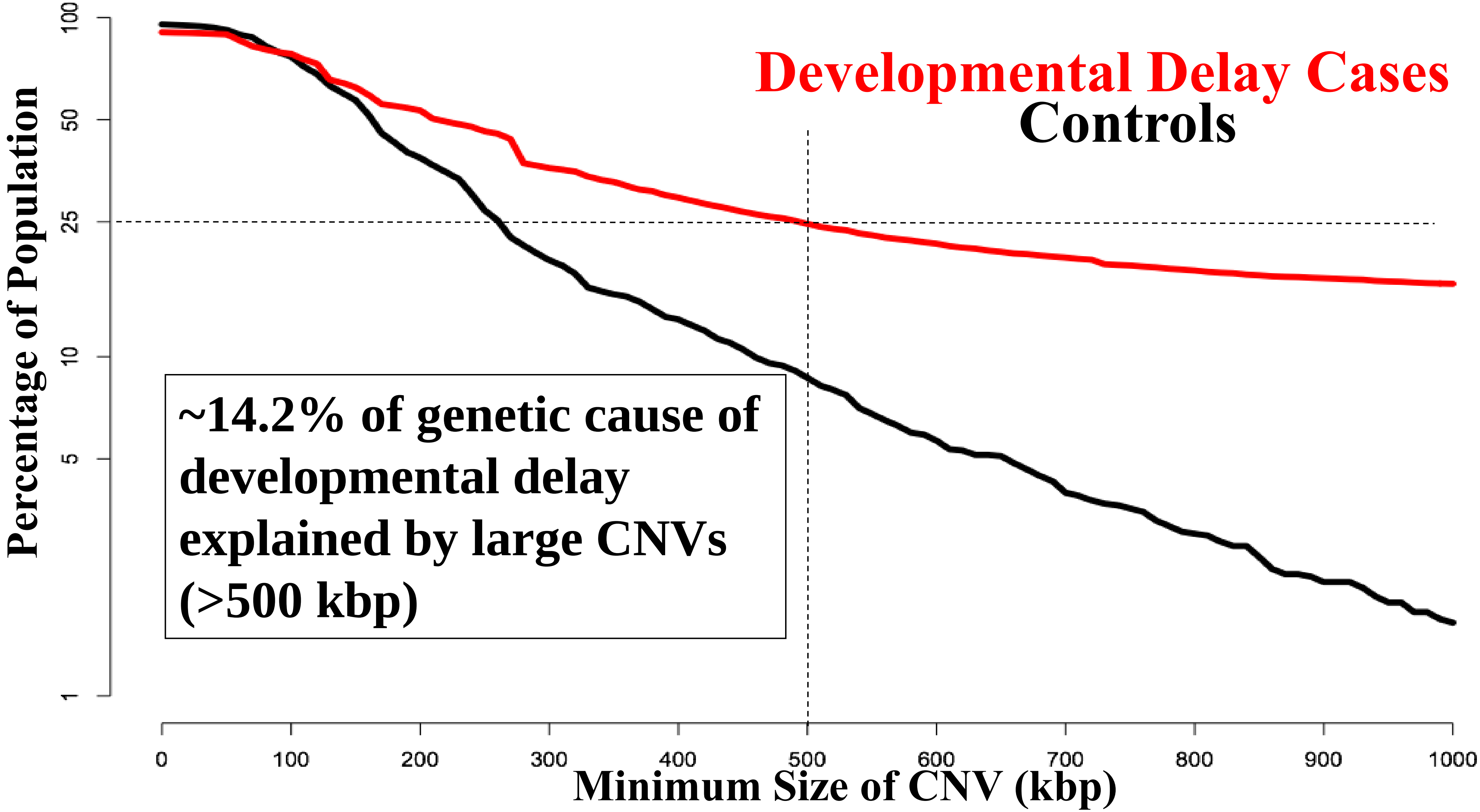

A
B
C
D
Chromosome 17


A
B
C
Chromosome 15


Chromosome 15

Genome Wide CNV Burden

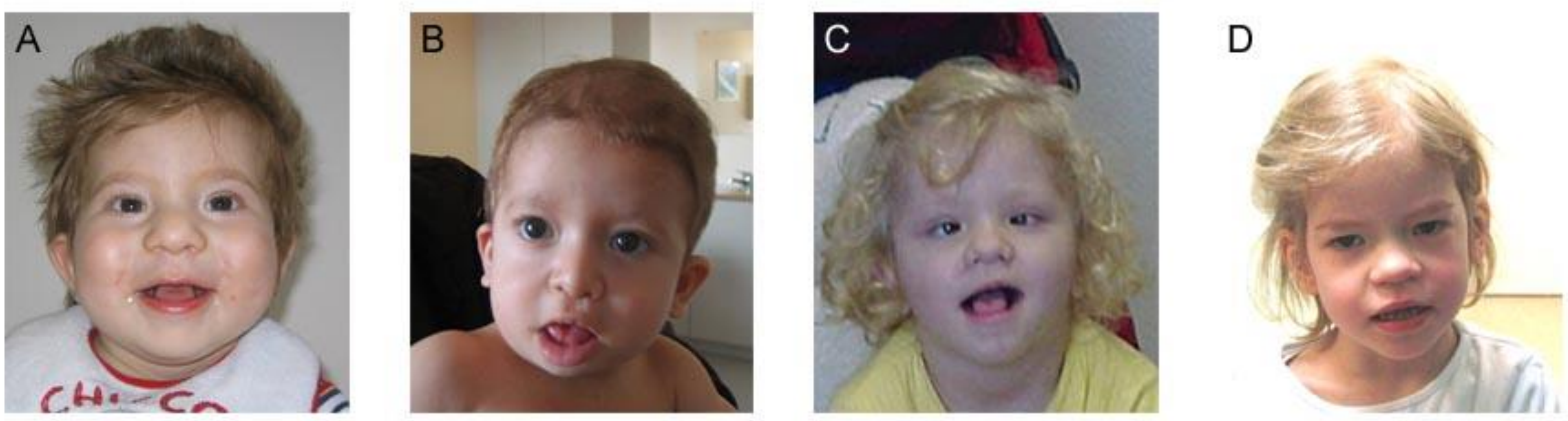
(15,767 cases of ID,DD,MCA vs. 8,328 controls)

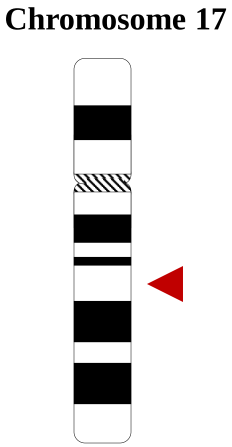


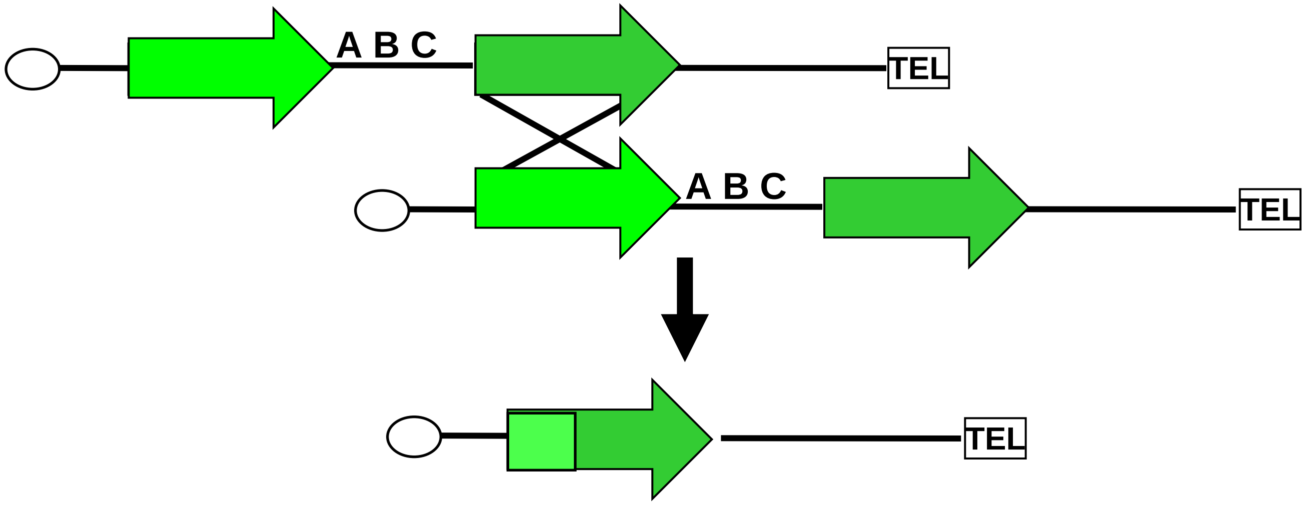
~14.2% of genetic cause of developmental delay explained by large CNVs (>500 kbp)

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

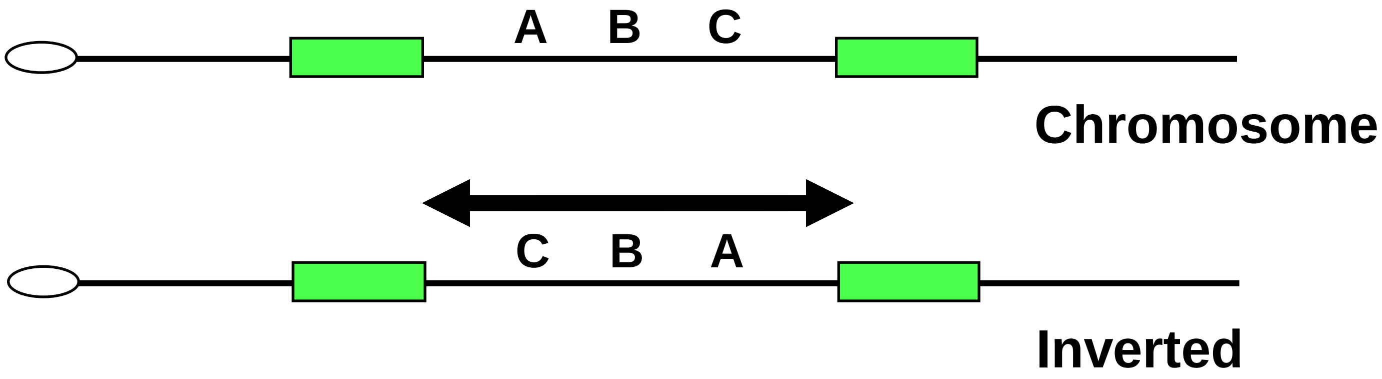
Common and Rare Structural Variation are Linked
17q21.31 Deletion Syndrome







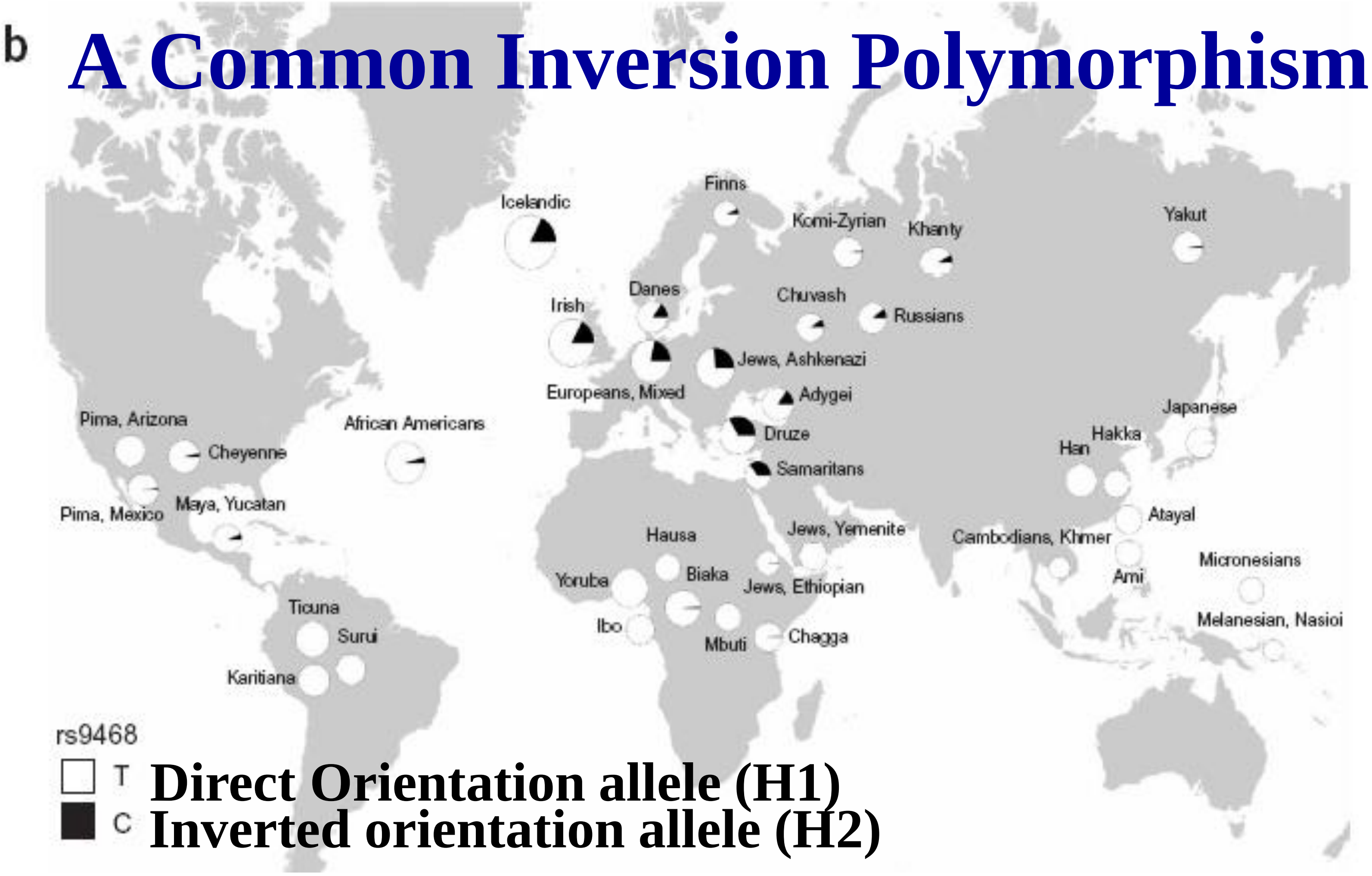
17q21.31 Inversion



- 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

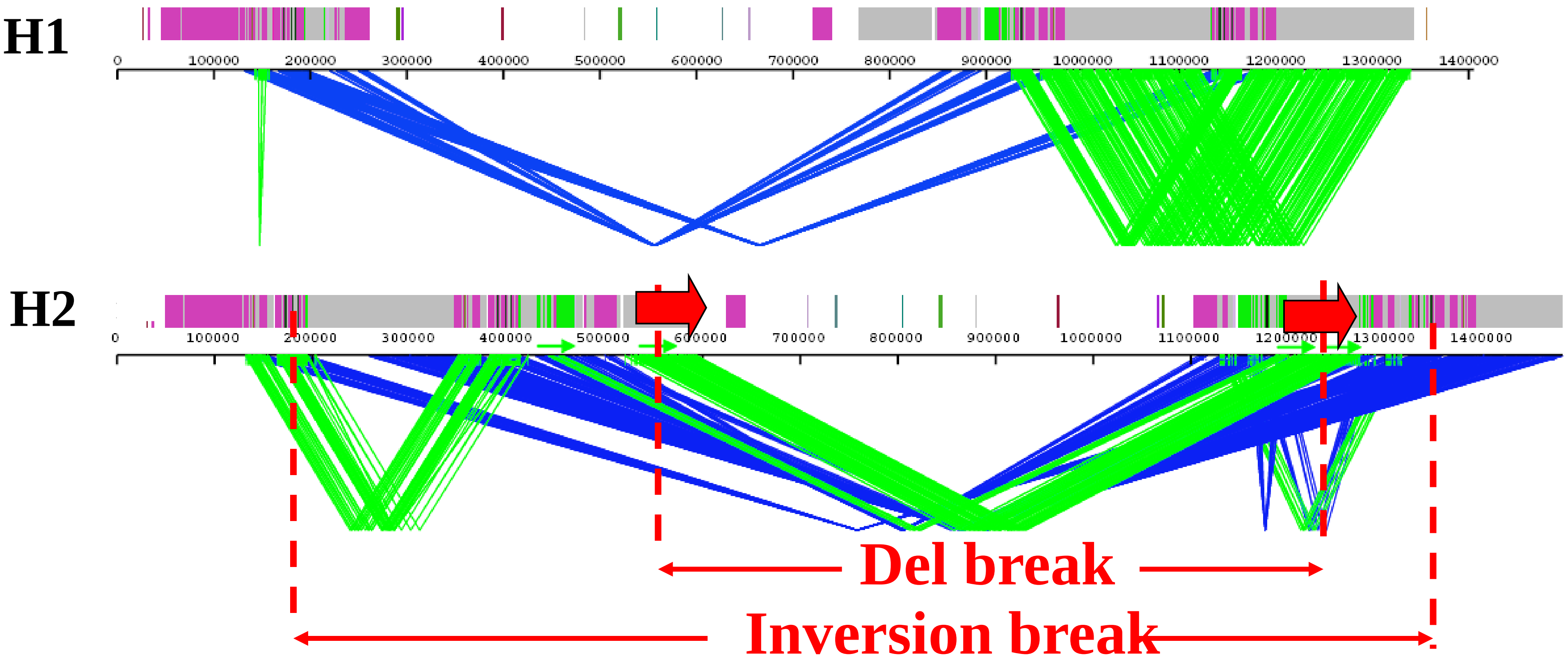


Direct Orientation allele (H1)

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



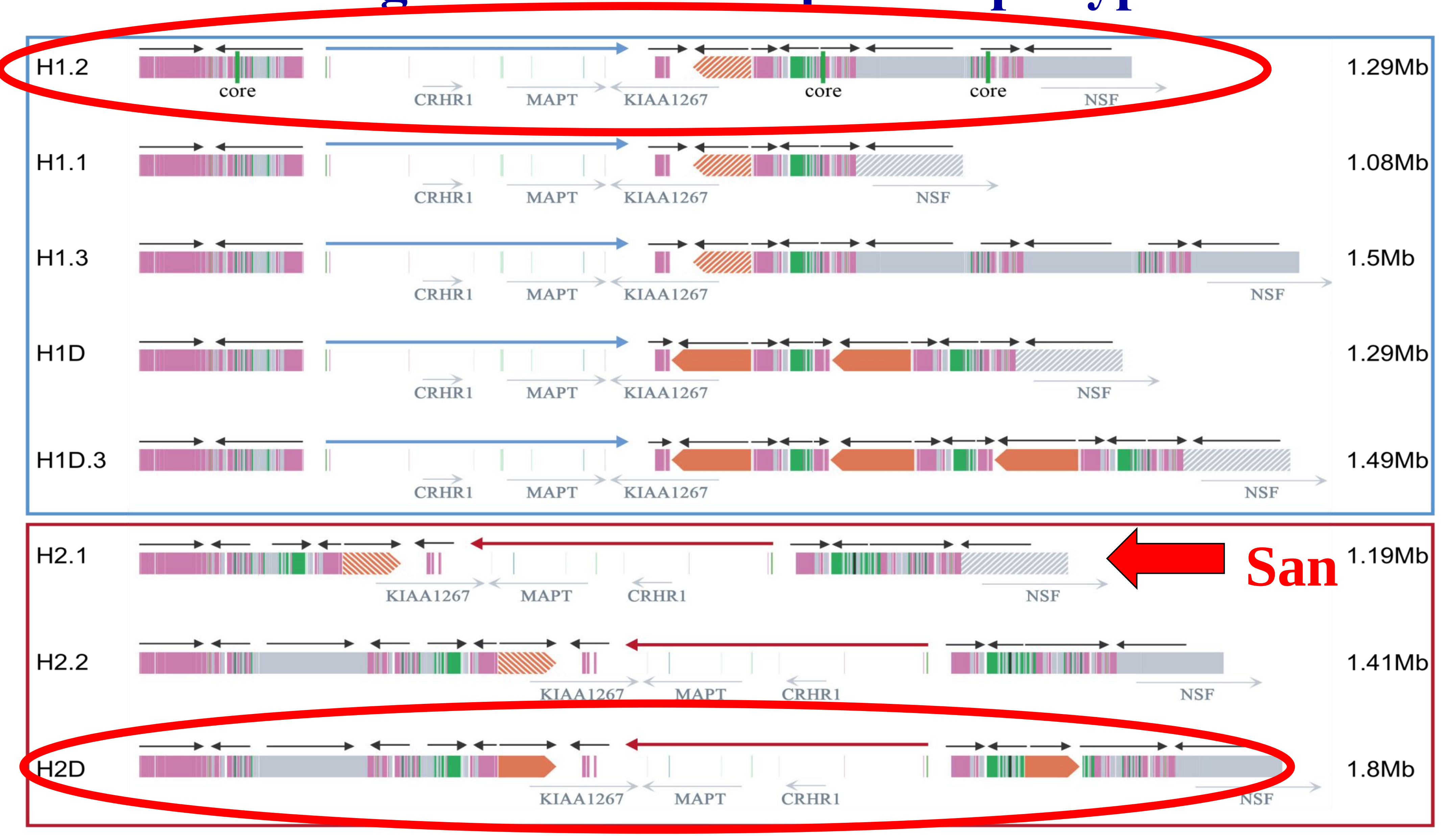
Del break

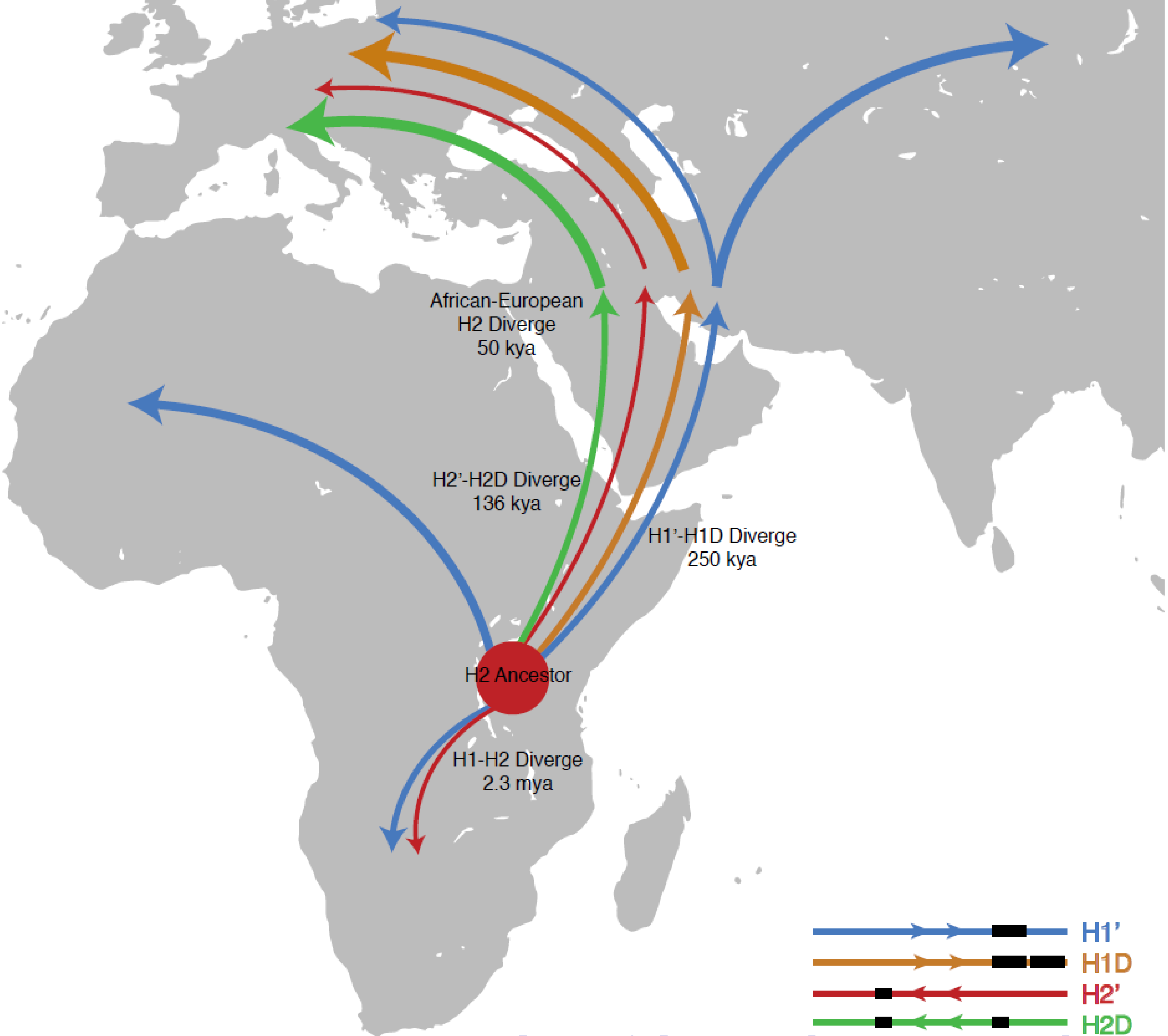
Inversion break

- 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

Structural Variation Diversity
Eight Distinct Complex Haplotypes





Meltz-Steinberg *et al*., Boettger *et al*., *Nat. Genet*. 2012

Summary

- Human genome is enriched for segmental duplications which predisposes to recurrent large CNVs during germ-cell production
- 15% of neurocognitive disease in intellectual disabled children is “caused” by CNVs—8% of normals carry large events
- Segmental Duplications enriched 10-25 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

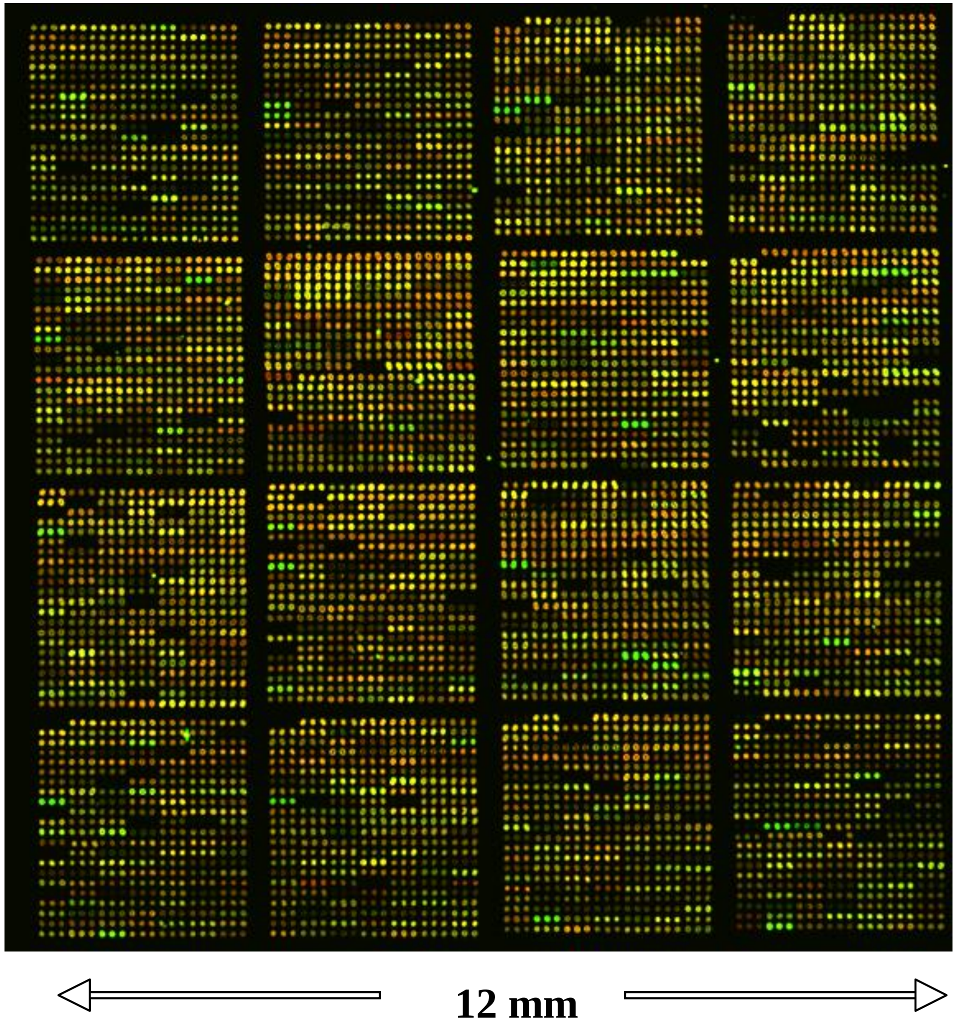
Single molecule mapping

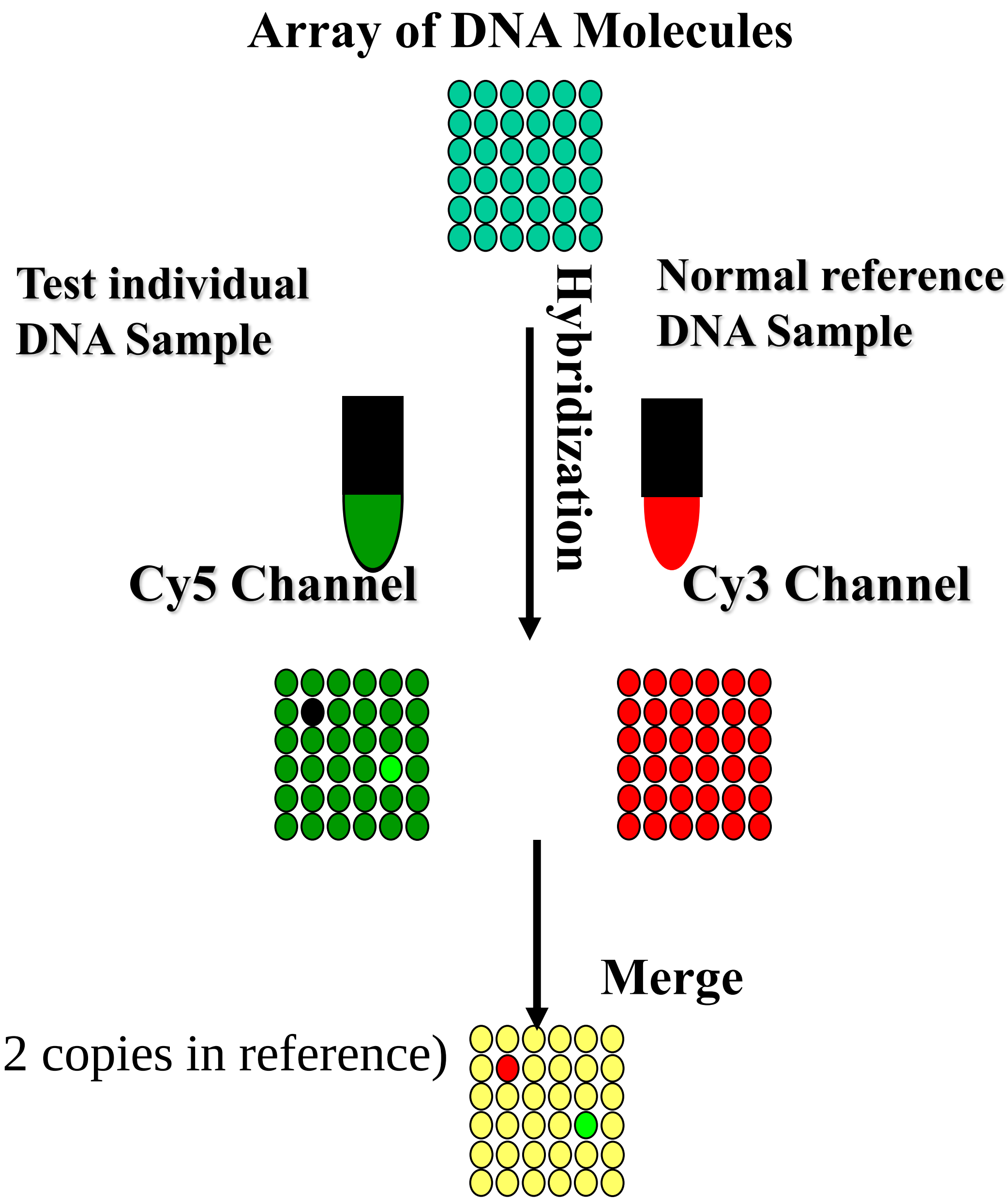
- Optical mapping**:** Teague et al., 2010
- Bionnano Genomics: Levy-Sakin et al, 2019

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,
- 3rd generation –Long-read Sequencing: Chaisson *et al*., 2015, 2019, Pendleton *et al*., 2015, Sedlazeck et a., 2018 Audano *et al*, 2019

Array Comparative Genomic Hybridization

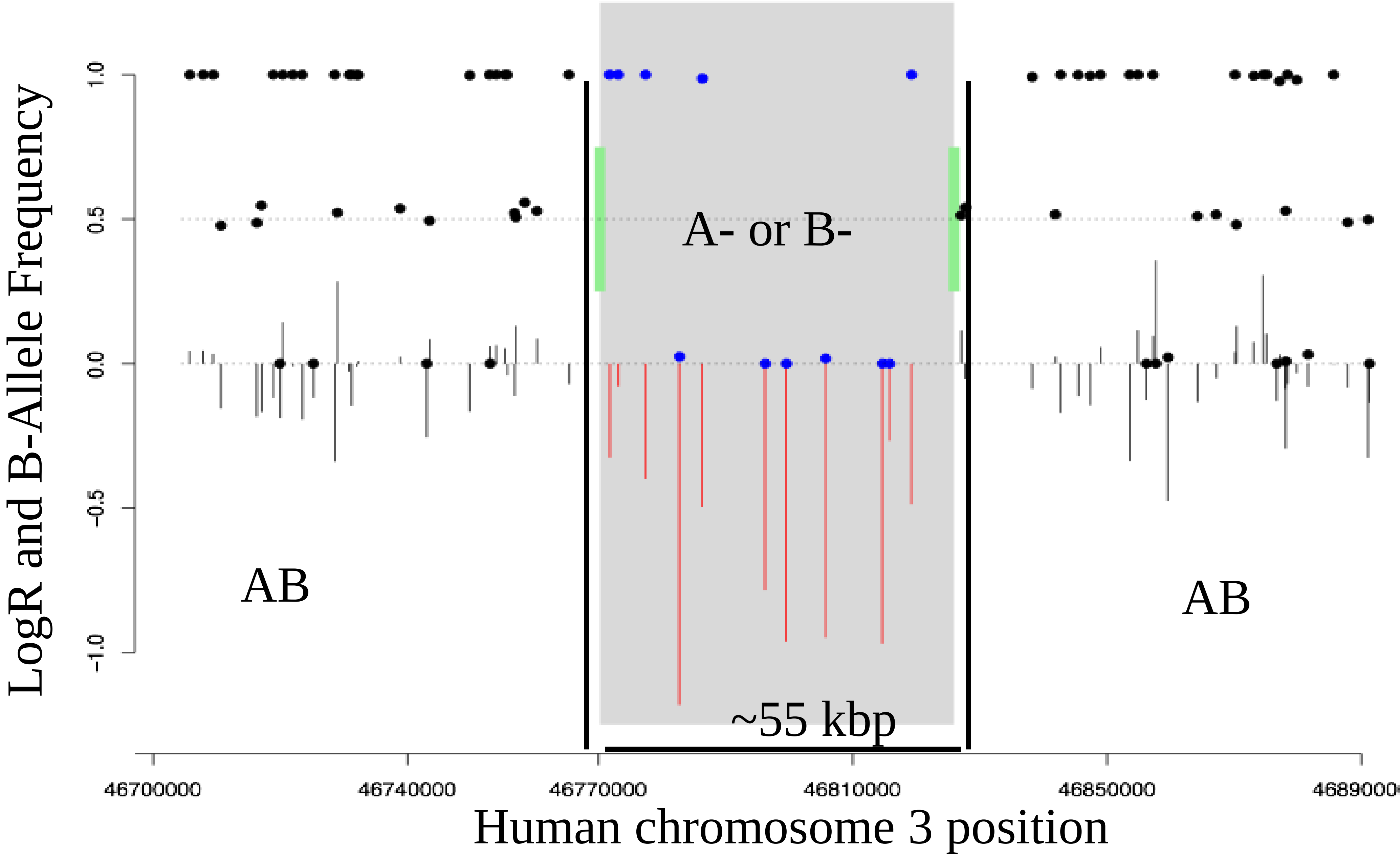




One copy gain = $\log_2(3/2) = 0.57$ (3 copies vs. 2 copies in reference)

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

SNP Microarray detection of Deletion (Illumina)



LogR and B-Allele Frequency

A- or B-

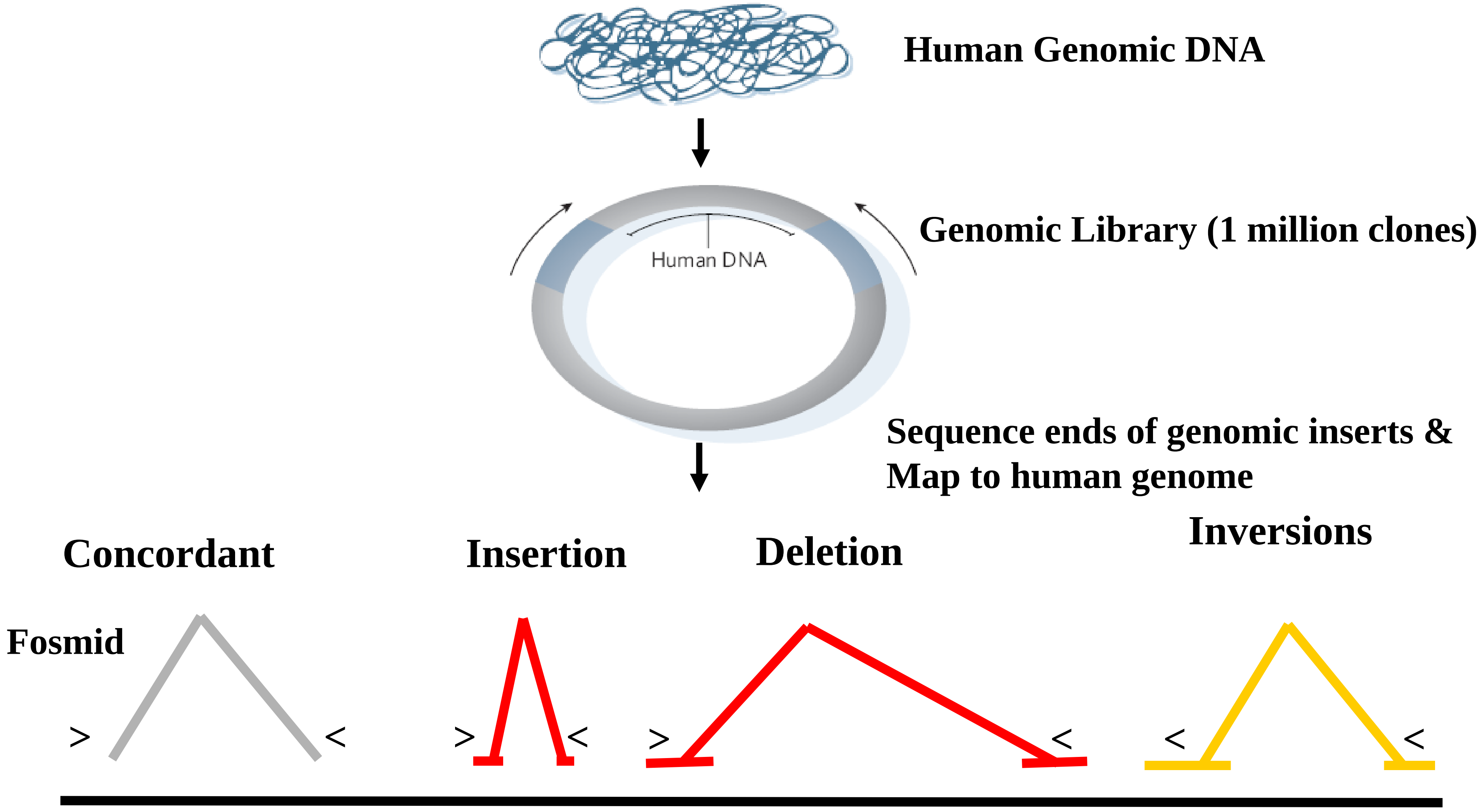
AB

AB

~55 kbp

Human chromosome 3 position

Using Read Pairs to Resolve Structural Variation



Human Genomic DNA

Genomic Library (1 million clones)

Sequence ends of genomic inserts &
Map to human genome

Concordant

Insertion

Deletion

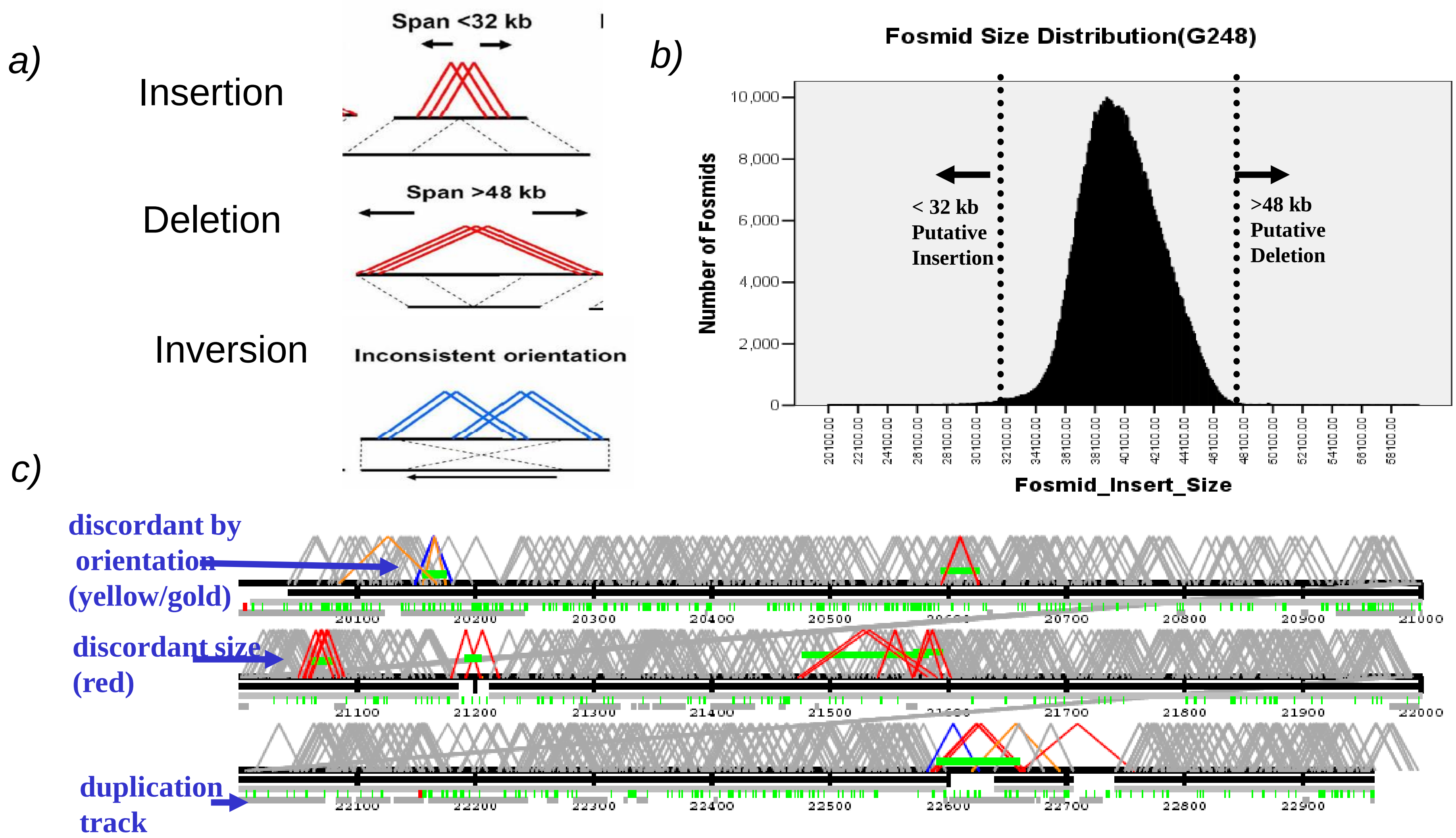
Inversions

Fosmid

Build35

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



discordant by orientation (yellow/gold)

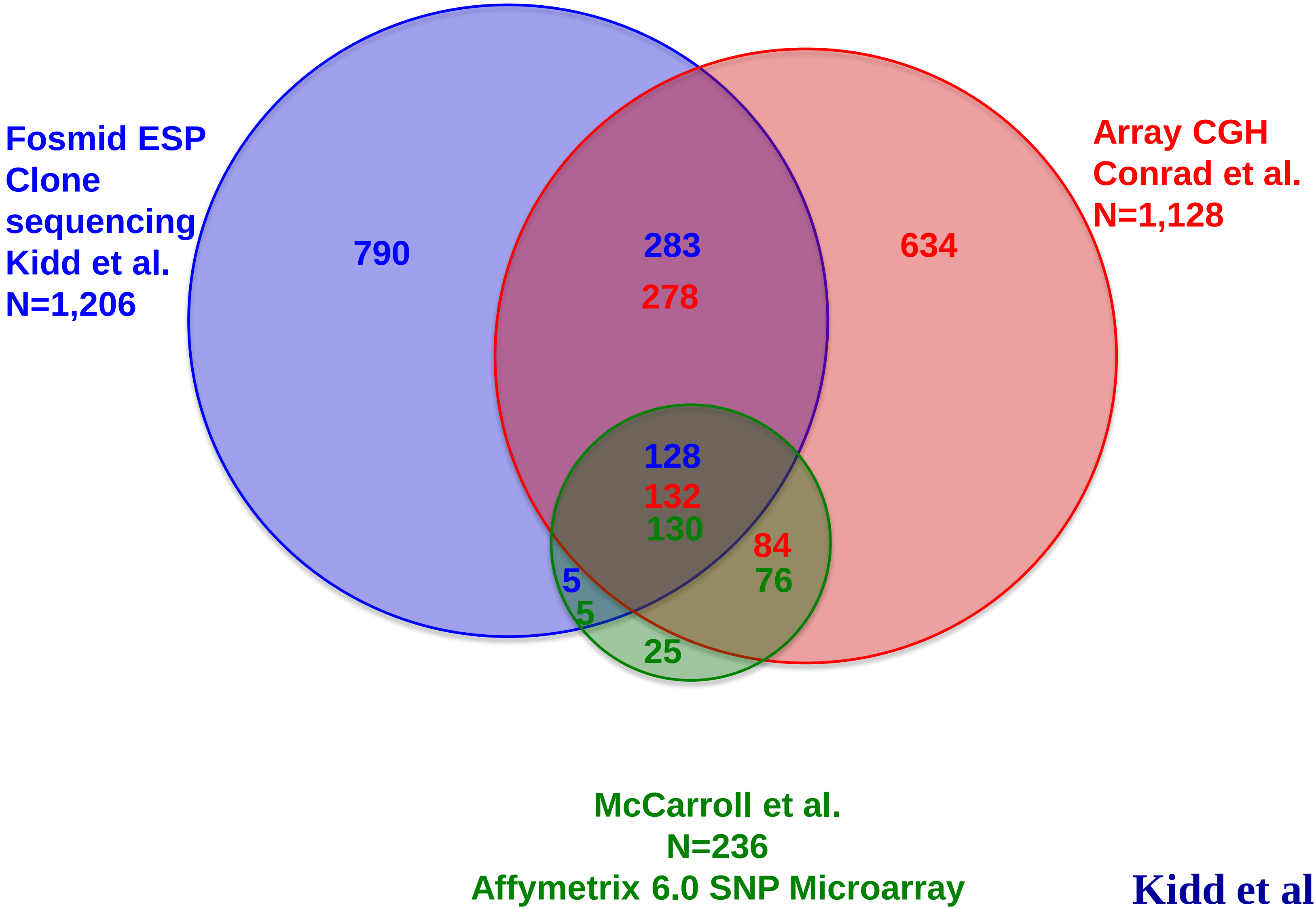
discordant size (red)

duplication track

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

Experimental Approaches Incomplete

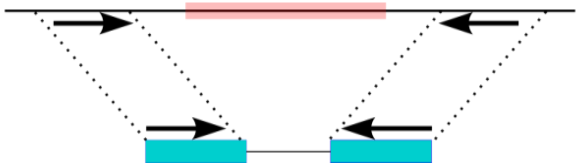
(Examined 5 identical genomes > 5kbp)



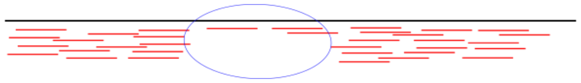
Kidd et al., *Cell* 2010

Next-Generation Sequencing Methods

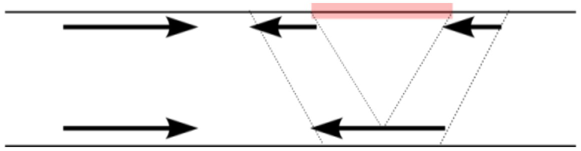
- **Read pair analysis**
  - Deletions, small novel insertions, inversions, transposons
  - Size and breakpoint resolution dependent to insert size



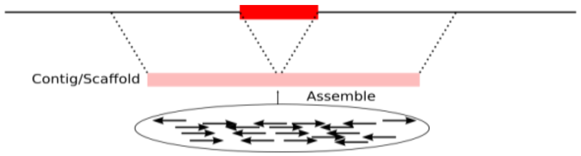
- **Read depth analysis**
  - Deletions and duplications only
  - Relatively poor breakpoint resolution



- **Split read analysis**
  - Small novel insertions/deletions, and mobile element insertions
  - 1bp breakpoint resolution

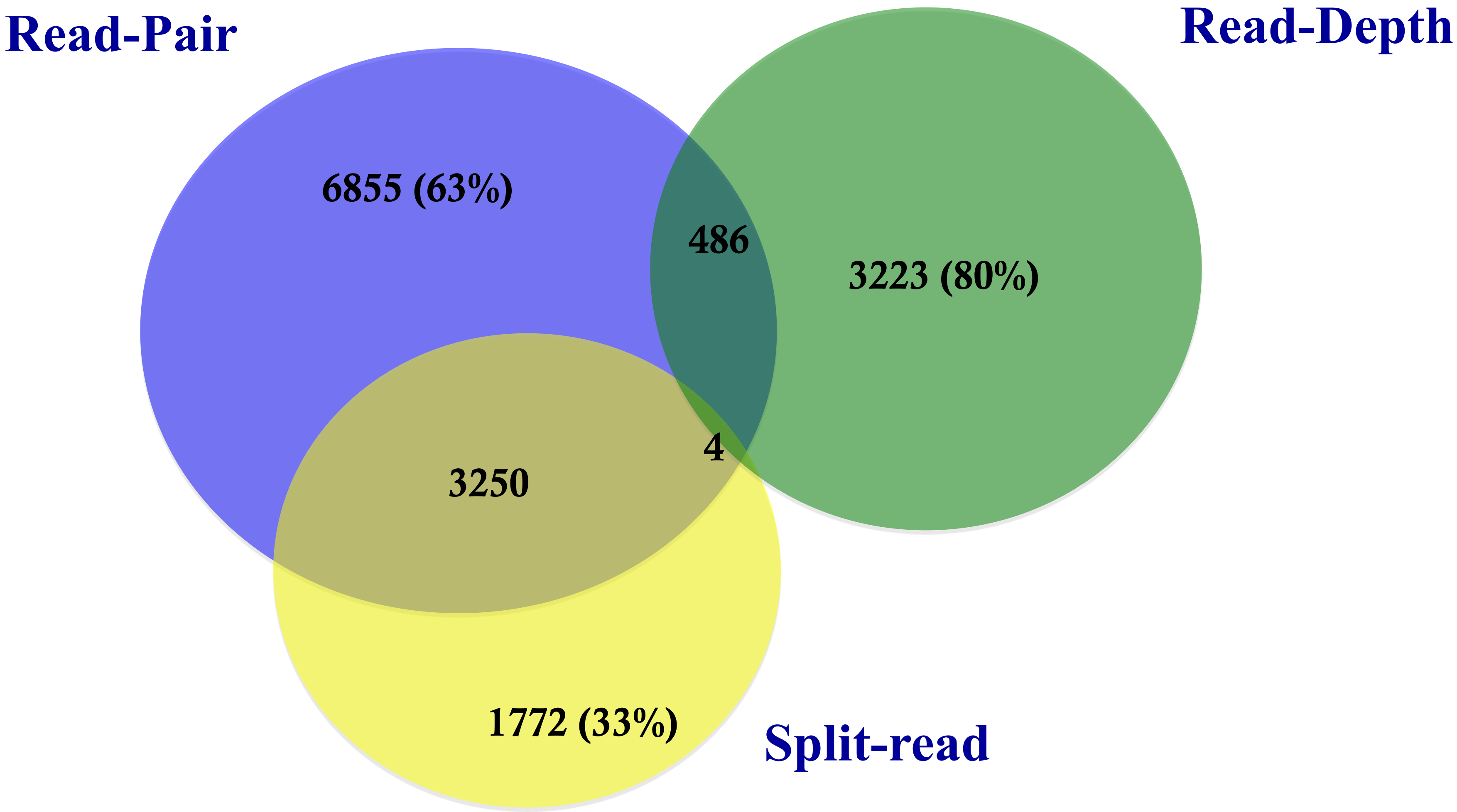


- **Local and *de novo* assembly**
  - SV in unique segments
  - 1bp breakpoint resolution



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

Computational Approaches are Incomplete
159 genomes (2-4X) (deletions only)



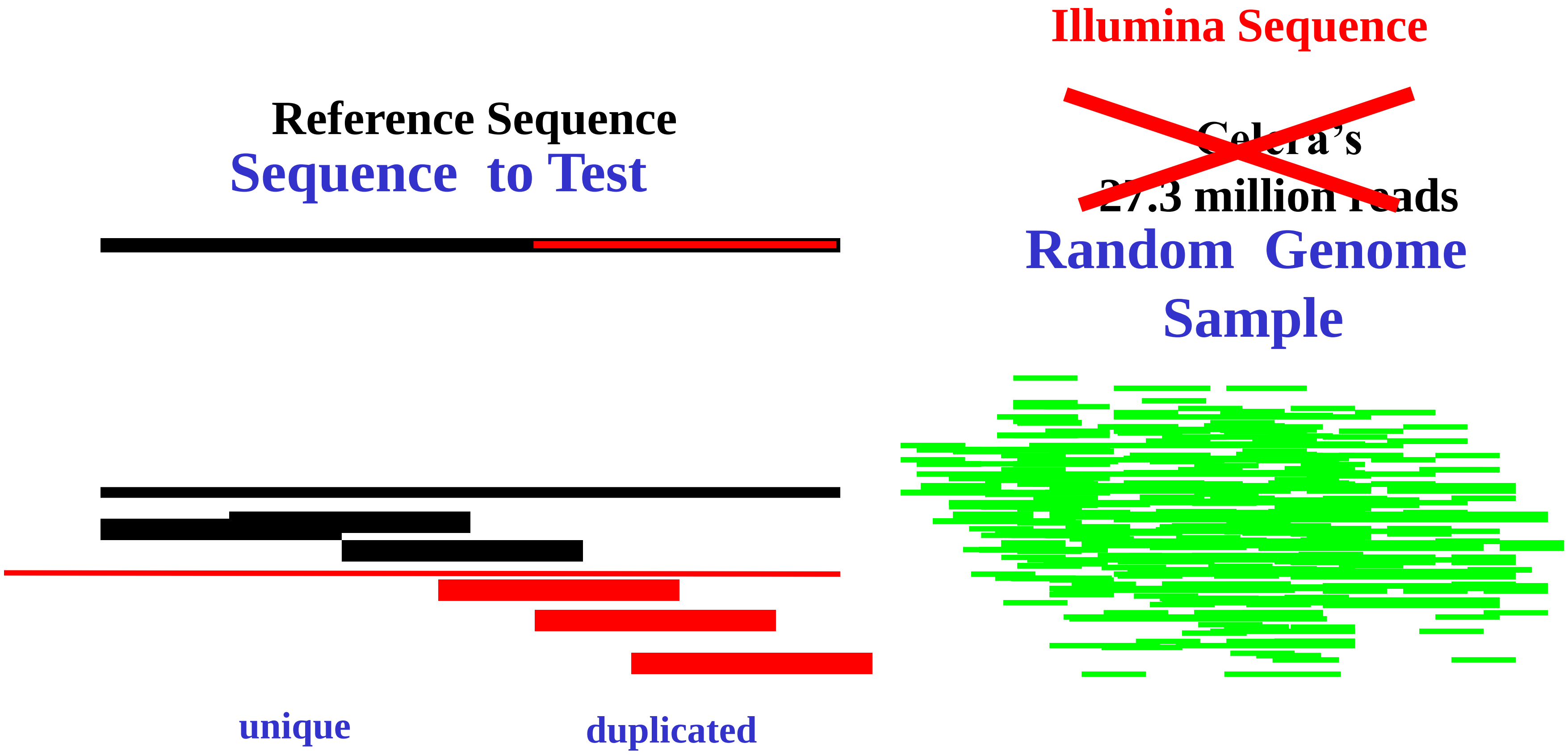
Mills *et al*., Nature 2011

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.

Using Sequence Read Depth

- Map whole genome sequence to reference genome
  - Variation in copy number correlates linearly with read-depth
- Caveat: need to develop algorithms that can map reads to all possible locations given a preset divergence (eg. mrFAST, mrsFAST)



Illumina Sequence

~~Celera's 27.3 million reads~~

Random Genome Sample

Reference Sequence

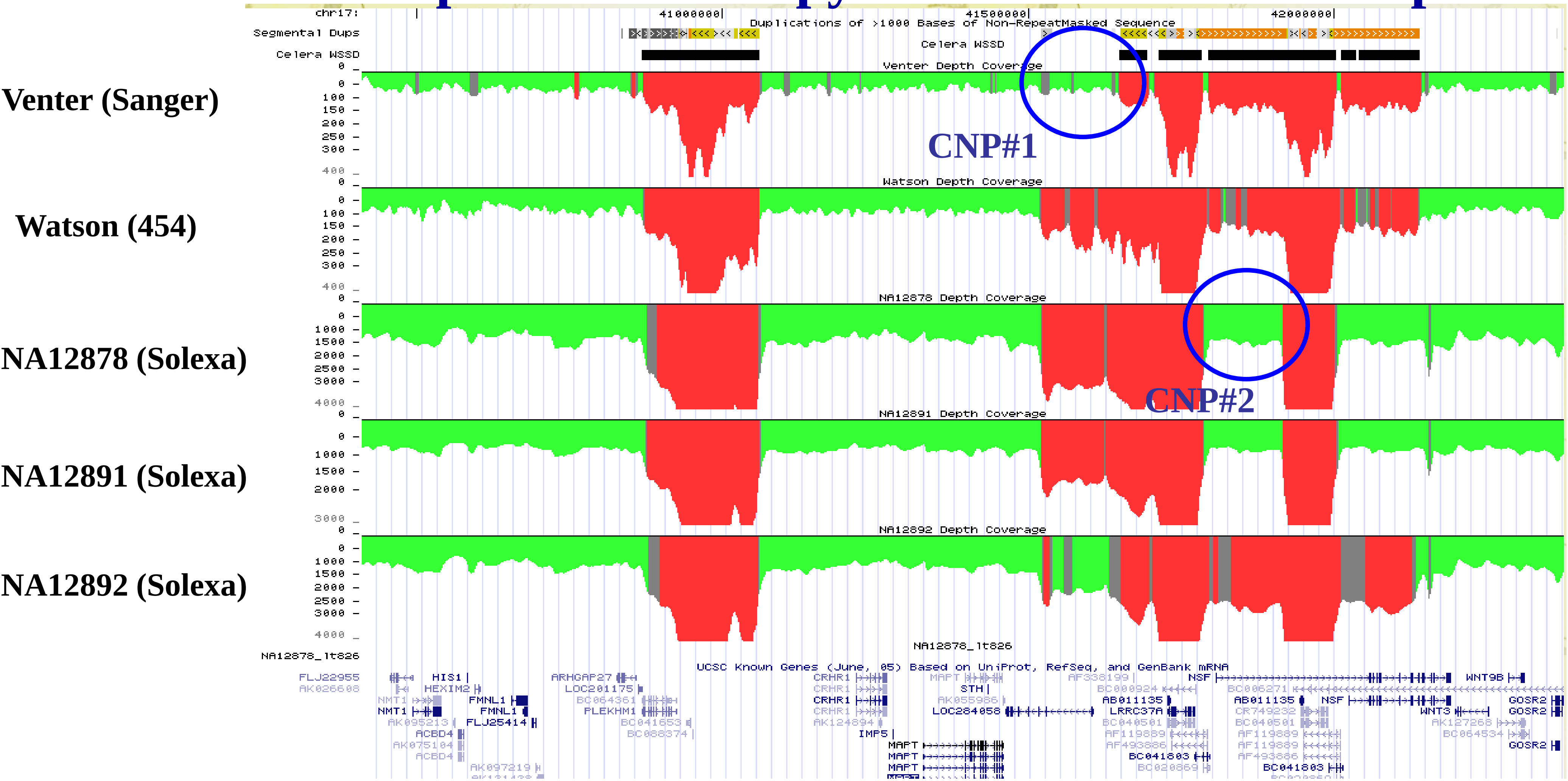
Sequence to Test

unique

duplicated

Bailey et al., Science, 2002

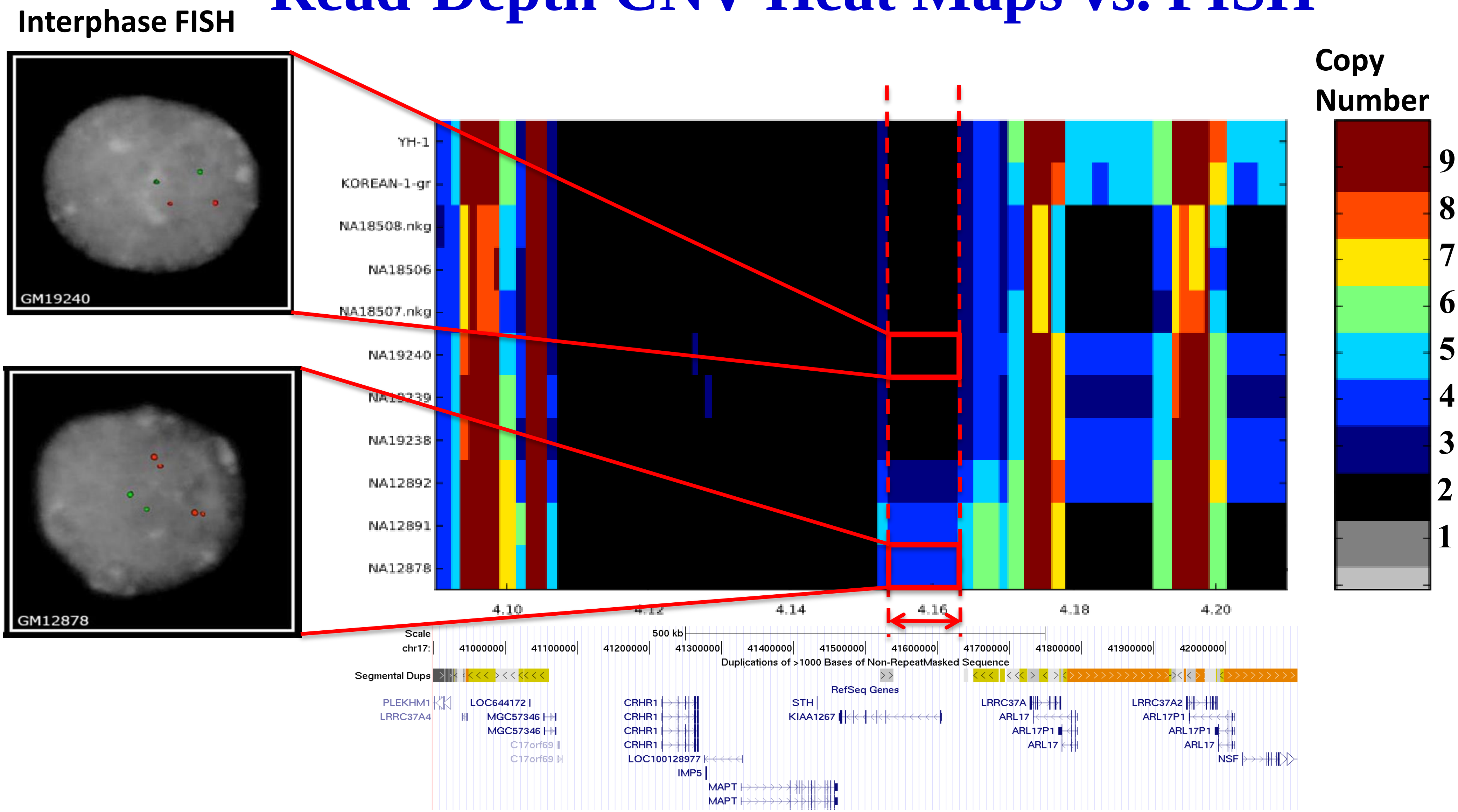
Personalized Duplication or Copy-Number Variation Maps



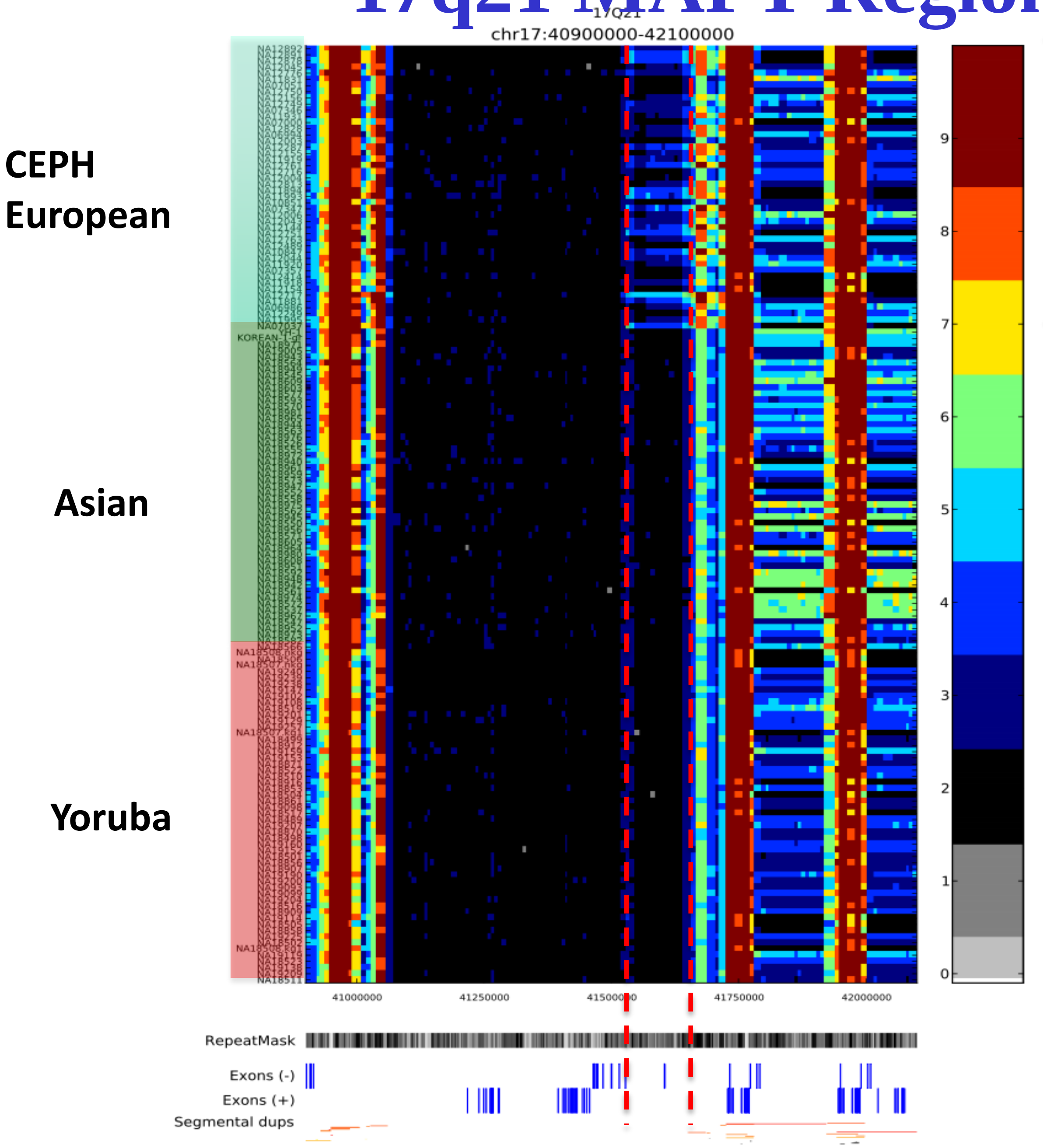
•Two known ~70 kbp CNPs, CNP#1 duplication absent in Venter but predicted in Watson and NA12878, CNP#2 present mother but neither father or child

Alkan, Nat. Genet, 2009

Read-Depth CNV Heat Maps vs. FISH



17q21 MAPT Region for 150 Genomes



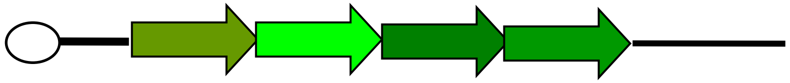
71% of Europeans carry at least Partial duplication distal (17q21 associated)—all inversions carry the duplication

24% of Asians are hexaploid for NSF gene N-ETHYLMALEIMIDE-SENSITIVE FACTOR potentially important in synapse membrane fusion; NSF (decreased expression in schizophrenia brains (Mimics, 2000), Drosophila mutants results in aberrant synaptic transmission)

Sudmant et al., 2010, Science

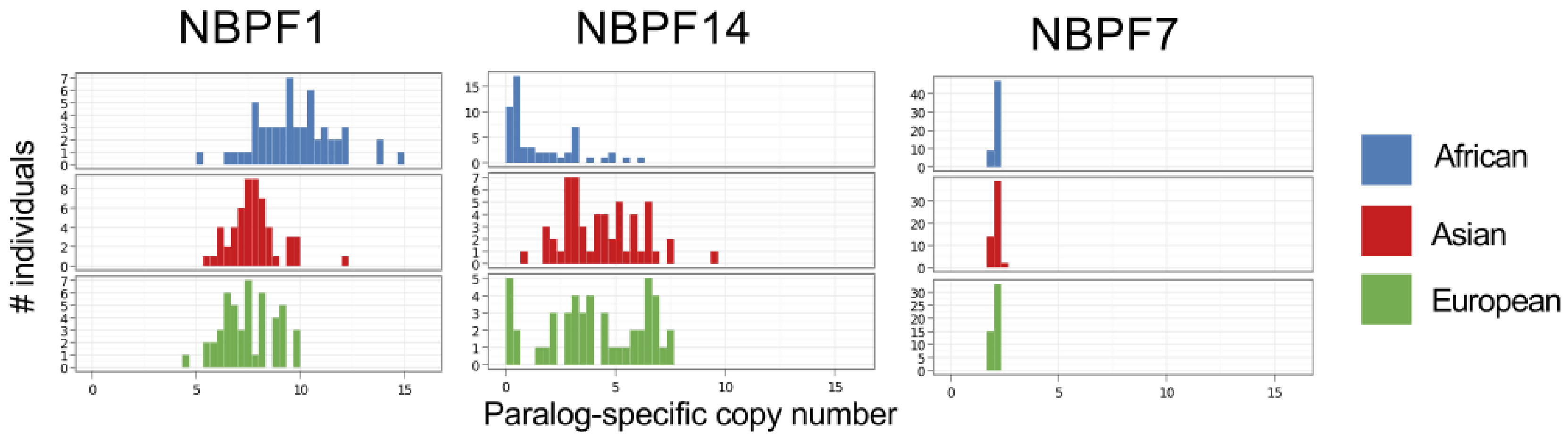
Unique Sequence Identifiers Distinguish Copies

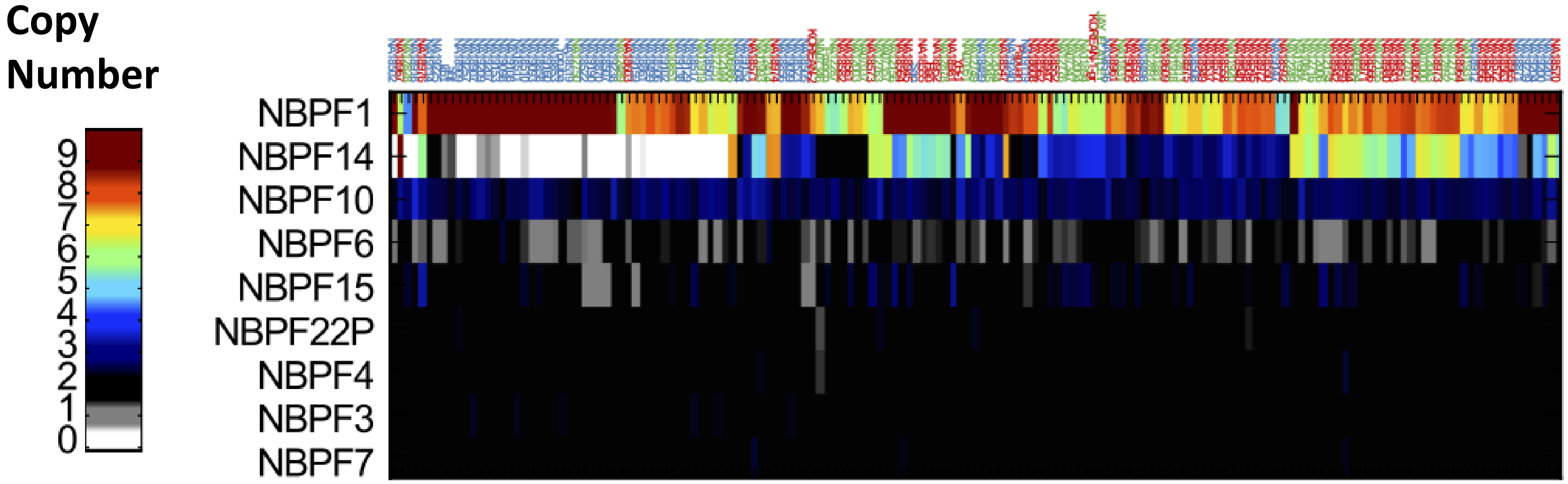
```
copy1   ATGCTAGGCATATAATATCCGACGATATACATATAGATGTTAG…
copy2   ATGCTAGGCATAGAATATCCGACGATATACATATACATGTTAG…
copy3   ATGCTACGCATAGAATATCCCACGATATACATATACATGTTAG…
copy4   ATGCTACGCATATAATATCCGACGATATAC--ATACATGTTAG.
```



- Self-comparison identifies 3.9 million singly unique nucleotide (SUN) identifiers in duplicated sequences
- Select 3.4 million SUNs based on detection in 10/11 genomes=informative SUNs=paralogous sequence variants that are largely fixed
- Measure read-depth for specific SUNs--genotype copy-number status of specific paralogs

NBPF Gene Family Diversity

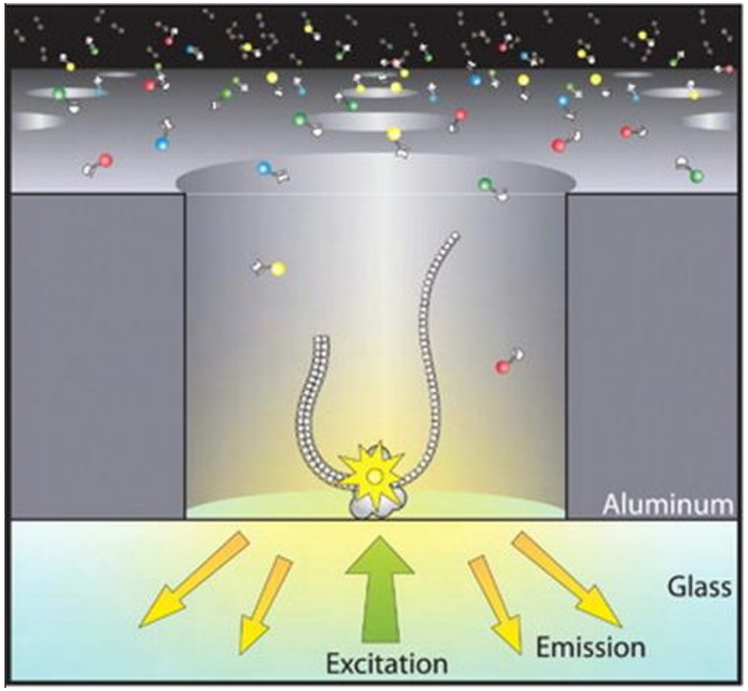




Future of SV Detection

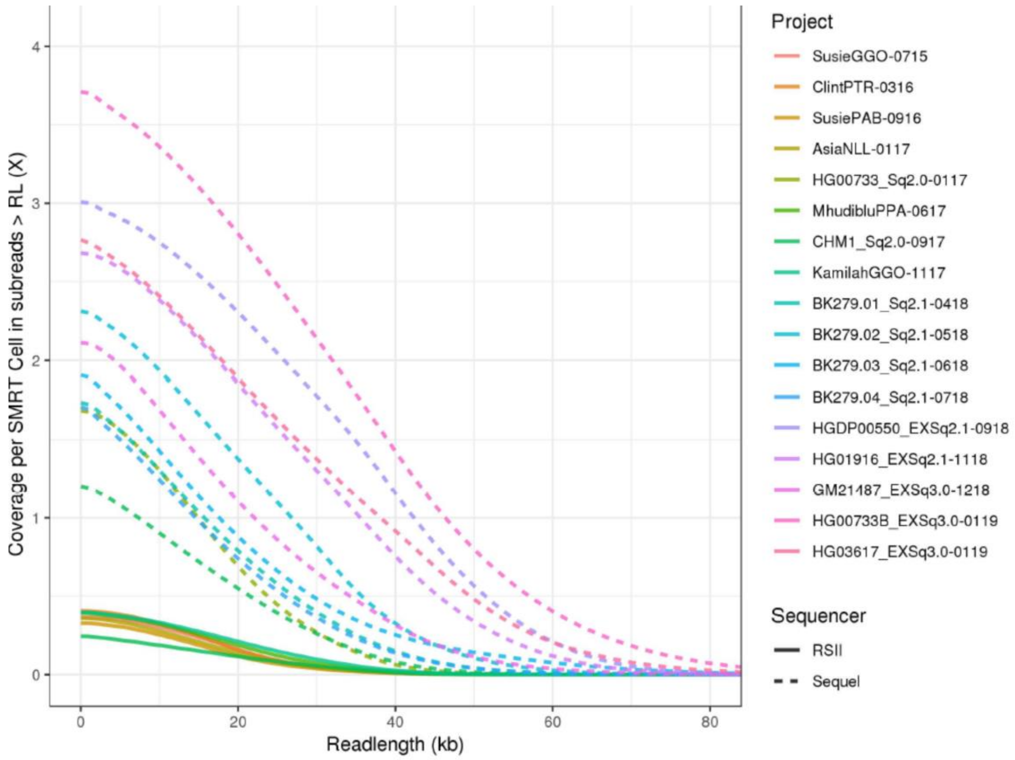
1) **Focus on comprehensive assessment of genetic variation**— large portions of human genetic variation are still missed
2) **Current NGS methods are indirect** and do not resolve structure but provide specificity and excellent dynamic range response.
3) **High quality sequence resolution of complex structural variation to establish alternate references/haplotypes—**often show extraordinary differences in genetic diversity
4) **Technology advances in whole genome sequencing "Third Generation Sequencing"**: Long-read sequencing technologies with NGS throughput in order to sequence and assemble regions and genomes *de novo*

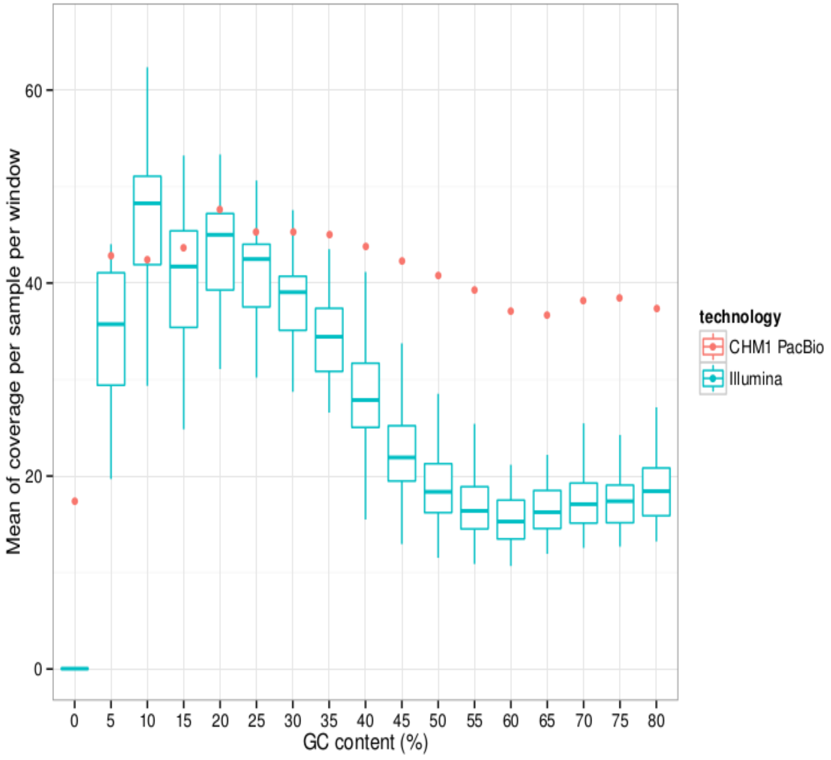
Single-Molecule Real-Time Sequencing (SMRT)
a.k.a. PacBio sequencing



CLR—Continuous Long Reads--no cloning, low throughput, 15% error rate
CCS—Circular Consensus Sequencing—no cloning, high throughput 0.1% error rate

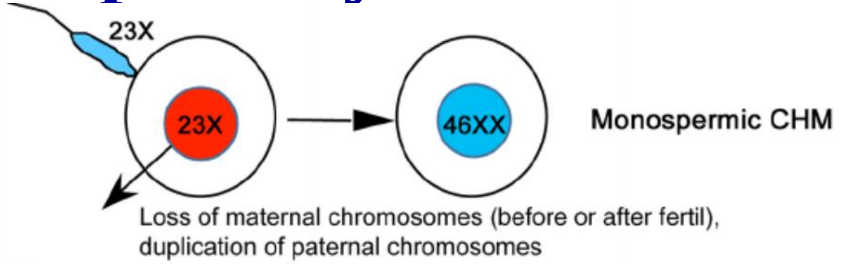
PacBio sequence reads are long, uniformly distributed with near-random error

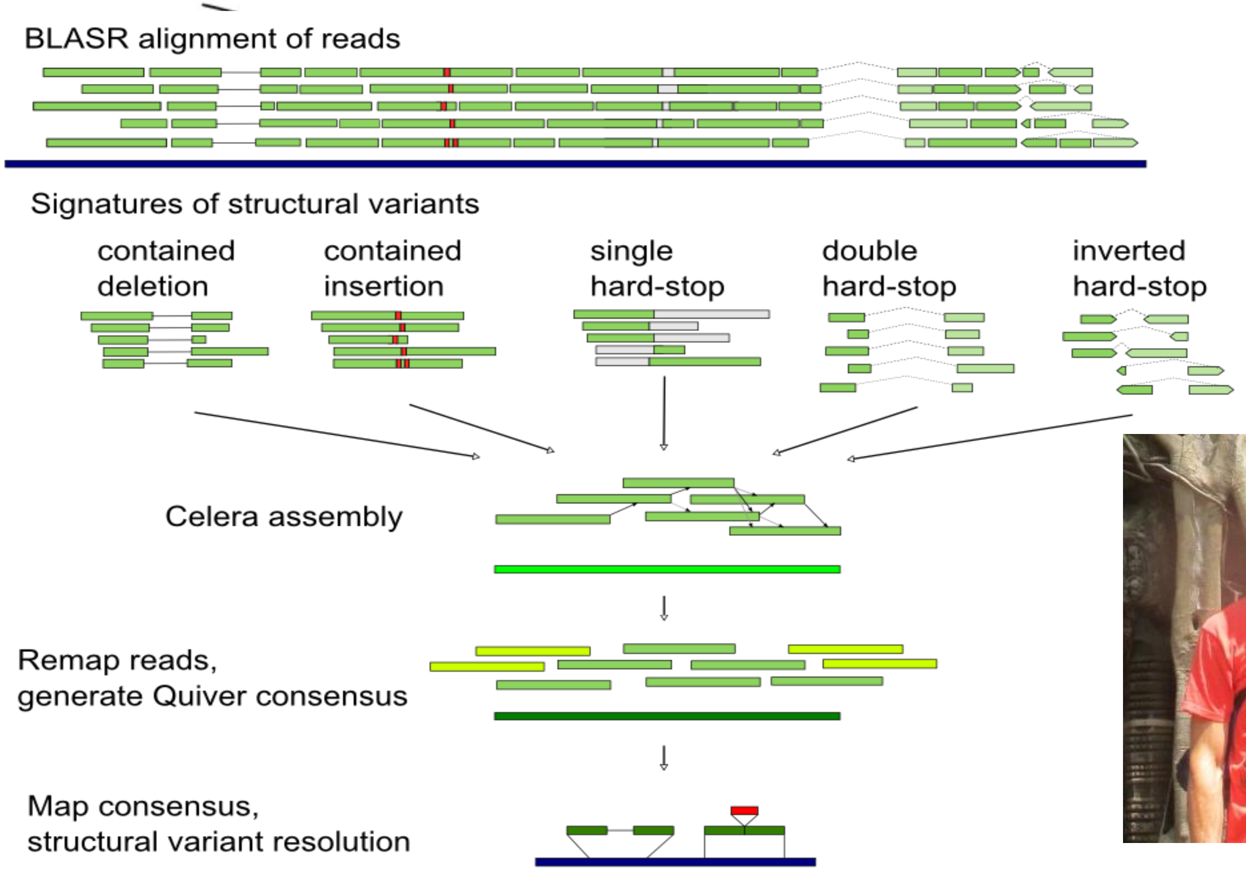




- P6C4 chemistry—30-40 kbp libraries
- Mean 15-25 kbp read (6 hr movies)
- Max 120-130 kbp

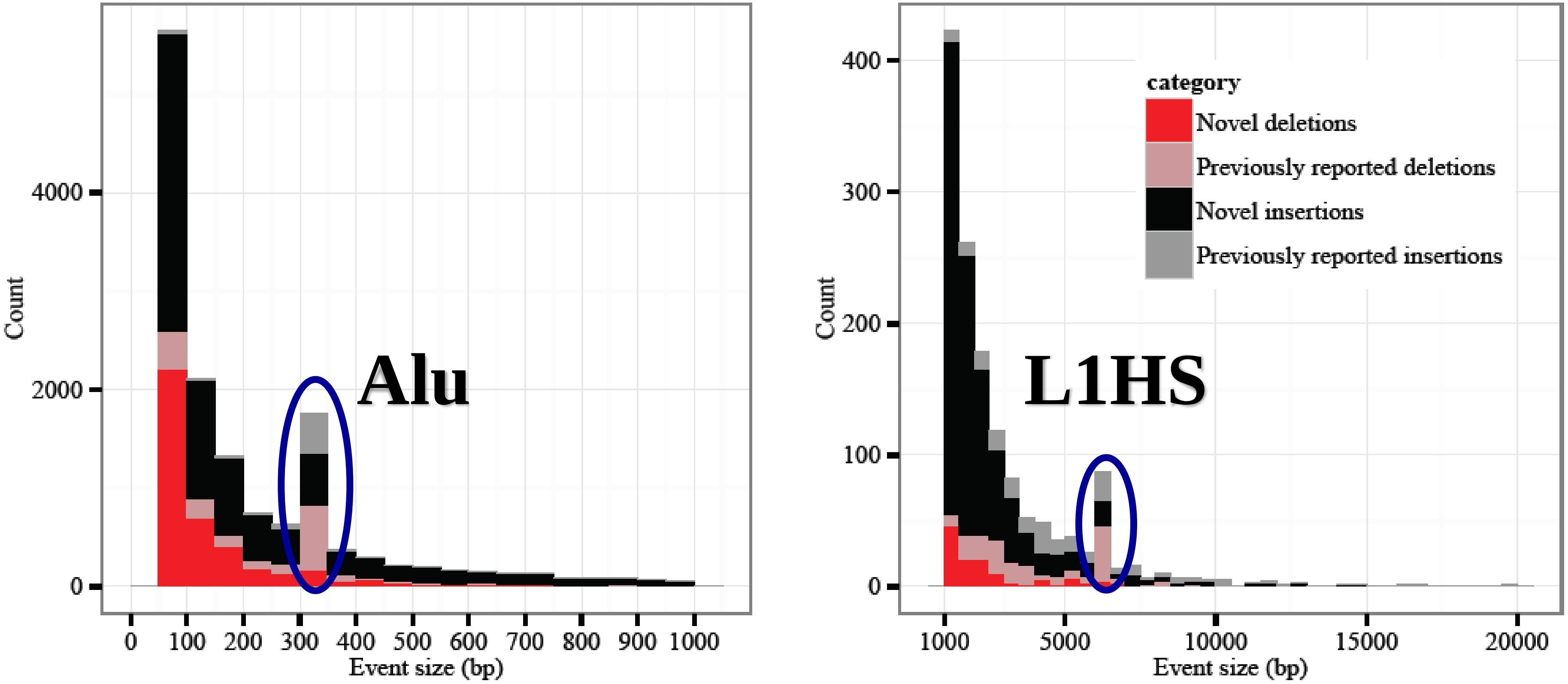
Structural variation detection using SMRT-SV on complete hydatidiform moles


23X
23X
46XX
Monospermic CHM
Loss of maternal chromosomes (before or after fertil),
duplication of paternal chromosomes



Chaisson et al, Nature, 2015

Increased Resolution of Structural Variation



92% of insertions and 60% deletions (30- 5,000 bp) are novel

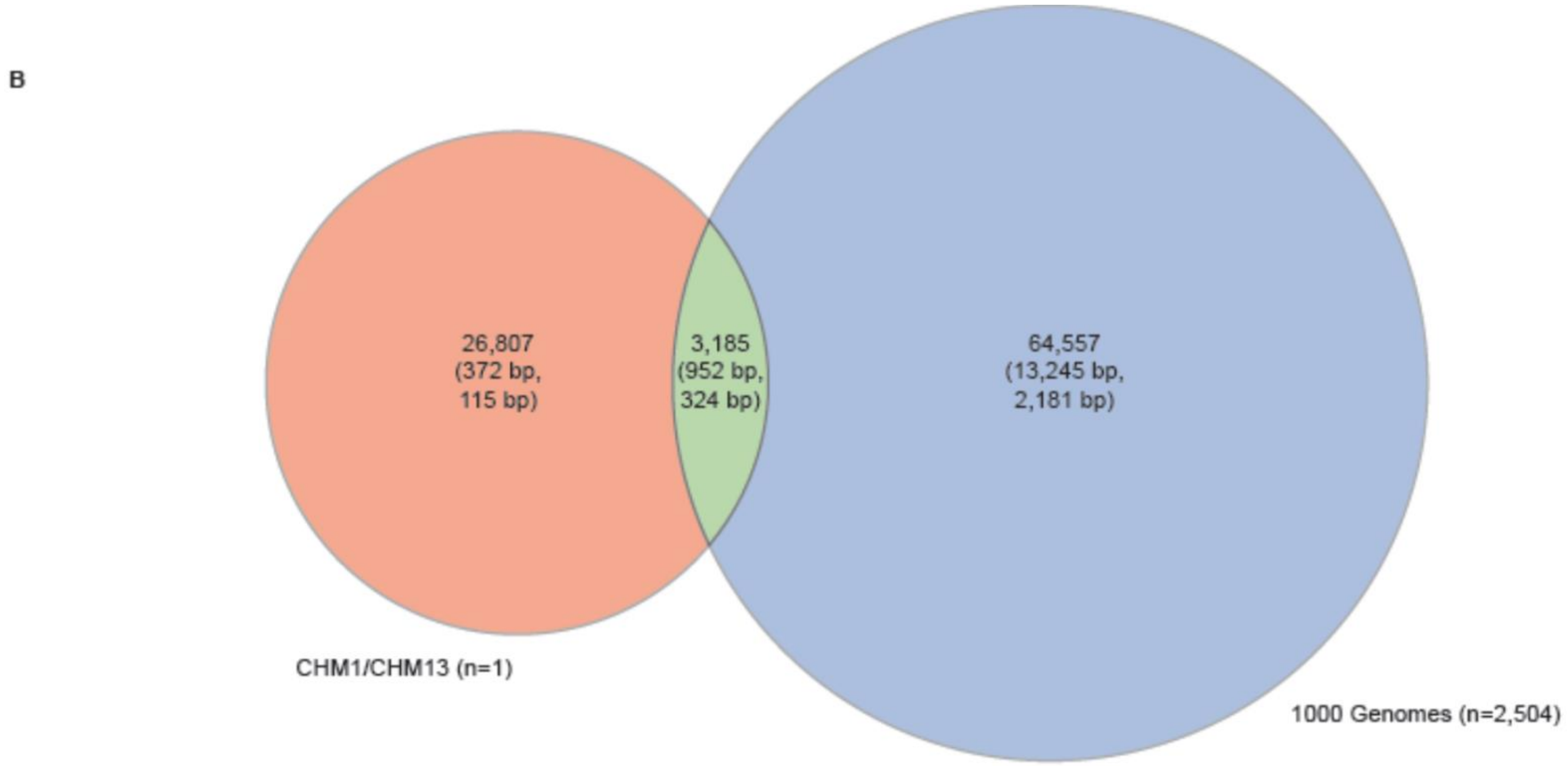
22,112 novel genetic variants corresponding to 11 Mbp of sequence

6,796 of the events map within 3,418 genes

169 within coding sequence or UTRs of genes

Chaisson et al, Nature, 2015

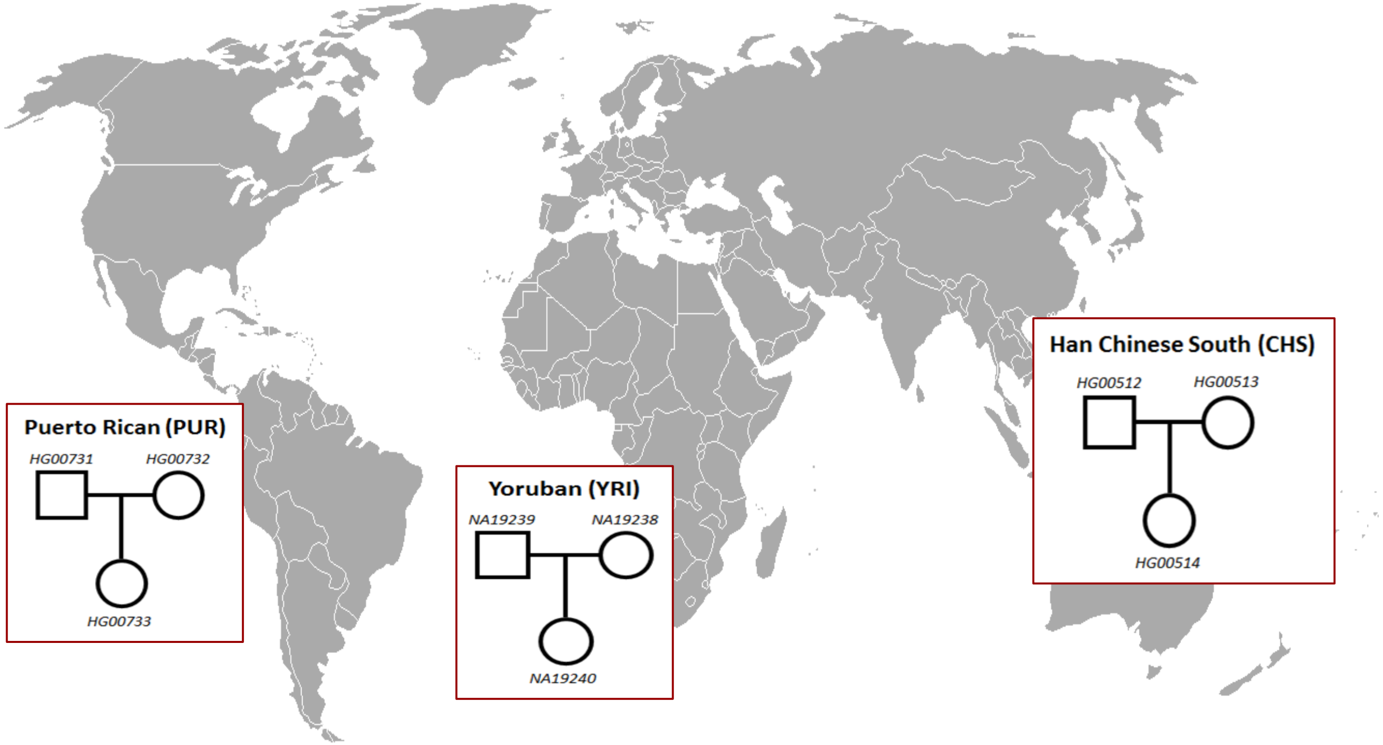
In Silico Diploid Genome: CHM1+CHM13

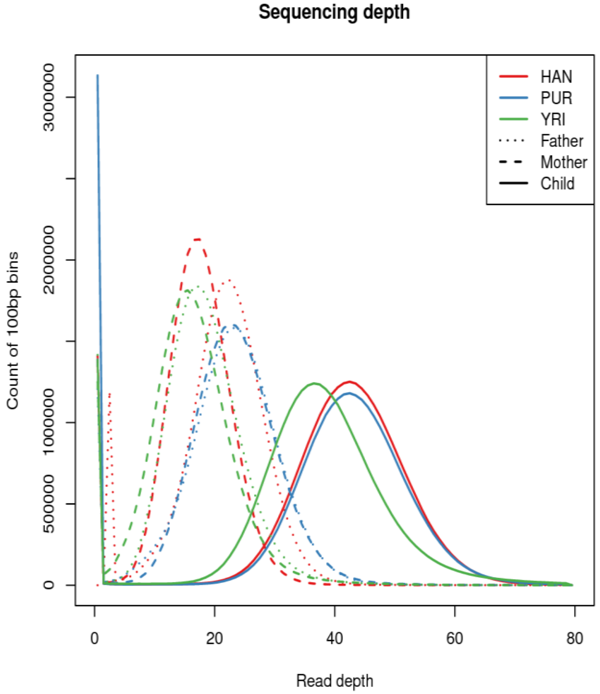


- two haploid human genomes full phased = 29,992 distinct SV events
- 30% of it missed by a naïve SMRT-SV caller that did not phase
- 89% of variants missed by the 1000 Genomes Project even after adjusting for common variants (MAF>1%)

Huddleston et al, *Genome Res*, 2016

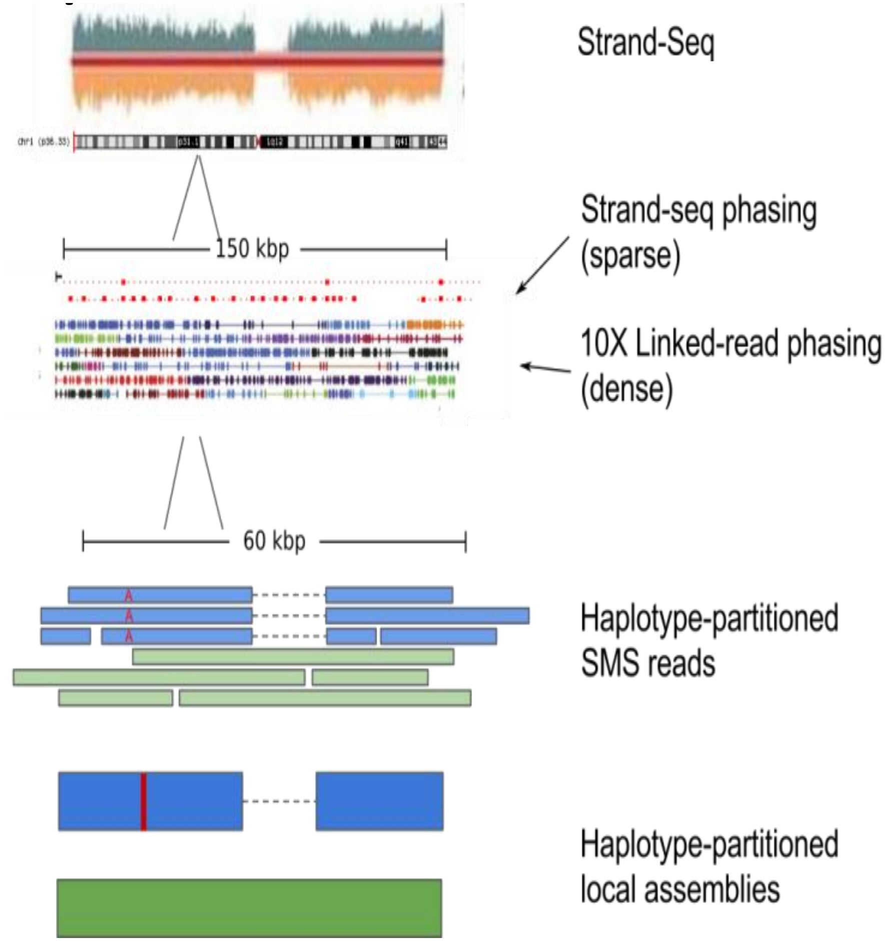
Human Genome Structural Variation Consortium (HGSVC)





- Establish gold standards for human genome SV
- Sequence three trios deeply with multiple platforms (Illumina, PacBio, 10X, Strand-seq, Bionano Genomics and one with ONT)

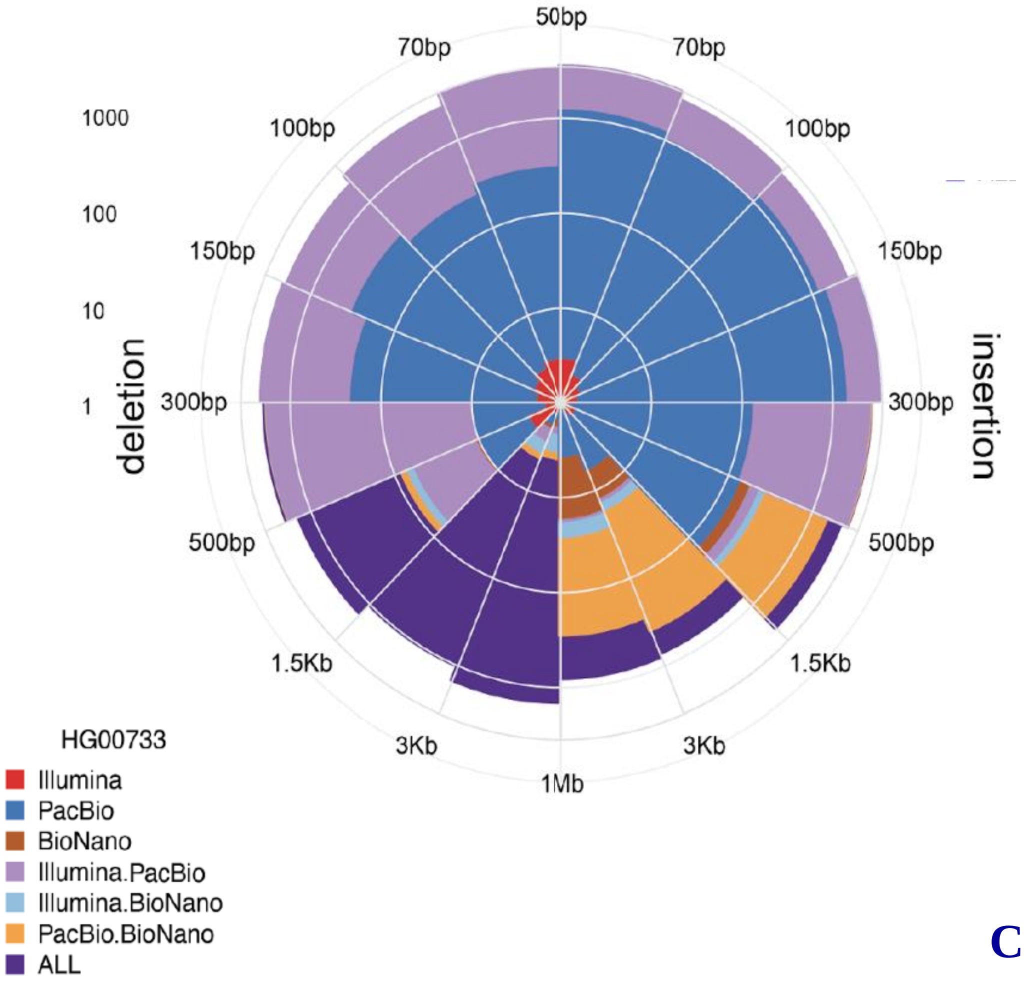
Phased-SV: Comprehensive SV Detection of a Diploid


Strand-Seq
150 kbp
Strand-seq phasing (sparse)
10X Linked-read phasing (dense)
60 kbp
Haplotype-partitioned SMS reads
Haplotype-partitioned local assemblies

- Strand-seq and 10-X linked read data are used to phase 70% of all PacBio Reads
- SVs are called using haplotype-type partitioned reads that are locally assembled
- 3-fold increase in sensitivity compared to 11-Illumina callers (30,000 vs. 11,000 events)

Chaisson *et al, Biorxiv, 2017/Nat Comm,2019*

Sequencing Platform Comparison for SV Detection

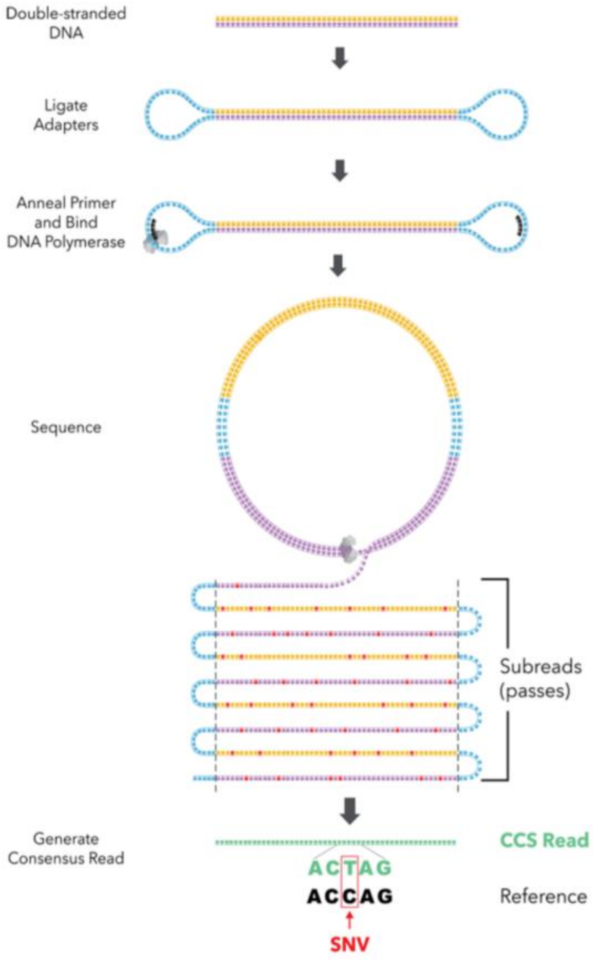


- ~30,000 PB vs. 11,000 Illumina SVs
- Illumina WGS at 30-40 fold sequence coverage combining results from 11 different SV callers (including Lumpy, GenomeStrip, Manta, WhamG etc) detects **a maximum of 49% of deletions and 11% of insertions in a human genome**
- **Large scale studies of WGS are identifying ~27% of SV variation events**
- Most of missing variation between 50-500 bp

Chaisson *et al, Biorxiv, 2017/Nat Comm,2019*

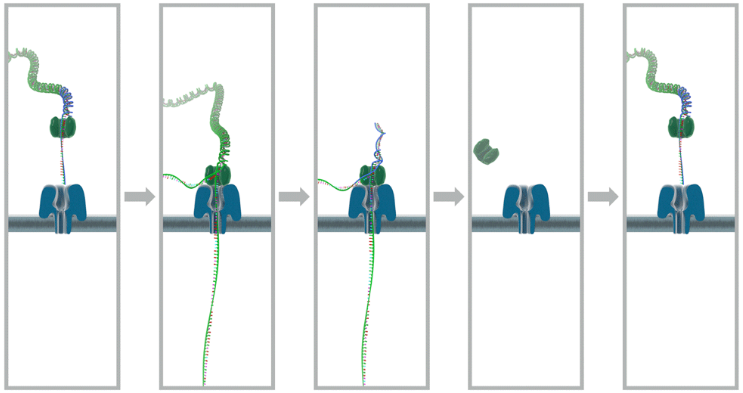
Advances in Long-Read Sequencing

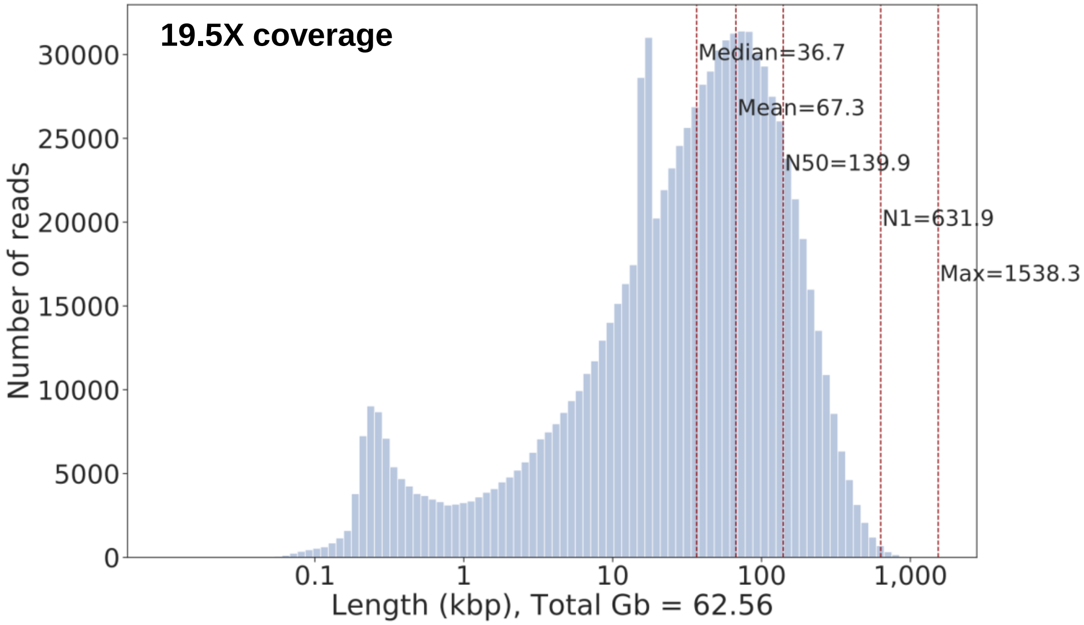
HiFi Pac Bio Sequencing



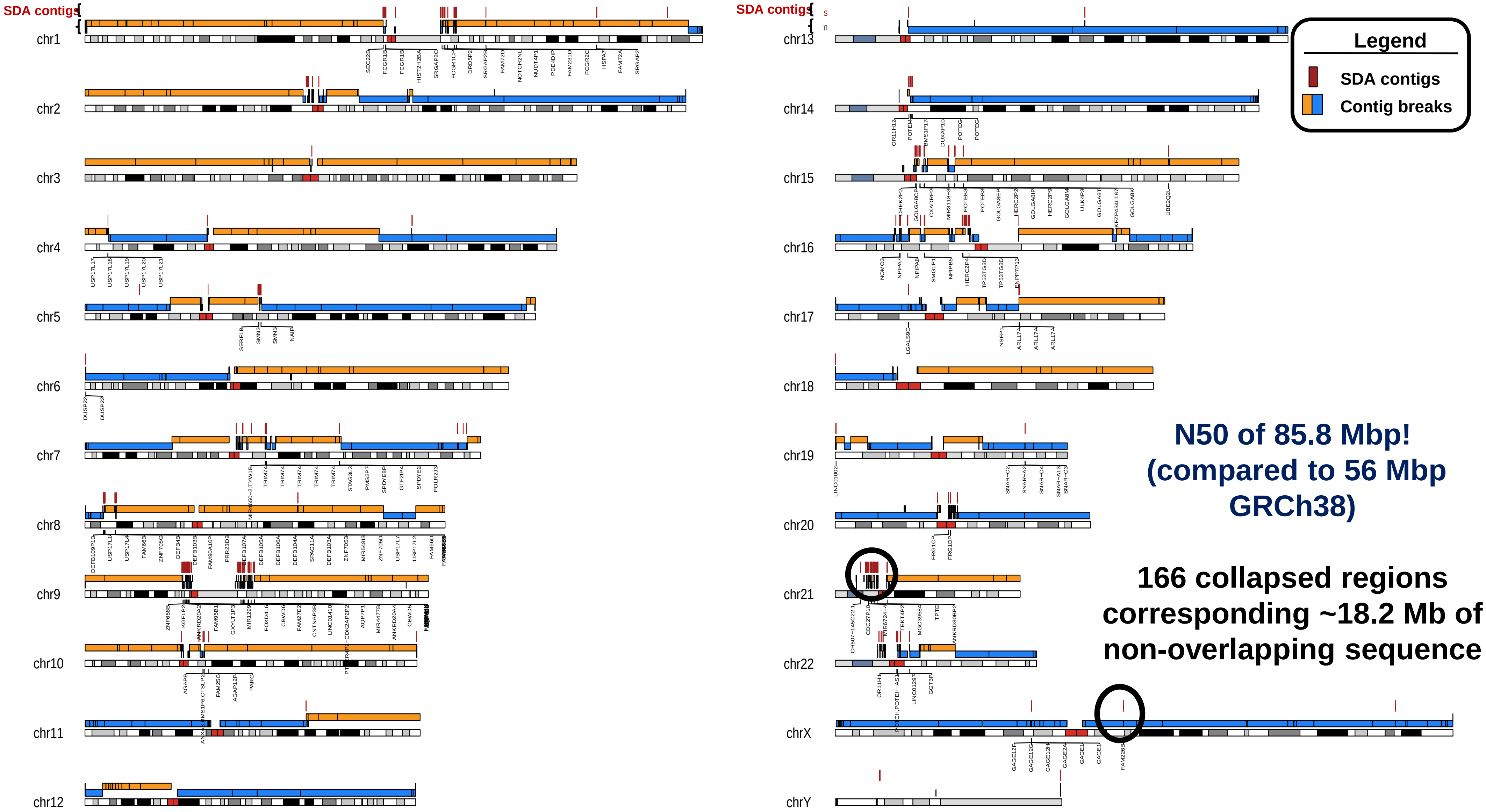
99.9% accurate 18 kbp reads

Ultra-long reads ONT





Telomere-to-telomere assembly of CHM13

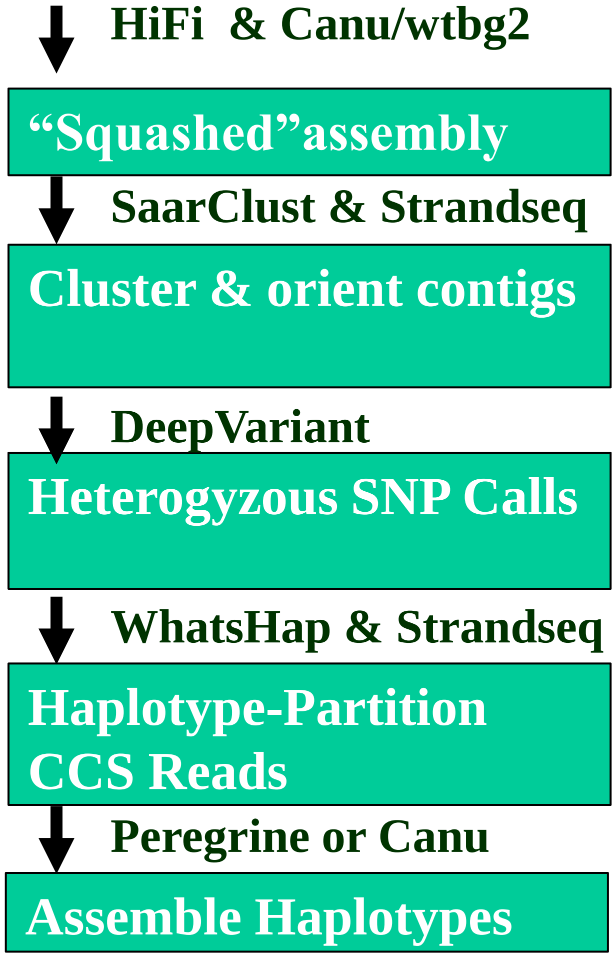


N50 of 85.8 Mbp!
(compared to 56 Mbp GRCh38)

166 collapsed regions corresponding ~18.2 Mb of non-overlapping sequence

Miga et al, biorxiv , 2019

Reference-free long-read phased diploid genomes (HiFi & Strandseq)



HiFi & Canu/wtbg2

“Squashed”assembly

SaarClust & Strandseq

Cluster & orient contigs

DeepVariant

Heterogyzous SNP Calls

WhatsHap & Strandseq

Haplotype-Partition CCS Reads

Peregrine or Canu

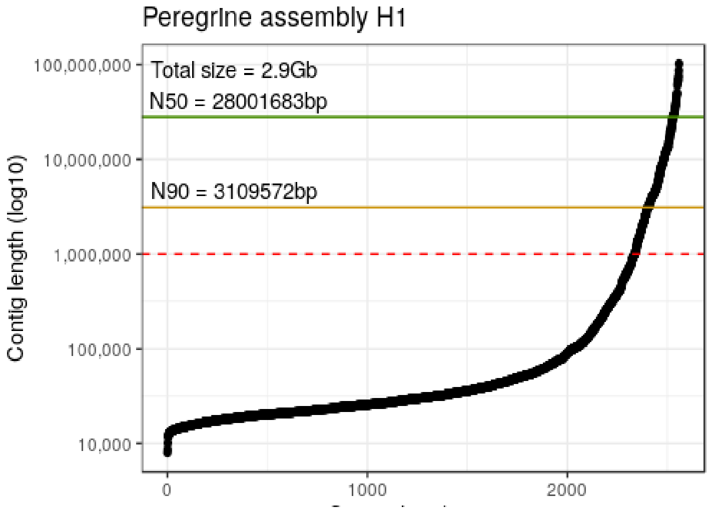
Assemble Haplotypes

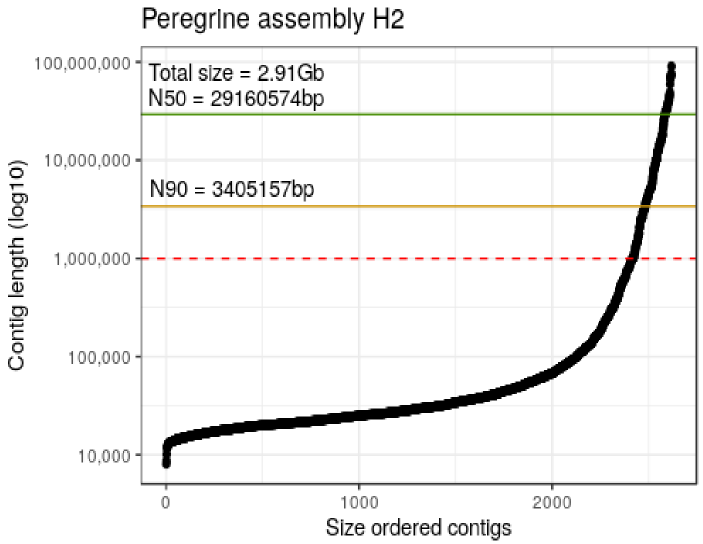
- 33.4-fold HiFi coverage from a 1000 Genomes Project Puerto Rican Genome HG00733 (sequence N50=13.4 kbp)
- Strand-seq: 2.87 X of linked reads (115 single-cell libraries) that allow chromosomal phasing
- 23 clusters where contigs are orientated without guidance from reference
- 95% of SNPs phased
- 81% of HiFi reads assigned to one of two haplotypes H1/H2
- ~5000 cpu-hours

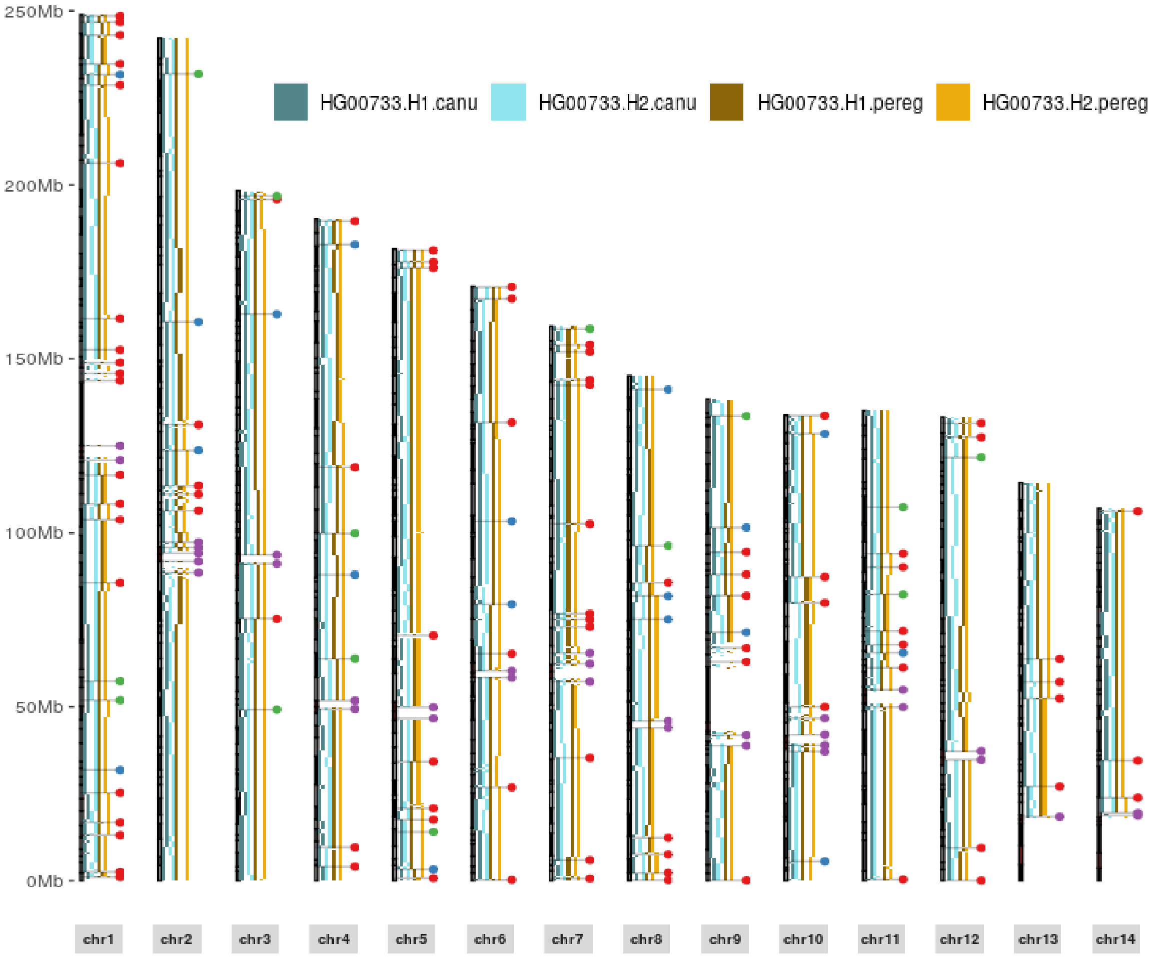
Porubsky, Ebert, Marschall et al. Biorxiv, 2019

Phased Assembly Contiguity

(Contig N50 H1=28.0 Mbp & H2=29.2 Mbp)

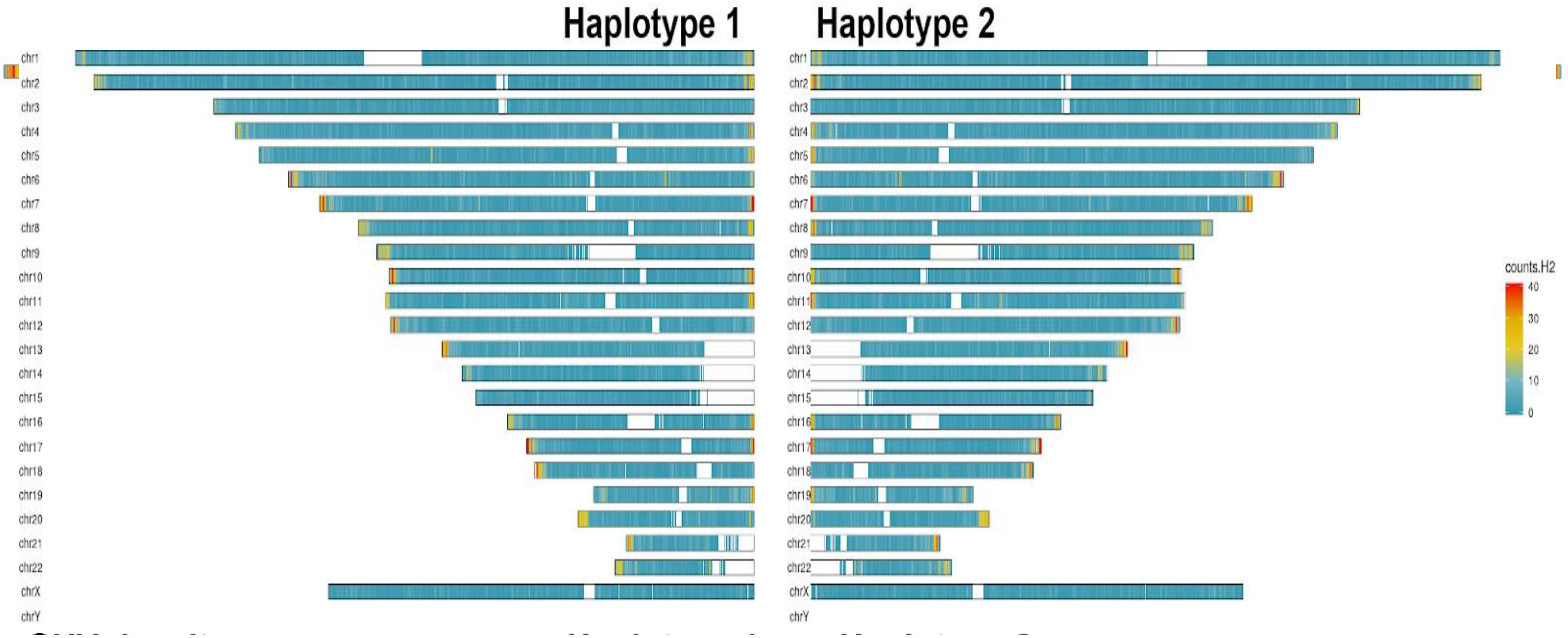






Contig N50 : the sequence length of the shortest contig at 50% of the total genome length

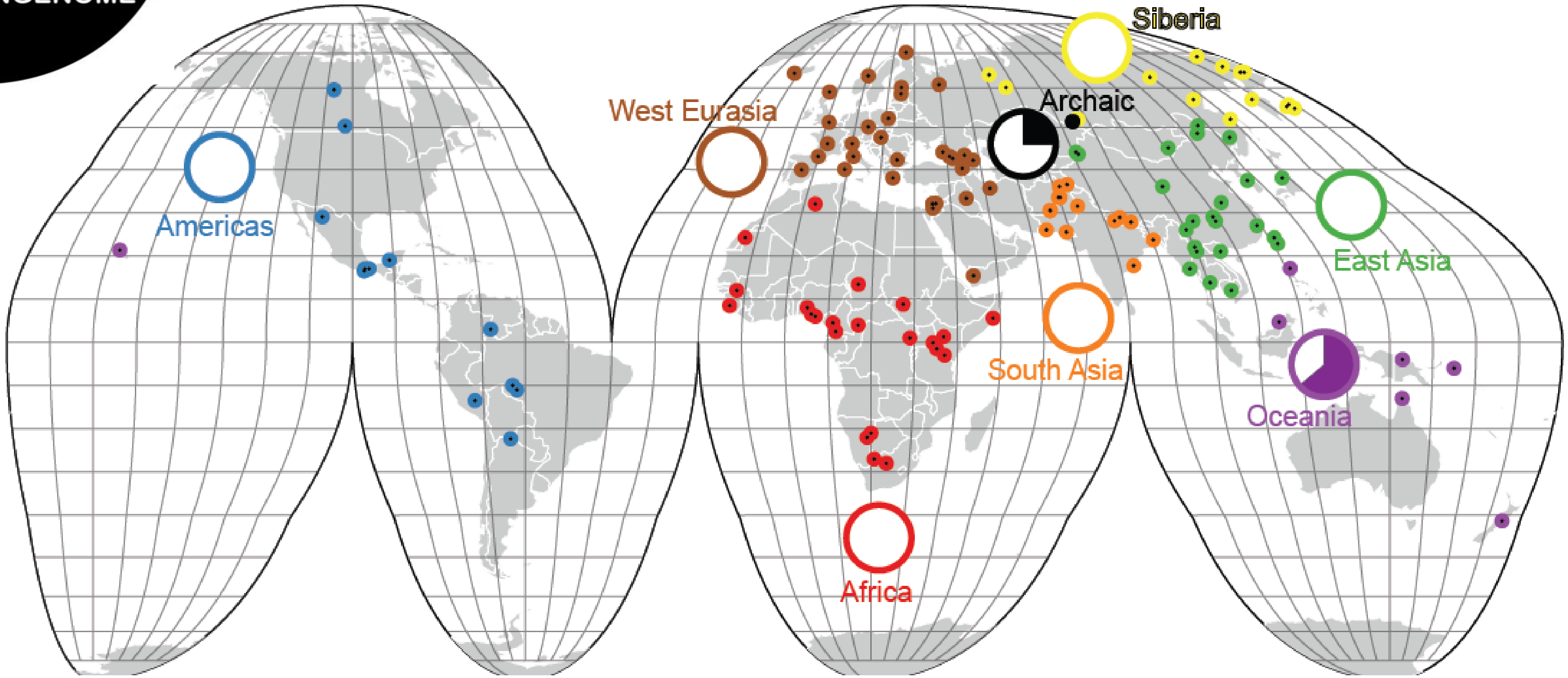
A 6 Gbp Human Genome Assembly



Porubsky et al. Biorxiv, 2019

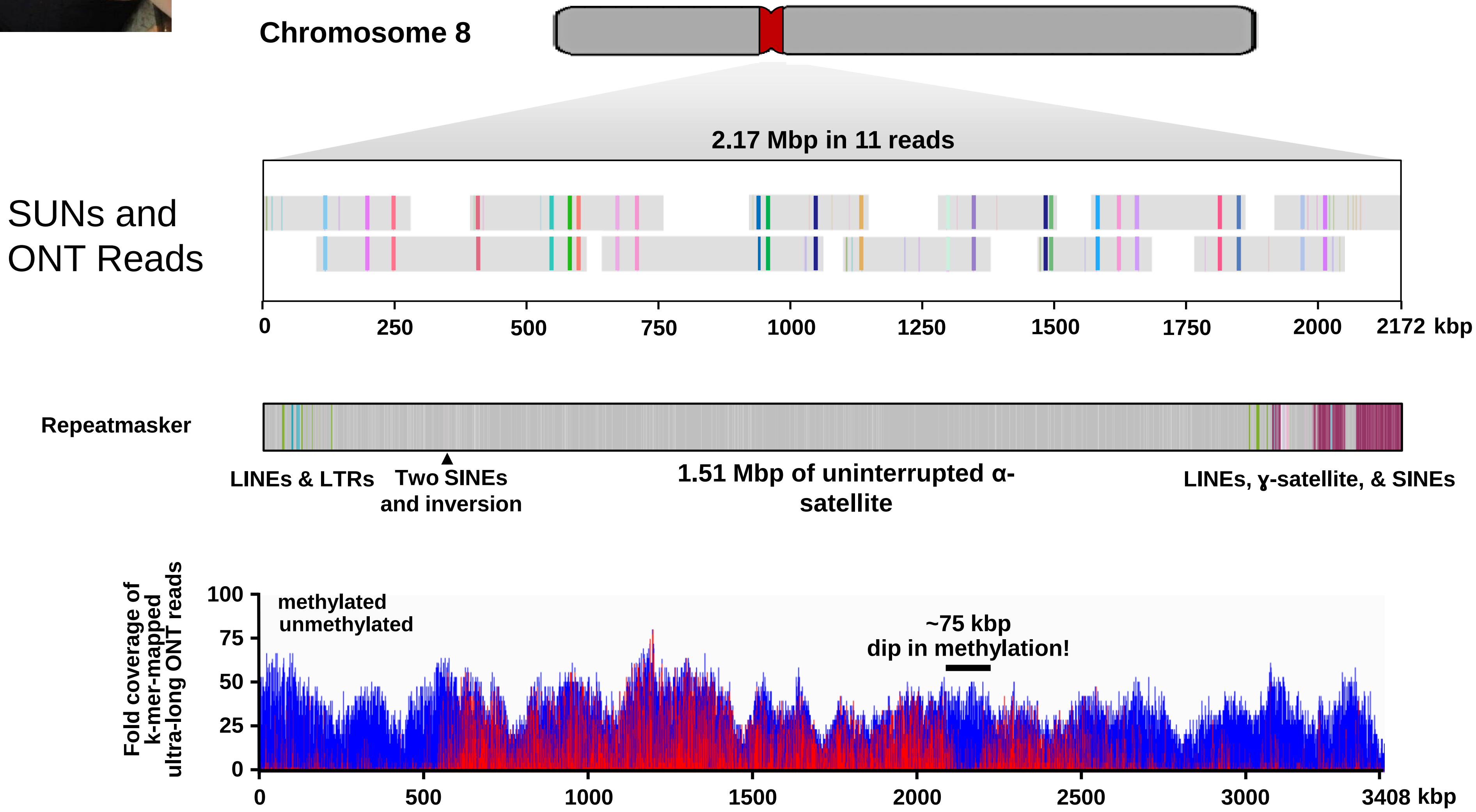
HUMAN PANGENOME

Human PanGenome Project



Goal: Telomere-to-telomere assembly of 350 human genomes over the next five years that represents the diversity of humanity

Sequence and assembly of chromosome 8 centromere

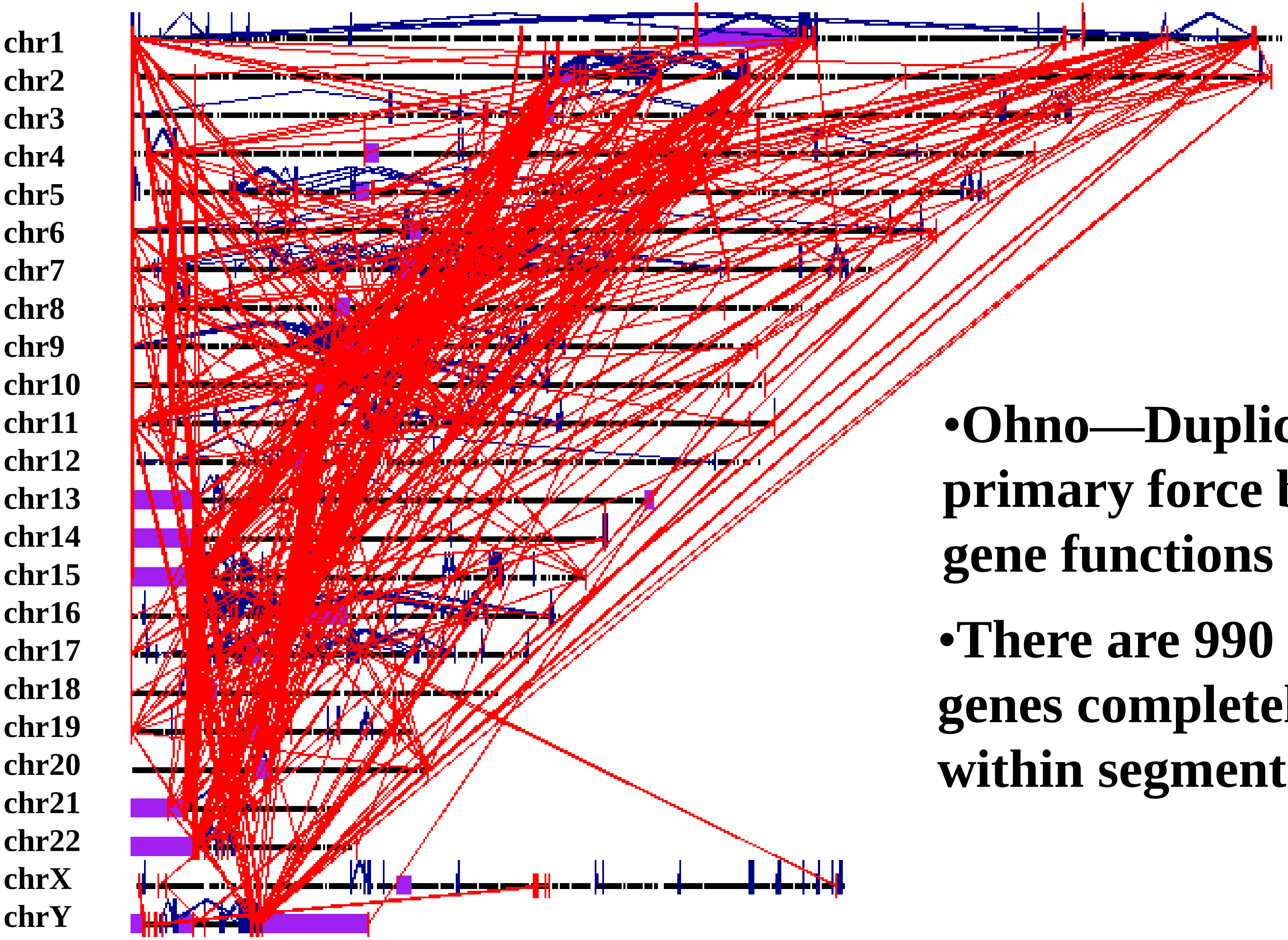


Logsdon and T2T, unpublished

Summary

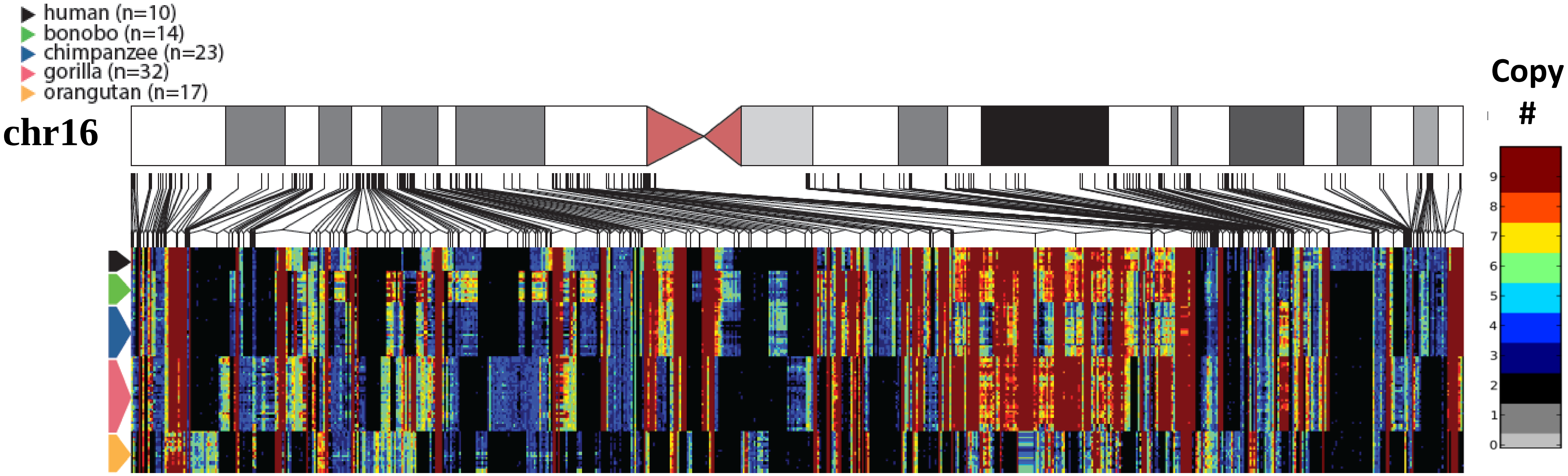
- Short read NGS approaches
  - Multiple methods need to be employed with short reads—Readpair+Read-depth+SplitRead coupled to an orthogonal method such as SNP microarray for validation
  - Tradeoff between sensitivity and specificity
  - 25% of SVs can be reliably detected because SVs is non-randomly distributed to repetitive regions
  - Read-depth approaches allow prediction of copy number in more complex regions but do not provide structure
- Third generation sequencing methods provide comprehensive assessment but limited throughput
  - Initial methods based on detection of specific signatures and local assembly
  - Ultimate is haplotype-resolved assembled genomes

III. Why?



- Ohno—Duplication is the primary force by which new gene functions are created
- There are 990 annotated genes completely contained within segmental duplications

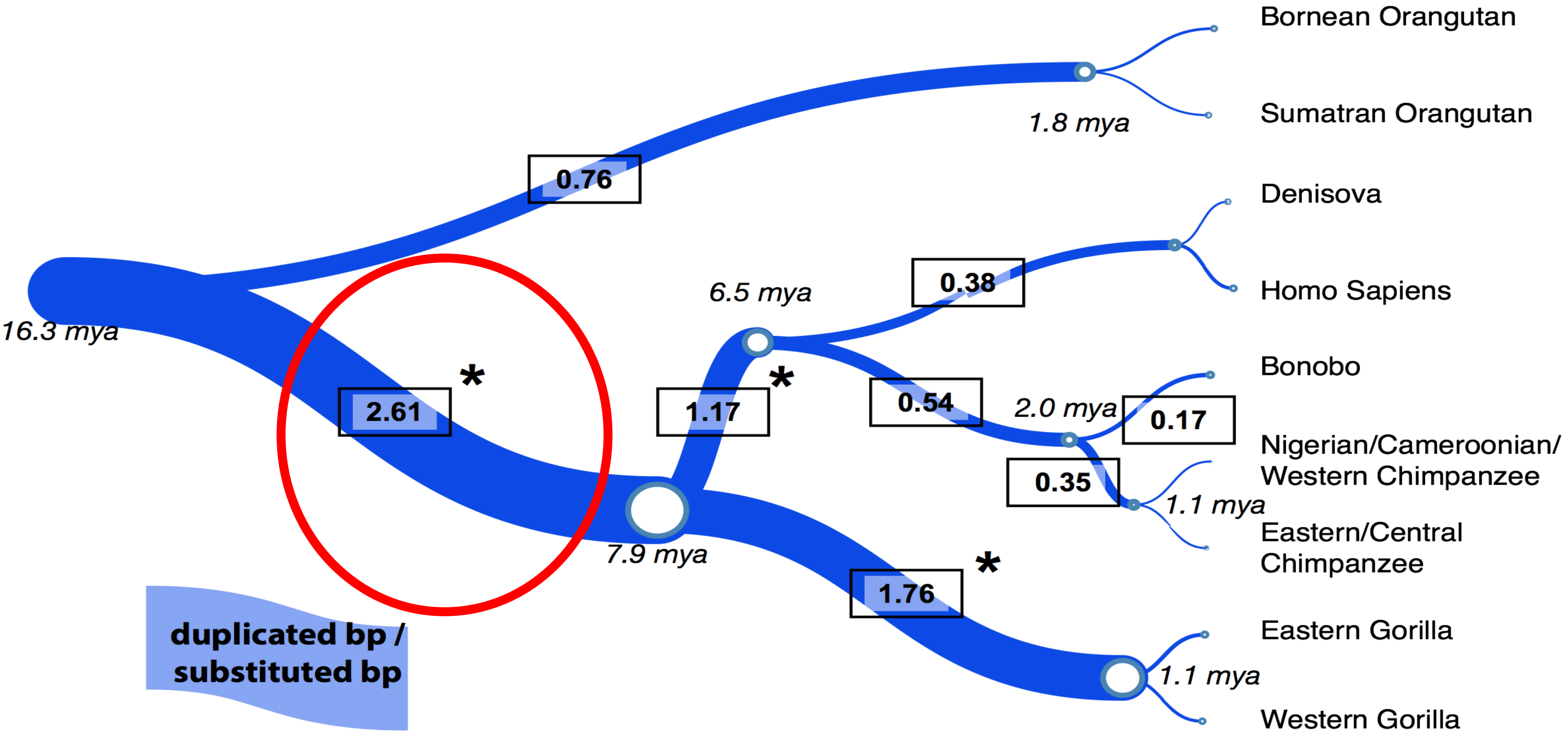
Dynamic Genetic Variation



- **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*

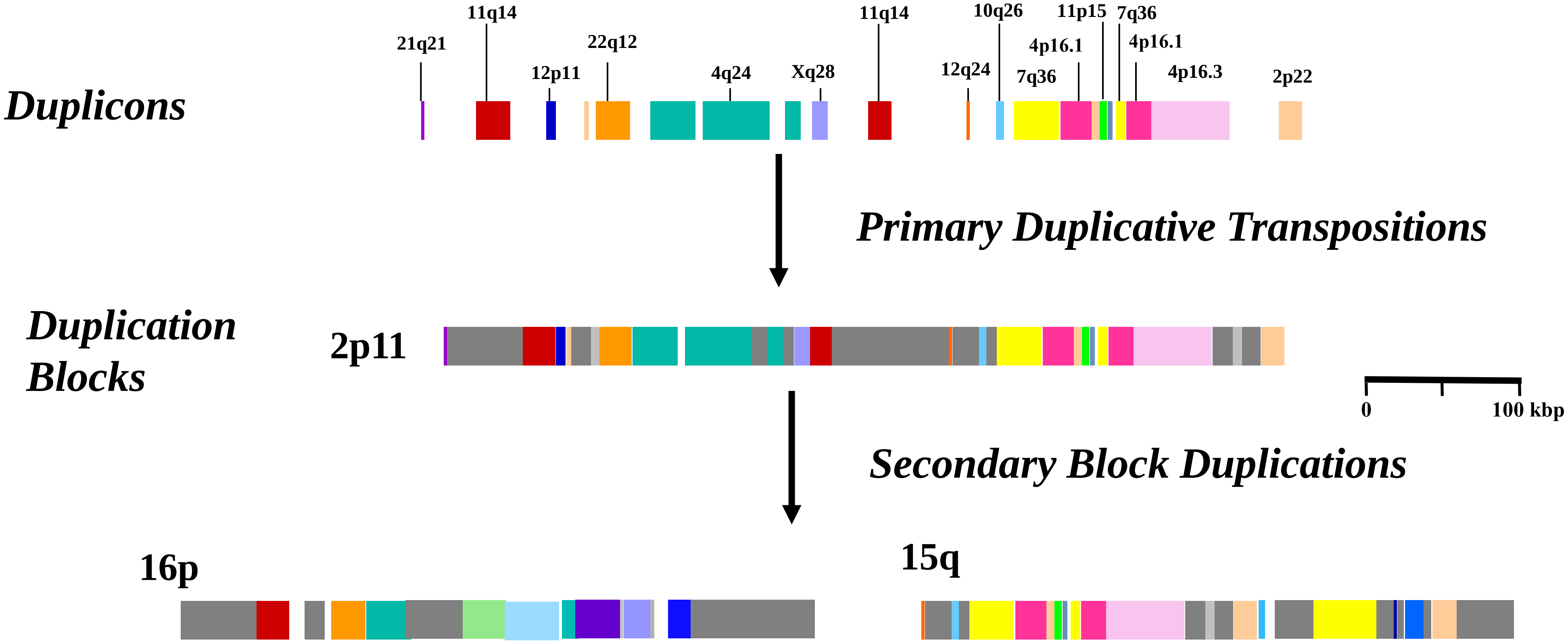
Rate of Duplication



p=9.786 X 10^{-12}

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

Mosaic Architecture


Duplicons

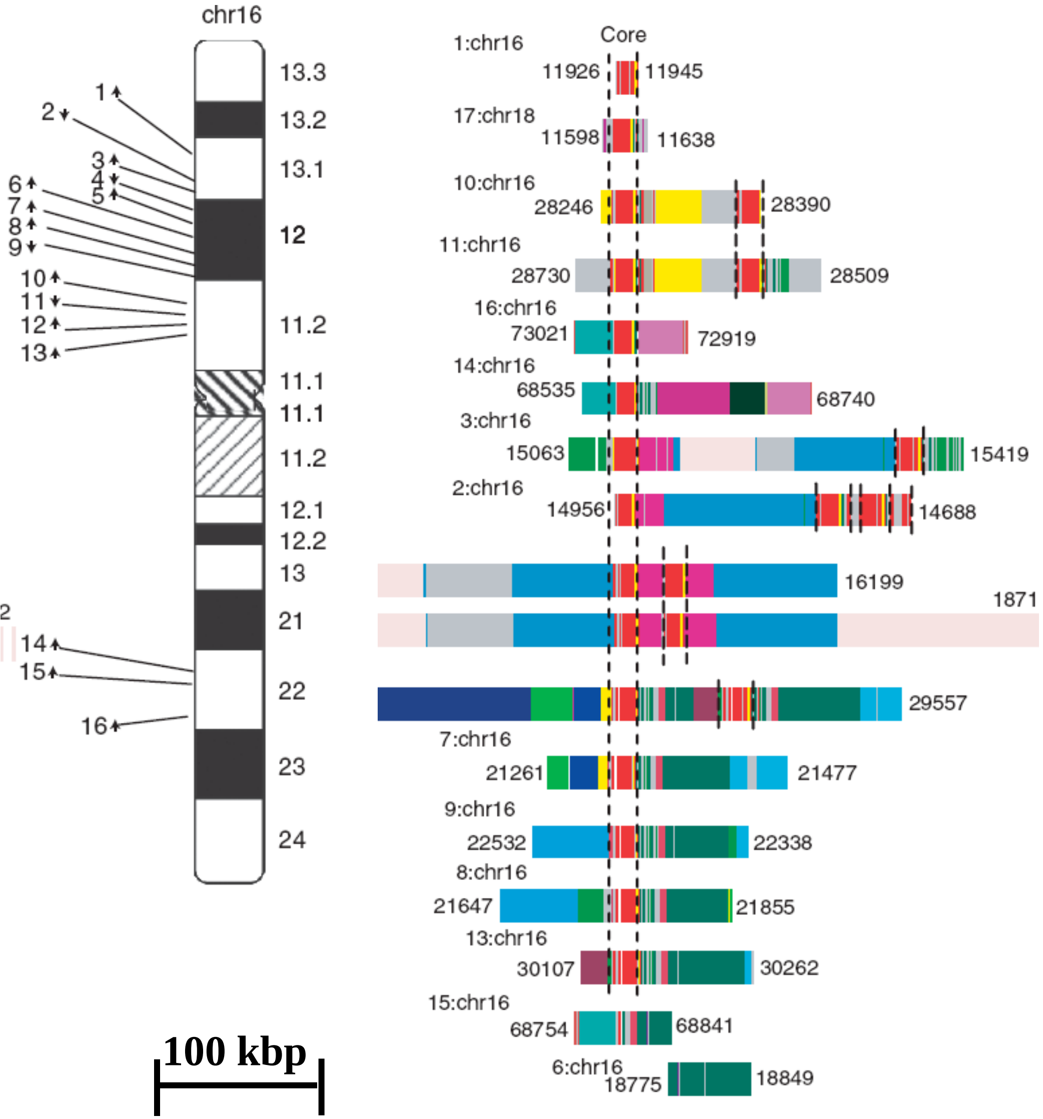
Primary Duplicative Transpositions

Duplication Blocks

Secondary Block Duplications

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

Human Chromosome 16 Core Duplicon

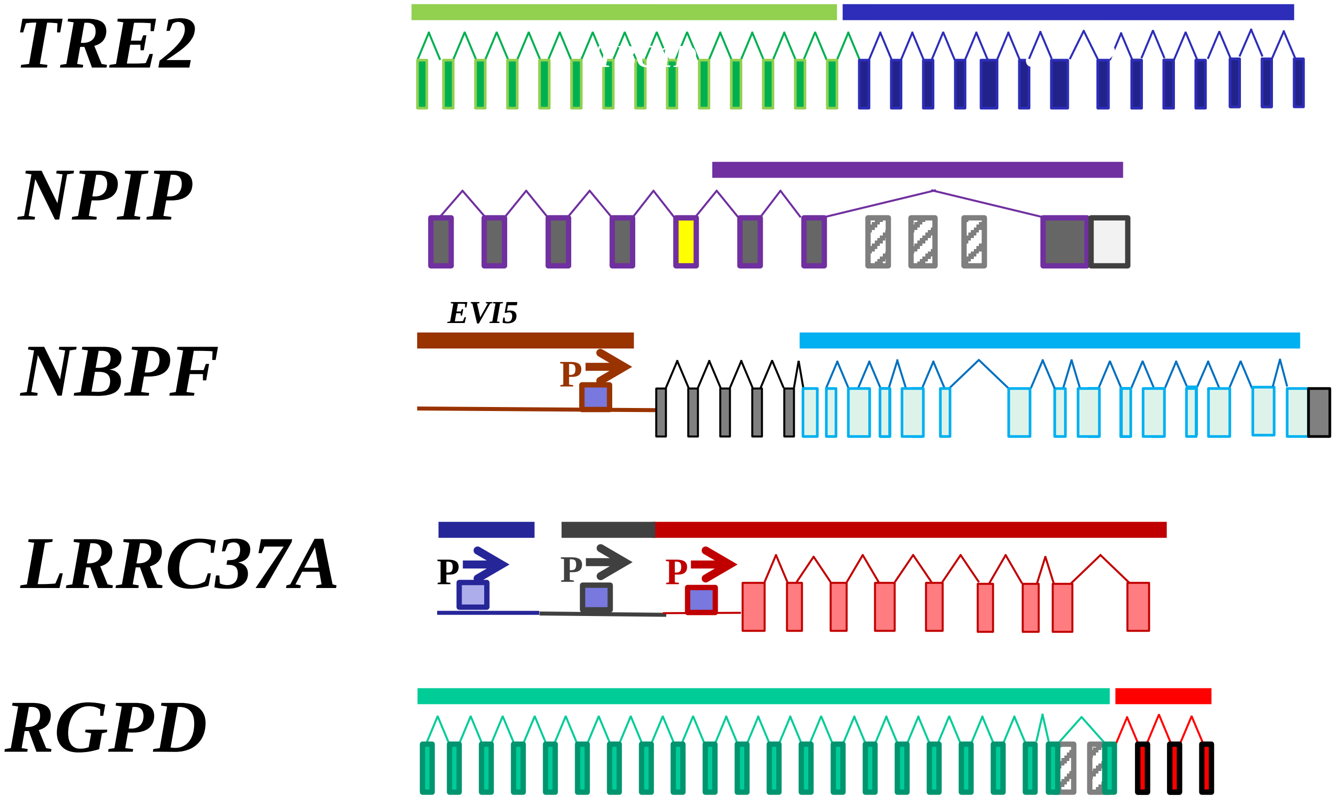


- The burst of segmental duplications 8-12 mya corresponds to core-associated duplications which have occurred on six human chromosomes (chromosomes 1,2, 7, 15, 16, 17)

- Most of the <u>recurrent</u> 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



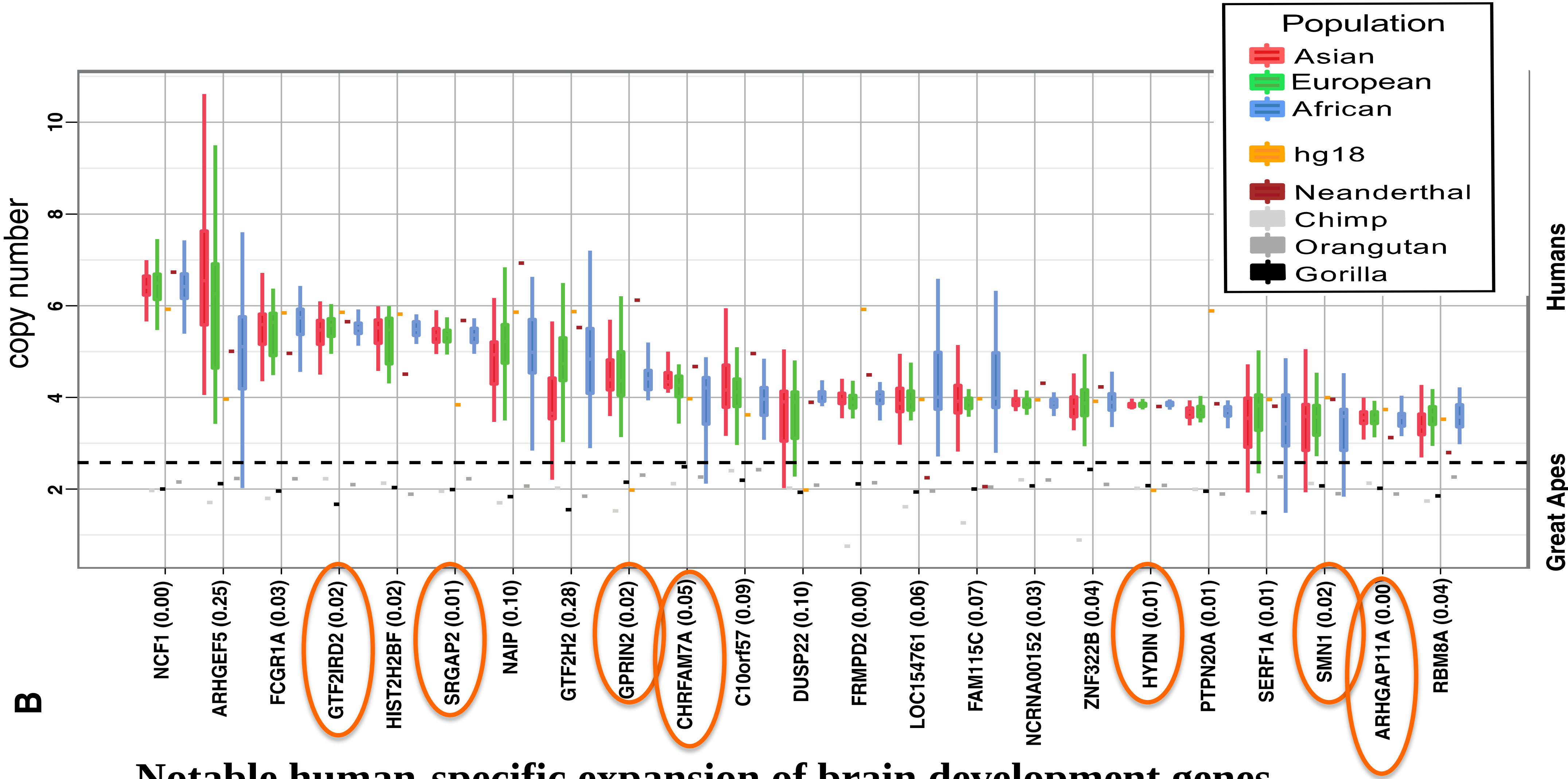
Features: No orthologs in mouse; multiple copies in chimp & human dramatic changes in expression profile; signatures of positive selection

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

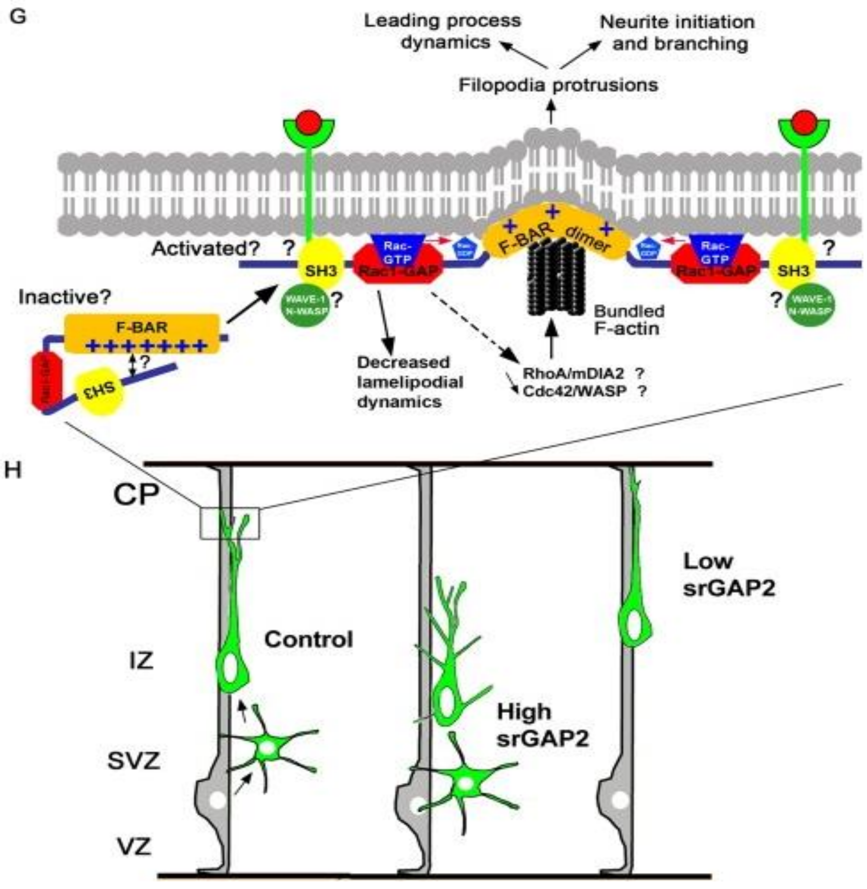


Notable human-specific expansion of brain development genes.
Neuronal cell death: p=5.7e-4; Neurological disease: p=4.6e-2

Sudmant et al., *Science*, 2010

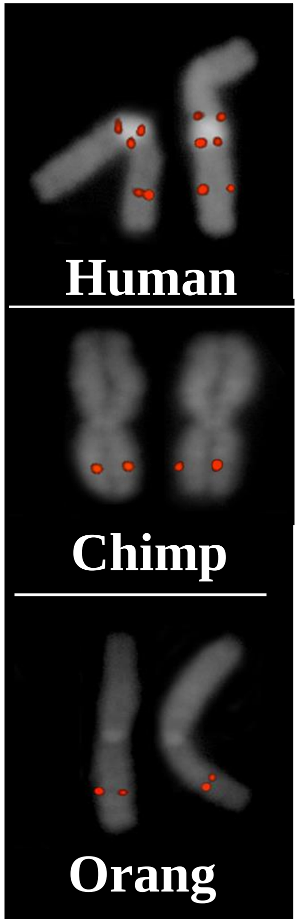
SRGAP2 function

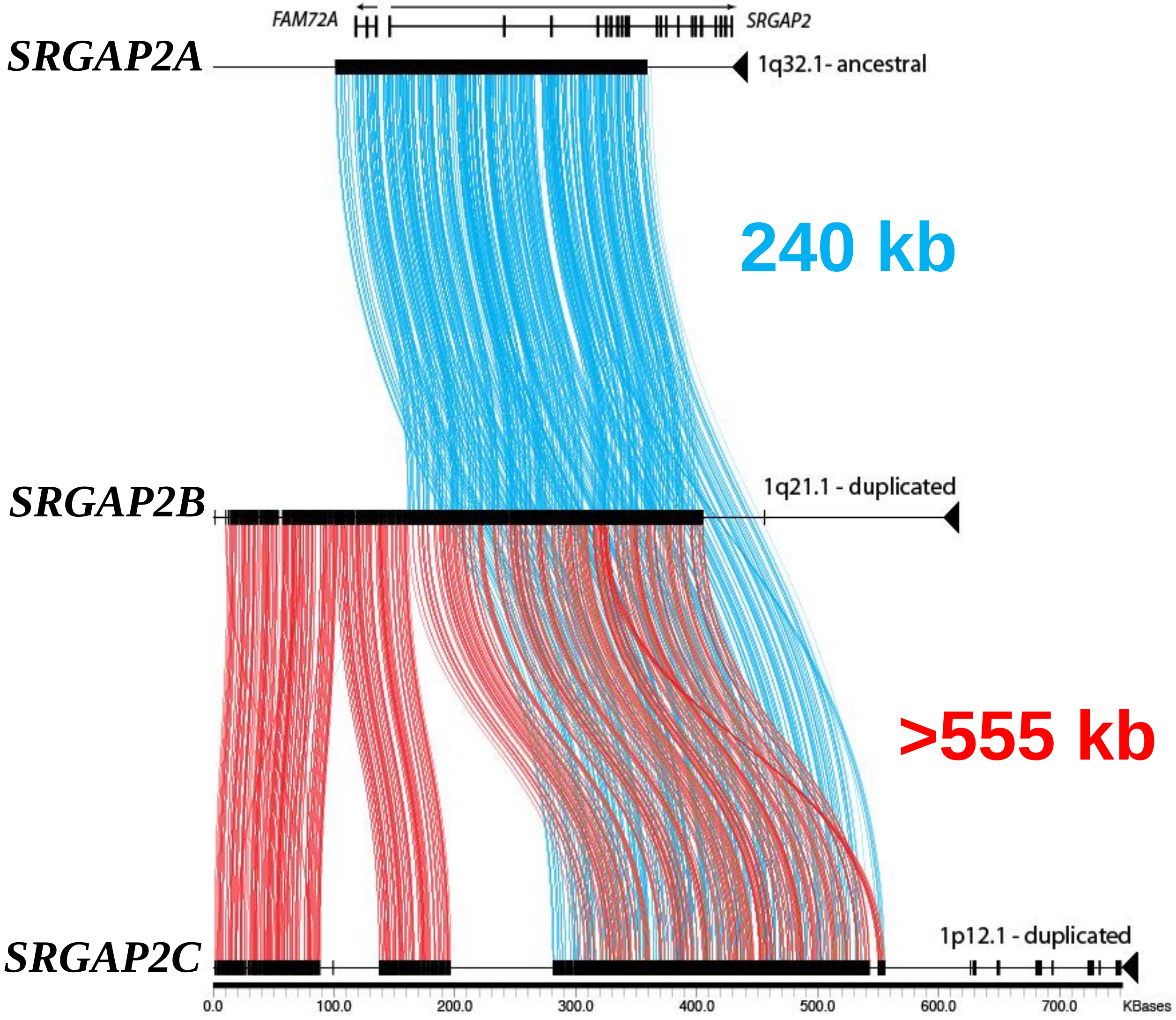
- *SRGAP2* (**SLIT**-ROBO Rho GTPase activating protein 2) functions to control migration of neurons and dendritic formation in the cortex
- Gene has been duplicated three times in human and no other mammalian lineage
- Duplicated loci not in human genome

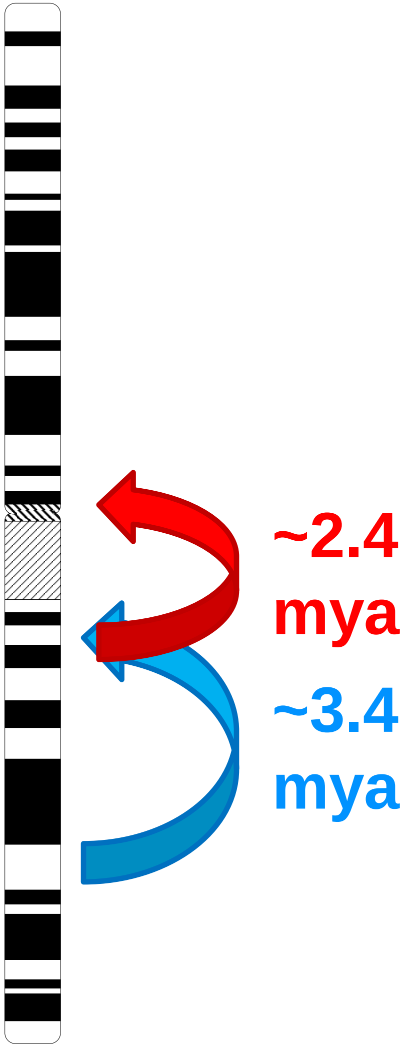


Guerrier et al., *Cell*, 2009

SRGAP2 Human Specific Duplication



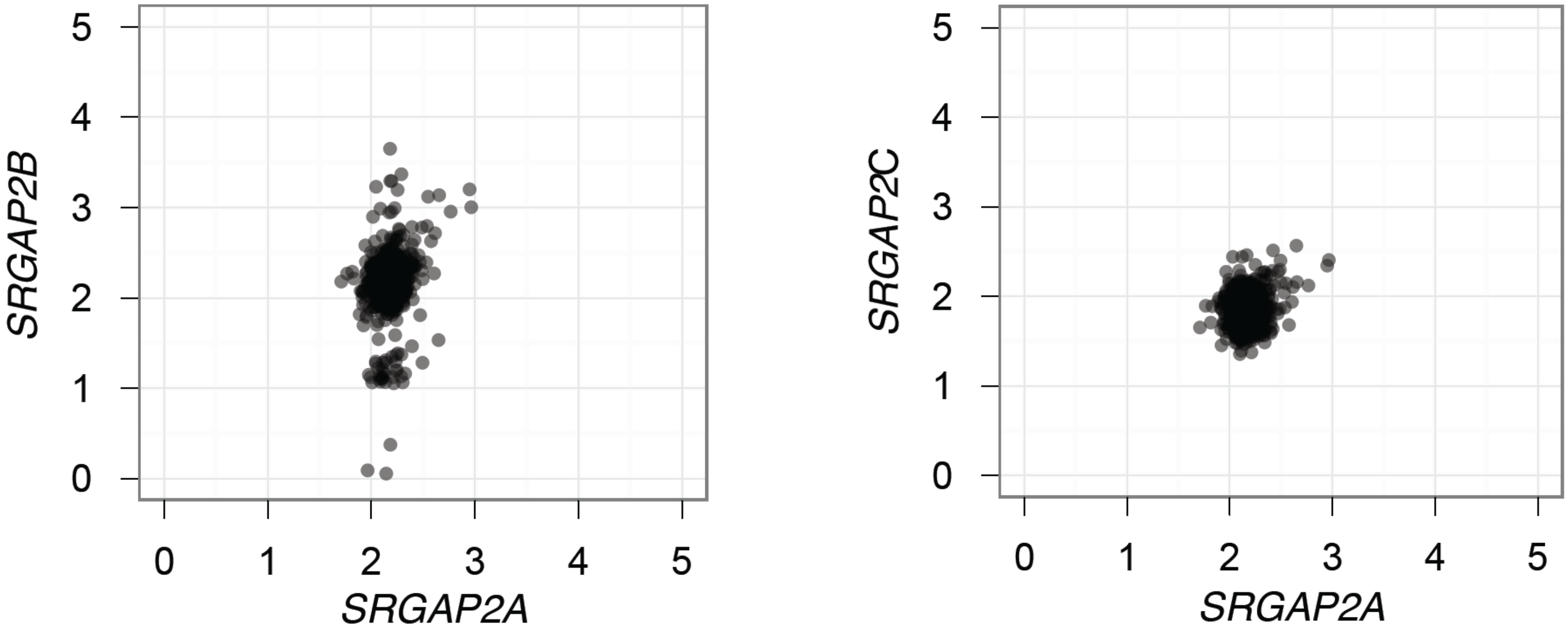




Dennis, Nuttle et al., *Cell*, 2012

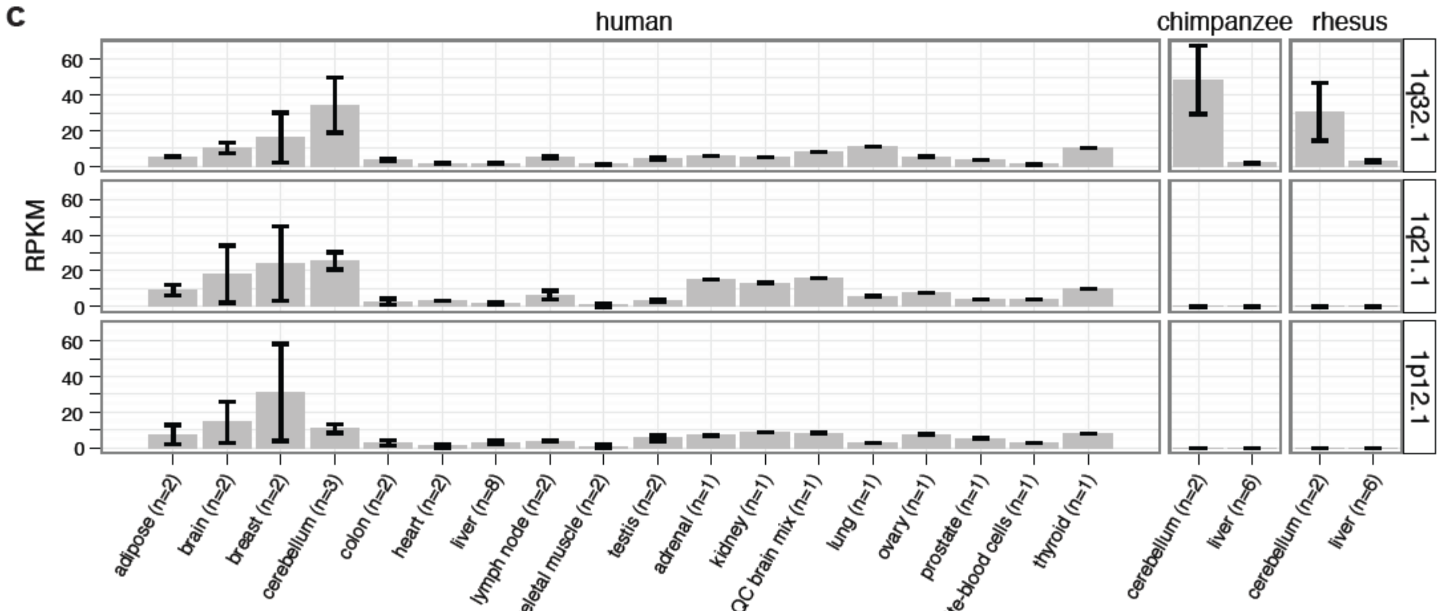
SRGAP2C is fixed in humans

(n=661 individual genomes)

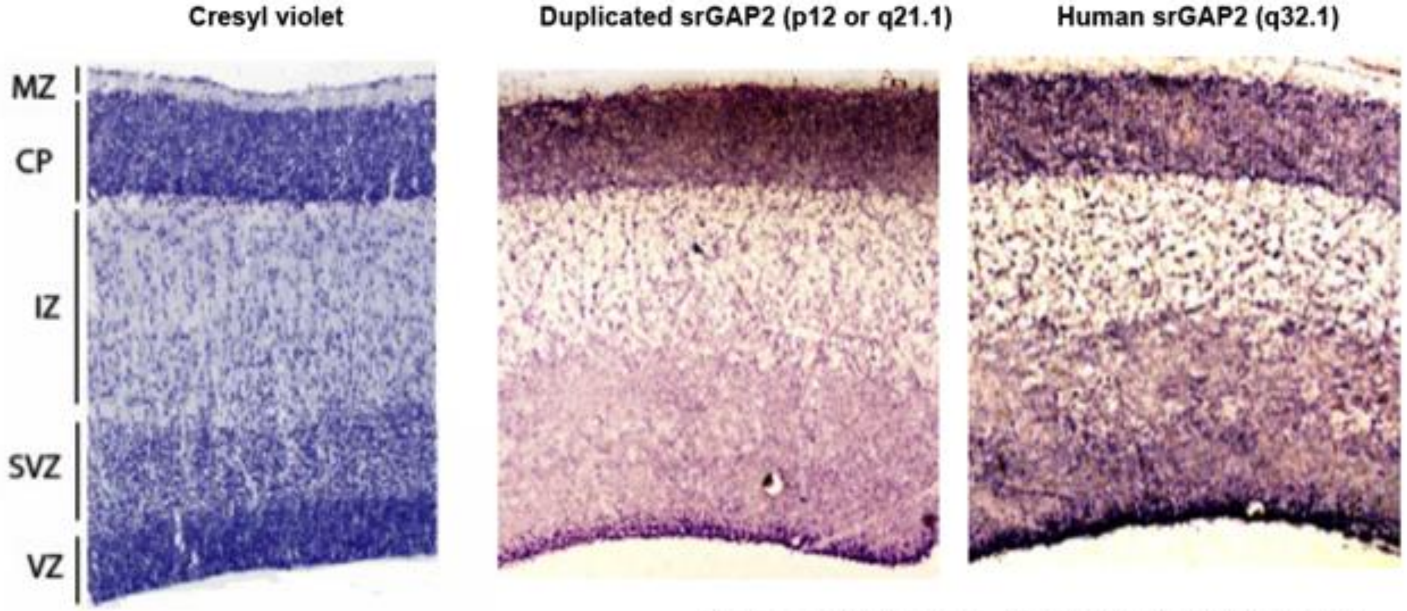


SRGAP2 duplicates are expressed

RNAseq

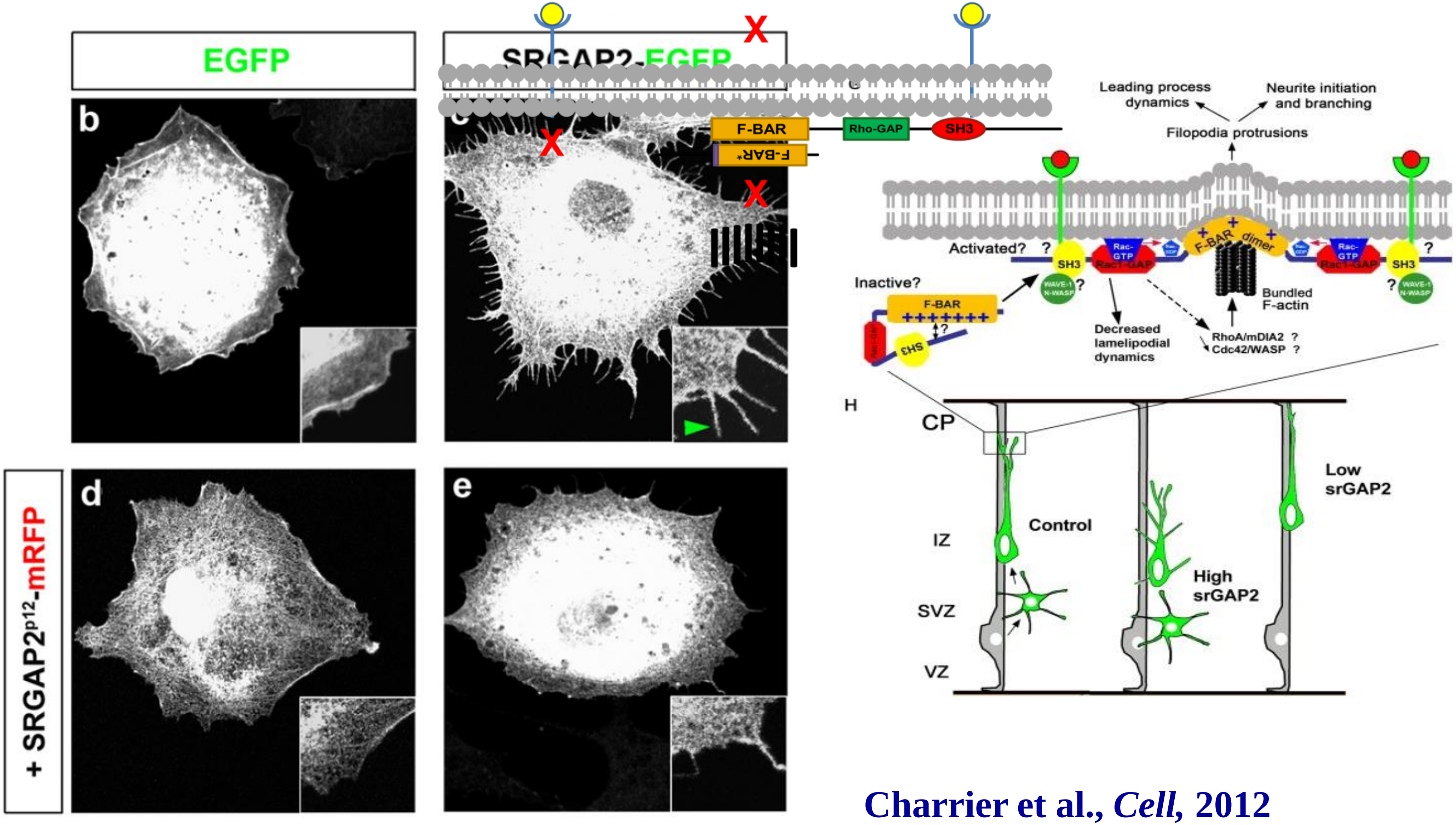


In situ

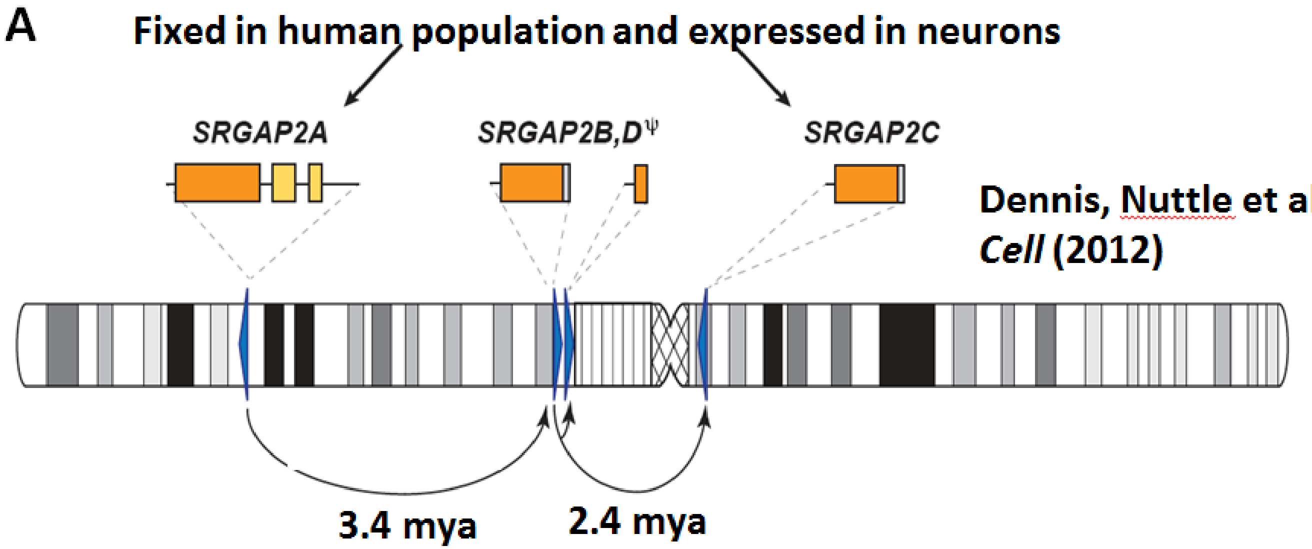


Human embryos Gestational Week 12

SRGAP2C duplicate antagonizes function

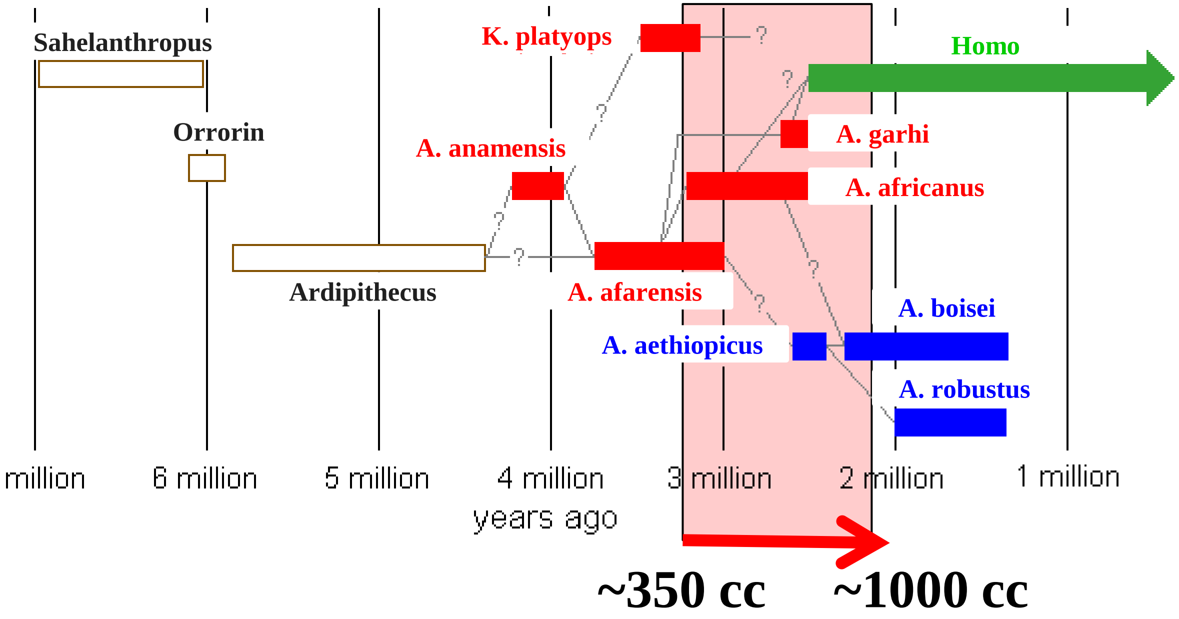


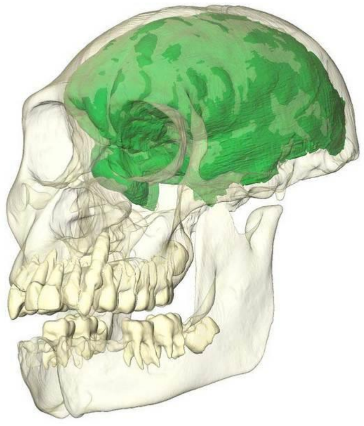
Charrier et al., *Cell*, 2012



Dennis, Nuttle et al.
Cell (2012)

Australopithecus

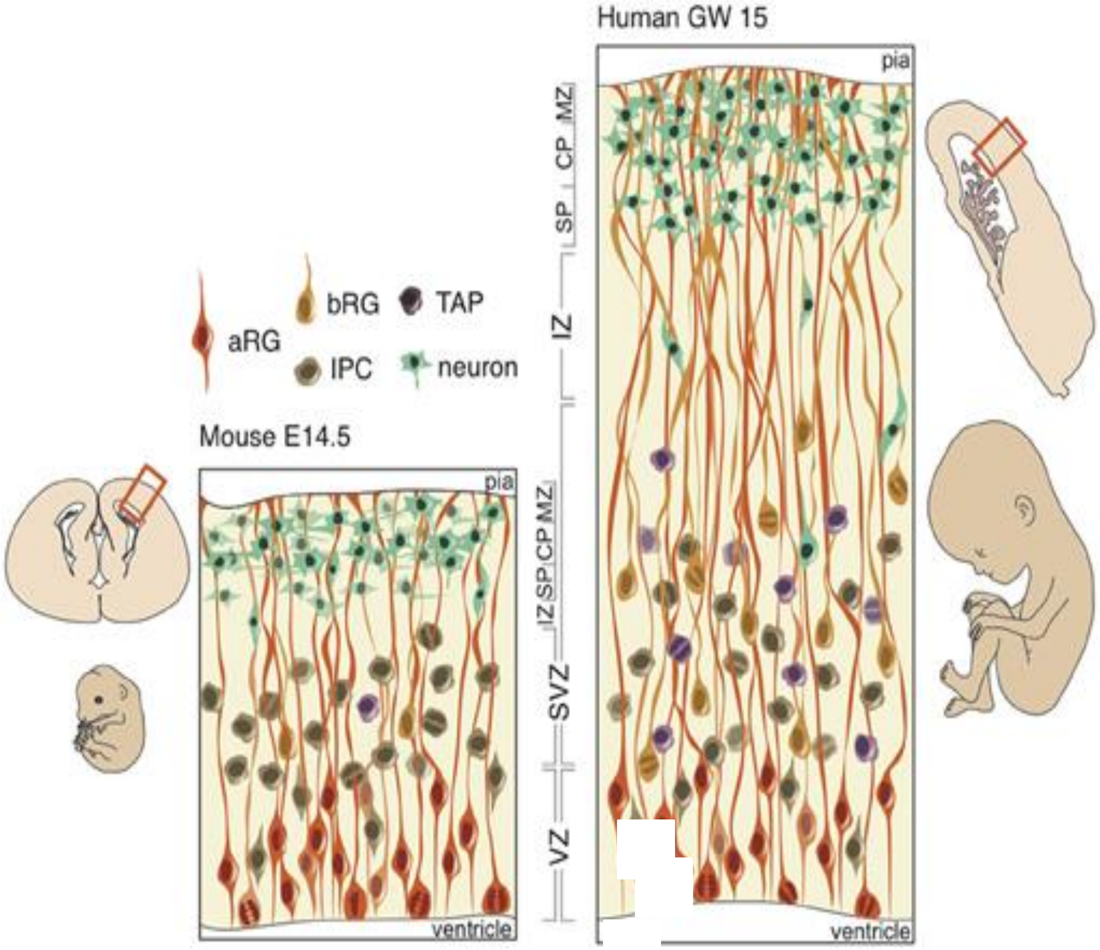




Homo habilis

Example 2: Human-specific Duplication of *ARHGAP11B*

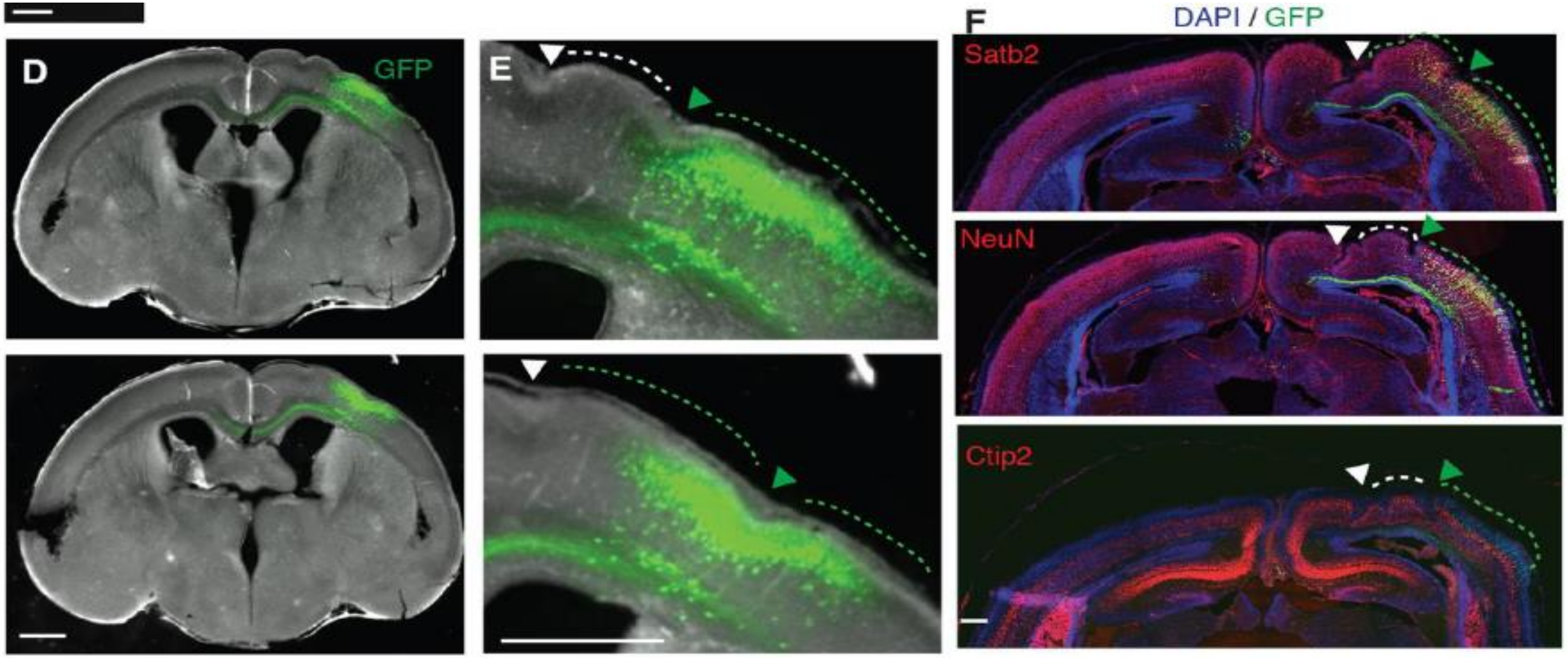
- A human-specific duplicated Rho GTPase activating protein that is truncated (5.3 mya)
- Predisposes to the most common cause of epilepsy
- Increase in number of basal radial glial hypothesized to lead to enlargement of the subventricular zone in humans.
- *ARHGAP11B* is expressed specifically in basal radial glial cells



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

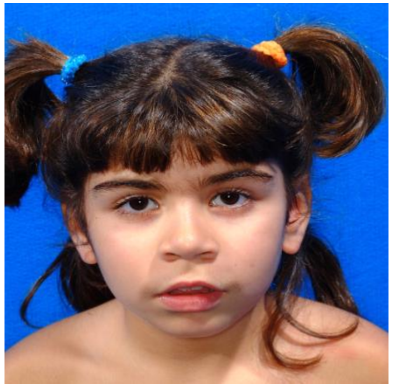
ARHGAP11B induced gyrification of mouse brain

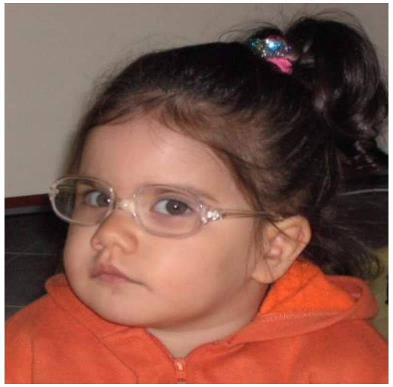
- E13.5 microinjection of *ARHGAP11B* induced folding in the neocortex by E18.5 in ½ of the cases– a significant increase in cortical area.

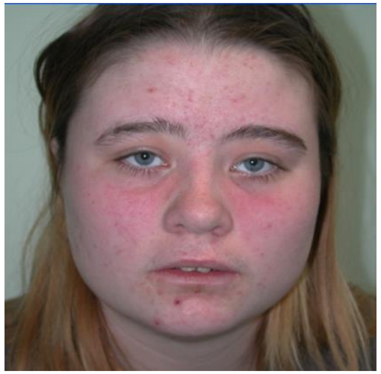


Florea *et al.*, *Science* 2015

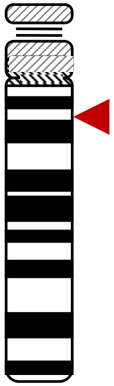
Duplication of *ARHGAP11B* and 15q13.3 Syndrome

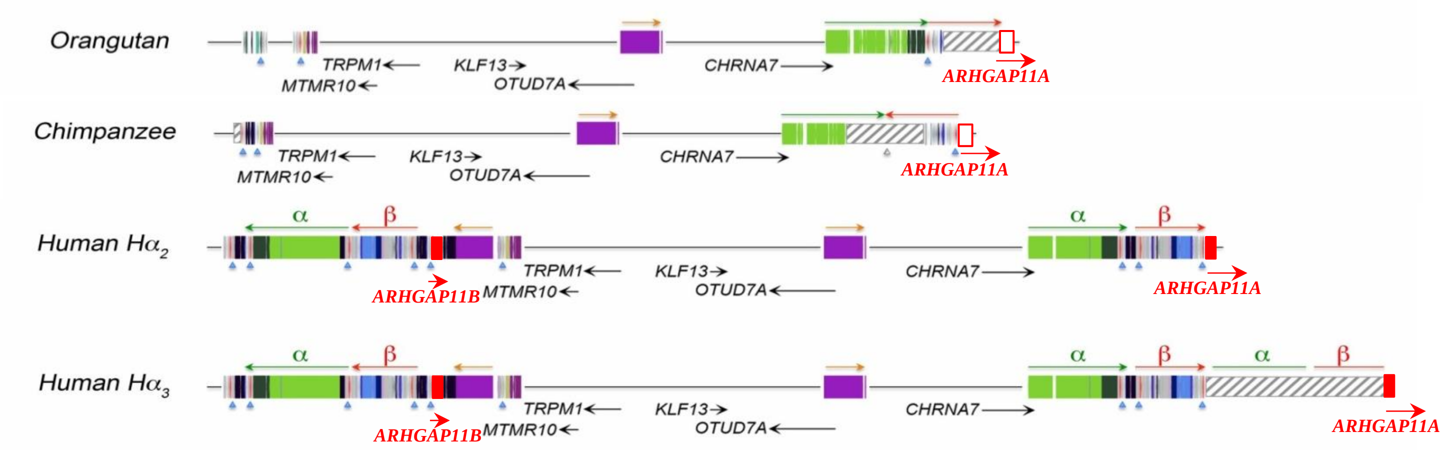


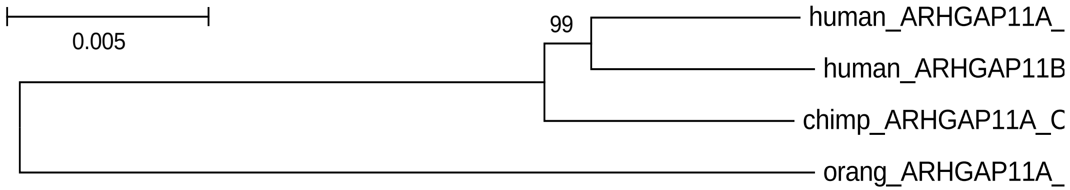




Chromosome 15





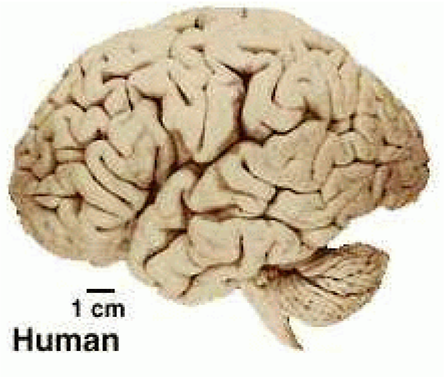


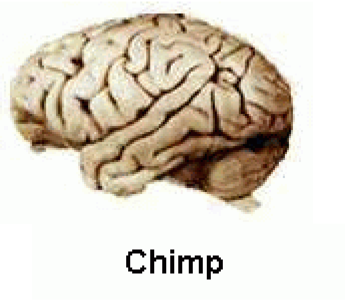
Duplication from*ARHGAP11A* to *ARHGAP11B* estimated to have occurred 5.3 +/- 0.5 million years ago.

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

Human-Specific Gene Innovations and Duplications

- *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—a genomic bias in location and gene type
- **II. Methods**: NGS Read-pair and read-depth methods to characterize SVs within genomes—long-read genomes that fully phase and assemble promise comprehensive characteriztion
- **III**: **Evolution**: Rapid evolution of complex human architecture that predisposes to disease coupled to gene innovation

Disease

Evolution

Eichler Lab

http://eichlerlab.gs.washington.edu/ genguest

Glossary

SV-structural variation

CNV- copy number variation

CNP—copy number polymorphism

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

Indel-insertion/deletion event

SD—segmental duplication

SUN-singly-unique nucleotide identifier

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

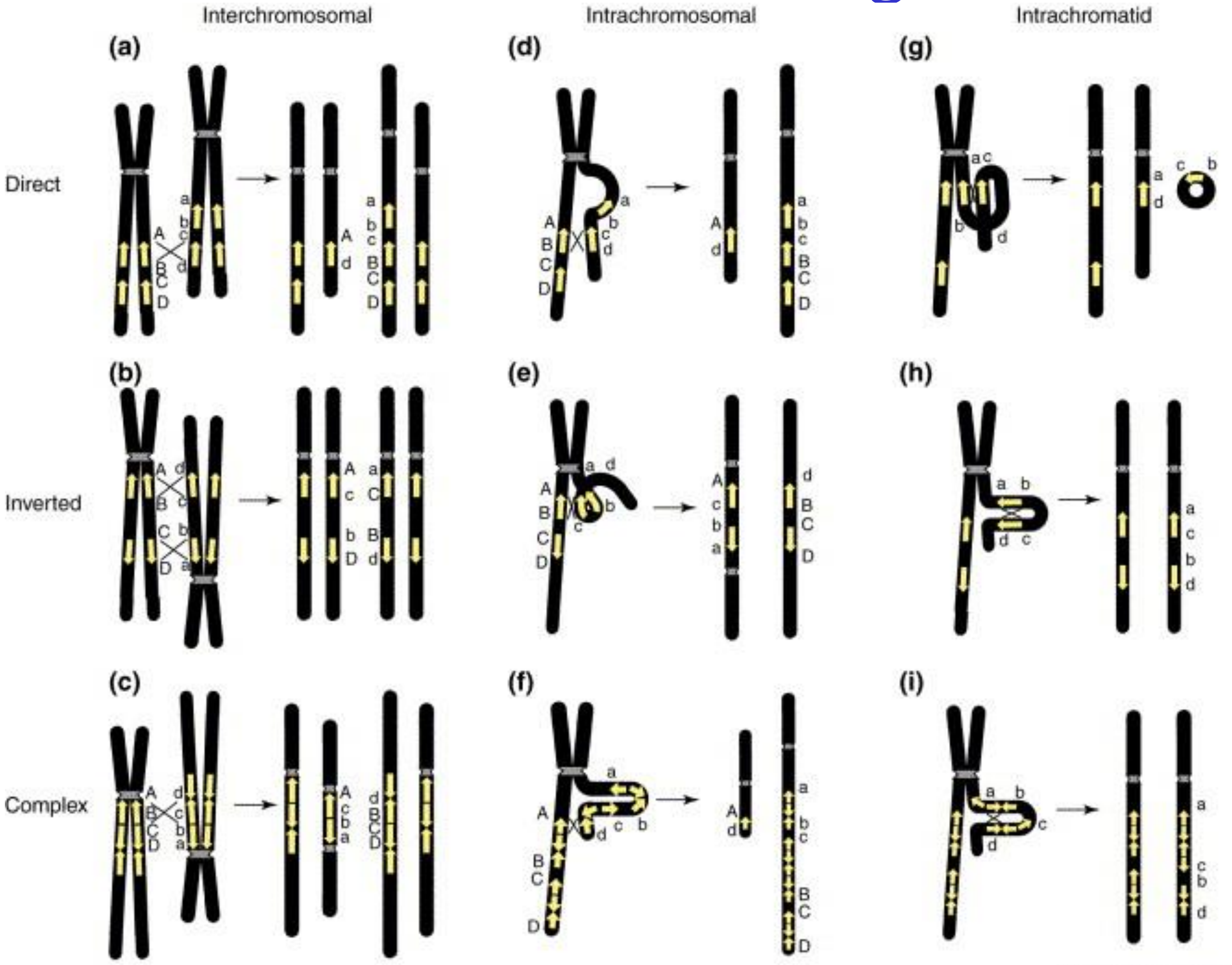
PacBio—Pacific Biosciences

ZMW-zero-mode wave guide

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 to increase sensitivity and specificity
- *Lumpy* --Quinlan/Hall—uses probabilistic framework to integrate multiple structural variation signals such as discordant paired-end alignments and split-read alignments
- Conifer and XHMM— Krumm/Eichler & Frommer/Purcellcalling CNVs from exomes
- *SMRT-SV2 & Phased-SV*—Chaisson/Eichler—maps SMRT long reads (BLASR/minimap) to reference, detects signatures of SV and generates local assembly
- *PBSV*—Aaron Wenger (PacificBiosciences software) signatures from pbmm2 alignments
- *SNIFFLES*—Sedlacek/Schatz– NGLMR mapping of PacBio or ONT data using split-read alignments, high-mismatch regions, and coverage analysis

SD-Mediated Rearrangements



TRENDS in Genetics