Genome Structural Variation

Evan Eichler
Howard Hughes Medical Institute
University of Washington

January 11th, 2016, Genomics Workshop, Český Krumlov

Genetic Variation

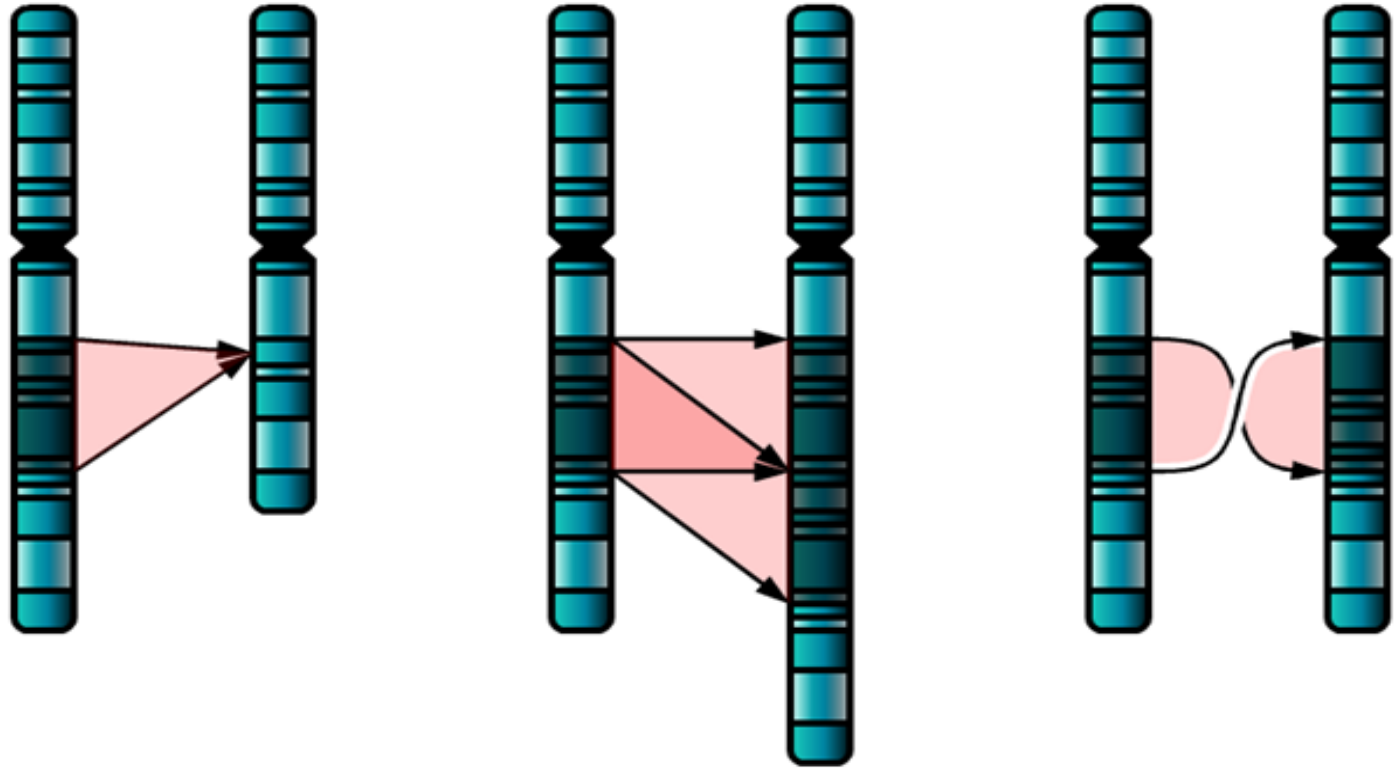
Types

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** includes copy-number variation (CNV) and balanced events such as inversions and translocations—originally defined as > 1 kbp but now >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

Rare chromosomal deletions and duplications increase risk of schizophrenia

The International Schizophrenia Consortium*

Nature 455:237-41 2008

Large recurrent microdeletions associated with schizophrenia

Nature 455:232-6 2008

Hreinn Stefansson[1]*, Dan Rujescu[2]*, Sven Cichon[3,4]*, Olli P. H. Pietiläinen[5], Andres Ingason[1], Stacy Steinberg[1], Ragnheidur Fossdal[1], Engilbert Sigurdsson[6], Thordur Sigmundsson[6], Jacobine E. Buizer-Voskamp[7]

Strong Association of De Novo Copy Number Mutations with Autism

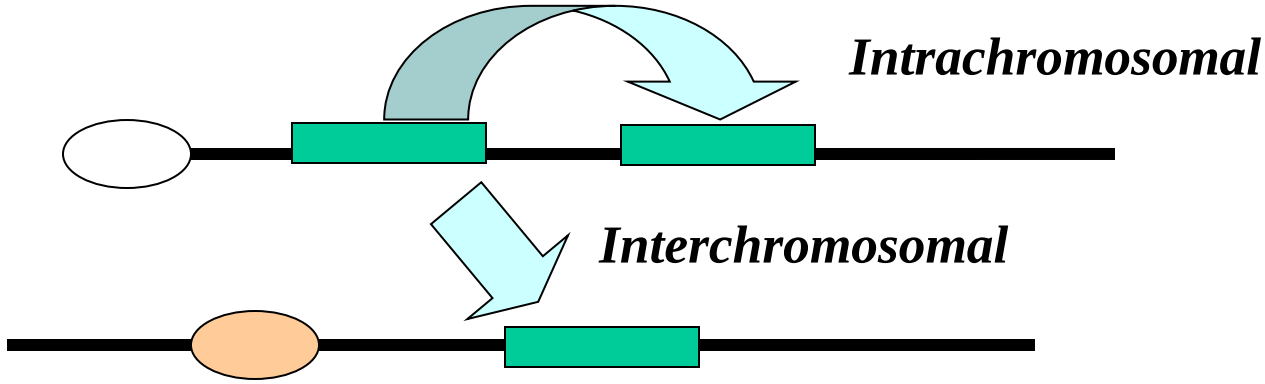
NCE

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]†

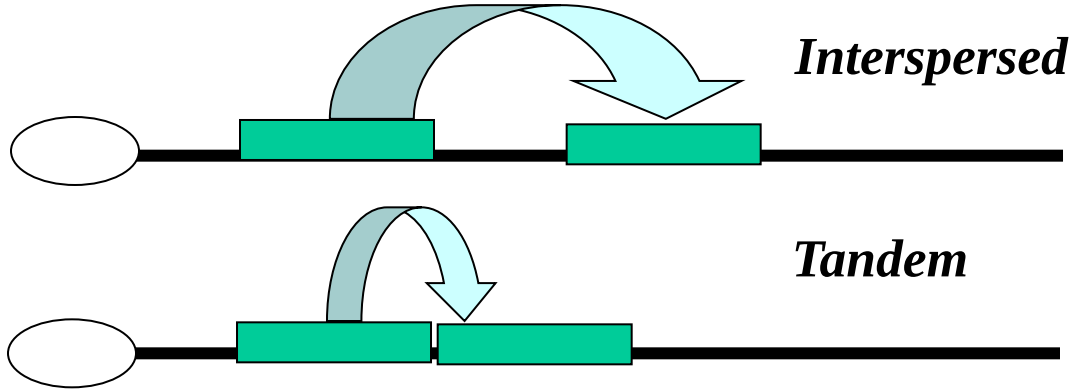
SCIENCE VOL 316 20 APRIL 2007

Perspective: Segmental Duplications (SD)

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

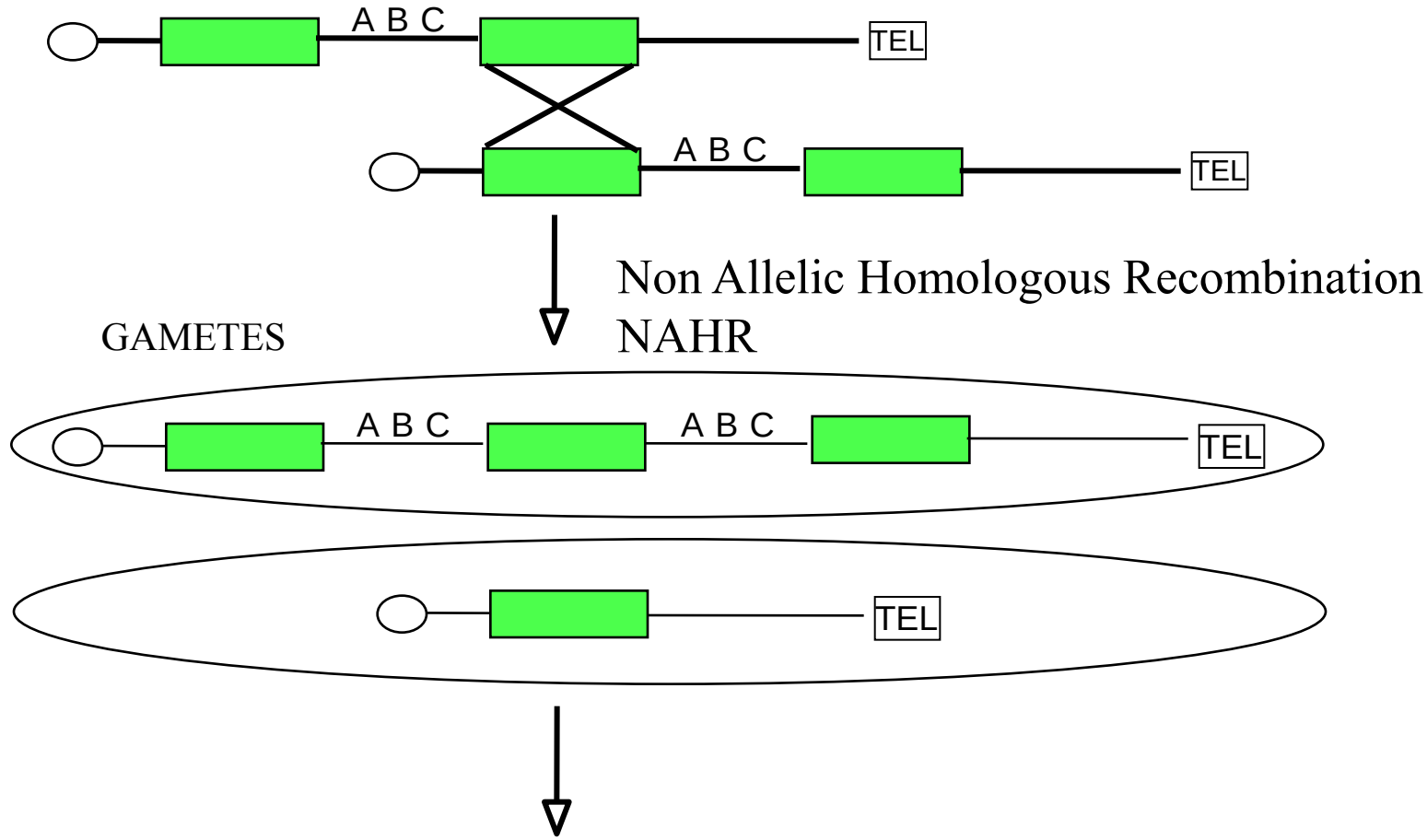


Distribution



Configuration

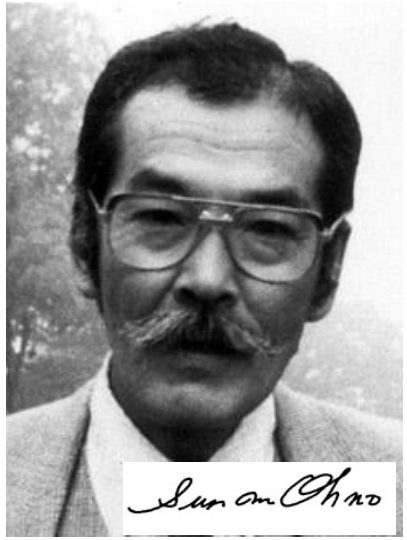
Importance:
SDs promote Structural Variation



Human Disease

Triplosensitive, Haploinsufficient and Imprinted Genes

Importance: Evolution of New Gene Function



GeneA

Duplication

GeneA'

Acquire New/
Modified Function

Mutation

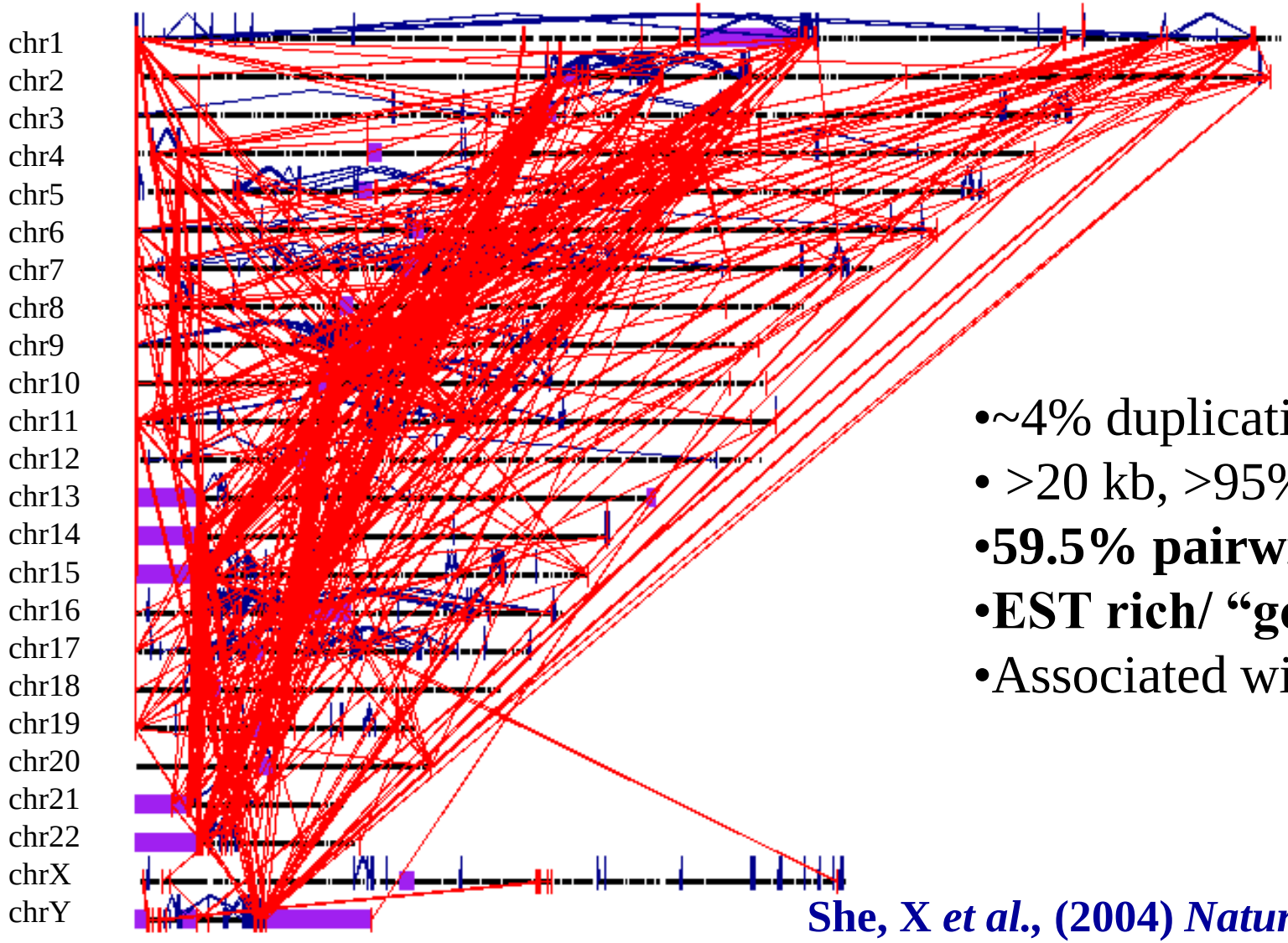
Mutation

Maintain old
Function

Mutation

Loss of Function

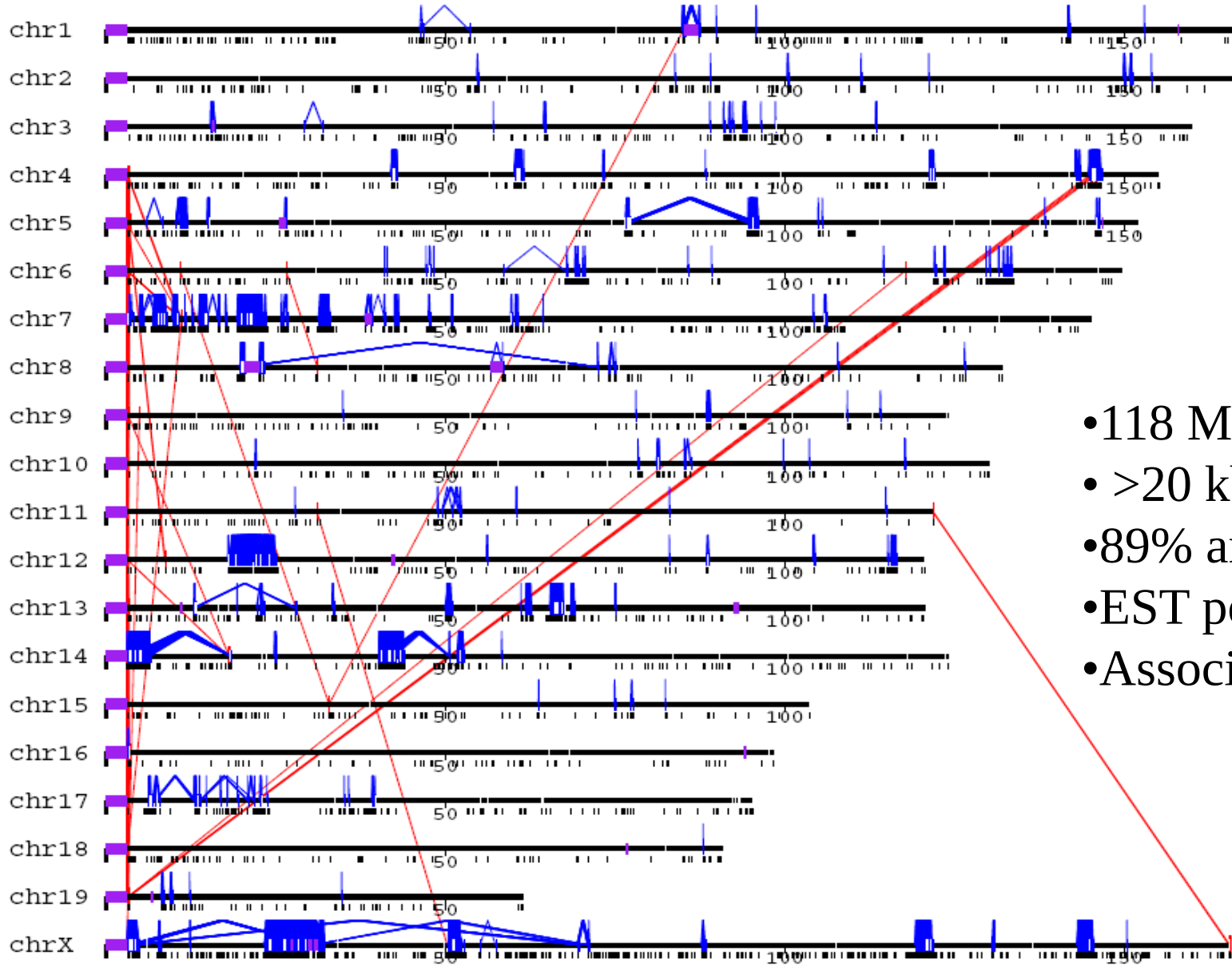
I. Human Genome Segmental Duplication Pattern



- ~4% duplication (125 Mb)
- >20 kb, >95%
- **59.5% pairwise (> 1 Mb)**
- **EST rich/ "gene" rich**
- Associated with Alu repeats

She, X *et al.*, (2004) *Nature* 431:927-30
http://humanparalogy.gs.washington.edu

Mouse Segmental Duplication Pattern



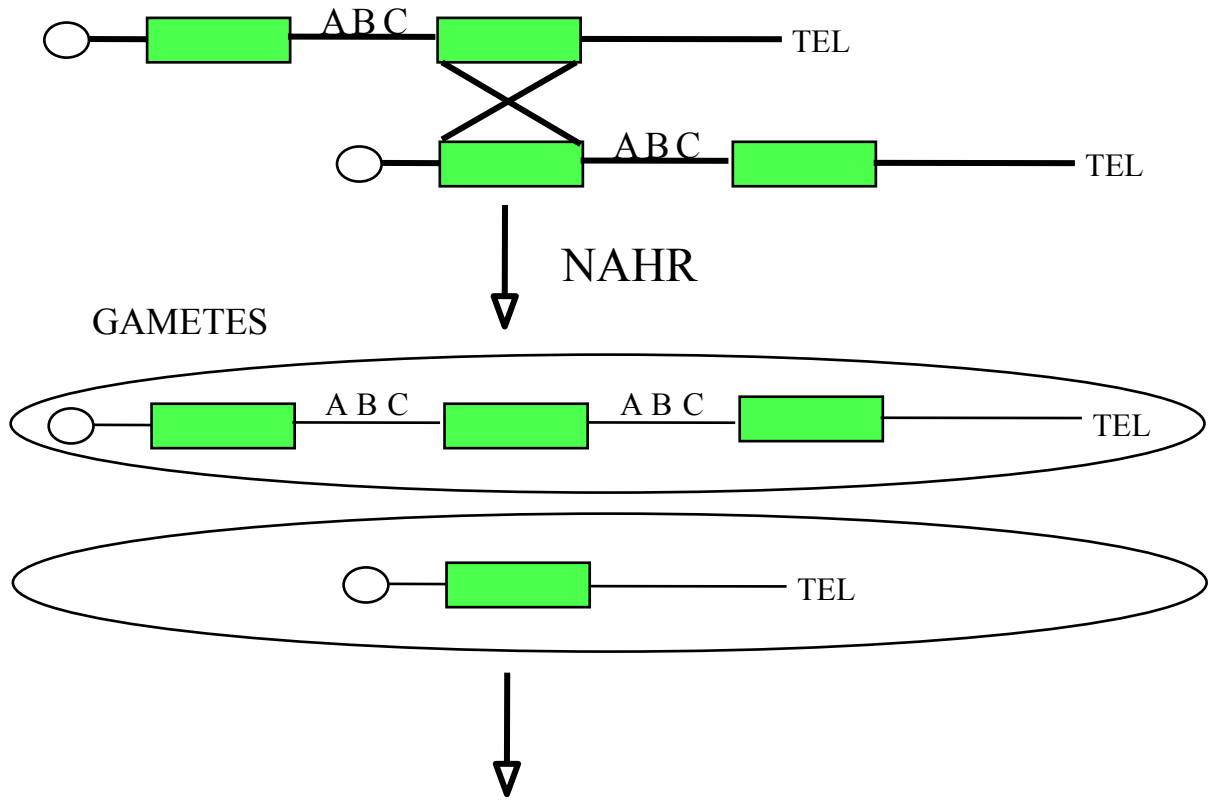
- 118 Mb or ~4% dup
- >20 kb, >95%
- 89% are tandem
- EST poor
- Associated with LINEs

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

Human Segmental Duplications Properties

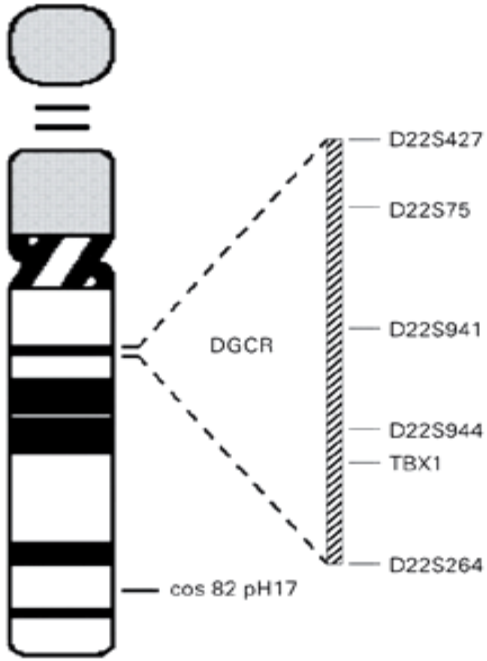
- Large (>10 kb)
- Recent (>95% identity)
- **Interspersed (60% are separated by more than 1 Mb)**
- Modular in organization
- Difficult to resolve

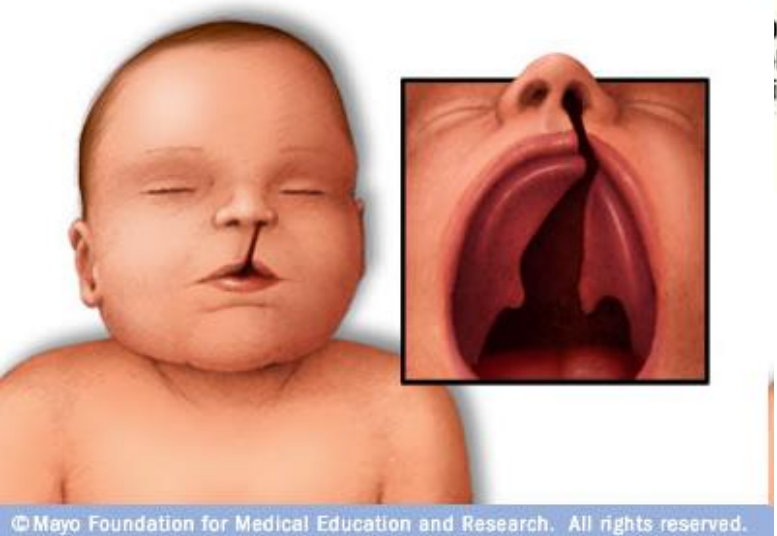
Rare Structural Variation & Disease

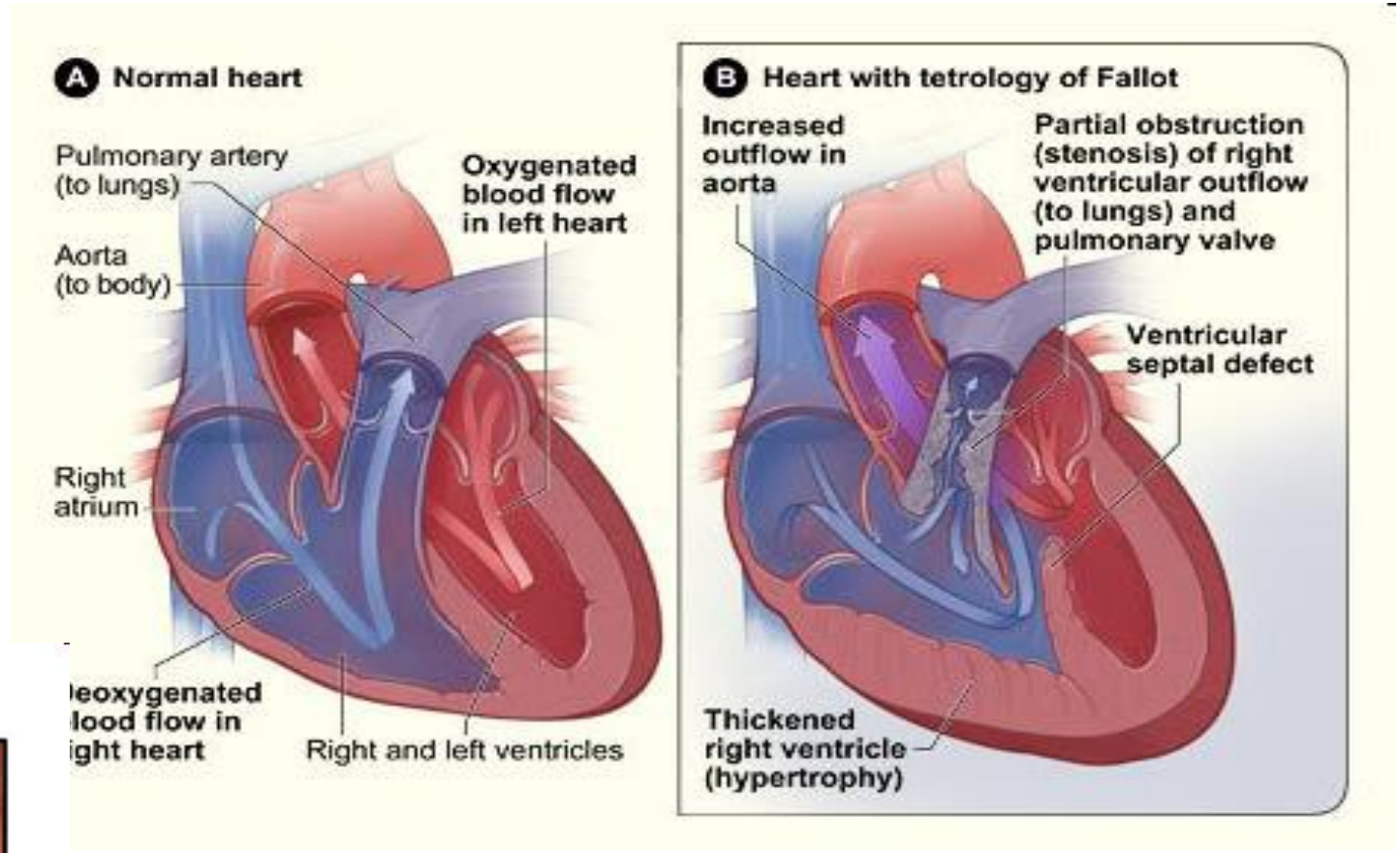


•**Genomic Disorders:** A group of diseases that results from genome rearrangement mediated mostly by non-allelic homologous recombination**. (***Inoue & Lupski , 2002***).**

DiGeorge/VCFS/22q11 Syndrome



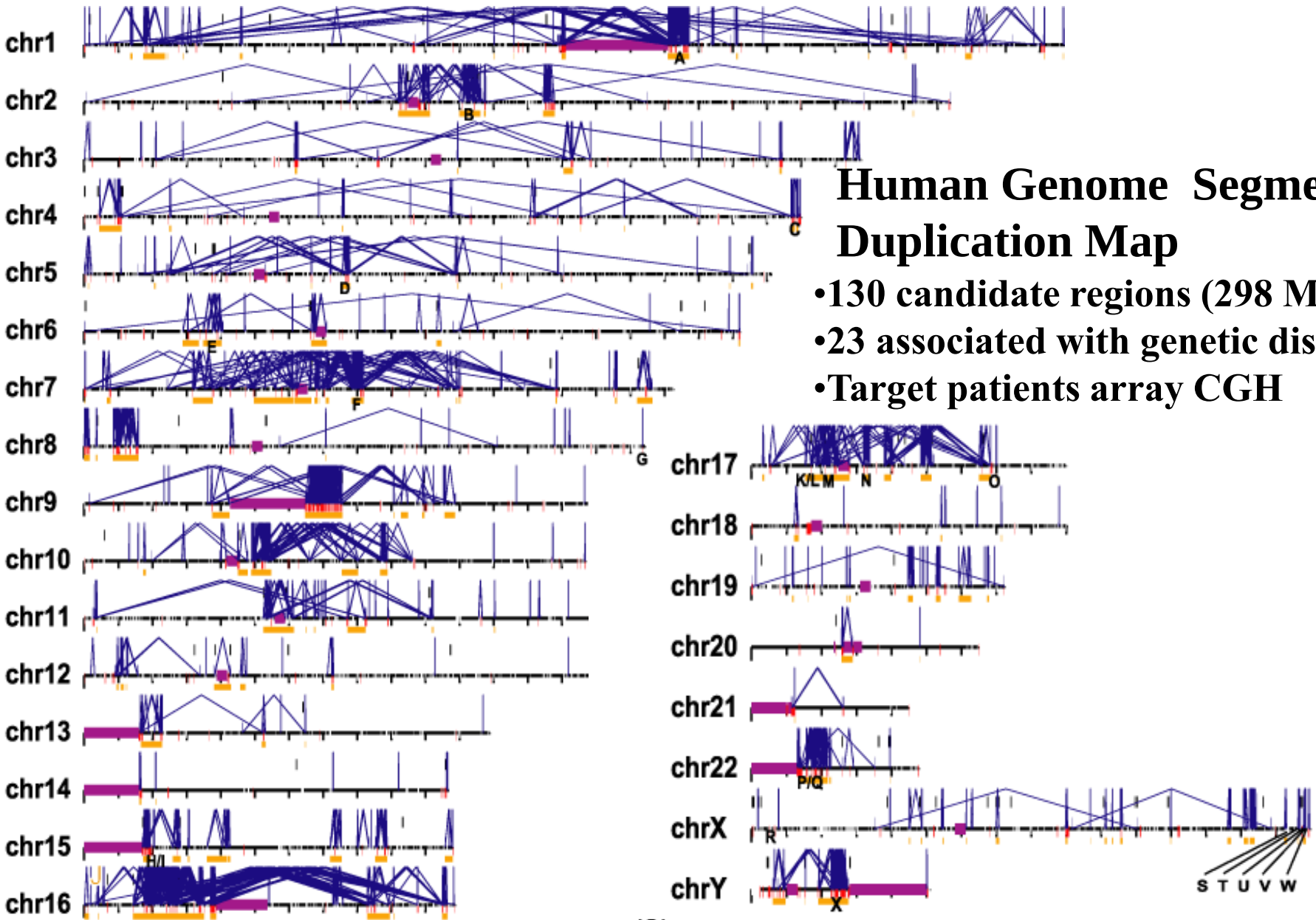




1/2000 live births

180 phenotypes

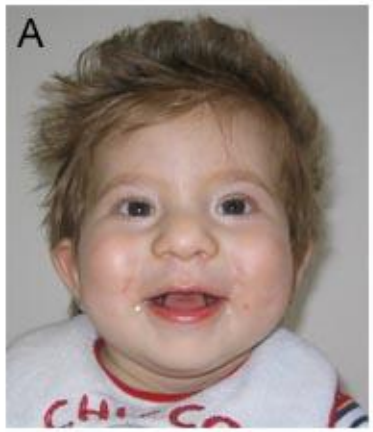
75-80% are sporadic (not inherited)

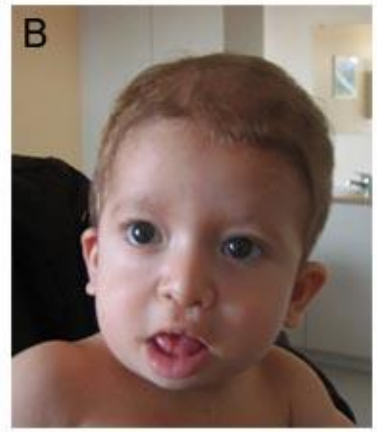


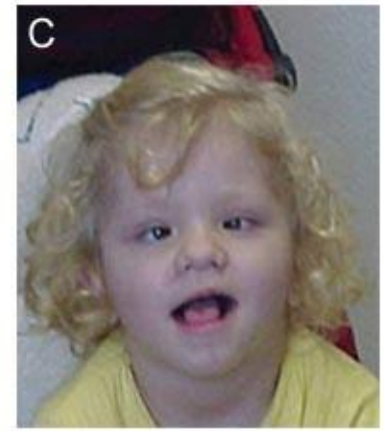
Human Genome Segmental Duplication Map

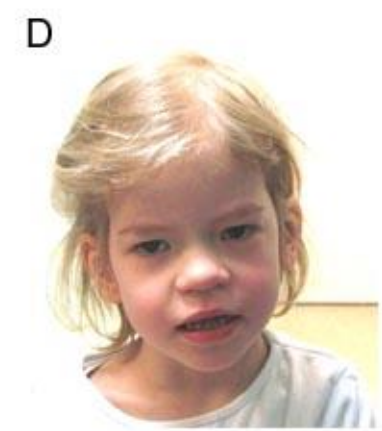
- 130 candidate regions (298 Mb)
- 23 associated with genetic disease
- Target patients array CGH

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

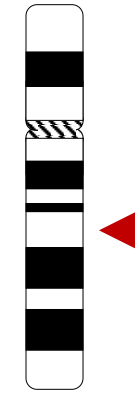


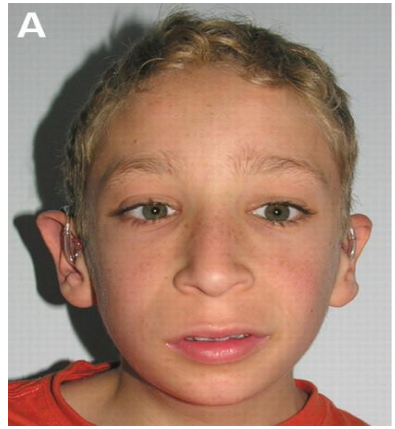


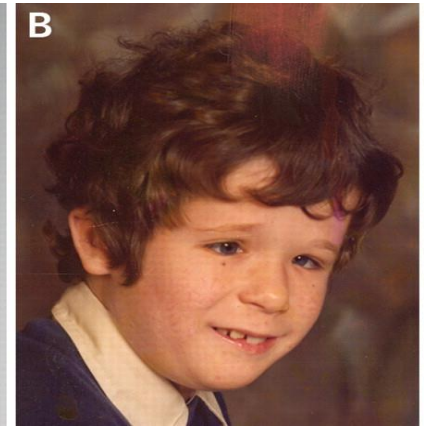


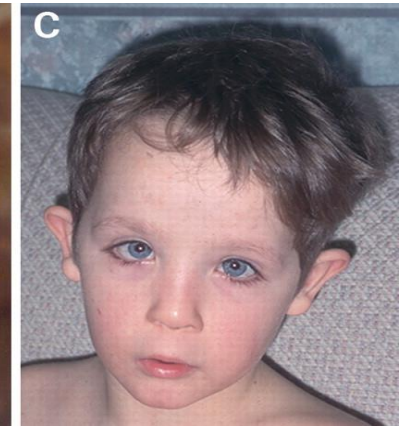


Chromosome 17

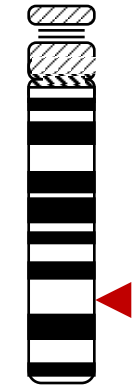


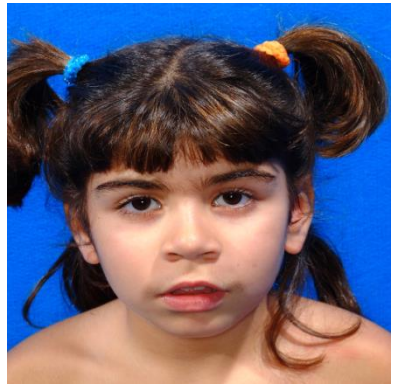


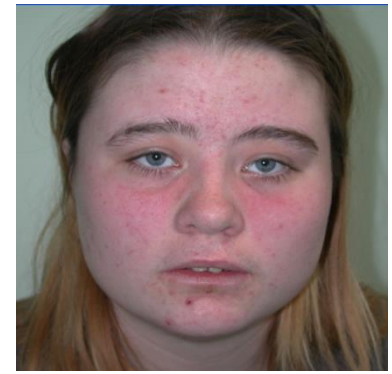




Chromosome 15



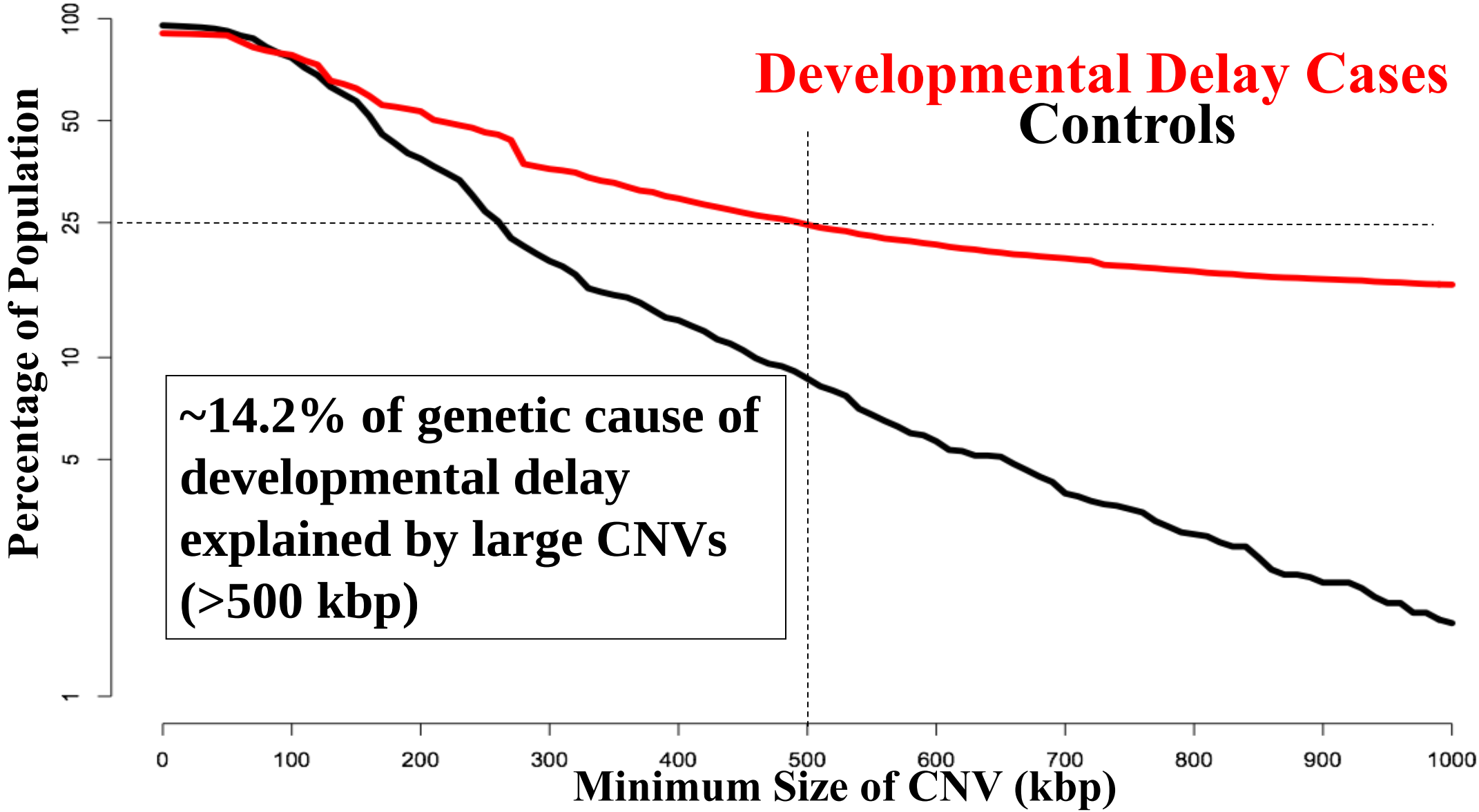




Chromosome 15

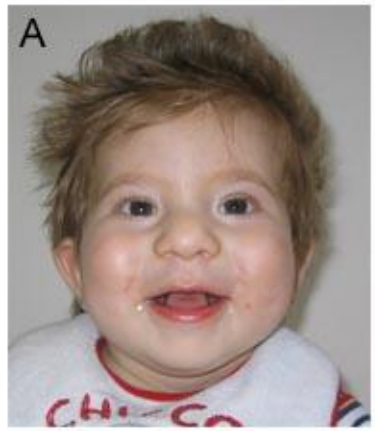
Genome Wide CNV Burden

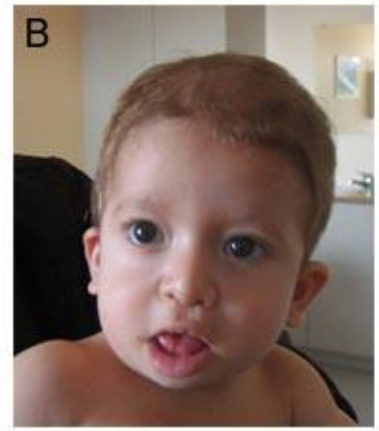
(15,767 cases of ID,DD,MCA vs. 8,328 controls)

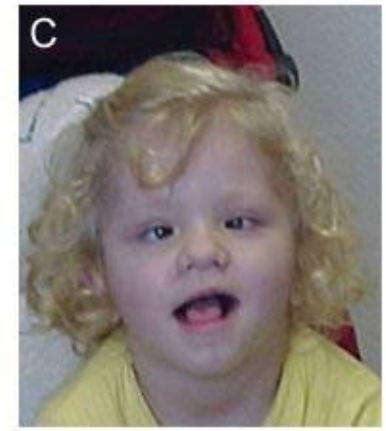


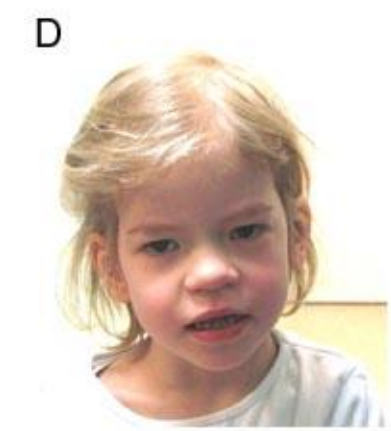
Cooper et al., *Nat. Genet*, 2011

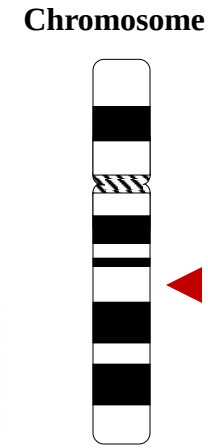
Common and Rare Structural Variation are Linked 17q21.31 Deletion Syndrome

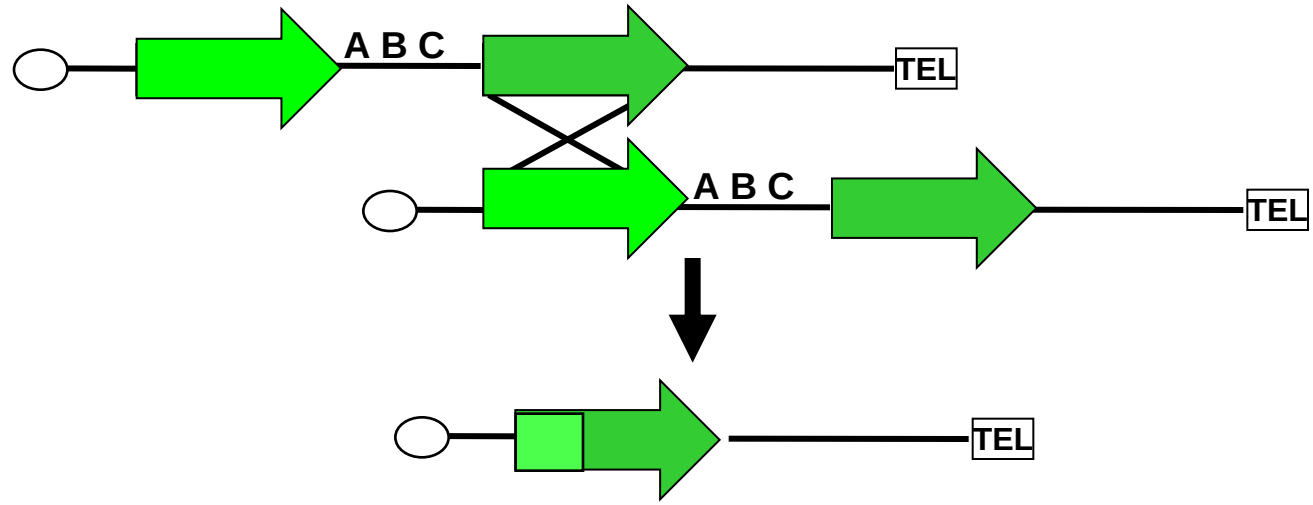




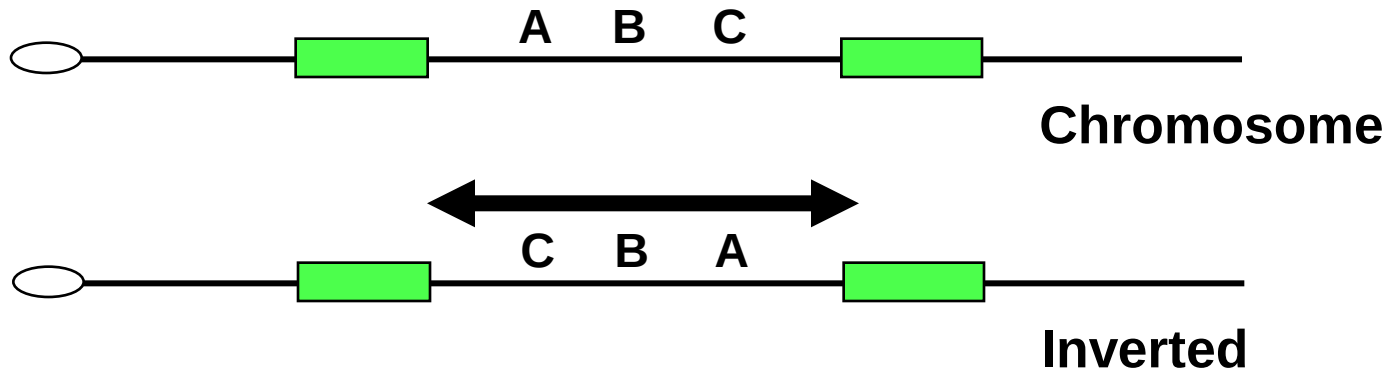








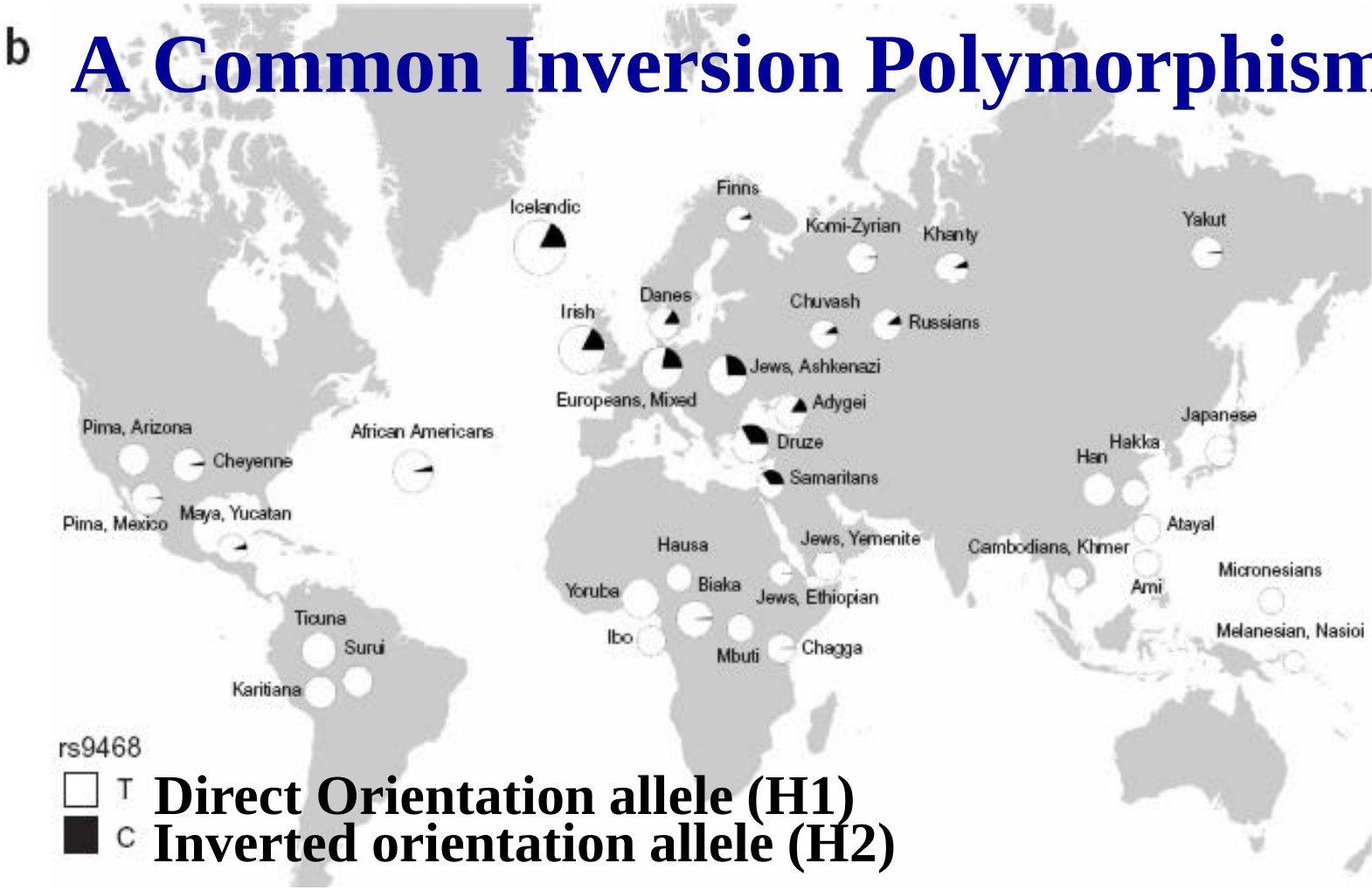
17q21.31 Inversion



- 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**

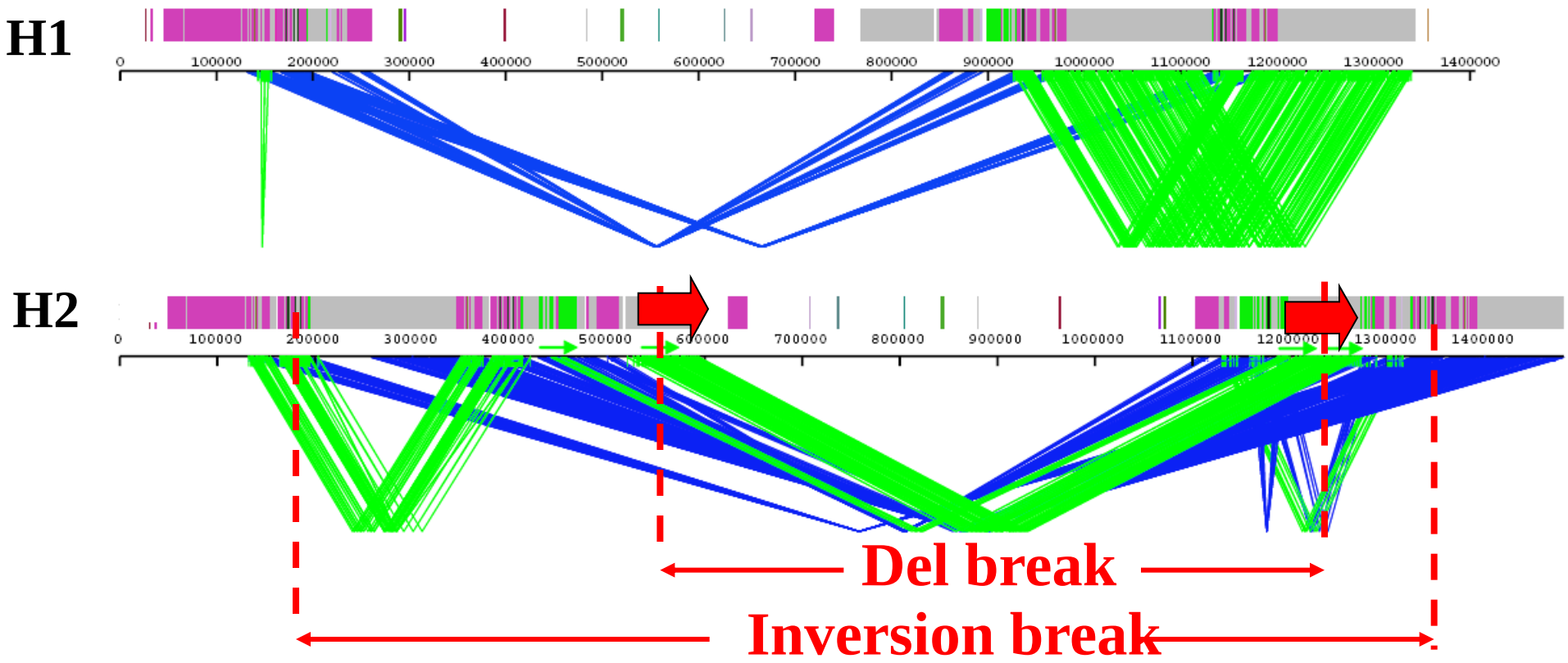
Stefansson, K *et al.*, (2005) *Nature Genetics*

A Common Inversion Polymorphism



- **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**

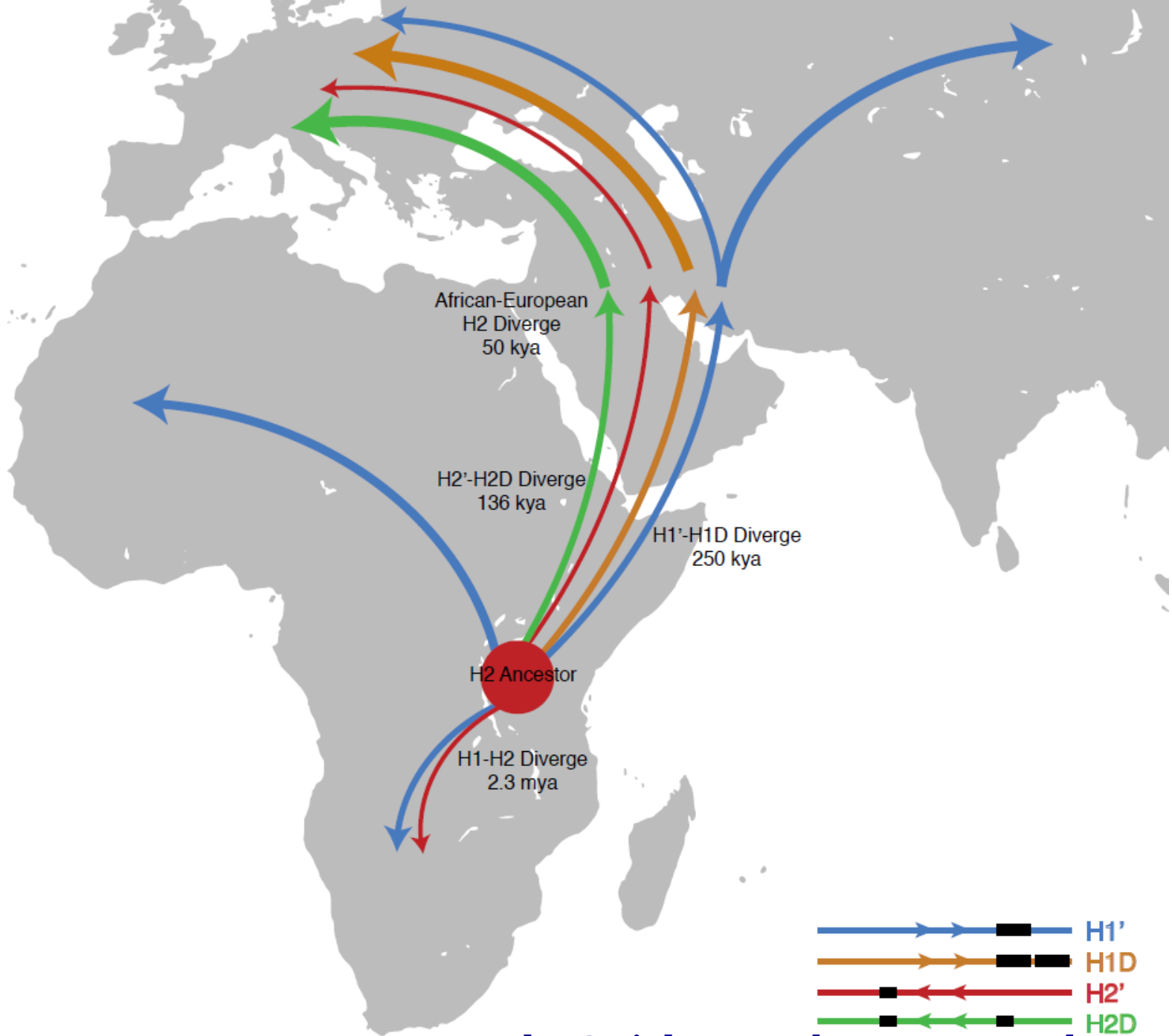
Duplication Architecture of 17q21.31 Inversion (H2) vs. Direct (H1) Haplotype



- 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

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

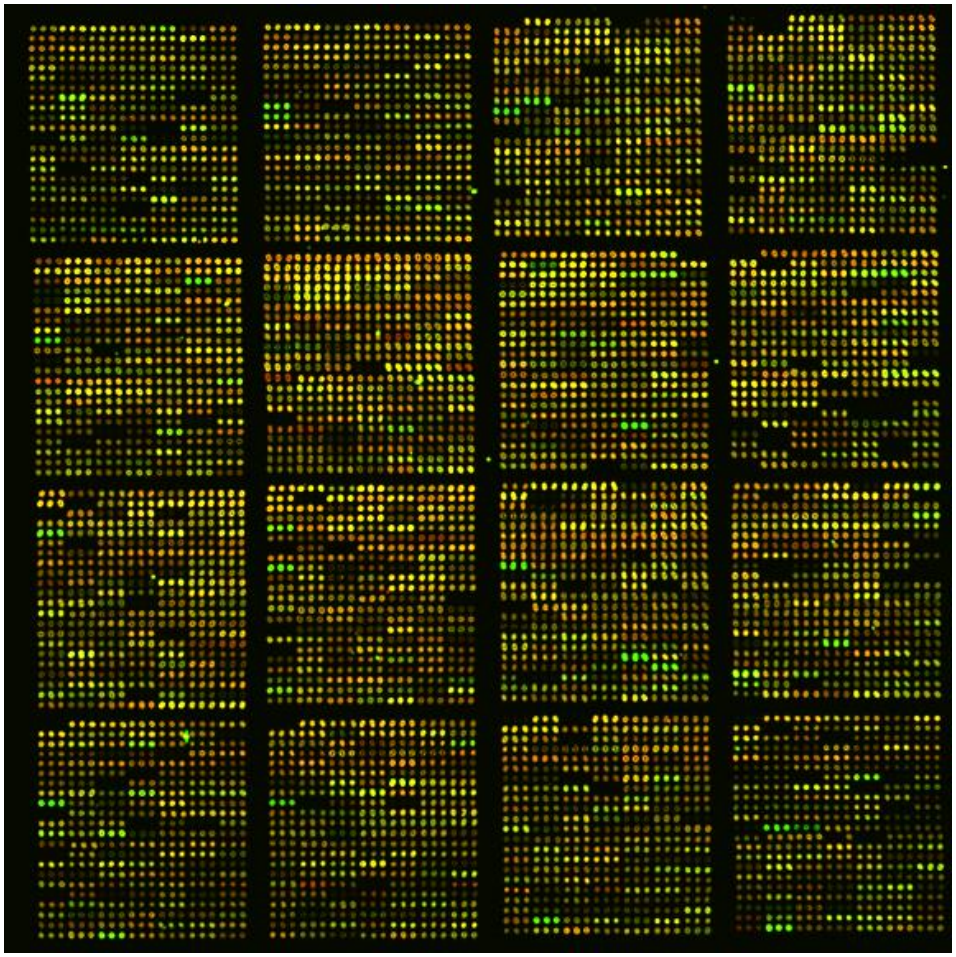
Single molecule mapping

- **Optical mapping:** Teague et al., 2010

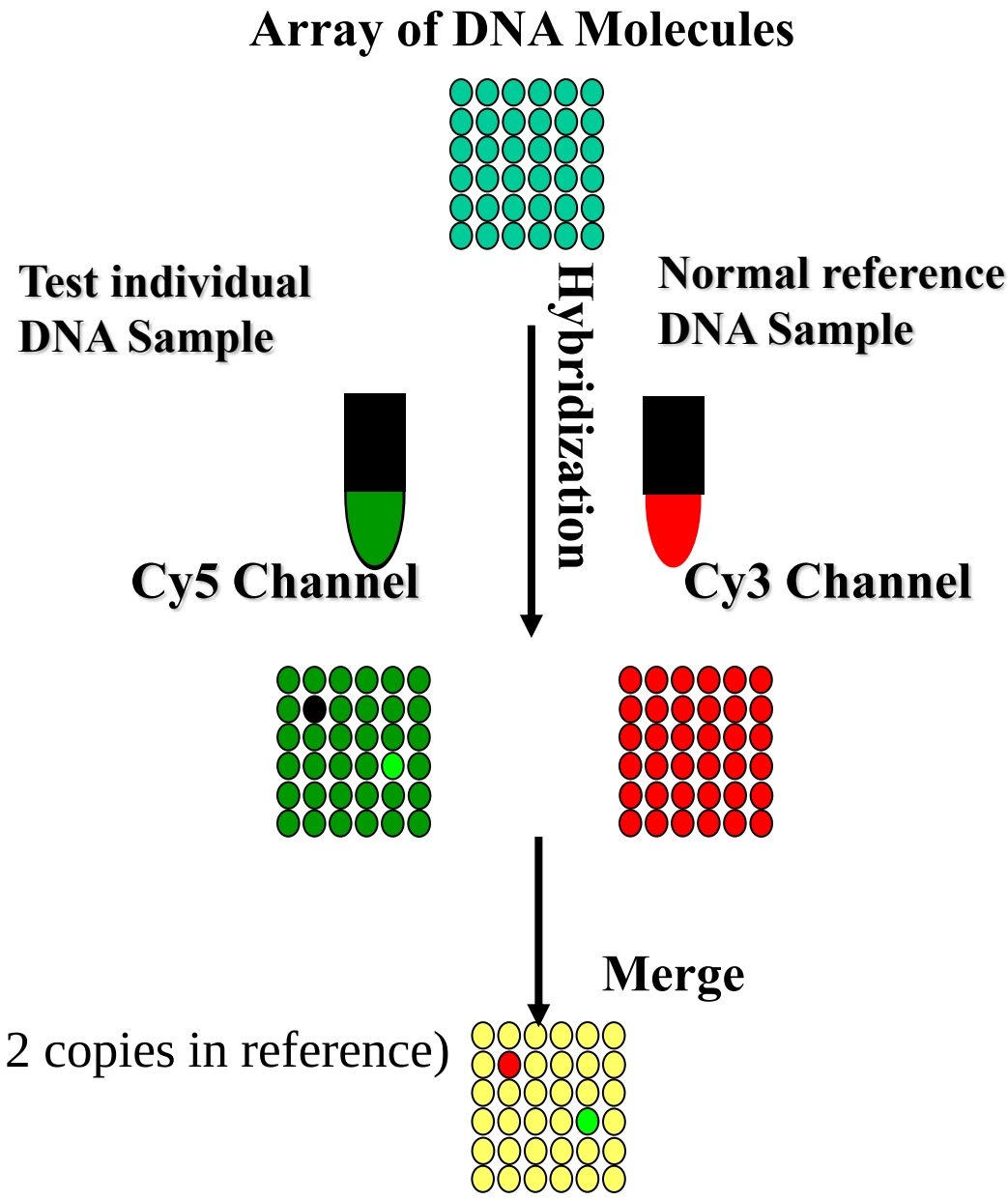
Sequencing-based

- Read-depth: Bailey et al, 2002
- Fosmid ESP: Tuzun *et al.* 2005, Kidd *et al.* 2008
- Sanger sequencing: Mills *et al.*, 2006
- Next-gen sequencing: Korbel *et al.* 2007, Yoon *et al.*, 2009, Alkan et al., 2009, Hormozdiari *et al.* 2009, Chen *et al.* 2009; Mills 1000 Genomes Project, Nature, 2011, Sudmant 2015
- 3rd generation --long-reads: Chaisson et al., 2015

Array Comparative Genomic Hybridization



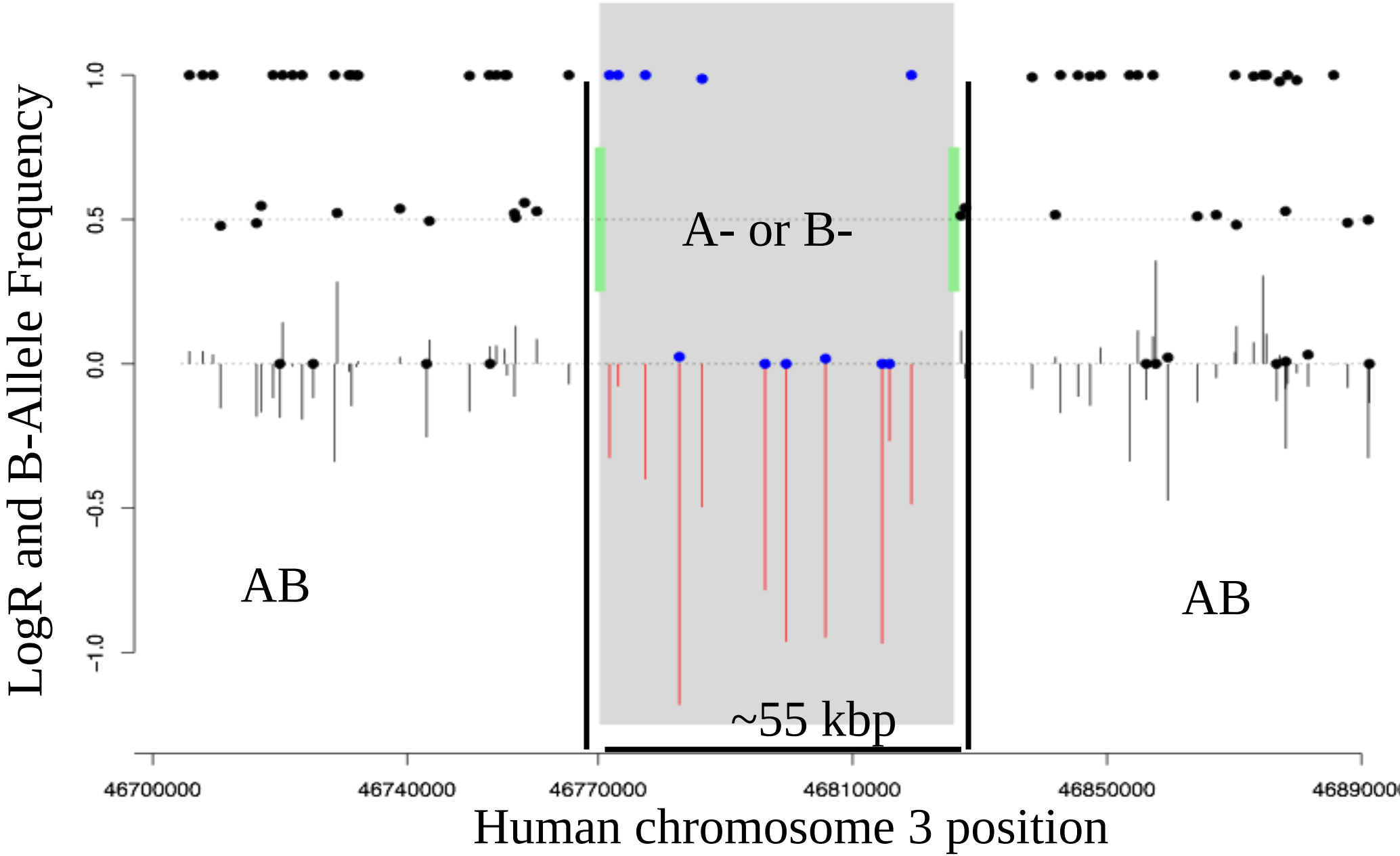
12 mm



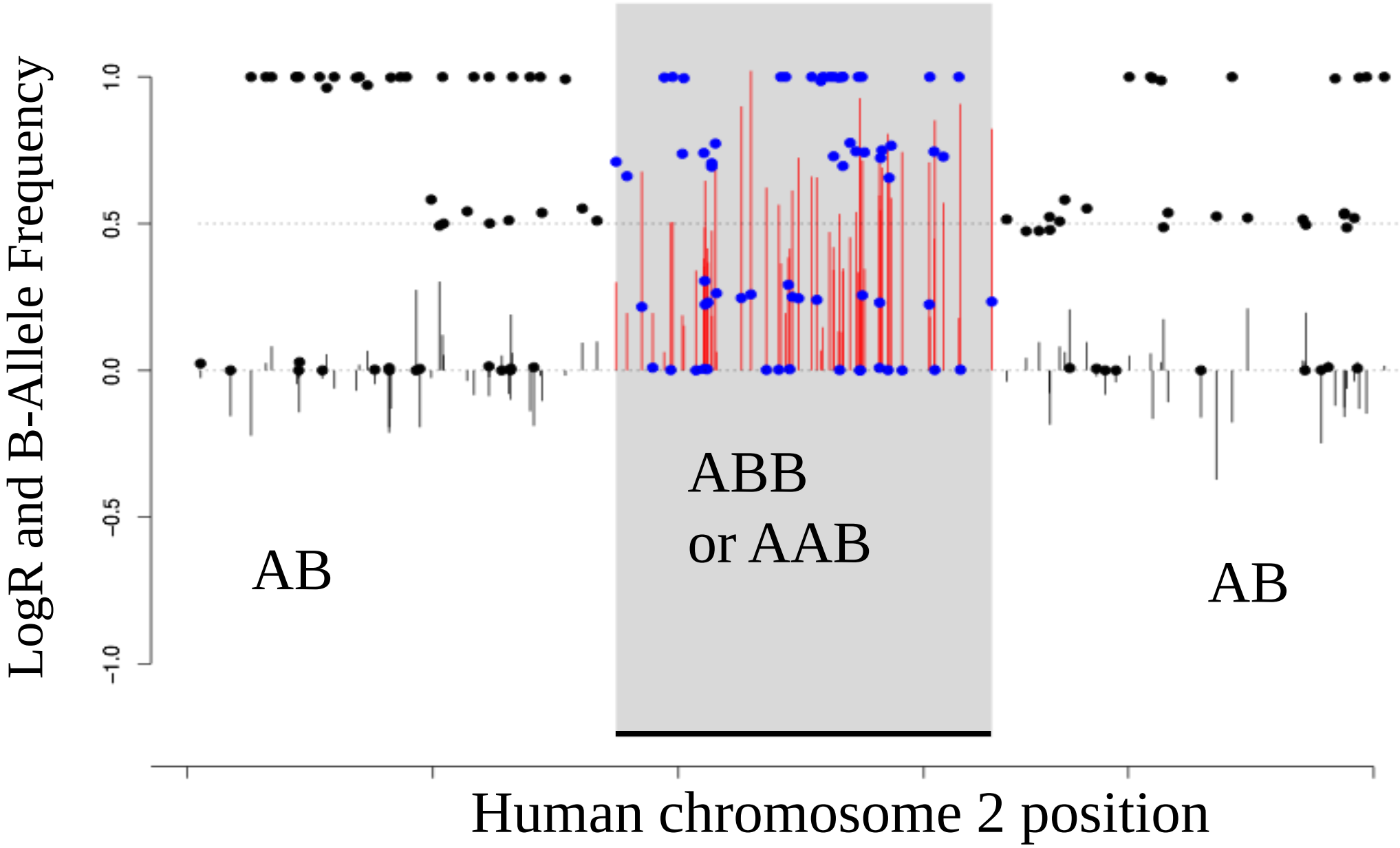
One copy gain = $\log_2(3/2) = 0.57$ (3 copies vs. 2 copies in reference)

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

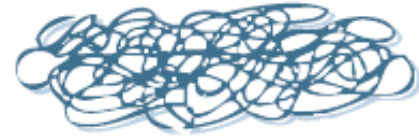
SNP Microarray detection of Deletion (Illumina)



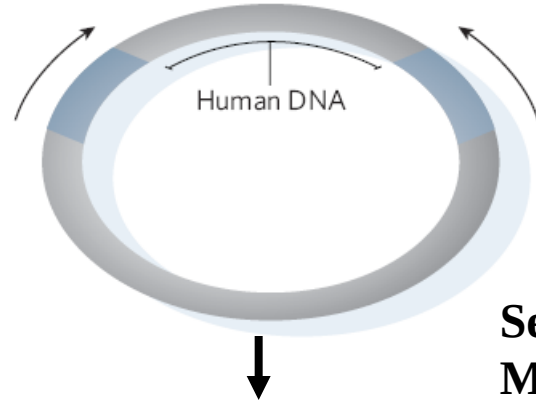
SNP Microarray detection of Duplication (Illumina)



Using Read Pairs to Resolve Structural Variation

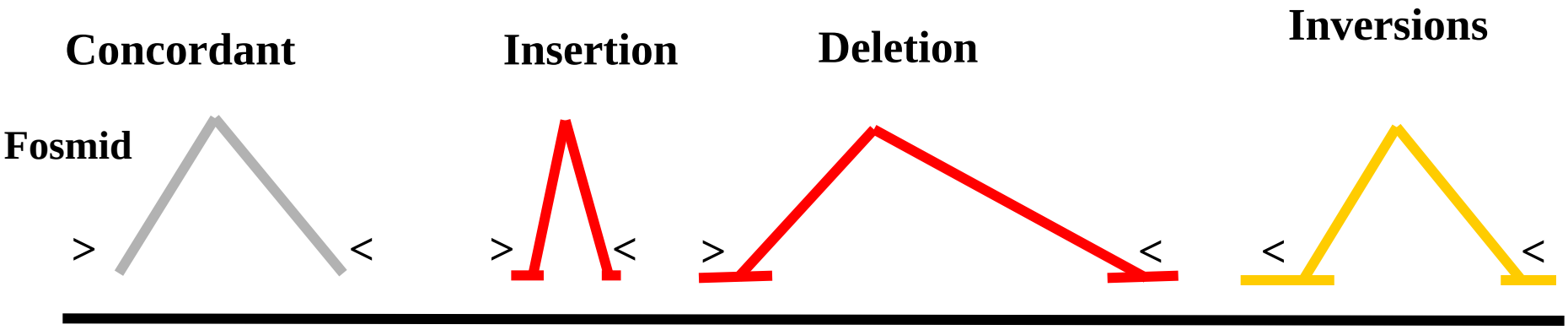


Human Genomic DNA



Genomic Library (1 million clones)

Sequence ends of genomic inserts &
Map to human genome



Build35

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

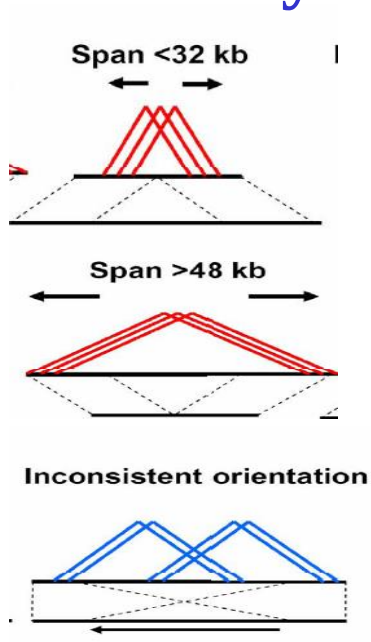
Genome-wide Detection of Structural Variation (>8kb) by End-Sequence Pairs

a)

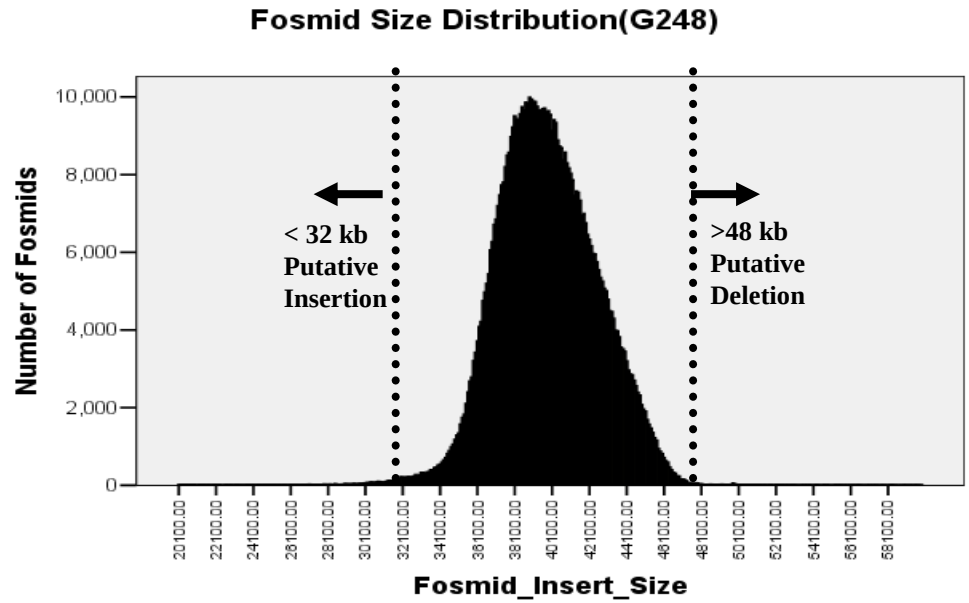
Insertion

Deletion

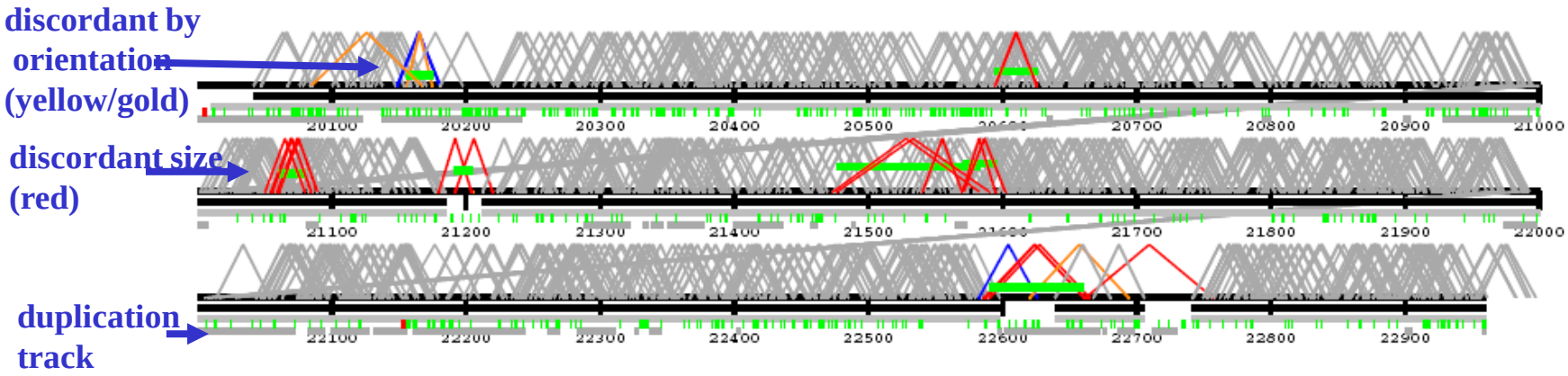
Inversion



b)

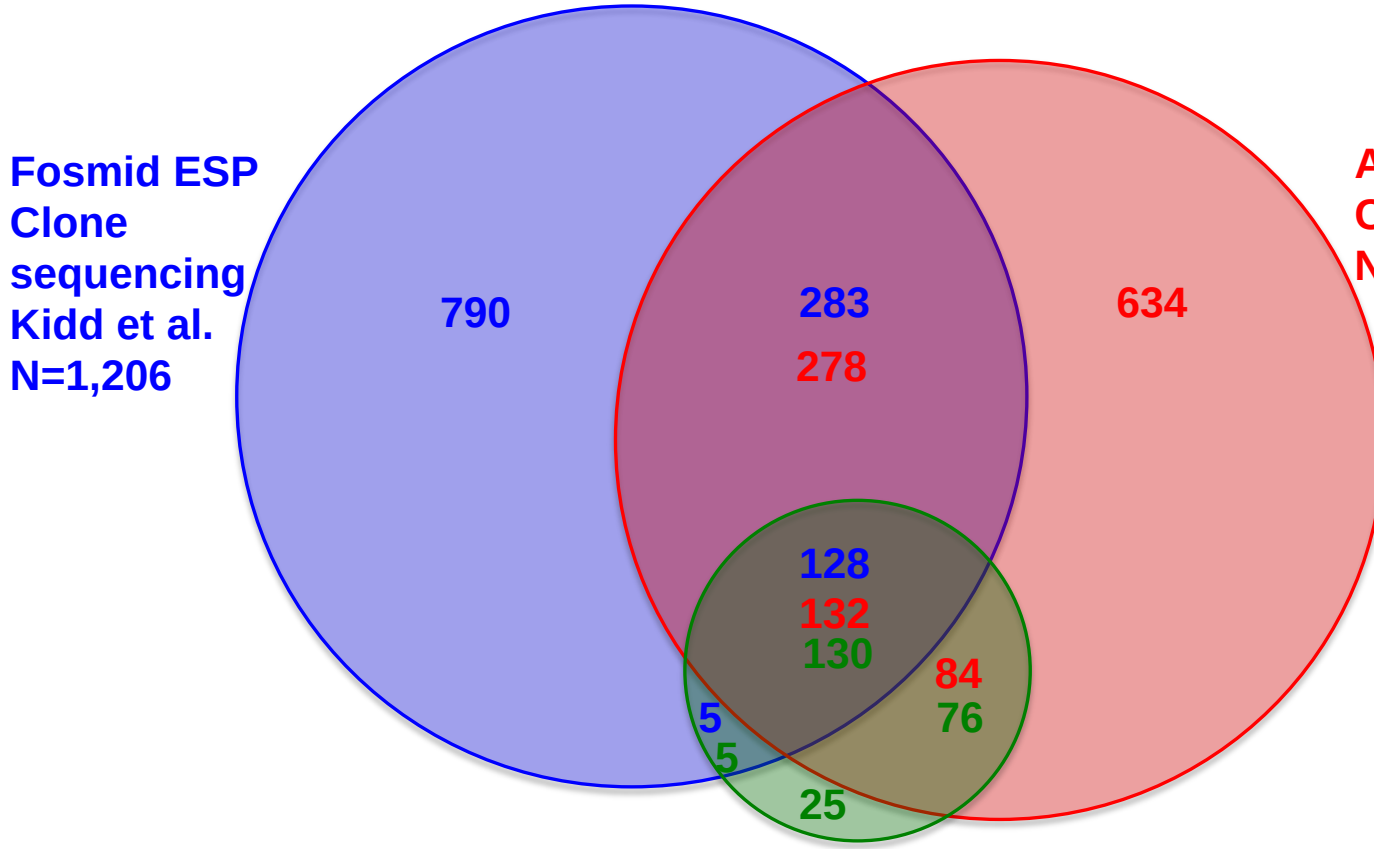


c)



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

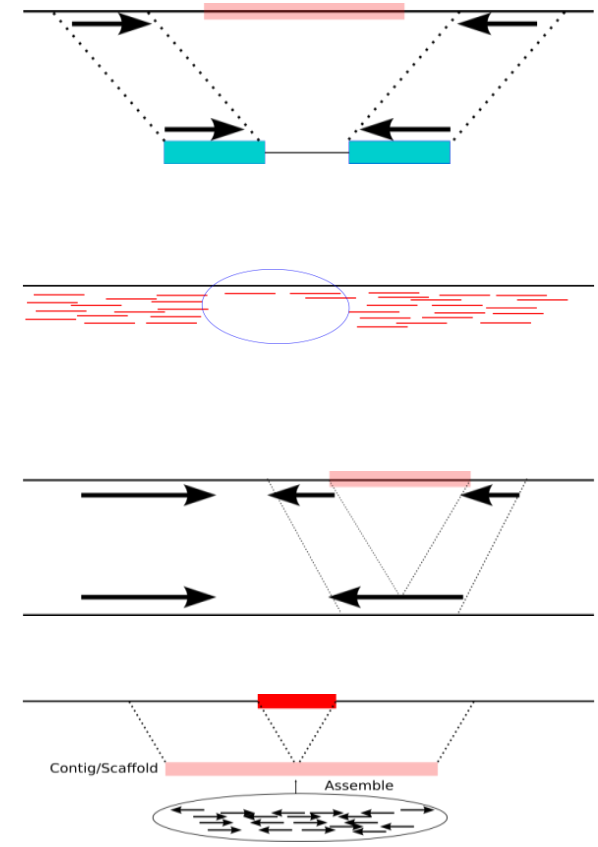
Experimental Approaches Incomplete
(Examined 5 identical genomes > 5kbp)



Kidd et al., *Cell* 2010

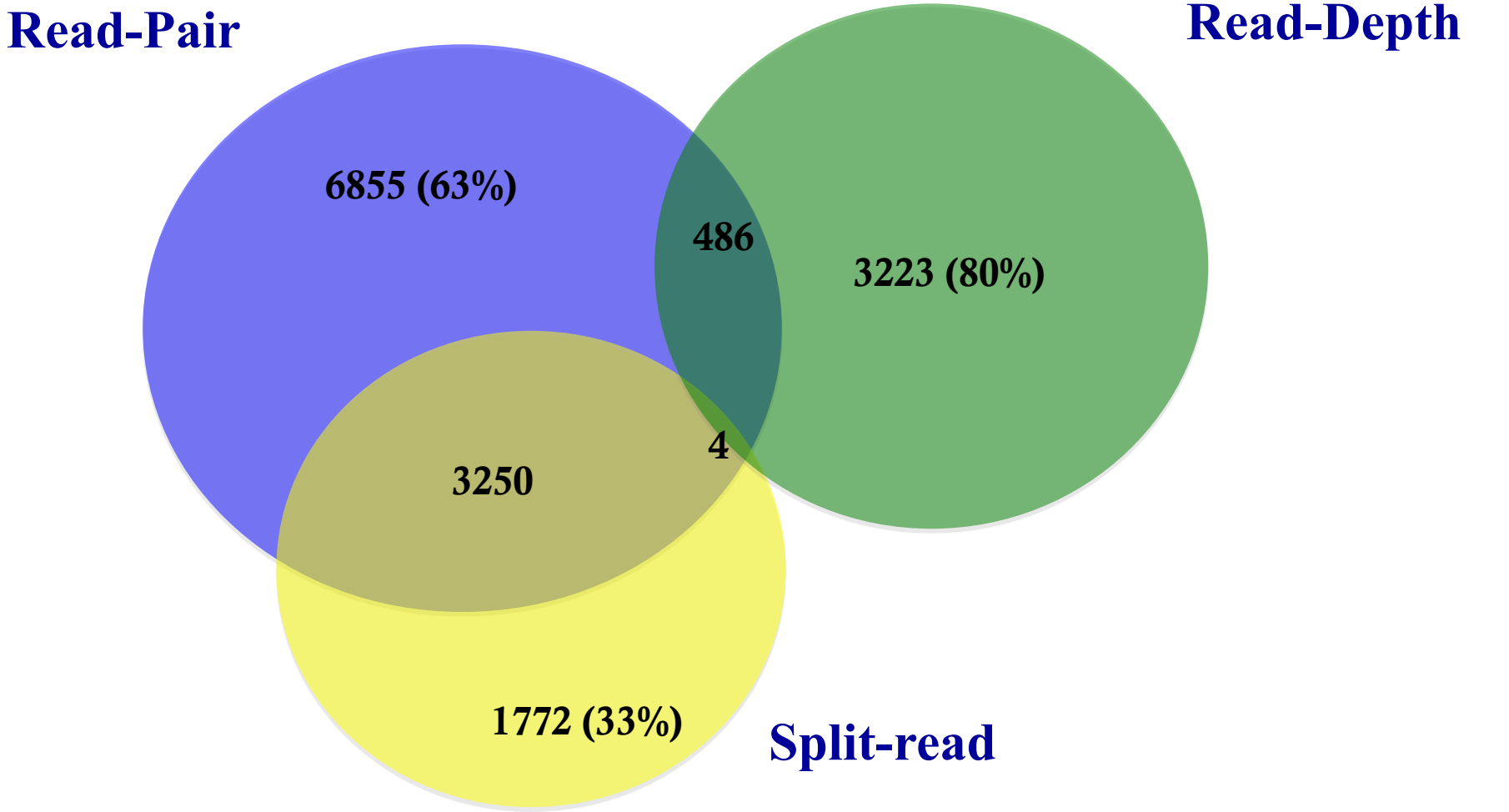
Next-Generation Sequencing Methods

- **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 eg. dC
- **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)



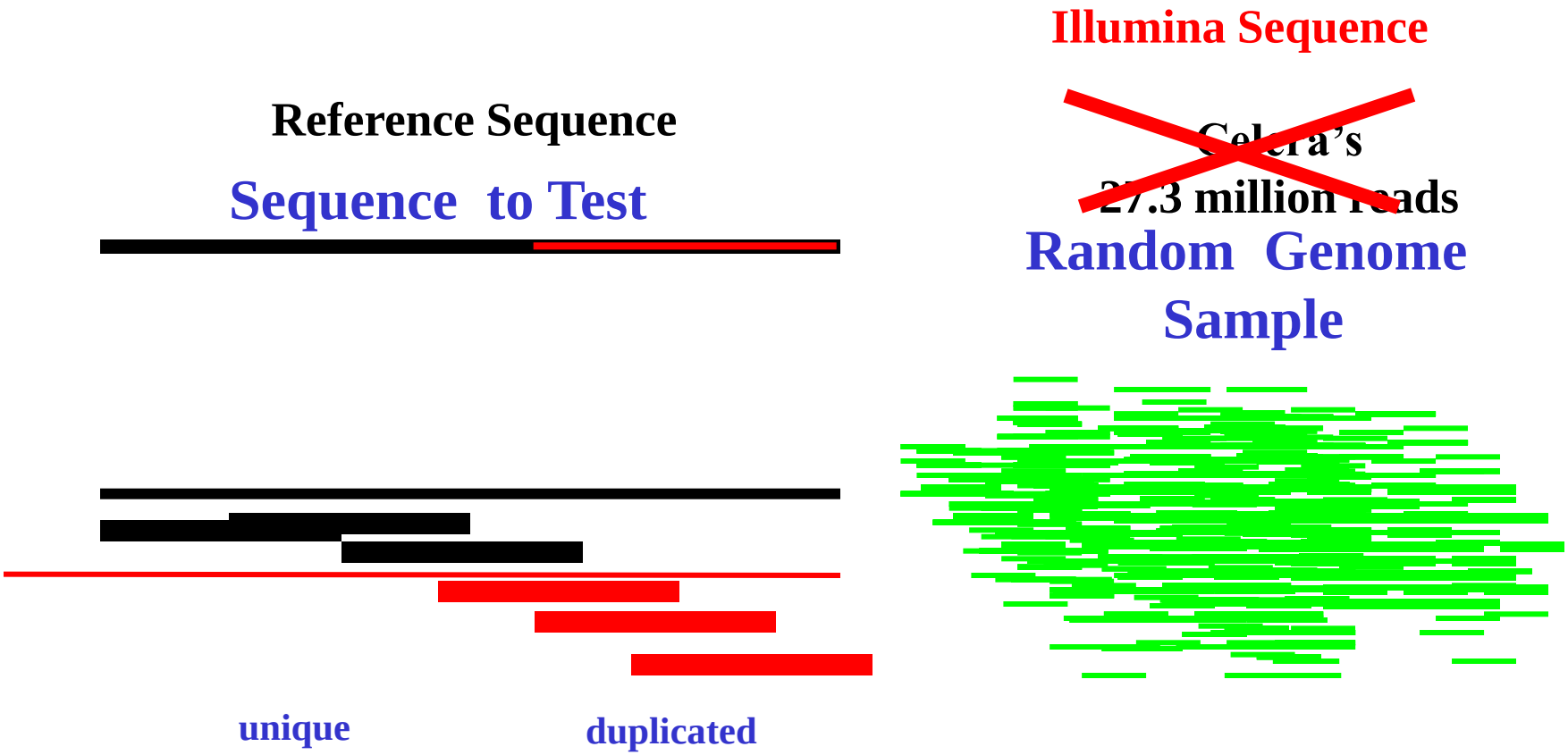
Mills *et al.*, Nature 2011

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.

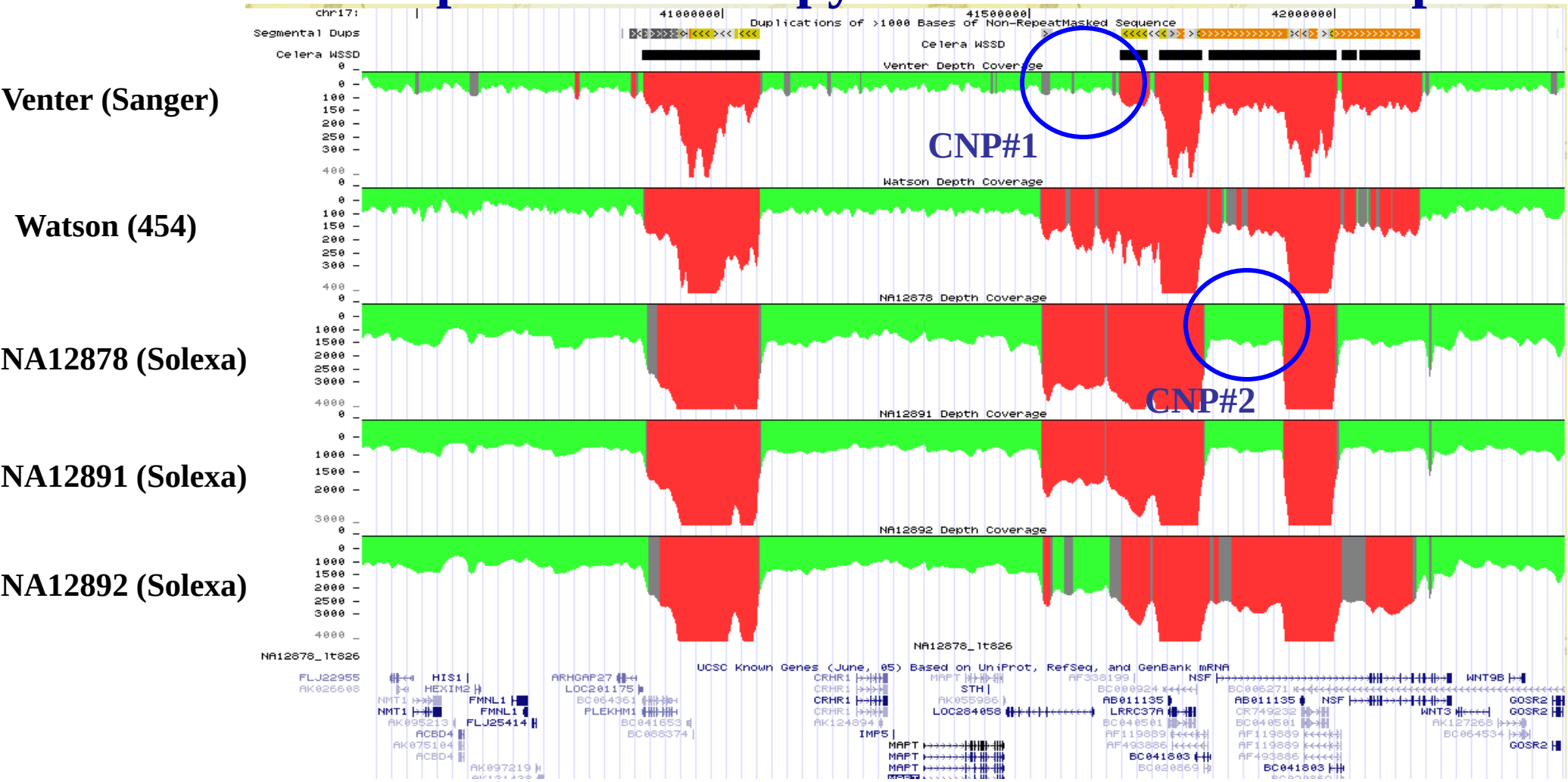
Using Sequence Read Depth

- 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)



Bailey et al., Science, 2002

Personalized Duplication or Copy-Number Variation Maps

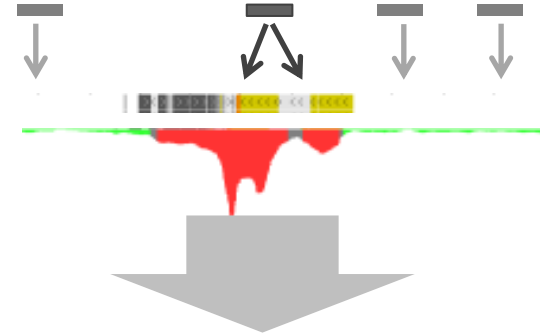


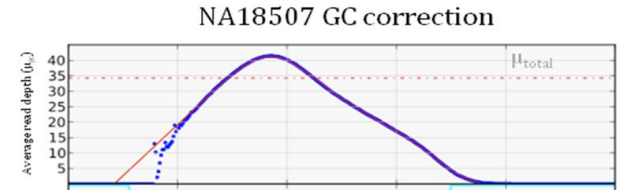
•Two known ~70 kbp CNPs, CNP#1 duplication absent in Venter but predicted in Watson and NA12878, CNP#2 present mother but neither father or child

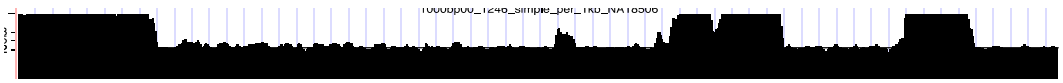
Alkan, Nat. Genet, 2009

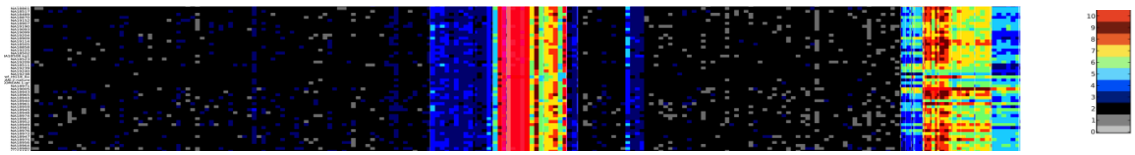
Copy number from short read depth

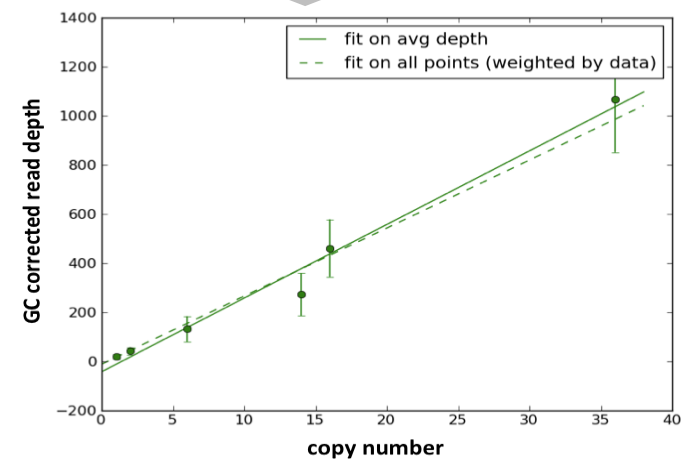
- Map reads to reference with *mrsFAST*
  - Records all placements for each read
  - http://mrsfast.sourceforge.net
- Per-library QC, (G+C)-bias correction
- Train estimator using depths at regions of known, invariable copy
- 1 kbp-windowed CN genomewide heatmap





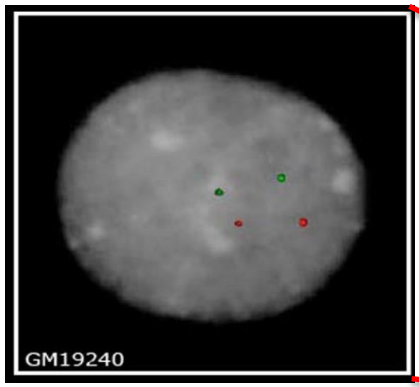


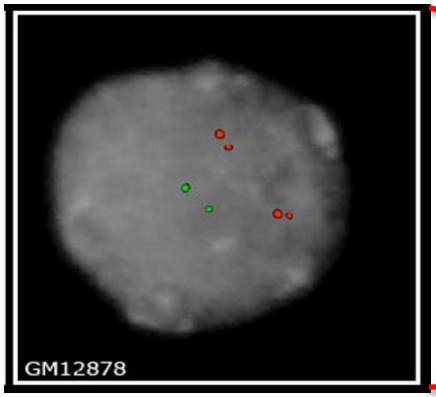


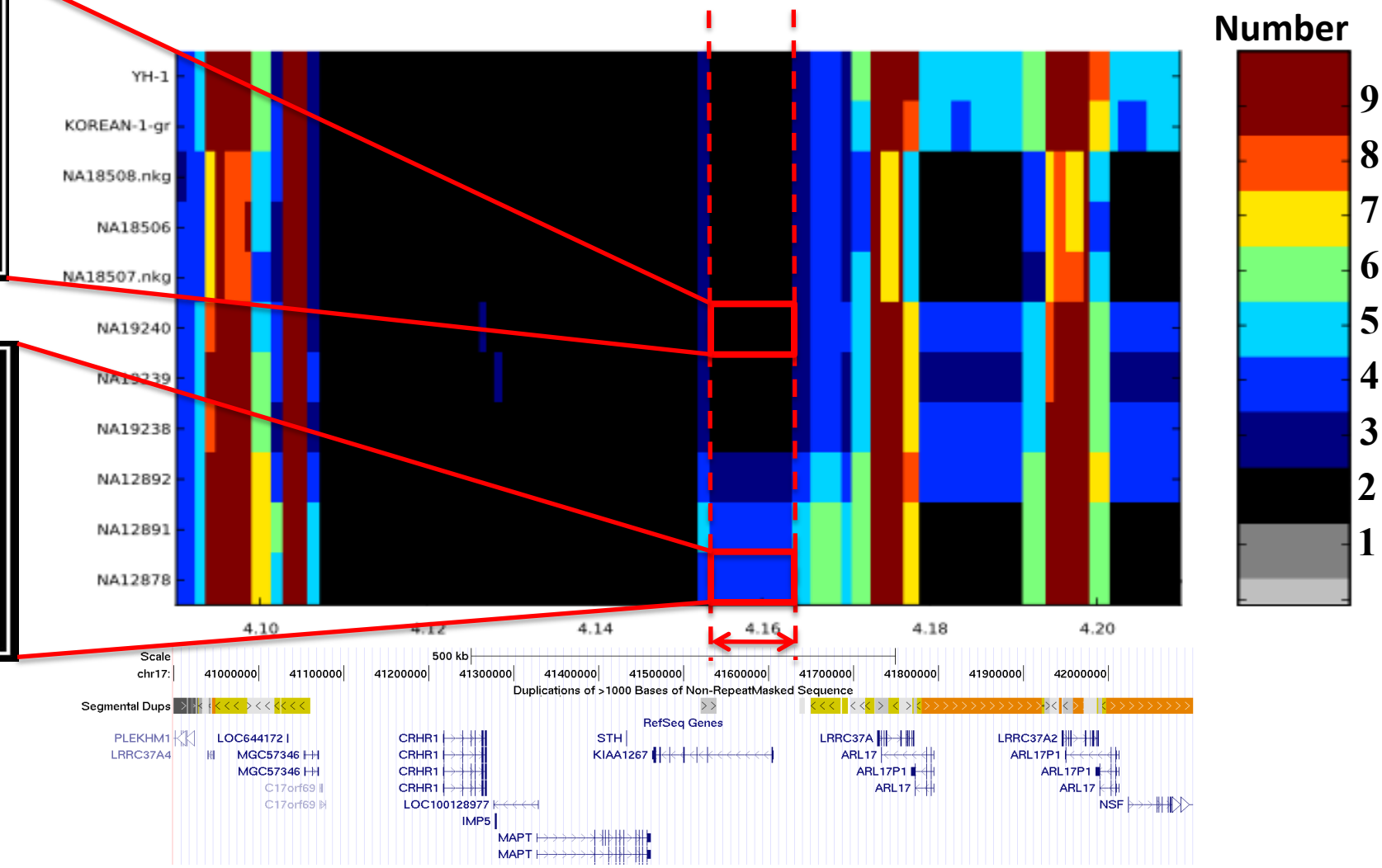


Read-Depth CNV Heat Maps vs. FISH

Interphase FISH

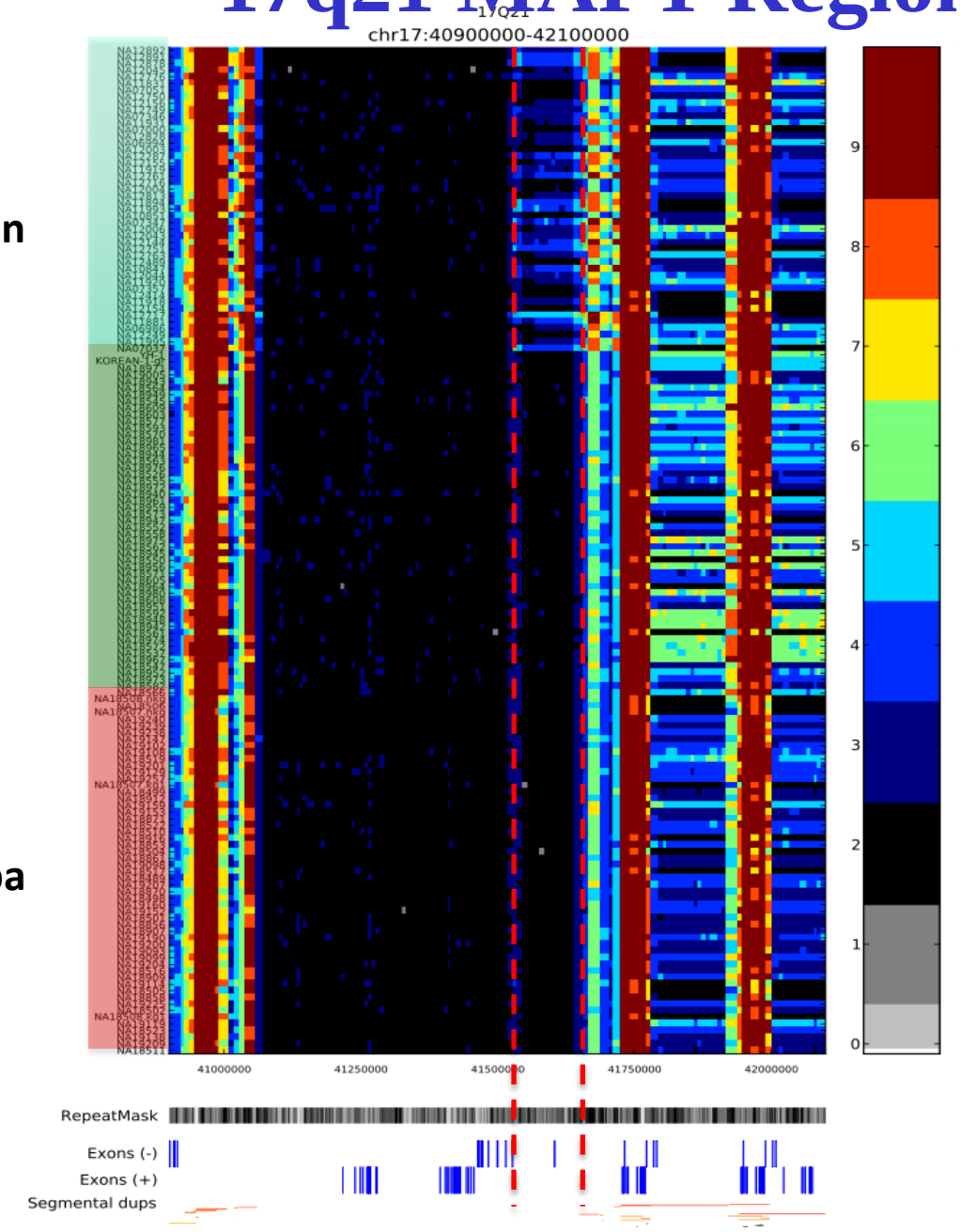






- 72/80 FISH assays correspond precisely to read-depth prediction (>20 kbp)
- 80/80 FISH assays correspond precisely to+/- 1 read-depth prediction

17q21 MAPT Region for 150 Genomes

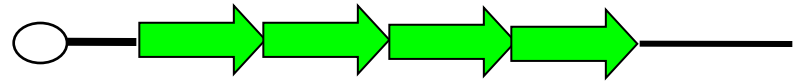


71% of Europeans carry at least Partial duplication distal (17q21 associated)—all inversions carry the duplication

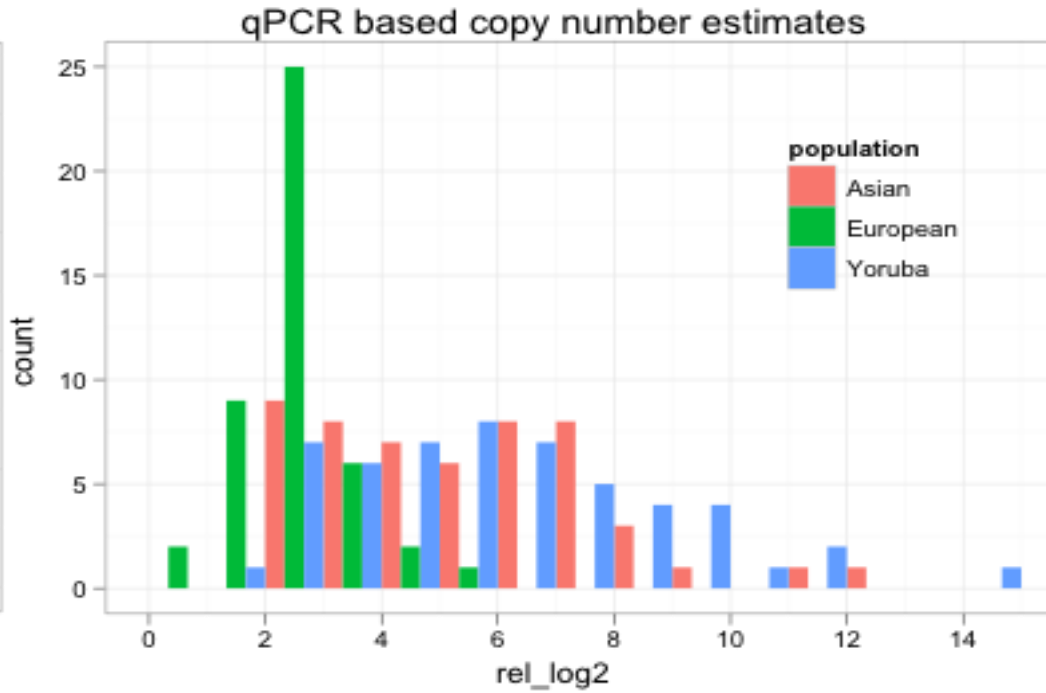
24% of Asians are hexaploid for NSF gene N-ETHYLMALEIMIDE-SENSITIVE FACTOR potentially important in synapse membrane fusion; NSF (decreased expression in schizophrenia brains (Mimics, 2000), Drosophila mutants results in aberrant synaptic transmission)

Sudmant et al., 2010, Science

Read-Depth vs. Quantitative PCR

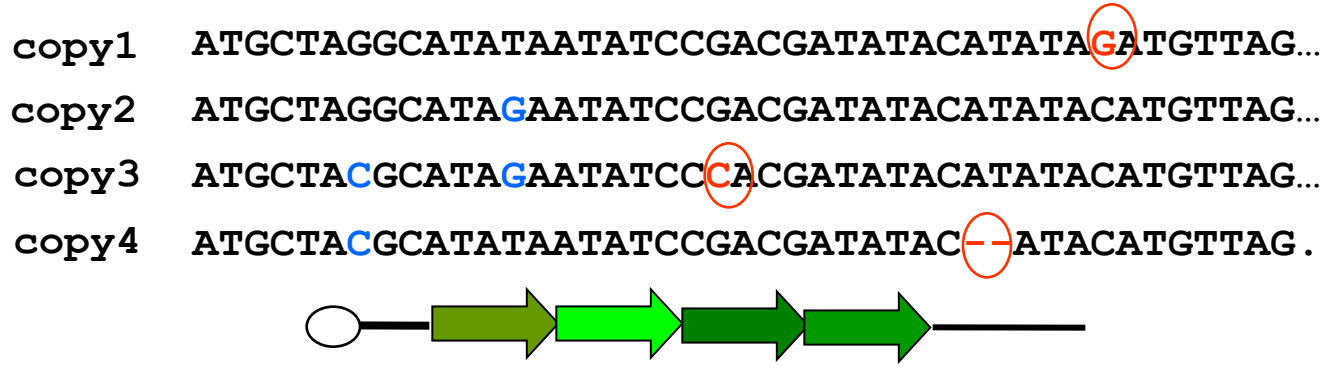


CCL3L1—chemokine ligand 3-like (1.9 kbp)



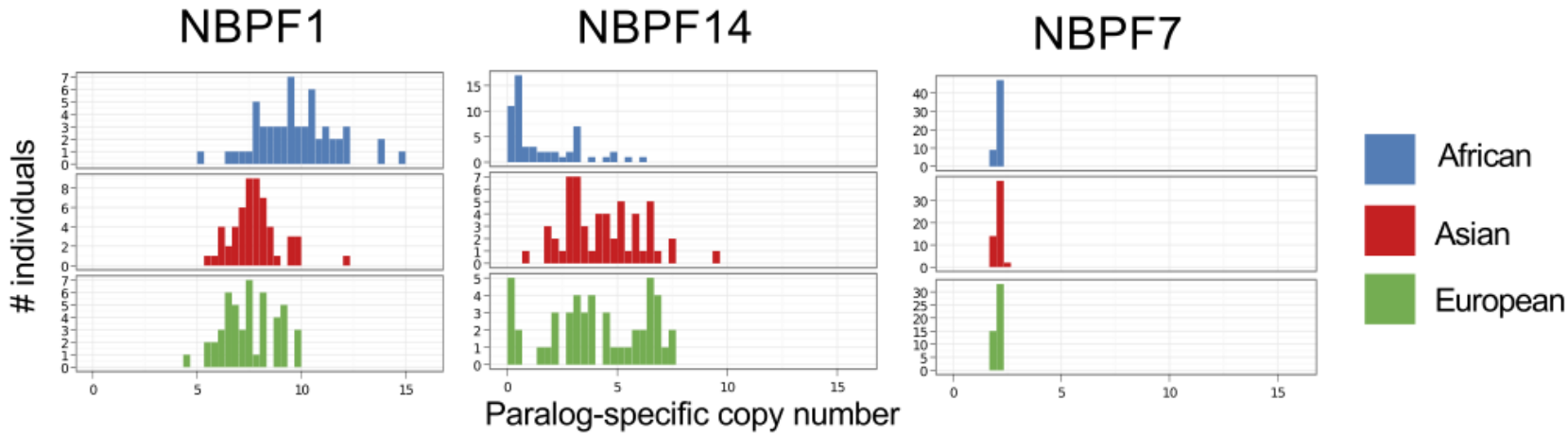
- **Tested 155 genomes read-depth (1-2 X coverage) vs. QPCR**
- **r^2=0.93 between sequence and quantitative PCR estimates**

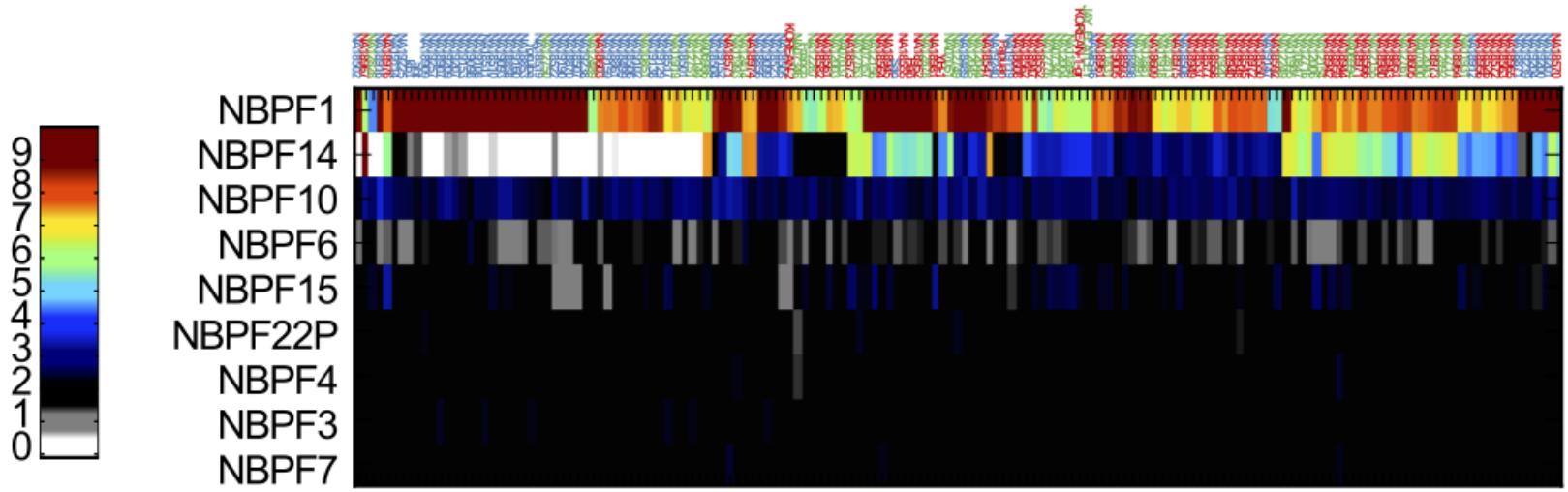
Unique Sequence Identifiers Distinguish Copies



- 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

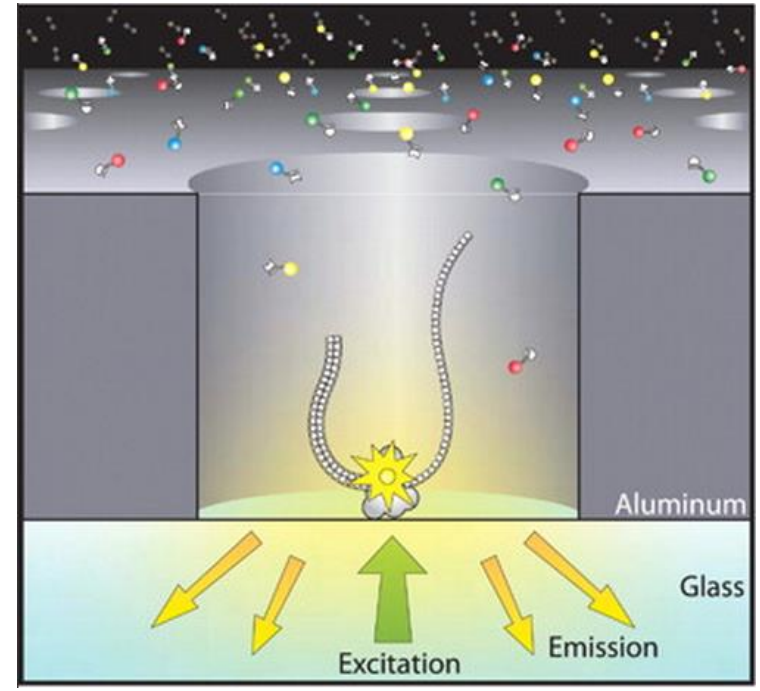




Going Forward

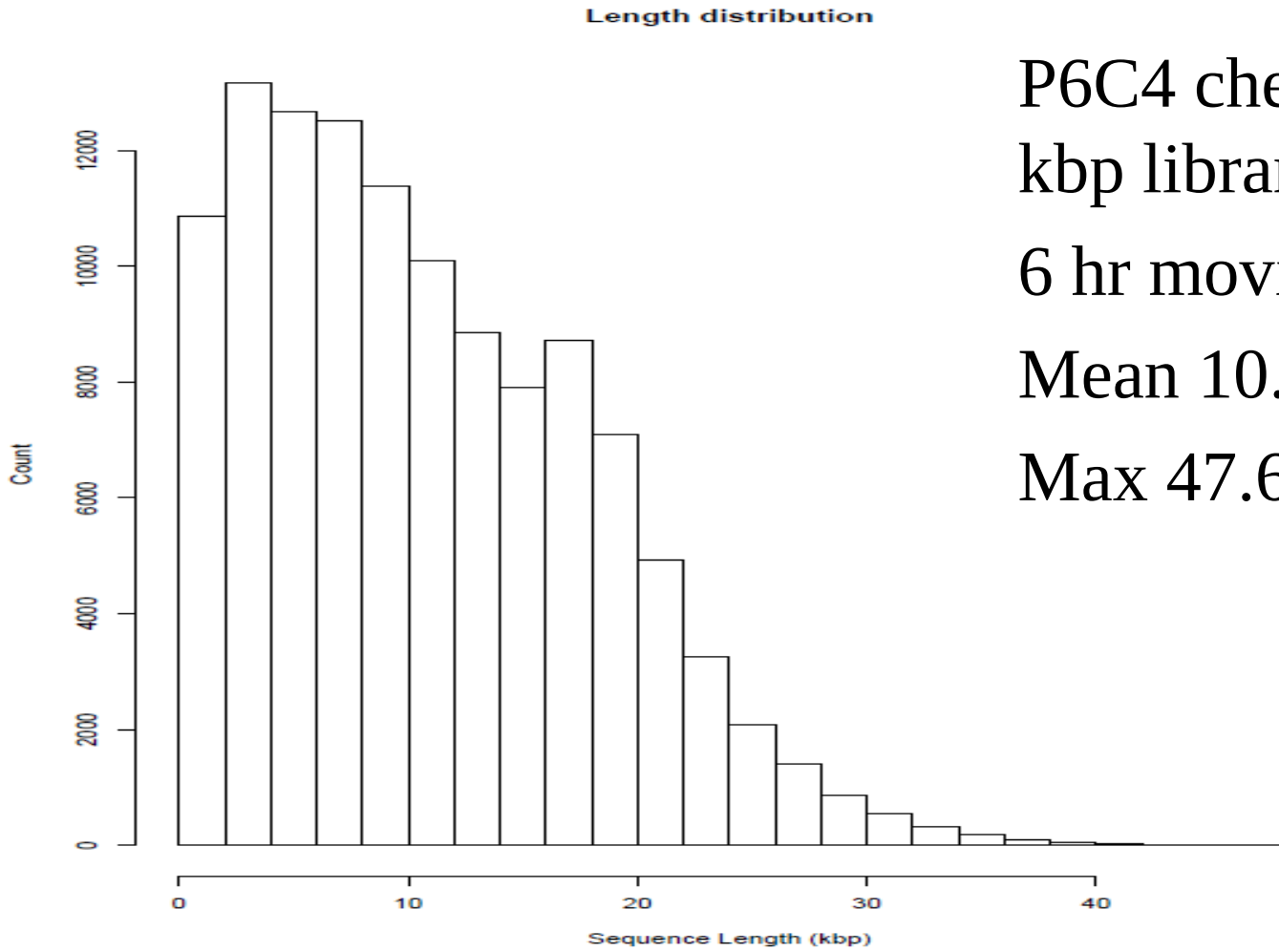
1) **Focus on comprehensive assessment of genetic variation**—large portions of human genetic variation are still missed
2) **Current NGS methods are indirect** and do not resolve structure but provide specificity and excellent dynamic range response.
3) **High quality sequence resolution of complex 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)



Long reads no cloning or amplification but lower throughput and 15% error rate

PacBio Sequence Reads are long



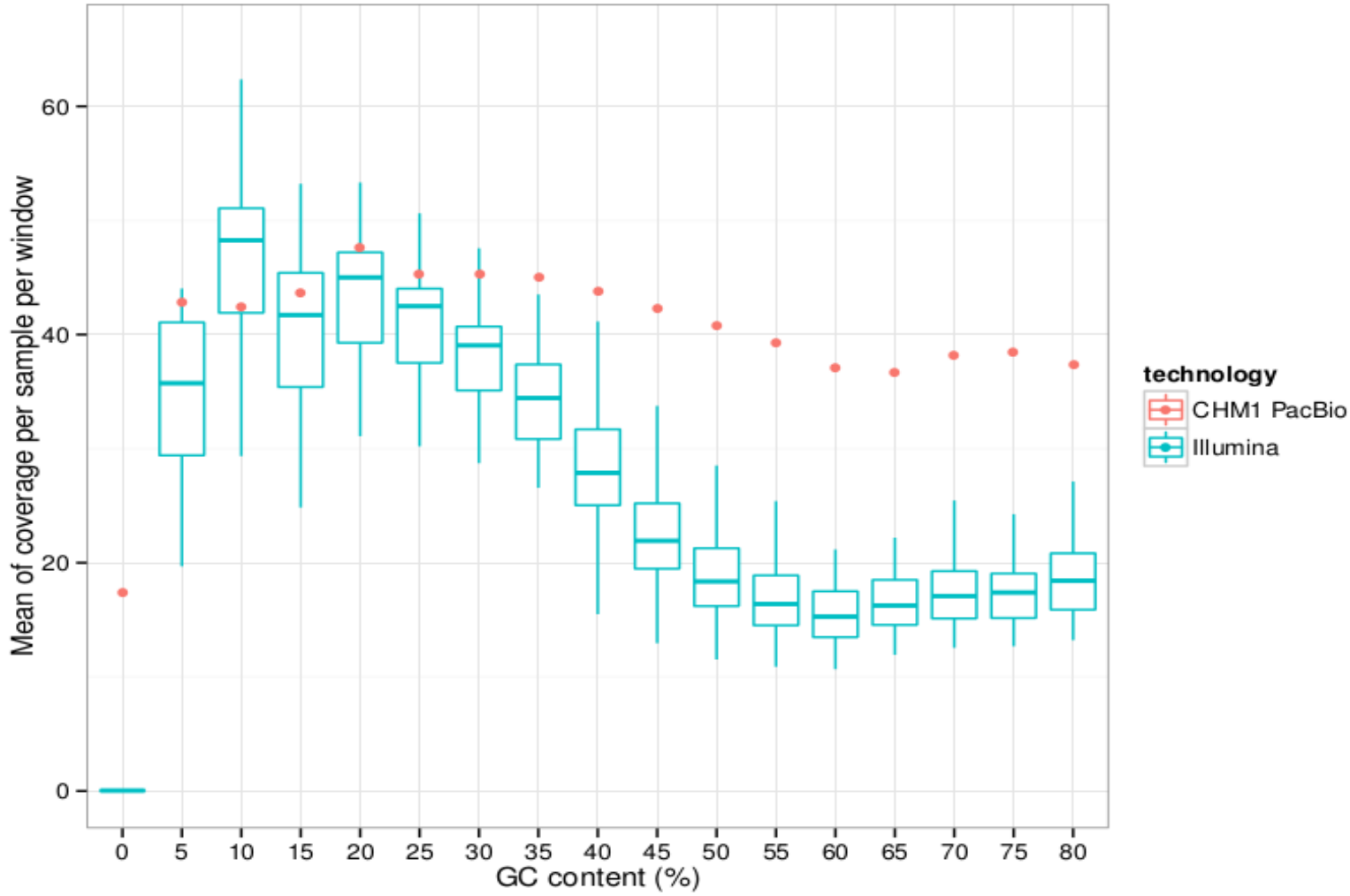
P6C4 chemistry—30-40 kbp libraries

6 hr movie

Mean 10.8 kbp read

Max 47.6 kbp

PacBio Sequence Reads are Uniform

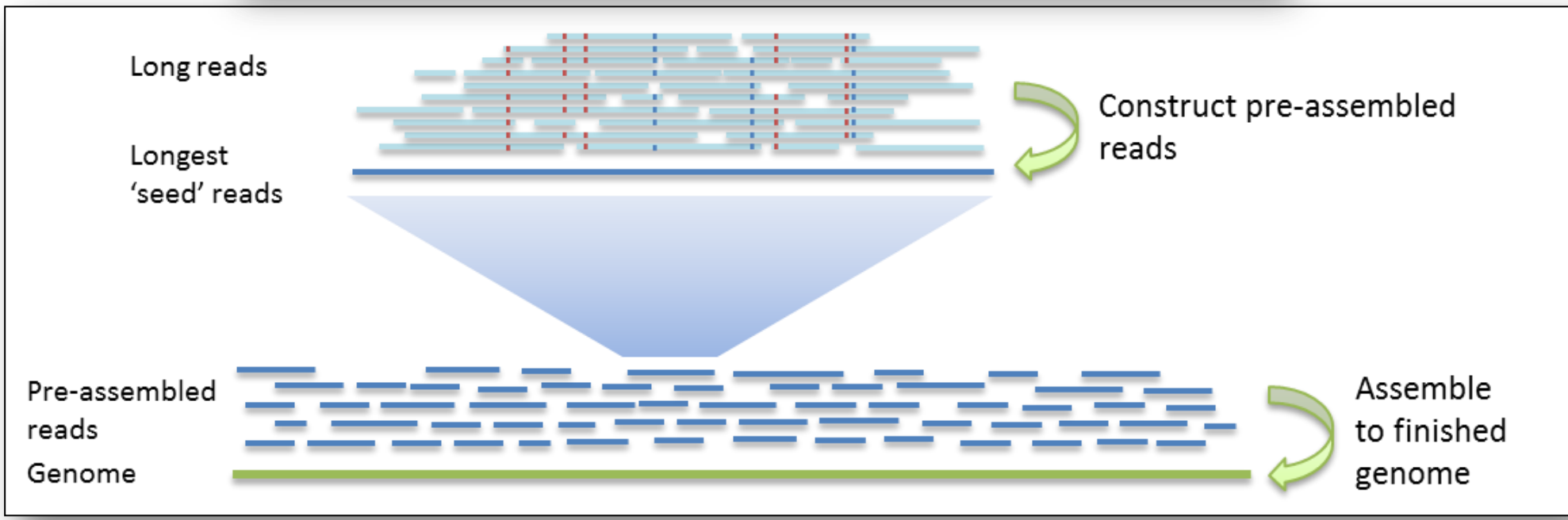


Algorithms: HGAP and QUIVER

ARTICLES

Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data

Chen-Shan Chin[1], David H Alexander[1], Patrick Marks[1], Aaron A Klammer[1], James Drake[1], Cheryl Heiner[1], Alicia Clum[2], Alex Copeland[2], John Huddleston[3], Evan E Eichler[3], Stephen W Turner[1] & Jonas Korlach[1]

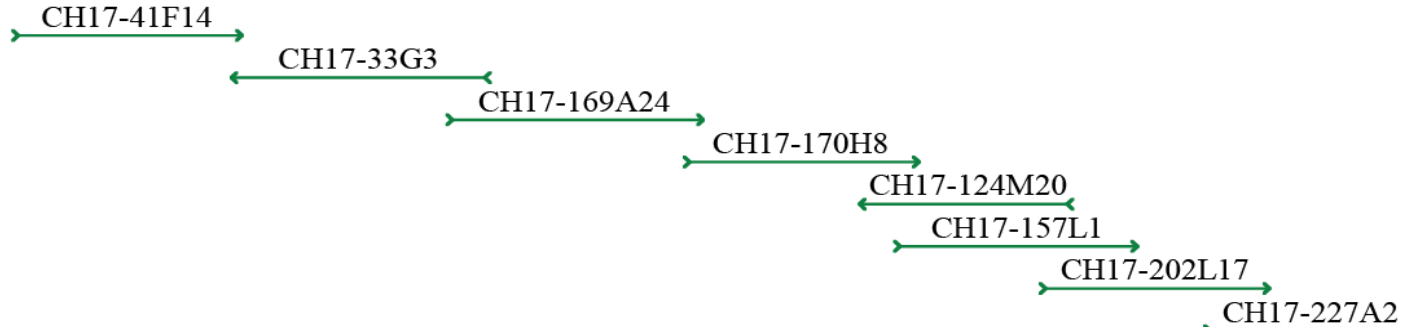


https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP

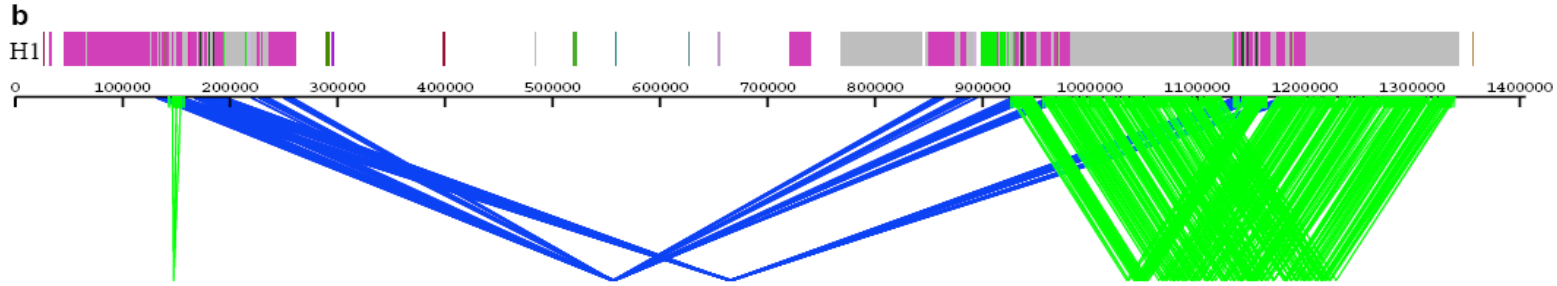
Chin et al. *Nat. Methods, 2013*

Clone Based Resolution of SV

BAC Tiling Path



Seg Dup Organization

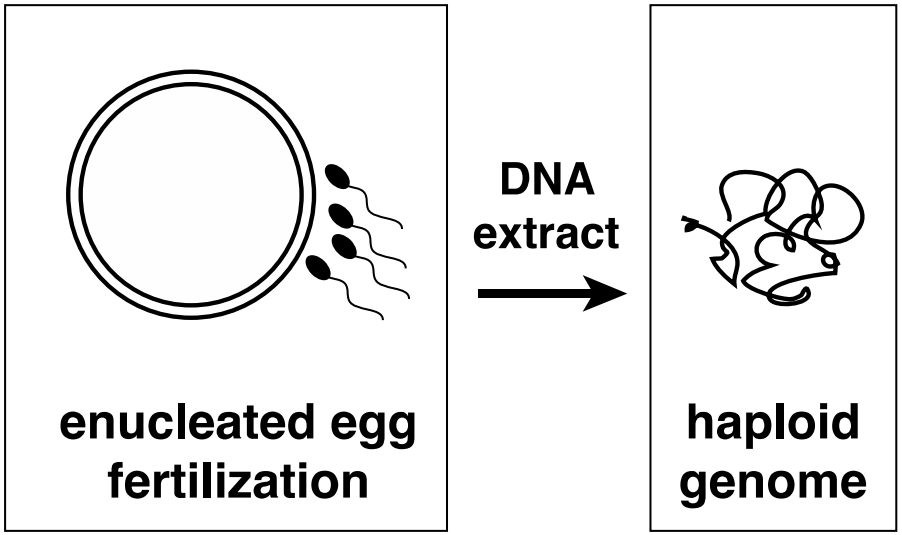


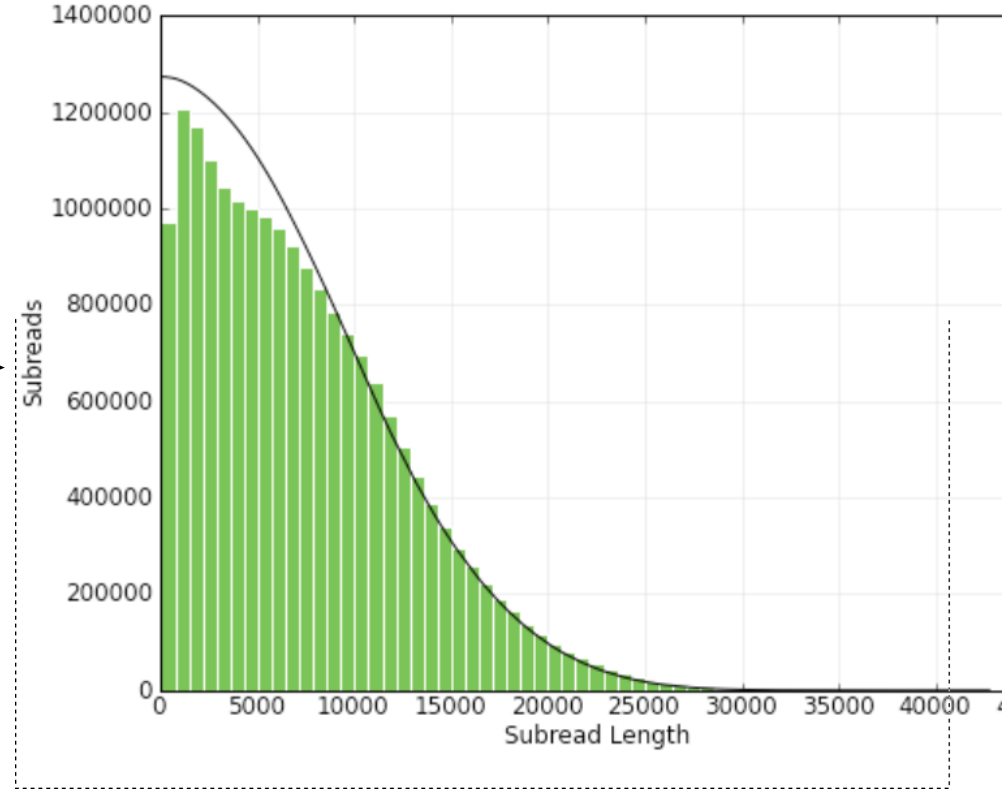
- Select tiling path of BAC clones corresponding to a complex region previously sequenced using Sanger
- Sequence each clone (~200 fold) using on average 1 SMRT Cell and assemble using HGAP and QUIVER
- Compare Sanger and Pacbio assembly using BLASR shows accurate (QV>45) assembly of complex region of human genome by BAC– 125 differences—31/44 favor PacBio over Sanger

Huddleston et al. *Genome Res, 2014*

PacBio Whole Genome Sequencing

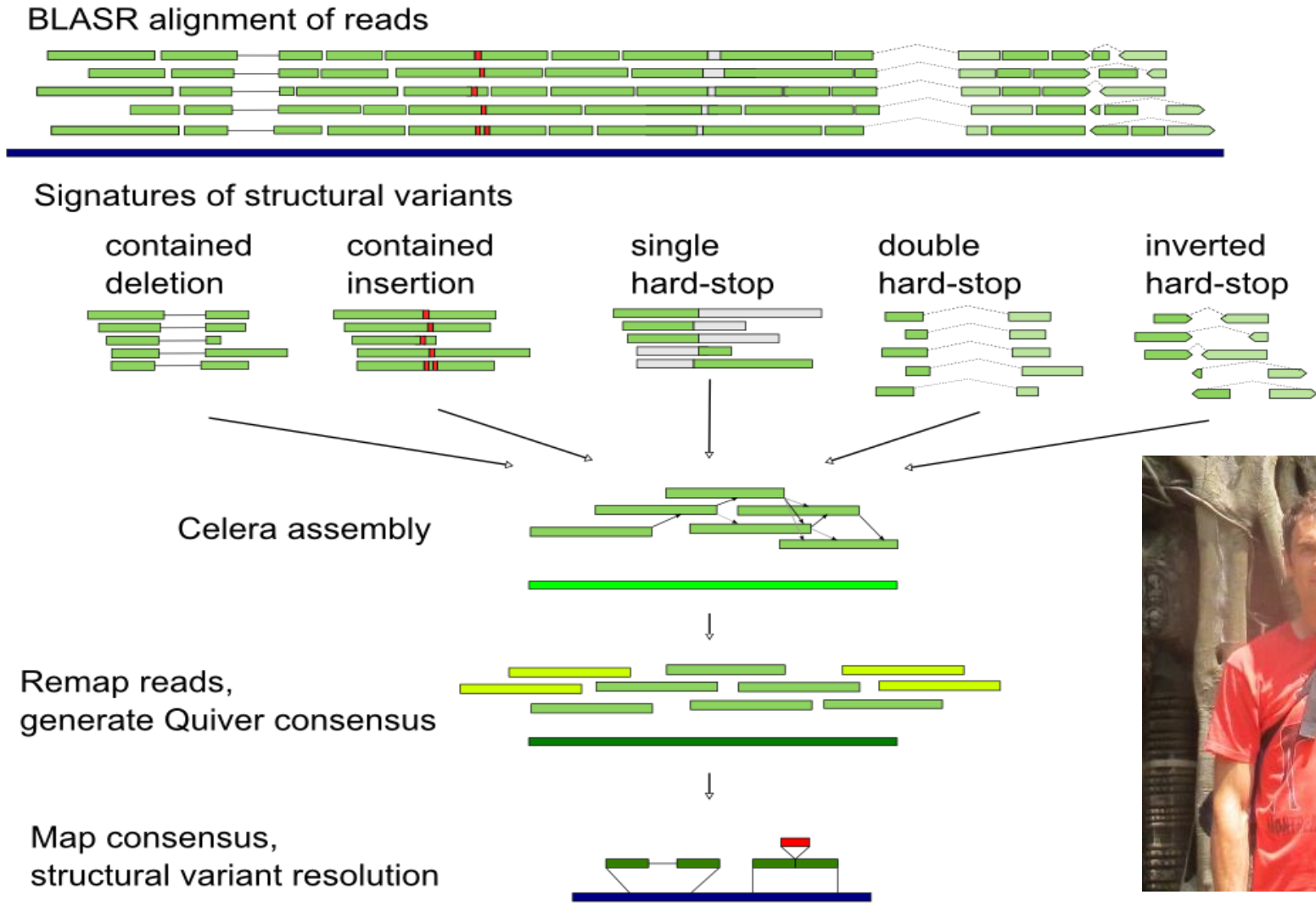
- **CHM1—complete hydatidiform mole (CHM1)- “Platinum Genome Assembly”**
- **45.8X Sequence coverage using RSII P5/C3 chemistry**
- **SMRT read lengths of ~9 kbp with 15% error.**





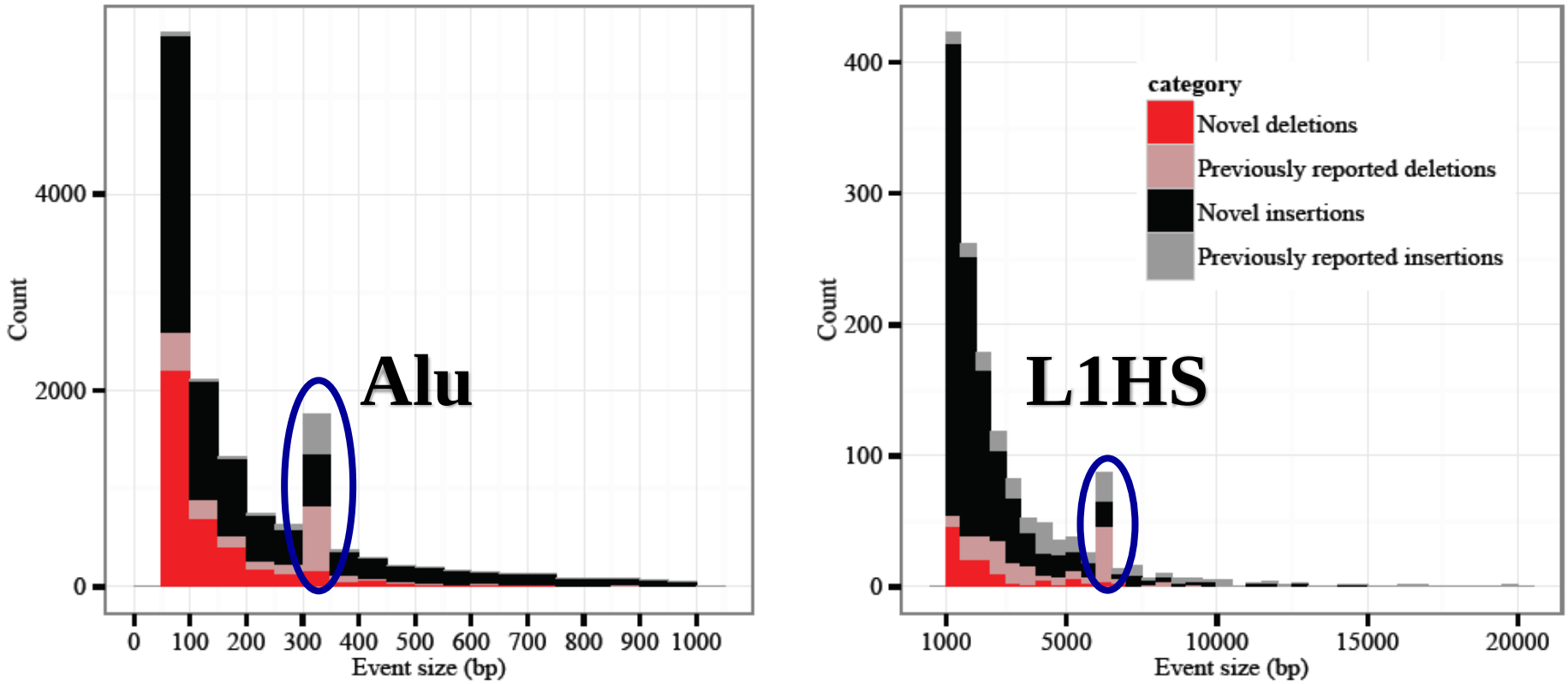
http://datasets.pacb.com/2013/Human10x/READS/index.html

SMRT-SV
Structural Variation Detection using PacBio



Chaisson et al, Nature, 2014/2015

Increased Resolution of Structural Variation



92% of insertions and 60% deletions (50- 5,000 bp) are novel

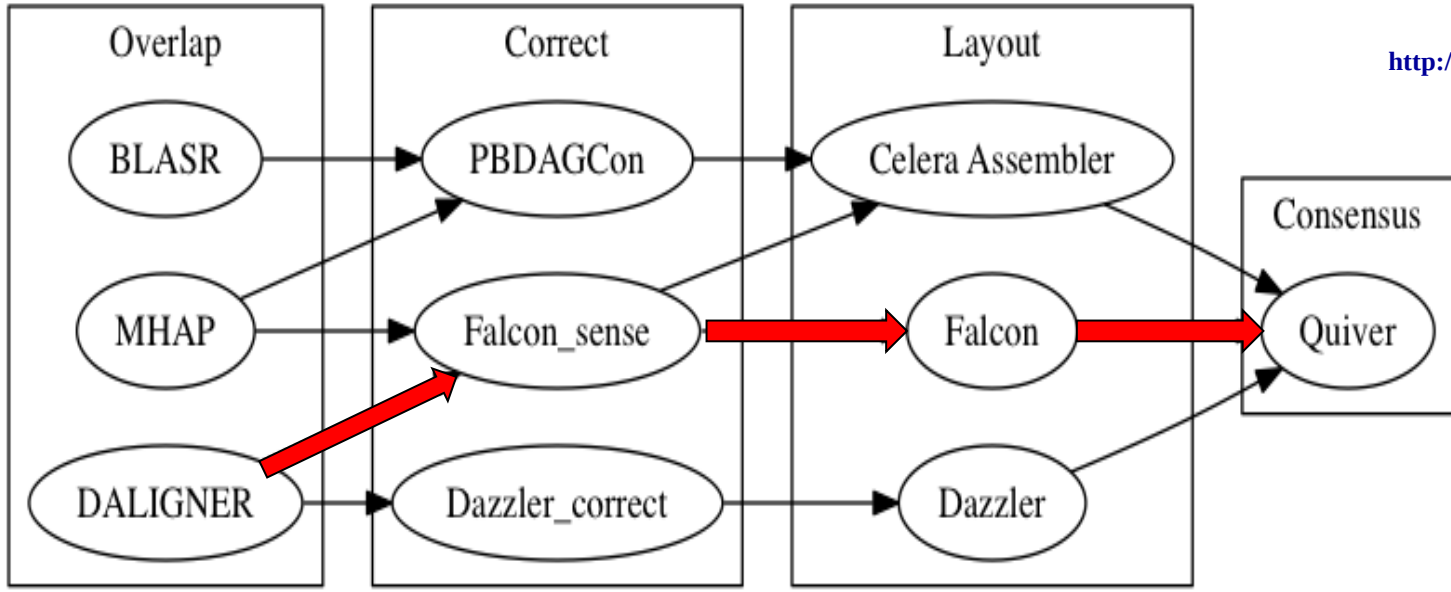
22,112 novel genetic variants corresponding to 11 Mbp of sequence

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

169 within coding sequence or UTRs of genes

Falcon SMRT Genome Assembly

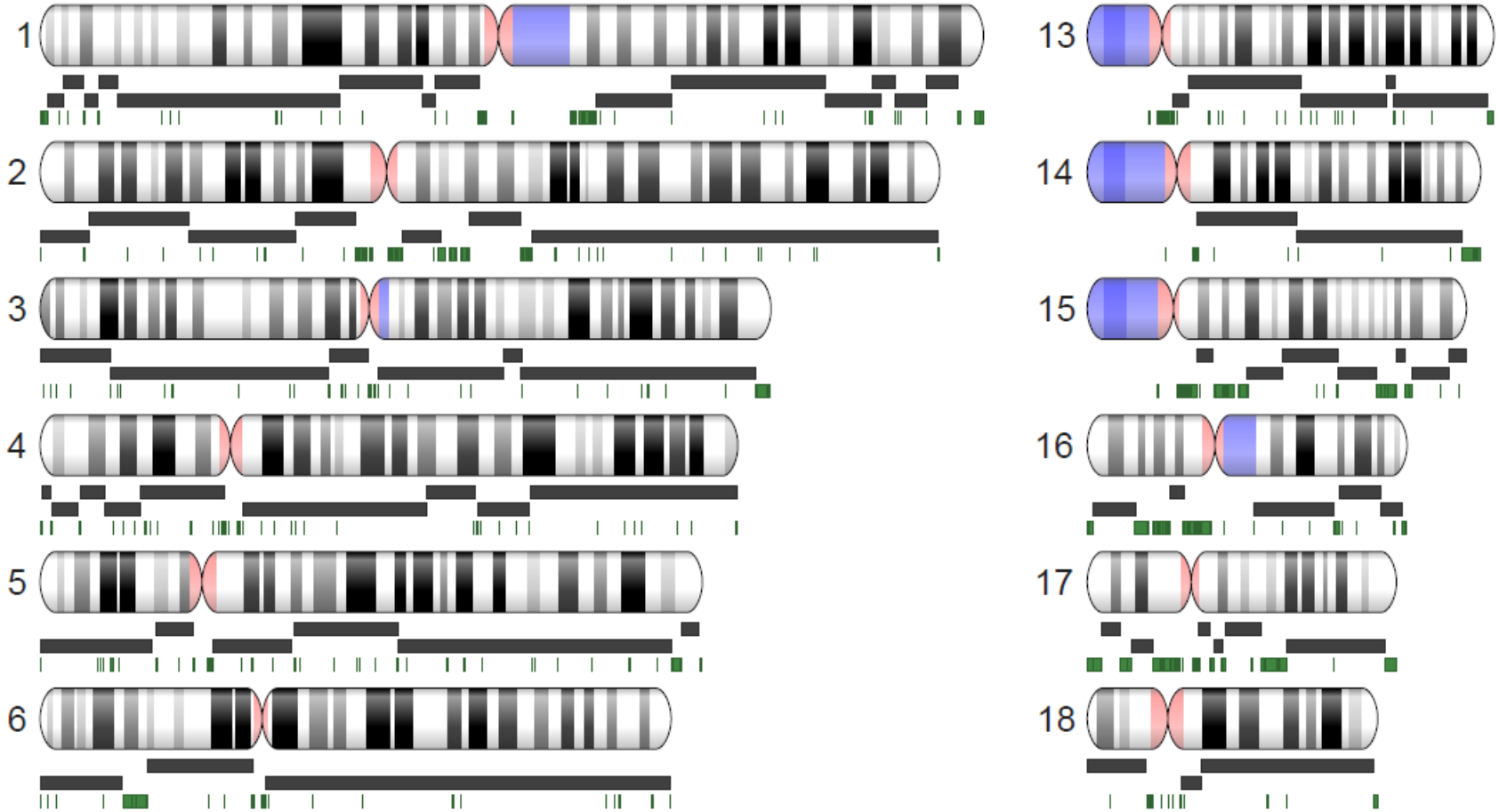
http://oukami4.deviantart.com



The stages of single-molecule sequencing assembly

- two phases: long reads are corrected and overlapped to generate a string graph—third phase “repeat unitig bridging”
- By Jason Chin http://github.com/PacificBiosciences/FALCON

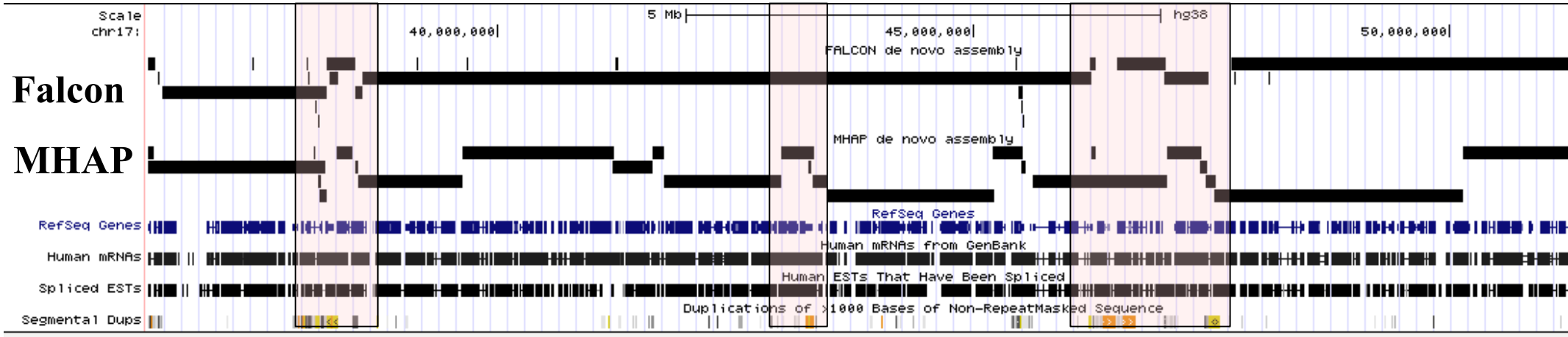
CHM1 Human Genome Assembly



- **67 X sequence coverage— Contig N50 27. 9 Mbp**
- **3,777 Contigs**

Chin *et al*, unpublished

Future: *De novo* Human Genome Assembly with SMRT WGS



- 125/167 Mbp of SD unresolved
- Contigs shatter over segmental duplications because 20 kbp reads are still not long enough.

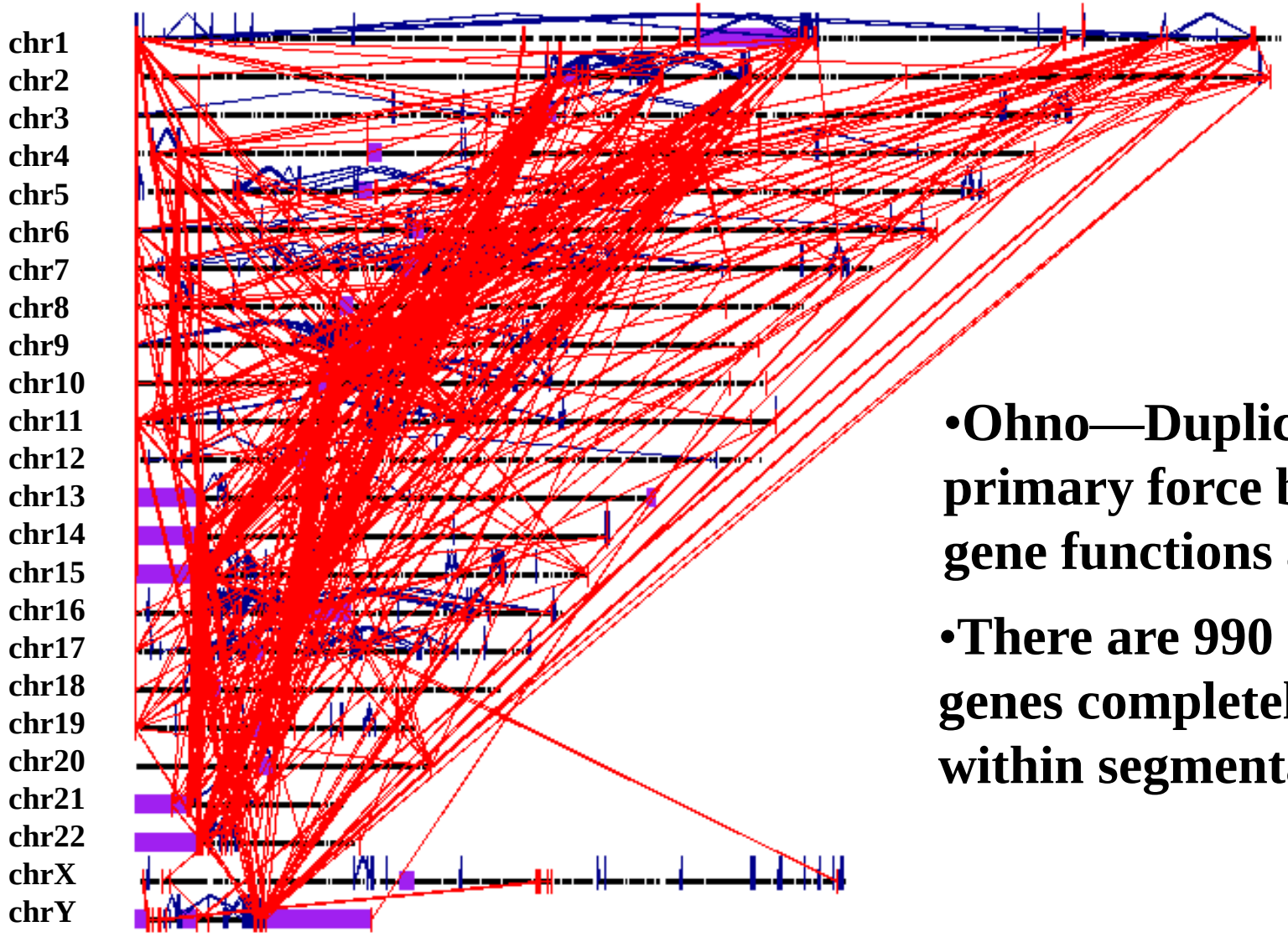
De novo Human Genome Assemblies PacBio/BAC Hybrid Assembly

- Platinum—higher quality than human reference genome=PacBio sequence >50X sequence coverage plus + BAC based sequencing of SD regions (CHM1 & CHM13)
- Continental References –2 African, 1 European, 1 Asian and 1 American genome
- PacBio trio – parent/child trios (40-20-20X).

Summary

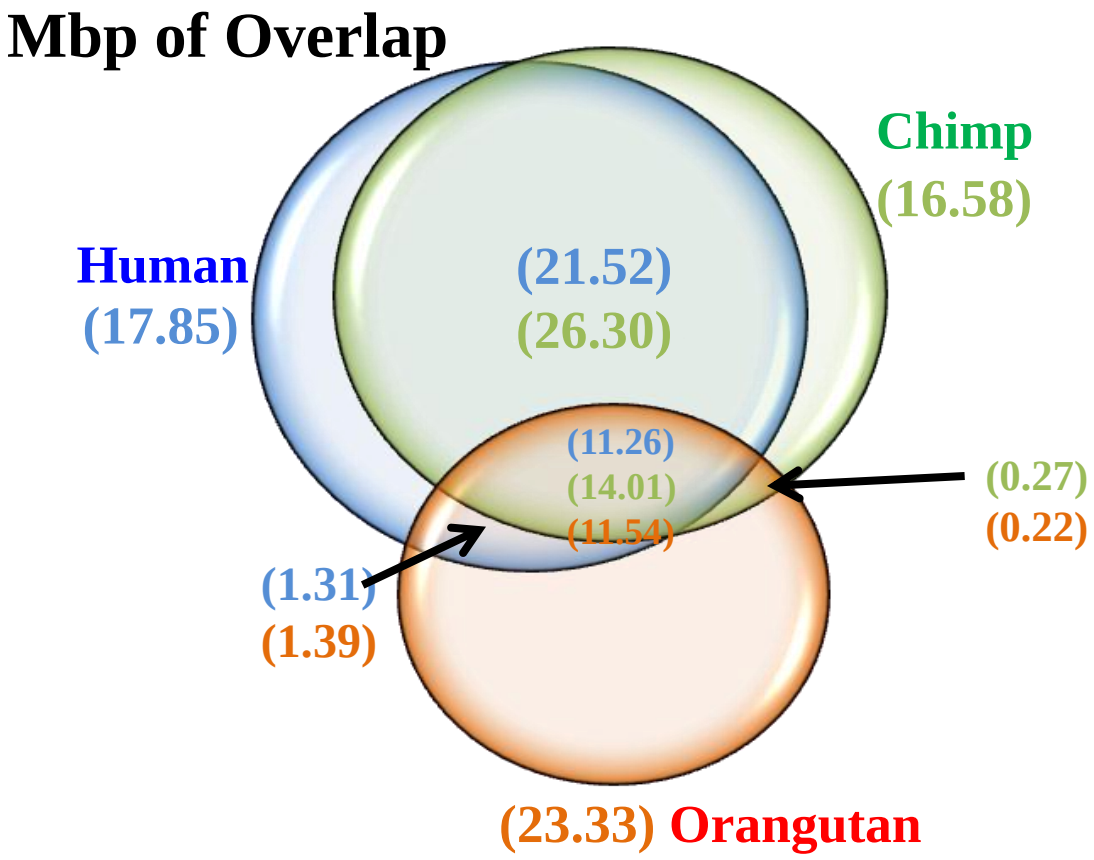
- Approaches
  - Multiple methods need to be employed—Readpair+Read-depth+SplitRead and an experimental method
  - Tradeoff between sensitivity and specificity
  - Complexity not fully understood
- Read-pair and read-depth NGS approaches
  - narrow the size spectrum of structural variation
  - lead to more accurate prediction of copy-number
  - unparalleled specificity in genotyping duplicated genes (reference genome quality key)
- Third generation sequencing methods hold promise but require high coverage—still expensive. ***Sequel***?

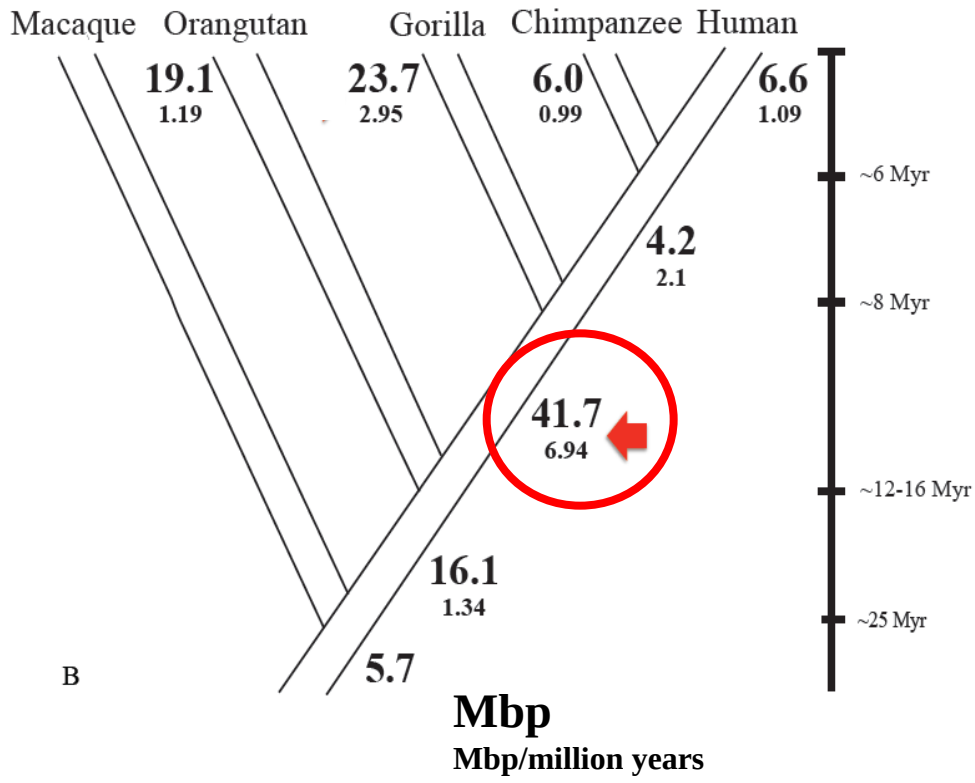
III. Why?



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

Duplication Acceleration in Human Great Ape Ancestor

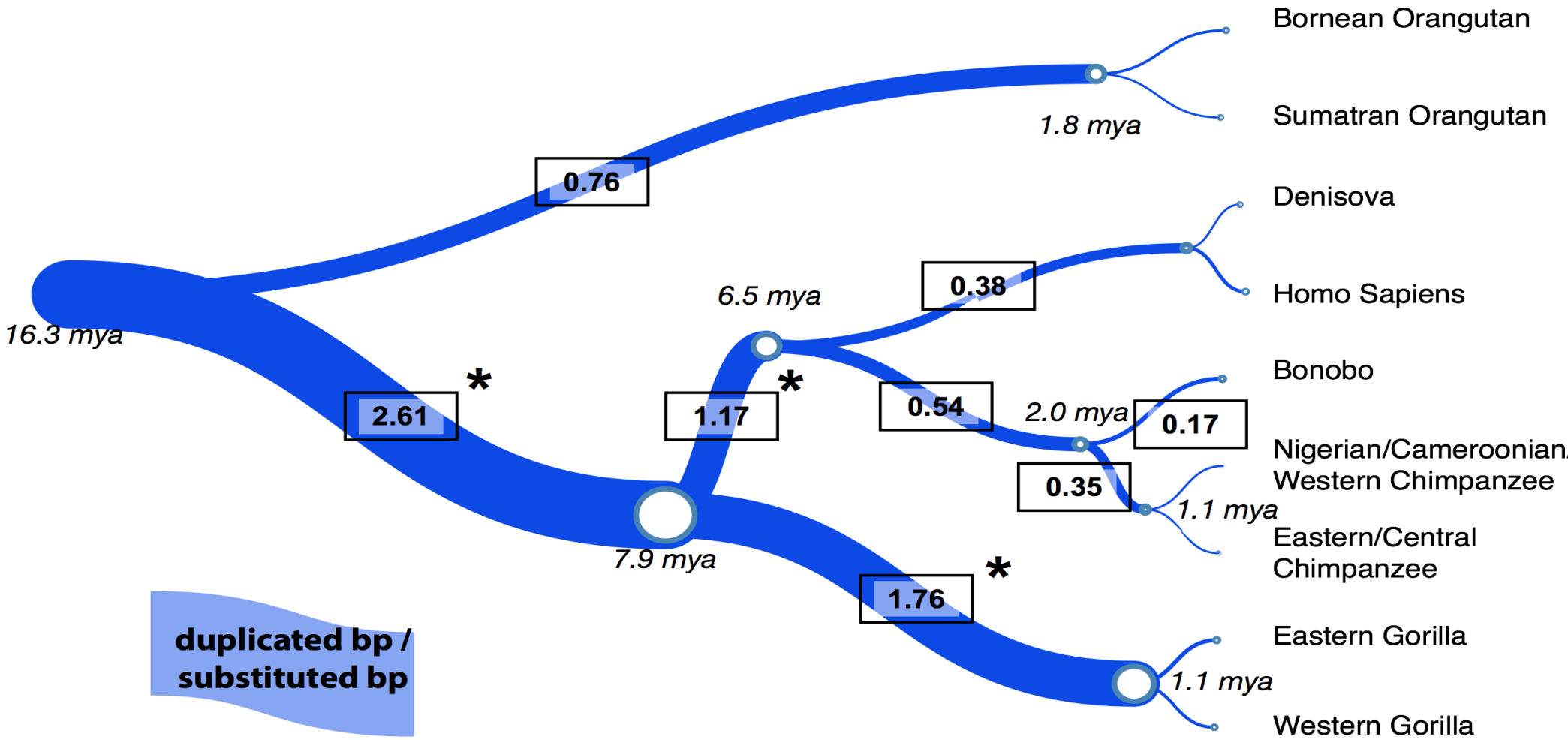




- A 3-4 fold excess in *de novo* segmental duplications in common ancestor of human, chimp and gorilla but after divergence from orangutan
- Not a continuous accumulation

Marques-Bonet et al., *Nature*, 2009; Ventura et al., *Genome Res.* 2011

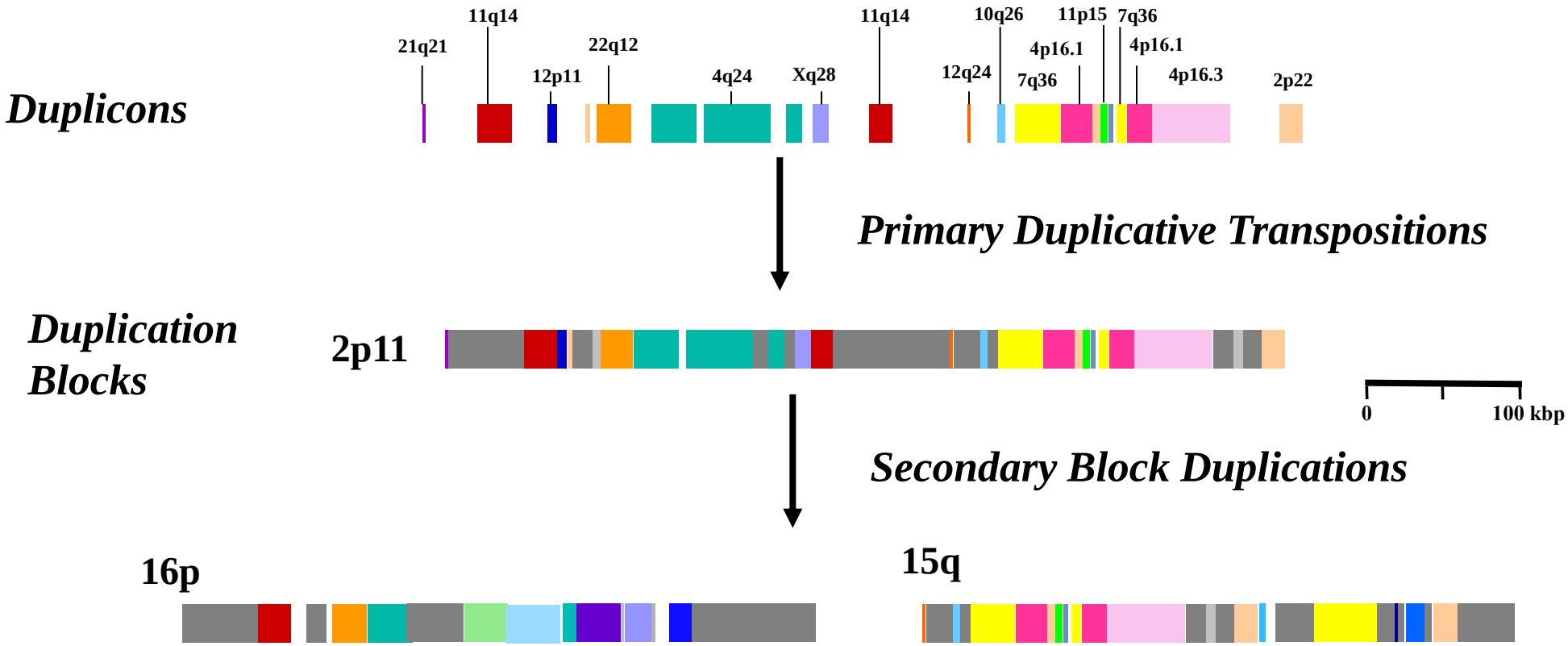
Rate of Duplication



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

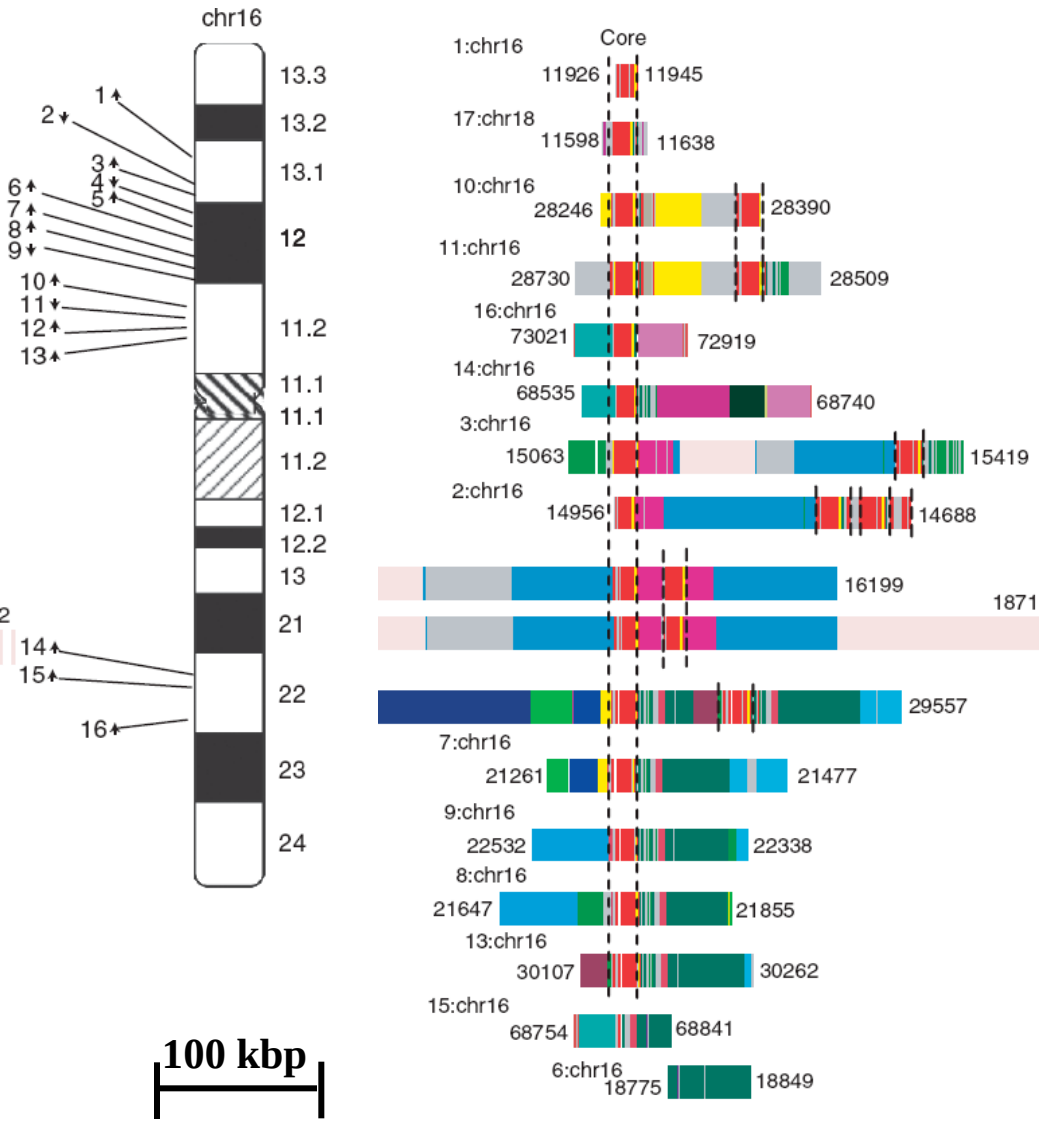
Sudmant PH et al. , *Genome Res.* 2013

Mosaic Architecture



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

Human Chromosome 16 Core Duplicon

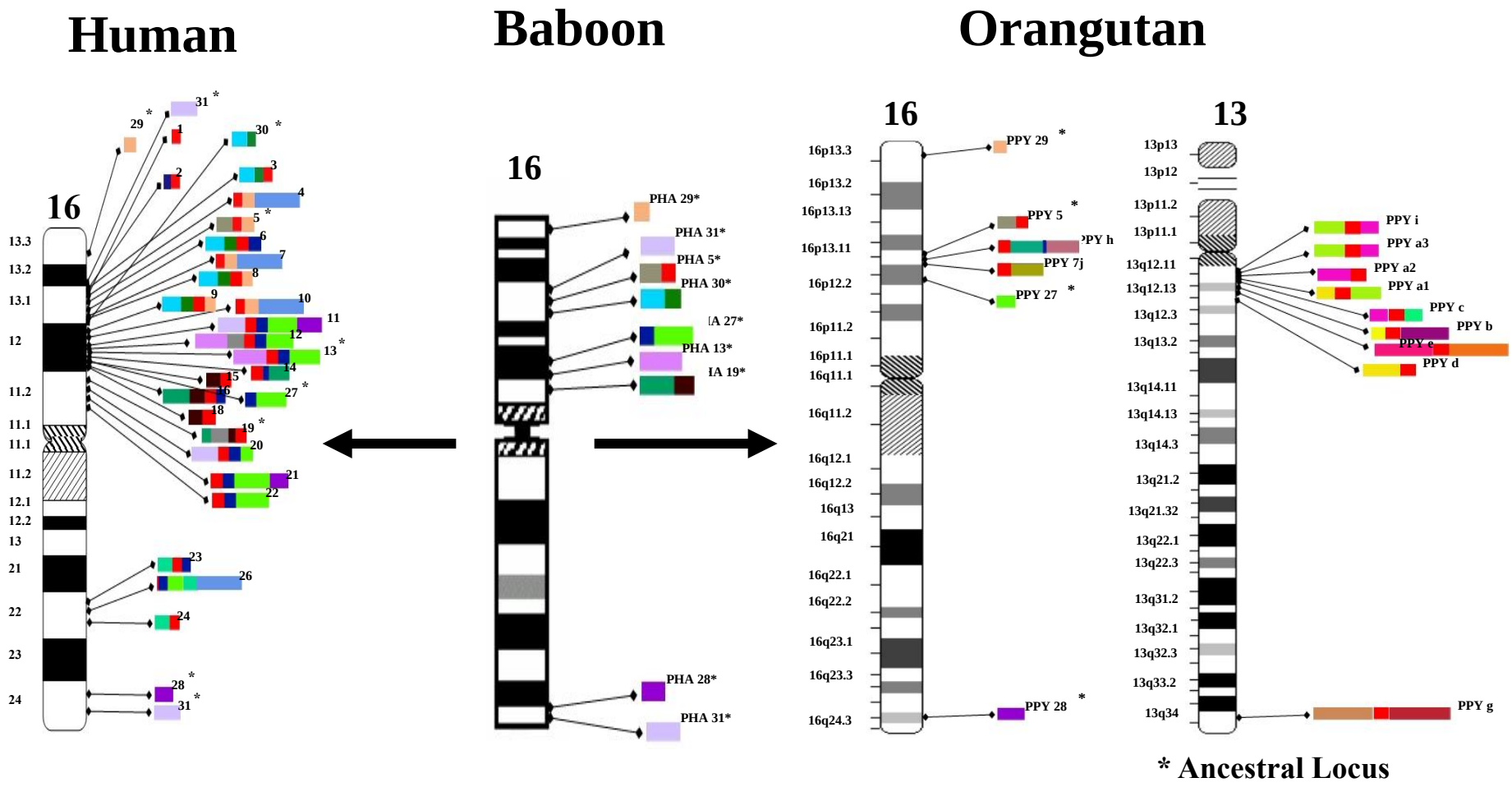


- **The burst of segmental duplications 8-12 mya corresponds to core-associated duplications which have occurred on six human chromosomes (chromosomes 1,2, 7, 15, 16, 17)**

- **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

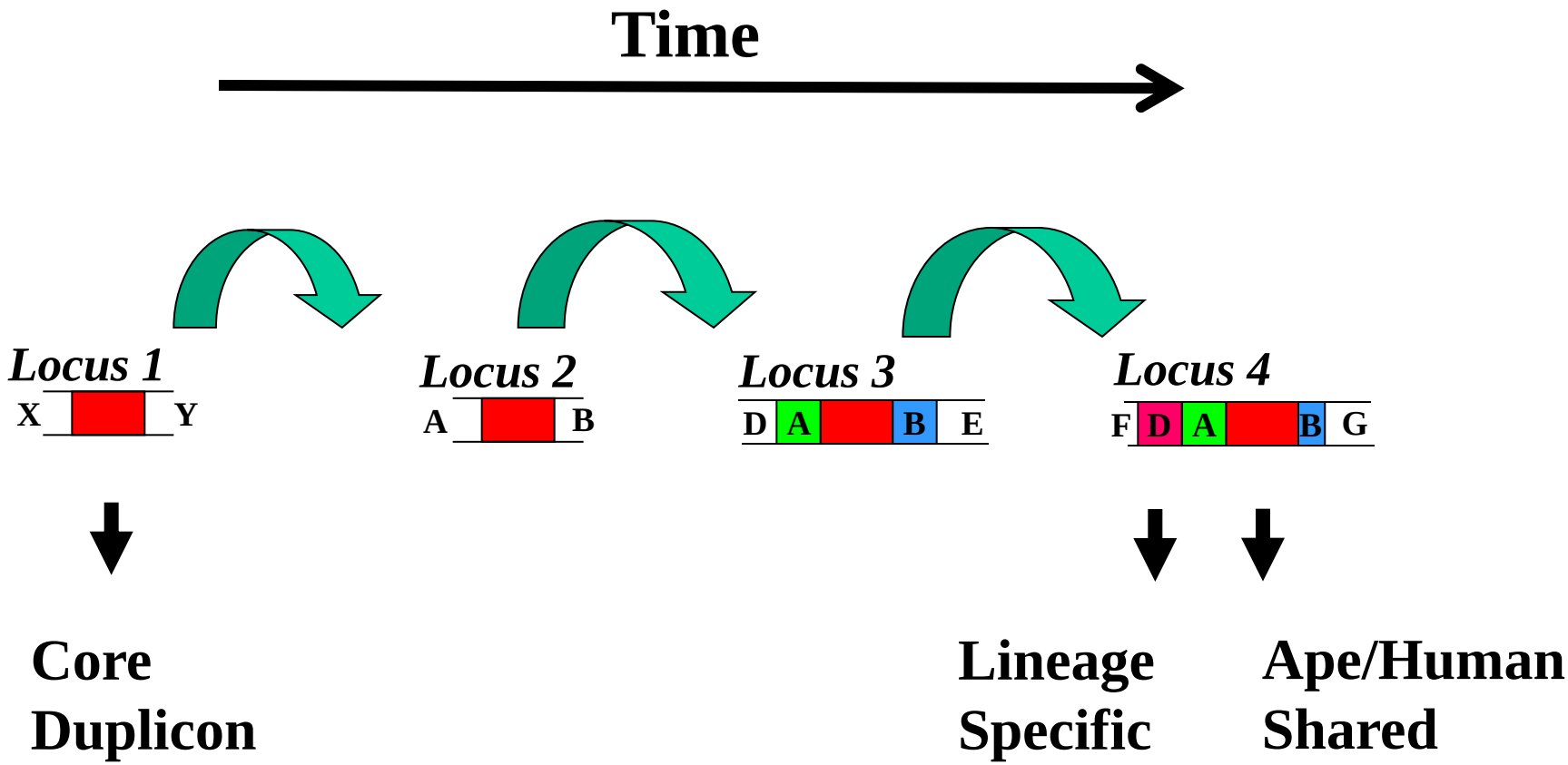
Increasing Duplication Complexity and Recurrence



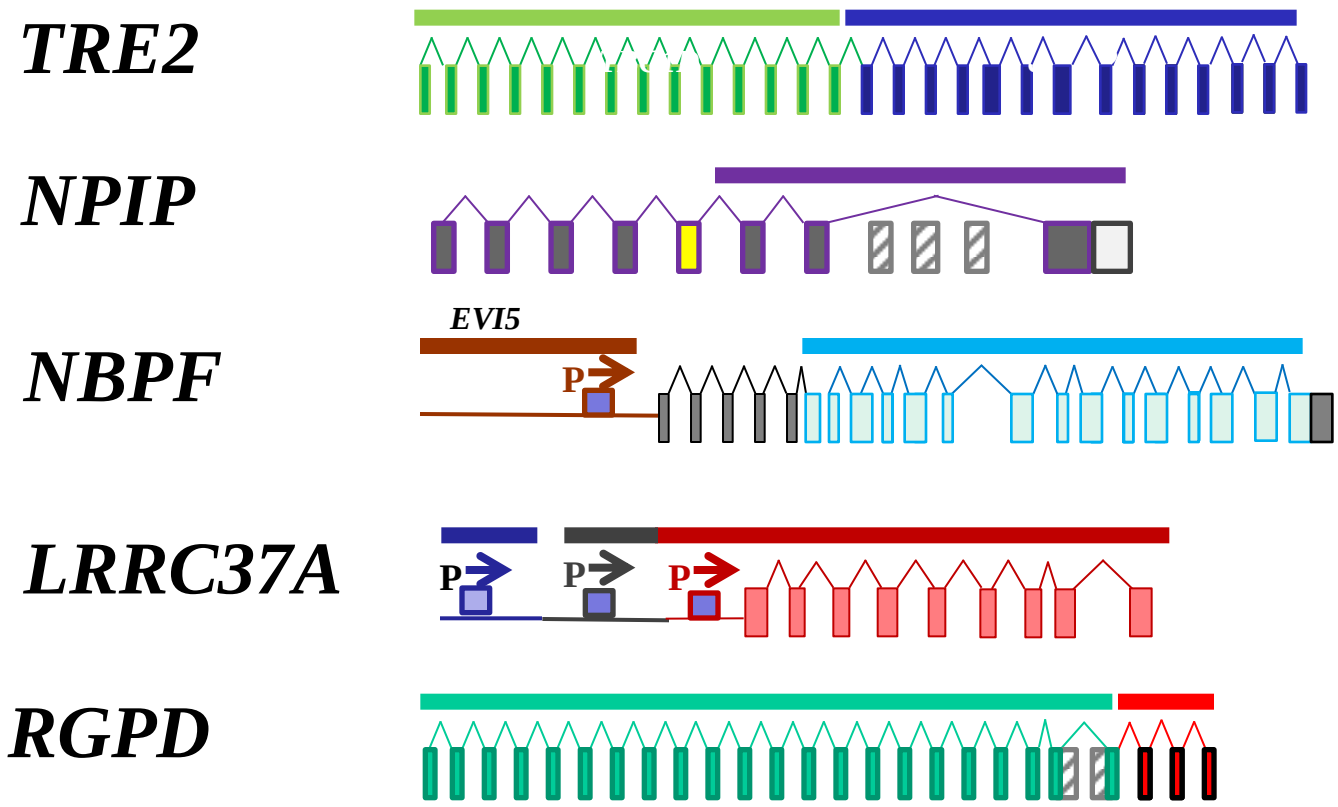
- Duplication blocks have become increasingly more complex (more duplicons) and have expanded in an interspersed fashion over the last 25 million years.
- Duplication blocks of different flanking content with exception of core

Johnson et al., *PNAS*, 2006

Core Expansion Model



Human Great-ape "Core Duplicons" have led to the Emergence of New Genes



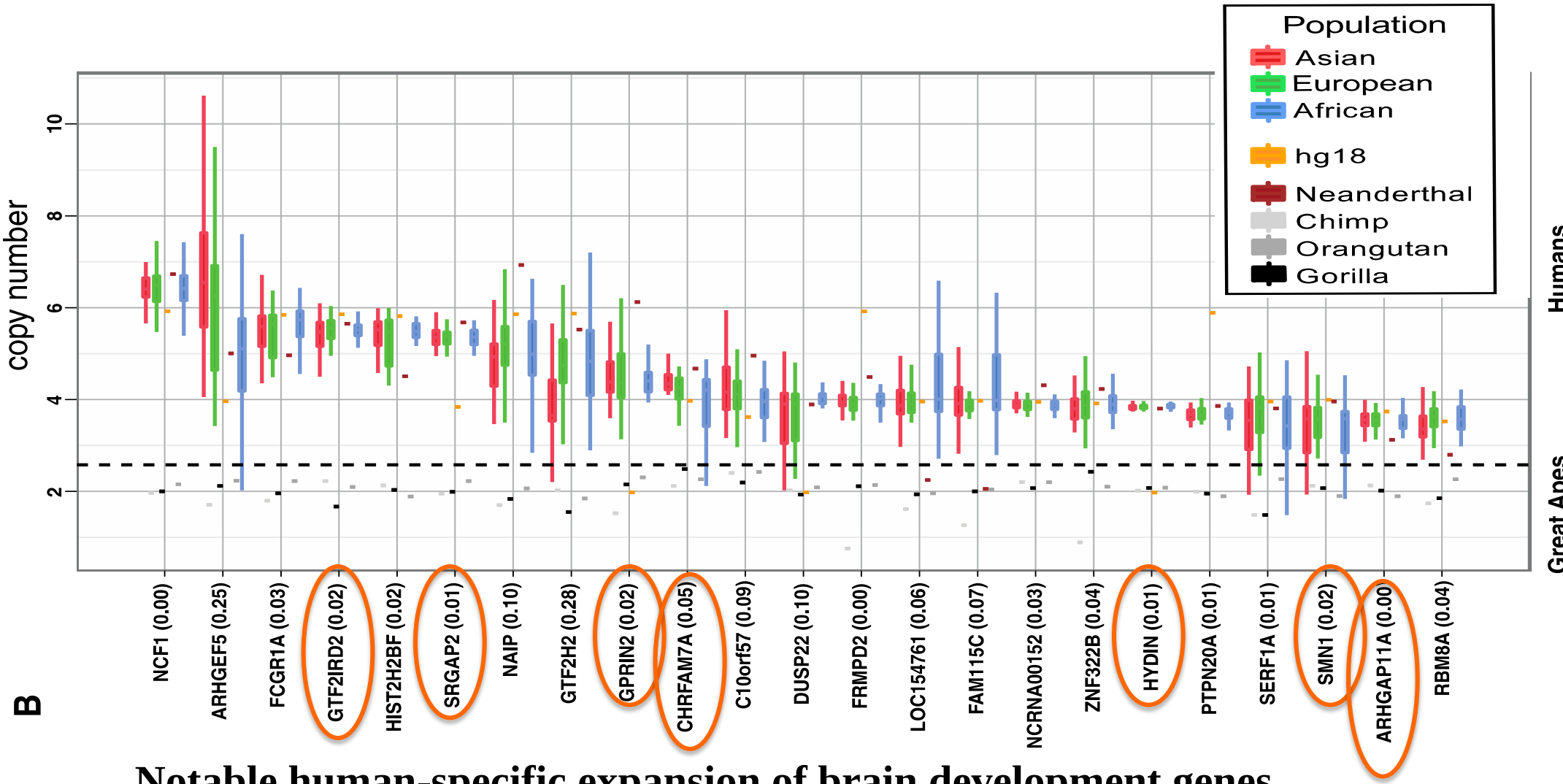
Features: No orthologs in mouse; multiple copies in chimp & human dramatic changes in expression profile; signatures of positive selection

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

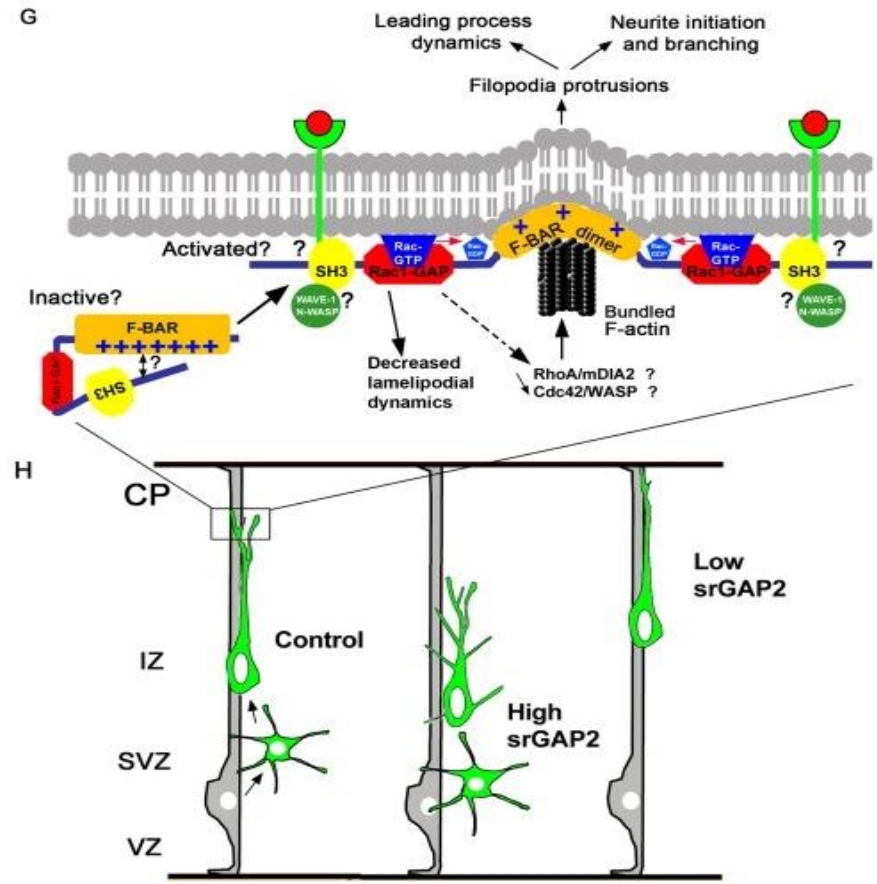


Notable human-specific expansion of brain development genes.
Neuronal cell death: p=5.7e-4; Neurological disease: p=4.6e-2

Sudmant et al., *Science*, 2010

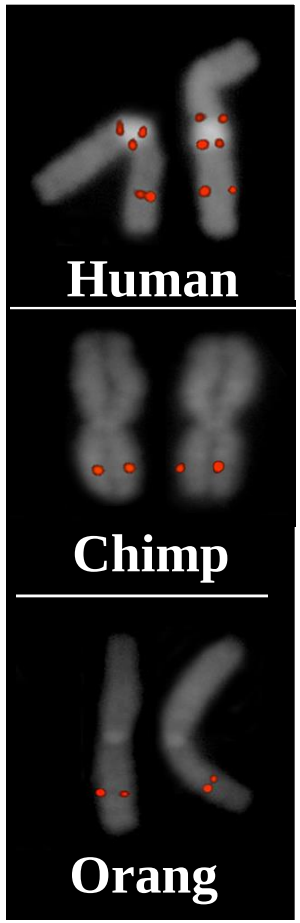
SRGAP2 function

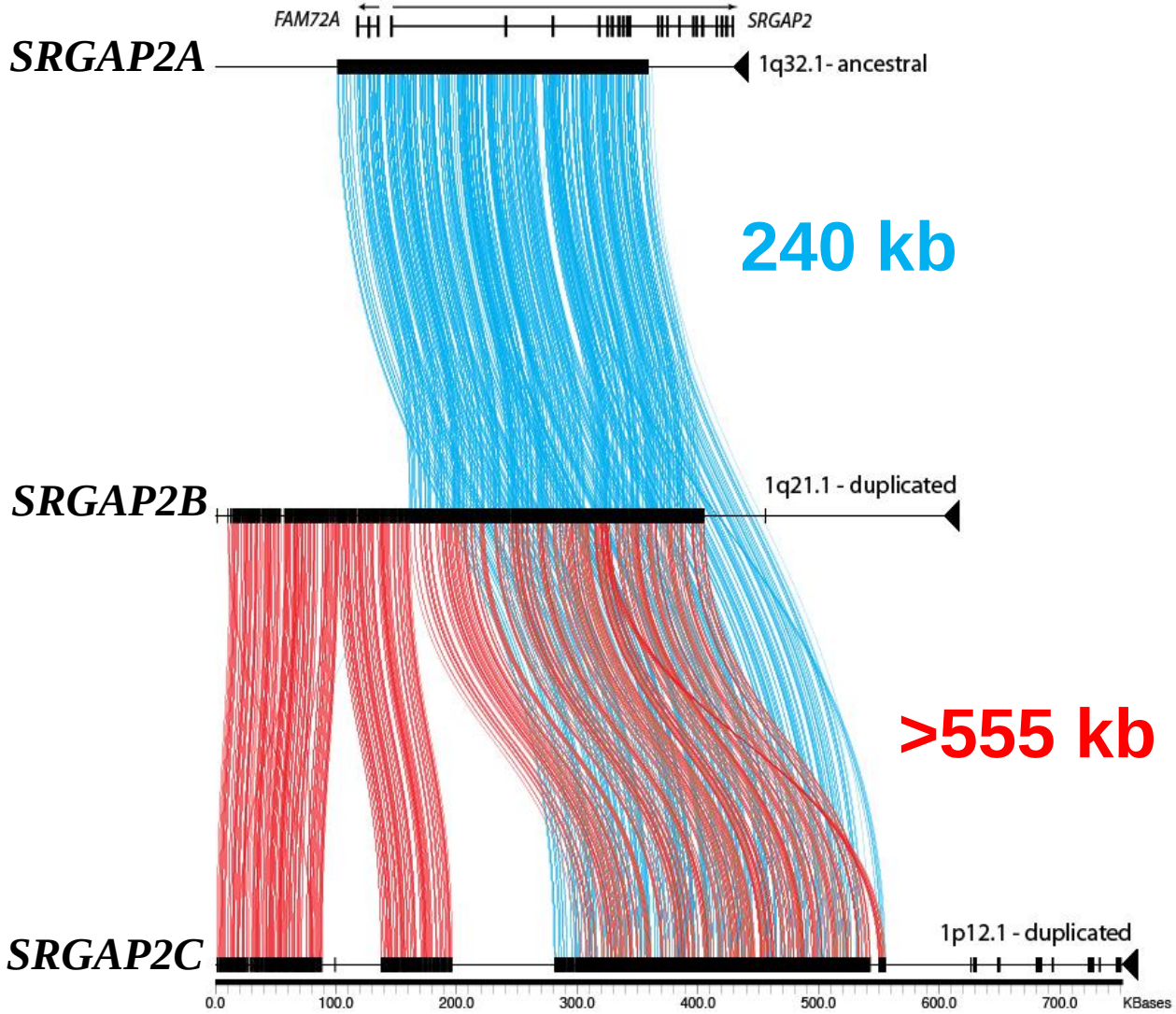
- *SRGAP2* (**SLIT**-ROBO Rho GTPase activating protein 2) functions to control migration of neurons and dendritic formation in the cortex
- Gene has been duplicated three times in human and no other mammalian lineage
- Duplicated loci not in human genome

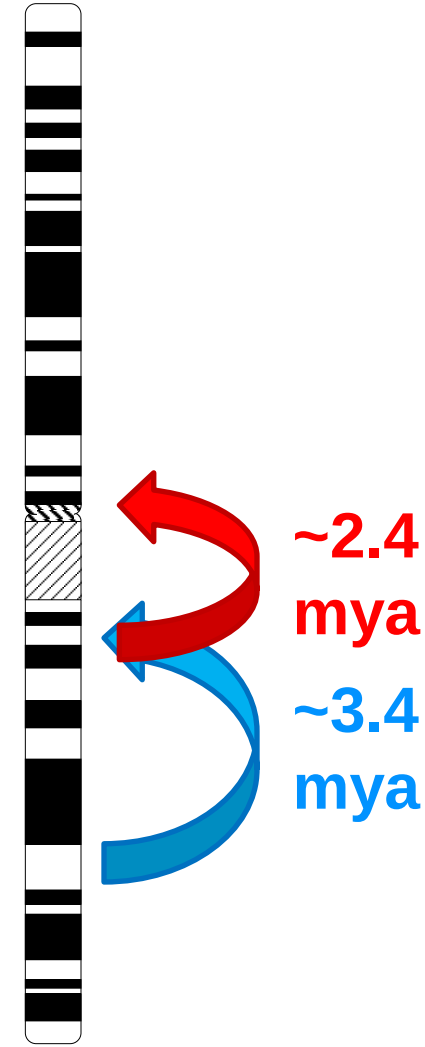


Guerrier et al., *Cell*, 2009

SRGAP2 Human Specific Duplication



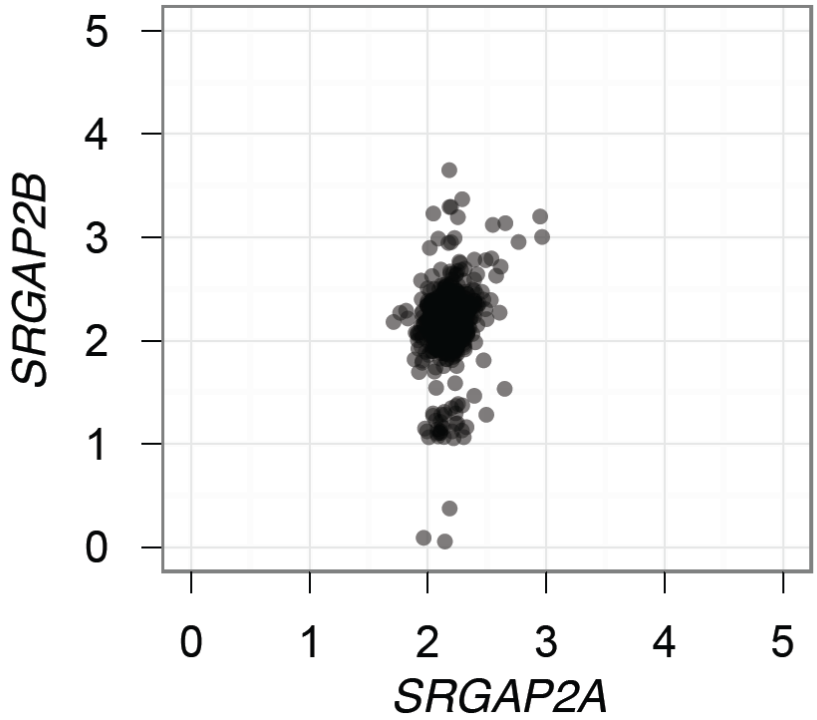


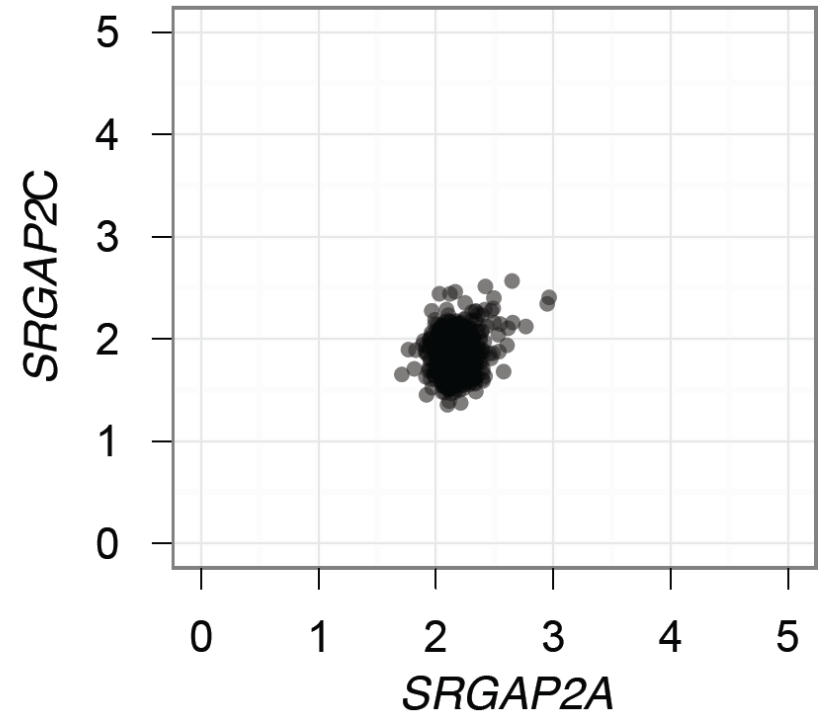


Dennis, Nuttle et al., *Cell*, 2012

SRGAP2C is fixed in humans

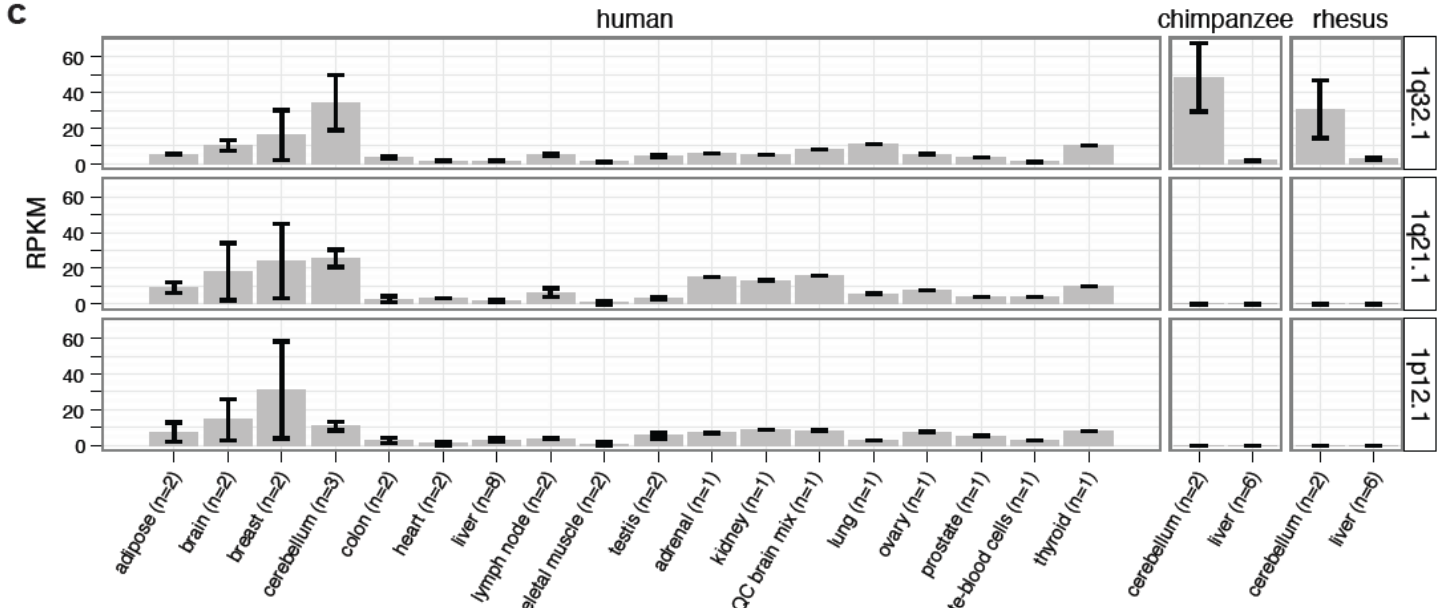
(n=661 individual genomes)



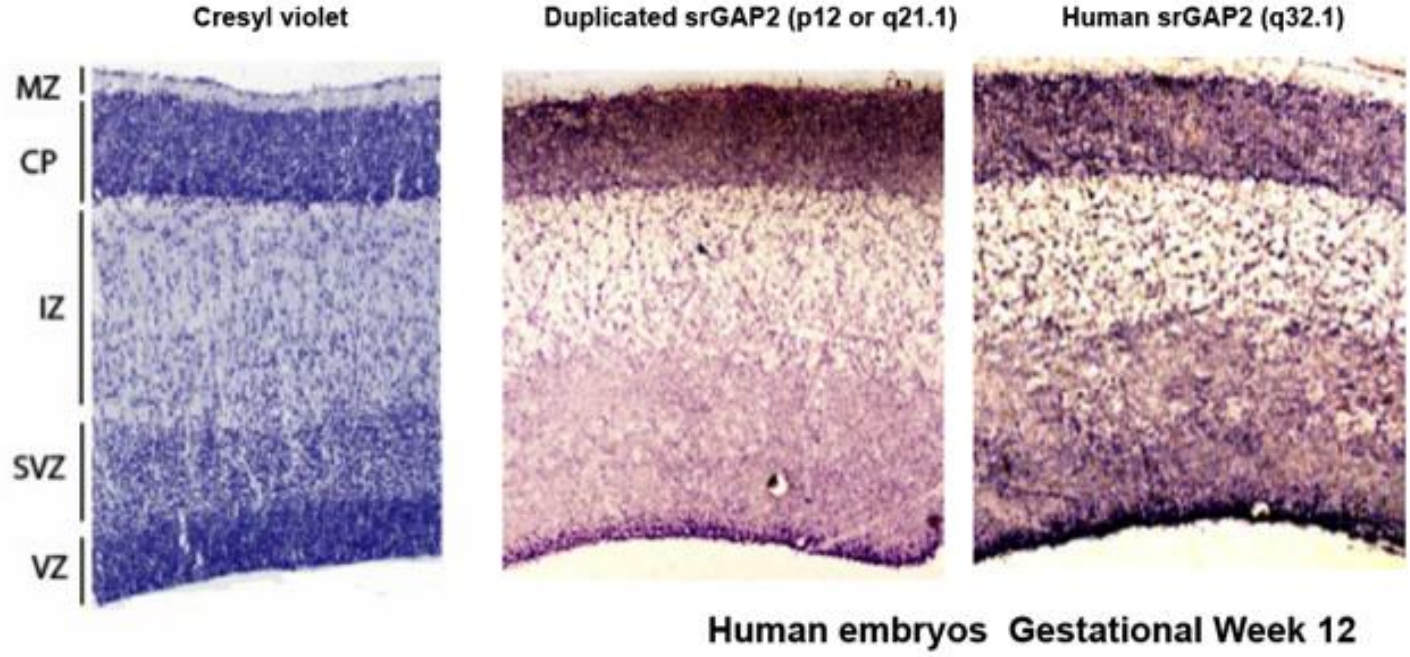


SRGAP2 duplicates are expressed

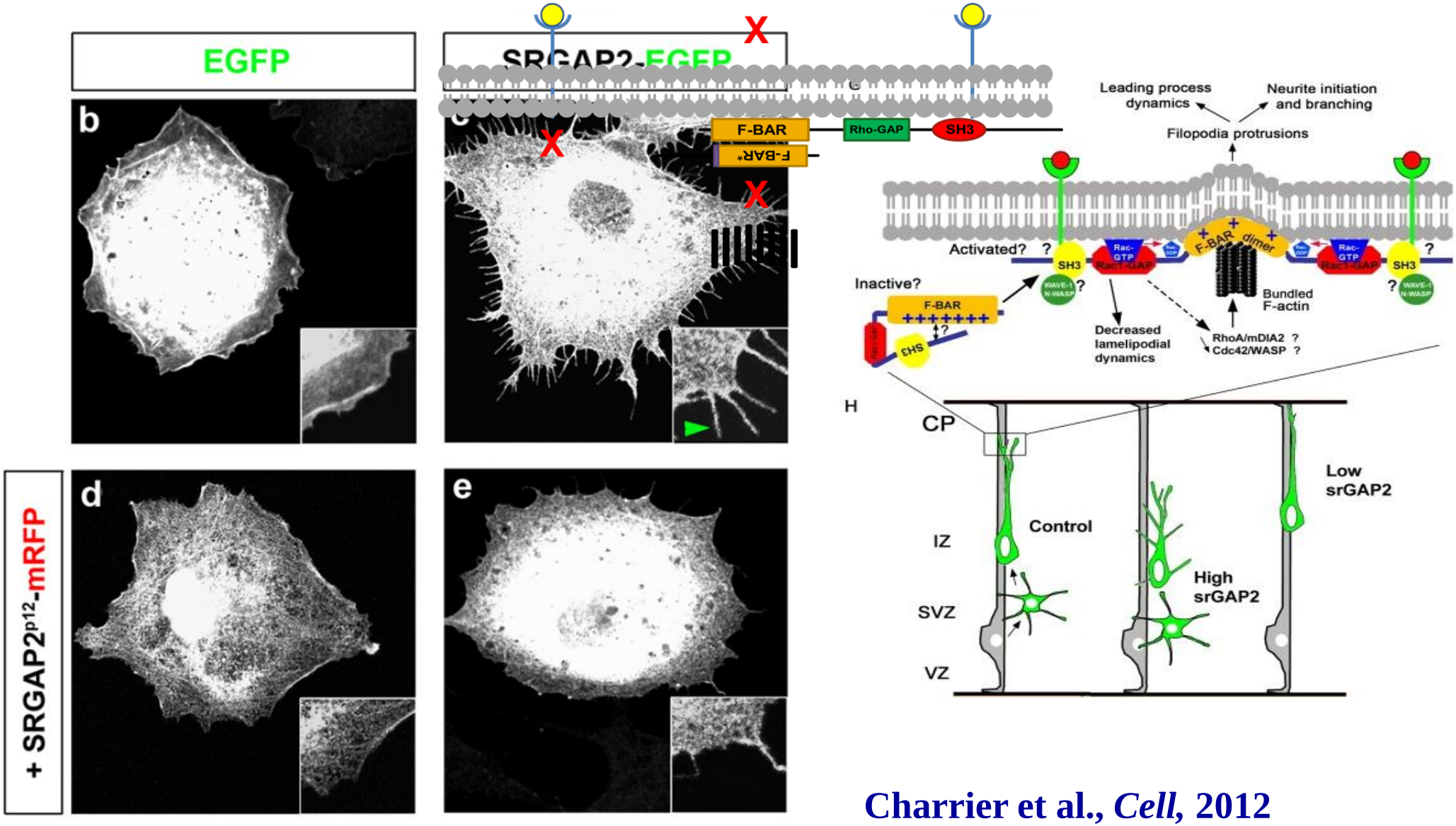
RNAseq



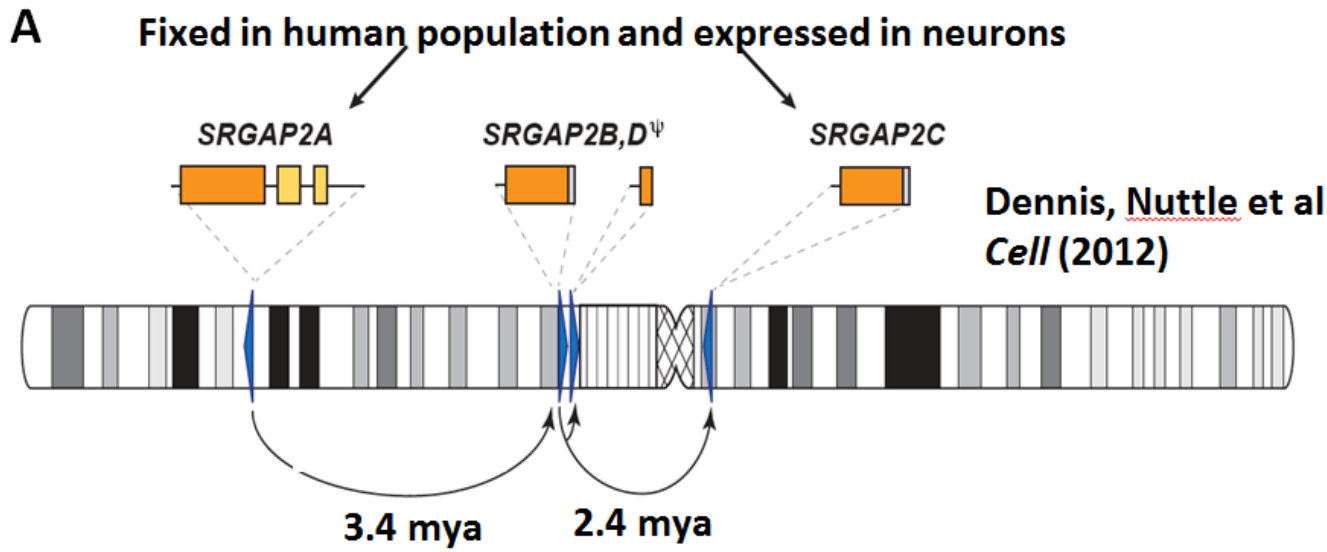
In situ



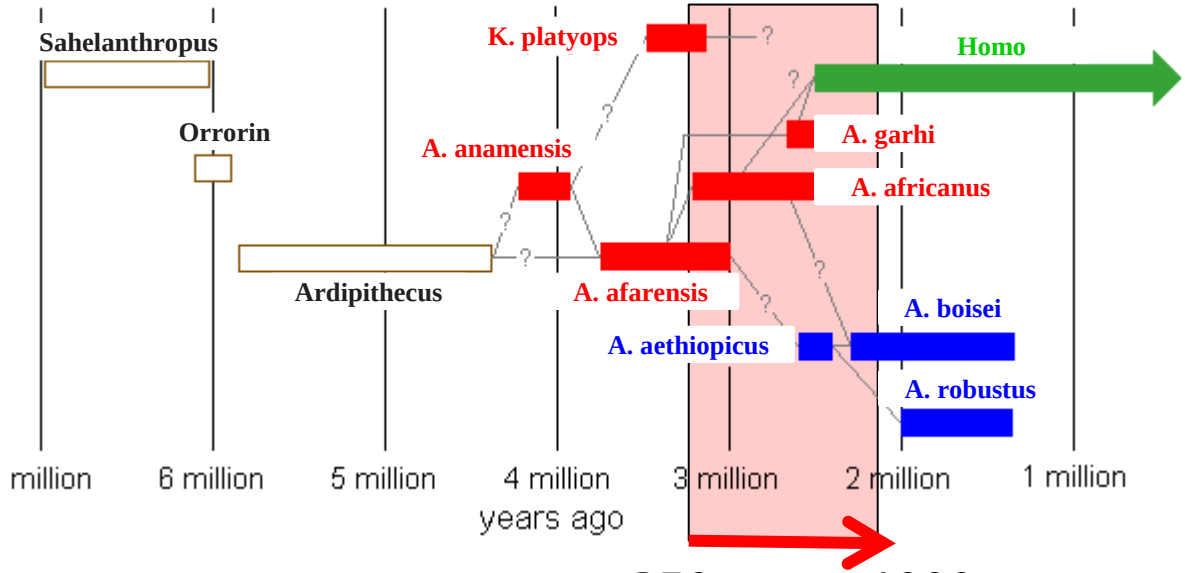
SRGAP2C duplicate antagonizes function



Charrier et al., *Cell*, 2012



Australopithecus

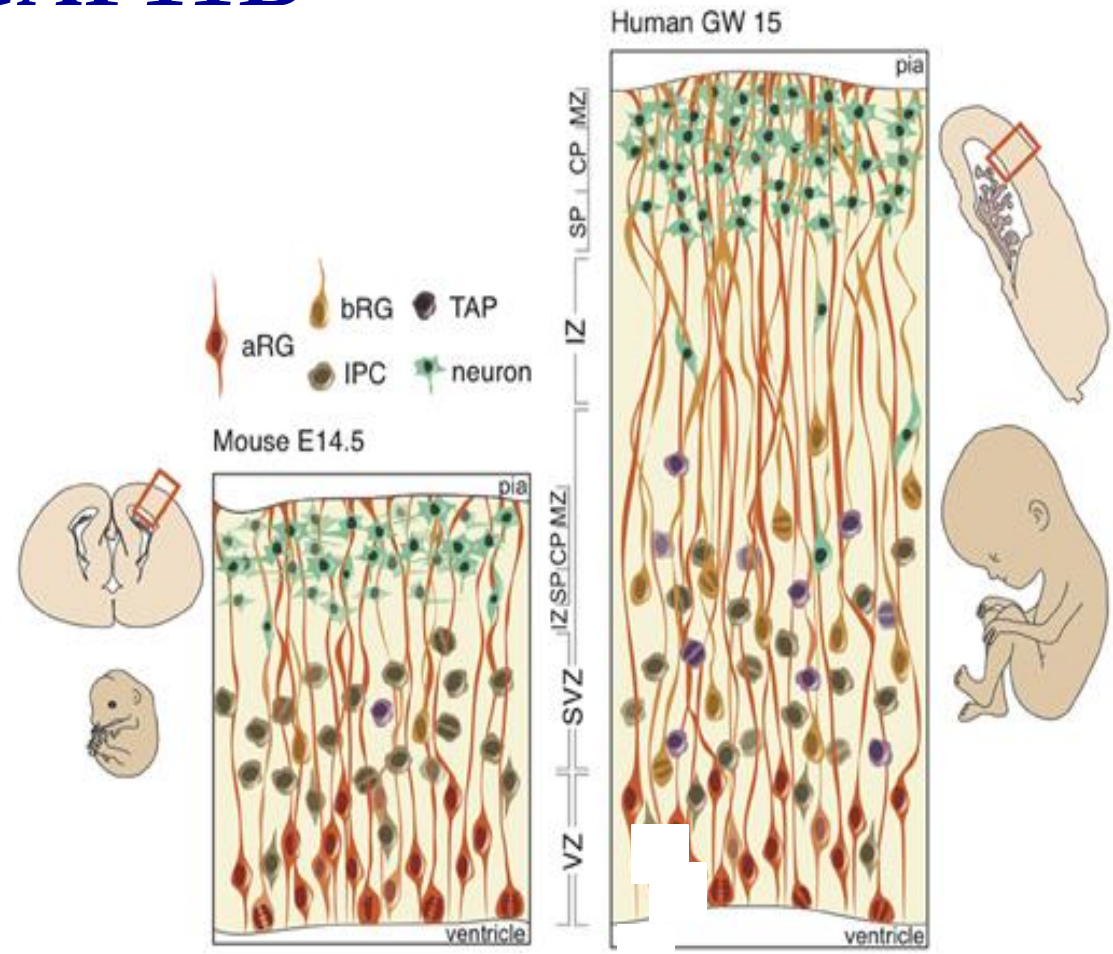


~350 cc **~1000 cc**

Homo habilis

Example 2: Human-specific Duplication of *ARHGAP11B*

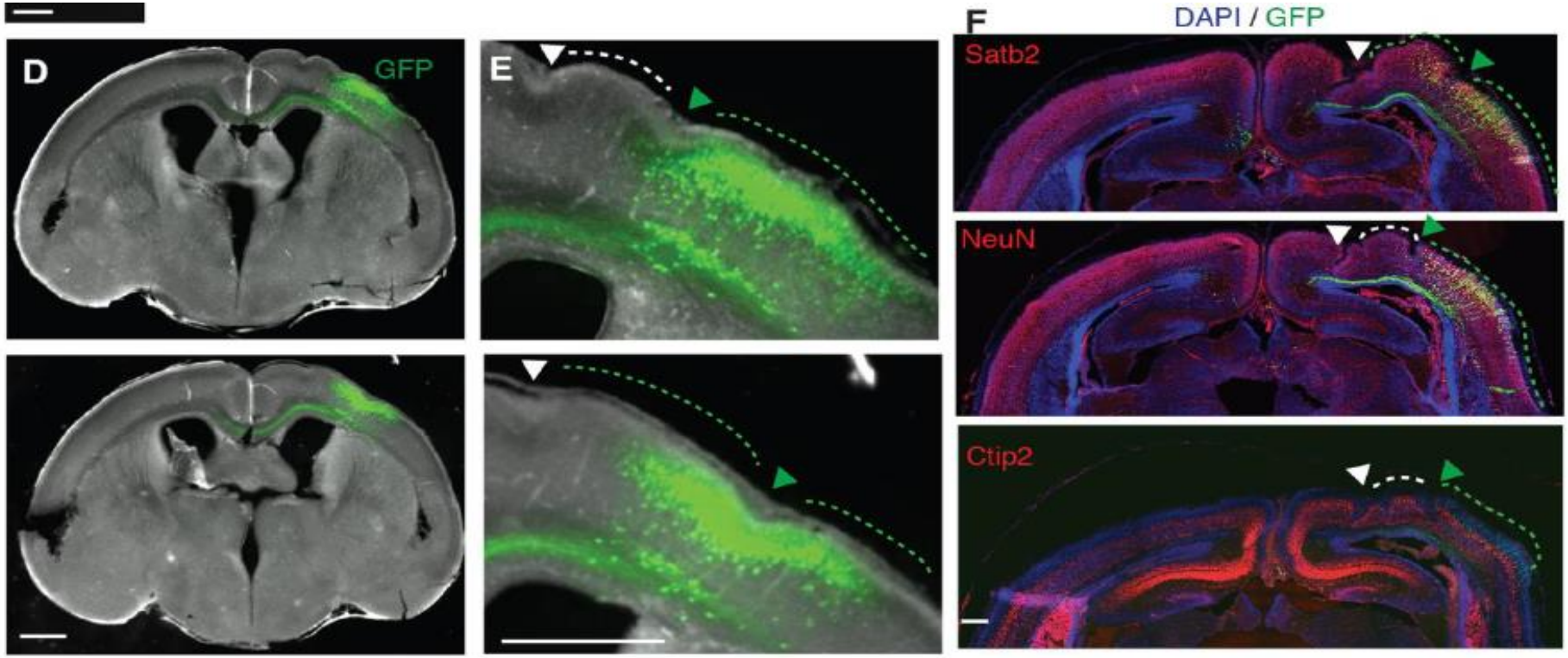
- A human-specific duplicated Rho GTPase activating protein that is truncated (5.3 mya)
- Predisposes to the most common cause of epilepsy
- Increase in number of basal radial glial hypothesized to lead to enlargement of the subventricular zone in humans.
- *ARHGAP11B* is expressed specifically in basal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

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**: Read-pair and read-depth methods to characterize SVs within genomes—need a high quality reference—not a solved problem.
- **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

Acronyms

SV-structural variation

CNV- copy number variation

CNP—copy number polymorphism

Indel-insertion/deletion event

SD—segmental duplication

SUN-singly-unique nucleotide identifier

SMRT-single-molecule real-time sequencing

WGS—whole genome shotgun sequencing

SV Software

- *Genomestrip*—Handsaker/McCarroll—combines read-depth and readpair data to identify potential sites of SV data from population genomic data
- *dCGH*—Sudmant/Eichler—measurre 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
- *VariationHunter*—Hormozdiari/Alkan—uses readpair & multiple mapping to discover SV
- *Lumpy* --Quinlan—uses probabilistic framework to integrate multiple structural variation signals such as discordant paired-end alignments and split-read alignments
- *PINDEL*—Kai Ye-- breakpoints of large deletions, medium sized insertions, inversions, tandem duplications and other structural variants at single-based resolution from next-gen sequence data. It uses a pattern growth approach to identify the breakpoints of these variants from paired-end short reads.
- *SMRT-SV*—Chaisson/Eichler—maps SMRT long reads (BLASR) to reference, detects signatures of SV and generates local assembly of SV

SD-Mediated Rearrangements