

# Genome Structural Variation

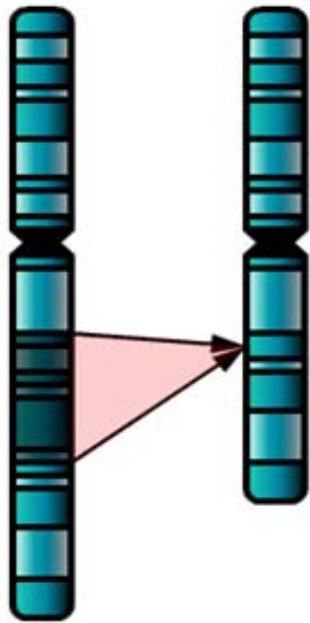
Evan Eichler

Howard Hughes Medical Institute

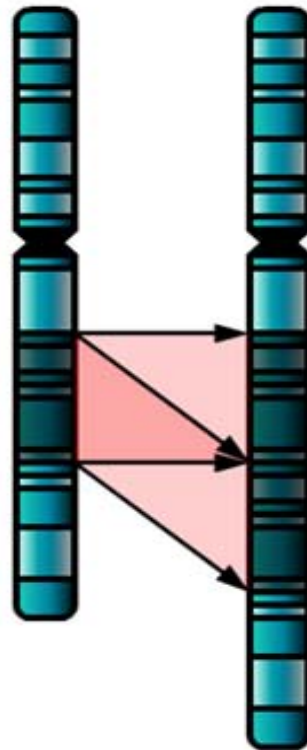
University of Washington

*January 17<sup>th</sup>, 2015, Genomics Workshop, Český Krumlov*

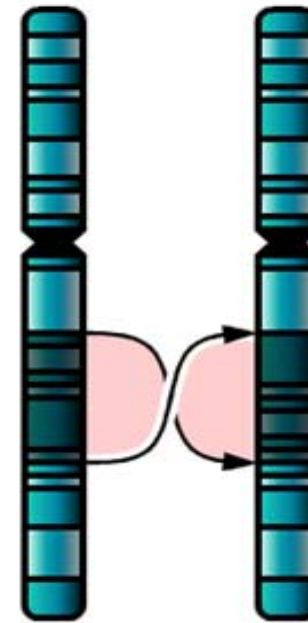
# Genome Structural Variation



**Deletion**



**Duplication**



**Inversion**

# Genetic Variation

## Types.

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

*Sequence*

*Cytogenetics*

# 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<sup>1</sup>, Rong Dong<sup>1\*</sup>, Timothy J. Vyse<sup>2\*</sup>, Penny J. Norsworthy<sup>1\*</sup>, Michelle D. Johnson<sup>1</sup>, Jennifer Smith<sup>3</sup>, Jonathan Mangion<sup>1</sup>, Cheri Robertson-Lowe<sup>1,2</sup>, Amy J. Marshall<sup>1</sup>, Enrico Petretto<sup>1</sup>, Matthew D. Hodges<sup>1</sup>, Gurjeet Bhangal<sup>3</sup>, Sheetal G. Patel<sup>1</sup>, Kelly Sheehan-Rooney<sup>1</sup>, Mark Duda<sup>1,3</sup>, Paul R. Cook<sup>1,3</sup>, David J. Evans<sup>3</sup>, Jan Domin<sup>3</sup>, Jonathan Flint<sup>4</sup>, Joseph J. Boyle<sup>5</sup>, Charles D. Pusey<sup>3</sup> & H. Terence Cook<sup>5</sup> [Nature](#) 2006

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

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

## Rare chromosomal deletions and duplications increase risk of schizophrenia

The International Schizophrenia Consortium\* [Nature](#) 455:237-41 2008

## Large recurrent microdeletions associated with schizophrenia

Hreinn Stefansson<sup>1\*</sup>, Dan Rujescu<sup>2\*</sup>, Sven Cichon<sup>3,4\*</sup>, Olli P. H. Pietiläinen<sup>5</sup>, Andres Ingason<sup>1</sup>, Stacy Steinberg<sup>1</sup>, Ragnheidur Fossdal<sup>1</sup>, Ennlhart Sturudsson<sup>6</sup>, Thorur Sturudsson<sup>6</sup>, Isachina F. Ruiz-Verdugo<sup>7</sup>

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

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

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

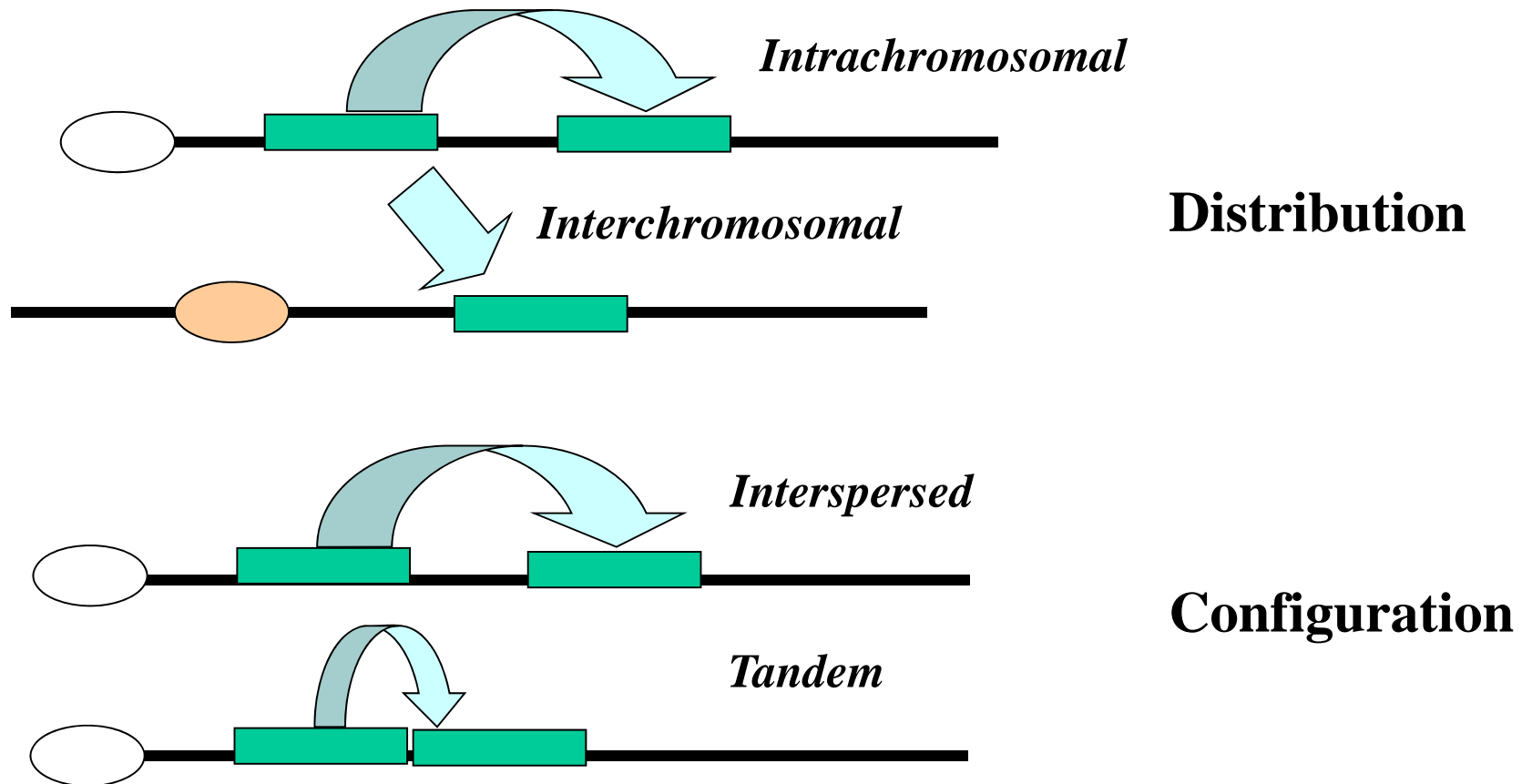
Lauren A. Weiss, Ph.D., Yiping Shen, Ph.D., Joshua M. Korn, B.S., Dan E. Arking, Ph.D., David T. Miller, M.D., Ph.D., Ragnheidur Fossdal, B.Sc., Evald Saemundsen, B.A., Hreinn Stefansson, Ph.D., Manuel A.R. Ferreira, Ph.D., Todd Green, B.S., Orah S. Platt, M.D., Douglas M. Ruderfer, M.S., Christopher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

## Strong Association of De Novo Copy Number Mutations with Autism

Jonathan Sebat,<sup>1\*</sup> B. Lakshmi,<sup>1</sup> Dheeraj Malhotra,<sup>1\*</sup> Jennifer Troge,<sup>1\*</sup> Christa Lese-Martin,<sup>2</sup> Tom Walsh,<sup>3</sup> Boris Yamrom,<sup>1</sup> Seungtae Yoon,<sup>1</sup> Alex Krasnitz,<sup>1</sup> Jude Kendall,<sup>1</sup> Anthony Leotta,<sup>1</sup> Deepa Pai,<sup>1</sup> Ray Zhang,<sup>1</sup> Yoon-Ha Lee,<sup>1</sup> James Hicks,<sup>1</sup> Sarah J. Spence,<sup>4</sup> Annette T. Lee,<sup>5</sup> Kaija Puura,<sup>6</sup> Terho Lehtimäki,<sup>7</sup> David Ledbetter,<sup>2</sup> Peter K. Gregersen,<sup>5</sup> Joel Bregman,<sup>8</sup> James S. Sutcliffe,<sup>9</sup> Vaidehi Jobanputra,<sup>10</sup> Wendy Chung,<sup>10</sup> Dorothy Warburton,<sup>10</sup> Mary-Claire King,<sup>3</sup> David Skuse,<sup>11</sup> Daniel H. Geschwind,<sup>12</sup> T. Conrad Gilliam,<sup>13</sup> Kenny Ye,<sup>14</sup> Michael Wigler<sup>1†</sup> [SCIENCE](#) VOL 316 20 APRIL 2007

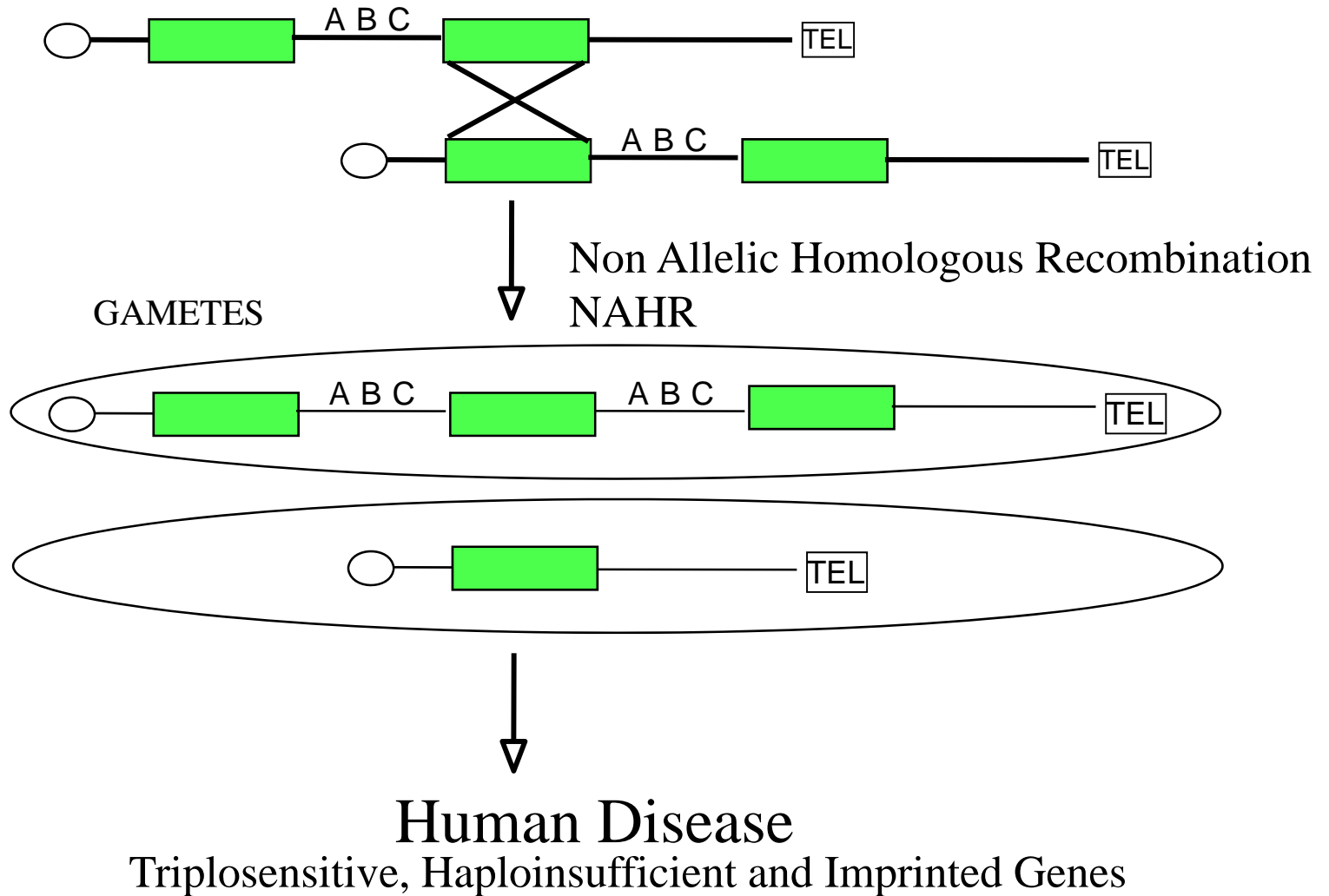
# Perspective: Segmental Duplications (SD)

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

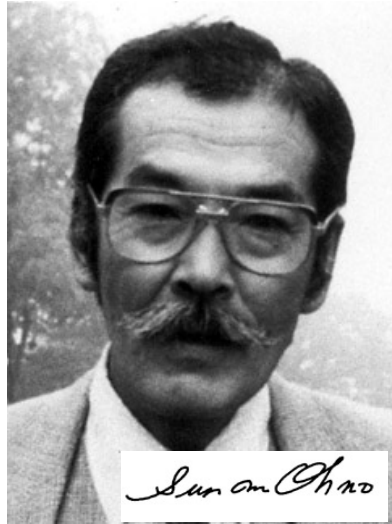


# Importance:

## SDs promote Structural Variation



# Importance: Evolution of New Gene Function



**GeneA**

Duplication

Acquire New/  
Modified Function

Mutation

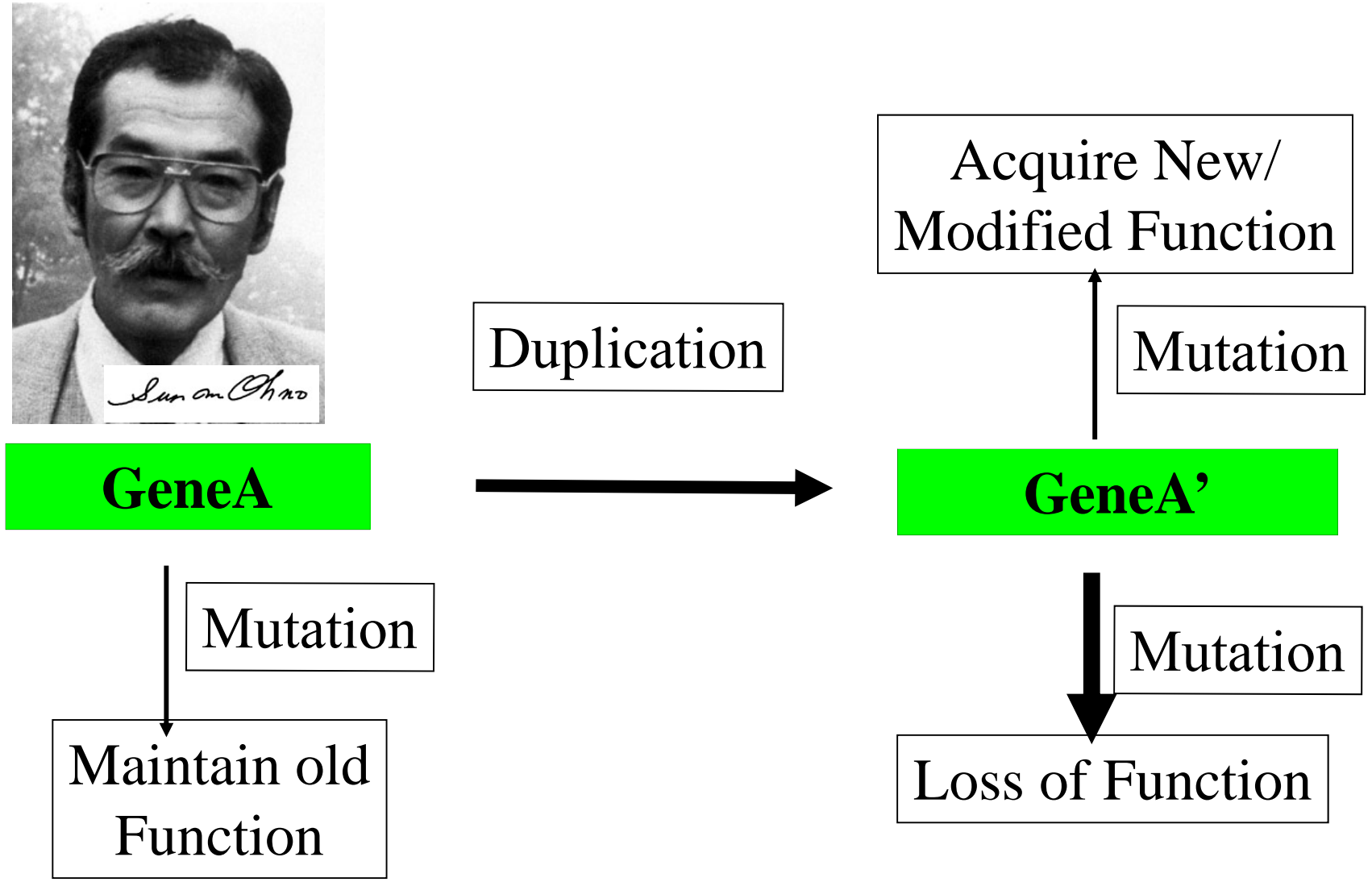
**GeneA'**

Mutation

Maintain old  
Function

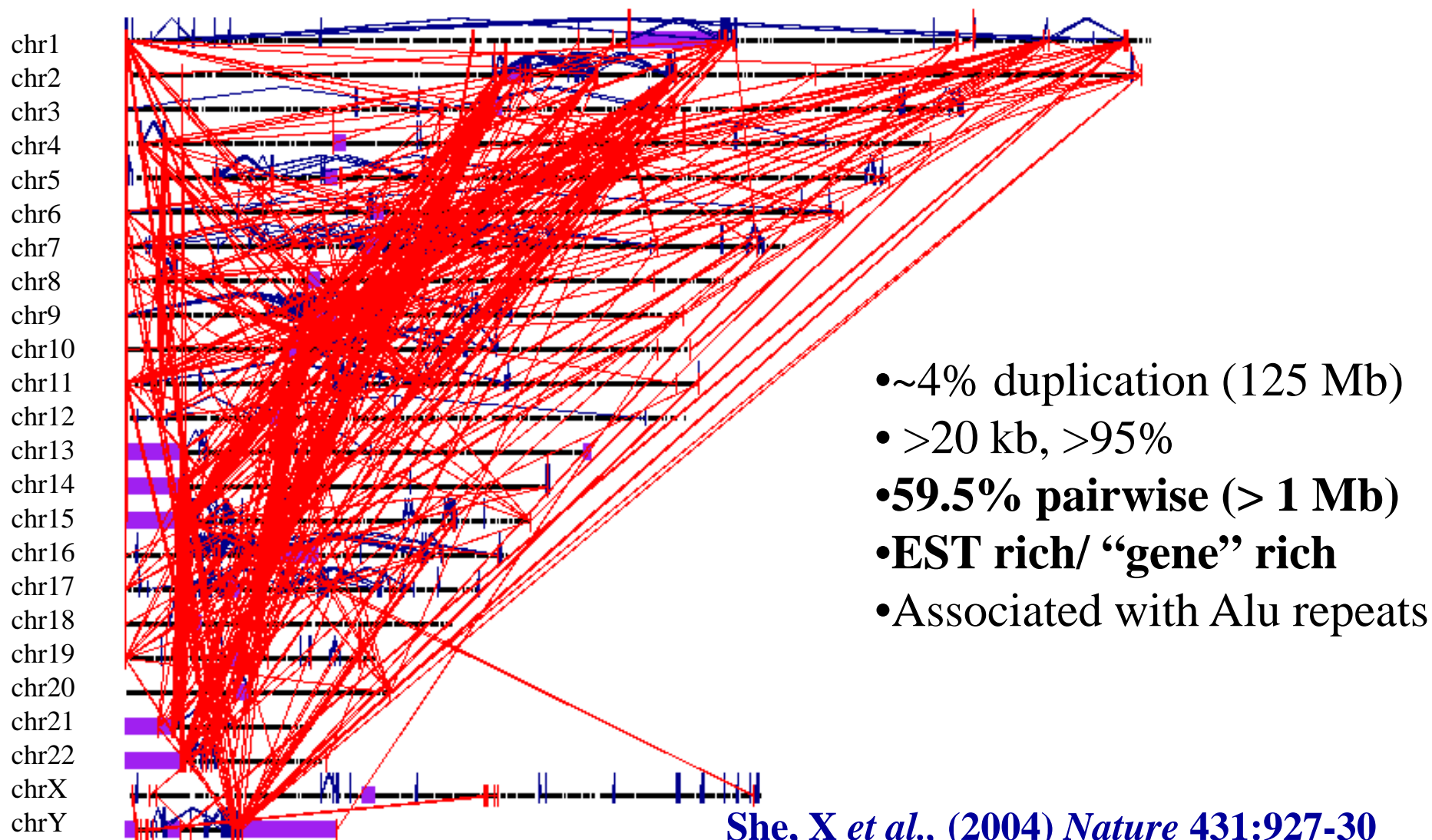
Mutation

Loss of Function



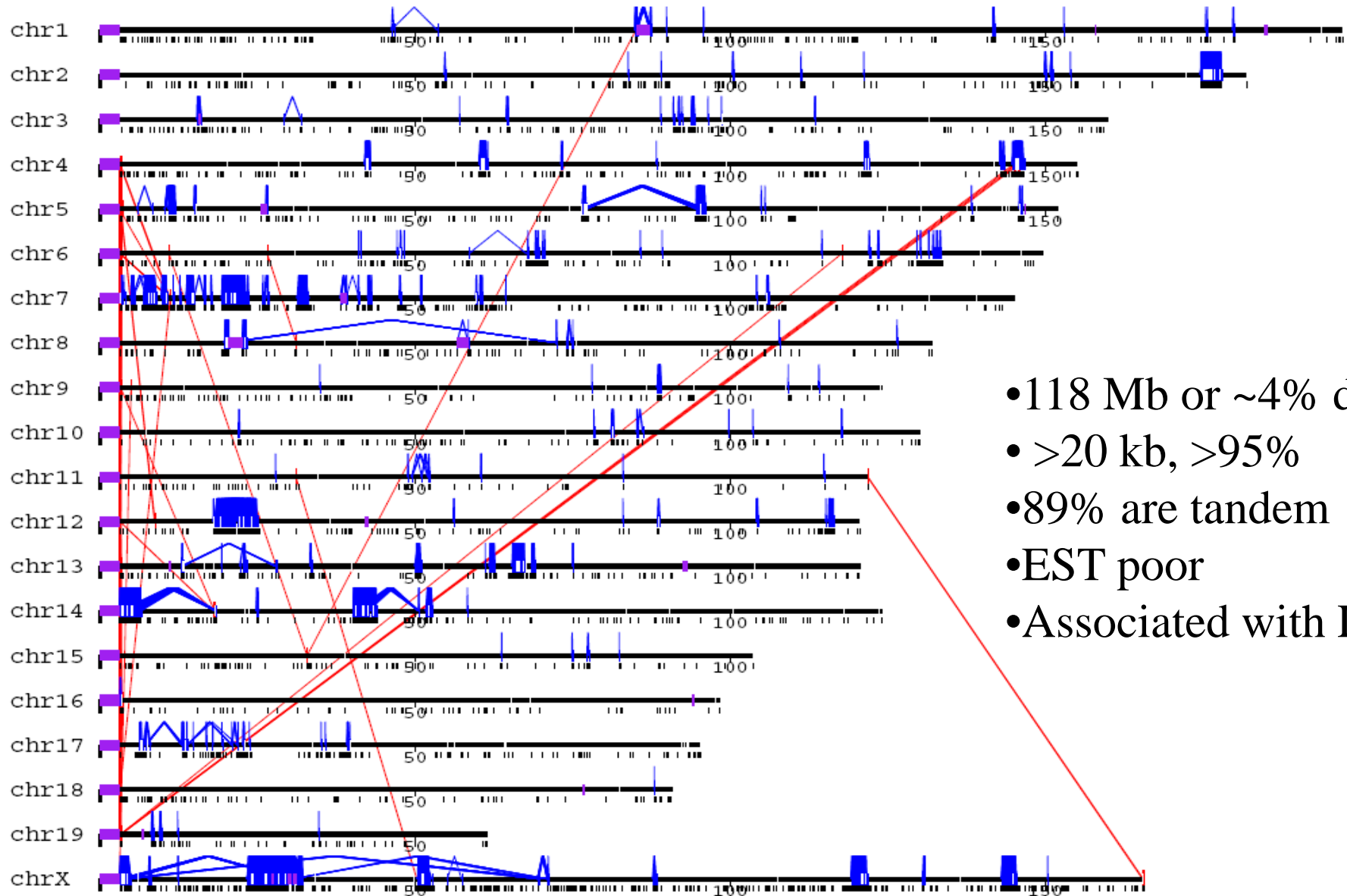


# I. Human Genome Segmental Duplication Pattern



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

# Mouse Segmental Duplication Pattern



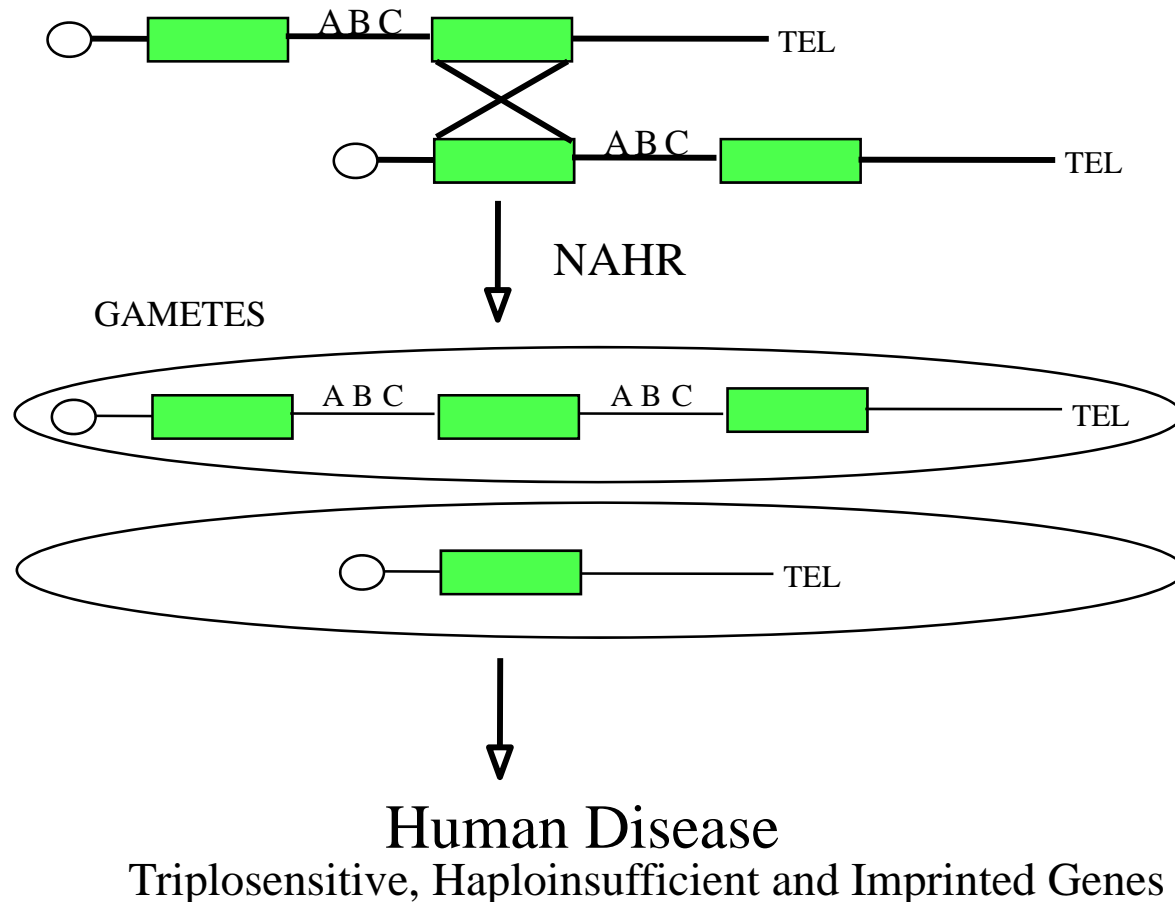
- 118 Mb or ~4% dup
- >20 kb, >95%
- 89% are tandem
- EST poor
- Associated with LINES

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

# Human Segmental Duplications Properties

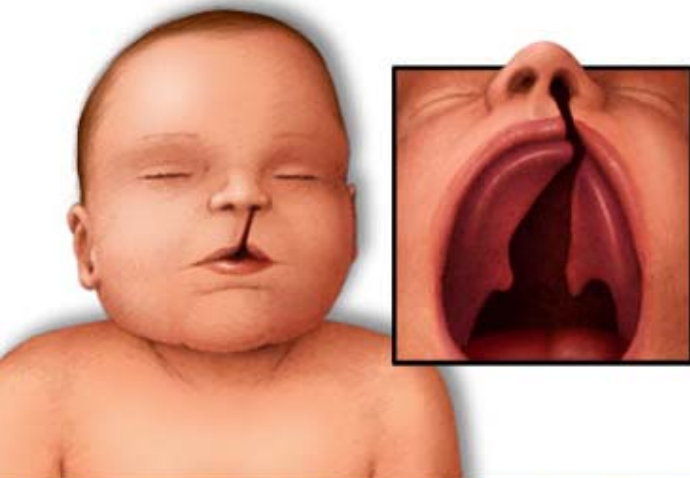
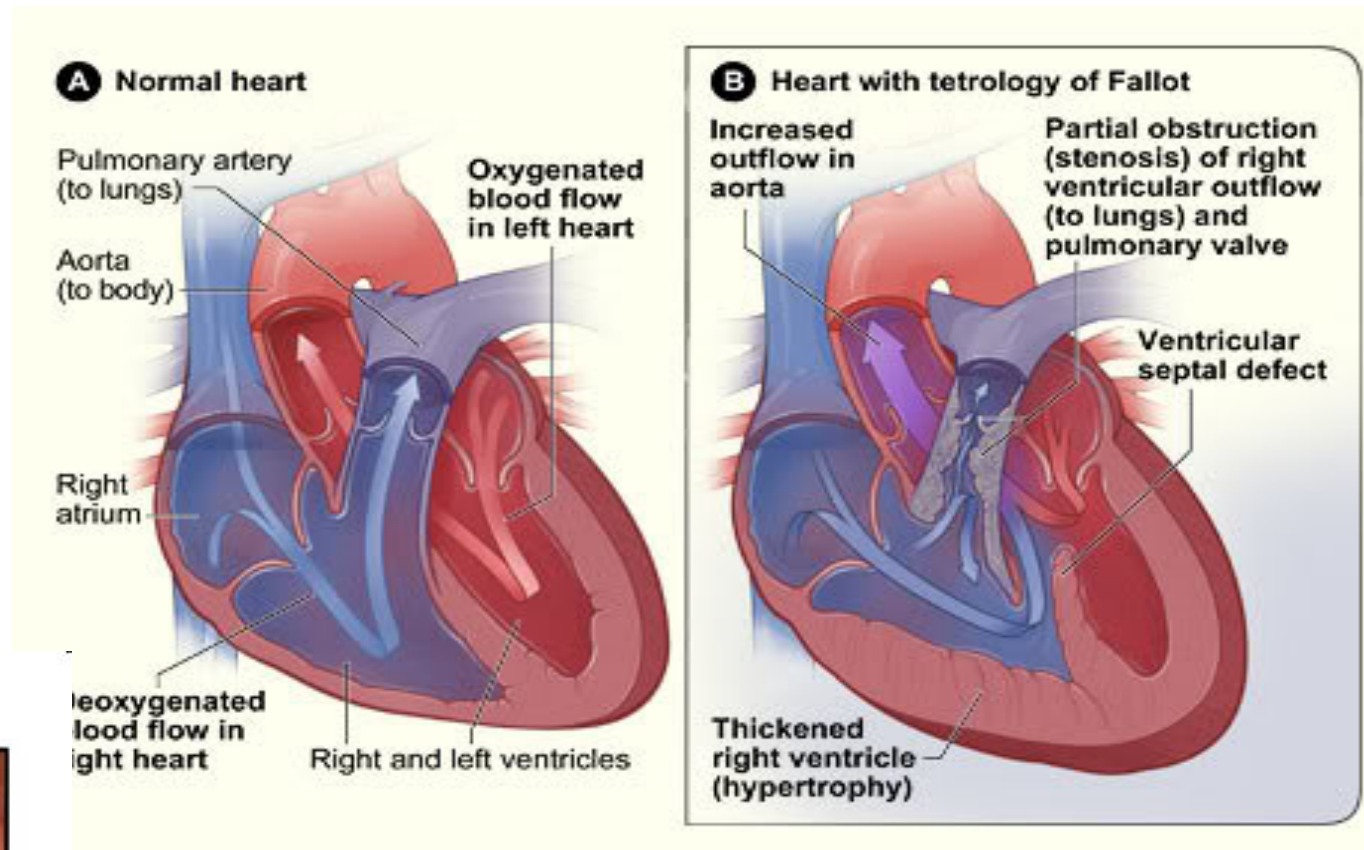
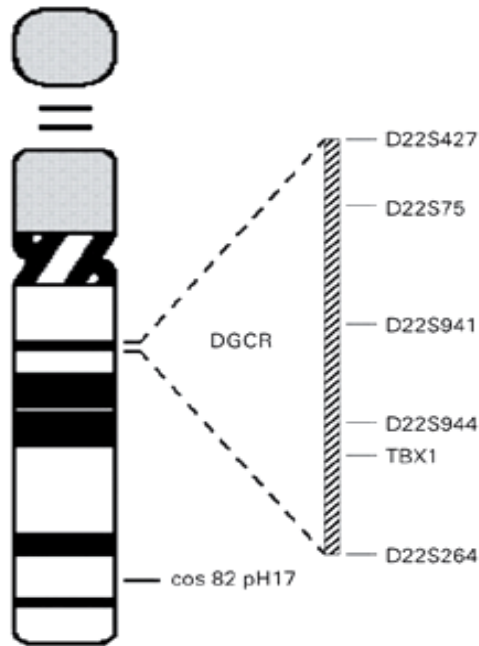
- Large (>10 kb)
- Recent (>95% identity)
- **Interspersed (60% are separated by more than 1 Mb)**
- Modular in organization
- Difficult to resolve

# Model #1: Rare Structural Variation



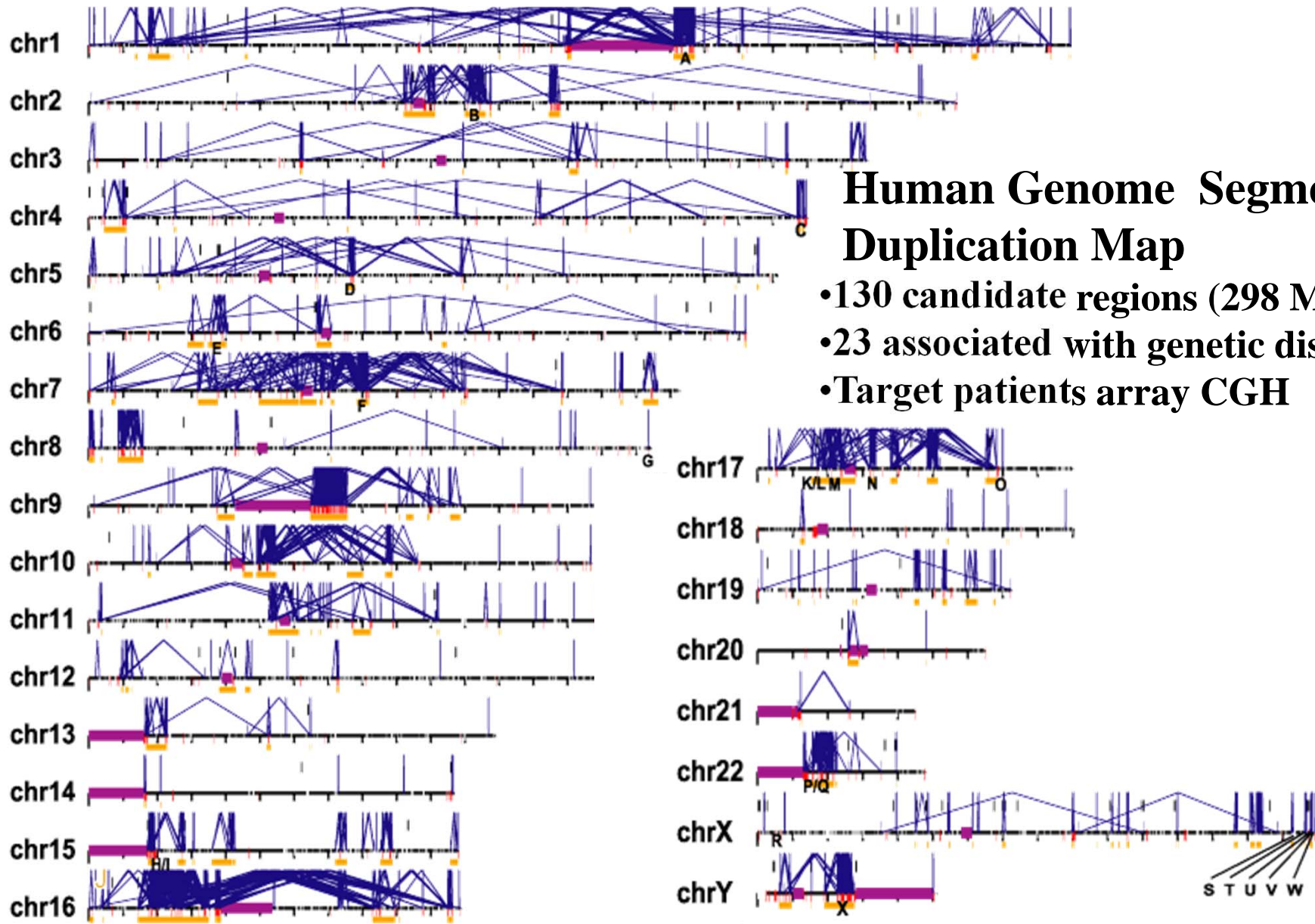
•**Genomic Disorders:** A group of diseases that results from genome rearrangement mediated mostly by non-allelic homologous recombination. (*Inoue & Lupski , 2002*).

# DiGeorge/VCF/22q11 Syndrome



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



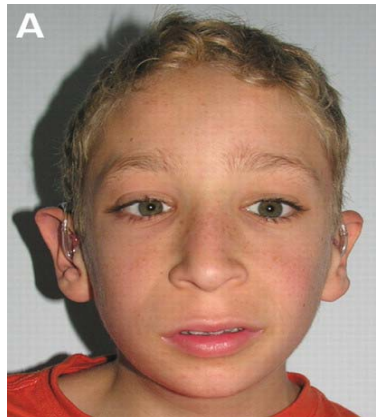
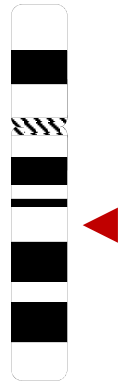


## Human Genome Segmental Duplication Map

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



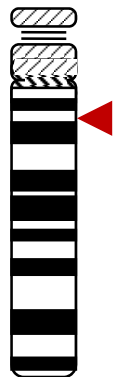
Chromosome 17



Chromosome 15

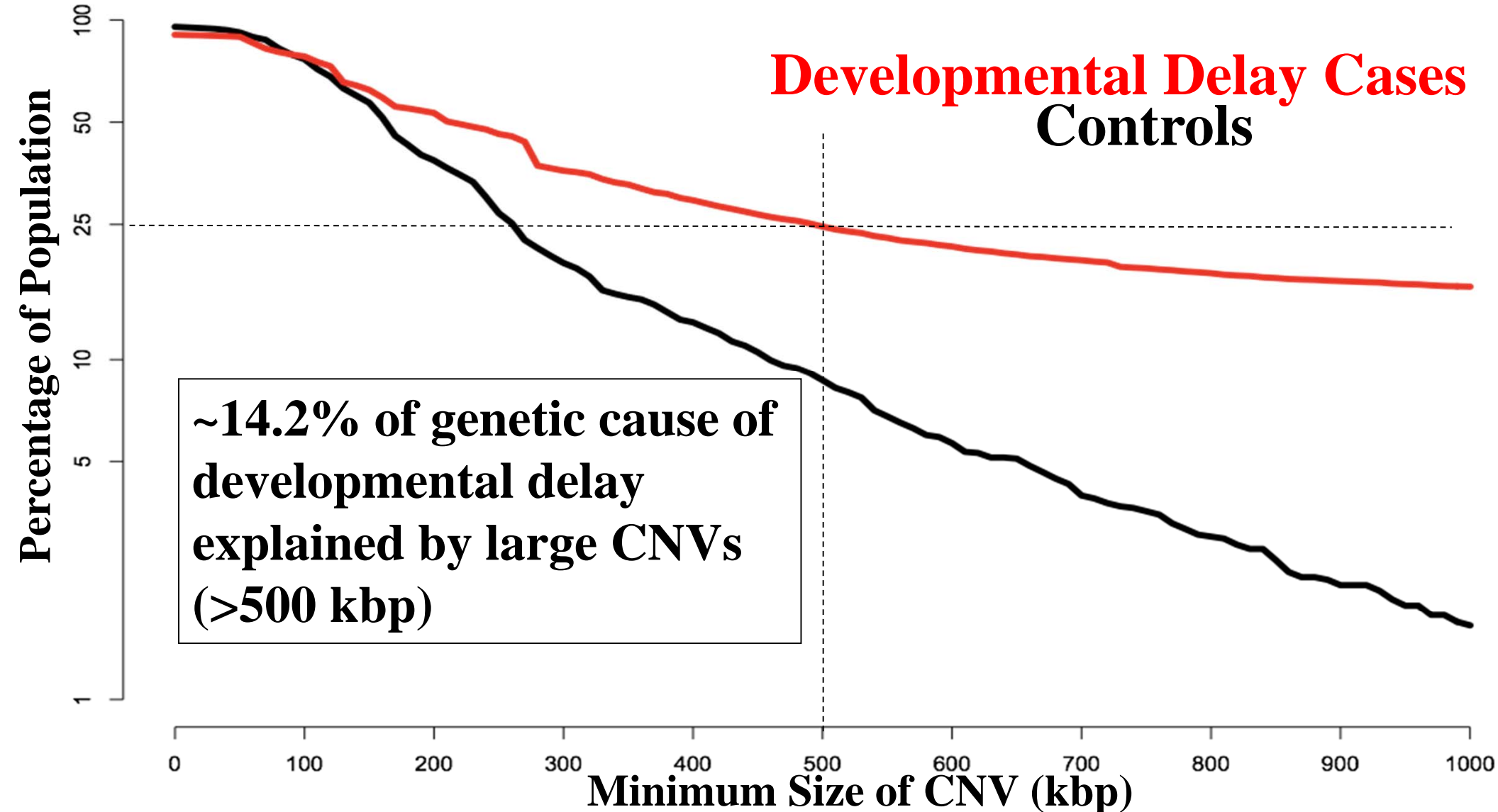


Chromosome 15



# Genome Wide CNV Burden

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



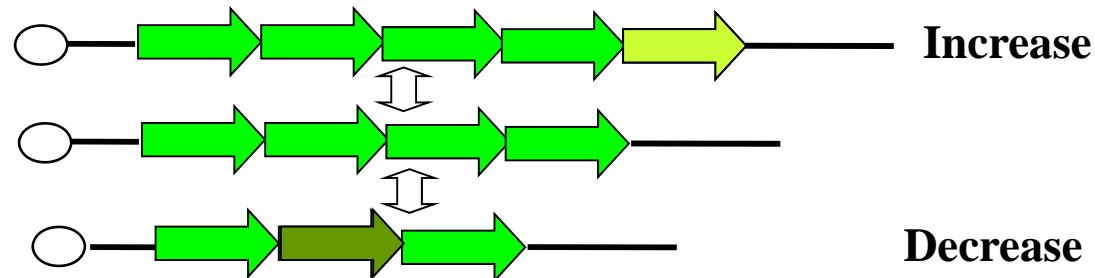
**Developmental Delay Cases**  
**Controls**

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



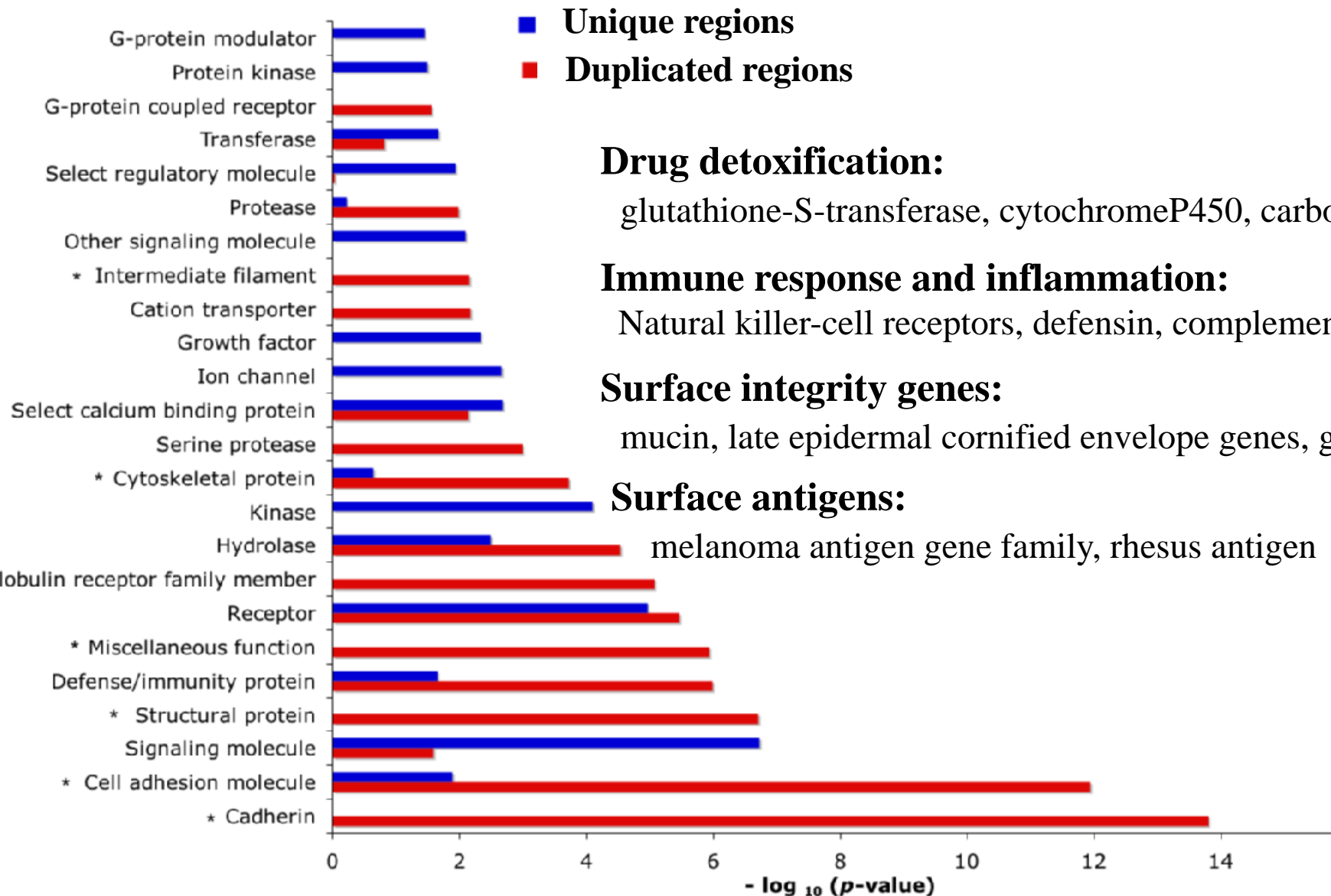
# Model #2: Copy Number Polymorphisms and Disease

Gene	Type	Locus	Seg. Dup	Phenotype
GSTT1	Decrease	22q11.2	54.3 kb	halothane/epoxide sensitivity
GSTM1	Decrease	1p13.3	18 kb	toxin resistance, cancer susceptibility
CYP2D6	Increase	22q13.1	5kb	antidepressant sensitivity
CYP21A2	Increase	6p21.3	35 kb	Congenital adrenal hyperplasia
LPA	Decrease	6q27	5.5*n kb	Coronary heart disease risk
RHD	Decrease	1p36.11	~60 kb	Rhesus blood group sensitivity
C4A/B	Decrease	6p21.33	32.8 kb	Lupus (SLE)
DEFB4	Decrease	8p23.1	~310 kb	Crohn Disease
DEFB4	Increase	8p23.1	~310 kb	Psoriasis



- **Multicopy or multiallelic CNPs associated with SDs**

# Structural Variation and Enriched Gene Functions



## Drug detoxification:

glutathione-S-transferase, cytochromeP450, carboxylesterases

## Immune response and inflammation:

Natural killer-cell receptors, defensin, complement factors

## Surface integrity genes:

mucin, late epidermal cornified envelope genes, galectin

## Surface antigens:

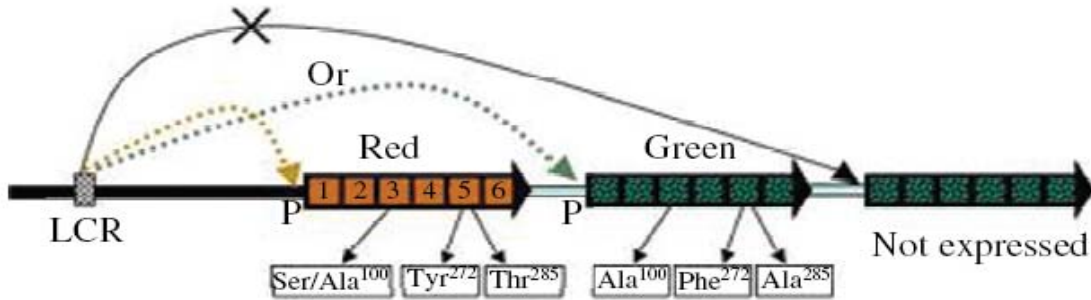
melanoma antigen gene family, rhesus antigen

• **Environmental interaction and cell-cell signaling molecules enriched**

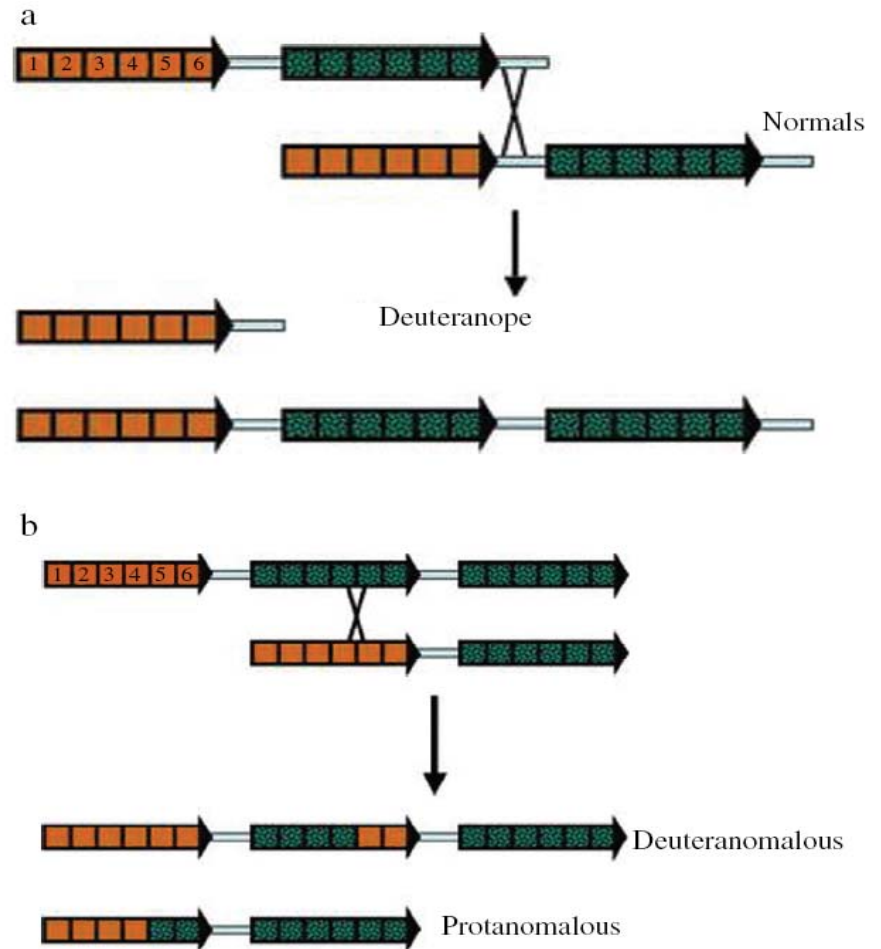
*Cooper et al., 2007*

# Copy-Number Detection is not Sufficient!

## Color-Blindness in Humans: The Opsin Loci



- Normal phenotypic variation
- Red-green color vision defects, X-linked
- 8% of males and 0.5% females. NEur.

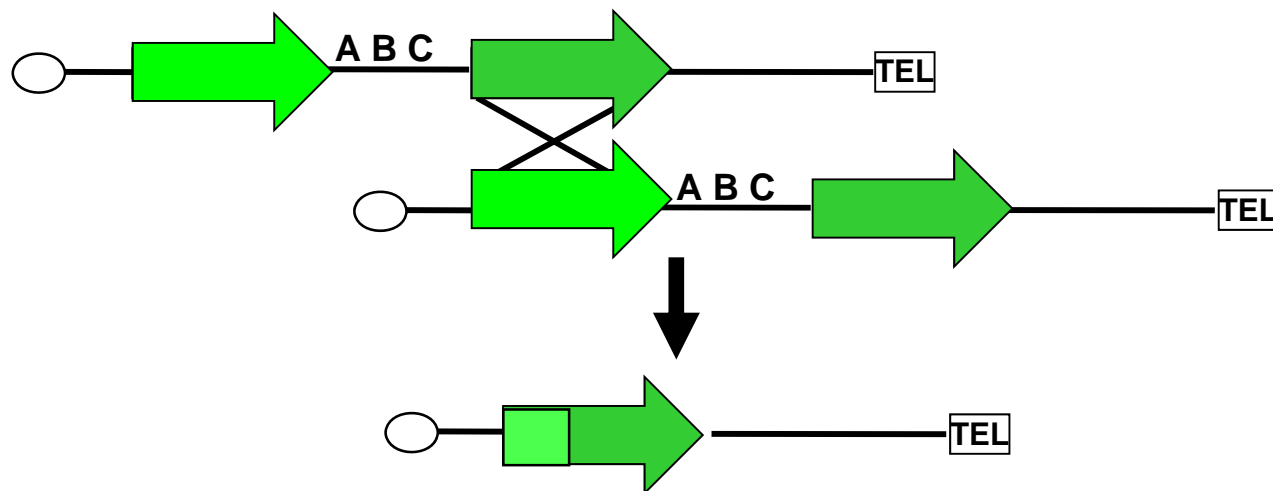
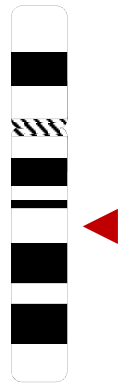


Deeb, SS, Clin. Genet, 2005

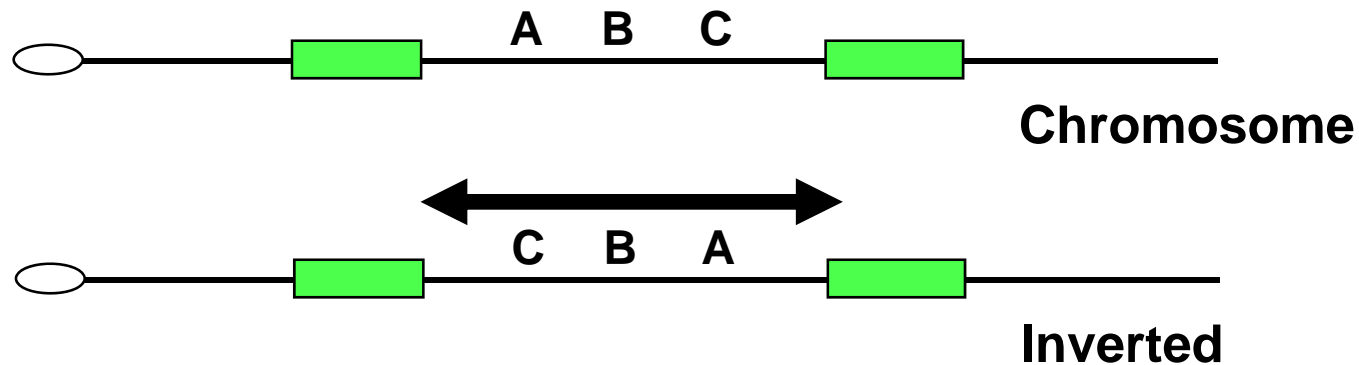
# Common and Rare Structural Variation are Linked 17q21.31 Deletion Syndrome



Chromosome 17



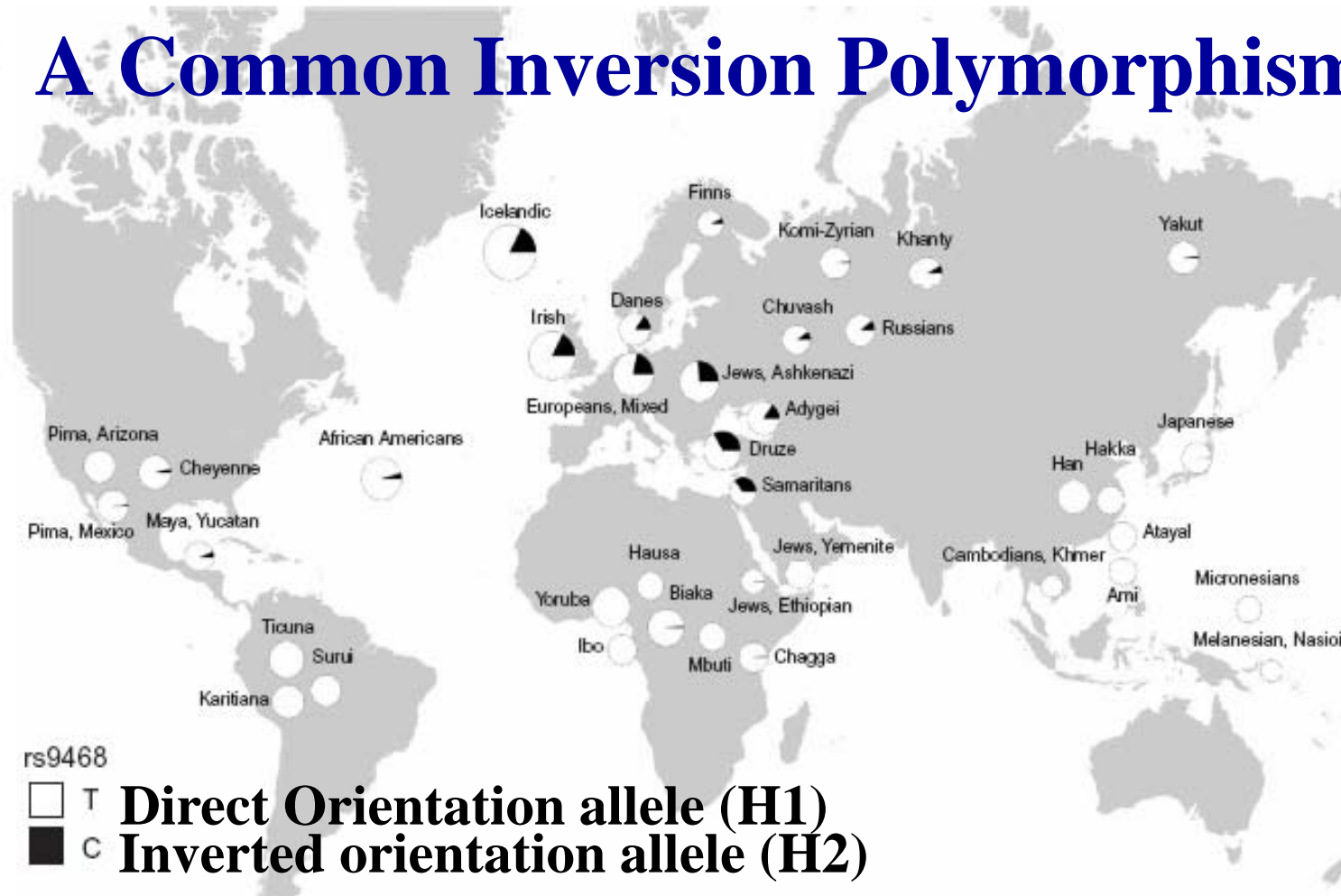
# 17q21.31 Inversion



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

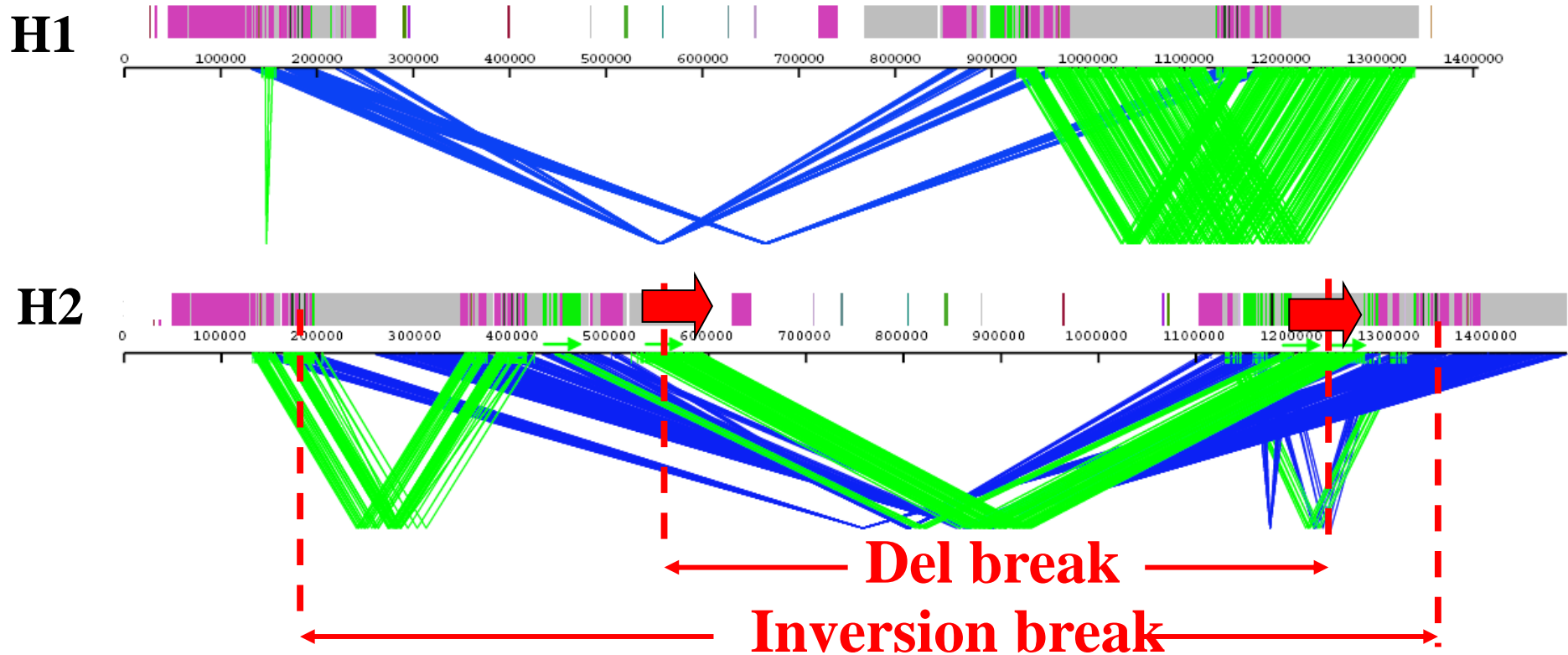
*Stefansson, K et al., (2005) Nature Genetics*

## b A Common Inversion Polymorphism



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

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



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

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

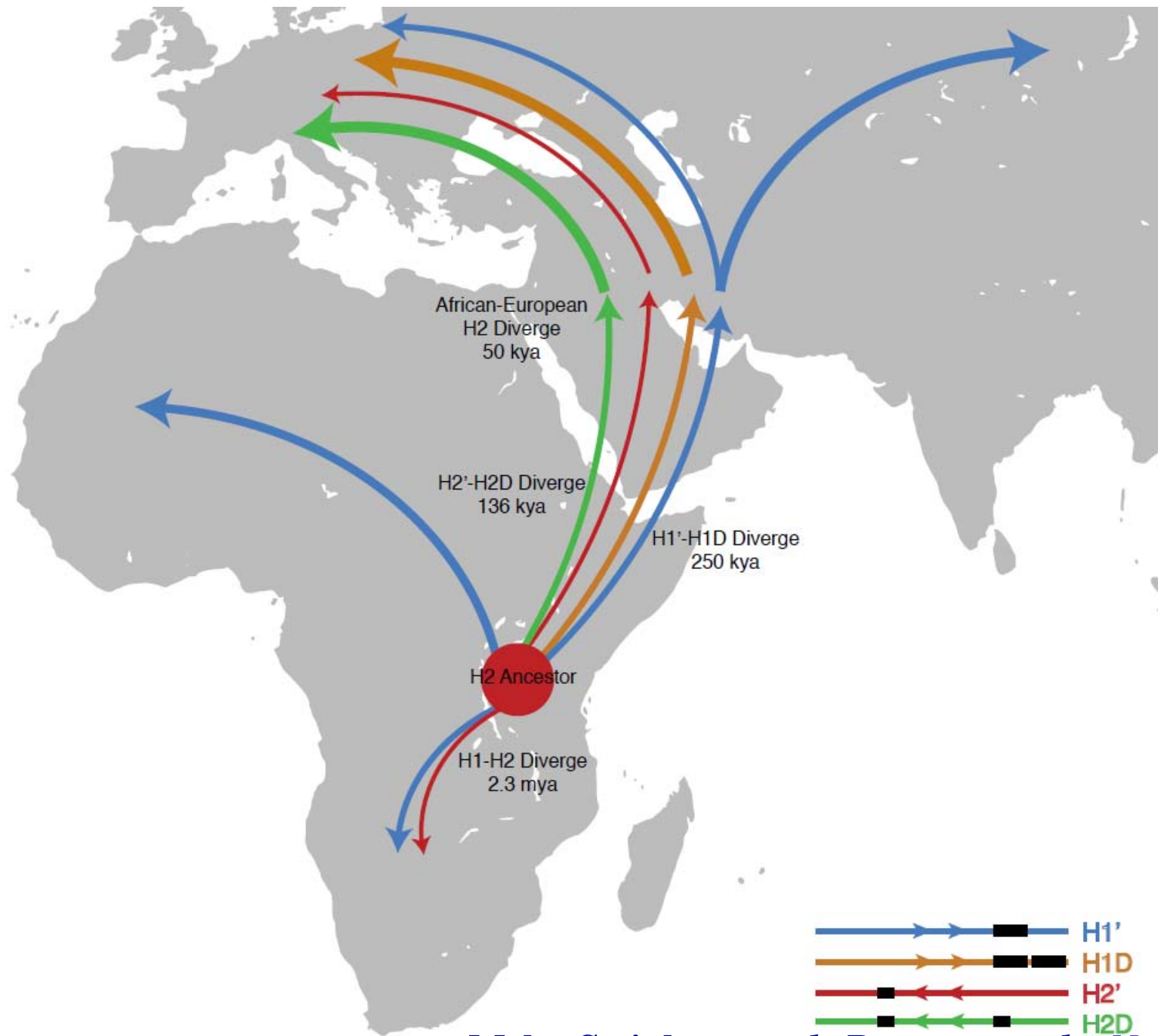


# Structural Variation Diversity

## 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

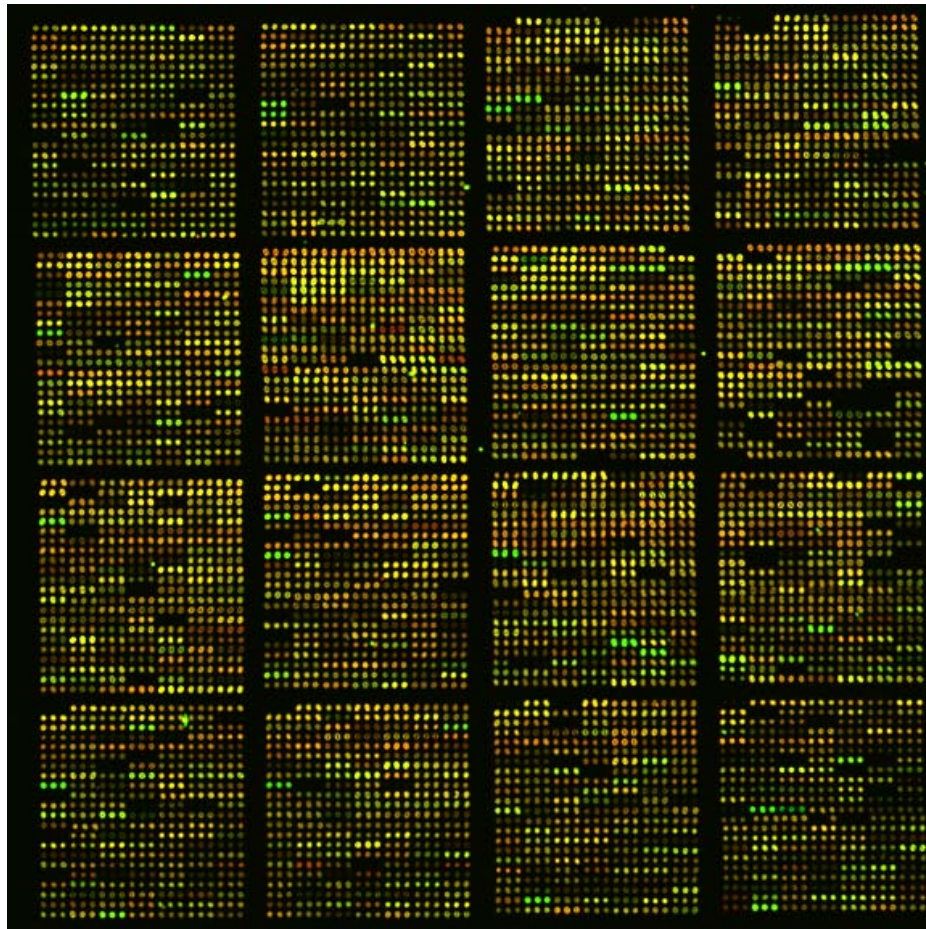
### Single molecule mapping

- **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: Korbelt *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

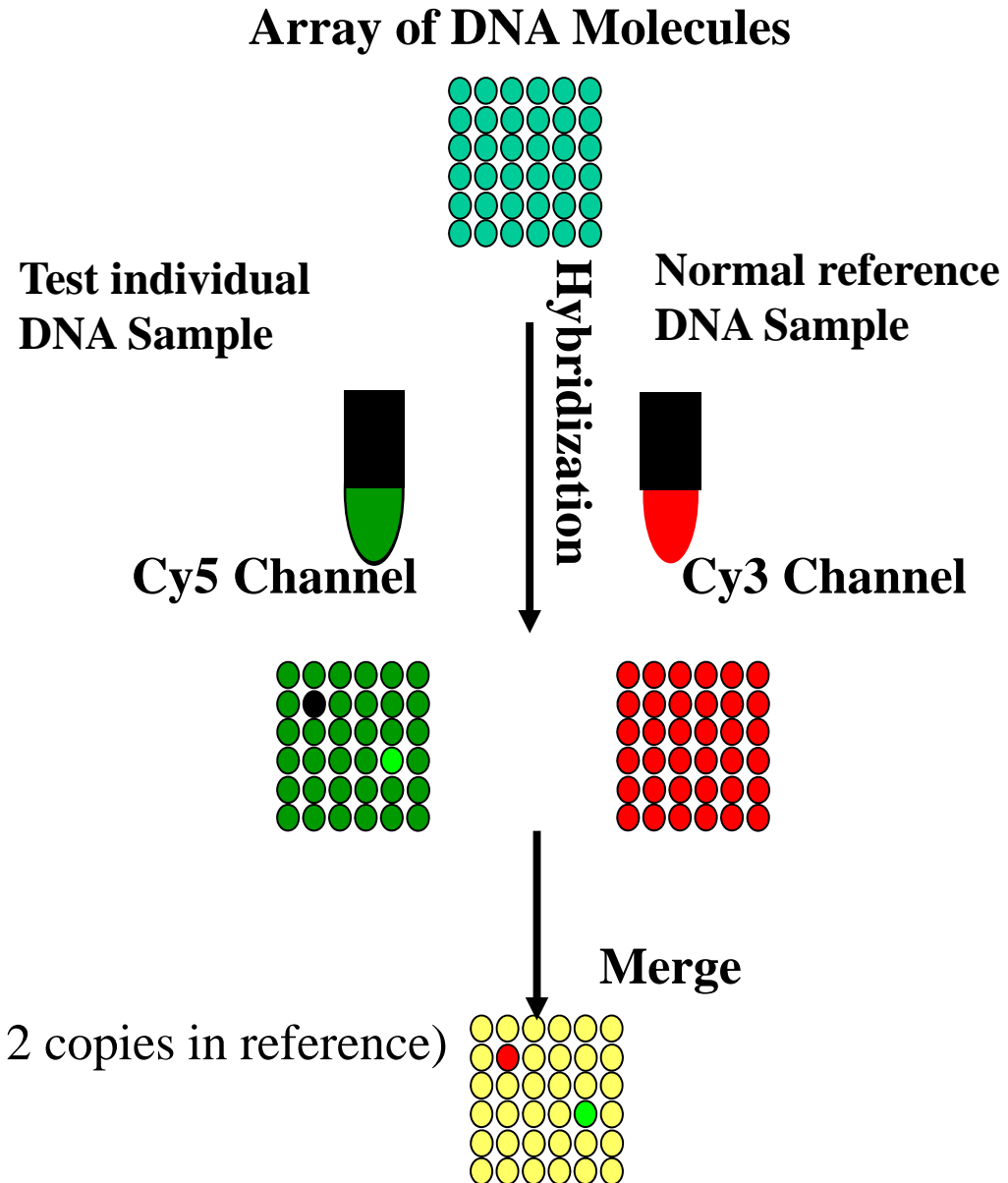
# Array Comparative Genomic Hybridization



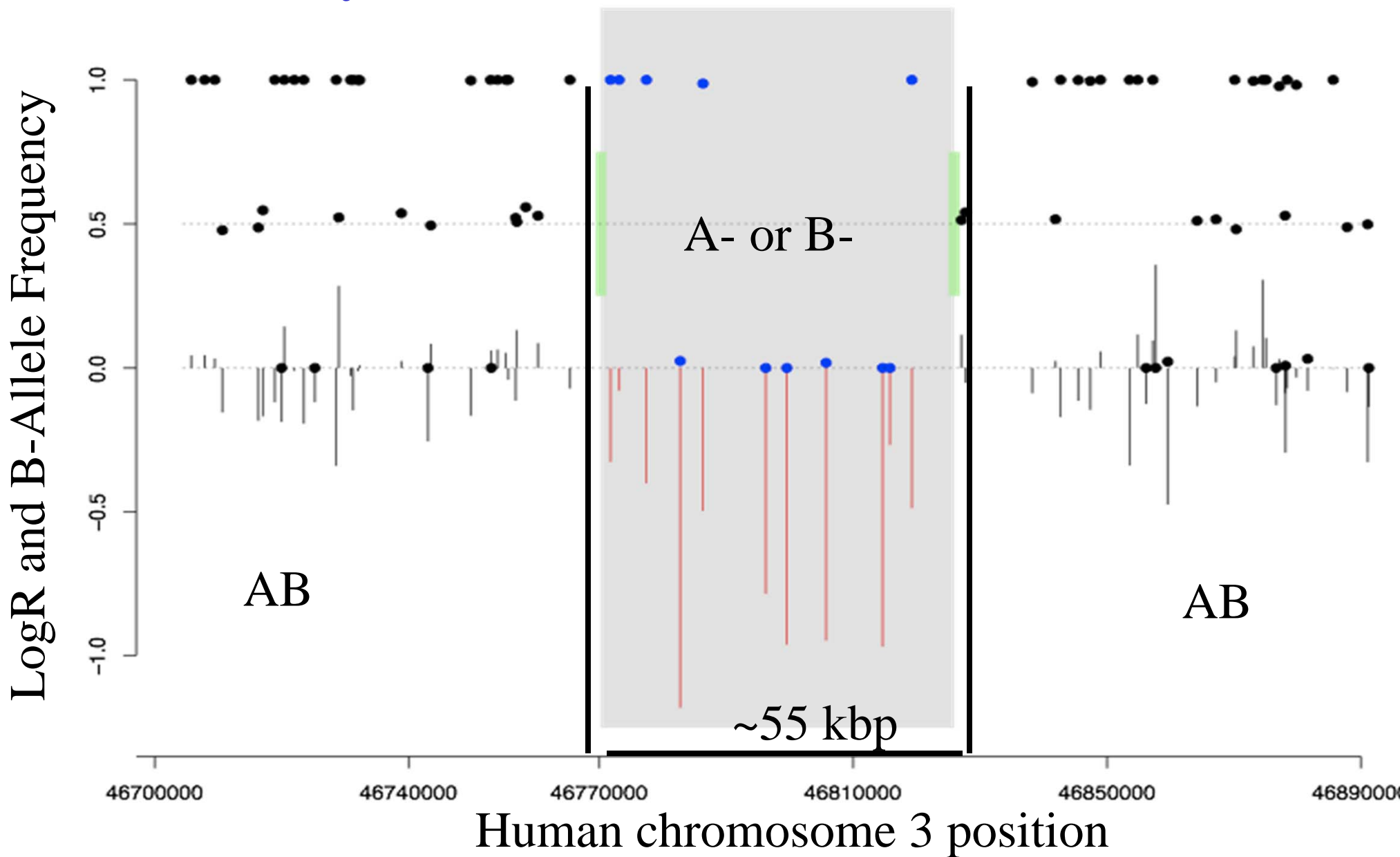
← 12 mm →

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

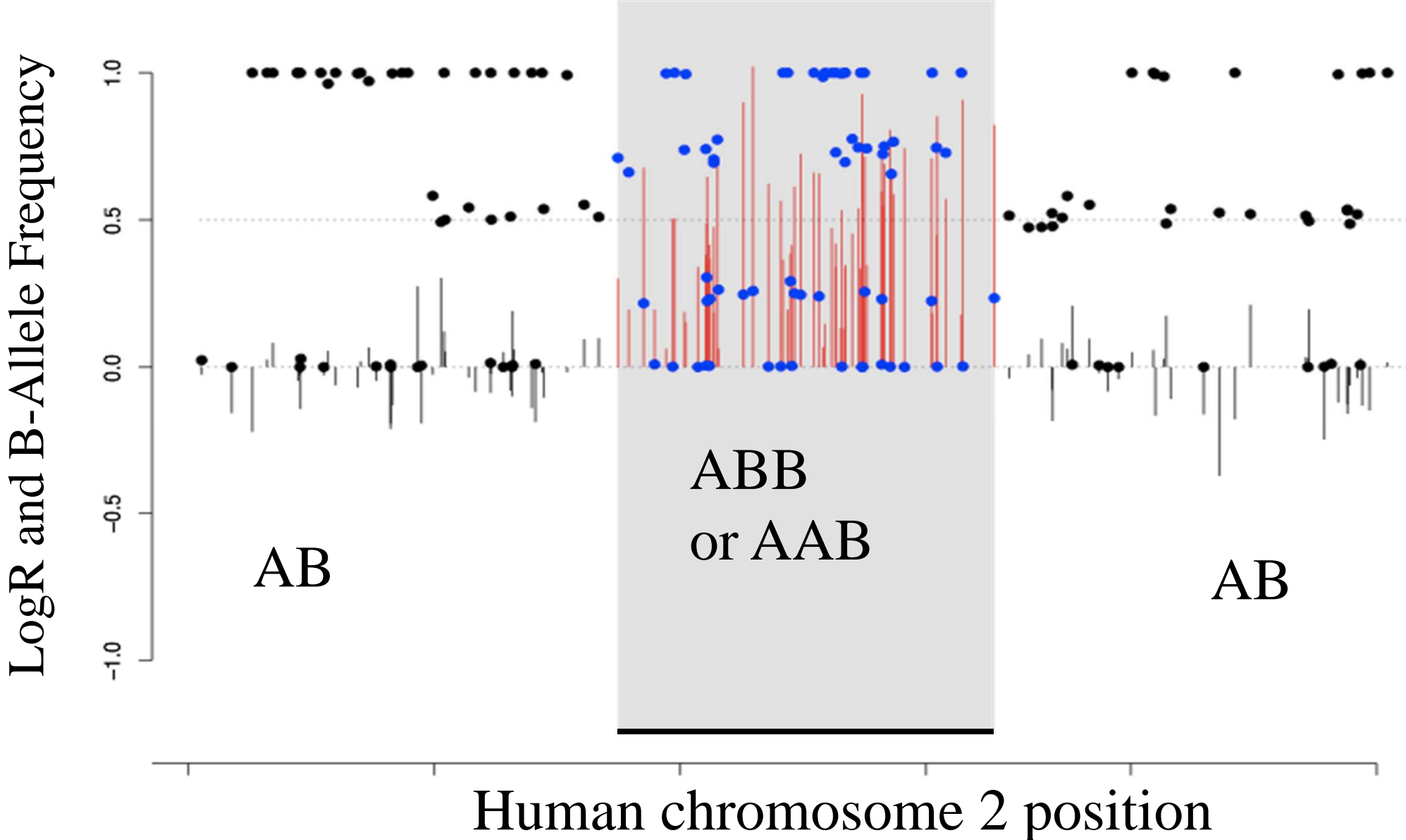
One-copy loss =  $\log_2(1/2) = -1$



## SNP Microarray detection of Deletion (Illumina)

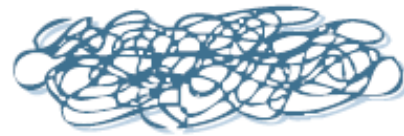


# SNP Microarray detection of Duplication (Illumina)

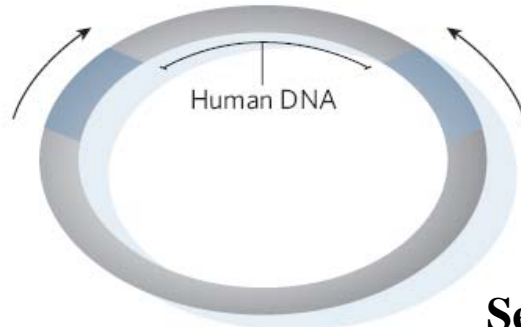




# Using Read Pairs to Resolve Structural Variation



**Human Genomic DNA**



**Genomic Library (1 million clones)**



**Sequence ends of genomic inserts & Map to human genome**

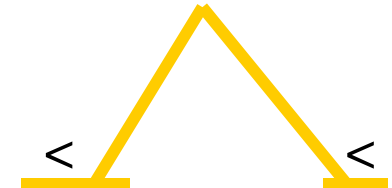
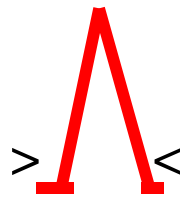
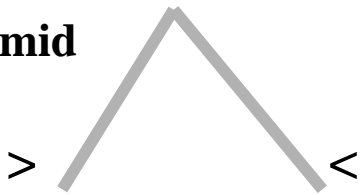
**Concordant**

**Insertion**

**Deletion**

**Inversions**

**Fosmid**

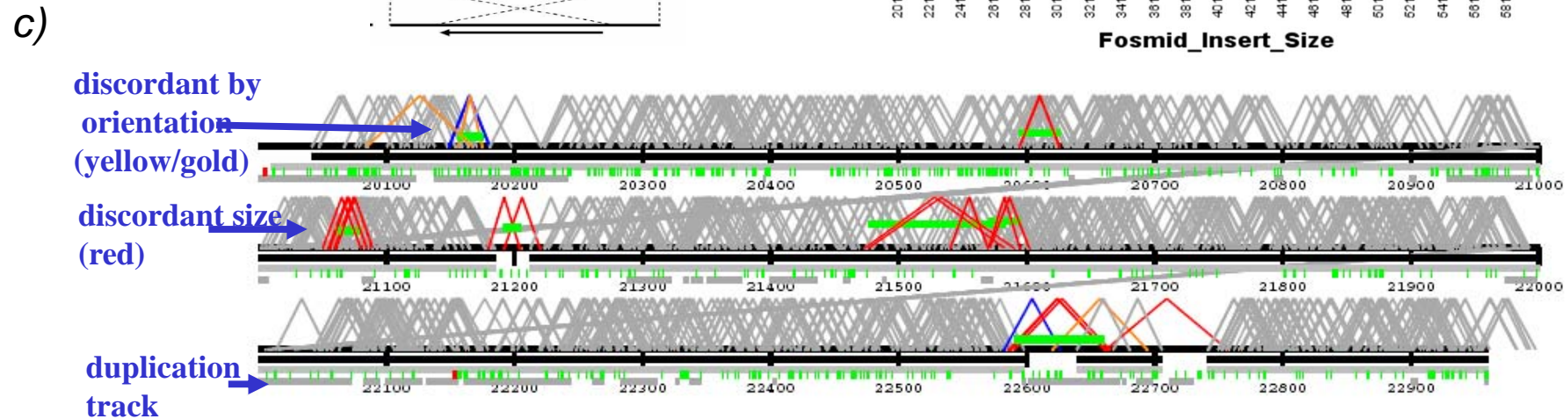
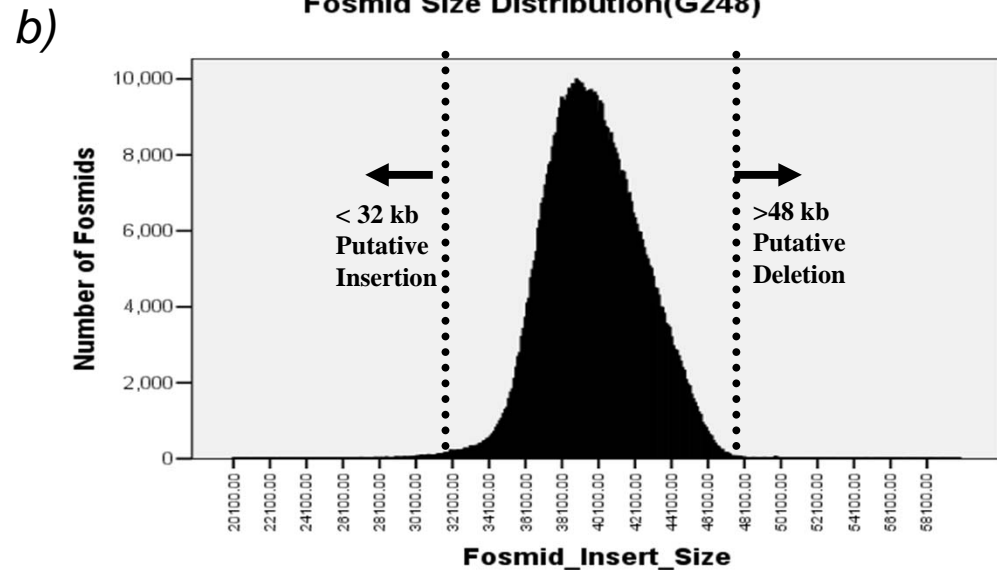
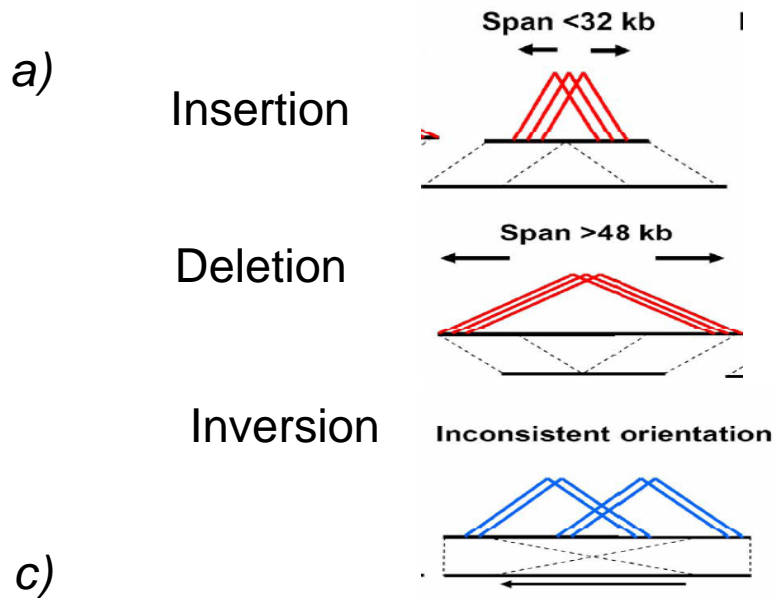


**Build35**

**Dataset: 1,122,408 fosmid pairs preprocessed (15.5X genome coverage)**

**639,204 fosmid pairs BEST pairs (8.8 X genome coverage)**

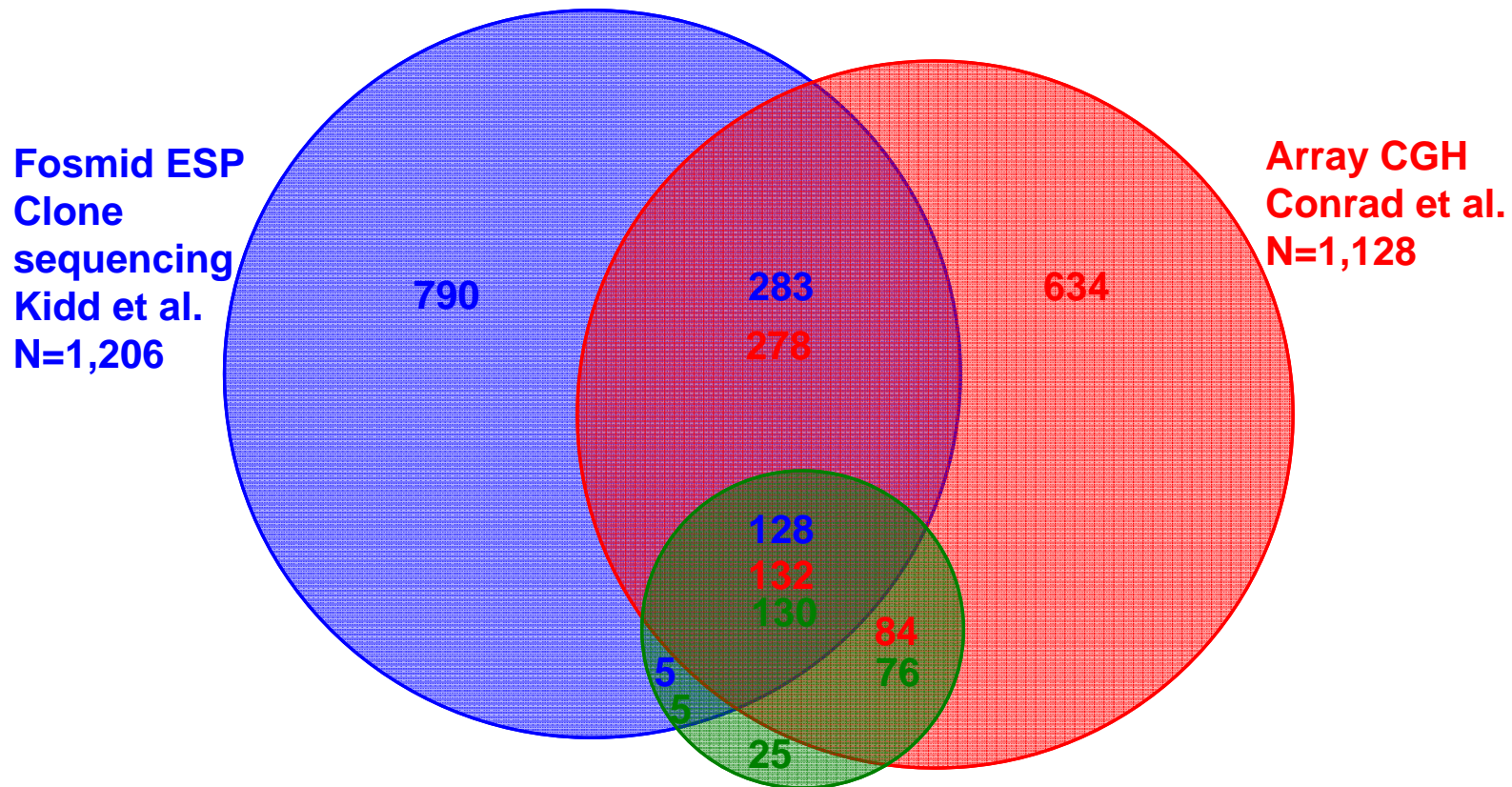
# Genome-wide Detection of Structural Variation (>8kb) by End-Sequence Pairs



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



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

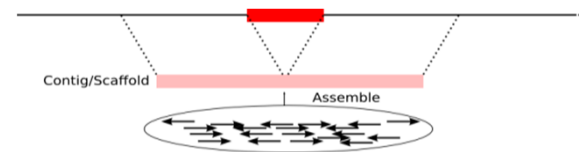
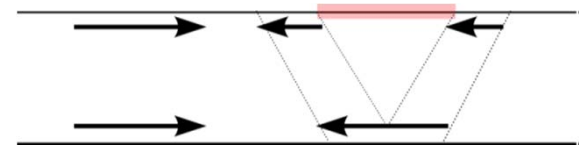
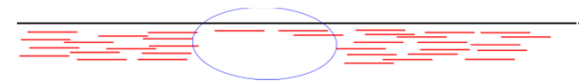
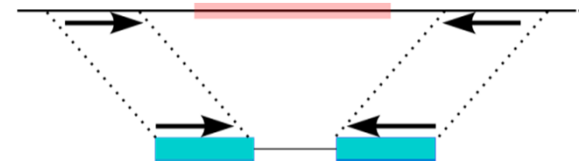


McCarroll et al.  
N=236  
Affymetrix 6.0 SNP Microarray

Kidd et al., *Cell* 2010

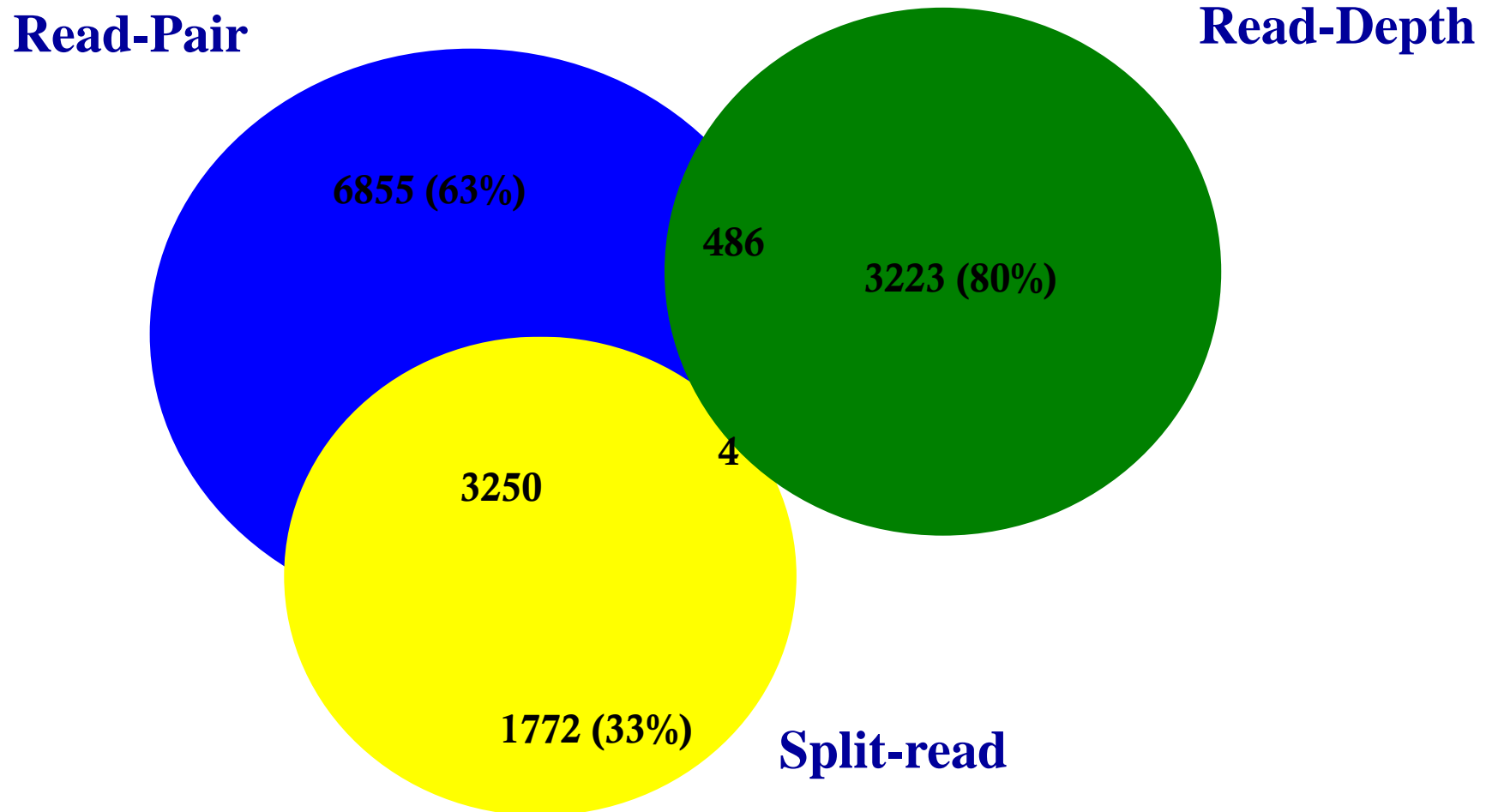
# Next-Generation Sequencing Methods

- **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



# Computational Approaches are Incomplete

159 genomes (2-4X) (deletions only)



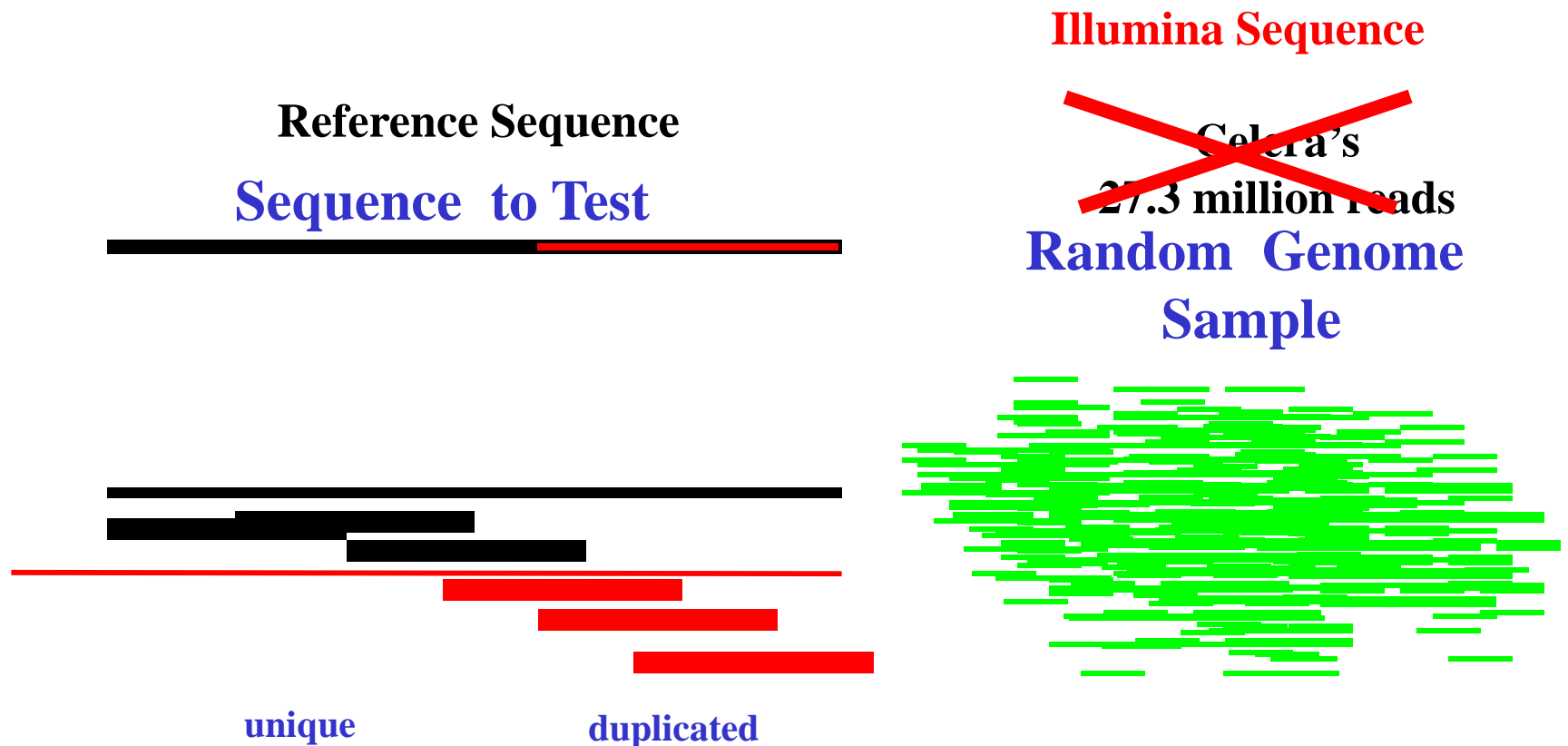
Mills *et al.*, Nature 2011

# Challenges

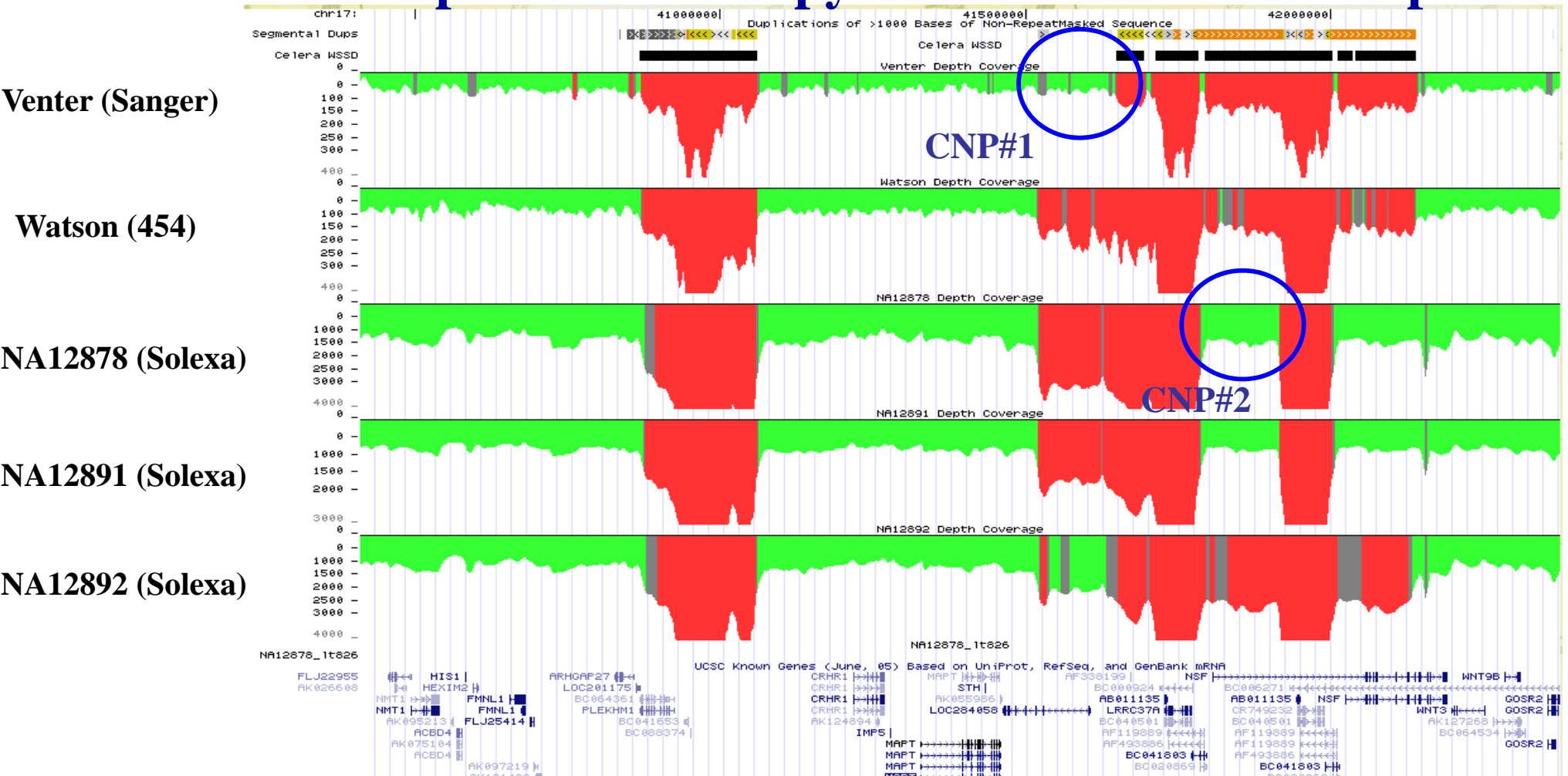
- Size spectrum—>5 kbp discovery limit for most experimental platforms; NGS can detect much smaller but misses events mediated by repeats.
- Class bias: deletions>>> duplications>>>> balanced events (inversions)
- Multiallelic copy number states—incomplete references and the complexity of repetitive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



# Personalized Duplication or Copy-Number Variation Maps

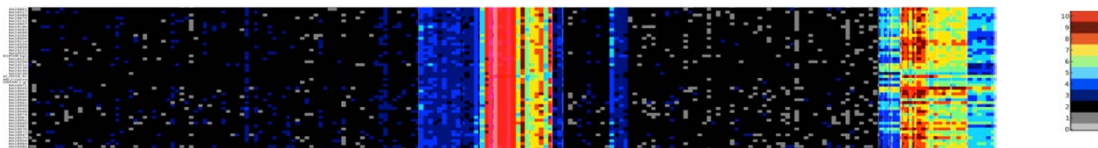
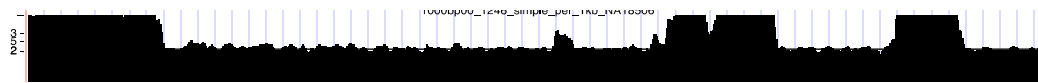
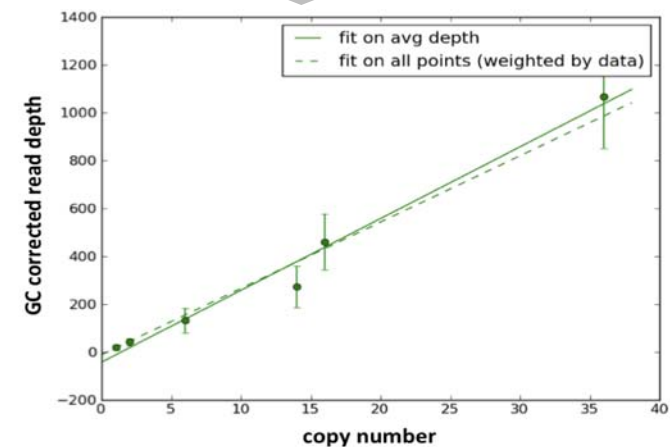
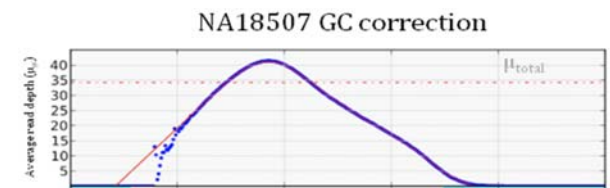
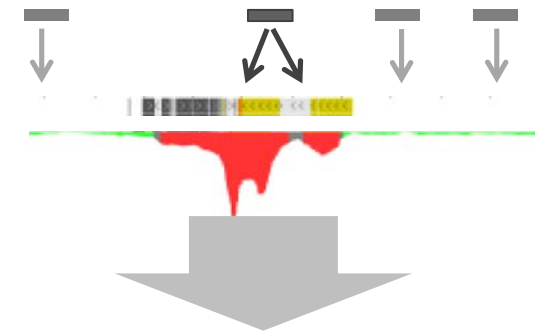


• Two known ~70 kbp CNPs, CNP#1 duplication absent in Venter but predicted in Watson and NA12878, CNP#2 present mother but neither father or child

*Alkan, Nat. Genet, 2009*

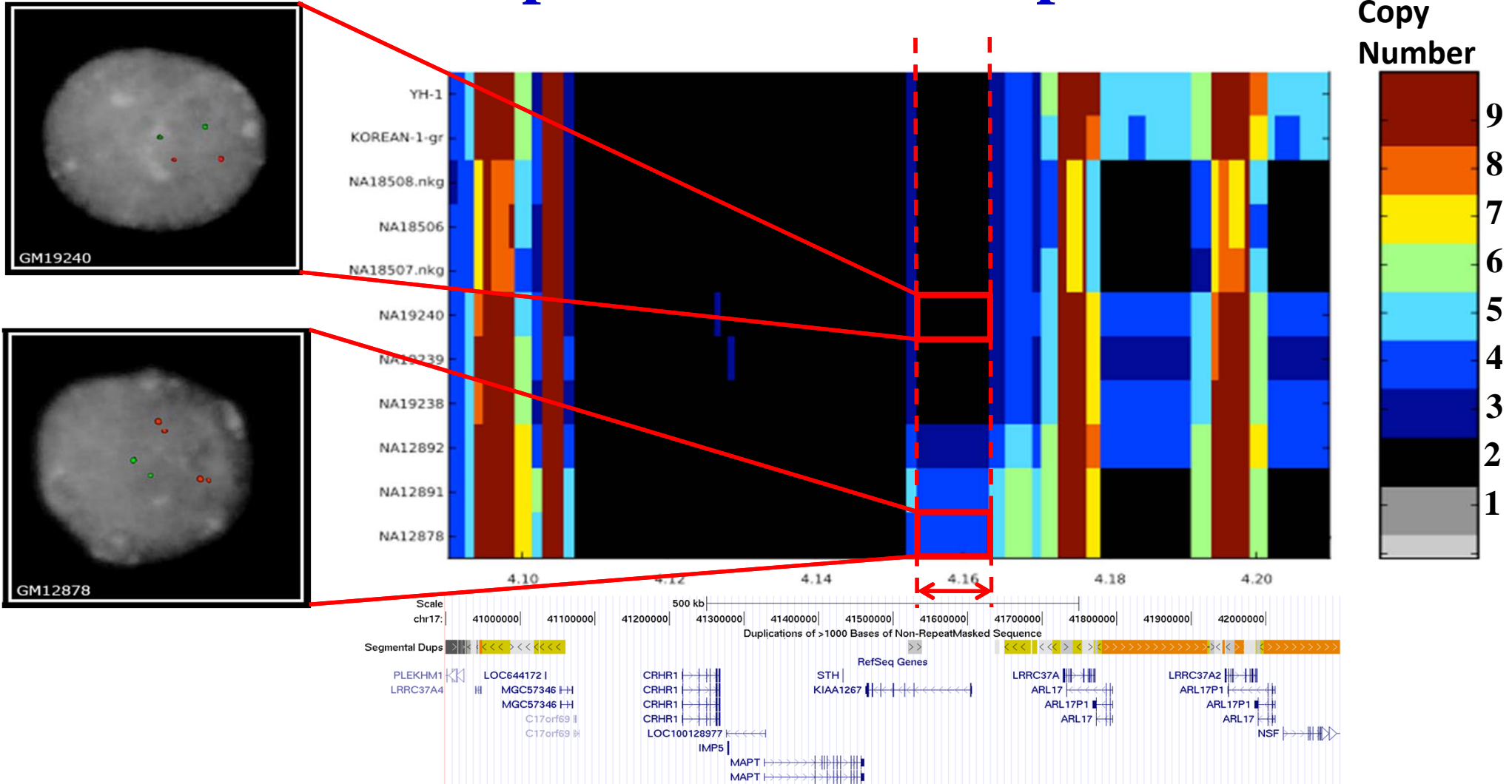
# Copy number from short read depth

- 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



Interphase FISH

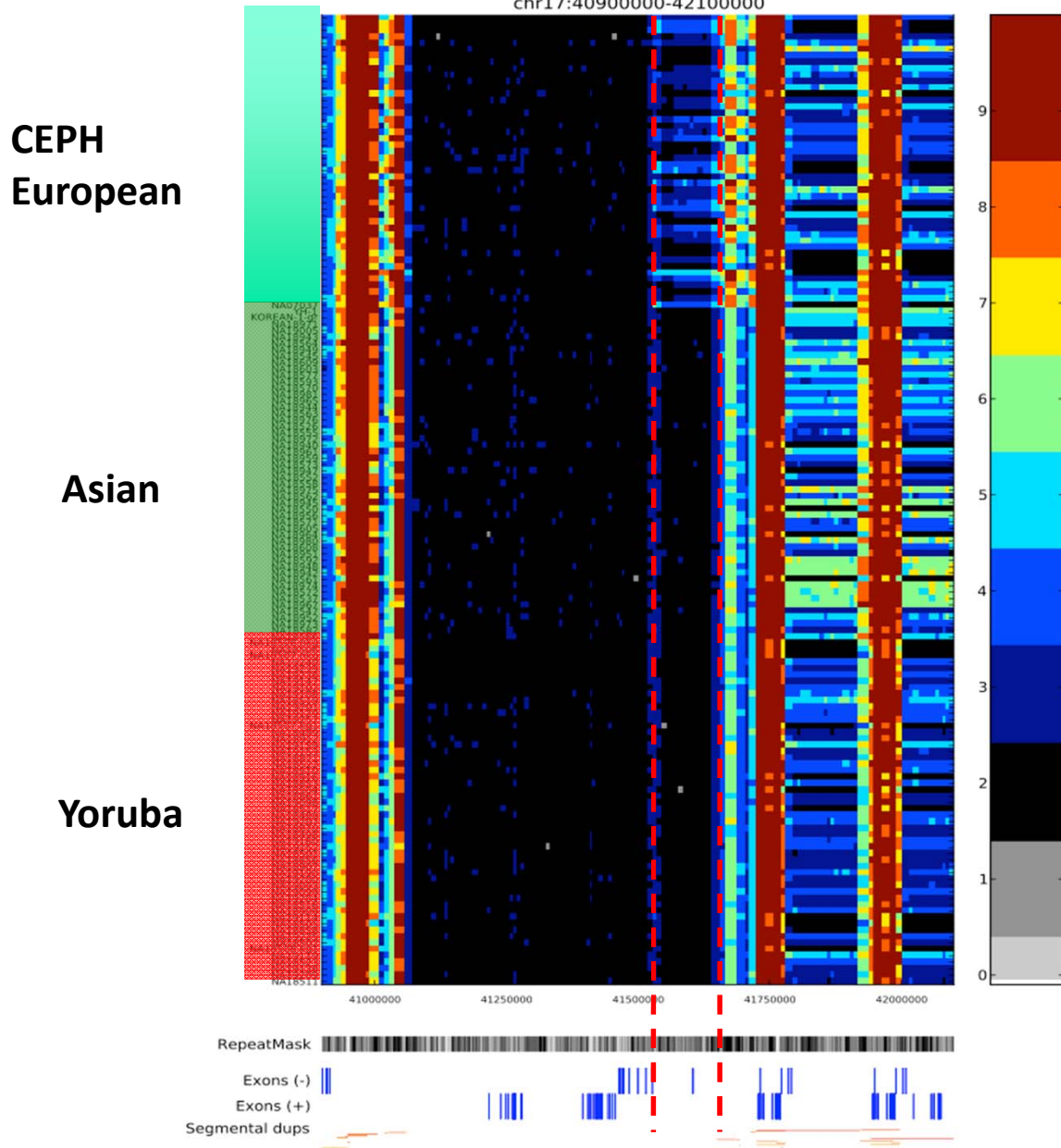
# Read-Depth CNV Heat Maps vs. FISH



- 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

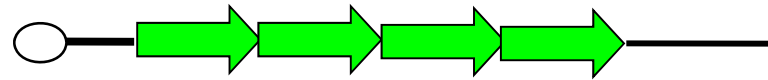


71% of Europeans carry at least Partial duplication distal (17q21 associated)—all inversions carry the duplication

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

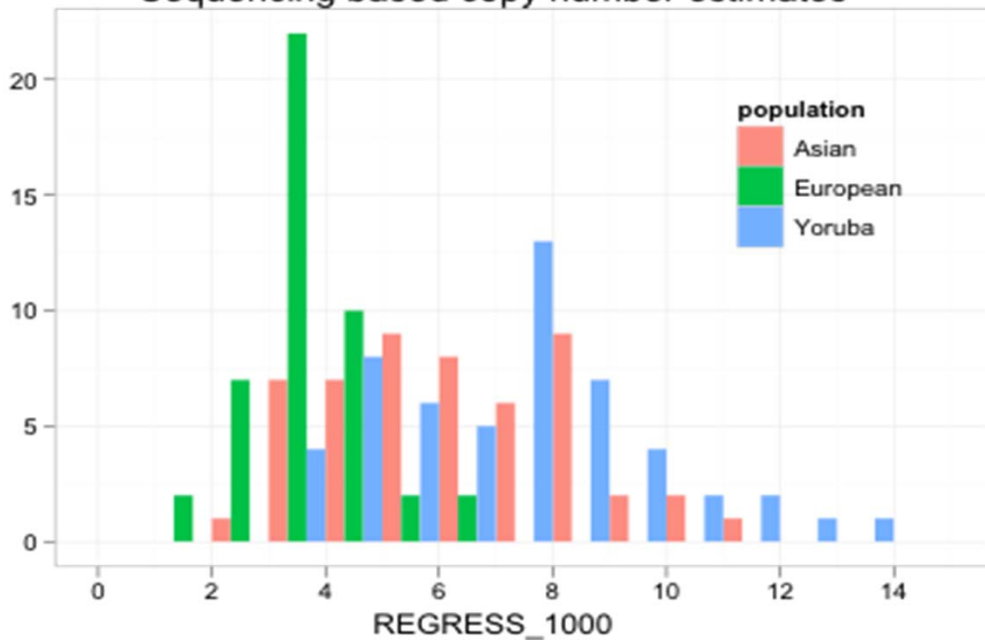
Sudmant et al., 2010, Science

# Read-Depth vs. Quantitative PCR

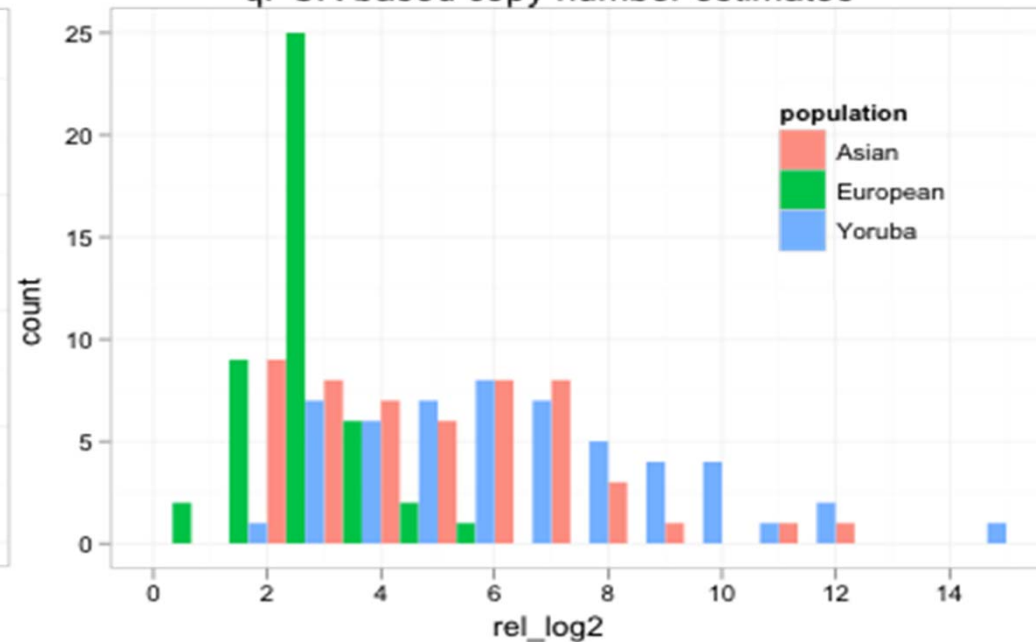


CCL3L1—chemokine ligand 3-like (1.9 kbp)

Sequencing based copy number estimates



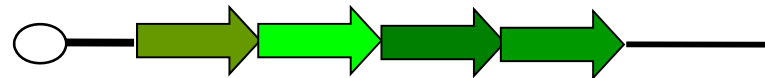
qPCR based copy number estimates



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

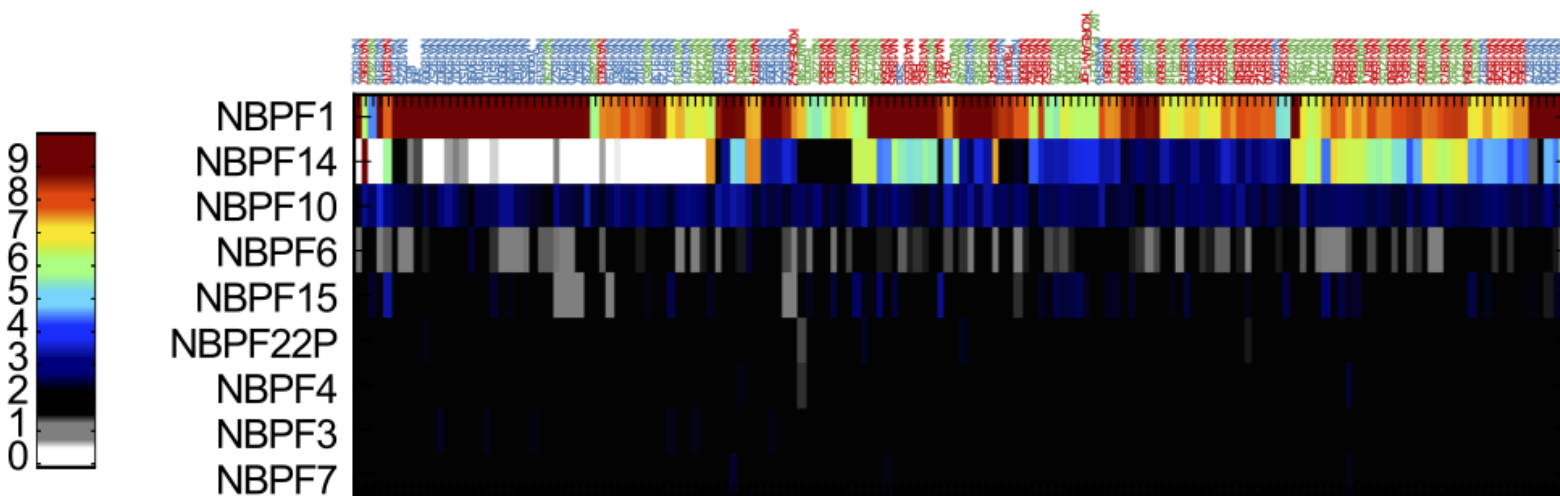
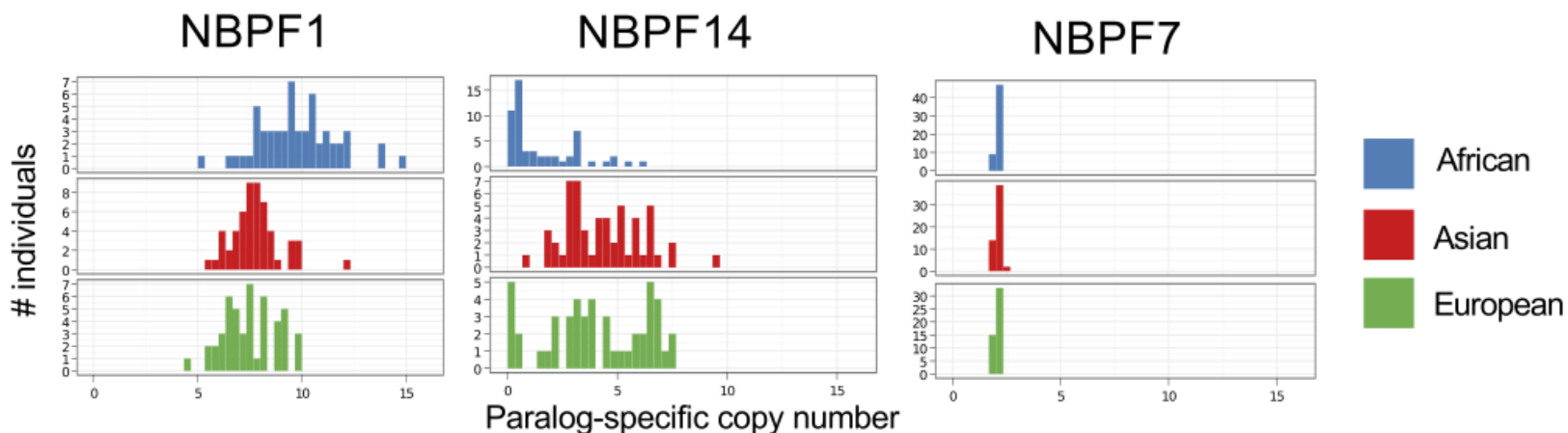
# Unique Sequence Identifiers Distinguish Copies

copy1 ATGCTAGGCATATAATATCCGACGATATACATATAGATGTTAG...  
copy2 ATGCTAGGCATAGAATATCCGACGATATACATATACATGTTAG...  
copy3 ATGCTACGCATAGAATATCCACGATATACATATACATGTTAG...  
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

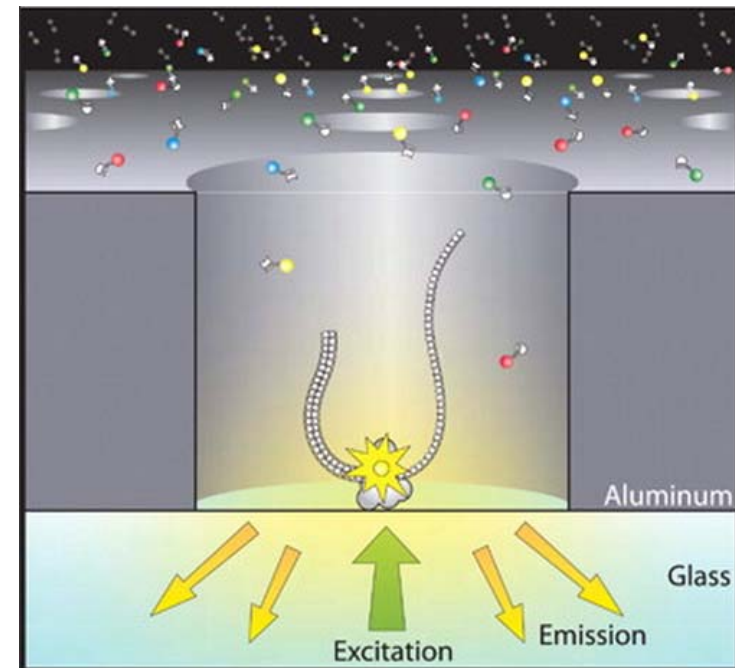
# NBPF Gene Family Diversity



# 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*

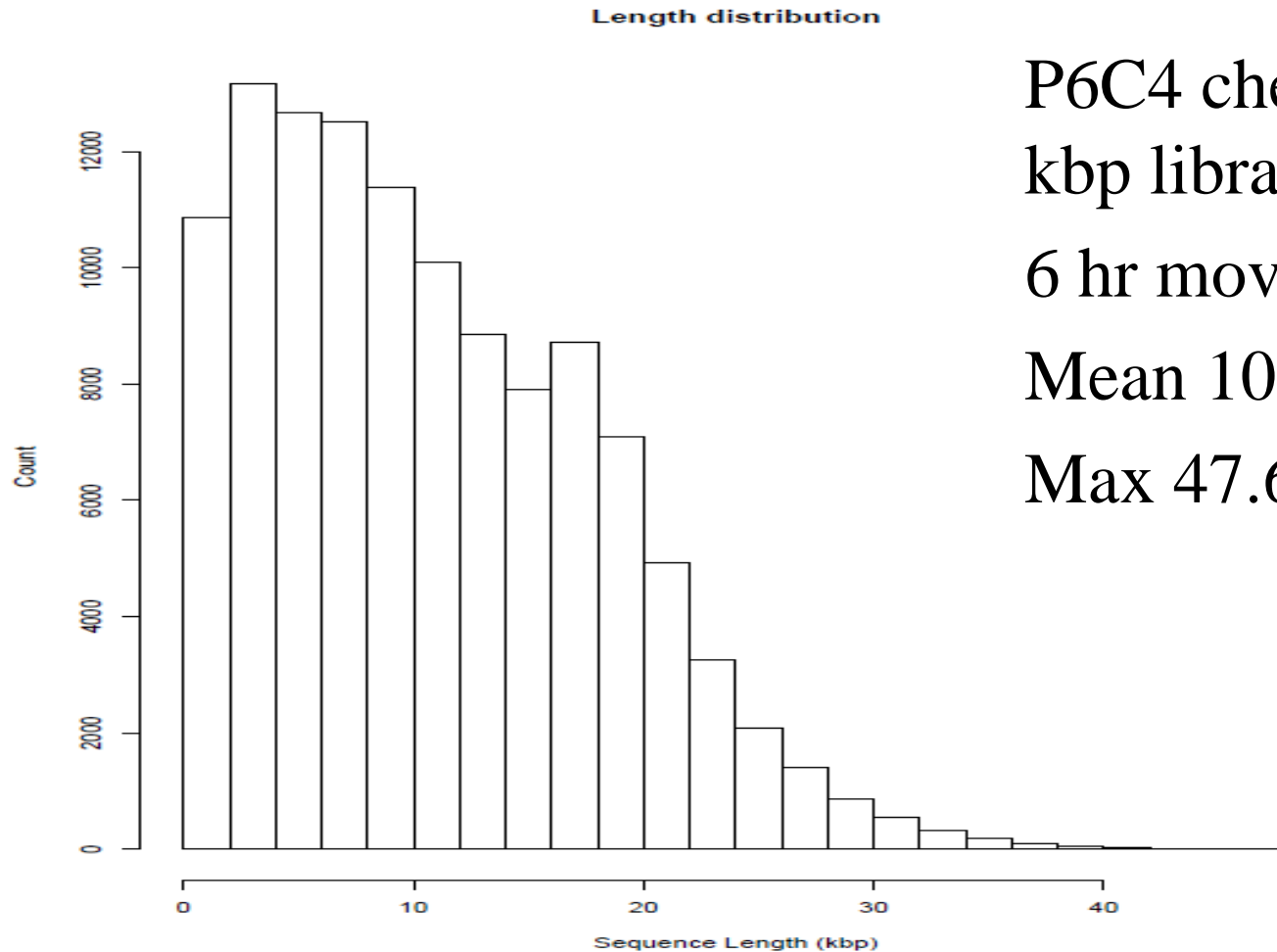
# Single-Molecule Real-Time Sequencing (SMRT)



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



# PacBio Sequence Reads are long



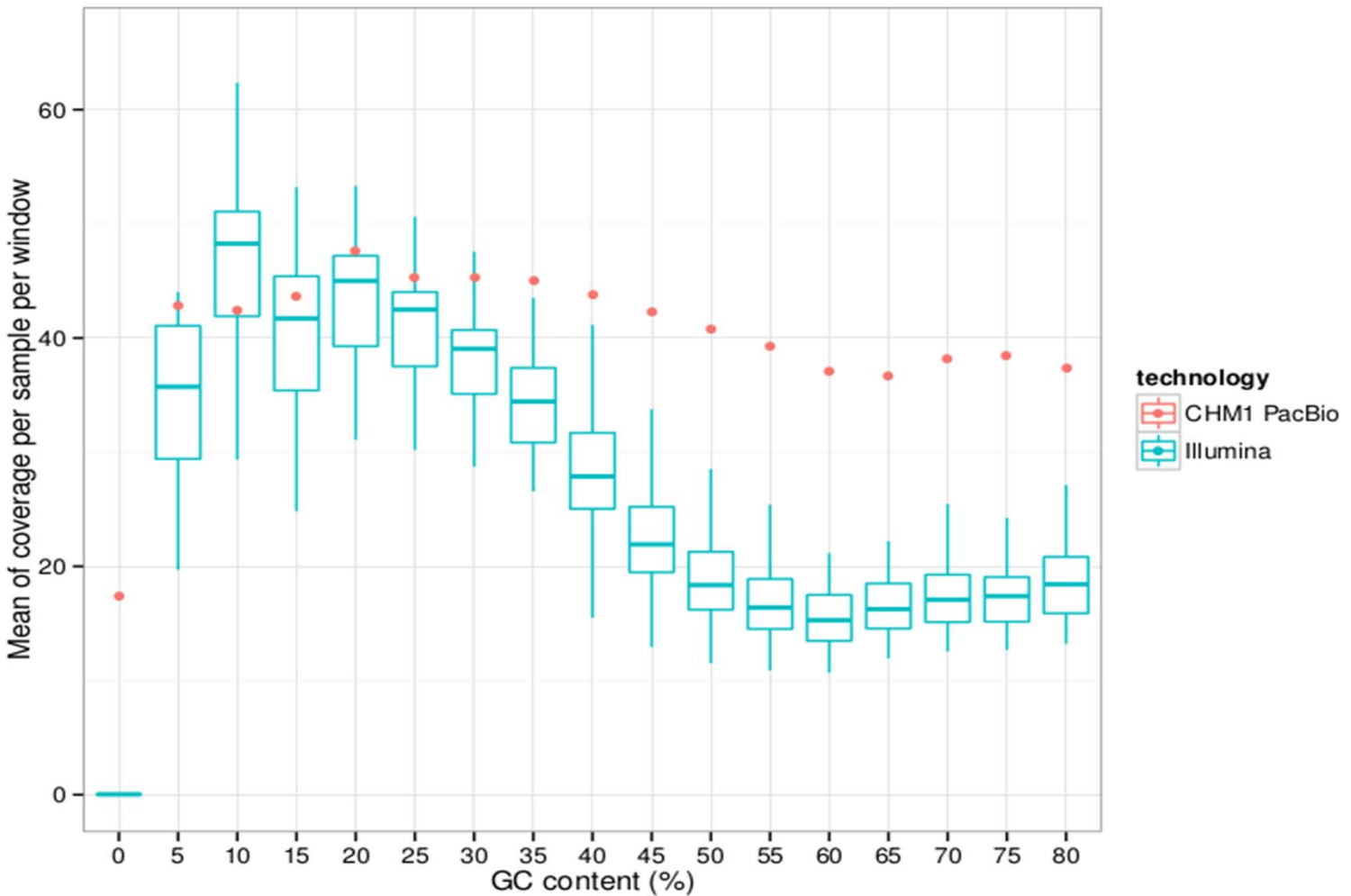
P6C4 chemistry—30-40  
kbp libraries

6 hr movie

Mean 10.8 kbp read

Max 47.6 kbp

# PacBio Sequence Reads are Uniform

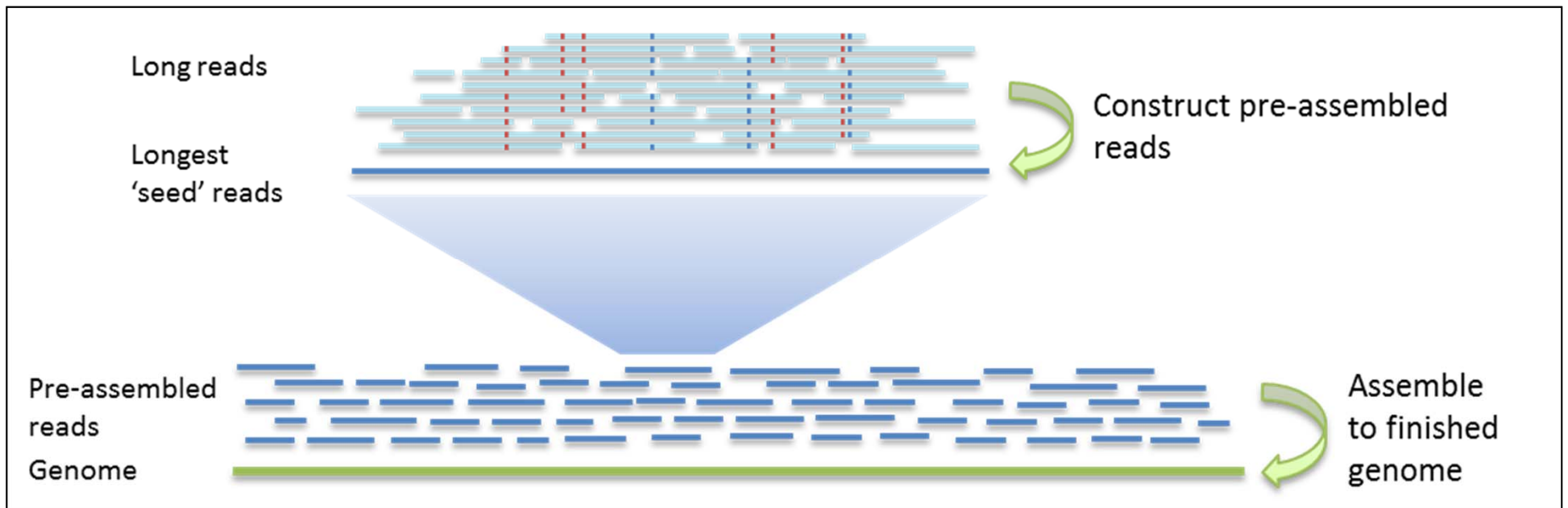


# Algorithms: HGAP and QUIVER

ARTICLES

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

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

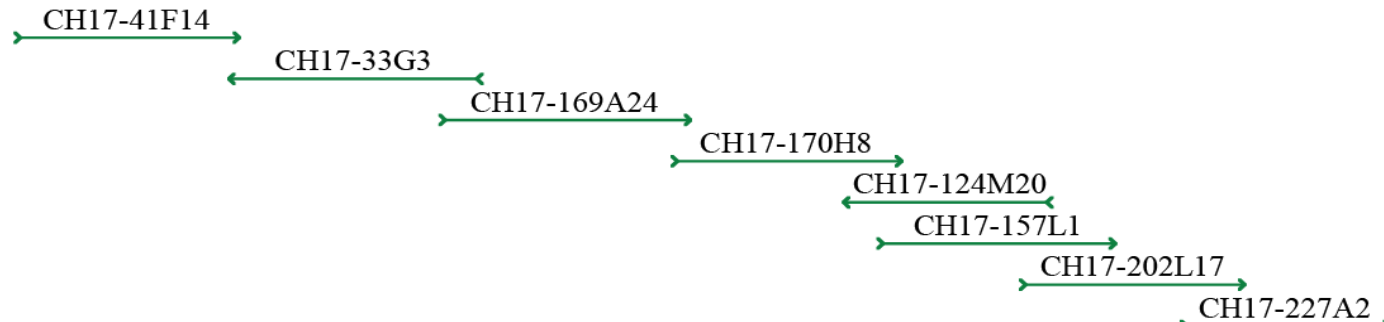


<https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP>

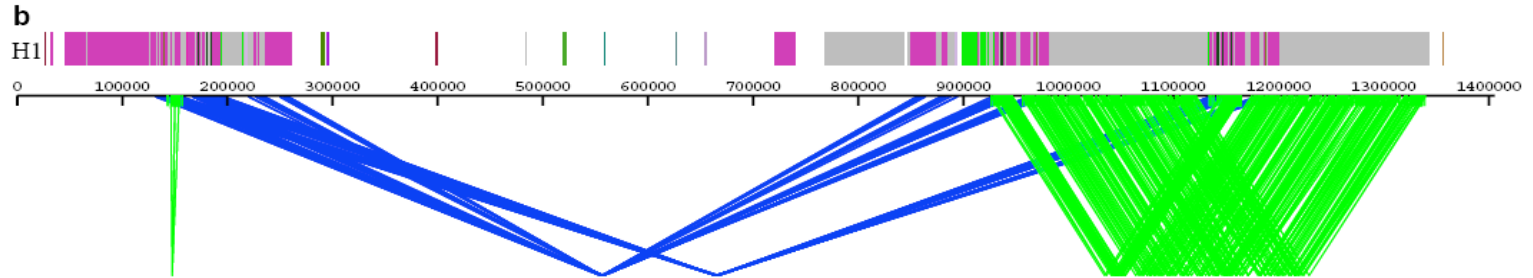
Chin et al. *Nat. Methods*, 2013

# Clone Based Resolution of SV

## BAC Tiling Path



## Seg Dup Organization

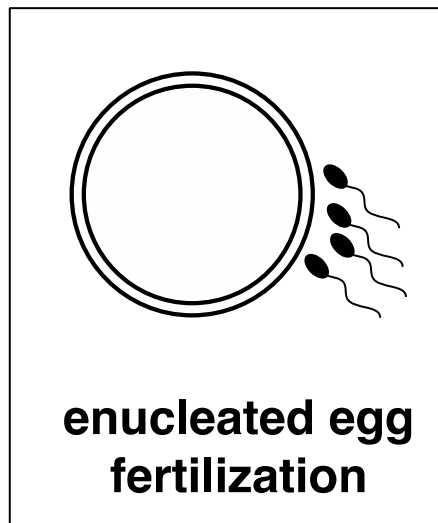


- Select tiling path of BAC clones corresponding to a complex region previously sequenced using Sanger
- Sequence each clone (~200 fold) using on average 1 SMRT Cell and assemble using HGAP and QUIVER
- Compare Sanger and Pacbio assembly using BLASR shows accurate ( $QV > 45$ ) assembly of complex region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

Huddleston et al. *Genome Res*, 2014

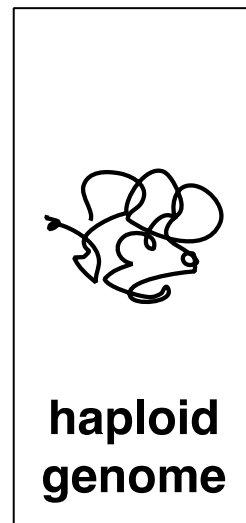
# PacBio Whole Genome Sequencing

- 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.

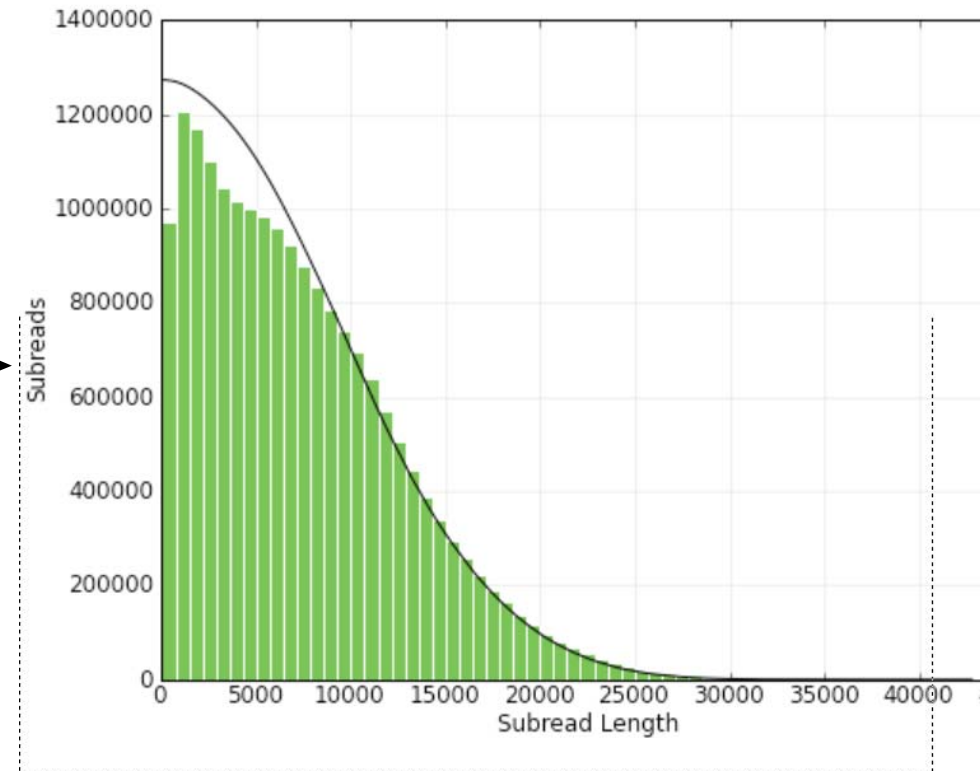


DNA  
extract

→

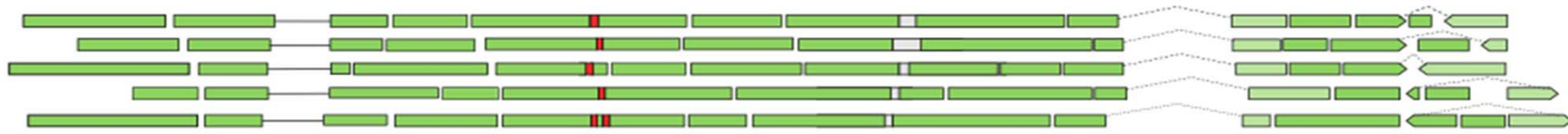


→

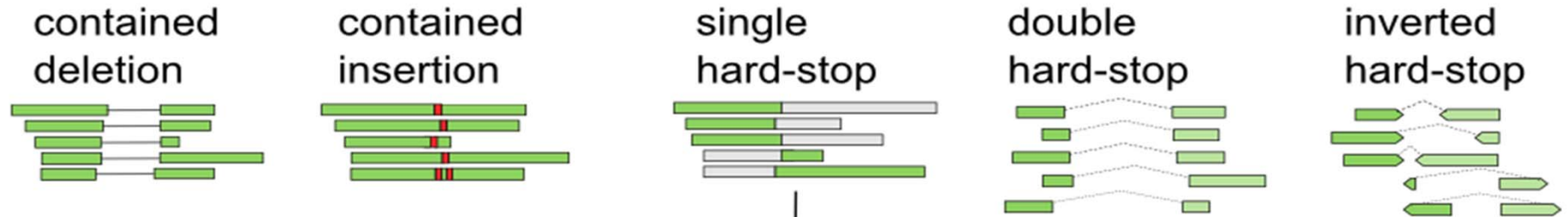


<http://datasets.pacb.com/2013/Human10x/READS/index.html>

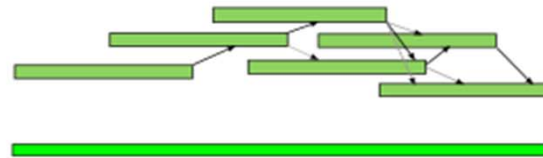
## BLASR alignment of reads



## Signatures of structural variants



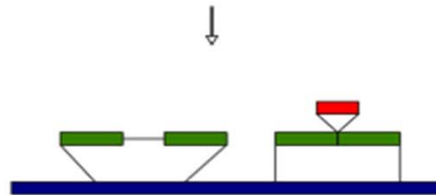
## Celera assembly



## Remap reads, generate Quiver consensus

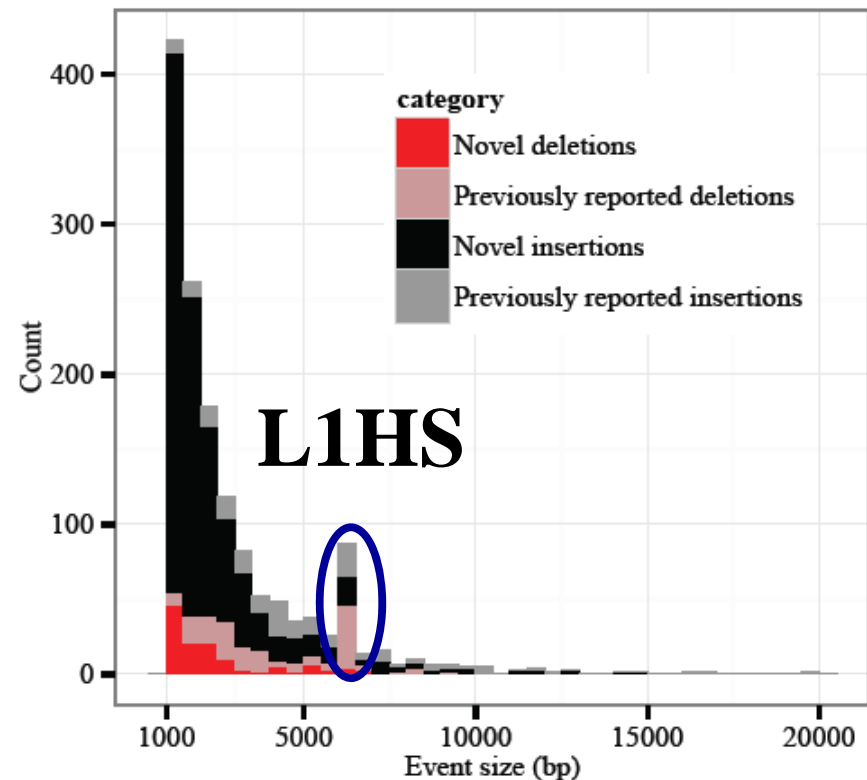
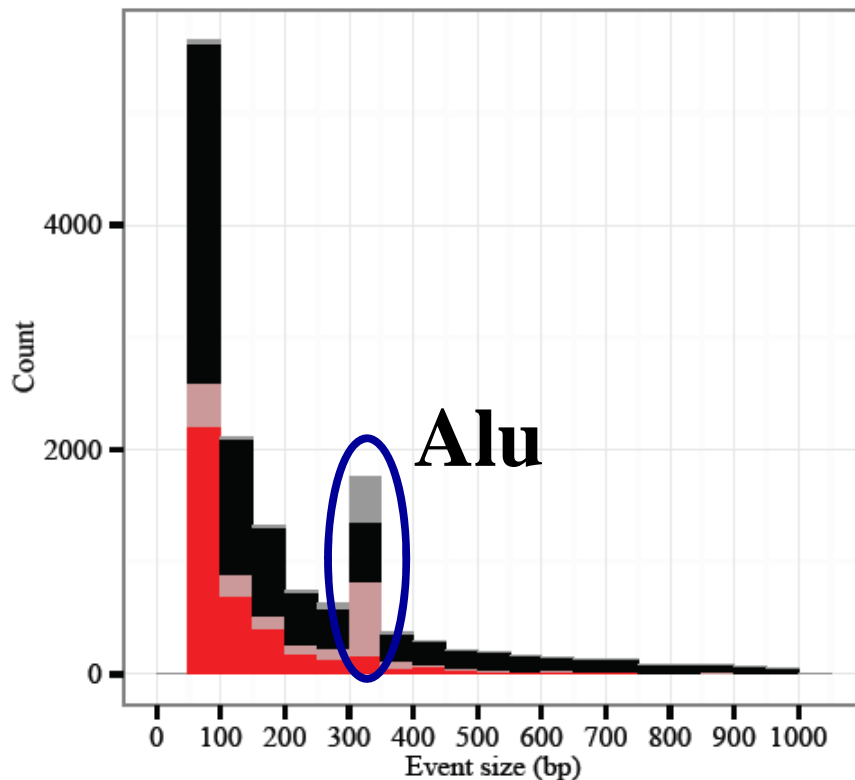


## Map consensus, structural variant resolution



Chaisson et al, Nature, 2014

# Increased Resolution of Structural Variation



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

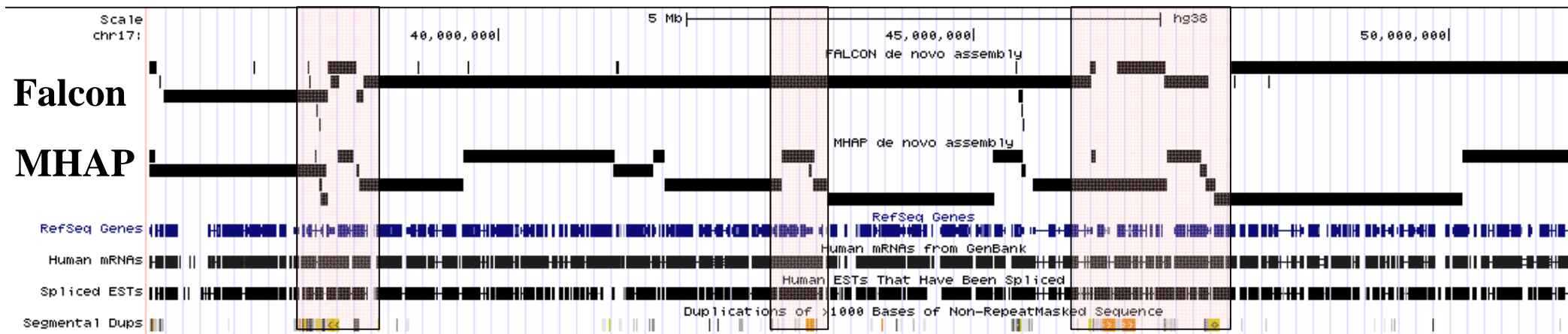
**22,112 novel genetic variants corresponding to 11 Mbp of sequence**

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

169 within coding sequence or UTRs of genes



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



- **Falcon Assembly (Jason Chin) and MHAP Assembly (Berlin/Philippy) N50 is ~ 5 Mbp**
- 125/167 Mbp of SD unresolved
- Contigs shatter over segmental duplications

# 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).

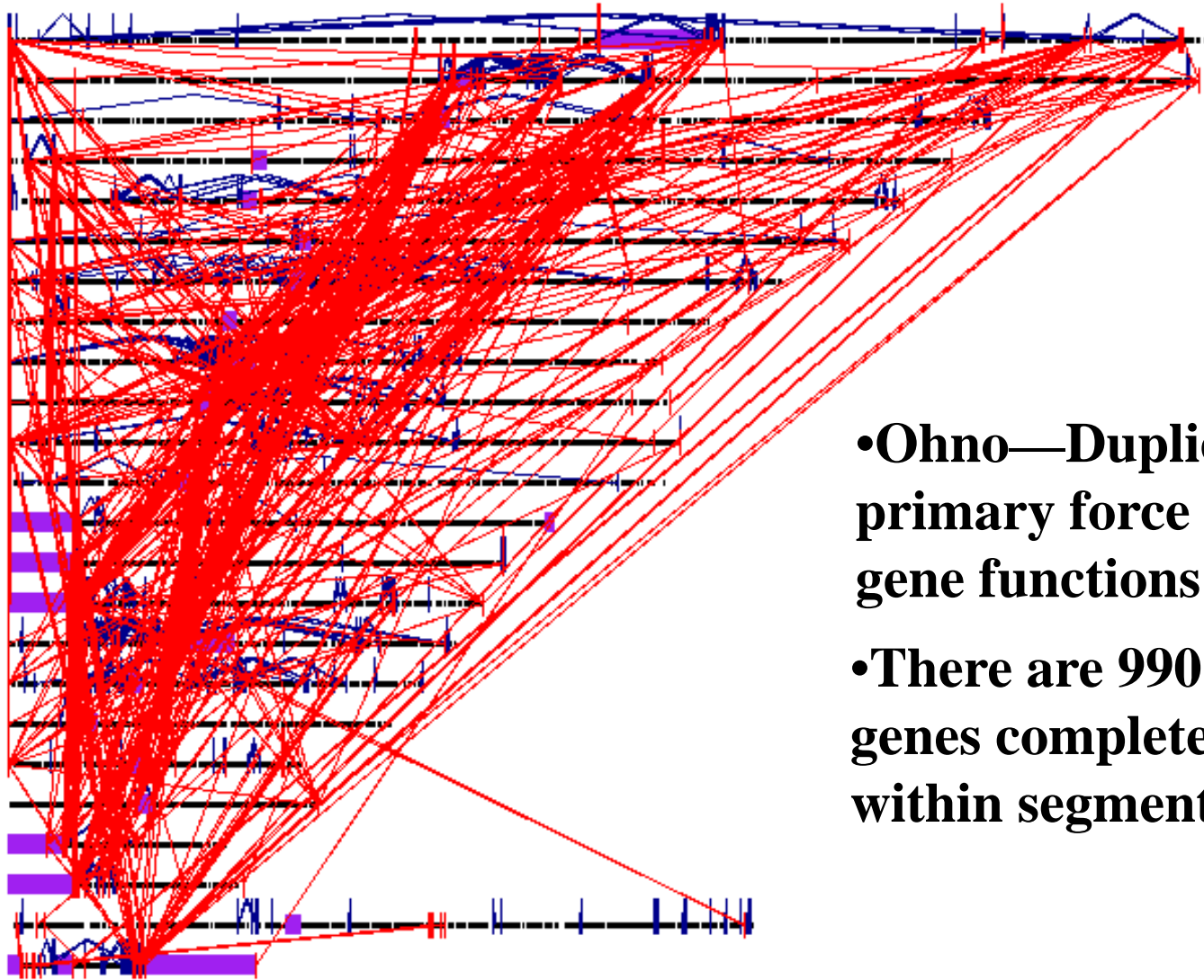


# Summary

- 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

# III. Why?

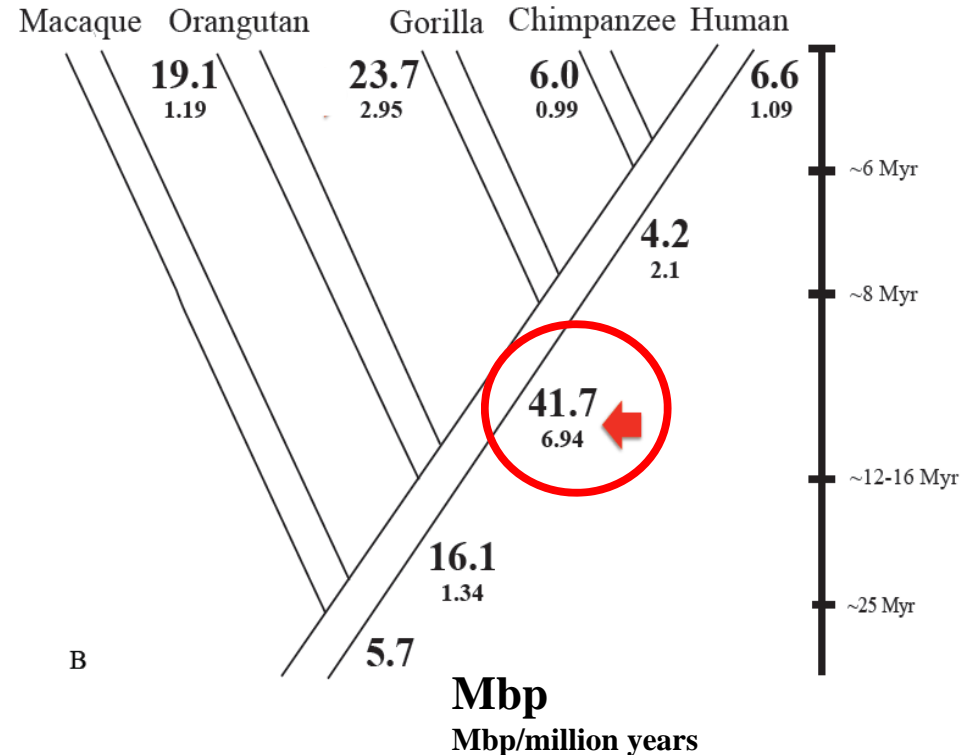
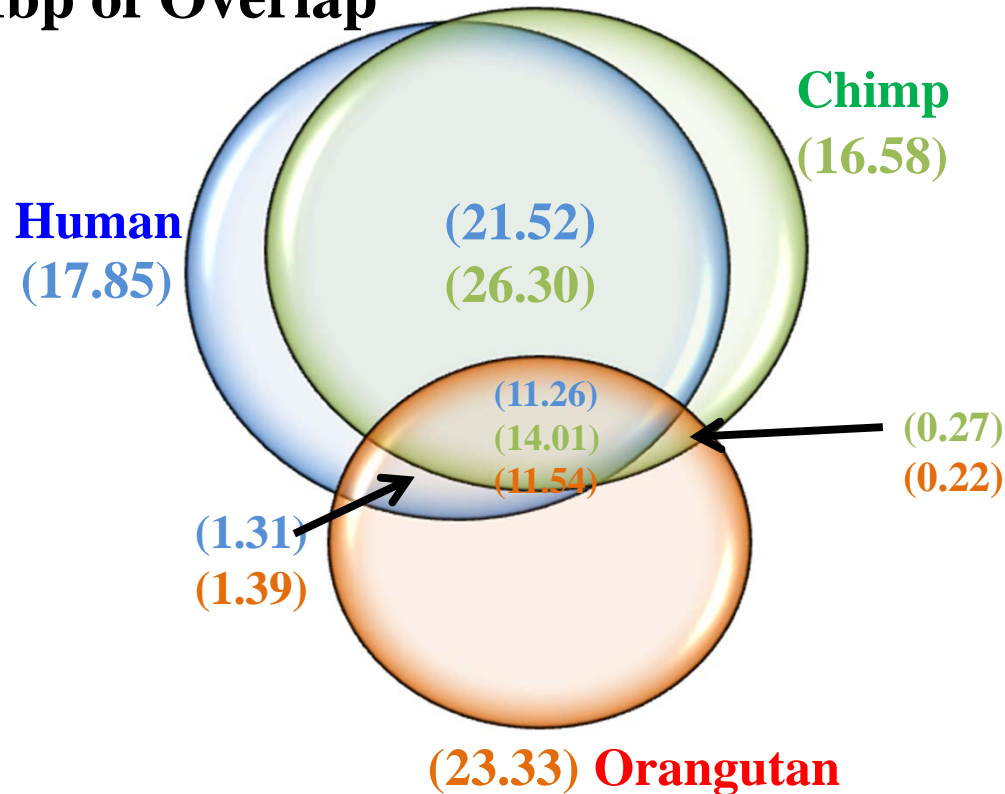
chr1  
chr2  
chr3  
chr4  
chr5  
chr6  
chr7  
chr8  
chr9  
chr10  
chr11  
chr12  
chr13  
chr14  
chr15  
chr16  
chr17  
chr18  
chr19  
chr20  
chr21  
chr22  
chrX  
chrY



- **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

Mbp of Overlap

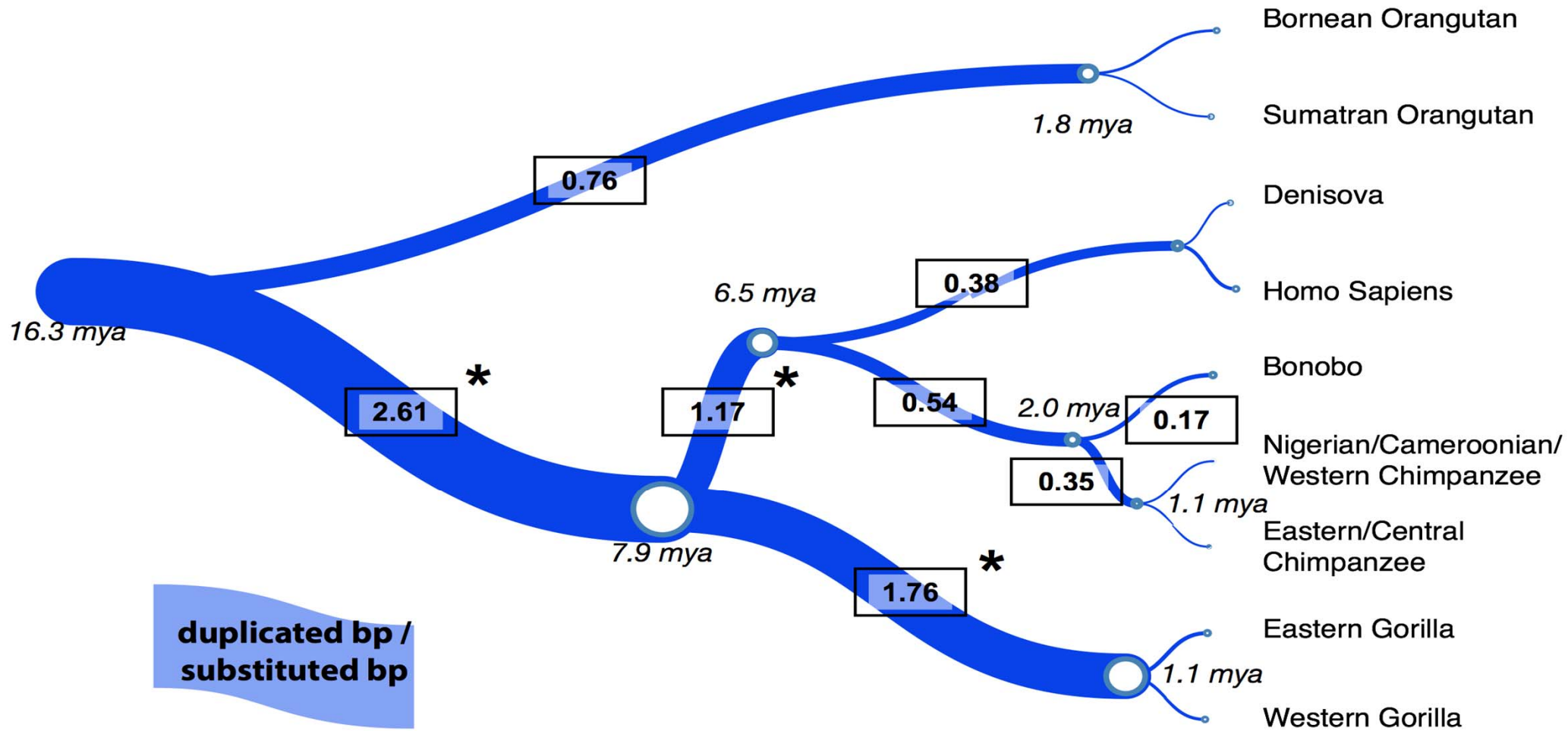


- A 3-4 fold excess in *de novo* segmental duplications in common ancestor of human, chimp and gorilla but after divergence from orangutan

- Not a continuous accumulation

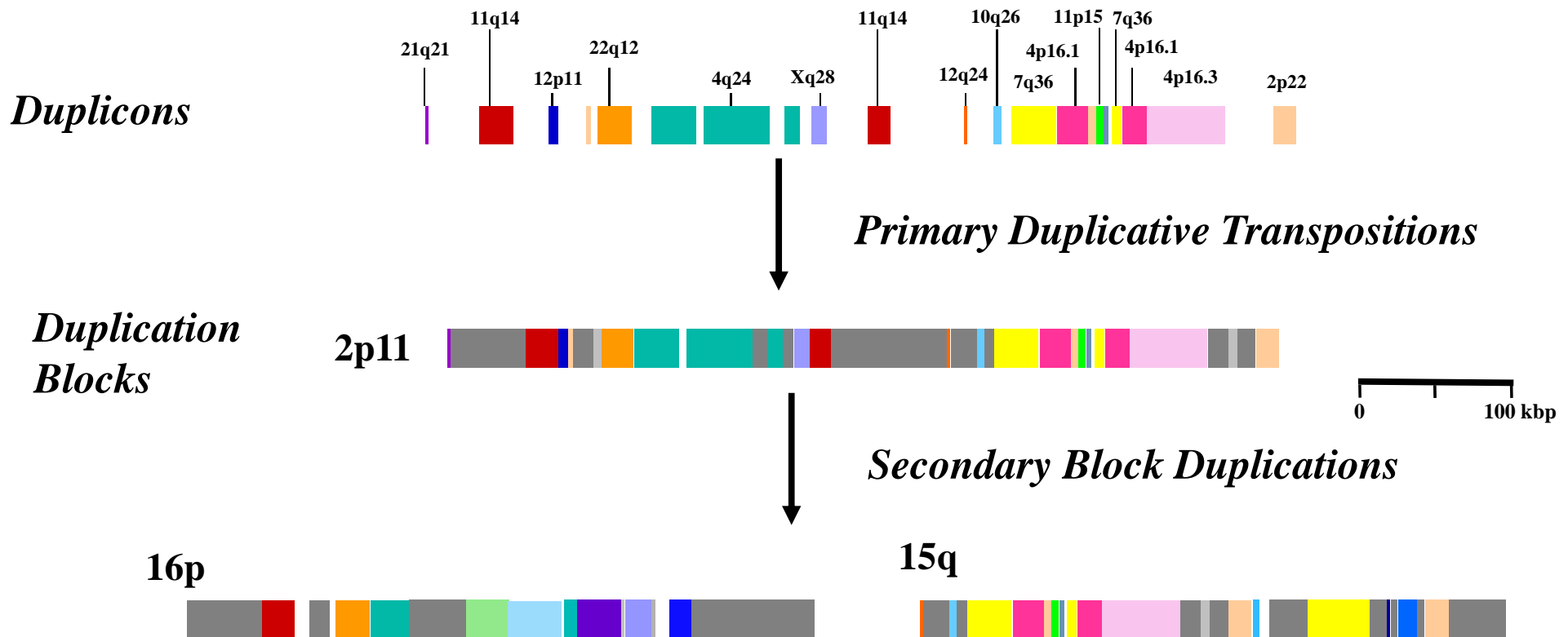
Marques-Bonet et al., *Nature*, 2009; Ventura et al., *Genome Res.* 2011

# Rate of Duplication



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

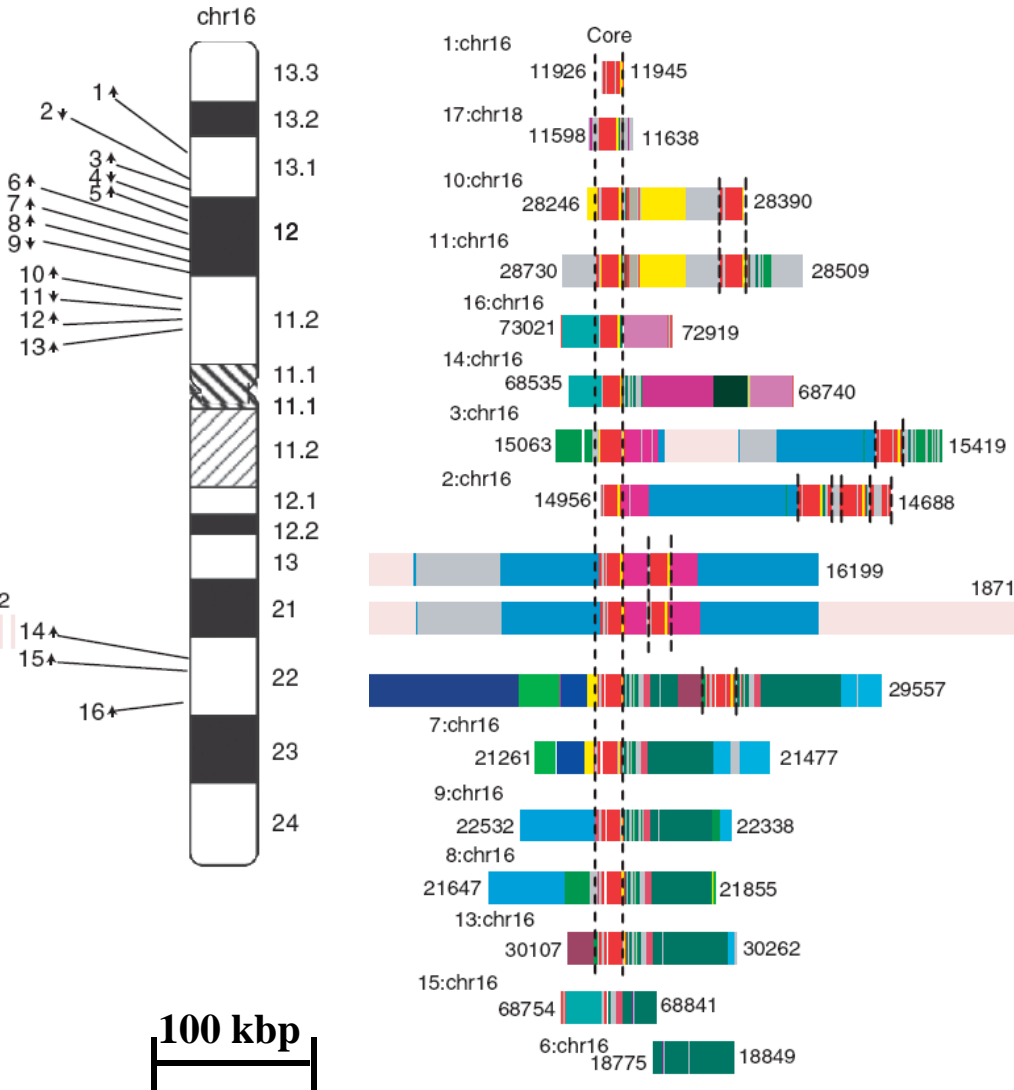
# Mosaic Architecture



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



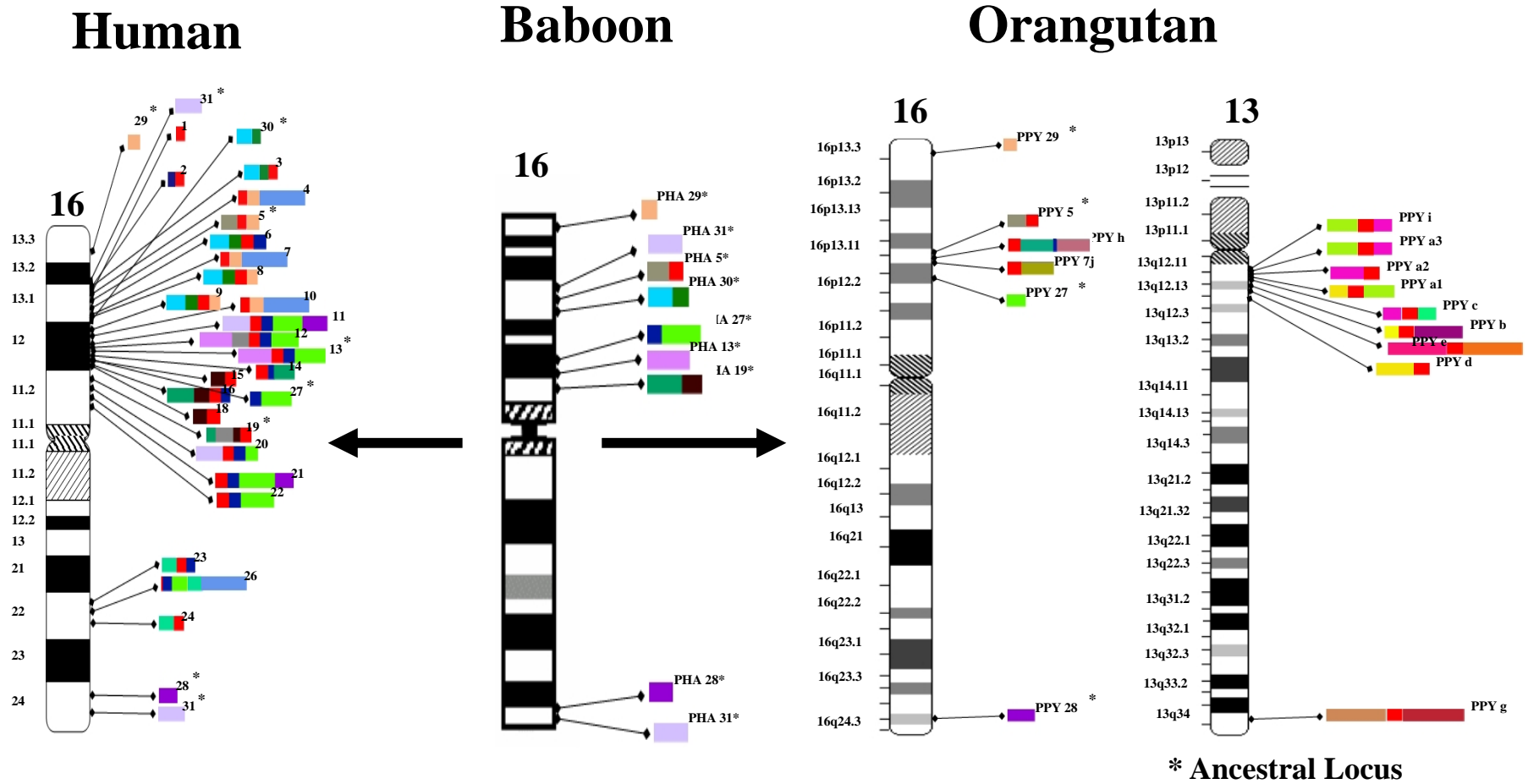
# Human Chromosome 16 Core Duplicon



•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)

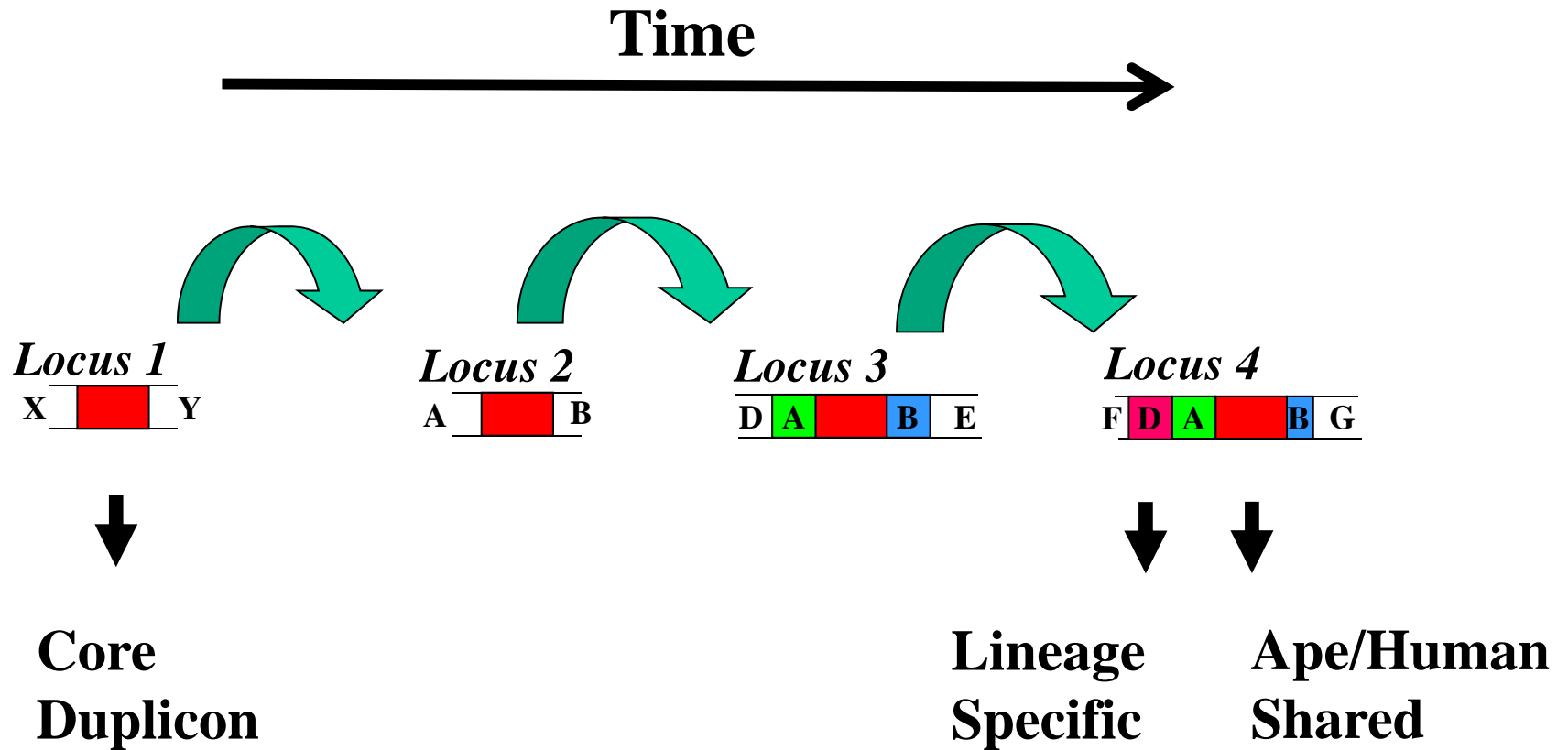
•Most of the recurrent genomic disorders associated with developmental delay, epilepsy, intellectual disability, etc. are mediated by duplication blocks centered on a core.

# Increasing Duplication Complexity and Recurrence



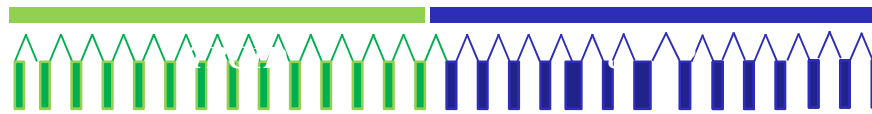
- Duplication blocks have become increasingly more complex (more duplicons) and have expanded in an interspersed fashion over the last 25 million years.
- Duplication blocks of different flanking content with exception of core

# Core Expansion Model



# Human Great-ape “Core Duplicons” have led to the Emergence of New Genes

*TRE2*

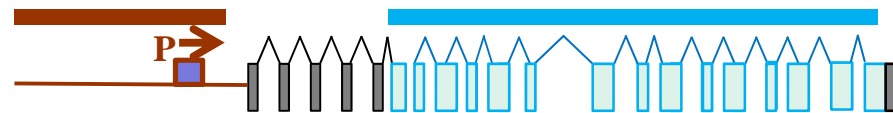


*NPIP*

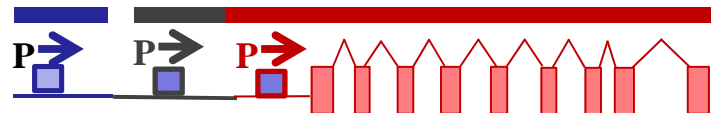


*EVI5*

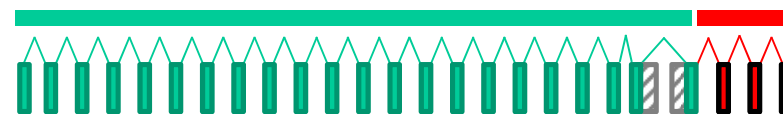
*NBPF*



*LRRC37A*



*RGPD*



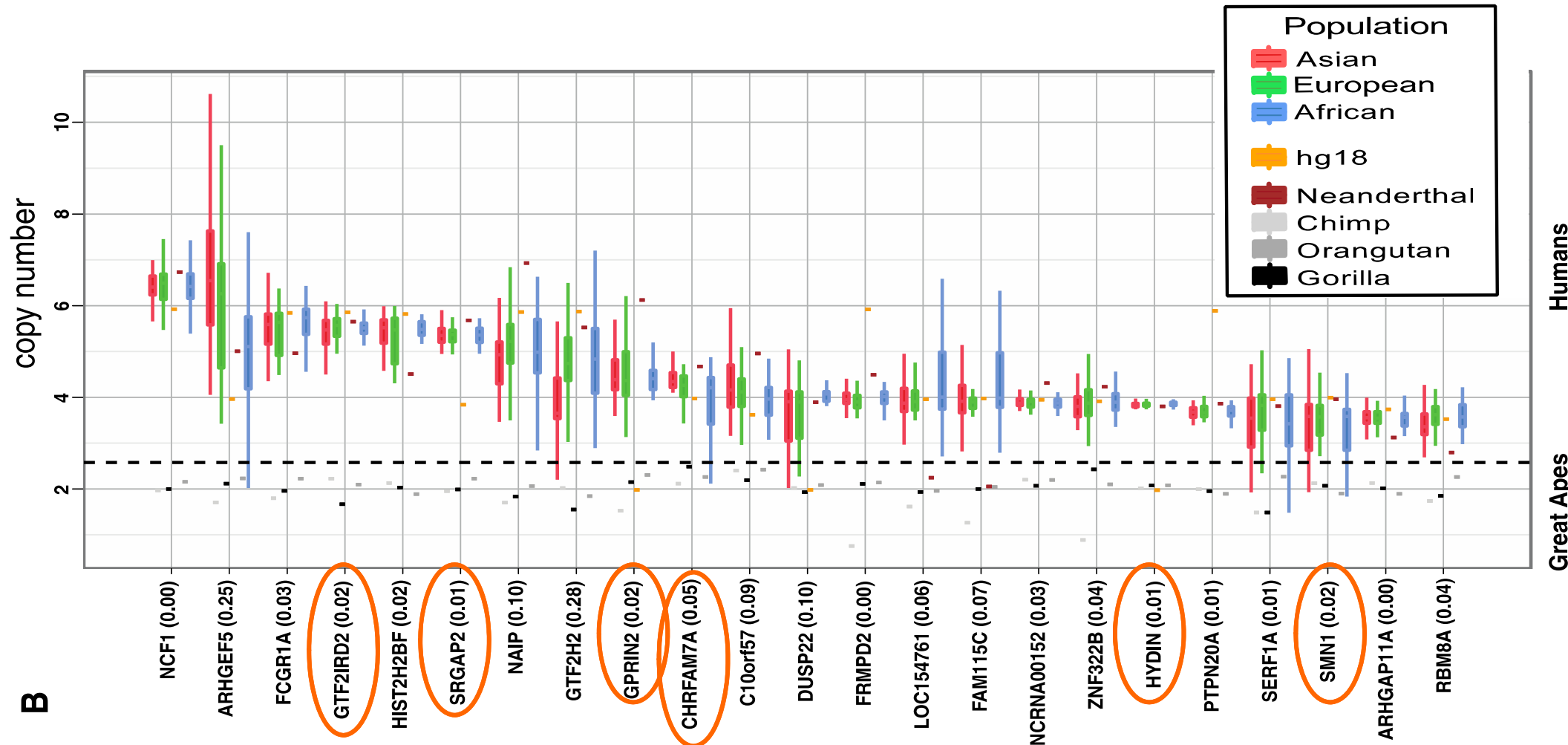
**Features: No orthologs in mouse; multiple copies in chimp & human  
dramatic changes in expression profile; signatures of positive selection**

# Core Duplicon Hypothesis

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

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

# Human-specific gene family expansions



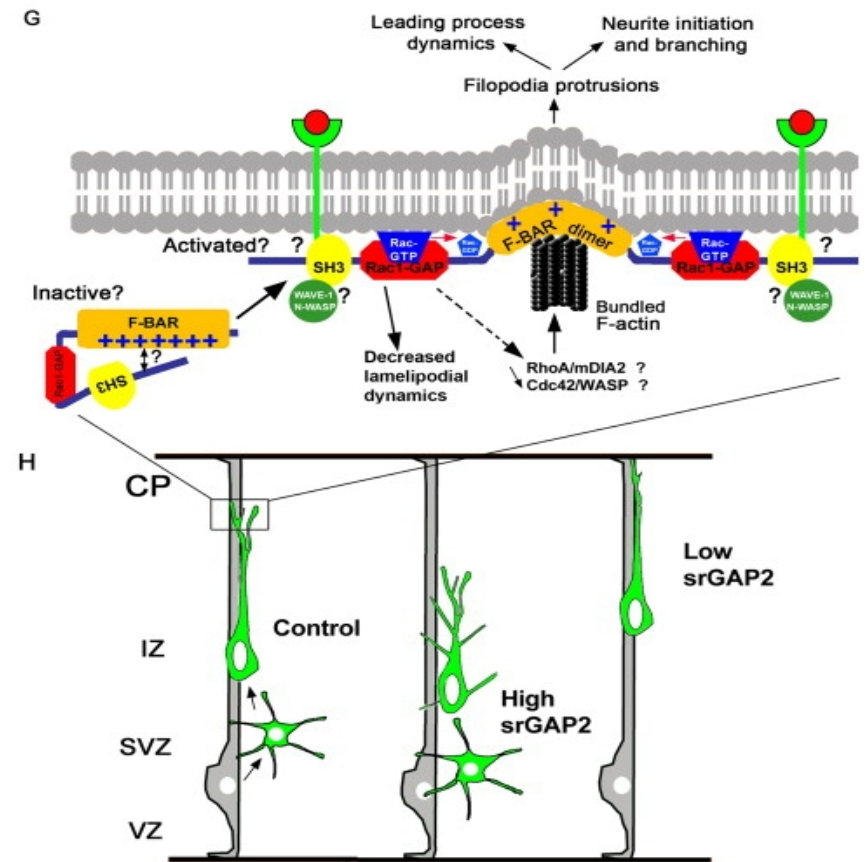
**Notable human-specific expansion of brain development genes.**

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

**Sudmant et al., *Science*, 2010**

# SRGAP2 function

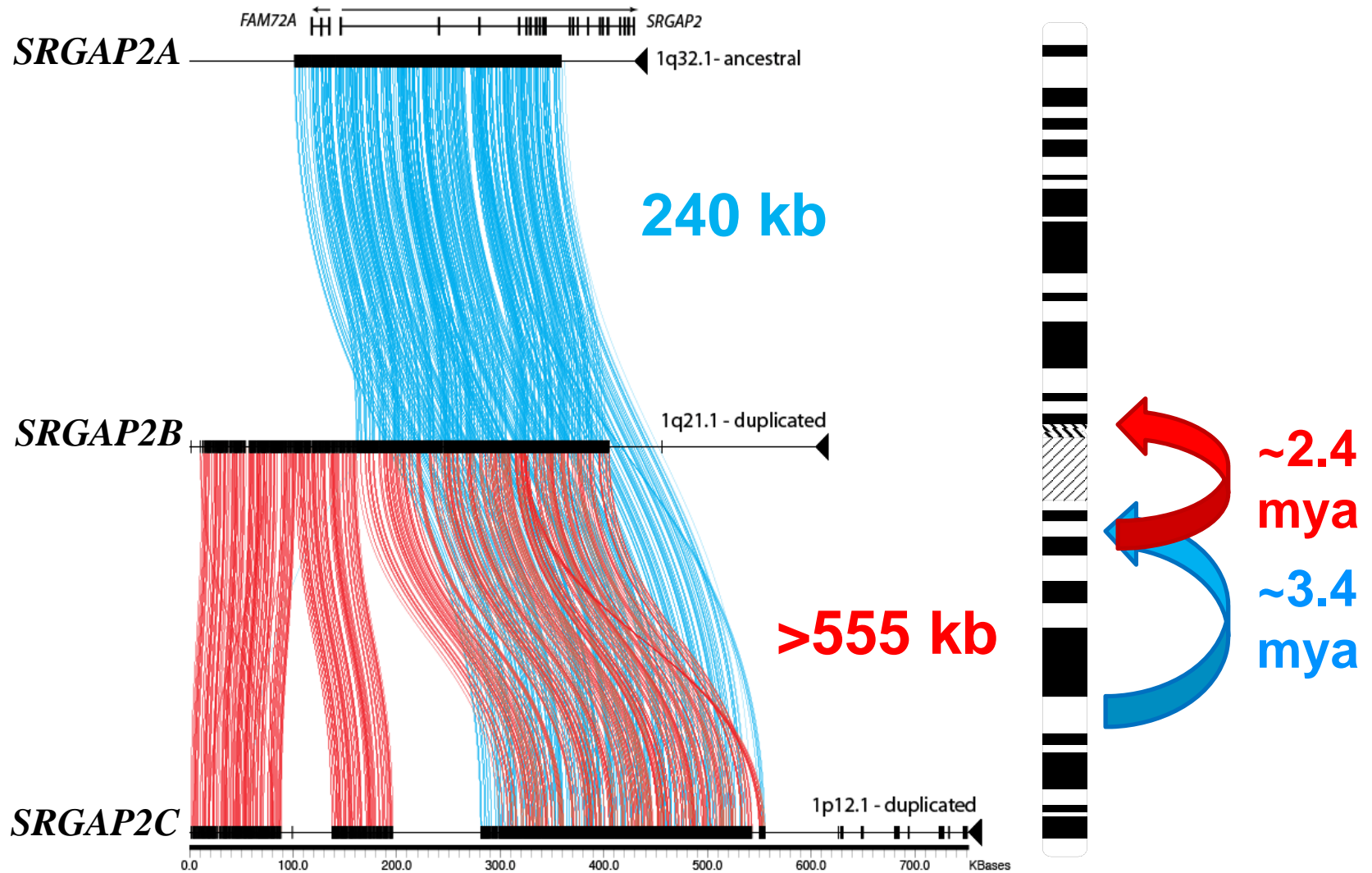
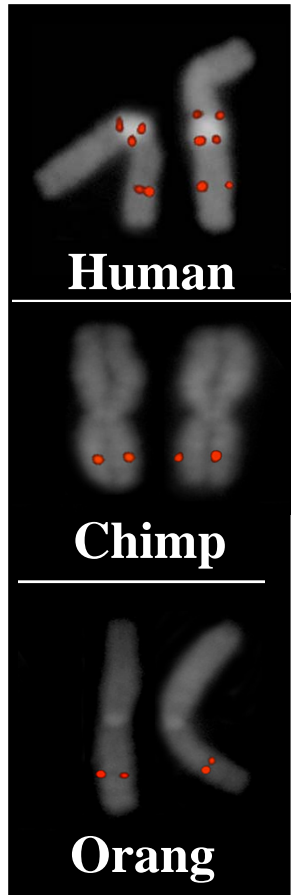
- *SRGAP2* (SLIT-ROBO Rho GTPase activating protein 2) functions to control migration of neurons and dendritic formation in the cortex
- Gene has been duplicated three times in human and no other mammalian lineage
- Duplicated loci not in human genome



Guerrier et al., *Cell*, 2009



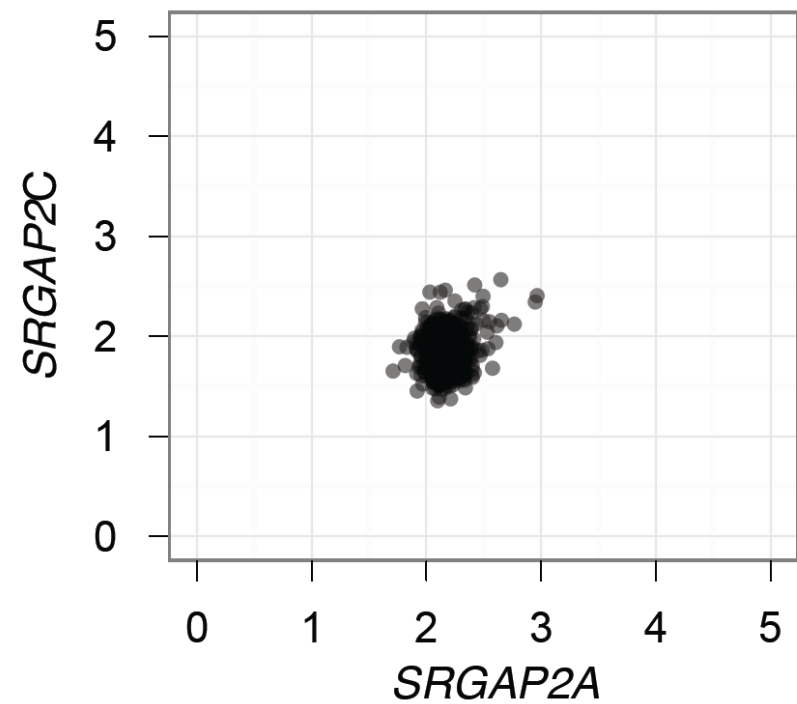
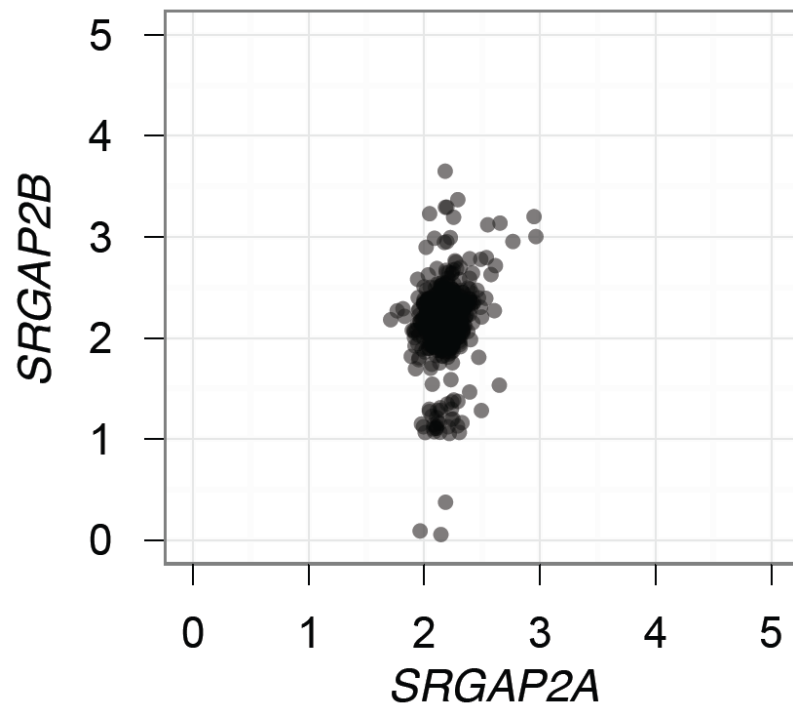
# SRGAP2 Human Specific Duplication



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

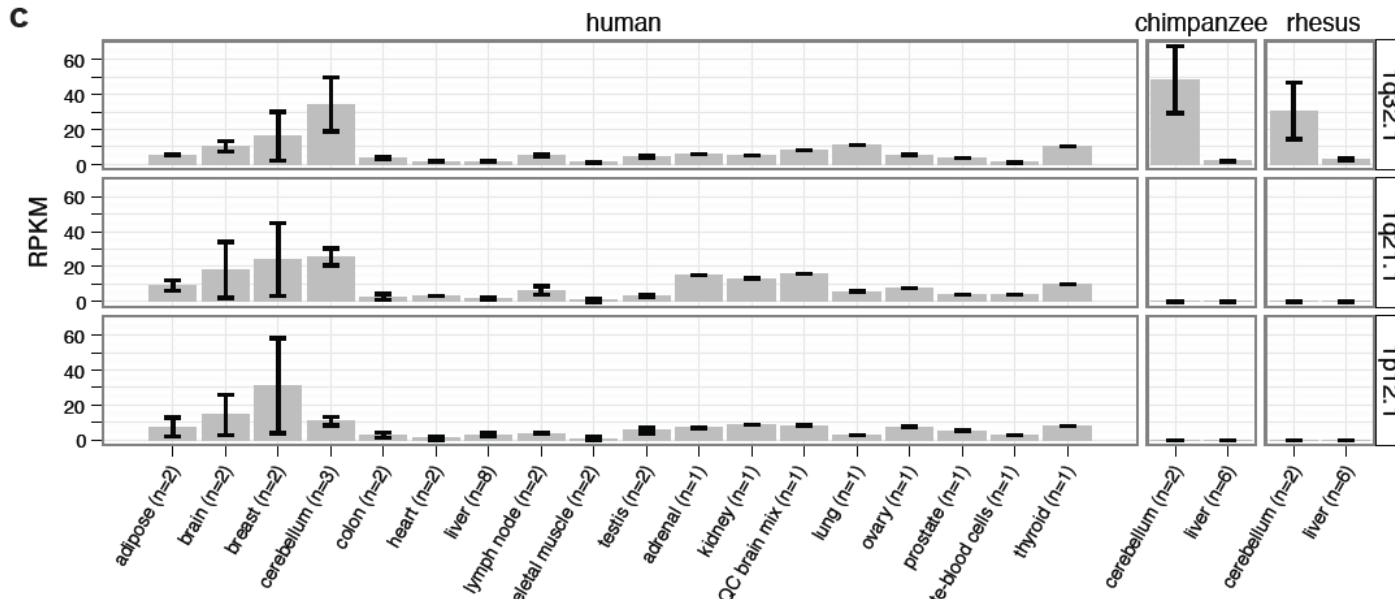
# SRGAP2C is fixed in humans

(n=661 individual genomes)

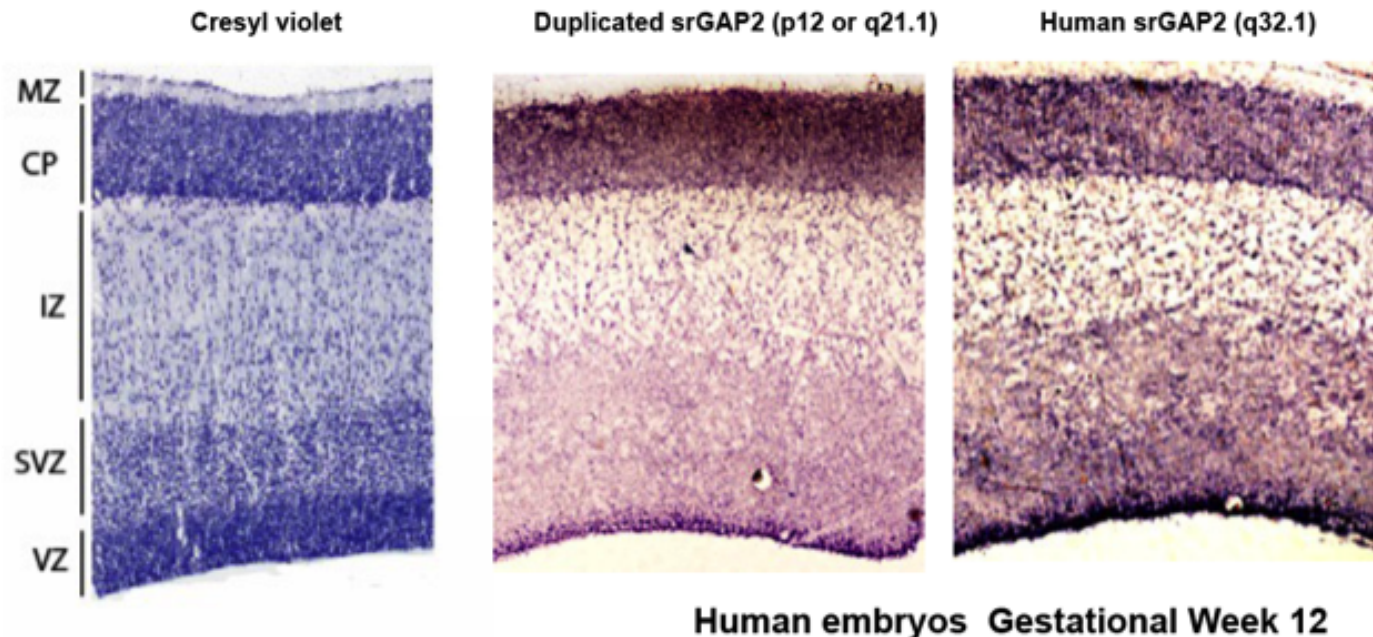


# SRGAP2 duplicates are expressed

**RNAseq**

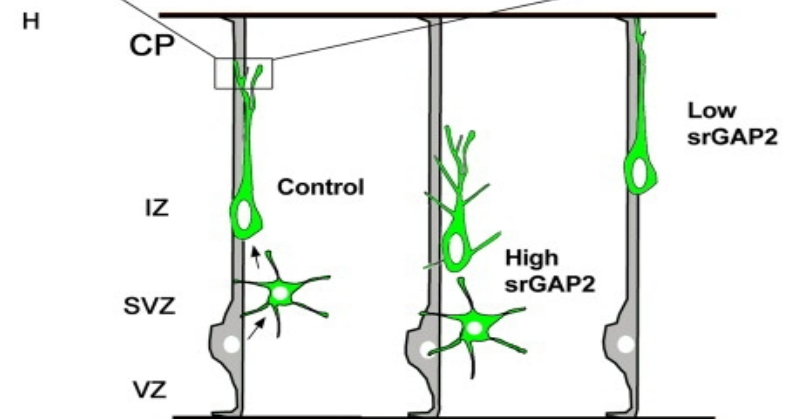
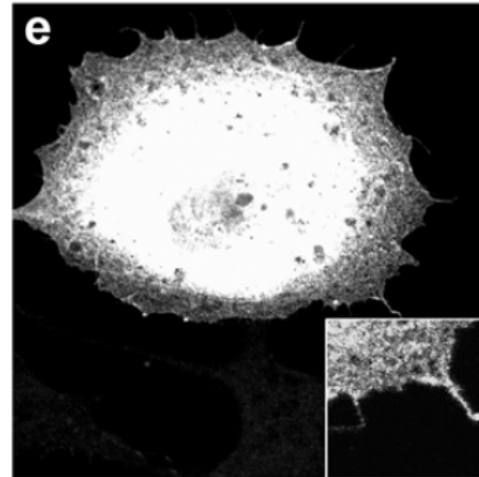
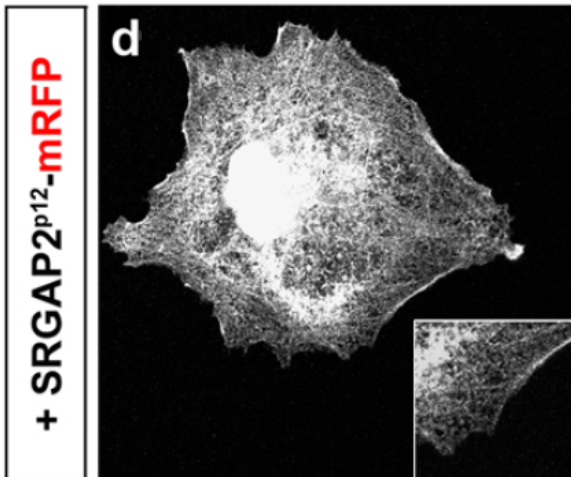
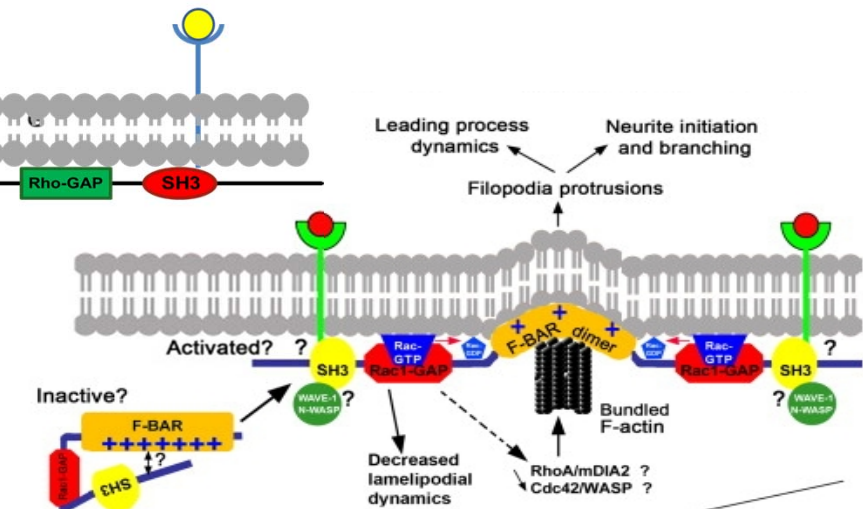
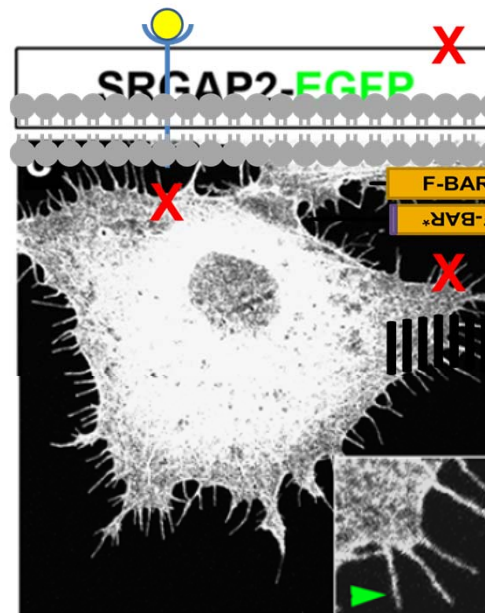
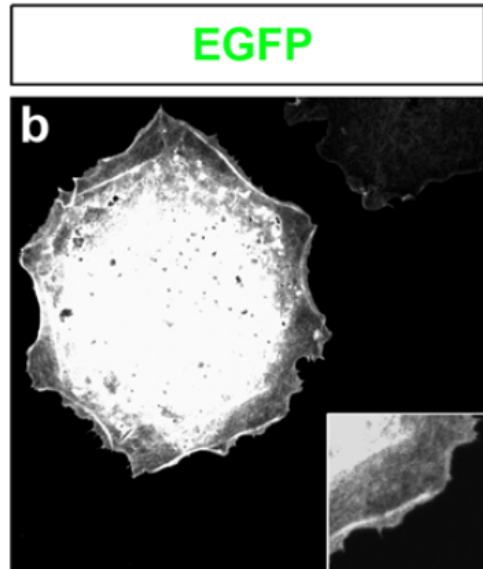


*In situ*



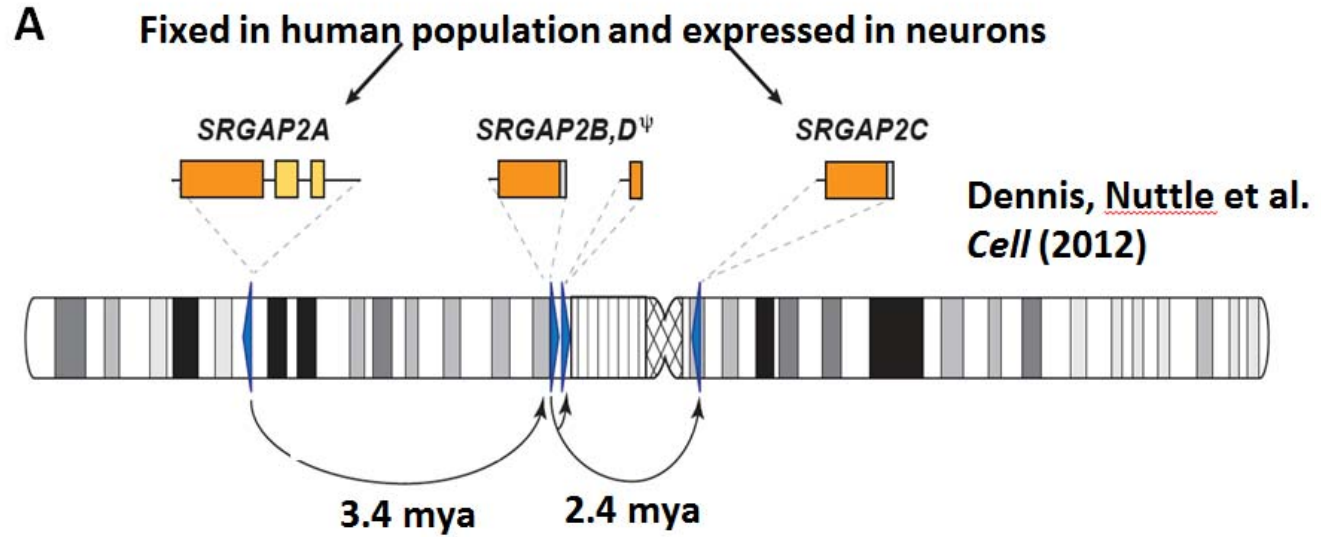
Human embryos Gestational Week 12

# SRGAP2C duplicate antagonizes function

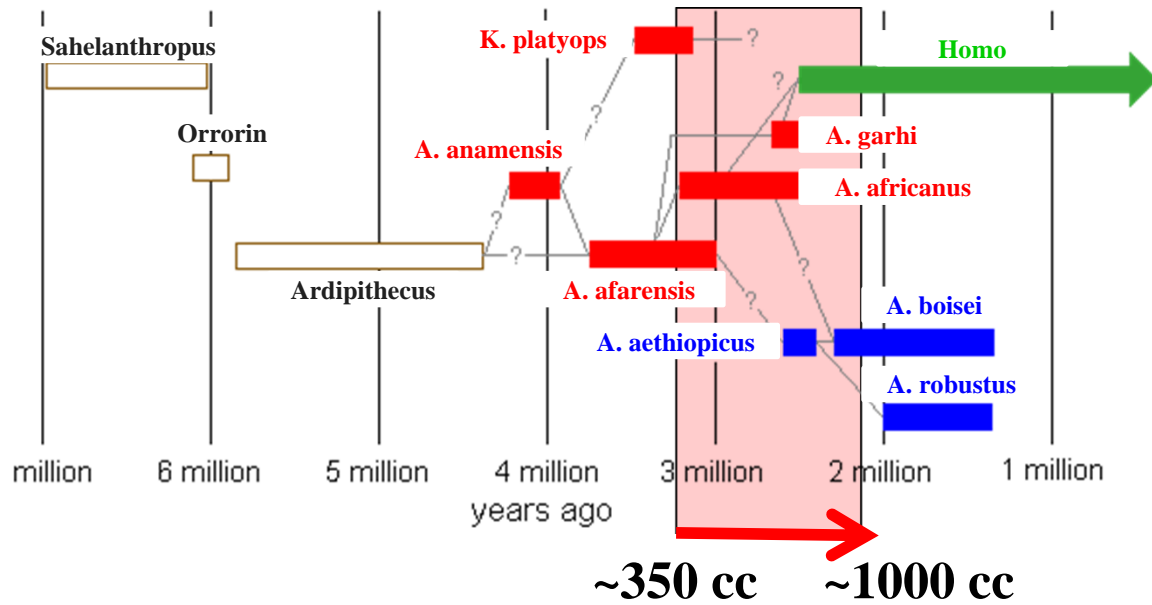


Charrier et al., *Cell*, 2012





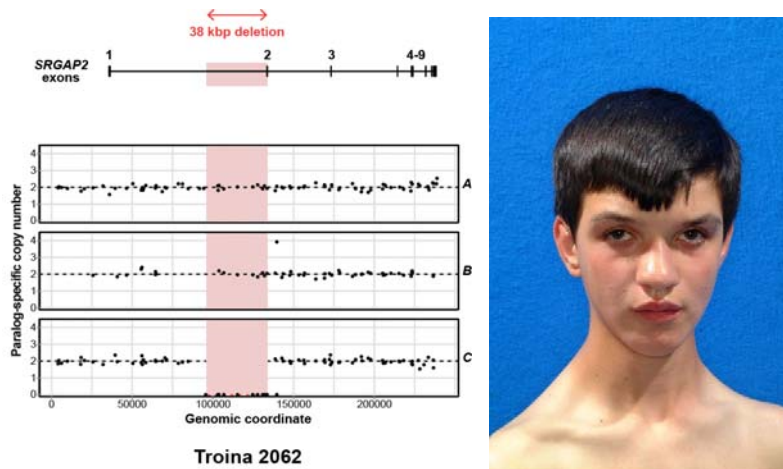
*Australopithecus*



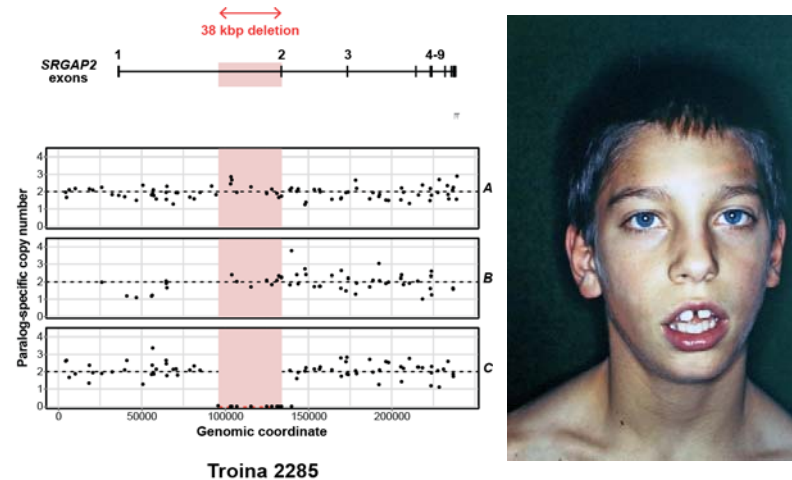
*Homo habilis*

# Homozygous Deletions of *SRGAP2C*

- 5/2711 patients with ID vs. 0/740 controls



- Severe intellectual disability
- Microbrachycephaly
- Orbitofrontal cortex 52.7 cm at age 18 (-2.3 standard deviations from mean)
- Also has inherited 7q11.22 microdeletion including *AUTS2*



- Moderate intellectual disability
- Microcephaly
- Orbitofrontal cortex 53 cm at age 29 (-2.1 standard deviations from mean)
- Partial agenesis of corpus callosum

Nuttle X, unpublished

# 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—two models common and rare—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



# SD-Mediated Rearrangements

