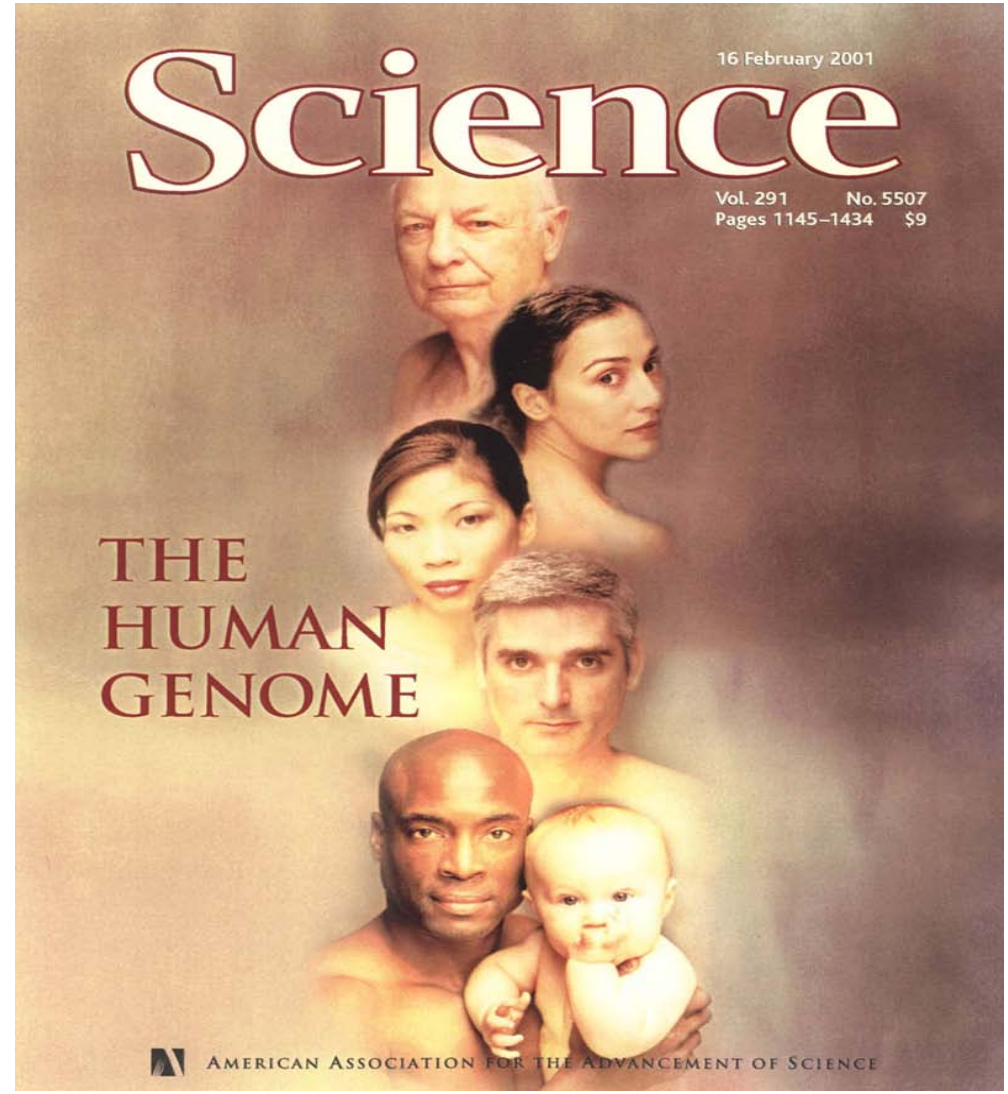
Genome Structural Variation Disease and Evolution

Evan Eichler
Howard Hughes Medical Institute
University of Washington

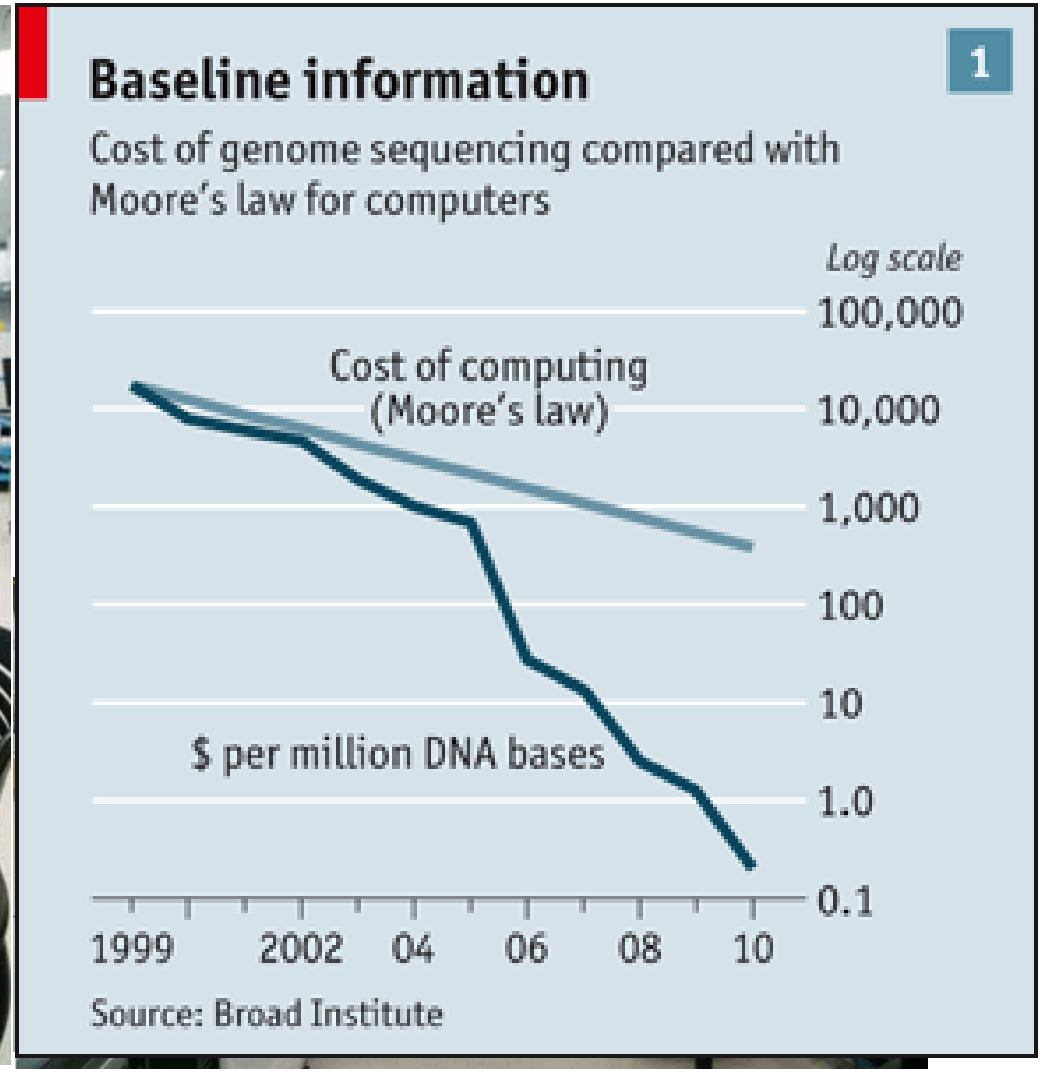
July 11th, 2011, Comparative Genomics, Fort Collins

A Watershed Moment in Genomics



February 15, 2001

Next Generation Sequencing Revolution



- **Genome Sequencing 30-fold coverage $4500/genome**

Bioinformatics: The Challenge

Genetic Variation

Types.

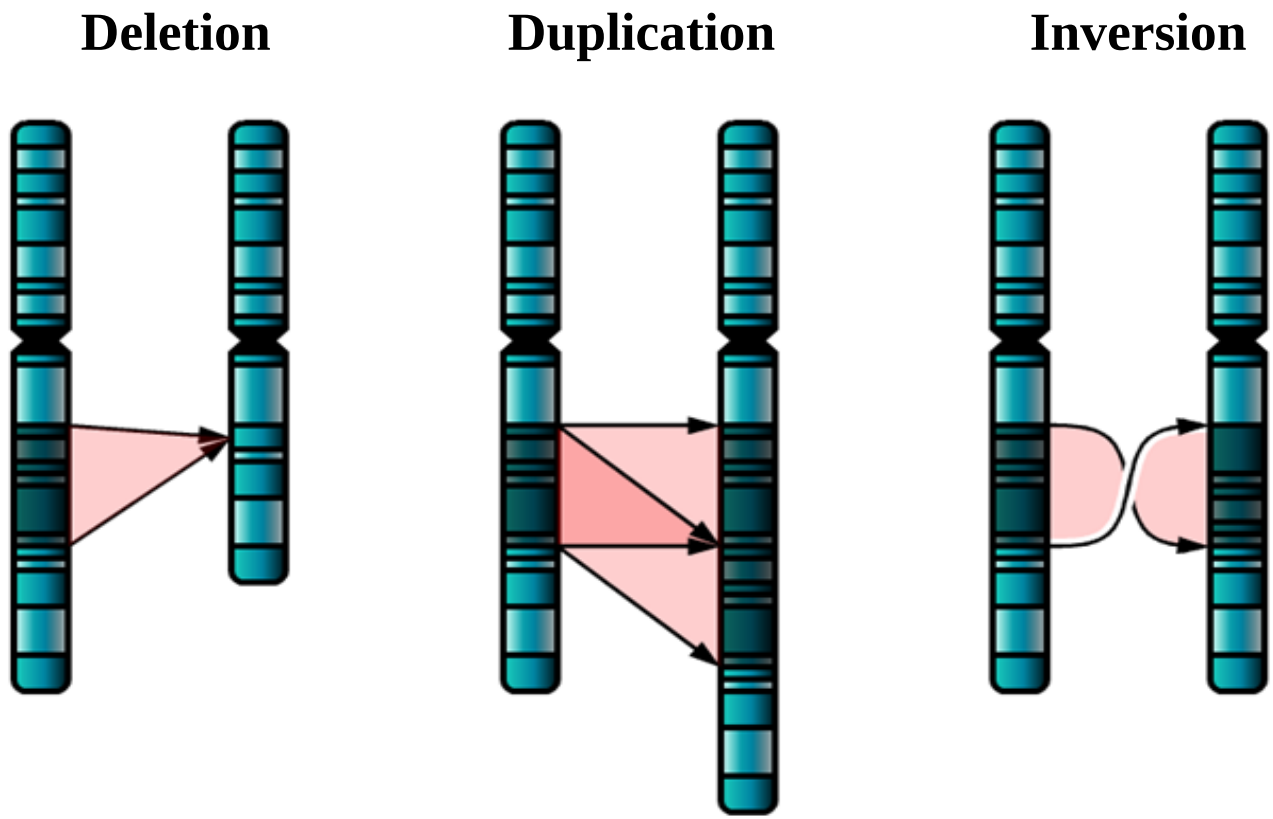
Sequence

- 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

- **Definition:** copy-number variation (CNV) and balanced rearrangement events such as inversions and translocations—originally defined as > 1 kbp but now >50 bp



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

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

NCE

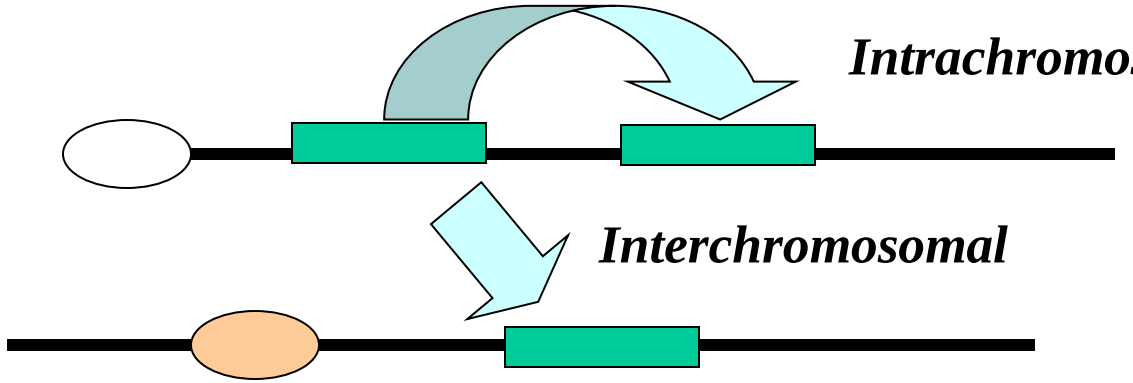
SCIENCE VOL 316 20 APRIL 2007

Objectives

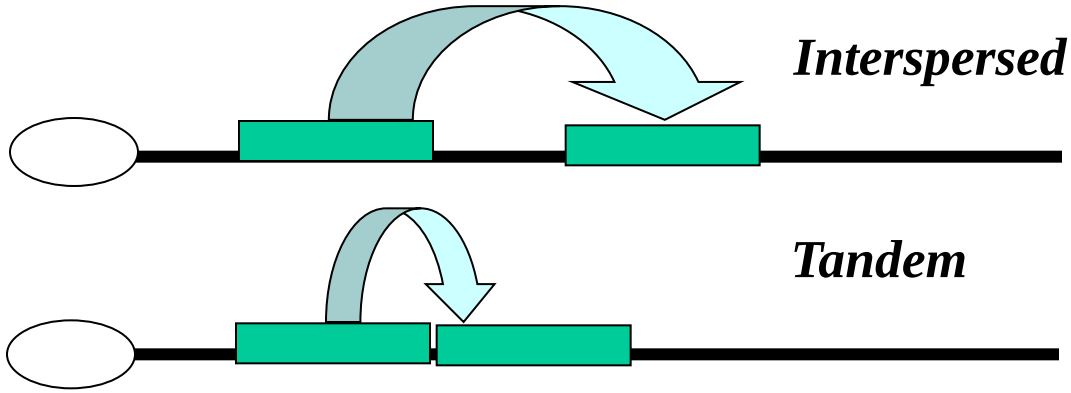
1. Genomic architecture and disease impact.
2. Sequence-based detection methods
3. Primate genome evolution and the core duplicon hypothesis.

Perspective: Segmental Duplications

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

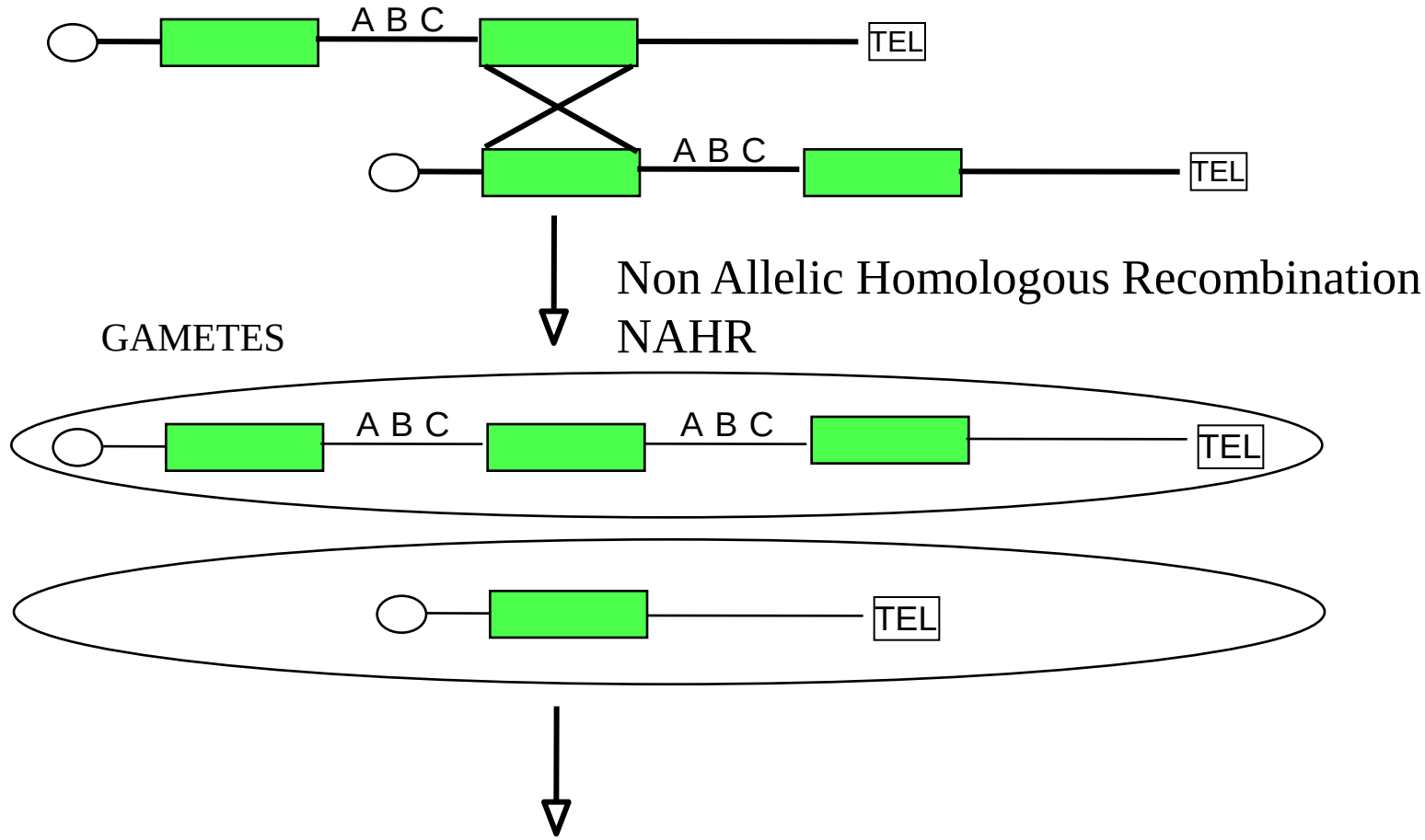


Distribution



Configuration

Importance: Structural Variation



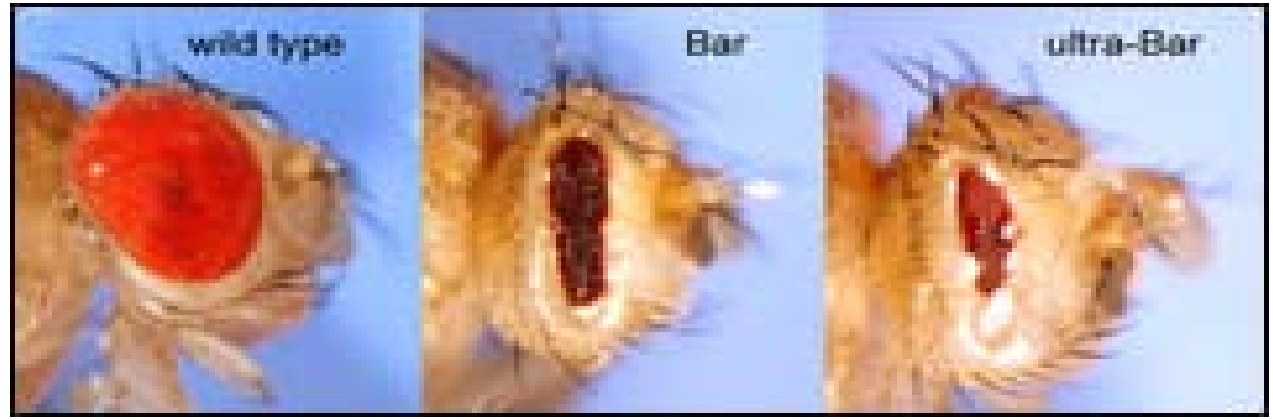
Human Disease

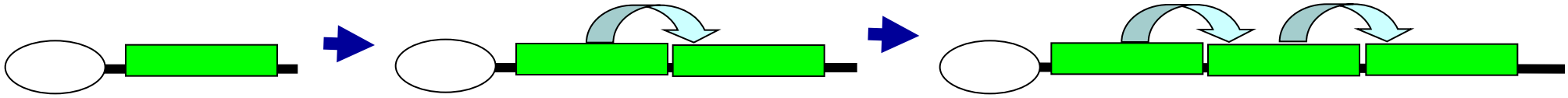
Triplosensitive, Haploinsufficient and Imprinted Genes

Calvin Bridges

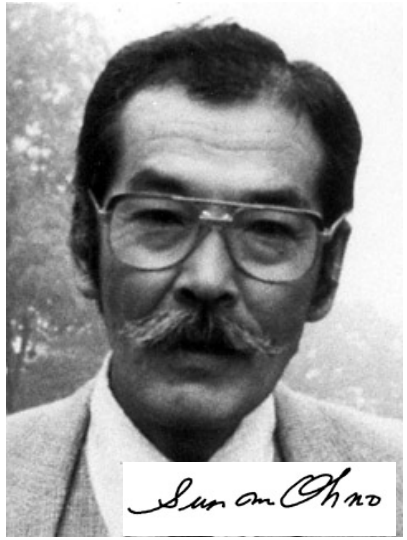
Alfred Sturtevant

H.J. Mueller





Importance: Evolution of New Gene Function



GeneA

Duplication

GeneA'

Acquire New/
Modified Function

Mutation

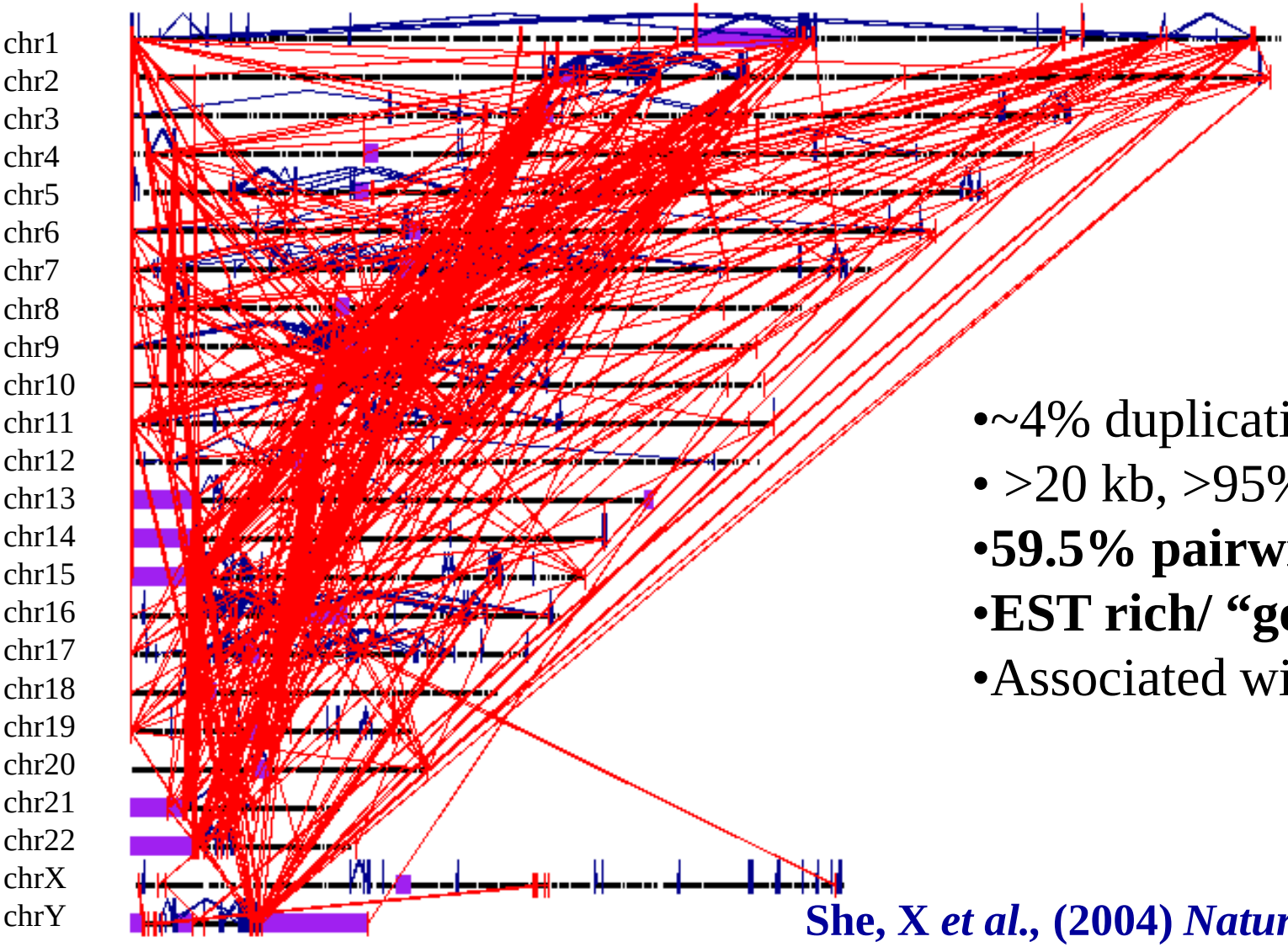
Mutation

Maintain old
Function

Mutation

Loss of Function

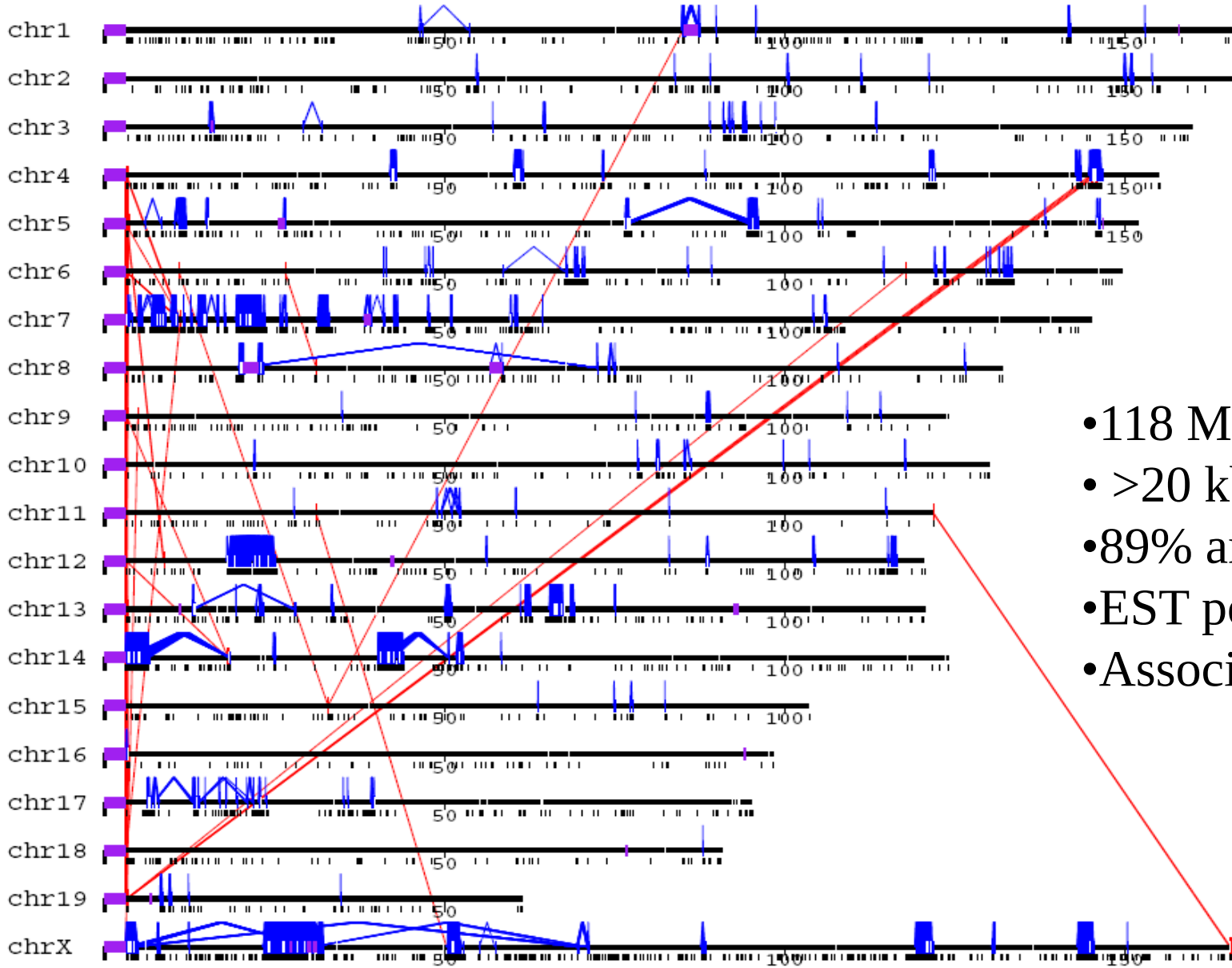
Human Genome Segmental Duplication Pattern



- ~4% duplication (125 Mb)
- >20 kb, >95%
- **59.5% 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

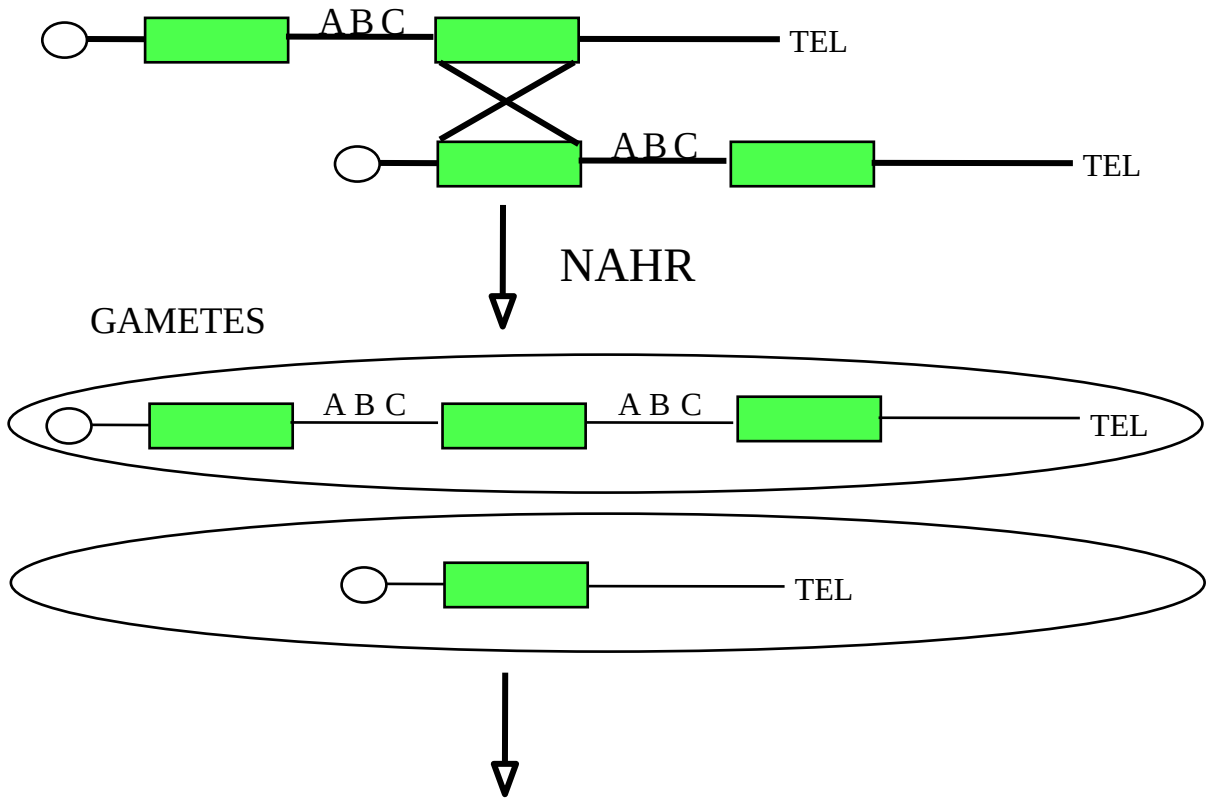
She, X *et al.*, (2008) *Nature Genetics*

Human Segmental Duplications Properties

- Large (>10 kb)
- Recent (>95% identity)
- Interspersed (60% are separated by more than 1 Mb)
- Modular (duplicon architecture) ~389 acceptor regions
- 2.7% Genetic Difference, human vs. chimpanzee

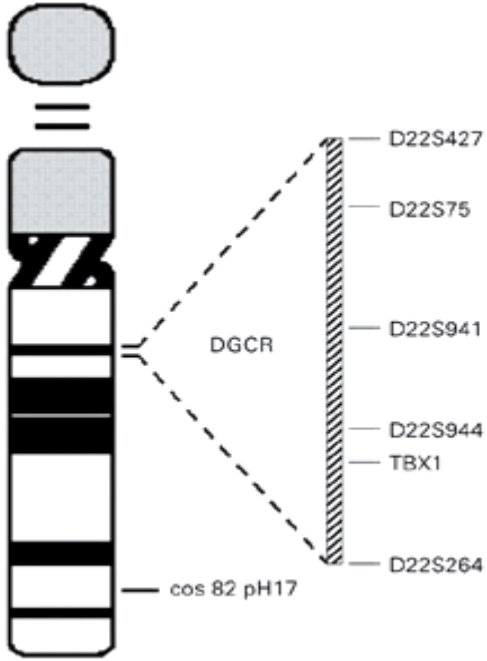
What impact in terms of human variation?
Is it still ongoing?

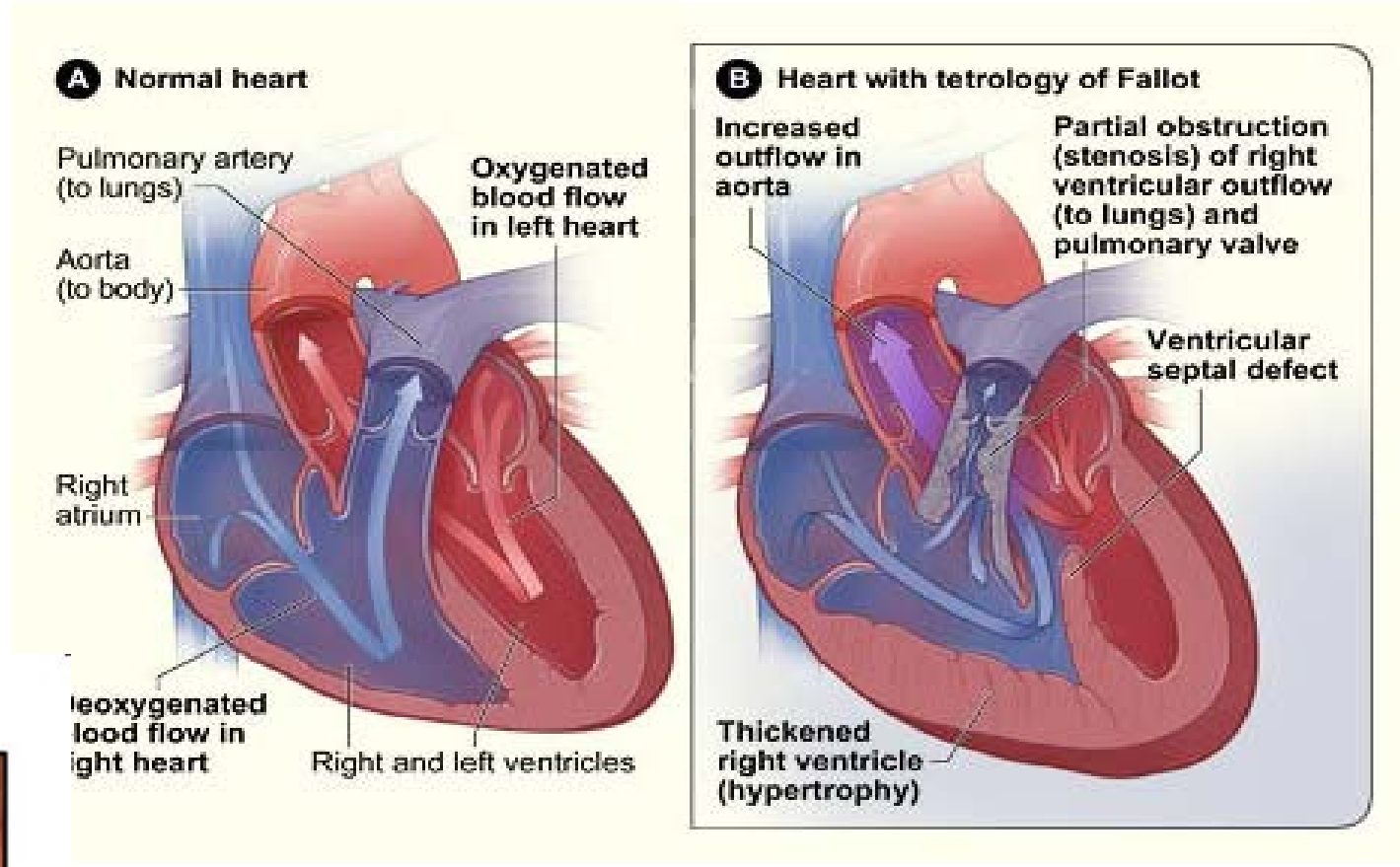
Model of Genomic Disease/Variation

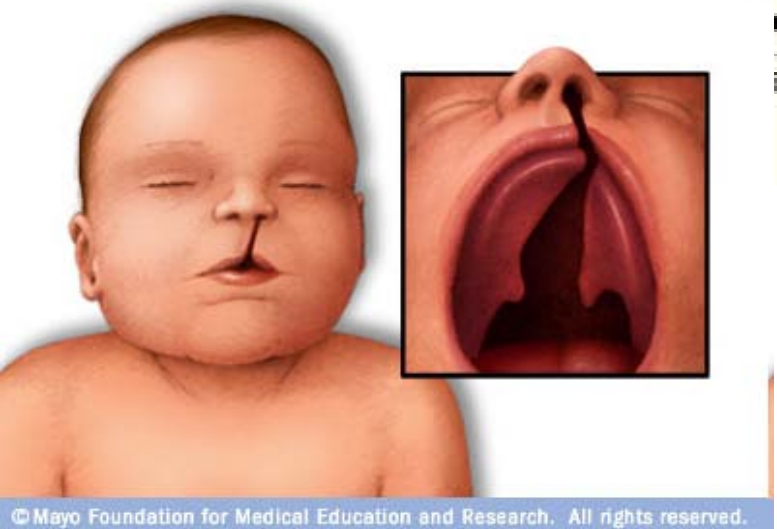


•**Genomic Disorders:** A group of diseases that results from genome rearrangement mediated mostly by non-allelic homologous recombination**. (***Inoue & Lupski , 2002***).**

DiGeorge/VCFS/22q11 Syndrome







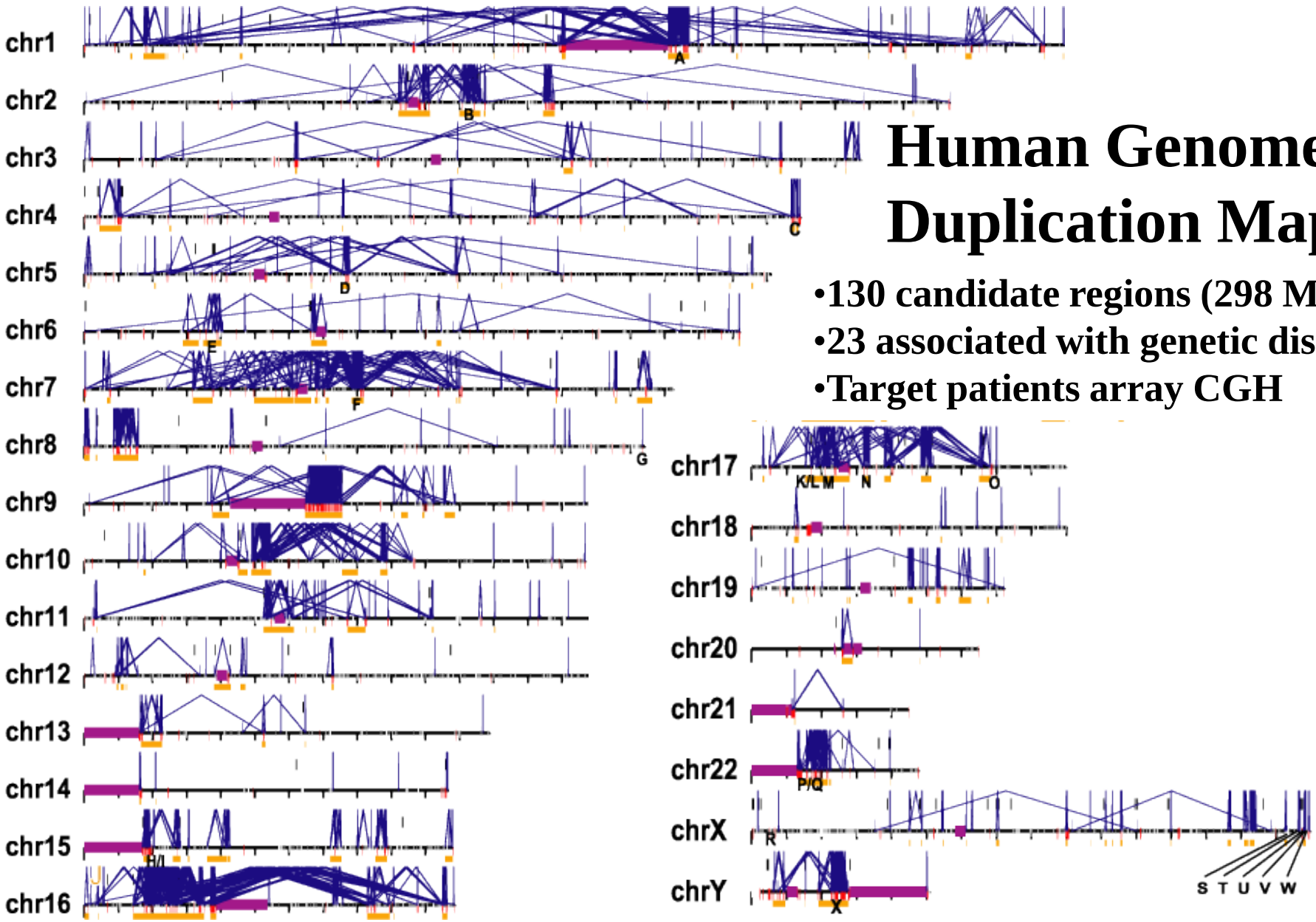
1/2000 live births

180 phenotypes

75-80% are sporadic (not inherited)

Genomic Disorders (mediated by SDs)

Genomic Disorder	Brain	Congenital Anomalies	Locus	Interval kb	LCR size kb	Duplicon	%identity	Incidence (%)	Incidence (MR)
Williams-Beuren syndrome	Severe MR	craniofacial, heart disease	7q11.23	1,600	>320	PMS2/GTFI2	96-99	**0.01**	**0.5**
Prader-Willi syndrome	Severe MR	small hands, feet, hypotonia, obesity, short stature	15q11.2-q13	3500	400	HERC2	92-99	**0.007**	**0.35**
Angelman syndrome	Severe MR	microcephaly, hyoptonia, seizures	15q11.2-q13	3500	400	HERC2	92-99	**0.007**	**0.35**
Smith-Magenis syndrome	Severe MR	crainiofacial, peripheral neuropathy	17p11.2	4000	200	SMSREP	98.2-99	**0.004**	**0.2**
dup17p11.2	mild MR	peripheral neuropathy	17p11.2	4000	200	SMSREP	98.2-99	**0.001**	**0.05**
Velocardiofacial syndrome	mild MR	cardiac, craniofacial defects	22q11.2	~3000	~300	LCR22	98-99	**0.03**	**0.7**
Cat Eye Syndrome	Severe MR	craniofacial,colobo ma	22q11	3000	400	LCR22	98-99	**0.003**	**0.15**
Inv dup(15)	Mild/Severe	mild facial, seizures	15q11/q14	4000	400	HERC2	98	**0.01**	**0.5**
Neurofibromatosis	Mild MR	fibromatous tumours, visual defects	17q11.2	1500	85	NF1REP	98.4	**0.003**	**0.03**
CMT1A	no MR	peripheral neuropathy	17p12	1400	24	CMT1A-REP	98.7	**0.01**	**NA**
HNPP	no MR	peripheral neuropathy	17p12	1400	24	CMT1A-REP	98.7	**0.001**	**NA**
								0.089	**2.80%**



Human Genome Duplication Map

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

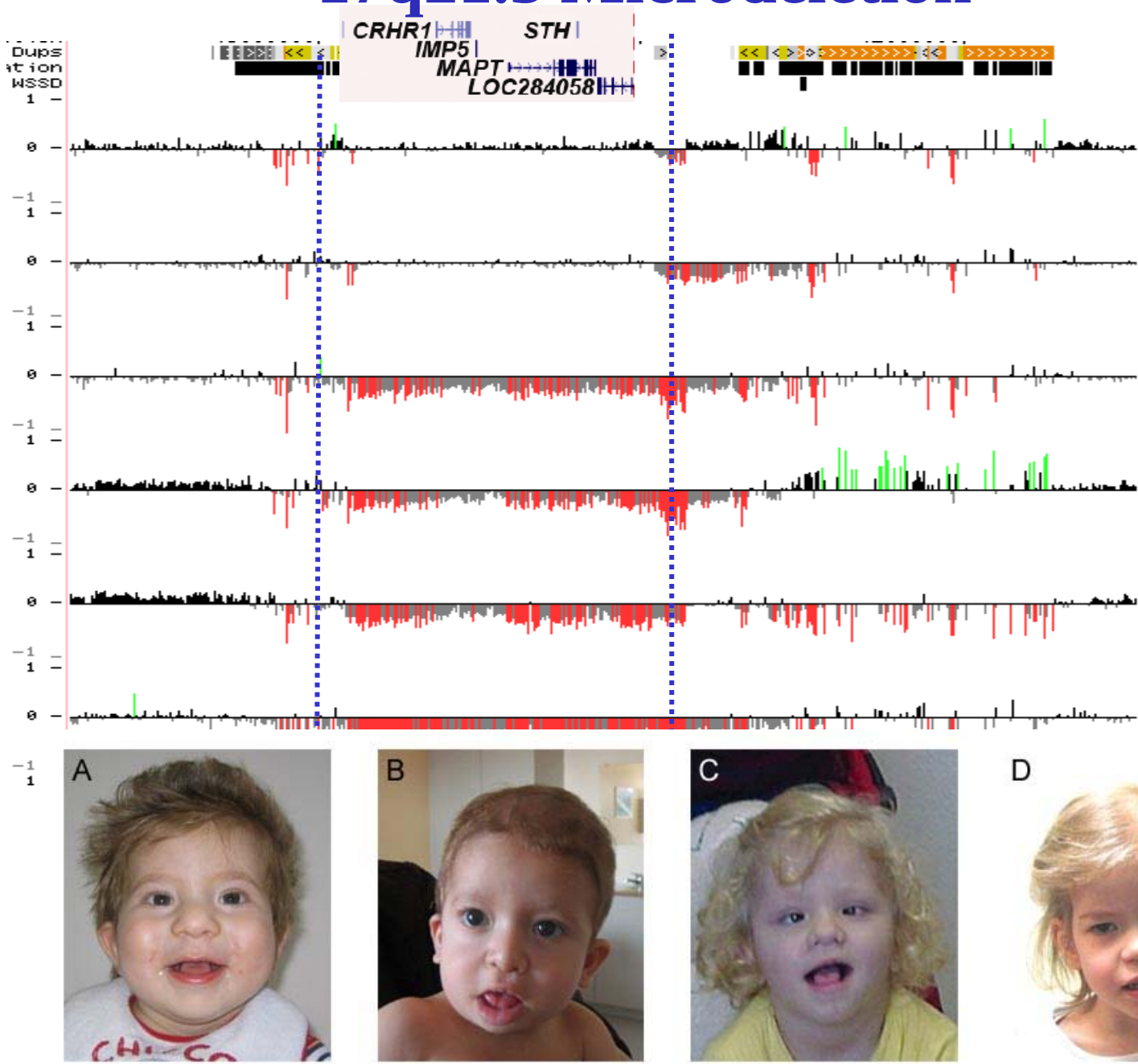
Bailey *et al.* (2002), *Science*

Large CNVs and Neurocognitive Diseases

- Pediatric intellectual disability (or idiopathic mental retardation) ~15-17% of cases (>500 kbp)
- Simplex autism—10% (10-fold when compared to normal controls)
- Schizophrenia—increase in frequency of gene disruptive CNVs when compared to controls
- 30-40% of events are mediated by SegDups
- Most pathogenic events that have been identified show breakpoints within SD (eg. 16p11.2, 15q13.3 and 17q21.3), however, there are several hundred singleton occurrences

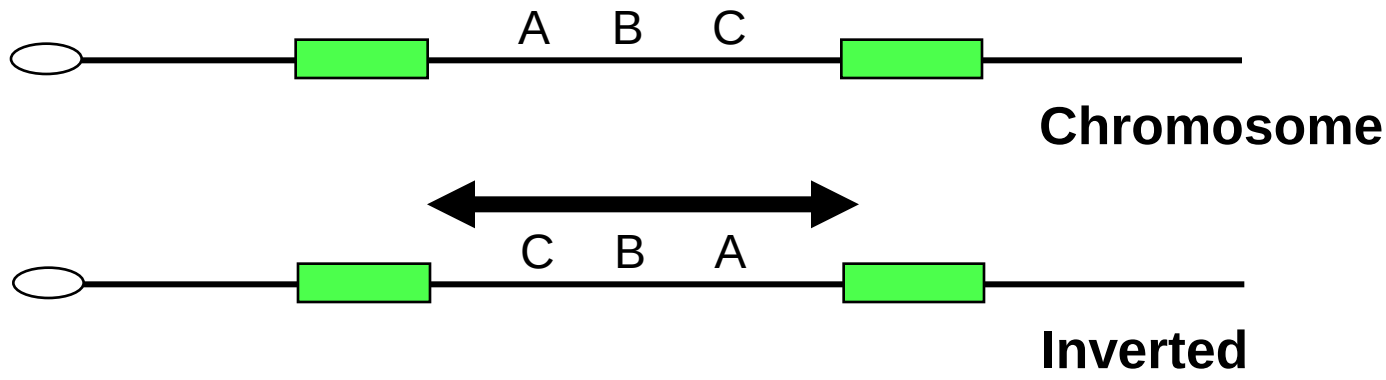
DeVries, *AJHG*, 2005; Sharp, Nat. Genet, 2006; Sebat, Science, 2007; Walsh, Science, 2008, Hilgers-Ropers Review *COGD*, 2008

17q21.3 Microdeletion



Sharp, A *et al*; Koolen, D *et al*; Shaw-Smith *et al*; (2006) *Nature Genetics*

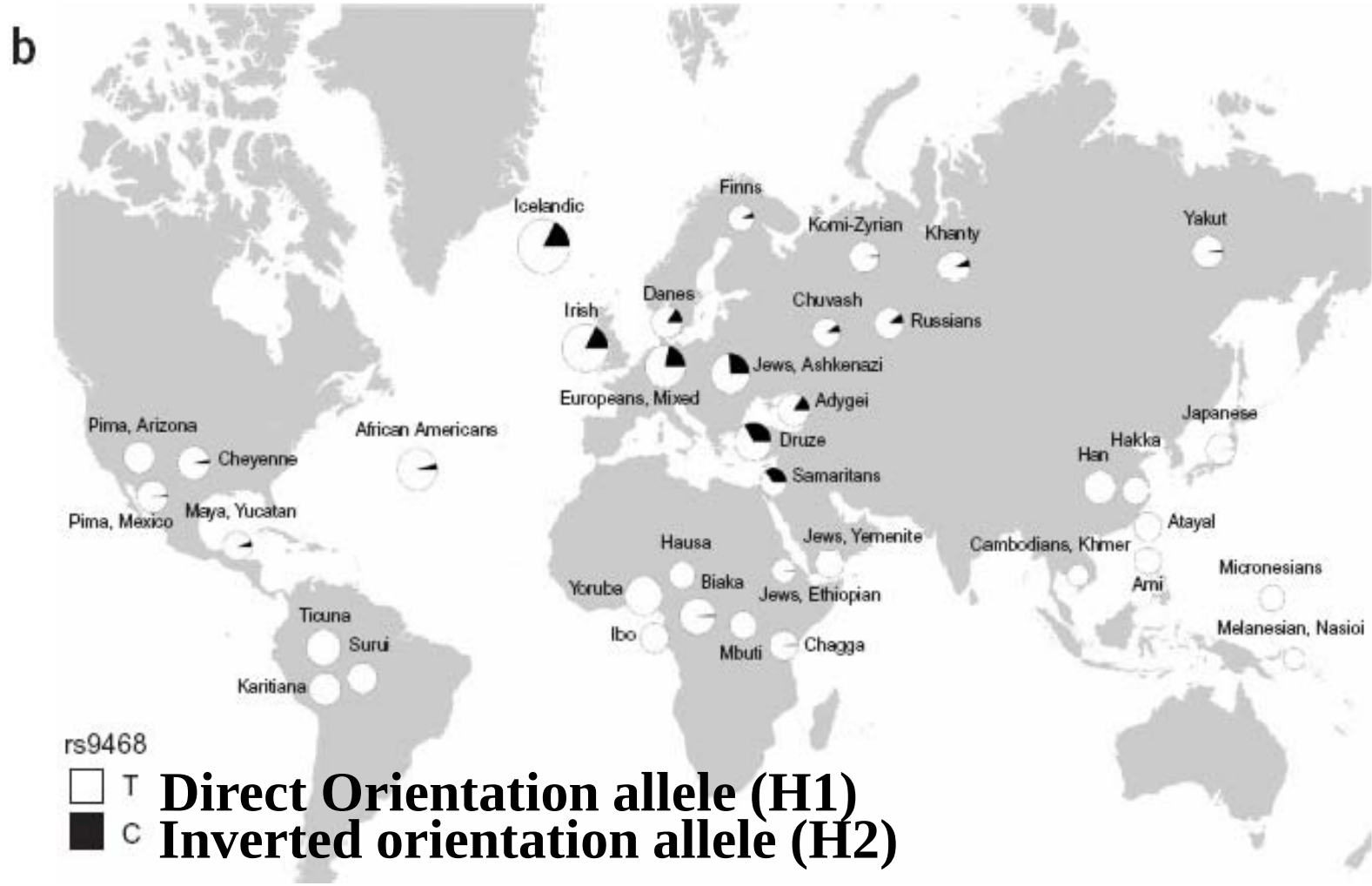
17q21.31 Inversion



- Region of spontaneous 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 under selection in Europeans

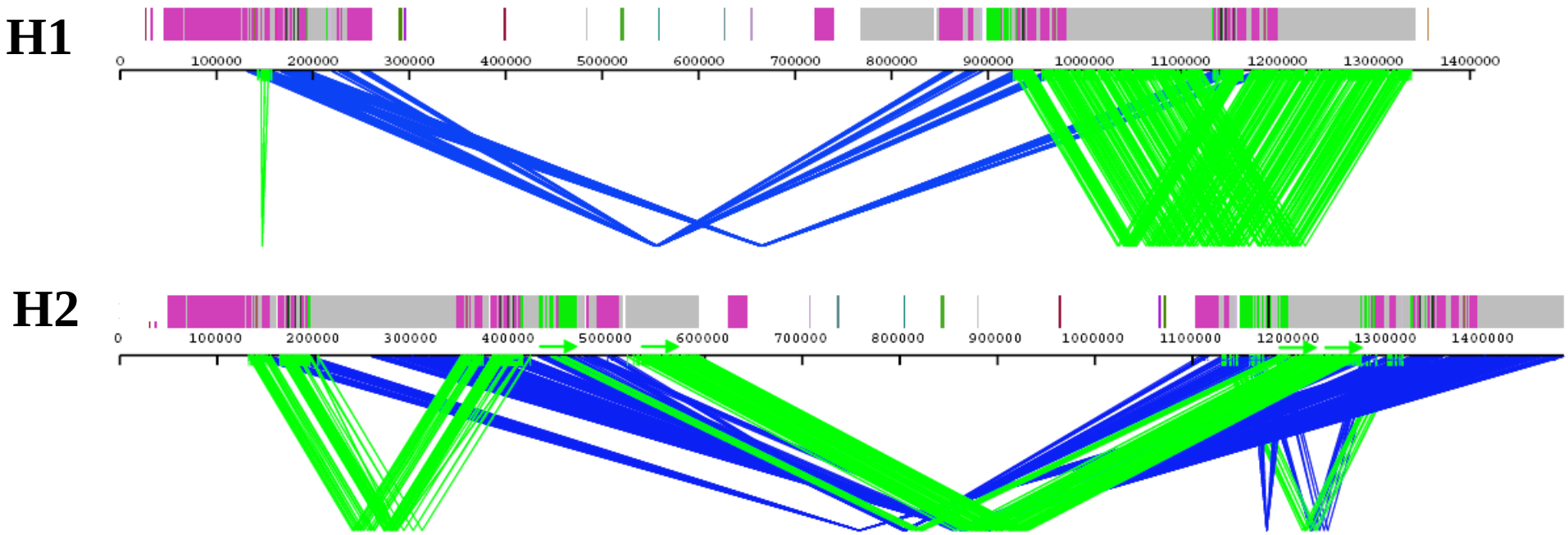


Direct Orientation allele (H1)

Inverted orientation allele (H2)

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

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



- H1 Flanking SD: 169 kbp, seq ID=98.3%, 0 kbp in direct
- H2 Flanking SD: 441 kbp, seq ID=99.3%, 95 kbp in direct
- H2 duplication architecture is a “premutation” haplotype

Zody et al., Nat. Genet, 2008

15q13.3 Recurrent Microdeletion

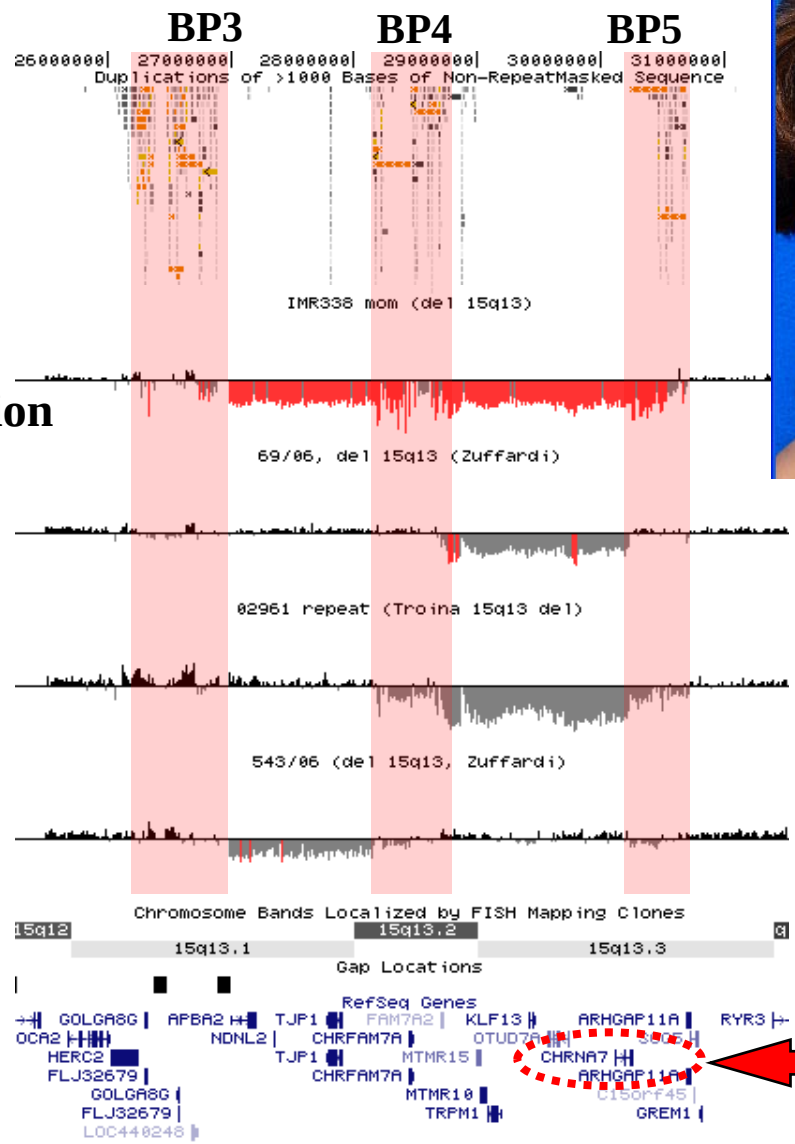
Segmental Duplications

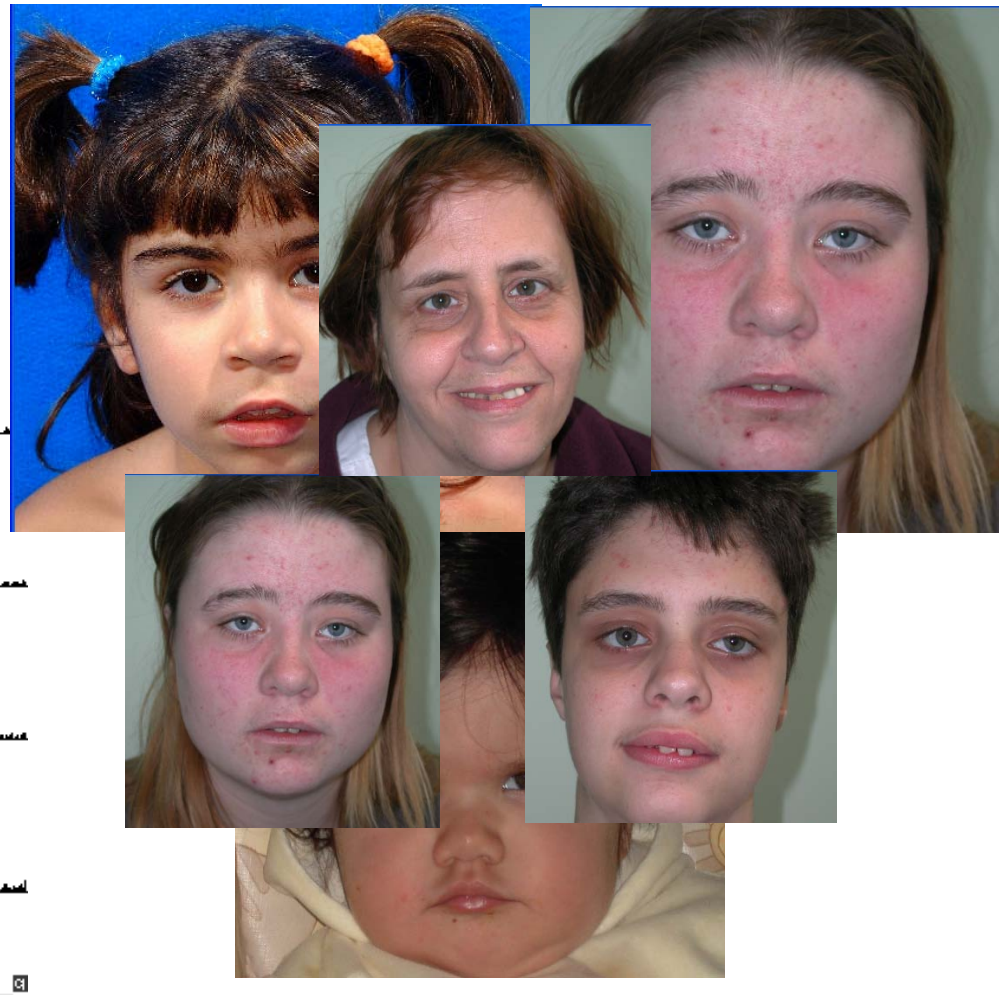
15q13.1-13.3 deletion

15q13.3 deletion

15q13.3 deletion

15q13.1 deletion



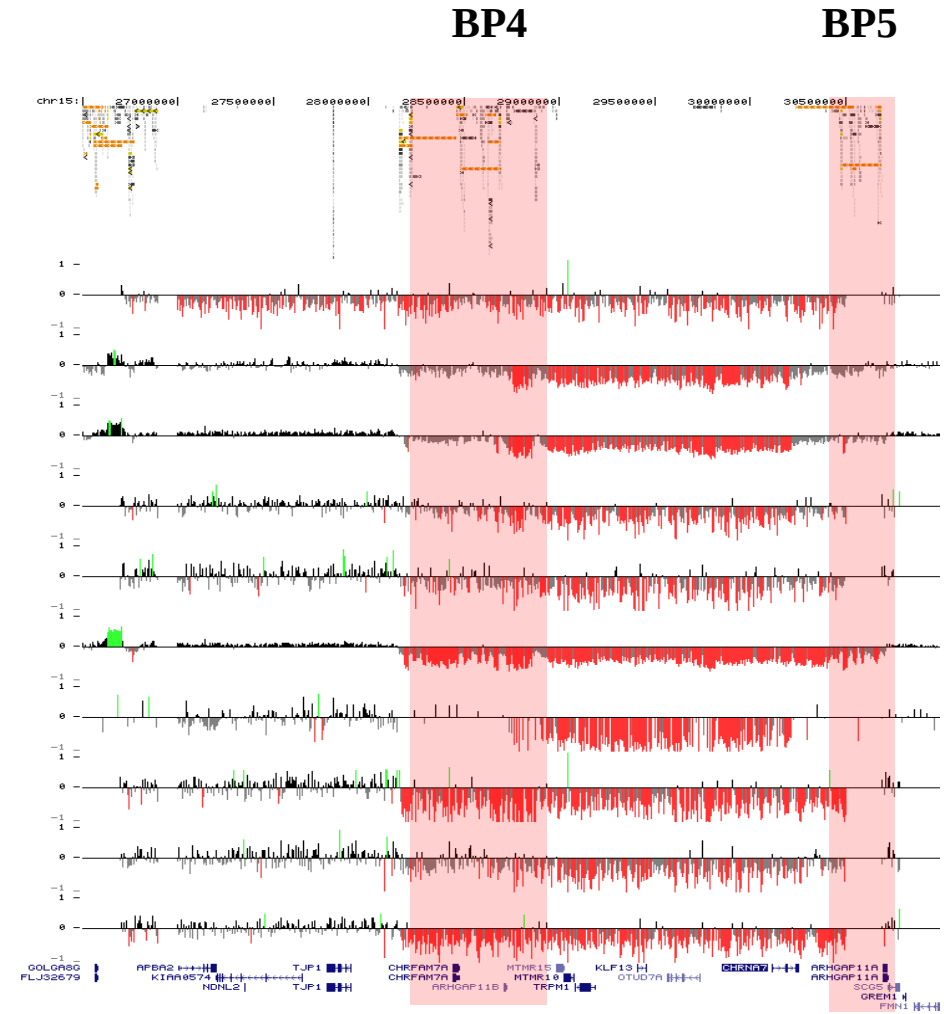


- **Three deletion patients presented with epilepsy/abnormal EEG**
- ***CHRNA7* (gated ion channel protein) linkage to myoclonic epilepsy**

Sharp *et al.*, *Nat. Genet, 2008*

Epilepsy and 15q1.3.3 Microdeletion

- Idiopathic generalized epilepsy without mental retardation—screened two independent cohorts of North-Western European ancestry (12/1,223 IGE patients and 0/3699 population controls; joint $p=5.32 \times 10^{-8}$).
- 10/12 patients have breakpoints consistent with BP4-BP5 segmental duplication mediated rearrangement that was identified in children with mental retardation.



Helbig et al.,*Nat. Genet*, 2009.

Schizophrenia and 15q13.3 Microdeletion

Large recurrent microdeletions associated with schizophrenia

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], Thomas Hansen[8,9], Klaus D. Jakobsen[8,9], Pierandrea Muglia[10], Clyde Francks[10], Paul M. Matthews[11], Arnaldur Gylfason[1], Bjarni V. Halldorsson[1], Daniel Gudbjartsson[1], Thorgeir E. Thorgeirsson[1], Asgeir Sigurdsson[1], Adalbjorg Jonasdottir[1], Aslaug Jonasdottir[1], Asgeir Bjornsson[1], Sigurborg Mattiasdottir[1], Thorarinn Blondal[1], Magnus Haraldsson[6], Brynja B. Magnusdottir[6], Ina Giegling[2], Hans-Jürgen Möller[2], Annette Hartmann[2], Kevin V. Shianna[12], Dongliang Ge[12], Anna C. Need[12], Caroline Crombie[13], Gillian Fraser[13], Nicholas Walker[14], Jouko Lonnqvist[15], Jaana Suvisaari[15], Annamarie Tuulio-Henriksson[15], Tiina Paunio[5,15], Timi Toulopoulou[16], Elvira Bramon[16], Marta Di Forti[16], Robin Murray[16], Mirella Ruggeri[17], Evangelos Vassos[16], Sarah Tosato[17], Muriel Walshe[16], Tao Li[16,18], Catalina Vasilescu[3], Thomas W. Mühleisen[3], August G. Wang[19], Henrik Ullum[20], Srdjan Djurovic[21,22], Ingrid Melle[22], Jes Olesen[23], Lambertus A. Kiemeney[24], Barbara Franke[25], GROUP†, Chiara Sabatti[26], Nelson B. Freimer[27], Jeffrey R. Gulcher[1], Unnur Thorsteinsdottir[1], Augustine Kong[1], Ole A. Andreassen[21,22], Roel A. Ophoff[7,27], Alexander Georgi[28], Marcella Rietschel[28], Thomas Werge[8], Hannes Petursson[6], David B. Goldstein[12], Markus M. Nöthen[3,4], Leena Peltonen[5,29,30], David A. Collier[16,18], David St Clair[13] & Kari Stefansson[1,31]

***Nature* 455:237-41 2008**

- 7/4,213 cases vs. 8/39,800 controls, OR 11.54 , $p=5.3 \times 10^{-4}$

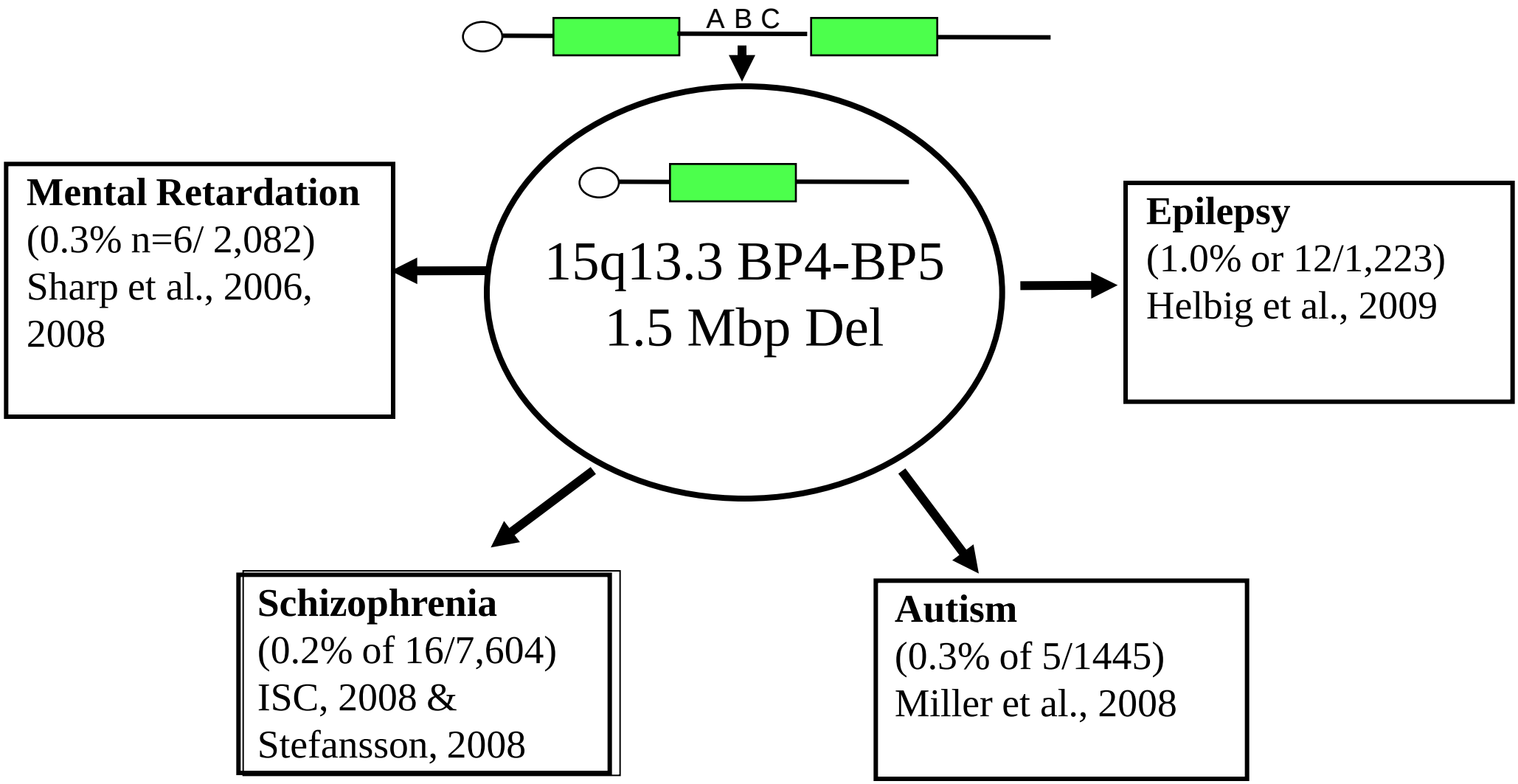
Rare chromosomal deletions and duplications increase risk of schizophrenia

***Nature* 455:232-6 2008**

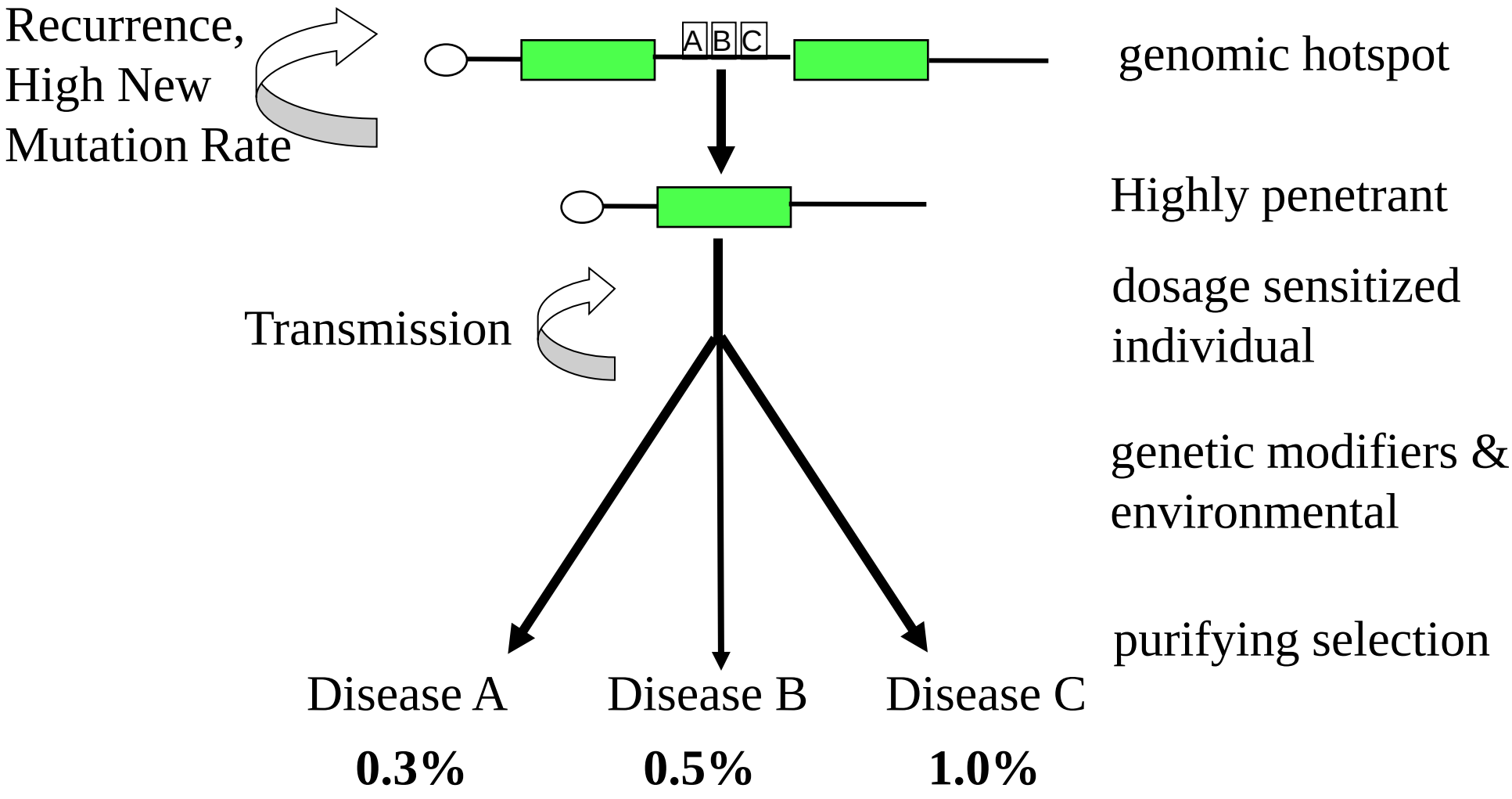
The International Schizophrenia Consortium*

- 9/3,931 cases vs. 0/3,181 controls, OR 17.9, p=0.0029

Disease Heterogeneity



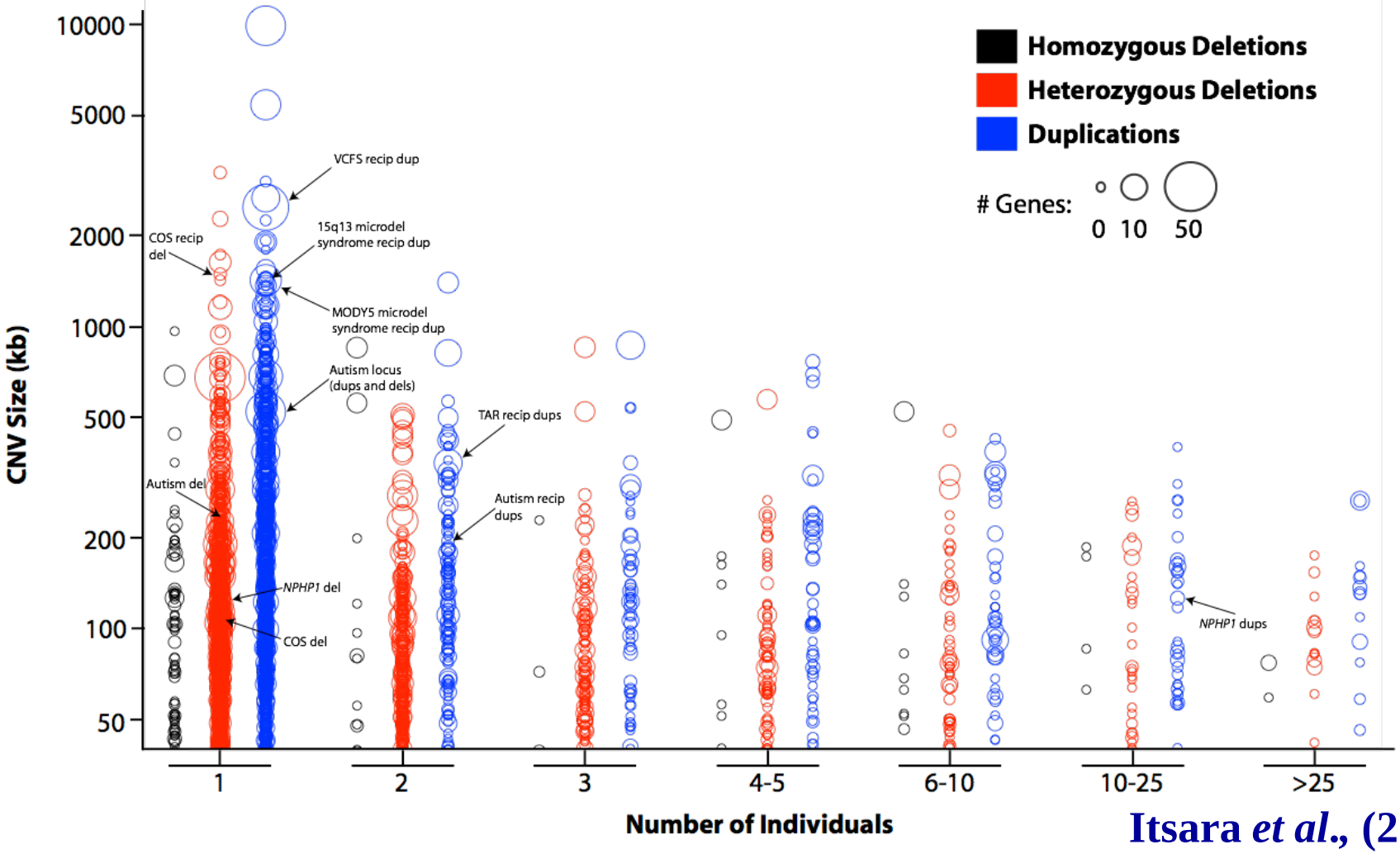
Genomic Hotspot Model of Disease



Mefford and Eichler, *COGD*, 2009

Individually Rare Collectively Common

(N=2,493 normal individuals Illumina SNP Microarrays)

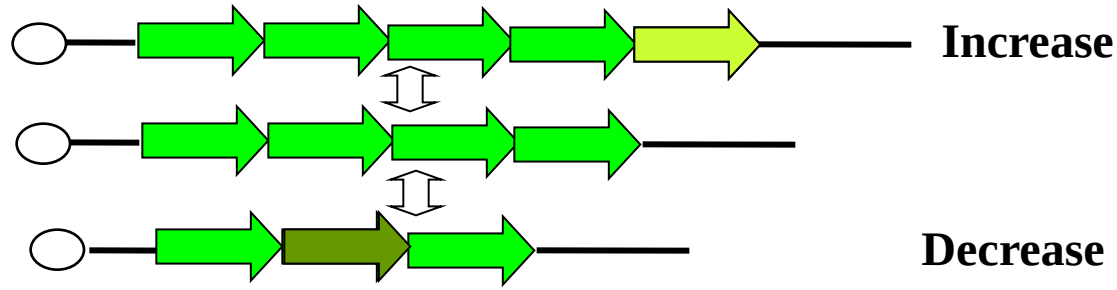


Itsara *et al.*, (2009) *AJHG*

- 8% of normals carry an event > 500kbp
- 90.2% of variant bp are below 1% polymorphism

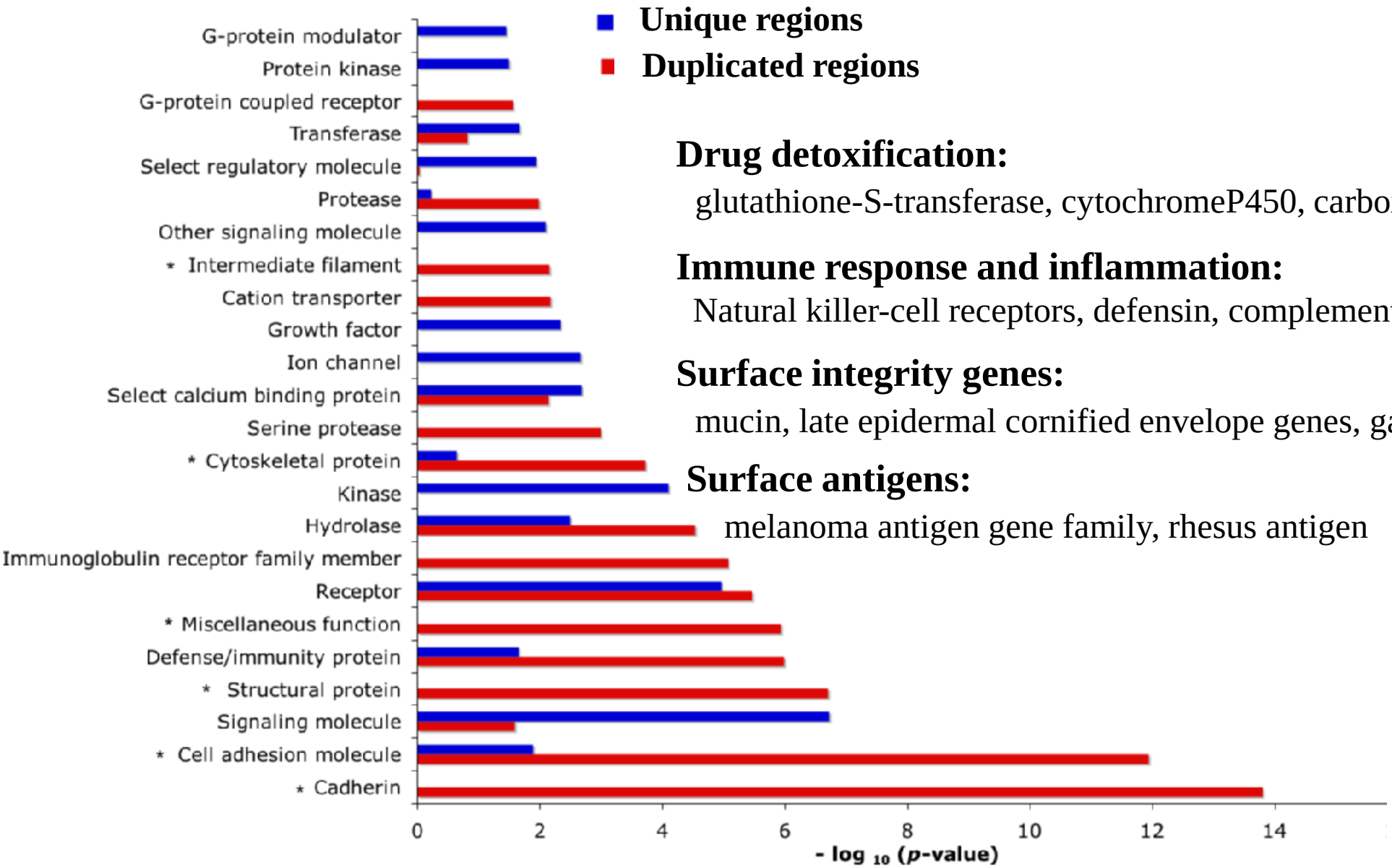
Model #2: Copy Number Polymorphisms and Disease

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



- **Disease CNPs enriched within duplicated sequences.**

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:

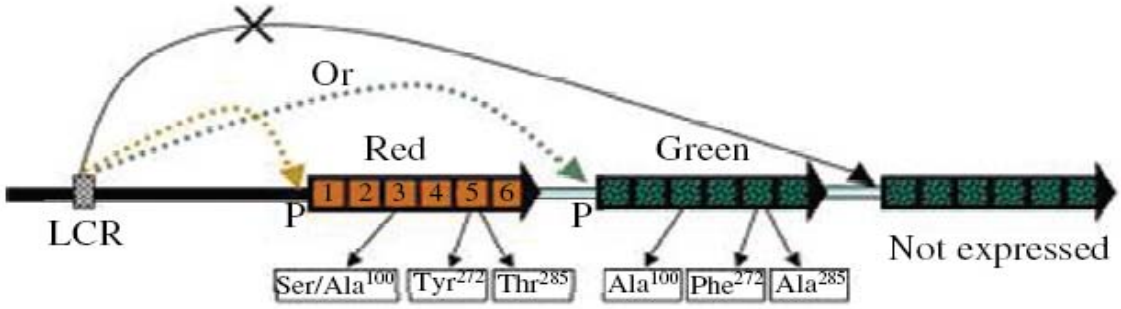
melanoma antigen gene family, rhesus antigen

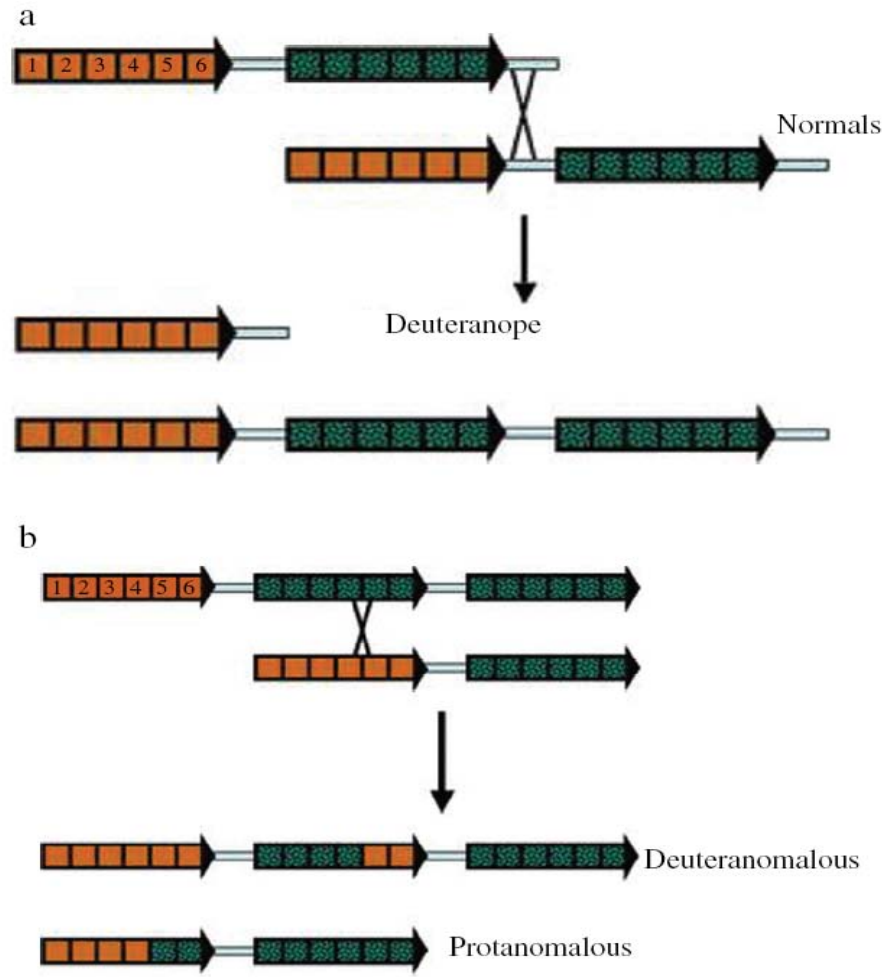
•Environmental interaction and cell-cell signaling molecules enriched

Cooper et al., 2007

Copy-Number Detection is not Sufficient!

Color-Blindness in Humans: The Opsin Loci

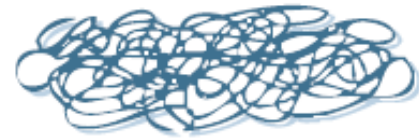




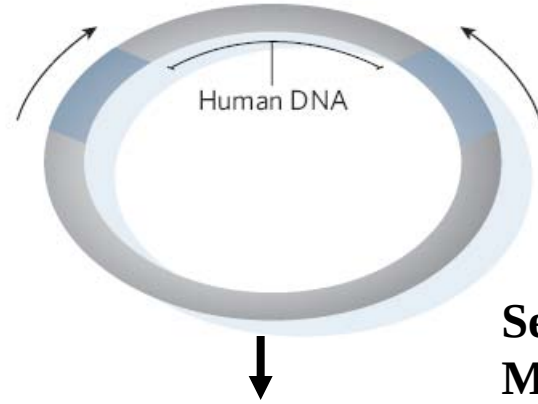
- Normal phenotypic variation
- Red-green color vision defects,X-linked
- 8% of males and 0.5% females. NEur.

Adapted from Deeb, SS, Clin. Genet, 2005

Sequence-Based Resolution of Structural Variation

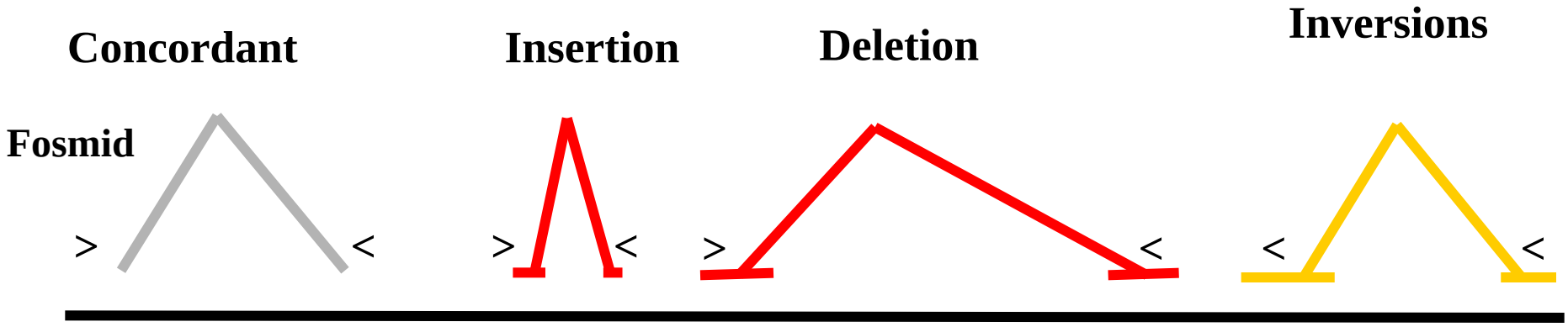


Human Genomic DNA



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)

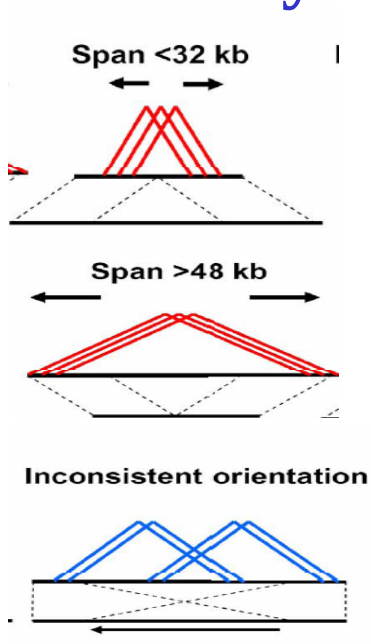
Genome-wide Detection of Structural Variation (>8kb) by End-Sequence Pairs

a)

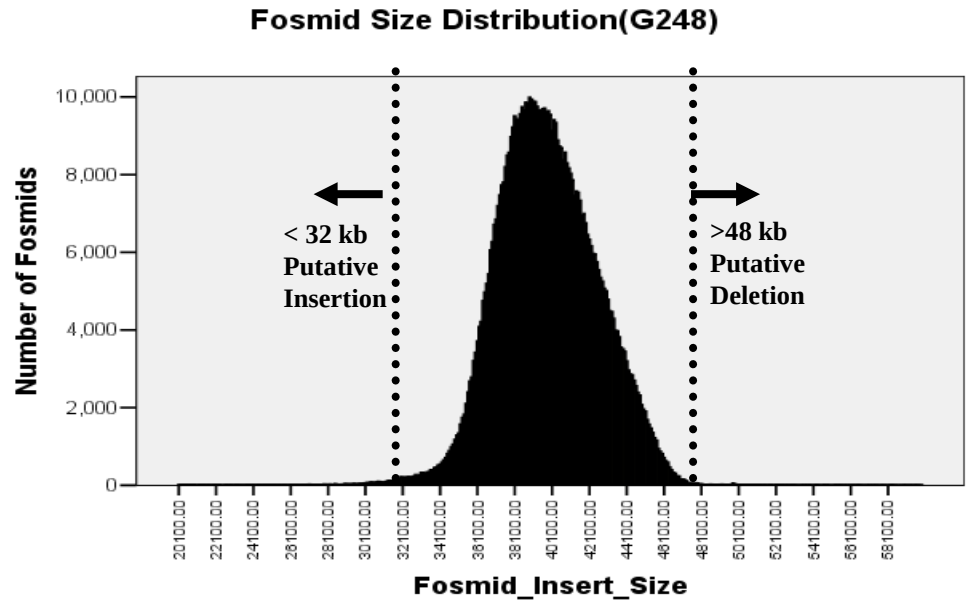
Insertion

Deletion

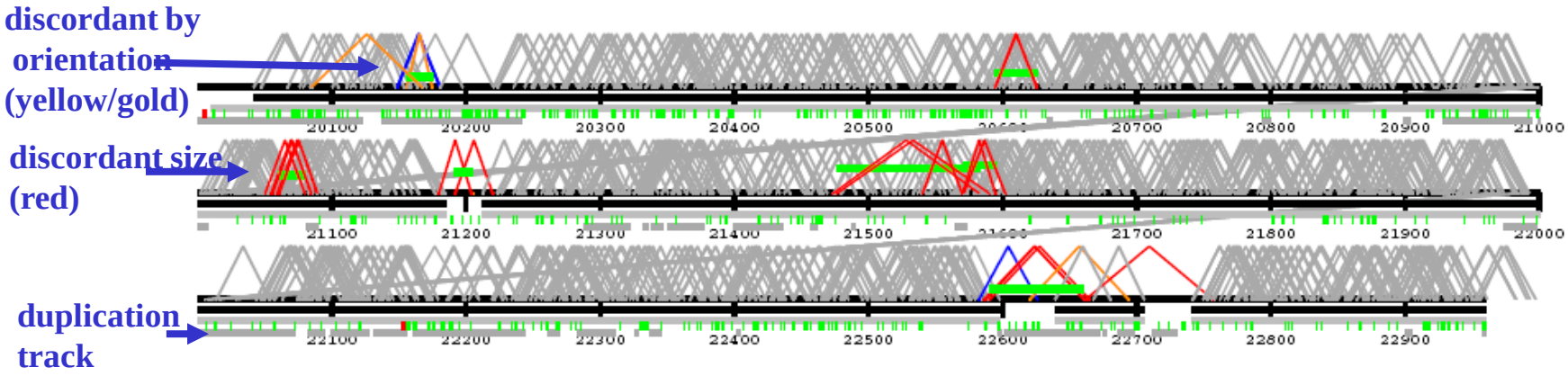
Inversion



b)

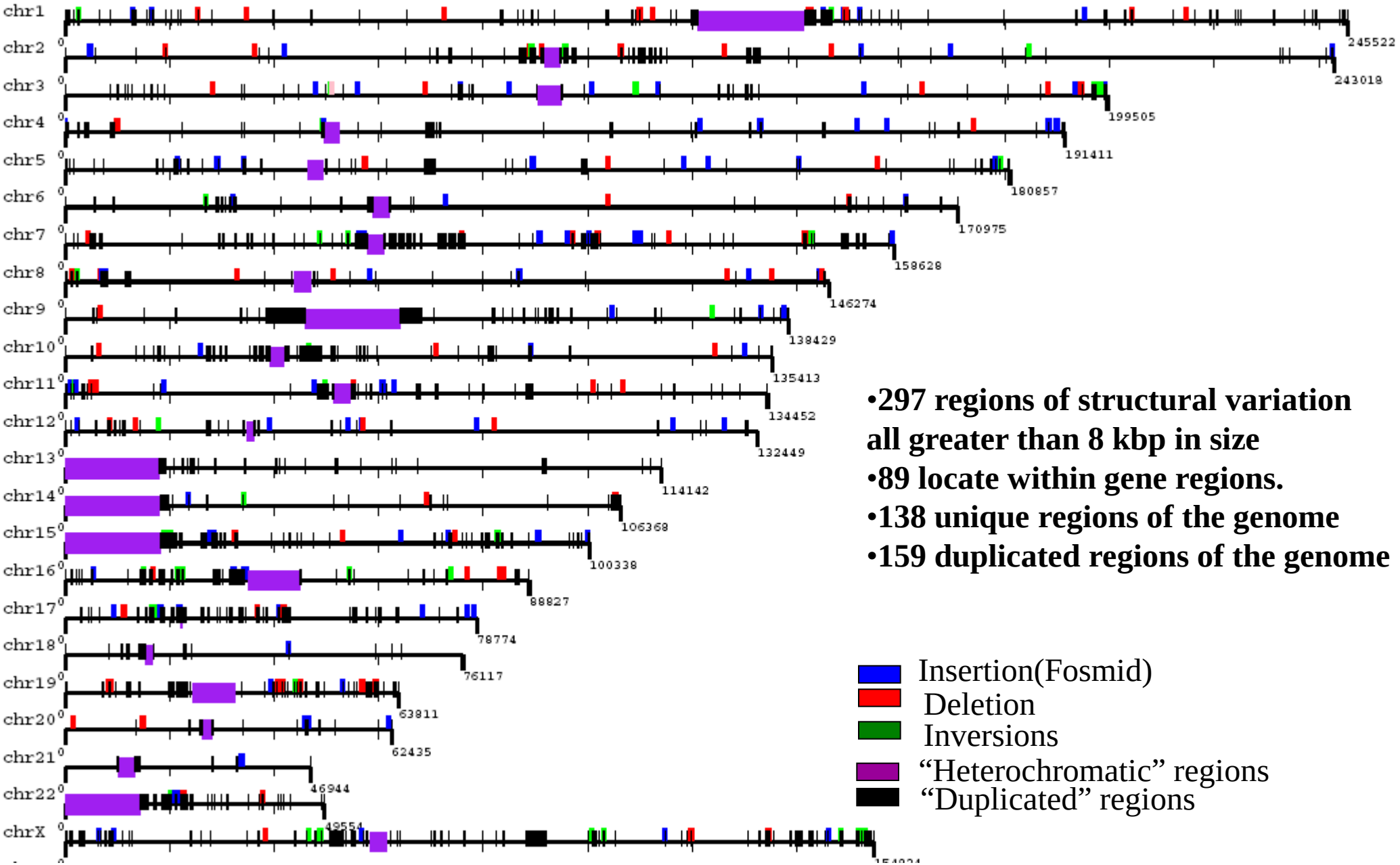


c)



Tuzun et al, *Nat. Genet.*, 2005

Fine-Scale Structural Variation Map: (build35 vs. Fosmids)



- **297 regions of structural variation all greater than 8 kbp in size**
- **89 locate within gene regions.**
- **138 unique regions of the genome**
- **159 duplicated regions of the genome**

Insertion(Fosmid)
Deletion
Inversions
"Heterochromatic" regions
"Duplicated" regions

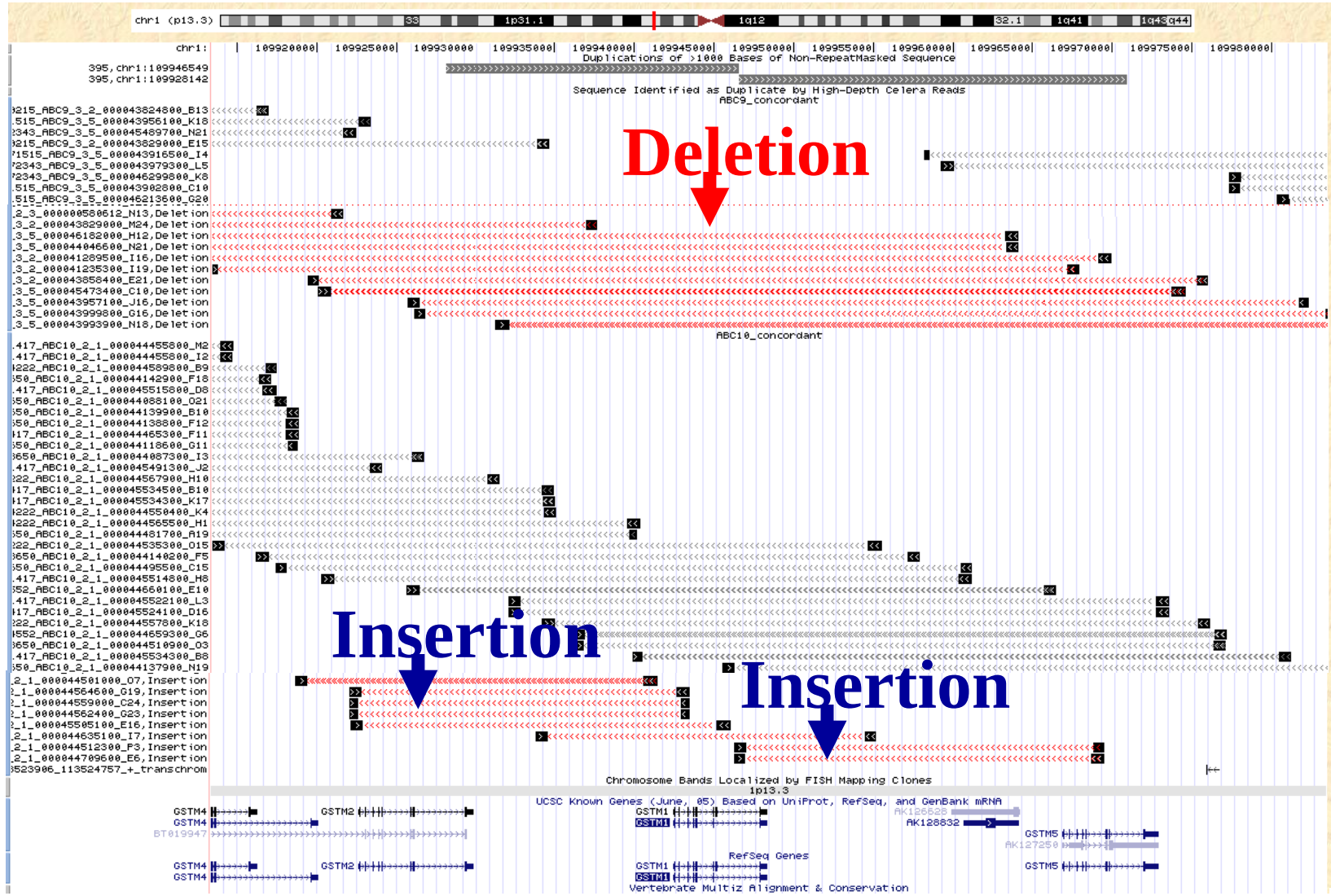
Structural Variation Project Summary

- 17 human genomes sequenced to 0.3 X Sanger sequence (8-10 X physical clone coverage)—combined 135X
- 6 Yoruban, 5 CEPH, 5 Asian (3 Japanese (JPT), 2 Chinese (CHB)
- Identifies ~2051 sites of structural variation and 525 novel insertions

Library	Population	Sample	Number of Reads	Number of Paired Clones	Average Q30 Bases	Total Q30 Bases	Estimated Sequence Coverage	Clones with 'Best' Placement	Mean *in silico* insert size (kb)	Estimated Physical Coverage (Best Clones)	StdDev (kb)	Size Threshold (kb)	Deletions	Insertions	Inversions
G248		NA15510	2,298,774	1,141,942	308	708,952,412	0.25	594,609	39.9	8.21	2.75	8.25	106	112	52
ABC7	Yoruba	NA18517	2,076,237	992,218	405	841,505,234	0.29	616,947	37.6	8.02	3.88	11.64	88	44	64
ABC8	Yoruba	NA18507	3,331,676	1,588,700	465	1,549,030,580	0.54	1,050,579	36.7	13.34	3.85	11.55	164	113	74
ABC9	Japan	NA18956	2,076,828	1,007,581	497	1,032,726,952	0.36	738,786	39.5	10.10	2.26	6.78	174	205	83
ABC10	Yoruba	NA19240	2,118,546	1,032,070	483	1,022,392,331	0.35	741,949	41.0	10.51	1.84	5.52	263	320	95
ABC11	China	NA18555	1,966,644	951,157	478	939,700,332	0.33	724,998	40.0	10.04	1.77	5.31	307	286	72
ABC12	CEPH	NA12878	2,168,656	1,039,478	464	1,005,800,422	0.35	755,087	39.8	10.39	1.4	4.2	290	299	83
ABC13	Yoruba	NA19129	2,053,392	1,009,530	528	1,083,273,138	0.37	757,837	39.3	10.30	1.77	5.31	268	291	91
ABC14	CEPH	NA12156	2,021,844	995,384	537	1,086,415,113	0.38	782,310	39.4	10.68	1.72	5.16	303	297	104
JVI1	China	NA18552	1,992,678	990,194	520	1,036,706,925	0.36	726,247	34.3	8.62	1.82	5.46	201	248	78
ABC16	Japan	NA18947	1,530,585	622,504	370	560,604,744	0.19	500,049	38.3	6.63	1.45	4.35	164	175	38
ABC18	CEPH	NA10847	1,203,756	444,663	369	444,126,739	0.15	361,412	39.2	4.90	1.81	5.43	21	20	6
ABC21	CEPH	NA11993	678,662	239,748	381	258,415,799	0.09	202,343	39.6	2.77	1.65	4.95	48	53	13
ABC22	CEPH	NA11840	777,200	256,967	343	266,283,260	0.09	212,681	38.7	2.85	1.91	5.73	41	24	11
ABC23	Yoruba	NA18523	1,504,416	566,334	379	569,946,866	0.20	467,955	38.7	6.26	1.72	5.16	107	113	32
ABC24	Yoruba	NA18502	1,380,651	544,470	368	507,621,602	0.18	453,587	38.9	6.11	1.52	4.56	167	165	39
ABC27	Japan	NA18942	1,227,159	458,599	350	429,089,775	0.15	374,887	38.9	5.05	1.43	4.29	126	131	25
Total			**30,407,704**	**13,881,539**	**439**	**13,342,592,224**	**4.62**	**10,062,263**		**134.79**			**2,838**	**2,896**	**960**
Total (non-redundant)													**899**	**908**	**244**

Kidd *et al.*, *Nature* 2008, Kidd *et al.*, *Cell, 2010*

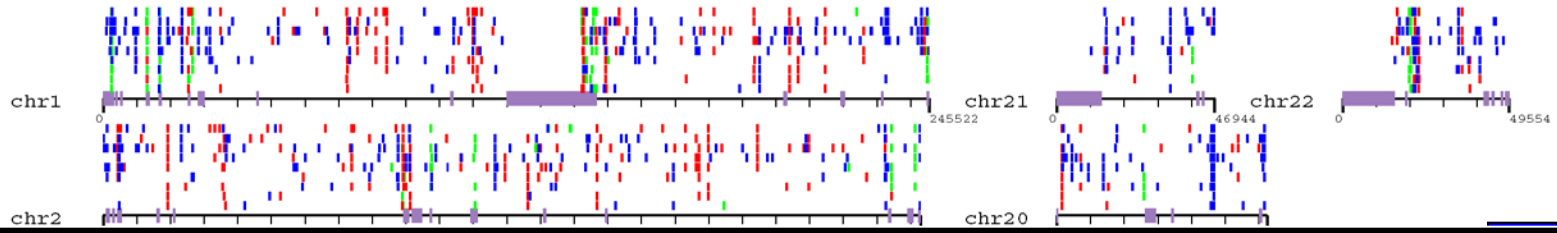
An End-Sequence Pair Clone Resource



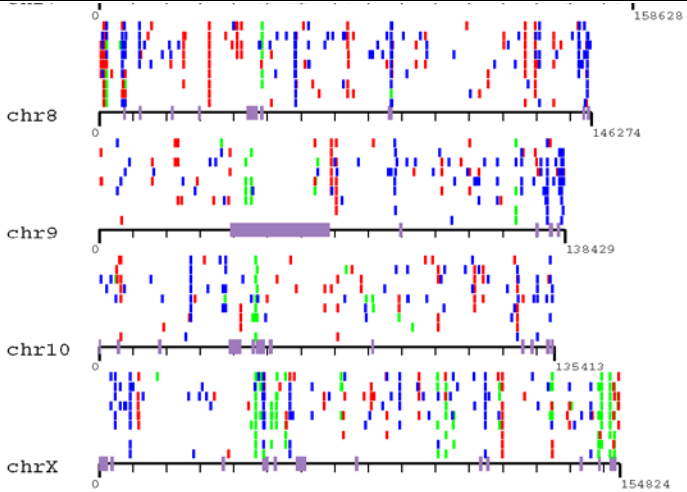
- 13.9 million clones mapped to the reference genome (1 clone every 500 bp)
- Publicly available from UW Genome Sciences and Coriell (last 8 libraries)

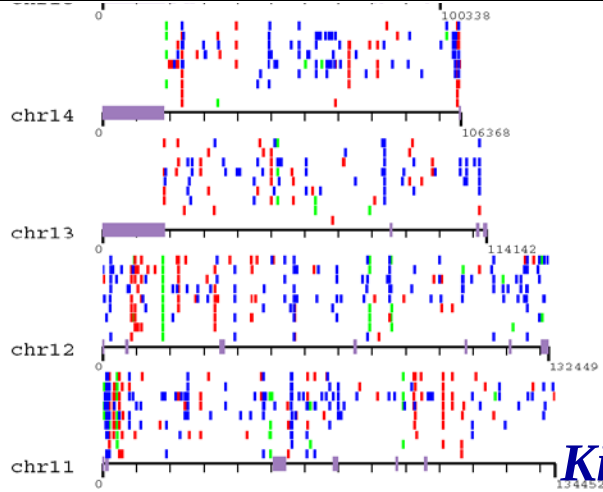
browser view (http://hgsv.washington.edu)

Sequence Map of Structural Variation in Normals (n=8)



27/1720 (1.56%) are known risk factors of human disease including: coronary heart disease, fetal anemia, obesity, psoriasis, Crohn disease, color-blindness, lupus, lupus, halothane/epoxide sensitivity, etc.





ABC10 (Yoruba)

ABC9 (Japan)

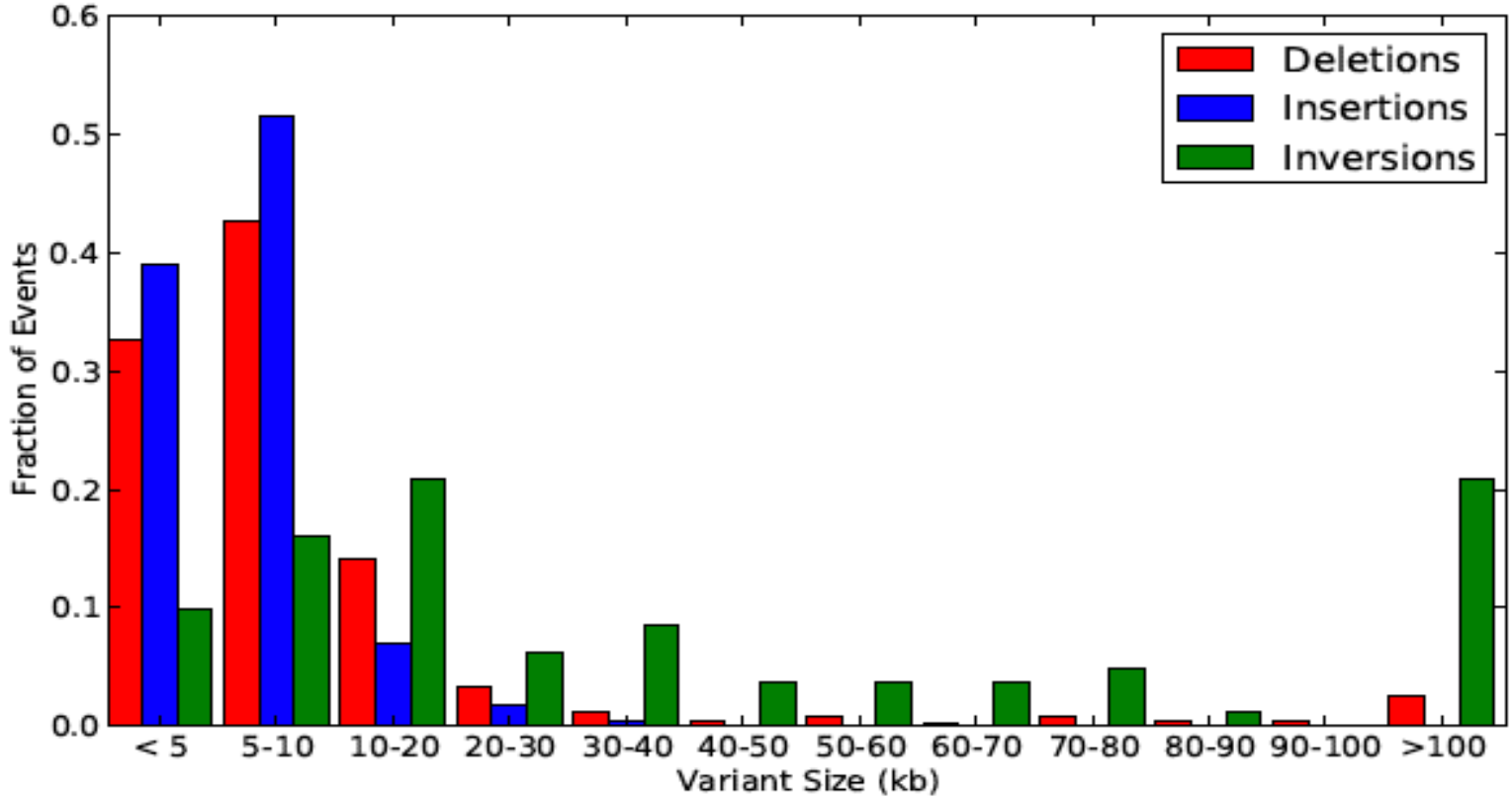
ABC8 (Yoruba)

ABC7 (Yoruba)

G248

Kidd et al., Nature, 2008.

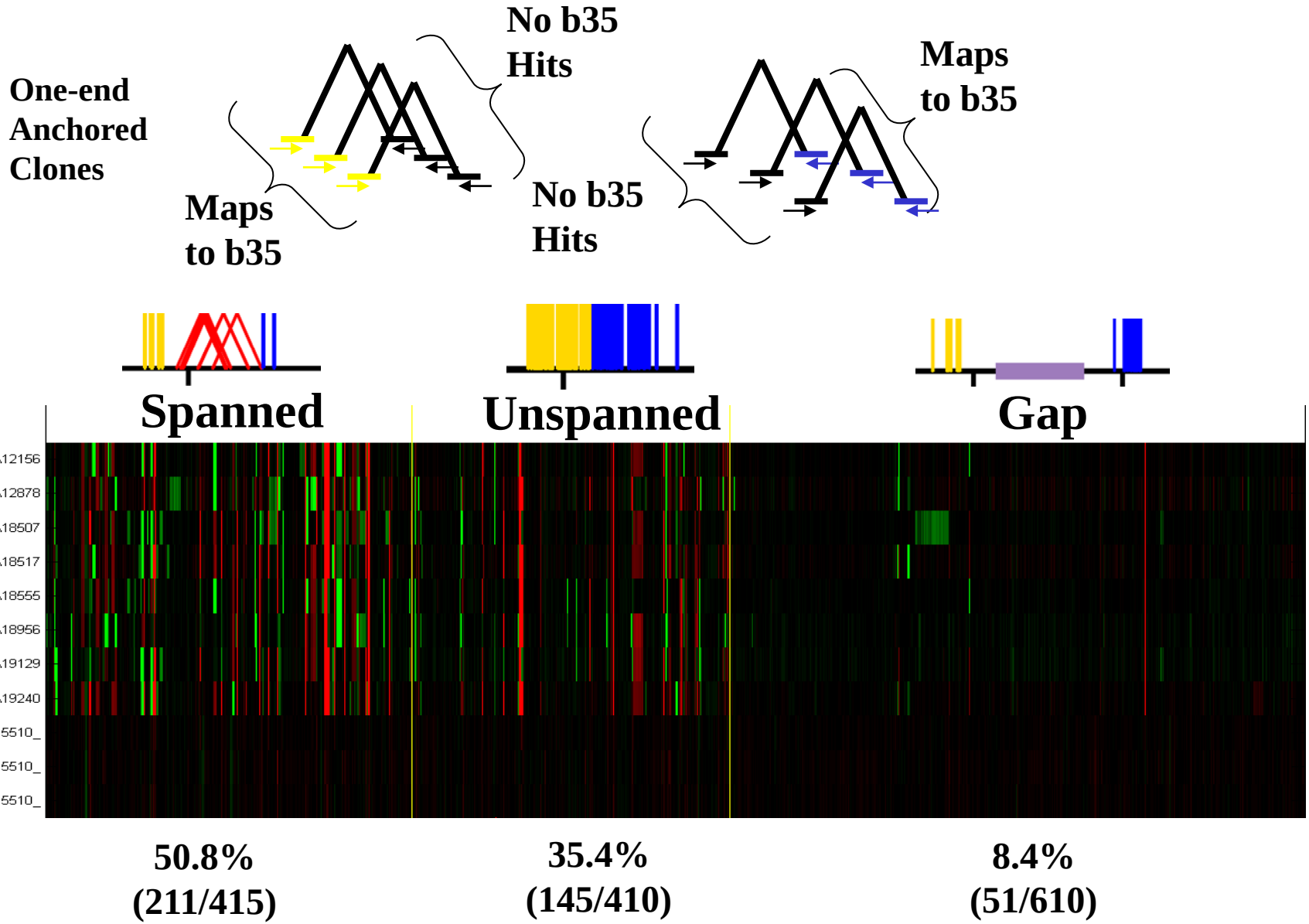
Size Distribution of Sequenced SVs



- **Targeted 1135 loci for fosmid insert capillary sequencing, 1054 resolved**
- **589 deletions,384 insertions and 81 inversions**
- **Mean size 14.1 kbp (del); 6.1 kbp (insertion) and 196 kbp (inversion)**
- **Analyzed 2,081 breakpoints and inferred mechanism(s)**

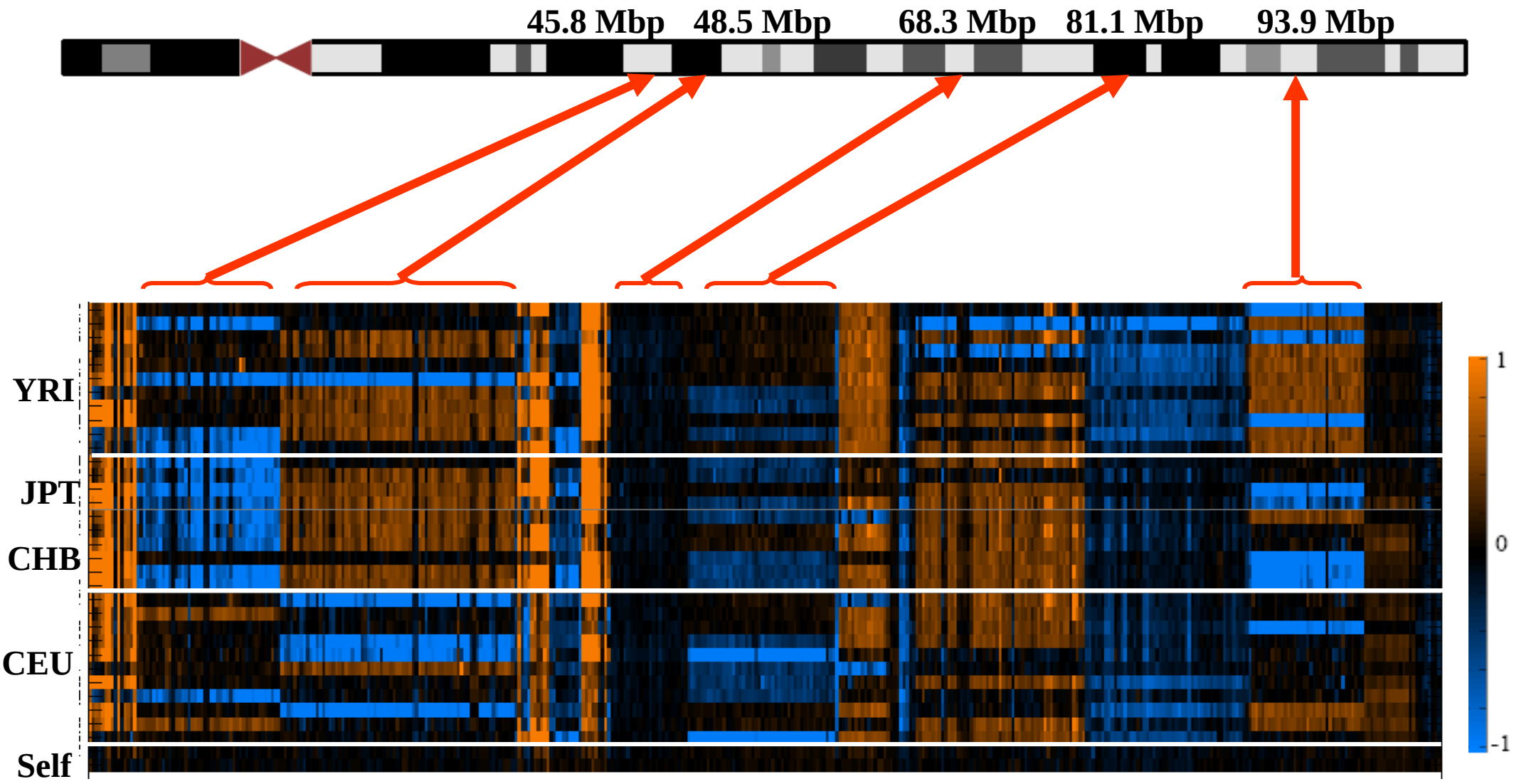
Kidd et al.,, *Cell*, 2010

Discovery of "New" Human Genomic Sequences



•identify 26,556 OEA clones; 4996 cluster into 525 locations

Chr14 Copy-number polymorphic Novel Sequences



- **Identified ~3,013 novel human sequences that are CNP.**

Kidd et al., *Nat. Meth.*, 2010

Summary

- Human genome is enriched for segmental duplications which predisposes to recurrent large CNVs during germ-cell production
- 17% of neurocognitive disease in intellectual disabled children is caused by CNVs—8% of normals carry potentially pathogenic rare event
- Ancestral duplication>>inversion polymorphism>>predisposition to microdeletion
- Read-pair approach allows to discover and completely sequence normal CNVs—estimate 300-500 inversions, deletions and duplications (>5 kbp) per genome.

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

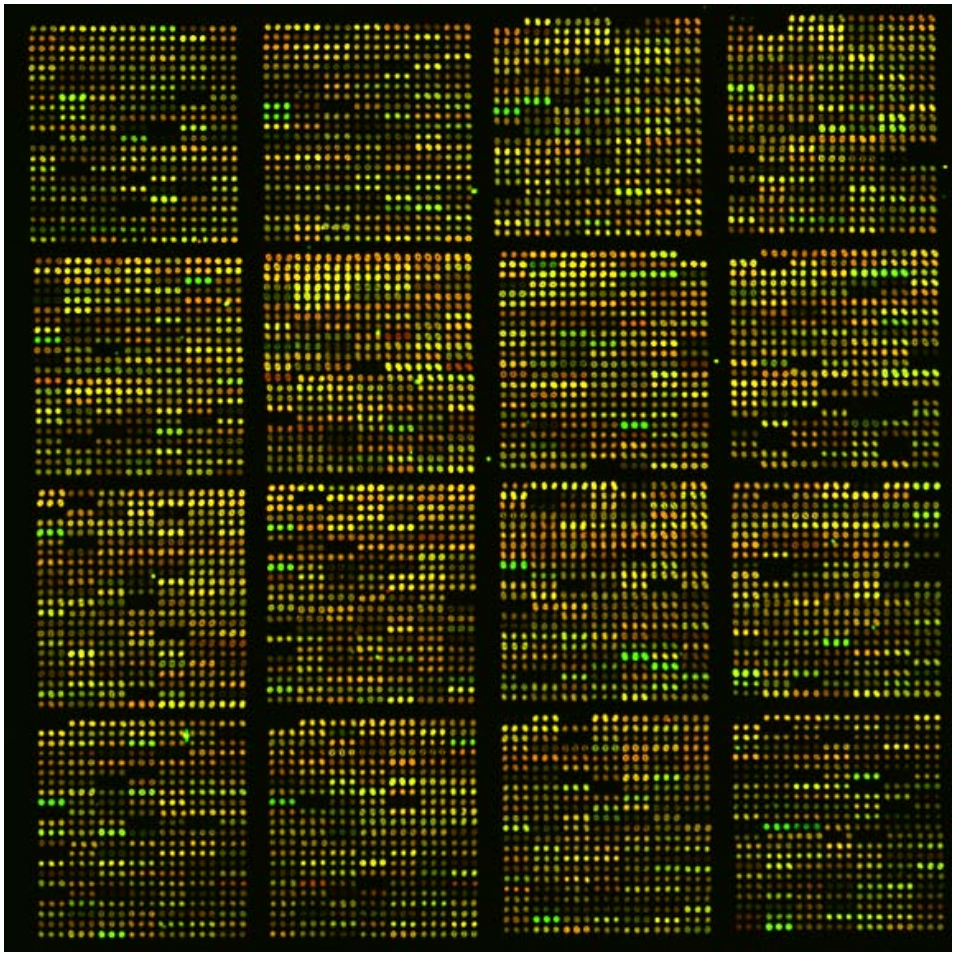
Single molecule analysis

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

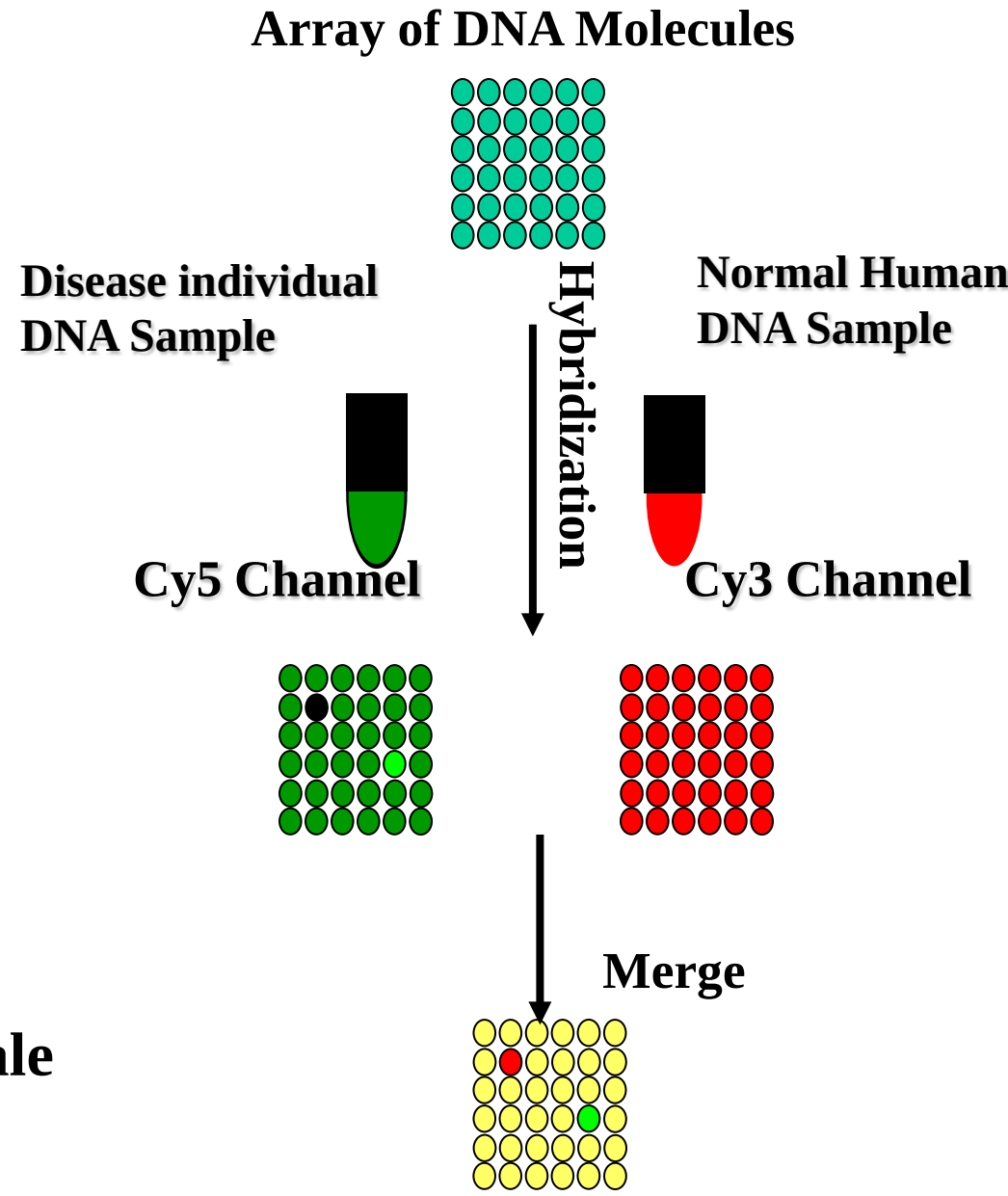
Sequencing-based

- 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,
  - 1000 Genomes Project

Array Comparative Genomic Hybridization

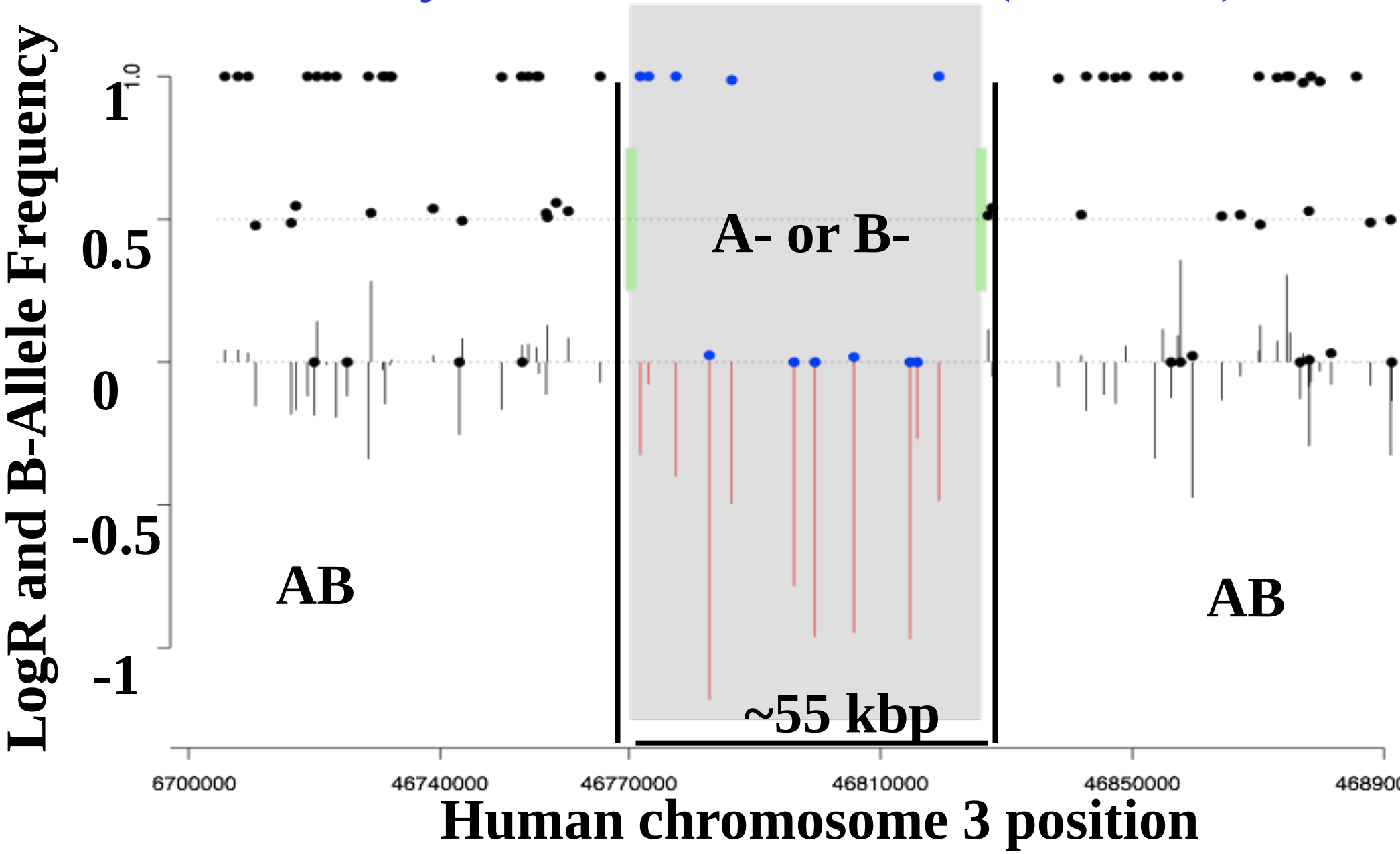


12 mm

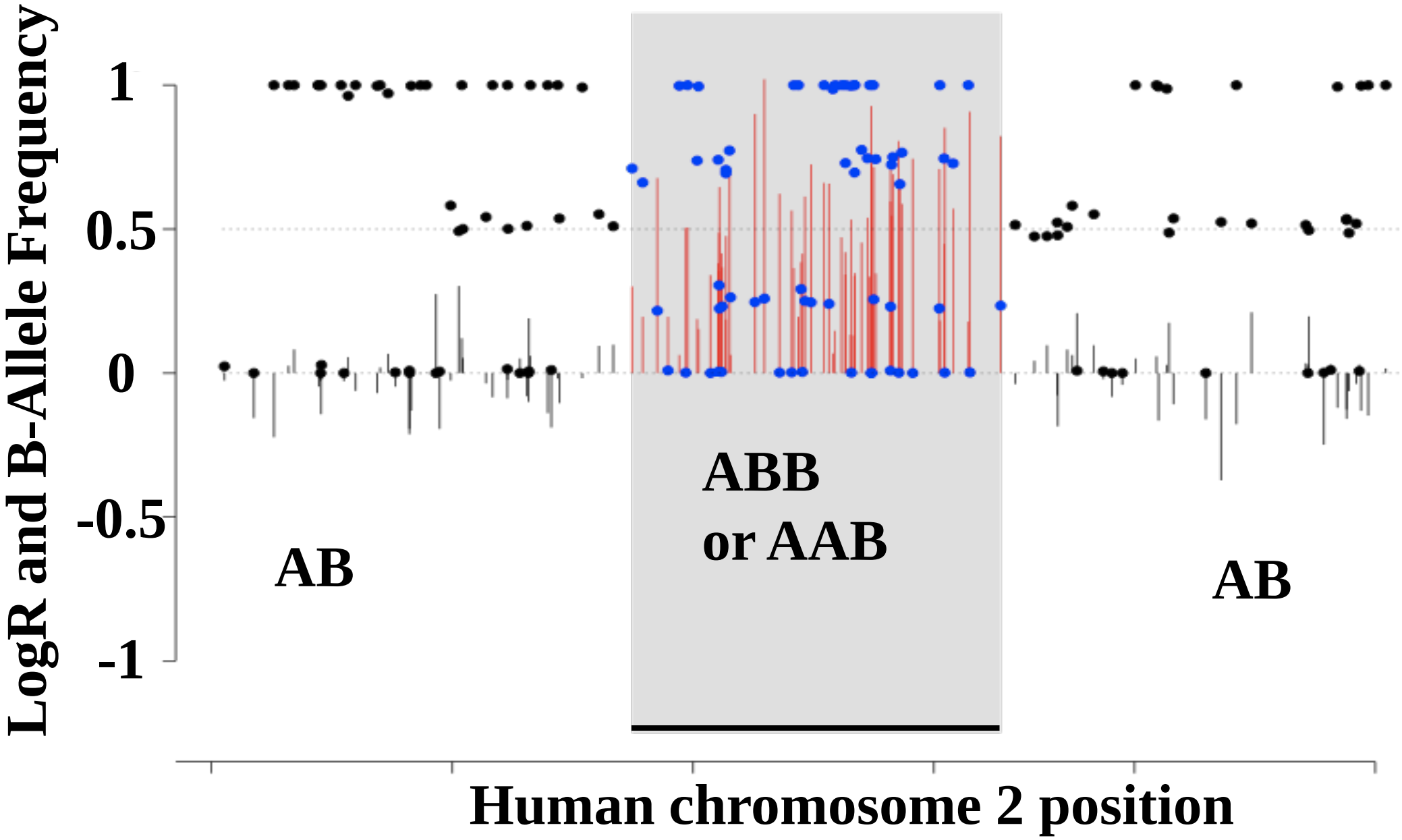


- High-throughput detection of large-scale Deletions and Duplications.

SNP Microarray detection of Deletion (Illumina)



SNP Microarray detection of Duplication (Illumina)

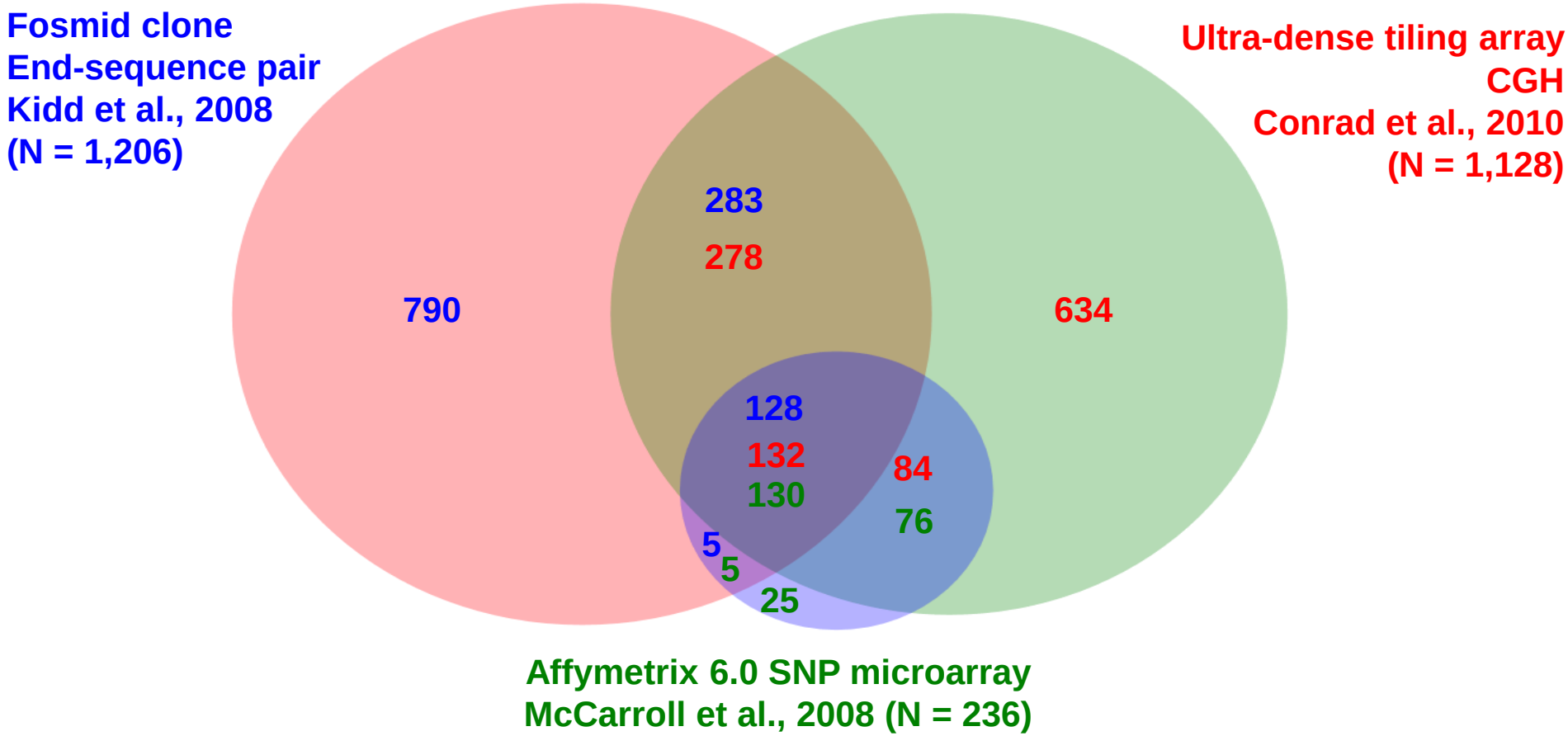


Challenges

- Size spectrum—>10 kbp discovery limit for most commercial platforms.
- Class bias—deletions>>>duplications>>novel insertions>>balanced events (inversions)
- Multiallelic copy number states.
- Repeat/duplication bias
- Throughput
- False negatives.

Method Comparison

(gains and losses > 5 kbp on same 5 individuals)

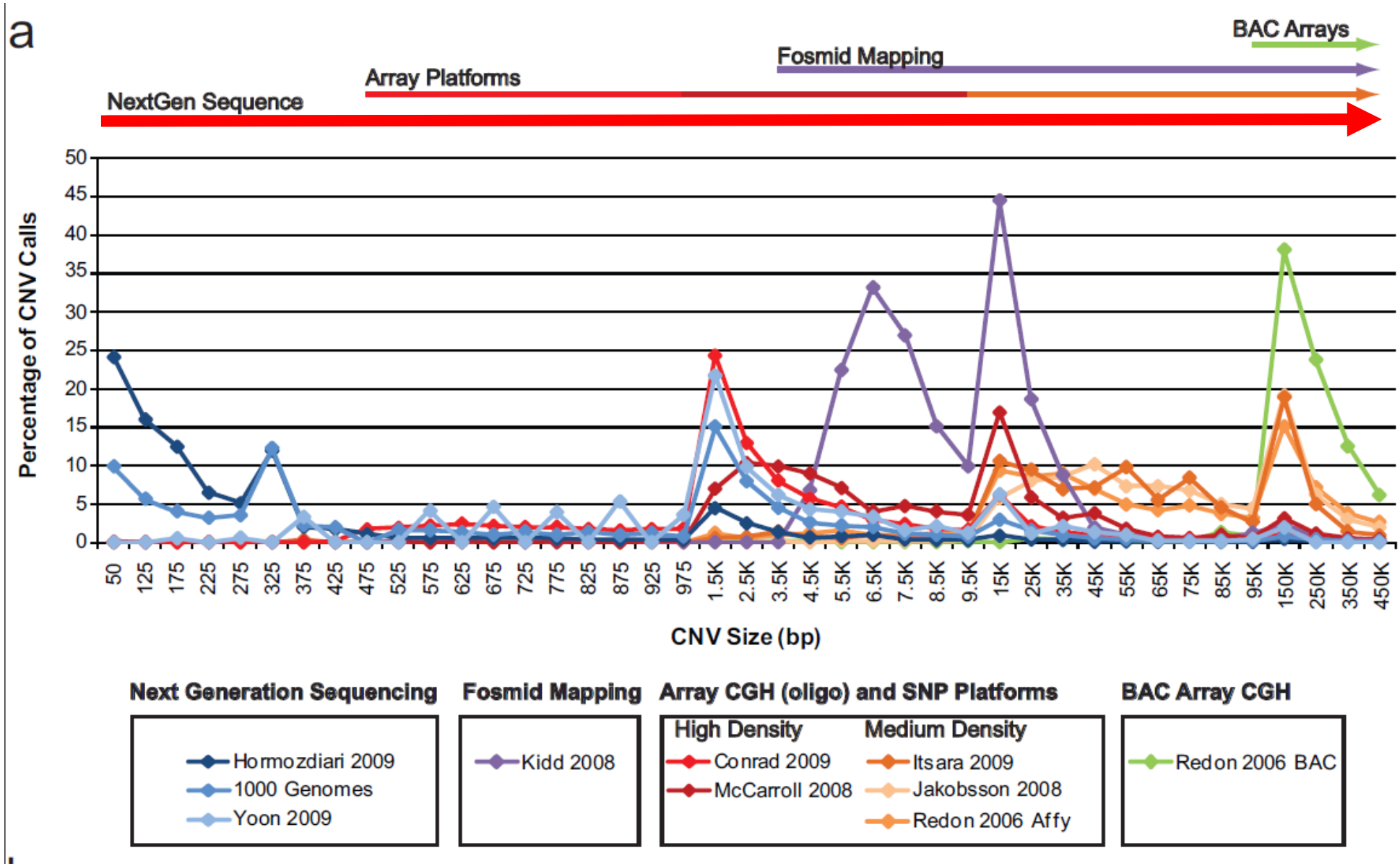


Kidd et al., *Cell 2010*

Next-Generation Sequence Methods

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

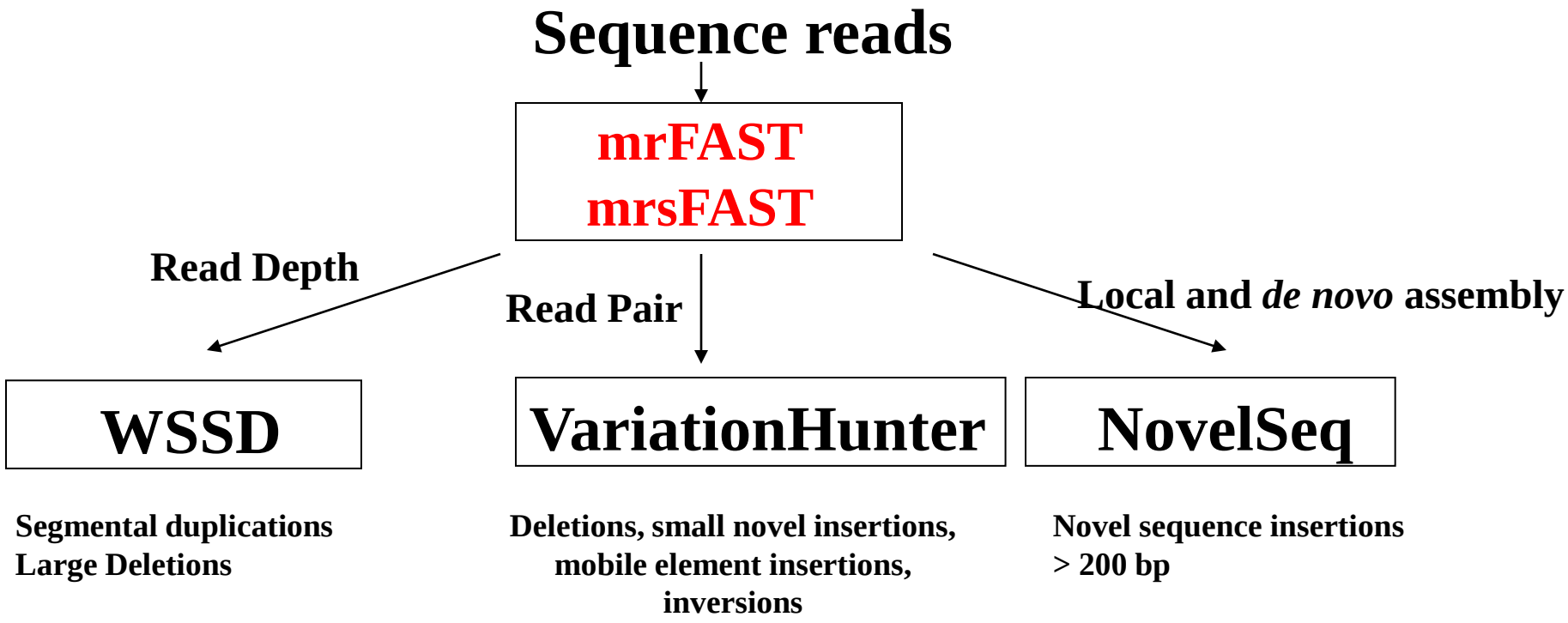
CNV Detection Resolution



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

Challenges and algorithms for NGS

- Short read mapping
  - Discard non-unique: lose sensitivity and detection power.
  - Try to utilize all information: higher sensitivity, less specificity.
    - *micro-read Fast Alignment Search Tools*



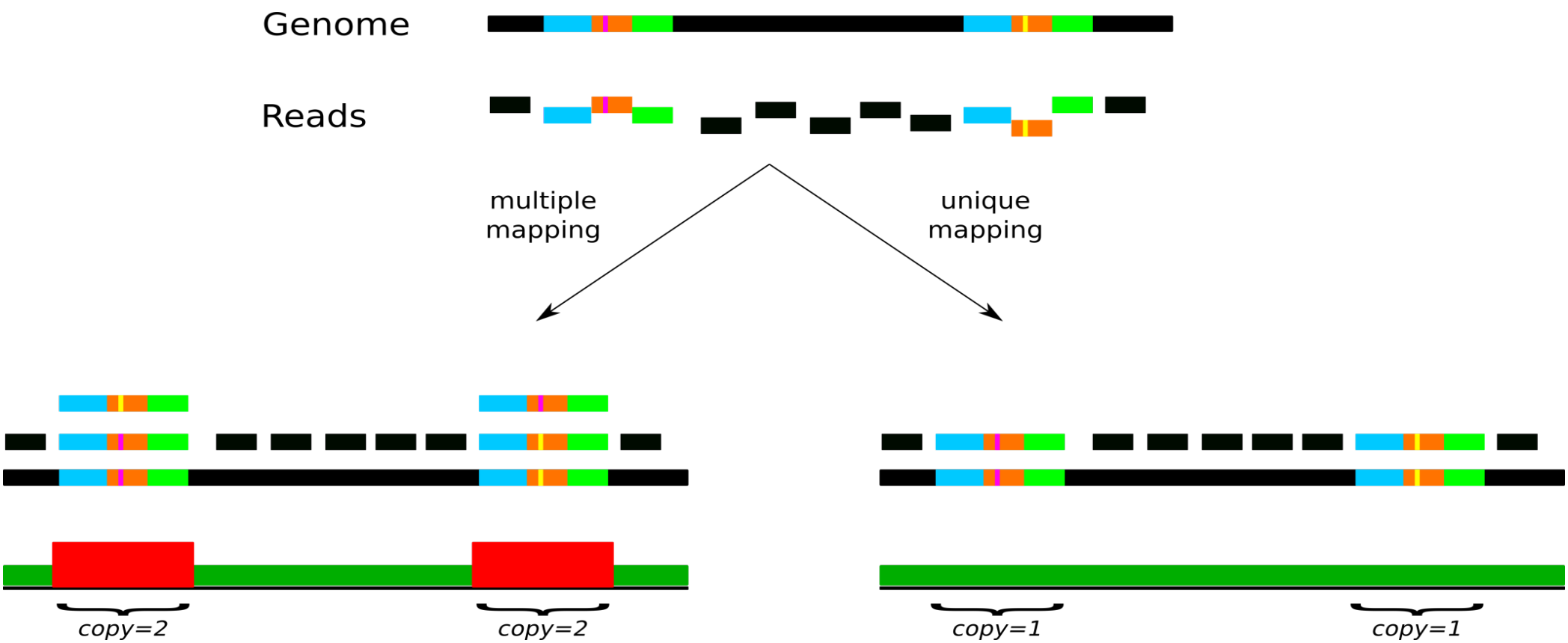
mrFAST algorithm

- Designed for reads generated with the Illumina platform
- Small indels up to 6 bp supported
- Iterative search: ***All*** locations and underlying sequence variation are returned
- Guaranteed sensitivity within user-specified edit distance threshold
- Most other algorithms discard multi-mapping reads:
  - Reduced “mappability” in complex regions of the genome
  - Good performance in unique sequence
- mrFAST-based algorithms:
  - Access to duplications and repeats
  - Higher sensitivity comes with higher FDR.

http://mrfast.sourceforge.net

Alkan et al., Nat. Genet, 2009
Hach et al., Nature Meth.2010

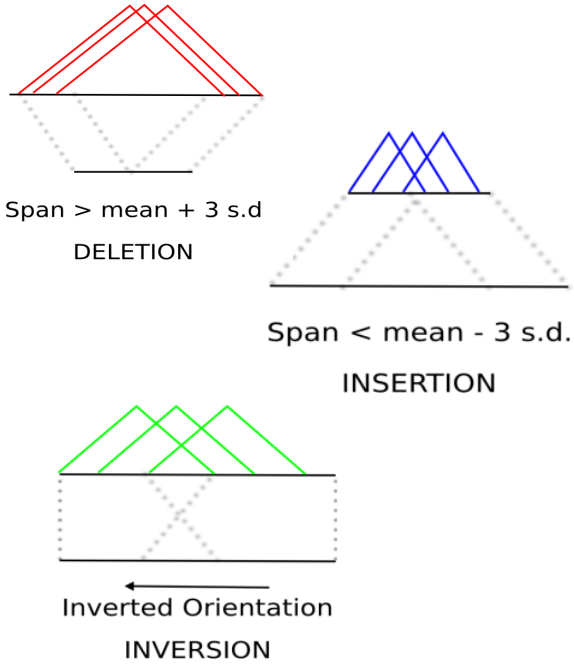
Multiple vs. unique mapping

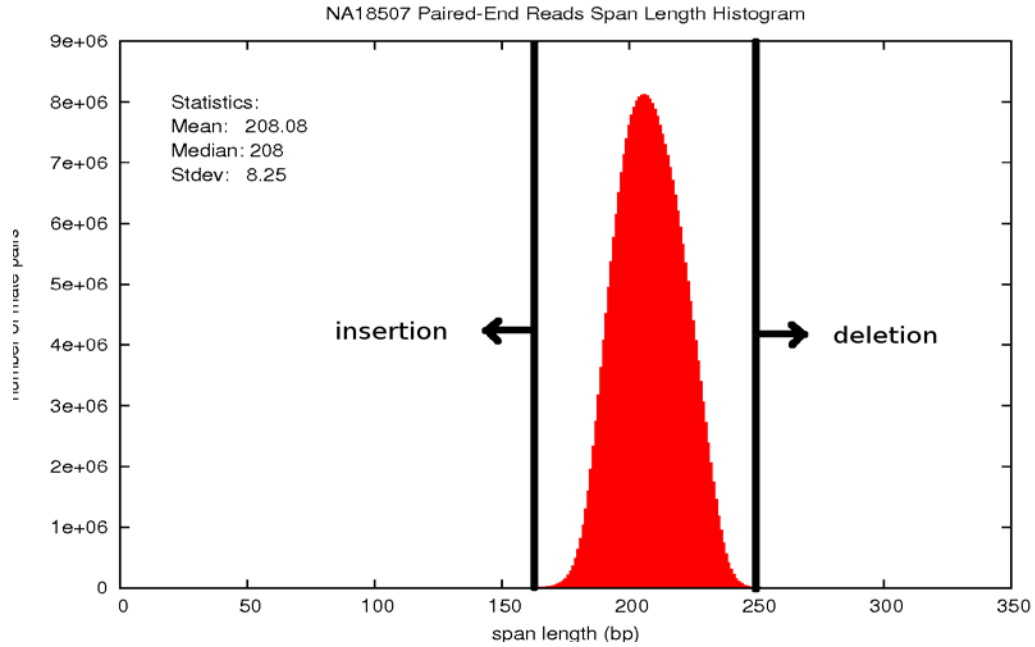


Modified from Chiang & McCarroll, Nat Biotech, 2009

Read-Pair Analysis: Variation Hunter

- VariationHunter-SC: Maximum parsimony approach; using all **discordant** map locations; finds an optimal set of SVs through a combinatorial algorithm based on ***set-cover***
- NA18507: Yoruba male, sequenced with the Illumina Genome Analyzer
  - 42X sequence coverage Read length: 36bp; average insert size: 208bp, standard deviation: 8.25bp

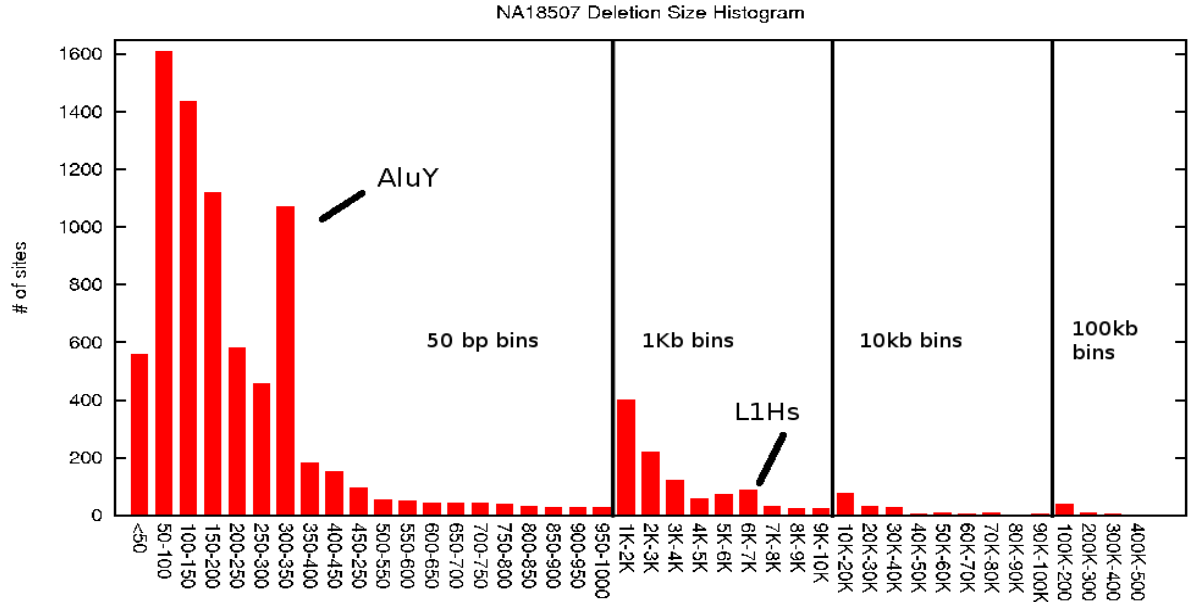


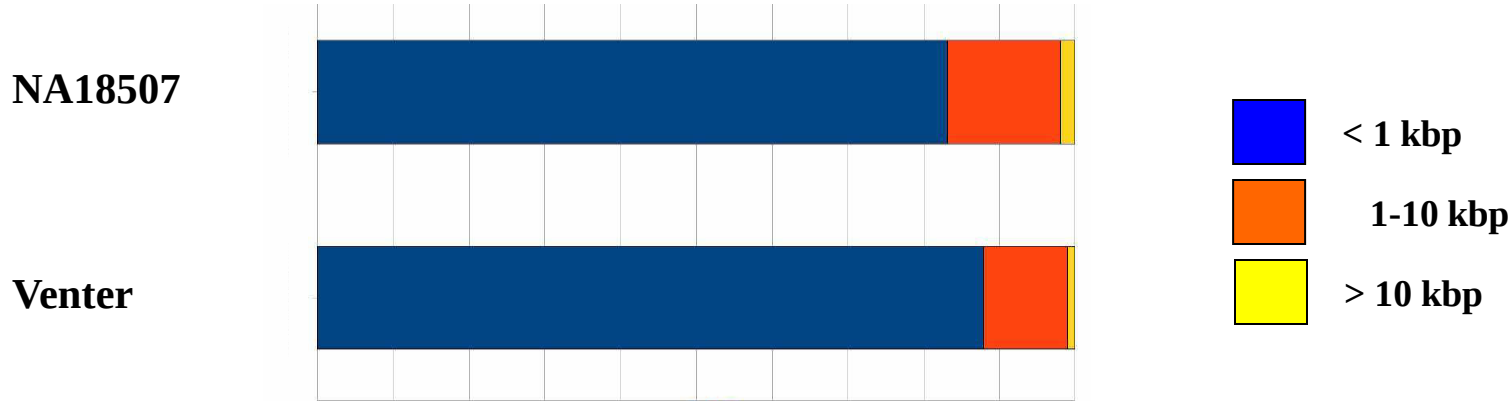


	Validated (Kidd 2008)	VariationHunter-SC		VariationHunter-Pr		Bentley 2008	
		Predicted	Overlap	Predicted	Overlap	Predicted	Overlap
Deletion	143	8,959	85	8,537	85	5,704	67
Inversion	82	504	23	181	11	*NA*	*NA*
Insertion	*NA*	5,575	*NA*	7,142	*NA*	*NA*	*NA*

Hormozdiari, Alkan, et al, 2009

Case Study: NA18507

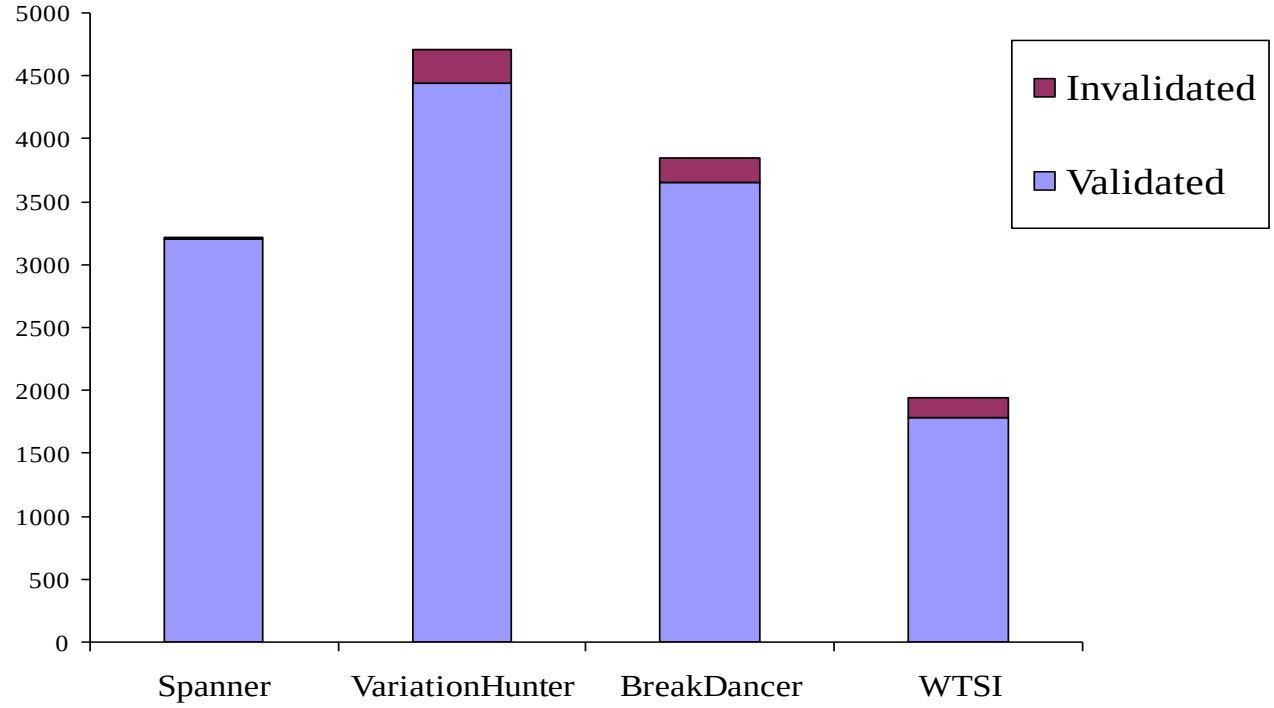




1000 Genomes: Validation

YRI trio: NA19238, NA19239, NA19240
Deletion calls with Illumina read pair analysis
Validation: arrayCGH and PCR

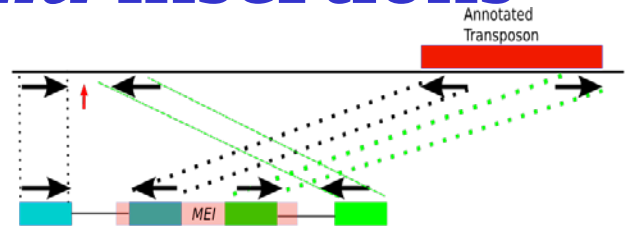
- **Higher sensitivity than unique-location based methods**
- **Higher false discovery rate**
  - **Assumes complete reference genome**
  - **Paralogous variants that do not exist in the reference genome will be invalidated**



1000G Consortium 2010, Mills *et al.*, 2011

Variation Hunter for *Alu* insertions

- 8 high-depth genomes
  - One trio (YRI: NA18506, NA18507, NA18508)

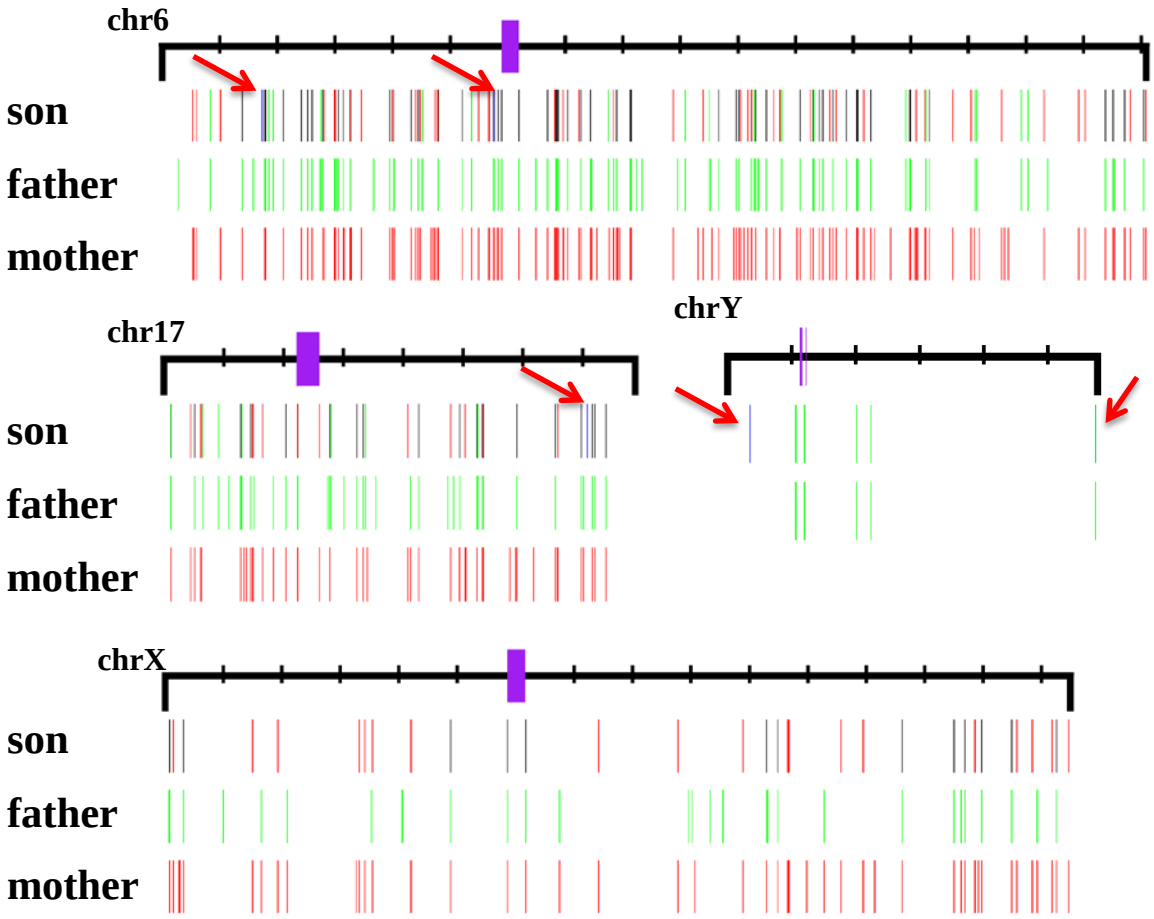


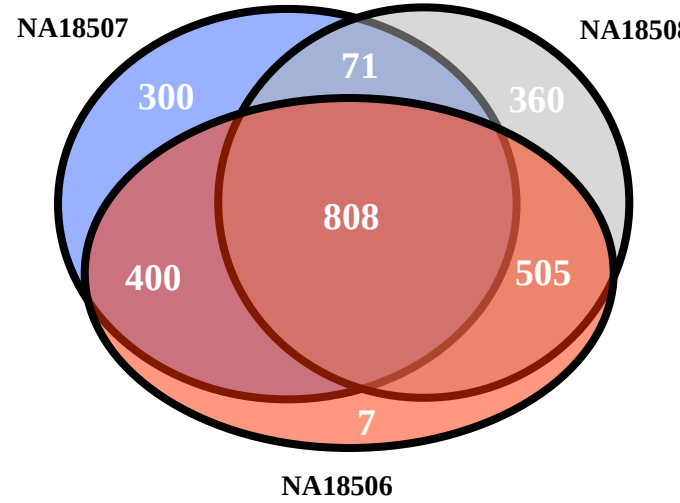
Individual	Population	Seq. Coverage	Phys. Coverage	# *Alu*	Novel
NA18506	YRI	40.1x	255x	1,720	1,280
NA18507	YRI	27.1x	157x	1,579	1,144
NA18508	YRI	37x	214x	1,744	1,293
NA10851	CEU	22x	160x	1,282	781
AK1	Korean	22.5x	49x	909	582
YH	Han Chinese	11.4x	27x	1,160	698
KB1	Khoisan	21x	25x	457	313
HGDP01029	Khoisan	4x	12x	307	214
Non Redundant Total				4,342	3,432

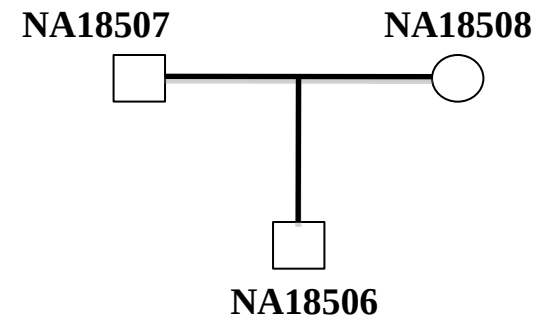
- **63/64 (98%) sites tested with PCR are validated.**
- **7 *de novo* predictions, 0/5 validated as de novo but are inherited**
- **1,437/4,342 (33.1%) map within genes (RefSeq May 2010)**

Hormozdiari *et al.*, *Genome Res.*, 2010

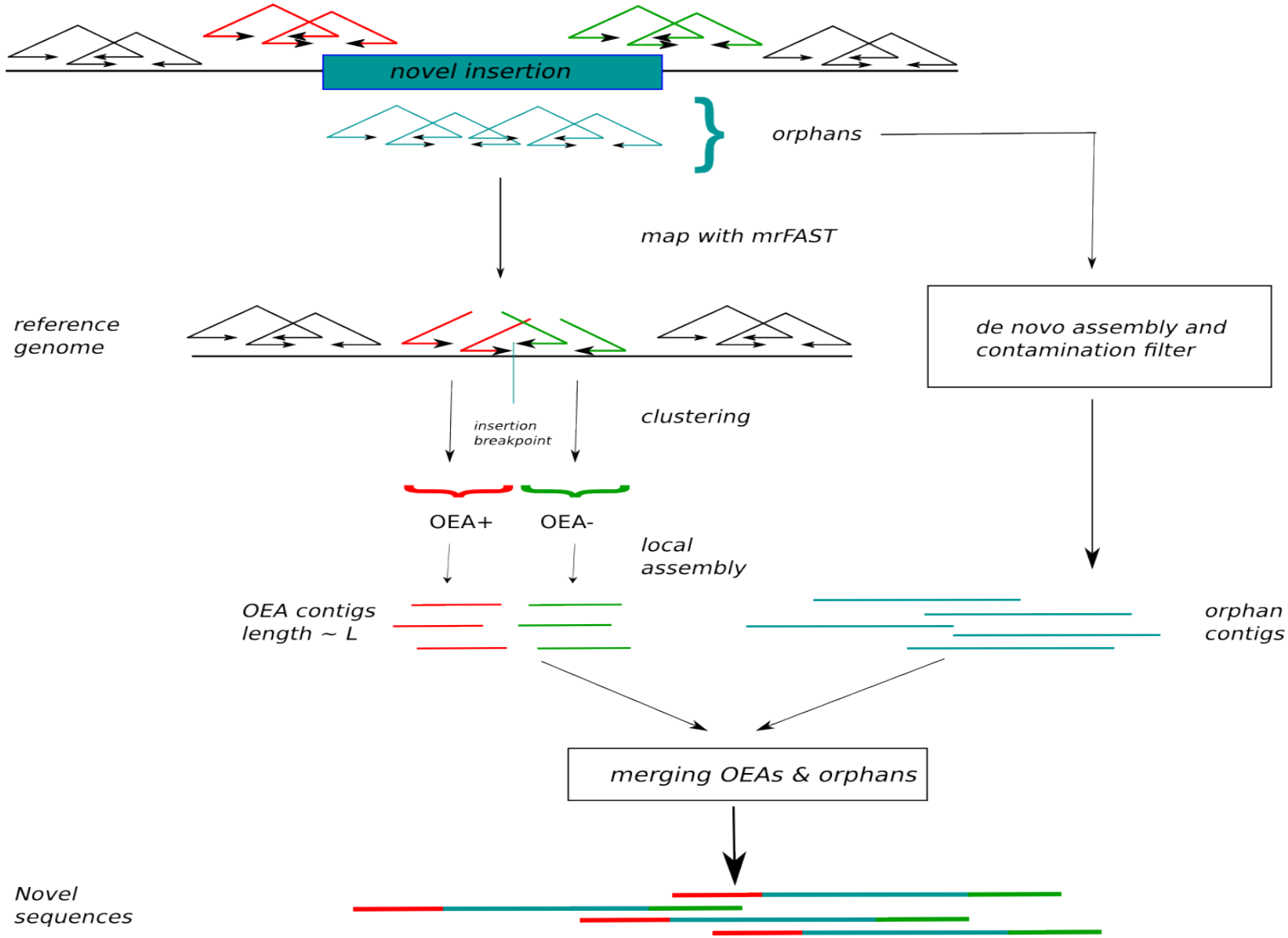
Familial transmission of *Alus*







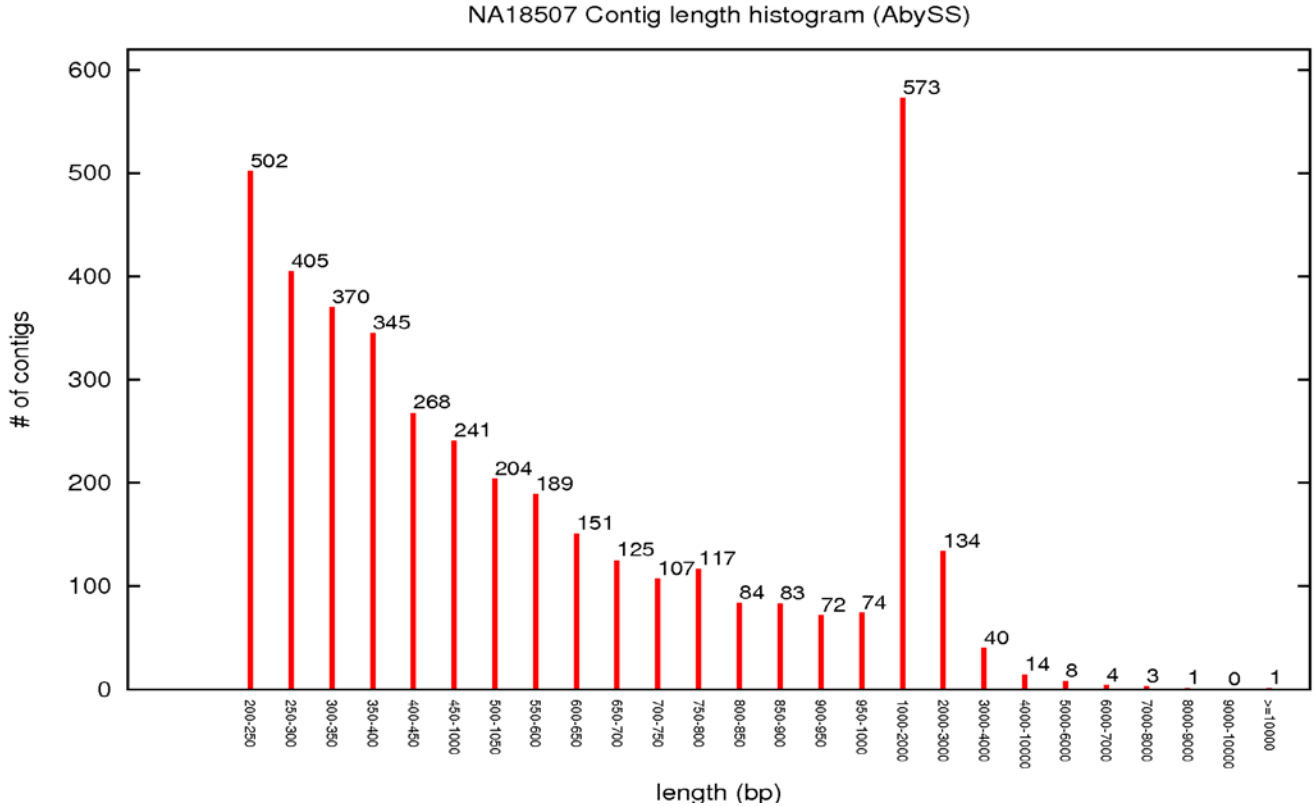
Assembly: NovelSeq

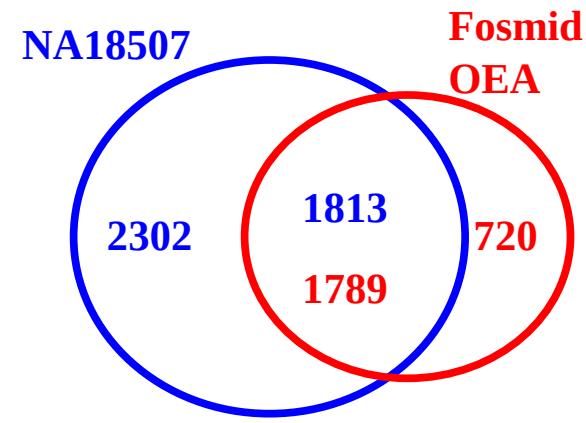


Hajirasouliha et al., Bioinformatics, 2010

NovelSeq: NA18507

Unmapped reads assembled into 4,154 contigs (≥200bp) with AbySS.
Total 2.9 Mb; N50 size: 955 bp

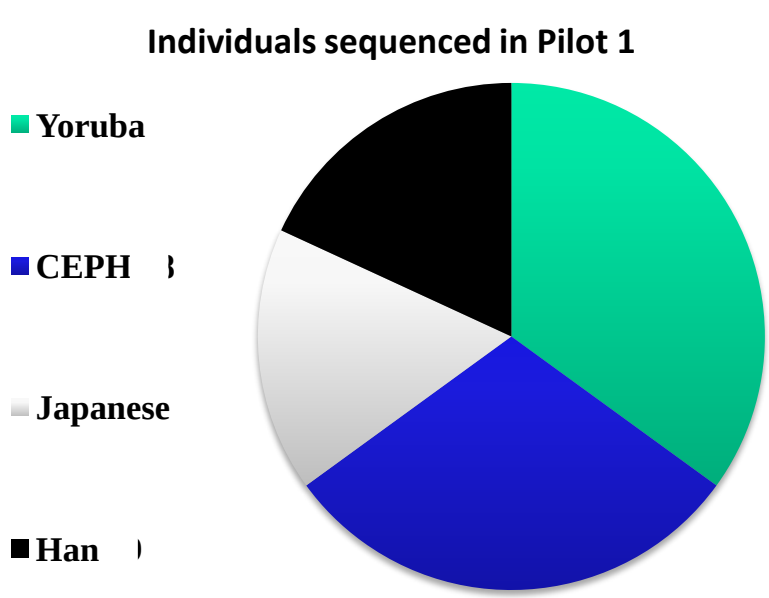


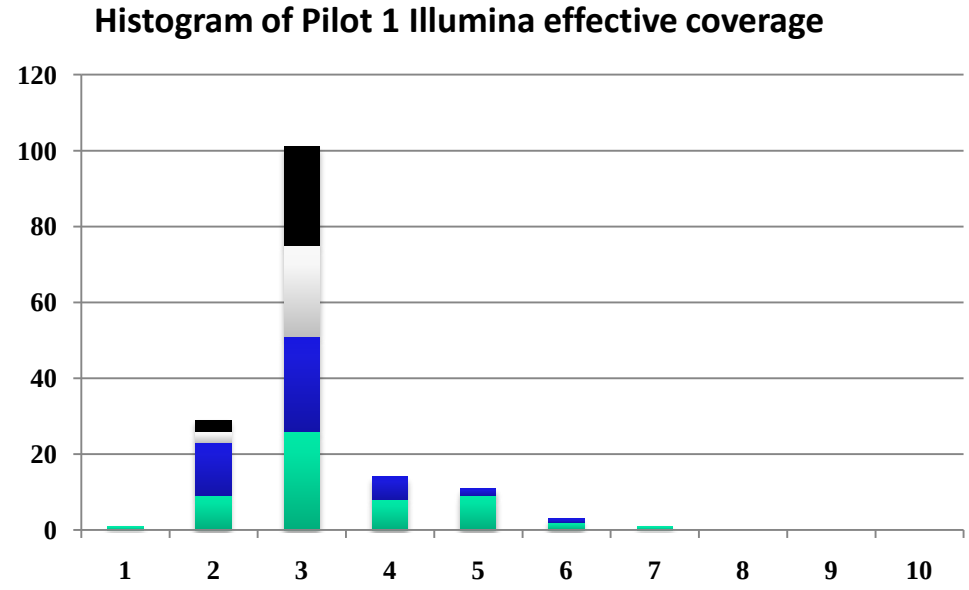


Number of total contigs, and contigs map to various sequence databases with >=99% identity

	Total	Build36	Venter WGS	HuRef	NA18507 Fosmid WGS
>=1 kb	779	13	744	43	337
>=200bp (<1Kb)	3336	110	2957	69	654

Scaling up: 1000 Genomes Project





- SV Subgroup applies 19 distinct algorithms
- 160 genomes 2-40X

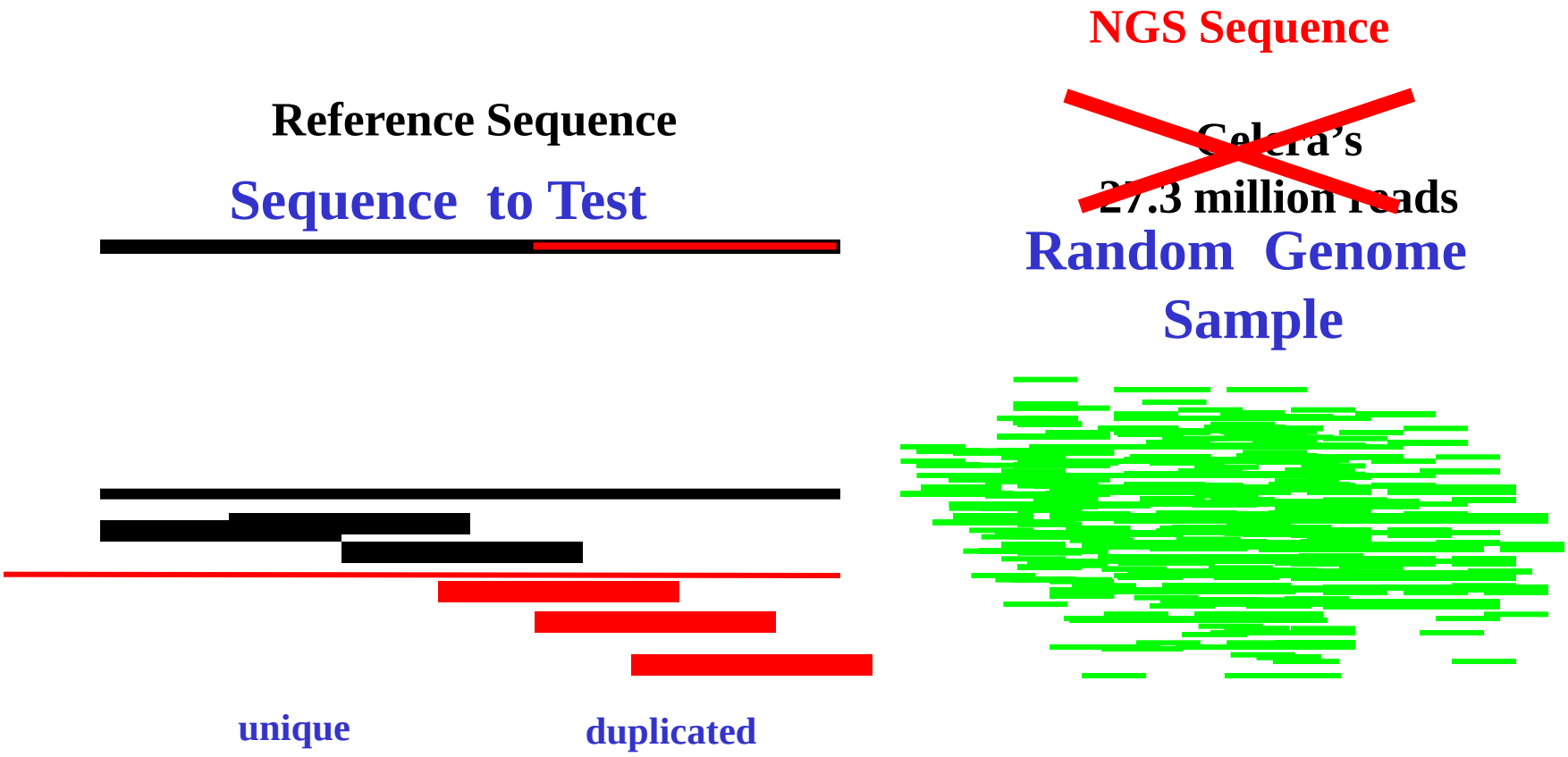
SV Summary

	Validated	predicted
Deletions	10,810	22025
Tandem Dups	-	501
Mobile Elements	-	5371
Novel Ins	128	128
	Total	28025

Mills et al., Nature, 2011

Using Sequence Read Depth

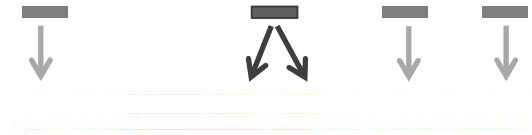
- Map whole genome sequence to reference genome
  - Variation in copy number correlates linearly with read-depth
- Caveat: need to develop algorithms that can map reads to all possible locations given a preset divergence (eg. mrFAST, mrsFAST)

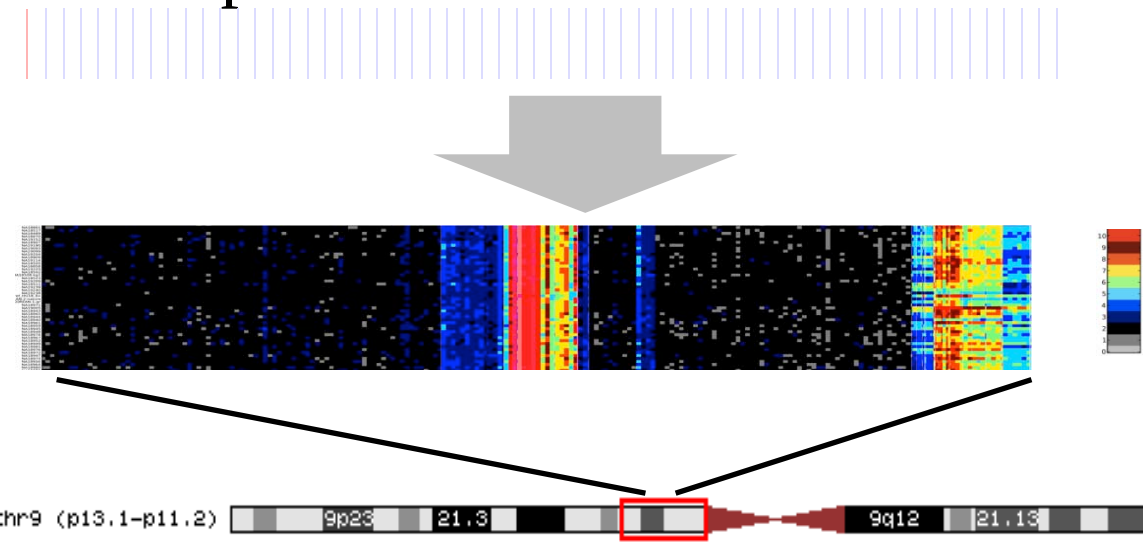


Bailey et al., Science, 2002

Copy number from short read depth

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





Personalized Duplication or Copy-Number Variation Maps

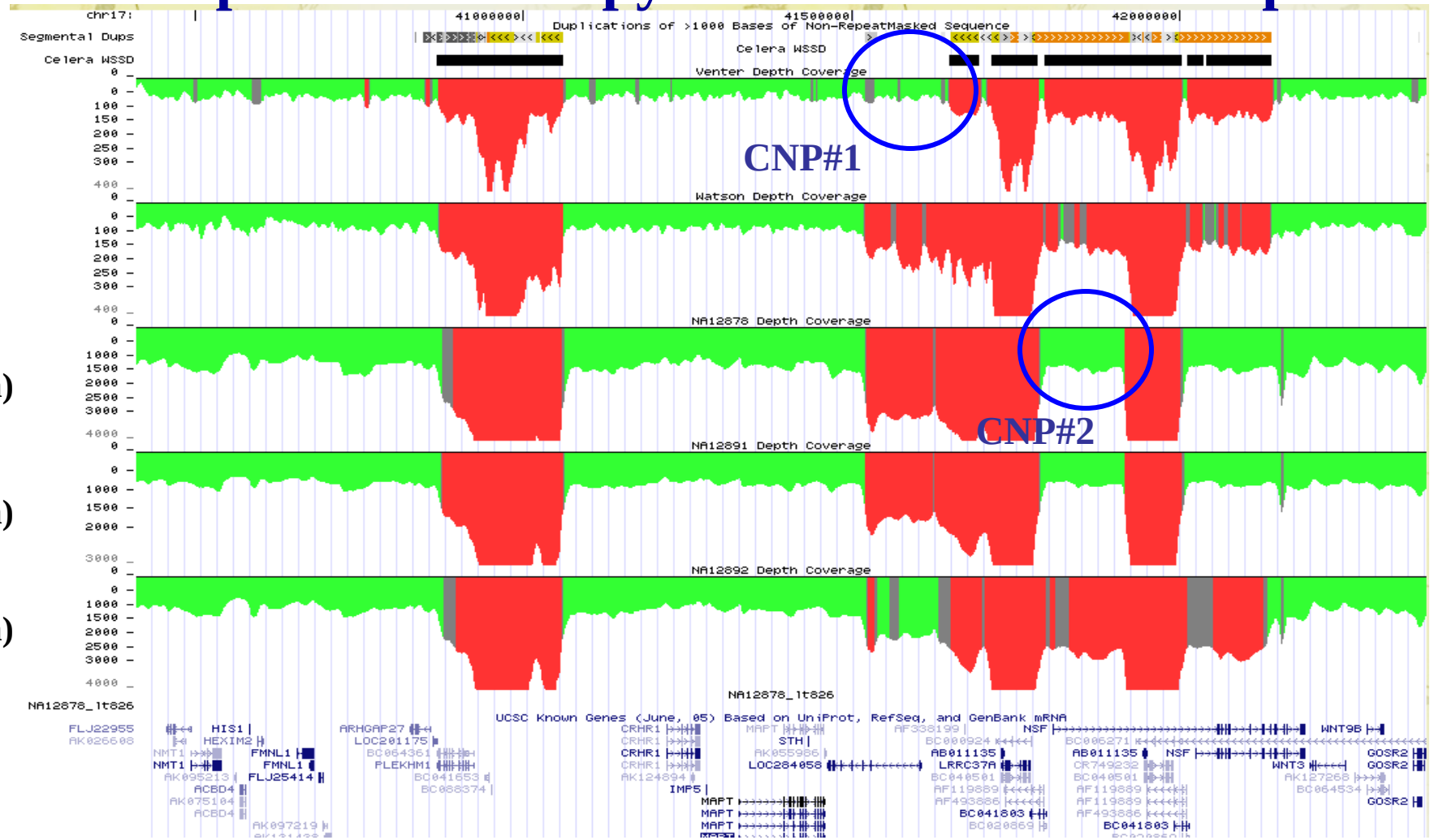
Venter (Sanger)

Watson (454)

NA12878 (Solexa)

NA12891 (Solexa)

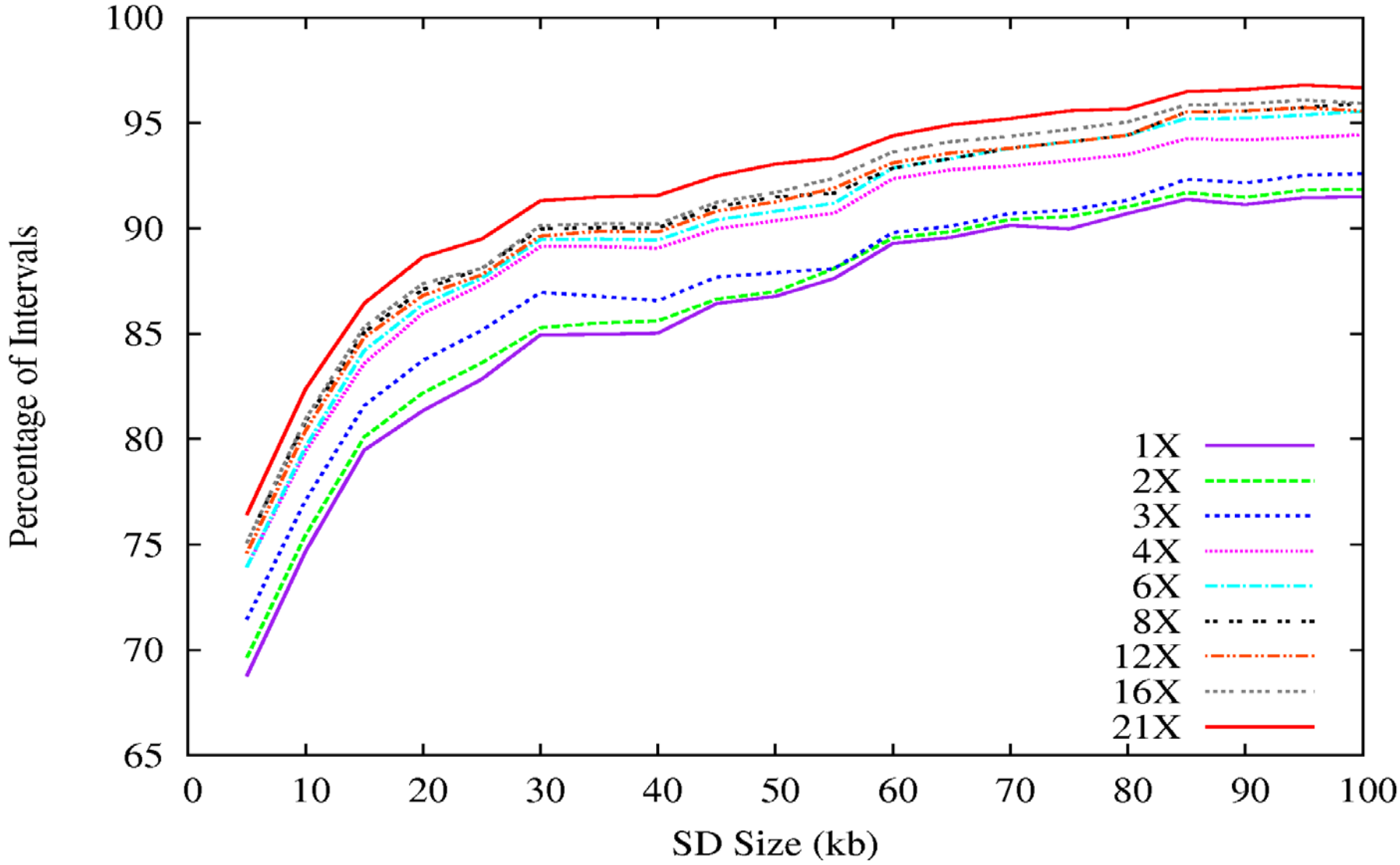
NA12892 (Solexa)



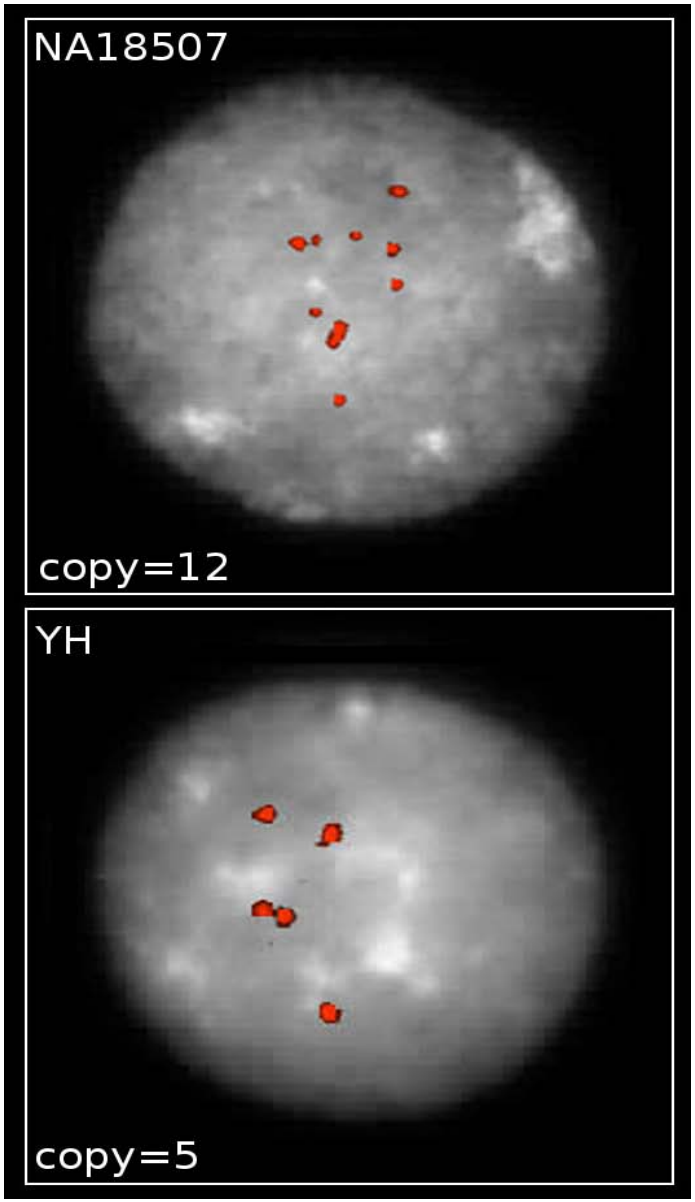
•Two known ~70 kbp CNPs, CNP#1 duplication absent in Venter but predicted in Watson and NA12878, CNP#2 present mother but neither father or child

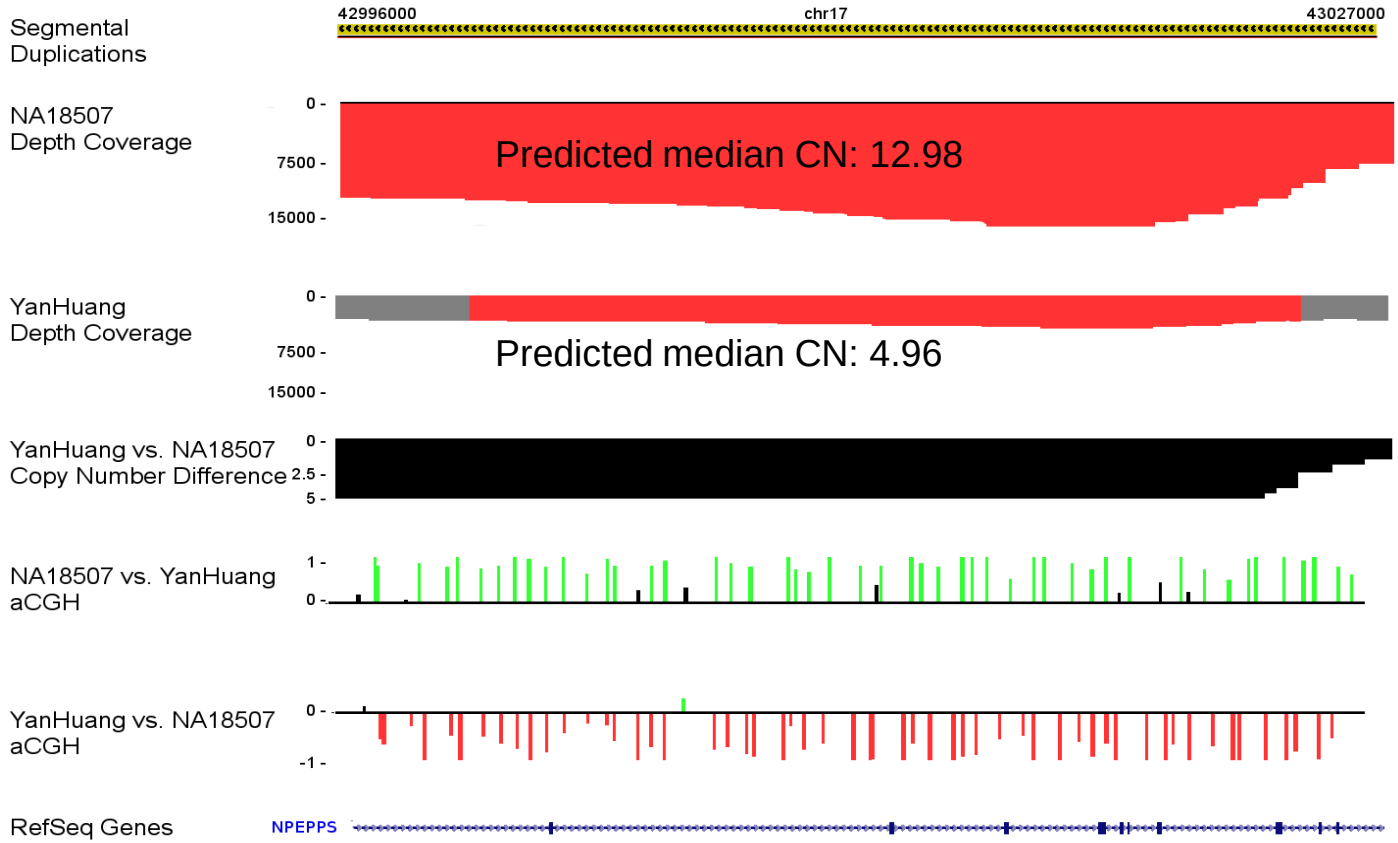
Alkan, Nat. Genet, 2009

Sequence coverage and detection power



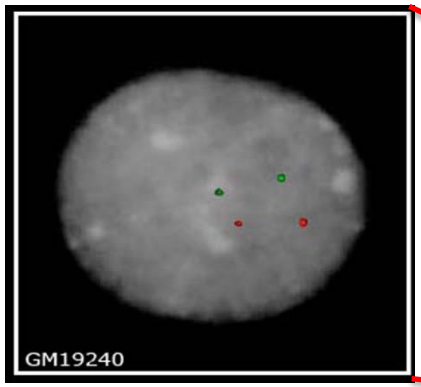
Absolute Copy Number Validation

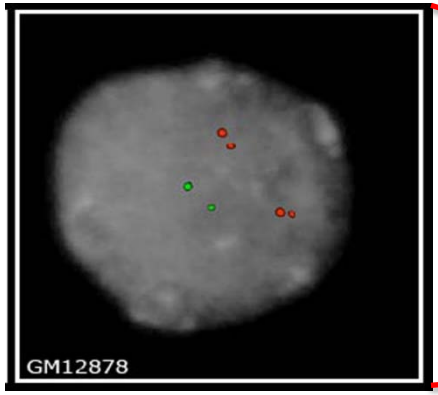


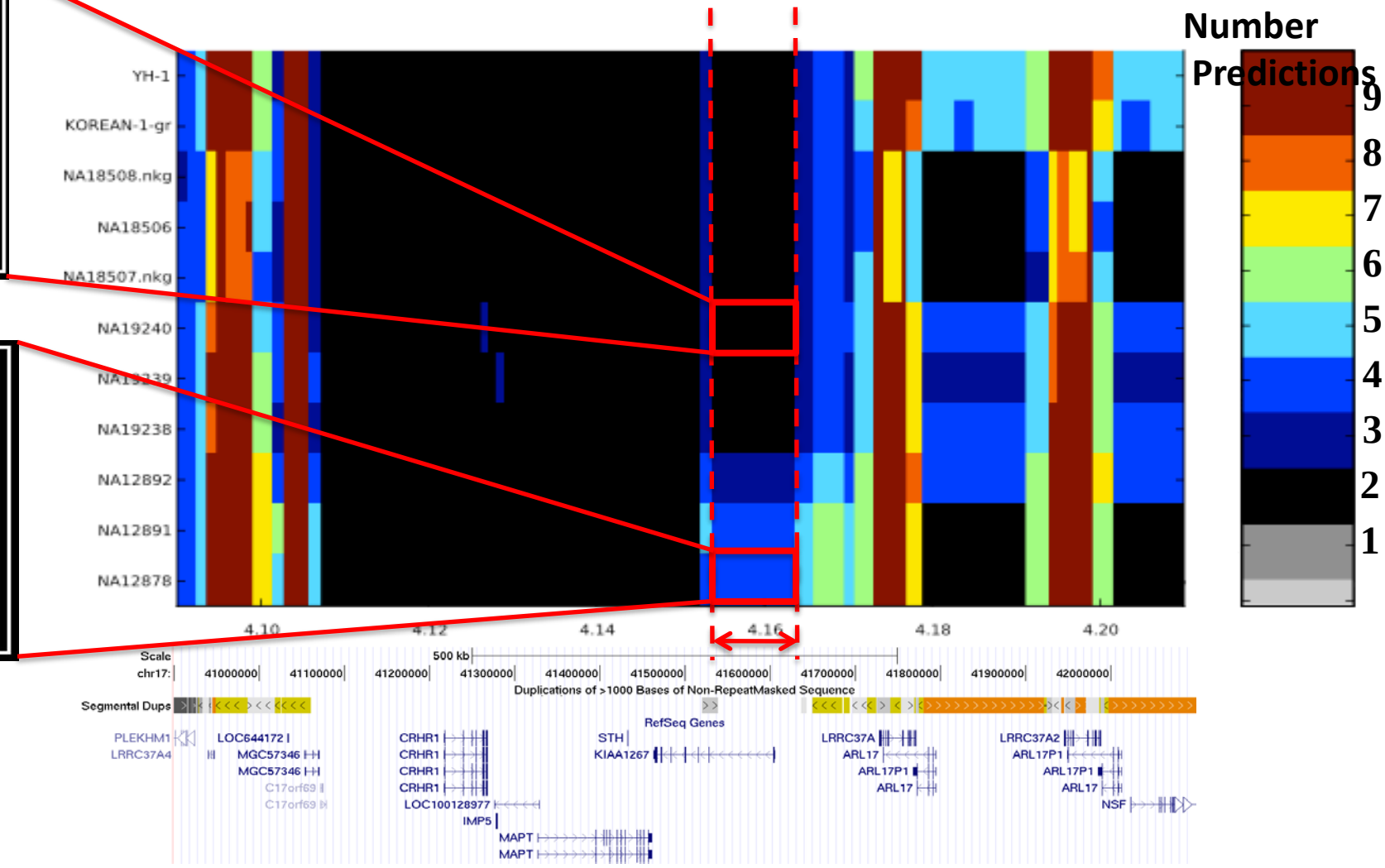


Read-Depth CNV Heat Maps vs. FISH

Interphase FISH

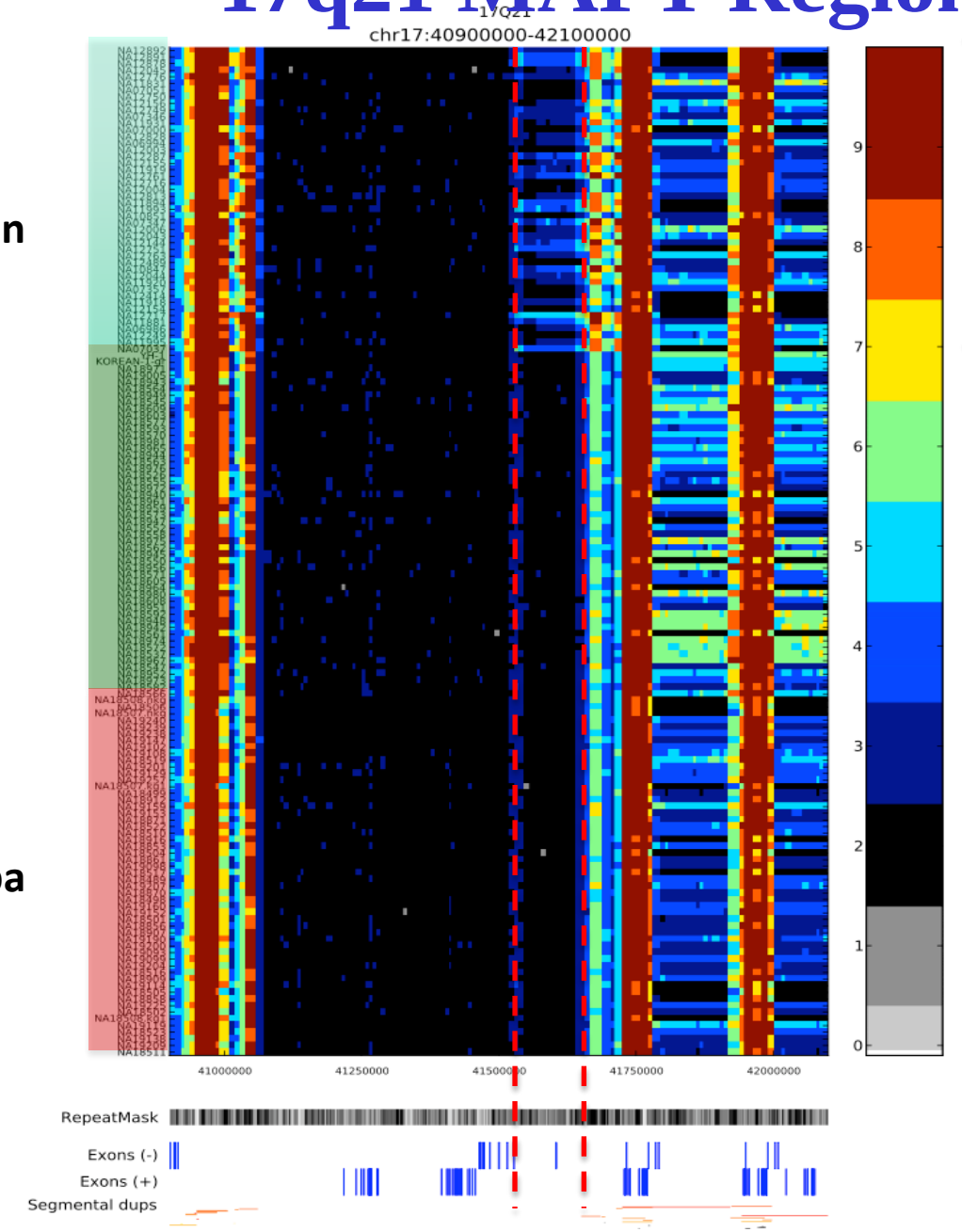






- 72/80 FISH assays correspond precisely to read-depth prediction (>20 kbp)
- 80/80 FISH assays correspond precisely to+/- 1 read-depth prediction

17q21 MAPT Region for 150 Genomes

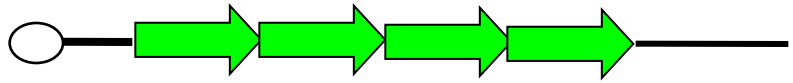


71% of Europeans carry at least Partial duplication distal (17q21 associated)—all inversions carry the duplication

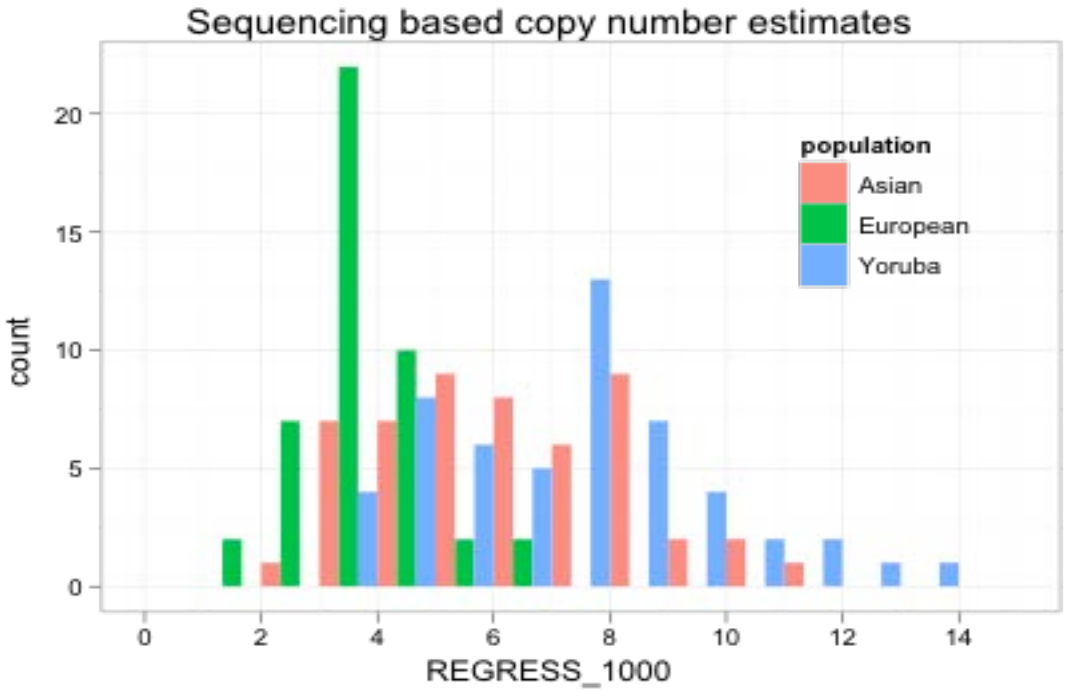
24% of Asians are hexaploid for NSF gene N-ETHYLMALEIMIDE-SENSITIVE FACTOR potentially important in synapse membrane fusion; NSF (decreased expression in schizophrenia brains (Mimics, 2000), Drosophila mutants results in aberrant synaptic transmission)

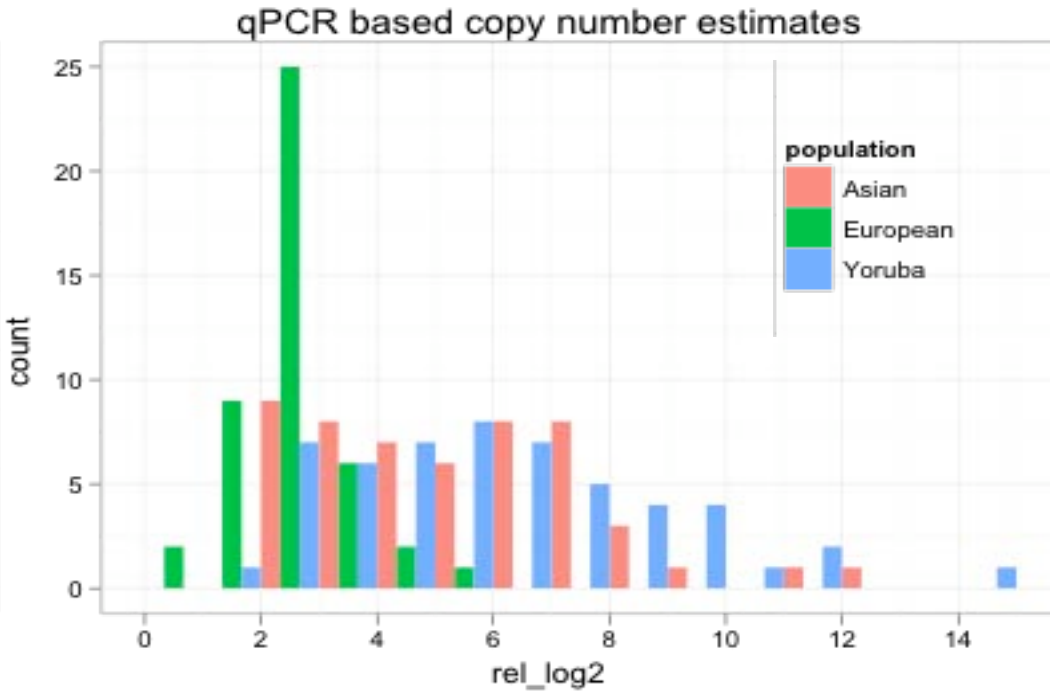
Sudmant et al., 2010, Science

Read-Depth vs. Quantitative PCR



CCL3L1—chemokine ligand 3-like (1.9 kbp)

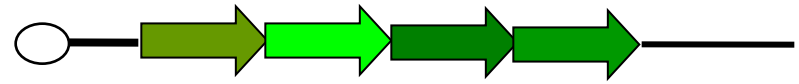




- **Tested 155 genomes read-depth (1-2 X coverage) vs. QPCR**
- **r^2=0.93 between sequence and quantitative PCR estimates**

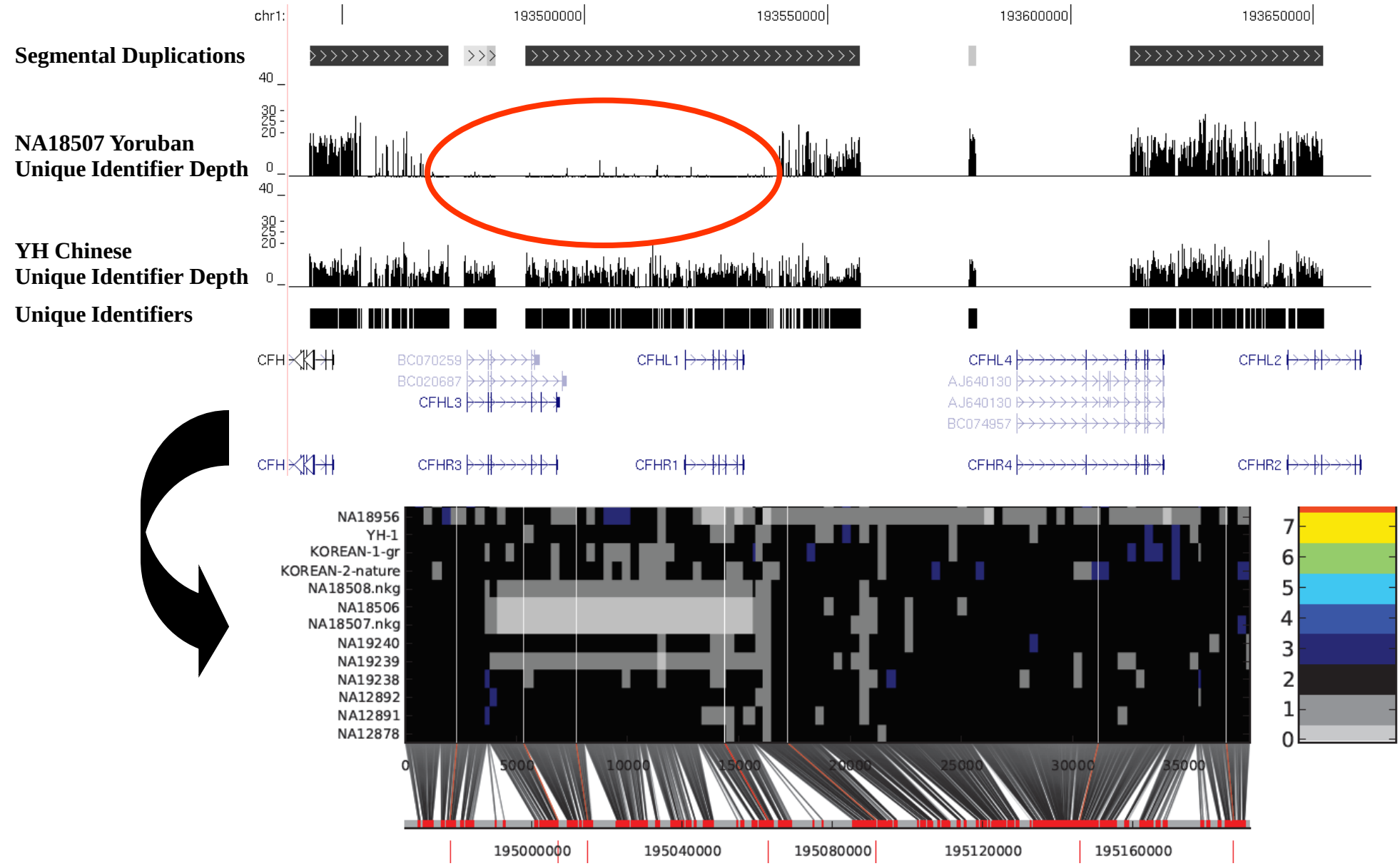
Unique Sequence Identifiers Distinguish Copies

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

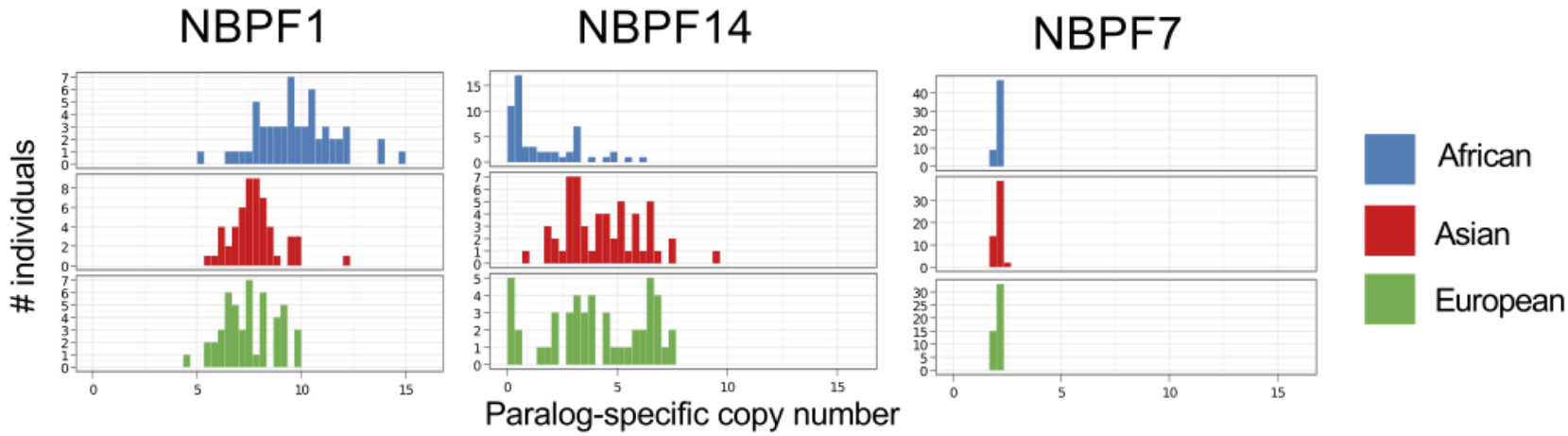


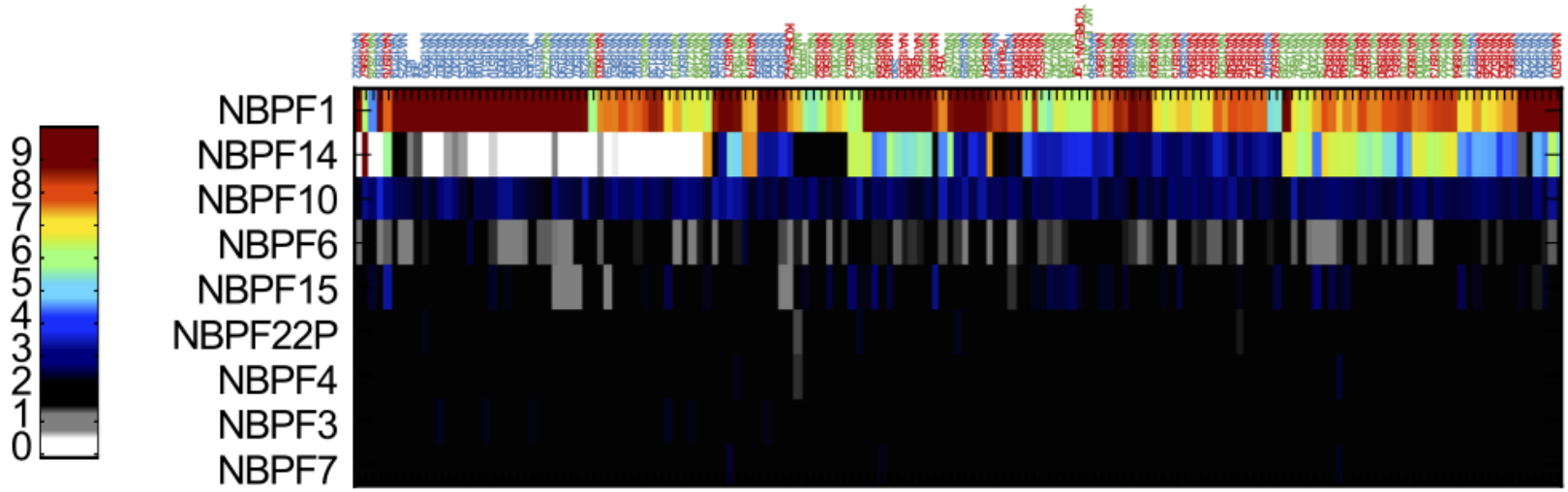
- Self-comparison identifies 3.9 million singly unique nucleotide (SUN) identifiers in duplicated sequences
- Select 3.4 million SUNs based on detection in 10/11 genomes=informative SUNs=paralogous sequence variants that are largely fixed
- Measure read-depth for specific SUNs--genotype copy-number status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

SUN Read Depth for CFHR Gene Cluster

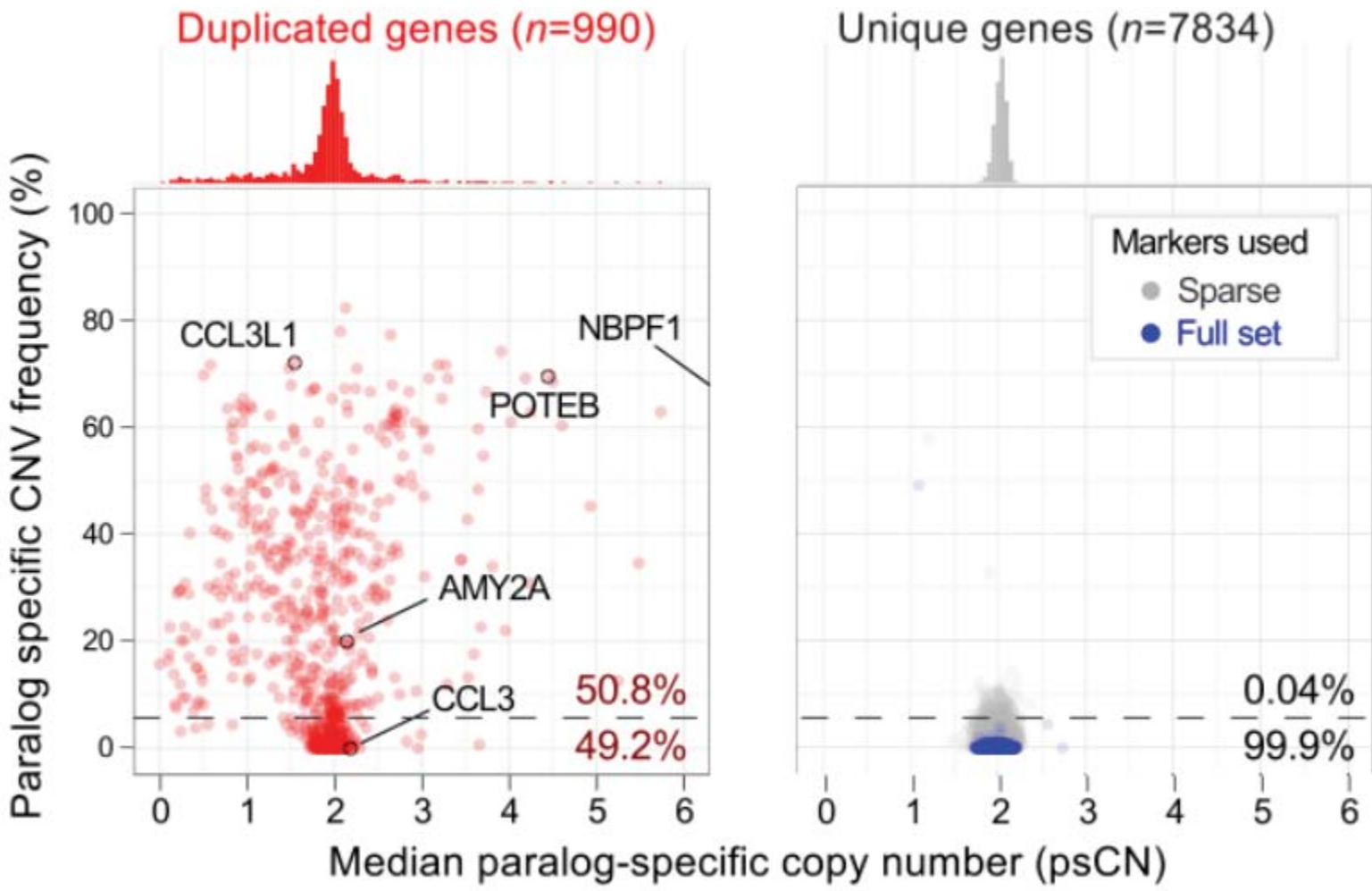


NBPF Gene Family Diversity

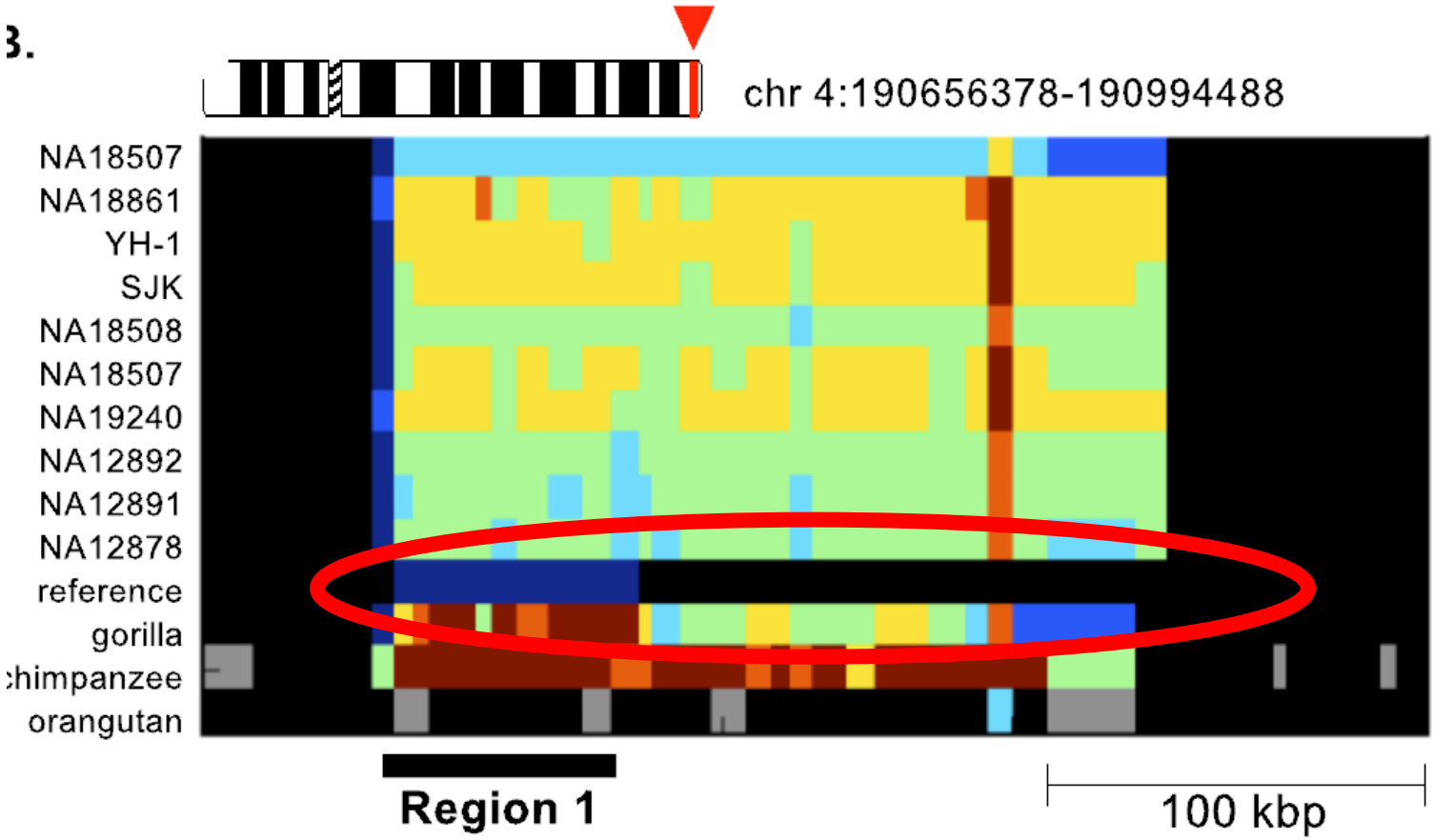


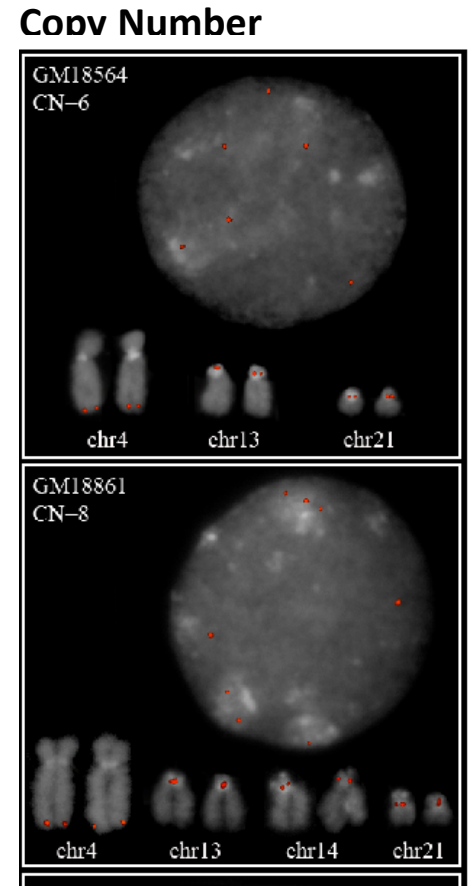


Copy Number Polymorphism Landscape
Duplicated vs. Unique Gene



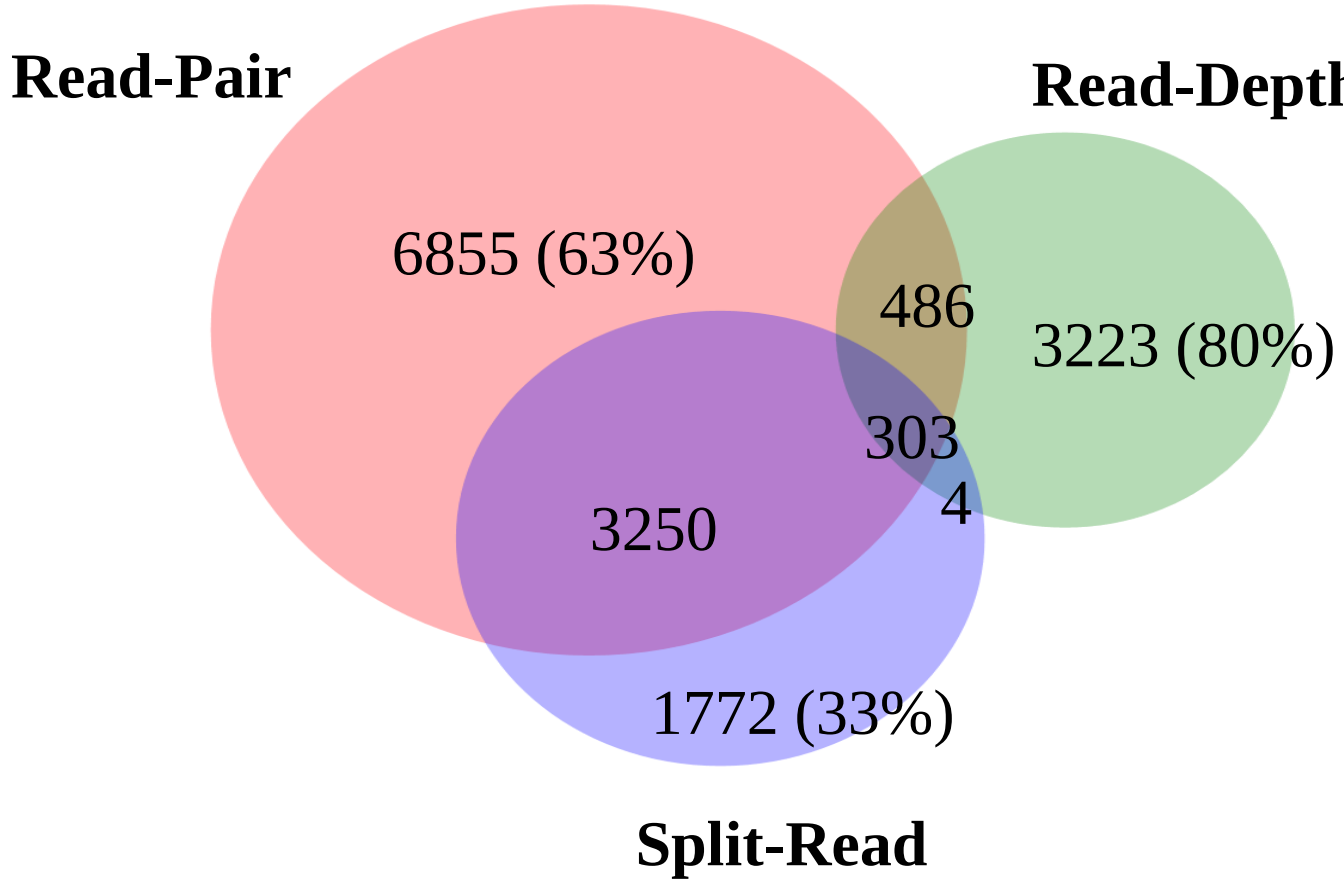
Missing Human Genomic Sequence





- 310 kbp segment 4-7 copies in humans but represented once in assembly
- 173 regions of collapse identified (10 are > 100 kbp)
- 44 “hidden” gene families.

No NGS method is comprehensive



1000G Consortium 2010, Mills *et al.* 2011

Fosmid SV Project vs. 1000 Genomes Project

	Deletions		Insertions		Inversions	
Sample	Sequenced Fosmid Variants	Found by 1000 Genomes 4X Pilot	Sequenced Fosmid Variants	Found by 1000 Genomes 4X Pilot	Sequenced Fosmid Variants	Found by 1000 Genomes 4X Pilot
NA12156	55	45	19	4	8	0
NA12878	59	50	31	5	10	0
NA18507	22	16	4	0	6	0
NA18517	13	10	2	0	11	0
NA18555	80	66	40	7	12	0
NA18956	182	158	174	19	7	0
NA19129	69	57	41	9	12	0
NA19240	89	75	89	13	14	0
Total	**559**	**447**	**400**	**57**	**82**	**0**

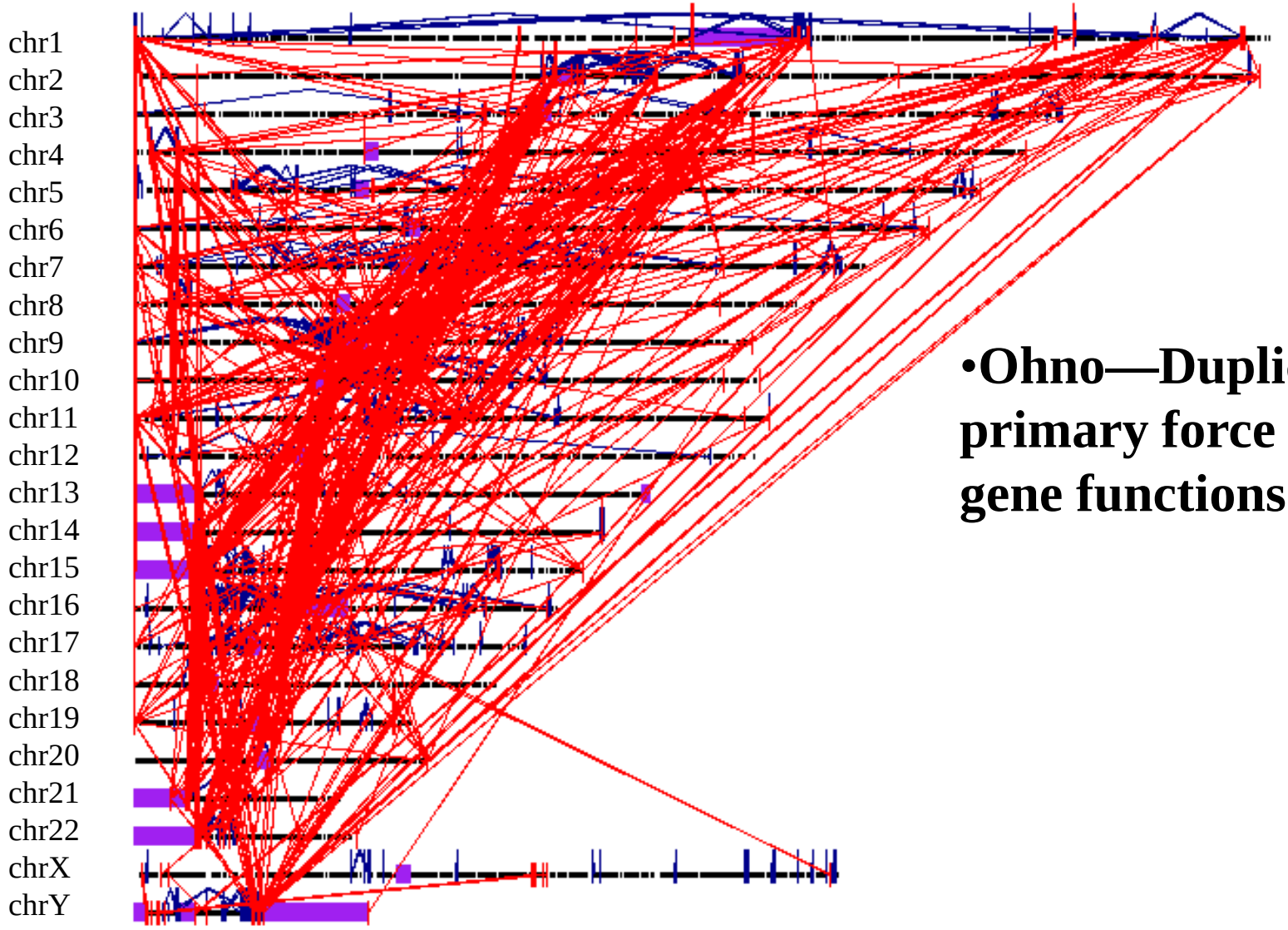
80% **14%** **0%**

- **Half of the events that overlap are resolved at the sequence level**
- **Need: increased coverage, longer-reads, larger libraries and further algorithmic development**

Summary

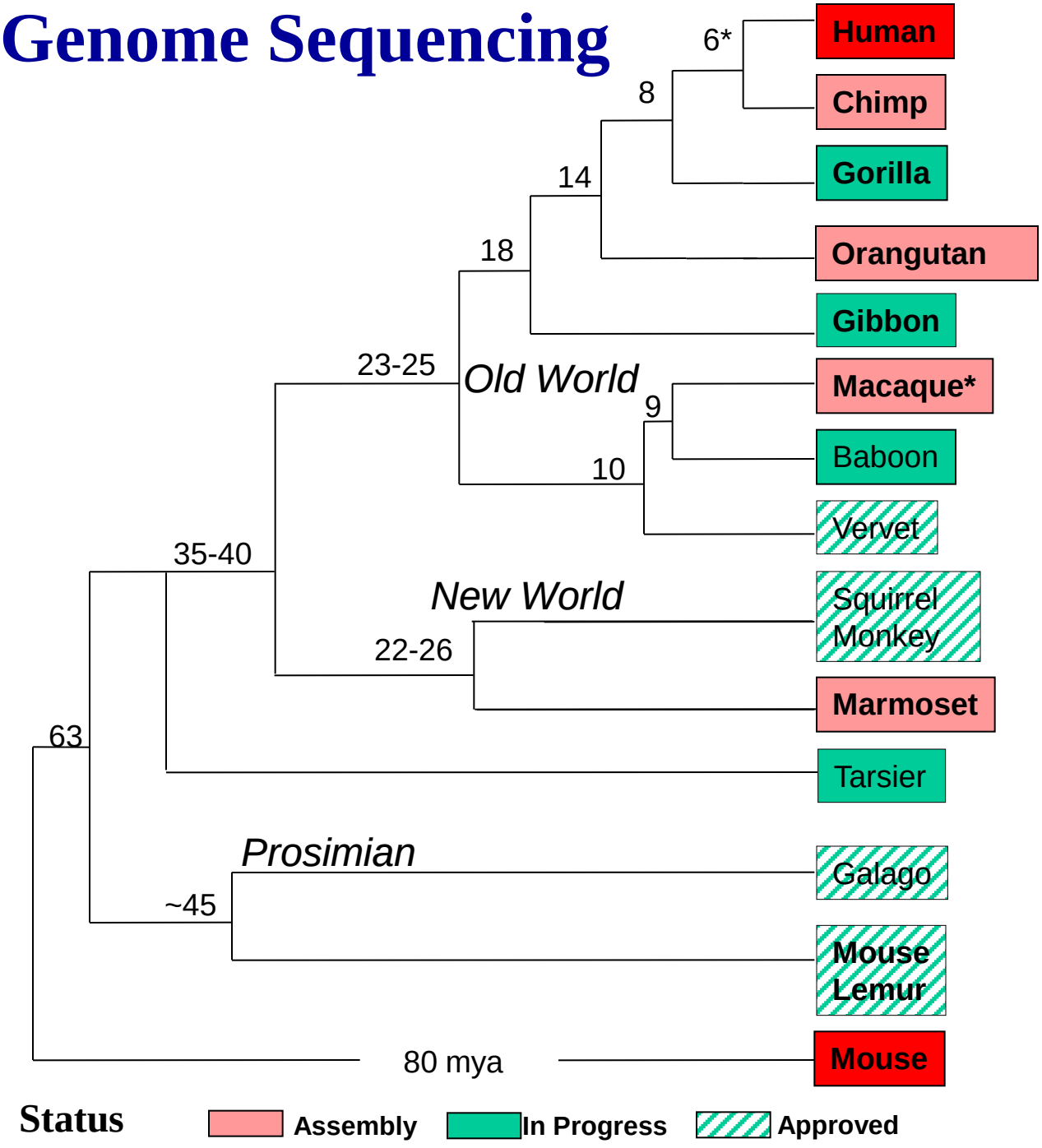
- Read-pair and read-depth NGS approaches
  - narrow the size spectrum of structural variation
  - lead to more accurate prediction of copy-number
  - unparalleled specificity in genotyping duplicated genes (reference genome quality key)
- Approaches
  - Multiple methods need to be employed—often methods to a specific class of variant
  - Tradeoff between sensitivity and specificity
  - Complexity not fully understood

When, How and Why?

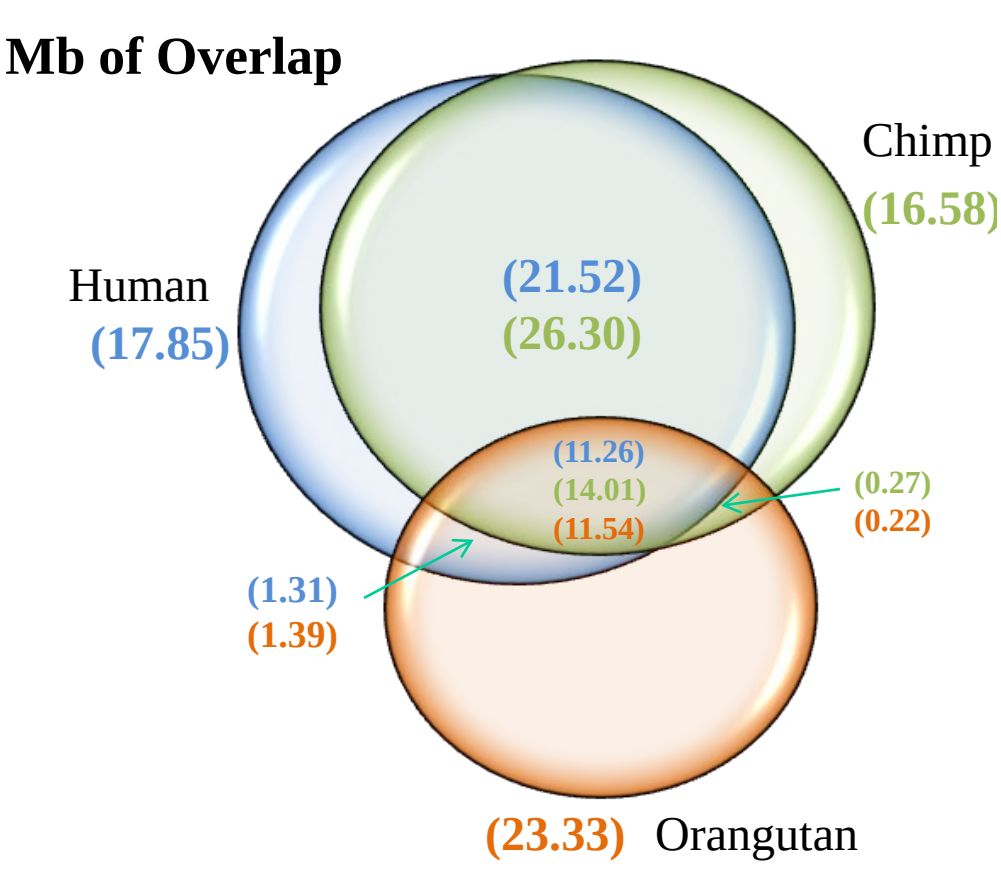


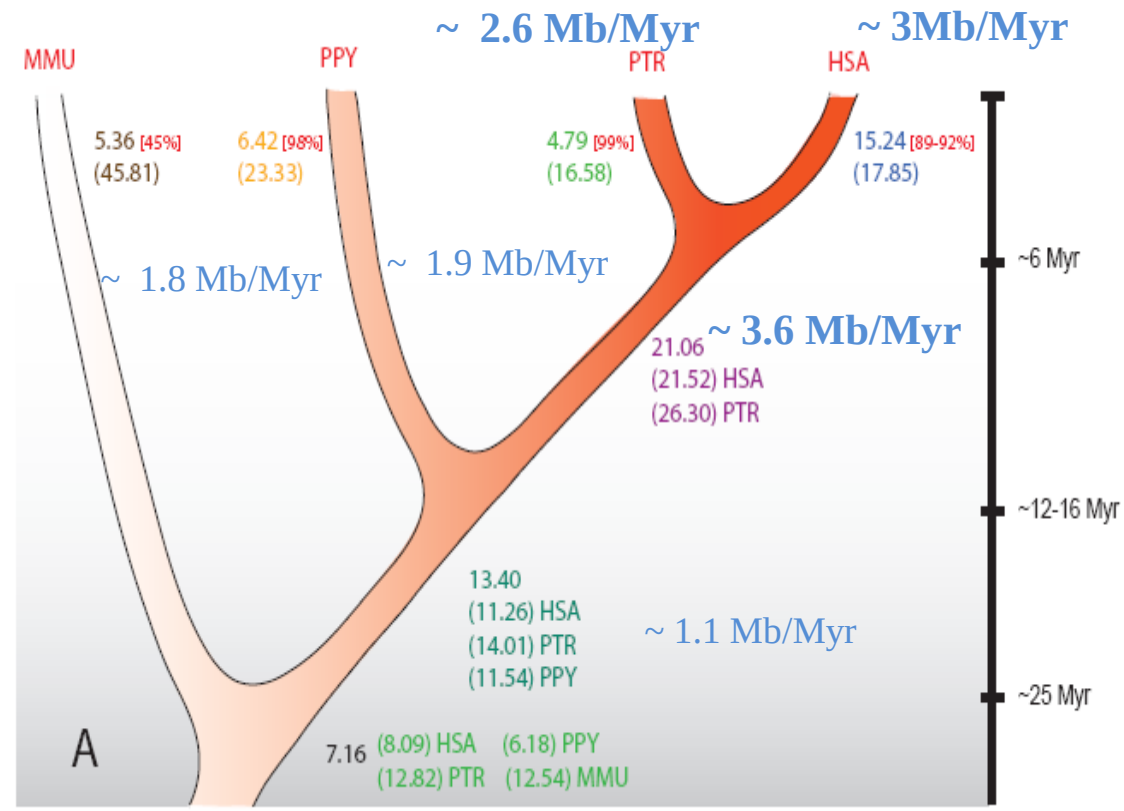
- **Ohno—Duplication is the primary force by which new gene functions are created**

Primate Genome Sequencing Status



A Hominoid Duplication Acceleration Hypothesis



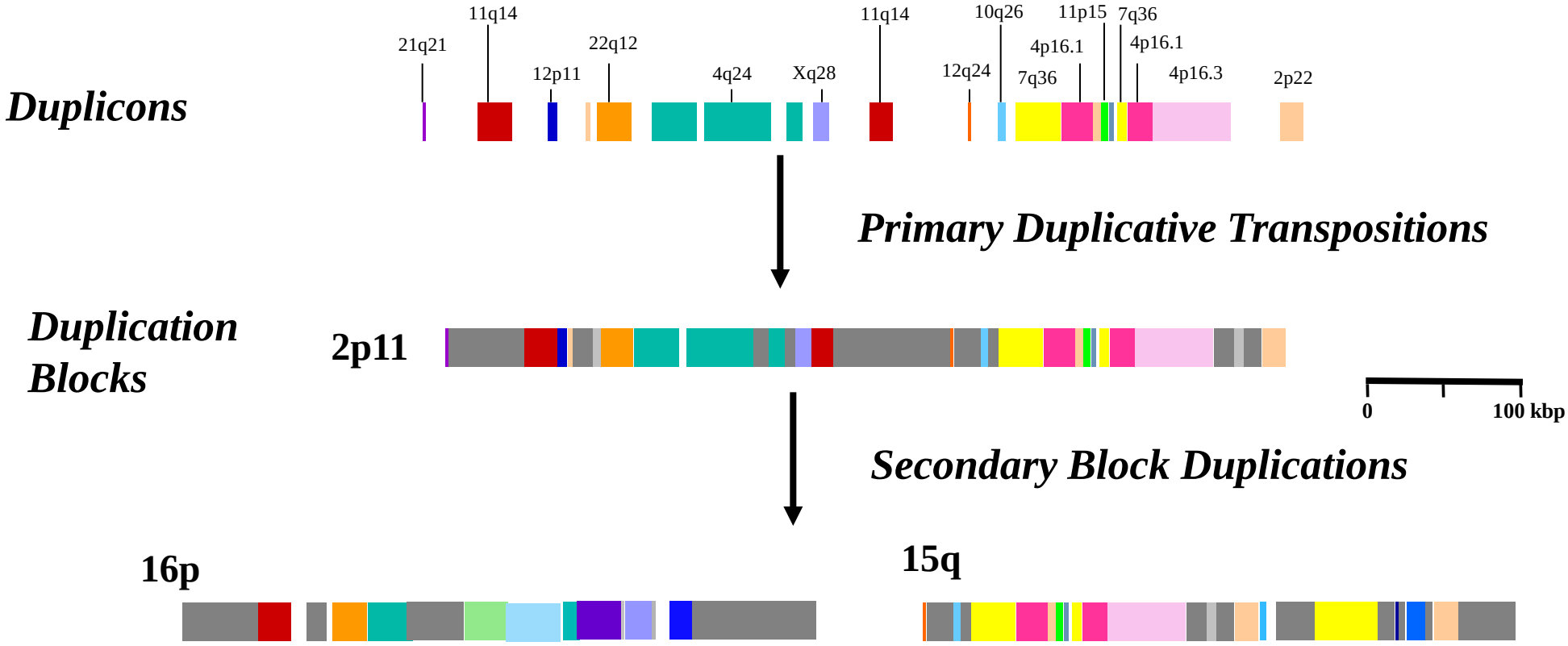


SDs > 20 kb

- **Experimental data suggest a 2X acceleration in de novo segmental duplications in common ancestor of human, chimpanzee and gorilla but after divergence from orangutan**
- **Develop a maximum likelihood estimating 20% homoplasy**

Tomas-Marques et al., Nature, 2009.

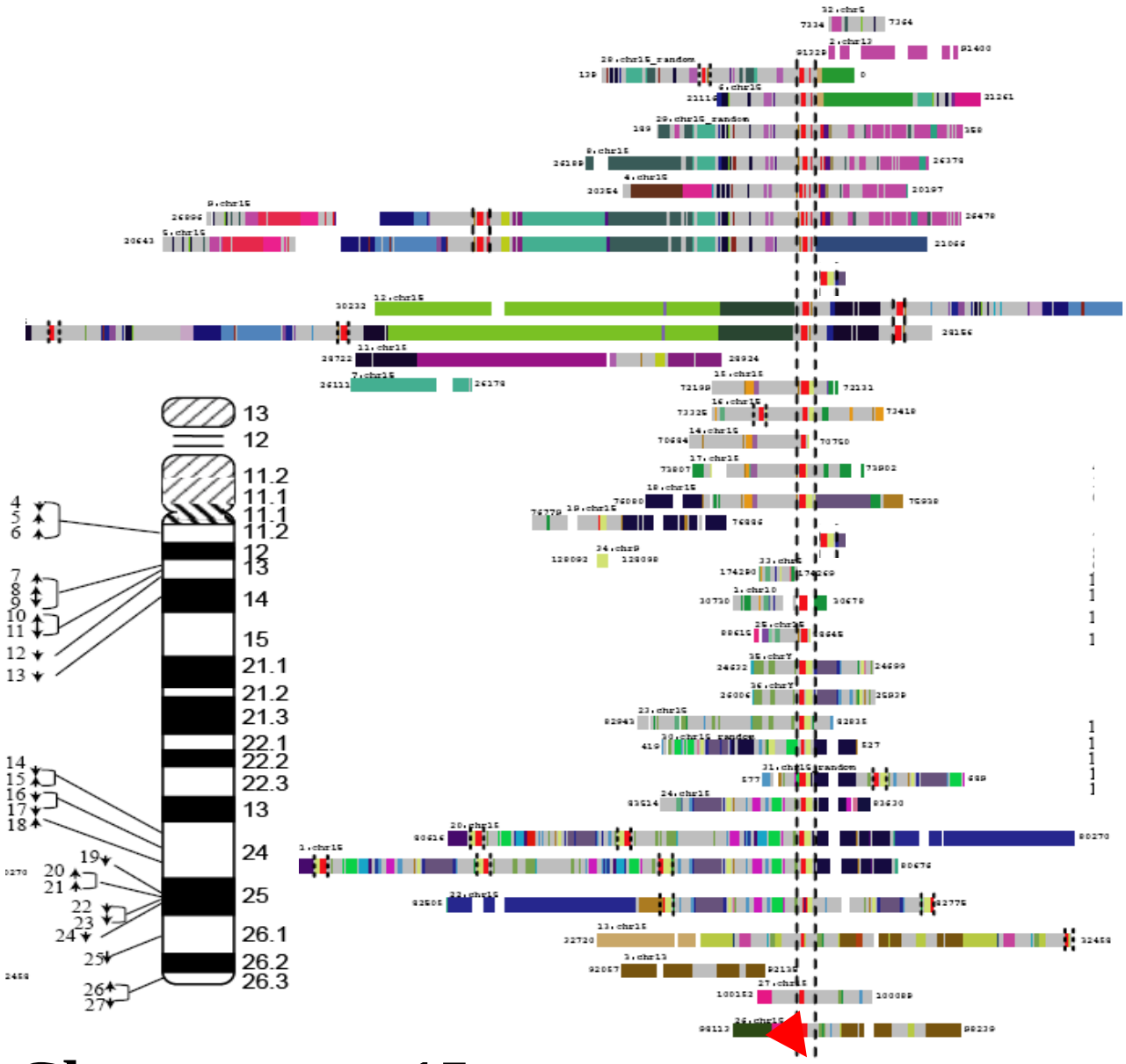
Architecture of Human Duplication Blocks



- A mosaic of recently transposed duplications
- Duplications within duplications.

Goal: Identify Ancestral States.

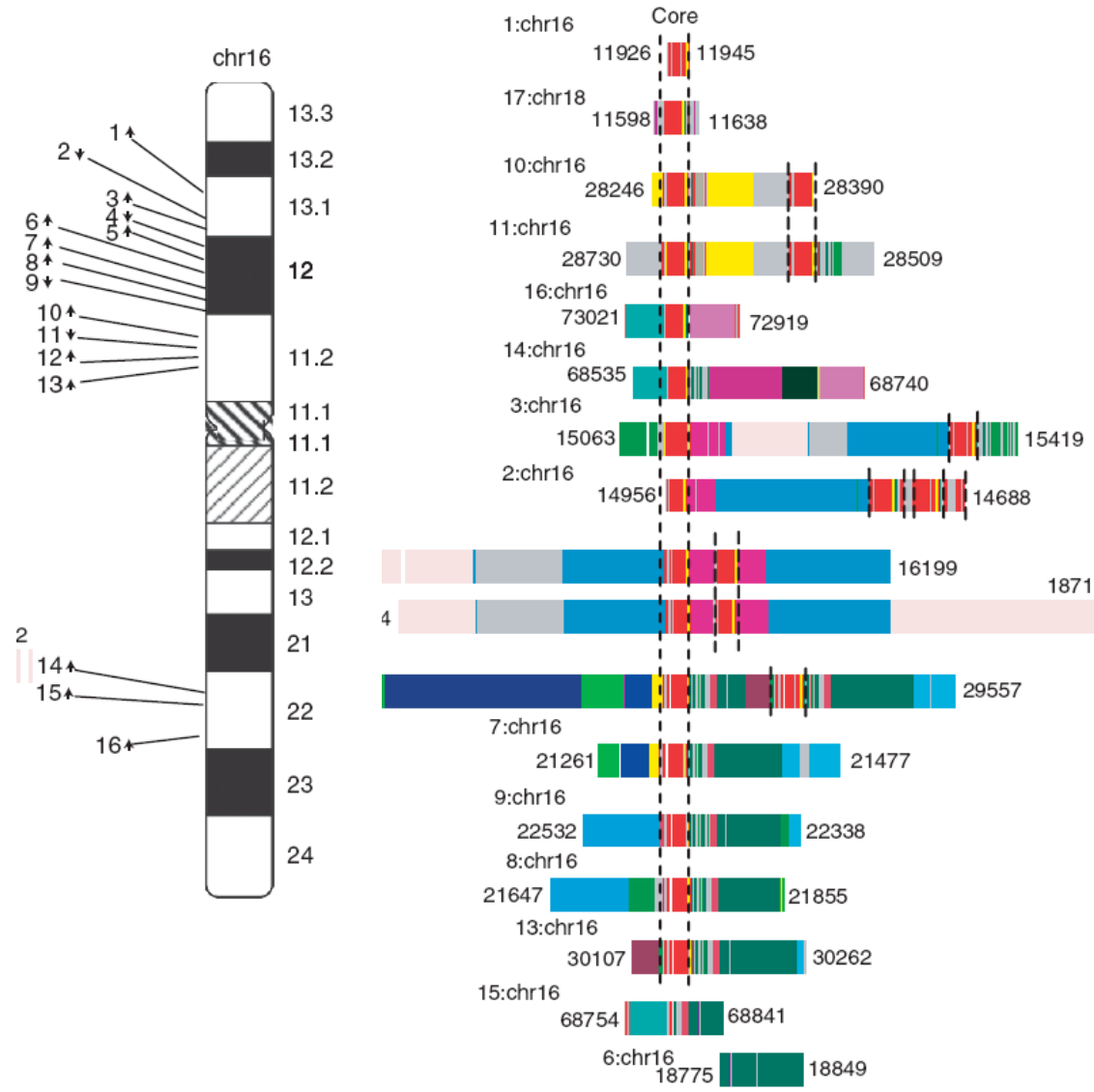
Duplication Blocks are organized around Core Duplicons



- Core sequences defined as top 90% of duplicon density
- Define the "focal" point for intrachromosomal duplications
- Cores frequently duplicated as solo elements, flanking duplicons almost never exist without core.
- Cores are enriched 4-5 fold for annotated genes and spliced ESTs when compared to flanks

Jiang and Tang et al., Nat. Genet, 2007.

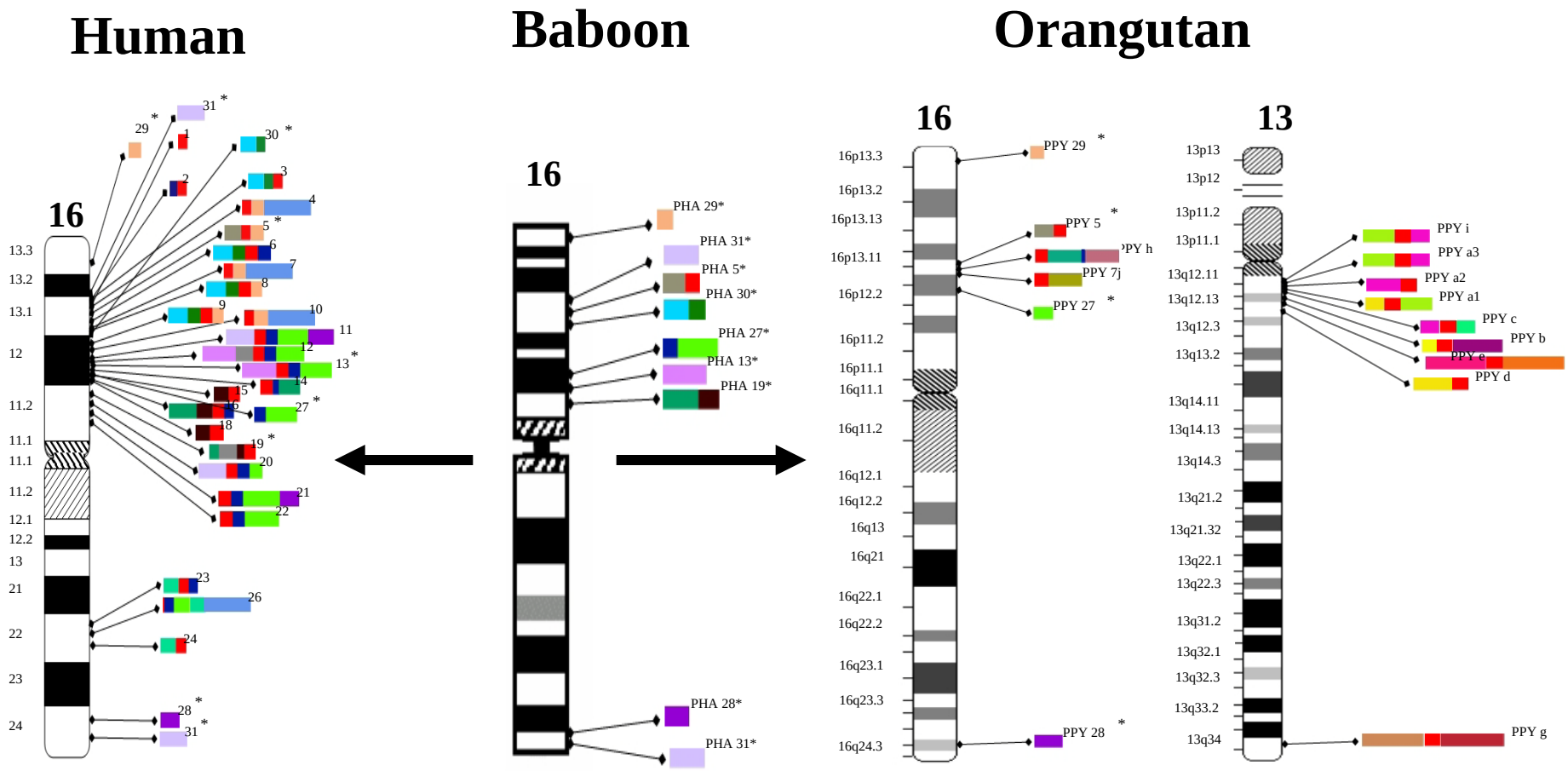
Core Duplicon on Human Chromosome 16 (LCR16a)



- Chromosome 16 specific
- Duplications are 19-20 kb in size
- Core duplicon harbors the NPIP gene family which shows positive selection in humans and great apes.
- NPIP gene family has evolved 50X faster in terms of aa Composition than a gene under Purifying selection

Johnson *et al.*, (2001) *Nature*

LCR16a Core Structural Evolution



*** Location of Ancestral Locus**

- **LCR16a duplicon mobilized to different chromosomes and independently transduced flanking sequences.**
- **Duplication blocks of different flanking content with exception of core**

Great-Ape/Human Specific Gene Families Map to Core Duplicons

- **NPIP** (Nuclear Pore Interacting Protein) (Johnson et al, 2001)—evolved 50 X faster than genes under purifying selection—expressed in regions of active neurogenesis
- **TRE2** Fusion protein –juxtaposition of two duplications (TBC1D3 & USP6) expressed only in great apes—marker for bladder cancer (Paulding et al., 2003)—regulates EGF-mediated cellular growth
- **RANGP2** (Ran GTPase binding) ancestral progenitor nuclear pore (Ciccarelli, 2005) –function unknown
- **DUF1220 or NBPF11** (neuroblastoma breakpoint gene family) expressed preferentially in specific neurons (Popescu et al., 2006) associated with pediatric neuroblastoma

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

Summary

- Human recent duplication architecture predisposes our genome to high frequency of *de novo* or rare mutations of large effect on neurocognitition
- Most of human duplication architecture has emerged and spread between 10-15 million years—data are consistent with a burst around the time of speciation.
- Great-ape specific gene families –function not understood but show signatures of selection and radical changes in expression
- **Core Duplicon Hypothesis**: intrachromosomal expansion focused on a small subset of sequences; disasdvantage offset by the advantage of newly minted genes embedded within the core—neuro-oncogenes?

Overall Summary

- **Prologue**: Role of CNVs in human disease—two models common and rare—recurrence, hotspot bias and variability in expressivity
- **Methods**: Read-pair and read-depth methods to characterize SVs within genomes—need a high quality reference—not a solved problem.
- **Epilogue**: evolution of the complex human architecture that predisposes to disease coupled to gene innovation

Acknowledgements

Eichler Lab
Francesca Antonacci
Peter Sudmant
Can Alkan
Tomas Marques
Megan Dennis
Xander Nuttle
Karyn Meltz-Steinberg
Saba Sajjadian
Laura Vives
Carl Baker
Ze Cheng
Lin Chen
Maika Malig

WashU
Wes Warren
Tina Graves
LaDeana Hillier
Rick Wilson

Broad
Michael Zody

Bari
Mariano Rocchi
Nicoletta Archidiacono

Yale
Ken Kidd
Michael Donnelly

U. Nijmegen
Bert de Vries
David Koolen

Signature Genomics
Lisa Shaffer
Jill Rosenfeld

1000 Genomes Project

U Penn
Sarah Tishkoff
Laura Scheinfeldt
Bill Beggs