Genome Structural Variation

Evan Eichler
Howard Hughes Medical Institute
University of Washington

January 18th, 2013, Comparative Genomics, Český Krumlov

Disclosure: EEE was a member of the Pacific Biosciences Advisory Board (2009-2013)
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.

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. Aitman1, Rong Dong1, Timothy J. Vyse2, Penny J. Norsworthy1, Michelle D. Johnson1, Jennifer Smith1, Jonathan Mangion1, Cheri Roberton-Lowe1, Amy J. Marshall1, Enrico Petretto1, Matthew D. Hodges1, Gurjeet Bhangal3, Sheetal G. Patel1, Kelly Sheehan-Rooney1, Mark Duda1, Paul R. Cook1, David J. Evans3, Jan Domen3, Jonathan Flint4, Joseph J. Boyle5, Charles D. Pusey5 & H. Terence Cook5

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* Raquel 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. Ahuja1§

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

Andrew J Sharp1, Sierra Hansen1, Rebecca R Selzer2, Ze Cheng1, Regina Regan3, Jane A Hurst4, Helen Stewart1, Sue M Price4, Edward Blair4, Raoul C Hennekam5, Carrie A Fitzpatrick6, Rick Segraves8, Todd A Richmond2, Cheryl Guiver2, Donna G Albertson6,8, Daniel Pinkel8, Peer S Eis2, Stuart Schwartz7, Samantha J L Knight8 & Evan E Eichler1

Associate 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., Ragheb J. Fossdal, B.Sc., Eivind Saemundsson, B.A., Heini 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., Ph.D., Rudolph E. Tanzi, Ph.D., 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.Ed., Ph.D., and Mark J. Daly, Ph.D., for the Autism Cons N Engl J Med 2008;358:667-75

Rare chromosomal deletions and duplications increase risk of schizophrenia

The International Schizophrenia Consortium*


Large recurrent microdeletions associated with schizophrenia

Nature 455:232-6 2008

Strong Association of De Novo Copy Number Mutations with Autism

Jonathan Sebat,2* B. Lakshmi,2 Dheeraj Malhotra,3* Jennifer Troge,4* Christa Lese-Martin,2 Tom Walsh,5 Boris Yamrom,6 Seungtae Yoon,7 Alex Krasnitz,8 Jude Kendall,3 Anthony Leotta,9 Deepa Pati,9 Ray Zhang,3 Yoon-Ha Lee,3 James Hicks,3 Sarah J. Spence,3 Annette T. Lee,9 Kajia Puura,3 Terho Lehtimäki,7 David Ledbetter,7 Peter K. Gregersen,1 Joel Bregman,1 James S. Sutcliffe,3 Vaidelah Jobanputra,10 Wendy Chung,10 Dorothy Warburton,10 Mary-Clare King,3 David Skuse,6 Daniel H. Geschwind,12* Conal Gilliam,3†

Kenny Ye,14 Michael Wigler1†

SCIENCE VOL 316 20 APRIL 2007
**Perspective: Segmental Duplications (SD)**

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

Non Allelic Homologous Recombination (NAHR)

Human Disease
Triplosensitive, Haploinsufficient and Imprinted Genes
Importance: Evolution of New Gene Function

GeneA

Mutation

Maintain old Function

Duplication

GeneA’

Acquire New/Modified Function

Mutation

Loss of Function

Mutation
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

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**

- **GAMETES**
- **NAHR**
- **Human Disease**
  - Triplosensitive, Haploinsufficient and Imprinted Genes

**Genomic Disorders:** A group of diseases that results from genome rearrangement mediated mostly by non-allelic homologous recombination. (*Inoue & Lupski*, 2002).
DiGeorge/VCFS/22q11 Syndrome

1/2000 live births
180 phenotypes
75-80% are sporadic (not inherited)
• 130 candidate regions (298 Mb)
• 23 associated with genetic disease
• Target patients array CGH

Human Genome Segmental Duplication Map

Bailey et al. (2002), Science
~14.2% of genetic cause of developmental delay explained by large CNVs (>500 kbp)

Cooper et al., Nat. Genet, 2011
### Model #2: Copy Number Polymorphisms and Disease

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

- Disease CNPs enriched within duplicated sequences.
Structural Variation and Enriched Gene Functions

- Drug detoxification:
  - glutathione-S-transferase, cytochrome P450, 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
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
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

### Structural Variation Diversity

**Eight Distinct Complex Haplotypes**

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Description</th>
<th>Length (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1.2</td>
<td></td>
<td>1.29</td>
</tr>
<tr>
<td>H1.1</td>
<td></td>
<td>1.08</td>
</tr>
<tr>
<td>H1.3</td>
<td></td>
<td>1.50</td>
</tr>
<tr>
<td>H1D</td>
<td></td>
<td>1.29</td>
</tr>
<tr>
<td>H1D.3</td>
<td></td>
<td>1.49</td>
</tr>
<tr>
<td>H2.1</td>
<td></td>
<td>1.19</td>
</tr>
<tr>
<td>H2.2</td>
<td></td>
<td>1.41</td>
</tr>
<tr>
<td>H2D</td>
<td></td>
<td>1.80</td>
</tr>
</tbody>
</table>

The diagram illustrates the diversity of structural variation in the genome, with each haplotype showing the distribution of genes such as CRHR1, MAPT, and KIAA1267, with variants ending in NSF (Non-Synonymous Frameshift). The San population is highlighted in the H2.1 haplotype, indicating specific genomic diversity in this group.

---

**San**
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
# 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

## Sequencing-based
- Read-depth: Bailey et al, 2002
- Sanger sequencing: Mills et al., 2006

## Single molecule analysis
- **Optical mapping:** Teague et al., 2010
Array Comparative Genomic Hybridization

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)

Human chromosome 3 position ~55 kbp
SNP Microarray detection of Duplication (Illumina)
Clone-Based Sequence Resolution of 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

a) Insertion
Deletion
Inversion

b) Fosmid Size Distribution (G248)

<32 kb Putative Insertion
>48 kb Putative Deletion

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

- Fosmid ESP Clone sequencing (Kidd et al., N=1,206)
- Array CGH (Conrad et al., N=1,128)
- 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

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

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

Alkan et al., *Nat Rev Genet*, 2011
Computational Approaches are Incomplete
159 genomes (2-4X) (deletions only)

Read-Pair

6855 (63%)

3250

1772 (33%)

Read-Depth

3223 (80%)

486

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

• Exome vs. Genome

• 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 (e.g., mrFAST, mrsFAST)

Illumina Sequence

Celera’s
27.3 million reads

Random Genome Sample

Reference Sequence

Sequence to Test

unique
duplicated

Bailey et al., Science, 2002
Personalized Duplication or Copy-Number Variation Maps

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

Alkan, Nat. Genet, 2009
Copy number from short read depth

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

Interphase FISH

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

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

CCL3L1—chemokine ligand 3-like (1.9 kbp)
Unique Sequence Identifiers Distinguish Copies

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

**NBPF1**

- African
- Asian
- European

**NBPF14**

- African
- Asian
- European

**NBPF7**

- African
- Asian
- European

**Paralog-specific copy number**

- Number of individuals

---

**Genomic Region Visualization**

- NBPF1
- NBPF14
- NBPF10
- NBPF6
- NBPF15
- NBPF22P
- NBPF4
- NBPF3
- NBPF7
CNV Detection by Exome Read-Depth

Discard first 10-12 components of variance

Krumm et al., Genome Res., 2012
Detecting Smaller CNVs

- 5-fold increased sensitivity for CNVs $\leq 10$ kbp than high density SNP microarray.

CoNIFER software: http://conifer.sourceforge.net/index.html
Going Forward

1) **Focus on comprehensive assessment of genetic variation**—NGS are indirect and do not resolve structure by provide specificity and excellent dynamic range response.

2) **High quality sequence resolution of complex structural variation to establish alternate references/haplotypes**—often show extraordinary differences in genetic diversity

3) **Technology advances in whole genome sequencing “Third Generation Sequencing”**: Long-read sequencing technologies with NGS throughput in order to sequence and assemble genomes *de novo*
WG Sequencing Recent Gene Duplicates is difficult.

She et al. Nature, 2005:
Shorter-Read Technologies further Limit.

- SOAP-de novo Assembly YH—93% of SDs missing
- Subsequent improvements in algorithms, Illumina read length, reads from longer inserts, fosmid pools all improve continuity but leave 75-81% of SDs missing or mis-assembled

Alkan et al. Nat. Methods, 2011, Chaisson et al. unpublished
Single-Molecule Real-Time Sequencing (SMRT)

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

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

Chen-Shan Chin¹, David H Alexander¹, Patrick Marks¹, Aaron A Klammer¹, James Drake¹, Cheryl Heiner¹, Alicia Clum², Alex Copeland², John Huddleston³, Evan E Eichler³, Stephen W Turner¹ & Jonas Korlach¹

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

Chin et al. Nat. Methods, 2013
A Simple Experiment

- Select tiling path of BAC previously sequenced using Sanger and corresponding to region of Complex SD
- Sequence each clone (~200 fold) using on average 1 SMRT Cell and assemble using HGAP and QUIVER
- Compare Sanger and Pacbio assembly using BLASR
Results

- Accurate (QV>45) assembly of complex region of human genome by BAC– 125 differences—31/44 favor PacBio over Sanger
- Most differences are indels but one large scale collapse of 20 kbp region to 12 kbp

Huddleston et al. Genome Res, 2014
Strategy for Resolving Complex Regions

Select Clones by BAC-end Sequence Data

Nextera Illumina Sequencing (96-well format)

PacBio Sequencing of Tiling Path
Whole Genome Sequencing with PacBio

- CHM1—complete hydatidiform mole (CHM1)- “Platinum Genome Assembly”
- 10X Sequence coverage using RSII P5/C3 chemistry

http://datasets.pacb.com/2013/Human10x/READS/index.html
Validated Breakpoint-Resolved Deletions

114.2 kb
Transitioning into the Centromeric Satellite

- Single 31.8 kb read mapping to edge of centromere on chromosome 16:
  - HSAT2RS anchor and extends 25 kbp into centromeric
  - Site of extensive copy number polymorphism and potential hotspot for rearrangements associated with cancer
  - Data suggest that 5 bp HSAT2 is organized into a 2.8 kbp HOR
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
• 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

Human (17.85)
(21.52)
(26.30)

Chimp (16.58)
(11.26)
(14.01)
(11.54)

(0.27)
(0.22)

(23.33) Orangutan
(1.31)
(1.39)

• 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
Great Ape Genome Diversity Project

- Deep genome sequencing of 79 wild and captive born great apes (6 species and 7 subspecies) and 10 human genomes
- 167 Mbp (83.6 million SNPs and 84.0 fixed SNVs)
- 469 Mbp affected by copy number
- 745 CNV; 1080 indels; 806 SNVs affect gene structure

Ape Segmental Duplication Patterns

- Sumatran Orangutan
- Bornean Orangutan
- Gorilla
- Bonobo
- Chimpanzee
- African American
- Luhya
- Yoruban
- Japanese
- Southern Han Chinese
- Beijing Han Chinese
- Mexican
- Columbian
- Puerto Rican
- British
- Finnish
- CEPH Europeans
- Toscani

[Graph showing segmental duplication patterns with frequency and copy number]
Rate of Duplication

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

\[ p = 9.786 \times 10^{-12} \]
Rate of Deletion

\[ *p = 4.79 \times 10^{-9} \]
- A mosaic of recently transposed duplications
- Duplications within duplications.
- Potentiates “exon shuffling”, regulatory innovation
• 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

Johnson et al., PNAS, 2006
Core Expansion Model

Time

Locus 1
X \[\text{Core Duplicon}\]
Y

Locus 2
A \[\text{Lineage Specific}\] B

Locus 3
D A \[\text{Ape/Human Shared}\] B E

Locus 4
F D A \[\text{Ape/Human Shared}\] B G
Human Great-ape “Core Duplicons” have led to the Emergence of New Genes

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

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

Marques-Bonet and Eichler, CSHL Quant Biol, 2008
Notable human-specific expansion of brain development genes.
Neuronal cell death: $p=5.7 \times 10^{-4}$; Neurological disease: $p=4.6 \times 10^{-2}$

Sudmant et al., *Science*, 2010
**SRGAP2 function**

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

Guerrier et al., *Cell*, 2009
SRGAP2 Human Specific Duplication

Dennis, Nuttle et al., Cell, 2012
SRGAP2C is fixed in humans
(n=661 individual genomes)
SRGAP2 duplicates are expressed

**RNAseq**

**In situ**
SRGAP2C duplicate antagonizes function

Charrier et al., Cell, 2012
Australopithecus

- Sahelanthropus
- Orrorin
- Ardipithecus
  - A. anamensis
  - A. afarensis
  - A. aethiopicus
  - A. robustus
  - A. boisei
- Homo
  - K. platyops
  - A. garhi
  - A. africanus

- Homo habilis

- Fixed in human population and expressed in neurons
  - SRGAP2A
  - SRGAP2B,D^p
  - SRGAP2C

- 3.4 mya
- 2.4 mya

- ~350 cc
- ~1000 cc

Dennis, Nuttle et al. Cell (2012)
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