Genome Structural Variation

Evan Eichler
Howard Hughes Medical Institute
University of Washington

January 13th, 2017, Genomics Workshop, Český Krumlov
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
Genome Structural Variation

Deletion

Duplication

Inversion
Introduction

• **Genome structural variation** includes copy-number variation (CNV) and balanced events such as inversions and translocations—originally defined as $> 1$ kbp but now $> 50$ bp

• **Objectives**
  1. Genomic architecture and disease impact.
  2. Detection and characterization methods
  3. Primate genome evolution
Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans


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


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

Andrew J Sharp, Sierra Hansen, Rebecca R Selzer, Ze Cheng, Regina Regan, Jane A Hurst, Helen Stewart, Sue M Price, Edward Blair, Raoul C Hennekam, Carrie A Fitzpatrick, Rick Segraves, Todd A Richmond, Cheryl Guiver, Donna G Albertson, Daniel Pinkel, Peer S Eis

Stuart Schwartz, Samantha L Knight & Ivan E Eichler

Nature Genetics Volume 38 | Number 9 | September 2006

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., Ragnar Fossdal, B.Sc., Evald Saemundsson, B.A., Heimn Stefansson, Ph.D., Manuel A.R. Ferreira, Ph.D., Todd Green, B.S., Orah S. Platt, M.D., Douglas M. Rudfer, 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.Ed., Ph.D., and Mark J. Daly, Ph.D., for the Autism Consor

Rare chromosomal deletions and duplications increase risk of schizophrenia

The International Schizophrenia Consortium


Strong Association of De Novo Copy Number Mutations with Autism


Science Volume 316 | 20 April 2007
Perspective: Segmental Duplications (SD)

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

Non Allelic Homologous Recombination (NAHR)

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

- **GeneA**
  - Maintain old Function
  - Mutation

- Duplication

- Acquire New/Modified Function

- **GeneA’**
  - Mutation
  - Loss of Function

- Mutation
I. Human Genome Segmental Duplication Pattern

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

http://humanparalogy.gs.washington.edu
Mouse Segmental Duplication Pattern

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

She, X et al., (2008) Nature Genetics
Human Segmental Duplications Properties

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

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

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

Bailey et al. (2002), Science
Developmental Delay

<table>
<thead>
<tr>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of Population</td>
<td></td>
</tr>
<tr>
<td>~14.2% of genetic cause of developmental delay explained by large CNVs (&gt;500 kbp)</td>
<td></td>
</tr>
</tbody>
</table>

Genome Wide CNV Burden
(15,767 cases of ID, DD, MCA vs. 8,328 controls)

Cooper et al., Nat. Genet, 2011
Common and Rare Structural Variation are Linked  
17q21.31 Deletion Syndrome
17q21.31 Inversion

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

Stefansson, K et al., (2005) Nature Genetics
• 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 duplcon
- 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>Diagram</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1.2</td>
<td><img src="image1.png" alt="Diagram H1.2" /></td>
<td>1.29Mb</td>
</tr>
<tr>
<td>H1.1</td>
<td><img src="image2.png" alt="Diagram H1.1" /></td>
<td>1.08Mb</td>
</tr>
<tr>
<td>H1.3</td>
<td><img src="image3.png" alt="Diagram H1.3" /></td>
<td>1.5Mb</td>
</tr>
<tr>
<td>H1D</td>
<td><img src="image4.png" alt="Diagram H1D" /></td>
<td>1.29Mb</td>
</tr>
<tr>
<td>H1D.3</td>
<td><img src="image5.png" alt="Diagram H1D.3" /></td>
<td>1.49Mb</td>
</tr>
<tr>
<td>H2.1</td>
<td><img src="image6.png" alt="Diagram H2.1" /></td>
<td>1.19Mb</td>
</tr>
<tr>
<td>H2.2</td>
<td><img src="image7.png" alt="Diagram H2.2" /></td>
<td>1.41Mb</td>
</tr>
<tr>
<td>H2D</td>
<td><img src="image8.png" alt="Diagram H2D" /></td>
<td>1.8Mb</td>
</tr>
</tbody>
</table>

San
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

Sequencing-based

• Read-depth: Bailey et al, 2002
• Fosmid ESP: Tuzun et al. 2005, Kidd et al. 2008
• Sanger sequencing: Mills et al., 2006
• 3rd generation --long-reads: Chaisson et al., 2015

Single molecule mapping

■ 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)

LogR and B-A allele Frequency

Human chromosome 3 position

~55 kbp
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

a) Insertion
b) Deletion

c) Inversion

dishconcordant by orientation (yellow/gold)
dishconcordant size (red)
duplication track

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%)
- **Read-Depth**: 3223 (80%)
- **Split-read**: 1772 (33%)

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 (e.g. mrFAST, mrsFAST)

Reference Sequence
Sequence to Test

unique
duplicated

Illumina Sequence

Bailey et al., Science, 2002
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
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
• 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
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

NBPF14

NBPF7

# individuals

Paralog-specific copy number

African
Asian
European
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

Length distribution

P6C4 chemistry—30-40 kbp libraries
6 hr movie
Mean 10.8 kbp read
Max 47.6 kbp
PacBio Sequence Reads are Uniform

![Box plot showing mean coverage per sample per window against GC content for different technologies. CHM1 PacBio and Illumina technologies are distinguished by color.](image)
Algorithms: HGAP and QUIVER

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

Chin et al. Nat. Methods, 2013

https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP
PacBio Whole Genome Sequencing

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

http://datasets.pacb.com/2013/Human10x/READS/index.html
SMRT-SV
Structural Variation Detection using PacBio

BLASR alignment of reads

Signatures of structural variants
contained deletion  contained insertion  single hard-stop  double hard-stop  inverted hard-stop

Celera assembly

Remap reads, generate Quiver consensus

Map consensus, structural variant resolution

Increased Resolution of Structural Variation

92% of insertions and 60% deletions (50 - 5,000 bp) are novel novel genetic variants corresponding to 11 Mbp of sequence 22,112 novel genetic variants corresponding to 11 Mbp of sequence 6,796 of the events map within 3,418 genes 6,796 of the events map within 3,418 genes 169 within coding sequence or UTRs of genes 169 within coding sequence or UTRs of genes
Phased-SV: Comprehensive SV Detection of a Diploid

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

Chaisson et al. Biorxiv, 2017
Falcon SMRT Genome Assembly

- two phases: long reads are corrected and overlapped to generate a string graph—third phase “repeat unitig bridging”
- By Jason Chin [http://github.com/PacificBiosciences/FALCON](http://github.com/PacificBiosciences/FALCON)
CHM1 Human Genome Assembly

- 67 X sequence coverage — Contig N50 27.9 Mbp
- 3,777 Contigs

Chin et al., unpublished
Future: *De novo* Human Genome Assembly with SMRT WGS

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

- 71.7 X sequence coverage
- average contig N50 = 9.6 Mbp
- assembly size 3.1 Gbp

- 16,073 contigs
- 911 >= 100 kbp

12.3 kbp tandem repeat

- 180 Mbp corresponding to 92% of euchromatic gaps in gorGor3 were closed. (399,243/433,861 closed gaps)
• Recovered 10,779 of 12,757 (84.5%) exons mapping within the gap regions (based on Human RefSeq models)
• Estimate 3,269/3,697 (88%) of gorilla models resolved
• 8-9% of additional gorilla RNA-seq data maps to GSMRT3
Ape-Human Structural Variation Resolution

86% (101,109) of gorilla structural variants not previously reported—new insights into the evolution of our species
Summary

• Approaches
  – Multiple methods need to be employed—Readpair+Read-depth+SplitRead and an experimental method
  – Tradeoff between sensitivity and specificity
  – Complexity not fully understood

• Read-pair and read-depth NGS approaches
  – narrow the size spectrum of structural variation
  – lead to more accurate prediction of copy-number
  – unparalleled specificity in genotyping duplicated genes
    (reference genome quality key)

• Third generation sequencing methods hold promise but require high coverage—still expensive. *Sequel?*
• Ohno—Duplication is the primary force by which new gene functions are created
• There are 990 annotated genes completely contained within segmental duplications

III. Why?
Rate of Duplication

Sudmant PH et al., Genome Res. 2013
Mosaic Architecture

- 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.

Jiang et al, Nat. Genet., 2007
Human Great-ape “Core Duplicons” have led to the Emergence of New Genes

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

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

Marques-Bonet and Eichler, CSHL Quant Biol, 2008
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

In situ
SRGAP2C duplicate antagonizes function

Charrier et al., *Cell*, 2012
Australopithecus

Sahelanthropus

Orrorin

Ardipithecus

K. platyops

A. anamensis

A. afarensis

A. aethiopicus

A. boisei

A. robustus

Homo

- 3.4 mya
- 2.4 mya

~350 cc

~1000 cc

Dennis, Nuttle et al. Cell (2012)
Example 2: Human-specific Duplication of *ARHGAP11B*

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

*Florea et al., Science 2015, Antonacci et al., Nat. Genet., 2014*
**ARHGAP11B** induced gyrification of mouse brain

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

*Florea et al., Science 2015*
Duplication of $ARHGAP11B$ and $15q13.3$ Syndrome

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

Antonacci et al., Nat Genet, 2014,
Summary

• Interspersed duplication architecture sensitized our genome to copy-number variation increasing our species predisposition to disease—children with autism and intellectual disability.

• Duplication architecture has evolved recently in a punctuated fashion around core duplicons which encode human great-ape specific gene innovations (eg. NPIP, NBPF, LRRC37, etc.).

• Cores have propagated in a stepwise fashion “transducing” flanking sequences---human-specific acquisitions flanks are associated with brain developmental genes.

• Core Duplicon Hypothesis: Selective disadvantage of these interspersed duplications offset by newly minted genes and new locations within our species. Eg. SRGAP2C
Overall Summary

- **I. Disease**: Role of CNVs in human disease—relationship of common and rare variants—a genomic bias in location and gene type

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

- **III: Evolution**: Rapid evolution of complex human architecture that predisposes to disease coupled to gene innovation
Disease

Evolution
Eichler Lab

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

SV-structural variation
CNV-copy number variation
CNP—copy number polymorphism
Indel-insertion/deletion event
SD—segmental duplication
SUN-singly-unique nucleotide identifier
SMRT-single-molecule real-time sequencing
WGS—whole genome shotgun sequencing
SV Software

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

Interchromosomal (a) Direct
(a) Inverted
(c) Complex

Intrachromosomal (d)
(e)
(f)

Intrachromatid (g)