Copy Number Variants:

detection and analysis

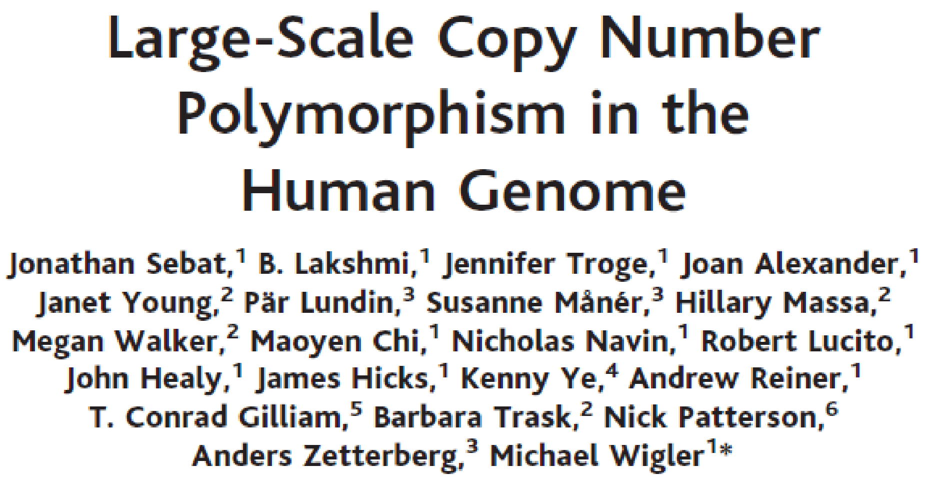
Manuel Ferreira & Shaun Purcell

Boulder, 2009

Large chromosomal rearrangements can cause sporadic disease

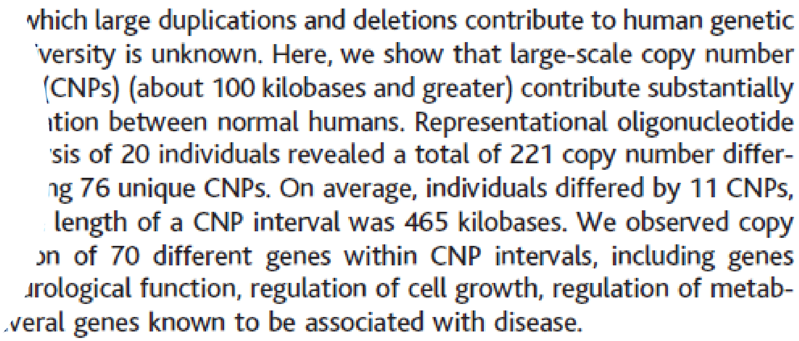
Down syndrome
Duchenne Muscular Dystrophy (DMD)
DiGeorge-Velo cardiofacial syndrome (VCFS)
...

Lupski 2007 Nat Genet 39: s43



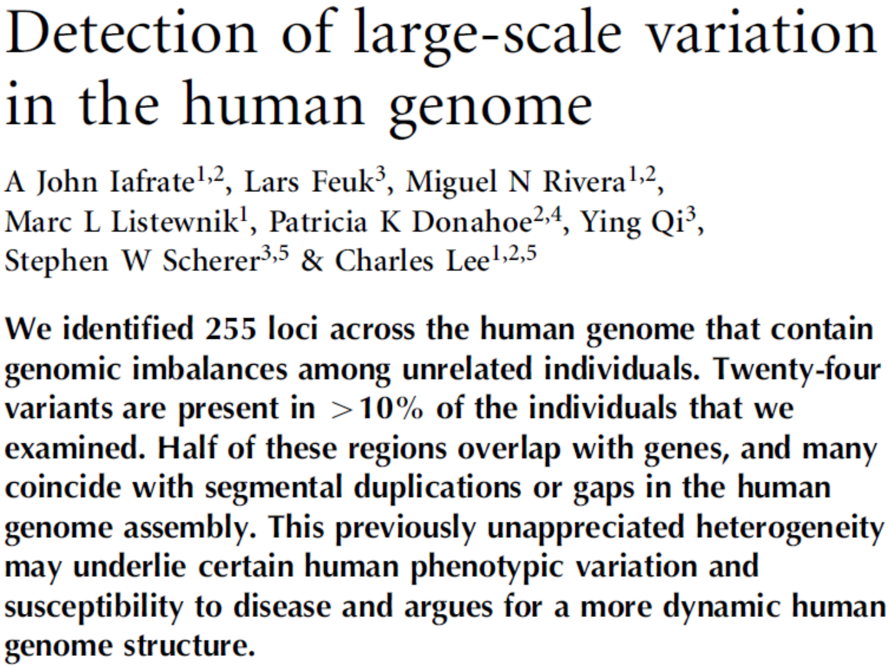
Large-Scale Copy Number Polymorphism in the Human Genome

Jonathan Sebat,[1] B. Lakshmi,[1] Jennifer Troge,[1] Joan Alexander,[1]
Janet Young,[2] Pär Lundin,[3] Susanne Månér,[3] Hillary Massa,[2]
Megan Walker,[2] Maoyen Chi,[1] Nicholas Navin,[1] Robert Lucito,[1]
John Healy,[1] James Hicks,[1] Kenny Ye,[4] Andrew Reiner,[1]
T. Conrad Gilliam,[5] Barbara Trask,[2] Nick Patterson,[6]
Anders Zetterberg,[3] Michael Wigler[1]*



which large duplications and deletions contribute to human genetic
versity is unknown. Here, we show that large-scale copy number
(CNPs) (about 100 kilobases and greater) contribute substantially
tion between normal humans. Representational oligonucleotide
sis of 20 individuals revealed a total of 221 copy number differ-
ng 76 unique CNPs. On average, individuals differed by 11 CNPs,
length of a CNP interval was 465 kilobases. We observed copy
on of 70 different genes within CNP intervals, including genes
urological function, regulation of cell growth, regulation of metab-
veral genes known to be associated with disease.

Sebat et al 2004 Science 305: 525



Detection of large-scale variation in the human genome

A John Iafrate[1,2], Lars Feuk[3], Miguel N Rivera[1,2],
Marc L Listewnik[1], Patricia K Donahoe[2,4], Ying Qi[3],
Stephen W Scherer[3,5] & Charles Lee[1,2,5]

We identified 255 loci across the human genome that contain genomic imbalances among unrelated individuals. Twenty-four variants are present in >10% of the individuals that we examined. Half of these regions overlap with genes, and many coincide with segmental duplications or gaps in the human genome assembly. This previously unappreciated heterogeneity may underlie certain human phenotypic variation and susceptibility to disease and argues for a more dynamic human genome structure.

Iafrate et al 2004 Nat Genet 36: 949

Outline

1. What is a Copy Number Variant (CNV)

2. Genome-wide detection of CNVs

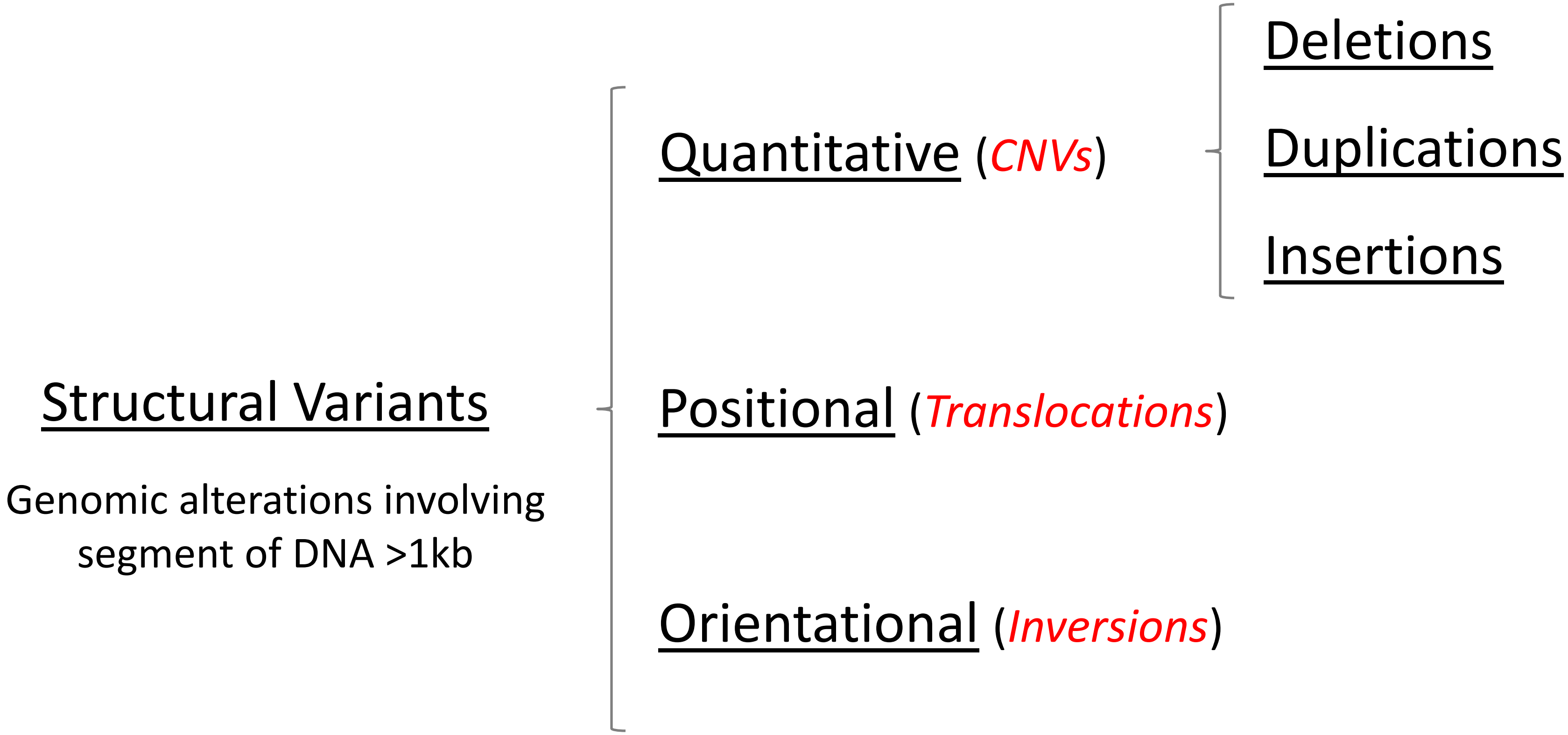
3. Association analysis of CNVs

4. Online databases

1. What is a CNV?

What is a CNV?

1. Classes of structural variants

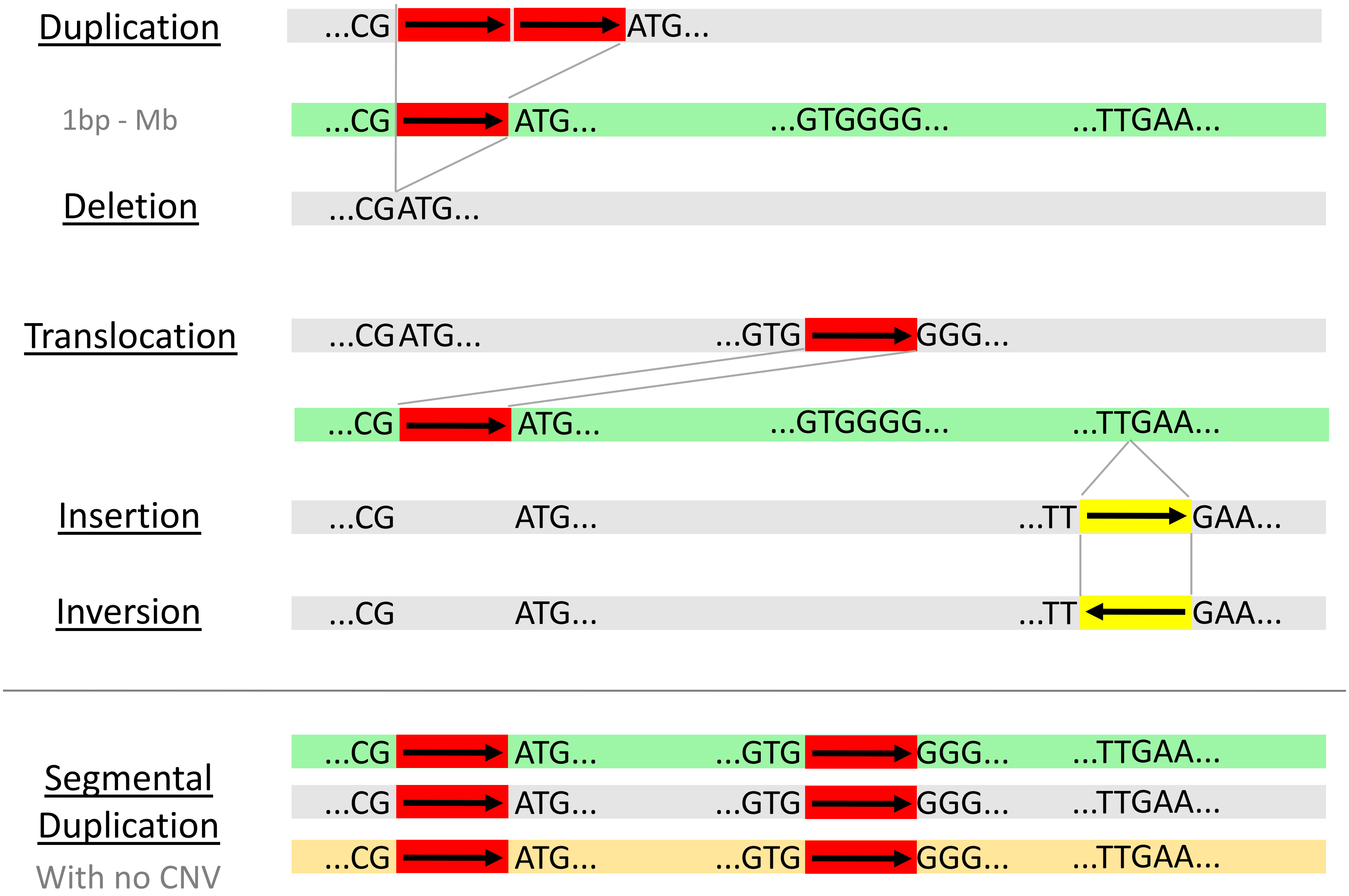


Structural Variants

Genomic alterations involving segment of DNA >1kb

- Quantitative (*CNVs*)
  - Deletions
  - Duplications
  - Insertions
- Positional (*Translocations*)
- Orientational (*Inversions*)

Copy Number Polymorphism (CNP) is a *CNV* that occurs in >1% population



Duplication

...CG ATG...

1bp - Mb

...CG ATG... ...GTGGGG... ...TTGAA...

Deletion

...CGATG...

Translocation

...CGATG... ...GTG GGG...

...CG ATG... ...GTGGGG... ...TTGAA...

Insertion

...CG ATG... ...TT GAA...

Inversion

...CG ATG... ...TT GAA...

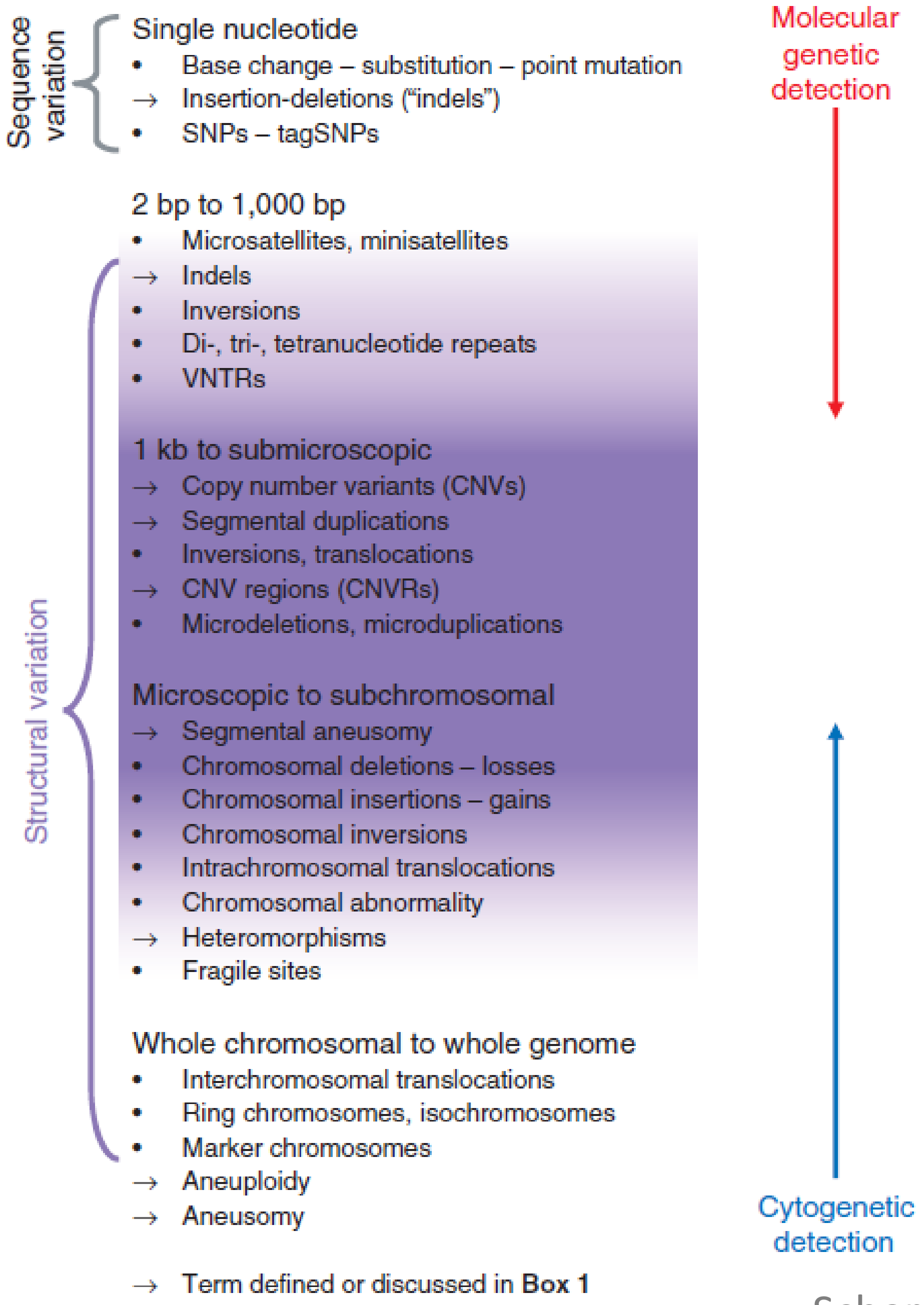
Segmental Duplication

With no CNV

...CG ATG... ...GTG GGG... ...TTGAA...

...CG ATG... ...GTG GGG... ...TTGAA...

...CG ATG... ...GTG GGG... ...TTGAA...



Single nucleotide

- Base change – substitution – point mutation

→ Insertion-deletions (“indels”)

- SNPs – tagSNPs

2 bp to 1,000 bp

- Microsatellites, minisatellites

→ Indels

- Inversions
- Di-, tri-, tetranucleotide repeats
- VNTRs

1 kb to submicroscopic

→ Copy number variants (CNVs)

→ Segmental duplications

- Inversions, translocations

→ CNV regions (CNVRs)

- Microdeletions, microduplications

Microscopic to subchromosomal

→ Segmental aneusomy

- Chromosomal deletions – losses
- Chromosomal insertions – gains
- Chromosomal inversions
- Intrachromosomal translocations
- Chromosomal abnormality

→ Heteromorphisms

- Fragile sites

Whole chromosomal to whole genome

- Interchromosomal translocations
- Ring chromosomes, isochromosomes
- Marker chromosomes

→ Aneuploidy

→ Aneusomy

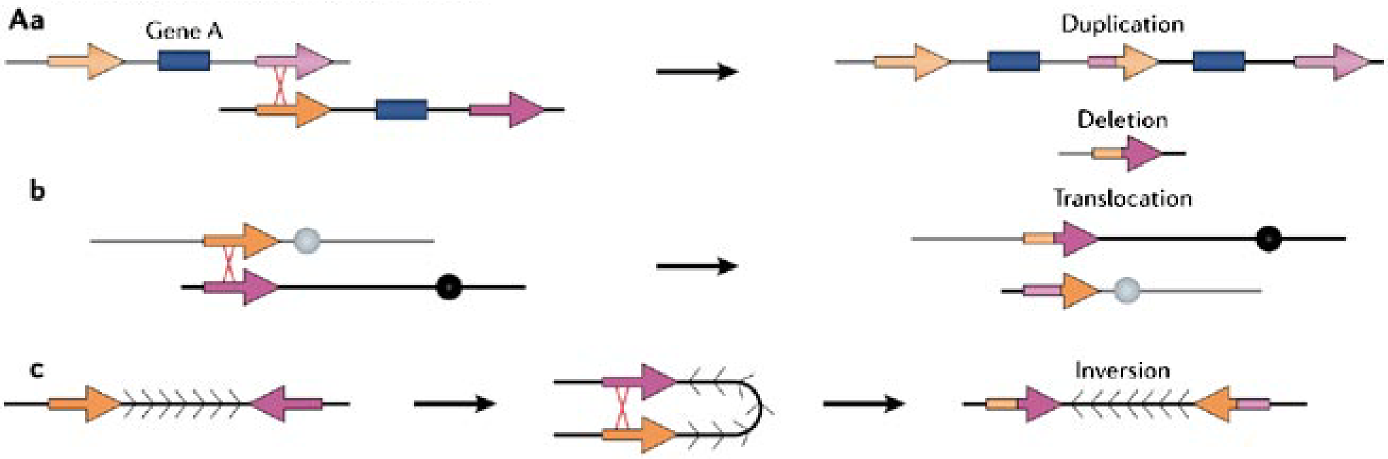
→ Term defined or discussed in **Box 1**

Scherer 2007 Nat Genet 39: s7

What is a CNV?

2. Origins of CNVs

(A) Non-allelic homologous recombination



(B) Non-homologous end joining

(C) Tandem repeat sequences

(D) Retrotransposons

Bailey & Eichler 2006 *Nat Rev Genet* **7:** 552

What is a CNV?

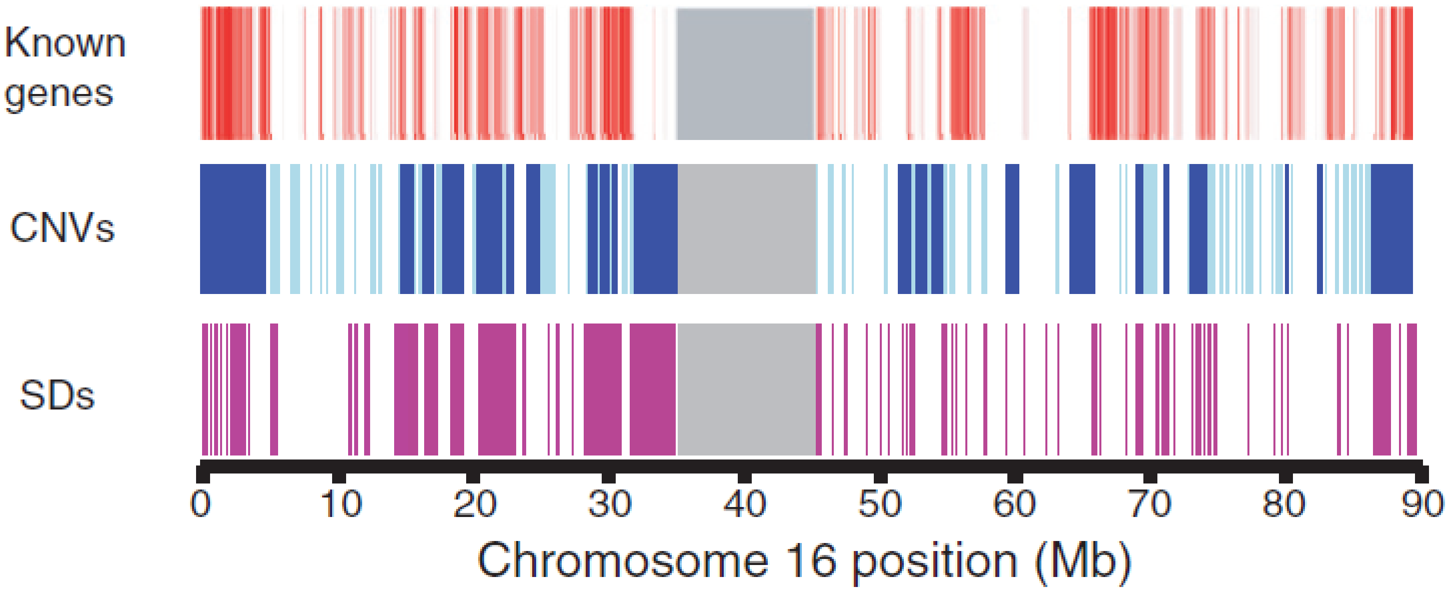
3. CNVs are abundant in the genome

Human vs Human	SNPs	CNVs
Base pairs	2.5 Mb	4 Mb
	1/1,200 bp	1/800
% genome	0.08%	0.12%

Sebat 2007 Nat Genet 39: s3

What is a CNV?

4. CNVs significantly overlap with known genes



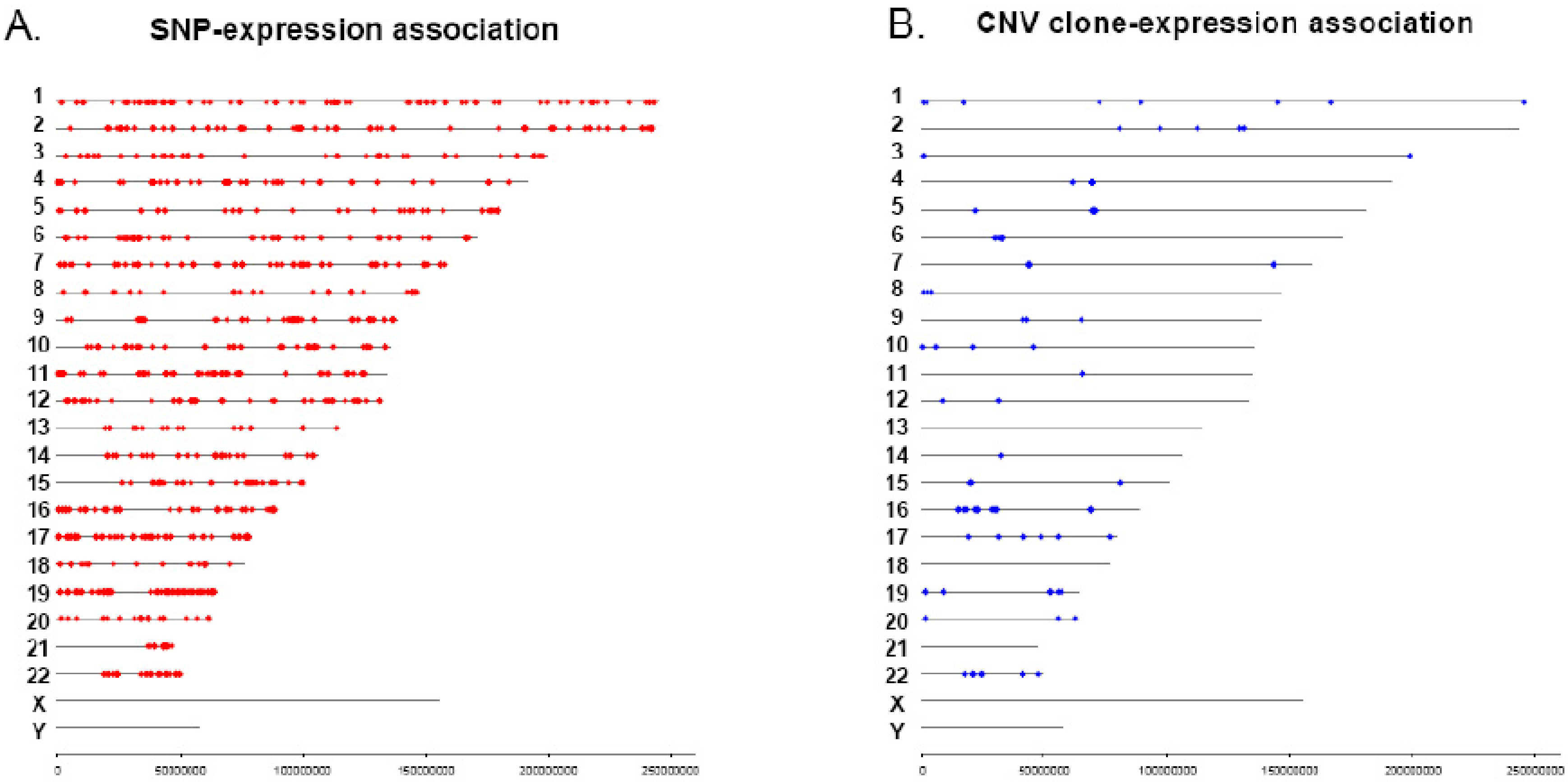
Cooper et al 2007 Nat Genet 39: s22

What is a CNV?

5. CNVs influence gene expression

83.6%

17.7%

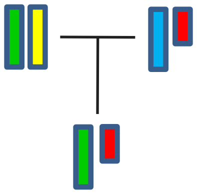


Stranger et al 2007 Science 315: 848

What is a CNV?

6. In healthy individuals, most CNVs are inherited…

>99% inherited

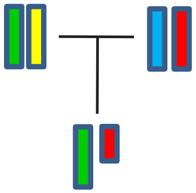


Rare CNVs 10%

Common CNVs 90%

>1% population

<1% de novo



McCarroll 2008 Hum Mol Genet 17: R135

McCarroll et al 2008 Nat Genet 40: 1166

2. Detection of CNVs

Detection of CNVs

A. Using intensity data from whole-genome arrays

SNPs → *Genotype known common variants*

CNVs

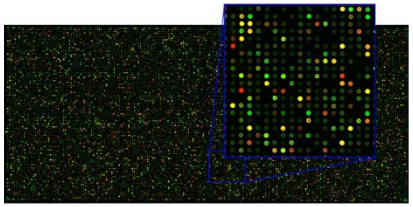
(A) *Genotype known common variants*

(B) *Identify and genotype new, potentially rarer variants*

Detection of CNVs

(A) Genotype known common CNVs using whole-genome arrays

Nimblegen



array-CGH, CNV only, test vs reference

custom or whole-genome (up to 2,1M probes)

Affy 6.0

>940,000 CNV non-polymorphic probes

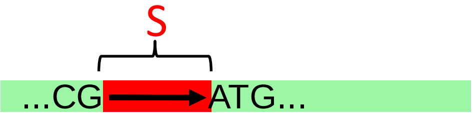
High-density in ~5,600 CNV regions in DGV + extended to whole-genome

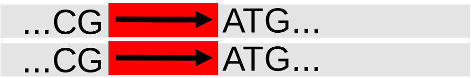
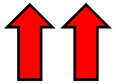
Illumina 1M

36,000 CNV non-polymorphic probes

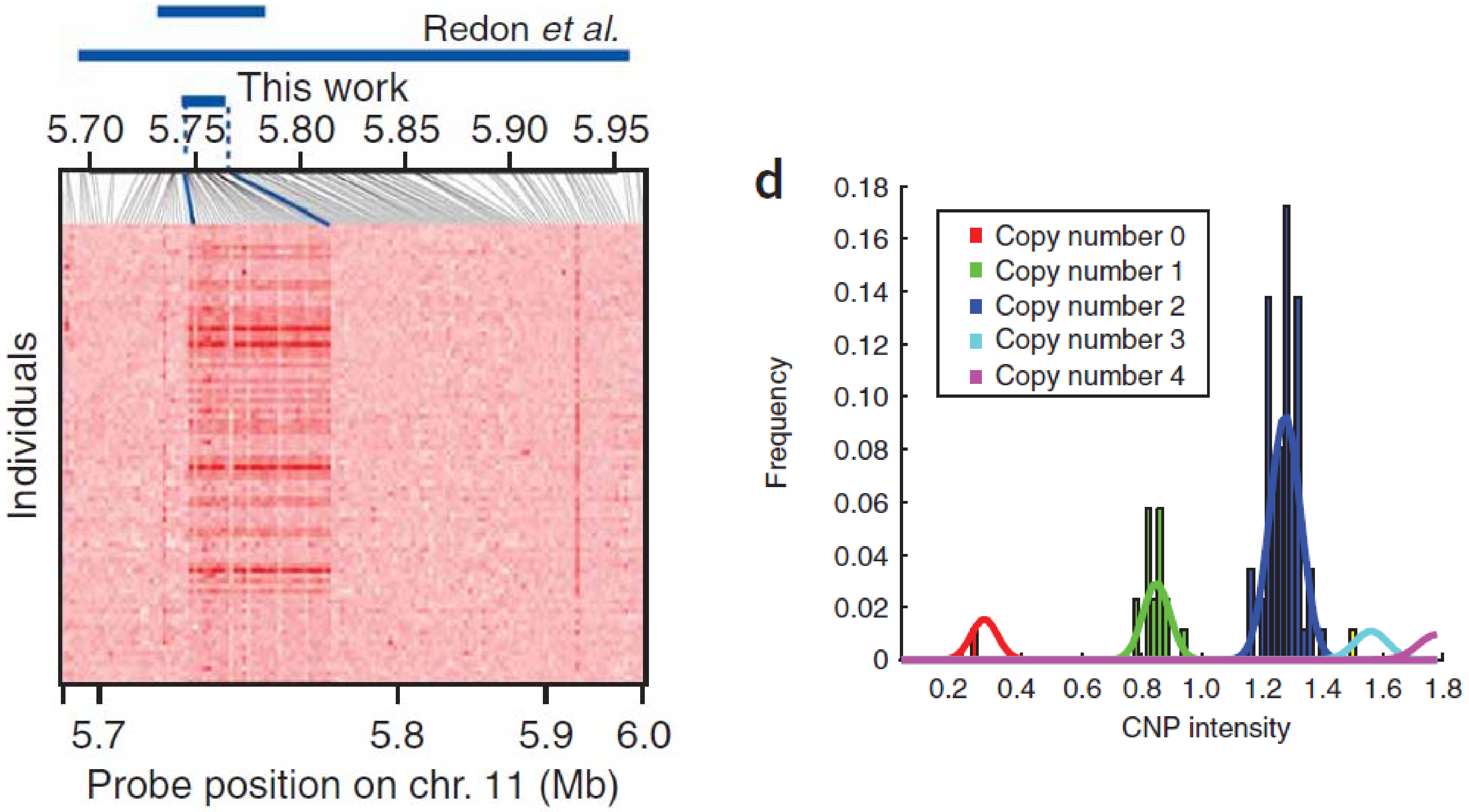
covering ~4,000 CNV regions in DGV

Detection of CNVs



Ind	Genotype Mat/Pat	Copy number at S	Amount of DNA at S
1	S/S	2 	

Detection of CNVs



Non-polymorphic probes

McCarroll et al 2008 Nat Genet 40: 1166

Detection of CNVs

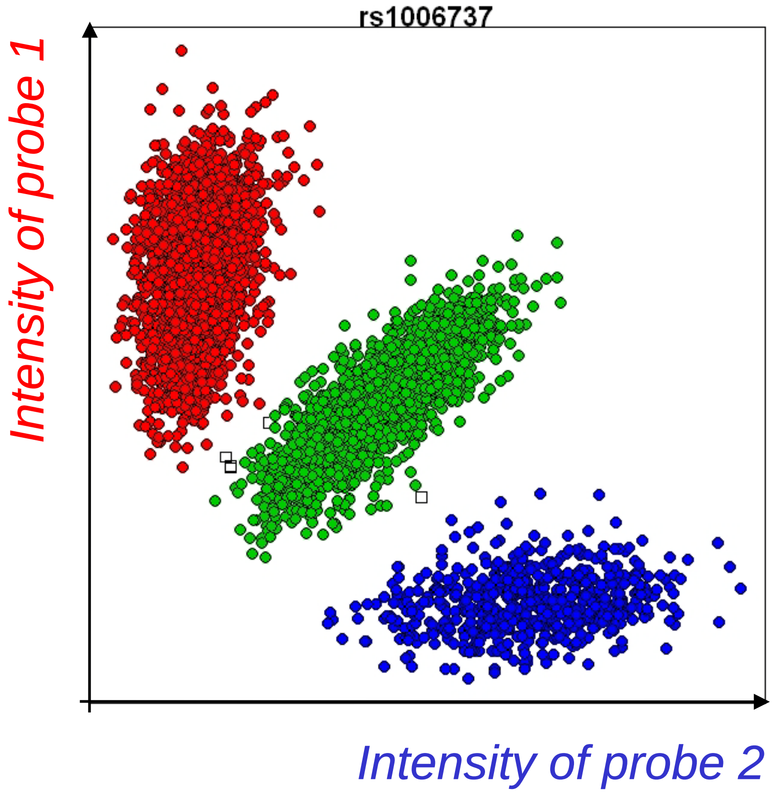
(B) Identify and genotype new, potentially rarer CNVs from whole-genome array data (CGH, Affymetrix/Illumina)

Example: rs1006737 **A**/**G**

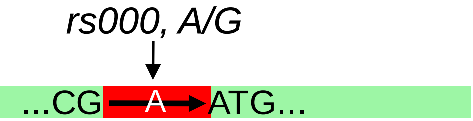
AA
AG
GG

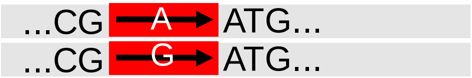
... AGCCCGAA<u>A</u>TGTTTTCAGA... *probe 1*

... AGCCCGAA<u>G</u>TGTTTTCAGA... *probe 2*

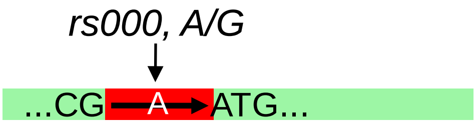


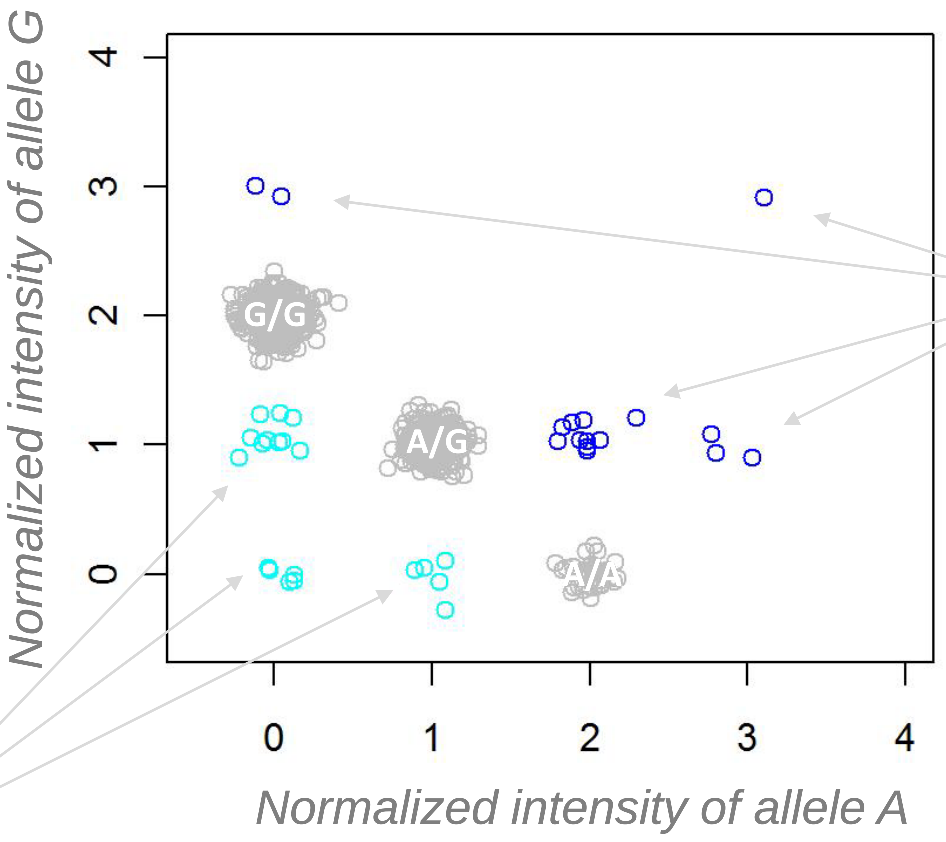
Detection of CNVs



Ind	Genotype Mat/Pat	Pattern	Copy number for:		
			A	G	Total
1	A/G		1	1	2

Detection of CNVs





Individuals with duplication(s)

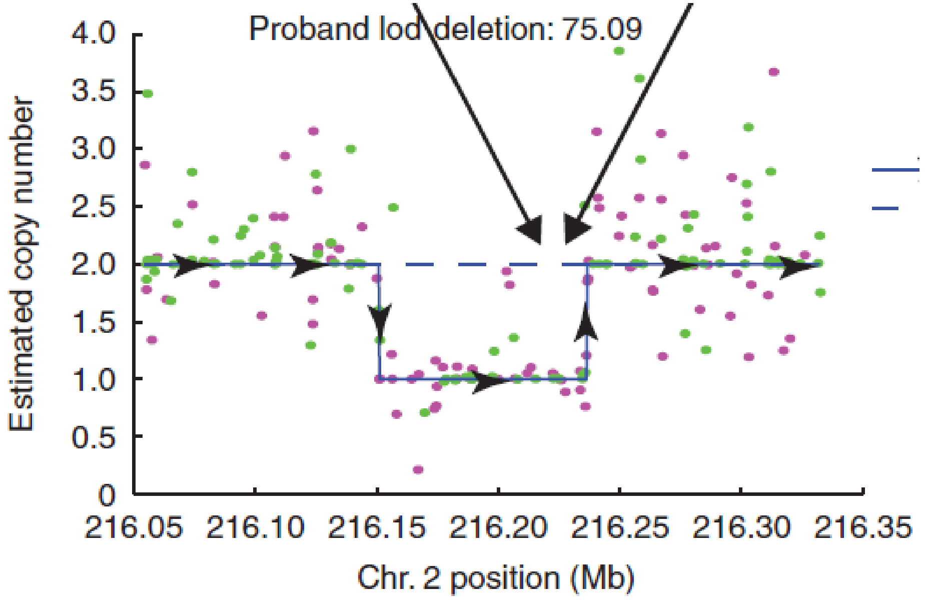
ie. total CN > 2

Individuals with deletion(s)

ie. total CN < 2

Polymorphic probe in CNV region

Detection of CNVs



Birdseye

Affy 5.0, 6.0

Korn et al 2008 Nat Genet 40: 1253

PennCNV

Affymetrix and Illumina

Wang et al 2007 Genome Res 17: 1665

Combine information across probes to identify new CNVs

For example...	Cases	Controls
100kb deletion chr. 2	10/5,000	1/5,000

Detecting CNVs through GWAS arrays is challenging…

Lots of confounders: DNA quality, concentration, source, batch effects, date effects.

Arrays have poor resolution for CNVs (>100kb).

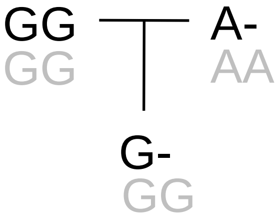
Genotype calling is computationally demanding, as it requires analysis of very large ‘raw’ cell files.

Genotype calling software often platform specific, not very user friendly.

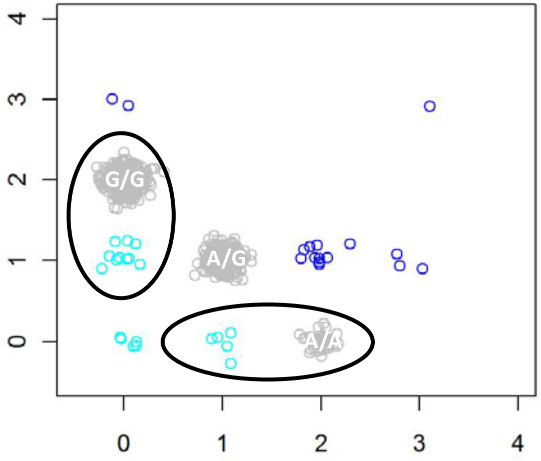
Detection of CNVs

B. Identifying CNVs through genotyping errors

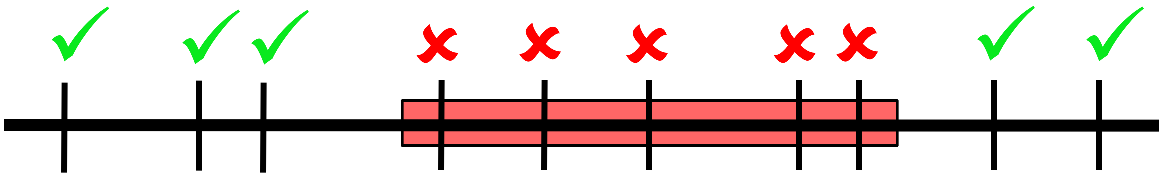
- Mendelian Inconsistencies



GG
GG
A-
AA
G-
GG



- Failure Hardy-Weinberg equilibrium

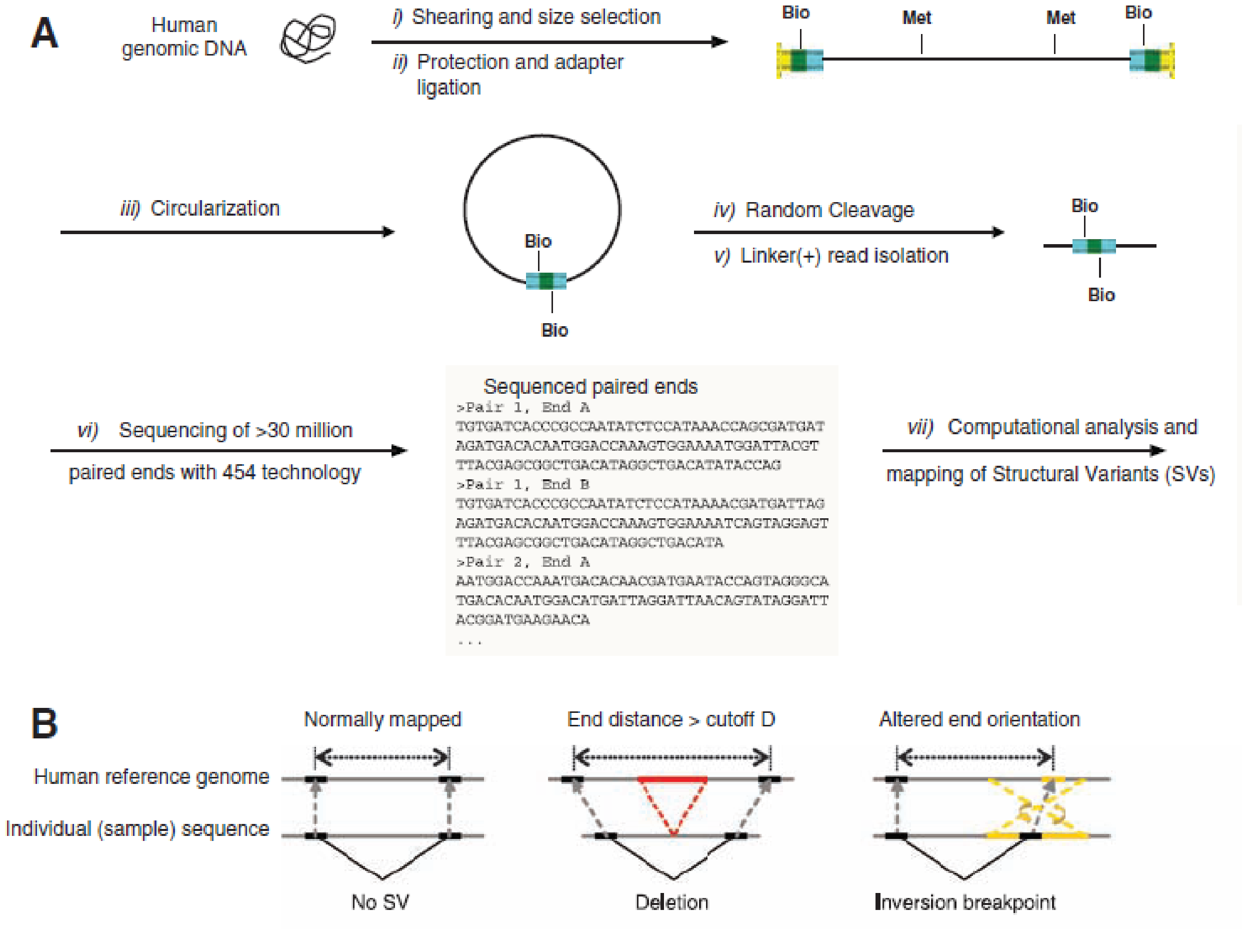


Conrad et al 2006 Nat Genet 38: 75

McCarroll et al 2006 Nat Genet 38: 86

Detection of CNVs

C. Targeted or whole-genome sequencing


A
Human genomic DNA
i) Shearing and size selection
ii) Protection and adapter ligation
Bio
Met
Met
Bio
iii) Circularization
Bio
Bio
iv) Random Cleavage
v) Linker(+) read isolation
Bio
Bio
vi) Sequencing of >30 million paired ends with 454 technology
Sequenced paired ends
vii) Computational analysis and mapping of Structural Variants (SVs)
B
Normally mapped
End distance > cutoff D
Altered end orientation
Human reference genome
Individual (sample) sequence
No SV
Deletion
Inversion breakpoint

```
>Pair 1, End A
TGTGATCACCCGCCAATATCTCCATAAACCAGCGATGAT
AGATGACACAATGGACCAAAGTGGAAAATGGATTACGT
TTACGAGCGGCTGACATAGGCTGACATATACCAG
>Pair 1, End B
TGTGATCACCCGCCAATATCTCCATAAAACGATGATTAG
AGATGACACAATGGACCAAAGTGGAAAATCAGTAGGAGT
TTACGAGCGGCTGACATAGGCTGACATA
>Pair 2, End A
AATGGACCAAATGACACAACGATGAATACCAGTAGGGCA
TGACACAATGGACATGATTAGGATTAACAGTATAGGATT
ACGGATGAAGAACA
...
```

Korbel et al 2007 Science 318: 420

Summary so far…

CNVs are abundant, often overlap genes, can influence gene expression and most are inherited in healthy individuals

Known and new CNVs can be identified and genotyped in large-scale studies using whole-genome genotyping arrays, such as the 6.0 and 1M. Low resolution (>100Kb) & low signal/noise ratio.

More accurate CNV genotyping maps/arrays/algorithms expected in the next few years.

What are the particular strategies and challenges for association analysis of CNVs?

3. Association analysis of CNVs

Association analysis of CNVs

1. Some of the relevant questions

(A) Are CNPs associated with variation in human traits or diseases?

(B) Can we identify rare CNVs associated with large increase in disease risk? Are these de novo or inherited in cases?

(C) When considering the whole-genome, do cases have more CNV events then controls, ie. increased burden?

(D) How to test SNPs in copy number regions?

(E) Are most CNVs tagged by SNPs in genotyping arrays?

Example 1: Autism whole-genome CNV analysis

Sample	16p11	Cases	Controls	*P*
Discovery	Del (600kb)	5/1,441	3/4,234	1.1 x 10^{-4}
[Affy 500K]	Dup	7/1,441	2/4,234	
Replication 1 (CHB)	Del	5/512	0/434	0.007
[array-CGH]	Dup	4/512	0/434	
Replication 2 (deCODE)	Del	3/299	2/18,834	4.2 x 10^{-4}
[Illumina]	Dup	0/299	5/18,834	

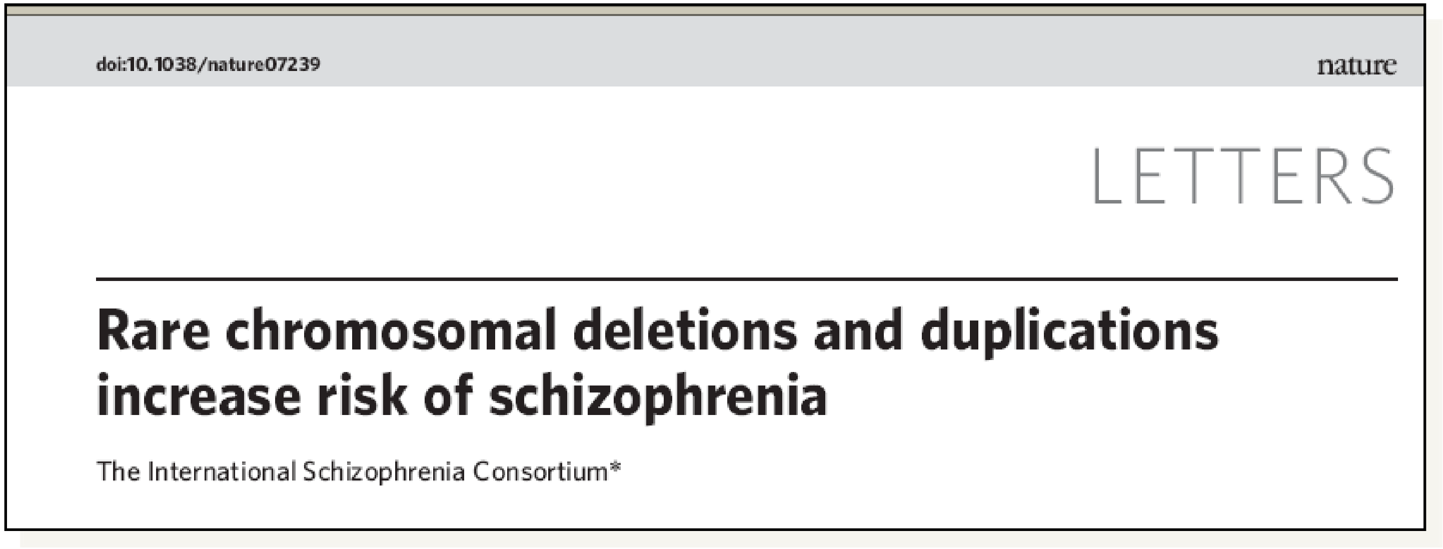
COPPER
Birdseye
CNAT

	Deletion frequency Iceland
Autism	1%
Psychiatric disorder	0.1%
General population	0.01%

	del	dup
inherited	2	6
de novo	10	1
unknown	1	4

Weiss et al. N Engl J Med 2008; 358: 667

Example 2: SCZ whole-genome CNV analysis



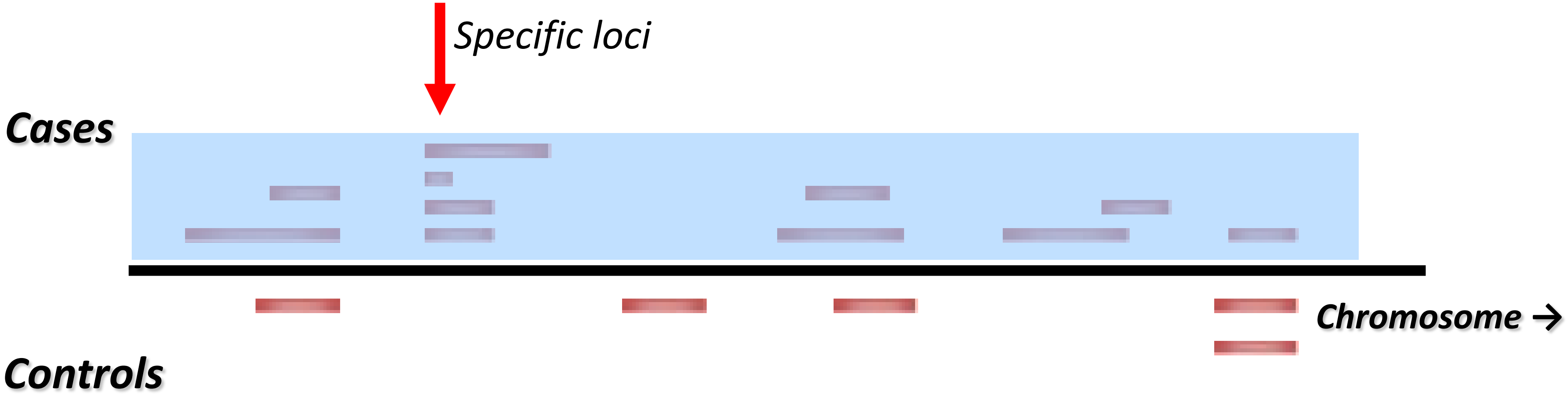
doi:10.1038/nature07239

nature

LETTERS

Rare chromosomal deletions and duplications increase risk of schizophrenia

The International Schizophrenia Consortium*



Specific loci

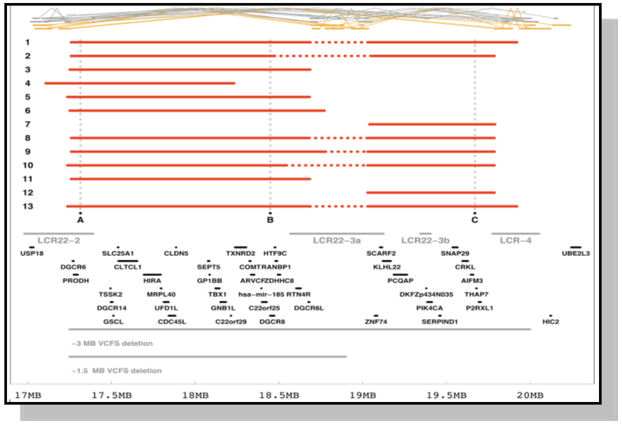
Cases

Chromosome →

Controls

Specific large (>500kb) rare deletions

22q11.2 (VCFS)



11 : 0

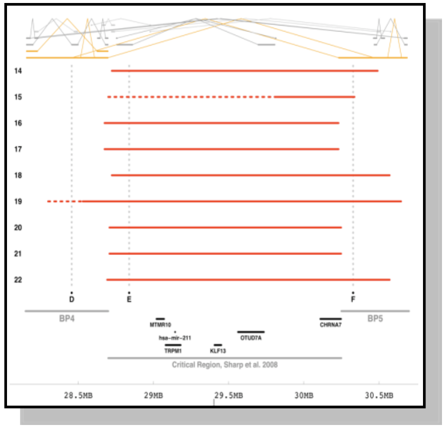
A "positive control"

1:4000 live births
~30% develop psychosis
In ~0.6-2% SCZ patients

3Mb and 1.5Mb variants

2 additional atypical deletions observed

15q13.3



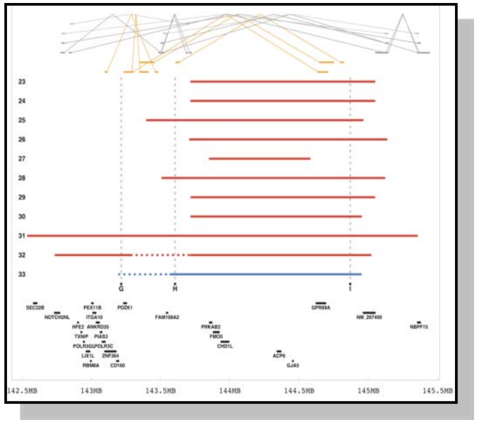
9 : 0

CHRNA7, alpha 7 nicotinic acetylcholine receptor

5 cases w/ impaired cognition; 1 w/ epilepsy

Previously seen in mental retardation with seizures

1q21.1



10 : 1

3 cases had cognitive abnormalities; 1 with epilepsy

Also seen in a patient with MR and seizures and two patients with autism.

Genome-wide burden of rare CNVs in SCZ

3,391 patients with SCZ, 3,181 controls

Filter for <1% MAF, >100kb

6,753 CNVs

Cases have greater rate of CNVs than controls
1.15-fold increase
P **= 3 × 10^{-5}**

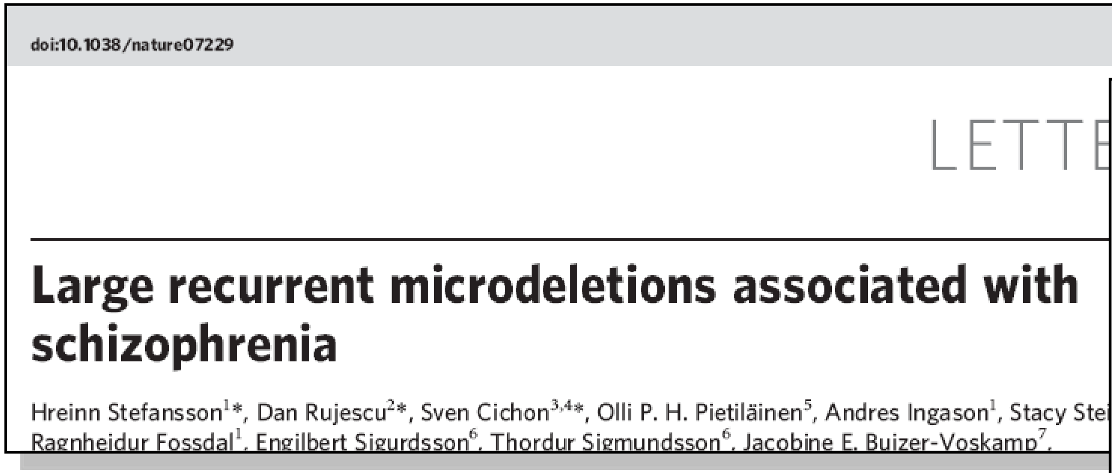
Cases have more genes intersected by CNVs than controls
1.14-fold increase
= 2 × 10^{-6}

True for *singleton* events (observed only once in dataset)
1.45-fold increase
(~15% cases versus 11% controls)
P **= 5 × 10^{-6}**

Rate of *genic* CNVs in cases v
1.18-fold increase
P **= 5 × 10^{-6}**

CNVs in cases versus controls
1.09-fold increase
P **= 0.16**

Results invariant to obvious statistical controls
Array type, genotyping plate, sample collection site, mean probe intensity



doi:10.1038/nature07229

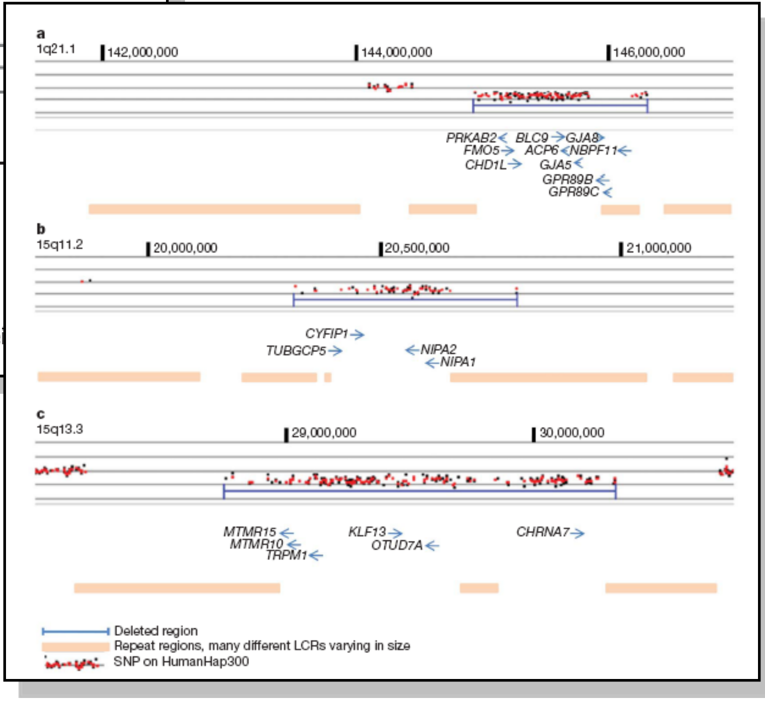
nature

LETTE

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 Stei

Ragnheidur Fossdal[1], Engilbert Sigurdsson[6], Thordur Sigmundsson[6], Jacobine E. Buizer-Voskamp[7],

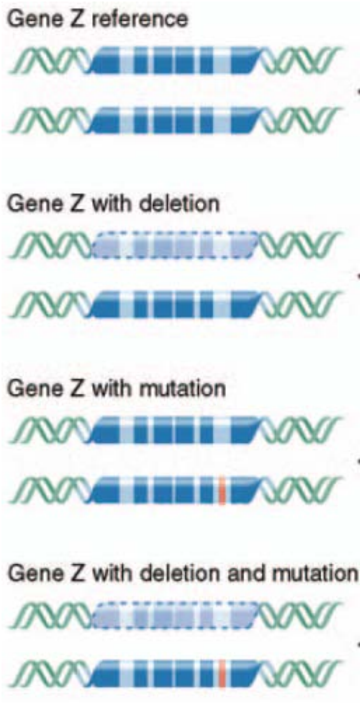


1q21.1 and 15q13.3 also identified by SGENE consortium

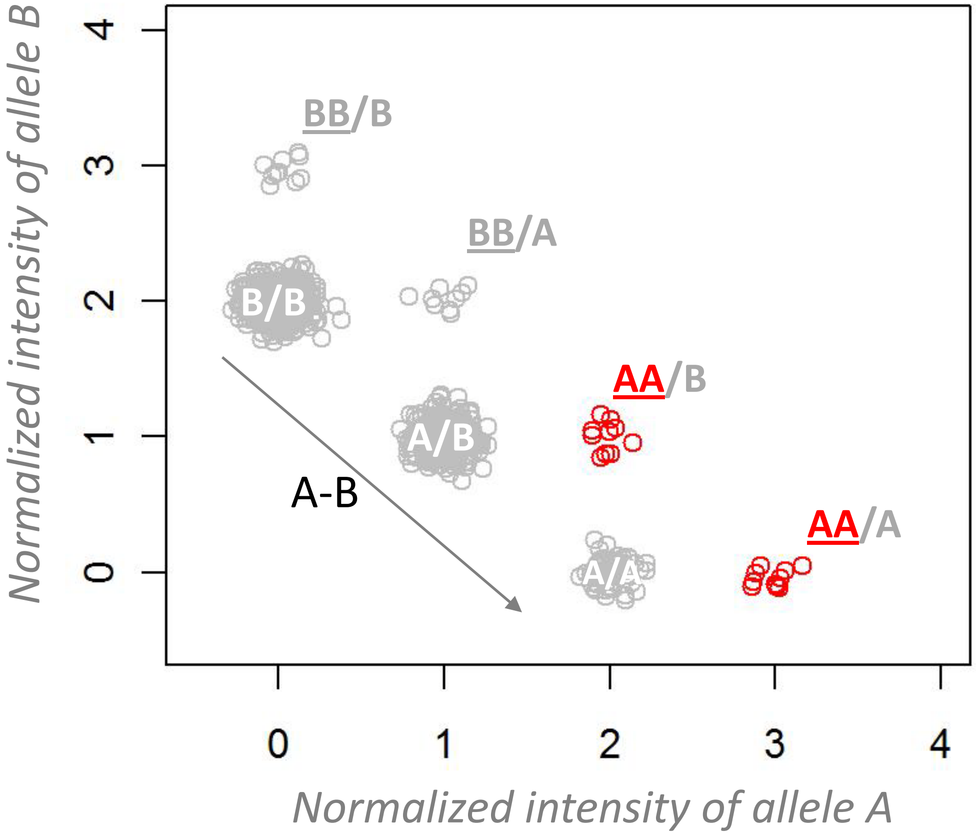
- Two other studies supporting a genome-wide increase in rare CNVs in schizophrenia
  - Walsh et al (2008) *Science*
    - 5% controls, 15% cases, 20% early onset cases
    - neurodevelopmental genes disrupted
  - Xu et al (2008) *Nature Genetics*
    - strong increased de novo rate in sporadic cases; but increased inherited rate also

Association analysis of CNVs

2. Testing SNPs in CNV regions



Gene Z reference	Healthy	Healthy
Gene Z with deletion	Disease	Healthy
Gene Z with mutation	Disease	Healthy
Gene Z with deletion and mutation	Disease	Disease
Individual analysis of SNPs or CNVs	✓	✗



$$y = \beta_1 \cdot \mathrm{SNP} + \beta_2 \cdot \mathrm{CNV}$$

$$\downarrow$$

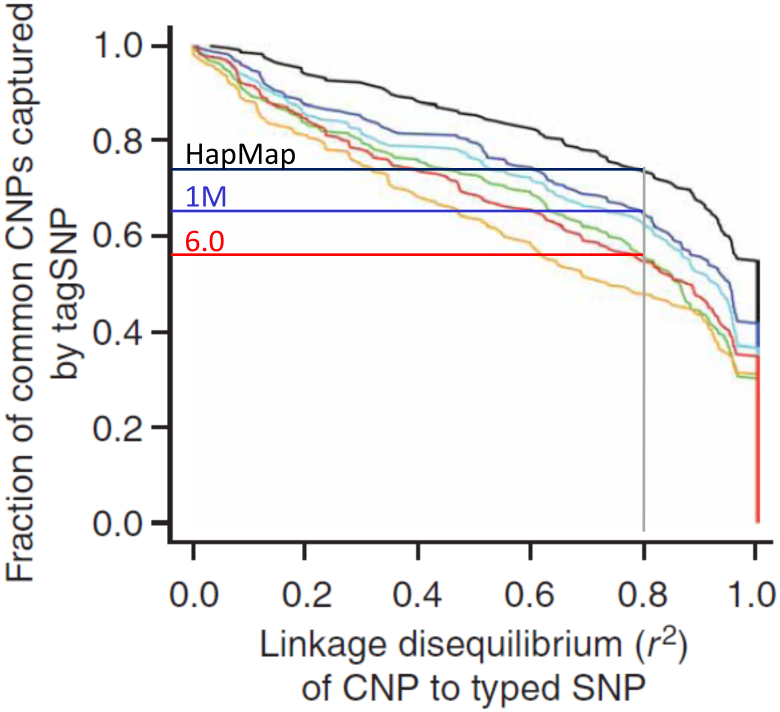
$$y = \beta_1 \cdot (\mathrm{A} - \mathrm{B}) + \beta_2 \cdot (\mathrm{A} + \mathrm{B})$$

Allele-specific risk CNV

Korn et al 2008 Nat Genet 40: 1253

Association analysis of CNVs

3. Testing CNVs through the analysis of SNPs in LD



Common CNVs

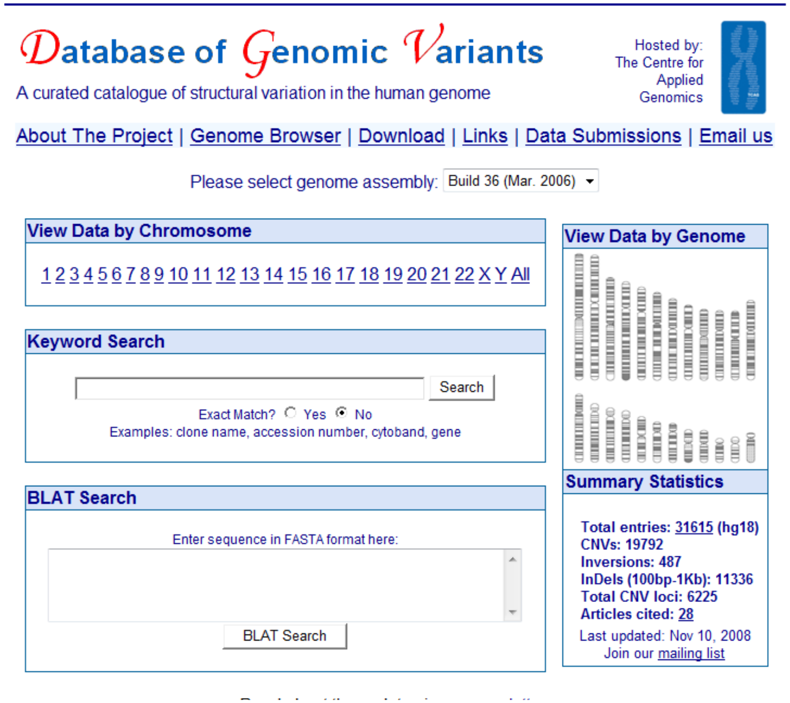
Coverage limited by lack of SNPs in CNV regions
(poor genotyping)

McCarroll et al 2008 Nat Genet 40: 1166

4. Online databases

Database of Genomic Variants

http://projects.tcag.ca/variation/



Database of Genomic Variants

A curated catalogue of structural variation in the human genome

Hosted by:
The Centre for
Applied
Genomics

About The Project | Genome Browser | Download | Links | Data Submissions | Email us

Please select genome assembly: Build 36 (Mar. 2006)

View Data by Chromosome

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y All

Keyword Search

Search

Exact Match? ○ Yes ● No

Examples: clone name, accession number, cytoband, gene

BLAT Search

Enter sequence in FASTA format here:

BLAT Search

View Data by Genome

Summary Statistics

Total entries: 31615 (hg18)
CNVs: 19792
Inversions: 487
InDels (100bp-1Kb): 11336
Total CNV loci: 6225
Articles cited: 28

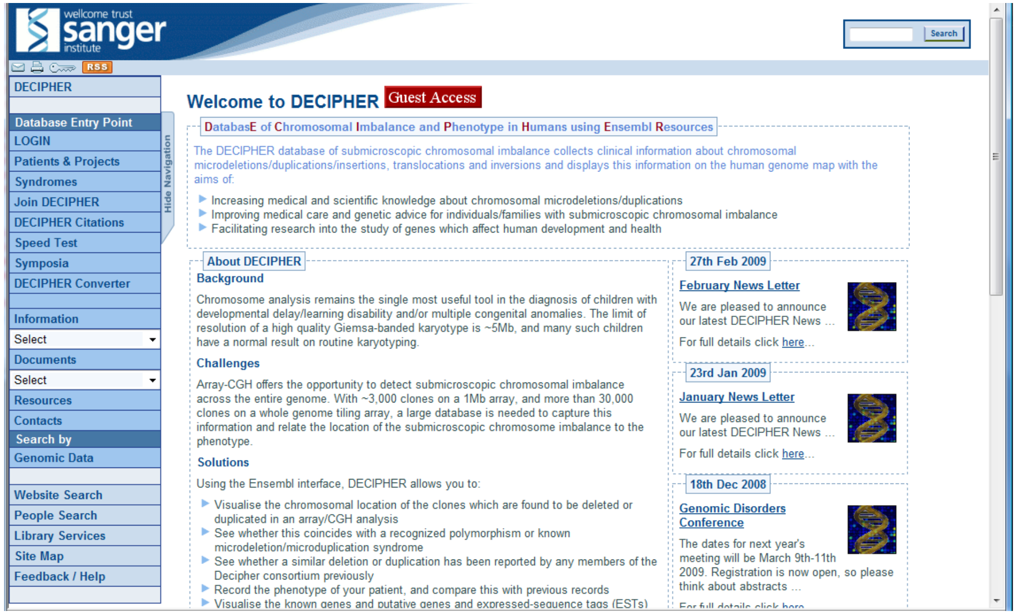
Last updated: Nov 10, 2008

Join our mailing list

Comprehensive summary of structural variation in the human genome. Healthy control samples

DECIPHER

https://decipher.sanger.ac.uk/



Welcome to DECIPHER

DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources

The DECIPHER database of submicroscopic chromosomal imbalance collects clinical information about chromosomal microdeletions/duplications/insertions, translocations and inversions and displays this information on the human genome map with the aims of:

- Increasing medical and scientific knowledge about chromosomal microdeletions/duplications
- Improving medical care and genetic advice for individuals/families with submicroscopic chromosomal imbalance
- Facilitating research into the study of genes which affect human development and health

About DECIPHER

Background

Chromosome analysis remains the single most useful tool in the diagnosis of children with developmental delay/learning disability and/or multiple congenital anomalies. The limit of resolution of a high quality Giemsa-banded karyotype is ~5Mb, and many such children have a normal result on routine karyotyping.

Challenges

Array-CGH offers the opportunity to detect submicroscopic chromosomal imbalance across the entire genome. With ~3,000 clones on a 1Mb array, and more than 30,000 clones on a whole genome tiling array, a large database is needed to capture this information and relate the location of the submicroscopic chromosome imbalance to the phenotype.

Solutions

Using the Ensembl interface, DECIPHER allows you to:

- Visualise the chromosomal location of the clones which are found to be deleted or duplicated in an array/CGH analysis
- See whether this coincides with a recognized polymorphism or known microdeletion/microduplication syndrome
- See whether a similar deletion or duplication has been reported by any members of the Decipher consortium previously
- Record the phenotype of your patient, and compare this with previous records
- Visualise the known genes and putative genes and expressed-sequence tags (ESTs)

27th Feb 2009

February News Letter

We are pleased to announce our latest DECIPHER News ...

For full details click here...

23rd Jan 2009

January News Letter

We are pleased to announce our latest DECIPHER News ...

For full details click here...

18th Dec 2008

Genomic Disorders Conference

The dates for next year's meeting will be March 9th-11th 2009. Registration is now open, so please think about abstracts ...

For full details click here

Database of submicroscopic chromosomal imbalances, from array-CGH data. Focuses on data from patients with developmental delay, learning disabilities or congenital anomalies.