Copy number variation in association studies of human disease

Steven McCarroll

Broad Institute of MIT and Harvard

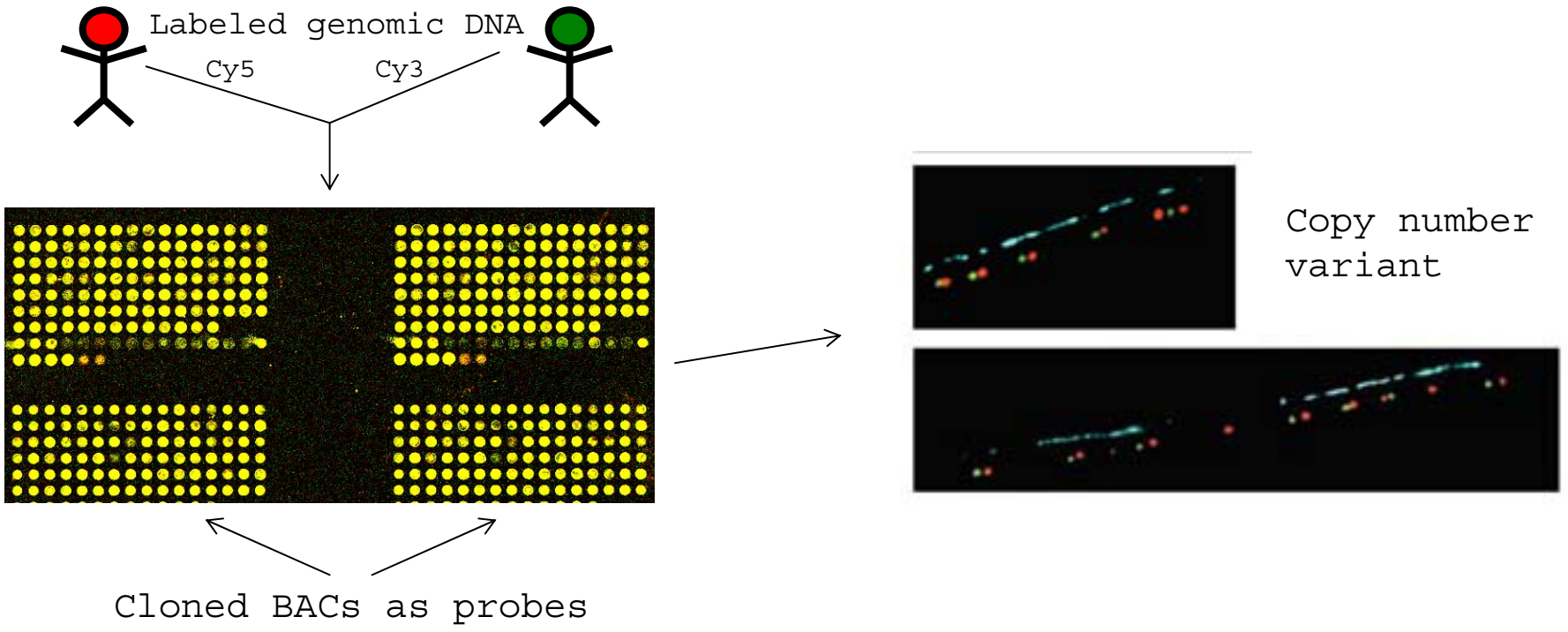
Outline

- Copy number variation
- Re-designing SNP arrays to interrogate copy number variation
- Analysis of copy number variation in whole genome association studies

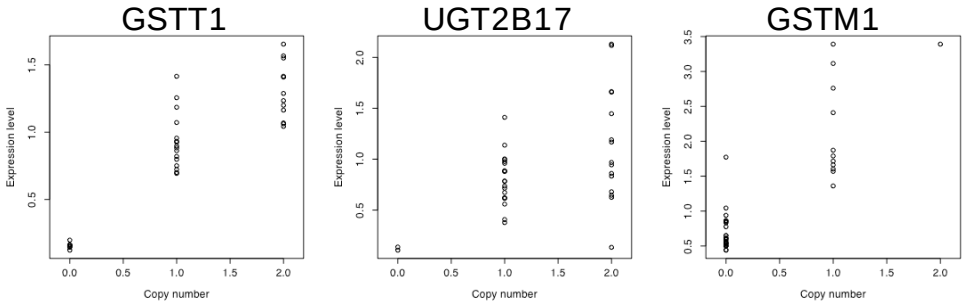
Copy number variation is common and extensive

Sebat et al., *Science* 2004

Iafrate et al., *Nature Genetics* 2004



Copy number variation shown to affect gene expression and some disease phenotypes



Deletions - McCarroll et al., 2006
>60 CNVs - Stranger et al., 2007

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

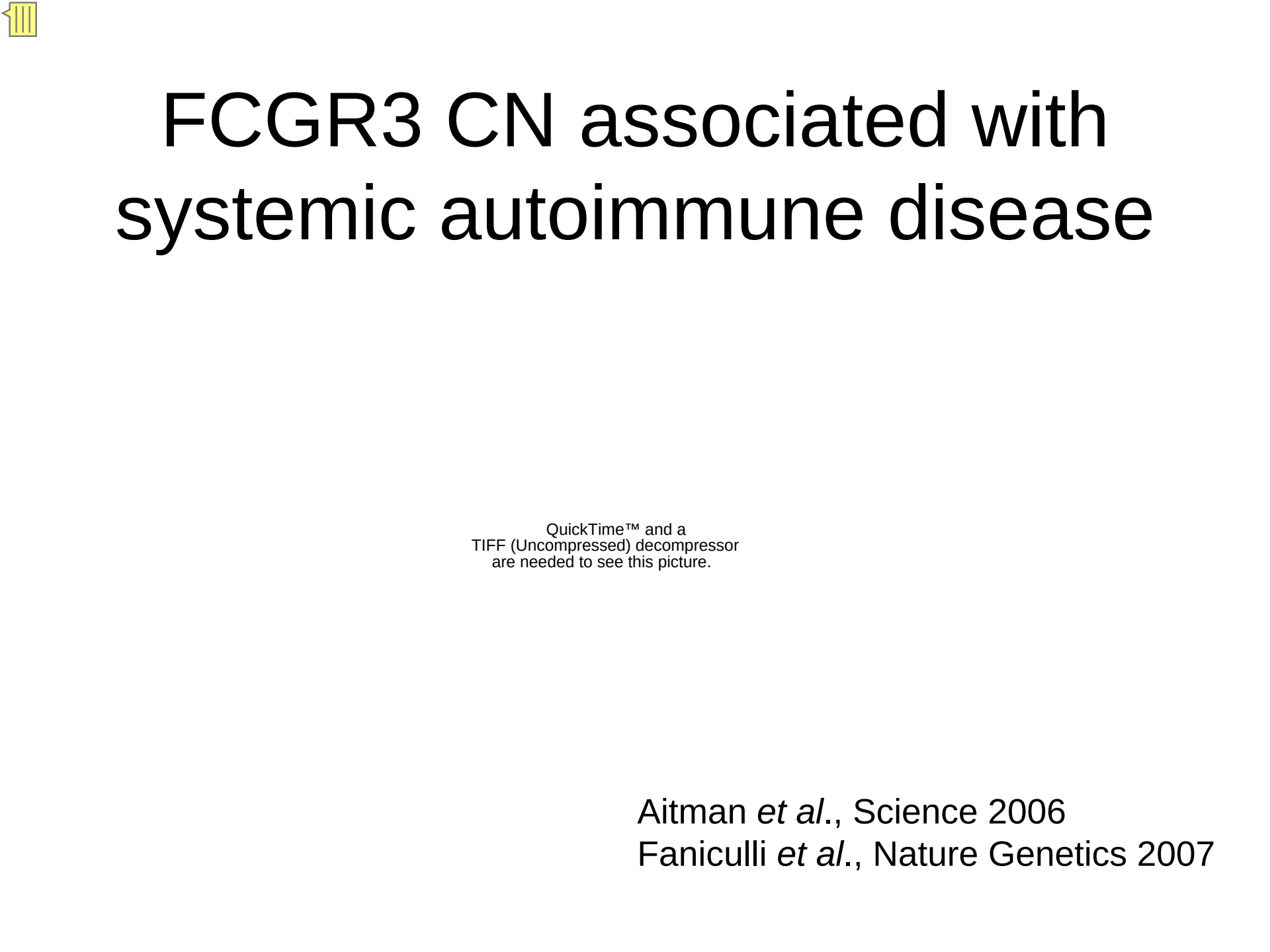
HIV progression - Gonzalez et al., 2005
Glomerulonephritis - Aitman et al., 2006
SLE - Faniculli et al., 2007

CCL3L1 CN associated with HIV progression to AIDS

Background:
CCL3L1 is the most potent known ligand for CC chemokine receptor 5 (CCR5), the major coreceptor for HIV, and it is a dominant HIV-suppressive chemokine

Gonzalez et al., Nature 2005

FCGR3 CN associated with systemic autoimmune disease

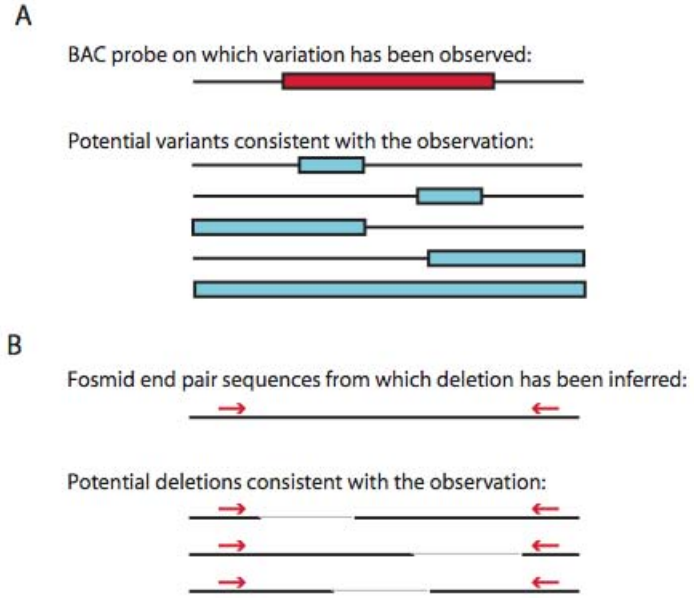


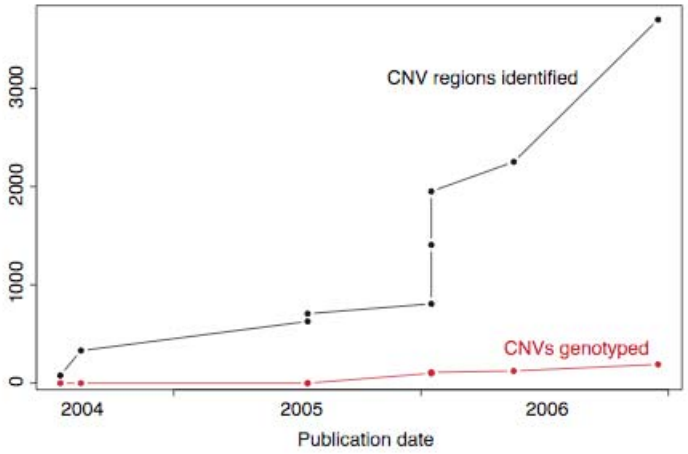
Aitman *et al*., Science 2006
Faniculli *et al*., Nature Genetics 2007

How broadly does copy number variation influence clinical phenotypes?

- A major outstanding question in the field

Need for high-resolution maps and genotyping technology

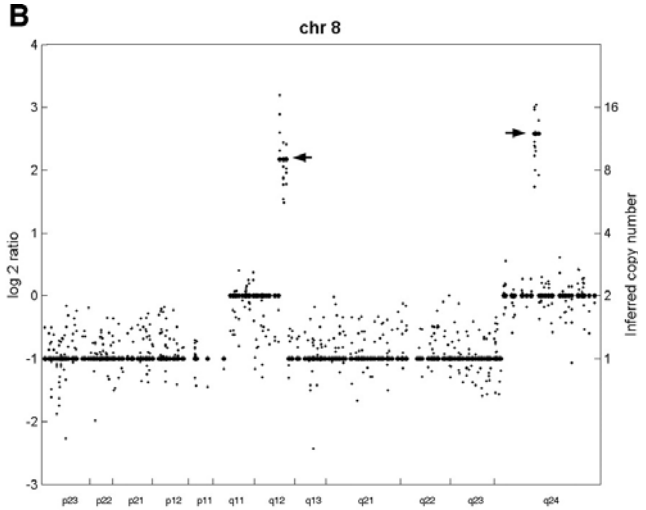




Outline

- Copy number variation
- Re-designing SNP arrays to interrogate copy number variation
- Analysis of copy number variation in whole genome association studies

CN analysis on earlier SNP arrays



Zhao et al., Cancer Res 2004

Little/no coverage of most common germline CNPs

SNP probes weren't optimized for copy number analysis

How could we improve CN analysis on SNP arrays?

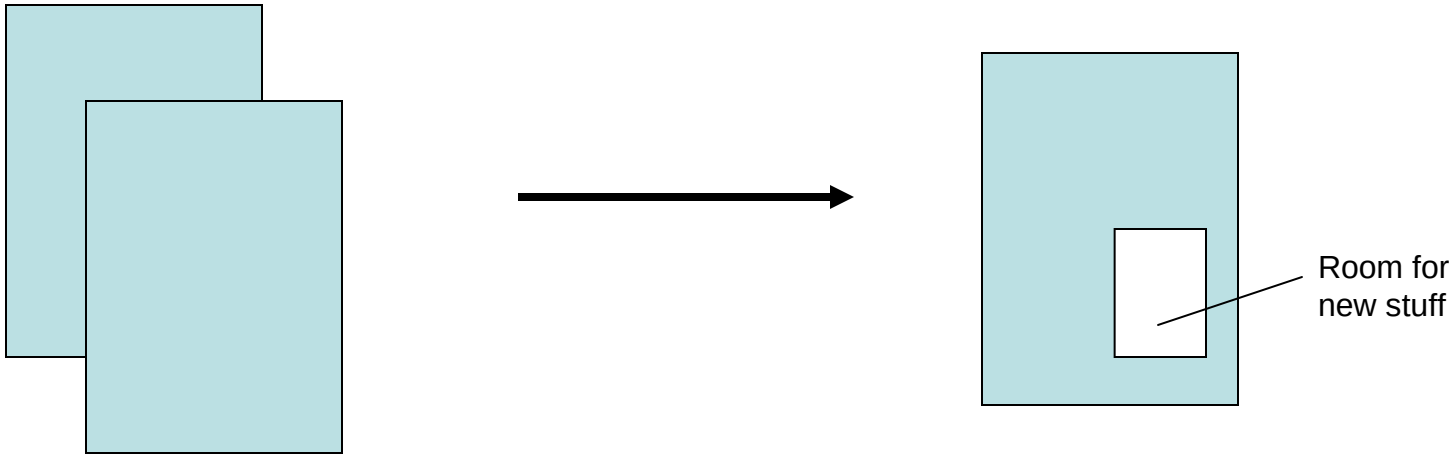
- Optimize probe locations in genome (within constraints of platform)
  - More-uniform coverage
  - Dense coverage in regions of interest (e.g. reported CNV)
- Optimize probe sequences for responsiveness
  - Unconstrained by SNP locations
  - Empirical probe selection

How could we improve CN analysis on SNP arrays?

- Optimize probe locations in genome (within constraints of platform)
  - More-uniform coverage
  - Dense coverage in regions of interest (e.g. reported CNV)
- Optimize probe sequences for responsiveness
  - Unconstrained by SNP locations
  - Empirical probe selection

---> an “expression” probe for DNA

Probe reduction made it possible to explore this idea



500K chip set

5.0 and 6.0 arrays

Selection of copy number probes (6.0 array)

- Selected
  - 800 K probes across genome -> more-uniform coverage
  - 140 K probes in “target regions”
    - Reported CNV regions
    - Regions of interest to cancer research

Copy number data on 6.0 array

- 1.8 M point measurements across the genome
  - 900 K from SNP probe sets
  - 940 K from individual copy number probes

Outline

- Copy number variation
- Re-designing SNP arrays to interrogate copy number variation
- Analysis of copy number variation in whole genome association studies

Copy number analyses

	Inherited CNPs	*De novo* CNVs
Mechanism	Inheritance	*De novo* mutation
# transmitted per meiosis	~ 100s *	~ 0.01 * (may be more in cases)
Distribution	Shared by many people	Unique or rare
Size	Vast majority < 100 kb	Up to several Mb

(* estimate limited by current methods for ascertainment)

	Inherited CNPs	*De novo* CNVs
Mechanism	Inheritance	*De novo* mutation
# transmitted per meiosis	~ 100s *	~ 0.01 * (may be more in cases)
Distribution	Shared by many people	Unique or rare
Size	Vast majority < 100 kb	Up to several Mb

(* estimate limited by current methods for ascertainment)

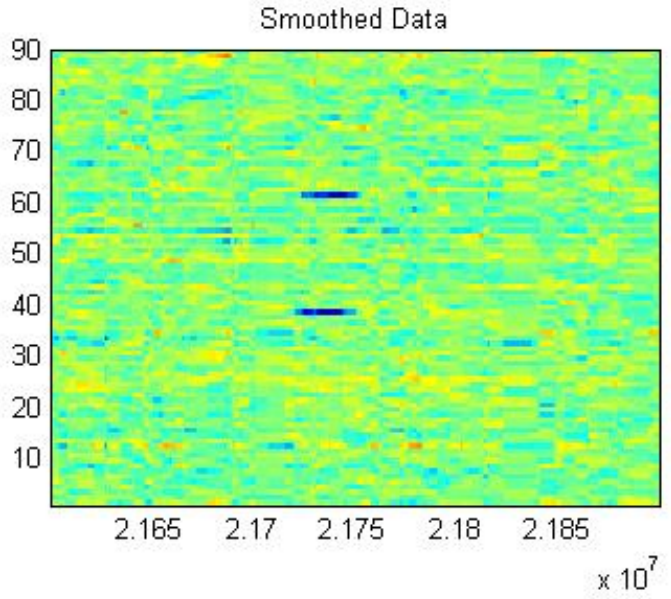
Finding copy number variants *de novo*

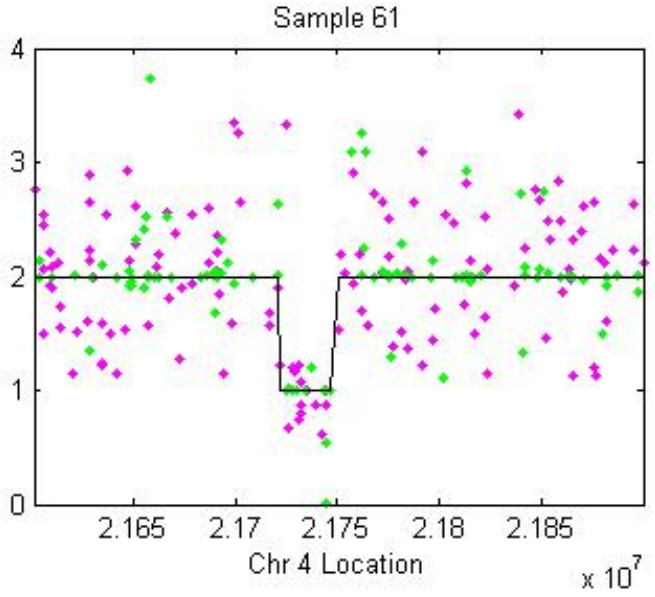
- In each sample or sample-pair, find genomic regions across which a series of probes indicates reduced (or increased) intensity
- Evidence for CNV must be strong (genome-wide significance)
- Various existing approaches
  - HMMs
  - Binary segmentation
  - Smith-Waterman

Birdseye (Josh Korn)

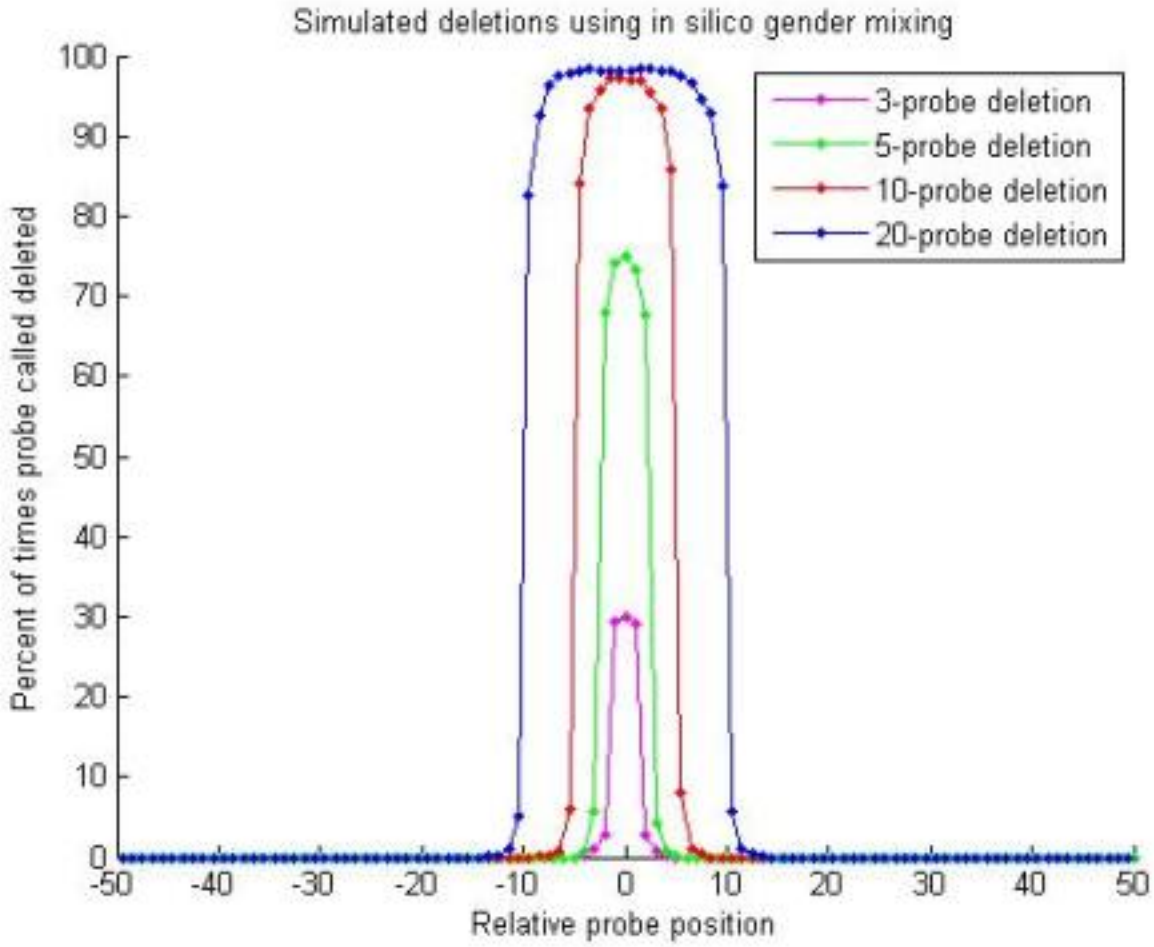
Sophisticated HMM
Uses CN and SNP probes together
Models the response properties of each probe
Discovers variants *de novo*

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.





Performance on in silico gender mixing experiment

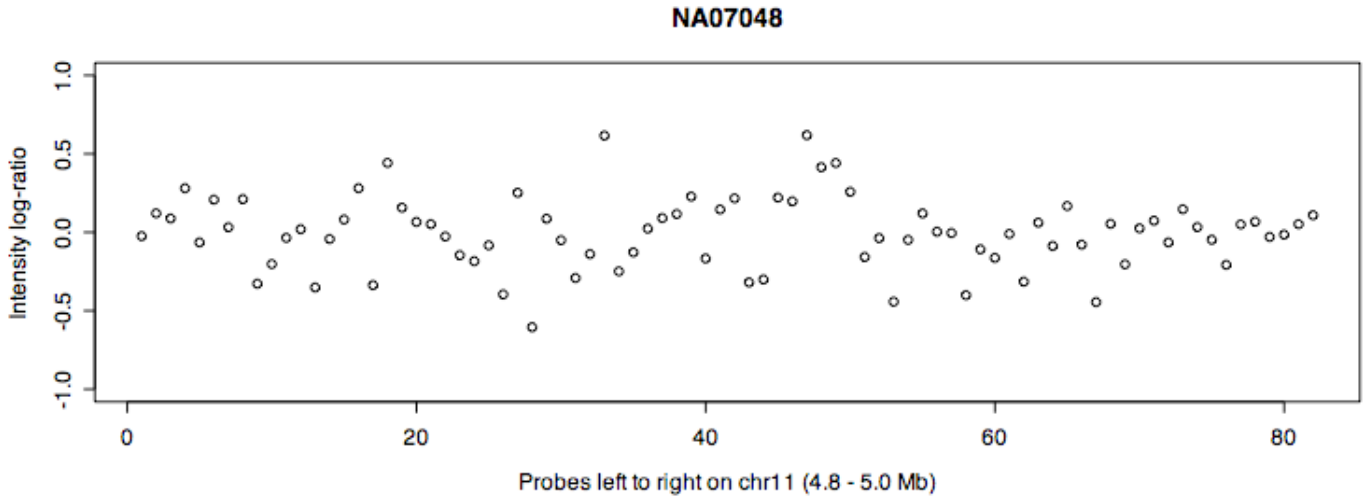


Potential association analyses for rare and *de novo* CNVs

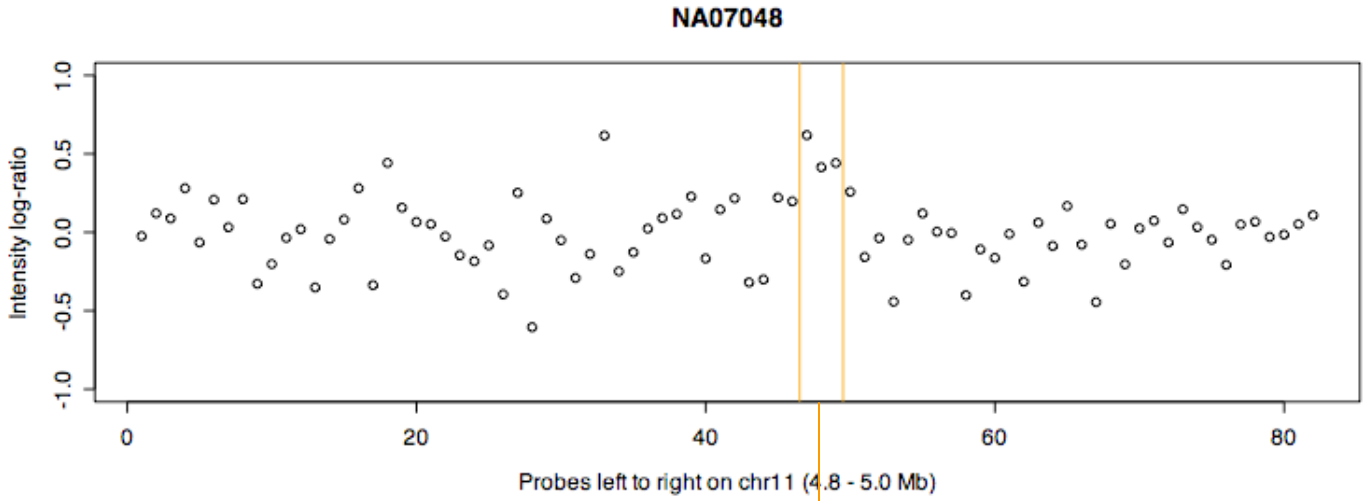
- Pileup at a particular genomic location
  - IMR, Sharp *et al*., Nat Genet 2006
- Other ways of summing CNVs
  - Across entire genome (sporadic autism, Sebat *et al*., Science 2006)
  - Critical to define criteria in advance
  - *Post hoc* criteria invented to fit observed data are unlikely to replicate in independent study
  - Precedent in analyses that group rare sequence variants (e.g. Cohen et al., Science 2004, NEJM 2006)
  - Discussion in McCarroll and Altshuler, Nat Genet 2007

	Inherited CNPs	*De novo* CNVs
Mechanism	Inheritance	*De novo* mutation
# transmitted per meiosis	~ 100s *	~ 0.01 * (may be more in cases)
Distribution	Shared by many people	Unique or rare
Size	Vast majority < 100 kb	Up to several Mb

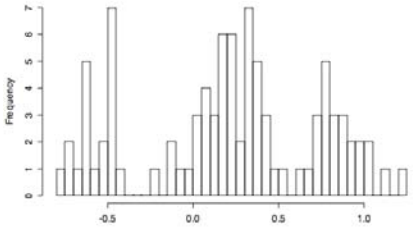
(* estimate limited by current methods for ascertainment)


NA07048
Intensity log-ratio
1.0
0.5
0.0
-0.5
-1.0
0
20
40
60
80
Probes left to right on chr11 (4.8 - 5.0 Mb)

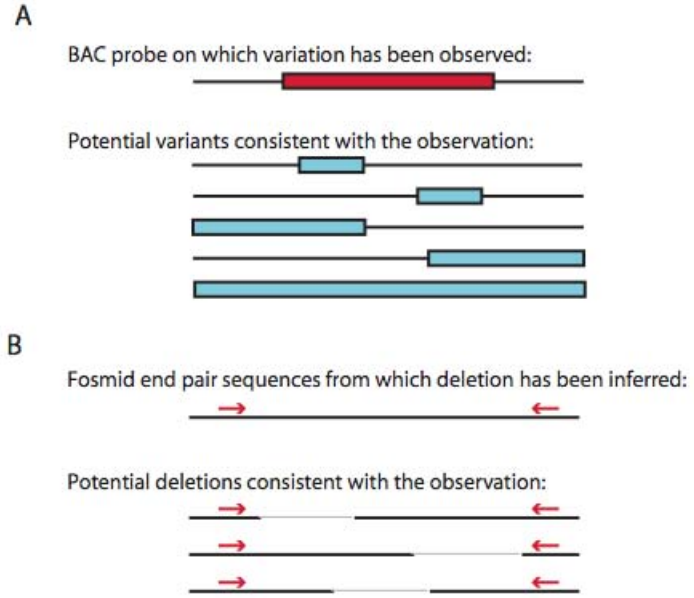
CNP analysis using prior information about CNP locations

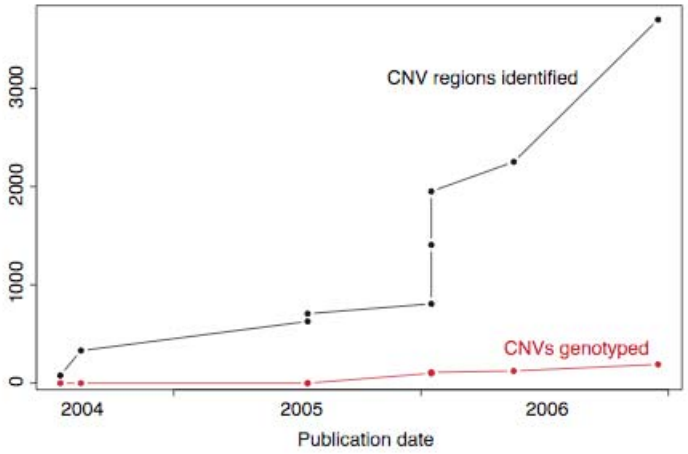


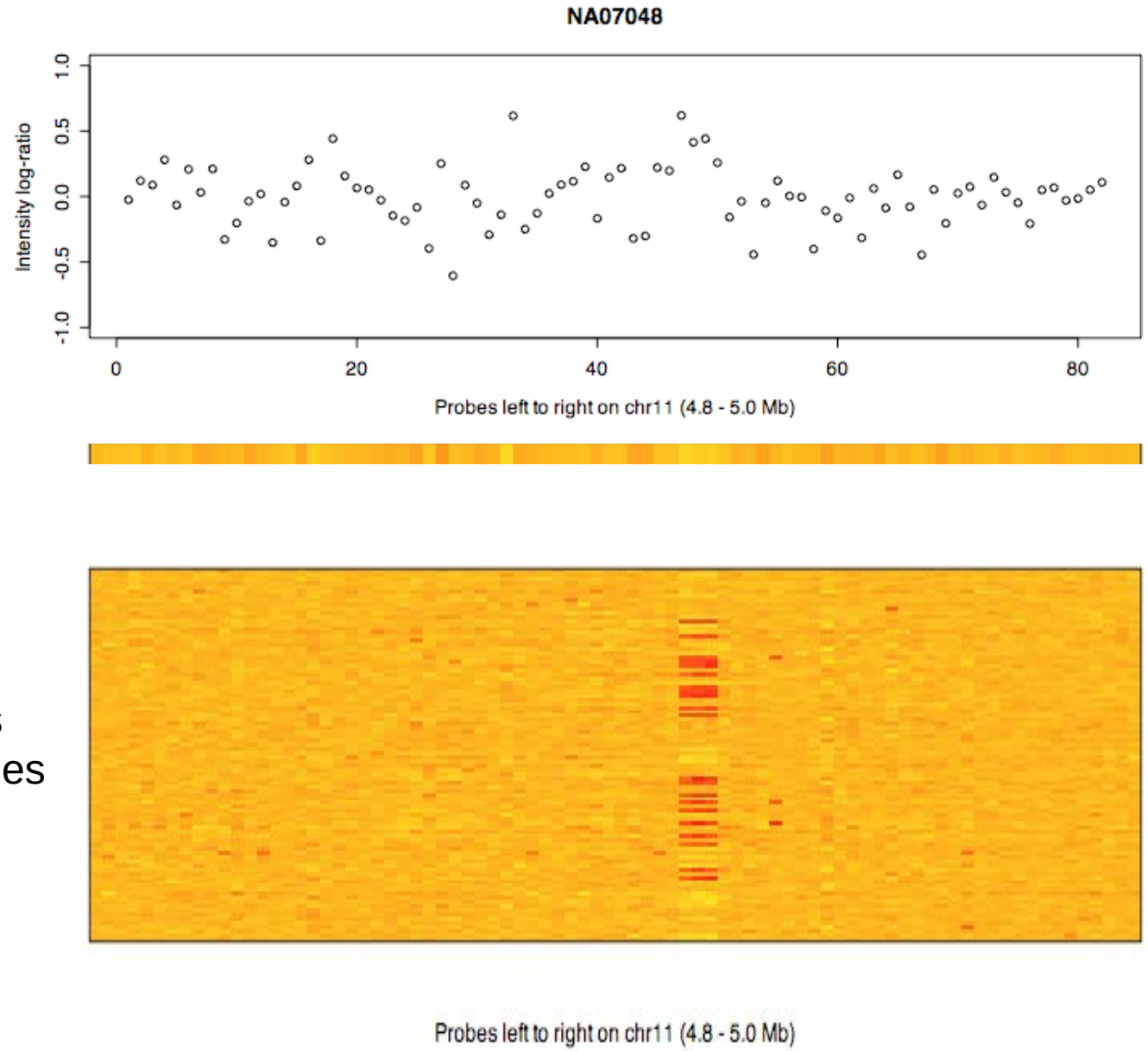
Distribution across population



Need for high-resolution maps and genotyping technology





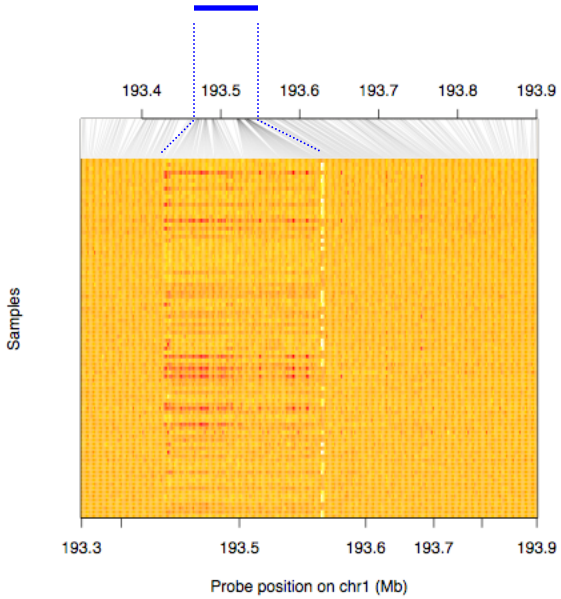


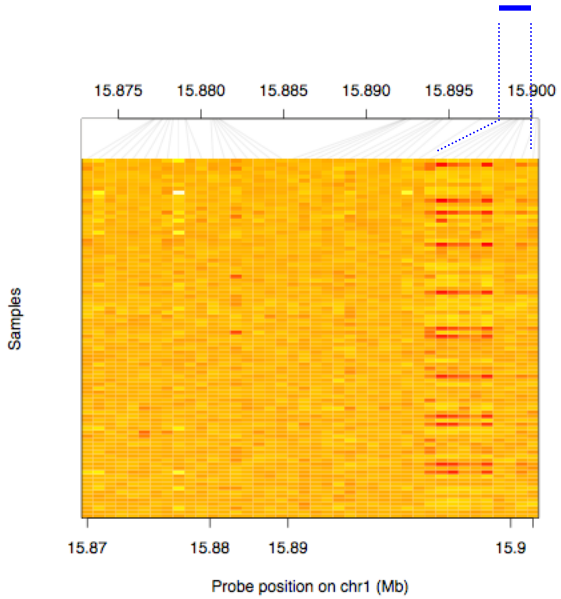
Across
90 samples

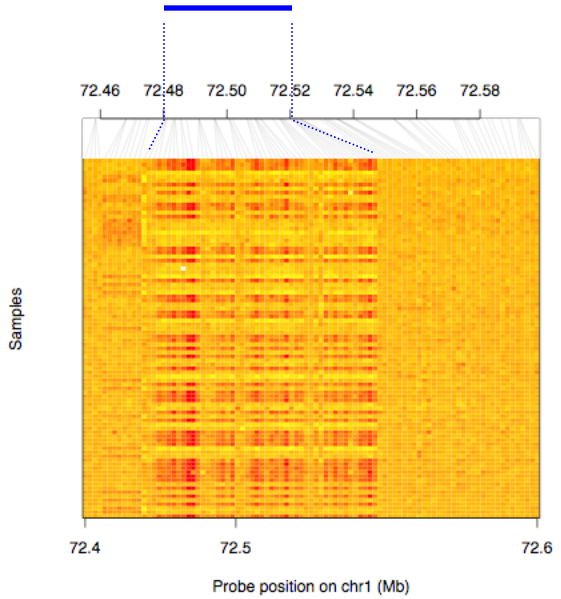
High-resolution map of CNPs

Reported CVNR:

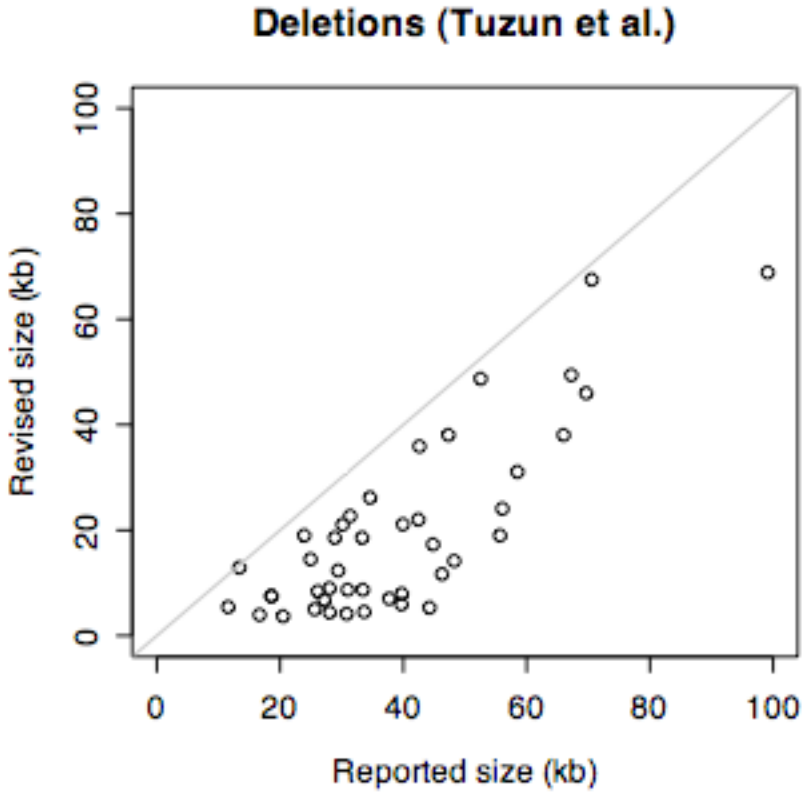
Revised CVNR:

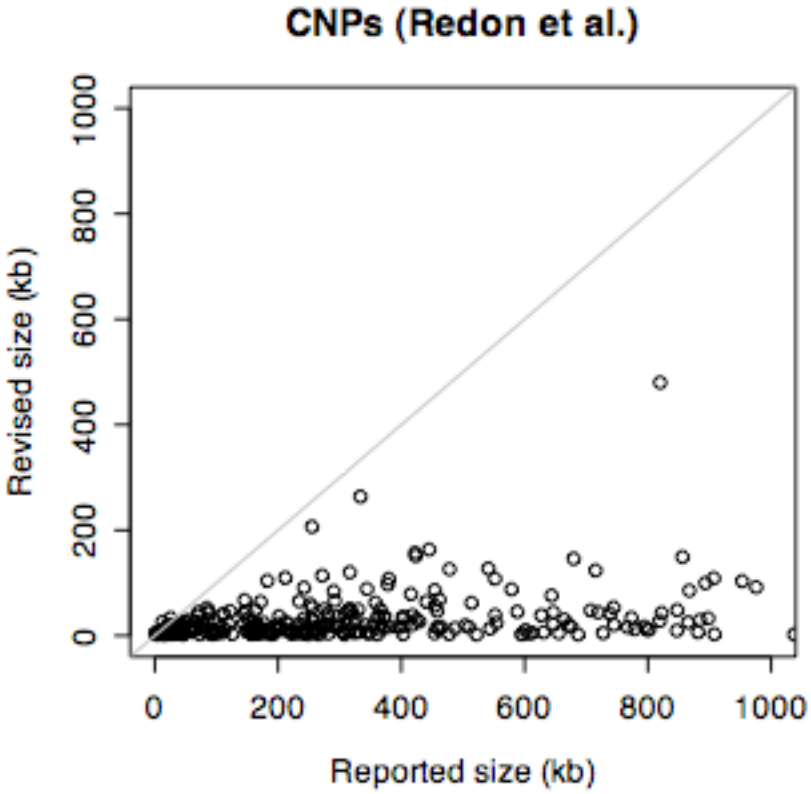






Most CNVs smaller than reported





Compared to earlier survey, more CNVs, but less of genome affected

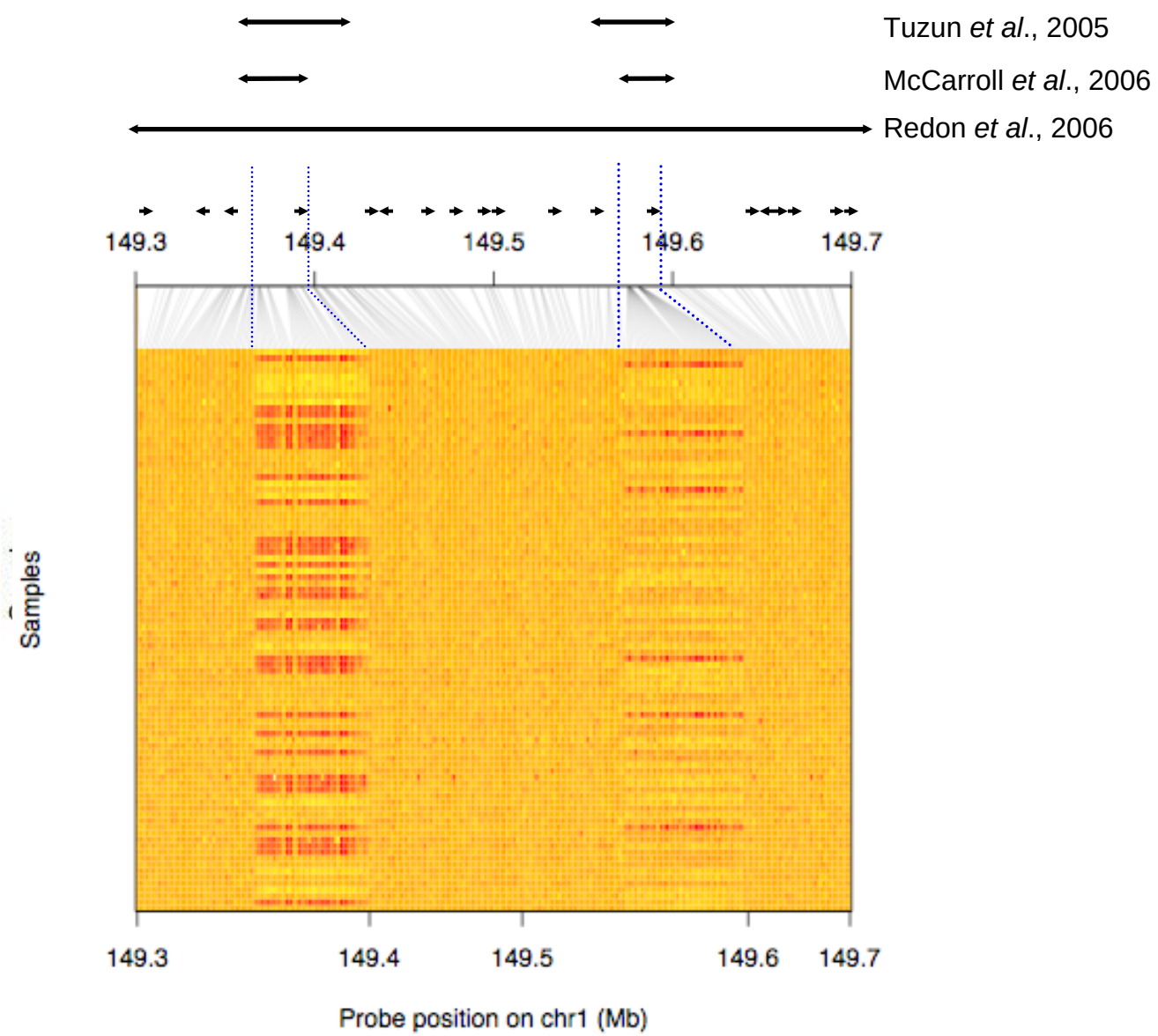
- Recent survey (Redon *et al.*, 2006)
  - 70 loci at different copy number levels between any two people
  - Mean size 228 kb
  - Cover 16 MB of genome
- This survey
  - 250 loci at different copy number levels between any two people
  - Mean size 20 kb
  - Cover 5 MB of genome

Revising gene content of CNVs

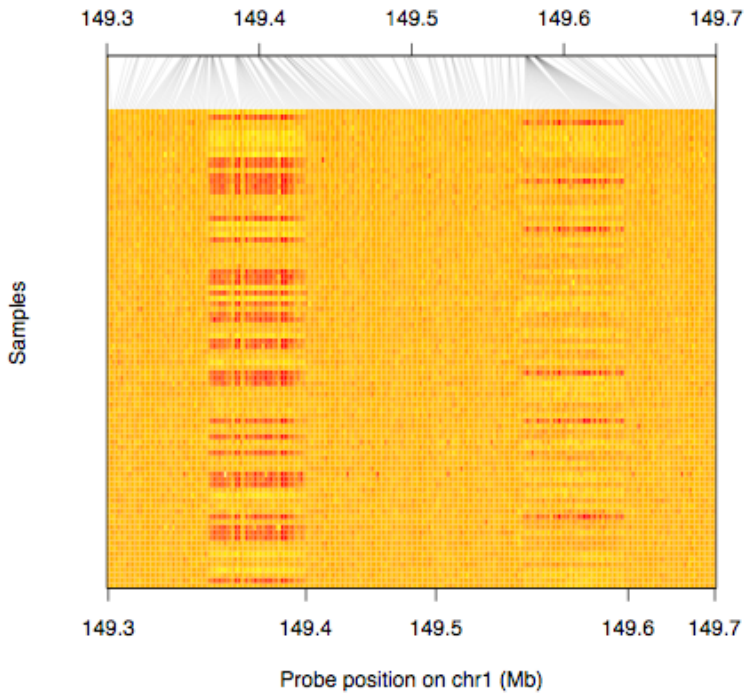
Earlier reports of CNVs

Genes

6.0 Array

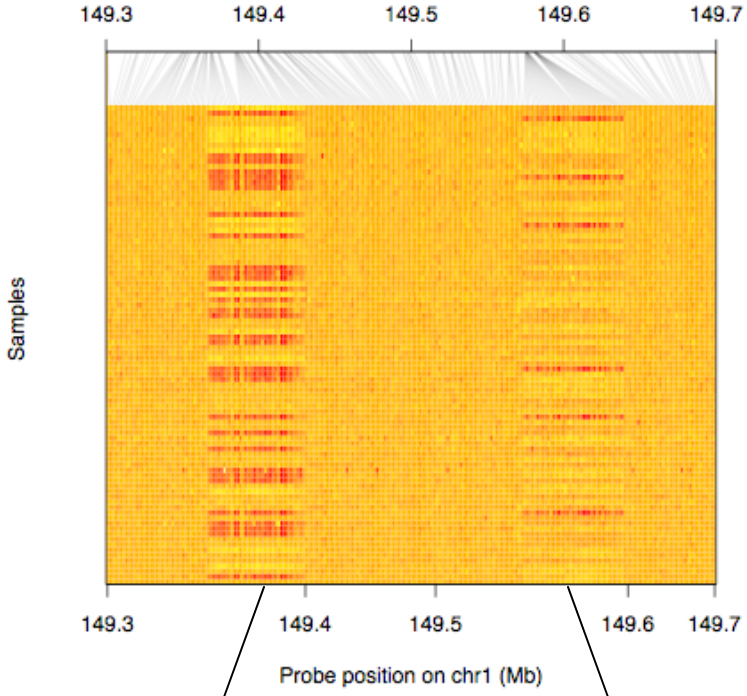


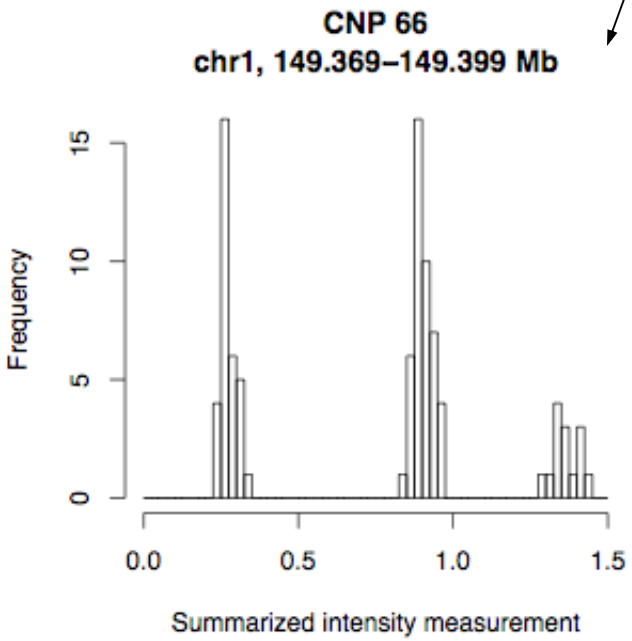
Most CNPs aren't directly interrogated by SNP assays (or by earlier platforms)

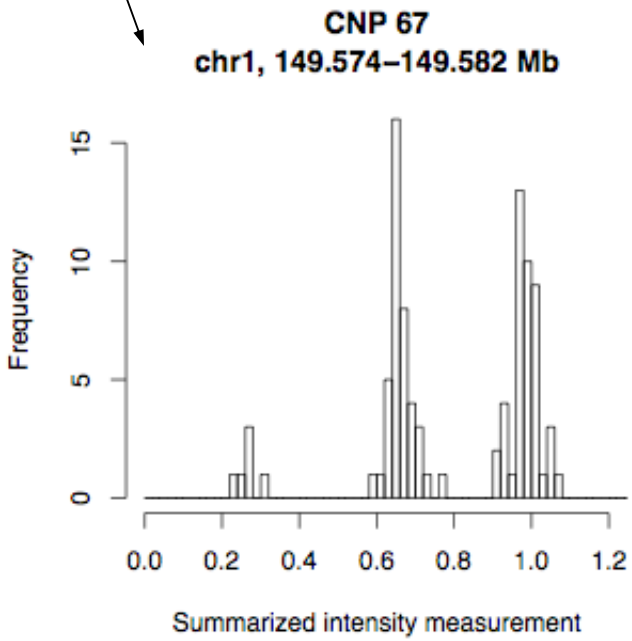


	SNP assays in "CNV region" (DGV)	SNP assays in actual CNVs within region
Affymetrix 500K	110	0
Illumina 300K	24	0

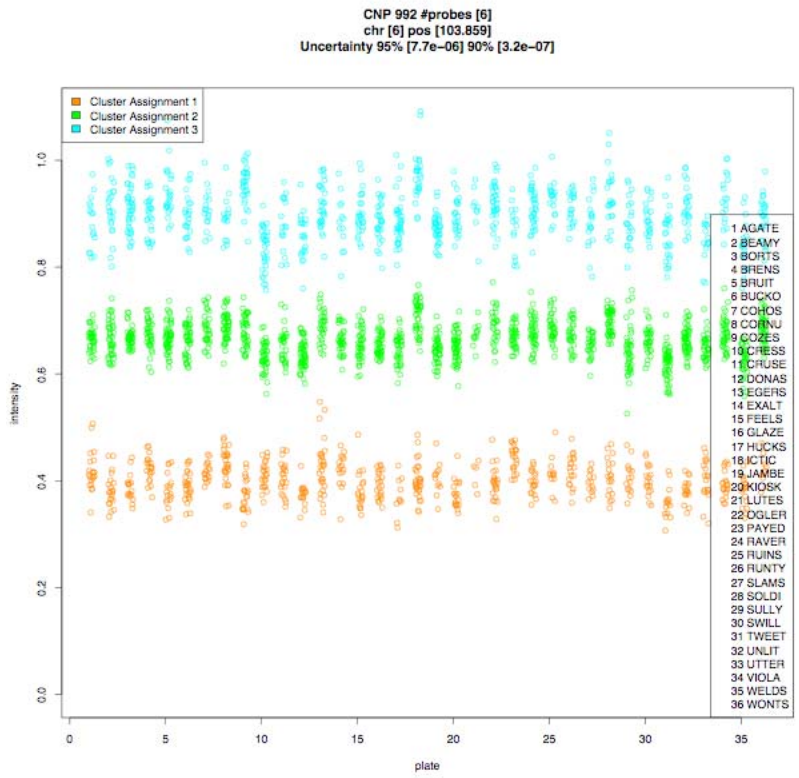
From CNVs
to *genotypes*

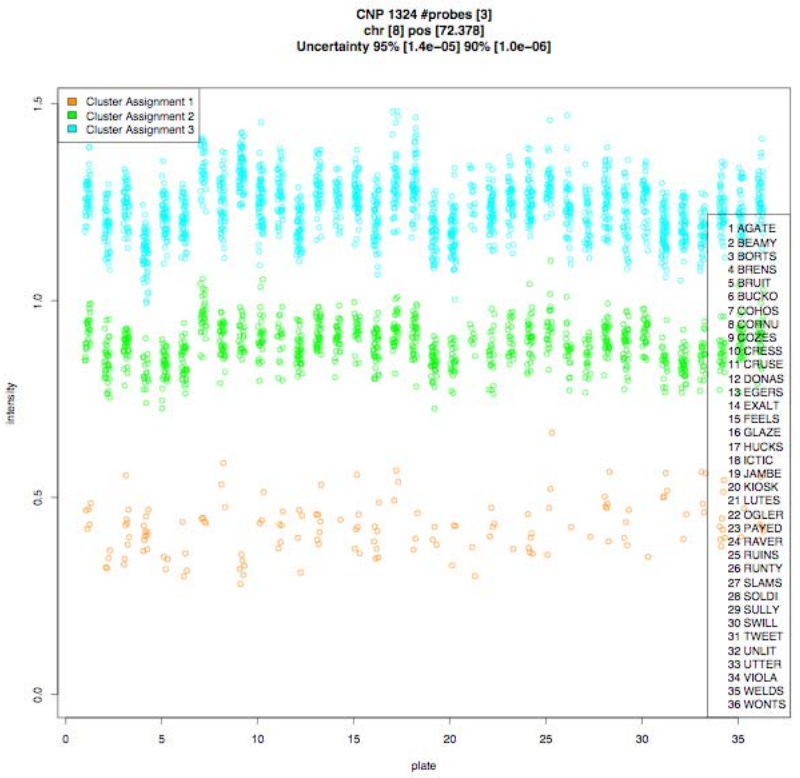






CNP genotyping in large studies





Evaluating the accuracy of CNP genotypes by inheritance

Minor allele frequency	Population	# of CNVs	Mendel failures per trio
1% - 5%	CEU	136	0.008
	YRI	209	0.004
5% - 15%	CEU	219	0.007
	YRI	364	0.008
15% - 50%	CEU	340	0.008
	YRI	387	0.008

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

- Accomplished in CANARY
  - Genotype samples for several hundred common, inherited CNPs

How much does copy number variation influence clinical phenotypes?

- A major outstanding question in the field
- Enabled by new array technology, will soon be possible to learn

Finny Kuruvilla

Josh Korn

Alec Wysoker
Jim Nemesh
Paul de Bakker
Casey Gates
Marcia Nizarri

Simon Cawley
Steve Lincoln
Keith Jones

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Stacey Gabriel

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Mark Daly

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

David Altshuler

Acknowledgements

Affymetrix
Simon Cawley
Steve Lincoln
Keith Jones
Earl Hubbell
Teresa Webster
Sean Walsh
Rui Mei
Xiaojun Di
Geoff Yang
Hajime Matsuzaki
Guoying Liu
Alan Williams
Harley Gorrell
Chuck Sugnet
Fan Shen
Michael Shapero

Finny Kuruvilla

Josh Korn

Alec Wysoker
Jim Nemesh
Paul de Bakker
Casey Gates
Marcia Nizarri

Simon Cawley
Steve Lincoln
Keith Jones

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Stacey Gabriel

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Mark Daly

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

David Altshuler