Estimation problems in high throughput SNP platforms

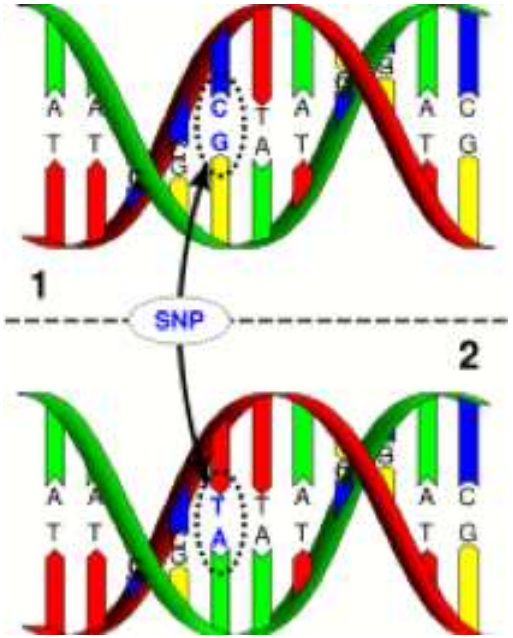
Rob Scharpf

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

November 24, 2008

Outline

Single nucleotide polymorphisms



- occur every 100 to 1000 bp
- minor allele frequence $\geq$ 1%

Copy number

- Autosomal copy number: 2
- Chromosome X: women have 2 copies / men have 1 copy
- copy number alterations:
  - homozygous deletion (0 copies)
  - hemizygous deletions (1 copy)
  - amplification (3 or more copies)
- copy number-neutral alterations: loss of heterozygosity (LOH)
- copy number variants can be rare or common (polymorphisms)
-

Copy Number Variation (CNV)

Copy number variants are segments of DNA typically > 1000 basepairs with altered copy number (e.g., hemizygous deletion, amplification)

Frequency:

- Thousands of genes are variable in copy number
- The more than 6000 CNVs currently in the Toronto Database of Genomic Variants is likely an underestimate

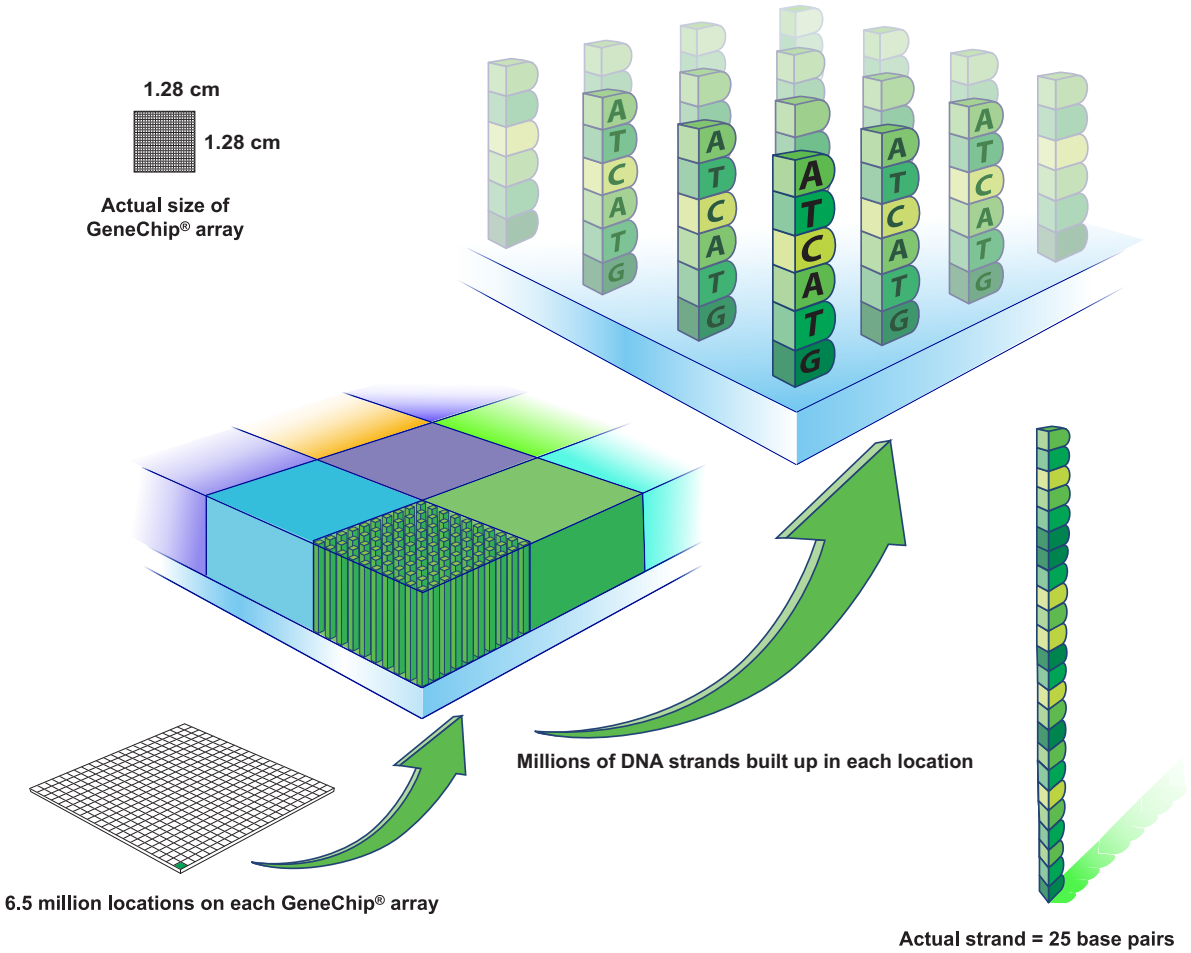
Biology:

- CNVs have been implicated in diseases such as autism, cancer, and diabetes
- Statistical methods to assess the contribution of CNVs to disease susceptibility are under development

Mechanism of copy number alterations

- See recent review by Gu *et al.*(2008), PathoGenetics

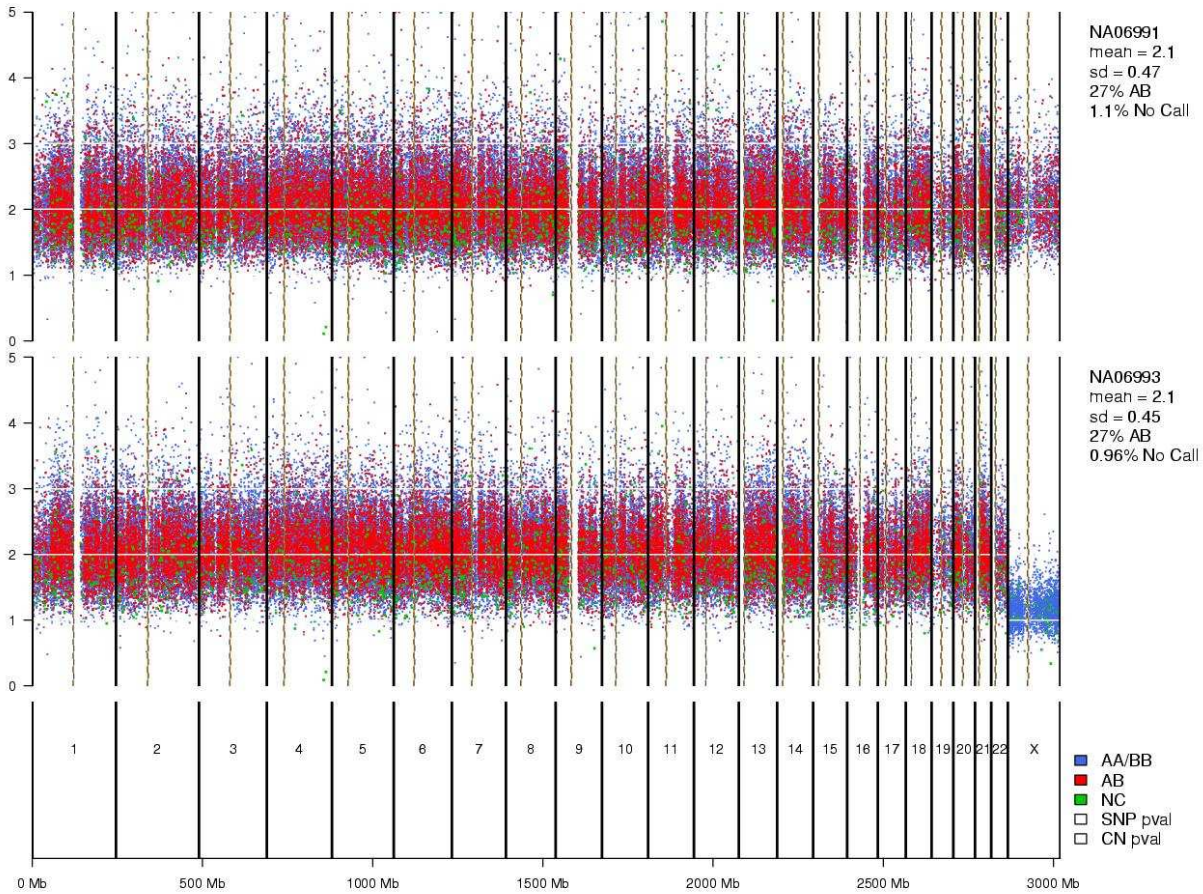
Affymetrix


1.28 cm
1.28 cm
Actual size of GeneChip® array
6.5 million locations on each GeneChip® array
Millions of DNA strands built up in each location
Actual strand = 25 base pairs

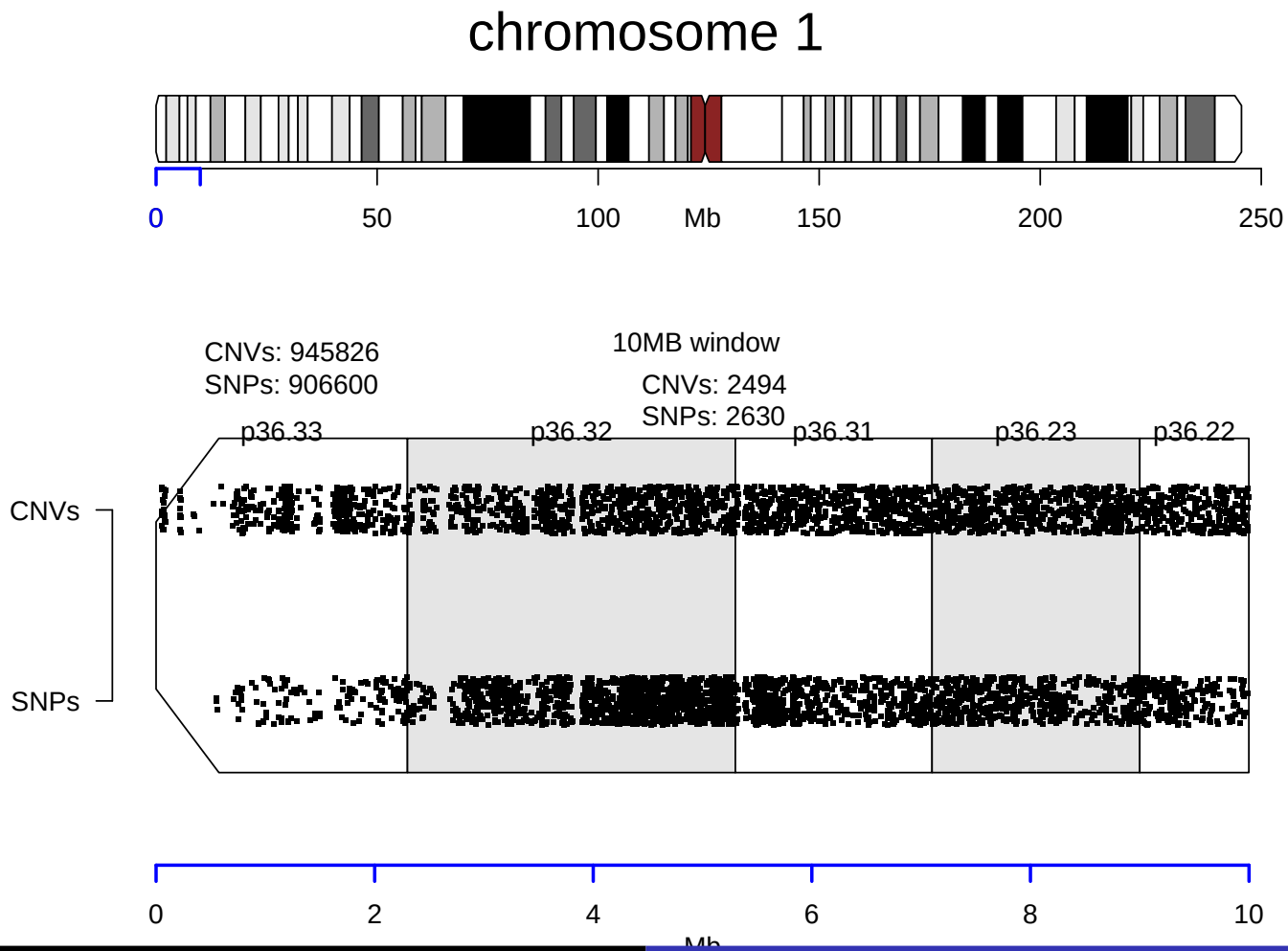
Illumina

- Gunderson *et al.* (2005), Nature Genetics
- References: Matt Richie (see recent Expressionists talk for overview)
- R Packages: beadarray, beadarraySNP

Affymetrix data (after some processing)

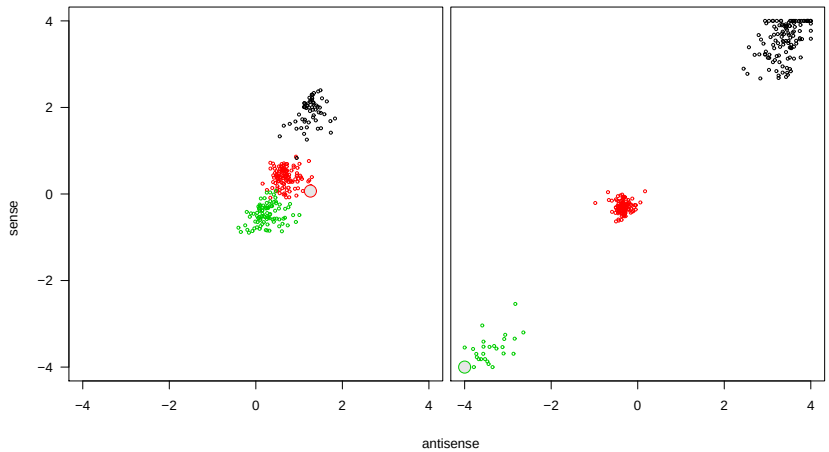


Polymorphic and nonpolymorphic probes



Uncertainty of point estimates

- Most genotyping algorithms are accurate and concordant for over 99.9% of diallelic SNPs. However, there is uncertainty in the genotype estimates that differs by SNP.



Multiple levels of variation

- the entire chromosome (e.g. trisomy 21)
- segmental changes such as insertions, deletions, inversions, and translocations
- small genomic regions including SNPs

Preprocessing

- Ideally, preprocessing method should provide normalized intensities that are robust to differences in labs and batch (this may not always be true)
- SNP-RMA (oligo package at Bioconductor) quantile normalizes the raw fluorescence intensities to a Hapmap reference distribution
- the normalized intensities can then be fed into algorithms for genotype calling (e.g., CRLMM) or copy number

Three tiers of estimation problems for germline diseases

- By locus: How can we use the low-level data to optimally estimate the genotype and DNA copy number for each locus in the array?
- By sample: how can we borrow strength between neighbouring loci, and infer regions of loss of heterozygosity (LOH) and CNV in the genome of the subject studied?
- Between samples: how can we compare the genotypes and copy numbers of many subjects, infer common regions of alterations, and assess differences between affected subjects and normal controls?

Tier 1: By locus

How can we use the low-level data to optimally estimate the genotype and DNA copy number for each locus in the array?

Available software for genotype calling

Affymetrix:

- Affymetrix Power Tools (APT)
  Affymetrix Inc. (2006), White Paper

  `http://www.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx`
- BRLMM (now the default used by APT)
  Rabbee and Speed (2006), Bioinformatics
- R package oligo – CRLMM
  Carvalho *et al.* (2007), Biostatistics

Illumina:

- BeadStudio
- . . .

Available software for copy number

Affymetrix:

- Affymetrix Power Tools (APT)
- R package aroma-affymetrix
  H. Bengtsson *et al.*(2008), Bioinformatics
- Others: Golden Helix (for more info, attend a webinar. Also supports Illumina)
- ITALICS (an iterative method for normalization. Available for 6.0?)
  G. Rigaill *et al.* (2008), Bioinformatics
- PLASQ (allele-specific copy number estimation)
  T. LaFramboise *et al.* (2007), Biostatistics
- dCHIP

Illumina:

- BeadStudio (log R ratios)
- Golden Helix
- R package: beadarraySNP
- ...

Some copy number tools are not listed here because they smooth across SNPs (e.g., ITALICS, CBS, GLAD, VanillaICE, PennCNV). More about these in Tier 2.

Basic approach to locus-specific estimation of copy number

Statistical model:

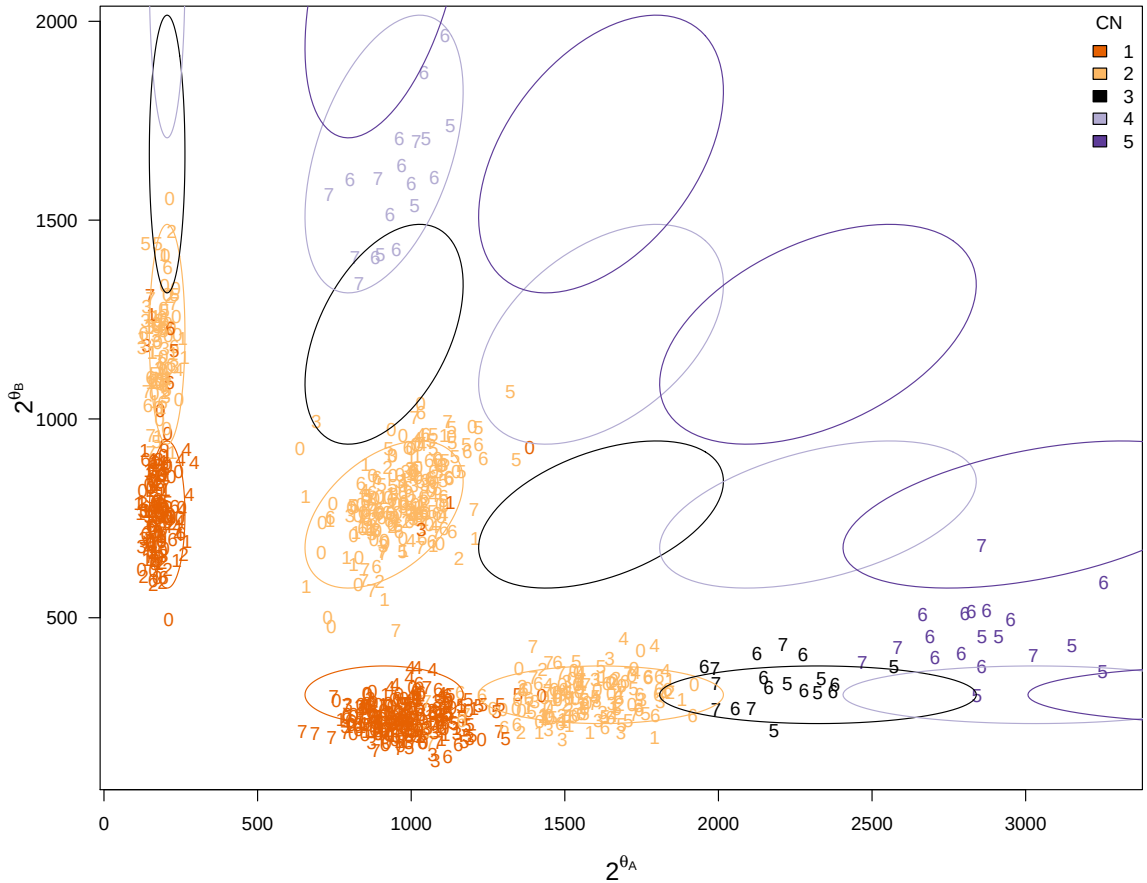
$$\text{Observed} = \text{Background} + \text{Nonspecific} + \text{Specific}$$

Algorithm

1. Derive locus-specific estimates of these parameters from a large training set where individuals are assumed to have two copies at each locus.
2. Using locus-specific parameters estimated from training data, estimate the copy number and provide quantifications of the uncertainty in a new dataset.

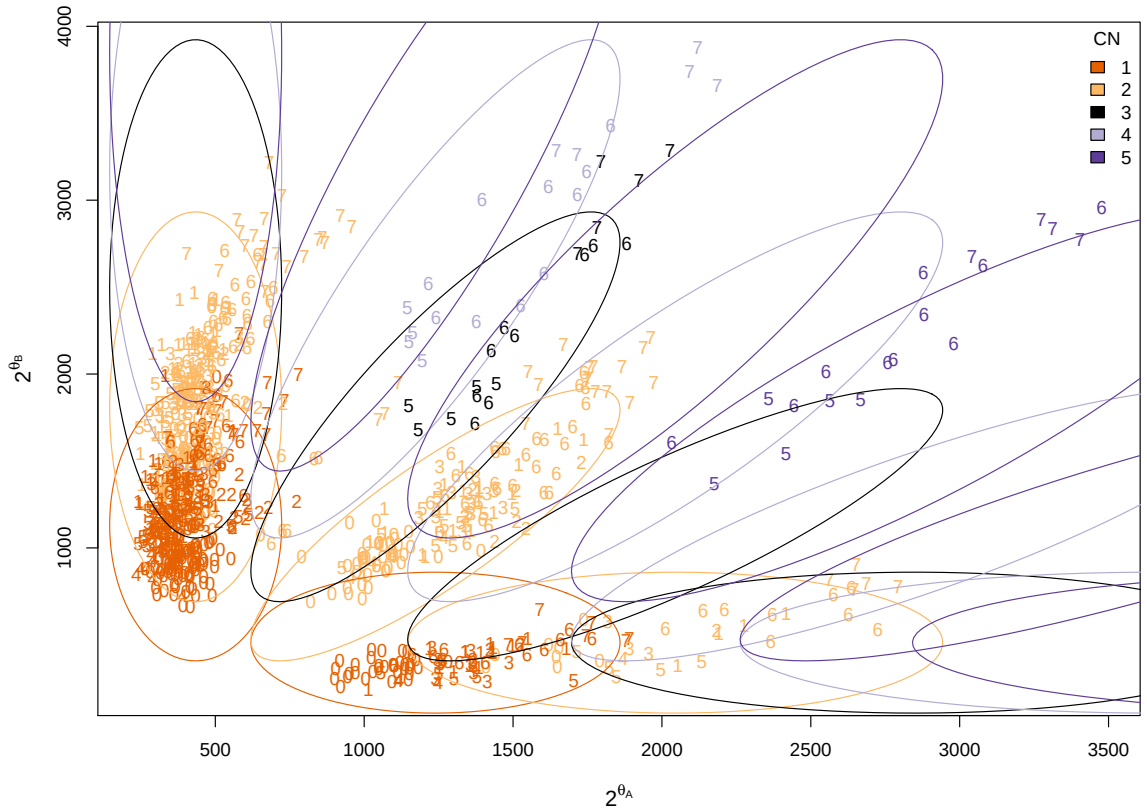
- Improve locus estimates by modeling the spatial dependence across SNPs – Tier 2.

Low uncertainty



Observed intensities for the B (y-axis) and A (x-axis) alleles for a single SNP

High uncertainty



Important points

- Probes differ in their ability to quantify copy number
- Important to propogate uncertainty estimates to downstream analyses
- Important to investigate differences between batches and labs

Tier 2: By sample

- How can we *borrow strength* between loci to infer regions of loss of heterozygosity (LOH) and CNV in the genome of the subject studied?
- The choice of statistical methodology for smoothing/segmenting locus-specific estimates of copy number may depend on disease etiology
  - nonparametric segmentation approaches are particularly useful for cancer samples when the copy number may be noninteger (attributable, for instance, to an impure biopsy of the tumor)
  - for germline diseases in which all cells are believed to have the same DNA, hidden Markov models have the advantage of making joint inference from the genotype calls and copy number estimates, inferring both copy number neutral (LOH) and nonneutral alterations (deletions, amplifications)

Considerations

- Motivation: Copy number and genotype estimates are correlated in high density SNP platforms due to underlying haplotype structure and linkage disequilibrium
- Probes differ in their ability to quantify copy number
  - When borrowing strength across loci, we should weight by the inverse of the variances of the locus-level estimates
- When trios of parents and proband are available, we can distinguish between *de-novo* and inherited events and use Mendelian rules of inheritance to improve the precision of breakpoints

Available Software (incomplete)

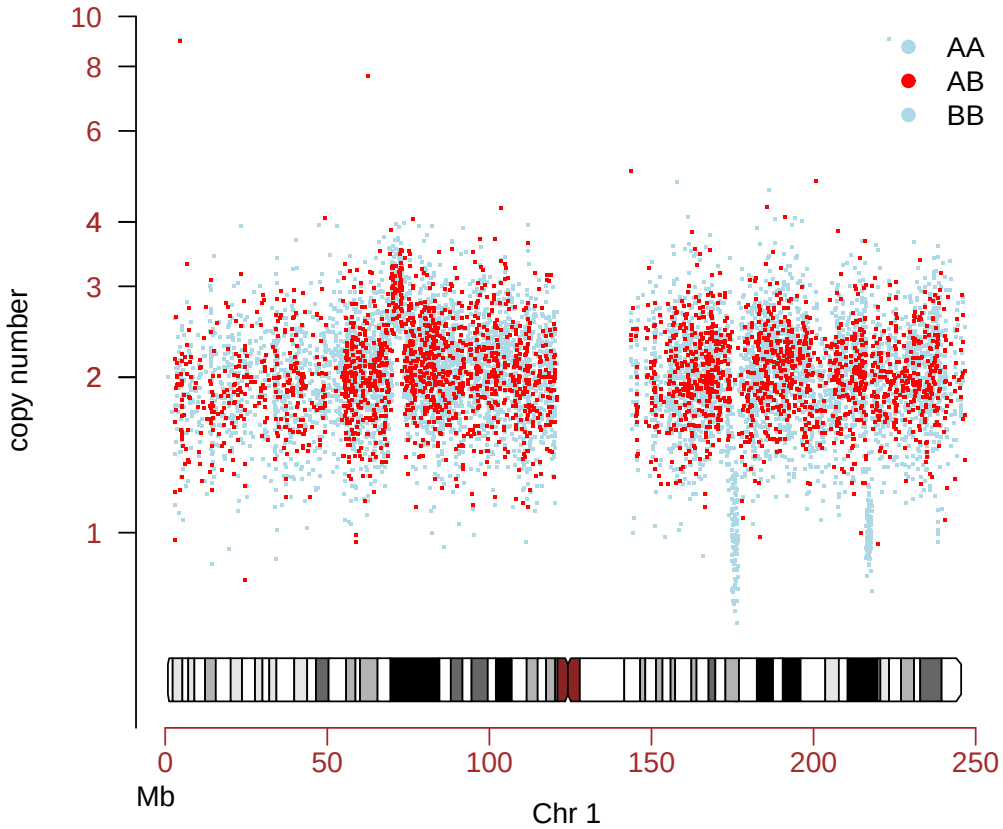
Affymetrix:

- APT uses a hidden Markov model
- R package VanillaICE
  HMM described in R. Scharpf *et al.* (2008) Annals of Applied Statistics
- R package DNAcopy
  Circular Binary Segmentation described in Olshen *et al.* (2004) Biostatistics

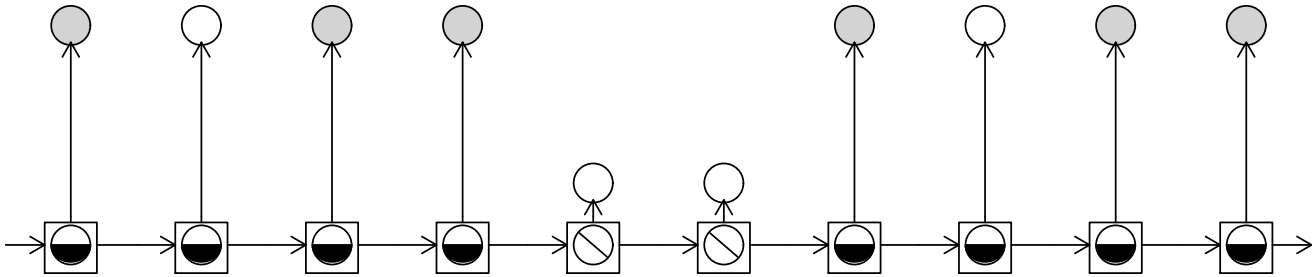
Illumina:

- QuantiSNP
  S. Colella *et al.* (2007), Nucleic Acids Research
- PennCNV
  K. Wang *et al.* (2008), Nucleid Acids Research
- ...

Simulated data for chromosome 1 of one sample

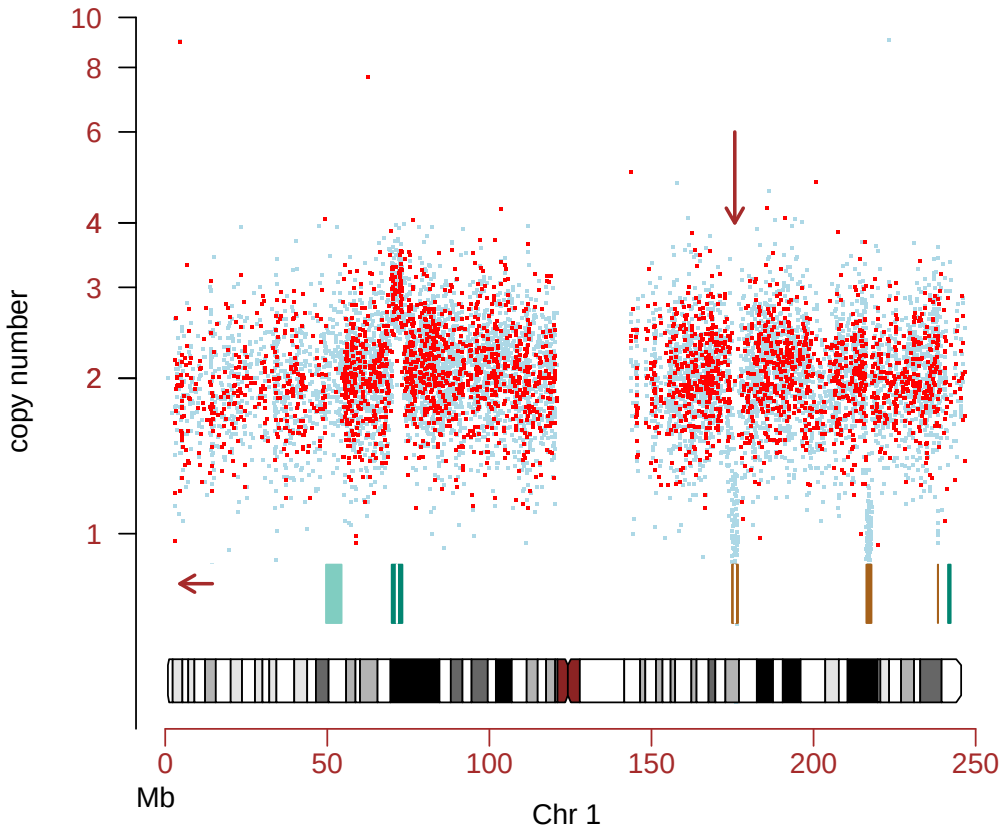


Hidden Markov Models (HMM)

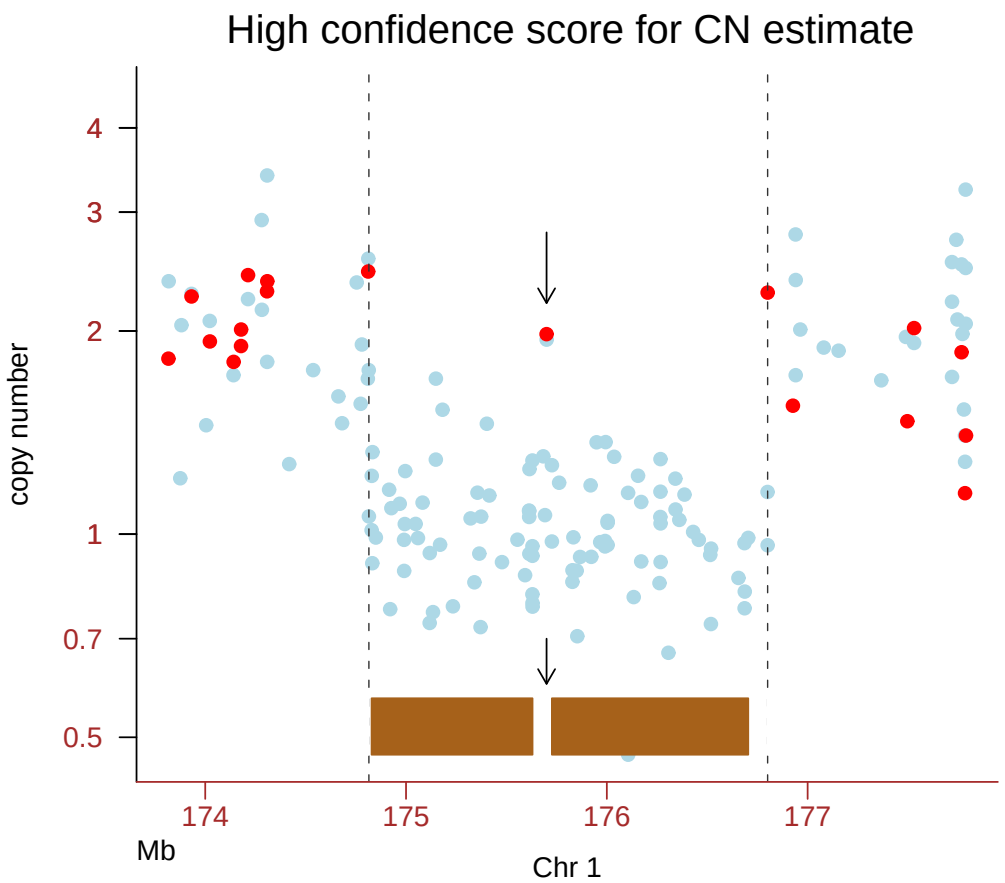


- Emission probabilities (vertical arrows): the probability of what we observe given the underlying hidden state
- Transition probabilities (horizontal arrows): the hidden state (in boxes) at a SNP is dependent on the true state of the previous SNP due to haplotype structure and linkage disequilibrium

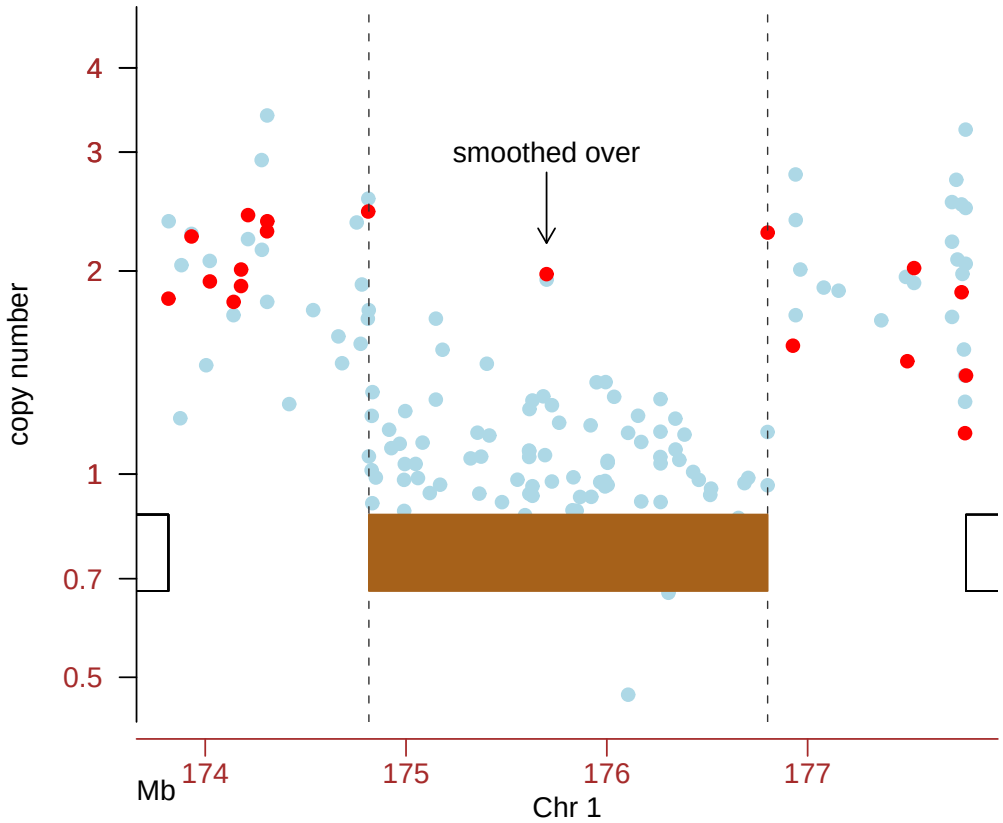
HMM Predictions



Low uncertainty

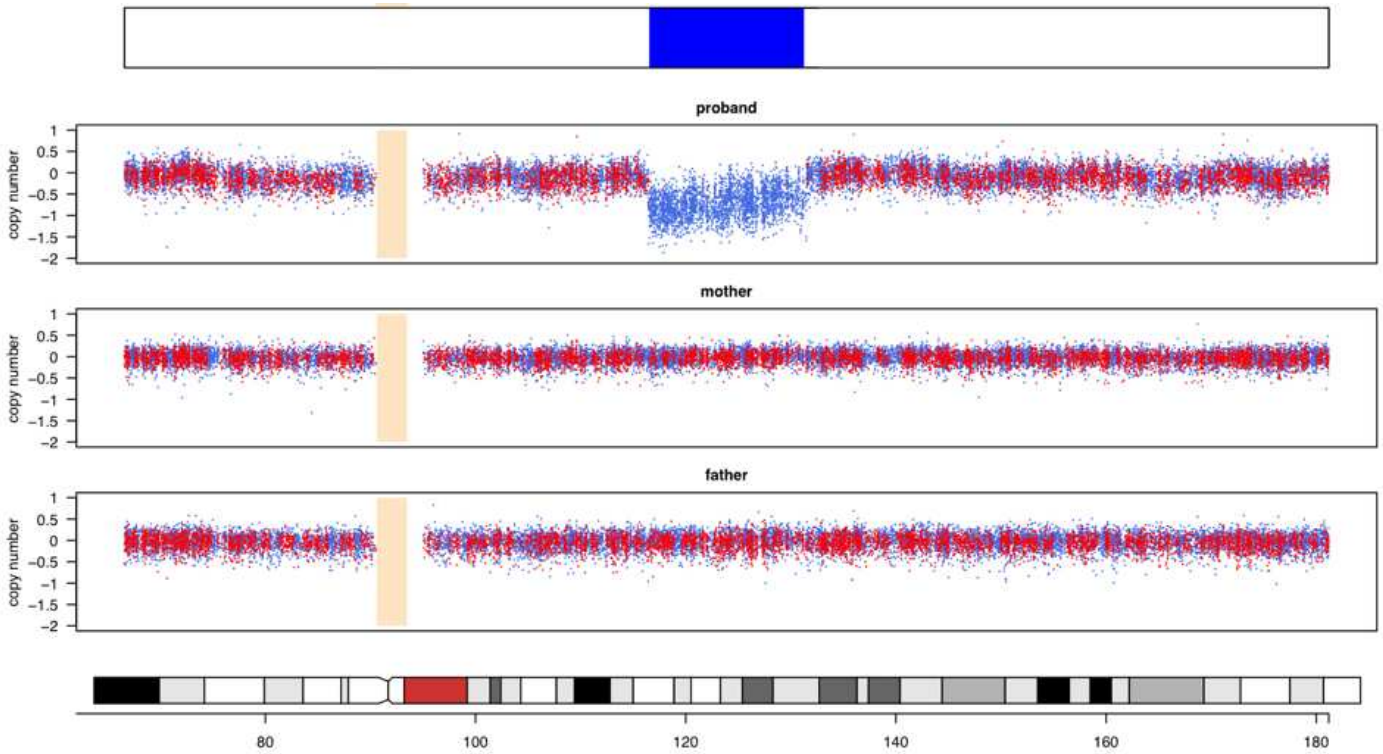


High uncertainty



A HMM for Trios

Idea: model the copy number and genotype estimates for the trio jointly to directly infer *de-novo* versus inherited regions of structural variation



Important points

Germline diseases

- Integer copy numbers (0, 1, 2, 3, . . .)
- Stochastic process well captured by HMM
- Analysis: Tier 1 $\rightarrow$ Tier 2 $\rightarrow$ Tier3

Somatic cell diseases

- Noninteger copy numbers (e.g., mosaicism)
- Stochastic process imposed by HMM not always appropriate
- Nonparametric techniques may be preferable
- Analysis: Tier 1 $\rightarrow$ Tier 3 ?

Tier 3: Between samples

How can we compare the genotypes and copy numbers of many subjects, infer common regions of alterations, and assess differences between affected subjects and normal controls?

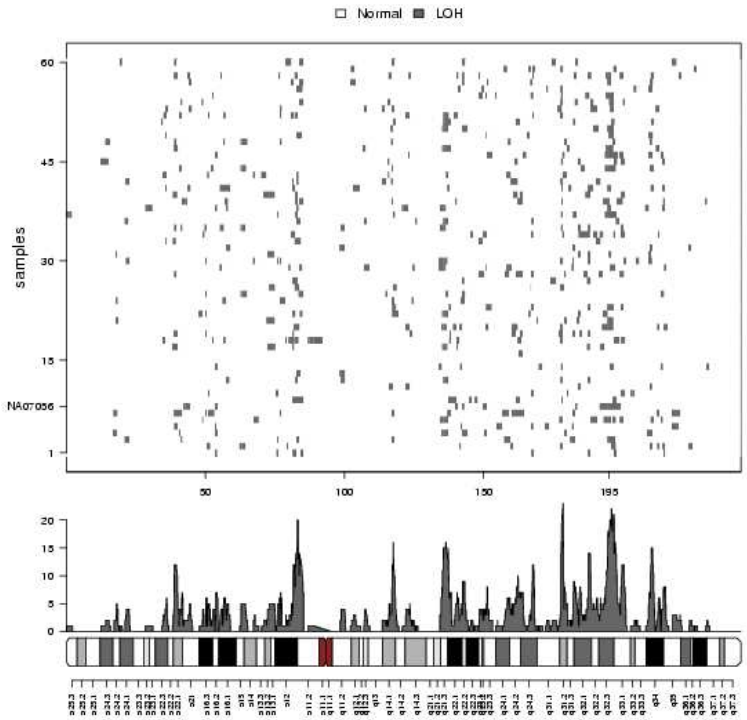
Considerations:

- Breakpoints of variants identified by HMM will differ across subjects
- For common variants, we can calculate sliding test statistics
- For diseases believed to be caused by multiple rare variants, inference for *de-novo* versus inherited variants will be important

Software

- R package CNVtools
  http://cnv-tools.sourceforge.net/CNVtools.html
  Barnes *et al.* (2008), Nature Genetics
- Plink http://pngu.mgh.harvard.edu/~purcell/plink/
  Purcell *et al.* (2007), American Journal of Human Genetics

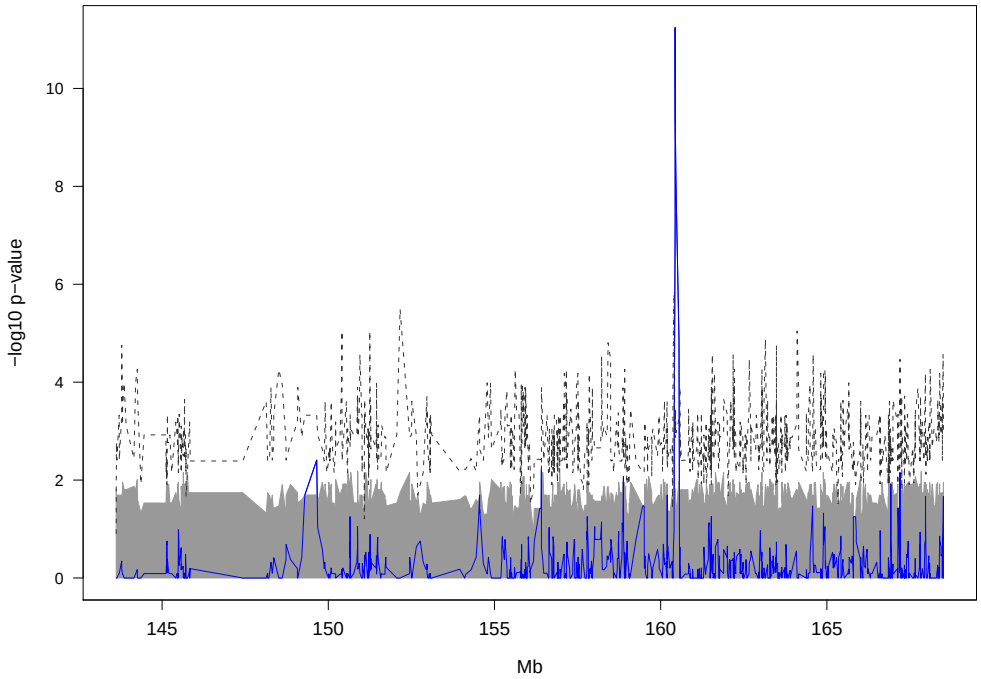
Loss of heterozygosity (LOH) in many subjects



Top: LOH regions identified by fitting a hidden Markov model.

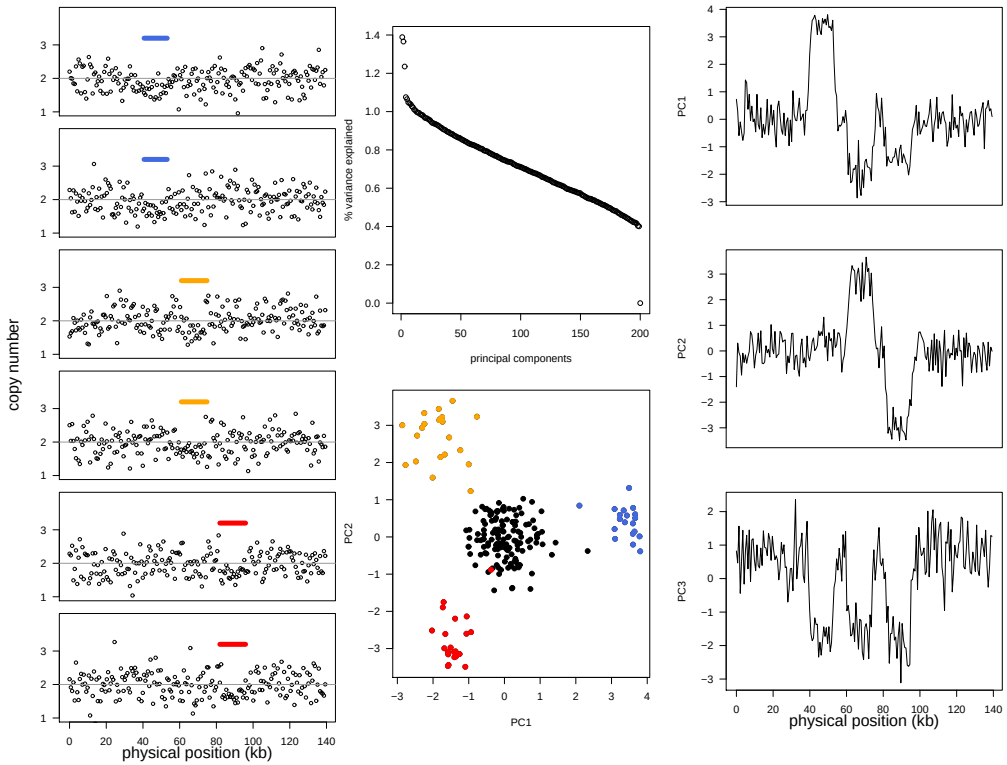
Bottom: frequency of LOH regions.

Statistical significance for common variants



Plotted in blue are the -log10 pvalues (y-axis) from a Fisher's exact test in a simulated dataset of 1000 cases and 1000 controls at each of 1000 loci (x-axis). A common variant that was in 10% of the cases and 1% of the controls was inserted at 161 Mb.

Principal components analysis (PCA)



Copy number for 450 controls and 450 cases were simulated. Left: 6 representative cases with 3 different types of deletions. Segmentation methods would smooth over the deletions (average across all samples 1-95)

Closing remarks

- Genotypes and copy number estimates that are plotted in Biology journals are heavily processed.
- Higher level analyses can be improved by propogating the uncertainty from lower level analyses
- Multiple tools are available for both CNV and genotyping. Ask for the rawest form of data whenever possible (CEL files for Affymetrix, raw X and raw Y intensities for Illumina)
- GWAS for copy number are starting to appear in the literature

Contributors and expertise

- Ingo Ruczinski: http://www.biostat.jhsph.edu/~iruczins
  copy number, hidden Markov models, genome-wide association methods
- Rafael Irizarry: http://rafalab.jhsph.edu
  copy number estimation, genotyping, Bioconductor developer
- Benilton Carvalho:
  http://www.biostat.jhsph.edu/people/student/carvalho.shtml
  genotyping, Bioconductor developer, R package: oligo
- Jonathan Pevsner: http://pevsnerlab.kennedykrieger.org/
  bioinformatics, SNP* web-based utilities, developmental disorders