Analysis of Oligonucleotide Single Nucleotide Polymorphism (SNP) Array Data

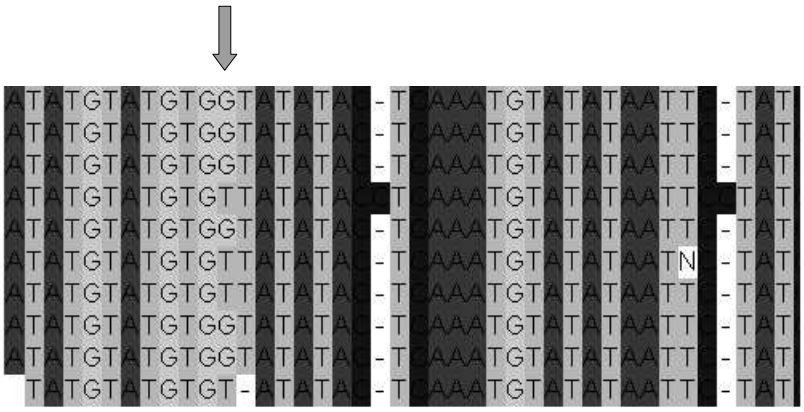
Cheng Li

2/12/04

Department of Biostatistical Science, Dana-Farber Cancer Institute
Department of Biostatistics, Harvard School of Public Health

SNP array technology

Single nucleotide polymorphisms (SNPs)

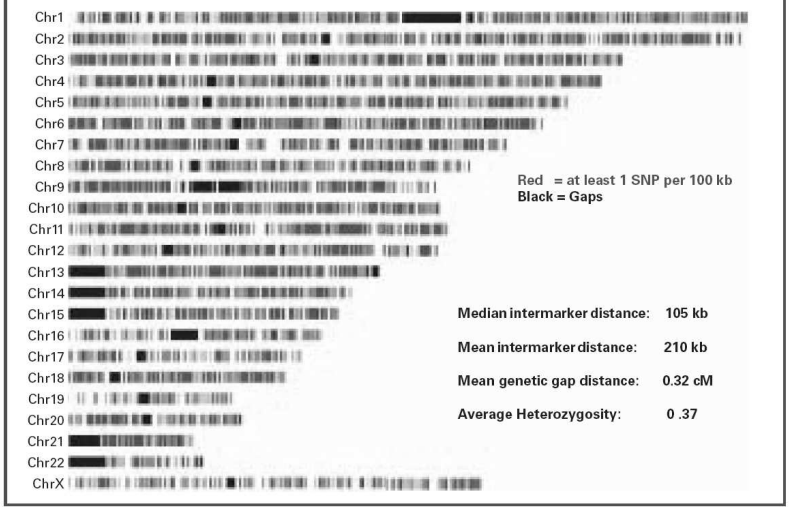


* Most common genetic variation in human genome. Occur every 1350 base pairs on average in genome

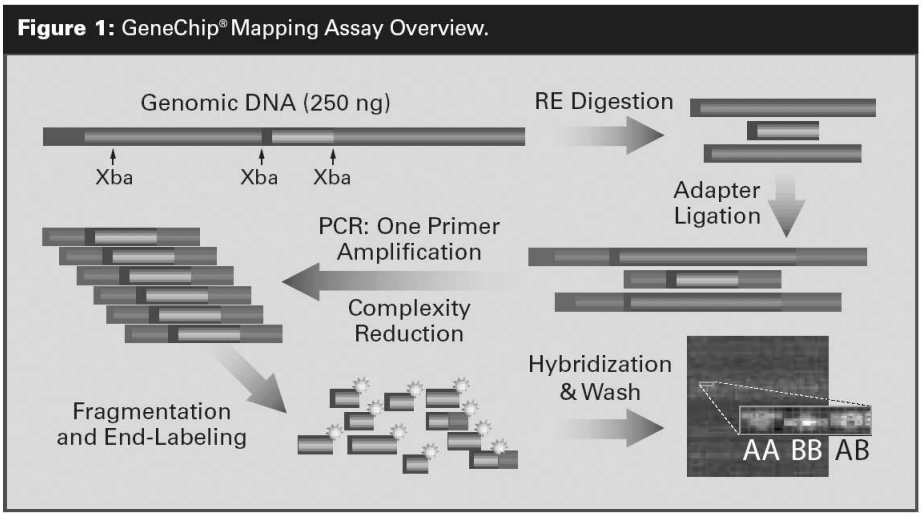
* Genome-wide SNP maps now available (millions in database)

GeneChip® Human Mapping 10K Array

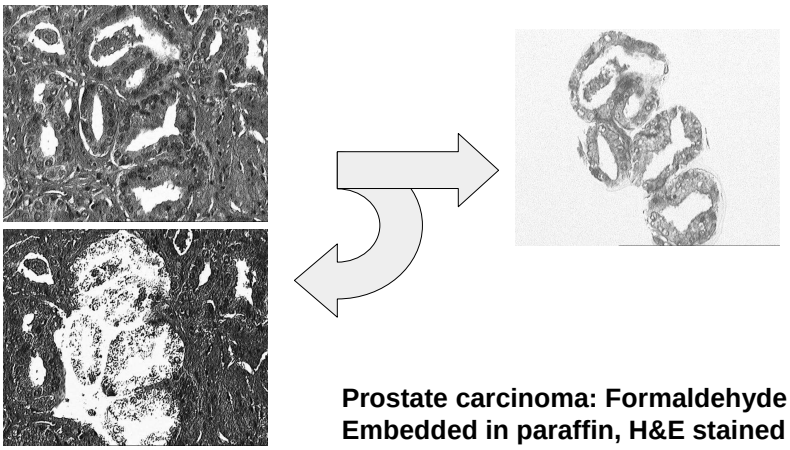
Figure 2: Genome Coverage of Mapping 10K SNPs by Chromosome. Black areas represent gaps in the human genome sequence, primarily centromeres and telomeres.



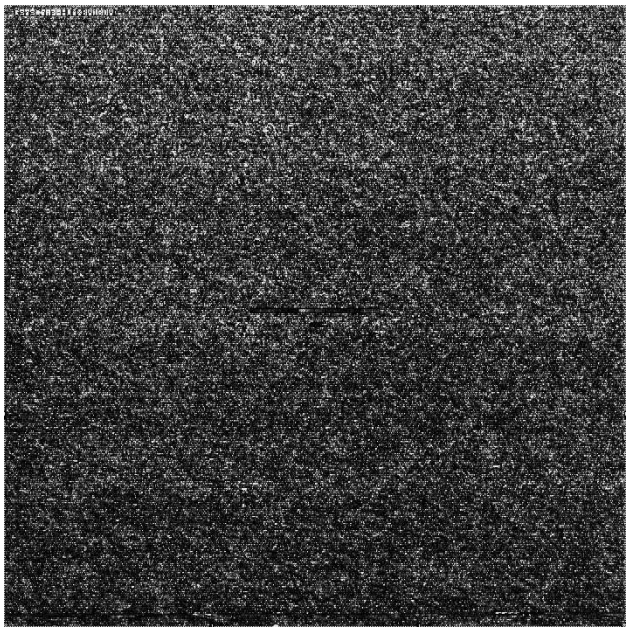
GeneChip® Human Mapping 10K Array

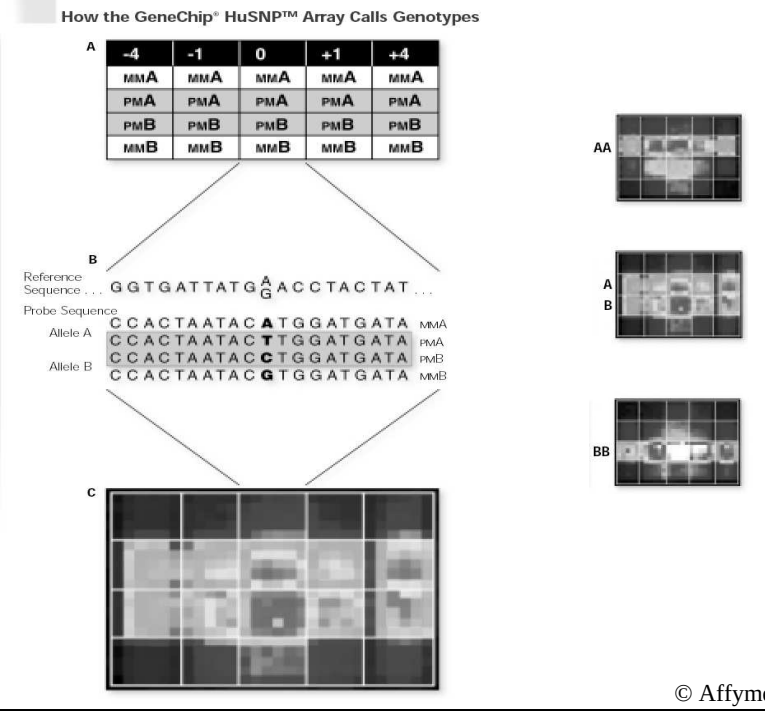


Laser Capture Microdissection

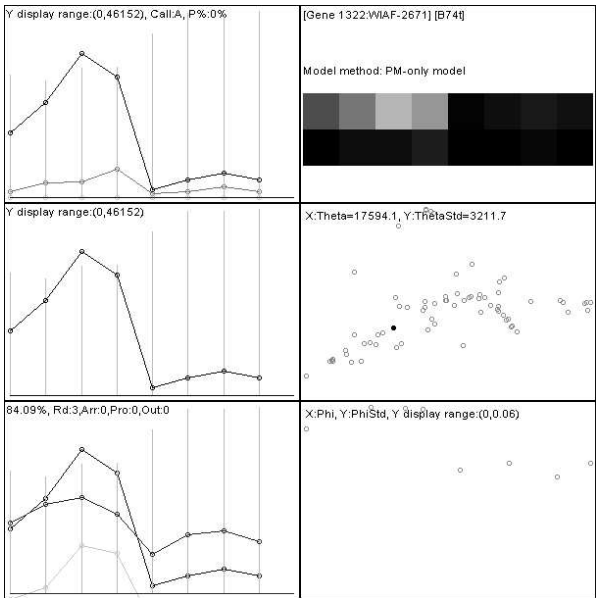


Prostate carcinoma: Formaldehyde fixed, Embedded in paraffin, H&E stained

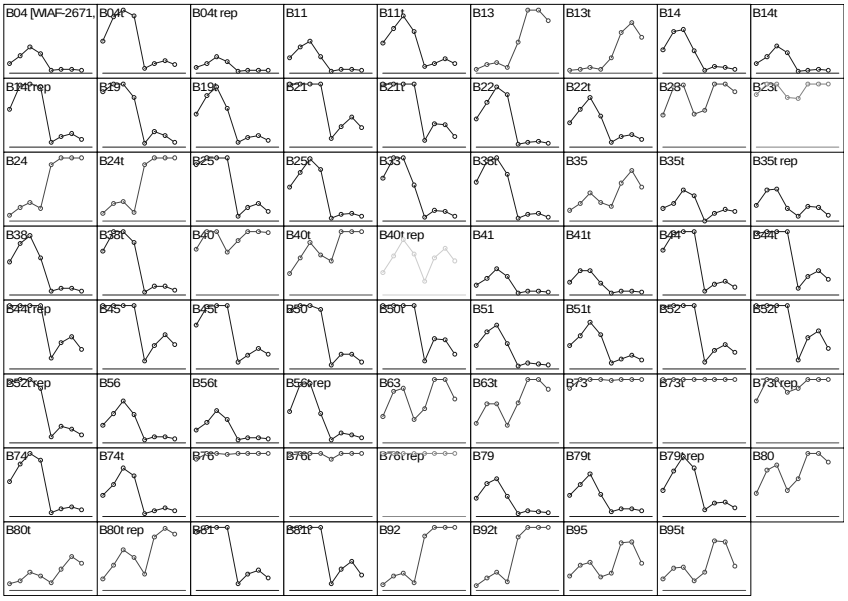




dChip PM/MM data view



dChip Variation View: Unsupervised clustering may recognize different PM patterns or SNP genotypes



WIAF-2671, Breast cancer data

dChip SNP view: project probe data of a SNP to 2D

- For each MiniBlock i = 1 … M, compute

  Diff_A = max (pmA – mmA, 1)

  R_i = Diff_A / (Diff_A + Diff_B)

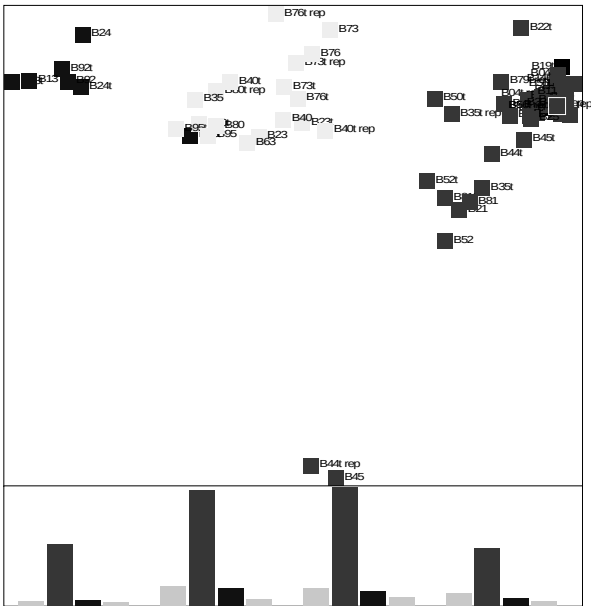
- The data of one SNP in one sample is (R_1, R_2 ,… R_M)
- Use principle component analysis (PCA) to project S data points (for S samples) into two dimension to visualize

Clustering SNP

AA

BB

AB



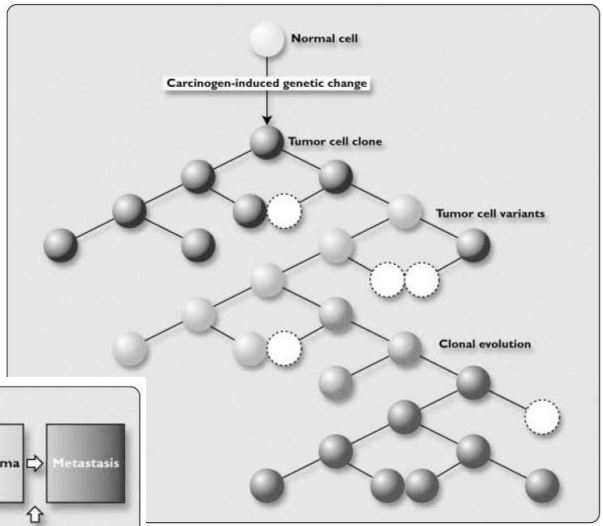
WIAF-2671
Breast cancer data

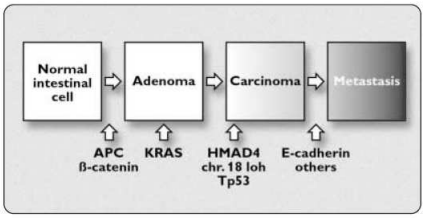
Loss of heterozygosity (LOH)

by SNP array

Genetic evolution model

Cancer is caused by deregulation of growth controlling molecular pathways.





Motivation

- Despite apparent locally confined disease, up to 30% of prostate cancer patients undergoing radical prostatectomy will develop recurrence.
- Initial Hypothesis: The differing clinical outcomes of prostate cancer arise from differentially expressed genes
- New hypothesis: Correlating gene expression data with specific genetic alterations may identify biologically relevant gene expression patterns.

Is there evidence that a genetic lesion can produce a global alteration in gene expression?

- BRCA

-Mixed Lineage Leukemia

- ALL, MLL, AML
- Upregulation of genes in MLL (FLT3)
- Expression profile classifies independent tumor set

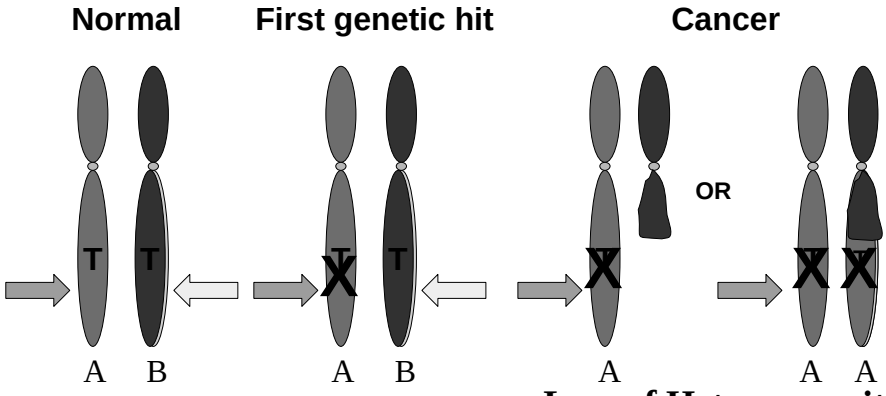
Chromosome Alterations

- Non-reciprocal translocations
- Aneuploidy
- Chromosomal amplifications
- Chromosomal deletions
  - loss of tumor suppressor via a point mutation followed by a deletion

Loss-of-Heterozygosity (LOH)

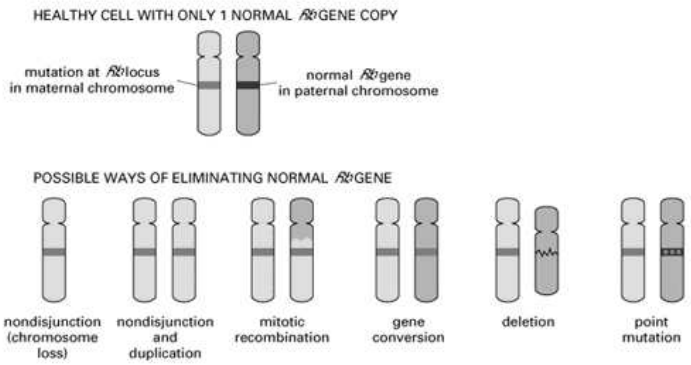
- If a marker (SNP, micro-satellite) has heterozygous genotype in the normal sample but has homozygous genotype in the tumor sample from the same patient.
- Indicates chromosomal alteration; often related to tumor-suppressor genes

Paradigm for Tumor Suppressor Gene Inactivation by Allelic Loss in Cancers



Loss of Heterozygosity (LOH)

Six ways of losing the remaining good copy of a tumor suppressor gene (Rb)



W.K. Cavenee et al., *Nature* 305:779-784, 1983.

Alberts et al. 1994 *Molecular Biology of the Cell*, 3rd ed.

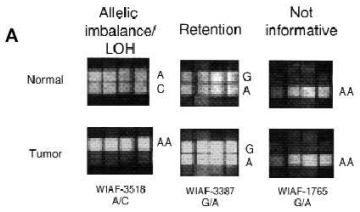
Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays

Kerstin Lindblad-Toh[1]*, David M. Tanenbaum[2–4]*, Mark J. Daly[1], Ellen Winchester[1], Weng-Onn Lui[5], Anuradha Villapakkam[1], Sasha E. Stanton[2], Catharina Larsson[5], Thomas J. Hudson[1,6], Bruce E. Johnson[2,3], Eric S. Lander[1,7] and Matthew Meyerson[2,4]

[1]Whitehead Institute/MIT Center for Genome Research, Whitehead Institute for Biomedical Research, Cambridge, MA 02139. [2]Department of Adult Oncology, Dana-Farber Cancer Institute, Boston, MA. [3]Departments of Medicine and [4]Pathology, Harvard Medical School, Boston, MA. [5]Department of Molecular Medicine, CMM, Karolinska Hospital, Stockholm, S-171 76 Sweden. [6]Montreal Genome Centre, McGill University Health Centre, Montréal, Québec. [7]Department of Biology, Massachusetts Institute of

the SNP array, stained with streptavidin–phycoerythrin, and assayed by fluorescence detection. The principles underlying genotyping with SNP arrays were described in an earlier study[4]. Briefly, the detector for each SNP locus contains four rows of 25-mer oligonucleotides, two of which contain oligonucleotides that perfectly match either SNP allele A or SNP allele B, whereas the other two contain single-base mismatches at various positions. The allelotype at a locus is determined by fluorescence intensity ratios in an automated fashion. The approach dramatically decreases the work involved in assaying 1,500 loci, as well as the amount of DNA required (to a total of only 120 ng DNA, corresponding to ~20,000 diploid human genomes) in comparison to both SSLPs and CGH.

The call rate (the proportion of loci to which genotypes could be assigned) was 80.7% ± 3.0% over all samples, yielding ~1,205 SNPs scored per sample (Table 1). The rate did not differ between normal and tumor samples. Many SNPs performed in a robust fashion,



Lindblad-Toh et al. *Nature Biotechnology* 2000

Compare normal and tumor samples

Uninvolved SV (snap frozen) Normal

↓

Genomic DNA Purification

↘

paraffin-embedded tumor

↓

Laser-Capture Microdissection

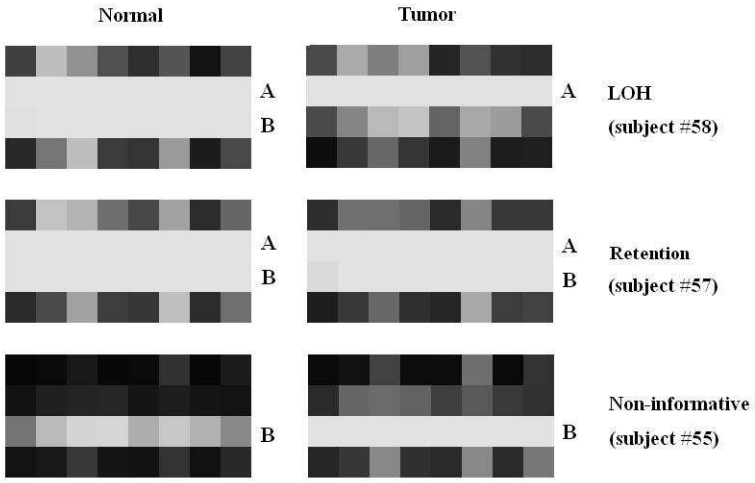
↓

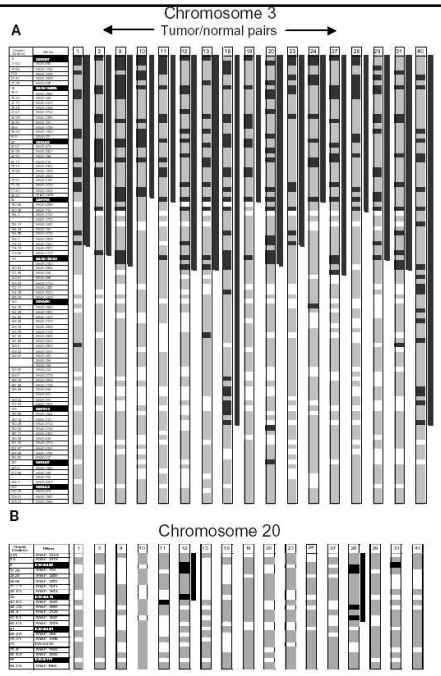
Genomic DNA Purification

↙

SNP array analysis

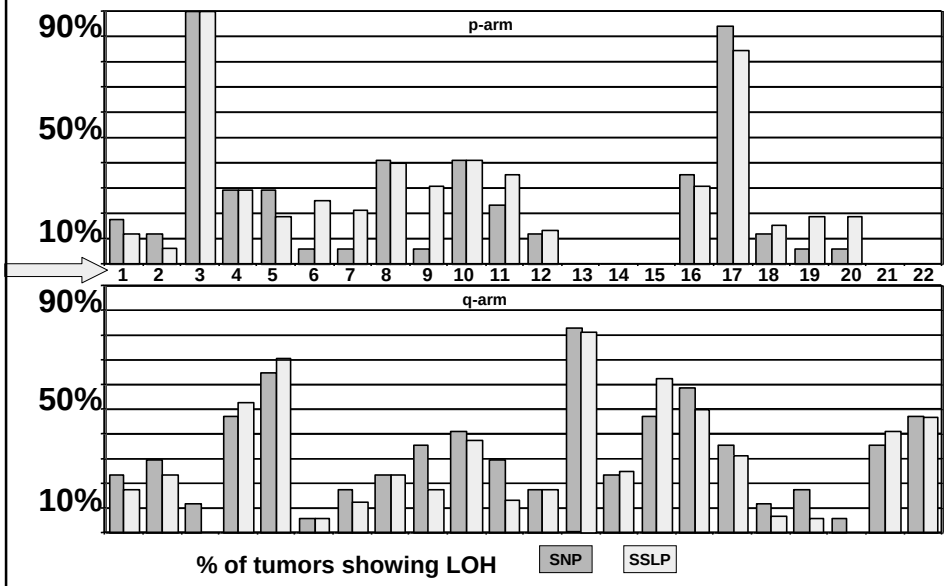
Making LOH calls





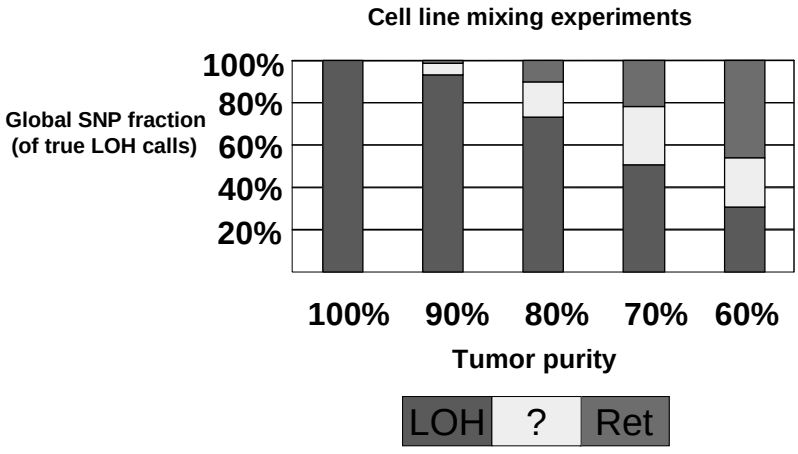
Lindblad-Toh et al., 2000

SNP and SSLP-based Analyses Give Comparable Results

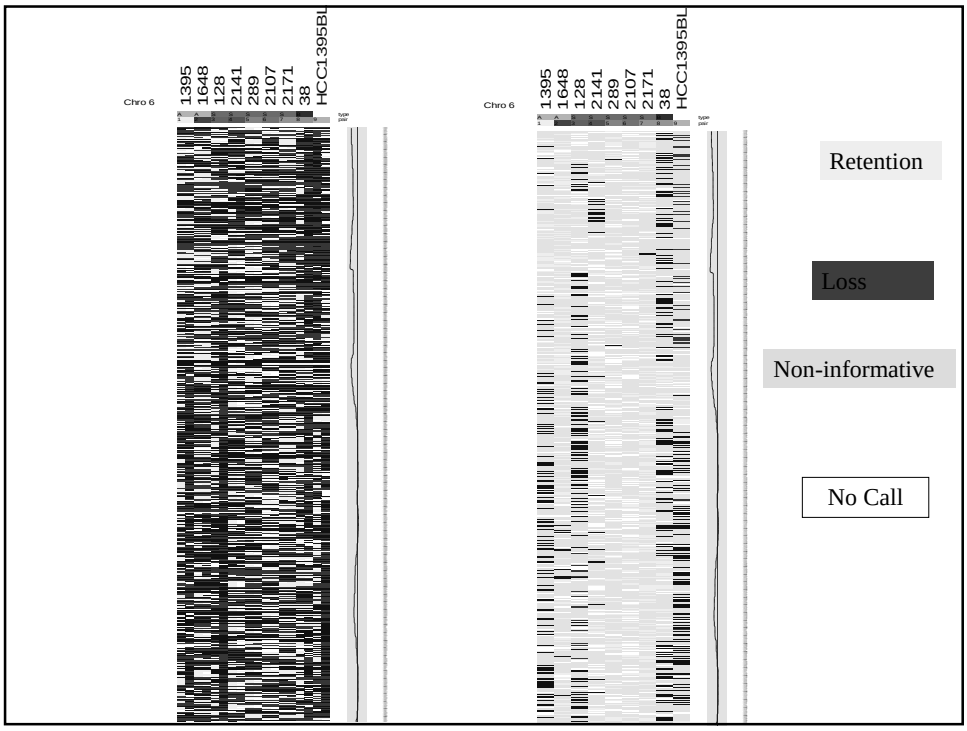


Lindblad-Toh et al., 2000

Global Effects of Purity on SNP-based LOH Analysis

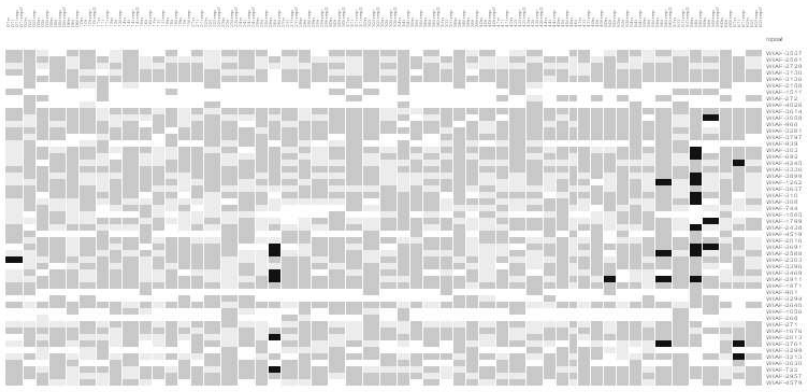


Lindblad-Toh et al., 2000



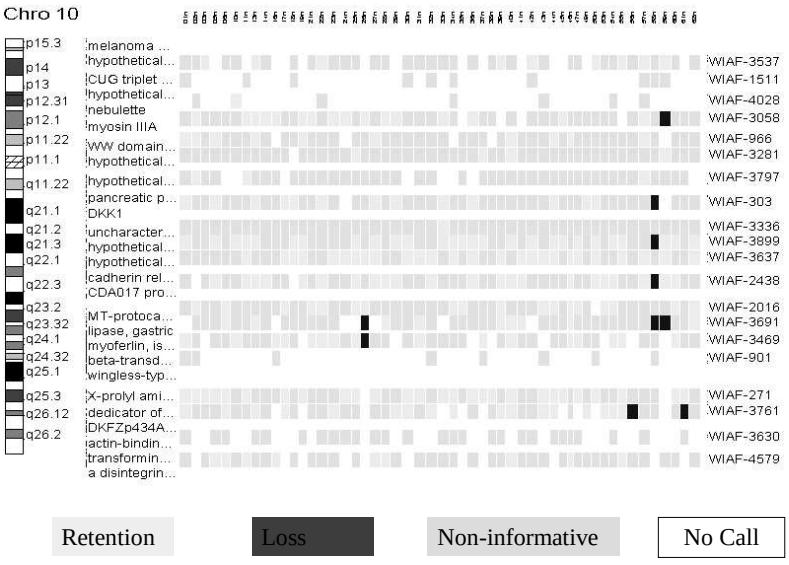
SNP Markers in chromosome 10

Observed LOH events

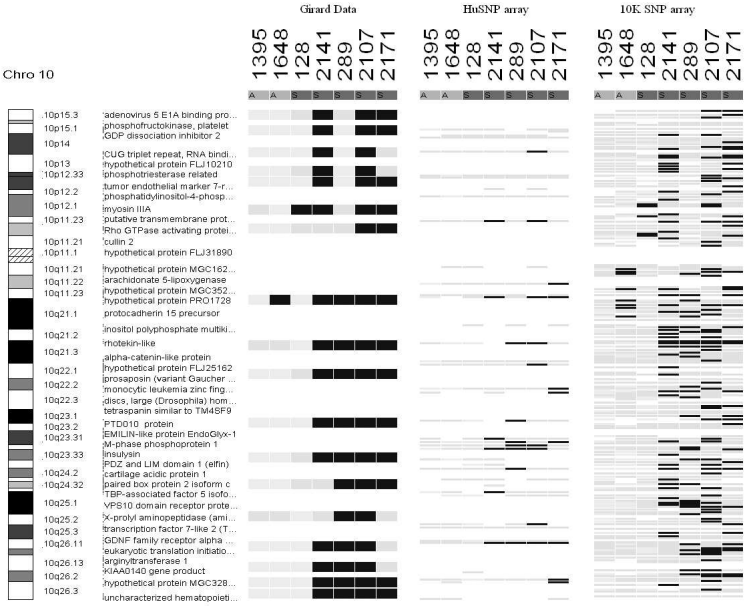


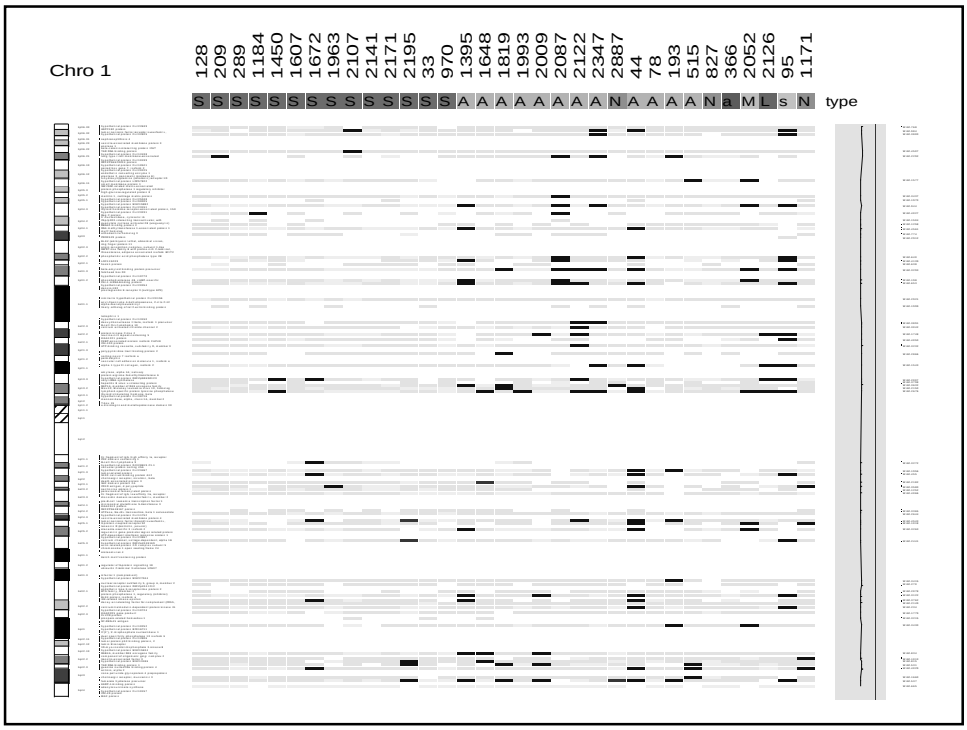
Retention | Loss | Non-informative | No Call

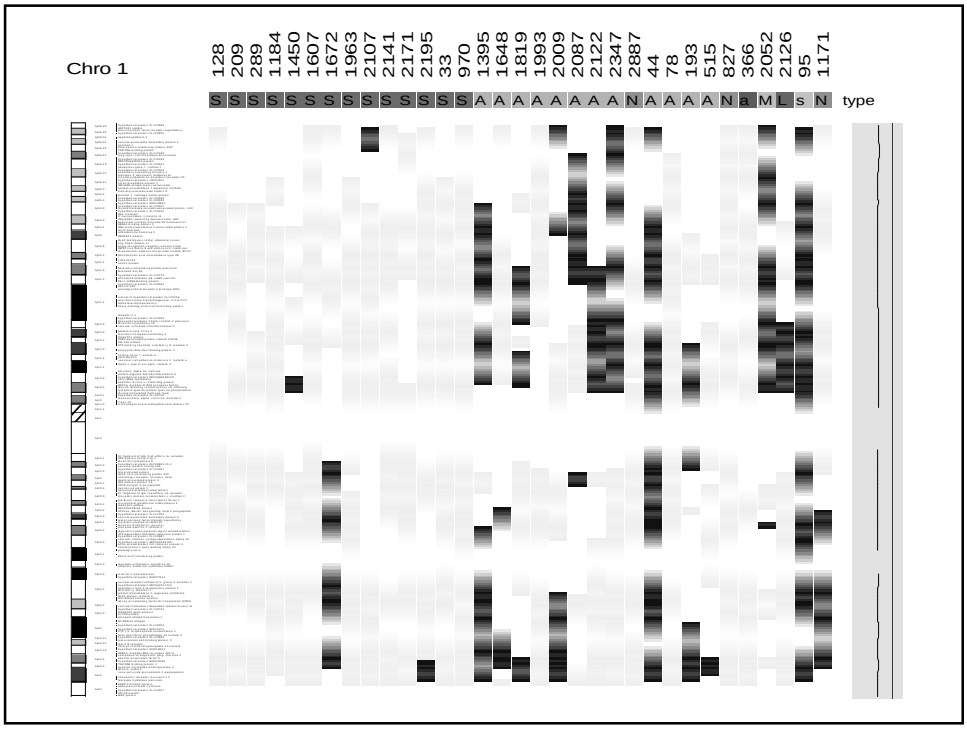
dChip Chromosome View with proportional distance



Compare LOH based on SSLP, HuSNP and 10K SNP

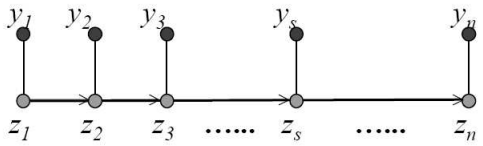



Chro 1
128 209 289 1184 1450 1607 1672 1963 2107 2141 2171 2195 33 970 1395 1648 1819 1993 2009 2087 2122 2347 2887 44 78 193 515 827 366 2052 2126 95 1171
S S S S S S S S S S S S S S A A A A A A A A N A A A A N a M L s N
type


Chro 1
128 209 289 1184 1450 1607 1672 1963 2107 2141 2171 2195 33 970 1395 1648 1819 2009 2087 2122 2347 2887 44 78 193 515 827 366 2052 2126 95 1171
S S S S S S S S S S S S S S A A A A A A A A N A A A A N a M L s N
type

Hidden Markov Models

- Markov: $x_1 \to x_2 \to \cdots \to x_n$
- HMM:



  - We only observe the y's;
  - The z's governed by a Markov transition rule

$$p_{ij} = P(z_t = j \mid z_{t-1} = i)$$

  - The y can be viewed as a "noisy copy" of the z

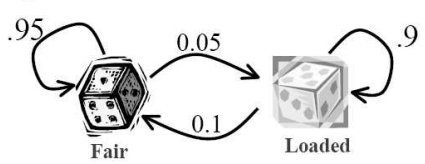
$$y_t \mid z_t \sim f_t(\cdot \mid z_t)$$

  - Example: z_i indicates intron or exon; y_i observed sequence

An example

- Occasionally dishonest casino

1:	1/6
2:	1/6
3:	1/6
4:	1/6
5:	1/6
6:	1/6



1:	1/10
2:	1/10
3:	1/10
4:	1/10
5:	1/10
6:	5/10

Simulted a sequence of 100 "die usages"

1 1 1 1 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 1 6 2 1 6 6 5 6 6 6 3 5 2 3 2 1 2 6 4 6 2 2 5 3 3 3 1 4 3 1 5 1 3 6 1 6 3 5 1 6 3 1 2 3 1 4 6 3 6

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2
5 1 3 3 5 6 1 3 5 5 4 6 3 2 4 1 6 2 5 4 2 4 4 2 1 2 3 2 6 3 6 6 6 4 5 6 2 2 4 6 6 1 4 6 3 4 2 6 4 6

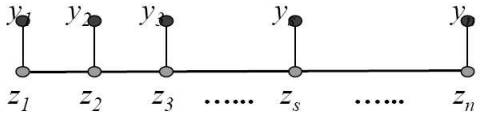
Question: what's the long-run fraction of each type?

However, we only observe the result of the rolls of a die: can we tell?

How do we find the most probable path?

- Optimal probability of the observed rolls $y_1 y_2 \cdots y_n$

$$f_i(z) = P(y_i \mid z_i = z) \times \max_{z'} \{ f_{i-1}(z') P(z_i = z \mid z_{i-1} = z') \}$$

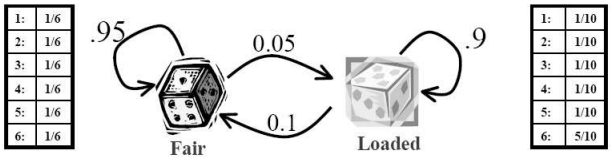


$f_i(z)$ is the optimal prob value for the first i observations with z_i =z.

- Tracing backward to find the optimal path

$$\hat{z}_i = \arg\max_z \{ f_i(z) P(z_{i+1} = \hat{z}_{i+1} \mid z_i = z) \}$$

Hidden Markov Model

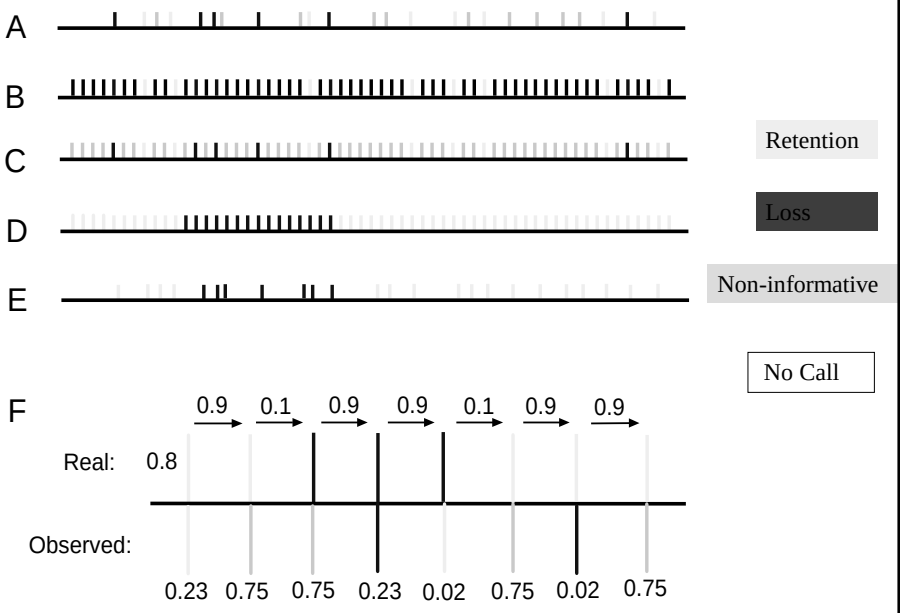


Simulated a sequence of 100 die tosses

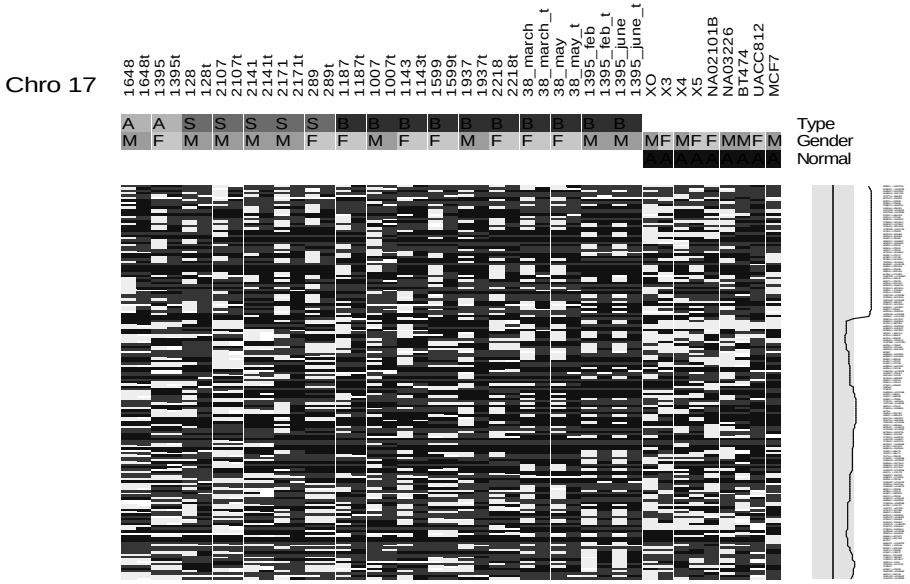
2 1 6 2 1 6 6 5 6 6 6 3 5 2 3 2 1 2 6 4 6 2 2 5 3 3 3 1 4 3 1 5 1 3 6 1 6 3 5 1 6 3 1 2 3 1 4 6 3 6
2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

5 1 3 3 5 6 1 3 5 5 4 6 3 2 4 1 6 2 5 4 2 4 4 2 1 2 3 2 6 3 6 6 6 4 5 6 2 2 4 6 6 1 4 6 3 4 2 6 4 6
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2

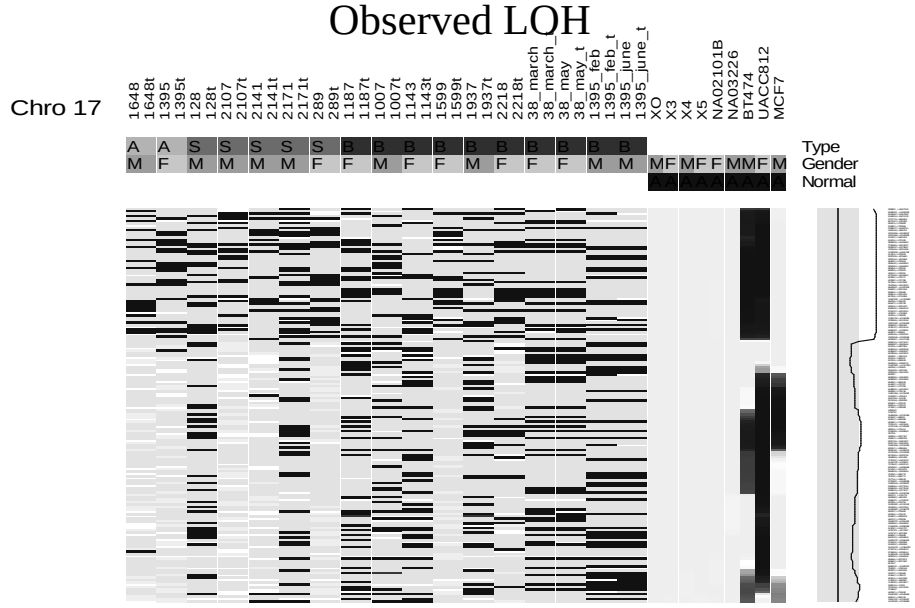
Hidden Markov Model to infer non-informative markers



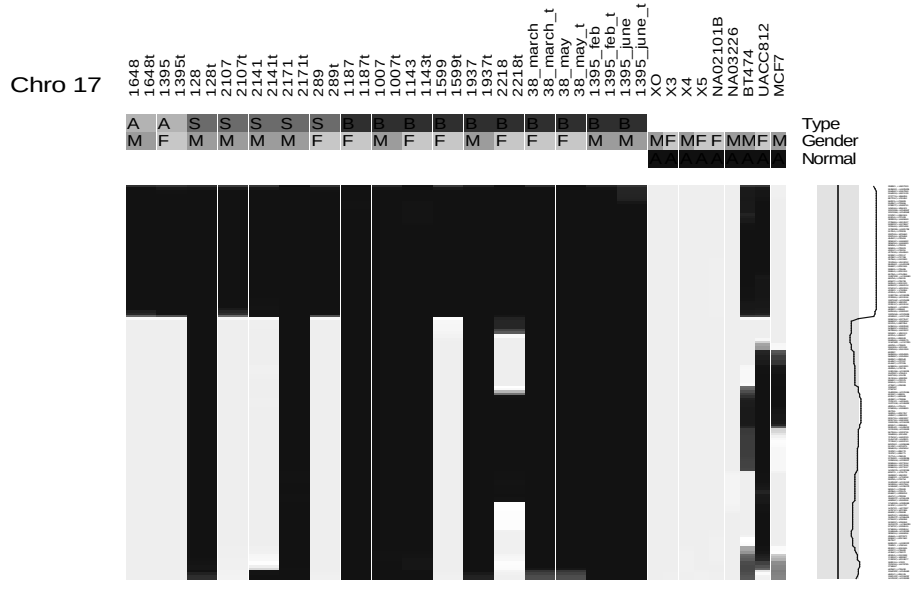
SNP call data



Observed LOH

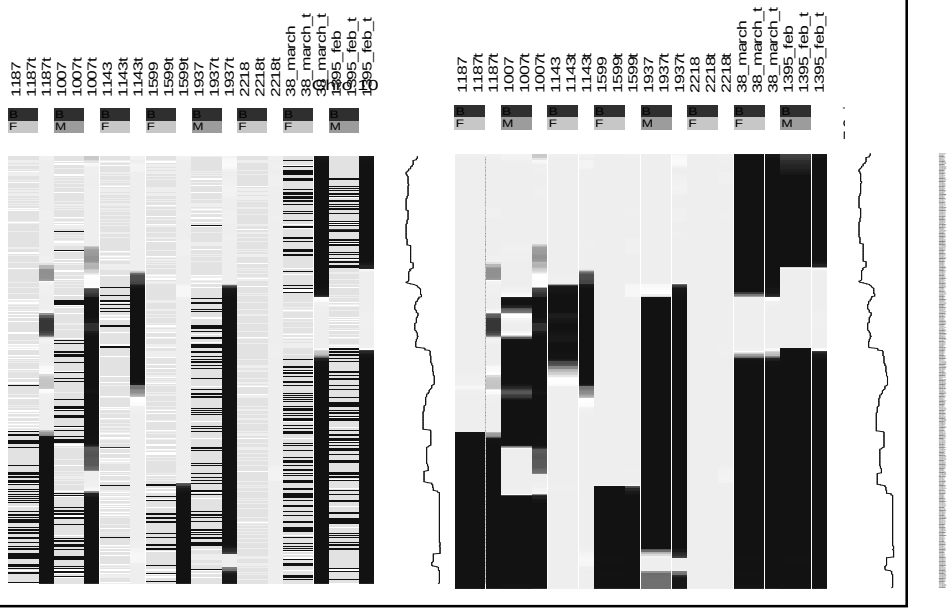


Inferred LOH

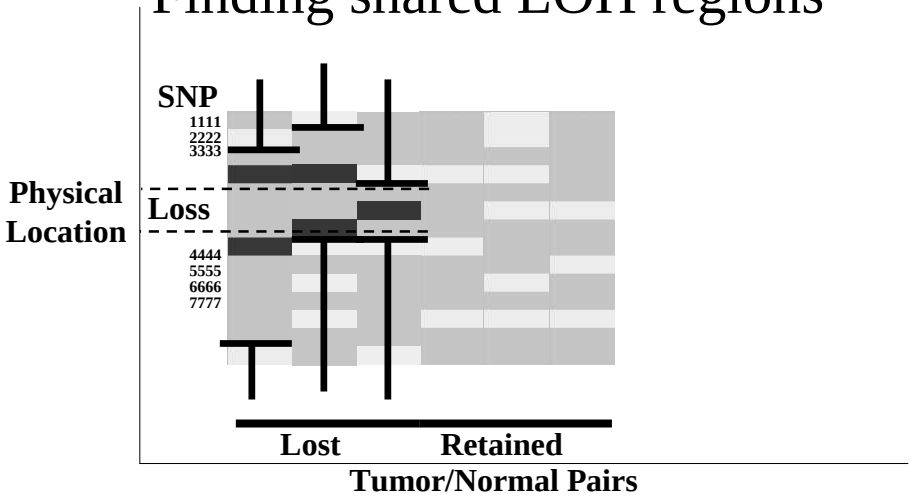


Comparing LOH obtained from paired normal and tumor samples and only tumor samples

Chro 10



Finding shared LOH regions



The SNP markers are not directly useful as single variables

Finding shared LOH regions

- Shared LOH regions might contain tumor suppressor genes
- Complication:

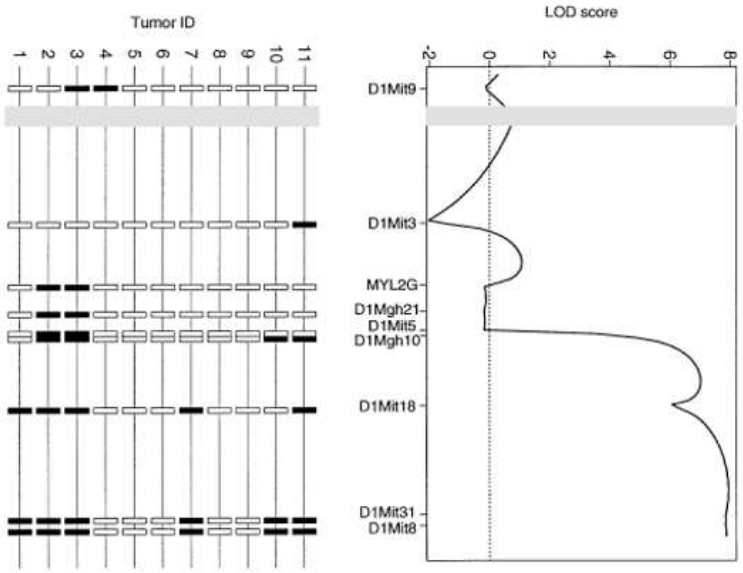
  1. Markers are 300Kb apart, and many of them are non-informative

  2. Call errors, mapping errors

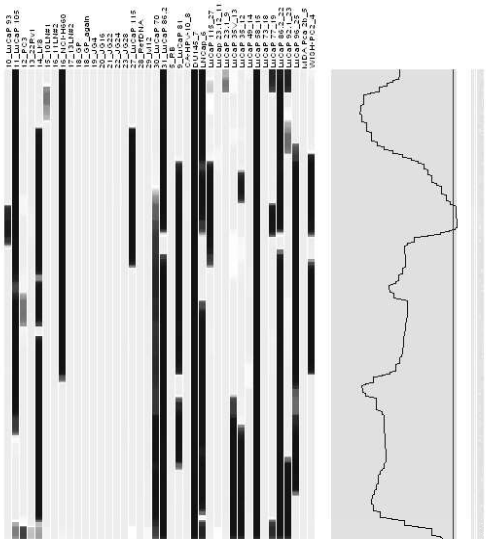
  3. Observed LOH events may be due to genetic instability of the tumor and are not cancer-related.

LOH LOD score

M. Newton et al. 1998 *Statist. Med*

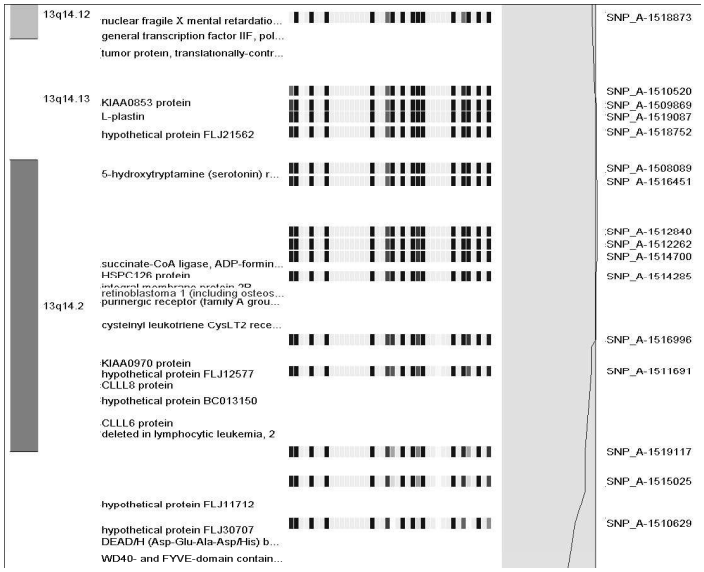


LOH score



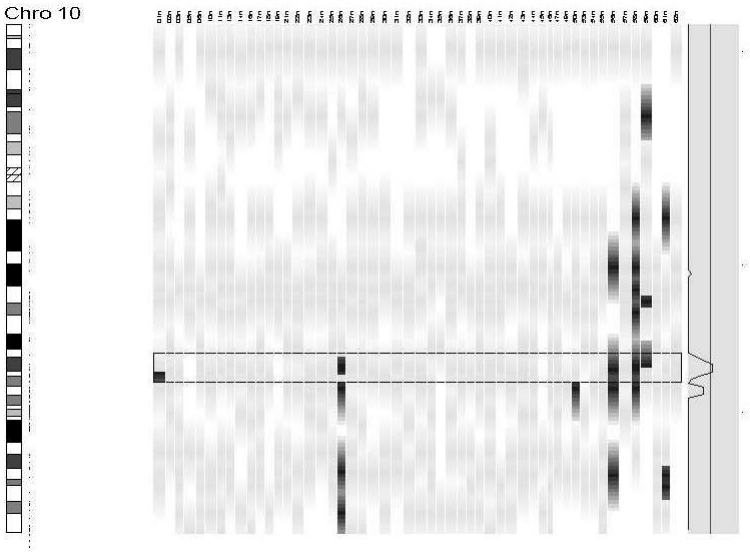
The inferred LOH status of 40 samples in chromosome 13, with LOD score curve on the right (gray area from –2 to 5)

LOH LOD score



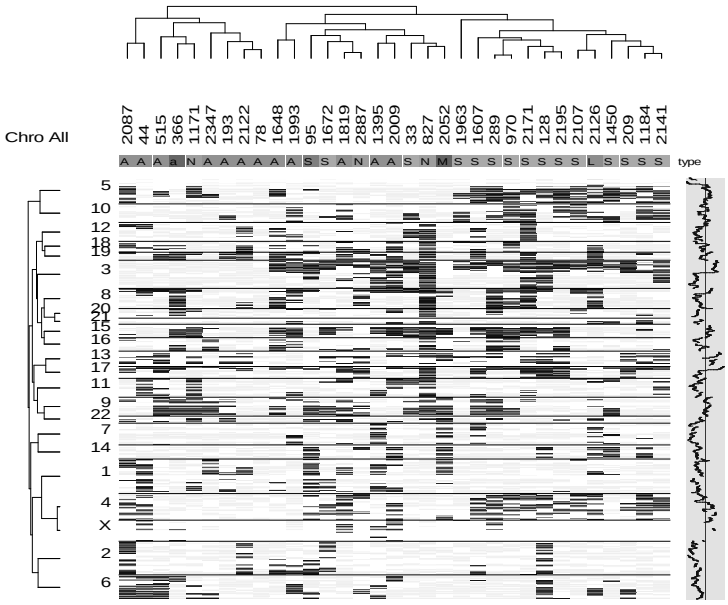
The 3 Mb area with peak curve value contains the RB gene.

Permutation to compute p-value of cancer-relatedness

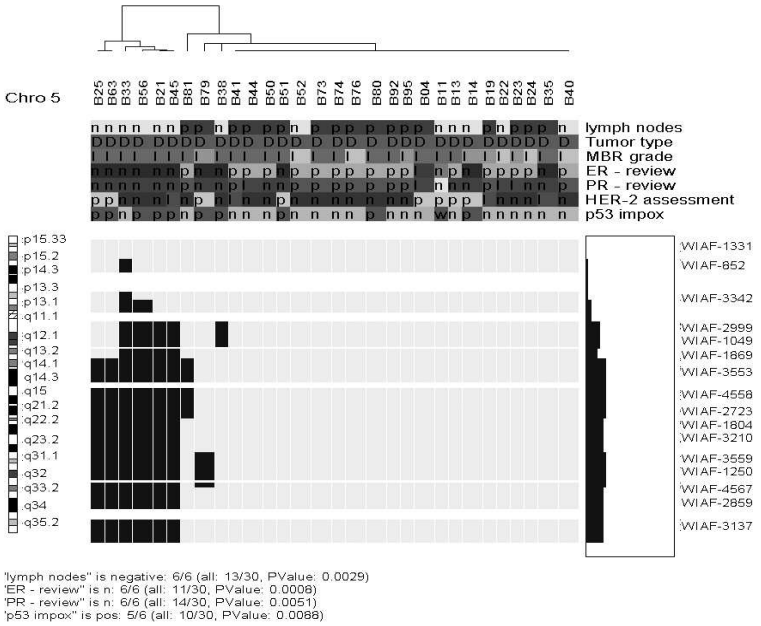


M. Lin et al. *Bioinformatics*, in press

Clustering samples and chromosomes

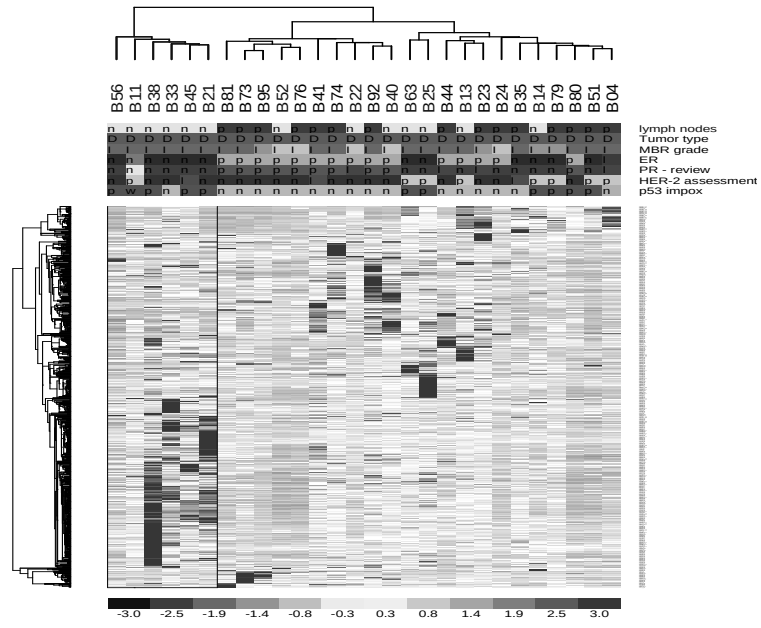


Correlate LOH data with expression data

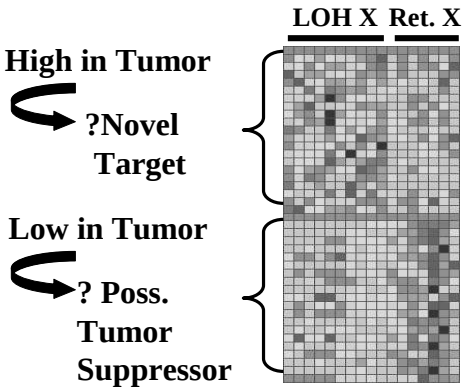


"lymph nodes" is negative: 6/6 (all: 13/30, PValue: 0.0029)
"ER - review" is n: 6/6 (all: 11/30, PValue: 0.0008)
"PR - review" is n: 6/6 (all: 14/30, PValue: 0.0051)
"p53 impox" is pos: 5/6 (all: 10/30, PValue: 0.0088)

Correlate LOH data with expression data

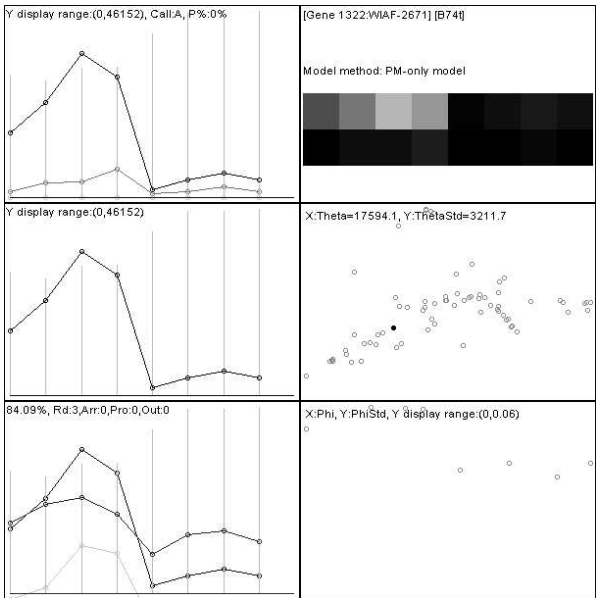


Applying supervised clustering using an LOH as a class distinction

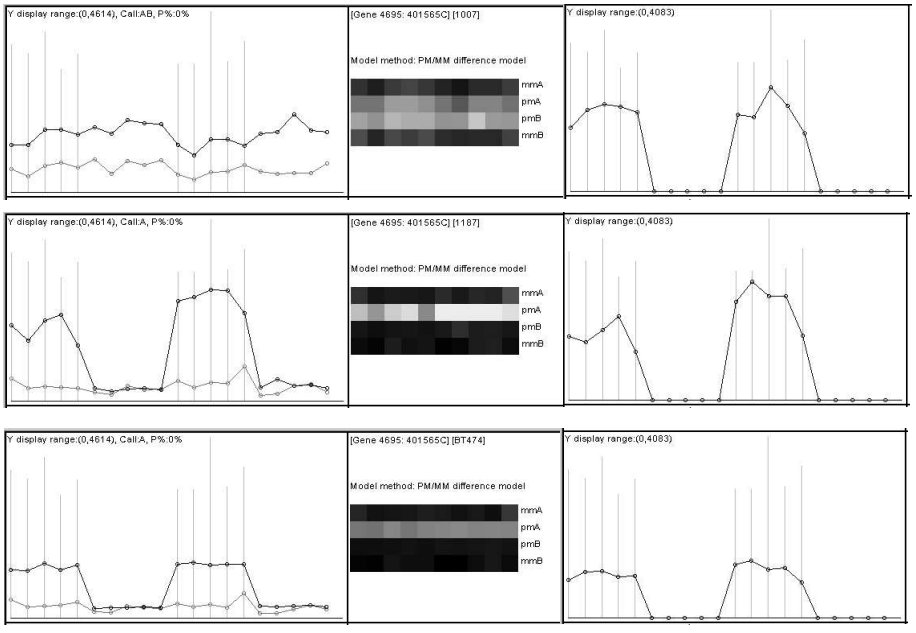


Array CGH by SNP array

dChip PM/MM data view



Making signal of AB and AA comparable

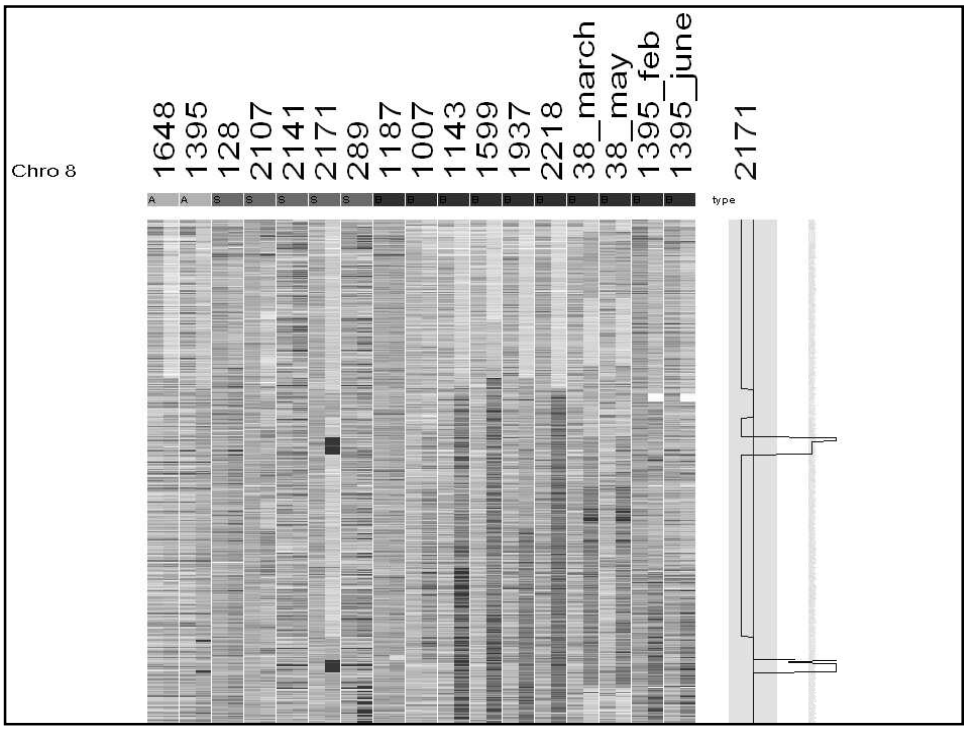


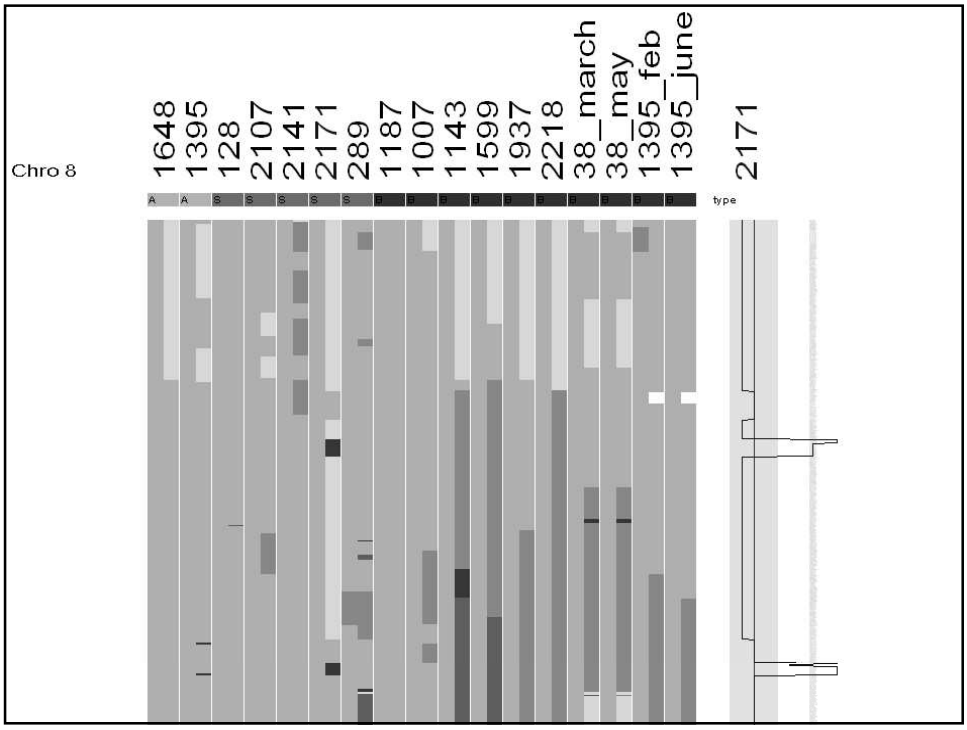
Copy number analysis using SNP array

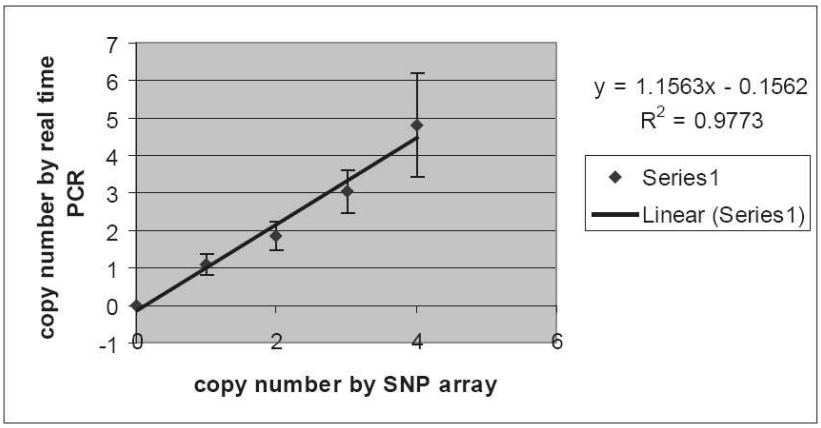
- Normalization and model-based signal for each array and SNP
- For a SNP, the signal values of all normal cell lines were averaged to obtain the mean signal of 2 copy; observed copy number = (observed signal / mean signal of two copy) * 2
- HMM to infer real copy number by best path

Copy number analysis using SNP array

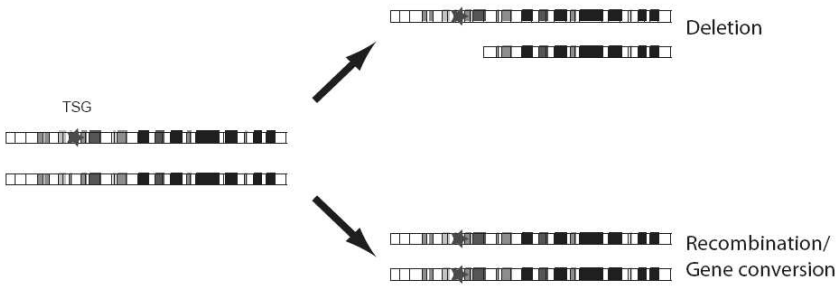
- Emission probabilities: for a SNP the observed signal values are random values drawn from $(\frac{Signal - Mean \bullet Fold}{Std \bullet Fold}) \sim t(40)$
- Transition probabilities: The Haldane's map function $\theta = \frac{1}{2}(1 - e^{-2d})$ convert the genetic distance *d* between two SNP markers to the probability (*2*theta*) that the copy number of the 2nd marker will return to the background distribution of copy numbers in this sample and thus independent from the copy number of the 1st marker
- Initial probabilties: The proportion of chromosome regions that have a particular copy number is set to fixed values in the first round (0.9 for 2 copy, 0.1/(N-1) for copy 0 to MaxCopy except 2)

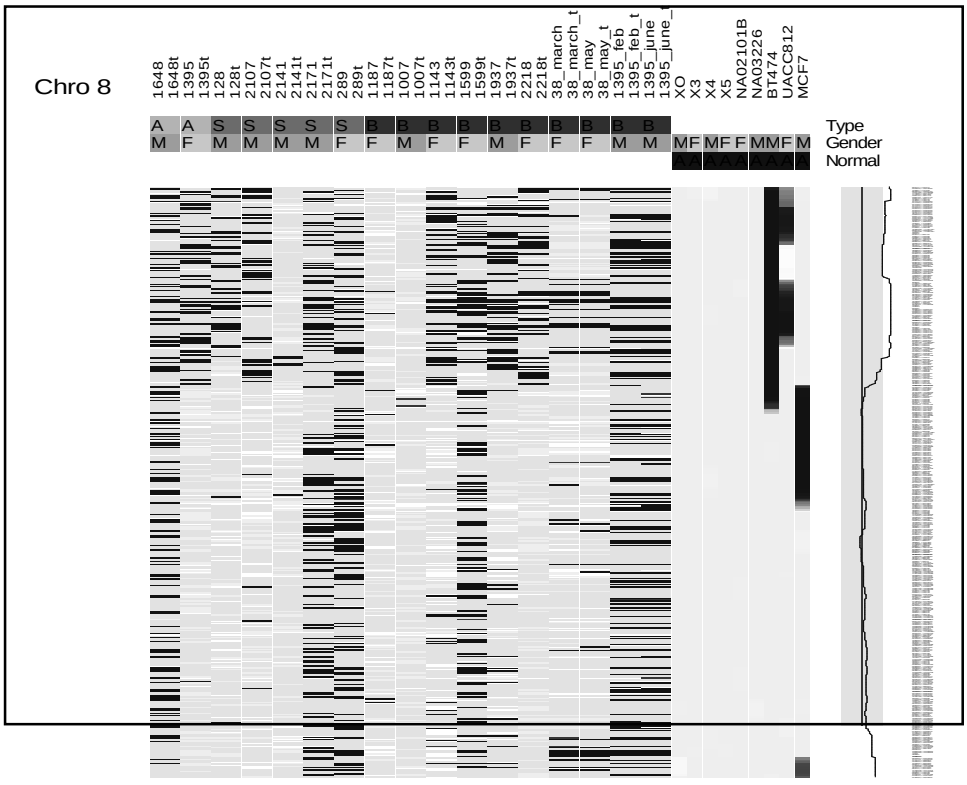

Chro 8
1648
1395
128
2107
2141
2171
289
1187
1007
1143
1599
1937
2218
38_march
38_may
1395_feb
1395_june
2171
type

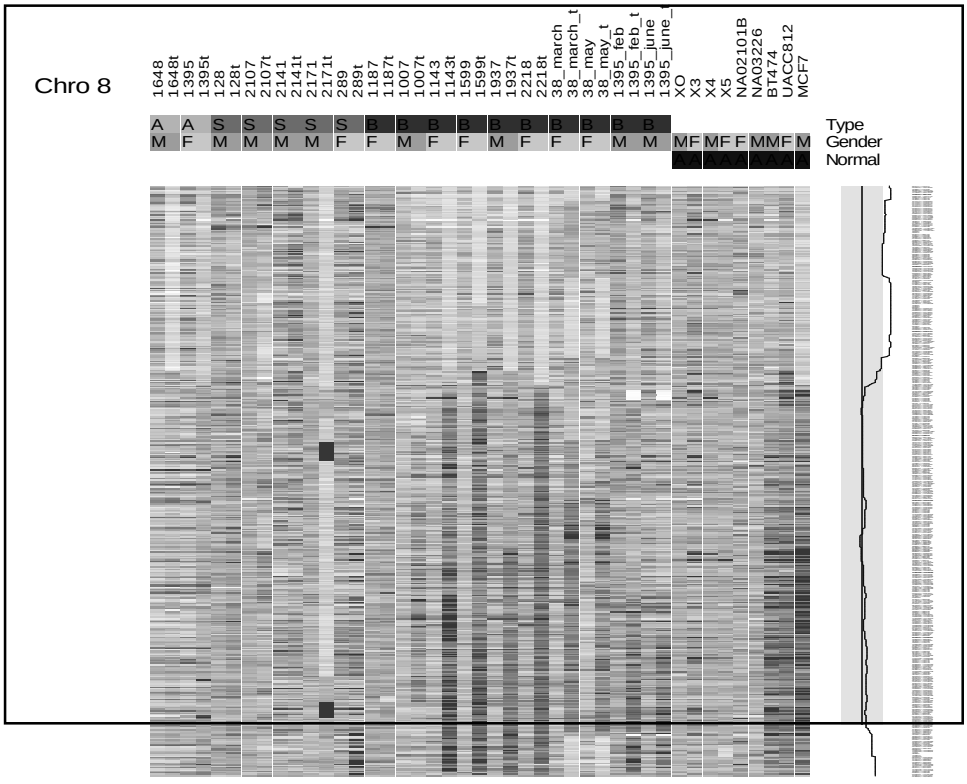

Chro 8
1648
1395
128
2107
2141
2171
289
1187
1007
1143
1599
1937
2218
38_march
38_may
1395_feb
1395_june
2171
type



Two mechanisms of LOH




Chro 8
1648
1648t
1395
1395t
128
128t
2107
2107t
2141
2141t
2171
2171t
289
289t
1187
1187t
1007
1007t
1143
1143t
1599
1599t
1937
1937t
2218
2218t
38_march
38_march_t
38_may
38_may_t
1395_feb
1395_feb_t
1395_june
1395_june_t
XO
X3
X4
X5
NA02101B
NA03226
BT474
UACC812
MCF7
A M
A F
S M
S M
S M
S M
S F
B F
B M
B F
B F
B M
B F
B F
B F
B M
B M
MF
MF
F
MM
F
M
Type
Gender
Normal


Chro 8
1648
1648t
1395
1395t
128
128t
2107
2107t
2141
2141t
2171
2171t
289
289t
1187
1187t
1007
1007t
1143
1143t
1599
1599t
1937
1937t
2218
2218t
38_march
38_march_t
38_may
38_may_t
1395_feb
1395_feb_t
1395_june
1395_june_t
XO
X3
X4
X5
NA02101B
NA03226
BT474
UACC812
MCF7
A M
A F
S M
S M
S M
S M
S F
B F
B M
B F
B F
B M
B F
B F
B F
B M
B M
MF
MF
F
MM
F
M
Type
Gender
Normal

Acknowledgment

Harvard School of Public Health

Ming Lin

Ke Hao

Lee-Jen Wei

Wing Hung Wong

Carsten Rosenow (Affymetrix)

Richard Smith (U. of Iowa)

Dana-Farber Cancer Institute

Matthew Meyerson
William R. Sellers

Pasi Antero Janne
Marshall Lieberfarb
Andrea Richardson
Zhigang C. Wang
Edward Fox
Xiaojun Zhao

Cheng Li: cli@hsph.harvard.edu