

The background of the slide is a photograph of a city square, likely in Padua, Italy. In the foreground, there is a large, ornate fountain with multiple tiers and statues. In the background, a large church with a prominent dome and a tall bell tower is visible. The sky is overcast and grey.

Analysis of gene expression profiles at chromosomal level

Andrea Callegaro, Silvio Bicciato

Department of Chemical Process Engineering
University of Padova

Motivations

□ **Integration**

□ Among others, integrating:

- ↳ **transcriptional profiles**
- ↳ **other types of gene information**, such as
 - locus
 - sequence
 - genome structure
- ↳ **DNA copy number alterations**

Motivations

□ In particular, integration between expression profiles and locus information could be effective in:

↳ **relating structural abnormalities** (e.g., allelic imbalances) **to physiological states**

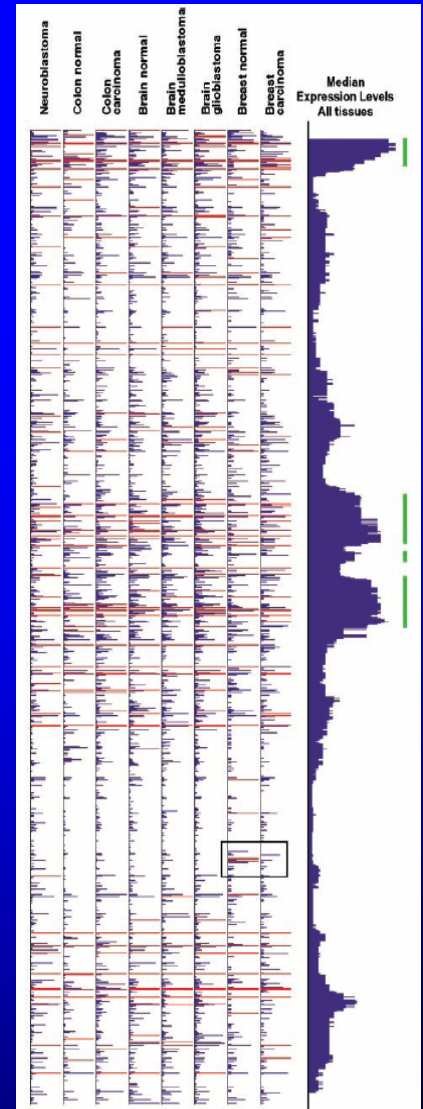
- there are evidences of relationship between genomic structural abnormalities and ***regional* expression imbalances** in malignant phenotypes

↳ **unveiling the transcriptional machinery** (*active chromatin domains*)

- transcriptional machinery may access co-expressed genes more efficiently if they are neighbors than if they are far apart
- co-expressed genes appear in clusters because neighboring genes share a single DNA enhancer element

Motivations

- ❑ Stronger correlation between the expression levels of genes that are located close to each other than between those of distant genes
 - ↳ The Human Transcriptome Map: Clustering of Highly Expressed Genes in Chromosomal Domains [Caron et al, *Science*, 2001]
 - strong clustering of highly expressed genes
 - large regions of highly expressed genes, **RIDGEs** (**regions of increased gene expression**)
 - e.g., a cluster of five matrix metalloproteinases on chromosome 11 that are down-regulated in breast cancer tissue



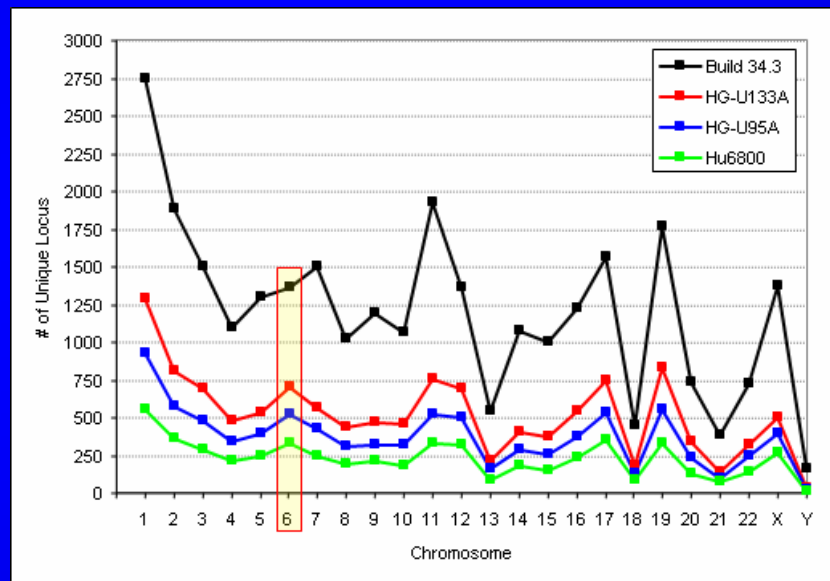
Objective and Tools

□ A computational procedure to:

- ↳ identify chromosomal regions of transcriptional modulation in specimen comparisons
- ↳ compare **regional expression profiles** to genomic structural changes (e.g., chromosomal gains and losses) and genotyping data
- ↳ mRNA and DNA profiling
 - high-density chips (e.g.: Affymetrix Human Genome U133 Set)
 - DNA array chips (e.g.: Affymetrix Mapping 10K 2.0 Array)
- ↳ Computational tools
 - R (Bioconductor libraries)
 - GDAS and Chromosome Copy Number Analysis Tool (CNAT)
 - Integrated Genome Browser (IGB)

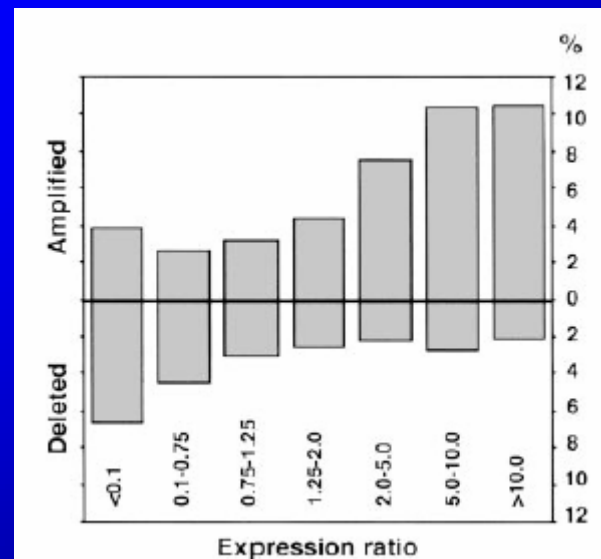
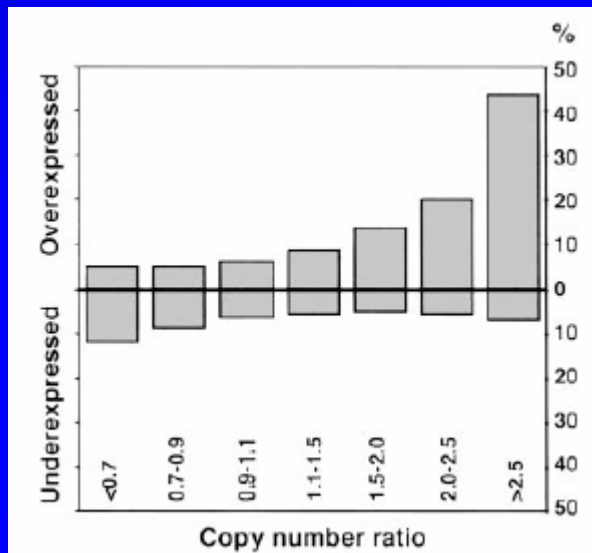
Assumptions (1)

- The systematic relationship between gene co-expression and chromosomal location, if any, is not associated with inherent chip artifacts. In Affymetrix microarrays:
 - ↳ probe order on the chips is not related to gene chromosomal order
 - ↳ the chromosomal distribution of the total gene set is representative of the distribution of known genes



Assumptions (2)

- There is a considerable influence of copy number on gene expression patterns
 - ↳ organizing gene-expression data by genomic mapping location and scanning for regions that contain a statistically modulated gene-expression signals may allow detecting chromosomal amplifications and deletions



Methods

□ **Positional re-ordering**

- ↳ probe sets are re-annotated in terms of LocusLink ID and then sorted according to their position (in bp) on chromosomes

□ **Statistical comparison of transcriptional levels**

- ↳ standard statistics (i.e., t-test, z-score, SAM) for each probe g in chromosome c

□ **Detection of transcriptional modulated regions**

- ↳ chromosomal regions of transcriptional modulation are identified
 - smoothing the transcriptional statistic on each chromosome
 - comparing the smoothed statistics to an empirical smoothed null distribution generated randomly permuting probe positions

Methods

□ Cubic splines are utilized as the smoothing function

↳ splines efficiently estimate the **local bandwidth**

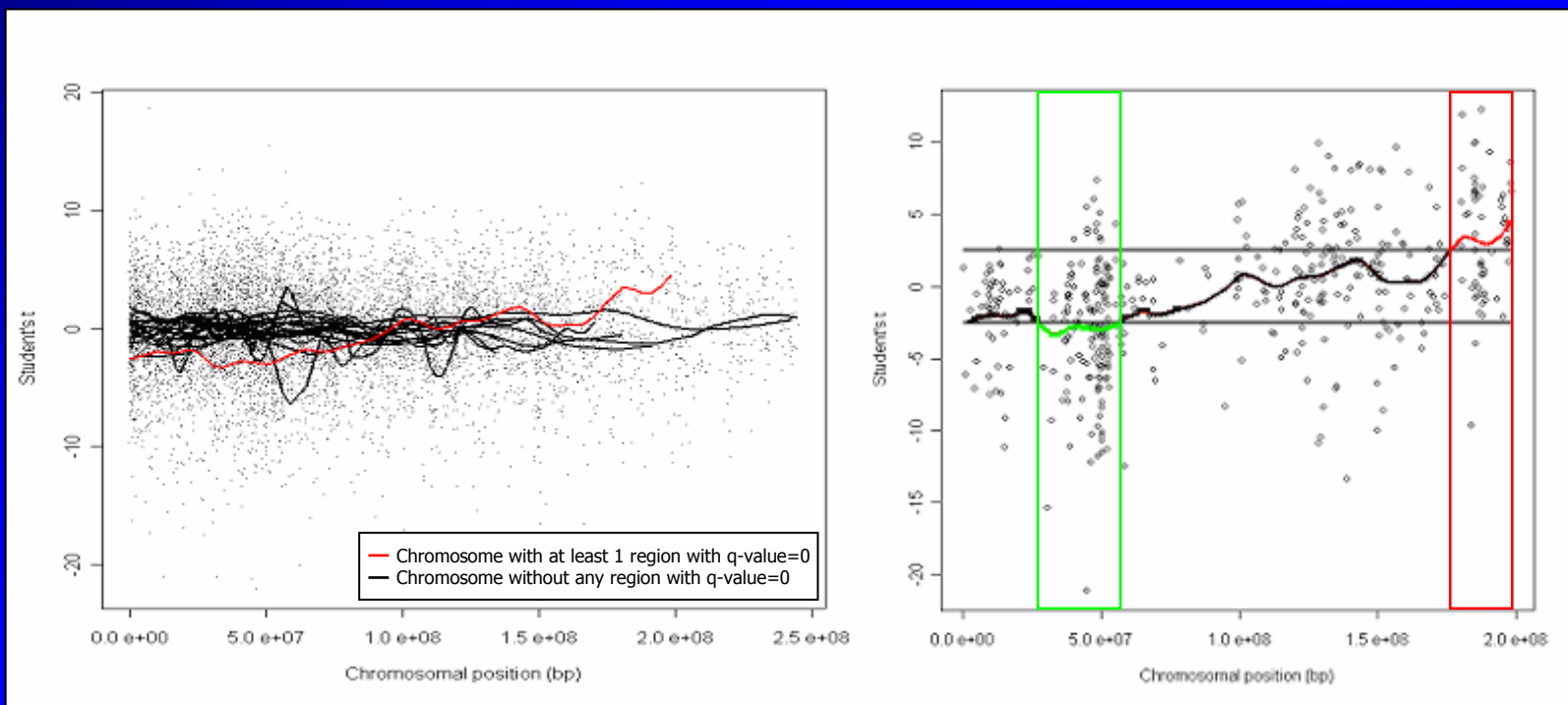
↳ determining the spline function $s(t)$, such that $y(t)=s(t)+\varepsilon$, implies minimizing a functional $F(s)$, given a twice-differentiable function s and a smoothing parameter $\alpha>0$

$$F(s) = \sum_{i=1}^n \{y_i - s(t_i)\}^2 + \alpha \int_a^b \{s''(t)\}^2 dt; \quad \alpha = \left\{ \left(\max t_j - \min t_j \right)^3 n^{-1} \sum_{i=1}^n w_i \right\} \alpha_s$$

Methods

- Differentially expressed regions are identified using a permutation procedure
 - ↳ ordering of probe positions is randomly shuffled ($P=1 \times 10^5 \div 1 \times 10^6$)
 - ↳ for each of these permutations the smoothing of probe-specific statistic is recalculated
 - ↳ smoothed randomly generated statistics generate the smoothed null distribution for scores over chromosomal regions
- The null distribution is used to assess significance of smoothed statistic for probe g on chromosome c observed in a relevant sample grouping
 - ↳ q -values are estimated to correct for multiple comparison [Storey and Tibshirani, *PNAS*, 2003]

Methods



Case studies

- ❑ Translocations in pediatric leukemia [Ross et al., *Blood*, 2003]
- ❑ Imbalances in squamous cell lung carcinomas as compared to normal lung tissue [Bhattacharjee et al., *PNAS*, 2001]
- ❑ Imbalances in renal cell tumors as compared to adjacent normal tissue [Lenburg et al., *BMC Cancer*, 2003]
- ❑ **Gene expression profiling of plasma cell dyscrasias** [Mattioli et al., *Oncogene*, 2005]
 - 13q deletion in MM samples
 - chromosome 11 polysomy in MM samples

Case study: MM

□ Expression data set:

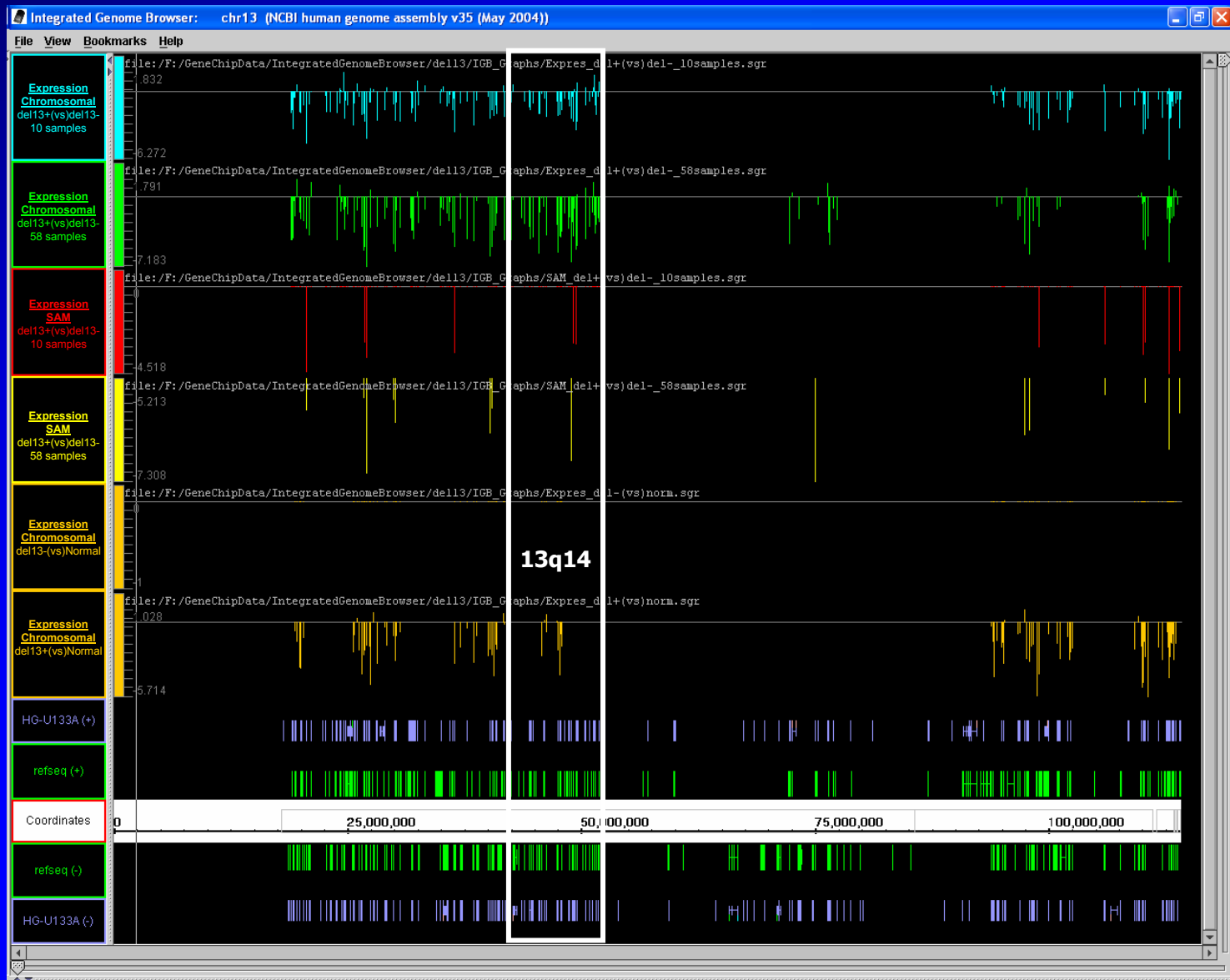
- ↳ Affymetrix HG-U133A
- ↳ multiple myeloma samples
- ↳ FISH analysis
 - 33 samples positive to 13q-
 - 25 samples negative to 13q-
- ↳ 4 normal samples

□ Genotyping data set:

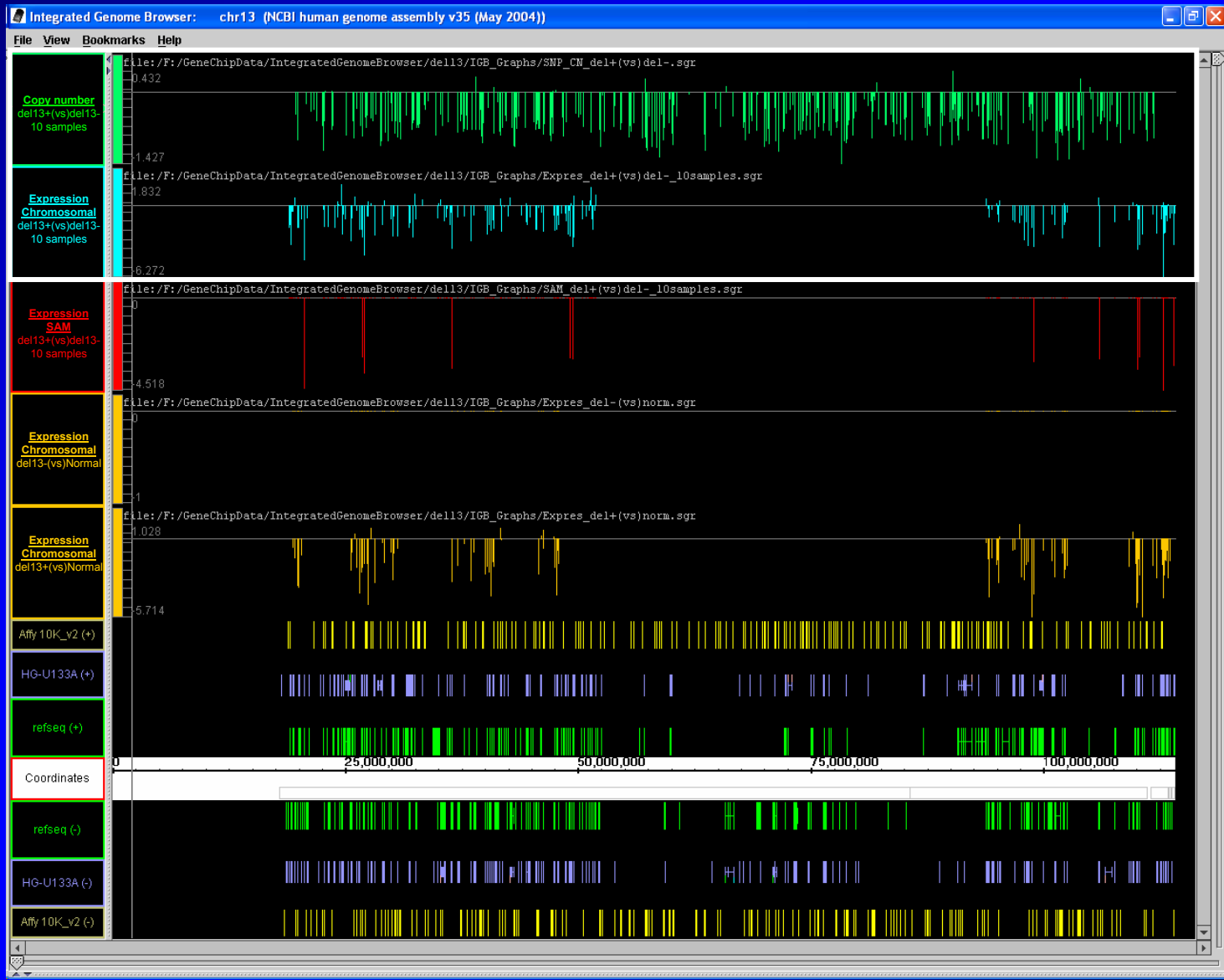
- ↳ Affymetrix Mapping 10K 2.0 Array
- ↳ multiple myeloma samples
- ↳ FISH analysis
 - 5 samples positive to 13q- and negative to 11+
 - 5 samples negative to 13q- and positive to 11+

Sample name	Stage	Stage Grade	TC	beta2-microglobulin	Bone lesions	13q-	11+	t(4;14)	t(6;14)	t(11;14)/t(11;7)	t(14;16)	t(14;20)/t(20;14)
MM-019	II	A	1	P	2	+	+	3/4	-	-	+	-
MM-021	II	A	4	P	0	-	+	-	+	-	-	-
MM-025	I	A	5	N	0	-	+	4	-	-	-	+
MM-027	I	A	3	P	0	-	+	-	-	-	-	-
MM-036	II	A	3	P	0	-	+	4	-	-	-	-
MM-037	II	A	1	N	1	+	+	-	-	-	+	-
MM-038	II	A	2	P	1	+	+	-	-	-	-	-
MM-040	III	B	3	P	2	+	+	-	-	-	-	-
MM-042	III	A	4	P	1	+	+	-	+	-	-	-
MM-047	III	A	3	P	1	+	+	-	-	-	-	-
MM-048	I	A	3	P	1	+	+	3	-	-	-	-
MM-052	I	A	1	N	1	+	+	-	-	-	+	-
MM-063	II	B	4	N	0	-	+	-	(+) lack FGFR3	-	-	-
MM-067	III	A	4	P	1	+	+	-	+	-	-	-
MM-069	II	A	5	P	1	+	+	4	-	-	-	+
MM-070	III	B	1	P	2	+	+	-	-	+	-	-
MM-072	I	A	3	N	0	-	+	-	-	-	-	-
MM-074	III	A	4	P	2	+	+	-	+	-	-	-
MM-078	III	A	3	P	1	+	+	3	-	-	-	-
MM-083	I	A	4	N	0	-	+	-	+	-	-	-
MM-087	III	A	4	P	2	+	+	-	+	-	-	-
MM-089	II	A	4	P	0	-	+	-	+	-	-	-
MM-094	I	A	3	P	0	-	+	-	-	-	-	-
MM-100			1	N			+	-	-	+	-	-
MM-101	I	A	3	N			+	4	-	-	-	-
MM-103	III	A	3	P	1	+	+	4	-	-	-	-
MM-104			4	P	2	+	+	on going	(+)	-	-	-
MM-107			3?	P	0	-	+	-	-	-	-	-
MM-109			4	P			+	-	+	-	-	-
MM-113			4				+	-	(+)	-	-	-
MM-114			3	P			+	-	-	-	-	-
MM-115			1				+	3	-	-	+	-
MM-117			3	P			+	-	-	-	-	-
MM-004	I	A	5	P	0	-	-	n.d.	-	-	-	+
MM-014	II	A	2	P	2	+	-	3	-	-	-	-
MM-015	II	A	1	P	1	+	-	3/4	-	-	+	-
MM-016	III	B	3	P	1	+	-	4	-	-	-	-
MM-026	III	B	1	P	2	+	-	-	-	-	+	-
MM-030	II	A	2	P	0	-	-	3	-	-	-	-
MM-031	III	A	1	P	2	+	-	4	-	-	+	-
MM-032	III	B	1	P	0	-	-	3	-	-	+	-
MM-034	I	A	2	P	1	+	-	-	-	-	-	-
MM-035	I	A	2	P	0	-	-	4	-	-	-	-
MM-039	II	A	2	P	1	+	-	3	-	-	-	-
MM-043	I	A	2	N	1	+	-	3	-	-	-	-
MM-049	III	B	2	N	2	+	-	3	-	-	-	-
MM-050	I	A	3	N	0	-	-	-	-	-	-	-
MM-054	I	A	1	P	1	+	-	3	-	-	+	-
MM-055	III	A	1	P	0	-	-	-	-	-	+	-
MM-056	I	A	2	N	1	+	-	3	-	-	-	-
MM-066	III	A	4	P	0	-	-	-	(+)	-	-	-
MM-077	II	A	2	N	1	+	-	3	-	-	-	-
MM-079	II	A	2	P	1	+	-	3	-	-	-	-
MM-082	I	A	3	N	0	-	-	-	-	-	-	-
MM-106			?	N	1	+	-	-	-	-	-	-
MM-111	I	A	1	P	0	-	-	-	-	-	+	-
MM-119			1				-	-	-	-	+	-
MM-121			2?				-	on going	-	-	-	-
MM-033	I	A	4	P	0	-	n.d.	n.d.	+	-	-	-
MM-081	I	A	3	N	0	-	n.d.	n.d.	-	-	-	-
MM-082	II	A	3	P	1	+	n.d.	-	-	-	-	-
MM-064	II	A	1	N	0	-	n.d.	-	-	-	+	-
MM-045	II	A	3	N	1	+	n.d.	n.d.	-	-	-	-
MM-068	III	A	3	N	0	-	n.d.	n.d.	-	-	-	-
MM-118			2?				n.d.	n.d.	-	-	-	-

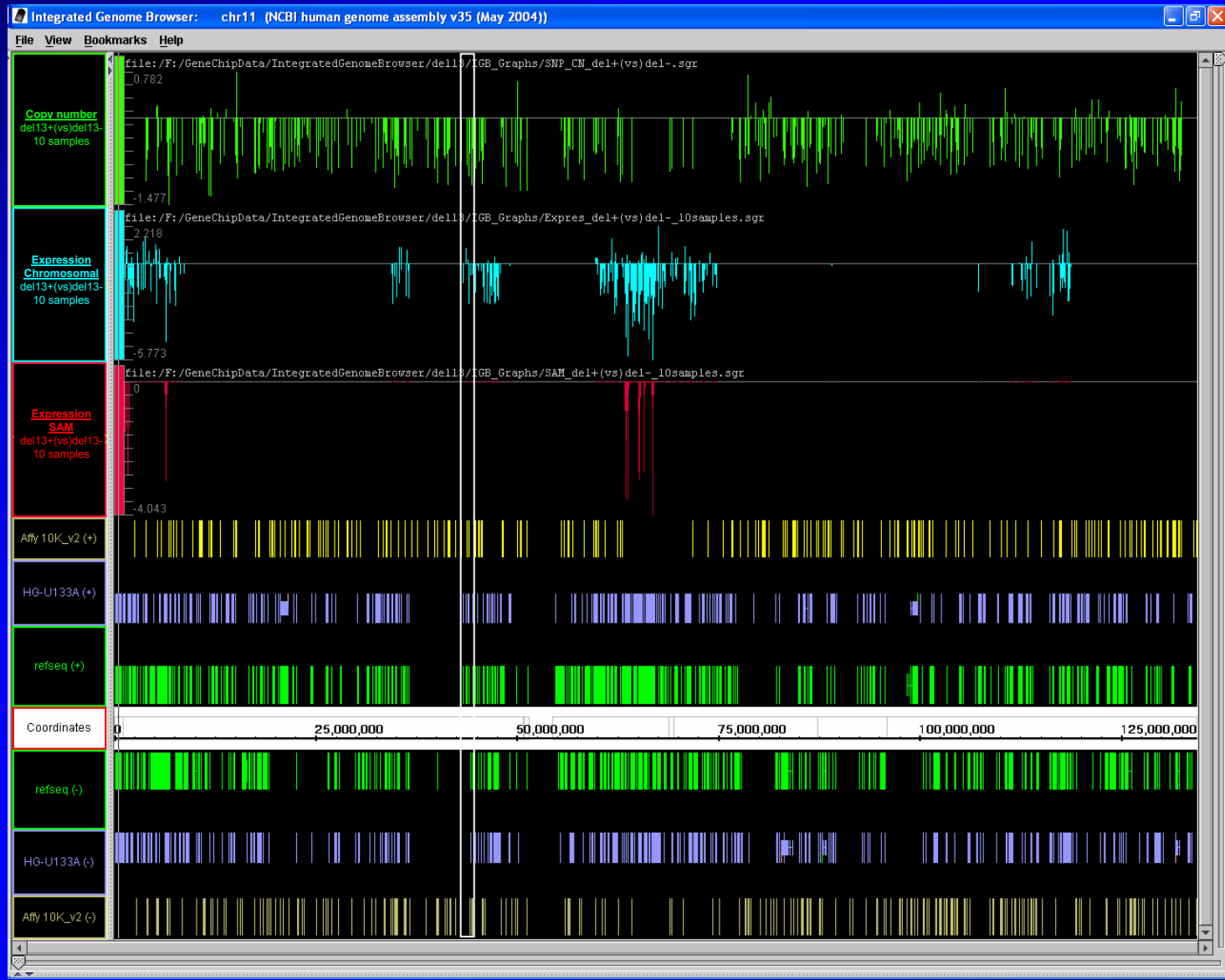
Case study: chr13



Case study: chr13



Case study: chr11

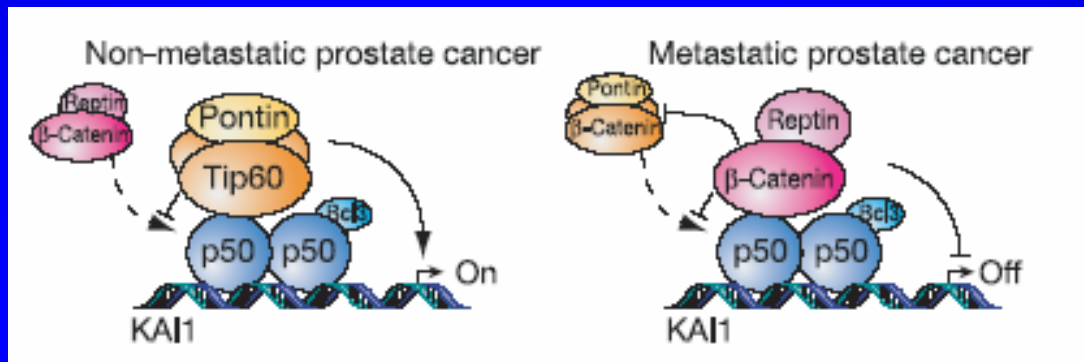


Case study: chr11



Case study: chr11

- Kim et al., *Nature*, 2005 Apr 14;434(7035):921-6
 - ↳ KAI1 is capable of inhibiting the progression of tumor metastasis without affecting primary tumorigenicity in vivo
 - ↳ molecular mechanisms of differential expression of KAI1 at the transcriptional level (chromatin immunoprecipitation assay)



- Tip60 coactivator complex recruited in non-metastatic cells

Case study: chr11

- Kim et al., *Nature*, 2005 Apr 14;434(7035):921-6

Tip60 were downregulated in metastatic cancer cells



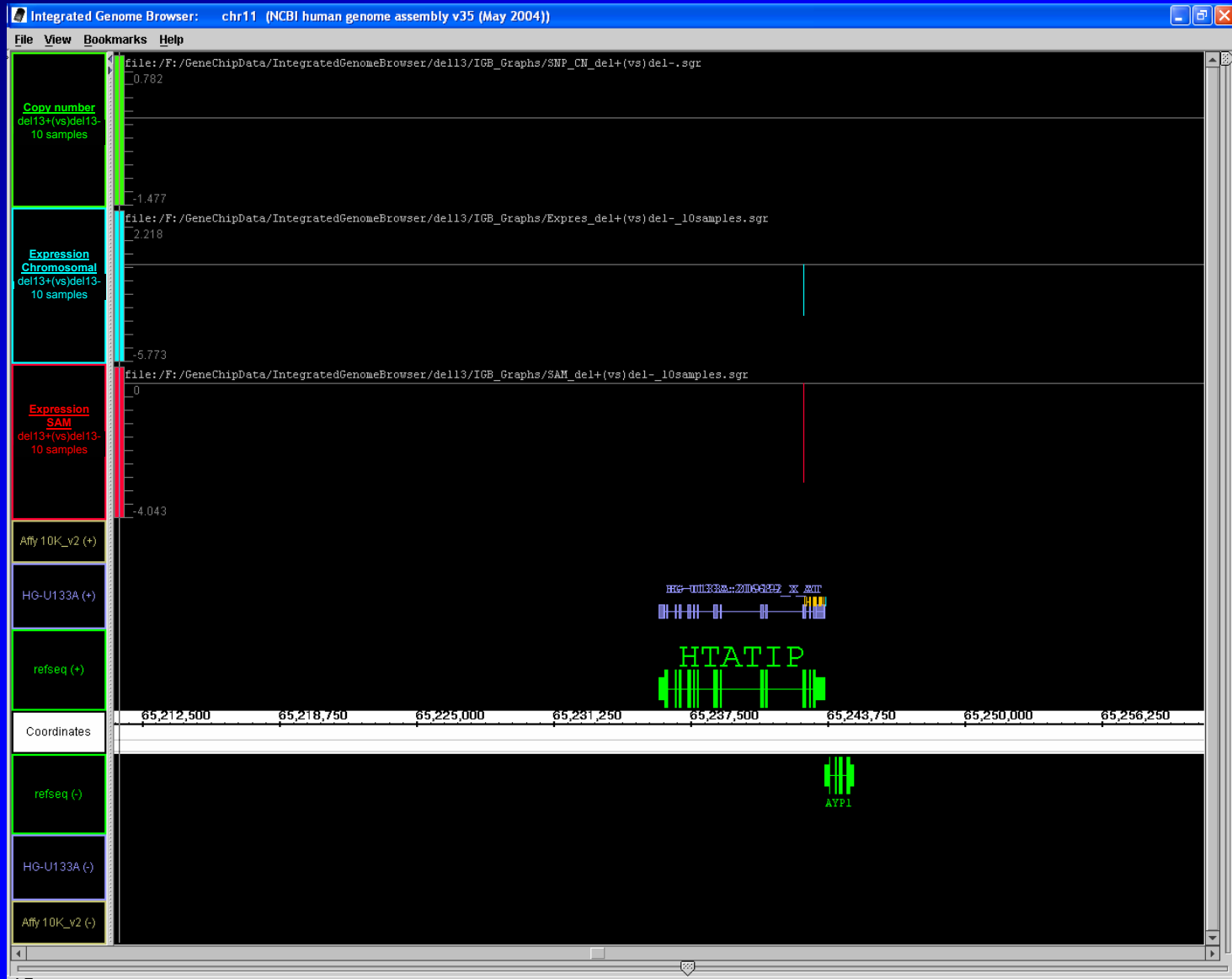
overexpression of Tip60



KAI1 mRNA levels were upregulated

- What about HTATIP expression level in del13+/- samples?

Case study: chr11

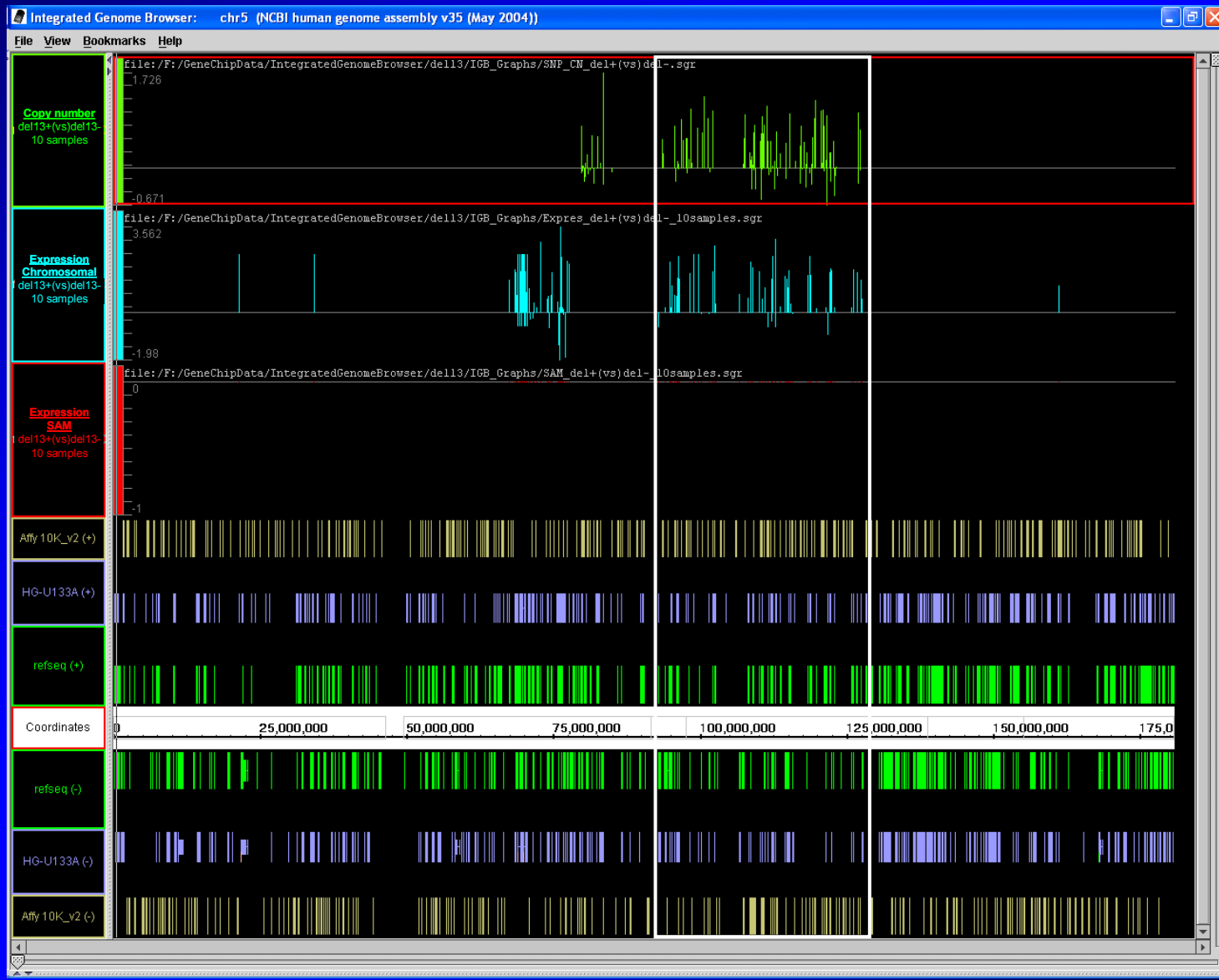


Case study: chr11

- What about HTATIP expression level in del13+/- samples?
 - ↳ down-regulated in del13+
 - ↳ detected by both SAM and chromosomal analysis
 - ↳ without KAI1 down-regulation, would HTATIP have remained a gene *in isolation*?

- What about genotype data at KAI1 locus?
 - ↳ the role for KAI1 in suppressing tumor metastasis seems not to involve either an allelic loss or a mutation [Dong et al., *Cancer Res*, 1996]
 - ↳ however, CNAT analysis of 10k data indicates CN lowering at KAI1 locus

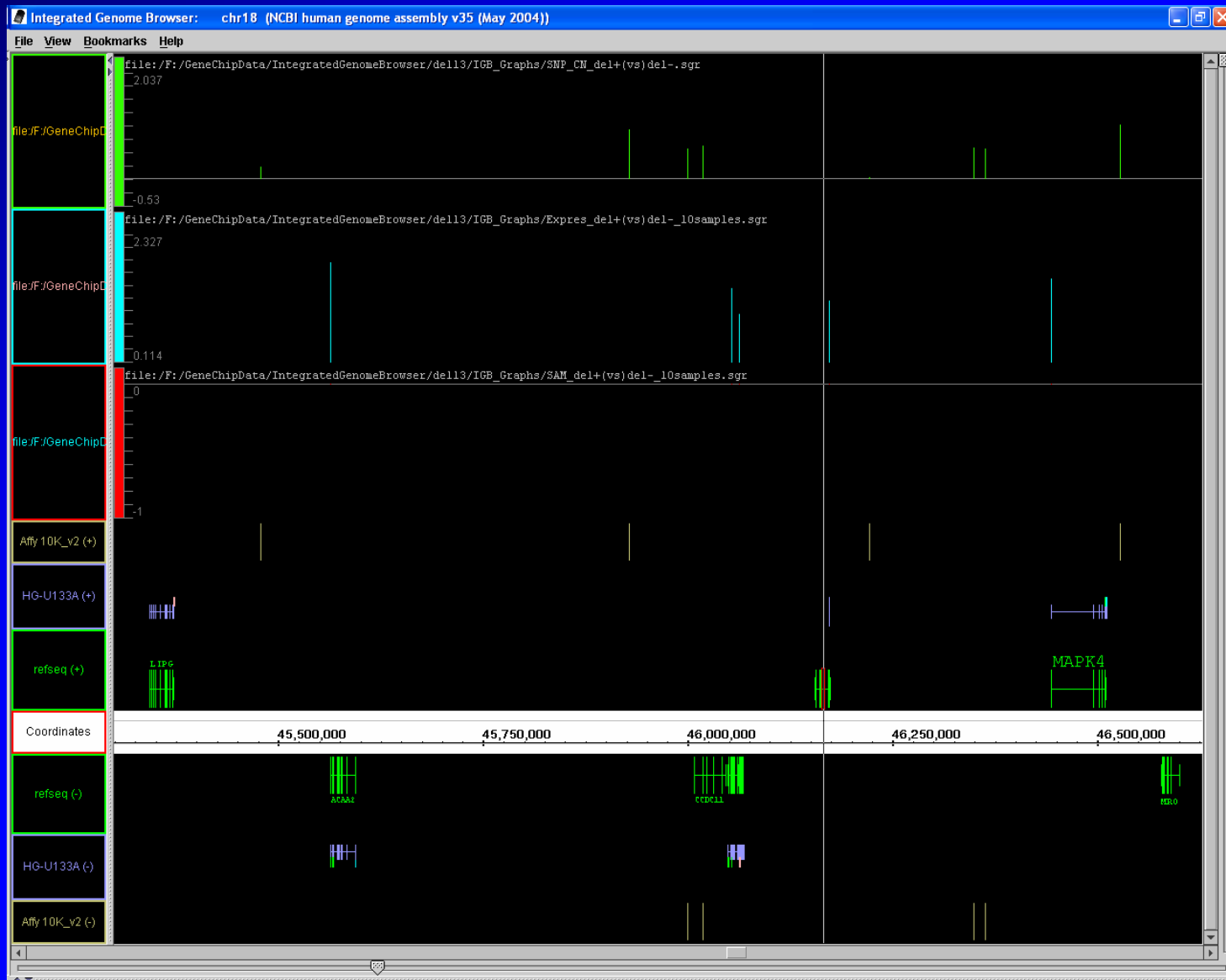
Case study: chr5



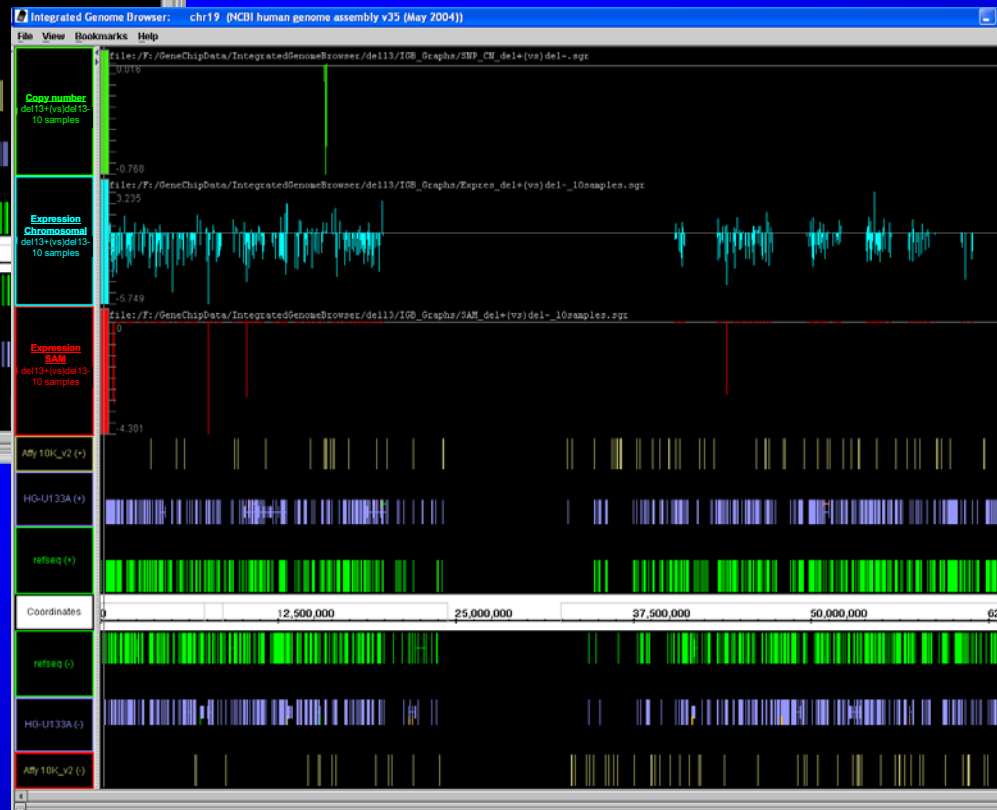
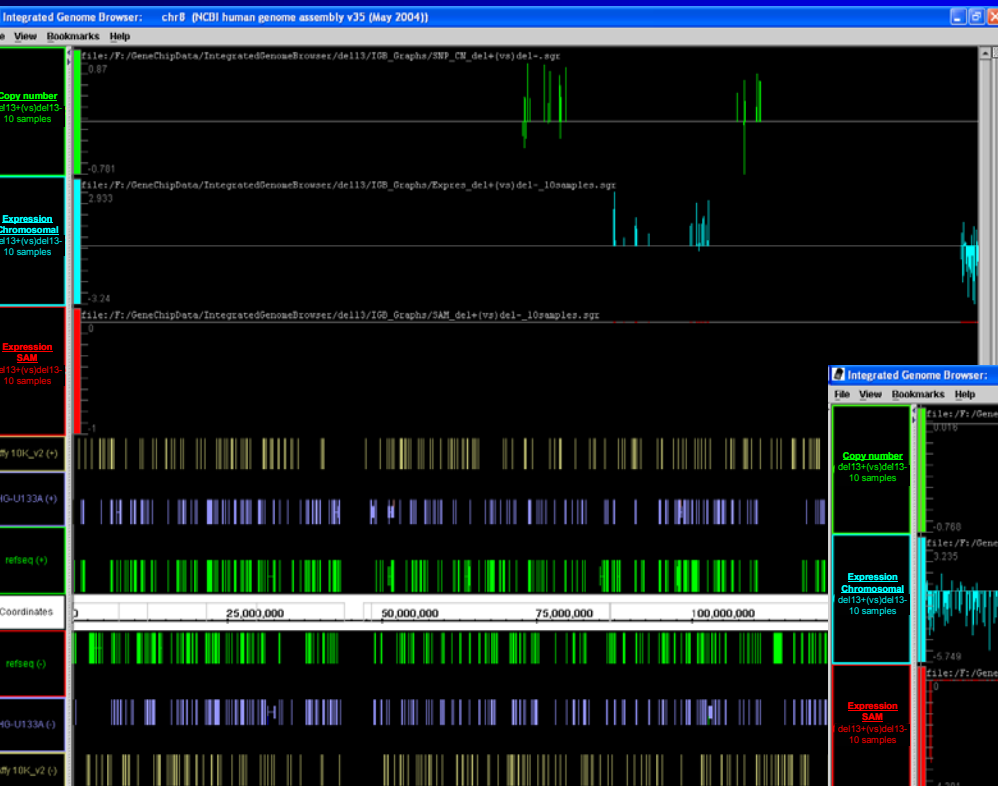
Case study: chr18



Case study: chr18



Case study: chr8 and chr19



What's next ...

□ Computational aspects:

- ↳ methods for pooling SNP array replicates to make LOH calls
 - HMM in dChip [Lin et al, *Bioinformatics*, 2004]
- ↳ unsupervised analysis
- ↳ extension to multi-class problems

□ Higher density expression chips and DNA mapping arrays

- ↳ Affymetrix GeneChip Human Genome U133 Plus 2.0 Array
- ↳ All exon chips
- ↳ Affymetrix GeneChip Mapping 100K Set

□ Result interpretation and validation

- ↳ refer to enhancer and regulatory elements databases
- ↳ pathways

Multiple myeloma project

□ Gene expression

- ↳ U.O. Ematologia 1 e 2, Dipartimento di Scienze Mediche, Università degli Studi di Milano, Ospedale Maggiore IRCCS, Milano, Italy
 - Antonino Neri
 - Luigia Lombardi
 - Michela Mattioli
 - Luca Agnelli
 - Sonia Fabris
 - Daniela Intini
 - Donata Verdelli
 - Luca Baldini
- ↳ Divisione di Ematologia e Centro Trapianto di Midollo, Azienda Ospedaliera "Bianchi-Melacrino-Morelli", Reggio Calabria, Italy
 - Fortunato Morabito
 - Vincenzo Callea

□ Genotyping

- ↳ Laboratory of Experimental Oncology, Oncology Institute of Southern Switzerland (IOSI), Bellinzona, Switzerland
 - Francesco Bertoni
 - Ivo Kwee
 - Andrea Rinaldi
 - Giulia Poretti
 - Carlo Catapano

□ Bioinformatics

- ↳ Dipartimento di Processi Chimici dell'Ingegneria, Università degli Studi di Padova, Padova, Italy
 - Andrea Callegaro
 - Alessandra Luchini
 - Silvio Bicciato

Methods

- Scores between measured values must be interpolated to compute differential expression statistics over chromosomal regions, since
 - microarrays contain only a discrete number of probes per each chromosome
 - distance in bp between measured probes varies significantly
- This interpolation does not aim at assigning statistics to non-represented locations, but to provide a smooth estimate of differential expression over chromosomal regions
- Formally it can be seen as a regression problem of the transcriptional statistic over a chromosomal coordinate

Methods

- For the given value α , minimizing $F(s)$ will result in the best compromise between smoothness and goodness-of-fit
- The amount of smoothing can be specified in terms of the smoothing parameter α
 - ↳ smoothing optimization can be made by cross-validation
 - ↳ the optimal estimator $s(t)$ would be the one that minimizes residuals $\{y_i - s_i(t)\}^2$ for a new (left out) observation y at the point t
- For any chromosome, α is estimated leaving out randomly selected statistics (i.e., probe scores) and minimizing spline residuals