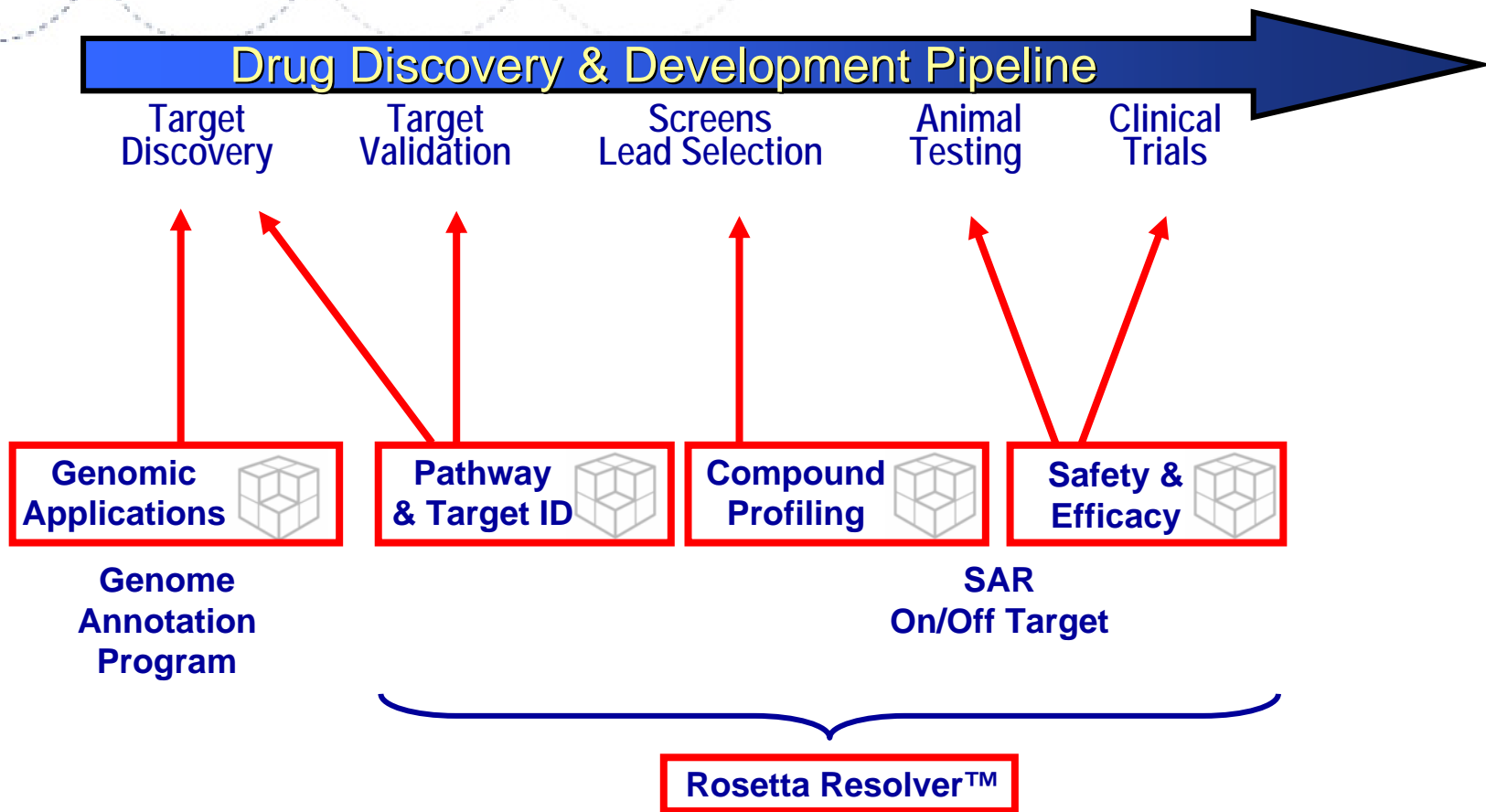

ROSETTA

INPHARMATICS



Bringing the genome to life.

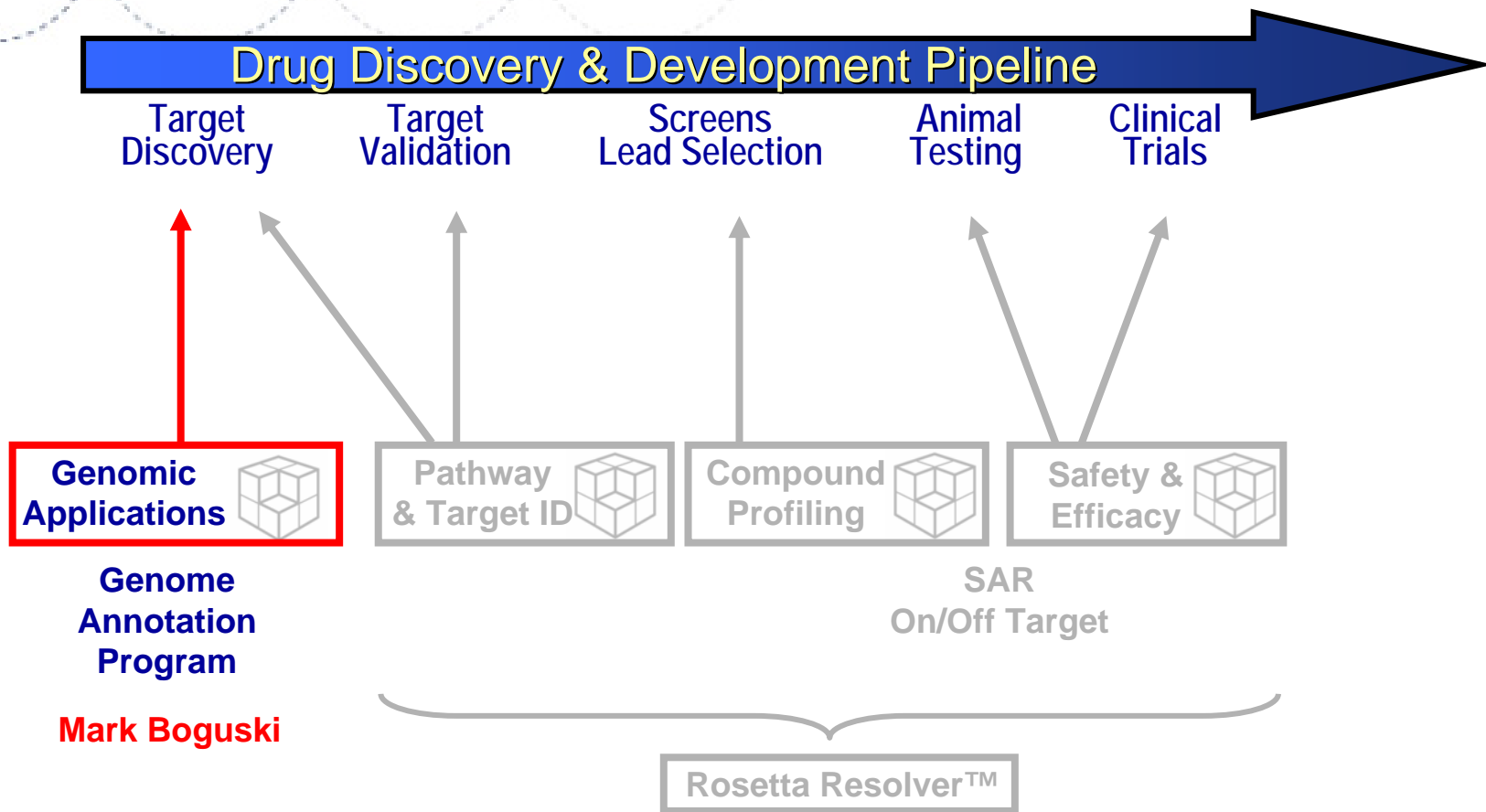
An Integrated Solution: Informational Genomics



Rosetta technologies have applications at each stage in the drug development pipeline



An Integrated Solution: Informational Genomics



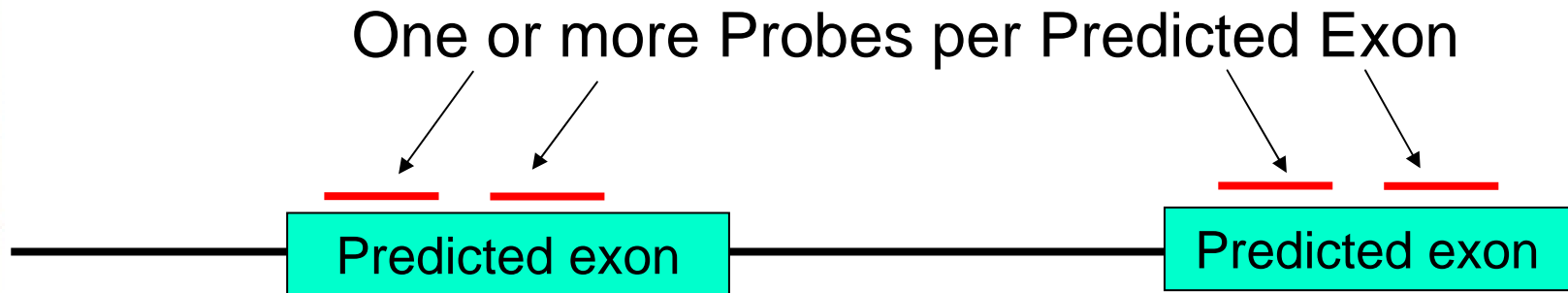
Genomic Applications
Human Genome Annotation Program

Mark S. Boguski, MD, Ph.D.
January, 2001



ROSETTA
INPHARMATICS

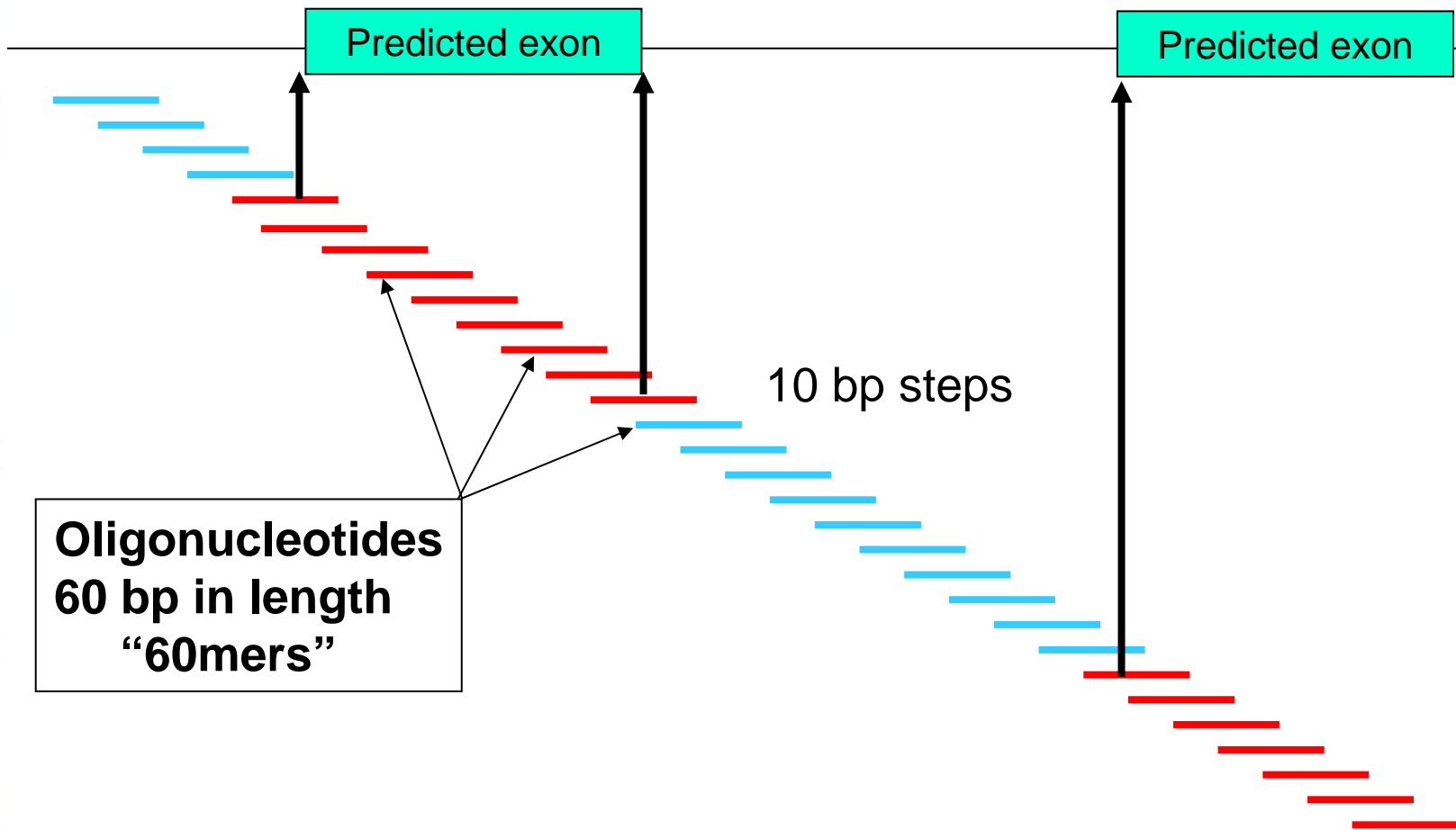
FlexJet™ Exon Arrays can validate Exon Predictions and assemble Gene Structures



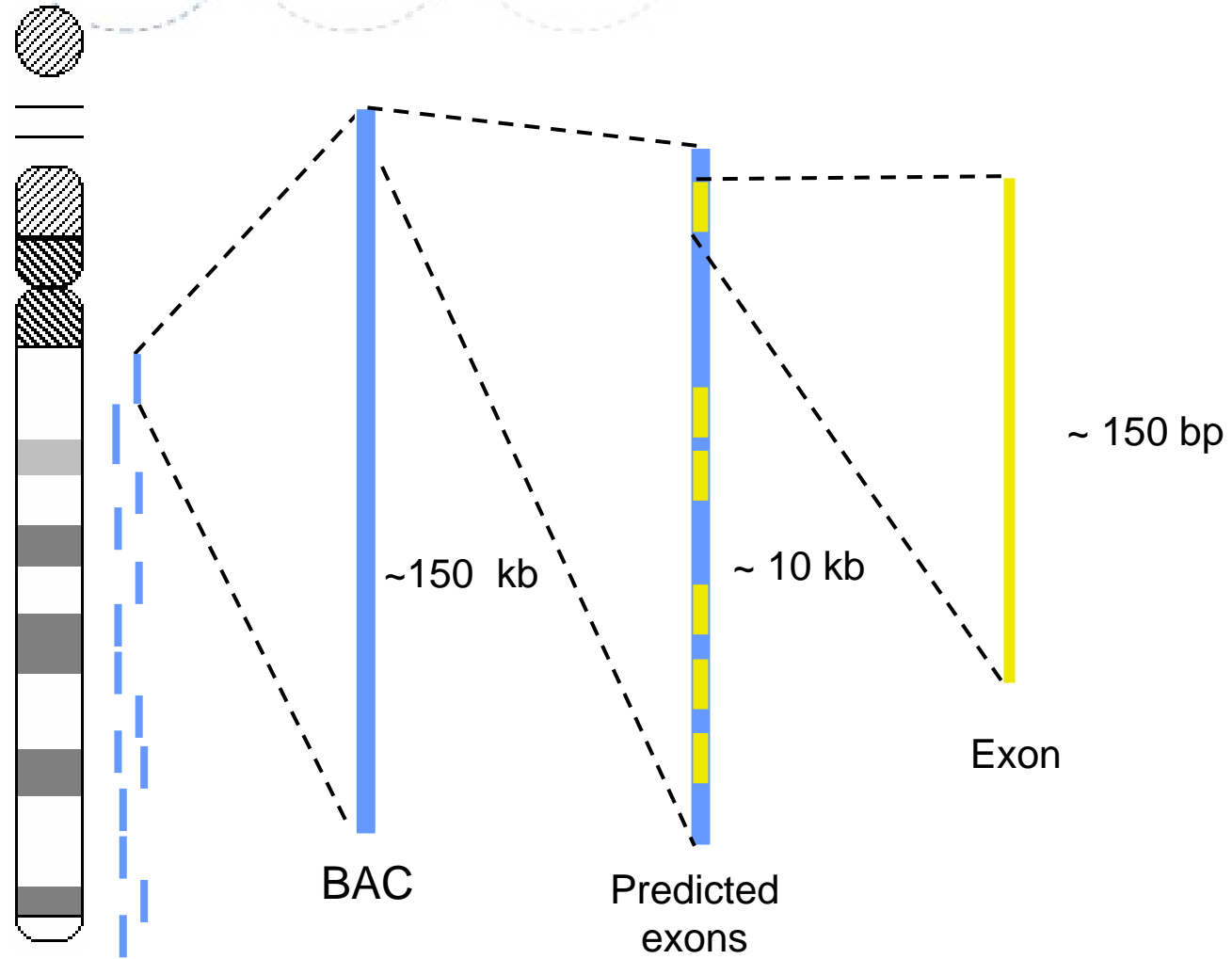
- & Verify predicted exons on a genome-wide scale.
- & Group exons into genes via co-regulation.



FlexJet™ Tiling Arrays can identify Exons and refine Gene Structures



Step 1: Identify predicted exons on chromosome 22q



Chr 22 (33 Mb)



Step 2: Design probes for each of the predicted exons

10,031 exons (confirmed + GenScan)

- Remove duplicates
- RepeatMasker

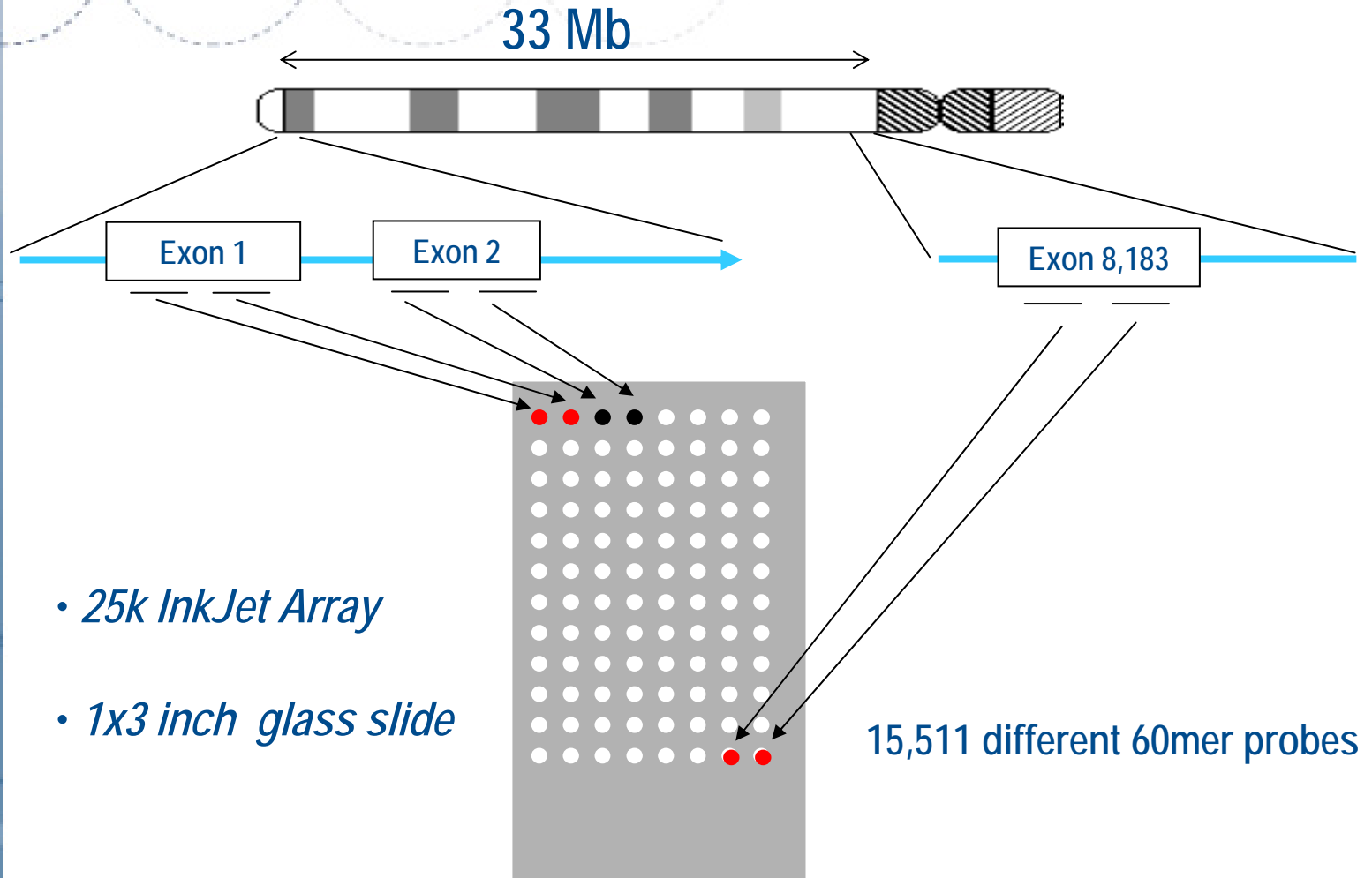
8,183 exons

- Probe selection algorithms
- Two probes per exon

15,511 60mer probes



Step 3: Generate a chromosome 22 exon array



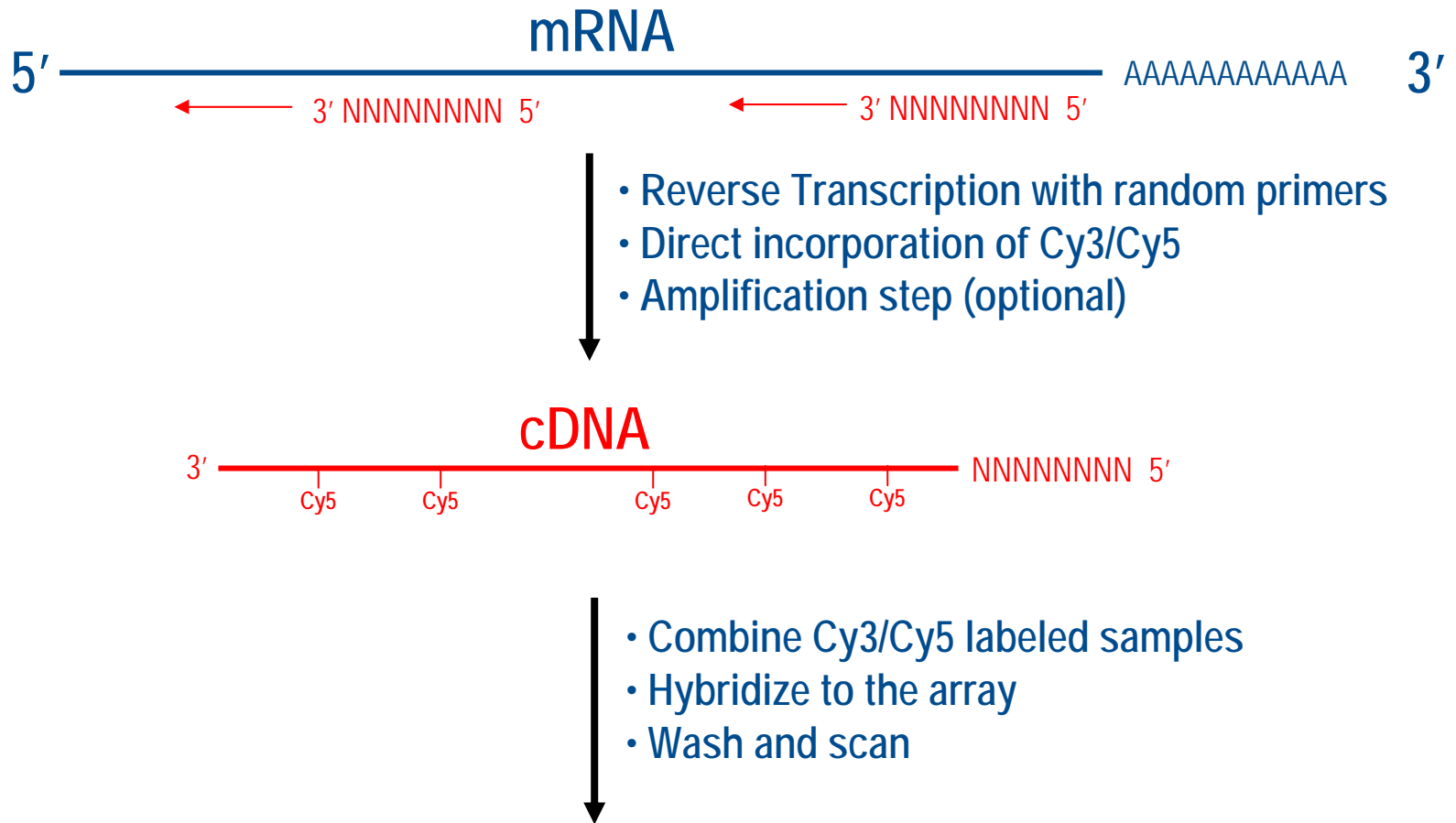
- 25k InkJet Array

- 1x3 inch glass slide

- Probes are synthesized in a linear fashion across Chromosome 22



Step 4: Label mRNA using a random prime protocol

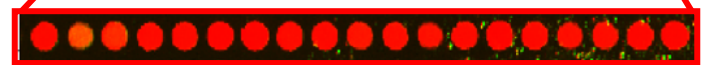
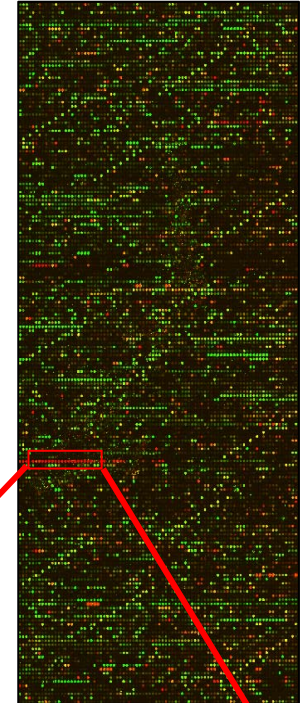
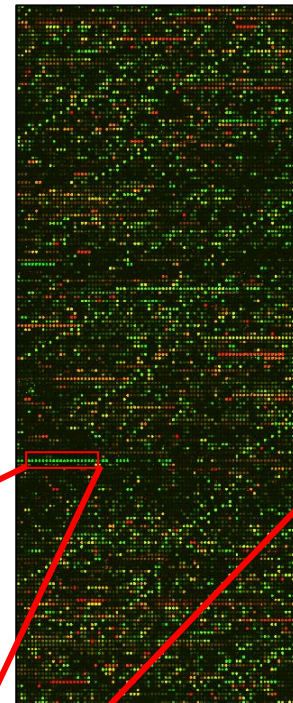


Step 4: Perform two-color hybridizations

- 25K Chromosome 22 Exon Array
- Placenta vs Pool fluor reversal
- Red and green strips = genes.

Placenta vs Pool

Pool vs Placenta



Fibulin (FBLN1)



Brain

Testes

Liver

Kidney

Heart

Bone Marrow

Lymphocytes

Endothelial Cells

1. More exons are experimentally verified.

2. Verified exons are grouped into genes (co-regulation).



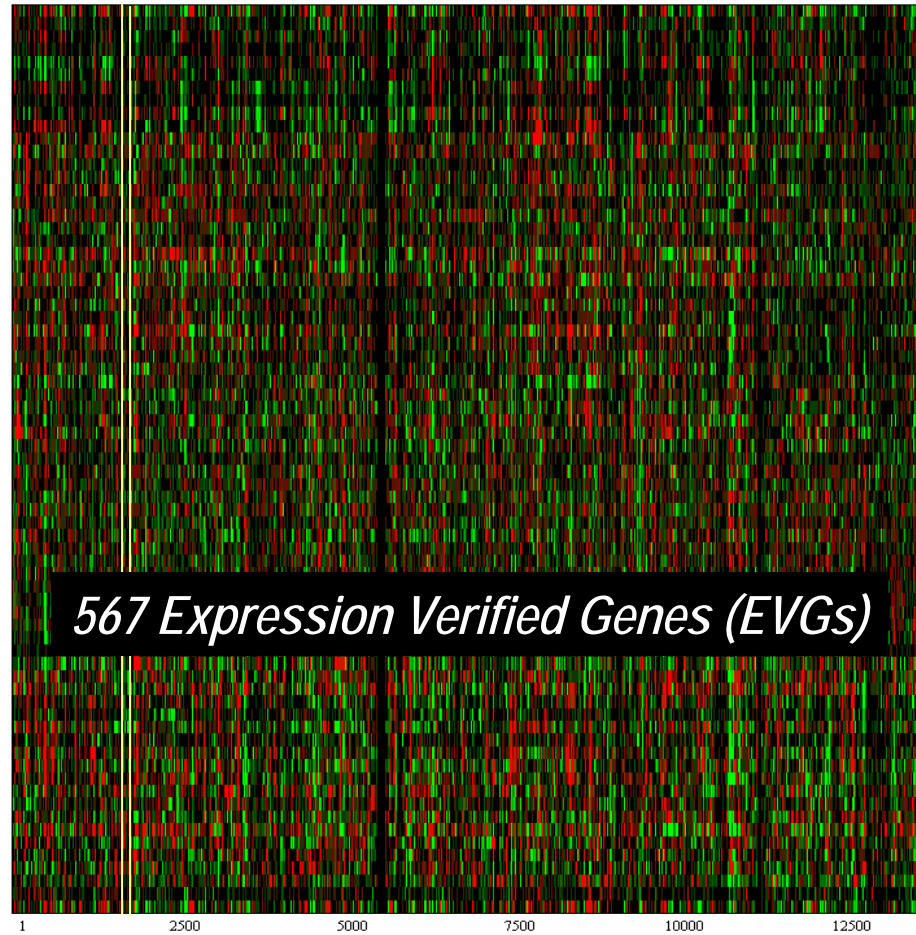
Chromosome 22, 33 Mb



Exons

1  8,183

69 Experiments



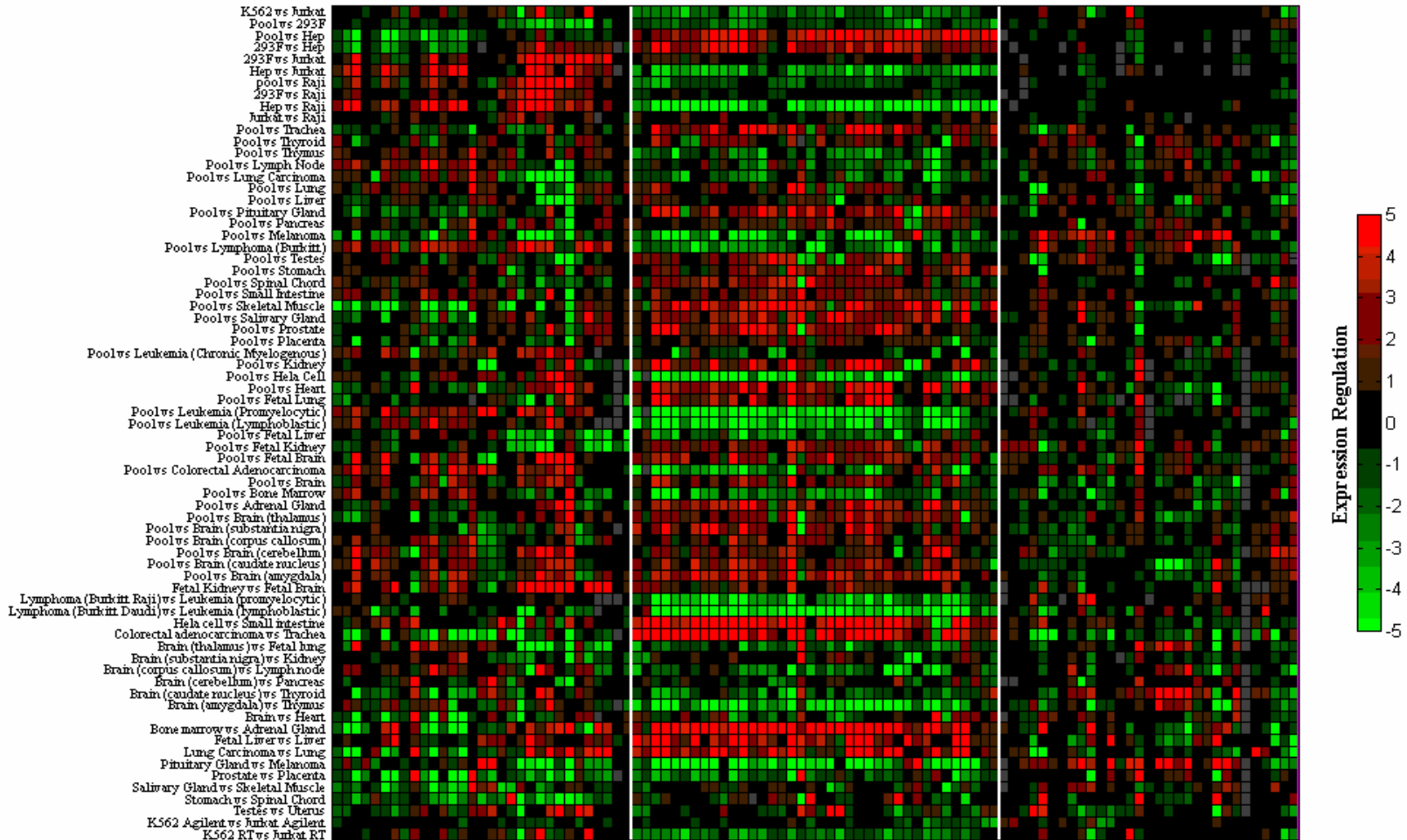
Expression Regulation
5
4
3
2
1
0
-1
-2
-3
-4
-5

1 2500 5000 7500 10000 12500

Zoom in on specific regions



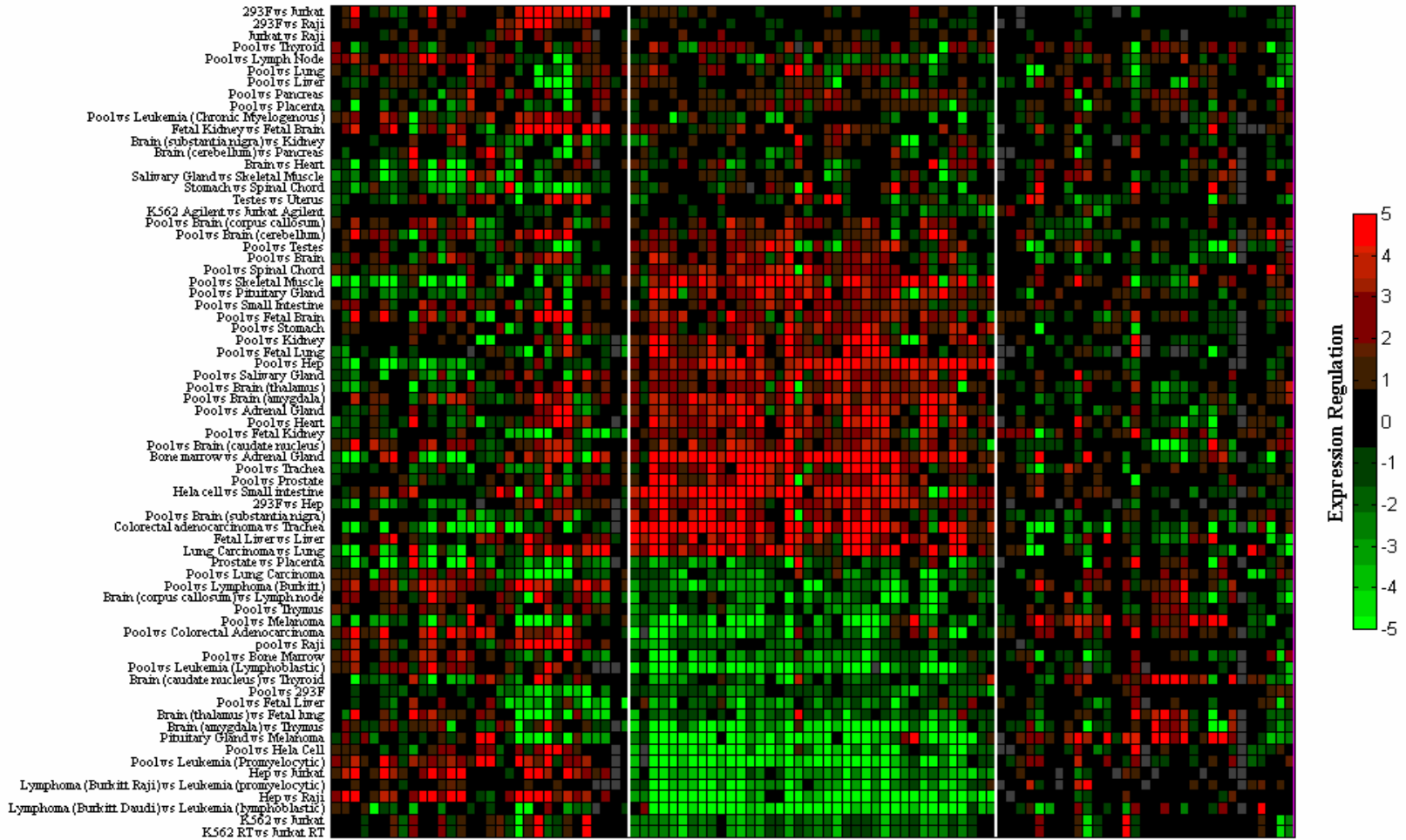
Strand = 1, BP = 19298511 - 19322816 [Z82244.00001 57558 - Z82244.00001 81922], N = 38, p-value = NaN



Exons →

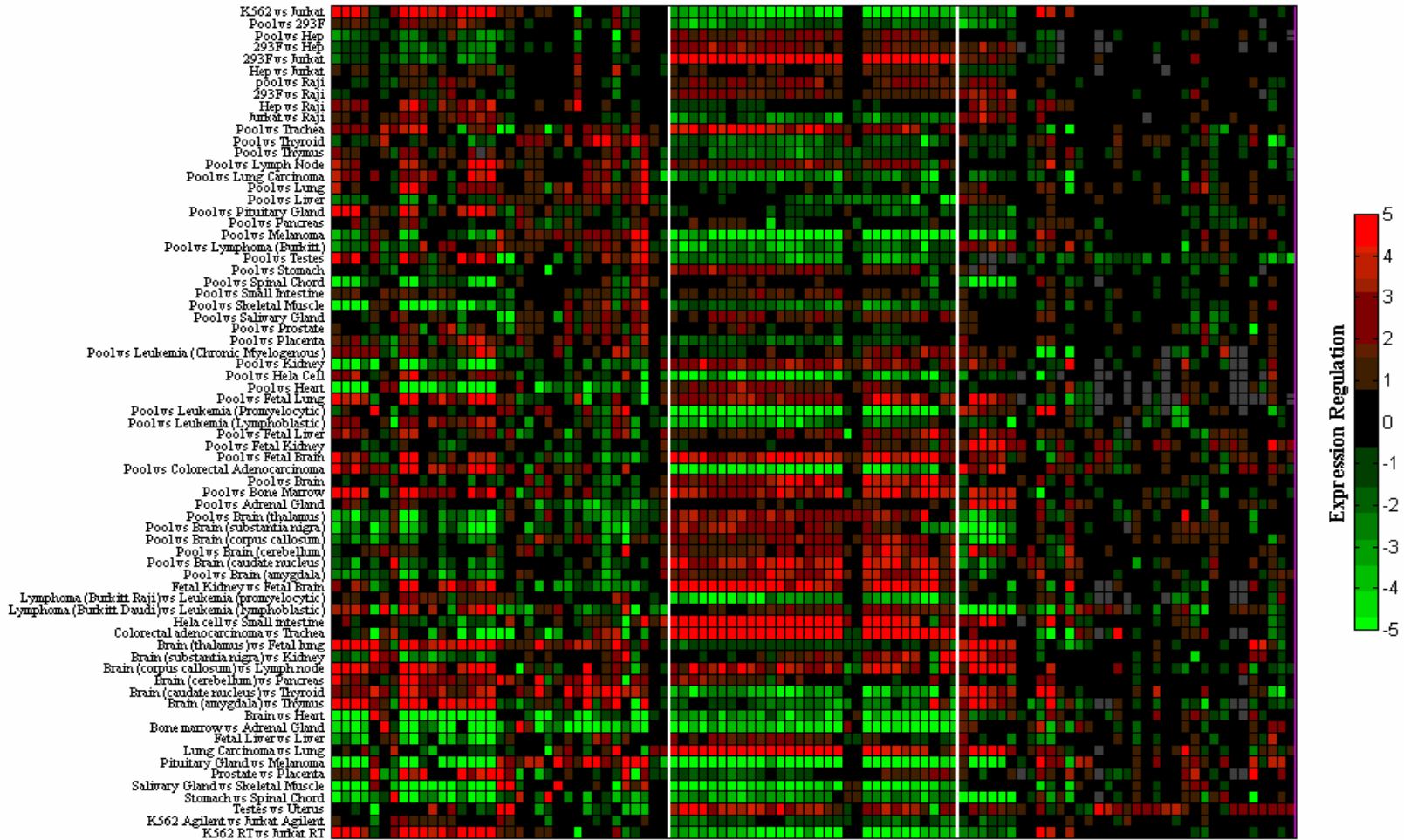
69 Experiments

Strand = 1, BP = 19298511 - 19322816 [Z82244.00001 57558 - Z82244.00001 81922], N = 38, p-value = NaN



CDC46/Mcm5

Strand = 1, BP = 25438861 - 25480776 [Z83840.00001 17145 - Z83840.00001 59119], N = 30, p-value = NaN



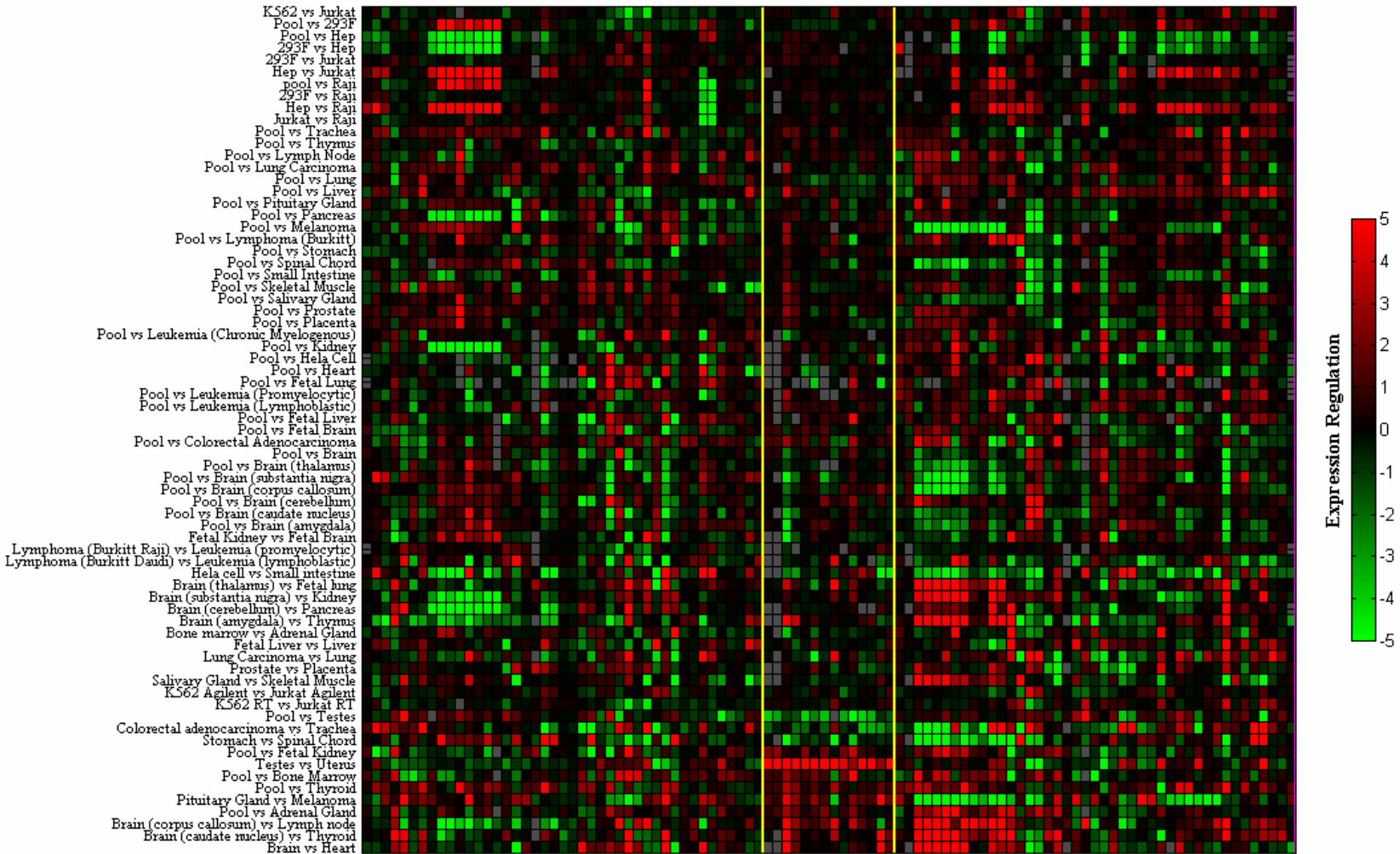
70KD Thyroid Autoantigen

Strand = 1, BP = 25438861 - 25480776 [Z83840.00001 17145 - Z83840.00001 59119], N = 30, p-value = NaN



70KD Thyroid Autoantigen

Strand = -1, BP = 21760605 - 21769974 [AL031587.00001 1008 - AL031587.00001 10436], N = 14, p-value = 0.0019608



Tissue specific expression

Performance Summary from Chromosome 22

	Dunham et al. (1999)	EVGs	Validation fraction
Known genes*	247	208	85%



Performance Summary from Chromosome 22

	Dunham et al. (1999)	EVGs	Validation fraction
High Known genes*	247	208	85%
Medium Related genes*	150	97	66%
Low Predicted genes*	148	77	67%



Level of experimental support



Performance Summary from Chromosome 22

	Dunham et al. (1999)	EVGs	Validation fraction	
High	Known genes*	247	208	85%
Medium	Related genes*	150	97	66%
Low	Predicted genes*	148	77	67%
No	<i>Ab initio</i> genes*	325	185	57%

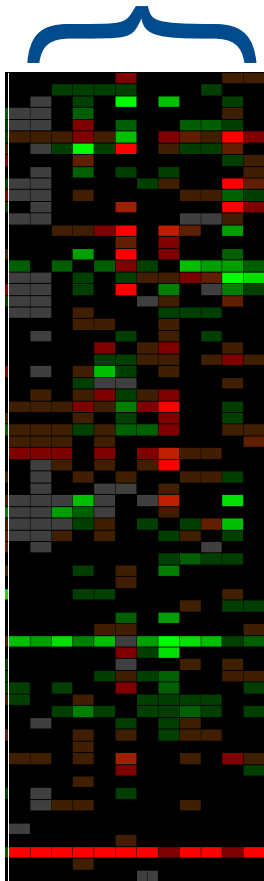


Level of experimental support



Using Tiling Arrays to Refine Gene Structure

EVG #438 (6 exons, 10 kb genomic interval)

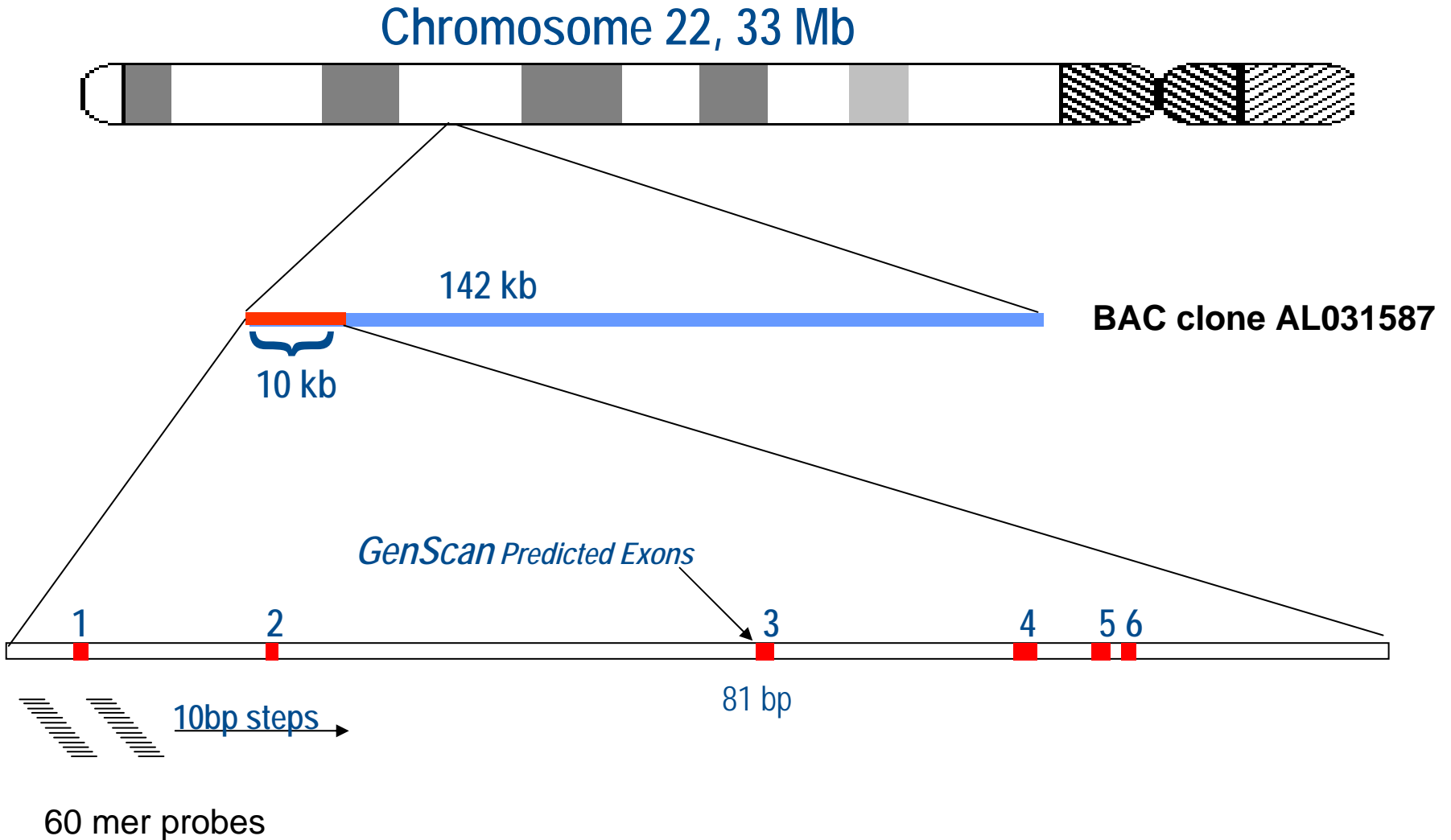


- *Were any exons in this region missed?*
- *Are the predicted splice junctions correct?*
- *Are the terminal exons correct?*

Testis



Using Tiling Arrays to Refine Gene Structure



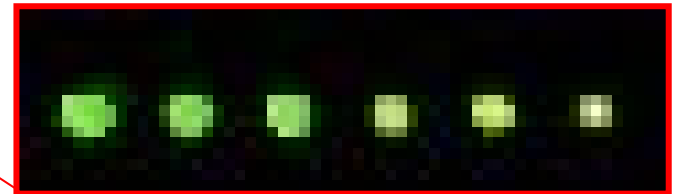
Example of a 25K tiling array



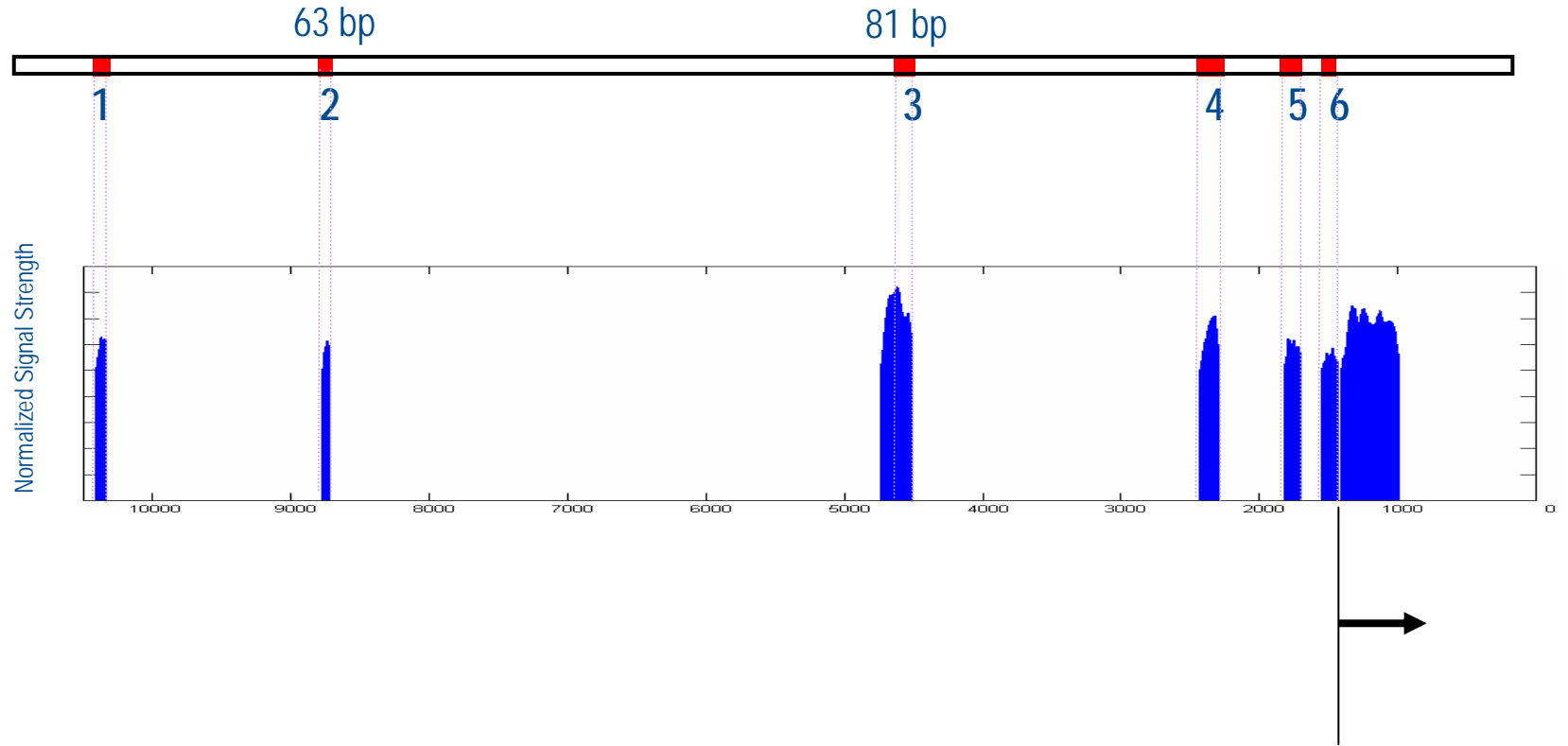
Control probes

Introns (dark)

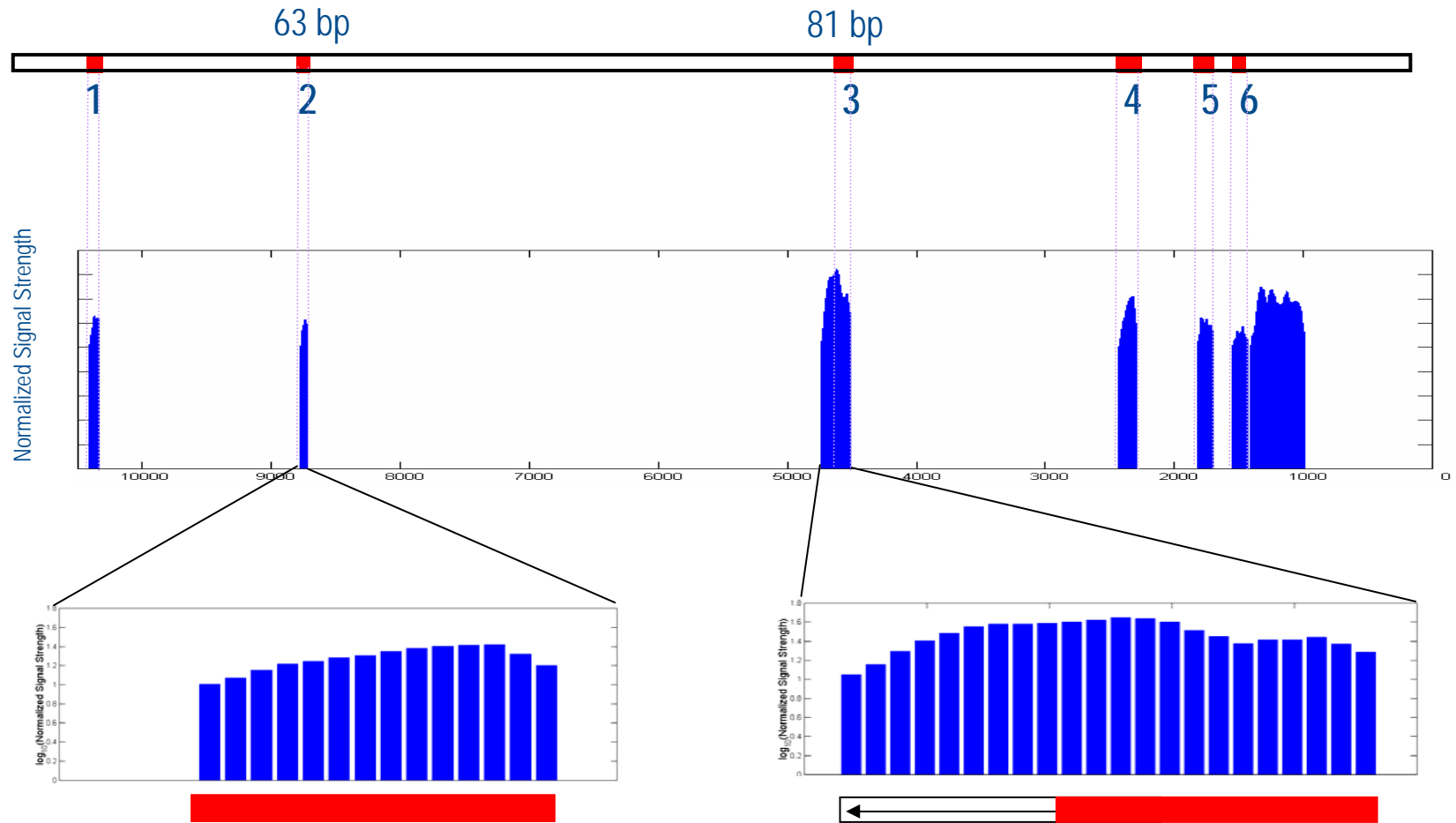
Exons (green strips)



Using Tiling Arrays to Refine Gene Structure



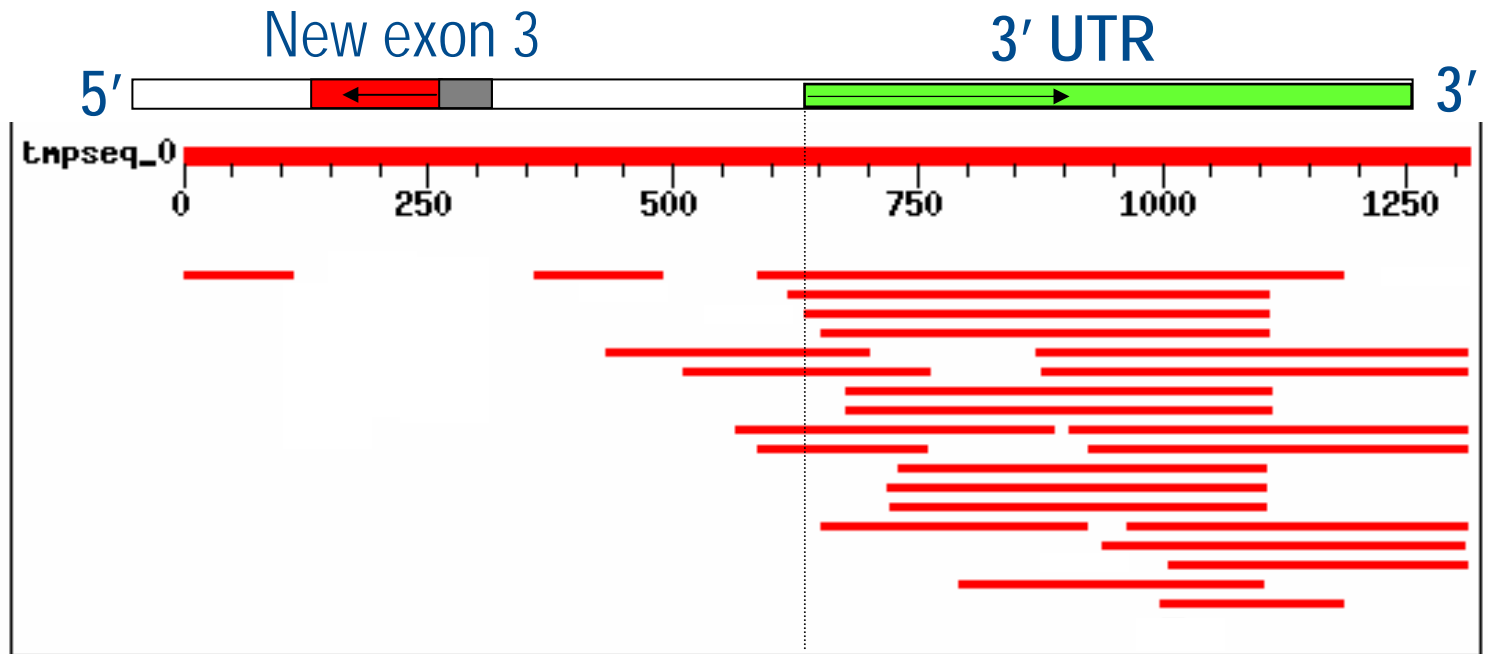
Using Tiling Arrays to Refine Gene Structure



Agrees with GenScan prediction

102 bp extension

Using Tiling Arrays to Refine Gene Structure



- Most of the EST support is in the 3' UTR.
- 104 bp extension does not have EST support.



Whole Genome Exon Analysis

Step 1. Download all exons from the Ensembl Database.

- 628,635 (June 15, 2000 freeze).

Step 2. Remove duplicates and repetitive sequences.

- 442,785 predicted and confirmed exons.

Step 3. Design two probes for every exon.

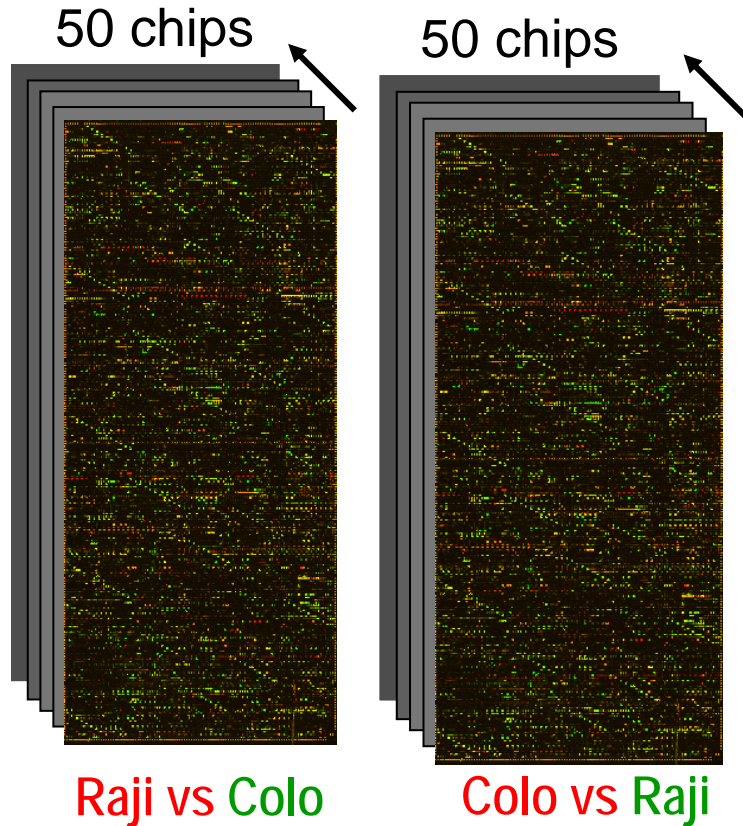
- 1,090,408 probes.

Step 4. Generate Ink-Jet arrays and hybridize under two conditions.

~ 50 arrays (exon probes + controls).



Whole Genome Exon Analysis

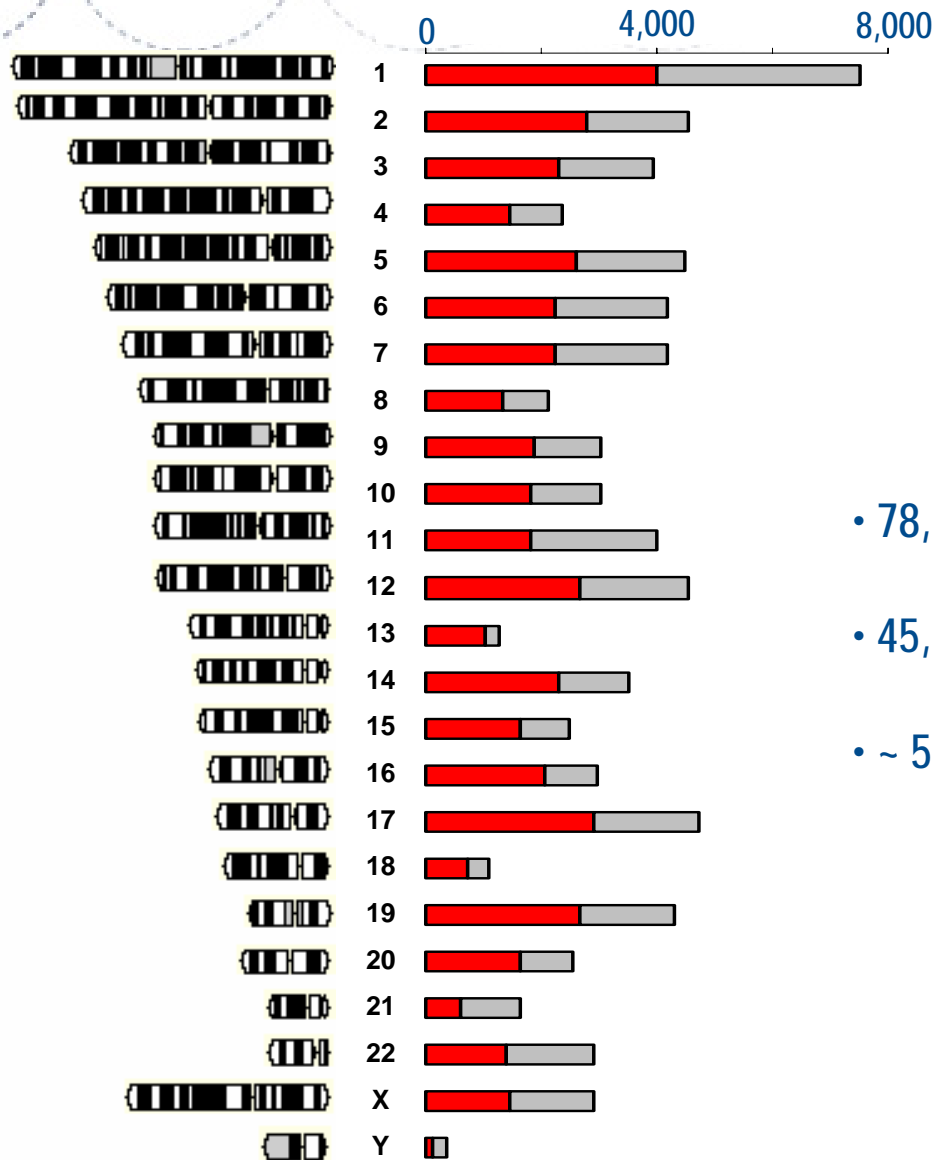


Which predicted exons are experimentally verified ?

- Burkitt's lymphoma
- Colorectal adenocarcinoma



58% of the Confirmed Exons were Verified

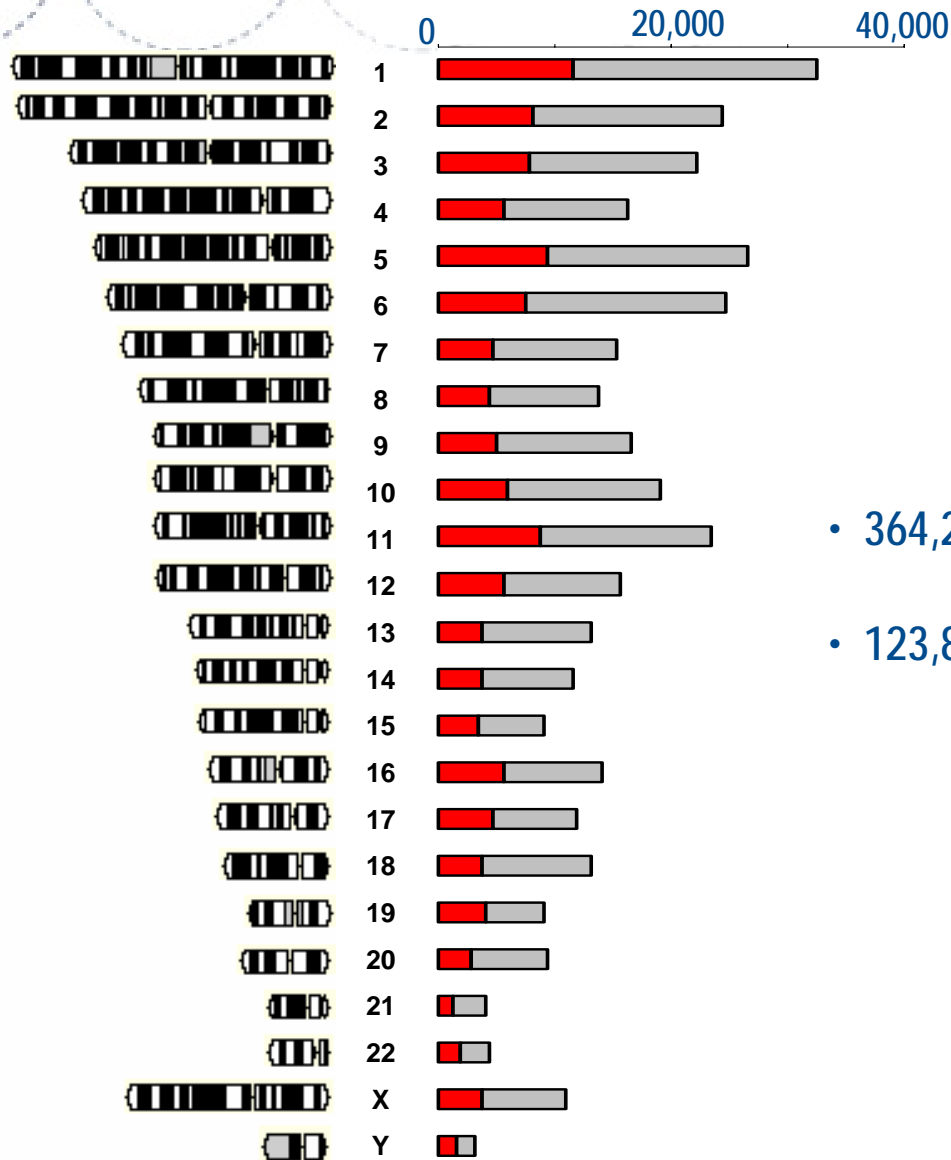


58 %

- 78,486 Total confirmed exons.
- 45,522 Verified with two conditions
- ~ 5% false positive rate..



Only 34% of the Predicted Exons were Verified



34 %

- 364,299 Total predicted exons.
- 123,861 Verified with two conditions.



Whole Genome Analysis

With additional conditions we can:

- Verify a larger fraction of the predicted exons.
- Group exons into genes via co-regulation.
- Generate an expression body atlas.
- Catalog alternative splicing.
- Suggest gene function based on co-regulation.



Gene level analysis

- Analyze disease linkage regions for novel genes.
- Extend EST's to obtain full-length clones.
- Comprehensively catalog alternative splicing for important genes.
- Use exon and tiling data to direct RT-PCR cloning efforts.

