

P. falciparum:

Examination of Correlation Between Spatial
Location and Temporal Expression of Genes

CAMDA Conference

11 November 2004

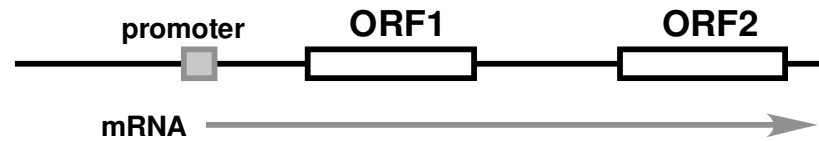
JB Christian, C Shaw, J Noyola-Martinez, MC
Gustin, DW Scott and R Guerra

Motivations:

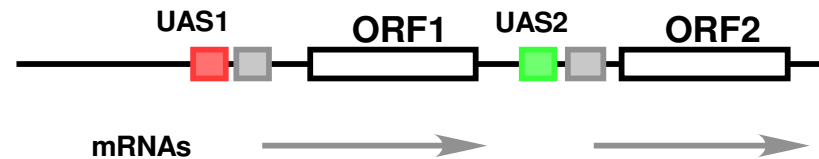
- Evidence for correlation in literature
 - Printing artifact
 - Biological
- Improving on Bozdech threshold
- Develop a visualization and statistical testing methodology

Biological Motivations

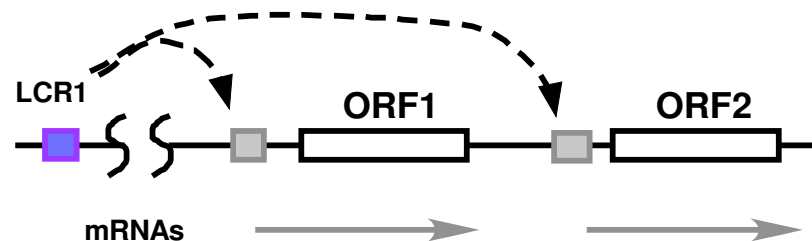
Operon control (bacteria)



Upstream Activating Sequences (yeast)



Locus Control Region (mammalian globin cluster)

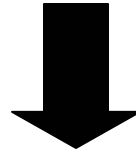


Hypothesis and Statistic

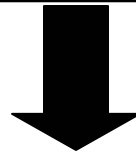
- **Statistical:** Correlation between chromosomal location and gene expression?
- **Biological:** Gene order random?
- H_0 : no correlation between location on chromosome and expression
- Consider correlations in partitions

Approach

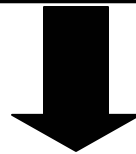
Covariogram: General Tool



Partition Chromosome, Develop Statistic



Permutation Testing Framework



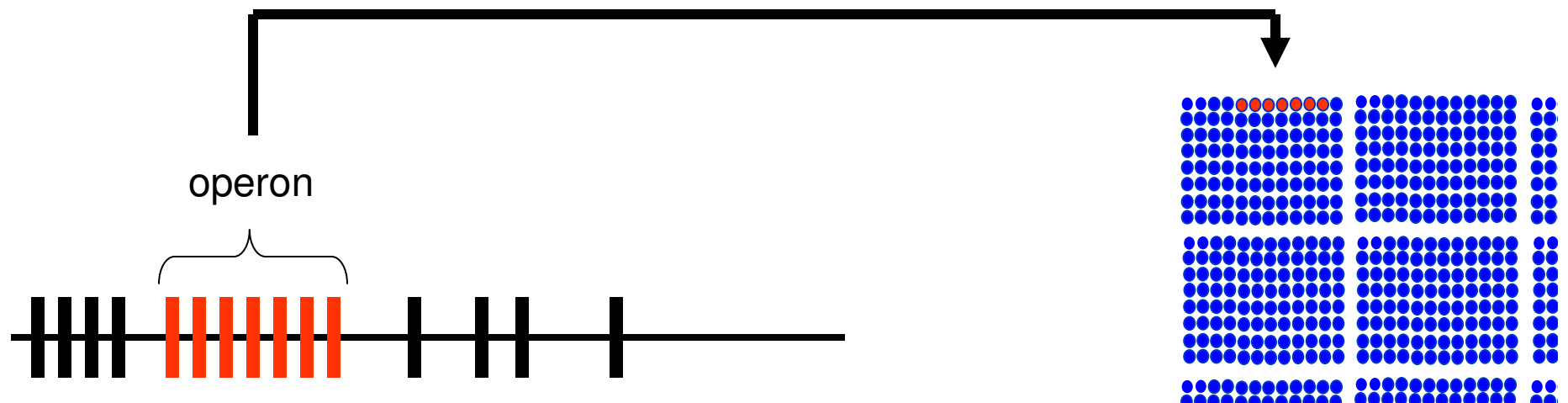
Check for Confounding Factors



Biological Significance

Issues

- Confounding (printing) or other artifacts
- Account for inter-gene distances (as opposed to adjacent pairwise correlation)
- Significance of correlation



Methods: Data

- Need gene information (plasmodb.org has annotated fastA files):

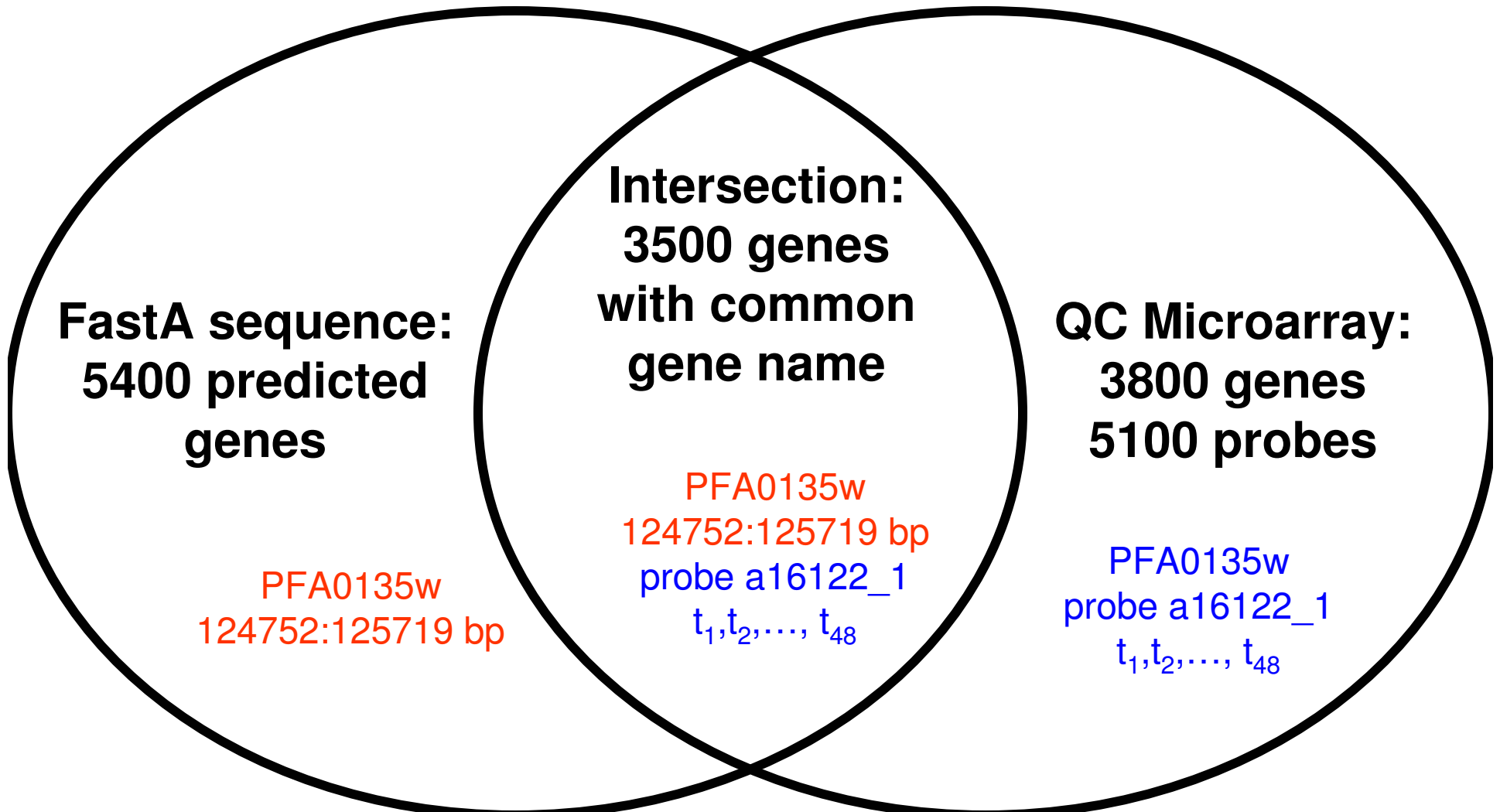
```
TCAAGCAATTGTTAGATGAGAACAATAGGAAGAATTTAAATTTTAATGAT
CTGGTTATACACCCTTGGTGGTCTTATAAGAATTAA
```

```
>Pfa3D7|pfal_chr1|PFA0135w|Annotation|Sanger (protein
coding) hypothetical protein
Location=join(124752..124823,124961..125719)
```

```
ATGATATTTTCATAAATGCTTTAAAATTTGTTCGCTCTCTTGTACTGTTTT
ATGGGTTACCGCCATATCATCGATCATTCAACCAGACAAACAACAAGAAA
```

- Normalized `gpr` files (2-D loess, centered and scaled)

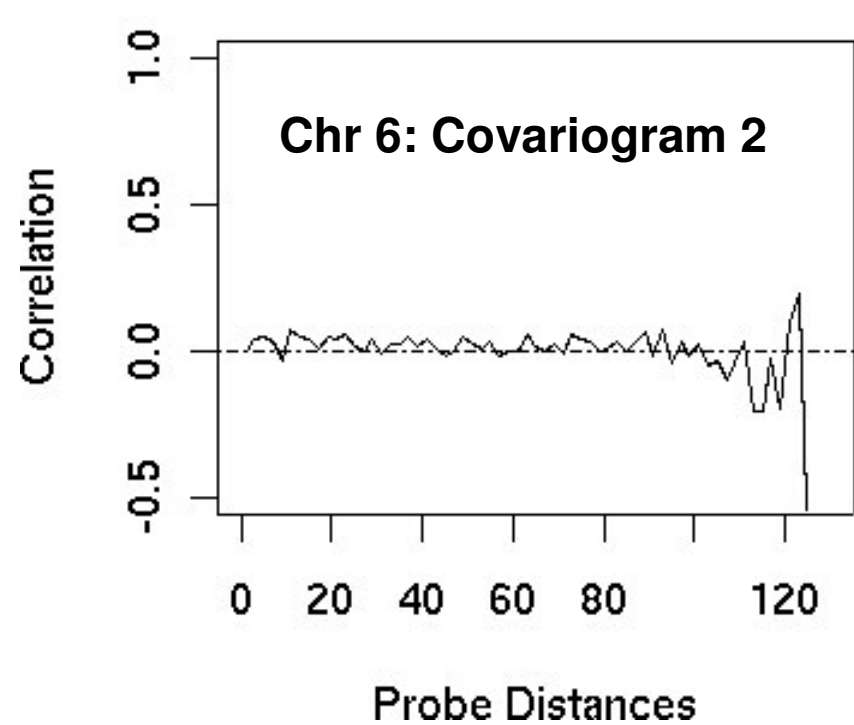
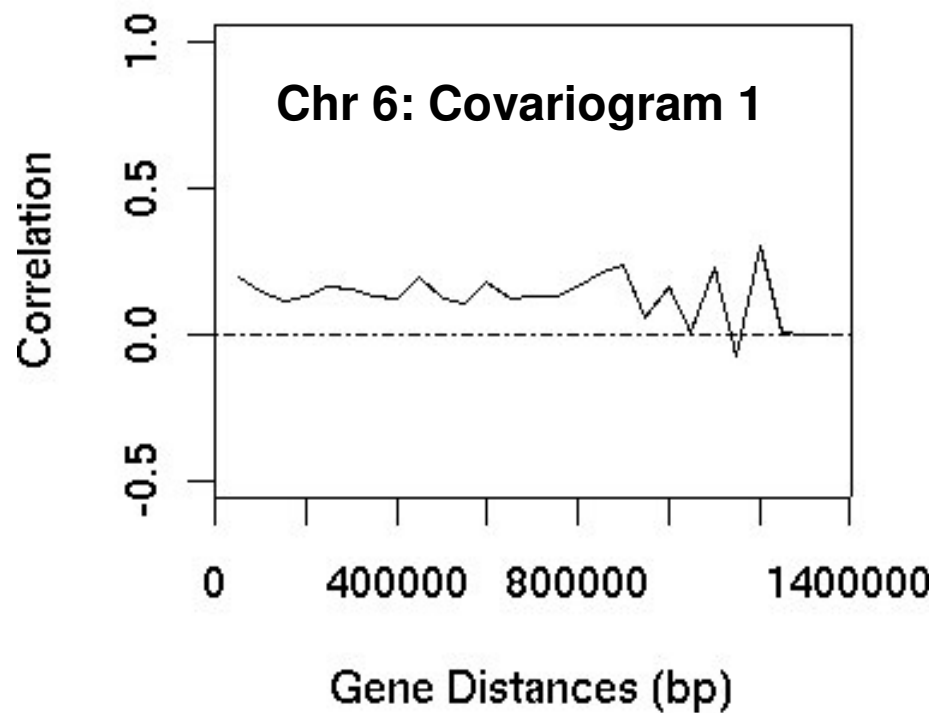
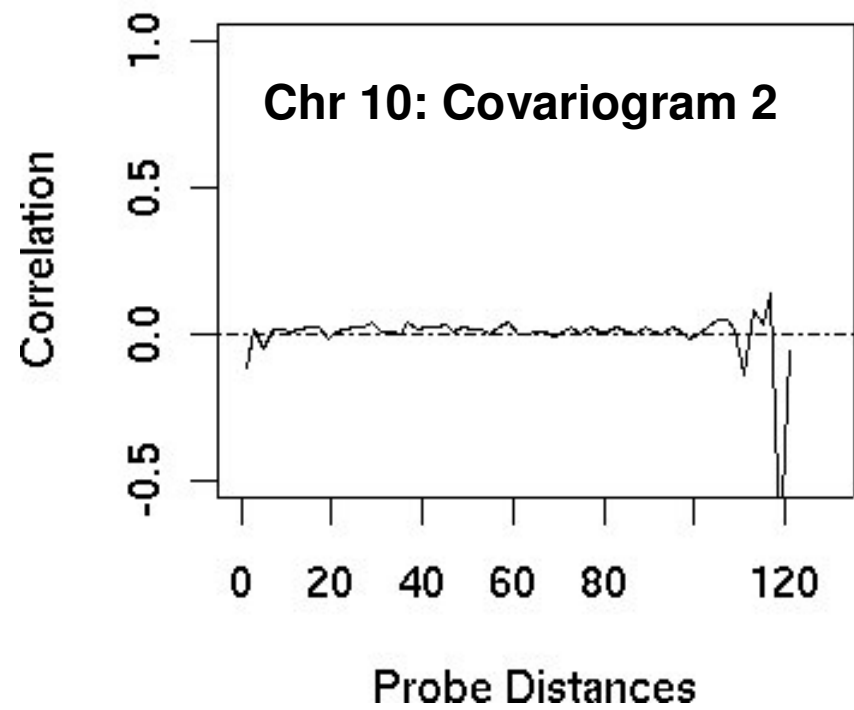
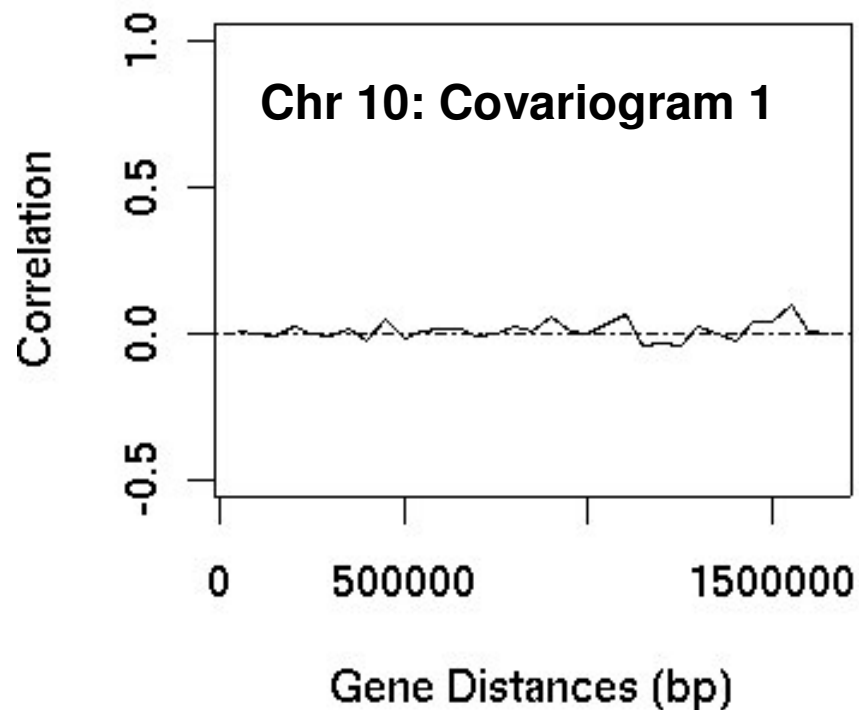
Methods: Data



Methods: Covariograms

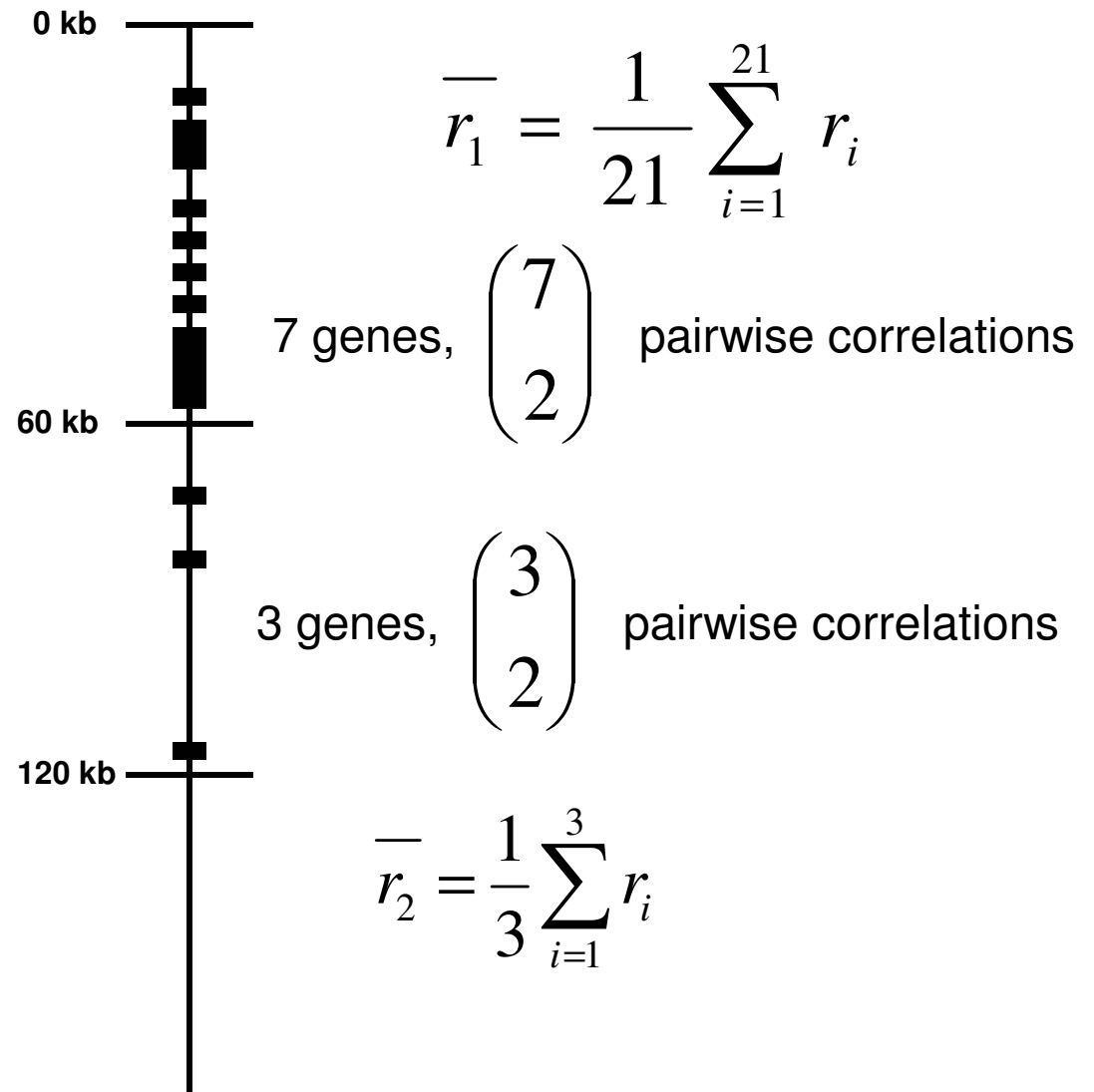
$$\gamma(x, y; d_a, d_b) = Ave[\rho(x, y \mid d_a \leq \text{dist}(x, y) < d_b)]$$

- Covariogram 1: distance is **chromosomal**
location: $d(g_i, g_j) = |g_{i, \text{midpt}(\text{chrloc})} - g_{j, \text{midpt}(\text{chrloc})}|$
- Covariogram 2: distance is **printed microarray**
location: $d(g_i, g_j) = \sqrt{((g_{i,x} - g_{j,x})^2 + (g_{i,y} - g_{j,y})^2)}$

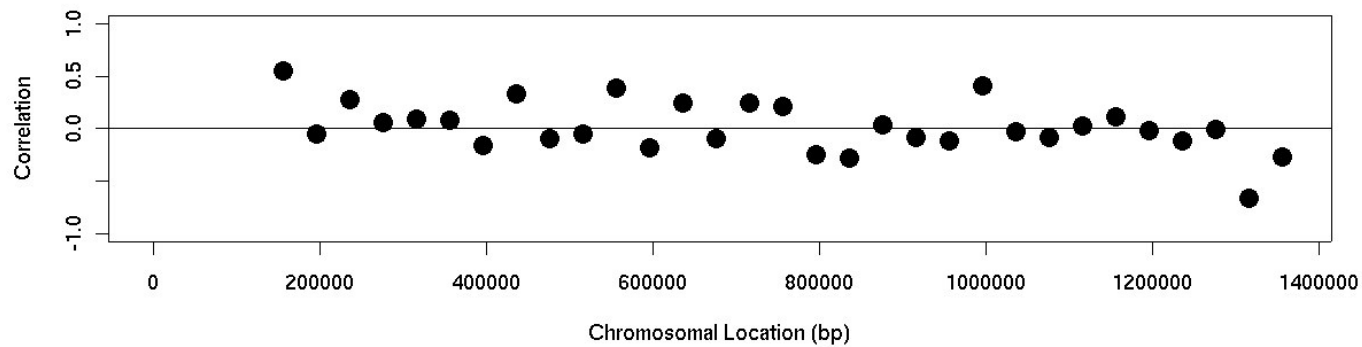


Methods: Partitioning

- Partition
- Avg of all pairwise Pearson correlations



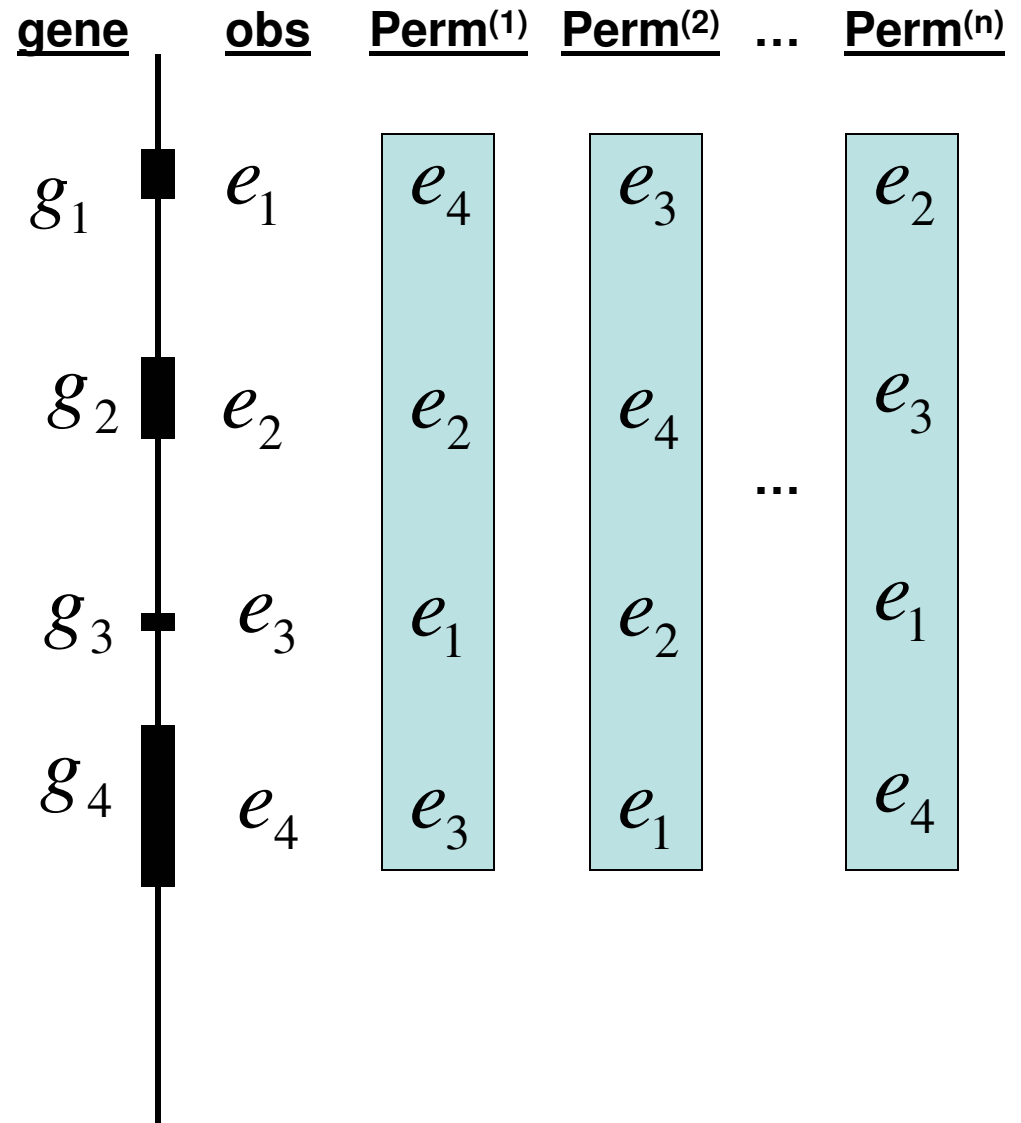
Methods: Partitioning



- Chr 6, 40 kb partition
- Significant?

Methods: Permutation Test

- $\bar{r} = .50$ in a 40kb interval on chr 6
- Permutation test
- Null distribution
- Estimated p-values



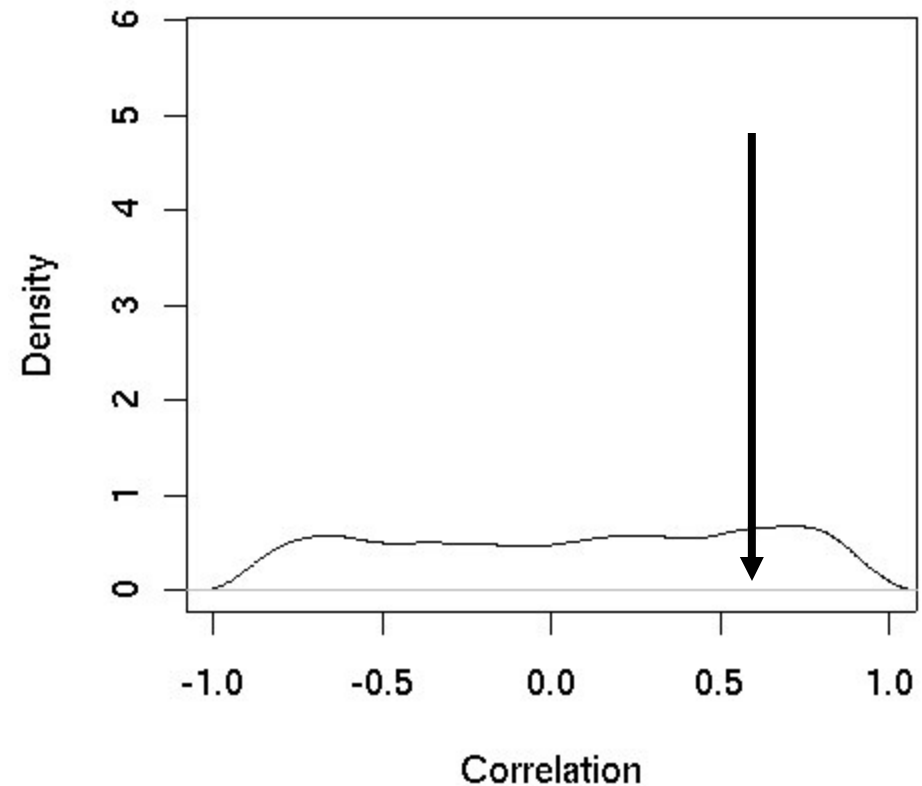
Methods: Permutation Test

- Distribution of $\overline{r}$ in 40 kb interval

$$\overline{r}_{obs} = 0.57$$

$$n_{genes} = 2$$

$$p - val = 0.22$$



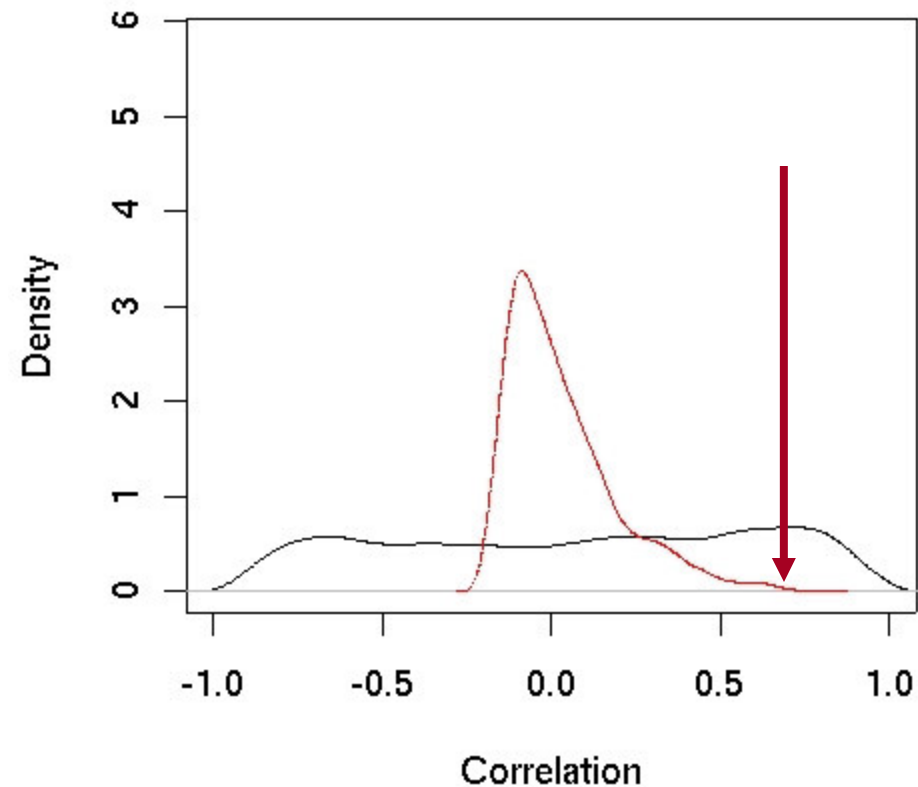
Methods: Permutation Test

- Distribution of $\overline{r}$ in 40 kb interval

$$\overline{r}_{obs} = 0.72$$

$$n_{genes} = 6$$

$$p - val \leq 0.001$$



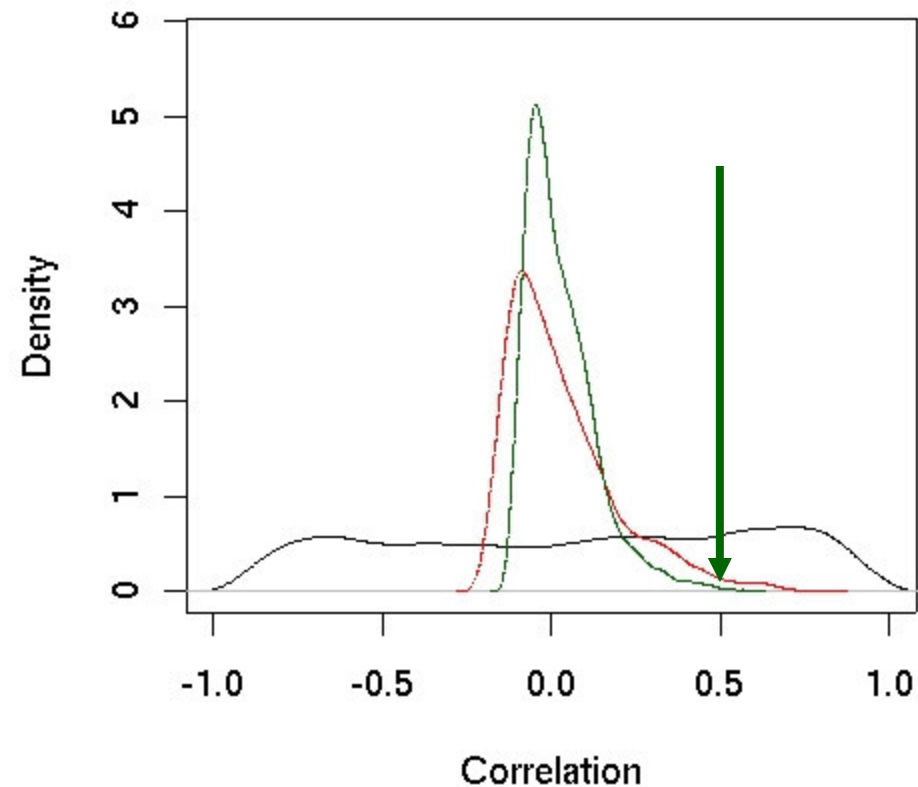
Methods: Permutation Test

- Distribution of $\bar{r}$ in 40 kb interval

$$\bar{r}_{obs} = 0.49$$

$$n_{genes} = 9$$

$$p - val = 0.002$$



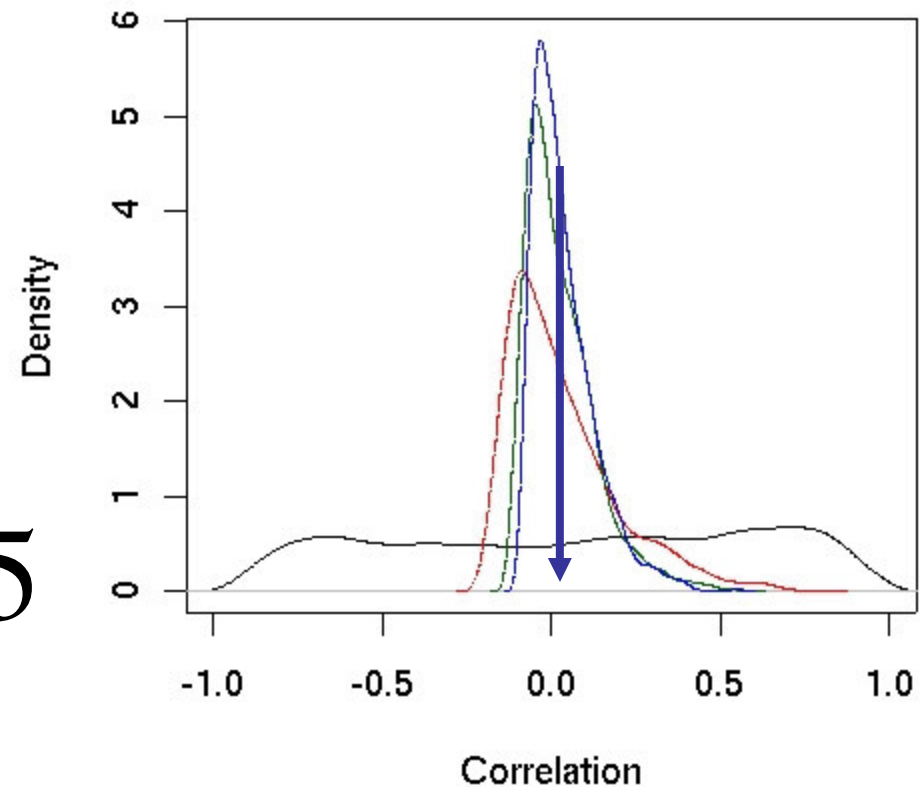
Methods: Permutation Test

- Distribution of $\overline{r}$ in 40 kb interval

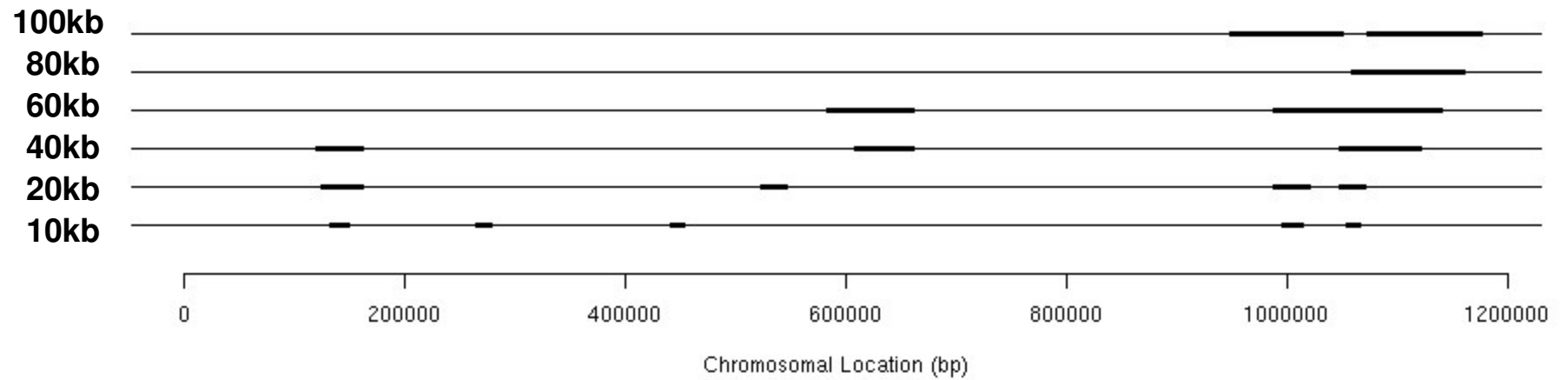
$$\overline{r}_{obs} = 0.018$$

$$n_{genes} = 12$$

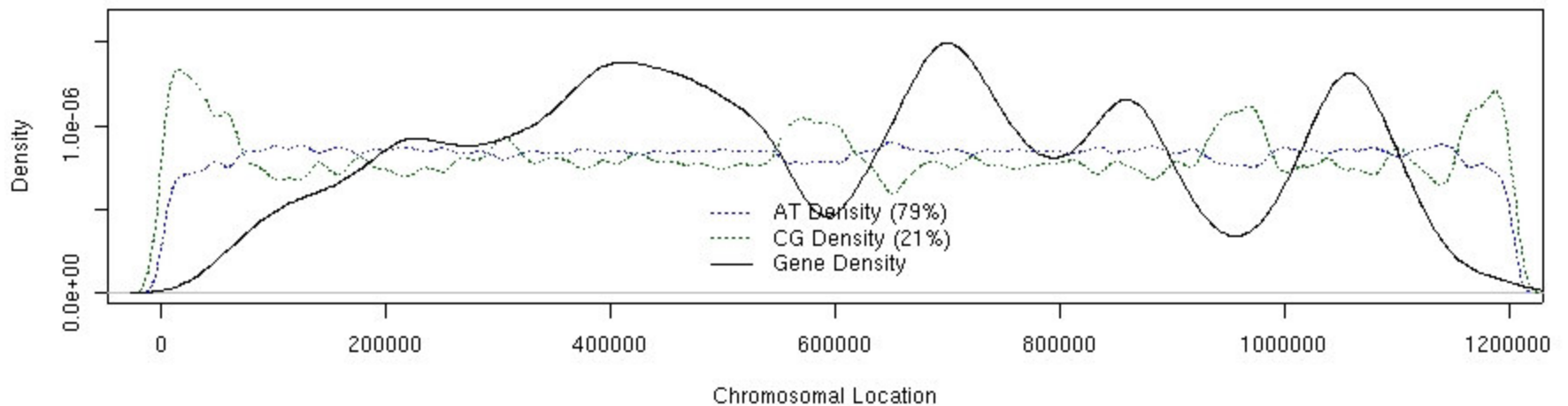
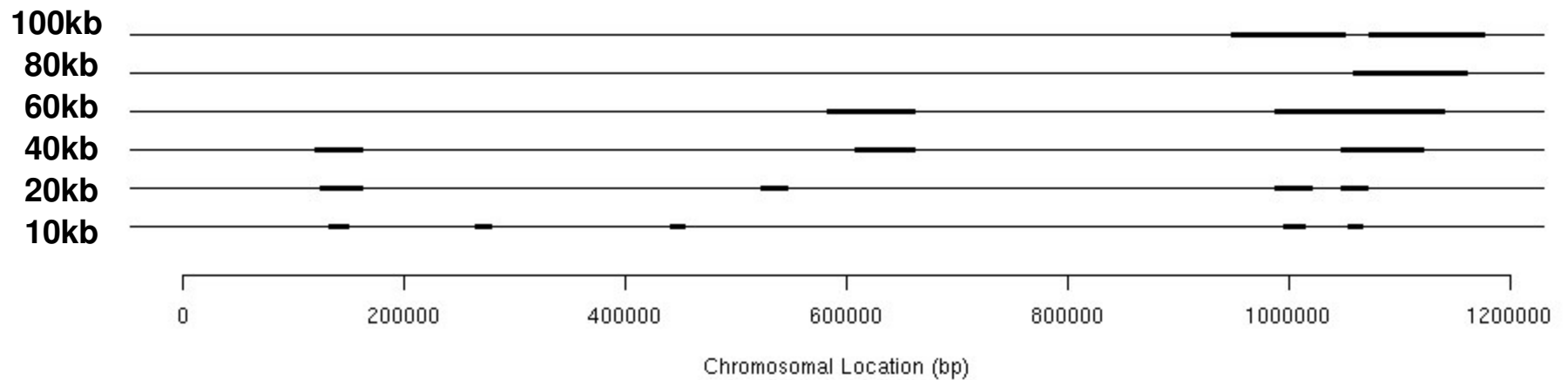
$$p - val = 0.475$$



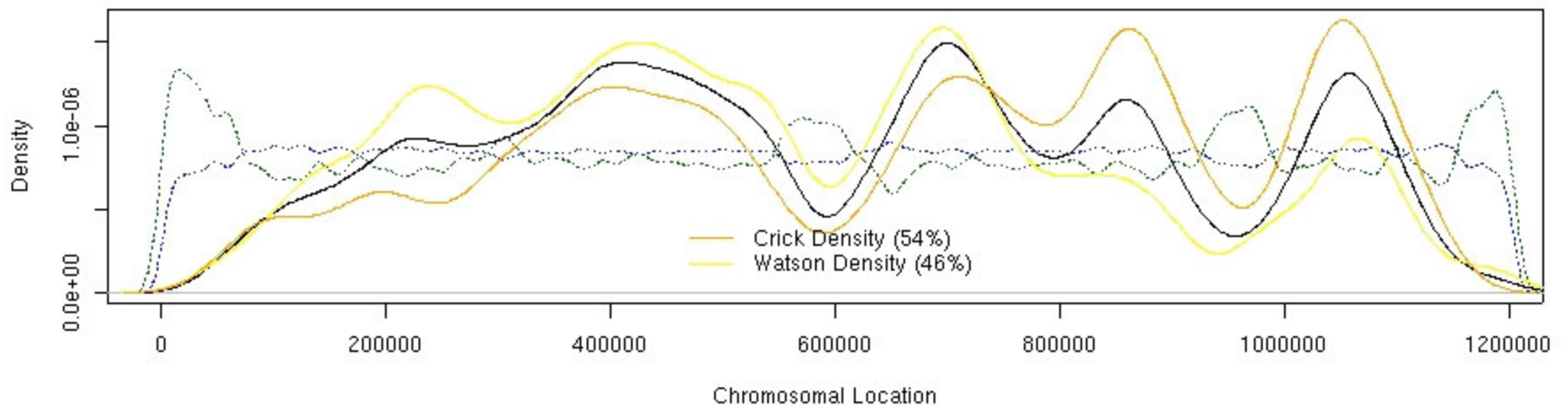
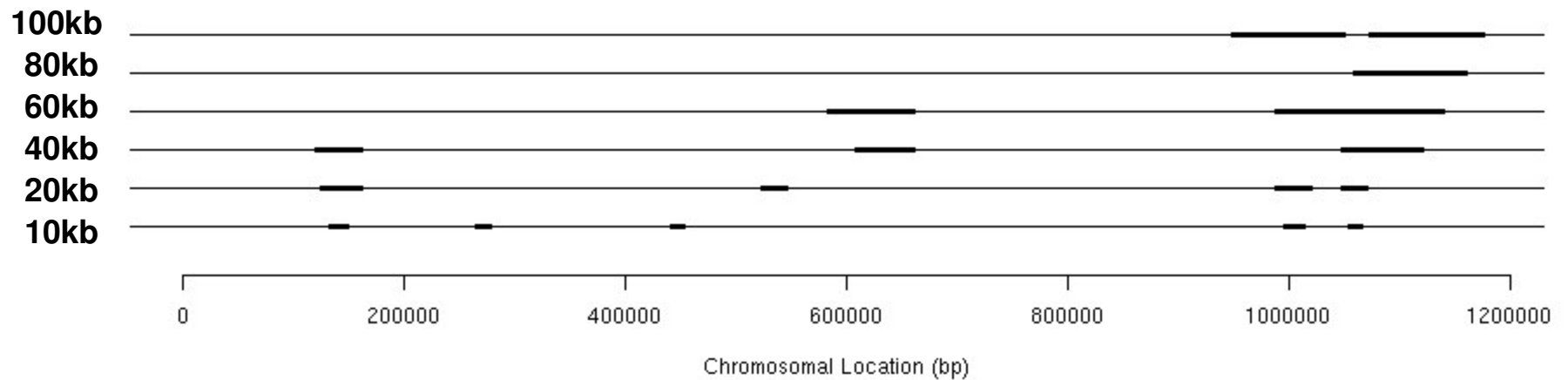
Significant Intervals (Chr 7)

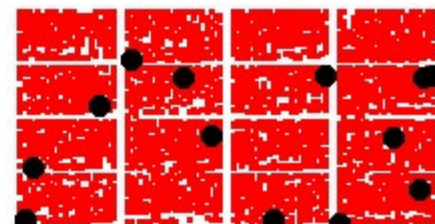
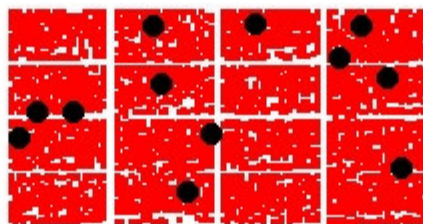
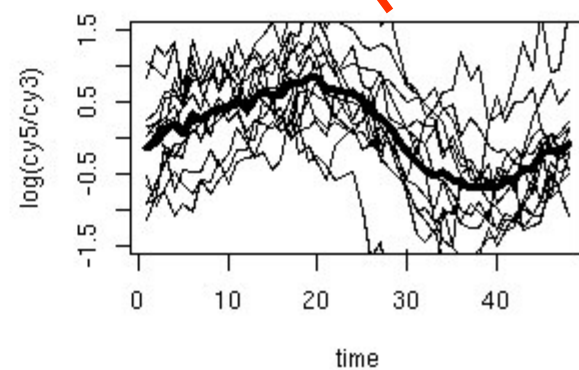
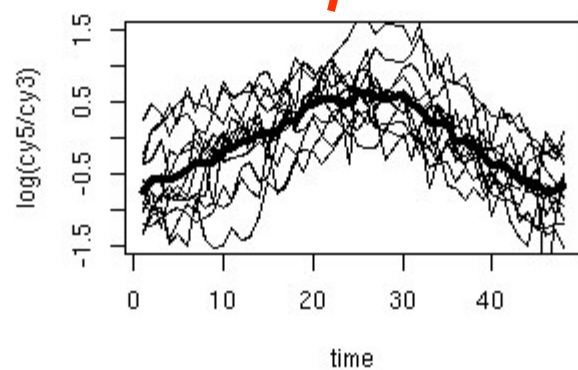
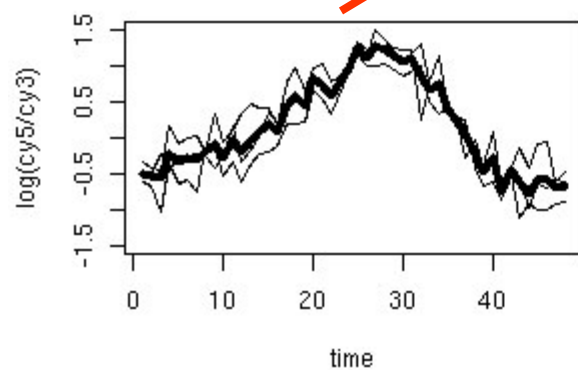
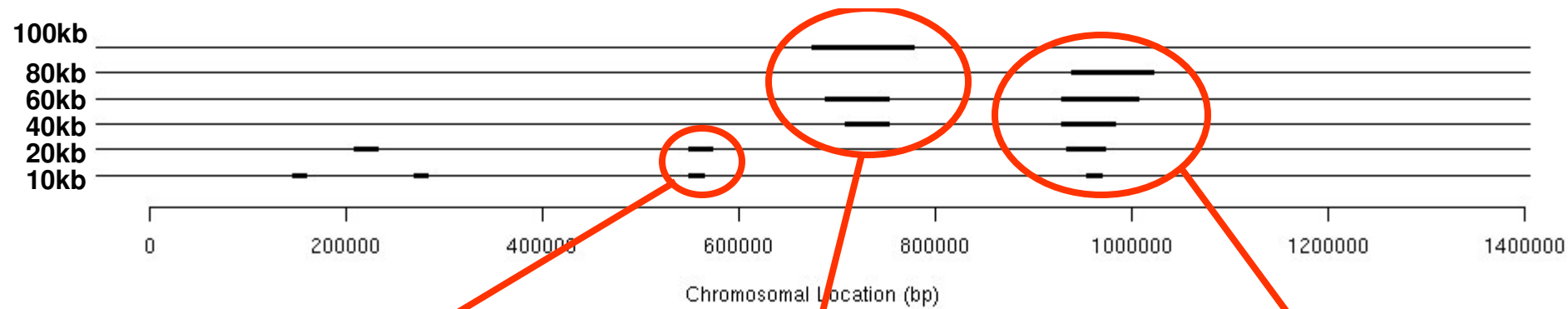


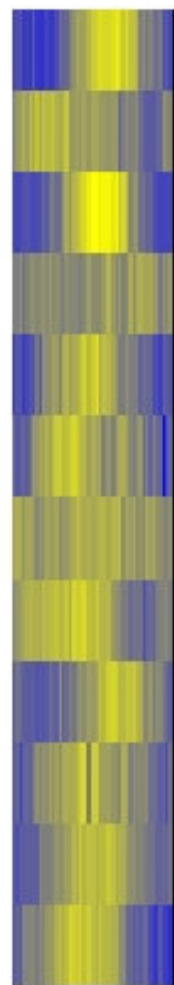
Significant Intervals (Chr 7)



Significant Intervals (Chr 7)







MAL6P1.257: hypothetical protein

MAL6P1.258: malate:quinone oxidoreductase

MAL6P1.259: hypothetical protein

MAL6P1.260: hypothetical protein

MAL6P1.263: hypothetical protein

MAL6P1.265: pyridoxine kinase

MAL6P1.266: hypothetical protein

MAL6P1.267: hypothetical protein

MAL6P1.268: hypothetical protein

MAL6P1.271: cdc2-like protein kinase

MAL6P1.272: ribonuclease

MAL6P1.273: hypothetical protein

10 40

Time (h)

Intervals (Chr 6)

p-val	Avg Cor	n _{genes}	Start kb	End kb	Size kb	Start Loc
0.003	0.86	3	550	560	10	0
0.004	0.86	3	550	570	20	10000
0.002	0.86	3	552.5	562.5	10	2500
0.003	0.27	14	675	775	100	75000
0.003	0.44	10	690	750	60	30000
0.004	0.39	9	710	750	40	30000
0.001	0.76	5	930	970	40	10000
0.001	0.57	8	930	990	60	30000
0	0.96	2	935	955	20	15000
0.002	0.64	5	940	980	40	20000
0.003	0.39	11	940	1020	80	60000
0.002	0.76	4	945	965	20	5000
0	0.51	9	945	1005	60	45000
0.002	0.76	4	950	970	20	10000
0.004	0.87	3	955	965	10	5000
0.002	0.87	3	957.5	967.5	10	7500

Results: Summary Table

	10kb	60kb	100kb	10kb in 60kb
Chr 3	3/400	0/68	0/40	0
Chr 4	10/476	5/80	2/48	4
Chr 5	6/528	1/88	3/56	0
Chr 14	4/1304	2/220	1/132	0

Conclusions

- Statistical: Significance for both **small regions of strong correlation** and **large regions of weak correlation**
- Biological: Evidence for regulation at multiple levels