Array Informatics

Mark Gerstein

CEGS Informatics Developing Tools and Technical Analyses Related to Genome Technologies

- Main Genome Technologies
  - ◊ Tiling Arrays
  - ◊ Next Generation Sequencing
- Main Applications
  - ◊ Transcript mapping
  - ◊ Protein-DNA Binding
  - ◊ CGH
- Transitioning to Seq....

Tools & Tech. Analyses for Processing of Genome Technology Data

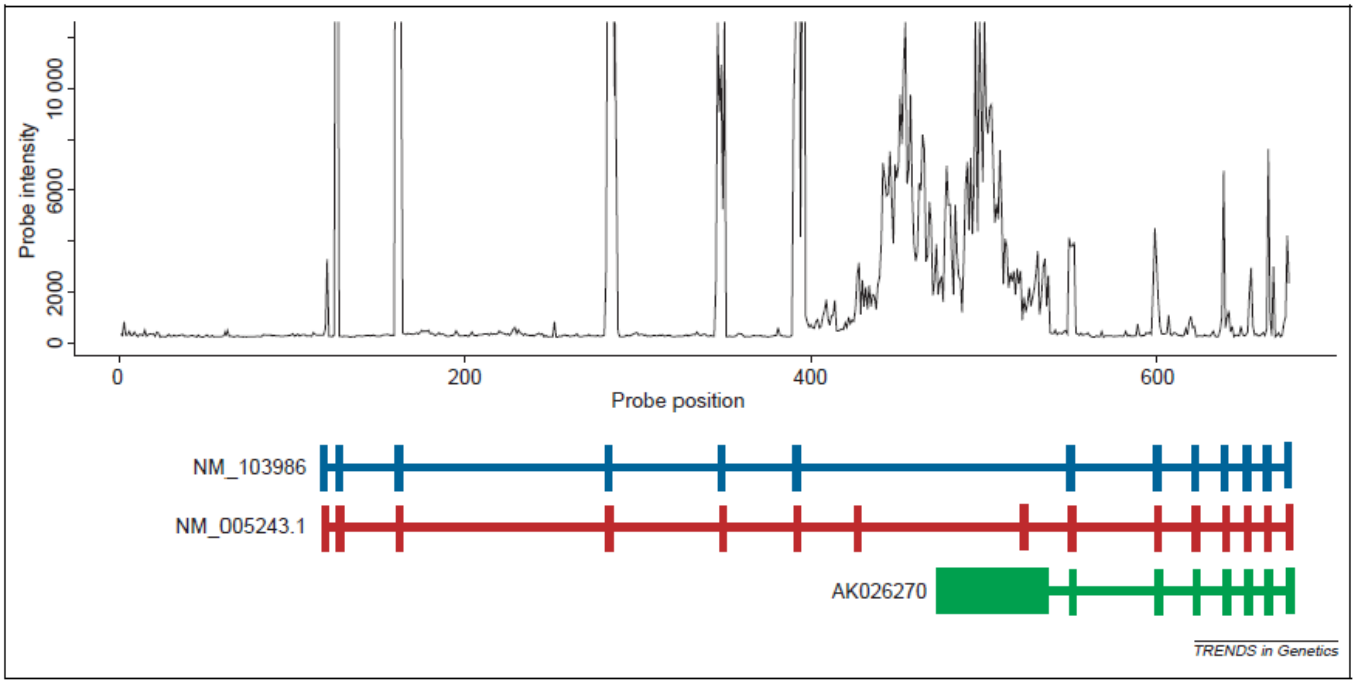
- Normalizing Arrays and Measuring & Correcting Artifacts
  - ◊ **COP** - Correcting positional artifacts [Yu et al. NAR '07]
  - ◊ **Efficient Pseudomedian** Calculation - for Tiling Array Scoring [Royce et al., BMC Bioinfo. '07]
  - ◊ **Measuring Mismatch Effects** [Seringhaus et al., BMC Genomics (submitted)]
  - ◊ **Removing Seq. Effects** [Royce et al., Bioinfo. '07]
  - ◊ **NN Prediction of Probe Intensity** - measuring & exploiting specific cross-hyb [Royce et al. NAR '07]
- Simulating NextGen Sequencing
  - ◊ **ChipSeqSim** - simulating ChIP Seq [Zhang et al., PLoS CB '08]

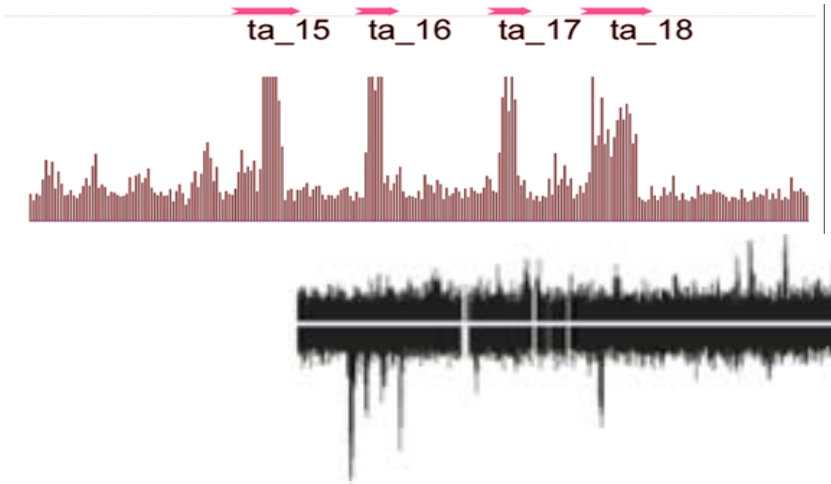
Tools & Tech. Analyses for Genome Structural Variation

- **Breakptr** - HMM-based Array Segmentation for CNV detection [Korbel et al., PNAS '07]
- **MSB** - Mean-shift-based Array Segmentation for CNV detection with extension to sequencing [Wang et al. Gen. Res. (submitted)]
- **PEMer** - Paired-end Mapping for SV Dectection with simulation calibration and breakpoint DB [Korbel et al., GenomeBiol. (submitted)]
- **Long-SV-Assembly** Simulations [Du et al., Nat. Meth. (submitted)]
- **SD-CNV-CORR** - Approach for correlating the occurrence of CNVs and SDs with genomic features (particularly repeats) [Kim et al., Genome Res. (submitted)]

A Starting Point: Noisy Raw Signal from Tiling Arrays (Transcription)

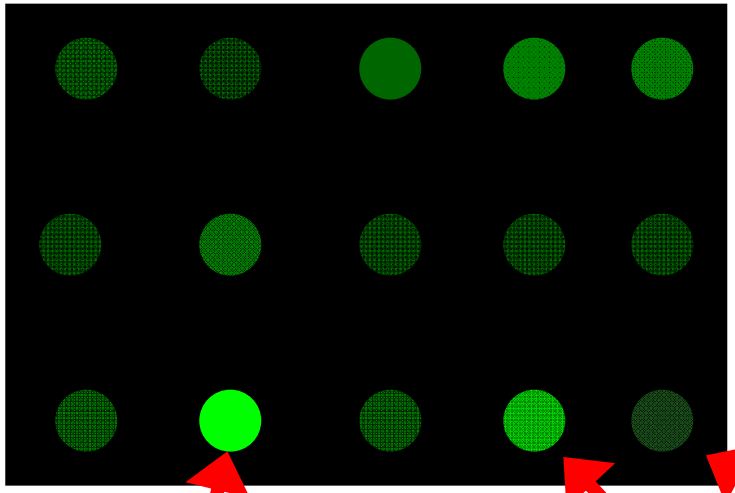
Johnson et al. (2005) TIG, 21, 93-102.



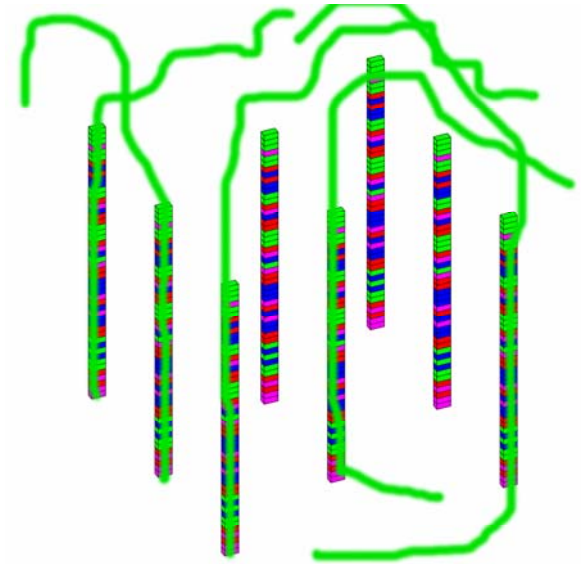


Li et al., PLOS one (2007)

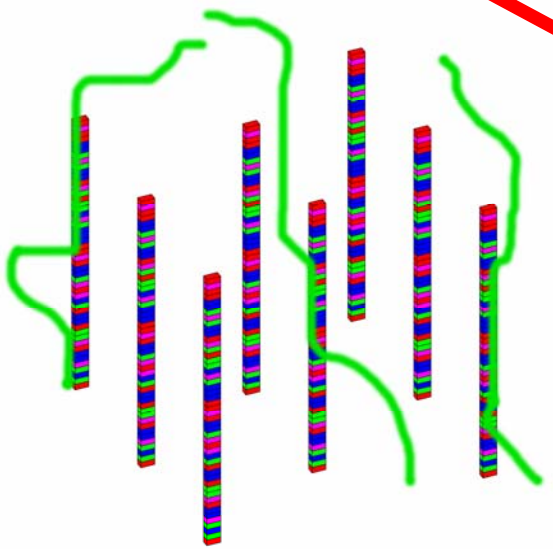
Specific & Non-specific Cross-Hyb.



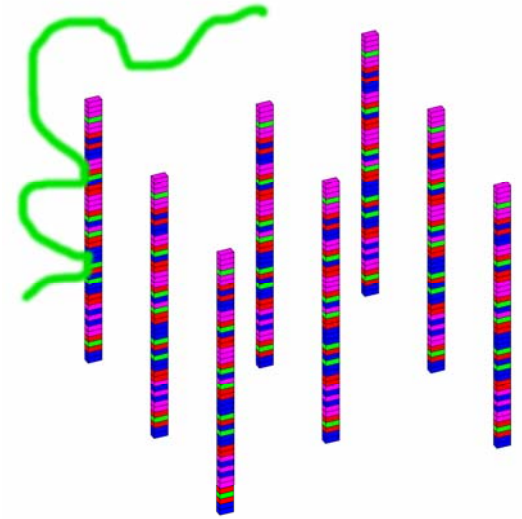
- Perfect match (PM):
  probe binding intended target
- Specific cross-hyb.: probes binding non-PM targets with a small number of mismatches
- Non-specific cross-hyb.: probes binding targets with many mismatches, due to general stickiness of oligos



Perfect Match



Specific Cross-hyb.

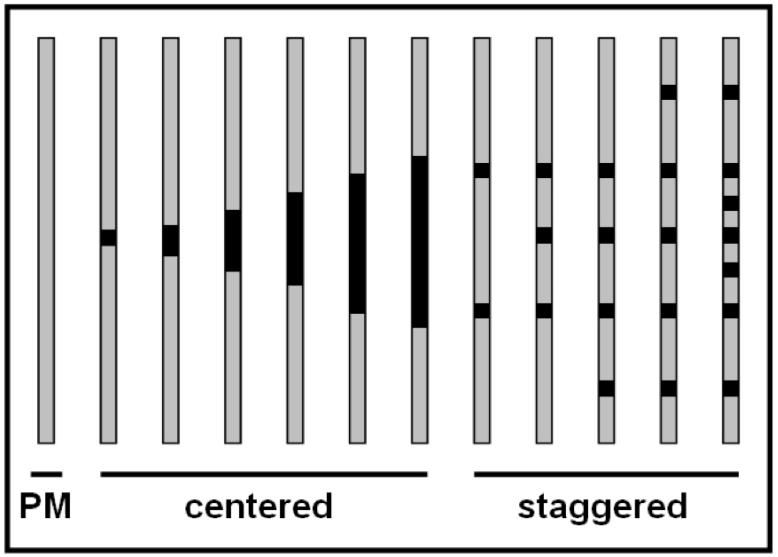


Non-specific Cross-hyb.

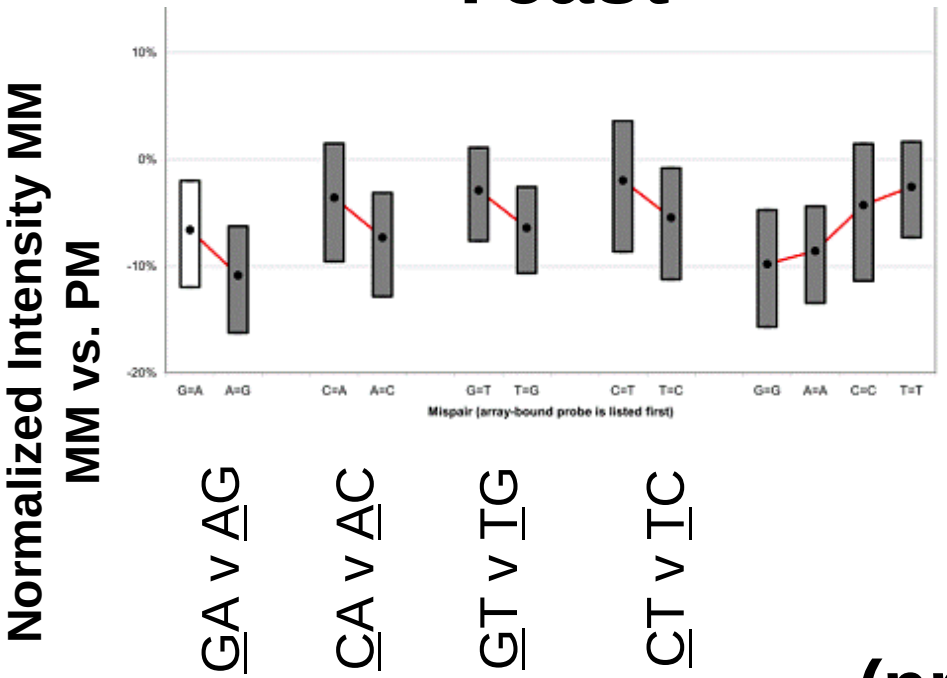
Non-Specific Cross Hyb. (Sequence Effects)

Creation of Standardized Datasets for Quantifying Effect of Mismatches

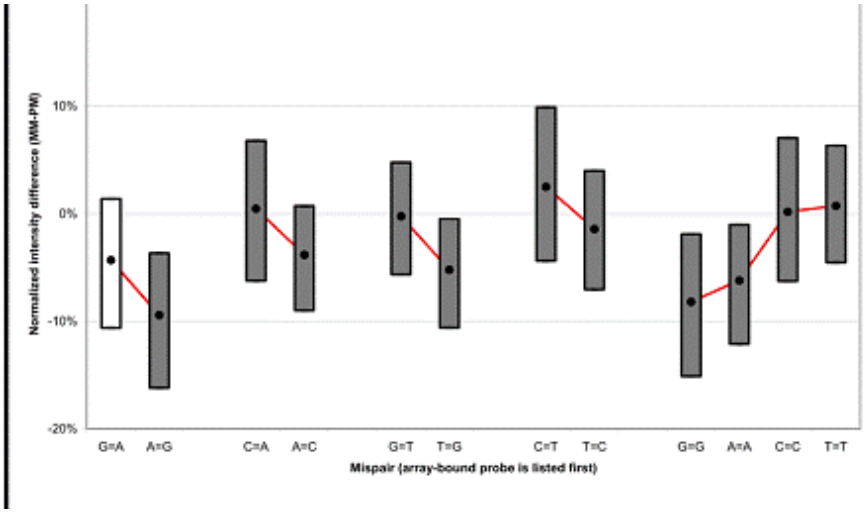
[Seringhaus et al., BMC Genomics (in press)]



Yeast

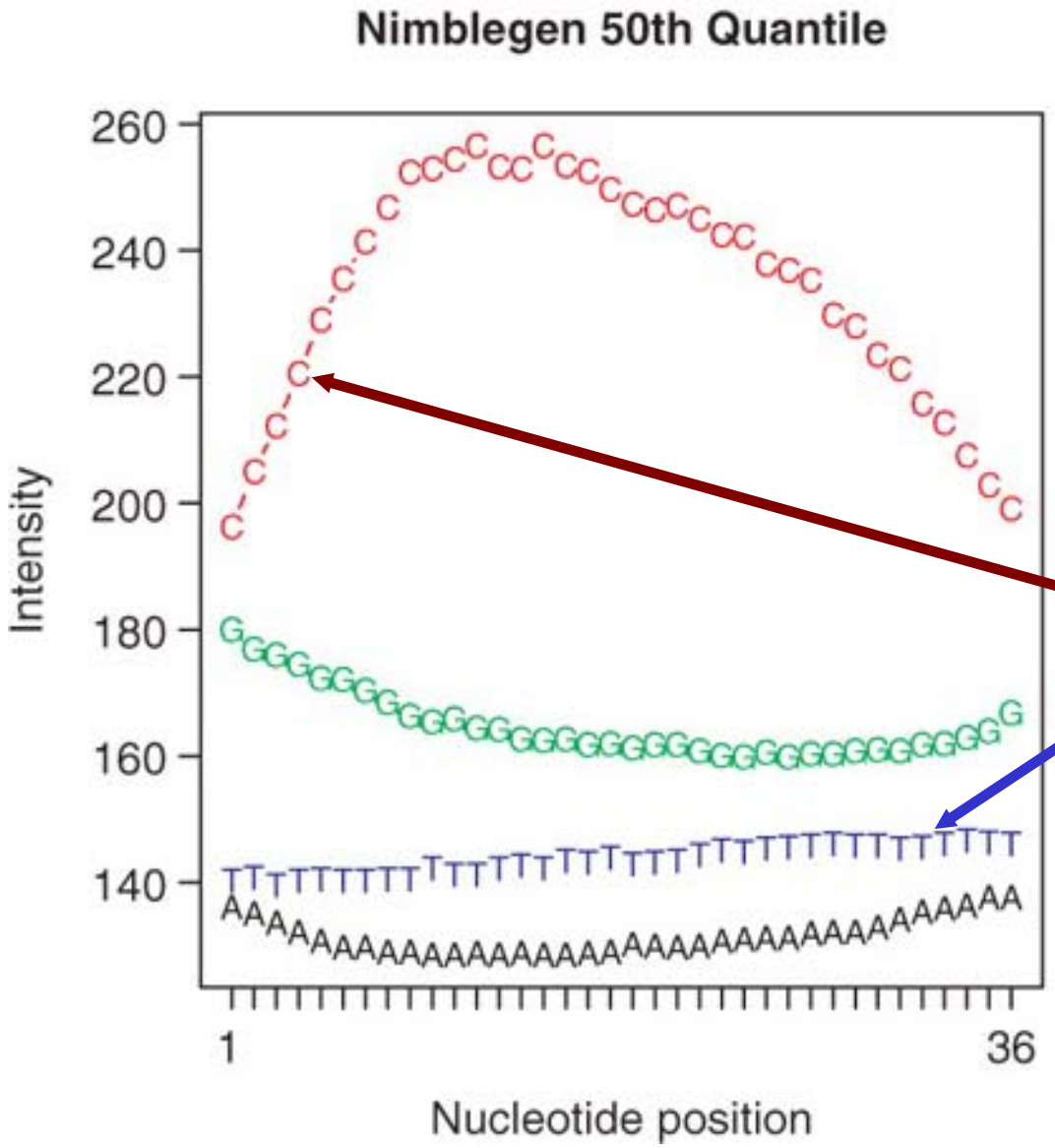


Human



Types of Mispairs (probe on array is first)

Observing Non-specific Cross-hyb. (Probe sequence effects)

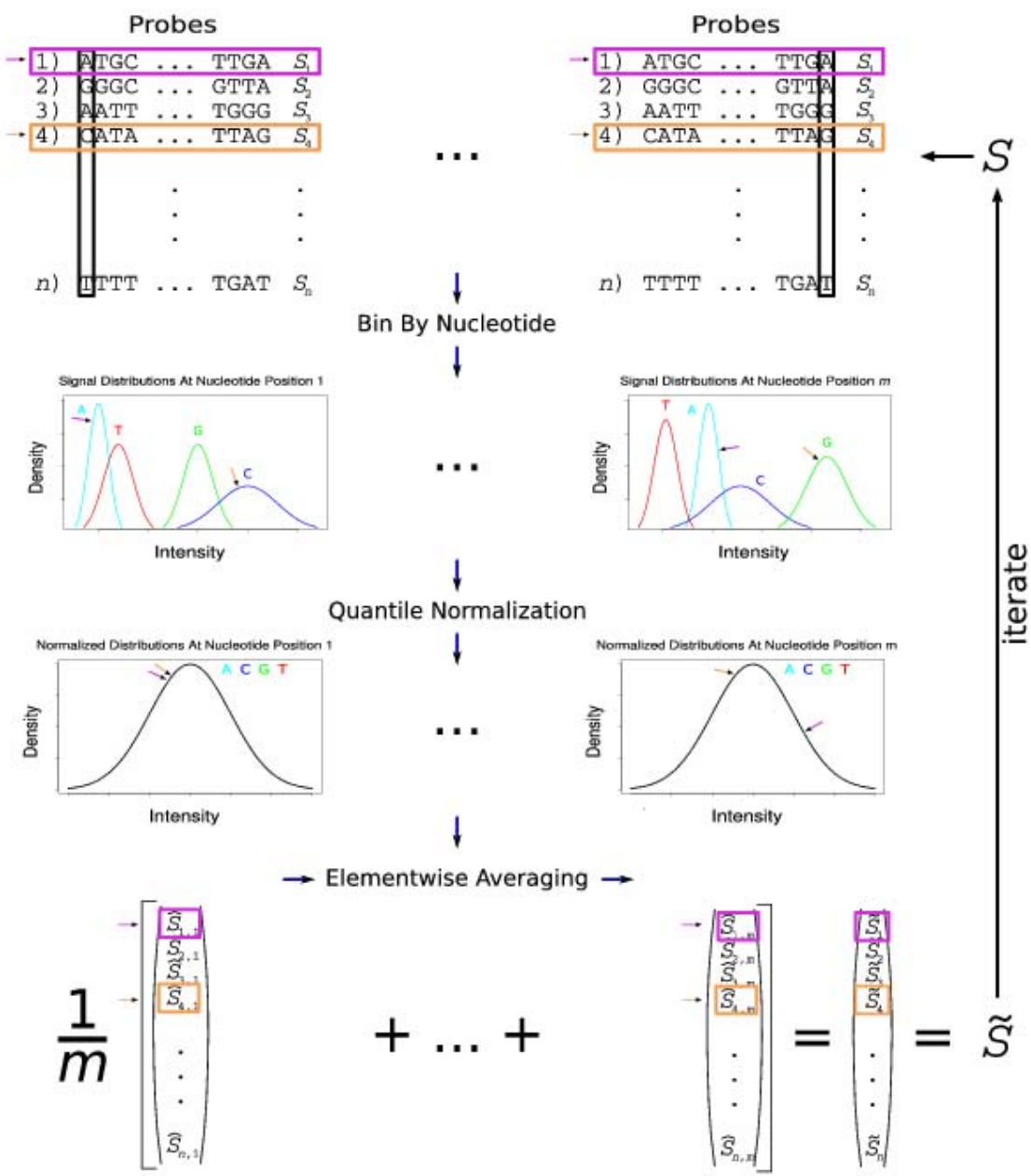


Avg. intensity of all background probes with a C at position 4

Avg. intensity of all background probes with a T at position 33

Source: Royce, T.E., et al (2007), *Bioinformatics*, **23**, 988-97

Iterated Quantile Normalization to Correct for Non-specific Cross-hyb.



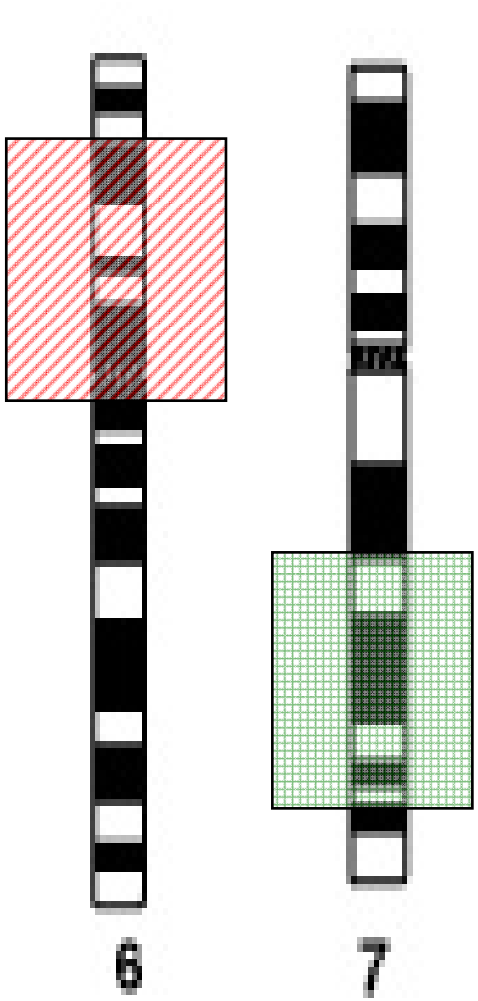
- Adapt Bolstad et al (2003) approach to tiling arrays
- Force distributions with a given nt at each position to be same
- Distributions at other positions now different so iterate
- Also, robust adaptation of Naef & Magnasco (2003)

T Royce et al (2007), *Bioinformatics*, **23**, 988-97

Measuring Specific Cross-Hyb

Source: Royce, T.E., et al (2007), *Nucleic Acids Res.*, **23**, 98-97

Proof of principle test to "exploit" this

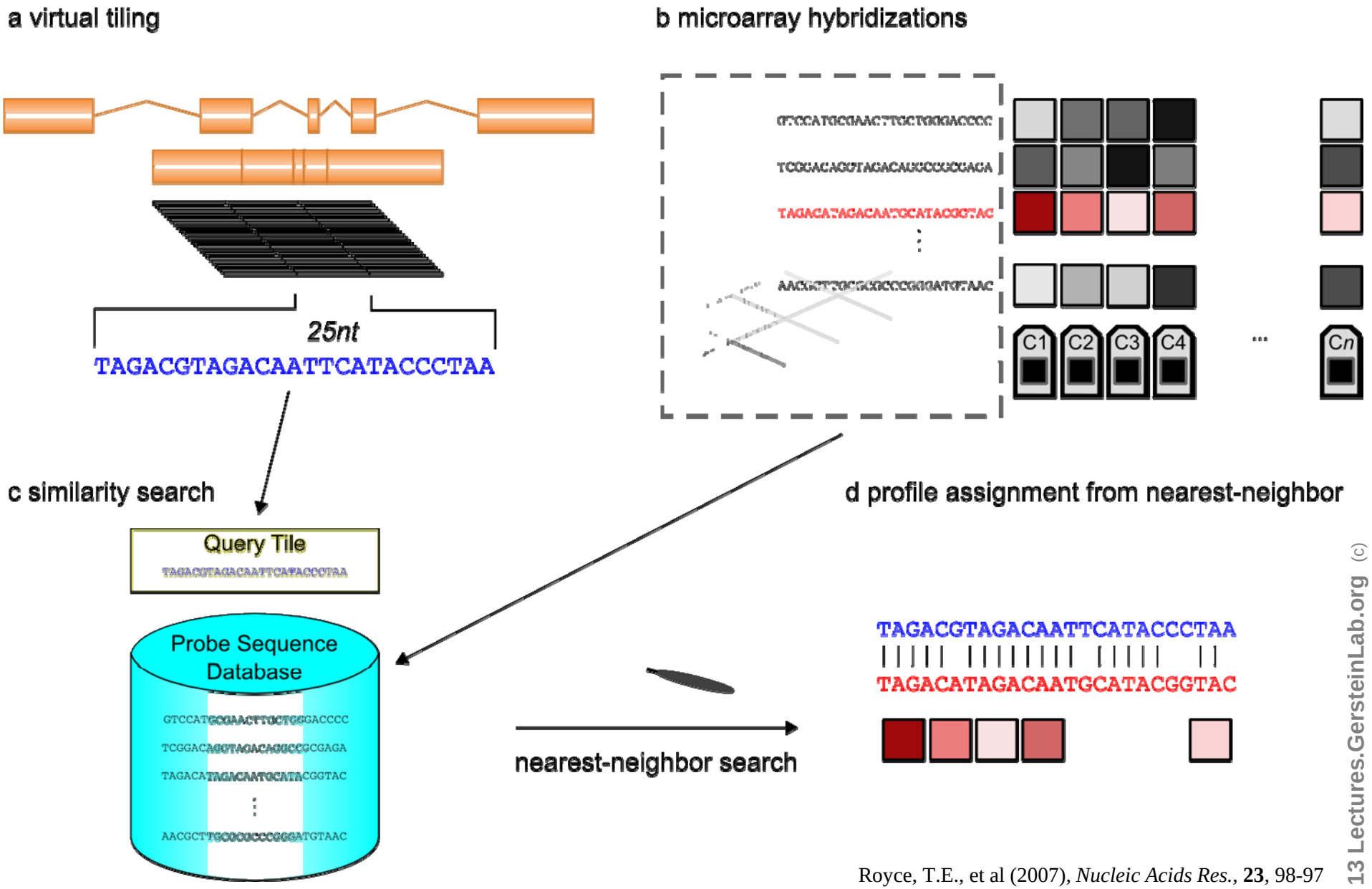


- Using Cheng et al. (2005), predict gene expression levels (and profiles across tissues) for genes on part of chr. #6
- ...Based on closest cross-hyb tiles on part of chr. #7
- Then compare to measured levels and profile on #6

Figure from http://www.members.cox.net/amgough/Fanconi-genetics-genetics-primer.htm

Source: Royce, T.E., et al (2007), *Nucleic Acids Res.*, **23**, 98-97

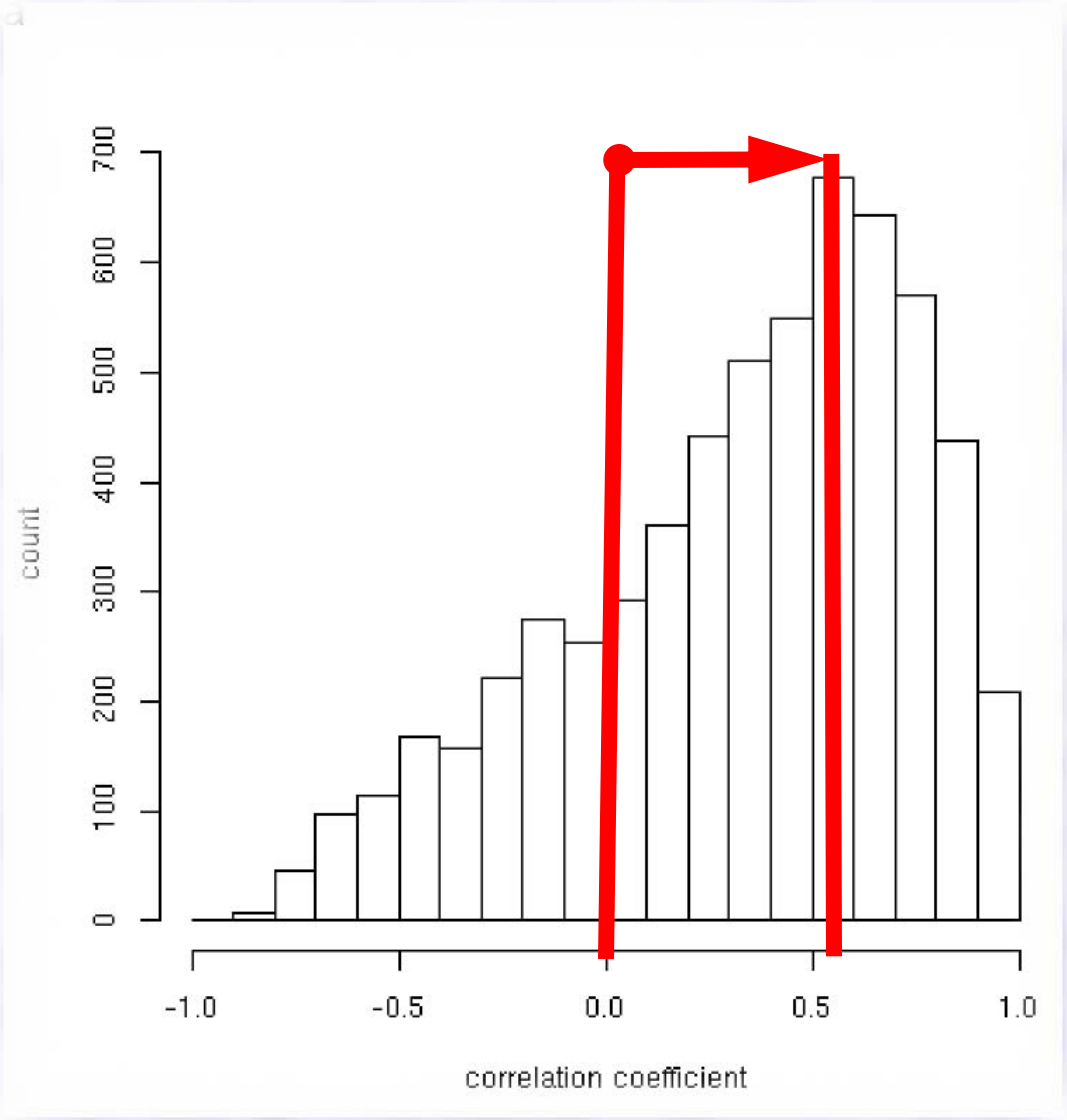
Nearest Nbr Search on Virtual Tiling



Royce, T.E., et al (2007), *Nucleic Acids Res.*, **23**, 98-97

Agreement between predicted tile expression profile and actual one

- Correlated predicted profiles with the actual profiles of gene expression across cell lines
- Much more correlation than expected by chance (dist. centered on 0)

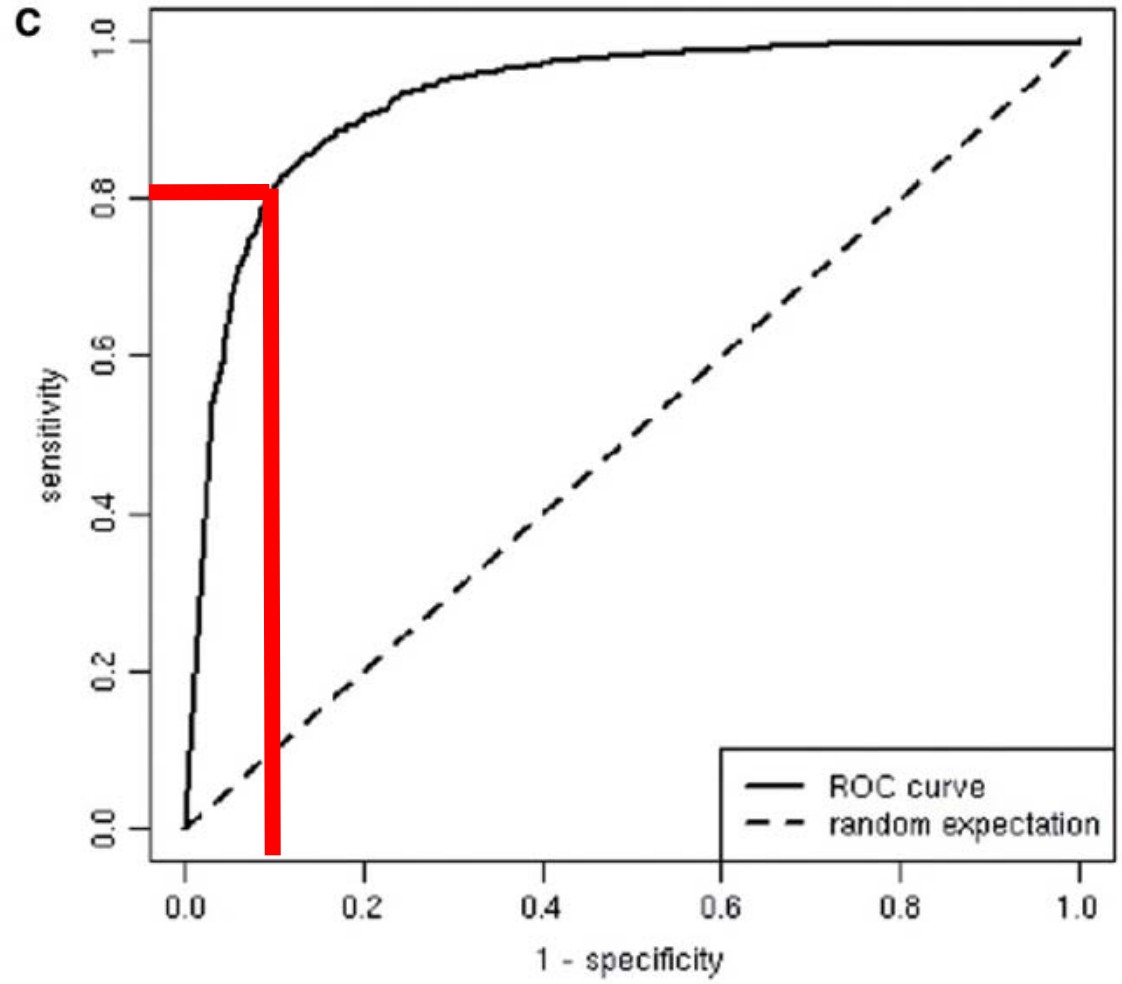


Source: Royce, T.E., et al (2007), *Nucleic Acids Res.*, **23**, 9

Very Strong ROC Curve: Most genes are accurately detected using nearest-neighbor features' signals

- **Illustrates great magnitude of cross-hyb. on hi-density arrays**
- High feature density arrays inadvertently resurrecting generic n-mer concept (van Dam & Quake, 2003)
- Suggests that tiling arrays could be exploited to create **universal arrays**

- Gold std. set of known expressed genes. How well do we find.
- A set of known positives was defined as the Refseq genes with at least 75% transfrag coverage. A set of known negatives was constructed by permuting the sequences in the set of known positives. For various thresholds, sensitivity and specificity were computed and then plotted.



Royce, T. E. et al. Nucl. Acids Res. 2007 35:e99

CEGS Informatics Credits

- Array Corrections
  - ◊ J Rozowsky
  - ◊ T Royce
  - ◊ M Seringhaus
- PEMer, SD-CNV, BreakPtr
  - ◊ P Kim
  - ◊ J Korbel
  - ◊ J Du
  - ◊ X Mu
  - ◊ A Abyzov
  - ◊ N Carriero
- Experimental
  - ◊ M Snyder
  - ◊ S Weissman
  - ◊ A Urban