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 Detection 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
Non-Specific Cross Hyb.
(Sequence Effects)
Creation of Standardized Datasets for Quantifying Effect of Mismatches

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

Types of Mispairs
(probe on array is first)

Yeast

Human

<table>
<thead>
<tr>
<th>MM vs. PM</th>
<th>Normalized Intensity MM</th>
</tr>
</thead>
<tbody>
<tr>
<td>G v A</td>
<td>A v G</td>
</tr>
<tr>
<td>C v A</td>
<td>A v C</td>
</tr>
<tr>
<td>G v T</td>
<td>T v G</td>
</tr>
<tr>
<td>G v T</td>
<td>T v C</td>
</tr>
<tr>
<td>G v C</td>
<td>C v G</td>
</tr>
<tr>
<td>A v C</td>
<td>C v A</td>
</tr>
<tr>
<td>G v C</td>
<td>C v T</td>
</tr>
<tr>
<td>G v T</td>
<td>T v C</td>
</tr>
<tr>
<td>G v T</td>
<td>T v C</td>
</tr>
<tr>
<td>G v C</td>
<td>C v G</td>
</tr>
<tr>
<td>A v C</td>
<td>C v A</td>
</tr>
<tr>
<td>G v C</td>
<td>C v T</td>
</tr>
<tr>
<td>G v T</td>
<td>T v C</td>
</tr>
<tr>
<td>G v T</td>
<td>T v C</td>
</tr>
<tr>
<td>G v C</td>
<td>C v G</td>
</tr>
<tr>
<td>A v C</td>
<td>C v A</td>
</tr>
</tbody>
</table>

PM centered staggered
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

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)

Measuring Specific Cross-Hyb

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

Nearest Nbr Search on Virtual Tiling

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)

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**

---

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