The Quest for functional ncRNAs

Kristin Reiche
RNomics group
Fraunhofer Institute for Cell Therapy and Immunology

October 2009
Pervasive transcription in mammalian genomes

DNA

ncRNA → snoRNA → protein-coding mRNA → 3' 5'
short RNAs → antisense ncRNA

Kristin Reiche
Quest for RNAs
Known functions of small and long ncRNAs

- Translational silencing: miRNAs, siRNAs (RNA interference)
- Epigenetic control:
  - Recruitment of chromatin modifying complexes by *HOTAIR*, and several lincRNAs
  - X-chromosome inactivation by *XIST*
- Transcriptional regulation:
  - Examples of ncRNAs controlling promoter usage by preventing the binding of transcription factors
  - Co-activators for transcription factors like Evf2 RNA for DLX2
- Long transcripts that generate many small RNAs:
  - RNA duplex of *XIST* and *Tsix* generates small interfering RNAs
  - *GAS5* generates small nucleolar RNAs in its introns (mouse)
- Guiding chemical modifications of other RNAs: snoRNAs
- mRNA splicing: snRNAs
- Specific expression patterns (developmental stages, tissues)
How to identify functional ncRNAs?

- **Computational methods:**
  - Known class: Homology search or classification (tRNAscan-SE, RNAmicro, etc.)
  - Novel ncRNAs: Conserved in sequence and/or structure are evidences for function (blast, RNAz, Evofold, etc.)

- **Experimental strategies:**

  ![Diagram showing steps for identifying functional ncRNAs]

  1. Manipulate tumor relevant pathways in cell line models of oncologic diseases
  2. Identify regulated ncRNAs
     - Tiling arrays – long ncRNAs
     - Transcriptome sequ. – short RNAs
  3. Develop custom arrays to allow efficient detection
Experimental strategies (long transcripts)

**Tiling Arrays:**
- Cover the whole human genome by 25-mer probes at a spacing of approximately 35 bases in an unbiased fashion.
- Detection of novel ncRNAs
- Detection of different splice variants of protein-coding gene
- Variation of the whole transcriptome is monitored at a high level of resolution.
- Expensive technique, hence in general no replicates possible
- No strand information

**Custom microarrays:**
- (High-throughput) validation of regions differentially expressed in tiling arrays (run in triplicates, rather good estimates of variance)
- Detection of reading direction (each transcript spotted in sense and antisense)
General outline of experimental strategies (long transcripts)

Tiling arrays (1 for each condition)

Putative regulated transcripts ...

... are summarized on a custom microarray

Validation in context of various conditions (3 for each condition)
Genome-wide expression studies for two essential pathways (p53 induced, STAT-3 induced) and one basal cellular mechanism (cell cycle)

<table>
<thead>
<tr>
<th></th>
<th>Cell line</th>
<th>Tiling Array</th>
<th>Custom Array</th>
</tr>
</thead>
<tbody>
<tr>
<td>p53</td>
<td>HFF</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>STAT-3</td>
<td>INA6</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Cell Cycle</td>
<td>HFF</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

HFF: human foreskin fibroblast cells (no chromosomal aberrations, normal set of chromosomes). INA6: myeloma cell line

- p53:
  - induces growth arrest
  - activates DNA repair
  - initiates apoptosis

- STAT-3:
  - suppresses apoptosis
  - leads to cellular proliferation

Kristin Reiche  
Quest for RNAs
Results of tiling arrays: Overall numbers

- **Controlled transcripts** = In at least one condition high expressed and differentially expressed
- Highest number of controlled transcripts in cell cycle. (expected as it is a basal cellular mechanism)
- Less transcripts controlled by p53 and STAT-3
- Specific expression of transcripts (however overlap of transcripts regulated by p53 and in cell cycle)

<table>
<thead>
<tr>
<th></th>
<th>Cell Cycle</th>
<th>p53</th>
<th>STAT-3</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>all transcripts</td>
<td>13601</td>
<td>793</td>
<td>36</td>
<td>14440</td>
</tr>
<tr>
<td>(at least 50%</td>
<td>6981</td>
<td>61</td>
<td>616</td>
<td>7845</td>
</tr>
<tr>
<td>overlap) - 10</td>
<td>3542</td>
<td>72</td>
<td>8</td>
<td>3622</td>
</tr>
<tr>
<td>controlled in all</td>
<td>1299</td>
<td>15</td>
<td>158</td>
<td>1386</td>
</tr>
<tr>
<td>intergenic</td>
<td>542 (26%)</td>
<td></td>
<td></td>
<td>181</td>
</tr>
<tr>
<td>transcripts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(at least 50%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>overlap) - 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>controlled in all</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Kristin Reiche  
Quest for RNAs
Most transcripts differentially expressed between resting phase G0 and G1

However, less pronounced observation for intergenic transcripts
Results: Genomic distribution

- Specific differential expression in genomic regions
- Pervasive transcription in 3’UTRs (compared to 5’UTRs)
- Most transcripts controlled by p53 are found in introns
Results: Transcription in 3’UTRs?

- Observed differential transcription just due to different length of UTRs?
- 3’ UTRs in general much longer than 5’ UTRs
Results: Transcription in 3’UTRs?

- What we need: Statistical test if the number of observed transcripts in UTRs is larger than expected for random genomic regions of the same length.
- How to choose random regions: No repeats, same length as UTRs (on average), and uniformly distributed among genome.
- Applied Wilcoxon Rank Sum Test: Observation of increased differential transcription in 3’UTRs more pronounced than in 5’UTRs.

<table>
<thead>
<tr>
<th></th>
<th>Cell cycle</th>
<th>p53</th>
<th>STAT-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>5’UTR</td>
<td>$p \approx 0.03$</td>
<td>$p \approx 0.03$</td>
<td>$p \approx 0.01$</td>
</tr>
<tr>
<td>3’UTR</td>
<td>$p &lt; 0.0001$</td>
<td>$p &lt; 0.0001$</td>
<td>$p &lt; 0.0001$</td>
</tr>
</tbody>
</table>

$p$-values according to 1000 random intervals. At least 99% overlap required.
Results: Transcription in introns

- Increased number of transcripts in introns than expected by chance for cell cycle and p53
- Pervasive transcription of exons (as expected for tiling arrays)

<table>
<thead>
<tr>
<th></th>
<th>Cell cycle</th>
<th>p53</th>
<th>STAT-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introns</td>
<td>$p &lt; 0.0001$</td>
<td>$p &lt; 0.0001$</td>
<td>$p \approx 0.07$</td>
</tr>
<tr>
<td>Exons</td>
<td>$p &lt; 0.0001$</td>
<td>$p &lt; 0.0001$</td>
<td>$p &lt; 0.0001$</td>
</tr>
</tbody>
</table>

$p$-values according to 1000 random intervals. At least 99% overlap required.
Results: Transcription close to protein-coding genes?

- Test if distance of regulated intergenic transcripts to protein-coding genes is less than distances to random chosen regions

- Result: For all conditions: $p$-value < 0.0001
Summary - Tiling Arrays

- More transcripts regulated in cell cycle as by p53 and STAT-3
- Increased number of transcripts in 3’UTRs
- Increased number of transcripts in introns
- Increased transcription in close proximity to protein-coding genes
Functional validation of regulated transcripts

- Functional validation by spotting regulated transcripts on a custom microarray
- Studying expression of transcripts under various cellular conditions has the power to reveal ncRNAs that are regulatory functional
- Detection of reading direction (each transcript spotted in sense and antisense)
- nONCOchip: Array for cancer related transcripts
  - Experimentally identified ncRNAs regulated by oncogenes (STAT-3, prostate related), tumorsuppressors (p53) and cyclins (cell cycle)
  - All human mRNAs (RefSeq)
  - Known ncRNAs from public databases (RNAdb, NONCODE, RefSeq)
  - Computationally predicted ncRNAs (RNAz, Evofold, Foldalign)
Intergenic ncRNAs can discriminate between different disease stages
Determination of protein coding potential of controlled non-coding transcripts

Conservation patterns

Controlled regions associated to
  - known promoters or transcription factor binding sites (p53 and STAT-3 binding site)
  - open chromatin structures

Comprehensive homology search to known ncRNAs

Structural clustering to identify structural classes of RNAs
Thanks to RNomeics group at FH IZI

- Jörg Hackermüller
- Kerstin Ullmann
- Katharina Schutt
- Stephan Schreiber
- Anne-Sophie Krakovic

as well as

- Peter F. Stadler
- Friedemann Horn
Kampa et. al. 2004: Pilot study of Affymetrix tiling array

Chromosome
Affymetrix tile path
(on avg. every 35bp of 25bp length)

Pseudo−median

Pseudo−median > INTENSITY = positive probe

< MAXGAP

> MINRUN

Transfrags (contiguously transcribed elements)

2.9% false positive rate:

BW = 35bp
INTENSITY = 150
MINRUN = 90bp

MAXGAP = 40bp
Significance calculation for single signals: Wilcoxon signed ranked test on PM-MM values within bandwidth area (from all replicate arrays)

- $H_0$: No difference between PM and MM

Signal calculation (Hodges-Lehmann estimator):
- Calculate PM-MM values
- Calculate all pairwise averages (Walsh Averages)
- Signal = median of Walsh averages (log2)

- Bandwidth for single analysis: 35bp
- Bandwidth for differential analysis: 150bp
Overview

x: minimal single intensity = 75
x: minimal single intensity = 150

no RefSeq mRNAs

overlap with p53 regulated RNAs

p.value for differential expression < 0.05
QQ-Plots for sampled intervals

Kristin Reiche  
Quest for RNAs