Evolution of the Cell-division Cycle in Woodland Populations of Budding Yeast

Outlook at the midpoint

Daniel F. Simola
Laboratory of Junhyong Kim
17 April 2007
Genetic variability among woodland isolates

- Sampling of 27 oak tree-dwelling strains
- Introns from 9 unlinked loci sequenced (4.9 kb)
- Nucleotide diversity of 0.189 +/- 0.062%
- Sequencing of noncoding loci near 6 cell-cycle genes revealed elevated variation of 0.812 ± 0.55% (3.7 kb)

<table>
<thead>
<tr>
<th>ID</th>
<th>HO-Kan Strain ID</th>
<th>Origin</th>
<th>Haplotype clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>YPS2067</td>
<td>Tyler Arboretum, PA</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>YPS2055</td>
<td>Tyler Arboretum, PA</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>YPS3137</td>
<td>Jenkins Woods, PA</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>YPS2060</td>
<td>Mettlers Woods, NJ</td>
<td>B</td>
</tr>
<tr>
<td>7</td>
<td>YPS2066</td>
<td>Mettlers Woods, NJ</td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>YPS2079</td>
<td>Westtown School Woods, PA</td>
<td>B</td>
</tr>
<tr>
<td>4</td>
<td>YPS2073</td>
<td>Mettlers Woods, NJ</td>
<td>C</td>
</tr>
<tr>
<td>5</td>
<td>YPS3060</td>
<td>Jenkins Woods, PA</td>
<td>C</td>
</tr>
</tbody>
</table>
Variability in cell-cycle related phenotypes

Cell Size vs Cycle Length for SC Strains

- Origin: Tyler, Mettlers, Westtown, Jenkins
- 18° C, SD medium
- YPS Strain Neutral haplotype Cln3 haplotype

F Ge, S Brundage (unpublished)

p < 0.011
Molecular phenotyping with microarrays

- Numerous microarray studies comprehensively characterize molecular variability in *S. cerevisiae* (over 1000 published arrays)
- Others have used microarrays to investigate transcriptome evolution, e.g. *Rifkin et al.* (2005)
- No study to date has combined both approaches in a high-resolution manner
Goals

• Improve on previous cell-cycle analyses using our collection of closely related strains

• Uncover developmental heterochronies among populations

• and use them to recover *S. cerevisiae* transcriptional regulatory network

• as well as map genetic sources of CGP variation...
Overview

- Experimental protocol
- Data normalization
- QA & Sanity checks
- Visualization
In other words...

How to get from this to this
Experimental design

For each strain

- Synchronize culture with 4 µM α-factor
- Sample 18 points through 1.3 cell cycles (18° C) (roughly every 18 min)
- Dye-swapped replicates of each time sample
- Obtain two-channel measurement by hybridizing sync. samples with unsync. lab strain as common reference

- **Total slides: $12 \times 18 \times 2 = 432$**
Status

• 65 200 slides processed
• Complete data for 4 strains; halfway done another 2
• Just sampled another 3 strains. Need to harvest RNA, generate cDNA, and hybridize.
1. Yeast synchronization and sampling
2. RNA extraction of samples
3. cDNA polymerization and labeling
4. Hybridization
5. Scanning
6. Image analysis
Data processing SOP

- For each slide
  - Filter bad spots
  - Down-weight missing spots (based on S/N ratio)
  - Normalize red and green intensity distributions
  - Filter genes with high within-slide replicate variance
  - Average spot replicates for each gene
  - Compute ratios for each gene
  - Normalize log ratio distribution between slides

- Average replicate slides—biological, technical
- Impute missing values
- Python script with calls to R
- Process one or two channel data (in silico hybridization)
- Arbitrary grouping via map file
- Generates slide diagnostics, gene x condition matrices, ANOVA tables, and plots
# Map file annotations

<table>
<thead>
<tr>
<th>Barcode</th>
<th>Strain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage</td>
<td>Dyes</td>
</tr>
<tr>
<td>Dye swap</td>
<td>Duration of hyb</td>
</tr>
<tr>
<td>Scan date</td>
<td>Authors</td>
</tr>
<tr>
<td>Cell sampling dates</td>
<td>RNA dates</td>
</tr>
<tr>
<td>cDNA dates</td>
<td>Scanner</td>
</tr>
</tbody>
</table>
Background filter

Remove dropped spots (technical) but not low intensity spots (biological)
1a. Lg F(ch1) YPS2073t0s0_25
stdev 1.201

Background filtered Lg F(ch1) [Median spot]
stdev 1.201

4% of spots below cutoff on this channel

Cutoff = 7.6 bits
Spot below cutoff on both channels?

3. Prop. gene spots below 555 bg
   avg = 0.265

3. Prop. gene spots below 647 bg
   avg = 0.254

2. Prop. gene spots below bg
   avg = 0.122
Background filter

Not effective with highly fluorescent surfaces!

• Flag the spots in region as BAD
• Redo slide
• Automate: local background cutoff with auto flagging
When to redo a slide

1. B(555) cutoff
   avg = 7.454

1. B(647) cutoff
   avg = 6.393

9. Spot: Gene flagged bad
   avg = 70.976
Between channel scale normalization
Between slide scale normalization
Final data set: YPS\textsuperscript{2073}

6.2: Final Scaled ln(ch1)–ln(ch2)

Intensity

Data points for different samples are shown in the diagram.
Histograms of scaled \( \log(\text{ratios}) \)

Histogram of YPS2073t0s0_25

\[ \text{stdev } 0.993 \]

Histogram of YPS2073t0s0_76

\[ \text{stdev } 0.992 \]

Intensity

Frequency

0 400 800 1200

Histogram of YPS2073t135s0_32

\[ \text{stdev } 0.989 \]

Histogram of YPS2073t135s0_32

\[ \text{stdev } 0.995 \]

Intensity

Frequency

0 400 800 1200
Imputation of missing values

Each gene $g_i$ has a measurement from each array...
Use information from other similarly expressed genes
to fill in the missing value in array $j$ for the gene: $g_{i,j}$

- **K-nearest neighbors** use row information
  - Find $k$ genes most similar to $g_i$
  - Set $g_{i,j}$ to the average of $k$ genes in array $j$
    weighted by their distance to $g_i$

- **Cubic spline interpolation** use column information
  - Temporal data should be correlated across arrays
Cubic spline interpolation EX.
Imputation of missing values

• Throw out rows with > 50% missing data
• Fill in each missing value as
  \[ p \text{ KNN}(g_i, k) + (1-p) \text{ CS}(g_i) \]
• Weight \( p \) reflects amount of information in rows (genes)/cols (time)
• How to choose weight \( p \) and neighbors \( k \)?

SS Error vs. KN

J Weiss
QA & sanity checks

- Technical QA: diagnostics reported by normalization routine
- Biological QA: timepoints and distance
Total oligo coverage

- Overall coverage > 99.7%
- 18 NAs per slide, on average ~10 variance, 10 flagging
- Coverage of published arrays varies between .95 and .98
- Imputation used minimally—have some flexibility in choosing background cutoff
Extreme variance

6.4b Extreme replicate variance
avg = 5.098

6.4b Extreme replicate variance
avg = 14.967

6.4b Extreme replicate variance
avg = 11.605

6.4b Extreme replicate variance
avg = 4.792

<ev> = 8.8 genes

Dynamic range

5. Dynamic range lg(555/647)
avg = 10.643

5. Dynamic range lg(555/647)
avg = 9.444

5. Dynamic range lg(555/647)
avg = 11.474

5. Dynamic range lg(555/647)
avg = 11.672

5. Dynamic range lg(555/647)
avg = 11.588

<dr> = 10.2 bits
### Detecting differential expression

#### 6.3c Avg red–green diff stdev

- **avg = 0.228**

#### 6.3d. stdev(ratios)

- **avg = 0.62**

<table>
<thead>
<tr>
<th>Oligo</th>
<th>SD 183</th>
<th>2060</th>
<th>2073</th>
<th>2079</th>
<th>3060</th>
<th>3137</th>
<th>ref</th>
<th>2 sd</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.19</td>
<td>.29</td>
<td>.23</td>
<td>.28</td>
<td>.24</td>
<td>.27</td>
<td>.15</td>
<td>.50</td>
</tr>
<tr>
<td>Slide</td>
<td>.41</td>
<td>.82</td>
<td>.62</td>
<td>.84</td>
<td>.69</td>
<td>.75</td>
<td>.34</td>
<td>1.38</td>
</tr>
<tr>
<td></td>
<td>183</td>
<td>2060</td>
<td>2073</td>
<td>2079</td>
<td>3060</td>
<td>3137</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>log ratios</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.46</td>
<td>.67</td>
<td>.57</td>
<td>.75</td>
<td>.75</td>
<td>.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>log prods</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.85</td>
<td>.90</td>
<td>.74</td>
<td>.94</td>
<td>.88</td>
<td>.86</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\rho(\text{log ratios})
\]

\[
\rho(\text{log prods})
\]

\[
\langle \rho \rangle = .64
\]

\[
\langle \rho \rangle = .86
\]
Integration of time series batches

- Synchronized samples were processed in two batches, with interleaved time points.

<table>
<thead>
<tr>
<th>Original</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermed.</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>15</td>
</tr>
</tbody>
</table>

- Are these two samplings comparable?
- Use Magwene et al. (2003) path reconstruction algorithm to infer temporal ordering of samples.
Path reconstruction

• Estimated ordering from normalized data matrix (top 150 periodic genes)


gene order: [3, 15, 5, 17, 7, 9, 11, 13, 1, 0, 16, 12, 14, 8, 4, 6, 2, 10]

RMSE = 120.9
Path reconstruction with PCA(X)

- Estimated ordering from PCA of normalized data matrix (top 150 periodic genes)

\[ [9, 11, 13, 7, 3, 15, 10, 12, 0, 5, 1, 2, 14, 4, 16, 6, 8] \]

RMSE = 113.5
Singular value decomposition

- On what frequency does cell-cycle operate in expression space?

\[
\text{SVD}(X) = U \times S \times V^t = [gxg] [gxc] [cxc]^t
\]

- Singular values in S indicate variance of data projected along respective axes of U
Singular values

Eigenvalue distribution: OPUS/imputed_data/imputed_spellman_alpha (Entropy=3.494)

Eigenvalue distribution: OPUS/imputed_data/imputed_spellman_cdc15 (Entropy=3.597)

Eigenvalue distribution: OPUS/imputed_data/imputed_spellman_cdc28 (Entropy=3.411)

Eigenvalue distribution: OPUS/imputed_data/imputed_YPS183_expression (Entropy=2.289)
Singular values: rank

Cumulative eigenvalue distribution:

- OPUS/imputed_data/imputed_spellman_alpha (Entropy=3.494)
- OPUS/imputed_data/imputed_spellman_cdc15 (Entropy=3.597)
- OPUS/imputed_data/imputed_spellman_cdc28 (Entropy=3.411)
- OPUS/imputed_data/imputed_YPS183_expression (Entropy=2.289)
Singular values: rank

Cumulative eigenvalue distribution: OPUS/imputed_data/imputed_YPS2060_expression (Entropy=1.863)

Cumulative eigenvalue distribution: OPUS/imputed_data/imputed_YPS2073_expression (Entropy=2.401)

Cumulative eigenvalue distribution: OPUS/imputed_data/imputed_YPS2079_expression (Entropy=1.572)

Cumulative eigenvalue distribution: OPUS/imputed_data/imputed_YPS3060_expression (Entropy=1.381)

Cumulative eigenvalue distribution: OPUS/imputed_data/imputed_YPS3137_expression (Entropy=2.115)
Static visualization: PCA

Spellman et al. data
Static visualization: PCA

SVD projection: OPUS/imputed_data/imputed_YPS183_expression
SVD projection: OPUS/imputed_data/imputed_YPS2060_expression
SVD projection: OPUS/imputed_data/imputed_YPS2073_expression
SVD projection: OPUS/imputed_data/imputed_YPS2079_expression
SVD projection: OPUS/imputed_data/imputed_YPS3060_expression
SVD projection: OPUS/imputed_data/imputed_YPS3137_expression

KimLab data
Dynamic visualization: time-delayed phase plots

\[ \psi_p(t) = \mathbf{v} \cdot \mathbf{x}(t) \]

Time-delayed variable evaluated using

\[ \psi_p(t+nd), \]

\[ n = 0, 1, k-1, \]
\[ d = 1, k = 3 \]

Which \( \mathbf{v} \) to use?

Rifkin & Kim (2002)
Dynamic visualization: time-delayed phase plots

Spellman et al. data
Dynamic visualization: time-delayed phase plots

Smoothed temporal dynamics: OPUS/imputed_data/imputed_YPS183_expression (SVD axis 7)

Smoothed temporal dynamics: OPUS/imputed_data/imputed_YPS2060_expression (SVD axis 9)

Smoothed temporal dynamics: OPUS/imputed_data/imputed_YPS2073_expression (SVD axis 6)

186.3m/2.8c  
201.5/2c  
221.4m/1.5c  

KimLab data
Dynamic visualization: time-delayed phase plots

Smoothed temporal dynamics: OPUS/imputed_data/imputed_YPS2079_expression (SVD axis 5)

Smoothed temporal dynamics: OPUS/imputed_data/imputed_YPS3060_expression (SVD axis 6)

Smoothed temporal dynamics: OPUS/imputed_data/imputed_YPS3137_expression (SVD axis 10)

KimLab data
Plot interpretation

<table>
<thead>
<tr>
<th>Strain</th>
<th>Cycles</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>183</td>
<td>1.5</td>
<td>221</td>
</tr>
<tr>
<td>2060</td>
<td>2</td>
<td>201</td>
</tr>
<tr>
<td>2073</td>
<td>2.8</td>
<td>186</td>
</tr>
<tr>
<td>2079</td>
<td>2</td>
<td>197</td>
</tr>
<tr>
<td>3060</td>
<td>1.5</td>
<td>220</td>
</tr>
<tr>
<td>3137</td>
<td>2.1</td>
<td>202</td>
</tr>
</tbody>
</table>

\[ R^2 = 0.9032 \]

\[ p < 0.0036 \]
Future work

- Finish arrays
- Finalize normalization options
- Correct genes differentially expressed across common reference batches
- Correct for cycle length variation across strains