Tipping the balance

Introducing data management on a centre-wide level

Tomasz Zieliński, Eilidh Troup, Andrew Millar
Data life-cycle

- Creating data
- Processing data
- Analysing data
- Preserving data
- Access to data
- Re-using data
Data life-cycle

- Creating Data
- Processing Data
- Analysing Data
- Preserving Data
- Access to Data
- Re-using Data
Data life-cycle

1. CREATING DATA
2. PROCESSING DATA
3. ANALYSING DATA
4. PRESERVING DATA
5. ACCESS TO DATA
6. RE-USING DATA

DATA SHARE
DATA STORE
Data life-cycle

- Creating Data
- Processing Data
- Analysing Data
- Preserving Data
- Access to Data
- Re-using Data

Data Share
Data Store
Data life-cycle

- Creating Data
- Processing Data
- Analysing Data
- Preserving Data
- Access to Data
- Re-using Data
Data life-cycle

**Data Creator:**
Martin, Sarah F.; Noordally, Zeenat B.; van Ooijen, Gerben; Barrios-Llerena, Martin E.; Simpson, Ian; Millar, Andrew J.; Hindle, Matthew M.; Thierry Le Bihan

**Date Available:**
2014-06-24

**Citation:**

**Dataset Description (abstract):**

Background
The current knowledge of eukaryote signalling originates from phenotypically diverse organisms. There is a pressing need to identify conserved signalling components among eukaryotes, which will lead to the transfer of knowledge across kingdoms. Two useful properties of a eukaryote model for signalling are (1) reduced signalling complexity, (2) conservation of signalling components. The alga Ostreococcus tauri is described as the smallest free-living eukaryote. With less than 8,000 genes, it represents a highly constrained genomic palette. Results
Our survey revealed 133 protein kinases and 34 protein phosphatases (1.7% and 0.4% of the proteome). We conducted phospo-proteomic experiments and reconstructed domain structures and phylogenies for the catalytic protein-kinases. For each of the major kinases families we review the completeness and divergence of O. tauri representatives in comparison to the well-studied kinomes of the laboratory models Arabidopsis thaliana and Saccharomyces cerevisiae, and of Homo sapiens. Many kinase clades in O. tauri were reduced to a single member, in preference to the loss of family diversity, whereas TKL and ABC1 clades were expanded. We also identified kinases that have been lost in A. thaliana but retained in O. tauri. For three, contrasting eukaryotic pathways – TOR, MAPK, and the circadian clock – we established the subset of conserved components and demonstrate conserved sites of substrate phosphorylation and kinase motifs. Conclusions
We conclude that O. tauri satisfies our two central requirements. Several of its kinases are more closely related to H. sapiens orthologs than S. cerevisiae is to H. sapiens. The greatly reduced kinome of O. tauri is therefore a suitable model for signalling in free-living eukaryotes.
The reduced kinase of *Ostreococcus tauri*: core eukaryotic signalling components in a tractable model species

Matthew M Hindle1,4, Sarah F. Martin1,2, Zeenat B Noordally1,2, Gerben van Ooijen1,2, Martin E Barrios-Llerena1,2, T. Ian Simpson3, Thierry Le Bihan1,2 and Andrew J Millar1,2

**Abstract**

**Background:** The current knowledge of eukaryotic signalling originates from phenotypically diverse organisms. There is a pressing need to identify conserved signalling components among eukaryotes, which will lead to the transfer of knowledge across kingdoms. Two useful properties of a eukaryotic model for signalling are (i) reduced signalling complexity, and (ii) conservation of signalling components. The alga *Ostreococcus tauri* is described as the smallest free-living eukaryote. Within less than 8,000 genes, it represents a highly constrained genomic palette.

**Results:** Our survey revealed 133 protein kinases and 34 protein phosphatases (1.7% and 0.4% of the proteome). We conducted phosphoproteomic experiments and constructed domain structures and phylogenies for the catalytic protein kinases. For each of the major kinase families we review the completeness and divergence of *O. tauri* representatives in comparison to the well-studied kinomes of the laboratory models *Arabidopsis thaliana* and *Saccharomyces cerevisiae*, and of *Homo sapiens*. Many kinase clades in *O. tauri* were reduced to a single member, in preference to the loss of family diversity, whereas TKL and ABC1 clades were expanded. We also identified kinases that have been lost in *A. thaliana* but retained in *O. tauri*. For three, contrasting eukaryotic pathways - TOR, MAPK, and the circadian clock – we established the subset of conserved components and demonstrate conserved sites of substrate phosphorylation and kinase motifs.

**Conclusions:** We conclude that *O. tauri* satisfies our two central requirements. Several of its kinases are more closely related to *H. sapiens* orthologues than to *S. cerevisiae* to *H. sapiens*. The greatly reduced kinase of *O. tauri* is therefore a suitable model for signalling in free-living eukaryotes.

**Keywords:** Conserved eukaryote signalling, Protein kinase phylogeny, *Ostreococcus tauri*, Model kinase, Phosphorylation, TOR signalling, MAPK cascade, Circadian clock.

**Background**

Protein kinases are a major component of the complex signalling networks that coordinate all fundamental cellular processes, including transcription, cell cycle and metabolism. Protein kinases and phosphatases elicit reversible phosphorylation, which enable the rapid cellular responses that are crucial for survival in a continually changing environment. Protein kinases activate and deactivate proteins by addition of the gamma-phosphate from ATP to serine (S), threonine (T), tyrosine (Y), aspartate (D) or histidine (H) amino acid residues [1]. Cascades of consecutive kinase-mediated phosphorylation events constitute the backbone of signalling pathways [2]. The complexity of the signalling networks scales with size. Part of this complexity is constrained by the number of genes encoding protein kinases, also known as the kinase. The number of encoded protein kinases in free-living eukaryotes ranges from as little as 126 kinases in *Saccharomyces cerevisiae* [3] to ~1000...
Getting metadata

Species: Arabidopsis thaliana
Background: Columbia (Col)
Genotype: ztl-105

Assigned regions:
Reg 24-31
Reg 32-39
Reg 40-47
Reg 48-55
Reg 56-63
Reg 64-71
Reg 72-79
Reg 80-81

Marker: CAB
Tag type: Luc
Data cat: Transcriptional fusion
Lines code: 4F2;4H1
Sample type: group of seedlings
Sample origin: group of seedlings
Growth stage: seedling
Group size: 8
Growth cond: 22C LD
Exp cond: 15C RD+RR
Regions from: 1 to: 2

Region description:
(Transcriptional fusion)
Columbia (Col)
Genotype: ztl-105 line (4F2;4H1)
ztl-105
Growth cond: 22C LD
Exp cond: 15C RD+RR
Getting metadata
Getting meta data

- No one likes describing data for repositories
Getting metadata

No one likes describing data for repositories
Getting metadata

- No one likes describing data for repositories
Tipping the balance
Data management as part of the workflow

General repositories: SEEK and OpenBIS

- fine grained metadata,
- logical structure, data relationships
- customizable
- hooks for data processing
Workflow
Workflow
Workflow
## Workflow

<table>
<thead>
<tr>
<th>Code</th>
<th>Reporters</th>
<th>Resistance</th>
<th>Backbone</th>
<th>Description</th>
<th>ORI of Rep</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E.coli DH10B</td>
<td>NONE</td>
<td>no vectors</td>
<td></td>
<td>NA</td>
<td>Standard E. coli</td>
</tr>
<tr>
<td>10</td>
<td>BBa_80034</td>
<td>CM</td>
<td>pSB1A2</td>
<td>Strong RBS (tctagaGAAAGAGCCAC)</td>
<td>COLE1</td>
<td>strong 1009</td>
</tr>
<tr>
<td>11</td>
<td>BBa_20050</td>
<td>CM</td>
<td>pSB1C3</td>
<td>EYFP</td>
<td>PMB2</td>
<td>enhanced yellow</td>
</tr>
<tr>
<td>12</td>
<td>BBa_20050</td>
<td>CM</td>
<td>pSB1C3</td>
<td>mRFP1</td>
<td>PMB1</td>
<td>monomeric</td>
</tr>
<tr>
<td>13</td>
<td>BBa_E0040</td>
<td>AP</td>
<td>pSB1A2</td>
<td>GFPmut3b</td>
<td>COLE1</td>
<td>Green fluorescent</td>
</tr>
<tr>
<td>14</td>
<td>BBa_J06702</td>
<td>AP</td>
<td>pSB1A2</td>
<td>mCherry (yeast version also)</td>
<td>COLE1</td>
<td>B0034 RBS/</td>
</tr>
<tr>
<td>15</td>
<td>BBa_J23114</td>
<td>AP</td>
<td>pJ61002</td>
<td>Weak promoter (tttatgcttagctc)</td>
<td>COLE1</td>
<td>weak promoter</td>
</tr>
<tr>
<td>16</td>
<td>BBa_J23100</td>
<td>AP</td>
<td>pJ61002</td>
<td>Strong promoter (tttagagctagctc)</td>
<td>COLE1</td>
<td>strong promoter</td>
</tr>
<tr>
<td>17</td>
<td>BBa_J23114</td>
<td>AP</td>
<td>pSB1A2</td>
<td>weak RBS (ttctagaGAAAGACACAGGA)</td>
<td>COLE1</td>
<td>weak 12%</td>
</tr>
<tr>
<td>18</td>
<td>E. coli cc118 Lambda Redmid</td>
<td>NONE</td>
<td>no vectors</td>
<td>cloning strain to maintain RK6 pl</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>P. putida KT2440</td>
<td>NONE</td>
<td>no vectors</td>
<td>standard P. putida strain without</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>HB101 RK600</td>
<td>CM</td>
<td>no vectors</td>
<td>E. coli HB101 helper strain to modulate</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>pBBFLP</td>
<td>TC</td>
<td>pBBFLP</td>
<td>Plasmid source Flipase</td>
<td>UNKNOWN</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>pSEVA 121</td>
<td>AP</td>
<td>pSEVA 121</td>
<td>low copy broad host range plasm</td>
<td>RK2</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>pSEVA 221</td>
<td>KM</td>
<td>pSEVA 221</td>
<td>low copy broad host range plasm</td>
<td>RK2</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>pSEVA 321</td>
<td>CM</td>
<td>pSEVA 321</td>
<td>low copy broad host range plasm</td>
<td>RK2</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Magnetospirillum ma...</td>
<td>UNKNOWN</td>
<td>NA</td>
<td>glicerol prepared from sample control</td>
<td>UNKNOWN</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>pSEVA 521</td>
<td>TC</td>
<td>pSEVA 521</td>
<td>low copy broad host range plasm</td>
<td>RK2</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>pSEVA 621</td>
<td>CM</td>
<td>pSEVA 621</td>
<td>low copy broad host range plasm</td>
<td>RK2</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>pUC18NotI</td>
<td>AP</td>
<td>pUC18NotI</td>
<td>high copy shuttle vector for E. co</td>
<td>UNKNOWN</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>BBA_80034</td>
<td>CM</td>
<td>pSB1C3</td>
<td>pSB1C3</td>
<td>PMB2</td>
<td>enhanced yellow</td>
</tr>
</tbody>
</table>

[Image of workflow diagram]
## Workflow

### Table

<table>
<thead>
<tr>
<th>Code</th>
<th>Reporters</th>
<th>Resistance</th>
<th>Backbone</th>
<th>Description</th>
<th>ORI of Rep</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E.coli DH10B</td>
<td>NONE</td>
<td>no vectors</td>
<td>NA</td>
<td>Strandard E.coli DH10B</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>BbA_B0034</td>
<td>CM</td>
<td>no vectors</td>
<td>Strong RBS (tctagaGAAAGAGC) COLE1</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BbA_J06702</td>
<td>CM</td>
<td>pSB1C3</td>
<td>EYFP</td>
<td>PMB2</td>
<td>enhanced yellow fluorescent protein</td>
</tr>
<tr>
<td>3</td>
<td>BbA_J23114</td>
<td>CM</td>
<td>pSB1C3</td>
<td>ECFP</td>
<td>PMB2</td>
<td>monomeric cyan fluorescent protein</td>
</tr>
<tr>
<td>5</td>
<td>BbA_J23100</td>
<td>CM</td>
<td>pSB1C3</td>
<td>mRFP1</td>
<td>PMB1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>BbA_E0040</td>
<td>CM</td>
<td>pSB1C3</td>
<td>LpRFP</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>BbA_J06702</td>
<td>CM</td>
<td>pSB1C3</td>
<td>LpRFP</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>BbA_J23114</td>
<td>CM</td>
<td>pSB1C3</td>
<td>LpRFP</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>BbA_J23100</td>
<td>CM</td>
<td>pSB1C3</td>
<td>LpRFP</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>BbA_J23114</td>
<td>CM</td>
<td>pSB1C3</td>
<td>LpRFP</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>BbA_J61101</td>
<td>CM</td>
<td>pSB1C3</td>
<td>LpRFP</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>E.coli cc118 LD</td>
<td>CM</td>
<td>pSB1C3</td>
<td>LpRFP</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>P.putida KT24</td>
<td>CM</td>
<td>pSB1C3</td>
<td>LpRFP</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>H8101 RK600</td>
<td>CM</td>
<td>pSB1C3</td>
<td>LpRFP</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>pBBFLP</td>
<td>CM</td>
<td>pSB1C3</td>
<td>LpRFP</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>pSEVA 121</td>
<td>CM</td>
<td>pSB1C3</td>
<td>LpRFP</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>pSEVA 221</td>
<td>CM</td>
<td>pSB1C3</td>
<td>LpRFP</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>pSEVA 321</td>
<td>CM</td>
<td>pSB1C3</td>
<td>LpRFP</td>
<td>PMB2</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Magnetospirillum marni</td>
<td>UNKNOWN</td>
<td>NA</td>
<td>glicerol prepared from sample cu</td>
<td>UNKNOWN</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>pSEVA 521</td>
<td>TC</td>
<td>pSEVA 521</td>
<td>low copy broad host range plasm</td>
<td>RK2</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>pSEVA 621</td>
<td>GM</td>
<td>pSEVA 621</td>
<td>low copy broad host range plasm</td>
<td>RK2</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>pUC18Notl</td>
<td>AP</td>
<td>pUC18Notl</td>
<td>high copy shuttle vector for E. co</td>
<td>UNKNOWN</td>
<td></td>
</tr>
</tbody>
</table>
Workflow
Lowering the barrier
Successful data management?

- is part of the research workflow not an extra burden
- solves user problems not creates new ones
- provides extra value for the user
- focuses on both data consumers and producers
The missing link

DATA STORE

DATA SHARE

CREATING DATA

PROCESSING DATA

ANALYSING DATA

PRESERVING DATA

ACCESS TO DATA

RE-USING DATA