S1 Data integration

S1.1 Ondex integration workflow XML file

```xml
<?xml version="1.0"?>
<Ondex xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:noNamespaceSchemaLocation="ONDEXParameters.xsd">
  <DefaultGraph name="sce_integrated" type="memory">
    <Parameter name="ReplaceExisting">true</Parameter>
  </DefaultGraph>

  <Parser name="yeasthomology">
    <Parameter name="InputFile">${homologyfile}</Parameter>
  </Parser>

  <Parser name="yeastgenome">
    <Parameter name="InputFile">${sgdfile}</Parameter>
    <Parameter name="TranslationFile">${sgdmapping}</Parameter>
    <Parameter name="CreateRNAConcepts">true</Parameter>
    <Parameter name="InferRelativeLocations">true</Parameter>
  </Parser>

  <Parser name="tf2">
    <Parameter name="InputFile">${regulationfile}</Parameter>
  </Parser>

  <Parser name="sbml2">
    <Parameter name="InputDir">${sbmldir}</Parameter>
    <Parameter name="DataSource">UMan</Parameter>
  </Parser>

  <Parser name="goa2">
    <Parameter name="InputFile">${goafile}</Parameter>
    <Parameter name="TranslationFile">${goamapping}</Parameter>
  </Parser>

  <Parser name="biogrid">
    <Parameter name="InputFile">${biogridfile}</Parameter>
    <Parameter name="TranslationFile">${biogridmapping}</Parameter>
    <Parameter name="TaxIDRestriction">4932</Parameter>
  </Parser>

  <Transformer name="yeastmerger"/>

  <Export name="oxl" datafile="sce_integrated.v4.1.xml.gz">
    <Parameter name="GZip">true</Parameter>
  </Export>
</Ondex>
```
S1.2 The integration workflow

The semantic data integration performed by Ondex is a product of the interplay between three factors: The underlying definitions of a metadata structure form the basic semantics of the biological domain (S1.2.1); Parsers interpret given source data and express it in terms of this basic domain description, thus generating semantic subnetworks (S1.2.2); Finally, transformation methods join and interweave these subnetworks to form a fully connected knowledge network, while simultaneously controlling the data’s consistency (S1.2.3).

More information about the general mechanisms of data integration with Ondex can be found in previous Ondex publications [1, 2, 3].

S1.2.1 The structure of the metadata

The Ondex metadata is a controlled vocabulary of concept types, relation types, evidence codes and provenance codes. In particular, the concept types and relation types are organised in subtype hierarchies and thus each bear a computationally amenable meaning.

Figure S1 shows the hierarchy of concept types used in this work. The complete metadata set can be found by downloading the Ondex workflow engine (described below) and opening the file “data/xml/metadata.xml” using the Ondex metadata editor tool.

Figure S1: The hierarchy of concept types in the Ondex metadata as seen in the Ondex metadata editor tool.
S1.2.2 Parsing

Parsers are an important contributor to the process of data integration. Their role goes beyond simple conversions of one data format into another. Within their code parsers bear their respective author’s interpretation of the subject data and are thus able to elevate these data to the semantic level. Data points from the data sources are assigned types from the controlled metadata sets and are then formed into concepts and relations in the knowledge network.

The workflow shown in the above XML file starts by creating a new OndexGraph with the name “sce_integrated”. It then proceeds to parse a tab-delimited homology file, creating two Gene concepts connected by a homologue relation for each entry.

The “yeastgenome” parser proceeds to parse SGD’s feature table [4]. Creating concepts of various subclasses of Nucleotide_Feature for each feature in the file. Names and descriptions are copied, cross references are saved to Ondex accessions, Genomic positions and exon-intron structures are recorded as concept attributes (GDSs). The parser also creates connected RNA concepts for all genes that are known to be expressed. Finally it infers relative locations between features and creates adjacency and overlap relations between appropriate Nucleotide_Feature concepts.

The regulatory network parser now parses Balaji and colleagues’ “tf2” source file [5]. It creates a Transcription_Factor concept and a Gene concept connected by a regulation relation for each entry.

Subsequently the SBML parser reads the jamboree yeast metabolic network file [6]. Based on SBO annotations it creates concepts of appropriate types for each species (e.g. Enzyme, Metabolite) and reaction (e.g. Reaction, Transportation) and connects them with relations of appropriate subtypes of participates_in. It also creates concepts of type Cellular_Compartment and connects them with located_in relations to the correct species concepts.

The Gene Ontology annotation parser proceeds by reading in SGD’s annotation files [4, 7]. It creates concepts of type Gene connected to concepts of the types Biological_Process, MolecularFunction and Cellular_Compartment using relations of type participates_in, not_participates_in, has_function, not_has_function, located_in and not_located_in. It also annotates these relations with concepts representing the reporting publications.

The BioGRID parser now parses the BioGRID yeast interaction data file [8]. It creates concepts of appropriate types (Gene, Polypeptide, RNA) as well as concepts of various genetic and physical interaction types and connects them correctly using subtypes of participates_in relations. It also creates concepts of type Publication connected to the interaction concepts. In order to prevent cross-species interactions to be parsed from the BioGRID file as well it checks whether the NCBI taxonomy IDs of each interaction participant equals “4932” (for Saccharomyces cerevisiae and discards unqualified interactions.)
S1.2.3 Data joining and interlinking

Since every parser creates new concepts regardless of the potential existence of equivalent concepts there are now many duplicates within the datasets. By eliminating these duplicates through merging, the actual joining of the data will take place. The “yeastmerger” transformer tool will perform this function here. It indexes all concepts according to their cross references and thus identifies equivalence sets. These are then merged if their respective types allow for this operation to safely proceed. The merger will automatically infer the most specific type for the resulting merge products. Concepts that can not be merged because of incompatible typing are reported to the log file, as they usually represent errors or inconsistencies in the data sources. For example, a concept cannot be both gene and a pseudogene.

Furthermore the transformer identifies groups of derivative concepts and connects them with appropriate relations: Genes, mRNAs and Polypeptides with the same ORF identifier are connected with transcription, translation and encoding relations. Protein complexes containing the same gene names as Polypeptides are connected with part_of relations. This way the previously independent subnetworks are systematically interlinked with each other, forming a unified knowledge network.

S1.3 Workflow execution

To run the workflow described above the Ondex workflow engine needs to be downloaded and installed. A detailed instruction can be found on the Ondex.org webpage.

The workflow requires the following plug-ins: iah | sbml | interaction. To install the plug-ins simply copy them into the workflow engine’s “plugins/” directory.

When executing the workflow with the above xml file, make sure to initialise all variables declared in the file, pointing them at the appropriate data source files.
## S1.4 Data sources

### Table S1: SGD yeast genome. Version 11/02/2010

| Entity type               | # entities |
|---------------------------|------------|
| Gene                      | 6773       |
| Unclassified feature      | 516        |
| Long terminal repeat      | 383        |
| ARS                       | 337        |
| tRNA gene                 | 313        |
| ARS Consensus             | 99         |
| snoRNA gene               | 77         |
| Retrotransposon           | 50         |
| XCore                     | 32         |
| Telomeric Repeat Region   | 32         |
| Subtelomeric binding site | 31         |
| Telomeric repeat          | 31         |
| Subtelomeric repeat       | 28         |
| rRNA gene                 | 27         |
| Repeat                    | 19         |
| Y' element                | 19         |
| CDEI                      | 16         |
| CDEII                     | 16         |
| CDEIII                    | 16         |
| Centromere                | 16         |
| mRNA gene                 | 13         |
| snRNA gene                | 6          |
| X region                  | 3          |
| Y region                  | 3          |
| Z1 region                 | 3          |
| Multigene locus           | 3          |
| W region                  | 2          |
| Z2 region                 | 2          |
| Gene cassette             | 2          |
| Mating locus              | 1          |

### Table S2: Jamboree yeast metabolic network (Herrgard et al. 2008). Version 1.0

| Entity type       | # entities |
|-------------------|------------|
| Complex           | 96         |
| Enzyme            | 764        |
| Transporter       | 124        |
| Metabolite        | 1168       |
| Transport         | 545        |
| Reaction          | 1312       |
| Compartment       | 15         |

### Table S3: Yeast regulatory network (Balaji et al. 2006)

| Entity type      | # entities |
|------------------|------------|
| Transcription factor | 157        |
| Gene             | 4410       |
| Regulation       | 12873      |

### Table S4: BioGRID yeast genetic and physical interactions. Version 2.0.61

| Entity type       | # entities |
|-------------------|------------|
| Publication       | 7838       |
| Gene              | 4230       |
| Protein           | 5105       |
| Physical interaction | 54165      |
| Phenotypic enhance- ment | 24637      |
| Synthetic lethality | 12425      |
| Synthetic growth defect | 12007      |
| Phenotypic suppress- tion | 9267       |
| Dosage rescue     | 3336       |
| Synthetic rescue  | 2478       |
| Dosage lethality  | 463        |
| Synthetic haploinsufficiency | 201       |
| Dosage growth defect | 121        |

### Table S5: SGD yeast GO annotations. Version 11/02/2010

| Entity type       | # entities |
|-------------------|------------|
| Gene              | 6353       |
| Biological process | 1877       |
| Molecular function | 1581       |
| Cellular component | 609        |
| located in        | 25911      |
| is not located in | 7          |
| participation     | 27943      |
| does not participate | 29         |
| has function      | 21239      |
| does not have function | 36         |
S1.5 Integrated concepts by class

Table S6: Sequence features

| Concept class                  | # Concepts |
|-------------------------------|------------|
| Gene                          | 6773       |
| Unclassified feature          | 516        |
| Long terminal repeat          | 383        |
| ARS                           | 337        |
| tRNA gene                     | 313        |
| ARS Consensus                 | 99         |
| snoRNA gene                   | 77         |
| Retrotransposon               | 50         |
| XCore                         | 32         |
| Telomeric Repeat Region       | 32         |
| Subtelomeric binding site     | 31         |
| Telomeric repeat              | 31         |
| Subtelomeric repeat           | 28         |
| rRNA gene                     | 27         |
| Repeat                        | 19         |
| Y' element                    | 19         |
| CDEI                          | 16         |
| CDEII                         | 16         |
| CDEIII                        | 16         |
| Centromere                    | 16         |
| nRNA gene                     | 13         |
| snRNA gene                    | 6          |
| X region                      | 3          |
| Y region                      | 3          |
| Z1 region                     | 3          |
| Multigene locus               | 3          |
| W region                      | 2          |
| Z2 region                     | 2          |
| Gene cassette                 | 2          |
| Mating locus                  | 1          |

Table S7: Interactions

| Concept class                  | # Concepts |
|-------------------------------|------------|
| Physical interaction          | 54165      |
| Phenotypic enhancement        | 24637      |
| Synthetic lethality           | 12425      |
| Synthetic growth defect       | 12007      |
| Phenotypic suppression        | 9267       |
| Dosage rescue                 | 3336       |
| Synthetic rescue              | 2478       |
| Dosage lethality              | 463        |
| Synthetic haploinsufficiency  | 201        |
| Dosage growth defect          | 121        |

Table S8: Molecules

| Concept class                  | # Concepts |
|-------------------------------|------------|
| Polypeptide                    | 6256       |
| mRNA                           | 5877       |
| Metabolite                     | 1168       |
| Enzyme                         | 805        |
| mRNA                           | 299        |
| Transcription factor           | 157        |
| Transporter                    | 124        |
| snoRNA                         | 77         |
| rRNA                           | 27         |
| ncRNA                          | 13         |
| snRNA                          | 6          |
| RNA                            | 1          |

Table S9: Other

| Concept class                  | # Concepts |
|-------------------------------|------------|
| Publication                    | 15872      |
| Biological process             | 1877       |
| Molecular function             | 1581       |
| Cellular component             | 609        |
| Reaction                       | 1312       |
| Transportation                 | 545        |
Figure S2: Edited screenshot from OndexView showing the neighbourhood of BMH1 and BMH2. Edges represent physical interactions (blue), homology (red), regulation (green) and epistasis (ochre). Nodes represent genes, some of which weak (red) and strong (dark red) suppressors of cdc13-1. Genes which show lethal or slow growing phenotypes upon deletion are marked with dashed outlines, as their epistatic behaviour can be expected to be largely unknown.
S3 Differential regulation hypothesis

Figure S3: Gene expression of BMH1 and BMH2 (Spellman et al. 1998)
References

[1] Jacob Kohler, Jan Baumbach, Jan Taubert, Michael Specht, Andre Skusa, Alexander Ruegg, Chris Rawlings, Paul Verrier, and Stephan Philippi. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22(11):1383–1390, June 2006.

[2] J. Taubert, M. Hindle, B. Hoekman, R. Winnenburg, J. Weile, K. P Sieren, J. Baumbach, S. Ernst, S. Elsner, K. Hassani-Pak, A. Lysenko, S. Philippi, T. Yates, P. Verrier, C. Rawlings, and J. Koehler. The ONDEX data integration framework. In *Bioinformatics Open Source Conference (BOSC 2007)*, 2007.

[3] Jan Taubert, Klaus P Sieren, Matthew Hindle, Berend Hoekman, Chris Rawlings, and Jacob Köhler. The OXL format for the exchange of integrated datasets. In *Proceedings of the 4th International Symposium on Integrative Bioinformatics*, volume 4(3), page 62, Ghent, Belgium, 2007. Journal of Integrative Bioinformatics.

[4] J M Cherry, C Ball, S Weng, G Juvik, R Schmidt, C Adler, B Dunn, S Dwight, L Riles, R K Mortimer, and D Botstein. Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature*, 387(6632 Suppl):67–73, May 1997. PMID: 9169866.

[5] S. Balaji, M. Madan Babu, Lakshminarayan M. Iyer, Nicholas M. Luscombe, and L. Aravind. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *Journal of Molecular Biology*, 360(1):213 – 227, 2006.

[6] Markus J Herrgard, Neil Swainston, Paul Dobson, Warwick B Dunn, K Yakin Arga, Mikko Arvas, Nils Buthgen, Simon Borger, Roeland Costenoble, Matthias Heinemann, Michael Hucka, Nicolas Le Novere, Peter Li, Wolfram Liebermeister, Monica L Mo, Ana Paula Oliveira, Dina Petranovic, Stephen Pettiler, Evangelos Simeonidis, Kieran Smallbone, Irena Spasie, Dieter Weichert, Roger Brent, David S Broomhead, Hans V Westerhoff, Betul Kurdar, Merja Penttila, Edda Klipp, Bernhard O Palsson, Uwe Sauer, Stephen G Oliver, Pedro Mendes, Jens Nielsen, and Douglas B Kell. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotech*, 26(10):1155–1160, October 2008.

[7] M. Ashburner, C. A Ball, J. A Blake, D. Botstein, H. Butler, J. M Cherry, A. P Davis, K. Dolinski, S. S Dwight, J. T Eppig, M. A Harris, D. P Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C Matses, J. E Richardson, M. Ringwald, G. M Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000.

[8] Bobby-Joe Breitkreutz, Chris Stark, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, Michael Livstone, Rose Oughtred, Daniel H Lackner, Jürg Bähler, Valerie Wood, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2008 update. *Nucleic Acids Research*, 36(Database issue):D637–640, January 2008. PMID: 18000002.