

# The Automation of Science

Ross D. King<sup>1\*</sup>, Jem Rowland<sup>1</sup>, Stephen G. Oliver<sup>2</sup>, Michael Young<sup>3</sup>, Wayne Aubrey<sup>1</sup>, Emma Byrne<sup>1</sup>, Maria Liakata<sup>1</sup>, Magdalena Markham<sup>1</sup>, Pnar Pir<sup>2</sup>, Larisa N. Soldatova<sup>1</sup>, Andrew Sparkes<sup>1</sup>, Kenneth E. Whelan<sup>1</sup>, Amanda Clare<sup>1</sup>

<sup>1</sup>Department of Computer Science, University of Wales, Aberystwyth, SY23 3DB, UK

<sup>2</sup>Cambridge Systems Biology Centre, Department of Biochemistry, University of Cambridge, Sanger Building, 80 Tennis Court Road, Cambridge CB2 1GA, UK

<sup>3</sup>Institute of Biological, Environmental and Rural Sciences, University of Wales, Aberystwyth University, SY23 3DD, UK

\*Corresponding author: Email [rdk@aber.ac.uk](mailto:rdk@aber.ac.uk) Tel. +44 (0) 1970 622432 Fax. +44 (0) 1970 622455

One line summary: We report the automation of the discovery of novel scientific knowledge.

## Abstract

The basis of science is the hypothetico-deductive method, and the recording of experiments in sufficient detail to enable reproducibility. We report the development of the Robot Scientist “Adam” which advances the automation of both. Adam has autonomously generated functional genomics hypotheses about the yeast *Saccharomyces cerevisiae*, and experimentally confirmed these hypotheses using laboratory automation. We have manually confirmed Adam's conclusions using additional experiments. To describe Adam's experiments we have developed an ontology and logical language. The resulting formalisation involves over 10,000 different research units in a nested tree-like structure, ten levels deep, that relates the 6.6 million biomass measurements to their logical description. This formalisation describes novel scientific knowledge discovered by a machine.

Computers are playing an ever-greater role in the scientific process (1). Their use to control the execution of experiments contributes to a vast expansion in the production of scientific data (2). This growth in scientific data, in turn, requires the increased use of computers for modelling and analysis. The use of computers is also changing the way that science is described and reported. Scientific knowledge is best expressed using formal logical languages (3). Only formal languages provide the semantic clarity to ensure the reproducibility of results and the free exchange of scientific knowledge. Despite the advantages of logic most scientific knowledge is expressed using natural languages. There is however growing interest in formalising scientific knowledge through such developments as the Semantic Web (4), and ontologies (5).

A natural extension of the trend to ever-greater computer involvement in science is the concept of a Robot Scientist (6). This is a physically implemented laboratory automation system that exploits techniques from the field of artificial intelligence (7-9) to execute cycles of scientific experimentation. A Robot Scientist automatically originates hypotheses to explain observations, devises experiments to test these hypotheses, physically runs the experiments using laboratory robotics, interprets the results, and then repeats the cycle.

High-throughput laboratory automation is transforming biology and revealing vast amounts of new scientific knowledge (10). However, current high-throughput methods are insufficient to build comprehensive models of cellular systems. This is because, even though very large numbers of experiments can be executed, each individual experiment cannot be designed to test a hypothesis about a model. Robot Scientists have the potential to overcome this fundamental limitation.

The complexity of biological systems forces necessitates the recording of experimental metadata in as much detail as possible. Acquiring these metadata has often proved problematic. With Robot Scientists the production of comprehensive metadata is a natural by-product of the way they work. As the experiments are conceived and executed automatically by computer, it is possible to completely capture and digitally curate all aspects of the scientific process (11, 12).

To demonstrate that the Robot Scientist methodology can both be fully automated and be made effective enough to discover new scientific knowledge we have developed the Robot Scientist, "Adam" (13) (Fig. 1). Adam's hardware is fully automated such that it only requires a technician to periodically add laboratory consumables and to remove waste.

Adam's hardware is designed to automate the high-throughput execution of individually designed microbial batch growth experiments in microtitre plates (14). Adam measures growth curves (phenotypes) of selected microbial strains (genotypes) growing in defined media (environments). Cell culture growth can be easily measured in high-throughput, and growth curves are sensitive to

changes in genotype and environment.

We applied Adam to the discovery of genes encoding orphan enzymes in *S. cerevisiae*: enzymes catalysing biochemical reactions thought to occur in yeast, but for which the gene(s) encoding them have not been identified (15). To set-up Adam for this application required: *i.* A comprehensive logical model encoding knowledge of *S. cerevisiae* metabolism (~1,200 ORFs, ~800 metabolies) (15), expressed in the logic programming language Prolog. *ii.* A general bioinformatic database of genes and proteins involved in metabolism (also in Prolog). *iii.* Software to abduce hypotheses about the genes encoding the orphan enzymes: this is done using a combination of standard bioinformatic software and the database. *iv.* Software to deduce experiments that test the observational consequences of hypotheses (based on the model). *v.* Software to plan and design the experiments, which are based on the use of deletant mutants and the addition of selected metabolites to a defined growth medium. *vi.* Laboratory automation software to physically execute the experimental plan, and to record the data and meta-data in a relational database. *vii.* Software to analyse the data and meta-data (generate growth curves and extract parameters). *viii.* Software to relate the analysed data to the hypotheses: for example, statistical methods are required to decide on significance. Once this infrastructure is in place, no human intellectual intervention is necessary to execute cycles of simple hypothesis-led experimentation. (For more details of the software, and their application to a related functional genomics problem see Supporting Online Material SOM).

Adam formulated and tested 20 hypotheses concerning genes encoding 13 orphan enzymes (see SOM) and obtained novel results (Table 1). The weight of the experimental evidence for the hypotheses varied (based on observations of differential growth), but twelve novel hypotheses were confirmed with  $P < 0.05$  for the null hypothesis.

Because Adam's experimental evidence for its conclusions are indirect, we tested Adam's conclusions with more direct experimental methods. The enzyme 2-aminoadipate:2-oxoglutarate aminotransferase (2A2OA) catalyses a reaction in the lysine biosynthetic pathways of fungi. Adam hypothesised that three genes encode this enzyme (YER152C, YJL060W, YGL202W), and observed results consistent with all three hypotheses (Table 1). To test Adam's conclusions, we purified the protein products of these genes and used them in *in vitro* enzyme assays confirming Adam's conclusions (Fig. 2). (See SOM for further details, and additional experimental evidence)

To further test Adam's conclusions we examined the scientific literature on the 20 genes investigated (Table 1) (see SOM). This revealed the existence of strong empirical evidence for the correctness of six of the hypotheses, i.e. the enzymes were not actually orphans (see Table 1). The reason that Adam considered them to be orphans was due to the use of an incomplete bioinformatic database. These six genes therefore constitute a positive control for Adam's methodology. A possible error was also

revealed (see SOM).

To better understand the reasons why the identity of the genes encoding these enzymes has remained obscure for so long we investigated their comparative-genomics in detail (see SOM). The likely explanation is a combination of three complicating factors: gene duplications with retention of overlapping function, enzymes that catalyse more than one related reaction, and existing functional annotations. Adam's systematic bioinformatic and quantitative phenotypic analyses were required to unravel the web of their functionality.

Use of a Robot Scientist enables all aspects of a scientific investigation to be formalised in logic. For the core organisation of this formalisation we used the ontology of scientific experiments: EXPO (11, 12). This ontology formalises generic knowledge about experiments. For Adam we developed LABORS, a customised version of EXPO, expressed in the description logic language OWL-DL (19). Application of LABORS produces experimental descriptions in the logic-programming language Datalog (20). In the course of its investigations Adam observed 6,657,024 OD<sub>600nm</sub> measurements (from 26,495 growth curves). These data are held in a MySQL relational database. Use of LABORS resulted in a formalisation of the scientific-investigations/argument involving over 10,000 different research units. This has a nested tree-like structure, ten levels deep, that logically connects the experimental observations to the experimental metadata. (Fig. 2). This structure resembles the trace of a computer program, and takes up 366 Megabytes - see SOM. Making such experimental structures explicit renders scientific research more comprehensible, reproducible, and reusable. This paper may be considered as simply the human-friendly summary of the formalisation.

A major motivation for the formalisation of experimental knowledge is the expectation that such knowledge is easily reused to answer other scientific questions. To test this we investigated whether we could reuse Adam's functional genomic research (see SOM). An example question investigated was the relative growth-rate ( $\mu_{\max}$ ) in rich and defined media of the deletant strains v the wild-type. What was observed, in both media, was an asymmetric distribution of differences, with a few deletant strains having a much lower  $\mu_{\max}$  than the wild-type, but most having a slightly higher  $\mu_{\max}$ . These observations question the common assumption that wild-type *S. cerevisiae* is optimised for  $\mu_{\max}$ , and provide quantitative test data for yeast Systems Biology models (21).

It could be argued that the scientific knowledge “discovered” by Adam is implicit in the formulation of the problem, and is therefore not novel. This argument that computers cannot originate anything is known as “Lady Lovelace's objection” (22): “The Analytical Engine has no pretensions to *originate* anything. It can do *whatever we know how to order it to perform*” (Lady Lovelace's italics). We accept that the knowledge automatically generated by Adam is of a modest kind. However, this knowledge is not trivial. Moreover, in the case of the genes encoding 2A2OA, it sheds light on, and

perhaps solves, a 50 year-old puzzle (17).

Adam is a prototype and could be greatly improved. Adam's hardware and software are “brittle”, so although Adam is capable of running for a few days without human intervention, it is advisable to have a technician nearby in case of problems. The integration of Adam's AI software also needs to be enhanced so that it works seamlessly. To extend Adam we have developed software to enable external users to propose hypotheses and experiments, and we plan to automatically publish the logical descriptions of automated experiments. The idea is to develop a way to allow teams of human and robot scientists to work together. The greatest research challenge will be to improve the scientific intelligence of the software. We have shown that a simple form of hypothesis-led discovery can be automated, but what remains to be determined are the limits of automation.

## References

1. Various authors. *Nature* **440**, 383 (2006).
2. Hey, T. Trefethen, A. In *Grid Computing - Making the Global Infrastructure a Reality*, **36** 809 (John Wiley & Sons, New York, 2003).
3. Toulmin, S. The Philosophy of Science. In *Encyclopaedia Britannica Deluxe Edition 2004 CD* (Encyclopaedia Britannica UK, London, 2003).
4. Berners-Lee *et al.*, *Scientific American*, May 17, 2001. [www.sciam.com](http://www.sciam.com) (2001).
5. Ashburner, M. *et al.*, *Nature Genetics* **25**, 25 (2000).
6. King, R.D. *et al.*, *Nature* **427**, 247 (2004).
7. Buchanan, B.G. *et al.*, In *Machine Intelligence Vol. 5* (Eds. Meltzer, B. & Michie, D.) 253 (Edinburgh University Press, Edinburgh, 1969).
8. Langley, P. *et al.*, *Scientific Discovery: Computational Explorations of the Creative Process* (The MIT Press, Cambridge, Massachusetts, 1987).
9. Zytkow, J.M. *et al.*, In *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90)*, 889 (AAAI Press, Menlo Park, CA, 1990).
10. Hood, L. *et al.*, *Science* **306**, 640 (2004).
11. Soldatova, L.N. King, R.D. *J. Roy. Soc. Interface* **3**, 795 (2006).
12. Soldatova L.N. *et al.*, *Bioinformatics* **22**, e464 (2006).
13. King *et al.*, The Robot Scientist Project. <http://www.aber.ac.uk/compsci/Research/bio/robotsci/> (2008).
14. Warringer, J. *et al.*, *Proc. Natl. Acad. Sci. USA* **100**, 15724-15729 (2003).
15. Whelan, K.E. King, R.D. *BMC Bioinformatics* **9**:97 (2008).
16. Wogulis, M. *et al.*, *Biochemistry*. **47**, 1608 (2008).
17. Zabriski, T.M. Jackson, M.D. *Nat. Prod. Rep.* **17**, 85 (2000).
18. Cannon, J.F. *et al.*, *J. Biol. Chem.* **265**, 11897 (1990).
19. Horrocks, I. *et al.*, *Web Semantics: Science, Services and Agents on the World Wide Web.* **1**, 7 (2003).
20. Ullman, J.D. *Principles of Database and Knowledge-Base Systems, Vol I: Classical Database Systems* (Computer Science Press, New York, NY, 1988).
21. Herrgard M.J. *et al.*, *Nature Biotechnology* **26** 1155 (2008)
22. Turing, A. *Mind.* **236**, 433 (1950).
23. Young, M. *J. Exp. Bot.* **24**, 1172 (1973).
24. This work was funded by grants from the BBSRC to RDK and SGO, by a SRIF 2 award to RDK, by Fellowships from the Royal Commission for the Great Exhibition of 1851 and the Royal Academy of Engineering to AF, and by a RC-UK Fellowship to LNS. We thank Michael Benway for help with Adam.

## Figure Legends

### Fig. 1

The Robot Scientist, Adam. The advances that distinguish Adam from other complex laboratory systems are the individual design of the experiments to test hypotheses, and the utilisation of complex internal cycles. Adam's basic operations are: selection of specified yeast strains from a library held in a freezer, inoculation of these strains into microtitre plate wells containing rich medium, measurement of growth curves on rich medium, harvesting of a defined quantity of cells from each well, inoculation of these cells into wells containing defined media (minimal SD media plus up to four added metabolites from a choice of six), and measurement of growth curves on the specified media. To achieve this functionality Adam has the following components: [a] an automated -20°C freezer, [b] three liquid-handlers (one of which can separately control 96 fluid channels simultaneously), [c] three automated +30°C incubators, [d] two automated plate-readers, [e] three robot arms, [f] two automated plate slides, [g] an automated plate centrifuge, [h] an automated plate washer, [i] two HEPA air filters, and [j] a Perspex sterile enclosure. There are also two barcode readers, seven cameras, twenty environment sensors and four PCs, as well as the software. Adam is capable of designing and initiating over a thousand new strain/defined-growth-medium experiments each day (from a selection of thousands of yeast strains), with each experiment lasting up to 5 days. The design enables measurement of optical density ( $OD_{600nm}$ ) for each experiment at least once every 30 minutes (more often if running at less than full capacity), allowing accurate growth curves to be recorded (typically we take over a hundred measurements a day per well), plus recording associated metadata. See the SOM for a video of Adam in action.

### Figure 2

Assay results for 2A2OA activity. The proteins encoded by YGL202W, YJL060W, YER152C, & YDL168W were expressed from OpenBiosystems Yeast ORF clones and purified. Activity was tested in an assay of NADPH production, based on (23). L- $\alpha$ -amino adipic acid and 2-oxoglutarate were provided as substrates, and pyridoxal phosphate as co-factor. Glutamate production was assayed using commercially available yeast glutamate dehydrogenase, which uses NADP as cofactor and deaminates glutamate producing ammonia & NADPH and regenerating 2-oxoglutarate (see SOM). Also consistent with 2A2OA activity is experimental evidence indicating a higher activity with L- $\alpha$ -amino adipic acid over either alanine and aspartate (see SOM).

### Figure 3

a) Structure of the Robot Scientist investigation (a fragment). It consists of two main parts: an investigation into the automation of science, and an investigation into the reuse of formalised experiment information. The top levels involve AI research (red), while requires research in functional genomics (blue), and systems biology (yellow). Each level of research unit (studies, cycles, trials, tests, and replicates) is characterised by a specific set of properties (see fig S1 and SOM). Such a nested structure is typical of many scientific experiments, where the testing of a top-level hypothesis requires the planning of many levels of supporting work. What is atypical in Adam's work is the scale and depth of the nesting.

### Table 1 The orphan enzymes and Adam's hypotheses

The hypothesised genes are those which Adam's bioinformatics-based hypothesis formation method abduced encoded the orphan enzyme. **Prob** is the Monte-Carlo estimate of the probability of obtaining the observed discrimination accuracy or better using a random labelling of replicates. The discrimination is between the differences in growth curves observed with the addition of specified

metabolites to the wild type and the deletant. **Acc** is the highest accuracy for a metabolite species in discriminating between the growth curves observed with the addition of specified metabolites to the wild type and the deletant. **No.** is the number of metabolites tested. **Existing Annotation** is the summary from SGD of the annotation of the ORF. **Dry** is the summary of whether the annotated function is the same as predicted by Adam. If a gene has already an associated function, we do not consider this to be contradictory to Adam's conclusions unless this function is capable of explaining the observed growth phenotype, e.g. *BCY1*. (ida - inferred from direct assay; iss - inferred from sequences or structural similarity) (5). **Wet** is the result of our manual enzyme assays. (See SOM for details).

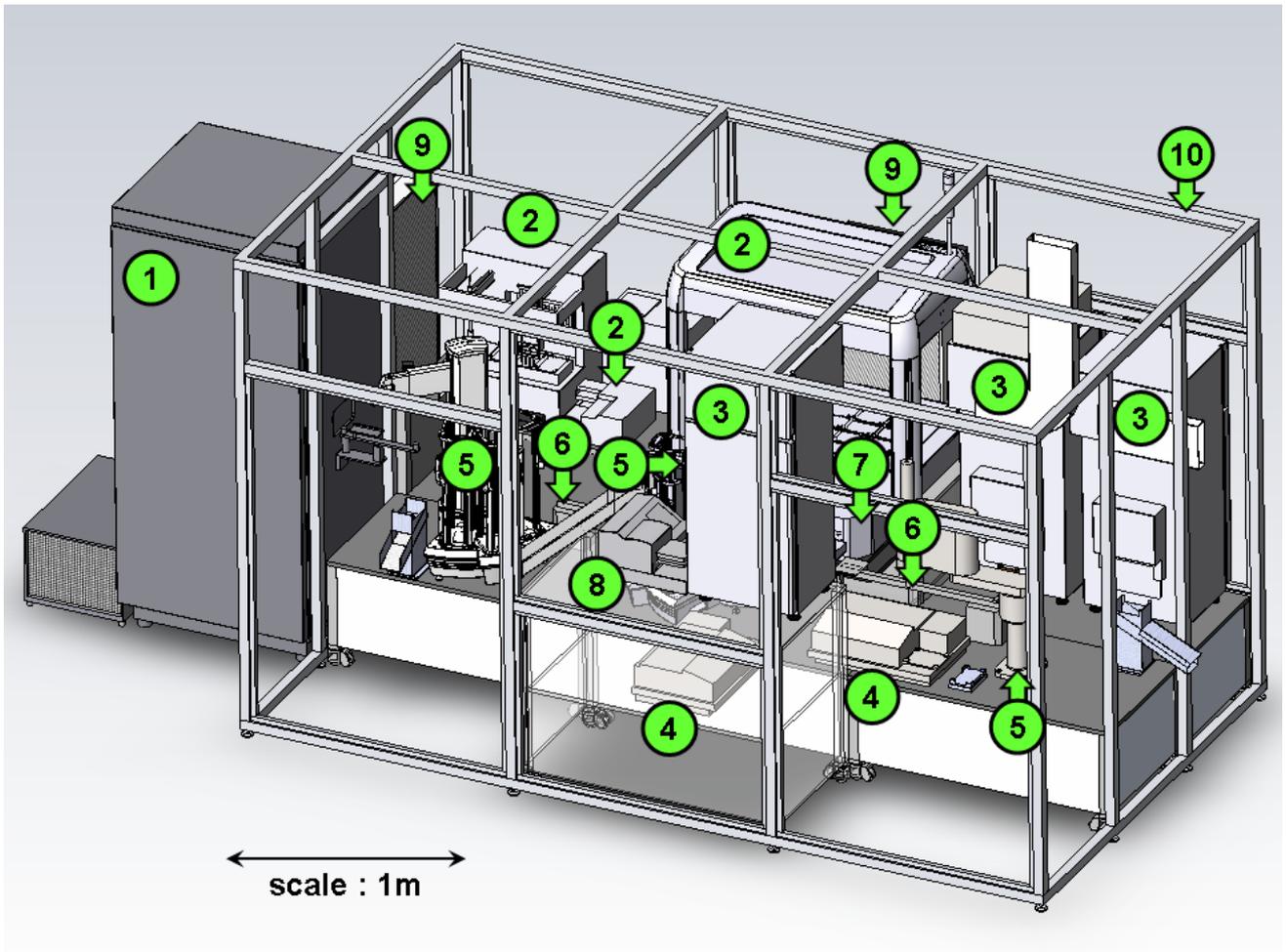


Figure 1

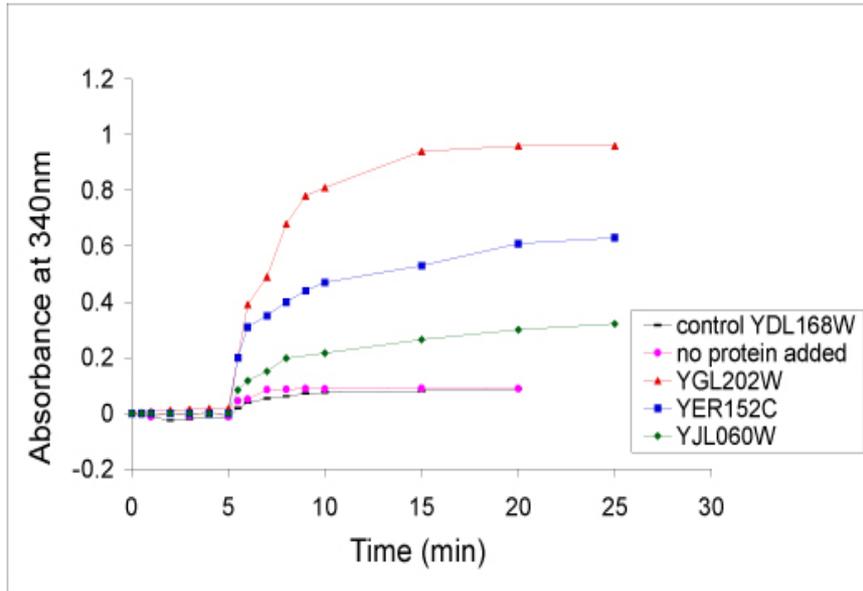


Figure 2

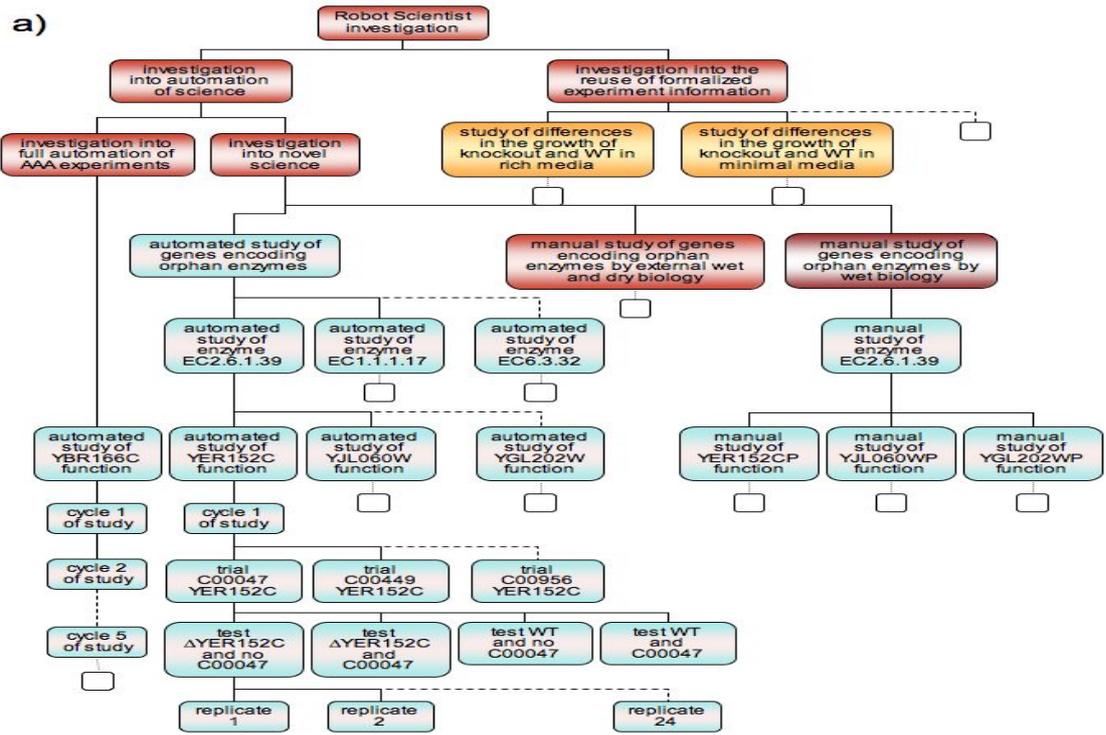


Figure 3

Orphan Enzyme		Hypothesised Gene	Prob.	Acc.	No.	Existing Annotation	Dry	Wet
1	glucosamine-6-phosphate deaminase (3.5.99.6)	YHR163W (SOL3)	<10 <sup>-4</sup>	97	8	'6-phosphogluconolactonase' ida	-	-
2	glutaminase (3.5.1.2)	YIL033C (BCY1)	<10 <sup>-4</sup>	92	11	'cAMP-dependent protein kinase inhibitor' ida	x ?	-
3	L-threonine 3-dehydrogenase (1.1.1.103)	YDL168W (SFA1)	<10 <sup>-4</sup>	83	6	'alcohol dehydrogenase' ida	-	-
4	purine-nucleoside phosphorylase (2.4.2.1)	YLR209C (PNP1)	<10 <sup>-4</sup>	82	11	'purine-nucleoside phosphorylase' ida	✓	-
5	2-aminoadipate transaminase (2.6.1.39)	YGL202W (ARO8)	<10 <sup>-4</sup>	80	3	'aromatic-amino-acid transaminase' ida	✓	✓
6	5,10-methenyltetrahydrofolate synthetase (6.3.3.2)	YER183C (FAU1)	<10 <sup>-4</sup>	80	4	'5,10 formyltetrahydrofolate cyclo-ligase' ida	✓	-
7	glucosamine-6-phosphate deaminase (3.5.99.6)	YNR034W (SOL1)	<10 <sup>-4</sup>	79	2	'possible role in tRNA export'	-	-
8	pyridoxal kinase (2.7.1.35)	YPR121W (THI22)	<10 <sup>-4</sup>	78	1	'phosphomethylpyrimidine kinase' iss	-	-
9	mannitol-1-phosphate 5-dehydrogenase (1.1.1.17)	YNR073C	<10 <sup>-4</sup>	78	6	'putative mannitol dehydrogenase' iss	-	-
10	1-acylglycerol-3-phosphate O-acyltransferase (2.3.1.51)	YDL052C (SLC1)	0.0001	80	6	'1-acylglycerol-3-phosphate O-acyltransferase' ida	✓	-
11	glucosamine-6-phosphate deaminase (3.5.99.6)	YGR248W (SOL4)	0.0002	78	2	'6-phosphogluconolactonase' ida	-	-
12	maleylacetoacetate isomerase (5.2.1.2)	YLL060C (GTT2)	0.0003	76	3	'glutathione S-transferase' ida	-	-
13	serine O-acetyltransferase (2.3.1.30)	YJL218W	0.0005	78	2	'unknown function'	-	-
14	L-threonine 3-dehydrogenase (1.1.1.103)	YLR070C (XYL2)	0.0052	75	6	'xylitol dehydrogenase' ida	-	-
15	2-aminoadipate transaminase (2.6.1.39)	YJL060W (BNA3)	0.0084	73	3	'kynurenine aminotransferase' ida	-	✓
16	pyridoxal kinase (2.7.1.35)	YNR027W	0.0259	76	2	'involved in bud-site selection' iss	-	-
17	polyamine oxidase (1.5.3.11)	YMR020W (FMS1)	0.0289	78	4	'polyamine oxidase' ida	✓	-
18	2-aminoadipate transaminase (2.6.1.39)	YER152C	0.0332	74	3	'uncharacterized'	-	✓
19	L-aspartate oxidase (1.4.3.16)	YJL045W	0.1300	72	1	'succinate dehydrogenase isozyme' iss	-	-
20	purine-nucleoside phosphorylase (2.4.2.1)	YLR017W (MEU1)	0.1421	72	6	'methylthioadenosine phosphorylase' ida	✓	-

**Table 1**