ISBE Infrastructure for Systems Biology Europe

WP9

Technology and Science Watch

ISBE WP9 REPORT

Continuous technology forecasting report

D 9.1

Rüdiger Ettrich

Thomas Höfer

Dimitris Thanos

Sarah Butcher

Tomas Rolinek

Project ref. no.	INFRA-2012-2.2.4: 312455
Project title	ISBE – Infrastructure for Systems Biology Europe
	R= Report
Contractual date of delivery	Month 15
Actual date of delivery	Month 15
Deliverable number	D9.1
Deliverable title	Continuous technology forecasting report
Dissemination Level	PU
Status & version	Version 1
Number of pages	14
WP relevant to deliverable	WP9
Lead Participant	C4SYS, prof. RNDr. Rüdiger Ettrich, PhD.
Author(s)	Rüdiger Ettrich, Sarah Butcher, Dimitris Thanos, Thomas Höfer, Tomas Rolinek
Project coordinator	Richard Kitney
EC Project Officer	Maria Douka

Dissemination level: PU = Public, RE = Restricted to a group specified by the Consortium (including Commission services), PP = Restricted to other programme participants (including Commission Services), CO= Confidential, only for members of the Consortium (including the Commission Services)

Nature of Deliverable: P= Prototype, R= Report, D=Demonstrator, O = Other.

ISBE Infrastructure for Systems Biology Europe

I. TABLE OF CONTENT

II. Introduction 4

A. Objectives 4

B. 4 specific questions 5

III. The report 6

A. Microarray technologies 7

B. Next generation sequencing technologies 8

C. Single cell technologies 8

D. Proteomics technologies 9

E. Metabolomics technologies 9

F. Image technologies 11

G. Dynamic modelling 11

IV. Current demands for integration of technologies into the future infrastructure 13

V. Next steps 14

II. INTRODUCTION

Workpackage 9 is focused on Technology and Science Watch and one of the key deliverables is continuous technology forecasting reports. In this initial period of the ISBE preparatory phase these reports might serve as a guide for building up the infrastructure, and thus mainly serve the workpackages that deal with the planning of the future centres or the engagement of the relevant community, while in the future these reports will serve as a basis for continuous integration of new methodology or technologies into the existing infrastructure. This report does not aim to provide fully representative data covering all fields on the same level, but rather to highlight technology or science demands coming from the scientific community towards the infrastructure. The Science and Technology watch committee, that will consist of appointed scientists or experts from all relevant fields, will start working from January next year and will play a key role in later reports. Thus this initial report shall serve as a starting point that, on the one hand, shall aid the identification of potential members of the future committee, by defining the areas and expertise that need currently to be covered, and on the other hand shall give a first global overview of the existing state-of-the-art with respect to technology or methodology and possible near future directions. Additionally, a chapter dealing with the possible role of the future infrastructure to foster systems biology research is incorporated to fulfil the purpose of inspiring the planned infrastructure as above mentioned. This report represents the essence from tech literature watch, a series of interviews held with scientists from Europe and the United States that work at the forefront of systems biology and are regularly invited as plenary speakers to systems biology conferences, data gained from the initial survey, and data gained from a broad analysis of recent conference proceedings and abstracts.

A. OBJECTIVES

- ➔ Examining and evaluating of the existing state-of-the-art of available technologies
- ➔ Determining whether future technological and scientific developments in the scientific areas of systems biology should integrate these new technologies
- ➔ Estimating which technologies should be provided by the ISBE infrastructure

B. 4 SPECIFIC QUESTIONS

Systems biology research needs collection and processing of data from large numbers of biological experiments using automated procedures and requires the ability to obtain, integrate and analyze complex data sets from multiple experimental sources using interdisciplinary tools.

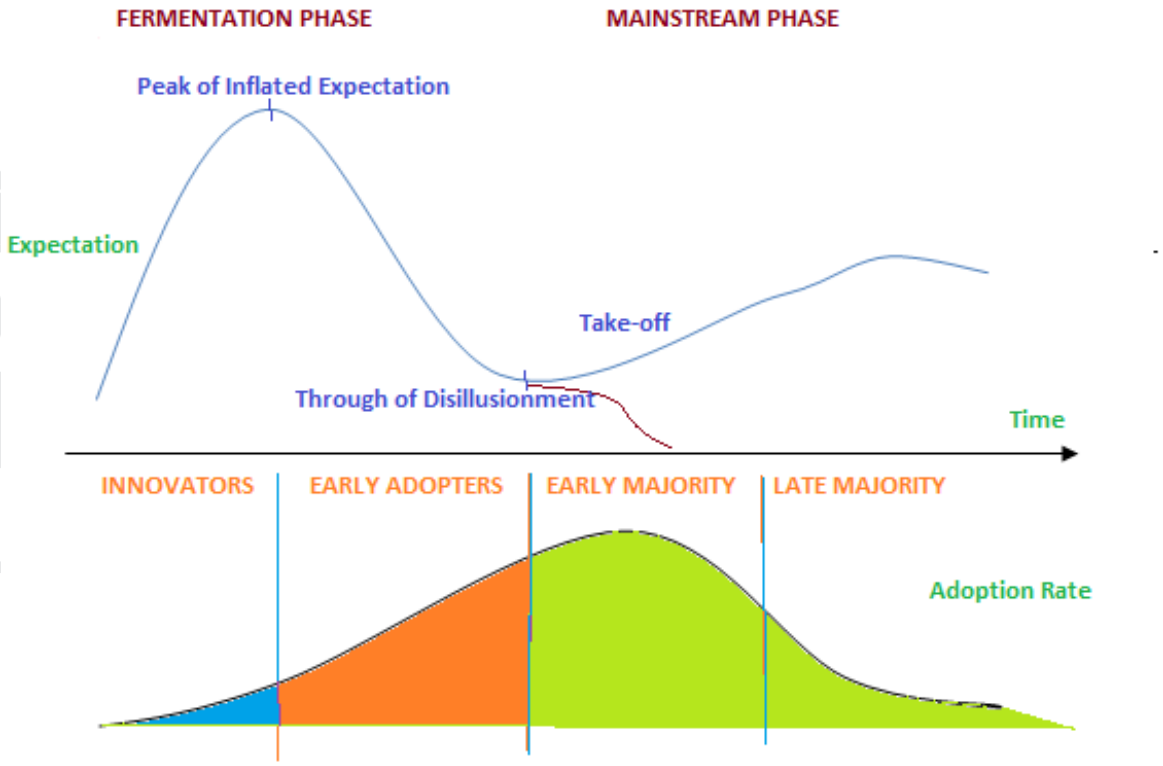
The report addresses four specific questions:

1. What existing technologies are considered as state-of-the-art or key technology in the specific fields of systems biology?
2. Would it make sense to have these available within an infrastructure?
3. What emerging technologies will be important in the near future?
4. How can these new technologies be integrated into Systems biology and how an infrastructure might help with this?

Additionally to tech literature watch, data extracted from the survey and by assembling statistical data of systems biology conference proceedings and abstracts, interviews (lasted approximately 30 min and was initiated with the following questions and the consequent discussions) were conducted with selected scientist that each are on the forefront in at least one of the following fields: Modelling, microscopy & image analysis, live single-cell imaging & modelling, mass spectrometry, proteomics, RNAi screens, genomics & sequencing, metabolomics.

III. THE REPORT

Every new technology that finds its way into science follows a sigmoidal curve, first we have a fermentation phase in which only a small number of scientists are using the respective technique, then we see a take-off, when the technique becomes generally accepted and becomes mainstream, followed by the consequent stagnation when the technique reaches maturity.

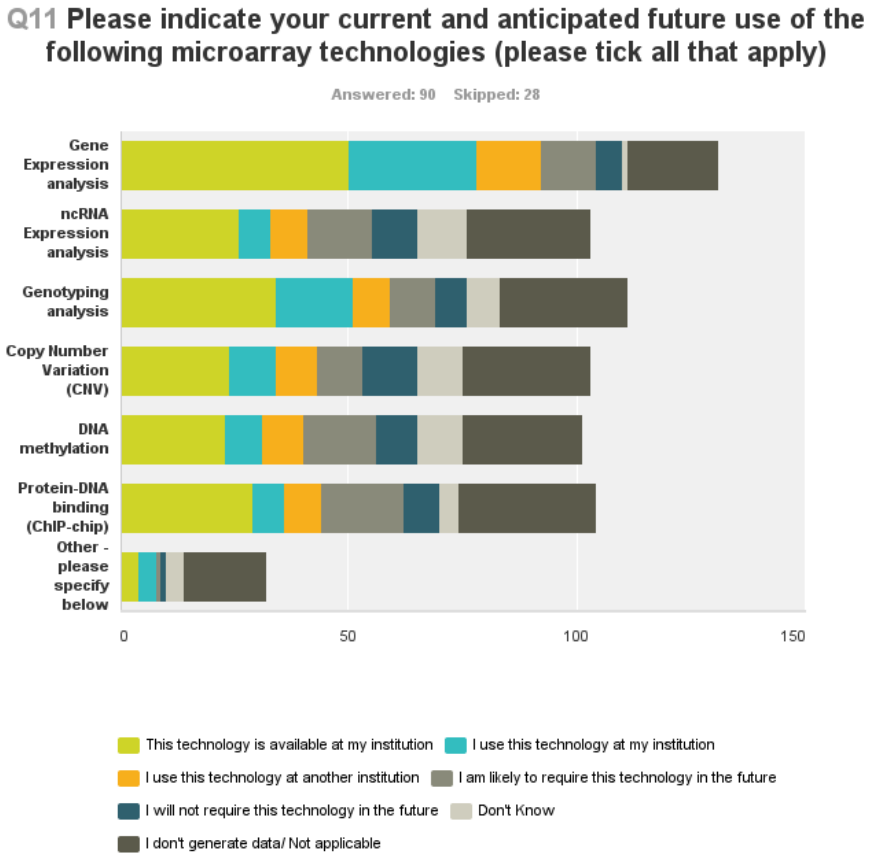


Expectation and Technology Adoption Lifecycle plotted together

This then can be followed by either discontinuity, when the technique is replaced by a newer one, or by an extension, in case the technique is modernized or upgraded. Hereby it is important to note that the revolutionary step, in which a new technique is invented doesn't necessarily change the market and leads to the take-off, but is in general much earlier and is followed by the fermentation that can take a significantly long time. The disruptive step then is not the invention itself but the time when the science is ripe for the new technique and demands it. This makes it extremely difficult to predict which of the new techniques that are developed and reported daily does actually make it to the take-off and will be demanded by the systems biology community. This will therefore be the most difficult task of the science watch committee and our continuous forecasting reports.

A. MICROARRAY TECHNOLOGIES

While a large number of **microarray technologies** is available in a majority of institutions that conduct systems biology and in principal would cover the demand fully (Copy number variant, Gene Expression analysis, Genotyping analysis, ChIP chip) for some of the technologies, as is gene expression analysis, more than 16% of survey respondents use this technology at a different institution and additional 14% require the usage in the future. The reasons for having the experiments done externally need to be analyzed in this case, as this might be a potential task for the infrastructure to cover.

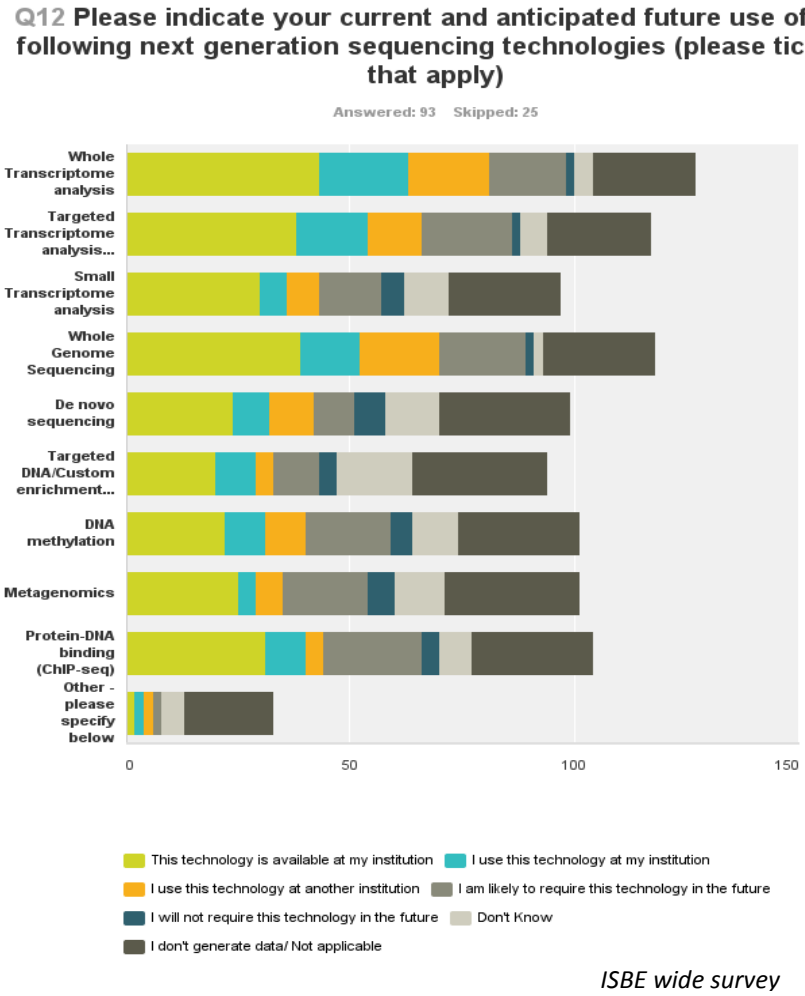


ISBE wide survey

Though currently only a minimum of contributions at systems biology conferences mention **DNA methylation** as their used technique, it is likely to be required by over 19% of the responding scientists in the future and more than 10% use this technology at another institution, however it is available only in less than 28% of the institutions. We can identify this as a potential technique that is not yet used widely but might become a mainstream in the future and should be taken care of by the infrastructure. We can identify the similar situation with **Protein – DNA binding (ChIP chip)**, when it is likely required by over 21% of the responding scientists in the future.

B. NEXT GENERATION SEQUENCING TECHNOLOGIES

With respect to **next generation sequencing technologies** on a first glance it seems that Small Transcriptom analysis, Whole Transriptom Analysis, Targeted Transcriptome analysis (e.g. mRNA-seq or miRNA-seq), Whole Genome Sequencing and De novo sequencing is available in institutions and covering their demand. What we can identify as potential techniques that are not provided in sufficient amount by home institutions and could be taken care of by the infrastructure are: **Metagenomics** which is likely to be required by more than 22% of survey respondents and used by 6% at different institutions, and by 4% of respondents at their institutions and **DNA methylation** is likely to be required by almost 23% of the respondents, used by 11% of correspondents at other institutions and is available only in 26% of institutions.



ISBE wide survey

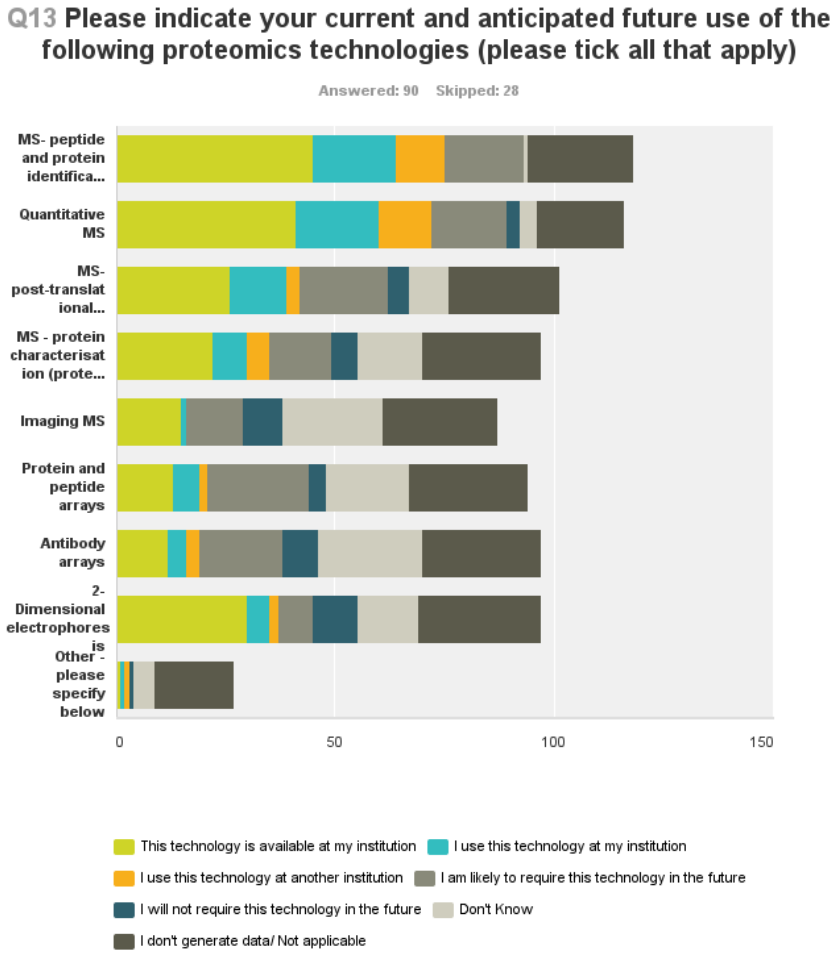
C. SINGLE CELL TECHNOLOGIES

Currently, a larger number of scientists are really pushing cutting edges on single cell sequencing and on RNA level. Single cell is increasingly getting important for proteomics, RNA, DNA. It seems that though the techniques are available in principle that price is still a big issue and the infrastructure might provide grants that would allow the scientists to pick up more transcripts for a critically lower price because if one wants single cell sequencing of RNA, there are thousands and ten thousands cells. In general, scientist agree that next generation sequencing is not still at the peak of its usage, and near-future development will be again Next gen sequencing based but pushing the cost standard (developing library technologies libraries to get more accurate, how to read a longer bit of DNA, to build

better single cell RNA sequencing accurate as possible to capture as many transcripts per single cell). A dream list technology might be a technology that reads of histom markup of DNA. In the opinion of most respondents also current technologies – illumino machines, high and low level analysis – have still a solid future.

D. PROTEOMICS TECHNOLOGIES

Mass spectrometry based techniques have become widely available in recent years. Nowadays, 2 out of 3 papers in the Nature-Science group are using mass spectrometry and from 2007, when the orbitrap technique became available, mass spectrometry papers have tripled. MS-peptide and protein identification, Quantitative MS, MS posttranslational modifications, Protein and peptide arrays, Antibody arrays and 2-Dimensional electrophoresis for **proteomics** are available in a majority of responding institutions and fully cover their demand. However, there are two techniques: **Protein and eptide arrays** – is available only in 16% of institutions but more than 29% require the usage in the future and **Antibody arrays -** is available only in 15% of institutions but more than 23% require the usage in the future. These techniques should be monitored in the near future and might be worth to implement into the infrastructure.



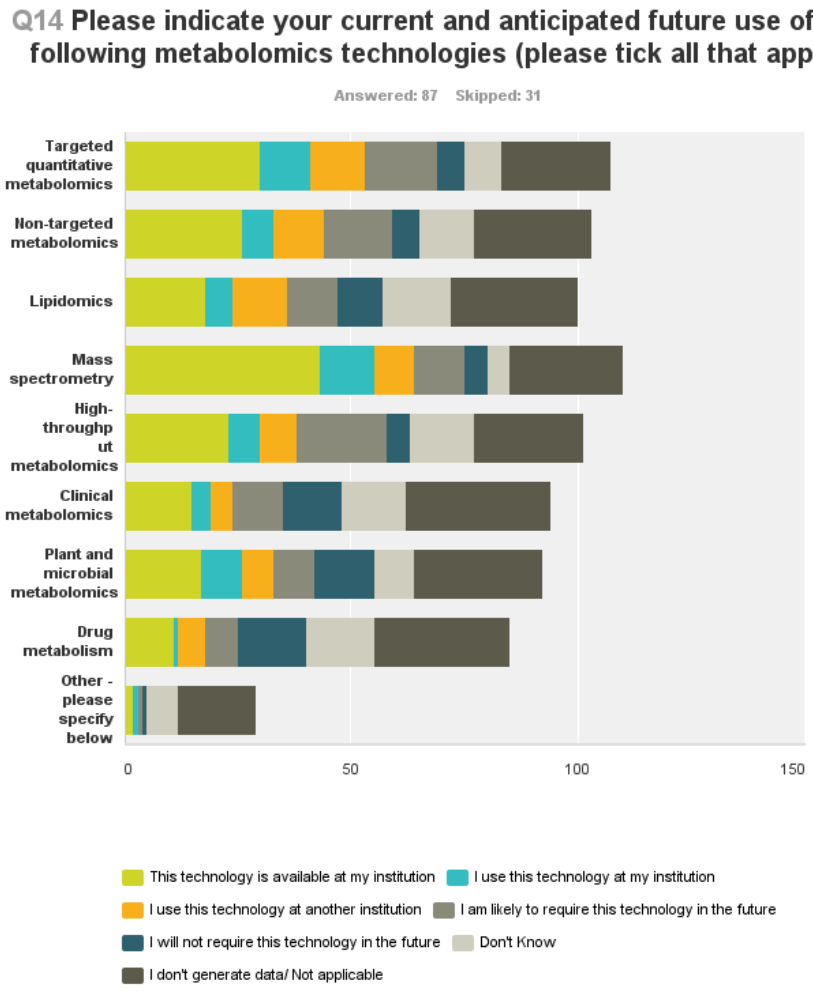
ISBE wide survey

E. METABOLOMICS TECHNOLOGIES

Targeted quantitative metabolomics, and Mass spectrometry are **metabolomics technologies** available in more than 36% of institutions. We can identify Mass spectrometry (52%) as the mainstream of available technique at institutions, but still only

15% (Mass spectrometry) and 13% (Targeted quantitative metabolomics) of survey respondents use these techniques at their institution. The reasons for this need to be understood and might open the question if the infrastructure should interface these techniques in the future. Other techniques (Non-targeted metabolomics, Plant and microbial metabolomics, Highthroughput metabolomics, Clinical metabolomics) are in balance between their availability in their institutions (33%) and being likely to require in the future. **Highthroughput metabolomics** might give opportunity for the infrastructure to take care of because 24% of correspondents is likely to require this technology in the future and 10% use this technology at another institution.

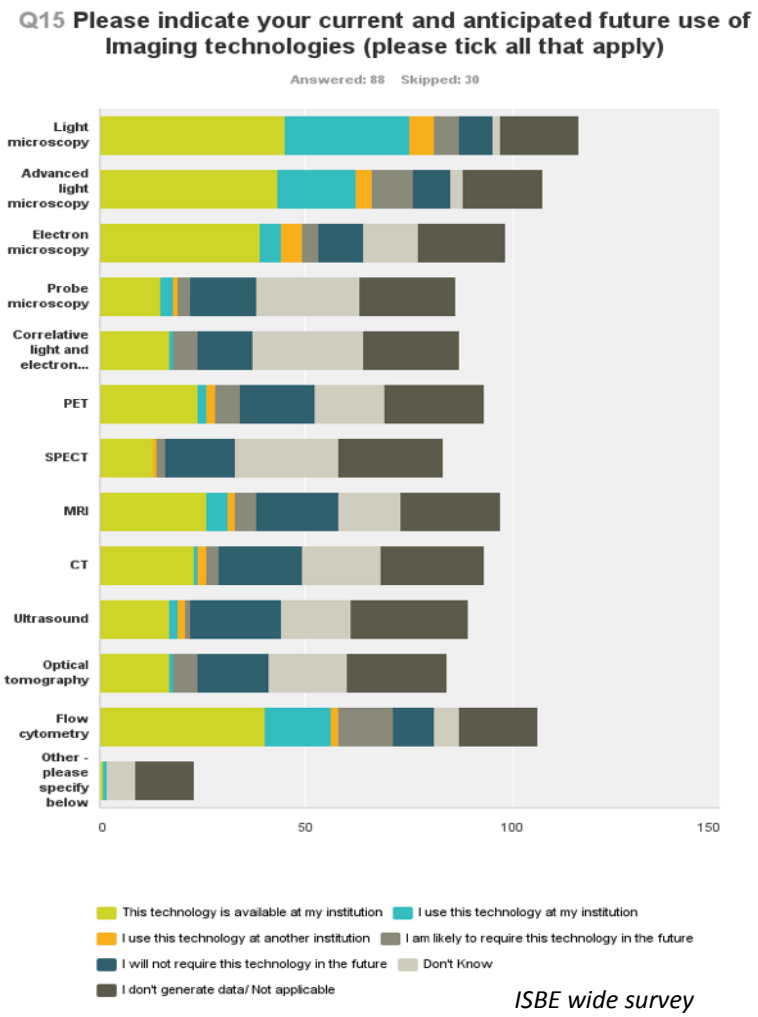
Future directions for protein or protein metabolite interaction might be monitoring of the cooperative mechanism. This would be about cooperativity rather than charting interactions one by one, quantitative deregulation implies allosterism, not only interactions happen all the time. Upcoming high-throughput techniques might be lab scale surface plasmon resonance (SPR), or signal machine like electrophoresis machines where you can really measure ass/disassociation constants on the lab scale of a very tiny material. Characterization of protein complexes going back to bench "old fashioned" chemical methods like gel filtration might become important again, however miniaturized and implemented on a micro platform to be used in large-scale screenings.



ISBE wide survey

F. IMAGE TECHNOLOGIES

Most technologies (Light microscopy – 52%, Advanced Light microscopy – 52%, Electron microscopy – 48%) are available in their institutions and cover the demand fully. In comparison with these technologies other **Image technologies** (Probe microscopy, Correlative light and electron, PET, SPECT, MRI, CT, Ultrasound, Optical tomography) are not generally available in the responding institutions but they seem to cover the demand fully, at least the scientists are currently not aware of a required demand in the future. In microscopy there are couple of techniques that might have taken the revolutionary step (as **2PPM**) and are currently in the fermentation phase, and it will be necessary to monitor if their potential is indeed demanded by future science in systems biology.



ISBE wide survey

G. DYNAMIC MODELLING

Key technologies are mathematical modelling and analysis software (SW) for simulation (such as Matlab, Mathematica, Maple, xppaut and diverse systems biology add-ons for those such as the SBPOP PACKAGE (formerly called Systems Biology toolbox). SW will include: Parallel implementations of deterministic and stochastic simulators and analysis tools; model editing/annotation/visualisation tools; standard model exchange language, both textual and graphical. Data integration: standard data description language: tools for integrating proteomics and genomics data from existing databases (DBs) and new experiments.

Experimental perspective: multiplex assays that can measure several intracellular concentrations in one sample, since these facilitate the generation of high-density time-course and perturbation data for model calibration. Single-cell mass spectroscopy (also known as mass cytometry). Imaging flow cytometry. Both, single-cell mass spectroscopy and imaging flow cytometry are not well established and are being constantly improved, hence their importance will increase in the near future.

Rule based or related for large scale modelling, use of multiple data sources: quantitative techniques like proteomics and phosphoproteomics, reverse protein arrays, plus genomics data and single cell measurement techniques take important role. We will need efficient computational methods to extract the information from the ever increasing datasets and DBs.

Clearly Genome-Scale Metabolic models find the widest application in industrial biotechnology in terms of mathematical modelling. Besides this any omics technology that enables rapid phenotypic characterization is useful - in particular RNAseq and proteomics. Metabolomics used less as it is the most difficult set of data to integrate and use for guiding metabolic engineering.

IV. CURRENT DEMANDS FOR INTEGRATION OF TECHNOLOGIES INTO THE FUTURE INFRASTRUCTURE

Between the respondents and the views represented in various literature reports is a remarkable agreement in the opinion that the big projects are heavily funded for data collection and underfunded on data analysis. **A useful infrastructure thus should not just provide `huge machines`, but rather complex expertise with a strong emphasis on informatics necessary for sharing and analysing data and modelling.** In the context of genomics data basing and the ability to share data seems to be an issue, while instrumentation is relatively available. A significant challenge in this field is seen is to enable access to existing data. Herby, an infrastructure can help to make experiments by teams of scientists from different fields to get from one experiment data they can use in their specific research area. **Data sharing is generally seen as a critical limitation.**

Better integration of large scale-data and DBs into the modelling process is needed. Most DBs lack the kinetic information, including rate parameters. The standards are importantly, release of raw data and models in standard format upon publications, implementation of easy to use tools for automatic import/export. Single-cell mass spectroscopy allows for multiplexed measurement of up to 100 molecules and phenotypes on the single-cell level with high throughput. The quantitative data obtained with such a measurement can be used to reconstruct topology of signalling networks and their dynamics. Combining flow cytometry with imaging makes it possible to correlate the molecular state of the cell with its morphological changes. Additionally, spatial localisation of molecules can be tracked which provides valuable information for computational modelling of signalling networks. **The key is to make things standard, for exchange and reusability of both models and data.**

Generally, **there is agreement that core facilities in informatics are not meant in terms of storage but in terms of staff people who offer their expertise in limited time projects** (several months) and will be working on analysing and giving data. It needs specific types of people to do that and the problem could be a tension between giving community service versus their career, which is then an issue that needs to be solved in the implementation of the infrastructure. These people (employees of ISBE) nevertheless cannot be just service personal but need to be embedded in the scientific process, too, not to miss the development and new trends in new technologies usage.

A great chance for ISBE infrastructure is seen not only in the access to very expensive technologies, but also in enabling access to very simple technologies for a longer period to conduct massive parallel experiments or large-scale repetitions to get more robust data. Consolidation of methods and concept makes also a lot of sense. There is for example one repository on genome-scale metabolic models run by us that is gaining wide use (BioMet toolbox - found via www.sysbio.se). Further expansion of this will surely be useful for the community.

Technology must be widely accessible. It is not sustainable if high performances analysis can be done only in max 5 labs, nowhere else and thus the analysis relies on specific scientific collaboration only. Some private institutions offer experiments with data generation and analysis nowadays to their communities – EMBO [there are some genomic core facilities], EMBL. If you are not the member of these institutions, it is really difficult to get access to their data and these paid surveys are also quite expensive.

Infrastructure is good for genomic and proteomic data generation but it is more important to do the same with the analysis of data which will be more difficult because they have to be customised. Technology should deal with it. This will also demand forecasting the specific requirements and future needs for hardware and storage, which will be dealt with in a separate report devoted to this topic. BGI Americas (the largest genomics centre in the world) providing comprehensive sequencing and bioinformatics services for medical, agricultural and environmental applications is a good example of a working core facility.

V. NEXT STEPS

The next steps are planned:

(i) to establish science and watch committee

(ii) To identify the reasons why correspondents use techniques at other institutions although these techniques are also available in their institutions

(iii) to monitor ant put stress on techniques which is likely to require in the future and now used by correspondents at other institutions and identify their potential to become mainstream techniques

(iv) to enable access to existing data. Data sharing is generally seen as a critical limitation

(v) to analyse Better integration of large scale-data and DBs into the modelling process

(vi) to enable access not only to very expensive technologies, but also to very simple technologies for a longer period to conduct massive parallel experiments or large-scale repetitions to get more robust data in compliance with the idea that Technologies must be widely accessible

(VII) to concentrate on the Technology should deal with the analysis of data which is generally seen in comparison with data generation more difficult because they have to be customised