ISBE Infrastructure for Systems Biology Europe

WP9

Technology and Science Watch

ISBE WP9 REPORT

Continuous technology forecasting report

D 9.3

Committee members as co-authors

Rüdiger Ettrich

Dimitris Thanos

Sarah Butcher

Marcela Kotrcova

Natalie J Stanford

Carole Goble

Angela Oberthuer

Project ref. no.	INFRA-2012-2.2.4: 312455
Project title	ISBE – Infrastructure for Systems Biology Europe
	R= Report
Contractual date of delivery	Month 33
Actual date of delivery	Month 36
Deliverable number	D9.3
Deliverable title	Continuous technology forecasting report
Dissemination Level	PU
Status & version	Version 1
Number of pages	29
WP relevant to deliverable	WP9
Lead Participant	C4SYS, prof. RNDr. Rüdiger Ettrich, PhD.
Author(s)	Rüdiger Ettrich, Sarah Butcher, Dimitris Thanos, Marcela Kotrcova, Natalie J Stanford, Carole Goble, Angela Oberthuer
Project coordinator	Richard Kitney
EC Project Officer	Keji-Alex Adunmo

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

TABLE OF CONTENT

I. INTRODUCTION

The third continuous technology report is the result of the joint effort of the WP9 members and input from the Technology and Science Watch committee appointed by the steering committee in April 2014. This updated report shall direct the interface/discussion with the Research infrastructure centres that may provide the experimental methodologies or technologies utilised by the modelling and data stewardship canters during systems biology experimental design.

The comments from reviewers in individual areas of expertise are not themselves recommendations for an exact infrastructure. However these, together with the systems biology community surveys, continue to highlight technologies that will be required both immediately and in the near future

We also provide a fine-grained view from the appointed experts within the Science and Technology watch committee within their relevant fields. Thus the report gives a first global overview of the existing state-of-the-art of molecular systems biology with respect to technology or methodology and possible near future directions.

METHODOLOGY

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

This report represents the essence from

- Technological literature watch (The Sociology of Expectations in Science and Technology)
- Reports from the Science and Technology Watch Committee members (see appendix)
- A series of interviews held with scientists from Europe and the United States that work at the forefront of systems biology and who are regularly invited as plenary speakers to systems biology conferences
- Data obtained from a broad analysis of recent conference proceedings and abstracts.

In the report we address the following fields:

Modelling, microscopy & image analysis, live single-cell imaging & modelling, mass spectrometry, proteomics, RNAi screens, genomics & sequencing, metabolomics.

A. OBJECTIVES

➔ Examination and evaluation of the existing state-of-the-art available technologies

➔ Determination of whether future technological and scientific developments in the scientific areas of systems biology should integrate these new technologies

II. THE REPORT

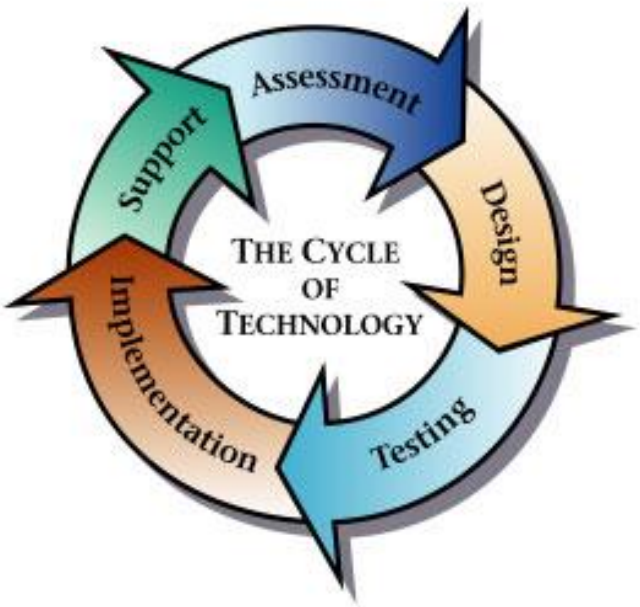
In order to delimit the use of different state-of-the-art technologies, we also discussed and researched the Technology cycle, its stages and what plays an important role before the technology becomes generally adopted. Before we can forecast important technology in Science we need to take into consideration expectations and visions that are also important for other groups beyond scientists, and engineers.

They play a central role in mobilizing resources both at the macro level, for example in national policy through regulation and research patronage, and at the middle level of sectors and innovation networks, and at the micro-level within engineering and research groups and in the work of the single scientist or engineer. For these and other reasons, analysing the dynamics of expectations is a key element in understanding scientific and technological change. One of the main reasons for this is, we would argue, because expectations frequently serve to bridge or mediate across different boundaries and otherwise distinct (though overlapping) dimensions and levels. Expectations are fundamental in the coordination of different communities. They also change over time in response and adaptation to new conditions or emergent problems - temporal coordination. (http:// Utwente.nl)

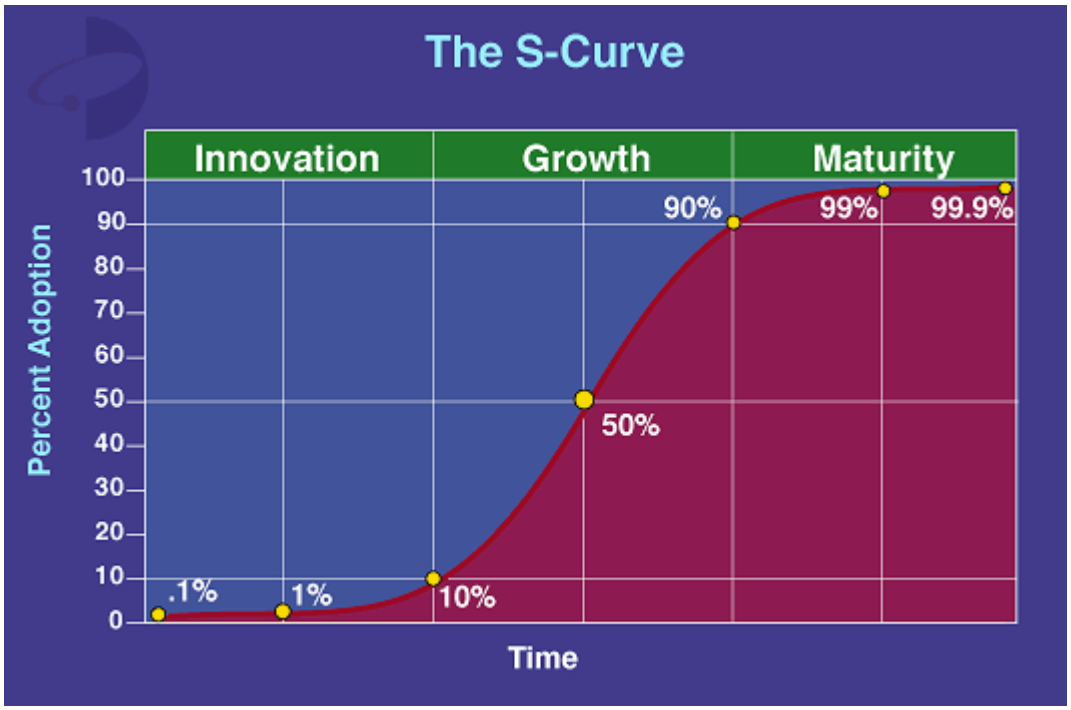
Our update from members of the Technology and Science Watch Committee (TSWC) is a collection of scientists' inputs coming from various background and communities (Germany, Ireland, UK and Slovenia) therefore we could expect that the state of-art-technology examples mentioned by members of TSWC could be tied to their specific country and experience.

Expectations link technical and social issues, because expectations and visions refer to images of the future, where technical and social aspects are tightly intertwined. Finally, expectations constitute 'the missing link' between the inner and outer worlds of techno-scientific knowledge communities and fields. At the same time, expectations and visions are often developed and reconstructed in material scientific activities and disseminated in obdurate and durable forms. In a sense, expectations are both the cause and consequence of material scientific and technological activity.

Expectations and visions in science are closely connected and driven by need and expectation throughout the technology lifecycle as shown in the figure below:



Every new technology that finds its way into science follows a sigmoidal curve of adoption,. This starts with fermentation phase in which only a small number of scientists are using the respective technique, this is followed by a take-off phase, when the technique becomes generally accepted and becomes mainstream, and finally by the consequent stagnation phase when the technique reaches maturity. (Sandstrom)



For our report we collected information on state of-art-technology from experts in the following fields:

- Microarray technologies
- Next Generation sequencing technologies
- Single cell technologies

- Proteomics technologies
- Metabolomics technologies
- Image technologies
- Dynamic modelling technologies

A. MICROARRAY TECHNOLOGIES

Microarrays were developed as one of the first high throughput molecular biology techniques nearly 20 years ago and are still in widespread use today. They enabled the lab scientist to determine in parallel the expression profiles of all genes in a biological sample. Microarray technology has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of phenotypic anomalies in many organisms. Microarrays continue to be used for a wide range of applications including systems biology, developmental biology, gene discovery, drug discovery, disease diagnosis and toxicological research. To date over 1.5 million samples (126,145 to date in 2015 alone) from 58,000 experiments using over 10,000 different microarray types have been submitted to the public microarray data repository Gene Expression Omnibus (GEO).

The microarray has been adopted for a number of different techniques including:

- **ChIP-chip:** determination of protein binding site occupancy and histone modifications.
- **SNP detection (SNV)**: Identifying single nucleotide polymorphism among alleles within or between populations, which are used for a number of applications including genotyping, forensic analysis, measuring predisposition to disease, identifying drug-candidates, evaluating germline mutations in individuals or somatic mutations in cancers.
- **Copy Number variation (CNV):** detection of abnormal number of copies of genes, or sections of DNA
- **Comparative genomic hybridisation**: Assessing genome content in different cells or closely related organisms
- **Methylation arrays:** a technique to map methylation changes in human DNA.

And more recently microarrays have been developed for number of different domain types including tissue, protein, peptide and carbohydrate arrays.

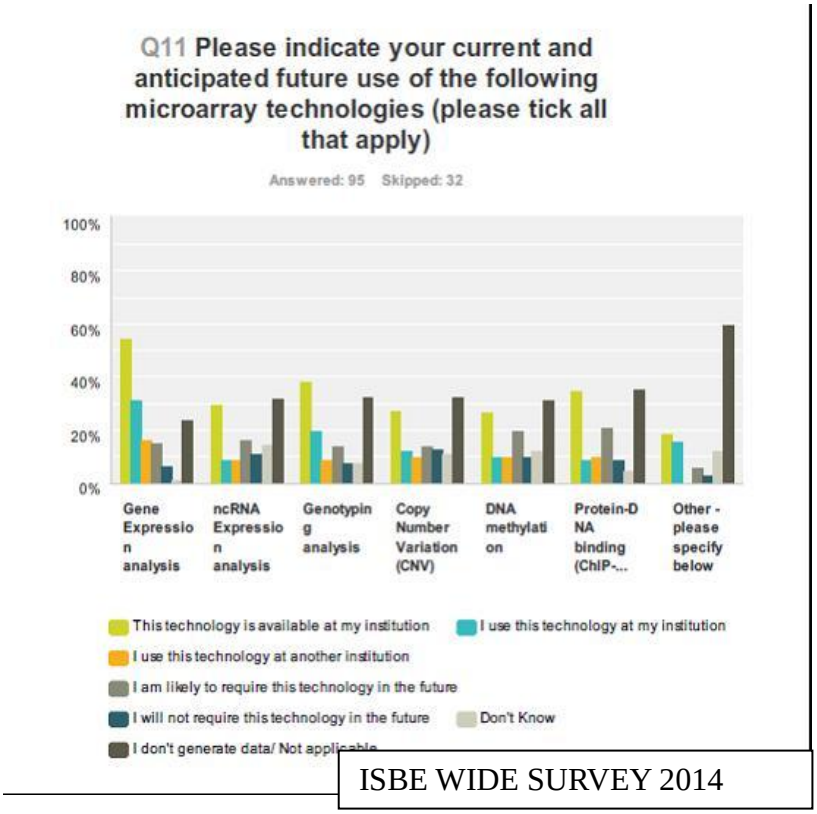
High-throughput transcriptomic technologies (microarrays, RNA-seq) usually require additional confirmation of results using other methods. This is mostly accomplished by using real-time (RT) PCR on a subset of the selected candidate genes. Recently RT-PCR is itself undergoing a high throughput revolution, with the development of digital PCR (dPCR), allowing greatly increased numbers of genes for this confirmation stage. An example is the Biomark HD System from Fluidigm, which enables performance of real-time PCR reactions in 96 sample sets, where each sample is tested in 96 assays. Other available dPCR platforms are state-of-art technology with respect to absolute quantification, however throughput in terms of samples and assays is not as high as with the previously mentioned Biomark HD. It is highly likely that other platforms for dPCR will become even more efficient in handling common systems biology experimental setups in the near future (Dobnik, appendix I).

All microarrays require the generation and binding of specific DNA target fragments to a solid support in known patterns – the array. A typical microarray experiment involves the specific hybridization of an mRNA molecule generated under the experimental conditions to the array-based corresponding DNA template. A single microarray contains many

hundreds of *thousands of DNA template fragments,* together with suitable controls. Measurement of mRNA binding to each DNA fragment (probe) is quantitative and is related to the expression level of the various genes under study. Once suitable controls are applied, the data are analyzed statistically and a profile is generated for gene expression in the cell.

In reference to our updated reviews from TSWC and latest ISBE Wide Survey, a large number of **microarray technologies** are available in the majority of institutions that conduct systems biology and in principal would cover the demand fully for some of the requested services of microarrays.

However there were a substantial proportion of survey respondents (16%) who use an external service provider, and additional 15% who will require the usage [of microarray technologies or external service provider] in the future. Several reasons were cited for having the experiments done externally, including accessibility, costs, bureaucracy (internal), easier to use (some private companies include basic analysis). This could point to a potential mediating role for the ISBE infrastructure.



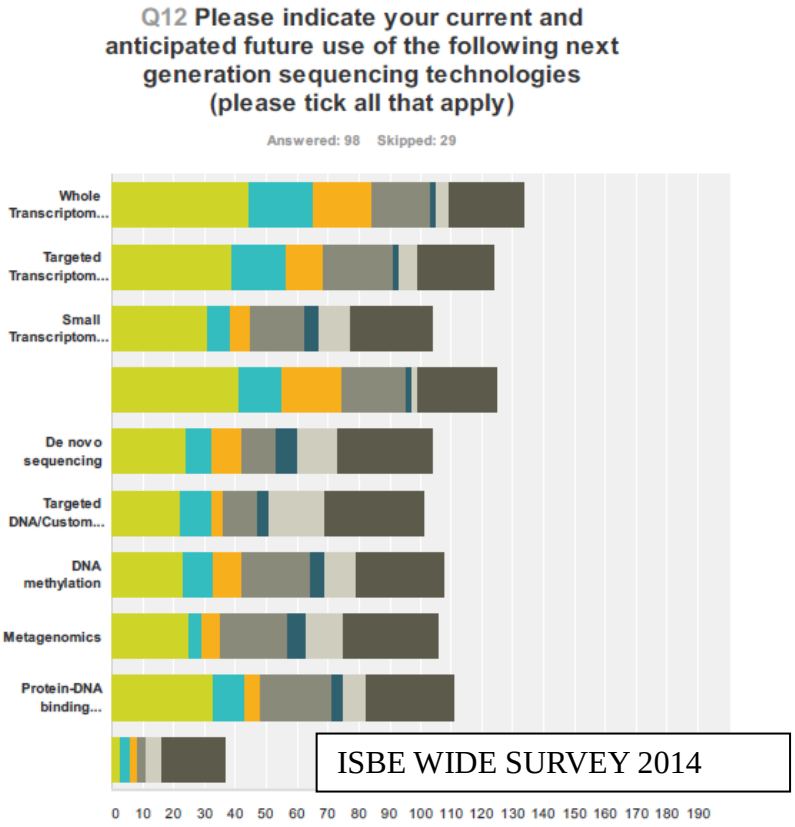
ISBE WIDE SURVEY 2014

Though currently only a minimum number of contributions at systems biology conferences cite **DNA methylation** as their used technique, it is likely to be required by over 20% of the responding scientists in the future and over 10% already use this technology at another institution. In contrast, it is available in less than 26% of the institutions. We thus identify DMA Methylation arrays as a technique that is not yet used widely but might become mainstream in the future and thus should be taken care of by the infrastructure. Via the ISBE Wide survey in answers to question 11, 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

Since the so-called Next generation sequencing (NGS) technologies became available for practical use in the late 1990's there has been a paradigm shift in the way that sequencing is used in all fields of molecular biosciences. The field continues to change relatively rapidly, with numerous companies bringing products to market in waves; with the continual promise of decreased cost (per base), speed of data aquisition and/or increasing read length allowing ever-more high throughput experimentation. Improvements in sample preparation techniques, together with decreasing requirements in input DNA levels are also driving single-cell studies, where the entire DNA (or indeed RNA) complement of a single cell or small group of cells is sequenced. Furthermore, some newer types offer the promise of direct characterisation of DNA base modificatons including methylation (Mavis *et al* 2013. In fact, we are already utilising the third generation of sequencing technology, with further technology changes already showing promise in methods development publications. Broadly, co-existing technologies may differ in the template preparation methodologies (shearing, size selection, presence/absence and number of amplification steps), the method used to produce a measurable 'signal' at each base (ligation of specific linkers, fluorophores, amplification steps), the actual overal mechanism itself - pyro-sequencing, sequencing by synthesis etc, and the physico-chemical method of signal measurement (fluorphore light emission followed by image analysis, electrical voltage change). These differences influence the relative effectiveness of different sequencing technologies for some experimental types and this is expected to lead to the continued requirement for access to more than one NGS platform. For example, the ability to produce read-lenths of over 10kb such as currently in production with Pacific Biosciences SMRT may be crucial for closing bacterial genomes rapidly, also improving assembly of problem repetitive regions and scaffolding other shorter reads. even though the cost per base is considerably more, and the error rate far higher than Illumina sequencing.

Next generation sequencing technologies (NGS) are used for whole-genome and targeted sequencing, small RNA discovery, transcriptome analysis, metagenomics, methylation profiling (epigenetics), and genome-wide protein-nucleic acid interaction analysis. Illumina dominates the next generation sequencing market with its suite of systems (MiSeq, HiSeq 2500/ X Ten and Nextseq 500). In 2012 Illumina had a market share of 56%, rising to 80% in late 2014 *(*Forbes 2014). Life Technologies (Ion Torrent), Roche (454) and Pacific Biosciences (PACBIO RS II, a)ll) also provide state-of-the-art systems and have recently been joined by Oxford Molecular's MinIon system. As each company provides systems with varying strengths and weakness it may be important for ISBE to provide or broker access to a number of different systems, as the best system depends on the required application. For example, each of the following applications RNA-sSq, exome sequencing, other targeted sequencing or novel genome assembly are each best suited to different systems. In addition, the most cost effective system partly depends on the number of samples being run (e.g. Illumina HiSeq for large numbers and Ion PMG or Illumina MiSeq for small sample numbers) and the read depth required. (Duffy, appendix II)



An initial investigation of the requirements survey results for **next generation sequencing technologies** suggests that Small Transcriptome analysis, Whole Transcriptome Analysis, Targeted Transcriptome analysis (e.g. mRNA-seq or miRNA-seq), Whole Genome Sequencing and De novo sequencing are already available in institutions and satisfying current demand. What we can identify as potential techniques that are not provided in sufficient amount by home institutions and which could be facilitated by the infrastructure are: **Metagenomics** which is likely to be required by more than 23% of survey respondents and used by more than 6% at different institutions, and by 4% of respondents at their own institutions; and **DNA methylation studies** which are likely to be required by almost 25% of the respondents, used by 11% of correspondents at other institutions and which is available only in 26% of institutions.

The high costs of methylation studies using NGS are seen as a major current obstacle to uptake. However, the history of other similar techniques has demonstrated that cost might come down very fast once the technique is established, and so it is likely that this will start to change soon for this case, too. At the horizon we already can get a glimpse some knowledge of third generation sequencing technologies e.g. Pacific Bioscience systems for this type of sample/study. The methylation status of DNA can be sequenced straight from the DNA with minimal preparation of the DNA sample. This technology is currently prohibitively expensive for individual labs however it may be worth collaboratively outsourcing samples or identifying a central point to which samples can be sent. (Kenny, Appendix II)

NGS technologies are capable of producing vast amounts of genomic data rapidly, and analyses often also require the storage of large-scale reference datasets (e.g. reference genome assemblies, known variants) and interim data types that may run to 10's of terabytes per experiment. Our TSWC members stated in their inputs that local data storage at research institutions may become a severe limitation unless carefully managed, despite the decreasing cost of storage. In addition to improved algorithms for data compression (e.g. reference-based compression (Cochrane *et al* 2012), some sequencers are already able to ameliorate this by directly uploading data to a cloud. Like with all use of public or indeed private clouds, the issue of data ownership and security needs to be carefully addressed; nevertheless the potential of having a central repository/pipeline for data storage and analysis utilizing cloud computing seems attractive.

A recent publication described RNA sequencing in situ (Je Hyuk Lee et al., 2014) Highly Multiplexed Subcellular RNA Sequencing in Situ, Science vol. 343, March 2014). This technique provides a combination of *in situ* library preparation and sequencing. At this time it has a limited amplicon length of 27 base pairs and the number of sequenced amplicons is low. The correlation coefficient, when compared to Illumina is still quite low (0.5 – 0.7), however the technique offers the advantage of subcellular localization, so you can visualize the localization of the original RNA template for each sequenced amplicon. This technique is one example of one at the rise of expectation area within the fermentation phase of the adoption lifecycle and is not yet fully tested. (i.e. not sufficiently mature to be offered as part of ISBE infrastructure).Nevertheless this kind of technique might be of great importance in the future. (Dobnik, appendix I).

C. SINGLE CELL TECHNOLOGIES

Technologies enabling studies of individual cells are a fast-moving growth area.

A large number of scientists are pushing back the boundaries on single cell DNA sequencing and on RNA profiling. Single cell protocols are also increasingly important for proteomics. It seems that although the techniques are available in principle, that price is still a big issue for uptake to become more widespread. A considerably lower price could increase throughput, for instance RNA sequence profiling of thousands or 10's of thousands of individual cells per experiment rather than the 10's to hundreds now possible. One suggestion for facilitating this via ISBE infrastructure is for the nationally funded services centres to provide grants that would allow the scientists to leverage lower prices, for enhanced throughput in critical areas.

In general, scientist agree that next generation sequencing is not still at the peak of its usage, and near-future development will push the cost standard forwards (chiefly by developing library technologies, increased sequencing accuracy, increased read length, and in particular improved single cell RNA sequencing accuracy) all of which will facilitate experiments that seek to capture as many transcripts per single cell as possible. A blue sky list might include be a technology that directly reads histone mark-up of DNA. In the opinion of most respondents, current technologies for both platforms and analysis types still have a solid future.

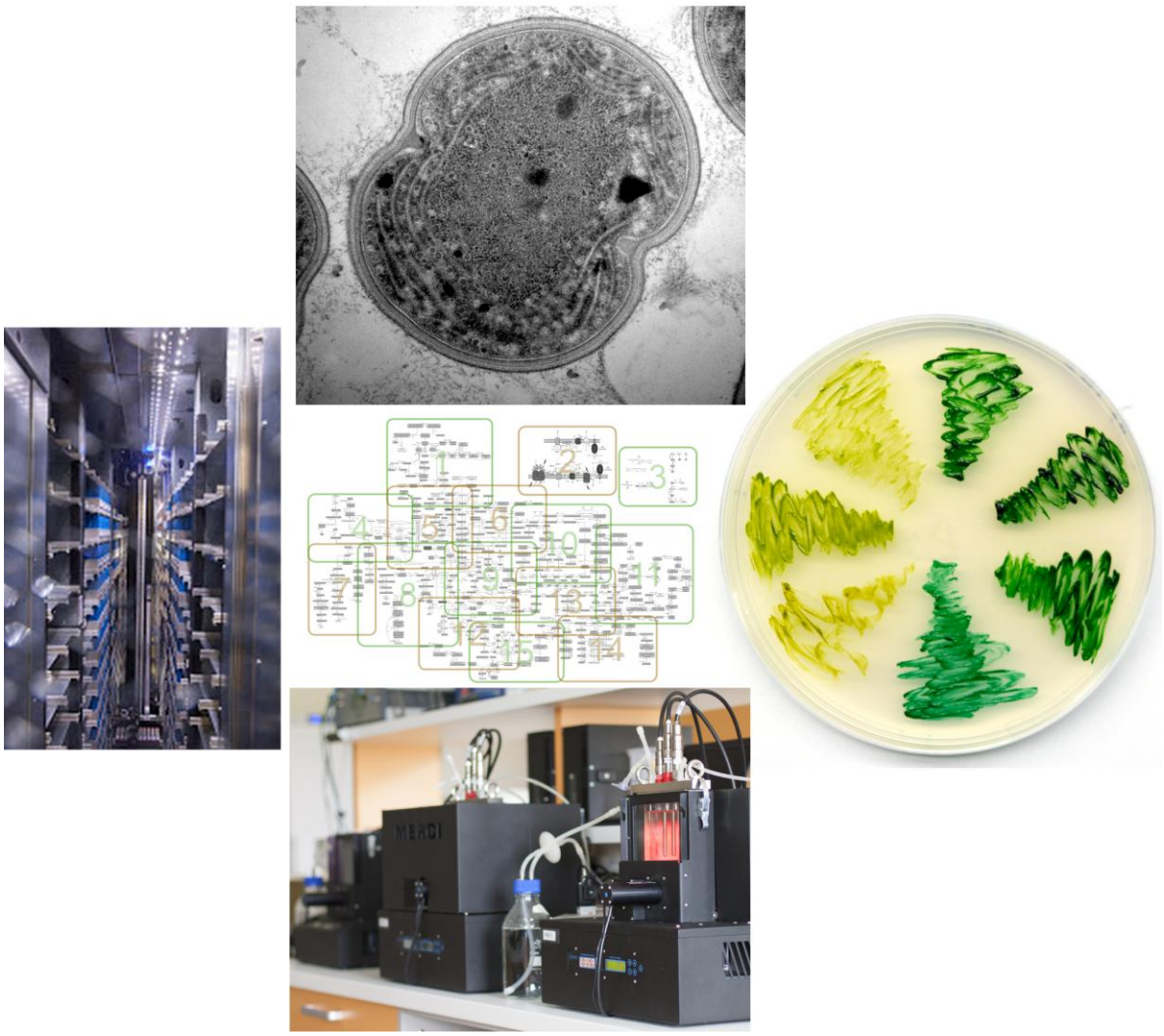


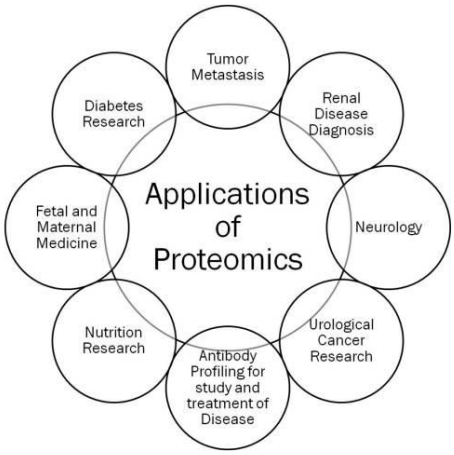
Image by Institute of Nanobiology and Structural Biology in Czech Republic

Recently at the University of North Carolina School of Medicine, researchers have developed technology that dissects the properties of single stem cells following isolation and growth of thousands of elusive intestinal stem cells at one time. This high throughput technological advance could give scientists the ability to study stem cell biology gastrointestinal disorders

Single cell isolation technologies from tissues and cells might be added as a key access technology for single cell analysis. Laser microdissection, like the “LMD6500” and LMD7000 from Leica or different technologies like the “CellSelector from ALS Automated Lab Solutions are state of the art.. Due to the high cost of equipment and requirement for well-trained personnel, it would make sense to have this technology available within an infrastructure. In the near future, we predict that solutions combining different technologies in one set-up, like AFM (atomic force microscopy), microinjection and single cell analysis will become increasingly important. Such systems allow population-based analysis of cellular events, e.g. RNA interference or cDNA over-expression and thus might form a key data producer for the modelling-experimentation feed-back loop. (Erfle, Appendix III)

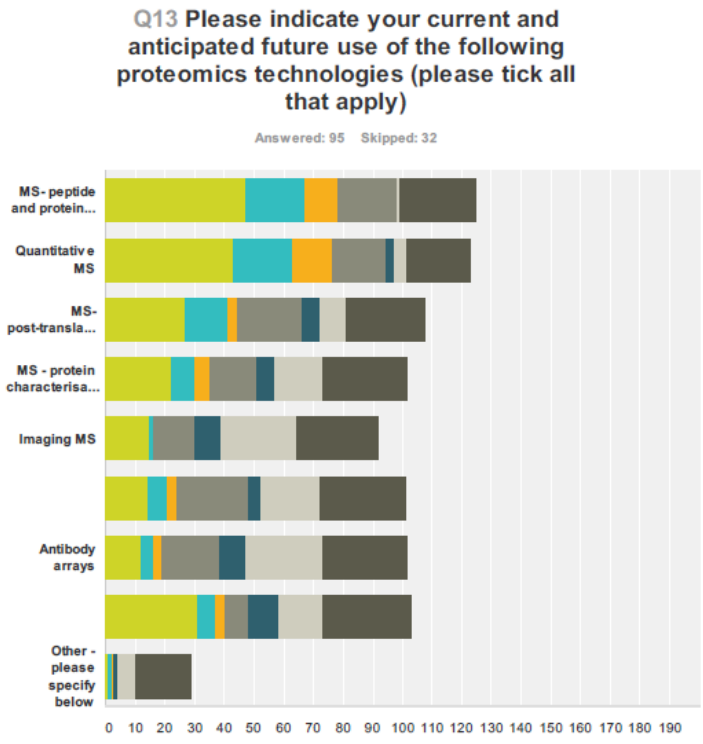
D. PROTEOMICS TECHNOLOGIES

This part of the report describes and evaluates the proteomic technologies that will play an important role in drug discovery, molecular diagnostics and practice of medicine in the post-genomic era - the first decade of the 21st century.



The most commonly used technologies currently are 2D gel electrophoresis for protein separation and analysis of proteins by mass spectrometry. Micro analytical protein characterization with multidimensional liquid chromatography/mass spectrometry improves the throughput and reliability of peptide mapping. Matrix-Assisted Laser Desorption Mass Spectrometry (MALDI-MS) has become a widely used method for determination of biomolecules including peptides and proteins. Functional proteomics technologies include yeast two-hybrid system for studying protein- protein interactions. (K. K. Jain 2015)

Giving the fact that Proteomics is the large-scale study of proteins, particularly their structures and functions, the two flavors of proteomic technologies are of critical importance. The first is discovery proteomics, also referred to as shotgun proteomics. This technology is used to identify the components of a biological system. The second proteomic approach, exemplified by targeting proteomics methods, aims at quantifying sets of proteins with high consistency across multiple samples. In systems biology such sample-sets are exemplified by differentially perturbed cells or tissues. Targeting methods include those based on affinity reagents (e.g. reverse arrays) and the mass spectrometric techniques selected/multiple reaction monitoring (S/MRM - *selected/multiple reaction monitoring*) and SWATH-M-SWATH MS - *a data independent acquisition (DIA) method.* (Aebersold, Appendix IV)



Mass spectrometry (MS) based techniques for protein profiling 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,

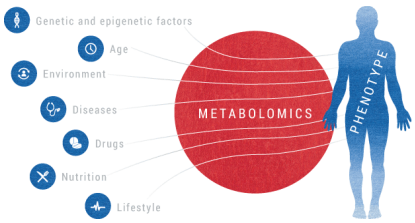
when the orbitrap technique became available, mass spectrometry papers have tripled. MS-peptide and protein identification, Quantitative MS, MS posttranslational modifications, Protein and peptide arrays, Antibody arrays and 2-Dimensional electrophoresis for **proteomics** are available in a majority of responding institutions and fully cover their demand. However, there are two not so widely available techniques: **Protein and petide arrays** – are available only in 16% of institutions but more than 28% require their usage in the future, and **Antibody arrays –** which are available only in 14% of institutions but more than 22% require their usage in the future. These 2 techniques should be monitored in the near future and might be worth consideration for inclusion into the infrastructure, should their availability become lower than demand.

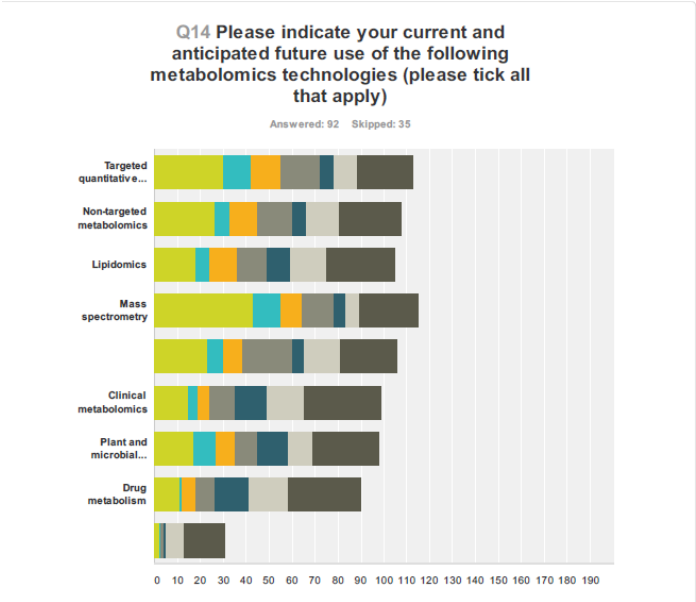
A point that might be addressed in future discussions is separating proteomic technology into 'instrumentation' and 'expertise', since many proteomic methods can be performed on the same MS instrumentation within the infrastructure at an institution, but the data generated often requires both skilled experimentalists and the necessary software to get the most out of a data set. (Hitchin, Appendix IV)

E. METABOLOMICS TECHNOLOGIES

Metabolomics is the global analysis of all or a large number of cellular metabolites. Like other functional genomics research, metabolomics generates large amounts of data.

Metabolomics Technologies use rather complex detection methods that require analytical and extensive data processing. There are two state-of-the-art or key technological approaches in metabolomics. The first employs NMR (Nuclear Magnetic Resonance) and the latter mass spectrometry (MS). MS has become mainstream, because of several advantages over NMR. Here the trend is towards high-resolution, accurate mass technology (1-2ppm mass accuracy). Targeted metabolomics focuses on known compounds, while untargeted aims to analyze all mass features from one sample. (Schauer, Appendix V)





than

14%

13%

Targeted quantitative metabolomics, and Mass spectrometry are **metabolomics technologies currently** available in more 34% of institutions. We can identify Mass spectrometry (52%) as the mainstream of available technique at institutions, but still only (Mass spectrometry) and (Targeted quantitative

metabolomics) of survey respondents use these techniques at their own institution. The reasons for this need to be understood and might raise the question of whether or not the infrastructure should interface these techniques in the future. Other techniques (Non-targeted metabolomics, Plant and microbial metabolomics, high throughput metabolomics, Clinical metabolomics) are in balance between their availability in local institutions (33%) and their likely level of requirement in the future. **High throughput metabolomics** , in contrast, may be a suitable target area for infrastructural assistance since 25% of correspondents are likely to require this technology in the future and 10% already use this technology at another institution.

Future directions for protein or protein metabolite interaction mapping might be monitoring of the 'cooperative mechanism'. This would focus on studying the overall interaction networks rather than charting interactions one by one. Upcoming high-throughput techniques might include lab scale surface plasmon resonance (SPR), or electrophoresis systems that allow the measurement of association/disassociation constants on very small samples. In this area of study i.e. characterization of protein complexes, going back to bench-based "old fashioned" chemical methods like gel filtration might become important again; however these will require miniaturization and implementation on a micro platform (using microfluidics) to be used in large-scale screenings.

Based on the current requests and the complexity from the analytical and raw data processing side it may be a low priority to look at facilitation mechanisms for metabolomics within ISBE infrastructure. As seen earlier, for gene expression profiling, metabolomics may rather be serviced in from specialized companies providing higher throughput, improved quality and shorter turn-around times; although this is a decision to be taking based on demands, investment, time and level of routine laboratory desired. (Schauer, Appendix V)

F. IMAGE TECHNOLOGIES

There are valid points for integrating imaging technologies and tools with more conventional approaches to analyze the biological circuits of microorganisms, plants, and animals. Light microscopy methods seem most suited for the systems biology field for a quick validation of proposed models; particularly due to achievable throughput levels and ease in experimental planning and running. (Erfle, Appendix VI)

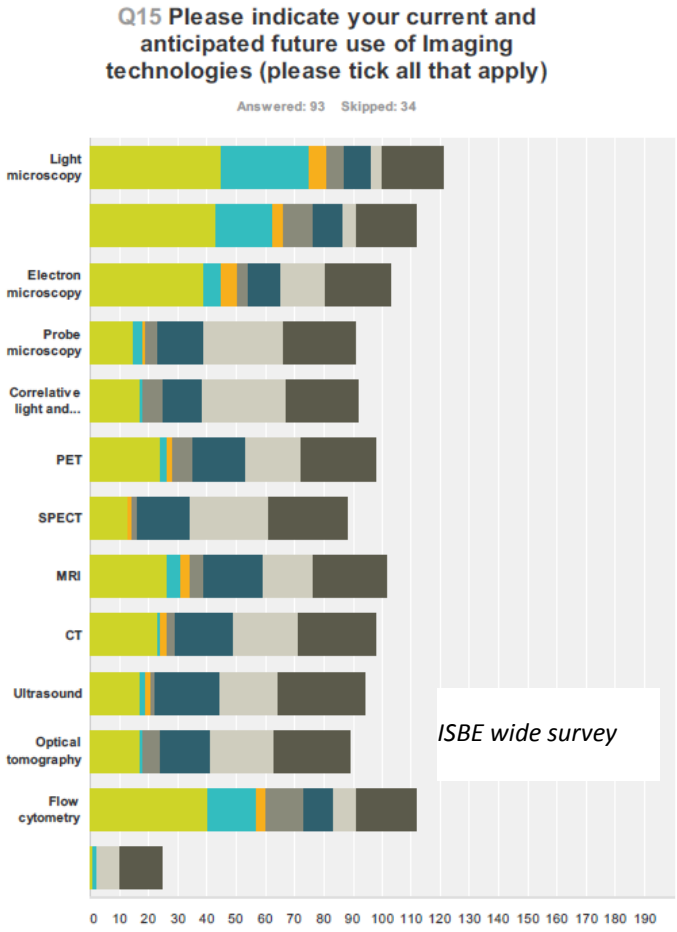
Image by Institute of Nanobiology and Structural Biology in Czech Republic

Imaging stands out amongst the many technologies used in systems biology as being (almost) the only one compatible with *in vivo* study, rather than a *post-mortem* analysis of biological systems (e.g. like that generally required for sequencing, proteomics, metabolomics, most other –omics), it allows the investigation of information flow in its biological context and change in space and time." (Spiteler, Appendix VI)

In imaging the state-of-the-art technologies are the following:

High-resolution:

- **Localisation microscopy (PALM, STORM):** most relevant for systems biology, allowing single-molecule measurements
- **Structured illumination (SIM):** 100nm resolution limit in 3D, but well suited to bridge the gap between single molecules and whole cells (50µm) at reasonable speed (seconds)
- **Laser overlay (STED):** high resolution (50nm) at relatively high speed (milliseconds- seconds), but at the cost of destructive laser power (mostly incompatible with life)
- **High-speed detectors:** sub-millisecond time frames



The current ISBE wide survey indicates that most technologies (Light microscopy – 49%, Advanced Light microscopy – 49%, Electron microscopy – 44%) 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 unmet demand in the future. In microscopy there are a few techniques that might have taken the revolutionary step (such 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.

Although the feeling is that it would make sense to have these available within an infrastructure, there is a warning sign that all imaging technology developments (listed in Appendix VI) absolutely require a well-managed infrastructure.They only work if all the different technologies are integrated, (moreover, together with the appropriate storage and compute), which is beyond most researcher's means (currently some leading Centres offer integrated facilities e.g. EMBL in Heidelberg, Max Planck-Institute of Molecular Cell Biology and Genetics in Dresden." (Spiteler, Appendix VI)

G. DYNAMIC MODELLING

The key current technologies in the area of Dynamic Modelling are mathematical modelling and analysis software packages (SW) for simulation (such as Matlab, Mathematica, Maple, Xppaut) and diverse systems biology add-ons for these such as the SBPOP PACKAGE (formerly called Systems Biology toolbox). These packages and add-ons can be categorised into specific areas as follows:

- 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 tools including standard data description language
- tools for integrating proteomics and genomics data (and indeed other –omics data) from existing databases (DBs) and new experiments

From an experimental perspective, advances in specific areas of experimentation are likely to be important for producing datasets informative for modelling (and model verification). These include 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. Additionally, rule based or related methods will be needed for large scale modelling, use of multiple data sources, including quantitative techniques like proteomics and phosphor-proteomics, reverse protein arrays, plus genomics data and single cell measurement techniques are all expected to take an important role. We will also need more efficient computational methods to extract the information from the ever increasing datasets and databases.

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

Image from the data department of computational biology / Institute of Nanobiology and Structural Biology of GCRC ASCR,v.v.i.

With the increasingly advanced understanding that diseases like cancer are a manifestation of deregulation of multiple pathways, and with availability of multiplex data on multi-pathways, large-scale mathematical dynamic models that account for pathways cross-talk rather than single pathways should become crucial in the future.

If integrated within the infrastructure, the data stewardship standards dealing with model sharing, storing and annotating will be particularly important in enabling transparency in the community and speeding up the modeling process.

To this point, the majority of sections above on Technology Watch have been concerned with the techniques and instrumentation required to produce new primary data. Whilst the ISBE infrastructure itself will no longer include specific data-generation centres, a technology watch in these areas is still critical for 2 reasons:

- Researchers will require access to appropriate data-generation technologies to support their research, and a 'brokering' service could facilitate identification of centres of expertise and instrumentation, together with a suitable access framework, since the diversity of experimental methods used in systems biology are unlikely to all be sufficiently well represented in every systems biology centre (or team).
- New technologies produce new data types and generally-speaking; tend to increase data volume and speed of acquisition, in addition to introducing new data formats (and potentially standards). The ISBE infrastructure will need to be sufficiently flexible and well-informed, to be able to adapt to effectively exploit these new data at every level

III. ADDED VALUE OF A STEWARDSHIP INFRASTRUCTURE

Between the respondents and the views represented in various literature reports is a remarkable agreement in the opinion that large projects are heavily funded for data collection and underfunded for data analysis. **A useful infrastructure for systems biology thus should not provide `huge machines` or data generation facilities but rather complex expertise and stewardship with a strong emphasis on informatics necessary for sharing and analysing data and modelling.** In the context of genomics databasing and the ability to share data seems to be a real issue, while instrumentation is relatively available.

Enabling access to existing data is seen as a significant challenge in this field. . An infrastructure for stewardship can facilitate multi-team projects to all generate re-useable data suitable for their own fields of research from the same experiments **Data sharing is generally seen as a critical limitation.**

Better integration of large scale-data and databases into the modelling process is clearly needed. Most current databases lack kinetic information, including rate parameters. The standards for improvement are importantly, release of appropriately annotated raw data and models in standard formats upon publication, and implementation of easy to use tools for automatic data and model import/export. Once this is achieved, different modality data can be brought together to address complex biological systems. For example, 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 large-scale physical storage but in terms of people (staff) who offer their expertise in limited time projects in the modelling centres** (several months) and who will be working on analysing and data. This type of service provision requires specific types of people and a risk could be a tension between giving community service versus their career progression, which could become an issue that needs to be addressed in the implementation of the infrastructure. These people (employees of ISBE) nevertheless cannot be exclusively service personnel but need to be embedded in the full scientific process, too, not to miss the development and new trends in new technologies usage.

Technology must be widely accessible, and the ISBE nodes should be able to negotiate or mediate access to experimental facilities provides by either other RIs or participating institutions. 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 a member of these institutions however, it is really difficult to get access to their data and these paid surveys are also quite expensive. **The availability of standardized data that is readily available to the SB community is a step towards a resource to test the robustness of models in a variety of experimental conditions.**

IV. REVIEW FROM TECHNOLOGY AND SCIENCE WATCH COMMITTEE

Updates from Science and Technology Watch Committee

David J. Duffy, Postdoctoral Researcher with Systems Biology, Ireland

Elaine Kenny, Co-founder of SME -Elda Biotech, Ireland

David Dobnik, National institute of biology Ljubljana, Slovenia

Paul Hitchin, Department of Life Sciences Faculty of Natural Sciences, , Imperial College

Ruedi Aebersold, ETH Zurich, Switzerland

Nicolas Schauer, CEO of Metabolomic Discoveries, Potsdam-Golm, Germany (updated 2015)

Holger Erfle, Head of the BioQuant RNAi screening Facility, Germany

Martin Spitaler, previously Facility for Imaging by Light Microscopy [FILM], Imperial College London, currently Light Microscopy Facility, Max Planck Institute for Biochemistry, Munich, Germany.

Lan K. Nquyen, Systems Biology at Conway Institute, Ireland

V. APPENDIX I. - MICROARRAY TECHNOLOGIES

Section 1: By David Dobnik, National institute of biology Ljubljana, Slovenia

High-throughput transcriptomic technologies (like **microarrays** or RNA-seq) usually require additional confirmation of results. This is mostly done by real-time PCR only on few of the selected candidate genes. Recently, with the development of digital PCR (dPCR) this kind of confirmations could gain on throughput. Specifically I have in mind the Biomark HD System from Fluidigm, which enables performing the real-time PCR reaction for 96 samples, where each sample is tested for 96 assays. Other dPCR platforms available are of course the state-of-art technology when speaking of absolute quantification, however the throughput in terms of samples and assays is not as high as with Biomark HD. It could happen that in near future also other platforms for dPCR would become handier for handling the systems biology experimental setup. The availability of this machine within an infrastructure would make sense, if such analyses would prove to be needed.

APPENDIX II. NEXT GENERATION SEQUENCING TECHNOLOGIES

Section 1: By David J. Duffy, Postdoctoral Researcher with Systems Biology Ireland

1. What existing technologies are considered as state-of-the-art or key technology in the specific fields of systems biology?

Illumina dominates the next generation sequencing (NGS) market with its suite of systems (MiSeq, HiSeq 2500/ X Ten and Nextseq 500), in 2012 Illumina had a market share of 56%. Life Technologies (Ion Torrent), Roche (454) and Pacific Biosciences (PACBIO RS II) all also provide state-of-the-art systems. As each company provides systems with varying strengths and weakness it would be important for ISBE to provide access to a number of different systems, as the best system depends on the required application. For example, each of the following applications mRNA-seq, exome sequencing, targeted sequencing or novel genome assembly are each best suited to different systems. In addition, the most cost effective system depends on the number of samples being run (e.g. Illumina HiSeq for large numbers and Ion PMG for small sample numbers) and the read depth required.

2. Would it make sense to have these available within an infrastructure?

While commercial companies and some academic institutes do currently supply access to these systems it would make sense to make them available in the infrastructure, especially if they are coupled to the supply of state-of-the-art bioinformatic and data handling support. The purchase of commercial bioinformatics and data storage/handling are prohibitively expensive and often the quality is quite limited. Given the nature of Systems Biology projects it is useful to be able to have on-going discussions and collaborations with bioinformaticians, as opposed to the purchase of a one-time only, locked analysis.

In addition, access should be provided to clinical diagnostic grade sequencers such as the MiSeqDx (FDA approved) and the Ion PGM Dx. Access to such equipment will be key to the application of Systems Medicine (a key emerging branch of Systems Biology) approaches to the clinical setting. Without diagnostic grade instruments sequencing results can be used for research only, rather than being directly applicable to individual patient diagnostics and facilitating the advent of precision medicine.

3. What emerging technologies will be important in the near future?

The current NGS technologies continue to be upgraded and improved with incremental advances being made. Some of these advances require no further investment, such as the release of improved software and sequencing flowcells/chips. However, incrementally improved systems are also released (primarily by Illumina). Therefore, funds should be budgeted to allow the periodical updating of ISBE equipment, as opposed to only investing in current systems.

Also funds should be benchmarked for the next NGS systems which will become available in the short to medium term. Nanopore sequencing (Oxford Nanopore Technologies) currently appears to be the closest technology to market, of a new generation of sequencing technologies. Nanopore machines have been accessible since spring 2014 to a limited number of applicants through an access program (https://nanoporetech.com/community/the-minion-access-programme), with their use already leading to a number of peer-reviewed publications. Nanopore sequencing, once refined, will offer a number of revolutionary improvements over current NGS systems. For instance, read lengths of up to tens of kilo bases and the ability to sequence RNA directly (no cDNA conversion or PCR enrichment).

4. How can these new technologies be integrated into Systems biology and how an infrastructure might help with this?

NGS technologies are currently integrated into Systems Biology, but this currently happens primarily at a more haphazard local level. An improved infrastructure could make this integration over an EU wide level, providing the required solutions and saving time, effort and money all of which are currently duplicated by every Systems Biology lab who conduct NGS experiments. Any ISBE initiative in this area should also provide open access to standardized reported NGS results using an intuitive interface, to enable the maximum use of the generated data by having it interrogatable by any researcher. The integration of infrastructure to facilitate cost effective sequencing, data management and bioinformatics analyses (both at the initial primary research level and later stage meta-analyses) would be hugely beneficial to future Systems Biology research in Europe. In addition, given the rate of continuous advance of NGS technologies, the centralization of these technologies into larger infrastructure centres would facilitate the purchase and more rapid adoption of the latest equipment by European researchers. Smaller less centralized labs cannot maintain pace with the continually advancing iterations of these technologies.

Section 2: By Elaine Kenny, Co-founder of SME called Elda Biotech, Ireland

Whilst all of the institutions have access to next generation sequencing (NGS) technologies it's unclear where specific expertise reside. Identifying the key expertise of institutes will also identify partner/collaborator groups for relevant projects and vastly reduce the cost associated with poor data generation. Unlike microarray technology for example the sample preparation protocol employed can have a huge impact on the quality and type of data generated. Whilst there's no need to centralize this technology, possibly a central repository for how samples are prepared (library prep protocol/kits/adjustments etc) for each of the studies would help in the interpretation of the results. This is certainly something that could be provided quite easily by the infrastructure, where basic but often very important information about each experiment is stored.

The report mentioned that the study of DNA methylation was identified as likely required in future by 23% of respondents to the questionnaire. Currently methylation studies using NGS can be quite expensive to run, however as the cost comes down this will certainly start to change. It's also worth looking at some of the third generation sequencing technologies e.g. Pacific Bioscience systems for this type of sample/study. The methylation status of DNA can be sequenced straight from the DNA with minimal preparation of the DNA sample. This technology is currently prohibitively expensive for individual labs however it may be worth collaboratively outsourcing samples or identifying a central point to which samples can be sent.

With the generation of vast amounts of genomic data using this technology data storage and analysis is always going to be an issue. Some of the newer NGS machines come equipped capable of uploading data to the cloud. I see cloud computing becoming quite vital in NGS projects and it has certainly increased our collaborative ability. The use and reliance on it will continue, especially in the research environment. Many users like to have ownership of their data and their analysis; however it's worth looking at the possibility of having a central repository/pipeline for data analysis utilizing cloud computing. The key to making such a thing work however would be the turnaround time of analysis. The idea being that all NGS data could be uploaded and QC passed/checked to ensure a minimum standard is met.

Section 3: By David Dobnik, National institute of biology Ljubljana, Slovenia

Another new sequencing technology, which was given a lot of attention, was nanopore sequencing (e.g. MinION from Oxford Nanopore Technologies), especially suitable for cases where long reads are needed. After initial problems with accuracy (sequence identity of 66% in June 2014), new chemistry and base calling algorithms releases improved this to a better level (to up to 85% identity in November 2014; reported in Jain et al. 2015. Nature Methods vol. 12 no. 4). However, all of the NGS techniques are using some kind of medium to translate the signal into the base (protons, electrical current, etc.). A completely new approach might come from the idea of the G. Schneider and his group (Leiden University), who are working on establishment of sequencing through the graphene nanopores, where each base of the DNA molecule coming through the pore would be read as such, without additional chemical reactions taking place. Nevertheless, in my opinion the bioinformatic tools will still play the greatest role in any kind of sequencing applications now and in the future.

APPENDIX III. SINGLE CELL TECHNOLOGIES

Section 1: By Holger Erfle, Head of the BioQuant RNAi screening Facility

Recent technological advances in single-cell analyses allow to study heterogenity, signaling, and stochastic gene expression.

The report tries to address in each area the following:

1. What existing technologies are considered as state-of-the-art or key technology in the specific fields of systems biology?

Recent: convergence of genomics and single-cell biology

a) Sequencing genomic DNA from single cells – herein improvements on whole-genome amplification methods are important.
b) Single-cell RNA sequencing (RNA-seq) e.g. reverse transcription directly on cell lysates from individual cells using oligo-dT primers (Tang et al., 2009). cDNA library PCR amplified, fragmented, and subjected to sample preparation for deep sequencing
c) high-throughput or parallelized real-time PCR - high-throughput real-time PCR cannot be probes at one like in sequencing.
d) Single-cell mass cytometry (Bendall et al., 2011) allows parallel detection of a large number of proteins in single cells by using specific antibodies labeled with heavy metals.
e) Imaging-based techniques such as single-molecule FISH or GFP-based approaches can been used and depending on application coupled to Single cell removal technologies from tissues and cells might be added as a key access technology for single cell analysis. Laser

microdissection, like the “LMD6500” and LMD7000 from Leica or different technologies like the “CellSelector from ALS Automated Lab Solutions are state of the art technologies.

2. Would it make sense to have these available within an infrastructure?
   Yes, it would make sense to have these technologies available within an infrastructure as purchase of equipment is costly and running the site demands experienced and well-trained personal.

3. What emerging technologies will be important in the near future?
   Depending on application all 5 technologies mentioned a) –e) under point 1) will find their own niche.

4. How can these new technologies be integrated into Systems biology and how an infrastructure might help with this?
   Those technologies might answer questions, otherwise not resolvable in the feedback-loop of model and experimentation.

Section 2: By David Dobnik, National institute of biology Ljubljana, Slovenia

In regard to the single cell technologies, there has been a publication recently describing the RNA sequencing in situ (Highly Multiplexed Subcellular RNA Sequencing in Situ, Je Hyuk Lee et al., Science vol. 343, March 2014). The described technique provides a combination of in situ library preparation and sequencing. At time it is limited with amplicon length of 27 bp and the number of sequenced amplicons. The correlation coefficient, when compared to Illumina is still quite low (0.5 – 0.7), however the technique offers the advantage of subcellular localization, so you can see where was the RNA of each of the sequenced amplicons localized. I see this technique at the rise of expectation in fermentation phase and it is yet not fully tested (i.e. cannot be included to infrastructure). Recently, a protocol for this technique was also published (J.H. Lee et al. 2015, Nature Protocols 10, 442–458). Nevertheless this kind of techniques might be of great importance in the future.

VI. APPENDIX IV. PROTEOMICS TECHNOLOGIES

Section 1: By Paul Hitchin, Faculty of Natural Sciences, Department of Life Sciences, Imperial College

Having read through the report on continuous technology forecasting with respect to Proteomics technologies, I can largely agree with the findings from the survey. It seems that many of the proteomic technologies are in place for researchers at their institutions or are available to use at another institution but the survey has identified two techniques: Protein and peptide arrays and antibody arrays, that might need to be implemented in any infrastructure in the near future. A point that might like to be addressed in future discussions is separating the proteomic technology into 'instrumentation' and 'expertise', since many proteomic technologies can be performed on the MS instrumentation within the infrastructure at an institution, but the data generated often requires both skilled experimentalists and the necessary software to get the most out of a data set.

Section 2: By Ruedi Aebersold, ETH Zurich

1. What existing technologies are considered as state-of-the-art or key technology in the specific fields of systems biology?

For systems biology two flavors of proteomic technologies are of critical importance. The first is discovery proteomics, also referred to as shotgun proteomics. This technology is used to identify

the components of a biological system. The second proteomic approach, exemplified by targeting proteomics methods, aims at quantifying sets of proteins with high consistency across multiple samples. In systems biology such sample-sets are exemplified by differentially perturbed cells or tissues. Targeting methods include those based on affinity reagents (e.g. reverse arrays) and the mass spectrometric techniques selected/multiple reaction monitoring (S/MRM) and SWATH-MS. Targeting methods generally require the development and validation of specific assays for each targeted protein (e.g. an antibody for immunodetection; a reference fragment ion spectrum for MS based techniques) and the one time development for the community of proteome-wide assay libraries would be a particularly fruitful endeavor for ISBE.

2. Would it make sense to have these available within an infrastructure?

Supporting these techniques as infrastructure platforms would certainly generate a very high impact. This is particularly the case for the above described targeting techniques which would allow a large number of systems biologists to accurately quantify essentially any protein with a high degree of reproducibility across multiple samples, e.g. sample sets representing differentially perturbed cells or tissues. Considering that data driven systems biology studies to date are for the most part based on transcript measurements and the well-known fact that transcripts do neither predict the quantity nor the activity state of proteins, quantitatively accurate protein data would greatly advance the field of systems biology.

If the technology is to be supported by an infrastructure/facility, it will be important to make an integrated technology platform available.

VII. APPENDIX V METABOLOMICS TECHNOLOGIES

Section1: By Nicolas Schauer, CEO of Metabolomic Discoveries, Potsdam-Golm, Germany

1. What existing technologies are considered as state-of-the-art or key technology in the specific fields of systems biology?

Two technological approaches in metabolomics exist. The first employs NMR and the latter mass spectrometry (MS). MS has become mainstream, because of several advantages over NMR. Here the trend is towards high-resolution, accurate mass technology (1-2ppm mass accuracy). Targeted metabolomics focuses on known compounds, while untargeted is analyzing all mass features from one sample.

2. Would it make sense to have these available within an infrastructure?

Metabolomics will become of more importance over the next years and an integral part of systems biology. Based on the current requests and the complexity from the analytical and raw data processing side it may not be beneficial to implement metabolomics. As seen in gene expression profiling, metabolomics may rather be serviced in from expert companies providing higher throughput, improved quality and shorter turn-back times. Though this is a decision to be taking based on demands, investment, time and level of routine laboratory desired.

3. What emerging technologies will be important in the near future?
In the future the coupling of MS to ion mobility is most promising, though this is still in its infancy.

4. How can these new technologies be integrated into Systems biology and how an infrastructure might help with this?

Metabolome data provides rich information on top level and thus gives the most insights into biological mechanisms. Data can be easily integrated into systems biology approaches, as KEGG and other identifiers allow pathway and network building and thus provides a close link between in-silico and experimental data.

VIII. APPENDIX VI IMAGE TECHNOLOGIES

Section1: By Holger Erfle, Head of the BioQuant RNAi screening Facility

The report tries to address in each area the following:

1. What existing technologies are considered as state-of-the-art or key technology in the specific fields of systems biology?

   Quantitative fluorescence-microscopy allows investigating functional molecules in living cells with ever-growing spatial and temporal resolution. Easy to apply assays have recently shown to allow genome-wide functional analyses by high-throughput-microscopy. In addition, the existence of processes and standardized reagents that interfere with cellular functions like RNAi, GFP-tagging and transgene approaches allowing direct characterization of the quantity, localization, dynamics and interaction of proteins in intact and even living cells, strengthen the influence of imaging-based proteomics in systems biology.

   High-throughput and high-content microscopy can provide data for systems-biology by a high level of automation, available tools for quantification and precise integration of multiple steps in an integrated workflow.

   Developments in super-resolution microscopy are pushing the limits of resolution of Light microscopy. In addition, Light microscopy analysis can be automated for high-throughput analysis in screening protein knockdown in a morphomics approach (John M. Lucocq et al, Trends in Cell Biology, 2015).
   In addition combining high-throughput with super-resolution microscopy is latest state of the art in high-content microscopy (Gunkel et al 2014, Histochem Cell Biol and Flottmann et al 2013, Biotechniques).

2. Would it make sense to have these available within an infrastructure?

   Yes, it would make sense to have this technology available within an infrastructure as purchase and maintenance of equipment are costly and carrying out experiments and teaching users demand experienced and well-trained personal.

3. What emerging technologies will be important in the near future?

   Here one might add super resolution techniques, like STED, PALM or STORM and light-sheet imaging for imaging in novel cell culture systems, like 3D culture. In addition correlating high-throughput with high-resolution methods are of highest value.

4. How can these new technologies be integrated into Systems biology and how an infrastructure might help with this?

   Those new technologies allow a quick and easy link between modeling and experimental validation. Light microscopy has several advantages over other microscopy techniques, ranging from in vivo analysis to wide sampling fields. Due to throughput and ease in experiment planning and running, light microscopy methods are most suited for the

systems biology field for a quick validation of proposed models. In addition, scientists can be relatively easy taught to perform themselves individual experiments.

Section 2: By Martin Spitaler, Facility for Imaging by Light Microscopy [FILM], Imperial College London

Imaging stands out amongst the many technologies used in systems biology as being (almost) the only one compatible with live: Rather than a post-mortem analysis of biological systems (e.g. like in sequencing, proteomics, metabolomics, most other –omics), it allows to investigate information flow in its biological context and change in space and time. Two major limitations have limited its use for systems biology:

- (1) limited resolution in space and time
- (2) difficulties extracting unambiguous information from unstructured data

Both these limits are currently being overcome at dramatic speed (see below), although many logs of improvements will still be required to reach a 'saturation' level, at which no more improvements could be expected (maybe ~µsec for speed, nm for dimensions, especially the combination is still utopia). Which leaves as a new limit the handling and processing of data, both from the logistic point of view (annotation, transfer, visualisation) and hardware / software capacities.

1. What existing technologies are considered as state-of-the-art or key technology in the specific fields of systems biology?

- High-resolution:

○ **Localisation microscopy (PALM, STORM):** most relevant for systems biology, allowing single-molecule measurements

○ Structured illumination (SIM): 100nm resolution limit in 3D, but well suited to bridge the gap between single molecules and whole cells (50µm) at reasonable speed (sec)

○ **Laser overlay (STED):** high resolution (50nm) at relatively high speed (msec-sec), but at the cost of destructive laser power (mostly incompatible with life)

○ **High-speed detectors:** sub-msec time frames

- Data handling:

○ **Data storage and annotation:** first functional solutions (e.g. OMERO), but still rather limited solution (handling of large data, usability, integration with software tools); content-based search in early experimental stage

○ **Data analysis:** Pattern recognition becoming established in light microscopy; single-molecule localisation and statistics at the stage of ongoing community-based reviewing and standardisation

○ **On-the-fly processing:** slowly moving from developmental stage to early adopters, with emerging support from commercial microscope manufacturers

1. Would it make sense to have these available within an infrastructure?

All developments above absolutely decide on well-managed infrastructure: They only work if all developments are integrated, which is beyond most researcher's means (currently lead e.g. by EMBL in Heidelberg, Max Planck-Institute of Molecular Cell Biology and Genetics in Dresden).

2. What emerging technologies will be important in the near future?

A) Overcoming resolution limits in space and time

Two directions will change the way systems biology works: Higher resolution and detector sensitivity allowing single-molecule observations at high speed, and whole-organism (zebrafish, embryos) / whole-organ imaging of physiological process.

Higher resolution in space close to single molecule-scale is now possible thanks to super-resolution microscopy, made possible by a parallel development of novel microscopy techniques, detectors (CCD and CMOS) with single-photon sensitivity and novel fluorescent markers (photoswitchable proteins and chemical fluorophores). These techniques (under the acronyms of PALM and STORM) already allow studying the cellular signalling circuitry at molecular level, or physiological events down to sub-millisecond speed. However, there is still a strong trade-off between spatial and temporal resolution, ranging from 30nm precision of whole-cell cross-sections (10-30µm length) at 3-20min per frame to the other extreme of 50µsec per frame, but of small areas (1µm length) with 1µm precision. At the other end of the spectrum, novel microscopy techniques (2-photon intravital microscopy, light-sheet microscopy) in combination with new markers allow visualisation of mm-scale organs or organism at cellular precision, thereby quantifying interaction between cells and tissues, movements over time, decision trees in development (e.g. embryo development, haematopoiesis) etc.

B) Improvements in the extraction of information from unstructured data

The main value of microscopy images is it's high-content, multidimensional, unstructured information, but this makes it difficult to translate into computer-readable formats. However, many of these difficulties are based on a slow translation of established technologies from other areas to biological applications: While biologists use mobile phones with smile detection in private life, they still mostly rely on archaic intensity thresholding for object detection in their scientific work.

We are currently witnessing a massive push in this re-adaptation of technologies, be it astronomy algorithms for single-molecule localisation or pattern recognition to track cells in noisy 3D data of whole organs.

(1+2)=(3) intelligent image acquisition

A major development currently in an early experimental state (although applied in selected labs for over a decade) will be intelligent acquisition, i.e. rather than generating huge amounts of meaningless data, to incorporate the biological question in the image acquisition. This will drastically improve the data quality while in parallel drastically reducing the data volume, or rather the ratio data volume per scientific information (the total volume will keep moving on the limits of the technical possibilities). On-the-fly-analysis will work in two ways:

- rather than saving unstructured data, only the information of interest is saved; for example, if studying cell movements in development, the XYZT coordinates of a few thousand cells (Mbytes) would be saved, rather than GBytes of raw images per time point
- low-resolution screening in space and time, then switching to high-resolution mode when encountering an event / object of interest; this will also reduce the amount of data (or increase the number of observable events / objects) by many logs

1. How can these new technologies be integrated into systems biology and how an infrastructure might help with this?

The main need to integrate them into Systems Biology are:

- standardised annotation of unstructured data:
  - on the hardware side, the solution is on its way with the Open Microscopy Environment (OME) data standard, now supported by most commercial vendors and open-source tools
  - on the sample side, standardised protocols for sample preparation and annotation are still needed; only user education will be able to bridge the gap

- on the analysis side, standardisation of algorithms is on its way, especially in the super-resolution field, but it will take another few years to find a common sense

- uptake by systems biologists / mathematicians:
  - especially modellers tend to shy away from the unstructured, multi-dimensional nature of microscopy images; in combination with above efforts (improving the data quality), education will be needed to help them understand the huge potential (and some pitfalls) of these technologies

IX. APPENDIX VII DYNAMIC MODELLING

Section1: By Lan K. Nquyen, Systems Biology at Conway Institute, Ireland

The authors of the report have discussed many key areas that are important for an infrastructure with regards to Dynamic Modeling, both in terms of the modeling techniques and data integration required for the modeling process. I have a number of additional points on both of these aspects:

- With the increasingly advanced understanding that diseases like cancer are a manifestation of deregulation of multiple pathways, and with availability of multiplex data on multi-pathways, large-scale mathematical dynamic models that account for pathways crosstalks rather than single pathways should become crucial in the future. Hence, new computational frameworks that make it easy and time-efficient to build, integrate and maintain these large models are strongly required. These frameworks should allow integration of information like mutational landscape/epigenetics on top of existing network models so that models can be adapted for different cancers and/or patients, thereby pushing modeling towards personalized medicine. Standards dealing with model sharing, storing, annotating would be particularly important in enabling transparency in the community and speeding up the modeling process.
- A key related need is the development of focused databases on kinetic information and protein concentrations that are essential for model calibration and parameter estimation. Good annotation of these databases would be important for modelers to extract cellular-context specific information for model adaptation. Efficient parallel
parameter estimation methods capable of running on clusters (which could be Sharp among the Infrastructure institutions) should be available and integrated into modeling software for access by the community.
- As models are multi-dimensional, novel methods for efficient analysis and visualization of the model dynamics in multi-dimensional settings are crucial for better “global” understanding of the networks being modeled. This would provide a more truthful picture of the network dynamics and facilitate therapeutic strategies. Current analysis methods are limited in this aspect.
- Regarding data as input for modeling process, the report has included the key experimental technologies. In addition, technology that is capable of obtaining (multiplex) data directly on patients sample tissues such as tissue FRET imaging would be particularly useful in the future to adapt models from cell-based towards patient based. I expect that these techniques would be quite challenging to develop but would be of enormous applicability.

Cited Works

Jain, K. K. "Proteomics - Technologies, Markets and Companies." *Proteomics - Technologies, Markets and Companies*. Jain PharmaBiotech, 1 June 2015. Web. 21 June 2015.

Management, Technology Analysis & Strategic, and Nos. Vol. 18. "The Sociology of Expectations in Science and Technology." *Vol. 18, Nos. 3/4, 285 –298, July – September 2006 The Sociology of Expectations in Science and Technology* (n.d.): n. pag. *Utwente*. 01 Apr. 2015. Web. May-June 2006.
http://www.utwente.nl/bms/steps/people/scientific/konrad/editorial.pdf

Metzker, Michael L. "Sequencing Technologies — the next Generation." *Nature*. N.p., Nov. 2014. Web.

"Microarray Technology: An Introduction to DNA Microarray." *Microarray Technology: An Introduction to DNA Microarray*. PREMIER Biosoft, 02 Feb. 2015. Web. 21 June 2015.

Sandstrom, Chris. "Technology S-curves." *Technology S-curves*. Slideshare.net, 27 Jan. 2013. Web. 03 July 2015.

Mavis, B.M., Chao, M.C., and Waldour, M.K. "Entering the era of bacterial epigenomics with single molecule real time DNA sequencing" Current Opinion in Microbiology. Volume 16, Issue 2, April 2013, Pages 192–198

Herper M. "Flatley's Law: The Company Speeding A Genetic Revolution" Forbes September 8, 2014 http://www.forbes.com/sites/matthewherper/2014/08/20/flatleys-law-how-one-company-became-the-force-behind-medicines-genetic-revolution/ "September 8, 2014 issue of Forbes

Cochrane Cook C.E. and Birney, E. "The future of DNA sequence archiving – review" GigaScience 2012, 1:2 http://www.gigasciencejournal.com/content/1/1/2

Je Hyuk Lee *et al*. "Highly Multiplexed Subcellular RNA Sequencing in Situ", Science vol. 343, March 2014.