Gomez-Cabrero et al. BMC Systems Biology 2014, 8(Suppl 2):I1
http://www.biomedcentral.com/1752-0509/8/S2/I1

BMC Systems Biology

INTRODUCTION Open Access

Data integration in the era of omics: current and future challenges

David Gomez-Cabrero[1*], Imad Abugessaisa[1], Dieter Maier[2], Andrew Teschendorff[3], Matthias Merkenschlager[4], Andreas Gisel[5], Esteban Ballestar[6], Erik Bongcam-Rudloff[7], Ana Conesa[8], Jesper Tegnér[1]

From High-Throughput Omics and Data Integration Workshop
Barcelona, Spain. 13-15 February 2013

Abstract

To integrate heterogeneous and large omics data constitutes not only a conceptual challenge but a practical hurdle in the daily analysis of omics data. With the rise of novel omics technologies and through large-scale consortia projects, biological systems are being further investigated at an unprecedented scale generating heterogeneous and often large data sets. These data-sets encourage researchers to develop novel data integration methodologies. In this introduction we review the definition and characterize current efforts on data integration in the life sciences. We have used a web-survey to assess current research projects on data-integration to tap into the views, needs and challenges as currently perceived by parts of the research community.

Introduction

Data integration is now a very commonly used notion in life sciences research. As of 2006 there were 1,062 papers explicitly mentioning "*data integration*" in their abstract or title, whereas this number has more than doubled in 2013 (2,365). However, there is still no unified definition of data integration, nor taxonomy for data-integration methodologies despite some recent efforts on this topic [1-5]. In February 2013, the FP7 STATegra project (http://stategra.eu/) and the COST Action SeqAhead (http://seqahead.eu/), two EU-funded initiatives on the bioinformatics of high-throughput data, organized in the city of Barcelona the "Workshop of Omics and Data Integration", with the aim of reviewing current technologies on omics data production and the available methods for their integrative analysis. The workshop consisted of contributed talks, sessions for open discussion and we included an on-line survey to investigate the current opinions of the research community on this topic. Three major conclusions were extracted from the Barcelona workshop. First, there is a clear need for revisiting the concepts of data integration and stating available resources in this field; second, it was advantageous to extend our survey to a broader audience of scientists in life sciences, and third the commitment of organizers to publish the discussed topics, contributions and outcome of the public survey in a relevant journal is an important driver to spearhead further discussion in the community. In this supplement we discuss these three conclusions in some detail. In this introductory article we review current definitions of data integration and describes it formally as the combination of two challenges: data discovery and data exploitation [5]. We briefly list major public efforts in creating resources (datasets, methods and workshops) for data integration. We also present the results of the extended community survey, which took place between February and March 2013 and on the basis of the survey we extract a couple of conclusions which warrant further elaboration in the community. Finally we introduce the contributions of the papers collected in this supplement within the context of the discussed data integration topics and stated community needs.

Challenges of data integration in life sciences

Research in life sciences has the generic goal to identify the components that make up a living system (G1) and to understand the interactions among them that result in the (dys)functioning of the system (G2). Collection of

[1]Unit of Computational Medicine, Center for Molecular Medicine, Department of Medicine, Karolinska Institute and Karolinska University Hospital, Stockholm, Sweden
Full list of author information is available at the end of the article

BioMed Central

© 2014 Gomez-Cabrero et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

biological data is therefore a method to catalogue the elements of life, but the understanding of a system requires the integration of these data under mathematical and relational models that can describe mechanistically the relationships between their components. We can illustrate the state of affairs on data integration in life science research using a simple example taken from metabolic modeling. Let us consider the glycolysis pathway (GLY), which consists of the conversion of glucose into pyruvate to release energy (see Figure 1 and [7]). In the study of GLY, G1 is considered "*to be known*" as there are a detailed set of genes, proteins and metabolites already described; however we are not yet certain that this list contains all involved elements, for example the list does not incorporate the epigenetic marks that may be associated to the regulation of GLY. When we consider G2, Figure 1 again depicts the current knowledge of the system and may erroneously imply that the system - defined as a set of interactions - is fully known. However, pathway elements and relations may be missing (see for instance the recent work on synthetic non-oxidative glycolysis [7]) and this representation does not allow us to determine completeness. Once more, the figure does not depict all the regulatory mechanisms involved or the rates of the reactions. This brings us to the first question of: *"What are the available data that can be used to fully characterize the GLY metabolic pathway*?"

The present situation is very fortunate since over the last decades several different types of data were generated and huge efforts were dedicated to create database repositories for different data-types where investigators were encouraged to deposit and share datasets associated with scientific publications. The benefits of this are twofold: on the one hand it enables or support researchers in reproducing and validating the analysis of other labs, and on the other hand it allows researchers to analyze data in novel ways and/or with different methodologies that were not originally considered by the team who generated the data. To illustrate this we investigated the availability of GLY-related datasets in Gene Expression Omnibus (GEO [8,9]) as an example of a major gene expression data repository and we readily made two observations. There exist a small number of datasets pertaining to the direct investigation of the GLY pathway, but the majority of microarray and NGS datasets contain information about the GLY pathway at the mRNA level. Moreover, it is possible to complement such information with enzyme kinetics data from databases such as BRENDA [10]. These observations bring us to the next questions. Once relevant data sources have been identified, "*How do we integrate all (or part of) the available datasets in order to improve our definition of the GLY system*?" and "*How do we re-use all this information when designing new and novel experiments*?"

All the above questions and challenges intuitively define the notion of "*data integration*".

Data integration challenges

The term *data integration* refers to the situation where, for a given system, multiple sources (and possible types) of data are available and we want to study them integratively to improve knowledge discovery. In the GLY example

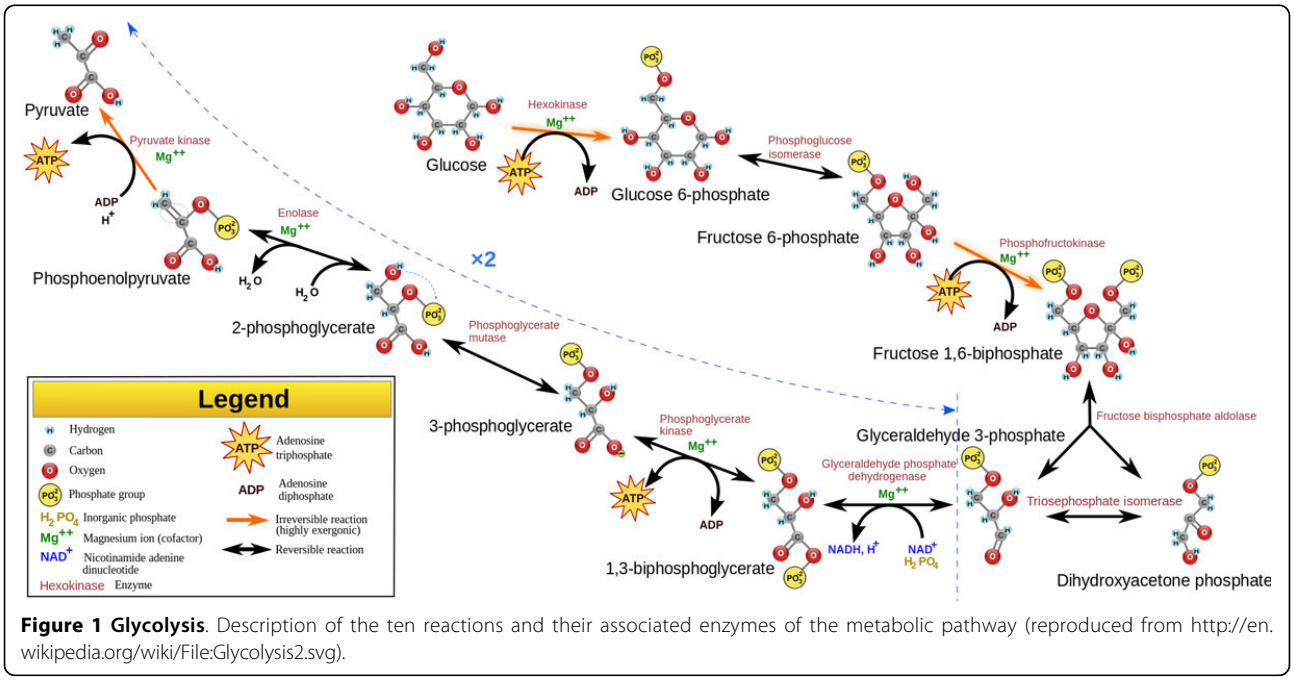

Figure 1 Glycolysis. Description of the ten reactions and their associated enzymes of the metabolic pathway (reproduced from http://en.wikipedia.org/wiki/File:Glycolysis2.svg).

system we could have two datasets describing the system, one containing information about gene expression at the mRNA level and the other describing the CpG DNA methylation profile. In several studies [11,12] where gene expression and DNA methylation data were available, the genome-wide relationships between DNA methylation and gene expression have been investigated in order to infer *generic* rules to questions such as: "Does DNA methylation regulation occurs at CpG islands and/or shores?", or "How does DNA methylation in promoters/gene-bodies/enhancers regulate gene expression?" [13]. These kinds of analyses have advanced our understanding of gene regulation by providing "generic rules yet with several exceptions" that associate epigenetic modifications with transcription [11,12]. For instance, as a general rule CpG methylation in promoters in mammals was found to be anti-correlated with gene expression, while CpG methylation in gene bodies in mammals was positively correlated; yet these generic rules are observed *as a trend*, but are not necessarily true for all genes and/or for all biological situations.

To understand the challenges of data integration it is first required to define the term. The term "*data integration*" first appeared from the need to access different databases with overlapping content to provide "*a redundancy free representation of information from a collection of data sources with overlapping content*" [14] which describes a need that appeared when the first databases were designed [15] and it was required to connect several of them: "*integration of multiple information systems aims at combining selected systems so that they form a unified new whole and give users the illusion of interacting with one single information system*". The aims of database integration were to make data more comprehensively available, and to increase the value of existing data by allowing previously difficult queries to be made upon it. Data mining (as a step in Knowledge Discovery in Databases [16,17]) is a major beneficiary from database integration. However this definition considers only *access* to data, and not *exploitation* of data, hence this definition of data integration is not fully applicable to life sciences research.

We define *data integration* as the use of multiple sources of information (or data) to provide a better understanding of a system/situation/association/etc; hence data integration, as defined here, is an action performed on a daily basis by most individuals, and a critical element in research.

Data integration in the life sciences becomes a more complex challenge considering the current "*data explosion*". This "*added*" challenge has been already been duly recognized; for instance in 2010 the National Research Council of the National Academies in US organized a workshop to "*explore alternative visions for achieving large-scale data integration in fields of importance to the federal government*" [5]. The workshop's aims and main results were reported in [5]; at the beginning of the document two main challenges associated with data integration were defined: *data discovery* and *data exploitation*. We followed the same structure in the present review and the next sub-sections briefly detail these challenges in the life sciences.

Data discovery

Data source discovery is defined as the identification of relevant data sources. Discovery of publicly available biological data sources is easy ("*just google it*" albeit with some exceptions, e.g. neuroscience [18]), whereas discovering the "*appropriate data*" is a more complicated task. One problem is the diversity of existing data types and formats, each one compliant to a different standard, which results in data heterogeneity and what has been called a "*loose federation of bio-nations*" [2]. The publication of specialized web databases has flourished in the last decade due to the relative ease of creation and maintenance and the reputation that it brings to the developers [2]. While specialized platforms may answer specific needs of the research community they may also introduce biases that affect data analysis. Two examples of this problem are the pathways and miRNA databases. The early 2000s witnessed the beginning of the generation of many pathway databases and their number has been increasing ever since, but has stabilized in the 2010s [2,4]. By 2013, *Pathguide* [19] reported a list of 547 biological pathways and molecular interaction related resources. These resources are not simply complementary, but often define similar signaling and metabolic pathways with different boundaries and components. This different specification is not irrelevant as many genome analysis methods are based on pathways and are therefore affected by how these are defined (see for instance [20] in this supplement). A second example relates to the storage of miRNA information [21]; this field has observed the development of generic purpose databases (e.g. miRBase and miRNAmap), many specialized databases (e.g. miRWalk, mirDB and Tarbase among others), and even standards for miRNA annotation [22]. In order to cope with this heterogeneity additional resources were developed such as catalogs of all available resources (e.g. *Pathguide* in pathways) and novel and larger databases developed in a joint effort between the developers of many older miRNA databases (see for instance RNAcentral [22]). We foresee two possible future scenarios: in the first one *developers of novel data-type resources, learning from previous experiences, will join efforts to consolidate data and create standards at earlier stages; in the second one we will accept redundant overlaps and solve them with data integration and knowledge management approaches.*

The rise of database resources certainly helps but does not solve entirely the problem of easy access to relevant

data. An example is the Gene Expression Ombinus (GEO [8,9]); GEO (similarly to ArrayExpress [23]) is a data repository for microarray and NGS data that requires the data producers to submit data following the Minimum Information About a Microarray Experiment (MIAME) guidelines [24]. MIAME was originally designed to provide standards for microarray data sharing to ensure that "*data can be easily interpreted and that results from its analysis can be independently verified*" [24]; GEO requires that both raw and normalized data be available, samples are annotated (including experimental design) and laboratory and data processing protocols are described. Enforcing MIAME allowed many researchers to reexplore datasets from novel perspectives and "*more and more research is now built on the analysis of data that were not collected by the researchers themselves, and many of the extant data have not been utilized to their full potential*" [5]. However, annotation of experimental data in GEO still makes little use of controlled vocabularies (e.g ontologies), which is necessary for automated retrieval of relevant datasets for specific large-scale studies. Therefore finding datasets for a specific condition is possible, but using all samples associated to that condition without previous manual curation is still unfeasible.

We consider that the integration of Laboratory Information Management systems (LIMS) and/or Experiment Management Systems (see in this supplement [25]) in lab-life operations of omics data, and its standardization (such as the use of ontology-derived nomenclatures) and use in submission to public data repositories will smooth the path towards efficient data discovery and sharing.

Data exploitation

Data exploitation refers to the effective use of collective information to obtain new insights [5]. We can classify data exploitation according to the type of data used (similar or heterogeneous data types) or the information considered (all data points from all studies or summary results of individual studies, i.e. meta-analysis [26,27]). However, no classification will fully characterize contemporary research as researchers are blurring the boundaries by developing hybrid methodologies to optimize data analysis outcomes. We next develop some examples in current research.

If we consider datasets of similar data types, *meta-analysis* (that is, combining summary information from independent studies [26,27]) is a widely used statistical tool, as in many recent GWAS studies [28]. Importantly, we consider meta-analysis as a sub-type of data integration methodologies.

Data integration of heterogeneous data types is currently an active field of research where biostatisticians are constantly proposing hybrid approaches to improve data utilization and scientific discovery. Concepts such as the classification of data as "similar" or "heterogeneous" are still sometimes an open question [29] which clearly depends on the specific context. Hamid and collaborators define data as similar if they are from the "*same underlying source*" (e.g. all gene expression) and as heterogeneous if at least two fundamentally different data sources are involved (e.g. SNP and gene expression). Nevertheless other aspects such as technology may make integration complex, for example, when integrating RNA-seq and microarray based mRNA profiling. Following these definitions, and considering exploitation of datasets with heterogeneous data types involved (either across studies or within studies) then tools such as Co-Inertia Analysis [30,31], Generalized Singular Value Decomposition [32] and Integrative Bi-Clustering [33] among others are relevant. A comparison between these three methodologies in the integrative analysis of mRNA and protein abundance from a study of Plasmodium falciparum is included in this supplement [34]. In this supplement, Reverter et al. [35] propose a kernel PCA methodology that first selects the appropriate kernel for each data type and second combines the kernels from the different data types for a given statistical task.

Moreover, data exploitation in biological research involves not only actual datasets but also previous knowledge (sometimes referred to as Biological Domain Data [29]) which is captured in knowledge databases such as Gene Ontology [36] or the many biological pathway databases such as KEGG [37], or Reactome [38]. Gene Set Enrichment Analysis (GSEA, [39,40]) is a popular approach for integrating previous biological knowledge in the analysis of transcriptomics, which has been extended to other domains such as genomics and proteomics (e.g. [41] in this supplement) and the analysis of genomic regions (GREAT, [42]). Interestingly novel methods are still appearing that incorporate the biological domain knowledge also in the analysis of heterogeneous datasets. This supplement reviews the mathematical background of different methodologies that improve the integration of high-throughput transcriptomics and metabolomics data by incorporating prior knowledge in the form of gene sets and pathways [43].

Brief overview of current approaches to data integration

Data integration is both a challenge and an opportunity and most certainly an increasing reality in genome research. Scientists have acknowledged that biological systems cannot be understood by the analysis of single-type datasets as the regulation of the system certainly occurs at many levels (see [29,44] and in this supplement [45]). Therefore projects have appeared aiming to investigate biological systems at several levels and *create large heterogeneous data-sets*. In several cases, such efforts ended in the design of *novel methodologies to analyze the data*. Furthermore *workshops and conference*s focused on the

topic are starting to proliferate. These three aspects are detailed below.

Data sources

The Human Genome Project [46,47] is probably the most well known biological project before 2000, but during the beginning of 21st century numerous other even more *data-intense* biological projects have been granted research funding. We aim to describe briefly a few of the most relevant projects, and prioritizing those where the resulting datasets are (or will be) publicly available. Other projects of interest not discussed here include the suite of Phantom Projects [48], TRANSFAC database [49] or the previously described GEO.

1000 Genomes Project [50,51] aims to identify those generic genetic variants that have frequencies of at least 1% in the human population by sequencing many individuals with the novel NGS technologies. The project presented a technical challenge of how to store and manage not just the 1000 resulting genomes but the raw and processed data associated with them. The 1000 Genome Project is not as such a data integration driven project but certainly provides useful information on the identification of conserved regions and in GWAS studies.

Encyclopedia of DNA Elements Project (ENCODE, [52-54]): considering that the genomes of several models organisms were nearly completed, ENCODE (*Homo Sapiens*), modENCODE (*C. elegans* and *D. melanogaster* [44]), and mouseENCODE (*Mus musculus* [55,56]) projects were launched with the common goal of identifying all functional elements within the genome, including "*protein-coding genes, non-protein-coding genes, transcriptional regulatory elements, and sequences that mediate chromosome structure and dynamics*" among them. These projects represent truly integration-based approaches which aim to characterize for a set of "animal models and/or tissues and/or cell lines" the profile of mRNA expression (e.g. RNA-seq, CAGE), histone marks and transcription factor binding profiling (ChIP-seq), DNA methylation (RRBS), chromatin conformation (e.g. ChIA-PET, 5C) and the location of active regulatory regions (DNAse-seq) among others. In September of 2012 the ENCODE consortium launched a synchronized publication effort with the preliminary analysis of the data.

The Cancer Genome Atlas Project (TGCA): TGCA's major aim is to generate insights into the heterogeneity of different cancer subtypes by creating a map of molecular alterations for every type of cancer at multiple levels [57]. For instance the endometrial carcinoma has been characterized by mRNA, miRNA, protein, DNA methylation, copy number alterations and somatic chromosomal aberrations [58].

Immunological Genome Project (ImmGen [59]) aims to characterize the mouse immunological system. ImmGen used microarrays to profile the mRNA of most immune cell types under carefully standardized conditions. Interestingly, ImmGen identified the project as a combined effort between immunologists and computational biologists, and is intended as a public resource. Not surprisingly, ImmGen has become a key resource in numerous investigations of the murine and human immune system research (*e.g.* [60]).

Method development

Most of the previous data-intensive projects required the development of novel methodologies to analyze the data. Within ENCODE there has been a considerable effort to identify the relationship between combinations of histone marks and the activity state of DNA elements; Dynamic Bayesian Networks [61] have been used to classify intervals of the genome of K562 into specific classes (e.g. Protein Coding Transcription Start Sites) and more recently self-organizing maps [62] have been used for a similar purpose. Network analysis have also been addressed at ENCODE by the investigation of DNase-seq data, which allows the identification of active regulatory DNA elements, and its integration with Position Weight Matrixes to generate regulatory networks for each ENCODE cell-type [63,64]. To visualize networks circular plots were generated with Circos [65].

Immgen is the data-intensive project where the most advanced network inference methodology has been applied. In [66] authors developed Ontogenet to identify Transcription Factors (TF) acting as differentiation stage-specific regulators of mouse hematopoiesis. The methodology first identified 81 coarse- and 334 fine-grained expression modules, and secondly associates a set of TFs (among a pool of 580 candidates) to each one of these modules by defining the expression level of a module as the weighted linear combination of the associated regulatory TFs; the assignment uses a methodology similar to the Elastic Net [67] or Lasso, but adds penalty functions during the reconstruction of the network that prioritizes similarity (at the TF-module association stage) between cell lines that are closer in the lineage tree.

There is also a relevant need for the development of methodologies aiming to integrate omics and clinical data, both as network-based approaches [68,69] and as both network and data-driven approaches [70]. Overall, previous examples are just the tip of the iceberg of what has been developed, and we expect many more novel developments in the near future.

Conferences, workshop and projects

Scientific meetings on data integration have proliferated in the last decade either as specialized stand-alone conferences or as part of a larger congress. To our knowledge the first International Workshop on Data Integration in the Life Sciences (DILS) took place in Germany in 2004,

and last year, 2013, it was in Montreal; the Workshop aims "*to foster discussion, exchange, and innovation in research and development in the areas of data integration and data management for the life sciences*". This conference, which has a strong computational background [3], has consolidated as a major meeting point in data integration research. Also conferences such as the International Conference on Systems Biology has established workshops on integration-related topics such as metadata or data visualization (ICSB2013). The International Work-Conference on Bioinformatics and Biomedical Engineering of 2014 (IWBBIO2014) contains a special session devoted to *"integration of data, methods and tools in biosciences"*. Recent one-time events of interest are the session on Data Integration hosted at BioMedBridges in 2013; the Statistical Data Integration Challenges in Computational Biology: Regulatory Networks and Personalized Medicine Workshop organized at BIRS [71]; the Workshop in Genomic Data Integration 2013 [72] located at Imperial College, The Next NGS Challenge Conference: Data Analysis and Integration (Valencia, May 2013) and the High-throughput Omics and Data Integration Workshop in February 2013 from which this supplement originated.

Canvassing the research community - a survey on data integration

From February to March 2013 we launched a web inquiry (see Additional file 1), that continued the survey initiated during the *Omics and Data Integration Workshop*, where we investigated major data-integration challenges for the research community in the field of life sciences. The results of this analysis are presented in this section.

Survey: dissemination and biases

By conducting a massive emailing effort among many institutions, the individuals that completed the survey (n = 125) more than doubled the number of registrations in relation to the workshop. Still, most participants were from Europe (80.8%) followed by US (5.6%), mostly from the academic sector (78.4%) and with major expertise in RNA-seq analysis (punctuation: 3/5) Complete DNA Sequencing (punctuation: 2.74/5). We obtained a proper balance between senior (37.6%) and junior (35.2%) researchers, and since the survey was answered by a limited number of individuals (125) we did not consider further stratification. Overall, we acknowledge that the present survey may not represent the views of the entire research community but it does highlight relevant questions and provides initial insights into the opinion of researchers dealing with data integration issues.

Survey: main results

An objective of the STATegra project (and also relevant to the wider scientific community) is to identify current and upcoming needs w.r.t data-analysis thus accelerating the development of novel integrative approaches. To investigate this, we included in the survey a question (4) to identify the major interests in single data types (see Additional file 1) for which individuals were able to select more than one answer. The following aspects of NGS data types caught the largest interest among the responders to the survey: RNA-seq (66.1%) and complete DNA-sequencing (36.3%). The second place was for clinical data (37.9%) followed by proteomics (35.5%). Most individuals were interested in the integrated analysis of multiple data types (72.8%) and this result was independent of participation in the Workshop in Omics Data Integration (Table 1) but significantly correlated with the researcher's expertise (p-value < 0.01).

We next asked which integration schemes for two or more datasets were considered most relevant (Figure 2a upper matrix). We observed that the regulation of gene expression is a major goal and the integration of RNA-seq with all other data-types attracts great attention. Notably integration of clinical data was stated as very important, and this is relevant since this result does not particularly associate with the expertise and interests shown in Supp Table 1 in Additional file 2, but we believe reflects the continuous growth of translational research even in groups devoted to basic science. Integration of proteomics and RNA-seq was considered of high interest, together with a cluster formed by histone marks, transcription factor binding and CpG DNA methylation. We also investigated if these same integration priorities were maintained when thinking of clinical environments (Figure 2a, lower matrix). Clinical data and RNA-seq was the most frequently selected combination, also Clinical Data was highly associated to exome sequencing, complete DNA sequencing followed by metabolomics, proteomics and CpG DNA methylation. Not surprisingly co-morbidities was selected also as a very interesting data type for this setting.

Finally we observed that integration of same-type datasets was also highlighted (Figure 2b). Once more RNA-seq (14.4% basic science; 5.6% clinical environment) and clinical data (4.0%; 5.6%) were considered relevant (Figure 2c). Notably, only integration of several RNA-seq datasets is as highly prioritized as the integration of heterogeneous data types. Results were similar if the analysis was performed after stratification by individuals that "*participated or did not participate*" in the Workshop (results not shown).

Present tools in omics research After stating the interest of the research community in data integration we surveyed their opinion in the availability of appropriate analysis tools. We designed a set of questions where "5" was associated to complete agreement, and "1" to complete disagreement. When considering the analysis of single data types there was an overall consensus (average score = 4.01) on the availability of proper tools, it was considered

Table 1 Scientific interest(s) of survey participants.

	ALL	Workshop participant	Not a participant
Progress in experimental data production methods/technology	25.60%	22.03%	28.79%
Single data-type analysis methods.	29.60%	37.29%	22.73%
Multiple data-type integrated analysis	72.80%	76.27%	69.70%
Biomarker discovery	35.20%	28.81%	40.91%
Understanding of biological mechanisms	56.80%	50.85%	62.12%
Decision support for clinical care	25.60%	16.95%	33.33%

This table summarizes Question 3 results (*Select the developments you are more interested in*). Survey participants were allowed to select more than one answer. The percentages of selected questions are shown for all participations and after stratification by their participation in the workshop

that most software was mainly available for researchers with a programming background (3.45). There was no clear consensus on the availability of user-friendly tools (2.72) but there was on the necessity of developing novel analysis tools in the field (4.55). The average opinion when asked about methods for integrative analysis of multiple data types was slightly different as there was no clear consensus on the availability of proper tools (2.57). Other aspects such as exclusivity of tools for programmers (3.84), existence of user-friendly software (2.21), and need of new analysis tools (4.72) had similar scores.

Future tools in omics research When asked about what should be the major focus in the future the only and almost complete consensus was in the need of developing novel tools in explorative data analysis (4.45), causal discovery tools (4.50), knowledge-bases (4.29) and tools for making public data available and properly organized (4.51). A major requirement was the development of tools first as user-friendly software (4.60) and secondly as Open-Source software such as Bioconductor packages (4.16).

Funding and research participants were asked where funding agencies would be required to invest in order to

	RNA-Seq	ncRNA	ChIP-Seq Histone	ChIP-Seq TF	CpG DNA Methylation	DNase-Seq	Complete DNA sequencing	Exome sequencing	Proteomics	Metabolomics	Chromatin Conformation	Clinical Data	Co-morbidities	Other
RNA-Seq		29.6%	24.8%	29.6%	32.8%	16.0%	21.6%	22.4%	36.8%	21.6%	14.4%	28.0%	10.4%	0.0%
ncRNA	6.4%		8.0%	7.2%	10.4%	4.0%	6.4%	8.0%	5.6%	4.0%	1.6%	10.4%	4.0%	0.0%
ChIP-Seq Histone	6.4%	0.8%		16.0%	16.0%	11.2%	3.2%	4.8%	7.2%	4.0%	8.8%	5.6%	2.4%	0.0%
ChIP-Seq TF	6.4%	0.8%	0.8%		12.0%	16.0%	5.6%	7.2%	9.6%	4.0%	10.4%	7.2%	2.4%	0.0%
CpG DNA Methylation	11.2%	2.4%	3.2%	2.4%		8.8%	9.6%	7.2%	6.4%	4.0%	9.6%	12.0%	4.8%	0.0%
DNase-Seq	4.0%	0.8%	1.6%	2.4%	4.8%		4.0%	5.6%	4.8%	4.0%	10.4%	9.6%	2.4%	0.0%
Complete DNA sequencing	8.8%	1.6%	1.6%	1.6%	2.4%	4.0%		10.4%	13.6%	10.4%	2.4%	20.0%	5.6%	0.0%
Exome sequencing	17.6%	0.8%	1.6%	0.8%	2.4%	0.8%	6.4%		12.0%	8.8%	0.0%	20.0%	7.2%	0.0%
Proteomics	15.2%	1.6%	0.8%	0.8%	1.6%	2.4%	4.8%	8.0%		27.2%	5.6%	16.8%	5.6%	1.6%
Metabolomics	16.8%	2.4%	2.4%	1.6%	3.2%	2.4%	6.4%	4.8%	10.4%		2.4%	17.6%	6.4%	0.8%
Chromatin Conformation	0.8%	0.0%	2.4%	2.4%	0.8%	0.0%	0.8%	0.0%	0.0%	0.8%		4.0%	2.4%	0.0%
Clinical Data	31.2%	8.0%	7.2%	9.6%	15.2%	9.6%	20.0%	21.6%	16.8%	20.0%	4.0%		14.4%	3.2%
Co-morbidities	8.8%	4.0%	3.2%	5.6%	6.4%	4.8%	7.2%	5.6%	2.4%	5.6%	0.8%	16.0%		1.6%
Other	0.8%	0.0%	0.0%	0.0%	0.8%	0.0%	0.8%	0.0%	0.0%	0.0%	0.0%	2.4%	0.8%	

Same data Type in Basic Science	14.4%	6.4%	5.6%	6.4%	4.8%	3.2%	5.6%	4.0%	7.2%	4.8%	2.4%	4.0%	3.2%	1.6%
Same data type in Clinical Environment	5.6%	0.0%	0.0%	0.8%	0.8%	0.0%	2.4%	0.0%	1.6%	4.0%	0.0%	5.6%	0.8%	0.0%

Figure 2 Relevance of integration schemes. (a) Each matrix location (i,j) shows the percentage of survey participants that selected as relevant the integration of data type i and data type j in basic (upper matrix) and clinical (lower matrix) research. (b) shows the percentage of participants that selected as relevant the integration of the same data type for the data types included in the list.

support the coming future of omics integrative analysis. Three questions with a 1 (least interesting) to 4 (most interesting) answer options were provided. Three funding goals were indicated as most relevant: *large publicly available data-sets* (2.35), *large data-sets from cohorts of selected diseases* (2.15) and new tools for data analysis (1.99). Still many participants indicated that other funding priorities were required such as education, a more focused tool development proposal on tools for integration with clinical data, data curation, and "*unification and standardization of all available omics data bases*".

Data standardization Not included in the questionnaire design but mentioned by many responders in questions 16 (*Describe what you think is the most pressing/urgent/important research problem w.r.t data-integration*) and 17 (*Any comment you would like to add*?) was that data standardization is still an open issue. Standardization requirements were identified as two different but linked topics: the need to defined standard formats for every data type - which has been partially successfully managed by the several normalization efforts (e.g. MIAME), and the standardization of metadata. We acknowledge that, despite the enormous effort involved in providing annotated data repositories, the metadata included in many of them is still not sufficiently consistent or comprehensive enough to support large data approaches. *The editorial team agrees that resources must be committed to the developing and continuous support of public data repositories, while focusing not only in the challenges of storing the massive data files but also for more efficient annotation of the data involved. We believe that this goal will be facilitated by journal* policies requesting and controlling submission not only of data but also standardized metadata prior to publication.

Open challenges and discussion

Data integration in the life sciences is not a new challenge, but it is a recurring one that has only recently been unfolded as a major challenge in part driven by technology development producing increasing amounts and type of data. However it is become increasingly clear that to be able to integrate across different types of is not only an opportunity but also a competitive advantage within the biological research community. While the availability of genomics data is reasonably well provided for by publicly accessible and well-maintained data repositories (with the relevant exception of clinical data), there is a need for improved (and novel) annotation standards and requirements in data repositories to enable better integration and reuse of publically available data.

The data exploitation aspect of data integration is probably the one that requires most attention, as it involves (1) the use of prior knowledge - and its efficient storage, (2) the development of statistical methods to analyze heterogeneous data sets and (3) the creation of data explorative tools that incorporate both useful summary statistics and new visualization tools.

We investigated in a survey with 125 responders what the most urgent questions of the research community are regarding data integration. Two relevant observations stand out: first, the need for user-friendly tools targeting integration of heterogeneous datasets; and secondly the relevance of translational medicine, as shown by the interest of incorporating clinical data in most integrative omic studies.

One aspect that we have not discussed in this editorial is that efficient data integration in life sciences may require the creation of novel research profiles. Most bioinformaticians engaged in the analysis of genomics data are either "*trained computer scientists or statisticians devoted to biology*", or "*trained biologists that were required to learn the basics of programming in order to dig deeper into their data*". While both are necessary and have pushed the field forward, it is increasingly recognized that the growth of computational biology requires the reformulation of the teaching system and the appearance of new wider syllabuses that cover all aspects of this interdisciplinary research filed in equivalent detail [73]). This is a major challenge, to raise a new generation of computer savvy researchers with a good understanding of the biology thus enabling development and application of relevant methods for intergration.

A second aspect we have not discussed is the impact of BIGdata analytics in the life sciences. The term BIGdata intuitively describes a situation present in many research fields: the amount of data generated by instruments is exploding, and in many cases doubling over short periods of time. Biology is not an exception: "*since* 2008, *genomics data is outpacing Moore's Law by a factor of 4*" [74]. This situation results in the requirement for developing scalable infrastructures able to manage these quantities of data while making it available for efficient access and indexing. But more interestingly, big data have provided new ways to exploit data in many disciplines, such as economics (see *Data Economy*), business (as in Amazon or Google, [75]), high-energy physics [76] and even biology [77]. The main summary of BIGdata analysis is that even minor changes or low-level associations may be uncovered by the use of (very) large numbers of data points; therefore it remains to be seen how big data concepts will further reshape data integration in the life sciences.

A final aspect is that data integration is also seen as a commercial product and well-established companies (such as Ingenuity or Biomax) are competing with novel companies (such as Anaxomics or LifeMap) in a rapidly advancing field where the commercial edge is constantly being updated.

What is evident is that the era we are living in is nothing else than a paradise for integrative data analysis.

Additional material

Additional file 1: Survey details: *The needs & future in Omics & Data Integration*.

Additional file 2: Supplementary Table 1. Interests (Question 4) and knowledge (Question 7) of participants on different research areas.

Competing interests
No conflict of interest.

Acknowledgements
We would like to sincerely thank PhD Gordon Ball and PhD Ali Mortazavi for the constructive review of the manuscript. The supplement originated thanks to a Workshop co-organised by EU FP7 306000 STATegra and SeqAhead COST Action BM1006. The contribution of DGC, IA, DM, MM, EB, AC and JT was supported by EU FP7 306000 STATegra. The contribution of DGC was also supported by BILS (http://www.bils.se). The contribution of AG, EB and AC was supported by EU COST Action BM1006: SeqAhead. The contribution of AG and EB-R was supported by EU FP7 289452 ALLBIO. The contribution of JT was also supported by Stockholm County Council, and the Swedish Research Council. The contribution of IA was also supported by Åke Wibergs Stiftelsen medicine research Diarienr: 719593091 (http://ake-wiberg.se/).

Declarations
This article has been published as part of *BMC Systems Biology* Volume 8 Supplement 2, 2014: Selected articles from the High-Throughput Omics and Data Integration Workshop. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/8/S2.

Authors' details
[1]Unit of Computational Medicine, Center for Molecular Medicine, Department of Medicine, Karolinska Institute and Karolinska University Hospital, Stockholm, Sweden. [2]Biomax Informatics AG, Munich, Germany. [3]Statistical Cancer Genomics, UCL Cancer Institute; Centre for Mathematics and Physics in the Life Sciences and Experimental Biology, University College London, London WC1E 6BT, UK. [4]Lymphocyte Development Group, MRC Clinical Sciences Centre, Imperial College London, London W12 0NN, UK. [5]Istituto di Tecnologie Biomediche (CNR), Unità Organizzativa di Bari, Via Amendola 122/D, 70126 Bari, Italy. [6]Chromatin and Disease Group, Cancer Epigenetics and Biology Program (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [7]Department of Animal Breeding and Genetics, SLU Global Bioinformatics Centre, Swedish University of Agricultural Sciences, Uppsala, Sweden. [8]Computational Genomics Program, Centro de Investigaciones Príncipe Felipe, Valencia, Spain.

Published: 13 March 2014

References

1. Bairoch A, Cohen-Boulakia S, Froidevaux C: **Review of the selected proceedings of the Fifth International Workshop on Data Integration in the Life Sciences 2008.** *BMC bioinformatics* 2008, **9(Suppl 8)**:S1, doi:10.1186/1471-2105-9-S8-S1.
2. Goble C, Stevens R: **State of the nation in data integration for bioinformatics.** *Journal of biomedical informatics* 2008, **41(5)**:687-93, doi:10.1016/j.jbi.2008.01.008.
3. Philippi S: **Data and knowledge integration in the life sciences.** *Briefings in bioinformatics* 2008, **9(6)**:451, doi:10.1093/bib/bbn046.
4. Stein L: **Creating a bioinformatics nation A web-services model will allow biological data to be fully exploited .** *Nature* 2002, **000**:119-120.
5. Weidman S, Arrison T: **Steps Toward Large-Scale Data Integration in the Sciences: Summary of a Workshop.** *National Research Council of the National Academies* 2010.
6. Berg JM, Tymoczko JL, Stryer L: *Biochemistry.* 7 edition. W.H. Freeman & Company; 2010, (24 Dec 2010).
7. Bogorad IW, Liao JC: **Synthetic non-oxidative glycolysis enables complete carbon conservation.** *Nature* 2013, **502(7473)**:693-7, doi:10.1038/nature12575.
8. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic acids research* 2002, **30(1)**:207-10.
9. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Soboleva A: **NCBI GEO: archive for functional genomics data sets–update.** *Nucleic acids research* 2013, **41(Database issue)**:D991-5, doi:10.1093/nar/gks1193.
10. Schomburg I, Chang A, Placzek S, Söhngen C, Rother M, Lang M, Schomburg D: **BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA.** *Nucleic acids research.* 2013, , **41 Database issue:** D764-72, doi:10.1093/nar/gks1049.
11. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Lander ES: **Genome-scale DNA methylation maps of pluripotent and differentiated cells.** *Nature* 2008, **454(August)**:766-771, doi:10.1038/nature07107.
12. Rakyan VK, Down Ta, Balding DJ, Beck S: **Epigenome-wide association studies for common human diseases.** *Nature reviews Genetics* 2011, **July, doi:10.1038/nrg3000.**
13. Jones Pa: **Functions of DNA methylation: islands, start sites, gene bodies and beyond.** *Nature Reviews Genetics* 2012, **May, doi:10.1038/nrg3230.**
14. Ulf Leser, Felix Naumann: 2007, I-XIII, 1-464.
15. Patrick Ziegler, Dittrich RKlaus: **"Three decades of data integration-All problems solved?".** *University of Zurich* 2004.
16. Fayyad U, Piatetsky-shapiro G, Smyth P: **From Data Mining to Knowledge Discovery in Databases.** *AI Magazine* 1996, 37-54.
17. Abugessaisa I: **Knowledge discovery in road accidents database integration of visual and automatic data mining methods.** *The International Journal of Public Information Systems, ISSN 1653-4360.* 2008, **1**:59-85.
18. Akil H, Martone ME, Van Essen DC: **Challenges and opportunities in mining neuroscience data.** *Science (New York, N.Y.)* 2011, **331(6018)**:708-12, doi:10.1126/science.1199305.
19. **Pathguide [pathguide.org].** .
20. Ponzoni I, Nueda MJ, Tarazona S, Götz S, Montaner D, Dussaut JS, Dopazo J, Conesa A: **Pathway network inference from gene expression data.** *BMC Syst Biol* 2014, **8(Suppl 2)**:S7.
21. Pritchard CC, Cheng HH, Tewari M: **MicroRNA profiling: approaches and considerations.** *Nature reviews. Genetics* 2012, **13(5)**:358-69, doi:10.1038/nrg3198.
22. Bateman A, Agrawal S, Birney E, Bateman A, Agrawal S, Birney E, Stadler PF: **RNAcentral?: A vision for an international database of RNA sequences RNAcentral?: A vision for an international database of RNA sequences.** 2011, doi:10.1261/rna.2750811.22.
23. Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, Sarkans U: **ArrayExpress update–trends in database growth and links to data analysis tools.** *Nucleic acids research* 2013, **41(Database issue)**:D987-90, doi:10.1093/nar/gks1174.
24. Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Vingron M: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nature genetics* 2001, **29(4)**:365-71, doi:10.1038/ng1201-365.
25. Hernández R, Boix-Chova N, Gómez-Cabrero D, Tegner J, Abugessaisa I, Conesa A: **STATegra EMS: an experiment management system for complex next-generation omics experiments.** *BMC Syst Biol* 2014, **8(Suppl 2)**:S9.
26. Van Houwelingen HC, Arends LR, Stijnen T: *TUTORIAL IN BIOSTATISTICS Advanced methods in meta-analysis?: multivariate approach and meta-regression* 2002, **624(June 2001)**:589-624, doi:10.1002/sim.1040.
27. Normand ST: *TUTORIAL IN BIOSTATISTICS META-ANALYSIS?: FORMULATING, EVALUATING, COMBINING, AND REPORTING* 1999, **359(January 1998)**:321-359.
28. Evangelou E, Ioannidis JPa: **Meta-analysis methods for genome-wide association studies and beyond.** *Nature reviews. Genetics* 2013, **14(6)**:379-89, doi:10.1038/nrg3472.
29. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CMT, Beyene J: **Data integration in genetics and genomics: methods and challenges.** *Human genomics and proteomics?: HGP* 2009, **2009, doi:10.4061/2009/869093.**

30. Dolédec S, D C: **Co-inertia analysis: an alternative method for studying species-environment relationships.** *Freshwater Biology* 1994, , **(31):** 277-294.
31. Fagan A, Culhane AC, Higgins DG: **A multivariate analysis approach to the integration of proteomic and gene expression data.** *Proteomics* 2007, **7(13)**:2162-71, doi:10.1002/pmic.200600898.
32. Alter O, Brown PO, Botstein D: **Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100(6)**:3351-6, doi:10.1073/pnas.0530258100.
33. Kaiser S: **Biclustering: Methods, Software and Application.** *PhD thesis* Ludwig-Maximilians-University Munich, Department of Statistics; 2011.
34. Tomescu O, Mattanovich D, Thallinger GG: **Integrative omics analysis. A study based on *Plasmodium falciparum* mRNA and protein data.** *BMC Syst Biol* 2014, **8(Suppl 2)**:S4.
35. Reverter F, Vegas E, Oller JM: **Kernel-PCA data integration with enhanced interpretability.** *BMC Syst Biol* 2014, **8(Suppl 2)**:S6.
36. Gene Ontology Consortium O, Ashburner M, Ball CA, Blake JA, Botstein D, Sherlock G: **Gene Ontology?: tool for the unification of biology.** *Nature Genetics* 2000, **25(1)**:25-29, doi:10.1038/75556.Gene.
37. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic acids research* 2004, **32(Database issue)**:D277-80, doi:10.1093/nar/gkh063.
38. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, D'Eustachio P: **The Reactome pathway knowledgebase.** *Nucleic acids research* 2014, **42(1)**: D472-7.
39. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL: **Gene set enrichment analysis?: A knowledge-based approach for interpreting genome-wide.** 2005.
40. Efron B, Tibshirani R: **On testing the significance of sets of genes.** *The Annals of Applied Statistics* 2007, **1(1)**:107-129, doi:10.1214/07-AOAS101.
41. Schmidt A, Forne I, Imhof A: **Bioinformatic analysis of proteomics data.** *BMC Syst Biol* 2014, **8(Suppl 2)**:S3.
42. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Bejerano G: **GREAT improves functional interpretation of cis-regulatory regions.** *Nature biotechnology* 2010, **28(5)**:495-501, doi:10.1038/nbt.1630.
43. Reshetova P, Smilde AK, van Kampen AHC, Westerhuis JA: **Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data.** *BMC Syst Biol* 2014, **8(Suppl 2)**:S2.
44. Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Waterston RH: **FEATURE Unlocking the secrets of the genome.** 2009, **459(June)**:927-930.
45. Conesa A, Mortazavi A: **The common ground of genomics and systems biology.** *BMC Syst Biol* 2014, **8(Suppl 2)**:S1.
46. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, FitzHugh W: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409(6822)**:860-921, doi:10.1038/35057062.
47. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Holt Ra: **The sequence of the human genome.** *Science (New York, N.Y.* 2001, **291(5507)**:1304-51, doi:10.1126/science.1058040.
48. Suzuki H, Forrest ARR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, de Hoon MJL: **The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line.** *Nature genetics* 2009, **41(5)**:553-62, doi:10.1038/ng.375.
49. Wingender E, Dietze P, Karas H, Knuppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996, **24**:238-241.
50. Gonçalo R, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs Ra, McVean Ga: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467(7319)**:1061-73, doi:10.1038/nature09534.
51. Goncalo R, Auton A, Brooks LD, M. a, Durbin RM, Handsaker RE, McVean Ga: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491(7422)**:56-65, doi:10.1038/nature11632.
52. Ecker JR, Bickmore WA, Barrose I, Segal E: **ENCODE explained.** *Nature* 2012, **489**:, 52-55.
53. Encode T, Consortium P: **A user's guide to the encyclopedia of DNA elements (ENCODE).** *PLoS biology* 2011, **9(4)**:e1001046. doi:10.1371/journal.pbio.1001046.
54. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Hubbard TJ: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome research* 2012, **22(9)**:1760-74, doi:10.1101/gr.135350.111.
55. Shen Y, Yue F, Mccleary DF, Ye Z, Edsall L, Kuan S, Ren B: **A map of the cis-regulatory sequences in themouse genome.** *Nature* 2012, **488(7409)**:116-120, doi:10.1038/nature11243.
56. Mouse ENCODE Consortium, Stamatoyannopoulos Ja, Snyder M, Hardison R, Ren B, Gingeras T, Kaul R: **An encyclopedia of mouse DNA elements (Mouse ENCODE).** *Genome biology* 2012, **13(8)**:418. doi:10.1186/gb-2012-13-8-418.
57. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Clark Graham, Pickering Lisa, Stamp Gordon, Gore Martin, Szallasi Zoltan, Downward Julian, Andrew Futreal P, Swanton Charles, *et al*: **Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing.** *The New England journal of medicine* 2012, **366(10)**:883-892.
58. The Cancer Genome Atlas Research Network: **Integrated genomic characterization of endometrial carcinoma.** *Nature* 2013, **497(7447)**:67-73, doi:10.1038/nature12113.
59. Shay T, Kang J: **Immunological Genome Project and systems immunology.** *Trends in immunology* 2013, **34(12)**:602-9, doi:10.1016/j.it.2013.03.004.
60. Yosef N, Shalek AK, Gaublomme JT, Jin H, Lee Y, Awasthi A, Regev A: **Dynamic regulatory network controlling TH17 cell differentiation.** *Nature* 2013, doi:10.1038/nature11981.
61. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes Ja, Noble WS: **Unsupervised pattern discovery in human chromatin structure through genomic segmentation.** *Nature methods* 2012, **9(5)**:473-6, doi:10.1038/nmeth.1937.
62. Mortazavi A, Pepke S, Jansen C, Marinov GK, Ernst J, Kellis M, Wold BJ: **Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps.** *Genome research* 2013, **0000**:2136-2148, doi:10.1101/gr.158261.113.
63. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos Ja: **Circuitry and dynamics of human transcription factor regulatory networks.** *Cell* 2012a, **150(6)**:1274-86, doi:10.1016/j.cell.2012.04.040.
64. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Stamatoyannopoulos Ja: **An expansive human regulatory lexicon encoded in transcription factor footprints.** *Nature* 2012b, **489(7414)**:83-90, doi:10.1038/nature11212.
65. Krzywinski M, Birol I, Jones SJM, Marra Ma: **Hive plots–rational approach to visualizing networks.** *Briefings in bioinformatics* 2012, **13(5)**:627-44, doi:10.1093/bib/bbr069.
66. Jojic V, Shay T, Sylvia K, Zuk O, Sun X, Kang J, Turley S: **Identification of transcriptional regulators in the mouse immune system.** *Nature immunology* 2013, **14(6)**:633-643, doi:10.1038/ni.2587.
67. Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005, **67(2)**:301-320, doi:10.1111/j.1467-9868.2005.00503.
68. Vidal M, Cusick ME, Barabási A-L: **Interactome networks and human disease.** *Cell* 2011, **144(6)**:986-98.
69. Park J, Lee D-S, Christakis Na, Barabási A-L, Data S: **The impact of cellular networks on disease comorbidity.** *Molecular systems biology* 2009, **5(262)**:262. doi:10.1038/msb.2009.16.
70. Menche J, Sharma A, Cho MH, Mayer RJ, Rennard SI, Celli B, Miller BE, Locantore N, Tal-Singer R, Ghosh S, Larminie C, Bradley G, Riley JH, Agusti A, Silverman EK, Barabási A-L: *A divisive shuffling approach (VIStA) for gene expression analysis to identify subtypes in chronic obstructive pulmonary disease.*
71. **Regulatory Networks and Personalized Medicine Workshop.** [http://www.birs.ca/events/2013/5-day-workshops/13w5083].
72. **Workshop in Genomic Data Integration.** 2013 [http://www2.imperial.ac.uk/~gmontana/data_integration/genomic_data_integration.html].
73. Rost B: **ISCB: past-present perspective for the International Society for Computational Biology.** *Bioinformatics (Oxford, England)* 2014, **30(1)**:143-5.
74. O'Driscoll A, Daugelaite J, Sleator RD: **"Big data", Hadoop and cloud computing in genomics.** *Journal of biomedical informatics* 2013, **46(5)**:774-81.
75. Mayer-Schönberger V, Cukier K: **Big Data: A Revolution That Will Transform How We Live, Work, and Think.** *Eamon Dolan/Houghton Mifflin Harcourt editorial* 2013.
76. Brumfiel BG: **Down the Petabyte Highway.** *Nature* 2011, **469(20)**:282-283.
77. Swarup V, Geschwind DH: **From big data to mechanism.** *Nature* 2013, **500**:4-5.

doi:10.1186/1752-0509-8-S2-I1
Cite this article as: Gomez-Cabrero *et al.*: Data integration in the era of omics: current and future challenges. *BMC Systems Biology* 2014 **8**(Suppl 2):I1.