The Use and Abuse of *-Omes*

Sonja J. Prohaska & Peter F. Stadler

March 18, 2010

Abstract

The diverse fields of *-omics* research share a common logical structure combining a cataloging effort for a particular class of molecules or interactions, the underlying *-ome*, and a quantitative aspect attempting to record spatio-temporal patterns of concentration, expression, or variation. Consequently, these fields also share a common set of difficulties and limitations. In spite of the great success stories of *-omics* projects over the last decade much remains to be understood, not only at the technological but also at the conceptual level. The focus of this contributions is on the dark corners of *-omics* research, where the problems, limitations, conceptual difficulties, and lack of knowledge are hidden.

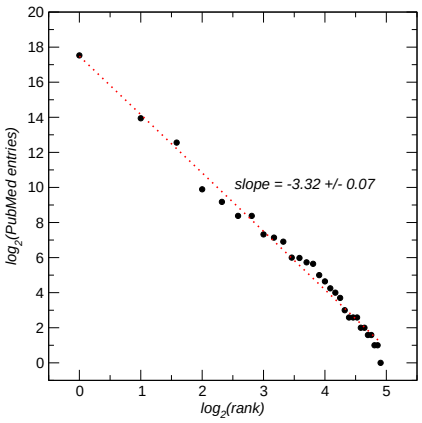
1 Introduction

In cellular and molecular biology, the suffix *-ome* refers to "all constituents considered collectively". The first *-ome* of this kind was defined in the early 20th century. The term "genome" combined the "gene" with the "chromosome" (i.e. "colored body", from the Greek "chromo","color", and "soma", "body") which was known to carry the genes. Bioinformaticians and molecular biologists over the last couple of decades started to widely use *-ome*, adding the suffix to all sorts of biological concepts from epigenes to transcripts, to refer to large data sets produced in one shot by high-throughput technologies. An entertaining commentary on the history of the *-omics* words is [41]. The emerging research field of digesting the resulting pile of data was readily labeled as *-omics*, in analogy to the already time-honored fields of genomics, transcriptomics, and proteomics. The Gerstein Lab keeps a list of the most popular *-omes* on their web page. We have updated their citation data and added several other *-omes* and *-omics* that have made it into PubMed in Tab. 1.

The High throughput *-omics* approaches have profoundly changed Molecular Biology over the past decade. Many exciting success stories have been highlights in top-ranking journals, reviewed and elaborated upon in the following chapters of this book. Here, however, we want to focus on the flipside of the coin, the limitations and shortcomings of the current practice of "*-omics* biology". Being closest to the authors' own work, we use transcriptomics as a paradigmatic example; as we shall see in the next section, however, the issues are generic ones.

Table 1: Usage of *-ome* and *-omics* terms in the scientific literature. PubMed was queried on Fri Jan 29 2010 for "`*ome or *omes`" and "`*omics`" for each of the terms below. The distribution of *-ome* and *-omics* terms follows a power law. Only a handful of top-ranking terms are commonly used.

	PubMed entries		
	-ome[s]		*-omics*
	number	since	number
genome	189019	1943	55750
proteome	15756	1995	23343
transcriptome	6022	1997	778
metabolome	950	1998	1686
interactome	578	1999	43
epigenome	375	1987	189
secretome	333	2000	8
peptidome	160	2001	158
phenome	141	1989	102
glycome	120	2000	479
lipidome	64	2001	279
orfeome	63	2000	1
degradome	53	2003	22
cellome	32	2002	68
fluxome	25	1999	21
regulome	19	2004	2
variome	16	2006	–
toponome	13	2003	7
transportome	8	2004	–
modificome	6	2006	3
translatome	6	2001	2
localizome	6	2002	–
ribonome	4	2002	10
RNome	4	2005	54
morphome	3	1996	1
recombinome	3	2006	–
signalome	2	2001	–
expressome	2	2007	–
foldome	1	2009	1



Transcriptomics aims at cataloging the transcriptome by listing all transcripts. Transcripts of what? The transcriptome of an individual organism, say a mouse, is the set of transcripts that can be found in all of the organisms' cells taken together. In distinction from genomes, transcriptomes vary between cell types, developmental state, and in dependence of environmental stimuli. As a consequence, one may say that a mouse has more then one transcriptome. Distinct individual transcriptomes can be derived from different samples, and such a transcriptome needs to be interpreted in the context of the cells sampled (Figure 1). Emerging single cell methods are expected to show variation even between individual cells of the same cell type. Some of these differences may still be functional, sub-dividing the cell types further into subtypes, with another part of variance constituting neutral stochastic variation.

The power of high-throughput techniques to measure many constituents concurrently in general comes hand-in-hand with a lack of selectivity, sensitivity, and precision. The difficulty of the task is to keep similar but different objects separate and to accurately measure concentrations in parallel that vary by several orders of magnitude. Low-abundance transcripts are either not detected at all, or the signal cannot be separated from background noise inherent in the measurement technology. Quite inconveniently, however, rarity of a functional element is not synonymous with negligibility. In Figure 1,

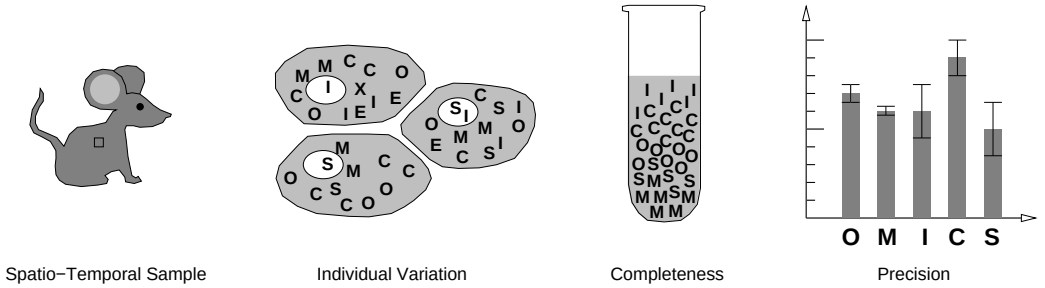


Figure 1: Transcriptomics is the attempt to list and quantify all transcripts and their relative or absolute quantities in the sample. The interpretability of the results depends (1) on the sample that was taken e.g. from a whole mouse. Given some sample from different mice, developmental stages and/or tissues, the transcriptome can be expected to be different. In addition, (2) individual cells can contribute further variation (see "C" and "S"). (3) The completeness of the list is strongly dependent on the measurement technique (in the figure only "one-stroke characters" are analyzed). Transcripts that are not recognized (see "E" and "X") will systematically be missed. Several orders of magnitude separate the high and low abundant transcripts and complicate the quantification process. The measurement technique is also a source of imprecision, making it difficult to distinguished true variation from uncertainty.

the functional elements "X" and "E" are rare, but they are essential to form words of length six or more (e.g. "MEXICO" and "MEIOSIS").

As if the technical issues would not be enough, the biological concept of the entities under investigation may be inadequate and introduce unintended biases. Early transcriptomics, for example, was performed on a poly-T primed cDNA sample, biasing the sample towards mRNA-like transcripts while overlooking even the most highly abundant non-coding RNAs and as well as the mRNAs of histones. A fundamental question in *omics*-style approaches should be: How complete is the catalog really? And in all honesty we have to answer, even for the best understood biological systems, "we don't really know".

Long before the rise of high-throughput techniques we have learned to address the limited precision of measurements. Repetitions of experiments, statistical methods to evaluate significance, and comparative control experiments help us to handle unavoidable imprecision and identify relevant signals. The combination of poorly understood biological variation, of poorly understood technological biases and artifacts (from library preparation to hybridization on chips or sequencing chemistry), and the high costs that in practice limit the number of replicates challenge the available statistical methods and limit to accuracy and confidence of the results. It is still fair to say that there is no high-throughput method that can accurately measure absolute or relative concentrations of transcripts. Indeed, it has become standard operating procedure to validate measurements by means of classical "gold standard", such as Northern blots (to demonstrate the existence of an RNA) or quantitative PCR (to measure concentrations).

The mass of *-omics* data available to day has a profound impact on the way how we think about and organize large-scale data in the life-sciences. Moving away from a molecular towards a systems view of cells or larger functional entities we are expected to handle a single *-ome* with ease. Computational and statistical methods are therefore quickly becoming ubiquitous and indispensable in experimental laboratory environments. Data analysis and interpretation focuses on the comparison of related experimental settings as well as the large-scale integration of different *-omics* on the same or related settings, a task that we see at the very heart of Systems Biology. We argue below that data organization is not just a technical detail. On the contrary, it has a profound impact on the biological interpretation whether transcriptomics data are stored, interpreted, and utilized as expression profiles of DNA intervals (defined by a probe on an array or a piece of sequence), entire transcripts (ESTs), exons (using exon arrays), or entire genes (as defined by various GeneChips).

The eventual goal of all the *-omics* technologies is to measure each of the part in full spatio-temporal resolution for a complete collection of environmental conditions. Wouldn't it be great, say, to have a complete list of all transcripts with exact abundances for every single cell in Mr. Spurlock[1] in 1s time intervals starting from the first bite of his Super Sized fast food meal? (Instead of before and after aggregates of entire organs [68]). Technical limitations will most likely prevent such a SciFi-style experiment, in particular if we stipulate that it should be performed in way that is non-destructive for Mr. Spurlock. We shall see in the following, however, that technical limitations are not the only obstacles, current *-omics* experimentation can also be restricted by problematic experimental design, flawed concepts, and sloppy analysis.

Clearly, the measurement of, say, a spatio-temporally resolved transcriptome is not an ultimate goal in itself. Eventually, the data are to be utilized to infer knowledge of biochemical mechanisms, biological processes, to diagnose an illness, or to assist a therapeutic regimen. Obviously, therefore, *-omics* is eventually about function on a large scale. Just as obviously, a transcript does not make its function explicit by measuring a single, or even a series, of transcriptomes. The task of inferring biological functions, usually known as *annotation*, requires the interlinking of data for different types of experiments. In the *-omics* context, because of the sheer amount of data, this requires the systematic integration of diverse *-omics* data sets.

Before we proceed to this topic in more detail, however, let us analyze the current state of affairs in a somewhat more formal setting.

[1] `http://www.youtube.com/watch?v=I1Lkyb6SU5U`

2 Omology[2]

2.1 Generic Properties

Many of the modern high-throughput methods in the Life-Sciences have claimed a term ending in *-omics*, proclaiming the aim to provide comprehensive descriptions of their subject of study. With the advent of alternative technologies to detect and measure the same, or at least similar, objects, the term has usually shifted to refer to the study of the corresponding *-ome*, i.e., the comprehensive collection of particular type of relevant objects, irrespective of the particular technology that is employed. Not surprisingly, this has also lead to to quite a bit of confusion in the language. *Transcriptomics*, for instance is often used, in particular in a biomedical context, to refer specifically to gene-chip data.

We argue here that a full-fledged *-omics* has three components that make it a coherent scientific endeavor:

1. A suite of (typically high-throughput) technologies address a well-defined collection of biological objects, the pertinent *-ome*.

2. Both technologically and conceptually there is a cataloging effort to enumerate and characterize the individual objects — hopefully coupled with an initiative to make this catalog available as a database.

3. Beyond the enumeration of objects, one strives to cover the quantitative aspects of the *-ome* at hand.

A surprisingly large number of *-omes* adhere to this general outline, Tab 1. Some of them, such as peptidome (dealing with small peptides), secretome (dealing with secreted peptides), or degradome (dealing with proteases) are clearly identifiable as specialized subsets, in this case of the proteome. Only a few *-omes* are pervasively used in the literature. Not surprisingly, the publication statistics follows a power law[3]

Two classes of sub-fields, or flavours, of *-omics* are recurring topics in the literature. *Functional -omics* summarize the attempts to make biological sense out of the wealth of data generated for the *-ome*, typically by ascribing biological functions to the individual objects. One might also say that the goal of *-omics* is to "functionally" annotate the objects of the *-ome*. *Comparative -omics*, on the other hand, makes use of cross-species comparisons of *-omics* data to leverage functional information that has already been determined for one species in order to gain, usually functional, insights on an homologous object in another organism. The four canonical *-omics* that deal with best studies cellular constituents have reached this state, Tab. 2.

[2]We just could not think of a good catchy title for this section. Unfortunately we are not quite immune to the pleasure of coining an utterly useless term and seeing it appear in print. The term *omology* thus designates the systematic study of *-omes*, their associated *-omics*, and interrelations of *-omes*. We express our sincere hope that the field of omology will not survive beyond the end of this section.

[3]We include this observation mostly in the hope that it will impede the appearance of research articles with titles like *"Power-laws in Omology"*.

Table 2: Structural comparison of *-omics* fields. The last two columns list the number of PubMed articles returned by querying the phrase *"comparative -omics"* and *"functional -omics"*.

Field	Objects	quantitative	comparative	functional
Genomics	DNA sequence	variation	2448	5215
Transcriptomics	transcripts	expression	31	4
Proteomics	polypeptides	expression	340	424
Metabolomics	metabolites	concentration	11	32

The wide-spread use of *-omics* data has also fundamentally changed the work-flow in bio-data analysis. Computational and statistical methods are therefore quickly becoming ubiquitous and indispensable in experimental laboratory environments. Even simple routine analyses of Next Generation Sequencing, Tandem-Mass-Spectrometry, or time-resolved imaging already require computational capabilities that go way beyond spreadsheets. Sophisticated bioinformatics has thus become an integral part of *-omics* research, bringing with it algorithmic challenges and the requirement to provide a formal data model that is amenable to efficient processing e.g. in databases. Many of the *-omics* fields in Tab. 1 thus come with an associated "*computational -omics*" that deals specifically with these aspects.

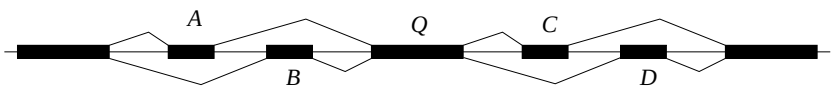
2.2 Limitations

In practice, all *-omics* studies are subject to inaccuracies and bounds on prior knowledge that necessarily restrict our efforts to approximations of what we aspire to measure. An issue of particular importance is the set of – often unspoken – assumptions that underlie the measurement technology, the design of the experiments, and the subsequent data analysis. Due to their similar conceptual construction, all *-omics* fields are confronted with essentially the same set of issues, that are, depending on the maturity of measurement technology and computational analysis techniques, addressed only partially, implicitly, or not at all in the scientific literature. In the following paragraphs we address the issues at a general level, individual examples will be briefly discussed in the subsequent survey.

1. **Technical Limitations** are caused by the measurement technology itself. Notorious examples are the uncertain connections between actual RNA expression levels and the signals produced by the various micro-array technologies [4, 5], biochemical issues that e.g. leave genomic DNA libraries incomplete, or problems in detecting the very small peptides in current proteomics protocols.

2. **Limitations in the Experimental Design** preclude access to certain information not just as a practical matter. Obvious examples are the limitation of GeneChips to mostly coding regions, the implicit assumption in high throughput sequencing that nucleic acids use the canonical

A



$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} c_{AB} \\ c_{BC} \\ c_{AD} \\ c_{BD} \end{pmatrix} = \begin{pmatrix} [A] \\ [B] \\ [C] \\ [D] \end{pmatrix}$$

B

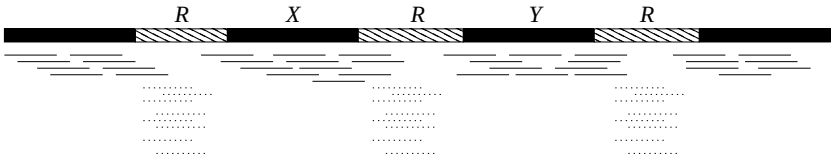


Figure 2: Two examples of design limitations. (**A**) Certain combinations of splice variants cannot be disentangled by exon chips. For instance, the concentrations c_{AB} through cBD of the four transcripts arising from a combination of two pairs A, B and C, D of mutually exclusive exons cannot be inferred from measuring expression levels of the individual exons (or probes across the splice junctions), since the linear equation relating exon signals $[A]$ through $[C]$ to transcript concentrations is not invertible. (**B**) Repetitive elements that are longer than the shotgun reads make it impossible to determine the relative order of the non-repetitive genome fragments X and Y interleaved with the repeats R. Reads located entirely within the repeats could belong to any one of the copies. Unambiguous contigs in pure shotgun assembly are therefore restricted to intervals devoid of long repetitive elements.

4-letter alphabet despite sometimes abundant chemical modifications, or the selection of poly-adenylated mRNAs in EST library construction. There are many less obvious cases as well: Exon Chip approaches cannot disentangle certain combinations of splice variants, and sufficiently long repetitive elements make it impossible to complete genome assemblies based exclusively on shotgun data, Fig. 2. Expression profiles of a mix of different tissues, developmental stages, or even species also fall into this category. Obviously, one better be aware of such limitations in subsequent data analysis and in particular when phrasing the biological interpretation.

3. **Conceptual Limitations** affect both technology and experimental design. The notion of a "gene", for instance, carries a particular burden in that it is not well defined and a closer analysis (see below) shows that its practical application e.g. in GeneChips does not correspond very well with our current understanding of transcriptome structures, Fig. 3. It leaves the nagging question "what is it that we are measuring here?" and raises issues in the biological interpretation of data. Measurements that necessarily produce complex aggregated data that

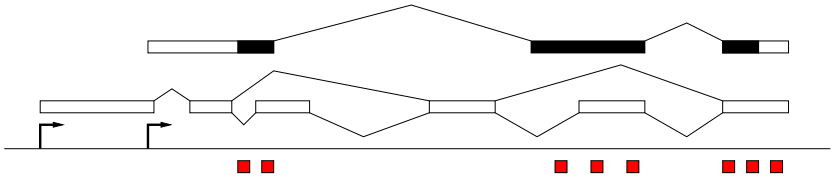


Figure 3: GeneChip design causes difficulties in data interpretation. Each "gene" is assayed by a collection of probes whose signals are averaged to determine the expression level. In this example, the eight probes (little squares) respond to different combinations of splice variants that derive from two different transcripts. The "gene expression level" is therefore a complex combination of the different transcripts that cannot be disentangled without a detailed understanding of both the transcript structures and the concentration-response curve of each individual probe. Such arrangements of overlapping transcripts are common in eukaryotic genomes [70, 32], implying that "gene expression levels" will *typically* be composite signals.

represent a mixed signal arising from multiple distinct physical entities may be hard or even impossible to map to biological reality.

4. **Limitations in the Analysis** cause a partial loss of the measured information at the level of computational analysis. This may have many causes. It may be the case that the analytical tasks are inherently difficult algorithmically or in practice. An example is the analysis of the *in-situ* hybridization images of the BDGP [72] for which several basic image processing problems so far have not been solved satisfactorily, such as the registration of expression patterns from different individuals onto each other. Computational resources may be too limited for more accurate predictions as e.g. in the case of large scale protein structure predictions [54]. In many cases, it is just the convenience of quick-and-dirty analysis with the aim of harvesting the low-hanging fruits[4].

A particularly important issue is that of the *Incompleteness of the Catalog*, i.e., the inability to completely enumerate the members of the *-ome* in questions. Indeed, all *-omics* fields are still in a discovery phase in the sense that none of them can claim to possess a complete catalog of a particular *-ome* (with the possible exception of a few "finished" genomes).

More often than not, holes in the catalog are a composite of several or all of the types of limitations introduced above. Technical limitations make us miss low abundance transcripts because they fall below the detection limits. Implicit assumptions on the transcripts, e.g., on a particular chemistry of their 3' and 5' end [60] may focus on or exclude certain molecules by design. Conceptually, one may have set out to measure concentration of "genes" (whatever these are), leading in practice to a restriction to certain ORF-carrying RNAs in array experiments. The deliberate choice to exclude unspliced ESTs (e.g. because their reading direction cannot be determined)

[4]We are not going to single out particular examples here. These are so numerous that we're confident that the readers of this book have encountered many examples already.

may have been a limitation by design, by computational considerations, or because of a misleading concept stipulating that "interesting genes" arise from spliced mRNAs.

Pragmatic choices, that eventually may turn out to be misleading may also result from the legitimate desire to avoid false positives. If we presuppose that trans-splicing is a rare exotic phenomenon in mammals (see [25, 42] for a few well-documented case studies), we may be inclined to remove *all* chimeric sequences that map to different chromosomal regions or chromosomes as artifacts. From the published literature it is by far not always clear why such choices are made, and whether they have been taken by design, or just because this is how others proceeded before.

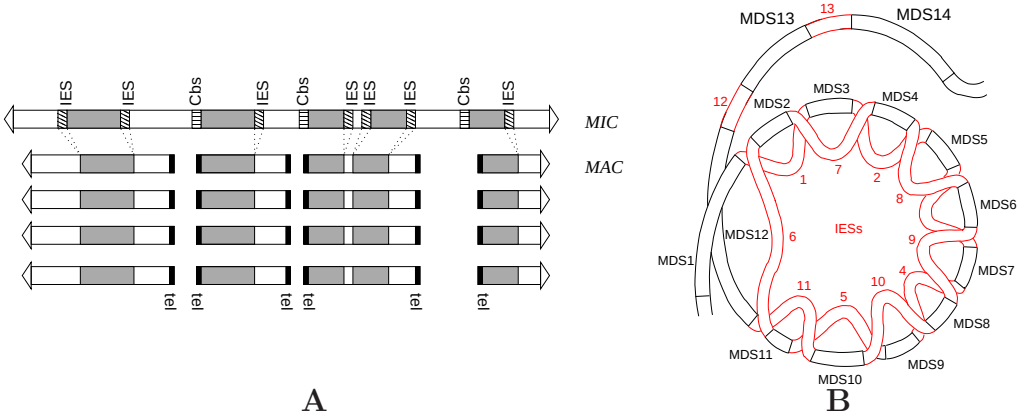
A second serious issue is the comparability and reproducibility of data among different technologies. Here, the high costs of high throughput experiments is the limiting factor on the number of replicates, and the employment of different technologies to measure the same system is usually completely out of reach. The ENCODE Pilot Project [70] may in fact serve a good example how different the results of different technologies can be in detail, and to what extent different method produce complementary rather than congruent results. We do not see this as a problem *per se* — as long as the researcher is aware of these issues and deals with them appropriately in the interpretation of the results. We shall return to this point briefly in the following section.

2.3 What can go wrong, will go wrong

By far the frequently invoked *-omics* approach is **genomics**. Defined as the study of complete genomes, the terminology is the prime exception to the rule outlined above. Genomics is not (or should not be) concerned with the collection of all "genes", since we have learned that genes are just a small subset of the genomic DNA at best [52], and an ill-defined concept at worst [56]. Genomics is not only the oldest and most mature *-omics* it has also the least problem to fulfill the promise to provide "complete catalogs", after all, we can quite safely assume that most prokaryotic genomes are complete and we have a least a few "finished" eukaryotic genomes in at our disposal. Still, the devil is in the details. For instance, the microchromosomes are are notably under-represented in the chicken genome assembly, so that there are still genes known to exist in chicken that are not represented, see e.g. [14]. With most plant and animal genomes still at draft stage, similar artifacts may well have gone unnoticed in some cases.

A related issue of practical relevance it sequence assembly itself. Since genomes are of of course not sequenced as whole, but broken down into pieces that are palatable for the various sequencing technologies (from a few dozen nucleotides for emerging next generation sequencing approaches to about a kilobase for Sanger sequencing) reconstructing the correct genomic sequence is still a formidable computational problem [62]. Current toolkits are still plagued by all sorts of limitations that, in combination with the unavoidable biases in the data generation itself, adversely affect the final result. The *Hox* cluster regions of the chicken genome may serve as an instructive example of the difficulties [59].

Sequence variation across individuals in a populations or among tissues



Ciliates exhibit complex processing of the DNA in the transition from their micronucleus to the macronucleus. (**A**) In *Tetrahymena*, the macronuclear chromosomes (MAC) derive from the micronuclear (germline) DNA by deletion of the MIC specific IES sequences, site specific cleavage at the Cbs sites (15 nt sequence signals) and subsequent attachment of new telomers. Most MAC chromosomes are amplified about 45 times [17]. (**B**) In *Oxitrycha*, some of the micronuclear genes are "encrypted" in addition: they consist of "macronuclear-destined segments" (MDS) that are out of order relative to the functional gene sequence. Upon removal of the IES, the MDS are rearranged in the correct order [55]. Only the pre-processed and amplified macronuclear chromosomes are transcriptionally active.

in the same individuals, i.e., single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) may be seen as a quantitative aspect of the genome. In some cases, developmental variations can be dramatic. Ciliates have long been known for their elaborate processing of genomic DNA during the construction of the functional macronucleus and the transcriptionally inactive micronucleus [33, 16], Fig. 2.3. Chris Amemiya and collaborators recently reported that such mechanisms are also at work much closer to home [67]: The sea lamprey, a jawless vertebrate, undergoes a dramatic remodeling of its genome, resulting in the elimination of hundreds of millions of base pairs from many somatic cell lineages, suggesting that genomes can be much more dynamical than commonly perceived. This opens up yet another can of worms: current genome browsers and genome databases do not seem well equipped to deal with such situations.

Transcriptomics started out as the attempt to quantify the expression levels of mRNAs [63], laying claim to reflecting a sample's transcriptional state in its entirety. Unbiased approaches to cataloging transcripts, such as the large scale cDNA sequencing project FANTOM (mouse) and the *Encyclopedia of DNA elements* (ENCODE, human) which set out to complete the list of protein coding genes, amassed evidence for pervasive transcription of the genome, massive differential expression patterns of non-protein coding transcripts and a much more complex organization of the "transcriptome" [70, 45]. The recent advent of high throughput sequencing technologies not only lead to an ever-increasing flood of transcriptomics data, for instance as part of organized efforts such as the ModENCODE projects and the scale-up phase of ENCODE, but keeps changing our picture of genome/transcriptome

organization by refining and expanding both the complexity of primary transcription and the scope of processing. Even for the best-studied organisms the catalog of transcripts is far from complete and novel genes, including protein-coding ones, as well as a plethora of widely different ncRNAs keep being discovered, see e.g. [15, 28, 32, 13]. Large-scale whole-mount *in-situ* hybridization (ISH) [72, 20] opens up the way to spatially resolved transcriptome analysis. So far, however, broad applications of this approach are hampered by unresolved technical difficulties in both image processing (such as the registration of images from different biological samples onto each other) and subsequent data processing [38, 75, 27].

Proteomics Tandem mass spectrometry followed by database search is currently the predominant technology for peptide sequencing in shotgun proteomics experiments [1]. In practice, this endeavor is not hypothesis-free: Rather, databases of putative polypeptide sequences or previously determined mass-spectra underlie the recognition of fragmentation patterns [31, 44, 39] effectively limiting the scope of most experiments to what is already known or predicted to exist. There are also technical limitations: Most, but by no means all proteins of an organism are detectable by current MS-based proteomics. Small peptides in complex protein mixtures are a particular challenge. Compared to the overall protein expression levels, short peptides often show low abundance, they are easily lost using standard proteomic protocols, and only a limited number of proteolytic peptides can be obtained [22, 36]. Many proteins are modified posttranslationally in chemically very diverse manners, ranging from simple phosphorylation or acetylation to the complex carbohydrate structures of glycoproteins. Specialized methods are necessarily to address these modifications in practice [58, 21]. MELK [65] is an ultrasensitive topological proteomics technology capable of analyzing protein co-localizations with subcellular spatial resolution. It addresses the higher level order in a proteome, referred to as the toponome, coding cell functions by topologically determined webs of interacting proteins. The computational techniques for the analysis of such data are currently under rapid development.

Metabolomics by definition deals with tens of thousands of molecules with small molecular weight, many of which are still of unknown identity. The compounds in question have a wide variety of different functional groups, physicochemical properties, and chemical reactivities, and they appear in distinct pathways in abundances that vary by many orders of magnitudes, from sugars and salts at mM concentrations to vitamins and metabolic intermediates in a nanoMol or even lower concentration range. All these issues limit metabolomics techniques in practice [61] and make incompleteness of data a crucial issue.

Epigenomics is concerned with position-specific information that sits "on top of" the genome, in particular DNA methylation and histone variants and modifications. Both seem to be relevant for the cross-talk between genome and environment and the definition of genomic/cellular states. DNA methylation marks, i.e., methylated cytidines, are also studied under the name **methylome**. MeDIP (methylated DNA immunoprecipitation) is one technique that allows 90-fold enrichment of methylated, genomic DNA fragments due to a antibody specific for methylated cytosine [73]. Sequence

identification of methylated fragments can be carried out by hybridization to a microarray (MeDIP-chip) or by sequencing (MeDIP-seq). To obtain a signal, the fragments need to be long enough and span more than one methylated cytosine for the antibody to bind. The maximal resolution of the MeDIP technology is therefore limited by the size of the fragments sequenced and/or the chip resolution. It is unlikely to fall below 30nt. In theory, single nucleotide resolution can be achieved with bisulfite sequencing. In practice, however, incomplete conversion of unmethylated cytosine to uracil and DNA degradation are limiting [19].

The information contained in nucleosomes and their positions is even more difficult to analyze. The nucleosome is a DNA-protein complex built from 2 times 4 histones $(H3-H4)_2(H2A-H2B)_2$ that wrap up 146nt of DNA. For a human nucleosome, more than hundred distinct marks are possible, which may appear in complex combinations, see [57] for a summary. To date, it is known that histone marks can index the genome, marking transcribed regions, active promoters, direction of transcription, and exon/intron boundaries [2, 49, 66, 71]. Ideally, a map of the epigenome would list the type of the mark, the marked site, distinguish between the two copies of the four histone types, take histone isoforms/paralogs into account, and capture the genomic positioning of the nucleosome *in vivo*. Chromatin immunoprecipitation in combination with DNA sequencing (ChIP-seq) is, currently, the standard high-throughput method to study the genomewide distribution of selected histone modifications. Modified histones can be assigned to a genomic location with a resolution of $\pm$ 50nt. This constitutes an increase by a factor of 10 compared to the resolution of ChIP-chip technologies, but still provide only approximate nucleosome positions. It is hard to evaluate the accuracy of individual modification maps since the antibodies used in the immunoprecipitation step recognize the chemical groups in their histone context and hence might be disturbed by complex modification patterns. At present, no high-throughput method can analyze the histone modification state without relying on the specificity of antibodies.

A quickly increasing number of less well-known *-omics* fields is currently entering the literature, Tab. 1. Some of them are well-defined in the sense of the discussion of this section. **Glycomics** [3, 77], for instance, deals with glycan, i.e., oligo- and polysaccharides. Glycans are branched rather than linear polymers, they frequently incorporate chemically modified sugars, and in contrast to the overwhelming majority of peptides, they are not coded in easily accessible information molecules. Already the determination of their structural formula, i.e., the analog of sequencing presents substantial challenges. Others *-omes*, such as the ORFeome, are probably better understood as particular aspects of more inclusive research areas, in this case proteomics. A similar case is RNomics (focusing on small ncRNAs in its original definition [30] and used in broader sense by many authors since then), which we see as a sub-field of transcriptomics.

2.4 Interactomics

This fast-growing field is concerned with the complex web of interactions that link biological molecules in a cell [35]. In contrast to the *-omics* disciplines

discussed in the previous section, it does not deal with material objects but rather with relationships between them. Nevertheless, it is well-defined to speak of "physical interactions of two or more biomolecules", giving a concrete meaning to the *interactome*. A broad array of techniques is employed to assay different types of interactions. Despite the technological advances of the last years, only certain aspects of the complete interactome are accessible to experimental approaches at present, however.

- Protein-protein interactions are most frequently assessed using yeast-two-hybrid (Y2H) assays, reviewed in [6]. Modern Y2H tools can screen nearly the entire cellular proteome for interactions, and in particular deal with membrane proteins, transcriptionally active proteins, and different localization. Nevertheless, the method suffers from high false positive and false negative rates.
  An alternative to Y2H is affinity purification of protein complexes and subsequent mass spectroscopic analysis (AP/MS) of the intact protein complexes, see e.g. [26]. Again, technical issues limit the accuracy of the attainable data. At present, Y2H should be seen as complementary to emerging AP/MS techniques, and combination of different techniques together with bioinformatics stands a chance to lead to an accurate description of large interaction networks [6].

- DNA-protein interaction can be assayed by large-scale chromatin immunoprecipitation, reviewed e.g. in [74]. The attached nucleic acid sequences are then determined either by quantification on a microarray [76] or by deep sequencing [50]. A similar technology, known as HITS-CLIP, maps RNA-protein binding sites *in vivo* by crosslinking immunoprecipitation and subsequent sequencing of the RNAs [43]. Again, an corresponding microarray-based approach, known as RIP-chip, is possible [34]. This type of approaches is limited by the availability, sensitivity and specificity of antibodies against the protein component for the immunoprecipitation step. The size of the crosslinked DNA or RNA fragments, furthermore, imposes as limitation on the positional resolution.

- Interactions between two nucleic acids molecules are typically mediated by specific base pairing (hybridization). Direct experimental verification of DNA-DNA, DNA-RNA, or RNA-RNA binding interactions is based on detailed chemical probing of the cofolded structure, an approach that is still beyond routine procedures even in a low-throughput scale [7]. Nucleic acid interactions, however, can be predicted computationally with decent (but by far not perfect) accuracy [48, 8, 9, 29]. We are not aware, furthermore, of high-throughput approaches to measuring the interactions of biological macromolecules with small molecules such as metabolites.

In contrast to interactomics, which studies measurable interactions, the related term "regulomics" refers to the study of biological regulation at system-wide scales. As such, regulomics is not at all an *-omics* field in the sense of this article, but rather consitutes a part of systems biology that attempts to reconstruct regulatory networks and mechanisms.

3 Integration of *-Omics* = Systems Biology

3.1 Comparability of *-Omics* Data

Biological systems are neither stationary in time nor homogeneous in space. In fact, much of the success of *-omics* approaches comes from the comparison of variation between individuals, environmental stimuli, tissue types, or cell-lines. One may say, as we did in the introduction, that each sample defines its own transcriptome, proteome, metabolome, etc. From a data management point of view, however, it is much more convenient to speak e.g. of the "mouse transcriptome" as the union of all the transcripts detectable in *Mus musculus*. An particular transcriptomics measurement thus assays the *state* of transcriptome in a particular sample. Interpreting the *-ome* the collection of objects (here: transcripts) that can in principle be observed, it is justified to represent the state as a vector, providing a convenient mathematical starting point for analyzing quantitative data. Of course, this is the setting in which e.g. GeneChip experiments have always been analyzed anyway.

Logically, the comparison of two *-omics* measurements thus requires that we have a way of identifying when two of the assayed objects are the same. This may seem trivial at the first glance. In the case of micro-array experiments this issue is typically hidden in the manufacturer's software, which produces expression levels for "genes" as output. It is easy to conceive examples, however, where different platforms, using different probe locations, interrogate different transcripts. To our knowledge, this issue has not been investigated systematically. That there is an issue, however, is exemplified by several attempts to re-compute the assignment of probes to genes for Affymetrix GeneChips in the light of improving genome assemblies and gene annotations [11].

In the case of sequencing data, this is much less obvious: individual reads first need to be assembled before reconstructed transcripts can be compared. As for ESTs, there is no guarantee that such contigs are complete at both ends hence in general ambiguities remain, including those of Fig. 2. In the words of President Clinton, *it depends what the meaning of "is" is*[5], when we say that object A in experiment I *is* object B in experiment II. The meaning of "is" depends explicitly on a *model* of both the physical entities and the measurement process. In many cases, this model remains implicit, hidden in the computational voodoo of data integration pipelines. One instantiation could be a collection of tables that map probe IDs to gene IDs and gene IDs of different nomenclature and annotation systems to each other.

A particularly striking case are the ambiguities of gene annotations. The concept of the *gene* itself has come under intense scrutiny in response to the recognition that the "standard" model of genes as beads on a genomic DNA string is inconsistent with the findings of high-throughput transcriptomics [51, 53]. As a consequence, several modifications of the concepts of the gene have been explored, ranging from purely structural definitions in terms of groups of transcripts [23], the consideration of transcripts themselves as the central operational units of the genome [24], to functional notions [64], and attempts to reconcile functional and structural aspects [56, 69].

[5] Grand jury testimony, August 17, 1998

In practice, different annotation schemes (GENCODE, RefSeq, VEGA, ENSEMBL) subsume different transcription and protein products under their gene symbols, and they do so based on the expert judgment of individual annotators. Database information linked to gene symbols thus cannot be traced unambiguously to the actual physical entities involved in the original measurements (without going back to the original literature).

3.2 Data Models for *Computational -Omics*

The definition of the entities that constitute the *-ome* naturally determines how particular *-omics* data are stored in data repositories and how they are handled in data integration efforts. If we think of genes as the natural entities of which a transcriptome is composed, it is completely legitimate to represent (possibly spatio-temporal) expression levels of genes as the basic data, and all integration of data, such as the interpretation of variational data from genome resequencing, will have to refer to these basic entities. Modern transcriptomics, however, has shown beyond reasonable doubt, that "genes" are a theoretical construct whose exact meaning remains ambiguous at best and ill-defined at worst, suggesting that "genes" are not a particularly appropriate basis for genome annotation in the first place [56].

We suspect that other *-omics* disciplines suffer to varying degrees from similar issues. Genome browsers, for example, invariably view data relative to a single reference genome, which is entirely adequate for *Escherichia coli* or *Homo sapiens*, makes is difficult to deal with organisms that show a large degree of regulated DNA remodelling. This is not just a matter of technological convenience, but of the deep-sitting preconception that what needs to be understood is how transcripts map to genomic locations, because transcripts *directly* derive from there in the few model organisms that are typically studied. It may not come as a surprise that the very same technology is also used to browse *Tetrahymena*[6], where a more detailed model incorporating the relationships of micro-nucleus and macro-nucleus, Fig. 2.3, might be desirable.

As we argue below, the integration of *-omics* data calls for a data representation that is less focused on a particular task but rather can be viewed as abstractions of biological processes. For instance, transcripts are processed in complex ways, from primary to mature transcripts and further to degradation products which may or may not be functional, Fig. 4. It would be extremely helpful, for example, to have an explicit representation of the `processed from` relation, linking genomic DNA (more precisely, germ-line DNA to accommodate DNA-remodelling as well) to products it encodes. So far, the only connection that is routinely available is the link between a protein-coding mRNA and the polypeptide it encodes.

3.3 *Functional -Omics* = Annotation

The goal of *-omics* analysis is, in many cases, to infer functional information. By definition, this information does not come from within the *-omics* experiment but rather by linking it to other experiments of different type.

[6] www.ciliate.org

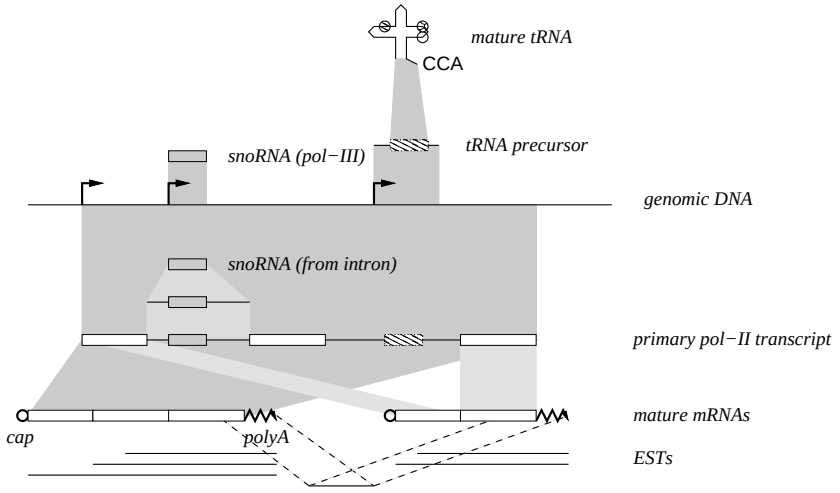


Figure 4: Relationships of processing products are not always obvious and ideally could be stored explicitly. The snoRNA, for instance, might be processed from an intron, via its own promoter, or possibly through either pathway in different contexts. The tRNA is cleaved from a precursor (by RNAse P and RNAse Z), a CCA tail is added (e.g. in human), and several positions are chemically modified. The pol-II transcript is spliced (here in two alternative forms), capped and polyadenylated, while the snoRNA-carrying intron is spliced out. The sequence of processing stages is often known (or can be inferred from the sequences, albeit with some effort) but for the most part these simple relations (gray shades) are not explicitly represented, and hence cannot by utilized.

Typically *function* is expressed in terms of direct interactions ("X is part of the Y-complex"), as involvement in a particular process of action ("Y is part of the apoptosis pathway", "Z contributes to the risk for coronary heart diseases", "W catalyzes the dehydrogenation of malate in the citric acid cycle"), as membership in a particular class of molecules ("Q is a tyrosine kinase") or, in not too many cases, in terms of a mechanistic description of a catalytic activity.

Much of this information is derived from correlations between *-omic* experiments and/or transferred from the annotation of one biological system to another by means of comparative arguments. In particular, sequence homology is routinely used to annotate newly sequenced genomes, using sequence homology as a predictor for functional similarity. Despite the power of homology in spreading annotation information from system to system, eventually this information must be grounded in experimentally verified knowledge, derived, for instance, from related *-omics* experiments.

3.4 Integration of *-Omics* Data

The combination of different *-omics* data is necessarily based on the assumption that we know the relation of different types of data. In some cases this is straightforward: from mature mRNAs we can predict the encoded proteins with rather high certainty, allowing the direct comparison of transcriptomics

and proteomics data (despite all the complications of multiple transcripts and splice variants that may eventually lead to the same protein). The observation that transcript levels are surprisingly poor predictors for protein levels, for instance, was the first indication for the pervasive scale of post-transcriptional regulation. Another example is the combination of genomic variation (SNP) and gene expression data, known as eQTL studies. These provide indications for the effect of mutations on particular regulatory pathways. In practice, however, the interpretation of such data requires a rather detailed model of the pathways in question. In combination with comparative genomics, correlated expression and the presence of common conserved sequence motifs can be utilized to infer cis-regulatory networks, see e.g. [12].

Efforts to large-scale integration of *-omics* data can be seen as an effort to make rather vague statements ("Y is part of the apoptosis pathway") more precise and to place them in the context of a pathway ("Y is down-regulated by X and inhibits Z") by observing correlations and reactions to perturbations of the system. Knowledge about the molecular properties of X, Y, and Z, furthermore may help to make the statement even more precise ("The microRNA Y targets the mRNA of Z. The expression of the primary precursor of Y is inhibited by the transcription factor X").

The model-dependent integration of *-omics* data constitutes one facet of Systems Biology. The other one is the creation and modification of the models themselves, a task that can be rephrased as "generating biological insights". As emphasized e.g. in [40], this requires that the underlying conceptual assumptions are made explicit in the form of a *computational* model so that it can be used in the course of data processing and eventually becomes amenable to rigorous analysis. Conversely, *every* computational integration of *-omics* data explicitly or implicitly presupposes an underlying model of the relationships of the data. The efforts in *-omics* research are, to a large extent, geared towards the construction of a model of biological reality that is as faithful as possible: it strives to provide a *mechanistic explanation* of the molecules and their interactions that is as complete and as accurate as possible.

Many systems biologists describe their field in a somewhat different way: *Systems biology combines experimental techniques and computational methods in order to construct predictive models* [18]. Naïvely, one might think that predictive models are even better than "just" descriptions of the mechanisms. In fact, the inference of the parameters implicitly also determines the (significant) interactions as those with non-zero values. Mathematically, this is formulated as an *inverse problem*, where the unknown network topology and parameters are fitted to reproduced the measures data. This inverse problem, however, is *ill-posed*, i.e., under-determined, in general, so that additional constraints, so-called regularizations, need to be employed. Most commonly, one asks for "sparsity", that is, the number of non-zero parameters should be as small as possible. A similar approach is often taken in the statistical analysis of gene expression data, when "candidate gene" or "gene set" approaches focus on the (usually small) subsets of the strongest response to stimuli or differences in conditions.

It cannot be overemphasized that the interpretation of any reconstruction or parameter estimation is highly model dependent. For example, the dis-

crepancies between transcript and peptide concentrations implies that there are mechanisms of post-transcriptional regulation at work. It may well be possible to represent much or even all of these observed data for example by a complex networks of protein influencing each other's lifetime and to estimate parameters for such a model. The point is that expression data for poly-adenylated mRNAs and polypeptides, even taken together do not imply or even hint at the existence and importance of microRNAs, which as we now know, play an important role in post-transcriptional regulation. Indeed, we are not aware of many hints in the literature before the discovery of microRNAs that a complex additional regulation circuitry such as the RNAi machinery could be missing in our mechanistic understanding of gene regulation, let alone that its specificity would be based on RNAs rather than proteins. To reach this conclusion, the discovery of microRNAs and their characterization as wide-spread regulators was necessary.

At least with the present arsenal of mathematical and statistical methods, we stand little chance to determine the completeness and mechanistic correctness of models by considering time-series, stimulus-response data, or healthy/diseased comparisons. The reason is that nature appears to have evolved robust redundant networks [10].

For this type of systems, the details of the architecture appear to be very hard to infer. Already in simple cases, such as Bayesian Network models of tiny signaling networks, the answer to these question can be disappointing: Huge amounts of data may be required [47]. We should be aware, therefore, that regularized sparse models, whether we obtained them as solution of an inverse problems or whether we constructed them "by hand" (as has been done with most transcription factor networks and signaling pathways) based on *-omics* data and biological insight might work well as predictors for a certain range of circumstances and phenomena, while at the same time being quite far from the mechanistic reality of the biological system.

The Systems Biology approach to solve this problem is to create a sufficiently large data pool focusing on simultaneous experimental approaches in conjunction with improved, model driven approaches to integrating heterogeneous experiments [37, 46]. Unbiased cataloging efforts, such as ENCODE, are another part of the remedy. The discovery of a huge diversity of non-coding RNAs provides us with new pieces, many of which still need to be placed in the puzzle, but at least we know that puzzle is much bigger than anticipated and we can try to stick them into various different models to see where they might fit. Similarly, in order to understand the the role of epigenetic marks in gene expression, it will be necessary to systematically acquire such data as completely as possible.

Even complete parts lists and spatio-temporal correlations between the parts will not be sufficient, however, to build a complete computational model. We strongly suspect that the systematic measurement of physical interactions as constraints on correlation-based data will be indispensable to make the step from sufficient predictive representations to mechanistic explanations.

References

[1] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, 2003.

[2] R. Andersson, S. Enroth, A. Rada-Iglesias, C. Wadelius, and J. Komorowski. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res.*, 19:1732–1741, 2009.

[3] A. Apte and N. S. Meitei. Bioinformatics in glycomics: glycan characterization with mass spectrometric data using SimGlycan. *Methods Mol Biol.*, 600:269–281, 2010.

[4] H. Binder, T. Kirsten, M. Löffler, and P. F. Stadler. The sensitivity of microarray oligonucleotide probes — variability and the effect of base composition. *J. Phys. Chem.*, 108:18003–18014, 2004.

[5] H. Binder and S. Preibisch. "hook" calibration of GeneChip-microarrays: Theory and algorithm. *Alg. Mol. Biol.*, 3:12, 2008.

[6] A. Brückner, C. Polge, N. Lentze, D. Auerbach, and U. Schlattner. Yeast two-hybrid, a powerful tool for systems biology. *Int J Mol Sci.*, 10:2763–2788, 2009.

[7] C. Brunel and P. Romby. Probing RNA structure and RNA-ligand complexes with chemical probes. *Methods Enzymol.*, 318:3–21, 2000.

[8] A. Busch, A. Richter, and R. Backofen. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24:2849–2856, 2008.

[9] H. Chitsaz, R. Salari, S. Sahinalp, and R. Backofen. A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, 25:i365–i373, 2009.

[10] S. Ciliberti, O. C. Martin, and A. Wagner. Innovation and robustness in complex regulatory gene networks. *Proc. Natl. Acad. Sci. USA*, 104:13591–13596, 2007.

[11] W. C. de Leeuw, M. J. Rauwerda, H Jonker, and T. M. Breit. Salvaging affymetrix probes after probe-level re-annotation. *BMC Res Notes*, 1:66, 2008.

[12] A. Dean, S. E. H. Harris, I. Kalajzic, and J. Ruan. A systems biology approach to the identification and analysis of transcriptional regulatory networks in osteocytes. *BMC Bioinformatics*, 10 (Suppl 9):S5, 2009.

[13] M. E. Dinger, P. P. Amaral, T. R. Mercer, and J. S. Mattick. Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. *Brief Funct Genomic Proteomic.*, 8:407–423, 2009.

[14] M. Douaud, K. Fève, M. Gerus, V. Fillon, S. Bardes, D. Gourichon, D. A. Dawson, O. Hanotte, T. Burke, F. Vignoles, M. Morisson, M. Tixier-Boichard, A. Vignal, and F. Pitel. Addition of the microchromosome GGA25 to the chicken genome sequence assembly through radiation hybrid and genetic mapping. *BMC Genomics*, 9:129, 2008.

[15] Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450:203–218, 2007.

[16] S. Duharcourt, G. Lepère, and E. Meyer. Developmental genome rearrangements in ciliates: a natural genomic subtraction mediated by non-coding transcripts. *Trends Genetics*, 25:344–350, 2009.

[17] J. A. Eisen, R. S. Coyne, M. Wu, D. Wu, M. Thiagarajan, J. R. Wortman, J. H. Badger, Q. Ren, P. Amedeo, K. M. Jones, L. J. Tallon, A. L. Delcher, S. L. Salzberg, J. C. Silva, B. J. Haas, W. H. Majoros, M. Farzad, J. M. Carlton, R. K. Smith Jr., J. Garg, R. E. Pearlman, K. M. Karrer, L. Sun, G. Manning, N. C. Elde, A. P. Turkewitz, D. J. Asai, D. E. Wilkes, Y. Wang, H. Cai, K. Collins, B. A. Stewart, S. R. Lee, K. Wilamowska, Z. Weinberg, W. L. Ruzzo, D. Wloga, J. Gaertig, J. Frankel, C.-C. Tsao, M. A. Gorovsky, P. J. Keeling, R. F. Waller, N. J. Patron, J. M. Cherry, N. A. Stover, C. J. Krieger, C. del Toro, H. F. Ryder, S. C. Williamson, R. A. Barbeau, E. P. Hamilton, and E. Orias. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.*, 4:e286, 2006.

[18] H. W. Engl, C. Flamm, P. Kügler, J. Lu, S. Müller, and P. Schuster. Inverse problems in systems biology. *Inverse Problems*, 25:123014, 2009.

[19] M. F. Fraga and M. Esteller. DNA methylation: a profile of methods and applications. *BioTechniques*, 33:632649, 2002.

[20] E. Frise, A. S. Hammonds, and S. E. Celniker. Systematic image-driven analysis of the spatial *Drosophila* embryonic expression landscape. *Mol Syst Biol.*, 6:345, 2010.

[21] Y. Fu, W. Jia, Z. Lu, H. Wang, Z. Yuan, H. Chi, Y. Li, L. Xiu, W. Wang, C. Liu, L. Wang, R. Sun, W. Gao, X. Qian, and S. M. He. Efficient discovery of abundant post-translational modifications and spectral pairs using peptide mass and retention time differences. *BMC Bioinformatics*, 10 (Suppl 1):S50, 2009.

[22] S. Garbis, G. Lubec, and M. Fountoulakis. Limitations of current proteomics technologies. *J. Chromatography A*, 1077:1–18, 2005.

[23] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder. What is a gene, post-ENCODE? history and updated definition. *Genome Res.*, 17:669–681, 2007.

[24] T. R. Gingeras. Origin of phenotypes: Genes and transcripts. *Genome Res.*, 17:682–690, 2007.

[25] T. R. Gingeras. Implications of chimaeric non-co-linear transcripts. *Nature*, 461:206–211, 2009.

[26] A. J. Heck. Native mass spectrometry: a bridge between interactomics and structural biology. *Nat Methods*, 5:927–933, 2008.

[27] A. Heffel, P. F. Stadler, S. J. Prohaska, G. Kauer, and J.-P. Kuska. Process flow for classification and clustering of fruit fly gene expression patterns. In *Proceedings of the 15'th IEEE International Conference on Image Processing, ICIP 2008*, pages 721–724. IEEE, 2008.

[28] M. Hiller, S. Findeiß, S. Lein, M. Marz, C. Nickel, D. Rose, C. Schulz, R. Backofen, S. J. Prohaska, G. Reuter, and P. F. Stadler. Conserved introns reveal novel transcripts in *Drosophila melanogaster*. *Genome Res.*, 19:1289–1300, 2009.

[29] F. W. D. Huang, J. Qin, C. M. Reidys, and P. F. Stadler. Target prediction and a statistical sampling algorithm for RNA-RNA interaction. *Bioinformatics*, 26:175–181, 2010.

[30] A. Hüttenhofer, J. Brosius, and J. P. Bachellerie. RNomics: identification and function of small, non-messenger RNAs. *Curr Opin Chem Biol.*, 6:835–843, 2002.

[31] R. S. Johnson, M. T. Davis, J. A. Taylor, and S. D. Patterson. Informatics for protein identification by mass spectrometry. *Methods*, 35:223–236, 2005.

[32] P. Kapranov, J. Cheng, S. Dike, D. Nix, R. Duttagupta, A. T. Willingham, P. F. Stadler, J. Hertel, J. Hackermüller, I. L. Hofacker, I. Bell, E. Cheung, J. Drenkow, E. Dumais, S. Patel, G. Helt, G. Madhavan, A. Piccolboni, V. Sementchenko, H. Tammana, and T. R. Gingeras. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316:1484–1488, 2007.

[33] L. A. Katz. Evolution and implications of genome rearrangements in ciliates. *J. Euk. Microbiol.*, 52:7S–27S, 2005.

[34] A. M. Khalil, M. Guttman, M. Huarte, M Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander, and J. L. Rinn. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. USA*, 106:11675–11680, 2009.

[35] L. Kiemer and G. Cesareni. Comparative interactomics: comparing apples and pears? *Trends Biotech.*, 25:448–454, 2007.

[36] C. Klein, M. Aivaliotis, J. V. Olsen, M. Falb, H. Besir, B. Scheffer, B. Bisle, A. Tebbe, K. Konstantinidis, F. Siedler, F. Pfeiffer, M. Mann, and D. Oesterhelt. The low molecular weight proteome of *Halobacterium salinarum*. *J Proteome Res*, 6:1510–1518, 2007.

[37] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems Biology in Practice. Concepts, Implementation, and Application*. Wiley, Weinheim, DE, 2005.

[38] S. Kumar, K. Jayaraman, S. Panchanathan, R. Gurunathan, A. Marti-Subirana, and S. J. Newfeld. BEST: a novel computational approach for comparing gene expression patterns from early stages of *Drosophila melanogaster* development. *Genetics*, 162:2037–2047, 2002.

[39] H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, S. E. Stein, and R. Aebersold. Building consensus spectral libraries for peptide identification in proteomics. *Nat Methods*, 5:873–875, 2008.

[40] R. Laubenbacher, V. Hower, A. Jarrah, S. V. Torti, V. Shulaev, P. Mendes, F. M. Torti, and S. Akman. A systems biology view of cancer. *Biochim Biophys Acta.*, 1796:129–139, 2009.

[41] J. Lederberg and A. T. McCray. 'ome sweet 'omics– a genealogical treasury of words. *The Scientist*, 15 [7]:7–8, 2001.

[42] H. Li, J. Wang, X. Ma, and J. Sklar. Gene fusions and RNA trans-splicing in normal and neoplastic human cells. *Cell Cycle*, 8:218–222, 2009.

[43] D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, J. C. Darnell, and R. B. Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456:464–469, 2008.

[44] J. Liu, A. W. Bell, J. J. M. Bergeron, C. M. Yanofsky, B. Carrillo, C. E. H. Beaudrie, and R. E. Kearney. Methods for peptide identification by spectral comparison. *Proteome Science*, 5:3, 2007.

[45] N. Maeda, T. Kasukawa, R. Oyama, J. Gough, M. Frith, P. G. Engström, B. Lenhard, R. N. Aturaliya, S. Batalov, K. W. Beisel, C. J. Bult, C. F. Fletcher, A. R. Forrest, M. Furuno, D. Hill, M. Itoh, M. Kanamori-Katayama, S. Katayama, M. Katoh, T. Kawashima, J. Quackenbush, T. Ravasi, B. Z. Ring, K. Shibata, K. Sugiura, Y. Takenaka, R. D. Teasdale, C. A. Wells, Y. Zhu, C. Kai, J. Kawai, D. A. Hume, P. Carninci, and Y. Hayashizaki. Transcript annotation in FANTOM3: Mouse gene catalog based on physical cdnas. *PLoS Genetics*, 2:e62, 2006. doi:10.1371/journal.pgen.0020062.

[46] F. Marcus. *Bioinformatics and Systems Biology*. Springer, Berlin, 2008.

[47] K. Missal, M. A. Cross, and D. Drasdo. Gene network inference from incomplete expression data: transcriptional control of hematopoietic commitment. *Bioinformatics*, 22:731–738, 2006.

[48] U. Mückstein, H. Tafer, J. Hackermüller, S. B. Bernhard, P. F. Stadler, and I. L. Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22:1177–1182, 2006.

[49] S. Nahkuri, R. J. Taft, and J. S. Mattick. Nucleosomes are preferentially positioned at exons in somatic and sperm cells. *Cell Cycle*, 8:3420–3424, 2009.

[50] P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet.*, 10:669–680, 2009.

[51] H. Pearson. Genetics: What is a gene? *Nature*, 441:398401, 2006.

[52] E. Pennisi. A low number wins the genesweep pool. *Science*, 300:1484, 2003.

[53] E. Pennisi. DNA study forces rethink of what it means to be a gene. *Science*, 316:15561557, 2007.

[54] U. Pieper, N. Eswar, B. M. Webb, D. Eramian, L. Kellz, D. T. Barkan, H. Carter, P. Mankoo, R. Karchin, M. A. Marti-Renom, F. P. Davis, and A. Sali. MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, 37:D347–D354, 2009.

[55] J. D. Prescott, M. L. DuBois, and D. M. Prescott. Evolution of the scrambled germline gene encoding α-telomere binding protein in three hypotrichous ciliates. *Chromosoma*, 107:293–303, 1998.

[56] S. J. Prohaska and P. F. Stadler. "Genes". *Th. Biosci.*, 127:215–221, 2008.

[57] S. J. Prohaska, P. F. Stadler, and D. C. Krakauer. Innovation in gene regulation: The case of chromatin computation. *J. Theor. Biol.*, 2010. in press.

[58] J. Reinders and A. Sickmann. Modificomics: posttranslational modifications beyond protein phosphorylation and glycosylation. *Biomol Eng.*, 24:169–177, 2007.

[59] M. K. Richardson, R. P. Crooijmans, and M. A. Groenen. Sequencing and genomic annotation of the chicken (*Gallus gallus*) *Hox* clusters, and mapping of evolutionarily conserved regions. *Cytogenet Genome Res.*, 117:110–119, 2007.

[60] J. G. Ruby, C. Jan, C. Player, M. J. Axtell, W. Lee, C. Nusbaum, H. Ge, and D. P. Bartel. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell*, 127:1193–1207, 2006.

[61] A. Scalbert, L. Brennan, O. Fiehn, T. Hankemeier, B. S. Kristal, B. v. Ommen, E. Pujos-Guillot, E. Verheij, D. Wishart, and S. Wopereis. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*, 5:435–458, 2009.

[62] K. Scheibye-Alsing, S. Hoffmann, A. M. Frankel, P. Jensen, P. F. Stadler, Y. Mang, N. Tommerup, M. J. Gilchrist, A.-B. N. Hillig, S. Cirera, C. B. Jørgensen, M. Fredholm, and J. Gorodkin. Sequence assembly. *Comp. Biol. Chem.*, 33:121–136, 2009.

[63] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[64] K. Scherrer and J. Jost. The gene and the genon concept: A conceptual and information-theoretic analysis of genetic storage and expression in the light of modern molecular biology. *Th. Biosci.*, 126:65–113, 2007.

[65] W. Schubert, B. Bonnekoh, A. J. Pommer, L. Philipsen, R. Böckelmann, Y. Malykh, H. Gollnick, M. Friedenberger, M. Bode, and A. W. Dress. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nat Biotechnol*, 24:2006, 12701278.

[66] S. Schwartz, E. Meshorer, and G. Ast. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol.*, 16:990–995, 2009.

[67] J. J. Smith, F. Antonacci, E. E. Eichler, and C. T. Amemiya. Programmed loss of millions of base pairs from a vertebrate genome. *Proc Natl Acad Sci USA*, 106:11212–11217, 2009.

[68] M. Somel, H. Creely, H. Franz, U. Mueller, M. Lachmann, P. Khaitovich, and S. Pääbo. Human and chimpanzee gene expression differences replicated in mice fed different diets. *PLoS ONE*, 3:e1504, 2008.

[69] P. F. Stadler, S. J. Prohaska, C. V. Forst, and D. C. Krakauer. Defining genes: A computational framework. *Th. Biosci.*, 128:165–170, 2009.

[70] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, 2007.

[71] H. Tilgner, C. Nikolaou, S. Althammer, M. Sammeth, M. Beato, J. Valcárcel, and R. Guigó. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol.*, 16:996–1001, 2009.

[72] P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S. E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. E. Celniker, and G. M. Rubin. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biology*, 3:R0088, 2002.

[73] M. Weber, J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, and D. Schübeler. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet.*, 37:853–862, 2005.

[74] E. Wong and C. L. Wei. ChIP'ing the mammalian genome: technical advances and insights into functional elements. *Genome Med.*, 1:89, 2009.

[75] J. Ye, J. Chen, R. Janardan, and S. Kumar. Developmental stage annotation of drosophila gene expression pattern images via an entire solution path for lda. *ACM Trans. Knowl. Discov. Data*, 2:1–21, 2008.

[76] S. J. Yoder and S. A. Enkemann. ChIP-on-Chip analysis methods for Affymetrix tiling arrays. *Methods Mol Biol*, 523:367–381, 2009.

[77] J. Zaia. Mass spectrometry and the emerging field of glycomics. *Chem Biol.*, 15, 2008.

Author Addresses

Sonja J. Prohaska

Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany

sonja@bioinf.uni-leipzig.de

Peter F. Stadler

Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany

Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany

Fraunhofer Institut für Zelltherapie und Immunologie – IZI Perlickstraße 1, D-04103 Leipzig, Germany

Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria

Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

studla@bioinf.uni-leipzig.de