Biocuration 2016 - Presentations

Source: http://www.sib.swiss/events/biocuration2016/oral-presentations

Evolution and the proteome: Insights into protein function from deeply conserved gene modules

Edward Marcotte

Transferring annotations between species requires some assurance that the annotated genes are indeed still performing the same functions. I'll describe our attempts to test the extent to which deeply homologous genes and pathways are predictive across distant eukaryotic species, including our search for new models of human disease among phenotypes of distant organisms, our attempts to systematically humanize yeast cells, and our program to apply high-throughput protein mass spectrometry in order to measure conserved physical interactions among the thousands of proteins shared across the eukaryotic tree of life. Studies such as these reveal the evolutionary basis for traits and diseases, help annotate the expressed proteomes of major eukaryotic lineages, and reveal the evolutionary conservation, divergence, and rewiring of protein complexes across eukaryotic proteomes.

iPTMnet: An Integrated Resource for Protein Post-Translational Modification Network Analysis

Cathy Wu, Cecilia Arighi, Hongzhan Huang, Karen Ross, Jia Ren, Sheng-Chih Chen, Gang Li, Catalina O Tudor and K Vijay-Shanker

Protein post-translational modification (PTM) is an essential cellular regulatory mechanism and disruptions in PTM have been implicated in disease. PTMs are an active area of study in many fields, leading to a wealth of PTM information in the scientific literature. There is a need for user-friendly bioinformatics resources that capture PTM information from the literature and support analyses of PTMs and their functional consequences. We have developed iPTMnet (http://proteininformationresource.org/iPTMnet), a resource that integrates PTM information from text mining, curated databases, and ontologies and provides sequence alignment and network visualization tools for exploring PTM networks, PTM crosstalk, and PTM conservation across species. Our text mining tools RLIMS-P and eFIP provide full-scale mining of PubMed abstracts to identify phosphorylation (kinase-substrate-site) information and phosphorylation-dependent protein-protein interactions (PPIs). Experimentally observed PTMs are incorporated from well-curated PTM databases. Proteins and PTM protein forms (proteoforms) are organized using the Protein Ontology, enabling representation and annotation of proteoforms modified on combinations of PTM sites and orthologous relationships between proteoforms. The iPTMnet database currently covers seven PTM types (phosphorylation, acetylation, ubiquitination, methylation, glycosylation, sumoylation and myristoylation) with more than 250,000 PTM sites in more than 45,000 modified proteins, along with more than 1,000 PTM enzymes for human, mouse, rat, yeast, Arabidopsis and several other model organisms. The website supports online search and visual analysis for scientific queries such as: i) finding the substrates for given PTM enzymes; ii) finding PTM sites for given proteins; iii) finding the PTM enzymes that modify given proteins; iv) comparing modified sites and proteoforms for given proteins; v) finding phosphorylation-dependent PPIs, vi) visualizing modification sites in multiple sequence alignment of proteins across species, including known proteoforms; and vii) visualizing networks of PTM enzyme-substrate relations along with sites, proteoforms and PPIs. We further provide a number of use cases illustrating iPTMnet as a gateway for biologists to search, browse, visualize and explore PTM networks, thereby enabling PTM discovery.

Knowledge driven research by combining large-scale genomic data and functional annotations in UniProt.

Andrew Nightingale, Jie Luo, Maria-Jesus Martin, Peter McGarvey and Uniprot Consortium

Over the last decade, life science research has become a data driven scientific field. The challenge for the next decade is to add biological context to these data, transforming the life sciences into a knowledge driven research field. UniProt contributes high quality and comprehensive protein information. By providing additional functional annotations and a new resource to view functional annotations within genome browsers UniProt is facilitating knowledge driven biomedical research. Researchers now have the capability to investigate how genomic alterations can contribute to modifications in the translated protein and evaluate how these can result in a disease or syndrome. In collaboration with Ensembl and other genomic resources; UniProt has mapped all the protein sequences in UniProtKB to the Ensembl reference genomes and imported protein altering variants available from Ensembl for the mapped genomes. For example, for human UniProt has incorporated the 1000 Genome, Catalogue Of Somatic Mutations In Cancer (COSMIC), The Exome Aggregation Consortium (ExAC); consisting of rare disease variants from exome sequencing projects and the NHLBI GO Exome Sequencing Project (ESP); composed of phenotyped populations with heart, lung and blood disorders protein altering variants. UniProt has, in addition, undertaken to define the human mapped protein sequences to genomic coordinates along with important functional positional annotations such as active and metal binding sites, Post Translational Modifications (PTMs), disulfide bonds and UniProtKB/Reviewed variants. The genomic coordinates for UniProtKB human annotations and sequences are distributed in a binary Big bed format allowing users to add UniProtKB annotations as additional tracks in genome browsers. The mapped protein altering variants are distributed as tab delimited files that are available for download at the UniProt FTP site. The UniProt genomic tracks now make it possible to examine the annotations of a disease associated protein within genome browsers. For example, in Alzheimer disease, and Cerebral Amyloid Angiopathy (CAA), specific variants have been associated with driving the amyloidosis. The location of these variants within the genome can now be correlated to structural and functional annotations of the protein. The consequences of more complex genome alterations, found in developmental syndromes or cancers, can also be examined at the translated protein level within genome browsers. For example the deletion or alternative expression of exons can remodel the final protein product and potentially alter its function, its interaction with other proteins or its modulation within a pathway.

TopAnat: GO-like enrichment of anatomical terms mapped to genes by expression patterns

Julien Roux, Mathieu Seppey, Komal Sanjeev, Valentine Rech De Laval, Philippe Moret, Panu Artimo, Severine Duvaud, Vassilios Ioannidis, Heinz Stockinger, Marc Robinson-Rechavi and Frederic B. Bastian

Gene Ontology (GO) enrichment analyses have become a standard to discover information about large lists of genes. They allow to find GO categories that are significantly over-represented or under-represented in a gene list, using the curated associations between GO terms and genes. Here, we propose a new application of gene set enrichment analyses, based on relations between genes and anatomical structures described in the Uberon ontology, computed from gene expression data. This approach is implemented in TopAnat (http://bgee.org/?page=top_anat). Such a test allows to discover in which anatomical structures genes from a given set are preferentially expressed. Note that this is not to be confused with a differential gene expression analysis, where gene expression levels are compared between two conditions, to detect changes in expression. Rather, TopAnat retrieves the anatomical structures where genes are expressed, and for each anatomical structure, tests whether genes from the list of interest are over-associated with this structure, as compared to a background list of genes (by default, all genes with data in Bgee for the species considered). This is exactly the same approach as for GO enrichment tests, applied to anatomical structures.This approach is possible thanks to the integration of gene expression data into the Bgee database. Bgee provides a reference of normal gene expression in animals, comparable between species, currently comprising human and 16 other species. Importantly, Bgee integrates datasets from multiple data types (RNA-Seq, microarray, EST and in situ hybridization data), and transforms all these data into comparable present / absent calls.Annotated gene expression patterns expressed as present / absent are surprisingly powerful to detect biologically relevant information, and the results are very specific. Moreover, all types of data contribute such signal. Thus for example a TopAnat analysis on all mouse genes annotated to the GO term "neurological systems" with RNA-seq data only recovers as top three hits Ammons horn, cerebral cortex and cerebellum; entorhinal cortex, perirhinal cortex and anterior amygdaloid area with microarray only; ganglionic eminence, olfactory epithelium and hypothalamus with in situ hybridization only; and nasal cavity epithelium, organ segment and head organ with EST data only. Because only healthy wild-type gene expression is annotated in Bgee, the enrichment patterns can be readily interpreted.TopAnat shows how the combination of a complex ontology (Uberon) and careful annotation of expression data, both quantitative (RNA-seq, microarrays, ESTs) and qualitative (in situ hybridizations), can be made useful to the biological community.

Using glycobiology in the evaluation of undiagnosed diseases

Jean-Philippe Gourdine, Thomas Metz, David Koeller, Matthew Brush and Melissa Haendel

While only one percent of the genome belongs to genes involved in glycans related processes (i.e. synthesis and degradation of N-glycans, O-glycans, glycolipids, GPI-anchored proteins or free oligosaccharides, glycan-binding protein, etc.), hundreds of diseases are related to glycobiology and can be observed through glycans defects (glyco-phenotypes). Glycan diversity surpasses amino acid, nucleic acid, or lipid diversity but are an underrepresented class of biomolecules in the Human Phenotype Ontology (HPO). Our goal is to integrate data from animal models and known human genetic diseases (disease name, gene, OMIM number, clinical and molecular phenotypes, techniques, etc.) to enable cross-species phenotypic comparison. Towards these ends, we have been curating glycan-related diseases by adding content to community ontologies such as the HPO and CHEBI and curating published articles and textbooks. Along with a hundred identified diseases related to glycobiology, new diseases involving glycan related genes have been discovered within the past decade. The US Undiagnosed Disease Network collects bodily fluids from patients with unknown diseases in order to identify possible new markers and genes by running omics experiments (Exome, Glycome, Proteome, Metabolome, etc.). Using the glycomics data and an ontology for glycobiology, we will be able to compare unknown glyco - phenotypes in order to aid variant prioritization for diagnosis and suggest new phenotyping or omics assays, and garner a better insight of these unknown diseases.

The Cellosaurus, a wealth of information on cell lines

Amos Bairoch

The Cellosaurus (http://web.expasy.org/cellosaurus) is a thesaurus of cell lines. It attempt to list all cell lines used in biomedical research. it scope includes immortalized and naturally immortal cell lines as well as finite cell lines when those are distributed and used widely. It cater for vertebrate cell lines with an emphasis on human, mouse and rat as well as invertebrate (insects and ticks) cell lines. Currently it contains manually curated information on 55'000 cell lines from 517 different species. It includes more than 100'000 cross-references to 47 different ontologies, databases, cell line catalogs, etc. as well as 9'000 distinct publication references and 30'000 synonyms. We are constantly expanding the Cellosaurus, not only in term of the number of cell lines but also in the wealth of information that we provide on these cell lines. Recently we have started adding information on the resistance of cancer cell lines to drugs and other chemical compounds, the protein target of the monoclonal antibodies produces by hybridoma and the transformation method used to immortalize a cell line.We are working in close collaborations with the International Cell Line Authentication Committee (http://iclac.org/) to identify problematic (misidentified or contaminated) cell lines and to foster the use of well behaved nomenclature guidelines.

PAM: A standards-based database for integrating and exchanging pediatrics-specified information from multitude of biomedical resources

Jinmeng Jia and Tieliu Shi

With a significant number of children around world suffer from the consequence of the misdiagnosis and ineffective treatment of various diseases, there is in urgent need for global sharing and exchange of pediatric clinical data for clinical applications and basic researches. A standardized representation of pediatric disease related data is on demand for the purpose. To facilitate the pediatric disease diagnosis and treatment, we built a standards-based database called Pediatrics Annotation & Medicine (PAM) that realizes standardized name and classification of pediatric disease based on International Nomenclature of Diseases (IND), Standard Nomenclature of Diseases and Operations (SNDO), Disease Ontology (DO) and ICD-10. PAM provides both comprehensive disease textual description and standardized conceptual phrases. Meanwhile, it is also a convenient information exchange platform for pediatricians, clinicians and other medical professionals around world to catalog the existing knowledge, integrate either biomedical resources or clinical dada from Electronic Medical Records (EMRs) and to support the development of computational tools which will enable robust data analysis and integration. PAM standardized disease-specified concepts through Unified Medical Language System (UMLS), Disease Ontology (DO) and Human Phenotype Ontology (HPO). In addition, we used disease-manifestation (D-M) pairs from existing biomedical ontologies as prior knowledge to automatically recognize D-M-specific syntactic patterns from full text articles in MEDLINE, and extracted drugs and their dose information in pediatric disease from records reported in Clinical Trials. Both the automatic extracted phenotypes and drug related information were curated manually and standardized using UMLS Concept Unique Identifiers. Currently, PAM contains 3,332 unique pediatric diseases with 14,527 synonyms, 127,048 phenotypes and 208,692 symptoms. Each record-based disease term in PAM now has 6 annotation fields containing definition, synonyms, symptom, phenotype, reference and cross-linkage. Our ongoing task is to extract standardized conceptual-phrases from the textual descriptions using Metamap, Overall, PAM provides a comprehensive standards-based concept list for each disease record and a standardized platform for the information exchange between different biomedical resources.

The OBO Foundry in 2016

Lynn Schriml, Colin Batchelor, Mathias Brochhausen, Melanie Courtot, Janna Hastings, Suzanna Lewis, Chris Mungall, Darren A. Natale, James A. Overton, Bjoern Peters, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard Scheuermann, Barry Smith, Christian J. Stoeckert, Ramona Walls and Jie Zheng

The Open Biomedical Ontologies (OBO) Foundry (http://obofoundry.org) promotes access and sharing of biomedical ontologies and vocabularies, supporting collaborative development of new ontologies and utilization of ontology design principles. The updated OBO Foundry site currently lists over 130 biomedical vocabularies in the OBO Library, those ontologies intending to follow the OBO Foundry principles. The OBO Foundry is governed by a collective of ontology developers that are committed to collaboration and adherence to shared principles, serving on the Operations Committee and Editorial, Technical and Outreach working groups. Collaborative development of ontologies using best practices principles provide multiple advantages for biocurators. The OBO Foundry promotes use of a consistent identifier and metadata scheme for its resources and has set up and actively maintains a system of Permanent URLs for those identifiers. This in turn enables resolution of ontology terms in a web browser for human curators, as well as providing a standard, computable way of referencing each term. Terms need to have a clear label, definition, source of definition (e.g. PubMed ID) as well as an Ontology Coordinator which can be contacted for further clarification or updates. It also ensures terms never disappear: while they can become obsolete, their identifiers remain valid, allowing curators to always find the term corresponding to their annotation. The OBO Foundry recently switched to a new GitHub-backed website, which allows ontology editors to directly update the resource information, thus ensuring continual access to the latest information.In addition to the resources in the OBO Library section of the updated website, the Editorial committee manually reviews and provides feedback on compliance with OBO Foundry principles and whether the ontology can be promoted to OBO Foundry ontology status. The OBO Foundry guidelines and review process have recently been overhauled to provide greater transparency of the review process, including the introduction of an OBO principles compliance self-assessment for ontology editors to complete prior to review. "Foundry" status ontologies comprise a set of interoperable ontologies that are logically well-formed, scientifically accurate and compliant with the OBO Foundry principles including open use, multiplicity of users, collaborative development, non-overlapping and strictly-scoped content, and common syntax and relations. Over the past year, the OBO Foundry Operations Committee has updated documentation, clarified Foundry goals and established regular working group meetings to drive ongoing progress. The OBO Foundry actively encourages ontology developers to join the OBO Foundry community, submit their ontologies for review, and participate in discussions on shaping the OBO Foundry principles.

ICEPO: the Ion Channel ElectroPhysiology Ontology

Valerie Hinard, Aurore Britan, Jean-Sebastien Rougier, Amos Bairoch, Hugues Abriel and Pascale Gaudet

Ion channels are transmembrane proteins that selectively allow ions to flow across the plasma membrane and play key roles in diverse biological processes. A multitude of diseases are due to ion channel mutations such as epilepsies, muscle paralysis, pain syndromes, cardiac arrhythmias, hypoglycemia or autoimmune diseases. A wide corpus of literature is available on ion channels, covering both their functions and their roles in disease. The research community needs to access this data in a user-friendly, yet systematic manner. However, extraction and integration of this increasing amount of data have been proven to be difficult because of the lack of a standardized vocabulary that describes the functioning of ion channels at the molecular level. To address this, we have developed Ion Channel ElectroPhysiology Ontology (ICEPO), an ontology that allows one to annotate the electrophysiological parameters of the voltage-gated class of ion channels. This ontology is based on a three-state model of ion channel gating describing the three conformations/states that an ion channel can adopt: open, inactivated and closed. This ontology supports the capture of voltage-gated ion channel electrophysiological data from the literature in a structured manner and thus enables other applications such as querying and reasoning tools. Here, we present ICEPO, as well as examples of its use.

Comparison of ontology mapping techniques to map traits

Marie-Angelique Laporte, Leo Valette, Laurel Cooper, Christopher Mungall, Austin Meier, Pankaj Jaiswal and Elizabeth Arnaud

The development of improved crop varieties relies on both traditional breeding methods and next-generation methods such as high-throughput sequencing, molecular breeding and automated scoring of traits. Data interoperability issues have led to the development of a number of ontologies of traits over the last years. They fulfil the needs of specific communities but lead to the creation of species or clade-specific ontologies, which limits our ability to perform cross-species analyses. Because these ontologies grow in size and number, manual mapping becomes time-consuming, leading to a need for reliable automated concept mapping techniques to help curators to perform a complete semantic integration, thus opening channels for data integration and discovery.The Crop Ontology (CO, www.cropontology.org) is a suite of 19 species-specific trait ontologies, and is the standard for the international crop breeding and biodiversity community. The CO is developed by Bioversity International in collaboration with breeders and plant biologists in multiple countries. The CO is used both in the field, to directly record crop traits, and for annotation by curators. The Planteome platform (www.planteome.org) aims at supporting comparative plant biology by providing an integrated access to annotated datasets. To address this objective, the Planteome project is currently developing and promoting the use of a set of reference ontologies for plants, proposing species-neutral concepts as well as common data annotation standards. In order to harmonize between the species-specific ontologies and the Planteome reference ontologies, we are currently mapping Crop Ontology traits to the reference ontology Plant Trait Ontology (TO).In this objective, our study shows the benefit of the ontology matching technique based on formal definitions and shared ontology design patterns, compared to standard automatic ontology matching algorithm, such as AML (AgreementMakerLight). First, patterns have been created for each type of traits (e.g Simple Traits, Stress Traits, Phenological Traits) based on Entity-Quality (EQ) statements, accordingly to the patterns introduced in the last release of TO (purl.obolibrary.org/obo/to.owl). In its simplest form, a trait, such as leaf color, can be decomposed as an Entity, leaf, which is well defined in the Plant Ontology, and a Quality, color, coming for instance from Phenotypic Quality Ontology (PATO). The patterns have then been applied to the species-specific ontologies of CO: the CO concepts were decomposed into EQ statements, supported by AML and its AML-Compound algorithm. Finally, ontology reasoners (e.g. HermiT, ELK) automatically inferred the mappings between the ontologies, based on the axioms contained in the defined patterns. The mapping using formal definitions resulted in the alignment of 80% of the CO classes to TO while only 25% were mapped using AML. This increase of successful mappings is mainly due to (i) the use of specialized terms in CO and, (ii) the ability to map a CO class to a TO subclass whereas AML only deals with equivalent classes.As a result of this mapping, TO is being enriched with well defined crop-specific terms of Crop Ontology and Planteome can integrate additional data annotated in a unified way by the breeding and the genetic communities.

Publishing articles and making linked data available: steps towards seamless research reporting

Theodora Bloom

In some fields of study, it is now the norm to report on research in well-structured documents and annotated datasets that are optimally useful for future re-use. In others, there are still many broken links between research reports and the underlying data. Experience over the past few years from a journal editor's perspective yields some lessons for how we can move closer to the ideal for more biomedical fields. I will discuss specific examples including development of the PLOS journals data policy, the Dryad repository working with journals and authors, an RDA working group aiming to highlight best practice in data publishing, and some of the special issues around clinical trials and patient-level data.

UniCarbKB: building a semantic framework for glycomics

Matthew Campbell, Sophie Zhao, Ian Walsh, Miguel Macias, Elisabeth Gasteiger, Julien Mariethoz, Frederique Lisacek, Niclas Karlsson, Peiqing Zhang, Pauline Rudd and Nicolle Packer

The UniCarb KnowledgeBase provides access to a growing, curated database of information on the glycan structures of glycoproteins. It is an international effort that aims to improve our understanding of structures, pathways and networks involved in glycosylation and glyco-mediated processes by integrating structural, experimental and functional glycoscience information. The database consists of two levels of literature - based annotation (i) global-specific data on oligosaccharides released and characterised from single purified glycoproteins and (ii) information pertaining to site-specific glycan heterogeneity. Since its launch in 2012 we have continued to improve the organisation of our data model and expanded the coverage of the human glycome by selecting high quality data collections. This biocuration program has increased the number of annotated glycoproteins from appropriately 270 to over 600 entries with shared UniProt/SwissProt links. Additionally, we have started to collate structural and abundance changes of oligosaccaharides in different human disease states. As the database grows larger and more complex, standardisation has become an important topic. Here, we describe efforts to standardise structural and experimental metadata by using existing ontologies and extending the newly established GlycoRDF vocabulary. To improve interoperability, we have introduced a SPARQL engine that allows users to query the annotations available in UniCarbKB, and perform federated queries with supported glycomics resources (the mass spectrometry database UniCarb-DB and a newly released experimental database relating structure to separation) and other relevant databases including UniProt and neXtProt. We discuss how semantic technologies are improving search capabilities within the glycomics domain, and also facilitating cross-disciplinary connections in the proteomics and genomics domains.

'What does this gene do': Strategies for Summary Data Presentation in MODs

Judith Blake and On Behalf Of The Mouse Genome Informatics Team

The Model Organism Databases (MODs) have a strong history of i) gathering relevant data from the biomedical literature, and by data loads from other sources, ii) strenuously integrating that data for heterogeneous data types, and iii) providing that data to the scientific community in a variety of forms including web interfaces, APIs, dataMines, ftp files, and more. The Mouse Genome Informatics (MGI) project (www.informatics.jax.org), one such MOD, is the community resource for the laboratory mouse, a premier model for the study of genetic and genomic systems relevant to human biology and disease. MGI data includes data from over 220,000 publications for almost 23,000 protein-coding genes. These data include information about 46,000 mutant alleles with over 11,000 genes with mutant alleles in mice. There are over 14,000 genes with detailed developmental expression data, and over 24,000 genes with GO annotations. In addition, MGI captures homology data, especially data about human orthologs (>17,000) and mouse models for human diseases (>4,800). Recently, MGI completed a comprehensive revision of data summation and presentation to provide detailed yet interpretable overviews of information available for mouse genes. MGI Gene Detail pages have always provided links to all the information we have on the mouse gene. With this release, the Gene Detail pages display more information and provide more ways to view subsets of data and access details. New graphical displays provide a synopsis of a gene's functions, where it is expressed, and the phenotypes of mutant alleles. Particular emphasis is on the homology between human genes and diseases, with details about mouse models for these diseases. Links to curated Wikipedia pages for the human genes provide textual summation of available data.Matrix grids that highlight the highest-level categories of gene function, phenotype, and expression data (sometimes referred to as 'slims') represent carefully considered high-level terms that may be transferable to other MODs and -like resources. Each square in the grid drills down to the details of underlying experimental data:o

The SwissLipids knowledge resource for lipid biology

Lucila Aimo, Robin Liechti, Nevila Hyka-Nouspikel, Anne Niknejad, Anne Gleizes, Lou G"otz, Dmitry Kuznetsov, Fabrice P.A. David, Gisou van der Goot, Howard Riezman, Lydie Bougueleret, Ioannis Xenarios and Alan Bridge

Lipids are a large and diverse group of biological molecules involved in membrane formation, energy storage, and signaling. The lipid complement or lipidome of an individual cell, tissue or organism may contain tens of thousands of lipid structures, whose composition and metabolism is tightly regulated in response to changes in cellular signaling and nutritional status. The perturbation of these processes in cancer, hypertension, allergy, diabetes and degenerative diseases, among others, leads to profound alterations in lipidome composition, a fact which underlies the important role of lipids as disease biomarkers and potential diagnostic tools. Modern analytical methodologies such as high-throughput tandem mass-spectrometry provide a means to analyze lipidomes, but a complete understanding of the roles of lipids in human health requires the integration of lipidomic data with biological knowledge. To facilitate this task we have developed a knowledge resource for lipids and their biology - SwissLipids. SwissLipids provides a hierarchical classification that links mass spectrometry analytical outputs to almost 500,000 lipid structures. Integration with metabolic pathways is facilitated by the provision of expert-curated data on enzymatic reactions, interactions, functions, and localization, with supporting links to primary literature and evidence codes (ECO) describing the type of supporting evidence. These annotations are provided using namespaces and ontologies such as UniProtKB, ChEBI, Rhea and Gene Ontology (GO), supplemented by links to matching structures in other reference lipid and metabolite databases such as Lipid Maps and HMDB; they cover the lipids of human and widely studied model organisms including fungi, bacteria, and invertebrates. In summary, SwissLipids provides a reference namespace for lipidomic data publication, data exploration and hypothesis generation. It is updated daily with new knowledge and all data is freely available to search or download from http://www.swisslipids.org/.

LEGO: expressing complex biological models using the Gene Ontology

Paul D. Thomas, Christopher J. Mungall, Seth Carbon, David P. Hill, Suzanna E. Lewis, Huaiyu Mi and Kimberly Van Auken

Accurate descriptions of gene function can be complex, and capturing this information in a computational form is an ongoing challenge. The Gene Ontology (GO) annotation paradigm has dominated the function annotation landscape for years. The advantages of a controlled, semantically precise expression, as well as complete traceability to the underlying scientific evidence, are clear. But GO annotations are currently very simple, and generally associate a single gene product with a single functional concept. We describe a new framework for linking together multiple, simple GO annotations into more complex statements, or models, of gene function. This framework maintains the advantages of the current GO annotation paradigm while opening new opportunities. The GO Consortium has tested this framework extensively, and has developed a tool, called Noctua (noctua.berkeleybop.org), for creating linked GO annotations, or "LEGO models." We describe the framework and tool, and how they specifically address several key barriers that have made GO annotation a challenging enterprise.

HUMAN IMMUNODEFICIENCY VIRUS LIFE CYCLE AND HOST INTERACTION RESOURCE AT VIRALZONE

Patrick Masson, Chantal Hulo, Megan Druce, Lydie Bougueleret, Tulio De Oliveira, Ioannis Xenarios and Philippe Le Mercier

Human Immunodeficiency Virus (HIV) infects 35 million people and is the cause of AIDS, one of the most deadly infectious diseases with over 39 million fatalities. The medical and scientific communities have produced a number of outstanding results in the past 30 years, including the development of over 25 anti-retroviral drugs and over 82,000 publications. To help users to access virus and host genomic data in a useful way, we have created an HIV resource that organizes knowledge and aims to give a broad view of the HIV life cycle and its interaction with human proteins. This resource has been created following an extensive review of the literature. The lifecycle is annotated with controlled vocabulary linked to dedicated pages in ViralZone, Gene Ontology, and UniProt resources. The HIV proteins in the cycle are linked to UniProt and the BioAfrica proteome resource. In total, approximately 3,400 HIV-host molecular interactions have been described in the literature, which represents about 240 partners for each viral protein. This list has been reduced to 57 essential human-virus interactions by selecting interactions with a confirmed biological function. These are all described in a table and linked to publication, sequence database and ontologies. Lastly, this resource also summarizes how antiretroviral drugs inhibit the virus at different stages of the replication cycle. We believe that this is the first online resource that links together high-quality content about virus biology, host-virus interactions and antiviral drugs. This resource is publically accessible at ViralZone website (http://viralzone.expasy.org/).

From paper to the cloud: the mechanics of biocuration may change but not the logic

Donna Maglott

As data are generated, they must be organized to be communicated effectively to others. The primary researcher writes a paper and/or prepares a database submission; peer reviewers and/or database staff validate the content, and multiple downstream processes harvest the information and package it for diverse end users. Over more years than I choose to remember, I have participated in all aspects of this data stream, starting when the method of reporting was completing paper forms via a manual typewriter. I will review aspects of multiple biocuration projects, based on my experiences with Index Medicus, the Hybridoma Data Bank (HDB), the American Type Culture Collection (ATCC) catalog, and several resources at the National Center for Biotechnology Information (NCBI). I will pay special attention to my current work on ClinVar and MedGen, as examples of the challenges of standardizing information in the areas of phenotype, sequence variation, and sequence-phenotype relationships.

Curation in the 100,000 Genomes Project

Ellen M. Mcdonagh, Emma Baple, Mark J. Caulfield, Andrew Devereau, Tom Fowler, Eik Haraldsdottir, Tim Hubbard, Matthew Parker, Augusto Rendon, Antonio Rueda-Martin, Richard Scott, Damian Smedley, Katherine R. Smith, Ellen R.A. Thomas, Clare Turnbull and Caroline Wright

Genomics England was established to deliver the sequencing and interpretation of 100,000 whole genomes from rare disease and cancer patients and their families within the National Health Service England (NHSE). Its four main aims are; to create an ethical and transparent programme based on consent; to bring benefit to patients and set up a genomic medicine service for the NHSE; enable new scientific discovery and medical insights; and to kick start the development of a UK genomics industry. Genomic Medicine Centres (GMCs) were established across England to recruit patients and to collect samples and clinical data required for analysis of the genomes, and validate findings before reporting back to participants. Genomics England Interpretation Partnerships (GeCIPs) span groups of diseases or cross-cutting themes to bring together researchers and clinicians from the NHS and academia to investigate the data generated within the 100,000 Genomes Project and to help improve the interpretation of genomes.Curation is a key element in the analysis and interpretation of the genomes within the 100,000 Genomes project, and has developed in order to adapt to the needs of the project and the analysis pipelines. It is also essential to engage members of the GMCs and GeCIPs in the curation process. Current curation for the project includes establishing virtual gene panels for specific diseases to help the interpretation and tiering of variants found within the genomes, and collecting known pathogenic variants and actionable information in the context of relevance within the NHS and clinical trials available to patients within the UK.To harness the clinical and scientific knowledge of the GMCs, GeCIPs and the wider international Scientific Community, we have created a publicly-available crowdsourcing database "PanelApp" (https://bioinfo.extge.co.uk/crowdsourcing/PanelApp/) to facilitate the curation of gene panels. Gene panels for all the rare diseases within the project were initially established using four sources, and ranked using a traffic light system to provide an initial level of confidence in the gene-disease association. All are available to view and download from PanelApp. Using guidelines based upon the ClinGen and Development Disorder Genotype - Phenotype Database (DDG2P) criteria, we are asking experts to review the gene panels on PanelApp, provide their opinion on whether the gene should be on a clinical-grade diagnostic panel, and contribute further important information such as the mode of inheritance. The reviews made can be seen publicly, with the aim of encouraging open debate and establishing a consensus gene panel for each disorder. Working closely with the Validation and Feedback GeCIP domain, a set of rules have been established for internally evaluating the results of the crowdsourcing effort and subsequent curation requirements, to revise the gene panels for genomic analysis.Key curation issues being tackled within the project include the normalisation of information, mapping phenotypes to standardised ontologies, collating of information where no key resource exists currently, keeping up-to-date with actionable information and relevance to the NHSE. Curation therefore has a critical role in the delivery of the 100,000 Genomes Project, and integration of genomics into the NHS.

Leveraging text mining, expert curation and data integration to develop a database on psychiatric diseases and their genes

Alba Gutierrez Sacristan, `Alex Bravo, Olga Valverde, Marta Torrens, Ferran Sanz and Laura I. Furlong

During the last years there has been a growing research in psychiatric disorders' genetics, supporting the notion that most psychiatric disorders display a strong genetic component. However, there is still a limited understanding of the cellular and molecular mechanisms leading to psychiatric diseases, which has hampered the application of this wealth of knowledge into the clinical practice to improve diagnosis and treatment of psychiatric patients. This situation also applies to psychiatric comorbidities, which are a frequent problem in these patients. Some of the factors that explain the lack of understanding of psychiatric diseases etiology are the heterogeneity of the information about psychiatric disorders and its fragmentation into knowledge silos, and the lack of resources that collect this wealth of data, integrate them, and supply the information in an intuitive, open access manner to the community along with analysis tools. PsyGeNET (http://www.psygenet.org/) has been developed to fill this gap, by facilitating the access to the vast amount of information on the genetics of psychiatric diseases in a structured manner, providing a set of analysis and visualization tools. PsyGeNET is focused on mood disorders (e.g. depression and bipolar disorder), addiction to substances of abuse and schizophrenia.In this communication we describe the process to update the PsyGeNET database, which involves i) extraction of gene-disease associations (GDAs) from the literature with state-of-the-art text mining approaches, ii) curation of the text-mined information by a team of experts in psychiatry and neuroscience, iii) integration with data gathered from other publicly available resources. BeFree, a text mining tool to extract gene-disease relationships, is used to identify genes associated to the psychiatric diseases of interest from a corpus of more than 1M publications. BeFree has a performance of 85% F-score for the identification of genes associated to diseases by exploiting morpho-syntactic features of the text. In addition, it normalizes the entities to standard biomedical ontologies and vocabularies. The text-mined data is then reviewed by a team of experts to validate the GDAs, following specific curation guidelines. Each expert is assigned a disease area according to her/his area of expertise. A web-based annotation tool was developed to assist the curation process. The tool supports a multi-user environment by user and password assignment. It displays the evidence that supports the association for each GDA to the curator. More specifically, it shows the sentences that support the association between the gene and the disease, highlighted (both the sentence and the entities involved in the association) in the context of the MEDLINE abstract. The curator has to validate the particular association based on the evidence of each publication, and select an exemplary sentence that states the association. We also describe the protocol designed to assign the curation tasks to the different experts and the method to assess the inter-annotator agreement. Finally, we describe the approach to integrate the expert-curated data with GDAs identified from other publicly available resources.

Leveraging Wikidata for crowd curation

Andra Waagmeester, Elvira Mitraka, Sebastian Burgstaller-Muehlbacher, Tim Putman, Julia Turner, Justin Leong, Paul Pavlidis, Lynn M. Schriml, Andrew I. Su and Benjamin M. Good

The process of creating, maintaining, updating, and integrating biological databases (biocuration) is, by all accounts, time and resource intensive. Because of this, many within the biocuration community are seeking help from the communities that they serve to extend the scope of what they can accomplish. In principle, such 'community curation' scales well with the rate of knowledge production in science, yet it has proven highly difficult to achieve. In fact, it is fair to say that most wiki-based projects have failed to generate a sufficient critical mass of contributors to make them effective. One approach that has proven successful is to leverage the very large, pre-existing editing community of Wikipedia. Up until now, Wikipedia has provided the means to gather contributions of unstructured, textual content (e.g. reviews of gene function), but it has offered no effective means of crowdsourcing the generation or verification of the structured data that is the principal interest of the biocuration community. With the recent introduction of Wikidata, this has changed. Wikidata is a new Semantic Web compatible database, operated by the MediaWiki foundation as a means to manage the links between articles on the 290+ different language Wikipedias and as a way to structure information content used within the articles. Yet it can be much more than a backend database for the Wikipedias. It is openly editable - by humans and machines - and can contain anything that is of interest. Here, we suggest that the biocuration community can use Wikidata as (1) an outreach channel, (2) a platform for data integration and (3) a centralized resource for community curation.Our group has initiated the process of integrating data about drugs, diseases, and genes from a variety of different authoritative resources directly in the (SPARQL-accessible) Wikidata knowledge base. In our efforts, scripts are developed and applied in close collaboration with curators. They run in frequent update cycles, where data is compared and updated. Input from Wikidata (i.e. disagreement with other sources or users) is in turn looped back to the curators. As a next step, we are looking for more systematic approaches to capture input from the Wikidata community (and through it the Wikipedia community) that can be fed back to the curators of the data sources that are being integrated. We look forward to work with more curation teams to develop mature and effective curation cycles, leveraging the full potential of both professional and amateur curators worldwide.

The Gene Ontology Annotation project: community curation standards and best practices

Melanie Courtot, Aleksandra Shypitsyna, Elena Speretta, Alexander Holmes, Tony Sawford, Tony Wardell, Maria Martin and Claire

The Gene Ontology Annotation (GOA) project at EMBL-EBI is the largest and most comprehensive source of GO annotations to the Gene Ontology (GO) Consortium (245 million as of January 29th 2016).The Protein2GO curation tool we develop, supporting over 24 groups in their GO annotations, has now been extended to support annotations of RNAs via RNAcentral IDs and to macromolecular complexes, identified by IntAct Complex Portal IDs in addition to proteins.While we focus on annotating experimental data, for many species electronic annotations are the only source of information for biological investigation, and it is therefore critical that solid pipelines for data integration across multiple resources be implemented. GOA leverages predictions from other groups to generate new electronic annotations; these include predictions based on species orthologs from Ensembl, and on InterPro signatures, which identify proteins with conserved function or location. In September 2015, an additional 1.5 million annotations from the UniProt Unified Rule (UniRule) system were added electronically; this number has since risen to 2.3 million.In addition to increasing the number of annotations available, GOA also supports, as part of manual curation, the addition of information about the context of a GO term, such as the target gene or the location of a molecular function, via annotation extensions. For example, we can now describe that a gene product is located in a specific compartment of a given cell type (e.g., a gene product that localizes to the nucleus of a keratinocyte). Annotation extensions are amenable to sophisticated queries and reasoning; a typical use case is for researchers studying a protein that is causative of a specific rare cardiac phenotype: they will be more interested in specific cardiomyocytes cell differentiation proteins than all proteins involved in cell differentiation.GOA leads the way in community curation. Because annotations from other international GO curator groups are collected, curators can see what has already been annotated (for gene products or papers), preventing duplication of effort.In collaboration with those curators' groups, we provide guidelines and establish best practices for annotation quality, as well as participating in curation consistency exercises. Many quality checks are implemented; at the syntactic level, annotation rules are enforced at editing time by Protein2GO, as well as custom scripts upon data submission, while at a biological level, submissions of new annotation files are manually checked for correctness and completeness. GOA curators are strictly trained, and, in turn, provide training for others in an effort to ensure best practices for community curation.Thanks to tight collaboration between the GOA team and the EBI GO editorial office, newly developed sections of the ontology often give rise to new annotations and collaboration with new groups, such as ExoCarta and VesiclePedia. Conversely, curation projects inform the ontology editors of missing terms, for example in the recent Antimicrobial peptide project.Annotation files for various reference proteomes are released monthly. The GOA dataset can be queried through our user-friendly QuickGO browser or downloaded in a parsable format via the EMBL-EBI and GO Consortium FTP sites.

From data repositories to submission portals: rethinking the role of domain-specific databases in CollecTF

Sefa Kilic, Dinara Sagitova, Shoshannah Wolfish, Benoit Bely, Melanie Courtot, Stacy Ciufo, Tatiana Tatusova, Claire O'Donovan, Marcus Chibucos, Maria Martin and Ivan Erill

Domain-specific databases are essential resources for the biomedical community, leveraging expert knowledge to curate published literature and provide access to referenced data and knowledge. The limited scope of these databases, however, poses important challenges on their infrastructure, visibility, funding and usefulness to the broader scientific community. CollecTF is a community-oriented database documenting experimentally-validated transcription factor-binding sites in the Bacteria domain. In its quest to become a community resource for the annotation of transcriptional regulatory elements in bacterial genomes, CollecTF has moved away from the conventional data-repository paradigm of domain-specific databases. Through the adoption of well-established ontologies, identifiers and collaborations, CollecTF has progressively become a portal for the annotation and submission of information on transcriptional regulatory elements to major biological sequence resources (RefSeq, UniProtKB and the Gene Ontology Consortium). This fundamental change in database conception capitalizes on the domain-specific knowledge of contributing communities to provide high-quality annotations, while leveraging the availability of stable information hubs to promote long-term access and provide high-visibility to the data. As a submission portal, CollecTF generates transcription factor-binding site information through direct annotation of RefSeq genome records, definition of transcription factor-based regulatory networks in UniProtKB entries and submission of functional annotations to the Gene Ontology. As a database, CollecTF provides enhanced search and browsing, targeted data exports, binding motif analysis tools and integration with motif discovery and search platforms. This innovative approach allows CollecTF to focus its limited resources on the generation of high-quality information and the provision of specialized access to the data.

Biocuration with insufficient resources and fixed timelines

Raul Rodriguez-Esteban

Biological curation, or biocuration, is often studied from the perspective of creating and maintaining databases that have the goal of mapping and tracking certain areas of biology. However, much biocuration is, in fact, dedicated to finite and time-limited projects in which insufficient resources demand trade-offs. This typically more ephemeral type of curation is nonetheless of importance in biomedical research. I propose a framework to understand such restricted curation projects from the point of view of return on curation (ROC), value, efficiency and productivity. Moreover, I suggest general strategies to optimize these curation efforts, such as the "multiple strategies" approach, as well as a metric called overhead that can be used in the context of managing curation resources.

Principles of metadata organization at the ENCODE Data Coordination Center

Cricket Sloan, Eurie Hong, Esther Chan, Jean Davidson, Idan Gabdank, J Seth Strattan, Benjamin Hitz, Jason Hilton, Aditi Narayanan and J Michael Cherry

The ENCODE Data Coordinating Center (DCC) is responsible for organizing, describing,and providing access to the diverse data generated by the ENCODE project. Thedescription of these data, known as metadata, includes the biological sample used asinput, the protocols and assays performed on these samples, the data files generatedfrom the results, and the computational methods used to analyze the data. Here weoutline the principles and philosophy used to define the ENCODE metadata in order tocreate a metadata standard that can be applied to diverse assays and multiple genomicprojects. In addition, we present how the data are validated and used by the ENCODEDCC in creating the ENCODE portal (https://www.encodeproject.org/).

SciCura: a new tool for intuitive annotation

Steven Vercruysse, Marcio Luis Acencio, Astrid Laegreid and Martin Kuiper

Only a relatively small fraction of knowledge from scientific literature makes it into curated databases, despite many large curation projects. This situation keeps deteriorating as the flow of scientific publications continues to increase. Still, the availability of comprehensive knowledge in well-structured formats is the only way in which core background knowledge from scientific research can be used effectively for the production and analysis of new experimental data. This challenge calls for bold approaches, and we believe that partnerships between volunteering scientific domain experts and professional curators can increase the capacity of precise knowledge capture from literature. We demonstrate such a partnership with our work on DNA binding transcription factors (DbTF). Here we jointly designed a set of curation guidelines, which we used to link proteins and DbTF-type GO terms while documenting sufficient amounts of supporting experimental evidence. We are now engaged in finishing the curation of all remaining candidate DbTFs (some 500) from human, mouse and rat for which literature holds experimental evidence.While a variety of curation tools exists for this type of work, we are using our newly developed SciCura, an intuitive curation platform that enables users to compose functional statements of named biological entities and relationships; and where the full content of these statements is supported by ontologies and semantic IDs. In addition, a key feature of SciCura is that users can write any detailed statement as a 'sentence' and indicate its syntax in a straightforward way. A small set of simple rules guides this process, no matter how long or complex the sentence is, or what kind of information it describes. This uniform approach to versatile knowledge capture highly facilitates our efforts to curate diverse experimental evidence in detail.To welcome new users, SciCura supports the definition of 'template' sentences with boxes where words or terms are expected. These boxes offer autocompletion that pre_ranks certain terms (gene ID, interaction type, experiment type, etc) based on which type of information is expected there in the statement. In our use case of curating about a 1000 papers, we defined several detailed templates, each of which take only about a minute to complete after reading the information from a paper. Next, as SciCura is a web-application that enables cooperation, information can be reviewed or adjusted by other curators. Other social interactions such as discussion or rating by a community are possible future extensions for the software. Finally, we implemented several output scripts that query SciCura's API. Particular scripts export facts to several formats, such as tab_delimited for the GO database, or the PSI_MITAB format for the IntAct database.Ultimately hosting this knowledge in well-established databases like GO and IntAct has several advantages: knowledge becomes available to all analysis approaches that use GO annotations and IntAct interaction data; knowledge will be subjected to the time-tested quality and consistency standards implemented by these databases; and all data is further maintained and synchronized with underlying reference sequence databases and controlled vocabularies.

Creating a Standard Schema for Evidence and Provenance in Data Sharing

Kent Shefchek, Matthew Brush, Tom Conlin, Mary Goldman, Mark Diekhans, Melissa Haendel and Pascale Gaudet

In many fields but especially within genomic medicine, aggregating data from multiple sources is fundamental in conducting comprehensive analyses. The data integration challenge is not only in identifying and navigating the diversity of data sources, but in combining differing schemata and data at varying levels of granularity. For genotype-to-phenotype data, there is a spectrum of phenotypic specificity, from asserting a genomic variation confers susceptibility to a disease, to asserting a variation is causal for a single phenotype. For example, ClinVar includes information about gene variants with a four level classification system from "benign" to "pathogenic" and an additional classification of unknown significance. Model organism databases may contain assertions relating a collection of variations, or a genotype, to a set of phenotypes, while protein databases contain statements about cellular processes being associated with specifically modified proteins. Given the clinical impact of such data, the supporting and refuting evidence of these assertions must be made accessible and computable. The combining of sources for a particular assertion can reinforce confidence or inject doubt when different sources are conflicting. For example, in ClinVar, of the records containing clinical significance assertions made from multiple submitters, 21% contain differing assertions. Although the vast majority of disagreement is within one level of classification, for example pathogenic and likely pathogenic, it is imperative that researchers and clinicians are able to investigate the varying lines of evidence used to make a clinical assertion. In addition to lines of evidence, the provenance, or history, of those lines of evidence is of major importance when an assertion is derived from a collection of agents, such as multiple tools and sources. Storing and sharing data provenance allows attribution and promotes reproducibility when the data is used in bioinformatic applications or in aggregate. Despite existing standards for describing provenance and evidence for scientific claims, most do not support the degree of granularity needed for the development of algorithms and tools to evaluate clinical significance for genotype-to-phenotype associations. In partnership with the Monarch Initiative, the BRCA Challenge, and the Global Alliance for Genomics and Health (GA4GH), we propose a standard data model to capture evidence and provenance metadata to support genotype-to-phenotype evaluation. Our approach was to review variant classification protocols from clinical assertions made from variants involved in cancer pathology, model organism genotype-to-phenotype data, and molecular phenotype data. The model proposed is available as an RDF schema for triple and graph stores or JSON schema for documented oriented data stores. It leverages standard IDs and logical definitions from the Evidence Ontology, Statistics Ontology, and the Information

Artifact Ontology. This model is being implemented with data from ClinVar and CIViC and a public API will be available in Spring 2016.

Exploring human-machine cooperative curation

Tonia Korves, Christopher Garay, Robyn Kozierok, Matthew Peterson and Lynette Hirschman

Expert curation can no longer keep up with the huge volume of published literature, making it critical to find ways to speed up and partially automate curation. This raises the question: can humans and machines cooperate to help break the curation bottleneck? We report here on experiments done in the context of the DARPA Big Mechanism program, which focuses on automated extraction and assembly of mechanisms for cancer signaling pathways at scale. A key challenge has been to understand how well automated systems perform in terms of accuracy and throughput. The Big Mechanism year one evaluation focused on capture of pathway information in relation to a given model. In order to evaluate accuracy and throughput, these interactions were captured from papers in three ways: by humans, by automated systems and in a subsequent experiment, by a combination of humans and systems. For the human capture, eight researchers (with minimal training on the curation task) identified interactions found in 39 open access articles, along with supporting evidence from the text. The motivation was to provide a baseline for comparison to machine output and to create a "silver standard" - a set of human-reviewed interactions that were correct (although not necessarily complete). To evaluate performance, we created a small reference set by correcting interactions provided by the humans. The results showed that the human-generated cards needed substantial editing to make them accurate; in particular, the humans did a poor job of providing the correct UniProt identifiers for the proteins in the interactions; overall, only 7.5% of the human extracted interactions were complete and correct, although the humans were good at identifying interactions in terms of participants and interaction type (85% correct). We then performed a follow on experiment, which asked whether machines could improve upon the human-generated results. The experiment was done in two phases. For Phase I, machines were given the sentence from the paper that the human selected as evidence for an interaction. By combining the output from humans and machines, the results rose to correctly cover 48% of the reference set of interactions. In Phase II, machines were given both the human-generated interaction and the evidence sentence. The Phase II results showed that machine-editing was very effective in adding the correct links to UniProt identifiers which improved the percent of correctly captured interactions from 7.5% for humans alone to 73% for the pooled machine-edited cards. This improvement could be attributed to the humans' ability to accurately identify types of interactions and the participating entities, coupled with automated linkage of proteins to the correct UniProt identifiers. These results suggest a way to leverage automated systems and non-expert human curation (for example, through crowdsourcing) to augment expert curation. In the context of the Big Mechanism program, this approach could be used for more efficient creation of a reference set for evaluating advances in machine reading.This work was supported under the DARPA Big Mechanism program, contract W56KGU-15-C-0010.Approved for Public Release. 2016 The MITRE Corporation. All Rights Reserved.

A text-mining method for identifying a scientific program generated data use

Jiao Li, Si Zheng, Hongyu Kang, Li Hou and Qing Qian

The scientific community benefits from the data sharing. By using previous research data, researchers are able to advance scientific discovery far beyond their original analysis. It is challenging to identify data use in full text literature and track the scientific data, although there is a long tradition of partnership between scientific literature and public data in the field of medical sciences. In this study, we conducted an investigation to track the use of scientific data generated by a long-term and government-funded program. We selected the TCGA program to track via analyzing over five thousand full-text articles collecting from PMC. We constructed a benchmark dataset that truly used TCGA data, and compared it with the full-text articles retrieved from PMC. Furthermore, we built up a controlled vocabulary tailed for the TCGA program describing cancer type and high-throughput platform. As a result, we found the number of TCGA related articles increases along the program moving forward. The TCGA publications has increased significantly since 2011, although the TCGA project was launched in 2005. The comparison result shows the similar TCGA feature distribution in the retrieved PMC article set and benchmark dataset, but the article proportions are lower in the retrieved PMC article set than the benchmark dataset. It is because that some articles in the retrieved set merely mentioned the TCGA term rather than actually using the data. Furthermore, we found that glioblastoma (28%), lung cancer (18%) and breast cancer (11%) are the hottest cancer types which data were frequently used. The data generated by the RNA-Seq platform is the most widely used (48%). Our efforts may help develop an automatic method to identify recent publication using the TCGA data. It would facilitate cancer genomics researchers learn the latest progress cancer molecular therapy, as well as promote data sharing and data-intensive scientific discovery.

Text mining genotype-phenotype associations at PubMed scale for database curation and precision medicine

Ayush Singhal, Michael Simmons and Zhiyong Lu

To provide personalized health care it is important to understand patients' genomic variations and the effect these variants have in protecting or predisposing patients to disease. Several projects aim at providing this information by manually curating such genotype-phenotype relationships in organized databases using data from clinical trials and biomedical literature. However, the exponentially increasing size of biomedical literature and the limited ability of manual curators to curate the "hidden" genotype-phenotype relationships in text has led to delays in keeping such databases updated with the current findings. The result is a bottleneck in leveraging the valuable information that is currently available to develop personalized health care solutions. In the past, a few computational techniques have attempted to speed up the curation efforts by using text mining techniques to automatically mine genotype-phenotype information from biomedical literature. However, such previous approaches suffered from limited accuracy and scalability such that they are insufficient for practice use. In this work, we present a novel end-to-end system for mining complete genotype-phenotype relationships (i.e. disease-gene-variation triplets). To achieve high accuracy, our approach exploits state-of-the-art named entity recognition tools for tagging relevant entities (e.g. genes) and advanced machine-learning techniques for extracting relationships. Our approach is also unique because we not only use the local text content (individual articles) but also a global context (from both the Internet and the entire biomedical literature). To validate of our approach, we first performed a benchmarking test on two human-annotated gold-standard test sets. The proposed approach achieved 0.79 and 0.74 in F1-measure for the ternary relationships (disease-gene-variation), which represents approximately 50% improvement over the previous state of the art. Further, to assess the potential utility of our approach, we compared our text-mined results against curated relationships in SwissProt for a total of ten different diseases. This large-scale analysis first confirmed the accuracy of our overall approach, as reported in the benchmarking test. Furthermore, it revealed that a significant portion of text-mined relationships is not currently captured by human curation, but may serve as potential candidates for triage. We conclude that our work represents an important and broadly applicable improvement to the state of the art for mining genotype-phenotype relationships. We believe it will provide necessary support for the implementation of personalized health care using genomic data.

TextpressoCentral: A System for Integrating Full Text Literature Curation with Diverse Curation Platforms including the Gene Ontology Consortium's Common Annotation Framework

Kimberly Van Auken, Yuling Li, Seth Carbon, Christopher Mungall, Suzanna Lewis, Hans-Michael Muller and Paul Sternberg

Manual, full-text literature curation underlies much of the annotation found in biological databases, but the process of finding relevant papers and specific supporting evidence for annotations can be very time-consuming. To ameliorate the cost and tedium of manual curation, we have developed the TextpressoCentral text mining system to allow for direct annotation of full text and integration of that annotation into diverse curation systems. TextpressoCentral allows for sophisticated literature queries using keywords and semantically related categories, but enhances the user experience by providing search results in the context of full text. Resulting full text annotations can readily be integrated into any user-defined curation platform and also used to develop training sets to further improve text mining performance. The current implementation of TextpressoCentral includes all articles from the PMC Open Archive and uses the Unstructured Information Management Architecture (UIMA) framework as well as Lucene indexing. To maximize the utility of the system, TextpressoCentral also allows for processing and displaying of results from third-party text mining and natural language processing algorithms. As a first step towards integrating TextpressoCentral into a real-world curation pipeline, we are collaborating with the Gene Ontology Consortium (GOC) to integrate TextpressoCentral into the GOC's Common Annotation Framework, including the Noctua curation tool developed to support GO's more expressive LEGO curation paradigm.

The UniProtKB guide to the human proteome

Lionel Breuza, Michele Magrane and Uniprot Consortium

Researchers now routinely perform whole genome and proteome analysis thanks to high-throughput and advanced technologies. To help them in this task, they need high quality resources providing comprehensive gene and protein sets. They use these sets both as references to map and compare their data and as a source of knowledge. Using the example of the human proteome, we will describe the content of a complete proteome in the UniProt Knowledgebase (UniProtKB). We will show how we collect and combine information on proteins including function, localization, interactions, sequence, structure and variations. We will explain why literature-based expert curation is central to capture this knowledge. We will show how we use expert-driven automatic pipelines to filter reliable high-throughput data, extract knowledge and curate it into UniProtKB without reducing quality. Finally, we will describe the way manual expert curation of UniProtKB/Swiss-Prot is complemented by expert-driven automatic annotation to build a comprehensive, high quality and traceable resource.

GO annotation in InterPro: why stability does not indicate accuracy in a sea of changing annotations

Amaia Sangrador, Alex Mitchell, Hsin-Yu Chang, Siew-Yit Yong and Robert D. Finn

The removal of annotation from biological databases is often perceived as an indicator of prior erroneous annotation. As a corollary, annotation stability is considered to be a measure of reliability. However, diverse data-driven events can affect the stability of annotations in both primary protein sequence databases and the protein family databases that are built upon the sequence databases and used to help annotate them. The InterPro database integrates 11 protein families databases (CATH-Gene3D, HAMAP, PANTHER, Pfam, PRINTS, ProDom, Prosite Patterns and Profiles, SMART, SUPERFAMILY, and TIGRFAM) into a single resource to provide a single, comprehensive resource for understanding protein families, domains and functional sites. As part of the value added information provided by InterPro, curators annotate each database entry with literature referenced abstracts and additional functional annotations, including Gene Ontology (GO) terms where possible. Currently, with more than 108 million term associations to UniProt sequences at release 2016_01, InterPro continues to be one of the main sources of GO annotation. InterPro GO annotations are largely stable with ~98% of terms remaining from release to release. However, the dynamic nature of the underlying data sources integrated within InterPro means that some changes are necessary. Here, we describe the underlying events driving the changes and their consequences for the InterPro database. We demonstrate that annotation removal or reassignment is rarely linked to incorrect annotation by the curator, but are necessitated by both changes in the underlying data and improved published scientific knowledge. The annotation changes underline the vital importance of continued curation, both within InterPro and the resources that it is built upon. Within InterPro the annotation of database entries should never be considered completely resolved.

From one to many: Expanding the Saccharomyces cerevisiae reference genome panel

Stacia Engel, Shuai Weng, Gail Binkley, Kelley Paskov, Giltae Song and Mike Cherry

In recent years, thousands of Saccharomyces cerevisiae genomes have been sequenced to varying degrees of completion. The Saccharomyces Genome Database (SGD) has long been the keeper of the original eukaryotic reference genome sequence, which was derived primarily from S. cerevisiae strain S288C. Because new technologies are pushing S. cerevisiae annotation past the limits of any system based exclusively on a single reference sequence, SGD is actively working to expand the original S. cerevisiae systematic reference sequence from a single genome to a multi-genome reference panel. We first commissioned the sequencing of additional genomes and their automated analysis using the AGAPE pipeline. We will describe our curation strategy to produce manually reviewed high-quality genome annotations in order to elevate 11 of these additional genomes to Reference status.

Discovering Biomedical Semantic Relations in PubMed Queries for Database Curation and Literature Retrieval

Chung-Chi Huang and Zhiyong Lu

Identifying relevant papers from the literature is a common task in biocuration. Most current biomedical literature search systems primarily rely on matching user keywords. Semantic search, on the other hand, seeks to improve search accuracy by understanding the entities and contextual relations in user keywords. However, past research has mostly focused on semantically identifying biological entities (e.g. chemicals, diseases, and genes) with little effort on discovering semantic relations. In this work, we aim to discover biomedical semantic relations in PubMed queries in an automated and unsupervised fashion. Specifically, we focus on extracting and understanding the contextual information (or context patterns) that is used by PubMed users to represent semantic relations between entities such as "CHEMICAL-1 compared to CHEMICAL-2." With the advances in automatic named entity recognition, we first tag entities in PubMed queries and then use tagged entities as knowledge to recognize pattern semantics. More specifically, we transform PubMed queries into context patterns involving participating entities, which are subsequently projected to latent topics via latent semantic analysis (LSA) to avoid the data sparseness and specificity issues. Finally, we mine semantically similar contextual patterns or semantic relations based on LSA topic distributions. Our two separate evaluation experiments of chemical-chemical (CC) and chemical-disease (CD) relations show that the proposed approach significantly outperforms a baseline method, which simply measures pattern semantics by similarity in participating entities. The highest performance achieved by our approach is 0.89 and 0.83 respectively for the CC and CD task when compared against the ground truth in terms of normalized discounted cumulative gain (NDCG), a standard measure of ranking quality. These results suggest that our approach can effectively identify and return related semantic patterns in a ranked order. To assess the potential utility of those top ranked patterns of a given relation in semantic search, we performed a pilot study on 12 frequently sought semantic relations in PubMed (49 chemical-chemical and chemical-disease patterns in total) and observed improved search results based on post-hoc human relevance evaluation. Further investigation in larger tests and in real-world scenarios is warranted.

Centralizing content and distributing labor: a community model for curating the very long tail of microbial genomes.

Timothy Putman, Sebastian Burgstaller-Muehlbacher, Andra Waagmeester, Chunlei Wu, Andrew Su and Benjamin Good

The last 20 years of advancement in DNA sequencing technologies have led to the sequencing of thousands of microbial genomes, creating mountains of genetic data. While our efficiency in generating the data improves almost daily, applying meaningful relationships between the taxonomic and genetic entities on this scale requires a structured and integrative approach. Currently, the knowledge is distributed across a fragmented landscape of resources from government-funded institutions such as NCBI and UniProt to topic-focused databases like the ODB3 database of prokaryotic operons, to the supplemental table of a primary publication. A major drawback to large scale, expert curated databases is the expense of maintaining and extending them over time. No entity apart from a major institution with stable long term funding can consider this, and their scope is limited considering the magnitude of microbial data being generated daily. Wikidata is an, openly editable, semantic web compatible framework for knowledge representation. It is a project of the Wikimedia Foundation and offers knowledge integration capabilities ideally suited to the challenge of representing the exploding body of information about microbial genomics. We are developing a microbial specific data model, based on Wikidata's semantic web compatibility, that represents bacterial species, strains and the gene and gene products that define them. Currently, we have loaded over 28000 gene and protein items for 16 bacterial genomes including two strains of the human pathogenic bacteria Chlamydia trachomatis. We used this subset of data as an example of the empowering utility of this model. In our next phase of development, we will expand by adding another 118 bacterial genomes and their gene and gene products, totaling over 900,000 additional entities. This aggregation of knowledge will be a platform for uniting the efforts of the entire microbial research community, data and domain experts alike.

Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model

Leonore Reiser, Tanya Berardini, Donghui Li, Robert Muller, Emily Strait, Qian Li, Yarik Mezheritsky, Andrey Vetushko and Eva Huala

Databases and data repositories provide essential functions for the research community by integrating, curating, archiving and otherwise packaging data to facilitate discovery and reuse. Despite their importance, funding for maintenance of these resources is increasingly hard to obtain. Fueled by a desire to find long term, sustainable solutions to database funding, staff from the Arabidopsis Information Resource (TAIR), founded the non-profit organization, Phoenix Bioinformatics, using TAIR as a test case for user-based funding. Subscription-based funding has been proposed as an alternative to grant funding but its application has been very limited within the non-profit sector. Our testing of this model indicates that it is a viable option, at least for some databases, and that it is possible to strike a balance that maximizes access while still incentivizing subscriptions. One year after transitioning to subscription support, TAIR is self-sustaining and Phoenix is poised to expand and support additional resources that wish to incorporate user based funding strategies. Websites: www.arabidopsis.org, www.phoenixbioinformatics.org.

Why the world needs phenopacketeers, and how to be one

Melissa Haendel

The health of an individual organism results from complex interplay between its genes and environment. Although great strides have been made in standardizing the representation of genetic information for exchange, there are no comparable standards to represent phenotypes (e.g. patient disease features, variation across biodiversity) or environmental factors that may influence such phenotypic outcomes. Phenotypic features of individual organisms are currently described in diverse places and in diverse formats: publications, databases, health records, registries, clinical trials, museum collections, and even social media. In these contexts, biocuration has been pivotal to obtaining a computable representation, but is still deeply challenged by the lack of standardization, accessibility, persistence, and computability among these contexts. How can we help all phenotype data creators contribute to this biocuration effort when the data is so distributed across so many communities, sources, and scales? How can we track contributions and provide proper attribution? How can we leverage phenotypic data from the model organism or biodiversity communities to help diagnose disease or determine evolutionary relatedness? Biocurators unite in a new community effort to address these challenges.

David meets Goliath: Using curated data to infer functional boundaries in large datasets

Gemma Holliday, Eyal Akiva, Rebecca Davidson and Patricia Babbitt

In the genomic age there are new sequences being catalogued every day, resulting in too much data with no experimentally characterised function. Thus, computational methods must be employed to give investigators the tools to focus their efforts in the most rewarding direction; for example in inferring potential functions for uncharacterised proteins. A useful theoretical framework for function prediction is enzyme superfamilies, i.e. groups of evolutionarily related proteins that share functional and mechanistic aspects. We use the Structure-Function Linkage Database (SFLD) to capture and catalogue these superfamilies. Because of the inherent homology in enzyme superfamilies, we can infer the function of a protein of interest using annotation of multiple other superfamily members. To this end, we integrate sequence similarity and functional annotation using sequence similarity networks - a tool that allows intuitive visualisation and inference of correlations between sequence similarity and functional similarity. The information from the Gene Ontology (GO) plays a crucial role here. It contains both manual and computationally-derived annotations for gene products relating to function, cellular location and biological processes. The manual annotation is often related to experimental evidence which only accounts for around one percent of all the annotations in GO. By bringing together the small quantity of high quality data and the high quantity of lower quality data we can begin to draw functional boundaries: similarity thresholds at which the reliability of functional annotation transfer is high. The delineation of safe inference rules give researchers clues as to what their protein of interest might do. This marriage of data can also help curators identify potential misannotation or where new functions may exist. Here we will explore what makes an enzyme unique and how we can use GO to infer aspects of protein function based on sequence similarity. These can range from identification of misannotation in a predicted function to accurate function prediction for an enzyme of entirely unknown function. Although GO annotation applies to all gene products, we describe here the use of the SFLD as a guide for informed utilisation of annotation transfer based on GO terms for enzymes.

HAMAP - leveraging Swiss-Prot curation for the annotation of uncharacterized proteins

Ivo Pedruzzi, Catherine Rivoire, Andrea H. Auchincloss, Elisabeth Coudert, Guillaume Keller, Patrick Masson, Edouard de Castro, Delphine Baratin, Beatrice A. Cuche, Lydie Bougueleret, Sylvain Poux, Nicole Redaschi, Ioannis Xenarios and Alan Bridge

HAMAP (High-quality Automated and Manual Annotation of Proteins) is a rule-based automatic annotation system for the functional annotation of protein sequences. It consists of a collection of family profiles for determining protein family membership, and their associated annotation rules for attachment of functional annotation to member sequences. Both HAMAP family profiles and annotation rules are created and maintained by experienced curators using experimental data from expertly annotated UniProtKB/Swiss-Prot entries. Part of the UniProt automatic annotation pipeline, HAMAP routinely provides annotation of Swiss-Prot quality for millions of unreviewed protein sequences in UniProtKB/TrEMBL. In addition, HAMAP can be used directly for the annotation of individual protein sequences or complete microbial proteomes via our web interface at http://hamap.expasy.org. Originally developed to support the manual curation of UniProtKB/Swiss-Prot records describing microbial proteins, the scope and content of HAMAP has been continually extended to cover eukaryotic and lately also viral protein families.

MetaNetX/MNXref - reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks

Marco Pagni

MetaNetX is a repository of genome-scale metabolic networks (GSMNs) and biochemical pathways from a number of major resources imported into a common namespace of chemical compounds, reactions, cellular compartments-namely MNXref-and proteins. The MetaNetX.org website (http://www.metanetx.org/) provides access to these integrated data as well as a variety of tools that allow users to import their own GSMNs, map them to the MNXref reconciliation, and manipulate, compare, analyze, simulate (using flux balance analysis) and export the resulting GSMNs. MNXref and MetaNetX are regularly updated and freely available.

The gene repertoire of eukaryotes: a focus on the missing knowledge of current genome annotations

Betina Porcel, Laura Do Souto, Benjamin Noel, Corinne Da Silva, Jean-Marc Aury, Arnaud Couloux, Simone Duprat, Eric Pelletier and Patrick Wincker

Genome projects of the past decades have supplied a gigantic amount of information on the gene content of many organisms. Still, uncertainties in gene functional annotations have seriously hampered the final goal of describing what an organism can do from its genome sequence. The most challenging example corresponds to the large fraction of novel genes discovered in every new genome sequenced to date, with no detectable homologues to known sequences nor close relatives in databases. Interestingly, most of the publications dealing with the analysis of a new genome only describe the functional potential of the "known" fraction of genes, letting aside this large part of the repertoire. These genes, very restricted in their phylogenetic distribution, make up a large proportion of the encoded genes in almost all sequenced eukaryote genome, frequently accounting 10 to 30% of the gene predictions and even more in poorly sampled taxa. Not much is known about their possible biological functions. Their lack of homology indicates that many of these genes appear to code for specific functions, which are probably more linked to the lifestyle of the organism bearing them, than to central functions for basic cellular processes. Indeed, these taxonomically-restricted genes are considered of special interest since expected to be linked to particular life history traits, playing an important role in unravelling the ecological adaptations to specific niches. The fungal kingdom comprises a vast diversity of taxa with diverse ecologies, life cycle strategies, and morphologies from unicellular yeasts, to highly organized multicellular structures like mushrooms. The characterization of the ""genetic signatures"" of certain branches of the tree of life through bioinformatics analyses will be discussed.

Harvesting cancer genome data - Current opportunities and future challenges

Michael Baudis

While the analysis of cancer genomes using high-throughput technologies has generated tens of thousands of oncogenomic profiles, meta-analyses of datasets is greatly inhibited through limited data access, technical fragmentation and a multitude of raw data and annotation formats.For the arrayMap cancer genome resource, our group collects, re-processes and annotates cancer and associated reference data from genomic array experiments, retrieved from public repositories (e.g. NCBI GEO, EBI ArrayExpress) as well as from publication supplements and through direct requests to the primary producers. So far, our resources provide pre-formatted data for more than 50'000 cancer genome profiling experiments, derived from 340 array platforms and representing more than 700 original publications.Here, I will present some aspects of the shifting landscape of cancer genome data production and publication, an overview about the data accessible through our resources, and an outlook into developments especially with regard of work performed in the context of the Global Alliance for Genomics and Heath.

The BIG Data Center: from deposition to integration to translation

Zhang Zhang and On Behalf Of Big Data Center Members

The rapid advancement of high-throughput sequencing technologies has resulted in an unprecedentedly exponential growth in the volumes and types of biological data generated. Considering that China is now becoming a powerhouse in data generation, the need to collect massive biological data and provide easy access to these data in China that further allows efficient integration and translation of big data into big discoveries is greater than ever. Towards this end, the BIG Data Center (BIGD; http://bigd.big.ac.cn) was founded in January 2016 as part of Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, with the aim for big data deposition, integration and ultimately, translation. BIGD features three major focuses: (1) construction of Genome Sequence Archive (GSA, http://gsa.big.ac.cn; a raw sequence data repository in China that is compliant with extant archives in the International Nucleotide Sequence Database Collaboration), (2) incorporation of a research cohort containing Chinese population genome data and health data (e.g., Virtual Chinese Genome Database;

http://vcg.cbi.ac.cn), and (3) integration of diverse omics data for economically important species in China (e.g., Information Commons for Rice; http://ic4r.org). Hence, BIGD will construct and maintain biological databases by big omics-data integration and value-added curation, perform basic research by developing advanced methods to aid translation of big data into big discovery, and provide freely open access to a variety of databases and tools in support of worldwide research activities in both academia and industry.

Compendium of gene-expression data sets to facilitate machine-learning benchmarks

Anna I Guyer and Stephen R Piccolo

High-throughput assays can be used to profile expression levels for thousands of genes at a time. Clinical researchers attempt to use such data to derive biomarkers that can predict biomedical outcomes, such as disease development, prognosis time, and treatment responses. Due to the large number of predictor variables and dependencies that often exist among variables, traditional statistical models are often unsuitable for such data. However, the computer-science community has developed hundreds of general-purpose algorithms that are designed to be applied to such large and complex data sets. A few of these "machine-learning" algorithms have been applied broadly within the biomedical research community. However, a researcher's choice of algorithm may suffer from either or both of the following limitations: 1) the researcher may arbitrarily choose an algorithm that is not optimally suited for the data to which it is applied, or worse, 2) the researcher may apply a large number of algorithms and risk over-optimization, resulting in a choice that may work well on that particular data set but that does not generalize well to other relevant data sets.To help address this problem, we have compiled a collection of gene-expression data sets from the public domain that other researchers have used to develop and test biomarkers. For each data set, we have curated annotation files that indicate the biomedical outcomes, clinical covariates, and batch information (when available) associated with each patient. To ensure consistency across the data sets, we have renormalized the raw data using the same normalization method (Single Channel Array Normalization) for each data set. Each data file is represented in a consistent, tabular data format (Wickham, 2014) and is stored in the public domain so that other researchers may use these data more easily for their own purposes (https://figshare.com/authors/Stephen_Piccolo/505953).Using these data, we performed a systematic benchmark comparison across 24 classification algorithms that span the major types of algorithm that have been developed by the machine-learning community. We also applied 8 different types of "ensemble" method that combined evidence across all 24 algorithms. Using Monte Carlo cross validation (100 iterations), we found that the Support Vector Machines algorithm (Vapnik, 1998) considerably outperformed all other algorithms; this was true for all three variants of this algorithm (linear, polynomial, and radial basis function kernels) that we tested. The ensemble-based methods and the Random Forests algorithm (Breiman, 2001) also performed consistently well. These findings are consistent with prior studies (Statnikov, 2008). However, we also observed substantial variability across the data sets-the overall top-performing algorithms performed quite poorly on some data sets. This finding confirms the importance of using evidence-based methods to inform algorithm selection. As others have stated previously (Wolpert and Macready, 1997), no single algorithm is optimal in every condition. We are currently evaluating these findings to better understand the conditions under which a given algorithm performs best and will use this information to help researchers match algorithms to their data, given specific characteristics of their data.

Mining Large Scale Proteomics LC-MS/MS Data for Protein Modifications

Markus Muller, Oliver Horlacher and Frederique Lisacek

The functional annotation of post translational modifications (PTMs) lags behind the discovery of new modification sites. This gap in annotations of PTMs could be narrowed by mining suitable existing proteomic MS/MS data, since the presence or absence of a PTM could be correlated to the biological condition of the samples. MS/MS database search results presented in the original publications most often include only a very limited set of PTMs and ignore the large number of other PTMs also detectable in these data. Researching data stored in publicly accessible repositories using an unrestricted set of modifications (open modification search OMS) can reveal the presence of unexpected PTMs or known modifications not considered in the original search. Here we present a workflow to search large proteomic MS/MS datasets for a unrestricted set of modifications. The workflow consists of a combination of standard MS/MS search tools and the in-house OMS tool MzMod [1], which finds PTMs within a predefined mass range at a controlled error rate. MzMod implements a spectrum library approach and guarantees a high accuracy, which is important in error prone OMS strategies. MzMod uses the Apache Spark framework, which facilitates the implementation of parallel code, and is able to analyse very large datasets of 10s of millions of spectra within a day on a medium sized cluster. We present the application of this workflow to two different settings: first we searched MS/MS data of human kidney tissues, which had been stored by formalin fixation and paraffin embedding (FFPE). It was suspected that the FFPE storage and subsequent sample processing induce chemical modifications on proteins, but the identity and extent of these modifications in real medical samples was unknown. Our workflow followed by statistical analysis revealed methylation on lysine as the most abundant modification induced by FFPE [2] - a result which could be confirmed by different studies of independent groups. In a second project we researched a dataset of 25 million MS/MS spectra from 30 human tissues for PTMs, which allowed us to investigate the tissue profiles of the detected PTMs. A preliminary analysis based on spectral counting revealed interesting global properties of PTMs. Succinylation for example was strongly present only in brain, liver and heart tissues. Phophorylation seemed to be more abundant in brain and immune system cells. In this talk we would like to present more detailed results on the tissue specificity of PTMs. We use label free MS1 quantification to find which pathways are differently expressed in different tissues and correlate this information with quantitative information of PTM sites on proteins. [1] Horlacher O, Lisacek F, M"uller M (2016) Mining Large Scale Tandem Mass Spectrometry Data for Protein Modifications Using Spectral Libraries. Journal Proteome Research, doi: 10.1021/acs.jproteome.5b00877[2] Zhang Y, M"uller M, Xu B, et al. (2015) Unrestricted modification search reveals lysine methylation as major modification induced by tissue formalin fixation and paraffin-embedding, Proteomics, doi: 10.1002/pmic.201400454