

CRANFIELD UNIVERSITY

ELEANOR JULIET STANLEY

IDENTIFICATION, ORGANISATION AND VISUALISATION OF COMPLETE PROTEOMES IN UNIPROT  
THROUGHOUT ALL TAXONOMIC RANKS: ARCHAEA, BACTERIA, EUKARYOTE AND VIRUS.

CRANFIELD HEALTH

MSc by Research THESIS

CRANFIELD UNIVERSITY

CRANFIELD HEALTH

MSc by Research THESIS

Academic Year 2011-2012

Eleanor Juliet Stanley

Identification, organisation and visualisation of complete proteomes in UniProt throughout all taxonomic ranks: archaea, bacteria, eukaryote and virus.

Supervisors: Dr. Fady Morareb  
Dr. Maria Jesus Martin

April 2012

© Cranfield University, 2012. All rights reserved. No part of this publication may be reproduced without the written permission of the copyright owner.

## Abstract

Users of uniprot.org want to be able to query, retrieve and download proteome sets for an organism of their choice. They expect the data to be easily accessed, complete and up to date based on current available knowledge. UniProt release 2012\_01 (25<sup>th</sup> Jan 2012) contains the proteomes of 2,923 organisms; 50% of which are bacteria, 38% viruses, 8% eukaryota and 4% archaea. Note that the term 'organism' is used in a broad sense to include subspecies, strains and isolates. Each completely sequenced organism is processed as an independent organism, hence the availability of 38 strain-specific proteomes *Escherichia coli* that are accessible for download.

There is a project within UniProt dedicated to the mammoth task of maintaining the “Proteomes database”. This active resource is essential for UniProt to continually provide high quality proteome sets to the users. Accurate identification and incorporation of new, publically available, proteomes as well as the maintenance of existing proteomes permits sustained growth of the proteomes project. This is a huge, complicated and vital task accomplished by the activities of both curators and programmers.

This thesis explains the data input and output of the proteomes database: the flow of genome project data from the nucleotide database into the proteomes database, then from each genome how a proteome is identified, augmented and made visible to uniprot.org users. Along this journey of discovery many issues arose, puzzles concerning data gathering, data integrity and also data visualisation. All were resolved and the outcome is a well-documented, actively maintained database that strives to provide optimal proteome information to its users.

## **Acknowledgements**

I would like to thank my previous project supervisor, Dr. Lee Larcombe, for his invaluable guidance and encouragement during the development of this thesis. Thanks must go to Daniel Barrell and Benoît Bely for their passion and knowledge of Perl they shared at times of need. Also, much affection to Finlay Aitchison, a truly amazing young man who, due to the constant love and care of his dedicated parents, reached the fine age of four in April 2012, that is four amazing years living without a pancreas.

## List of Contents

Abstract.....	i
Acknowledgements.....	ii
List of Contents .....	iii
List of Figures .....	iv
1. INTRODUCTION.....	1
1.1 European Bioinformatics Institute, EMBL-EBI, hosts UniProt.....	1
1.2 Nuclear, organellar and virus genomes .....	7
1.3 Sequencing complete genomes .....	9
1.4 DNA Variation .....	15
1.5 Transcriptomes, proteomes and integration of the data .....	17
1.6 History of Proteomes database .....	24
1.7 Aims and objectives .....	25
2. METHODS.....	27
2.1 Input of data to proteomesDB.....	27
2.2 Input of data to UniProtKB.....	33
3. RESULTS.....	37
3.1 Proteome editor.....	37
3.2 Reference and representative proteomes.....	42
3.3 Further improvements to proteomes database and UniProtKB entries .....	43
3.4 Quality assurance and error reporting.....	46
3.5 Complete proteome representation at uniprot.org .....	50
4. DISCUSSION.....	53
4.1 Proteogenomics .....	53
4.2 Homology: orthologs and paralogs.....	54
4.3 The genetics of disease and personal genomics.....	63
4.4 EMBL-EBI and ELIXIR .....	67
5. CONCLUSIONS .....	73
5.1 Database improvements .....	73
5.2 New data inputs to proteomesDB and UniProtKB.....	73
REFERENCES .....	77
APPENDIX.....	90

## List of Figures

Figure 1. The diversity of EBI databases, taken from Brooksbank <i>et al.</i> , 2010.....	3
Figure 2. The Tree of Life, taken from David Hillis, Derrick Zwickl and Robin Gutell, University of Texas (Image is freely available for non-commercial educational use). It is based on analysis of small sub-unit rRNA sequences sampled from about 3,000 species from throughout the Tree of Life. ....	6
Figure 3. The cell nucleus contains chromosomes that are made up of DNA that is present as a double-stranded molecule called a double helix, taken from The New Genetics. NIH Publication No. 10-662, 2010. ....	8
Figure 4. Growth in complete genomes, taken from Karsch-Mizrachi <i>et al.</i> , 2012. The layered chart shows the number of new species with genomes entered into INSDC databases over time by taxonomic group. The 2011 time point includes data released in the first 9 months. ....	13
Figure 5. Taxonomic coverage, growth in number of taxa with sequence data, taken from Cochrane <i>et al.</i> , 2011. ....	15
Figure 6. Anatomogram for WNT1 available from the Gene Expression Atlas (GXA), taken from Kapushesky <i>et al.</i> , 2012. ....	18
Figure 7. Snapshot of the UniProtKB complete proteome query page, <a href="http://www.uniprot.org/taxonomy/complete-proteomes">http://www.uniprot.org/taxonomy/complete-proteomes</a> . Accessed Feb 2012. ....	21
Figure 8. Expansion of the hierarchy view of the UniProtKB Complete proteomes, <a href="http://www.uniprot.org/taxonomy/?query=complete:yes&amp;by=parent#131567,2157,2759">http://www.uniprot.org/taxonomy/?query=complete:yes&amp;by=parent#131567,2157,2759</a> . Accessed Feb 2012. ....	22
Figure 9. Gene view of human gene SLC24A5 (ENSG00000188467), <a href="http://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000188467">http://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000188467</a> . Accessed Feb 2012....	23
Figure 10. The module that checks each INSDC genome entry to identify if it defines a new proteome or if it is an update to an existing proteome. Viral proteomes have specific rules due to the multitude of the genomes available. ....	31
Figure 11. Description of the genome assembly and annotation for Rat on the Ensembl website, <a href="http://www.ensembl.org/Rattus_norvegicus/Info/Index">http://www.ensembl.org/Rattus_norvegicus/Info/Index</a> . Accessed Feb 2012. ....	33
Figure 12. A UniProtKB entry augmented with proteome annotation. Proteome annotation is highlighted in yellow. ....	35
Figure 13. System architecture diagram of the platforms involved in the redesign of the Proteome editor. ....	38
Figure 14. Top page to view all properties of the fungus <i>Kluyveromyces lactis</i> proteome.....	39
Figure 15. Taxonomy of the fungus <i>Kluyveromyces lactis</i> . ....	39
Figure 16. The genome of the fungus <i>Kluyveromyces lactis</i> . ....	40
Figure 17. The publications associated with the genome of the fungus <i>Kluyveromyces lactis</i> . ....	41
Figure 18. Data flow diagram of the proteome post processing script. ....	48
Figure 19. Areas of research of scientists attending the two UniProt usability workshops. ....	51
Figure 20. The yellow emperor (ymp, residing at 96E on the right arm of chromosome 3) and Alcohol dehydrogenase (Adh, residing at 35B on the left arm of chromosome 2) genes from <i>Drosophila melanogaster</i> are fused during a speciation event to generate the chimerical structure of <i>Drosophila</i>	

<i>teissieri</i> jingwei (jgw) gene, taken from Llopart <i>et al.</i> , 2002. Boxes symbolize exons and the lines between exons represent introns. Exon 2 and 3 are fused in the intron-absent copy of jingwei (jgw) because of a polymorphic genomic deletion. ....	56
Figure 21. The definition of orthologs, inparalogs and outparalogs, taken from Sonnhammer <i>et al.</i> , 2002. ....	57
Figure 22. G-butyrobetaine hydroxylase inparalogs, taken from Sonnhammer <i>et al.</i> 2002. The points of speciation and duplication are colour coded so easily identifiable.....	58
Figure 23. Ensembl Compara gene view provides displays for data associated at the gene level such as orthologs, paralogs, regulatory regions and splice variants. Gene tree displayed is for the frataxin (FXN), <a href="http://www.ensembl.org/Homo_sapiens/Gene/Compara_Tree?g=ENSG0000016506">http://www.ensembl.org/Homo_sapiens/Gene/Compara_Tree?g=ENSG0000016506</a> . Access Feb 2012. ....	59
Figure 24. Understanding how genes and their products impact the health of humans and other species requires interdisciplinary approaches that incorporate a broad spectrum of demanding technologies and resources. ELIXIRs data integration will allow the knowledge generated to be transformed into technical and industrial developments. Image is freely available with EMBL copy write. ....	68
Figure 25. Genomic positions of the GATA6 mutations on the exon/intron gene structure and the resulting sequence modifications for the resulting protein, taken from Allen <i>et al.</i> , 2011. The recognised mutations include; missense mutations at highly conserved residues within the DNA binding domain, mutations at canonical splice sites reducing the strength of these, an insertion causing premature termination of translation and frameshift mutations. ....	70
Figure 26. Clinical characteristics of the pancreatic agenesis cohort, taken from Allen <i>et al.</i> , 2011. ....	71

# 1. INTRODUCTION

Until the year 2000 only four eukaryotic genomes (yeast, fly, worm and *Arabidopsis*) were sequenced, as well as a few dozen bacteria. At this time the International Human Genome Sequencing consortium (IHGSC) estimated there were about 30,000 - 35,000 protein-coding genes in the human genome with a few classical non-coding RNAs and repetitive DNA elements that were thought to be parasites and junk (Lander *et al.*, 2001).

Over 10 years later, the public sequence databases hold in excess of 140 eukaryotic genomes, 4,400 bacteria, archaea and viruses, metagenomic projects and many hundreds of human genomes. From studying patterns of evolutionary conservation in vertebrates, the human gene count is currently thought to be nearer 21,000 ([http://www.ensembl.org/Homo\\_sapiens/Info/](http://www.ensembl.org/Homo_sapiens/Info/), accessed Feb 2012), much less than originally thought, and the non-coding DNA is known to be the key to understanding evolution.

Next-generation sequencing technologies have allowed the cost of sequencing to fall 100,000 fold in past decade, vastly faster than Moore's Law. With this high rate of innovation in sequencing, predicting the future becomes very difficult. Will this huge amount of sequence data provide a more complete picture of how a genome shapes an organism? From the genotype we hope to explain the phenotype; if an individual's phenotype is altered, can the sequence data be used to aid diagnosis and advance clinical medicine? As common variants do not yet completely explain complex disease genetics, there is no doubt that less common and rare alleles must also contribute. Next-generation platforms would allow the resequencing of many 'normal' human genomes to better capture the spectrum of variability and to establish an important baseline for complex disease studies.

Viruses are not only the most abundant biological entities on the planet (Suttle, 2007); they are also the most represented taxonomic group in UniProtKB. The HIV genome encodes about 9 proteins, but for HIV-1 virus alone there are approximately 350,000 UniProtKB entries corresponding to the equivalent of 35,000 complete genomes! While this may reflect the tremendous sequence diversity of viruses it also makes it impossible for a user to navigate the database looking for their protein of interest. There are many other species that are of huge economic importance for which UniProt needs to provide data to assist scientific research. In an attempt to resolve these dominant issues UniProt has the Complete proteome project.

## 1.1 European Bioinformatics Institute, EMBL-EBI, hosts UniProt

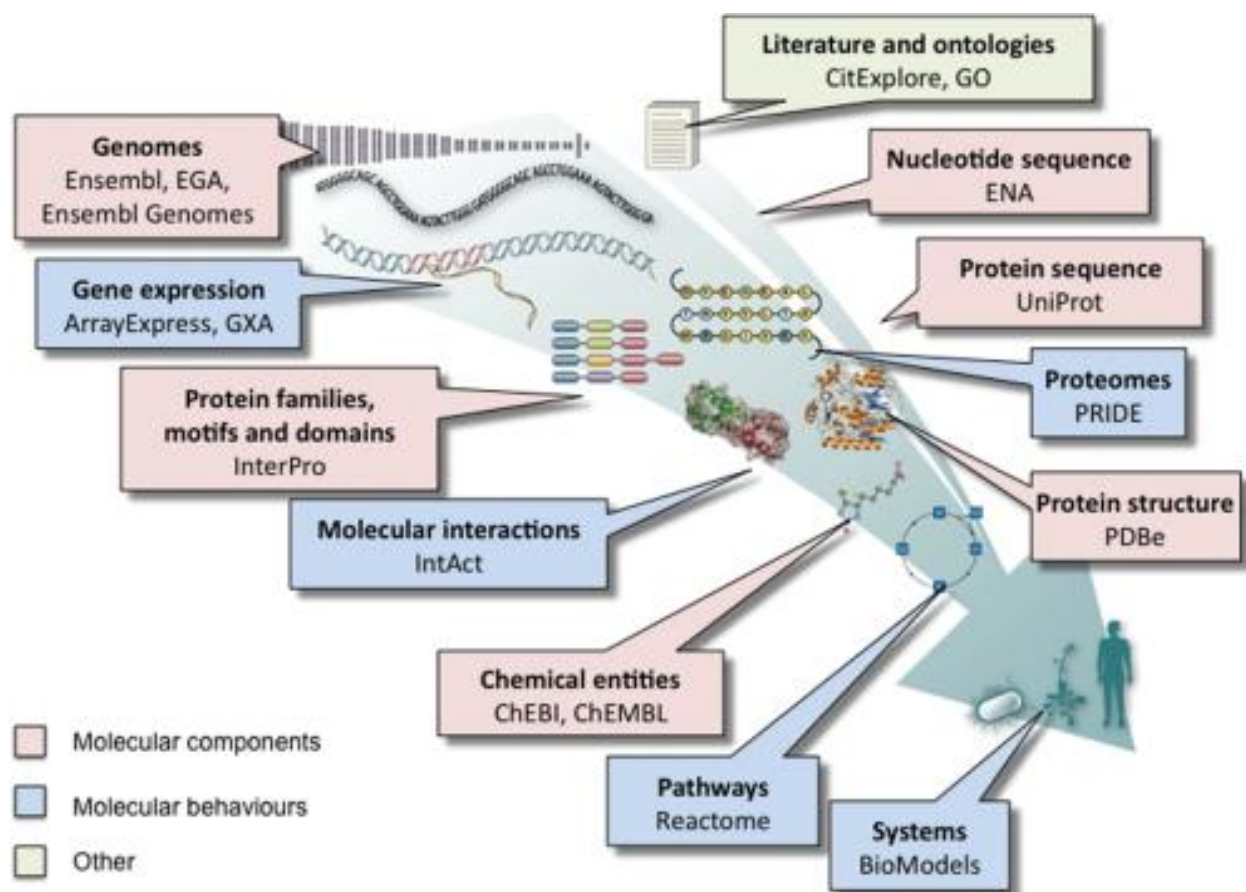
The EMBL Nucleotide Sequence Data Library (now known as the European Nucleotide Archive, ENA, Amid *et al.*, 2012) was established in 1980 at the European Molecular Biological Laboratories (EMBL) in Heidelberg, Germany. EMBL's pioneering work to provide public biological databases to the research community started with the inception of the EMBL Data Library, the world's first nucleotide sequence database whose original goal was to establish a central computer database of DNA sequences. The tasks of the data library grew in scale with the start of the genome projects, and the library grew in visibility as



the data became relevant to research in the commercial sector. To ensure its long-term viability, the EMBL Data Library needed better financial security and collaboration with global partners to both support the project and provide assistance to industry. To achieve this, the EMBL Council voted in 1992 to establish the European Bioinformatics Institute (EMBL-EBI) and to locate it at the Wellcome Trust Genome Campus in Hinxton, United Kingdom where it would be in close proximity to the major sequencing efforts at the Wellcome Trust Sanger Institute (WTSI). From 1992 through to 1995, a gradual transition of the activities from EMBL Heidelberg to EMBL-EBI took place, until in September 1995 the EMBL-EBI opened.

Since 1995, the Wellcome Trust Genome Campus has hosted the EMBL-EBI and the WTSI making it one of the world's largest concentrations of expertise in genomics and bioinformatics. The EMBL-EBI is a non-profit academic organisation that forms an outstation of the European Molecular Biology Laboratory (EMBL) (Brooksbank *et al.*, 2010). It is an integral outstation playing a vital role in achieving EMBL's mission; providing a top-quality research environment that also develops new technologies, and providing services and training to Europe's molecular life scientists. Like the other EMBL sites, the EMBL-EBI has an extremely cosmopolitan staff base, and alumni who have moved on to successful careers all over the world.

In 1995, the EMBL-EBI hosted two databases, one for nucleotide sequences (ENA, Cochrane *et al.*, 2011) and one for protein sequences (Swiss-Prot and TrEMBL, now known as UniProt, The Universal Protein resource, UniProt Consortium, 2012). Since this time the EMBL-EBI has diversified, the number and scope of database has grown hugely to host some of the world's most important collections of biological data, including, complete genomes (Ensembl, Flicek *et al.*, 2012), three-dimensional structures (PDBe, Velankar *et al.*, 2012, European resource for the worldwide Protein databank, wwwPDB, PDBe was formerly known as the Macromolecular Structure Database), data from gene expression experiments (ArrayExpress, Parkinson *et al.*, 2011), protein-protein interactions (IntAct, Kerrien *et al.*, 2012) and pathway information (Reactome, Croft *et al.*, 2011) (Fig. 1). As a consequence the EMBL-EBI has become one of the few places in the world that has the resources and expertise to begin understanding biology at the systems level.



**Figure 1. The diversity of EBI databases, taken from Brooksbank *et al.*, 2010.**

Technologies such as genome-sequencing, microarrays, proteomics and structural genomics provide the platform to understand the requirements for many living organisms, and researchers can focus on how individual components fit together to build an entire system. Europe has always been at the forefront of bioinformatics research, but as we move towards the European Union's goal of a single European Research Area, there is a greater need than ever for bioinformatics experts and experimental biologists throughout Europe to work together towards common goals. The hope is that scientists will be able to translate their insights into improving the quality of life for everyone. However, the high-throughput revolution also threatens to drown us in data. There is an on-going and growing need to collect, store and curate all this information in ways that allow its efficient retrieval and exploitation.

In response to these challenges, EMBL-EBI is coordinating the preparatory phase of the pan-European ELIXIR project (<http://www.elixir-europe.org/>) whose missions' statement reads: "The purpose of ELIXIR is to build a plan for a sustainable infrastructure for biological information in Europe. This plan focuses on generating stable funding for Europe's most important publicly accessible databases of molecular biological information, and the development of a compute infrastructure that can cope with the biological data deluge." The ELIXIR infrastructure will be critical for improving coordination of life science research across Europe. One of the main aims is to link biomedical and biological data resources

to facilitate the understanding of old age diseases in a hope to drive earlier diagnosis, preventive strategies and improve disease management. This improved drug discovery plan will require communication to both pharma and biotech industries to facilitate a pre-competitive collaboration. Another aim is to provide information on plant genomes, insect pests and plant pathogens to enable researchers to develop healthier, more productive crops that could meet the food requirements of a rapidly growing population. And finally, the project also wants to support environmental scientists by improving the way we monitor life in the oceans, understand the effects of climate change on species diversity and develop new methods to tackle pollution and waste. This can be achieved and maintained only with sustainable funding that ensures the longevity of these public resources and support Europe's researchers in their use of these biological data resources.

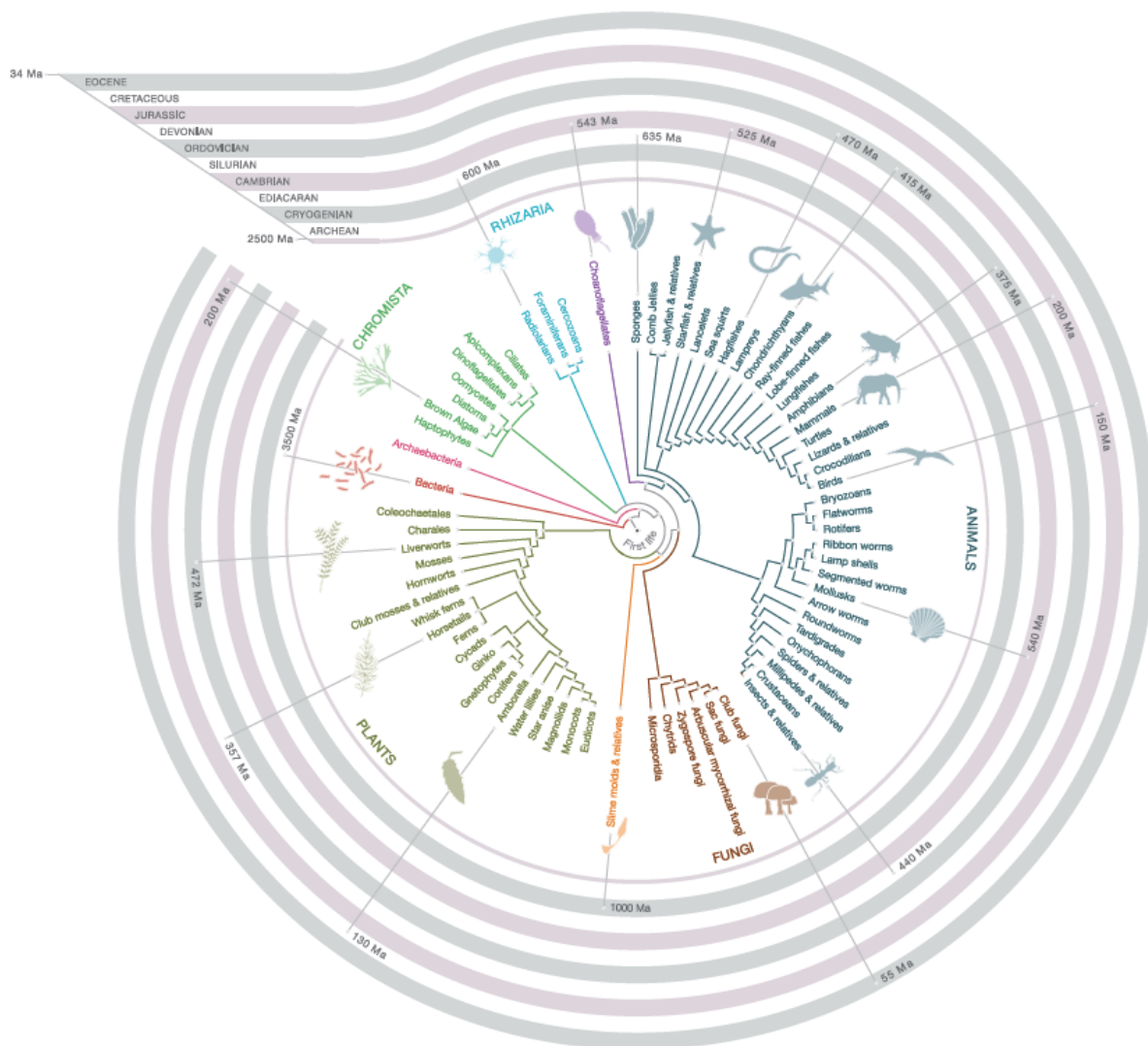
9<sup>th</sup> February 2011 saw an announcement that funding has been earmarked from the UK's Large Facilities Capital Fund for ELIXIR. This will allow the construction of ELIXIR's central hub at EMBL-EBI ensuring the maintenance and expansion of essential biological data resources to support bioscience researchers working in many disciplines. ELIXIR will make important information freely available to researchers across academia and industry through a network of nodes distributed throughout Europe and coordinated at EMBL-EBI.

EMBL-EBI hosts nearly 30 databases, one of the largest being the Universal Protein Resource, UniProt. UniProt provides a centralised repository and comprehensive catalog of protein sequences and their functional annotation maintained by the UniProt consortium (UniProt consortium, 2012). The [www.uniprot.org](http://www.uniprot.org) website (Jain *et al.*, 2009) is freely accessible to the scientific community and is the primary access point to the data and documentation, and to tools such as full text and field-based text search, sequence similarity search, multiple sequence alignment, batch retrieval and database identifier mapping.

The UniProt Consortium is collaboration between the EMBL-EBI, the Swiss Institute of Bioinformatics (SIB) in Geneva and the Protein Information Resource (PIR) at Georgetown University Medical Center. UniProt is comprised of four components. Firstly, the expertly curated UniProt Knowledgebase (UniProtKB) which is the centerpiece of the UniProt Consortium's activities, providing an expertly and richly curated protein database consisting of two sections. UniProtKB/Swiss-Prot contains manually curated information extracted from literature and curator-evaluated computational analysis for well-characterised proteins. It contains a minimal level of redundancy and a high level of integration with other databases. UniProtKB/TrEMBL contains automatically annotated information on protein sequences sourced from the International Nucleotide Sequence Database Collaboration (INSDC, Karsch-Mizrachi *et al.* 2012), Ensembl and protein sequences extracted from the literature or submitted to UniProtKB. Entries are enriched with automated classification and annotation. The INSDC is collaboration between the DNA Databank of Japan (DDBJ) at the National Institute for Genetics in Mishima, Japan; EMBL-EBI; and the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland, USA. Together, the INSDC partner databases (DDBJ (Kodama *et al.*, 2012a), ENA (Amid *et al.*, 2012) and GenBank (Benson *et al.*, 2012)) set out to provide a globally comprehensive collection of public domain nucleotide sequence and associated metadata. Secondly, the UniProt archive (UniParc),

into which new and updated sequences are loaded on a daily basis. UniParc (Leinonen *et al.*, 2009) is a comprehensive repository of protein sequences, providing a mechanism by which the historical association of database records and protein sequences can be tracked. It is non-redundant at the level of sequence identity, but may contain semantic redundancies. Thirdly, the UniProt Reference clusters (UniRef) provide non-redundant reference data collections based on UniProtKB in order to obtain complete coverage of sequence space at several resolutions: 100, 90 and 50% sequence similarity (Suzek *et al.*, 2007). The UniRef clusters are generated in a hierarchical manner; the UniRef100 database combines identical sequences and sub-fragments into a single UniRef entry, UniRef90 is built from UniRef100 clusters and UniRef50 is built from UniRef90 clusters. Each individual member sequence can exist in only one UniRef cluster at each identity level and have only one parent or child cluster at another identity level. UniRef100, UniRef90 and UniRef50 yield database size reductions of ~11, 40 and 72%, respectively. The reduced size of the UniRef90 and UniRef50 datasets provide faster sequence similarity searches and reduce the research bias in similarity searches by providing a more even sampling of sequence space. The UniRef clusters merge closely related sequences based on sequence identity to speed up searches. Fourthly, the UniProt Metagenomic and Environmental Sequences database (UniMES), which is available on the FTP site, was created in response to the expanding area of metagenomic data. Data arising from metagenomic studies is from environmental samples and as such the species may not be known or is unidentified. For this reason, the predicted proteins from UniMES are not included in UniProtKB or UniRef, but they are included in UniParc. The protein entries do have automatic classification by InterPro to enhance the original information with further analysis. UniMES includes data from the Global Ocean Sampling Expedition (GOS) (Yooseph *et al.*, 2007).

All UniProtKB entries specify the organism which is the source of the stored sequence. The organism can be captured at the species level or at the strain, cultivar, and isolate level depending on the information available. These organisms are all present in taxonomic tree structure which represents the taxonomic lineage. The position of each node on a tree is determined by its rank in the taxonomy hierarchy. Higher ranks (phylum, order and family) are placed higher on the tree and the lower ranks (species, sub-species or strain) represent the leaves on the tree's branches (Fig. 2). The ordered list of the nodes forms the lineage. Taxonomy lies at the uneasy interface between biology and logic. The processing of information follows somewhat different rules in these two systems and the role of taxonomy is to reconcile them as tidily as possible.



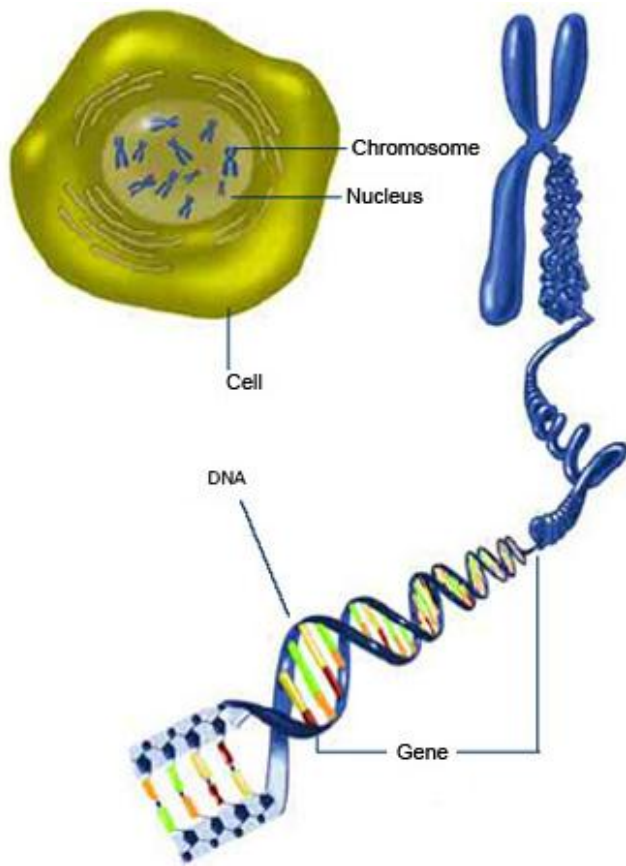
**Figure 2. The Tree of Life, taken from David Hillis, Derrick Zwickl and Robin Gutell, University of Texas (Image is freely available for non-commercial educational use). It is based on analysis of small sub-unit rRNA sequences sampled from about 3,000 species from throughout the Tree of Life.**

The NCBI taxonomy database (Federhen, 2012) is not a primary source for taxonomic or phylogenetic information but instead serves as a central organising principal for the Entrez biological databases and provides links to all data for each taxonomic node, from superkingdoms to subspecies. The database attempts to incorporate phylogenetic and taxonomic knowledge from a variety of sources, including the published literature, web-based databases, and the advice of taxonomy experts. Consequently, the database is growing at the rate of 2,200 new taxa per month and indexes almost 409,000 organisms named at the genus level or lower that are represented in an INSDC entry by at least one nucleotide.

In contrast to cellular organisms, a virus is an inert particle outside its host. The virion on its own has neither metabolism, nor any replication capability, nor autonomous evolution, and therefore cannot be considered a living organism. The Virology Division of the International Union of Microbiological Societies (IUMS) charged the International Committee on Taxonomy of Viruses (ICTV) with the task of developing, refining, and maintaining a universal virus taxonomy (Fauquet *et al.*, 2005). The goal of this undertaking is to categorise the multitude of known viruses into a single classification scheme that is supported by verifiable data and expert consensus and that reflects their evolutionary relationships. This is not a simple task as viruses defined as being in the same family can infect a wide range of hosts, from mammals to insects. Note that the nature of the host does not always appear in the virus name, for example the host of the “Yellow Head Virus” is the shrimp!

## **1.2 Nuclear, organellar and virus genomes**

All cellular organisms possess a nucleus which is the control centre of the cell. The nucleus is a membrane-enclosed organelle found in eukaryotic cells that contains most of the cell's genetic material. This material is organised as multiple long linear DNA molecules in complex with a large variety of proteins, such as histones, to form chromosomes (Fig. 3). The function of the nucleus is to maintain the integrity of these genes and to control the activities of the cell by regulating gene expression. The genes within these chromosomes constitute the cell's nuclear genome.



**Figure 3. The cell nucleus contains chromosomes that are made up of DNA that is present as a double-stranded molecule called a double helix, taken from The New Genetics. NIH Publication No. 10-662, 2010.**

Mitochondria and plastids are membrane-bound organelles, again found in eukaryotic cells, which convert energy from foodstuffs (mitochondria and non-photosynthetic plastids) or sunlight (chloroplasts) into cellular energy. Some plastids may also be used for starch storage and the synthesis of fatty acids and terpenes. Organelles have their own independent genome that encodes a range of genes directly related to producing energy for the cell (Daley and Whelan, 2005).

Eukaryotic nuclear genomes can be distinguished from organelle and prokaryotic genomes by size and complexity. Metazoan, plant, fungal, and other mitochondrial and plastid genomes tend to vary greatly in size and gene content. For example, a typical higher plant nuclear genome contains about  $5 \times 10^9$  base pairs of DNA per haploid set of chromosomes. This is about 30,000 times as much as in a single chloroplast genome and some 10,000 times as much as in a moderately sized plant mitochondrial genome. It is also 1000 times more than that of bacterial DNA present in *Escherichia coli*.

All bacteria are haploid, *i.e.* possess only one chromosome. Most bacterial genomes are organised into a circular chromosomal structure which enables its DNA replication to start and stop at the same location, a feature not seen in the linear genomes of other organisms. Some bacterial genomes contain less than 160,000 base pairs (for example the psyllid symbiont *Carsonella ruddii*, Nakabachi *et al.*, 2006); while

others contain over 9 million base pairs (for example industrial microorganism *Streptomyces avermitilis*, Ikeda et al., 2003). Consequently, the number of genes found in bacteria is highly variable, anywhere between 180 and 7,660 genes. As a comparison, the haploid human genome contains over 3 billion base pairs and approximately 21,000 genes (Levy et al., 2007).

The composition and structure of a virus genome is more varied than any of those seen in the entire bacterial, plant or animal kingdoms. The nucleic acid comprising the genome may be single- or double-stranded, it may be DNA or RNA, and in a linear, circular or segmented configuration. Single-stranded virus genomes may be positive sense, negative sense or ambisense (a mixture of the two). Virus genomes range in size from approximately 3,200 nucleotides (for example Hepadnaviruses) to approximately 1.2Mbp (for example Mimivirus). Many of the DNA viruses of eukaryotes closely resemble their host cells in terms of the biology of their genomes, a requirement as they are obligate intracellular parasites and only able to replicate inside the appropriate host cells.

### 1.3 Sequencing complete genomes

#### 1.3.1 What genomes have been sequenced?

Genome sequences are made publically available via submission to the INSDC (Brunak et al., 2002). With the current lack of a controlled vocabulary, specific feature key or consistent semantics within and between INSDC nucleotide entries, it is very hard to define and identify these complete genomes. Various resources try to list all publically available genomes; each uses specific and different criteria.

EMBL-EBI has a list of complete genomes (<http://www.ebi.ac.uk/genomes>) for all taxonomic ranks. The species included in this list are required to have a genome assembly, and therefore exclude those genomes that are captured within a WGS (whole genome shotgun) project. The NCBI Entrez Genome Project database (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj>) is intended to be a searchable collection of complete and incomplete (in-progress) large-scale sequencing, assembly, annotation, and mapping projects for cellular organisms. It provides an umbrella view of the status of each genome project, links to project data in the other Entrez databases, and links to a variety of other NCBI and external resources associated with a given genome project. Genomes OnLine Database, GOLD, (<http://www.genomesonline.org/>, Pagani et al., 2012) contains information for more than 9,235 genome projects, of which 1,548 are complete and their sequence data deposited in a public repository. Since 1997, GOLD has grown to become a comprehensive repository of metadata information and will evolve into a universal genome project core catalog/indexer charged with the task of providing data interconnectivity, exchange and dissemination.

Sequencing institutes also provide lists of genomes they are actively working on, for example the Broad institute (<http://www.broadinstitute.org/science/projects/projects>) and the Baylor College of Medicine (<http://www.hgsc.bcm.tmc.edu/project-species-x-organisms.hgsc>).



### 1.3.2 Up to the publication of the reference human genome

The first complete genome was sequenced in 1977, Sanger and his colleagues published the 5,386 nucleotides that comprise the single stranded circular genome of *Enterobacteria phage phiX174* (Sanger *et al.*, 1977). Sanger used a sequencing strategy based on random unselected pieces of DNA with cloned restriction enzyme fragments. Nine genes were annotated; two pairs of genes coded by the same region of DNA using different reading frames, and are available as a proteome in UniProtKB (<http://www.uniprot.org/uniprot/?query=+J02482&sort=score>). This same year the first gene in the human genome was sequenced (Seeburg *et al.*, 1977); Chorionic somatomammotropin hormone (<http://www.uniprot.org/uniprot/P01243>). Significantly, three years later, Sanger's group also sequenced the 16,569-base pairs that constitute the complete human mitochondrial genome (Anderson *et al.* 1981). They noted the extreme economy of the organellar genome, in that the genes have none or just a few noncoding bases between them and in many cases the termination codons are not coded in the DNA but are created post-transcriptionally by polyadenylation of the mRNAs. Several further viral and organellar genomes were completely sequenced before the first complete genome of a free living organism was available in 1995. *Haemophilus influenzae* Rd / H175 is a small, non-motile Gram-negative bacterium whose natural host is human. Its genome comprises 1.83 Mb and 1,743 open reading frames (Fleischmann *et al.*, 1995). A shotgun sequencing strategy was used to sequence the whole genome; independent random sequences assembled into a single assembly. The proteome is available in UniProtKB (<http://www.uniprot.org/uniprot/?query=organism:727+keyword:181>). The first eukaryotic genome sequenced was *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996). The combination of a large number of chromosomes, 16, and a small genome size, 14Mb, meant that it was possible to divide sequencing responsibilities conveniently among different international groups involved in the project. The proteome is available in UniProtKB (<http://www.uniprot.org/uniprot/?query=organism:4932+keyword:181>).

In early 1998, PE Biosystems (now Applied Biosystems) developed an automated, high throughput capillary DNA sequencer. Discussions between Applied Biosystems and TIGR scientists resulted in the formation of Celera Genomics, a biotechnology company with the mission to sequence the human haploid genome, the first vertebrate complete genome, and provide clients with early access to the resulting data. A test case for whole-genome assembly, using the capillary sequencer and whole-genome shotgun sequencing techniques, on a large and complex eukaryotic genome was chosen: *Drosophila melanogaster*. In collaboration with Gerald Rubin and the Berkeley Drosophila Genome Project, the nucleotide sequence of the 120Mbp euchromatic portion of the Drosophila genome was determined using the Celera "shotgun" method. The project was completed over a 1-year period providing a genome with 13 fold coverage (Adams *et al.*, 2000).

Armed with the knowledge gained, Celera initially proposed to do ten-fold sequence coverage of the human genome over a three year period and to make interim assembled sequence data available quarterly. DNA from five different individuals was used for sequencing. The lead scientist of Celera Genomics, Craig Venter, later acknowledged (in a public letter to the journal Science) that his DNA was one of 21 samples in the pool from which five were selected. The sequencing plan was changed in

response to the initiation of a public international human genome sequencing effort, the International Human Genome Sequencing Consortium (IHGSC). This consortium represented a collaboration involving 20 groups from the United States, the United Kingdom, Japan, France, Germany and China. Researchers collected blood (female) and sperm (male) samples from a large number of donors and only a few of many collected samples were processed as DNA resources to protect the donor identities. It has been informally reported, and is well known in the genomics community, that much of the DNA for the public Human Genome project (HGP) came from a single anonymous male donor from Buffalo, New York (code name RP11) (Osoegawa *et al.*, 2001).

Acceleration to the Celera's project was required and achieved by performing random shotgun sequencing to 5-fold BAC sequence fragments and subassemblies published by the IHGSC to the INSDC. Quarterly announcements were also abandoned due to the absence of any interim assemblies to report. The time frame from the initiation of sequencing to the completion of the first assembly was 13 months, finishing in October 2000 and published in February 2001 (Venter *et al.*, 2001).

Also in February 2001, the draft sequence of the human genome was made available by IHGSC (Lander *et al.*, 2001). The draft genome sequence was generated from a physical map covering more than 96% of the euchromatic part of the human genome and additional sequences in public databases. The sequence was produced over a relatively short period of fifteen months. The initial genome sequence had approximately 10% of the euchromatic genome missing, around 150,000 gaps and the order and orientation of many segments within local regions had not been established. As the human genome is full of dispersed repeats and large segmental duplications, this made generation of the reference human genome sequence a huge challenge. By 2004, IHGSC improved sequence coverage to 99% and accuracy to an error rate of approximately one event per 100,000 bases. This release of the human genome provided a complete genome of 2.85 billion nucleotides interrupted by only 341 gaps (International Human Genome Sequencing Consortium, 2004).

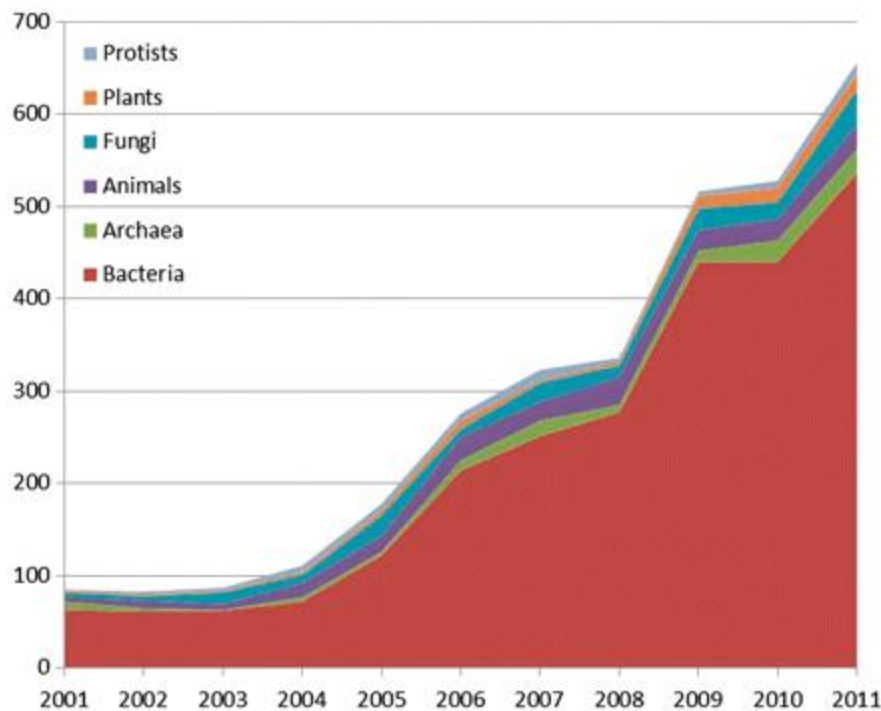
There are a number of regions of the human genome that can be considered unfinished. Firstly, the centromeres, the central region of each chromosome, are millions of base pairs of highly repetitive DNA sequences that are difficult to sequence. Secondly, the telomeres, the 46 chromosome ends, are also highly repetitive, of undefined length, and difficult to sequence. Thirdly, there are several loci in each individual's genome that contain members of multigene families that are difficult to disentangle with shotgun sequencing methods. These multigene families often encode proteins important for immune functions and require focused manual curation to resolve.

Since March 2010, the Genome Reference Consortium (GRC, <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc>) has been releasing genome patches to address some of these issues. These patches are contig or scaffold sequences released outside of the full genome assembly release cycle and constitute part of a minor release. The sequences either correct errors in the assembly (a fix patch) or add additional alternate loci (a novel patch). All patches are rolled into the next major assembly. One of the first patches covers the ABO region, originally this reflected an allelic variant never seen in a human population and now it represents the functional A2 allele.

The human sequence now serves as a foundation for biomedical research, allowing comparative genomics to identify conserved functional features and recognise new innovations in specific lineages. Sequencing additional primates and other organisms will help define key developments along the vertebrate and non-vertebrate lineages. Sequencing costs have dropped 100-fold over the last ten years, corresponding to a roughly twofold decrease every eighteen months. This rate is similar to 'Moore's law' concerning improvements in semiconductor manufacture. In both sequencing and semiconductors, such improvement does not happen automatically, but requires aggressive technological innovation fuelled by major investment. Improvements are constantly being made, the new generation sequencing machines being a perfect example of this.

### **1.3.3 Advancements after the release of the reference human genome**

Since the first days of DNA sequencing performed using the chain termination method developed by Sanger there has been a huge change in the paradigm of DNA sequencing, namely the advent of 'next-generation' technologies. The power of next-generation sequencing lies in the ability to process hundreds of thousands to millions of DNA templates in parallel, resulting in a low running cost per base of generated sequence and a throughput on the gigabase (Gb) scale. Next-generation sequencing technologies such as Roche/454 pyrosequencing, Illumina (Solexa) Genome Analyzer and ABI/SOLiD sequencing have led to previously unimaginable amounts of data being deposited in the public nucleotide sequence databases. The Sequence Read Archive (SRA, Kodama *et al.*, 2012b) of the ENA is a newly established repository for publically available raw data from next generation sequencing platforms. This is the fastest growing part of the ENA. Entries from different data classes are connected through high-level sample and project information. The continued growth in the number of complete genomes is shown in Fig. 4.



**Figure 4. Growth in complete genomes, taken from Karsch-Mizrachi *et al.*, 2012. The layered chart shows the number of new species with genomes entered into INSDC databases over time by taxonomic group. The 2011 time point includes data released in the first 9 months.**

Several large-scale microbial genome sequencing initiatives have been launched using the new technologies. The Human Microbiome Project (<http://www.hmpdacc.org/>, Peterson *et al.*, 2009) studies samples from multiple body sites to investigate the role of changes in the resident human microbiome in disease and health. The project spans 9,949 organisms and the status of each range from “Awaiting DNA” (where an organism has been selected, but the DNA has not yet arrived at the DNA sequencing centre) to “Complete” (where DNA sequencing has been completed). The Genomic Encyclopaedia of Bacteria and Archaea (<http://www.jgi.doe.gov/programs/GEBA/>, Wu *et al.*, 2009) is a collaborative pilot project between JGI (DOE Joint Genome Institute) and DSMZ (German Collection of Microorganisms and Cell Cultures) that aims to sequence 100 bacterial and archaeal genomes based on the phylogenetic positions of organisms in the tree of life. The long-term goal of the GEBA project would be to generate reference genomes for every major and minor group of bacteria and archaea, approximately 5,000 organisms.

The 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010) is an international collaboration that started in 2008 which aims to produce an extensive public catalog of human non-pathogenic genetic variation. This includes SNPs and structural variants, and their haplotype contexts as a foundation for investigating the relationship between genotype and phenotype in each of five major population groups (populations in or with ancestry from Europe, East Asia, South Asia, West Africa and the Americas). This “wild-type” resource will support genome-wide association studies and other

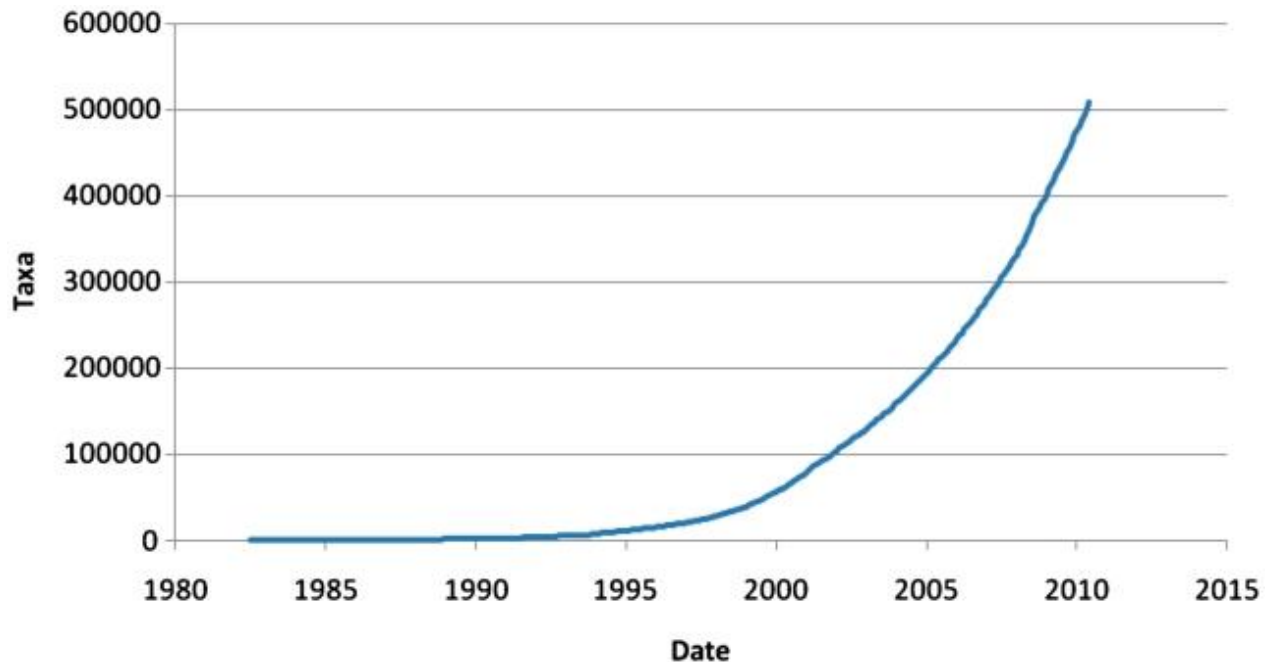
medical research studies as a means to filter all variations to discover those that potentially have pathogenic properties.

To complement the genetic only 1000 genome project, the UK10K project (<http://www.uk10k.org/>), initiated in March 2010, aims to identify low-frequency and rare genetic changes and link their effects to human disease caused by the harmful changes to the proteins the body makes. As not all changes are harmful, the project will first sequence the genomes of 4,000 people whose physical characteristics are well documented. The aim is to identify those changes that have no discernable effect and those that may be linked to a particular disease. This set will be compared to the exomes of 6,000 people with extreme health problems in a hope to find only those changes in DNA that are responsible for the particular health problems observed. The UK10K project will exploit the increasing speed of DNA sequencing and computer processing, coupled with the fall in cost, to conduct a large-scale genome-wide study of thousands of people's DNA sequences to explore rare variants in different types of disease at an unprecedented level. Analysis will begin with obesity and neurodevelopmental disorder cohorts and a further 8 disease areas will be studied.

The Genome 10K Community of Scientists (G10KCOS) propose to create a collection of tissue and DNA specimens for 10, 000 vertebrate species specifically designated for whole-genome sequencing in the very near future (Genome 10K Community of Scientists, 2009). G10KCOS are assembling and allocating a bio specimen collection of some 16,203 representative vertebrate species spanning evolutionary diversity across living mammals, birds, non-avian reptiles, amphibians, and fishes allowing a truly comprehensive study of vertebrate evolution.

The 1001 Genomes Project (Weigel and Mott, 2009), launched at a similar time to the 1000 Genome Project, has a goal to discover the whole-genome sequence variation in 1001 strains of the reference plant *Arabidopsis thaliana*. The study combines large-scale association studies in wild strains with forward genetic analyses in experimental crosses. This will allow identification of alleles underpinning phenotypic diversity across the entire genome and the entire species. Each of the strains in the project is an inbred line with seeds that will be freely available from a stock centre. Unlimited numbers of plants with an identical genotype can be grown and phenotyped for each strain, in as many environments as desired, and so the sequence information can be used directly in association studies at biochemical, metabolic, physiological, morphological, and whole plant-fitness levels.

The rapid growth in number of taxa with available sequence information is illustrated in Fig. 5. These genomes both gap-fill portions of the taxonomy where no genome sequence had been deciphered and generate data for variation in populations of species of particular interest.



**Figure 5. Taxonomic coverage, growth in number of taxa with sequence data, taken from Cochrane *et al.*, 2011.**

## 1.4 DNA Variation

For the first time, an individual human diploid genome was published in September 2007 (Levy *et al.*, 2007). Craig Venter published his six-billion-nucleotide genome. It was sequenced using previous-generation Sanger sequencing technology at 7.5-fold coverage and a cost of \$70 million. The first full diploid genome to be sequenced using next-generation rapid-sequencing technology, Roche/454 pyrosequencing, took just four months, a handful of scientists and less than \$1.5 million. The genome is that of DNA pioneer James Watson (Wheeler *et al.*, 2008) and is provided at 7.4-fold coverage. The achievement is first proof of principle that these rapid-sequencing machines can decipher large, complex genomes. One deliberate omission from Watson's sequence is that of the ApoE4 genotype, associated with Alzheimer's disease, which Watson, born in 1928, asked not to know about. This perfectly illustrates the ethics surrounding knowledge of your personal genome; Alzheimer's disease is incurable and claimed one of his grandmothers.

Both Venter and Watson are of European descent, which leaves gaps in knowledge about how people of different ethnic backgrounds could be susceptible, or alternatively immune, to inherited diseases or respond to medicine. To address this issue, the diploid genome of an anonymous male Han Chinese individual, who has no known genetic diseases, has been sequenced by next-generation Illumina short-read sequencing technology (Wang *et al.*, 2008). The read lengths averaged 35bp, and the two paired-end libraries had a span size of 135bp and 440bp, respectively. Aligning the short reads onto the human reference genome allowed the consensus sequence of the genome to be built with 36 fold coverage, at

a cost of less than \$500,000. Illumina short-read sequencing technology was also used to sequence a member of the Yoruba ethnic group in West Africa (Bentley *et al.*, 2008) with 30-fold coverage. In 2011 Illumina dropped its price to sequence an individual genome to \$19,500.

Geographical isolation and genetic impact on populations can now be studied with the availability of the first Irish genome sequence (Tong *et al.*, 2010), also sequenced using Illumina short-read sequencing technology to 11-fold coverage.

Studies of DNA variation continue in the International HapMap Project, whose goal is to develop a haplotype map (<http://hapmap.ncbi.nlm.nih.gov/>, HapMap) of the human genome, which will describe the common patterns of human genetic variation (Altshuler *et al.*, 2010). The DNA samples for HapMap came from a total of 270 individuals: Yoruba people in Ibadan, Nigeria; Japanese people in Tokyo; Han Chinese in Beijing; and the French Centre d'Etude du Polymorphismes Humain (CEf) resource, which consisted of residents of the United States having ancestry from Western and Northern Europe. HapMap is expected to be a key resource for researchers to find genetic variants affecting health, disease and responses to drugs and environmental factors. The information produced by the project is made freely available to researchers around the world. It is hoped that the study of genomics will help us learn why some people get sick from certain infections, environmental factors, and behaviors, while others do not. Better understanding of the interactions between genes and the environment will help us find better ways to improve health and prevent diseases.

MapSeq/pf is a database of genome variation in the malaria parasite *Plasmodium falciparum* in populations around the world (<http://www.sanger.ac.uk/MapSeq/>). Malaria researchers and collaborators submit the blood of individuals with malaria infection. From this, parasite DNA is extracted and sequenced at the WTSI using 'next-generation' sequencing technologies by Illumina/Solexa. This generates very large numbers of sequence reads of *Plasmodium* DNA which are aligned to a reference genome. A sequencing pipeline analyses the vast quantities of sequencing data generated from the samples, identifying SNPs, indels and other forms of variation across the whole genome, and genotyping every sample at those positions. The resulting genotyping data can be browsed and analysed using MapSeq. The aim of the project is to detect and describe all variations in the *Plasmodium falciparum* genome, and to provide tools that will facilitate their analysis.

Distinguishing the genetic differences between individuals of the same species and linking these genotypic differences to phenotypic differences provides important leads for medical and agricultural research. In July 2008 the Vertebrate Genomics team at the EMBL-EBI launched the European Genome-phenome Archive (EGA), <http://www.ebi.ac.uk/ega>. EGA is a secure repository for all types of potentially identifiable data types including the array-based genotype data from genome-wide association studies, for example DNA sequence arising from re-sequencing, transcriptomics and epigenomics. The EGA stores the raw data from many types of experiments including case control studies, cancer sequencing, and family and population studies. Available data types include single nucleotide polymorphism (SNP) and copy number variation (CNV) genotypes, whole genome sequence and phenotype data. Each data type is stored at the EGA using methods designed to ensure that the storage and distribution is done in accordance with the consent and confidentiality agreements that the research participants agreed to at

the time of entry into the study. This is important as the data stored could potentially make individuals identifiable. Where permitted, some data is publically available. This includes data from Wellcome Trust Case Control Consortium (WTCCC).

## **1.5 Transcriptomes, proteomes and integration of the data**


### **1.5.1 Transcriptomes**

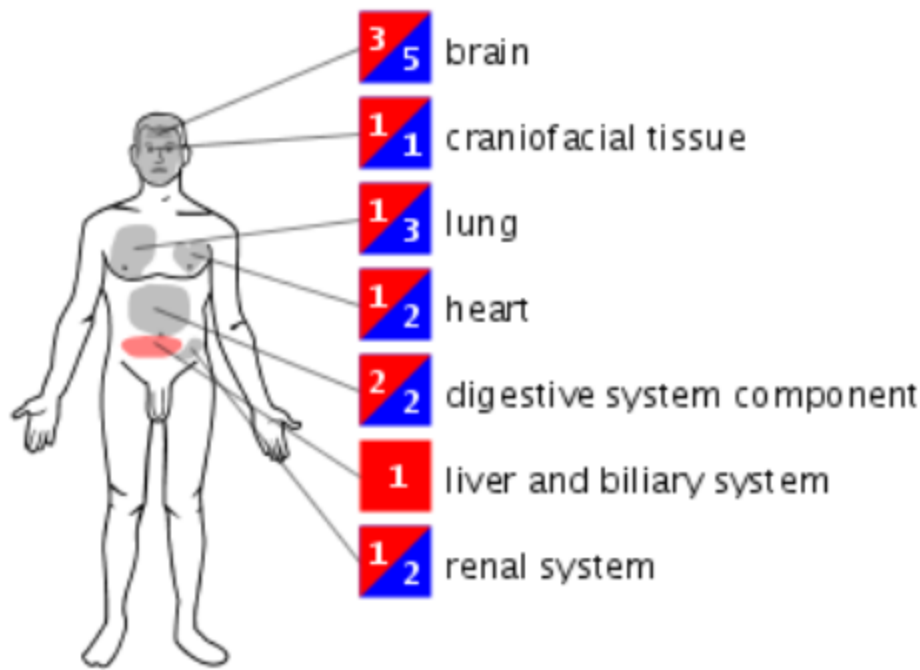
A genome is the entire DNA sequence of an organism. In contrast, the transcriptome represents the small percentage of the genome (less than 5 percent in humans) that is transcribed into RNA. The transcriptome constitutes the RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA, that can be produced in one cell or a population of cells. A gene may produce many different types of mRNA molecules, so a transcriptome is much more complex than the genome that encodes it. Unlike the genome, which is roughly fixed for a given cell line (excluding mutations), the transcriptome can vary with external environmental conditions. Because a transcriptome includes all mRNA transcripts in the cell, it reflects the genes that are being actively expressed at any given time. This “expression profiling” can be used to identify genes involved in certain functions. For example a gene whose expression levels are dramatically higher in cancer cells than in healthy cells might suggest that the unknown gene may play a role in cell growth.

DNA Microarray studies are the most common form of expression profiling. To capture this, and data from other profiling techniques, The ArrayExpress Archive (Parkinson *et al.*, 2011) was launched by the EMBL-EBI in 2002 as the world’s first open-access, standards-compliant repository for high-throughput transcriptomics assays. In compliance with the MIAME initiative (Brazma *et al.*, 2001), most scientific journals now require publication-related microarray gene expression data to be deposited in ArrayExpress or the NCBI’s Gene Expression Omnibus (GEO) (Barrett *et al.*, 2011). Data from near 22,000 studies are available from these archives, but using these data to answer biological questions is not straightforward.

The Gene Expression Atlas (GXA) (Kapushesky *et al.*, 2012), also at EMBL-EBI, simplifies the analysis of gene expression data. It is a tool that allows users to query condition-specific gene expression in different organisms based on multiple independent gene expression studies. Differential expression of a gene of interest can be viewed (Fig. 6) as well as the user discovering which genes are differentially expressed in a particular condition or site of interest. Both types of query can be combined to focus on particular genes and their role in a specific condition; for example, GXA makes it straightforward to search for members of the Wnt signalling pathway that are expressed in colorectal adenocarcinoma.



 Number of published studies where the gene is over/under expressed compared to the gene's overall mean expression level in the study.



**Figure 6. Anatomogram for WNT1 available from the Gene Expression Atlas (GXA), taken from Kapushesky *et al.*, 2012.**

GXA takes a subset of the data from the ArrayExpress Archive, including data imported from GEO and subjects it to rigorous curation. Mapping of genes to the latest genome-builds ensures that each gene in GXA has an unambiguous reference point. Mapping of conditions to a purpose-built ontology, the Experimental Factor Ontology (EFO, Malone *et al.*, 2010), ensures that users retrieve all the results relevant to their query, not just those that exactly match the text of their query.

Analysis of cDNAs by next-generation sequencing provides an accurate picture of active transcriptional patterns in an organism. This RNA sequencing (RNA-Seq) is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies. RNA-Seq provides a far more precise measurement of levels of transcripts and their isoforms than other methods, so these studies demonstrate the true complexity of eukaryotic transcriptomes.

### 1.5.2 Proteomes

The term “proteome” was first introduced in the mid-1990s by Wilkins and Williams to indicate the entire “PROTEin” complement expressed by a “genOME” of a cell, tissue, or entire organism (Wilkins *et*

*al.*, 1996). As with transcriptomes, while there is one definitive finite genome for an organism, this genome defines a proteome that can change between different cell types, different tissue types, different stages of development, different alternative splice variants, different post translational modifications, and the list goes on. A large number of proteomes can be produced during the lifetime of a cell or organism. A vivid visualisation of proteome changes in a single organism and its single genome is the development of an egg, to caterpillar, to nymph and finally an adult butterfly: many different proteomes with strikingly different phenotypes.

For UniProtKB, a complete proteome is defined as the entire set of proteins expressed by a specific organism, the majority of proteome sets are based on the translation of a completely sequenced genome. UniProtKB complete proteomes aims to provide users with as much information as possible for each proteome and its constituent proteins. To identify a complete proteome is problematic due to several considerations that have to be taken into account. Annotation of gene models on the genome is the first challenge. This process relies on the genome being complete and having no or a minimum number of gaps and having the correct assembly. Assuming this criteria is met, the architecture of gene models can be convoluted and must allow for all possibilities in biology which include overlapping/nested genes within and between DNA strands, gene families resulting from tandem duplications or transpositions in the genome, non-consensus splice sites and long introns to name a few.

Secondly, proteome annotation relies on the stability of the underlying genome sequence. The reference genome may undergo continual changes, for example release of a new assembly, modifications to an automated pipeline for gene model annotation, adjustments to parameters used by a gene prediction program or due to manual gene model updates. All updates must be captured and incorporated into the proteome as efficiently as possible.

Thirdly, how do we define a proteome for an organism? One extreme example of an alternative proteome is the Gram-negative bacillus *Bacteroides fragilis*. This microbe is an obligate anaerobe of the human colon that is able to modulate its surface antigenicity by producing at least eight distinct capsular polysaccharides and is able to regulate their expression in an on-off manner by the reversible inversion of DNA segments containing the promoters for their expression. This reversible surface diversity allows the organism to exhibit a wide array of distinct surface polysaccharide combinations, contributing to huge pathogenic potential and ability to maintain an ecological niche in the intestinal tract (Krinos *et al.*, 2001 and Cerdeno-Tarraga *et al.*, 2005).

Fourthly, ideally a comprehensive description of a proteome should include information on such post translational modifications as they frequently indicate the functional state of the protein and also give clues about its cellular location. Looking at the human proteome available in UniProtKB/Swiss-Prot, about 25% of those proteins have not yet been studied experimentally. For the remainder, the information available is often scarce. Many proteins have not been completely analyzed with respect to their abundance, distribution, subcellular localization and interactions with other biomolecules, post-translational modifications or, even more critical, function.

Fifthly, identification of pseudogenes in a genome can be difficult. Improving the identification of pseudogenes can have a huge effect on the quality of the resulting proteome. Extreme numbers of pseudogenes in a genomes include *Rickettsia prowazekii*, where approximately one-quarter of the genome is noncoding (Andersson *et al.*, 1998) and less than one-half of the 3.27-Mb genome of *Mycobacterium leprae* contains functional genes (Cole *et al.*, 2001).

Across the entire taxonomic range, including viruses, UniProtKB provides complete proteome sets for each species where data is publically available from the INSDC nucleotide database. For an organism to be included there are two main requirements; the organism must have a completely sequenced genome (fully closed and exhibiting either good gene prediction models or good quality transcriptome/proteome data) and proteins in the set are mapped to the genome. UniProtKB entries for organisms that meet these criteria are augmented with consistent annotation relevant to the taxonomy, including the strain (where available) and the publication of the genome (literature or sequence submission). Entries are also tagged with the keyword 'Complete proteome' allowing the easy retrieval of the proteome set from the database.

A first draft of the human proteome, comprising 20,325 protein-coding sequences, was released in September 2008. This data set has now been re-annotated to improve the depth and quality of the information provided. New splice variants and polymorphisms have been added to existing records, and records have been created for newly discovered protein sequences. UniProt consortium has joined the Consensus coding sequence (CCDS) project (Pruitt *et al.*, 2009), a collaborative effort including the WTSI, the University of California, Santa Cruz, the US National Center for Biotechnology Information and the EMBL-EBI, to identify a core set of consistently annotated and high-quality human and mouse protein-coding regions. The long-term goal is to support convergence towards a standard set of gene and protein annotations.

The development of new sequencing techniques is generating a flood of genomes to the databases. These ever-growing numbers of genomes often have submission problems that prevent the production of a non-redundant protein set or have problems regarding the gene model predictions. Unfortunately, this includes some important model organisms, such as *Danio rerio* (zebrafish) and *Chlamydomonas reinhardtii*. UniProtKB overcomes this issue by generating the proteome in collaboration with Ensembl (Flicek *et al.*, 2012) and RefSeq (Pruitt *et al.*, 2012), for example mouse, rat, cow, chicken, dog and zebrafish.

Currently, a few species are represented by more than one complete proteome, for example *Escherichia coli* and *Streptococcus agalactiae*, and this number is likely to grow in the near future due to continuing developments in high-throughput sequencing technologies. Historically, UniProtKB merged sequence and annotation data from different strains and complete proteomes into a single entry. Such merging will no longer be performed due to the potentially large differences in protein composition that may exist between different strains of the same species. The final outcome of all required curation policy changes will be the provision of clearly delineated and correctly annotated proteomes for individual strains or isolates in UniProtKB.

Where multiple proteomes cluster within a small taxonomic range, UniProtKB will help users select a proteome by defining a reference or representative proteome for the species with the greatest amount of relevant functional annotation. Key model organisms will be manually selected as reference proteomes and an automatic procedure that assesses annotation content will define the representative proteomes.

In order to give users access to these complete proteomes there is the 'Complete proteomes' web page (Fig. 7).

UniProt > Taxonomy Downloads · Contact · Documentation/Help

Search Blast Align Retrieve ID Mapping

Search in Taxonomy Query \* AND complete:yes Search Advanced Search > Clear

### COMPLETE PROTEOMES AND REFERENCE PROTEOMES

A **complete proteome** consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced.

A **reference proteome** is the complete proteome of a representative, well-studied model organism or an organism of interest for biomedical research.

These organisms can be searched via the taxonomy pages, which provide links to download complete and reference proteome sets when available, as well as links to the HAMAP web site.

Browse or list organisms with:

Complete proteomes	Reference proteomes
<ul style="list-style-type: none"> <li>Browse by hierarchy</li> <li>List all Bacteria</li> <li>List all Archaea</li> <li>List all Eukaryota</li> <li>List all Viruses</li> </ul>	<ul style="list-style-type: none"> <li>Browse by hierarchy</li> <li>List all Bacteria</li> <li>List all Archaea</li> <li>List all Eukaryota</li> <li>List all Viruses</li> </ul>

Search organisms with complete proteomes:

 Search

Search organisms with reference proteomes:

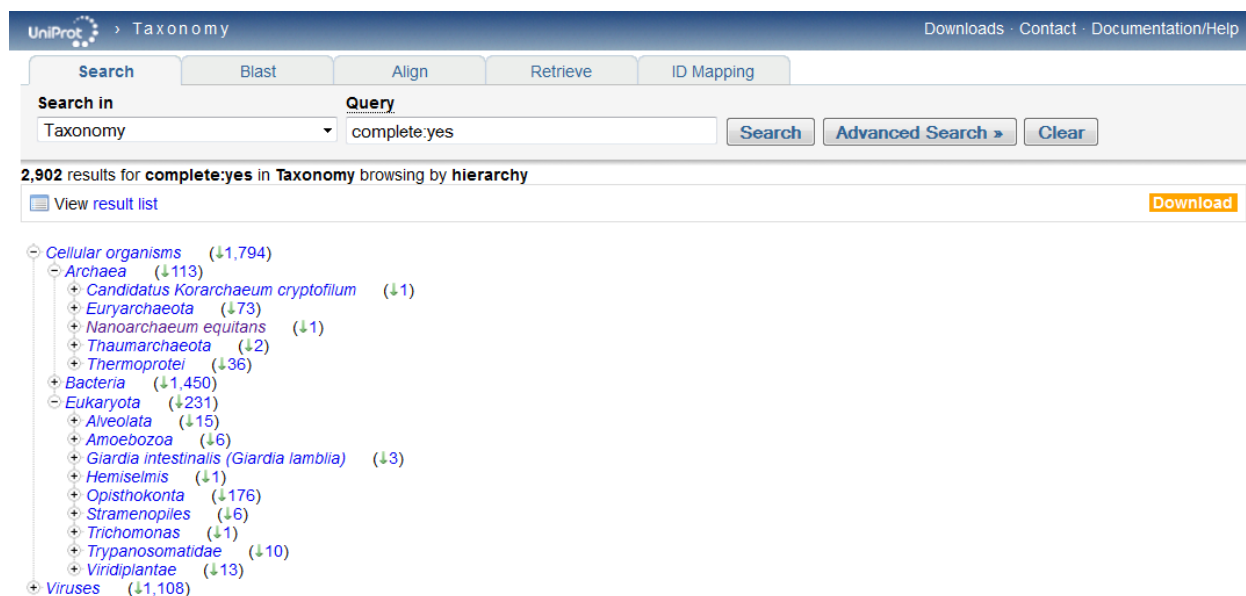
 Search

### FAQ

- > What are complete proteome sets?
- > What are reference proteome sets?
- > How to retrieve sets of protein sequences?
- > What is HAMAP?  
HAMAP is a system, based on manual protein annotation, that identifies and semi-automatically annotates proteins... [More](#)

**Figure 7. Snapshot of the UniProtKB complete proteome query page, <http://www.uniprot.org/taxonomy/complete-proteomes>. Accessed Feb 2012.**

Currently this page allows users to search for and retrieve the proteome of a species of interest. The spread of the species across the taxonomic nodes can be viewed using the hierarchy feature (Fig. 8). Access to data on complete proteomes, including an improved search interface and downloads, will be provided via a new portal which is currently under development. This portal will provide users with information and simple statistics for both complete proteomes and their individual components, such as chromosomes and plasmids.



**Figure 8. Expansion of the hierarchy view of the UniProtKB Complete proteomes, <http://www.uniprot.org/taxonomy/?query=complete:yes&by=parent#131567,2157,2759>. Accessed Feb 2012.**

The Entrez Protein Clusters database ([www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters](http://www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters), Klimke *et al.*, 2009), contains over 492,000 sets of RefSeq proteins (Release Dec 02 2010) encoded by complete prokaryotes, bacteriophages and mitochondrial and chloroplast organelles, organised in a taxonomic hierarchy. Protein Clusters provides annotations, publications, domains, structures, external links and analysis tools, including multiple sequence alignments and phylogenetic trees. ProtClustDB contains both curated and uncurated clusters of proteins grouped by sequence similarity. PubMed identifiers and external cross references are collected for all clusters and provide additional information resources.

### 1.5.3 Integration of all genome, transcriptome, and proteome data

EMBL-EBI, in collaboration with the Wellcome Trust Sanger Institute, developed the Ensembl genome browser in 2001. Ensembl's original purpose was to facilitate navigation and analysis of the human genome, focusing on the annotation of known genes and predicting the location of previously uncharacterised ones. As the number of species has grown Ensembl has expanded its focus to all sequenced chordates owing to their ability to help us understand human biology and evolution. Over the past eleven years, Ensembl's coverage has grown to over 50 genomes. The Ensembl system has been extended to the rest of the taxonomic tree with the arrival of Ensembl Genomes (Kersey *et al.*, 2012). Ensembl Genomes provides a companion service to Ensembl in the form of five new sites: Ensembl Bacteria, Ensembl Protists, Ensembl Fungi, Ensembl Plants and Ensembl Metazoa. The launch of Ensembl Genomes provides a consistent framework for inter-species analyses across the whole of taxonomic space. One of the major goals of Ensembl is to provide gene sets which are as accurate and

complete as possible and these continue to be used as reference gene sets in analysis of new vertebrate genomes.

Querying Ensembl with a gene of interest produces a results page organised into four classes: Location, Gene, Transcript and Variation, which can be easily navigated through tabs at the top of each web page. The location class includes views of the genome sequence at a range of resolutions and genome sequence based comparative views. Gene based views include textual information about the gene, views of its local genomic environment, views of the gene in the context of its orthologs and paralog relationships with other genomes in the Ensembl system and views of sequence variation within that population (Fig. 9). Transcript based views are similar to the gene based ones, but focus around individual transcript structures with more detail. Variation based views display information focused around individual SNPs.

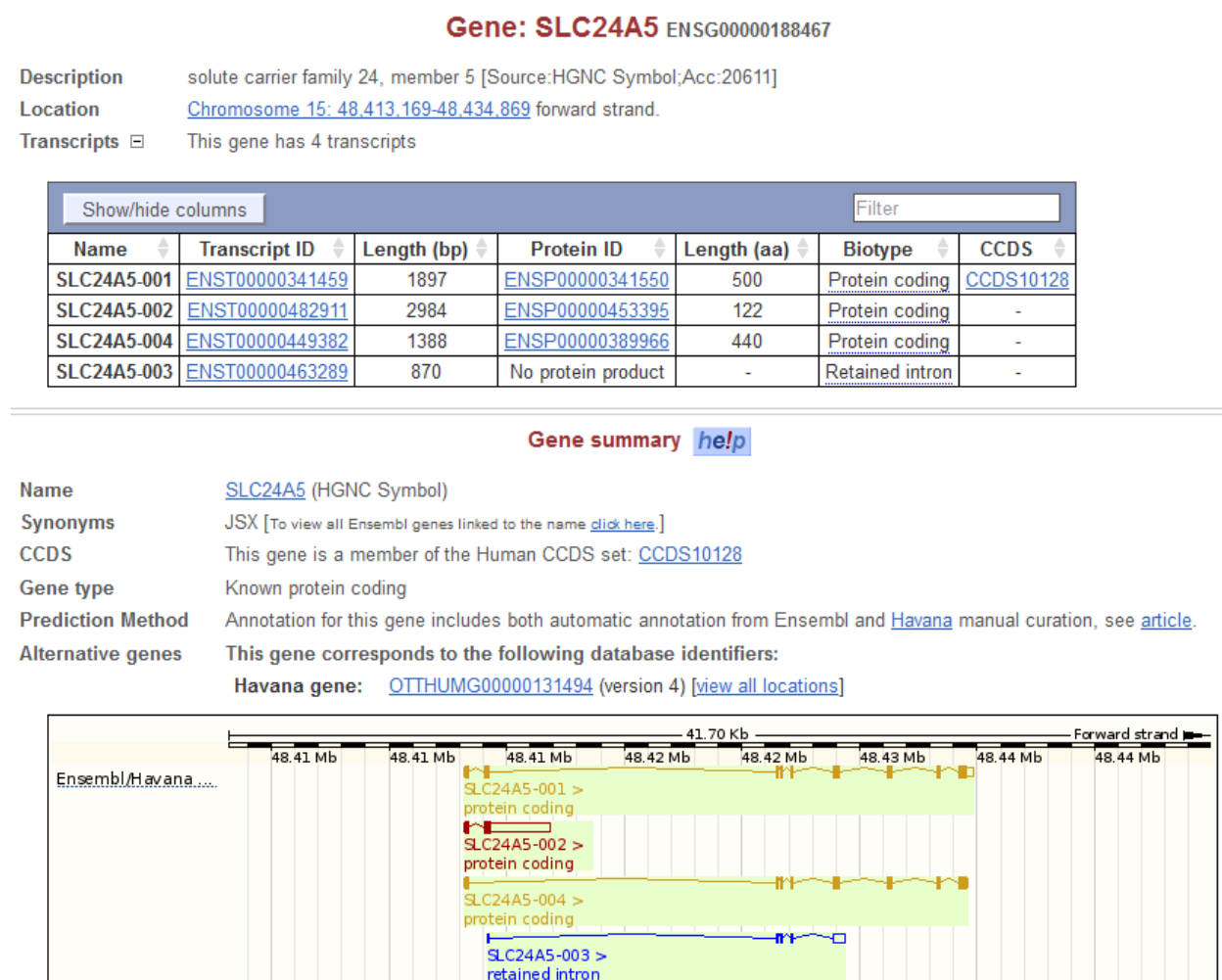


Figure 9. Gene view of human gene SLC24A5 (ENSG00000188467), [http://www.ensembl.org/Homo\\_sapiens/Gene/Summary?g=ENSG0000018846](http://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG0000018846). Accessed Feb 2012.

## 1.6 History of Proteomes database

The Proteomes database (proteomesDB) was initially developed by the Integr8 project as part of the TEMPLOR grant from the European Union RTD program "Quality of Life and Management of Living Resources" and later as part of the FELICS grant within the Research Infrastructure Action of the FP6 "Structuring the European Research Area" program. The outcome was a database replete with tables, over 240 of which less than 10 are core tables, the remainder are vocabulary tables, trigger tables, proteome import related tables and tables to support other databases. Over time the database was used to serve three public EMBL-EBI hosted resources: Integr8 (Kersey *et al.*, 2005), International protein index (IPI, Kersey *et al.*, 2004) and Genome Reviews (Kersey *et al.*, 2005). It is also the source of data for two public SIB hosted databases: HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes, Lima *et al.*, 2009) and ViralZone (Hulo *et al.*, 2011). While these SIB databases continue those at EMBL-EBI are closing in response to the EBI reorganisation of its processing of biomolecular sequence information ([http://www.ebi.ac.uk/Information/News/pdf/EBI\\_reorg\\_070607.pdf](http://www.ebi.ac.uk/Information/News/pdf/EBI_reorg_070607.pdf)) and will be replaced by UniProtKB, Ensembl Genomes (Kersey *et al.*, 2012) and ENA Genomes (a new database to be announced that will cover those species with a complete genome that were in Integr8 and are not within Ensembl Genomes).

In February 2009, with the inception of Ensembl Genomes, UniProtKB took over the maintenance and development of proteomesDB. This was the beginning of a new challenge for UniProtKB: what is proteomesDB, who needs the data, when do they need it, why do they need it, what is the input data, how is the data retrieved, how does the data get processed and what data is made public and, crucially, why is there no documentation for any of these procedures? A lot of questions and issues had to be resolved to enable the current project to blossom, accomplish the deliverables dictated by the National Institutes of Health (NIH) grant 1U41HG006104-01 and grow into the high profile UniProt consortium initiative it must become. To achieve this, I was chosen as project leader to drive this development having strong bioinformatic credentials; a thorough understanding of the biology of proteomes, experience in handling data from large databases and previously demonstrated managerial skills to lead and work with a team of computer scientists. Alongside managing the proteomes development team, I am also a senior UniProtKB and proteome curator of 14 years. My deep knowledge of UniProtKB gives me the insight to accurately and efficiently steer the proteomes project in the right direction to the benefit of all.

One consequence of the change in database dependency for the proteomesDB described above is that the entire pipeline has been rewritten in Java. The direct editing of the code was performed by programmers in consultation with me at all stages to ensure all functionality was retained and any queries concerning data handling could be resolved and, where required, improved upon. Java was chosen as this is the preferred language of the UniProtKB software engineers and this expertise has been used to rewrite all UniProtKB production processes. It is essential that the components of the proteomeDB fit within the UniProtKB Java framework to ensure successful integration of data with each UniProtKB release, as well as the projects durability and longevity. The advantages of writing it in this

framework are that it is modular which in turn allows testing and maintenance, and, importantly, it is extensible. The proteomesDB pipeline has been incorporated into the UniProtKB Java framework, JAPI, which allows programmatic access to UniProtKB (Patent *et al.*, 2008).

The proteomes database (the schema including core tables is provided in Appendix 2) has been designed and maintained to serve several purposes within UniProtKB and external teams, for example Ensembl Genomes who use the database to link a species name to its genome, and therefore its proteome. To achieve this, the database needs to be constantly updated to reflect the current status of the complete genomes. The following section describes how this is performed; the proteomes team is responsible for the perpetual evolution of the database and its documentation so that it reliably delivers complete proteomes to its internal and external users now and for all future release of UniProtKB.

## **1.7 Aims and objectives**

Maintenance of a complete and robust proteomes database for UniProtKB is a huge and constant challenge. The aim of this project is to investigate methods to improve capture of proteome data and implement these within the UniProtKB release production cycle. This will be accomplished by achieving the following objectives.

### **1: Identify new proteomes for incorporation into proteomesDB and maintenance of existing proteomes**

In the absence of a consistent and systematic way to identify a complete genome within INSDC, proteomesDB must develop and maintain a procedure that can identify complete genomes for individual organisms. For each genome, proteomesDB stores the appropriate INSDC accession(s). Any aspect of an INSDC nucleotide entry can change between releases: NCBI taxonomy identifier, accession number, CDS base span or a gain or loss in the number of CDS features. These updates must be correctly identified and propagated to proteomesDB to maintain an up to date proteome. Where a genome exists in INSDC and the gene model annotations are present in another database (for example Ensembl or Ensembl Genomes), proteomesDB must link both resources to correctly identify the proteome. Between Ensembl releases protein translations may be changed, deleted or added as new. Again this must be synchronized with UniProtKB to maintain the proteome.

### **2: Update of UniProtKB entries for proteome data within the UniProtKB release production**

A proteome in proteomesDB links to a subset of UniProtKB entries via the INSDC accession(s). The import of data from proteomesDB to UniProtKB must perform a plethora of annotation updates, including addition and removal of proteome data in UniProtKB entries as appropriate, at an optimal time in release production. Reference and representative complete proteomes are defined within proteomesDB and data propagated to UniProtKB entries.

### **3: Database improvements, quality assurance and error reporting**



While extensive automatic procedures are in place to capture and maintain proteomes manual edits are still required, principally correction to data, addition of the UniProtKB 3-5 letter organism code assigned to all species of interest (which is called the “oscode”) and toggling of proteome status flags. The proteome editor has been developed for this purpose.

This extensive reworking of all pipelines has identified many database improvements required including removal of unnecessary tables, to removing redundant columns, adding new columns and overhaul of the script logs.

The principle QA is to ensure proteomes are maintained between UniProtKB releases, a report must be generated that details a loss or gain of a proteome between the previous and current release, as well as reporting a change in the number of UniProtKB entries per proteome. This must be run as the last step in proteome import so any problems can be addressed before the end of the UniProtKB release production cycle. Further reports are generated to guarantee stability of complete proteomes to our users.

#### **4: Visualisation of proteomes in uniprot.org**

Complete proteomes are made public to users via the UniProt website uniprot.org. UniProt user workshops highlighted the proteome pages were sub-optimal so improvements to the user experience are required.

## 2. METHODS

### 2.1 Input of data to proteomesDB

#### 2.1.1 From INSDC to proteomes database: EMBL import

The complete proteomes within UniProtKB are the translations of all protein coding genes annotated on an INSDC submitted genome assembly. To gather this data accurately and consistently it is vital the proteomeDB input procedures are maintained in line with the data structures in source databases. The data flow diagram available in Appendix 3 is a visualization of all procedures providing an understanding of how each piece of the proteomesDB jigsaw is used to build the complete proteome sets.

The EMBL to proteome import code was originally written in Perl. The debug and rewrite of all original data input code into Java was a vast undertaking. Working closely with the developer dedicated to this task ensured all decision trees were fully understood within the context of the pipeline and biologically relevant. If these criteria were met the decision tree was maintained and, where possible, simplified. The outcome of our work is a newly designed, production level code that is fully documented, efficiently and constantly maintained. It is run nightly to incorporate new proteomes from the entire scope of life, alongside perpetual maintenance of all INSDC accessions within existing proteomes. The steps in the procedure are now described. The bulk of the genome data is captured automatically; those genomes that fail to be gathered for reasons that will be explained later in this section can be manually added via the proteome editor.

Maintenance of proteomes can be very complicated as an INSDC entry may be a few hundred base pairs long and constitute a complete genome (for example the genome of *Hepatitis delta virus genotype II* is 866bp, <http://www.ebi.ac.uk/ena/data/view/D90192>), or be a contig entry that is several million base pairs long and span just a portion of a chromosome (for example a section of the human chromosome 1 is 9,224,644bp, <http://www.ebi.ac.uk/ena/data/view/GL000004>). Some entries represent a linear piece of chromosomal genomic DNA and others a circular genome for a bacterial plasmid. A genome can be defined by one INSDC entry or thousands of INSDC entries within a WGS project. For an idea of the volume of sequence data within ENA and the data captured within each data class please look here: <http://www.ebi.ac.uk/ena/about/statistics>. This diversity of architecture for each genome adds to the complicated maintenance of the data for each species.

The initial step of the import procedure is to fetch all the INSDC genome entries that are candidates for inclusion into proteomesDB. The first source is the “classical” INSDC entries that ENA curators identify as being complete genomes. The curators execute a script that examines each entry looking for key elements such as the words chromosome or complete genome within the description line, the complete genome keyword or consistent annotation between entries that suggest the genome of interest has many components. This is a laborious and error prone process but essential as there is no single way to accurately define a complete genome entry. The ENA complete genomes are published here: (<http://www.ebi.ac.uk/genomes/>).

The second source of genome entries is the ENA WGS project database (<http://www.ebi.ac.uk/genomes/wgs.html>). WGS projects are highly variable in nature and can reflect progressive releases of a genome and its annotation by a submitter, or be a one off dump of data never to be touched again. Each WGS project is considered a complete genome at the current time. To determine if it is also a complete proteome the total number of gene models annotated on the genomic DNA is compared to the average proteome size of its taxonomic neighbours. This check is run by the “Complete proteome detector” script. Within a taxonomic clade, a margin of error is allowed for a proteome to be considered complete and if the size falls within this margin the proteome is integrated into proteomesDB. If there are no neighbourhood proteomes then the species is flagged for manual verification. Inclusion of this proteome relies on manual authentication by a proteome curator.

Once the entries are gathered, the next task is to identify the taxonomy of the genome component(s). Taxonomy is a dynamic system and, for each organism, proteomesDB must track any possible changes to the NCBI taxonomic classification of a species, namely the tax\_id, scientific name and the UniProtKB 3-5 letter organism code assigned to all species of interest which is called the “oscode”. These oscodes can be newly defined or they may move between tax\_ids (consequence of a proteome moving from species level to strain/isolate level organism code). In addition to the public taxonomy data, UniProtKB can also request tax\_ids from NCBI (Federhen, 2012) that remain private until UniProtKB publishes the data. The taxonomy decision tree is shown in Fig. 10. The outcome of this decision tree is that where possible a proteome is linked to an oscode. This code is essential for a proteome to be visible within UniProtKB – no oscode, no complete proteome for the user. Proteomes that do not have an oscode remain within the database but are considered “unassigned” and sit in the tables awaiting a curator to manually assign the appropriate code. This is a significant rate limiting step within the production of proteomes (Fig. 11) and one for which no alternative solution has yet been discussed within the project. The principle reason for this is because oscodes are essential to identify an organism throughout the whole of the TrEMBL production process and therefore all UniProtKB entries. A change to this would have enormous repercussions for all production scripts, all datasets and consequently a huge impact on users too.

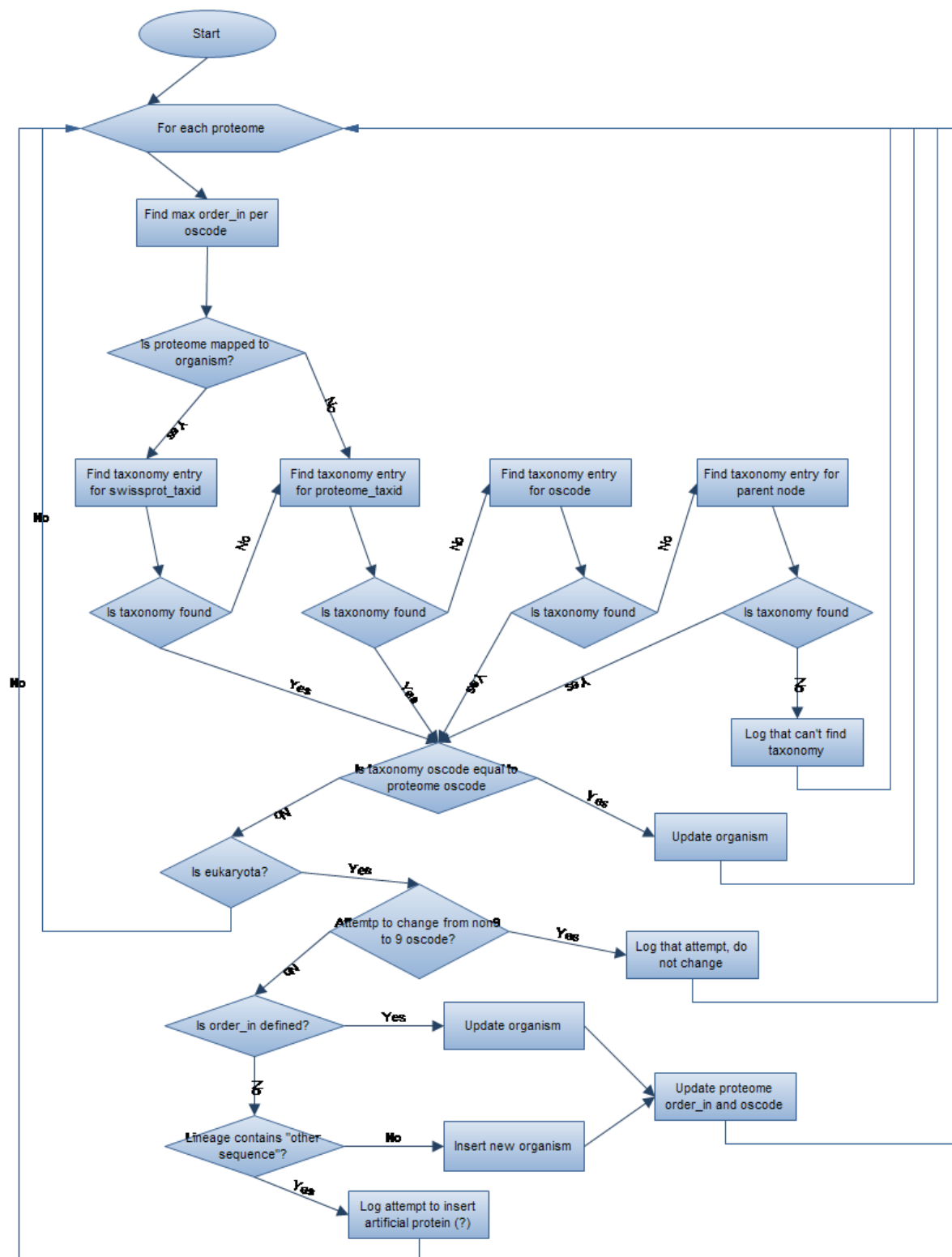
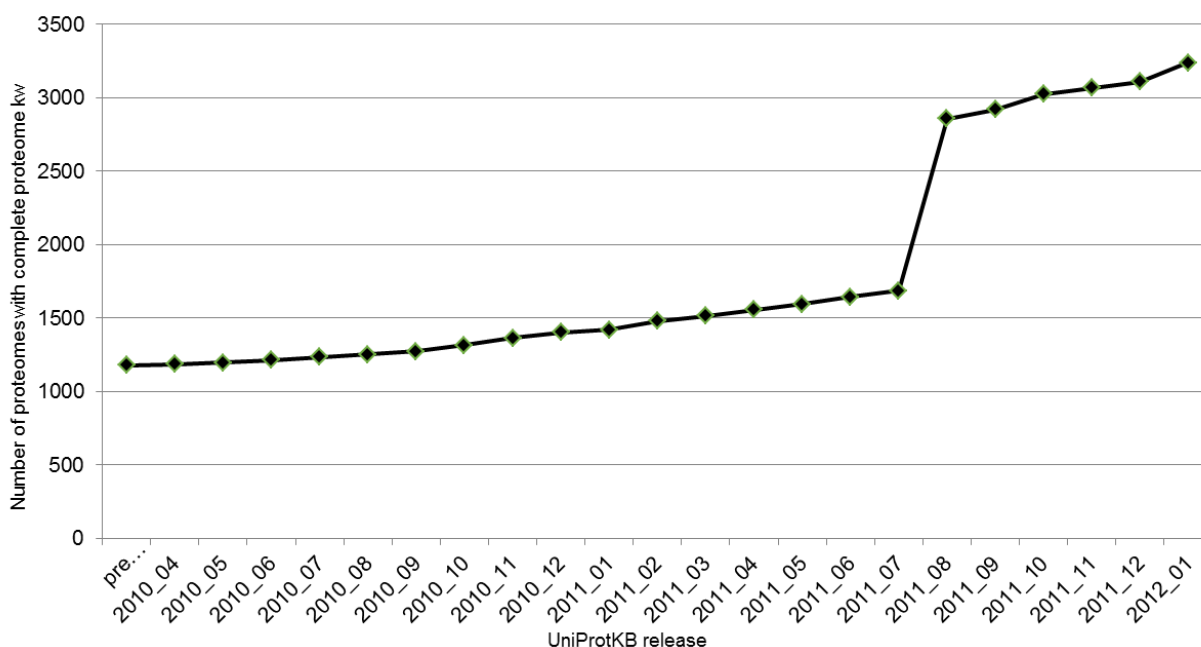


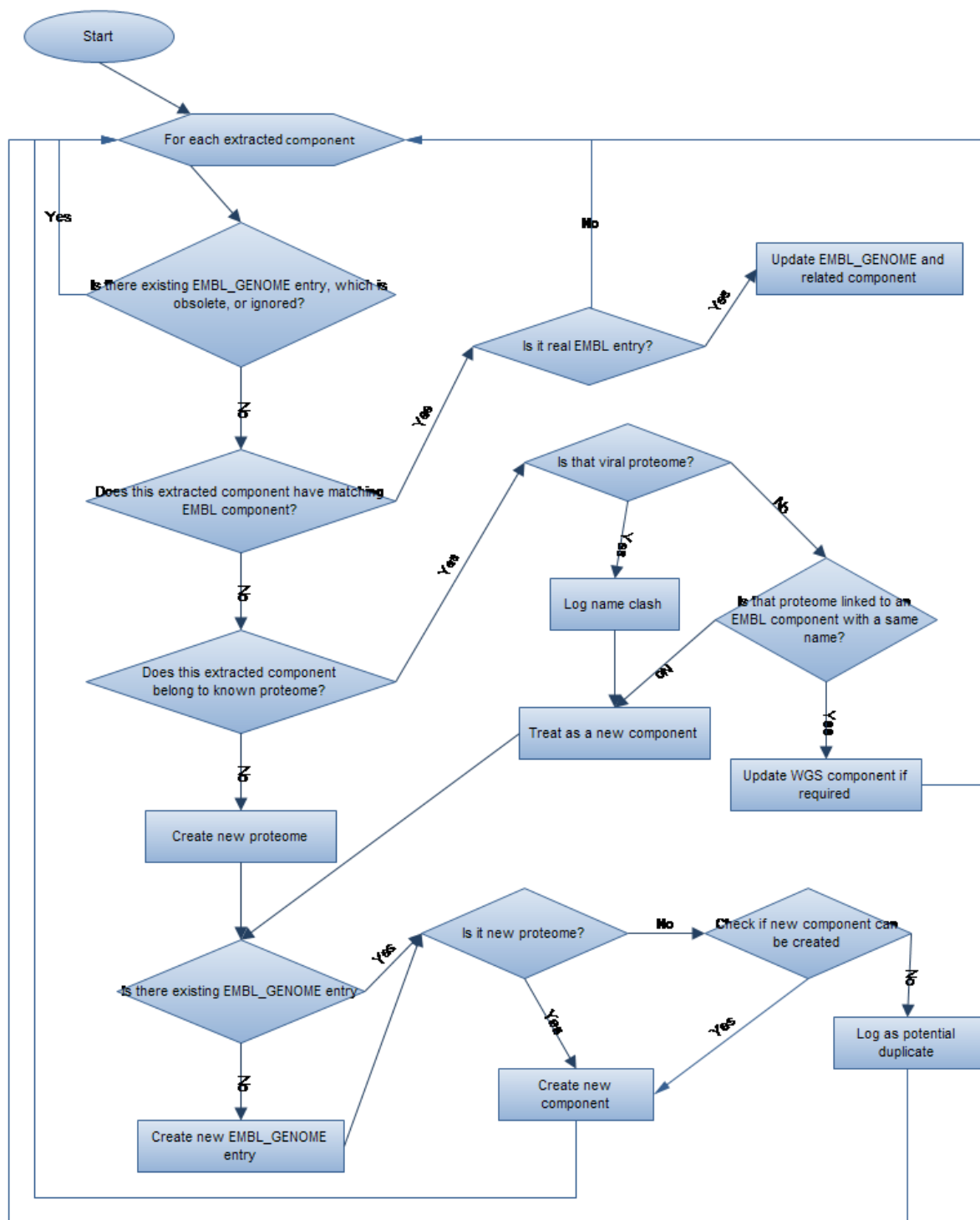
Figure 10. Taxonomy decision tree module. For each new organism that enters proteomesDB, the decision tree determines if it is new to the database, if it already exists, and if it does, is there already a proteome for that species (this is known using the order\_in column).



**Figure 11. Linear growth in UniProtKB Complete proteomes. The significant leap between UniProtKB release 2011\_08 and 2011\_09 is the inclusion of viral proteomes in preparation of the first release of the UniProtKB reference proteomes.**

Within Fig. 10, one of the decisions within the tree is “is eukaryote?” This is part of the original code before it was rewritten in Java and reflects the history of the database; one of its original intentions was to serve the HAMAP curation project. Over time, as eukaryotic genomes became more plentiful, it was decided that the database should extend to eukaryotes, hence the “unstreamlined” additions to the code to deal with these. Now there are no such specific rules, each tax\_id of cellular organisms undergoes identical scrutiny before inclusion or exclusion from the database.

An accurate identification of the taxonomic source is essential to determine if candidate INSDC entry(ies) are of a new genome or, less frequently, a duplication of a genome. The entries could also be an update to an existing genome (addition of missing components, removal of deleted components or a transfer of annotation from one to another ENA component). These decisions are made using the tree in Fig. 12.



**Figure 10.** The module that checks each INSDC genome entry to identify if it defines a new proteome or if it is an update to an existing proteome. Viral proteomes have specific rules due to the multitude of the genomes available.

### 2.1.2 From Ensembl to proteomes database: Ensembl import

Despite the thoroughness of the import procedure already described above, there are key model organisms that do not possess a full annotated genome and will therefore be absent from proteomesDB. These species, for example *Rattus norvegicus*, instead have excellent transcriptomes that constitute thousands of entries in INSDC. Defining a transcriptome instead of a genome is wrong biologically and maintaining the thousands of mRNA INSDC entries that constitute the transcriptome would cause a logistical nightmare if they were added to proteomesDB. Fortunately, to avoid this chaos, the Ensembl gene build pipeline (Flicek *et al.*, 2012) uses these mRNAs to annotate gene models on a reference genome. In collaboration with Ensembl, an Ensembl import pipeline has to be developed as part of this research: proteomes defined and maintained by Ensembl are incorporated into UniProtKB. As this pipeline covers organisms for which some sequences in UniProtKB are available, these existing sequences have to be reconciled with those imported from Ensembl. Within in each species/strain, the procedure works in the following way:

- Ensembl sequences are mapped to their UniProtKB counterparts requiring 100% sequence identity over 100% of the length of the two sequences
- Those that are a 100% match gain the keyword 'Complete proteome' and a cross reference to Ensembl
- Ensembl sequences that are absent from UniProtKB are imported *de novo* into UniProtKB/TrEMBL and tagged with the keyword 'Complete proteome' and gain a cross reference to Ensembl
- A complete proteome is formed from all UniProtKB/Swiss-Prot entries (irrespective of whether they map to Ensembl) plus those UniProtKB/TrEMBL entries with a cross reference to Ensembl.

The first high profile proteomes that benefited from this process were human and mouse and these were available in UniProtKB release 2011\_05 (<http://www.uniprot.org/news/2011/05/03/release>, Headline: Complete proteome sets for *Homo sapiens* and *Mus musculus*). In release 2012\_01, twenty five other chordate species with high assembly coverage are available and include anole lizard, chicken, cow, dog, elephant, gibbon, gorilla, horse, human, little flying bat, macaque, marmoset, mouse, opossum, panda, pig, platypus, rabbit, rat, sea squirt, stickleback, tasmanian devil, turkey, xenopus and zebrafish. All Ensembl import proteomes are updated with each UniProtKB release to ensure consistency between Ensembl and UniProtKB.

Researching candidate Ensembl import species that should be considered for inclusion in UniProtKB requires evaluation of many factors, namely:

- Genome coverage (preferably >6x for Sanger sequencing methods),
- date of the genome assembly and consequent full gene build,
- contig and scaffold N50 (these give an rough idea of the number of truncated contigs and sequence gaps within a genome, though the number can be skewed by genome submissions containing short sequences), and


- number of species specific mRNAs that are available in INSDC that can be mapped onto the genome to build a gene model and how many gene models are generated purely by projection from a well annotated chordate (human is used predominantly).

All this information can be found within the Ensembl database and within the Ensembl description page for each species, Rat is shown in Fig. 13.

Description

**Rat (*Rattus norvegicus*)**

**Assembly**



The Rat Genome project is an international collaboration to sequence the genome of the brown rat (*Rattus norvegicus*). The DNA sequence is generated by [Baylor College of Medicine](#), [Celera Genomics](#), [Genome Therapeutics](#), [The Institute for Genome Research](#), and [The University of British Columbia](#).

The rat data used on this site can be downloaded directly from [Baylor College of Medicine](#).

**Annotation**

Ensembl is working with the broader rat genomics community ([RGD](#)) to provide annotation of the rat genome. We have also participated in the [STAR consortium](#) to help identify and map single nucleotide polymorphisms in the rat.

The Rat RGSC 3.4 assembly was annotated using the standard Ensembl GeneBuild pipeline. To improve the gene set, we have incorporated new data resources which have become available since the last RGSC 3.4 genebuild (August 2006), including an updated rat-specific repeat library, additional RefSeq and Uniprot protein sequence data for predicting the coding regions of protein-coding genes, as well as new cDNAs and ESTs for annotating untranslated regions (UTRs) of protein-coding genes. This results in the extension of previously partially-predicted genes, merging of genes which were previously mis-annotated as two distinct neighbouring genes, and the recovery of new rat genes with mammalian orthologues.

**Figure 11. Description of the genome assembly and annotation for Rat on the Ensembl website, [http://www.ensembl.org/Rattus\\_norvegicus/Info/Index](http://www.ensembl.org/Rattus_norvegicus/Info/Index). Accessed Feb 2012.**

Next-generation sequence (NGS) assemblies can have a high coverage, for example 30x, but this can mean a fragmented and low-quality assembly as NGS assemblies are more influenced by short read length (70-350bp depending on the technology) and the quality of the assembly algorithm that aligns the huge amounts of sequence data. If the Ensembl gene models are consequently fragmented then the assembly is considered of a low quality and UniProtKB will not import the proteome.

## 2.2 Input of data to UniProtKB

Proteome import is the process that propagates data from proteomesDB to the appropriate UniProtKB entries to build the entire set of entries that define a proteome. This update runs with each UniProtKB release and is responsible for the addition and removal of proteome annotation to and from UniProtKB entries, respectively. It is vital this proteome import step is highly optimized, with post-process checking procedures in place, as the output of this pipeline is the publically available complete proteome sets that are visible and accessible to users via uniprot.org. The original import code was written in Perl (Wall *et al.*, 2000) and has been rewritten into Java by a member of the proteomes team, overseen by me. As for the EMBL to proteome code rewrite, each step is fully debugged, documented and described here.

A complete proteome in proteomesDB awaiting import of data into UniProtKB will have all the following values:



- An ocode assigned to the tax\_id of the proteome
- organism.annotation\_status = Y
- proteome.is\_complete = 1.

For every proteome that will be new to or require updates within the next UniProtKB release will have these values and the ocode will be listed in the proteome import trigger table. Complete genome INSDC accessions and their associated protein\_ids are used to identify those UniProtKB entries that should receive proteome annotation. Protein\_ids are the protein identifiers for each annotated gene model within an INSDC entry. These protein\_ids are used by the TrEMBL production procedure to retrieve the appropriate INSDC annotations that are used to generate UniProtKB/TrEMBL entries.

For each proteome that needs to be processed import tables are populated and the appropriate UniProtKB entries will receive the following augmentations:

- The correct taxonomic specification of the species, strain or isolate as appropriate. In UniProtKB entries this equates to the ocode, tax\_id and organism name affecting the ID, OX and OS lines, respectively,
- capture of the genome project locus names as ORFNames or OrderedLocusNames in the gene name (GN) line,
- genome project reference will be standardised to be the submission reference or, where possible, a scientific publication for that organism. This updates the reference, R, lines, and
- addition of the “Complete proteome” keyword (KW) to the KW line.

Fig. 14 is an example of a UniProtKB entry within a complete proteome that has received proteome annotations.

```

ID   C3PDN0_BACAA          Unreviewed;      316 AA.
AC   C3PDN0;
DT   16-JUN-2009, integrated into UniProtKB/TrEMBL.
DT   16-JUN-2009, sequence version 1.
DT   14-DEC-2011, entry version 26.
DE   RecName: Full=L-lactate dehydrogenase 3;
DE       Short=L-LDH 3;
DE       EC=1.1.1.27;
GN   Name=ldh3; OrderedLocusNames=BAA_5272;
OS   Bacillus anthracis (strain A0248).
OC   Bacteria; Firmicutes; Bacillales; Bacillaceae; Bacillus;
OC   Bacillus cereus group.
OX   NCBI_TaxID=592021;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RA   Dodson R.J., Munk A.C., Bruce D., Detter C., Tapia R., Sutton G.,
RA   Sims D., Brettin T.;
RT   "Genome sequence of Bacillus anthracis A0248.";
RL   Submitted (APR-2009) to the EMBL/GenBank/DDBJ databases.
CC   -!- CATALYTIC ACTIVITY: (S)-lactate + NAD(+) = pyruvate + NADH.
CC   -!- PATHWAY: Fermentation; pyruvate fermentation to lactate; (S)-
CC       lactate from pyruvate: step 1/1.
CC   -!- SUBUNIT: Homotetramer (By similarity).
CC   -!- SUBCELLULAR LOCATION: Cytoplasm (By similarity).
CC   -!- SIMILARITY: Belongs to the LDH/MDH superfamily. LDH family.
DR   EMBL; CP001598; ACQ49366.1; -; Genomic_DNA.
DR   RefSeq; YP_002869240.1; NC_012659.1.
DR   ProteinModelPortal; C3PDN0; -.
DR   STRING; C3PDN0; -.
DR   EnsemblBacteria; EBBACT00000129898; EBBACP00000125238; EBBACG00000127114.
DR   GeneID; 7849720; -.
DR   GenomeReviews; CP001598_GR; BAA_5272.
DR   KEGG; bai:BAA_5272; -.
DR   PATRIC; 18776452; VBIBacAnt132916_5510.
DR   GeneTree; EBG00050000001153; -.
DR   ProtClustDB; PRK00066; -.
DR   GO; GO:0005737; C:cytoplasm; IEA:UniProtKB-SubCell.
DR   GO; GO:0004459; F:L-lactate dehydrogenase activity; IEA:HAMAP.
DR   GO; GO:0000166; F:nucleotide binding; IEA:InterPro.
DR   GO; GO:0006096; P:glycolysis; IEA:HAMAP.
DR   HAMAP; MF_00488; Lactate_dehydrog; 1; -.
DR   InterPro; IPR001557; L-lactate/malate_DH.
DR   InterPro; IPR011304; L-lactate_DH.
DR   InterPro; IPR018177; L-lactate_DH_AS.
DR   InterPro; IPR022383; Lactate/malate_DH_C.
DR   InterPro; IPR001236; Lactate/malate_DH_N.
DR   InterPro; IPR015955; Lactate_DH/Glyco_Ohase_4_C.
DR   InterPro; IPR016040; NAD(P)-bd_dom.
PE   3: Inferred from homology;
KW   Complete proteome; Cytoplasm; Glycolysis; NAD; Oxidoreductase.
FT   NP_BIND      14      42      NAD (By similarity).
FT   ACT_SITE     178     178     Proton acceptor (By similarity).
FT   BINDING      91      91      Substrate (By similarity).
FT   BINDING     123     123     NAD or substrate (By similarity).
FT   BINDING     154     154     Substrate (By similarity).
FT   BINDING     233     233     Substrate (By similarity).
SQ   SEQUENCE      316 AA;  34771 MW;  575157416E72F03E CRC64;
      MKRHTRKIAI IGTGLVGSSC AYSIVNQGIC EELLLIDINH ERAVGEAMD L SHCINFNTNR
      TKVYAGSYED CKMDIVIIIT AGPAFKPGQS RLDTLGASAK IMESVVGVM ESGFDGIFLL
      ASNPVDIITY QVWKLGLPR NRVI GTG TSL DSSRLRTILS EMLHVDPRSI HGYSLGEHGD
      SQMVANSHVT VGGKPILQIL EEQKERFGEI DLDEIVEKTA KAGWEIYKRK GTTYYGIGNS
      LAYIASSIFN DDHRVIAVSA ILDGEYGEYD ICTGVPPIIT RDGIREIVEL NLTEDEESRF
      AKSNDILRDY MKTIGY
//

```

**Figure 12. A UniProtKB entry augmented with proteome annotation. Proteome annotation is highlighted in yellow.**

The procedure also writes data in proteomesDB and the current UniProtKB release database, respectively:

- Proteome.kw\_added\_release\_id column within the proteomes table keeps track of the first public release of a complete proteome, and
- proteome2uniprot table is populated with data required for data integrity checking within and between UniProtKB releases.

### 3. RESULTS

#### 3.1 Proteome editor

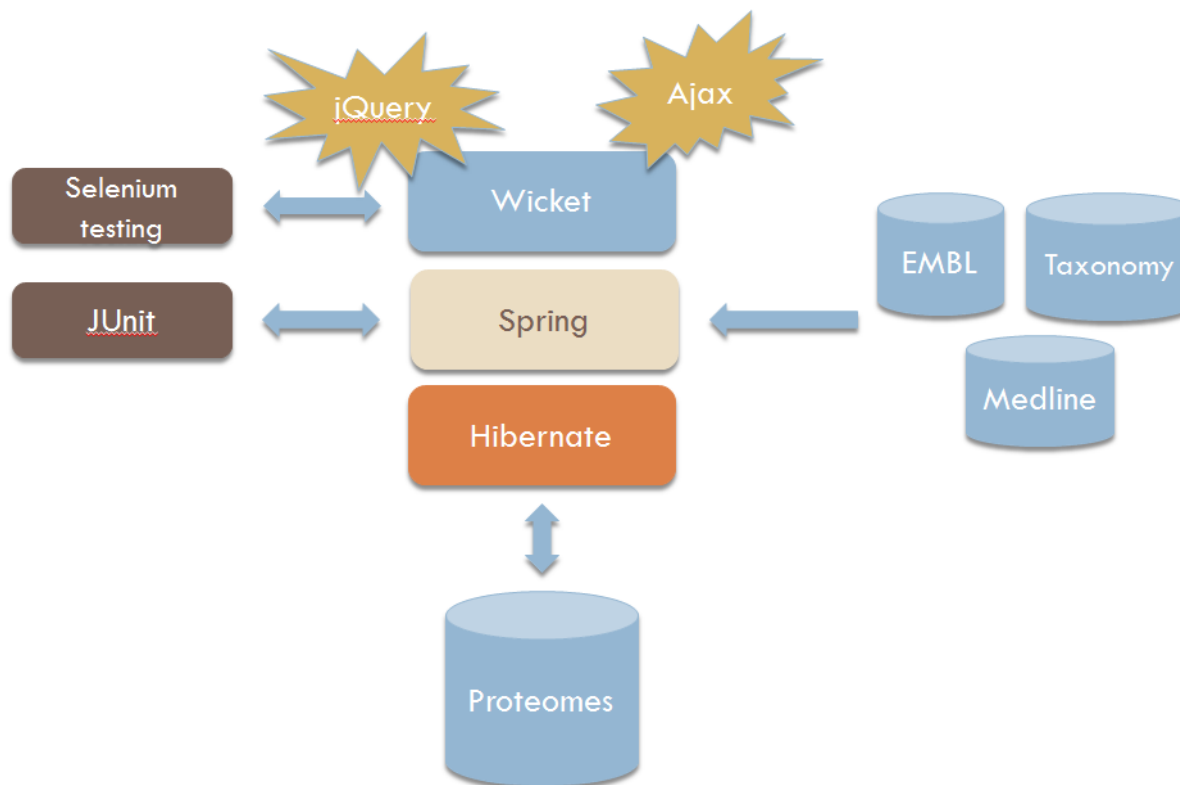
While the proteomes pipeline successfully captures the vast majority of the complete genomes, some may be missed. An example of this is where a genome comes along that has a surprising number of gene models compared to its taxonomic neighbours, a consequence of an evolutionary response to environmental conditions. Two distinct examples are *Trichomonas vaginalis* and *Mycoplasma genitalium*.

Trichomoniasis is a common sexually transmitted infection (STI) caused by a tiny flagellate parasite called *Trichomonas vaginalis*. When the proteome of *T. vaginalis* was made available (Carlton *et al.*, 2007) the protein count of 50,189 far exceeded that expected; the nearest taxonomic neighbour being *Phytophthora infestans* with 17,602 proteins (Hass *et al.*, 2009). The genome was found to be approximately 160 Mb in size and as much as two-thirds of the sequence consisted of repetitive and transposable elements. This reflects a massive, evolutionarily-recent expansion of the genome. The total number of predicted protein-coding genes is nearly 60,000, which include ~38,000 'repeat' genes (virus-like, transposon-like, retrotransposon-like, and unclassified repeats) all with high copy number and low polymorphism.

*Mycoplasma genitalium* is an obligate intracellular parasitic pathogen which has one of the smallest genomes known for a free living organism at 0.58 Mb (Fraser *et al.*, 1995) translating 483 proteins. Its nearest neighbour, *Mycoplasma mobile*, has 628 proteins (Jaffe *et al.*, 2004a), a 30% increase. Due to its environmental niche in the human urogenital tract, *M. genitalium* lacks genes encoding enzymes required for amino acid biosynthesis, the peptidoglycan cell wall, tricarboxylic acid (TCA) cycle enzymes and many other biosynthetic genes. Instead the parasite acquires these components from the host. The genome preferentially maintains genes necessary for this host dependent mode of life so a significant portion of its genome is devoted to the transport of nutrients from its host such as glucose and fructose, adhesins for attachment and gene for antigenic variation to evade the host immune system.

While these proteomes are translations of true complete genomes, they are considered incomplete by automatic procedures and as a consequence not imported into proteomesDB. For proteomes that fall into this scenario we have the proteome editor which allows manual addition of all details and edit all flags in the proteomes database to generate a UniProtKB complete proteome. The proteome editor is an annotation tool that was initially developed at the Swiss Institute of Bioinformatics (SIB) in Geneva to give HAMAP curators a means to inspect a genome of interest and, if required, make manual edits. Over time more curators were using the tool, wanting to edit a broader taxonomic range of genomes and demanding improved functionality. This, along with the expectation of thousands of new complete genomes to be submitted to the public domain, prompted a redesign of the proteome annotation platform to improve curator efficiency, usability and user experience. To achieve this, the tool was migrated to EMBL-EBI and became the responsibility of a developer within the proteomes team. During development all updates were tested, bugs and required modifications were communicated directly to the developer. To optimise the tool, it was adapted to become a Java Web application built with the

open-source frameworks Apache Wicket (Vaynberg, 2011), Spring (Walls, 2010) and Hibernate (Linwood and Minter, 2010). The front end makes extensive use of Ajax and Javascript libraries, such as jQuery. Fig. 15 illustrates this standard structure: a layer to communicate with the database - Hibernate, a middle layer to handle all the business logic - Spring and a layer for web interface - Wicket. This allows for flexibility with regards to future developments, for instance changing technologies for the web interface requires redoing only the web layer.



**Figure 13. System architecture diagram of the platforms involved in the redesign of the Proteome editor.**

The proteome editor has become a user friendly and intuitive tool. To use the editor a proteome curator must have a login account. Once logged in a proteome can be retrieved from proteomesDB via a free text search using the scientific name or oscore (Fig. 16) or via an advanced query using other data types, for example an INSDC accession. Proteomes are also available from browsable lists where the proteomes are divided into their superregnum and listed alphabetically. The data is displayed in a tabulated view; taxonomy (Fig. 17), genome component(s) (Fig. 18) and genome reference(s) (Fig. 19). All these items can be viewed, edited or deleted as required. Unassigned proteomes are available as a browsable list and the editor allows these proteomes to become linked to a new oscore which therefore makes the proteome a candidate for import into UniProtKB.

**PROTEOME EDITOR** beta Account Log out

Add Organism Advanced Search Browse Manage Cross-Refs  Search


---

KLULA: *Kluyveromyces lactis* (strain ATCC 8585 / CBS 2359 / DSM 70799 / NBRC 1267 / NRRL Y-1140 / WM37) (Yeast) ✖ Delete  
(*Candida sphaerica*)

Taxonomy Proteomes References

Figure 14. Top page to view all properties of the fungus *Kluyveromyces lactis* proteome.

### Taxonomy

 **Edit**








 <b>KLULA</b>	<i>Kluyveromyces lactis</i> (strain ATCC 8585 / CBS 2359 / DSM 70799 / NBRC 1267 / NRRL Y-1140 / WM37) (Yeast) ( <i>Candida sphaerica</i> )
 <b>Type</b>	eukaryota
 <b>Lineage</b>	Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; <i>Kluyveromyces</i>
 <b>TaxID</b>	284590
 <b>Scientific Name</b>	<i>Kluyveromyces lactis</i> (strain ATCC 8585 / CBS 2359 / DSM 70799 / NBRC 1267 / NRRL Y-1140 / WM37)
 <b>Common Name</b>	Yeast
 <b>Synonym</b>	<i>Candida sphaerica</i>
<b>Use for annotation</b>	true
<b>ORF Regexp</b>	

Figure 15. Taxonomy of the fungus *Kluyveromyces lactis*.

## Kluyveromyces lactis

Re-assign
 Edit
 Remove

Strain: ATCC 8585 / CBS 2359 / DSM 70799 / NBRC 1267 / NRRL Y-1140 / WM37  
 Complete: Y  
 Reference: N  
 Status 2

### COMPONENTS

**Chromosome A - CR382121**

TaxID: 28985 Strain: Type: EMBL Primary

Regexp: KLLA0A(\d{5})g? Sequential: Unknown Shift:

References: 15229592 x + Add

**Chromosome B - CR382122**

TaxID: 28985 Strain: Type: EMBL Primary

Regexp: KLLA0B(\d{5})g? Sequential: Unknown Shift:

References: 15229592 x + Add

**Chromosome C - CR382123**

TaxID: 28985 Strain: Type: EMBL Primary

Regexp: KLLA0C(\d{5})g? Sequential: Unknown Shift:

References: 15229592 x + Add

**Chromosome D - CR382124**

TaxID: 28985 Strain: Type: EMBL Primary

Regexp: KLLA0D(\d{5})g? Sequential: Unknown Shift:

References: 15229592 x + Add

**Chromosome E - CR382125**

TaxID: 28985 Strain: Type: EMBL Primary

Regexp: KLLA0E(\d{5})g? Sequential: Unknown Shift:

References: 15229592 x + Add

**Chromosome F - CR382126**

TaxID: 28985 Strain: Type: EMBL Primary

Regexp: KLLA0F(\d{5})g? Sequential: Unknown Shift:

References: 15229592 x + Add

**Mitochondrion - AY654900**

TaxID: 28985 Strain: ATCC 76492 / CBS 2359/152 / CLIB 210 Type: EMBL Primary

Regexp: Sequential: Unknown Shift:

References: 15691736 x + Add

**Plasmid pGKI-1 - X00762**

TaxID: 28985 Strain: Type: EMBL Primary

Regexp: Sequential: Unknown Shift:

References: 6473099 x + Add

**Plasmid pGKI-2 - X07776**

TaxID: 28985 Strain: Type: EMBL Primary

Regexp: Sequential: Unknown Shift:

References: 3041369 x + Add

+ Add

Figure 16. The genome of the fungus *Kluyveromyces lactis*.

## References

Complete nucleotide sequence of the mitochondrial DNA from *Kluyveromyces lactis*.

 Edit  Delete

15691736

*FEMS Yeast Res.* 5:315-322(2005)

10.1016/j.femsyr.2004.09.003

Zivanovic Y, Winkler P, Vacherie B, Bolotin-Fukuhara M, Fukuhara H.

ASSIGNED TO

*Kluyveromyces lactis*

[Mitochondrion](#)

Nucleotide sequence and transcription analysis of a linear DNA plasmid associated with the killer character of the yeast *Kluyveromyces lactis*.

 Edit  Delete

6473099

*Nucleic Acids Res.* 12:6011-6030(1984)

10.1093/nar/12.15.6011

Stark M J R, Milleham A J, Romanos M A, Boyd A.

ASSIGNED TO

*Kluyveromyces lactis*

[Plasmid pGK-1](#)

Genome evolution in yeasts.

 Edit  Delete

15229592

*Nature* 430:35-44(2004)

10.1038/nature02579

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Labatolre I, de Montigny J, Marck C, Neuvéglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich J-M, Beyne E, Bleykasten C, Boltrame A, Boyer J, Cattolico L, Confanioli F, de Daruvar A, Despons L, Fabre E, Fairhead C, Ferry-Durazeth H, Groppi A, Hantraye F, Hennequin C, Jaumaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Müller H, Nicaud J-M, Nikolski M, Ozols S, Ozier-Kalogeropoulos O, Pellenz S, Poter S, Richard G-F, Straub M-L, Suleau A, Swennen D, Telata F, Wesolowski-Louvel M, Westhof E, Wirth B, Zenlou-Meyer M, Zivanovic Y, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissbach J, Winkler P, Souciet J-L.

ASSIGNED TO

*Kluyveromyces lactis*

[Chromosome A](#) [Chromosome B](#) [Chromosome C](#) [Chromosome D](#) [Chromosome E](#) [Chromosome F](#)

Genome organization of the killer plasmid pGKL2 from *Kluyveromyces lactis*.

 Edit  Delete

3041369

*Nucleic Acids Res.* 16:5863-5878(1988)

10.1093/nar/16.13.5863

Tomasino M, Ricci S, Galeotti C L.

ASSIGNED TO

*Kluyveromyces lactis*

[Plasmid pGK-2](#)

**Figure 17. The publications associated with the genome of the fungus *Kluyveromyces lactis*.**

The evolution of the new proteome editor has been assisted through persistent beta testing and error reporting by proteome curators. The efficiency of the developer has ensured that existing bugs were resolved and suggested new functionality is fully operational. The result is an excellent annotation tool that gives visualization of a proteome allowing intuitive editing as required.



### 3.2 Reference and representative proteomes

UniProtKB needs to reorganise its complete proteomes so that users are able to navigate their way through the taxonomic tree of life to seek out organisms of interest and not get lost in the plethora of UniProtKB entries that are available. To achieve this, UniProtKB has manually defined a subset of the complete proteomes as being “Reference Proteomes” and the PIR members of the UniProt consortium have developed an automatic procedure to define “Representative Proteomes” (Chen *et al.*, 2011).

Proteome curators, in collaboration with Ensembl and NCBI Reference Sequence collection (RefSeq, Pruitt *et al.*, 2012), chose an initial list of reference proteomes for UniProtKB release 2011\_09.

Reference proteomes have been selected to provide broad coverage of taxonomy (including viruses), and constitute a representative cross-section of the diversity to be found within UniProtKB. A species chosen to become a Reference proteome has to fulfill certain criteria, they must be:

- the standard for a particular user community,
- extensively studied to understand biological phenomena with the expectation that discoveries will provide insight into the workings of other organisms, and
- of interest for biomedical and biotechnological research.

Species of particular importance may be represented by numerous reference proteomes for specific ecotypes or strains of interest. Reference proteomes will be a target for manual annotation when resources permit.

After some discussions the list was finalized for release 2011\_09 at 455 proteomes with the understanding that the species chosen will be continuously reviewed as new proteomes of interest become available and as existing taxonomic classifications are revised (release 2012\_01 has 512 reference proteomes). The goal of the reference proteome collaboration is that the same consensus sets will be provided by all three resources in the future.

To accommodate this new definition of proteomes, a new column has been added to the proteome table: `is_reference`. Currently two values are allowed (1 and 0). With proteome import, any proteome with `proteome.is_reference = 1` will add the keyword “Reference proteome” to the appropriate UniProtKB entries, in addition to the “Complete proteome” keyword. Reference proteomes were made available in UniProtKB release 2011\_09 (<http://www.uniprot.org/news/2011/09/21/release>).

In contrast to the manually chosen reference proteomes, representative proteomes are automatically defined as being the best representative of a taxonomic grouping in terms of the majority of the sequence space and annotated information. The automatic selection is made by first creating Representative Proteome Clusters (RPGs). RPGs are groups of proteomes that are co-members within a UniRef50 cluster. Those RPGs that include a defined reference proteome are identified and selected out of the automatic procedure. For those RPG remaining the most information-rich proteome (based on a collective annotation score of its entries) within an RPG is chosen as being the Representative. The RPGs and their representative proteomes (RPs) are visible at PIR (<http://pir.georgetown.edu/rps/>) and will

soon be integrated into UniProtKB. RPs are calculated at 75, 55, 35 and 15% UniRef50 co-membership thresholds using a top-down approach that ensures an RP at a lower threshold is also an RP at a higher threshold. The 55% threshold (RP55) will be used for prokaryotes and RP75 for eukaryotes as these levels most closely follow the standard taxonomic classifications and preserves the majority of the annotation and sequence diversity of the entire UniProtKB, while reducing the sequence space by more than 80%.

The annotation score is a metric for protein annotation quality that can provide an overview of an entire proteome. Specification of these metrics relies on capturing the full complexity of protein annotation: nomenclature, functions and processes in time and space, sequence annotation and alternative products (isoforms and natural variants) and defining the evidence level of each annotation item. To each of these a score is assigned depending on how they contribute to the understanding of the protein, a heavier weighting being given to experimental functional properties than to predicted features. This provides the definition of "maximal" annotation and the highest score achievable for a protein entry. The annotation score of a proteome is the mean average of the sum of all the constituent entry scores.

UniProtKB is in the process of evaluating this data from PIR and aims to make it public via uniprot.org at the end of 2012. In line with this, data will be added to proteomesDB that will result in the new keyword "Representative proteome" to appropriate UniProtKB entries.

### **3.3 Further improvements to proteomes database and UniProtKB entries**

#### **3.3.1 Database and input procedures**

A clean-up of data within proteomesDB has been performed. Each update identified (listed below) is small in itself but as part of a larger process they all contribute to increased quality of the data within proteomes, and therefore UniProtKB too. The process of cleaning is ongoing and with each UniProtKB release data quality and reliability improves.

Essential database clean ups that ensure we have no "floating" or unassigned data include:

- deletion of organisms lacking a proteome,
- deletion of proteomes with no components, and
- deletion of references not assigned to a component.

These updates are reported to me for manually verification securing no loss of data or introduction of data inconsistency within a proteome.

Other clean-ups include edits to fields that are imported from proteomesDB into UniProtKB and are causing syntax errors in the resulting UniProtKB entries. The bulk of these edits are to the reference fields: presence of double spaces or unexpected characters in the author list; the journal title in the citation line being incorrect UniProtKB format and more. These were allowed into the database during a

testing phase of the proteome editor and have now been corrected so that such errors can no longer be loaded into the database.

The pipeline has been improved to extract only the data from ENA that is required for proteome import into UniProtKB. For example the translation table and molecule type present in an INSDC entry were captured but are of no use within proteomesDB or a UniProtKB entry so are no longer captured.

Import code that is no longer required for production has been deprecated. These were SGD- and TAIR-specific import pipelines that were over complicated and problematic. Now the respective MODs have recently submitted an annotated genome to INSDC these are no longer needed.

Some decisions made by the import code were inconsistent between superregnum, for example `annotation_status` (the flag that defines if a proteome is imported into UniProtKB) was set to 'yes' for bacteria and archaea but 'no' for eukaryotes. This resulted in a continuous increase in the number of proteomes for bacteria and archaea in UniProtKB but a minimal increase for other kingdoms. This has now been resolved so all superregnum are treated equally.

The database triggers and auditing of tables has been greatly improved and simplified to track all changes to the database. The following tables are audited: `COMPONENT`, `COMPONENT2EMBL`, `DB_XREF`, `EMBL_GENOME`, `EXTERNAL_DB`, `ORGANISM`, `PROPERTY`, `PROTEOME`, `REFERENCE`, `REFERENCE2COMPONENT`, and `XREF2PROTEOME`. The audit table records the three main events namely insert, update and delete. Previously only update and delete events were recorded which made auditing more complicated than required.

In addition to an edit made by the import pipeline, every edit made by the proteome editor is now captured. Certain edits may require proteome import to be rerun, as this process requires a list of oscodes as an input, it is the oscode that is loaded into the trigger table. This is much simpler than recording the data type that was edited as this information is not as meaningful as the oscode itself. The previous situation where all automatic but few manual updates threw a trigger and the log of the triggers were over complicated is now resolved.

Additional columns have been made to accommodate recent developments in UniProtKB complete proteomes:

- `is_reference`, defines if the proteome has been manually chosen as a reference proteome, with the introduction of a database constraint that `is_reference=1` is only allowed if `is_complete=1`. Consequence being that only complete proteomes can become reference proteomes. Reference proteomes were introduced in UniProtKB release 2011\_09.
- `proteome_map_type`, defines how the proteome in proteomes database is mapped to UniProtKB. `proteome_map_type = 0` is used for classical proteomes, `proteome_map_type = 1` is for Ensembl import species and `proteome_map_type = 2` is for classical proteomes that require slightly different behaviour in UniProtKB/Swiss-Prot.

### 3.3.2 Input procedure warnings file

In Fig. 11 and Fig. 12 you may observe some decisions that result in output written to a log, for example “Log that can’t find taxonomy” and “Log as potential duplicate”. These errors can only be resolved by manual inspection of the proteome and its components. This log file of errors had grown so large it had become unreadable and unusable. Work has been performed to resolve this; a combination of improvements to the data as well as the input pipeline to reduce the number of warnings.

The simplest update was to the text of each warning type making it more intuitive and informative so a curator could actually understand the warning being reported and therefore the action required. Other changes included defining the difference between a warning and an error, for example an INSDC accession that had been suppressed or deleted should not be a primary accession in proteomesDB, so these are now reported as errors instead of warnings.

A warning was being issued for the component\_name of some WGS projects. The component\_name is created by looking for text strings within the description line of an INSDC entry. If none of the expected strings appears then a warning is reported and the default name “Chromosome” is assigned. WGS projects were being incorrectly interpreted and given the default component name; this has been improved to capture each WGS component as “Unassembled WGS sequence” which no longer causes a warning.

Errors in taxonomy were also discovered: incorrect tax\_ids for WGS entries were being retrieved and private tax\_ids were considered false as they were not public yet. For both of these the bug identified was that the procedures were querying the wrong table. Correction was essential to ensure the correct taxonomic assignment for every proteome.

Those updates in INSDC that can impact on an entire genome are:

- Changing the version number of a WGS project, e.g. AACQ01 to AACQ02
- Movement of a WGS project into contig entries, e.g. AAEU02001436 into CM000157
- Identification of new entries to an existing genome
- Loss of annotations from an existing genome.

Each of these is tracked, monitored and reported as they occur. Fortunately, given their significance, these are low frequency updates.

The complete proteome detector that is used to evaluate new WGS proteomes for integration into the database is now also run over all WGS proteomes in proteomesDB to ensure they continue to qualify as complete proteomes. Those that are reported as incomplete require research by me to ascertain if the annotations have changed detrimentally or if in fact the proteome remains complete.

### 3.3.3 Cleanup of UniProtKB entries

Manual annotation of UniProtKB entries began in 1986. This was a time when the availability of the human genome was a dream, and little did the curators know of the avalanche of complete genomes that would fall into the sequence databases over the next 25 years. Annotation policies were defined according to the data available and to fulfill achievable objectives. For example, UniProtKB assigned oscodes at species-level and merged all proteins or proteomes from child strains to that oscode. The entries may contain sequences and experimental information from one strain with a complete proteome as well as sequence and experimental information from other strains. This curation legacy has caused many complications in the capture and output of complete proteomes. To illustrate this problem let us consider the *Bacillus anthracis* proteome; unbeknownst to a user, the *B. anthracis* proteome is the composite of three different strains Ames, Ames ancestor and Sterne. As a consequence, the proteome has 6,614 entries, which is surprising if the user is expecting approximately 5,600 entries. The current *B. anthracis* proteome must be demerged into its individual strain proteomes to resolve this issue with the current proteome.

To achieve this, UniProtKB must define a new oscode for each strain level NCBI tax\_id. This desynchronisation of a proteome from its former tax\_id and/or oscode is announced for high profile proteomes to help prevent confusion to users. UniProtKB release 2011\_05 saw the change in taxonomy of the model fungal organisms *Saccharomyces cerevisiae* (YEAST) and *Schizosaccharomyces pombe* (SCHPO) (News item: <http://www.uniprot.org/news/2011/05/03/release>). For lower profile species, the proteome project need to define a priority list and work through these oscodes one by one, this is where being a proteome as well as a UniProtKB curator are essential skills.

The new oscode is the tip of the iceberg; the data within the entries has to be divided. Demerging a UniProtKB/TrEMBL entry is an automatic process, tagging of data within the database allows easy identification of which item belongs to which tax\_id. Manual demerging of a UniProtKB/Swiss-Prot entry requires the references, biochemical characterisation and feature annotations to be split into the individual proteome strain entries. This requires complicated rules that reflect the complexity of the biological data captured. The scale of the problem is different for each organism, for example 724 UniProtKB/Swiss-Prot entries need to be demerged to create the three proteomes for *B. anthracis*: <http://www.uniprot.org/uniprot/?query=organism%3A1392+keyword%3A181&sort=score>.

These demerges are and will continue to be performed by proteome curators so that successive UniProtKB releases will see cleaner proteome sets.

### 3.4 Quality assurance and error reporting

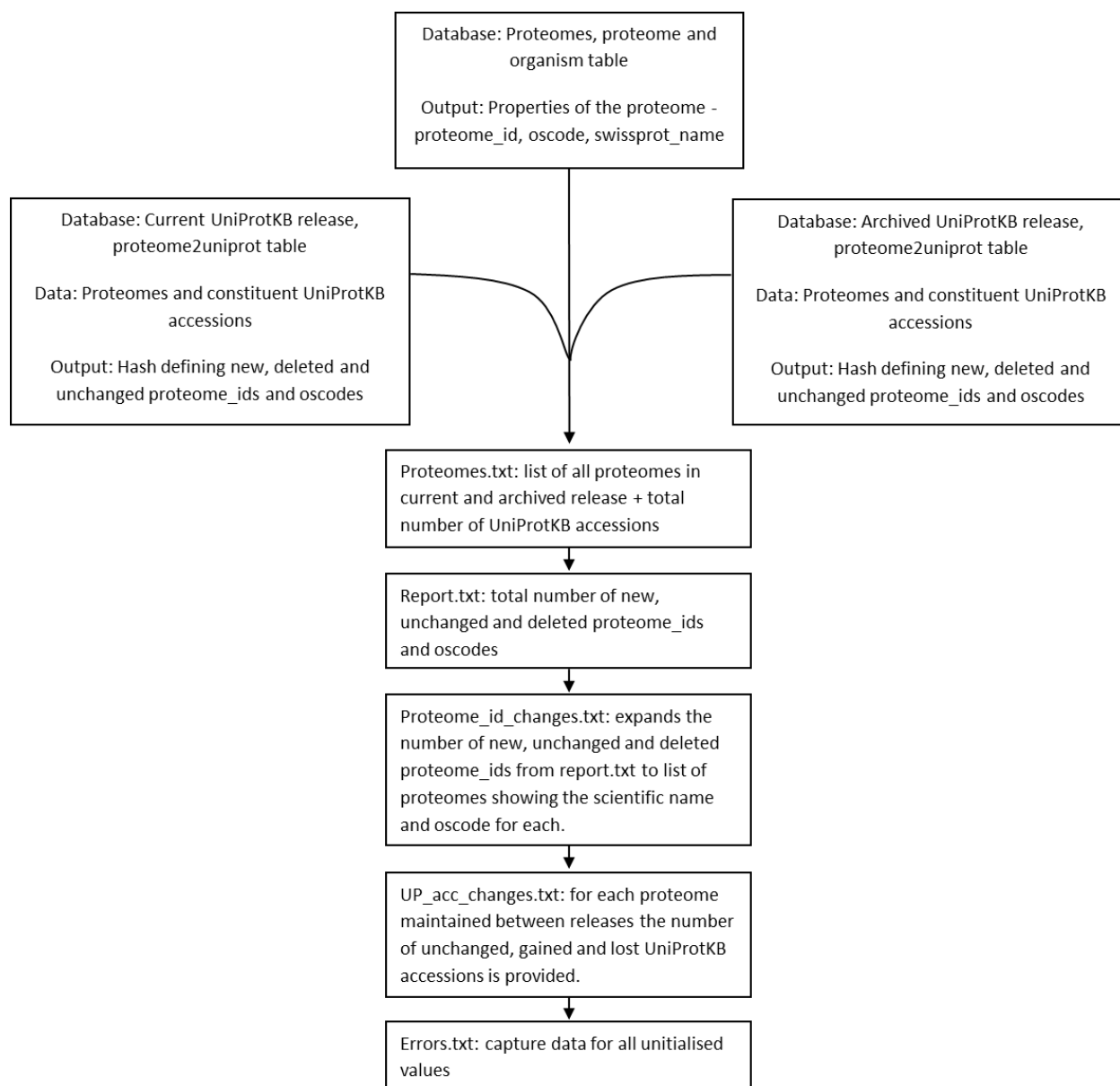
It is important for UniProtKB complete proteomes project to maintain consistency between releases – guard that each proteome is available from one release to the next and assure no proteome has been lost. To do achieve this, a post processing procedure has been written in Perl (Wall *et al.*, 2000). This is

run after the import of proteome data into UniProtKB entries. The script is written in Perl for several reasons:

- Perl DBI is a simple database interface that allows direct interaction with DBMS by executing native SQL queries and retrieving results back in a list form,
- String processing and pattern matching is a major component of the script and is one of Perl's strengths that outperforms Java,
- Multidimensional hashes of key value pairs are easier to implement and more flexible in Perl than Java, and
- Perl is less verbose than Java.

As is good practice, and in accordance with other UniProtKB production Perl scripts, the script has been written with the warning and strict pragmas. Also, as Perl is notorious for having several different ways to perform a task, the code is commented throughout to ensure other programmers know what the code does and how it does it, thereby easing future debugging and maintenance of the script.

The script (described in Fig. 20) compares the proteome2uniprot table, which is populated with proteome\_ids and their constituent UniProtKB entries, from the current UniProtKB production database to a copy of the same table from the previous archived UniProtKB release. All proteomes and their constituent UniProtKB accession numbers are retrieved from the new and archived releases. A comparison is made between releases to identify the new, deleted and unchanged proteomes. For those proteomes present in both releases it is interesting to know how the UniProtKB entries that build the proteome have changed so pattern matching of the accessions is used to address this question. As proteome\_ids themselves convey little information, data from proteomesDB is used to supplement these with the scientific name and the oscode of the organism for the generated reports.



**Figure 18. Data flow diagram of the proteome post processing script.**

The synopsis of the changes between releases is shown in report.txt. Appendix 4 contains the report between release 2012\_01 and 2012\_02. Note that while 2012\_01 is the current public release, release 2012\_02 is internally available and undergoes numerous production procedures before it is released 4 weeks after the previous release. It is important for the report to show the changes in proteome\_ids and oscodes between releases as proteome\_ids are a direct reflection of the data in proteomesDB and oscodes a reflection of the proteomes available in UniProtKB. One or more proteome\_ids can exist for one organism as a nuclear and organellar proteome can be assigned different proteome\_ids for a variety of reasons (one may not have the strain defined so require different data propagated by proteome import, an Ensembl nuclear proteome requires different pipeline to a mitochondrial INSDC proteome).

The new and deleted oscodes are interesting to know in more detail as they show how the database has changed between releases. These are detailed in proteome\_id\_changes.txt, a snippet of changes between release 2012\_01 and 2011\_02 are shown below:

UP proteome changes between releases

=====

Proteomes that are new to the release:

PYRF1 264555 *Pyrolobus fumarii* (strain DSM 11204 / 1A)

MYCPK 264912 *Mycoplasma putrefaciens* (strain ATCC 15718 / NCTC 10155 / C30 KS-1 / KS-1)

MURRD264914 *Muricauda ruestringensis* (strain DSM 13258 / LMG 19739 / B1)

ENTAL 265106 *Enterobacter asburiae* (strain LF7a)

Proteomes that are deleted from the release:

CLVK 77918 African cassava mosaic virus (isolate West Kenyan 844) (ACMV) (Cassava latent virus (isolate West Kenyan 844))

CPSMV 77937 Cowpea severe mosaic virus (strain DG) (CPSMV)

TMVKR 77941 Tobacco mosaic virus (strain Korean NC 82) (TMV)

CXA24 78058 Coxsackievirus A24 (strain EH24/70)

For those proteomes that are unchanged between releases (same proteome\_id and oscode), a report is provided that has details of how the ingredients of each proteome has changed with respect to UniProtKB accession numbers:

HUMAN, proteome\_id: 25 has 58327 unchanged UP entries

gained 170 entries

and lost 131 entries

DROPS, proteome\_id: 25396 has 16013 unchanged UP entries

gained 0 entries

and lost 0 entries

Human is an Ensembl import species and these entry changes will reflect updates within the latest Ensembl release as well as the usual loss of accession numbers as a consequence of manual UniProtKB annotation merging one or more child entries into one fully annotated master that becomes a UniProtKB/Swiss-Prot entry. DROPS is the oscode for *Drosophila pseudoobscura pseudoobscura* whose genome annotation is maintained by FlyBase (McQuilton *et al.*, 2012). The status quo of this proteome shows no genome annotation updates have been made by FlyBase and no UniProtKB entry merging by UniProtKB curators. This report supplements QA checks already made within TrEMBL production at an earlier stage of the UniProtKB release cycle as gain or loss of a proteome protein\_ids potentially has a very significant impact on proteomes later in the release cycle.

Further reports made by the proteomes team are:



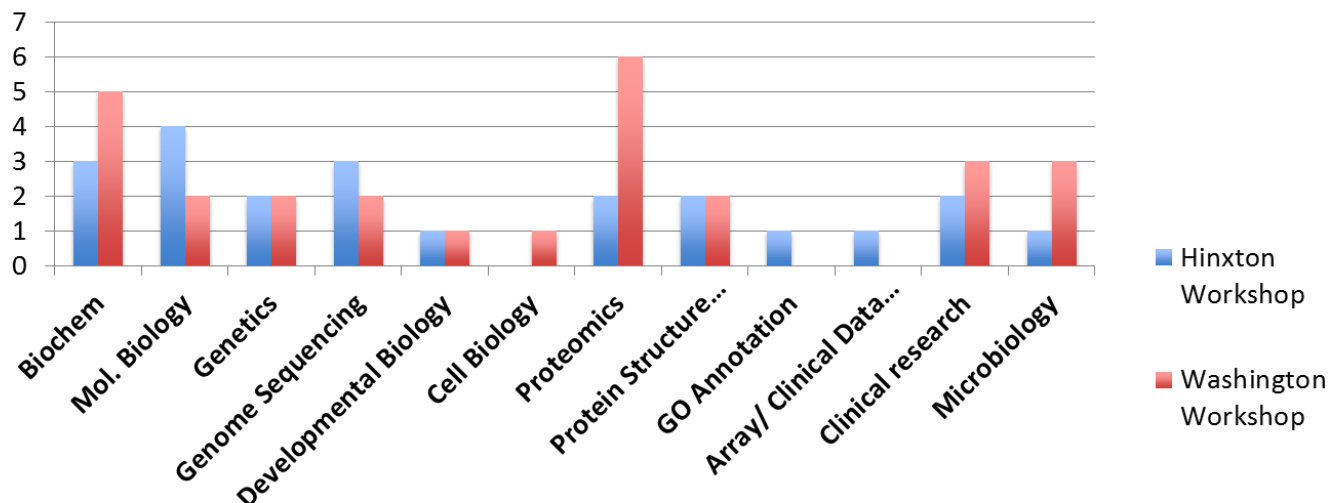
- All taxonomy changes are written to a table (oscode\_tax\_id\_update) and a report is emailed each release, and
- Sql script is run to check correct assignment of Complete proteome keyword within UniProtKB entries, if there are any discrepancies between the proteomesDB and UniProtKB concerning addition or removal of the Complete proteome KW then a flag is toggled that states the oscode must be immediately rerun by proteome import.

These files and reports are all used internally and by external collaborators, PIR, for their generation of representative proteomes.

### **3.5 Complete proteome representation at uniprot.org**

UniProt has hosted two usability workshops for users from industry and academia, the first in PIR Georgetown, Washington in June 2011 and the second at EMBL-EBI in September 2011. These are part of a larger effort by the UniProt Consortium to focus on its user experience. The workshops had shared areas of interest as well as distinct aims in the hope of discovering how our American and European users differ. The activities were designed to understand why and how people use uniprot.org, identify gaps and usability issues and to determine users' priorities and requirements to help guide future development. The first concentrated on discovering if the uniprot.org website was intuitive to its users; can they find information without confusion, can they query the data efficiently and understand the results returned. The results prompted a revamp of the UniProtKB entry page and this redesign was tested at the second workshop. In addition, the second workshop aimed to get feedback and ideas on upcoming pages for the website. One of these was the complete proteome page. We wanted discussions on page design and content, and the expected workflow from initial query through to entries within a proteome.

Attendees to each workshop were personally invited. Each attendee must be familiar with UniProt (or used UniProt databases in the past) and currently work in a wet lab or have close links with wet lab personnel. Ideally they would be a hands-on biomedical researcher or a bioinformatics specialist working directly with biomedical researchers to answer questions and solve problems. Before each meeting the participants filled in a survey to provide details of their current scientific research, their species of interest and the experiments they regularly performed. Fig. 21 shows the results from both workshops.



**Figure 19. Areas of research of scientists attending the two UniProt usability workshops.**

In addition participants were asked what databases they are familiar with to retrieve biological information, what they find most frustrating about online resources and if they could have a wish list of three items what would be the most useful data for them concerning genomes, transcript profiling and protein structure. All the surveys were used as a guide to gauge how the group would respond to certain activities during the workshop.

One of the outcome of the first workshop identified that users wanted UniProt to have complete proteome data; they thought it would be a really useful component. It seems the majority were oblivious to the fact that this data was already present and public. This was a huge incentive to the consortium to improve the website to increase visibility of the proteomes project.

Another outcome was that users really liked the idea of reference proteomes, but at the same time UniProt must not forget about the proteomes of non-model organisms and those organisms with a lower public profile. This was great news considering that the UniProtKB reference proteome initiative was made public with release 2011\_09 on 21<sup>st</sup> September 2011 (<http://www.uniprot.org/news/2011/09/21/release>).

I attended the complete proteome section of the second workshop whose aim was to extend the findings and ask the participants what they would most like from a UniProtKB complete proteome: what data do they think is important to visualise, download and have immediate access to via a query? To achieve this the participants were given a blank canvas and in groups they brain stormed ideas on the best layout, data expected on the tope page, queries they would like to do to receive optimal results, and other tasks. One of the most intriguing outcomes is that users were unaware of exactly what UniProtKB complete proteomes were and the data they would expect to find from the website. This highlighted that we need to raise user awareness of the data, how it can be queried, retrieved and downloaded. Another interesting point is that users were concerned that by restricting their view of the species to the proteome, instead of all the protein translations available for a species, that they would

miss proteins that would be important to them. This is not the case as those supernumerary UniProtKB entries do not belong to the proteome as they are not translations of the reference genome and introduction of these would potentially add redundancy to the existing proteome. UniProt users come from a wide variety of backgrounds and we need to cater for them all. Proteomics users may prefer to download every available sequence for an organism and not be restricted by the proteome.

## 4. DISCUSSION

The aim of this project was to have a complete understanding of all aspects of the generation and maintenance of the proteomes database within the UniProt consortium. The elemental use case for proteomes was to develop a means to help organize the complete proteomes to cope with the flood of expected new sequences and to assist users in their navigation through the information to find what they need for whatever their purpose. The solution is the definition of reference and representative proteomes; a combination of minimal manual curation with a significantly higher automatic contribution, whose overall production fits within the strict UniProtKB release schedule.

With the knowledge gained from this extensive disentanglement of proteomes, areas of improvement, redesign and redevelopment for future projects have been identified. UniProtKB is now in a strong position to fulfill the original use case as well as visualise future use cases for complete proteomes and ensure the project is prepared and poised to provide new and improved functionality to users.

### 4.1 Proteogenomics

Historically, the proteomic and genomic communities have operated independently. The genomic community taking charge of annotation efforts and the resulting predicted proteome passed over to the proteomics community for validation and identification of posttranslational events. The task of annotating the genome for protein coding genes is difficult and requires substantial effort. Most annotation pipelines utilize nucleotide centric information, such as cDNA or homology to known genes, to refine computational predictions. Unfortunately error rates can be high both in terms of genes which are mispredicted and gene models that are wholly missing from the annotation. Incorporating peptides, obtained from mass spectrometry, into the genomics pipeline to annotate the genome would be enormously beneficial, not least as detection of a protein with mass spectrometry allows removal of the “hypothetical” tag associated with many currently annotated open reading frames (ORFs) in biological databases. This new field, proteogenomics, has the potential to dramatically improve the accuracy and completeness of genome annotation by providing an orthogonal data source to predict gene models with levels of sensitivity that are complementary to cDNA sequencing.

Working with *Mycoplasma pneumonia*, Jaffe and colleagues (Jaffe *et al.*, 2004b) endeavoured to build a set of gene predictions based on observations of peptides from expressed proteins; mass spectral data from whole cell lysates of strain FH were correlated to the published genomic sequence annotation of strain M129 via the technique “proteogenomic mapping”. Strain FH is a less virulent, but very closely related strain of *M. pneumoniae* strain M129. The overall organization of the genome with respect to ORF order is substantially the same. This work identified new ORFs, extensions of existing ORFs, and suggested removal of questionable predicted ORFs. The study was able to verify the existence of many heretofore hypothetical proteins as well as add to the suspicion that some gene models do not exist as translated protein products. Some of these differences may be a result of the strain analyzed but overall the study demonstrated the robustness of protein analysis across closely related genomes.

Using MS/MS spectra from *Arabidopsis thaliana*, Castellana and colleagues (Castellana *et al.*, 2008) reannotated the TAIR7 release of the *Arabidopsis* genome. They found that 18,024 of 144,079 peptides did not match current gene models suggesting that 13% of the *Arabidopsis* proteome is incomplete. This was due to approximately equal numbers of missing and incorrect gene models; 778 were missing from the genome annotation, and 695 need to be refined or corrected. After the initial study was completed, TAIR released their next revision of the genome/proteome, TAIR8. Only a small number of the novel peptides (3%) appeared in the TAIR8 release indicating that the proteogenomic approach is complementary to computer-based annotation.

A similar study was performed by Zhao and colleagues (Zhao *et al.*, 2011). High-throughput shotgun proteomic data was used to explore the comprehensive protein expression profile of *Shigella flexneri* 2a strain 301. The outcome was validation of 823 protein products (including hundreds of hypothetical proteins), correction to several start sites and several novel open reading frames (ORFs) were confirmed by combining MS analysis and RT-PCR. This analysis allowed detection of annotation errors in the genome annotation, such as incorrect start sites assignment, sequencing errors, and wrongly annotated pseudogenes. The findings of novel ORFs provide a new clue to conduct functional research. Moreover, some of the novel ORFs were identified as overlapping genes, which increases our understanding of the complexity of the genome structure and reveals the underestimation of such gene arrangements within prokaryote species. Correction to all gene model annotations prevents propagated misannotation of gene models in future releases of the *S. flexneri* 2a genome or in the genome of any orthologous species.

Currently UniProtKB complete proteomes are the predicted proteomes from genome annotations. There is potential to expand these proteomes with proteogenomic supported gene models either via submitter reannotation of the genome entries in INSDC or via an import pipeline from one of the main repositories of MS derived proteomics data, the Proteomics Identifications Database (PRIDE, Vizcaino *et al.*, 2009). An import pipeline from PRIDE could be a future consideration for the development of UniProtKB.

## 4.2 Homology: orthologs and paralogs

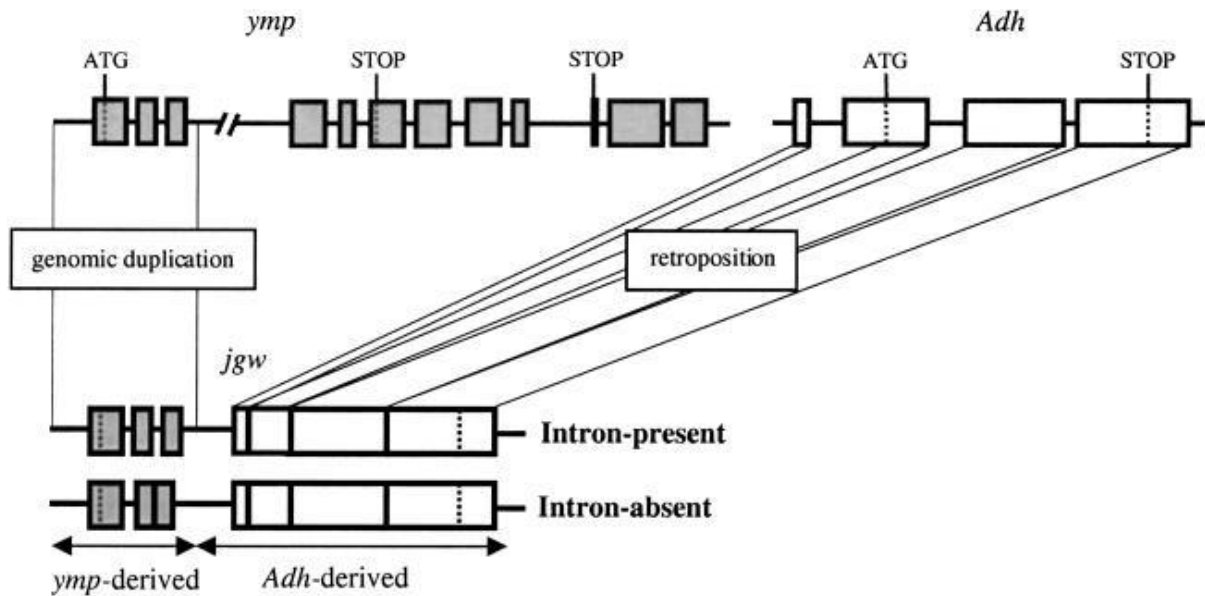
The determination of orthologs within UniProtKB would also be highly beneficial to users. While users are requesting this information, they do acknowledge it is very difficult to accurately predict. This being the case, would users trust ortholog data provided by UniProtKB? It is very important that UniProtKB does not reinvent the wheel with regards to ortholog detection when other projects responsible for developing such algorithms are becoming more trusted and improved with each release. Instead uniprot.org could attempt to unravel the complexity and display ortholog determination for its users. So ... what is an ortholog? How does an ortholog differ from a paralog and are they all just homologs?

Homology traditionally refers to evolution from a common ancestor by vertical descent. This is a very simplified view that works well for single domain proteins, but multidomain proteins evolve via both

vertical descent and lateral transfer of domains among unrelated families. The concept of a homologous gene family is a major focus in the field of molecular evolution to reconstruct evolutionary relationships across species and is the basis for many function prediction strategies.

The distinction between orthologs and paralogs is defined within the broader context of a homologous family. The definitions of orthology and paralogy provided by Fitch over 40 years ago (Fitch *et al.*, 1970) distinguish between these two classes of homology, those descended from a common ancestor by virtue of a speciation event (orthologs) *versus* those that diverged by gene duplication (paralogs). This definition is purely evolutionary and does not take into account the fact that orthologs often have similar function. Indeed, people often use the term ortholog to refer to genes with a conserved function. Also an ortholog is used by some to describe the “original” gene in a gene duplication event; the ortholog will be the gene that remains in its original syntenic context.

Given the complexity of multidomain proteins, should the entire sequence length be used to predict orthologous relationships or a local domain within a translation? As homology traditionally refers to evolution from a common ancestor by vertical descent should subgenic sequence fragments be considered as the units of interest? Many ortholog detection applications require a definition of homology where the entire gene is the basic unit and as a result important families are frequently excluded from genomic analyses. An example of this is shown in Fig. 22. A model of homology that can be applied to multidomain families would be a great utility, never forgetting that conserved domain architecture between orthologs does not necessarily mean common function.

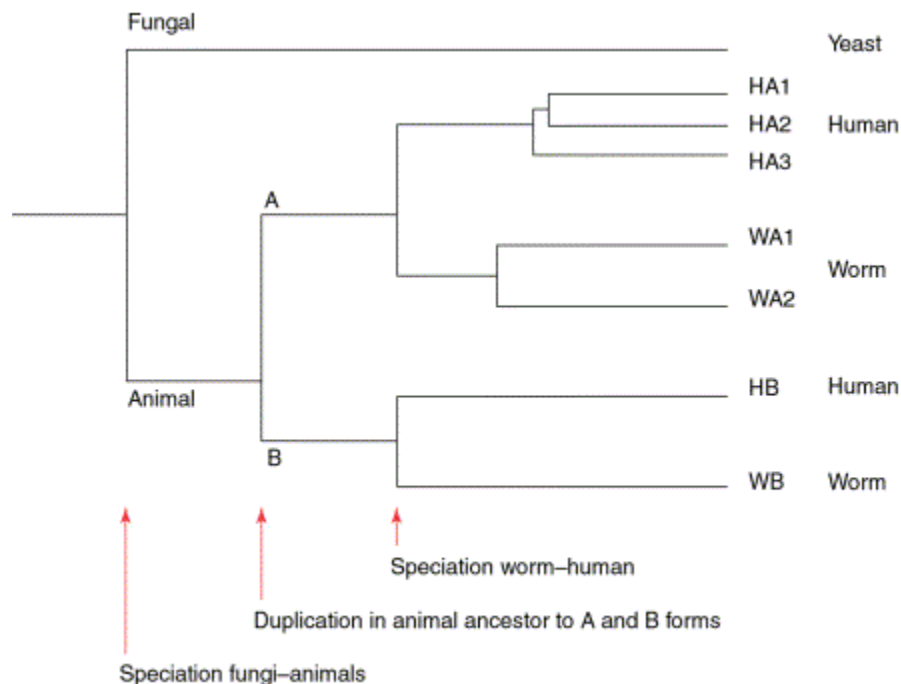


**Figure 20. The yellow emperor (*ymp*, residing at 96E on the right arm of chromosome 3) and Alcohol dehydrogenase (*Adh*, residing at 35B on the left arm of chromosome 2) genes from *Drosophila melanogaster* are fused during a speciation event to generate the chimerical structure of *Drosophila teissieri jingwei* (*jgw*) gene, taken from Llopart *et al.*, 2002. Boxes symbolize exons and the lines between exons represent introns. Exon 2 and 3 are fused in the intron-absent copy of *jingwei* (*jgw*) because of a polymorphic genomic deletion.**

The “ortholog conjecture” that, at a similar degree of sequence divergence, orthologs are generally more conserved in function than paralogs, has been a prevailing paradigm in comparative genomics originally supported by theory rather than empirical studies. This derives from the knowledge that gene duplication has a role in functional divergence so paralogs would be expected to have differing functions. But is this functional diversity the more common outcome? Consider two recently duplicated paralogs, are these are likely to be more similar in function than two distant orthologs? Forslund (Forslund *et al.*, 2011a) has recently shown that orthologs exhibit greater domain architecture conservation, and therefore assumed function, than paralogs at the same evolutionary distance. Henricson (Henricson *et al.*, 2010) has shown that intron positions are also conserved between orthologous pairs, though the maintenance of their position over long evolutionary timescales is seen for only a fraction of the dataset for two main reasons: the difficulty in correctly assigning intron positions and the movement of introns plays a role in evolution. Also, Huerta-Cepas (Huerta-Cepas *et al.*, 2011a) demonstrates that gene duplication is specifically associated with higher levels of tissue expression divergence and that a significant part of this divergence was acquired shortly after gene duplication, when comparing human and mouse. In contrast, Nehrt (Nehrt *et al.*, 2011) demonstrates that paralogs are often a much better predictor of function than orthologs, even at lower sequence identities; among paralogs, those found within the same species are consistently more functionally similar than those found in a different species. This study is based on annotation of Gene ontology (GO, The Gene Ontology Consortium, 2012) function terms to human and mouse proteins and in conclusion

they state that the most important factor in the evolution of function is not amino acid sequence, but rather the cellular context in which proteins act.

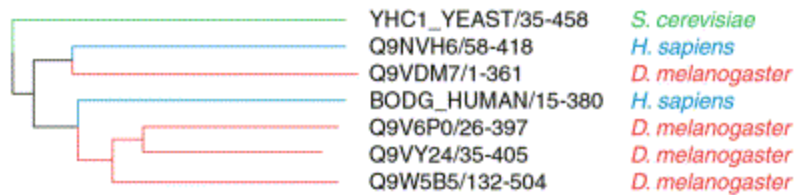
In today's world where the number of fully sequenced genomes is growing by the day, accurate and efficient automated prediction of orthology is essential, and to do this there is a need for the unambiguous definition of orthologs and paralogs to dispel the widespread confusion about the exact meanings of key terms. Sonnhammer has provided definitions that are being accepted by the community (Sonnhammer *et al.*, 2002).



**Figure 21. The definition of orthologs, inparalogs and outparalogs, taken from Sonnhammer *et al.*, 2002.**

Fig. 23 shows a hypothetical ancient gene inherited in the yeast, worm and human lineages. The gene was duplicated early in the animal lineage, before the human–worm split, into genes A and B. After the human–worm split, the A form was in turn duplicated independently in the human lineage generating genes HA1, HA2 and HA3 and in the worm lineage generating genes WA1 and WA2. The yeast gene is orthologous to all worm and human genes, which are all co-orthologous to the yeast gene. When comparing the human and worm genes, all genes in the HA\* set are co-orthologous to all genes in the WA\* set. If we look at the paralogs, the definition is an inparalog describes a paralog within an ortholog group and an outparalog is a paralog between ortholog groups, an event prior to a species split. The HA\* genes are 'inparalogs' to each other when comparing human to worm and by contrast, the HB and HA\* genes are 'outparalogs'. However, HB and HA\*, and WB and WA\* are inparalogs when comparing with yeast, because the animal–yeast split pre-dates the HA\*–HB duplication. A real-life example of inparalogs is shown in Fig. 24.





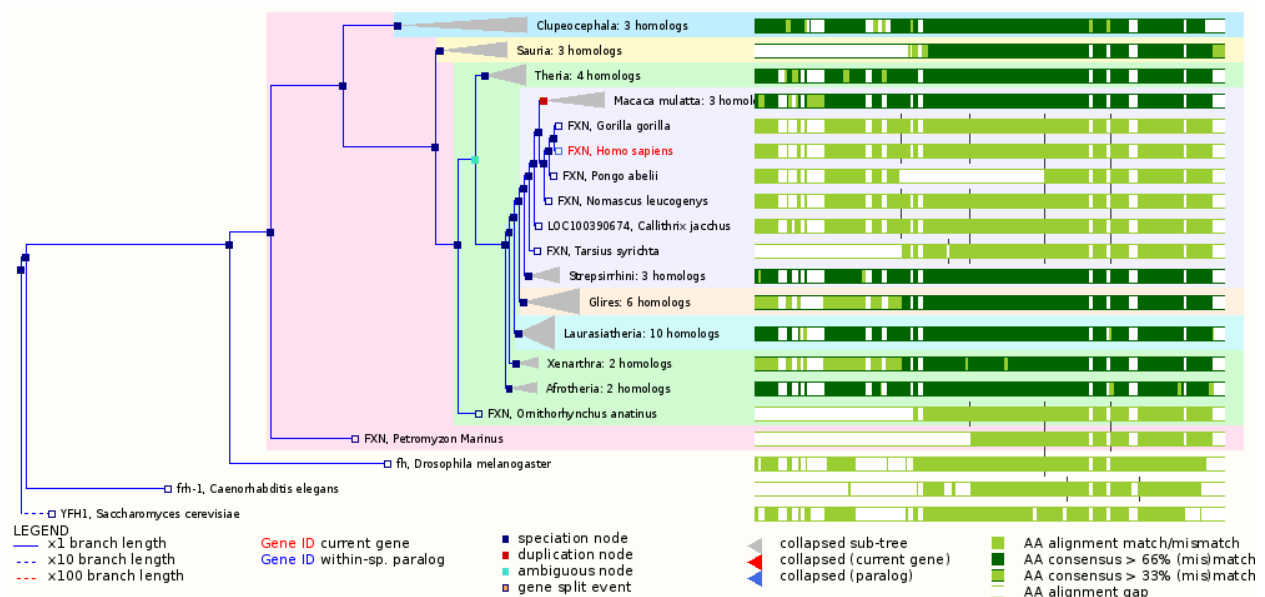
**Figure 22. G-butyrobetaine hydroxylase inparalogs, taken from Sonnhammer *et al.* 2002. The points of speciation and duplication are colour coded so easily identifiable.**

Orthology has the potential to allow inferences from easily studied model systems to much less malleable systems of interest, such as ourselves, humans. The large-scale delineation of gene genealogies is a challenging task, and the numerous approaches to the problem reflect the importance of the concept of orthology, important not least in the study of genetic diseases. There is a plethora of ortholog databases available, each having their own algorithm and pipelines to generate ortholog predictions and each having a distinct focus to the analysis. To illustrate this, a synopsis of a few databases follows.

InParanoid (Ostlund *et al.*, 2010) finds orthologous genes and inparalogous genes that arose after a speciation event. The project uses proteomes of approximately 100 completely sequenced eukaryotic species plus *Escherichia coli* and calculates pairwise ortholog relationships among them using Blast. Analysis of these datasets suggests that orthologs confer a high degree of conservation of functionally important features such as domain architecture and intron positions, and that these are more similar than outparalogs. This supports the “ortholog conjecture”.

OrthoDisease (Forslund *et al.*, 2011b) is an InParanoid-based disease orthology database that surveys the taxonomic distribution of human gene orthologs involved in different disease categories. The hypothesis that paralogs can mask the effect of deleterious mutations predicts that known heritable disease genes should have fewer close paralogs. Large-scale study supports this hypothesis as significantly less duplication is observed for disease genes in the OrthoDisease ortholog groups.

Ensembl Compara (Vilella *et al.*, 2009) is a multi-species database of 47 chordates and three species outgroups (*Caenorhabditis elegans*, *Drosophila melanogaster* and *Saccharomyces cerevisiae*), that stores the results of genome-wide species comparisons calculated for each Ensembl data release. The database includes comparative genomics, for example whole genome alignments and synteny regions and comparative proteomics, for example ortholog and paralog predictions using reconciled phylogenetic trees (Fig. 25) and protein family clusters.



**Figure 23. Ensembl Compara gene view provides displays for data associated at the gene level such as orthologs, paralogs, regulatory regions and splice variants. Gene tree displayed is for the frataxin (FXN), [http://www.ensembl.org/Homo\\_sapiens/Gene/Compara\\_Tree?g=ENSG0000016506](http://www.ensembl.org/Homo_sapiens/Gene/Compara_Tree?g=ENSG0000016506). Access Feb 2012.**

Examination of genome alignments and the resulting trees is essential to identify split genes, these can be caused by the gene build algorithm running over a genome assembly that has errors (missing, inverted or misplaced contigs) and these partial genes would cause additional duplication nodes in a phylogenetic tree. Ignoring these would lead to incorrect conclusion in orthology determination and in the detection of gene family contractions or expansions.

OrthoMCL (Chen *et al.*, 2006) houses ortholog group predictions for 55 species representing phylogenetically diverse lineages from bacterial, archaeal and eukaryotic genomes. Proteins are clustered based on sequence similarity, using an all-against-all BLAST search of each species' proteome, followed by normalization of inter-species differences, and Markov clustering. An incremental method is used to add new genomes to ortholog groups, minimizing the need to recompute results as new datasets are added.

OrthoDB (Waterhouse *et al.*, 2011) is a catalog of eukaryotic orthologs of vertebrates, arthropods and fungi from over 100 species available from the UniProtKB complete proteomes project. Uniform analysis across lineages this different, with divergence levels varying from several to hundreds of millions of years, provides essential data for uncovering and quantifying long-term trends of gene evolution. As expected, the defined orthologous groups confirm that essential genes are highly retained and exhibit strong constraints on gene sequence evolution.

Clusters of Orthologous Groups of proteins (COGs, Tatusov *et al.*, 2003) are delineated by comparing protein sequences encoded in 66 complete genomes of prokaryotes and unicellular eukaryotes that represent major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs

from at least 3 lineages and thus corresponds to an ancient conserved domain. The eukaryotic orthologous groups (KOGs) include proteins from 7 eukaryotic genomes: *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Encephalitozoon cuniculi*. Compared to the coverage of the prokaryotic genomes with COGs, a considerably smaller fraction of eukaryotic genes could be included into the KOGs; addition of new eukaryotic genomes is expected to result in substantial increase in the coverage of eukaryotic genomes with KOGs. Examination of the phyletic patterns of KOGs reveals a conserved core represented in all analyzed species and consisting of approximately 20% of the KOG set. This conserved portion of the KOG set is much greater than the ubiquitous portion of the COG set (approximately 1% of the COGs). In part, this difference is probably due to the small number of included eukaryotic genomes, but it could also reflect the relative compactness of eukaryotes as a clade and the greater evolutionary stability of eukaryotic genomes.

The growing availability of complete genomic sequences from diverse species has brought about the need to scale up phylogenomic analyses, including the reconstruction of large collections of phylogenetic trees. PhylomeDB (Huerta-Cepas *et al.*, 2011b) is the largest phylogenetic repository and hosts 17 phylomes, comprising 416,093 trees and 165,840 alignments. More than 300 million proteome pairs from over 829 fully-sequenced genomes are predicted. For each protein-coding gene in a seed genome (human, yeast, *Escherichia coli*, *Arabidopsis thaliana*, *Candida albicans*, or the pea aphid *Acyrtosiphon pisum*), the alignments and derived phylogenetic trees are available. PhylomeDB is currently being used by several genome sequencing projects that couple the genome annotation process with the reconstruction of the corresponding phylome, a strategy that provides relevant evolutionary insights.

PLAZA (Proost *et al.*, 2009) is a platform for plant comparative genomics. Paleopolyploidy and a multitude of small and large scale duplications in the history of most plant species often results in multiple paralogs per gene. This causes a huge challenge as reciprocal blast hit detection cannot cope with a one-to-many or many-to-one relationships and construction of phylogenetic trees while having the highest confidence it has the lowest coverage as the trees cannot be generated for all gene families. PLAZA has been developed that has an integrative meta-method approach; orthologs are detected on a gene-by-gene basis using gene families, reconciled phylogenetic trees, colinearity information and gene based multi-species families.

As the number and diversity of ortholog databases grows, it becomes more difficult to compare orthologs between the resources due to a lack of standardized data source and incompatible representations of ortholog relationships. To facilitate the first point, UniProtKB is responsible for providing one gene one gene product reference proteome sets for 66 species chosen by the Quest for ortholog (QfO) consortium (<http://questfororthologs.org/>). In an attempt to address the second point, a standardised data exchange format is being developed, two XML schemas (Schmitt *et al.*, 2011): SeqXML, a lightweight format for sequence records, for the input sequences and OrthoXML for the output ortholog clusters. OrthoXML was designed to represent ortholog assignments from any source in a consistent and structured way, yet cater to specific needs such as scoring schemes or meta-

information. A unified format is particularly valuable for ortholog consumers that want to integrate data from numerous resources, e.g. for gene annotation projects. Reference proteomes for the QfO defined organisms are already available in SeqXML, and 10 orthology databases have signed on to OrthoXML. Adoption by the entire field would substantially facilitate exchange and quality control of sequence and orthology information.

The availability of standardized datasets (in terms of format and content) should significantly ease the challenge of sourcing complete proteomes, a problem faced by all providers of ortholog detection, and holds great promise for orthology inference benchmarking. Manually constructed reference gene trees are an ideal benchmark for homolog and ortholog detections. Three well conserved protein families have been manually examined and captured as 3783 gene relationships (Boeckmann *et al.*, 2011). These can be used as benchmark to calculate sensitivity and precision scores for the predictions provided by phylogenomic databases. A web server prototype for orthology benchmarking has been presented to QfO consortium members. This service allows users to easily assess the accuracy of their orthology predictions using a large set of different tests, such as: an assessment of how well the predictions satisfy a standard definition of orthology (Fitch, 1970), discrimination of homologs of multidomain proteins, functional conservation of orthologs and a test assessing accuracy in predicting gene ontology (GO, The Gene Ontology consortium, 2012) function annotations (du Plessis, *et al.*, 2011). It is hoped this will become a public resource well-used by all ortholog providers.

One of the chief benefits of accurate ortholog group assignment is the potential for inferring putative function. With new sequencing methodologies making it increasingly possible to assemble genomes and define genes from species where experimental data is lacking, predicting function is becoming more important. Protein Analysis THrough Evolutionary Relationships (PANTHER, Me *et al.*, 2010) is a comprehensive software system for inferring the functions of genes based on their evolutionary relationships. Reconciled phylogenetic trees of gene families for 48 species are annotated with ontology terms, as part of the Gene Ontology Reference Genome project, describing the evolution of gene function from ancestral to modern day genes. These trees are used to predict the functions of uncharacterized genes, based on their evolutionary relationships to genes with functions known from experiment.

Automatic annotation systems within UniProtKB are striving to provide accurate annotations to UniProtKB/TrEMBL protein sequences that might never be experimentally characterized. One system, UniRule, which incorporates the HAMAP (Lima *et al.*, 2009), RuleBase (Fleischmann *et al.*, 1999) and Protein Information Resource (PIR) (Natale *et al.*, 2004) systems, uses annotation rules created and monitored by experienced curators. Each annotation rule specifies a number of annotations, and conditions which must be satisfied for that annotation to be applied. The other system, the Statistical Automatic Annotation System (SAAS, previously named Spearmint (Kretschmann *et al.*, 2001)) supplements the labour-intensive UniRule system and to ensure scalability of computational annotation. UniRule and SAAS currently predict both protein properties, such as function, catalytic activity, pathways, sub-cellular location and sequence-specific information, such as location of active sites. Addition of ortholog determination as a condition to both systems would greatly enhance their

specificity, allowing high-quality predictions to be added to UniProtKB/TrEMBL entries and prevent propagation of potentially erroneous data.

Orthology inference has been traditionally focused on the study of protein coding genes, but there is increasing interest in applying similar analyses to non-coding RNAs (ncRNAs); Ensembl and miROrtho (Gerlach *et al.*, 2009) have started to provide orthology predictions for a subset of ncRNAs, largely based on synteny. Phylogenetic models used for protein coding genes usually assume that sites evolve independently, but ncRNAs often violate this assumption, owing to the importance of secondary structure conservation. Other limitations hindering phylogenetic study of ncRNAs, include the difficulty in reliably detecting these genes. The RFam database (Gardner *et al.*, 2011) contains a high-quality set of ncRNA families, but its scope is limited to families for which an expert multiple alignment is available. A central repository for RNA sequences has been recently proposed (Bateman *et al.*, in press) and this will be essential to drive evolutionary studies on RNA sequences.

Uniprot.org hopes to display ortholog relationships as part of the interface redevelopment for the complete proteomes web page, along with the identification of orthologs, uniprot.org could provide a “pan-proteome”. In 2005, Tettelin and colleagues introduced a new concept, the “pan-genome” (Tettelin *et al.*, 2005). A pan-genome includes a core genome containing genes present in all strains of a particular species and a dispensable genome composed of genes absent from one or more strains and genes that are unique to each strain. Genomes of multiple, independent isolates are required to understand the global complexity of the species of interest. Some pan-genomes may be “open” and require a multitude of available genomes to establish the core genome, which would represent just a small fraction of the pan-genome. Other pan-genomes may be “closed”, for example a species that occupies an isolated environmental niche like *Bacillus anthracis*. The pan-genome concept has been widely used to analyse the evolution of *Streptococcus pneumoniae* (Hiller *et al.*, 2007), *Haemophilus influenzae* (Hogg *et al.*, 2007), *Escherichia coli* (Rasko *et al.*, 2008) and many others. It has also been used to detect strain specific virulence factors for some pathogens, *Legionella pneumophila* (D’Auria *et al.*, 2010) and develop vaccines against bacterial pathogens (Serruto *et al.*, 2009). To extend this principle further, we can determine the “pan-proteome” for isolates or strains of a species of interest which allows determination of the global protein repertoire; the core and dispensable proteomes. These dispensable proteins would be potential candidates for drug related studies.

Inferring orthology is a non-trivial task. Given the current diversity in algorithms used by ortholog providers, it is very difficult to determine which ortholog prediction out performs the others, is regularly updated and will be maintained for future UniProtKB releases. To provide a reliable service, uniprot.org must display trusted predictions such that our users will be satisfied with the data displayed. Currently UniProt consortium feels this field is too young for this decision to be made at this moment. With the QfO consortium making a concerted effort towards benchmarking and standardising data formats, UniProt consortium hopes that this is something that could be displayed in the future.

### 4.3 The genetics of disease and personal genomics

Hand in hand with the display of orthologous relationships, uniprot.org would have the facility to display those proteins that are unique to a specific taxonomic clade, or species, namely the dispensable proteome. When considering the highly complex and diverse field of drug discovery, identification of these species specific proteins could be enormously beneficial to guide further research; knowing which metabolic pathway controls or is controlled by a protein of interest within a living organisms is key to the development of a therapy. Identification of one disease target can lead to a number of alternative drug targets in the same pathway and increase the possibilities for a novel therapeutic.

To the delight of research medical geneticists, association studies of thousands of individuals are now feasible due to the high-throughput dense genotyping systems such as those produced by Affymetrix and Illumina. These genome-wide association studies (GWAS) are thought to be one of the best ways to identify genomic regions that contribute to the genetic risk of complex diseases such as diabetes, heart disease and asthma. Small genetic differences between individuals that affect their predisposition for certain diseases can be distinguished, giving promise for future healthcare advancements and disease prevention. An explosion in the number of individual genotyping experiments is expected over the next decade due to the work performed by several European projects including the Wellcome Trust Case Control Consortium (WTCCC, <http://www.wtccc.org.uk/>). Hand in hand with this there is a real concern about personally identifiable information entering the public domain. As with DNA sequences and other similar data, there are clear benefits to having open scientific access. However, ethical considerations associated with large-scale medical studies that include the potential for inferring individual phenotype information require restricted data access agreements. The EGA at EMBL-EBI is the central and permanent European archive for genotypes, high resolution genome sequence and phenotypic data. Data protection and provision of anonymised data is a high priority to maintain patient confidentiality. Where permitted by the research subjects, information can be made publicly available and this is provided through Ensembl. In the future, UniProtKB will capture this variant annotation within the respective proteome sets.

23andme (<https://www.23andme.com/>) is a private company that provides genotyping for individuals from a saliva sample. Customers have the opportunity to investigate their genetic information using SNPs from over 200 disease risk genes, and maternal (mitochondrial) and paternal (Y-chromosome) haplotyping allows their ancestry to be traced. Controversial genotypes that have a potentially serious health impact, for example your Apolipoprotein E (APOE) status, are only disclosed if the user accepts the conditions of a genotype warning. This is essential as the information may have an impact on the health of your parents, siblings, partner and children. DNA profiling is not 100% sensitive because a disorder may be caused by an unidentified mutation; therefore, not finding a mutation does not necessarily exclude a diagnosis. Conversely, there may be false positives because detected sequence changes may be non-pathogenic. As new discoveries in health related genetics are published, 23andme send these as monthly reports to keep customers informed.

Disease caused by a Mendelian condition, a single gene with a large effect (Sobreira *et al.*, 2010), can be avoided if there is prior knowledge or a visible phenotype of an individual harbouring the gene. An example would be Friedreich ataxia (FRDA) (Clark *et al.*, 2004), an autosomal recessive neurodegenerative disease caused by hyperexpansion of a polymorphic GAA triplet repeat localized within an Alu sequence (GAA-Alu) in the first intron of the frataxin (FXN) gene. In contrast, some diseases are hidden in the genome and do not reveal themselves as easily. Huntington's disease is an autosomal dominant progressive neurodegenerative disorder; a child has a 50% risk of inheritance from an affected parent (Walker, 2007). Again, this is caused by expansion of a trinucleotide repeat. When the length of the region exceeds a minimum threshold the resulting protein, Huntingtin, is altered and the differing functions of this proteins is the cause of pathological changes which in turn cause the disease symptoms.

A large number of diseases are not caused by the mutation of a single gene, but rather a number of genes that together determine a person's risk of developing a particular disease. For example, certain mutations in the BRCA gene family raise the risk for cancer (Easton *et al.*, 1995). However, this risk does not always equal 100% certainty, and individuals bearing certain BRCA mutations may never develop cancer. Other known allelic variants can increase susceptibility to diseases; ApoE4 mutations increase the chance of the manifestation of Alzheimer's (Saunders *et al.*, 1993).

Progress in DNA sequencing technology has enabled rapid identification of disease genes through genetic screening. Having the knowledge of potential diseases harboured in your genome does not necessarily mean a patient can expect an accurate diagnosis or that the medical help required to control or manage the disease symptoms is available. Early diagnosis and prevention of a disease is of course beneficial to the individual and the health care system, for example a link between haemochromatosis (an inherited disease in which too much iron builds up in your body ) and an increased risk of a ischemic stroke is now known (Ellervik *et al.*, 2007). Within the health care system, cascade screening is being used more to identify biological relatives of a patient known to be affected with a disorder; a working example of this being familial hypercholesterolemia (FH). FH is an autosomal dominant disease (Huijgen *et al.*, 2010) and is one of the most common inherited disorders. Because of the high prevalence of FH among family members, cascade screening has been shown to be a cost-effective method of identifying people with FH providing early detection. Treatment with statins has been shown to reduce morbidity and mortality (Neil *et al.*, 2008).

Although UniProtKB is not a medical-oriented database, about 3,300 human proteins within the proteome set contain manually curated information related to their involvement in pathologies. This data is captured as the position of the single amino acid polymorphisms known to cause the disease, and cross-references to variant databases and genomic resources, such as dbSNP and Ensembl. While clearly of value, such information is not easily accessible for the clinical management of patients as UniProtKB does not use standard medical vocabularies to describe diseases associated to proteins and their variants (Mottaz *et al.*, 2008). In the medical and clinical domain, there have been numerous and successful efforts to implement controlled vocabularies for pathologies; Medical Subject Headings (MeSH terms, the controlled vocabulary thesaurus used for biomedical and health-related documents

indexing, Nelson *et al.*, 2004), International Classification of Diseases (ICD , the official disease classification provided by the World Health Organisation (WHO) for diagnostic information, <http://www.who.int/en/>), and SNOMED clinical terminology (used for clinical information, Donnelly, 2006). All have served well in their respective domain of application and most of these terminologies are collected and organised into concepts in the umbrella resource Unified Medical Language System, UMLS, a major repository of biomedical standard terminologies (Bodenreider, 2004). These common terminologies can act as a metadata layer to provide the missing links between protein information and disease information, overcoming the main obstacle that is the compartmentalization of data in different databases and helping to bridge the gap between clinical medicine and molecular biology for the benefit of both research and public health. The mapping between UniProtKB human entries and to MeSH and ICD is available (<http://research.isb-sib.ch/unimed>).

Environmental factors such as diet, toxic exposures (tobacco), trauma, stress, and other life experiences (exposure to UV light) are assumed to interact with genetic susceptibility factors to result in disease. How do genetics and the environment influence an individual's phenotype, behaviour, intelligence and personality? This nature versus nurture debate is on-going. It is likely that both genetic and environmental components contribute.

Diet should be considered key when studying obesity, though it is now known that copy number variation contributes significantly to the genetic architecture of obesity (Bochukova *et al.*, 2010). There are enormous social implications concerning obesity; childhood obesity being a potential pointer for abuse. A Times article written by M. Henderson in 2009 highlights that two children have been taken off the child protection register, to the immense relief of their parents, when it was discovered that their obesity was not due to parental negligence but in fact due to deletion of the gene SH2B1. SH2B1 is known to have a role in controlling both body weight and glucose homeostasis (Ren *et al.*, 2007).

Cigarette smoking is a major preventable cause of disease worldwide and tobacco addiction is a major cause of death world-wide. Maternal smoking during pregnancy is associated with a number of deleterious outcomes including miscarriage/perinatal mortality, foetal growth restriction and various pregnancy complications, but despite these risks many women still smoke during pregnancy. A study by Freathy demonstrated that genetic factors have a role in influencing smoking cessation during pregnancy (Freathy *et al.*, 2009); the rs1051730 variant in the nicotinic acetylcholine receptor gene cluster (*CHRNA5–CHRNA3–CHRNA4*) is associated an increased likelihood of continued smoking in pregnancy as the variant affects both smoking quantity and strength of addiction.

Succumbing to a communicable disease is the result of transmission and infectivity of an infectious agent. In addition, genetic factors may play a role in the sensitivity to these diseases; a twin study implicated inherited susceptibility as a major risk factor for tuberculosis in humans (Kallmann and Reisner, 1943) though later studies (van der Eijk *et al.*, 2007) suggest that environmental factors such as the intensity of exposure to the tubercle bacilli and the context of transmission should be given more emphasis when studying inter-individual and population differences.



Given the complexity of nature vs. nurture, how does an individual logically process the information provided by their genome. Interpretation is very difficult and can have implications not only for themselves but also close relatives, partner and medical practitioners. Reasonable questions to ask would be what disease do I have, might I get, might I transmit to my progeny, might I have inherited from my parents, and what drugs should I, can I take? Clinical records can and should be used to correlate genotype to phenotype but a platform is not available to do this yet. One tool that starts to address this issue is an online catalogue of reproducible and common SNP-trait associations from published genome-wide association studies, [www.genome.gov/gwastudies](http://www.genome.gov/gwastudies) (Hindorff *et al.*, 2009). This systematic capture and summarizing of key characteristics and the trait/disease associated SNPs (TASs) underlying a variant provides an opportunity to examine the potential impact of common genetic variants on complex diseases. This catalogue is a significant step in identifying genetic variants likely to be of clinical or public health importance, particularly for developing preventive or therapeutic interventions.

There is an expectation that society will responsibly use the public genome data available. DNA profiling is a technique that is used to aid enquiries associated with criminal offences, disaster victim identification and missing person's investigations. It is possible that the surnames of a male DNA specimen could be guessed at using published genotyping data and public databases of names and haplotypes; studies of the Y-chromosome haplotypes within surnames reveal high levels of co-ancestry among surname cohorts (King and Jobling, 2009). This has theoretical implications for forensics, where an autosomal DNA profile may yield no matches; a Y chromosome profile could suggest one or more surnames that could prioritise a suspect list. This is yet to be used in practice as it works best for intermediate frequency surnames, the link between surname and Y-haplotype is weak for common surnames. While this could be beneficial for forensics, the identity of an individual contributing DNA anonymously for medical research needs to be protected, for example the donors of HapMap.

Is the diagnosis of Down syndrome in an unborn foetus and subsequent termination of the pregnancy a form of eugenics? Some would say yes, and others would defend the parental choice saying it is socially acceptable. Given the controversy surrounding genetic testing, it is not surprising that the public feel that an insurer could request results of genetic tests and if adverse, they could be used to penalise an individual by refusing to hire them or terminating a position. In 2008, the American Congress passed the Genetic Information and Nondiscrimination Act (GINA). This act was limited to employment and health insurance coverage only until a recent extension, SB 559, which now states genetic information cannot be used to discriminate against a person in housing, employment, education, health insurance, life insurance, mortgage lending and elections. In the UK, the Association of British Insurers (ABI) announced that a moratorium on the use of genetic test results by UK insurance companies has been extended until 2017, with a review in 2014. This agreement prevents insurance companies from requesting predictive genetic test results from customers which could be used to deny or increase the cost of cover. Currently the only genetic test that is deemed accurate, clinically reliable and actuarially relevant is Huntington's disease. In addition there is early onset familial Alzheimer's disease and hereditary breast and ovarian cancer, who knows how this list could grow as the knowledge surrounding the general indicators of increased morbidity or mortality grows.

Given that so much DNA is shared between individuals within a species and even between species within a taxonomic clade, is personal genomics the correct term for this new era in biology? Maybe unique genomics would be a more appropriate catchphrase? Regardless of this detail, why would an individual want their genome sequenced? Is it pure curiosity, is it with an expectation that the data can be used to improve quality of life now or is it to have the data available hoping that within the next few years the potential for therapy, diagnosis or treatment would far exceed all current expectations? A medical practitioner would say never do a test if the result will neither inform a patient nor suggest a treatment. This could be taken a step further to say never sequence a genome if the information is not going to benefit an individual or the population by being part of an anonymous collection for research purposes or to be used by forensics. But also, we are not alone, as humans we are a composite of many species (Qin *et al.*, 2010) and should these not be considered too; while an individual's DNA may suggest lactose intolerance, persistent gut bacteria may completely compensate for this. The more complete our understanding of human proteins is the better equipped we will be to understand the functioning of the human body at molecular level. It is possible that all the answers to these questions are already available but buried in the multitude of biological databases currently available, how can we navigate these?

#### 4.4 EMBL-EBI and ELIXIR

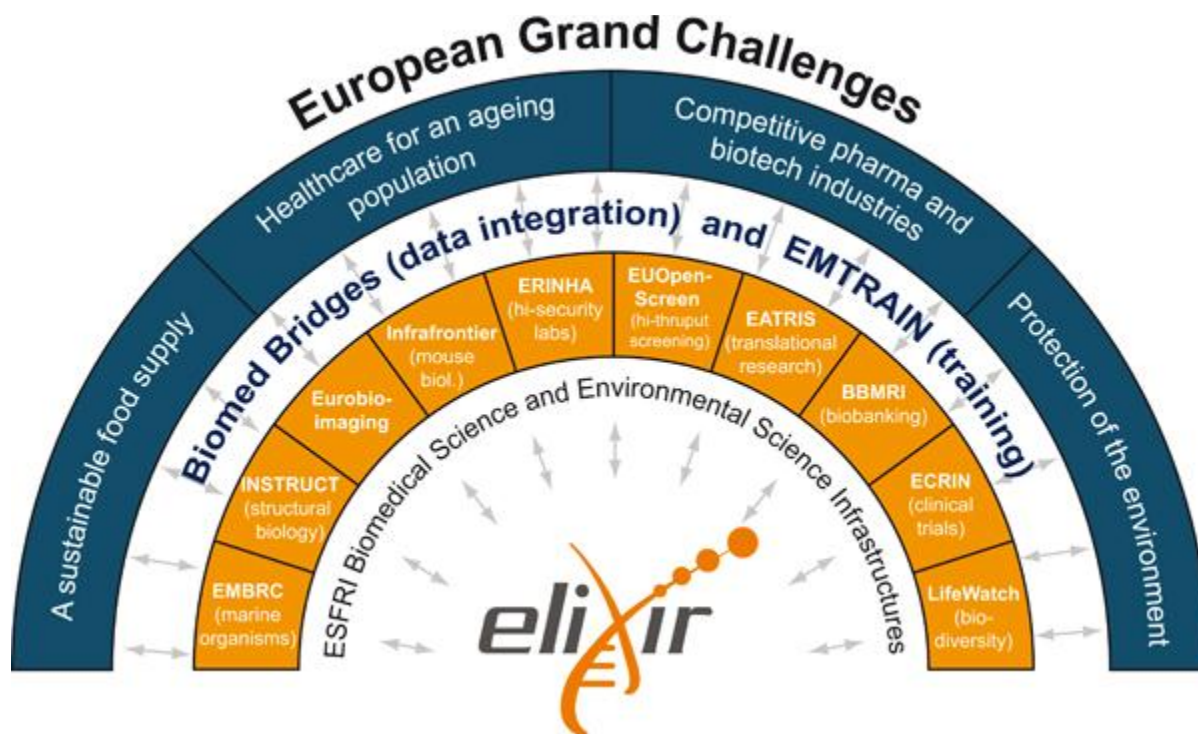
The hundreds of biological databases in Europe range from major core datasets that are maintained by a team of researchers (for example UniProt) to very small specialist collections overseen on a part-time basis by individuals (for example RESID, Farriol-Mathis *et al.*, 2004). The number and variety of these is confusing to the majority of potential users. Who can provide a data structure that will encompass these in a meaningful way to provide a single, transparent interface to a world of widely distributed resources? In addition to this fragmentation of scientific communities, globally life scientists are faced with huge challenges every day that include:

- maintaining open access to biological data to enhance competitiveness and innovation,
- managing the data avalanche,
- integrating the data to reduce fragmentation of effort and research, and
- exploiting new types of data.

Analysis of all the data is now the bottleneck in life-science research. The new ELIXIR project (<http://www.elixir-europe.org/>) is a new and realistic distributed model of data integration and output that will ease this bottleneck by uniting Europe's leading life science organisations in managing and safeguarding the elephantine amounts of data being generated every day by publicly funded research (Fig. 26). It is a pan-European effort, which will be built on existing data resources and services, to safeguard and foster data generated in life-science experiments whose core objective is to ensure that Europe can continue to handle the rapidly growing volume and variety of data. ELIXIR will also provide the facilities necessary for all biologists, from bench scientists to cheminformaticians, to make the most of the enlarging store of information about living systems. Proper management of this information

promotes knowledge-based economic growth, and facilitates the translation of research into innovations that meet global challenges in food security, energy and health. ELIXIR has the potential to enhance the development of Europe-based R&D business in fields ranging from pharmaceuticals to agriculture.

The era of personal genome sequencing is upon us and it is becoming clear that every national health system will need to build and maintain expertise to interpret this information to treat patients accordingly. In conjunction, the personal biological data of each individual must be kept private and secure. An effective informatics infrastructure will take these needs into account and develop comprehensive solutions that will benefit clinicians, organisations and patients equally.

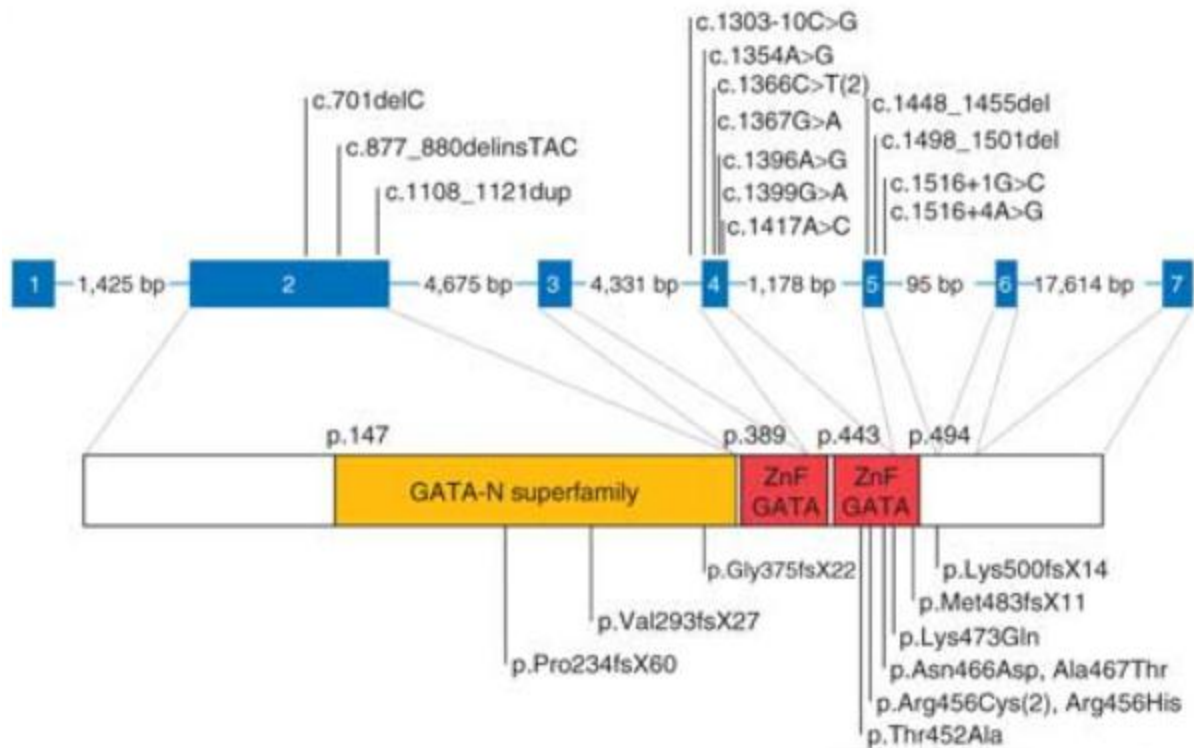


**Figure 24. Understanding how genes and their products impact the health of humans and other species requires interdisciplinary approaches that incorporate a broad spectrum of demanding technologies and resources. ELIXIRs data integration will allow the knowledge generated to be transformed into technical and industrial developments. Image is freely available with EMBL copy write.**

By the end of 2012, ELIXIR will have completed a five-year preparatory phase funded by the EU's Seventh Framework Programme as part of the European Strategy Forum on Research Infrastructures (ESFRI) process. November 2011, EMBL-EBI and the Biotechnology and Biological Sciences Research Council (BBSRC), on behalf of the life science community, learnt that the UK Government has committed £75 million from the Department for Business, Innovation and Skills' Large Facilities Capital Fund (LFCF) for the ELIXIR research infrastructure. The project is also supported by the Medical Research Council, Natural Environment Research Council and The Wellcome Trust. EMBL-EBI will host the future central

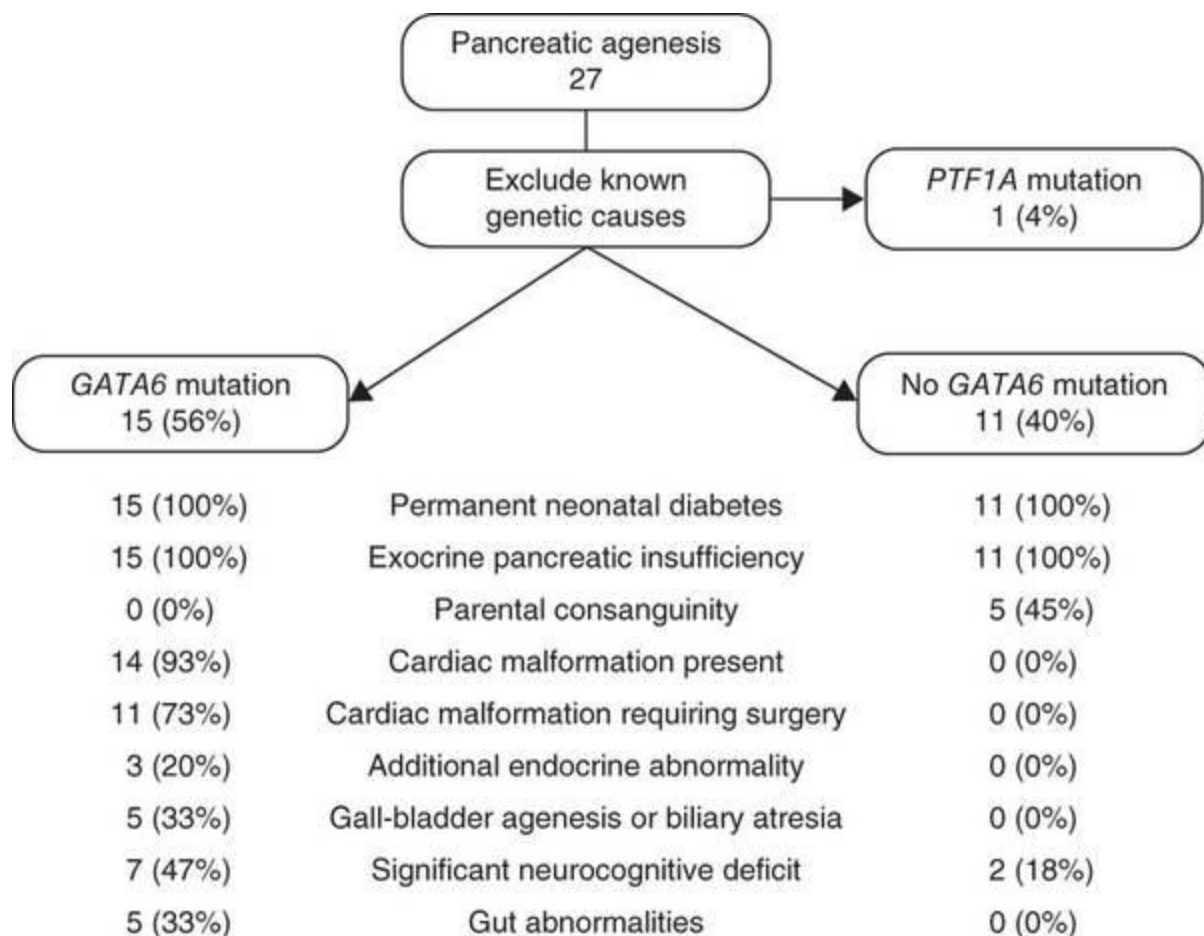
hub of ELIXIR and BBSRC is the leading funding body for its construction. The new funding will allow the construction of ELIXIR's central hub at EMBL-EBI on the Wellcome Trust Genome campus in Hinxton, Cambridge. The hub will be the nerve centre for bioinformatics in Europe, coordinating the delivery of services and user training from several centres of excellence Europe-wide. The hub will also establish a robust pan-European computing infrastructure that can handle the rising tide of life science data; it will open up unlimited opportunities for serendipitous discovery.

An emotive example of just this is a recent publication concerning a discovery linking haploinsufficiency of GATA6 to pancreatic agenesis in humans (Allen *et al.*, 2011). This is a study of 27 individuals who were born to non-diabetic parents and exhibit neonatal diabetes requiring insulin treatment and exocrine pancreatic insufficiency requiring enzyme replacement therapy. These subjects have a complete absence or marked hypoplasia of the pancreas. As 26 of the individuals have consanguineous parents and none have affected siblings it is hypothesised that the agenesis results from de novo heterozygous mutations. To reveal these mutations, the exome of two individuals and their parents was sequenced. The exomes covered 90% of the Consensus coding sequence (CCDS) bases with at least ten-fold coverage so were considered complete. The variants were identified using the Genome Analysis Toolkit (McKenna *et al.*, 2010) which includes filters to remove synonymous variants, variants present in dbSNP (Sayers *et al.*, 2012) or 1000 genome project database (January 2011 release) and those present in the parents. This filtering reduced the number of potentially pathogenic de novo mutations to a single heterozygous GATA6 mutation. Sanger sequencing was used to verify the presence of the mutation in the child and absence in the parents (Fig. 27).



**Figure 25. Genomic positions of the GATA6 mutations on the exon/intron gene structure and the resulting sequence modifications for the resulting protein, taken from Allen *et al.*, 2011. The recognised mutations include; missense mutations at highly conserved residues within the DNA binding domain, mutations at canonical splice sites reducing the strength of these, an insertion causing premature termination of translation and frameshift mutations.**

The region was sequenced in further individuals and a mutation of GATA6 was found in the majority (15/27) of subjects. In addition to pancreatic agenesis, the most common phenotype was congenital cardiac defects (Fig. 28).



**Figure 26. Clinical characteristics of the pancreatic agenesis cohort, taken from Allen *et al.*, 2011.**

This study by the International Pancreatic Agenesis consortium was the culmination of clinical techniques and studies, and bioinformatics studies using several databases including UniProt, CCDS, 1000 Genomes and dbSNP. In the future, it is hoped that ELIXIR would make this combined study approach simpler and more accessible to a wider scientific community. The current efforts to develop replacement therapies for diabetes, which focus on inducing functional endocrine cells from adult somatic cells through the expression of key transcription factors (Zhou *et al.*, 2008) or through recapitulation of human pancreatic development from pluripotent cells (Kroon *et al.*, 2008), may now take a new direction. This newly discovered essential function of the GATA6 transcription factor in human pancreatic organogenesis provides new knowledge that can be used to develop tools for regenerative medicine in diabetes.

With the exponential increase in sequence data, it is becoming vitally important to structure the data in an automatic fashion that will provide a global genome and proteome display of the sequence space. To achieve this, UniProt consortium is reorganising the data to create a UniProtKB core subset. The core will consist of reference and representative proteomes which are a subset of the complete proteome sets and the manually reviewed UniProtKB/Swiss-Prot section. This combination of entries aims to provide both completeness and expert literature curation while eliminating redundancy. Within the

core, it is hoped that users will only find the most relevant and best annotated sequences instead of drowning in reports of redundant sequences. Those redundant sequences will continue to be provided in the UniProtKB non-core subset.

Reference proteomes were made available on the UniProt web and ftp sites in September 2011. It is planned that the UniProtKB core set will be available by the end of 2012 for FTP download, similarity searches and searching. New keywords 'Reference proteome' and 'Representative proteome' have been and will be created, respectively, in addition to the existing 'Complete proteome' keyword, to allow the easy retrieval of these proteome sets.

Provision of the UniProtKB core will allow future use cases for proteomes to be fulfilled:

- Providing a sort criterion for similarity search results,
- Selection of a representative proteome from a set of related proteomes,
- Selection of a representative entry from a protein entry cluster,
- Tracking the evolution of proteome and entry annotation over time,
- Setting annotation priorities; select low scoring UniProtKB entries with large amounts of computationally-mapped literature, and
- Phylogenetic tree builders; require one representative isoform for each gene within a proteome.

Browsing complete proteomes via the website will be easier with a new portal which is currently under development at <http://www.uniprot.org/taxonomy/complete-proteomes>. This new Complete proteome page will provide users with simple and advanced query facilities, proteome information and simple statistics for both complete proteomes and their individual components, such as chromosomes and plasmids as well as download options in several formats.

UniProt consortium is constantly improving the database and web services in terms of accuracy and representation. All user feedback is extremely valuable can be a sent to us via [www.uniprot.org/contact](http://www.uniprot.org/contact) or by email to [help@uniprot.org](mailto:help@uniprot.org).



## 5. CONCLUSIONS

### 5.1 Database improvements

With the expected closure of IPI, Integr8 and Genome reviews and with proteomesDB\_Java code now in production, it is possible to identify and implement further improvements to the database tables. A data collection activity was performed within UniProtKB to determine which tables and which columns within the tables were required by scripts running at EMBL-EBI and SIB. With this information the proteomes team can reduce the schema from 241 tables to just 21 tables and from the remaining tables we can also remove unwanted columns. In comparison to the old schema the reduced database (Appendix 5) includes the following changes.

- Core table DICTIONARY will disappear,
- Core control vocabulary tables as CV\_SCOPE, CV\_STATUS, CV\_IPI\_TYPE and others will disappear,
- Reduced number of fields in PROTEOME, COMPONENT and EMBL\_GENOME tables, and
- REFERENCE table must contain additional fields to store original REFERENCE\_ID. This is necessary to track a reference back in case additional information is needed.

To synchronise these changes in all applications will be difficult so we aim gradually implement the changes one at the time. Currently the reduced schema is loaded into a development schema and all procedures are being tested to guarantee functionality of the proteomes database is maintained.

Along with the reduced schema, it will be possible to simplify the assignment of “component\_type”. Currently there are 11 component\_types that provide mappings from the INSDC genome entries to the constituent UniProtKB entries. This is defined automatically during the EMBL2proteome import procedure and manually by a proteome curator when adding a new proteome to the database via the proteome editor. The wrong choice of component\_type can lead to an empty or incomplete proteome, neither being acceptable. This choice can be simplified from 11 types to just 4. This change will be implemented when the reduced schema is in production.

### 5.2 New data inputs to proteomesDB and UniProtKB

The following section discusses improvements that are required within the proteomes database to maintain the high standard of the database and to ensure the longevity and reliability of the data therein.

#### 5.2.1 Genome collections

After informal discussions held at the Biology of Genomes meeting at Cold Spring Harbor, NY in May of 2008, a Browser genome release agreement was written:

[http://www.ensembl.org/info/about/legal/browser\\_agreement.html](http://www.ensembl.org/info/about/legal/browser_agreement.html). This document stipulates that



Ensembl, NCBI and UCSC browsers (Dreszer *et al.*, 2012) and annotation groups must publically display the same reference genome data for all organisms. To fulfill this agreement the “Genome” database is being developed at NCBI as a repository of all reference genomes. The database (known as Genome collections) is now in production mode and is made publically available at NCBI, <http://www.ncbi.nlm.nih.gov/sites/genome>. All NCBI submitted genomes are available and work is ongoing to complete the backfill of genomes from other INSDC members. A huge factor that must be remembered is that the current scope of Genome collections excludes archaea, viroids and viruses.

The database is dumped nightly at EMBL-EBI with an immensely complicated schema. After much unravelling, a simplified schema of the EMBL-EBI Genome collections is available in Appendix 6 ([http://www.ebi.ac.uk/ena/about/genome\\_collection\\_database](http://www.ebi.ac.uk/ena/about/genome_collection_database)). Within the database, all the genomes can be considered complete at the current time with the current knowledge and with current resources. Data about assembly status is available which ranges from an assembly having no scaffolds, no chromosomes to all sequences in the assembly being complete chromosomes. As well as maintaining this schema, ENA are actively developing a Genome collection submission tool at EMBL-EBI so that all genome submissions from this time onwards fulfill all required criteria for inclusion in Genome collections. The aim is for Genome collections to become a centralised repository for all genomes.

The proteomes team is investigating how we can use this data to identify complete proteomes to ensure that we are using the same reference genomic data as the other major genome databases. We will have to significantly change our input procedures to accurately identify the complete genomes that have full gene model annotation. ENA are working with UniProtKB to realise this goal.

Human, mouse and zebrafish are the only reference genome assemblies agreed on by all sites so far: <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/data.shtml>.

### 5.2.2 BioProjects and BioSamples

NCBI have given advance notice of a dramatic change in their policy regarding assignment of taxonomic identifiers, tax\_ids; they plan to stop assigning tax\_ids in for strain level organisms in 2013. EMBL-EBI was not prepared for this policy change and now need to update all pipelines and procedures to allow an alternative organism identifier. Despite EMBL-EBI encouraging a delay in this policy change NCBI have already stopped issuing strain level nodes to higher plants or animals and shortly will stop assigning strain level ids for microbe isolates. In the future, a submitter of a genome project to INSDC will be issued a BioProject accession number (PRJ prefix), an assembly identifier from the Genome Collections database (GCA prefix) and BioSample accessions for each organism (SAM prefix). BioSample accessions will replace tax\_ids and with these their genomes will be available in INSDC.

NCBI and EBI both have a BioSample database, BioSDn (<http://www.ncbi.nlm.nih.gov/biosample>, Barrett *et al.*, 2012) and BioSDe (<http://www.ebi.ac.uk/biosamples>, Gostev *et al.*, 2012), respectively. These store information about biological samples used in molecular experiments, such as sequencing, gene expression or proteomics. BioSamples have two levels of hierarchy, there are reference biosamples that

are commonly and repeatedly used cell lines that may also be distributed commercially and assay biosamples that are small scale samples used within a laboratory for research and then disposed of. Both databases currently house distinct datasets but there is an agreement for data exchange so in the long term there will be a central repository for all BioSamples.

Unfortunately many UniProtKB production pipelines rely on tax\_ids – TrEMBL production, proteome production and the process that merges 100% identical UniProtKB/TrEMBL entries within a tax\_id (performed to remove unwanted redundancy). In the absence of tax\_ids, UniProtKB must look into using an alternative accession as its replacement which will mean a huge overhaul of very important scripts. This is a massive shift in the fundamental assignment of a unique organism identifier.

### 5.2.3 Ensembl Genome import

Ensembl Genomes (Kersey *et al.*, 2012) were developed to extend the limited taxonomic diversity offered by Ensembl (chordates and three outliers, fruit fly, yeast and worm). Ensembl Genomes displays the genomic DNA and its annotations using data from ENA or those imported from an approved consortium that have a funded responsibility to maintain the annotation of a genome. For those proteomes present in ENA, Ensembl Genomes and UniProtKB complete proteomes display identical data. But, for those derived from consortium annotation, UniProtKB will have an absence of a proteome or possibly an out of date, incomplete proteome from submitter annotation of INSDC entries that is no longer maintained.

For UniProtKB release 2011\_12, a comparison of the species covered by Ensembl Genomes with those in proteomesDB was performed and identified that the following species are completely absent in proteomesDB and would be ideal candidates for an Ensembl Genome import pipeline. These are the species of interest:

- Fungi - *Puccinia triticina*
- Metazoa - *Acyrtosiphon pisum*, *Amphimedon queenslandica*, *Apis mellifera*, *Atta cephalotes*, *Bombyx mori*, *Caenorhabditis japonica*, *Pristionchus pacificus* and *Strongylocentrotus purpuratus*
- Plants - *Oryza glaberrima*
- Protists - *Phytophthora ramorum* and *Pythium ultimum*.

For the Ensembl genome species that exist in ENA, we need to compare the protein sequences of these species from Ensembl Genome with those from ENA. For those that show a difference in protein translations, these are further candidate species for import via Ensembl Genomes as these illustrate the differences between consortium gene models and ENA submitter gene models, respectively. If comparisons demonstrate that the consortium annotation contains improved gene models then the species should be imported into proteomesDB. Proteome annotation would be removed from those UniProtKB entries sourced from ENA and added to those generated by Ensembl genome import

The existing Ensembl import framework can be extended to include the Ensembl genome database. Prototyping has been implemented and tested, and the first four species will be available in UniProt release 2012\_04.

## REFERENCES

The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 467:1061-1073(2010).

Adams M.D., Celniker S.E., Holt R.A., Evans C.A., Gocayne J.D., Amanatides P.G., Scherer S.E., Li P.W., Hoskins R.A., Galle R.F., George R.A., Lewis S.E., Richards S., Ashburner M., Henderson S.N., The genome sequence of *Drosophila melanogaster*. *Science* 287:2185-2195(2000).

Allen H.L., Flanagan S.E., Shaw-Smith C., De Franco E., Akerman I., Caswell R., Ferrer J., Hattersley A.T., Ellard S. GATA6 haploinsufficiency causes pancreatic agenesis in humans. *Nat. Genet.* 44:20-22(2011).

Altshuler D.M., Gibbs R.A., Peltonen L., Altshuler D.M., Gibbs R.A., Peltonen L., Dermitzakis E., Schaffner S.F., Yu F., Peltonen L., Dermitzakis E., Bonnen P.E., Altshuler D.M., Gibbs R.A., de Bakker P.I., Deloukas P., Gabriel S.B., Gwilliam R., Hunt S., Inouye M., Jia X., Palotie A., Parkin M., Whittaker P., Yu F., Chang K., Hawes A., Lewis L.R., Ren Y., Wheeler D., Gibbs R.A., Muzny D.M., Barnes C., Darvishi K., Hurles M., Korn J.M., Kristiansson K., Lee C., McCarroll S.A., Nemesh J., Dermitzakis E., Keinan A., Montgomery S.B., Pollack S., Price A.L., Soranzo N., Bonnen P.E., Gibbs R.A., Gonzaga-Jauregui C., Keinan A., Price A.L., Yu F., Anttila V., Brodeur W., Daly M.J., Leslie S., McVean G., Moutsianas L., Nguyen H., Schaffner S.F., Zhang Q., Ghorri M.J., McGinnis R., McLaren W., Pollack S., Price A.L., Schaffner S.F., Takeuchi F., Grossman S.R., Shlyakhter I., Hostetter E.B., Sabeti P.C., Adebamowo C.A., Foster M.W., Gordon D.R., Licinio J., Manca M.C., Marshall P.A., Matsuda I., Ngare D., Wang V.O., Reddy D., Rotimi C.N., Royal C.D., Sharp R.R., Zeng C., Brooks L.D., McEwen J.E. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52-58(2010).

Amid C., Birney E., Bower L., Cerdano-Tarraga A., Cheng Y., Cleland I., Faruque N., Gibson R., Goodgame N., Hunter C., Jang M., Leinonen R., Liu X., Oisel A., Pakseresht N., Plaister S., Radhakrishnan R., Reddy K., Riviere S., Rossello M., Senf A., Smirnov D., Ten Hoopen P., Vaughan D., Vaughan R., Zalunin V., Cochrane G. Major submissions tool developments at the European nucleotide archive. *Nucleic Acids Res.* 40:D43-D47(2012).

Anderson S., Bankier A.T., Barrell B.G., de Bruijn M.H., Coulson A.R., Drouin J., Eperon I.C., Nierlich D.P., Roe B.A., Sanger F., Schreier P.H., Smith A.J., Staden R., Young I.G. Sequence and organization of the human mitochondrial genome. *Nature* 290:457-465(1981).

Andersson S.G., Zomorodipour A., Andersson J.O., Sicheritz-Ponten T., Alsmark U.C., Podowski R.M., Naslund A.K., Eriksson A.S., Winkler H.H., Kurland C.G. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396:133-140(1998).

Barrett T., Troup D.B., Wilhite S.E., Ledoux P., Evangelista C., Kim I.F., Tomashevsky M., Marshall K.A., Phillippy K.H., Sherman P.M., Muertter R.N., Holko M., Ayanbule O., Yefanov A., Soboleva A. NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Res.* 39:D1005-D1010(2011).

Barrett T., Clark K., Gevorgyan R., Gorelenkov V., Gribov E., Karsch-Mizrachi I., Kimelman M., Pruitt K.D., Resenchuk S., Tatusova T., Yaschenko E., Ostell J. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.* 40:D57-D63(2012).

Bateman, A., *et al.* (in press) RNAcentral: a vision for an international database of RNA sequences, *RNA*.

Benson D.A., Karsch-Mizrachi I., Clark K., Lipman D.J., Ostell J., Sayers E.W. GenBank. *Nucleic Acids Res.* 40:D48-D53(2012).

Bentley D.R., Balasubramanian S., Swerdlow H.P., Smith G.P., Milton J., Brown C.G., Hall K.P., Evers D.J., Barnes C.L., Bignell H.R., Boutell J.M., Bryant J., Carter R.J., Keira Cheetham R., Cox A.J., Ellis D.J., Flatbush M.R., Gormley N.A., Humphray S.J., Irving L.J., Karbelashvili M.S., Kirk S.M., Li H., Liu X., Maisinger K.S., Murray L.J., Obradovic B., Ost T., Parkinson M.L., Pratt M.R., Rasolonjatovo I.M., Reed M.T., Rigatti R., Rodighiero C., Ross M.T., Sabot A., Sankar S.V., Scally A., Schroth G.P., Smith M.E., Smith V.P., Spiridou A., Torrance P.E., Tzonev S.S., Vermaas E.H., Walter K., Wu X., Zhang L., Alam M.D., Anastasi C., Aniebo I.C., and more. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53-59(2008).

Bochukova E.G., Huang N., Keogh J., Henning E., Purmann C., Blaszczyk K., Saeed S., Hamilton-Shield J., Clayton-Smith J., O'Rahilly S., Hurles M.E., Farooqi I.S. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* 463:666-670(2010).

Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32:D267-D270(2004).

Boeckmann B., Robinson-Rechavi M., Xenarios I., Dessimoz C., Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief. Bioinform.* 12:423-435(2011).

Brazma A., Hingamp P., Quackenbush J., Sherlock G., Spellman P., Stoeckert C., Aach J., Ansorge W., Ball C.A., Causton H.C., Gaasterland T., Glenisson P., Holstege F.C., Kim I.F., Markowitz V., Matese J.C., Parkinson H., Robinson A., Sarkans U., Schulze-Kremer S., Stewart J., Taylor R., Vilo J., Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29:365-371(2001).

Brooksbank C., Cameron G., Thornton J. The European Bioinformatics Institute's data resources. *Nucleic Acids Res.* 38:D17-D25(2010).

Brunak S., Danchin A., Hattori M., Nakamura H., Shinozaki K., Matise T., Preuss D. Nucleotide sequence database policies. *Science* 298:1333-1333(2002).

Carlton J.M., Hirt R.P., Silva J.C., Delcher A.L., Schatz M., Zhao Q., Wortman J.R., Bidwell S.L., Alsmark U.C., Besteiro S., Sicheritz-Ponten T., Noel C.J., Dacks J.B., Foster P.G., Simillion C., Van de Peer Y., Miranda-Saavedra D., Barton G.J., Westrop G.D., Muller S., Dessi D., Fiori P.L., Ren Q., Paulsen I., Zhang H., Bastida-Corcuera F.D., Simoes-Barbosa A., Brown M.T., Hayes R.D., Mukherjee M., Okumura C.Y., Schneider R., Smith A.J., Vanacova S., Villalvazo M., Haas B.J., Pertea M., Feldblyum T.V., Utterback T.R.,

Shu C.L., Osoegawa K., de Jong P.J., Hrdy I., Horvathova L., Zubacova Z., Dolezal P., Malik S.B., Logsdon J.M. Jr., Henze K., Gupta A., Wang C.C., Dunne R.L., Upcroft J.A., Upcroft P., White O., Salzberg S.L., Tang P., Chiu C.H., Lee Y.S., Embley T.M., Coombs G.H., Mottram J.C., Tachezy J., Fraser-Liggett C.M., Johnson P.J. "Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* 315:207-212(2007).

Castellana N.E., Payne S.H., Shen Z., Stanke M., Bafna V., Briggs S.P. Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc. Natl. Acad. Sci. U.S.A.* 105:21034-21038(2008).

Cerdeno-Tarraga A.M., Patrick S., Crossman L.C., Blakely G., Abratt V., Lennard N., Poxton I., Duerden B., Harris B., Quail M.A., Barron A., Clark L., Corton C., Doggett J., Holden M.T., Larke N., Line A., Lord A., Norbertczak H., Ormond D., Price C., Rabbino-witsch E., Woodward J., Barrell B., Parkhill J. Extensive DNA inversions in the *B. fragilis* genome control variable gene expression. *Science* 307:1463-1465(2005).

Chen C., Natale D.A., Finn R.D., Huang H., Zhang J., Wu C.H., Mazumder R. Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLoS ONE* 6:E18910-E18910(2011).

Chen F., Mackey A.J., Stoeckert C.J. Jr., Roos D.S. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34:D363-D368(2006).

Clark R.M., Dalgliesh G.L., Endres D., Gomez M., Taylor J., Bidichandani S.I. Expansion of GAA triplet repeats in the human genome: unique origin of the FRDA mutation at the center of an Alu. *Genomics* 83:373-383(2004).

Cochrane G., Karsch-Mizrachi I., Nakamura Y. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* 39:D15-D18(2011).

Cole S.T., Eglmeier K., Parkhill J., James K.D., Thomson N.R., Wheeler P.R., Honore N., Garnier T., Churcher C., Harris D., Mungall K., Basham D., Brown D., Chillingworth T., Connor R., Davies R.M., Devlin K., Duthoy S., Feltwell T., Fraser A., Hamlin N., Holroyd S., Hornsby T., Jagels K., Lacroix C., Maclean J., Moule S., Murphy L., Oliver K., Quail M.A., Rajandream M.A., Rutherford K.M., Rutter S., Seeger K., Simon S., Simmonds M., Skelton J., Squares R., Squares S., Stevens K., Taylor K., Whitehead S., Woodward J.R., Barrell B.G. Massive gene decay in the leprosy bacillus. *Nature* 409:1007-1011(2001).

Croft D., O'Kelly G., Wu G., Haw R., Gillespie M., Matthews L., Caudy M., Garapati P., Gopinath G., Jassal B., Jupe S., Kalatskaya I., Mahajan S., May B., Ndegwa N., Schmidt E., Shamovsky V., Yung C., Birney E., Hermjakob H., D'Eustachio P., Stein L. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* 39:D691-D697(2011).

Daley D.O., Whelan J. Why genes persist in organelle genomes. *Genome Biol.* 6:110-110(2005).

D'Auria G., Jimenez-Hernandez N., Peris-Bondia F., Moya A., Latorre A. *Legionella pneumophila* pangenome reveals strain-specific virulence factors. *BMC Genomics* 11:181-181(2010).

Donnelly K, SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Techno Inform* 2006, 121:79-90.

Dreszer T.R., Karolchik D., Zweig A.S., Hinrichs A.S., Raney B.J., Kuhn R.M., Meyer L.R., Wong M., Sloan C.A., Rosenbloom K.R., Roe G., Rhead B., Pohl A., Malladi V.S., Li C.H., Learned K., Kirkup V., Hsu F., Harte R.A., Guruvadoo L., Goldman M., Giardine B.M., Fujita P.A., Diekhans M., Cline M.S., Clawson H., Barber G.P., Haussler D., James Kent W. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.* 40:D918-D923(2012).

Easton D.F., Ford D., Bishop D.T. Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.* 56:265-271(1995).

van der Eijk E.A., van de Vosse E., Vandenbroucke J.P., van Dissel J.T. Heredity versus environment in tuberculosis in twins: the 1950s United Kingdom Prophit Survey Simonds and Comstock revisited. *Am. J. Respir. Crit. Care Med.* 176:1281-1288(2007).

Ellervik C., Tybjaerg-Hansen A., Appleyard M., Sillesen H., Boysen G., Nordestgaard B.G. Hereditary hemochromatosis genotypes and risk of ischemic stroke. *Neurology* 68:1025-1031(2007).

Farriol-Mathis N., Garavelli J.S., Boeckmann B., Duvaud S., Gasteiger E., Gateau A., Veuthey A.L., Bairoch A. Annotation of post-translational modifications in the Swiss-Prot knowledge base. *Proteomics* 4:1537-1550(2004).

Fauquet C.M., Mayo M.A., Maniloff J., Desselberger U., Ball L.A. Virus Taxonomy: VIIIth Report of the International Committee on Taxonomy of Viruses, Elsevier Academic Press(2005).

Federhen S. The NCBI Taxonomy database. *Nucleic Acids Res.* 40:D136-D143(2012).

Fitch W.M. Distinguishing homologous from analogous proteins. *Syst. Zool.* 19:99-113(1970).

Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.F., Dougherty B.A., Merrick J.M., Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd. *Science* 269:496-512(1995).

Fleischmann W., Moller S., Gateau A., Apweiler R. A novel method for automatic functional annotation of proteins. *Bioinformatics* 15:228-233(1999).

Flicek P., Amode M.R., Barrell D., Beal K., Brent S., Carvalho-Silva D., Clapham P., Coates G., Fairley S., Fitzgerald S., Gil L., Gordon L., Hendrix M., Hourlier T., Johnson N., Kahari A.K., Keefe D., Keenan S., Kinsella R., Komorowska M., Koscielny G., Kulesha E., Larsson P., Longden I., McLaren W., Muffato M., Overduin B., Pignatelli M., Pritchard B., Riat H.S., Ritchie G.R., Ruffier M., Schuster M., Sobral D., Tang Y.A., Taylor K., Trevanion S., Vandrovcova J., White S., Wilson M., Wilder S.P., Aken B.L., Birney E., Cunningham F., Dunham I., Durbin R., Fernandez-Suarez X.M., Harrow J., Herrero J., Hubbard T.J., Parker A., Proctor G., Spudich G., Vogel J., Yates A., Zadissa A., Searle S.M. Ensembl 2012. *Nucleic Acids Res.* 40:D84-D90(2012).

Forslund K., Pekkari I., Sonnhammer E.L. Domain architecture conservation in orthologs. BMC Bioinformatics 12:326-326(2011a).

Forslund K., Schreiber F., Thanintorn N., Sonnhammer E.L. OrthoDisease: tracking disease gene orthologs across 100 species. Brief. Bioinform. 0:0-0(2011b).

Fraser C.M., Gocayne J.D., White O., Adams M.D., Clayton R.A., Fleischmann R.D., Bult C.J., Kerlavage A.R., Sutton G., Kelley J.M., Fritchman R.D., Weidman J.F., Small K.V., Sandusky M., Fuhrmann J., Nguyen D., Utterback T.R., Saudek D.M., Phillips C.A., Merrick J.M., Tomb J.F., Dougherty B.A., Bott K.F., Hu P.C., Lucier T.S., Peterson S.N., Smith H.O., Hutchison C.A. III, Venter J.C. The minimal gene complement of *Mycoplasma genitalium*. Science 270:397-403(1995).

Freathy R.M., Ring S.M., Shields B., Galobardes B., Knight B., Weedon M.N., Smith G.D., Frayling T.M., Hattersley A.T. A common genetic variant in the 15q24 nicotinic acetylcholine receptor gene cluster (CHRNA5-CHRNA3-CHRNA4) is associated with a reduced ability of women to quit smoking in pregnancy. Hum. Mol. Genet. 18:2922-2927(2009).

Gardner P.P., Daub J., Tate J., Moore B.L., Osuch I.H., Griffiths-Jones S., Finn R.D., Nawrocki E.P., Kolbe D.L., Eddy S.R., Bateman A. Rfam: Wikipedia, clans and the 'decimal' release. Nucleic Acids Res. 39:D141-D145(2011).

The Gene Ontology Consortium, The Gene Ontology: enhancements for 2011. Nucleic Acids Res. 40:D559-D564(2012).

Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J. Hered. 100:659-674(2009).

Gerlach D., Kriventseva E.V., Rahman N., Vejnar C.E., Zdobnov E.M. miROrtho: computational survey of microRNA genes. Nucleic Acids Res. 37:D111-D117(2009).

Goffeau A., Barrell B.G., Bussey H., Davis R.W., Dujon B., Feldmann H., Galibert F., Hoheisel J.D., Jacq C., Johnston M., Louis E.J., Mewes H.W., Murakami Y., Philippsen P., Tettelin H., Oliver S.G., Life with 6000 genes. Science 274:546, 563-546, 567(1996).

Gostev M., Faulconbridge A., Brandizi M., Fernandez-Banet J., Sarkans U., Brazma A., Parkinson H. The BioSample Database (BioSD) at the European Bioinformatics Institute. Nucleic Acids Res. 40:D64-D70 (2012).

Haas B.J., Kamoun S., Zody M.C., Jiang R.H., Handsaker R.E., Cano L.M., Grabherr M., Kodira C.D., Raffaele S., Torto-Alalibo T., Bozkurt T.O., Ah-Fong A.M., Alvarado L., Anderson V.L., Armstrong M.R., Avrova A., Baxter L., Beynon J., Boevink P.C., Bollmann S.R., Bos J.I., Bulone V., Cai G., Cakir C., Carrington J.C., Chawner M., Conti L., Costanzo S., Ewan R., Fahlgren N., Fischbach M.A., Fugelstad J., Gilroy E.M., Gnerre S., Green P.J., Grenville-Briggs L.J., Griffith J., Grunwald N.J., Horn K., Horner N.R., Hu C.H., Huitema E., Jeong D.H., Jones A.M., Jones J.D., Jones R.W., Karlsson E.K., Kunjeti S.G., Lamour K., Liu Z., Ma L., Maclean D., Chibucos M.C., McDonald H., McWalters J., Meijer H.J., Morgan W., Morris



P.F., Munro C.A., O'Neill K., Ospina-Giraldo M., Pinzon A., Pritchard L., Ramsahoye B., Ren Q., Restrepo S., Roy S., Sadanandom A., Savidor A., Schornack S., Schwartz D.C., Schumann U.D., Schwessinger B., Seyer L., Sharpe T., Silvar C., Song J., Studholme D.J., Sykes S., Thines M., van de Vondervoort P.J., Phuntumart V., Wawra S., Weide R., Win J., Young C., Zhou S., Fry W., Meyers B.C., van West P., Ristaino J., Govers F., Birch P.R., Whisson S.C., Judelson H.S., Nusbaum C. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* 461:393-398(2009).

Henderson M. Obese children taken off at-risk register after genes found to be at fault. *The Times* (2009) <http://www.timesonline.co.uk/tol/news/uk/health/article6946615.ece>

Henricson A., Forslund K., Sonnhammer E.L. Orthology confers intron position conservation. *BMC Genomics* 11:412-412(2010).

Hiller N.L., Janto B., Hogg J.S., Boissy R., Yu S., Powell E., Keefe R., Ehrlich N.E., Shen K., Hayes J., Barbadora K., Klimke W., Dernovoy D., Tatusova T., Parkhill J., Bentley S.D., Post J.C., Ehrlich G.D., Hu F.Z. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. *J. Bacteriol.* 189:8186-8195(2007).

Hindorff L.A., Sethupathy P., Junkins H.A., Ramos E.M., Mehta J.P., Collins F.S., Manolio T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106:9362-9367(2009).

Hogg J.S., Hu F.Z., Janto B., Boissy R., Hayes J., Keefe R., Post J.C., Ehrlich G.D. Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol.* 8:R103-R103(2007).

Huerta-Cepas J., Dopazo J., Huynen M.A., Gabaldon T. Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. *Brief. Bioinform.* 12:442-448(2011a).

Huerta-Cepas J., Capella-Gutierrez S., Pryszcz L.P., Denisov I., Kormes D., Marcet-Houben M., Gabaldon T. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.* 39:D556-D560(2011b).

Huijgen R., Kindt I., Fouchier S.W., Defesche J.C., Hutten B.A., Kastelein J.J., Vissers M.N. Functionality of sequence variants in the genes coding for the low-density lipoprotein receptor and apolipoprotein B in individuals with inherited hypercholesterolemia. *Hum. Mutat.* 31:752-760(2010).

Hulo C., de Castro E., Masson P., Bougueleret L., Bairoch A., Xenarios I., Le Mercier P. ViralZone: a knowledge resource to understand virus diversity. *Nucleic. Acids. Res.* 39:D576-D582(2011).

Ikeda H., Ishikawa J., Hanamoto A., Shinose M., Kikuchi H., Shiba T., Sakaki Y., Hattori M., Omura S. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat. Biotechnol.* 21:526-531(2003).

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 431:931-945(2004).

International Statistical Classification of Diseases and Health Related Problems In (The) ICD-10. Second Edition edition. WHO Press, Geneva, <http://www.who.int/en/>

Jaffe J.D., Stange-Thomann N., Smith C., DeCaprio D., Fisher S., Butler J., Calvo S., Elkins T., FitzGerald M.G., Hafez N., Kodira C.D., Major J., Wang S., Wilkinson J., Nicol R., Nusbaum C., Birren B., Berg H.C., Church G.M. The complete genome and proteome of *Mycoplasma mobile*. *Genome Res.* 14:1447-1461(2004a).

Jaffe J.D., Berg H.C., Church G.M. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* 4:59-77(2004b).

Jain E., Bairoch A., Duvaud S., Phan I., Redaschi N., Suzek B.E., Martin M.J., McGarvey P., Gasteiger E. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* 10:136-136(2009).

Kallmann F.J., Reisner D. Twin studies on the significance of genetic factors in tuberculosis. *Am Rev Tuberc* 47:549-574(1943).

Kapushesky M., Adamusiak T., Burdett T., Culhane A., Farne A., Filippov A., Holloway E., Klebanov A., Kryvych N., Kurbatova N., Kurnosov P., Malone J., Melnichuk O., Petryszak R., Pultsin N., Rustici G., Tikhonov A., Travillian R.S., Williams E., Zorin A., Parkinson H., Brazma A. Gene Expression Atlas update-- a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 40:D1077-D1081(2012).

Karsch-Mizrachi I., Nakamura Y., Cochrane G. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* 40:D33-D37(2012).

Kerrien S., Aranda B., Breuza L., Bridge A., Broackes-Carter F., Chen C., Duesbury M., Dumousseau M., Feuermann M., Hinz U., Jandrasits C., Jimenez R.C., Khadake J., Mahadevan U., Masson P., Pedruzzi I., Pfeifferberger E., Porras P., Raghunath A., Roechert B., Orchard S., Hermjakob H. The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 40:D841-D846(2012).

Kersey P.J., Duarte J., Williams A., Karavidopoulou Y., Birney E., Apweiler R. The International Protein Index: an integrated database for proteomics experiments. *Proteomics* 4:1985-1988(2004).

Kersey P.J., Bower L., Morris L., Horne A., Petryszak R., Kanz C., Kanapin A., Das U., Michoud K., Phan I., Gattiker A., Kulikova T., Faruque N., Duggan K., McLaren P., Reimholz B., Duret L., Penel S., Reuter I., Apweiler R. Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.* 33:D297-D302(2005).

Kersey P.J., Staines D.M., Lawson D., Kulesha E., Derwent P., Humphrey J.C., Hughes D.S., Keenan S., Kerhornou A., Koscielny G., Langridge N., McDowall M.D., Megy K., Maheswari U., Nuhn M., Paulini M.,

Pedro H., Toneva I., Wilson D., Yates A., Birney E. Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.* 40:D91-D97(2012).

King T.E., Jobling M.A. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. *Trends Genet.* 25:351-360(2009).

Klimke W., Agarwala R., Badretdin A., Chetvernin S., Ciufo S., Fedorov B., Kiryutin B., O'Neill K., Resch W., Resenchuk S., Schafer S., Tolstoy I., Tatusova T. The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res.* 37:D216-D223(2009).

Kodama Y., Mashima J., Kaminuma E., Gojobori T., Ogasawara O., Takagi T., Okubo K., Nakamura Y. The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments. *Nucleic. Acids. Res.* 40:D38-D42(2012a).

Kodama Y., Shumway M., Leinonen R. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* 40:D54-D56(2012b).

Kretschmann E., Fleischmann W., Apweiler R. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics* 17:920-926(2001).

Krinos C.M., Coyne M.J., Weinacht K.G., Tzianabos A.O., Kasper D.L., Comstock L.E. Extensive surface diversity of a commensal microorganism by multiple DNA inversions. *Nature* 414:555-558(2001).

Kroon E., Martinson L.A., Kadoya K., Bang A.G., Kelly O.G., Eliazer S., Young H., Richardson M., Smart N.G., Cunningham J., Agulnick A.D., D'Amour K.A., Carpenter M.K., Baetge E.E. Pancreatic endoderm derived from human embryonic stem cells generates glucose-responsive insulin-secreting cells in vivo. *Nat. Biotechnol.* 26:443-452(2008).

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., and more. Initial sequencing and analysis of the human genome. *Nature* 409:860-921(2001).

Leinonen R., Diez F.G., Binns D., Fleischmann W., Lopez R., Apweiler R. UniProt archive. *Bioinformatics* 20:3236-3237(2009).

Levy S., Sutton G., Ng P.C., Feuk L., Halpern A.L., Walenz B.P., Axelrod N., Huang J., Kirkness E.F., Denisov G., Lin Y., MacDonald J.R., Pang A.W., Shago M., Stockwell T.B., Tsiamouri A., Bafna V., Bansal V., Kravitz S.A., Busam D.A., Beeson K.Y., McIntosh T.C., Remington K.A., Abril J.F., Gill J., Borman J., Rogers Y.H., Frazier M.E., Scherer S.W., Strausberg R.L., Venter J.C. The diploid genome sequence of an individual human. *PLoS Biol.* 5:E254-E254(2007).

Lima T., Auchincloss A.H., Coudert E., Keller G., Michoud K., Rivoire C., Bulliard V., de Castro E., Lachaize C., Baratin D., Phan I., Bougueleret L., Bairoch A. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res.* 37:D471-D478(2009).

Linwood J., Minter D. Beginning Hibernate. Apress. 400pp (2010) ISBN 1430228504.

Llopart A., Comeron J.M., Brunet F.G., Lachaise D., Long M. Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. *Proc. Natl. Acad. Sci. U.S.A.* 99:8121-8126(2002).

Malone J., Holloway E., Adamusiak T., Kapushesky M., Zheng J., Kolesnikov N., Zhukova A., Brazma A., Parkinson H. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 26:1112-1118(2010).

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytzky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M.A. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297-1303(2010).

McQuilton P., St Pierre S.E., Thurmond J. FlyBase 101 - the basics of navigating FlyBase. *Nucleic. Acids. Res.* 40:D706-D714(2012).

Mi H., Dong Q., Muruganujan A., Gaudet P., Lewis S., Thomas P.D. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.* 38:D204-D210(2010).

Mottaz A., Yip Y.L., Ruch P., Veuthey A.L. Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC Bioinformatics* 9:S3-S3(2008).

Nakabachi A., Yamashita A., Toh H., Ishikawa H., Dunbar H.E., Moran N.A., Hattori M. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* 314:267-267(2006).

Natale D.A., Vinayaka C.R., Wu C.H. Large-scale, classification-driven, rule-based functional annotation of proteins. In: Subramaniam S, editor. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. 2004. Bioinformatics Volume. John Wiley & Sons, Ltd, NY.

Nehrt N.L., Clark W.T., Radivojac P., Hahn M.W. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput. Biol.* 7:E1002073-E1002073(2011).

Neil A., Cooper J., Betteridge J., Capps N., McDowell I., Durrington P., Seed M., Humphries S.E. Reductions in all-cause, cancer, and coronary mortality in statin-treated patients with heterozygous familial hypercholesterolaemia: a prospective registry study. *Eur. Heart J.* 29:2625-2633(2008).

Nelson S.J., Schopen M., Savage A.G., Schulman J.L., Arluk N. The MeSH translation maintenance system: structure, interface design, and implementation. *Stud. Health. Technol. Inform.* 107:67-69(2004).

Osoegawa K., Mammoser A.G., Wu C., Frengen E., Zeng C., Catanese J.J., de Jong P.J. A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* 11:483-496(2001).

Ostlund G., Schmitt T., Forslund K., Kostler T., Messina D.N., Roopra S., Frings O., Sonnhammer E.L. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38:D196-D203(2010).

Pagani I., Liolios K., Jansson J., Chen I.M., Smirnova T., Nosrat B., Markowitz V.M., Kyrpides N.C. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 40:D571-D579(2012).

Parkinson H., Sarkans U., Kolesnikov N., Abeygunawardena N., Burdett T., Dylag M., Emam I., Farne A., Hastings E., Holloway E., Kurbatova N., Lukk M., Malone J., Mani R., Pilicheva E., Rustici G., Sharma A., Williams E., Adamusiak T., Brandizi M., Sklyar N., Brazma A. ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.* 39:D1002-D1004(2011).

Patient S., Wieser D., Kleen M., Kretschmann E., Jesus Martin M., Apweiler R. UniProtJAPI: a remote API for accessing UniProtKB data. *Bioinformatics* 24:1321-1322(2008).

Peterson J., Garges S., Giovanni M., McInnes P., Wang L., Schloss J.A., Bonazzi V., McEwen J.E., Wetterstrand K.A., Deal C., Baker C.C., Di Francesco V., Howcroft T.K., Karp R.W., Lunsford R.D., Wellington C.R., Belachew T., Wright M., Giblin C., David H., Mills M., Salomon R., Mullins C., Akolkar B., Begg L., Davis C., Grandison L., Humble M., Khalsa J., Little A.R., Peavy H., Pontzer C., Portnoy M., Sayre M.H., Starke-Reed P., Zakhari S., Read J., Watson B., Guyer M. The NIH Human Microbiome Project. *Genome Res.* 19:2317-2323(2009).

du Plessis L., Skunca N., Dessimoz C. The what, where, how and why of gene ontology--a primer for bioinformaticians. *Brief. Bioinform.* 12:723-735(2011).

Proost S., Van Bel M., Sterck L., Billiau K., Van Parys T., Van de Peer Y., Vandepoele K. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell* 21:3718-3731(2009).

Pruitt K.D., Harrow J., Harte R.A., Wallin C., Diekhans M., Maglott D.R., Searle S., Farrell C.M., Loveland J.E., Ruff B.J., Hart E., Suner M.M., Landrum M.J., Aken B., Ayling S., Baertsch R., Fernandez-Banet J., Cherry J.L., Curwen V., Dicuccio M., Kellis M., Lee J., Lin M.F., Schuster M., Shkeda A., Amid C., Brown G., Dukhanina O., Frankish A., Hart J., Maidak B.L., Mudge J., Murphy M.R., Murphy T., Rajan J., Rajput B., Riddick L.D., Snow C., Steward C., Webb D., Weber J.A., Wilming L., Wu W., Birney E., Haussler D., Pruitt K.D., Tatusova T., Brown G.R., Maglott D.R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* 40:D130-D135(2012).

Hubbard T., Ostell J., Durbin R., Lipman D. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* 19:1316-1323(2009).

Qin J., Li R., Raes J., Arumugam M., Burgdorf K.S., Manichanh C., Nielsen T., Pons N., Levenez F., Yamada T., Mende D.R., Li J., Xu J., Li S., Li D., Cao J., Wang B., Liang H., Zheng H., Xie Y., Tap J., Lepage P., Bertalan M., Batto J.M., Hansen T., Le Paslier D., Linneberg A., Nielsen H.B., Pelletier E., Renault P., Sicheritz-Ponten T., Turner K., Zhu H., Yu C., Li S., Jian M., Zhou Y., Li Y., Zhang X., Li S., Qin N., Yang H., Wang J., Brunak S., Dore J., Guarner F., Kristiansen K., Pedersen O., Parkhill J., Weissenbach J., Bork P.,

Ehrlich S.D., Wang J. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59-65(2010).

Rasko D.A., Rosovitz M.J., Myers G.S., Mongodin E.F., Fricke W.F., Gajer P., Crabtree J., Sebaihia M., Thomson N.R., Chaudhuri R., Henderson I.R., Sperandio V., Ravel J. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J. Bacteriol.* 190:6881-6893(2008).

Ren D., Zhou Y., Morris D., Li M., Li Z., Rui L. Neuronal SH2B1 is essential for controlling energy and glucose homeostasis. *J. Clin. Invest.* 117:397-406(2007).

Sanger F., Air G.M., Barrell B.G., Brown N.L., Coulson A.R., Fiddes C.A., Hutchison C.A., Slocombe P.M., Smith M., Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265:687-695(1977).

Saunders A.M., Strittmatter W.J., Schmechel D., George-Hyslop P.H., Pericak-Vance M.A., Joo S.H., Rosi B.L., Gusella J.F., Crapper-MacLachlan D.R., Alberts M.J. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* 43:1467-1472(1993).

Sayers E.W., Barrett T., Benson D.A., Bolton E., Bryant S.H., Canese K., Chetvernin V., Church D.M., Dicuccio M., Federhen S., Feolo M., Fingerman I.M., Geer L.Y., Helmberg W., Kapustin Y., Krasnov S., Landsman D., Lipman D.J., Lu Z., Madden T.L., Madej T., Maglott D.R., Marchler-Bauer A., Miller V., Karsch-Mizrachi I., Ostell J., Panchenko A., Phan L., Pruitt K.D., Schuler G.D., Sequeira E., Sherry S.T., Shumway M., Sirotkin K., Slotta D., Souvorov A., Starchenko G., Tatusova T.A., Wagner L., Wang Y., Wilbur W.J., Yaschenko E., Ye J. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 40:D13-D25(2012).

Schmitt T., Messina D.N., Schreiber F., Sonnhammer E.L. Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Brief. Bioinform.* 12:485-488(2011).

Seeburg P.H., Shine J., Martial J.A., Ullrich A., Goodman H.M., Baxter J.D. Nucleotide sequence of a human gene coding for a polypeptide hormone. *Trans. Assoc. Am. Physicians* 90:109-116(1977).

Serruto D., Serino L., Masignani V., Pizza M. Genome-based approaches to develop vaccines against bacterial pathogens. *Vaccine* 27:3245-3250(2009).

Sobreira N.L., Cirulli E.T., Avramopoulos D., Wohler E., Oswald G.L., Stevens E.L., Ge D., Shianna K.V., Smith J.P., Maia J.M., Gumbs C.E., Pevsner J., Thomas G., Valle D., Hoover-Fong J.E., Goldstein D.B. Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genet.* 6:E1000991-E1000991(2010).

Sonnhammer E.L., Koonin E.V. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* 18:619-620(2002).

Suttle C.A. Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.* 5:801-812(2007).

Suzek B.E., Huang H., McGarvey P., Mazumder R., Wu C.H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23:1282-1288(2007).

Tatusov R.L., Fedorova N.D., Jackson J.D., Jacobs A.R., Kiryutin B., Koonin E.V., Krylov D.M., Mazumder R., Mekhedov S.L., Nikolskaya A.N., Rao B.S., Smirnov S., Sverdlov A.V., Vasudevan S., Wolf Y.I., Yin J.J., Natale D.A. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41-41(2003).

Tettelin H., Masignani V., Cieslewicz M.J., Donati C., Medini D., Ward N.L., Angiuoli S.V., Crabtree J., Jones A.L., Durkin A.S., Deboy R.T., Davidsen T.M., Mora M., Scarselli M., Margarit y Ros I., Peterson J.D., Hauser C.R., Sundaram J.P., Nelson W.C., Madupu R., Brinkac L.M., Dodson R.J., Rosovitz M.J., Sullivan S.A., Daugherty S.C., Haft D.H., Selengut J., Gwinn M.L., Zhou L., Zafar N., Khouri H., Radune D., Dimitrov G., Watkins K., O'Connor K.J., Smith S., Utterback T.R., White O., Rubens C.E., Grandi G., Madoff L.C., Kasper D.L., Telford J.L., Wessels M.R., Rappuoli R., Fraser C.M. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U.S.A.* 102:13950-13955(2005).

Tong P., Prendergast J.G., Lohan A.J., Farrington S.M., Cronin S., Friel N., Bradley D.G., Hardiman O., Evans A., Wilson J.F., Loftus B. Sequencing and analysis of an Irish human genome. *Genome Biol* 11:R91-R91(2010).

UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 40:D71-D75(2012).

Vaynberg I., Apache Wicket Cookbook. Packt Publishing. 312pp (2011) ISBN 1849511608

Velankar S., Alhroub Y., Best C., Caboche S., Conroy M.J., Dana J.M., Fernandez Montecelo M.A., van Ginkel G., Golovin A., Gore S.P., Gutmanas A., Haslam P., Hendrickx P.M., Heuson E., Hirshberg M., John M., Lagerstedt I., Mir S., Newman L.E., Oldfield T.J., Patwardhan A., Rinaldi L., Sahni G., Sanz-Garcia E., Sen S., Slowley R., Suarez-Uruena A., Swaminathan G.J., Symmons M.F., Vranken W.F., Wainwright M., Kleywegt G.J. PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* 40:D445-D452(2012).

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., Gocayne J.D., Amanatides P., Ballew R.M., Huson D.H., Wortman J.R., Zhang Q., The sequence of the human genome. *Science* 291:1304-1351(2001).

Vilella A.J., Severin J., Ureta-Vidal A., Heng L., Durbin R., Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19:327-335(2009).

Vizcaino J.A., Cote R., Reisinger F., Foster J.M., Mueller M., Rameseder J., Hermjakob H., Martens L. A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* 9:4276-4283(2009).

Walker F.O. Huntington's disease. *Lancet* 369:218-228(2007).

Wall L., Christiansen T., Orwant J. Programming Perl, 3rd Edition. O'Reilly Media. 1104pp (2000) ISBN 978-0-596-00027-1.

Walls C. Spring in Action. Manning. 700pp (2010) ISBN 1935182358.

Wang J., Wang W., Li R., Li Y., Tian G., Goodman L., Fan W., Zhang J., Li J., Zhang J., Guo Y., Feng B., Li H., Lu Y., Fang X., Liang H., Du Z., Li D., Zhao Y., Hu Y., Yang Z., Zheng H., Hellmann I., Inouye M., Pool J., Yi X., Zhao J., Duan J., Zhou Y., Qin J., Ma L., Li G., Yang Z., Zhang G., Yang B., Yu C., Liang F., Li W., Li S., Li D., Ni P., Ruan J., Li Q., Zhu H., Liu D., Lu Z., Li N., Guo G., Zhang J., Ye J., Fang L., Hao Q., Chen Q., Liang Y., Su Y., San A., Ping C., Yang S., Chen F., Li L., Zhou K., Zheng H., Ren Y., Yang L., Gao Y., Yang G., Li Z., Feng X., Kristiansen K., Wong G.K., Nielsen R., Durbin R., Bolund L., Zhang X., Li S., Yang H., Wang J. The diploid genome sequence of an Asian individual. *Nature* 456:60-65(2008).

Waterhouse R.M., Zdobnov E.M., Tegenfeldt F., Li J., Kriventseva E.V. OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res.* 39:D283-D288(2011).

Weigel D., Mott R. The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol.* 10:107-107(2009).

Wheeler D.A., Srinivasan M., Egholm M., Shen Y., Chen L., McGuire A., He W., Chen Y.J., Makhijani V., Roth G.T., Gomes X., Tartaro K., Niazi F., Turcotte C.L., Irzyk G.P., Lupski J.R., Chinault C., Song X.Z., Liu Y., Yuan Y., Nazareth L., Qin X., Muzny D.M., Margulies M., Weinstock G.M., Gibbs R.A., Rothberg J.M. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872-876(2008).

Wilkins M.R., Sanchez J.C., Gooley A.A., Appel R.D., Humphery-Smith I., Hochstrasser D.F., Williams K.L. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol. Genet. Eng. Rev.* 13:19-50(1996).

Wu D., Hugenholtz P., Mavromatis K., Pukall R., Dalin E., Ivanova N.N., Kunin V., Goodwin L., Wu M., Tindall B.J., Hooper S.D., Pati A., Lykidis A., Spring S., Anderson I.J., D'haeseleer P., Zemla A., Singer M., Lapidus A., Nolan M., Copeland A., Han C., Chen F., Cheng J.F., Lucas S., Kerfeld C., Lang E., Gronow S., Chain P., Bruce D., Rubin E.M., Kyrpides N.C., Klenk H.P., Eisen J.A. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462:1056-1060(2009).

Yooseph S., Sutton G., Rusch D.B., Halpern A.L., Williamson S.J., Remington K., Eisen J.A., Heidelberg K.B., Manning G., Li W., Jaroszewski L., Cieplak P., Miller C.S., Li H., Mashiyama S.T., Joachimiak M.P., van Belle C., Chandonia J.M., Soergel D.A., Zhai Y., Natarajan K., Lee S., Raphael B.J., Bafna V., Friedman R., Brenner S.E., Godzik A., Eisenberg D., Dixon J.E., Taylor S.S., Strausberg R.L., Frazier M., Venter J.C. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* 5:E16-E16(2007).

Zhao L., Liu L., Leng W., Wei C., Jin Q. A proteogenomic analysis of *Shigella flexneri* using 2D LC-MALDI TOF/TOF. *BMC Genomics* 12:528-528(2011).

Zhou Q., Brown J., Kanarek A., Rajagopal J., Melton D.A. In vivo reprogramming of adult pancreatic exocrine cells to beta-cells. *Nature* 455:627-632(2008).



## APPENDIX

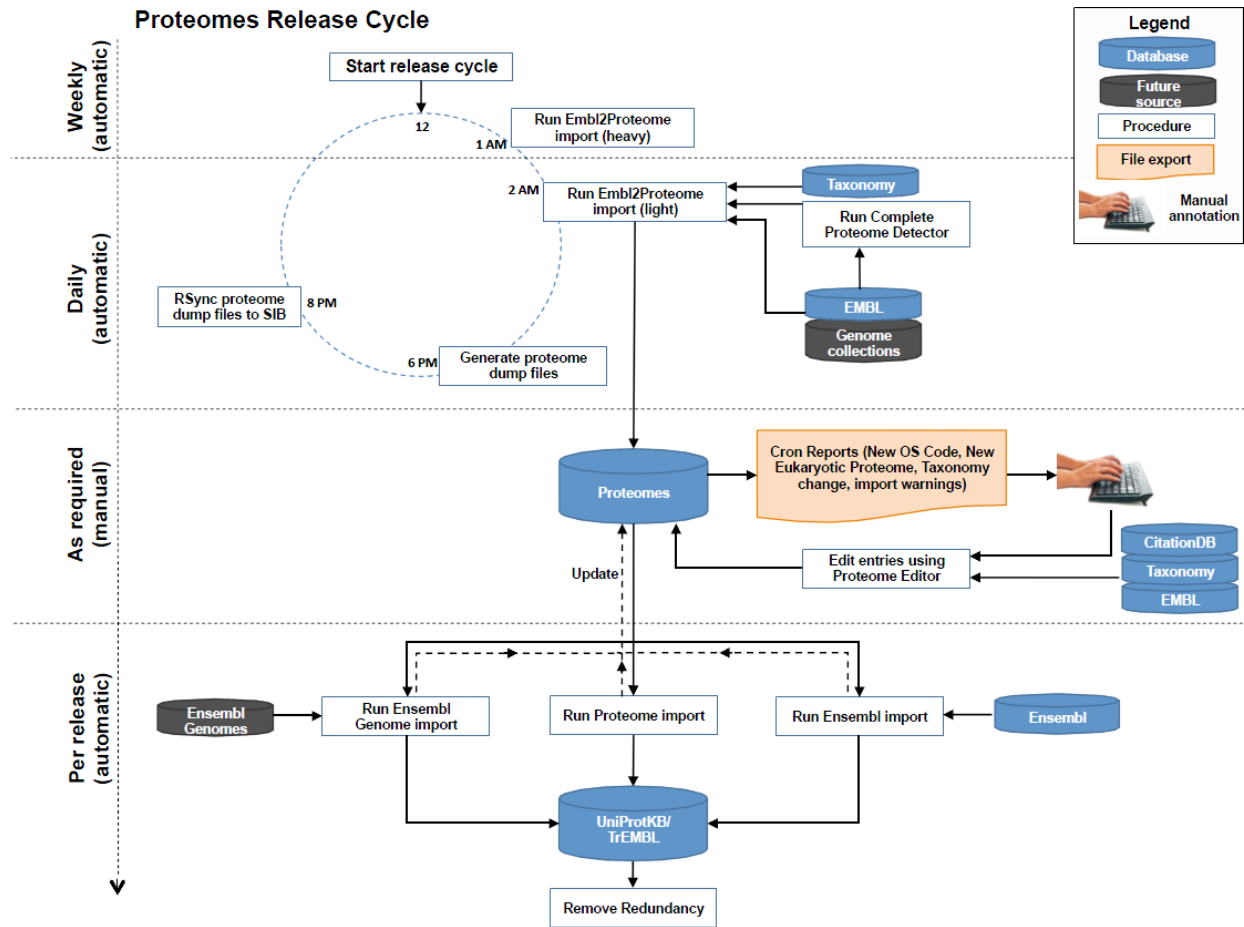
### Appendix 1. Table of Database and Consortium abbreviations

Abbreviation	Expansion
DDBJ	DNA Databank of Japan
EGA	European Genome-phenome Archive
ELIXIR	European Life Sciences Infrastructure for Biological information
EMBL-EBI	European Bioinformatics Institute
EMBL	European Molecular Biological Laboratories
ENA	European Nucleotide Archive
G10KCOS	Genome 10K Community of Scientists
GEBA	Genomic Encyclopedia of Bacteria and Archaea
GOLD	Genomes OnLine Database
GXA	The Gene Expression Atlas
HAMAP	High-quality Automated and Manual Annotation of microbial Proteomes
HapMap	International Haplotype Map Project
HGP	Human Genome project
ICTV	International Committee on Taxonomy of Viruses
IHGSC	International Human Genome Sequencing Consortium
INSDC	International Nucleotide Sequence Database Collaboration
IUMS	International Union of Microbiological Societies
NCBI	National Center for Biotechnology Information
PDBe	Protein Databank in Europe
PIR	Protein Information Resource
PRIDE	Proteomics Identifications Database
RefSeq	NCBI Reference Sequence
SGD	Saccharomyces Genome Database
SRA	Sequence Read Archive
TAIR	The Arabidopsis Information Resource
UniMES	UniProt Metagenomic and Environmental Sequences database
UniParc	UniProt archive
UniProt	The Universal Protein resource consortium
UniRef	UniProt Reference clusters
UniProtKB	The Universal Protein resource Knowledgebase
WTSI	Wellcome Trust Sanger Institute
WTCCC	Wellcome Trust Case Control Consortium

## Appendix 2. Proteomes database schema



### Appendix 3. Proteomes workflow diagram



**Appendix 4. Report.txt for the comparison of proteomes between UniProt release 2012\_01 and 2012\_02.**

Metrics to compare the new and old release

=====

Number of proteome\_ids the same between databases: 3095

Number of proteome\_ids new to the database: 48

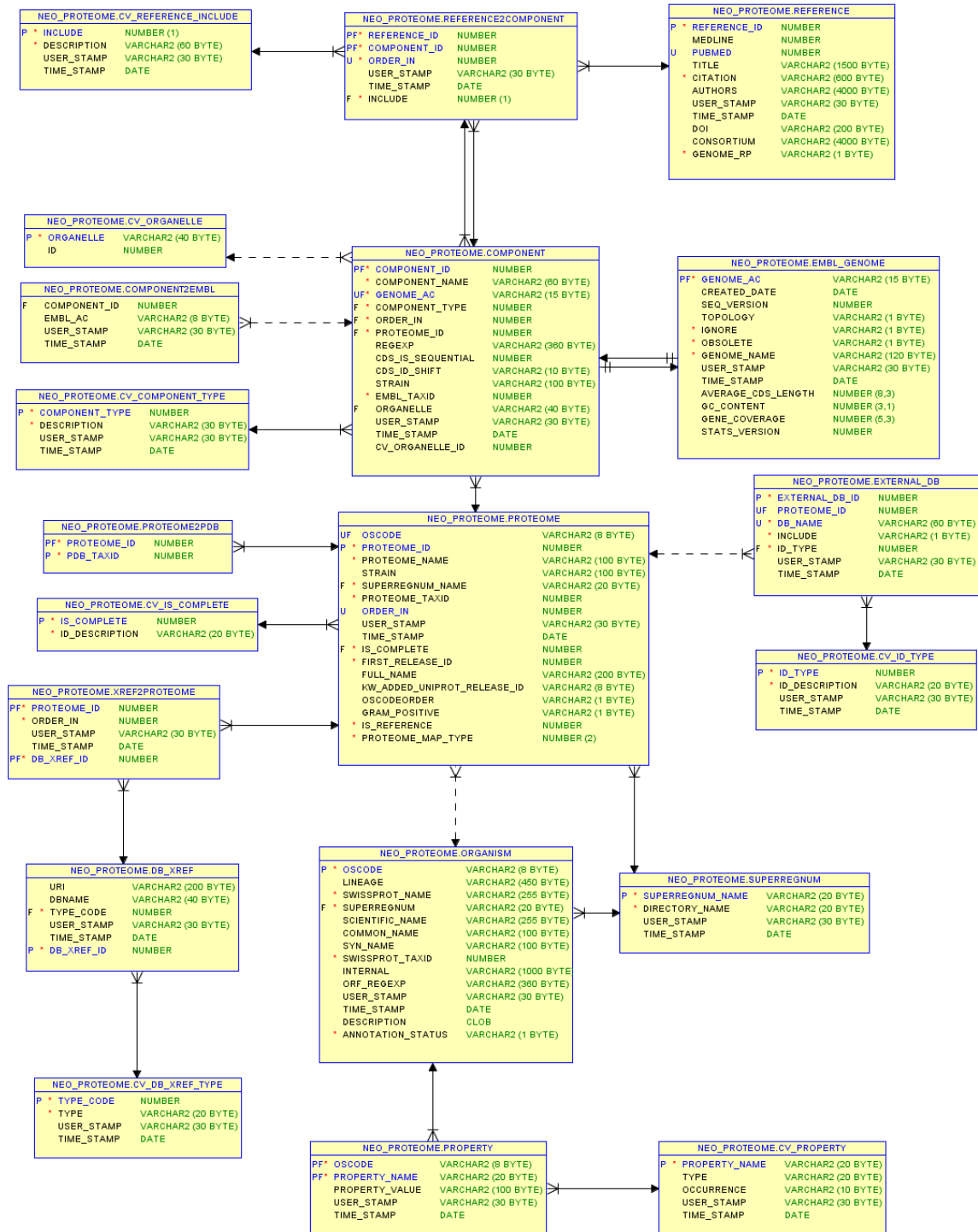
Number of proteome\_ids deleted from the new database: 24

Number of oscodes the same between databases: 2923

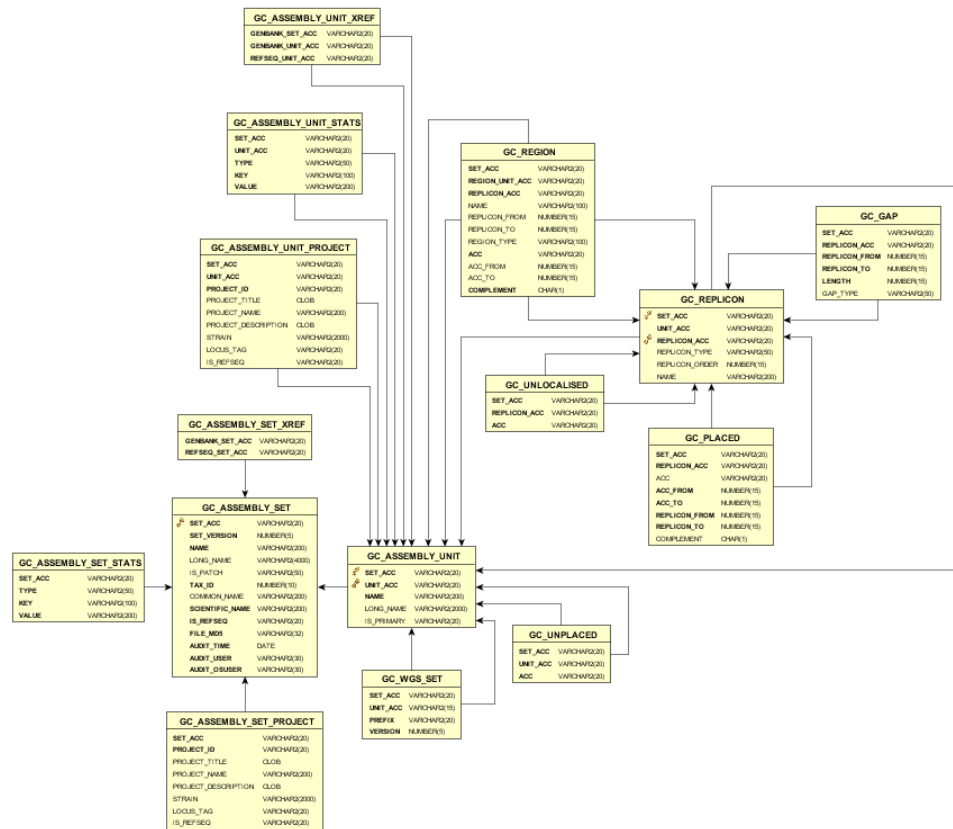
Number of oscodes new to the database: 25

Number of oscodes deleted from the new database: 24

## Appendix 5. Reduced Proteomes database schema



## Appendix 6. Genome Collections schema



files