Biological Databases

Don't read here – download as a PDF!
But I'm still working on it...

Table of Contents

Database Basics

Basic questions for any database

I. Using this databases
II. Publications?
    A. how to use this resources
III. Learn which kind of databases exists
IV. Database entries questions
    A. What kind of datasets / data types / entity types it store
    B.
V. What type of analyses you are able to perform using this databases

Databases: general info

I. Why we need a database? External lecture
    A. available for other scientists
        1. through web-interface
    B. biological data in computer-readable format
        1. computer-based analysis
        2. They contain a set of data, often in the form of rules that describe the knowledge in a logically consistent manner.

        a) Using forms and grammar
    3. Ideally –> Having automated deductive reasoning = Artificial Intelligence
        a) Reasoning = New facts are deduced logically from old ones
C. handle and share large volume of data
    1. defined formats
    2. automated storage and retrieval of experimental data
D. link knowledge with external resources

II. Technical design of databases

A. Flatfile db
    1. Simple and fast
    2. it has no structure for indexing and there are usually no structural relationships between the records
    3. Like excel sheet
        a) 65k lines only
        b) Scalability of the system reaches it's limits
    4. Examples
        a) NCBI GenBank
        b) PDB-file
        c) Swiss-Prot in the beginning
B. Relational db (=table oriented db)
    1. Unlimited storage space
    2. Pros
        a) Easy data retrieval
        b) Easy to update data and data type
        c) Easy to add or delete records from the data
        d) Consistent[1] data
        e) Stable enables multiple users to work at the same time Security
        f) Avoids data duplication
    3. MySQL
        a) Academic bioinformatic community use it
        b) In SCAI | by biologist can be used as Front-end db
    4. E.g. Swiss-Prot
C. object-oriented db
D. NoSQL =(non relational | not {only}* SQL) db
    1. Document-oriented db
        a) XML
            (1) Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. Simplicity, generality and usability across the Internet.
            (2) XSD (XML schema): road map for the XML document.
            (3) How to transfer information from XML to database: Use script to translate hierarchical information in XML file into SQL statements.
        b) JSON
    2. Graph db
        a) Coming, but not yet going
        b) Pros
            (1) graph databases are often faster for associative data sets and map more directly to the structure of object-oriented applications.
            (2) They can scale more naturally to large data sets as they do not typically require expensive join operations.

[1] unchanging in achievement or effect over a period of time.: "manufacturing processes require a consistent approach".

(3) As they depend less on a rigid schema, they are more suitable to manage ad hoc and changing data with evolving schemas.

c) Cons

(1) Conversely, relational databases are typically faster at performing the same operation on large numbers of data elements.

III. Access to the data

A. Accession code

1. Primary key for the entry
2. Supposed to be stable
3. You should give this entry in discession

B. Identifier

1. Interpretable by human
2. Can be changed: both stay, but old as secondary entry
3. e.g. locus in GenBank; entry name in UniProtKB/Swiss-Prot

IV. Handling of data

A. primary data

1. experimental results directly into database

B. secondary or derived data

1. results of analysis of primary databases

V. **Types of biological databases** (based on scope and level of curation) from NCBI & Hofmann's lecture

A. Comprehensive

1. Nucleotide
   a) NCBI GenBank
   b) EMBL ENA
   c) DDBJ: DNA Data Bank of Japan
2. Expression data
   a) ArrayExpress
   b) GEO
3. Protein sequence
   a) UniProtKB/Swiss-Prot
   b) UniProtKB/TrEMBL
   c) PIR-International: Protein Information Resource
   d) InterPro
4. Protein structure
   a) RCSB-Protein Data Bank (RCSB-PDB)
   b) PDBe
   c) PDBj
   d) MMDB: Molecular Modeling Database (based on PDB)
5. Other molecules
   a) Lipids
      (1) Lipid Bank
         (a) http://lipidbank.jp/
   b) conjugates and polymers
      (1) Macromolecular structure databases (MSD) by EBI
6. Genomes & Maps
   a) Entrez Genomes
7. Genetic variants db
8. Metabolic pathways
9. ...

B. Specialized

1. Organism Specific
   a) Human Genome Sequencing
   b) GDB: Genome Database (human mapping information)

c) MGD: Mouse Genome Database
d) SGD: Saccharomyces Genome Database

2. Functional
    a) TRANSFAC: Transcription Factors
    b) Vector Database
3. Sequencing technology
    a) EST: Expressed Sequence Tags
    b) GSS: Genome Survey Sequences
    c) STS: Sequence Tagged Sites
    d) HTG: High Throughput Sequences
4. Disease-specific

C. Biological portals
1. EBI
    a) Call themself "The portal to knowledge"
        (1) data deposition and exploitation
        (2) - integrating biological and molecular annotations (semantic knowledge as a cross-links
        (3) between databases)
        (4) - “gateway” to BioDatabases = EBI combines several Bio-databases
        (5) - Biomedical knowledge is classified by specific area and stored in databases.
2. Expasy.org
    a) SIB Bioinformatics Resource Portal which provides access to scientific databases and software tools (i.e., resources) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc.
3. NCBI Entrez?
4. Closed ones
    a) SRS

VI. Level of Curation

A. Preliminary
1. unfinished sequence data are often made available on the web site of the sequencing center before the sequence data are finished and deposited to a database

B. Archival
1. repository of information
2. redundant; might have many sequence records for the same gene, each from a different lab
3. submitters maintain editorial control over their records

C. Curated
1. non-redundant; one record for each gene, or each splice variant
2. each record is intended to present an encapsulation of the current understanding of a gene or protein, similar to a review article
3. records contain value-added information that have been added by an expert(s)
    a) In data repository responsibility for the data is on the submitter
4. examples:
    a) RefSeq: NCBI Database of Reference Sequences (mRNAs, proteins, genomic contigs, and complete genomes/chromosomes)
    b) Swiss-Prot (protein sequences)
    c) Clusters of Orthologous Groups (COGs) (natural system of gene families from complete genomes)

D. Peer Reviewed

1. each record subject to review and comments from members of the scientific community
2. exampleomes:
    a) e.g., PROW: Protein Resources on the Web, short, structured reviews on proteins and protein families

VII. How to access the data?

A. Web interface
    1. web-based
    2. small scale
B. Web service
    1. SOAP
    2. COBRA
C. Flatfiles
    1. script-based
    2. large scale
D. Database dump
    1. script-based
    2. large scale

VIII. Databases of databases

A. Nucleic Acid Research – database issue every year
B. Database Journals
    1. Database: The Journal of Biological Databases and Curation
    2. Science Magazines websites
        a) sciencemag.org
C. Database portals
    1. DBD (database of biological database)
    2. Pathguide
    3. Expasy.org
        a) SIB Bioinformatics Resource Portal which provides access to scientific databases and software tools (i.e., resources) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc.
    4. NCBI Entrez?
    5. Closed ones
        a) SRS

IX. Database vs Knowledge base vs data repository

A. Database: essentially to store and organize data and only limited for this. Has no context, meaning in relation to other data.
B. Knowledge base: a base to store a knowledge. It can use more databases or can combine and enrich it with information from public data resources. Connection to other data.
    1. It is a highly curated database having as much annotations as possible.
    2. - is a special kind of database for knowledge management
    3. - knowledge base provides a means for information to be collected, organized, shared, searched and utilized.
    4. - Has accuracy and is not redundant
    5. - use more databases or can combine and enrich it with information from public data resources. Connection to other data.
C. Data repository: data warehouse, more databases.
    1. the difference between (general purpose) databases and repositories was the difference between "data" and "metadata".
    2. So, a database stores data. A repository is a special class of database which is designed to store meta-data, that is, data that describes other data.
        a) Metadata is data that describes other data. Meta is a prefix that in most information technology usages means "an underlying definition or

description." Metadata summarizes basic information about data, which can make finding and working with particular instances of data easier.

3. Example: BioGRID

Fundamental Principles of Databases

I. Annotation
   A. – the process of adding information to a particular entity, OR
   B. – a combination of comments, notations, references and citations (properties, features of any given object), that describe all the experimental and inferred information about a gene, protein, patient, etc.
      1. such kind of useful information such as:
         a) **molecular function**
         b) **biological process**
         c) **cellular localisation**
            (1) **GO**
      2. properties, features of any given object, that are not necessarily obvious (explicit knowledge); or implicit knowledge and information, that make it reasonable to classify or tag that with some term
   C. – Description of features and properties of any given object in a very standardized way. The features are not always so obvious, for example, for Alzheimer disease - dementia also.
   D. – The process of attaching additional information to biological entities. Annotation can be
      1. structural (i.e. identification of the elements from a sequence, such as protein coding regions or the location of regulatory motifs) or
      2. functional (i.e. adding biological information to the identified elements, such as the biological function of a protein domain or an entire protein, or the molecular interactions or regulatory role of a nucleotide sequence).
      3. Annotation can either be applied automatically or can be manually added (in a process called 'curation') from various sources, such as the scientific literature. Annotation can either be applied automatically or it can be curated (manually) from the scientific literature. At EMBL-EBI, we use a combination of automatic and manual annotation to enrich our databases.
   E. Swiss-Prot
      1. When a new protein sequence enters into SwissProt from TrEMBL it is manually annotated follows: Sequence curation → Sequence analysis → Literature curation → family-based Curation (Homology) → Evidence Attribution → Quality assessment
   F. Gene Annotation
      1. You annotate splice sites, exons/introns; position in the genome; the neighborhood; binding sites

II. Provenance = source of info
   A. ability to trace back the original source of information
   B. e.g. evidence

III. Curation
   A. why?
      1. eliminating redundancies and removing outliers
      2. unifying semantics
      3. check quality control, reliability of information
      4. Standardization
         a) To create the database; curators need to extract and organize data from literature. They also need to describe the data in terms of standards, protocols, vocabulary.

IV. Versioning
   A. accession numbers

1. generated automatically
2. Are stable from release to release

B. submission process
1. why you want an accession number
a) unique identifier = readable by human
2. you have submitted; for publication
a) the mechanism to force people to share info;

V. Data integration
A. done by shared key = standards
1. the same table level, which identical in 2 databases
B. # Uniprot = hub: have links, pointers to other DBs

VI. Usage of biodb (2 types)
A. humans – interface
B. used by machines (machine-machine)
1. web services

VII. Ownership of info: distinguish db with
A. primary data
B. derived data
1. derived: submitter = owner

VIII. Staging of db / updating mechanism
A. archive
B. gene db: annual update and packaged together with previous years
C. builds of db (context of human genome); build = version
1. complex object like genome;
2. you assemble it and refined
a) many different sequences, context = areas, continuous DNA stretch;
b) problem with repetitive sequence
(1) clarify: one, two or more repeats;
3. consensus about info

IX. Information fusion and aggregation
A. # ensembl
1. documentation! http://www.ensembl.org/info/index.html
B. strategies to aggregate different types of information
C. need sort of shared keys -– smth you have to map on

X. Distribution looks awfully exponential, or cut gamma

Definitions

(taken mainly from Oxford Dict – select a word + command+shift+Y – and PDF)

I. Data
A. facts and statistics collected together for reference or analysis
B. *Philosophy* things known or assumed as facts, making the basis of reasoning or calculation.
C. from Latin, literally **'something given'**, neuter past participle of *dare* **'give'**.

II. Metadata
A. a set of data that describes and gives information about other data

III. Information
A. facts provided or learned about something or someone
B. *Computing* data as processed, stored, or transmitted by a computer
C. from Latin *informare* **'shape, fashion, describe'**, from *in-* **'into'** + *forma* **'a form'**.

IV. Knowledge
A. facts, information, and skills acquired through experience or education; the theoretical or practical understanding of a subject
1. the sum of what is known: the transmission of knowledge.
2. information held on a computer system.

B. Old English *cnāwan* (earlier *gecnāwan*)'recognize, identify'

V. Dictionary

A. List of terms (in alphabetical order) with their meaning or equivalents, used in a particular subject
B. Latin *dictio* 'word', Latin *dictinarium* 'manual or book of words'

VI. Glossary

A. an alphabetical list of terms or words found in or relating to a specific subject, text, or dialect, with explanations; a brief dictionary
B. Latin *glossa* 'explanation of a difficult word', from Greek *glōssa* 'word needing explanation, language, tongue'.

VII. Vocabulary

A. the body of words used in a particular language
B. Latin *vocabularius*, from Latin *vocabulum*, from *vocare* 'call'

VIII. Catalogue

A. a complete list of items, typically one in alphabetical or other systematic order – organized & detailed description of datasets location
B. from Greek *katalogos*, from *katalegein* 'pick out or enrol'.

IX. Index

A. a set of items each of which specifies one of the records of a file and contains information about its address
B. *Computing* a set of items each of which specifies one of the records of a file and contains information about its address
C. from Latin *index*, *indic-* 'forefinger, informer, sign', from *in-* 'towards' + a second element related to *dicere* 'say' or *dicare* 'make known'

X. Hierarchy

A. an arrangement or classification of things according to relative importance or inclusiveness
B. from Greek *hierarkhia*, from *hierarkhēs* 'sacred ruler'

XI. Controlled vocabulary

A. is an organized arrangement of terminology values used to index content and/or to retrieve content through browsing or searching.
    1. It typically includes preferred and variant terms and has a defined scope or describes a specific domain.
    2. The purpose of controlled vocabularies is to organize information and to provide terminology to catalog and retrieve information.
    3. The most important functions of a controlled vocabulary are to gather together variant terms and synonyms for concepts and to link concepts in a logical order or sort them into categories.

B. Types of controlled vocabularies:

1. Relationships in General

a) relationship means a state of connectedness (= an association between two things in a database—in this case, fields or tables in a database) for a controlled vocabulary

2. Subject Heading List (Heading)

a) uniform words or phrases intended to be assigned to books, articles, or other documents in order to describe the subject or topic of the texts and to group them with texts having similar subjects
    (1) are typically arranged in alphabetical order, with cross-references between the preferred, nonpreferred, and other related headings
    (2) Differs from other controlled vocabularies by pre coordination of terminology, which is a characteristic of subject headings in that they combine several unique concepts together in a string
b) E.g. Medical Subject Heading (MeSH) – is used for indexing journal articles and books on medical science. MeSH incorporates a thesaurus structure with subject headings.

3. Taxonomy

a) is an orderly classification for a defined domain.

b) Is-a relationship between class and subclasses
    (1) It may also be known as a faceted vocabulary. It comprises controlled vocabulary terms (generally only preferred terms) organized into a hierarchical structure. Each term in a taxonomy is in one or more parent/child (broader/ narrower) relationships to other terms in the taxonomy. There can be different types of parent/child relationships, such as whole/part, genus/ species, or instance relationships.
    (2) A taxonomy may differ from a thesaurus in that it generally has more obvious (=shallower) hierarchies and a less complicated structure. For example, it often has no equivalent (synonyms or variant terms) or related terms (associative relationships).
    (3) In common usage, the term taxonomy may also refer to any classification or placement of terms or headings into categories, particularly a controlled vocabulary used as a navigation structure for a Web site.
c) Classes and subclasses (before Homo sapiens: wiki 28 classes and 7 unranked = 35 + H. s. sapiens = 36)
d) No multiple inheritance!
e) from Greek *taxis* 'arrangement' + *-nomia* 'distribution'

4. Thesaurus (pl. -ri)

a) A book in which words are put into groups of synonyms and related concepts.
    (1) – semantic network of unique concepts, including relationships between synonyms, broader and narrower (parent/child) contexts, pointers and other related concepts.
    (2) Thesauri may contain three types of relationships: equivalence (synonym), hierarchical (whole/ part, genus/species, or instance), and associative.
    (3) Thesauri may also include additional peripheral or explanatory information about a concept, including a definition (or scope note), bibliographic citations, and so on. A thesaurus is more complex than a simple list, synonym ring list, or simple taxonomy.
    (4) The term thesaurus may also be used for any controlled vocabulary arranged in a known order, displayed with standardized relationship indicators, and generally used for browsing in post-coordinated information storage and retrieval systems.
b) Pointers – instance in a class of liver disease, e.g. malaria, and is also infectious disease. =Multiple inheritance
c) Greek *thēsauros* 'treasure, collection, storehouse'

C. Ontology

1. Classification & organization & formalization of information
2. One more article about it
3. What ontologies are for?
4. In common usage in computer science, an ontology is a formal, machine-readable specification of a conceptual model in which concepts, properties, relationships, functions, constraints, and axioms are all explicitly defined.
    a) Another def: the description of what exist specifically within a determined field. This includes the relationship and hierarchy between these parts.
        (1) They are focused on naming parts and processes and grouping similar ones together into categories. More general at the top, more specific at the bottom.
    b) Such an ontology is not a controlled vocabulary, but it uses one or more controlled vocabularies for a defined domain and expresses the vocabulary in a representative language that has a grammar for using vocabulary terms to express something meaningful. Ontologies generally divide the

realm (=kingdom) of knowledge that they represent into the following areas: individuals, classes, attributes, relations, and events.

c) The grammar of the ontology links these areas together by formal constraints that determine how the vocabulary terms or phrases may be used together. There are several grammars or languages for ontologies, both proprietary and standards-based. An ontology is used to make queries and assertions.

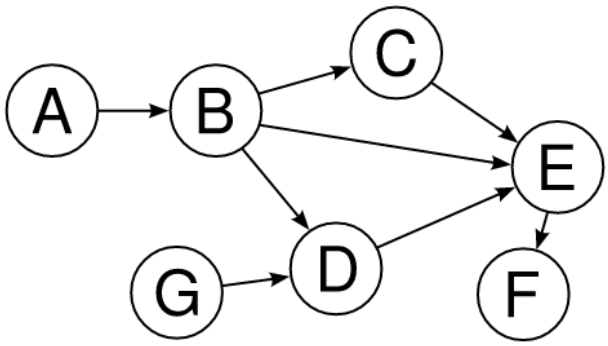
d) Ontologies have some characteristics in common with faceted taxonomies and thesauri, but ontologies use strict semantic (in language or logic) relationships among terms and attributes with the goal of knowledge representation in machine-readable form, whereas thesauri provide tools for cataloging and retrieval.

e) Ontologies are used in the Semantic Web, artificial intelligence, software engineering, and information architecture as a form of knowledge representation in electronic form about a particular domain of knowledge.

f) Relationships in ontologies are defined according to strict rules, which are different than the equivalence, hierarchical, and associative relationships used for thesauri and other vocabularies discussed in this book.

g) *Specification* – an act of describing or identifying something precisely or of stating a precise requirement

5. Components of ontology
    a) Individuals: Instances, objects
    b) Classes: Sets
    c) Attributes: Properties of individuals
    d) Relations: Between classes and individuals
    e) Functional terms made from several classes
    f) Rules (if-then statement), restrictions (description of relations), axioms, events (changes)
6. Directed acyclic graph (DAG) – graph, edges of which are directed and contains no cycle.
    a) An ontology is structured as a directed acyclic graph (DAG) where each term has defined relationships to one or more other terms in the same domain, or in other domains. A logic connection of terms can never lead back to the term itself. E.g. in the picture below, each number stands for a term (protein, biological function or DNA) and the arrows stand for a connection between them.



7. Can have multiple inheritance (multiple parents) and multiple childs
8. Latin *ontologia*, from Greek *ōn, ont-* 'existence or being real'
    a) In philosophy, is the study of what exists, in general.
        (1) What are fundamental parts of the world and how are they related to each other?
    b) Used to discuss challenging questions to build theories and models, and to better understand the ontological status of the world.

c) Ontological materialism – material things are more real than the human mind; reality exists regardless of human observer.
d) Ontological idealism – immaterial phenomenon (mind and consciousness) are more real than material things. Reality is constructed in the mind of the observer.

9. Ontology vs. Taxonomy
10.

Ontology	Taxonomy
controlled vocabulary with explicit grammar rules à meaningful	Controlled vocabulary
Ontology is needed because of a big amount of different nomenclatures. And we should organize and standardize knowledge by sorting it in knowledge domains.	
Ontology is mostly presented as a DAG structure	tree structure
Interactions are not necessarily in a hierarchical way	A taxonomy can only sot the terms in a hierarchical way
broader scope of information	Less data info

11. Synonyms vs. Homonyms
    a) Synonyms are different words with identical meaning. If multiple terms are used to mean the same thing, one of the terms is identified as the preferred term in the controlled vocabulary and the other terms are listed as synonyms.
    b) If the same term is commonly used to mean different concepts in different contexts, then this term is a homonym. Homonyms are identical words with different meaning.

XII. Conceptual model

A. models which are formed after a conceptualization or generalization process.Conceptual models are often abstractions of things in the real world whether physical or social.
B. Semantics studies are relevant to various stages of concept formation and use as Semantics is basically about concepts, the meaning that thinking beings give to various elements of their experience.
C. Temporal changes
    1. metadata (experimental conditions), ID, last actualization
    2. results (row data), curation, ID

XIII. Annotation

A. a note by way of explanation or comment added to some entity
B. from Latin annotat- ‘marked’, from the verb annotare, from ad- ‘to’ + nota ‘a mark’.

XIV. Curation

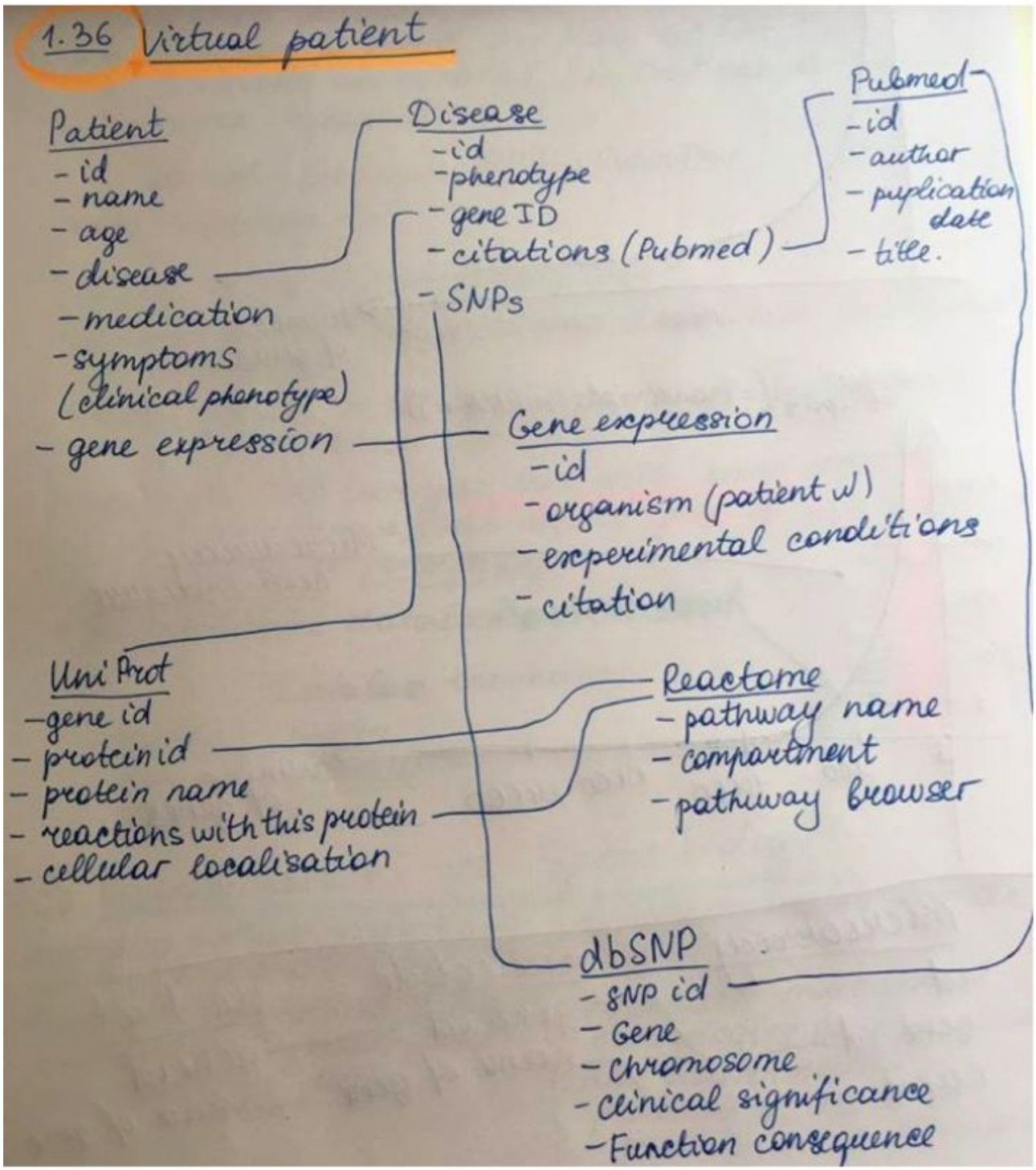
A. quality control of primary biological research data intended for publication, extracting and organizing data from original scientific literature, and describing the data with standard annotation protocols and vocabularies that enable powerful queries and biological database interoperability.
B. biocurator is a professional scientist who curates, collects, annotates, and validates information that is disseminated by biological and model organism databases.
C. from Latin curare ‘take care of’, from cura ‘care’

XV. Web services

A. (www.w3.org/ws) are software systems designed to support interoperable machine-to-machine interaction over a network. To ensure that software systems from different sources work well together, they are built using open standards such as SOAP.

Questions

1. **Develop a conceptual model of a database that stores information about a "virtual patient". Think first! What elements does a "virtual patient" have? What sort of information do you need to represent in the conceptual model to be able to use that model for the purpose of "personalized medicine"?**



2. **You are doing your Master Thesis with some clinical researcher at Venusberg. The medical researchers expect you to help them with in silico methods to identify molecular determinants of Alzheimer disease. What databases will you access and mine to support them and what sort of information do you get from these databases?**

a. **PubMed** – papers, citations, related to the disease. We have found what is already done and what can we discover further. Related genes and proteins are found.
b. In gene databases we can look only for the related genes (either with the gene name, or with the disease relation).
    i. Gene databases: **DDBJ, GenBank, EMBL** - sequence, chromosome location, links to variations (SNPs), interactions, pathways, protein id, genomes, gene ontologies (GO)
c. From gene databases we can connect directly to the other databases types, like protein DBs, SNPs DBs, pathways DBs, enzymes DBs, genome DB, **GeneRef**.
    i. GO – molecular function, cellular localization, biological function
    ii. Genome DBs (**ENSEMBL** for human, **MGI** for mouse knockout strains for example, ...)
    iii. **dbSNP** – information about this SNP, chromosome location, gene name, consequence, clinical significance.
    iv. **UniProt** – sequence of protein, function of protein, mutations of this protein, cellular localization of this protein, links to other DBs, like GO, ArrayExpress.
    v. **PDB** – protein database with the 3D structure of desired protein.
    vi. **KEGG** and **Reactome** – pathways related to gene, connections of pathways.
    vii. **GeneRef** – gene references into functions.
    viii. **ArrayExpress** and **GEO** – expression of genes, and experimental conditions of the screening experiments.
    ix. **OMIM** – genes and genetic disorders are connected. You get the phenotype, disease information.

Alzheimer's (AD), Parkinson's disease (PD), Posttraumatic stress disorder (PTSD)

I. How do you extract and present information on AD from databases?
II. Ontology of AD – different aspects of disease
    A. Molecular level
    B. Cellular level
    C. Clinical aspects

    Which are organized in different categories = try to break down the knowledge:
    - Genomics
    - Proteomics
    - Metabolomics

III.

Biology Basics

Models in biology

I. Model a disease – used no understand biological phenomena
A. animal model
    1. Mammalia -> mouse, rat, rabbit
    2. Yeast = Saccharomyces cerevisiae
        a) ~ 6k gene
        b) over 40,000 research papers, and the number of yeast researchers exceeds the number of genes =)
    3. Worm -> Caenorhabditis elegans
    4. Plant -> Arabidopsis thaliana
B. advantages:

1. fast reproducible
2. you are able to track genetically their genotype, generation by generation

C. what kind of molecular analysis they do?
   1. Omics!

D. why not humans model in biology?
   1. ethical issues

E. Interesting facts
   1. Fugu fish – almost no introns; all exonic;
      a) Number of genes in Human Genome Project were estimated from the reassociation and hybridisation of entire genome;
         (1) ~ 30k
   2. Start to shrink – fragments, which belongs together. Annotation was wrong – 1 big peace instead of one
      a) ~ 24.800 of genes
      b) 19k in articles

Omics = Cellular Level

- **Genomics**
  - **Epigenomics**
  - also studying the mutations
- **Transcriptomics**
  - level of mRNA
- **Proteomics**
- **Metabolomics**
  - – the scientific study of chemical processes involving metabolites.
  - – "systematic study of the unique chemical fingerprints that specific cellular processes leave behind", the study of their small-molecule metabolite profiles.
  - The metabolome represents the collection of all metabolites in a biological cell, tissue, organ or organism, which are the end products of cellular processes. One can find them in blood, body liquids and spinal fluid (S. cord).
  - We can use metabolites as biomarkers – chemical compounds, which are secreted in the body and can be used in diagnosis.
  - **Metabonomics**
    - the study in the change of metabolites. Be careful when talking to a researcher on metabolites. Whether they identify with the L or the N crew is incredibly important to their ego.
  - **Lipidomics**
    - could be included in the metabolomics (most definitely, people who study lipids just want their own name)
  - **Glycomics** = study of sugars
    - Actually pretty interesting to Cembio peeps because of the way that each organism has its own glycosylation pattern on proteins. Important when making antibodies in a non-human cell line, since they have to be fixed so human immune systems don't go after them

Itself

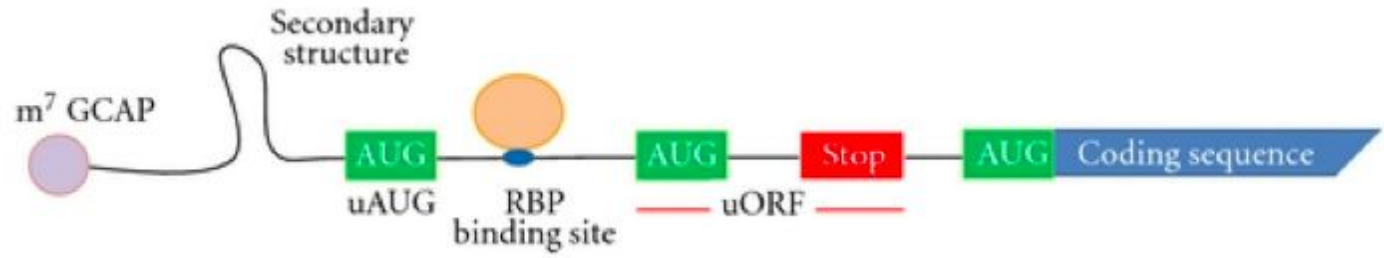
1. **RNA**

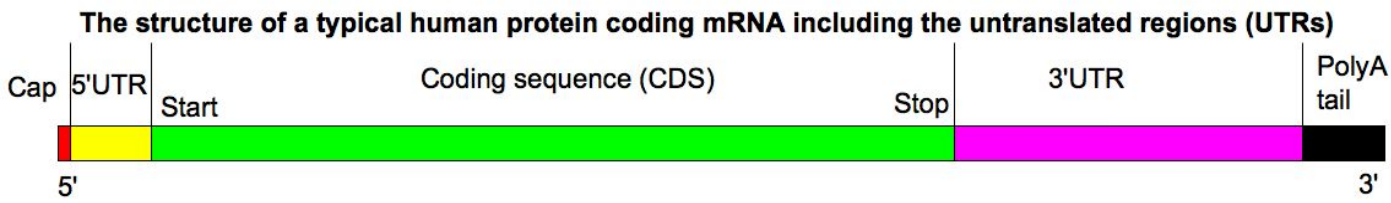
a. cDNA (complementary DNA)
   i. is a piece of DNA copied from a mature mRNA.

b. non-coding RNA (ncRNA) or non-protein-coding RNA (npcRNA) or non-messenger RNA (nmRNA) or functional RNA (fRNA)
   i. Heterogenous nuclear RNA (hnRNA)
      1. a precursor RNA, i.e. an RNA transcript before it is processed into mRNA,

ii. Small interfering RNA (siRNA)
    1. sometimes known as short interfering RNA or silencing RNA, is a class of double-stranded RNA molecules, 20-25 base pairs in length. siRNA plays many roles, but it is most notable in the RNA interference (RNAi) pathway, where it interferes with the expression of specific genes with complementary nucleotide sequences.
iii. micro RNA (miRNA)
    1. a small non-coding RNA molecule (containing about 22 nucleotides) found in plants, animals and some viruses, that functions in RNA silencing and post-transcriptional regulation of gene expression.[1][2]
iv. Small nucleolar RNAs (snoRNAs)
    1. are a class of small RNA molecules that primarily guide chemical modifications of other RNAs, mainly ribosomal RNAs, transfer RNAs and small nuclear RNAs.
v. Small nuclear ribonucleic acid (snRNA)
    1. also commonly referred to as U-RNA, is a class of small RNA molecules that are found within the splicing speckles and Cajal bodies of the cell nucleus in eukaryotic cells.
vi. Long non-coding RNAs (long ncRNAs, lncRNA)
    1. are non-protein coding transcripts longer than 200 nucleotides.
vii. The scaRNAs[1] (Small Cajal body RNA genes)
    1. resemble snoRNAs and perform a similar role in RNA maturation, but their targets are spliceosomal snRNAs and they perform site-specific modifications of spliceosomal snRNA precursors in the Cajal bodies of the nucleus.
viii. Extracellular RNA (also known as exRNA or exosomal RNA)
    1. describes RNA species present outside of the cells from which they were transcribed. In Homo sapiens, exRNAs have been discovered in bodily fluids such as venous blood, saliva, breast milk, urine, semen, menstrual blood, and vaginal fluid.
ix. Piwi-interacting RNA (piRNA)
    1. is the largest class of small non-coding RNA molecules expressed in animal cells

c. Structure
    i. mRNA consist of: 5-Cap, 5-UTR, Coding sequence, 3' UTR, 3-PolyA





    1. 3 ÚTR is untranslated region.
        a. • Influence the polyadenylation, translation efficiency, localization, stability of mRNA

b. • Has binding sites for the miRNA à decreasing expression of the mRNA
c. • Silencer genes are also there

2. 5 ÚTR
   a. - regulation of translation
   b. - uORF
   c. - Kozak sequence (initiation codon)
   d. - High GC content hairpin loops

2. What is a transcription factor site and why would you collect information on these sites in a database?

a. Transcription factors:
   i. • Interact with DNA
   ii. • Regulate the transcription process (initiate or inhibit)
   iii. • They are bound to the transcription factor site (either enhancer or promoter regions of
   iv. DNA)
   v. • Consist of domains: DNA-binding domain (DBD), trans-activating domain for the other cofactors (TAD), an optional signal sensing domain (SSD) (e.g., a ligand binding domain).
   vi. Transcription factor site is responsible for binding of transcription factors

b. Why should you collect information about that?
   i. • Transcription factors are related to some diseases and play an important role in life development.
   ii. • Manipulating TFs to reverse the cell differentiation process is the basis of methods for deriving stem cells from adult tissues.
   iii. • TF sites - information of artificial sequences resulting from mutagenesis, in vitro selection procedures from random nucleotide mixtures or from specific theoretical considerations.
   iv. There are some databases of TF of eukaryotic genes ranging from Humans to Yeast.
      1. DBD

3. Name at least three different classes (types) of transcription factors.

a. Transcription factors may be classified by their:
   i. - Mechanism of action
   ii. - Regulatory function (constitutively active/ signal dependent/ nuclear factors/cytoplasmic factors/ cell specific)
   iii. - Structural (Zinc fingers, helix-loop-helix, Zipper).

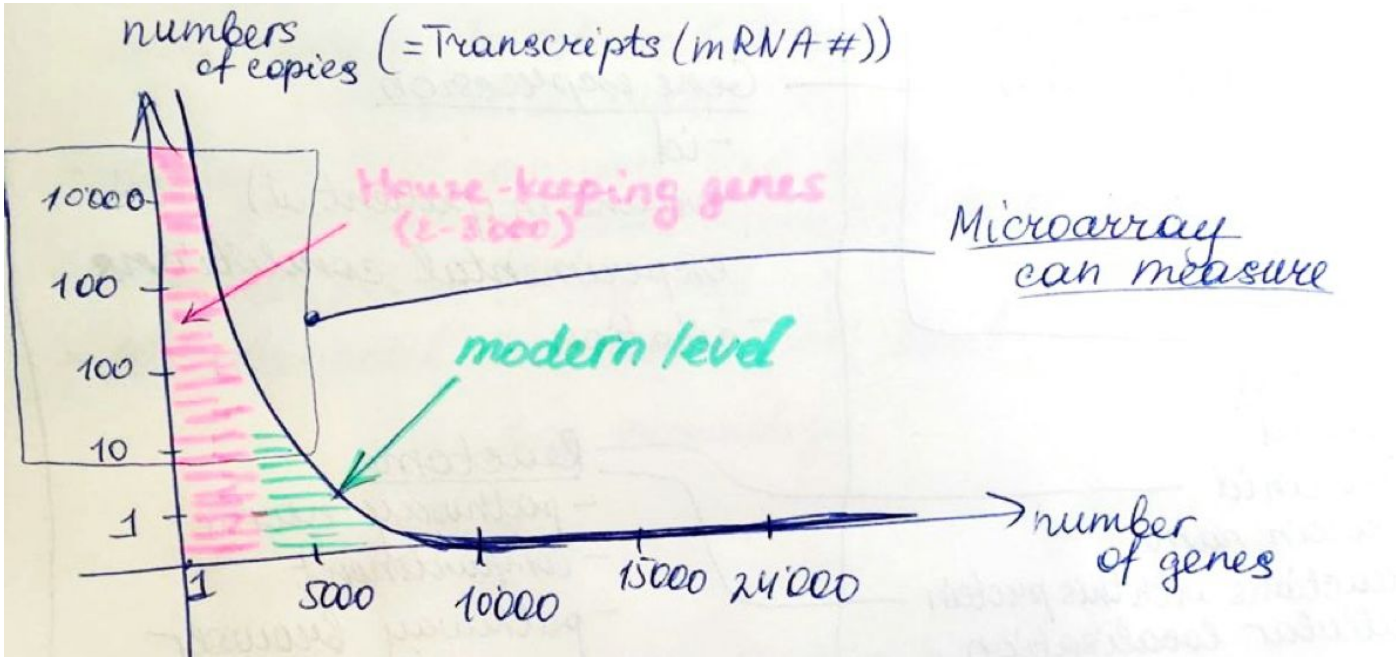
4. Explain how in silico prediction of transcription factor binding sites can be validated through molecular biology experiments. How come that a computer is able to predict the start and the end of a gene?

a. There are several databases, where you can find information about the binding sites and TFs, like DBD. After the prediction of the binding site in silico, you can check it in molecular biology experiments, like:
   i. - ChIP-on-chip analysis
   ii. - Yeast 2 hybrid
   iii. - site-directed mutagenesis
   iv. - 2-D electrophoresis
   v. - Southwestern blotting

b. Gene prediction:
   i. - ab initio:
      1. automated process
      2. computer is given instruction for finding genes in sequence
      3. computer looks for common sequences known to be found at the start and the end of gene (like promoter, start or stop)

ii. - evidence based:
    1. Technique relies on the non-DNA-data (mRNA and protein)
    2. mRNA and protein à backwards throught transcription and translation à idea of the original DNA sequence

5. **Explain the fundamentals of gene regulation (activation of transcription; features of DNA that mediate and control transcription) and sketch a simple conceptual model of entities that describes this biological process.**
    a. DNA:
        i. • Regulatory sequences
        ii. • TFs (activating TF/inhibitory repressor)
        iii. • Enhancer
        iv. • Repression, induction
        v. • Modification of chromatin (histones)
        vi. • DNA-Methylation - epigenetic regulation
    b. Transcription:
        i. • Transcription efficiency
    c. RNA:
        i. • RNA Processing
        ii. • Alternative splicing
        iii. • Stability of RNA
        iv. • miRNA
        v. • siRNA
    d. Translation:
        i. • translation efficiency
        ii. • binding on the ribosome
    e. Protein:
        i. • posttranslational modification (stability, degradation, ubiquitination)
6. **How does the transcriptional machinery know about the beginning and the end of a "gene" (a transcript)?**
    a. Beginning:
        i. - Promoters:
            1. • TSS – transcription start site
            2. • Binding site for RNA-polymerase
            3. • TF binding sites, e.g. TATA-box
        ii. - START codon – ATG
    b. Termination:
        i. - rho-independent (when the synthesized RNA forms a hairpin loop)
        ii. - rho-dependent (occurrence of the rho protein factor) in bacteria
        iii. - UGA for STOP codon
7. **Definition of housekeeping genes**
    a. - constitutive genes
    b. - for the maintenance of basic cellular function
    c. - expressed in all cells of an organism under normal and patho-physiological conditions -
    d. Examples: actin, GAPDH and ubiquitin.
8. **What is the typical distribution of all mRNA species expressed in a cell?**
    a. - Each gene: 1 -25000 copies per cell
    b. - Microarray method is not so sensitive because of the fluorescent labelling can measure only till 10 copies per cell. RNA seq can detect even 1 copy per cell!!!!
    c. - 2-3 thousands genes are highly expressed housekeeping genes
    d. - only 10000 genes from 24000 are expressed other are embryonic example
9. **"abundantly expressed" genes?**
    a. Genes that are highly expressed/over expressed (more than usual) in a sample.



10.
11.

Tissue, organ, organism layers

- 1st one is important in cancer research
  - oncology – science/study of treatment of tumors

High-throughput & other technologies

- https://docs.google.com/document/d/1Ybj4gGO-IPI562VrnI8EiGujhsx0JFkMLXkEASz0ZFg/edit#heading=h.vyco0vbw41j0

Types of mutations (causative genetic perturbations? or polymorphism)

- Single nucleotide polymorphism (SNP)
- Insertion and deletion (InDel)
- Copy number variations (CNV)

Questions

1. **What features would you assign to "Bioinformatics" and how does it differ from "Systems Biology"?**
   a. Bioinformatics:
      i. It's the application of computer science and information technology to the field of biology and medicine.
      ii. It deals with algorithms, databases, and information systems and soft computing data mining etc.
      iii. Genome, transcriptome and proteome are the features that can be assigned to Bioinformatics. Since Bioinformatics gives information about them.
   b. Systems Biology:
      i. Study of an organism viewed as an integrated and interacting network of genes (=graphs), proteins and biochemical reactions which give use to life.
      ii. Aims to explain how higher level properties of complex biological systems arise from the interaction among their parts.

iii. For System Biology, in additional to genome, transcriptome and proteome, metabolome can also be assigned as a feature. System Biology is a computational and experimental research group while Bioinformatics is advanced genome information technology research group.

2. Biomarker
    a. Biomarkers are often measured and evaluated to examine normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.
3.

Database comparison

1. **What sort of discrepancy (=difference) exists between Bio-Databases that represent information on genes and genomes as opposed to Bio-Databases that store information on gene expression and what are the consequences for the database design? EXAM**
    a. Sequences that are displayed by genes and genomes databases have a consistent representation inter-species wise.
    b. Gene expression data is more confined(=restricted) to the special sample, experiment reflecting the status of our bio-sample at a specific time i.e. the results cannot be generalized over the whole species.

BioDBs – gene and genome	BioDBs - gene expression
➢ static information, does not change with time	➢ dynamic information, changes with time, conditions and org
➢ contains data regarding gene sequence, location on the chromosome etc.	➢ contains data about experiment conditions, cell types, time etc on which gene expression
➢ comparatively less volume of data	➢ huge volume of data, increases with time (splicing, posttranscriptional modifications)

Gene Expression Databases

Gene Expression databases hold information about the levels of RNA expression. Data is acquired through microarray, RNA-seq, and other platforms. EMBL-EBI has a good primer on functional genomics to read first.

Data Platforms

Expressed sequence tags (EST) is a short subsequence of a cDNA sequence that's useful to each of these techniques in identifying transcripts.

- - expressed sequence tags (100-400 base pairs)
- - GenBank - dbEST, UniLib, UniGene comprises information on EST sequences
- - ESTs are produced via one-short sequencing of a cloned cDNA
- - with informatics tools you cluster them based on overlapping redundancy to form a contig that should represent the mRNA sequence.

EST & SAGE & other technologies for gene expression determination

Hybridization Methods - Microarray

Sequencing Methods - RNA-Seq ("Next-Gen Sequencing")

Chromatin Immunoprecipitation Sequencing (ChIP-Seq)

Data Standardization

The Functional Genomics Data Society (FGED) published these standards to improve interpretability and reproducibility of functional genomics and expression experiments. They've also published some standards on data formats, but that wasn't really mentioned in BioDB class; I just found it on their site's projects directory.

Minimum Information About a Microarray Experiment (MIAME)

This standard was established to ensure microarray experimental results are less ambiguous, easier to interpret, and more reproducible. It consists of the following points:

1. Raw Data
2. Normalized Data
3. Sample Annotation - compounds used, concentrations, etc.
4. Experimental Design and Sample/Data Relationships
5. Data Annotation (gene identifiers, genomic coordinates, probe sequence, reference array catalog number)
6. Laboratory and data processing protocols

MIAME Notation in Markup Language (MINiML)

MINiML (MIAME Notation in Markup Language, pronounced 'minimal') is a data exchange format optimized for microarray gene expression data, as well as many other types of high-throughput molecular abundance data. MINiML assumes only very basic relations between objects: Platform (e.g., array), Sample (e.g., hybridization), and Series (experiment). MINiML captures all components of the MIAME checklist, as well as any additional information that the submitter wants to provide. MINiML uses XML Schema as syntax.

Sources

- http://dx.doi.org/10.1038/ng1201-365
- http://fged.org/projects/miame/
- MINiML: http://www.ncbi.nlm.nih.gov/geo/info/MINiML.html

Minimum Information about a high-throughput nucleotide SEQuencing Experiment (MINSEQE)

This standard was established to ensure RNA-Seq experimental results are less ambiguous and easy to reproduce. Mufasra specifically said this would be on the test during her lecture on it, so yeah. It consists of the following points:

1. Description of biological system, samples, and experimental variables
2. Sequence read data for each assay (FASTQ)
3. Final data for each assay
4. Data-sample relationships, associated publication, summary of experiment
5. Experimental and data processing protocols

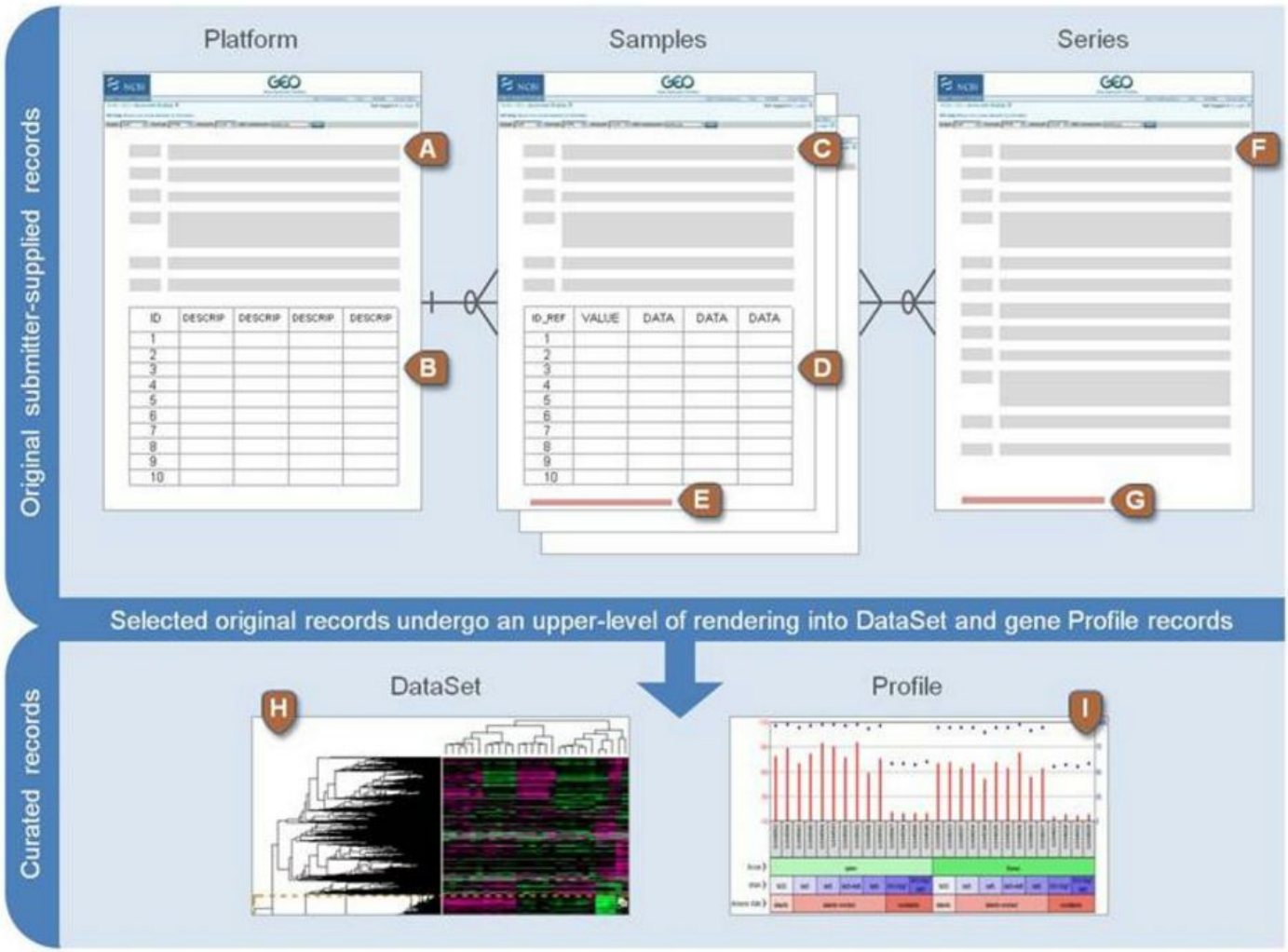
Links

- http://fged.org/projects/minseqe/

Gene Expression Omnibus (GEO)

A NCBI public repository of functional genomics data supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles. GEO stores lots of different types of expression data. However, ArrayExpress stores microarray data as well as and is better for access.

Organization schema of GEO records



A	Text description of the array or sequencer
B	Text tab-delimited table of the array template
C	Text description of the biological sample and protocols to which it was subjected
D	Text tab-delimited table of processed hybridization result or sequence counts
E	Raw data file, or processed sequence data file
F	Text description of the overall experiment
G	Tar archive of raw data files, or processed sequence data files

1. A Platform record is composed of a summary description of the array or sequencer and, for array-based Platforms, a data table defining the array template.
2. A Sample record describes the conditions under which an individual Sample was handled, the manipulations it underwent, and the abundance measurement of each element derived from it.

3. A Series record links together a group of related Samples and provides a focal point and description of the whole study. Series records may also contain tables describing extracted data, summary conclusions, or analyses.
4. How to submit:
    a. http://www.ncbi.nlm.nih.gov/geo/info/submission.html
5. Data types:
    a. Gene expression (see example)
    b. High throughput quantitative sequence data (see example) (see specific instructions)
    c. ChIP-chip (see example)
    d. ArrayCGH (Comparative Genomic Hybridization) (see example)
    e. SNP arrays (see example)
    f. SAGE (see example)
    g. Protein arrays (see example)

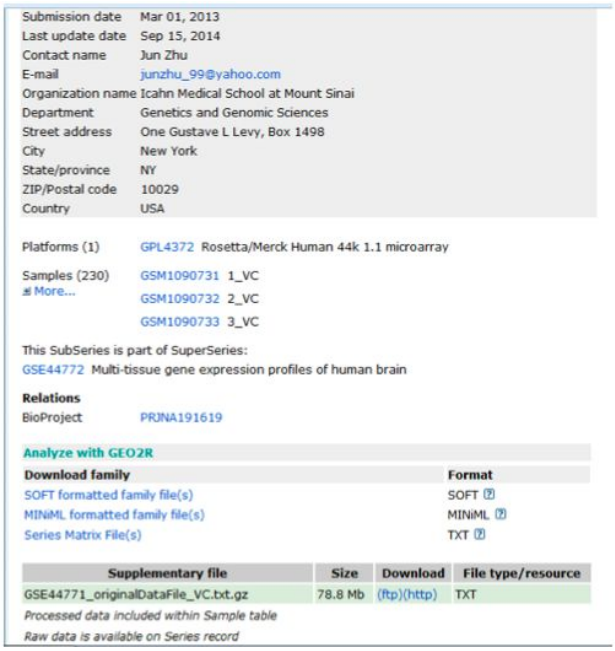
To quickly locate data relevant to your interests, search GEO DataSets and GEO Profiles:

- GEO DataSets is a *study-level* database which users can search for studies relevant to their interests. The database stores descriptions of all original submitter-supplied records, as well as curated DataSets. DataSet records are assembled by GEO curators.
    - http://www.ncbi.nlm.nih.gov/geo/info/datasets.html
- GEO Profiles is a *gene-level* database which users can search for gene expression profiles relevant to their interests. A Profile consists of the expression measurements for an individual gene across all Samples in a DataSet.
    - http://www.ncbi.nlm.nih.gov/geo/info/profiles.html

Questions

1. **How are experimental series represented in GEO?**

Series GSE44771	Query DataSets for GSE44771
Status	Public on Apr 25, 2013
Title	Multi-tissue gene expression profiles of human brain (VC)
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	The genetics of complex disease produce alterations in molecular interactions of cellular pathways which collective effect may become clear through the organized structure of molecular networks. To characterize molecular systems associated with late-onset Alzheimer´s disease (LOAD), we constructed gene regulatory networks in hundreds of autopsied brain tissues from LOAD patients and non-demented subjects. We demonstrate that LOAD reconfigures specific portions of the molecular interaction structure, and via an integrative network-based approach we rank ordered these sub-networks (modules) for relevance to LOAD pathology, highlighting the immune/microglia module as the top ranking. Through a Bayesian inference approach we identified multiple key causal regulators for LOAD brains.
Overall design	Autopsied tissues from dorsolateral prefrontal cortex (PFC), visual cortex (VC) and cerebellum (CR) in brains of LOAD patients, and non-demented healthy controls, collected through the Harvard Brain Tissue Resource Center (HBTRC), were profiled on a custom-made Agilent 44K array (GPL4372_1). All subjects were diagnosed at intake and each brain underwent extensive LOAD-related pathology examination. Gene expression analyses were adjusted for age and sex, postmortem interval (PMI) in hours, sample pH and RNA integrity number (RIN). In the overall cohort of LOAD and non-demented brains the mean ± SD for sample PMI, pH and RIN were 17.8±8.3, 6.4±0.3 and 6.8±0.8, respectively. 230 samples with all PFC, VC, and CR tissue profiled were included for further multi-tissue analysis.
Contributor(s)	Zhang B, Gaiteri C, Bodea L, Wang Z, McElwee J, Zhang C, Xie T, Tran L, Dobrin R, Fluder E, Clurman B, Narayanan M, Suver C, Shah H, Hahajan M, Lamb JR, Molony C, Stone DJ, Gudnason V, Myers AJ, Schadt EE, Neumann H, Zhu J, Emilsson V
Citation(s)	Zhang B, Gaiteri C, Bodea LG, Wang Z et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* 2013 Apr 25;153(3):707-20. PMID: 23622250



Submission date	Mar 01, 2013
Last update date	Sep 15, 2014
Contact name	Jun Zhu
E-mail	junzhu_99@yahoo.com
Organization name	Icahn Medical School at Mount Sinai
Department	Genetics and Genomic Sciences
Street address	One Gustave L Levy, Box 1498
City	New York
State/province	NY
ZIP/Postal code	10029
Country	USA

Platforms (1) GPL4372 Rosetta/Merck Human 44k 1.1 microarray

Samples (230) GSM1090731 1_VC
GSM1090732 2_VC
GSM1090733 3_VC
More...

This SubSeries is part of SuperSeries:
GSE44772 Multi-tissue gene expression profiles of human brain

Relations
BioProject PRJNA191619

Analyze with GEO2R

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE44771_originalDataFile_VC.txt.gz	78.8 Mb	(ftp)(http)	TXT

Processed data included within Sample table
Raw data is available on Series record

Platform → info about the array, how many, which genes, firm → link!!

Sample → info about each sample in the experiment → link!

a. - Series records are supplied by submitters.
b. - Serie has an accession number (GSExxx)
c. - Series record = group of related Samples
d. - focal point and description of the whole study
e. - contain summary of study, analysis = 1 experiment!
f. - Same like “ArrayExpress experiment”

2. **What fields can be searched in GEO and what fields can be browsed?**
   a. accession number – to platform, sample, serie, or data sets! (GPLxxx, GSMxxx, GSExxx, GDSxxx)
   b. - keywords, author
   c. - gene name
3. **What are GEO profiles?**
   a. This database stores individual gene expression profiles from curated DataSets in the Gene Expression Omnibus (GEO) repository. Search for specific profiles of interest based on gene annotation or pre-computed profile characteristics.
   b. - In comparison to DataSets, that have all experimental metadata, Profiles have individual gene expression
   c. - derived from curated GEO DataSets
   d. - each Profile is presented as:
   e. - chart that displays the expression level of one gene across all Samples within a DataSet.
   f. - GEO DataSets - study-level, Profiles - gene- level.
4. **Is there an accession number for microarray data?**
   a. GEO: more accession numbers:
      i. GSExxx (Series),
      ii. GSMxxx (Samples),
      iii. GPLxxx (Platforms)
      iv. GDSxxx (DataSet)
   b. ArrayExpress: Example, E-GEOD-60582
      i. E-“source”-xxx
      ii. Source: indicates where the experiment is coming from.
5.

Links

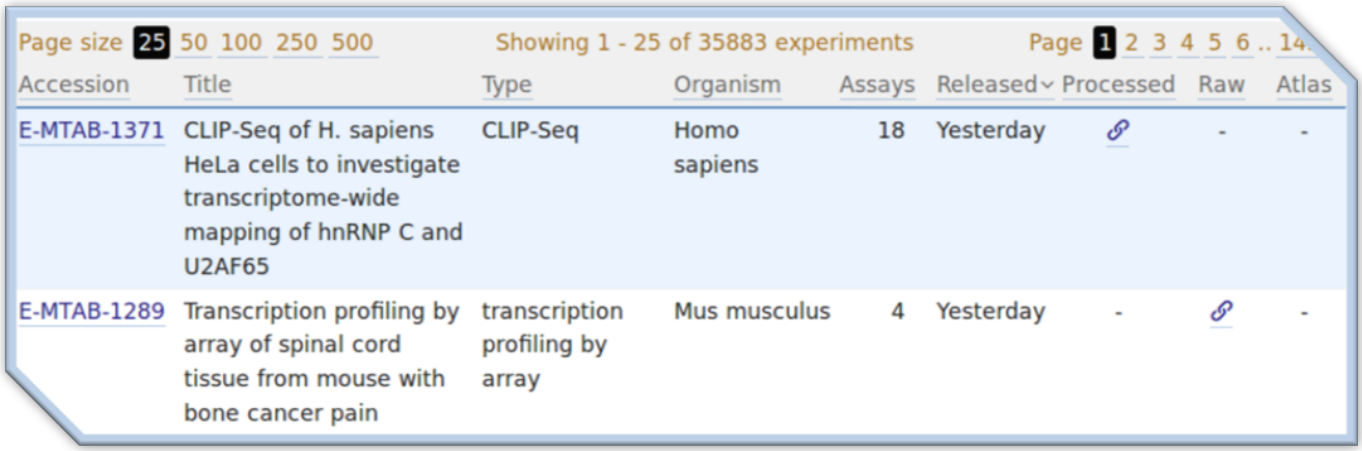
- http://www.ncbi.nlm.nih.gov/geo/
- Overview: http://www.ncbi.nlm.nih.gov/geo/info/overview.html

ArrayExpress

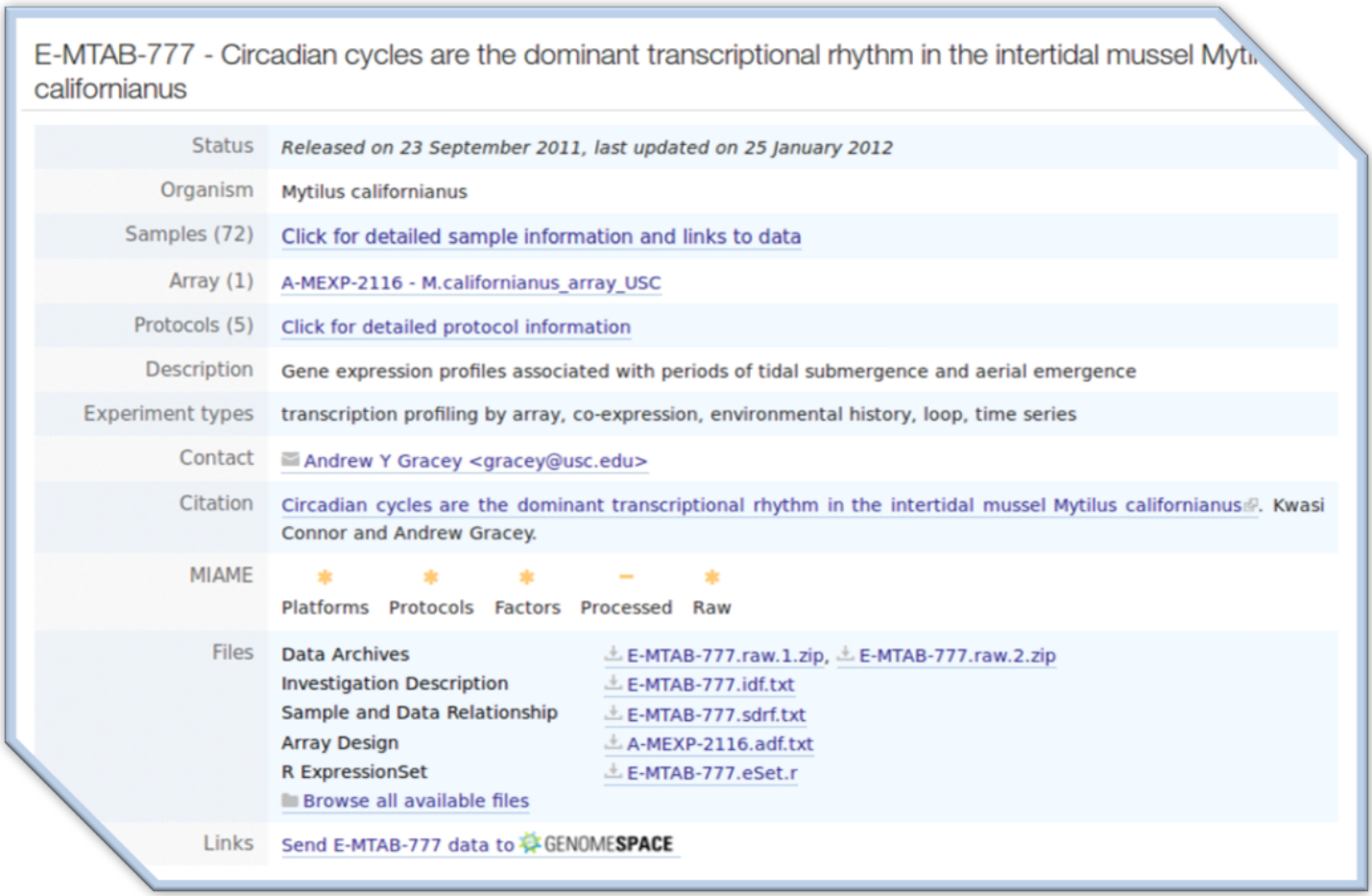
A database of functional genomics data, part of EMBL-EBI. Data in ArrayExpress is gathered from researchers (using Annotare tool) and imported from GEO. It contains information on experiments, assays, raw data files of microarray and high-throughput sequencing (HTS) analysis that are described and archived according to the community guidelines for microarray (MIAME) and HTS (MINSEQE).

QuickStart Guide

1. Experiments table
    a. One row per experiment

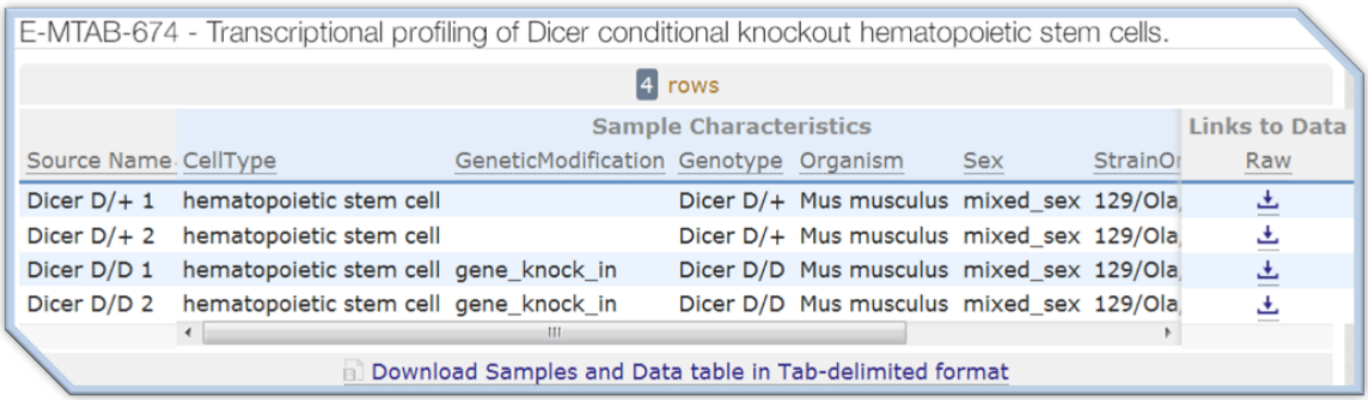


2. Single experiment overview
    a.



3. Samples table

a. Here, samples are in rows and sample attributes are in columns. You'll also find direct link to data files for each sample in the last column.



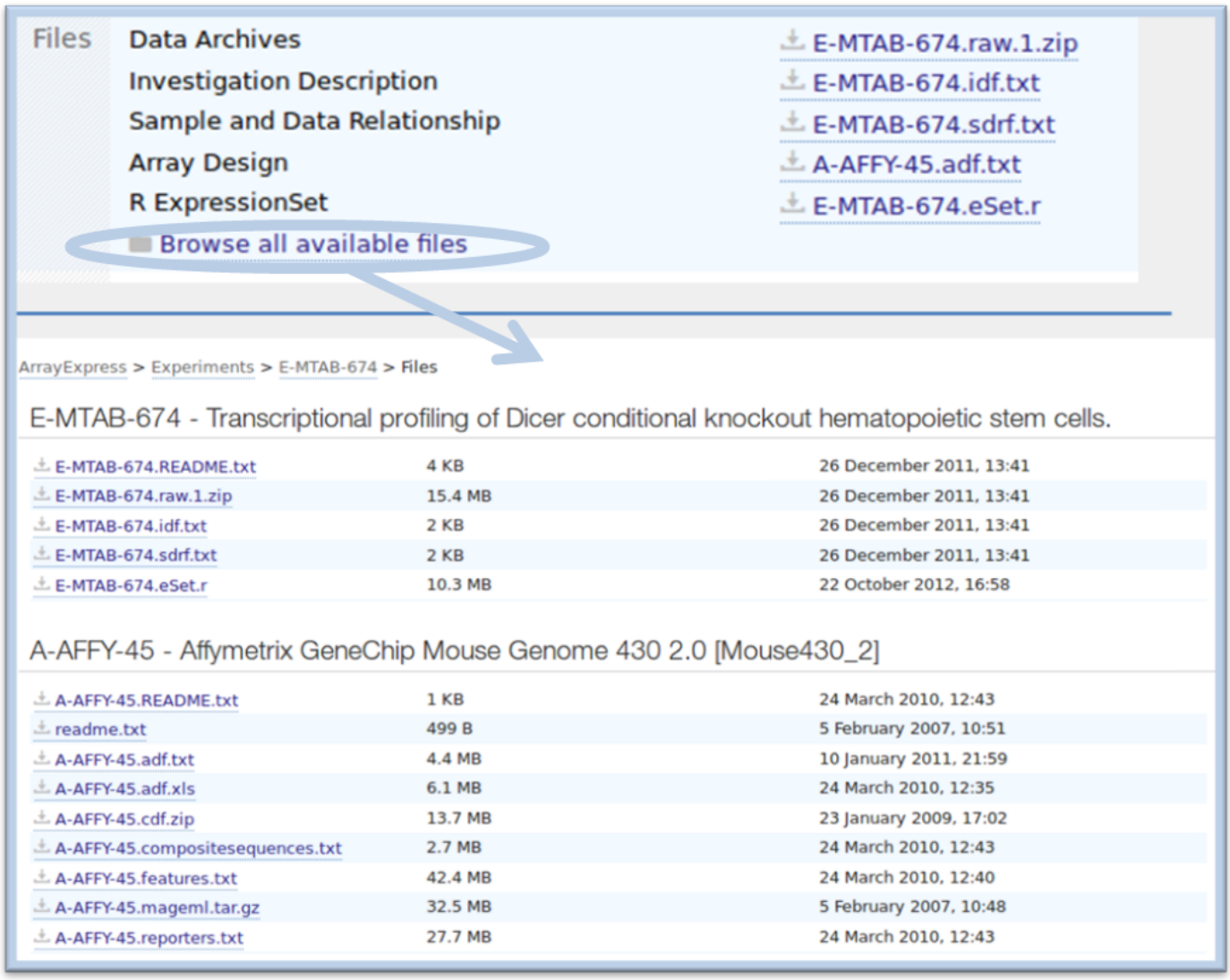
E-MTAB-674 - Transcriptional profiling of Dicer conditional knockout hematopoietic stem cells.

4 rows

	Sample Characteristics						Links to Data
Source Name	CellType	GeneticModification	Genotype	Organism	Sex	StrainOr	Raw
Dicer D/+ 1	hematopoietic stem cell		Dicer D/+	Mus musculus	mixed_sex	129/Ola	
Dicer D/+ 2	hematopoietic stem cell		Dicer D/+	Mus musculus	mixed_sex	129/Ola	
Dicer D/D 1	hematopoietic stem cell	gene_knock_in	Dicer D/D	Mus musculus	mixed_sex	129/Ola	
Dicer D/D 2	hematopoietic stem cell	gene_knock_in	Dicer D/D	Mus musculus	mixed_sex	129/Ola	

Download Samples and Data table in Tab-delimited format

4. Files for download
   a. On the Single experiment overview page, you can download all files (e.g. sample annotations, raw data files) for the experiment using links in the "Files" section



Files

Data Archives — E-MTAB-674.raw.1.zip
Investigation Description — E-MTAB-674.idf.txt
Sample and Data Relationship — E-MTAB-674.sdrf.txt
Array Design — A-AFFY-45.adf.txt
R ExpressionSet — E-MTAB-674.eSet.r
Browse all available files

ArrayExpress > Experiments > E-MTAB-674 > Files

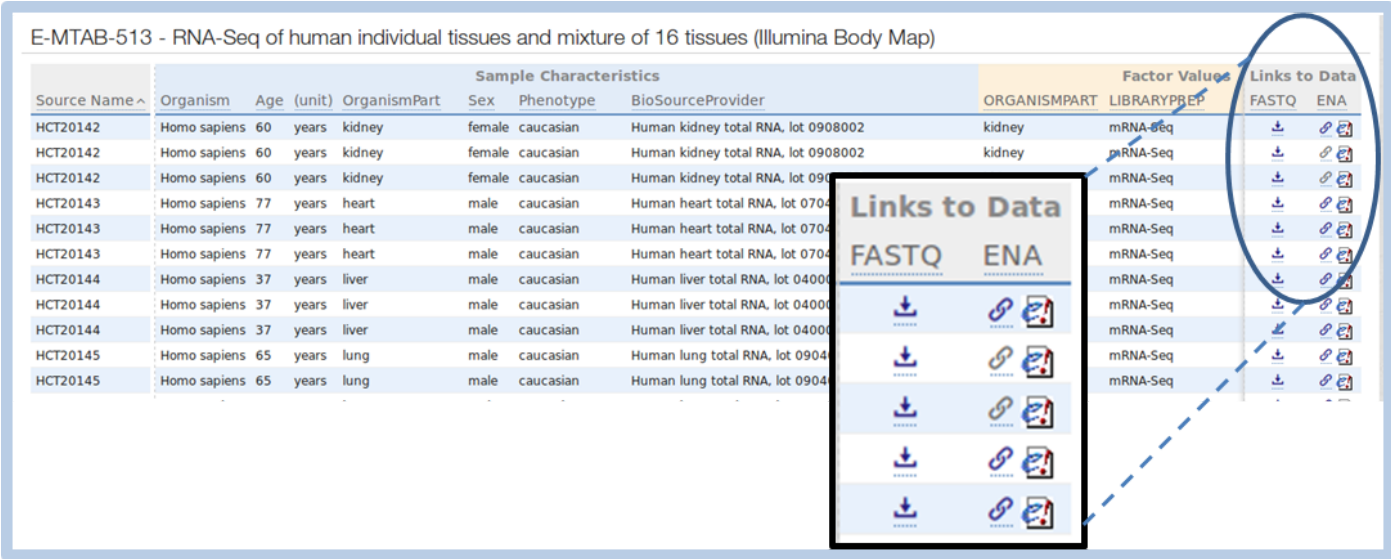
E-MTAB-674 - Transcriptional profiling of Dicer conditional knockout hematopoietic stem cells.

E-MTAB-674.README.txt	4 KB	26 December 2011, 13:41
E-MTAB-674.raw.1.zip	15.4 MB	26 December 2011, 13:41
E-MTAB-674.idf.txt	2 KB	26 December 2011, 13:41
E-MTAB-674.sdrf.txt	2 KB	26 December 2011, 13:41
E-MTAB-674.eSet.r	10.3 MB	22 October 2012, 16:58

A-AFFY-45 - Affymetrix GeneChip Mouse Genome 430 2.0 [Mouse430_2]

A-AFFY-45.README.txt	1 KB	24 March 2010, 12:43
readme.txt	499 B	5 February 2007, 10:51
A-AFFY-45.adf.txt	4.4 MB	10 January 2011, 21:59
A-AFFY-45.adf.xls	6.1 MB	24 March 2010, 12:35
A-AFFY-45.cdf.zip	13.7 MB	23 January 2009, 17:02
A-AFFY-45.compositesequences.txt	2.7 MB	24 March 2010, 12:43
A-AFFY-45.features.txt	42.4 MB	24 March 2010, 12:40
A-AFFY-45.mageml.tar.gz	32.5 MB	5 February 2007, 10:48
A-AFFY-45.reporters.txt	27.7 MB	24 March 2010, 12:43

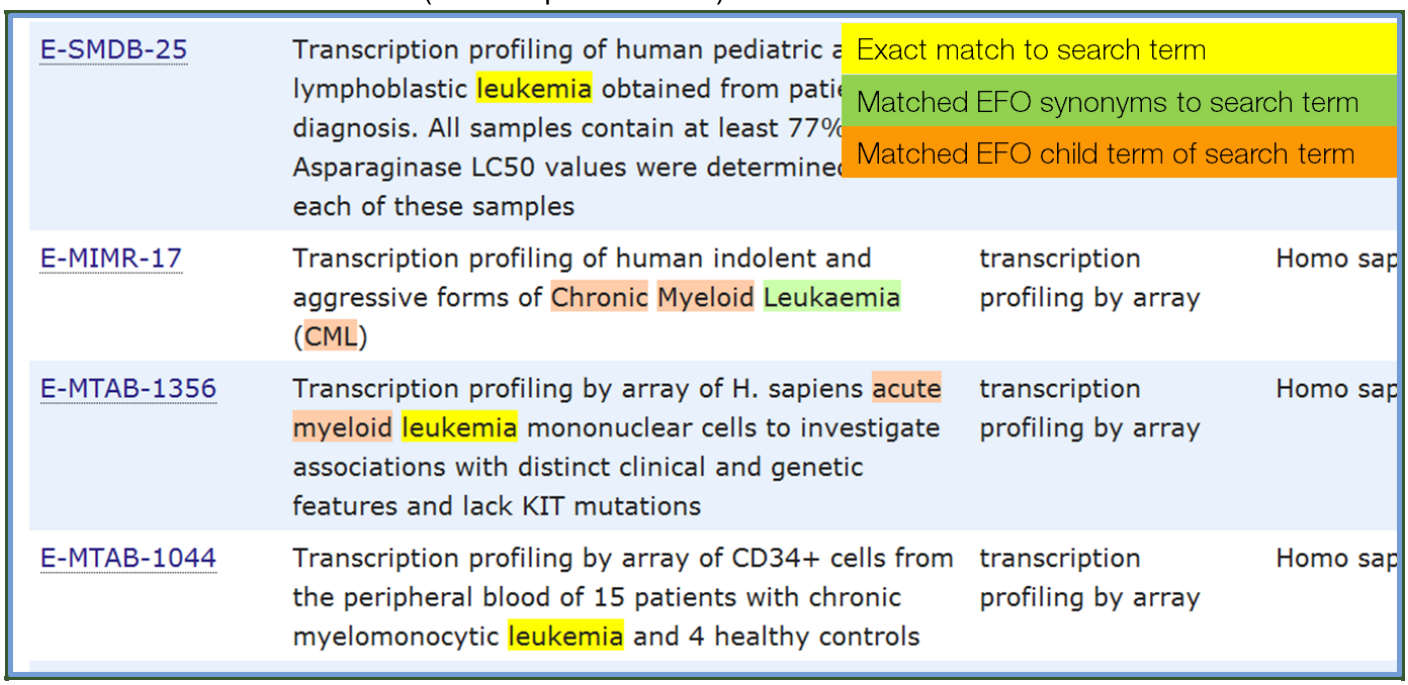
5. BAM files for RNA-seq experiments
   a. For a subset of RNA-seq experiments in ArrayExpress, e.g.https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-822/samples/, we created BAM files by processing the experimental raw FASTQ data files with the ArrayExpressHTS BioConductor package as part of the RNA-seq data processing pipeline. The BAM files are available for download and can also be viewed in Ensembl.



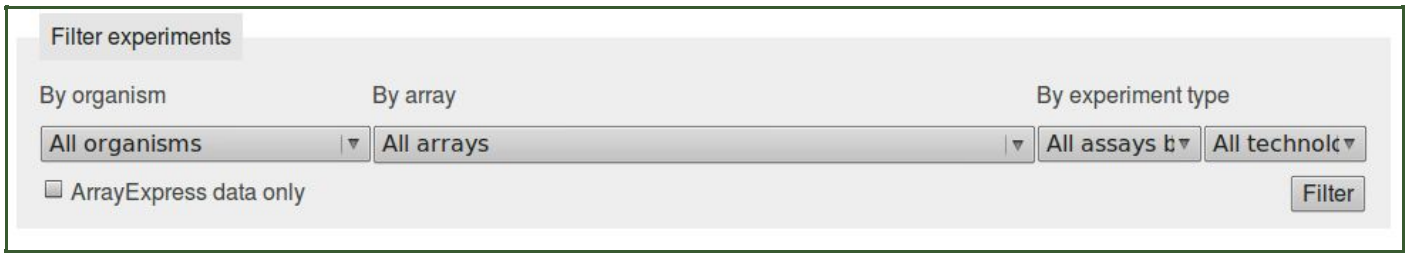
6. More help

One can use ArrayExpress to:

- Search by keywords or experiment's properties (e.g. citation, transcriptomics platform, species or sample annotation) and identify experiments of interest;
    - 1.1 Search by accession or keyword
    - 1.2 Search term expansion by Experimental Factor Ontology (EFO)
        - extend your query to synonyms (e.g. "cerebral cortex" and "adult brain cortex") and EFO child-terms
        - Search terms are expanded using EFO by default for keyword-based searches and all relevant fields:
            - organism
            - exptype (for "experiment type")
            - expdesign (for "experiment design")
            - ef (for "experimental factor")
            - efv (for "experimental factor value")
            - sa (for "sample attribute")



    - http://www.ebi.ac.uk/arrayexpress/help/how_to_search.html
- Filtering results

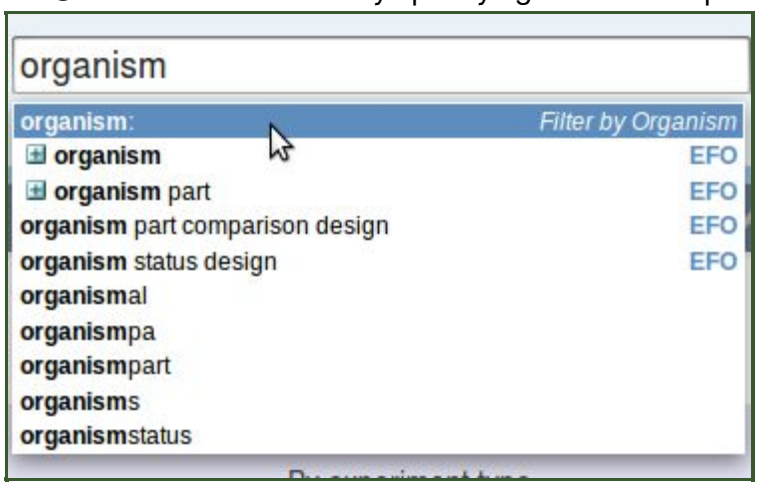


Experiments can be filtered using the drop down menus at the top of the results table. They can be filtered by:

- organism
- array design used
- molecule (DNA, RNA, amplicon, metabolite, protein)
- **technology (array, high-throughput sequencing, mass spectrometry)**
- ArrayExpress data only - experiments submitted directly to ArrayExpress, not imported from the NCBI Gene Expression Omnibus (GEO). Use the 'ArrayExpress data only' checkbox to activate this filter. For more information about the data we import from GEO see the GEO data help page.
- Advanced search
  - Boolean

Search terms	How the terms are interpreted	
1	heart brain	Search for experiments which mention both "heart" and "brain". The "AND" relationship between search terms is implicit.
2	heart, brain	The same as (1).
3	heart AND brain	The explicit way of writing the same query as in (1) and (2).
4	"heart brain"	Search for experiments which mention "heart brain" (two terms adjacent to each other).
5	"heart and brain"	Search for experiments which mention "heart and brain" (three words strung together as such).
6	heart OR brain	Search for experiments which mention either "heart" or "brain".
7	heart NOT brain	Search for experiments which mention "heart" but not "brain".

  - 2.2 Restrict search by specifying the search space



  - Searchable field:

Field name	Search scope	Example use case

accession	Experiment primary or secondary accession	accession:E-MTAB-1234
array	Array design accession or name	array:A-AFFY-33
ef	Experimental factor, the name of the main variable under study in an experiment. E.g. if the factor is "sex" in a human study, the researchers would be comparing between male and female samples, and "sex" is not merely an attribute the samples happen to have. Has EFO expansion.	ef:genotype
efv	The value of an experimental factor. E.g. The values for "genotype" factor can be "wild type", "p53-/-". Has EFO expansion.	efv:"wild type"
expdesign	Experiment design type, related to the questions being addressed by the study, e.g. "time series design", "stimulus or stress design", "genetic modification design". Has EFO expansion.	expdesign:"time series"
exptype	Experiment type, related to the assay technology used. Has EFO expansion.	exptype:"RNA-seq of coding RNA"
gxa	Presence/absence of an ArrayExpress experiment in the Expression Atlas. Use values "true" and "false" respectively.	gxa:true
pmid	PubMed identifier for a publication.	pmid:16553887
sa	Sample attribute values. Has EFO expansion.	sa:fibroblast
organism	Species of the samples. Can use common name (e.g. "mouse") or binomial nomenclature/Latin names (e.g. "Mus musculus"). Has EFO expansion.	organism:"homo sapiens"

- Filtering results by counts

Experiments fulfilling certain count criteria can also be searched for. E.g. Those having more than 10 assays (hybridizations). Here are some examples:

Filter	Query format	What is filtered	Example
Number of assays	assaycount:[x TO y]	Filter on the number of of assays where x <= y and both values are between 0 and 99,999 (inclusive) . To count excluding the values given, use curly brackets e.g. assaycount:{1 TO 5} will find experiments with 2-4 assays. Single numbers may also be given e.g. assaycount:10 will find experiments with exactly 10 assays.	assaycount:[1 TO 5]
Number of experimental factors	efcount:[x TO y]	Filter on the number of experimental factors (the main variables under study in an experiment, e.g. "sex", "genotype", "strain".)	efcount:[1 TO 5]
Number of samples	samplecount:[x TO y]	filter on the number of samples	samplecount:[1 TO 5]
Number of sample attribute categories	sacount:[x TO y]	filter on the number of sample attribute categories. A category can be "patient ID", "treatment", "sex", "diet".	sacount:[1 TO 5]
Raw data files	raw:true/false	filter on the presence/absence of raw data files (native files obtained directly from microarray scanner or sequencing machine)	raw:true
Processed data files	processed:true/false	filter on the presence/absence of processed data files (e.g. normalised/transformed data)	processed:true

Presence of adequate meta-data (MIAME score)	miamescore:[x TO y]	filter on the MIAME compliance score (maximum score is 5)	miamescore:[1 TO 5]
Release date	date:yyyy-mm-dd	filter by release date ○ date:2009-12-01 - will search for experiments released on 1st of Dec 2009 ○ date:2009* - will search for experiments released in 2009 ○ date:[2008-01-01 2008-05-31] - will search for experiments released between 1st of Jan and end of May 2008	date:[2008-01-01 2008-05-31]

- Sorting asc/desc
  - accession
  - name
  - assays
  - species
  - releasedate
  - fgem (for "final gene expression matrix", i.e. processed data)
  - raw
  - atlas

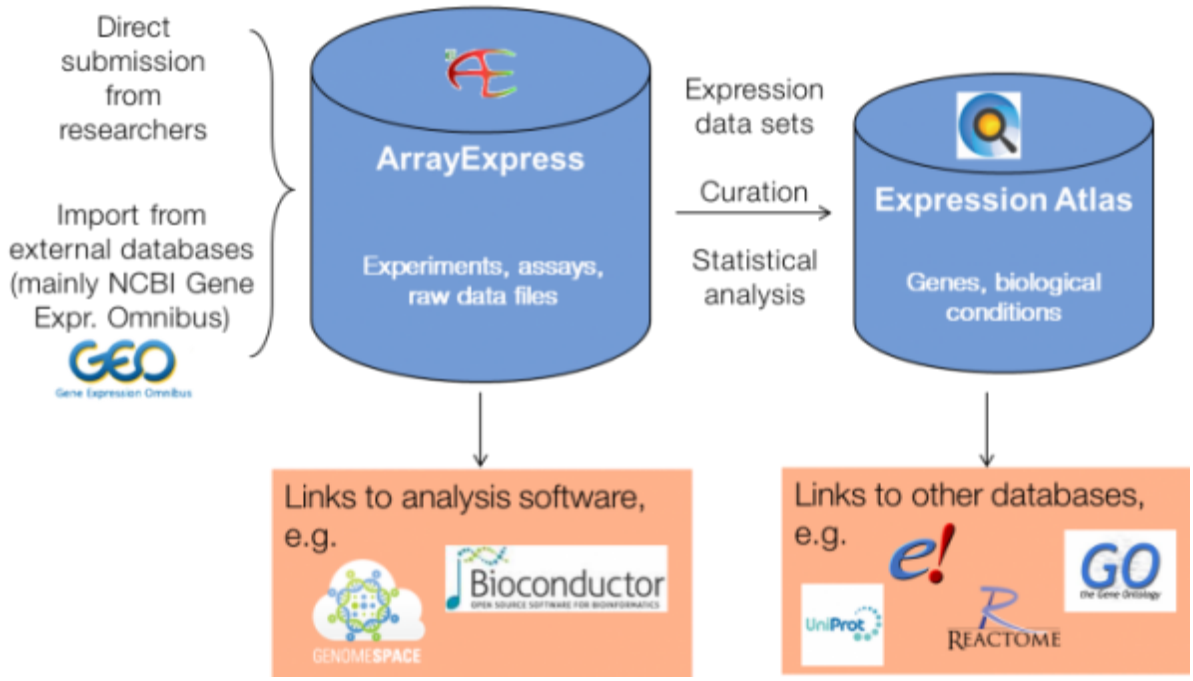
Using AE one can:

- Download data associated with experiment(s), alongside its annotation, for your own analysis;
- Submit microarray or HTS data that you want to publish.
  - Formats of the data:
    - MTAB = data submitted via the MAGE-TAB (discontinued since September 2014) or Annotare submission route
      - http://fged.org/projects/mage-tab/
    - GEOD = data imported from NCBI Gene Expression Omnibus
    - discontinued:
      - MEXP = data submitted via the MIAMExpress submission route (discontinued since July 2014)
      - TABM = data submitted via the Tab2MAGE submission route (discontinued since January 2012)
-

Curated data is stored into **Expression Atlas**. A baseline is calculated for normal/untreated tissues/common cell lines. Data is then compared to the baseline and annotations for genes being up-regulated, down-regulated, or not regulated are added for differential experimental conditions (like a cancer cell vs normal cell).
Contains links to analysis software, like **Bioconductor** – a project which also provides tools for comprehension of high-throughput genomic data. It is based primarily on the R programming language.
Links

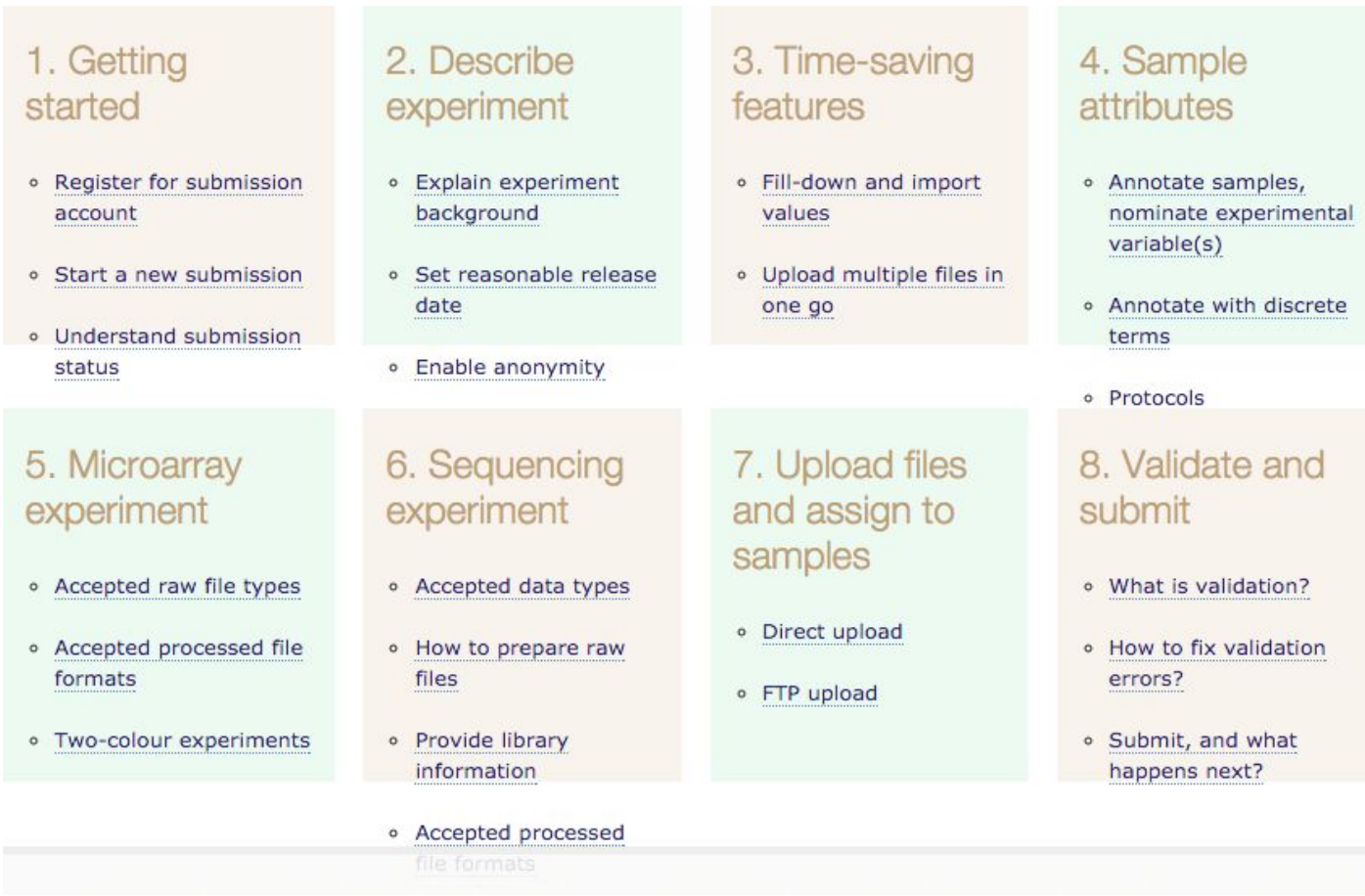
- https://www.ebi.ac.uk/arrayexpress/
- Training course: http://www.ebi.ac.uk/training/online/course/arrayexpress-quick-tour-1

Schema of relationship between ArrayExpress and Expression Atlas



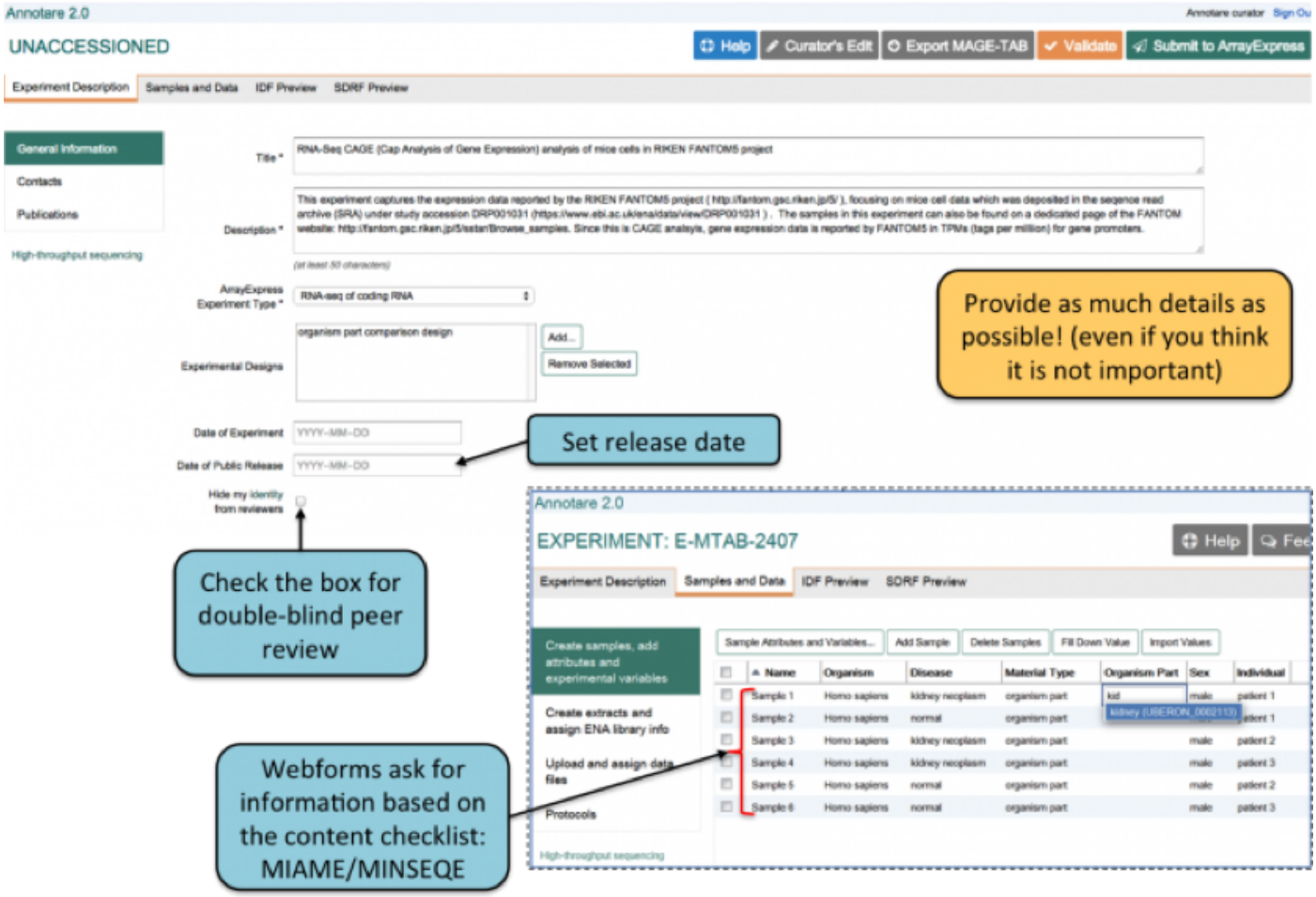
Questions

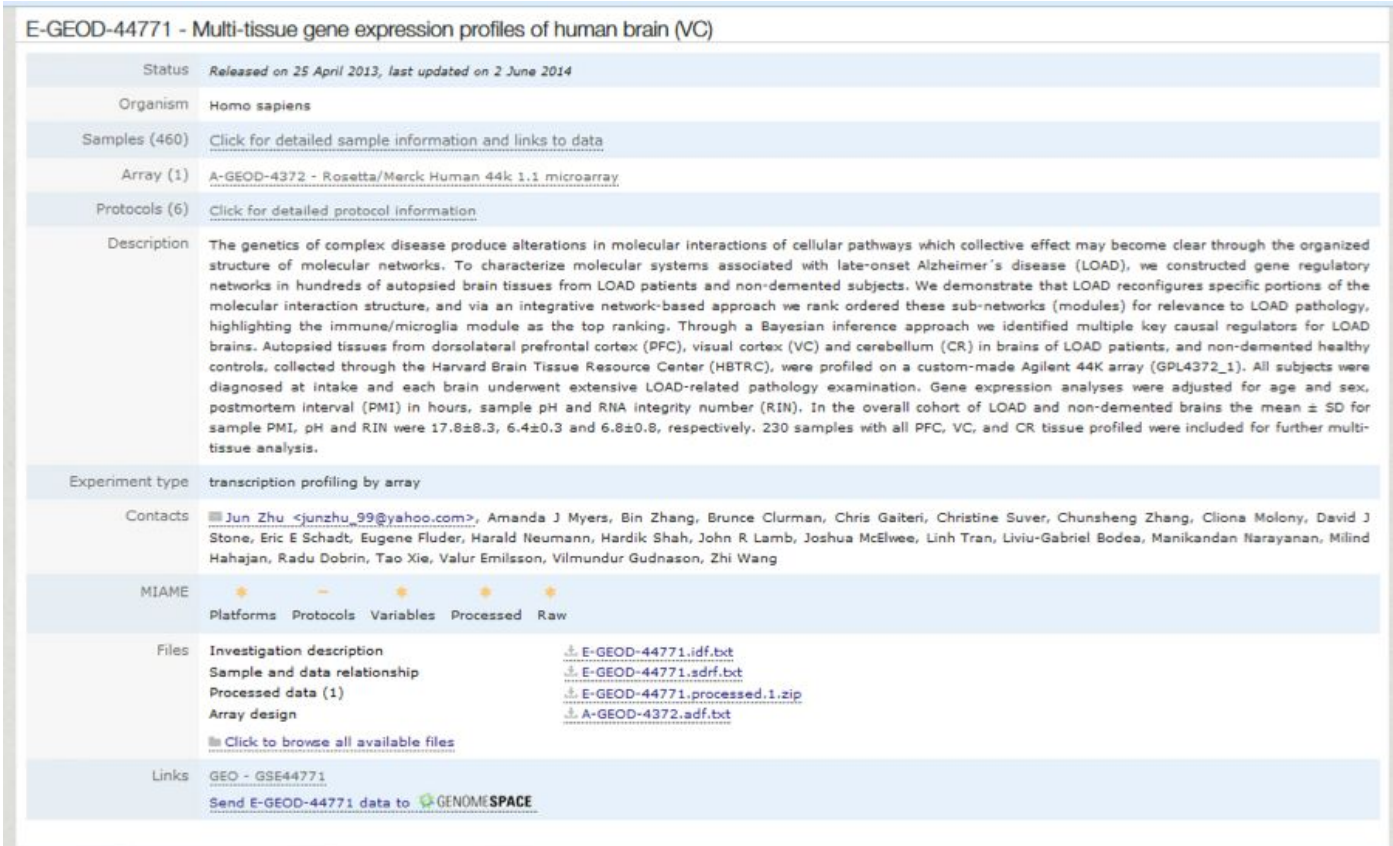
1. **Submission**
    a. ArrayExpress accepts all functional genomics data, e.g. expression profiling, genotyping, chromatin immunoprecipitation, copy-number variation, from microarray and sequencing technologies.
2. Through Annotare
    a. Annotare is a tool for submitting functional genomics experiments to ArrayExpress in MAGE-TABformat.



Following the MIAME/MINSEQE guidelines:

1) **Metadata**: Background biology of your experiment, aim of the experiment, biological materials/samples used and their characteristics, wet-lab and dry-lab procedures (protocols).
2) **Raw data files**: These are unprocessed data files obtained from the microarray scanner, or from the sequencing machine.
3) **Processed data files**: These are data files generated from the raw data files, often involving e.g. normalisation or background-subtraction. In sequencing experiments, the processing can also involve alignment of sequencing reads to a reference genome, calculation of RPKM/FPKM values, etc.





E-GEOD-44771 - Multi-tissue gene expression profiles of human brain (VC)

Status	*Released on 25 April 2013, last updated on 2 June 2014*
Organism	Homo sapiens
Samples (460)	Click for detailed sample information and links to data
Array (1)	A-GEOD-4372 - Rosetta/Merck Human 44k 1.1 microarray
Protocols (6)	Click for detailed protocol information
Description	The genetics of complex disease produce alterations in molecular interactions of cellular pathways which collective effect may become clear through the organized structure of molecular networks. To characterize molecular systems associated with late-onset Alzheimer's disease (LOAD), we constructed gene regulatory networks in hundreds of autopsied brain tissues from LOAD patients and non-demented subjects. We demonstrate that LOAD reconfigures specific portions of the molecular interaction structure, and via an integrative network-based approach we rank ordered these sub-networks (modules) for relevance to LOAD pathology, highlighting the immune/microglia module as the top ranking. Through a Bayesian inference approach we identified multiple key causal regulators for LOAD brains. Autopsied tissues from dorsolateral prefrontal cortex (PFC), visual cortex (VC) and cerebellum (CR) in brains of LOAD patients, and non-demented healthy controls, collected through the Harvard Brain Tissue Resource Center (HBTRC), were profiled on a custom-made Agilent 44K array (GPL4372_1). All subjects were diagnosed at intake and each brain underwent extensive LOAD-related pathology examination. Gene expression analyses were adjusted for age and sex, postmortem interval (PMI) in hours, sample pH and RNA integrity number (RIN). In the overall cohort of LOAD and non-demented brains the mean ± SD for sample PMI, pH and RIN were 17.8±8.3, 6.4±0.3 and 6.8±0.8, respectively. 230 samples with all PFC, VC, and CR tissue profiled were included for further multi-tissue analysis.
Experiment type	transcription profiling by array
Contacts	Jun Zhu <junzhu_99@yahoo.com>, Amanda J Myers, Bin Zhang, Brunce Clurman, Chris Gaiteri, Christine Suver, Chunsheng Zhang, Cliona Molony, David J Stone, Eric E Schadt, Eugene Fluder, Harald Neumann, Hardik Shah, John R Lamb, Joshua McElwee, Linh Tran, Liviu-Gabriel Bodea, Manikandan Narayanan, Milind Hahajan, Radu Dobrin, Tao Xie, Valur Emilsson, Vilmundur Gudnason, Zhi Wang
MIAME	Platforms – Protocols Variables Processed Raw
Files	Investigation description: E-GEOD-44771.idf.txt Sample and data relationship: E-GEOD-44771.sdrf.txt Processed data (1): E-GEOD-44771.processed.1.zip Array design: A-GEOD-4372.adf.txt Click to browse all available files
Links	GEO - GSE44771 Send E-GEOD-44771 data to GENOMESPACE

1. **Are images taken from microarray scanners part of the database schema of ArrayExpress?**
    a. Yes, see database schema. This is very important because the readout from the images is the first step in the analysis where different methods lead to different results.
2. **In ArrayExpress, you will find data sets with the designator "tiling array" or "genome tiling experiment". What is the difference between a "classical" microarray experiment and a genome tiling experiment? Explain!**
    a. (genome) tiling array
        i. A tiling array is an array which has short fragments of nucleic acid immobilized on a substrate. These are designed to cover the whole genome of the target species. Tiling arrays are used to determine genome binding in ChIP assays or to identify transcribed regions.
    b. genome tiling experiment – not exists in EFO. Maybe they meant, that they are synonyms.
    c. Another answer...
        i. e.g. Human genome Project
        ii. - microarray in tiling experiment consists of many pieces of genome (overlapping probes)
        iii. - represent a genomic region of interest
        iv. - the don't select few genes, but represent the whole genome cut into small pieces
        v. - found out, for example, that junk DNA is also expressed)))
        vi. - used for: analysis of alternative splicing, characterisation of methylation of DNA, SNPs analysis
3. **What is Experimental Factor Ontology (EFO)?**
    a. EFO combines terms from a subset of ontologies which are: well designed, actively maintained, have definitions, provide suitable coverage and which are compatible with EFO or which are the definitive resource. This includes:
        i. ChEBI - chemical compounds and their roles
        ii. Units ontology - SI and combinatorial units
        iii. Gene Ontology - cellular component and biological process terms
        iv. Ontology for Biomedical Investigation - experimental design terms, protocols

v. NCBI taxonomy - Species and strains
vi. Basic Formal Ontology - Upper level terms though we simplify the use of this ontology to a small subset for our needs

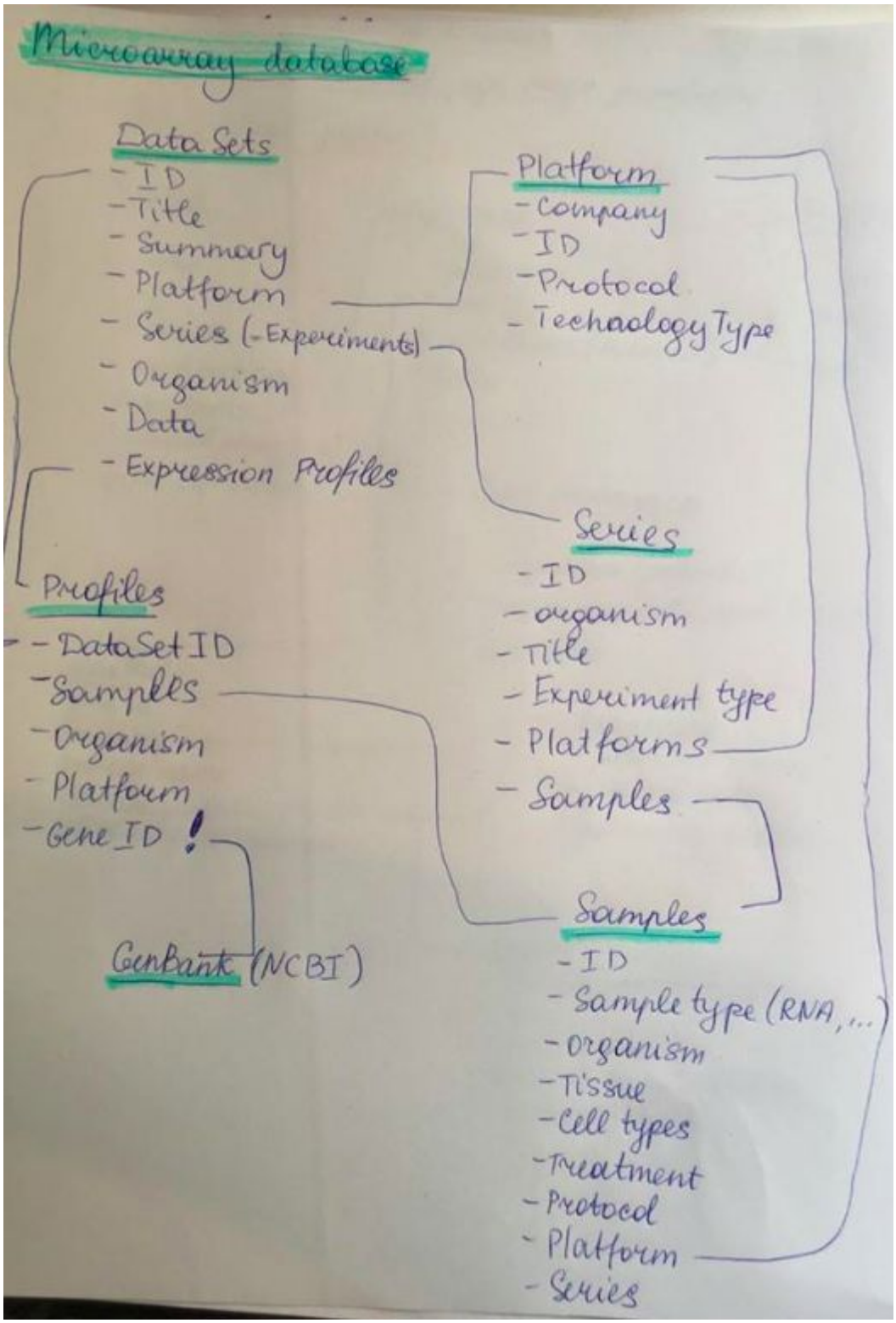
4. **What are the major entity types used in the object model of ArrayExpress?**
    a. Experiment as entity type, and you can also open from it the protocols, samples. You can't search genes!! All information in the experiments!
5. **ArrayExpress vs. GEO**

ArrayExpress	GEO
accession number – one experiment, in each experiment set of samples	More accession numbers: Information from a submitter: - Series = general info about the study, what was done, why and how. [GSE] - Samples = Data for each sample [GSM] - Platform = Info about the array [GPL] - DataSets - the processed (curated) experiment, not always. - Profile - comparison of gene expression
File formats: MAGE-TAB -idf - experiment information TXT -sdrf - metadata TXT -adf - platform information TXT -processed data - ZIP	File formats: - GEOarchive (spreadsheets, e.g., Excel) - SOFT (plain text) - MINiML (XML) - CEL, CHP files
connected: only to GEO (Series)	connected: to other NCBI databases, like Genbank, UniGene
Analysing tool: Genomespace	Analyzing tools: GEO2R

6. **Name at least two foreign keys that could link a microarray database to a nucleotide sequence database or UniProtKB.**
    a. accession number, gene/protein name
7. **Microarray and Gene Expression Data Ontology (MGED Ontology)?**
    a. MGED: organization of biologists, computer scientists, that aims to facilitate biological and biomedical discovery through data integration. They establish standards for data quality, management, annotation and exchange, creation of tools that leverage these standards.
    b. MGED Ontology – provides a standard terminology for describing components of a DNA microarray experiment.
8. **MAGE-TAB?**
    a. MicroArray Gene Expression - Tabular format (MAGE-TAB) is a MIAME-compliant, tab-delimited format used to annotate microarray data.
    b. In order to provide a common platform for sharing characterization data within the research community, the Microarray Gene Expression Data (MGED) Society developed the Minimum Information About a Microarray Experiment (MIAME) standard.
    c. For more information on MAGE-TAB, visit the FGED Society website.

File Type	Platform	Description
Investigation Description Format(IDF) .idf	mage-tab	Provides general information about the investigation, including its name, a brief description, the investigator's contact details, bibliographic references, and free text descriptions of the protocols used in the investigation.
Sample and Data Relationship Format(SDRF) .sdrf	mage-tab	Describes the relationships between samples, arrays, data, and other objects used or produced in the investigation, and providing all MIAME information that is not provided elsewhere. In TCGA SDRF files, a row represents an analyzed element (often an aliquot) in its most basic electronic form (i.e. raw data file) and the production of higher-level data files (Level 2 and 3) as protocols (e.g. normalization) are applied to the file and its derivatives. These protocols correspond to those listed in the IDF.
Array Design Format(ADF) .adf Not mandatory	mage-tab	Defines each array type used. An ADF file describes the design of an array, e.g., what sequence is located at each position on an array and what the annotation of this sequence is. An ADF may exist in the MAGE-TAB archive or through the Data Portal on the Platform Design page.

9. **Sketch a high level conceptual schema of a microarray database**



10. **Is there an absolute expression unit? How can we compare data from different technical platforms?**
    a. There are few purpose-built cross-platform normalization strategies. This feat data to an idealized data structure.
    b. e.g. Yugene
11. **Sketch the workflow from deep sequencing (RNA seq) to submission to ArrayExpress and shed some light on the impact that RNA seq has on standardization of experiment description.**

Microarrays	RNAseq

robust, reliable method, proven!!!	provides a comprehensive view of the transcriptome
high throughput method	not dependent on any prior sequence knowledge
can be easily automated	can detect structural variation such as gene fusions and splicing
lower cost	

a. RNA seq:
   i. - sample extraction from the tissue
   ii. - RNA extraction and purification
   iii. - Reverse transcription à cDNA
   iv. - Illumina sequencing
   v. - Read mapping à the long sequenced sequence of RNA (translated in cDNA)
   vi. - Statistical analysis based on raw count
   vii. - Submission of data to the ArrayExpress:
   viii. - Standardization using MIAME/MINSEQE guidelines (organism, samples, raw data, final data, experiment data, metadata, protocols)
b. Standardization: the RNA seq doesn't need the prior knowledge of genome, in comparison with microarray – maybe that's why standardize? Accuracy?

12. How are RNA seq data being integrated in Gene Builds in ENSEMBL?

a. - What is GenBuilds? Is the integrated presentation of the genome sequence where more databases are integrated (RNA seq ENA, dbSNP, ..)
b. - Raw RNASeq data is submitted only to ENA, and from there by curator to the ENSEMBL.
c. - In the ENSEMBL by the genome view ("Location") –you see the RNA seq annotated sequences.
d. - Raw data are integrated into the BWA (SOFTWARE)
e. - GeneBuilds in ENSEMBL using Ensembl Genes present in ENA as foreign key

13. Relate microarray DB to ENTREZ feature

a. ENTREZ – searching NCBI databases.
   i. 1) GEO datasets
   ii. 2) GEO profiles

14. Single spotted array process (e.g. Illumina)

a. • Explain just the normal process of the microarray
b. • Single (for one channel), spotted (the probes are spotted on the glass)

15. Explain the procedure for Expression databases

a. Annotation, Curation, normalization
b. Statistical analysis: like in Geo Profiles, the expression patterns, ... or Expression Atlas! Topological analysis: like RNA seq data are integrated into ENSEMBL.

Expression Atlas (the 'Atlas')

A database containing analysed gene expression data derived from sets (gene expression patterns under different biological conditions, like different cell types, organisms parts, diseases, compound treatments and genotypes) stored in ArrayExpress. Contains two components:

- Baseline[2] Atlas - containing exclusively RNA-seq data, displays expression levels of gene products under 'normal' conditions (e.g. normal human tissues).

[2] a minimum or starting point used for comparisons

- Differential Atlas - allowing queries on genes that are up- or down- regulated in different experimental conditions, , e.g. 'in Arabidopsis shoots, what genes are upregulated in plants treated by X?'
    - containing both microarray and sequencing data

Unlike ArrayExpress which focuses on experiments, the Atlas focuses on genes and biological conditions, allowing you to ask biological questions such as:

- What genes are expressed in normal human liver?
- What genes are expressed across a panel of ENCODE cell lines?
- What genes are up- or down-regulated in drought and salt tolerance (DST) mutant Japanese rice plants vs wild type controls?

How the Atlas is produced

- The Atlas is composed of a sub-set of datasets from ArrayExpress, namely those on expression profiling, which are manually curated and then analysed in-house by a standard statistical pipeline.
- The manual curation step ensures only well-annotated data sets from well-designed experiments are included in the Atlas.
    - For example, for an experiment to be considered for the differential atlas, it must have at least three biological replicates for each condition for proper downstream statistical analysis, and the intent of the experiment must be clear.
- Various quality-control metrics are also implemented during statistical analysis to discard sub-standard data, e.g. microarray data with lots of background noise.
    - Microarray data must be gene expression data (rather than ChiP-chip or CGH for example), it must have a sufficient number of replicates to allow robust statistical analysis, and it should be possible to re-annotate the array design against Ensembl.
    - For RNA-seq data the sequences must be of high quality, there should a good quality reference genome build available, there must be sufficient replicates for statistical analysis for differential experiments and the tissues must be 'normal' if the experiment is be used in the Baseline Atlas.
        - We sometimes only include part of an experiment in the Expression Atlas because (1) there are not sufficient replicates of all the sample groups within an experiment, or ...
        - ... (2) the hybridization or sequencing was not of high enough quality. If there are still enough assays in the experiment after the removal of those with too few replicates or low quality then we continue processing the experiment for the Expression Atlas.
- **Quality control**
    - Microarray data quality
        - is assessed using the ***arrayQualityMetrics*** package in R. Outlier arrays are detected using distance measures, boxplots, and MA plots. If an array is classed as an outlier by all three methods, it is excluded from further analysis. Please see the arrayQualityMetrics documentation for more details on the methods used.
    - RNA-seq reads
        - are discarded based on several criteria. First, reads with quality scores less than Q10 are removed.
        - Second, the reads are mapped against a contamination reference genome (E. coli for animal data, fungal and microbial non-redundant reference for plants). Any reads that map to the contamination reference are removed.
        - Third, reads with "uncalled" characters (i.e. "N"s) are discarded.
        - Lastly, for paired-end libraries, any reads whose mate was lost in the previous three steps are also discarded. Please see the iRAP documentation for more details on the methods used.
- **Data analysis**
    - Raw single-channel microarray intensities are normalized using RMA via the oligo package from Bioconductor (Affymetrix data) or using quantile normalization via the

limma package (Agilent data). Two-channel Agilent data is normalized using LOESS via the limma package. Pairwise comparisons are performed using a moderated t-test for each gene using limma.

- RNA-seq data is analysed using the iRAP pipeline. FASTQ files are mapped to the reference genome from Ensembl using TopHat. Raw counts are generated using htseq-count. FPKMs are calculated from the raw counts by iRAP pipeline. Pairwise comparisons are performed using a conditioned test based on the negative binomial distribution, using DESeq.

Searching

How do I find out which genes are expressed in my favourite tissue?

Use the Experimental conditions search box on the home page to search all of Expression Atlas for organism parts e.g. kidney. Your query is expanded using EFO, so that this search will also return results matching synonyms and child terms of kidney in EFO. You will see results about baseline and differential expression of genes in the organism part(s) you searched for. You can narrow down the search to specific genes by also typing gene identifiers in the Gene query search box.

Use the Organism part box on a Baseline experiment page (e.g. Illumina Body Map) to search within a single experiment. See the Baseline Atlas help for more information about searching within a Baseline experiment.

How do I search for multiple conditions at once?

Searching with space-separated experimental variables or sample characteristics finds experiments with any one of the terms you entered, or all of them. For example, searching with liver heart will find all experiments with both liver and heart as well as ones with either liver or heart.

If you want to only find experiments with all terms you entered, separate them with AND. For example, searching for "wild type" AND Col-0 will find only experiments annotated with wild type and Col-0, but not those with only one of the terms.

How do I search for my favourite gene?

Use the Gene query search box on the home page to search all of Expression Atlas for genes. You will see results about baseline and differential expression for the gene or genes you searched for. You can narrow down the search to specific experimental conditions by also typing something in the Experimental conditions search box.

Use the Gene query search box on an experiment page (e.g. Illumina Body Map) to search for genes within a single experiment.

How do I search for multiple genes at once?

Enter gene IDs and/or keywords separated by spaces. Put multi-word search terms in quotes, e.g. "transcription factor binding".

What gene identifiers can I use to search?

You may use the following identifiers to search using the Gene query box:

- EMBL e.g. AC000015
- Ensembl Gene e.g. ENSG00000171658
- Ensembl Protein e.g. ENSP00000000233
- Ensembl Transcript e.g. ENST00000000233
- Gene Ontology ID e.g. GO:0008134
- Gene Ontology term e.g. "transcription factor binding" (please put multi-word query terms in quotes)
- Interpro e.g. IPR017892
- RefSeq e.g. NM_212505
- UniProt e.g. Q8IZT6
- UniProt Metabolic Enzyme e.g. O15269
- gene or protein name e.g. WNT1 or Wnt-1
- keyword e.g. transcription
- gene name synonym e.g. Calmbp1
- Ensembl gene biotype e.g. protein_coding or non_coding. See the glossary at Ensembl for more details.

How do I find a particular experiment?

All experiments currently in Expression Atlas are listed on the Experiments page. Click on the Description to see the experiment in Expression Atlas. Click on the ArrayExpress accession number (e.g. E-MTAB-1066) to see the experiment in ArrayExpress. If you know the ArrayExpress accession of the experiment you want to see, you can link to the experiment in Expression Atlas using the following format:

http://www.ebi.ac.uk/gxa/experiments/<ArrayExpress accession>

E.g. http://www.ebi.ac.uk/gxa/experiments/E-MTAB-1066

Check other questions on this page.

iRAP: RNA-seq analysis tool

iRAP is a flexible pipeline for RNA-seq analysis that integrates many existing tools for filtering and mapping reads, quantifying expression and testing for differential expression. iRAP is used to process all RNA-seq data in Expression Atlas

Questions

1. **ArrayExpress vs. The Atlas**
    a. The difference between the two databases is that
        i. ArrayExpress is built around experiments (containing information on data files, sample annotation and others),
        ii. whereas the Atlas is built around genes and biological conditions and is used to visualise changes in gene expression associated with different biological conditions.
2. **If you ever visited ArrayExpress, you should have read about "Gene Expression Atlas". What does the "Gene Expression Atlas" comprise?**
    a. provides information on gene expression patterns under different biological conditions
    b. both microarray and RNA-seq data
    c. curated!
    d. detect interesting expression patterns
    e. search: gene names, diseases, organisms, cell types
    f. Baseline Atlas:
        i. which gene products are present (and at what abundance) in "normal" conditions (e.g. tissue, cell type)
        ii. which genes are specifically expressed in kidney?
        iii. what is the expression pattern for gene SAA4 in normal tissues?
        iv. Highly curated and quality checked RNA seq
        v. Animal and plant species
    g. Differential Atlas:
        i. Microarray and RNA-seq experiments
        ii. Statistical analysis (to identify genes with a high probability of expression) – so only computer analysis

Links:

- https://www.ebi.ac.uk/gxa/home
- Quick start: https://www.ebi.ac.uk/gxa/help/index.html

UniGene

UniGene computationally identifies transcripts from the same locus; analyzes expression by tissue, age, and health status; and reports related proteins (protEST) and clone resources.

UniGene is an NCBI **database of the transcriptome** and thus, despite the name, not primarily a database for genes. Each entry is a set of transcripts that appear to stem from the same transcription locus (i.e. gene or expressed pseudogene). Information on protein similarities, gene expression, cDNA clones, and genomic location is included with each entry. (https://en.wikipedia.org/wiki/UniGene)

Features

1. **UniGene clusters often have an expression such as "ESTs, highly similar to ACTIN 1" or "weakly similar to..." How are the degrees of similarity defined?**
    a. **Basically, there are three distinctions of similarity:**
    b. 1. "Highly similar to" means >90% in the aligned region.
    c. 2. "Moderately Similar to" means 70-90% similar in the aligned region.
    d. 3. "Weakly similar to" means <70% similar in the aligned region.
2. **How are the protein similarities in the PROTSIM field of UniGene records calculated?**
    a. For each nucleotide sequence in UniGene, a search is made for sequence similarity to known proteins from eight organisms. This is done using Blastx.
    b. Organisms:
        i. *Homo sapiens,*
        ii. *Mus musculus,*
        iii. *Rattus norvegicus,*
        iv. *Drosophila melanogaster,*
        v. *Caenorhabditis elegans,*
        vi. *Saccharomyces cerevisiae,*
        vii. *Escherichia coli.*
        viii. *Arabidopsis thaliana.*
3. **I have seen cases where multiple genes, as identified in Entrez Gene, are linked to a UniGene cluster. Can you explain why this happens?**
    a. This can happen for several different reasons.
        i. When UniGene generates clusters based on alignment to the sequence of a genome (genome-based build), transcripts identified in Entrez Gene as being in different genes may all align to the same location (share intron-exon boundaries with the annotated gene). These co-placement data are used by RefSeq staff to review the curated GeneID/transcript relationship.
        ii. Multiple GeneIDs can also be associated with a UniGene cluster when UniGene uses a transcript-based build and some transcript sequences in the UniGene cluster are re-assigned to a different gene by Entrez Gene after the data freeze for the build.
4. Other questions here

- http://www.ncbi.nlm.nih.gov/unigene
- Typical Entry http://www.ncbi.nlm.nih.gov/nucest/BF732171.1

Nucleotide Sequence Databases

- Sequence submitted directly by scientists and genome sequencing group, and sequences taken from literature and patents
- entries are synchronized on a daily basis
  - Once your data has been accepted into one of these databases it will be included in the sequences aligned to the genome in the Ensembl gene build.
- accession numbers are managed in a constant manner
  - International Nucleotide Sequence Database Collaboration (INSDC)
  - [http://www.insdc.org/]
  - basically gives them the rights to hand out accession numbers for each sample.

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

- comparatively little error checking and fair amount of redundancy
- Gene
  - Locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions.
  - Attributes
    - Sequence, Organism, Locus in Chromosome, Gene name, Molecular weight, Coding regions, 5' and 3'

FASTA-format

text-based format for representing either nucleotide sequences or peptide sequences, where nucleotides or amino sequences are presented as a single letter.

>gi|31563518|ref|NP_852610.1| microtubule-associated proteins 1A/1B light chain 3A isoform b [Homo sapiens]

MKMRFFSSPCGKAAVDPADRCKEVQQIRDQHPSKIPVIIERYKGEKQLPVLDKTKFLVPDHVNMSEL V

IRRRLQLNPTQAFFLLVNQHSMVSVSTPIADIYEQEKDEDGFLYMVYASQETFGFIRENE

- ">" symbol is the identifier of the sequence, and the rest of the line is the description
- It is recommended that all lines of text be shorter than 80 characters.

European Nucleotide Archive (ENA)

– The ENA is part of EMBL-EBI. It contains sequences in the FASTA, FASTQ, CRAM, and other formats from different periods in their analyzed lifetimes. It contains short sequences and a section for whole genomes.

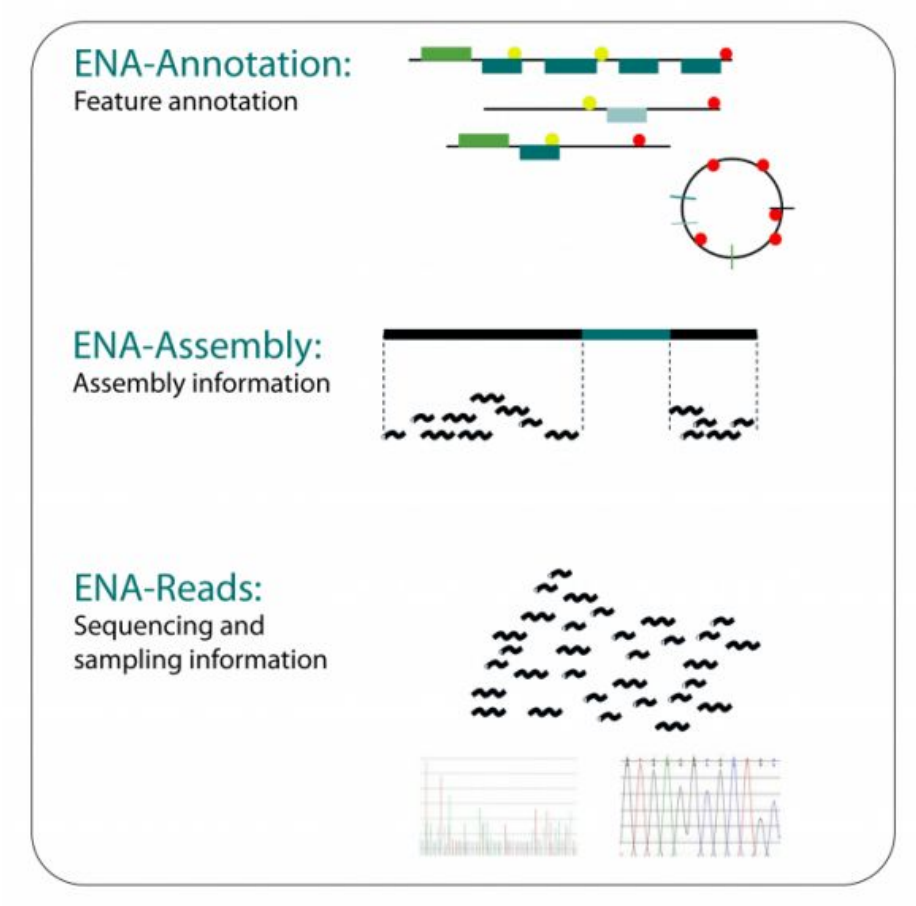
From Quick tour:

– The European Nucleotide Archive (ENA) provides a comprehensive, accessible and publicly available repository for nucleotide sequence data. The ENA attracts users from a multitude of research disciplines and serves as an underlying data infrastructure for other EBI services, including Ensembl, Ensembl

Genomes, UniProt and ArrayExpress. Data submitted to the ENA are validated by automated quality checking and, where possible, manual inspection and curation.

Once started as a primary database for assembled and annotated sequences, the ENA's remit has expanded enormously in response to advances in sequencing technology and the broad applications of sequence data. The ENA now incorporates raw data from electrophoresis-based sequencing machines as well as raw reads from next-generation sequencing platforms. By consolidating information from these three tiers, the ENA provides access to the whole scale of sequencing information: from raw data, through assembly and mapping information that relates very fragmented raw sequence reads into contigs and higher order structures, such as scaffolds and chromosomes, through to high-level functional annotation.

- **ENA data formats**
  - There are three tiers[3] within ENA providing a level of abstraction from the underlying infrastructure that has resulted from the integration of three legacy databases: the EMBL Nucleotide Sequence Database (EMBL-Bank), the Trace Archive and the Sequence Read Archive (SRA). The three ENA data tiers are:
    - **Reads**: sequencing machine output including base and colour calls, call qualities and signals.
    - **Assembly**: information relating overlapping fragmented sequence reads to contigs and higher order structures representing complete biological molecules, such as chromosomes.
    - **Annotation**: interpretations of biological function projected onto coordinate-defined regions of assembled sequence in the form of annotation.

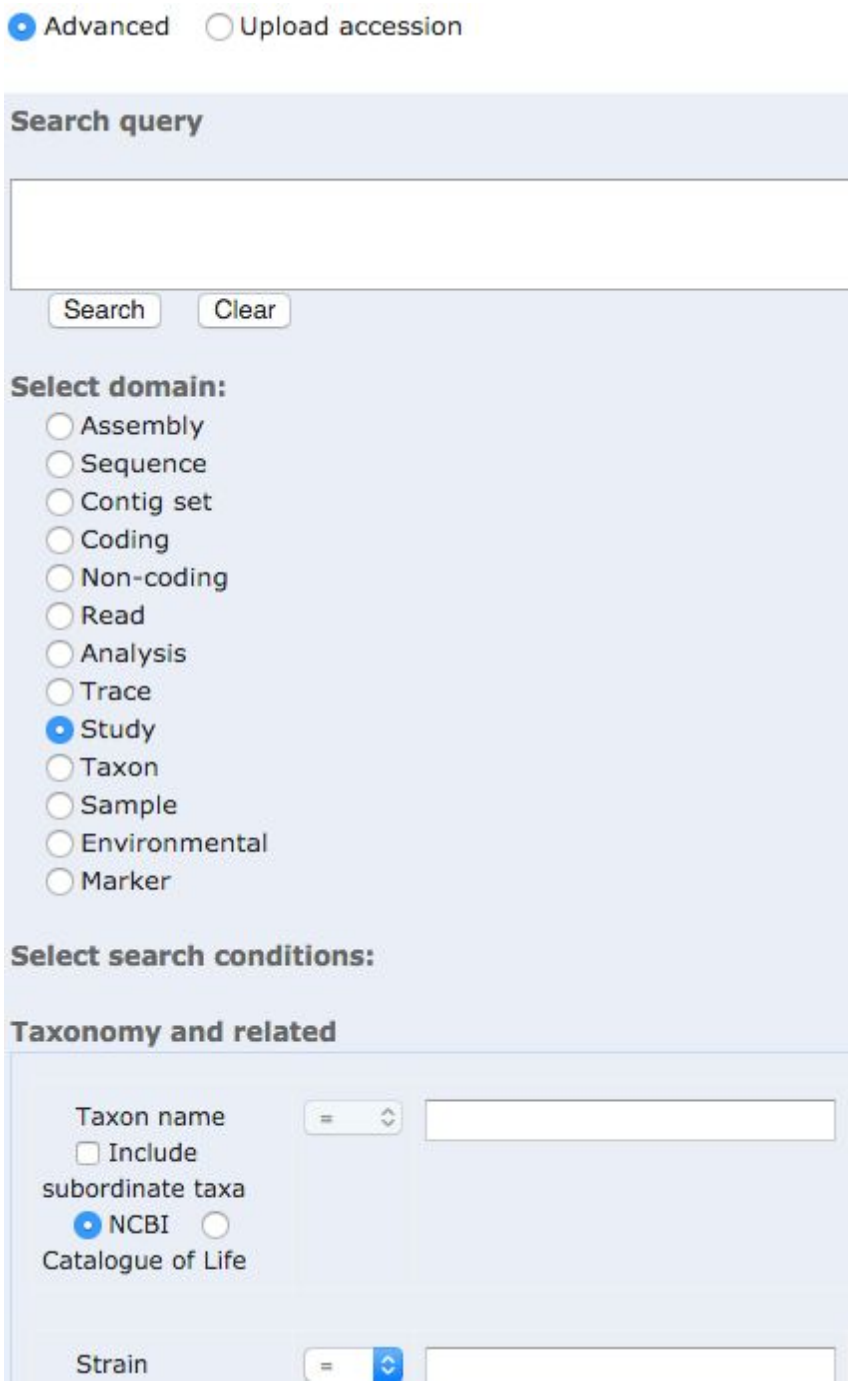


- **ENA data domains**
  - Data from the ENA tiers are organised into domains, each belonging typically to a single data tier but in some cases included in multiple tiers. Data types within ENA are:

[3] a level or grade within the hierarchy of an organization or system: companies have taken out a tier of management to save money

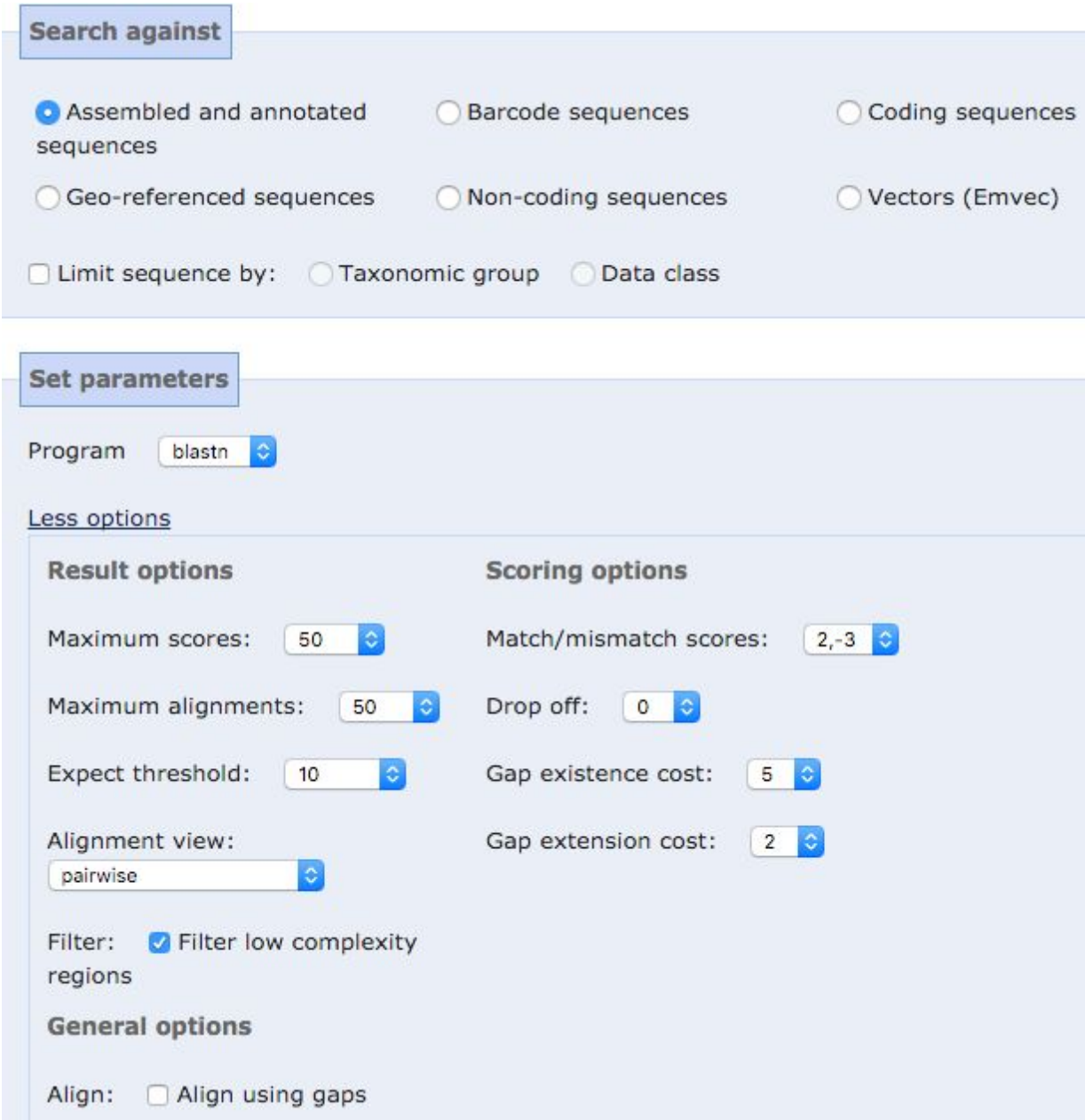
        - **Assembly**: information describing the construction of reads and sequence contigs into higher order scaffolds and chromosomes.
        - **Sequence**: assembled and, optionally, annotated assembled reads.
        - **Coding**: a virtual domain* comprising sequence regions reported by data providers as being protein-coding regions.
            - Virtual domains represent searchable and retrievable views of ENA data. Data in these domains are submitted as part of other domains from which the views are ultimately created.
        - **Non-coding**: a virtual domain* comprising sequence regions reported by data providers as representing non-protein-coding (RNA) genes.
        - **Marker**: a virtual domain* comprising information relating to phylogenetic, identification and molecular ecology marker data.
        - **Analysis**: derived data forms, such as recalibrated aligned reads and metabarcoding identifications.
        - **Read**: raw sequencing reads from next generation platforms.
        - **Trace**: raw sequencing data from capillary platforms.
        - **Taxon**: information relating to the organism that was the source of the sequenced biological sample.
        - **Sample**: information relating to the biological sample studied in the sequencing experiment.
        - **Study**: information relating to the scope[4] of the sequencing effort; also known as 'Project', the primary use of study is to unite content otherwise dispersed across the ENA domains.
        - (**Submission**): an accessory domain that serves to package submitted data; while useful for submitter-ENA communications, this domain has no lasting use beyond a submission transaction.
- **ENA data classes**
    - Data domains are further subdivided in some cases into data classes. Within a data class, data are presented uniformly. Please refer to the table below for a summary of ENA data classes and supported formats.
    - Table. Classes
- **Origin of data**
    - submissions of raw data,
        - Typical workflow includes
            - the isolation and preparation of material for sequencing,
            - a run of a sequencing machine in which sequencing data are produced and a subsequent bioinformatic analysis pipeline.
            - ENA records this information in a data model that covers input information (sample, experimental setup, machine configuration), output machine data (sequence traces, reads and quality scores) and interpreted information (assembly, mapping, functional annotation).
    - assembled sequences and annotation from small-scale sequencing efforts,
    - data provision from the major European sequencing centres and routine
    - comprehensive exchange with our partners in the International Nucleotide Sequence Database Collaboration (INSDC).
- **Searching**
    - **Free text search**
        - Free text search is provided from the search box in the header of all ENA web pages and through the search available at the top of all EMBL-EBI web pages. Advanced search options are available from the ENA Advanced Search page.

[4] the extent of the area or subject matter that something deals with or to which it is relevant: we widened the scope of our investigation



- **Programmatic data access**
  - The main programmatic interface for accessing ENA data is through the ENA Browser. The ENA Browser is designed to be accessed through REST URLs for easy programmatic access to retrieve data and metadata in a variety of formats.
  - **Search**
    - Both free text and advanced search can be accessed via REST URLs. These provide access to the complete functionality of ENA's advanced search as well as allowing users to download all data objects that match a given search.
  - **Data retrieval**
    - The main programmatic interface for accessing ENA data is through the ENA Browser. The ENA Browser is designed to be accessed through REST URLs for easy programmatic access to retrieve data and metadata in a variety of formats. The Taxon portal also has additional options to allow retrieval of all ENA data based on taxonomic classification.
  - **Sequence similarity search**

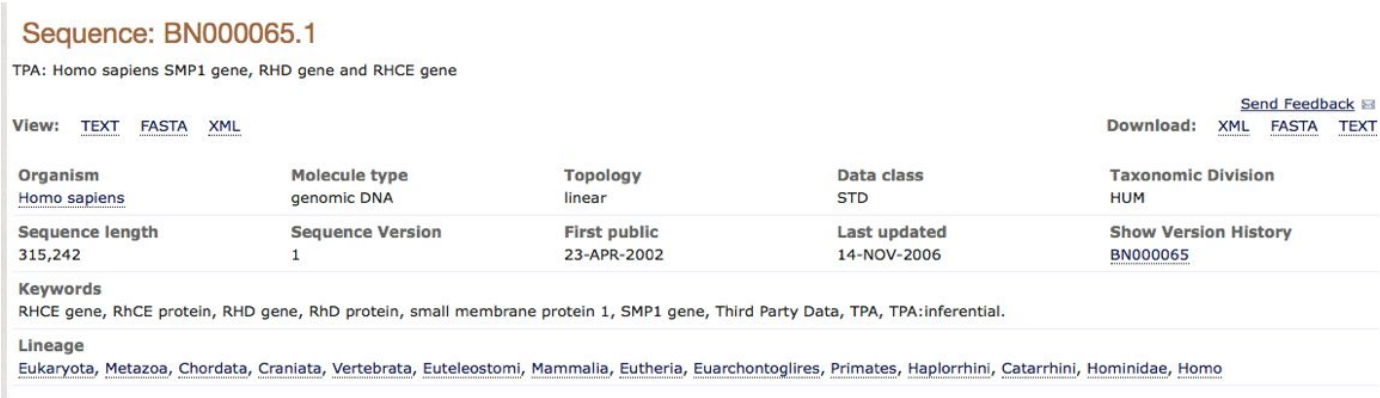
- EBI's central NCBI BLAST service can be accessed via REST and SOAP. For assistance matching options to those provided at ENA's sequence search



- **Bulk data download**
  - Most ENA data can be downloaded in bulk through FTP and Aspera protocols ...more information. The following datatypes are available for bulk download:
    - Assembled and annotated sequences
    - Read data
    - Taxonomy data
  - The main tool for downloading ENA data is the ENA Browser. The ENA Browser can be used both interactively and programmatically through REST URLs. All ENA data including assembled and annotated sequences is available for download through the ENA Browser.
  - Data in ENA can be searched via the search box in the header of all our pages. The search results are presented through the ENA Browser.

- **Sketch the major parts of a typical ENA entry: what categories does a “normal” ENA entry have?**
  - Molecule type:
    - linear genomic DNA
    - circular genomic DNA

        - other DNA
        - mRNA
        - other RNA
            - pre-RNA, rRNA, snoRNA, snRNA, tRNA, viral cRNA
    - 1) Sequence details
        - Length
        - version
    - 2) Released date
        - first public
        - last update
    - 3) Experimental design
    - 4) Source organism
    - 5) Sample details
    - 6) Lab and instrumental properties
    - 7) Environmental conditions
    - keywords



Sequence: BN000065.1

TPA: Homo sapiens SMP1 gene, RHD gene and RHCE gene

Send Feedback

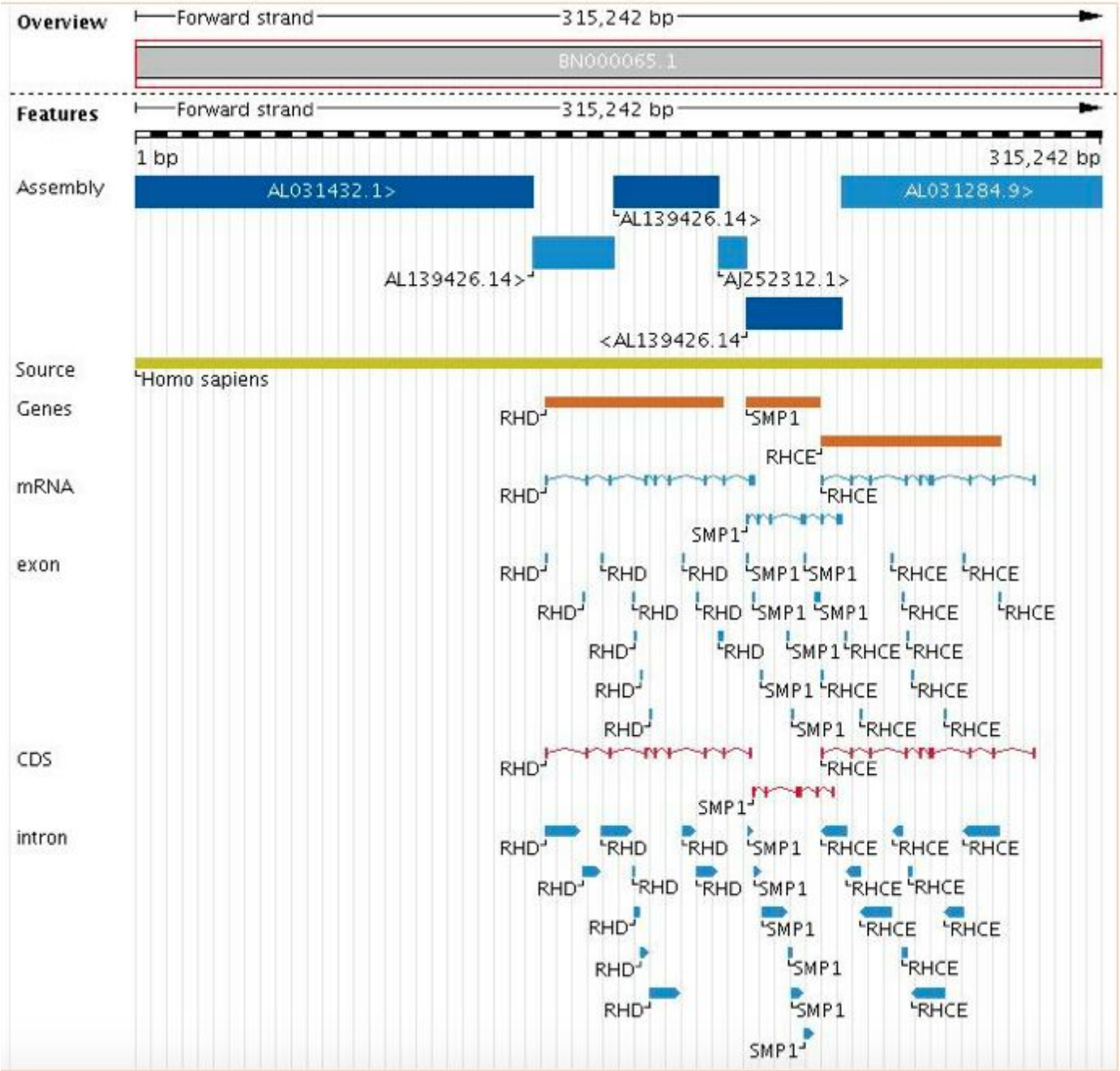
View: TEXT FASTA XML Download: XML FASTA TEXT

Organism	Molecule type	Topology	Data class	Taxonomic Division
Homo sapiens	genomic DNA	linear	STD	HUM
Sequence length	**Sequence Version**	**First public**	**Last updated**	**Show Version History**
315,242	1	23-APR-2002	14-NOV-2006	BN000065

Keywords
RHCE gene, RhCE protein, RHD gene, RhD protein, small membrane protein 1, SMP1 gene, Third Party Data, TPA, TPA:inferential.

Lineage
Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Euarchontoglires, Primates, Haplorrhini, Catarrhini, Hominidae, Homo

- Submitter is the owner of a sequence and he has the authority of changing and annotating sequences.
- **How is information on the exon-intron-structure of a gene represented in an ENA-entry?**
    - In form of annotations, using the FT (Feature Table) lines, which provide a mechanism for the annotation of the sequence data.
    - Different entities of molecule: genes, mRna, exons, CDS, intron.

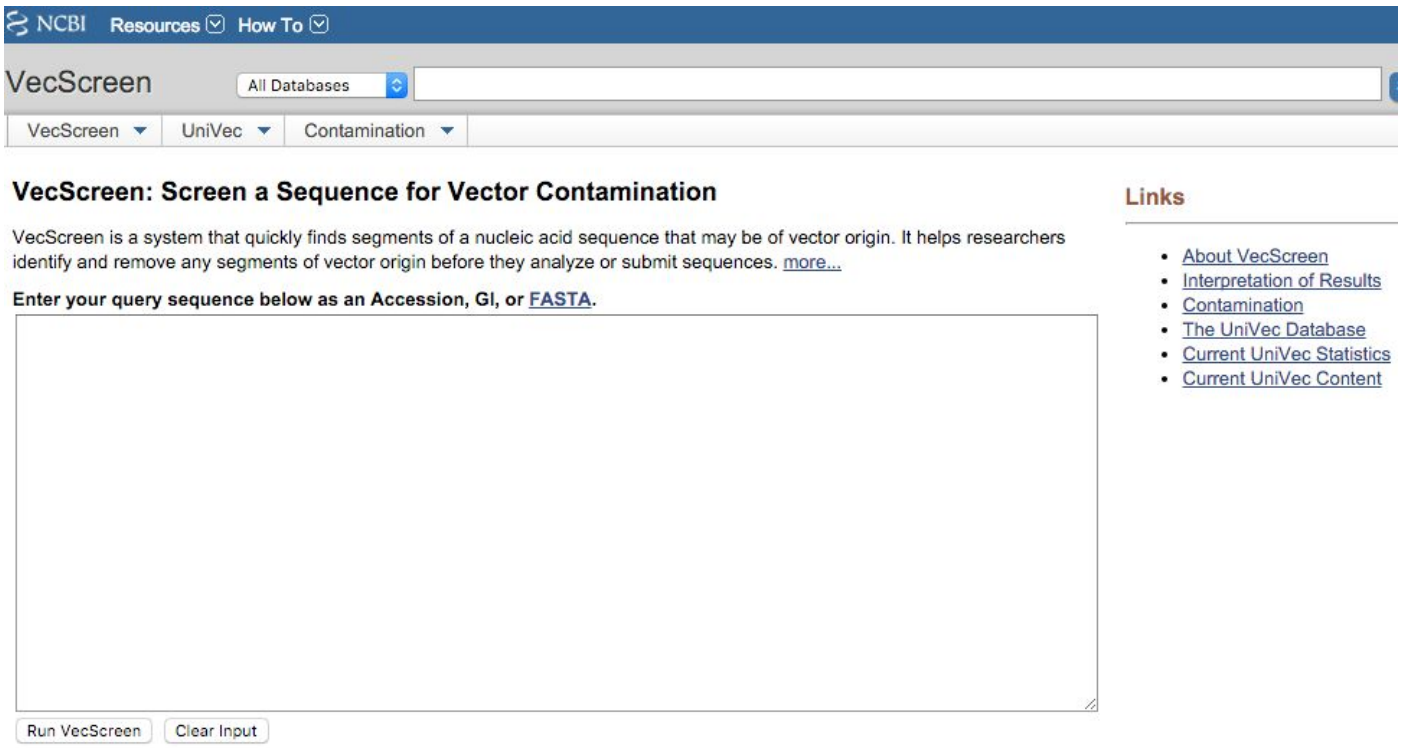


- **What can I do with the ENA?**
  - Permanently archive your sequence data and disseminate to the global research community.
  - Share your pre-publication data with collaborators in multi-centred sequencing studies.
  - Reduce your local hardware requirements for archiving next-generation sequence data.
  - Locate, retrieve and aggregate existing sequence data for analysis and meta-analysis using the EBI Toolbox and third party tools.
  - Report novel annotation relating to existing sequence as part of the Third Party Data policy.
  - Browse existing sequence and annotation referred to in the literature.
  - Find all sequence and annotation available for a gene of interest.
  - Use sequence similarity to search data (including unassembled raw data) and find out what is known about your new sequence.
  - Link through from nucleotide data to a host of integrated resources, such as genomes (Ensembl and Ensembl Genomes), the scientific literature (CiteXplore), protein products (UniProt) and protein families, motifs and domains (InterPro).
- **Sequence Read Archive (SRA)**
  - Reads of raw data consisting of short, unassembled fragments of sequence generated using Next Generation sequencing technology.

- **CRAM**
  - CRAM is a framework technology comprising file format and toolkit in which we combine highly efficient and tunable reference-based compression of sequence data with a data format that is directly available for computational use. In support of CRAM, we also provide the CRAM reference registry.
- Quick tour
- YouTube videos
  - Introduction
    - https://www.youtube.com/watch?v=ZeDB3X4G1gU
  - SRA submission 1: some theory
    - https://www.youtube.com/watch?v=x9jJyrUCeTk
  - Metagenomic data submission through EBI ENA Webin tool
    - https://www.youtube.com/watch?v=Zml8jTqfQPg
  - SRA submission 2: A walk-through of a sequence submission
    - https://www.youtube.com/watch?v=jFr11j1TJTY
- http://www.ebi.ac.uk/ena

GenBank

- VecScreen – for Vector clipping
  - separating the DNA segment of interest from the vectors DNA, when the DNA of interest is contaminated with the vectors DNA.



- http://www.ncbi.nlm.nih.gov/genbank/
- Tutorial: https://www.youtube.com/watch?v=j7hV10gYz1Q

DNA Databank of Japan (DDBJ)

- http://www.ddbj.nig.ac.jp/
- DDBJ Tutorial | How To Use DNA Data Bank Of Japan
  - https://www.youtube.com/watch?v=UF640aBB_c0

Human Genome Nomenclature Consortium (HGNC)

Standardize genes

1. Look for the various names
2. Choose one name as standard

    a. Official gene symbol
3. Make other names as synonyms

Questions

1. **Why is that possible that so many genes have more than one name?**
    a. - Gene names are not standardizing across species
    b. - gene can be present in 2 animals, with the same function
    c. - not been well established and curated gene naming system for a long time

Genome Databases

Genome databases contain the entire nucleotide sequences for an organism, and information about the assembly used.

Builds are basically the name for an entirely assembled chromosome, or genome. They have versions because they're not perfect and are always being improved with more reliable sequencing and analysis techniques.

Workflow of genome sequence submission

Plasmid with desired sequence (restriction digestion) --> a small amount of plasmid transforms the competent E.coli --> grow of bacteria on the agar plate with antibiotics --> pick the colonies in order to find the right one --> each colony grows in the liquid medium --> plasmid purification from this colonies --> plasmid DNA sequencing (Sanger) --> control of plasmids --> submission of data into ENA --> via web-based tool WebIn --> following information:

1) Sequence details
2) Released date
3) Experimental design
4) Source organism
5) Sample details
6) Lab and instrumental properties
7) Environmental conditions

File Formats

- FASTA: Annotations plus {ACTG}* with 60 characters per line
- FASTQ: extra information on top of FASTA – quality scores for each line

The Reference Sequence (RefSeq)

The Reference Sequence (RefSeq) collection provides a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. RefSeq sequences form a foundation for medical, functional, and diversity studies. They provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis (especially RefSeqGene records), expression studies, and comparative analyses.

- RefSeqGene
    - a subset of NCBI's Reference Sequence (RefSeq) project, defines genomic sequences to be used as reference standards for well-characterized genes.
    - These sequences, labeled with the keyword RefSeqGene in NCBI's nucleotide database, serve as a stable foundation for reporting mutations, for establishing conventions for numbering exons and introns, and for defining the coordinates of other variations.
- http://www.ncbi.nlm.nih.gov/refseq/

Questions

- **What was the intention to create the RefSeq database?**

- - comprehensive, integrated, non-redundant, well-annotated set of sequences (genomic DNA, transcripts, and proteins)
- - stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis (especially RefSeqGene records), expression studies, and comparative analyses.
- - provide separate and linked records for the genomic DNA, the gene transcripts, and the proteins arising from those transcripts)

EBI Genomes

within the European Nucleotide Archive (ENA)

Submission requires having annotations about samples, experiments, and runs. Like, what's in your cow? And I did a triplicate sequence on it. Nice. Includes viruses, bacteria, and eukaryotes.

- http://www.ebi.ac.uk/genomes/
- There will probably be a question about how to submit data to this database (https://www.youtube.com/watch?v=atkePLnse7A)

NCBI Genomes

- http://www.ncbi.nlm.nih.gov/genome/
- Help Page: http://www.ncbi.nlm.nih.gov/books/NBK3837/

Ensembl

Contains full genomes with alignments to entries in NCBI Reference Sequence and UniProt. Ensembl produces automated gene annotations by finding clusters of these entries.

Information such as gene sequence, splice variants and further annotation can be retrieved at the genome, gene and protein level. This includes information on protein domains, genetic variation, homology, syntenic[5] regions and regulatory elements. Coupled with analyses such as whole genome alignments and the effects of sequence variation on proteins, this powerful tool aims to describe a gene or genomic region in detail.

Ensembl imports genome sequences from consortia, which is consistent with many other bioinformatics projects. Each species in Ensembl has its own homepage, where you can find out who provided the genome sequence and which version of the genome assembly is represented. To see an example, visit the Ensembl home page for human.

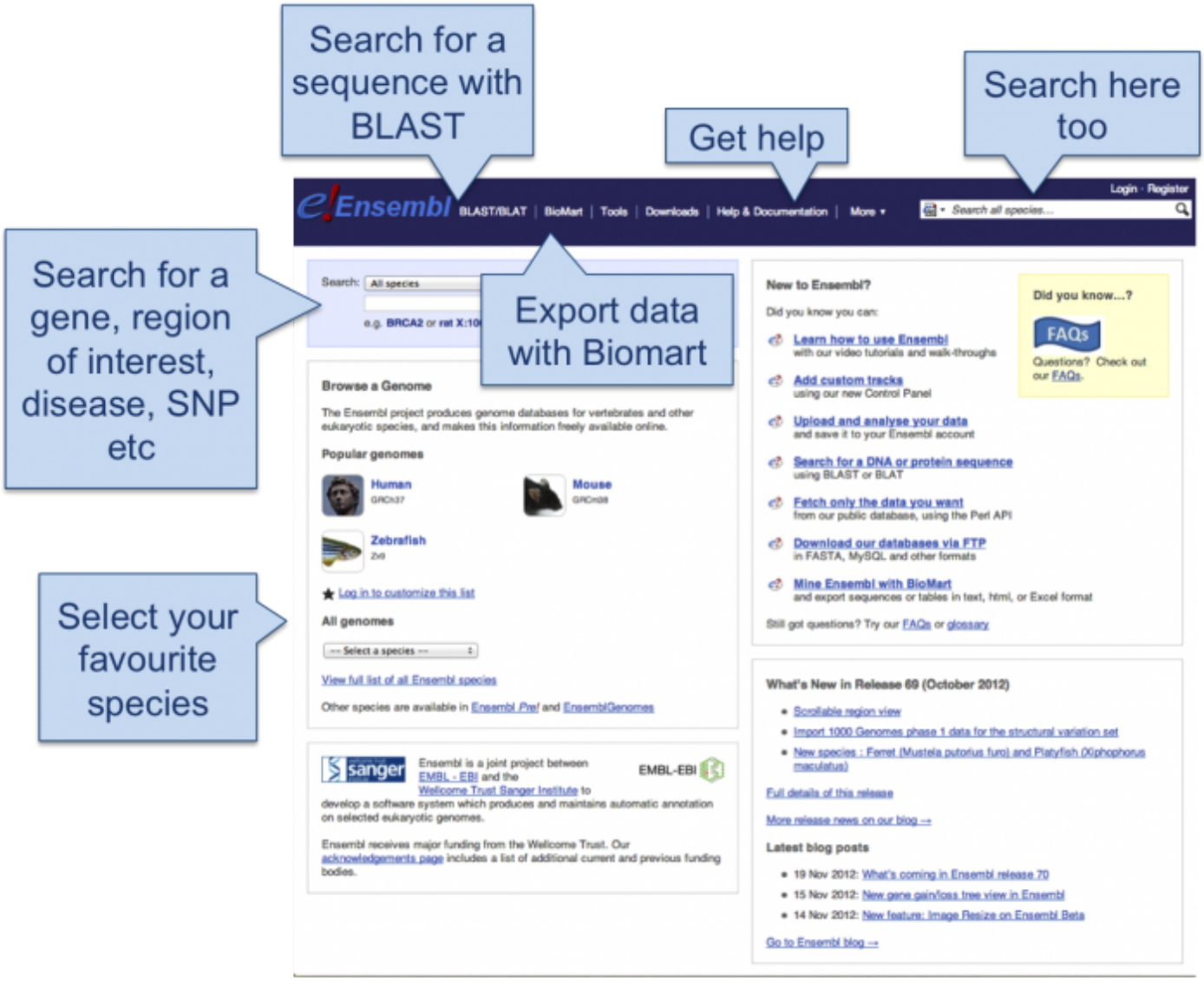
Its genome viewer also contains other powerful annotation overlays to contigs, sequence variations (SNPs) from NCBI dbSNP, comparative genomics, and functional genomics from ENCODE. One of these annotations is Human and Vertebrate Analysis aNd Annotation (HAVANA), a set of manually curated gene annotations for whole genomes. Manually curated sets are a good benchmark to include against an automated process.
The automated annotations Include annotations of introns, exons, and noncoding regions. All annotations are colored by source.

Ensembl has many features. Some of the things you can do are:

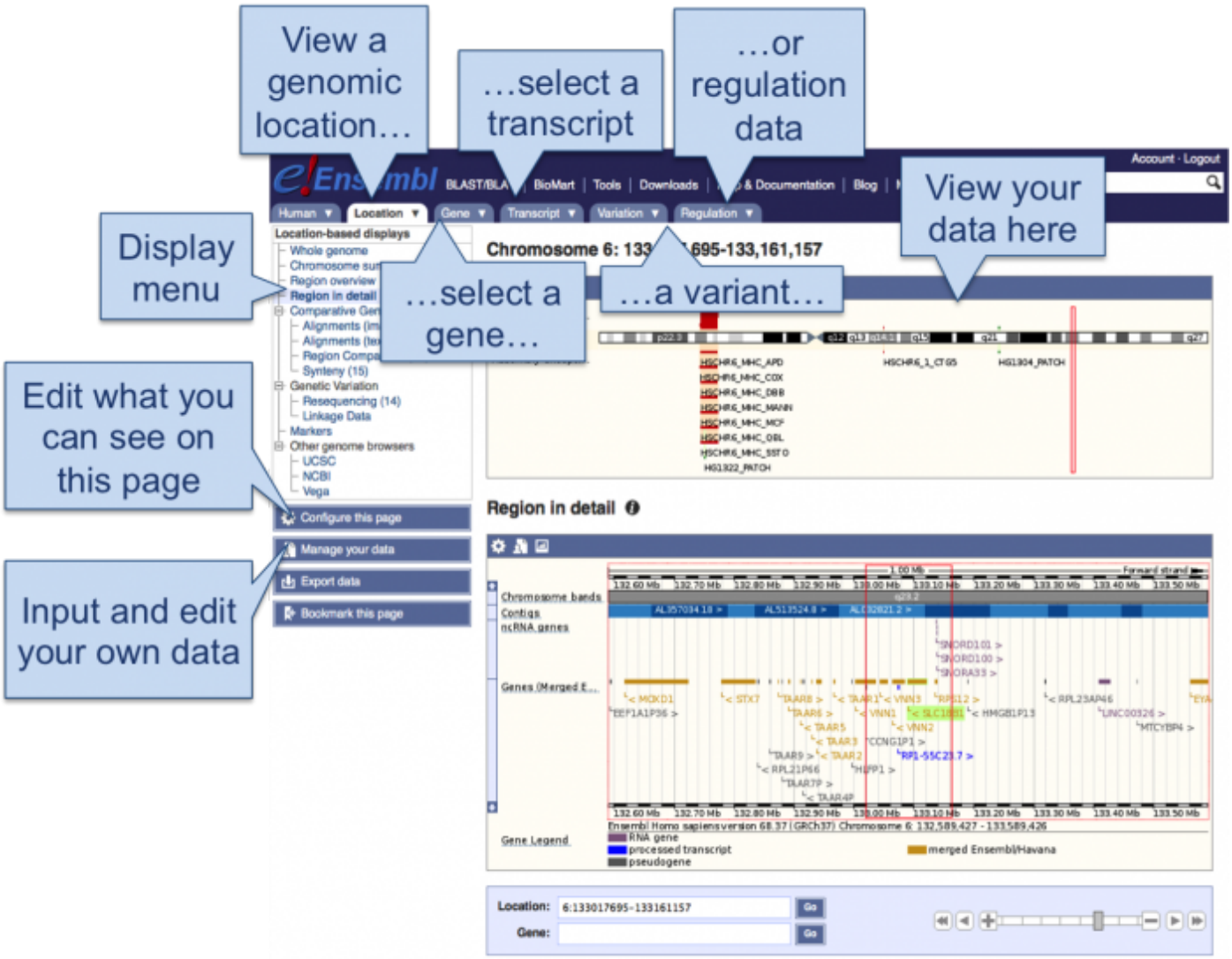
- Examine the characteristics of a region, such as genes, regulatory features or oligo probes.
- Retrieve genomic, cDNA and gene sequences.
- Align sequences against any genome in Ensembl.
- Study gene alignments between species.
- View gene transcripts and proteins.
- Export data such as sequences, tables and Single Nucleotide Polymorphism (SNP) data.
- Upload your own data to the browser.
- View sequence variation.

[5] The term synteny was originally defined to mean that two gene loci share the same chromosome. In a genomic context we refer to syntenic regions if both sequence and gene order is conserved between two (closely related) species

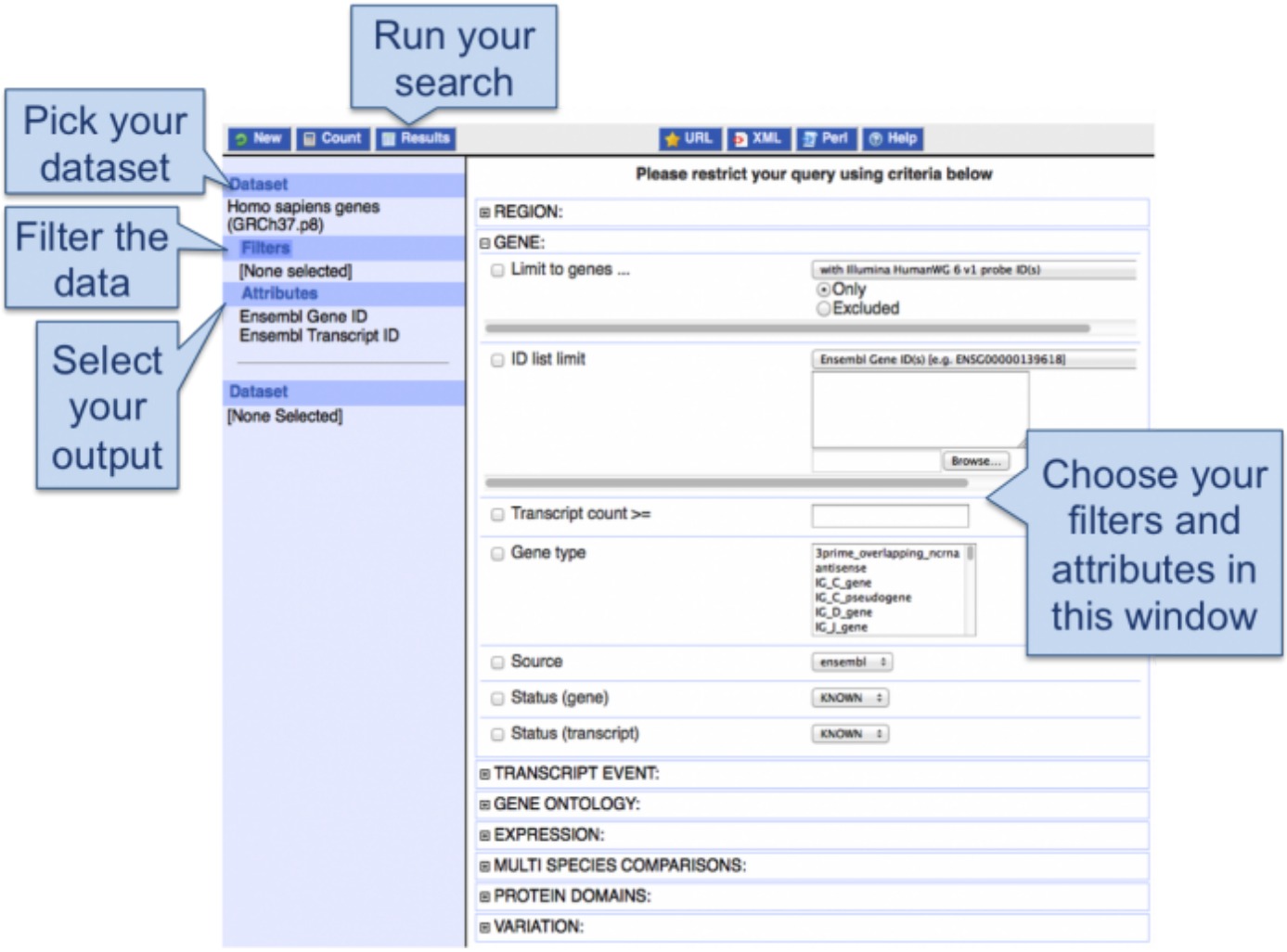
Access and navigate



Searching and visualizing data from Ensembl



Getting data from Ensembl using Biomart



YouTube video about BioMart

The UniProt BioMart allows for querying of UniProtKB data and linking it to data in other resources such as InterPro, PRIDE and Ensembl.

Links:

- http://www.ensembl.org/
- Overview: https://youtu.be/ZpnQOOxXufM
- Ensembl contains an interface called BioMart (http://www.ensembl.org/biomart/) that allows for bulk download of data

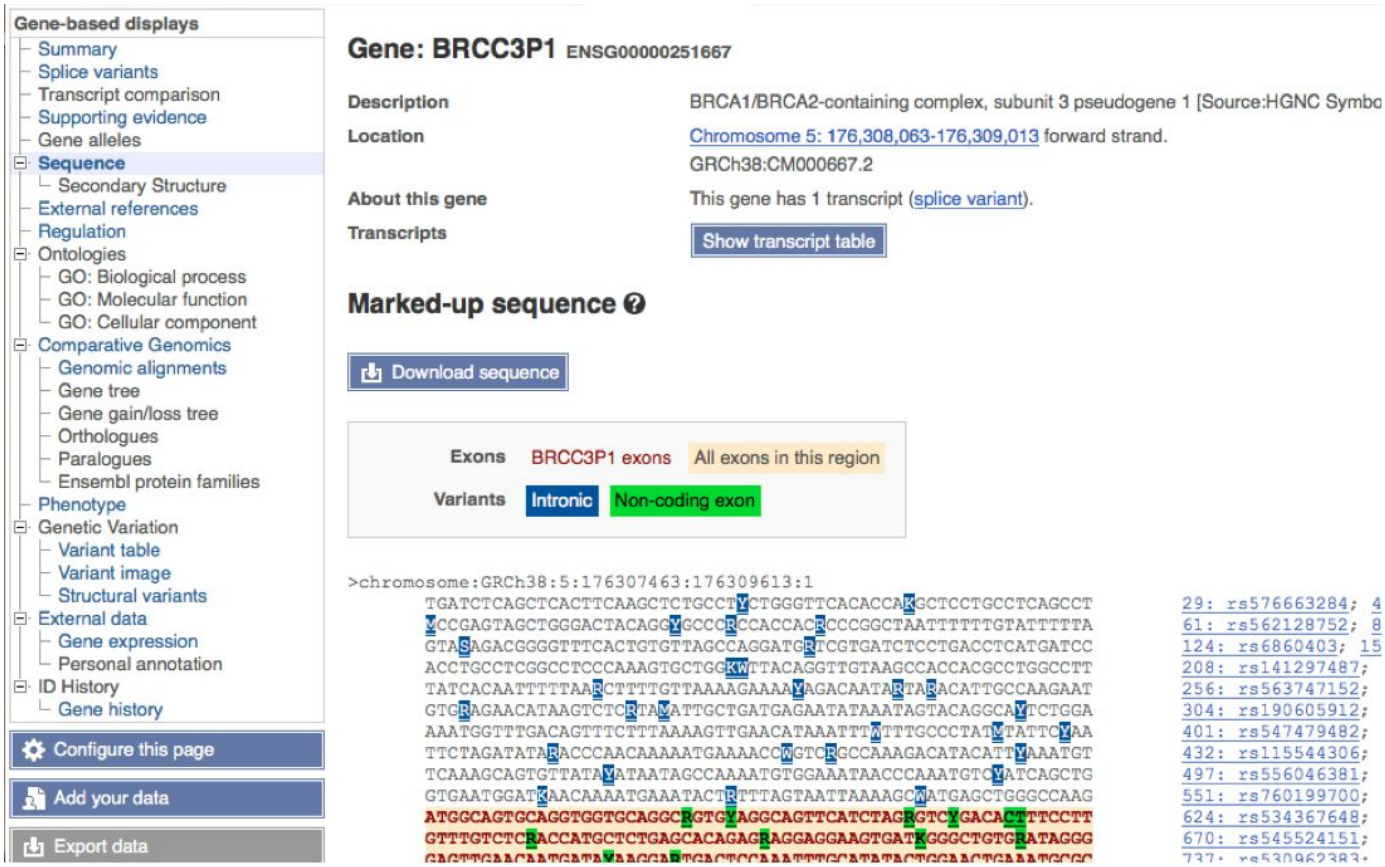
Questions

- **What differentiates EMBL from ENSEMBL?**
  -

ENA	ENSEMBL
sequence (molecule) focused database	genome (organism) focused database
experimental workflows of nucleotide sequencing	automated genome annotation and subsequent visualisation of the annotated genomes

- **How is gene polymorphism-information linked to sequence information in ENSEMBL? Where does SNP- Information come from and how is it technically linked?**
  - You click on the Sequence (you see the whole sequence), then configure this page, and turn on “show variants and links”. Then on the sequence you see directly the all possible variants and the links to it.

- This information comes from the dbSNP (NCBI) and technically is linked to the ENSEMBL via links.

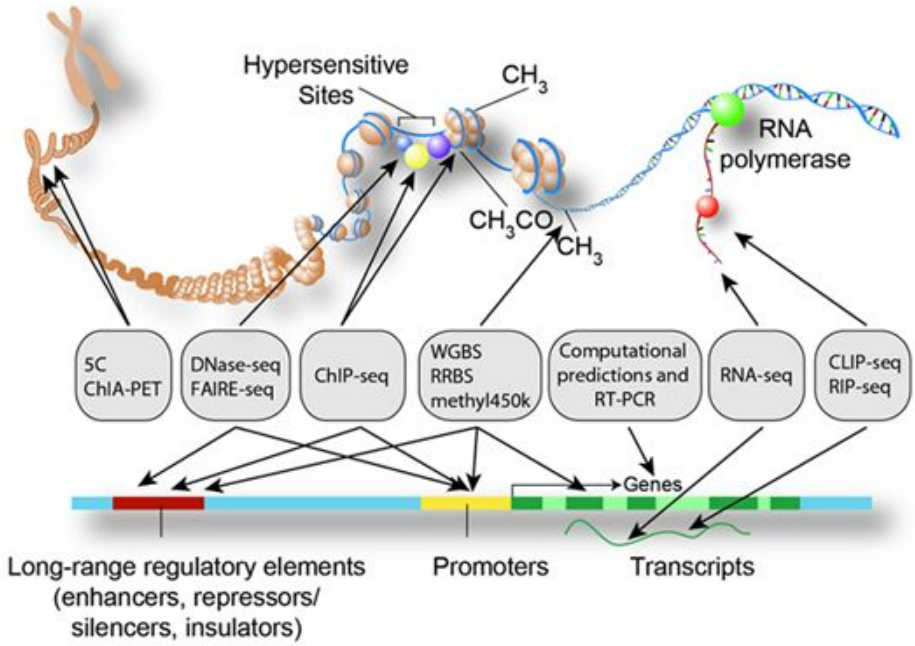


Encyclopedia of DNA Elements (ENCODE)

Encode is a a set of hella[6] experiments on regulatory data- available information about DNA Methylation, Histone Protein locations and transcription binding with chromatin immunoprecipitation (ChIP), open chromatin tests with DNAse1, methylation with H3L4Me, etc. Really impressive.

Regulatory elements are typically investigated through DNA hypersensitivity assays, assays of DNA methylation, and immunoprecipitation (IP) of proteins that interact with DNA and RNA, i.e., modified histones, transcription factors, chromatin regulators, and RNA-binding proteins, followed by sequencing.

[6] extremely (used for emphasis)



- https://www.encodeproject.org/

Other Organism Specific Genome Databases

- **Saccharomyces Genome Database (SGD)** http://www.yeastgenome.org/
  - a comprehensive integrated biological information for the budding yeast *S. cerevisiae* along with search and analysis tools to explore these data, enabling the discovery of functional relationships between sequence and gene products in fungi and higher organisms.
- **Rat Gene Database** http://rgd.mcw.edu/
  - the premier site for genetic, genomic, phenotypic, and disease data generated from rat research. In addition, it provides easy access to corresponding human and mouse data for cross-species comparisons.
- **Zebrafish Model Organism Database** http://zfin.org
- **Plant Genome Database (PlantGDB)** http://plantgdb.org/
  - PlantGDB provides sequence data for >70,000 plant species, custom EST assemblies (PUT) for over 150 species, web tools and plant genome browsers, as well as an outreach portal for plant genomics

Questions

1. What the difference between gene and genome database?

Gene Databases

Hugo Gene Nomenclature Committee (HGNC)

HGNC provides unique and stable identifiers for human loci, which includes protein coding genes, ncRNA genes, pseudogenes, and other stuff.

Links:

- http://www.genenames.org/

Entrez Gene

Pronounced ahn-tray. This is the NCBI's internal gene accession number. It's not so stable, or sensical[7], for that matter.

Links:

- http://www.ncbi.nlm.nih.gov/gene/

The ENTREZ documentation mentions "E-utilities". Please explain what E- utilities are and what they can be used for?

E-utilities = Entrez Programming Utilities

- - tools that provide access to Entrez data outside of the regular web query interface in all 38 bio databases of NCBI
- - use a fixed URL syntax (translates a standard set of input parameters into the values necessary for various NCBI software components to search for)
- - helpful for retrieving search results in another environment.
- 8 server-side programs

SNP Databases

Refresher on the difference between SNP versus Mutation:

- There needs to be a certain threshold frequency of a change in the population for a mutation to be considered a SNP,
- while a mutation is any change from one base to another. Mutations have a strong biological impact and are easier to speculate on than SNPs. SNPs are often found in the intergenic regions.
- A haplotype is a genetic pattern that is inherited together.

In the future, whole genome sequences will help to identify rare SNPs, ones that don't even make the cut[8] of normal SNP chips.

Laboratory Methods and Background

Genome-wide Association Study (GWAS)

A Genome-wide Association Study (GWAS) tries to build a statistical correlation between the occurrence of a certain SNP and a phenotype. It uses very simple statistics, like the Fisher's Exact Test and Bonferroni Correction for Multiple-Hypothesis Testing. This means that since many different conditions/phenotypes are being analyzed by scientists, we need to set the bar for significance very high so we don't get results by accident.

Interestingly, GWAS studies often find disease associated to intergenic SNPs.

Knockdown and Knockout Experiments

GWAS experiments are often followed up by a knockdown experiment in mice to help get a mechanistic understand of how a SNP contributes to the measured phenotype. This is pretty simple, and commonly done with RNAi, shRNA, and now CRISPR-Cas9. Further characterization is usually needed.

Expression quantitative trait loci (eQTL)

eQTLs are genomic loci that contribute to variation in expression levels of mRNAs.

While a GWAS can associate a phenotype to certain SNPs, eQTL studies attempt to link SNPs to gene expression (in a specific cell type).

https://en.wikipedia.org/wiki/Expression_quantitative_trait_loci

[7] t o make sense, to be understood

[8] Make the cut – To meet or come up to a required standard

Linkage Disequilibrium

In one of the links to "relevant background information", a primer on molecular biology mentions "linkage disequilibrium". What does this term mean?

"Linkage Disequilibrium" - the presence of the statistical = non random associations between alleles at different loci. Linkage disequilibrium looks at 2 SNPs together and asks how well they correlate.

In the HapMap, cliques[9] of SNPs who all have a correlation of > 0.8 are called LD-bins. These bins reflect areas with little or no recombination (if the SNPs occur close in physical location, though that's often not the case). Since these bins are experimentally determined though, we get results of physically distant SNPs correlating highly.

For a real experiment, it's hard to measure all of the SNPs, so a couple representative SNPs are picked from each LD-bin, called Tag-SNPs.

- Factors of linkage disequilibrium:
  - Selection
  - rate of recombination
  - Mutation
  - genetic drift.
- - it is the non-random association of alleles at 2 or more loci, not necessarily on the same chromosome.
- International HapMap Project aims to develop a haplotype map of the human
- genome, which will describe the common patterns of human genetic variation

NCBI dbSNP

dbSNP and Ensembl are both genetic variation databases. In addition to its genome data, Ensembl also contains information about structural genetic variation, such as copy number, inversions, and translocations. Like most NCBI websites, dbSNP looks terrible. This website lets you look up information about a SNP based on its RS number, like genomic position, chromosome number, related genes, etc.

Links:

- http://www.ncbi.nlm.nih.gov/SNP/

Genome.gov and GWAS Catalog

NCBI's Genome.gov and EBI's each hold catalogs for published GWASs.

Links:

- http://www.genome.gov/
- http://ebi.ac.uk/gwas

Regulatory Elements Database

Regulatory regions can be very far away from a gene, even 1 megabase. One element can even regulate multiple genes.

The Regulatory Elements Database is part of the ENCODE project. It's useful to start with a given genomic coordinate and find regulatory regions, or start with a gene and get regulatory elements.

Unmet need: what cell type is your gene mutation affecting most, causing a disease?

Links:

- http://dnase.genome.duke.edu/

[9] Clique – a small close-knit group of people who do not readily allow others to join them

RegulomeDB

RegulomeDB contains information about what regulatory elements in the intergenic space that a given SNP has been measured/predicted to interact with/be a part of. These regions shot DNAse hypersensitivity, are the binding sites of transcription factors, and/or are promoter regions.

Links:

- http://regulomedb.org/

HapMap

There are 8 million common SNPs that are matched with obvious traits (eyes, hair, ethnicity, etc.). Rare allele frequency is <10%, and if you have a <1% occurrence, it's not really an allele anymore.

The goal of HapMap is to identify the different haplotypes in different ancestries (African, European, Asian) and produce tags to go with each haplotype (see LD-Bins above)

Links:

- http://hapmap.ncbi.nlm.nih.gov/
- Finding Genes for Human Disease by Lynn Marquis
  - https://www.youtube.com/watch?v=w1IJO_0t7Jg

Functional SNP Database

This database/service relates data out of lots of other sources including Ensembl, dbSNP, HapMap, RegulomeDb to predict the functional effect of SNPs on transcription, translation, PTMs, etc.

Links:

- http://compbio.cs.queensu.ca/F-SNP/

Disease Specific Databases

Disease specific databases curate data, and add annotations from the data's sources that aren't included in the general databases. Useful tool is the MISO Sequence Ontology Browser.

- Catalog of Somatic Mutations in Cancer (COSMIC) [http://cancer.sanger.ac.uk/cosmic]

Questions

1. **Sketch the high level conceptual design of a database representing associations between gene polymorphisms and a disease phenotype!**

Proteins databases

1. Subcategories from issue of the journal NAR (Nucleic Acid Research) the category "protein sequence databases"
   a. General sequence databases (AA sequence)
   b. - Protein properties
   c. - Protein localization and targeting
   d. - Protein sequence motifs and active sites
   e. - Protein domain databases; protein classification
   f. - Databases of individual protein families

Protein structure databases

RCSB-Protein Data Bank (PDB)

The PDB contains 3D X-ray crystal structures and some NMR structures. It has its own implementation of the BLAST algorithm for sequence alignments, and it also has 3D similarity searches as well.

Links:

- http://www.rcsb.org/

Macromolecular Structure Database Group Overview

The MSD group is principally involved with the data associated with macromolecular structure associated with the metabolism of living organisms, that is, the atomic coordinates of proteins, nucleic acids and molecules that bind with these. We also maintain links to protein sequence information, textual information from scientific publications and a number of derived properties that augment the macromolecular structure information. The macromolecular coordinate data is collectively known as the "protein structure databank" (PDB) although we provide far more information by providing a searchable database of this and links to other information.

- http://www.ebi.ac.uk/msd/about/overview.html

Protein sequence databases

UniProt Knowledgebase

The UniProt Knowledgebase is the central resource for data on proteins. It contains:

- GO Annotations for function and biological processes
- Annotations to pathology

It also links to information about:

- Protein-protein interactions (Reactome)
- Inhibitors (ChEMBL, DrugBank, BindingDB)
- 3D Structure (PDB)
- Families and Domains (Pfam, InterPro, etc.)
- Amino Acid Sequence (UniParc)
- Genetic Variants (dbSNP)
- Tons of cross-references

Each annotation in UniProtKB also contains a link to its evidence, whether it be a database, manual curation, assertion by similarity, a publication, or automatic curation. UniProtKb consists of TrEMBL, which automatically annotates proteins, and SwissProt, which consists of manually curated annotations. The Protein Information Resource also is part of the UniProt Consortium.

TrEMBL

Trembl will look at cDNA sequences in both the 3'->5' and 5'->3' direction and look for start codons in all 3 possible frames. It translates in silico, and usually keep the longest as a putative protein.

Links:

- http://www.uniprot.org/uniprot
- http://pir.georgetown.edu/

InterPro

1. http://www.ebi.ac.uk/interpro/

UniParc and UniProt Reference Clusters (UniRef)

UniParc contains protein sequences. They are imported from different databases and have their own stable accession numbers.

UniRef is a collection of three neural networks (look up “radial basis function network”!!!) that cluster protein sequences from UniParc by three different conditions. Users can query it with a seed sequence to assign it to a cluster. Each entry in UniRef represents one cluster.

1. UniRef100 clusters identical sequences over 11 residues from any organism into a single UniRef entry
2. UniRef90 analyzes the longest sequence from each cluster in UniRef100, and builds new clusters of groups of sequences that all have at least 90% identity with every other element of the cluster. Each cluster is given an entry. Each element in the cluster must have at least 80% overlap with a seed sequence to be given as a result.
3. UniRef50 does the same as UniRef90 by taking the longest sequence from each cluster in UniRef90 and making new clusters sharing 50% identity. Each element in the cluster must have at least 80% overlap with a seed sequence to be given as a result

   Links:

- http://www.uniprot.org/uniparc/
- http://www.uniprot.org/uniref/

Questions

1. **Which parts of UniProt can be distinguished and what is the purpose of the partitioning of UniProt?**

UniProtKB (SwissProt + TrEMBL)	UniRef (100/90/50)	UniPArc
DB of all knowledge about proteins	combine closely related sequences into a single record – UniRef100 Then they can be combined with 90% and 50%	comprehensive, non-redundant repository
names, sequence, taxonomic and bibliographic data, annotations. Functional info, posttranslational modifications, diseases, structural info	Clustering makes the searching quicklier	all possible protein sequences in the database

   a. Purpose of partitioning
      i. - To clearly distinguish between the curated knowledge in the KB and the archive. Each part has its own purpose.
      ii. - TrEMBL 1996 in response to the increased data flow from genome projects
      iii. - to support biological research by maintaining a high quality database that serves a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces freely accessible to the scientific community.

2. **What are the original root databases of UniProt KB?**

SwissProt (550 000)	**TrEMBL (59 000 000)**	**PIR**

manually annotated and reviewed	automatically annotated and not reviewed	PIR-PSD – protein sequence database
non-redundant	?	
brings together experimental results, computer features and scientific conclusions	high quality computationally analyzed records that are enriched with automatic annotation and classification	database of protein sequences □and curated families
more reliable data		
	was created because of the huge amount of generated data	

3. **Why is UniProtKB called a "curated" database?**
    a. - high level of annotation
    b. - checked and curated by more experts
    c. - minimal redundancy
    d. - great deal of human effort
    e. - high level of information: functions of a protein, its domains structure, post-translational modifications, variants, etc
    f. curator even extract information from the literature and perform computational analysis
4. And other questions:
    a. 6.7 – 6.53
        i. Pp. 69 – 76 BioDB-2016

Structural Classification of Proteins (SCOP)

SCOP provides a hierarchical classification for proteins with PDB entries based on their 2D and 3D structural features.

The tree is divided in the following order:

1. Class
2. Fold
3. Superfamily
4. Family
5. Protein
6. Species

Links:

- http://scop.mrc-lmb.cam.ac.uk/scop/
- http://supfam.org/SUPERFAMILY/

Other protein databases

Pfam

The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).

- http://pfam.xfam.org/

IntAct Molecular Interaction Database

- http://www.ebi.ac.uk/intact/

Molecules databases?

ChEBI

Protein-Protein Interactions (PPIs)

PPIs are the final frontier of drug discovery. It's still incredibly novel to find small molecules that can disrupt the binding between two proteins. Maybe allostery can muck it up, but it would be really exciting to find molecules that get in the way of these interactions. On a biological level, these databases help make assertions about which proteins are doing what, and later help give evidence when building pathways.

Laboratory Techniques

- Forster Resonance Electron Transfer (FRET)
- Co-immunoprecipitation (Co-IP)
- Yeast 2 Hybrid
- Synthetic Gene Array

BioGrid

BioGrid is a curated database of protein-protein and protein-genetic interactions. This was the one presented in class. Does network analysis of connectivity coefficient of connected neighbors to a given gene.

Links:

- http://thebiogrid.org/

Other PPI Databases

- Biological Interaction Network Database (BIND) [https://www.bindingdb.org/]
- Database of Interacting Proteins (DIP) [http://dip.doe-mbi.ucla.edu/]
- IntAct [http://www.ebi.ac.uk/intact]
- Molecular Interactions Database (MINT) [http://mint.bio.uniroma2.it/mint]
- Human Protein Reference Dataset (HPRD) [http://www.hprd.org/]
    - PPI
    - Subcellular localization
    - Post-translational modification
    - Enzyme substrate relationships
    - Disease associations

Agile Protein Interaction DataAnalyzer (APID)

APID provides access to experimentally validated PPIs in BIND, BioGRID, DIP, HPRD, IntAct, and MINT. Allows for exploration of the "interactome" network. It's notable because it aggregates the results of all of these databases, which aren't synchronized. This was the one presented in class.

Links:

- http://bioinfow.dep.usal.es/apid/

Other PPI Data Aggregation Services

- Microbial Protein Interaction Database (MPIDB) [http://jcvi.org/mpidb/]
- Protein Interaction Network Analysis (PINA) [http://cbg.garvan.unsw.edu.au/pina/]

PPI Predictions

These are more web services than databases. They use information from PPI databases and machine learning to make predictions about binding surfaces and interactions. Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) uses network analysis, PIP uses naïve bayes (lol), and MiMI is discontinued.

- STRING [http://string-db.org/]
- PIP [http://www.compbio.dundee.ac.uk/www-pips/]
- MiMI [http://mimi.ncibi.org/]

Questions

1. Interaction DB Qq. 7.1 - 7.25
    a. Pp. 77 – 92

Pathway Databases

Pathway Databases store informations about biochemical, signaling, metabolic, and other pathways. Though pathways are all part of larger and more complex networks, these databases allow for reasoning over smaller units.

Kyoto Encyclopedia of Genes and Genomics (KEGG)

KEGG is a huge knowledge base, consisting of a database of pathways, enzymes, transformations/biochemical reactions, and tons of other stuff. It's like the Japanese EBI. It has lots of cross-links to other databases (within KEGG, albeit) but they provide lots of context that other databases can't. Go Japan. And GIF image maps. Hell yeah.

- KEGG Enzymes contains data about enzymes
- KEGG Reactions contains all data about reactions
- KEGG RPAIR contains compound pairs from enzymatic reactions and their enzyme

Broad Definitions of Reactions contains:

- Transportation
- Classical Biochemical Transformation
- Binding
- Dissociation
- Degradation
- Phosphorylating
- Dephosphorylation

Links:

- http://www.genome.jp/kegg/

WikiPathways

WikiPathways is built on the community editing, open sourced software behind MediaWiki. It has a custom editor that allows for community curation of pathways. They're stored in GPML, an extension of XML compatible with visualization software such as Cytoscape. WikiPathways also contains data from KEGG

Links:

- http://www.wikipathways.org/

Reactome

Reactome is an open-source, curated pathway database, with lots of cross-references to other databases. It has an incredibly powerful browser as well.

- http://www.reactome.org/
- http://www.reactome.org/PathwayBrowser/

Human Metabolomics Database (HMDB)

The Human Metabolome Database (HMDB) is a freely available electronic database containing detailed information about small molecule metabolites found in the human body.

- http://www.hmdb.ca/

Questions

2. Pathway DB; Qq. 8
    a. Pp. 93 – end

Bibliographic databases

– is a database of bibliographic records, an organized digital collection of references to published literature, including journal and newspaper articles, conference proceedings, reports, government and legal publications, patents, books, etc.

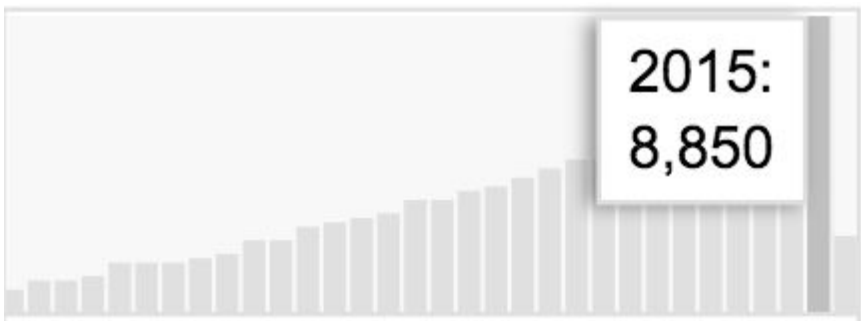
PubMed

PubMed comprises over 25 million citations for biomedical literature from MEDLINE, life science journals, and online books.
PubMed provides access to bibliographic information that includes MEDLINE, as well as:

- The out-of-scope citations (e.g., articles on plate tectonics or astrophysics) from certain MEDLINE journals, primarily general science and chemistry journals, for which the life sciences articles are indexed for MEDLINE.
- DIsplay abstracts only
    - PubMed Central specialises in full text.
- Citations that precede the date that a journal was selected for MEDLINE indexing.
- Some additional life science journals that submit full text to PubMed Central and receive a qualitative review by NLM.
- Citations for the NCBI Bookshelf collection.

A. Advanced search filter
    a. The Advanced search builder Show index list provides an alphabetical display of all terms in each PubMed search field.
        i. E.g. name; date; author; ISBN; location ID; MeSH
        ii. PMID
            1. unique identifier for each entry
            2. To exactly find paper and to get directly an abstract from paper
    b. With Boolean operators
    c.
B. PubMed Single Citation Matcher – to find a specific citation
C. Sorting
    a. By relevance
        i. Keywords appear more often
    b. By date
        i. Epub
        ii. Paper version
    c. But not relevance and date! → data mining technologies!

D. Filtering
    a. You can narrow your search results by article types, text availability,publication dates, species (biological model), languages, sex, subjects, journal categories, ages and search fields.
E. Send to citation manager
F. One can save in different formats
    a. Summary

    b. Abstract
        i. Used for data mining
        ii. ProMiner
G. Typical query
    a. "alzheimer disease"[MeSH Terms] OR ("alzheimer"[All Fields] AND "disease"[All Fields]) OR "alzheimer disease"[All Fields] OR "alzheimer"[All Fields]
H. Hofmann loves to ask questions about the publication on certain subject vs. time histogram. Thus, I put the plot for AD publications in PubMed:
I. 

    a. How relevant is the topic → easy to spot trends (~protein of the year)
    b. Measurement of investment in the field
J. http://www.ncbi.nlm.nih.gov/pubmed

Medical Subject Headings (MeSH)

The NLM Medical Subject Headings controlled vocabulary of biomedical terms that is used to describe the subject of each journal article in MEDLINE. MeSH contains approximately 26 thousand terms and is updated annually to reflect changes in medicine and medical terminology. MeSH terms are arranged hierarchically by subject categories with more specific terms arranged beneath broader terms. PubMed allows you to view this hierarchy and select terms for searching in the MeSH Database.

MeSH developers: not an ontology, but a thesaurus! (hierarchical, but with multiple inheritance). It is only related to an ontology. As papers are published, they get the appropriate MeSH terms annotated to them. This makes MeSH an invaluable tool for literature search on repositories like Medline and PubMed. The synonyms dictionary is also useful for text mining.

- MeSH – organized keywords, used for unification of concepts (=abstract idea)
- Classify documents according to bigger topics
    - So you will not get false positive results: you search for bacterial genes, but something similar is discussed in context of PTSD
- Single keyword does not represent the context
    - Search dementia – only 1% of relevant documents.
    - Google has one word index –> most people use just one word
- Relatedness of documents can exist On the level of
    - authors
    - Across different species or within the species
    - Different viewpoints
    - In GO (three fundamental roots)
        - Biological process

        - Molecular function
        - Cellular localization
- Contains synonyms
    - I.e. keywords represents a unification of concepts, that means the same entity, but different terms

Links:

- http://www.ncbi.nlm.nih.gov/mesh
- https://www.nlm.nih.gov/mesh/MBrowser.html

Information retrieval

Information retrieval (IR) is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, as well as that of searching structured storage, relational databases.

MeSH allows us to search the same thing but calling it with different terminologies. The more MeSH terms are there in an abstract, the higher is the relevance of the term to the entry in Medline.

Brief overview of journal regulations

Typical article (in Nature) contain:

1. Summary
    a. For readers outside the discipline/field
        i. Without references, numbers, abbreviations, acronyms
            1. Acronym – UNESCO, GO – pronounced as a whole
            2. Initialism – BBC, DNA – separately every letter
2. Content
    a. Introduction
        i. importance
    b. Methods
    c. methods
    d. Main conclusion = outcomes
    e. Discussions
    f. Main findings in the general context – results, which put the field forwards
    g. References

Medline

- The largest component of PubMed
- All abstracts of MEDLINE are indexed with MeSH terms.

MEDLINE is the bibliographic database that contains references to journal articles (medical literature, books, publications) in the Life Science (LS) that PubMed uses as a service to link to those documents. This is the first stop in all searches for informations about biological stuff. Faceted search[10] can be performed and MeSH terms can be used to make even more powerful queries.

MEDLINE records are indexed with NLM Medical Subject Headings (MeSH). The database contains citations from the late 1940s to the present, with some older material. New citations that have been indexed with MeSH terms, publication types, GenBank accession numbers, and other indexing data are available daily and display with the tag [PubMed - indexed for MEDLINE].

Links:

- https://health.ebsco.com/products/medline

[10] also called faceted navigation or faceted browsing, is a technique for accessing information organized according to a faceted classification system, allowing users to explore a collection of information by applying multiple filters. A faceted classification system classifies each information element along multiple explicit dimensions, called facets, enabling the classifications to be accessed and ordered in multiple ways rather than in a single, pre-determined,taxonomic order.[1]

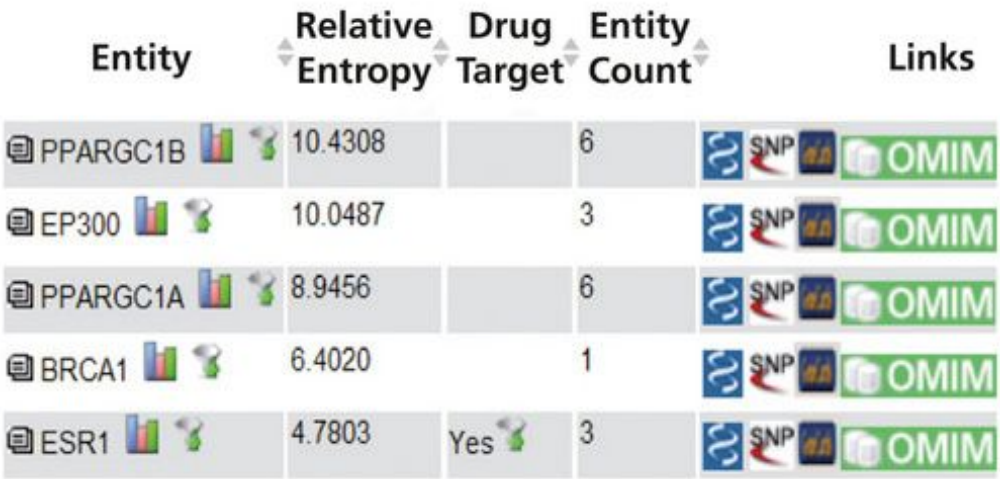
PubMed Central

PubMed Central (PMC) is the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life sciences journal literature.

ProMiner

Scientific publications found in abstract databases, full text journals or patents are the main and most up-to-date information source, but the amount of text is overwhelming for most life science areas. Recognition of life science terminology is a key prerequisite for performing automatic information retrieval and information extraction. Huge and complex terminologies with high numbers of synonymous expressions, ambiguous terminology and numerous generations of new names and classes present named entity recognition with a real challenge. ProMiner is a tool for specific terminology recognition and addresses several fundamental issues in named entity recognition in the field of life sciences:

- ProMiner can handle voluminous dictionaries, complex thesauri and large controlled vocabularies derived from ontologies
- regularly updated dictionaries through automatic curation followed by a manual evaluation process
- mapping of synonyms to reference names and data sources
- context dependent disambiguation of biomedical termini and resolution of
- acronyms
- specific handling of common English word synonyms
- spelling variants of expressions in the source dictionary can be recognized
- high speed tagging and parallel workflow for multiple dictionaries
- incorporation of regular expressions (e.g. for the recognition of SNP rs numbers)
- full text annotation in XML, HTML or PDF format
- patent annotation



SCAIView

is a knowledge discovery software for the life sciences. It facilitates the rapid identification of aggregated information from large text sources.

The information retrieval system SCAIView allows for semantic searches in large text collections by combining free text searches with the ontological representations of entities derived by ProMiner. SCAIView gives answers to questions such as “Which genes / proteins are related to a certain disease, pathway or epigenetics?”

SCAIView´s key features are:

- A user-friendly search environment with a query builder supporting semantic queries with biomedical entities

- Fast and accurate search and retrievals, based on the newest technologies of semantic search engines
- Visualization and ranking of the most relevant entities and documents
- Exportation of the search results in various file formats

Documents are retrieved by precisely formulated questions using ontological representations of biomedical entities. The entities are embedded in searchable hierarchies and span from genes, proteins, accompanied single-nucleotide polymorphisms to chemical compounds and medical terminologies. SCAIView supports the selection of the suitable entities by an autocompletion functionality and a knowledge base for each entity. This includes a description of the entity, structural information, pathways and links to relevant biomedical databases like Entrez Gene, dbSNP, KEGG, GO and DrugBank. SCAIView represents the search results using a color-coded highlighting of the different entity-classes, statistical search results and various ranking functions.

Online Mendelian Inheritance in Man (OMIM)

An Online Catalog of Human Genes and Genetic Disorders. Records that include the current articles as reference cited at the end of the OMIM record.

1. System, which is able to find indirect results: SNPs related to AD
2. Comprehensive / authority compendium of human genotype <–(focused on relationship)–> genetic phenotypes + cross-links
    a. Compendium –a collection of concise but detailed information about a particular subject, especially in a book or other publication
3. Object – all known mendelian disorders
4. **What is the content of OMIM? Sketch a simple entity model of OMIM comprising the major entity types represented in OMIM**
    a. Phenotype-gene relationships
    b. - Description of disease
    c. - Clinical features
    d. - Biochemical features
    e. - Pathogenesis
    f. - Inheritance
    g. - Animal model

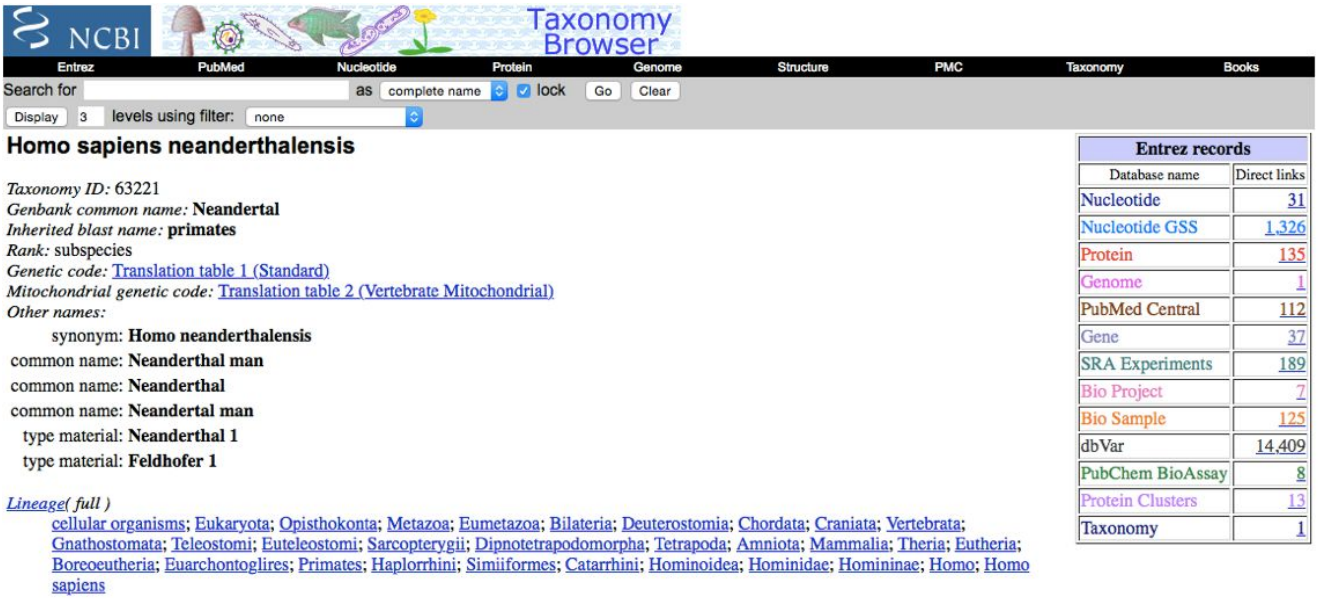
- Links:
    a. http://www.omim.org/

OMIM vs. Medline

OMIM	MEDLINE
reviews genes and their phenotypic appearance, diseases, and explains molecular etiology of the disease.	bibliographic database covering medicine, contains abstracts and citation over full articles.
literature-based	literature-based
Catalog of genes and diseases	premier bibliographic database

Taxonomy database

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.



- Nomenclature of all organisms, connections of roots, classes, families. Systematically.
- Entrez records: Nucleotide, Nucleotide GSS, Protein, Genome, PubMed Central, Gene, Sra, PubChem, Bio Sample, Bio Project.
  http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root

Ontologies

Definition

1) Def
2) Erfan Younesi's lecture
   a) representation of knowledge domain, which organizes and standardizes the knowledge specific to that domain (~ Alzheimer disease; genes; proteins etc.).
      i) You try to catch all the knowledge about this domain and organize it like a tree, or any other type of classification system.
   b) Is used in databases – before the data comes to make it standard
      i) You match it again ontology and standardize it
      ii) And only then you store it in database
   c) Interoperability
3) Martin's lecture
   a) Classification schema (put into classes or concept), organize information and its formalization (formal knowledge representation). A way to bring some order, which comes in the classes schema
   b) Why? Because it can be used by
      i) humans
      ii) Computers
4) Why is it useful? – Using GO example
   a) Questions:
      i) GO was designed to make the repertoire of gene
      ii) Is a liver cell in us slight/ little modification of yeast cell?
      iii) Describe the molecular function that protein has and cell localization – then you need to have annotation of proteins and genes; you need to use the mouse or gene-modified mouse; knockout and knockin mouse; you need to be able to describe the mouse;
      iv) Tumor & Annotation of genes in GO
         (1) Helps to observing a certain pattern of genes, that upregulated in tumor cells, when comparing totally different cell types

(2) gene ontology annotation of those genes - functional terms;
(3) by checking patterns you observe in primary data, the genes that you find in pattern - GO annotation, statistical method - treat annotation the same way

v) Helps to answer the questions about Gene X, like
(1) Part of a pathway?
(2) A phenotype?
(3) Associated with a disease?
(4) What does it do?

vi) Microarray experiment: you have a list of genes
(1) tons of literature to skim
(2) No experimentation done

b) Finding information is no time-consuming
i) Curators are already doing that work

c) Annotations capture information about gene function, subcellular localization and process

Gene Ontology (GO)

GO consists of annotations for proteins functions, their involvement in processes, and their cellular location. These annotations have been widely adopted across platforms describing proteins and gene products in many organisms. Each ontology is hierarchical to allow for powerful searches of variable granularity. You'll notice GO Terms in UniProtKB, Entrez, ...

Three components of GO:

1. GO Process
   a. Biological Process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms
2. GO Component
   a. Cellular Component (in class called cellular localisation), the parts of a cell or its extracellular environment;
3. GO Function
   a. Molecular Function, the elemental activities of a gene product at the molecular level, such as binding or catalysis;

Features from lectures:

1) Describes genes using **standardized controlled vocabulary**
   a) GO consortium – they have the authority to control the terminology to annotate gene and genomes
2) Major database players: we now bring order into the term space, semantic space
   a) Semantics – the science of the meaning of things; Need to be controlled and shared
3) one can identify cardinalities and communicate with other using GO
4) Additional data to the GO term
   a) An evidence code denoting the type of evidence upon which the annotation is based (IMP, IDA); Evidence codes of GO term
      i) Experimental; Inferred from Direct Assay (IDA) or Inferred from Experiment (EXP)
         (1) Using mouse or human gene or protein
      ii) Computational; Inferred from Mutant Phenotype (IMP)
         (1) Function predicted by computer – conserved domains
   b) The reference used to make the annotation (e.g. a journal article)
   c) • The date and the creator of the annotation
5) Not in GO
   a) Gene product
      i) ~~Cytochrome c~~
      ii) Oxidoreductase activity
   b) Process/function/components

i) Oncogenesis – not a normal function of a gene

c) Attribute of sequence
   i) intron/exon – part of SO
d) Protein domains / structural features
e) PPI
f) Environment, evolution, expression
g) No layers above cellular components

Links:

- http://geneontology.org/
- Gene Ontology Browser

The Sequence Ontology (SO)

The Sequence Ontology is a set of terms and relationships used to describe the features and attributes of biological sequence. SO includes different kinds of features which can be located on the sequence. Biological features are those which are defined by their disposition to be involved in a biological process. Examples are *binding_site* and *exon*. Biomaterial features are those which are intended for use in an experiment such as aptamer and *PCR_product*. There are also experimental features which are the result of an experiment.
The Sequence Ontologies are provided as a resource to the biological community. They have the following obvious uses:

- To provide for a structured controlled vocabulary for the description of primary annotations of nucleic acid sequence, e.g. the annotations shared by a DAS server (BioDAS, Biosapiens DAS), or annotations encoded by GFF3.
- To provide for a structured representation of these annotations within databases. Were genes within model organism databases to be annotated with these terms then it would be possible to query all these databases for, for example, all genes whose transcripts are edited, or trans-spliced, or are bound by a particular protein. One such genomic database is Chado.
- To provide a structured controlled vocabulary for the description of mutations at both the sequence and more gross level in the context of genomic databases.

Links:

- http://www.sequenceontology.org/
- wiki: http://www.sequenceontology.org/so_wiki/index.php/Main_Page

MISO: the Sequence Ontology Browser

Useful for the following:

- Search for a SO term by entering a SO term name or synonym in the query box above;
- Explore the structure of SO and browse for SO terms using the expandable, cascading tree on the left;
- Go to the detail page for a term where you can:
  - Get details about a term, its definition and relationships;
  - See graphical views of a term's place in the ontology and link to it's neighbors;
  - Export details about a term in a variety of formats;
  - And access and contribute detailed documentation about a term and it's history by linking through to the SO wiki.

Links:

- http://www.sequenceontology.org/browser/obob.cgi

The Ontology for Biomedical Investigations (OBI)

The Ontology for Biomedical Investigations (OBI) addresses the need for controlled vocabularies to support integration and joint ("cross-omics") analysis of experimental data, a need originally identified in the transcriptomics domain by the FGED Society, which developed the MGED Ontology as an

annotation resource for microarray data. OBI uses the Basic Formal Ontology upper level ontology as a means of describing general entities that do not belong to a specific problem domain. As such, all OBI classes are a subclass of some BFO class.

The ontology has the scope of modeling all biomedical investigations and as such contains ontology terms for aspects such as:

- biological material - for example blood plasma
- instrument (and parts of an instrument therein) - for example DNA microarray, centrifuge
- information content - such as an image or a digital information entity such as an electronic medical record
- design and execution of an investigation (and individual experiments therein) - for example study design, electrophoresis material separation
- data transformation (incorporating aspects such as data normalization and data analysis) - for example principal components analysis dimensionality reduction, mean calculation

Less 'concrete' aspects such as the role a given entity may play in a particular scenario (for example the role of a chemical compound in an experiment) and the function of an entity (for example the digestive function of the stomach to nutrient the body) are also covered in the ontology.
Links:

- http://obi-ontology.org/page/Main_Page

International Classification of Diseases (ICD)

The ICD is the World Health Organization's ontology for diseases. It is not the same as MeSH, though there is a bit of shared stuff between them.

Links:

- https://bioportal.bioontology.org/ontologies/ICD10

Disease Ontology / Mammalian Phenotype / Pathway Ontology

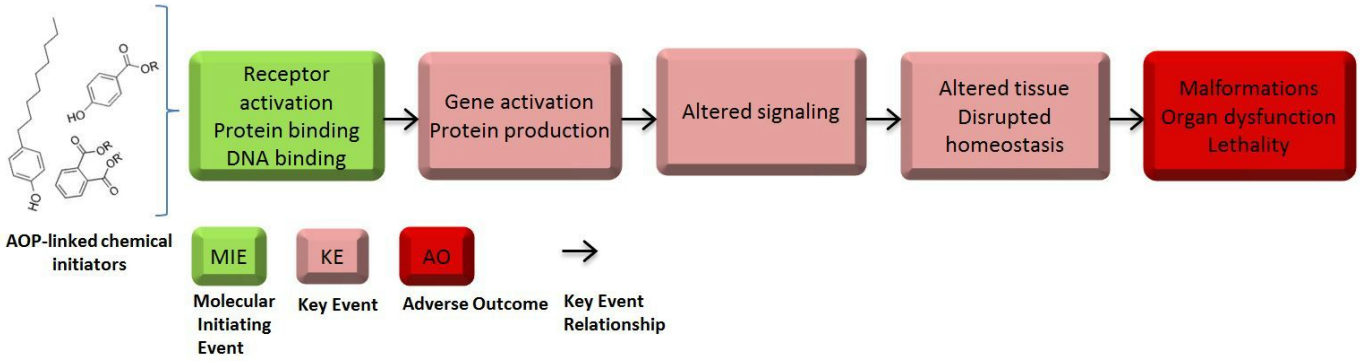
Biomedical Ontology

> *The National Center for Biomedical Ontology*

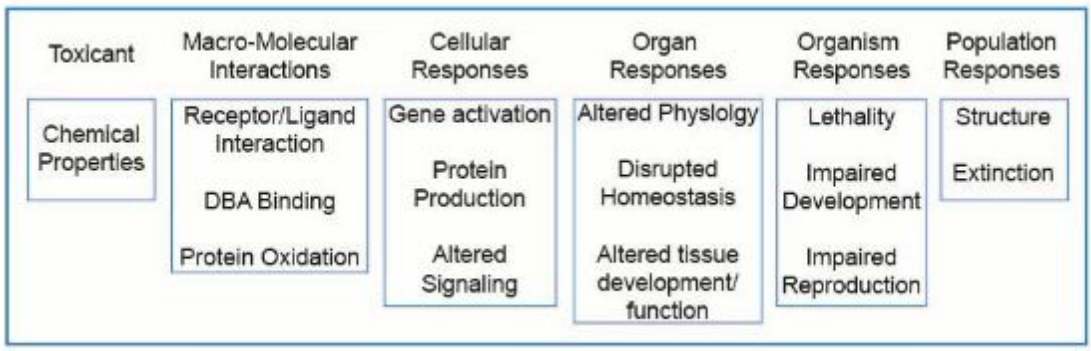
http://www.bioontology.org/mission

Challenges

Adverse Outcomes Pathology



Source: http://aopkb.org/background.html



Source: http://www.oecd.org/chemicalsafety/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm

Links:

- Adverse Outcome Pathway KB (AOP-KB) [http://aopkb.org/]
- http://www.oecd.org/chemicalsafety/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm
- http://aopkb.org/background.html - read this. it basically has the answer to hofmann's question

Systems Toxicology

- Toxicology - how chemicals mess stuff up. See ADME**T**
- How do chemicals modulate biological pathways
- Make models of this = systems toxicology
- Omis, Transcript, protein, and metabolite profiling

Useful Databases:

- BRENDA
- KEGG
- Reactome

Links:

- http://www.ncbi.nlm.nih.gov/pubmed/22562485
- http://www.nature.com/nrg/journal/v5/n12/execsumm/nrg1493.html
- http://pubs.rsc.org/en/content/articlelanding/2015/tx/c4tx00058g#!divAbstract