Relational DBs and Repositories

DAVID L. TABB, PH.D.

Overview

- Storing complex structures in files
- Scaling storage via relational databases
- Repositories for biomolecular data:
  - Sequences and variants
  - Transcriptomes
  - Structures
  - Proteomes

JSON: streamlined messages among tools

```
{
 "contentManagementSystems" : [
  {
  "name": "WordPress",
  "percentMarketShare": 58.9
  },
  {
  "name": "Joomla",
  "percentMarketShare": 6.1
  },
  {
  "name": "Drupal",
  "percentMarketShare": 4.9
  }
 ]
}
```

Array of objects

Name: Value pair

- Web services frequently communicate data by JSON.
- JavaScript Object Notation began at Netscape, but many languages now support it.

https://www.whoishostingthis.com/resources/json-resource/

UNIVERSITEIT STELLENBOSCH UNIVERSITY

XML labels contents through markup

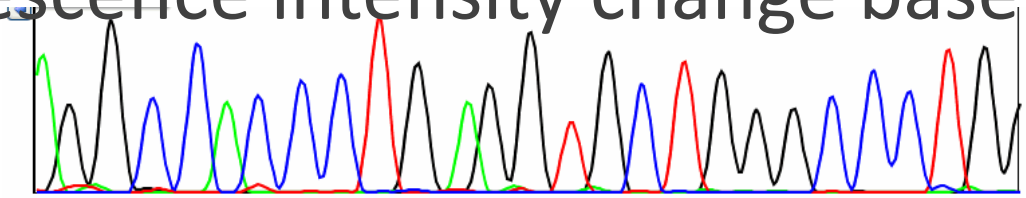
```
<?xml version="1.0" encoding="UTF-8"?>
<CATALOG>
 <CD>
    <TITLE>The very best of</TITLE>
    <ARTIST>Cat Stevens</ARTIST>
    <COUNTRY>UK</COUNTRY>
    <COMPANY>Island</COMPANY>
    <PRICE>8.90</PRICE>
    <YEAR>1990</YEAR>
  </CD>
  <CD>
    <TITLE>Unchain my heart</TITLE>
    <ARTIST>Joe Cocker</ARTIST>
    <COUNTRY>USA</COUNTRY>
    <COMPANY>EMI</COMPANY>
    <PRICE>8.20</PRICE>
    <YEAR>1987</YEAR>
  </CD>
</CATALOG>
```

- XML encloses data in tags that label content.
- This markup type is also used in HTML, a sibling.
- A separate DTD or schema defines allowable elements.

https://ccollins.wordpress.com/2008/03/03/a-brief-history-of-xml/

High-throughput Sequencing

- Electropherogram: a trace showing fluorescence intensity change base by base



- FASTQ (text): sequences of reads with per-base quality assessments

Modern sequencers routinely discard electropherograms due to size

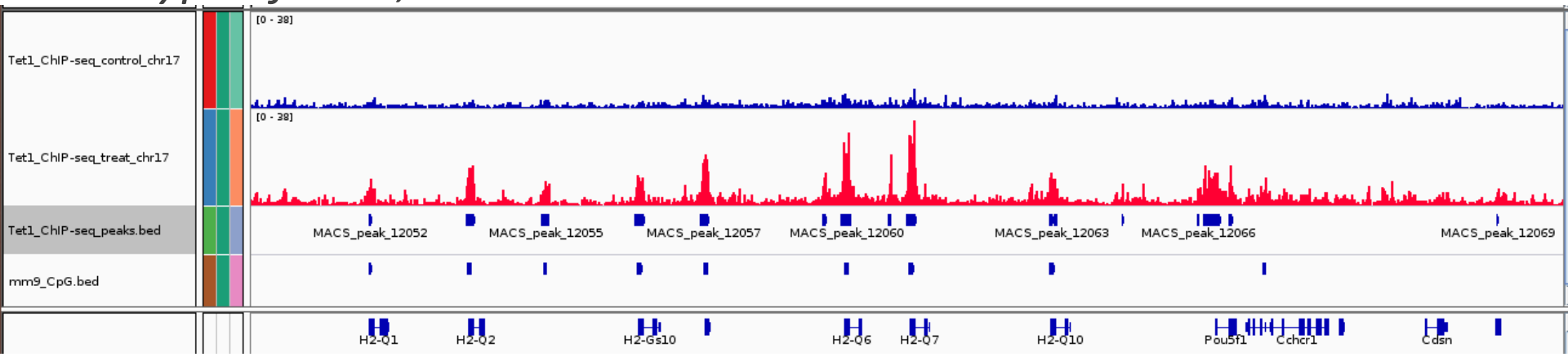
http://www.gendx.com/

P. J. A. Cock et al. *Nucl. Acids Res*. (2010) 38: 1767-1771

Reporting alignment products

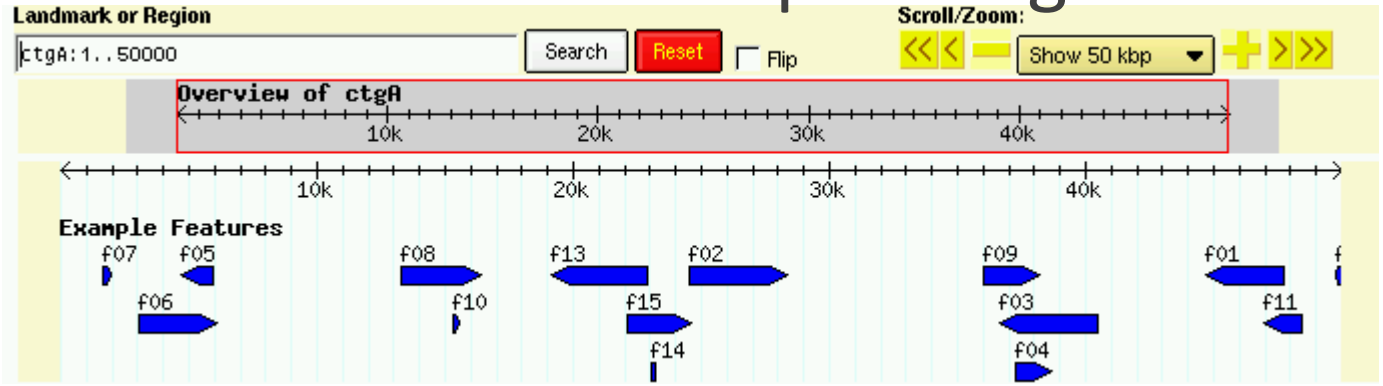
- SAM(text)/BAM(binary): Sequence Alignment Maps align reads to reference annotation
- BED (text): Browser Extensible Data defines genomic features for visualization

BAM and BED represent different emphases for the same type of data; converters exist.



Highly distilled information

- GFF (text): General Feature Format supplies coordinates of genetic features
- VCF (text)/BCF (binary): Variant Call Format details variant nucleotides (mutations or SNPs) for an individual from sequencing



https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

https://samtools.github.io/hts-specs/

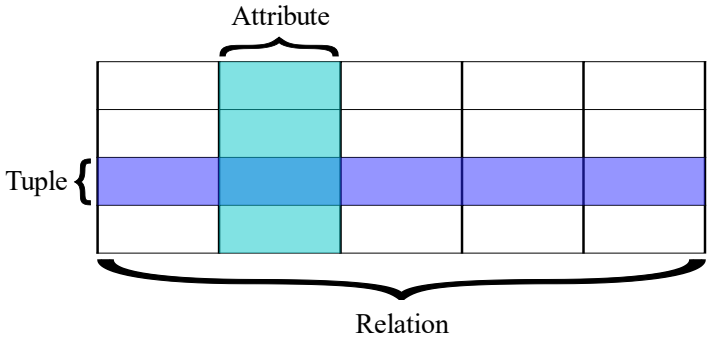
http://chlamycyc.mpimp-golm.mpg.de/gbrowse/tutorial/dbgff/tutorial.html

RDMS and SQL history

- Relational database management systems (RDMS) were defined in 1970 by E.F. Codd.
- IBM created the SEQUEL (later SQL) language to support RDMS manipulation.
- Essentially any modern database can support SQL queries: MS Access, SQLite, MySQL, PostgreSQL, Oracle…

https://www.w3schools.com/sql/

Keys ID each *tuple*

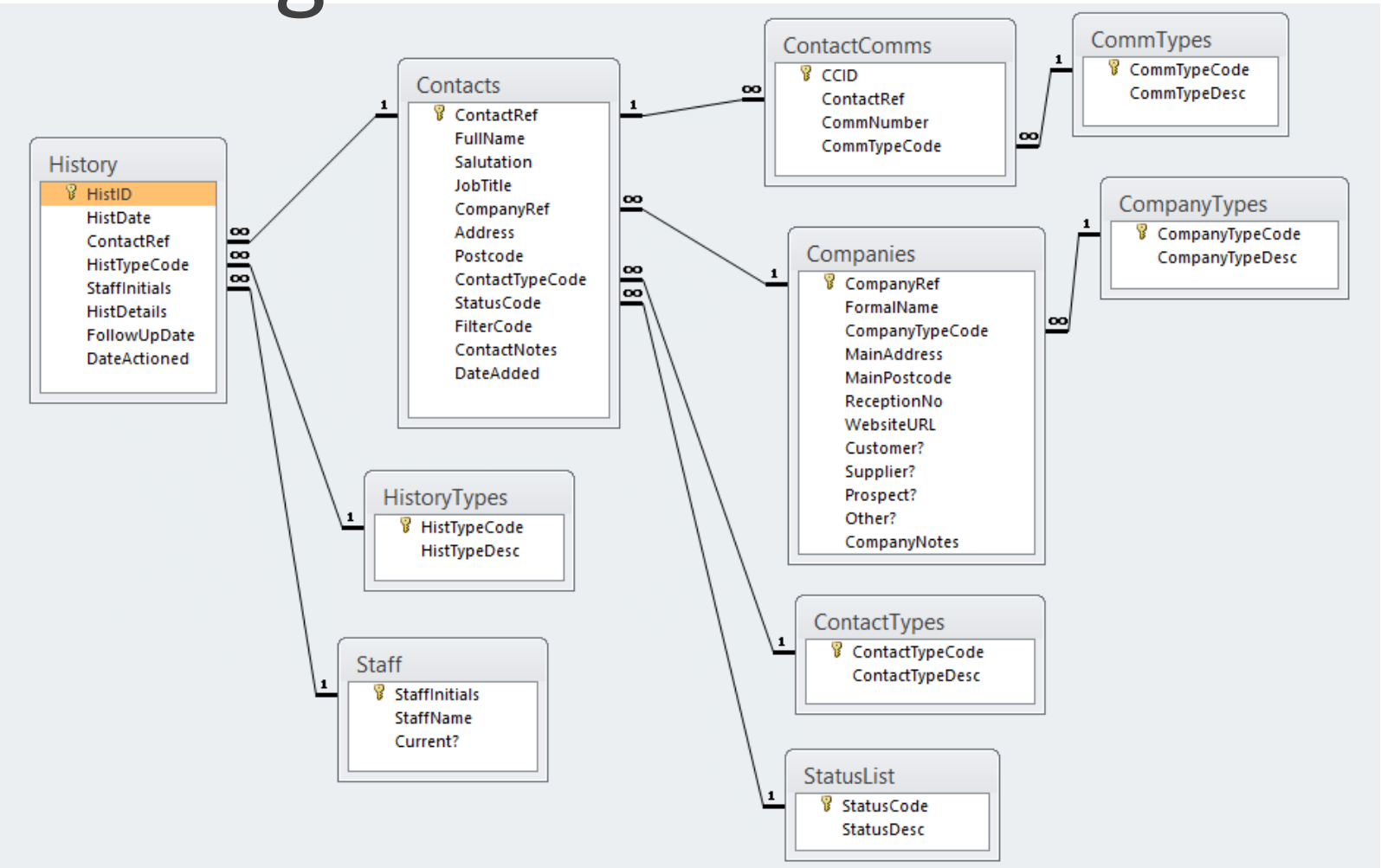


- A single record or row from a table is frequently called a *tuple*.
- Some tables have a column (attribute) that is unique for each row. This is a *primary key*.
- Keys make it possible to relate tables to each other very rapidly in a *join*.

http://www.databaseprimer.com/pages/keys/

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Joins connect information among tables



http://www.software-matters.co.uk/contact-management-database-image1.html

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Relational databases model complex structures

- Index: a lookup table of a unique value for each record that enables rapid retrieval.
- Inner Join: combine data from two tables, aligning on a field in common to both.
- Query: select fields of data from a body of data (whether single table or a “join”).

```
SELECT ID, NAME, AMOUNT, DATE FROM
CUSTOMERS INNER JOIN ORDERS ON
CUSTOMERS.ID = ORDERS.CUSTOMER_ID;
```


https://www.tutorialspoint.com/sql/

The case for normalization

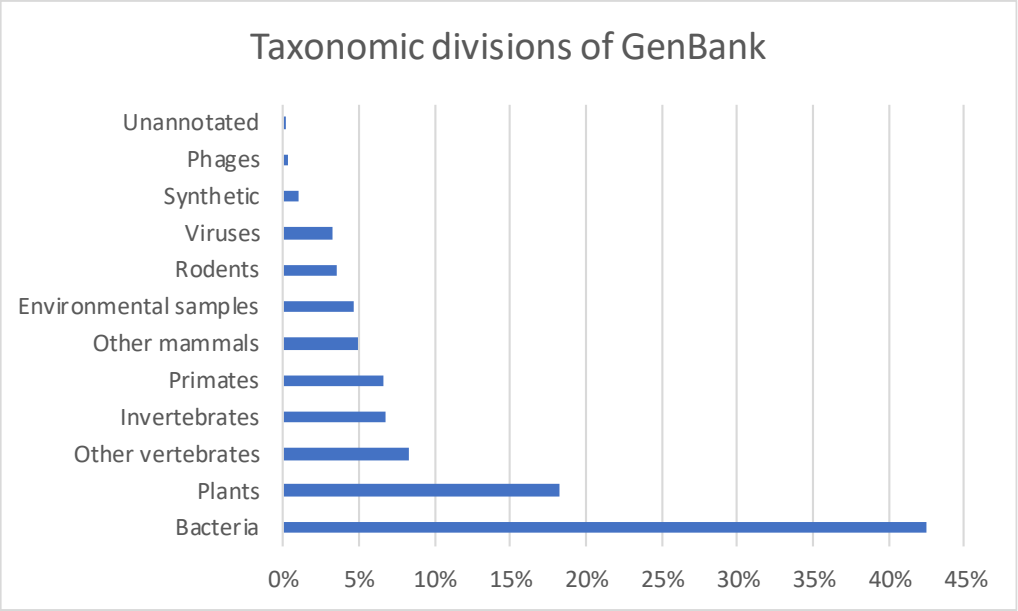
- Update anomaly: adding a record requires re-entry of extant data for consistency
- Insertion anomaly: we may not have all the fields we need when adding a record
- Deletion anomaly: when we remove a record, we lose more data than intended

http://www.studytonight.com/dbms/database-normalization.php

ID	Name	City	Subject
401	Adam	Durbanville	Bio
402	Alex	Bellville	Maths
403	Stuart	Cape Town	Maths
404	Adam	Durbanville	Physics

Repositories house vast collections of sequences.

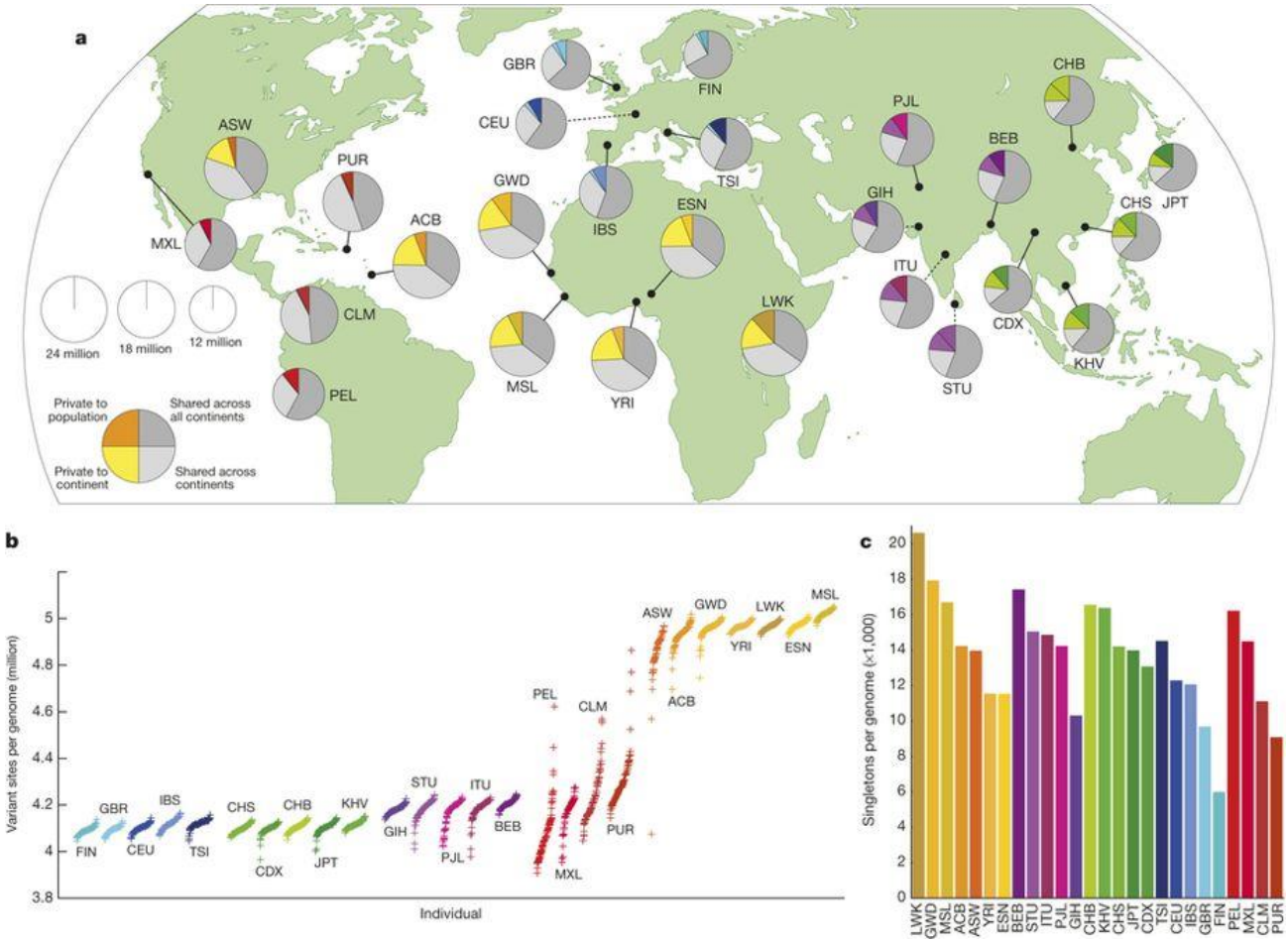
- GenBank / Euro Nucleotide Archive / DDBJ Synchronized DNA repositories at NCBI, European Molecular Biology Lab, and DNA Data Bank of Japan
- #1 species:
  Homo sapiens
  19 752 523 722 bp



EW Sayers et al. *Nucl. Acids Res*. (2019) 47: D94-D99

UNIVERSITEIT STELLENBOSCH UNIVERSITY

“1000 Genomes” project data: www.internationalgenome.org

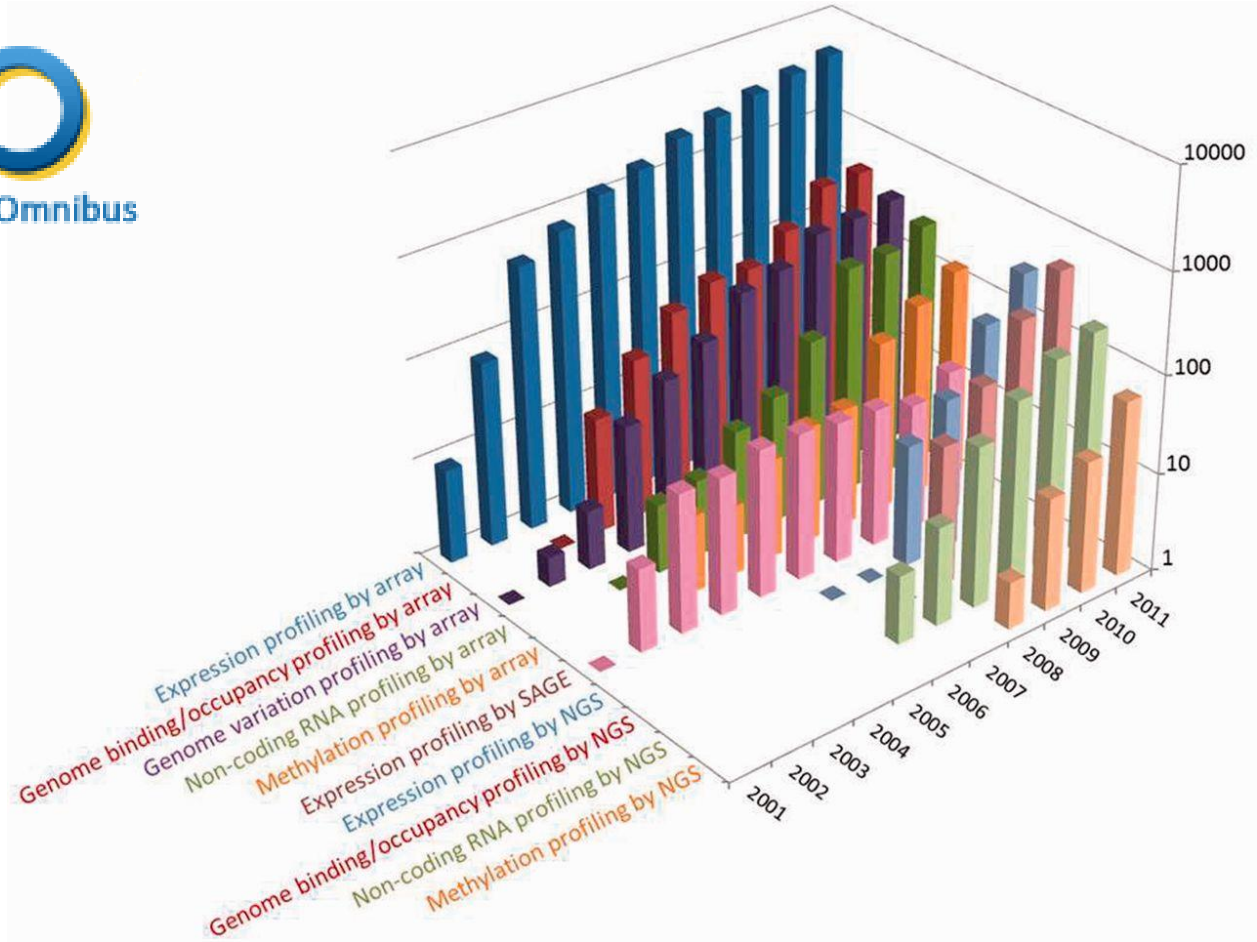


1000 Genomes Project Consortium. *Nature* (2015) 526: 68-74

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Gene Expression Omnibus: Functional genomics galore!

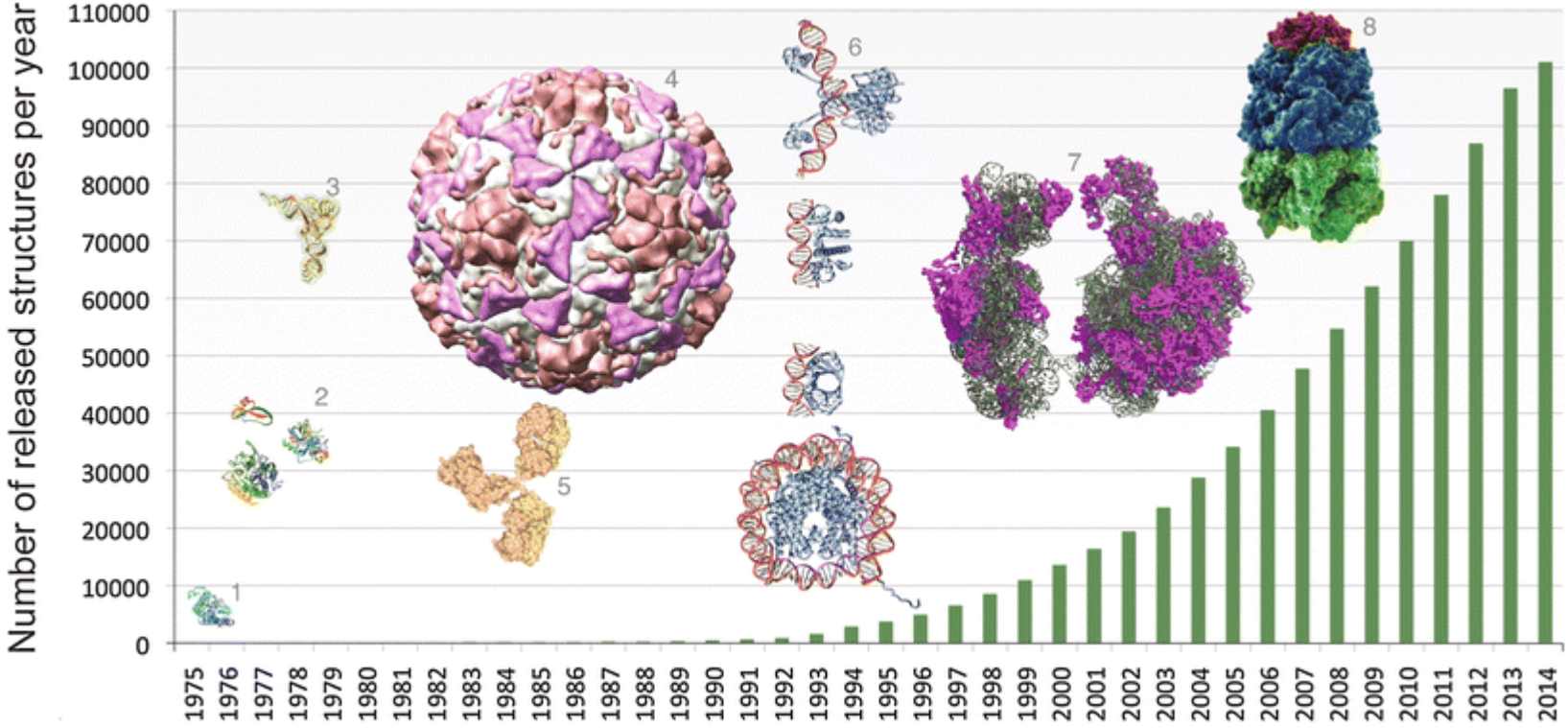
Gene Expression Omnibus



T Barrett et al. *Nucl. Acids Res*. (2013) 41: D991-D995

Protein structures since 1971

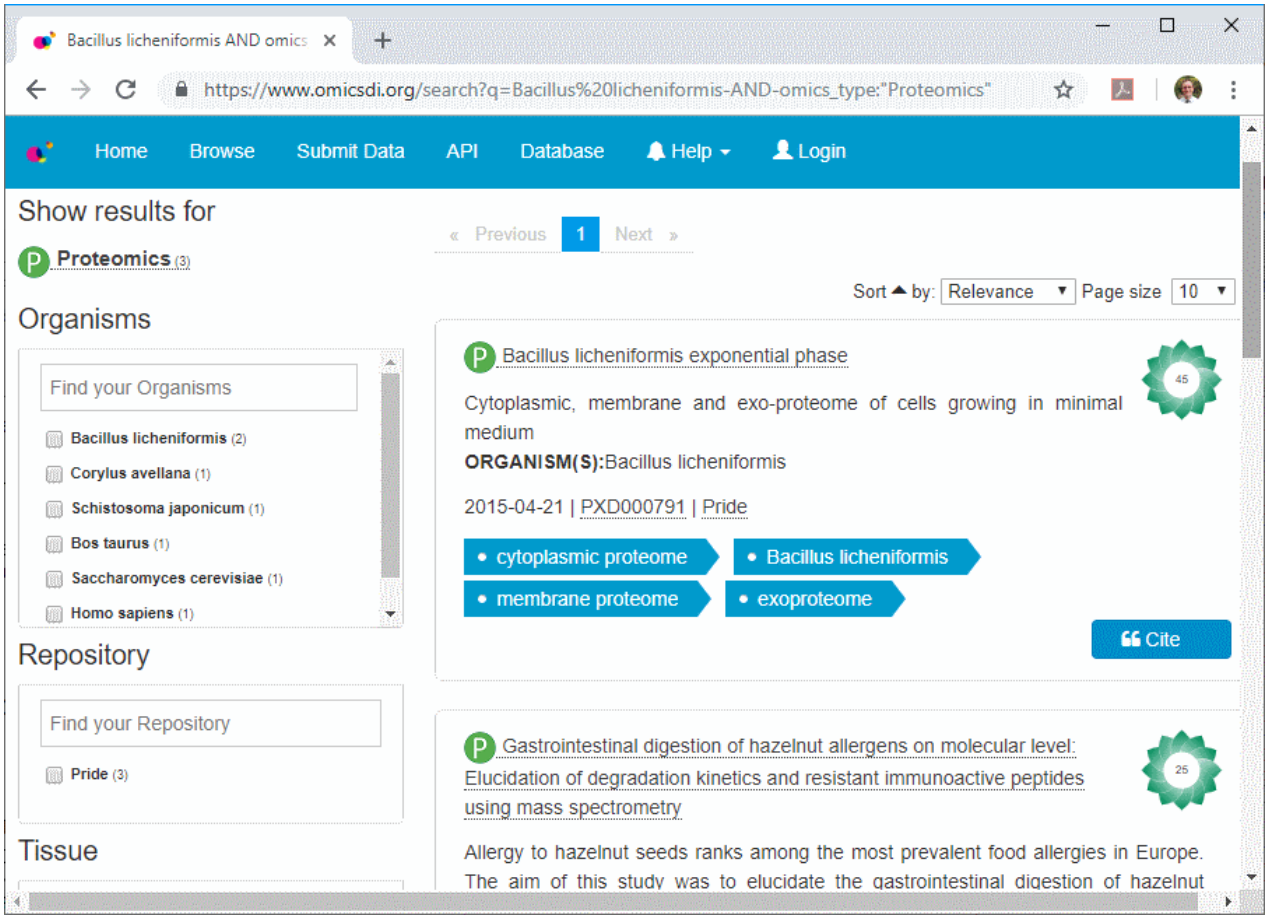
The Protein Data Bank is an international resource for structures.



http://www.wwpdb.org/

Berman et al. *J. Comp. Aid. Molec. Design* (2014) 28: 1009-1014

Proteomics data from via OmicsDI



EW Deutsch et al. *Nucl. Acids Res*. (2017) 45: D1100-D1106.

Closing thoughts

- Spreadsheets are fine for 2D tables of numbers, but knowing about databases empowers you to handle more complexity.
- Supporting your experimental data with those published to repositories makes your manuscript stronger.