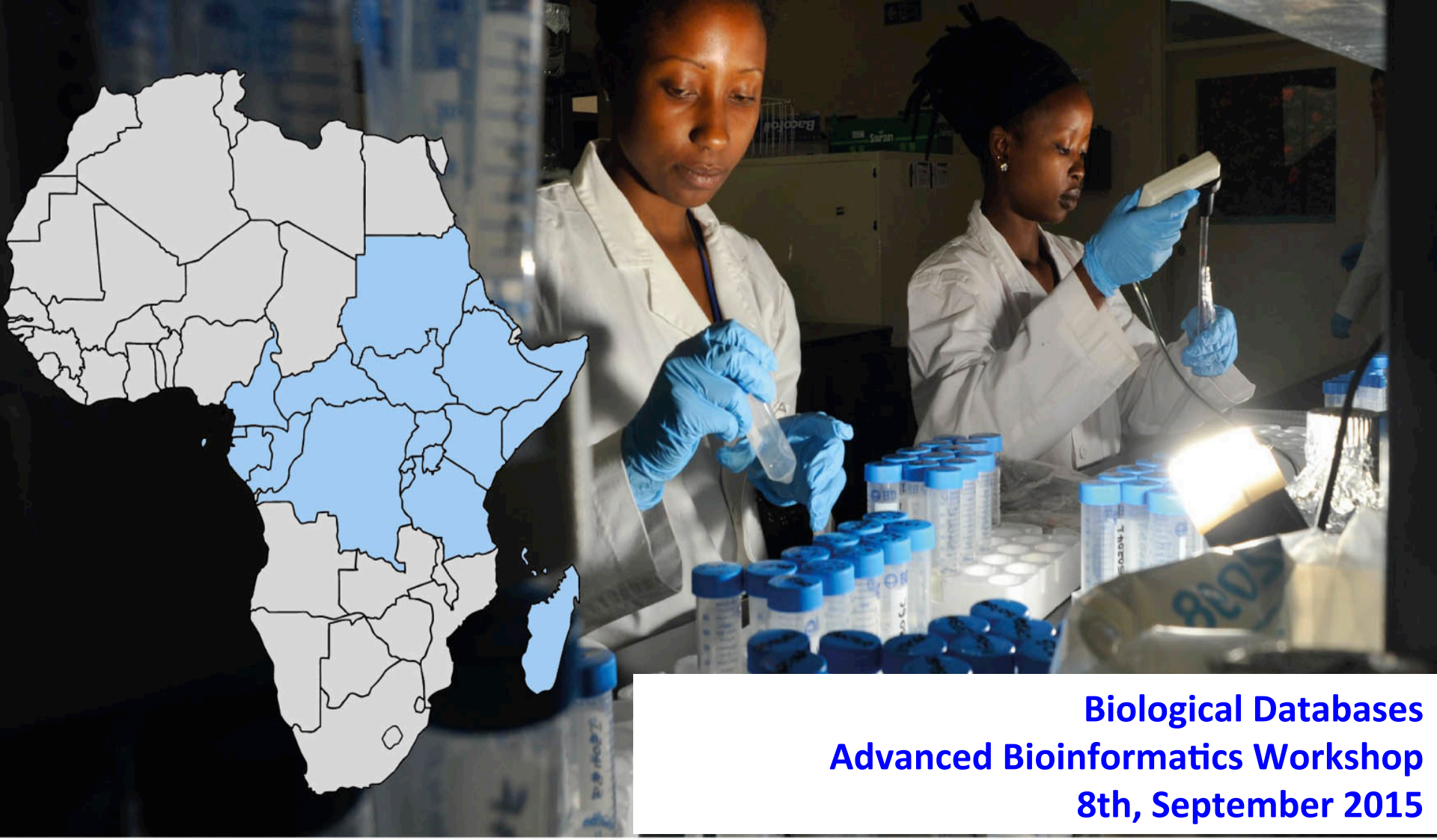


Biological Databases
Advanced Bioinformatics Workshop
8th, September 2015

Joyce Njoki Nzioki
BecA-ILRI Hub, Nairobi, Kenya
http://hub.africabiosciences.org/
http://www.Ilri.org/
j.n.njuguna@cgiar.org

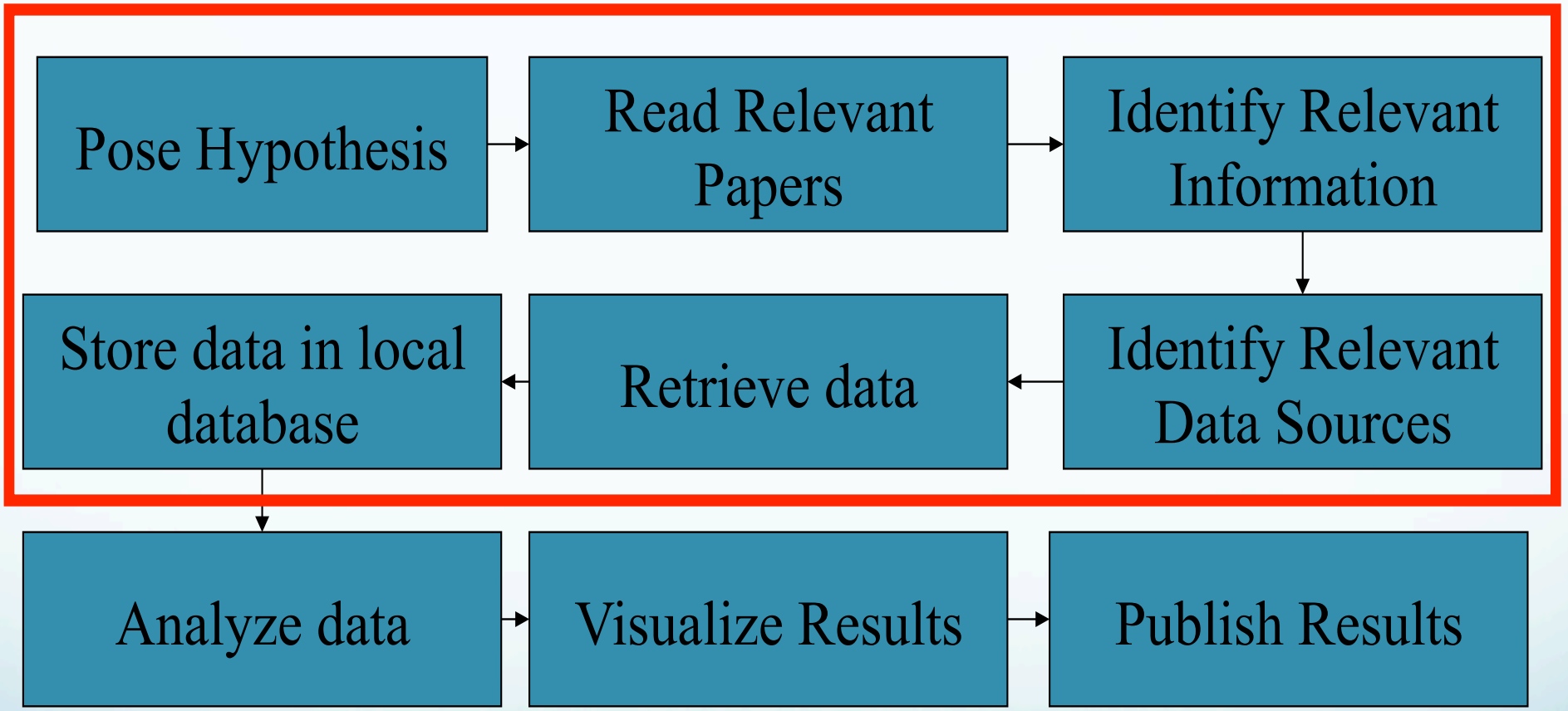
ILRI
INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

African Union

NEPAD
A PROGRAMME OF THE AFRICAN UNION

biosciences
eastern and central africa

Typical Bioinformatic Project



This is an iterative process. You may loop back at almost any step.

Data Acquisition

Identify relevant data

Data Acquisition

Identify relevant data

- In collaboration with researchers
- What is needed ?
  - ✓Gene sequences, amino acid sequences, structural information, literature
- What quality is required?
  - ✓No errors , some errors

Data Acquisition

Identify relevant databases

Data Acquisition

Identify relevant databases

- What databases contain the data you are interested in?
- Which ones have the required quality?
- Do you need general or specific data?

Data Acquisition

Retrieve Data

Data Acquisition

Retrieve Data

- What format is required?
- Track where and when data is retrieved
- Do you need to update the data frequently?
- How do updates impact your analysis

Biological Databases

- A biological database is a computerized archive used to store, organize and ease retrieval of sequence data

Biological Databases

- A biological database is a computerized archive used to store, organize and ease retrieval of sequence data
- A database typically supports the following operations
  - ✓Retrieval
  - ✓Insertion
  - ✓Updating
  - ✓Deletion

Biological databases

I. A database can be thought as a large table where row represents record and columns represents fields.

II. The organization of records allows for querying on the data to retrieve information from the databases

III. An ideal biological database has fields as shown below

Accession number	Name	Length	Sequence	Taxonomy	Reference
NR235462.1	MTGA	268	ACTTGC...	E.coli	A.Kelly et. al
NR235463.1	HKY	350	TGAGTA...	E.coli	J.Jone et. al
NR235464.1	THY	289	TGACGT...	S.Aurius	K.Moy et. al

Importance of Biological Databases

- Means to handle and share large volumes of biological data.

Importance of Biological Databases

- Means to handle and share large volumes of biological data.
- Biological databases represent an invaluable resource to support biological research. In some cases no need to sequence.

Importance of Biological Databases

- Means to handle and share large volumes of biological data.
- Biological databases represent an invaluable resource to support biological research. In some cases no need to sequence.
- Link knowledge to sequenced data – Cross referencing, most sequences databases are cross linked to load of biological information: gene/protein – information, Structural information, pathways and biological processes literature.

Types of Biological Databases

1. Primary databases: hold raw sequenced data
    - GenBank
    - EMBL (European Molecular Biology Laboratory)
    - DDBJ (DNA Data Bank of Japan)
    - PDB (Protein Data Bank)

Types of Biological Databases

1. Primary databases: hold raw sequenced data
    - GenBank
    - EMBL (European Molecular Biology Laboratory)
    - DDBJ (DNA Data Bank of Japan)
    - PDB (Protein Data Bank)
2. Secondary databases: they are curated and annotated
    - Swiss-Prot – detailed annotation if proteins
    - RefSeq – non-redundant , curated sequenced data

Types of Biological Databases

1. Primary databases: hold raw sequenced data
    - GenBank
    - EMBL (European Molecular Biology Laboratory)
    - DDBJ (DNA Data Bank of Japan)
    - PDB (Protein Data Bank)
2. Secondary databases: they are curated and annotated
    - Swiss-Prot – detailed annotation if proteins
    - RefSeq – non-redundant , curated sequenced data
3. Specialized databases: these focus on data of specific research interest
    - VectorBase
    - PlasmoDB

Where to look for Biological databases

- ✓ Search Engines (Google)

- ✓ Journals related to bioinformatics

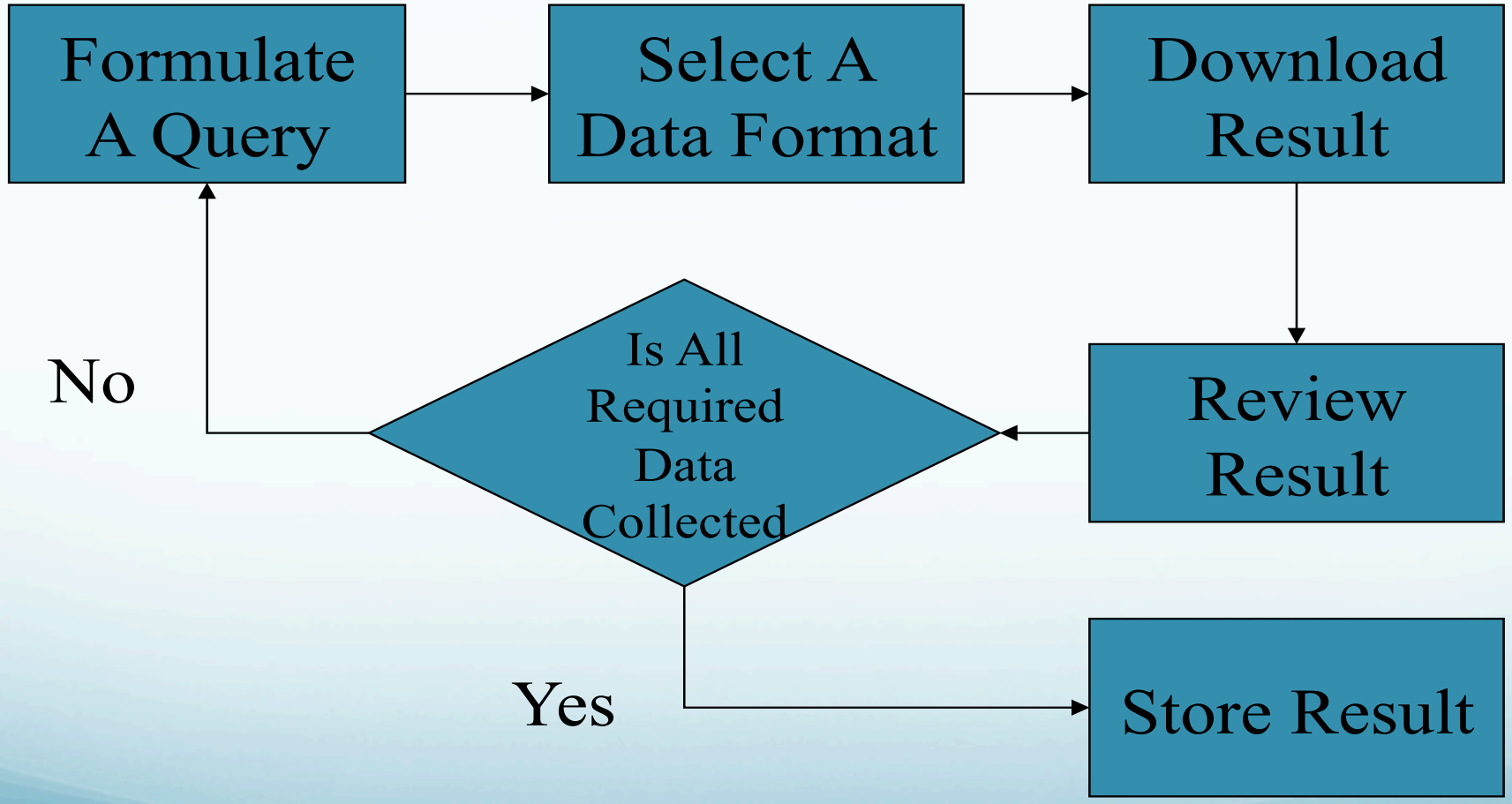
  Nucleic Acid research NAR online database issue

- ✓ Websites like; www.expasy.ch

Major Biological databases

- ✓ There are many public resources but few key resources
- ✓ Many databases are part of projects that have limited lifespan
- ✓ Use major public resources: **NCBI, EBI, Ensemble, PDB, KEGG**
- ✓ If your field is more specific, identify the major resource in that area

Retrieving Data



Accession Numbers

- ✓ Stable ways of identifying GenBank Entries.
- ✓ No biological meaning.
- ✓ Originally an uppercase letter followed by 5 digits U00002
- ✓ Now two uppercase letters followed by six digits BC037153
- ✓ Version of entry added later as a decimal BC037153.1

GenInfo (gi) IDs

- ✓ Identifier for a particular sequence only
  - Each entry gets a unique gi number
  - gi:22477487

- ✓ Not subject to versioning
  - The entry remains the same all the time
  - Different versions of the same sequence are managed by accession numbers.

Sequence formats

- ✓ This is the required arrangement of characters symbols and keywords in a sequence record.

Sequence formats

- ✓ This is the required arrangement of characters symbols and keywords in a sequence record.
- ✓ There many different sequence formats for the purpose of database integration and organizing sequenced data

Sequence formats

- ✓ This is the required arrangement of characters symbols and keywords in a sequence record.
- ✓ There many different sequence formats for the purpose of database integration and organizing sequenced data
- ✓ When considering a format for retrieval:
  - What is easy to parse
  - What format do the tools need
  - What information is needed

FASTA format

- Used by fasta tools
- Comment line > then sequence data

```
>gi|8547324|gb|AF271385.1| Fasciola hepatica cathepsin L mRNA, complete cds
GGGCAAACAATGAGATTGGTAATCCTAACCCTACTCATCGTCGGAGTGTTCGCCTCAAATGACGATTTGT
GGCATCAATGGAAGCGAATTTACAATAAAGAATACAATGGAGCTGACGATGACCACAGGAGAAATATTTG
GGAACAAAATGTGAAACATATCCAAGAACACAACCTGCGCCACGATCTCGGTCTCGTCACCTACAAGTTG
GGATTGAACCAATTCACCGATATGACATTCGAGGAATTCAAAGCCAAATATCTAACAGAAATGCCACGCG
CGTCTGAGTTACTCTCACACGGTATCCCATATAAGGCTAACAAGCGTGCTGTACCCGACAGAATTGACTG
GCGTGAATCCGGTTATGTGACGGAGGTGAAAGATCAGGGAGGCTGTGGTTCTTGTTGGGCTTTCTCAACA
ACAGGTGCTATGGAAGGACAGTATATGAAAAACCAAAGAACTAGTATTTCATTCTCTGAACAACAACTGG
TCGATTGTAGCCGTGATTTTGGCAATTATGGTTGTAATGGTGGACTAATGGAAAATGCATACGAATATTT
GAAACGATTTGGATTGGAAACCGAGTCTTCTTATCCTTACAGGGCTGTGGAAGGACAGTGTCGATACAAC
GAGCAGTTGGGAGTTGCCAAAGTGACTGGCTACTATACGGTACATTCTGGAGATGAGGTAGAATTGCAAA
ATCTAGTCGGTGCCGAAGGACCTGCTGCGGTCGCTTTGGATGTGGAGTCAGACTTCATGATGTACAGGAG
TGGTATTTATCAGAGCCAAACTTGTTCACCGGATCGTTTGAACCATGGAGTGTTGGCTGTCGGTTATGGA
ATACAGGATGGTACTGACTACTGGATTGTGAAAAACAGTTGGGGAACGTGGTGGGGTGAGGACGGTTACA
TTCGAATGGTTAGGAAAAGAGGTAACATGTGTGGAATTGCTTCTCTGGCCAGTGTCCCGATGGTGGCACA
ATTTCCGTGA
```

GenBank format

- Flat file format used by GenBank
- Has annotation, author, version etc

```
LOCUS       MMU35641                5538 bp    mRNA    linear   ROD 18-OCT-1996
DEFINITION  Mus musculus Brca1 mRNA, complete cds.
ACCESSION   U35641
VERSION     U35641.1  GI:1040960
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus.
REFERENCE   1  (bases 1 to 5538)
  AUTHORS   Sharan,S.K., Wims,M. and Bradley,A.
  TITLE     Murine Brca1: sequence and significance for human missense
            mutations
  JOURNAL   Hum. Mol. Genet. 4 (12), 2275-2278 (1995)
   PUBMED   8634698
REFERENCE   2  (bases 1 to 5538)
  AUTHORS   Bradley,A.
  TITLE     Direct Submission
```

Sequence formats

- Sequence are store in databases or in files as simple text (ASCI text)
- Microsoft Word format is not a sequence format
- Save sequence files as text .txt file !!!
- Use text editors like note-pad, text-pad to open such files

Problems of general sequence databases

i. Redundancy of sequence information
ii. Inadequate sequence
iii. Old sequences
iv. Partially annotated sequences
v. Inconsistent and outdated annotations
vi. Error sequences

Searching sequence databases

- ✓ Newly sequenced DNA data is compared to that already available in biological databases.

- ✓ Sequence comparison (of DNA / Protein data) is achieved through alignment, the process by which regions of similarity is searched between sequences.

- ✓ This eases annotation of new sequences as biological knowledge from well characterized homologs can be conferred

Searching sequence databases

1. Sequence similarity; this is when two sequences are very alike in base pair or amino acid sequence
   - ✓ Statistical measures like E-value. P-Value and bit score
   - ✓ Percentage identity (% of identical residues between sequences)
   - ✓ The length of sequence stretch that is similar

Some terminology

1. Sequence similarity; this is when two sequences are very alike in base pair or amino acid sequence
   - ✓ Statistical measures like E-value. P-Value and bit score
   - ✓ Percentage identity (% of identical residues between sequences)
   - ✓ The length of sequence stretch that is similar
2. Homology; homologs diverse from a common ancestor and homology is inferred by sequence, structural and functional similarity
   - ✓ Orthologs – arise due to a speciation event "same gene ini diff species"
   - ✓ Paralogs – arise due to gene duplication within the sequence

Searching sequence database

- ✓ A query sequence is searched against a database to look for homologs.
- ✓ The algorithm used aligns your query to those in the database and returns highly similar sequences.
- ✓ A scoring procedure is implemented on searches to measure the degree of similarity.
- ✓ Judgment needs to be made on whether the similar sequences are homologous to your query based on scientific knowledge
- ✓ There 2 programs for this:
  1. BLAST (Altschul et al. 1990)
  2. FastA (Pearson and Lipman 1988)

Conclusion

Biological data management has many challenges hence:

- Organize your data
- Use appropriate databases
- Know what king of information to expect
- Use appropriate tools

The End

Acknowledge Etienne for some of the slides on biological databases

Thank you