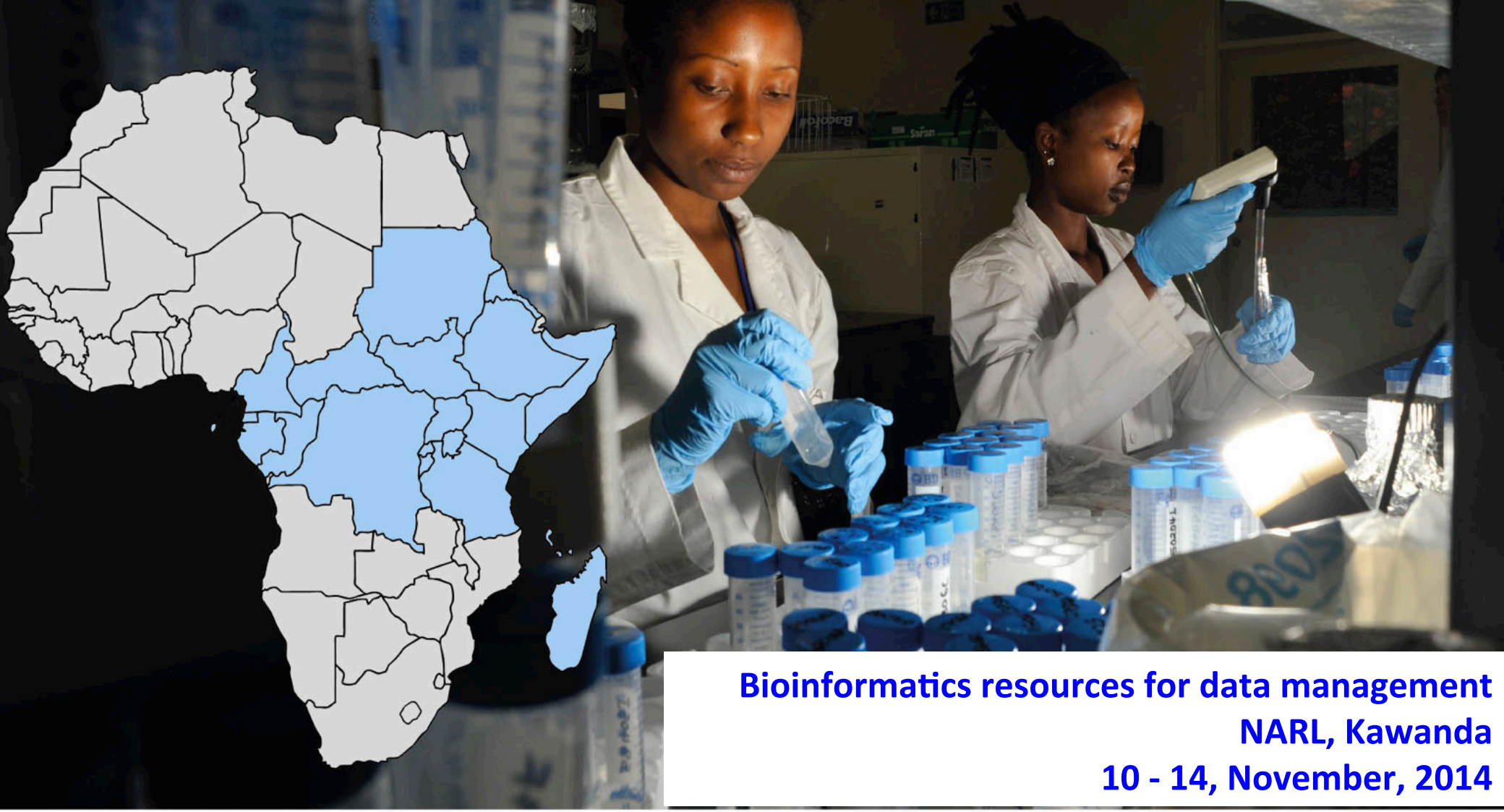


Bioinformatics resources for data management

NARL, Kawanda

10 - 14, November, 2014

Joyce Njoki Nzioki

BecA-ILRI Hub, Nairobi, Kenya

http://hub.africabiosciences.org/

http://www.Ilri.org/

j.n.njuguna@cgiar.org

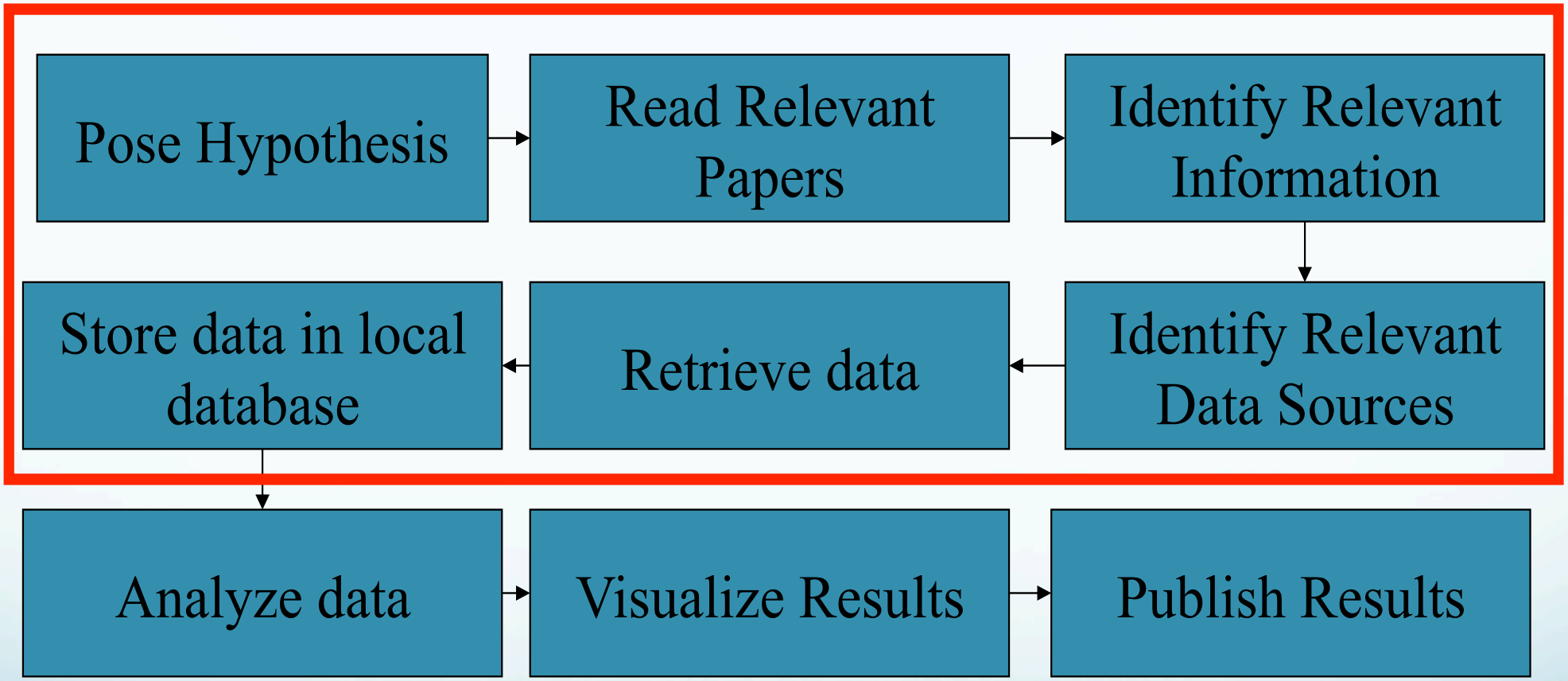
ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

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

ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE
African Union
NEPAD
biosciences eastern and central africa

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**
  - 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 and structural data.

- A database typically supports the following operations
  - Retrieval
  - Insertion
  - Updating
  - Deletion

Biological databases

I. A database can be thought of as a large table where row represent records and columns represent fields.

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

III. An ideal biological databases will have the fields sequences, accession number, taxonomic data, references, annotation, keywords, cross-references

Accession number	Name	length	Sequence	Taxonomy	Reference
NR235462.1	MTGA	268	ACTGTGCA..	E.coli	a.Kely et. al
NR235463.1	HKY	350	TGAGTATG...	E.coli	J.Jone et. al
NR235464.1	THY	289	TGACGTCA..	S.aurius	k.moy et. al

Biological databases

- There different categories of biological databases:
  1. **Primary databases**: that hold raw sequenced data
     - » GenBank.
     - » EMBL (The European Molecular Biology Laboratory)
     - » DDJB (The DNA Data Bank of Japan)
     - » PDB (Protein Data Bank)
  2. **Secondary databases**: that have curated and annotated data
     - » SWISS-PROT- detailed annotation of protein
     - » RefSeq- non-redundant, curated sequenced data
  3. **Specialized databases**: these focus on data of specific research interest
     - » Vector base
     - » 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

Nucleic Acids Research

Database issue

Volume 42 Issue D1 1 January 2014

For checked items
view abstracts download to citation manager Go Clear

I. Nucleic acid sequence, structure and regulation

Xosé M. Fernández-Suárez, Daniel J. Rigden, and Michael Y. Galperin
The 2014 *Nucleic Acids Research* Database Issue and an updated NAR online Molecular Biology Database Collection
Nucl. Acids Res. (1 January 2014) 42 (D1): D1-D6 doi:10.1093/nar/gkt1282
» Abstract » **FREE** Full Text (HTML) » **FREE** Full Text (PDF) » **FREE** Screen PDF
» Database Summaries » Permissions

NCBI Resource Coordinators
Database resources of the National Center for Biotechnology Information
Nucl. Acids Res. (1 January 2014) 42 (D1): D7-D17 doi:10.1093/nar/gkt1146
» Abstract » **FREE** Full Text (HTML) » **FREE** Full Text (PDF) » **FREE** Screen PDF
» Permissions

Catherine Brooksbank, Mary Todd Bergman, Rolf Apweiler, Ewan Birney, and Janet Thornton
The European Bioinformatics Institute's data resources 2014
Nucl. Acids Res. (1 January 2014) 42 (D1): D18-D25 doi:10.1093/nar/gkt1206
» Abstract » **FREE** Full Text (HTML) » **FREE** Full Text (PDF) » **FREE** Screen PDF
OPEN ACCESS (cc) BY

« Previous | Next Issue »

This Issue
1 January 2014 42 (D1)

» Index By Author
» Cover Image
» Table of Contents (PDF)

I. Nucleic acid sequence, structure and regulation
II. Protein sequence and structure, motifs and domains
III. Metabolic and signalling pathways, enzymes
IV. Viruses, bacteria, protozoa and fungi
V. Human genome, model

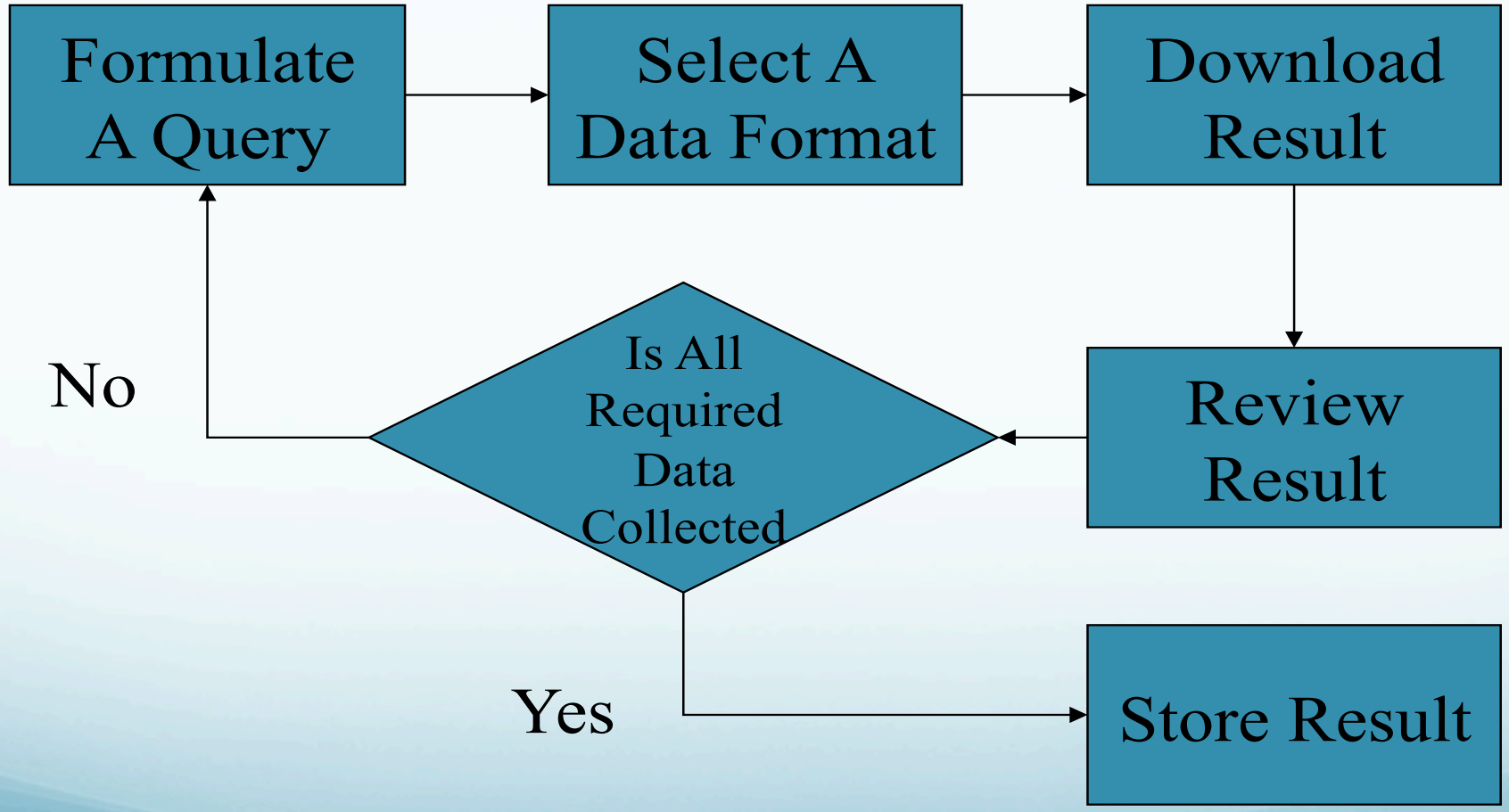
Presents new databases and updates of excising one

It is freely available online

Major Biological databases

- There are many public resources but few key resources
- Many databases are part of projects that have a limited lifespan
- Use major public resources: NCBI, EBI, Ensembl, PDB, KEGG
- If your field is more specific, identify the major resources in that area (e.g IMG for microbial genomics)

Retrieving Data



Accession Numbers

- Stable way 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 al the time
  - Different versions of the same sequence are managed by accession numbers

Sequence Formats

- These are required arrangements of characters, symbols and key words 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 (or Pearson)

- Used by FASTA tools
- Comment line followed by sequence data
- Example

```
>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

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

```
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

- Sequences are stored 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 test editors like note-pad, text pad or note pad++ to open such files

Problems of general sequence databases

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

CONCLUSION

<u>Biological data management has many challenges hence:</u>

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

Secondary/ Curated Databases

- **SWISS_PROT** – Manually annotated, non-redundant, cross-references, fully documented database of protein sequences.
- **TrEMBL** - Is automatically generated from annotated EMBL coding sequences (CDS) and annotated using software. It contains all that is not in SWISS-PROT before manual curating.
- **NCBI-RefSEq** – Non-redundant, annotated, cross referenced database holding genomic, transcript and protein data.
- **RDP** – provides annotated bacterial and Archeal 16S rRNA sequences.

123RF

Sequence Comparison

- Sequence Comparison is the start of structural and functional analysis of sequenced data.
- Newly sequenced DNA data is compared to that that is already available in biological databases.
- This eases annotation of new sequences as biological knowledge from well characterized homologs can be conferred.
- Sequence comparison is achieved through alignment, the process by which regions of similarity is searched between sequences.
- Nature acts in a conservative manner, where there is high level of similarity at the molecular level in organisms of the same common ancestor.

Some Terminology

1. **Sequence similarity: this is when two sequences are very alike in base pair or amino acid sequence. This is defined by**
   - Statistical measures like the E-value, P- value and Bit score
   - Percentage identity (% of identical residues between sequences)
   - The length of the sequence stretch that is similar.
2. **Homology: homologs diverge from a common ancestor and homology is inferred by the sequence, structural and functional similarity.**
   - Orthologs- arise due to a speciation event (genes that perform same function in different species)
   - Paralogs- arise due to gene duplication within the sequence. (homologous genes that perform different function)

If sequence similarity is high enough it can infer homology and consequently structural and functional similarity <u>but not in all cases</u>

Sequence Divergence

Sequences diverge over time and evolution is due to mutations

1. Substitutions

GATTCGTACG

GATGCGTACG

Mismatch due to substitution

2. Insertion or deletions

GATTCGT _ _ _ ACG

GATTCGTGGTACG

Gaps due to insertion or deletion

3. Duplication
   - Complete gene duplication
   - Partly gene duplication
   - Partial polysomy (part of a chromosome duplicates)
   - Poliploidy (a whole genome)

ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE | African Union | NEPAD | biosciences eastern and central africa

Sequence Alignment

- Pairwise sequence alignment: comparison of two sequences to establish conserved regions.
- **Why align sequences:**
  - Identify sequences with significant similarity / homology
  - Identify the domains with sequence, structure & function similarity.
  - Database similarity searching

- **There are two alignment strategies:**
  - **Global alignment**: finds optimal alignment over the entire length of the two compared sequences.
    - Best for highly similar sequence
    - May miss important biological relationships in low similarity
  - **Local alignment:** aligns short regions of similarity between sequences.
    - Useful when looking for domains in proteins and gene finding in DNA
    - Best for sequences of low similarity and different length.

Global vs local sequence alignment

1. **Global** (Needleman – Wunsch)

```
FTFTALILLAVAV
F--TAL-LLA-AV
```

2. **Local** (Smith – Waterman algorithm)

```
FTFTALILL-AVAV
--FTAL-LLAAV--
```

Global vs local Algorithms

1. **Needleman – Wunsch**
   - Exhaustive global alignment
   - most rigorous method when aligning conserved sequences of similar length

2. **Smith – Waterman algorithm**
   - Exhaustive local alignment
   - Alignment does not have to extend along the full length of the sequences
   - In contrast to N-W alignments initiating at all possible positions of the sequence-space will be considered
   - Can be very slow

Searching sequence databases

- A query sequence is searched against a database to look for homologs.
- The algorithms used align your query sequence to those in the database and return highly similar sequence.
- A scoring procedure is implemented in searches to measure degree of similarity.
- Judgment needs be made on whether the similar sequences are homologous to our query based on scientific knowledge.
- There are two programs that are effective in similarity searching:
  - FastA (Pearson and Lipman 1988)
  - BLAST (Altschul et al. 1990)

Acknowledgement for some of the slides

Thanks to Etienne de Villiers (KEMRI WELCOME TRUST)
Erik Bongcam-Rudloff (SLU)

THE END

ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE
African Union
NEPAD A PROGRAMME OF THE AFRICAN UNION
biosciences eastern and central africa