Next Generation Sequencing

Section of Population Genetics

Center for life and food sciences Weihenstephan

Introduction

- Name ?
- Department ?
- Study Interest ?
- Previous Experience ?
- Why NGS ?
- Expectations ?

What is NGS

NGS ~~ High Throughput sequencing

- Ilumina
- 454 (Roche)
- Solid (Abi)
- Pac Bio
- Ion torrent
- …..

Only Ilumina

Is NGS actually New?

Traditional Sequencer

~ 1000 base pairs in 1 run

Is NGS actually New?

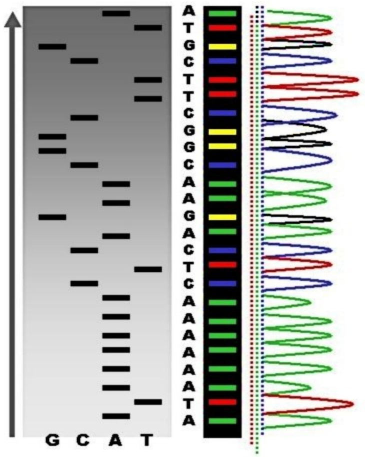
Think Big!

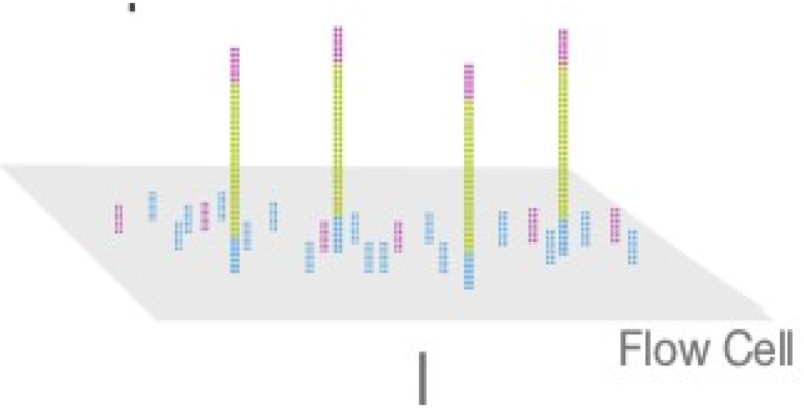
~ 1000 base pairs in 1 run

300 Billion base pairs in 1 run

How it's achieved?

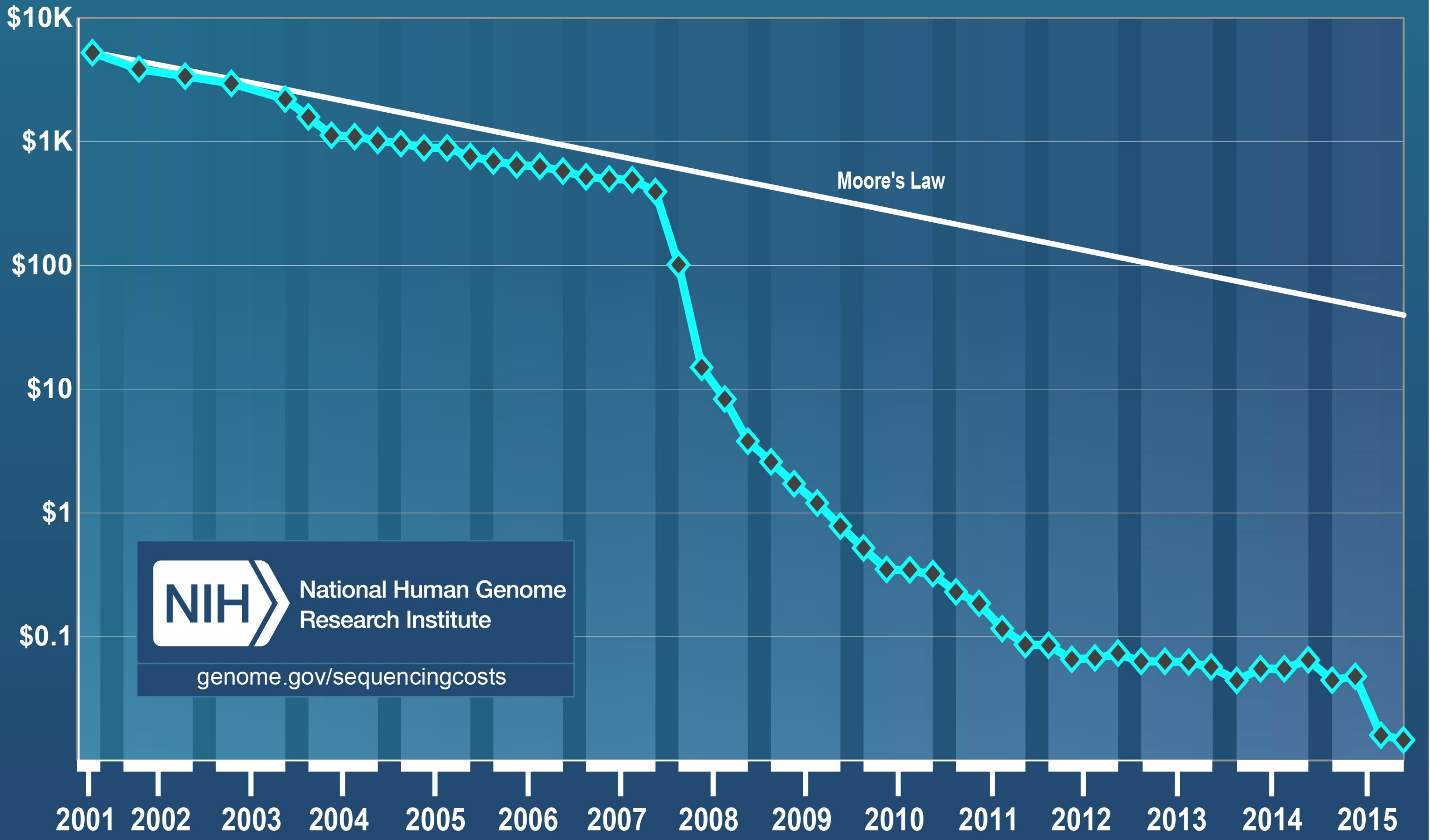
- NGS ~~ Massive Parallelization
- NGS ~~ Massive Miniaturization
- NGS ~~ Massive Cost Reduction





Moore's Law : Number of Transistors in ICs doubles every 2 years

Cost per Raw Megabase of DNA Sequence

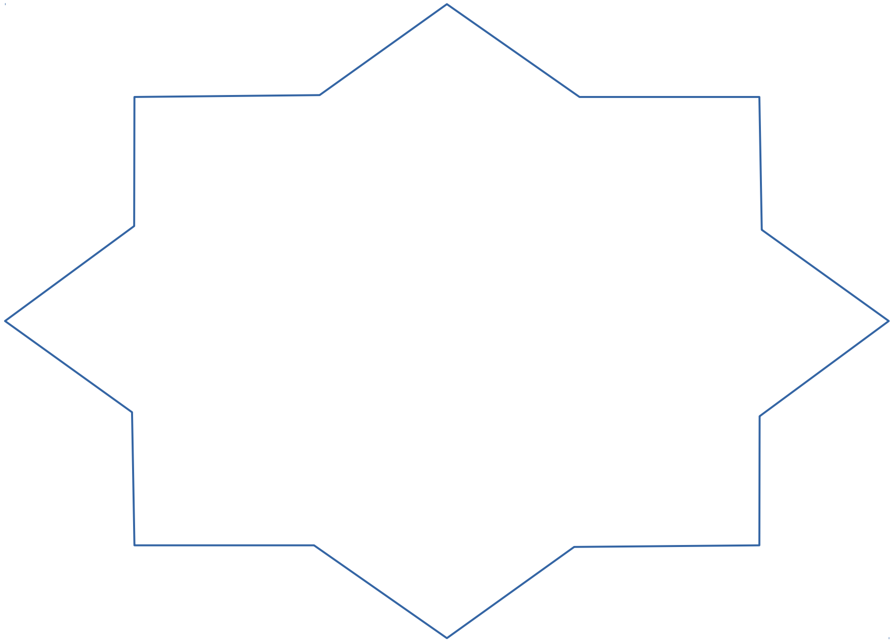


National Human Genome Research Institute

genome.gov/sequencingcosts

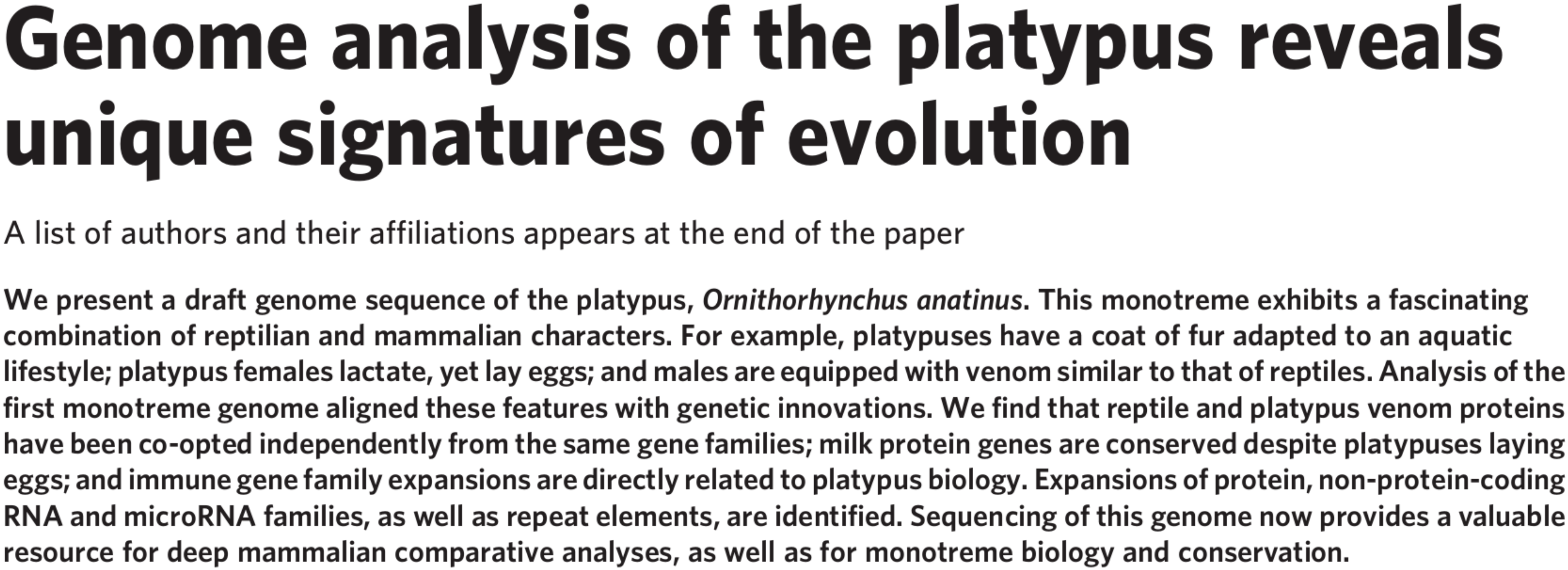
What can be done

De Novo
Genome Sequencing
(Assembly)



De Novo Genome Sequencing

- No Prior genome is available
- Massive coverages and Costs
- Optical and/or genetic maps needed
- Complicated!



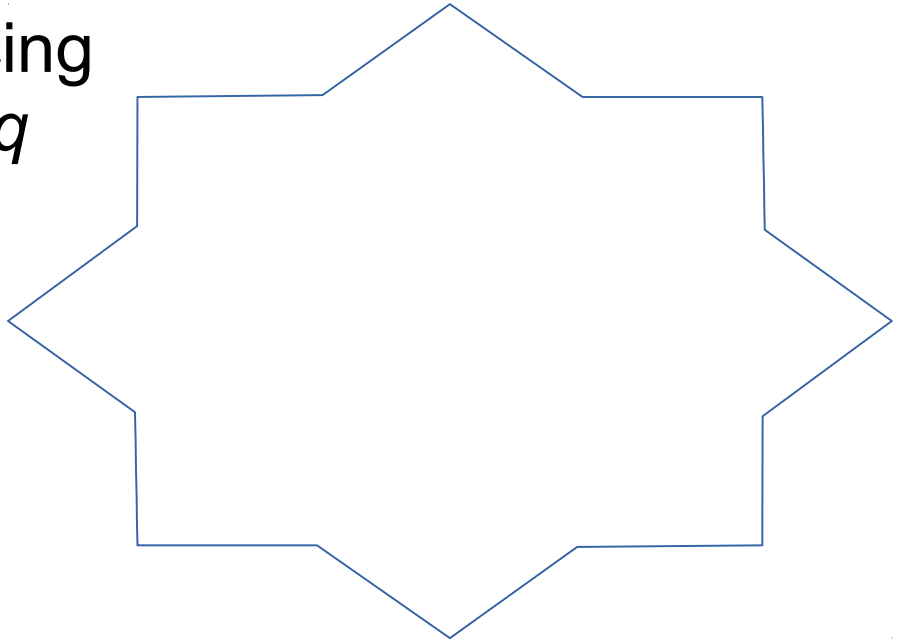
Genome analysis of the platypus reveals unique signatures of evolution

A list of authors and their affiliations appears at the end of the paper

We present a draft genome sequence of the platypus, *Ornithorhynchus anatinus*. This monotreme exhibits a fascinating combination of reptilian and mammalian characters. For example, platypuses have a coat of fur adapted to an aquatic lifestyle; platypus females lactate, yet lay eggs; and males are equipped with venom similar to that of reptiles. Analysis of the first monotreme genome aligned these features with genetic innovations. We find that reptile and platypus venom proteins have been co-opted independently from the same gene families; milk protein genes are conserved despite platypuses laying eggs; and immune gene family expansions are directly related to platypus biology. Expansions of protein, non-protein-coding RNA and microRNA families, as well as repeat elements, are identified. Sequencing of this genome now provides a valuable resource for deep mammalian comparative analyses, as well as for monotreme biology and conservation.

What can be done

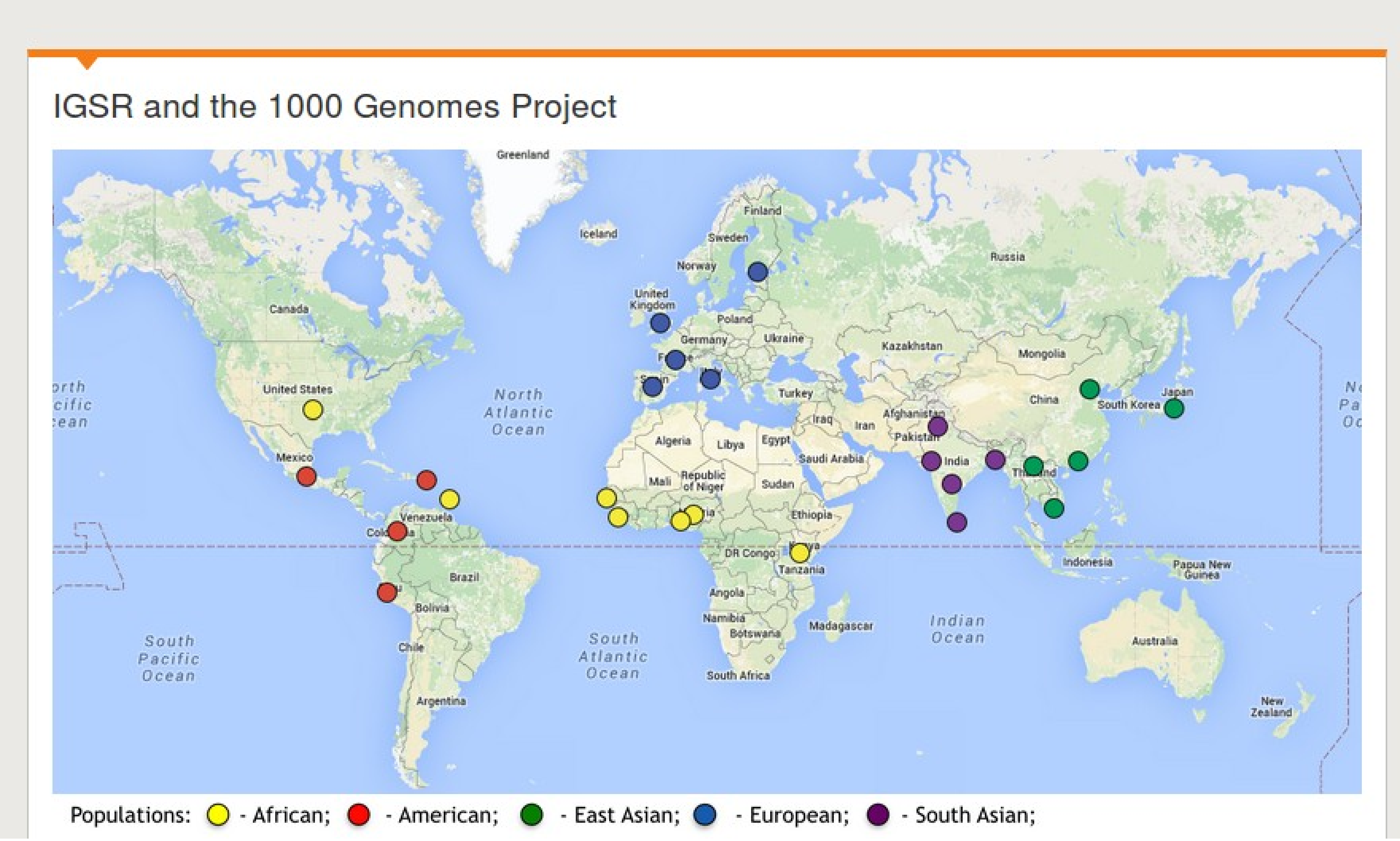
Resequencing
DNA-Seq



Resequencing

- Genome is already sequenced and used as reference
- A population or an individual sample is sequenced and mapped to the the reference
- Even a part of genome can be resequenced (Exomes)
- Case control, variant discovery, genotype -> phenotype, Causal variants
- Common, relatively cheap

1000 genomes project

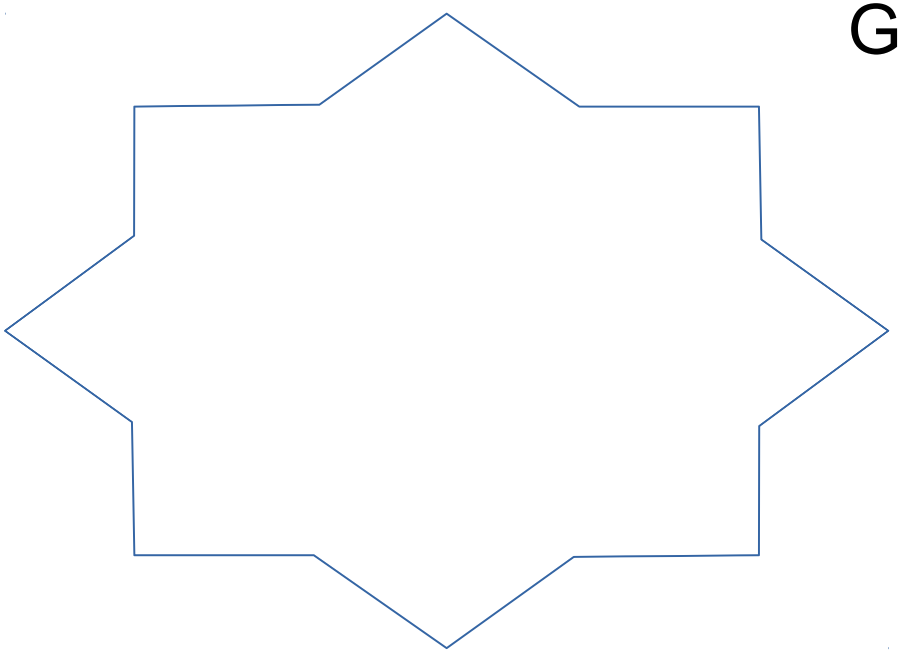


IGSR and the 1000 Genomes Project

Populations: ● - African; ● - American; ● - East Asian; ● - European; ● - South Asian;

http://www.1000genomes.org/

What can be done



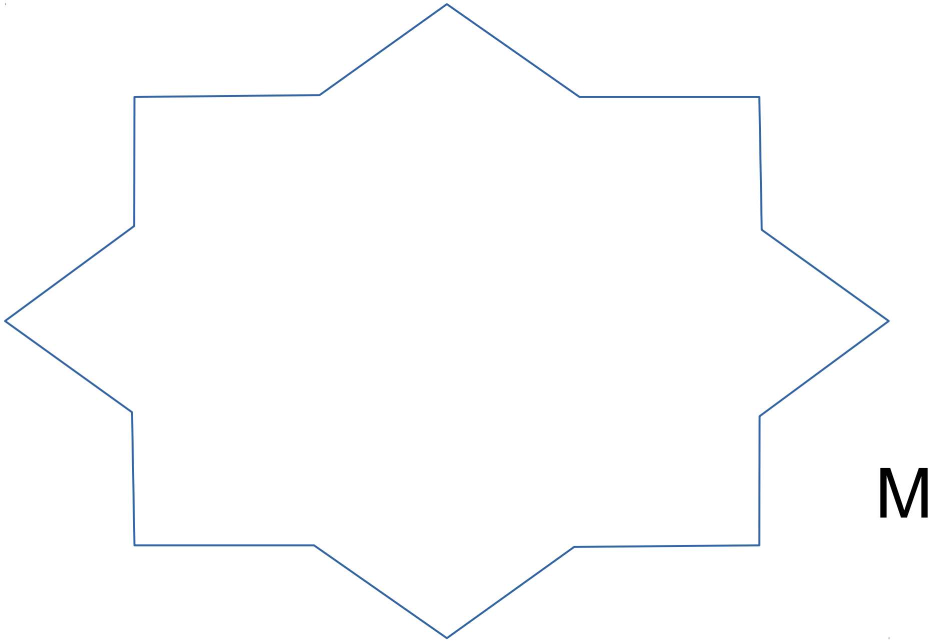
Gene Expression
RNA-Seq

Splicing
RNA-Seq

RNA-Seq

- Expression of genes is determined
- Is actually cDNA-Seq
- Can be done with/without reference
- Unbiased (No prior knowledge of genes needed)
- Splice variants can be detected
- New genes can be detected
- Case/control, survey

What can be done



Metagenomics
DNA-Seq

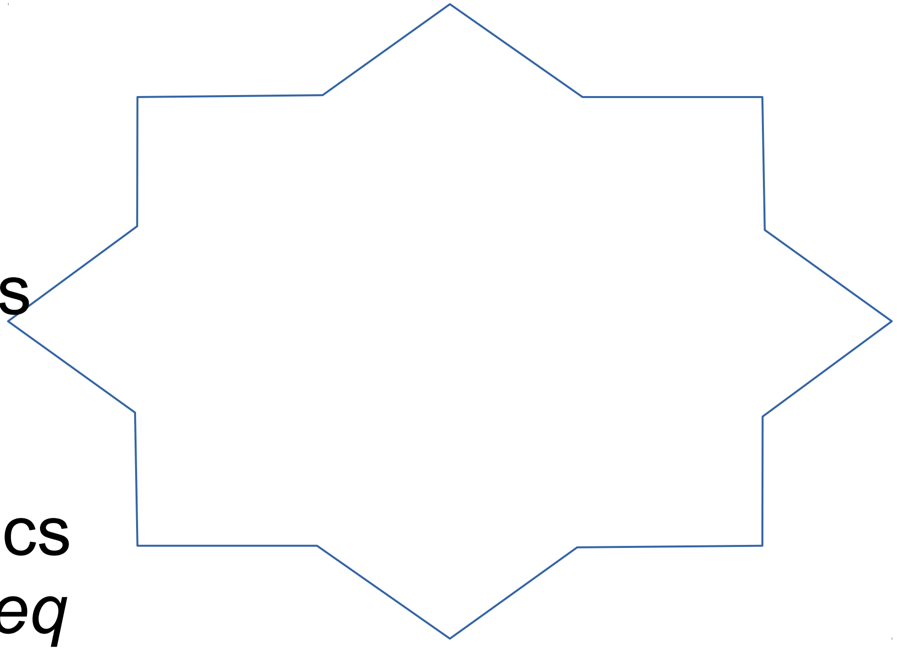
Metagenomics

- Culture free sequencing
- Snapshot of community of organisms living in the sample
- Typically used for microbiomes
- Biotechnological and medical applications

The coffee-machine bacteriome: biodiversity and colonisation of the wasted coffee tray leach

Cristina Vilanova[1], Alba Iglesias[1] & Manuel Porcar[1,2]

What can be done

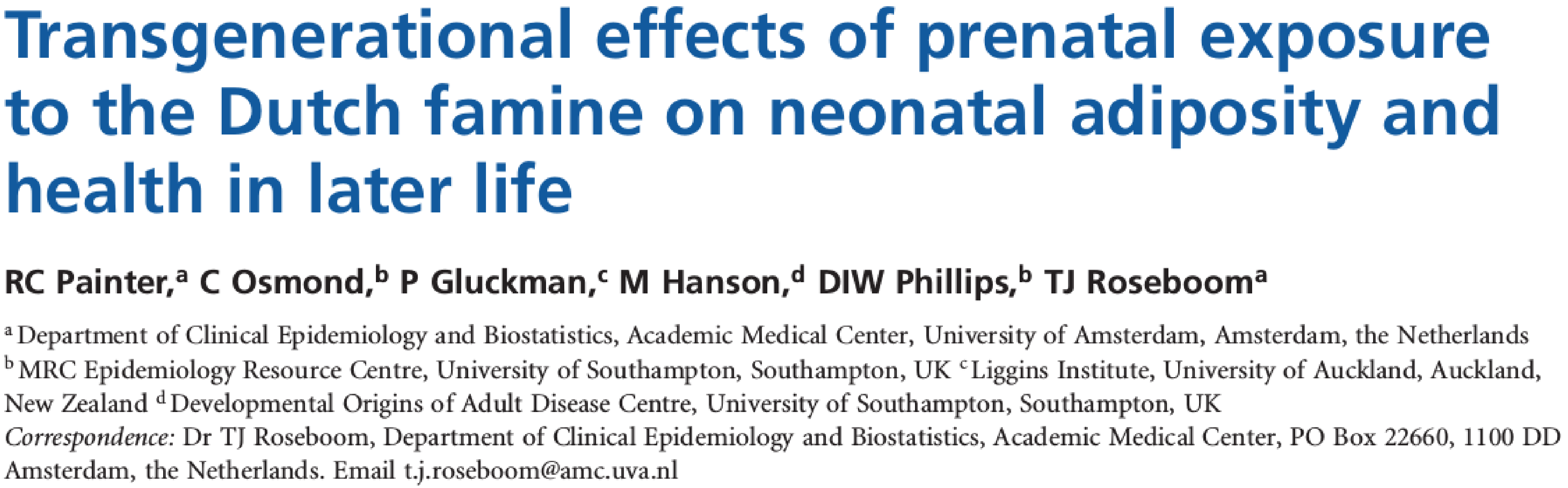


Translation Speeds
Ribo-Seq

Epigenetics
Methyl-Seq

Protein Binding
to DNA
Chip-Seq

Curious case of Dutch Famine



Transgenerational effects of prenatal exposure to the Dutch famine on neonatal adiposity and health in later life

RC Painter,[a] C Osmond,[b] P Gluckman,[c] M Hanson,[d] DIW Phillips,[b] TJ Roseboom[a]

[a] Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands [b] MRC Epidemiology Resource Centre, University of Southampton, Southampton, UK [c] Liggins Institute, University of Auckland, Auckland, New Zealand [d] Developmental Origins of Adult Disease Centre, University of Southampton, Southampton, UK

Correspondence: Dr TJ Roseboom, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, PO Box 22660, 1100 DD Amsterdam, the Netherlands. Email t.j.roseboom@amc.uva.nl

Famine affects development 2 generations after it was over

Problem

- Which aspect do you like the most?
- Your lab works in soil science. In one particular soil sample you find higher growth and plant yields. You suspect of something in addition to soil chemistry is at play. Which NGS technique will you use ?
- Weihenstephan-Kiwis are specially bred kiwi plants developed in Weihenstephan which can tolerate low temperatures in bavaria. But with a reduction in fruit size. You are interested in finding out what changes the breeders selected for. Which NGS technique will you use.

Problem

- Smoking has not yet associated with DNA damage in the germline. However studies are indicating poor child development in smoking mothers. Which molecular mechanism could you test as responsible by using NGS methodologies ?
- Your lab works on a insect which is a major crop pest causing massive crop damages each year. But not much is known about it. Which NGS methodology would fit best for a genomic analysis ?

Changes in Research from NGS

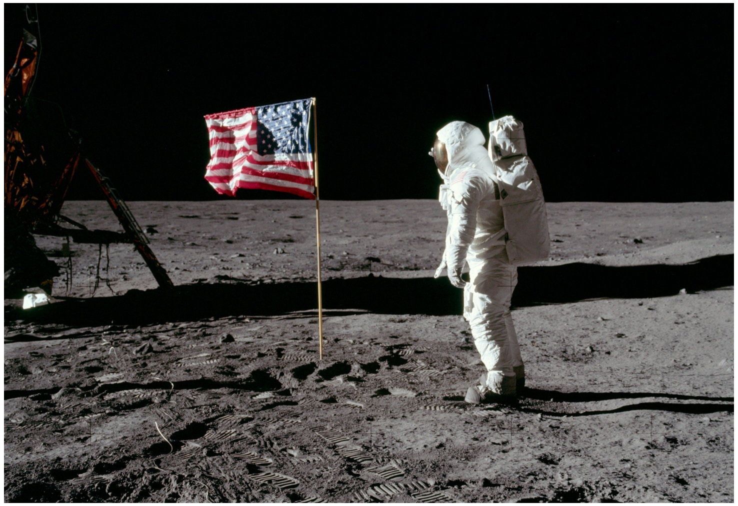
- Designing Studies
- Hypothesis testing
- Costs
- Systems level analysis
- Experiment Design

NGS

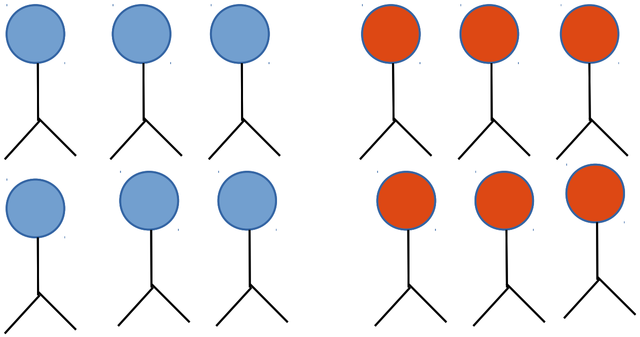
- Data processing
- Data Storage
- New Softwares
- Complex pipelines
- Computational Power
- Statistical testing
- NOISE

Combining Two Categories

Exploratory



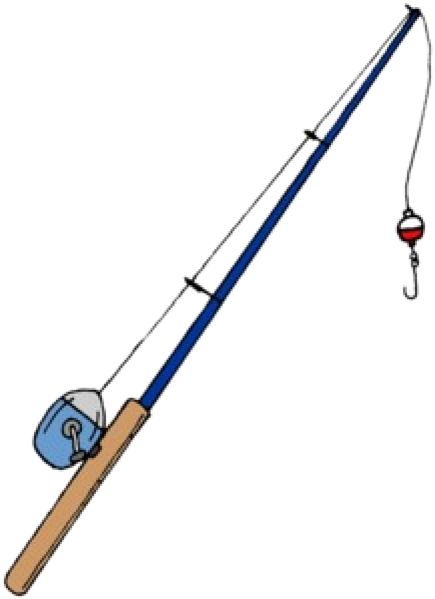
Hypothesis Driven



Healthy

Disease

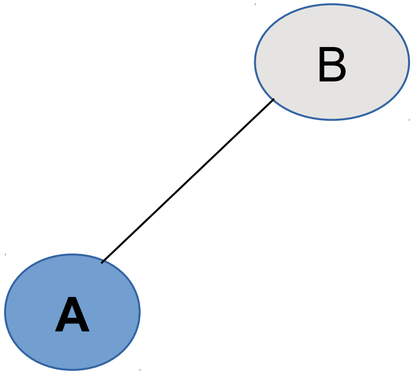
Combining categories



Is this gene responsible?

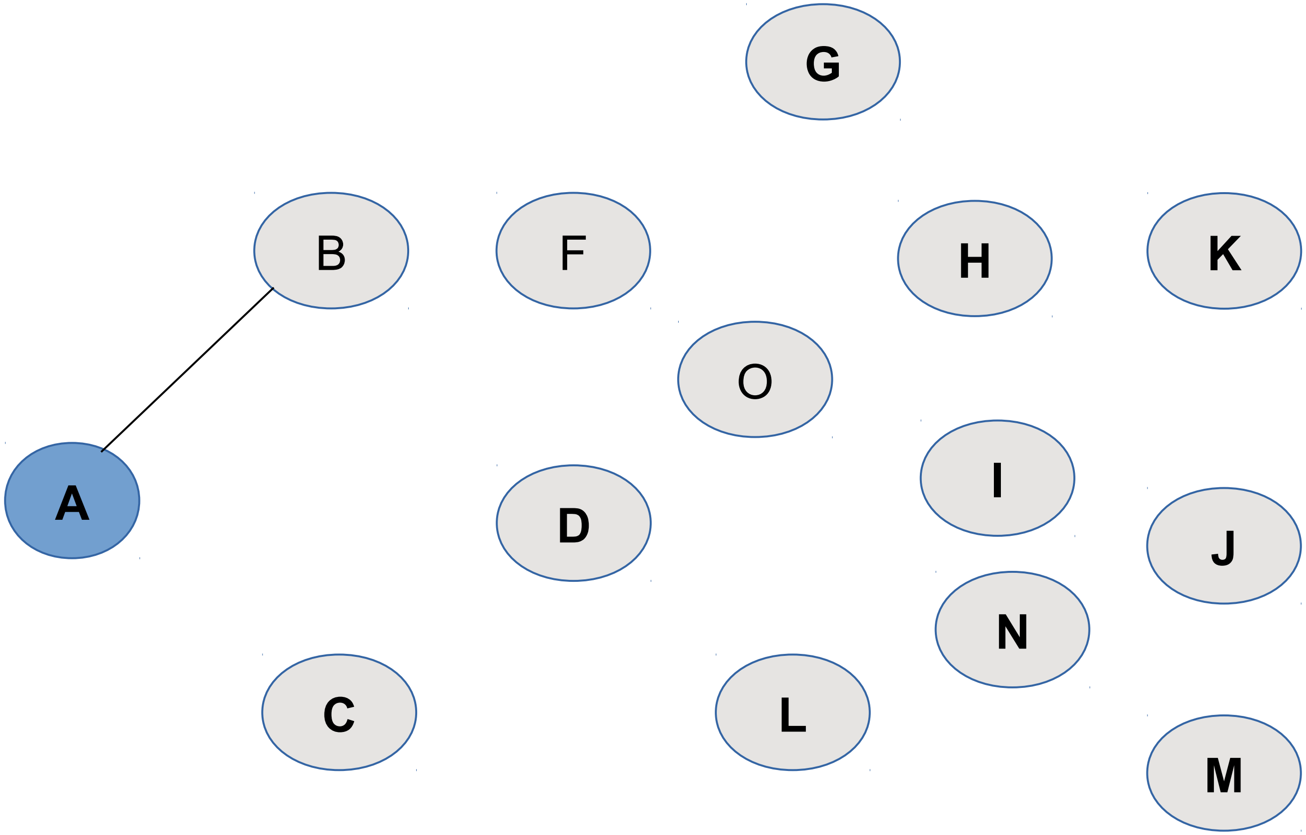
Which genes are responsible ?

Traditional Approach

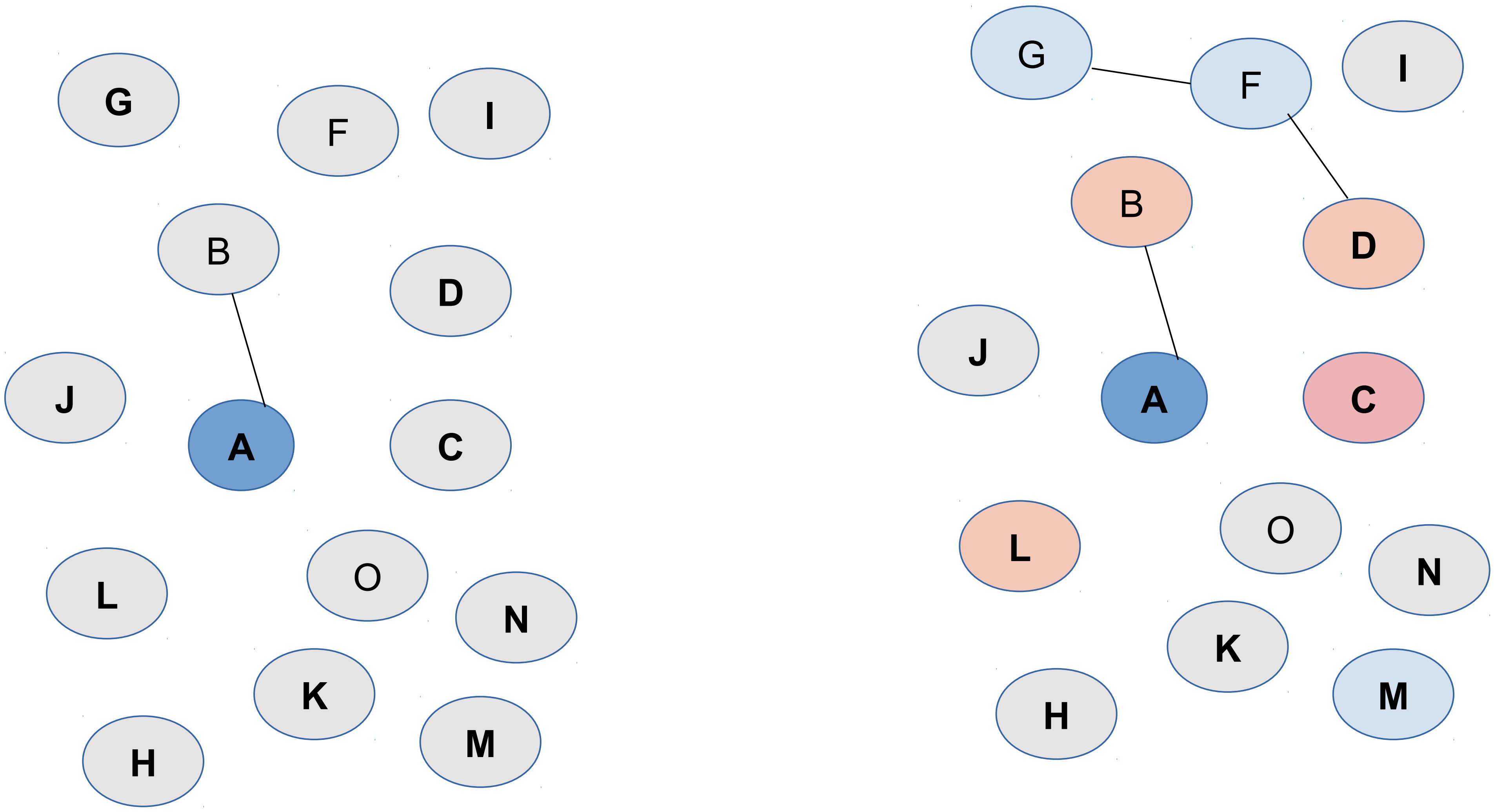


- Gene A is involved in Stress resistance
- Gene B interacts with gene A
- Gene B could be also involved in stress resistance ?
- Experimental validation
- Pathway A - B is responsible for stress resistance

Other Genes



Global Unbiased Screen



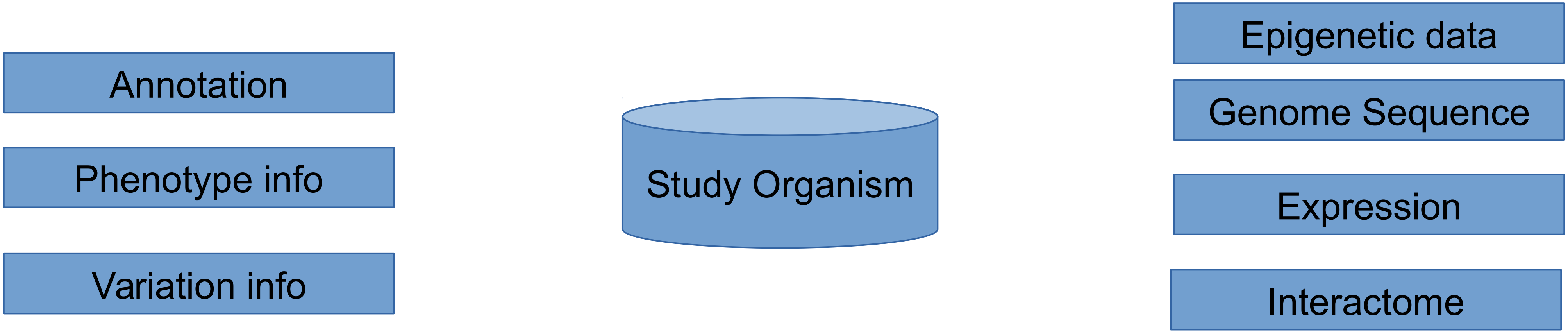
No Stress (Control)

Stress

Navigating the state of art

- What is already available
- Where is the organism in Tree of Life
- Is it's genome sequenced
- How well studied is it
- What information is available in related organisms
- Hypothesis development

Navigating the state of art



Annotation

Phenotype info

Variation info

Study Organism

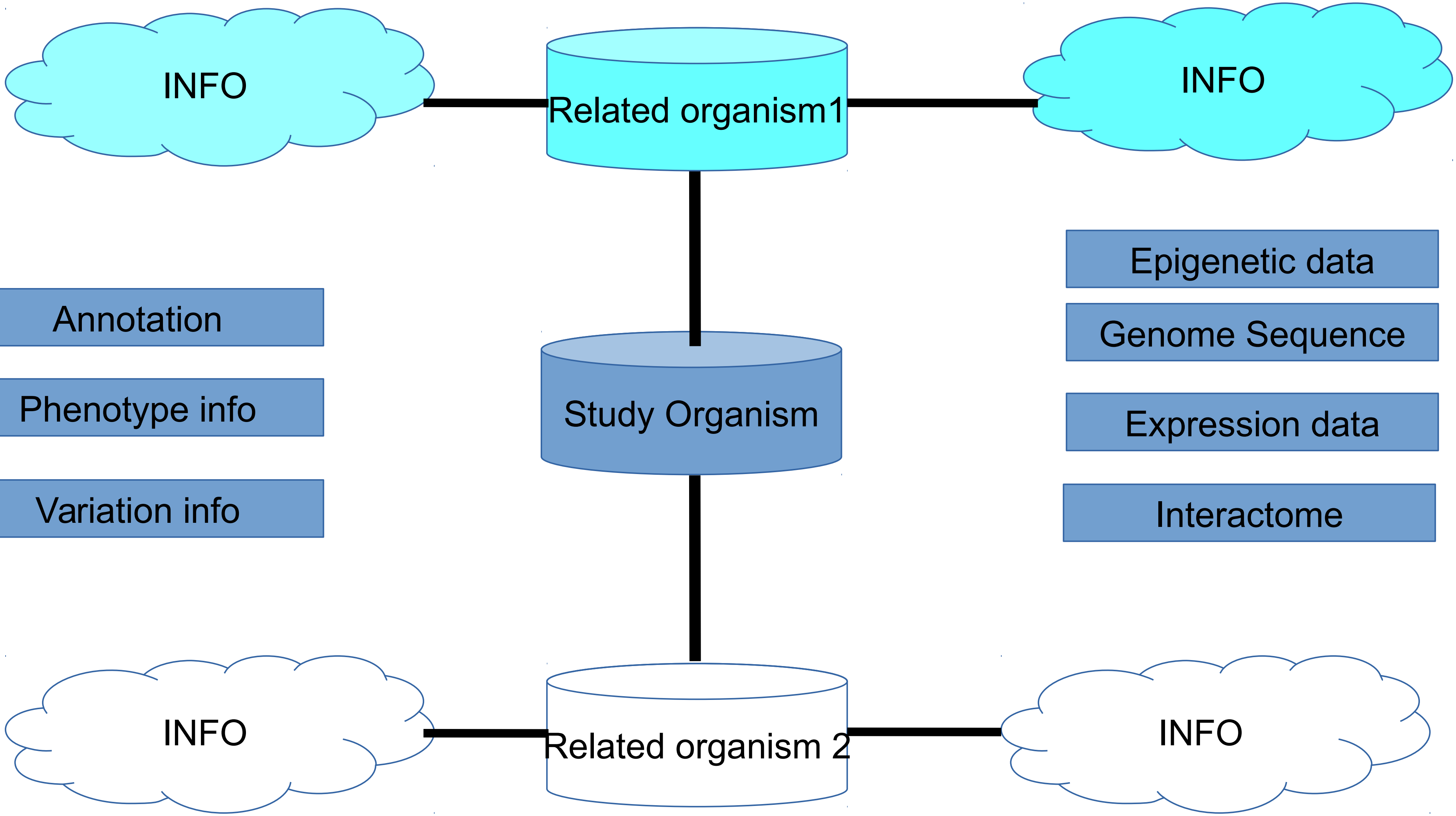
Epigenetic data

Genome Sequence

Expression

Interactome

Navigating the state of art



INFO

Related organism1

INFO

Annotation

Phenotype info

Variation info

Study Organism

Epigenetic data

Genome Sequence

Expression data

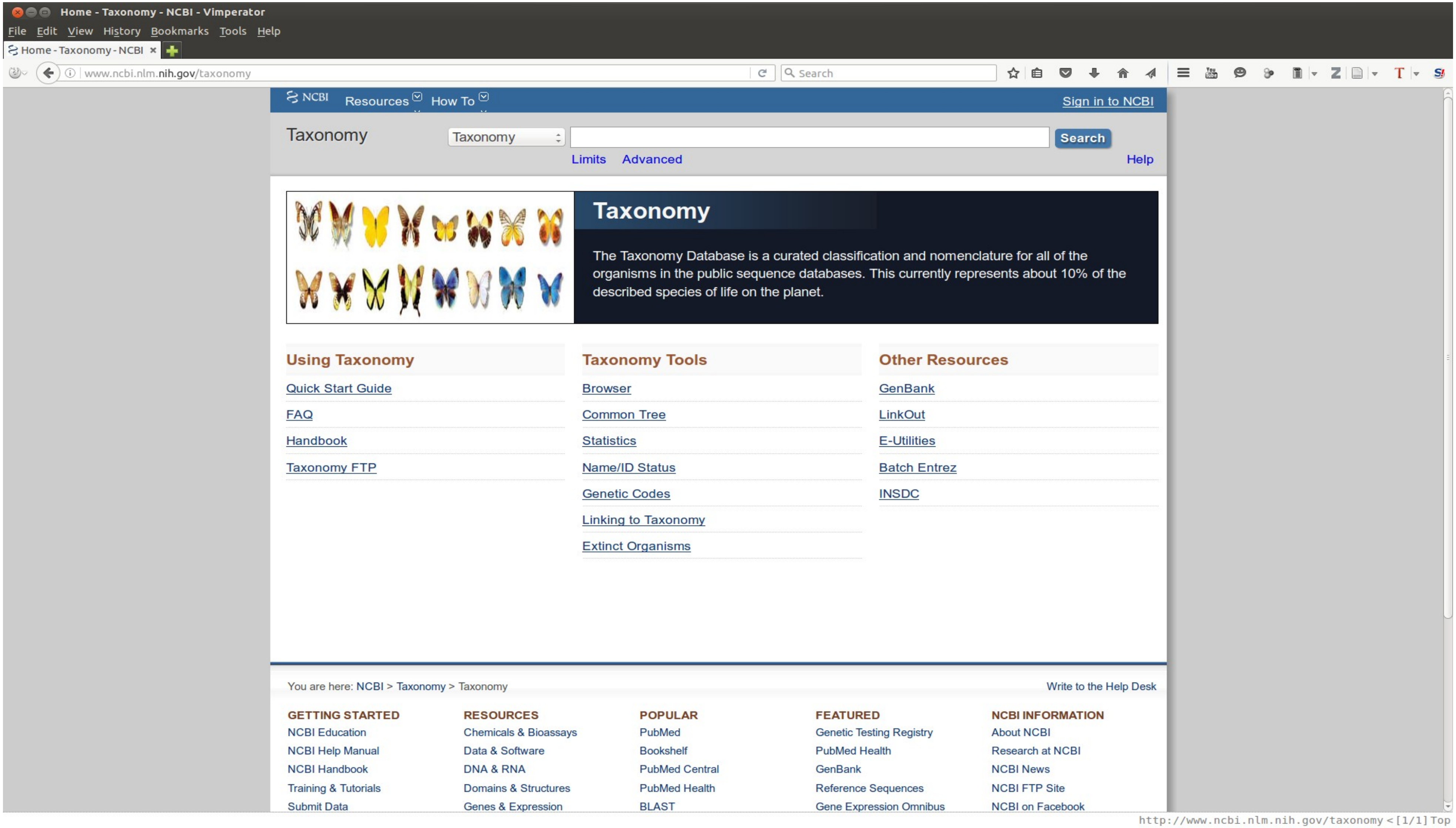
Interactome

INFO

Related organism 2

INFO

One stop shop: NCBI Taxonomy



Home - Taxonomy - NCBI - Vimperator
File Edit View History Bookmarks Tools Help
Home - Taxonomy - NCBI
www.ncbi.nlm.nih.gov/taxonomy
Search
NCBI Resources How To
Sign in to NCBI
Taxonomy
Taxonomy
Search
Limits Advanced
Help

Taxonomy

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.

Using Taxonomy

- Quick Start Guide
- FAQ
- Handbook
- Taxonomy FTP

Taxonomy Tools

- Browser
- Common Tree
- Statistics
- Name/ID Status
- Genetic Codes
- Linking to Taxonomy
- Extinct Organisms

Other Resources

- GenBank
- LinkOut
- E-Utilities
- Batch Entrez
- INSDC

You are here: NCBI > Taxonomy > Taxonomy

Write to the Help Desk

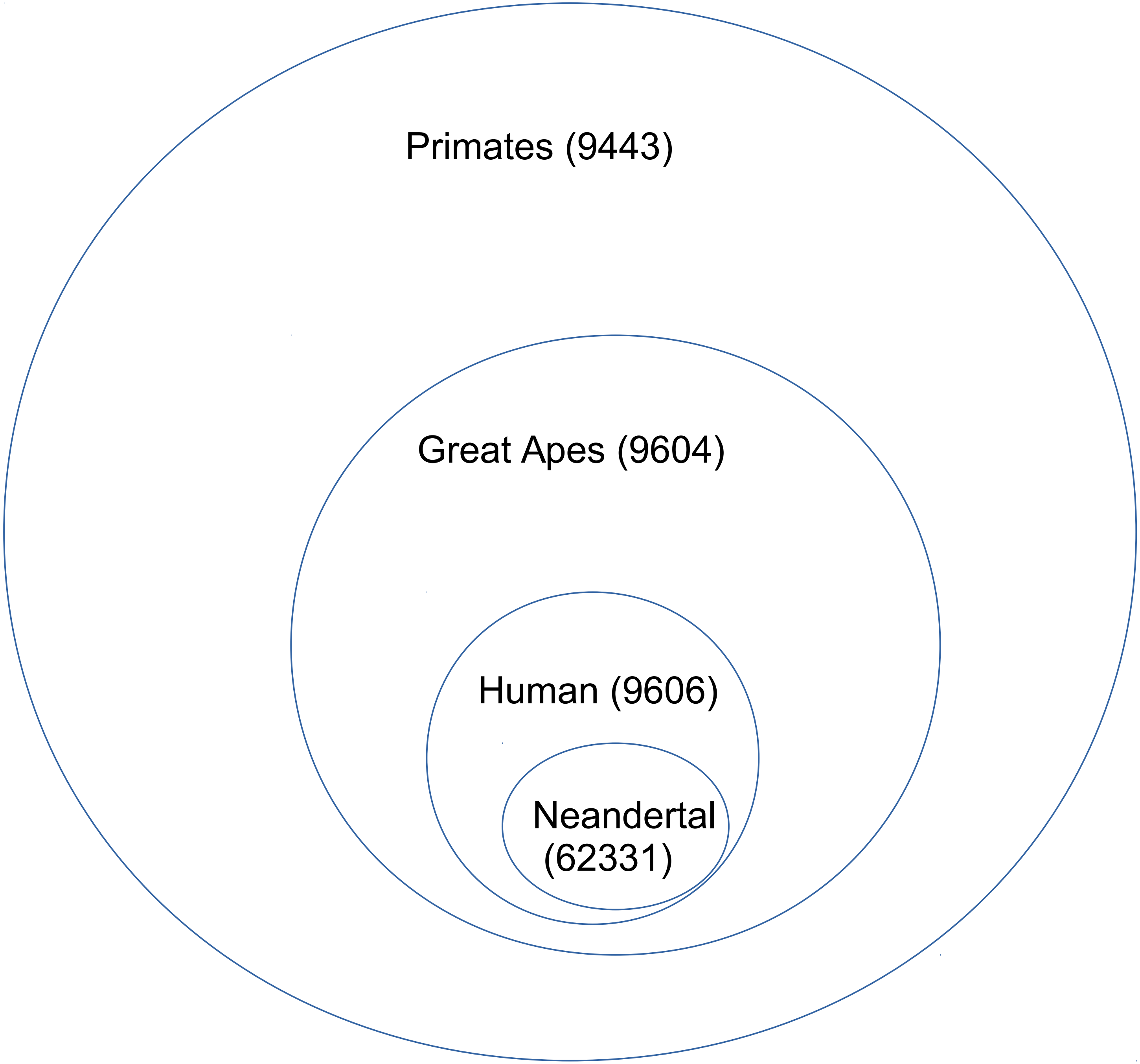
GETTING STARTED	RESOURCES	POPULAR	FEATURED	NCBI INFORMATION
NCBI Education	Chemicals & Bioassays	PubMed	Genetic Testing Registry	About NCBI
NCBI Help Manual	Data & Software	Bookshelf	PubMed Health	Research at NCBI
NCBI Handbook	DNA & RNA	PubMed Central	GenBank	NCBI News
Training & Tutorials	Domains & Structures	PubMed Health	Reference Sequences	NCBI FTP Site
Submit Data	Genes & Expression	BLAST	Gene Expression Omnibus	NCBI on Facebook

Why NCBI Taxonomy ?

Have I already been Scooped ?

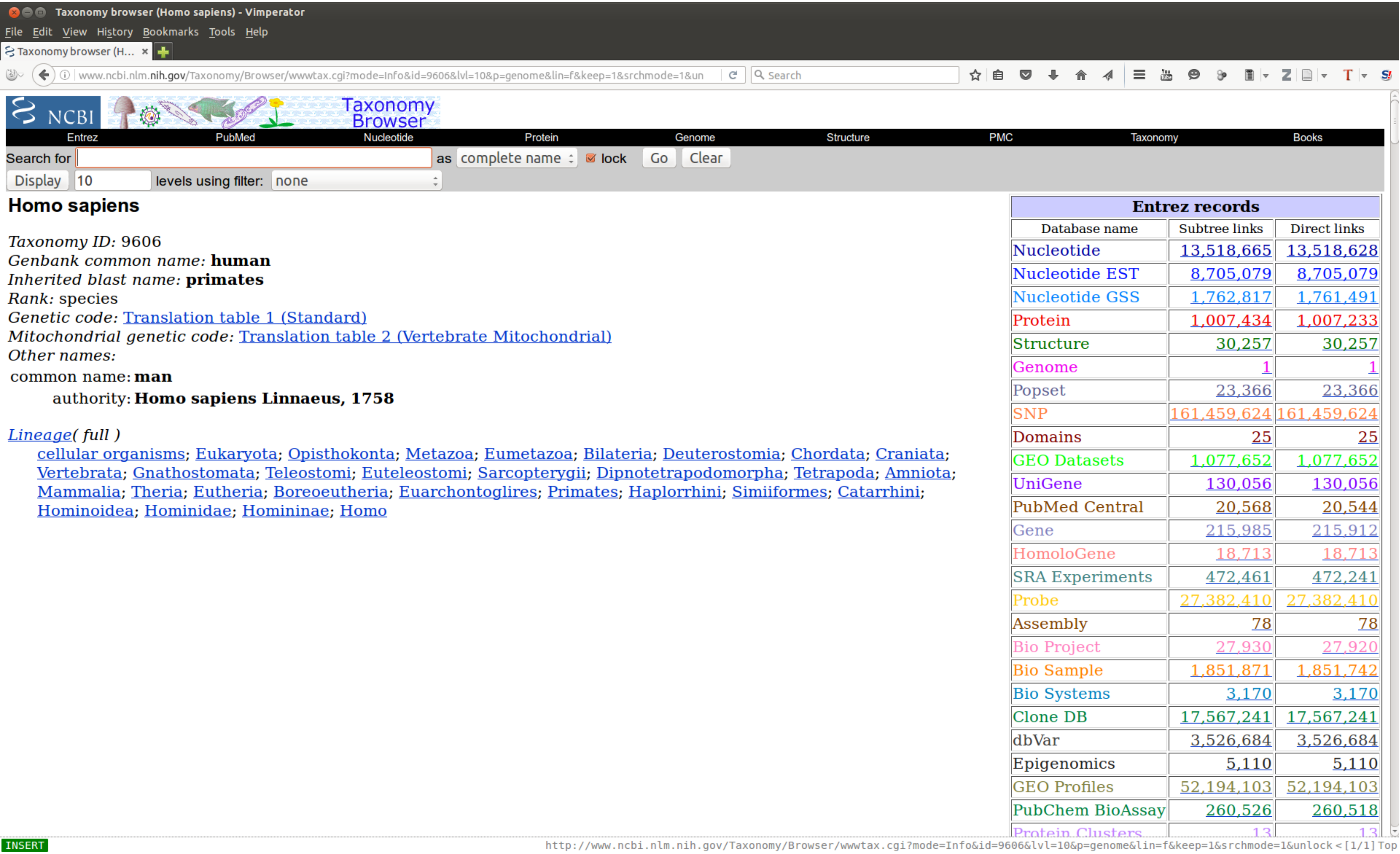
NCBI Taxonomy

- Every organism has a unique id (taxid)
- Internal nodes also have taxids
- Related organisms don't have related ids
- All related information is linked from other NCBI databases
- Taxids have a hierarchical structure


Primates (9443)
Great Apes (9604)
Human (9606)
Neandertal
(62331)

- **Hominidae** (great apes) *Click on organism name to get more information.*

  - **Homininae**
    - **Gorilla**
      - **Gorilla beringei** (eastern gorilla)
        - **Gorilla beringei beringei** (eastern mountain gorilla)
        - **Gorilla beringei graueri** (eastern lowland gorilla)
      - **Gorilla gorilla** (western gorilla)
        - **Gorilla gorilla diehli** (Cross River gorilla)
        - **Gorilla gorilla gorilla** (western lowland gorilla)
        - **Gorilla gorilla uellensis**
    - **Homo**
      - **Homo heidelbergensis** (Heidelberg man)
      - **Homo sapiens** (human)
        - **Homo sapiens neanderthalensis** (Neandertal)
        - **Homo sapiens ssp. Denisova** (Denisova hominin)
    - **Pan** (chimpanzees)
      - **Pan paniscus** (pygmy chimpanzee)
      - **Pan troglodytes** (chimpanzee)
        - **Pan troglodytes ellioti**
        - **Pan troglodytes schweinfurthii**
        - **Pan troglodytes troglodytes**
        - **Pan troglodytes vellerosus**
        - **Pan troglodytes verus**
        - **Pan troglodytes verus x troglodytes**
  - **Ponginae**
    - **Pongo**
      - **Pongo abelii** (Sumatran orangutan)
      - **Pongo abelii x pygmaeus**
      - **Pongo pygmaeus** (Bornean orangutan)
        - **Pongo pygmaeus pygmaeus**
      - **Pongo sp.**


Taxonomy browser (Homo sapiens) - Vimperator
File Edit View History Bookmarks Tools Help
Taxonomy browser (H...
www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=9606&lvl=10&p=genome&lin=f&keep=1&srchmode=1&un
Search
NCBI
Taxonomy Browser
Entrez
PubMed
Nucleotide
Protein
Genome
Structure
PMC
Taxonomy
Books
Go
Clear
Display
INSERT

Search for [ ] as complete name ☑ lock

Display 10 levels using filter: none

Homo sapiens

Taxonomy ID: 9606
Genbank common name: **human**
Inherited blast name: **primates**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
Other names:
common name: **man**
authority: **Homo sapiens Linnaeus, 1758**

Lineage(*full*)
cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Homo

Entrez records

Database name	Subtree links	Direct links
Nucleotide	13,518,665	13,518,628
Nucleotide EST	8,705,079	8,705,079
Nucleotide GSS	1,762,817	1,761,491
Protein	1,007,434	1,007,233
Structure	30,257	30,257
Genome	1	1
Popset	23,366	23,366
SNP	161,459,624	161,459,624
Domains	25	25
GEO Datasets	1,077,652	1,077,652
UniGene	130,056	130,056
PubMed Central	20,568	20,544
Gene	215,985	215,912
HomoloGene	18,713	18,713
SRA Experiments	472,461	472,241
Probe	27,382,410	27,382,410
Assembly	78	78
Bio Project	27,930	27,920
Bio Sample	1,851,871	1,851,742
Bio Systems	3,170	3,170
Clone DB	17,567,241	17,567,241
dbVar	3,526,684	3,526,684
Epigenomics	5,110	5,110
GEO Profiles	52,194,103	52,194,103
PubChem BioAssay	260,526	260,518
Protein Clusters	13	13

http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=9606&lvl=10&p=genome&lin=f&keep=1&srchmode=1&unlock < [1/1] Top

Issues

- Redundant information
- Incomplete information
- Organalle genomes are counted as sequenced genomes
- Resequenced genomes are not counted
- Only open data available
- Not an authoritative phylogenetic resource

Problem

- Get the taxid of your favorite organism
- Is it's genome sequenced ?
- What is the nearest species available with genome sequence ?
- How many expression datasets are available for it ?

Problem

Candida albicans is a well known human pathogen. It belongs to genus candida. Your advisor wants to do a comparative and gene expression analysis with other candida genus samples. Which other candida species you would select ? He also asks for a report with information about chosen samples with pictures and information on where you can order these these samples for lab analysis. How would you use NCBI taxonomy for this ?

Next Lecture

- Navigating NGS Data Repositories
- Downloading NGS data
- Understanding File formats
- Resequencing analysis
- Mapping or Read Alignment