Course Introduction

Introduction to BI229F

Autumn 2024

Course structure

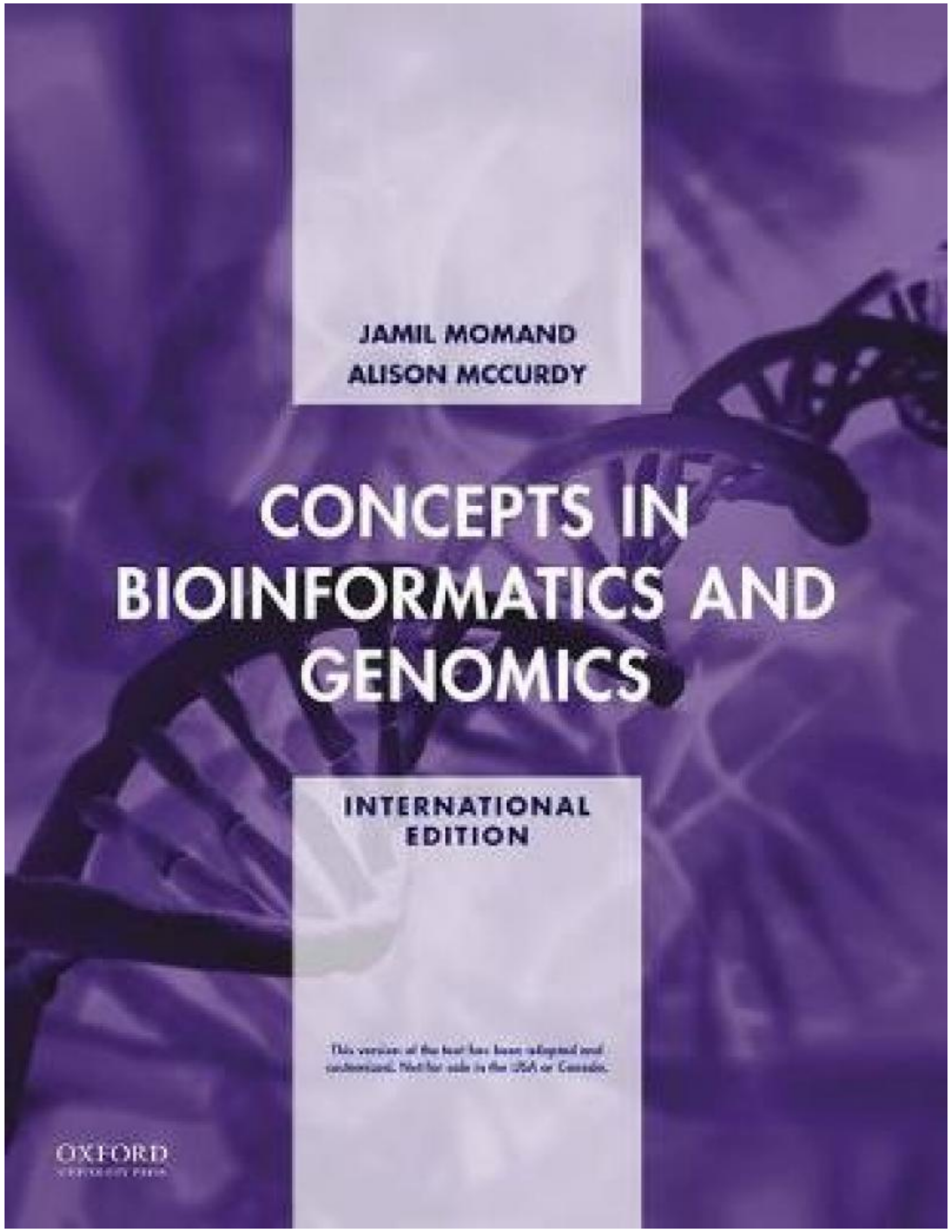
- A weekly mixture of lectures and practicals
  - Tuesday lecture
  - Thursday practical
- Extended practical sessions
  - 3-4 hours
  - prelim. dates set
- Mandatory project
  - Coding or data analysis
  - Designed with your input

Introductory course

- Introduction to:
  - Databases
  - Tools
  - Resources
- Details:
  - some Algorithms (eg. sequence alignment)
  - some Statistical methods
- Practical skills (writing code)
  - Data mangling
  - Statistical analysis
  - Project management

Content and schedule may be adjusted as we progress
Suggestions for changes are welcomed

Reference material


JAMIL MOMAND
ALISON MCCURDY
CONCEPTS IN BIOINFORMATICS AND GENOMICS
INTERNATIONAL EDITION
This version of the text has been adapted and customized. Not for sale in the USA or Canada.
OXFORD

There is a book

Good book,
but not that necessary:
almost everything can
be found on the web

Bioinformatics? Genomics?

What are these things and why do we care?

- Bioinformatics
  - Development of computational methods for biological data analysis
  - Computational analysis of biological data
- Genomics
  - The study of genomes
    - Structure
    - Evolution
    - ...
  - The use of genomics scale data
    - Genome wide gene expression data (transcriptomics)
    - Population genomics (eg. identification of selected loci)
    - ...

We care because data is cheap

Lots of data

Advances in technology allow us to define complete genome sequences and to measure the activities of 1000s of genes without spending large amounts of money or time

This makes genomics useful, and bioinformatics essential for almost all biological questions

How much data?

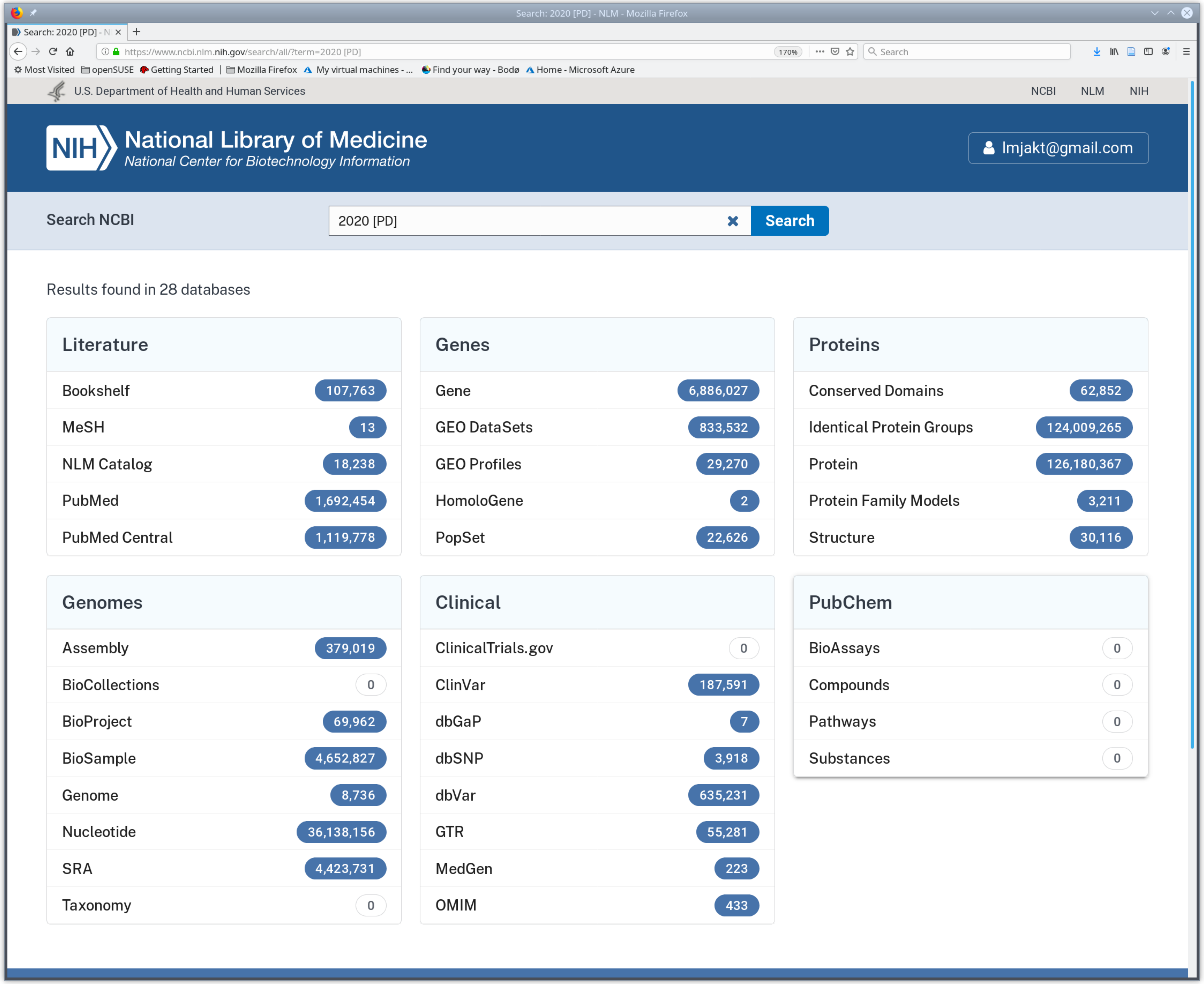
http://www.ncbi.nlm.nih.gov/

Has the answer

How much data?

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

Has the answer


Search: 2020 [PD] - NLM - Mozilla Firefox
Search: 2020 [PD] - N
https://www.ncbi.nlm.nih.gov/search/all/?term=2020 [PD]
170%
Search
Most Visited
openSUSE
Getting Started
Mozilla Firefox
My virtual machines - ...
Find your way - Bodø
Home - Microsoft Azure
U.S. Department of Health and Human Services
NCBI
NLM
NIH
NIH
National Library of Medicine
National Center for Biotechnology Information
lmjakt@gmail.com
Search

Search NCBI 2020 [PD]

Results found in 28 databases

Literature	
Bookshelf	107,763
MeSH	13
NLM Catalog	18,238
PubMed	1,692,454
PubMed Central	1,119,778

Genes	
Gene	6,886,027
GEO DataSets	833,532
GEO Profiles	29,270
HomoloGene	2
PopSet	22,626

Proteins	
Conserved Domains	62,852
Identical Protein Groups	124,009,265
Protein	126,180,367
Protein Family Models	3,211
Structure	30,116

Genomes	
Assembly	379,019
BioCollections	0
BioProject	69,962
BioSample	4,652,827
Genome	8,736
Nucleotide	36,138,156
SRA	4,423,731
Taxonomy	0

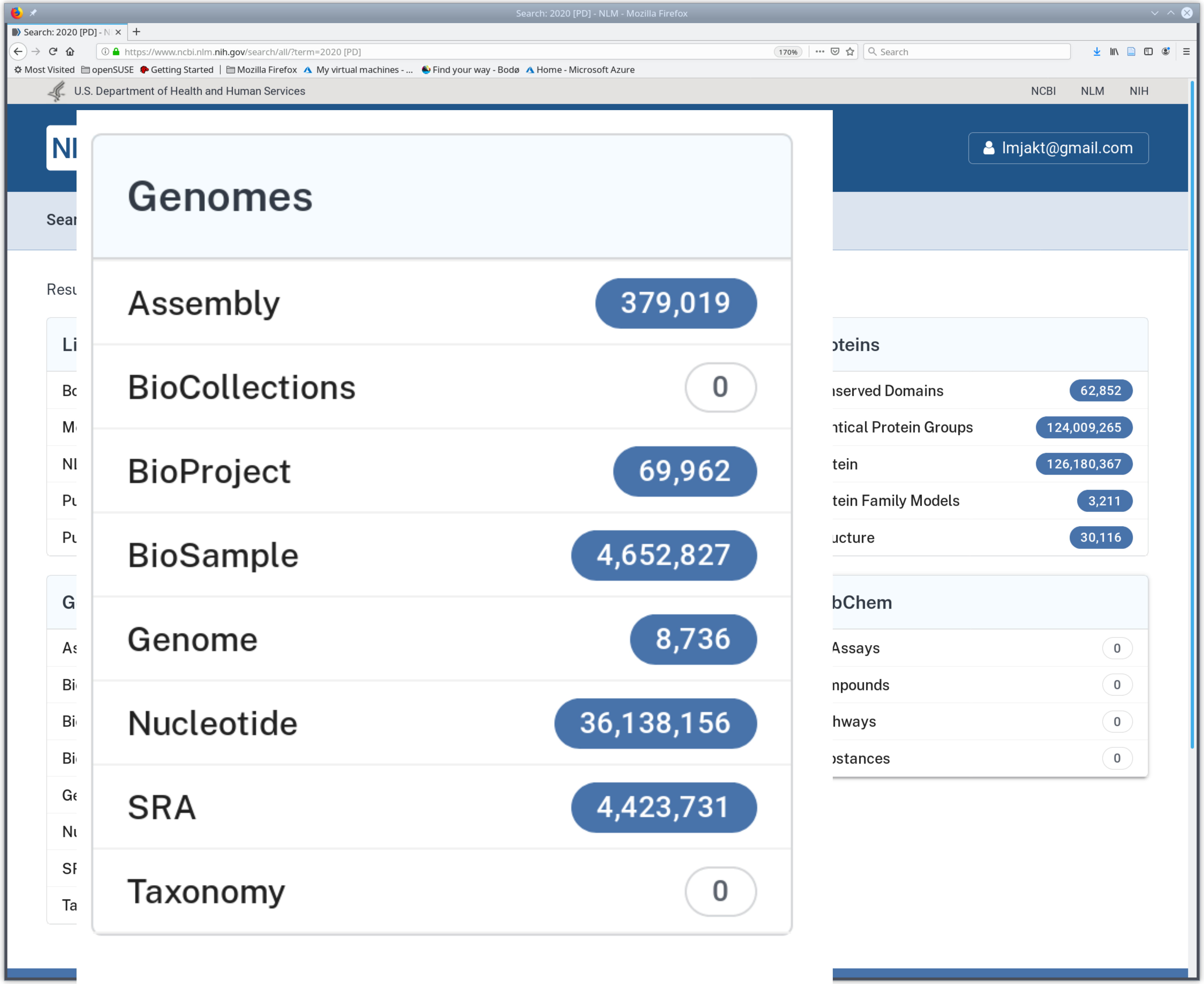
Clinical	
ClinicalTrials.gov	0
ClinVar	187,591
dbGaP	7
dbSNP	3,918
dbVar	635,231
GTR	55,281
MedGen	223
OMIM	433

PubChem	
BioAssays	0
Compounds	0
Pathways	0
Substances	0

How much data?

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

Has the answer


Search: 2020 [PD] - NLM - Mozilla Firefox
https://www.ncbi.nlm.nih.gov/search/all/?term=2020 [PD]
U.S. Department of Health and Human Services
NCBI
NLM
NIH
lmjakt@gmail.com

Genomes

Assembly	379,019
BioCollections	0
BioProject	69,962
BioSample	4,652,827
Genome	8,736
Nucleotide	36,138,156
SRA	4,423,731
Taxonomy	0

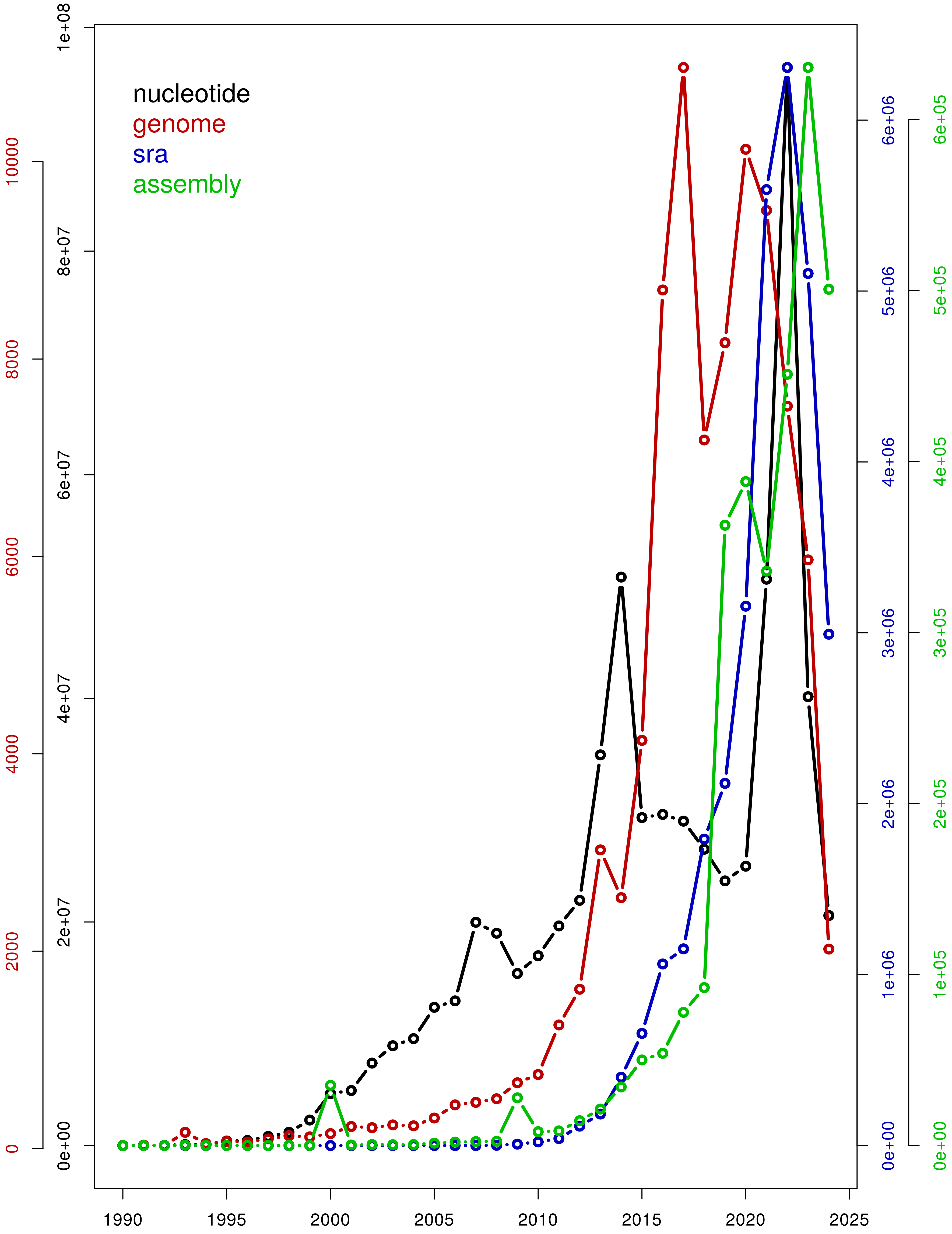
oteins

nserved Domains	62,852
ntical Protein Groups	124,009,265
tein	126,180,367
tein Family Models	3,211
ucture	30,116

bChem

Assays	0
npounds	0
hways	0
ostances	0

Change over time



A bit of R:

```
## prepare some publication trends data:
py.range <- 1990:2024
ncbi.url <- urlInfo()
py.counts <- cbind('year'=py.range, 'nucleotide'=NA)
for(i in 1:length(py.range)){
    tmp <- search.ncbi.py(ncbi.url, py.range[i], db='nucleotide', "")
    py.counts[i,2] <- extract.count( tmp[[1]] )
}

urlInfo  <- function(){
    list(base="https://eutils.ncbi.nlm.nih.gov/entrez/eutils/",
         search_suffix = "esearch.fcgi?",
         summary_suffix = "esummary.fcgi?",
         data_suffix = "efetch.fcgi?")
}

## other functions to take this list as a an argument

## terms is a character vector that will be combined
## into a single string
search.ncbi  <- function(url, db="pubmed", terms, type="id", max=0){
    query=paste(url$base, url$search_suffix, "db=", db, "&", sep="")
    query=paste( query, "term=", paste(terms, collapse="+"),
                "&rettype=", type, sep="" )
    if(max && max > 0)
        query = paste(query, "&retmax=", max, sep="")
    readLines(query)
}

search.ncbi.py  <- function(url, years, db="pubmed", terms, type="id", max=0){
    terms.list  <- paste( paste(terms, collapse="+"),
                          paste(years, "[pdat]", sep=""), sep="+AND+" )
    lapply(terms.list, function(x){
        search.ncbi(url, db=db, terms=x, type=type, max=max )
    })
}

extract.ids  <- function(lines){
    gsub("[^0-9]", "",  grep("<Id>([0-9]+)</Id>$", lines, value=TRUE))
}

extract.count  <- function(lines){
    as.numeric( sub(".+?<Count>([0-9]+)</Count>.+", "\\1",
                    grep("<Count>[0-9]+</Count>", tmp[[1]], value=TRUE))[1] )
}
```

Course topics

Molecular Biology	Information storage (DNA), transmission (RNA), functional molecules (RNA & proteins)
Practical bioinformatics	Resources and tools for looking at sequences and other data
Algorithms	Sequence alignment, database search
Computers	Hardware, operating systems, networks, applications
Big(ish) data analysis	Visualisation and analysis of large data sets
Statistics	Derivation of p-values and multiple testing
Doing the above	i.e. Programming
Genomes & transcriptomes	Structure and other properties

With adjustments along the way if necessary

Course objectives (1)

- Data handling (genomics / bioinformatics data types)
  - General data formats (eg. text / binary)
  - Specific data formats (sequences, alignments, etc)
  - Import / parsing of data
- Selected bioinformatics algorithms
- Sources of information (databases)
- Data visualisation
- Data analysis (statistical tests)

Course objectives (2)

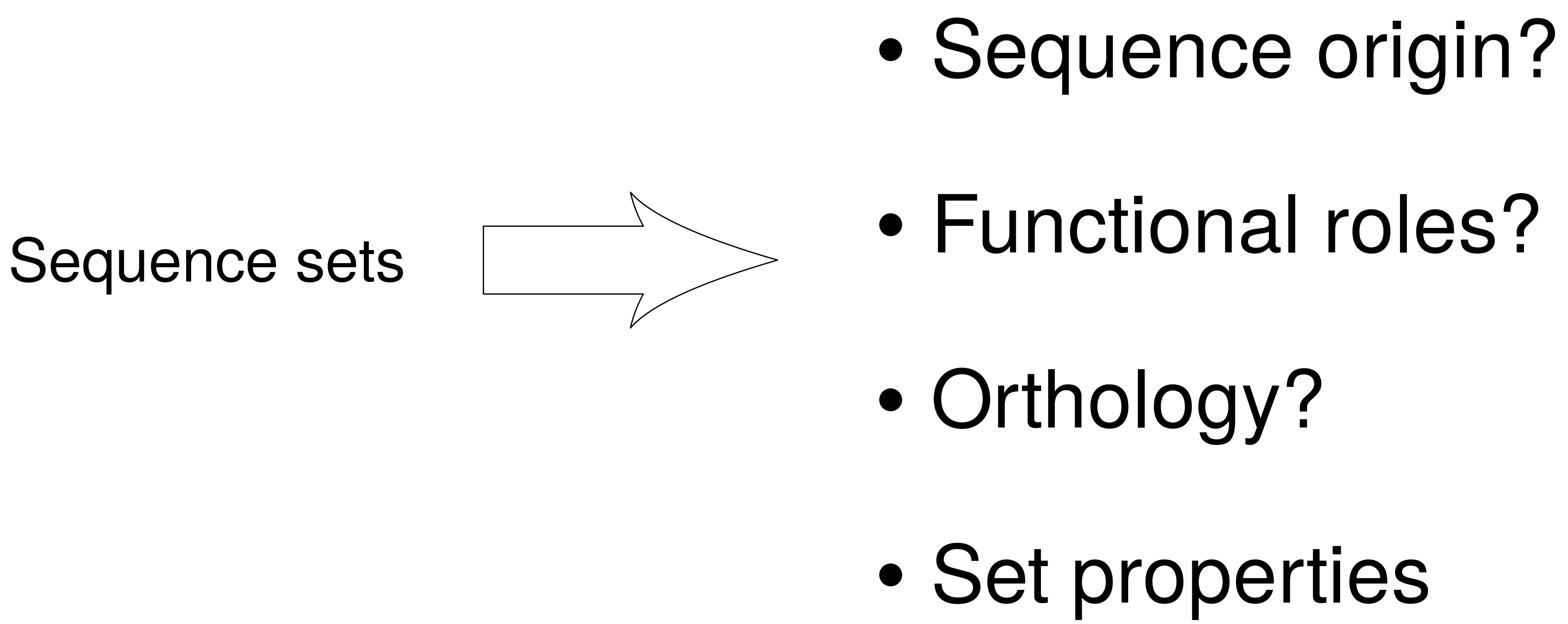
- Computational skills
  - File systems
  - General purpose programming (R!?*^%$)
  - Data extraction and munging
  - Method implementation

Learnt from programming exercises

Course objectives (3)

- Genomes
  - Anatomy (content)
  - Genes
  - Evolution of
  - Types (nuclear, organellular)
- Transcriptomes
  - Structure of (distribution)
  - Differential expression

A practical approach



Sequence sets

- Sequence origin?
- Functional roles?
- Orthology?
- Set properties

Tentative timetable

week	month	date	day	start	end	Topic
34	8	19	Mon	12:15	14:00	Lecture: course introduction
	8	23	Fri	10:15	12:00	Lecture: Essential molecular biology
35	8	26	Mon	12:15	14:00	Lecture: Introduction to R
	8	27	Tue	14:15	16:00	Practical: R exercises
36	9	2	Mon	12:15	14:00	Lecture: Pairwise sequence alignment
	9	3	Tue	14:15	16:00	Practical: Implementing pairwise sequence alignment in R
37	9	9	Mon	12:15	14:00	Lecture: Project description
	9	10	Tue	14:15	16:00	Practical and group work: Choosing your projects
	9	13	Fri	12:15	15:00	project
38	9	17	Tue	14:15	16:00	Lecture: Multiple sequence alignment
	9	19	Thu	12:15	14:00	Practical: Doing multiple sequence alignment in R
39	9	23	Mon	12:15	14:00	Lecture: Genome structure
	9	24	Tue	14:15	16:00	Practical: Exploring genome structures in R
	9	27	Fri	12:15	15:00	project
40	9	30	Mon	12:15	14:00	Lecture: Programs and scripts
	10	1	Tue	14:15	16:00	Practical: General programming challenges in R
41	10	8	Tue	14:15	16:00	Lecture: Biological databases and online tools
	10	9	Wed	12:15	14:00	Practical: Identifying sequences
	10	11	Fri	12:15	15:00	project
42	10	14	Mon	12:15	14:00	Lecture: Random numbers to p-values
	10	15	Tue	14:15	16:00	Practical: Making statistical tests
44	10	28	Mon	12:15	14:00	Lecture: Statistics for large data sets
	10	29	Tue	14:15	16:00	Practical: Analysing large-ish data
45	11	4	Mon	12:15	14:00	Lecture: Sequencing technologies
	11	5	Tue	14:15	16:00	TBA

Some questions for you

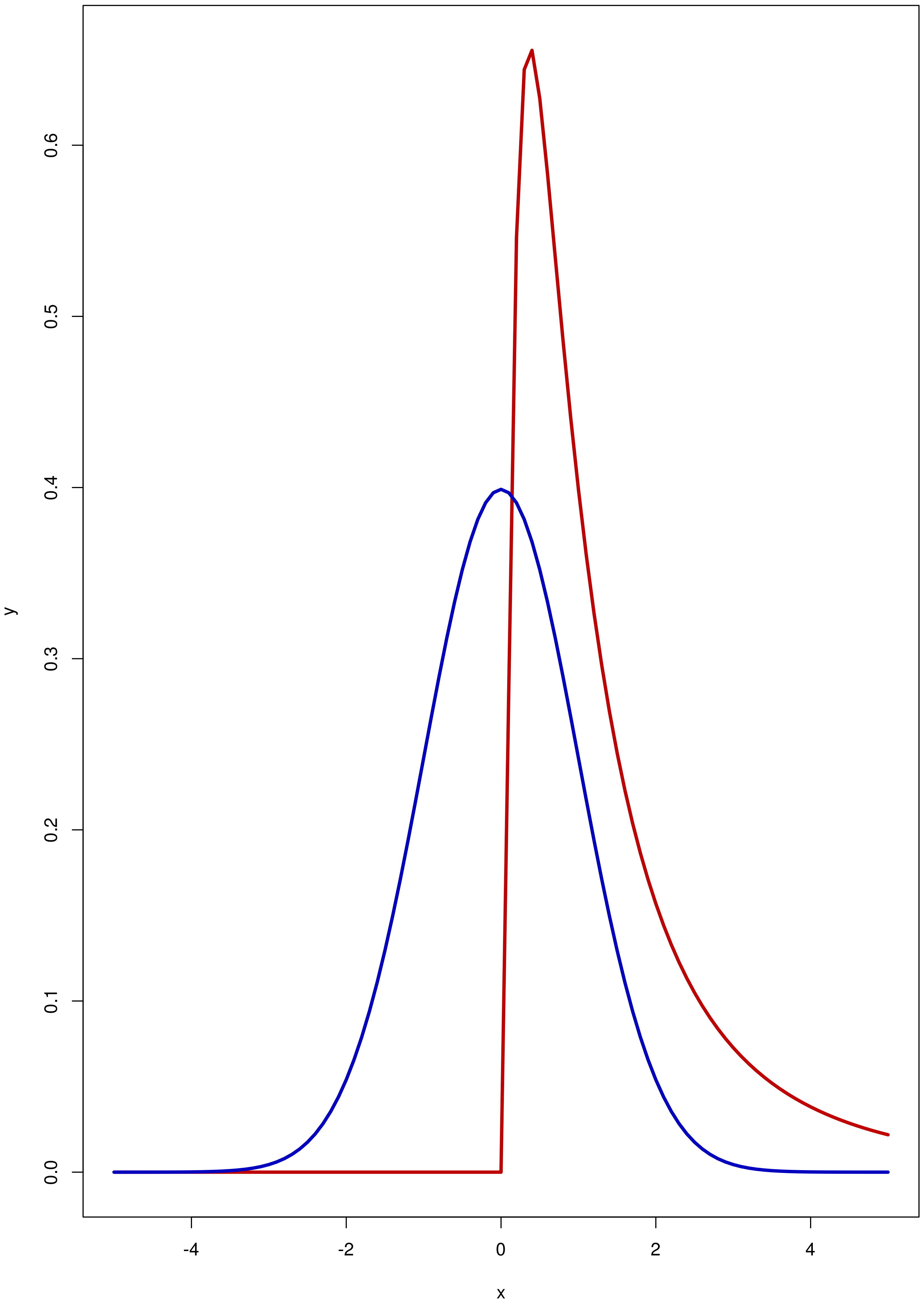
Molecular biology questions

- What is the most common type of RNA molecule?
- How many different ways can you translate
  ATGGTACTATAA ?
- And how about
  AUGGUACUAUAA ?
- What is the difference between RNA and DNA?
- What is a gene?
- Name an important consequence of double-strandedness?
- How many different codons are there?
- What are histones?
- What is gene expression?

Bioinformatics experience

- What is global and local alignment?
- How many of you have used BLAST?
- What is the problem with multiple sequence alignment?
- How do we estimate gene expression from RNA sequencing data?
- Are you familiar with: fasta, fastq, sam and bam files?
- Why is the following funny?
    - There are 10 types of people in the world
        - Those who understand binary
        - And those who don't
- How many of you have written shell scripts?

Distributions



What are these?

What kind of processes give rise to them?

R

```r
for(i in 1:10){
   cat(paste(i, i^2, sep='\t'), '\n')
}
```

```r
f1 <- function(x){
   sum(x) / length(x)
}
```

```r
f2 <- function(x){
   sort(x)[ length(x) / 2 ]
}
```

What do these do?

Main questions

- Why have you selected this course?
- What do you hope to get out of it?
- Do you care about the 'Learning Outcome Description'?