Lecture 1 - An Introduction to Bioinformatics

Ian Simpson

ian.simpson@ed.ac.uk

Learning Outcomes

By the end of our course you should be able to:

- Communicate between biological and computational domains to facilitate effective inter-disciplinary working.
- Use and/or implement Bioinformatics tools, services and software in practical research scenarios.
- Have sufficient background knowledge, skills and understanding to discover and apply additional bioinformatics techniques.

Foundations	Skills	Application
Bioinformatics Algorithms	Python Scripting	Case Studies
Biological Concepts	Using Web Tools & Services	Coursework Mini-projects
Data Types & Sources	Building Analytic Workflows	Inter-disciplinary communication in practice

ian.simpson@ed.ac.uk

Course Organisation

Teaching

Lectures
Weeks 1-10 on Friday 10:00-10:55
Lecture Theatre 1 Appleton Tower
Also streamed live and recorded
Slides released >24h prior to lecture

Computing Lab
Weeks 2, 4, 6, and 8 on Wednesday 14:00-15:55
Appleton Tower Lab 6.06

Reading List (called Resource List with all reading material)

Materials
Everything available on courseware site:-
https://opencourse.inf.ed.ac.uk/bio1

Assessment

- 2x Coursework Assignments, 50% each
- Submission via Gradescope
- CW1 released 06/10/23, submit 27/10/23
- CW2 released 30/10/23, submit 17/11/23
- dedicated discussion channel on Piazza

Optional Mini-Course

Introduction to Python

- Optional catch-up mini-course for those who don't have prior experience of scripting and/or Python
- practice computational notebooks
- Small challenge coding puzzles to aid
- reading list
- dedicated discussion channel on Piazza

Communication

Discussion Board on Piazza and Course Announcements

ian.simpson@ed.ac.uk

Course Topics

Week	Labs	CW	Date Commencing	Week Topic Title
1			18th September	Introduction to Bioinformatics
2	●		25th September	Pairwise Sequence Alignment
3*		6th	2nd October	Basic Local Alignment and Search Tool (BLAST)
4	●		9th October	Multiple Sequence Alignment
5			16th October	Exploring Biological Databases
6	●	27th	23rd October	Working with Biological Databases
7		30th	30th November	Ontologies & Functional Enrichment Analysis
8	●		6th November	Biological Network Analysis
9		17th	13th November	Introduction to Next Generation Sequencing
10			20th November	Course Wrap-Up, Q&A

*NB pre-recorded lecture

ian.simpson@ed.ac.uk

Objectives of the Coursework

Aim

- to apply methods
- to develop research skills
- gain experience using bioinformatics sources and literature
- learn scientific writing and reporting

Task

- questions to rehearse key concepts
- self-directed research
- strong focus on precise description of methods and sources
- experimentation
- analytical, showing insight and evidence of evidence gathering to support conclusions

ian.simpson@ed.ac.uk

The Extended Common Marking Scheme

https://web.inf.ed.ac.uk/infweb/student-services/ito/students/common-marking-scheme

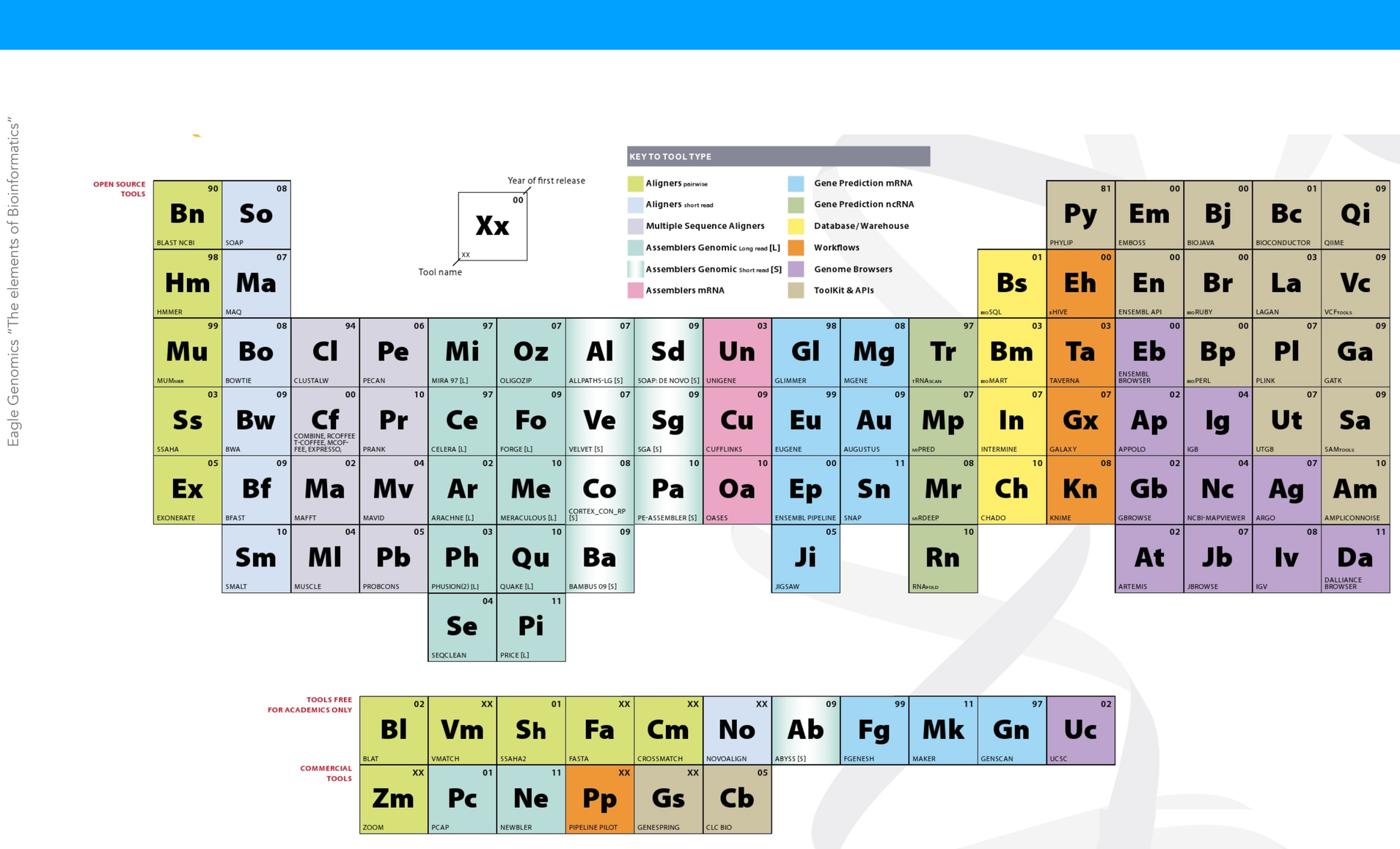
Grade	Mark	Degree Award	Description
A1	90-100	1st class or MSc with distinction	**Excellent**. Outstanding in every respect, the work is well beyond the level expected of a competent student at their level of study. The work should meet the criteria for an A2 grade and should also evidence a clear understanding of the limits of the state of knowledge, and their consequences, for the topic at hand.
A2	80-89	1st class or MSc with distinction	**Excellent**. Outstanding in some respects, the work is often beyond what is expected of a competent student at their level of study. Demonstrates that the student is actively extending their knowledge and capacity well beyond required materials and making new connections independently: for example, by showing a strong grasp of a range of related materials that are optional or not directly provided, or by demonstrating unusual creativity, depth of analysis, or synthesis with other areas of study.
A3	70-79	1st class or MSc with distinction	**Excellent**. Very good or excellent in most respects, the work is what might be expected of a very competent student. It indicates that the student has an excellent grasp of the required materials for the course, and may have demonstrated some limited knowledge of or fluency with additional optional materials, if provided.
B	60-69	2(i) or MSc with merit	**Very Good**. Good or very good in most respects, the work displays thorough mastery of the relevant learning outcomes.
C	50-59	2(ii) or MSc	**Good**. The work clearly meets requirements for demonstrating the relevant learning outcomes.
D	40-49	3rd class or PG Diploma/Cert	**Pass (undergraduate or Diploma level)**. The work meets minimum requirements for demonstrating the relevant learning outcomes. A satisfactory performance for undergraduate
E	30-39	Fail	**Marginal Fail**. The work fails to meet minimum requirements for demonstrating the relevant learning outcomes.
F	20-29	Fail	**Clear Fail**. The work is very weak and/or incomplete in important respects.
G	10-19	Fail	**Bad Fail**. The work is extremely weak or mostly incomplete/absent.
H	0-9	Fail	**Bad Fail**. The work is absent or of very little, if any, consequence to the area in question.

Computing Labs

- Structured sessions putting into practice topics we've covered in lectures
- Weeks 2, 4, 6, & 8, @14:00 AT6.06
- Python - Jupyter Notebooks & Worksheets
- Q&As in lab and on site discussion boards
- Opportunity to meet and discuss the topics with your peers and demonstrators
- Asynchronous working possible if you cannot make labs
- University Noteable System available 24/7

ian.simpson@ed.ac.uk

The "Elements" of Bioinformatics



ian.simpson@ed.ac.uk

What is Bioinformatics ?

bioinformatics, *n.*

View as: Outline | Full entry

OED Oxford English Dictionary
The definitive record of the English language

Pronunciation: Brit. /ˌbʌɪəʊɪnfəˈmatɪks/ , U.S. /ˈˌbaɪoʊˌɪnfərˈmædɪks/

Etymology: < BIO- *comb. form* + INFORMATICS *n.* Compare Dutch *bioinformatica* (1978 or earlier).... (Show More)

With *sing.* concord. The branch of science concerned with information and information flow in biological systems, esp. the use of computational methods in genetics and genomics.

Categories »

1976 *Acta Biotheoretica* **25** 3 We publish papers on the philosophy of biology, biomathematics and bioinformatics.

1987 *Science* 4 Sept. 1108/3 A new research program in bioinformatics. This is intended to bring together research in computing science, structural biology, and molecular genetics.

2001 *N.Y. Times* 4 Jan. B6/2 A leader in cutting-edge fields like bioinformatics, in which computers are used to decipher genes and proteins.

2007 *Wired* May 125/1 Delwart chops all the remaining genetic material into little pieces... He uses specially designed bioinformatics software to check them.

Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med. 2001;40(4):346-58.

ian.simpson@ed.ac.uk

Branches of Bioinformatics

"Statistics & Machine Learning"

- Mathematical & Statistical Modelling
- Dynamical modelling (simulation)
- Data integration methods

"Computational Biology"

- High Performance Computing (HPC)
- Database, software and algorithm development
- Web Services

"Classical Bioinformatics"

- Structural Bioinformatics
- Gene Prediction
- Motif Prediction & Pattern matching
- Gene expression analysis
- Genome Annotation
- Molecular phylogenetics

"Systems Bioinformatics"

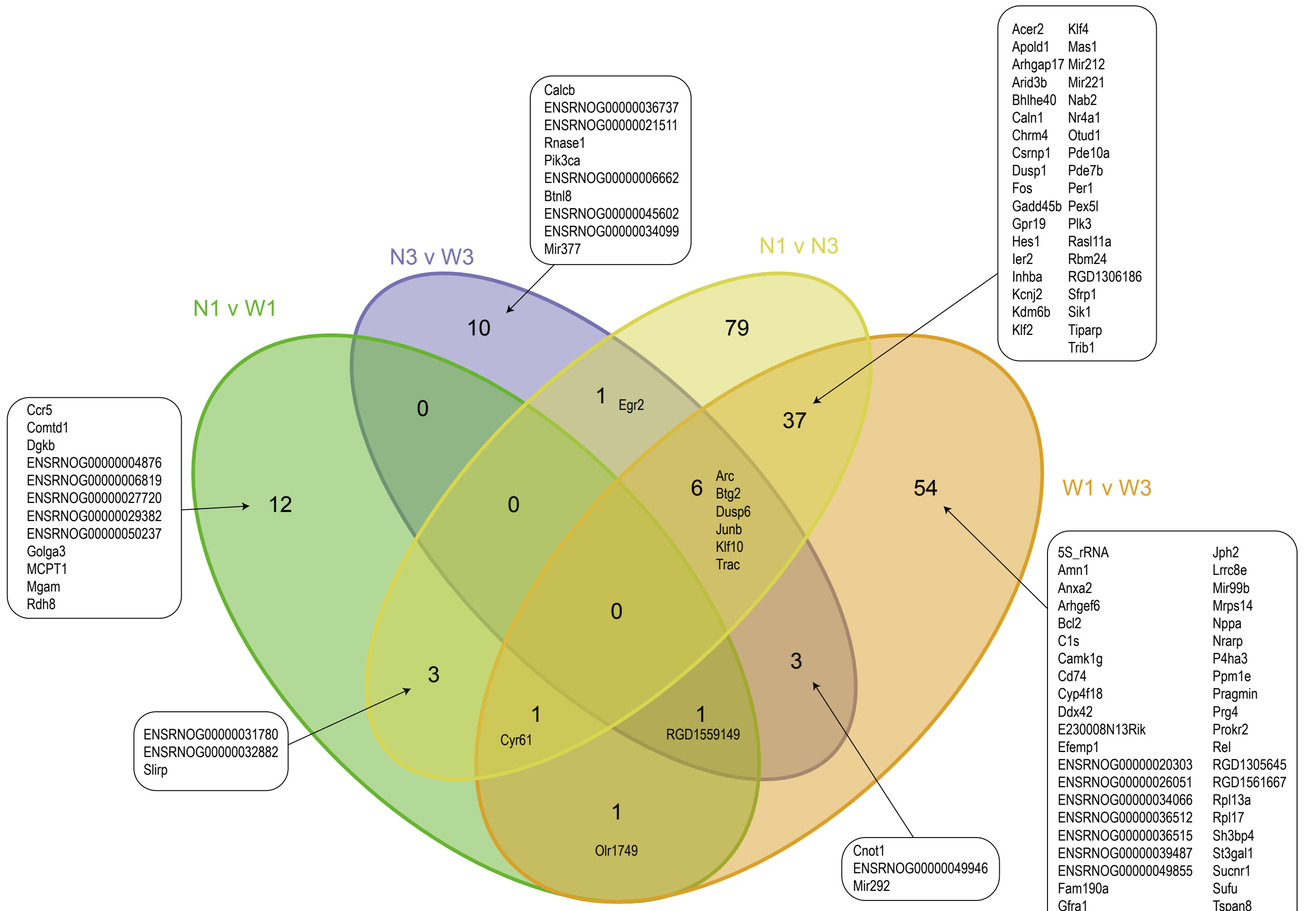
- Systems Biology
- Network Biology
- Drug Design
- Image Analysis

"Statistical Genetics"

- Genome wide linkage and association Studies
- Genotype Analysis
- Population Genetics

ian.simpson@ed.ac.uk

Integration of Gene Expression Data

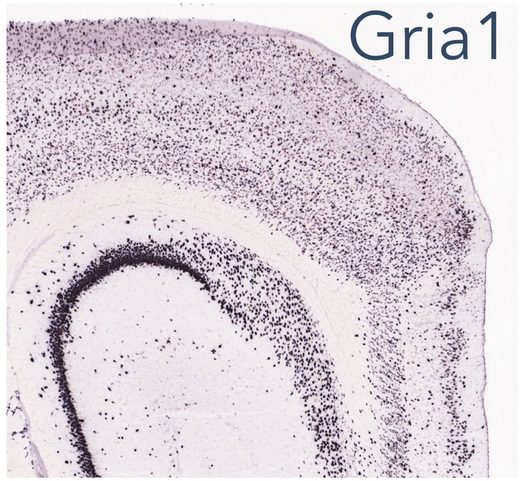


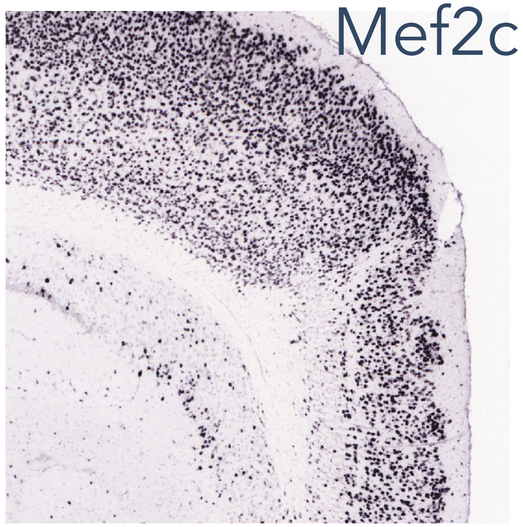
Conditioned Taste Aversion (CTA) learning paradigms

- compared gene expression in the gustatory cortex between rats subject to CTA with or without novel taste
- arrays were sampled 1 and 3 hours post-experience
- we have identified over 200 genes that are differentially expressed in the learning process
- key genes involved in the regulation of synaptic plasticity are amongst the lists and are being validated in vivo

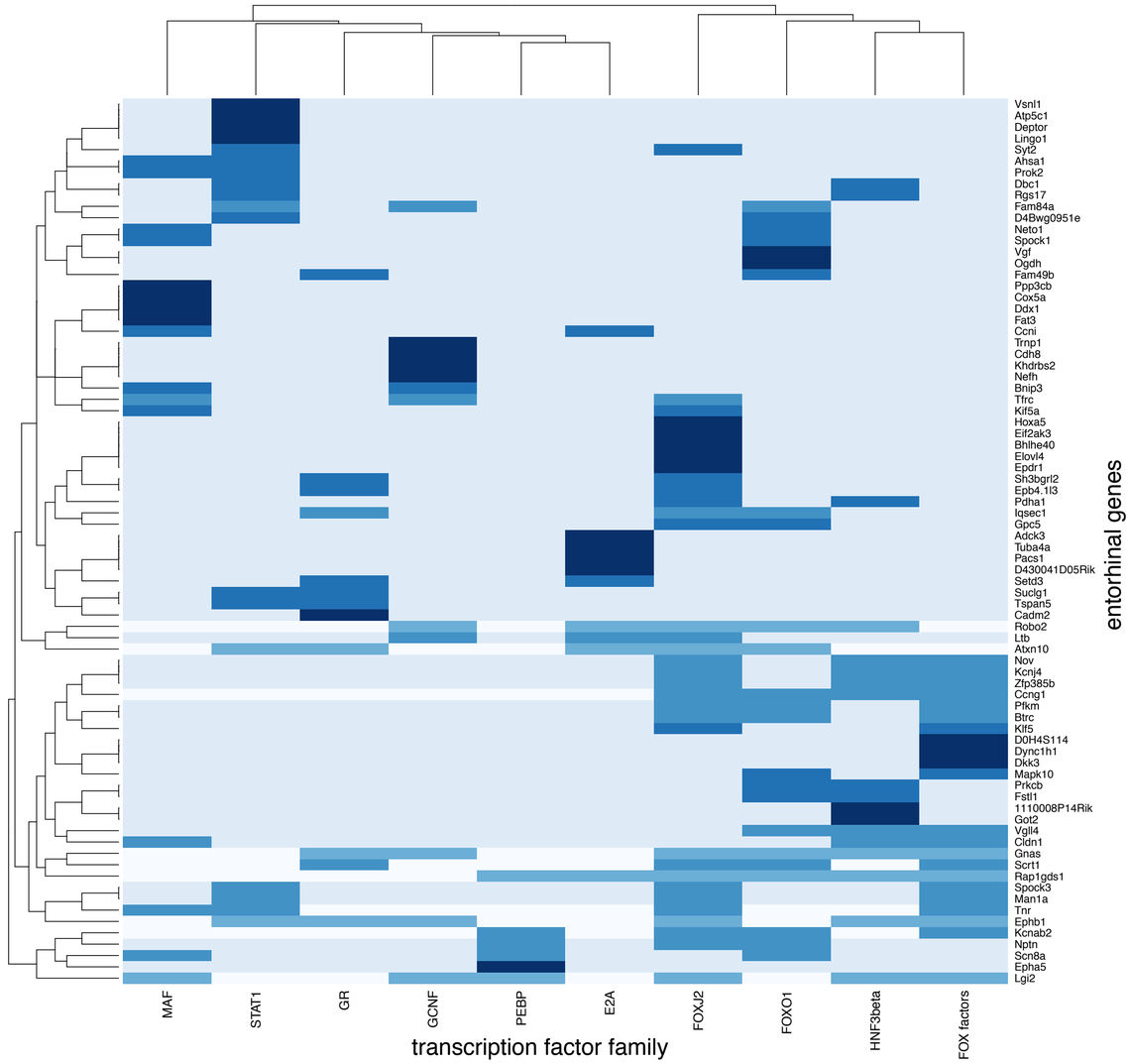
ian.simpson@ed.ac.uk

Gene Expression Regulation in the Brain





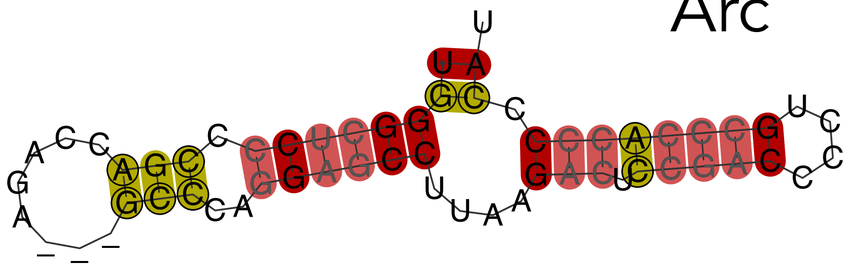
expression coding in the entorhinal cortex

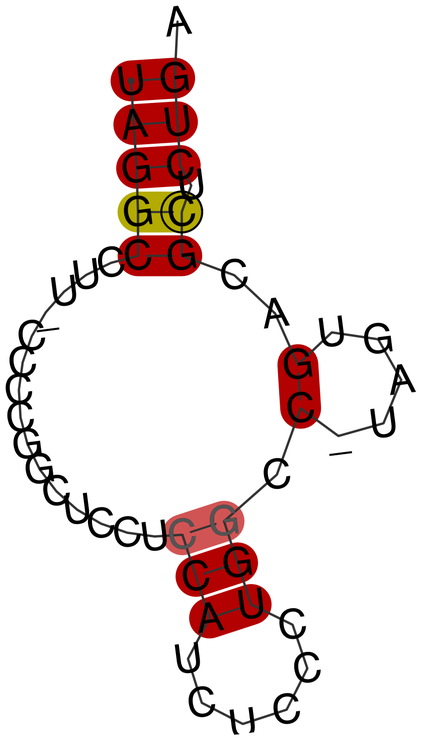


ian.simpson@ed.ac.uk

Predicting mRNA localisation signals

dendritic target signals



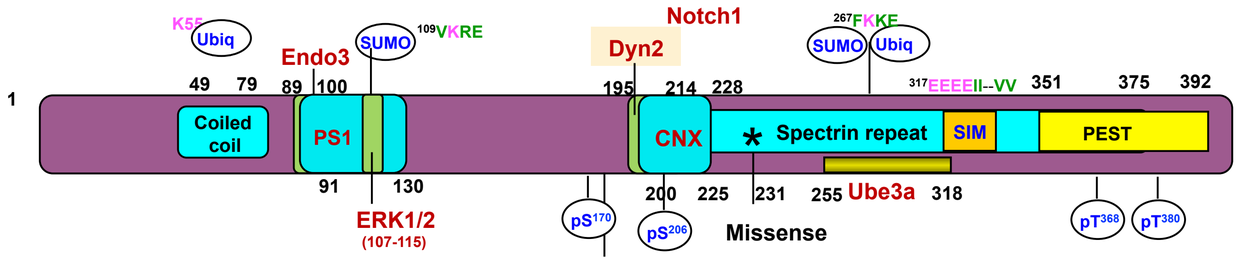


- align 3'UTRs of dendritically localised mRNAs
- predict secondary structure (Vienna, RNApromo, MC-Fold, LocARNA)
- combine sequence and structural (region) predictions (filter 60% id, >=35nt)
- rank candidates for experimental validation

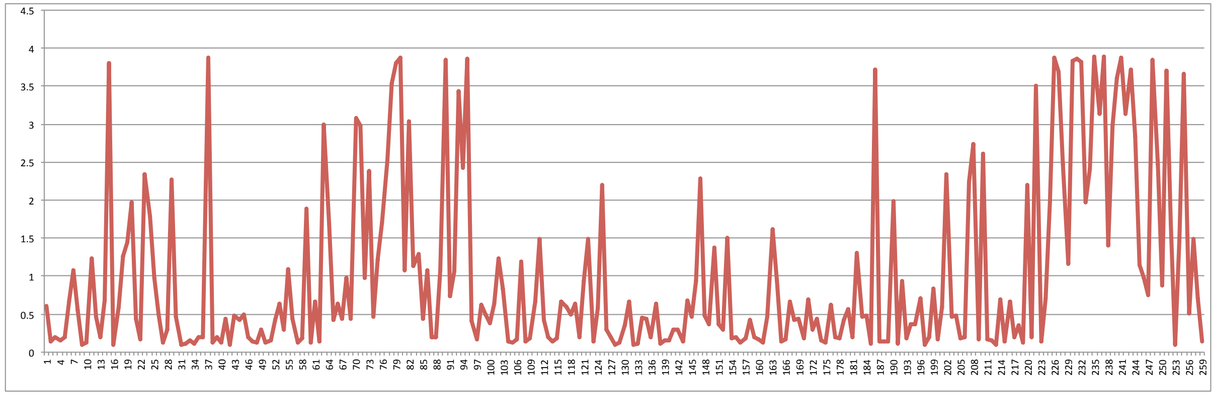
ian.simpson@ed.ac.uk

Evolutionary Sequence Conservation & Neuroregulation

Which sites are most likely to be functional ?



domain structure of Arc

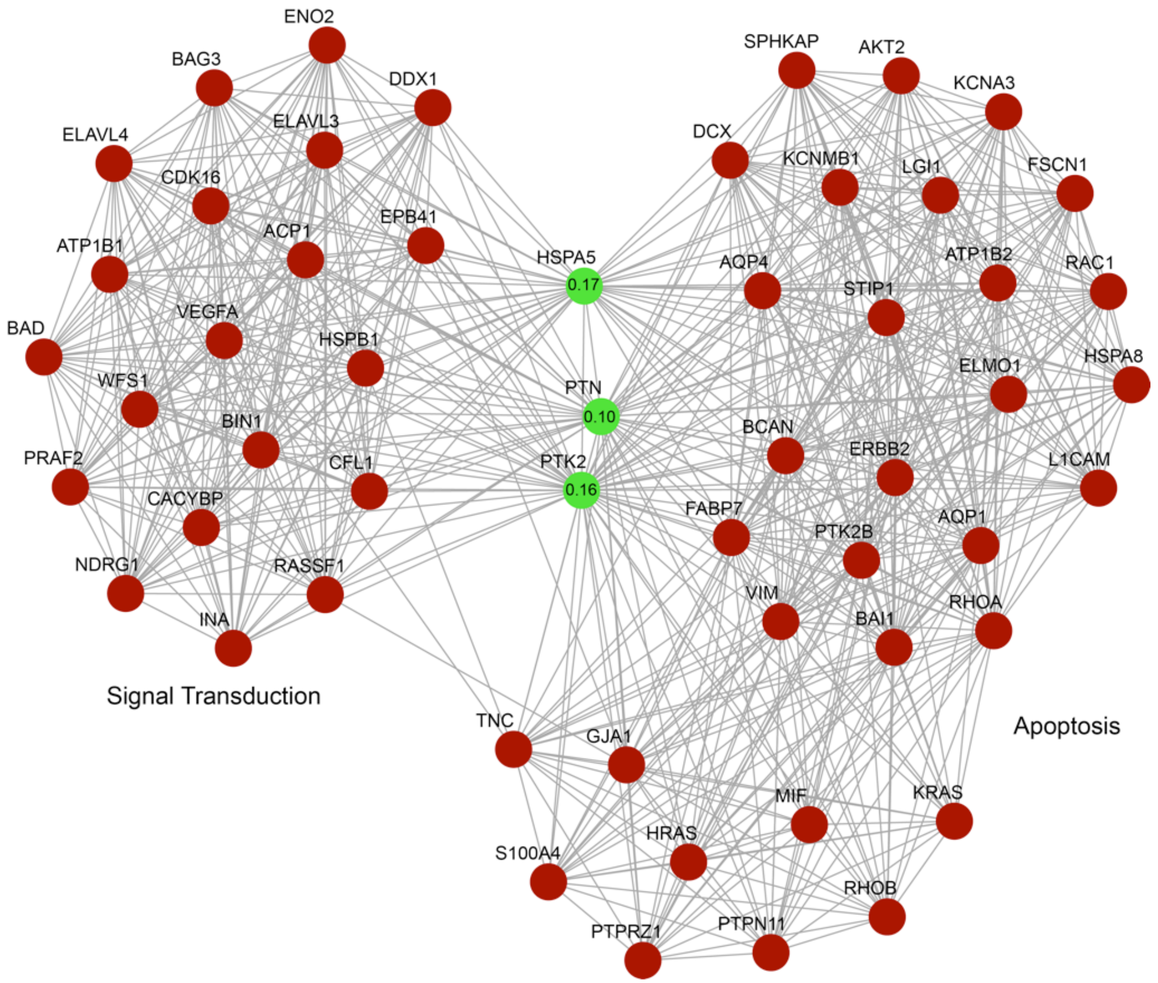


site-specific evolutionary rate

- identified Arc orthologues
- modeled sequence evolution (RaxML)
- mapped selection rates to Arc sites
- candidate sites to be validated in the lab

ian.simpson@ed.ac.uk

Network analysis reveals key disease genes



Disease Gene Network

- Mine gene disease associations

OMIM/HDO/GeneRIF

- Calculate pair-wise common disease counts between all gene pairs
- Set genes as nodes, disease counts as edges

Cluster network by community

- Newman and Girvan 2002

Functional annotation

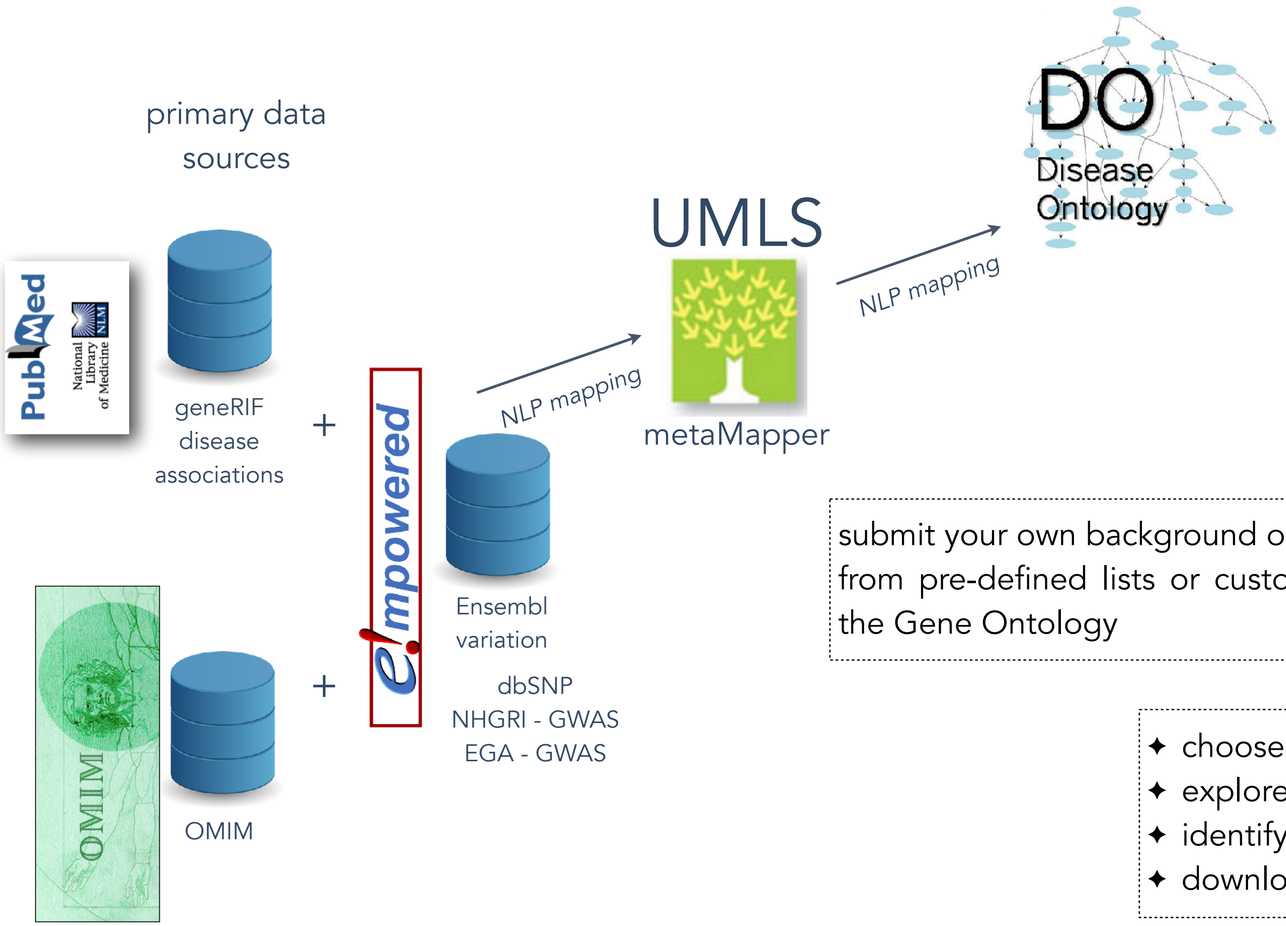
- semantic similarity, BINGO/GO, InterPro domain, KEGG, COG.

Network analyses

- betweenness centrality
- modularity
- clustering coefficients
- power-law
- stress testing
- critical genes (targets)

ian.simpson@ed.ac.uk

Computational Pipelines - DisEnT - Disease Enrichment Rool



discover enriched disease states from gene lists (Ensembl or Entrez) from 6 model organisms

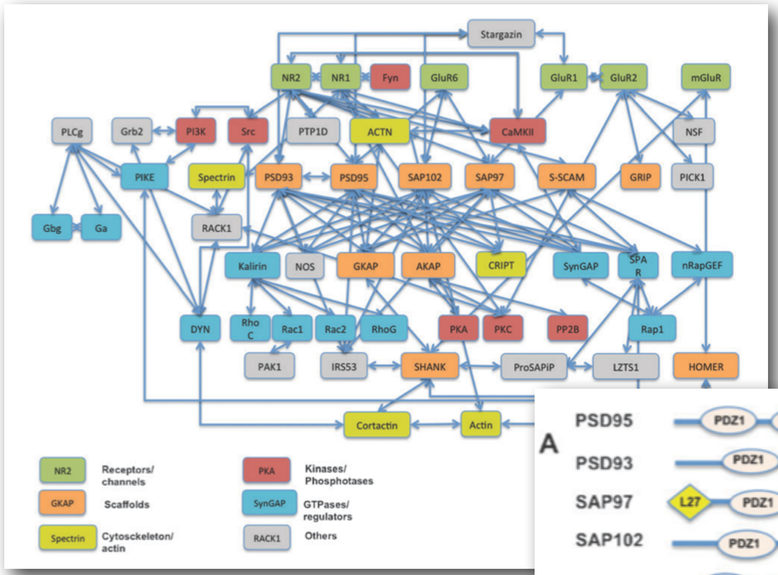
- human
- mouse
- rat
- zebrafish
- fly
- flatworm
- yeast

submit your own background or select from from pre-defined lists or custom defined from the Gene Ontology

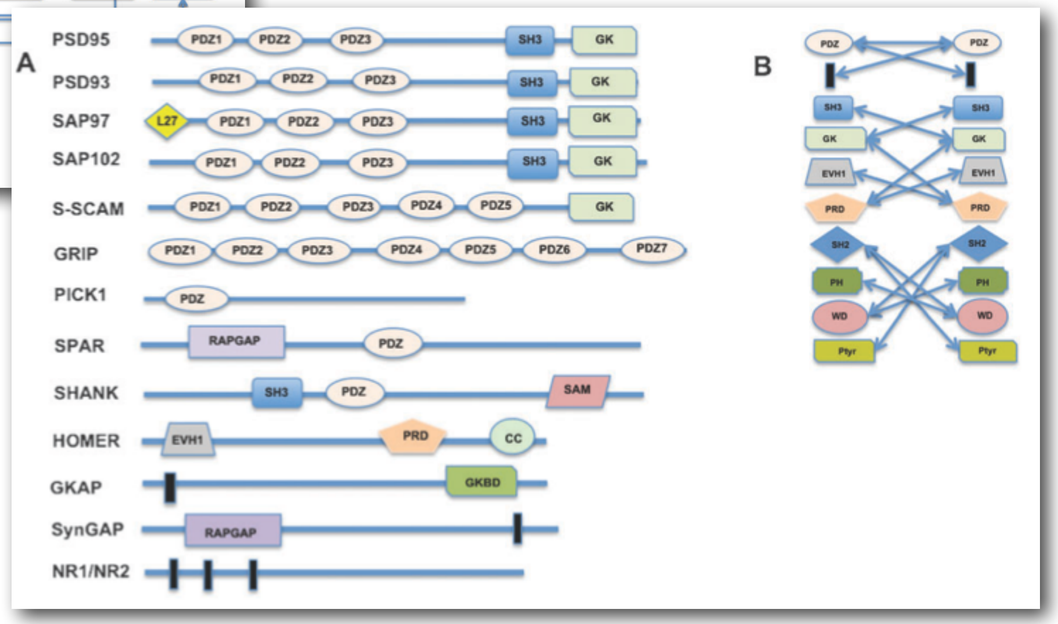
- choose p-value cut-off, with correction
- explore data through link outs
- identify contributing genes
- download results in CSV format

ian.simpson@ed.ac.uk

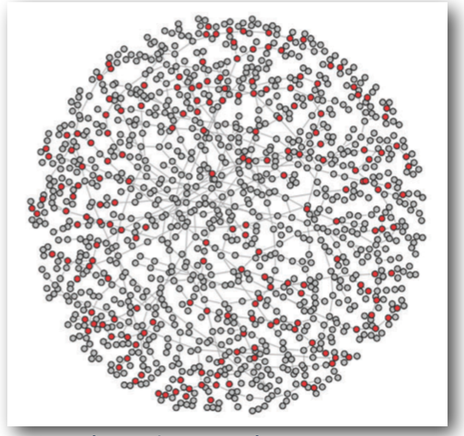
Dynamical Modelling of Protein Complexes



PSD core-complex



Protein domain binding preferences

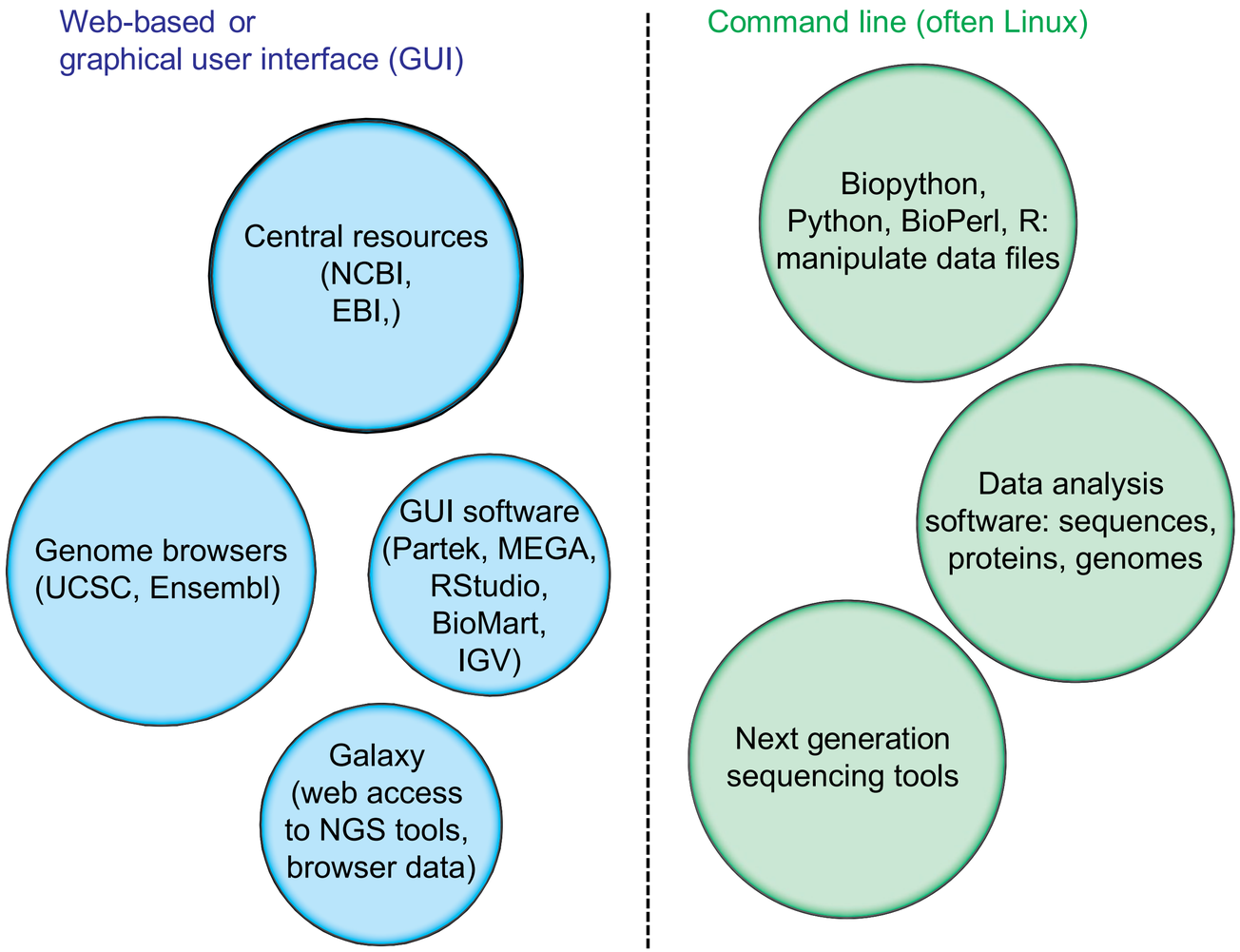


simulated complex

Sorokina et al. 2011

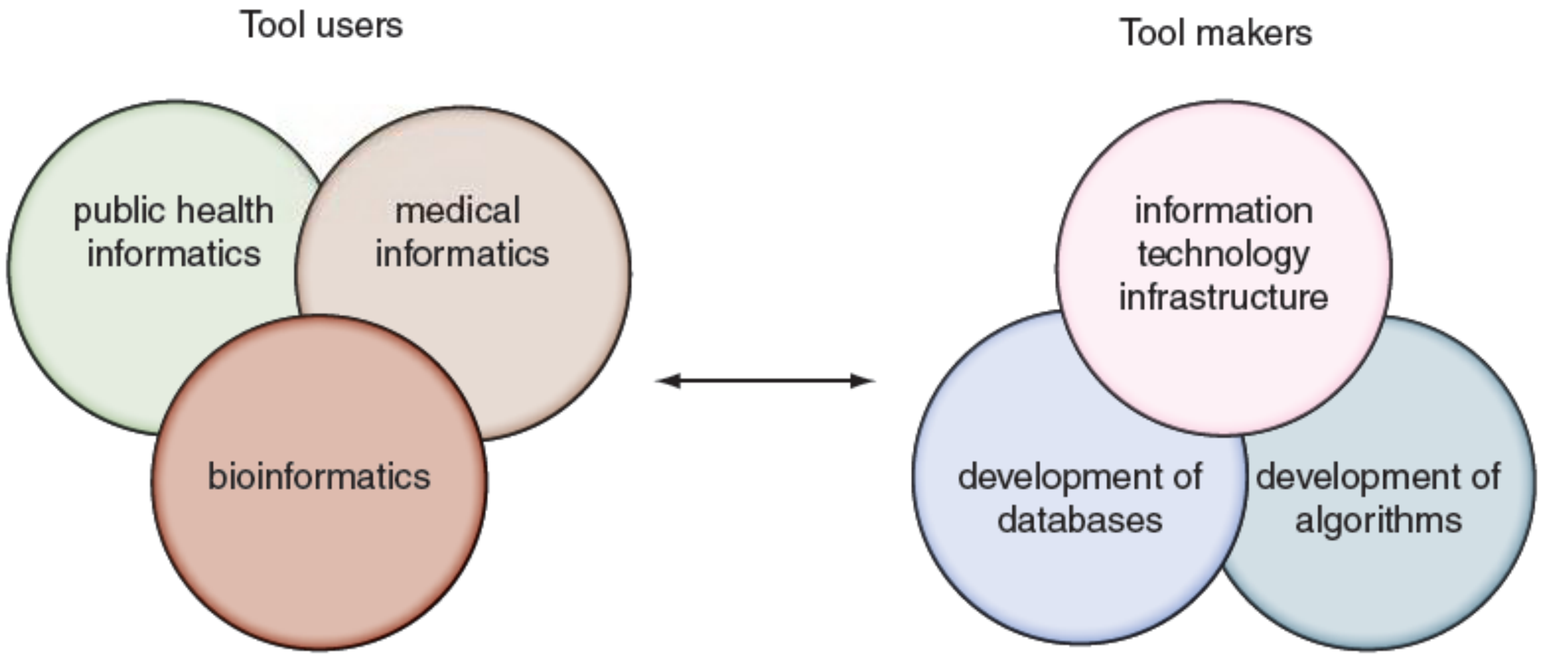
ian.simpson@ed.ac.uk

Web Tools or Programming?



ian.simpson@ed.ac.uk

Tool Users & Tool Makers



ian.simpson@ed.ac.uk

Coding in Bioinformatics

Python

R/Bioconductor

- Bioinformatics requires extensive programming skills in several languages
  - Dominated by
    - R (statistical programming) - https://cran.r-project.org, https://www.bioconductor.org
    - Python - https://www.python.org, https://biopython.org, Pandas, Numpy, Scipy etc.
  - Increasing use of Machine Learning and "Data Science" methods
    - Supervised/Unsupervised Classification
    - Training of Statistical Models
    - Modelling of Data in Graph (Network) Structures
    - Bayesian Inference
    - Computational Statistics
    - Simulation
  - Ability to integrate multiple packages/tools/data sources is key

ian.simpson@ed.ac.uk

Bioinformatics - crossing disciplines

Computer Science

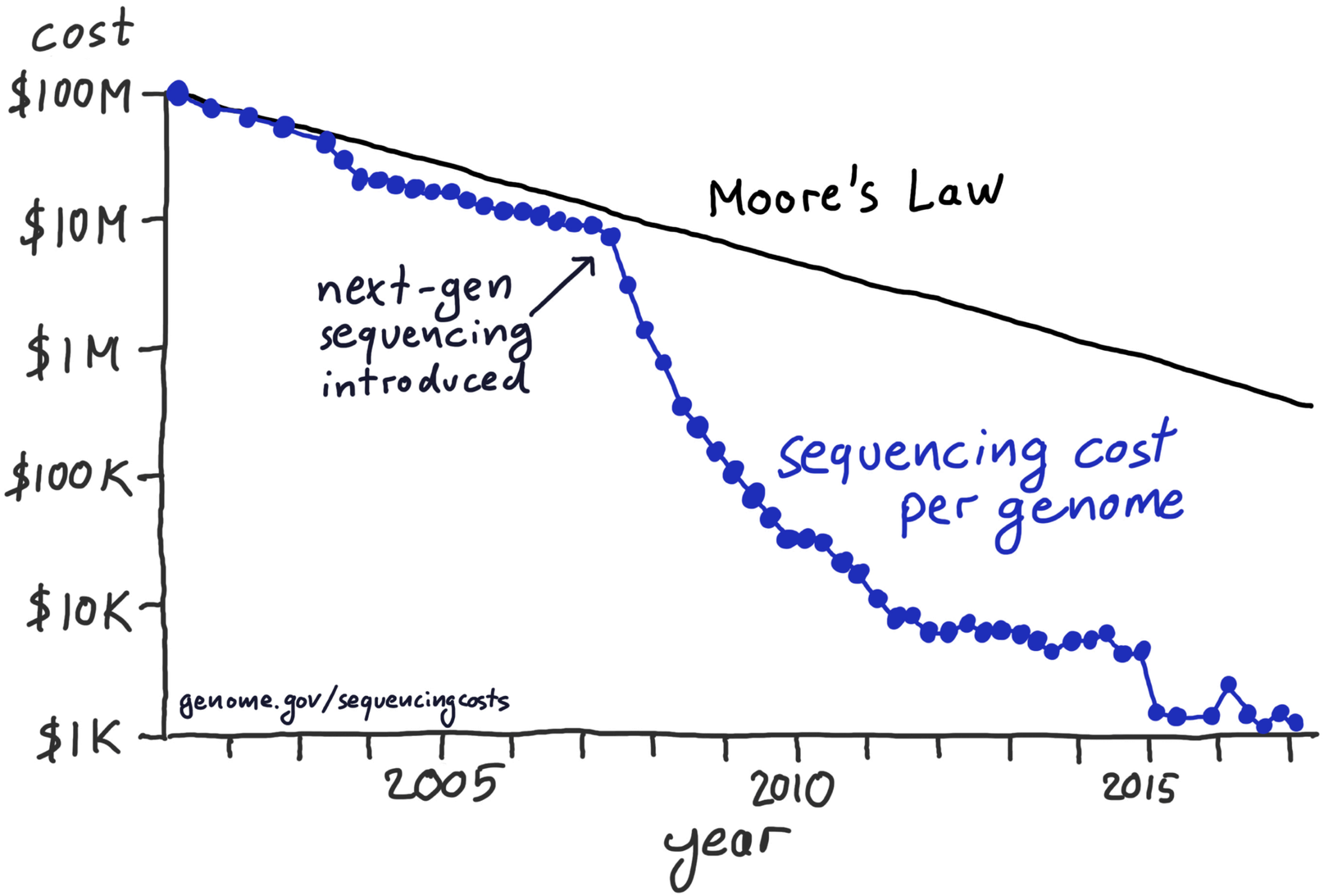
- Biology provides CS with new challenges with clear medical significance.
- Complex and large datasets, sometimes very noisy with hidden structures.
- Can biological solutions be used to inspire new computational tools and methods?

Biology

- Bioinformatics enables analysis of increasingly large and complex data sets generated in Biological experiments. (100k Genomes, UKB-Imaging, Microbiomes, Single-cell sequencing)
- Bioinformatics facilitates the structuring of often unstructured and heterogeneous data to allow data integration for insight
- Bridges CS to Biology especially when realised in Web services & Tools

ian.simpson@ed.ac.uk

Next-generation Sequencing



ian.simpson@ed.ac.uk

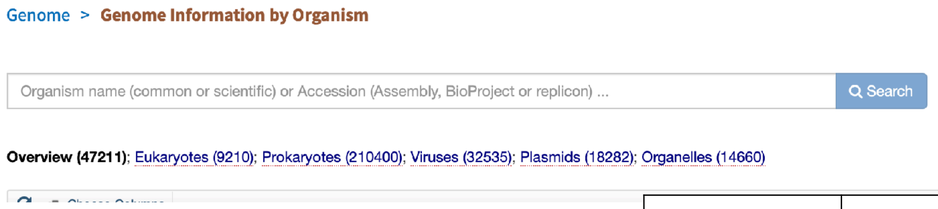
Genomic (Open) Data

2015

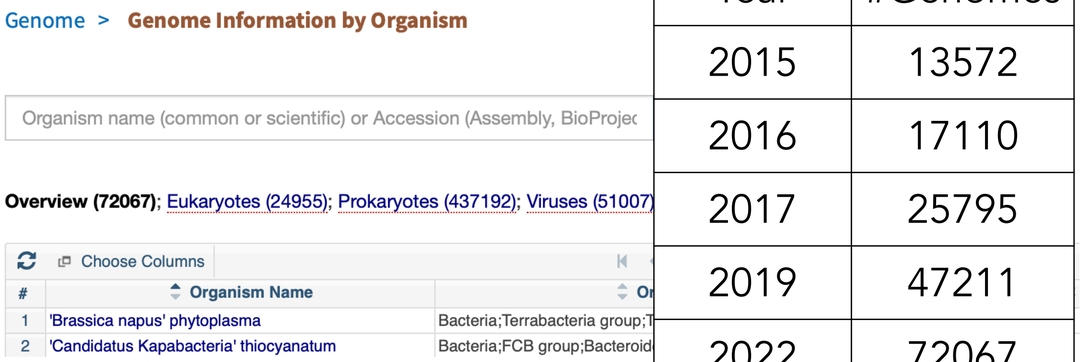
2016

2017

2019

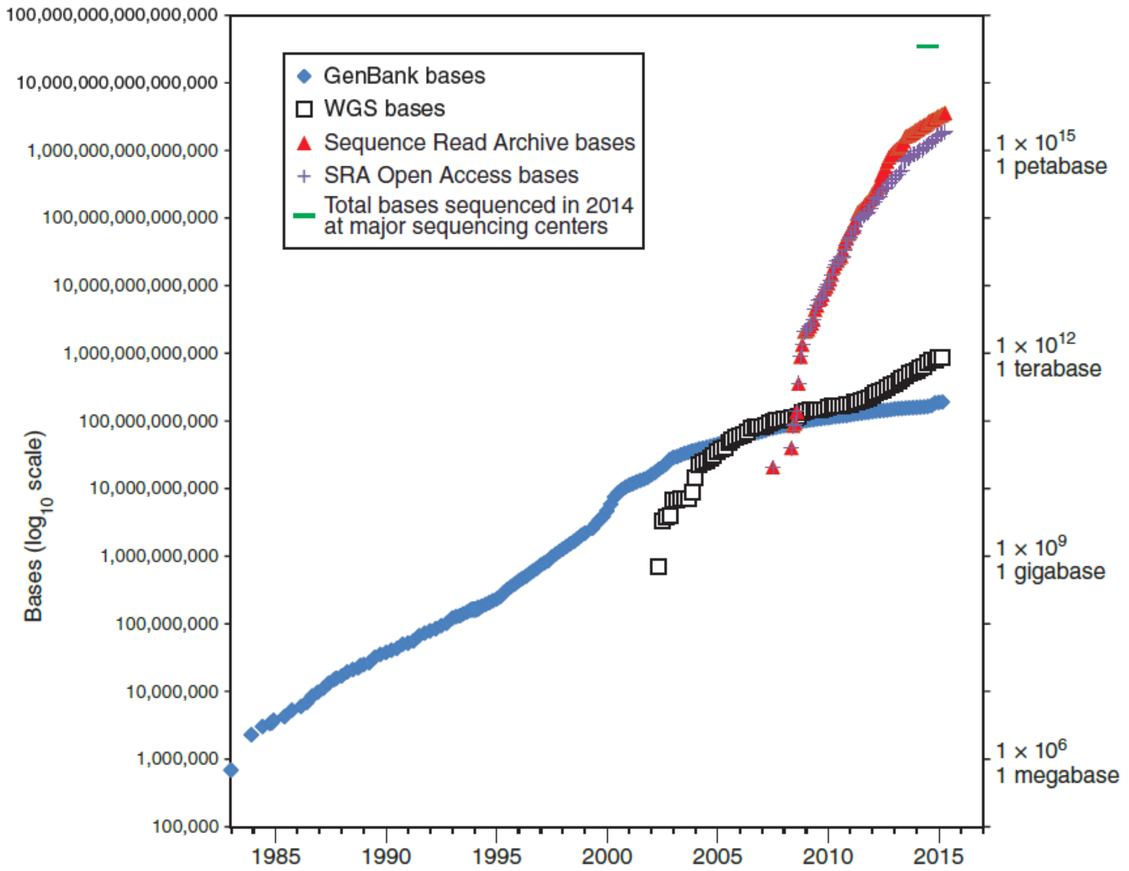


2022



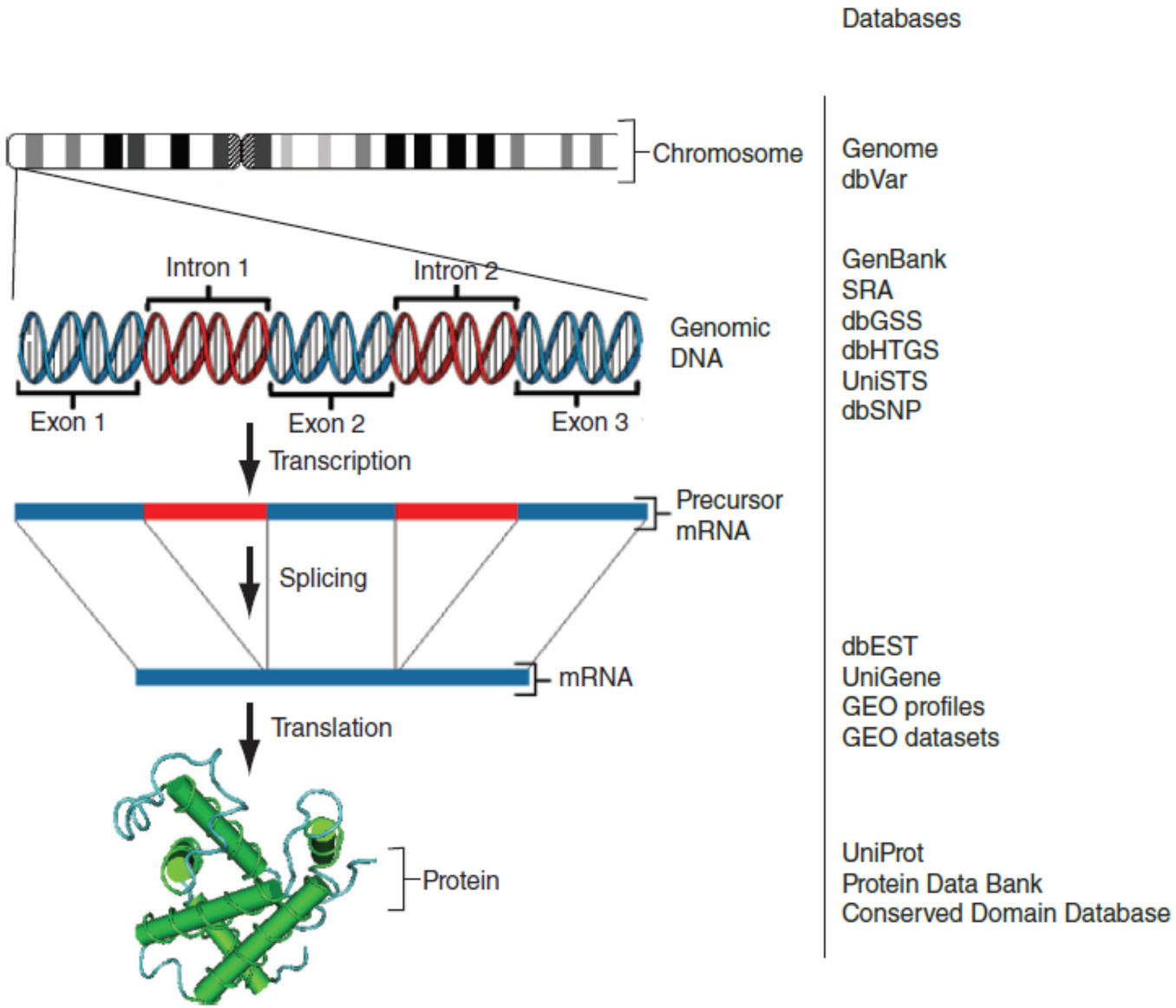
Year	#Genomes
2015	13572
2016	17110
2017	25795
2019	47211
2022	72067

Growth of Sequence Databases



ian.simpson@ed.ac.uk

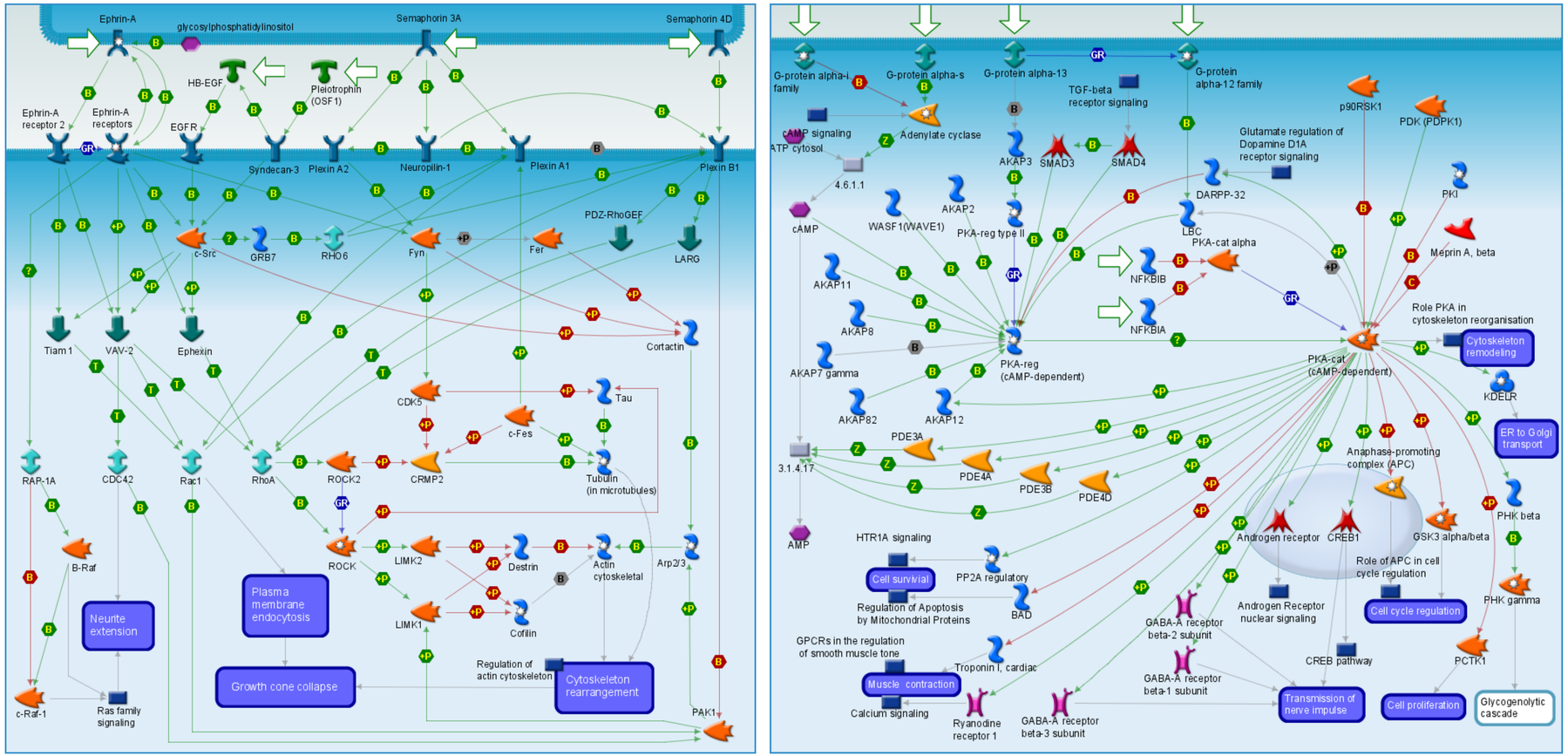
Data Types & Databases



B&FG3e

ian.simpson@ed.ac.uk

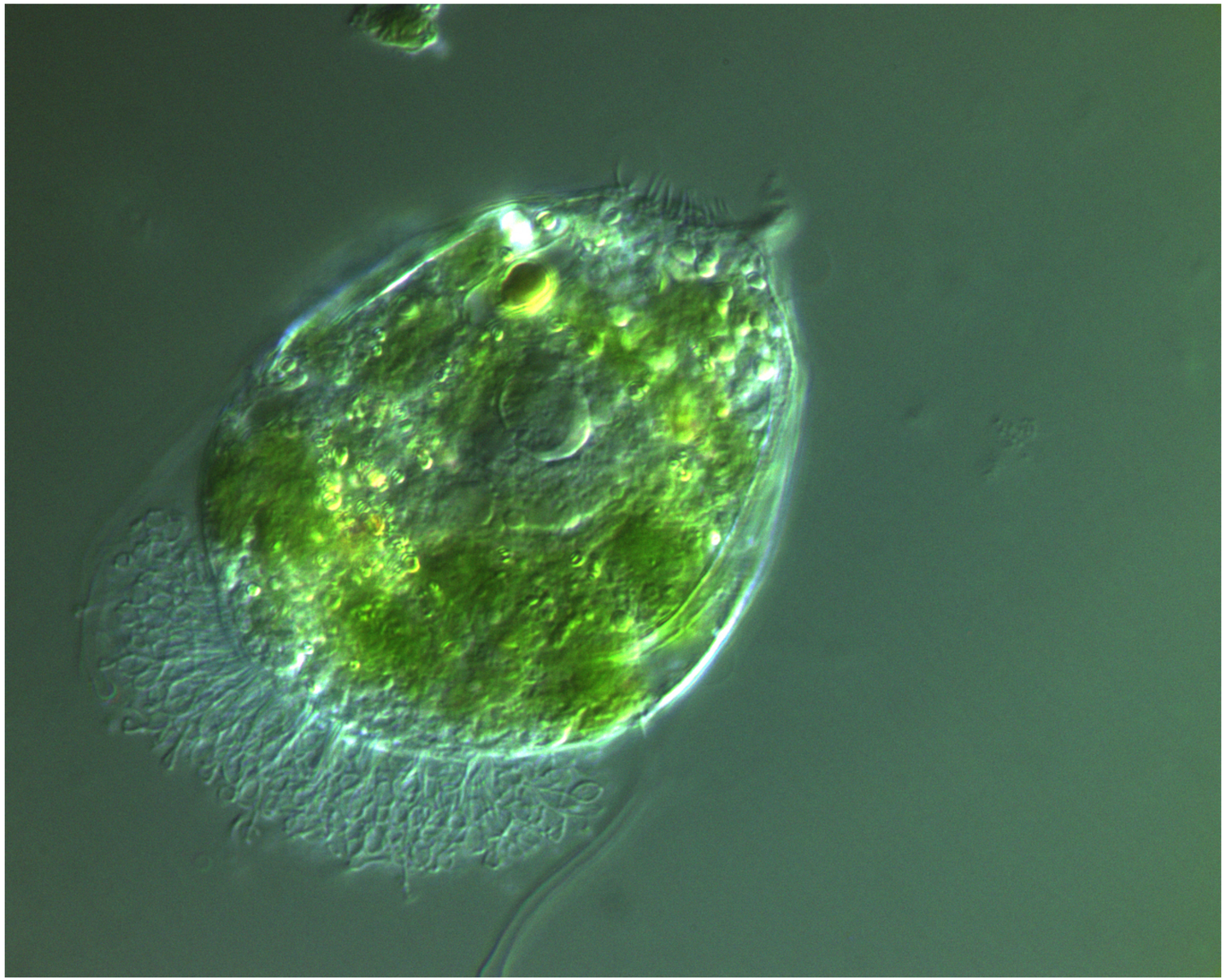
How much biology should the bioinformatician know?



Bioinformaticians and Computational Biologists **develop** and **test** new biological hypotheses, not just analyse data for others.

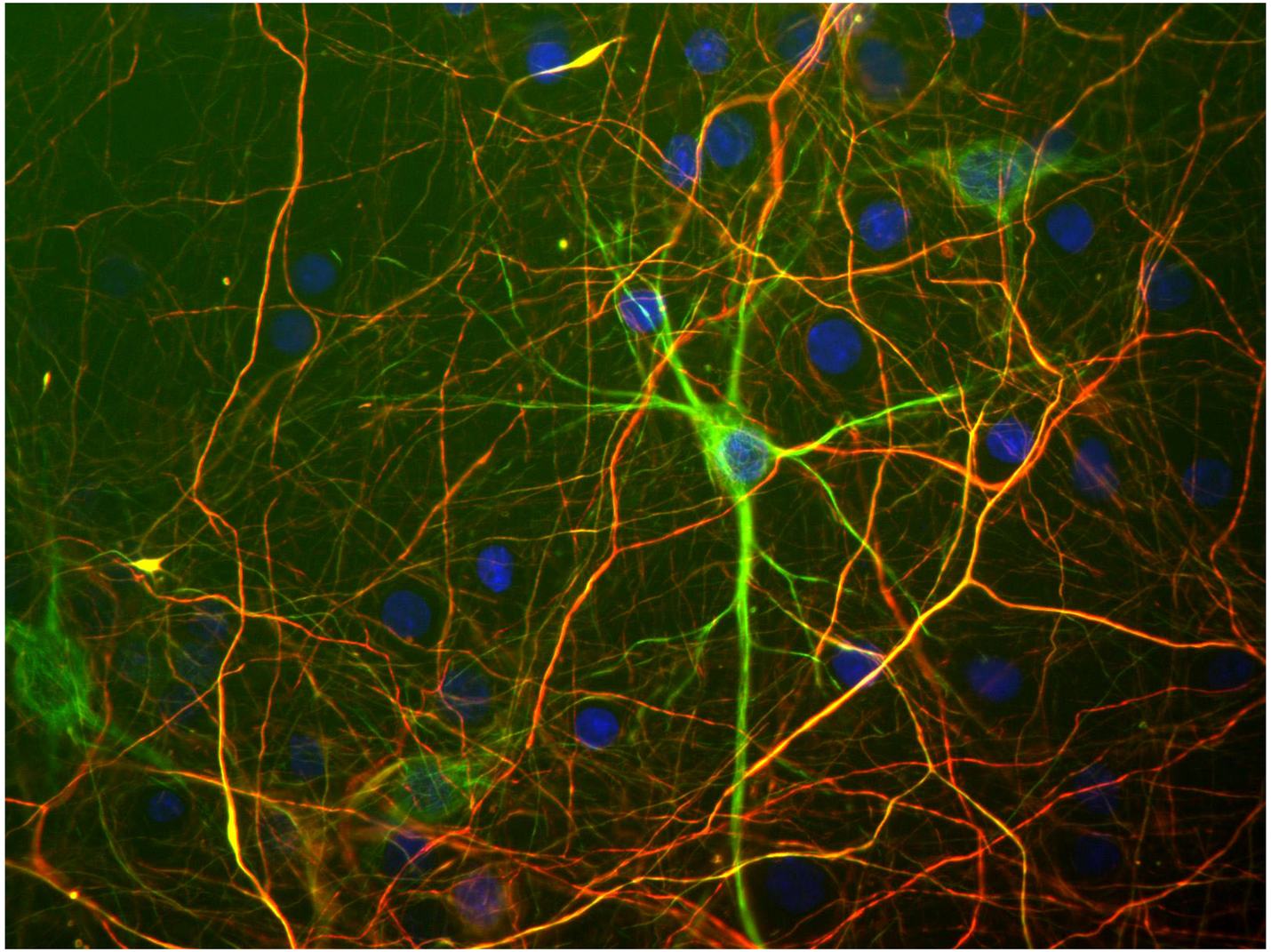
ian.simpson@ed.ac.uk

A Unicellular Organism - Strombidium stylifer



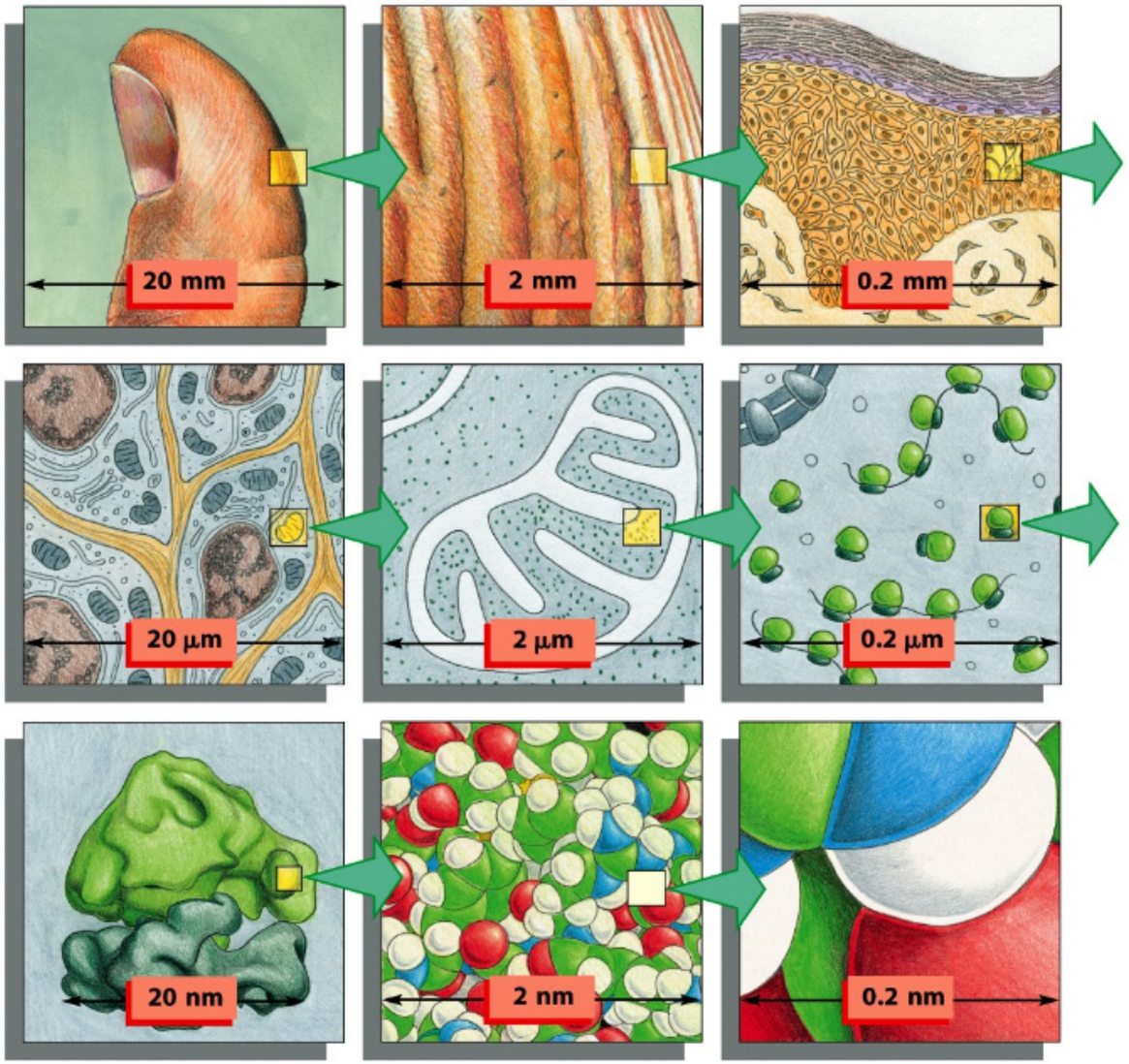
ian.simpson@ed.ac.uk

Cultured Neurons



ian.simpson@ed.ac.uk

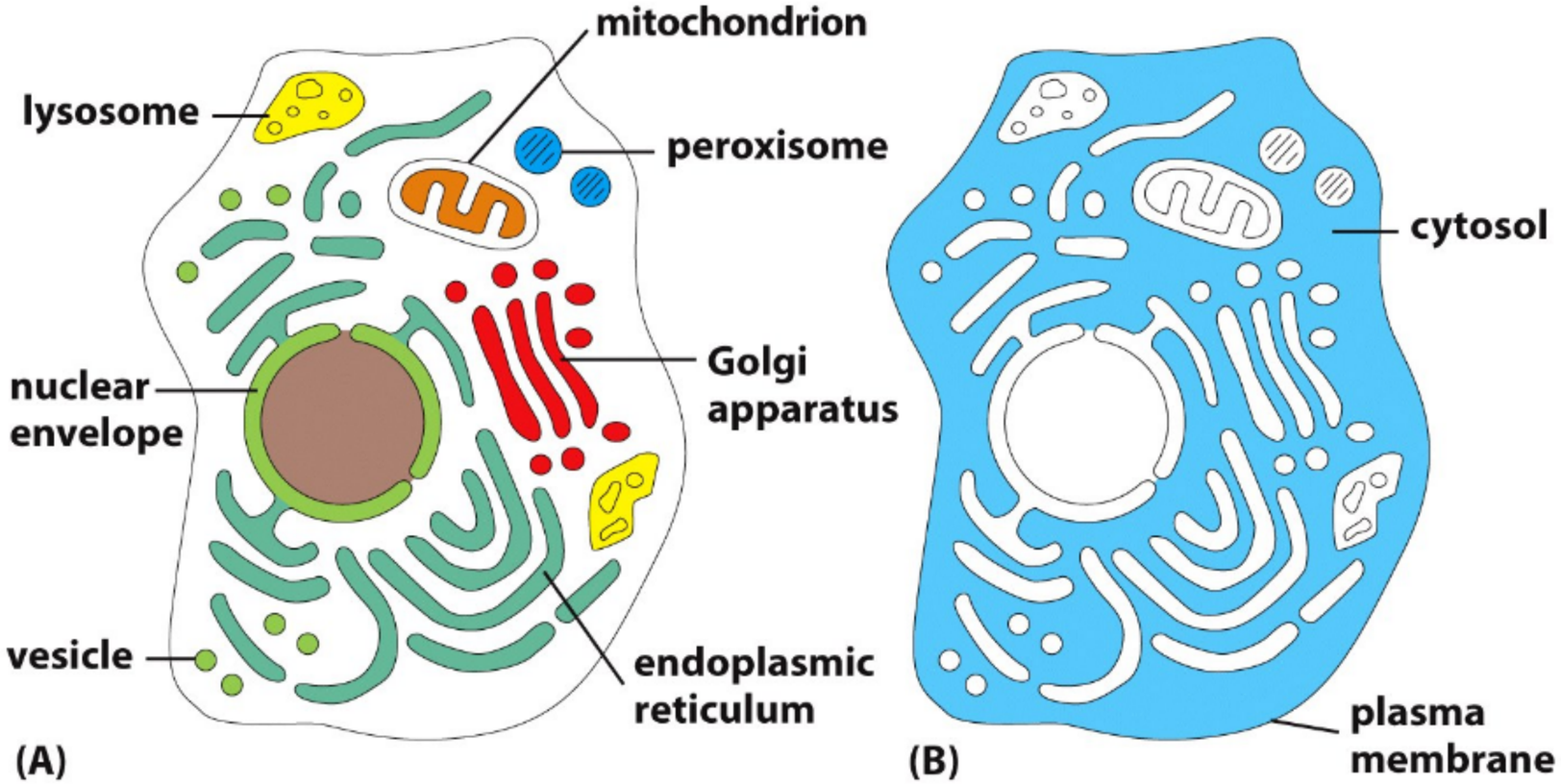
Studying Across Scales



Alberts, B. (2014). *Essential cell biology*. New York: Garland Science.

ian.simpson@ed.ac.uk

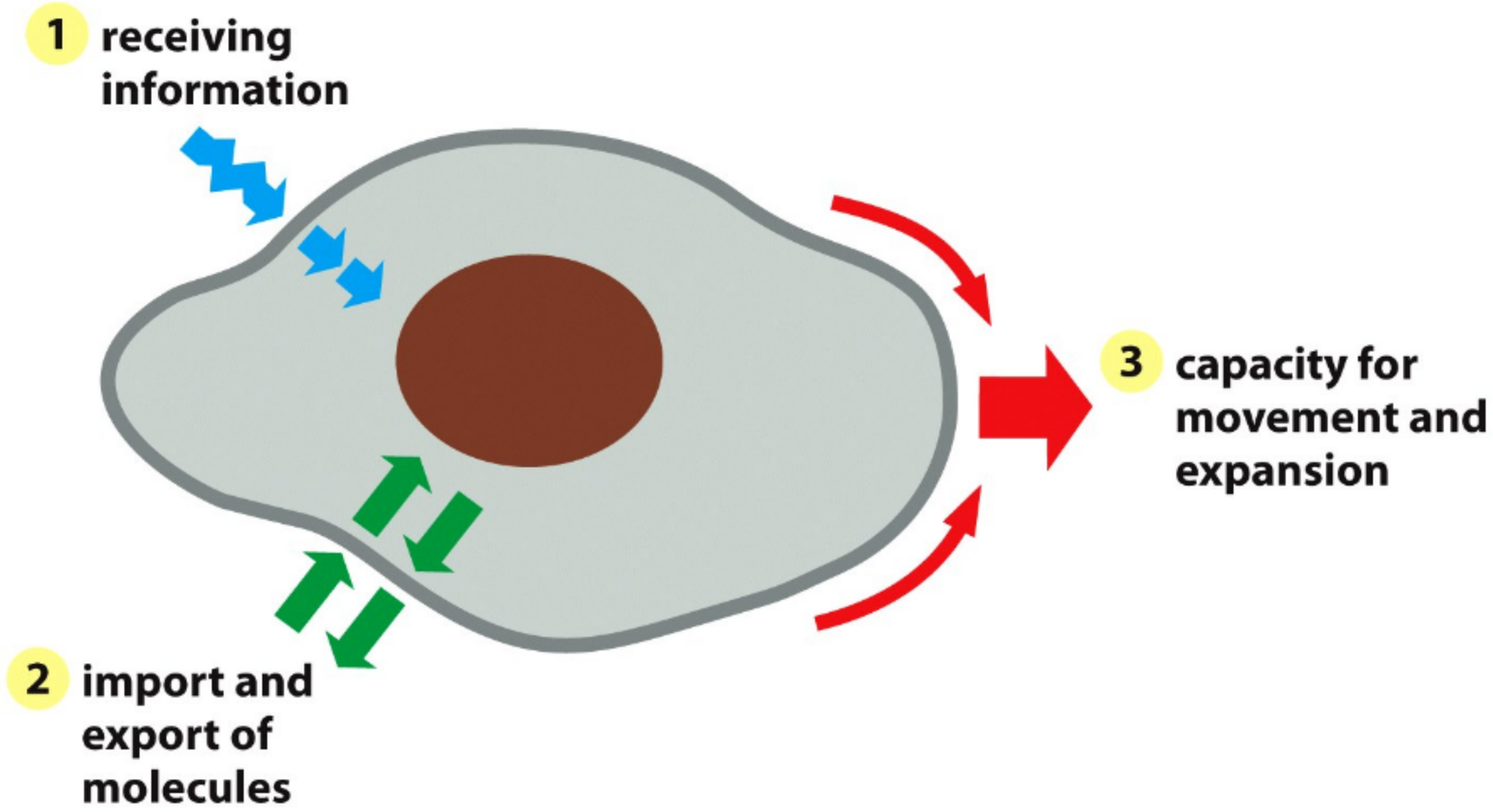
Cell Structure



Alberts, B. (2014). *Essential cell biology*. New York: Garland Science.

ian.simpson@ed.ac.uk

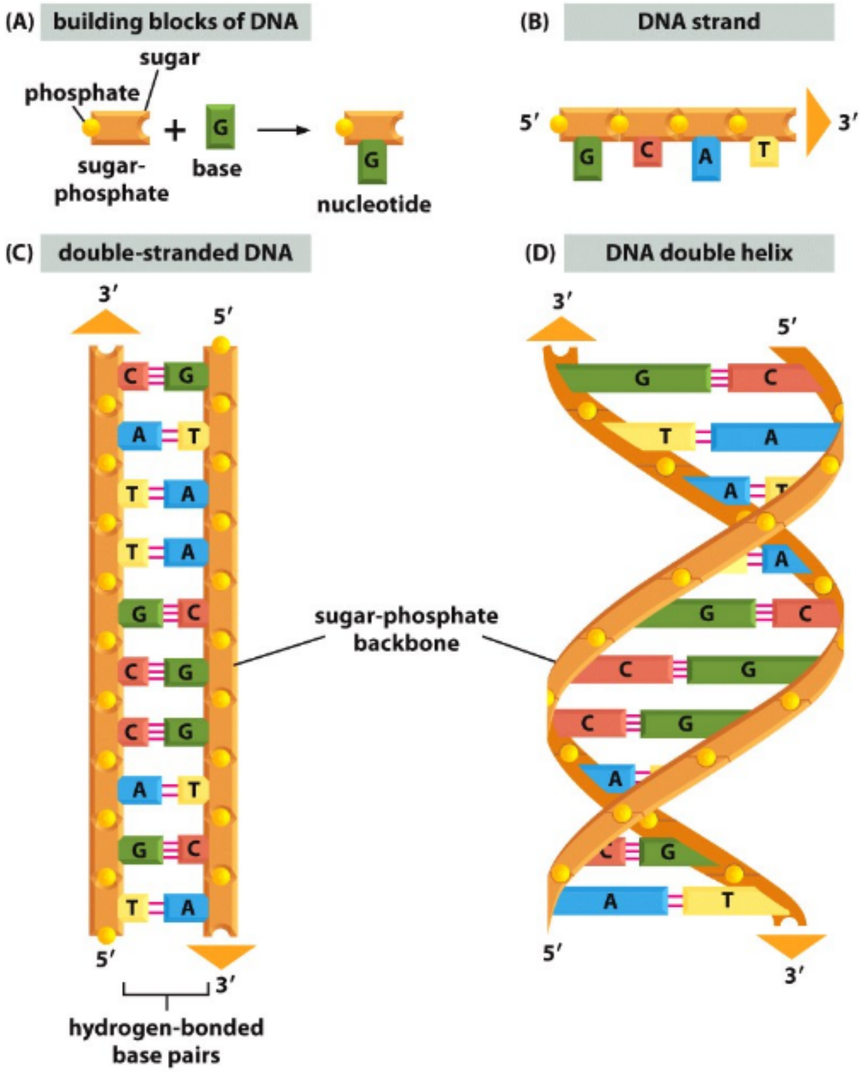
Cells Respond to their Environment

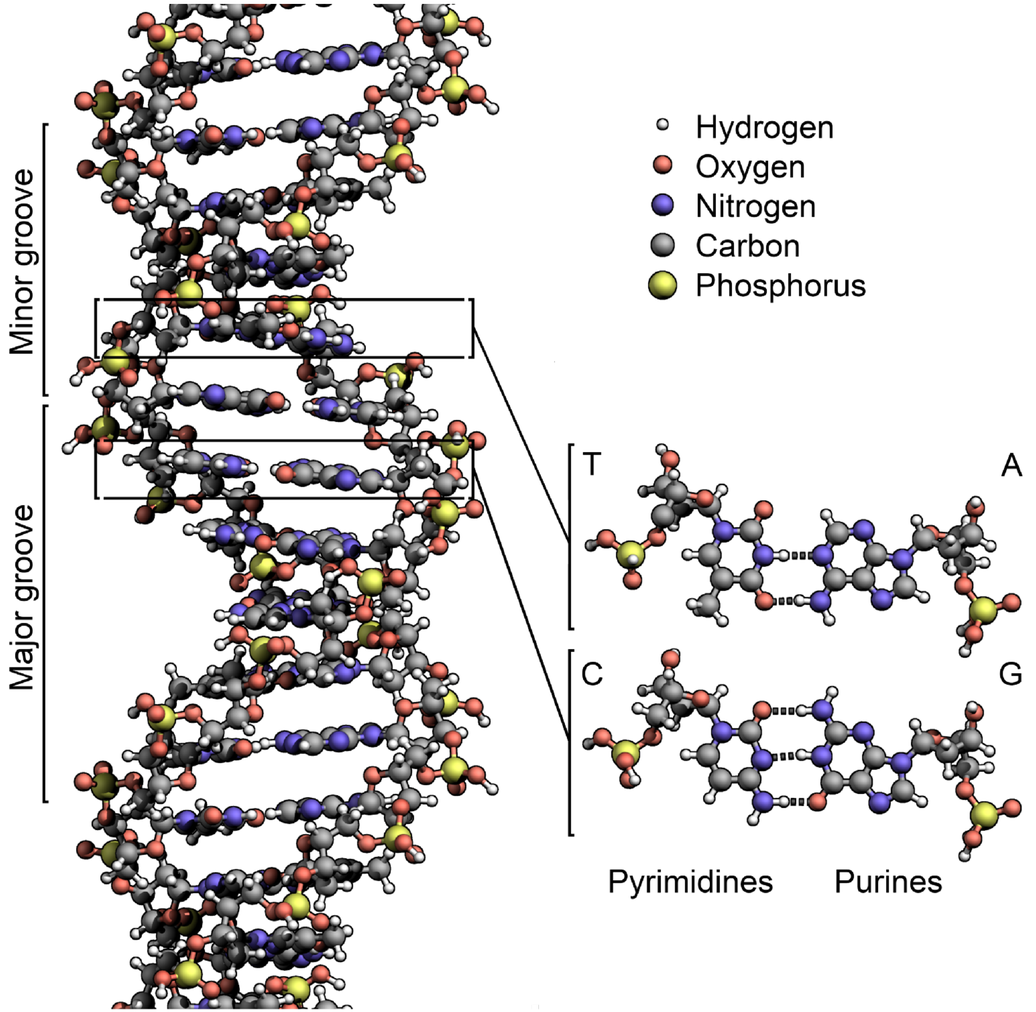


Alberts, B. (2014). *Essential cell biology*. New York: Garland Science.

ian.simpson@ed.ac.uk

Deoxy-riboNucleicAcid (DNA)

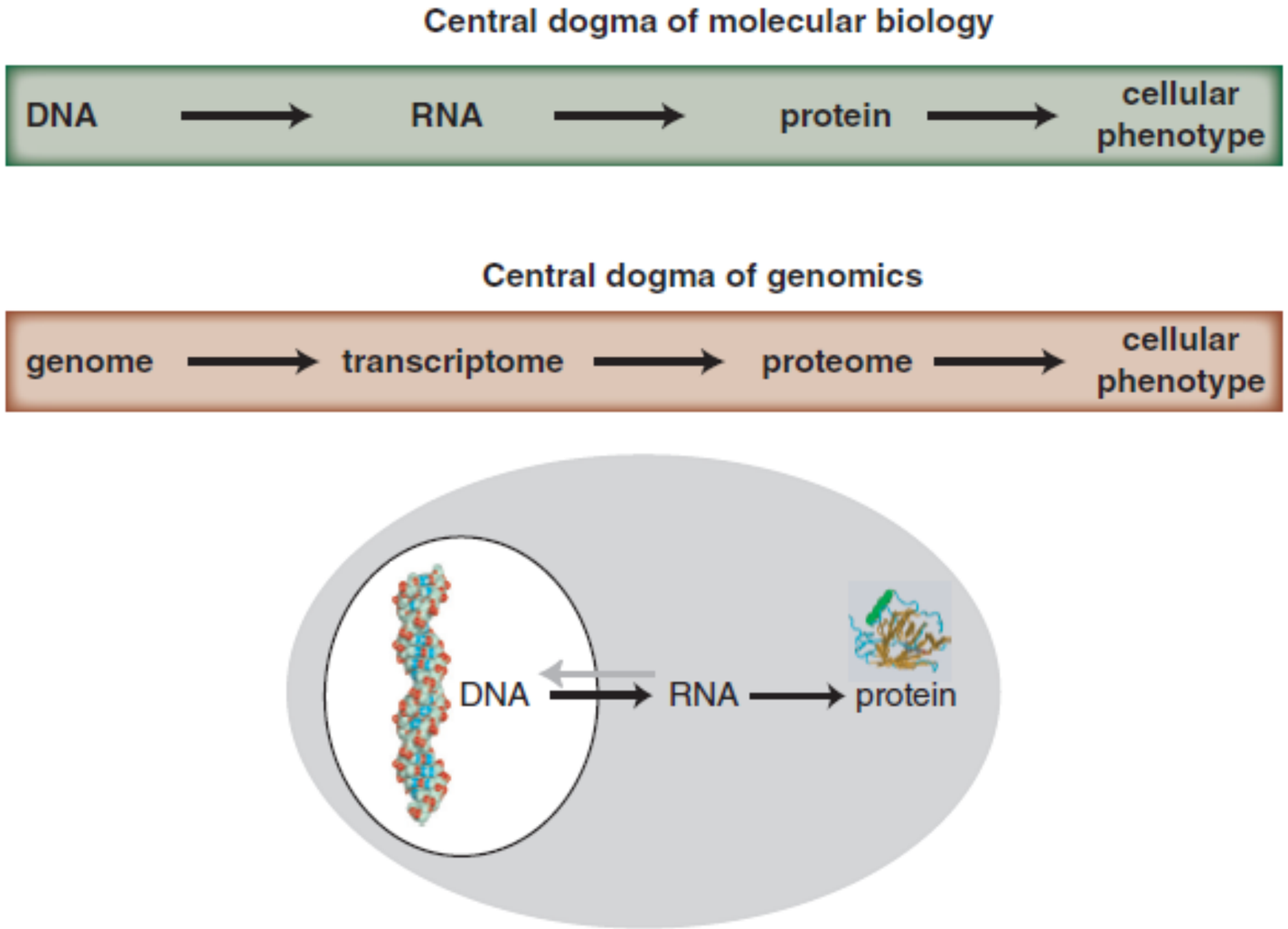




A - adenine
C - cytosine
G - guanine
T - thymine
U - uracil

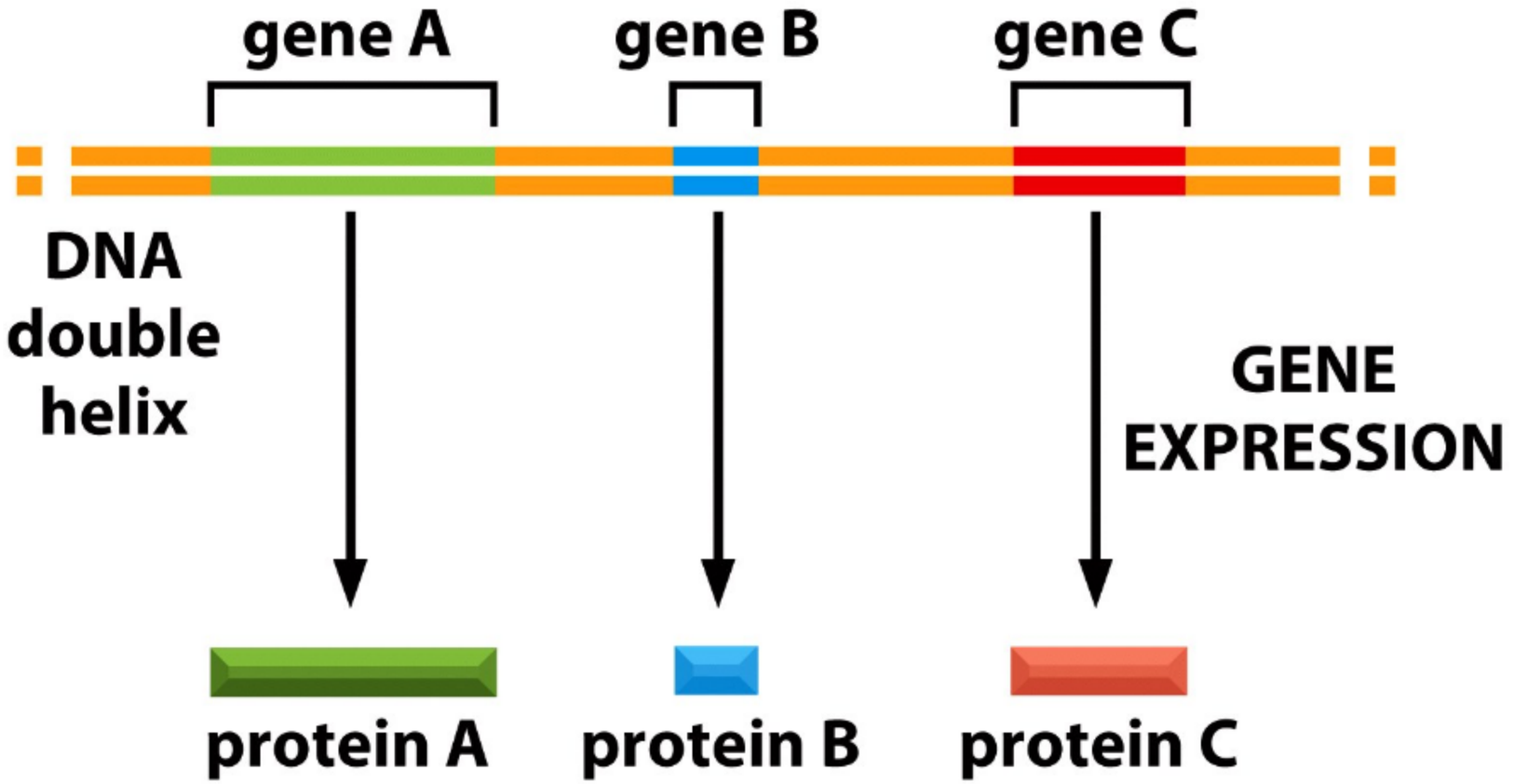
ian.simpson@ed.ac.uk

The "Central Dogma"



ian.simpson@ed.ac.uk

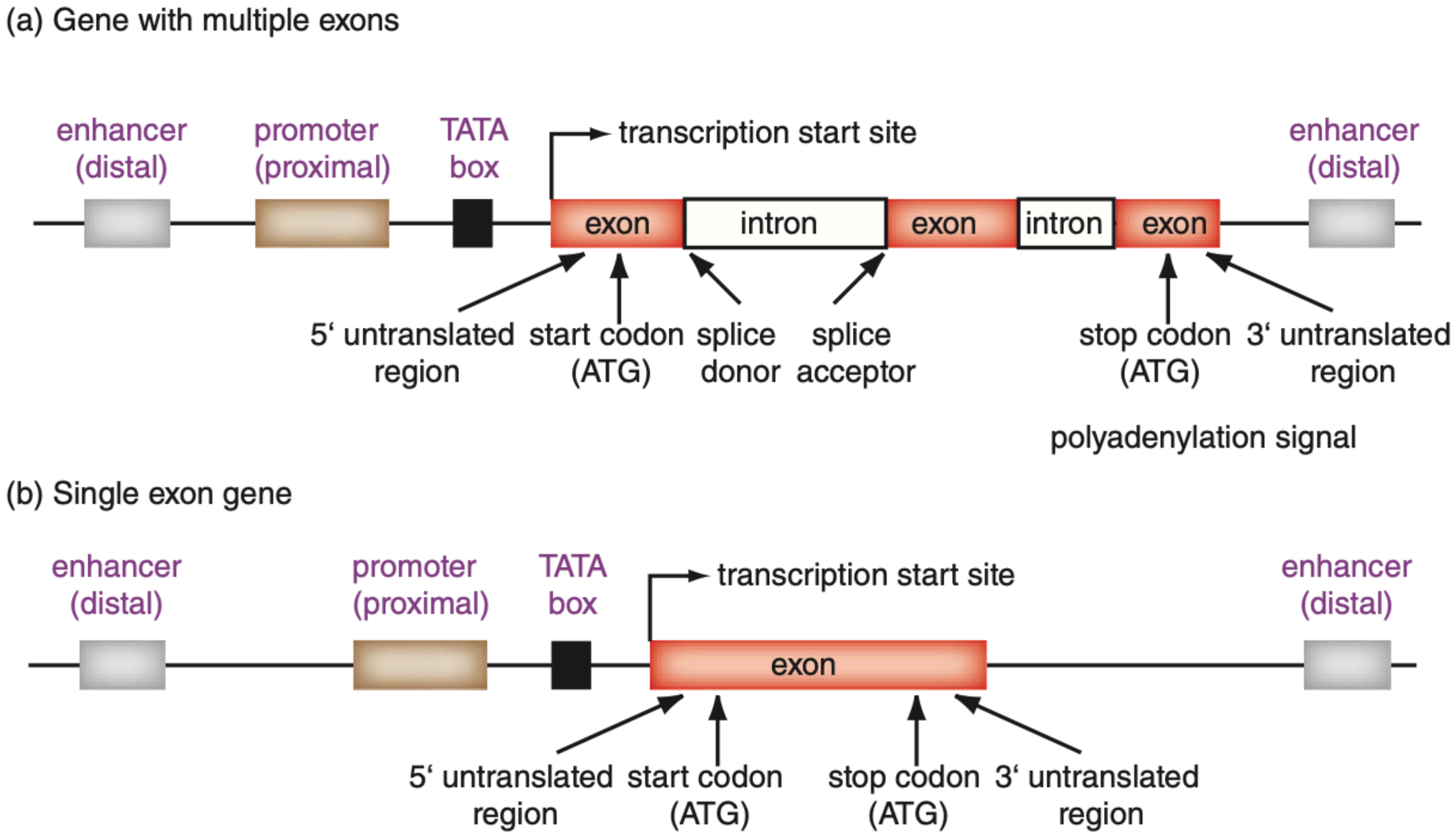
Protein Coding Genes



Alberts, B. (2014). *Essential cell biology*. New York: Garland Science.

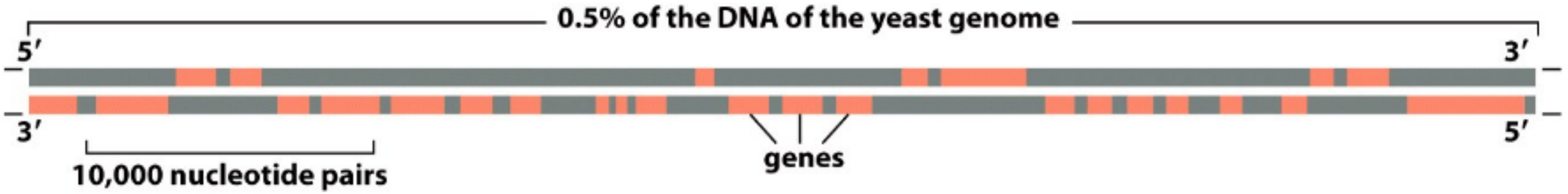
ian.simpson@ed.ac.uk

Gene Structures (Eukaryotic)



ian.simpson@ed.ac.uk

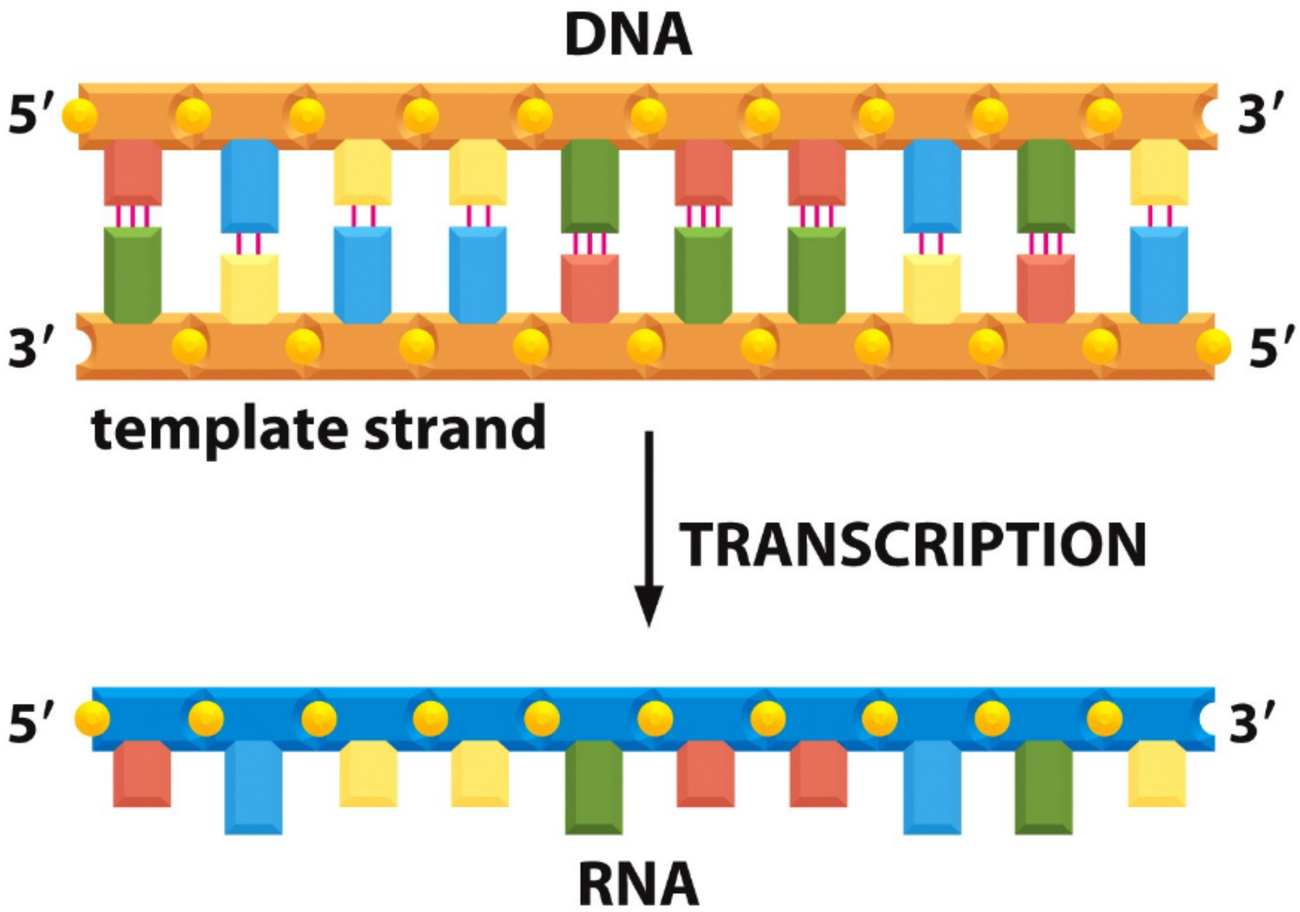
DNA Strandedness



Alberts, B. (2014). *Essential cell biology*. New York: Garland Science.

ian.simpson@ed.ac.uk

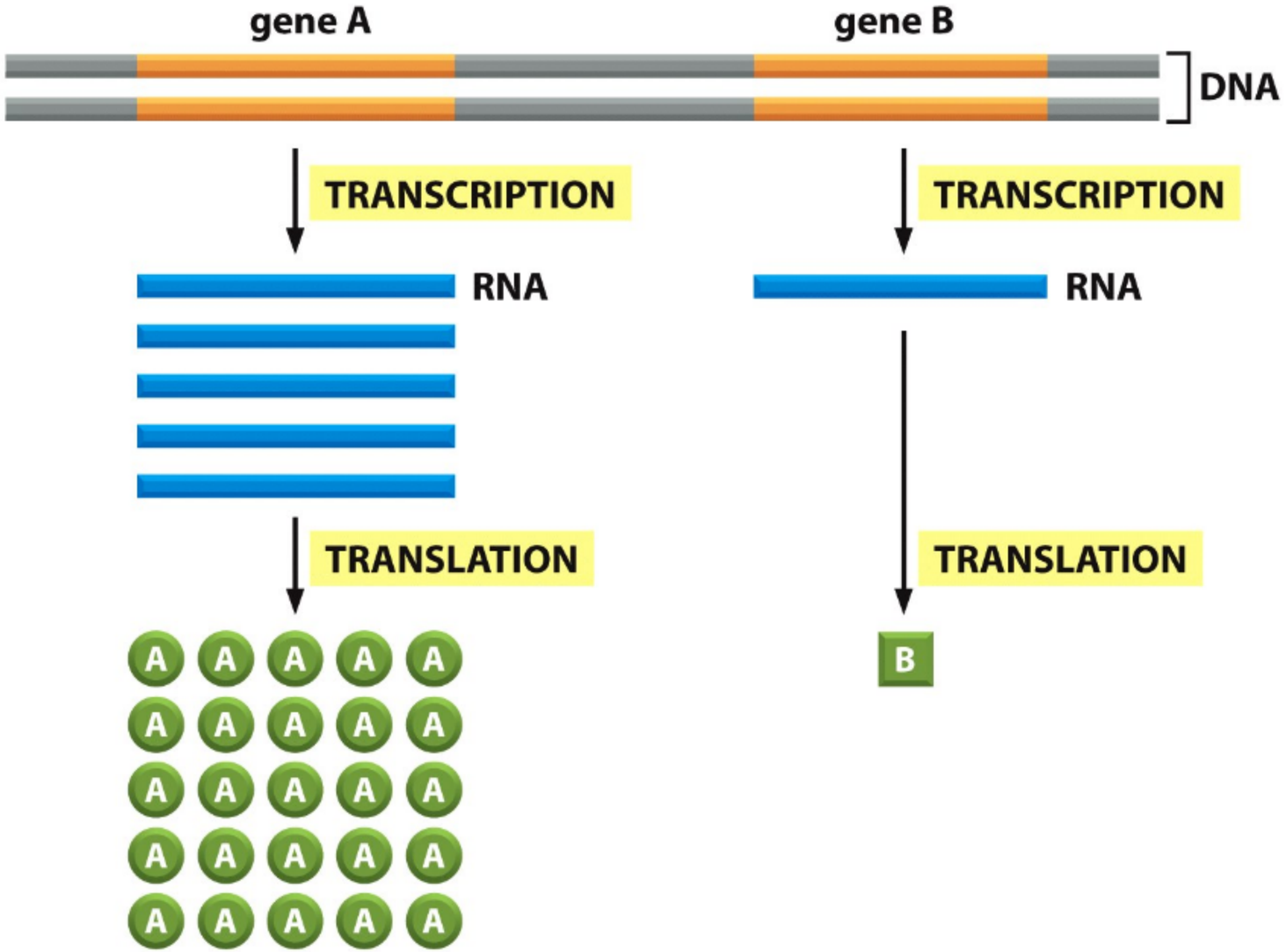
Transcription



Alberts, B. (2014). *Essential cell biology*. New York: Garland Science.

ian.simpson@ed.ac.uk

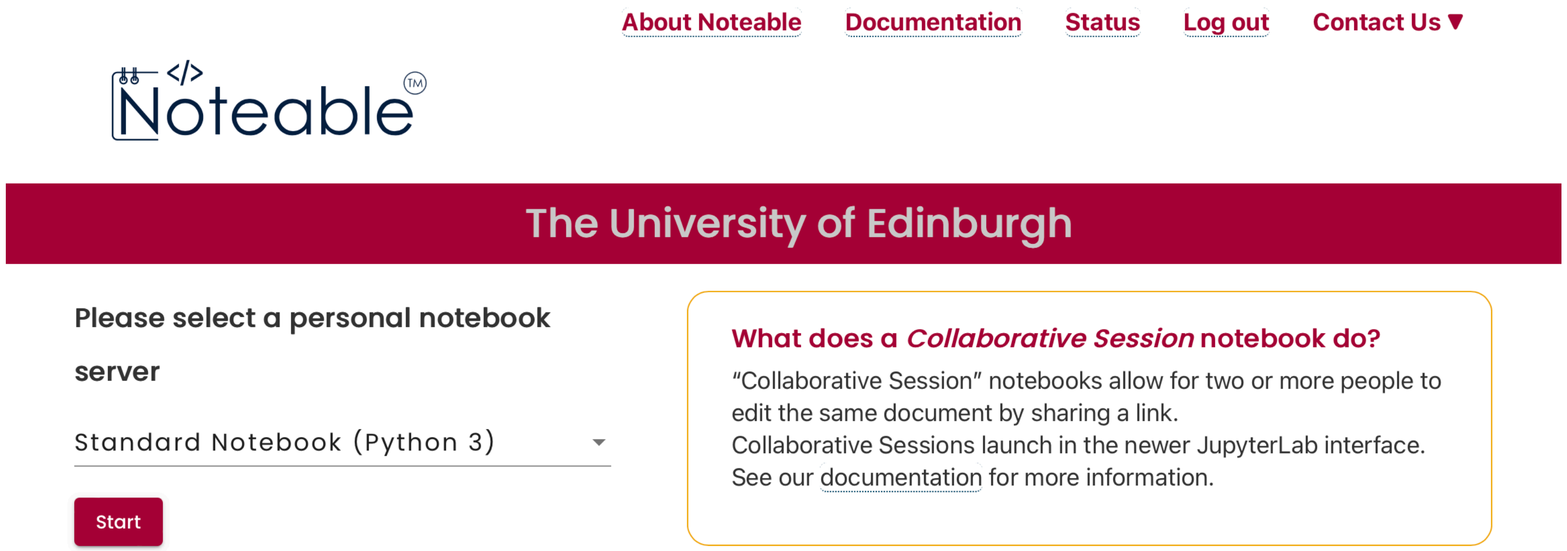
Translation



Alberts, B. (2014). *Essential cell biology*. New York: Garland Science.

ian.simpson@ed.ac.uk

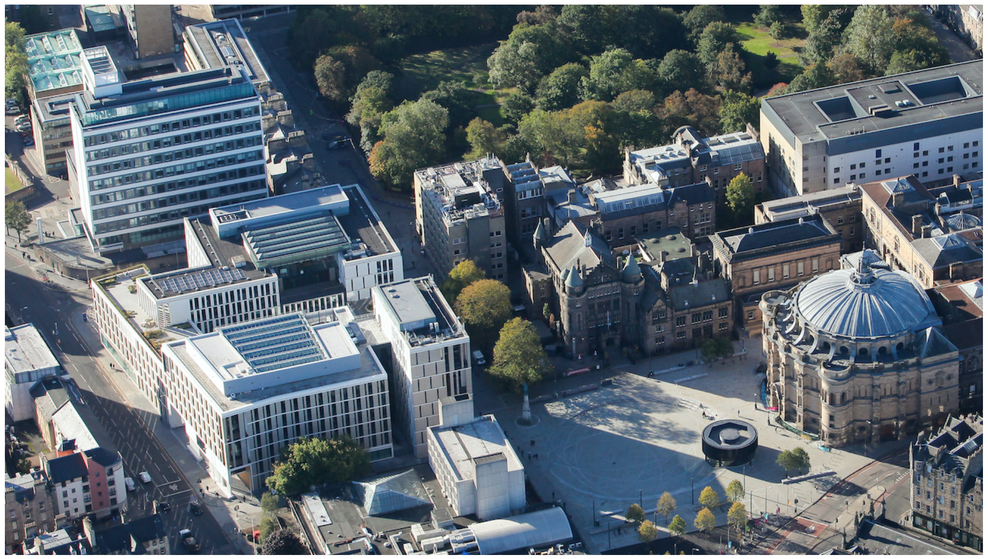
Noteable & Jupyter Notebooks



https://noteable.edina.ac.uk/

ian.simpson@ed.ac.uk

Bioinformatics at Edinburgh



Selected Research & Facilities
http://web.inf.ed.ac.uk/anc/research/bioinformatics
https://www.ed.ac.uk/biology/research/themes/research-theme-systems
https://www.ed.ac.uk/mrc-human-genetics-unit/research/semple-group
http://bifx-core.bio.ed.ac.uk/wiki/index.php/Main_Page
https://genomics.ed.ac.uk

Seminar series:
https://www.ed.ac.uk/biology/synthsys/news-and-events/events
http://www.ed.ac.uk/biology/news-events/events-and-seminars
https://www.ed.ac.uk/usher/news-events/seminars

Selected Topic Areas

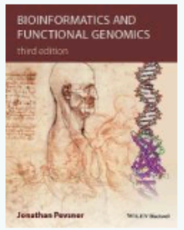
- Database integration
- Data ontologies and provenance
- Evolutionary and genetic computation
- Gene expression databases
- High performance data structures for semi-structured data
- Microarray data analysis
- Epigenomics
- Natural language and bio-text mining
- Neural computation, visualisation and simulation
- Protein complex modelling
- Systems Biology, modelling
- Synthetic Biology
- Machine learning
- Medical Image Analysis

ian.simpson@ed.ac.uk

Next Week

Week 2 - Pairwise Sequence Alignment

If you would like to read ahead then please look at the following chapter, if not we will cover this in next week's lecture
This is available from the Bio1 course "Resource List"



BOOK **Bioinformatics and functional genomics**

Pevsner, Jonathan, 1961-, 3rd ed.., Chichester, West Sussex, UK ; Hoboken, NJ, USA, John Wiley and Sons, Incorporated, 2015

Note: *Read Chapter 3, "Pairwise Sequence Alignment"*

Add tags to item

Complete

ian.simpson@ed.ac.uk