

Welcome to MSc Bioinformatics

2024

Prof Conrad Bessant
MSc Bioinformatics Programme Leader
c.bessant@qmul.ac.uk

A little biology ...

The human genome

YouTube Music | Course: MSc Bioinfo | Homo_sapiens - En...

ensembl.org/Homo_sapiens/Info/Annotation

Ensembl BLAST/BLAT | VEP | More

Search Human...

Human (GRCh38.p13)

Human assembly and gene annotation

Assembly

This site provides a data set based on the December 2013 *Homo sapiens* high coverage assembly GRCh38 from the [Genome Reference Consortium](#). This assembly is used by UCSC to create their hg38 database. The data set consists of gene models built from the genome alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exonerate.

This release of the assembly has the following properties:

- contig length total 3.4 Gb.
- chromosome length total 3.1 Gb (excluding haplotypes).

It also includes 261 alt loci scaffolds, mainly in the LRC/KIR complex on chromosome 19 (35 alternate sequence representations) and the [MHC region on chromosome 6](#) (7 alternate sequence representations).

[Watch a video on YouTube](#) about patches and haplotypes in the Human genome.

Patches

As the GRC maintains and improves the assembly, patches are being introduced. Currently, assembly patches are of two types:

- Novel patch: new sequences that add alternative sequence at a loci and will remain as haplotypes in the next major assembly release by GRC
- Fix patch: sequences that correct the reference sequence and will replace the given region of the reference assembly at the next major assembly release by GRC.

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart

Gene annotation

The Ensembl human gene annotations have been updated using Ensembl's automatic annotation pipeline. The updated

More information

General information about this species can be found in [Wikipedia](#).

Statistics

Summary

Assembly	GRCh38.p13 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.28 , Dec 2013
Base Pairs	3,096,649,726
Golden Path Length	3,096,649,726
Assembly provider	Genome Reference Consortium
Annotation provider	Ensembl
Annotation method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Apr 2022
Database version	107.38
Gencode version	GENCODE 41

Gene counts (Primary assembly)

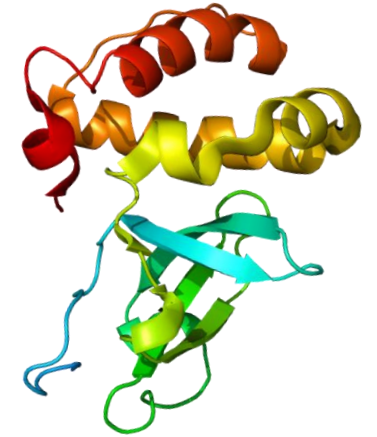
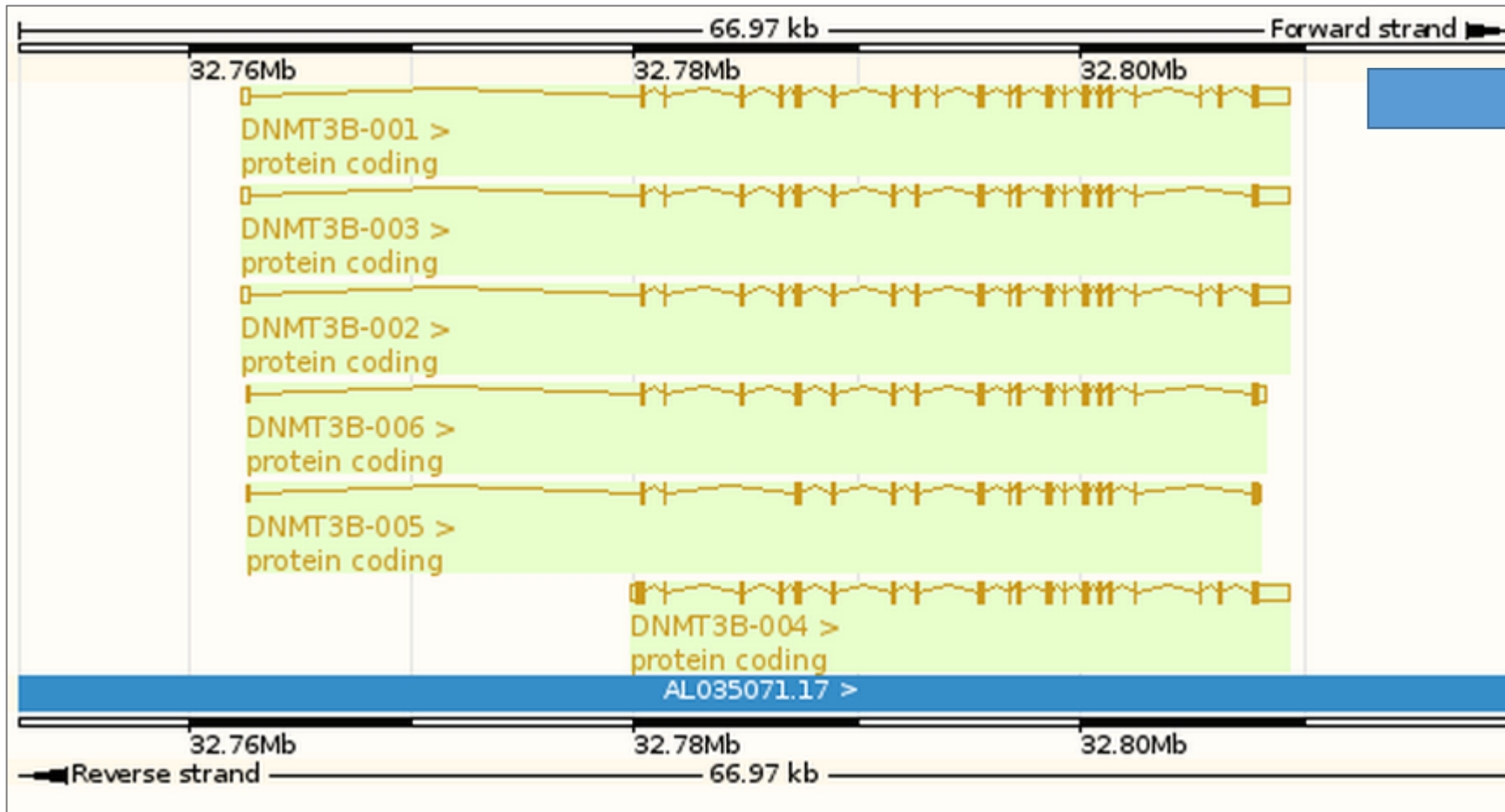
Coding genes	19,804 (excl 647 readthrough)
Non coding genes	25,134
Small non coding genes	4,864
Long non	18,049

DNMT3B: Transcript

ATGGAACCAAGTCCAGCCTCCAAGCTTGGAAAAGCATGAAGGGAGACACCAGGCATCTCAATGGAGAGG
AGGACGCCGGCGGGAGGGAAGACTCGATCCTCGTCAACGGGGCCTGCAGCGACCAGTCCTCCGACTCGCC
CCCAATCCTGGAGGCTATCCGCACCCCGGAGATCAGAGGCCGAAGATCAAGCTCGCGACTCTCCAAGAGG
GAGGTGTCCAGTCTGCTAAGCTACACACAGGACTTGACAGGCGATGGCGACGGGGAAGATGGGGATGGCT
CTGACACCCCAGTCATGCCAAAGCTCTTCCGGGAAACCAGGACTCGTTCAGAAAGCCCAGCTGTCCGAAC
TCGAAATAACAACAGTGTCTCCAGCCGGGAGAGGCACAGGCCTTCCCCACGTTCCACCCGAGGCCGGCAG
GGCCGCAACCATGTGGACGAGTCCCCCGTGGAGTTCCCGGCTACCAGGTCCCTGAGACGGCGGGCAACAG
CATCGGCAGGAACGCCATGGCCGTCCCCTCCCAGCTCTTACCTTACCATCGACCTCACAGACGACACAGA
GGACACACATGGGACGCCCCAGAGCAGCAGTACCCCTACGCCC GCCTAGCCCAGGACAGCCAGCAGGGG
GGCATGGAGTCCCCGCAGGTGGAGGCAGACAGTGGAGATGGAGACAGTTCAGAGTATCAGGATGGGAAGG
AGTTTGAATAGGGGACCTCGTGTGGGGAAAGATCAAGGGCTTCTCCTGGTGGCCCGCCATGGTGGTGTG
TTGGAAGGCCACCTCCAAGCGACAGGCTATGTCTGGCATGCGGTGGGTCCAGTGGTTTGGCGATGGCAAG
TTCTCCGAGGTCTCTGCAGACAAACTGGTGGCACTGGGGCTGTTTACGCCAGCACTTTAATTTGGCCACCT
TCAATAAGCTCGTCTCCTATCGAAAAGCCATGTACCATGCTCTGGAGAAAGCTAGGGTGCAGACTGGCAA
GACCTTCCCCAGCAGCCCTGGAGACTCATTGGAGGACCAGCTGAAGCCCATGTTGGAGTGGGCCACGGG
GGCTTCAAGCCCACTGGGATCGAGGGCCTCAAACCCAACAACACGCAACCAGAGAACAAGACTCGAAGAC
GCACAGCTGACGACTCAGCCACCTCTGACTACTGCCCCGCACCCAAGCGCCTCAAGACAAATTGCTATAA
CAACGGCAAAGACCAGGGGGATGAAGATCAGAGCCGAGAACAAATGGCTTCAGATGTTGCCAACACAAG
AGCAGCCTGGAAGATGGCTGTTTGTCTTGTGGCAGGAAAAACCCCGTGTCTTCCACCTCTCTTTGAGG
GGGGGCTCTGTCAGACATGCCGGGATCGCTTCCCTTGAGCTGTTTTACATGTATGATGACGATGGCTATCA
GTCTTACTGCACTGTGTGCTGCGAGGGCCGAGAGCTGCTGCTTTCAGCAACACGAGCTGCTGCCGGTGT
TTCTGTGTGGAGTGCCTGGAGGTGCTGGTGGGCACAGGCACAGCGGCCGAGGCCAAGCTTCAGGAGCCCT
GGAGCTGTTACATGTGTCTCCCGCAGCGCTGTTCATGGCGTCCCTGCGGCGCCGGAAGGACTGGAACGTGCG
CCTGCAGGCCTTCTTACCAGTGACACGGGGCTTGAATATGAAGCCCCCAAGCTGTACCCTGCCATTCCC
GCAGCCC GAAGGCGGCCCATTCGAGTCCTGTCATTGTTTGATGGCATCGCGACAGGCTACCTAGTCCTCA
AAGAGTTGGGCATAAAGGTAGGAAAGTACGTCGCTTCTGAAGTGTGTGAGGAGTCCATTGCTGTTGGAAC
CGTGAAGCACGAGGGGAATATCAAATACGTGAACGACGTGAGGAACATCACAAAGAAAAATATTGAAGAA
TGGGGCCCATTTGACTTGGTGAATTGGCGGAAGCCCATGCAACGATCTCTCAAATGTGAATCCAGCCAGGA
AAGGCCTGTATGAGGGTACAGGCCGGCTCTTCTTTCGAATTTTACCACCTGCTGAATTACTCACGCCCCAA
GGAGGGT GATGACCGGCCGTTCTTCTGGATGTTT GAGAATGTTGTAGCCATGAAGGTTGGCGACAAGAGG
GACATCTCACGGTTCCTGGAGTGTAATCCAGTGATGATTGATGCCATCAAAGTTTCTGCTGCTCACAGGG
CCCGATACTTCTGGGGCAACCTACCCGGGATGAACAGGCCCGTGATAGCATCAAAGAATGATAAACTCGA
GCTGCAGGACTGCTTGAATACAATAGGATAGCCAAGTTAAAGAAAGTACAGACAATAACCACCAAGTCG
AACTCGATCAAACAGGGGAAAAACCAACTTTTCCCTGTTGTTCATGAATGGCAAAGAAGATGTTTTGTGGT
GCACTGAGCTCGAAAGGATCTTTGGCTTTCCCTGTGCACTACACAGACGTGTCCAACATGGGCCGTGGTGC
CCGCCAGAAGCTGCTGGGAAGGTCCTGGAGCGTGCCTGTCATCCGACACCTCTTCGCCCTCTGAAGGAC
TACTTTGCATGTGAATAG

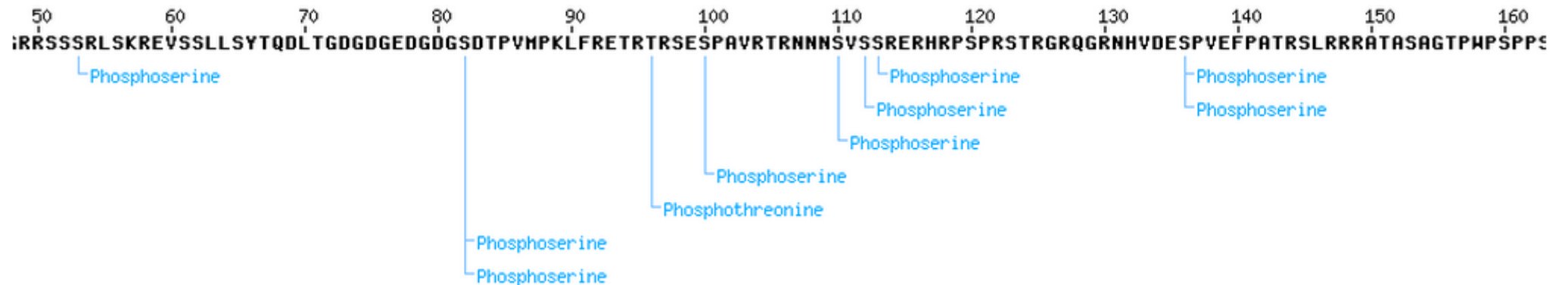
Added complexity: Alternate splicing

Example: The DNMT3B gene produces six distinct proteins.



More complexity: Proteins modified by PTM

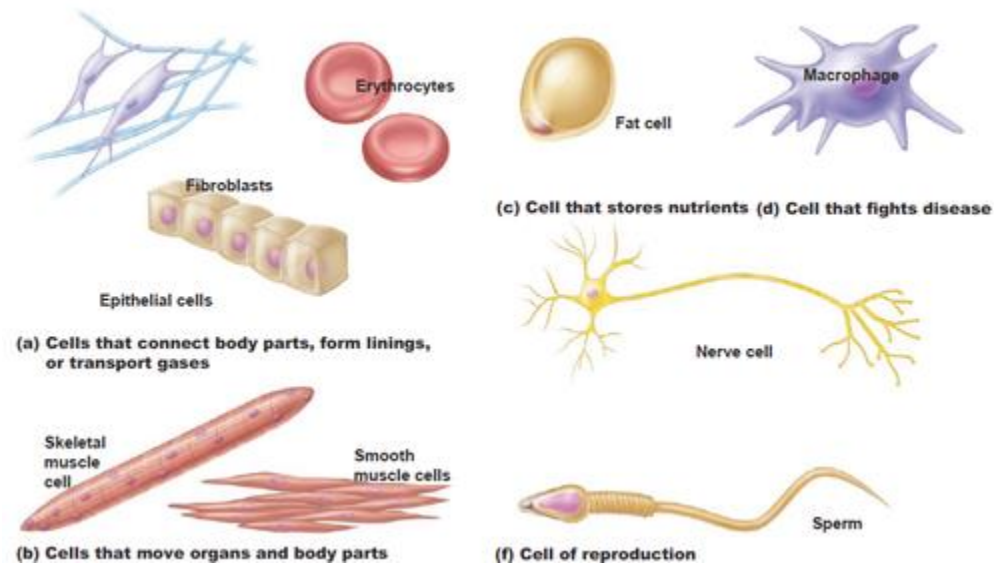
DNMT3B post translational modifications (PTMs):



Even more complexity: Abundance variation

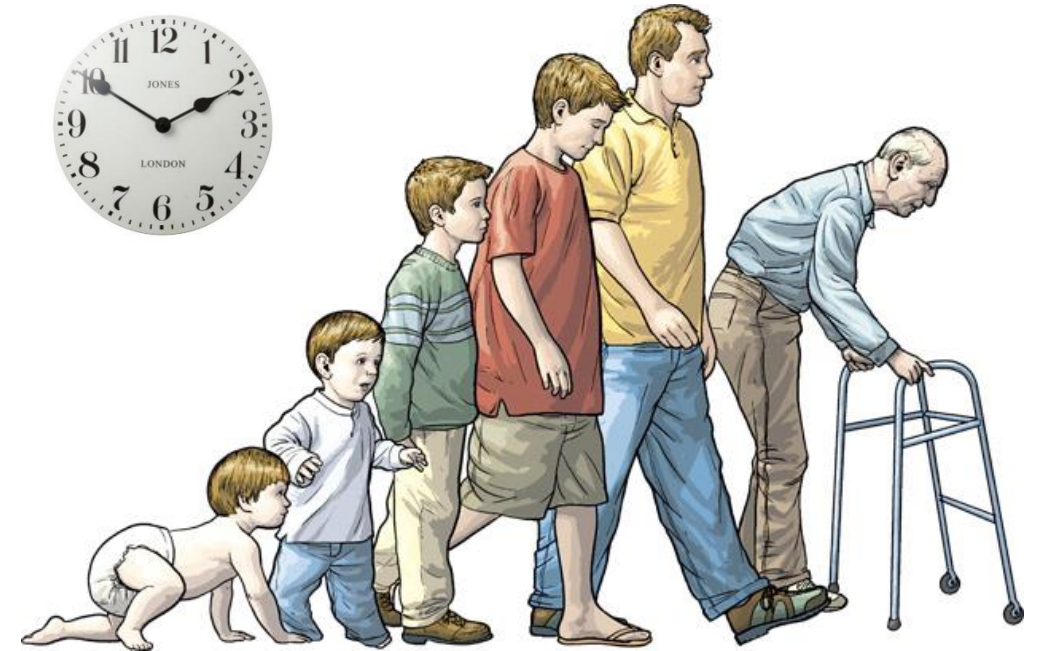
- Expression of genes (and hence proteins) varies according to:

Cell Type



anatomyandphysiology.com

Time



www.nicaraguaenvivo.com

High throughput analytical techniques

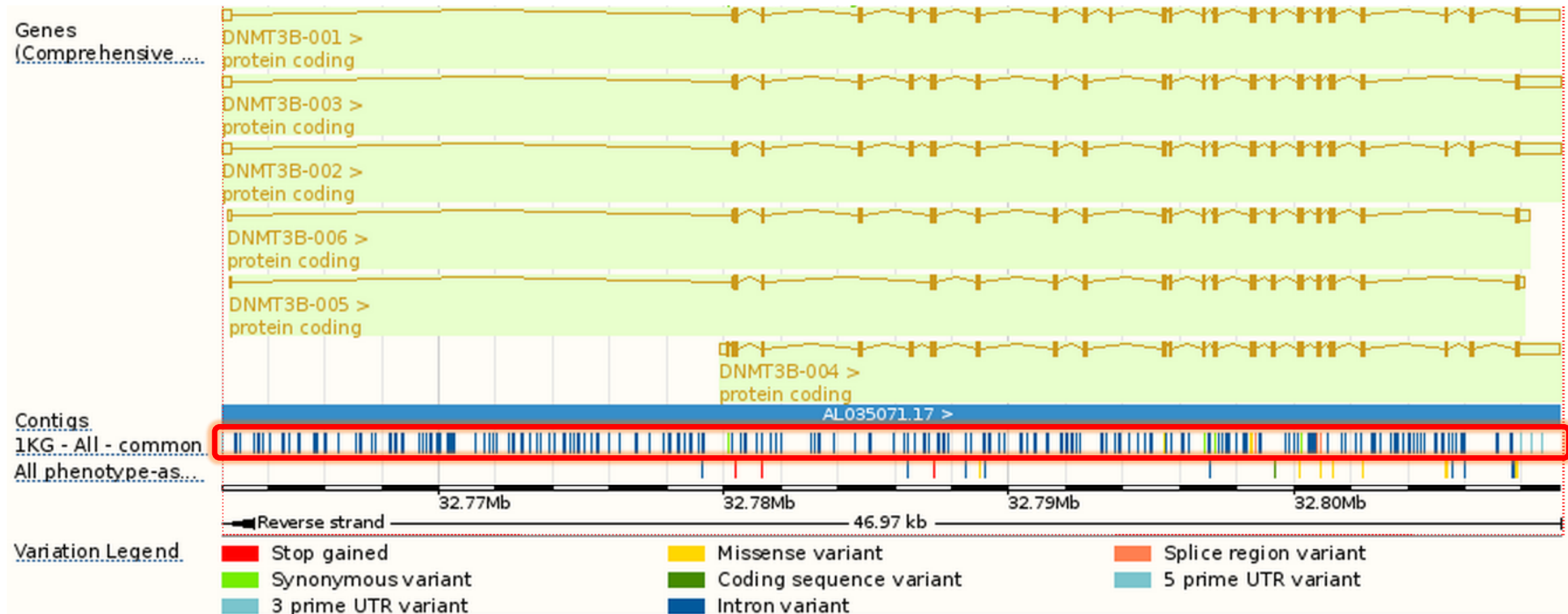


Sequencers to quantify gene expression (RNA-seq).

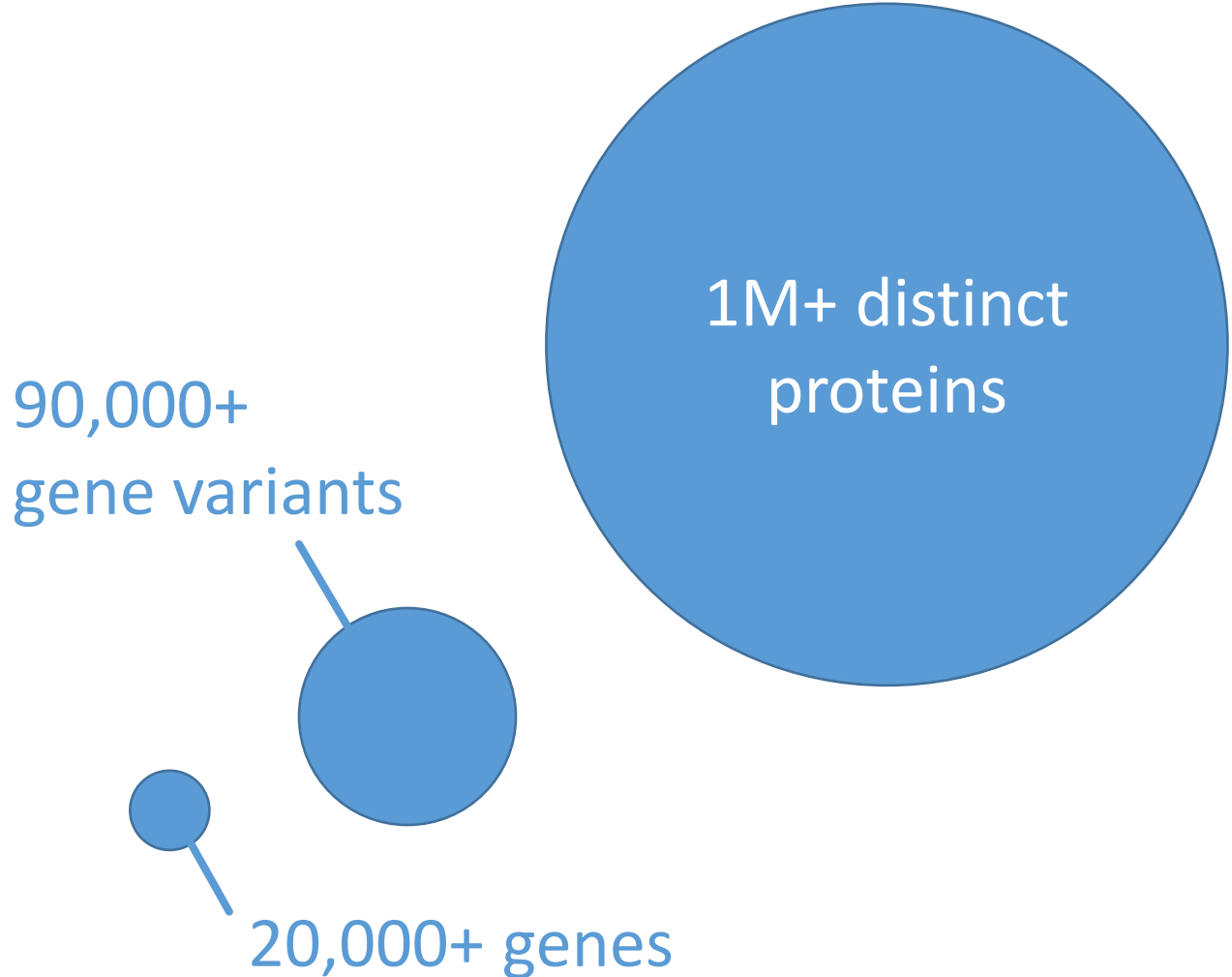


Mass spectrometers to quantify proteins and metabolites.

And there's more complexity: Genome variation



Just in humans ...



billions of different
states and individual
variations

Bioinformatics in 2024

High throughput laboratory methods are producing a tsunami of biological data.

Not only DNA sequence data. Also RNA transcripts, data on proteins, metabolites, etc. etc.

The bottleneck in making new discoveries in the biosciences has moved from the lab to the computer.

Good bioinformaticians are in high demand.



A modern sequencing lab

Nanopore Sequencing

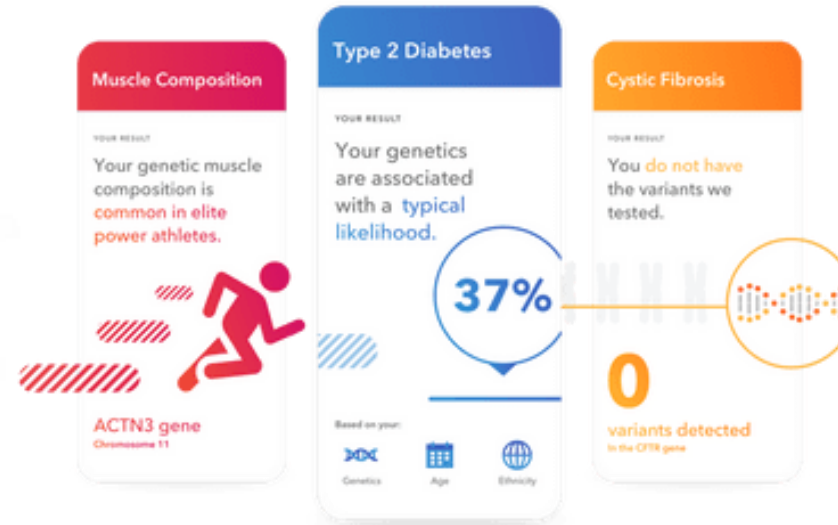
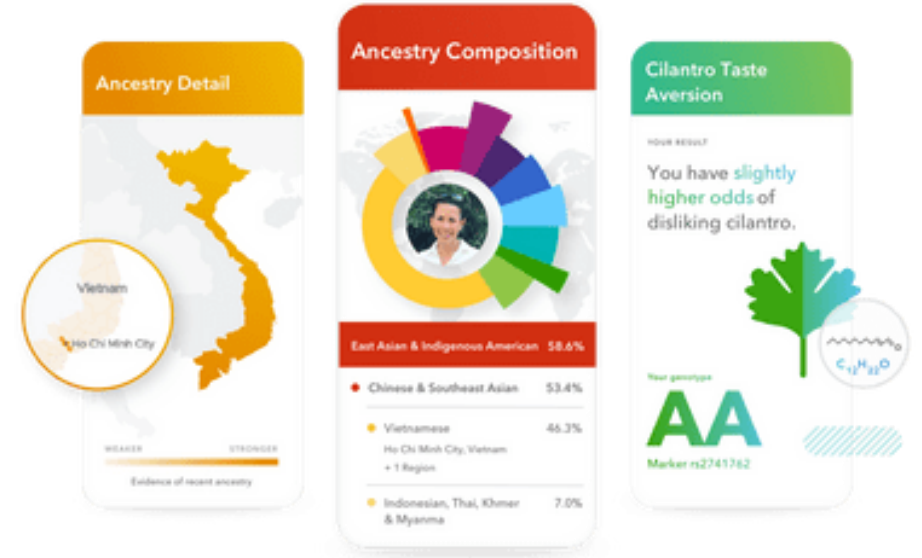
- ✓ Portable
- ✓ Low cost
- ✓ Whole genome
- ✓ Long read lengths



Personal genotyping



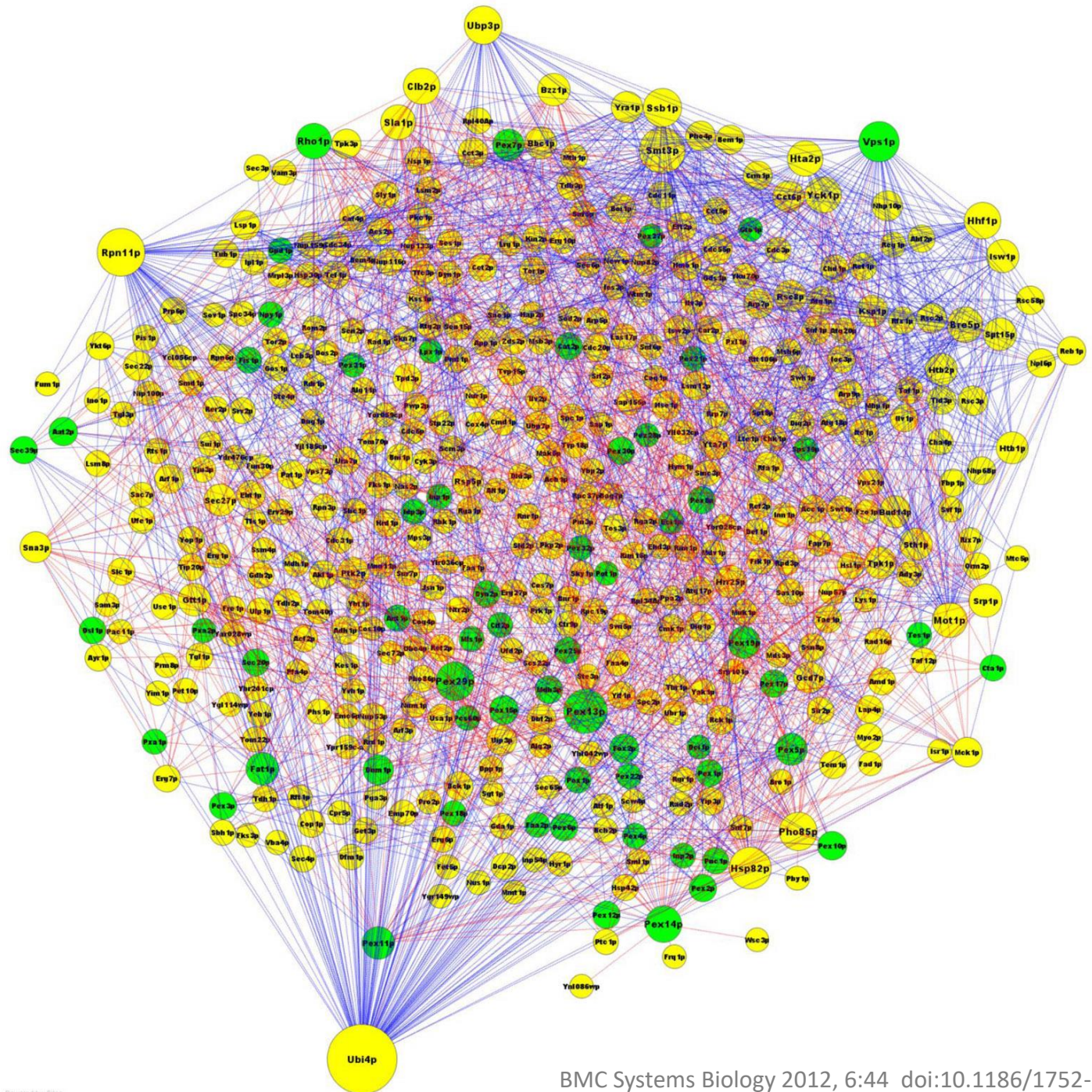
£159



Systems biology

Aims to understand what is actually happening at a molecular level.

Potentially millions of interactions between proteins, metabolites and other molecules.



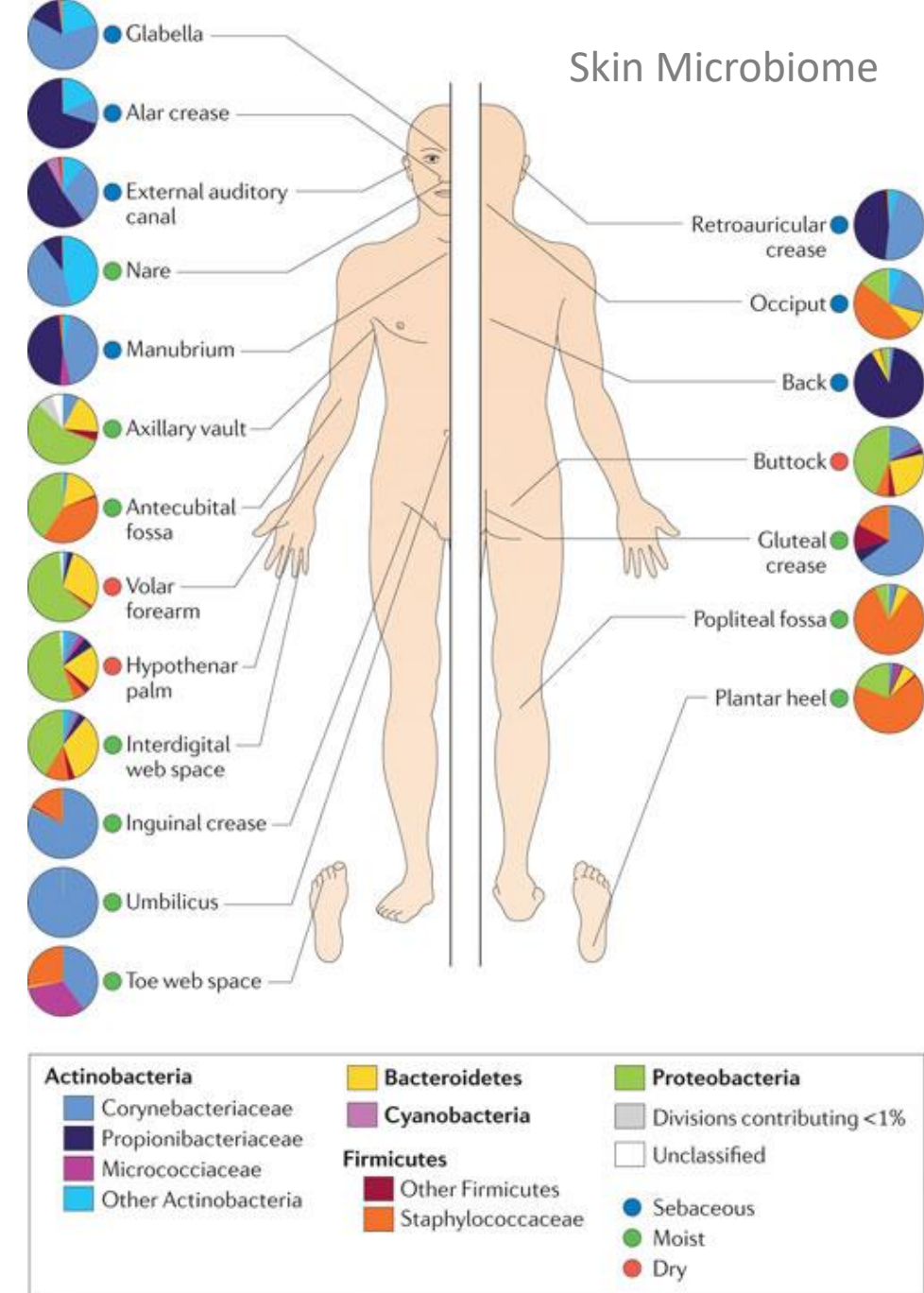
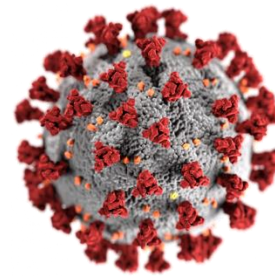
Even more complexity ...

We are more microbe than human:

- 100s of other genomes to consider.

Many other important species to study:

- Organisms of environmental and ecological importance.
- Pathogens.
- Food.
- Industrial organisms.



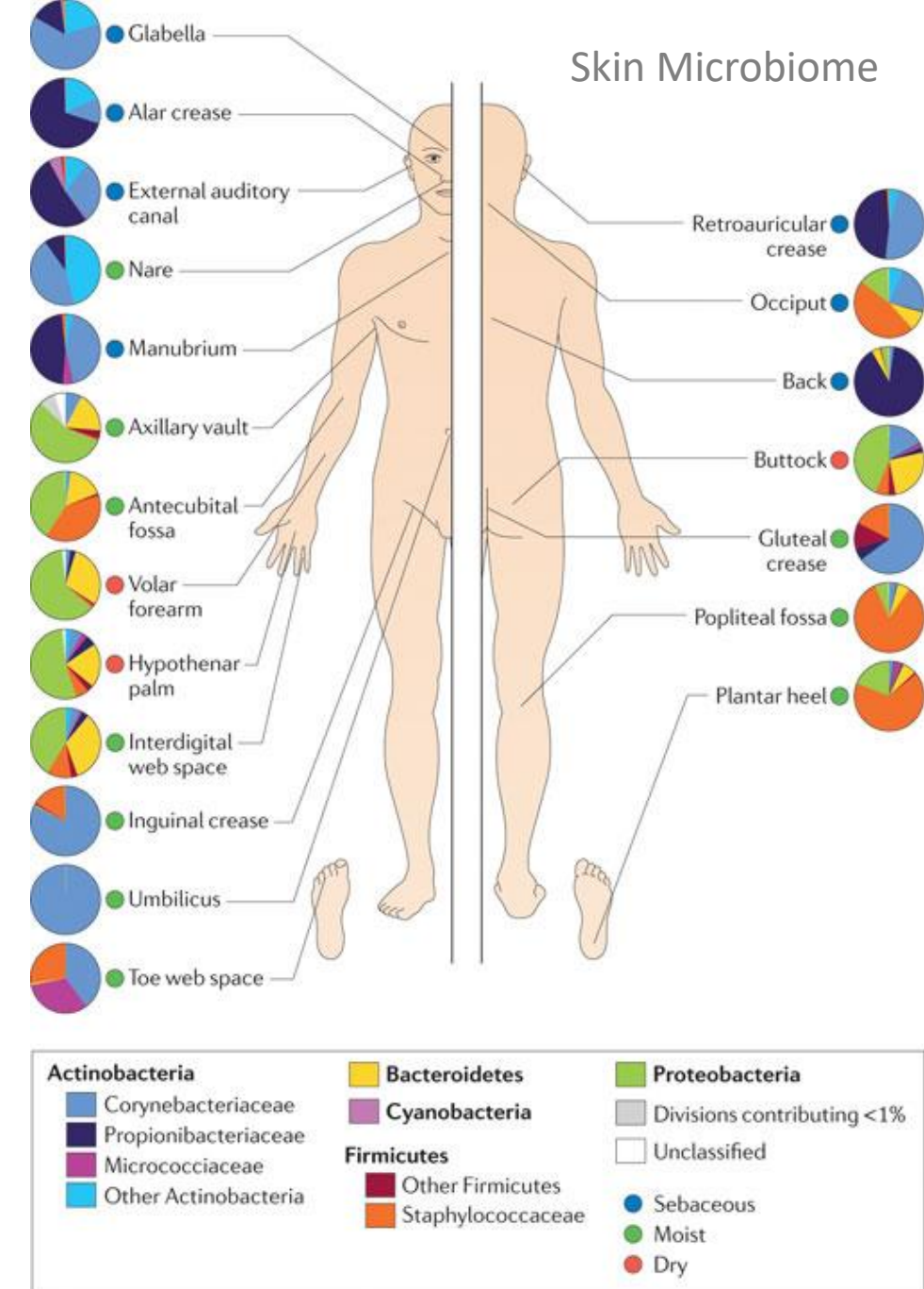
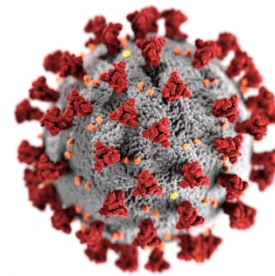
Even more complexity ...

We are more microbe than human:

- 100s of other genomes to consider.

Many other important species to study:

- Organisms of environmental and ecological importance.
- Pathogens.
- Food.
- Industrial organisms.



We need computation ... a lot of computation

Computational biology

Practicalities:

- Data storage and management.
- High performance computing.
- Visualisation.

Research:

- Application of existing algorithms to new data, to answer biological questions.
- Development of new algorithms.



2000 -



2010 -



2020 -

Essential computational biology skills

Technical

Molecular biology

Analytical science

Data analysis

Coding (programming)

Transferable

Scientific method

Teamwork

Communication

Attention to detail

Your MSc

Taught Modules

W1-3: Unix and analysis of large genomic datasets

W4-6: Coding for bioscientists

W7-9: Statistics for bioinformatics

W10-12: AI and data science in biology

Individual Research Project (5½ months)

A focussed piece of computer-based research within a research group at QMUL (or elsewhere)

23 Sept
2024

Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug



Yannick
Wurm



Fabrizio
Smeraldi



Matteo
Fumagalli



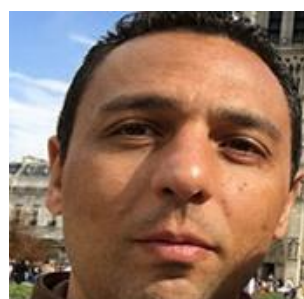
Conrad
Bessant

Group Project (six weeks)

Work in a team of 4-5 people to produce a piece of software to solve a biological problem

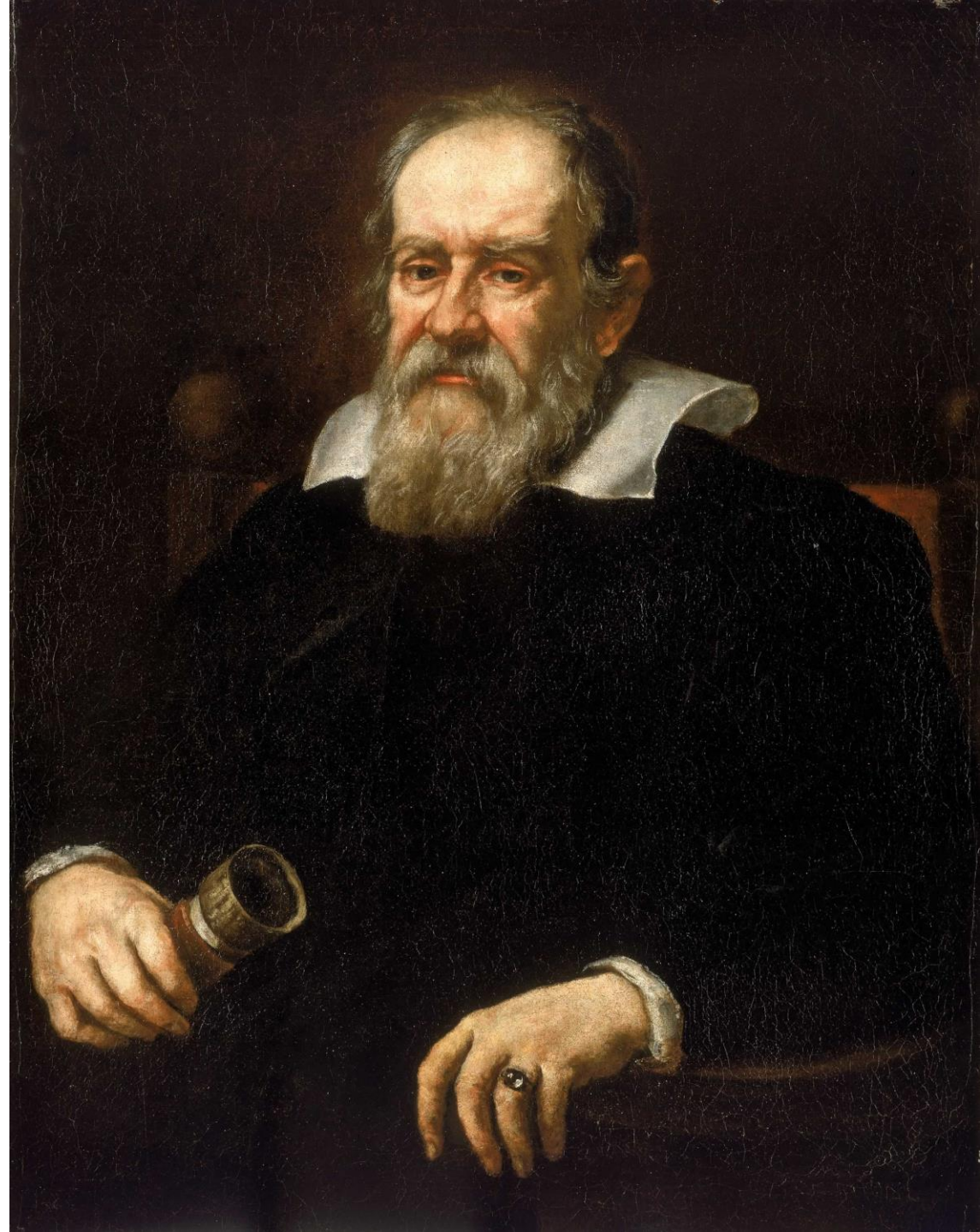
Examples of this previous MSc research projects

1. Tracking ongoing genome evolution using **single cell sequencing**
2. Can **heritable DNA methylations** increase disease resistance?
3. Inference of genomic signatures of natural selection using **deep learning**
4. Using **machine learning** for thyroid disease prediction and multimorbidity
5. Mapping **epigenomic data** to repetitive elements
6. **Integrative analysis** of miRNA and mRNA expression profiles in arthritis
7. Exploring how genome size impacts **plant extinction risk**
8. Multi-modal analysis of single cell data using **machine learning** methods
9. Systematic identification of biochemical **networks** in **cancer** cells



“We cannot teach
people anything; we
can only help them
discover it within
themselves”

Galileo Galilei
(1564-1642)



Less of this ...



More of this ...



King Size Homer (Episode 3F05), Simpsons Season 7, 1995.

A word about cheating

Plagiarism, collusion, outsourcing,
excessive use of chatGPT.

Don't do it!

**Ask lectures, supervisor
or student support for help instead.**

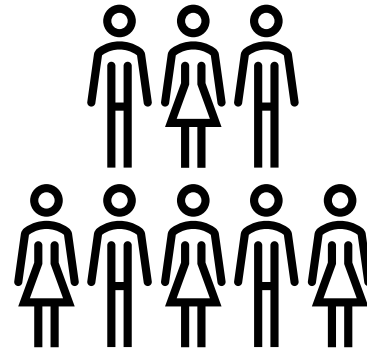
Supporting your learning

Programme Leader



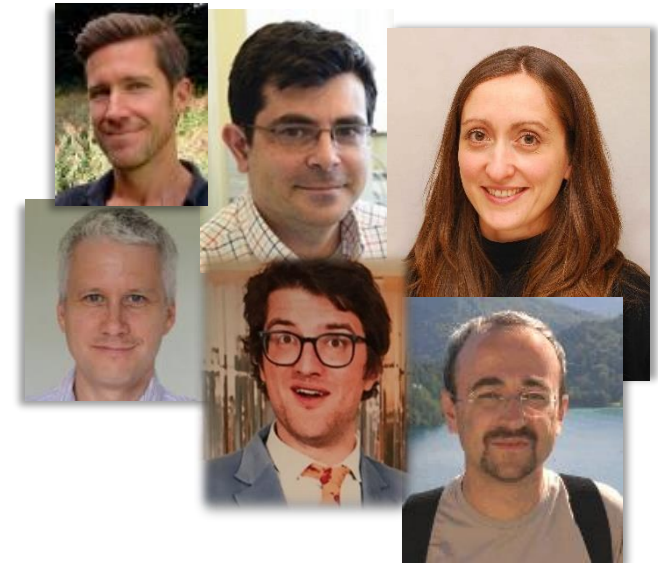
Conrad Bessant
c.bessant@qmul.ac.uk

Admin



SBBS Admin Team
sbbs-office@qmul.ac.uk

So Many Others

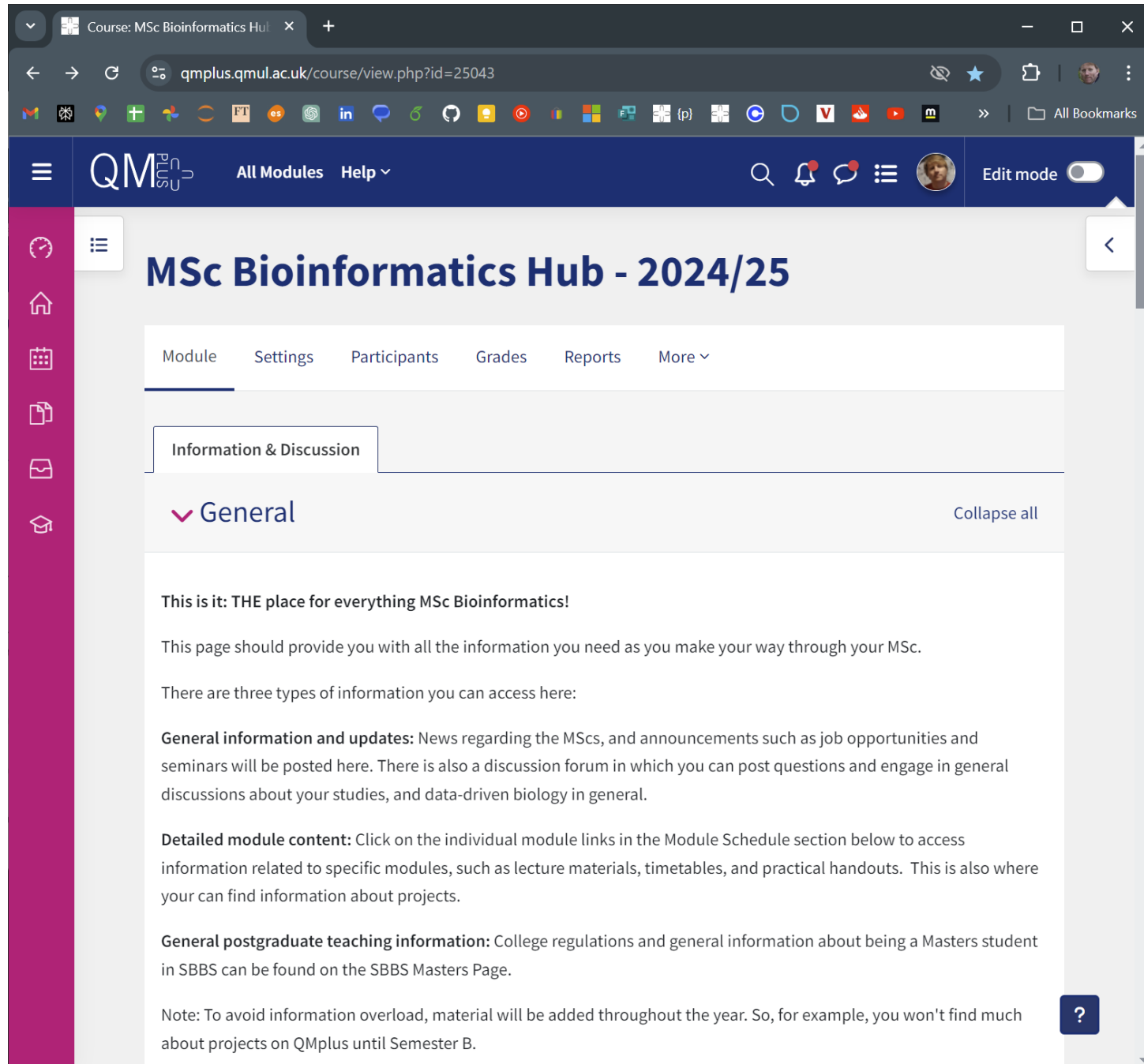


Module Leads
+ Lecturers
+ Demonstrators
+ Project Supervisors

MSc Bioinformatics Hub

bit.ly/bix24

(you need to be logged into QMplus to access)



The screenshot shows a web browser window displaying the MSc Bioinformatics Hub page on the QMplus platform. The browser's address bar shows the URL qmplus.qmul.ac.uk/course/view.php?id=25043. The page header includes the QMplus logo, navigation links for 'All Modules' and 'Help', a search bar, and an 'Edit mode' toggle. The main content area is titled 'MSc Bioinformatics Hub - 2024/25' and features a navigation menu with options: 'Module', 'Settings', 'Participants', 'Grades', 'Reports', and 'More'. The 'Information & Discussion' tab is selected, and the 'General' section is expanded. The page content includes a welcome message, a list of information types accessible, and detailed instructions for finding module content and postgraduate teaching information. A help icon is visible in the bottom right corner.

Course: MSc Bioinformatics Hub

qmplus.qmul.ac.uk/course/view.php?id=25043

QMplus All Modules Help

MSc Bioinformatics Hub - 2024/25

Module Settings Participants Grades Reports More

Information & Discussion

General Collapse all

This is it: THE place for everything MSc Bioinformatics!

This page should provide you with all the information you need as you make your way through your MSc.

There are three types of information you can access here:

General information and updates: News regarding the MScs, and announcements such as job opportunities and seminars will be posted here. There is also a discussion forum in which you can post questions and engage in general discussions about your studies, and data-driven biology in general.

Detailed module content: Click on the individual module links in the Module Schedule section below to access information related to specific modules, such as lecture materials, timetables, and practical handouts. This is also where you can find information about projects.

General postgraduate teaching information: College regulations and general information about being a Masters student in SBBS can be found on the SBBS Masters Page.

Note: To avoid information overload, material will be added throughout the year. So, for example, you won't find much about projects on QMplus until Semester B.

Get ready for next week ...

Unix and Analysis of Large Genomic Datasets
(BIO726P)

Starts on Monday

Head over to the BIO726P QMplus page **from Friday** to get up to speed – there are things you need to do!

First live session in on Monday: see Qmplus for details.



Tower Hamlets
History Library...

The Octagon
At Queen Mary...

University
of London

**Room QB209
Queens Building**

G. E. Fogg Building

G. E. Fogg Building

Queens' Building

QMUL School of
Engineering and...

A17

Harford St

A word about attendance

We expect everyone to be present in person at the hands-on practical sessions in Semester A.

We will not be streaming these sessions online.

Why?

Is it worth it?



Justyna Gredecka

Full Stack Software Engineer at Cambridge Cancer Genomics



Modupeh Betts

Bioinformatics fellow at Medical Research Council(UK) The Gambia



Rachael Turner

Data Engineer at Sainsbury's



Numaan Iqbal

Data Analyst at Knight Frank



Nazrath Nawaz

Senior Bioinformatician at Kinomica



yasmine benbrahim

Inside Sales Account Manager (Nordics) at illumina
Connected 2 years ago



Karol Antoniuk

Bioinformatics Associate at Exact Sciences



Alejandra Carriero

PhD Student at University of Sussex



Josephine Mensah-Kane

Data Scientist at Mologic



Dionysios Grigoriadis

Bioinformatician at European Bioinformatics Institute



Hajar Saihi

PhD at the Blizzard Institute (Centre for Immunobiology)

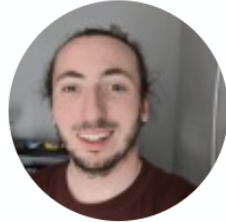


Suresh Hewapathirana

Bioinformatician/ Software Engineer
at European Bioinformatics Institute

Alumni Panel Q&A

2022



Liam Johnson

Bioinformatician - Clinical Genomics
The Royal Marsden NHS Foundation Trust

2022



Janeesh Kaur Bansal

PhD Student at Queen Mary University of London
London

2022



Shiloh Alleyne

Financial Model Developer at Gallagher Re
Ascot