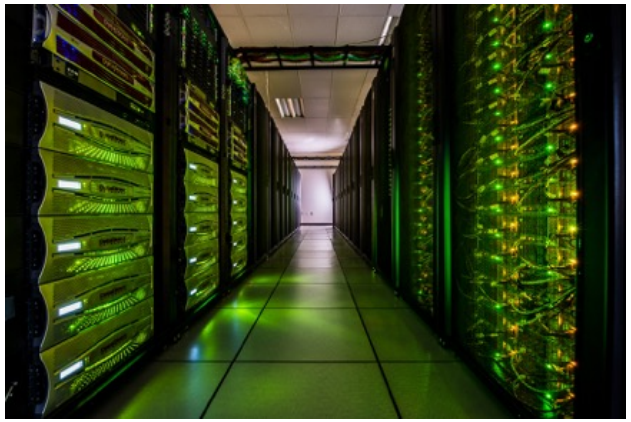
Life Sciences Data and Data-Centric Research

Sarah Butcher
s.butcher@imperial.ac.uk
www.imperial.ac.uk/bioinfsupport

Imperial College London

The World of Data

- Technology producing complex (interdisciplinary) data at exponential rates— data deluge
- Data are a resource BUT size and complexity are still overwhelming scientists' current practices to extract useful information
- Exploiting this resource requires better tools, practices and new solutions
- Need to combine scientific expertise, computational knowledge and statistical skills to solve critical problems and make new discoveries
- Requires new initiatives, institutional commitment, people-power and technology



Imperial College London

Data-Centric Science – It's All About the Data

"Hypotheses are not only tested through directed data collection and analysis but also generated by combining and mining the pool of data already available "

Goble and Roure (2009) from ***The Fourth Paradigm: Data-Intensive Scientific Discovery Edited by Hey, Tansley and Tolle).***

But In order to do this – data have to be discoverable and re-useable

Imperial College London

Summary - Questions

- Overview of work
- How did you start working with methodology side?
- Collaborative work with methodology side – shared benefits
- New research themes from your collaborative work and write technical papers?
- How do you educate/train pi-shaped scientists?

Imperial College London

Data as a Resource – The Rothamsted Park Grass Experiment

Picture from Rothamsted *e-RA*

- Oldest continuing experiment on permanent grassland in the world – started 1856
- Investigate ways of improving hay yield by using inorganic fertilisers and organic manure
- Measured species diversity and soil function also interactions with meteorological conditions
- Park Grass results are increasingly important to ecologists, environmentalists and soil scientists
- Being used in ways never imagined by the original scientists
- Possible as DATA and SAMPLES were kept, WE KNOW WHERE THEY ARE and samples can be re-analysed to provide missing data

Imperial College London

A Brief History Of Genome Sequencing

1977 first complete genome phage Φ-X174 (5,375bp)

1980 ~56 DNA gene sequences in public domain, ~180 by 1983

1995 first complete bacterial genome *Haemophilus influenzae*

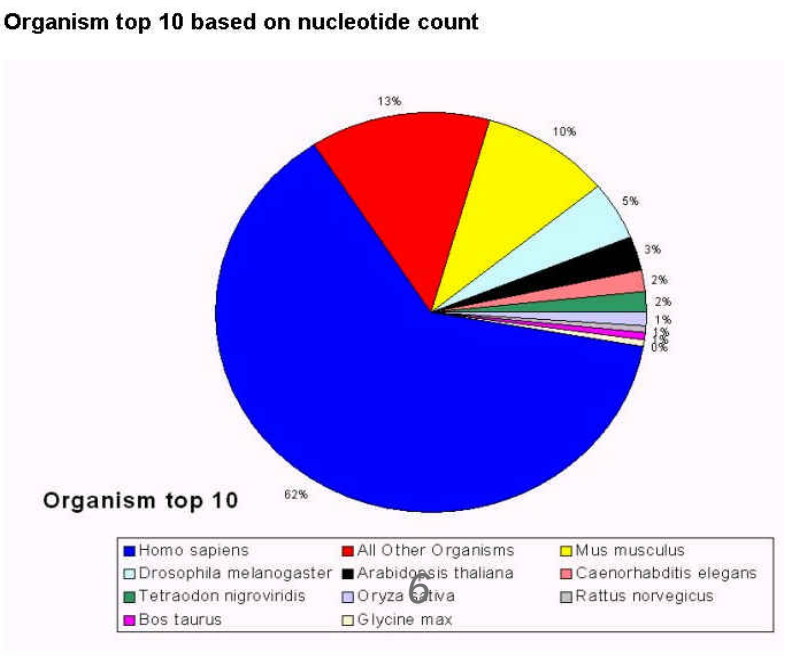
1996 first complete eukaryotic genome *Saccharomyces cerevisiae*

1998 first multicellular eukaryote genome *Caenorhabditis elegans* - (97Mb)

2001 Draft human genome published

over 11 million records in EMBL

2015 1939 completed eukaryotes, 31611 prokaryotes

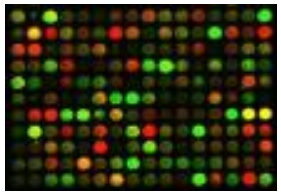


Bio-data Characteristics – The Basics

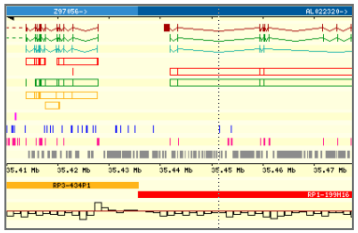
- Lack of structure, rapid growth but not (very) huge volume, **high heterogeneity**
- Multiple file formats, widely differing sizes, acquisition rates
- Considerable manual data collection
- Multiple format changes over data lifetime including production of (evolving) exchange formats
- Huge range of analysis methods, algorithms and software in use with wide ranging computational profiles
- Association with multiple metadata standards and ontologies, some of which are still evolving
- Increasing reference or link to patient data with associated security requirements

Imperial College London

Data Diversity And Volume



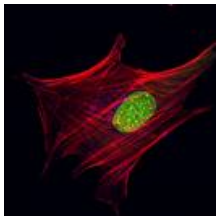
Transcriptome



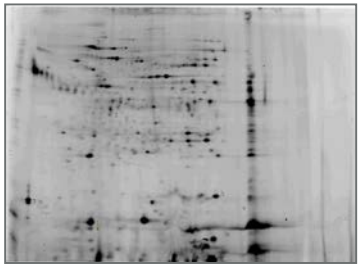
Genomes

?

Other -omes



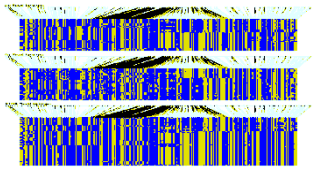
Bio-Imaging



Proteome

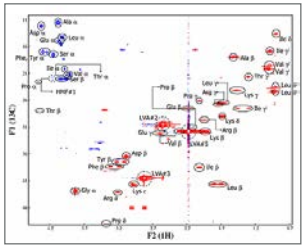
Improved understanding of complex biological system

Challenges in primary analyses (smaller) AND in meaningful integration (huge)

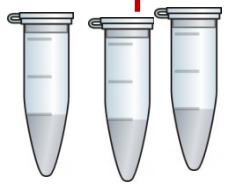


Variant analyses

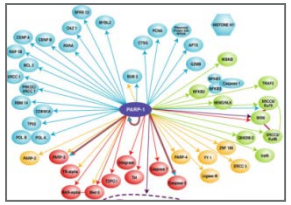
Large-scale field studies



Metabolomics



Clinical data, Sample-related data



Protein interactions

Imperial College London

Adding Complexity – Formats, Standards, Repositories

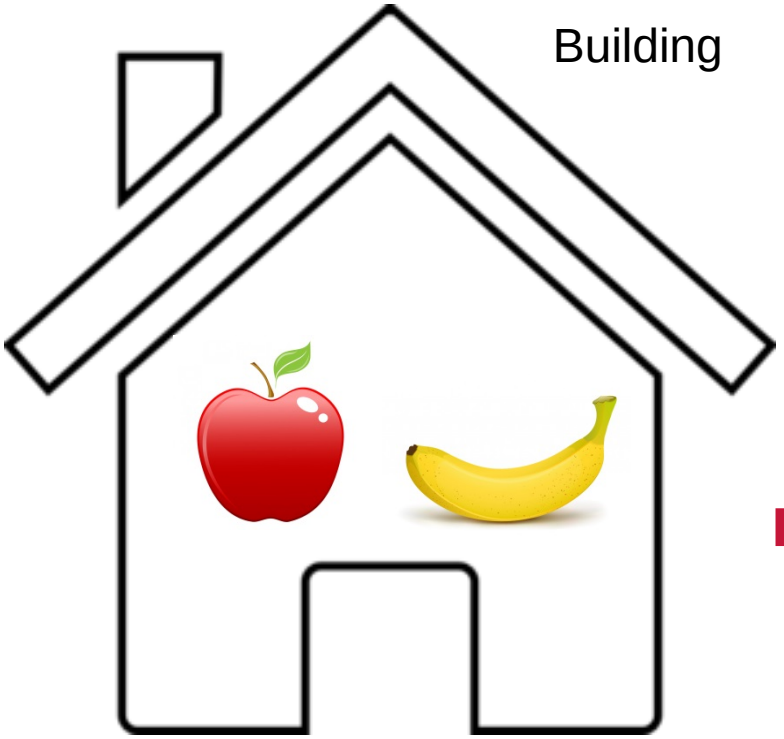
Next generation sequencing - including genome sequencing, re-sequencing and variant detection, RNA-Seq, ChIP-Seq	Binary alignment	BAM	Compressed (binary) version of SAM
	Sequence alignment/map	SAM	Created by alignment programs
	Defining annotation lines on a reference sequence	BED	For visualising annotations in genome browser
	'wiggle' format for continuous-valued data in a track format, also binary compressed version (BigWIG)	WIG BigWIG	e.g. visualisation of GC percent, probability scores, and transcriptome data on genome sequence
	Contains sequence and quality	FASTQ	Fasta format sequence and quality

- One raw data type BUT many file formats -may be human readable, require specific software, proprietary or open source
- Over 1552 different public databases, most limited by data domain, origin or both (NAR online Molecular Biology Database Collection)
- 30+ minimum reporting guidelines for bio/ biomedical data but few cross experimental types

= fragmentation, confusion for non-domain specialists

Imperial College London

The Systems Biology Lesson – Integration Takes Effort



Building

Bridging skills:
Understanding of
Programming
Data types
Some methodologies
Necessary software

?

interpreter

Data integration
Models
Full cycle systems biology

Biological data
Biologists

Models
Numerical scientists

Interdisciplinary training

Imperial College London

The Bioinformatics Support Service – What We Do

We support all stages in the **data lifecycle** - experimental design, data and metadata capture, primary and later stage analyses, data management, visualisation, sharing and publication

- Large-scale genomics & Next Generation Sequencing Analyses
- **Tools** for multiplatform data and metadata management
- Bespoke clinical and biological databases, tissue-banking
- **Software** and script development, data visualisation, mobile apps
- Full grant-based **collaboration** across disciplines
- **Brokering, skills sharing, advocacy**
- **New ways of high throughput working** – e.g. cloud, workflows
- Teaching, Workshops and One-to-One tutorials

Variety of skill-sets cover wet-lab bio, statistics, computer science

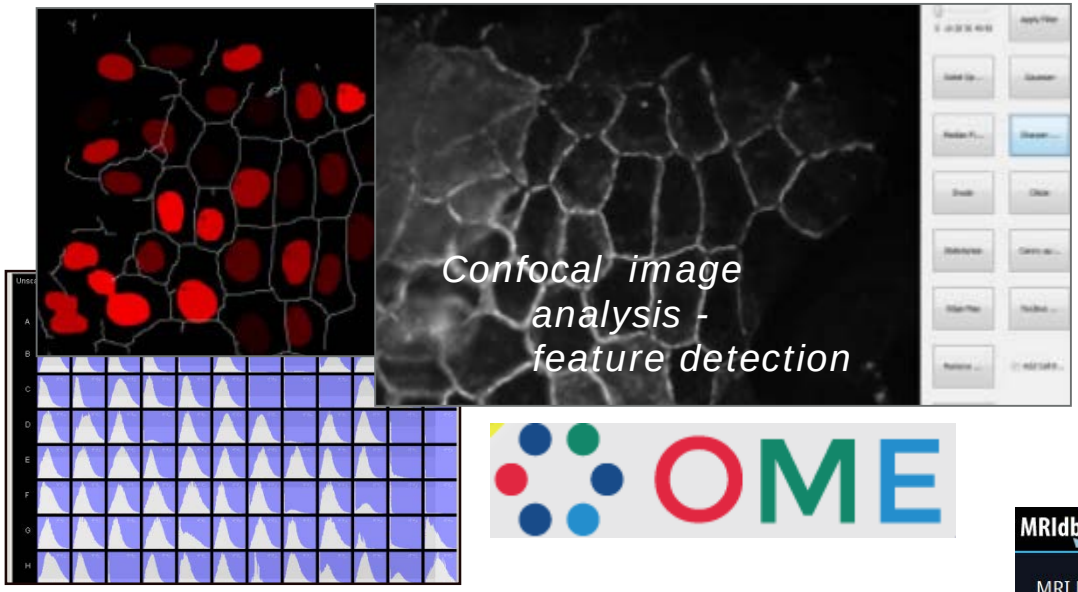
Imperial College London

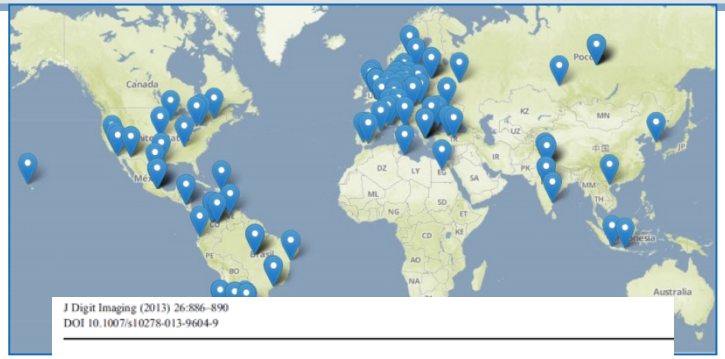
The Publication Complication

- Public bio-database formats lead to data fragmentation
- May cross-reference datasets across databases (good)
- Each has its own format and metadata requirements
- Quality assurance can be variable
- Data submission may be a requirement for journal publication (good)
- Large datasets can take weeks to prepare/validate and generate 100's of thousands of lines of XML, TB of data
- Automation complicated by regular changes to uploaders
- Where to put the other associated data – that may not be linked to a publication?

Imperial College London

Example - Bridging the Gaps In One Domain – Bio-imaging

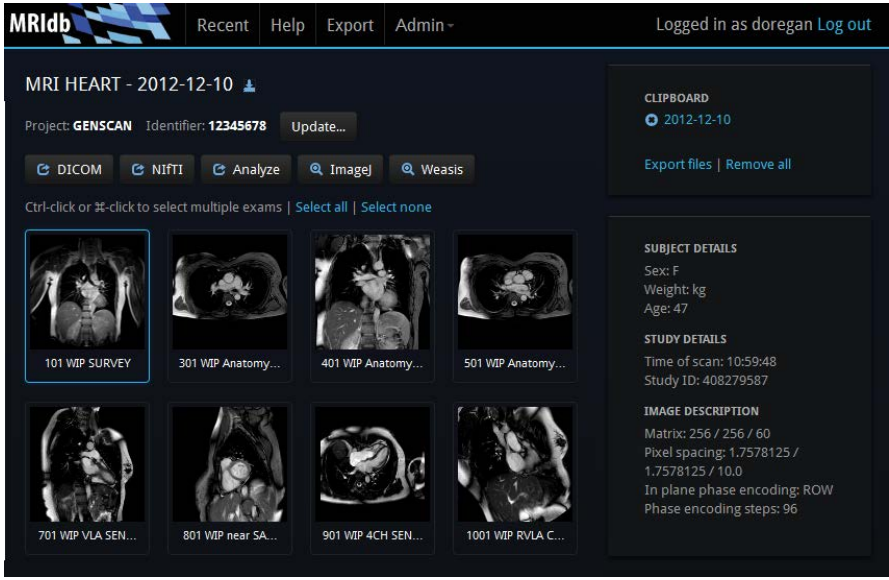




MRIdb: Medical Image Management for Biobank Research

Mark Woodbridge · Gianlorenzo Fagiolo · Declan P. O'Regan

OME



- Sample tracking for image analysis specialists
- Bespoke automated analysis systems for biologists
- Maintaining OMERO OME database for Photonics researchers
- MRI scan management solution for research groups

Imperial College London

Example - Encouraging Electronic Data Capture - Mobile applications For Data Input

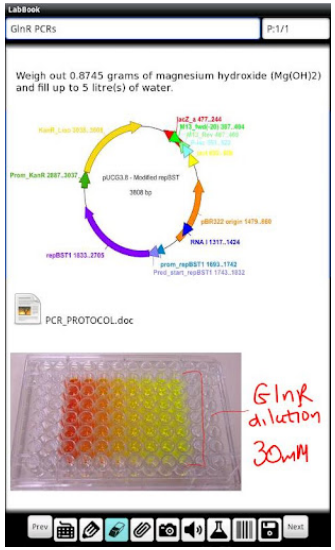
OPEN ACCESS Freely available online PLoS one

EpiCollect: Linking Smartphones to Web Applications for Epidemiology, Ecology and Community Data Collection

David M. Aanensen[1]*, Derek M. Huntley[2], Edward J. Feil[3], Fada'a al-Own[3], Brian G. Spratt[1]

1 Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom, 2 Centre for Bioinformatics, Imperial College London, London, United Kingdom, 3 Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

customisable geo-tagged data capture in the field automated remote database storage



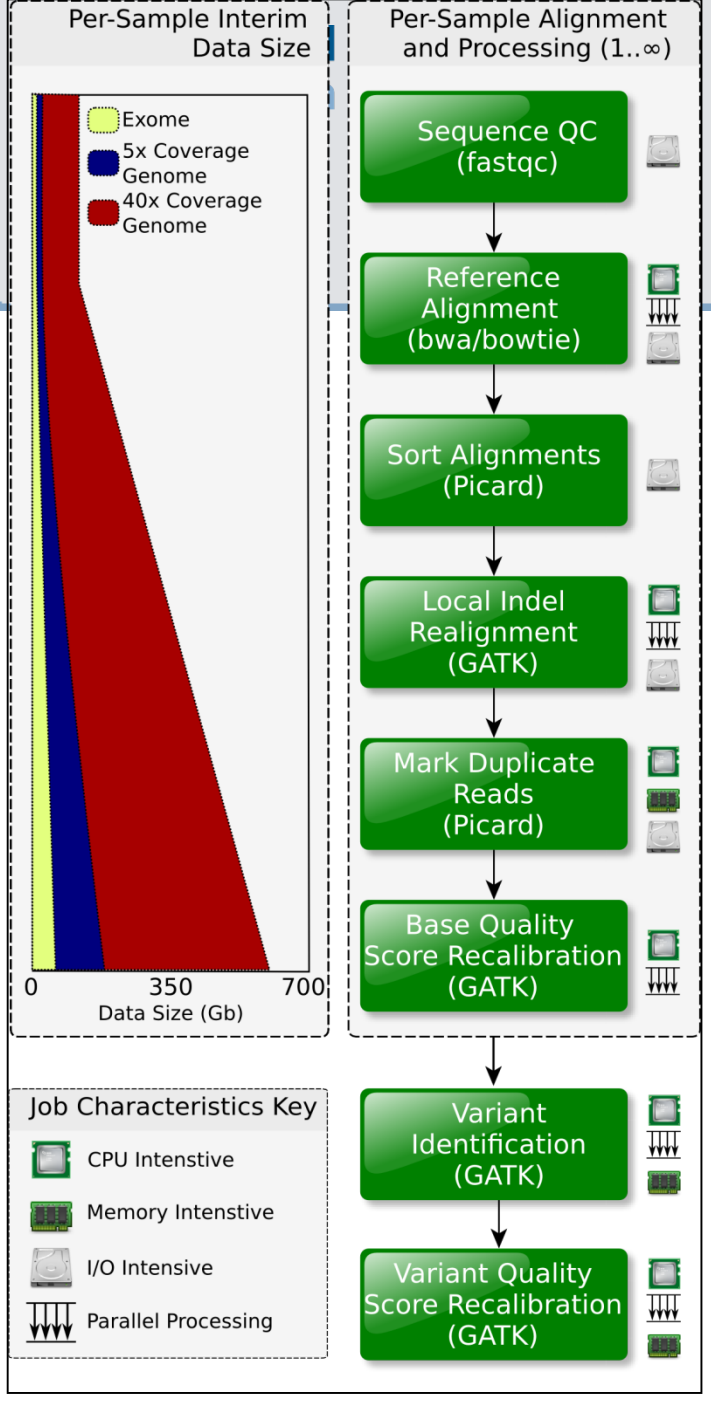
LabBook http://labbook.cc

Secure backup, sharing, search, version control via website

Handwritten notes, annotation

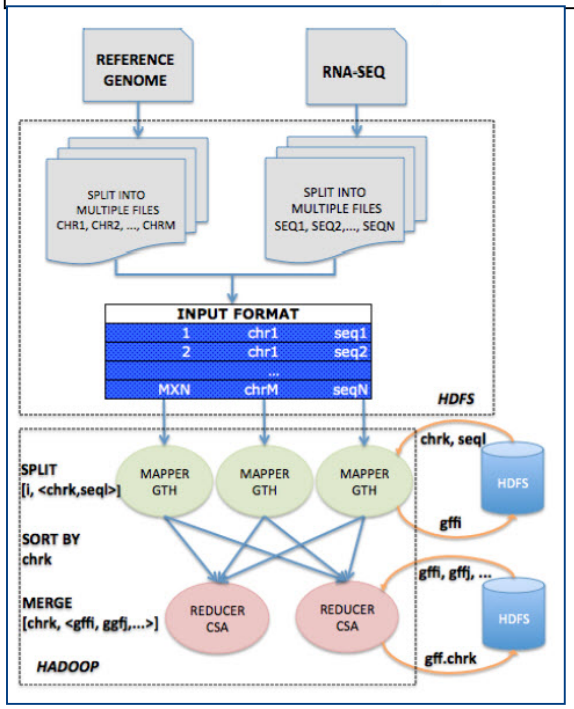
Supports photos, videos, file attachments, voice memos, barcode scanning

Practical Improvements For Increasingly Large Scale Data



RAPPORT: Running Scientific HPC Applications on the Cloud

Jeremy Cohen, Ioannis Filippis, Daniela Bauer, Brian Fuchs, Mike Jackson, Mark Woodbridge, Sarah Butcher, David Colling, John Darlington, Matt Harvey and Neil Chue Hong



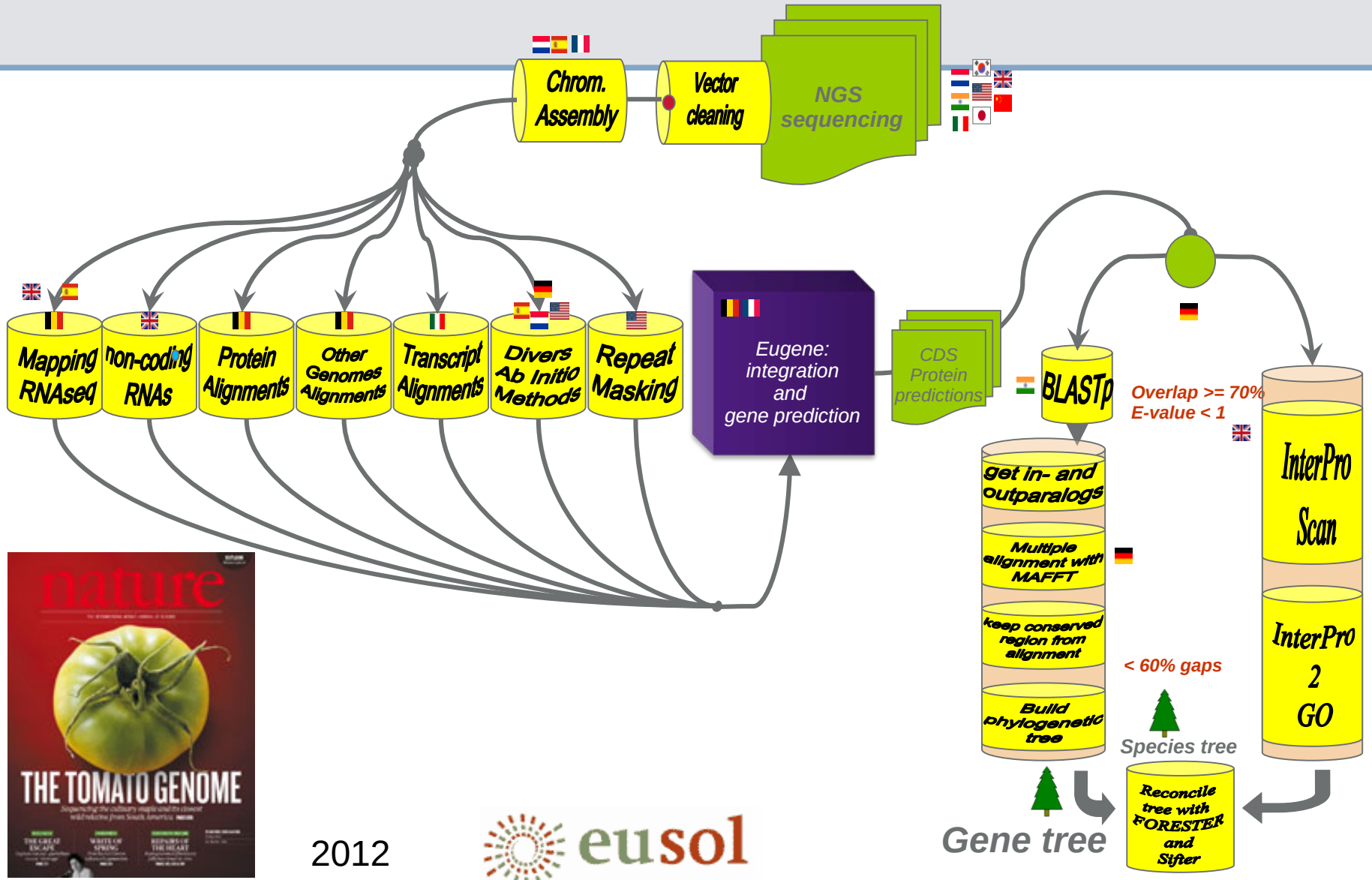
GenomeThreader in the MapReduce framework

What can we learn from Collaborators:

High Energy Physics
Astronomy
Photonics
Chemistry
Mathematics
Computer Science

Imperial College London

The iTAG Annotation Pipeline



2012

eusol

S. Rombauts, iTAG

Imperial College London

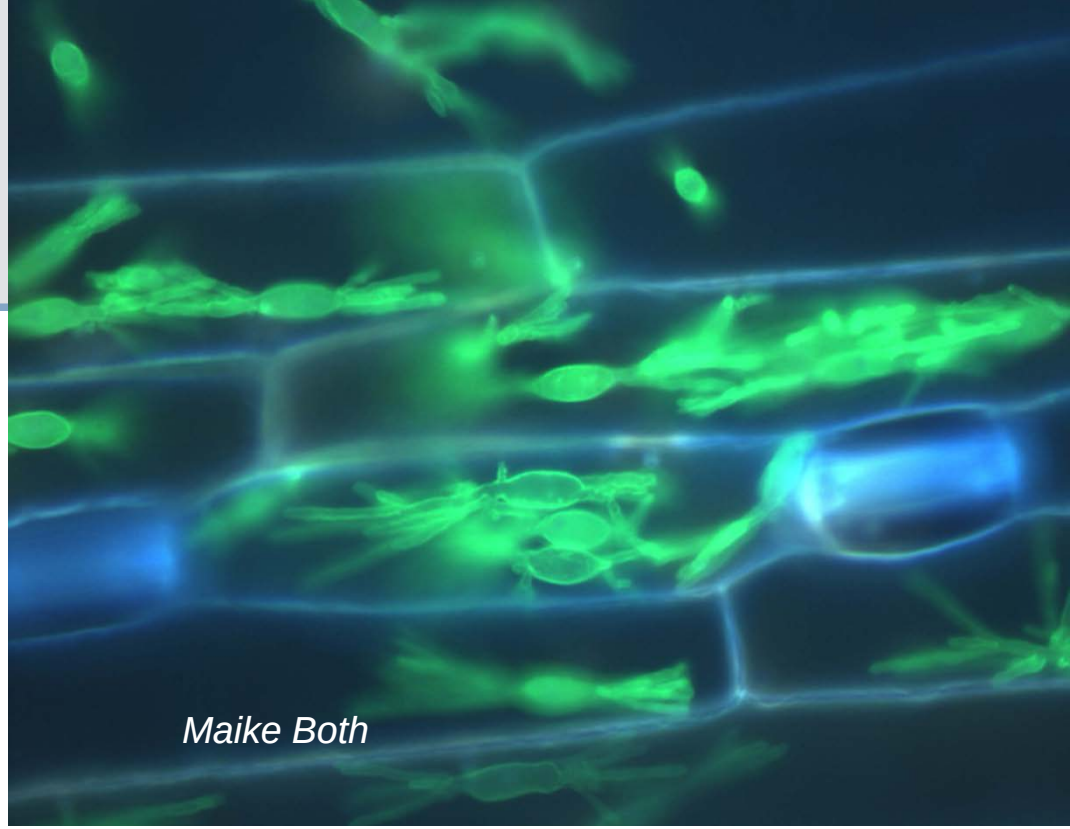
Grass Roots Challenges

- Integrative approaches repeatedly show that complete metadata are vital for optimal data reuse BUT
- Metadata capture still a complex time-consuming task
- Data fragmentation across multiple sites still a major barrier to uptake (*can't find it… can't use it…)*
- Practical aspects – cost of storage & curation, sheer volume of datasets
- Difficulty of obtaining consistent funding for fundamentals

  - maintaining core infrastructure, software, databases
- Staff – shortage of truly inter-disciplinary infrastructure & knowledge providers, career progression

Imperial College London

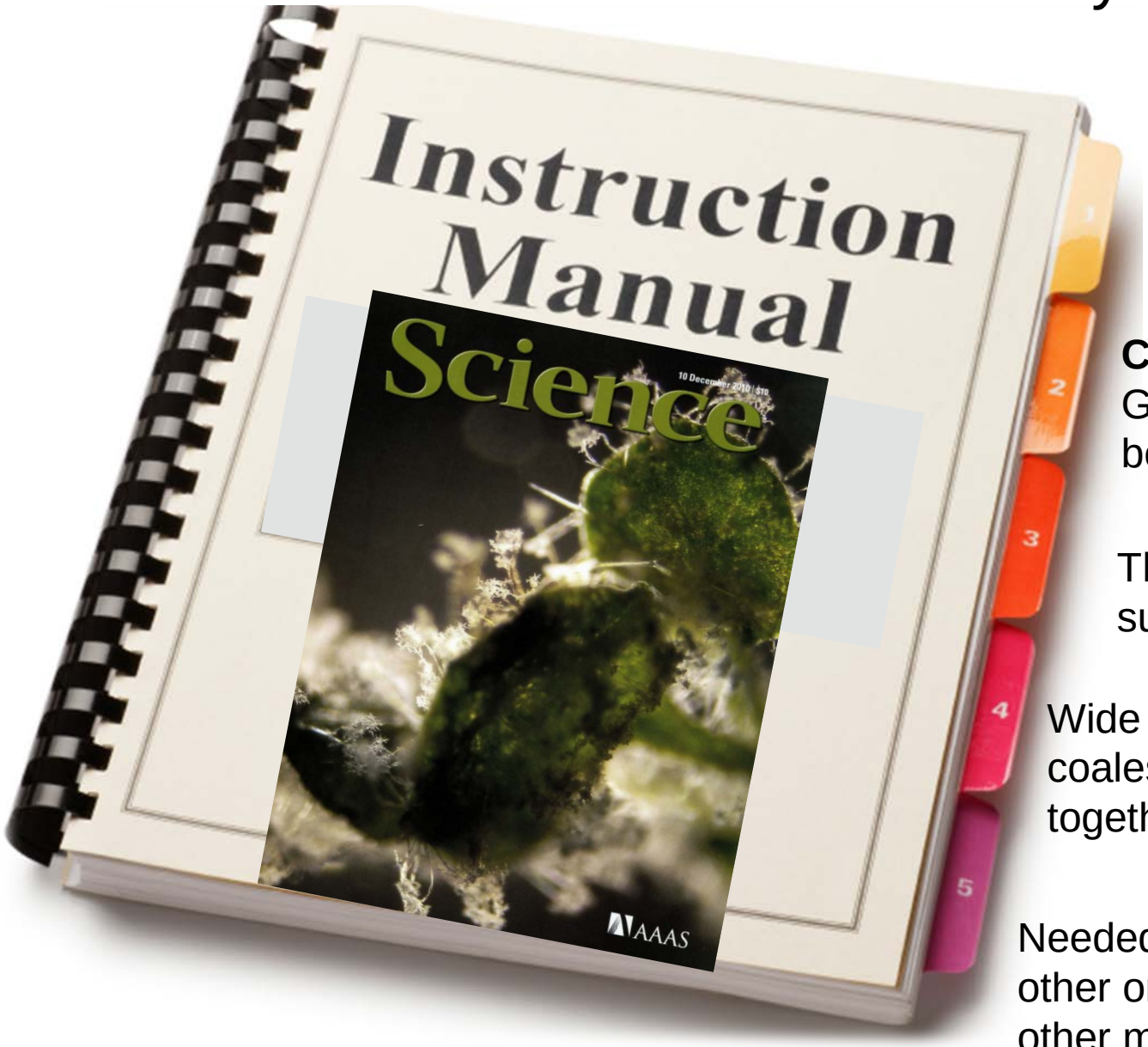
The Blumeria Story
Spanu *et al*

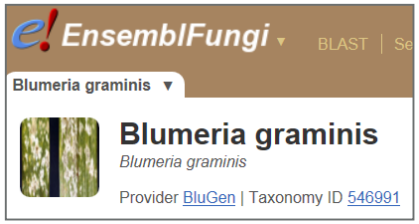


A Collaboration Story

- Cereal powdery mildews
- Obligate biotrophs of Wheat, Barley
- Fungal Haustoria fill the living plant cells and siphon off food
- Also may deliver the Effectors that turn off the Plant ‘immune’ response

How a wet-lab went multi –omic by collaboration





Changes in technology: Genome sequencing became cost-effective

The genome produced surprises

Wide team of Collaborators coalesced - still working together

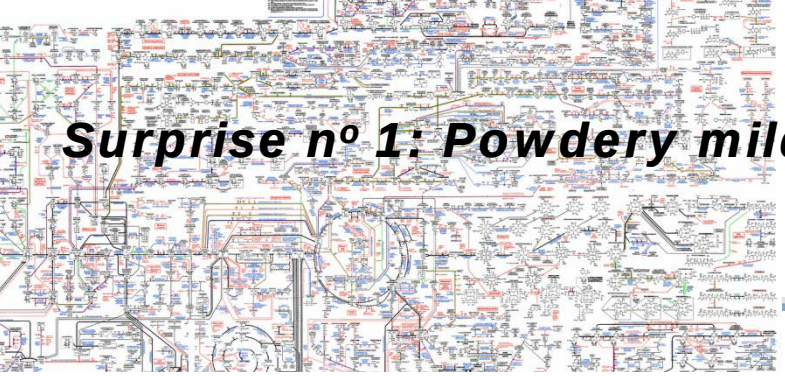
Needed input from many other organisms, other datasets, other methodologies to get the bigger picture

Spanu *et al* *DOI:10.1126/science.11 94573*

Imperial College London

Complex Heterogeneous Data

- Blumeria Genome - 5 different sequencing technologies required complex hybrid assemblies
- Annotation - automated pipeline AND extensive collaborative manual annotation across multiple countries
- Comparative analyses using data from 3 other species' genomes
- Integration across multiple data types:
  - RNA-seq data
  - Mass spec proteomics data
  - NMR data
  - Protein structural prediction AND AND AND……
- AND - originating lab had no informatics expertise



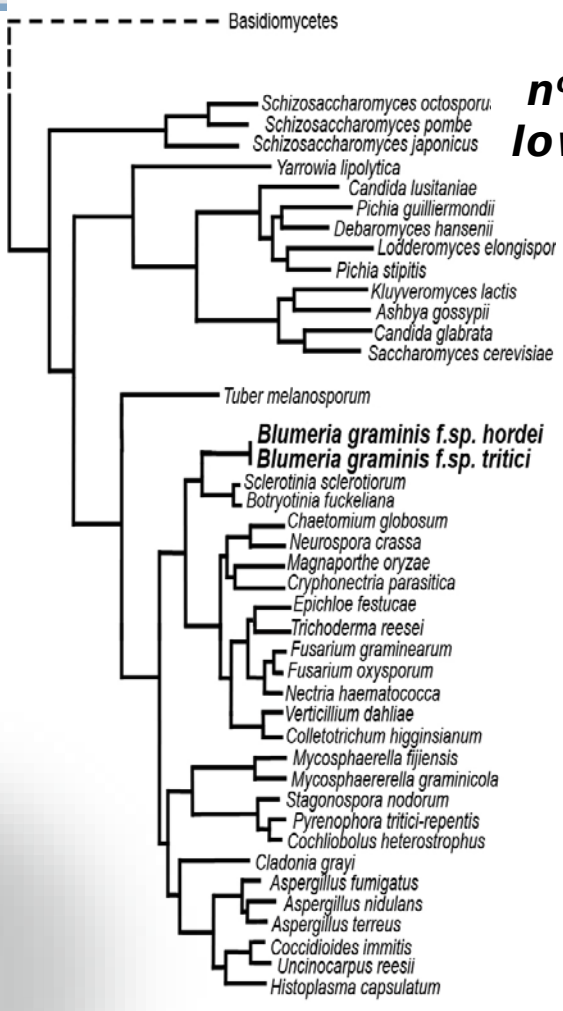
Surprise nº 1: Powdery mildew genome ~4 x larger than expected

Surprise nº 2: practically all primary metabolic pathways are conserved

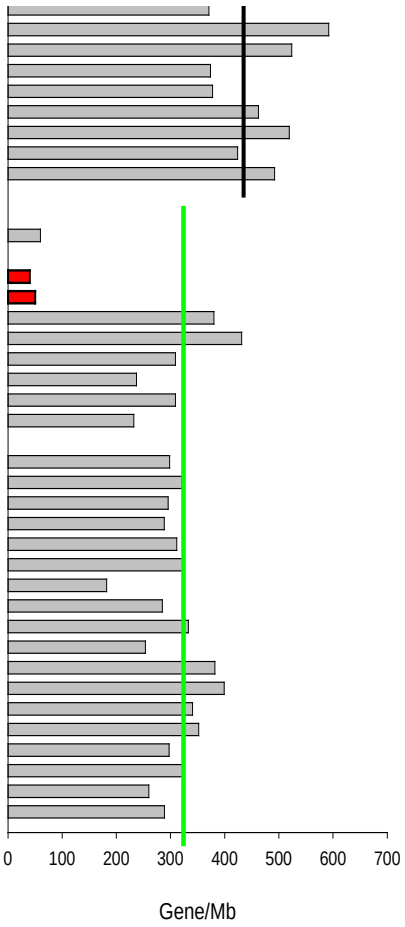
Surprise nº 3: Powdery mildews have big genomes with few genes

Number of genes in average fungal genome (~12,000)

Number of genes Cereal Powdery Mildew Genomes (~6500)



nº 4: surprising low gene density



Spanu et al, 2010

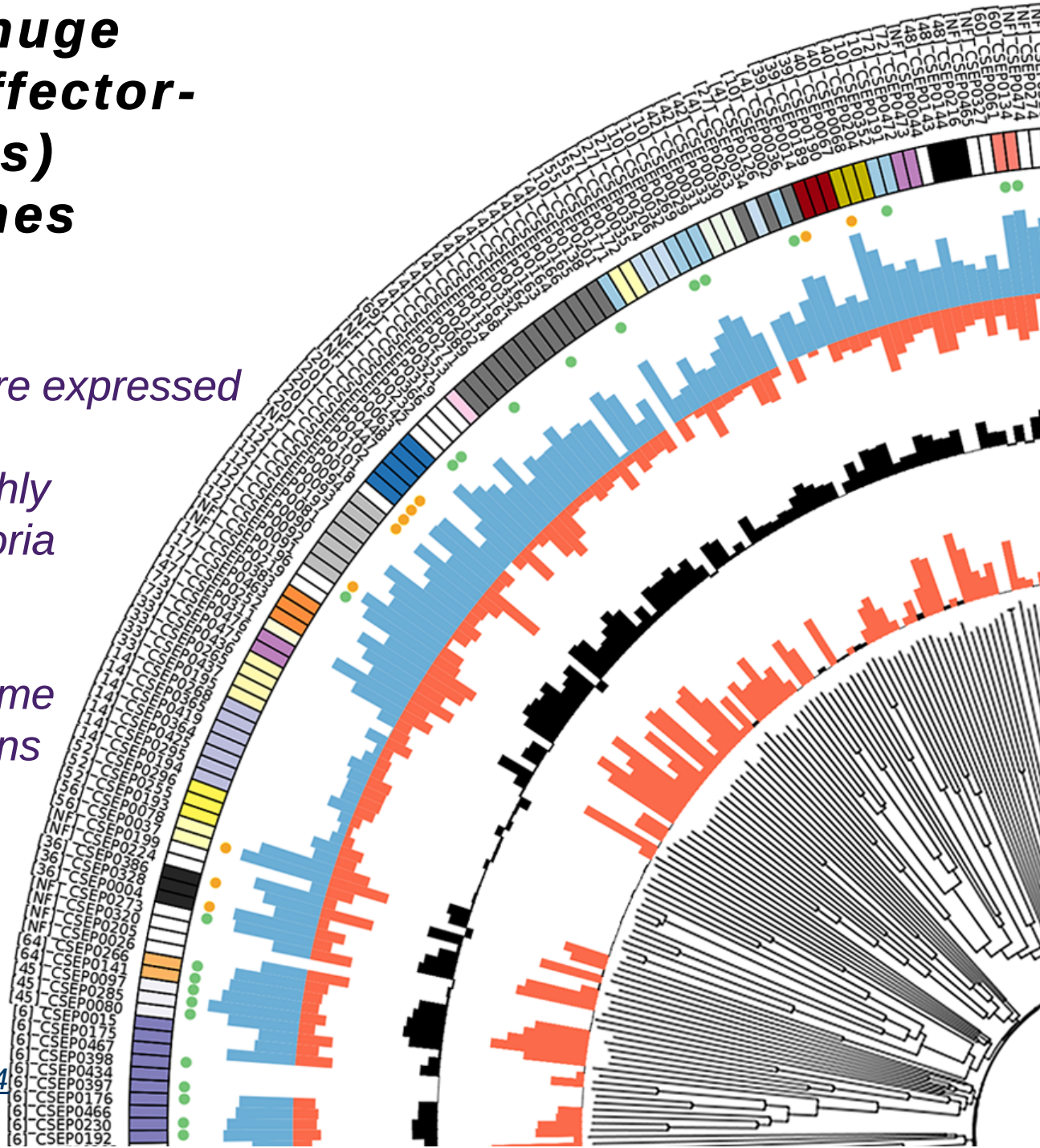
Surprise nº 5: a huge superfamily of effector-like genes (CSEPs) >7% of total genes

RNA-Seq shows:

- *vast majority of these are expressed at high levels*
- *the majority is more highly expressed in the haustoria*

Proteomics shows:

- *These proteins are some of the dominant proteins in haustoria*



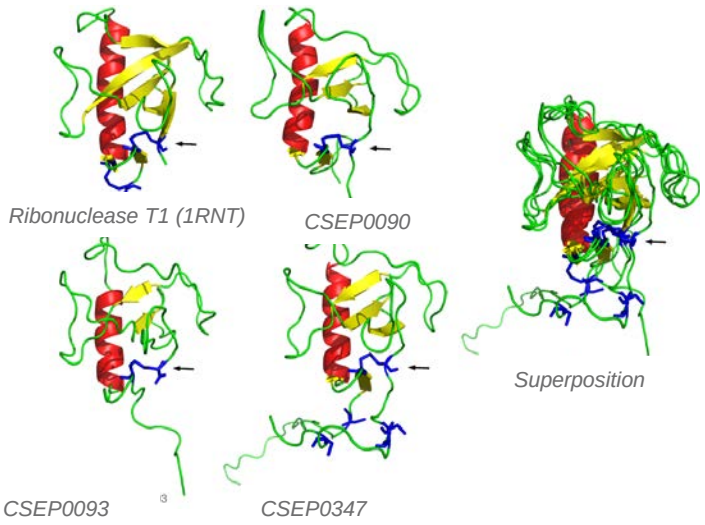
Pedersen *et al*. (2012)
doi:10.1186/1471-2164-13-694

Imperial College London

The End of the Beginning – Enabling New Investigations

A whole new theme of investigation - effectors:

- Host-Induced Gene Silencing to look at effects on pathogenicity
- Expression profiling during infection
- Transient expression in plants to study effect on susceptibility to some pathogens
- Structure prediction for RNAse-like ("RALPH") candidate effectors (PHYRE and INFOLD)
- Solved structure for some candidates
- RNA binding demonstration - Nucleic Acids induce NMR shift
- Ongoing studies on binding function



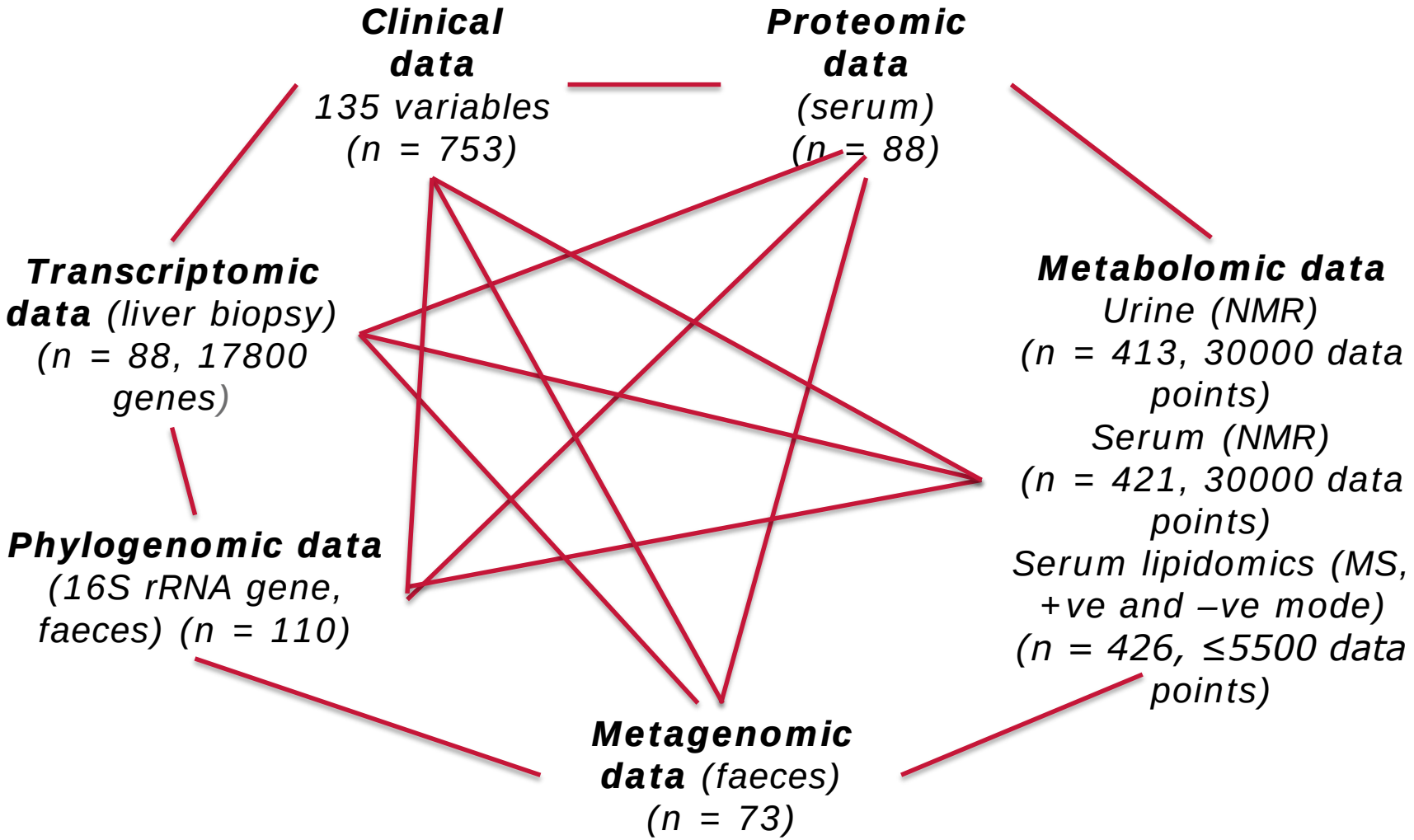
Pedersen et al (2012)

Imperial College London

Florinash

The role of intestinal microbiota in non-alcoholic fatty liver disease (NAFLD)

SEVENTH FRAMEWORK PROGRAMME

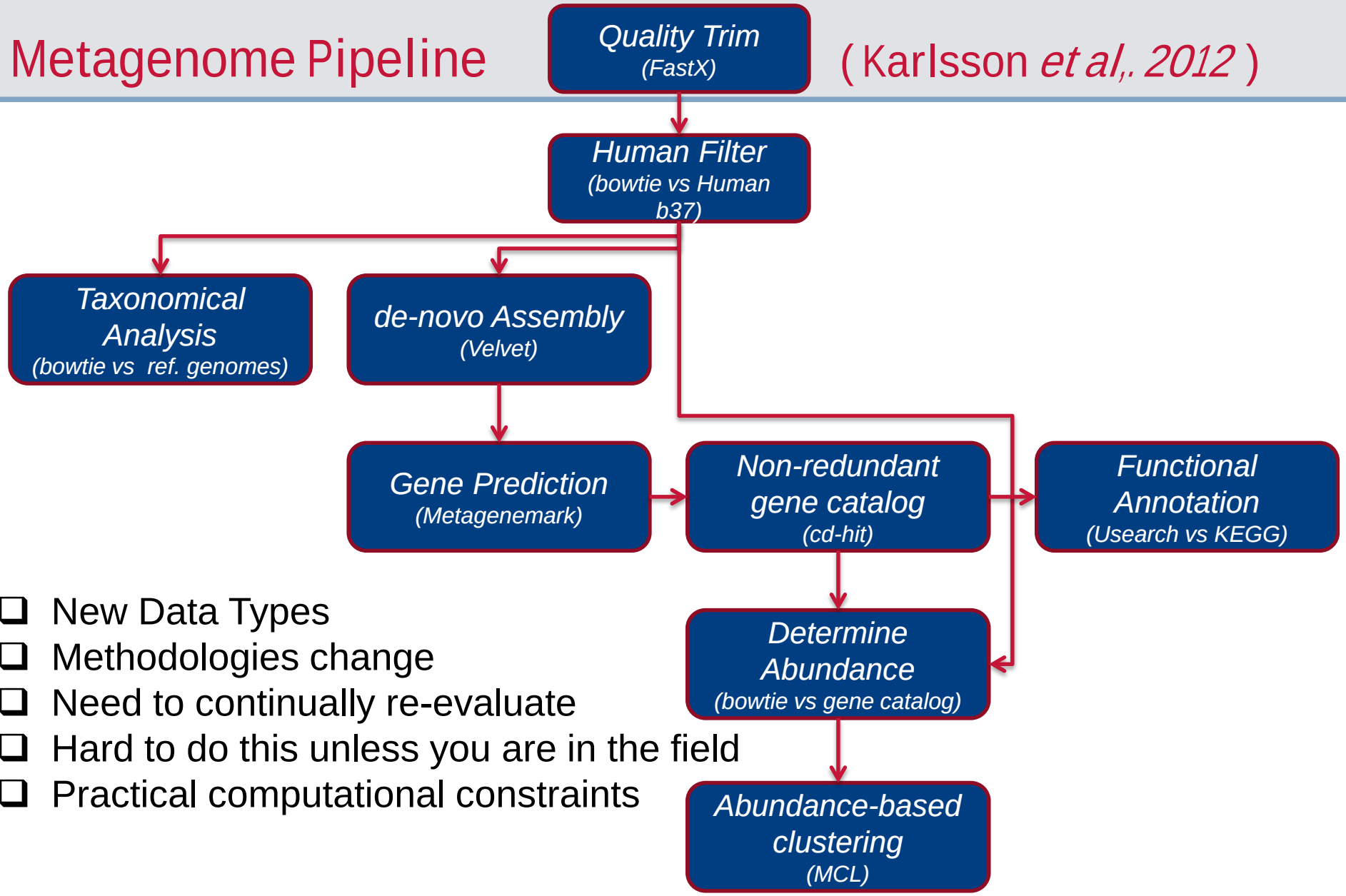


Not originally planned

L. Hoyles

Imperial College London

Metagenome Pipeline



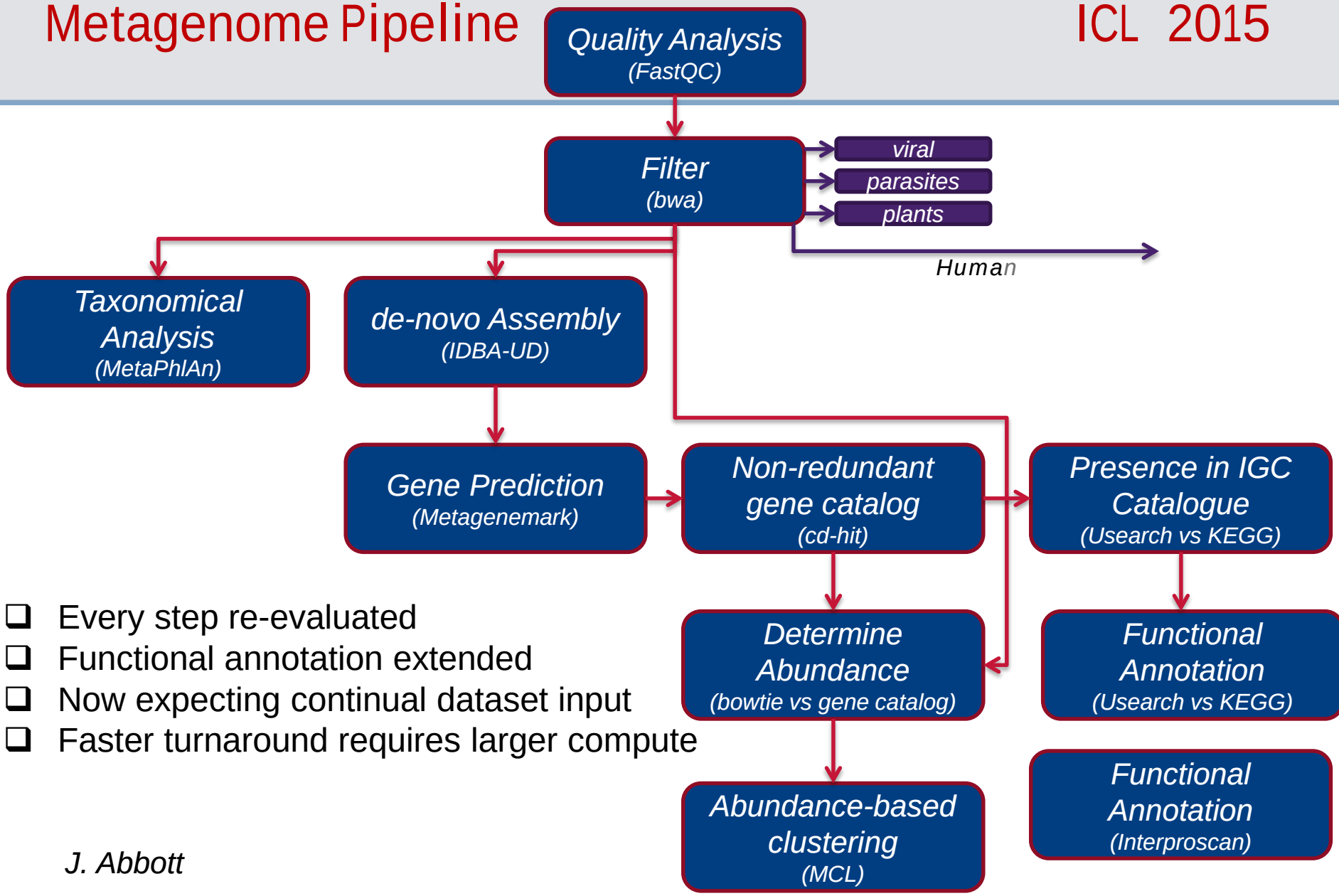
(Karlsson *et al,. 2012*)

- ❑ New Data Types
- ❑ Methodologies change
- ❑ Need to continually re-evaluate
- ❑ Hard to do this unless you are in the field
- ❑ Practical computational constraints

Imperial College London

Metagenome Pipeline

ICL 2015



- Every step re-evaluated
- Functional annotation extended
- Now expecting continual dataset input
- Faster turnaround requires larger compute

J. Abbott

Imperial College London

Better Instrumentation, Higher Throughput, More Integration

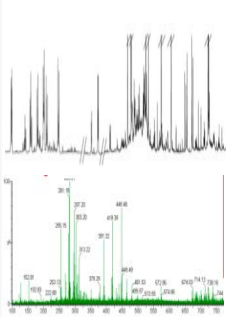
SPECTROSCOPIC PLATFORM

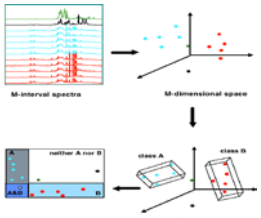
NMR

UPLC-MS

ANALYTICAL DATA

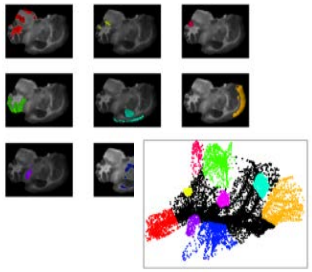
SPECTRA



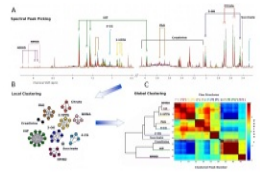


CLASSIFICATION & PREDICTION

IMAGE RECONSTRUCTION

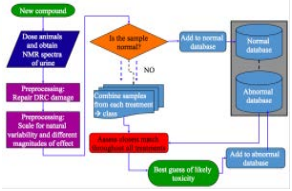


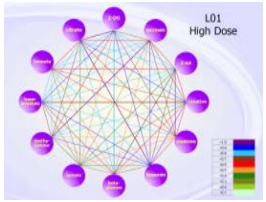
BIOINFORMATICS AND MODELLING



STATISTICAL SPECTROSCOPY

EXPERT SYSTEMS





OMICS INTEGRATION - THE INTERACTOME

Advancement & application of metabolic profiling methods & technologies

- *Undertake and develop state-of-the-art (mass spectrometric and NMR spectroscopic) analyses for metabolic finger-printing of biofluids*
- *Combine metabolic analyses with other clinical, lifestyle and –omics datasets*
- *A national resource and research capacity, enabling researchers to derive clinically-relevant insights to identify bio-markers or profiles*
- *Develop new methods and technologies*

MRC-NIHR National Phenome Centre

MRC | National Institute for Health Research | Imperial College London | KING'S College London | Waters THE SCIENCE OF WHAT'S POSSIBLE. | BRUKER

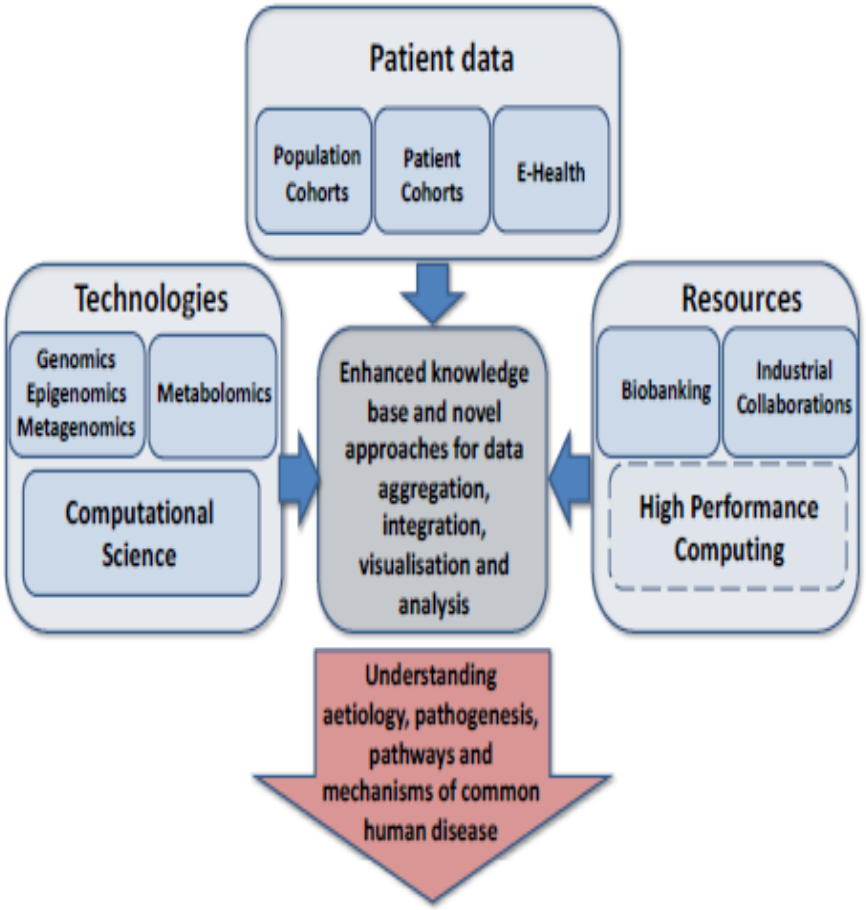
Imperial College London

UK MED-BIO: Aggregation, Integration, Visualisation and Analysis of Large, Complex Data

- **Example of newly funded multi-disciplinary initiatives**
- **1 of 6 national projects to improve infrastructure for medical informatics**
- **Multiple partner Institutions, multiple areas:**
  - *Imperial (population studies, GWAS, Metabolomics, data integration)*
  - Institute of Cancer Research (cancer informatics)
  - European Bioinformatics Institute (Metabolights database)
  - Centre for the Improvement of Population Health through E-health Research (e-health records)
  - MRC Clinical Sciences Centre (data integration, statistics)
  - MRC Human Nutrition Research (phospho-proteomics)
- **Multiple Industrial partners**

Imperial College London

MED-BIO – Complex Large Data



Largest primary data volume producer is metabolomics

Also:

- NGS (exomes, genomes, targetted)
- Proteomics (mass spec)
- Transcriptomics and methylation-based
- Gut metagenomics and meta-transcriptomics
- Genome wide association studies

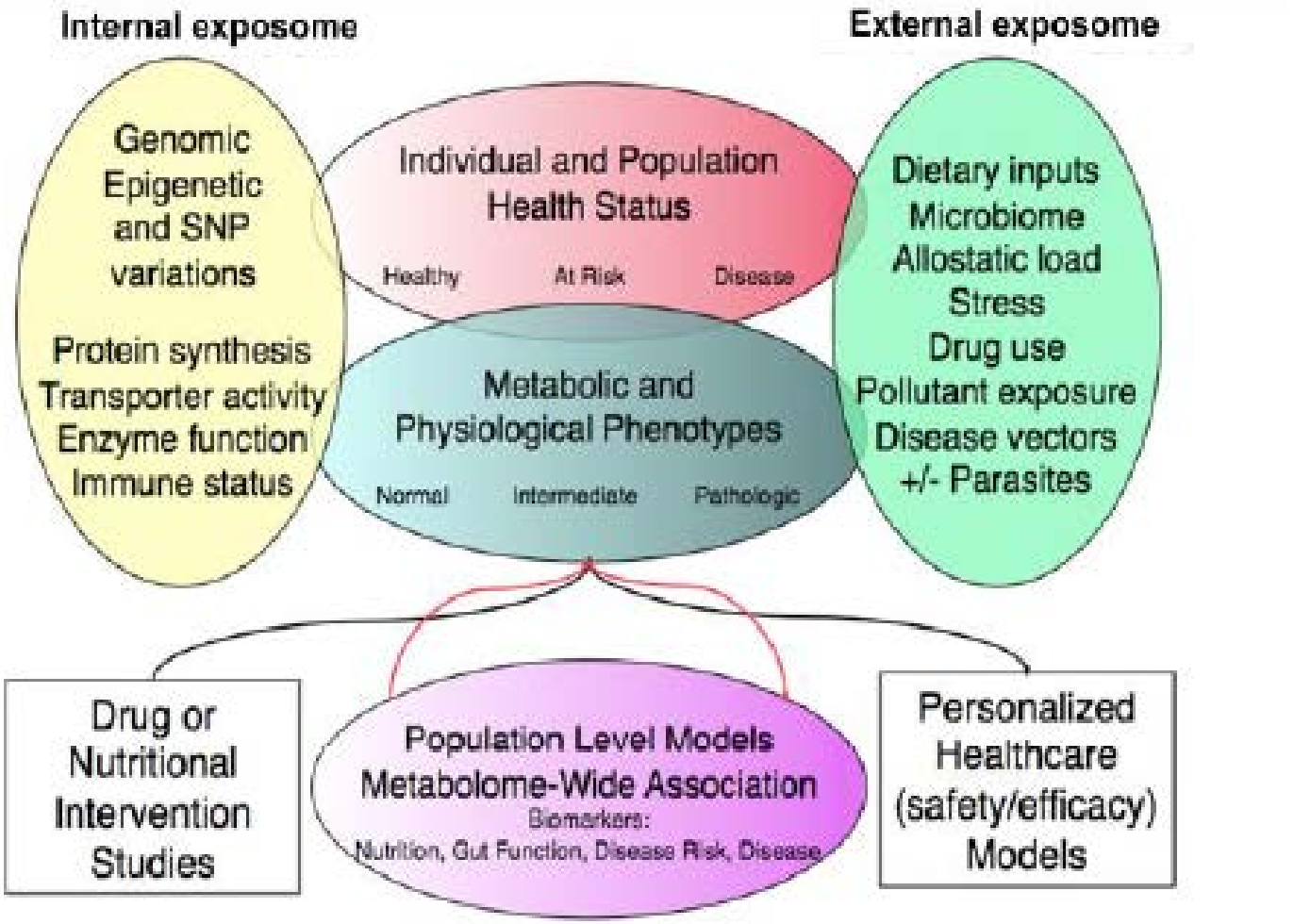
Need to support primary data analyses

AND Integration and intelligent data-mining of large, heterogeneous, high dimensional datasets (from all of above)

Also secure integration with patient data

Imperial College London

The Expososome



Holmes et al. 2008 *Cell* **134**:714-717.

Imperial College London

More Practical Challenges

- 1-off Capital funding to buy the big compute, big storage needed
  - BUT future needs are emergent – need flexibility and scaleability
- Little funding provision for staff to build and maintain (and help/support) the complex software/data infrastructures
  - Requires additional resources - or a bottleneck develops
- Funded mid-career Fellowships encourage innovation BUT
  - They also need integrative support
- Data and metadata management will be vital
  - BUT not 'trendy' or easily fundable and require domain-specific knowledge – automate as much as possible

Imperial College London

Scaling

- Support primary data analyses as well as later integration and mining
- Heterogeneous job profiles: standard cluster compute (3280 additional cores), cache-coherent memory (640 cores, 8 TB RAM), large memory nodes (40 cores, 1-2TB RAM each)
- Centralised active tiered storage – 800TB GPFS, 2 PB object store, 2 PB tape – duplicated across 2 sites
- Video wall, touch overlay, 3D projection capability for visualisation
- Centrally-managed software, scheduling, metadata capture
- BUSINESS MODEL for growth, sustainability

Imperial College London

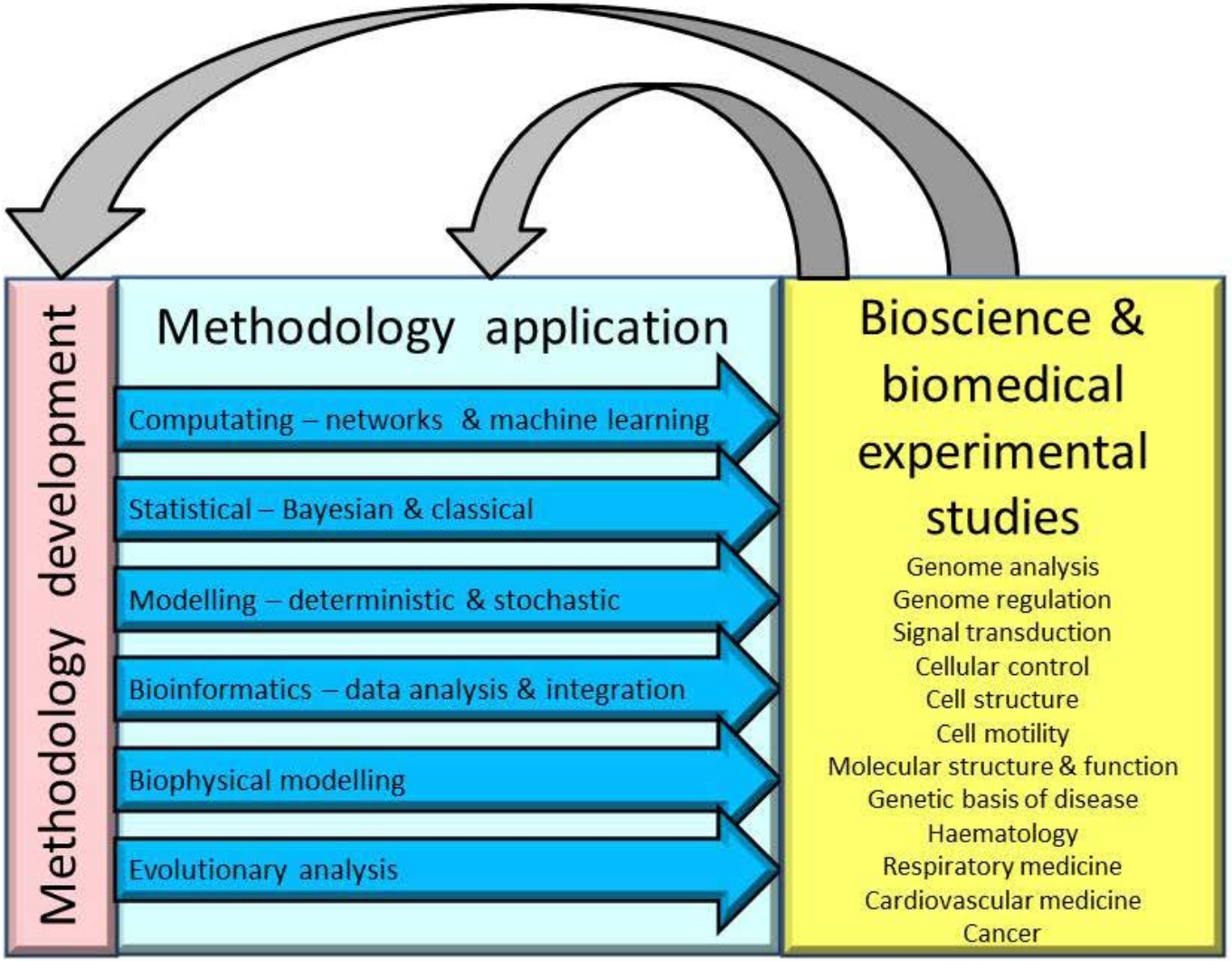
Training and Skills – The IT Scientist

A Recent survey of vulnerable skills and capabilities for UK Research Councils (BBSRC, MRC) identified:

- Lack of inter-disciplinary skills at postgraduate and postdoc. level, and need for depth as well as breadth of knowledge
- Data analytics especially bioinformatics vulnerable – but also general large scale data analysis skills – interpretation, storage, programming
- Maths, statistics and computational biology lacking at the postgrad and postdoc level – so recruiting difficult, not just in UK
- Quality and provision of operational and support roles an issue
- Bioinformatics now on Home Office's Shortage Occupation list

Imperial College
London

Over 30 Bioinformatics and Systems Biology Modelling Groups Across The College



Imperial College London

Formal Training - MSc Bioinformatics and Theoretical Systems Biology

- Aim - Train both numerical and biological undergraduates in bioinformatics and theoretical systems biology so they can progress to research posts in world leading academic, governmental and commercial centres
- Annual intake c. 15 students- always both numerical and biological
- Over 75% progress to PhDs in best institutions (Imperial, UCL, Cambridge, Oxford, ETH, EMBL)
- In last BBSRC funding round, this MSc was ranked top from all biological science proposals

http://www.imperial.ac.uk/study/pg/courses/life-sciences/bioinformatics/

Imperial College London

MSc in Bioinformatics and Theoretical Systems Biology - a 12 month course

- 1st three months formal training
  - Fundamentals of biology
  - Statistics and mathematical modelling
  - Bioinformatics and theoretical systems biology
  - Computer programming (Python, Java, MySQL)
- Project 1 – group database
- Project 2 – data analysis and web design
- Project 3 – research topic (sometimes published)
  - Over 30 groups provide research topics from many Imperial departments including clinical groups

Imperial College London

PhD Training Next Generation Computational Biologists

- Across departments, faculties and campuses
- With about 30 theoretical groups over 100 PhD students currently being trained
- Research supported by £25M grants
- Some purely theoretical, others mixed wet / dry
- Industrial partnership studentships – e.g. CASE
- BUT training, mentoring required for all stages – and not so easy to support or fund

Imperial College
London

EURATRANS