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
“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
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?
The Rothamsted Park Grass Experiment

- 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
A Brief History Of Genome Sequencing

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1977</td>
<td>First complete genome phage Ф-Х174 (5,375bp)</td>
</tr>
<tr>
<td>1980</td>
<td>~56 DNA gene sequences in public domain, ~180 by 1983</td>
</tr>
<tr>
<td>1995</td>
<td>First complete bacterial genome <em>Haemophilus influenzae</em></td>
</tr>
<tr>
<td>1996</td>
<td>First complete eukaryotic genome <em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td>1998</td>
<td>First multicellular eukaryote genome <em>Caenorhabditis elegans</em> - (97Mb)</td>
</tr>
<tr>
<td>2001</td>
<td>Draft human genome published over 11 million records in EMBL</td>
</tr>
<tr>
<td>2015</td>
<td>1939 completed eukaryotes, 31611 prokaryotes</td>
</tr>
</tbody>
</table>
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
Data Diversity And Volume

Transcriptome

Genomes

Other -omes

Bio-Imaging

Improved understanding of complex biological system

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

Proteome

Large-scale field studies

Metabolomics

Clinical data, Sample-related data

Variant analyses

Protein interactions
Adding Complexity – Formats, Standards, Repositories

- 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
The Systems Biology Lesson – Integration Takes Effort

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

Data integration
- Models
- Full cycle systems biology

Interdisciplinary training

Biological data
- Biologists

Models
- Numerical scientists

Building

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

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

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

What can we learn from Collaborators:

- High Energy Physics
- Astronomy
- Photonics
- Chemistry
- Mathematics
- Computer Science

GenomeThreader in the MapReduce framework
The iTAG Annotation Pipeline

- Chrom. Assembly
- Vector cleaning
- NGS sequencing
- Eugene: integration and gene prediction
- CDS
- Protein predictions
- BLASTp
- get in- and outparalogs
- Multiple alignment with MAFFT
- Keep conserved region from alignment
- Build phylogenetic tree
- InterPro Scan
- Overlap >= 70%
- E-value < 1
- < 60% gaps
- Species tree
- Reconcile tree with FORESTER and Sifter
- Gene tree
- InterPro 2 GO

Mapping RNaseq
non-coding RNAs
Protein Alignments
Other Genomes Alignments
Transcript Alignments
Divers Ab Initio Methods
Repeat Masking

S. Rombauts, iTAG
2012
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
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
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
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......
- 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

**Surprise no 4:** surprising low gene density

---

Spanu et al, 2010
**Surprise n° 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)
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)
The role of intestinal microbiota in non-alcoholic fatty liver disease (NAFLD)

Clinical data
135 variables
(n = 753)

Transcriptomic data (liver biopsy)
(n = 88, 17800 genes)

Phylogenomic data (16S rRNA gene, faeces) (n = 110)

Proteomic data (serum)
(n = 88)

Metabolomic data
Urine (NMR)
(n = 413, 30000 data points)
Serum (NMR)
(n = 421, 30000 data points)
Serum lipidomics (MS, +ve and –ve mode)
(n = 426, ≤5500 data points)

Metagenomic data (faeces)
(n = 73)

Not originally planned

L. Hoyles
Metagenome Pipeline

1. **Quality Trim** (FastX)
2. **Human Filter** (bowtie vs Human b37)
   - **Taxonomical Analysis** (bowtie vs ref. genomes)
   - **de-novo Assembly** (Velvet)
     - **Gene Prediction** (Metagenemark)
       - **Non-redundant gene catalog** (cd-hit)
         - **Determine Abundance** (bowtie vs gene catalog)
           - **Abundance-based clustering** (MCL)
         - **Functional Annotation** (Usearch vs KEGG)

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

- Quality Analysis (FastQC)
- Filter (bwa)
- Taxonomical Analysis (MetaPhlAn)
- de-novo Assembly (IDBA-UD)
- Gene Prediction (Metagenemark)
- Non-redundant gene catalog (cd-hit)
- Presence in IGC Catalogue (Usearch vs KEGG)
- Determine Abundance (bowtie vs gene catalog)
- Abundance-based clustering (MCL)
- Functional Annotation (Usearch vs KEGG)
- Functional Annotation (Interproscan)

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

J. Abbott
Better Instrumentation, Higher Throughput, More Integration

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
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
Largest primary data volume producer is metabolomics

Also:
- NGS (exomes, genomes, targeted)
- 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
The Expososome

Internal exposome
- Genomic
- Epigenetic and SNP variations
- Protein synthesis
- Transporter activity
- Enzyme function
- Immune status

External exposome
- Dietary inputs
- Microbiome
- Allostatic load
- Stress
- Drug use
- Pollutant exposure
- Disease vectors
- +/- Parasites

Individual and Population Health Status
- Healthy
- At Risk
- Disease

Metabolic and Physiological Phenotypes
- Normal
- Intermediate
- Pathologic

Drug or Nutritional Intervention Studies

Population Level Models
- Metabolome-Wide Association
- Biomarkers: Nutrition, Gut Function, Disease Risk, Disease

Personalized Healthcare (safety/efficacy) Models

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
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
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
Over 30 Bioinformatics and Systems Biology Modelling Groups Across The College
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/
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
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