Translational Science:
Current information challenges and solutions.

Phil Scordis, Director Translational Bioinformatics, UCB

Matt Page, James Snowden, Patrice Godard, Jonathan Van Eyll, Martin Armstrong, Fred Vanclef, Matthias Hoss, Alison Maloney, Ian White

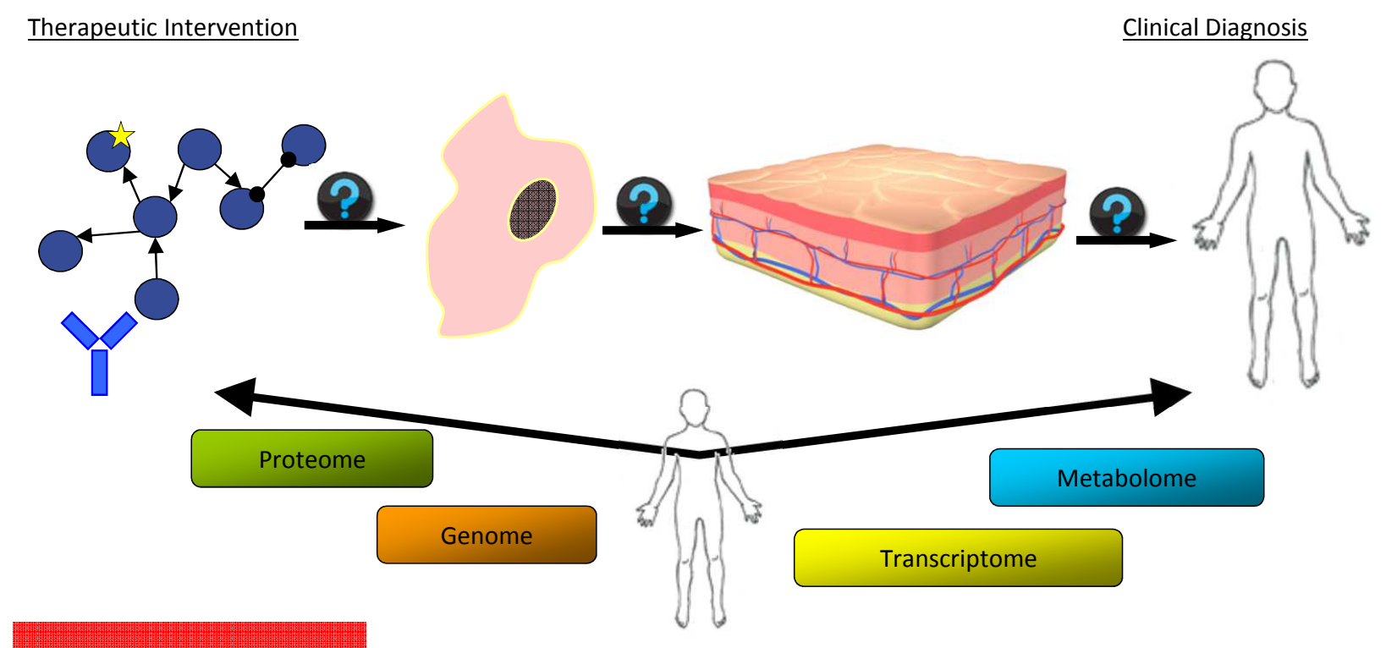
Abstract

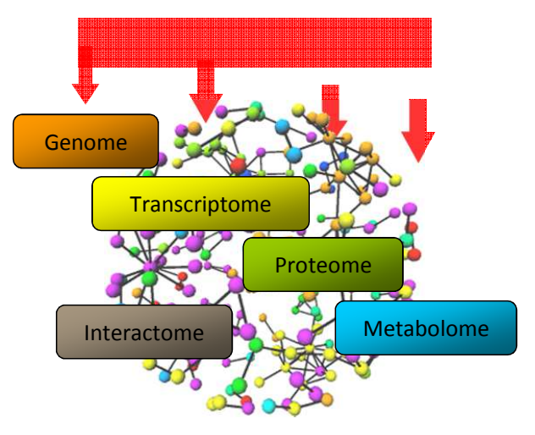
Translational Research relies on building bridges between clinical paradigms and fundamental biological models in order to bring insights into the complexity of human disease / treatment opportunities. As this is inherently reliant on bringing data from multiple domains together to address these challenges data integration continues to be a topic of great interest.

Exemplified through a set of internal use cases we highlight some of the challenges of enabling translational research activities, and the translational informatician at the heart of it. The emphasis is on exposing some of the internal approaches but we bring into this the context of a workshop recently hosted by the EBI in which some of the observations were magnified.

The poster aims to highlight some of the opportunities of Translational Research and provoke discussion around the challenges of bringing technologies to play in the context of a rapidly evolving ecosystem of data, analytical tools and infrastructure.

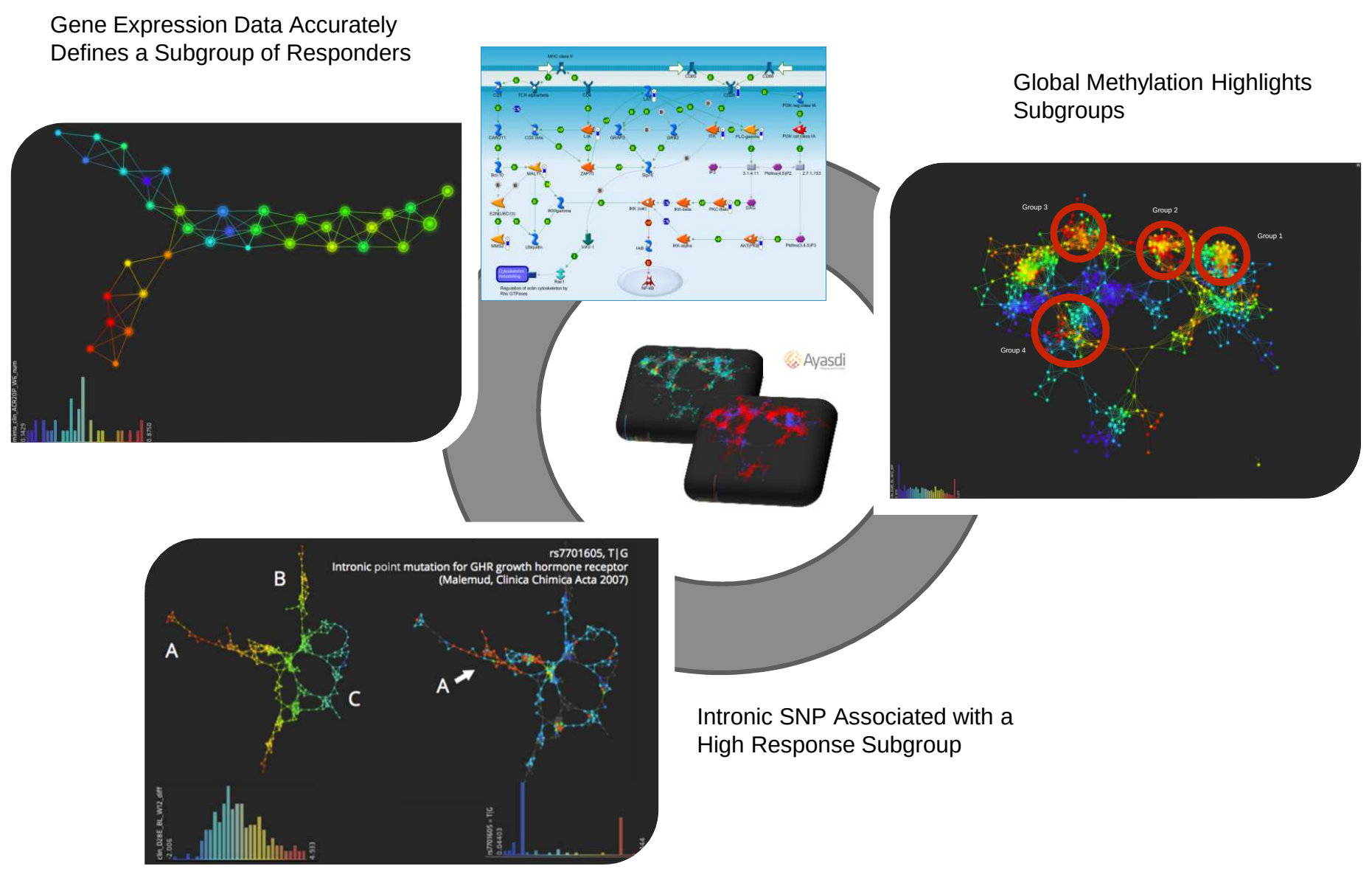
Understanding the Molecular Etiology of Disease





Multi-Omic Data Integration

Exploiting the interrelationships between Omic data sets, enable insight into underlying biology by combining complementary complex, large-scale omic data sets and knowledge frameworks as a basis for integration.

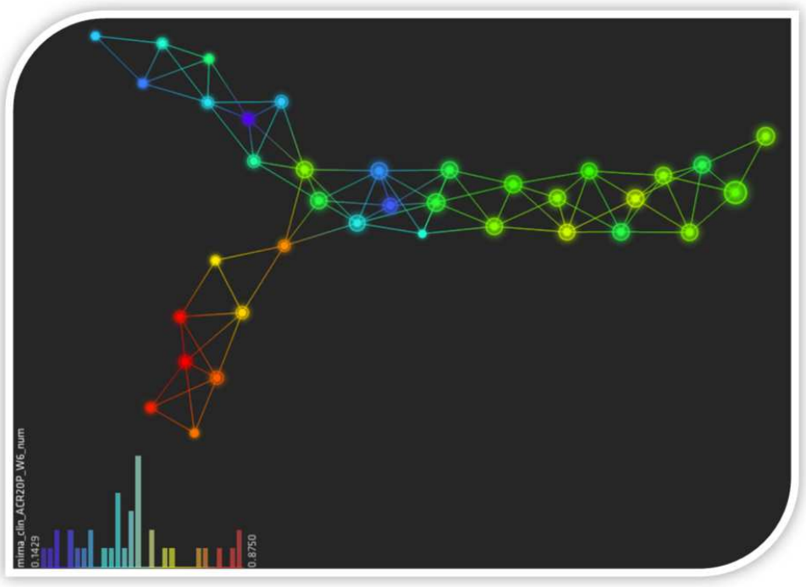


Gene Expression Data Accurately Defines a Subgroup of Responders

Global Methylation Highlights Subgroups

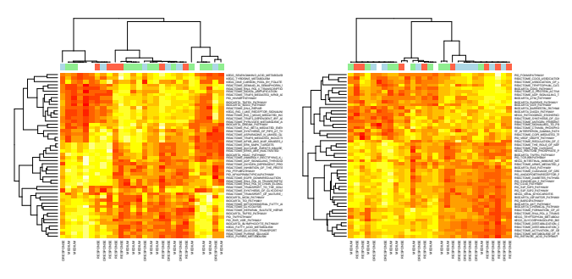
Intronic SNP Associated with a High Response Subgroup

Interrogating clinical data to understand patient heterogeneity



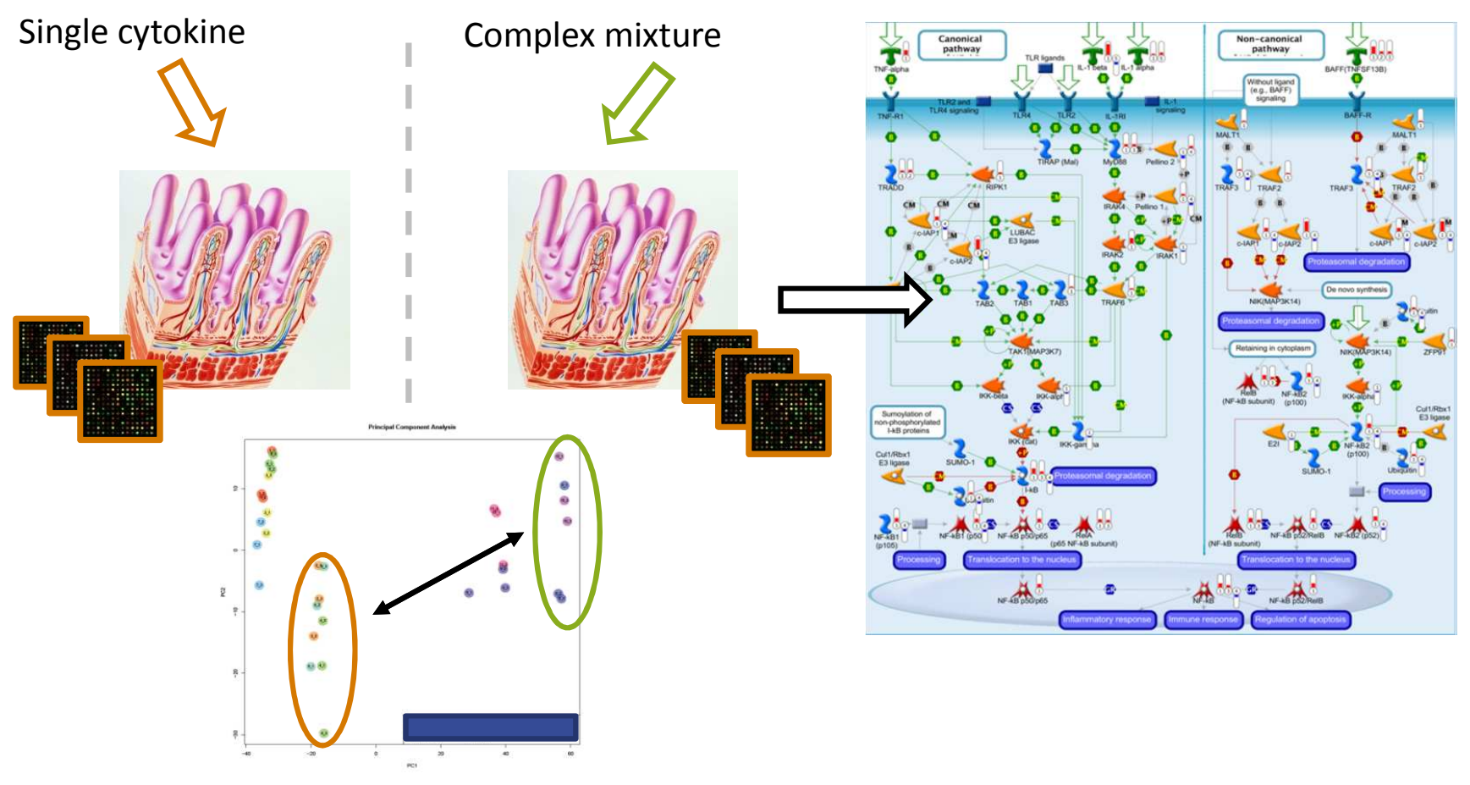
Identifying predictive gene signatures of patient response using existing clinical data in one of UCB's Phase 3 assets

Could lead to biomarker tools to guide patient stratification and treatment decisions



Creating translatable insight to select right indications

Replicating the complex human physiology of the gut in vitro for Phase 1 asset in order to construct a translatable mechanistic hypothesis.



EBI – Enabling the Translational Bioinformatician Workshop

Established with the following aims, to gather a greater understanding of the status of translational informatics and the expectations set for data generation/management in the public domain, an understanding of the academic/public funded initiatives in this domain, to probe opportunities for cross-industry pre-competitive collaborations, to convert the broad scope into future workshop opportunities able to drill down in greater detail. In particular the workshop focused on highlighting the opportunities of translational informatics, the challenges in achieving these from both a technical as well as broader perspective and a touch point on the emerging translational science that growth in data and computational capabilities is bringing to translational research.

Summary

The meeting had a diverse group of speakers from industry and academia who engage in translational informatics and covering a broad remit ranging from introductory discussions on the nature of translational research to its application in clinical and research settings in academia and industry, and the challenges of managing, standardizing, sharing and computing over data that supports these aims. Some of the key topics that became focal points in the discussions were the following:

- The reiteration of the simplicity of the central dogma that underlies much of the knowledge representation of the data environment in which translational informatics operates – both sparse data and simplistic models challenge the gene to protein or protein to function tenets.
- The growth of the human as the experimental organism brings great opportunities – the growth of Patient driven initiatives, Patient advocacy groups, patient reported outcomes, the quantified self and great challenges; consent, security, privacy and the complexity of regulatory diversity.
- Big Data – the ever expanding availability of data resources in the global environment – brings great hope of course as data driven knowledge models have the opportunity to grow iteratively. But the growth of data resources in the global community are inherently disparate, lack standards and have uncertain life-cycles. Genomics data are growing particularly rapidly, but few healthcare initiatives are scalable.
- A diverse skillset is required to operate in this domain, moving the eponymous Bioinformatician from a jack of all trades to a coordinator of a network of skilled practitioners and the rise of the Physician-Scientist a practitioner of translational science asking experimental questions in a clinical setting and appreciating the value of structured information.

Outcomes

Future workshop proposals:

- Technologies to support translational informatics: with a focus on leading examples of infrastructure investment in this area to explore how the community might address these challenges
- Consent and governance: bringing together the worlds of legal, clinical, and research
- Filling the Phenotype – Mechanism Gap: Building a Common Language, addressing the gap left by the lack of deep / endo phenotype descriptors and their connection to biological mechanisms
- Data standards
- Emerging Science in the Translational Domain.

The group saw opportunity in alignment across industry to address some of these challenges through pre-competitive collaboration/partnership identification, in particular the need to ensure there is some consistency between various working groups who have interest here – inc. Pistoia, PRISME, Cross Pharma Translational R&D IT/Informatics Network, etc.

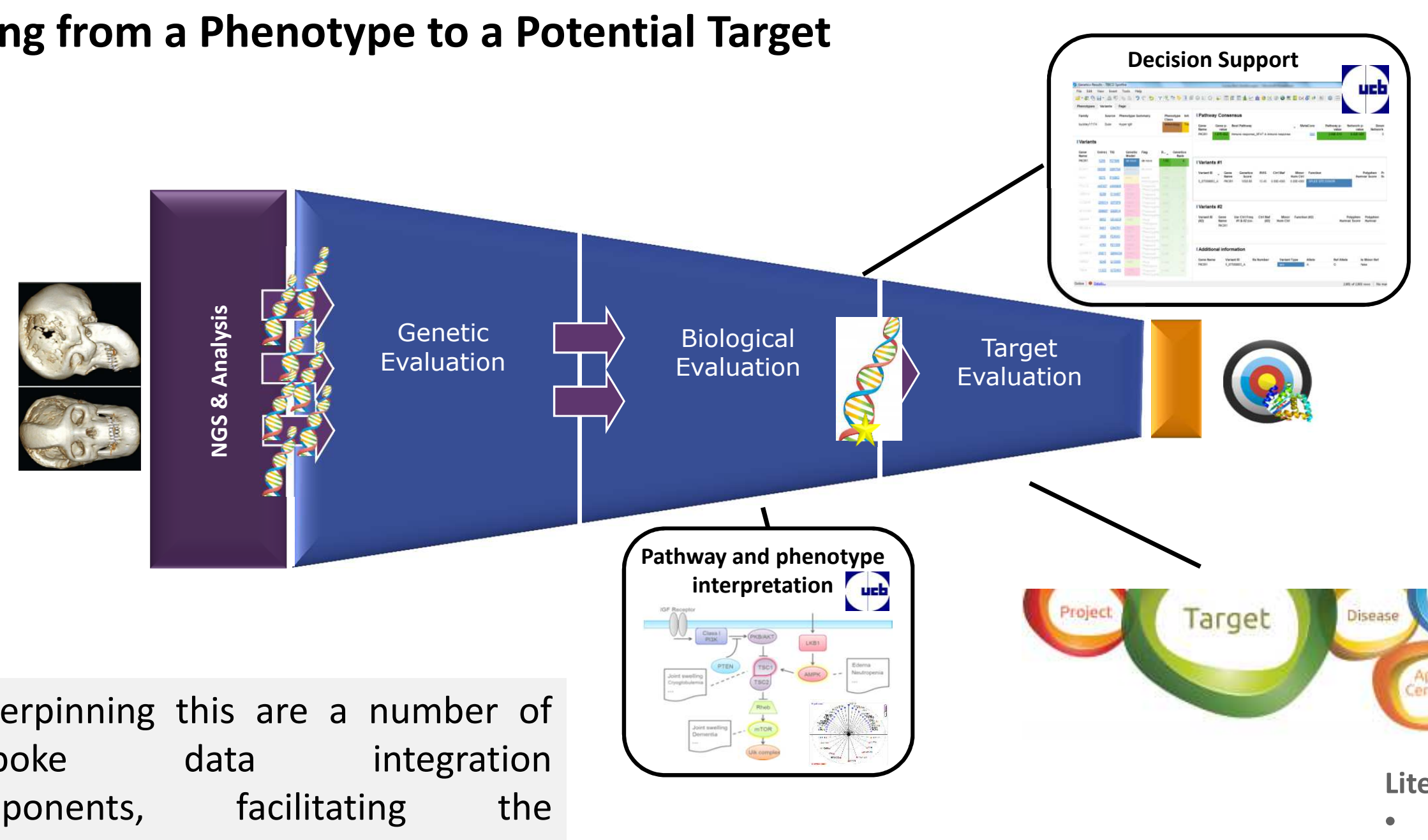
http://www.ebi.ac.uk/industry/private/industry-workshop/2015/09/enabling-translational-bioinformatician

The above generalises a number of activities and highlights feedback from a broad community – the challenges for translational research fall into a number of areas, some of which will require knowledge gaps to be closed, others that call for regulation and standards to be strengthened / established, while others are dependant on enhancing technical solutions and leveraging the experience we have generated to date in the industry to deal with the ever growing diversity, complexity and volume in the data space.

The internal use cases below exemplify problems like target identification, validation, biomarker identification and patient stratification which are encountered increasingly frequently and by a growing and diverse community, dealing with fragmented and varied data sources and requiring integration at multiple levels. The UCB data science community is a loosely coupled group distributed across a range of teams ranging from Exploratory and Global Clinical Statistics to Bioinformatics and Modelling and Simulation as well as inside the IT organisation and others.

Data integration technologies that facilitate these analytical approaches is as diverse as the problems themselves, varying in use across teams, from manual efforts, such as the use of excel, to the adoption of graph and semantic technologies. As the applications grow internally there is increasing demand for standardardizing data exchange mechanisms and stable technology platforms on which to build; therefore, the right investment choices are critical

Moving from a Phenotype to a Potential Target



Uncovering the causal mutations that underlie rare genetic disorders offers the opportunity to interrogate the mechanistic underpinnings of disease processes that are observed in a wider population of patients. Integrating a suite of variant calling and scoring mechanisms with a downstream annotation tool enables precise prediction of potential relationships.

Underpinning this are a number of bespoke data integration components, facilitating the visualisation of phenotype and network relationships is a graphdb environment which optimises the querying and provides an intuitive management of graph relations.

To support the integration of disparate types of data and to enable the flexible evolution of the specification an RDF triple store and bespoke ontology were constructed.

Target Information Aggregation

Literature
- Chronologically organised literature feeds.

Function
- Expert functional annotations.

Competitor Drugs
- Overview of clinical development activities.
- Detailed clinical trial information.

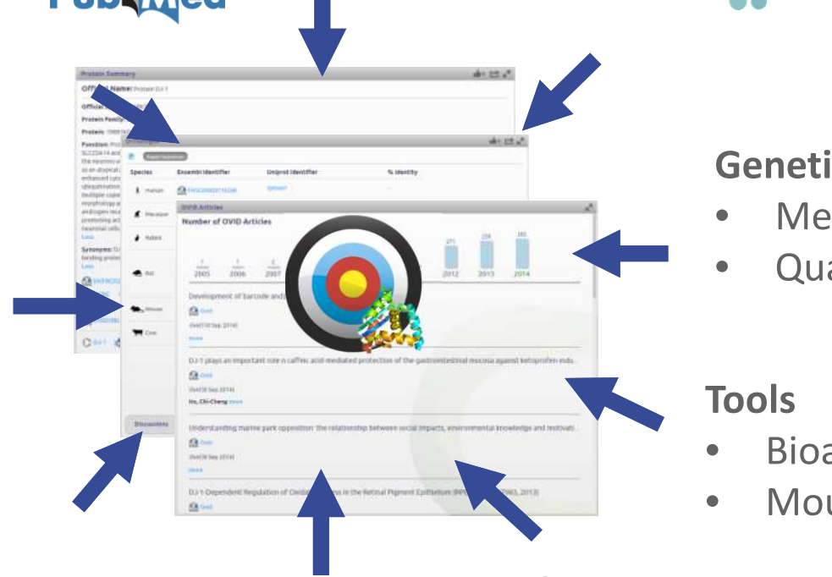
Genetics
- Mendelian traits
- Quantitative traits

Tools
- Bioactive molecules
- Mouse knock-outs

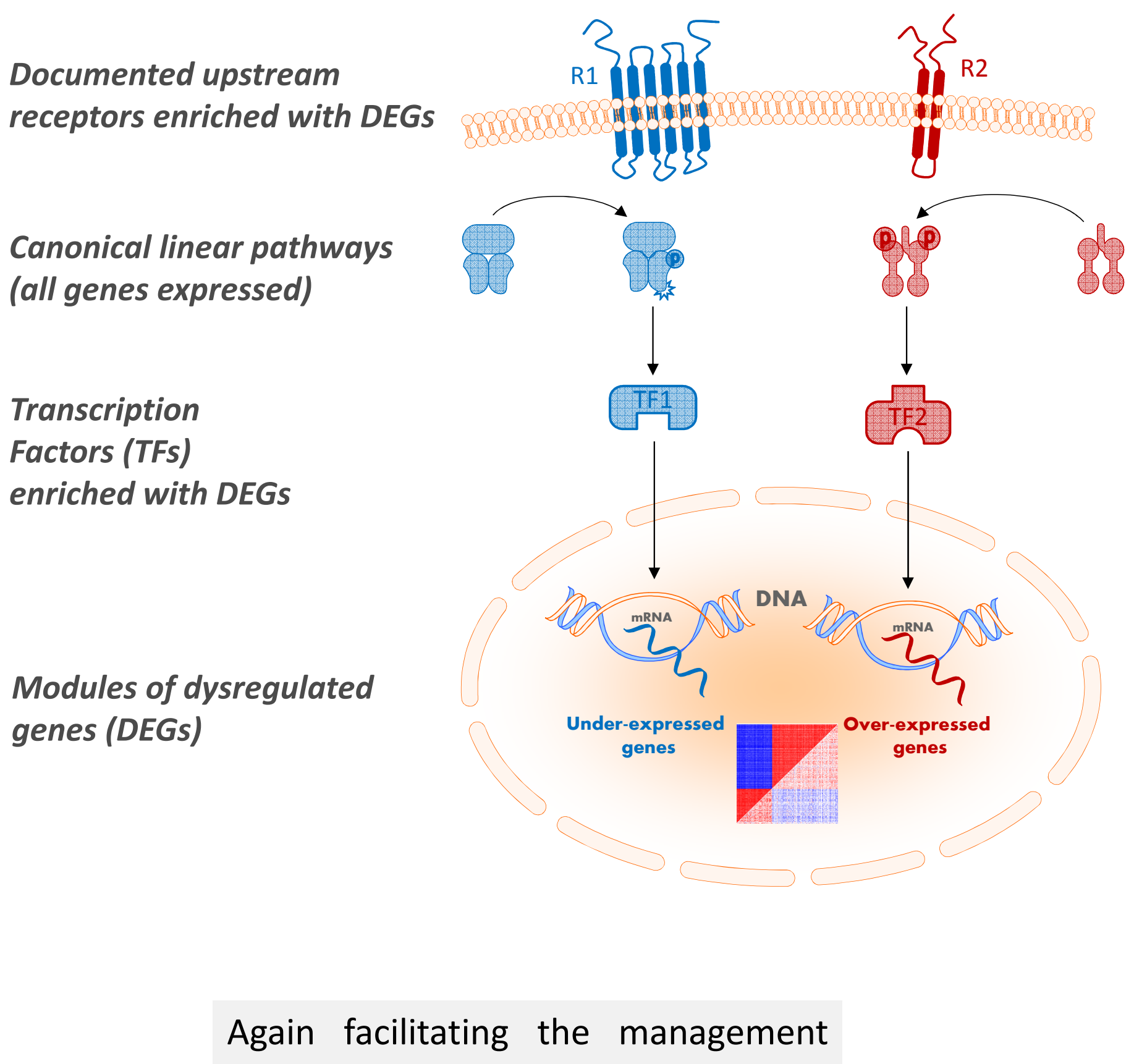
Pathways

Expression

Structure



Causal reasoning: identifying druggable targets acting through transcriptional regulation of gene modules


Documented upstream receptors enriched with DEGs
Canonical linear pathways (all genes expressed)
Transcription Factors (TFs) enriched with DEGs
Modules of dysregulated genes (DEGs)
R1
R2
TF1
TF2
DNA
mRNA
mRNA
Under-expressed genes
Over-expressed genes

Continuing the causal theme, this approach builds on the relationships between receptors, transcription factors and the proteins that are downstream of their influence.

In a study which integrated high throughput phenotype monitoring with gene expression data the expressed genes were classified into modules of transcriptional co-regulation. Then transcription factors and plasma membrane receptors potentially involved in the regulation of these modules were identified applying a simple causal reasoning approach based on gene set enrichment.

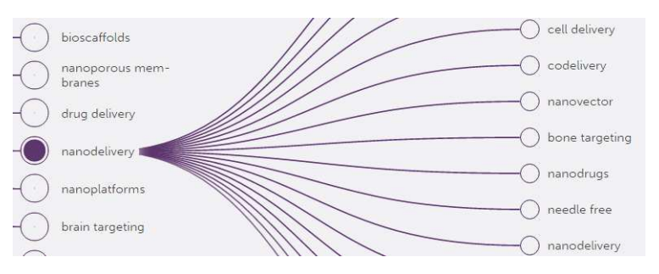
Those receptors controlling modules which exhibit differential behaviour in disease model cases compared to controls represent potential drug targets linked to disease progression.

Again facilitating the management and visualisation of network relationships is a graphdb environment.

Exploring content: trend mapping

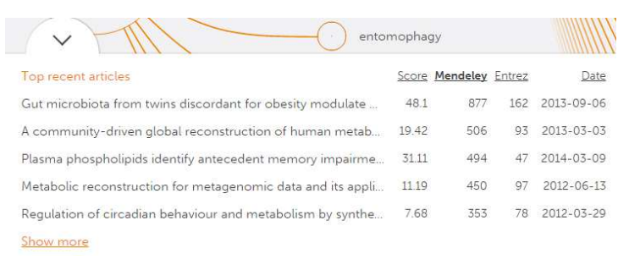
AWS and Apache Spark cluster facilitated rapid development and deployment of a computing environment capable of facilitation these analyses

Concept Explorer



- Clustering of related terms into topics and ability to explore a technology area by navigating topics & terms

Content Explorer



- Links to articles for a topic or term with information on their relevance / importance

Trend Explorer

- Compare trend information for a topic or term from 2010 -2015 for PubMed, PatBase & Google Trends

Institution Explorer

- Identification of companies & academic institutions involved in a topic or term area

Utilising NLP methods to explore topics to support navigation of content from multiple sources.
In this example an environment was constructed to facilitate the generation of word2vec vectors mapping concepts across articles.

Inspired by patients.
Driven by science.

Adaptability and flexibility have been important features of our investment, the evolution of technology in this space has moved rapidly and enabled bespoke development of efficient solutions to some of the challenges. As a counterpoint to this however, some of these solutions should be considered prototypes and there is a danger of growing reliance on these less-than-enterprise-ready components. Larger more comprehensive platform investments have not been made and we risk fragmentation and not being able to enable the growth in translational activity. Additional issues around the requirements for security and privacy that surround the use of clinically derived data are particularly challenging for an ad hoc infrastructure.

Currently, there are ongoing discussions to take steps to build federated approaches to bring the clinical and pre-clinical data environments together which will address some of the integration challenges but in doing this we potentially limit the innovation and flexibility that has been an enabling factor.

As we continue to struggle with the incompleteness of the underlying logical associations between data types, based on the evolving models of biology, flexibility should remain in the hands of the data scientists to continue to evolve their mapping and transformation strategies, the trick will be supporting this.