Neuroinformatics at GSK and Application to Target Identification and Validation

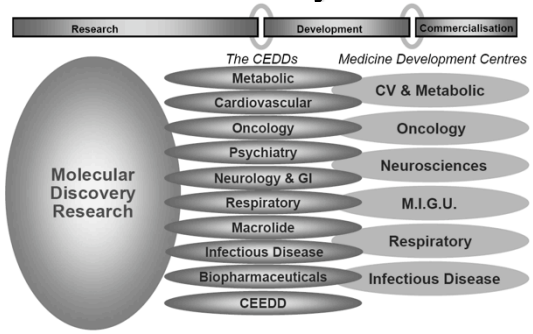
Chris Larminie

MDR Informatics

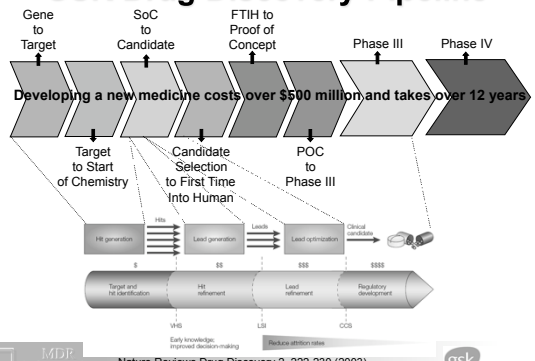
Presentation Summary

- QUICK overview of GSK R&D
- Integrating data
- Network / pathway analysis
- Neuroinformatics and issues pertinent to this field
- Informatics and Industry

R&D's Structure for Innovation & Flexibility



GSK Drug Discovery Pipeline



Nature Reviews Drug Discovery 2, 222-230 (2003)

MDR Informatics Areas

- Gene Identification, validation & classification
- HTS support, compound collections, (sub)structure searches, hard/soft filtering, HIT identification
- Data integration, application development
  - Unless integration is needed for a specific analysis project
- LIMS, data collection and pipelining
- Platform support (primer designs, SNP assay selection, etc.)
- Data analysis and interpretation
  - Sequence analysis
  - Platform data analysis
  - Network/Pathway analysis
  - Structural analysis
- New methods development, assessn

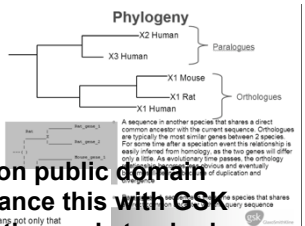
SYSTEMS BIOLOGY

- Progress in systems biology may be seen to rest on (*at least*) a three-way foundation:
  - **Pathways/Networks**: understanding the interactions that comprise biosystems
  - **Ontologies**: the effective representation of biological knowledge in all its richness
  - **Data Integration**: the ability to combine and analyse (*quality*) data from myriad sources

Data Integration:

What data is there to integ

- Gene Relevant Annotation
  - Names & Aliases
  - Putative function
  - Representative sequence
    - Genetic context
    - Tracking variants
- Ort
  - Pa
- Do
  - dif
  - ex

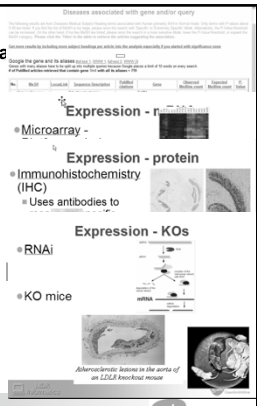


We are reliant on public domain sources but enhance this with GSK relevant information and standards

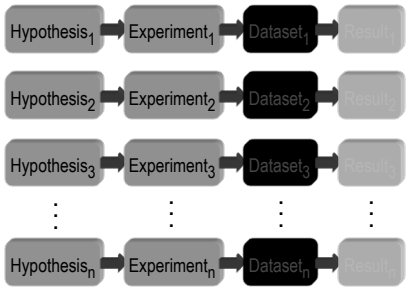
Data Integration:

What other data is there to integra

- Literature
  - key stimulant of interest
- Expression
  - is the gene expressed in tissues relevant to the disease?
  - Genome scale: Microarray technology
  - Focused follow up: TaqMan, ISH, IHC
- Phenotypic data:
  - Genetics
    - do people with *flaws* in this gene get sick more often?
    - Polymorphisms (SNPS) – genetic association, linkage analysis
  - Mouse KO/Transgenics
    - Identifying pathways from genetics
  - RNAi *in-vitro*/*in-vivo*
- Protein Structure analysis/modelling



Hypothesis-Driven Research

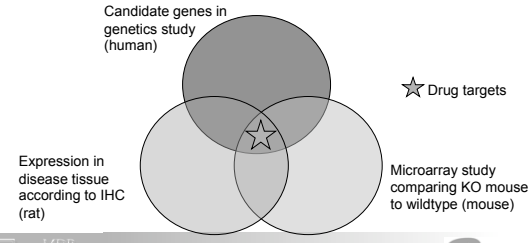


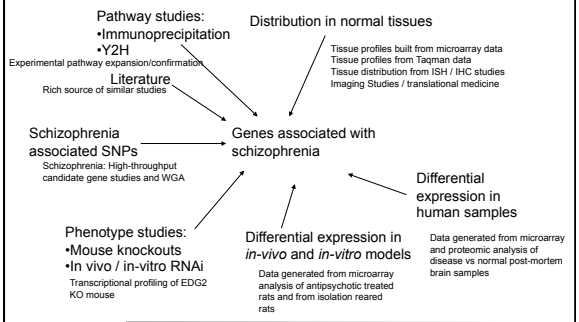
Combining validation data for finding targets – integrating validation data

Data from different sources can be combined to be more powerful than one source of data used in isolation

Permits prioritisation of targets to progress based on genetic, genomic and tractability data

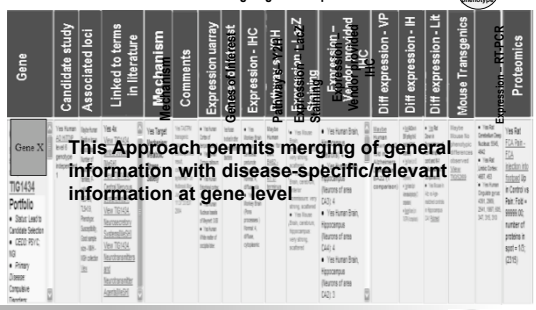
Represents a systematic and stringent approach





Disease view (PSTUD)

Different lines of evidence to find a drug target for a specific disease:

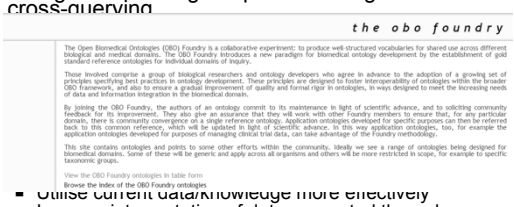


This Approach permits merging of general information with disease-specific/relevant information at gene level

What is required

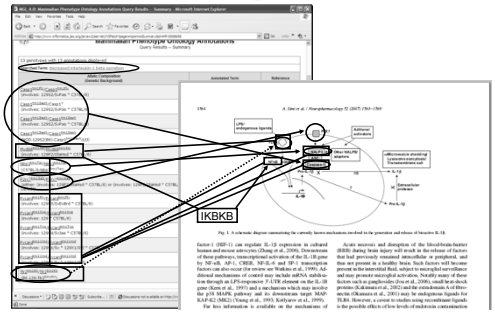
- Standard sample management/handling
- Standard data generation methods
- Standard QC analysis methods
- Standard significance/reporting cut-offs
- Agreed common standard for cross-reference
  - Gene?
  - Protein?
  - Genomic position?
- Pragmatism!
  - Balance standardisation with the need for the bespoke, avoid needless digression from established protocols
    - Some detail is inevitably lost through standardisation, determine cost-benefit and tactical versus strategic
    - If reasonable, report results from non-standard analyses and annotate as such
- Where one size does not fit all…

Ontologies

- Build standard vocabularies to accommodate these different data
- Effective and standardised representation of biological knowledge to permit mining and effective cross-querying
- 

- Utilise current data/knowledge more effectively
- Improve interpretation of data generated through new initiatives cited above

Ontologies are powerful tools for connecting data



RED: Genes with an IL-1B secretion KO phenotype that relate to the inflammasome
GREEN: Genes with an IL-1B secretion KO phenotype that link directly to components of inflammasome

Pathway/Network Biology and Drug Discovery

- Pathway identification, extension, inference, modelling *can* help for:
  - Novel targets and biomarkers identification
  - Disease understanding
  - Mechanisms of action (MOA) understanding
  - Potential safety concerns
  - Combination therapies
  - Alternative indications
- Permits collapse of pertinent data around a series of genes linked by a common biological context
- Permits identification of common, known pathways represented in multiple platform datasets
  - Pathways are often more stably represented than their constituent genes
- Permits identification of "novel" disease mechanisms from datasets
- Essential for data driven polypharmacological approaches

Pathway/Network Analysis

- Do we have sufficient data within Neurosciences to power pathway/network based analysis approaches?
  - Advent of platform based approaches to studying neuroscience has dramatically increased data space over recent years
    - Are these data of sufficient quality?
      - Reproducibility/variability
    - Can we gain sufficient access to it?
      - Data storage and structure
      - Data silos/ Data access
      - Data reporting, sharing and distribution
    - Data annotation: can we find it?
      - Semantics of data
- We need to generate more data and improve reporting, description and storage of these data to maximise its benefit
  - Better industry, academic partnering
  - Support ontologies
  - Enforce journal and grant awarding body rules

What data is particularly relevant to Neuroscience?

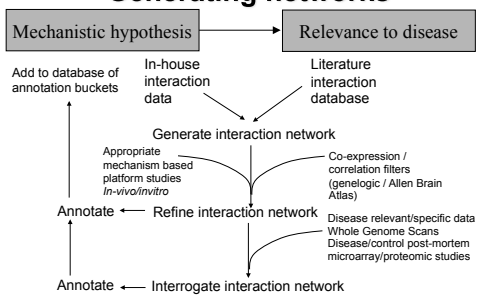
- **The 1990s: "Decade of the Brain"**
  - Recognition of social and economic burden of brain disease
  - Increasing confidence amongst research community that brain disease is now a tractable problem
- **Are we only now realising the benefits of this?**
  - Broad Institute
  - Allen Brain Atlas
  - Genetic studies
  - Genomic studies
  - (Jackson Lab KO ontologies)
  - Neuroimaging and other translational medicine approaches
  - Better integration between these data / sites

Nature Neuroscience **2**, 487 (1999)

Pathway/Network Informatics

- Pathway Informatics (*and platform data*) require tools to analyze gene *sets* rather than individual genes, in the context of very large databases of **known** relationships and interactions:
  - Relationship data compiled into gene 'buckets' from multiple sources in multiple 'universes'
    - *involved in the same pathway, part of the same network, coexpressed, genetically linked, phylogenetically related, …*
    - *In excess of 1m currently available*
  - Interaction data from different sources *binds, regulates, phosphorylates, degrades*
    - *Gene/protein/metabolite/compound*
- And determine if those observations are significant

Generating networks



Buckets generated through automated methods far outweigh those derived from manual processes outline above. Source data and assumptions affect quality

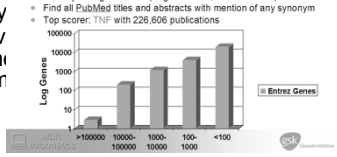
Literature

- Our molecular mechanistic understanding of CNS is improving
  - genes" returned (B4):
  - 1977-1986: 96
  - Gene Ontology
  - We can improv
    of molecular m
    elucidated from
  - BUT…

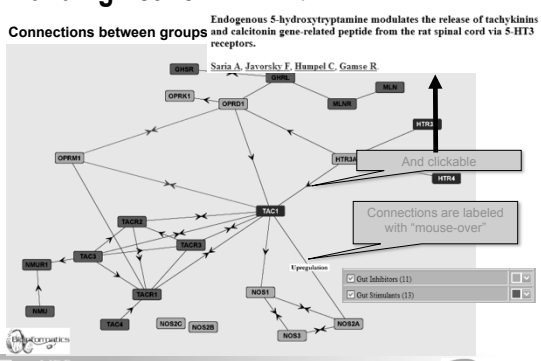
Using literature knowledge to identify CNS disease mechanisms will bias interpretation of your dataset

ure?

Rat orthologues
- Enumerate all synonyms for each, remove any synonym that may not only be a gene name (English words, Medical terms)
- Find all PubMed titles and abstracts with mention of any synonym
- Top scorer: TNF with 226,606 publications

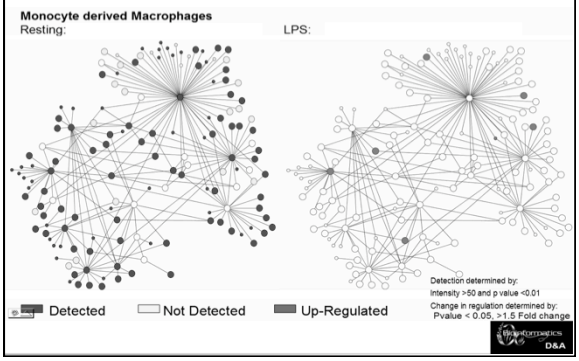


Building networ[illegible] from literature

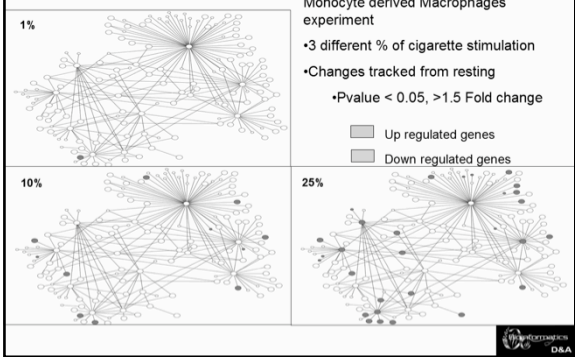


Identification of Pathways from Network

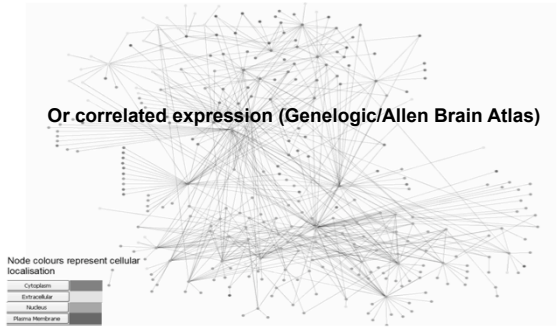
- Overlay specific cell type data to help with comparison of model systems/ cell lines



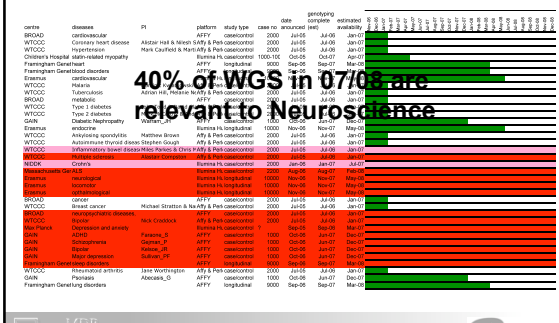
Dose Response Effect on Network



Compartmentalisation can help refine expanded Network

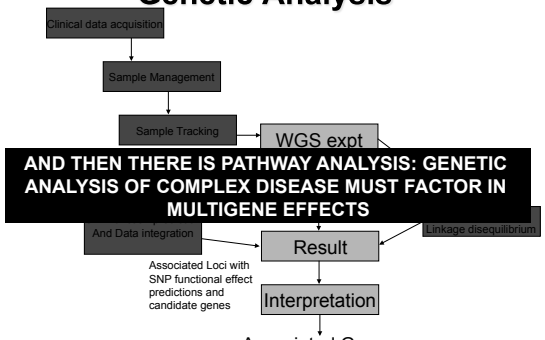


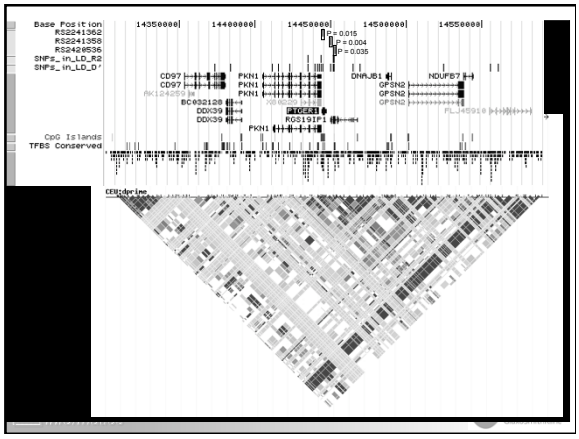
Whole Genome Scans



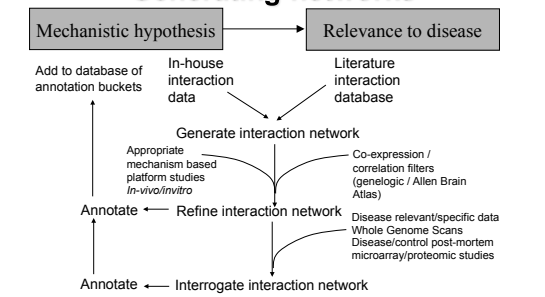
40% of WGS in 07/08 are relevant to Neuroscience

Genetic Analysis



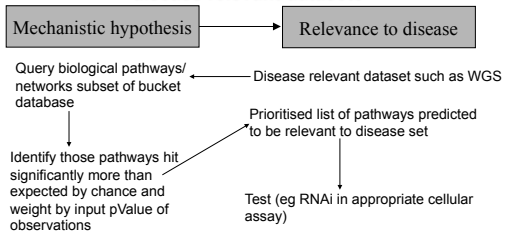


Generating networks

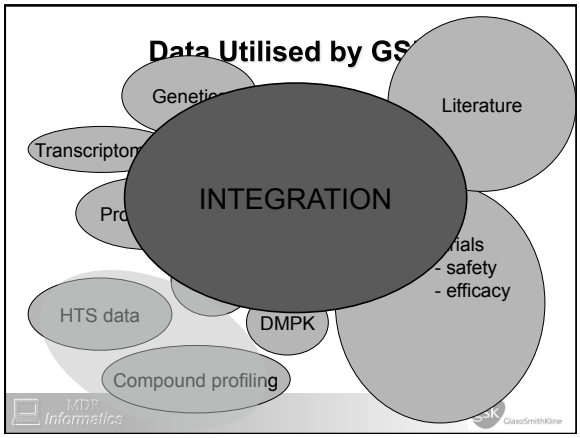


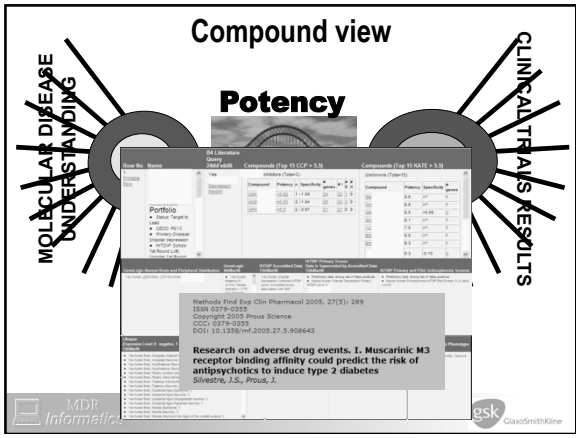
Buckets generated through automated methods far outweigh those derived from manual processes outline above. Source data and assumptions affect quality

Identifying significant biological processes within complex disease relevant datasets



BUT: Which gene(s)?
A pathway may be significant but individual genes may have small component effects, may need to knockout/down >1
Many significant pathways/networks identified, do we test in each?

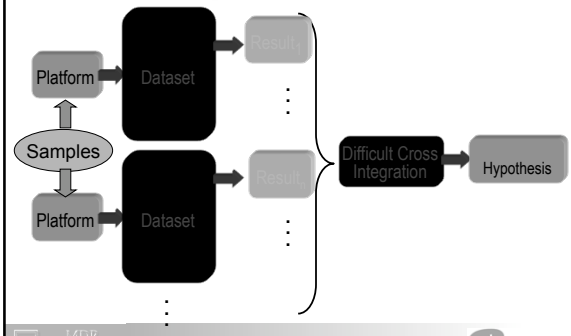




Thanks to

- Past and present
  - CB / MDR Ix staff
  - MDR IT staff

Multiple Platforms



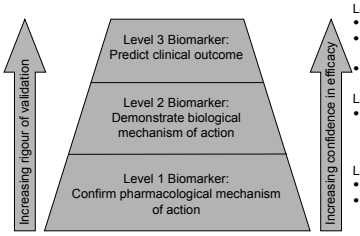
- BIOMARKERS / MULTIPLATFORM APPROACHES

Biomarkers

- "A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention"

  » NIH Biomarkers Definitions Working Group
- Improve understanding of the therapeutic area and disease pathophysiology
- Increase understanding of MOA of compound(s) and target both pre-clinically and clinically
- Identify profiles characteristic of unwanted toxicity in early drug candidate screening
- Provide evidence of drug efficacy and safety in early trials
- Enhance experimental and clinical design
- Provide means to make better clinical trial decisions earlier through use of surrogates

Clinical biomarkers in drug discovery



Level 3
- Surrogate endpoints
- Predictive power for a desired clinical outcome
- RNA viral loads in HIV

Level 2
- Confirm drug is producing desired pharmacodynamic effect directly related to it's potential for efficacy

Level 1
- Proof of mechanism biomarkers
- Show drug is modulating biochemical/molecular target *in vivo*

MDR Informatics

GlaxoSmithKline

Identifying biomarker panels

- Pilot studies:
  - Depression
  - Alzheimers
- Bloods from case/control populations
  - Clinical parameters
  - Genotype
  - Proteomics
  - Transcriptomics
- Machine Learning approaches and training sets
  - "minimum panel of 9 analytes capable of segregating case/control with 78% accuracy"

MDR Informatics

GlaxoSmithKline