Bioinformatics at GSK: Application to Target Identification and Asset Progression

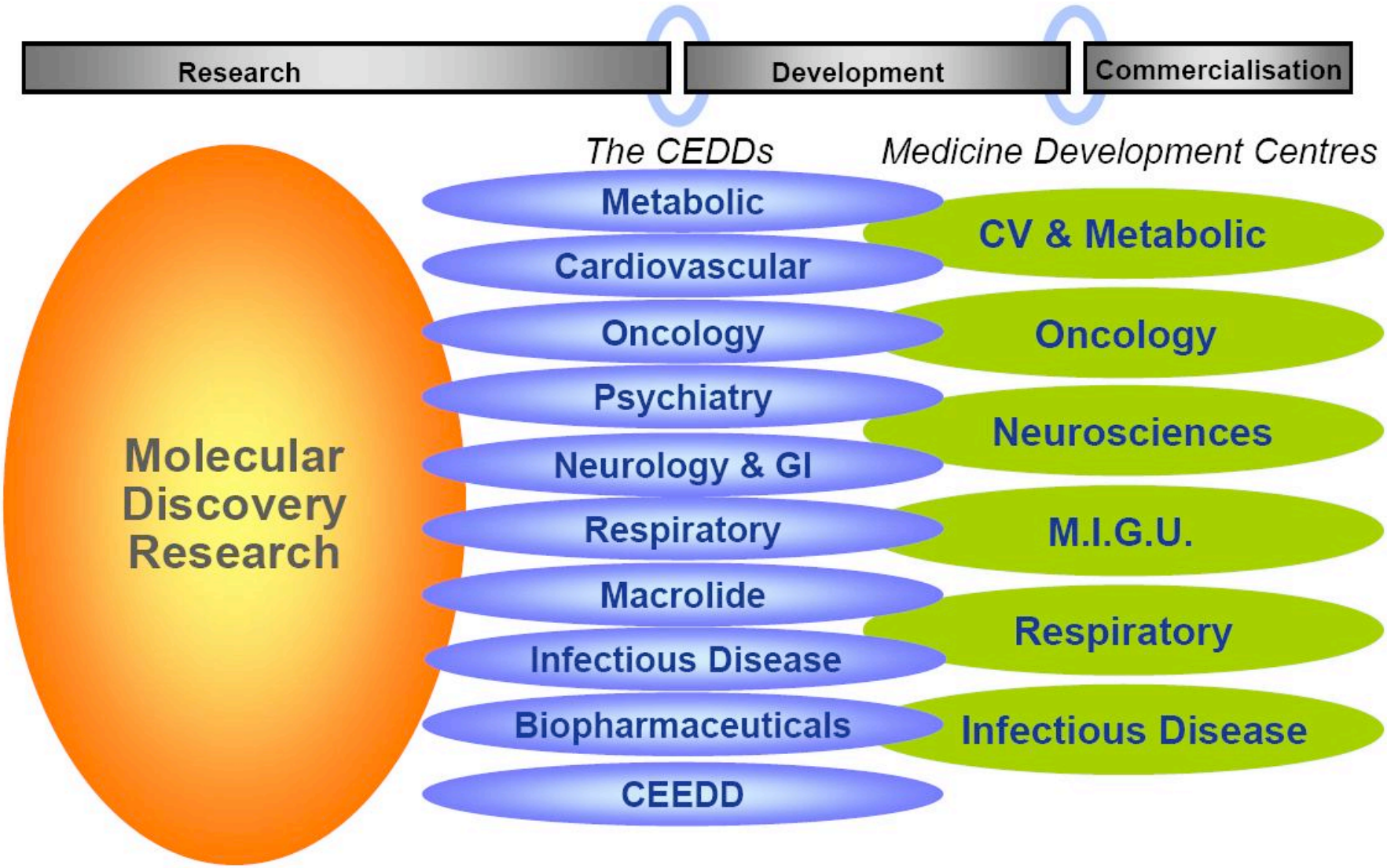
Chris Larminie

computational biology

Presentation Summary

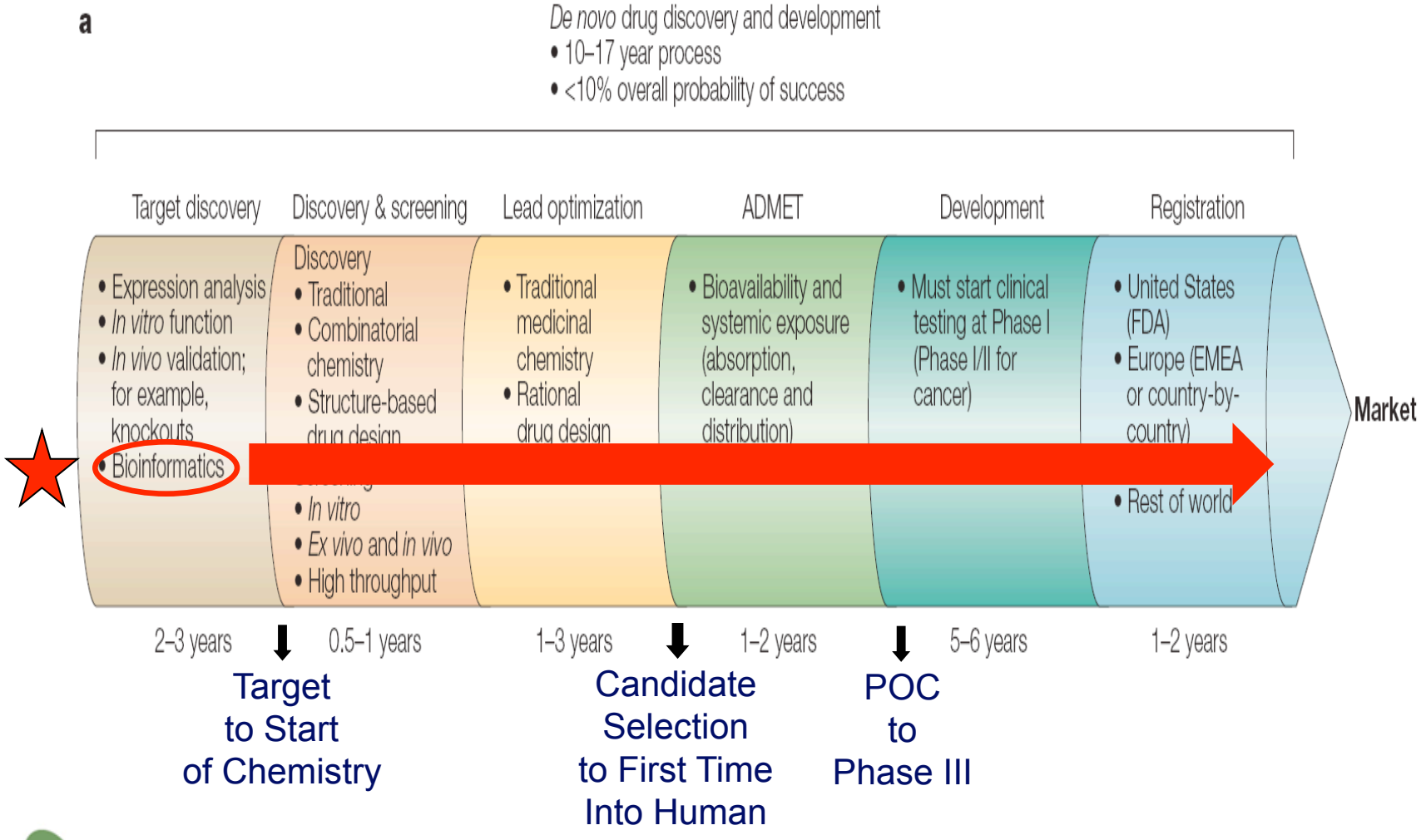
- QUICK overview of GSK R&D
- Integrating data
- Network / pathway analysis
- Neuroinformatics and issues pertinent to this field
- Informatics and Industry

R&D's Structure for Innovation & Flexibility



Drug Discovery Pipeline

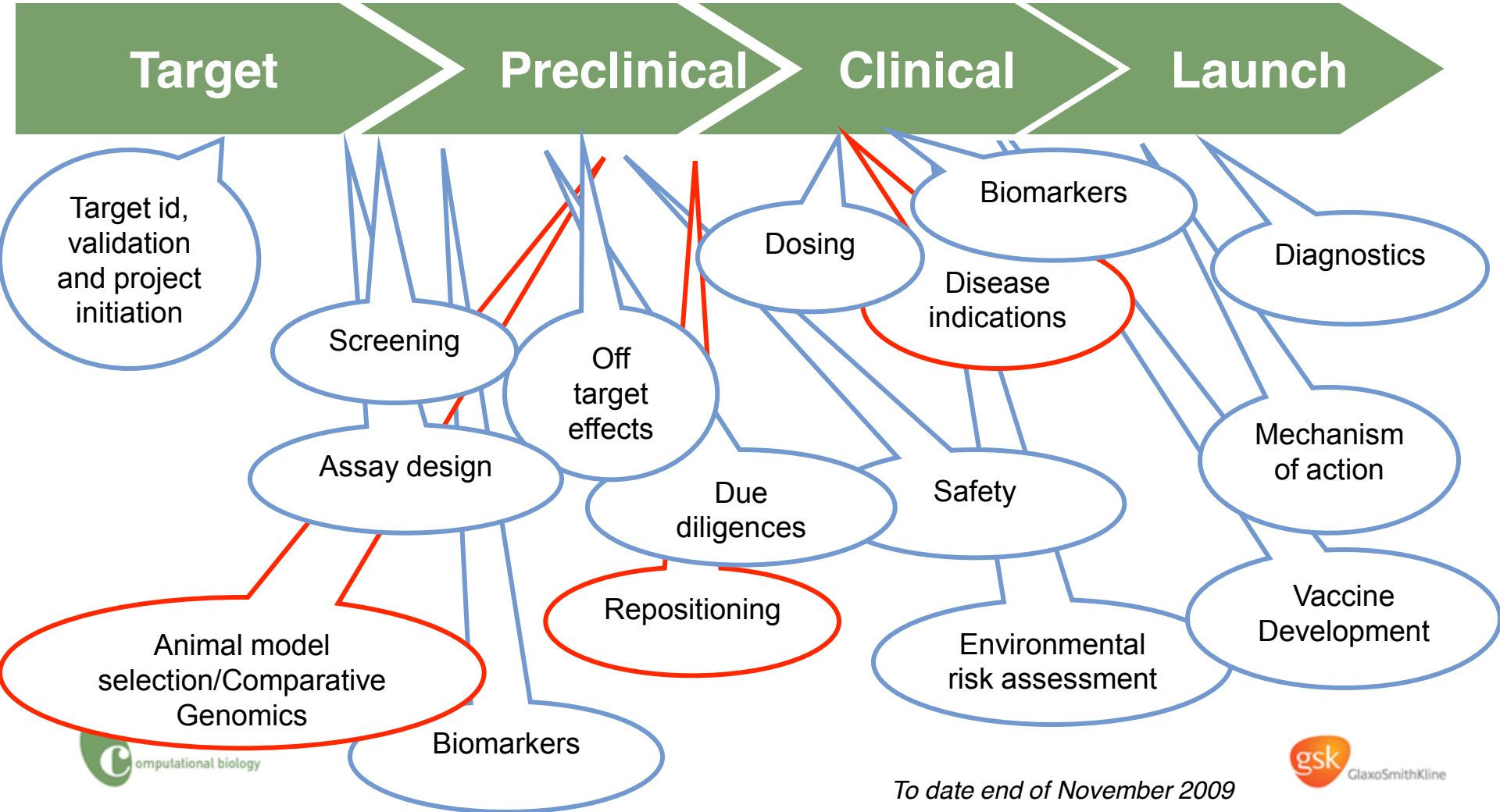
Developing a new medicine costs over $800 million



Adapted from Ashburn et al. Nature Rev. Drug Disc. 3:673

CB (Computational Biology) and the Drug Discovery Pipeline

- **CB is not only supporting early drug discovery: target identification, target validation etc…..**



To date end of November 2009

SYSTEMS BIOLOGY: A Driver for Integration

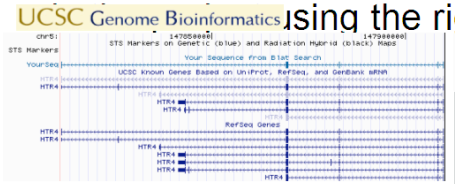
- Progress in systems biology may be seen to rest on (*at least*) a three-way foundation:
  - **Pathways/Networks**: understanding the interactions that comprise biosystems
  - **Ontologies**: the effective representation of biological knowledge in all its richness
  - **Data Integration**: the ability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

Data Integration:
What data is there to integ

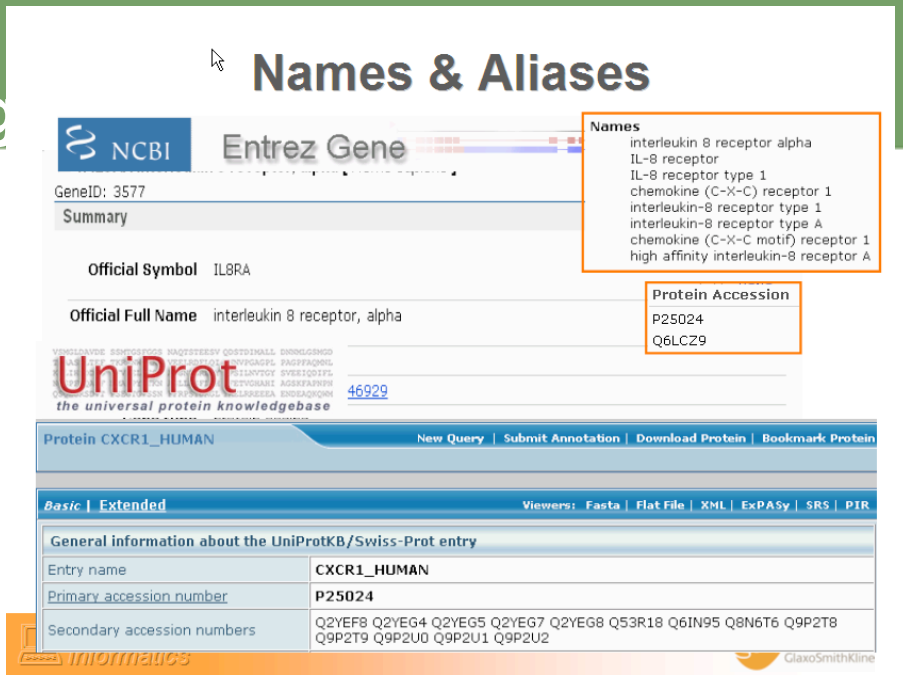
- Gene Relevant Annotation
  - Names & Aliases
  - Putative function
  - Representative sequence
    - Genetic context
    - Tracking variants
      - Polymorphisms / splice variants / mutations

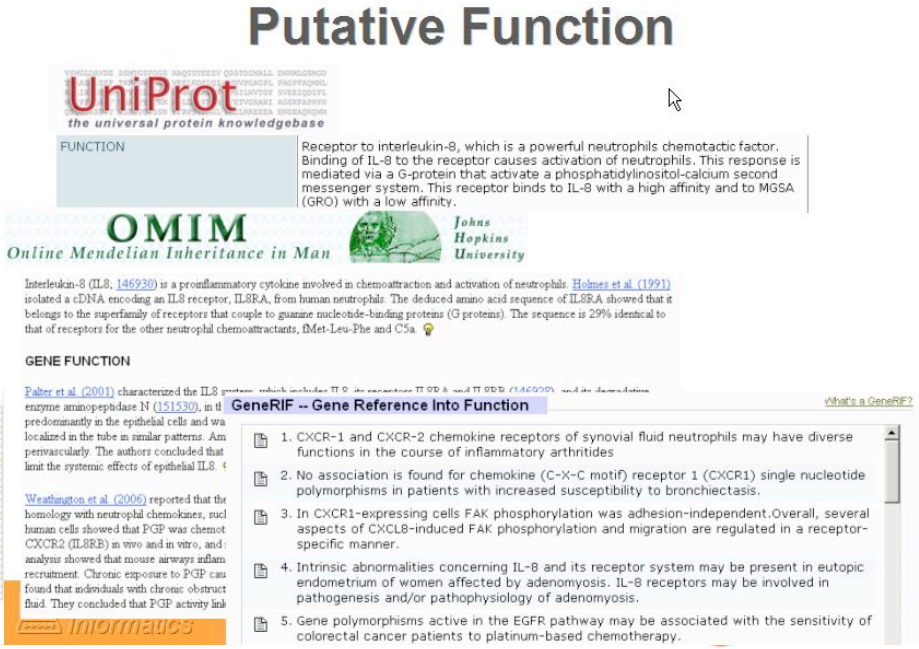
Representative Sequence

- Do splice variants exist? Do they have differing functions? Regulated expression? Polymorphisms?



using the right cloned



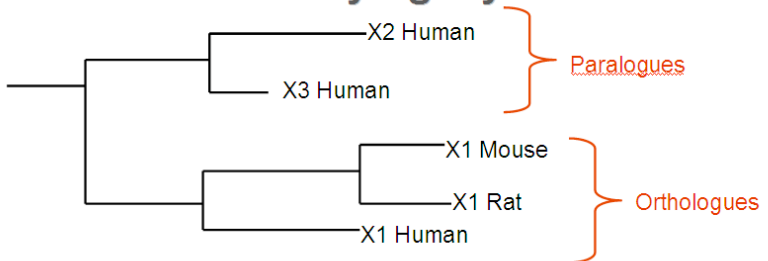


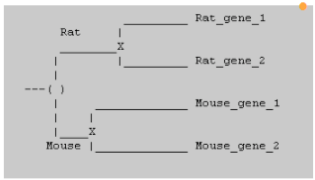
Data Integration:
What data is there to integrate?

- Gene Relevant Annotation
  - Names & Aliases
  - Putative function
  - Representative sequence
    - Genetic context
    - Tracking variants
      - Polymorphisms / splice variants / mutations

- Orthologues
  - Model organisms – allow study of the target in an animal to gain useful biological or genetic information relevant to the human gene
    - Function / disease relevance / toxicity / DMPK
  - Best defined by phylogeny and/or synteny, but reciprocal blast often used
  - Absence of clear orthologue may indicate organism is not a suitable model

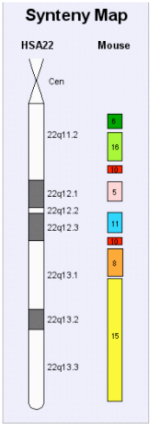
Phylogeny





- A sequence in another species that shares a direct common ancestor with the current sequence. Orthologues are typically the most similar genes between 2 species. For some time after a speciation event this relationship is easily inferred from homology, as the two genes will differ only a little. As evolutionary time passes, the orthology relationship becomes less obvious and eventually becomes ill-defined because of duplication and divergence
- Paralogue - A sequence in the same species that shares a direct common ancestor with the query sequence

Synteny

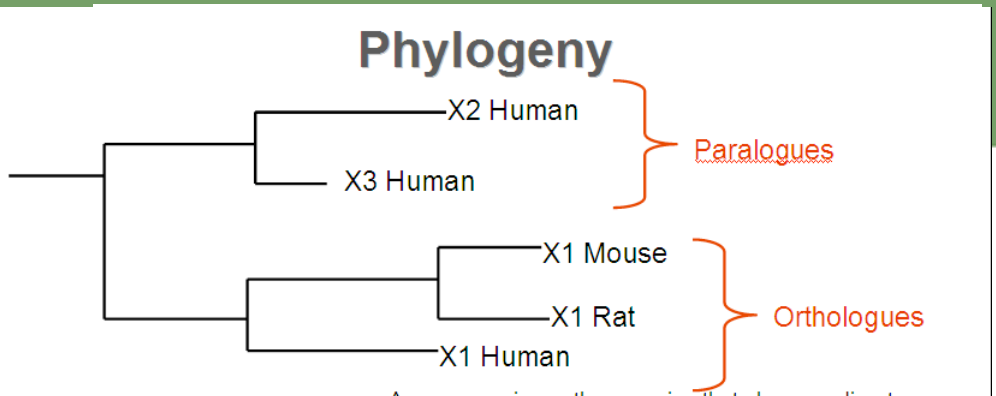


Synteny between species means not only that orthologous genes are present but that they are present in the same order on the genome, thus indicating common ancestry

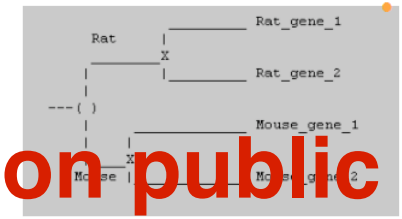
90.2% of the human genome and 93.3% of the mouse genome lie in conserved syntenic segments

It is possible to dismiss homology as arising from similar function, but it is not possible to dismiss synteny this way

Data Integration:
What data is there to i



- Gene Relevant Annotation
  - Names & Aliases
  - Putative function
  - Representative sequence
  - Genetic context

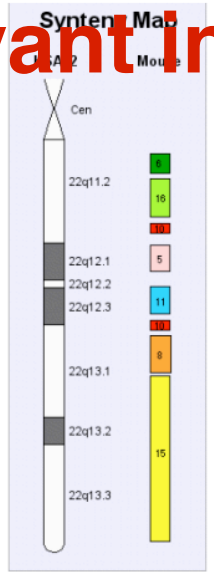


A sequence in another species that shares a direct common ancestor with the current sequence. Orthologues are typically the most similar genes between 2 species. For some time after a speciation event this relationship is easily inferred from homology, as the two genes will differ only a little. As evolutionary time passes, the orthology relationship becomes less obvious and eventually becomes ill-defined because of duplication and divergence

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

- Orthol
  - Mo in ge ge
  - Be bu
  - Ab or

Synteny



Synteny between species means not only that orthologous genes are present but that they are present in the same order on the genome, thus indicating common ancestry

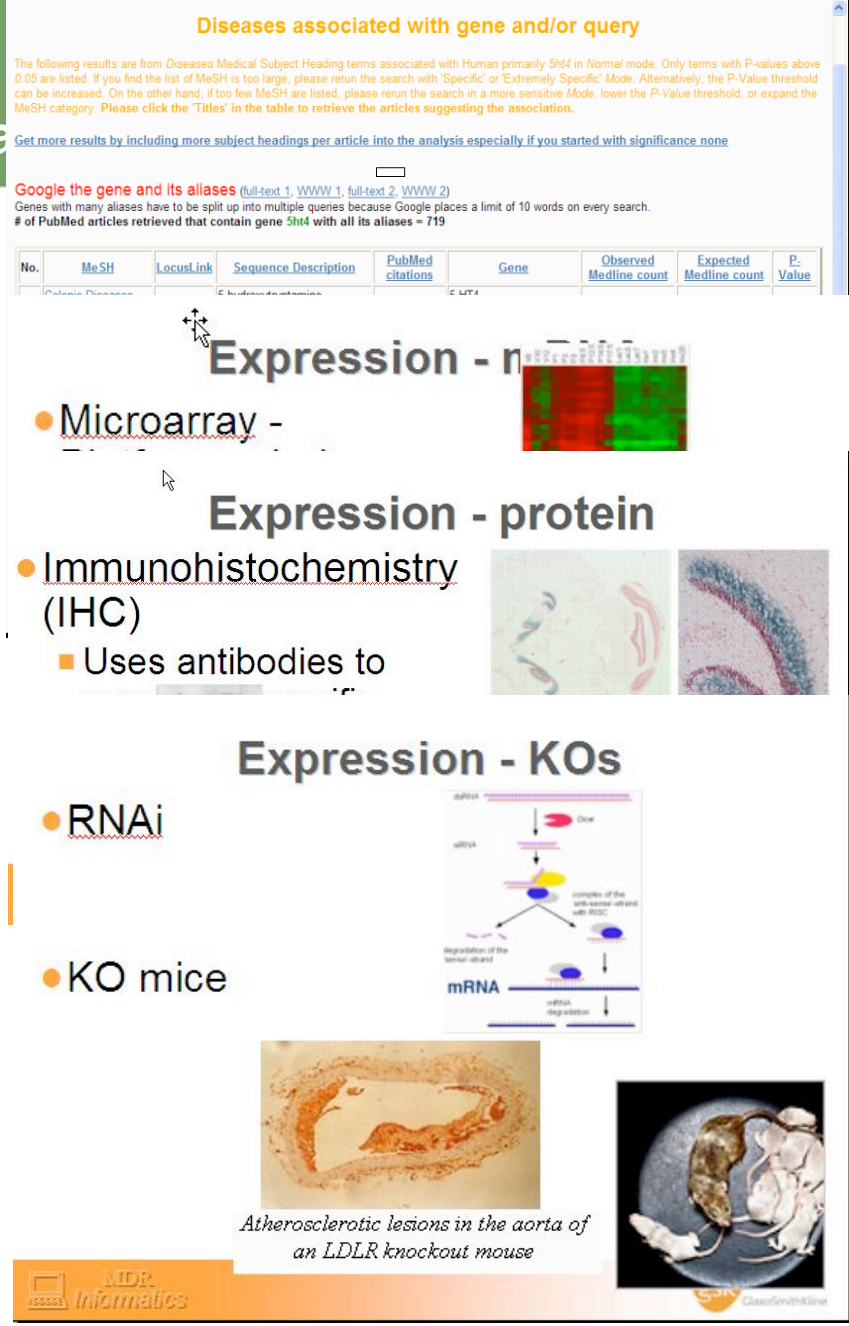
90.2% of the human genome and 93.3% of the mouse genome lie in conserved syntenic segments

It is possible to dismiss homology as arising from similar function, but it is not possible to dismiss synteny this way

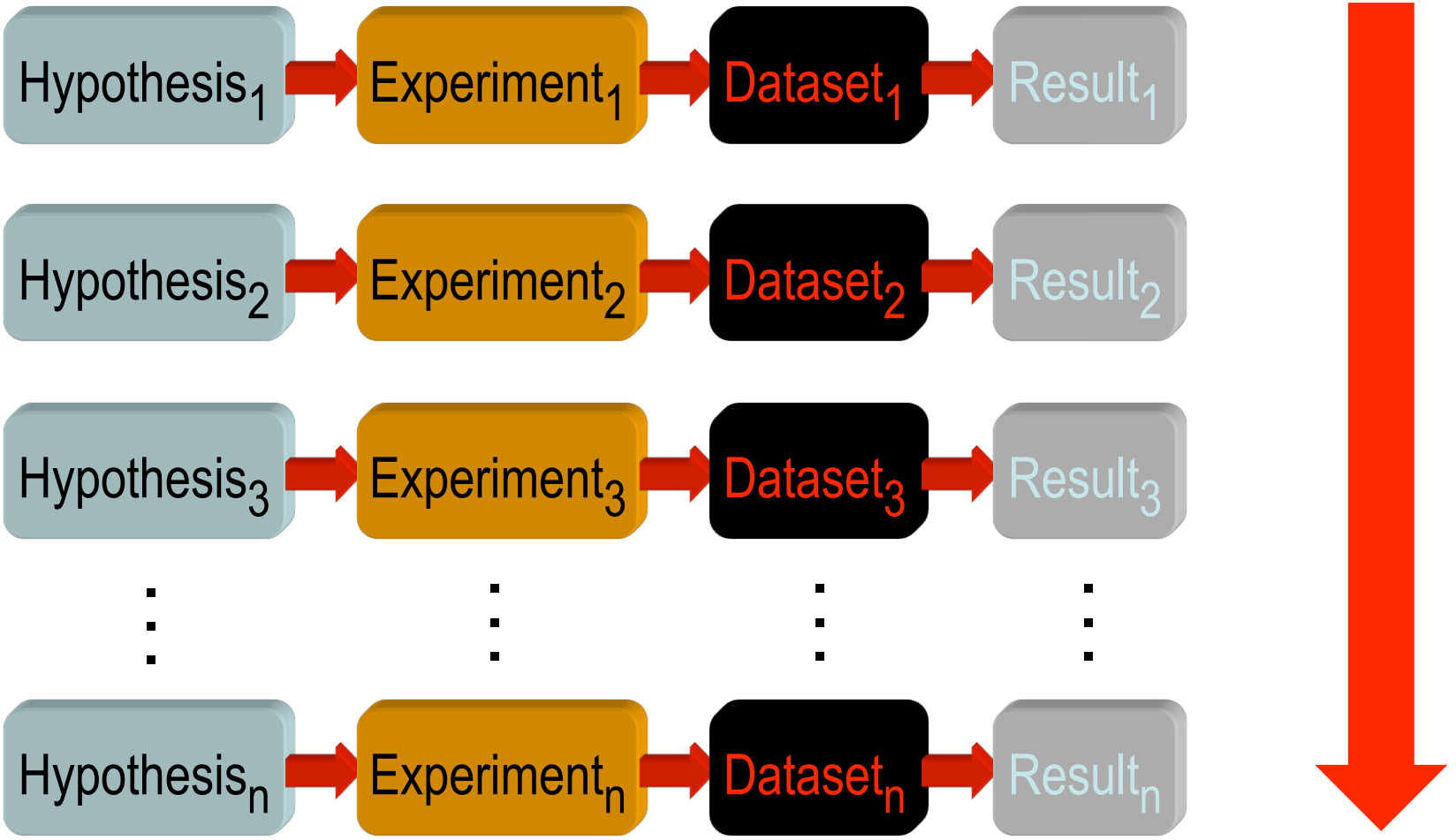
Data Integration:

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

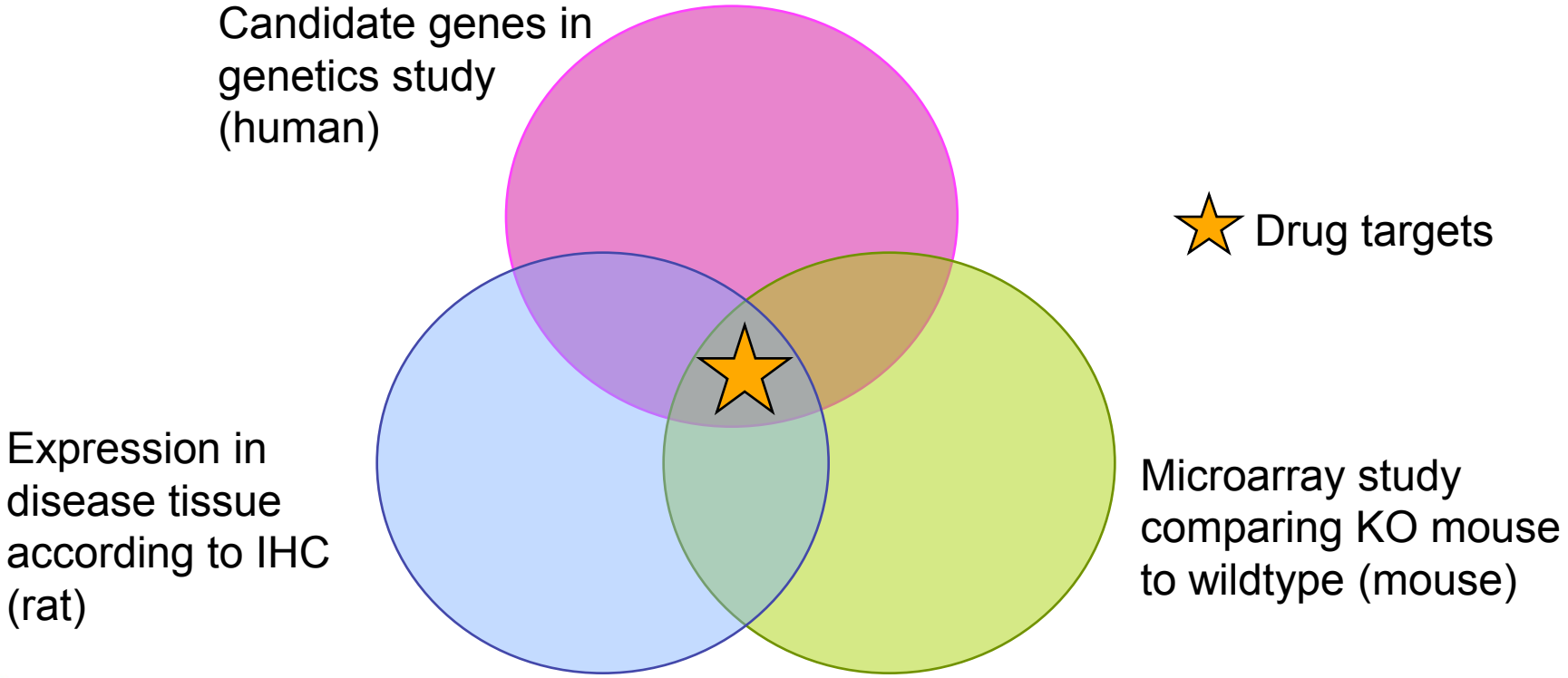


Hypothesis-Driven Research

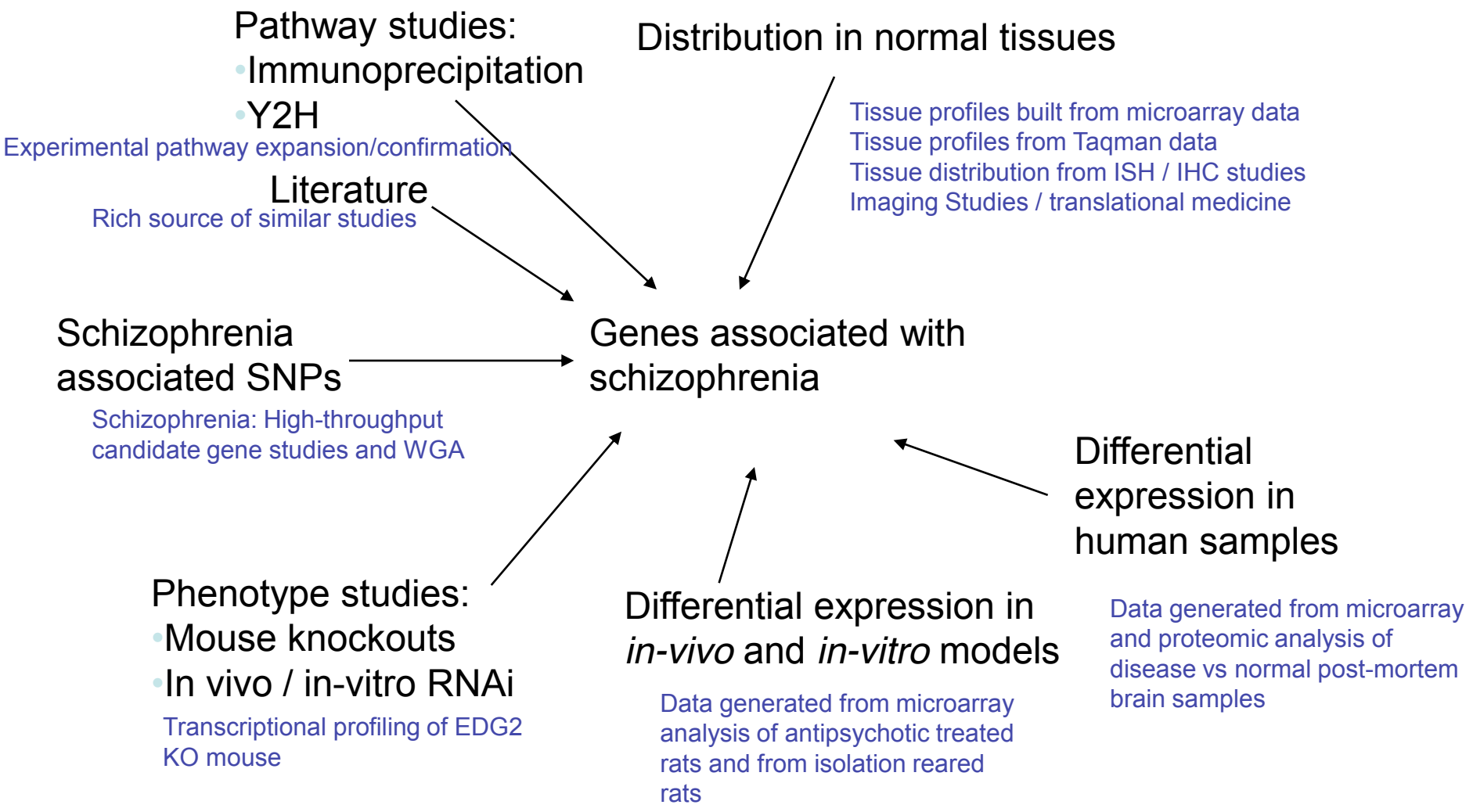


Combining validation data for finding targets – integrating validation data

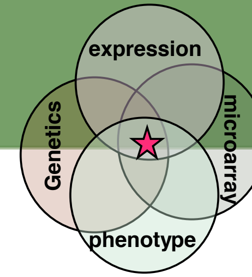
- Data from different sources can be combined to be more powerful than one source alone (reduced risk of false positives)
- Permits prioritisation of targets to progress based on genetic, genomic and tractability data
- Represents a systematic and stringent approach



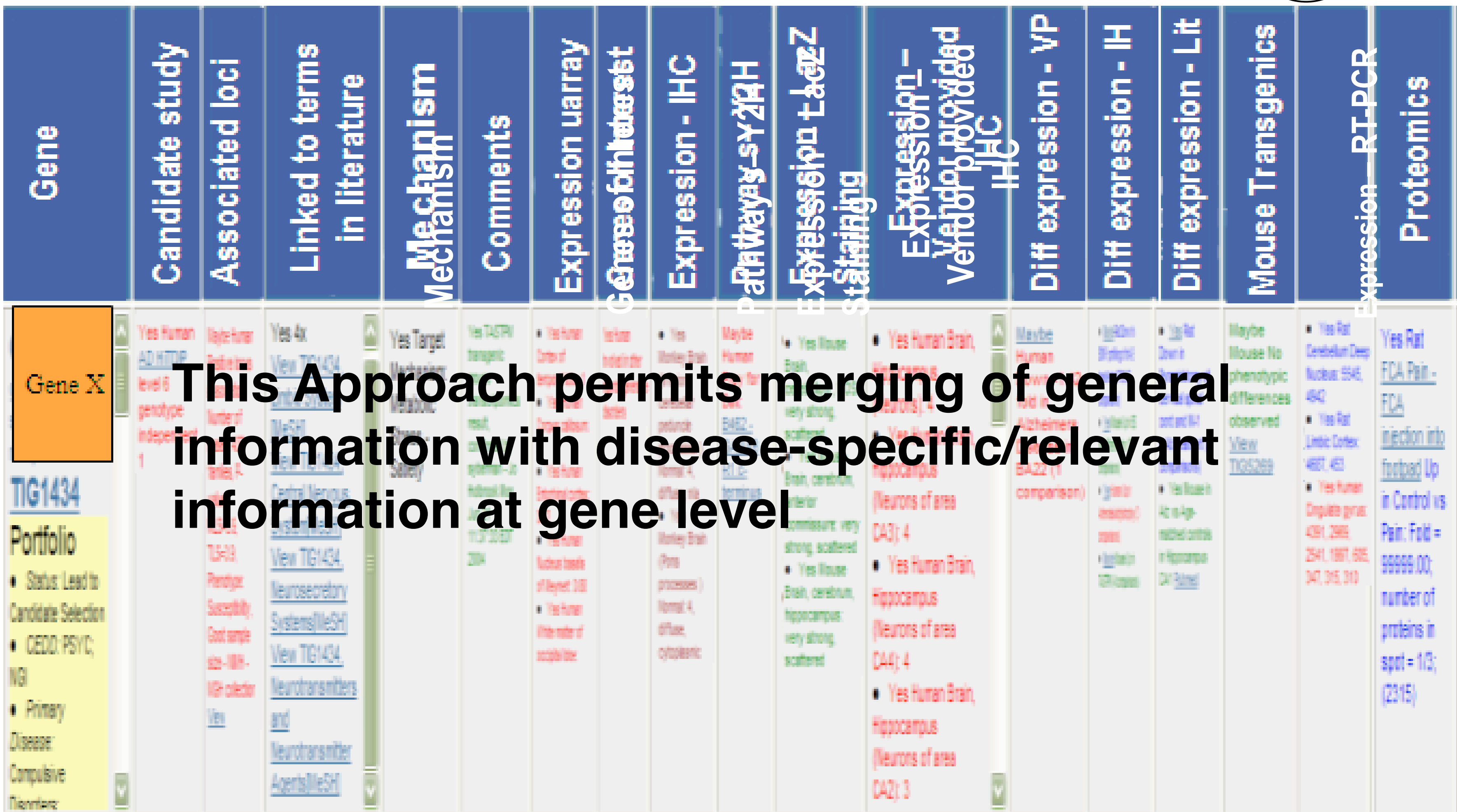
An example: Schizophrenia



Disease view (PSTUD)



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



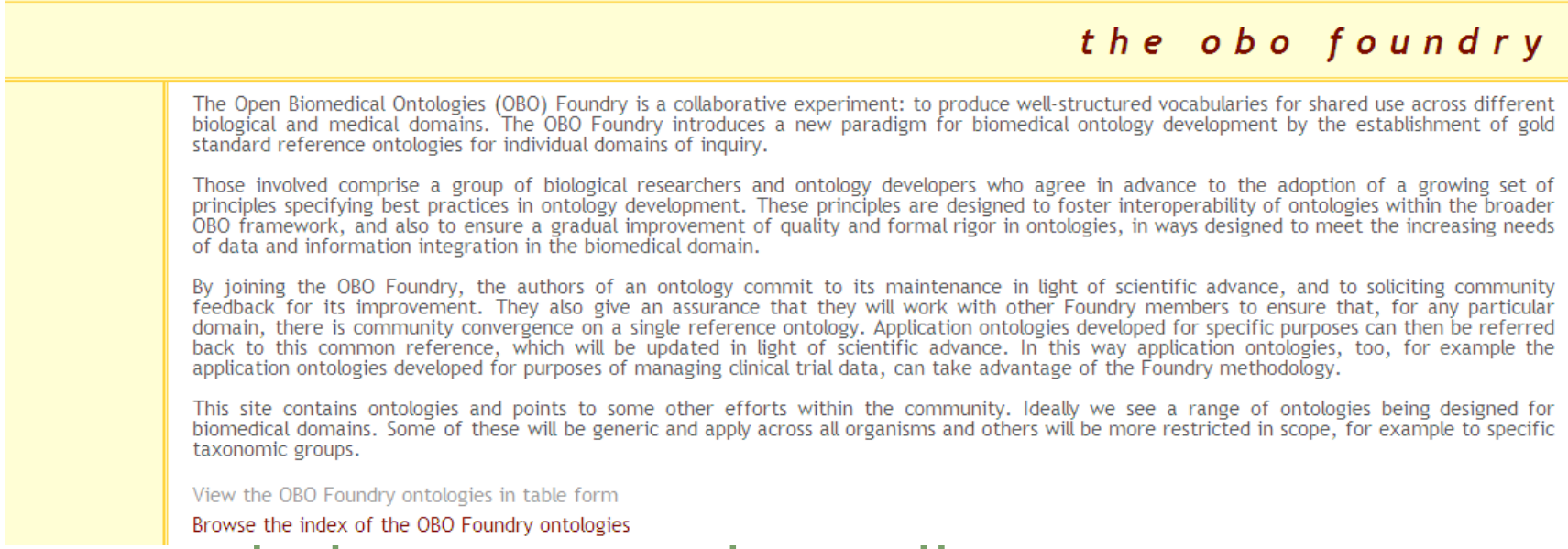
This Approach permits merging of general information with disease-specific/relevant information at gene level

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



the obo foundry

The Open Biomedical Ontologies (OBO) Foundry is a collaborative experiment: to produce well-structured vocabularies for shared use across different biological and medical domains. The OBO Foundry introduces a new paradigm for biomedical ontology development by the establishment of gold standard reference ontologies for individual domains of inquiry.

Those involved comprise a group of biological researchers and ontology developers who agree in advance to the adoption of a growing set of principles specifying best practices in ontology development. These principles are designed to foster interoperability of ontologies within the broader OBO framework, and also to ensure a gradual improvement of quality and formal rigor in ontologies, in ways designed to meet the increasing needs of data and information integration in the biomedical domain.

By joining the OBO Foundry, the authors of an ontology commit to its maintenance in light of scientific advance, and to soliciting community feedback for its improvement. They also give an assurance that they will work with other Foundry members to ensure that, for any particular domain, there is community convergence on a single reference ontology. Application ontologies developed for specific purposes can then be referred back to this common reference, which will be updated in light of scientific advance. In this way application ontologies, too, for example the application ontologies developed for purposes of managing clinical trial data, can take advantage of the Foundry methodology.

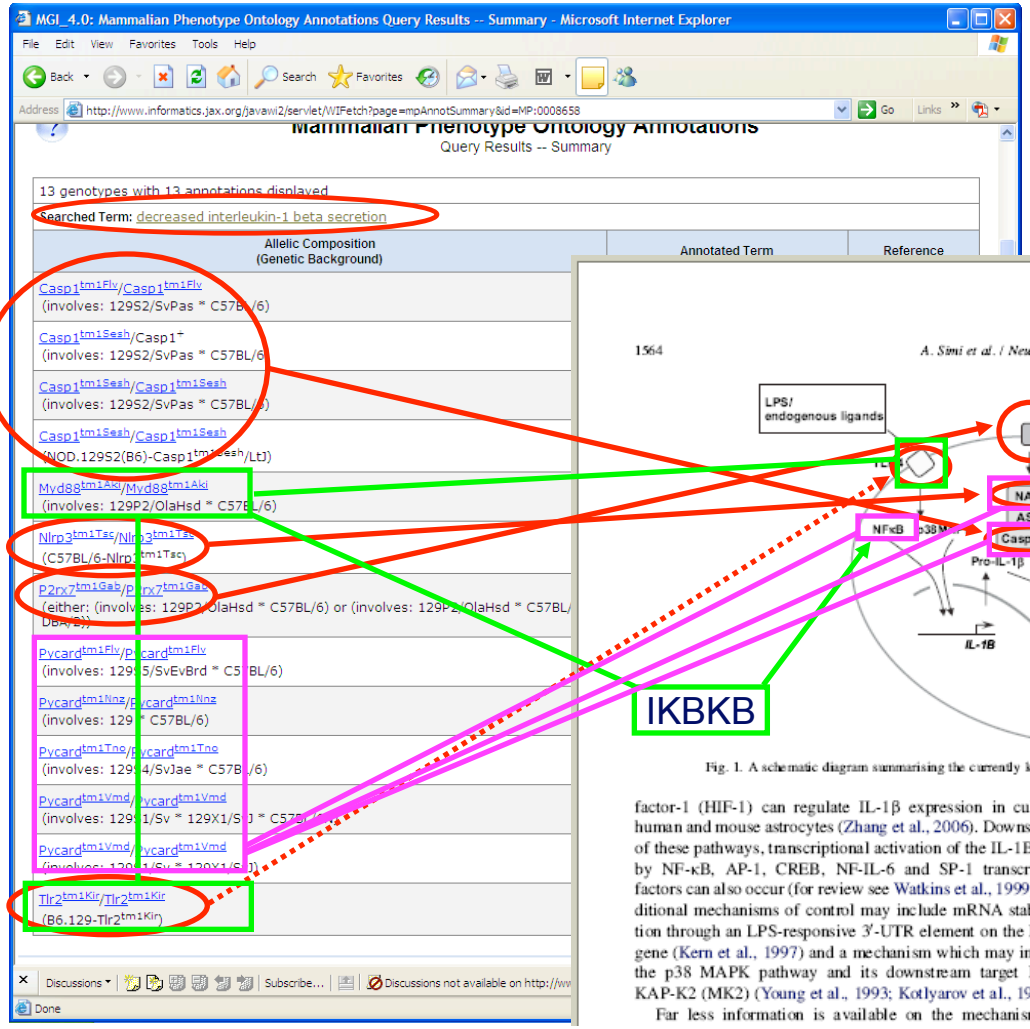
This site contains ontologies and points to some other efforts within the community. Ideally we see a range of ontologies being designed for biomedical domains. Some of these will be generic and apply across all organisms and others will be more restricted in scope, for example to specific taxonomic groups.

View the OBO Foundry ontologies in table form

Browse the index of the OBO Foundry ontologies

- generated in one area to another
  - 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
  - PISTOIA

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

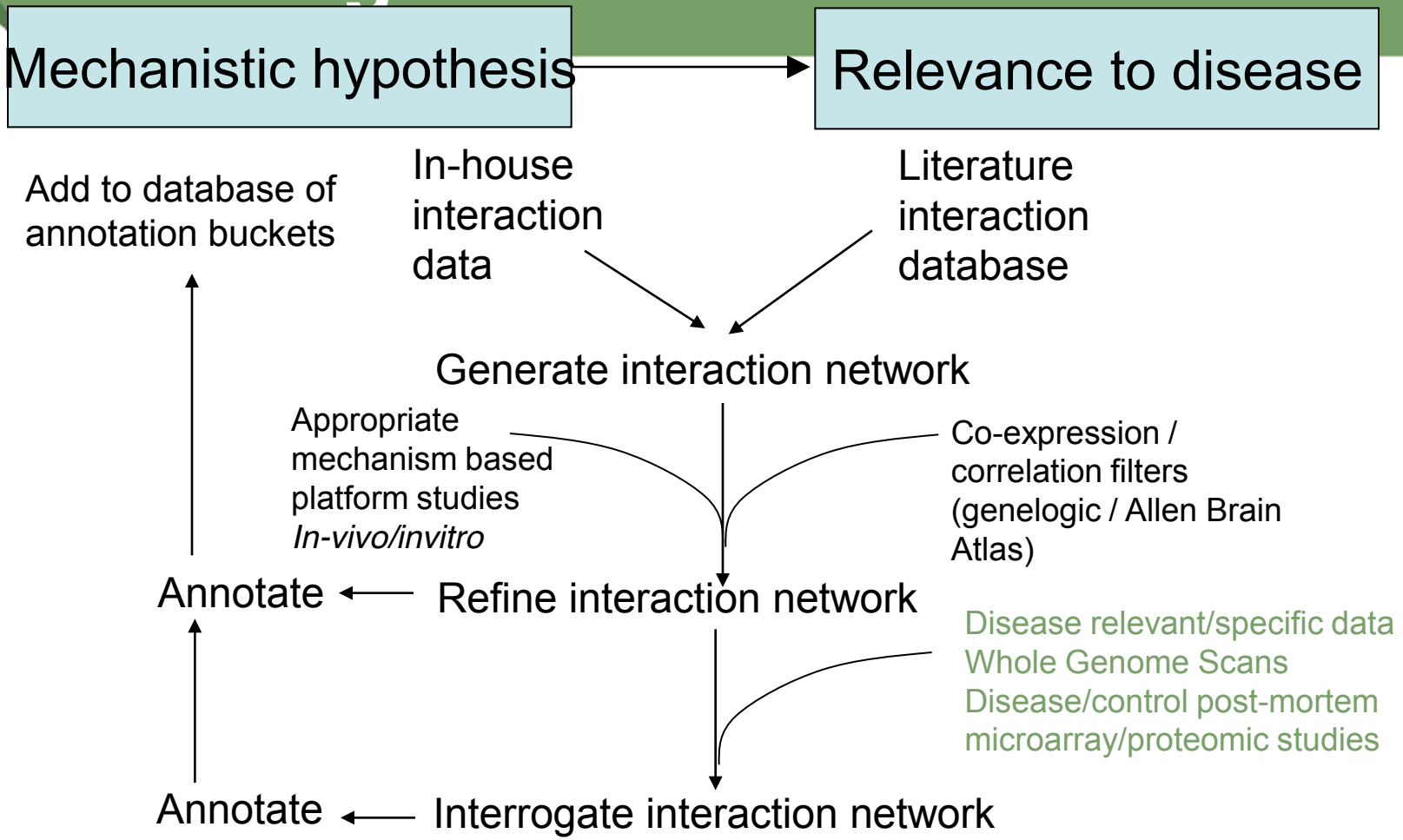
- **Are we only now realising the benefits of this?**
  - Broad Institute
  - Allen Brain Atlas
  - Genetic studies
  - Genomic studies
  - (Jackson Lab KO ontologies)
  - Neuroimaging and other translational medicine approaches
  - Better integration between these data / sites

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):
    - Query: "central nervous system[mh]" OR "central nervous system diseases"[mh]
    - 1997-2006: 1271
    - 1987-1996: 902
    - 1977-1986: 96
  - Gene Ontology annotation reflects this
  - We can improve on this by "hijacking" knowledge of molecular mechanisms and pathways elucidated from other tissues/diseases
  - BUT…

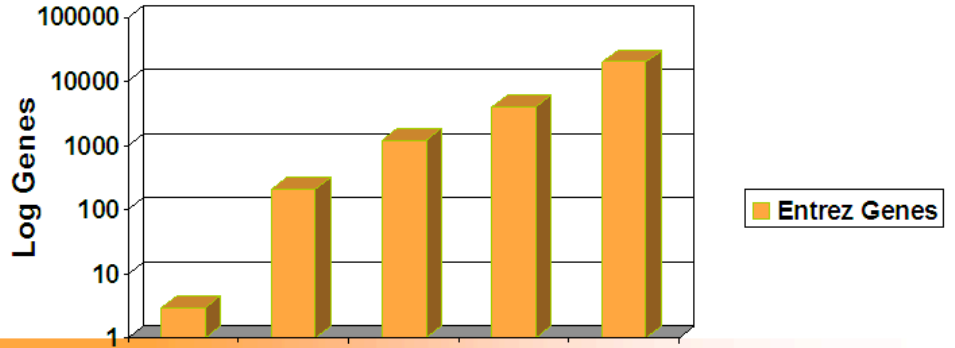
Literature

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

- Query: "central nervous systen
- 1997-2006: 12
- 1987-1996: 90
- 1977-1986: 96
- Gene Ontology
- We can improv knowledge of n pathways eluci tissues/disease
- BUT…

How much do we know from literature? (*as of Nov. 1st*)

- Take all Human Entrez genes (~25K) along with their Mouse and 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



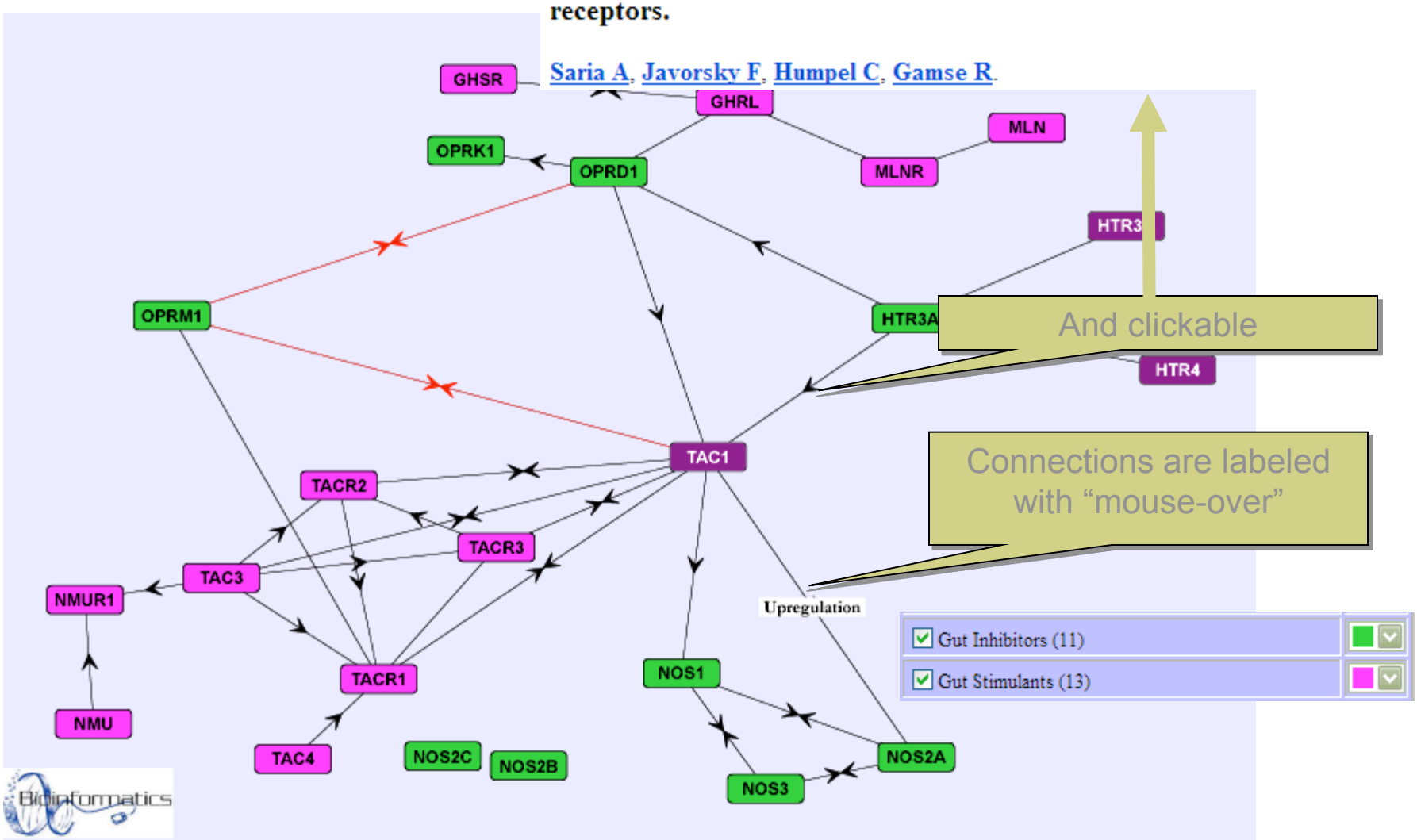
Building network from literature

Connections between groups

Ann N Y Acad Sci. 1991;632:464-5. Related Articles,

Endogenous 5-hydroxytryptamine modulates the release of tachykinins and calcitonin gene-related peptide from the rat spinal cord via 5-HT3 receptors.

Saria A, Javorsky F, Humpel C, Gamse R.



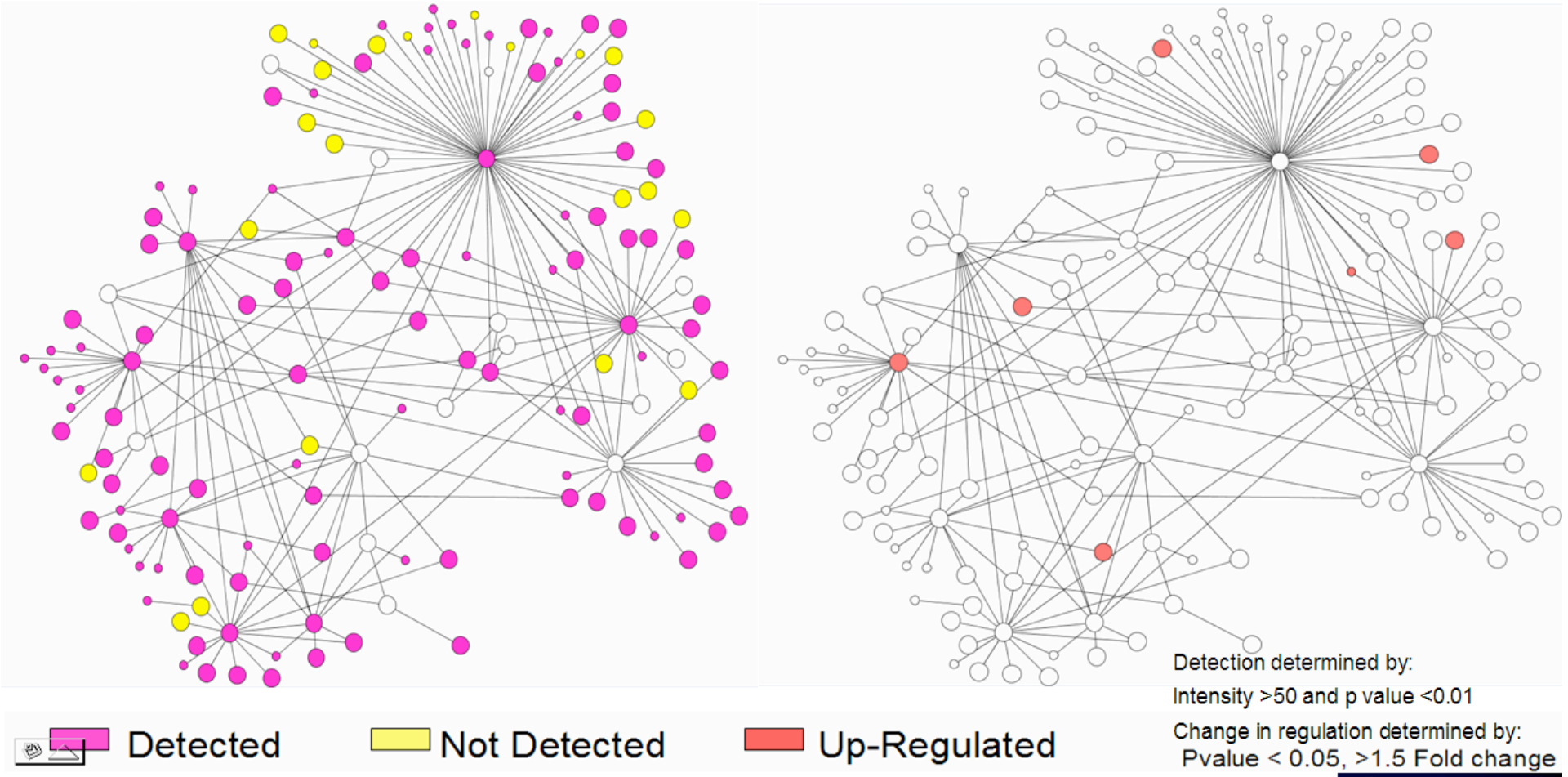
Identification of Pathways from Network

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

Monocyte derived Macrophages

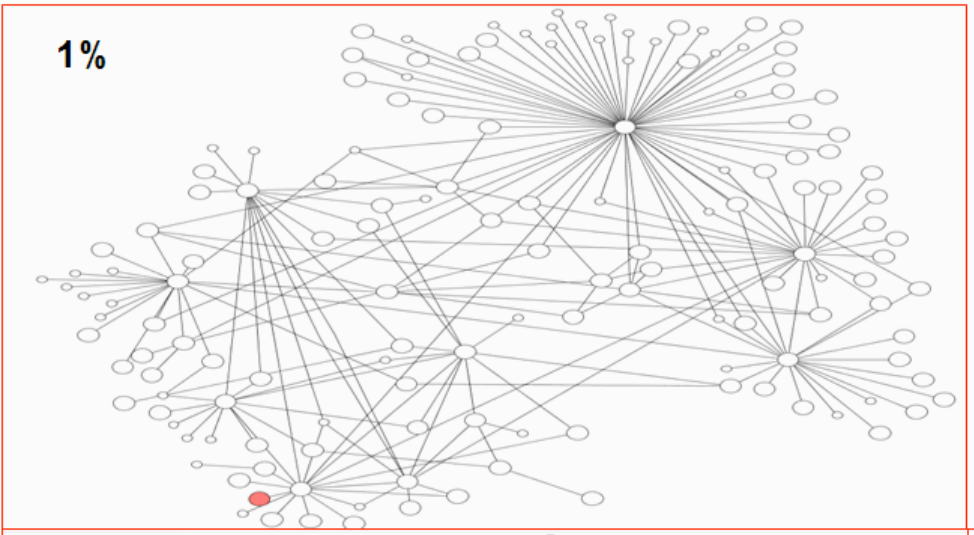
Resting:

LPS:



Bioinformatics D&A

Dose Response Effect on Network

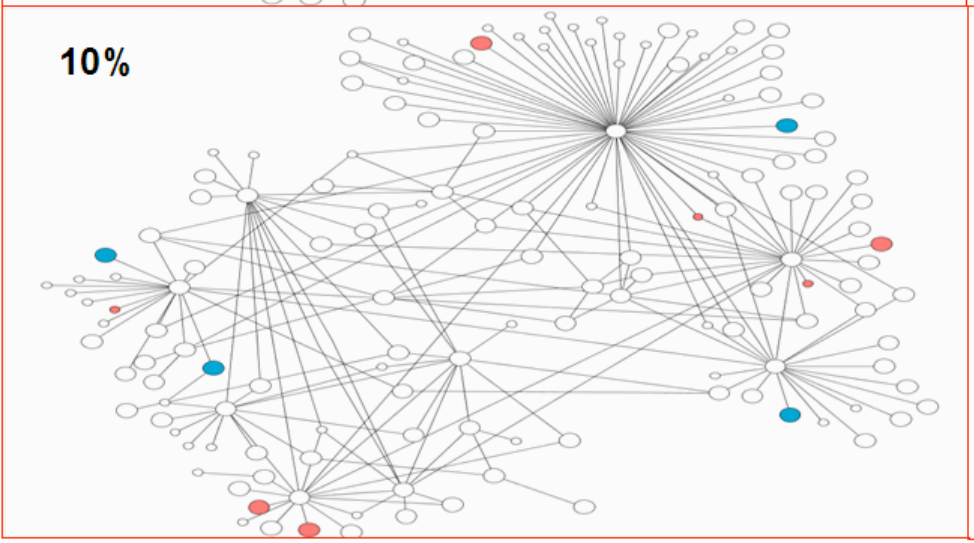


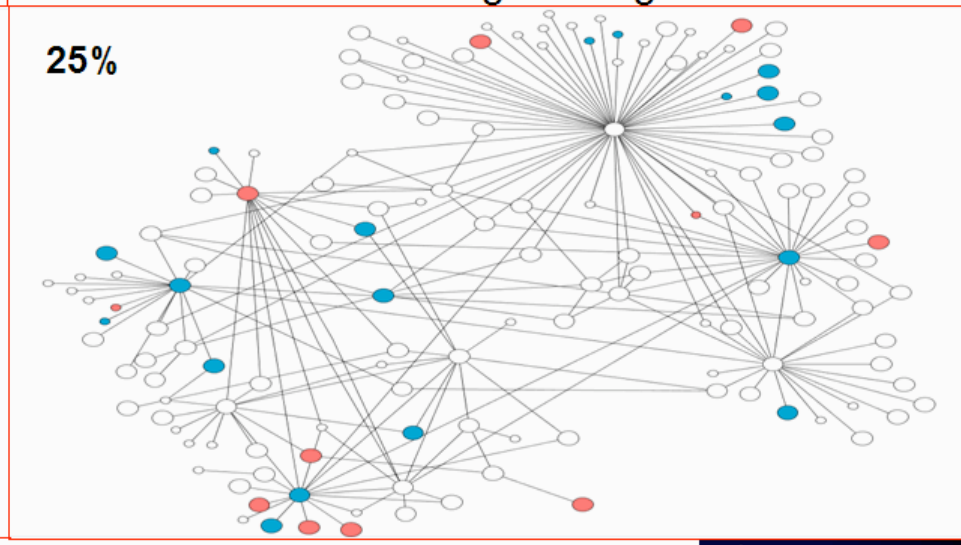
Monocyte derived Macrophages experiment

- 3 different % of cigarette stimulation
- Changes tracked from resting
  - Pvalue < 0.05, >1.5 Fold change

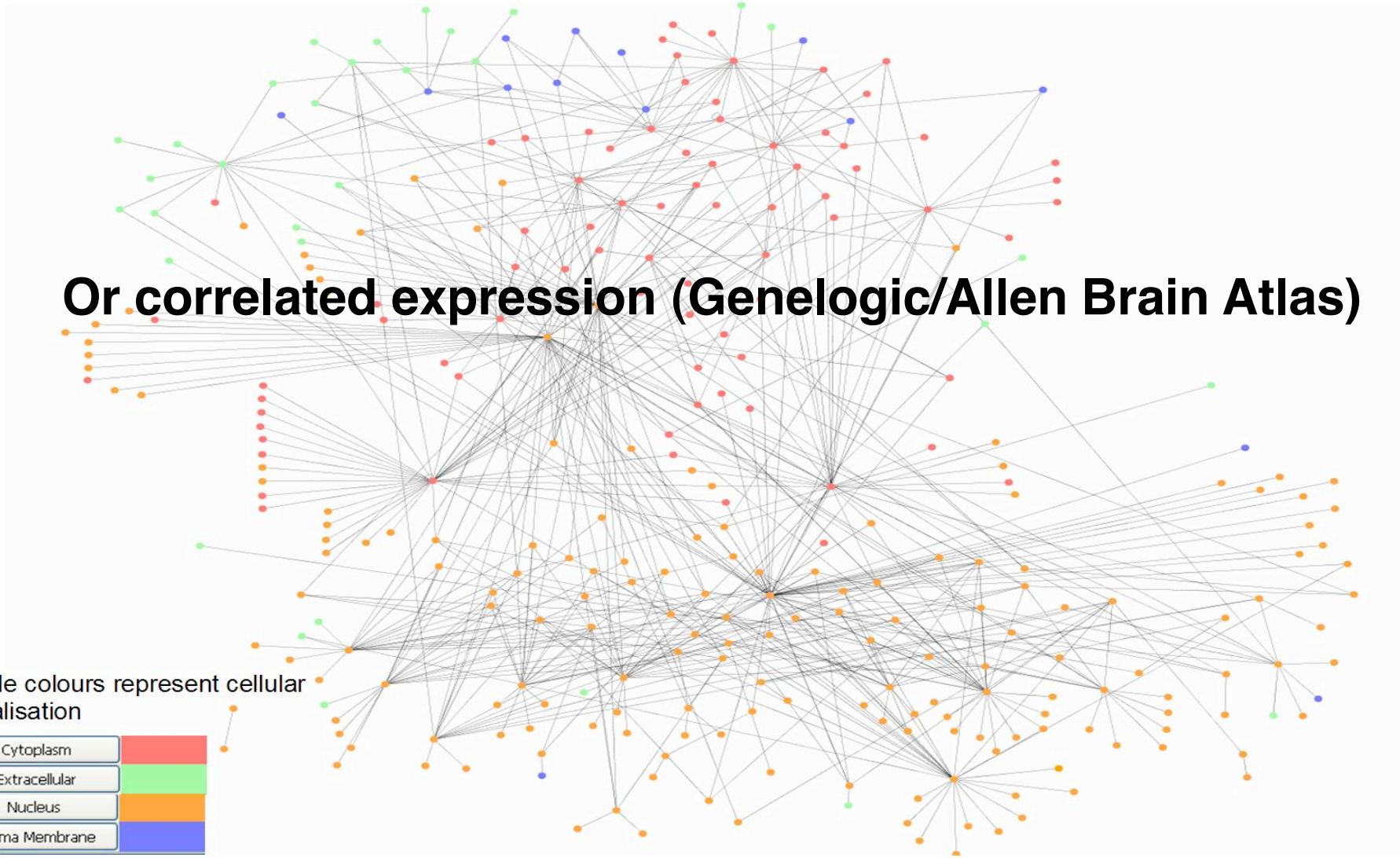
Up regulated genes

Down regulated genes





Compartmentalisation can help refine expanded Network



Whole Genome Scans

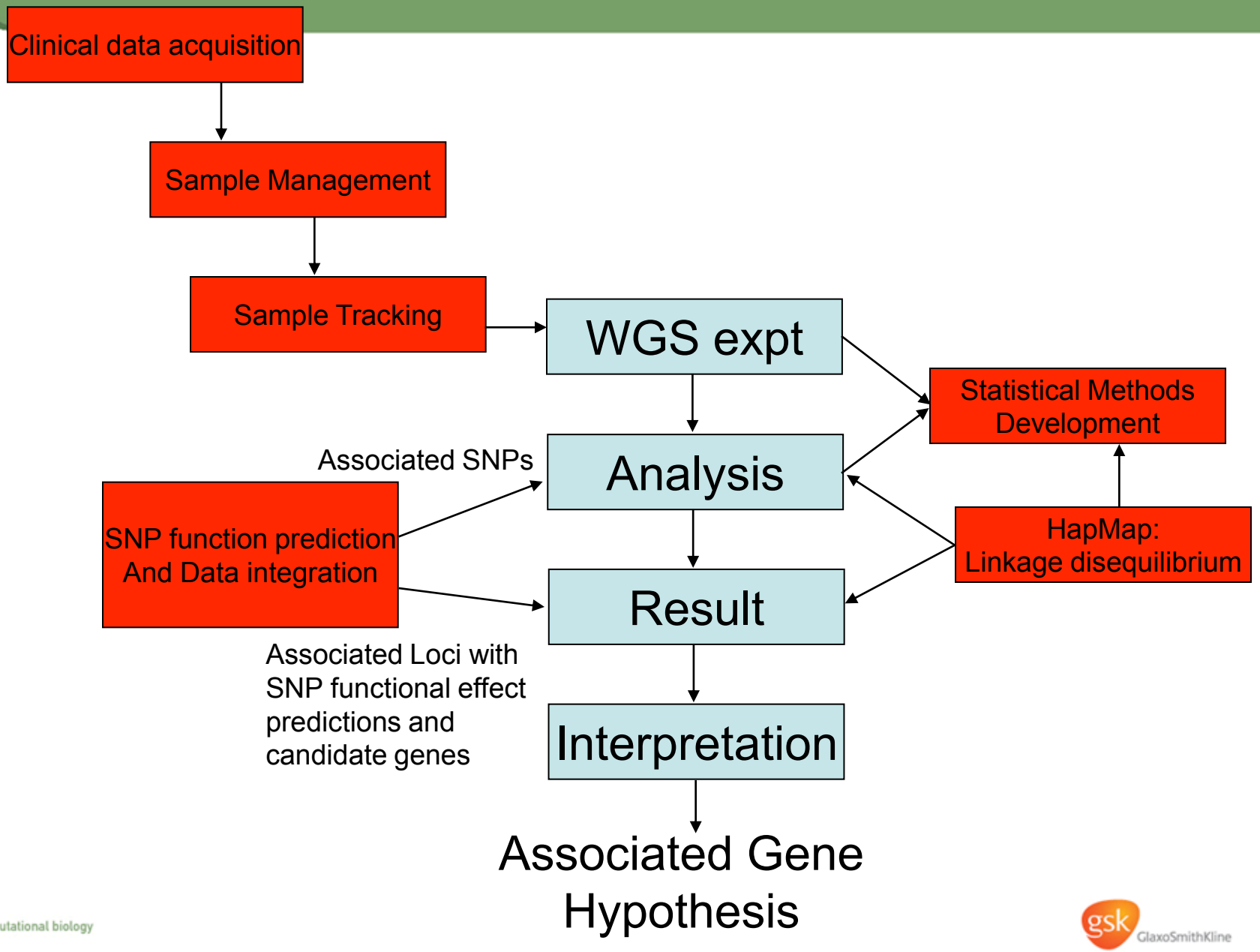
centre	diseases	PI	platform	study type	case no	date anounced	genotyping complete (est)	estimated availability
BROAD	cardiovascular		AFFY	case/control	2000	Jul-05	Jul-06	Jan-07
WTCCC	Coronary heart disease	Alistair Hall & Nilesh S	Affy & Perl	case/control	2000	Jul-05	Jul-06	Jan-07
WTCCC	Hypertension	Mark Caulfield & Marti	Affy & Perl	case/control	2000	Jul-05	Jul-06	Jan-07
Children's Hospital	statin-related myopathy		Illumina Hu	case/control	1000-100	Oct-05	Oct-07	Apr-07
Framingham Genet	heart		AFFY	longitudinal	9000	Sep-06	Sep-07	Mar-08
Framingham Genet	blood disorders		AFFY	longitudinal	9000	Sep-06	Sep-07	Mar-08
Erasmus	cardiovascular		Illumina Hu	longitudinal	10000	Nov-06	Nov-07	May-08
WTCCC	Malaria	Dominic Kwiatkowsk	Affy & Perl	case/control	2000	Jul-05	Jul-06	Jan-07
WTCCC	Tuberculosis	Adrian Hill, Melanie Ne	Affy & Perl	case/control	2000	Jul-05	Jul-06	Jan-07
BROAD	metabolic		AFFY	case/control	2000	Jul-05	Jul-06	Jan-07
WTCCC	Type 1 diabetes	John Todd & David Da	Affy & Perl	case/control	2000	Jul-05	Jul-06	Jan-07
WTCCC	Type 2 diabetes	Mark McCarthy & Andr	Affy & Perl	case/control	2000	Jul-05	Jul-06	Jan-07
GAIN	Diabetic Nephropathy	Warram_JH	AFFY	case/control	1000	Oct-06	Jun-07	Dec-07
Erasmus	endocrine		Illumina Hu	longitudinal	10000	Nov-06	Nov-07	May-08
WTCCC	Ankylosing spondylitis	Matthew Brown	Affy & Perl	case/control	2000	Jul-05	Jul-06	Jan-07
WTCCC	Autoimmune thyroid diseas	Stephen Gough	Affy & Perl	case/control	2000	Jul-05	Jul-06	Jan-07
WTCCC	Inflammatory bowel disease	Miles Parkes & Chris M	Affy & Perl	case/control	2000	Jul-05	Jul-06	Jan-07
WTCCC	Multiple sclerosis	Alastair Compston	Affy & Perl	case/control	2000	Jul-05	Jul-06	Jan-07
NIDDK	Crohn's		Illumina Hu	case/control	2000	Jan-06	Jan-07	Jul-07
Massachusetts Ger	ALS		Illumina Hu	case/control	2200	Aug-06	Aug-07	Feb-08
Erasmus	neurological		Illumina Hu	longitudinal	10000	Nov-06	Nov-07	May-08
Erasmus	locomotor		Illumina Hu	longitudinal	10000	Nov-06	Nov-07	May-08
Erasmus	opthalmological		Illumina Hu	longitudinal	10000	Nov-06	Nov-07	May-08
BROAD	cancer		AFFY	case/control	2000	Jul-05	Jul-06	Jan-07
WTCCC	Breast cancer	Michael Stratton & Na	Affy & Perl	case/control	2000	Jul-05	Jul-06	Jan-07
BROAD	neuropsychiatric diseases,		AFFY	case/control	2000	Jul-05	Jul-06	Jan-07
WTCCC	Bipolar	Nick Craddock	Affy & Perl	case/control	2000	Jul-05	Jul-06	Jan-07
Max Planck	Depression and anxiety		Illumina Hu	case/control	?	Sep-05	Sep-06	Mar-07
GAIN	ADHD	Faraone_S	AFFY	case/control	1000	Oct-06	Jun-07	Dec-07
GAIN	Schizophrenia	Gejman_P	AFFY	case/control	1000	Oct-06	Jun-07	Dec-07
GAIN	Bipolar	Kelsoe_JR	AFFY	case/control	1000	Oct-06	Jun-07	Dec-07
GAIN	Major depression	Sullivan_PF	AFFY	case/control	1000	Oct-06	Jun-07	Dec-07
Framingham Genet	sleep disorders		AFFY	longitudinal	9000	Sep-06	Sep-07	Mar-08
WTCCC	Rheumatoid arthritis	Jane Worthington	Affy & Perl	case/control	2000	Jul-05	Jul-06	Jan-07
GAIN	Psoriasis	Abecasis_G	AFFY	case/control	1000	Oct-06	Jun-07	Dec-07
Framingham Genet	lung disorders		AFFY	longitudinal	9000	Sep-06	Sep-07	Mar-08

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

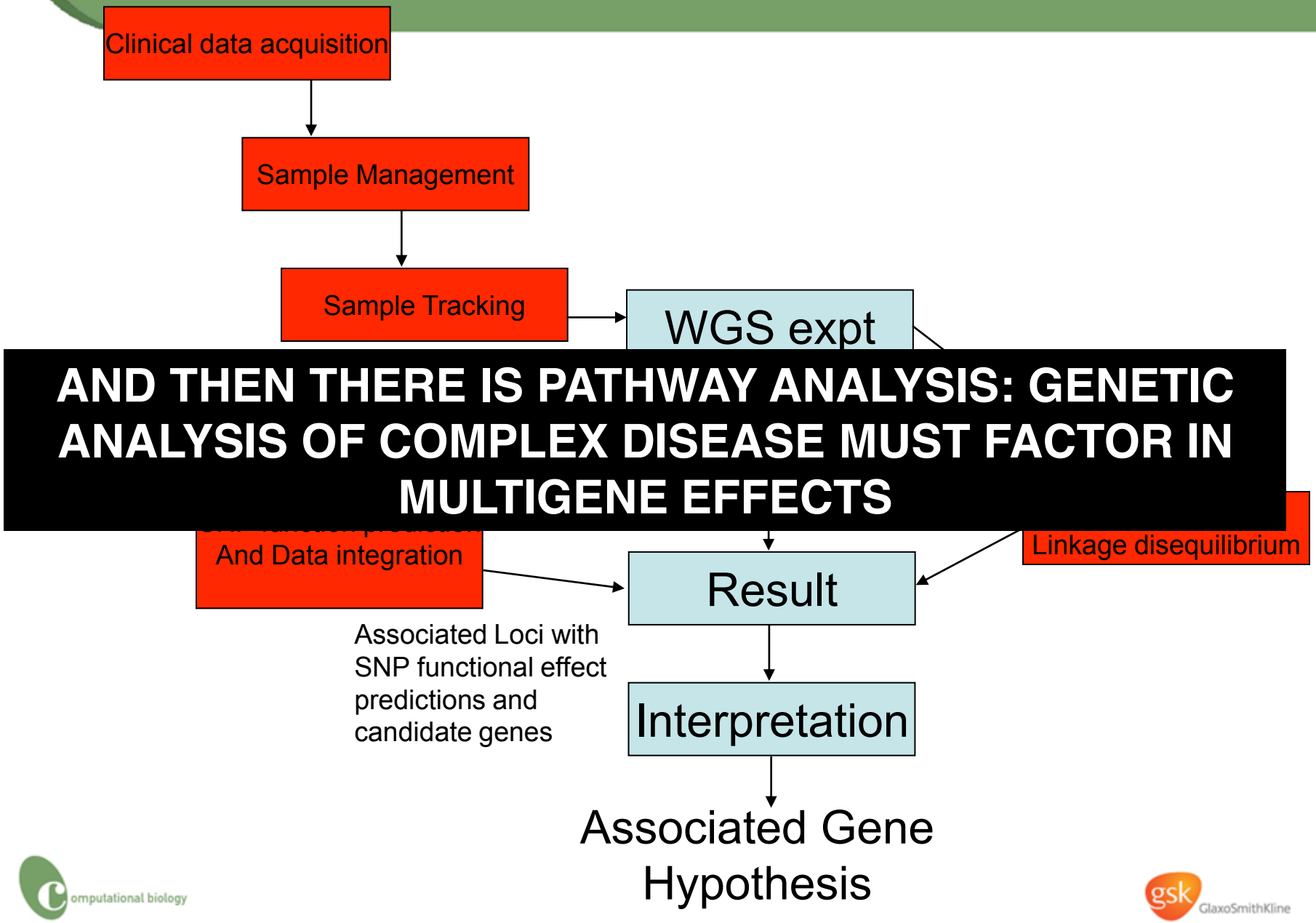
computational biology

GlaxoSmithKline

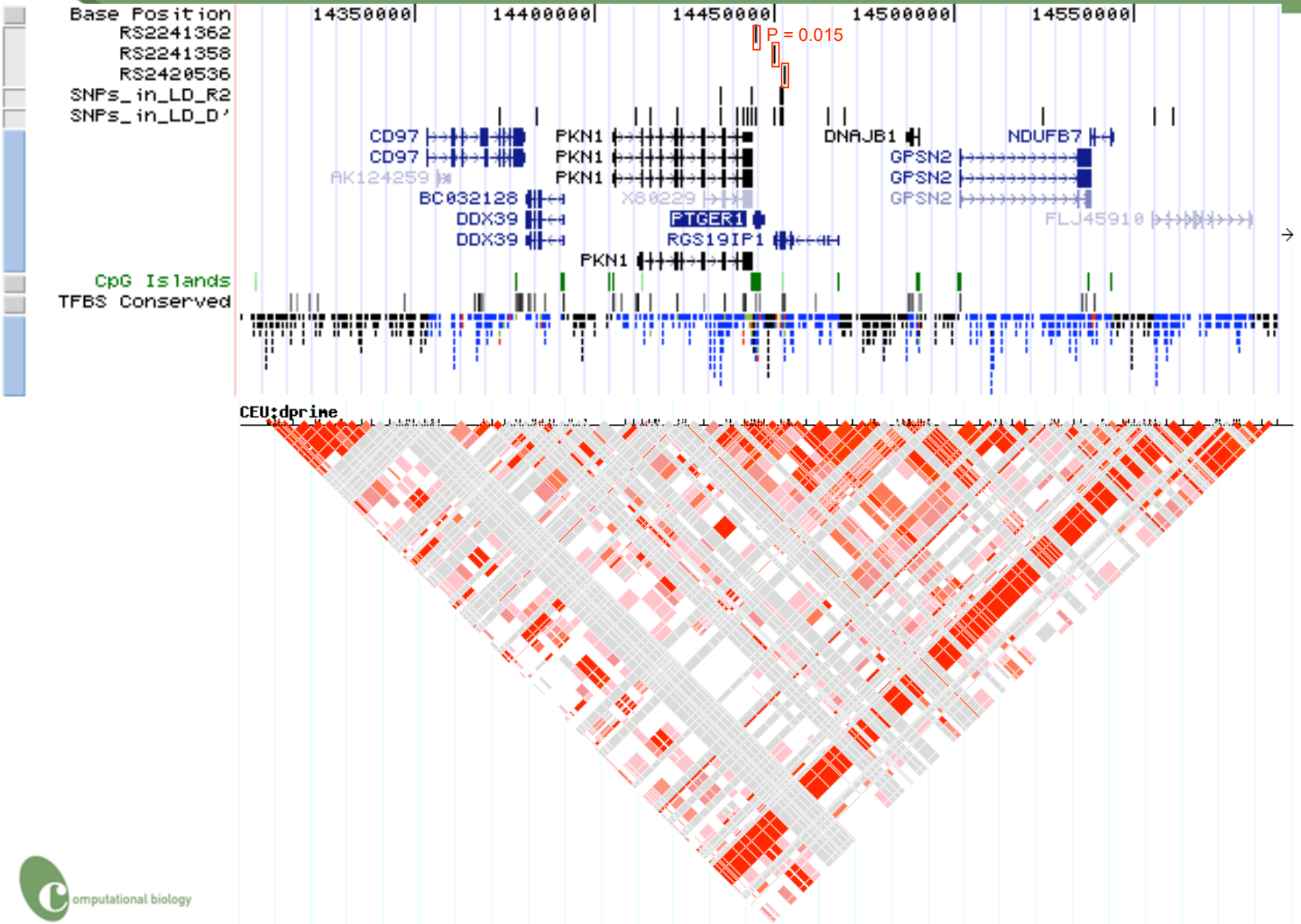
Genetic Analysis



Genetic Analysis

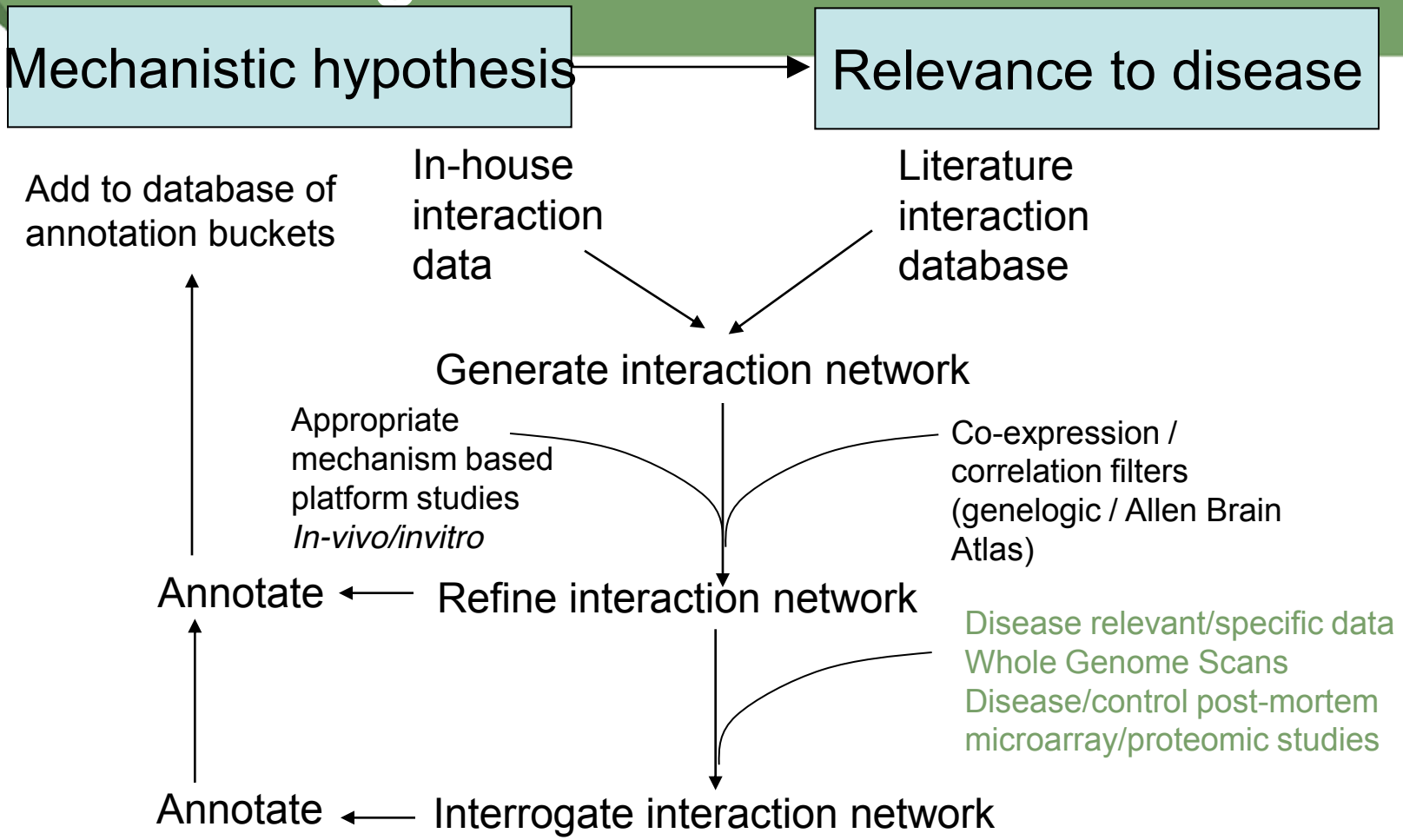


A word on linkage disequilibrium



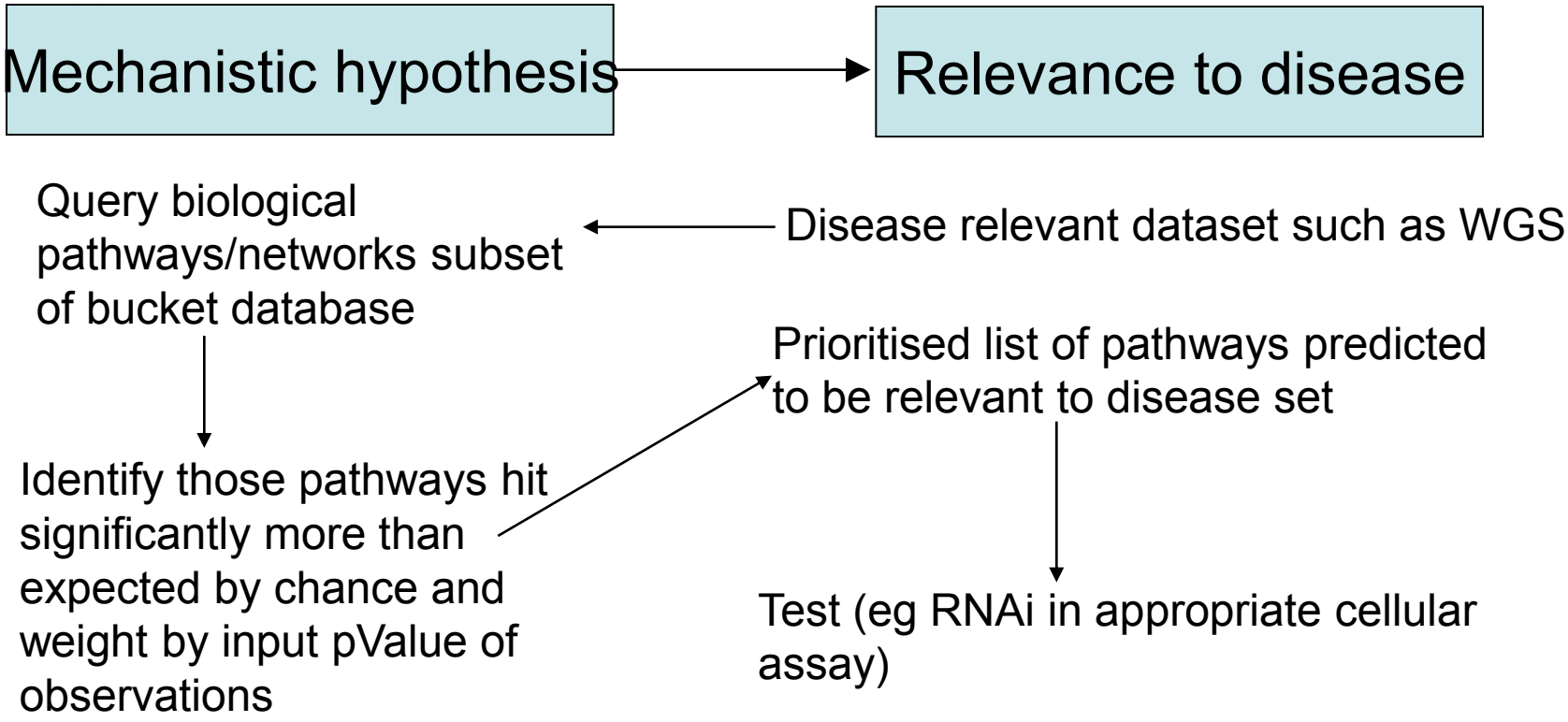
computational biology

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

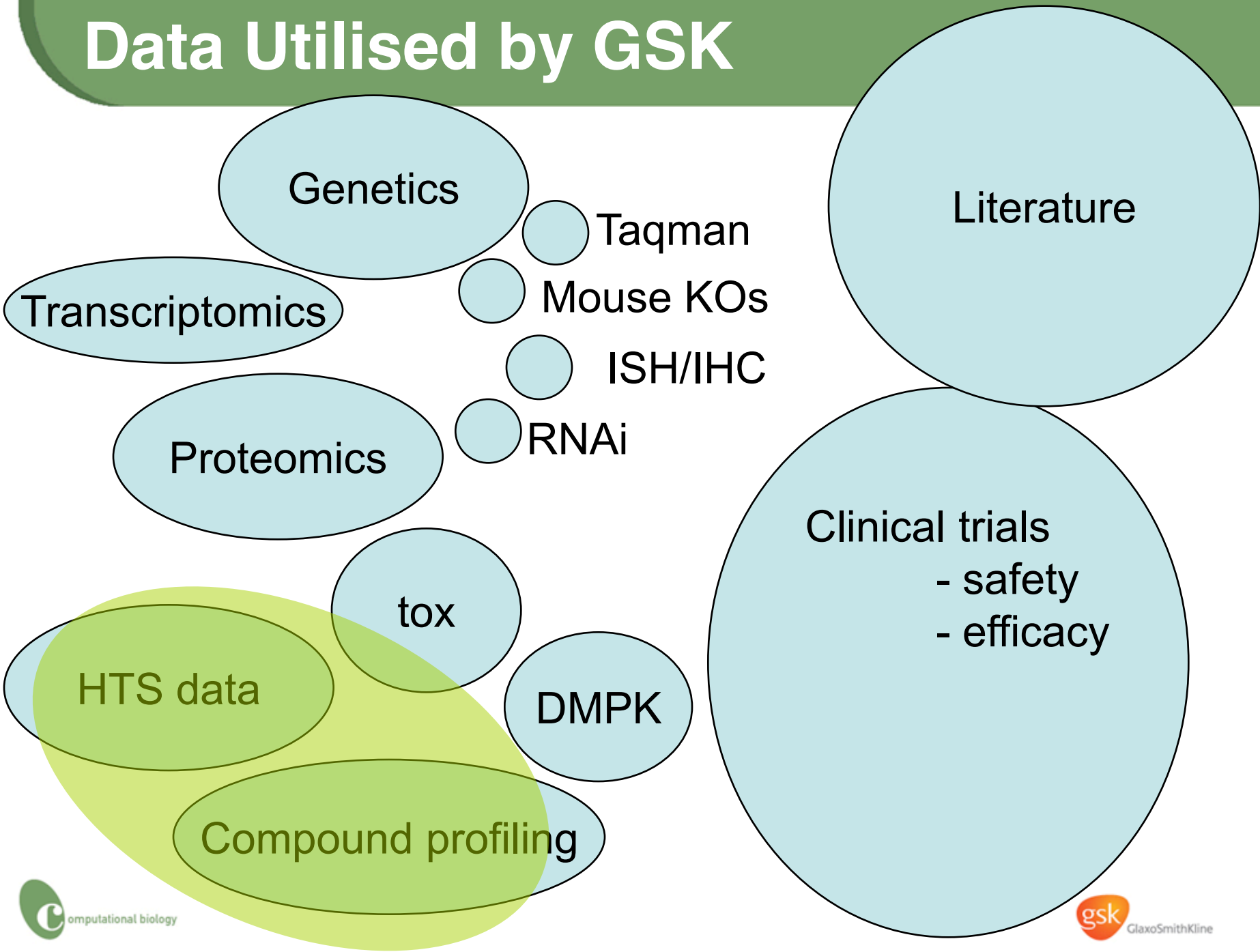


BUT: Which gene(s)?

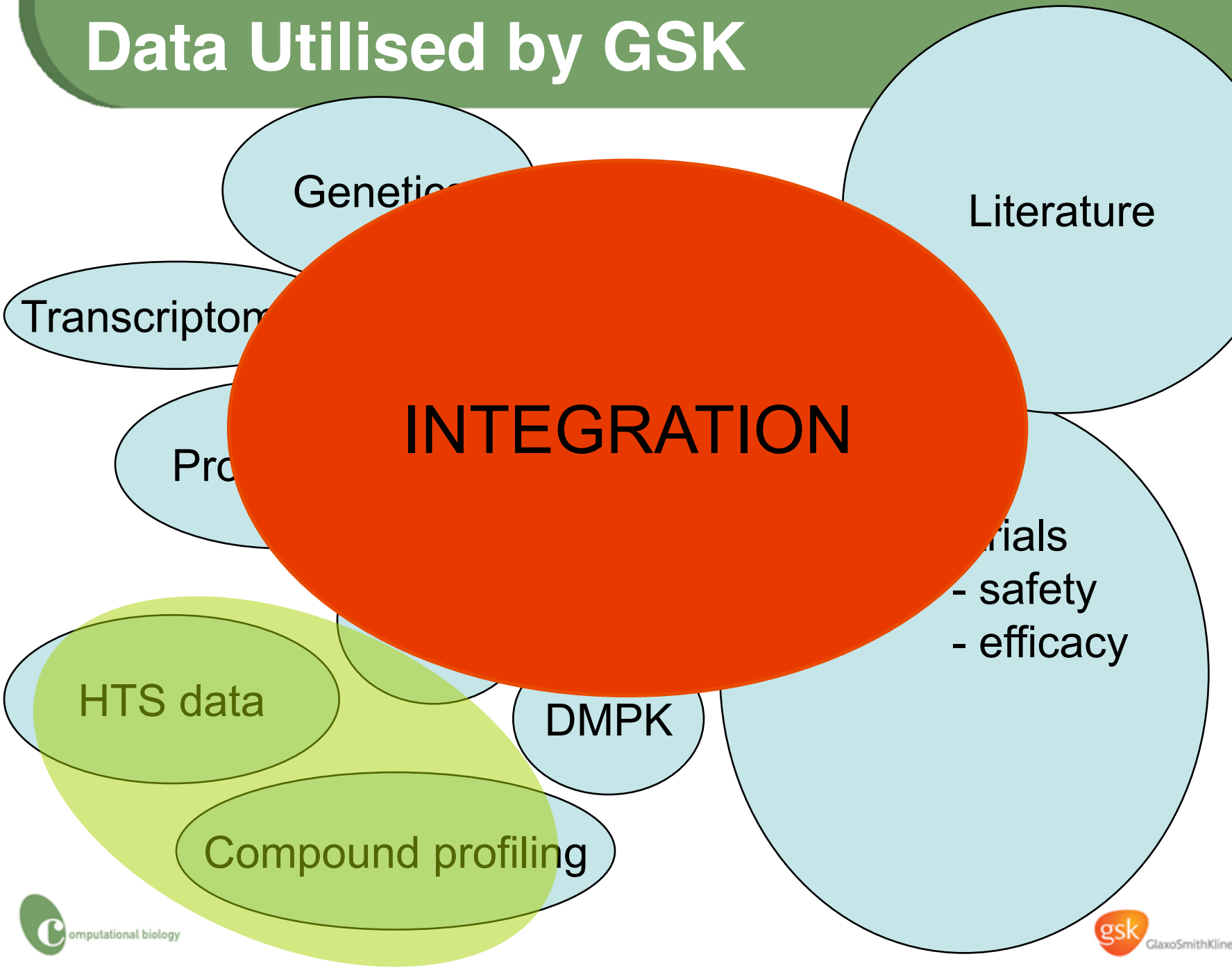
A pathway may be significant but individual genes may have small component effects, may need to knockout/down >1

Many significant pathways/networks identified, do we test in each?

Data Utilised by GSK



Data Utilised by GSK



Compound view

MOLECULAR DISEASE UNDERSTANDING

Genes

aliases

Potency

Specificity

Compounds

aliases

CLINICAL TRIALS RESULTS

Compound view

MOLECULAR DISEASE UNDERSTANDING

CLINICAL TRIALS RESULTS

Potency

B4 Literature Query 24thFeb06

Row No	Name	Compounds (Top 15 CCP > 5.5)	Compounds (Top 15 KATE > 5.5)
1 Printable Row	Yes	Inhibitors (Total=3):	Unknowns (Total=15)

Depression [MeSH]

Compound	Potency	n	Specificity	# genes	# I	# E	# X
GSK	<5.82	1	-1.89	24	23	1	0
AH2I	=5.63	2	-1.84	26	23	3	0
GR4	=5.5	2	-3.87	21	21	0	0

Compound	Potency	Specificity	# genes
SB	8.6	+?	1
GW	8.5	+?	1
SB	8.5	+0.60	3
SK	8.1	+?	1
12	7.5	+?	1
BR	6.5	+?	1
BR	6.3	+?	1
	6.3	-0.10	4

Portfolio

- *Status:* Target to Lead
- *CEDD:* PSYC
- *Primary Disease:* Unipolar depression
- *HITDIP:* Schizo 1st Round Lvl6; Unipolar 1st Round

GeneLogic Normal Brain and Peripheral Distribution | GeneLogic 8thMar06 | HiTDIP Accredited Data 12thMar06 | HiTDIP Primary Screen Data is Superceded by Accredited Data 12thMar06 | HiTDIP Primary and Pilot Schizophrenia Screens

Yes Human 100% Brain, 77% Non-brain

- Yes Human thalamus N (U133): Median Intensity = 1776, 217 (Potential

Yes Human Unipolar Depression Combined HITDIP Level: Accredited locus associated with SNP.

- Preliminary data: strong risk of false positives
- Maybe Human Unipolar Depression Primary HITDIP Level: 5

- Preliminary data: strong risk of false positives
- Maybe Human Schizophrenia HITDIP Plot Screen: 6, p value <=0.05

Lifespan (Expression Level: 0 - negative, 1 ... 13thMar06

s Phenotype

- Yes Human Brain, Amygdala (Adjacent t
- Yes Human Brain, Amygdala (Neurons):
- Yes Human Brain, Hypothalamus (Neuro
- Yes Human Brain, Hypothalamus (Neuro
- Yes Human Brain, Pituitary (Anterior pitu
- Yes Human Brain, Pituitary (Para interme
- Yes Human Brain, Thalamus (Astrocytes
- Yes Human Brain, Thalamus (Neurons):
- Yes Human Brain, Substantia Nigra (Ependyma): 3
- Yes Human Brain, Substantia Nigra (Neurons): 3
- Yes Human Brain, Substantia Nigra (Nonpigmented neurons): 3
- Yes Human Brain, Substantia Nigra (Pigmented neurons): 3
- Yes Human Brain, Medulla (Ependyma): 3
- Yes Human Brain, Medulla (Neurons): 3
- Yes Human Brain, Medulla (Neurons in the region of the cuneate nucleus): 3

Methods Find Exp Clin Pharmacol 2005, 27(5): 289
ISSN 0379-0355

CCC: 0379-0355
DOI: 10.1358/mf.2005.27.5.908643

Research on adverse drug events. I. Muscarinic M3 receptor binding affinity could predict the risk of antipsychotics to induce type 2 diabetes

Silvestre, J.S., Prous, J.

computational biology

GlaxoSmithKline

Informatics and Industry

- ***Application*** of computational methods towards drug discovery
- Looks to public science for methods development
  - Next Gen. Seq.
- More focus in public science on disease and drug discovery
  - ChEMBL
  - PubChem
    - Cheminformatics emerging in public science
- Move towards greater communication and collaboration between public science and industry sectors
  - *Better science applied to drug discovery*