# Data and Analysis in the Phenotypic World

OPhacts – Santiago de Compostela
16th-17th Feb 2015

# Ceara Rea

## Introduction

- Computational Chemist
  - Have been working @GSK for nearly 10 years
    - Based at the Stevenage site

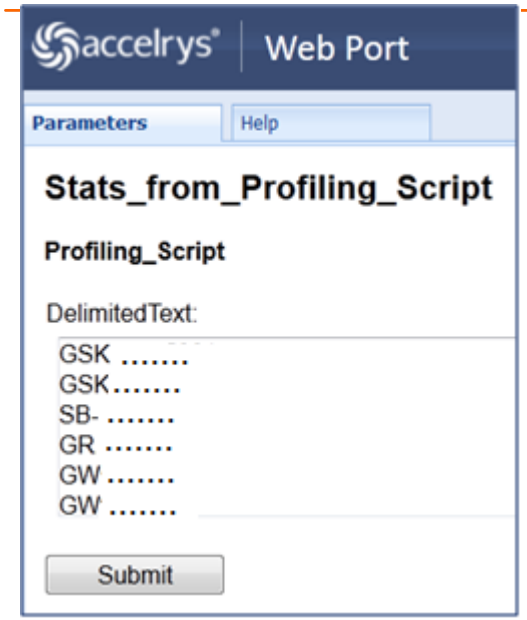
- I’ve worked on several phenotypic projects (data compilation & analysis)
- Skills:
  - Data storage (databases)
  - Programming (Python) & query language (SQL)

# Phenotypic Analysis

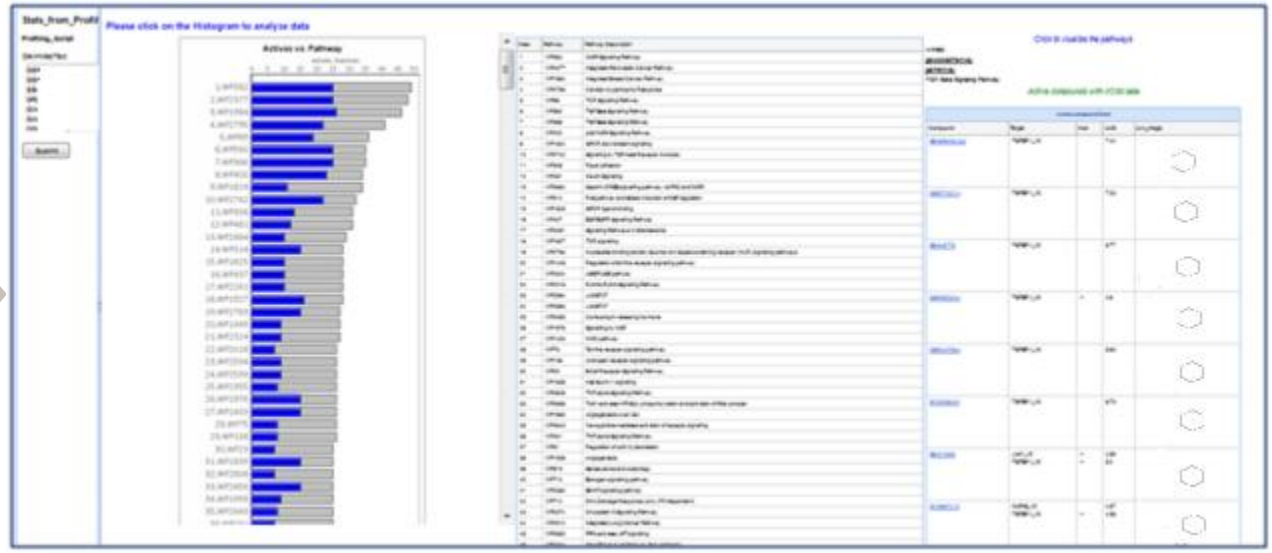
## Overview

- Getting the data - Data Integration
  - Workflows for :
    - Compound target profiling (target deconvolution)
    - Compound selecting (focussed screen, hypothesis validation)
    - Phenotype profiling
  - What issues are there?
    - Missing or unrecognisable data
    - Level of reporting (compound: salt/parent, target: transcript/gene)
    - Too much/irrelevant data
    - Standardisations/normalisations (identifiers, synonyms, controlled vocabularies)
    - Speed/technical difficulties
- Analysing/Visualising the data
  - What gives you **confidence** that a target/pathway is involved in a phenotypic mechanism?
    - A summary of activities can give first indication, but need to dig into the data
      - Expose info on specific assays with cells/technologies etc
        - » Assay interference/nuisance behaviour/endogenous activities
      - Think about the compounds involved
        - » Have we got different chemotypes for a target?, what about compounds near neighbours?
      - Investigate selectivity profiles for compounds
      - Can we correlate target activity with phenotypic response?
      - Can we show activity at multiple targets on a pathway?
      - If full panel of data available, what stats/probability scores can be generated

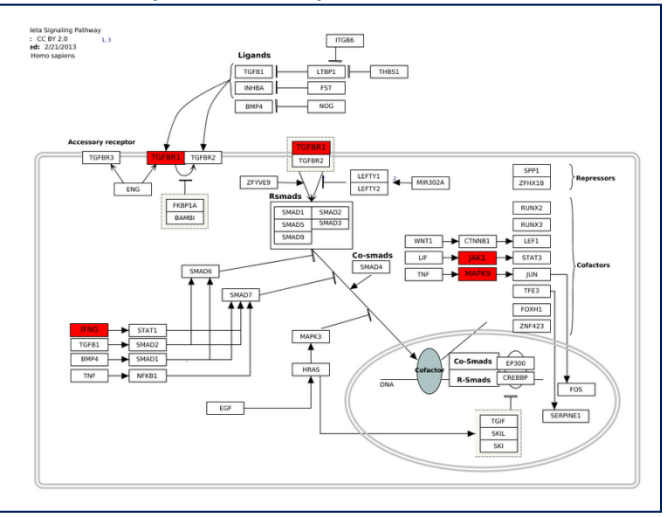
# Pipeline Pilot Protocol for Analysis of Phenotypic Screening Data




Pipeline Pilot protocol



Analysis from the Phenotypic screening




Pathway


Compound specific information

# Compound Profiling for Target Deconvolution

# Compound Profiling Hits from a Phenotypic Assay

## Workflow Overview

#### Currently not using OPhacts for Compound Information ####

- Take the hits from the phenotypic assay
- Compound Profiling:
  1. Expand an 'input' compound id to include all synonyms/salt ids & use a unique parent id:
     - **NICOTINE|**CHEMBL3|CHEMBL225057|CHEMBL1628647|CHEMBL3137669|CHEMBL151515|CHEMBL1201536|CHEMBL1448280|AUREUS10053|AUREUS230964|AUREUS566530|AUREUS566532|GR117011X|GR117011B|GR117011C|SKF-7925-A|BRL-9889NS
  2. Retrieve and collate all assay data and associated information for those ids
     - Datasources:
       - Activity data: Internal GSK, ChEMBL, Aureus
       - Compound associated data: Liabilities, projects, properties
       - Assay meta data: Internal GSK (technology/cell line)
       - Target & pathway data: Internal GSK, OPhacts, Wikipathways, GeneGo
  3. Summarise, pivot & bin activities
     - Aggregate at a target level
       - Best activity reported
     - Bins for activity @ '='. > 5.5, > 7
  4. Output summary files
     - Compound – top 3 targets, no. of targets/assays tested/no. actives
     - Targets – no. of compounds tested/no. of actives (Bins for activity @ '='. > 5.5, > 7)

# External Data Integration

- Objective:
  - Integrate external bioactivity data with GSK internal to provide an enriched profile of target activity for each compound
  - Datasources:
    - ChEMBL
    - Aureus

- How has this been achieved? ..... the good, the bad & the ugly.....

## The Good

- Direct compound lookups
  - InChi keys have been added to the compound information
    - Aureus contains InChi keys
    - You can do compound lookups using inChi keys in the ChEMBL web services
- Retrieving a compound name and external ids is a **big** plus
  - You can look up a name on wikipedia, and get all sorts of info about MoA
  - External ids can be used to browse the raw data

# External Identifiers

NICOTINE|GR 117011X|GR 117011X|GR 117011X|GR117011X|GR117011|GR 117011A|GR 117011A|GR 117011A|GR117011A|GR 117011B|GR 117011B|GR 117011B|GR117011B|GR 117011C|GR 117011C|GR 117011C|GR117011C|GR 117011D|GR 117011D|GR 117011D|GR117011D|GR 117011F|GR 117011F|GR 117011F|GR117011F|GR 117011G|GR 117011G|GR 117011G|GR117011G|GR 117011L|GR 117011L|GR 117011L|GR117011L|GR 117011H|GR 117011H|GR 117011H|GR117011H|GR 117011J|GR 117011J|GR 117011J|GR117011J|GR 117011K|GR 117011K|GR 117011K|GR117011K|SKF7925J|SKF-7925-J|SKF-7925|SKF7925A|SKF-7925-A|SKF7925F2|SKF-7925-F2|SKFS7925A2|SKF-S-7925-A2|SKF-S-7925|BRL9889NS|BRL-9889NS|BRL-9889|CHEMBL3|Nicoderm|Nicotrol Inhaler|Prostep|SID26752745|Nicotine|SID26752744|SID17389805|Nicotrol NS|Habitrol|Nicoderm CQ|Nicotrol|CHEMBL225057|CHEMBL1628647|CHEMBL3137669|CHEMBL151515|CHEMBL1201536|CHEMBL1448280|AUREUS10053|AUREUS236964|AUREUS566531|AUREUS36204|AUREUS566530|AUREUS566532|AUREUS270129|AUREUS777864|Nicotine bitartrate|Nicotine dihydrochloride

External IDs are provided

CompoundProfile.csv

| | A | B | C | D |
|---|---|---|---|---|
| 1 | **COMPOUND** | **SynonymsMerged** | **PrefPcn** | **Psmiles** |
| 2 | CCI133 | ASPIRIN\|CCI 16716\|CCI 16716\|CCI 16716\|CCI16716\|BRL9066\|BRL-9066\|AH 5311\|AH 5311\|A | CCI133 | CC(=O)Oc1 |
| 3 | GR117011A | NICOTINE\|GR 117011X\|GR 117011X\|GR 117011X\|GR117011X\|GR117011\|GR 117011A\|GR 11 | GR117011 | CN1CCC[C |
| 4 | AH14925AA | CAFFEINE\|SKF6053\|SKF-6053\|CCI 3994\|CCI 3994\|CCI 3994\|CCI3994\|BRL7867AA\|BRL-7867A | AH14925 | Cn1cnc2n( |

Preferred name from ChEMBL is given

# External Identifiers – Targets

## The Bad

- ChEMBL and Aureus use Uniprot Ids as a standard identifier
- In house (GSK) we use Tar Ids

### Link?

- Internal database contains ncbi (entrez) gene ids and refseq ids but not Uniprot ids
  - **Uniprot Id** → **RefSeq_p** → **Tar Id**
  - Initially implemented with OPhacts, but now using the Uniprot Id mapping service
    - Restful service, minimal data, batch input

# External Identifiers - Targets

## The Ugly

- Not all targets can be mapped back
- Aureus is very poorly annotated with Uniprot Ids

| TopTarget | TopTarName | TopTarClass | Top3Targets |
|---|---|---|---|
| AURTAR77590_ | PEPTIDYL-PROLYL CIS-TRANS ISOMERASE FKBP12 | | PEPTIDYL-PROLYL CIS-TRANS ISOMERASE FKBP12:9.48\|EIF4E:9.3\|F |
| tar72192 | EHMT2 | TRANSFERASE | EHMT2:4.85\|Cruzaine:4.7\|POLK:4.45 |
| tar5384 | IGF1R | TRANSFERASE_KINASE | IGF1R:8.66\|INSR:8.64\|ERBB4:7.99 |
| AURTAR46_9606 | ALPHA2 | | ALPHA2:8.72\|ALPHA-ADR_RAT:8.26\|ADRA2A:8.25 |
| CHEMBL5221 | IR1_RAT | | IR1_RAT:7.95\|IMIDAZOLINE 1 RECEPTOR_RAT:7.95\|ADRA2C:7.9 |

Can't always get a symbol

Vague target definitions

Can't get target class

Will get the same data from the different sources

- Current implementation is only using data with a UniProt Id

# External Identifiers Targets

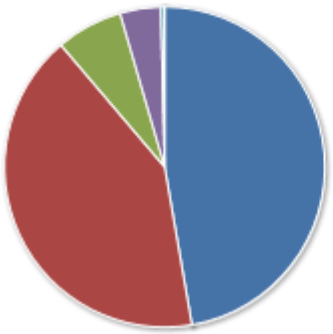
- CHEMBL612545

**Target Associated Bioactivities**

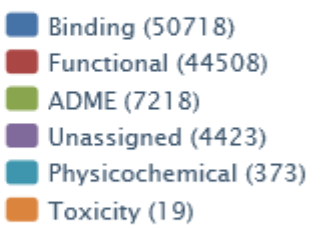
ChEMBL Act **Target Name and Classification**

Total: 1660480

| | |
|---|---|
| Target ID | CHEMBL612545 |
| Target Type | UNCHECKED |
| Preferred Name | Unchecked |
| Synonyms | |
| Organism | |
| Species Group | No |
| Protein Target Classification | Not applicable |



Total: 107259



# Associated Compound Data

## Liabilities (Internal data only)

- When being profiled a compound gets flagged as having a possible liability if:

  - It has a high hit rate
  - It has a common nuisance substructure
  - Compound degrades in DMSO
  - Compound has poor oxidative stability
  - Compound has measured auto-fluorescence
  - Interferes in a specific assay format

  → Internal lookup table available

  - Has been active in a cytotoxicity assay
  - Has been active in a specificity assay*

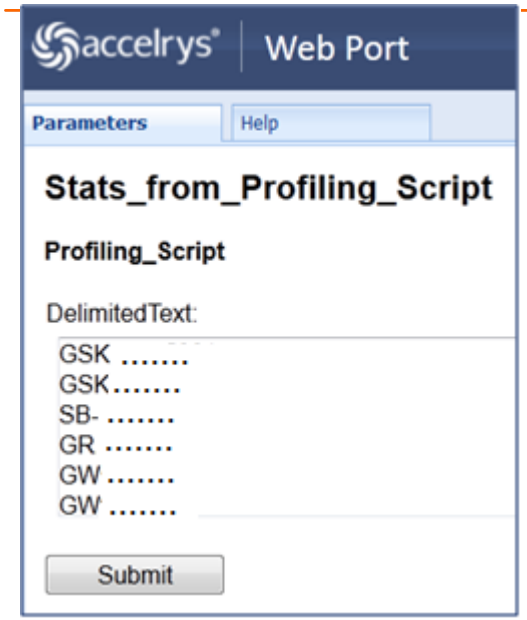
  → Hand curated internal lists of assays used

* Non-transfected cell line

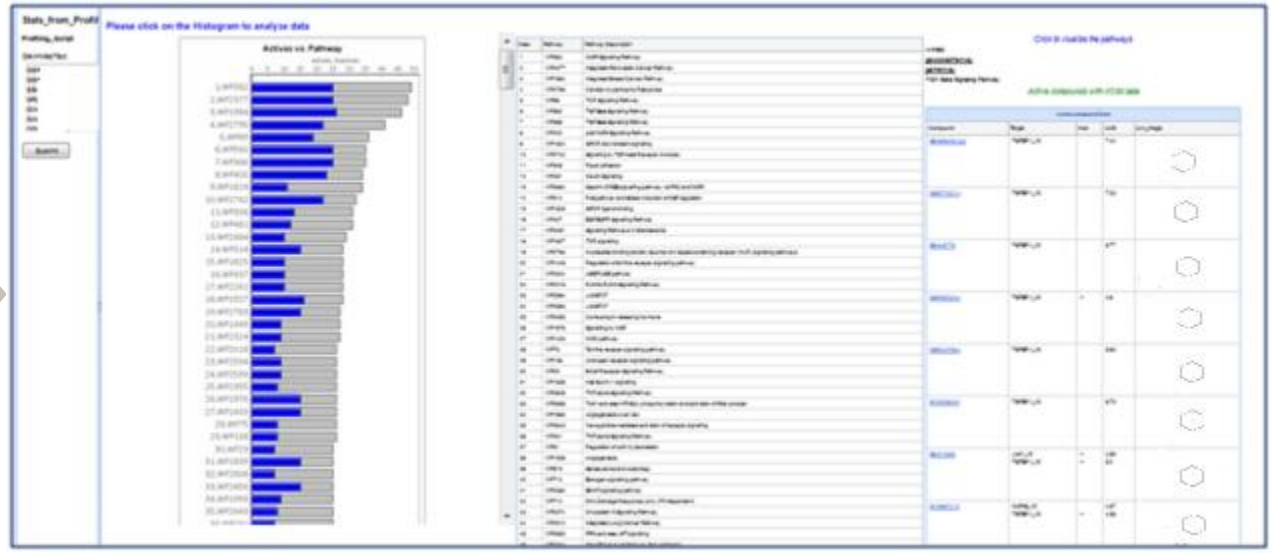
# Digging Into and Visualising the Data

- A Pipeline Pilot protocol has been developed to enable browsing of all data at an assay level for phenotypic assay hits
- Mechanistic biological information can be browsed via pathway maps
  - Compound target activities are overlayed onto wikipathways
  - Things to look for to improve confidence:
    - Are different targets involved?
    - Are there different chemotypes involved?
- Compound specific information, presented in terms of interactive html page, gives a comprehensive understanding of the profile of an interesting compound
  - Drill down allows you to see:
    - Activity at all targets (selectivity)
    - Behaviour in cells and technologies (nuisance? Endogenous activity?)

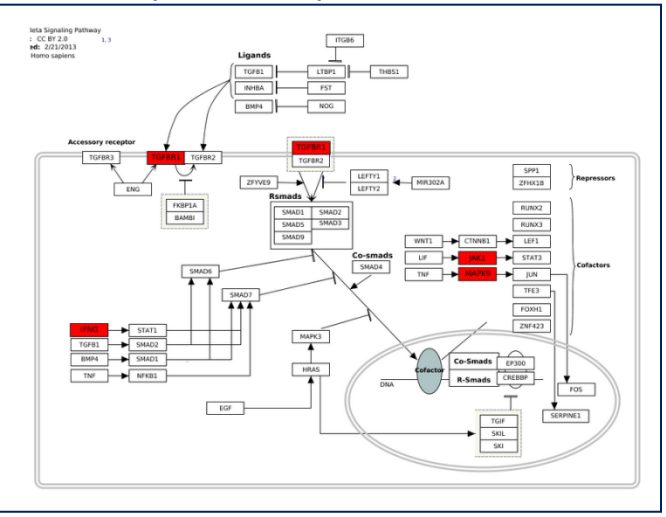
# Pipeline Pilot Protocol for Analysis of Phenotypic Screening Data




Pipeline Pilot protocol



Analysis from the Phenotypic screening

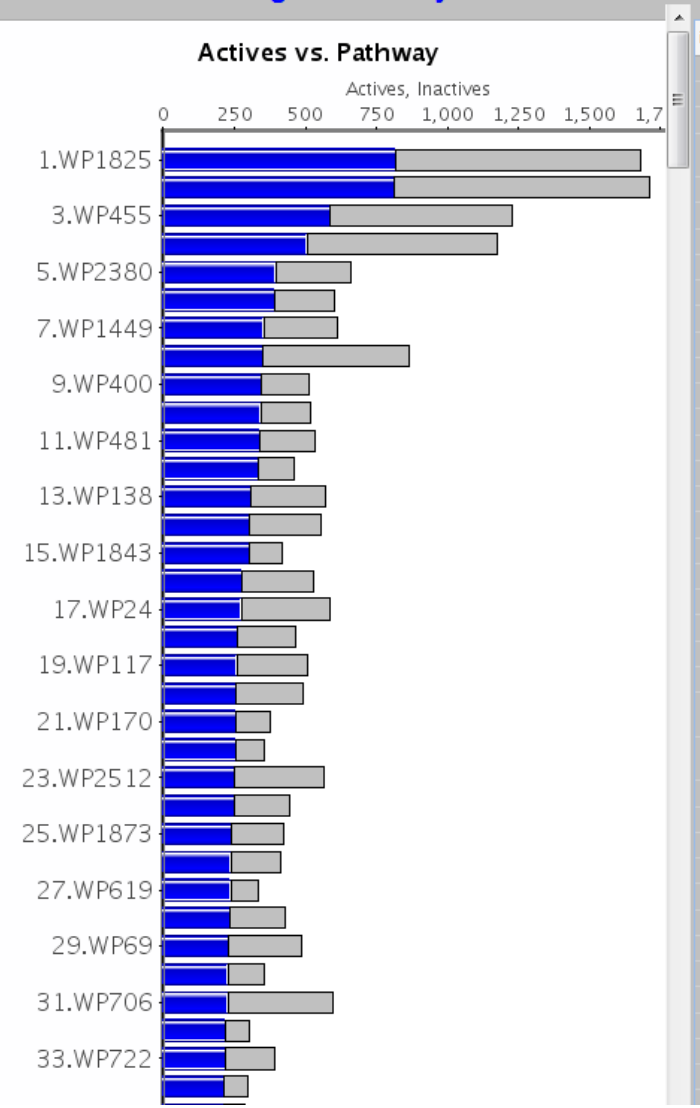



Pathway


Compound specific information

# Visualising Pathway Activities

Please click on the Histogram to analyze data




| index | Pathway | Pathway Description |
|---|---|---|
| 1 | WP1825 | GPCR ligand binding |
| 2 | WP1824 | GPCR downstream signaling |
| 3 | WP455 | GPCRs, Class A Rhodopsin-like |
| 4 | WP2664 | Gastrin-CREB signalling pathway via PKC and MAPK |
| 5 | WP2380 | BDNF signaling pathway |
| 6 | WP382 | MAPK Signaling Pathway |
| 7 | WP1449 | Regulation of toll-like receptor signaling pathway |
| 8 | WP2377 | Integrated Pancreatic Cancer Pathway |
| 9 | WP400 | p38 MAPK Signaling Pathway |
| 10 | WP585 | Interferon type I signaling pathways |
| 11 | WP481 | Insulin Signaling |
| 12 | WP437 | EGF/EGFR Signaling Pathway |
| 13 | WP138 | Androgen receptor signaling pathway |
| 14 | WP75 | Toll-like receptor signaling pathway |
| 15 | WP1843 | L1CAM interactions |
| 16 | WP1984 | Integrated Breast Cancer Pathway |
| 17 | WP24 | Peptide GPCRs |
| 18 | WP51 | Regulation of Actin Cytoskeleton |
| 19 | WP117 | GPCRs, Other |
| 20 | WP2059 | Alzheimers Disease |
| 21 | WP170 | Nuclear Receptors |
| 22 | WP1787 | Base Excision Repair |
| 23 | WP2512 | Integrated Lung Cancer Pathway |
| 24 | WP58 | Monoamine GPCRs |
| 25 | WP1873 | NGF signalling via TRKA from the plasma membrane |
| 26 | WP2864 | Apoptosis-related network due to altered Notch3 in ovarian cancer |
| 27 | WP619 | Type II interferon signaling (IFNG) |
| 28 | WP1976 | Signalling by NGF |
| 29 | WP69 | TCR Signaling Pathway |
| 30 | WP1433 | NOD pathway |
| 31 | WP706 | SIDS Susceptibility Pathways |
| 32 | WP734 | Serotonin Receptor 4/6/7 and NR3C Signaling |
| 33 | WP722 | Serotonin HTR1 Group and FOS Pathway |
| 34 | WP2637 | Structural Pathway of Interleukin 1 (IL-1) |
| 35 | WP2828 | Bladder Cancer |
| 36 | WP231 | TNF alpha Signaling Pathway |
| 37 | WP2808 | TNF alpha Signaling Pathway |
| 38 | WP2796 | Class I MHC mediated antigen processing &amp |
| 39 | WP2355 | Corticotropin-releasing hormone |

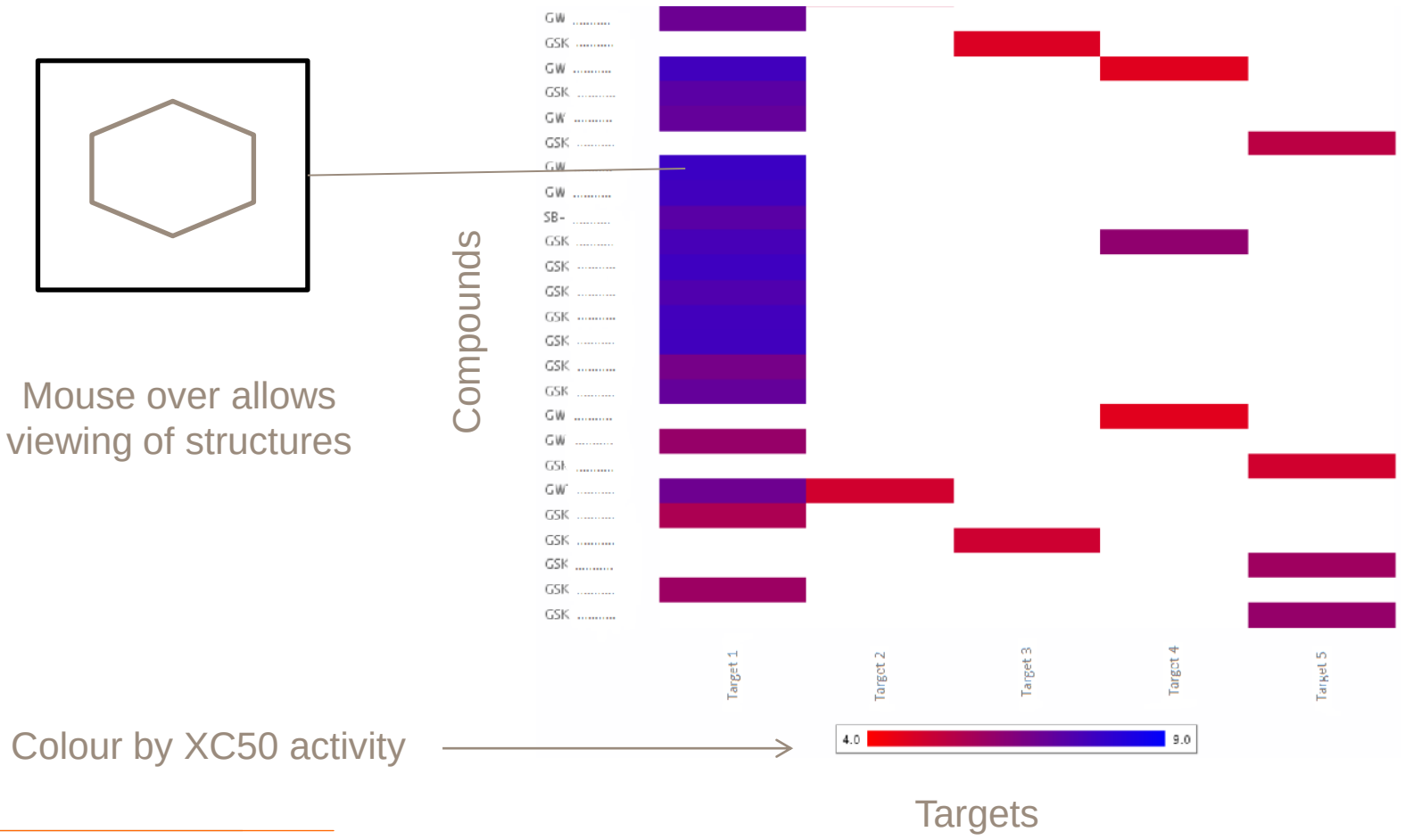
Some of the pathways aren't really pathways

Some duplication

Looking for bars like this:

# Visualising Pathway Activities

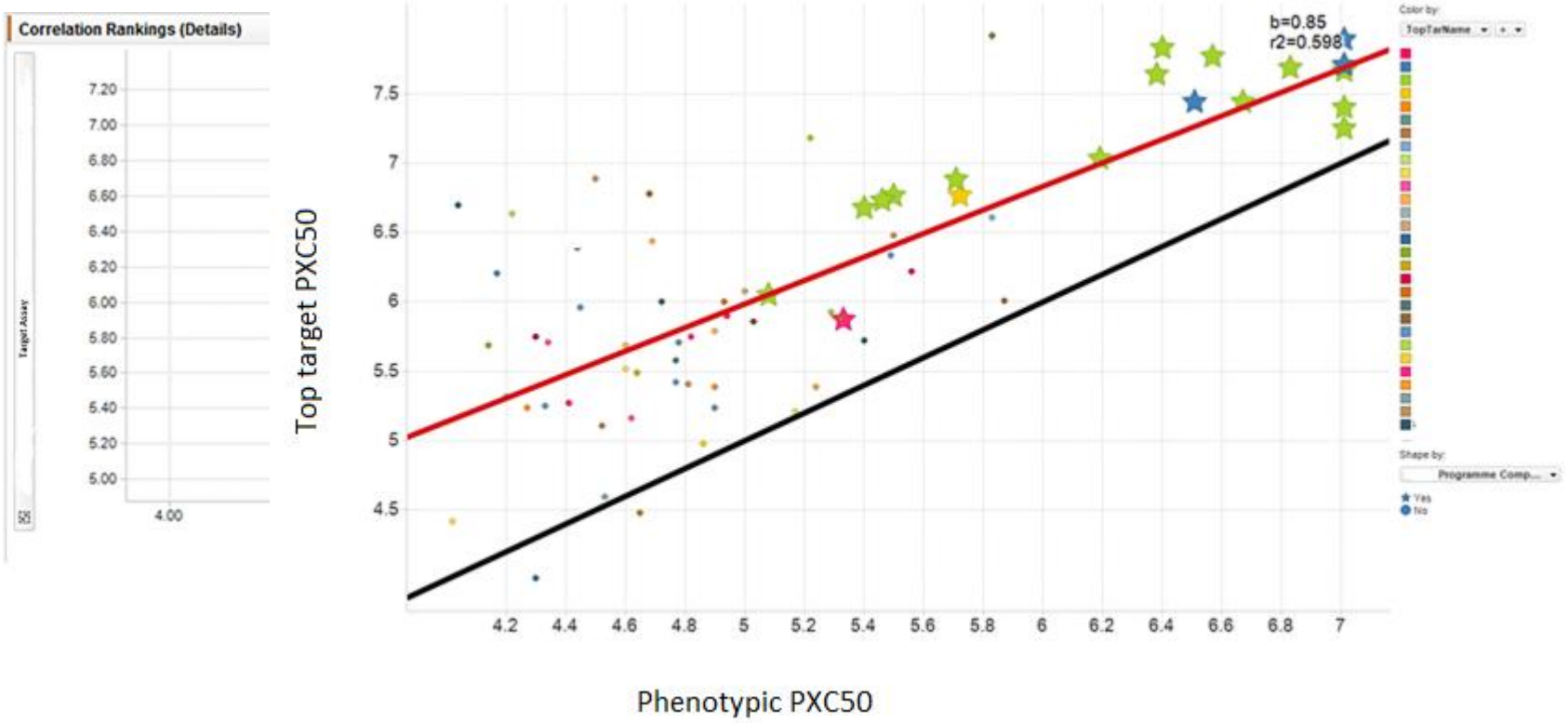
- Heat maps can display where compound activities lie in a pathway



# Correlations – Phenotypic vs Target Assays

Example

- Retrieving all assay data means that correlation analysis can be performed
  - Given a decent overlap between assays
    - Pearson/Spearman coefficient or probability can be calculated

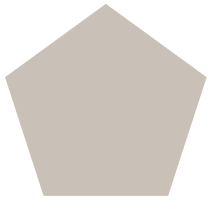



Can even correlate top target activity

# Compound Profiles - Similarity Expansion

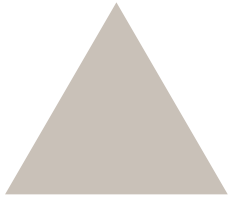
Example

- 3 Different chemotypes observed in hits for target X
  - Run similarity searching for near neighbours using cut off tanimoto 0.9
  - Assess the profiles of the near neighbours



Chemotype 1

No near neighbours
1 cmp has activity at target X



Chemotype 2

4 near neighbours
5 cmps have activity at target X

Chemotype 3

5 near neighbours
1 cmp has activity at target X
2 cmps have activity at target Y
1 cmp has activity at target Z

# Compound Selection

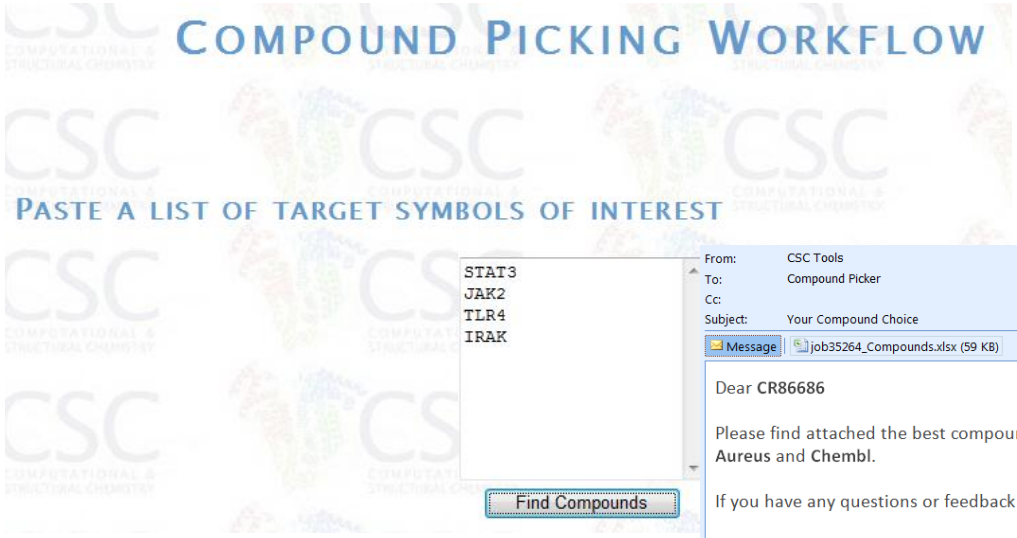
# Compound Selection

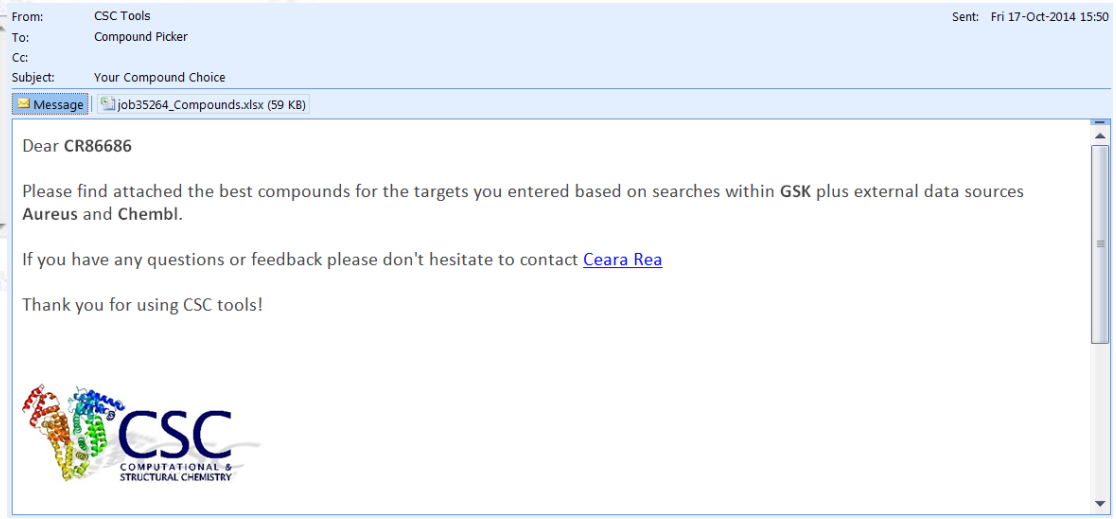
## Focussed Screening/testing hypotheses

- A list of targets is generated from a bioinformatics analysis
  - **INPUT**: Short gene name or list of (*deals with synonyms)
    - Queries for marketed drugs associated to that target
    - Queries for GSK candidates/leads for projects associated to that target
    - Queries tractable hits for projects associated to that target
    - Queries for compounds requested against a project associated to that target
    - Queries for compounds having a measureable result in an assay for that target
    - Queries the Aureus/Chembl for most potent compounds
- At this stage you have a list of compounds (which may be quite large). The top compounds are selected by a scoring triage:-
  - MarketedDrug: 20 points
  - Candidate: 15 points
  - Lead: 10 points
  - TractableHit: 5 points
  - Project code 3 points
  - Assays for target *potential (+3 points for each assay)
    - '=' Result 1 point
    - >5.5 1 point
    - >7 1 point

# Compound Selection

- Objective
  - Find ‘best’ compounds by target for a biologically focussed screen
- Website available

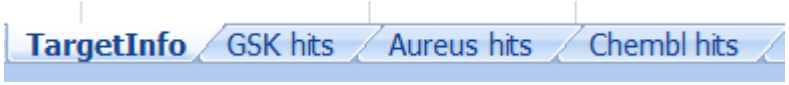


# Compound Selection

## Output File

- The data is split into different sheets for each data source with a master sheet giving extra annotation for GSK data




Data sources

Target IDs for Different Species/transcripts

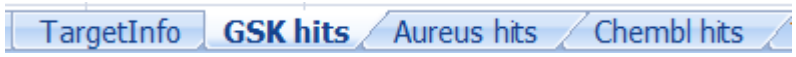
| | | | |
|---|---|---|---|
| **STAT3** | | | |
| | Found GSK Chembl | | |
| | | **GSK Tarids** | tar159650;tar159649;tar49695;tar30456;tar110565;tar64179;tar103011;tar151774 |
| | | **GSK Names** | STAT3_V0;STAT3_V1 |
| | | **GSK Projects** | |
| | | | Project_ID<1>:Project title for ID 1 (MoA) |
| | | | Project_ID<2>:Project title for ID 2 (MoA) |
| | | **GSK Assays** | |
| | | | Assay_ID<1>:Title for assay 1 (inc cell line and technology) |
| | | | Assay_ID<2>:Title for assay 2 (inc cell line and technology) |
| | | | Assay_ID<3>:Title for assay 3 (inc cell line and technology) |
| | | | Assay_ID<4>:Title for assay 4 (inc cell line and technology) |

Projects and Assays queried and used for the Scoring

# Compound Selection

## Output File

- For Compounds found in GSK:-




Sources are mdrugs/candidates/leads/tractable hits/assay hits

| 1 | TARGET | Source | Name | Compound Number | BestTarget XC50 |
|---|---|---|---|---|---|
| 941 | TARGET1 | mdrugs | DRUGNAME | CCI ..... | |
| 942 | TARGET1 | candidates | | GW ..... | |
| 943 | TARGET1 | leads | | GW ..... | |
| 944 | TARGET1 | activecmplist | | GSK ..... | 7.1 |
| 945 | TARGET1 | activecmplist | | GSK ..... | 6.5 |

Compounds may not have an XC50 value if they haven't been tested in house

# Compound Selection

## Output File

- For compounds found external to GSK, compound is mapped back to GSK & full availability info is given:

Inclusion in GSK compound sets

Internal availability

| MarketedDrugs | HTS | Available Solid | Available Solution | Reserved Status | eMolecules | PubMed_ID |
|---|---|---|---|---|---|---|
| | | | | | EM3386446 | 24177366 |
| | yes | no | yes | | EM4391607 | |
| | | | | | EM45532723 | |
| yes | yes | yes | yes | | EM500940 | |
| | yes | yes | yes | | EM6837214 | |
| no | no | yes | yes | RESERVED | | |
| | | | | | | 23061660 |
| | | | | | | 19945871 |

External availability

Reference to original paper

# Phenotypic Profiling

# Phenotypic Profiling

- There are many published phenotypic assays available
  - Possibility to profile across them
  - Current internal compound profiling workflow only looks at target assays
  - Biological fingerprinting of external data would be difficult as completely different sets of compounds tested in the assays
  - A hit in a similar (but not the same) phenotypic assay, may give valuable insight into a compound's behaviour

- Need to identify which are the phenotypic assays
  - Problem internally and externally
    - ChEMBL has assay type 'F' (functional), filter out those with a target?

- Need to classify types of phenotypic assay?

  ##### Not all phenotypic assays are the same #####

  - An imaging assay measuring neurite out growth is very different to one measuring gene expression by fluorescence/luminescence

# Phenotypic Profiling

## Panel Data

- GSK is working with an external partner to profile compounds in a range of human disease models (system mechanistic approach)
  - Primary cell lines are being used (diseased/non-diseased)
  - Cells are stimulated and then treated with compounds
  - Various phenotypic endpoints are measured in each system
    - Typically expression levels of biomarker(s) (up/down regulated)
    - Can be other more physical measurements, eg cell count (proliferation/cell death)

- Relevant meta-data captured for each system:

| Cell | Diseased Healthy | Stimuli | BioMarkers | Disease / tissue relevance | Predicted targets |
|---|---|---|---|---|---|