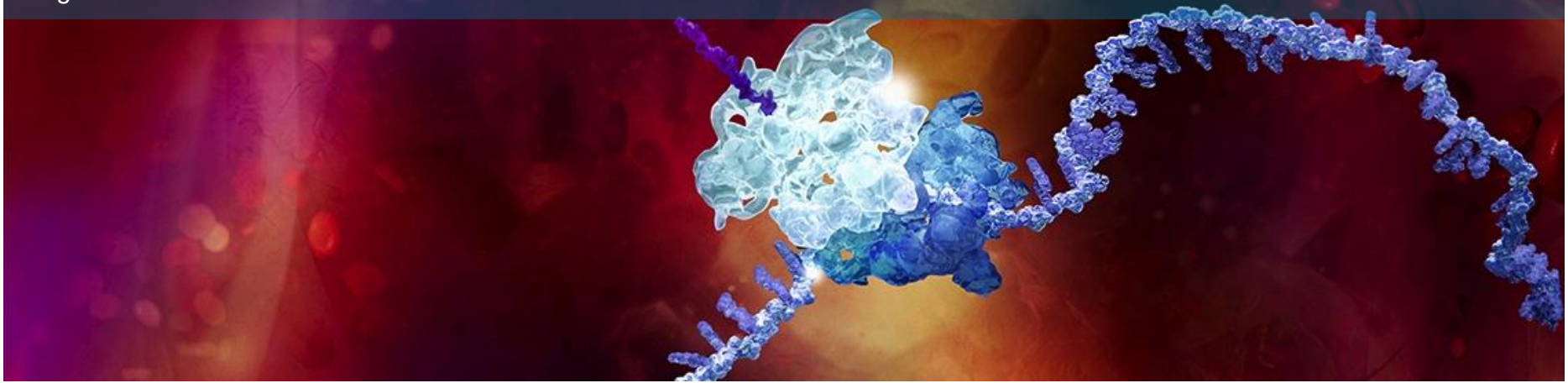
AstraZeneca
IMED Biotech Unit

An introduction to Phenotypic Screening

Ola Engkvist External Sciences, Discovery Sciences, AstraZeneca Mölndal

BigChem Lecture

2017-02-01



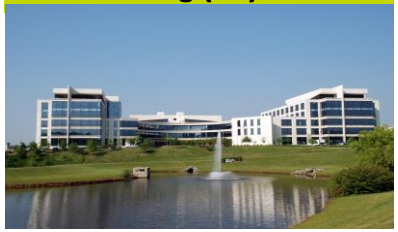
Global dimensions

- **$24.7bn** Total Revenue; $23.6bn Product Sales; $1.1bn Externalisation Revenue
- **61,500** employees
- **$5.6bn** invested in R&D with research across 5 countries
- **125 projects** in clinical development and **15 NMEs** (new molecular entities) in late-stage development; **18 NME approvals** in 2014 and 2015
- More than **850** collaborations and partnerships globally
- Manufacturing in **17** countries
- 4th fastest-growing top 10 multinational pharmaceutical company in emerging markets in 2015

As at 31 December 2015

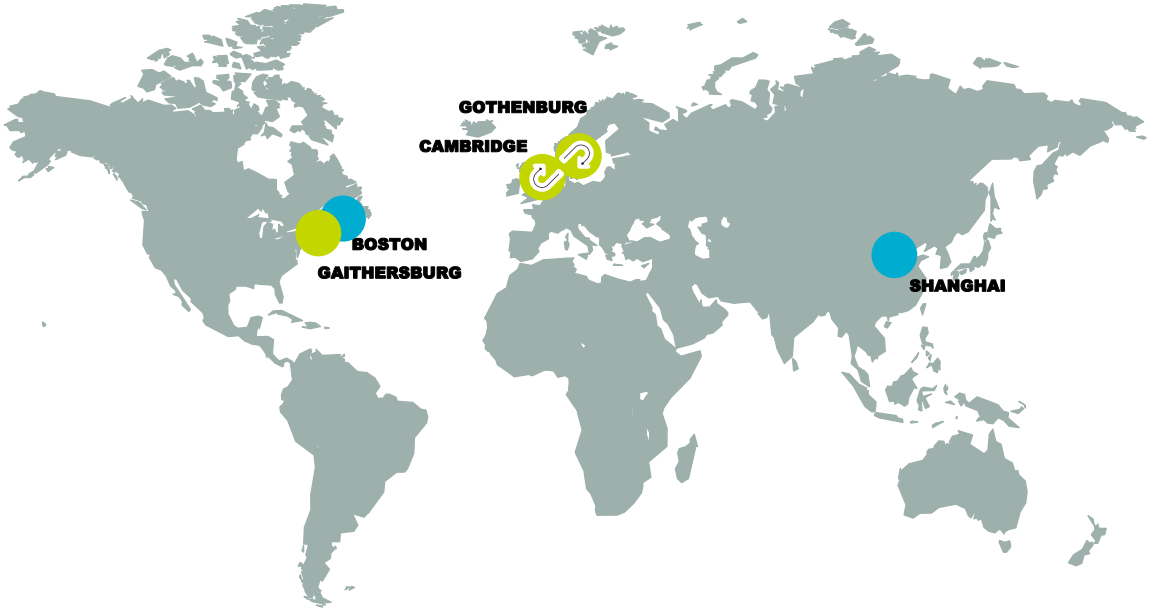
Three strategic R&D sites close to global bioscience clusters

Gaithersburg (US)

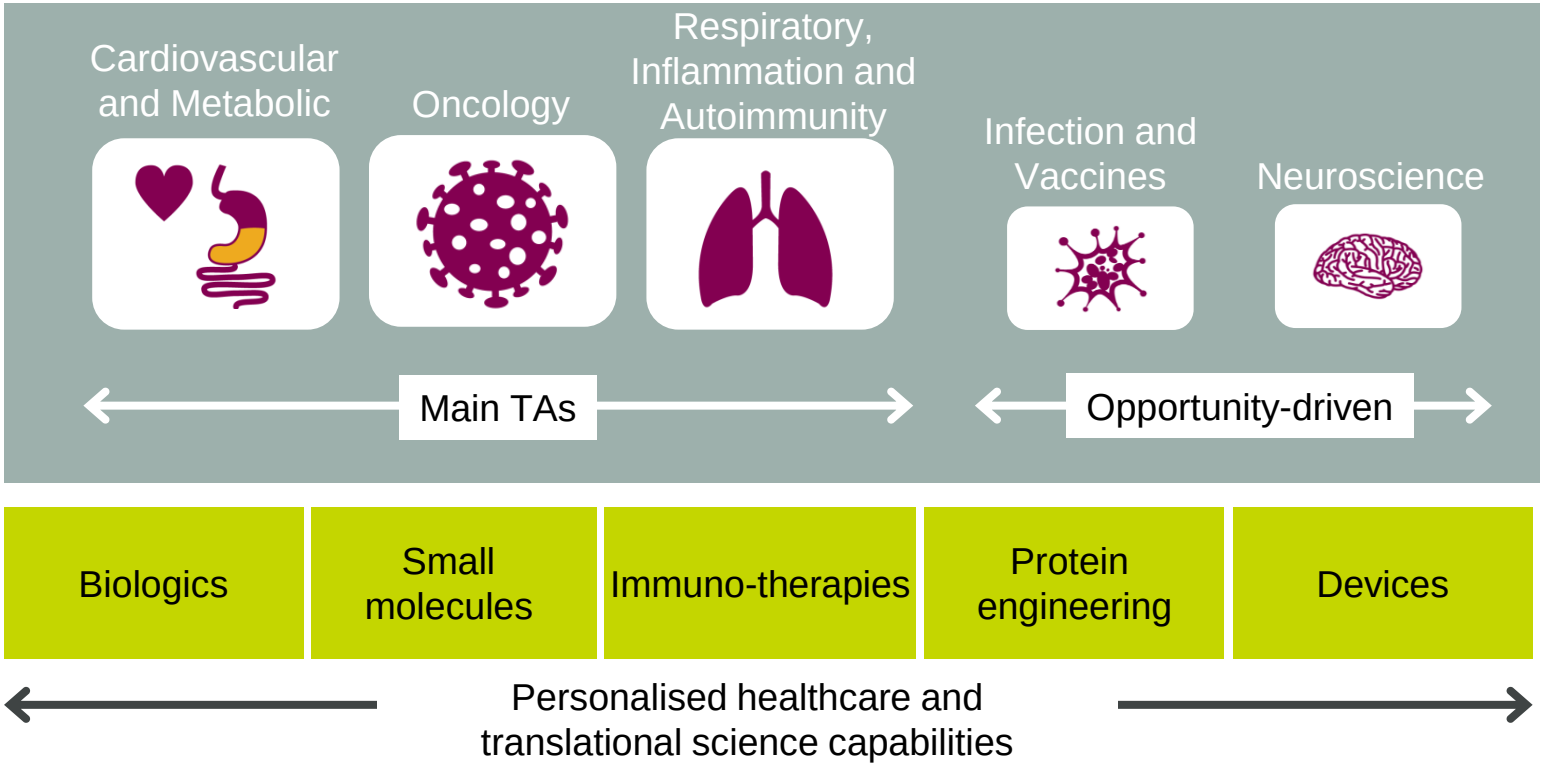


Gothenburg (SE)

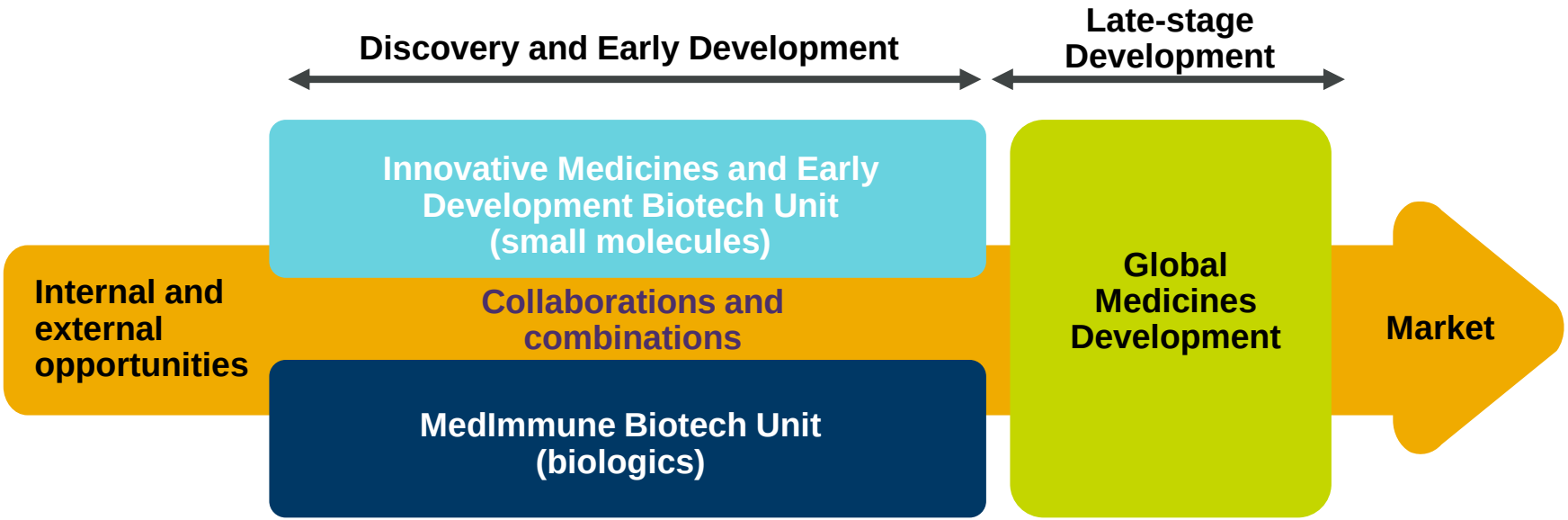
Cambridge (UK)



Focus on three main therapy areas across key platforms



Biotech units collaborating with each other and externally



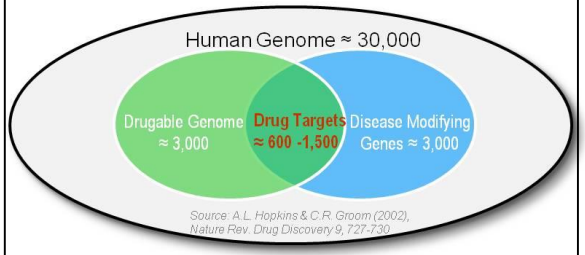
Drug Discovery – 21st Century
The Challenges

Success rate PhII very low at **< 20 %**
Success rate PhIII up to **57 %**

References:
J. Arrowsmith (2011) Nature Rev. Drug Discovery 10, 328
J. Arrowsmith (2013) Nature Rev. Drug Discovery 12, 569

Lack of Translatability:

Preclinical models don't match very well with patients

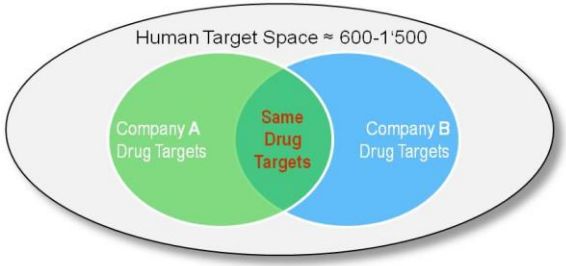


Total no. hu drug targets: 600-1'500
Total no. validated targets: ~ 680

References:
Drews, 2000; Hopkins & Groom, 2002; Russ & Lampel, 2005; Imming et al, 2006; Overington et al, 2006; Li et al, 2007; Plewczynski & Rychlewski, 2009; Mayr & Bojanic, 2009; Rask-Andersen et al, 2011;

Lack of Novel Modalities for Intervention:

Large parts of the genome seen as "undrugable"



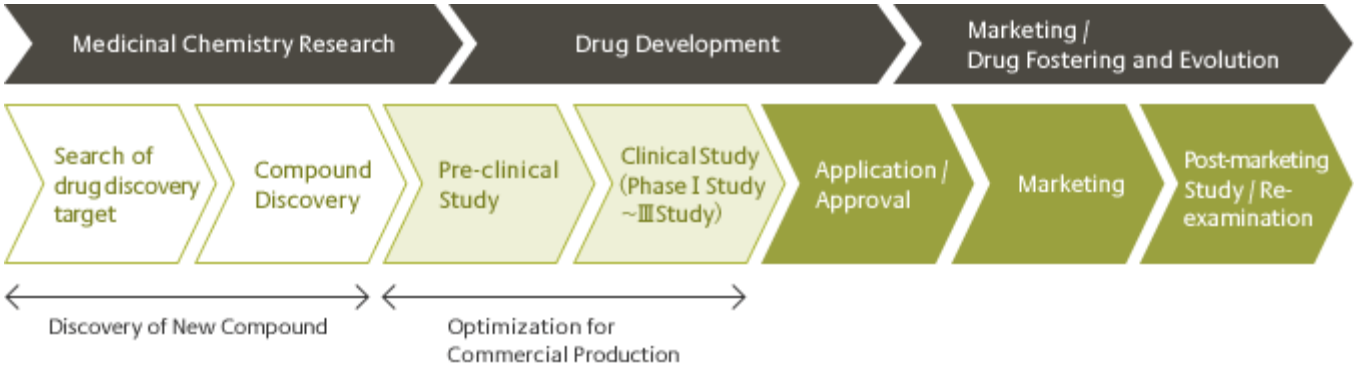
86 % overlap (target shared by >2 of 18 companies)
75 % targets from 3 classes (kinase, GPCR, protease)

References:
Paul L. Leeson & Stephen A. St. Galley (2011)
Nature Review Drug Discovery 10, 749-765.

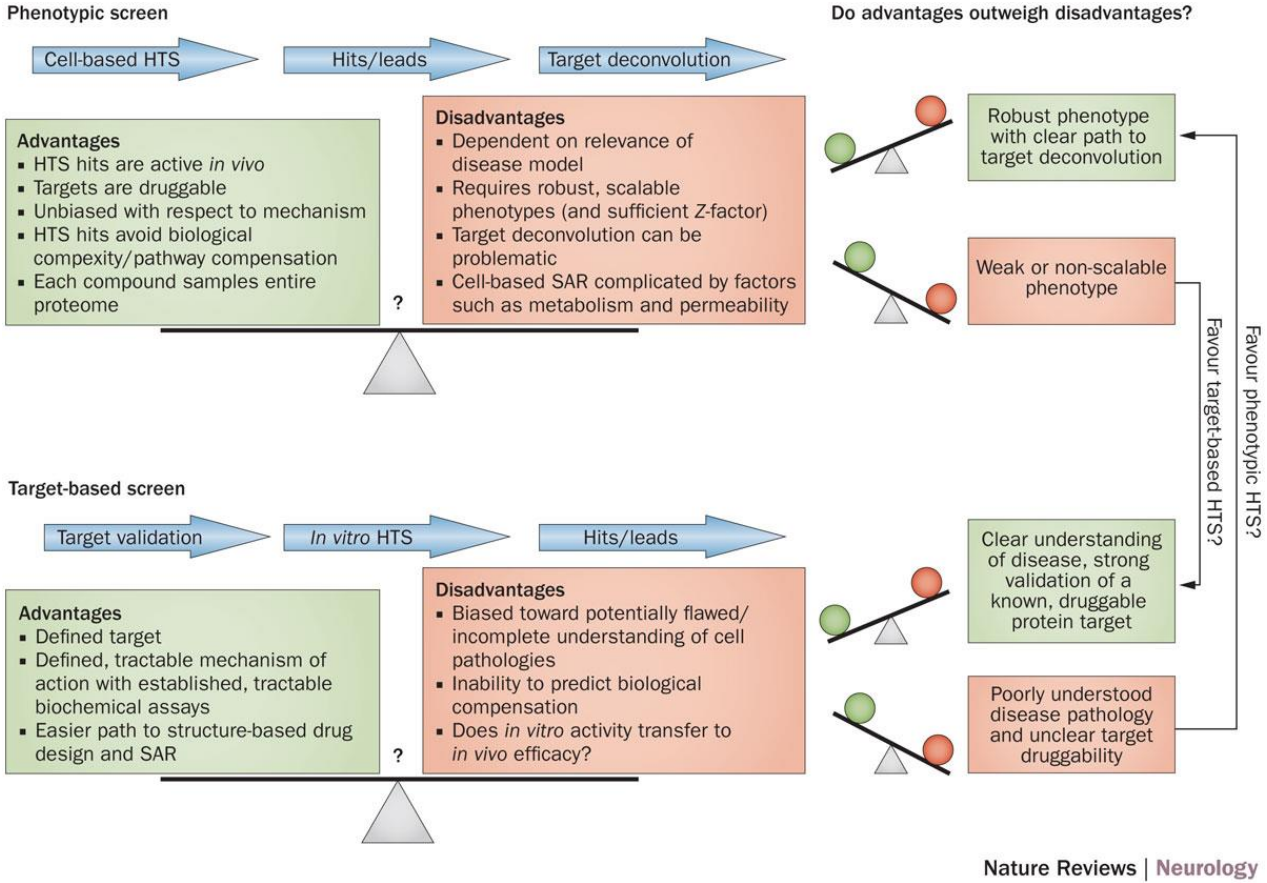
Lack of Novel Targets:

Shortage of novel, validated, molecular targets

Drug Discovery Process



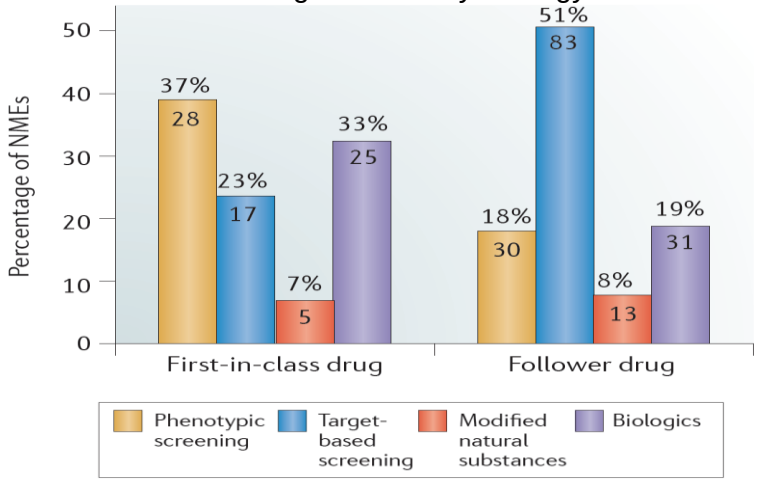
Phenotypic screening versus target based screening



Why Phenotypic Discovery?

Phenotypic Screening – Relevance for Drug Discovery

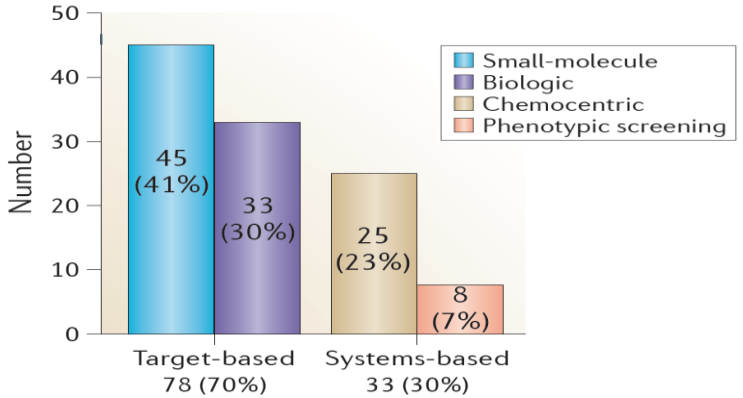
Distribution of new drugs discovered between 1999 and 2008, according to discovery strategy.



Reference:

- Swinney, D.C. & Anthony, J. (2011) How were new medicines discovered? *Nature Reviews Drug Discovery* 10, 507-519.

Discovery of first-in-class drugs approved by the FDA between 1999 and 2013.



Reference:

- Eder, J, Sedrani, R. & Wiesmann, C. (2014) The discovery of first-in-class drugs: origin and evolution. *Nature Reviews Drug Discovery* 13, 577-587.

Phenotypic screening is one of several important strategies for developing novel drugs

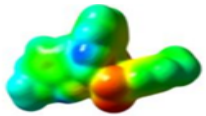
Integrating different target ID strategies

Phenotypic Screen Hits | Targeted leads with off-target effects | Literature cmpds with unknown MoA

Small molecules modulating unknown Targets

In Silico Prediction

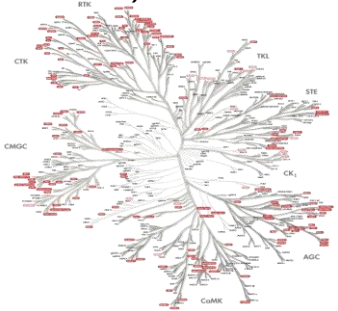
- Internal & External Bioannotation
- Target Predictions
- ID of near neighbors



Annotated targets

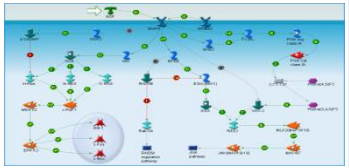
Biochemical Profile

- Protein families assays (eg, kinases, GPCRs)



Bio-Signatures

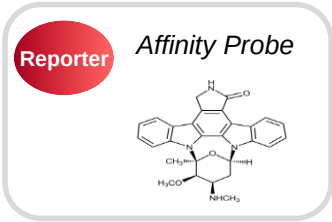
- Gene Signatures
- Signaling Pathways



Annotated targets

Affinity Chemo-Proteomics

- In-lysate affinity probes
- In-cell affinity probes
- Affinity matrices (eg Kinobeads™)



Small molecules modulating identified targets

How can informatics facilitate knowledge based phenotypic screening?

Chem- and Bioinformaticians part of the project team from start

Up-front work

- Collection and integration of relevant data
- Method development and validation of *in silico* target identification methods

Selection of screening set

- Maximize the target coverage
- Knowledge based

Analysis of screening data

- Target enrichment analysis
- *In silico* target prediction
- Pathway analysis

Data sources and its (difficult and time consuming) integration

Bioactivity data

Public data (ChEMBL)

Commercial data

In-house data (SAR, HTS)

ChemistryConnect (BAO)

Pathway data

-Omics data

Disease data

Semantic data integration needed

Open PHACTS

CORTELLIS DATA FUSION

THOMSON REUTERS

Which compounds to screen?

Screening set used depends on the assay throughput, project objective, available knowledge, translatability etc

Diversity set (up to 2M)
Chemical structure diversity

Phenotypic screening set (19K)
Biologically annotated – Maximally broad target coverage

Target based diversity subset

Knowledge based subset

in vivo compliant subset

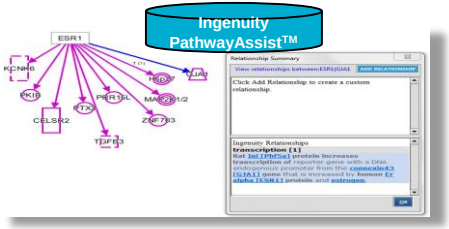
How is the phenotypic screening set composed?

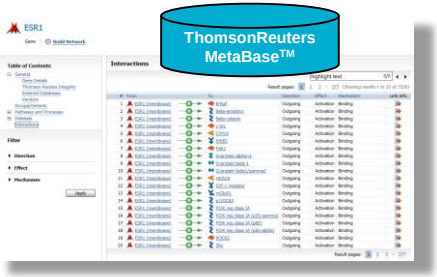
AZ Tool compounds (AZDs, panel profiled compounds)	4 K
Acquired external tool compounds	6K
Crowd sourced	7 K
Maximize coverage of target space	5K

- The set consists of in total ~19K available compounds
- The set is run in all HTS to improve bio-annotation
- All compounds have measured purity (>85%)
- Should be complemented with a text mining search for a final selection of compounds
- Part of the AZ Open Innovation initiative

Knowledge-led compound selection

- Querying of multiple information sources:



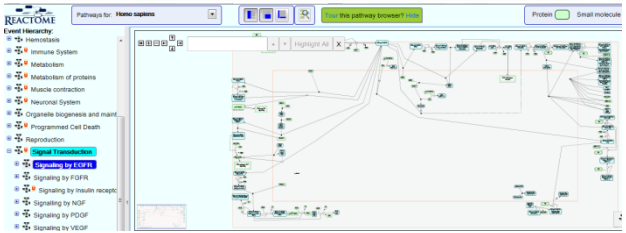


Proteins annotated with relevant bioprocess terms

Gene Ontology Consortium

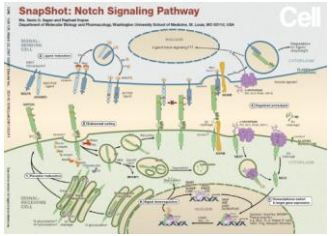
Upstream modulators of a key regulator

Key reviews, Disease experts



(e.g. Reactome)

"pathway" maps



(e.g. Cell journal "snapshot")

Medline abstract text mining

Proteins with a link to biology of interest

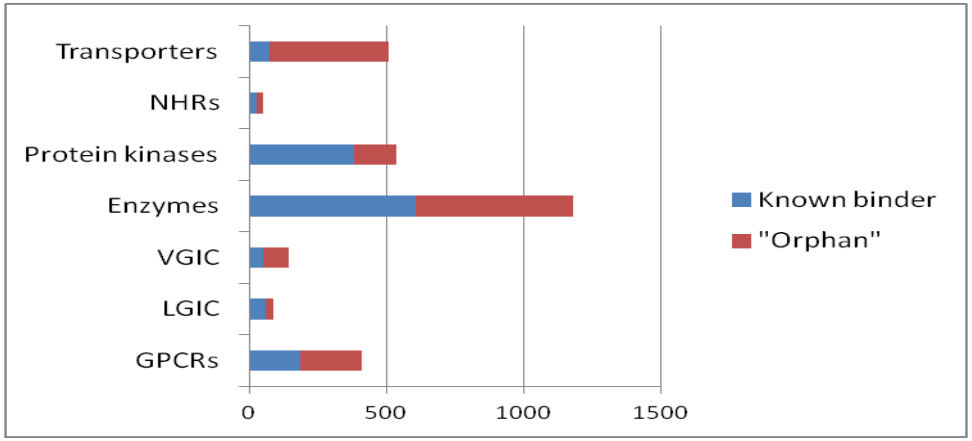
Expansion of extracellular proteins to their potential receptors (more likely to have compound hits)

Expansion of groups/complexes into their constituent proteins

How well does the set cover established target classes?

Definition a molecule with an activity better than 100nm

In total ~1600 targets are covered of 20K protein, however, large differences in target class coverage



Thus there exists many gaps also in the known “druggable” target classes

Many target classes doesn’t have any known small molecule modulators at all

Pro and cons with an annotated phenotypic screening set

✓Targets can be identified from a small screen

✓Targets can be identified quickly

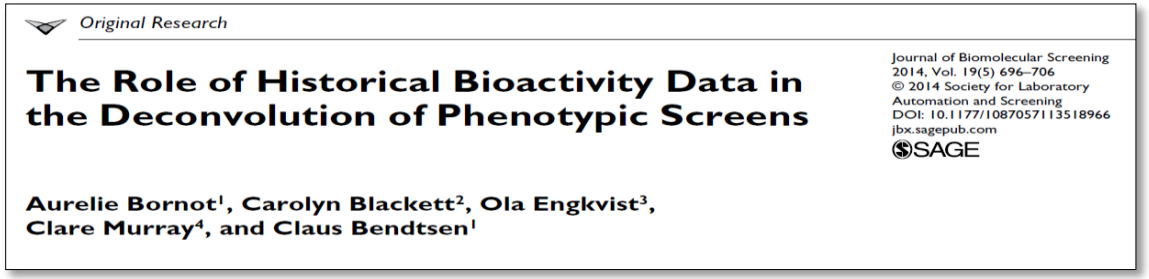
✓No-go targets, to be counter-screened for in a diversity screen, can be identified.

- Probability to identify a novel target is relatively low
- Needs to be validated by genetic methods

Target enrichment analysis of screening hits

- What are the targets of compound hits from our phenotypic screen?
  - Can we use historical bioactivity data of our compound set to find targets enriched in the hits, thus providing target hypotheses?

- Approach



Original Research

The Role of Historical Bioactivity Data in the Deconvolution of Phenotypic Screens

Aurelie Bornot[1], Carolyn Blackett[2], Ola Engkvist[3], Clare Murray[4], and Claus Bendtsen[1]

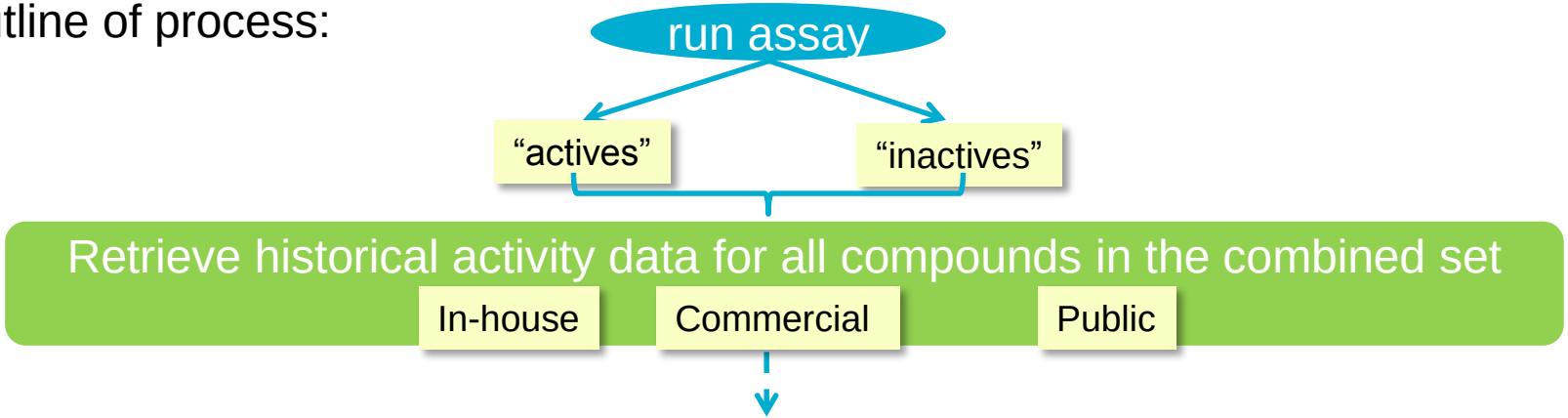
Journal of Biomolecular Screening
2014, Vol. 19(5) 696–706

DOI: 10.1177/1087057113518966
jbx.sagepub.com
SAGE

- Retrospective validation: TNFα production in lipopolysaccharide-stimulated THP-1 cells

Target enrichment analysis of screening hits

- Outline of process:



- Construct contingency table for each target

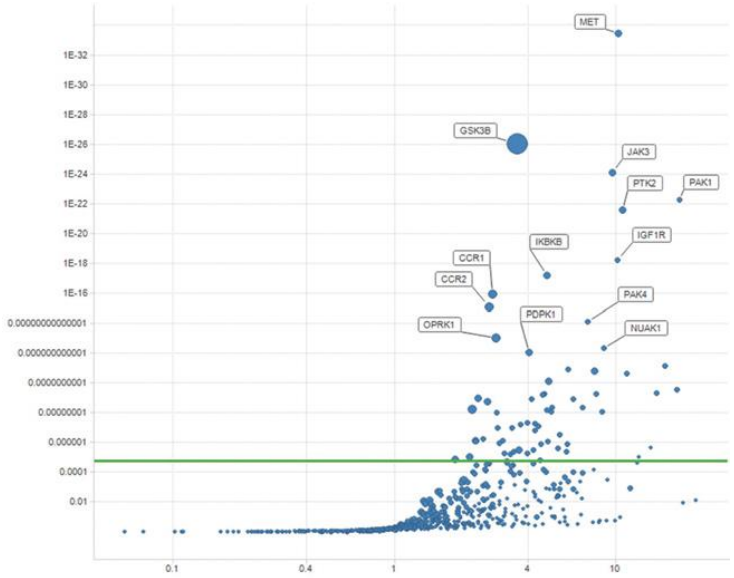
- Apply Fishers Exact test and Bonferroni correction

Target		Bioactivity	
		Active	Inactive
Screen	Active	576	66
	Inactive	2233	916

(p-value = 2.2e-26, odds ratio = 3.58)

Statistical enrichment scores for each target

Target enrichment analysis of screening hits



Log10 Fisher Test p-value

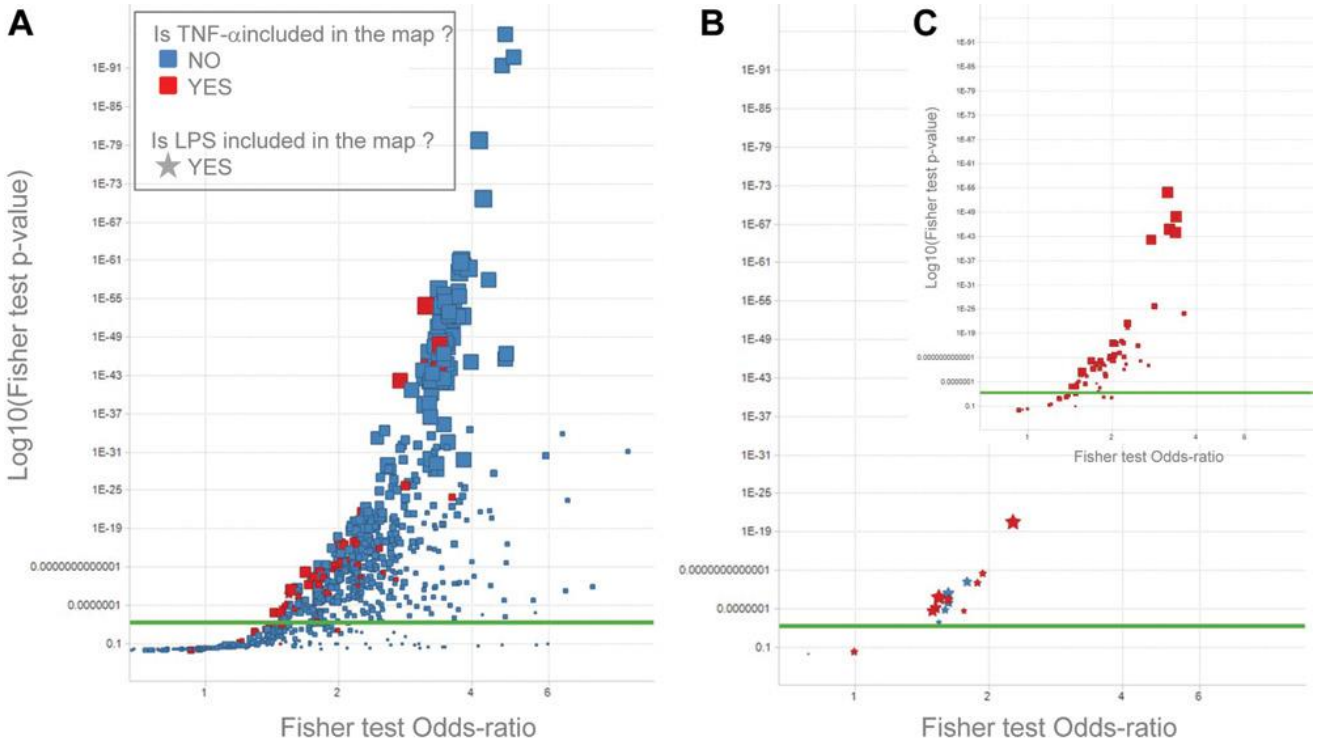
Fisher Test odds ratio

Targets passing an enrichment significance threshold

Things to take into account:

- Activity cut-off
- Inactive set
- Targets with few annotated compounds
- Correlation between different targets
- Include predicted actives

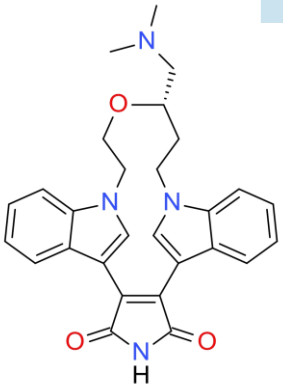
Pathway enrichment



Some enrichment seen, but needs to be validated further

In silico target fishing

Active compounds	Inactive compounds
Public, Licensed, in-house	**AZ HTS Datamart & PubChem**
Ligand 3 – Target 1	Ligand 1 – Target 1
Ligand 4 – Target 2	Ligand 1 – Target 2
Ligand 5 – Target 2	Ligand 2 – Target 2
...*n active* bioactivities = **8,505,197**	...*n inactive bioactivities* = **598,923,798**



Orphan compound(s)

- **Random Forest Classifiers**
  - 2,712 Targets
  - Balanced for 1:100 ratio of actives vs. inactives

Percentage of decision trees that predict compound as active

Target EGID	Random Forest Score
5590	1.00
3718	0.99
...*n targets* =***2,712***	...

Platt scaling converts output from the Random Forests into true probabilities

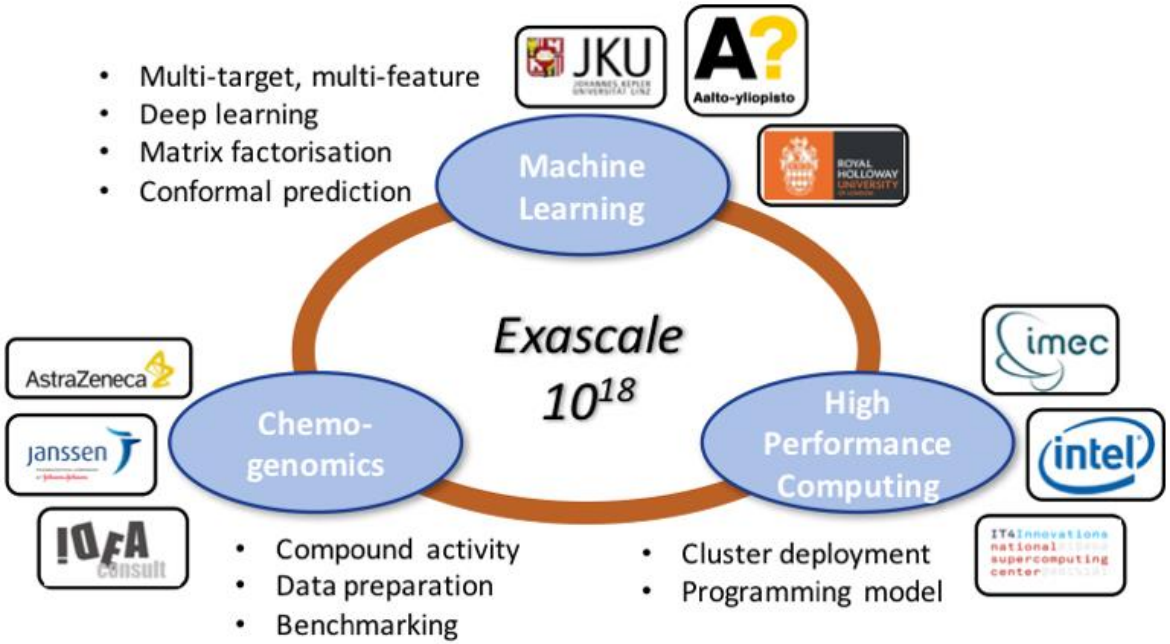
Scaled likelihood of true positive prediction

Target EGID	Probability Score
5579	0.99999
814	0.99922
...*n targets* = ***2,712***	...

IMED Biotech Unit I Discovery Sciences

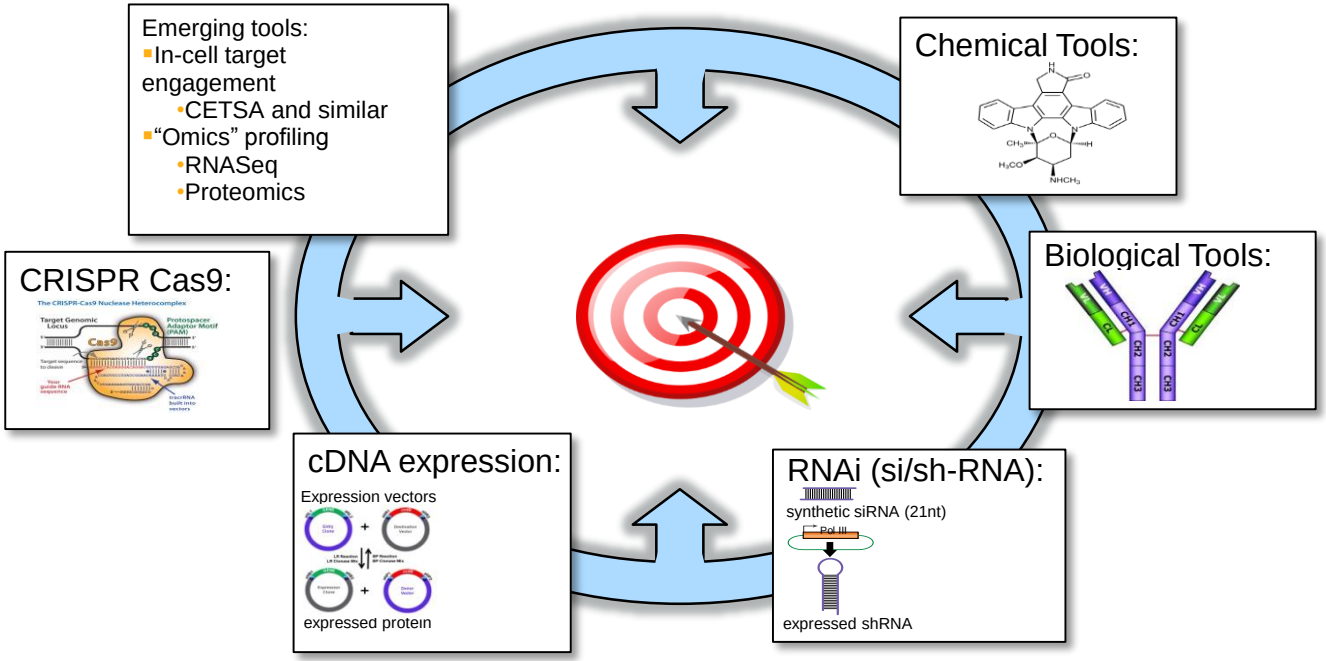
Lewis Mervin, Andreas Bender, University of Cambridge

Large scale chemogenomics prediction: Excape project



http://www.excape-h2020.eu/index.php

Target Validation: Building Confidence



Considerations and Conclusions:

- A large arsenal of technologies can be applied
- The choice(s) of technology depends on the question you are asking and the cells you are interrogating
- Combine data to build confidence in the target(s)

Conclusions

- Chem- and bio-informatics are integral components in a phenotypic screening cross disciplinary project team
- Progress have been made in data integration, but more to do
- Annotated tool compounds plus target enrichment analysis can deliver a target hypothesis very rapidly
- Limited possibilities to identify novel targets from a tool compound set
- Target hypothesis must be further experimentally validated

Confidentiality Notice