Genomics in Drug Discovery

John Whittaker
Outline

Scientific background
- Motivation
- Success and failure in drug discovery

What can we do differently?
- How can data science/genomics help?
- Open innovation and drug discovery

Summary
- Where are we?
- Where are we going?
Drug discovery
~6 years, 30% of cost

- Start with a therapeutic hypothesis; attempt to prove/refute/refine it
- Only definitive proof is in the clinic

<table>
<thead>
<tr>
<th>Disease or Target Hypothesis</th>
<th>Target Selection</th>
<th>Hit ID</th>
<th>Lead Generation</th>
<th>PreCandidate Evolution</th>
<th>Candidate Selection</th>
<th>Early Clinical Dev't</th>
</tr>
</thead>
</table>

**Hypothesis**
- Literature
- Compound that changes human clinical endpoint

- Gene expression
  - Animal models
  - Cell-based studies
- Tool compounds
  - Experimental med studies
- Biomarkers in humans (imaging, -omics)
- Micro-dosing

- Pathway context
- Mechanistic models
- Genetic association
- Systems biology predictions
Motivation

Eroom’s law and failure

Scannell et al, NRDD 2012

Hay et al, Nat Biotech 2014

Probability of success at target selection 3%
Genetics 101

From Genes to Proteins

Genes contain instructions for making proteins.

Proteins act alone or in complexes to perform many cellular functions.
Genetics 101

- Variant/polymorphism: variation in DNA sequence
  - Allele: one of two or more forms of the DNA sequence at a single location
  - Genotype: the pair of alleles at a specific location in an individual
- Concentrate on single basepair changes as shown (SNPs)
  - About 3 million common SNPs identified; many more rare ones
  - Total genome length: ~3 billion letters
- Two chromosomes so three possible genotypes e.g.: T/T, T/C, C/C
- Look for association of SNP with disease onset/progression/etc
- NB low prior probability of association implies high likelihood required for convincing evidence (typically $5 \times 10^{-8}$)
Many associations of genetic variants with disease

GWAS + sequencing of rare diseases

- 2015: 2,154 studies. 15,333 SNPs

Does genetic evidence predict success?

Harder than you think.....

- Integrate GWASdb, OMIM, PharmaProjects
- Map associations to genes through LD, gene expression (eQTL), and regulatory elements
- Score developed to quantify evidence for causality
  - LD
  - Functional information
  - Number of other possibilities
- Map traits to MESH terms and use MESH hierarchy to estimate similarity between genetic traits and drug indication

ie, genetic variant is an instrumental variable
The value of genetic evidence

Drugs with human genetic information >2x more likely to be successful

Proportion of "novel" targets with genetic support for ongoing indication (2015 data). GSK about average.

2016 R&D objective

>50% of targets with genetic evidence

Nelson et al, 2015 Nature Genetics

Genomics in drug discovery

12/07/17
Path to a medicine?

*NB: not the only way....*

Genetics drives target choice

- Genetic associations yield > 2x increase in POS
- >1 rare disease gene a week
- GWAS revolution
- Informatics allows integration

Cellular / tissue model

- Appropriate IPS or primary cells
- Molecular fingerprint via omics
- Understand genetic mode of action
- Gene editing to confirm hypothesis

Experimental medicine study

- Biological understanding drives population and endpoint choice
- May be genetically defined

Largely pre-competitive biology

**Genomics in drug discovery**
Experimental models

*Bias towards genome scale*
What’s a cellular model?

*NASH: Liver cell line treated with FFA accumulates lipids*
Research reproducibility

And the lack of it.....

Genomic responses in mouse models poorly mimic human inflammatory diseases

Two primary areas of critique:
- Translational relevance
- Reproducibility

The ability of animal studies to detect serious post-marketing adverse events is limited

Peter J.K. van Meer a,b*, Marloos Kooijman b, Christine C. Gispen-de Wied c, Ellen H.M. Moors b, Huub Schellekens a,b
Design and analysis

Example: single protein measured in serum

95% confidence interval on drug effect is (-1, 114), p-value = 0.053

Q. How many more serum samples do we need to demonstrate a drug effect?

Log transform, paired data A. Zero if we correct the analysis.

Correct 95% CI on drug effect is (11.3, 160), p-value = 0.007
Causes of irreproducibility

Open Targets
an innovative public private partnership

John Whittaker
Wellcome Genome Campus
www.opentargets.org
Mission

Aim: To be the world leader for human target discovery

We will combine large-scale genomic experiments with objective statistical and computational techniques to identify and validate the causal links between targets, pathways and diseases.

We will accelerate and enable research and innovation by making the evidence open and accessible to all.
The Partners

The partners shared the idea that target validation could be improved but that one institute could not necessarily do it alone.

A strong desire to collaborate based on
- highly complimentary skills set,
- existing strong relationships,
- real commitment to the endeavour.
Target Validation Knowledge Cycle

Public Databases and Pipelines

New experimental data
Physiologically relevant and at scale

Oncology  Immunity  Cross-Disease

Neuro

www.targetvalidation.org

Target Validation Platform

www.targetvalidation.org
Characterising Coding and Non-Coding Genome Regions Contributing to Cellular Phenotype

RNA-Seq
- Snap shot of whole genome RNA expression
- Captures known and novel coding & non-coding regions

ChIP-Seq
- Chromatin Immuno-Precipitation, requires antibodies to targets
  - H3K4me3 (active), H3K27ac (active promoters/enhancers), H3K27me3 (poised), CTCF (active or repressed)

ATAC-Seq
- Assay for Transposase-Accessible Chromatin
- Identifies open chromatin associated with active transcription

SNP6 Karyotyping
- Confirm cell type identity using SNPs and copy no. variants
Synthetic lethal screen in cancer cell lines (Mathew Garnett, Kosuke Yusa, Chris Carpenter)

Genome-wide lentiviral KO library

Cancer cell lines

\[ g_3 \delta_3 \xi_3 \delta \alpha_3 \pi_4 \phi \]

Mutants

Sequencing read-out

Identify molecular features correlated with lethality
Where are we going?

EHR + genetics + ‘platform’ experiments + better assays
Two current collaborations
UK Biobank

500,000 participants from the UK recruited 2006-2012

- Consent for all types of health research, as well as follow-up and some recontact
- Baseline questionnaires, measurements, and biospecimen banking
- Health outcomes measured
  - Through linkage to health records and disease registries
  - Repeated assessment of baseline measures (N=20,000)
- GSK/RGC funded exome sequencing

[i] NB: anonymised data shared with scientists

Genomics in drug discovery 12/07/17
How do we use this?

New targets and understanding existing drugs

GLP1R-agonists

- Effective and becoming widely used for T2D
- Cardiovascular risk?
- We found a (rare) variant that mimics drug mechanism of action

<table>
<thead>
<tr>
<th>Disease outcome</th>
<th>N cases</th>
<th>N controls</th>
<th>OR (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 2 diabetes</td>
<td>25,868</td>
<td>122,393</td>
<td>0.83 (0.76, 0.91)</td>
<td>9.4x10⁻⁵</td>
</tr>
<tr>
<td>CHD</td>
<td>61,846</td>
<td>163,728</td>
<td>0.93 (0.87, 0.98)</td>
<td>0.009</td>
</tr>
<tr>
<td>Pancreatic cancer</td>
<td>4987</td>
<td>8627</td>
<td>1.15 (0.82, 1.61)</td>
<td>0.43</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>1879</td>
<td>5118</td>
<td>0.98 (0.73, 1.31)</td>
<td>0.92</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>5157</td>
<td>4838</td>
<td>0.88 (0.70, 1.11)</td>
<td>0.28</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>3937</td>
<td>4423</td>
<td>1.16 (0.91, 1.48)</td>
<td>0.25</td>
</tr>
<tr>
<td>Parkinson's disease</td>
<td></td>
<td></td>
<td>1.07 (0.80, 1.43)</td>
<td>0.65</td>
</tr>
<tr>
<td>Alzheimer's disease</td>
<td></td>
<td></td>
<td>0.94 (0.81, 1.09)</td>
<td>0.40</td>
</tr>
</tbody>
</table>

OR per minor-allele
UK Biobank transforms genetic research

Can study 2000 phenotypes v 1m SNPS in 500k individuals.....

18 months and >1000 emails

1.5 hours and 3 emails

<table>
<thead>
<tr>
<th>Disease outcome</th>
<th>N cases</th>
<th>N controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 2 diabetes</td>
<td>25,668</td>
<td>122,393</td>
</tr>
<tr>
<td>CHD</td>
<td>61,846</td>
<td>163,728</td>
</tr>
<tr>
<td>Pancreatic cancer</td>
<td>4987</td>
<td>8627</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>1879</td>
<td>5118</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>5157</td>
<td>4838</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>3937</td>
<td>4423</td>
</tr>
<tr>
<td>Parkinson's disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alzheimer's disease</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N cases</th>
<th>N controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>5403</td>
<td>126,427</td>
</tr>
<tr>
<td>5134</td>
<td>128,260</td>
</tr>
<tr>
<td>174</td>
<td>133,220</td>
</tr>
<tr>
<td>311</td>
<td>133,083</td>
</tr>
<tr>
<td>3433</td>
<td>129,961</td>
</tr>
<tr>
<td>2383</td>
<td>131,011</td>
</tr>
<tr>
<td>402</td>
<td>132,992</td>
</tr>
<tr>
<td>162</td>
<td>133,232</td>
</tr>
</tbody>
</table>

OR per minor-allele
Genomics England: 100K genomes (RD, oncology)

Objectives: NHS transformation + science + drive biotech

Oversight:
- Department of Health

Funding:
- National Institute for Health Research
- Wellcome Trust
- MRC Medical Research Council
- Cancer Research UK

NHS Genomic Medicine Centres
- Clinical samples and hospital data
- Laboratory processing including molecular pathology
- Broad consent for research and re-contact

Biorepository
- DNA & samples for multi-omics

Sequencing
- Illumina

Participants
- NHS England

Clinical Data
- Identifiable clinical data
- Longitudinal
- Linked to genomic data

Research Data
- Pseudonymised
- GeCIP and industry partners work within data centre

Data

Existing Clinical Data
- Cancer & RD registries, HES, Mortality data, etc

Data and Analysis Improvement
- Annotation & QC
- Scientists/SMEs
- Product comparison

Fire wall

Clinicians & Academics

Training

Industry

NB: only summaries can be extracted from the database
Score card

Starting with the genetics

- Lots of associations
- May not be the most relevant traits
- Mechanism often uncertain
- EHR + genetics will be key

Model systems

- Many cell / tissue models available
- How do we validate? Understanding sources of variability?
- What can animal models tell us?
- Gene editing will be transformative
  - But not as easy as you might think.

Tractability

- Most targets are not tractable
- Pathways?
- New modalities?
Acknowledgements

Material stolen from:

- Lon Cardon
- Ricardo Macarron
- Philippe Sanseau
- Matt Nelson
- Mandy Bergquist
- Deepak Rajpal
- Chris Larminie
- Heather Madsen
- Chun-Fang Xu
- Jo Betts
- Enrico Ferrero
- Dan Felitsky
- Dave Cooper
Drug development
Stratified medicine and biological understanding

Pros:
- Genetic variants affecting safety/efficacy exist
- We expect 10% of drugs to have 'detectable' genetic predictors of efficacy
- We do PGx routinely in development

Cons:
- Trial programs are underpowered for PGx
- Very unlikely that genetics/genomics will rescue failed trials

Future:
- EHR/registries + biobanks
- But better to start in the right place
Random forests

Make many decision trees and average predictors

- Random Forest uses two types of randomness:
  - Each tree uses a different bootstrap sample of observations (bagging)
  - Each split of each tree considers only a random subset of the predictors

- RF is an ensemble of predictors (trees)
Genetic screens of immune cell function (Dan Gaffney, Andrew Powell)

Genome-wide lentiviral KO library

iPS derived macrophages phagocytosis

Sequencing read-out

Figure 1: Forward genetic screens in CTTV with IPS3 derived macrophages.
Genome-scale RNAi and CRISPR screens in a large number of cell lines

Extensive molecular characterization of cell lines

Dependencies

Pathway Activation

Mutational analysis

Copy number

Gene expression

Modified from: http://slideplayer.com/slide/4896413/
**shRNA Loss of Function Screens**

**Infect** a cell line with the shRNA library

*shRNA Library*

58K and 98K libraries ~ 5 motifs per gene

*Cancer Cell Line:*
(Each cell in the population gets one shRNA motif)

(40 days or 16 doublings)

**Measure Essentiality**
(proportion of shRNAs present at the end of the experiment compared to the start)

**Readout** - Next Generation Sequencing
(previously array-based hybridization) of shRNA barcode sequences to get counts per shRNA, per cell line

Validate targets with cellular models – initial studies with relevant cell lines -  WENSHENG XIE

TGF-β to induce fibrosis and inflammation

LX-2 cells → Fibrosis: Collagen Iα

Inflammation: cytokines

HuH-7 cells → Lipid accumulation

Free fatty acid treatment

34 Targets

Over-exp
siRNA
CRISPR

LX2, No TGF-β
LX2, 1ng/mL TGF-β

HuH-7, No FFA
HuH-7, 0.4mM FFA

Collagen1A (red) staining in LX2 cells
Lipid droplets (green) in HuH-7 cells

Target prioritization

Genetics
In vivo data
Biology
Caveat: association to gene not easy

Example: FTO

- Location of association is FTO
- But chromatin interactions span a broader region
- And gene expression//knockdown suggests IRX3 and IRX5