Recap of previous lectures

Sequence: understand target function using protein sequence. NLP to find targets (word sequence).
Graph: generate compound graph 2D structure (deep generative model)
Structure: modify structure according to 3D structure (geometric deep learning)
Genomics: side effects, personalized efficacy, repurposing, etc. (multi-modality)
• how to reuse an old drug
Precision medicine: the right patient, the right drug, the right time, the right dose

We don’t have so many “drugs”

- Discovery new drug?
  - Often not in the scope of precision medicine
  - New patient cannot wait for a new drug
- Drug repurposing
  - Drug A, which is used to treat disease X, is later used to treat disease Y
  - Well-documented side effects and less restriction from FDA
- Drug combination
  - Drug A is not effective. Drug B is not effective. Drug A and B used together is effective.
- Personalized dosage
  - Widely used in clinics. Use genomics data to determine dosage (regression).
### Table 1: Selected successful drug repurposing examples and the repurposing approach employed

<table>
<thead>
<tr>
<th>Drug name</th>
<th>Original indication</th>
<th>New indication</th>
<th>Date of approval</th>
<th>Repurposing approach used</th>
<th>Comments on outcome of repurposing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zidovudine</td>
<td>Cancer</td>
<td>HIV/AIDS</td>
<td>1987</td>
<td>In vitro screening of compound libraries</td>
<td>Zidovudine was the first anti-HIV drug to be approved by the FDA</td>
</tr>
<tr>
<td>Minoxidil</td>
<td>Hypertension</td>
<td>Hair loss</td>
<td>1988</td>
<td>Retrospective clinical analysis (identification of hair growth as an adverse effect)</td>
<td>Global sales for minoxidil were US$860 million in 2016 (Otsuka minoxidil sales report 2017; see Related links)</td>
</tr>
<tr>
<td>Sildenafil</td>
<td>Angina</td>
<td>Erectile dysfunction</td>
<td>1998</td>
<td>Retrospective clinical analysis</td>
<td>Marketed as Viagra, sildenafil became the leading product in the erectile dysfunction drug market, with global sales in 2012 of $20.5 billion</td>
</tr>
<tr>
<td>Thalidomide</td>
<td>Morning sickness</td>
<td>Erythema nodosum leprosum and multiple myeloma</td>
<td>1998 and 2006</td>
<td>Off-label usage and pharmacological analysis</td>
<td>Thalidomide derivatives have achieved substantial clinical and commercial success in multiple myeloma</td>
</tr>
<tr>
<td>Celecoxib</td>
<td>Pain and inflammation</td>
<td>Familial adenomatosus polyposis</td>
<td>2000</td>
<td>Pharmacological analysis</td>
<td>The total revenue from Celebrex (Pfizer) at the end of 2014 was $2.69 billion (Pfizer 2014 financial report; see Related links)</td>
</tr>
<tr>
<td>Acometaxine</td>
<td>Parkinson disease</td>
<td>ADHD</td>
<td>2002</td>
<td>Pharmacological analysis</td>
<td>Strattera (Eli Lilly) recorded global sales of $855 million in 2016</td>
</tr>
<tr>
<td>Duloxetine</td>
<td>Depression</td>
<td>SUI</td>
<td>2004</td>
<td>Pharmacological analysis</td>
<td>Approved by the EMA for SUI. The application was withdrawn in the US. Duloxetine is approved for the treatment of depression and chronic pain in the US</td>
</tr>
<tr>
<td>Raloxifene</td>
<td>Osteoporosis</td>
<td>Breast cancer</td>
<td>2007</td>
<td>Retrospective clinical analysis</td>
<td>Approved by the FDA for invasive breast cancer. Worldwide sales of $2.37 billion in 2015 (see Related links)</td>
</tr>
<tr>
<td>Fingelomod</td>
<td>Transplant rejection</td>
<td>MS</td>
<td>2010</td>
<td>Pharmacological and structural analysis</td>
<td>First oral disease-modifying therapy to be approved for MS. Global sales for fingelomod (Gilenya) reached $3.1 billion in 2017 (see Related links)</td>
</tr>
<tr>
<td>Dapoxetine</td>
<td>Analgesia and depression</td>
<td>Premature ejaculation</td>
<td>2012</td>
<td>Pharmacological analysis</td>
<td>Approved in the UK and a number of European countries; still awaiting approval in the US. Peak sales are projected to reach $750 million</td>
</tr>
<tr>
<td>Topiramate</td>
<td>Epilepsy</td>
<td>Obesity</td>
<td>2012</td>
<td>Pharmacological analysis</td>
<td>Osymia (Vivus) contains topiramate in combination with phentermine</td>
</tr>
<tr>
<td>Ketoconazole</td>
<td>Fungal infections</td>
<td>Cushing syndrome</td>
<td>2014</td>
<td>Pharmacological analysis</td>
<td>Approved by the EMA for Cushing syndrome in adults and adolescents above the age of 17 years (see Related links)</td>
</tr>
<tr>
<td>Aspirin</td>
<td>Analgesia</td>
<td>Colorectal cancer</td>
<td>2015</td>
<td>Retrospective clinical and pharmacological analysis</td>
<td>US Preventive Services Task Force released draft recommendations in September 2015 regarding the use of aspirin to help prevent cardiovascular disease and colorectal cancer</td>
</tr>
</tbody>
</table>

---

**Pushpakom et al. Drug repurposing: progress, challenges and recommendations**
Approaches used in drug repurposing

- Genetic association: Genes that are associated with a disease may prove to be potential drug targets.
- Pathway mapping: Network analysis using genetic, protein or disease data can aid identification of repurposing targets.
- Molecular docking: This is a structure-based computational strategy to predict binding site complementarity between a ligand (for example, a drug) and a therapeutic target (typically a protein).
- Retrospective clinical analysis: Systematic analysis of EHRs, clinical trial data and post-marketing surveillance data could inform drug repurposing.
- Signature matching: This involves comparing the 'signature' of a drug — characteristics such as its transcriptomic, structural or adverse effect profile — with that of another drug or disease phenotype.
- Novel data sources: Large-scale in vitro drug screens with paired genomic data, EHR-linked large biobanks and self-reported patient data are novel avenues to exploit for drug repurposing.
- Binding assays to identify relevant target interactions: Techniques such as affinity chromatography and mass spectrometry can be used to identify novel targets of known drugs.
- Phenotypic screening: High-throughput phenotypic screening of compounds using in vitro or in vivo disease models can indicate potential for clinical evaluation.

Gene target
3D structure
Gene expression
2D graph
EHR

Computational approaches
Experimental approaches
Drug repurposing strategy

- **Drug-based**
  - If drug A can cure disease X and is similar to drug B, then B might be also treat X

- **Disease-based**
  - If disease X and Y have similar profiles and indications, and drug R can cure X, then R can also cure Y.
Use gene expression after treatment

Drugs target on similar proteins or have similar Mode of Actions have similar (after treatment) expression.

Iorio et al. Discovery of drug mode of action and drug repositioning from transcriptional responses
Compare disease expression and drug expression

Sirota et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data
Expression-based drug repurposing

- People realized that the performance (accuracy, coverage) depends on the data, rather than the model
- How about we just generate the expression of X drugs on Y tissues
  - LINCS: Library of Integrated Network-based Cellular Signatures
  - 15 institutions, >1000 cell lines, >5000 drugs, 1000 genes
  - 1.3 million after treatment gene expression vectors
  - cMAP: 3 cell lines, but 20k genes
**Adverse drug reaction prediction** (Wang et al. Drug-induced adverse events prediction with the LINCS L1000 data)

**Drug target identification** (Xia et al. Target Predictions using LINCS Data)

**Expression signature comparison** (Xiao et al. SigMat: A Classification Scheme for Gene Signature Matching)

**Drug response prediction** (Lu et al. Drug-induced cell viability prediction from LINCS-L1000 through WRFEN-XGBoost algorithm)
We don’t have so many “drugs”

• Discovery new drug?
  - Often not in the scope of precision medicine
  - New patient cannot wait for a new drug

• Drug repurposing
  - Drug A, which is used to treat disease X, is later used to treat disease Y
  - Well-documented side effects and less restriction from FDA

• Drug combination
  - Drug A is not effective. Drug B is not effective. Drug A and B used together is effective.

• Personalized dosage
  - Widely used in clinics. Use genomics data to determine dosage (regression).
Synthetic lethality: Gene A or Gene B

Question: how to leverage SL in drug combination discovery?

Source: wikipedia, Jerby-Aron et al. Predicting Cancer-Specific Vulnerability via Data-Driven Detection of Synthetic Lethality
Drug combination therapy

- **Breast cancer**
  - an alkylating agent (cyclophosphamide) and antimetabolites (methotrexate and 5-fluorouracil)

- **Anti-HIV cocktail**
  - Use of three or more antiretroviral medicines

- We don’t have so many single drug candidate

- Drug combinations ($k \geq 2$) offer us more treatment plans
Drug treatment based on synthetic lethality

Goal: We want to make normal cells survive and kill cancer cells (BRCA deficient cancer cells)

Prior knowledge: PARP1 (off) + BRCA1 (off) -> cell death

Solution: Turn off PARP1 using Olaparib

Results:
- Normal cells: PARP1 (off) + BRCA1 (on) -> cell survive
- Cancer cells: PARP1 (off) + BRCA1 (off) -> cell death

Gilad et al. Drug Combination in Cancer Treatment—From Cocktails to Conjugated Combinations
Drug combination prediction

$E_{A,B} < E_A + E_B$

$E_{A,B} = E_A + E_B$

$E_{A,B} > E_A + E_B$

E(A) is the efficacy of using drug A (e.g., IC50)
Dose-response curve

IC(50): concentration of 50% viability
Drug combination dose-response curve

1D Dose-Response Curve

Proliferation Rate

0

Potent Drug

drug1

Efficacious Drug

drug2

Tumor Growth

Tumor Regression

2D Dose-Response Surface

Proliferation Rate

log(d2)

log(d1)
DeepSynergy: deep learning-based drug synergistic prediction

Table 3. Performance metrics for the classification task

<table>
<thead>
<tr>
<th>Performance Metric</th>
<th>ROC AUC</th>
<th>PR AUC</th>
<th>ACC</th>
<th>BACC</th>
<th>PREC</th>
<th>TPR</th>
<th>TNR</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Neural Networks</td>
<td>0.90 ± 0.03</td>
<td>0.59 ± 0.06</td>
<td>0.92 ± 0.03</td>
<td>0.76 ± 0.03</td>
<td>0.56 ± 0.11</td>
<td>0.57 ± 0.09</td>
<td>0.95 ± 0.03</td>
<td>0.51 ± 0.04</td>
</tr>
<tr>
<td>Gradient Boosting Machines</td>
<td>0.89 ± 0.02</td>
<td>0.59 ± 0.04</td>
<td>0.87 ± 0.01</td>
<td>0.80 ± 0.03</td>
<td>0.38 ± 0.04</td>
<td>0.71 ± 0.05</td>
<td>0.89 ± 0.01</td>
<td>0.43 ± 0.03</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.87 ± 0.02</td>
<td>0.55 ± 0.04</td>
<td>0.92 ± 0.01</td>
<td>0.73 ± 0.04</td>
<td>0.57 ± 0.04</td>
<td>0.49 ± 0.08</td>
<td>0.96 ± 0.01</td>
<td>0.48 ± 0.04</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>0.81 ± 0.04</td>
<td>0.42 ± 0.08</td>
<td>0.76 ± 0.06</td>
<td>0.73 ± 0.03</td>
<td>0.23 ± 0.04</td>
<td>0.69 ± 0.08</td>
<td>0.77 ± 0.07</td>
<td>0.24 ± 0.05</td>
</tr>
<tr>
<td>Elastic Nets</td>
<td>0.78 ± 0.04</td>
<td>0.34 ± 0.10</td>
<td>0.75 ± 0.05</td>
<td>0.71 ± 0.02</td>
<td>0.21 ± 0.03</td>
<td>0.65 ± 0.07</td>
<td>0.76 ± 0.06</td>
<td>0.22 ± 0.03</td>
</tr>
<tr>
<td>Baseline (Median Polish)</td>
<td>0.77 ± 0.04</td>
<td>0.32 ± 0.09</td>
<td>0.76 ± 0.04</td>
<td>0.70 ± 0.03</td>
<td>0.22 ± 0.03</td>
<td>0.62 ± 0.06</td>
<td>0.78 ± 0.04</td>
<td>0.22 ± 0.04</td>
</tr>
</tbody>
</table>
Problem setting

Menden et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen.
Use gene expression after treatment

Drugs target on similar proteins or have similar Mode of Actions have similar (after treatment) expression.

Iorio et al. Discovery of drug mode of action and drug repositioning from transcriptional responses
TAIJI: simulate post treatment expression

Li et al. Network Propagation Predicts Drug Synergy in Cancers
Drug combinations for treating COVID-19

# Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data

## Table 1. Demographic and Clinical Characteristics of the Derivation and Validation Cohorts.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Derivation Cohort (N = 4043)</th>
<th>Validation Cohort (N = 1099)</th>
<th>P Value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Warfarin dose — mg/wk</td>
<td></td>
<td></td>
<td>0.40</td>
</tr>
<tr>
<td>Median</td>
<td>28.0</td>
<td>28.0</td>
<td></td>
</tr>
<tr>
<td>Interquartile range</td>
<td>19.0–38.5</td>
<td>21.0–38.5</td>
<td></td>
</tr>
<tr>
<td>Genotype — no. (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VKORC1 rs9923231</td>
<td></td>
<td></td>
<td>0.97</td>
</tr>
<tr>
<td>G/G</td>
<td>1201 (29.7)</td>
<td>302 (29.9)</td>
<td></td>
</tr>
<tr>
<td>A/G</td>
<td>1444 (35.7)</td>
<td>363 (36.0)</td>
<td></td>
</tr>
<tr>
<td>A/A</td>
<td>1315 (32.5)</td>
<td>326 (32.3)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>83 (2.1)</td>
<td>18 (1.8)</td>
<td></td>
</tr>
<tr>
<td>CYP2C9†</td>
<td></td>
<td></td>
<td>0.38</td>
</tr>
<tr>
<td>*1/*1</td>
<td>2970 (73.5)</td>
<td>749 (74.2)</td>
<td></td>
</tr>
<tr>
<td>*1/*2</td>
<td>509 (12.6)</td>
<td>142 (14.1)</td>
<td></td>
</tr>
<tr>
<td>*1/*3</td>
<td>374 (9.3)</td>
<td>76 (7.5)</td>
<td></td>
</tr>
<tr>
<td>*2/*2</td>
<td>36 (0.9)</td>
<td>10 (1.0)</td>
<td></td>
</tr>
<tr>
<td>*2/*3</td>
<td>52 (1.3)</td>
<td>10 (1.0)</td>
<td></td>
</tr>
<tr>
<td>*3/*3</td>
<td>15 (0.4)</td>
<td>1 (0.1)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>87 (2.2)</td>
<td>21 (2.1)</td>
<td></td>
</tr>
<tr>
<td>Age — no. (%)</td>
<td></td>
<td></td>
<td>0.88</td>
</tr>
<tr>
<td>10–19 yr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20–29 yr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30–39 yr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40–49 yr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50–59 yr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60–69 yr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70–79 yr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80–89 yr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥90 yr</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Table 2. Predicted Warfarin Doses with the Pharmacogenetic Algorithm, Clinical Algorithm, and Fixed-Dose Approach as Compared with the Actual Stable Dose in the Derivation and Validation Cohorts.*

<table>
<thead>
<tr>
<th>Prediction Model</th>
<th>Derivation Cohort</th>
<th>Validation Cohort†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Absolute Error (95% CI)</td>
<td>R²</td>
</tr>
<tr>
<td></td>
<td>mg/wk</td>
<td>%</td>
</tr>
<tr>
<td>Pharmacogenetic algorithm‡‡</td>
<td>8.3 (8.1–8.6)</td>
<td>47</td>
</tr>
<tr>
<td>Clinical algorithm‡</td>
<td>10.0 (9.7–10.3)</td>
<td>27</td>
</tr>
<tr>
<td>Fixed-dose approach¶</td>
<td>13.3 (13.0–13.5)</td>
<td>0</td>
</tr>
</tbody>
</table>

The International Warfarin Pharmacogenetics Consortium. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data
### Example of precision medicine

<table>
<thead>
<tr>
<th>Condition</th>
<th>Gene</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mendelian disease</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cystic fibrosis</td>
<td>CFTR</td>
<td>Specific therapies such as ivacaftor and a combination of lumacaftor and ivacaftor</td>
</tr>
<tr>
<td>Long QT syndrome</td>
<td>KCNQ1, KCNH2 and SCN5A</td>
<td>Specific therapy for patients with SCN5A mutations</td>
</tr>
<tr>
<td>Duchenne muscular dystrophy</td>
<td>DMD</td>
<td>Ongoing phase III clinical trials of exon-skipping therapies</td>
</tr>
<tr>
<td>Malignant hyperthermia susceptibility</td>
<td>RYR1</td>
<td>Avoid volatile anaesthetic agents; avoid extremes of heat</td>
</tr>
<tr>
<td>Familial hypercholesterolaemia (FH)</td>
<td>PCSK9, APOB and LDLR</td>
<td>Heterozygous FH (HeFH): eligible for PCSK9 inhibitor drugs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Homozygous FH (HoFH): eligible for PCSK9 inhibitor drugs in addition to lomitapide and mipomersen</td>
</tr>
<tr>
<td>Dopa-responsive dystonia</td>
<td>SPR</td>
<td>Therapy with dopamine precursor l-dopa and the serotonin precursor 5-hydroxytryptophan</td>
</tr>
<tr>
<td>Thoracic aortic aneurysm</td>
<td>SMAD3, ACTA2, TGFBR1, TGFBR2 and FBN1</td>
<td>Customization of surgical thresholds based on patient genotype</td>
</tr>
<tr>
<td>Left ventricular hypertrophy</td>
<td>MYH7, MYBPC3, GLA and TTR</td>
<td>Sarcomeric cardiomyopathy, Fabry disease and transthyretin cardiac amyloid disease have specific therapies</td>
</tr>
<tr>
<td><strong>Precision oncology</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung adenocarcinoma</td>
<td>EGFR and ALK</td>
<td>Targeted kinase inhibitors, such as gefitinib and crizotinib</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>HER2</td>
<td>HER2 (also known as ERBB2)-targeted treatment, such as trastuzumab and pertuzumab</td>
</tr>
<tr>
<td>Gastrointestinal stromal tumour</td>
<td>KIT</td>
<td>Targeted KIT kinase activity inhibitors, such as imatinib</td>
</tr>
<tr>
<td>Melanoma</td>
<td>BRAF</td>
<td>BRAF inhibitors, such as vemurafenib and dabrafenib</td>
</tr>
<tr>
<td><strong>Pharmacogenomics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Warfarin sensitivity</td>
<td>CYP2C9 and VKORC1</td>
<td>Adjust dosage of warfarin or consider alternative anticoagulant</td>
</tr>
<tr>
<td>Clopidogrel sensitivity, post-stent procedure</td>
<td>CYP2C19</td>
<td>Consider alternative antiplatelet therapy (for example, prasugrel or ticagrelor)</td>
</tr>
<tr>
<td>Thiopurine sensitivity</td>
<td>TPMT</td>
<td>Reduce thiopurine dosage or consider alternative agent</td>
</tr>
<tr>
<td>Codeine sensitivity</td>
<td>CYP2D6</td>
<td>Avoid use of codeine; consider alternatives such as morphine and non-opioid analgesics</td>
</tr>
<tr>
<td>Simvastatin sensitivity</td>
<td>SLCO1B1</td>
<td>Reduce dose of simvastatin or consider an alternative statin; consider routine creatine kinase surveillance</td>
</tr>
</tbody>
</table>

Euan A. Ashley. Towards precision medicine.
Two key problems

- How to cluster patients
- We don’t have so many “drugs”
How to cluster patients

- Patient clustering = data integration
  - Find a “signature” vector for each patient
  - Signature is integrated from different data sources

- Heterogeneous data integration
  - General challenges: Heterogenous, missing values, noise, privacy

- Precision medicine specific data integration challenges:
  - Batch effects (different preprocessing pipelines, sequencing techniques, reference ranges)
  - Unpaired data (some patients only have genomics data, some patients only have EHR data, very few patients have both)
## Public biomedical databases

<table>
<thead>
<tr>
<th>DATA REPOSITORY</th>
<th>WEB LINK</th>
<th>DISEASE</th>
<th>TYPES OF MULTI-OMICS DATA AVAILABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Cancer Genome Atlas (TCGA)</td>
<td><a href="https://cancergenome.nih.gov/">https://cancergenome.nih.gov/</a></td>
<td>Cancer</td>
<td>RNA-Seq, DNA-Seq, miRNA-Seq, SNV, CNV, DNA methylation, and RPPA</td>
</tr>
<tr>
<td>International Cancer Genomics Consortium (ICGC)</td>
<td><a href="https://icgc.org/">https://icgc.org/</a></td>
<td>Cancer</td>
<td>Whole genome sequencing, genomic variations data (somatic and germline mutation)</td>
</tr>
<tr>
<td>Cancer Cell Line Encyclopedia (CCLE)</td>
<td><a href="https://portals.broadinstitute.org/ccle">https://portals.broadinstitute.org/ccle</a></td>
<td>Cancer cell line</td>
<td>Gene expression, copy number, and sequencing data; pharmacological profiles of 24 anticancer drugs</td>
</tr>
<tr>
<td>Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)</td>
<td><a href="http://molonc.bccrc.ca/aparicio-lab/research/metabric/">http://molonc.bccrc.ca/aparicio-lab/research/metabric/</a></td>
<td>Breast cancer</td>
<td>Clinical traits, gene expression, SNP, and CNV</td>
</tr>
<tr>
<td>TARGET</td>
<td><a href="https://ocg.cancer.gov/programs/target">https://ocg.cancer.gov/programs/target</a></td>
<td>Pediatric cancers</td>
<td>Gene expression, miRNA expression, copy number, and sequencing data</td>
</tr>
<tr>
<td>Omics Discovery Index</td>
<td><a href="https://www.omicsdi.org">https://www.omicsdi.org</a></td>
<td>Consolidated data sets from 11 repositories in a uniform framework</td>
<td>Genomics, transcriptomics, proteomics, and metabolomics</td>
</tr>
</tbody>
</table>

Subramanian et al. Multi-omics Data Integration, Interpretation, and Its Application
Personalized drug response prediction: multi-label regression problem

CCLE data: ~1000 cell lines, 20k genes, 400 drugs

Three settings

- Test patients: no drugs are observed for this patient
- Test drugs: no patients are observed for this drug
- Test <patient, drug> pairs
Cell line, xenograft, tumor, patient

- Cell line is a “copy” of a patient. We cannot test one patient with many drugs. But we can copy a cell line many times.
- Cell line is cheaper than xenograft. Xenograft is cheaper than patient data.
- TCGA has some patient data.
- ML question: how to integrate cell line data, xenograft data and patient data.
Batch effects: inconsistency or consistency?

ANALYSIS

doi:10.1038/nature12831

Inconsistency in large pharmacogenomic studies

Benjamin Haibe-Kains¹,², Nehme El-Hachem¹, Nicolai Juul Birkbak³, Andrew C. Jin⁴, Andrew H. Beck⁴*, Hugo J. W. L. Aerts⁵,⁶,⁷* & John Quackenbush⁵,⁸*

Two large-scale pharmacogenomic studies were published recently in this journal. Genomic data are well correlated between studies; however, the measured drug response data are highly discordant. Although the source of inconsistencies remains uncertain, it has potential implications for using these outcome measures to assess gene-drug associations or select potential anticancer drugs on the basis of their reported results.

ANALYSIS

doi:10.1038/nature15736

Pharmacogenomic agreement between two cancer cell line data sets

The Cancer Cell Line Encyclopedia and Genomics of Drug Sensitivity in Cancer Investigators*

Large cancer cell line collections broadly capture the genomic diversity of human cancers and provide valuable insight into anti-cancer drug response. Here we show substantial agreement and biological consilience between drug sensitivity measurements and their associated genomic predictors from two publicly available large-scale pharmacogenomics resources: The Cancer Cell Line Encyclopedia and the Genomics of Drug Sensitivity in Cancer databases.
Low correlation between drug response data

Integrating two datasets
High correlation between gene expression (features)
General framework: jointly decompose multiple data matrices

Key ideas
- One matrix capture batch effects
- One matrix capture common patterns

Detail implementations
- What distribution?
  - Mutation (Bernoulli)
  - Expression count (Poisson)
- How to decompose?
  - NMF, SVD, NN, MF

Argelaguet et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets
Mashup: integrating multiple networks

Cho et al. Compact Integration of Multi-Network Topology for Functional Analysis of Genes
Mashup improves protein function prediction

Protein function prediction is a good benchmark for machine learning algorithms because of it is high-quality and has many annotations. It can be used to evaluate:

- Network-based approach
- Sequence-based approach
- Few-shot/zero-shot learning
Mashup enables genetic interaction prediction

Synthetic Lethality
How to leverage SL to develop (personalized) drug (combination) therapy?

Integrate three sources:
• Mutation data of the patient (mutation A)
• SL network (Gene B has SL effect with Gene A)
• Drug target information (Drug X inhibits Gene B)
How to integrate network with a patient matrix

Patient matrix is very sparse
Use network to smooth it

Hofree et al. Network-based stratification of tumor mutations
How to integrate network with a patient matrix

Project each patient’s vector to the network
Use diffusion model to find overlapping regions

What diffusion models to use?
- Random walk with restart
- Heat diffusion
- Graph neural network (GCN, GAT)
- Network embedding (no node features)
- Other non-euclidean geometry (hyperbolic embedding)

Hofree et al. Network-based stratification of tumor mutations
Results on TCGA: use patient survival data as a benchmark to evaluate patient clustering model.
How to cluster patients

- Patient clustering = data integration
  - Find a “signature” vector for each patient
  - Signature is integrated from different data sources

- Heterogeneous data integration
  - General challenges: Heterogenous, missing values, noise, privacy

- Precision medicine specific data integration challenges:
  - Batch effects (different preprocessing pipelines, sequencing techniques, reference ranges)
  - Unpaired data (some patients only have genomics data, some patients only have EHR data, very few patients have both)

Grapov et al. Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine
How to handle unpaired data

Genomics

Patients with only genomics data

Patients with only EHR data

Patients with both

EHR

ML question: How to integrate all these patients?
Translation between features: generate expression from image

Schmauch et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images
Wu et al. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution.