Current Trends and Opportunities in Computational Biology

Lin Li, PhD
PharmaLex, Frederick, MD, USA

October 20th, 2022

NCS | Non Clinical Statistics Conference
Louvain-la-Neuve, Belgium / October 19-21, 2022

Disclaimer

The views and opinions expressed in this presentation are those of the speaker's and do not necessarily reflect the views or positions of the employer or its affiliates.

NCS | Non Clinical Statistics Conference

Louvain-la-Neuve, Belgium / October 19-21, 2022

Computational biology

- “Computational biology and bioinformatics is an interdisciplinary field that develops and applies computational methods to analyse **large collections of biological data**, such as **genetic sequences, cell populations or protein samples**, to make new predictions or discover new biology. The computational methods used include analytical methods, mathematical modelling and simulation.” https://www.nature.com/subjects/computational-biology-and-bioinformatics

- Omics
  - **Genomics** studies DNA that carries genetic information for the development and functioning of an organism.
  - **Transcriptomics** helps understand which genes are being expressed and which are silent
  - **Proteomics** provides a detailed view of proteins present in blood or tissue
  - …

The 5R framework

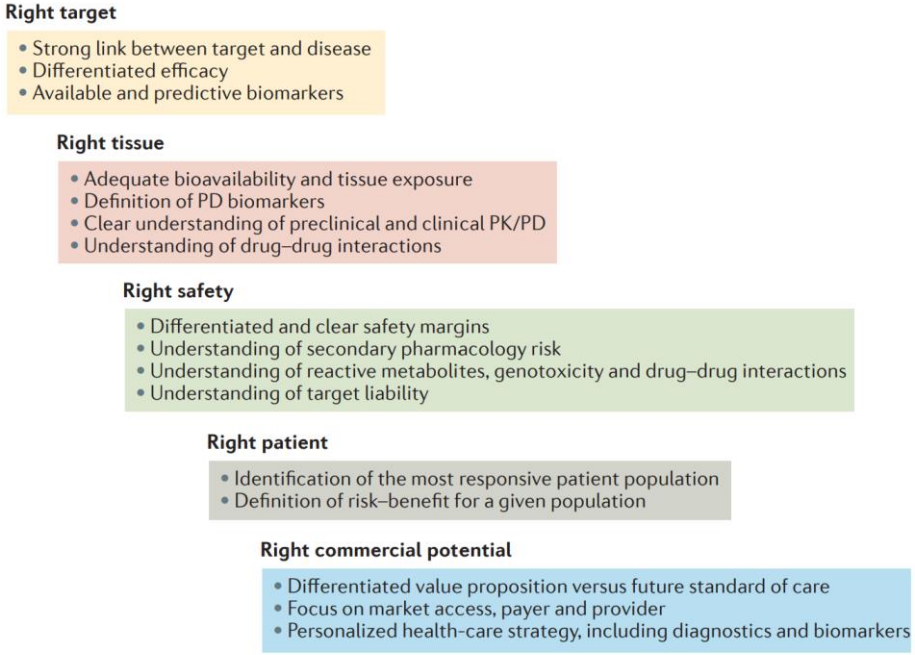
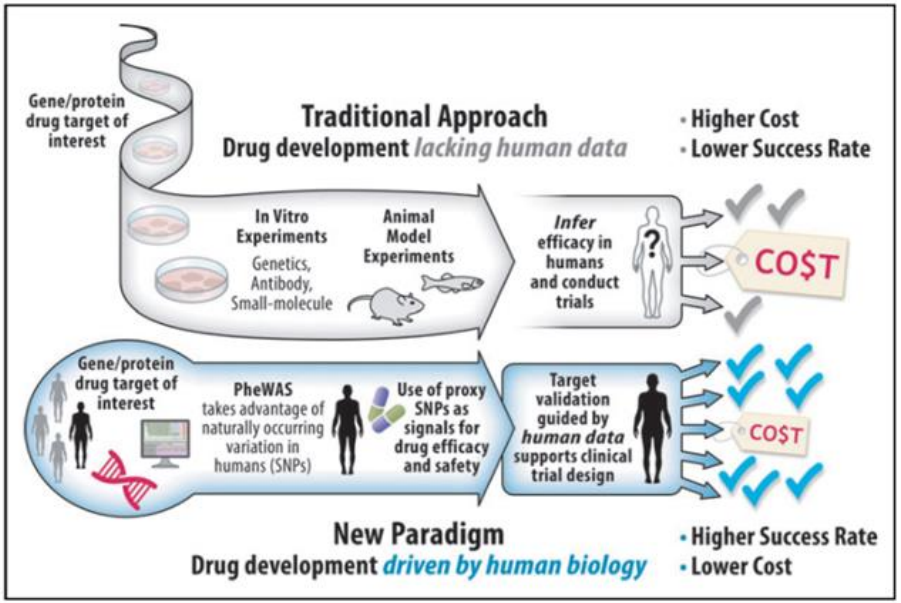


Figure 1 | **The 5R framework.** Summary of the key features of the five-dimensional framework: the right target, right tissue, right safety, right patient and right commercial potential. PD, pharmacodynamic; PK, pharmacokinetic.

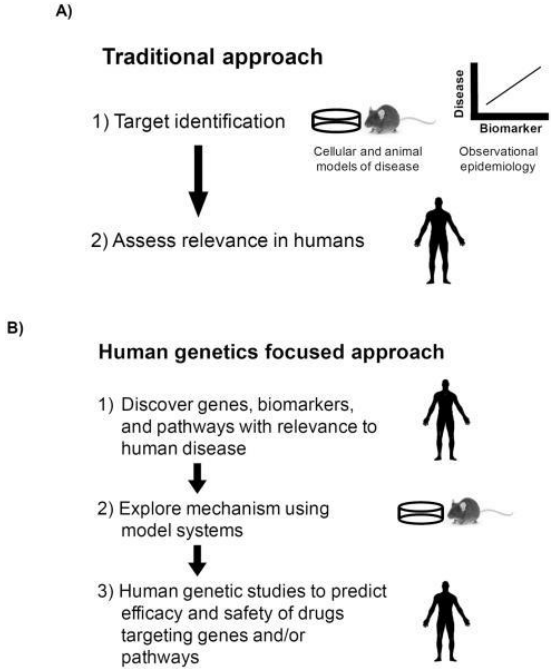
Morgan et al., (2018) *Nature Reviews Drug Discovery*, *17*(3), pp.167-181.

"Overall, the continued evolution and application of the 5R framework are beginning to have an impact, with success rates from candidate drug nomination to phase III completion improving from 4% in 2005–2010 to 19% in 2012–2016."

New paradigm of leveraging human genetics



Pulley et al, (2017) *Assay and Drug Development Technologies*, *15*(3), pp.113-119.



Stitziel and Kathiresan (2017) *Trends in Cardiovascular Medicine*, 27(5), pp.352-359.

The importance of human genetics in drug discovery

- Nelson et al. (2015) estimated that drug mechanisms with genetic support would succeed **twice** as often as those without it (from phase I to approval).

Table 1 The relative value of genetic support for the probability that a target-indication pair progresses along the drug development pipeline, based on historical drug trial information

	p(progress|genetic support)/(progress|no genetic support)		
Progression	GWASdb and OMIM	GWASdb	OMIM
Phase I to phase II	1.2 (1.1–1.3)	1.2 (1.1–1.3)	1.2 (1.1–1.3)
Phase II to phase III	1.5 (1.3–1.7)	1.4 (1.2–1.7)	1.6 (1.3–1.9)
Phase III to approval	1.1 (1.0–1.2)	1.0 (0.8–1.2)	1.1 (0.9–1.3)
Phase I to phase III	1.8 (1.5–2.1)	1.8 (1.4–2.1)	1.9 (1.5–2.3)
Phase I to approval	2.0 (1.6–2.4)	1.8 (1.3–2.3)	2.2 (1.6–2.8)

Values give the ratio of the probability of a target-indication pair progressing given genetic support to the probability of progressing without genetic support; 95% confidence intervals are given in parentheses.

Reference: Nelson et al. (2015), *Nature Genetics*

- Additional work “confirms drugs with genetically supported targets were more likely to be successful in Phases II and III. When causal genes are clear (Mendelian traits and GWAS associations linked to coding variants), we find the use of human genetic evidence increases approval by **greater than two-fold**, and, for Mendelian associations, the positive association holds prospectively.”

King, Davis, and Degner, (2019). *PLoS Genetics*, *15*(12), p.e1008489.

Partnerships around “big data”

- UK Biobank, the large-scale biomedical database and research resource, has been partnering with pharma companies in ambitious projects such as whole genome sequencing of ~500,000 participants as well as generating proteomics and longitudinal imaging data for tens of thousands of participants

- A global pharma company extended exclusive target discovery period of collaboration for a fifth year to discover and validate novel drug targets using a partner company’s proprietary genetic and health survey database

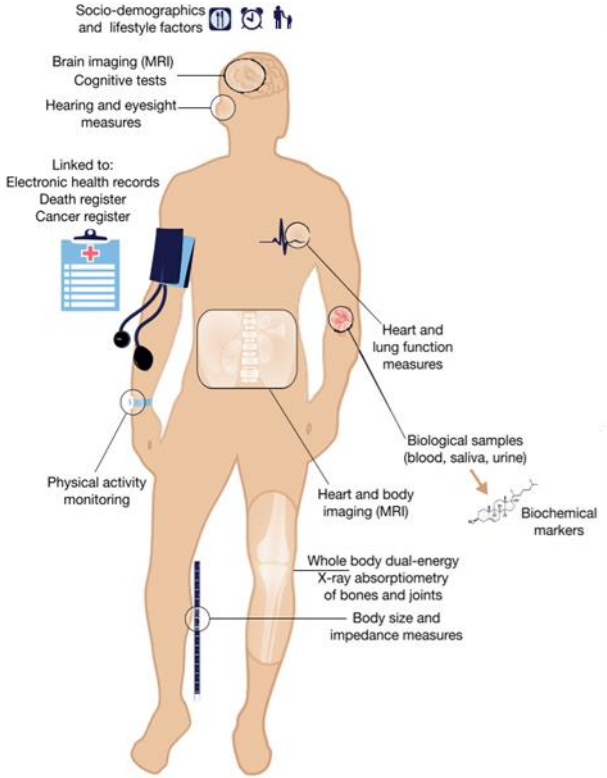
- A healthcare delivery network and pharma collaborated to analyze the genomes of up to 500,000 patients (Pharma company website)

- And more examples…

Overview of the UK Biobank phenotypic data

Extensive phenotypic information with **~8,500** data fields/variables (and growing)

- **~2,500** health related variables
  - First occurrences
  - Hospital inpatient
  - Cancer register
  - Death register
  - Primary care
  - Algorithmically-defined outcomes
- **4,000**+ quantitative traits
  - Imaging – Brain/Heart/Abdominal MRI, DXA
  - Blood biomarkers
  - Blood counts
  - Physical measures
    - Blood pressure, Spirometry, anthropometry, ...
  - Others

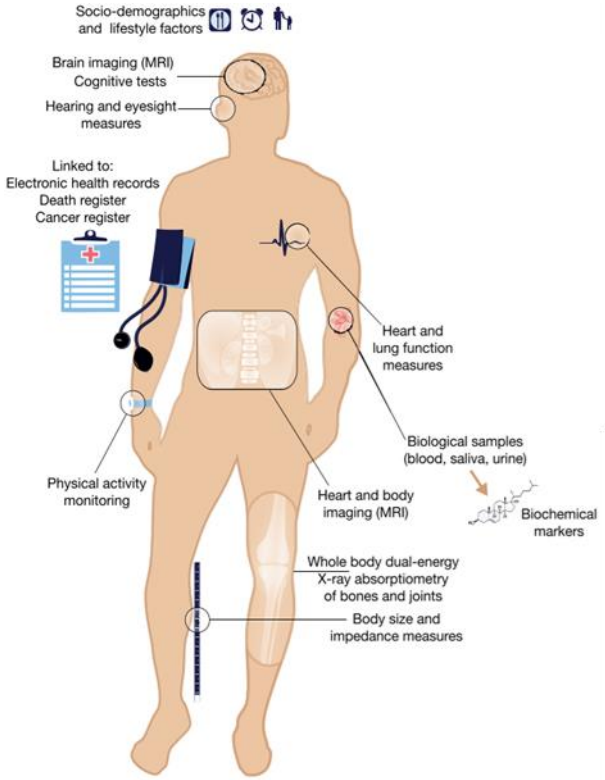


Source: Bycroft *et al. Nature* **562**, 203–209 (2018)

Drug target identification

- What target should be considered for an indication?
  - What genes are causal for a disease?
  - What gene is likely a good target for an indication with a desirable profile of efficacy and safety?

- To answer the questions:
  - **Right cohort and right endpoint**
  - **Association (Genome-wide association study; GWAS)**
  - **Fine-mapping and colocalization**
  - **Causality (Mendelian randomization)**
  - **Potential safety risk (phenome-wide association study; PheWAS)**



Source: Bycroft *et al. Nature* **562**, 203–209 (2018)

Right cohort and right endpoint

- Clinical diagnosis coming with different forms: ICD-9/10, SNOMED codes, self-reported diagnosis codes, operations codes
- Mapping between the types may not be one to one
- Some codes do not properly translate due to how diagnosis was characterized

Define cases

- ICD-9/10 designed mainly for billing purposes
  - Codes must be converted to ICD if the primary diagnosis was made using a different coding system, which can lead to reduced accuracy
- Biologically related diseases may be split into different groups based on affected organs rather than etiology
- ICD-9/10 often works well for defining cases, but control groups may be contaminated by similar diseases in other code groups

Defining Controls

- "Super" healthy controls
- "All-else“ controls
- Case-control matched
- Phecode controls
- …

An example of computational framework

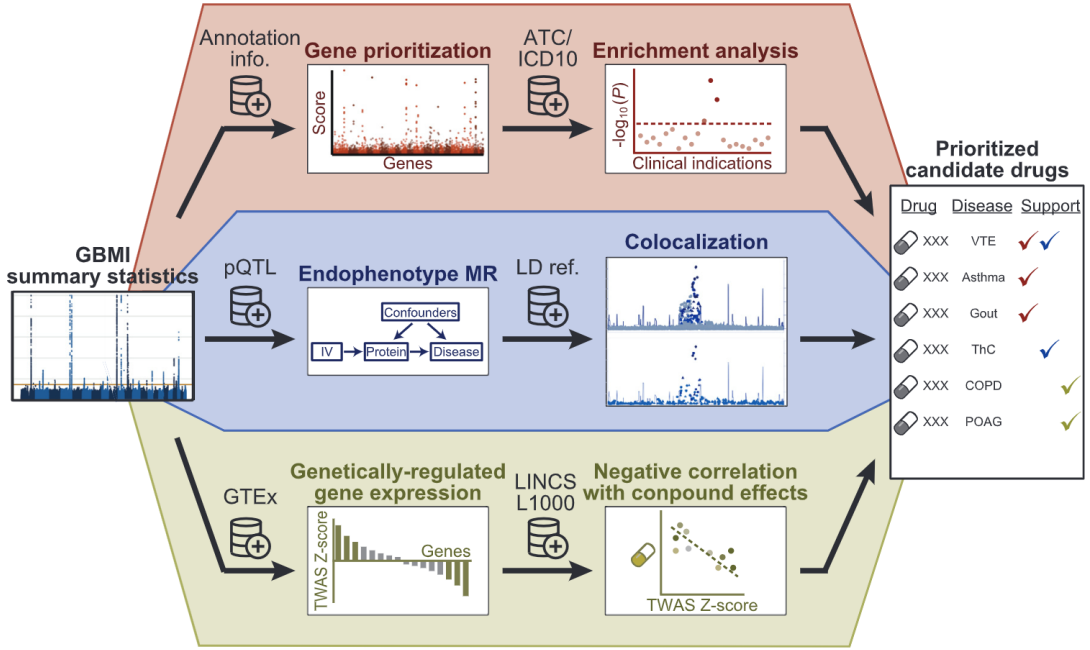
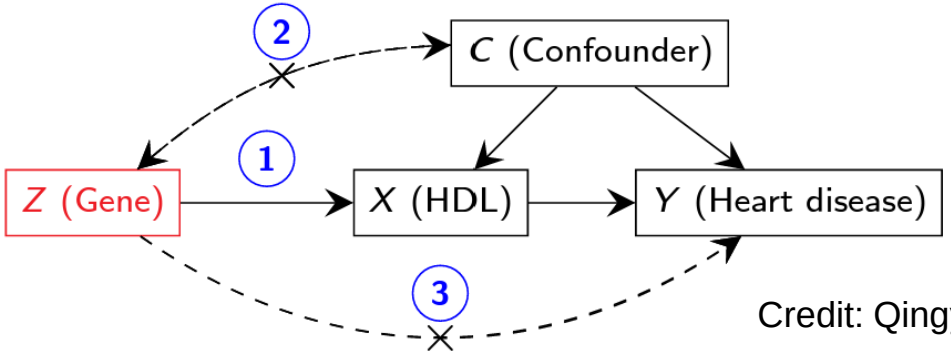


Figure 1. Overview of the genomics-driven drug discovery framework

Namba et al. (2022) *Cell Genomics*, 2(10), p.100190.

Mendelian randomization

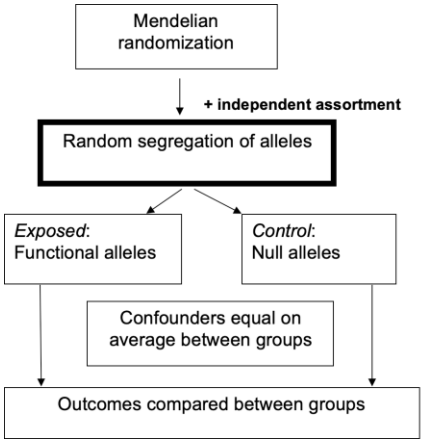


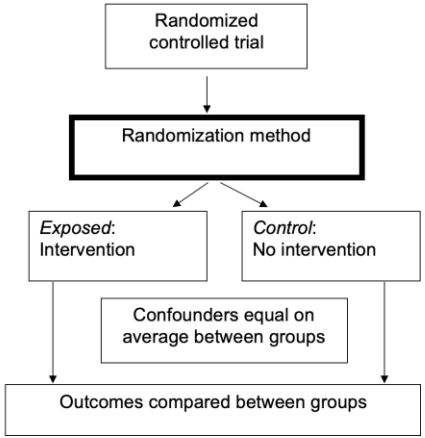
Credit: Qingyuan Zhao

Core IV assumptions

- **Relevance**: Gene (Z) is associated with the exposure (X)
- **Effective random assignment**: Gene (Z) is independent of the unmeasured confounder (C)
- **Exclusion restriction**: Gene (Z) cannot have any direct effect on the outcome (Y)

Mendelian randomization (cont'd)





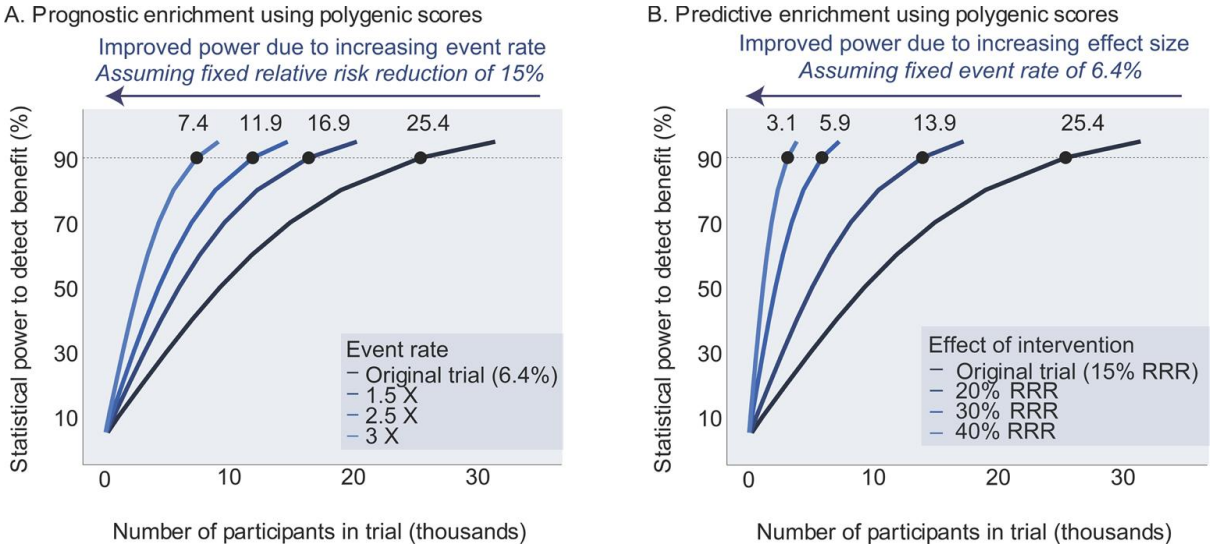
Ten simple rules for conducting a mendelian randomization study | PLOS Computational Biology

Rule 1: Have a clear research question
Rule 2: Keep in mind the core IV assumptions
Rule 3: Be attentive when selecting genetic variants to be used as instruments
Rule 4: Consider the possibility of reverse causality
Rule 5: Understand the pros and cons of using one- versus two-sample MR
Rule 6: Visualize results
Rule 7: Run sensitivity analyses to increase confidence in the validity of the results
Rule 8: Document code and ensure reproducibility
Rule 9: Carefully interpret results and acknowledge limitations
Rule 10: Disseminate findings to the research community

Gagliano Taliun and Evans (2021) *PLoS Computational Biology*, 17(8), p.e1009238.

Polygenic score

- A polygenic score (PGS) is a numeric value that summarizes the estimated effect of multiple genetic variants, usually in different genes, on an individual's phenotype.
- PGS has the potential to enable predictive or prognostic enrichment in clinical trials.

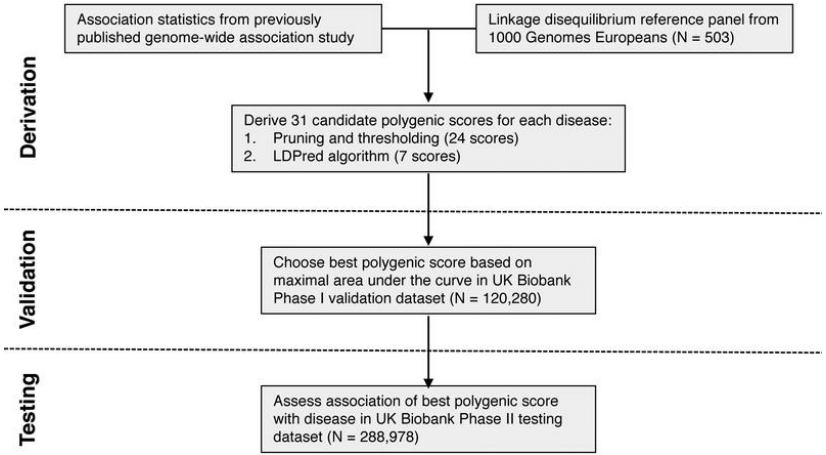


Reference: Fahed et al. (2022) Nature Communications

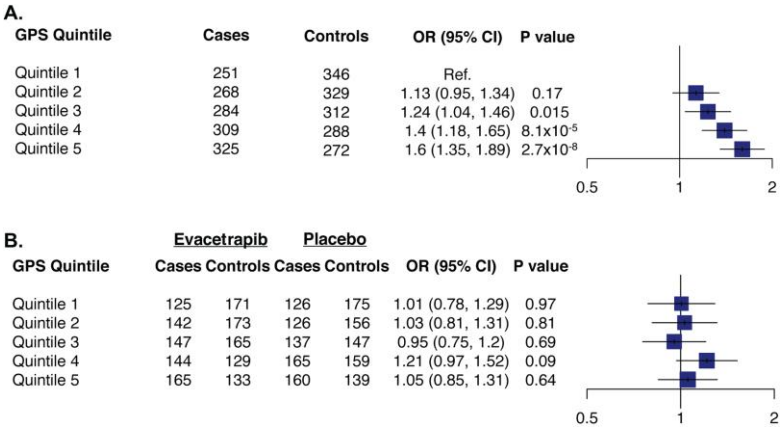
A polygenic score applied to clinical trial patients

A previously validated PGS for coronary artery disease → **Applied to patients with a major cardiovascular event observed in trial follow up**

60,801 cases; 123,504 controls



Khera et al. (2018) *Nature Genetics*, *50*(9), pp.1219-1224.



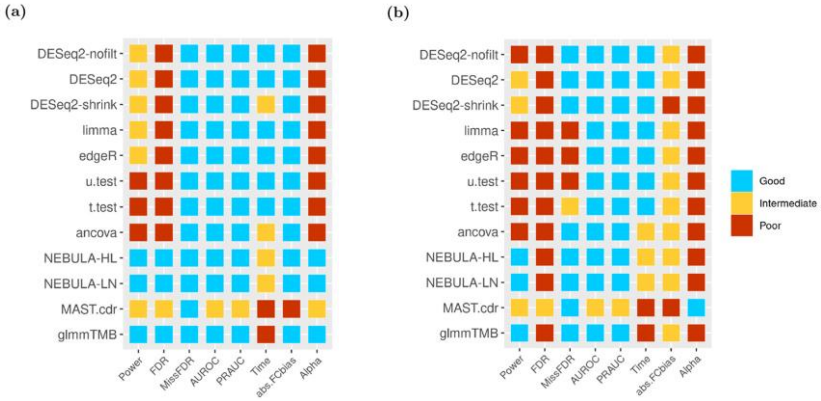
Emdin et al. (2020) *Circulation: Genomic and Precision Medicine*, *13*(1), p.e002767.

Potential bias in the use of polygenic score

- PGS usually requires large patient-level genomic studies and resources which are often enriched with participants with European ancestry
- This could lead to challenges in replication and utilization of PGS in other populations, especially minorities
- Active research has already been underway to improve performance of multi-ethnic PGS and other aspects of PGS
- However, it’s important for the entire research community to work together to address the issue and advocate diversity, equity, and inclusion in precision medicine research and implementation

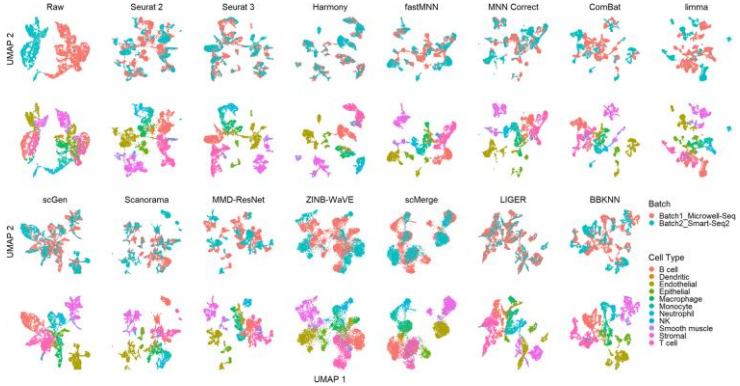
Examples of other omics research in discovery

- Recommendations of single cell RNA-seq (scRNA-seq) differential gene expression analysis based on comprehensive benchmarking



Gagnon et al. (2022). *Life*, *12*(6), p.850.

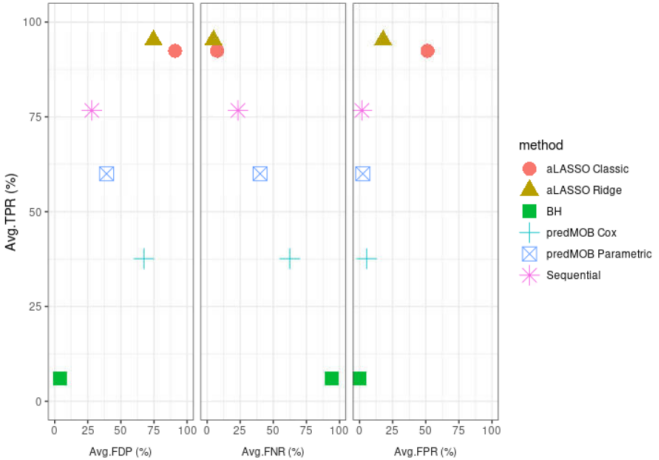
- Building atlas data by harmonizing public scRNA-seq datasets to
  - understanding upstream signaling regulators involved in certain function
  - Identify cell surface markers associated with certain phenotype



Genome Biology volume 21, Article number: 12 (2020)

Examples of other omics research in discovery & preclinical studies

- Predictive biomarkers discovery when candidate biomarkers are correlated
  - Permutation-based multiple testing (Michiels et al., 2011)
  - Multivariate (e.g., penalized regression) outperformed univariate (Ternes et al., 2017)
  - Variants of adaptive LASSO achieved best performance when candidate biomarkers can be grouped (Belhechmi et al., 2020)
  - Simulation results suggests some methods (e.g., sequential as in adaptive LASSO followed by predMOB) could achieve a good balance of TPR and FPR



Credits: Lira Pi, Yang Liao (Unpublished work)

- Active research in multi-omics methods and analysis

- …

Summary

- Computational biology involves analysis of large collections of biological data, e.g., genomics, transcriptomics, proteomics, …
- Increasing interests in leveraging human genetics to inform drug discovery and development
- The interdisciplinary field offers opportunities for statisticians to collaborate with various teams in e.g., target science, human genomics, computational biology, and biology