**3.1 Machine Learnings from HCA**

Dr Sarah Teichman

<http://www.teichlab.org/>

**Human Genome Project**

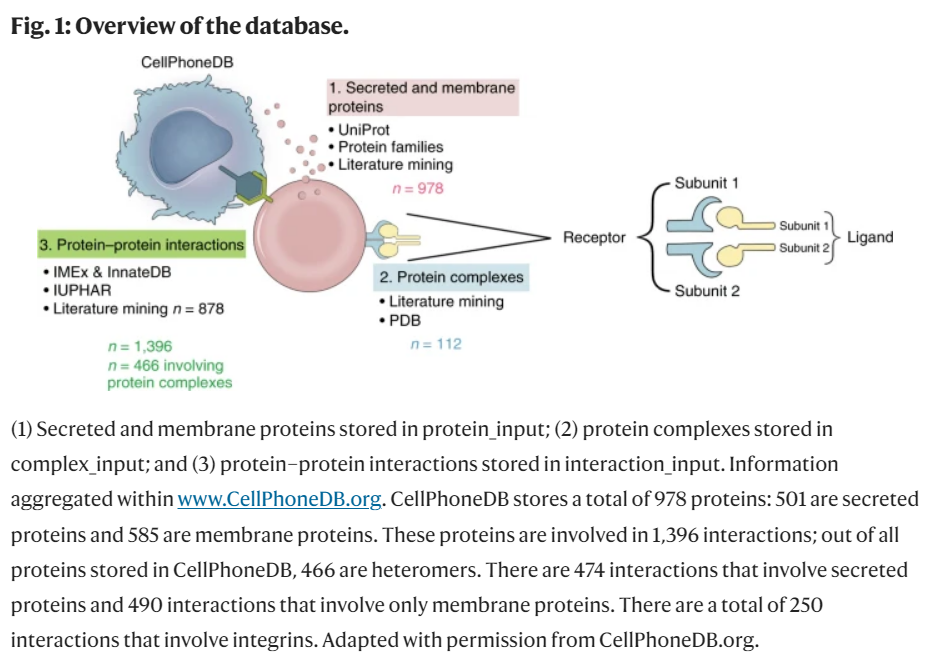
* Understanding how it is possible to encoding a multitude of cell types from a single genome?
  + Subset of genes that have been switched on/off defining the shape and functions of the cell types
* Cell mapping
  + “We need a programme of making maps of cells, and maps of how cells talk to each other… **The CellMap project, for which we don’t need a model organism, will be one of the things to occupy us for the next few decades**” – Sydney Brenners 2002
    - No measurement to quantify single cell
  + 2016
    - Aviv Regev (Broad Institute)
* Human Cell Atlas
  + To create a comprehensive reference mapof the types and properties of all human cells, the fundamental unit of life, as a basis for understanding, diagnosing, monitoring, and treating health and disease
  + Joining the initiative
    - <https://www.humancellatlas.org/join-hca/>
    - 2,414 members, 1,200 institutes, 79 countries
    - Multi-disciplinary

**Resolution revolution: single cell genomics, spatial transcriptomics**

* Combining sc genomics and spatial transcriptomics
  + WIth AI and ML
* Cell mapping at scale during development, physiology & disease
* Example projects from Teichmann lab
  + Maternal-fetal interface
    - Vento-tormo Nature 2018
      * <https://www.nature.com/articles/s41586-018-0698-6>
  + Immune development
    - Popescu Nature 2019
      * <https://www.nature.com/articles/s41586-019-1652-y>
    - Park Science 2020
      * <https://www.science.org/doi/10.1126/science.aay3224>
    - Jardine Nature 2021
      * <https://www.nature.com/articles/s41586-021-03929-x>
    - Suo, Dann Science, 2022
      * <https://www.science.org/doi/10.1126/science.abo0510>
  + Lung
    - Vieira Braga Nat Med 2019
      * <https://www.nature.com/articles/s41591-019-0468-5>
    - Madissoon, Oliver bioRxiv 2022
      * <https://www.biorxiv.org/content/10.1101/2021.11.26.470108v1>
  + Gut
    - James Nat Immunol 2020
      * <https://www.nature.com/articles/s41590-020-0602-z>
    - Elmentaite Dev Cell 2020
      * <https://linkinghub.elsevier.com/retrieve/pii/S1534-5807(20)30886-8>
    - Elmentaite Nature 2021
      * <https://www.nature.com/articles/s41586-021-03852-1>
  + Heart & skeletal muscle
    - Litvinukova Nature 2020
      * <https://www.nature.com/articles/s41586-020-2797-4>
    - Kedlian bioRxiv 2022
      * <https://www.biorxiv.org/content/10.1101/2022.05.24.493094v1>
  + Cross-tissue immunity
    - Dominguez-conde, Xu, Jarvis Science 2022
      * <https://www.science.org/doi/10.1126/science.abl5197?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed>

**Computational tool overview**

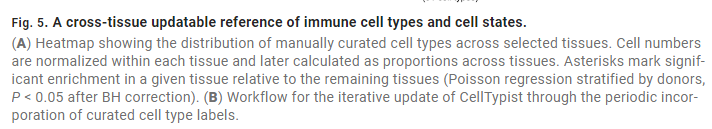
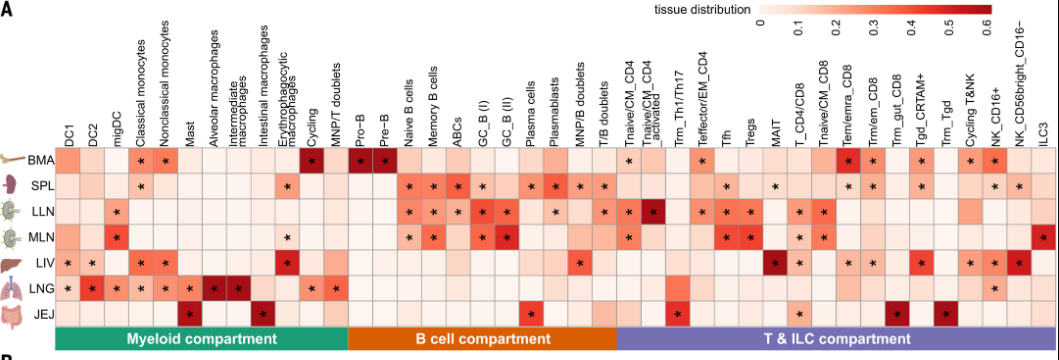
* Fundamental tools to understanding the cell
  + CellTypist: suspension cell
    - <https://www.celltypist.org/>
  + Cell2location: mapping tissue architecture
    - Integrating with spatial data
    - <https://cell2location.readthedocs.io/en/latest/>
  + CellphoneDB: Inferring cell-cell interactions
    - <https://www.nature.com/articles/s41596-020-0292-x>
  + Drug2cell:
    - Drug target exploration
* CellTypist
  + <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7612735/>
  + Precise and rapid annotation of single cell data
  + > 100 Immune cell type and states
    - One model encompassing all tissues
* Cell2location
  + <https://cell2location.readthedocs.io/en/latest/>
  + <https://www.nature.com/articles/s41587-021-01139-4>
* CellPhoneDB
  + <https://www.nature.com/articles/s41596-020-0292-x>



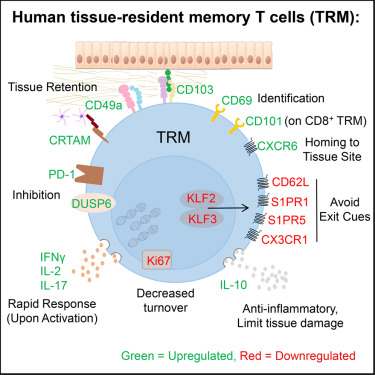
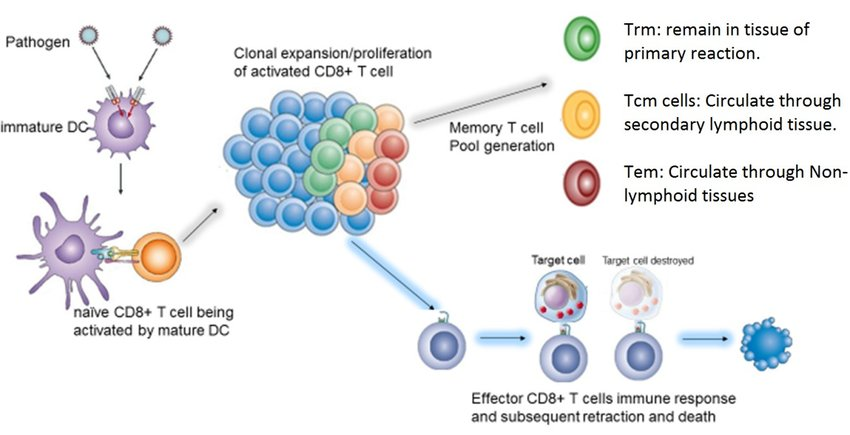
* Drug2cell
  + Based on ChEMBL
  + Finding drugs that act on cells
    - Clinically approved compounds wrt their targets predicted by transcriptomics
  + Manuscript under preparation
  + ChEMBL + Single-cell/nuclei RNAseq -> Filter drugs -> Drug scores -> Find drugs/ cells -> find target molecules

**CellTypist**

* What are the cell states across organs?
* What features are context-specific?
  + Is the related to the positioning of the data
* Categorizing the distributed immune system
  + 12 donors, ~330k immune cells from lymph nodes, blood samples, bone marrows, liver, thymus, lung, muscle, kidney, and gut
    - scRNA-seq or scVDJ-seq data
    - Generating scRNA-seq
  + With label curation and model training for CellTypist
    - Annotate cell in generated scRNA-seq data
* From knowledge-driven to data-driven
  + Knowledge and data are negatively correlated
    - More data, less knowledge
    - Less data, more knowledge required
  + Knowledge is required for manual annotation
  + Data is required for automatic annotation
  + AI/ML models
    - cross-tissue database and server integration?
    - Interpretable pipeline for label projection?
* Classical machine learning vs deep learning
  + High performance can be achieved with canonical machine learning methods
    - **A comparison of automatic cell identification methods for single-cell RNA sequencing data**
      * <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1795-z>
    - **Deep learning does not outperform classical machine learning for cell-type annotation**
      * <https://www.biorxiv.org/content/10.1101/653907v2.full>
  + Notably logistic regression models
    - **Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells**
      * <https://www.cell.com/cell/fulltext/S0092-8674(16)31309-5?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867416313095%3Fshowall%3Dtrue>
    - **Single-Cell Sequencing of Developing Human Gut Reveals Transcriptional Links to Childhood Crohn’s Disease**
      * <https://www.cell.com/developmental-cell/fulltext/S1534-5807(20)30886-8?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1534580720308868%3Fshowall%3Dtrue>
* Optimisation
  + Mini-batch training
    - Bypass memory overload and decrease execution time
    - 1,000 random cells as a training batch
    - 100 mini-batches per epoch \* 10 epochs
  + Stochastic gradient descent learning
    - Fast convergence for large datasets
    - Allow for online training (incorporating future new datasets)
* Cell type distribution across the body
  + Big picture of enrichment of specific cell types of specific tissues
  + <https://www.science.org/doi/10.1126/science.abl5197>



* Further insights into in vivo CD8+ T cell memory formation
  + Lower expression: migration to lymph node
  + Unbiased analysis with further understanding of the biology



* Summary
  + Cross tissue adult immunity
  + Integration of immune system across body
  + CellTypist: cell type encyclopedia
  + Towards an integrated Human Cell Atlas

**Focus on the Heart: tissue microenvironments and drug targets**

* Mapping how cells fit into the tissue microenvironment
* Expansion of the atlas:
  + V2 capturing other anatomical position to V1
    - Higher RA, PV, aorta…
* Understanding rare diseases
  + Atrial fibrillation
  + Aortopathies
  + Mitral valve prolapse
  + Coronary artery disease
  + Brugada syndrome
* Discovering spatial microenvironment
  + Visium slide -> Automated (cell2location) and structural (manual) annotation harmaonized -> identifying factors which refine manual automation -> CellPhone v3 + GPCR-DB
  + Understanding the electric conduction that flows from the pacemaker cells in the heart through all the cardiac walls of chambers
* Cell2Location telling us centric node
* Detailed understanding of where the spontaneously farms is
  + Cellular molecular level
  + Peripheral structure of particular cell populations
* Epicardium & fibrotic myocardial niches
  + Immune shielding around the edge of the heart
  + Fibrotic area associated to aging
    - Secretion of collagen
    - Activated, next to vessel
    - TGF and BMP signalling
* Pacemaker cells
  + Cardiac conduction system
  + Semi-nerounal and semi-muscular fibres
  + Singel cell analysis of pacemaker cells
  + Identifying SAN-region pacemaker cells
    - Unbiased clustering approaches
    - Working cardiac myocytes and Pacemaker cell separation
      * Low sodium channel SCN5A expression and high calcium channel CACNA1D expression
    - How are the ion channels affected by drugs?
      * Subclasses of drugs frequently targeted:
        + Which drug having an effect on these ion channels? Thus side effects on the heart?

ChEMBL

<https://www.ebi.ac.uk/chembl/>

Integration with ChEMBL – Drug2Cell

* + Drug target exploration with drug2cell
    - Identification of diabetic drugs which targets pacemaker cells
      * GLP1 analogues Perampanel targeting GLP1R and GRIA3 in pacemaker cells
    - Up to 6 beats per minute
      * Might be via the autonomic nervous system
      * This analysis pinpoints specifically– hypothesis generation
        + Homeostatic perturbation at autonomic nervous system

Brainstem and cord/ higher CNS centers

* Summary
  + Define cell type composition
  + Discover tissue microenvironments
  + Map drug targets

**Q&A**

* Data harmonisation, cleaning are mostly addressed not manually
* Classical ML vs DL
  + DL usually in normalization, integration tasks rather than hypothesis generation and elucidating mechanisms in single cell field
  + Gain for deep learning
    - Possibly in the future
    - Interpretability will be a major potential pitfall – obscure variables from DL
* CellTypist
  + Continuous improvement
  + Accuracy higher for immune area compared to stromal cells
    - Dependent of effort and data availability
* Single cell data direction
  + Multi-modal directions
* Single omics in understanding systemic impact
  + Cross-tissue integration, spatial microenvironment, cell-cell communication
  + Stitching cardiac conduction system
    - Has been performed
    - Understanding the overall system
  + Fetal trophoblast cells communicating with maternal feature
* Important skillsL
  + Python and R
  + Data handling and analysis
  + Framework
  + Asking questions on the data rather than new method development
    - Then going into complicated projects and questions