Advanced Topics in Bioinformatics

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2021

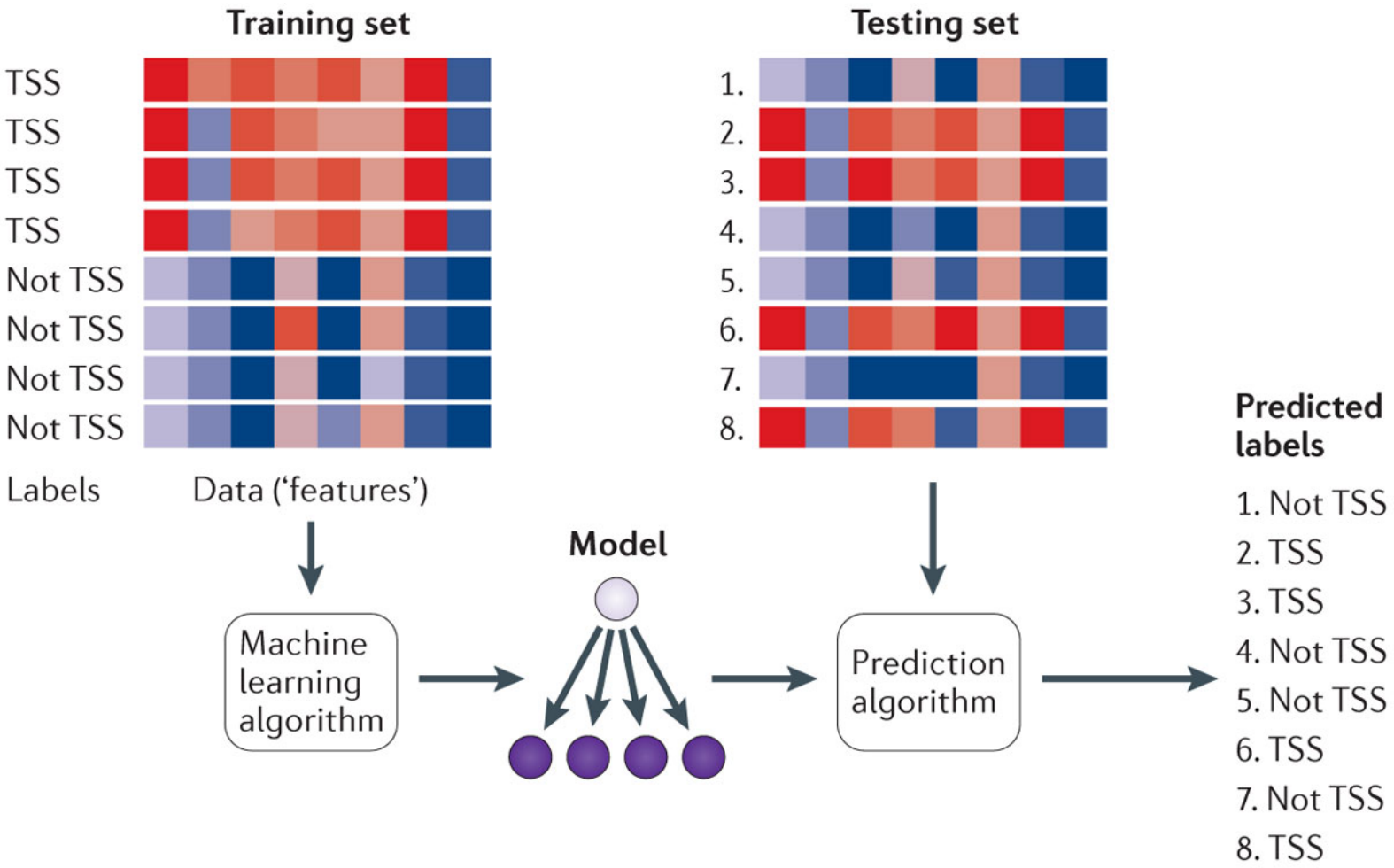
Daifeng Wang

daifeng.wang@wisc.edu

Outline

- More machine learning applications
- Machine learning challenges in bioinformatics
- Spatial transcriptomics
- Imaging genetics/genomics
- Artificial intelligence in drug discovery

Training and Testing



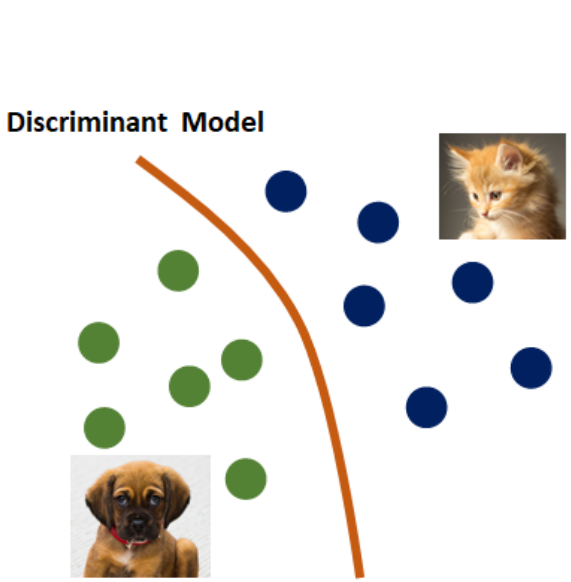
Libbrecht, M., Noble, W. Machine learning applications in genetics and genomics. *Nat Rev Genet* **16,** 321–332 (2015). https://doi.org/10.1038/nrg3920

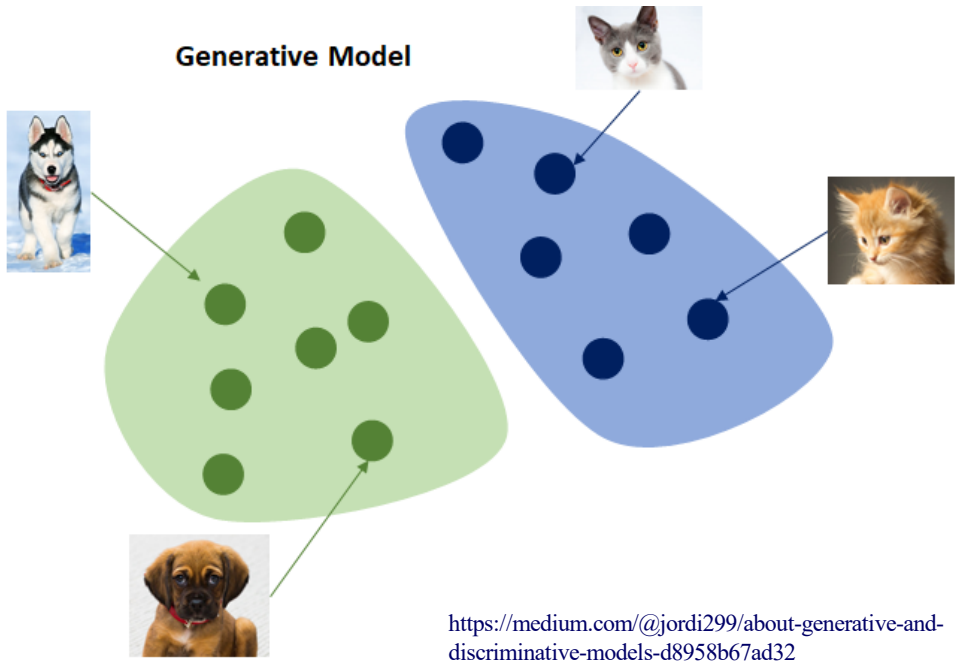
Nature Reviews | Genetics

Generative vs. Discriminative models

- Generative approaches model the joint probability p(x,y) for generating data

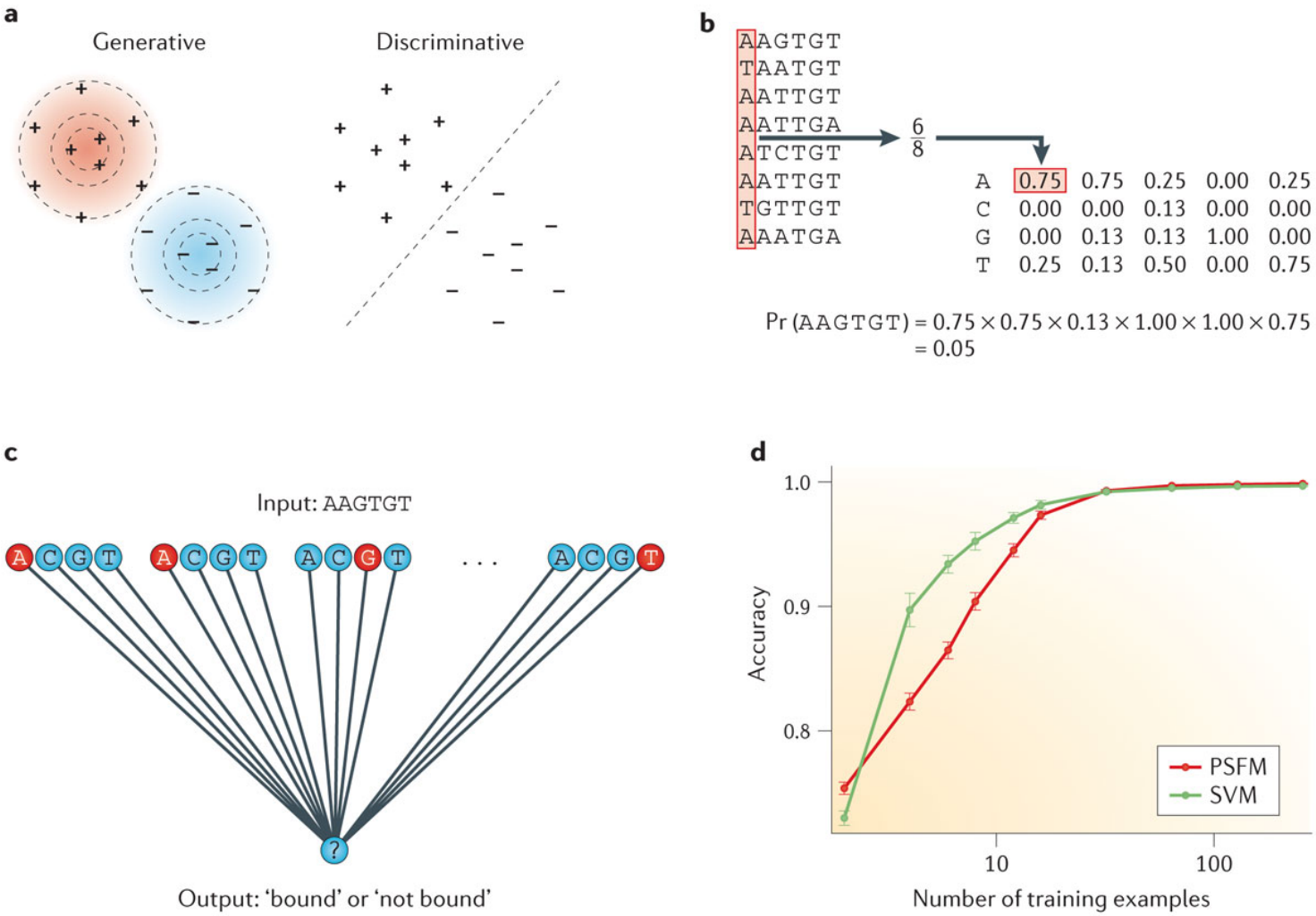
- Discriminative approaches directly model p(y|x) for classification





https://medium.com/@jordi299/about-generative-and-discriminative-models-d8958b67ad32

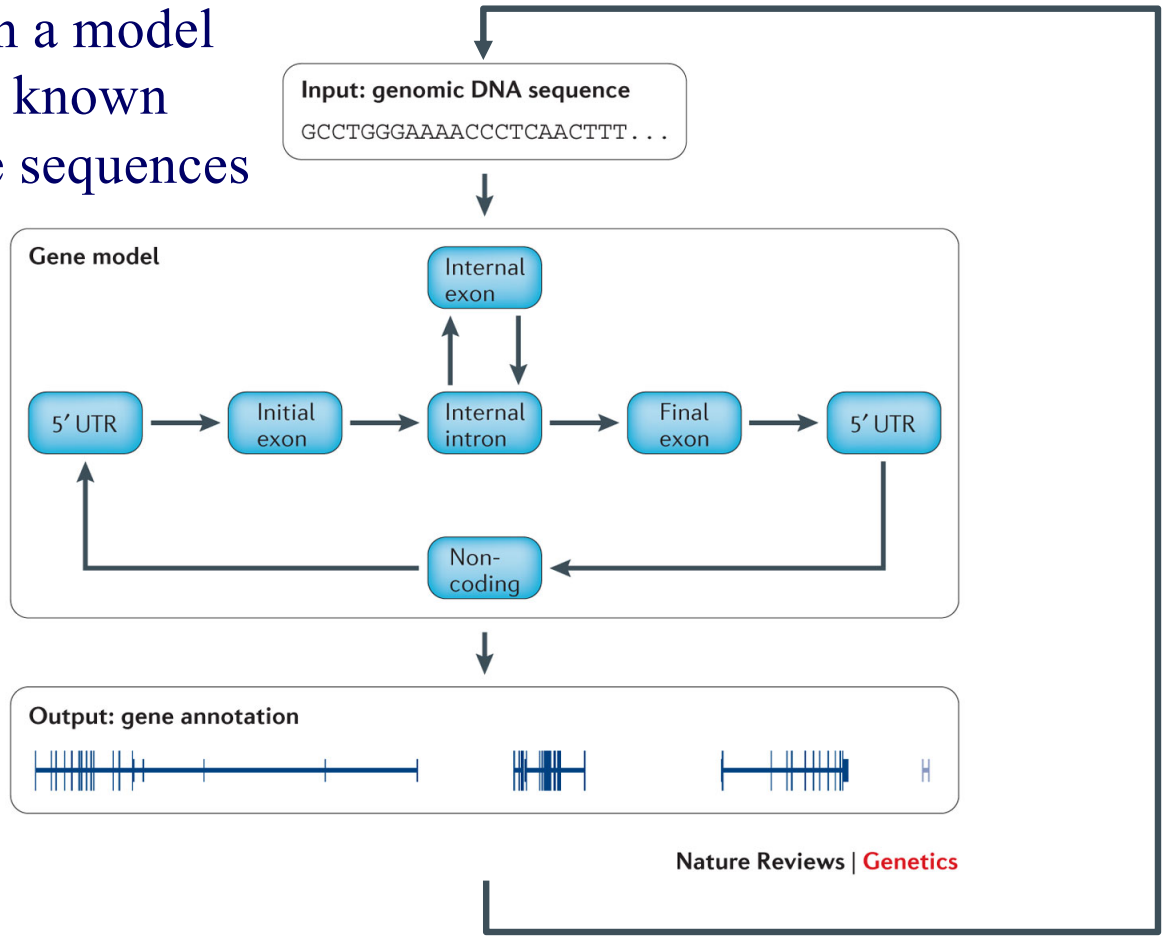
Predicting TF binding via Generative vs. Discriminative models



Libbrecht, M., Noble, W. Machine learning applications in genetics and genomics. *Nat Rev Genet* **16,** 321–332 (2015). https://doi.org/10.1038/nrg3920

Semi-supervised learning (e.g., gene finding)

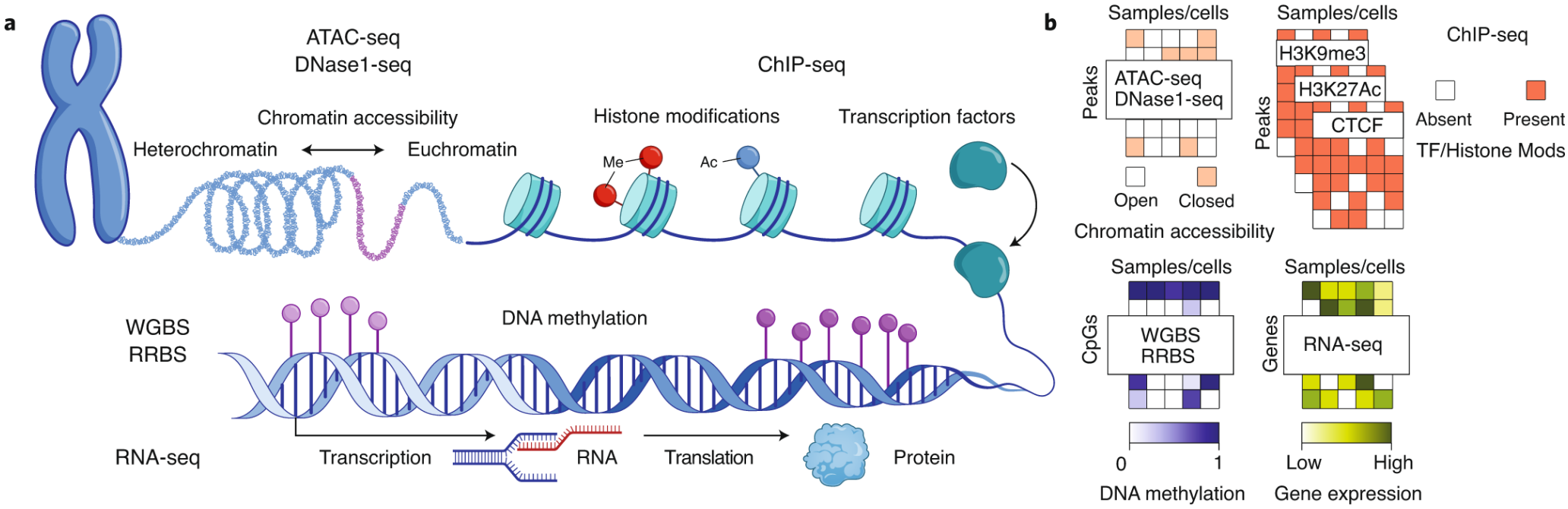
- Train a model with known gene sequences



- Predict labels for many unknown sequences
- Refine the model with known and predicted sequences

Libbrecht, M., Noble, W. Machine learning applications in genetics and genomics. *Nat Rev Genet* **16,** 321–332 (2015). https://doi.org/10.1038/nrg3920

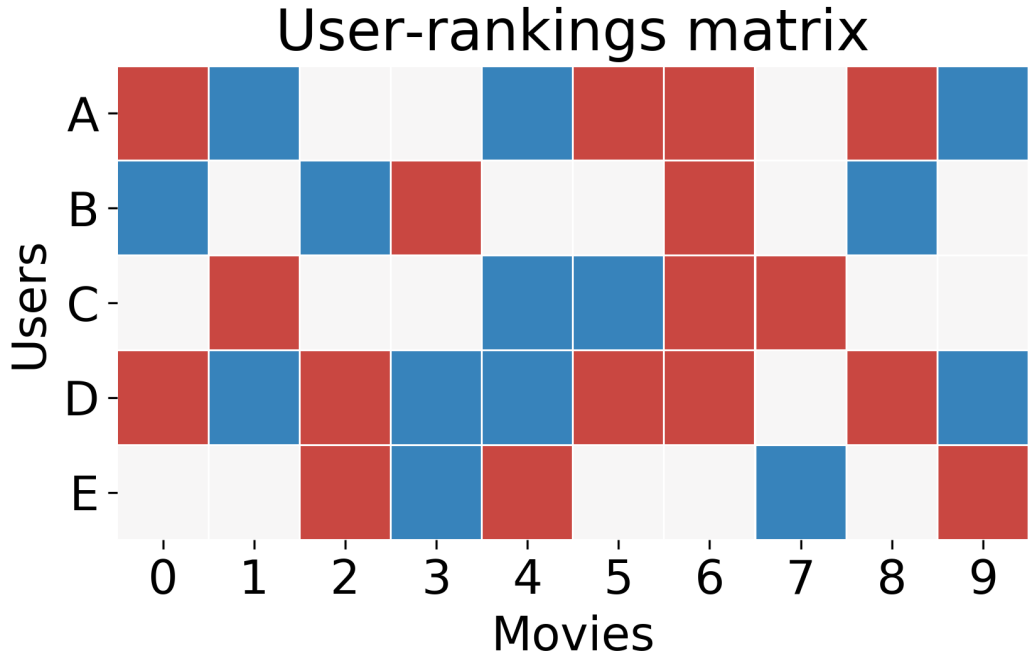
Data heterogeneity



Scherer, M., Schmidt, F., Lazareva, O. et al. Machine learning for deciphering cell heterogeneity and gene regulation. Nat Comput Sci 1, 183–191 (2021). https://doi.org/10.1038/s43588-021-00038-7

Missing data and imputation

- Remember Netflix Problem?



https://krishnaswamylab.github.io/tutorial/imputation_and_netflix/

- Biological data has many missing values
  - e.g., single cell dropouts

Spatial Transcriptomics

- Assessment of gene expression profiles and spatial organization for interrogation of complex, heterogeneous tissues

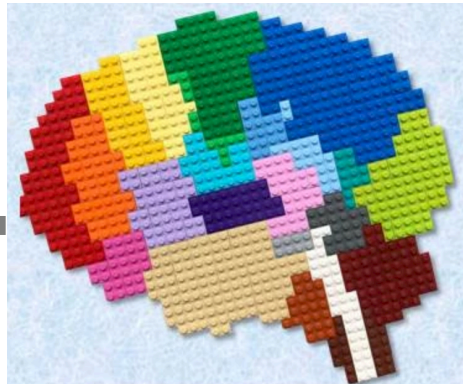
Bulk RNA-seq

- Average gene expression level, lacks cellular or spatial resolution

Single cell RNA-seq

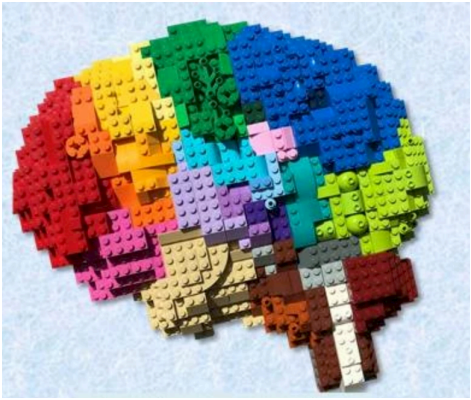
- Identify distinct cell types and their gene expression profiles

Spatial Transcriptomics



- Visualize gene expression profiles with tissue context

Functional tissue



- Spatial organization of distinct cell types and their gene expression profiles

https://boxia2018.wixsite.com/boxia/momemts https://www.youtube.com/watch?v=20-qNM4Ax3s

Genomics Visium platform

- Visium Targeted Gene Expression combines crucial spatial insights with the ease and breadth of targeted panels
- Accelerate the understanding of human health and disease with a more refined picture of the biology captured on a tissue slide

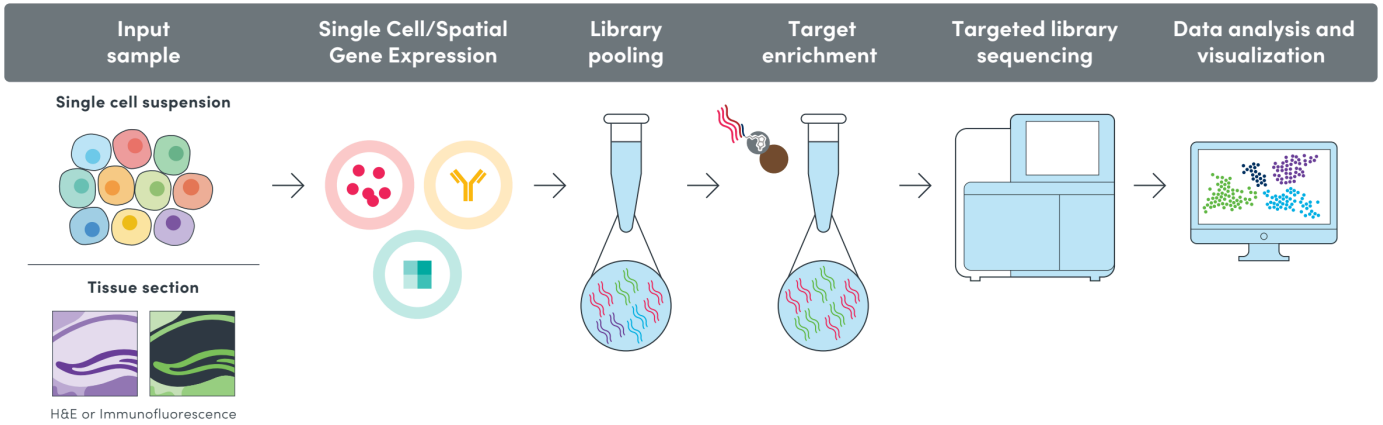


Figure 7. Targeted Spatial Gene Expression workflow with Visium Spatial solutions. Targeted Spatial Gene Expression enables the enrichment and analysis of a targeted set of mRNAs prepared from tissue sections. Starting with a final, barcoded 10x Genomics library, the workflow allows whole transcriptome and targeted gene expression on the same samples, while simultaneously examining morphology or co-detecting proteins.

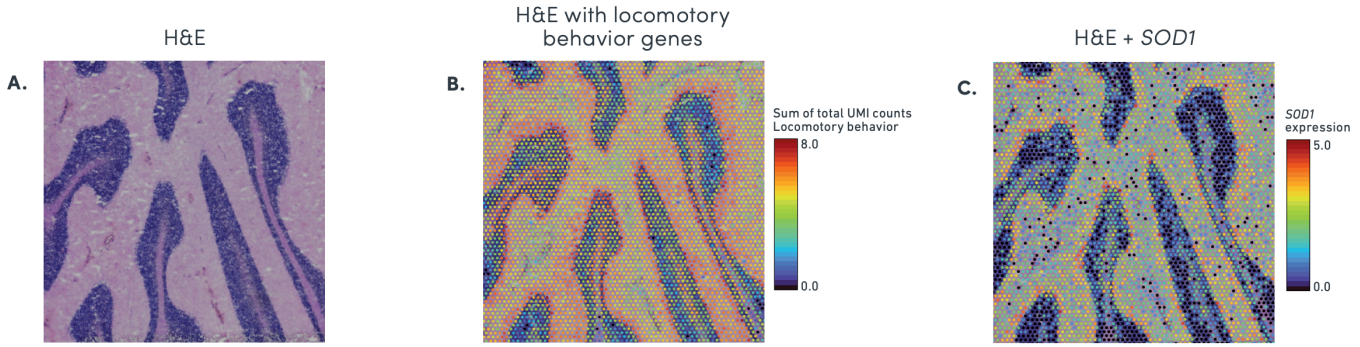


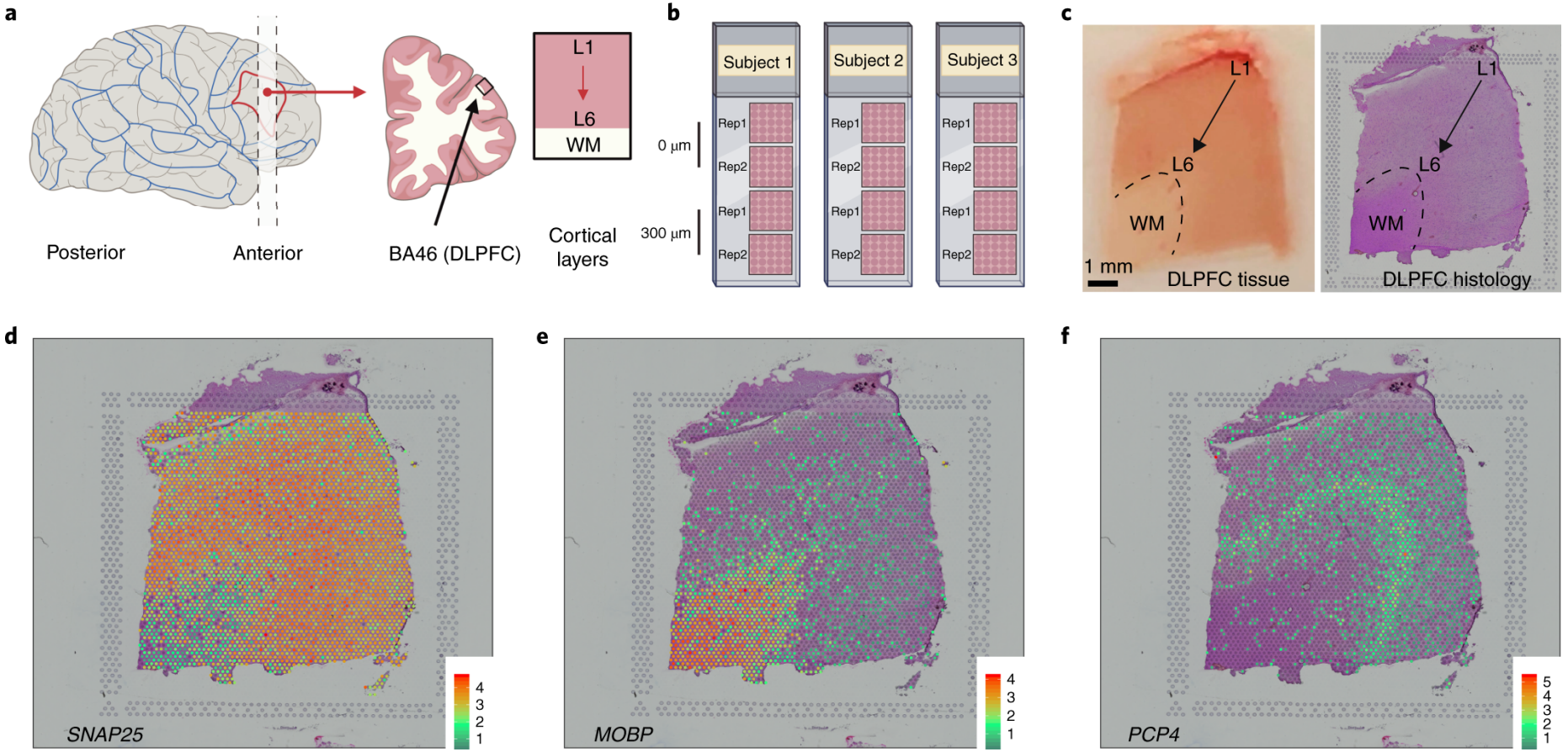
Figure 8. Spatially resolved targeted gene expression profiling with Visium Spatial solutions. A human cerebellum tissue section was H&E stained and processed using the Visium Spatial Gene Expression workflow, then enriched using Targeted Gene Expression with the Human Neuroscience Panel. Shown are the H&E image (A), H&E image overlaid with total UMI counts for 36 locomotory behavior genes from the neuroscience panel (B), and H&E image overlaid with *SOD1* expression level (C).

https://pages.10xgenomics.com/rs/446-PBO-704/images/10x_BR060_Inside_Visium_Spatial_Technology.pdf

Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex (DLPFC)

- Localize spatial gene expression in the human brain at cellular resolution will be critical to gain further insight into disease mechanisms

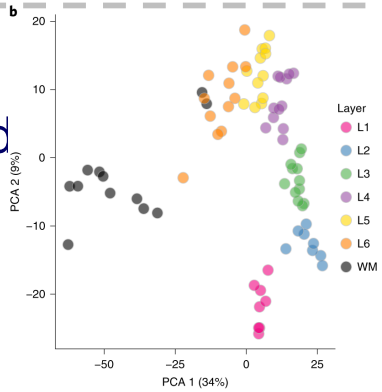
Spatial transcriptomics in DLPFC using Visium

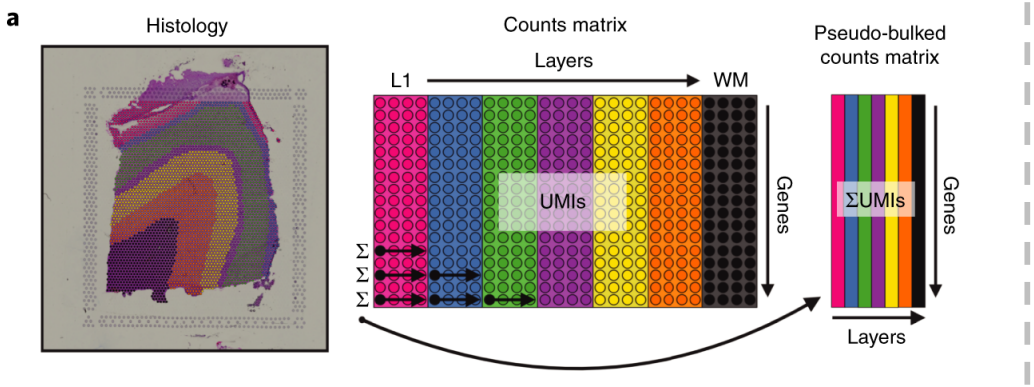


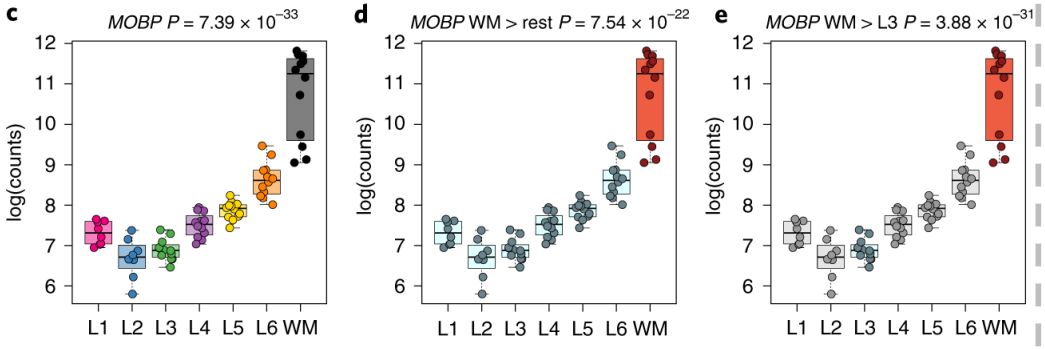
Maynard, K.R., Collado-Torres, L., Weber, L.M. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* **24,** 425–436 (2021). https://doi.org/10.1038/s41593-020-00787-0

Further applications of spatial gene expression

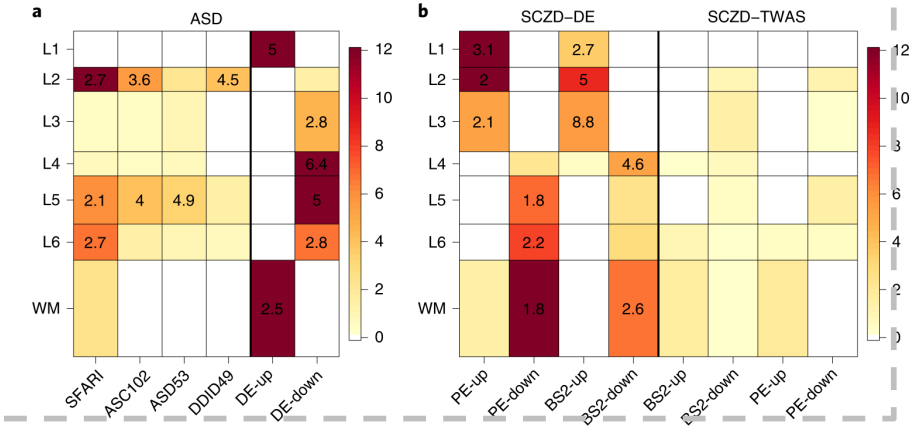
- Extensive layer-enriched expression signatures and refined associations to previous laminar markers



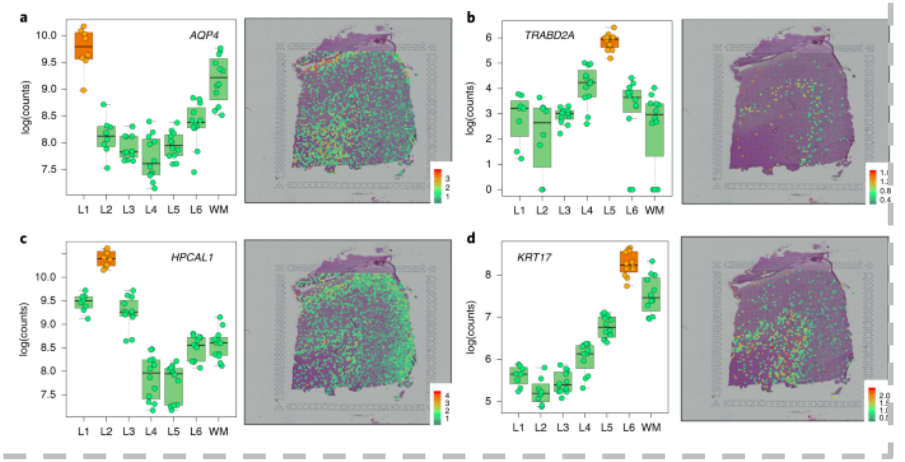




- Differential layer-enriched expression of genes associated with schizophrenia disorder (SCZD) and autism spectrum disorder (ASD), highlighting the clinical relevance of spatially defined



- Novel cortical layer-enriched genes

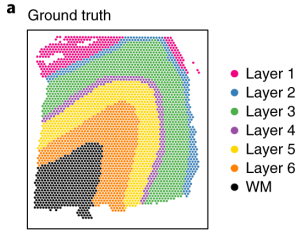


Maynard, K.R., Collado-Torres, L., Weber, L.M. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* **24,** 425–436 (2021). https://doi.org/10.1038/s41593-020-00787-0

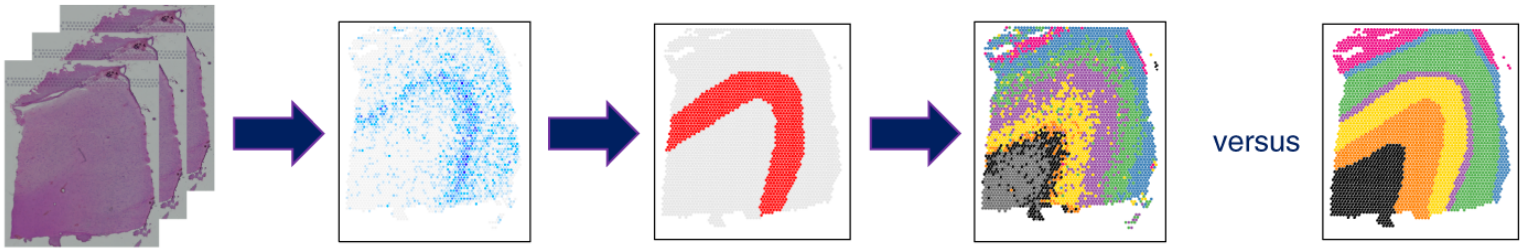
Further applications of spatial gene expression in DLPFC

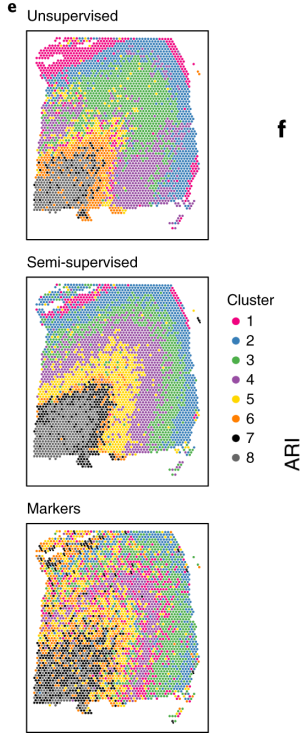
- A data-driven framework to define unsupervised clusters in spatial transcriptomics data, which can be applied to other tissues or brain regions in which morphological architecture is not as well defined as cortical laminae

Supervised annotation of DLPFC layers

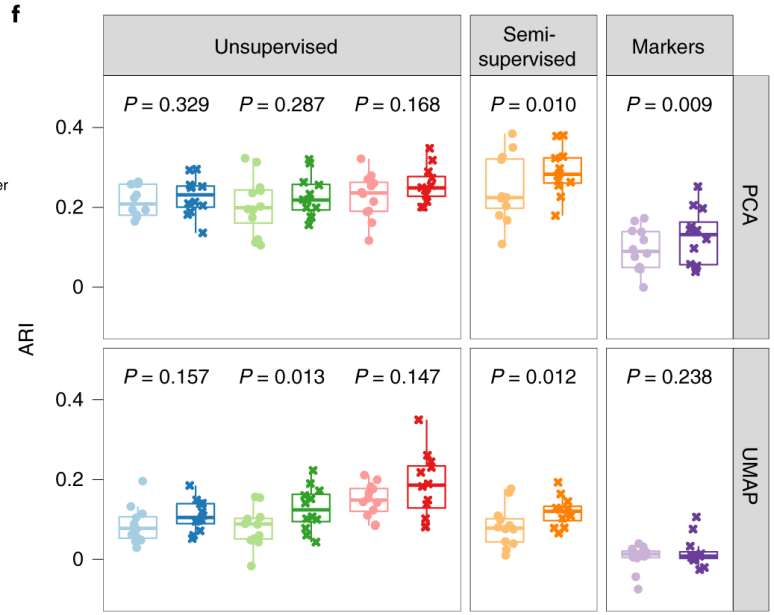


Schematic illustrating the data-driven clustering pipeline

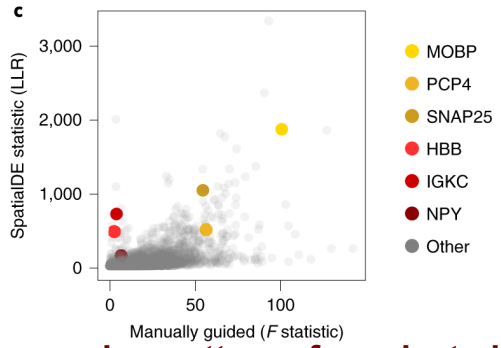




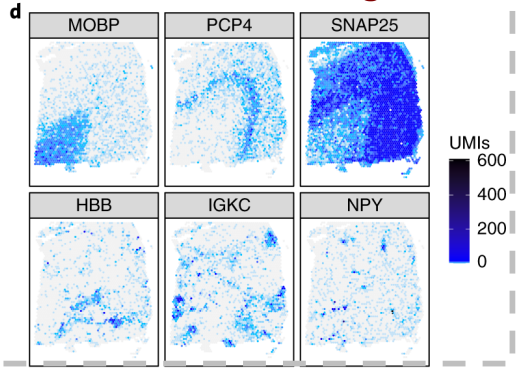
Evaluation of clustering performance



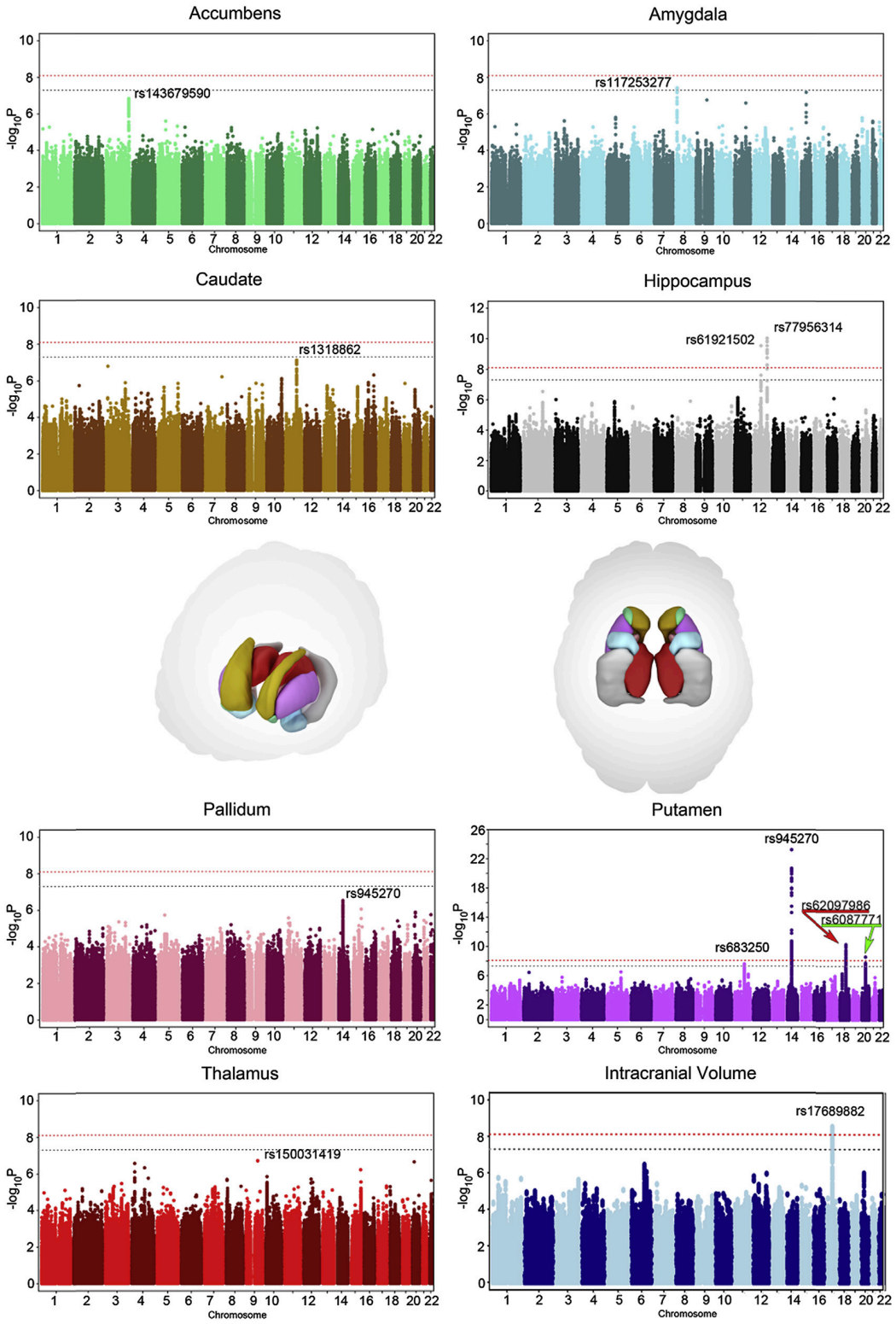
Comparison of gene-wise test statistics



Expression patterns for selected laminar and nonlaminar genes



Maynard, K.R., Collado-Torres, L., Weber, L.M. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* **24,** 425–436 (2021). https://doi.org/10.1038/s41593-020-00787-0



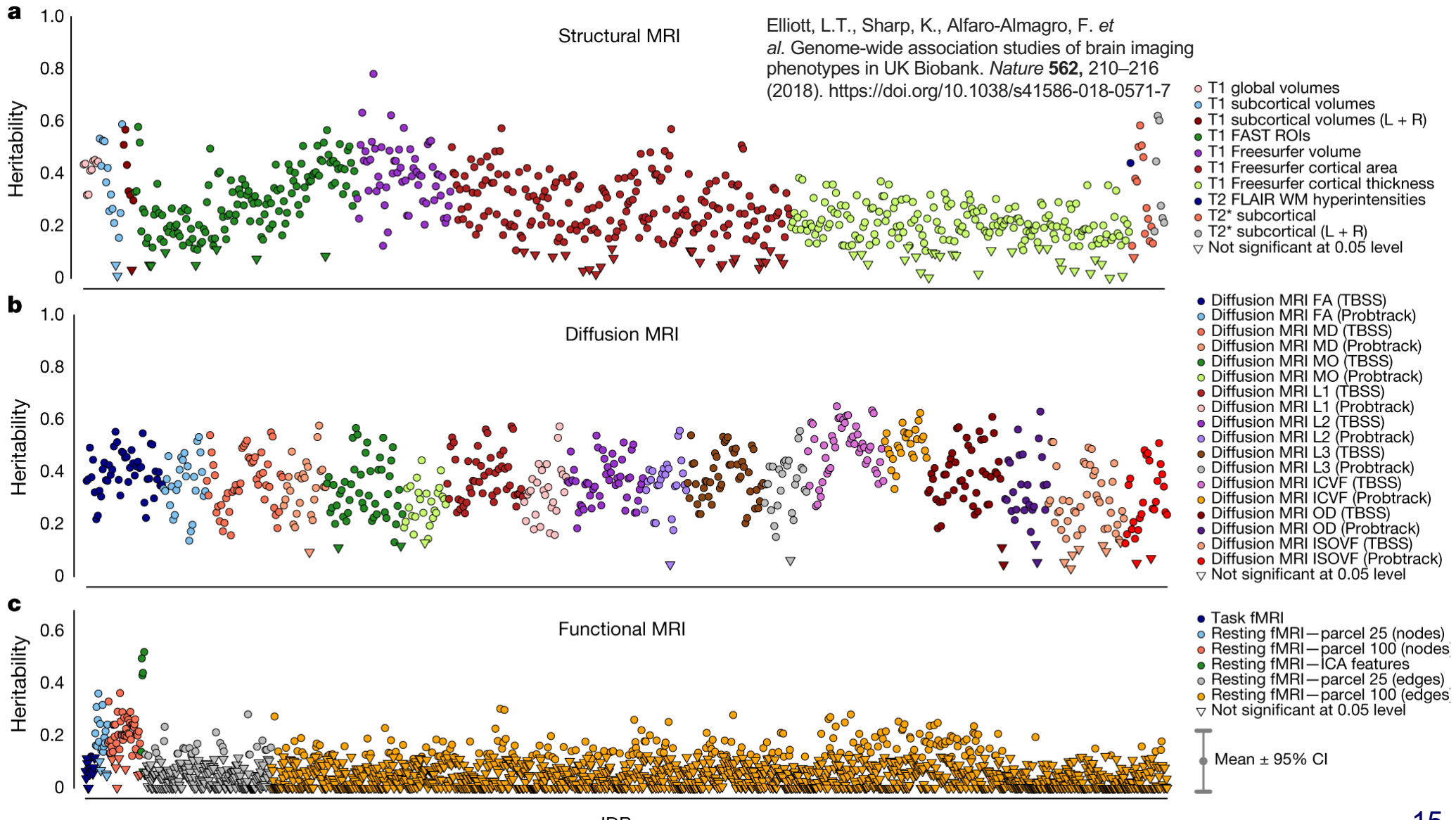
Imaging genetics

- GWAS for imaging phenotypes
  - Subcortical region volumes

Genomics, Circuits, and Pathways in Clinical Neuropsychiatry. http://dx.doi.org/10.1016/B978-0-12-800105-9.00007-X

UK BioBank

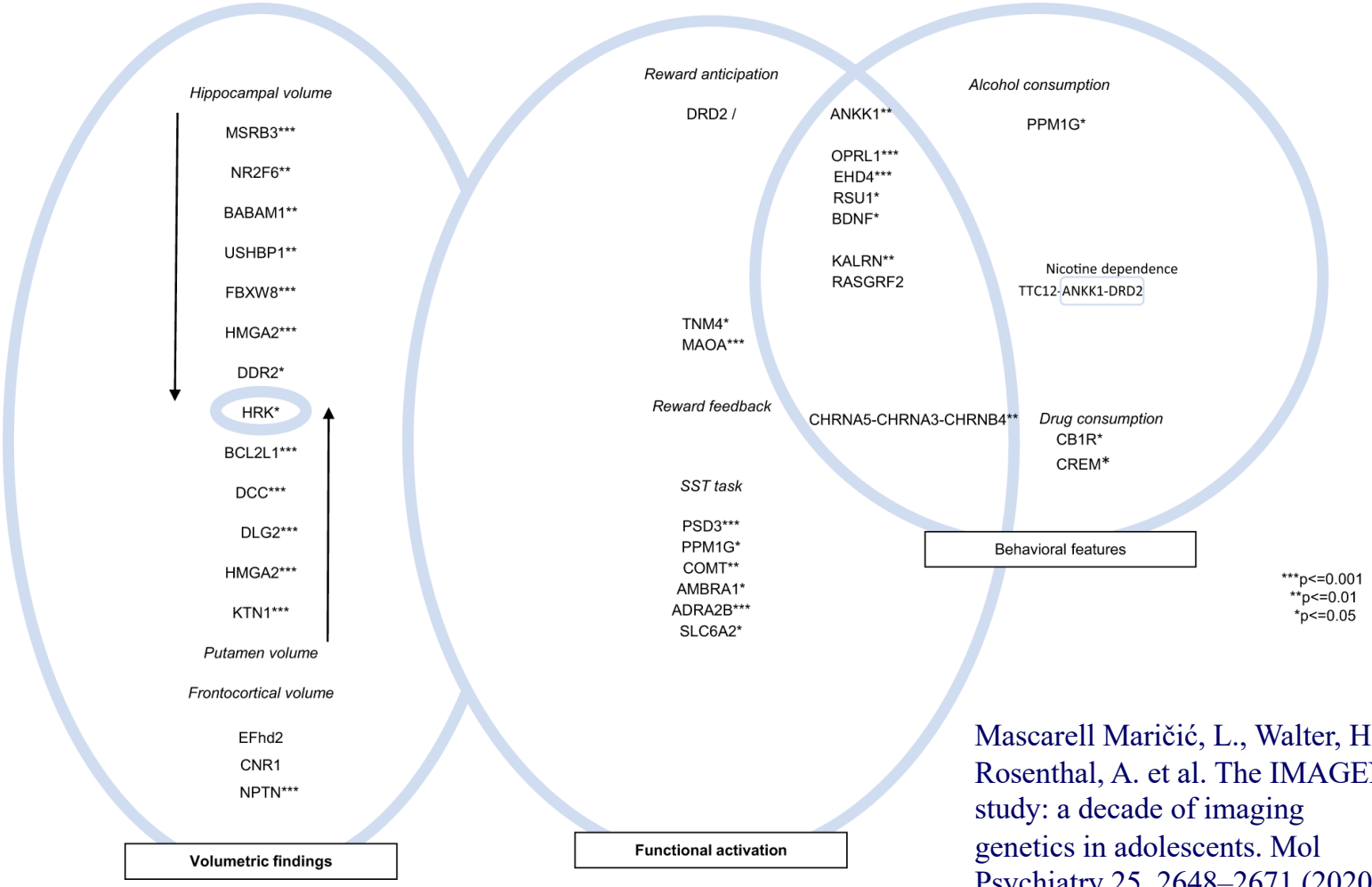
n = 8,428 subjects for 3,144 functional and structural brain imaging phenotypes



Elliott, L.T., Sharp, K., Alfaro-Almagro, F. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562,** 210–216 (2018). https://doi.org/10.1038/s41586-018-0571-7

Linking genes to brain phenotypes

- IMAGEN cohort



Mascarell Maričić, L., Walter, H., Rosenthal, A. et al. The IMAGEN study: a decade of imaging genetics in adolescents. Mol Psychiatry 25, 2648–2671 (2020).

Artificial Intelligence vs. Machine learning

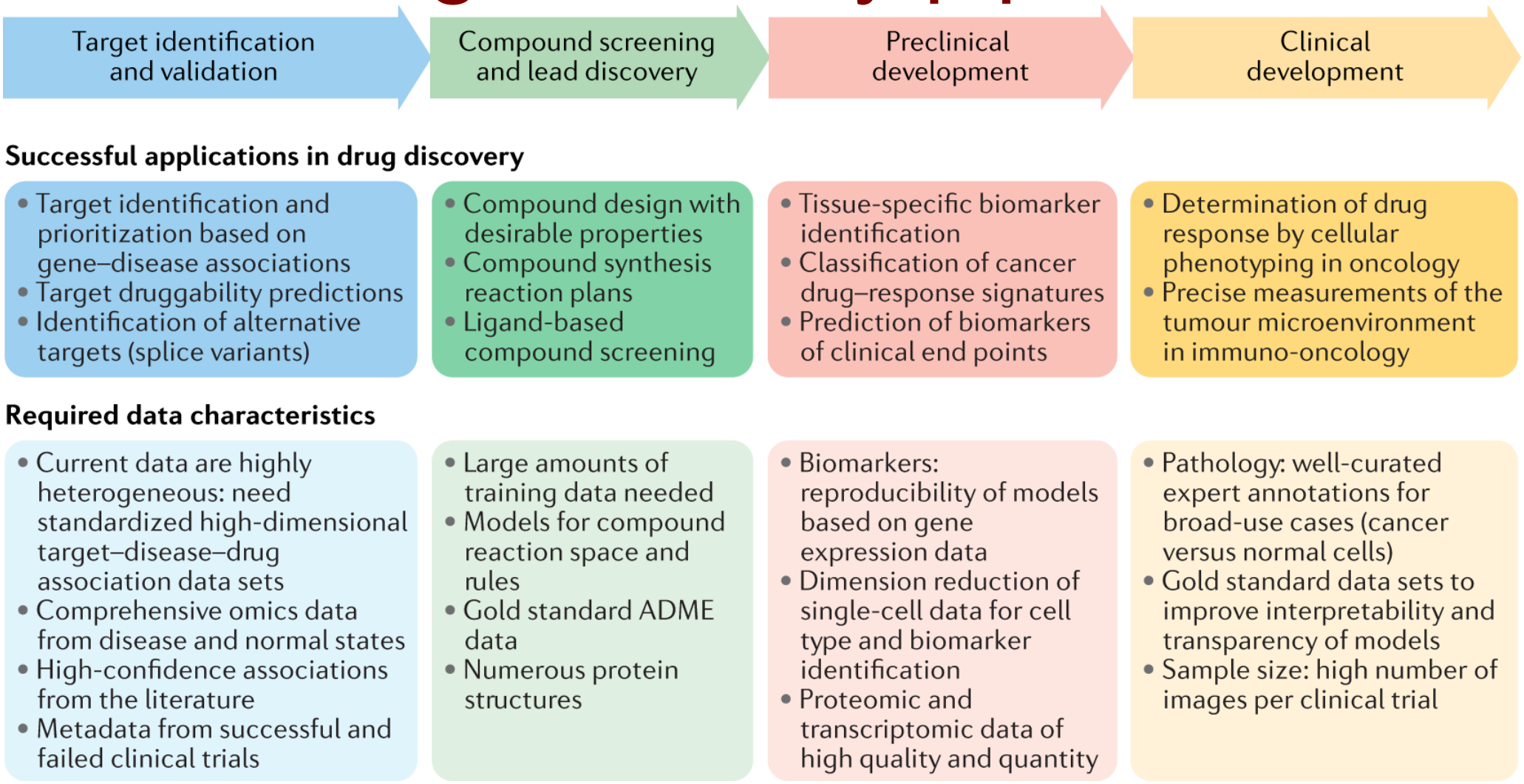
- Artificial Intelligence (AI)
  - Broad concept using machines to do human-intelligence tasks
  - Visual perception, speech recognition, etc.
- Machine learning (ML)
  - An AI subarea making machines (e.g., computers) automatically learn how to finish tasks

https://commonfund.nih.gov/bridge2ai/faqs

AI/ML in drug discovery

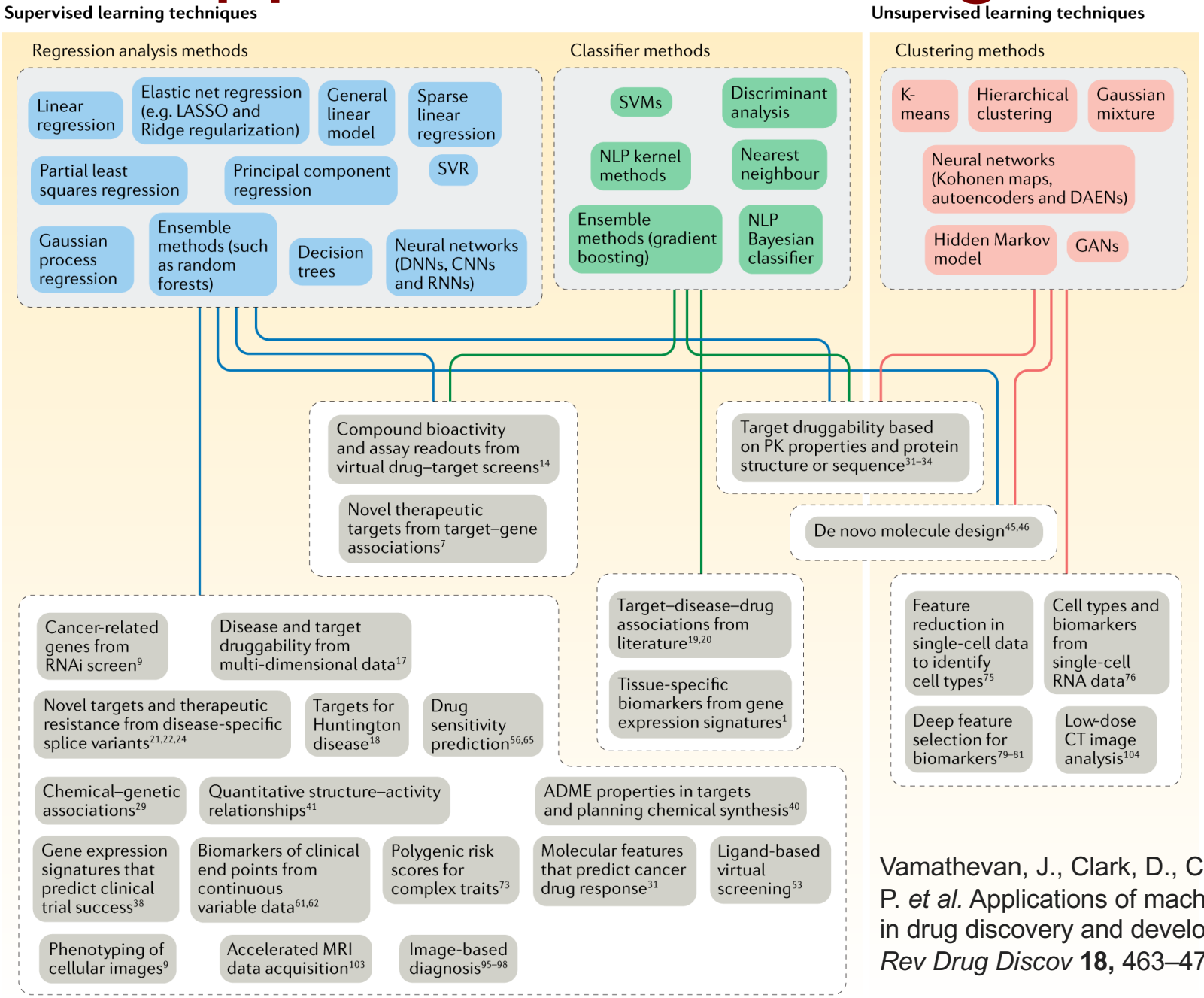
- Previously computer-aided drug design (CADD)
  - Molecular structures
  - Low successful rate (6.2%)
- Increasing data enables AI/ML application
  - Omics
  - Imaging
  - Diagnosis
  - Behaviors

Drug discovery pipeline



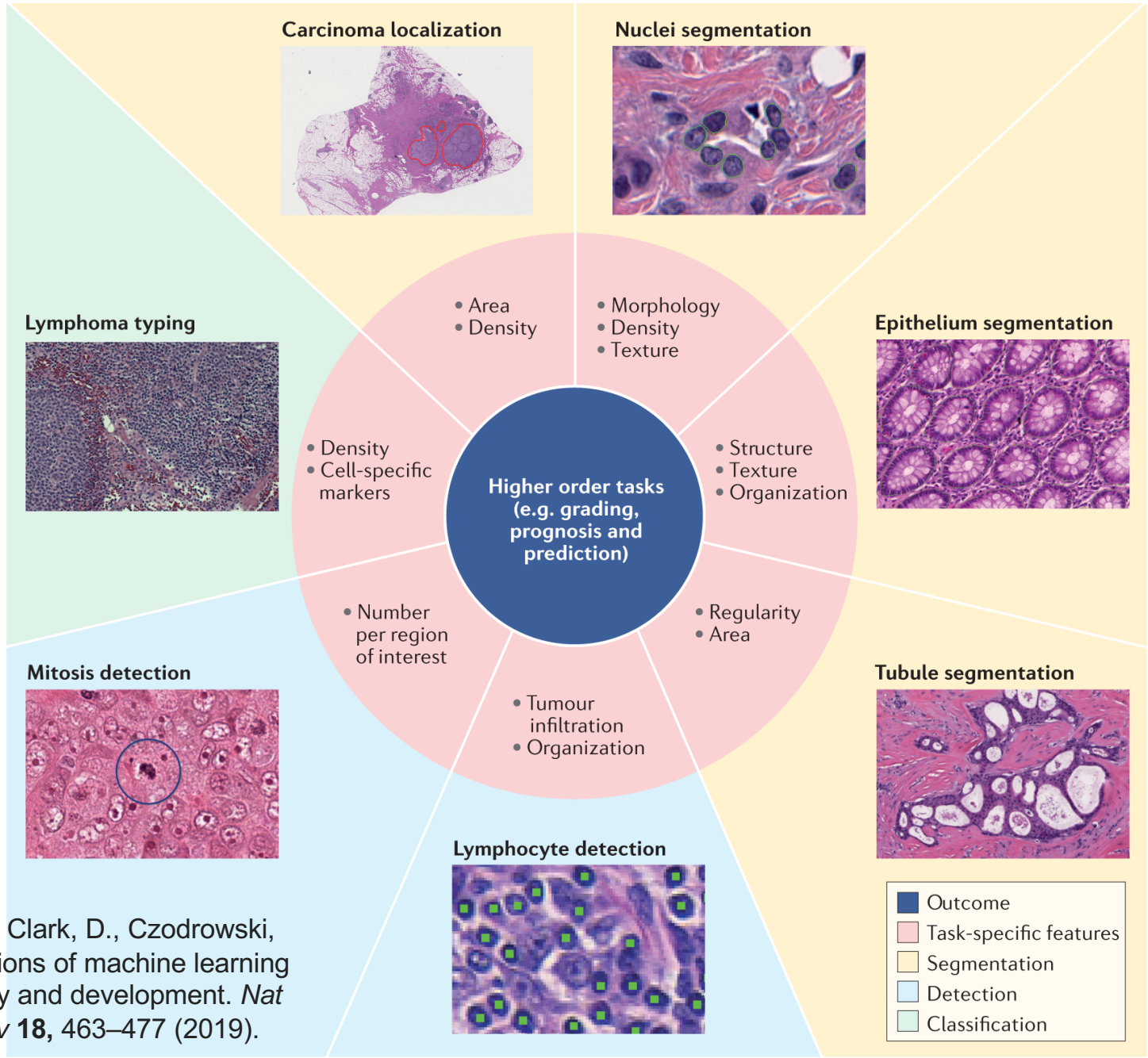
Vamathevan, J., Clark, D., Czodrowski, P. *et al.* Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* **18,** 463–477 (2019).

ML approaches for drug discovery



Vamathevan, J., Clark, D., Czodrowski, P. *et al.* Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* **18,** 463–477 (2019).

Deep learning applications in pathology



Vamathevan, J., Clark, D., Czodrowski, P. *et al.* Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* **18,** 463–477 (2019).

Final exam review

- Network biology
- Applied machine learning
- RNA-seq analysis
- Gene discovery