Thoughts on Perturb-Seq data analysis approaches and pipelines

Kirill Tsukanov
Senior Full Stack Developer & Data Engineer
ktsukanov@ebi.ac.uk

Perturbation Catalogue WP1/2/3 Technical Meeting

2025-05-25

Perturb-Seq data in the context of the Perturbation Catalogue

- For MAVE and CRISPR assays, we are lucky to have curated repositories (MaveDB, DepMap) with good quality, highly processed datasets.
- Perturb-Seq experiments are more complex: repositories like scPerturb aggregate dozens of studies but provide mostly raw expression counts.
- Interpretation and downstream analysis is left to the user, which can be quite complex.

Perturb-Seq data essence

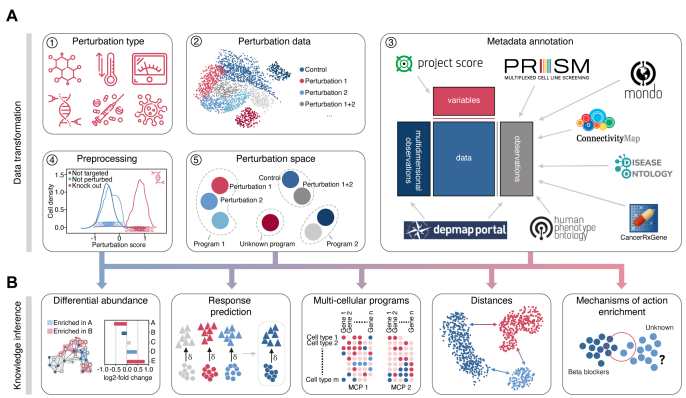
- Each observation is roughly: perturbing *gene X* in a specific cell type/tissue under set conditions yields a given gene expression profile per cell.
- Data characteristics:
  - High noise levels;
  - Pronounced batch effects;
  - Large scale: thousands of cells per perturbation, multiple conditions;
  - Raw counts require normalization, filtering, summarization, enrichment/differential expression/etc.
- Raw Perturb-Seq matrices need some systematic processing to become useful for Perturbation Catalogue users.

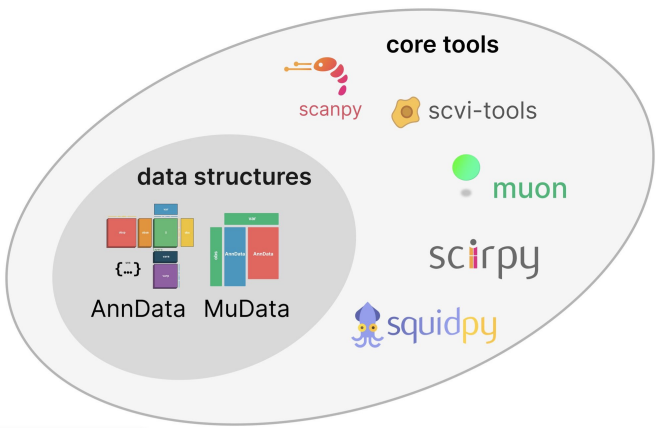
Possible processing approaches for Perturb-Seq data

- Specialized pipelines exist for rigorous statistical analysis:
  - **Python:** MIMOSCA, MAESTRO (with partial AnnData compatibility).
  - **R:** SCEPTRE, Mixscape.
- Challenges:
  - Tools are highly specialized, may require steep learning curves.
  - For many, limited maintenance past the initial publication.
  - Poor compatibility with the broader Python ecosystem for single cell analysis.

scverse ecosystem

- We are a small team and aim to deliver an MVP fast; diving into deep technical pipelines may slow progress.
- **scverse** offers a unified, well maintained, rapidly evolving ecosystem for single-cell analysis.
- Specifically, the **pertpy** tool is designed to handle single-cell perturbation workflows start to end.





Implementation Update: Curated Studies

- **Progress:** 4 studies from scPerturb curated by Aleks - huge thanks!
- Curation process is now well established, unified, and will proceed even quicker in the future.
- Currently all studies have 1 cell type per study; in the future we'll curate more and larger studies.

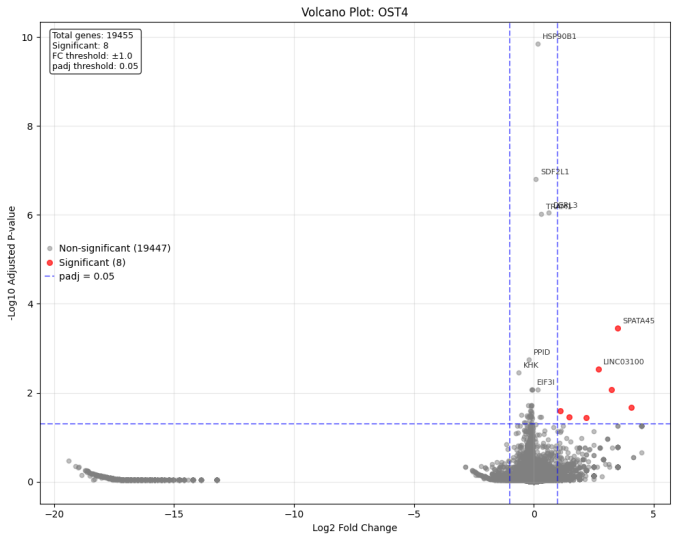
Study	Size	Genes	Cells/Gene	Cell Type
adamson_2016_pilot	117M	7	500	lymphoblast
adamson_2016_upr_epistasis	479M	15	8-1500	lymphoblast
adamson_2016_upr_perturb_seq	1.8G	90	250-750+	lymphoblast
datlinger_2017	132M	32	50-250	T cell

Processing approach: Pseudobulk Differential Expression

- **Input source:** scPerturb harmonised + curated to the common data schema (already implemented).
- **Strategy:** compute pseudobulk differential expression (simple and robust).
- **Workflow:**
  1. Group cells by control vs. perturbation within each cell type.
  2. Aggregate counts to pseudobulk profiles (mean expression per perturbation).
  3. Perform differential expression using t-tests with multiple testing correction.
  4. Apply filtering: adjusted p-value < 0.05, $|\text{log2FC}| > 1$.
- **Implementation:** Parallel processing with expression thresholding and robust statistical testing.

Analysis Results and Filtering

- **User-facing results:**
  - "Perturbing gene X induces significant changes in genes Y, Z..."
  - "Expression of gene A is most strongly altered by perturbations in genes B, C..."
- **Filtering statistics:**
  - Records written: 29,476
  - Records skipped: 2,803,366
  - Filter criteria: padj ≤ 0.05, $|\text{log2FC}| \geq 1.0$



Example volcano plot showing differential expression results for one perturbation.

Next Steps and Future Directions

- **Immediate:**
  - Continue curating additional studies from scPerturb.
  - Optimize processing pipeline for larger datasets.
  - Implement user interface for browsing results.
- **Future enhancements:**
  - Support for multi-cell-type studies.
  - Integration with pathway enrichment analysis.
  - Advanced visualization tools.
  - Cell-type-specific perturbation effects.
- **Questions/Discussion:**
  - Alternative analysis approaches?
  - Priority datasets for curation?