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

2025-04-29

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

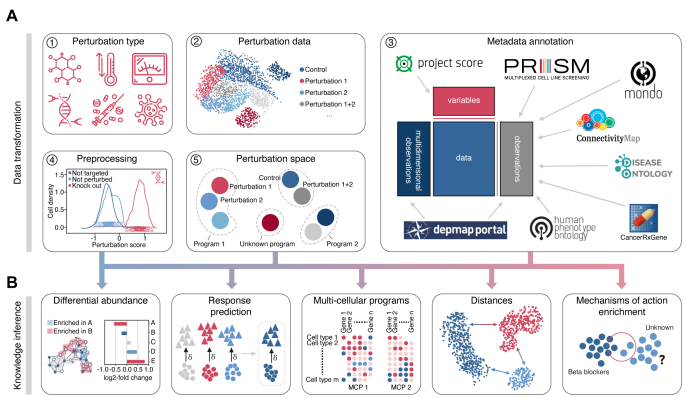
- Each observation is roughly: perturbing *gene X* in a specific cell type/tissue under set conditions yields a given gene expression profile per cell.
- Data characteristics:
  - High noise levels;
  - Pronounced batch effects;
  - Large scale: thousands of cells per perturbation, multiple conditions;
  - 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.

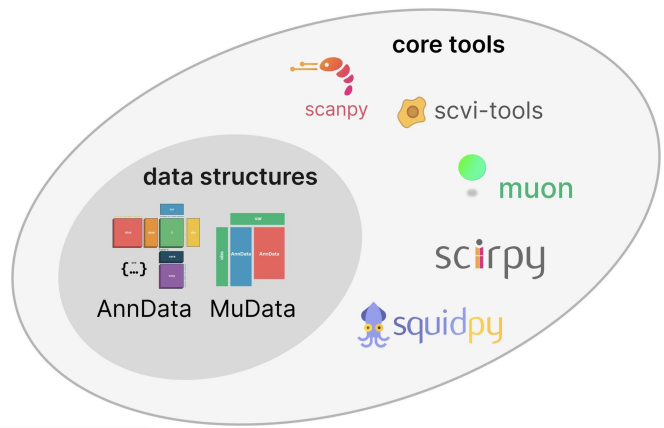
Possible processing approaches for Perturb-Seq data

- 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.





Suggested processing approach for Perturb-Seq

- **Input source:** scPerturb harmonised + curated to the common data schema (already in progress by Aleks).
- **Proposed strategy:** compute pseudobulk differential expression (because simple and robust).
- **Workflow:**
  1. Group cells by control vs. perturbation within each cell type.
  2. Aggregate counts to pseudobulk profiles.
  3. Perform differential expression using `pertpy` facilities.
- **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..."
- Any other ideas?