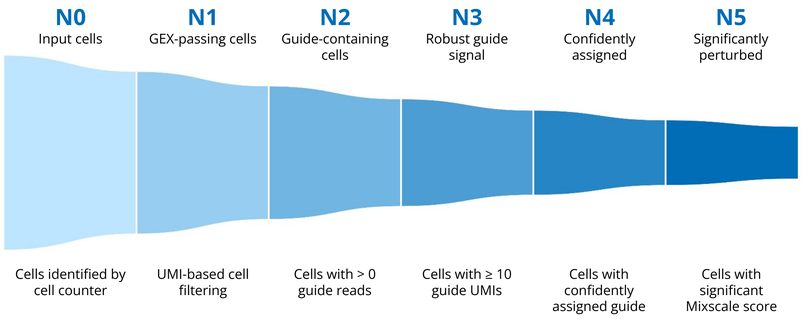
Establishing quality standards to facilitate creation of atlas-sized Perturb-seq datasets

Michael Gatzen, Hannah Rickner, Laura Drepanos, Dan Dubinsky, Yushan Xu, Pabodha Hettige, Yuming Cao, Junjie Xia, Toni Delorey, Elisa Donnard, Kathleen Yates, John Doench, Aziz Al'Khafaji, Sheila Dodge, Niall Lennon, Stacey Gabriel, Eric Lander



Background

- **Perturb-seq** is a powerful functional genomics tool for **understanding the cellular networks and dependencies** underpinning human disease, **combining CRISPR gene editing with a single-cell transcriptomic readout**
- With the emergence of new technologies, **Perturb-seq experiments can now be scaled to unprecedented levels**, enabling cost-effective targeted and genome-wide screens.
- **Consensus on key quality control (QC) criteria for Perturb-seq is lacking** in the field
- A **common understanding of key metrics and widely recognized standards are paramount** for ensuring the integrity of the generated data and the validity of the subsequent foundational models to accelerate biological interpretations
- **Experimental setups vary**, e.g. by cell type, single cell platform, or MOI. **The QC metrics need to be applicable to all modalities**

Perturb-seq at Broad Clinical Labs

- **At Broad Clinical Labs (BCL), we are generating Perturb-seq data for millions of cells per week**, and we anticipate larger and larger datasets to be generated going forward
- **We have several ongoing atlas-sized projects**, including an atlas for academic use, a commercial atlas, and internal cross-laboratory projects
- **We are using multiple single cell platforms** (10x & Illumina PIPseq) and experimental setups (e.g. CRISPR mechanism, MOI, cell hashing)

Highlights

1. **Framework of key process metrics** for quality control of large-scale Perturb-seq screens
2. **Nomenclature** for tracking cell attrition through processing
3. **Standards** applicable to **multiple platforms** (e.g. 10x & Illumina)
4. **Methods** for robust and scalable quantification of metrics
5. **Contribution to harmonization of standards** in the field

Key Process Metrics

Stage	Metric	Target
N0	Number of input cells	Varies by cell model and experimental need
	Cell viability	> 75%
	Sequencer metrics	Q30: > 85% %PF: > 80%
	Expected retention: 50%	
N1	Fraction of GEX reads in passing cells	> 70%
	Fraction of GEX mapped reads	> 80%
	Sequencing saturation	< 60%
	Mitochondrial percentage	< 10%
	Expected retention: 60-80%	
N2	Guide reads per cell	> 0 (useful for troubleshooting)
	Fraction valid guide reads in passing cells	> 80%
	Expected retention: 60-80%	
N3	Guide UMIs per cell	≥ 10
	Expected retention: 60-80%	
N4	Low MOI: Purity-based guide assignment	≥ 10 guide UMIs per cell and ≥ 75% guide purity
	High MOI: GMM-based guide assignment	e.g. Poisson-Gaussian mixture model
	Expected retention: *Varies by experiment*	
N5	Mixscale score of non-control cells	*Varies by experiment*

All values can vary by experimental design or platform.

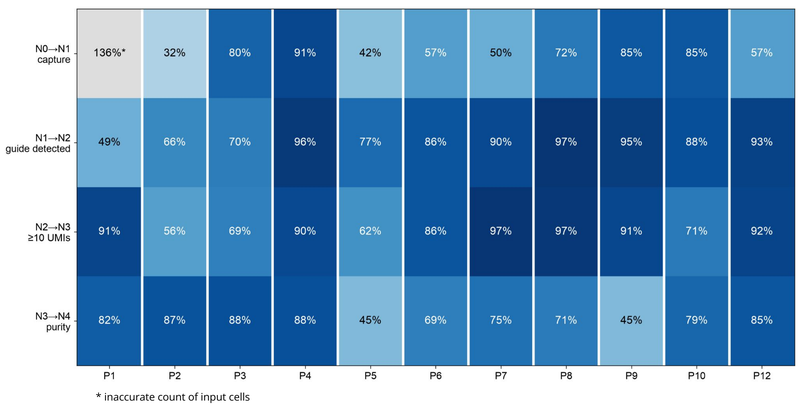
Library QC:

- **GEX:**
  - Concentration: > 5 ng/µL
  - Average fragment size: 450 bp
- **CRISPR**
  - Concentration: > 30 ng/µL

Sequencing depth targets:

- **GEX**: 30,000 transcriptomic reads per cell
- **CRISPR**: 5,000 CRISPR reads per cell

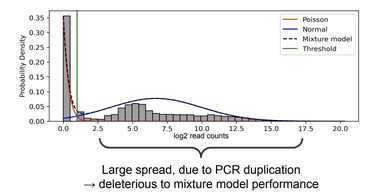
Cell retention (%) across pilot experiments:
Illumina single cell platform

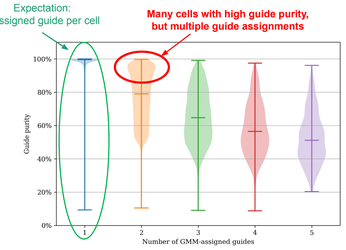


* inaccurate count of input cells

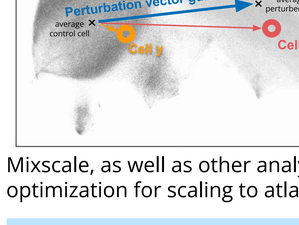
Guide Assignment (N3 → N4)

Typically, guide assignment is performed using Gaussian Mixture Models (GMM). However, GMMs have limitations for low multiplicity of infection (MOI) experiments:





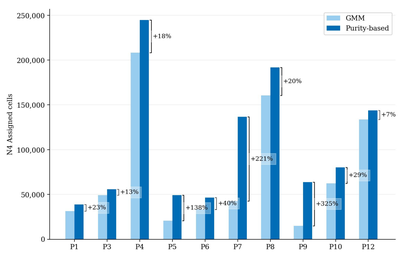
Alternative: **Purity-based guide assignment**

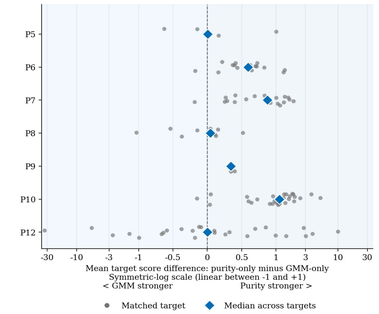


Criteria to assign a cell a guide:

1. Total guide UMIs ≥ 10
   and
2. Guide purity: $n_1 / (n_1 + n_2) \geq 75\%$
   n_1: Most abundant guide in cell
   n_2: Second most abundant guide in cell

Number and quality of cells added by purity-based guide assignment:





For low MOI experiments:
→ **Use purity-based guide assignment**
For high MOI experiments:
→ **Use Gaussian Mixture Model based guide assignment**

Perturbation Quantification (N4 → N5)

Use Mixscale [1] to measure perturbation in differentially expressed gene space:

|| **proj**$_A$ **x** ||: High magnitude
→ significantly perturbed cell

|| **proj**$_A$ **y** ||: Low magnitude
→ unperturbed cell

Mixscale, as well as other analysis software, required computational optimization for scaling to atlas-size datasets.

References

1. Jiang, L., Dalgarno, C., Papalexi, E. et al. Systematic reconstruction of molecular pathway signatures using scalable single-cell perturbation screens. Nat Cell Biol 27, 505–517 (2025). https://doi.org/10.1038/s41556-025-01622-z