Comparative evaluation of genomic footprinting algorithms for predicting transcription factor binding sites in single-cell data

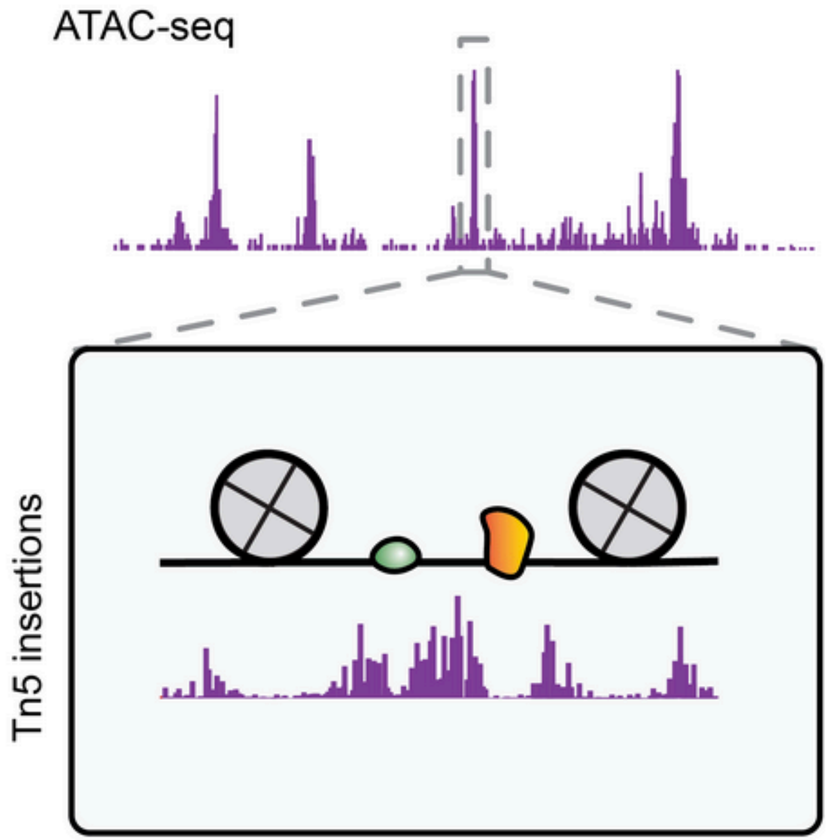
Amanda Everitt, Sean Whalen, Katherine S. Pollard

Chromatin Biology Club

Agata Smialowska 9 Jan 2026

TF footprinting concept

- genome-wide prediction of active TF binding sites (TFBS)
- PWM-centric or signal-centric
- sequence bias correction
- supervised training (ChIP-seq data)



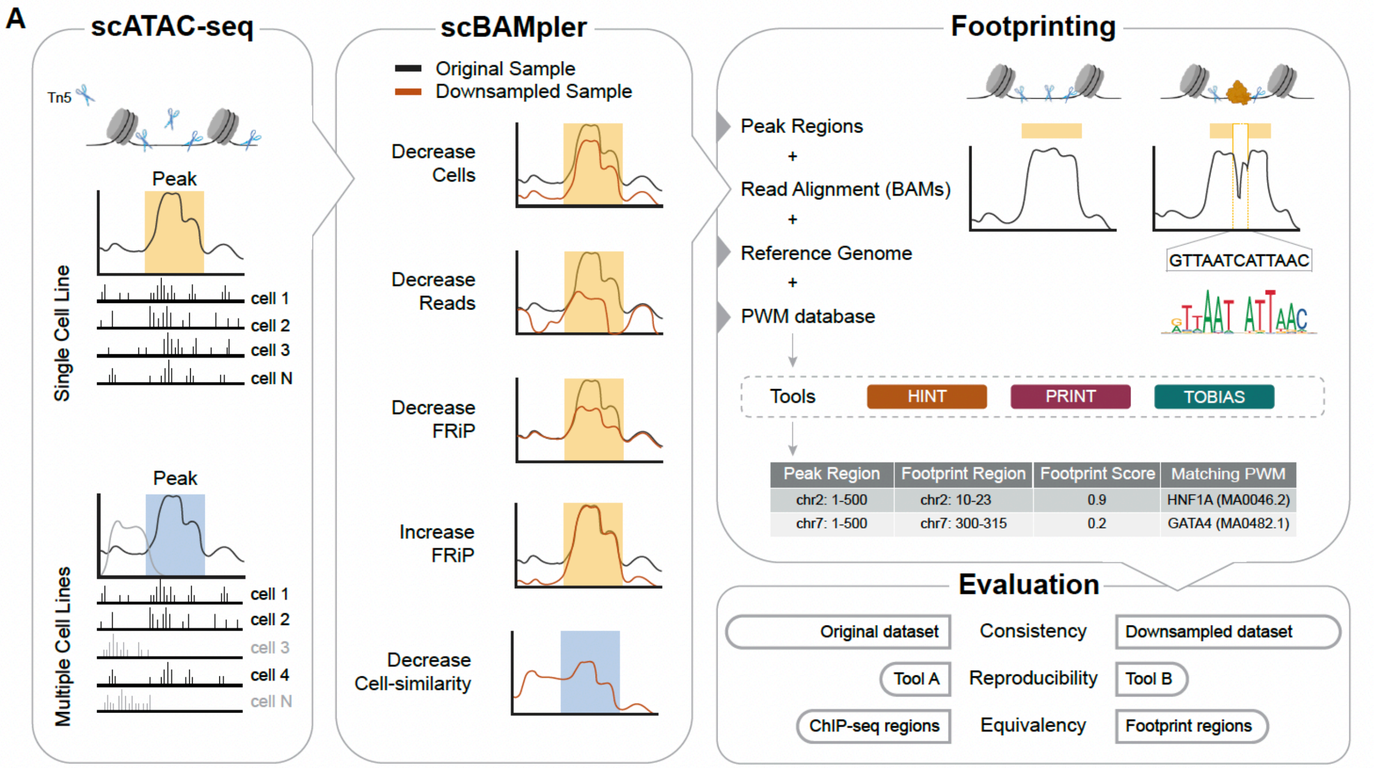
sc-ATAC-seq

- pseudobulking
- (-) data sparsity
- (+) context-dependent relationships between TFs and gene expression can be addressed
- (!) homogeneous vs. heterogeneous clusters (signal-to-noise)

image: https://github.com/buenrostrolab/PRINT

scATACseq TF footprinting benchmarking pipeline: scBAMpler

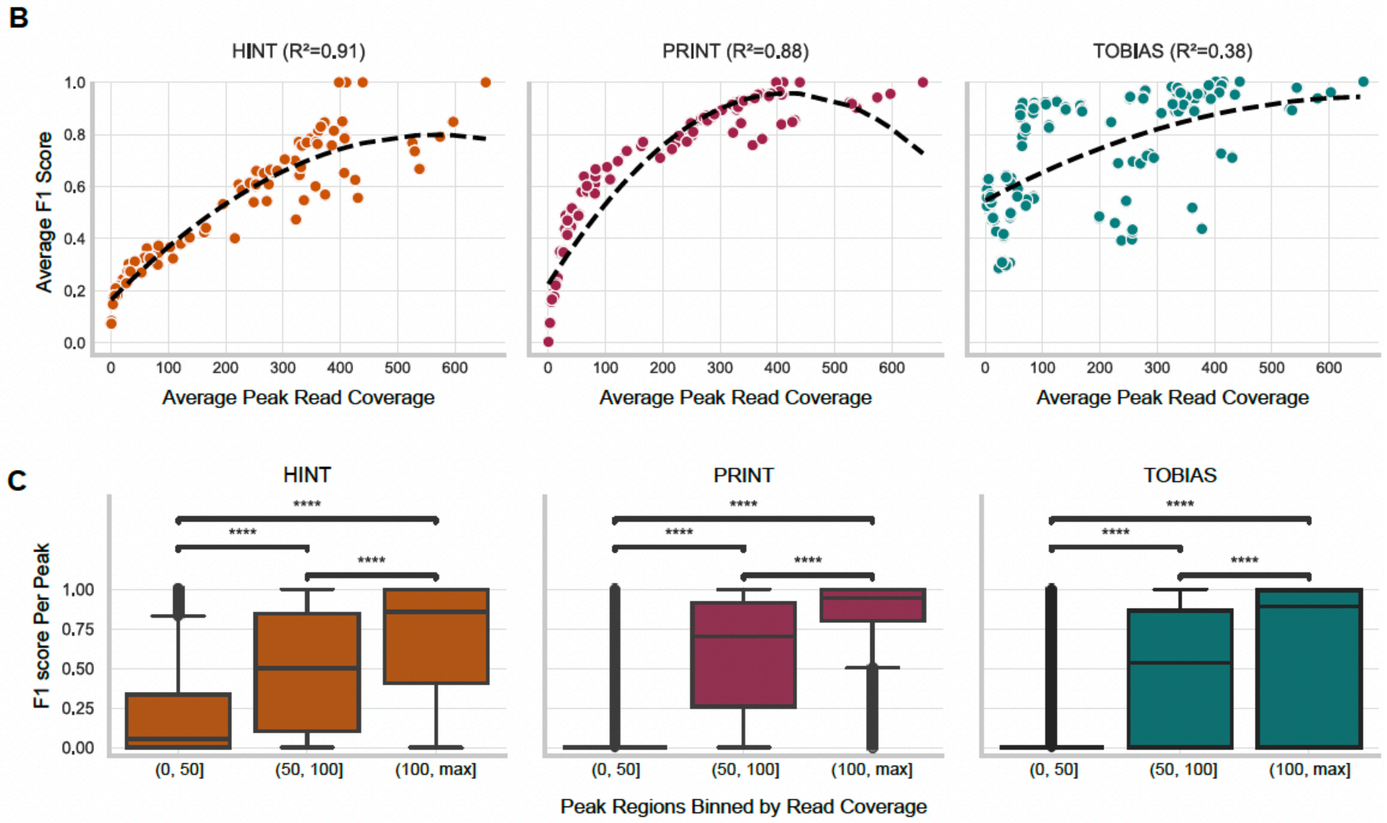
- alignment level downsampling (read count, cell count, FRiP, homogeneity)
- tool evaluation
    - TOBIAS (https://github.com/loosolab/TOBIAS/wiki/) - motif centric; insertion sites
    - PRINT (https://github.com/buenrostrolab/PRINT) - both motif-centric and *de novo* modes (motif mode evaluated); insertion sites
    - HINT (https://reg-gen.readthedocs.io/en/latest/hint/introduction.html) - footprint region w/o motif match; NFR reads



n=3 for each downsampled BAM

total ~400 BAM files

Peak read counts were the strongest factor in footprinting consistency



F1 score is the harmonic mean of precision and recall

0 (no prediction) - 1 (perfect)

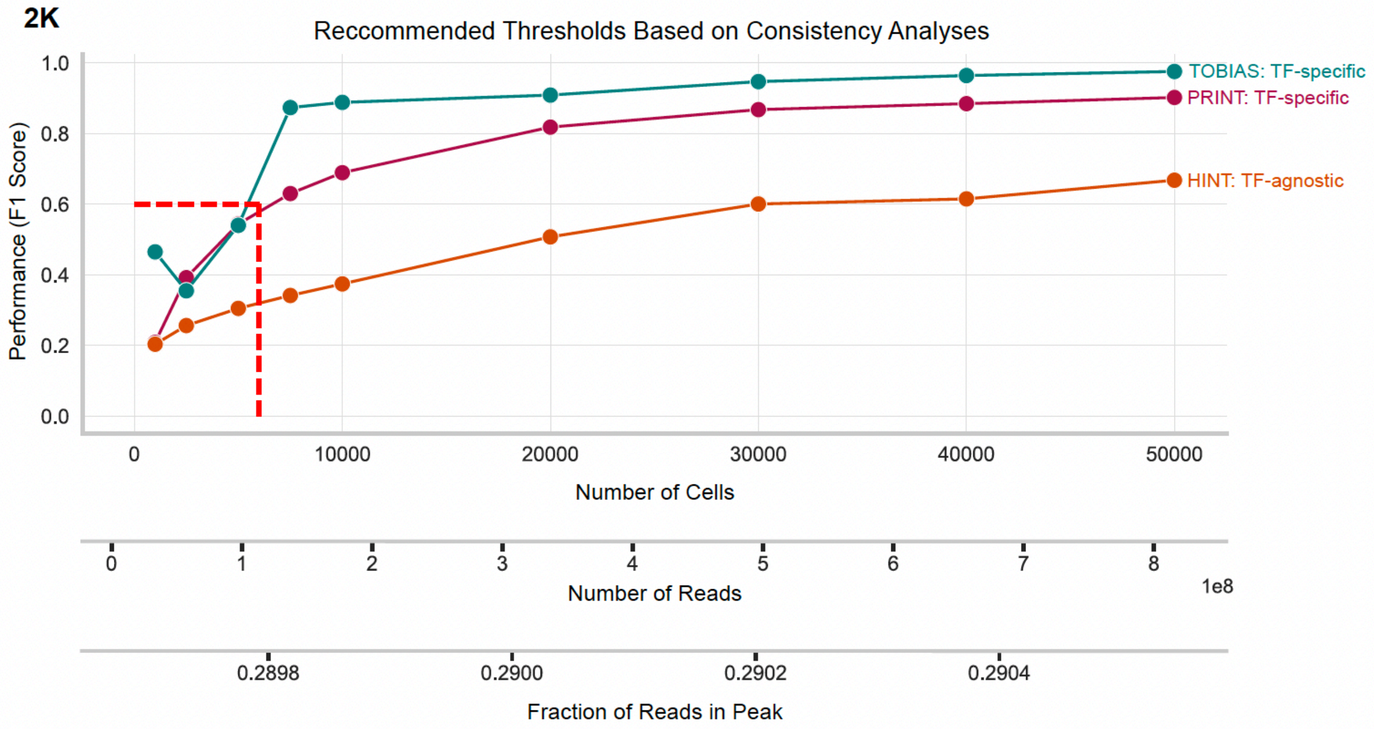
HINT & PRINT lost performance at 5k cells and 1e7 read pairs

peaks with <100 reads had lowest F1 scores

retain peaks with >100 reads

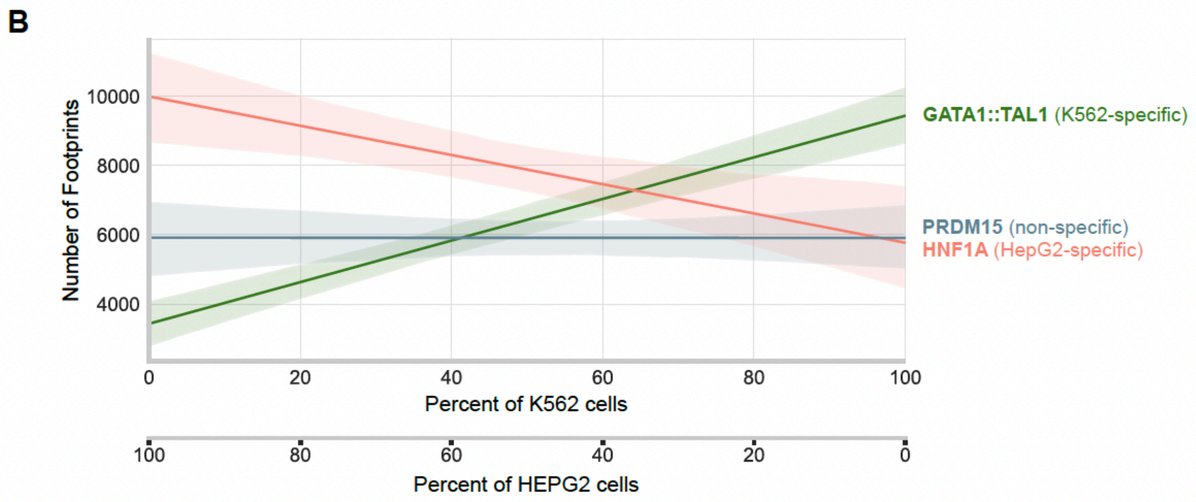
Different tools, different bottlenecks

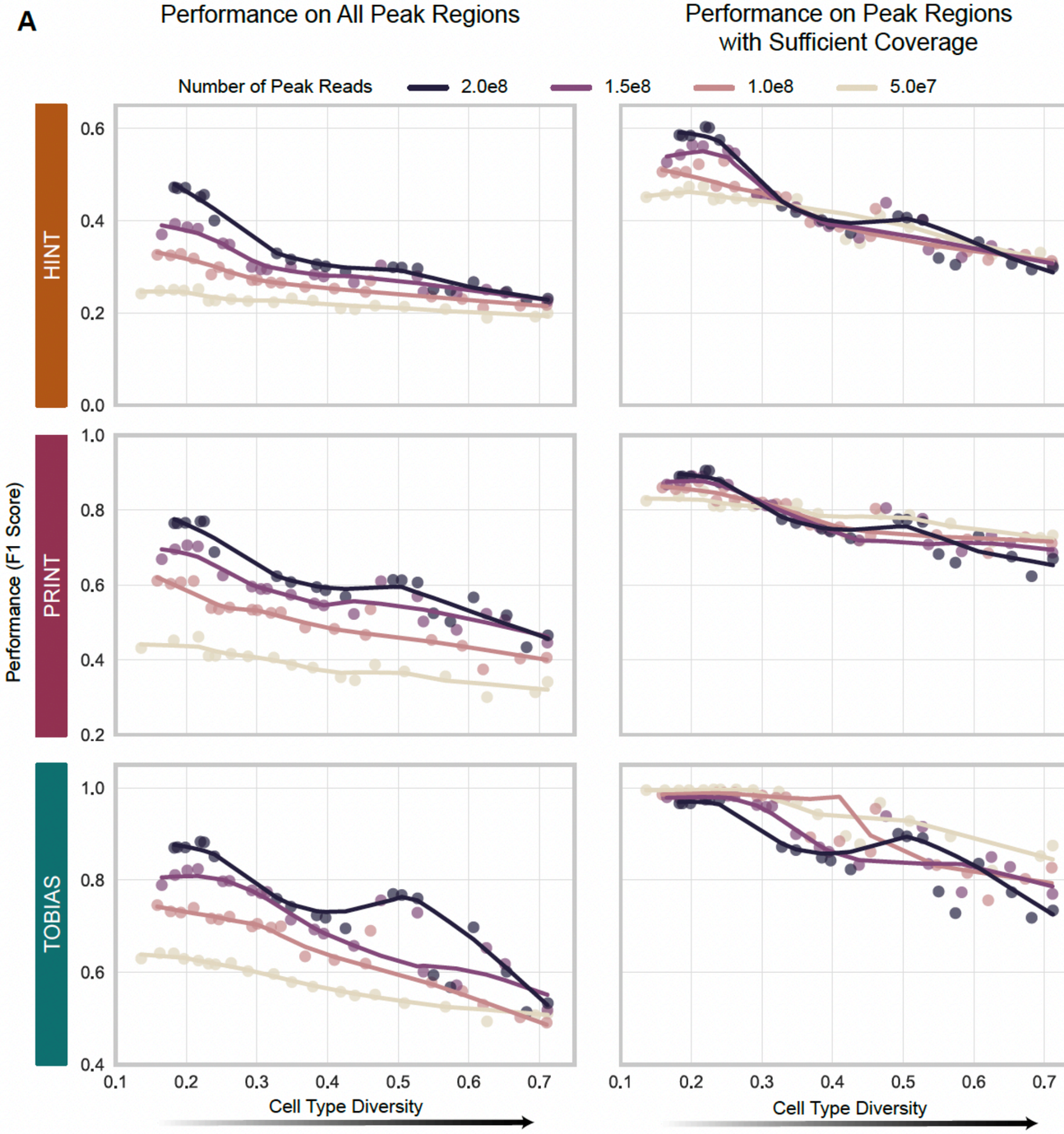
- *de novo* methods (HINT): stable footprint regions, small positional differences and hence resulting PWMs;
- TOBIAS has issues with estimating the *bound threshold* metric from low quality or sparse data; also sensitive to FRiP
- **100M PE reads / cell population** (20k fragments / cell, 0.29 FRiP pseudobulk), ca 6k cells



Cell diversity impacts footprint consistency

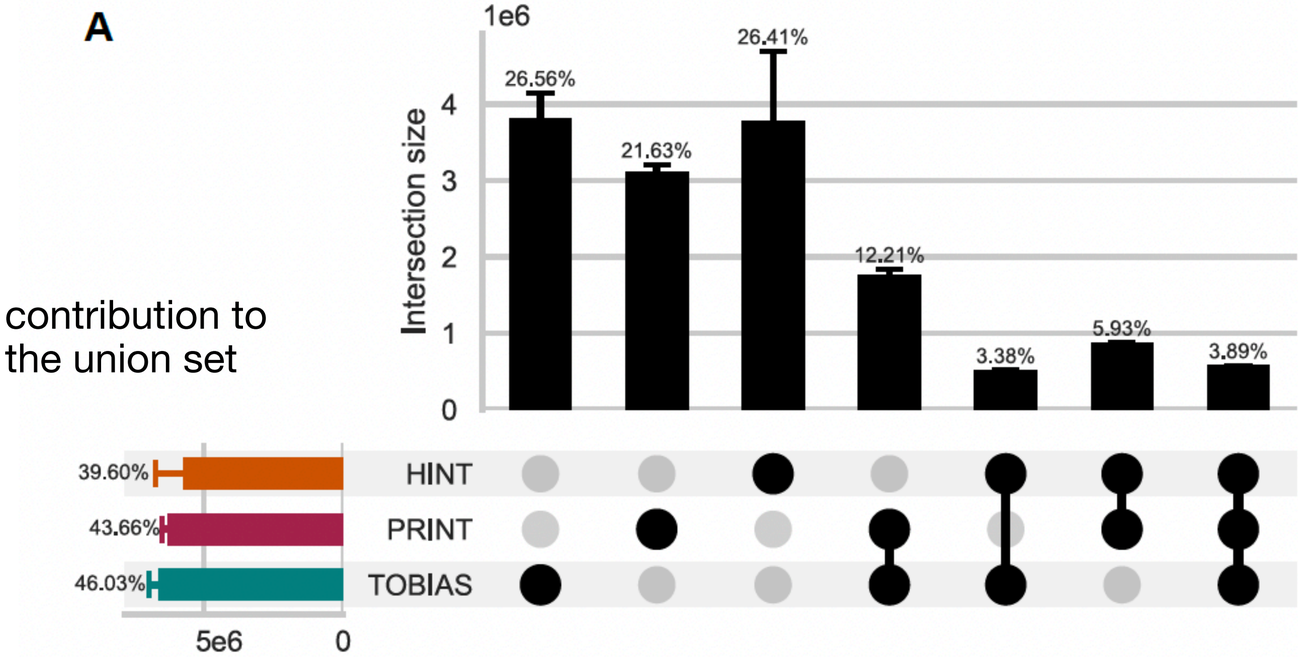
- Pseudobulk heterogeneity obscures footprinting results
- Clusters with different proportion of two cell types show different footprints
- Pseudobulking based on peak count similarity



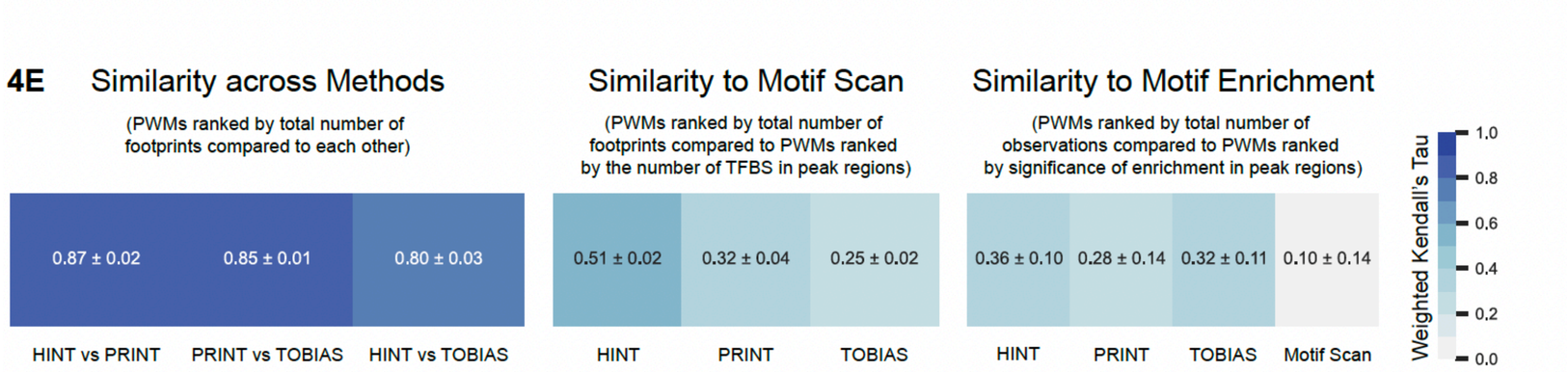


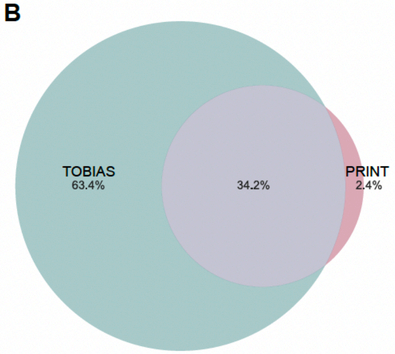
Method consistency

- low % of *footprint regions* shared between three tools; pairwise best between PRINT & TOBIAS
- motif-centric: TFBS candidate selection
- *de novo*: small differences in coordinates result in selection of a (related but different) PWM
- *PWM rankings* (“most frequent TF”) were concordant across tools and downsampling schemes
- PWM rankings across tools were more consistent than rankings based on motif scans (number of potential TFBS in peaks: monaLisa) or motif enrichment scores (HOMER)



PWM presence in the same OCR

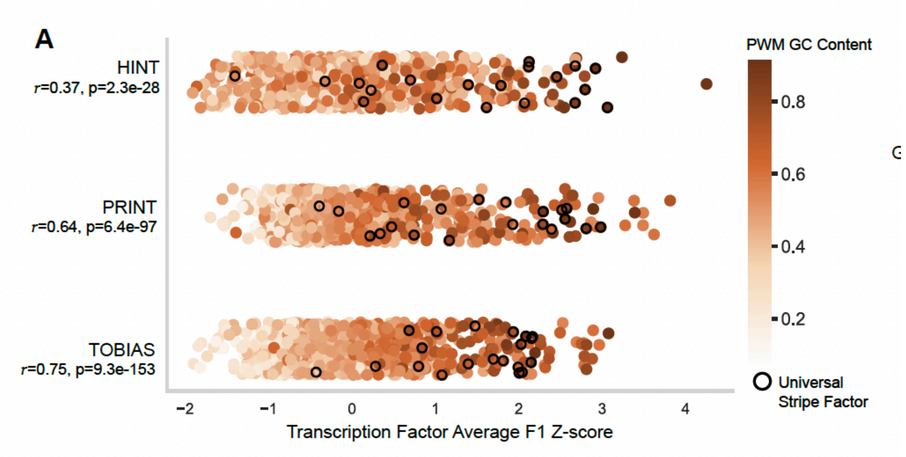


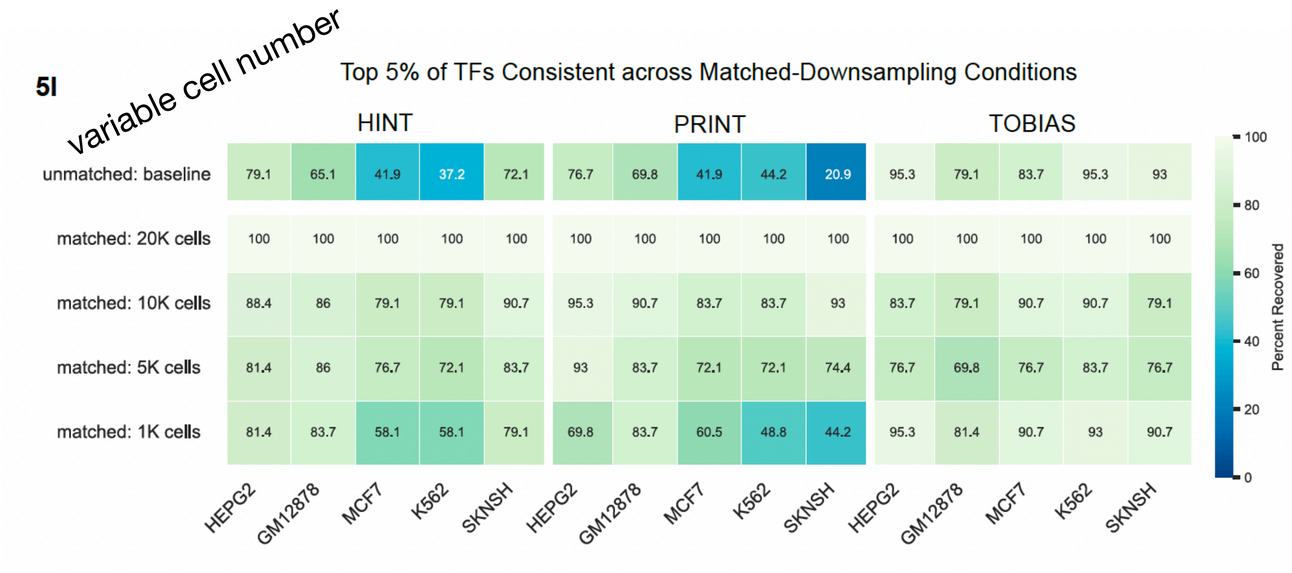


TFBS candidate selection

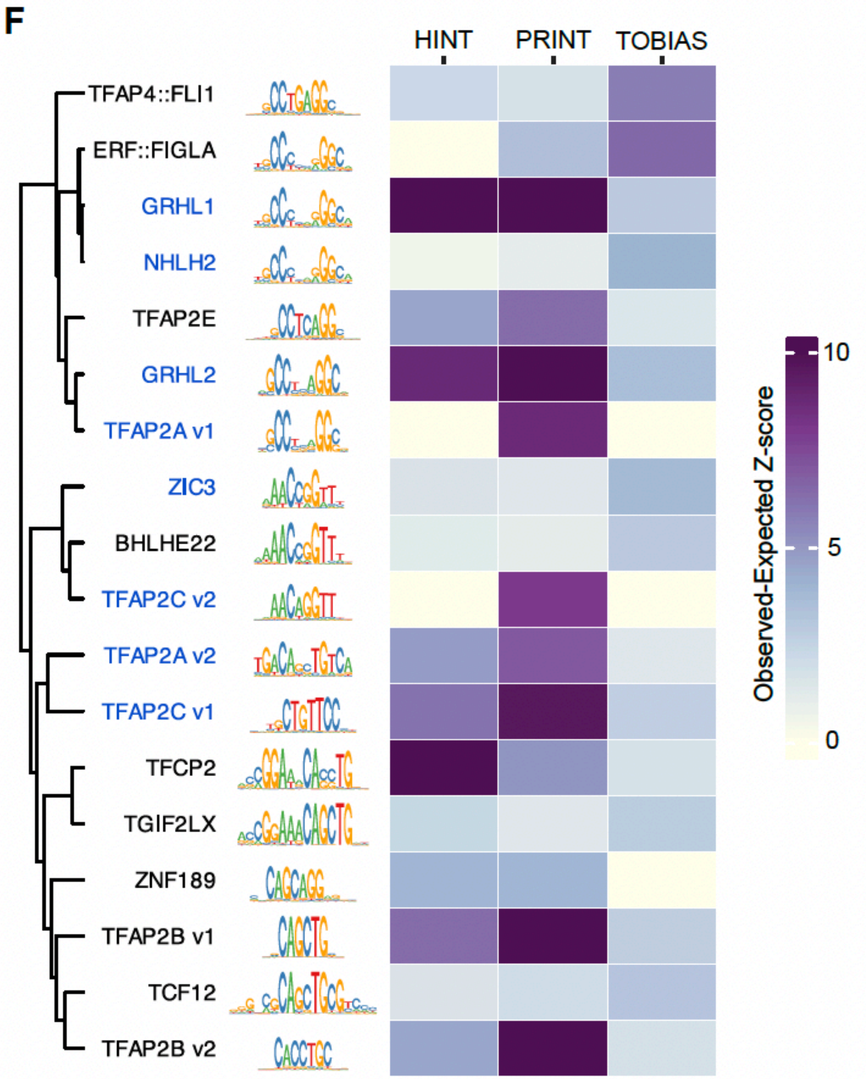
TFBS sequence vs footprinting results

- **GC**, information content, length, …





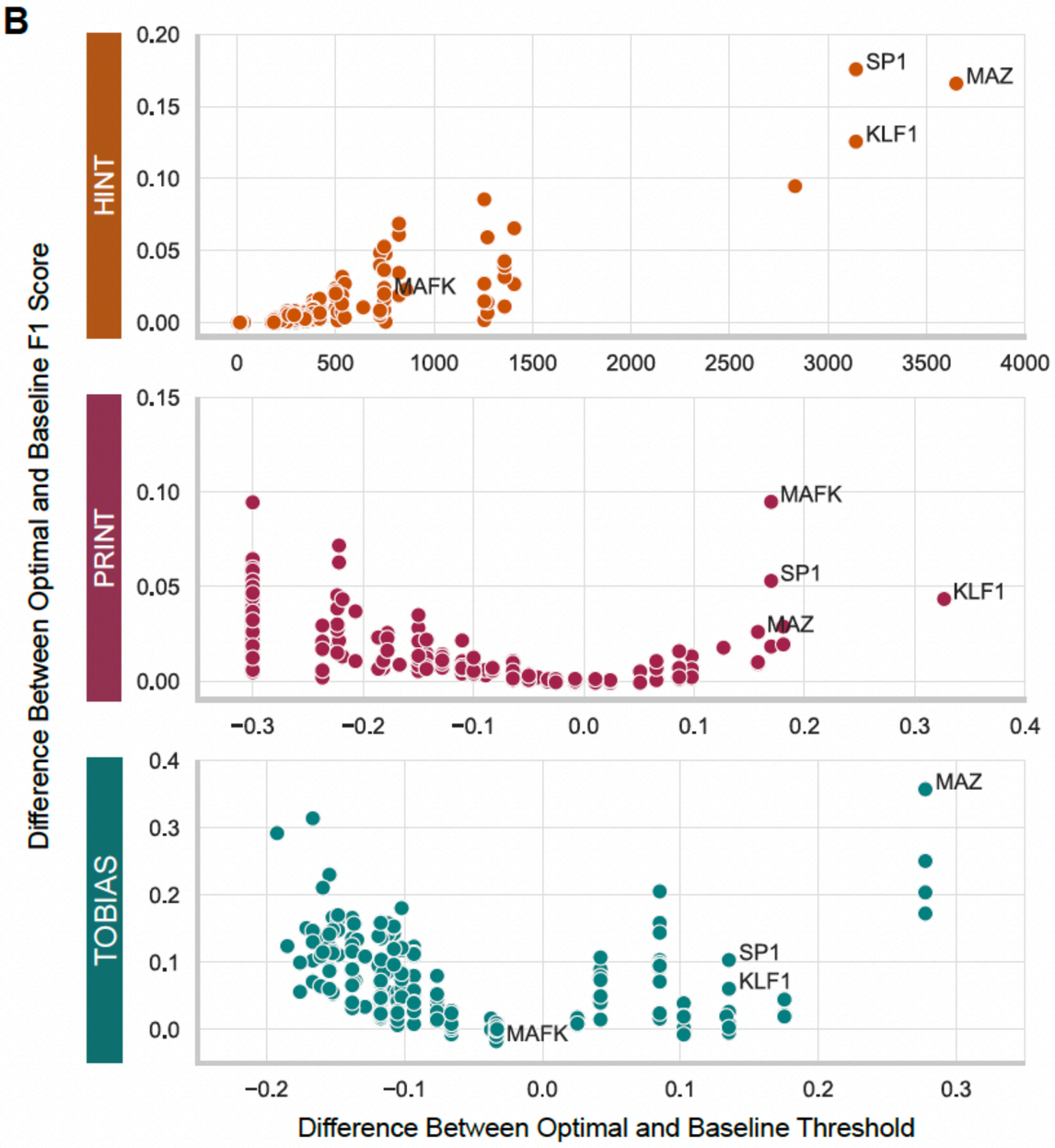
top TFs stable across downsampling conditions



top cell line specific TFs (MCF7)

Also discussed

- differential binding (TOBIAS, HINT)
- ChIP-seq overlaps
    - per TF *bound threshold* tuning
- Cell line effects



Take home message

- ATAC-seq signal quality had greater impact on footprinting performance than PWM quality or tool choice
- read depth (read per peak) dominant factor for footprinting performance (false negatives)
- performance decline was mainly due to increase in false negatives in <100 read peaks
- motif-centric tools seem to be more congruent with ChIP-seq signal (100M read pairs)
- recommendations
    - 100M read pairs / cell population (6k cells)
    - global different TF occupancy across cell populations: matching sample quality; downsampling (by cell count) all cell populations to match the smallest group (or split larger ones to pseudoreplicates)
    - merging similar clusters can be used to boost depth
    - focus on peaks >100 reads in both populations (skip cell type specific peaks): data quality matching less important (low cell number datasets - evaluate well covered regions, reduce high false negative low signal peaks)
    - motif scan more appropriate if: **<1k cells, FRiP < 0.1, low peak-read counts**
    - peak calling: conservative filtering (MACS3 peak summit filter 20)

Thanks for listening

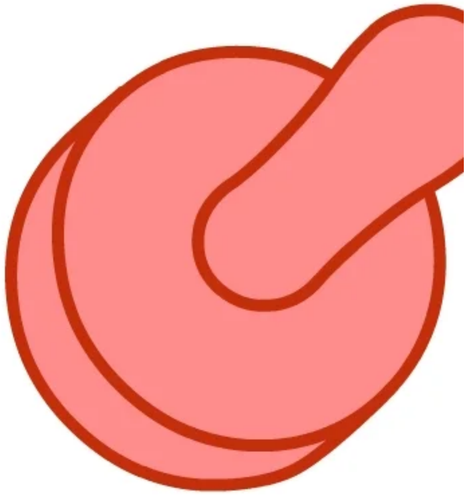


image:giphy.com