Accelerated Population-aware gRNA Candidate Selection Using High-throughput Nomination Methods

Olalekan Usman[1], Kyle J. Kinney[1], Thomas Osborne[1], **Morgan Sturgeon**[1], Andrea Sánchez-Peña[1], Ellen Schmaljohn[1], Bernice Thommandru[1], Karthik Murugan[1],Ariana Mostafa[1], Eugene Wyatt[1], Rolf Turk[1], Matthew McNeill[1], Ashley Jacobi[1], Garrett Rettig[1], & Gavin Kurgan*[1]

[1] Research & Development, Integrated DNA Technologies, Coralville, IA

*Corresponding author: Gavin Kurgan, gkurgan@idtdna.com

Interested in Off-target Services?

Contact us!

UNCOVERseq improves ***in cellulo*** nomination

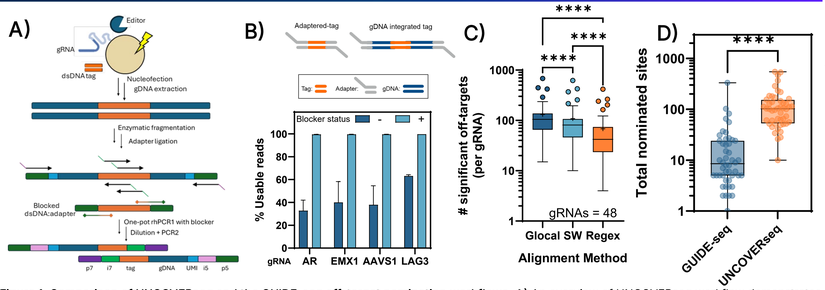

Figure 1. Comparison of UNCOVERseq and the GUIDE-seq off-target nomination workflows. A) An overview of UNCOVERseq workflow demonstrates that cells with a genomically integrated dsDNA tag have gDNA extracted and amplified with rhPCR in a single reaction, with dsDNA tag:adapter byproducts being blocked by a targeted oligo before being sequenced and analyzed using our described workflow. **B)** A depiction is shown regarding what kind of events are targeted by the blocking oligo, with the usable reads (non-dsDNA:adapter reads) measured across nominating off-targets for 4 gRNAs in K562 (n = 3 per gRNA) with (light blue) or without (dark blue) the blocking oligo. **C)** Comparison of different alignment methods used in publicly available nomination packages was performed on UNCOVERseq data to determine differences in levenshtein distance < 7 sites from a glocal alignment (Glocal), Smith-Waterman (SW; -100/-100 gap open/extension penalty as in original GUIDE-seq pipeline) or string-match method (Regex; same as current version "GUIDE-seq" pipeline). **D)** Tukey box plots show the end-to-end differences in total nominated OTE sites for the original GUIDE-seq method (wet lab protocol and alignment) compared with UNCOVERseq. Data is representative of 48 gRNAs (Targets: *PDCD1, LAG3, CTLA4, NRP1, IL2RA*, and *TIGIT*; 8 gRNAs per target). Statistical significance was determined for pairwise comparisons using Wilcoxon matched-pairs signed rank test and for multiple comparisons using a Friedman test with a post-hoc Dunn's test with Bonferroni correction. ****p < 0.0001.

UNCOVERseq outperforms other nomination methods

Figure 2. Comparative analysis of UNCOVERseq to other nomination technologies. Comparison of the sensitivity of other published accounts of nomination technologies to nominate 100% reproducible UNCOVERseq off-target sites using **A)** EMX1 (n = 6 nomination replicates) and **B)** FANCF (n = 6 nomination replicates). **C)** Total number of off-targets selected from variable Severity Bins of each nomination methodology nominated per nomination method for both EMX1 and **D)** FANCF. **E)** Confirmation of nominated off-targets was performed in HEK293–Cas9 (n = 3) to ensure true-positive were captured. Editing quantification at assays with >1,000x coverage. Each dot represents a gRNA on/off-target with the average raw frequency of indels of the control (x-axis) and treatment (y-axis) plotted. Blue dots indicate sites with no statistical significance while orange dots indicate significant sites (p adj<0.05) for EMX1 and **F)** FANCF. **G)** After confirmation, the cumulative frequency of significant off-target events not nominated per method for both gRNAs (n = 2) was quantified in addition to **H)** the total number of significant off-target sites missed. **I)** Method specific nomination frequency was quantified for both confirmed significant and confirmed not significant sites for the full off-target list (n = 2 gRNAs). **J)** Method-specific sensitivity (nominated true positives / confirmed true positives) and specificity (confirmed true positives / confirmed method true + false positives) was calculated for each method across both gRNAs (n = 2).

UNCOVERseq-pools enable high-throughput, sensitive off-target nomination that mirror non-pooled performance

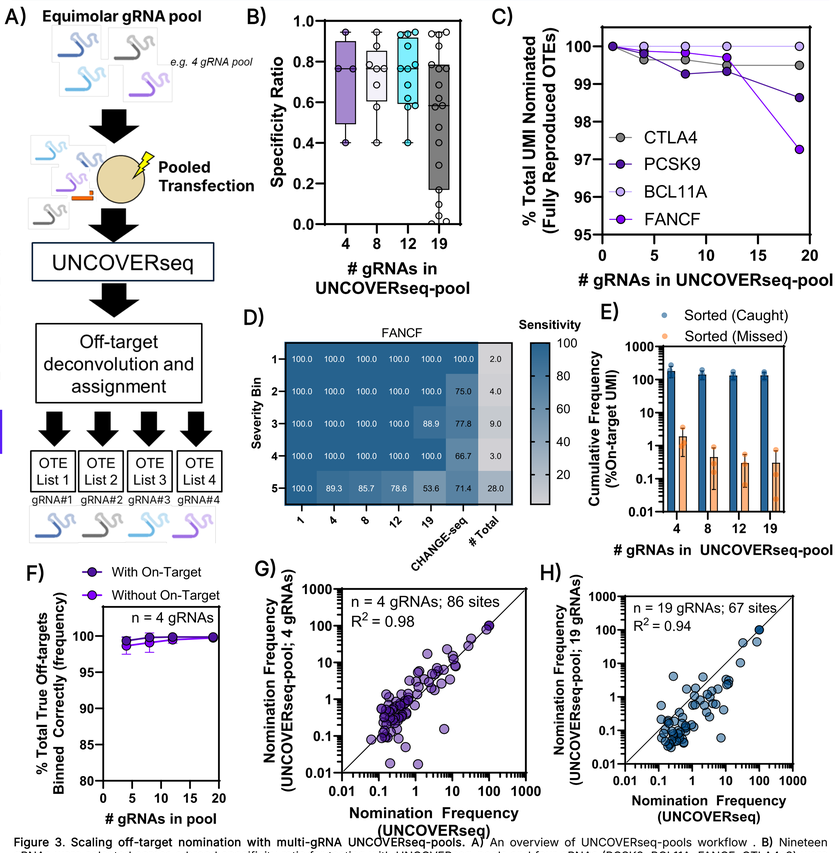

Figure 3. Scaling off-target nomination with multi-gRNA UNCOVERseq-pools. A) An overview of UNCOVERseq-pools workflow . **B)** Nineteen gRNAs were selected across a broad specificity ratio for testing with UNCOVERseq-pools and four gRNAs (PCSK9, BCL11A, FANCF, CTLA4s3) were monitored for performance across dilutions. **C)** To measure performance, the ability to identify fully reproduced UNCOVERseq sites in a non-pooled setting (n = 3 or more per replicates per gRNA) was measured as the % of total UMI frequency covered using the UNCOVERseq-pool strategy across pool sizes. **D)** Sensitivity for detecting the fully reproduced set of sites in each severity bin in HEK293-Cas9 for FANCF is shown for different pool sizes and published CHANGE-seq accounts for FANCF. **E)** To reduce spurious off-targets, off-targets are assigned to each gRNA using an off-target deconvolution algorithm (CATCH) and the total frequency of sites correctly sorted for each of the gRNAs (n =4) is shown along with **F)** the total % of off-targets correctly sorted based on frequency. **G)** To ensure frequency correlation is maintained across pool sizes, we measured the nomination frequency (as % on-target UMI events) for shared sites for the four described gRNAs in a **G)** 4 gRNA and **H)** 19 gRNA pool setting in HEK293-Cas9 (n=3 biological replicates).

Improving upon mismatch-based ***in silico*** nomination is necessary

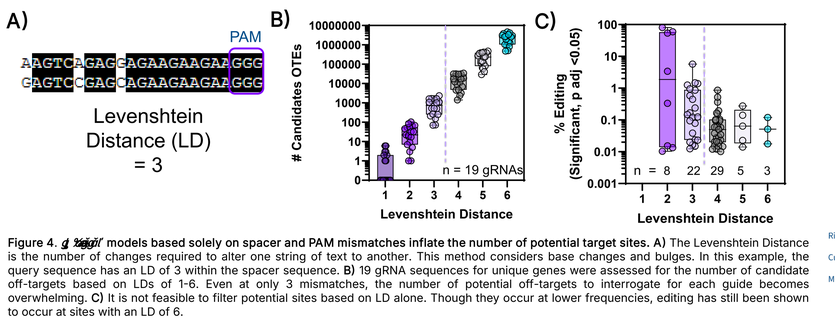

Figure 4. *gRNA* models based solely on spacer and PAM mismatches inflate the number of potential target sites. A) The Levenshtein Distance is the number of changes required to alter one string of text to another. This method considers base changes and bulges. In this example, the query sequence has an LD of 3 within the spacer sequence. **B)** 19 gRNA sequences for unique genes were assessed for the number of candidate off-targets based on LDs of 1-6. Even at only 3 mismatches, the number of potential off-targets to interrogate for each guide becomes overwhelming. **C)** It is not feasible to filter potential sites based on LD alone. Though they occur at lower frequencies, editing has still been shown to occur at sites with an LD of 6.

Comparative nomination performance of UNCOVER-DL

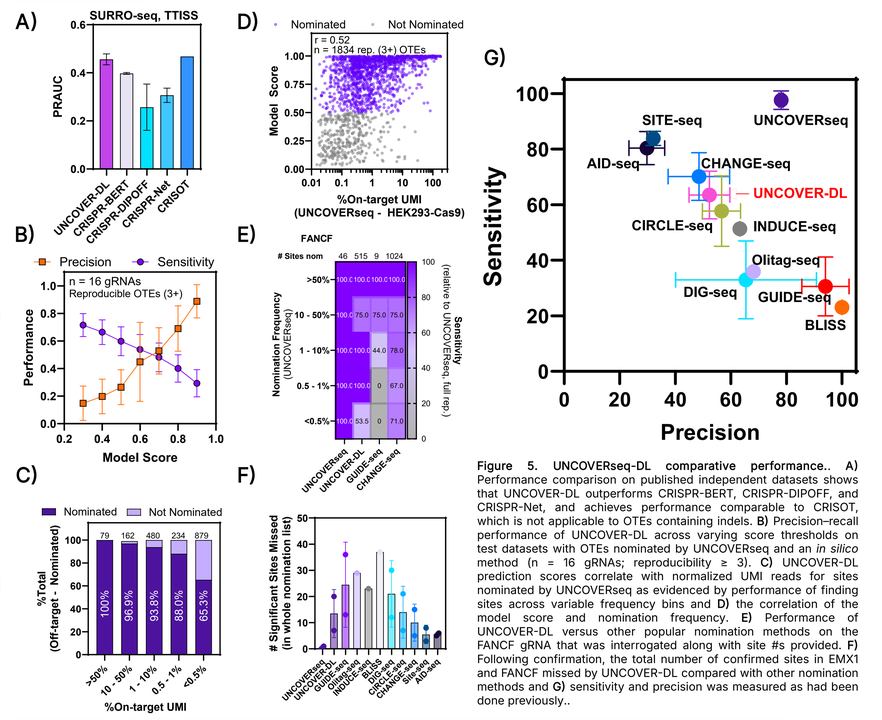

Figure 5. UNCOVERseq-DL comparative performance.. **A)** Performance comparison on published independent datasets shows that UNCOVER-DL outperforms CRISPR-BERT, CRISPR-DIPOFF, and CRISPR-Net, and achieves performance comparable to CRISOT, which is not applicable to OTEs containing indels. **B)** Precision–recall performance of UNCOVER-DL across varying score thresholds on test datasets with OTEs nominated by UNCOVERseq and an *in silico* method (n = 16 gRNAs; reproducibility ≥ 3). **C)** UNCOVER-DL prediction scores correlate with normalized UMI reads for sites nominated by UNCOVERseq as evidenced by performance of finding sites across variable frequency bins and **D)** the correlation of the model score and nomination frequency. **E)** Performance of UNCOVER-DL versus other popular nomination methods on the FANCF gRNA that was interrogated along with site #s provided. **F)** Following confirmation, the total number of confirmed sites in EMX1 and FANCF missed by UNCOVER-DL compared with other nomination methods and **G)** sensitivity and precision was measured as had been done previously..

Population variant-aware OTEs nomination with UNCOVER-DL

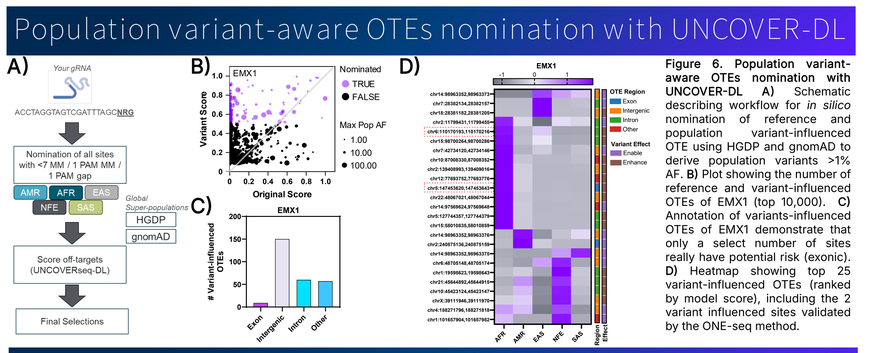

Figure 6. Population variant-aware OTEs nomination with UNCOVER-DL A) Schematic describing workflow for *in silico* nomination of reference and population variant-influenced OTE using HGDP and gnomAD to derive population variants >1% AF. **B)** Plot showing the number of reference and variant-influenced OTEs of EMX1 (top 10,000). **C)** Annotation of variants-influenced OTEs of EMX1 demonstrate that only a select number of sites really have potential risk (exonic). **D)** Heatmap showing top 25 variant-influenced OTEs (ranked by model score), including the 2 variant influenced sites validated by the ONE-seq method.

Rapid in silico risk assessment is possible using UNCOVER-DL

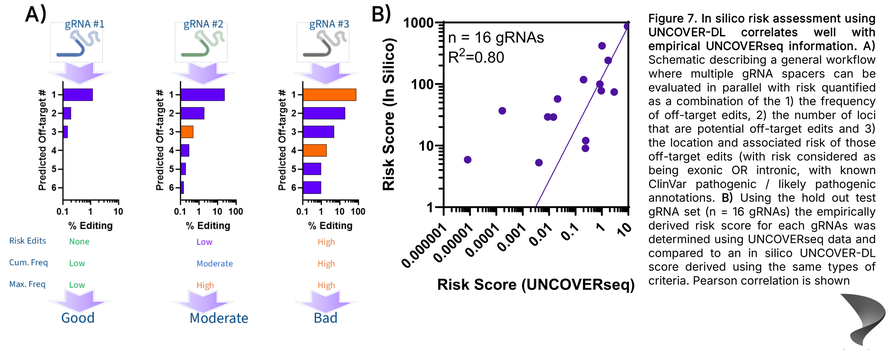

Figure 7. In silico risk assessment using UNCOVER-DL correlates well with empirical UNCOVERseq information. A) Schematic describing a general workflow where multiple gRNA spacers can be evaluated in parallel with risk quantified as a combination of the 1) the frequency of off-target edits, 2) the number of loci that are potential off-target edits and 3) the location and associated risk of those off-target edits (with risk considered as being exonic OR intronic, with known ClinVar pathogenic / likely pathogenic annotations. **B)** Using the hold out test gRNA set (n = 16 gRNAs) the empirically derived risk score for each gRNAs was determined using UNCOVERseq data and compared to an in silico UNCOVER-DL score derived using the same types of criteria. Pearson correlation is shown