GenoPipe: Identifying the genotype of origin within (epi)genomic datasets

Olivia Lang, Divyanshi Srivastava, B. Franklin Pugh, and William KM Lai

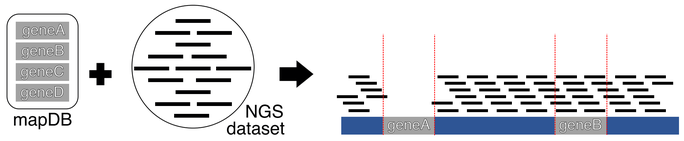
Check out the tool!

Abstract

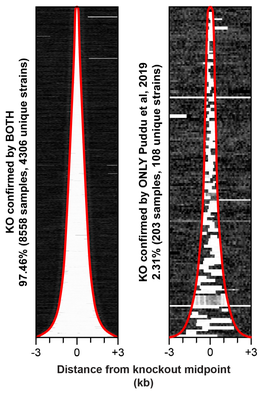
Confidence in experimental results is critical for discovery. As the scale of data generation in genomics has grown exponentially, experimental error has likely kept pace despite the best efforts of many laboratories. Technical mistakes can and do occur at nearly every stage of a genomics assay (i.e., cell line contamination, reagent swapping, tube mislabelling, etc.) and are often difficult to identify post-execution. However, the DNA sequenced in genomic experiments contains certain markers (e.g., indels) encoded within and can often be ascertained forensically from experimental datasets. We developed the Genotype validation Pipeline (GenoPipe), a suite of heuristic tools that operate together directly on raw and aligned sequencing data from individual high-throughput sequencing experiments to characterize the underlying genome of the source material. We demonstrate how GenoPipe validates and rescues erroneously annotated experiments by identifying unique markers inherent to an organism's genome (i.e., epitope insertions, gene deletions, and SNPs).

DeletionID

A powerful genetic modification of small genome organisms that researchers have used for the past few decades is full gene knockouts **(1)**. The DeletionID module surveys a large set of genomic intervals (e.g. gene coordinate annotations) and identifies intervals with significant depletion over the median coverage of the interval set. This allows the user to confirm genetic backgrounds from samples with whole gene knockouts.



Large scale detection of deletions from the Yeast Knockout Collection (YKOC)

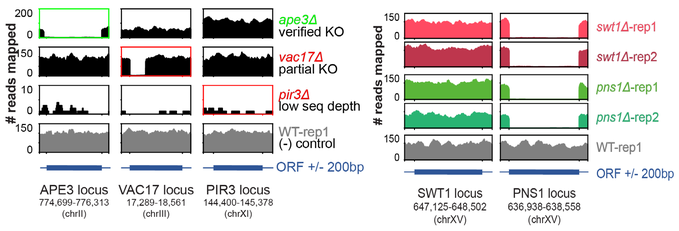


Over 9,000 validated samples of whole-genome sequencing (WGS) data from over 4,000 unique whole-gene knockouts were run through DeletionID **(2)**. The samples that confirmed the labeled deletion are visualized in the left heatmap while those not confirmed by DeletionID are shown on the right heatmap. Each row shows the read coverage of one sample centered on the expected gene knockout region. The rows are sorted by the length of the expected gene knockout region with the gene knockout regions outlined by the red line.

These heatmaps visualize the read coverage of each sample to show that DeletionID can flag incomplete knockout strains in real datasets at a large scale. Most of the samples flagged by DeletionID were discordant annotations of knockout region intervals among other reasons explored more deeply in the panel below.

Detection of deletions fails for discordant annotations or low sequencing depth

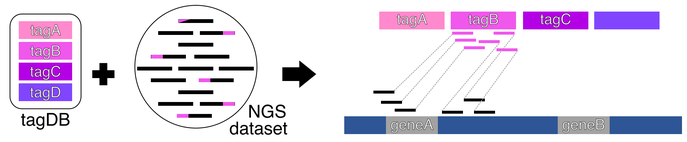
(Left) There are several factors to consider when DeletionID "fails" to detect a gene/genomic interval deletion. The figure above shows the read coverage from several examples including a "clean"/successful identification of a gene interval depletion (APE3), two samples for which DeletionID was unable to detect the knockout due to low sequencing coverage (VAC17) or the gene deletion annotation being discordant with the actual deletion interval (PIR3), a wild type no knockout control sample, and the gene annotations within each of the three genomic loci shown.



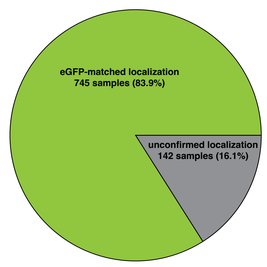
(Right) The first two rows show the read coverage for two sequencing replicates purportedly from strains with the SWT1 gene knocked out (shades of pink). DeletionID did not identify the expecte SWT1 knockout and instead called a PNS1 knockout. Below these two samples are two replicates from strains with PNS1 knocked out (shades of green) as positive controls, a replicate from a wildtype background (gray) as a negative control, and the annotated reference coordinates of the SWT1 and PNS1 genes.

EpitopeID

A popular genetic tool in contemporary research is the construction of strains with long insertion sequences like those encoding protein epitope tags **(3)**. The EpitopeID module identifies the presence of epitopes through alignment to user-provided epitope sequences (including the expected sequence of interest). For paired-end data, the inserted sequence can then be localized by determining where the mate-pairs of sequences that align to the inserted sequence map to the genome.



Localization of eGFP and 3xFLAG fusion protein backgrounds in ENCODE samples

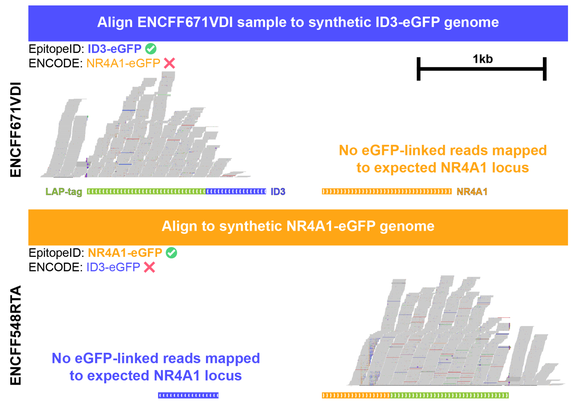


We ran EpitopeID on thousands of ChIP-seq datasets from ENCODE that were generated from strains with eGFP or 3xFLAG-tagged target backgrounds **(4)**. Across all single-end and paired-end eGFP datasets (1,150) and all 3xFLAG datasets that were all paired-end (984), only 8 total samples did not contain the epitope as reported by EpitopeID. In several samples this could be explained by low sequencing depth while others may warrant further quality control investigation.

Of the paired-end data for which EpitopeID could localize the eGFP tag to the gene target, EpitopeID successfully identified 745/887 datasets (83.9%) as matching the target gene metadata from ENCODE. Investigation of these samples with conflicting eGFP gene targets indicates that many of these samples are potentially mislabeled or that the epitope was localized to an off-target region while others may be unconfirmed simply due to insufficient sequencing depth.

Example eGFP fusion protein metadata swap between ENCODE samples

The two samples shown below are examples of a likely mixed up or mislabeling somewhere in the upstream processing steps. ENCFF671VDI is purportedly from an eGFP tagged NR4A1 genetic background and ENCFF548RTA is purportedly from an eGFP tagged ID3 genetic background. EpitopeID did not localize the eGFP epitope to a the respective loci matching the ENCODE metadata.



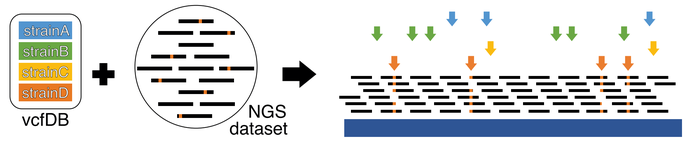
Instead, the EpitopeID report indicates epitope-tagged gene targets that would swap the target gene metadta between the samples. The browser shots above show the paired end read alignments to a synthetic ID3-eGFP genome (top) and alignments to a synthetic NR4A1-eGFP genome (bottom) where at least one read in each read pair displayed maps to the eGFP epitope sequence (LAP-tag).

Both datasets show alignment of reads to eGFP (LAP-tag), confirming the presence of the epitope. The top sample shows continuous alignment across the ID3-eGFP locus and no reads in the NR4A1 locus, despite being labled as having an NR4A1-eGFP genetic background. Conversely, the bottom sample shows continuous alignment across the NR4A1-eGFP locus and no reads in the ID3 locus, despite being labeled as having an ID3-eGFP genetic background. This demonstrates support for a mislabeling of the actual ID3-eGFP sample as "NR4A1-eGFP" and vise versa, validating the results of the calls made by EpitopeID.

This mixup explanation is further supported by other replicates submitted with simlar timestamps (presumably processed together) showing similar EpitopeID mix-up patterns while other replicates with different timestamps (presumably processed separately) show "correct" metadata patterns (EpitopeID identifies other NR4A1-eGFP samples as NR4A1-eGFP).

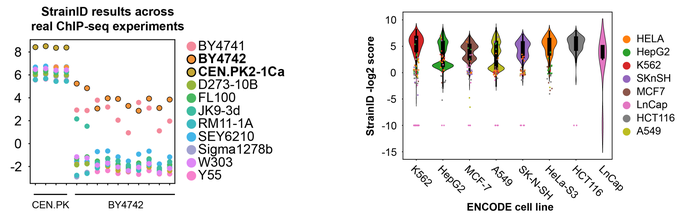
StrainID

Cell line contamination has been a challenge since the 1970s when the first human cell line was cultured and this continues to be raised as a concern in discussions of reproducibility in research **(5-9)**. We built StrainID to perform a substitution SNP-based identification of cell lines by leveraging VCF files of known variants (there are publicly available options for common cell lines). A score is calculated for each profile in a set of VCF files (i.e. cell line) based on tallies of reads containing alternate alleles, reads containing reference alleles, and a sampled background score of alternate and reference alleles **(10)**. The best (highest) score indicates the cell line variant profile with the best match.



Identification of yeast strain background in published ChIP-seq datasets

(Left) We tested the strain background of ChIP-seq samples from published small scale yeast studies and demonstrated StrainID identifies the appropriate strain background in every sample tested **(11-12)**. For large scale yeast studies, our lab has relies on GenoPipe to perform strain validation of thousands of ChIP-exo samples and has supported the work of several publications **(13-14)**.



(Right) We also ran StrainID on over ~13,000 ENCODE samples from various cell lines and matched its metadata to the cell line with the best StrainID score to flag samples with potential contamination or mislabeling. Samples with the best StrainID score conflicitng with its labeled background are represented by dots vertically aligned with the labeled cell line and colored by the cell line with the best StrainID score.

Runtime Performance

During the validation for this tool, we determined the minimum paired-end sequencing depth for reliable detection of each type of genetic background and the runtime performance across a series of sequencing depths. The average runtime performance of each module at the recommended sequencing depth using default human or yeast databases is as follows:

	DB	runtime	DB	runtime
DeletionID	sacCer3_Del (3M)	~1m 00s	*not recommended*	-
EpitopeID	sacCer3_EpiID (100K)	~0m 03s	hg19_EpiID (20M)	~4m 00s
StrainID	sacCer3_VCF (1M)	~0m 30s	hg19_VCF (1M)	~1m 30s

Citations

1. Giaever, G. and Nislow, C. (2014) The yeast deletion collection: a decade of functional genomics. Genetics, 197, 451-465.
2. Puddu, F., Herzog, M., Selivanova, A., Wang, S., Zhu, J., Klein-Lavi, S., Gordon, M., Meirman, R., Millan-Zambrano, G., Ayestaran, I. et al. (2019) Genome architecture and stability in the Saccharomyces cerevisiae knockout collection. Nature, 573, 416-420.
3. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. and Church, G.M. (2013) RNA-guided human genome engineering via Cas9. Science, 339, 823-826.
4. Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K. et al. (2020) New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res, 48, D882-D889.
5. Nelson-Rees, W.A., Daniels, D.W. and Flandermeyer, R.R. (1981) Cross-contamination of cells in culture. Science, 212, 446-452.
6. Masters, J.R. (2002) HeLa cells 50 years on: the good, the bad and the ugly. Nat Rev Cancer, 2, 315-319.
7. Horbach, S. and Halffman, W. (2017) The ghosts of HeLa: How cell line misidentification contaminates the scientific literature. PLoS One, 12, e0186281.
8. Health, N.I.o. (2007) Notice Regarding Authentication of Cultured Cell Lines.
9. Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P. et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res, 22, 1813-1831.
10. Song, G., Balakrishnan, R., Binkley, G., Costanzo, M.C., Dalusag, K., Demeter, J., Engel, S., Hellerstedt, S.T., Karra, K., Hitz, B.C. et al. (2016) Integration of new alternative reference strain genome sequences into the Saccharomyces genome database. Database (Oxford), 2016.
11. de Jonge, W.J., O'Duibhir, E., Lijnzaad, P., van Leenen, D., Groot Koerkamp, M.J., Kemmeren, P. and Holstege, F.C. (2017) Molecular mechanisms that distinguish TFIID housekeeping from regulatable SAGA promoters. EMBO J, 36, 274-290.
12. Cai, L., McCormick, M.A., Kennedy, B.K. and Tu, B.P. (2013) Integration of multiple nutrient cues and regulation of lifespan by ribosomal transcription factor Ifh1. Cell Rep, 4, 1063-1071.
13. Rossi, M.J., Kuntala, P.K., Lai, W.K.M., Yamada, N., Badjatia, N., Mittal, C., Kuzu, G., Bocklund, K., Farrell, N.P., Blanda, T.R. et al. (2021) A high-resolution protein architecture of the budding yeast genome. Nature, 592, 309-314.
14. Mittal, C., Lang, O., Lai, W.K.M., Pugh, B.F. (2022) An integrated SAGA and TFIID PIC assembly pathway selective for poised and induced promoters. Genes Dev, 36(17-18), 985-1001.

Funding

This poster presentation was supported by the National Science Foundation (NSF), the International Society for Computational Biology (ISCB), and the GLBIO23 Conference, and through the GLBIO23 Travel Fellowship. This work was supported by the National Institutes of Health (NIH) grant [R01ES034353] to BFP. This work was also supported by a National Science Foundation XSEDE and subsequent ACCESS award [BIO220026] to WKML.