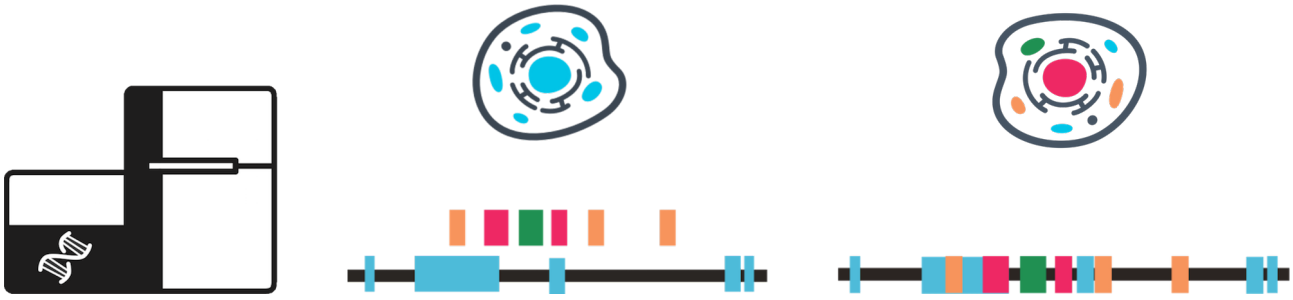
Model Characterization Improves the CRISPR Experimental Process

AJ Ajetunmobi

September 30, 2016



Model characterization is the process of sequencing an animal or cell model prior to carrying out gene editing experiments. Sequencing the experimental model is important for optimizing CRISPR guide design and analysis. This is a pivotal step in ensuring comparable analysis and clinical reproducibility across experimental datasets. This resource will outline the benefits of using model-specific genome data for guide design instead of relying on the reference genome.

Model-Specific Data Reveals Key Information

Imagine you've just landed at the airport of a city you last visited 10 years ago and you need to drive across town to check into your hotel. Unfortunately, the rental car has no GPS and your old school flip phone lacks a data connection. Your only option is to use the map you picked up the last time you were in town over a decade ago. Suddenly, a 30 minute trip spirals into a three-hour voyage of dead-ends, roadblocks and more than one near-death experience.

This scenario, while whimsical, is comparable to how many CRISPR labs carry out experiments. When it comes to experimental design, the status quo is to use outdated reference genome data instead of a model-specific genome. The assumption is that genetic variations between the reference genome and their investigative model are negligible with limited impact on design and analysis parameters. However, published literature as well as our own in-house data is beginning to challenge this dogma.

Whole genome sequencing has been used to characterize genome edits in several studies. One 2014 paper by Smith et al. compared five human induced pluripotent stem cell (hiPSC) lines to the human reference genome, Hg19. They found ≥4.2 million SNVs and >500,000 indels in both parental (BC1) and edited cell lines versus the reference. Further, they found 200-300 SNVs between the edited cell lines and BC1, none of which could be attributed to off-target cleavage.

In another study by Veres et al., differences between parental cell lines and alleged isogenic clones amounted to ~100 SNVs per clone, only 2-5% of which could be attributed

to nuclease activity. They suggest that more “rigorous” studies in the future would require whole genome sequencing for the model itself in order to ensure that mutations are due to genome edits and not inherent clone-to-clone mutations.

Whole Genome Sequencing Improves Guide Design

A pivotal study by Yang et al. 2014 demonstrated the importance of characterizing experimental models prior to performing a gene editing experiment. The team used whole genome and deep sequencing techniques to characterize *Streptococcus pyogenes* Cas9 (SpCas9) specificity in hiPSCs. While whole genome sequencing data generally showed low off-target SpCas9 activity, they discovered a germline single nucleotide variant (SNV) that creates a recurrent off-target site in their model.

Specifically, the PGP1 hiPSC line used in the study harbors a heterozygous G-C SNV labeled “Chr5_OT.” According to the reference (Hg19) genome, the targeting sgRNA used in the experiment has a potential off-target binding site with a three base pair mismatch at positions 11, 15 and 19. However, in the PGP1 genome, the SNV alters the mismatch at position 11. This reduces the number of mismatches to two base pairs on one of the Chr5_OT alleles. As a result, off-target cleavage at this site is significantly more likely in the model genome than in the reference genome.

The next step for Wang et al. was to verify this assumption by quantifying the frequency of the off-target event. This was accomplished through targeted amplicon deep sequencing analysis. They found that the variant allele with two mismatches had a 36.7% indel frequency while the reference allele had a 1% indel frequency. This highlights how a single germline SNV can create a recurrent off-target site undetected by *in silico* predictions based on reference genome sequence data.

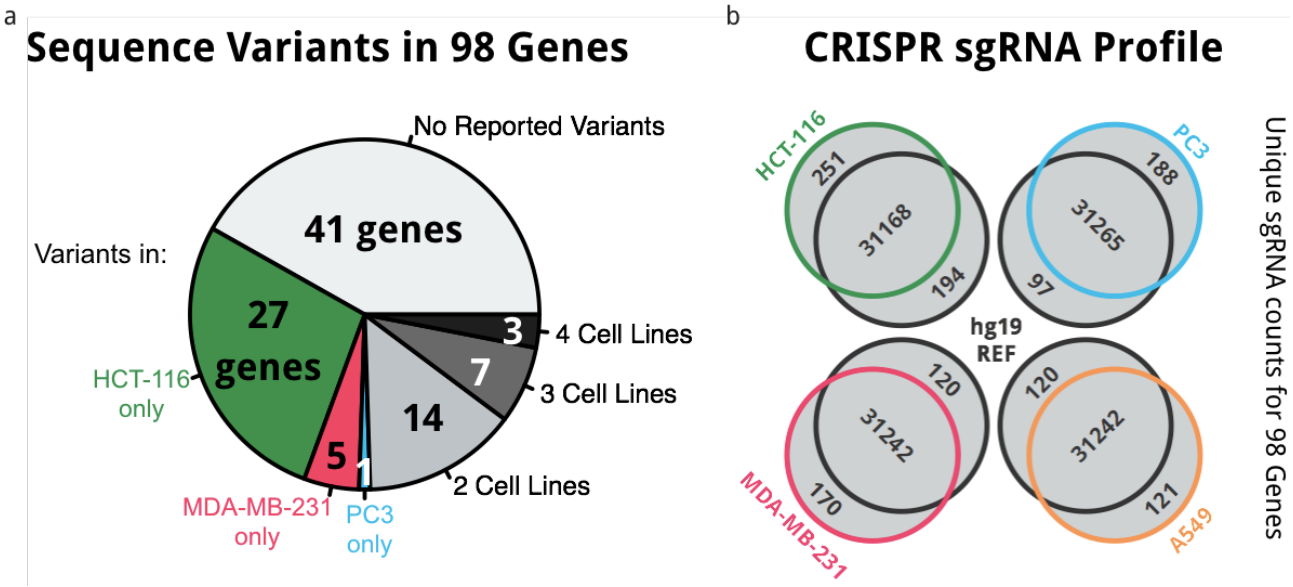


Figure 1. a) Describes overlapping variants in 98 genes across four human cell lines. b) Shows guide RNAs targeting 98 genes in each cell line and whether, depending on variants in the model genome, those guides overlap with hg19.

In an effort to further quantify the impact of SNVs on guide design and potential off-target effects, we carried out our own internal analysis. We investigated whole genome sequencing data on four patient-derived cancer cell lines and compared the change in guide behavior to reference data. We investigated unique guides targeting the whole genome and found that ~2% were impacted by SNVs in cell lines compared to the reference genome (Figure 1). This means that approximately 1 in 50 guides designed against the reference genome would have an altered or inaccurate predicted activity and specificity profile.

Moving CRISPR Toward Clinical Application

As CRISPR becomes more ubiquitous, there is a need to develop standardized approaches to assessing and quantifying activity and specificity. By employing whole genome sequencing and a model-specific guide RNA design prior to gene editing experiments, the accuracy and the reproducibility of experimental datasets can be improved. Further, concrete causal relationships between genotype and phenotype can be ascertained if the edits are validated at the nucleotide level.

Although some cell lines and animal models have been characterized with whole genome sequencing, many have not. Even cell line genome data may not match the clones used in the experiment. This means that an investigator using CRISPR in a disease model won't have all of the information necessary to inform quality guide design. In many instances, they are left using the human reference genome which may not match up with the specific characteristics (e.g. SNVs) of the model (Figure 2).

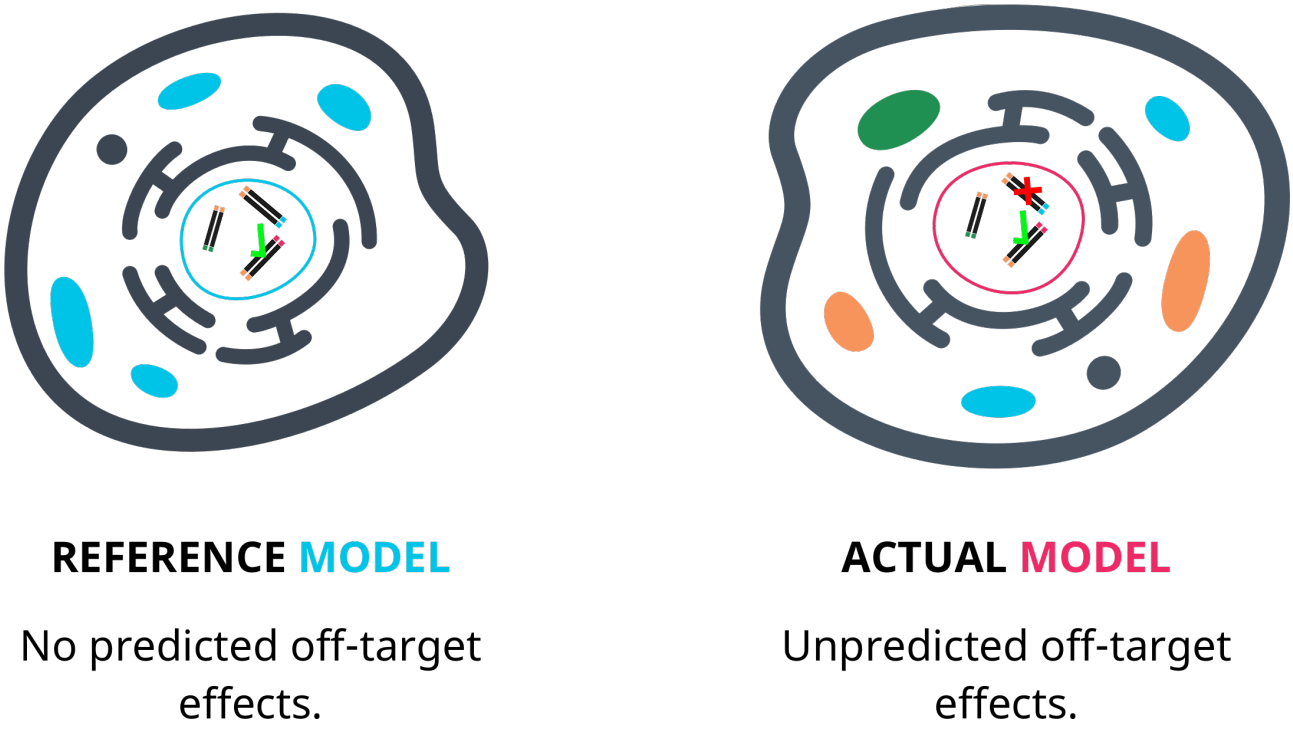


Figure 2. Experimental models may contain SNVs which lead to off-target effects. The SNVs present in the experimental model are not present in the reference and therefore remain undetected using prediction algorithms.

If an investigator uses whole genome sequencing to evaluate their model prior to an editing experiment, the benefits are twofold. First, they can engage in superior sgRNA design as aforementioned. Beyond that, they can use the initial sequencing data as a reference to compare against the edited model genome. This enables verification of CRISPR-mediated changes and confident assessment of the specificity of Cas9 in therapeutic applications.

Sources

Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov. 2011 Aug 31;10(9):712. doi: 10.1038/nrd3439-c1. PubMed PMID: 21892149.

Smith C, Gore A, Yan W et al. Whole-genome sequencing analysis reveals high specificity of CRISPR/Cas9 and TALEN-based genome editing in human iPSCs. Cell Stem Cell. 2014 Jul 3;15(1):12-3. doi: 10.1016/j.stem.2014.06.011. PubMed PMID: 24996165.

Veres A, Gosis BS, Ding Q et al. Low incidence of off-target mutations in individual CRISPR-Cas9 and TALEN targeted human stem cell clones detected by whole-genome sequencing. Cell Stem Cell. 2014 Jul 3;15(1):27-30. doi: 10.1016/j.stem.2014.04.020. Erratum in: Cell Stem Cell. 2014 Aug 7;15(2):254. Cowan, Chad A [added]. PubMed PMID: 24996167.

Yang L, Grishin D, Wang G et al. Targeted and genome-wide sequencing reveal single nucleotide variations impacting specificity of Cas9 in human stem cells. Nat Commun. 2014 Nov 26;5:5507. doi: 10.1038/ncomms6507. PubMed PMID: 25425480.

Ayokunmi Ajetunmobi, Søren Hough and Neil Humphryes-Kirilov contributed to this Resource.