

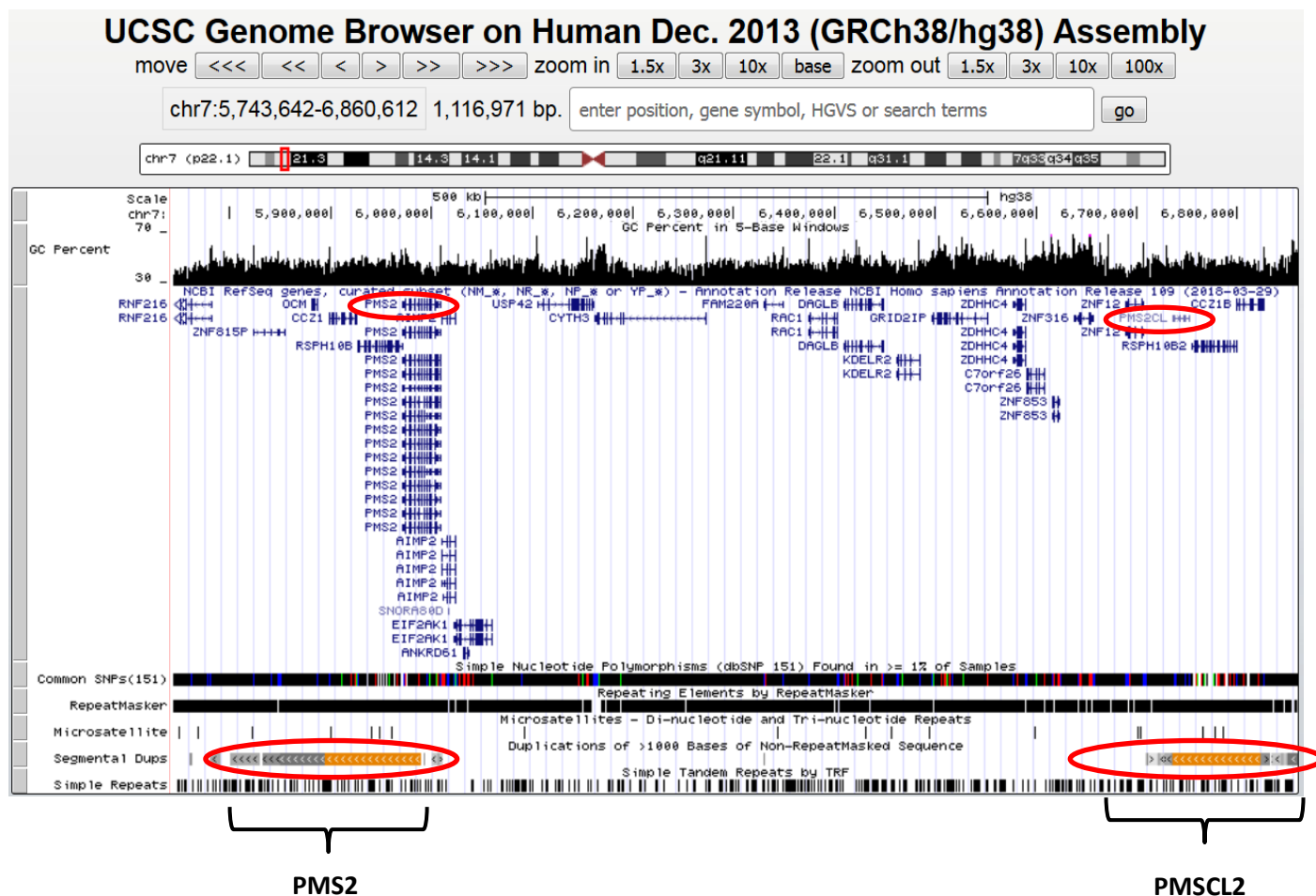


## **Cas9 guide-RNA Design Tutorial for Sage HLS-CATCH**

This document outlines recommendations for designing Cas9 guide-RNAs for HLS-CATCH purification of gene targets. These are based on methods used by Sage Science to select guides for internal testing. An example target gene (PMS2) will be used to illustrate the selection of suitable guide-RNAs.

PMS2 is a ~38 kb gene located on chromosome 7 that is involved in DNA repair. Mutations in the gene are associated with Lynch Syndrome and some cancers. The sequence of a large portion of the PMS2 is duplicated about 800kb to the right (in the 5' direction) of the gene on the same chromosome. These duplications in the pseudogene (PMSCL2) create difficulties in accurate sequence analysis of the PMS2 gene using hybridization capture or short-read whole genome sequence assembly.

HLS- CATCH can be used to isolate the PMS2 gene without contamination with PMS2CL2 sequence. In this example we'll design gRNAs to CATCH a human genomic fragment carrying only PMS2. The UCSC Genome Browser image below shows the position of PMS2 and PMSCL2, on chr. 7.



## Guide-RNA Design Approach

The general approach to guide-RNA (gRNA) design relies on reference sequences for human and mouse genomes and the highly featured annotations available from University of California Santa Cruz's [UCSC Genome Browser](https://genome.ucsc.edu/).

The goal is to identify unique genomic regions free of common SNPs and/or repeated sequences. Two strategies are used:

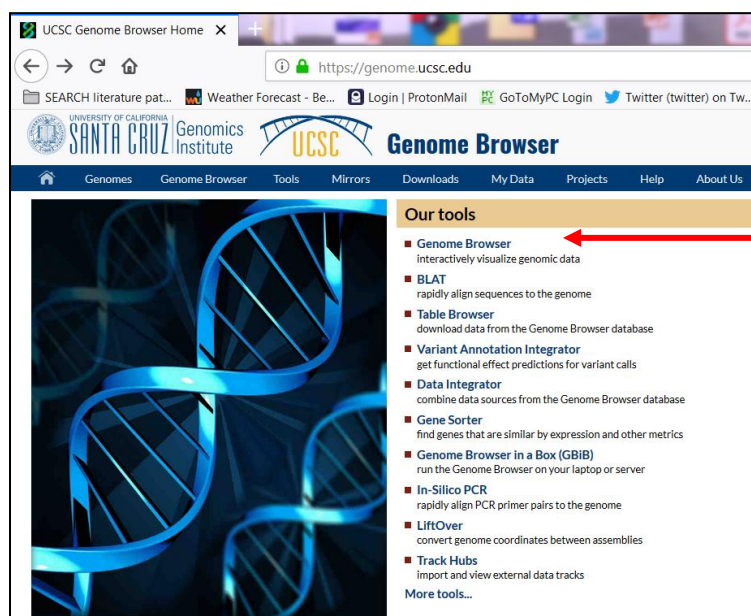
- **Strategy 1** (page 5) Uses the sequence coordinates (or FASTA sequences) of the repeat-free regions to search for gRNAs in databases, or in gRNA design packages. These databases and gRNA design packages are free and publicly available on the internet.
- **Strategy 2** (page 11) Uses the built-in CRISPR-10K track of the UCSC genome browser to find RNAs. This strategy is somewhat more limited in that the CRISPR-10k track only annotates Cas9 gRNAs within 10kb of known genes or exons. However, it is very convenient for gene-rich regions like the one used for our example, human PMS2.

**Specificity** is the prime consideration in designing gRNAs for HLS-CATCH, since very high Cas9 enzyme concentrations are used, and digestion conditions which bear little resemblance to the *in vivo* studies that most gRNA design sites use to determine gRNA cutting efficiency.

Based on internal Sage Science R&D, only 1 in 10 gRNAs designs showed a significant ( $\geq 2$ -fold) reduction in cutting efficiency (compared to neighboring alternative sites). And none of the gRNAs designs were completely inactive at the intended target site.

## UCSC Genome Browser Display Settings

1. Open the [UCSC Genome Browser](https://genome.ucsc.edu/) and click "[Genome Browser](#)"



2. Check that you are using the right reference for your project (for some genomic regions, it may be preferable to use hg19 rather than hg38).
3. Type in the name of the gene or coordinates to be targeted
4. Press "GO".

UCSC Genome Browser Gateway

SEARCH literature pat... Weather Forecast - Be... Login | ProtonMail GoToMyPC Login Twitter (twitter) on Tw... Sage Science UCSC Genome Bro

Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About Us

**Browse/Select Species**

POPULAR SPECIES

Human Mouse Rat Zebrafish Fruitfly Worm Yeast

Enter species or common name

REPRESENTED SPECIES

Human Chimp Bonobo Gorilla Orangutan Gibbon Green monkey Crab-eating macaque

**Find Position**

Human Assembly  
Dec. 2013 (GRCh38/hg38)

Position/Search Term  
PMS2  
Current position: chr7:5,973,239-6,009,125

GO

Human Genome Browser - hg38 assembly

UCSC Genome Browser assembly ID: hg38  
Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38.p12 (GCA\_000001405.27)  
Assembly date: Dec. 2013 initial release; Dec. 2017 patch release 12  
Assembly accession: GCA\_000001405.27  
NCBI Genome ID: 51 (Homo sapiens (human))  
NCBI Assembly ID: 5800238 (GRCh38.p12, GCA\_000001405.27)  
BioProject ID: PRJNA31257

5. Ordinarily there will be many matches for a well-researched gene like PMS2. Usually, the named gene will be one of the first few matches. Click on the link to the appropriate gene.

Human hg38 PMS2 UCSC Geno X

https://genome.ucsc.edu/cgi-bin/hgTracks?hgsid=75496

SEARCH literature pat... Weather Forecast - Be... Login | ProtonMail GoToMyPC Login Twitter (twitter) on Tw...

Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help

Your search resulted in multiple matches. Please select a position:

**Known Genes**

PMS2 (ENST00000265849.11) at chr7:5973239-6009125 - Homo sapiens PMS1 homolog 2, mismatch repair sys

PMS2 (ENST00000644110.1) at chr7:5973399-6004006 - The sequence shown here is derived from an Ensembl

PMS2 (ENST00000643595.1) at chr7:5973361-6009041 - Component of the post-replicative DNA mismatch r

PMS2 (ENST00000642456.1) at chr7:5973274-6009075 - Homo sapiens PMS1 homolog 2, mismatch repair sys

PMS2 (ENST00000642292.1) at chr7:5973247-6009097 - Homo sapiens PMS1 homolog 2, mismatch repair sys

PMS2 (ENST00000469652.1) at chr7:5982875-6009019 - PMS1 homolog 2, mismatch repair system component

PMS2 (ENST00000441476.6) at chr7:5973242-6009106 - Homo sapiens PMS1 homolog 2, mismatch repair sys

PMS2 (ENST00000415839.2) at chr7:6003310-6006066 - PMS1 homolog 2, mismatch repair system component

PMS2 (ENST00000406569.7) at chr7:5977674-6009019 - PMS1 homolog 2, mismatch repair system component

PMS2 (ENST00000382321.5) at chr7:5973399-6009019 - Component of the post-replicative DNA mismatch r

PMS2 (ENST00000380416.5) at chr7:6001966-6009098 - PMS1 homolog 2, mismatch repair system component

MLH1 (ENST00000539477.5) at chr3:36993777-37050842 - Heterodimerizes with PMS2 to form MutL alpha,

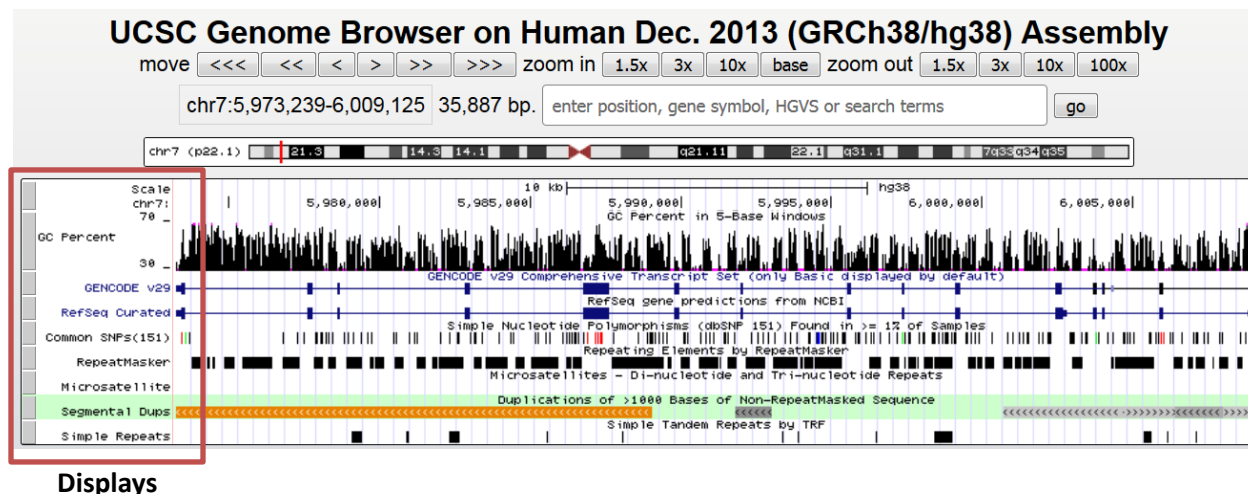
MLH1 (ENST00000458205.6) at chr3:36993776-37050846 - Heterodimerizes with PMS2 to form MutL alpha,

MLH1 (ENST00000455445.6) at chr3:36993798-37050706 - Heterodimerizes with PMS2 to form MutL alpha,

6. When the sequence assembly page loads, scroll down to the display settings. If necessary, disable any open displays by pulling down the display menus and selecting “hide”. Display settings are activated by pressing the “refresh” button in each category. Customize the display settings as follows:

- Mapping and Sequencing
  - Base Position (full)
  - GC Percent (full)
- Genes and Gene Predictions
  - All GENCODE (dense)
  - NCBI RefSeq (dense)
- Variation
  - Common SNPs 151 (dense)
- Repeats
  - Repeat Masker (dense)
  - Segmental Dups (dense)
  - Simple Repeats (dense)

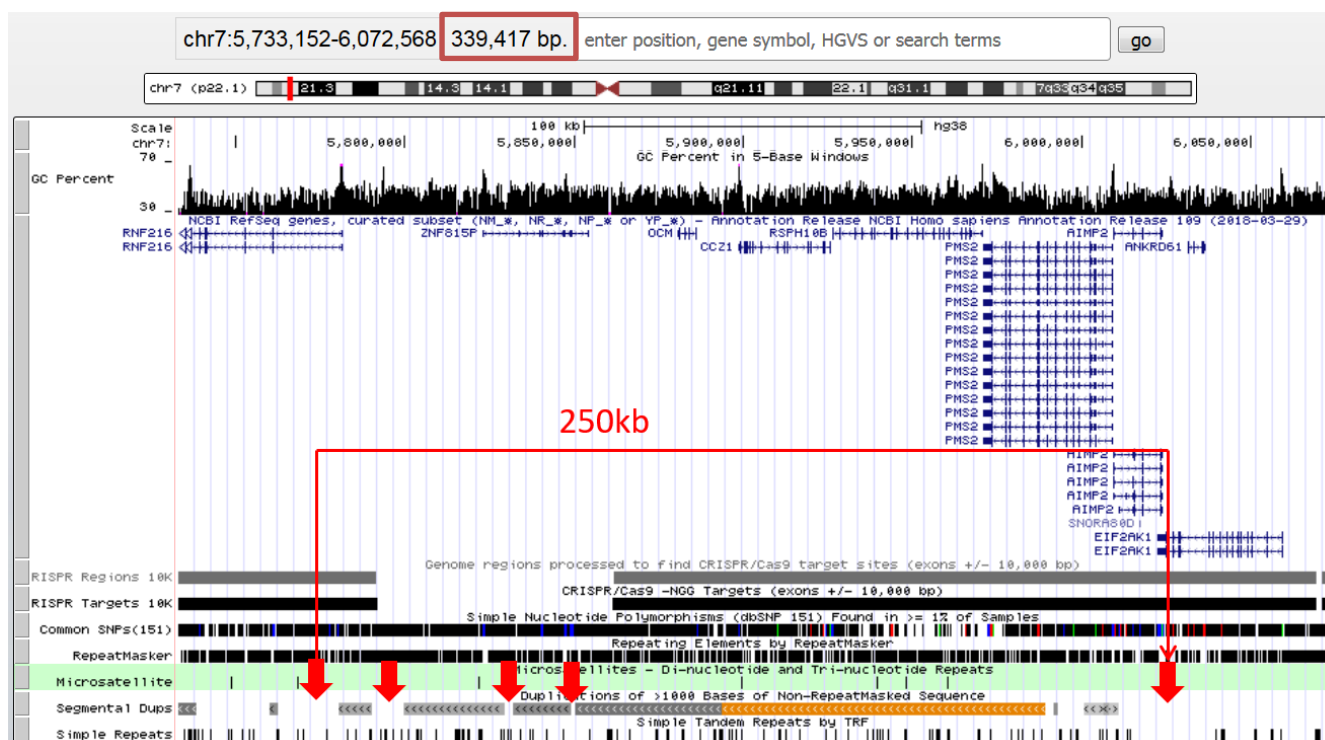
7. The Genome Browser display should be similar to the image shown.



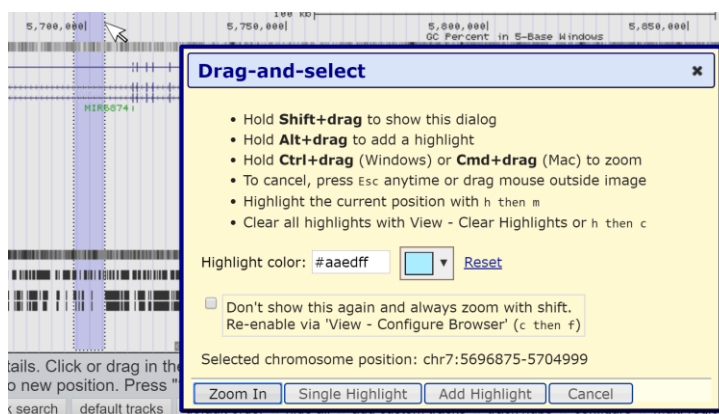
Displays

## Strategy 1 – Sequence Coordinates

1. When designing CATCH gRNAs, regions with skewed base composition and repeated sequences should be avoided. This view will allow selection of gRNA sites that are likely to be unique in the genome. Note that both ends of the gene show significant presence of segmental duplications.
2. To design gRNAs that flank the entire PMS2 gene, single-copy genomic sequences must be identified. To identify these regions, zoom out in the display. In the example below, the display has been zoomed out to 339 kb in order to identify region candidates, indicated below with red arrows.

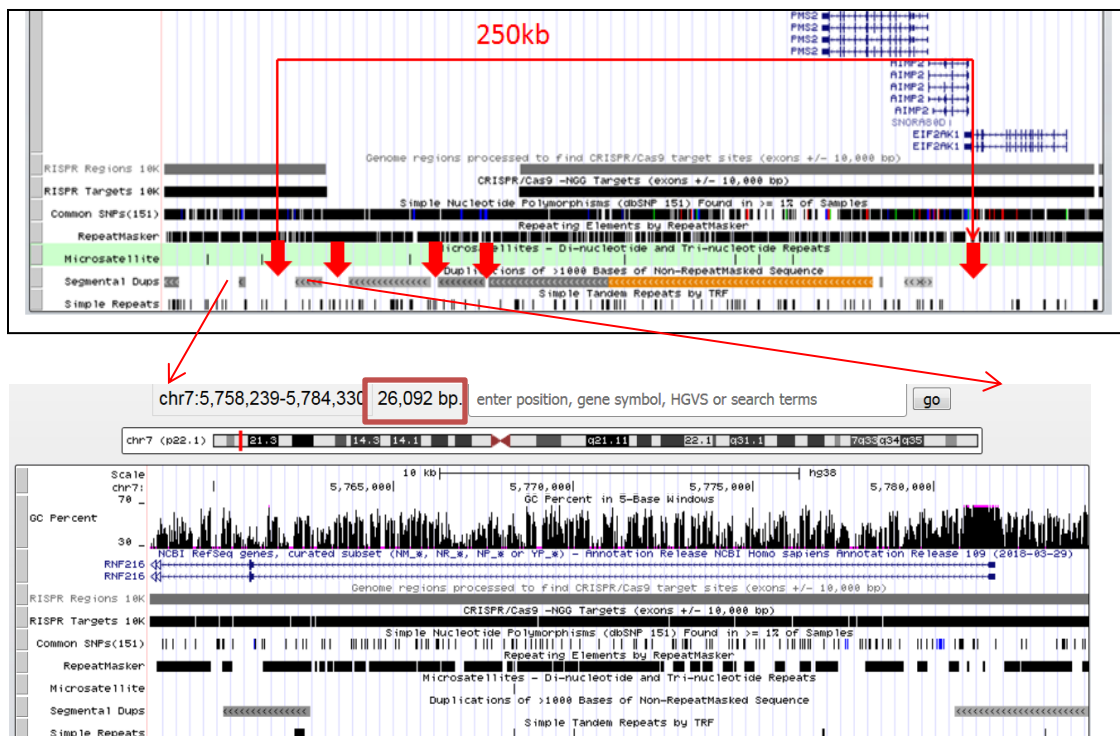


3. Once the region candidates are identified, they can be zoomed in on. To zoom, drag the mouse pointer across a region above the “Base Position” region on the assembly (this setting must be displayed). A dialogue box will appear. Press “Zoom In”.





4. Starting with the left boundary, zoom in for a closer evaluation.



5. The right side of this 26kb region has several regions without repeats or common SNPs. Zoom in further. Using criteria of balanced GC%, absence of SNPs and repeats, several promising regions for gRNAs are identified (circled in red). As an example, the right-most region will be zoomed in further.



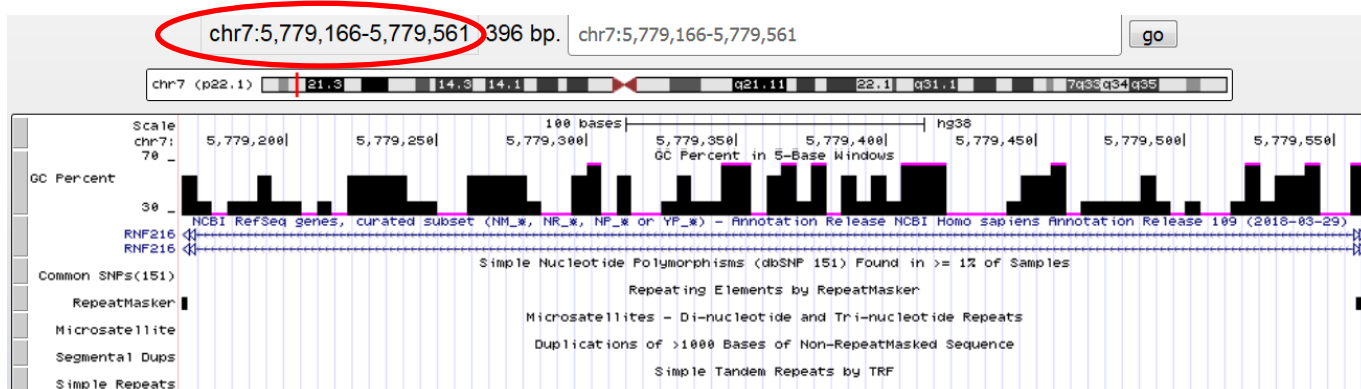
6. The coordinates of this repeat-free, common-SNP-free region can now be used to select gRNAs from a database of human gRNAs, such as:

<http://www.guidescan.com> (Sloan Kettering)

Or FASTA sequence from this region can be pasted into gRNA design modules such as those found at:

<https://www.benchling.com> (Benchling)

<https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design> (The Broad Institute)



7. Using [GuideScan](#) as an example, open the site in a browser, then 1) ensure the correct genome is selected, 2) select “spCas9” as the enzyme, 3) paste the genomic coordinates [FASTA or other files may also be uploaded] and 4) press “guide me!”.



8. After several seconds, scroll down the page. GuideScan will report the number on guideRNAs identified. Press “download results” to create a .csv file, and/or press “+” to view the results.


**OR**

BED, GFF/GTF, Fasta, TXT file upload


*TXT file contains single column of genomic coordinates in form chr#:start-end*

*Processing time scales with lines in file, 1000 queries takes ~5 minutes*

No file selected.



1 within queries processed, found gRNAs for 1 queried coordinates

  43 guideRNAs within chr7:5779166-5779561

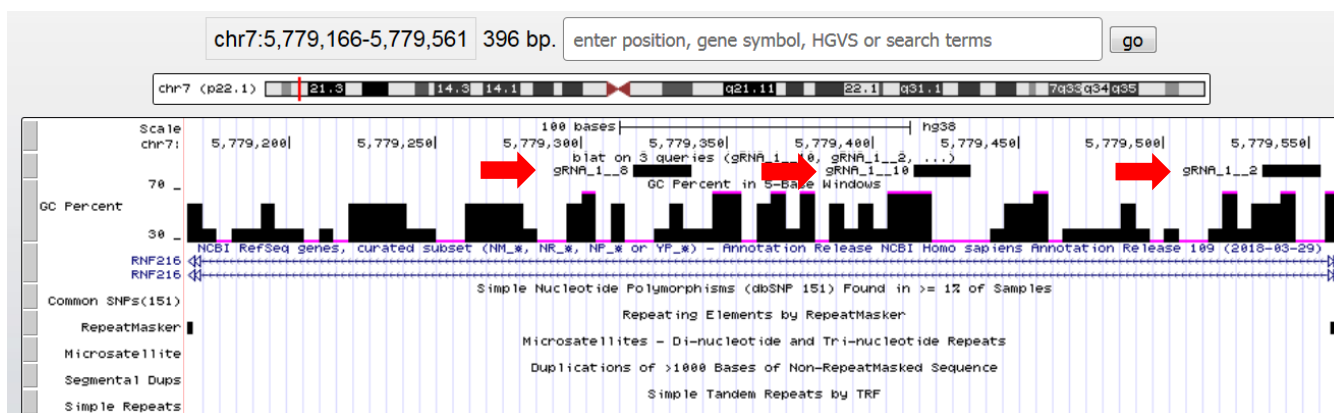
9. The columns (.csv file, below) show sequence coordinates, the crRNA 20b recognition sequence (in DNA bases), cutting efficiency and specificity scores, offtarget analyses, and a label. The cutting efficiency scores are based on *in vivo* editing data, and may not be very important to our (very different) *in vitro* CATCH conditions. However, the offtarget analyses are based on the frequency of matches to the reference genome (outside of the desired target region). gRNA positions with 0-1 mismatches to genomic sequences outside of the target region are not reported by Guidescan as unique gRNAs. Therefore, offtarget sequences reported by Guidescan have at least 2 mismatches to the gRNA. The left offtargets column shows total sequences with 2-3 mismatches, and the right offtargets column shows the number of offtarget sequences with 2 and 3 mismatches respectively.

For CATCH, we try to pick 2-3 non-overlapping gRNAs (based on the coordinates) with the best cutting efficiency scores and fewest mismatches. For this region, these are shown shaded in yellow. Overlapping gRNA positions are shaded gray.

chr7:5779166-5779561	chrom	target site start coordinate	target site end coordinate	gRNA	cutting efficiency	cutting specificity score	strand	offtargets	offtargets	annotation	gRNA label
	chr7	5779531	5779553	CATTCTTTACGCGCTGAGAA	56	0.68955224	-	1 2:0 3:1	*		gRNA_1_1
	chr7	5779532	5779554	ACATTCTTTACGCGCTGAGA	52	0.96466431	-	1 2:0 3:1	*		gRNA_1_2
	chr7	5779537	5779559	TCAGCGCGTAAAGATGTTTC	45	0.87491065	-	1 2:0 3:1	*		gRNA_1_3
	chr7	5779328	5779350	AGCCTCGAATGAGGTTTGCC	45	0.65511265	+	2 2:0 3:2	*		gRNA_1_4
	chr7	5779330	5779352	CCTCGAATGAGGTTTGCCAG	71	0.64673913	+	2 2:0 3:2	*		gRNA_1_5
	chr7	5779329	5779351	GCCTCGAATGAGGTTTGCCA	52	0.75186727	+	3 2:0 3:3	*		gRNA_1_6
	chr7	5779477	5779499	AATTAGTCGCGCTTGCTTTA	18	0.38267215	-	4 2:0 3:4	*		gRNA_1_7
	chr7	5779319	5779341	AAAATCTAAAGCCTCGAATG	56	0.77120096	+	5 2:0 3:5	*		gRNA_1_8
	chr7	5779330	5779352	CCCCTGGCAAACCTCATTG	53	0.46532889	-	6 2:0 3:6	*		gRNA_1_9
	chr7	5779412	5779434	TTCATAATTTATTGCGCCTG	50	0.71858366	-	6 2:0 3:6	*		gRNA_1_10
	chr7	5779405	5779427	TTTATTGCGCTGGGGTG	54	0.4039859	-	7 2:0 3:7	*		gRNA_1_11
	chr7	5779406	5779428	ATTTATTGCGCTGGGGTGT	40	0.3300318	-	8 2:2 3:6	*		gRNA_1_12
	chr7	5779414	5779436	ATTTATAATTTATTGCGCC	36	0.54663693	-	8 2:0 3:8	*		gRNA_1_13
	chr7	5779205	5779227	CAGATCGAAAGTTGATGAAT	39	0.20980347	-	9 2:0 3:9	*		gRNA_1_14

10. Most of the other gRNA selection tools will give output similar to this format, with cutting efficiency and specificity scores. However, most of the others are not as restrictive with respect to specificity as GuideScan, because they will show all gRNA sites, even those in repeated sequences.
11. Typically, 2 or 3 gRNA in a clustered cut region and perform CATCH with an equimolar mixture of the selected guides. We use a mixture to avoid the chance that a rare SNP will inactivate any single gRNA, thereby eliminating Cas9 cutting on that side of the fragment.

Finally, the selected gRNAs can be pasted into a custom BLAT search and displayed in our candidate cut region to confirm absence of overlaps, SNPs, and repeats.

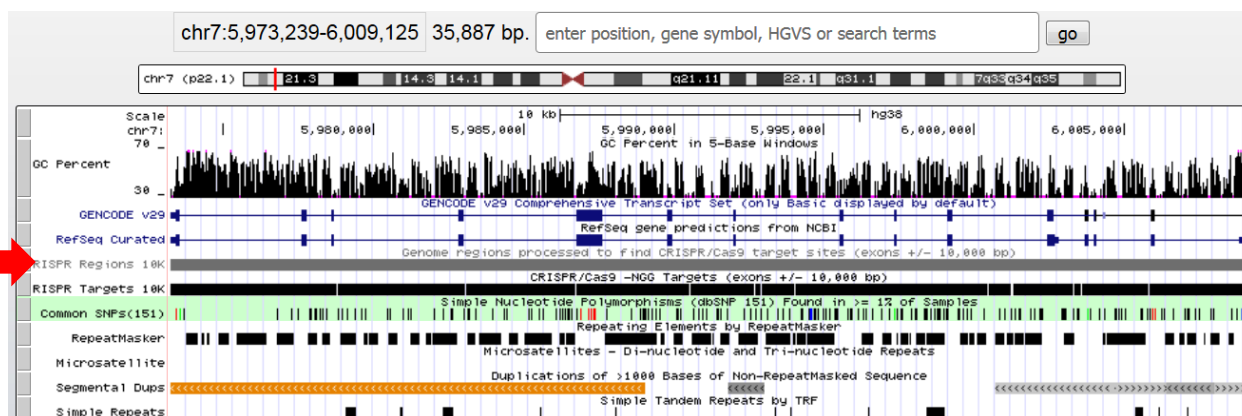


12. Repeat the process to select 2 or 3 gRNAs for the right side of the target region.

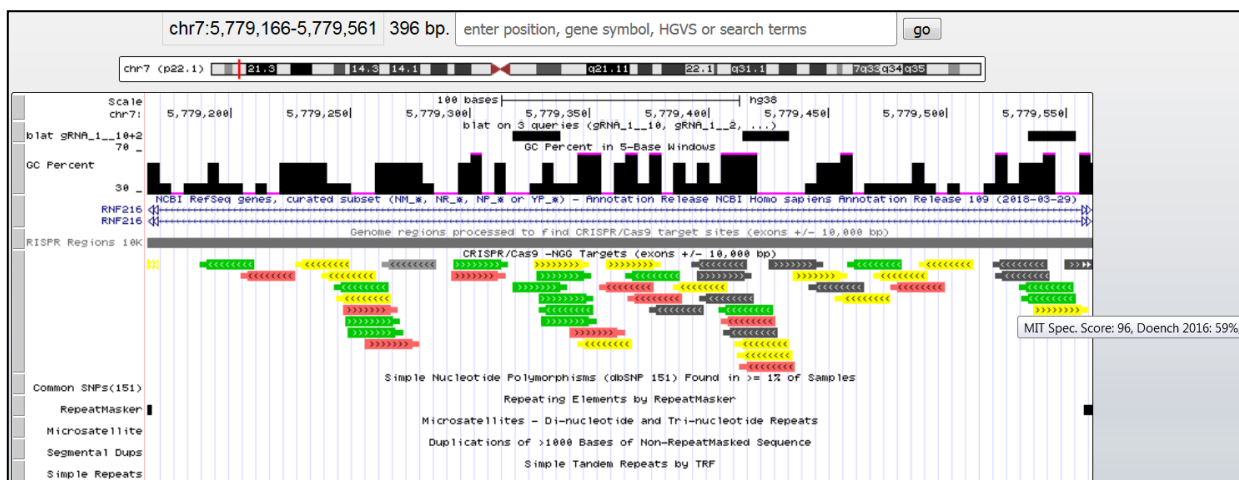
## Strategy 2 – Using the CRISPR-10K track

1. UCSC Browser also has a Cas9 gRNA track that shows gRNA positions within 10kb of exons. This tool is very useful for designing gRNAs very close to particular genes or specific exons within a gene. In the display setting for “Genes and Gene Predictions” select “pack” (or “full”) from the “CRISPR Targets” drop down menu.

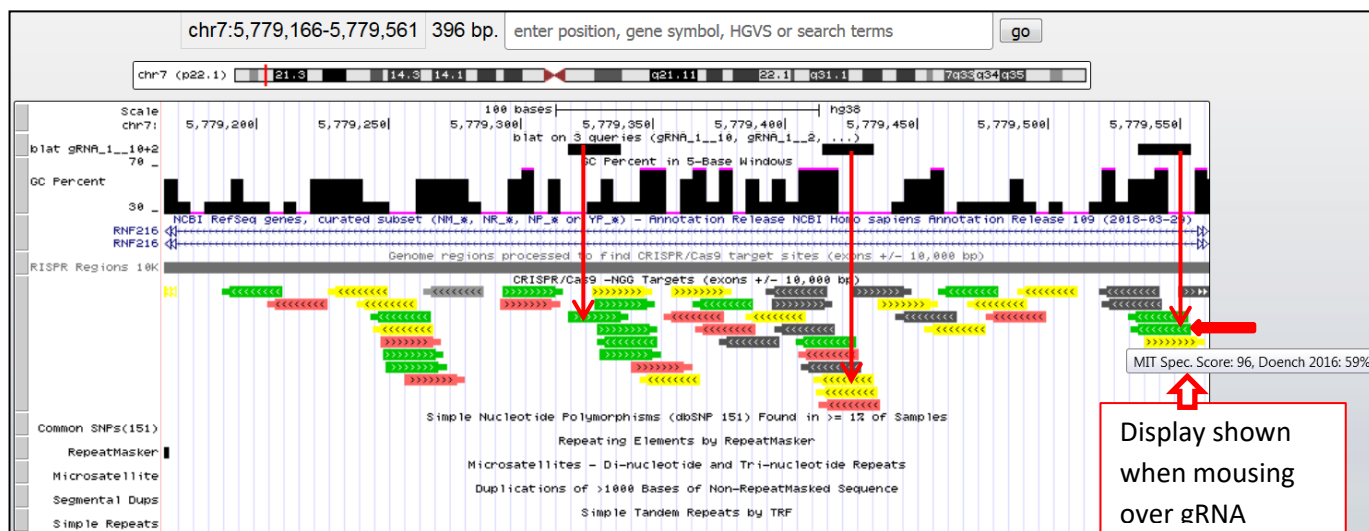
2. Returning to a 26kb view of the left-side candidate region, the gray CRISPR region indicates that the Genome Browser has mapped gRNA sites throughout this region.



3. Zooming in further to a 396 bp region within the previously identified region with no SNPs or repeats (chr7:5,779,166-5,779,561), the CRISPR-10K track shows gRNA sites color-coded for cutting efficiency and specificity.



3. Hovering the mouse pointer over the color-coded site gives a more detailed rating of specificity. You should give higher priority to the specificity scores close to 100. As expected, the gRNAs selected from Guidescan db are also found in the UCSC track, and the scores in the CRISPR track are some of the highest (based on the BLAT search of the GuideScan-select guides, shown below).



Shades of gray stand for sites that are hard to target specifically, as the 20mer is not very unique in the genome:

	impossible to target: target site has at least one identical copy in the genome and was not scored
	hard to target: many similar sequences in the genome that alignment stopped, repeat?
	hard to target: target site was aligned but results in a low specificity score <= 50 (see below)

Colors highlight targets that are specific in the genome (MIT specificity > 50) but have different predicted efficiencies:

	unable to calculate Doench/Fusi 2016 efficiency score
	low predicted cleavage: Doench/Fusi 2016 Efficiency percentile <= 30
	medium predicted cleavage: Doench/Fusi 2016 Efficiency percentile > 30 and < 55
	high predicted cleavage: Doench/Fusi 2016 Efficiency > 55

4. Repeat the process to select 2 or 3 gRNAs for the right side of the PMS2 target region.

## Summary

We have selected 2 sets of 3 gRNAs that will excise a unique ~250kb genomic fragment containing the PMS2 gene with contamination from the PMS2CL pseudogene:

Left side:

ACATTCTTTACGCGCTGAGA chr7:5779532-5779554

AAAATCTAAAGCCTCGAATG chr7:5779319-5779341

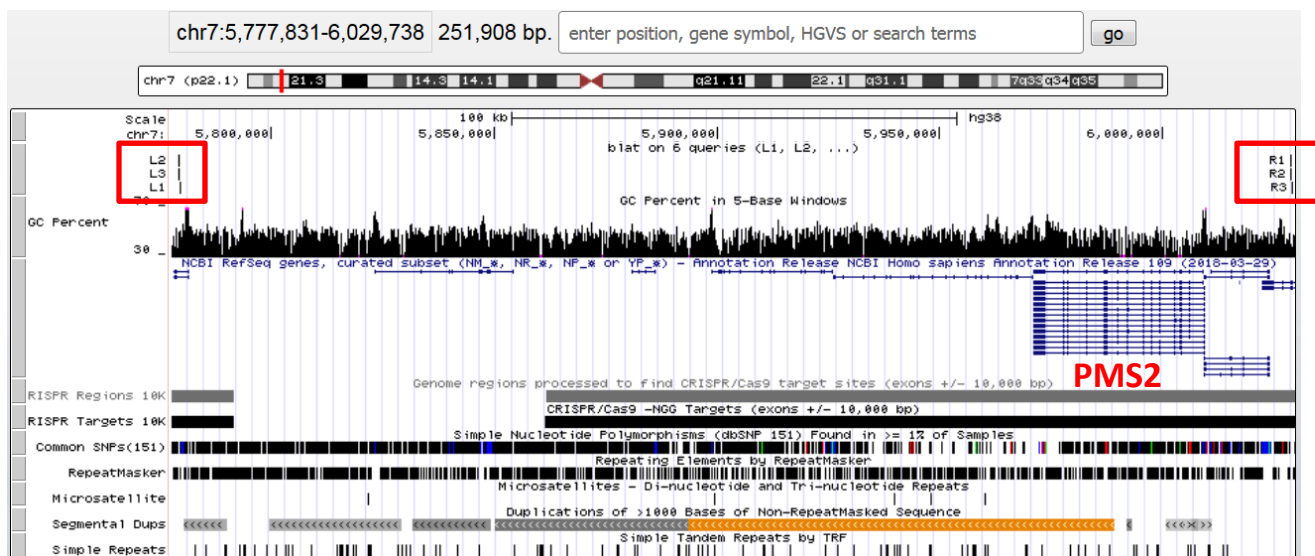
TTCATAATTTATTGCGCCTG chr7:5779412-5779434

Right side:

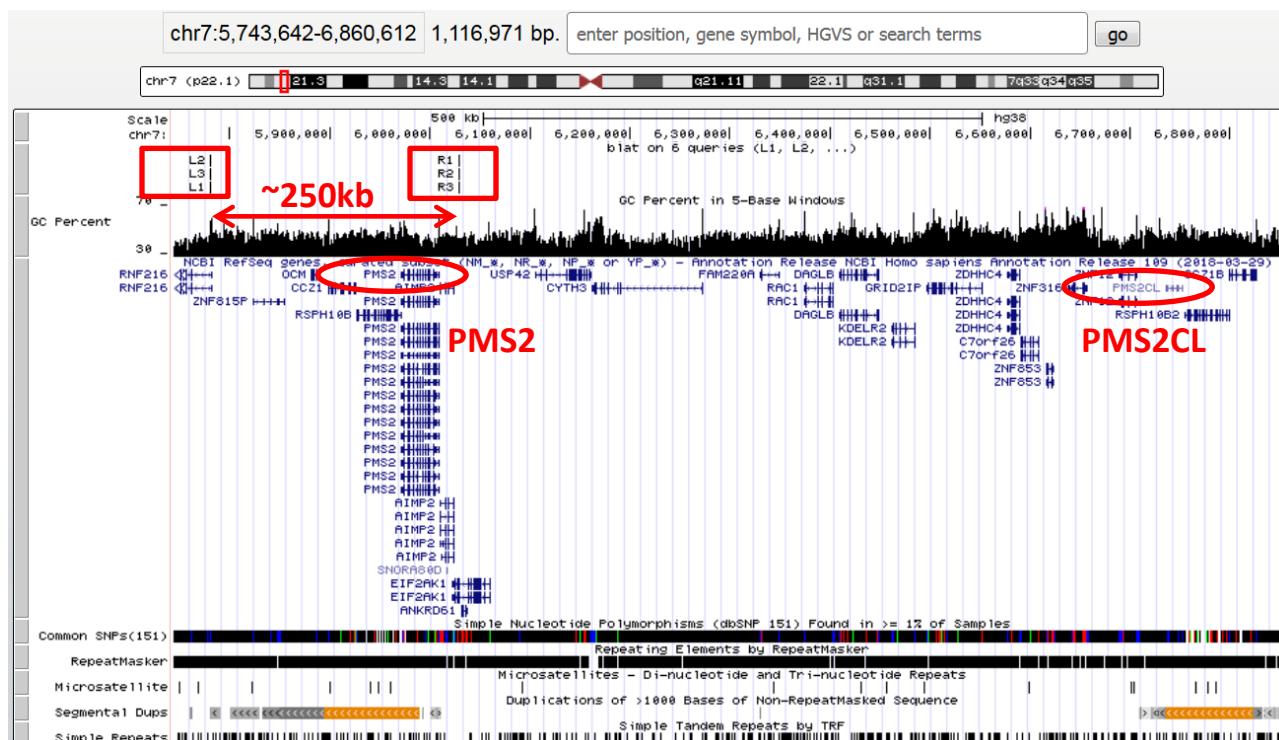
GTACGCTTCACCCGAACAGT chr7:6028638-6028660

CCCACTCTGGACGTATGTGT chr7:6028670-6028692

CTTCTGTAGGATGTCTGTGC chr7:6028949-6028971



UCSC overview of final design, showing PMS2 and PMS2CL:



### **Other challenges with gRNA design**

The UCSC Genome Browser is a valuable resource and it makes gRNA selection almost trivial for human and mouse genomes. For other organisms that lack such rich genome annotation, it should be possible to use a reference genome sequence with an oligonucleotide search tool like BLAT to assess gRNA specificity.

Another difficult problem is devising CATCH gRNA designs for regions that are embedded within segmental duplications. In such situations, it may be necessary to use non-unique gRNAs to excise the CATCH target. It may be possible to find islands of single copy sequence within the repeats surrounding region of interest that can be used. Another strategy is to look for gRNA sites surrounding the region of interest that are repeated only a few times, and see if the Cas9 digestion product for the region of interest can be separated from offtarget CATCH products during size-selection in the SageHLS workflow. When designing gRNAs for repeated genomic sequences, flexible gRNA design sites such Benchling or the Broad Institute gRNA design site must be used, since database sites like Guidescan.com only list unique gRNAs.