Project 2.2

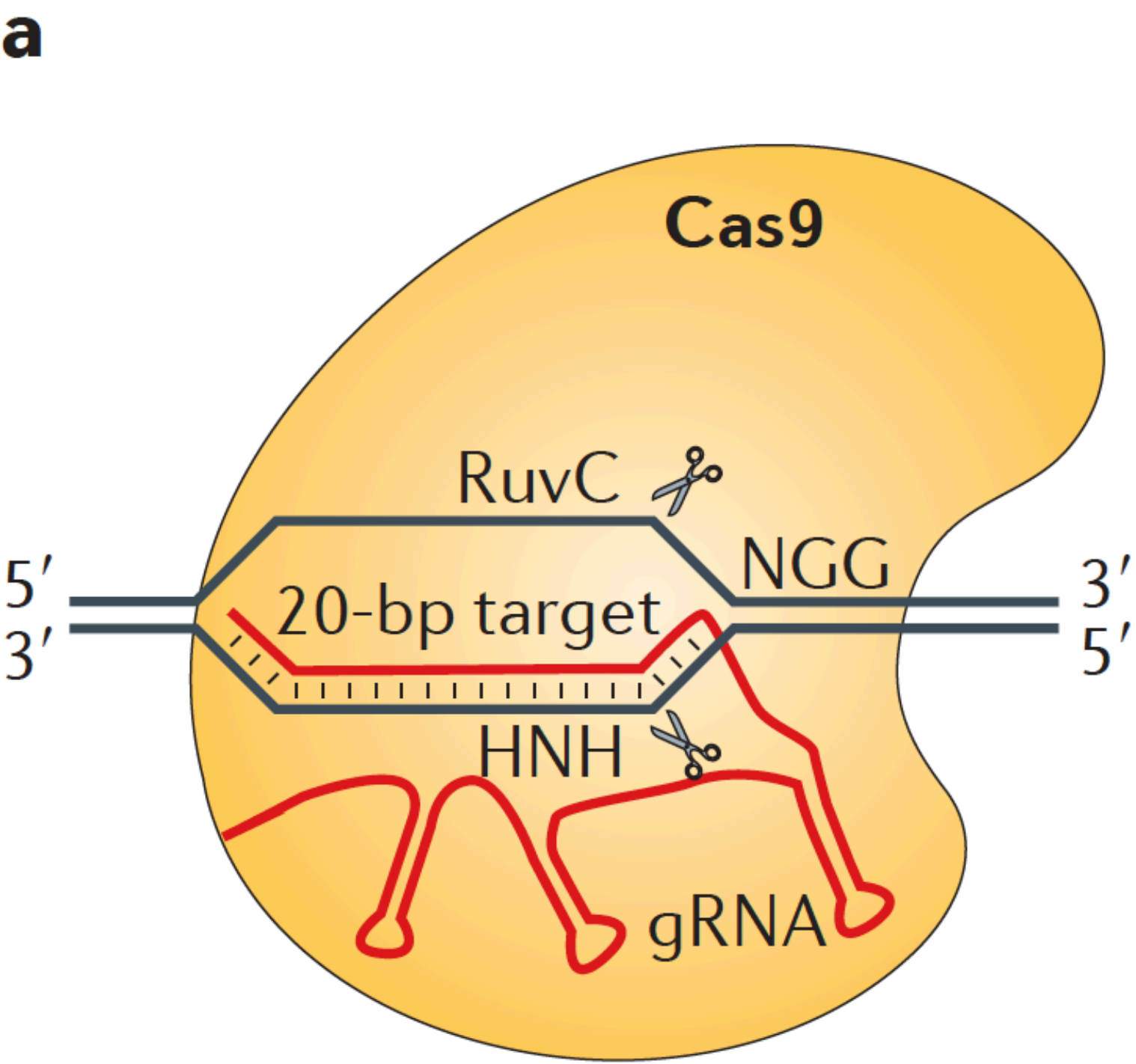
CBB 752

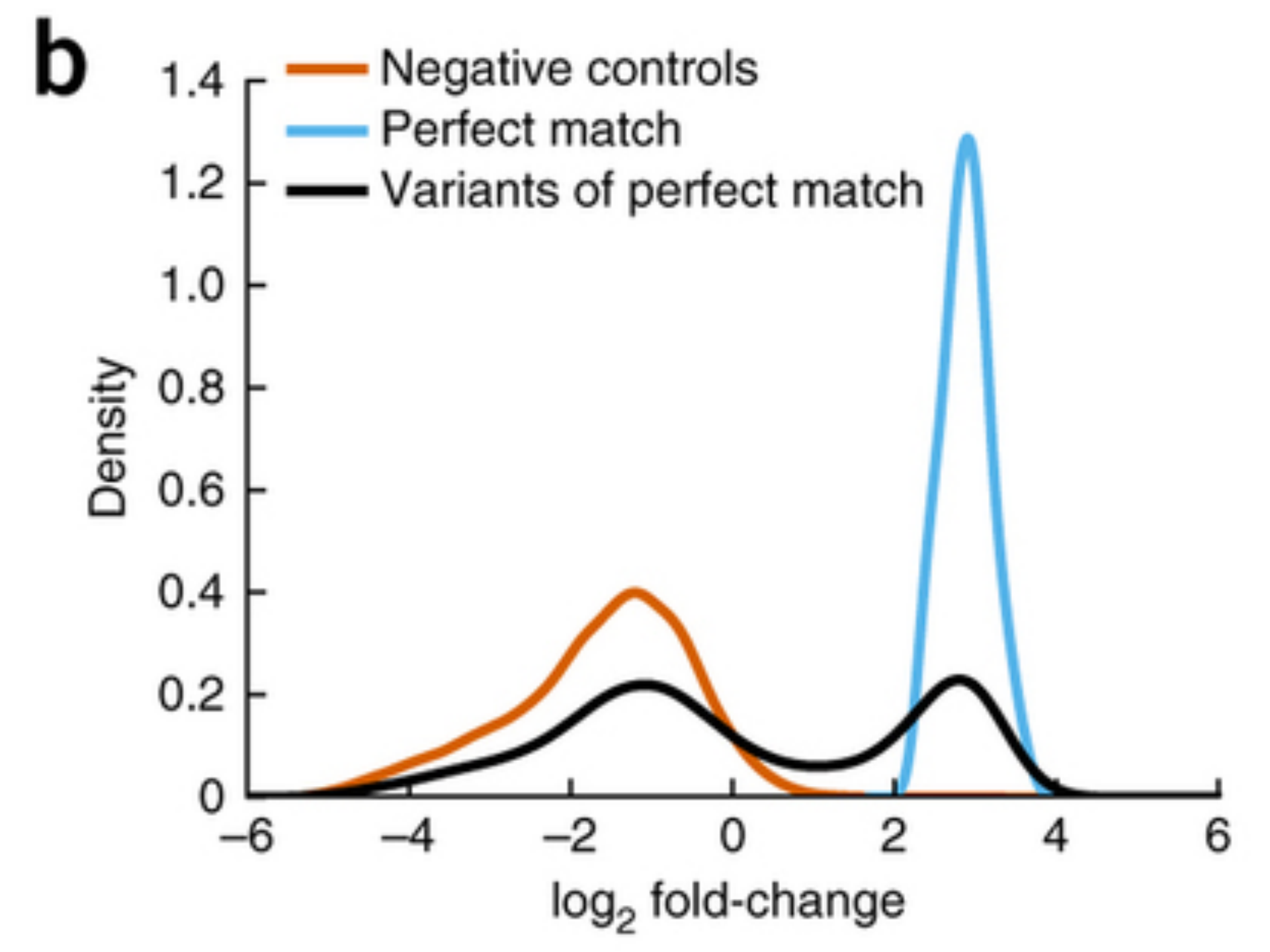
Final project

CRISPR and personal genomics: The impact of SNPs on sgRNA sets and off target mutations.

Writing

How might SNPs in Carl's genome impact the use of CRISPR as a treatment? Discuss how individual SNPs would impact the off-target effects in the presence of the SNP.


a
Cas9
RuvC
NGG
5′
3′
20-bp target
3′
5′
HNH
gRNA


b
Negative controls
Perfect match
Variants of perfect match
Density
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
−6
−4
−2
0
2
4
6
log2 fold-change

Overall, the best way to avoid SNPs causing both undesired off-target effects or decreased therapeutic efficiency is probably through thorough sequencing and screening of the patient's genome prior to construction of CRISPR gRNA sequences.

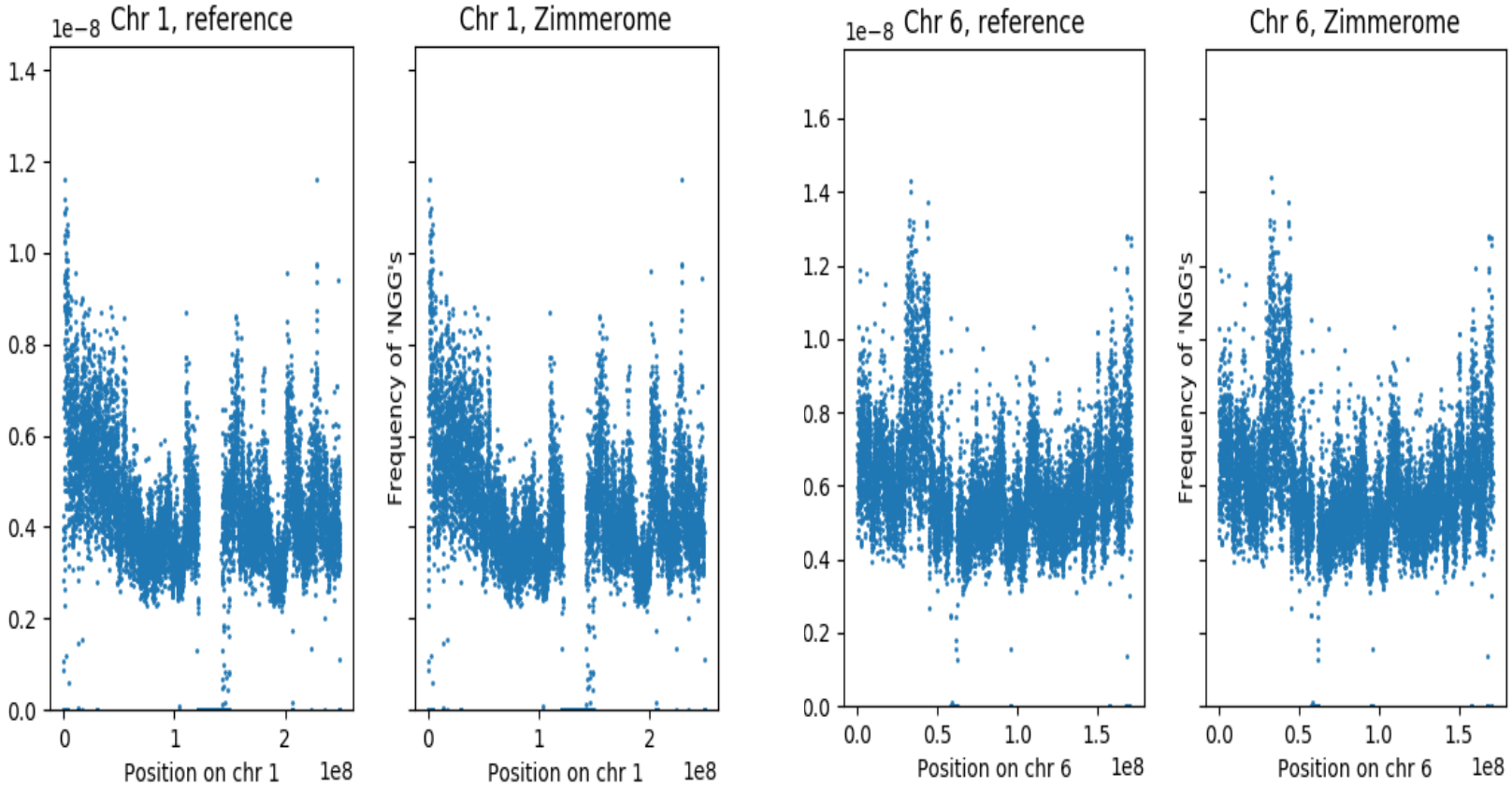
Coding

Propose a tool that finds PAM sites in the human reference genome as well as Carl's genome and compares the similarity of the two sets.

Coding idea

- Take the commonly used CRISPR/Cas9 PAM sites: *NGG* as example
- Find all the *NGG (GG)* sites on the reference genome (GRCh37)
- Find all the *NGG (GG)* sites on the altered genome (Zimmerome)
- Compare the *NGG* distribution on the two genomes

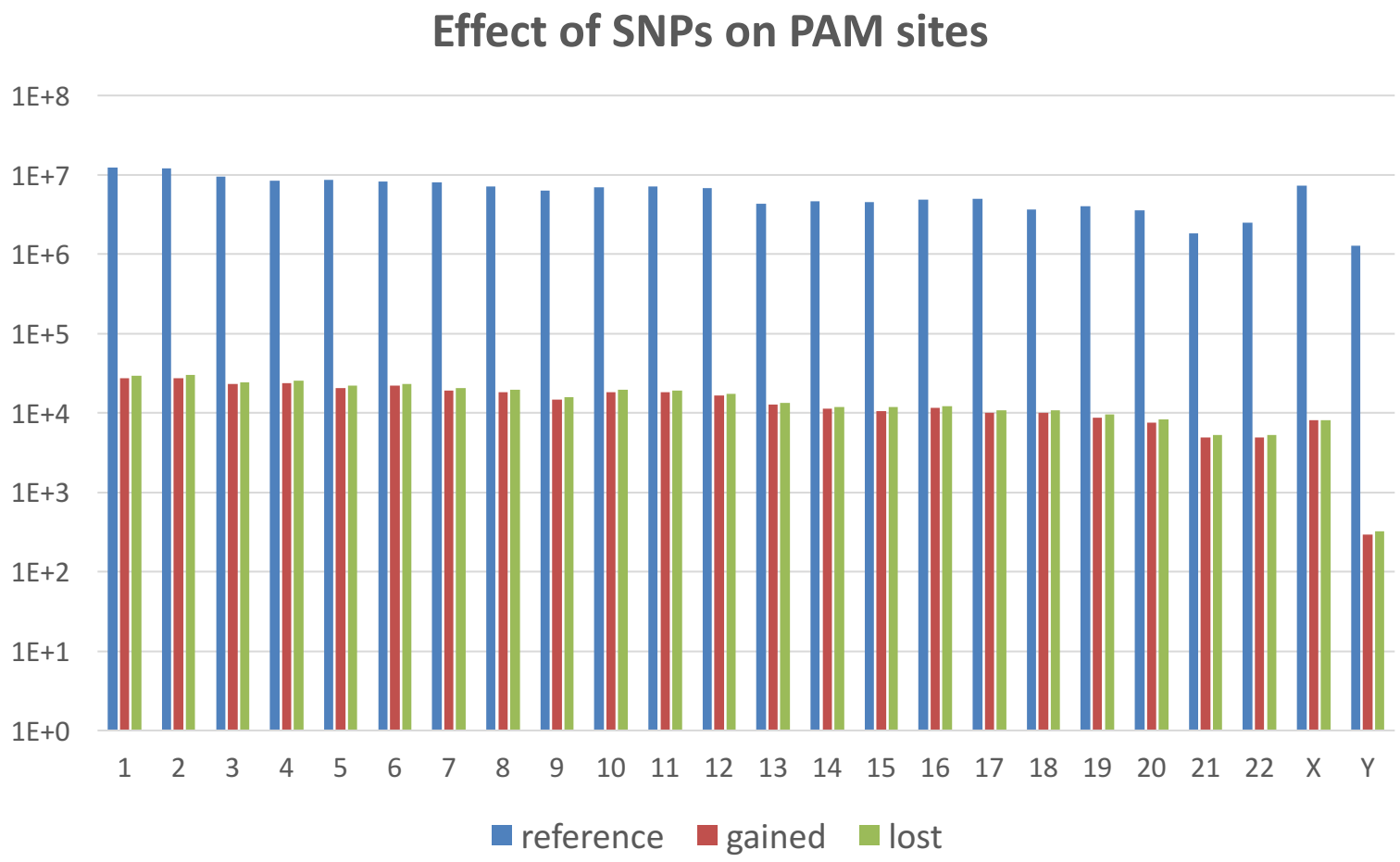
The uneven distribution of *NGG*s



Gained and lost PAM sites due to SNPs

#chrom	pos	change	#chrom	pos	change
1	58209	gained	1	79049	lost
1	65743	gained	1	88176	lost
1	74553	gained	1	91473	lost
1	122813	gained	1	91474	lost
1	127489	gained	1	101266	lost
1	230057	gained	1	133482	lost
1	250232	gained	1	239162	lost
1	537535	gained	1	239163	lost
1	537536	gained	1	242424	lost

Influence of SNPs on PAM sites

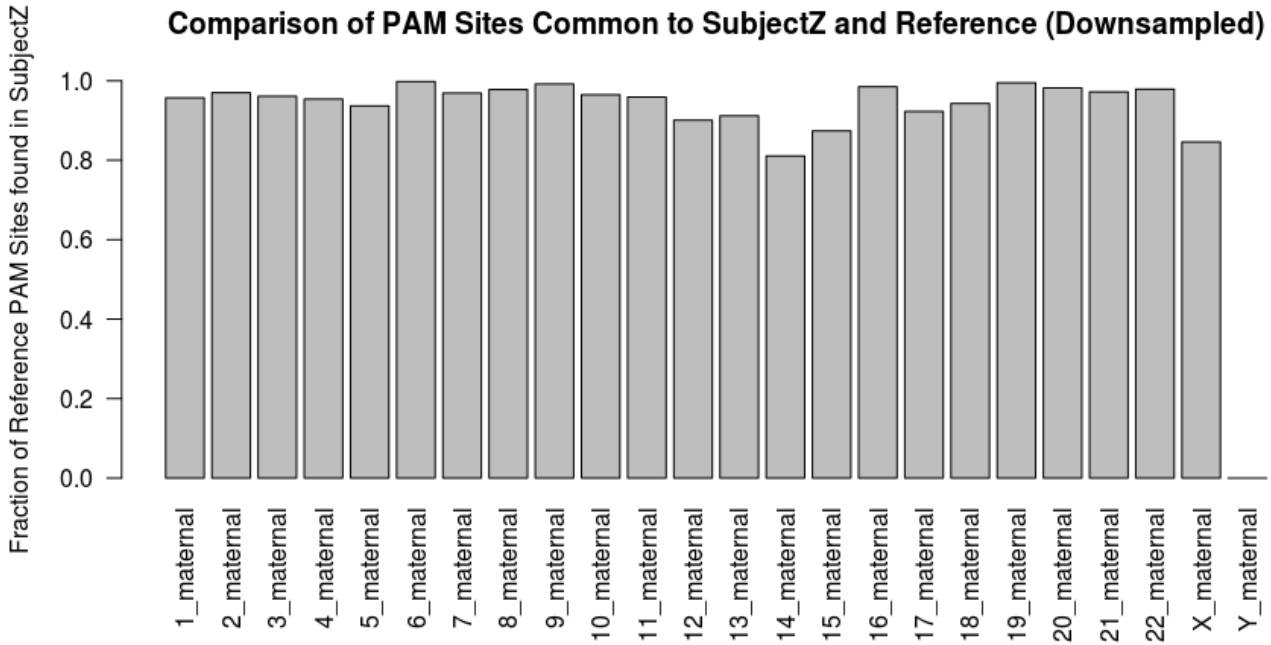


Coding: Approach

- Data I/O
- Find PAM Sites
  - Look for 5’-NGG-3’ and 3’-CCN-5’ motifs on each chromosome
- Compare Sets
  - Append 20 NTs upstream for specificity
  - Find intersection of Zimmerome and reference genome PAM+gRNA sequences

Coding Results

Many but not all reference PAM sites with matched upstream sequences are present in the Zimmerome

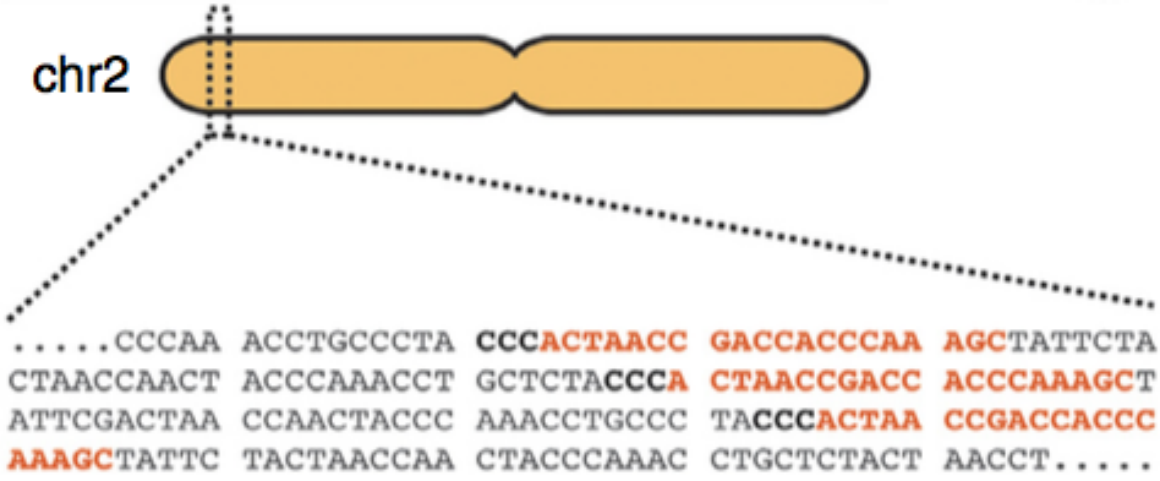


Pipeline

Calculate the sgRNA libraries for the reference genome and Carl's genome. How different are these two sets?

Goal of pipeline:

- Generate single guide RNA (sgRNA) libraries from Carl's genome and compare it to the reference genome
  - Two genome files for Carl: maternal and paternal
  - One reference genome: 37



GuideScan Software Implementation

- Dependencies:
  - samtools==1.3.1
  - pysam==0.8.3, pyfaidx==0.4.7.1, bx-python==0.7.3
  - biopython>=1.66
  - Sklearn==0.16.1

GuideScan: Inputs

- .fasta (or .fa) file of genome
- gRNA sequence length (20 – 100 bases, excluding PAM site)
- Alternative PAM sequence
- Chromosomes (list, comma separated)
- Min and Max allowed mismatches for off-target statistics

Perez et. al., Nature Biotechnology, 2017

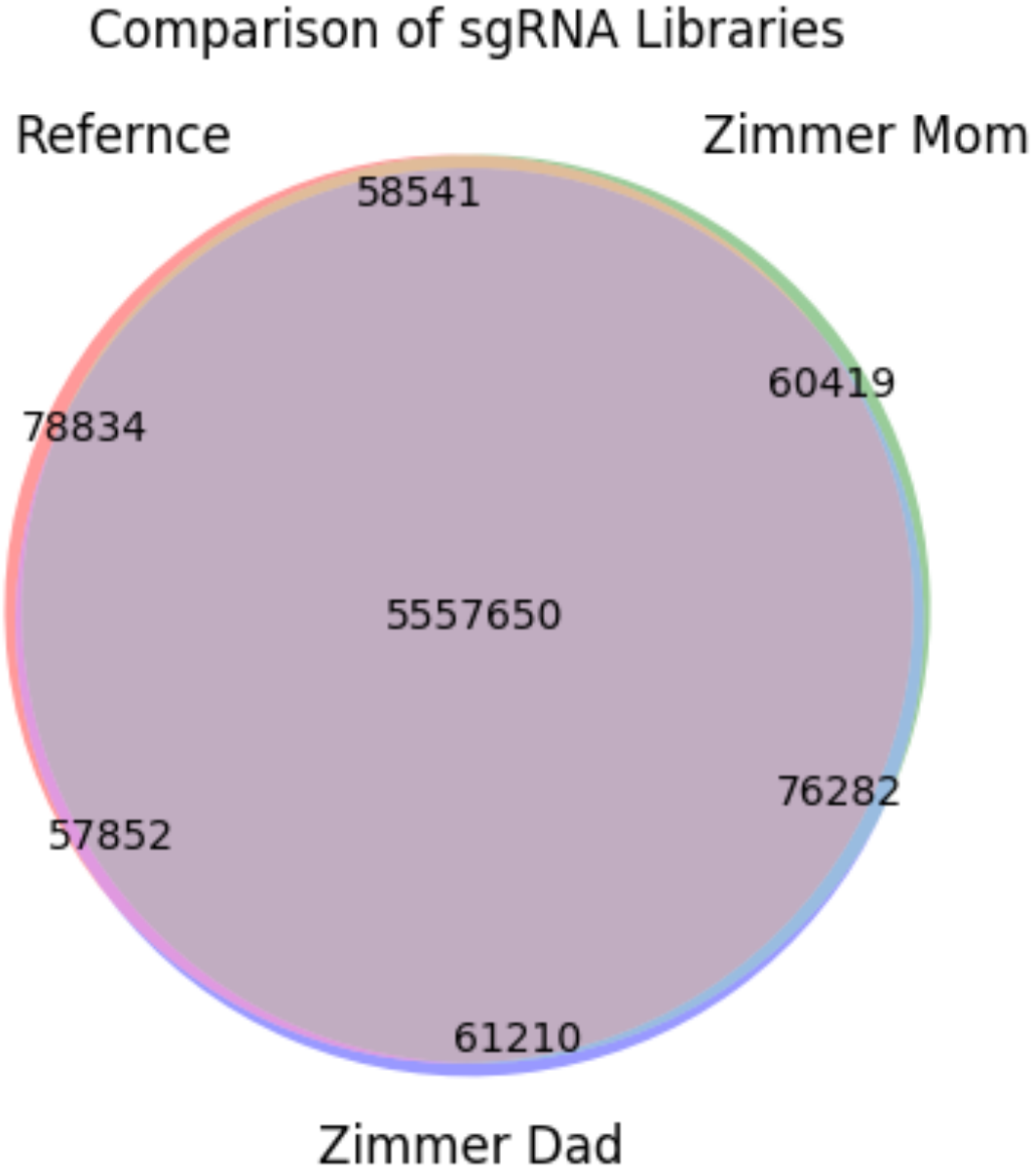
GuideScan: Inputs and Outputs

- Outputs:
  - Target site coordinates (start and end)
  - sgRNA sequence
  - For a given sgRNA:
    - Off target sites, gRNA sequence, ($M:N_M|Q:N_Q$)
      - M = number of allowed mismatches (min) within gRNA sequence to be considered the gRNA, N_M is the number of times it occurs
      - Q = number of mismatches to sequence to be considered off-target site, N_Q is the number of occurances
    - Cutting efficiency score, specificity score for ea. OT site

Our sgRNA Libraries

- Zimmer_mom
- Zimmer_dad
- Human reference genome 37

- Chromosome 18
- 20 base-long gRNA sequences
- Canonical PAM sequence NGG
- ~5.75 M gRNA for each


Comparison of sgRNA Libraries
Refernce
Zimmer Mom
58541
60419
78834
5557650
76282
57852
61210
Zimmer Dad

What we found

- Unique to each library:
  - Mom: 1.05% of 5,752,892
  - Dad: 1.06% of 5,752,994
  - Ref: 1.37% of 5,752,877

- 5.557 M gRNAs in common

What we found

- Maternal-Reference:
  - 58,541
- Paternal-Reference:
  - 57,852
- Maternal-Paternal
  - 76,282

Zimmer mom and dad have slightly more in common than with either does with the reference.