Make SNP/indel-masked genome

* inputs:
  + reference fasta
  + reference gtf
  + bam files with data aligned to this reference (4-6 works well)
* final outputs:
  + $out2fasta
  + $out2gtf
  + the helper/optional files listed in (5) that may or may not be needed for subsequent analyses
* steps:

1. Use bcftools to call SNPs and INDELS 🡪 .vcf file
2. Split .vcf into 3 separate files, uses grep and ParseAndFilterVCF.py:

🡪 (1) INDELs only

(2) SNPs passing quality score, but low in frequency (will be assigned N)

(3) SNPs passing quality score and allele frequency cutoff (will be assigned alternate allele)

3. Use gatk to modify fasta using vcfs (2) and (3) above 🡪 intermediate fasta ($out1fasta)

4. Use rf2m to modify the intermediate fasta along with the reference gtf using (1) above

- The first 2 steps here are to get files into the correct format and to filter the INDELs

a) Modify header in intermediate fasta

b) Modify INDEL only fasta (1) above to

- remove INDELs >50nt and closer than 50nt together (\*note this step was added during troubleshooting and is probably not necessary)

- add a ‘PASS’ flag needed for the next step

- The next 3 steps use perl scripts that make up the rf2m package

c) Format the vcf

d) Modify the intermediate fasta using the filtered INDELs \*Use genome\_creator\_**corrected**.pl

e) Modify the matching gtf

5. Make some more files that might be needed for future analyses

a) fasta index and sequence dictionary from the fasta

b) chrom.sizes file from the fasta

c) bed file from fasta for separating mouse/human/drosophila/ercc reads from future bam files

d) file for later converting ensembl gene IDs to gene names

\* note the script used is different for human and mouse, and $Celline needs to NIH3T3 to use mouse script. This will need to be modified if you use a different $Celline name for a mouse sample

e) bed file equivalent of gtf

f) bed file but with only one entry per gene, using the outside-most coordinates

g) compressed fasta and gtf for transferring and loading onto IGV