GENOME BUILDER USING SNPS AND INDELS

Requirements:

First, you will need to parse the VCF file and extract the relevant information. The VCF file contains information about genetic variations, including single nucleotide polymorphisms (SNPs), insertions, deletions, and structural variants. Each line in the file represents a variant, and the fields in the line contain information about the variant, such as its position in the genome, the reference and alternate alleles, and any additional annotations.

Next, you will need to read the reference genome into memory. The reference genome is typically stored in a file in FASTA format, which consists of a series of DNA sequences with associated labels.

Once you have extracted the variants from the VCF file and the reference genome, you can iterate over the variants and apply the necessary edits to the reference genome. For example, if the variant is a SNP, you can simply replace the base at the specified position with the alternate allele. If the variant is an insertion or deletion, you will need to insert or delete the appropriate number of bases at the specified position.

Finally, you can write the edited genome to a new file or return it as a string, depending on your requirements.

Sample files:

If this this a Reference Genome:

```
>chr1
CCCTGCCAGGGCTGCTGGTGATTCTCCACATCCTTAGGCTCCGCGGTGCTTACCTTCAGG
ACTCTCCAGTTGTAACCCCTTTGTTGGGATGCCTGGGAGCCAGACAAGGTCACCCCATTT
TTTAAGAGAGGACGAAGGTGAGAGGGAGACTACAATGAAAAGGTTGGGAGGGGCCCCAGG
CATGGCCCCTGTGTGTGGAAAACACAGGTGACCACCGGCACCCAGACTGTCTACACTATG
CCTCCAGAAGGCACTTTGCCTAGCAACAGGCCTGACCATGCAGCGCTGGTCCAATCTCTC
>chr2
ACCTGCCAGGGCTGCTGGTGATTCTCCACATCCTTAGGCTCCGCGGTGCTTACCTTCAGG
```

```
ACTCTCCAGTTGTAACCCCTTTGTTGGGATGCCTGGGAGCCAGACAAGGTCACCCCATTT

TTTAAGAGAGGACGAAGGTGAGAGGGAGACTACAATGAAAAGGTTGGGAGGGGCCCCAGG

CATGGCCCCTGTGTGTGGAAAACACAGGTGACCACCGGCACCCAGACTGTCTACACTATG

CCTCCAGAAGGCACTTTGCCTAGCAACAGGCCTGACCATGCAGCGCTGGTCCAATCTCTC

>chr3

TCCTGCCAGGGCTGCTGGTGATTCTCCACATCCTTAGGCTCCGCGGTGCTTACCTTCAGG

ACTCTCCAGTTGTAACCCCTTTGTTGGGATGCCTGGGAGCCAGACAAGGTCACCCCATTT

TTTAAGAGAGGACGAAGGTGAGAGGGAGACTACAATGAAAAGGTTGGGAGGGGCCCCAGG

CATGGCCCCTGTGTGTGGAAAACACAGGTGACCACCGGCACCCAGACTGTCTACACTATG

CCTCCAGAAGGCACTTTGCCTAGCAACAGGCCTGACCATGCAGCGCTGGTCCAATCTCTC
```

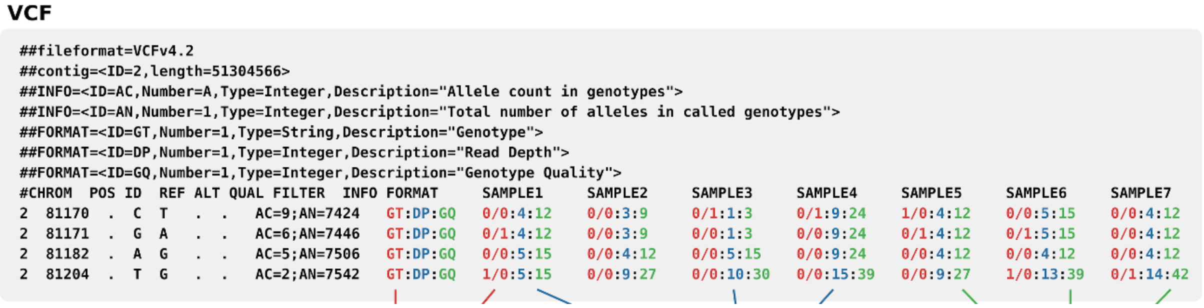
THIS IS A VCF FILE CONTAINING ALL THE VARIANTS

```
##fileformat=VCFv4.2

##fileDate=20090805

##source=myImputationProgramV3.1

##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta

##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="
Homo sapiens",taxonomy=x>

##phasing=partial

##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">

##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">

##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">

##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">

##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">

##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">

##FILTER=<ID=q10,Description="Quality below 10">

##FILTER=<ID=s50,Description="Less than 50% of samples have data">

##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
```

```
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM  POS    ID    REF    ALT    QUAL   FILTER  INFO   FORMAT  NA00001
chr1    3    .    C    A    100    PASS   .    GT    0/1
chr1    30    .    A    GT    100    PASS   .    GT    0/1
chr1    59    .    GG    T    100    PASS   .    GT    0/1
chr1    118    .    TTTTTT  T    100    PASS   .    GT    1/0
chr1    240    .    G    TAC    100    PASS   .    GT    0/1
chr2   120    .    T    AC    100    PASS   .    GT    0/1
chr3   5    .    G    C    100    PASS   .    GT    0/1
chr3   60    .    G    AAAAAA    100    PASS   .    GT    0/1
```

IF there is a 0/0 found in the sample (NA00001) coloum then that means the alternate is homozygous and it can be skipped else the variant is applied (0/1 , 1/0 , 1/1)

Sometimes there are Ns on the reference. If we encounter an N on the reference. Just continue and don't apply the variant.



VCF

```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS ID  REF ALT QUAL FILTER  INFO FORMAT     SAMPLE1     SAMPLE2     SAMPLE3     SAMPLE4     SAMPLE5     SAMPLE6     SAMPLE7
2  81170  .  C  T   .  .   AC=9;AN=7424  GT:DP:GQ  0/0:4:12    0/0:3:9     0/1:1:3     0/1:9:24    1/0:4:12    0/0:5:15    0/0:4:12
2  81171  .  G  A   .  .   AC=6;AN=7446  GT:DP:GQ  0/1:4:12    0/0:3:9     0/0:1:3     0/0:9:24    0/1:4:12    0/1:5:15    0/0:4:12
2  81182  .  A  G   .  .   AC=5;AN=7506  GT:DP:GQ  0/0:5:15    0/0:4:12    0/0:5:15    0/0:9:24    0/0:4:12    0/0:4:12    0/0:4:12
2  81204  .  T  G   .  .   AC=2;AN=7542  GT:DP:GQ  1/0:5:15    0/0:9:27    0/0:10:30   0/0:15:39   0/0:9:27    1/0:13:39   0/1:14:42
```

Variant files are tricky so be careful.

This is the successful outcome of after applying the program you created.

```
>chr1

CCATGCCAGGGCTGCTGGTGATTCTCCACGTTCCTTAGGCTCCGCGGTGCTTACCTTCAT

ACTCTCCAGTTGTAACCCCTTTGTTGGGATGCCTGGGAGCCAGACAAGGTCACCCCAT

AAGAGAGGACGAAGGTGAGAGGGAGACTACAATGAAAAGGTTGGGAGGGGCCCCAGG

CATGGCCCCTGTGTGTGGAAAACACAGGTGACCACCGGCACCCAGACTGTCTACACTATTAC

CCTCCAGAAGGCACTTTGCCTAGCAACAGGCCTGACCATGCAGCGCTGGTCCAATCTCTC

>chr2

ACCTGCCAGGGCTGCTGGTGATTCTCCACATCCTTAGGCTCCGCGGTGCTTACCTTCAGG

ACTCTCCAGTTGTAACCCCTTTGTTGGGATGCCTGGGAGCCAGACAAGGTCACCCCATTAC

TTTAAGAGAGGACGAAGGTGAGAGGGAGACTACAATGAAAAGGTTGGGAGGGGCCCCAGG

CATGGCCCCTGTGTGTGGAAAACACAGGTGACCACCGGCACCCAGACTGTCTACACTATG

CCTCCAGAAGGCACTTTGCCTAGCAACAGGCCTGACCATGCAGCGCTGGTCCAATCTCTC

>chr3

TCCTCCCAGGGCTGCTGGTGATTCTCCACATCCTTAGGCTCCGCGGTGCTTACCTTCAGAAAAAA

ACTCTCCAGTTGTAACCCCTTTGTTGGGATGCCTGGGAGCCAGACAAGGTCACCCCATTT

TTTAAGAGAGGACGAAGGTGAGAGGGAGACTACAATGAAAAGGTTGGGAGGGGCCCCAGG

CATGGCCCCTGTGTGTGGAAAACACAGGTGACCACCGGCACCCAGACTGTCTACACTATG

CCTCCAGAAGGCACTTTGCCTAGCAACAGGCCTGACCATGCAGCGCTGGTCCAATCTCTC
```

The actual Sequence is very Huge. In billions of bases and Gbs in Size.