**Jacky SNP Calling result final clean-up  
Canglin Wu & Zidain Xie**

**July 23, 2015**

**STEP 1:**

**Append two new columns into the csv file (save the new temp csv file):**  
1, zygosity

2, coverage

File name:

temp/WGC033595U\_snps\_temp.csv

temp/WGC033595U\_indels\_temp.csv

**STEP 2:**

**Extracting two columns from VCF file to CSV file (applied to both SNP and Indels files):**

1. In VCF column "INFO", extract "AF=". If AF=1.00, assign "homozygous"; if AF=0.500, assign "heterozygous" to new column "zygosity" in CSV file.
2. In VCF column "INFO", extract "DP=" into new column "coverage" in CSV file. For example, if "DP=19", put "19" in new column "coverage" in CSV file.

**Lookup rules from VCF to CSV files for Extraction** (applied to both SNP and Indels files)**:**

1. Primary lookup: If rs# exists in column "ID" in VCF file (NOTE, “.” means missing), match CSV column "snp138" (NOTE, “.” means missing). If multiple hit, only get the first hit.
2. Secondary lookup: If rs# is “.”in column "snp138" in CSV file. If multiple hit, only get the first hit.

VCF map\_id: CHROM:POS:REF:ALT

CSV map\_id: Chr:Start:Ref:Alt

For example, for the row with "#CHROM" is "chr1", "POS" is "69511", "REF" is "A", "ALT" is "G", its map\_id is "chr1:69511:A:G". Use this VCF map\_id to match CSV map\_id, which is composed of four columns "Chr", "Start", "Ref" and "Alt", such as "chr1:69511:A:G".

**STEP 3:**

**Combine snp and indel CSV files into one CSV file**

File name:

temp/WGC033595U\_combined\_temp.csv

**STEP 4:**

**Final CSV file format (column names may change) for uploading into Labkey should include:**  
chr\_position (Chr:Start, "Chr" column removes prefix "chr")  
dbsnp\_id (snp138)  
gene (Gene.refGene)  
coverage (coverage)  
exon\_number (empty)  
hgvs\_coding (GeneDetail.refGene)  
hgvs\_protein (AAChange.refGene)  
variant\_result\_type (ExonicFunc.refGene)  
transcript\_accession (empty)  
zygosity (zygosity)  
sift (SIFT\_pred)  
polyphen2\_hdiv (Polyphen2\_HDIV\_pred)  
polyphen2\_hvar (Polyphen2\_HVAR\_pred)  
ref (Ref)  
alt (Alt)  
variant\_location (Func.refGene)

**Result file name:**

WGC033595U\_variant\_calling\_final\_result.csv

Java class name:

JackyResultCleanup(subjectID, input\_file\_dir, output\_file\_dir)