Technical Report

This is a technical report about the analyses performed as part of the African genomic diversity based of the workplan in Fig 1.

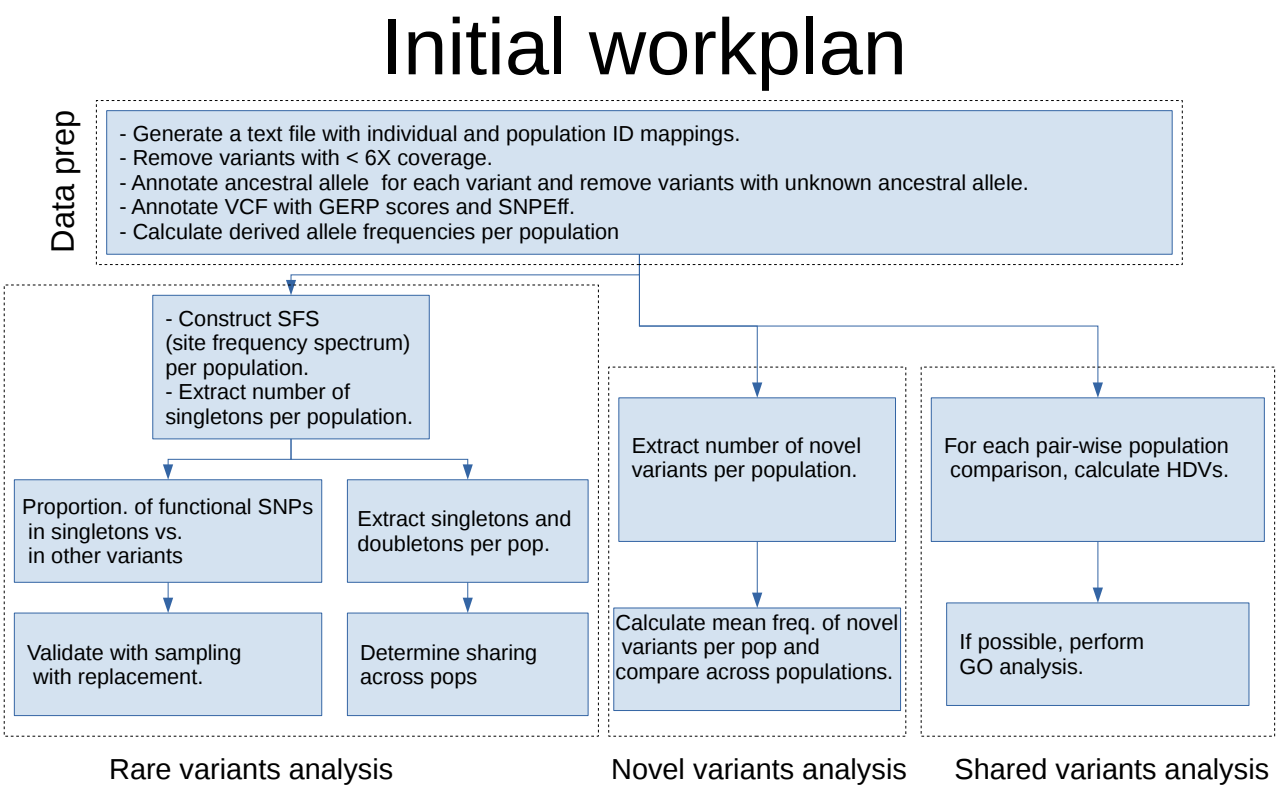

Fig 1. Initial workplan proposed by Laura.

Data preparation

Data to be used by our stream consist of Baylor data only. This dataset was recalled together with sanger and trypanogen, and the phased using Eagle.
The files shared with the genomic diversity analysis group were per chromosome (1 to 22) in VCF format as Eagle.baylor.${CHRM}.vcf.annoted.vcf.hg19_multianno.vcf.gz located in /popdata/gapw/GAPW_DATA/baylor/PHASED/

Generate a text file with individual and population IDs

- Input: /popdata/gapw/GAPW_DATA/baylor/UNPHASED/BAYLOR_UNPHASED.sample
- Output: /spaces/gapw/diversity/mamana/baylor_sample_info_ind_pop_id_only.tsv
- Command:

```
>  awk '{print $1"\t"$2}' ${input} > ${ouput}
```

Remove variants with < 6X coverage

- Input: /spaces/gapw/diversity/mamana/VCF/Eagle.baylor.${CHRM}.vcf.annoted.vcf.hg19_multianno.vcf.gz
- Output: /spaces/gapw/diversity/mamana/VCF_FILTERED/baylor_phased_chr${CHRM}_dp6.vcf.gz
- Command:

```
> bcftools view -i 'DP>6' ${input}  | bgzip -c > ${output}
```

Annotate ancestral allele for each variant and remove variants with unknown ancestral allele

- Input: /spaces/gapw/diversity/mamana/VCF_FILTERED/baylor_phased_chr${CHRM}_dp6.vcf.gz
- Output: /spaces/gapw/diversity/mamana/VCF_FILTERED/baylor_phased_chr${CHRM}_dp6_anc_f.vcf.gz
- Command:
  - Annotate the AA using house-made python script from Laura:

```
> ../scripts/add-ANC-to-vcf_new.py -g --in {input} --out
{output=output1} --genomedata genomedata_path
```

  - Filter for sites with AA only:

```
> bcftools view -i 'AA!="." & AA!="-" & AA!
="N"' {input=output1} | bgzip -c > {output}
```

Annotate VCF with GREP scores and snpEff

GREP scores where not annotate for anymore as there were already annotated for by Emile. We only annotated using snpEff and using snpSift for dbSNP IDs.

- Input: /spaces/gapw/diversity/mamana/VCF_FILTERED/baylor_phased_chr${CHRM}_dp6_anc_f.vcf.gz
- Output: /spaces/gapw/diversity/mamana/VCF_ANN/baylor_phased_chr${CHRM}_dp6_anc_f_dbsnp.vcf.gz and /spaces/gapw/diversity/mamana/VCF_ANN/baylor_phased_chr${CHRM}_dp6_anc_f_dbsnp_snpeff.vcf.gz
- Command:
  - Annotate for dbSNP IDs using snpSift:

```
> snpSift annotate {dbsnp_db} {input} > {output=output1} -v
```

  - Annotate using snfEff:

```
> snpEff -v {human_db} -stats {output}.html -csvStats
{output}.csv -dataDir {snpEff_database} {input=output1} >
{output=output2}
```

  - dbSNP database (dbsnp_db): dbSNP_human_9606_b150_GRCh37p13.vcf
  - human_db: hg19
  - snpEff_database: location of your snpEff database if not in default snpEff path

Calculate derived allele frequencies per population

1. Split sample file in population

- Input: /spaces/gapw/diversity/mamana/baylor_sample_info_ind_pop_id_only.tsv
- Output (pop_file): ../samples/{POP}.sample, for POP in Benini, Botswana, Burkina, Cameroon, Ghana, Mali, Nigeria, Zambia
- Command:

```
def split_sample_list_per_pop(POP, input_file, output_file):
    '''
    Read baylor_sample_info_ind_pop_id_only.tsv and
    split it by population
    '''
    i = 1 #line no
    data = []
    message("Extracting "+POP+' to '+output_file)
    for line in open(input_file):
        if i > 1:
            line = line.strip().split()
            POP_ = line[1].strip()
            if POP_ == POP:
                data.append('\t'.join(line)+'\n')
        i += 1
    if len(data) > 0:
        out = open(output_file, "w")
        for elt in data:
            out.writelines(elt)
        out.close()
```

2. Split into population
   - Input: /spaces/gapw/diversity/mamana/VCF_ANN/baylor_phased_chr${CHRM}_dp6_anc_f_dbsnp_snpeff.vcf.gz
   - Output: /spaces/gapw/diversity/mamana/VCF_POP/{POP}_phased_chr${CHRM}_dp6_anc_f_dbsnp_snpeff.vcf.gz
   - Command:

```
> vcftools \
    --gzvcf {input} \
    --keep {pop_file} \
    --recode --recode-INFO-all --mac 0 -c | \
    bgzip -c > {output}
```

3. Calculate derived allele frequencies
   - Input: /spaces/gapw/diversity/mamana/VCF_POP/{POP}_phased_chr${CHRM}_dp6_anc_f_dbsnp_snpeff.vcf.gz
   - Output: /spaces/gapw/diversity/mamana/VCF_POP/{POP}_phased_chr${CHRM}_dp6_anc_f_dbsnp_snpeff.daf.frq
   - Command:

```
> vcftools \
    --gzvcf {input} \
    --freq --derived \
    --out {output}
```