Copy of Creation of TWE AF vcf (TWE_POPAF_N500_chr1-22_220202.vcf.gz)

This page describes the process performed to create a vcf containing population AFs for the TWE assay.

Identification of suitable vcfs

List out all TWE projects:

```
for project in $(dx find projects --name "002_*_TWE" --brief); do
name=$(dx describe $project --name); echo -e "${project}\t${name}";
done | sort -k2V
```

Project ids of passing runs above saved to a file:

```
cat project_list_220201 (IDs removed)
project-ID
project-ID
project-ID
project-ID
project-ID
project-ID
project-ID
project-ID
```

Get vcf files ids from these projects:

```
dt=$(date '+%Y%m%d')

# For each target project
while read project; do

        # Find the dias single output folders and write them to a file
        dx ls --folders ${project}:/output/ | grep dias_single >
output_folders

        # Count how many were found
        single_count=$(cat output_folders | grep -c "dias_single")

        # For each dias single folder
        while read dias_single; do

                # Find Sentieon output vcf file ids
                command="dx find data --path ${project}:/output
/${dias_single}/sentieon-dnaseq/ --name 'X*r.vcf.gz' --brief"
                eval $command
        done < output_folders

# Write vcf file ids to a file
done > vcf_files_list_${dt} < project_list_220201
```

Check vcf files list for duplicate samples:

```
# Add extra fields for separate identifiers which will be checked for
duplicates
while read vcf_id; do
  name=$(dx describe $vcf_id --name)
  sampleid=$(echo $name | awk -F "-" '{print $1}')
  gm=$(echo $name | awk -F "-" '{print $2}')
  echo -e "${vcf_id}\t${name}\t${sampleid}\t${gm}"
done < vcf_files_list_20220201 > vcf_files_list_20220201_info

head vcf_files_list_20220201_info (IDs removed)
project-ID:file-ID        SampleID-PatientID-TWE-N-
EGG4_markdup_recalibrated_Haplotyper.vcf.gz        SampleID
PatientID
project-ID:file-ID        SampleID-PatientID-TWE-N-
EGG4_markdup_recalibrated_Haplotyper.vcf.gz        SampleID
PatientID
project-ID:file-ID        SampleID-PatientID-TWE-N-
EGG4_markdup_recalibrated_Haplotyper.vcf.gz        SampleID
PatientID
project-ID:file-ID        SampleID-PatientID-TWE-N-
EGG4_markdup_recalibrated_Haplotyper.vcf.gz        SampleID
```

```
PatientID
project-ID:file-ID        SampleID-PatientID-TWE-N-
EGG4_markdup_recalibrated_Haplotyper.vcf.gz        SampleID
PatientID
project-ID:file-ID        SampleID-PatientID-TWE-N-
EGG4_markdup_recalibrated_Haplotyper.vcf.gz        SampleID
PatientID
project-ID:file-ID        SampleID-PatientID-TWE-N-
EGG4_markdup_recalibrated_Haplotyper.vcf.gz        SampleID
PatientID
project-ID:file-ID        SampleID-PatientID-TWE-N-
EGG4_markdup_recalibrated_Haplotyper.vcf.gz        SampleID
PatientID
project-ID:file-ID        SampleID-PatientID-TWE-N-
EGG4_markdup_recalibrated_Haplotyper.vcf.gz        SampleID
PatientID
project-ID:file-ID        SampleID-PatientID-TWE-N-
EGG4_markdup_recalibrated_Haplotyper.vcf.gz        SampleID
PatientID
project-ID:file-ID        SampleID-PatientID-TWE-N-
EGG4_markdup_recalibrated_Haplotyper.vcf.gz        SampleID
PatientID

# Check for duplicate samples (IDs removed)
cut -f 4 vcf_files_list_20220201_info | sort | uniq -c | sort -k1n |
tail
      1 PatientID
      1 PatientID
      1 PatientID
      1 PatientID
      1 PatientID
      1 PatientID
      1 PatientID
      1 PatientID
      1 PatientID
      2 PatientID

# Just one sample is present twice - PatientID (ID removed)
grep PatientID vcf_files_list_20220201_info
project-ID-ID        SampleID-PatientID-TWE-F-
EGG4_markdup_recalibrated_Haplotyper.vcf.gz        SampleID
PatientID
project-ID-ID        SampleID-PatientID-TWE-F-
EGG4_markdup_recalibrated_Haplotyper.vcf.gz        SampleID
PatientID

# Download both files (IDs removed)
dx download file-ID -o file-ID.gz
dx download file-ID -o file-ID.gz
```

```
# Compare
diff file-ID1 file-ID2
26,27c26,27
< ##SentieonCommandLine.GVCFtyper=<ID=GVCFtyper,Version="sentieon-
genomics-201911",Date="2021-11-30T16:32:05Z",CommandLine="/usr/local
/sentieon-genomics-201911/libexec/driver -t 36 -r genome/hs37d5.fa --
interval ignore_decoy.bed --algo GVCFtyper -d resources/dbsnp_138.b37.
vcf.gz -v haplotyper.g.vcf.gz haplotyper.vcf.gz">
< ##SentieonCommandLine.Haplotyper=<ID=Haplotyper,Version="sentieon-
genomics-201911",Date="2021-11-30T16:27:52Z",CommandLine="/usr/local
/sentieon-genomics-201911/libexec/driver -t 36 -r genome/hs37d5.fa -i
markdup.bam -q recal_data_Sentieon.table --interval ignore_decoy.bed --
algo Haplotyper -d resources/dbsnp_138.b37.vcf.gz --emit_mode GVCF
haplotyper.g.vcf.gz">
---
> ##SentieonCommandLine.GVCFtyper=<ID=GVCFtyper,Version="sentieon-
genomics-201911",Date="2021-11-03T13:39:31Z",CommandLine="/usr/local
/sentieon-genomics-201911/libexec/driver -t 36 -r genome/hs37d5.fa --
interval ignore_decoy.bed --algo GVCFtyper -d resources/dbsnp_138.b37.
vcf.gz -v haplotyper.g.vcf.gz haplotyper.vcf.gz">
> ##SentieonCommandLine.Haplotyper=<ID=Haplotyper,Version="sentieon-
genomics-201911",Date="2021-11-03T13:35:21Z",CommandLine="/usr/local
/sentieon-genomics-201911/libexec/driver -t 36 -r genome/hs37d5.fa -i
markdup.bam -q recal_data_Sentieon.table --interval ignore_decoy.bed --
algo Haplotyper -d resources/dbsnp_138.b37.vcf.gz --emit_mode GVCF
haplotyper.g.vcf.gz">

# Files are identical except header timestamps, so we will remove one
from the vcf file list
grep -v file-ID vcf_files_list_20220201_info >
vcf_files_list_20220201_info_nodup

# Upload the vcf file list to DNAnexus (ID removed)
dx select project-ID

dx upload vcf_files_list_20220201_info_nodup --path /TWE/
ID                  file-ID
Class               file
Project             project-ID
Folder              /TWE
Name                vcf_files_list_20220201_info_nodup
State               closing
Visibility          visible
Types               -
Properties          -
Tags                -
Outgoing links      -
Created             Tue Feb  1 11:18:06 2022
Created by          garnerm
Last modified       Tue Feb  1 11:18:07 2022
```

```
Media type
archivalState       "live"
cloudAccount        "cloudaccount-dnanexus"
```

Creation of TWE pop_AF vcf

Now we use the vcf file list in a cloud workstation to merge the selected vcfs and calculate and add population AF annotation

```
# Clone past job as a shortcut to override instance type as I cannot
remember how to override it.
# This will need a big instance (IDs removed)
dx run cloud_workstation --clone job-ID --allow-ssh --priority=high -
imax_session_length=12h

# ssh to job (ID from output of above)
dx ssh job-ID

# Enable upload
unset DX_WORKSPACE_ID
dx cd $DX_PROJECT_CONTEXT_ID:

# Install bcftools
wget https://github.com/samtools/bcftools/releases/download/1.14
/bcftools-1.14.tar.bz2
tar -xvjf bcftools-1.14.tar.bz2
cd bcftools-1.14
make
cd -

# Install HTSlib
wget https://github.com/samtools/htslib/releases/download/1.14/htslib-
1.14.tar.bz2
tar -xvjf htslib-1.14.tar.bz2
cd htslib-1.14
make
cd -

# Download genome fasta
wget http://www.broadinstitute.org/ftp/pub/seq/references
/Homo_sapiens_assembly19.fasta
wget http://www.broadinstitute.org/ftp/pub/seq/references
/Homo_sapiens_assembly19.fasta.fai

# Download file list (ID of list from above)
dx download file-ID

# Make a folder for the vcf files we will download
mkdir vcfs
```

```
cd vcfs

# Download 500 vcfs
for file in $(cut -f 1 ../vcf_files_list_20220201_info_nodup | head -n
500); do dx download $file; done

# Index vcfs
for vcf in $(ls *vcf.gz); do ../bcftools-1.14/bcftools index $vcf; done
cd -

# Enable bcftools plugins
export BCFTOOLS_PLUGINS=/home/dnanexus/bcftools-1.14/plugins/

# Merge first 250 vcfs
command="bcftools-1.14/bcftools merge --output-type v -m none --missing-
to-ref"
# Add the vcf files names to the command
for vcf in $(ls vcfs/*vcf.gz | head -n 250); do command="${command}
$vcf"; done
command="${command} > merge1-250.vcf"
eval $command

# bgzip and index
htslib-1.14/bgzip merge1-250.vcf
bcftools-1.14/bcftools index merge1-250.vcf.gz

# Merge second 250 (depends on exact number needed)
command="bcftools-1.14/bcftools merge --output-type v -m none --missing-
to-ref"
# Add the vcf files names to the command
for vcf in $(ls vcfs/*vcf.gz | tail -n 250); do command="${command}
$vcf"; done
command="${command} > merge251-500.vcf"
eval $command

# bgzip and index
htslib-1.14/bgzip merge251-500.vcf
bcftools-1.14/bcftools index merge251-500.vcf.gz

# Merge the two batches, pipe additional steps for speed until sort
step
command="bcftools-1.14/bcftools merge --output-type u -m none --missing-
to-ref merge1-250.vcf.gz merge251-500.vcf.gz"

# Norm
command="${command} | bcftools-1.14/bcftools norm -m -any -f
Homo_sapiens_assembly19.fasta -Ou"

# Add tags
command="${command} | bcftools-1.14/bcftools +fill-tags --output-type v
```

```
-o merge_tag.vcf -- -t AN,AC,NS,AF,MAF,AC_Hom,AC_Het,AC_Hemi"

# Sort
command="${command} ; bcftools-1.14/bcftools sort merge_tag.vcf -Oz >
TWE_POPAF_N500_220202.vcf.gz"

# Index
command="${command} ; htslib-1.14/tabix -p vcf TWE_POPAF_N500_220202.
vcf.gz"

# Upload
command="${command}; dx upload --path /TWE TWE_POPAF_N500_220202.vcf.
gz; dx upload --path /TWE TWE_POPAF_N500_220202.vcf.gz.tbi"
eval $command
```

Remove X/Y chromosome variants

After creation of initial files the decision was made (with input from RD team) to exclude X/Y chroms since the assumption of diploid ploidy during variant calling can cause AFs on X/Y to be inaccurate.

```
dx run cloud-workstation

dx ssh job-ID (ID from output of above)

unset DX_WORKSPACE_ID
dx cd $DX_PROJECT_CONTEXT_ID:

# Download TWE pop AF vcfs
dx download file-ID # TWE_POPAF_N500_220202.vcf.gz
dx download file-ID # TWE_POPAF_N500_220202.vcf.gz.tbi

# Install bcftools
wget https://github.com/samtools/bcftools/releases/download/1.14
/bcftools-1.14.tar.bz2
tar -xvjf bcftools-1.14.tar.bz2
cd bcftools-1.14
make
cd -

# Install HTSlib
wget https://github.com/samtools/htslib/releases/download/1.14/htslib-
1.14.tar.bz2
tar -xvjf htslib-1.14.tar.bz2
cd htslib-1.14
make
cd -

# bcftools view regions to extract chroms we want
#
# Usage:
# -r, --regions chr|chr:pos|chr:beg-end|chr:beg-[,…]
# Comma-separated list of regions
#
# Keep chr 1-22, and drop X,Y,MT and the additional non-localised
contigs
bcftools-1.14/bcftools view -r
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22
TWE_POPAF_N500_220202.vcf.gz  -Oz > TWE_POPAF_N500_chr1-22_220202.vcf.gz

# Index
htslib-1.14/tabix -p vcf TWE_POPAF_N500_chr1-22_220202.vcf.gz

dx cd /TWE
dx upload TWE_POPAF_N500_chr1-22_220202.vcf.gz
dx upload TWE_POPAF_N500_chr1-22_220202.vcf.gz.tbi
```

Sanity check the final vcf

```
# First few records
zcat TWE_POPAF_N500_chr1-22_220202.vcf.gz | grep -v ^# | head | cut -f
1-8
1	10110	.	A	AAC	53.7	.
ExcessHet=3.0103;FS=0;MQ=45.1;QD=26.85;SOR=0.693;DP=4;AF=0.002;MLEAC=2;
MLEAF=1;AN=1000;AC=2;NS=500;MAF=0.002;AC_Het=0;AC_Hom=2;AC_Hemi=0
1	10146	rs375931351	AC	A
37.37	.	DB;ExcessHet=3.0103;FS=0;MQ=54;QD=16.35;SOR=0.693;
DP=5;AF=0.004;MLEAC=2;MLEAF=1;AN=1000;AC=4;NS=500;MAF=0.004;AC_Het=0;
AC_Hom=4;AC_Hemi=0
1	10390	.
CCCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA	C
43.7	.	ExcessHet=3.0103;FS=0;MQ=60;QD=15.35;SOR=2.303;
DP=4;AF=0.004;MLEAC=2;MLEAF=1;AN=1000;AC=4;NS=500;MAF=0.004;AC_Het=0;
AC_Hom=4;AC_Hemi=0
1	10403	.
ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC	A	53.7
.	ExcessHet=3.0103;FS=0;MQ=50.91;QD=26.85;SOR=0.693;DP=3;AF=0.
002;MLEAC=2;MLEAF=1;AN=1000;AC=2;NS=500;MAF=0.002;AC_Het=0;AC_Hom=2;
AC_Hemi=0
1	10409	.	ACCCTAACCCTAACCCTAACCCTAACCCTAAC
A	43.7	.	ExcessHet=3.0103;FS=0;MQ=52;QD=21.85;
SOR=0.693;ClippingRankSum=-0;MQRankSum=-0.431;DP=6;AF=0.003;MLEAC=1;
MLEAF=0.5;AN=1000;AC=3;NS=500;MAF=0.003;AC_Het=1;AC_Hom=2;AC_Hemi=0
1	10489	.	C	T	32.74	.
ExcessHet=3.0103;FS=0;MQ=46.82;QD=16.37;SOR=0.693;DP=2;AF=0.002;MLEAC=2;
MLEAF=1;AN=1000;AC=2;NS=500;MAF=0.002;AC_Het=0;AC_Hom=2;AC_Hemi=0
1	10492	rs55998931	C	T	62.74
.	DB;ExcessHet=3.0103;FS=0;MQ=50.91;QD=31.37;SOR=0.693;
BaseQRankSum=-0.385;ClippingRankSum=-0;MQRankSum=-0.674;ReadPosRankSum=-
0.674;DP=14;AF=0.005;MLEAC=2;MLEAF=1;AN=1000;AC=5;NS=500;MAF=0.005;
AC_Het=1;AC_Hom=4;AC_Hemi=0
1	10581	.	G	A	62.74	.
ExcessHet=3.0103;FS=0;MQ=50.99;QD=31.37;SOR=0.693;DP=2;AF=0.002;MLEAC=2;
MLEAF=1;AN=1000;AC=2;NS=500;MAF=0.002;AC_Het=0;AC_Hom=2;AC_Hemi=0
1	10583	rs58108140	G	A	62.74
.	BaseQRankSum=-1.282;ClippingRankSum=-0;DB;ExcessHet=3.0103;
FS=0;MQ=60;MQRankSum=-0;QD=11.15;ReadPosRankSum=-0;SOR=0.223;DP=9;AF=0.
005;MLEAC=2;MLEAF=1;AN=1000;AC=5;NS=500;MAF=0.005;AC_Het=1;AC_Hom=4;
AC_Hemi=0
1	10616	rs376342519	CCGCCGTTGCAAAGGCGCGCCG
C	143	.	DB;ExcessHet=3.0103;FS=0;MQ=45.39;QD=35.75;
SOR=0.693;DP=16;AF=0.01;MLEAC=2;MLEAF=1;AN=1000;AC=10;NS=500;MAF=0.01;
AC_Het=0;AC_Hom=10;AC_Hemi=0

# Records per chrom
zcat TWE_POPAF_N500_chr1-22_220202.vcf.gz | grep -v ^# | cut -f 1 |
sort | uniq -c | sort -k2V
1433839 1
```

```
1309741 2
1053907 3
 989697 4
 926027 5
 990255 6
 965863 7
 804805 8
 727826 9
 826377 10
 802416 11
 816609 12
 511497 13
 550443 14
 548733 15
 636415 16
 646763 17
 417637 18
 621449 19
 401274 20
 246947 21
 318050 22
```