eMERGEseq Mulitsample, Freeze 2 Final Genotypes Release, N=24,956

Ian Stanaway, Jodell Jackson, Kayla Howell, Gail Jarvik, & David Crosslin

e3helpme@uw.edu

September 9, 2019

Contents

List of Tables

List of Figures

1 Abstract

The data production and implementation of eMERGESeq is at the final freeze 2 release of sequencing based genotyping of 24,956 participants, to describe the genetic variation in the eMERGESeq cohort. The Genomic Data Coordinating Center quality control performed includes removing duplicates and mislabeled sample files, Principle Component Analysis (PCA) of genetic ancestry, qualitative comparison to self-reported ancestry, Identity by Descent (IBD) relatedness assessments, genetic to reported gender checks, genotype concordance to the clinical sequencing variants reported, and quantifying call rates in samples and variants. We provide data outlined in the following list.

1. eMERGESeq Participants' Variants in Multi-sample VCF format;
2. eMERGESeq Participants' Variants in Single-sample VCF format as provided by the Sequencing Centers (Broad and Baylor);
3. Principal Components Analysis of Participants' Variants;
4. Self-reported or Observed Ancestry;
5. Genetic Ancestry;
6. Variant Frequencies;
7. Low call rate variant sites;
8. IBD estimates;
9. Genetic to Reported Gender Checks;
10. Multisample Variants to Reported Clinical Variant Detection Concordance;

2 Data sets (including support files)

2.1 eMERGEseq Multisample VCF

2.2 SFTP Location of eMERGEseq Files

2.2.1 Multisample VCF ($n = 24,956$ participants)

Filename:
merged_eMERGEseq_samples.variants.chr-sorted.snps-recal.consented_clean.vcf.gz

2.2.2 Single Sample VCFs

Folder name:
single_sample_clinical_vcfs/*

Note: there are many files in this folder and it will take awhile to list them if you enter the 'ls' command. Just use the 'mget' command to avoid stalling the sftp connection.

Medical Center Site	Number of Participants
Columbia	2,580
CCHMC	2,956
CHOP	2,991
Geisinger	2,498
KP/UW	2,499
Harvard	2,493
Meharry	495
Mayo Clinic	3,020
Northwestern University	2,984
Vanderbilt University	2,440
Total	**24,956**

Table 1: Number of participants per eMERGE site

2.2.3 Call Rate Files

Filenames:

low_call_rate_variant_sites
low_call_rate_samples (0)
samples_variant_count
samples_variant_allele_hist
samples_missing_variant_allele_hist

2.2.4 Genetic Ancestry and Frequency Files

Filenames:

samples.genetic_ancestry.asian (n=4,157)
samples.genetic_ancestry.african (n=1,800)
samples.genetic_ancestry.european (n=18,999)

asian_site_counts.frq.count
asian_site_freqs.frq
african_site_counts.frq.count
african_site_freqs.frq
european_site_counts.frq.count
european_site_freqs.frq

site_counts.frq.count (Joint count)
site_freqs.frq (Joint frequencies)

2.2.5 Sample Annotation Files:

Eigenvectors 1-10 using MAF and MISS = 5% (see Section 3.2.2)

Filenames:
eMERGEseq_samples.csv
Data Dictionary
SUBJID eMERGE ID
SEX gender code
DECADE_BIRTH
YEAR_BIRTH
ETHNICITY ethnicity code
RACE self-reported race code
self_reported_race self-reported race
ethnicity2 self-reported ethnicity
sex2 gender (male/female)
site eMERGE medical center site
eig1
eig2
eig3
eig4
eig5
eig6
eig7
eig8
eig9
eig10

Eigenvectors 1-10 using no MAF and MISS filter (see Section 3.2.2)

Filenames:
eMERGEseq_samples_rare_variant_PCs.csv

Data Dictionary
SUBJID eMERGE ID
SEX gender code
DECADE_BIRTH
YEAR_BIRTH
ETHNICITY ethnicity code
RACE self-reported race code
self_reported_race self-reported race
ethnicity2 self-reported ethnicity
sex2 gender (male/female)
site eMERGE medical center site
eig1
eig2

eig3
eig4
eig5
eig6
eig7
eig8
eig9
eig10

3 Multi-sample VCF Methods

3.1 Current Multi-sample VCF Pipeline

Reference file: 1000 Genome phase two reference assembly sequence hs37d5.fa
Target file: eMERGE_target_combined_results.list (Note: This is the union of the Broad and Baylor target files.)
dbSNP build 138: dbSNP138.vcf

Files used for VQSR
hapmap_3.3.b37.vcf
1000G_omni2.5.b37.vcf
1000G_phase1.snps.high_confidence.b37.vcf
dbsnp_138.b37.vcf
dbsnp_138.b37.excluding_sites_after_129.vcf

1. Converting Bam file to Fastq file format:
Software: picard-tools; Version: 1.90
Tool: SamToFastq.jar

2. Map and process single samples:
##a. Map fastq files
Software: bwa; Version: 0.7.10
Tool: bwa mem

Software: Picard; Version: 1.90
Tool: SamFormatConverter.jar

##b. Sort Bam file
Software: Picard; Version: 1.90
Tool: SortSam.jar

Software: samtools; Version: 0.1.19
Tool: index

##c. Merge sample
Software: Picard; Version: 1.90
Tool: MarkDuplicates.jar

##d. List of suspicious indel intervals
Software: GATK; Version 3.5
Tool: GenomeAnalysisTK.jar -T RealignerTargetCreator

##e. Realign indels
Software: GATK; Version 3.5
Tool: GenomeAnalysisTK.jar -T IndelRealigner

##f. Calculate Recalibration Matrix
Software: GATK; Version 3.5
Tool: GenomeAnalysisTK.jar -T BaseRecalibrator

##g. Calling gVCF files
Software: GATK; Version 3.5
Tool: GenomeAnalysisTK.jar -T HaplotypeCaller
–output_mode EMIT_ALL_SITES
produces calls at any callable site regardless of confidence

3. Calling multisample VCFs:
If Number of samples per dataset is greater than 200 combine the individual gVCF files
##a. Combine gVCF files Software: GATK; Version 3.5 Tool: GenomeAnalysisTK.jar -T CombineGVCFs

##b. Joint genotyping on gVCF files
Software: GATK; Version 3.5
Tool: GenomeAnalysisTK.jar -T GenotypeGVCFs

##c. Merge VCF files Software: bcftools; Version 1.3.1
Tool: bcftools concat

4. Applying filters:
Software: GATK; Version 3.5 Tool: GenomeAnalysisTK.jar -T VariantRecalibrator
Software: GATK; Version 3.5 Tool: GenomeAnalysisTK.jar -T ApplyRecalibration

3.2 QC Performed

3.2.1 Principal Component Analysis (PCA)

Using the R package SNPRelate, we performed a principal component analysis on the multisample VCF. There were a total of 62,050 variants. We LD pruned (`snpgdsLDpruning`) the biallelic SNVs. We excluded SNVs with MAF $< 5\%$, or a missing rate $> 5\%$. We also pruned at $r = 0.84$. This

resulted in the selection of 1,571 SNPs. The scree plot (variance explained by eigenvector) and principal component plots for this analysis can be found in Figures 1 and 2, respectively. Next, we used the default `snpgdsLDpruning` settings that has no MAF or missing rate filters, and a LD threshold of $r = 0.2$. This resulted in the selection of 48,845 SNVs. The scree plot (variance explained by eigenvector) and principal component plots for this analysis can be found in Figures 3 and 4, respectively. For both analyses, the self-reported race and genetically-determined race appear to generally match. The first 10 eigenvectors from both analyses are provided in the sample annotation files. We used the kmeans() R function with three means and the common variant PC1 and PC2 to determine genetic ancestry groups.

Scree and PCA Plots for the eMERGEseq Multisample VCF

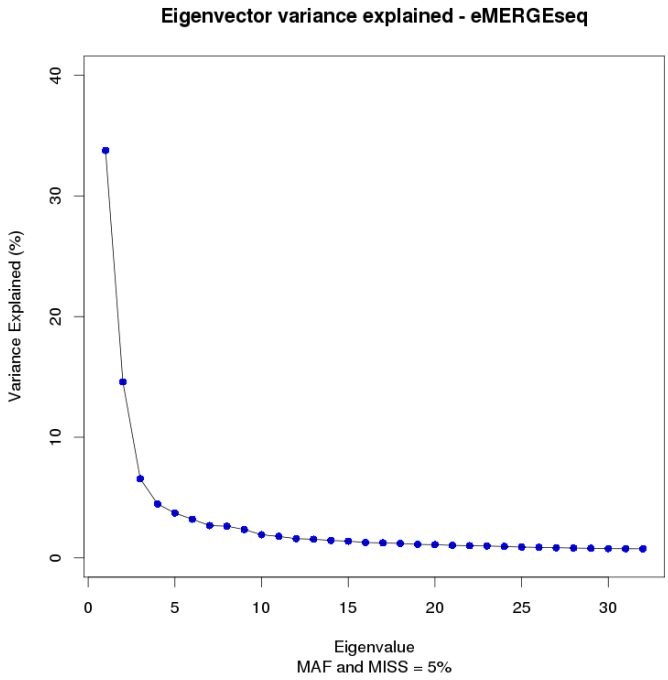

Figure 1: Scree plot of PCA using MAF and missingness of 5%

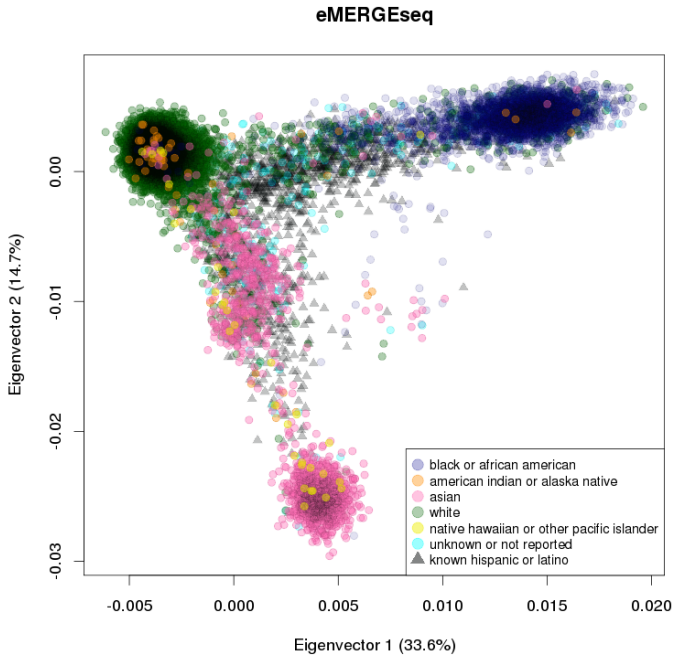

Figure 2: PCA plot using MAF and missingness of 5%

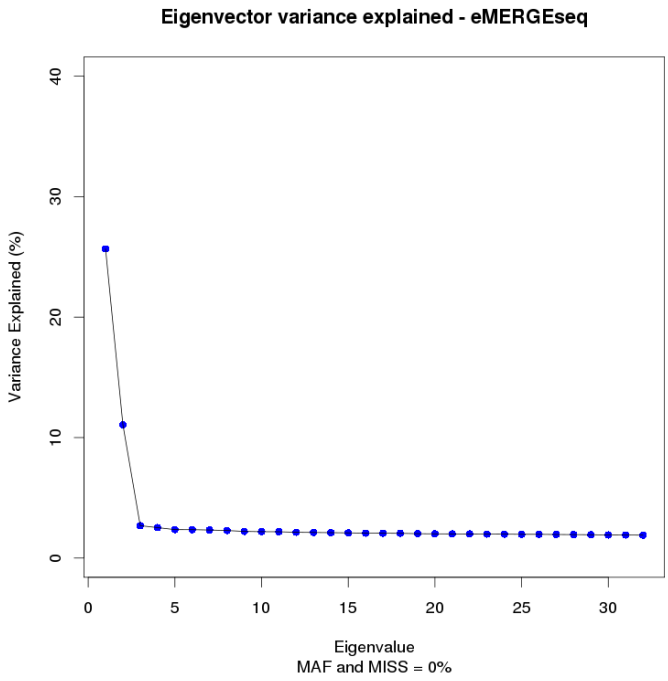

Figure 3: Rare variant scree plot of PCA using missingness of 5%

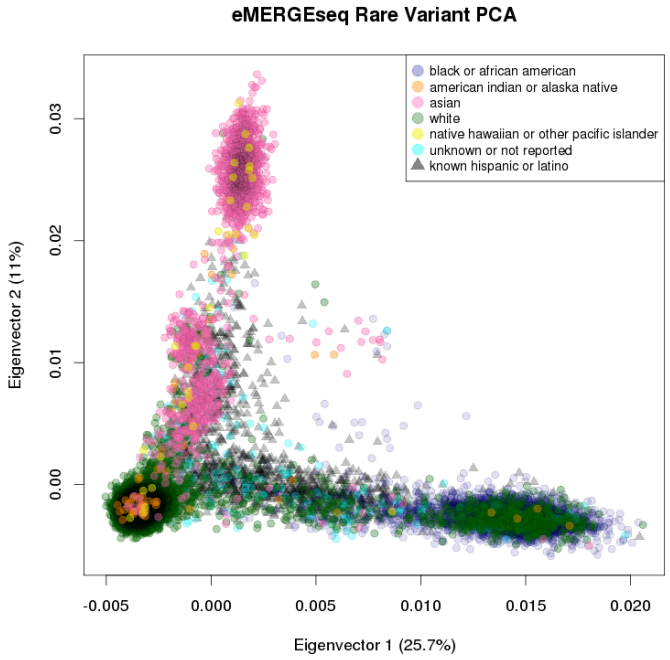

Figure 4: Rare variant PCA plot using missingness of 5%

3.2.2 Identity By Descent (IBD) Analysis

We performed IBD using the PLINK1.9 –genome function. First we pruned the VCF to 0.05 minor allele frequency and the LD pruned with the PLINK settings –geno 0.1 –mind 0.1 –indep-pairwise 1000 50 0.7 using the –exclude exclusion_regions_hg19.txt provided with the PLINK documentation. The resulting pruned file was then use to perform IBD calcualtions. We plotted the IBD in Figure 5.

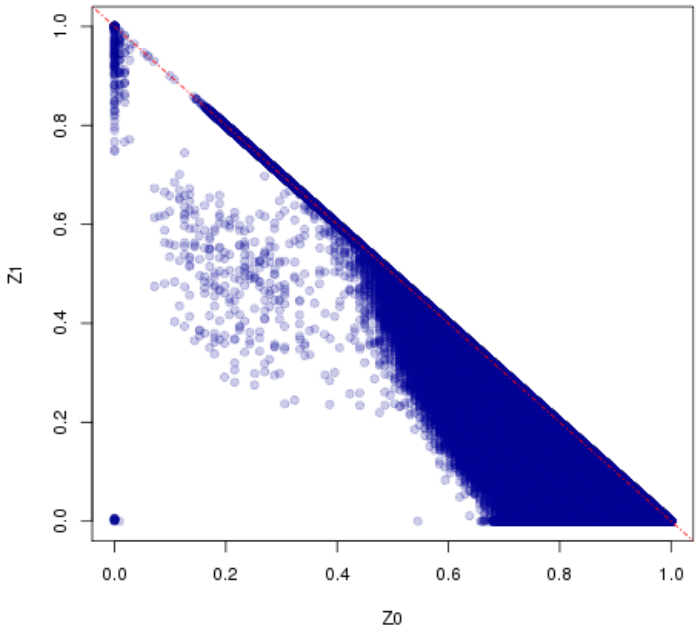

Figure 5: Identity by Descent (IBD) using PLINK

We provide the results files for IBD:

plink.ibd.genome
plink.ibd.genome.Z0_lt_0.15.Z1_lt_0.3.suspect_duplicates (14 pairs)
plink.ibd.genome.Z0_lt_0.15.Z1_gt_0.7.parent-child (43764 pairs)

Note, the relatedness IBD metric is inflated. This is visible at the wedge appearing in the bottom portion of the IBD plot.

4 Multi-sample VCF to Clinical Reported Variant Concordance

4.1 Baylor Sequenced Subjects

The Baylor Sequencing Center provided the file:

variantSearchExport_Baylor_20190513.csv

This file contains the participant subject eMERGE ID, chromosome, position, genotype calls and annotation information of 920 variants which were reported to participants as medically significant. A concordance analysis was analyzed using a custom perl script to interpret the results between this file and the Multi-sample VCF created for this release report. We were able to find 99.3% (914 / 920) of variant locations reported in the Baylor file. Of the 6 variants not found manual review of the Baylor calls show that 2 of the variants were somatic Mosaic and another 3 were Heterozygous calls, all of which had a low allele fraction of reads ($\sim 25\%$) contributing to the variants called. Given this low allele fraction, the non-detection in the multi-sample is not surprising. The last 1 of the 6 non-detect variants had a typical heterozygous allele balance ratio. These variants not called in the multi-sample VCF are summarized in this file:

Baylor_reported_variants_not_detected_in_multisample_vcf

We then inspected the allele variant identity at the 914 variants found. Raw concordance is $\sim 0.889 = 813/914$. Upon inspection of the 101 discordant genotypes it was found that 17 of the calls by the sequencing center were called homozygous non-reference, but but in error the homozygous reference allele genotypes were reported by the sequencing center rather than the alternate alleles as homozygous. Eighty-two (82) discordant genotypes reported were insertion-deletion variants that were correctly called as to their zygosity, but the left or right justification of the variants in the sequence or the repeat structure of the variants were slightly different. We found 2 variant subjects where the homozygous reference allele was called in the multi-sample VCF where the reported clinical variants were heterozygous insertion-deletions. These discordant and manully reviewed for condordant variant zygosity but differing allele identity variant subjects' variants are summarized in these files for the Baylor calls:

Baylor_reported_variants_detected_in_multisample_vcf.discordant
Baylor_reported_variants_detected_in_multisample_vcf.discordant.concordant
Baylor_reported_variants_detected_in_multisample_vcf.discordant.discordant

Given the non-detection of 6 variants and called discordance of 2 variants present in the file at the same genomic location the overall concordance is $\sim 0.991 = 912/920$.

4.2 Broad Sequenced Subjects

The Broad Sequencing Center provided the file:

Variants_reported_LMM_Broad_20190510_updated.csv
Variants_reported_LMM_Broad_20190510_updated.xlsx

This file contains the participant subject eMERGE ID, transcript cdna NMid, cdna position, genotype calls with reference to the cdna genomic strand and annotation information of 802 variants which were reported to participants as medically significant. A concordance analysis was analyzed using a custom perl script to interpret the results between this file and the Multi-sample VCF created for this release report. Since the genomic positions were not reported (we asked for this also), the analyst had to compute the genomic locations of the variants from the NCBI gene_seq.md reference file.

Among the variants reported we were not able to locate 152 of these in a programmatic automated fashion, the majority of these variants (137) were insertion-deletion variants. Manual inspection of the inferred locations based on the cdna position counting showed that adjacent insertion-deletion variants of similar sequence identity were detected but the locations were slightly offset and the left or right justification of them in the sequence context was different between the two call sets. We decided to exclude these 137 insertion deletions from our concordance analysis for summary purposes due to this location ambiguity. One sample had two variants reported in the same line: LDLR:c.[1690A>C;2397_2405delCGTCTTCCT] the second of which we are not able to find the location for programmatically. It is not known if these variants are in cis or trans and the second is an insertion-deletion. We will only search for the first given our ability to find it in the reference.

These variants not called in the multi-sample VCF are summarized in this file:

Broad_reported_variants_not_detected_in_multisample_vcf

Eleven (11) variants were not able to be searched at all due to ambiguous locations reported by the Broad sequencing center. These are additionally summarized in this file:

Variants_reported_LMM_Broad_20190510_updated_ambiguous_locations_not_searchable.csv

We were able to find $\sim$ 81.0% (650 / 802) of sample-variant genomic locations reported in the Broad file in the multi-sample VCF.

Inspection of the 650 the allele variants with found genomic locations detected showed a raw concordance of $\sim 0.969 = 630/650$. Upon inspection of the 20 discordant genotypes it was found that 11 of the calls by the sequencing center were correctly called as to their zygosity, but the allele identity of the variants were slightly different, mostly due to insertion-deletions. Only one of these 11 variants was a snv, and for that variant the reported text string was not standardized like the others and was actually concordant. We found 9 variant subjects where the homozygous reference allele was called in the multi-sample VCF where the reported clinical variants were heterozygous insertion-deletion variants. These discordant and manually reviewed for condordant variant zygosity, but differing allele identity among subjects' variants are summarized in these files for the Broad calls:

Broad_reported_variants_detected_in_multisample_vcf.discordant
Broad_reported_variants_detected_in_multisample_vcf.discordant.concordant
Broad_reported_variants_detected_in_multisample_vcf.discordant.discordant

We have excluded 137 insertion-deletions and 11 other variants with ambiguous genomic locations. For the total denominator we have 650 variants found, plus the remaining non-detection of 4 variants. Manual review of 20 programmatically called discordant showed 11 to be concordant and 9 variants to be discordant. Given this the overall concordance is $\sim 0.98 = 641/654$.

In summary, the clinically reported single nucleotide variants showed good concordance and detection across both data sets in the multi-sample VCF. Discordance and non-detection at both sequencing centers was enriched for insertion-deletion variants. Many of these insertion-deletion allele appeared to be concordant and/or present upon manual inspection but had slightly different genomic locations or allele identity based on left or right justified allele report decisions.

Harmonization of calling methods, position and naming conventions for insertion-deletion will facilitate the consensus of future efforts.

5 Genetic to Self-Report Gender Validation

To assess gender we looked at the zygosity of X chromosome variants and for the detection of Y chromosome variants while excluding the pseudoautosomal regions of X and Y.
We called a sample a 'genetic male' if all the X and Y chromosome markers were homozygous. We called a sample a 'genetic female' if heterozygous markers were detected on chromosome X and no markers were detected for the Y chromosome. If a sample was both heterozygous for chromosome X and had Y chromosome markers detected we marked it as genetic gender 'unknown'. We did this for both the multisample VCF and the single sample clinical VCFs produced by the sequencing centers. We additionally required the depth of sequencing to be greater than 50 reads. Among heterozygous variants, we required the variant allele depth to be greater than 15 reads to classify a variant as a heterozygote. Using these genetic gender calls we then compared to the reported gender from the clinical phenotyping. We summarize this here.

231 X chromosome variants and 13 Y chromosome variants were consistently called across the multisample VCF.

5.1 Multisample VCF Genetic Gender Calls

13,404 female
11,452 male
100 unknown

We provide the genetic gender calls, reported gender and chromosome X and Y variants counts and zygosty frequencies in the file:

eMERGEseq_samples.genetic_reported_gender

We only attempted to call concordance on the samples where gender was unambiguous and we could call male or female for both genetic and clinically reported values.
called_concordant_genders = concordant_count / total_called_count
0.99086908238391 = 24091 / 24313

There are 222 samples where the reported gender does not match the genetic gender. They are listed in this file:

multisample_gender_checks/eMERGEseq_samples.genetic_reported_gender_mismatch

Count	Reported Gender	Genetic Gender
111	female	male
111	male	female
356	NA	female
187	NA	male
3	NA	unknown
45	male	unknown
52	female	unknown

Table 2: Multisample VCF Gender Mismatch Category Counts. The top two lines, female-male and male-female were included in the concordance calculation.

5.2 Single Sample VCF Genetic Gender Calls

For the single sample VCFs we found that we had to include additional filters to call the gender correctly due to spurious alignments of variants from the CHEK2 gene on chr22 with a region of ChrY (Y:13483339-13842843) that made females appear to have chrY variants. We classified the samples and have these counts of genetically determined gender:

13,490 female
11,316 male
257 unknown

We provide the single sample genetic gender calls, reported gender and chromosome X and Y variants counts and zygosty frequencies in the file:

single_sample_gender_checks/eMERGEseq_samples.genetic_reported_gender_quality_exclude_CHEK2

called_concordant_genders = concordant_count / total_called_count
0.990428837787446 = 23904 / 24135

The mismatched gender samples are listed in this file:

single_sample_gender_checks/eMERGEseq_samples.genetic_reported_gender_mismatch_quality_exclude_CHEK2

Single Sample VCF Gender Mismatch Category Counts

Count	Reported Gender	Genetic Gender
112	female	male
119	male	female
358	NA	female
186	NA	male
7	female	NA
14	male	NA
2	NA	unknown
222	male	unknown
32	female	unknown

Table 3: Single Sample Clinical VCF Mismatch Category Counts, note the top two categories (female-male and male-female) were included in the gender concordance calculation.

6 Variant annotation

6.1 SeattleSeq

Variant annotation using the SeattleSeq server is provided in the file:

eMERGEseq.annotation_only.vcf

home page:
http://snp.gs.washington.edu/SeattleSeqAnnotation138/

how to use website:
http://snp.gs.washington.edu/SeattleSeqAnnotation138/HelpHowToUse.jsp

descriptions of columns (link to this is at the bottom of the HelpHowToUse.jsp page):
http://snp.gs.washington.edu/SeattleSeqAnnotation138/HelpInputFiles.jsp

7 Withheld Samples

Please remove the following samples from your analysis due to sample naming errors noted by the sequencing center.

52114489
52103075
52104602
52114475
52114465
52114468
27278320

8 SFTP Data Access

8.1 SFTP Data Downloading/Uploading Instructions

8.2 Login and Password Information

A. Please email e3helpme@uw.edu with your site, your name, phone number, and a convenient time to call you, and we will provide a login and password to access the data.

You will be able to download data via a Unix like command line interface and/or via a Windows application.

B. To download from the sftp server, you will need the following:

1. Log-in credentials provided by the University of Washington (UW);
2. If using Windows, install the Windows WinSCP software which can be accessed at the following URL:

https://winscp.net/eng/index.php

Examples of Unix sftp commands:

```
sftp crosslin-aspera1@sftp.gs.washington.edu

ls

cd eMERGEseq

get merged_eMERGEseq_samples.variants.chr-sorted.snps-recal.consented_clean.vcf.gz

mget *
```

```
quit
```