Pfizer/MGH CAMP QC Results for AMP T2D Knowledge Portal Data Submission

MGH Cardiology and Metabolic Patient Cohort (CAMP MGH)

- CAMP MGH Cohort consists of over 3,000 MGH Heart Center subjects with detailed phenotype and genotype information available.

Overview of CAMP data set

- CAMP cohort is the first pharma funded data set to be transferred to the DCC.
- CAMP is also the first data set to be transferred to the DCC for QC processing.
- Received the following data for CAMP:
  - Phenotype data for 3,863 individuals (590 T2D cases, 3273 T2D controls)
  - Genetic data for 3,732 individuals genotyped on Illumina Human Core Exome Chip. 26 samples look to be removed during the QC done by CAMP analysts and the remaining samples were removed from the analysis due to quality concerns.

Steps to bring data to the portal

1. Transfer data to DCC
2. QC of data
3. Association analysis
4. Deposition of data in portal

CAMP Data

1. Data transfer complete

2. QC of CAMP Data – This is an iterative process involving both the data submitters and the data intake team at the DCC. The data set is currently in this step and we have assessed the quality of the incoming data to ensure that it is ready to be used for future analysis.
    - We have also used this data set to develop a robust data QC process at the DCC that can be used for all incoming data sets.

3. Association Integrate the CAMP data set into the AMP T2D Knowledge Portal. This will happen over a number of steps:
    - Association Analysis performed on selected traits at the DCC
    - Association Analysis on all traits at the DCC. Analyses of CAMP data can come in freezes and be done over time as we review the traits.

Details of Data QC Process at DCC

- Check data for duplicates and cryptic relatedness that may indicate contamination or data collection errors
- Ancestry inference of the individuals using principal components derived from common ancestry informative variants in 1000 Genomes Project
- Detecting sample outliers
- Pedigree reconstruction
- Provide a completed QC report to the data submitters

Summary of QC Results

- All samples >95% call rate and all variants >98% call rate
- Identified 7 unique samples as outliers during ancestry inference and may be removed from future analysis.
- 6 samples failed the sex check
- 60 samples identified as outliers along either individual sample metrics or principal components of the metrics
- 98.2% samples proceeding on to association analysis

CAMP Variant Raw Data Summary by Minor Allele Frequency

3,732 samples genotyped on Illumina HumanCore Exome Chip. The table below shows the total number of variants detected.

Variants	*Autosomal*	*X*	*Y*	*X (Par)*	*Mitochondrial*	*Unplaced*	*Total*
Mono	103,117	2,132	1,352	0	39	0	**106,640**
(0, 0.01)	133,269	2,606	194	0	231	2	**136,302**
[0.01, 0.03)	16,496	491	38	0	68	0	**17,093**
[0.03, 0.05)	8,832	379	81	0	18	0	**9,310**
[0.05, 0.10)	19,311	674	34	1	26	0	**20,046**
[0.10, 0.50]	236,249	5,678	279	0	15	0	**242,221**
Total	**517,274**	**11,960**	**1,978**	**1**	**397**	**2**	**531,612**

Initial Data Preparation Steps

1. Removed 2 unplaced, 6,897 duplicate variants, and 106,641 monomorphic variants. Duplicate variants indicate that the same variant was on the same chip twice.

2. Checked for duplicate samples (PI_HAT > 0.9), 0 found

3. Checked for samples exhibiting “unusual” levels of ibd sharing (PI_HAT >= 0.25 with 10+ others. PI_HAT >= 0.25 is an approximate indicator of 2nd degree relatedness), 0 found

4. QC had already been performed, with sample callrate > 0.95 and variant callrate > 0.98

Ancestry Inference

Principal Components

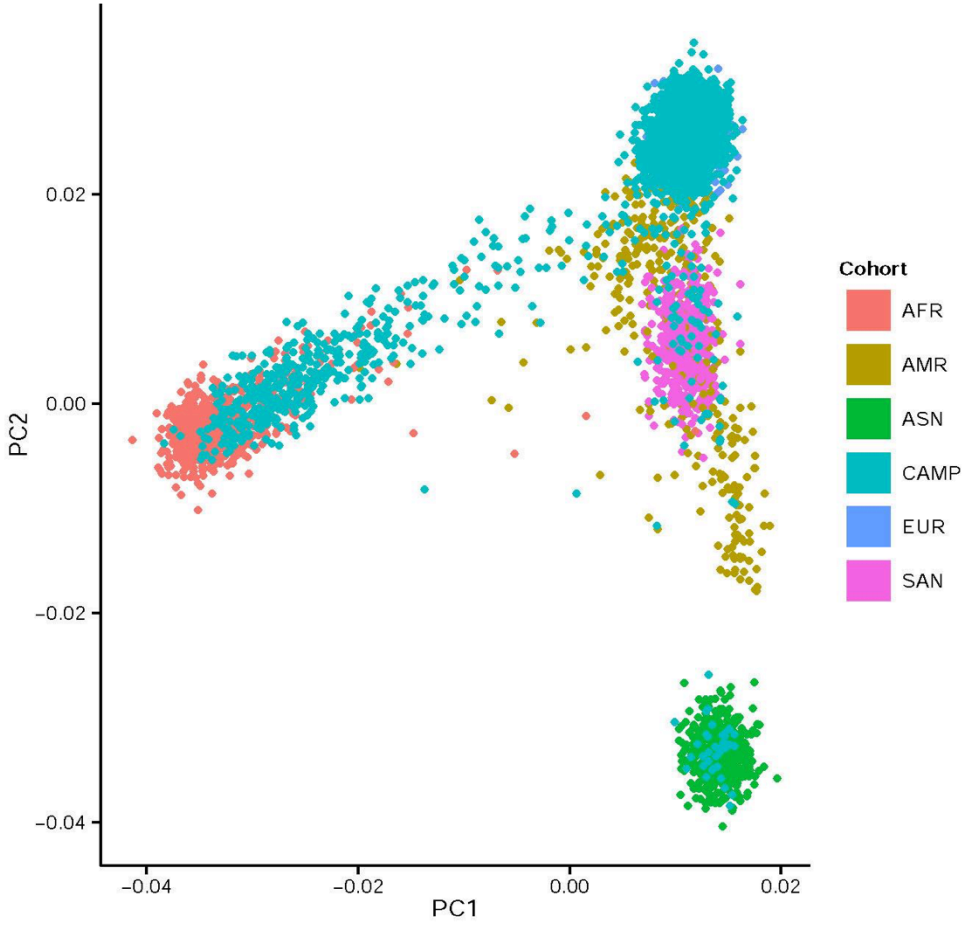
Merged common variants with 1000 genomes 5k ancestry inference sites

Inferred principal components using only 1000 Genomes samples

1KG 5k ancestry inference sites (Phase 3 v5)

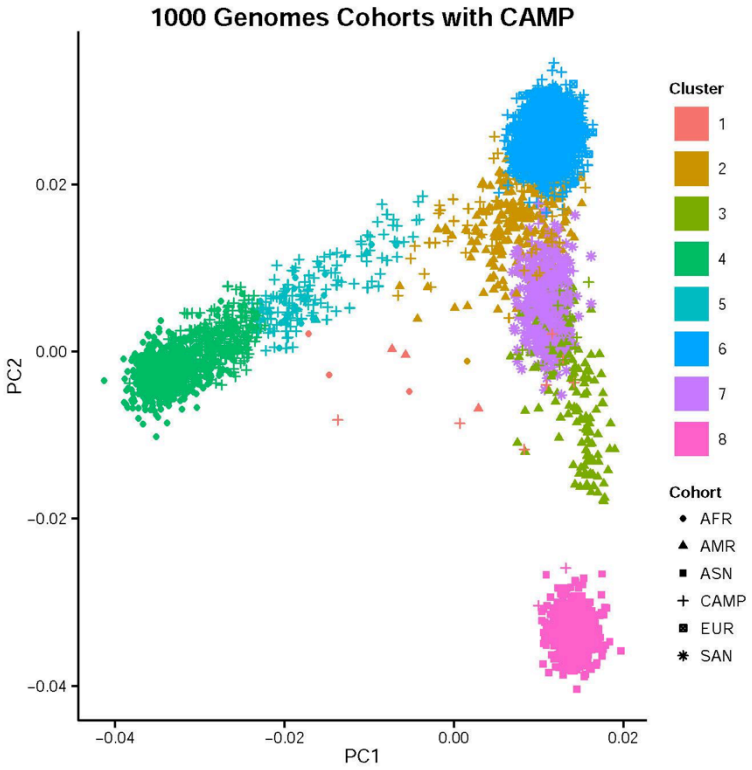
- 661 African (AFR)
- 503 European (EUR)
- 347 Ad Mixed American (AMR)
- 504 East Asian (ASN)
- 489 South Asian (SAN)

Principal Components based on 2,932 overlapping variants



Ancestry Inference after clustering 1000 Genome and CAMP together

Clustering (GMM algorithm (Klustakwik) and 3 PCs)



Cluster	AFR	AMR	ASN	CAMP	EUR	SAN	Assign
1	3	3	0	7	0	0	OUTLIER
2	1	183	0	71	4	0	AMR
3	1	138	0	16	0	0	AMR
4	637	0	0	283	0	0	AFR
5	19	3	0	133	0	0	AFR
6	0	20	0	3149	499	0	EUR
7	0	0	0	36	0	489	SAN
8	0	0	504	37	0	0	ASN

Sample Metrics and Principal Component Analysis (PCA) Approach

Calculate sample metrics

NALT, NMIN, NHET, NVAR, RATE, SING, TITV, DOUB, HET, HET_HIGH, HET_LOW, NHOM, HET_HOM

Calculate PC-adjusted residuals of metrics (PCARM) (NALT ~ PC1+PC2+PC3, etc)

Residuals are expected to be more normal

Adjusting for principal components hopefully eliminates multimodal distributions

Calculate PC's for the PCARM's

Center and scale

Box-Cox transformation to make more normal

Keep axes that capture 95% of the variation

Perform outlier detection for all samples using 2 methods

1. Use GMM on PC's for the PCARM's to capture any outliers that may fall outside on multiple metrics at once
2. Use GMM on PCARM's individually

Sample QC Summary Outlier Removal

Samples	*Initial Sample Count*	*Failed Sexcheck*	*Adj Metric PCA Outlier*	*Adj Metric Outlier*	*Overlap Metric Outliers*	*Total Outliers*	*Final Sample Count*
EUR	3,148	3	9	15	8	18	**3,130**
AFR	416	0	11	22	7	26	**390**
AMR	87	3	5	11	3	13	**74**
ASN	37	0	0	5	0	5	**32**
SAN	36	0	0	0	0	0	**36**
Total	**3,724**	**6**	**25**	**53**	**18**	**60**	**3,664**

Variant QC Summary

Remove variants with call rate < 0.98

Remove variants with Hardy Weinberg P-value < 1e-6 in unrelated samples only

- Calculate maximum unrelated set (PRIMUS)
- Calculate HWE P-value with maximum unrelated set

Variants	*Total*	*Unplaced*	*Duplicate*	*Monomorphic*	*Initial Variant Count*	*Failed Call Rate >= 0.98*	*Failed HWE P-value >= 1e-6*	*Final Variant Count*
EUR	**531,612**	2	6,897	106,640	418,560	0	7	**418,553**
AFR	**531,612**	2	6,897	106,640	418,560	0	20	**418,540**
AMR	**531,612**	2	6,897	106,640	418,560	0	0	**418,560**
ASN	**531,612**	2	6,897	106,640	418,560	0	0	**418,560**
SAN	**531,612**	2	6,897	106,640	418,560	0	0	**418,560**

Relatedness Summary

Calculate identify by descent

Used PRIMUS to reconstruct pedigrees

Family Structure	*Total sample size*	*Number of families*	*Mean family size*	*Largest family size*	*Smallest family size*	*Total Unrelated*
EUR	**3,132**	59	2.20339	5	2	3,061
AFR	**390**	16	2.0625	3	2	373
AMR	**74**	1	2	2	2	73
ASN	**32**	0	-	-	-	32
SAN	**36**	1	2	2	2	35
Total	**3,664**	**77**	**-**	**-**	**-**	**3,574**

Next Steps for CAMP

CAMP data looked great, with 3664/3732 samples passing our quality checks (>98% of the samples submitted passed!)

Our next step is to begin association analysis using a small number of selected traits. After reviewing the phenotype data we recommend proceeding forward with analysis on the following traits:

1. T2D status
2. Fasting glucose
3. Fasting insulin