Successful Illumina Array Genotyping on Serum Stored for Three Decades

Stanley H. Weiss, MD[1]; Daniel M. Rosenblum, PhD[1]; Andrew I. Brooks, PhD[2]; Christian J. Bixby, MS[2]; Charles Y. Hevi[2]; Eric Otto Johnson, PhD[3]

[1]Department of Medicine, Rutgers New Jersey Medical School, Newark NJ, [2]RUCDR Infinite Biologics, Piscataway NJ, Rutgers, The State University of New Jersey;
[3]Behavioral Health & Criminal Justice Division, RTI International, Research Triangle Park, NC.

Department of Medicine Research Day, 26 May 2020

Background

An integrated series of prospective cohort studies of drug users in medication assisted treatment across the US, begun in 1980's, linked with a research biospecimen repository, had >11,000 visits at which blood and sometimes other specimens were obtained.
These studies had all been started by Dr. Weiss as the PI. All subjects were interviewed by trained staff. Written consent included a provision for follow-up activities indefinitely and research testing.
There were several different sub-studies; repetitive enrollments over time among all studies combined were identified.

Summary:

Highly successful Illumina array genotyping is possible using limited serum volume that had been stored for three decades at -80°C.
The genotype and demographic & drug use phenotype data are being deposited by Dr. Weiss into dbGaP, will contribute to ongoing NGC analyses, and will significantly increase the number of samples available for joint analyses among Blacks and Hispanics.

Methods

Serum was collected in SST tubes, centrifuged locally, and shipped insulated overnight to a central lab for aliquoting; samples were then maintained continuously at −80°C.
At some study visits, anticoagulated peripheral blood (heparin, EDTA, or lithium) was Ficoll-Hypaque separated to obtain mononuclear cells, usually with controlled-rate freezing to maintain cell viability, and then continuous storage in liquid nitrogen.
The current study protocol for this prospective study was reviewed and approved by the Rutgers-Newark IRB.
We developed, based upon our review of our specimen inventory, two sets of specimens for initial processing by RUCDR. Original plans had been to run our specimens on the Affymetrix Smokescreen array, but NGC determined in prior work that specimens derived from serum did not yield sufficient DNA on that platform to achieve acceptable results. Since it was important for the same platform to be used on all our specimens, DNA derived from cells was also run on Illumina.
Serum. plasma and cell specimens were processed by RUCDR as per standard NIDA Genetics Consortium (**NGC**) & RUCDR protocols. RUCDR extracted DNA from the entire supplied vial followed by PCR amplification and performed genotyping using the Illumina Infinium OmniExpress-24 V1.3 BeadChip. Data on successfully genotyped specimens was returned to Dr. Weiss by the group at Washington Univ. in St. Louis (**WUStl**) which handles processing & data entry to the database of Genotypes and Phenotypes (**dbGaP**). Linkage of vial codes to subject information was retained exclusively by Dr. Weiss's group.
This initial manifest resulted in 5,033 coded vials of serum as well as 157 vials of lymphocytes being processed. A second manifest, chosen to include smaller volumes & to examine plasma, yielded 445 serum vials, 15 plasma vials, and 11 vials of lymphocytes. (14 [0.25%] serum samples in the combined manifests were not located at the expected inventory location, demonstrating high accuracy of the inventory database.) Serum and plasma volumes ranged from 0.5 to 2.0 ml, mean 1.54 ml, std. dev. 0.32 ml.
Although it had been thought that these lists contained unduplicated subjects, further review of study data in combination with the genomic results identified some subjects with serial specimens. In part, this was a function of strict segregation of personal identifying information and other study data. Our ongoing work of consolidation of multiple sub-studies remains in process. Basic demographic information (gender, race/ethnicity, age) was then returned on the genotyped specimens to WUStL for integration with the genotyping results.

WUStL sent to Dr. Weiss, for each sample, miss rates (which is the fraction of the 714,238 SNPs for which genotype data was missing), identification of gender and race/ethnicity on the basis of the genotype data, identification of putative relationships among these subjects, including apparent identical subjects, as well as first-degree relatives. A comparative review of all data identified instances of multiple enrollments. Still under review are a limited number of gender or race/ethnicity discrepancies. Data were reported back to WUStL.
In accordance with typical GWAS practices for the NGC, those specimens with miss rates ≥ 15% will not be included in dbGaP and NGC analyses.
We also display data utilizing a miss rate cutoff of 20%, which demonstrates the same relationship with volume.

Results

Of the 5,493 serum & plasma specimens, genotyping was successful on 5,485 (99.85%). Sample miss rates ranged from 0.17% to 71.2%, with a mean of 3.0% (median 1.2%, std. dev. 5.2 percentage points). Demographic information from study records showed that 33.0% of the 5,485 serum specimens were from females, 33.3% from non-Hispanic whites, 35.2% from non-Hispanic Blacks, 30.3% from Hispanics, and 1.2% from others.

Genotyping was successful on all 168 cell-derived specimens (100%). The sample miss rates were low, as expected; they ranged from 0.14% to 3.95% (mean 0.38%, median 0.31%, std. dev. 0.42 percentage points). For these 168 specimens, 41% were from females, 55% from whites, 33% from Blacks, 11% from Hispanics & 1% from other. These are not included in the table below

Determinants of sample miss rate in serum & plasma:

Sample miss rates in serum & plasma specimens, as expected, were negatively correlated with vial volume, $r = -0.29$ ($p < 0.0001$).
Grouping vial volumes into seven categories, the percentages of specimens with miss rates ≥ 20% and ≥ 15% decreased with increasing volume (ml):

Miss rate by ml of serum/ plasma	OVERALL	0.5-0.59	0.6-0.99 (all < 0.8)	1.0-1.3	1.35-1.45	1.5-1.55	1.6-1.65	1.7-2.0
# of samples	**5,485**	158	124	571	175	913	1,560	1,984
% missrate > 20%	**1.91 %**	5.06 %	12.10 %	4.73 %	2.86 %	1.53 %	1.03 %	1.01 %
% missrate > 15%	**2.75 %**	16.46 %	27.42 %	7.88 %	4.57 %	2.96 %	1.92 %	1.81 %

It was anticipated by us that HIV infection would be associated with higher miss rates, especially at lower vial volumes, with decreasing miss rates as volumes increase in both HIV-negative and HIV-positive subjects. Our data demonstrate those trends; 13% of specimens were HIV-positive.

% with miss rate > 15% by HIV status at study entry & volume of serum/plasma	OVERALL	0.5-0.59	0.6-0.99 (all < 0.8)	1.0-1.3	1.35-1.45	1.5-1.55	1.6-1.65	1.7-2.0
HIV-neg (4,766)	**3.42 %**	16.38 %	27.12 %	6.30 %	3.95 %	2.89 %	1.83 %	1.75 %
HIV-positive (699)	**6.15 %**	16.67 %	33.33 %	12.34 %	9.09 %	3.85 %	2.96 %	2.29 %

Conclusions:

- If only small volumes of serum or plasma are available, a potential bias in the final dataset could be a selective loss of HIV-positive subjects from the cohort. In contrast to serum or plasma, when starting from ***cells*** as a specimen source, the call rates are high in both HIV-positive and HIV-negative subjects. However, multivariate analyses showed:
- In a multiple logistic regression, specimen volume and Black race were significant predictors of miss rate > 15%, while gender, HIV status, and age were not. Investigations of collinearities among these predictors are ongoing to better delineate the effects of these predictors.
- **Our cohort subjects will be providing an unbiased large additional set of women and minorities to the NGC analyses.**

An earlier version was presented as poster B-20 on 13 January 2020 at the semiannual meeting of the NIDA Genetics and Epigenetics Cross-Cutting Research Team.

No benefit to rerunning samples with high initial miss rate:

In 115 serum vials (**2.3%**) from the first set of 5,033, the initial miss rate exceeded 15%. Those were re-run as per current standard NGC/RUCDR practices. On review of all data, there was no instance where, on the repeat run, the subject could now be included, because the miss rates always remained > 15%.
Scan failures: In 23 serum vials (**0.46%**) from the first set of 5,033, there was a complete scan failure in the initial run. The miss rate was above 15% in all repeated scans. Since this was limited to 23 samples, there is no clear recommendation with respect to whether standard practice might be changed, and this circumstance occurred relatively rarely.
Conclusion: rerunning samples may not be a maximally productive use of resources.
For the remaining ~3000 serum samples from this study that are anticipated to be genotyped, we are considering a change in standard practice: that the same sample not be routinely rerun, and instead that all residual eluent be preserved for possible other future studies.
In order to obtain a valid result on a maximal # of subjects, substitution of a larger serum volume or use of cells as the sample source or a specimen from a different draw date may be considered.

Repeat enrollments identified through genotyping:

Genotyping of the first set of specimens identified sets (91 pairs and 6 triples) of serum samples as having come from repeat enrollments. Each set included at least one specimen with a miss rate < 15%.
When the serum vial volumes differed (73 pairs), as expected there was a trend for the larger vial to have a lower miss rate. (The serial volumes were the same for 18 pairs. There were also two subjects with both a serum and cell specimen included in the manifest.)

Through detailed review of study data, we confirmed that 70 of the genotypically identified serum pairs and 5 of the triples, as well as two of the specimens from the remaining triple, are indeed all from repeat enrollments; phenotypic data will be sent on one specimen from each set for cataloging into dbGaP.
The remaining 22 apparently genotypically identical sets are still under review.

Summary:

From the samples so far processed from the first set of specimens for which the miss rate was < 15%, this cohort so far will be adding to the NGC analyses approximately:

- By gender 1,621 females and 3,302 males;
- By race/ethnicity 1,637 white, 1,777 Black, 1,452 Hispanic, 57 other;
- 602 known HIV-positive and 4,303 known HIV-negative at study entry.

Next Steps (in progress)

The total cumulative count of subjects from our cohorts anticipated to contribute to the NGC data pool will be over 8000 persons, with a similar demographic distribution as described above.
We are planning to use these data in follow-up outcome studies, with ascertainment through matching to vital statistics and cancer registries, to examine risk factors. Our most frequent cancer outcome is lung cancer. Our epidemiological analysis will incorporate our quantitated drug use data, which includes both tobacco and cannabis use, as well as data on many other potential risk factors,

We thank John Rice and Lingwei Sun at WUStL for their assistance with this study. Dr. E.O. Johnson is supported by NIDA Grant 1R01DA044014-01. RUCDR's genotyping is supported by NIDA Contract N01DA187789.