Impact of sample pretesting in a high through-put genotyping facility

B Marosy, C Boehm, B Craig, J Romm, C Ongaco, M Zilka, M Adams-Carr, Y Osimokun, I McMullen, J Zhang, M Jewell, K Hetrick, H Ling, EW Pugh, KF Doheny Center for Inherited Disease Research (CIDR) and Genetic Resources Core Facility (GRCF) SNP Center, IGM, JHUSOM, Baltimore, MD

Introduction:

The Center for Inherited Disease Research (CIDR) was established at the Johns Hopkins University in 1996. CIDR provides high quality genotyping services and statistical genetics consultation to investigators working to discover genes that contribute to common disease. Providing high quality data is a continued priority in the current era of Genome Wide Association studies (GWAs) and publically available data. DNA samples received by CIDR undergo pre-testing evaluation (if funding supports this step) prior to production genotyping. This pre-testing allows for samples and corresponding phenotypic data to be confirmed and includes:

- Confirmation of gender and expected duplicates - gender errors can identify sample swaps as well as plate-wide aliquoting or file annotation errors.
- Detection of unexpected duplicates – large GWAs studies composed of samples acquired from separate collections often include unexpected duplicated individuals
- Detection of any Mendelian inconsistencies – confirms parent-offspring pedigree information for family studies
- Assessment of sample quality – provides an initial screen for samples that may not perform well in the assay.
- Provision of a genotype barcode - allows for biological sample tracking throughout the production lab processing and data release process.

Any problem identified at pre-testing is reported to the investigator. In turn the investigator has the opportunity to replace a sample, drop a sample, or correct sample information. With the high cost associated with large scale genotyping it is imperative to be conscientious of cost. Identification of problematic samples aides in avoiding unnecessarily genotyping these samples and reduces costs. Here we present data on the impact of pre-testing for 12 GWAs projects (41,748 samples) and 16 linkage and custom SNP projects (21,582 samples) that were genotyped utilizing the Illumina® Infinium™ and GoldenGate™ platforms.

Methods:

Sample problems are identified and tracked at different phases of a sample's progress through the genotyping workflow. These phases include:

- Pre-testing – A small panel of 96 SNPs are genotyped on each individual. Data generated is used to identify and report problems related to gender inconsistencies, expected and unexpected duplicates, Mendelian inconsistencies and performance, in addition to providing a genetic barcode for sample tracking.
- Production – Once all samples are approved for production, large scale genotyping is performed on each sample with the array chosen for each project. Problems identified in this phase are primarily related to performance during production genotyping.
- Data Cleaning – Occurs post production genotyping and prior to data release. The genotype data generated at production is used to confirm and identify any problems related to gender inconsistencies, Mendelian inconsistencies and performance that may or may not have been previously identified.

CIDR utilizes a custom Laboratory Information Management System (LIMS) to collect information and data on samples at each phase of the workflow. The data presented here was collected on sample problems identified at pre-testing and prior to data release. Investigator responses to problems reported at the pre-testing phase were categorized based on the outcome for each sample problem and whether replacing a sample or changing sample information fixed the problem. The problem categories assigned are described below:

- Problem Fixed or Sample Dropped – Includes correction by replacing a sample, file correction by the PI or dropping samples upon request by PI or due to poor performance.
- Problem Not Fixed or not able to be evaluated - Includes replacement samples testing with the same problem, no replacement sample sent, or unable to correct problem.
- New problems at project release – Includes new problems identified after pretesting and samples which performed poorly in production.

Results:

Table 1. Summary of Pre-testing outcomes for GWAs and non-GWAs projects

	GWAs Projects			non-GWAs Projects		
Total samples pretested	41,748	Avg	Range per project	21,582	Avg	Range per project
Total samples with problems	1676	4%	0.3%-10.7%	822	3.8%	0.4%-8.6%
% problems fixed or sample dropped	1552	3.7%	1.0%-10.6%	631	2.9%	0.4%-6.1%
% problems not fixed or not able to be evaluated	124	0.3%	0%-0.8%	191	0.9%	0%-4.3%
Total new problems at project release	359	0.9%	0%-2%	467	2.2%	0%-9.3%

- 92.6% and 76.8% of samples with problems were either fixed or dropped for GWAs and non-GWAs projects respectively.
- Variation occurs between projects as different investigators provide a spectrum of sample source and quality for genotyping.

Table 2. Summary of problem types and investigator responses

	GWAs projects		non-GWAs projects	
Total number of samples w/problems	**1676**		**822**	
Poor Performance	**532**	**31.7%**	**557**	**67.8%**
Replaced	184	34.6%	247	44.3%
fixed	116	63.0%	55	22.3%
not fixed	68	37.0%	192	77.7%
Dropped	348	65.4%	140	25.1%
Used	-	-	170	30.5%
resolved	-	-	18	10.6%
not resolved	-	-	152	89.4%
Gender Inconsistency	**341**	**20.3%**	**105**	**12.8%**
Replaced	84	24.6%	22	21.0%
fixed	41	48.8%	13	59.1%
not fixed	43	51.2%	9	40.9%
Dropped	55	16.1%	10	9.5%
PI modified pedigree	188	55.1%	68	64.8%
Used	9	2.6%	3	2.9%
not resolved	9	100.0%	3	100.0%
NA	5	1.5%	2	1.9%
Mendelian Inconsistency	**782**	**46.7%**	**94**	**11.4%**
Replaced	118	15.1%	34	36.2%
fixed	73	61.9%	11	32.4%
not fixed	45	38.1%	23	67.6%
Dropped	192	24.6%	20	21.3%
PI modified pedigree	455	58.2%	38	40.4%
NA	17	2.2%	2	2.1%
Other	**21**	**1.3%**	**66**	**8.0%**
Replaced and Fixed	11	52.4%	31	47.0%
Dropped	10	47.6%	35	53.0%

For GWAs projects:

- Majority of problems (46.7%) found in GWAs projects were related to Mendelian inconsistencies.
- Replacement samples for poor performance in GWAs projects have the highest rate of correction (63%).
- Gender inconsistencies had the least amount of problems (20.3%) for GWAs projects.
- Mendelian inconsistencies overall had the highest rate of correction (92.1%) for GWAs projects.

For non-GWAs projects:

- Majority of problems (67.8%) found in non-GWAs projects were related to poor performance.
- Replacement samples for gender inconsistencies in non-GWAs projects had the highest rate of correction (59.1%)
- Mendelian inconsistencies had the least amount of problems (11.4%) for non-GWAs projects.
- Gender inconsistencies in non-GWAs projects overall have the highest rate of correction (86.7%).

Table 3. Summary of pretesting outcome and investigator responses by DNA source

	GWAs projects				non-GWAs projects			
	Blood	Buccal	Saliva	WGA	Blood	Buccal	Saliva	WGA
Fixed	827	16	28	13	150	-	-	84
Dropped	442	199	7	20	126	7	2	262
Not Fixed	99	0	1	2	153	1	-	33
Unable to be Evaluated	17	2	1	0	2	-	-	2
Poor Performance	315	188	6	23	262	2	1	292
Gender	317	6	14	4	63	-	-	42
Mendelian	733	23	16	8	71	1	1	21
Other	20	1	0	0	35	5	-	26
total # samples w/problems	1385	218	36	35	431	8	2	381
total # of samples	39566	1378	548	256	16969	824	45	3744
% of samples with pretesting problems by source type	3.5%	15.8%	6.6%	13.7%	2.5%	1.0%	4.4%	10.2%
Range per project - % of samples with pretesting problems by source type	0.3 - 8.5%	2.3 - 21.1%	0 - 14.3%	4.2 - 24.6%	0.2 - 8.7%	1.0 - 1.5%	9.1 - 20%	2.2 - 80%

- Blood is the sample source of choice. However, good source type does not always mean good data and is site dependant.

Table 4. Summary of sample performance by project

	GWAs projects		non-GWAs projects	
	Average	Range per project	Average	Range per project
Sample Call Rate	99.77%	99.65% - 99.86%	99.74%	98.82% - 99.94%
Max Sample Call Rate	99.97%	99.96% - 99.98%	99.99%	99.98% - 100%
Min Sample Call Rate	96.14%	89.75% - 98.28%	96.68%	86.62%-99.47%
Reproducibility Rate	99.99%	99.98% - 99.99%	99.98%	99.90% - 100%

Table 5. Summarization of cost savings

	GWAs Projects		non-GWAs projects	
	Average Project	Worst case scenario	Average Project	Worst case scenario
# of samples	3000	7500	1400	4100
Cost of pretesting	$24,000	$60,000	$11,200	$32,800
Cost of production genotyping	$1.35M	$3.375M	$280K	$820K
# of samples flagged w/problems	4%	10.7%	3.8%	8.6%
production genotyping cost for problem samples if not identified	$54,000	$361,125	$10,640	$70,520
Having high quality data	Priceless!			

Discussion:

Pre-testing has proven to be helpful in providing high quality genotyping data by validating sample information, evaluating sample quality and minimizing costs. As large scale studies continue and new technologies emerge, CIDR is committed to continuing stringent quality control of data and development of strategies to lower costs and provide high quality data.

Pre-testing allows for:

- Confirmation of sample file integrity – the sample 'is' what we think it 'is'. Gender matches, few or no family inconsistencies present (in the case of family studies, where family information is available) or no two samples are the same (identifying unexpected duplicates).
- Elimination of samples that are truly poor performers and will most likely not succeed in the large scale production genotyping.
- Generation of a biological barcode to track sample identity throughout the process, ensuring accurate data-to-sample association.
- Realization of cost savings – minimizing the number of problem samples that are genotyped in production = money saved.