

## Abstract

> The loci used in these assays may come from many sources, the most common of which is the repository of SNPs in the NCBI database, dbSNP (<http://www.ncbi.nlm.nih.gov/projects/SNP/>).

> 18,624 assays over 20 months (Feb. 2004-May 2005).

> Project size: 14 projects; 192 to 2,304 samples.

> Custom assay success rate: 75% to 96%.

> Goal: determine factors contributing to assay success.

> Factors tested:

- ✓ Illumina's SNP Score
- ✓ Reported average heterozygosity from dbSNP,
- ✓ Validation code from dbSNP
- ✓ Illumina's Validation Classes.

> Logistic regression was performed to examine the success of SNP loci.

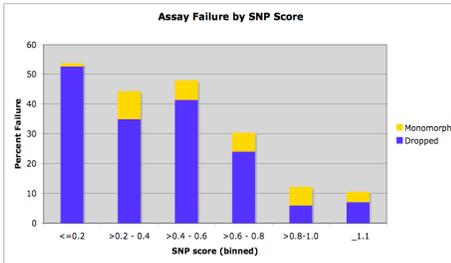
## Assay Failure

We considered two different measurements of assay failure:

- 1) The assay failed one or more quality control measures and was not released to the investigator
- 2) The assay worked, but the SNP was monomorphic (defined here as minor allele frequency <0.5%)

## Failed Assay vs. Binned Design Scores

|            | Odds Ratio | 95% CI      |
|------------|------------|-------------|
| ≤0.2       | Reference  | Reference   |
| >0.2 - 0.4 | 1.46       | 0.87, 2.42  |
| >0.4 - 0.6 | 1.26       | 0.82, 1.93  |
| >0.6 - 0.8 | 2.66       | 1.75, 4.04  |
| >0.8-1.0   | 8.34       | 5.51, 12.63 |
| 1.1        | 9.81       | 6.49, 14.82 |



## Failed Assay vs. Build 125 Reported Heterozygosity (binned)

|                                  | Odds Ratio | 95% CI       |
|----------------------------------|------------|--------------|
| Rare (Het <0.1)                  | Reference  | Reference    |
| Low Heterozygosity (Het 0.1-0.2) | 5.18       | 4.31, 6.23   |
| Common (Het >0.2-0.5)            | 12.41      | 10.84, 14.21 |
| Excess Heterozygosity (Het >0.5) | NA         | NA           |
| Heterozygosity missing           | 3.8        | 3.24, 4.32   |

| Failed Assay vs. Build 125 Binned Validation Codes |            |              | Failed Assay vs. Illumina Validation Codes |            |            |
|--|------------|--------------|--|------------|------------|
|  | Odds Ratio | 95% CI       |  | Odds Ratio | 95% CI     |
| Not Validated                                      | Reference  | Reference    | ILMN not validated                         | Reference  | Reference  |
| Cluster or Frequency                               | 3.23       | 2.52, 4.13   |  |            |            |
| Submitter  | 7.87       | 5.14, 12.06  |  |            |            |
| HapMap   | 4.06       | 3.28, 5.04   |  |            |            |
| HapMap & Submitter                                 | 22.43      | 16.96, 29.68 |  |            |            |
| Customer Identified                                | 12.33      | 9.81, 15.51  | ILMN validated                             | 1.9        | 1.72, 2.10 |
| Doublehit  | 7.06       | 5.75, 8.67   |  |            |            |
| Doublehit & Submitter                              | 23.14      | 16.16, 33.13 |  |            |            |
| HapMap & Doublehit                                 | 10.15      | 8.36, 12.32  |  |            |            |
| HapMap & Doublehit & Submitter                     | 34.14      | 27.56, 42.28 |  |            |            |
|  |            |              | ILMN ILMN validated                        | 3.17       | 2.87, 3.49 |

## SNP Selection

SNP selection criteria varied by project. For some projects, SNPs were selected by individual investigators, for others SNP Center personnel made the selections. In all cases, SNP scores supplied by Illumina were taken into consideration. Validation status from dbSNP and reported heterozygosity from dbSNP and/or HapMap were often also used as criteria.

SNP Scores are supplied by Illumina and range in value from 0 to 1.1. For this analysis, all SNPs were resubmitted to Illumina using the design algorithm as of February 2006.

Validation Codes came from two sources— Illumina provides a validation class score, and a Validation score can be obtained from dbSNP.

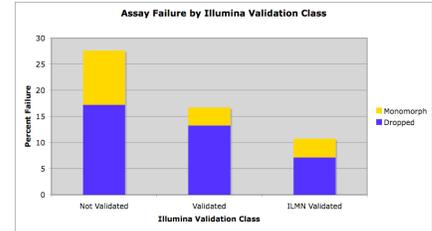
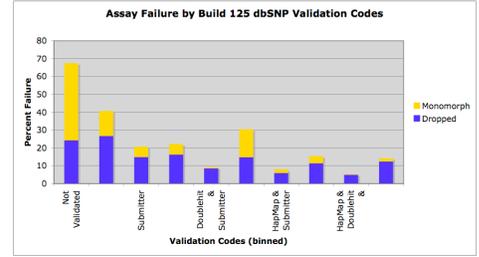
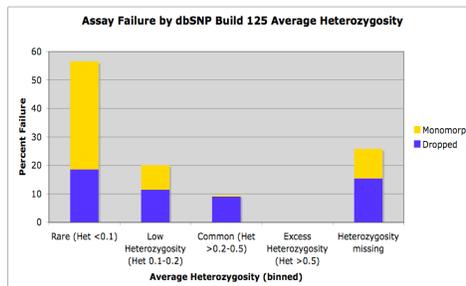
Illumina has three categories of validation: Not Validated, Validated and Illumina Validated. The first two categories are derived from dbSNP's Validation scores, the third indicates that Illumina has designed a successful assay for that SNP in the past. This is usually the same as a design score of 1.1.

The criteria for classification in dbSNP include:

- > Not validated - No validation information, or MAF of 0 in reported population.
- > Frequency - Presence of allele frequency data.
- > Cluster - Submitted by multiple labs.
- > Double-hit - Each allele has been seen independently in two or more samples.
- > Submitter - Submitter validated by two different methods.
- > Hapmap - Genotyped as part of the HapMap project.

A single SNP can fall into more than one category, which improves its validation score. Validation scores were derived from build 125 of dbSNP.

Average Heterozygosity is calculated by dbSNP based on reports received with each SNP submission. It does not take variation in ethnicity into consideration, but does consider the size of the population from which the submitted value was generated. The values used for this analysis were also derived from build 125 of dbSNP.



## Results

We used logistic regression to assay the utility of all factors, considering an assay to have failed if it was not released or if it was monomorphic.

- > Illumina's Validation class does predict success, but dbSNP's grouped individual categories give more information, when only Validation is considered as a predictor.
- > dbSNP categories containing the word "Submitter" have the best success rates.
- > A SNP score of 0.8-1.0 or 1.1 is the best predictor of assay success, when only SNP score is considered.
- > Average Heterozygosity values from dbSNP also predict success when considered alone, with common SNPs being 12.41 times more likely to succeed than rare SNPs.
- > When dbSNP Validation code, SNP score and heterozygosity are considered together in a multiple regression, all three are independent predictors of success. Because a SNP score of 1.1 is nearly perfectly correlated with "Illumina validation" class, a regression of containing both factors is not appropriate, and was not considered.

## Discussion

Based on these results, it is recommended that multiple factors be considered when selecting SNPs for use with this platform.

Our results suggest that where possible individuals selecting SNPs should:

- > Completely avoid SNPs which are not validated from dbSNP (dbSNP category = 0, our bin = 1).
- > Use a threshold of 0.8 for design score.
- > Use Double hit and/or Submitter validation from dbSNP or "Illumina validated" SNPs whenever possible.
- > Avoid SNPs with missing heterozygosity values. Note: Customer Validated SNPs worked well (PGA & Celera).