Designing Custom GoldenGate® Genotyping Assays

Guidelines for efficiently creating and ordering high-quality custom GoldenGate Genotyping Assays using the Illumina Assay Design Tool.

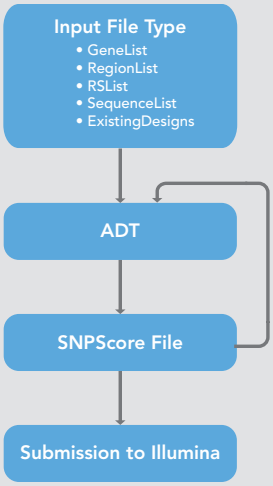
INTRODUCTION

The Illumina GoldenGate Assay is a widely used and robust assay for multiplexed genotyping. Along with fixed content GoldenGate products, Illumina offers researchers the ability to design custom genotyping panels for their organism of study. The GoldenGate Assay can be deployed on the BeadArray™ platform for up to 1,536-plex assays, or on the BeadXpress® platform using VeraCode® technology for up to 384-plex. Regardless of which platform fits their experimental plans better, researchers create their own genotyping panels with the assistance of the Assay Design Tool (ADT) and Illumina scientists.

Researchers create custom panels by selecting and submitting a requested list of loci to Illumina. To ensure fast and successful assay development, this list is evaluated with the Assay Design Tool. Researchers use the ADT SNPScore file output to refine an initial assay panel to include desired assays that are predicted to have a high likelihood of success. The SNPScore file also provides predicted success information, validation status, and minor allele frequencies from published studies. This information can be used to reject some SNPs and ultimately create a final SNPScore file for the an order.

This technical note describes the process of designing, analyzing, and ordering a custom panel of SNPs. Each of the file type options for ADT input and output are described with examples. Template files, similar to those in the examples, can be downloaded from iCom, from www.illumina.com on the downloads page[1] in the support tab, or from Illumina Technical Support[2] via email.

FIGURE 1: CUSTOM GOLDENGATE GENOTYPING ASSAY DESIGN WORKFLOW



PRELIMINARY INPUT FILES

ADT uses a separate file type for each of the four methods for preliminary evaluation of custom SNP loci: GeneList, RegionList, RSList, and SequenceList. Additionally, requests for probe designs from a previously ordered GoldenGate genotyping product use the ExistingDesigns file type. The ADT generates a SNPScore file that can, in turn, be used as an input file in subsequent rounds of evaluation or for ordering (*Figure 1*).

At this time, ADT returns only human sequences from GeneList, RegionList, or RSList input files. Assays for human and non-human genomes are scored using SequenceList or ExistingDesigns file submissions. It is important to note that ADT only supports one build of the human genome at a time. Illumina keeps the supported version of the human genome current and gives users at least two weeks notice before switching to a new version. Technical Support Scientists[2] can confirm which version of the human genome is in use.

Input files may be created or edited with any text editor or spreadsheet program. However, before submitting them to ADT, files must be saved in a comma-separated values (*.csv) format. The examples provided in this document show files created in Microsoft Excel and Notepad. Blank lines are generally not permitted in the data fields or between lines in the heading. The following formatting

TABLE 1: GENELIST FILE COLUMN HEADING DESCRIPTIONS

HEADING	DESCRIPTION
Gene_Name	Customer-supplied gene name. Can be a RefSeq accession ID or HUGO gene symbol.
Bases_Upstream	Number of bases upstream of the gene starting coordinate to search.
Bases_Downstream	Number of bases downstream of the gene ending coordinate to search.
Species	Valid entries are **human**, **man**, or ***Homo sapiens***.

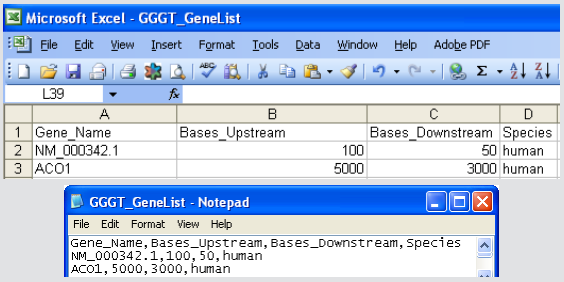
requirements must be followed precisely so ADT can properly evaluate requests.

- Format is comma-separated values with a *.csv file extension. Since the input file format is comma-delimited, no commas may be used within the values.
- File includes specific column headings for the data. As described below, each file type has different requirements for column headings.
- File contains fewer than 65,000 SNPs. If the number of SNPs exceeds this limit, the file must be split into batches of fewer than 65,000 SNPs for serial ADT evaluation.
- If the file is submitted by email rather than on iCom, it must include a file header section. File header format is the same for all file types (*Table 6* and *Figure 7*).

GENELIST

The GeneList file type provides a method for querying all SNPs within a gene and in the regions upstream and downstream from the indicated gene. A GeneList allows for the interrogation of the currently supported build of the human genome using RefSeq NM accession ID (preferred) or HUGO identifiers. ADT maps these accession numbers to the human genome to identify gene regions. The sizes of upstream and downstream regions queried by ADT are specified by the user. SNPs in overlapping gene regions will be listed in the SNPScore output file only once, but will be annotated as being present in both regions in the *Customer_Strand* field. The column headings and description information shown in Table 1 must be provided in the GeneList input file. Figure 2 provides examples of proper GeneList entries in Notepad and Excel. A general guideline is that a GeneList file with up to 100 genes with 10kb upstream and downstream will be below the SNP limit of 65,000.

FIGURE 2: GENELIST FILE FORMAT EXAMPLES

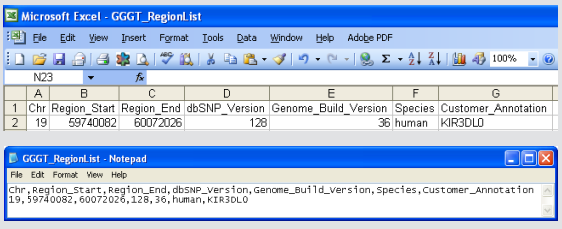


Example of properly formed entries in a GeneList file shown from Excel (top) and Notepad (bottom).

REGIONLIST

The RegionList file type provides a method for selecting SNPs between specified locations of a human chromosome. A RegionList file contains a list of regions in the human genome identified by chromosome and coordinate range that ADT will search and evaluate from among cataloged SNPs in a current Illumina-internal version of dbSNP. This internal database does not contain indels, MNPs, SSRs, or SNPs with ambiguous or multiple localizations. SNPs in overlapping regions will be listed in the SNPScore output file only once, but will be annotated as being present in both regions in the *Customer_Strand* field. Since ADT limits

FIGURE 3: REGIONLIST FILE FORMAT EXAMPLES



Example of properly formed entries in a RegionList file shown from Excel (top) and Notepad (bottom).

TABLE 2: REGIONLIST FILE COLUMN HEADING DESCRIPTIONS

HEADING	DESCRIPTION
Chr	Chromosome on which the SNP is located. Must be an integer, **X**, **XY**, or **Y**. Enter **0** if unknown.
Region_Start	First chromosome coordinate of region to search.
Region_End	Last chromosome coordinate of region to search.
dbSNP_Version	Source version number. Enter **0** if unknown.
Genome_Build_Version	Genome build that will be queried. Contact Technical Support[2] for the currently supported build.
Species	Valid entries are **human**, **man**, or ***Homo sapiens***.
Customer_Annotation	Customer comments. Limited to 30 characters.

TABLE 3: RSLIST FILE COLUMN HEADING DESCRIPTIONS

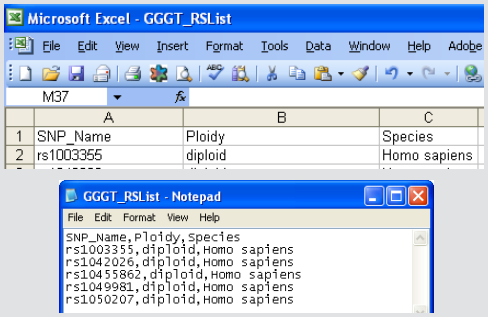
HEADING	DESCRIPTION
SNP_Name	rs number taken from dbSNP.
Ploidy	Since human is currently the only supported species, this entry must be **diploid**.
Species	Valid entries are **human**, **man**, or ***Homo sapiens***.

output to 65,000 SNPs, submitting less than 10 Mb of regions per file is recommended. The column headings and description information shown in Table 2 must be provided in the RegionList input file. Figure 3 provides examples of properly formed RegionList entries.

RSLIST

Known SNPs described in the current version of dbSNP can be requested specifically using the RSList file type. A current internal version of dbSNP is the source for rs SNP and flanking sequence data. The column headings and description information shown in Table 3 must be provided in the RSList input file. Figure 4 shows examples of properly formed RSList entries.

FIGURE 4: RSLIST FILE FORMAT EXAMPLES



Example of properly formed entries in a RSList file shown from Excel (top) and Notepad (bottom).

SEQUENCELIST

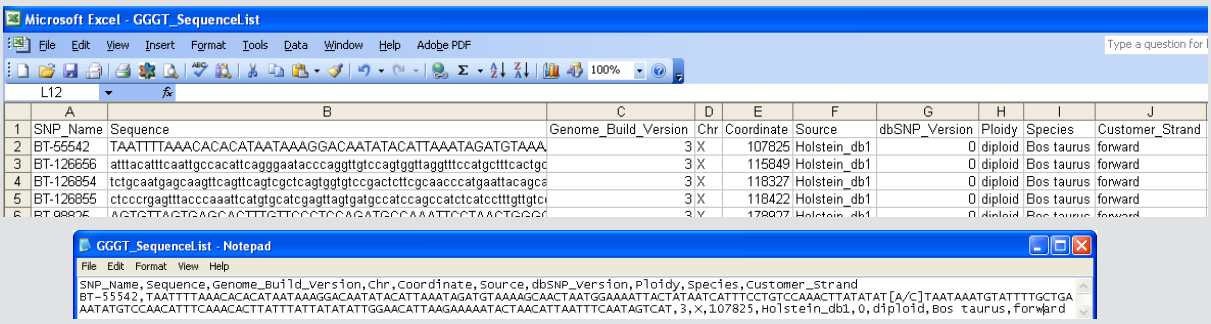
The SequenceList file format provides a method for evaluating SNPs from private databases or other sources, as well as from non-human species. The *SNP_Name* field is used to name sequences for easy identification. *SNP_Name* entries contained in this file must not begin with "rs" because that prefix designates rs ID names in the Illumina database and will trigger a database search.

To specify a SNP, put brackets around a polymorphic locus in the submitted sequence. Separate the two alleles with a forward slash (e.g., …TGC[A/C]CCG…). A minimum of 50bp of sequence on either side of the SNP is required, but 60bp flanking sequence is preferred. ADT will also accept IUPAC codes for degenerate bases in the flanking sequence and take these into consideration during design. ADT uses the *Lowercase_Weighting* checkbox on the iCom submission form (or file header value) to indicate whether lower-case nucleotides are considered for oligo design (unselected) or if lower-case nucleotides are masked (selected). In either case, an Illumina algorithm will identify repetitive or duplicated regions in the unmasked sequence. Since using lower-case nucleotides in public databases is

TABLE 4: SEQUENCELIST FILE COLUMN HEADING DESCRIPTIONS

HEADING	DESCRIPTION
SNP_Name	Customer-supplied name. Cannot begin with *rs*. Can only include letters, numbers, periods, and dashes.
Sequence	Limited to 10,000 bases. May only contain 1 bracketed SNP. Output will be ≤ 122 bases per line.
Genome_Build_Version	Customer-supplied version number. Otherwise, enter **0**.
Chr	Chromosome on which the SNP is located. Must be a valid chromosome for the species being analyzed. Enter 0 if unknown.
Coordinate	Chromosome coordinate of SNP. Enter **0** if unknown.
Source	The source of the sequence and annotation data. Must be completed. Enter **unknown** if no information is available.
dbSNP_Version	Source version number. Enter **0** if unknown.
Ploidy	Only **diploid** or **haploid** species are currently accepted.
Species	Contact Technical Support[2] for a list of currently supported species.
Customer_Strand	Must contain one of the following three values: **forward**, **reverse**, or **unknown**. Information is customer-supplied and is not validated.

FIGURE 5: SEQUENCELIST FILE FORMAT EXAMPLES



Example of properly formed entries in a SequenceList file shown from Excel (top) and Notepad (bottom).

not standardized to indicate masking, we recommend leaving *Lowercase_Weighting* unselected by default.

The column headings and description information shown in Table 4 must be provided in the SequenceList input file. Figure 5 provides examples of properly formed SequenceList entries.

USING PREVIOUS ASSAY DESIGNS

Illumina has created a separate method for conveniently ordering the exact same assays that were designed and used on a previous GoldenGate Genotyping product. As described in Table 5 and shown in Figure 6, an ExistingDesigns file only contains a list of the Ilmn_Id (from the original design manifest) and species for the requested SNP assays.

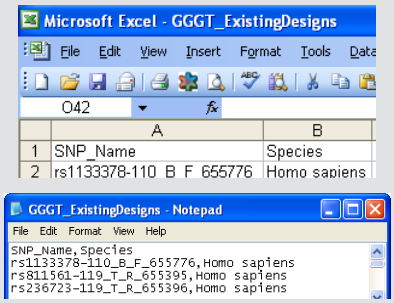
WEB SUBMISSION VIA ICOM

Preliminary input files can be submitted to ADT for evaluation either directly via a web-based interface, or by emailing the file to a Technical Support Scientist[2] who will submit the file to ADT. Submitting via the web is preferred since it provides rapid turnaround and 24-hour access. To submit a preliminary design file, log in to https://iCom.illumina.com and select *Prelim assay design tool (ADT)*. The ADT interface allows users to enter necessary file type information and attach a *.csv formatted

TABLE 5: COLUMN HEADINGS FOR EXISTINGDESIGNS FILE

HEADING	DESCRIPTION
SNP_Name	Ilmn_Id from original design manifest.
Species	Contact Technical Support[2] for a list of currently supported species.

FIGURE 6: EXISTINGDESIGNS FILE FORMAT EXAMPLES



Example of properly formed entries in a ExistingDesigns file shown from Excel (top) and Notepad (bottom).

input file. After the file has been scored, an email notification is sent to the user.

EMAIL SUBMISSION PROCESS

Preliminary ADT input files submitted by email must include an additional file header section (labeled [heading]) before the data entry section (labeled [data]). Headings and descriptions for the file header are listed in Table 6 with examples shown in Figure 7. The format of the file header is common to all preliminary input file types. The entire *.csv file should then be emailed to techsupport@illumina.com for SNP evaluation.

SNP SCORE OUTPUT FILE

If preliminary input files are submitted to ADT via iCom, an email notification is sent when scoring is complete. The results are returned as a SNPScore file that can be downloaded from iCom on the Prelim assay design files page. If an input file is submitted via email to Technical Support[2], an Illumina scientist will submit the file to ADT for processing. ADT generates the SNPScore output file, which is returned to the customer by email or secure FTP in 1–2 days.

The SNPScore file can be used to create a final order file or as an input file format for subsequent ADT submission. SNPScore files provide an important set of informative metrics for each scored SNP requested in the preliminary input file. These metrics should be used to preferentially select the assays that are most likely to be successfully designed for the final product. The SNPScore file header section includes additional summary information, such as the total number of SNPs in the file. A custom product using the GoldenGate Genotyping Assay on BeadArray technology requires 96 or 384–1,536 (in multiples of 96) attempted SNPs. The GoldenGate Genotyping Assay on VeraCode technology can accommodate up to 384 attempted SNPs (in fixed increments).

Following the SNPScore file header section, detailed information for each SNP is listed in the data section. All SNPScore file data section column headers are described in Table 7. Important performance values are also presented for each SNP. The *SNP_Score* indicates the expected success for designing a given assay, and may be supported with *Failure_Codes* for further information (*Table 8*). Validation status is also indicated to provide even greater confidence in design success (*Table 9*). To help researchers order the most applicable SNPs for their studies, minor allele frequencies (MAFs) in several populations are provided for SNPs when available from dbSNP. MAF from the largest study is reported, and is qualified based on peer-reviewed publication, study design and size, and verified results.

FILTERING AND SELECTING CUSTOM SNP LIST

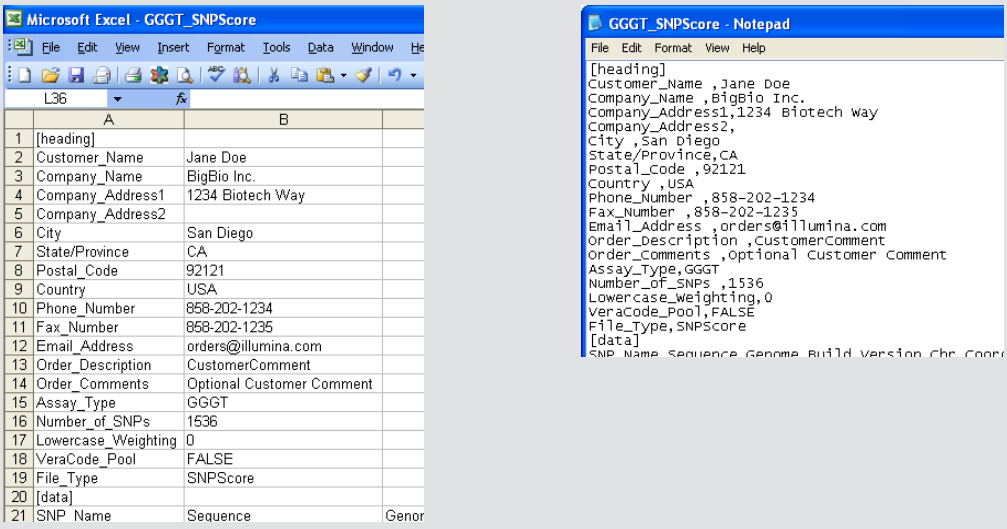
In addition to their use as an output file format, SNPScore files can be used as input files to ADT. Thus, users can easily create a filtered or edited output file (with SNPs removed or added) for iterative ADT analysis while determining the optimal set of loci to order. SNPs identified using more than one input search method (e.g., GeneList, RegionList, RSList, SequenceList, or ExistingDesigns) can be combined as one SNPScore file and resubmitted to ADT as an input file for evaluation as a single product.

Illumina recommends applying the following criteria for discriminating SNP lists to create a successful product that achieves the scientific aims of the experiment and has the highest chances of generating meaningful results.

TABLE 6: FILE HEADINGS FOR ADT INPUT FILE (ONLY REQUIRED FOR EMAIL SUBMISSION)

HEADING	DESCRIPTION	REQUIRED
Customer_Name	Name of person submitting the ADT file	Yes
Company_Name	Company name (no commas)	Yes
Company_Address1	Line 1 of customer's address	Yes
Company_Address2	Line 2 of customer's address (optional)	No
City	Customer's city	Yes
State/Province	Customer's state or province	Yes
Postal_Code	Customer's postal code	Yes
Country	Customer's country	Yes
Phone_Number	Customer's phone number	Yes
Fax_Number	Customer's fax number	Yes
Email_Address	Customer's email address	Yes
Order_Description	Description of work	Yes
Order_Comments	Additional comments (optional)	No
Assay_Type	**GGGT**	Yes
Number_of_SNPs	Number of SNP loci in file (may be **0** for GeneList, RegionList, RSList, and SequenceList file if the number of SNP loci is unknown)	Yes
Lowercase_Weighting	**0** for no masking or **1** to mask lower-case characters	No
VeraCode_Pool	**True** or **False**	No
File_Type	**GeneList**, **RegionList**, **SequenceList**, **RSList**, **ExistingDesigns**, or **SNPScore**	Yes

FIGURE 7: EXAMPLES OF FILE HEADER SECTION ONLY FOR EMAIL SUBMISSION



Examples of properly formed entries in the header of a SNPScore file shown from Excel (left) and Notepad (right). This header is only required for email submissions, and is formatted the same for any preliminary file format.

TABLE 7: SNPSCORE FILE COLUMN DESCRIPTIONS

HEADING	DESCRIPTION
SNP_Name	rs number or customer's unique name. Must be copied from Ilmn_Id value if ExistingDesigns file was used for preliminary scoring.
Sequence	The bracketed SNP site identified by the SNP_Name with > 50 bases of flanking sequence.
Genome_Build_Version	Genome build that will be queried. Contact Technical Support[2] for the currently supported build.
Chr	Chromosome on which the SNP is located. Must be a valid chromosome for the species being analyzed. Enter **0** if unknown.
Coordinate	Chromosome coordinate of SNP. Enter **0** if unknown.
Source	Identify the source of the sequence and annotation data. Enter **unknown** if no information is available.
dbSNP_Version	Source version number or **0** if unknown.
Ploidy	**Diploid** or **haploid**.
Species	Contact Technical Support[2] for a list of currently supported species.
Customer_Strand	Must contain one of the following three values: **forward**, **reverse**, or **unknown**. Information is customer-supplied and is not validated.
Customer_Annotation	Customer comments. Limited to 30 characters.
SNP_Score	Ranges from 0–1.1, with higher values reflecting greater ability to design a successful assay.
Designability_Rank	Simplified representation of SNPScore for easily sorting and filtering results. • If SNPScore < 0.4, rank = 0 • If SNPScore 0.4–0.6, rank = 0.5 • If SNPScore ≥ 0.6, rank = 1
Failure_Codes	If applicable, reasons why a successful assay at this SNP locus is unlikely. For a complete list of failure codes, see Table 8.
Validation_Class	Numerical representation of Validation_Bin (see Table 9).
Validation_Bin	Manner in which designed assays have been validated (see Table 9).
MAF_Default_SNPDome	Internal use only.
MAF_Caucasian ChrCount_Caucasian Study_Caucasian MAF_African ChrCount_African Study_African MAF_African_American ChrCount_African_American Study_African_American MAF_Japanese ChrCount_Japanese Study_Japanese MAF_Chinese ChrCount_Chinese Study_Chinese MAF_Other ChrCount_Other Study_Other	Minor allele frequency from the largest peer-reviewed study conducted, the study size in terms of number of chromosomes, and the study type. Data are retrieved from dbSNP for each population: • Caucasian • Yoruban • African-American • Han Chinese • Japanese • Unknown

TABLE 7: SNPSCORE FILE COLUMN DESCRIPTIONS CONTINUED

HEADING	DESCRIPTION
App_Version	Version of ADT used for scoring loci.
ILMN_ID	Unique identifier assigned by ADT for the designed assay.
Gene_ID*	Gene ID number from NCBI.
Gene_symbol	HUGO identifier.
Accession*	RefSeq Accession number.
Location*	Structural location of the SNP: intron, coding, flanking_5UTR, flanking_3UTR, 5UTR, 3UTR, UTR.
Location_relative_to_gene*	If the SNP does not fall within an exon, the value is the actual base pair distance from gene start. The absolute value of this number is the distance to the closest transcript. The negative sign is a formatting symbol and is not meant to imply strand or direction. If the SNP is within an exon, the two values separated by a '/' represent distances to the exon-intron boundaries.
Coding_status*	NONSYN or SYNON. If the SNP falls within an exon, this field notes a synonymous or non-synonymous amino acid change.
Amino_acid_change*	If the SNP falls within an exon, this field notes the actual change to the amino acid, followed by the GenBank protein sequence used in numbering the change.
Id_with_mouse*	Ratio of identical bases within 60 bp of flanking sequence compared to mouse sequence that have been aligned with the homologous human sequence and cover the SNP in question.
Phast_conservation*	Metric used by the UCSC Genome Browser to identify highly conserved SNPs among species.

*Additional gene annotation only in SNPScore output file from submitted GeneList, RSList, and RegionList files.

1) Remove SNPs that cannot be ordered (error codes in the 101–199 range).
2) Consider research requirements (e.g., tags, spacing, or MAF).
3) When appropriate, favor GoldenGate validated SNPs, since they have the highest chance of converting into functional assays.
4) Use two-hit or HapMap validated SNPs with a preference for SNPs with higher SNP_Scores.

FINAL SNPSCORE FILE

After ADT analysis and custom selection of SNPs that meet the research criteria, a final SNPScore file must be created to place an order. A preliminary SNPScore file is converted to a final SNPScore file by the completion of four header rows (white in Table 10): *Design_Iteration*, *Scale(Number_of_Tubes)*, *Purchase_Order_Number*, and *Product_Name*. It is important to ensure that the *Number of SNPs* value in the final file matches the number on the corresponding quotation or contract. If an ExistingDesigns file was used for ADT input, then all of the *Ilmn_Id* column values must be copied to the *SNP_name* column to create a final order file.

SUMMARY

Custom GoldenGate Genotyping products by Illumina allow researchers to create assays tailored directly to their specific needs for targeted region genotyping or fine-mapping of candidate disease association regions. The GoldenGate Assay may be deployed on BeadArray or BeadXpress platforms. The BeadArray platform supports the highest multiplex levels, and the BeadXpress platform is an ideal option for high-throughput biomarker screening. The ADT provides a simple and powerful method for evaluating individual loci and creating the most successful custom genotyping assays. By following the guidelines in this technical note, researchers can ensure that their orders are designed and placed quickly and easily.

TABLE 8: LIST OF FAILURE CODES FOR THE ADT

CRITICAL FAILURES (UNDESIGNABLE)	
101	Flanking sequence is too short.
102	SNP or sequence formatting error. SNP must match the format [A/B]. Possible causes: • Spaces in submitted sequence • More than one set of brackets in sequence • Missing brackets around SNP • SNP alleles not separated by "/"
103	TOP/BOT strand cannot be determined due to low sequence complexity.
104	SNP is not appropriate for Illumina platform. Possible causes: • Tri- or quad-allelic SNP • Insertion or deletion polymorphism • SNP contains characters other than A, G, C, and T
105	SNP is located in the mitochondrial genome. Not recommended for GoldenGate OPAs due to high copy number of target mitochondrial DNA.
106	Degenerate nucleotide(s) are in assay design region (e.g., W, R, S, N)
WARNINGS (DESIGNABLE)	
301	SNP is in duplicated/repetitive region.
302	T_m is outside assay limits.
340	Another SNP in the list is closer than 61 nucleotides away.
399	Multiple contributing issues.

TABLE 9: VALIDATION STATUS DESCRIPTIONS

VALIDATION_BIN	VALIDATION_CLASS	DESCRIPTION
GoldenGate validated	3	SNP has been previously designed and has successfully generated polymorphic results on the Illumina platform. Designed oligonucleotides have 100% sequence match to those previously designed.
Two-hit or HapMap validated	2	Both alleles of the SNP have been seen in two independent methods and populations, or have been validated by the HapMap Project.
Non-validated	1	SNP has been seen in only one method or population. Even if it has a high design score, there is an increased chance that it is monomorphic.
Unknown	0	SNP is not known within Illumina's database based on SNP name.

TABLE 10: HEADER SECTION FOR FINAL ORDER FILE

HEADING	DESCRIPTION	REQUIRED
Customer_Name	Name of person submitting the ADT File	Yes
Company_Name	Company name (no commas)	Yes
Company_Address1	Line 1 of customer's address	Yes
Company_Address2	Line 2 of customer's address (optional)	No
City	Customer's city	Yes
State/Province	Customer's state or province	Yes
Postal_Code	Customer's postal code	Yes
Country	Customer's country	Yes
Phone_Number	Customer's phone number	Yes
Fax_Number	Customer's fax number	Yes
Email_Address	Customer's email address	Yes
Order_Description	Description of work	Yes
Order_Comments	Additional comments (optional)	No
Assay_Type	**GGGT**	Yes
Number_of_SNPs	Number of SNP loci in file	Yes
Design_Iteration	**Final**	**Yes – for final file**
Scale(Number_of_Tubes)	**Must be 5 or greater**	**Yes – for final file**
Purchase_Order_Number	**Customer purchase order number**	**Yes – for final file**
File_Type	**SNPScore**	Yes

REFERENCES

(1) http://www.illumina.com/pagesnrn.ilmn?ID=75
(2) To contact Technical Support, email techsupport@illumina.com or call 1.800.809.4566.

ADDITIONAL INFORMATION

Visit our website or contact us to learn more about GoldenGate Custom Genotyping products from Illumina.

Illumina, Inc.
Customer Solutions
9885 Towne Centre Drive
San Diego, CA 92121-1975
1.800.809.4566 (toll free)
1.858.202.4566 (outside the U.S.)
techsupport@illumina.com
www.illumina.com

FOR RESEARCH USE ONLY

Pub. No. 370-2007-020 Current as of 5 August 2008