illumina®

Designing Custom GoldenGate® Genotyping Assays

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

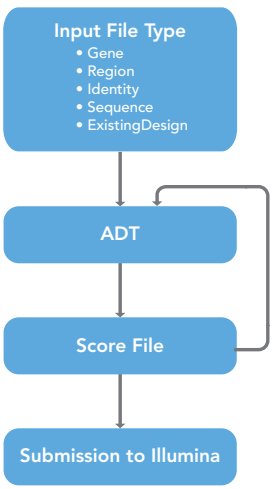
Introduction

The Illumina GoldenGate Assay offers researchers the ability to design custom panels for low- to mid-multiplex genotyping studies. The assay can be deployed on the BeadArray™ platform for 96-plex, or 384- to 1,536-plex assays (in multiples of 96) or on the BeadXpress® platform using VeraCode® technology for 48, 96, 144, 192, and 384 plex. For either platform, the Assay Design Tool (ADT) provides an easy and convenient method for researchers to create custom genotyping panels for their loci and organism of interest.

The process is initiated by selecting and submitting a list of requested loci to Illumina. Upon submission, Illumina will evaluate the list with ADT to ensure successful assay development. Metrics returned by ADT provide success prediction information, validation status, and minor allele frequencies from published studies. Researchers should use these metrics to select an initial assay panel that includes designs predicted to have a high likelihood of success for genetic analysis experiments.

This technical note explains how to desgin, analyze, and order a custom panel of designs. Each of the file type options for ADT input and output are described with examples. Template files can be downloaded from iCom, from the internet[1] (iCom.illumina.com), or by contacting Illumina Technical Support[2].

Figure 1: Custom Goldengate Genotyping Assay Design Workflow



Preliminary Input Files

ADT uses a separate file type for each of the four methods of evaluating preliminary designs: Gene, Region, Identity, and Sequence. Requests for probe designs from a previously ordered GoldenGate genotyping product use an ExistingDesign file. After preliminary evaluation with ADT, Illumina will return a Score file that can 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 Gene, Region, or Identity input files. Assays for human and non-human genomes are scored using Sequence or ExistingDesign 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.

Researchers interested in multi-sample custom panels for GoldenGate Indexing™ should contact Illumina Technical Support for template files; the standard templates should not be used for these requests. Up-to-date GoldenGate Indexing templates will be supplied upon request.

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 in this document show files created in Microsoft Excel. Blank lines are generally not permitted in the data fields or between lines in the heading. These following formatting 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.**
- **Each file type includes specific column headings for the data, as described below.**
- **File contains fewer than 250,000 designs. If the number exceeds this limit, the file must be split into batches of fewer than 250,000 designs 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).**

Gene File

The Gene file type provides a method for querying all loci within a gene and in the regions upstream and downstream from the indicated gene. A Gene file enables 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

Table 1: Gene 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.
Bases_Downstream	Number of bases downstream of the gene ending coordinate.

to identify gene regions. The sizes of upstream and downstream regions queried by ADT are specified by the user. Loci in overlapping gene regions will be listed in the Score output file only once, but will be annotated as being present in both regions in the Region_Description field. The column headings and description information shown in Table 1 must be provided in the Gene input file. Figure 2 provides examples of proper Gene entries in Excel.

Region File

The Region file type provides a method for selecting loci between specified locations of a human chromosome. A Region file contains a list of regions in the human genome identified by chromosome and coordinate range which ADT will search and evaluate from among cataloged markers in the current Illumina-internal version of dbSNP. This internal database does not contain MNPs, SSRs, or SNPs with ambiguous or multiple localizations. SNPs with a source molecule type of cDNA in dbSNP are also not included, as these may result in primers being inadvertently designed across intron-exon boundaries, resulting in a non-functional assay design. Markers in overlapping regions will be listed in the Score output file only once, but will be annotated as being present in both regions in the Region_Description field. The column headings and description information shown in Table 2 must be provided in the Region input file. Figure 3 provides examples of properly formed Region entries.

Identity File

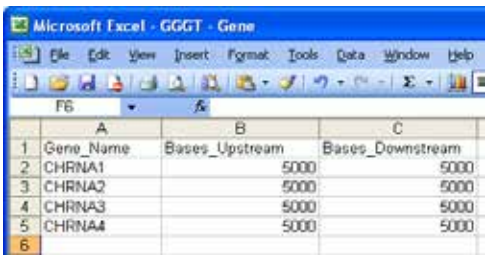
Known loci described in the current version of dbSNP can be requested specifically using the Identity file type. A current internal version of dbSNP is the source for rs loci and flanking sequence data. The column headings and description information shown in Table 3 must be provided in the Identity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

Sequence File

The Sequence file format provides a method for evaluating loci from private databases or other sources, as well as from non-human species. The Locus_Name field is used to name sequences for easy identification. Locus_Name entries contained in this file must not begin with “rs” because that prefix designates rs ID names in the Illumina database.

To specify a locus, put brackets around a polymorphic locus in the submitted sequence. Separate the two alleles with a forward slash (TGC[A/C]CCG). Similarly, to specify an indel, use a slash to separate a single minus sign indicating the deletion from the bases representing the insertion (TGC[-/AT]CCG). A minimum of 50 bp of sequence on either side of the variant is required; however, 60 bp flanking sequence is preferred. ADT will also accept IUPAC codes for degenerate bases

Figure 2: Gene File Format Examples

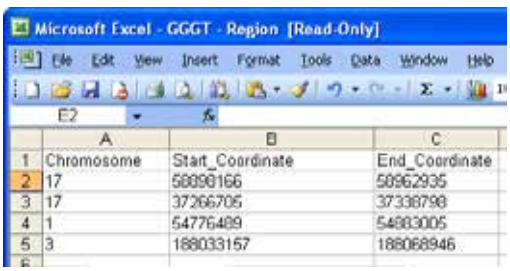


Example of properly formed entries in a Gene file shown from Excel.

in the flanking sequence and take these into consideration during design. If the Lowercase_Weighting checkbox on the iCom submission form (or file header value) is unchecked, lowercase nucleotides will be considered for oligo design; if it is checked, then lowercase nucleotides are masked. In either case, an Illumina algorithm will identify repetitive or duplicated regions in the unmasked sequence. Since lowercasing in public databases is not a standard way to indicate masking, we recommend clearing the Lowercase_Weighting checkbox by default.

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

Figure 3: Region File Format Examples



Example of properly formed entries in a Region file shown from Excel.

Table 2: Region File Column Heading Descriptions

Heading	Description
Chromosome	Chromosome containing the locus. Must be an integer, **X**, **XY**, or **Y**. Enter **0** if unknown.
Start_Coordinate	First chromosome coordinate of region to search.
End_Coordinate	Last chromosome coordinate of region to search.
User_Information*	Customer comments. Limited to 30 characters.

*** Any name can be added to a column except for reserved names (e.g. Ilmn_ID)**

Table 3: Identity File Column Heading Descriptions

Heading	Description
Locus_Name	RS number taken from dbSNP. Value must begin with "rs" (case insensitive) followed by an integer.

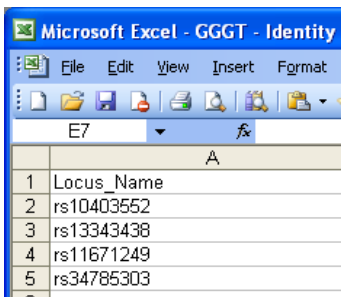
Using Previous Assay Designs

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 ExistingDesign file contains a list of the Ilmn_Ids from the original design manifest. Researchers planning to use this feature should contact Technical Support for additional instructions, especially when combining ExistingDesign assays with newly designed assays in an OPA.

Submitting Via iCOM

Preliminary input files can be submitted to ADT for evaluation either directly via iCom, or by emailing the file to a Technical Support Scientist[2] who will submit the file to ADT. Submitting via iCom is preferred since it provides rapid turnaround and 24-hour access. To submit a prelimi-

Figure 4: Identity File Format Examples

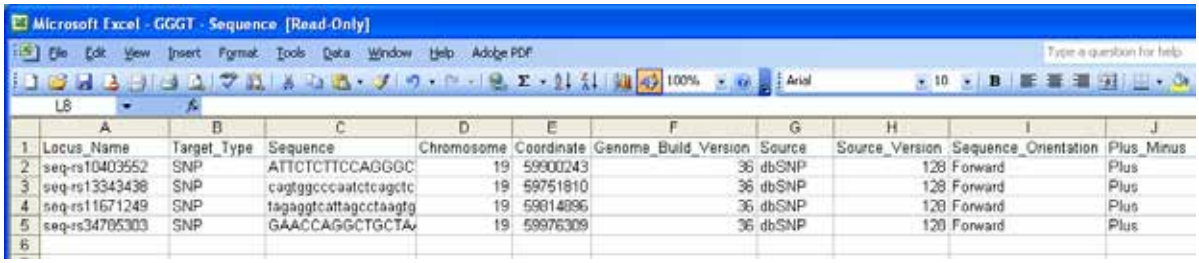


Example of properly formed entries in a Gene file shown from Excel.

Table 4: Sequence File Column Heading Descriptions

Heading	Description
Locus_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 one bracketed locus. Output will be ≤ 122 bases per line.
Target_Type	Must be either "SNP" or "Indel" (case insensitive).
Genome_Build_Version	Customer-supplied version number. If unknown, enter **0**.
Chromosome	Chromosome containing the locus. Must be a valid chromosome for the species being analyzed. Enter **0** if unknown.
Coordinate	Chromosome coordinate. Enter **0** if unknown.
Source	The source of the sequence and annotation data. Must be completed. Enter **unknown** if no information is available.
Source_Version	Source version number. Enter **0** if unknown.
Sequence_Orientation	Must contain one of the following three values: **forward**, **reverse**, or **unknown**. Information is customer-supplied and is not validated.
Plus_Minus	Must be either **Plus** or **Minus** (case insensitive).

Figure 5: Sequence File Format Examples



Microsoft Excel - GGGT - Sequence [Read-Only]

	A	B	C	D	E	F	G	H	I	J
1	Locus_Name	Target_Type	Sequence	Chromosome	Coordinate	Genome_Build_Version	Source	Source_Version	Sequence_Orientation	Plus_Minus
2	seq-rs10403552	SNP	ATTCTCTTCCAGGGC	19	59900243	36	dbSNP	128	Forward	Plus
3	seq-rs13343438	SNP	cagtggcccaatctcagctc	19	59751810	36	dbSNP	128	Forward	Plus
4	seq-rs11671249	SNP	tagaggtcattagcctaagtg	19	59814896	36	dbSNP	128	Forward	Plus
5	seq-rs34765303	SNP	GAACCAGGCTGCTA	19	59976309	36	dbSNP	128	Forward	Plus
6										

Example of properly formed entries in a Sequence file shown from Excel.

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

Submitting Via Email

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 evaluation.

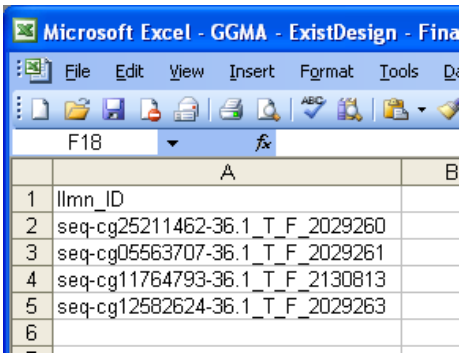
Score Output File

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 Score file that can be downloaded from iCom on the Prelim assay design files page. If an input file is submitted via email to Technical Support2, an Illumina scientist will submit the file to ADT for processing. ADT generates the Score output file, which is returned to the customer by email or secure FTP within two business days.

The Score file can be used to create a final order file or as an input file format for subsequent ADT submission. Score files provide an important set of informative metrics for each scored locus requested in the preliminary input file. These metrics should be used to preferentially select the assays that have a high likelihood for success in the final product design. The Score file header section will include the total number of designs 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 designs. The GoldenGate Genotyping Assay on VeraCode technology can accommodate up to 384 attempted SNPs in fixed increments.

Following the Score file header section, detailed information for each marker is listed in the data section. All Score file data section column headers are described in Table 7. Important performance values are also presented for each locus. The Final_Score indicates the expected success for designing a given assay, and may be supported with

Figure 6: Existing Design File Format Examples



Microsoft Excel - GGMA - ExistDesign - Fina

	A	B
1	Illmn_ID	
2	seq-cg25211462-36.1_T_F_2029260	
3	seq-cg05563707-36.1_T_F_2029261	
4	seq-cg11764793-36.1_T_F_2130813	
5	seq-cg12582624-36.1_T_F_2029263	
6		

Example of properly formed entries in a ExistingDesign file shown from Excel.

Failure_Codes for further information (Table 8). Validation status is indicated to provide even greater confidence in design success (Table 9). To help researchers order the most applicable designs for their studies, minor allele frequencies (MAFs) in several populations are provided 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 a Custom SNP Panel

In addition to being an output file format, Score files can be used as input files to ADT. Thus, users can easily create a filtered or edited output file (with designs removed or added) for iterative ADT analysis while determining the optimal set of loci to order. Loci identified using more than one input search method (e.g., Gene, Region, Identity, Sequence, or ExistingDesign) can be combined as one Score file and resubmitted to ADT as an input file for evaluation as a single product.

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

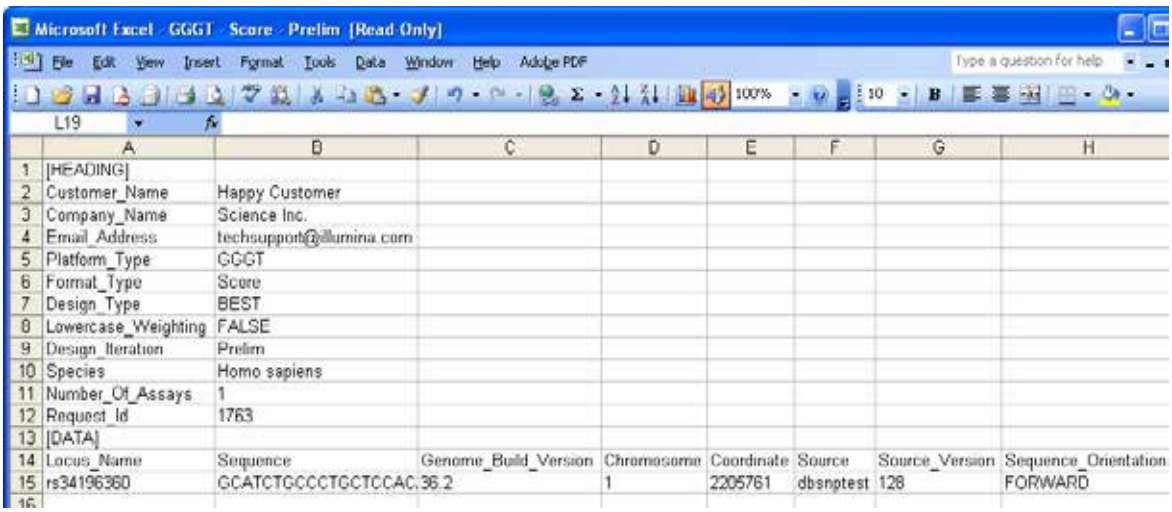
Table 5: Column Headings for Existing Design File

Heading	Description
Ilmn_Id	Ilmn_Id from original design manifest.

Table 6: File Headings for ADT Input File (Required Only for Email Submission)

Heading	Description	Required
Customer_Name	Name of person submitting the ADT file	Yes
Company_Name	Company name (no commas)	Yes
Email_Address	Customer's email address	Yes
Platform_Type	**GGGT**	Yes
Format_Type	**Gene**, **Region**, **Sequence**, **Identity, ExistingDesign**, or **Score**	Yes
Design_Type	**BEST, OTHER, TOP, BOT, PLUS** or **MINUS**	No
Lowercase_Weighting	**TRUE/FALSE**	No
Design_iteration	prelim	Yes
Species	**Species Name**	Yes
Number_of_Assays	Number of loci in file (may be **0** for Gene, Region, Identity, and Sequence files if the number of loci is unknown)	No
VeraCode_Pool	**FALSE**	No
Request_Id	Unique ADT identification number for file submission	No

Figure 7: Examples of File Header Section (Required Only for Email Submission)



Examples of properly formed entries in the header of a Score file shown from Excel. This header is only required for email submissions, and is formatted the same way for any preliminary file format.

Table 7: Score File Column Headers

Heading	Description
Locus_Name	RS number or customer's unique name.
Sequence	The bracketed site identified by the Locus_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.
Chromosome	Chromosome on which the marker is located. Must be a valid chromosome for the species being analyzed. Enter **0** if unknown.
Coordinate	Chromosome coordinate of marker. Enter **0** if unknown.
Source	Identify the source of the sequence and annotation data. Enter **unknown** if no information is available.
Source_Version	Source version number. Enter **0** if unknown.
Sequence_Orientation	Must contain one of the following three values: **forward**, **reverse**, or **unknown**. Information is customer-supplied and is not validated.
Region_Description	Description of the region of interest.
Final_Score	Ranges from 0 to 1.1, with higher values reflecting greater ability to design a successful assay.
Failure_Codes	If applicable, reasons why a successful assay at this marker locus is unlikely (Table 8).
Validation_Class	Numerical representation of validation_bin (Table 9).
Validation_Bin	Manner in which designed assays have been validated (Table 9).
MAF_Caucasian	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 African African-American Han Chinese Japanese Unknown
Chr_Count_Caucasian	
Study_Name_Caucasian	
MAF_African	
Chr_Count_African	
Study_Name_African	
MAF_African_American	
Chr_Count_African_American	
Study_Name_African_American	
MAF_Japanese	
ChrCount_Japanese	
Study_Japanese	
MAF_Chinese	
Chr_Count_Chinese	
Study_Name_Chinese	
MAF_Other	
Chr_Count_Other	
Study_Name_Other	

Table 7: Score File Column Descriptions Continued

Heading	Description
App_Version	Version of ADT used for scoring loci.
ILMN_ID	Unique identifier assigned by ADT to 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.
Coding_status*	NONSYN or SYNON. If the SNP falls within an exon, this field notes a synonymous or nonsynonymous amino acid change.
Amino_acid_change*	If the marker 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.
Exon_Location	If the locus is in a coding region, represents the location in the exon and the codon position.

*Additional gene annotation only in Score output file from submitted Gene, Identity, and Region files.

Table 8: List of Failure Codes for the ADT

Critical Failures (undesignable)	
101	Flanking sequence is too short.
102	Formatting error. Variant 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	Variant is not appropriate for Illumina platform. Possible causes: Tri- or quad-allelic variant Variant 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 marker in the list is closer than 61 nucleotides away.
360	SNP has a low score (<0.4)
399	Multiple contributing issues.

Table 9: Validation Status Descriptions

Validation_Bin	Validation_Class	Description
NonValidated	1	Locus 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.
OneKGenomeValidated	100	Locus has been sequenced in the 1000Genome project.
TwoHitValidated	110	Both alleles have been seen in two independent methods and populations.
HapMapValidated	120	Locus has been genotyped gy the HapMap project.
TwoHit_OneKGenomeValidated	200	Both alleles have been seen in two independent methods and populations. Locus has been sequenced in the 1000Genome project.
HapMap_OneKGenomeValidated	210	Locus has been genotyped gy the HapMap project. Locus has been sequenced in the 1000Genome project.
TwoHit_HapMapValidated	220	Locus has been genotyped gy the HapMap project. Both alleles have been seen in two independent methods and populations.
TwoHit_HapMap_OneKGenomeValidated	300	Locus has been genotyped gy the HapMap project. Both alleles have been seen in two independent methods and populations. Locus has been sequenced in the 1000Genome project.
GoldenGate_Validated	900	Design has been previously designed and successfully generated polymorphic results using the GoldenGate assay

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
Email_Address	Customer email address	Yes
Platform_Type	**GGGT**	Yes
Format_Type	**Gene, Region, Sequence, Identity, ExistingDesign or Score**	Yes
Design_Iteration	Final	Yes
Species	**Species Name**	Yes
Number_of_Assays	Number of loci in file (may be **0** for Gene, Region, Identity, and Sequence files if the number of loci is unknown)	No
VeraCode_Pool	**FALSE**	No
Purchase_Order_Number	**Customer purchase order number**	**Yes**
Scale(Number_of_Tubes)	**Must be 5 or greater**	**Yes**

1. **Remove designs 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 designs, since they have the highest chance of converting into functional assays.**
4. **Use two-hit or HapMap-validated loci with a preference for higher Final_Scores.**
5. **Avoid assays with Final_Scores < 0.4, which can decrease the overall performance of all assays.**
6. **Avoid assays assays containing SNPs with warning code 340, as these have a higher chance of failure**

Final Score File

After ADT analysis and custom selection of SNPs that meet the research criteria, a final Score file must be created to place an order. A preliminary Score or ExistingDesign file is converted to a final Score file by the completion of three header rows (white in Table 10): Design_Iteration, Scale (Number_of_Tubes), Purchase_Order_Number. It is important to ensure that the Number_of_Assays value in the final file matches the number on the corresponding quotation or contract. If an ExistingDesign file was used for ADT input, then all of the Ilmn_Id column values must be copied to the Ilmn_Id column to create a final order file. Final score files can be submitted via iCom or by emailing orders@illumina.com. Once the order is confirmed, the custom pool will take up to 60 days to manufacture.

Summary

Custom GoldenGate Genotyping products by Illumina allow researchers to create assays tailored directly to their specific needs for targeted region genotyping, fine-mapping of candidate disease association regions, and many more applications. The GoldenGate Assay can be deployed on either the BeadArray or BeadXpress platforms. The BeadArray platform supports the largest range of multiplex levels with options for LIMS and automation The BeadXpress platform is an ideal option for flexible, high-throughput genotyping. ADT provides a simple and powerful method for evaluating 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.

References

1. http://www.illumina.com/support.ilmn
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. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

illumina®