CpG Loci Identification

A guide to Illumina's method for unambiguous CpG loci identification and tracking for the GoldenGate® and Infinium® Assays for Methylation.

Introduction

There are more than 28 million CpG loci (aka CpG dinucleotides, CpG sites) in the human genome. Due to the increasing interest in epigenetic research—DNA methylation in particular—there is a strong need for a reference database where researchers can find information about individual CpG loci. Unlike other genomic loci, such as genes or SNPs, CpG loci lack a formal nomenclature. As a result, each CpG locus does not have a unique and unambiguous identifier to be used for clear reference to the locus in databases and other communications.

In addition to a unique CpG loci identifier, there is a need for a strand naming convention. CpG sequences are symmetric on forward and reverse strands of any double-stranded DNA. When identifying the methylation status of a CpG locus, it is necessary to unambiguously refer to the cytosine on the forward strand, reverse strand, or both.

Rather than using the evolving public databases to provide accurate CpG loci identifiers and strand orientation, Illumina has developed a method to consistently designate CpG loci based on the actual or contextual sequence of each individual CpG locus. The Illumina method takes advantage of sequences flanking a CpG locus to generate a unique CpG locus cluster ID with a similar strategy as NCBI's refSNP IDs (rs#) in dbSNP. This number is based on sequence information only and is unaffected by genome version. Illumina's standardized nomenclature also parallels the TOP/BOT strand nomenclature commonly used for SNP strand designation.

This document describes Illumina's CpG loci database and method for designating CpG loci IDs (cg#). The advantage of this method is that it will consistently designate the same CpG locus identifier and orientation calls even if public databases and genome assemblies change. This will enable all researchers to easily correlate the CpG loci identified today to research that may have been completed several years ago. Researchers can be confident that the same CpG loci are being analyzed across studies.

Designation of CG Identifier

The flanking sequences around the CpG dinucleotide are used to generate unique CpG cluster IDs (cg#), similar in concept to NCBI's refSNP cluster identifiers (rs#). Flanking sequences of 60 bases on each side of the CpG locus constitute a 122-base sequence used to define the locus. Any ambiguous nucleotide bases (e.g., N) in this flanking sequence are included. A unique "CpG cluster number" or cg# is assigned to each unique 122-base CpG locus.

The requirement for unique sequence differentiates the strategy of CpG cluster assignment from the concept of NCBI's refSNP clusters. A single CpG cluster can have multiple members that map onto different loci in a genome only if they have identical sequences. In contrast, the members of a refSNP cluster may or may not have identical sequences, but all are considered to be from the same location, or from a duplicated region of the genome.

Unique identification of CG Loci

Within a CpG cluster, three pieces of information are used to track individual member CpG loci: chromosome number, genomic coordinate, and genome build. Since a CpG locus contains two nucleotides, there are two genomic coordinates for a given site: one for C and the other for G. The lesser of the two coordinates is used as the coordinate of the CpG locus.

Of course, unique CpG sites are the most valuable when designing assays for methylation status. All Illumina standard methylation profiling products target CpG clusters that have only a single member. Custom methylation products should also be designed to target unique CpG sites. Data from the Illumina Assay Design Tool can be used for customer-designed methylation products, allowing easy targeting of unique CpG sites.

Table 1: CG Locus Designation In Illumina's CG Database

Cluster CG#	CHR #	Coordinate	Genome Build	Sequence	Top/Bot
cg00009407	14	88,360,674	36	...GGCG[CG]CTGC…	BOT
cg00003994	7	15,692,387	36	...TCTT[CG]TTGG...	TOP
cg00000292	16	28,797,601	36	...AATA[CG]GCCT…	TOP
cg00002426	3	57,718,583	36	...ACCA[CG]CTCT…	TOP
cg00005847	2	176,737,319	36	...ATGG[CG]CTTT…	BOT
cg00006414	7	148,453,770	36	...GGCG[CG]ATCC…	BOT

Figure 1: Sequencing Walking For CG Sites

The CpG dinucleotide is defined as position 'n', and pairs of bases are sequentially analyzed outward.

Designation of Strand

In addition to the cg# identifier assignment, the TOP/BOT strandedness of each CpG locus is determined. The designation of TOP or BOT strand for CG sites uses a sequence walking method similar to that used for SNPs. The difference is that both the C and G of the CpG locus are treated as a single unit (analogous to the SNP site). For this sequence walking method, the CpG dinucleotide is defined as position 'n'. The bases immediately before and after the CpG are 'n-1' and 'n+1', respectively (see Figure 1). Similarly, the second base before the CpG is 'n-2' and the second base after the CpG is 'n+2', etc. Using this method, sequence walking continues until an unambiguous pairing is present. An unambiguous pair is two bases equidistant from the CpG, one (and only one) of which is an A or T (i.e., A/G, A/C, T/C, or T/G). For the sequence shown in Figure 1, this occurs at the 'n-2'/'n+2' pairing, as the nucleotides in these positions are C and T, respectively. If the A or T in the first unambiguous pair is on the 5' side of the CpG, then the sequence is designated TOP. If the A or T in the first unambiguous pair is on the 3' side of the CpG (as in Figure 1), then the sequence is designated BOT.

Table 1 shows the identification information listed in Illumina's CpG database for several CpG loci out of the more than 27,000 present on the HumanMethylation27 BeadChip.

CG# Stability And Update Mechanism

As data about CpG loci accumulate, new members will be entered into the database. If the sequence of a new CpG locus is not found in the database, then a new cg# is assigned at random to the recently created CpG cluster. If the sequence is not new, then the CpG locus and its information are simply added to the existing cluster.

Some CpG loci may change cluster membership if the flanking sequence changes. When such changes occur, the lookup table in the CpG Cluster Tracking Database will be updated.

Conclusion

To unambiguously refer to CpG loci in any species, Illumina has developed a consistent and deterministic CpG loci database to ensure uniformity in the reporting of methylation data.

The advantage of Illumina's CpG database is that CpG cluster identifiers (cg#) and TOP/BOT strand orientations should be very stable despite frequent changes in RefSeq or other databases. By using the flanking sequences of a CpG locus, new loci can easily be included to update the database while simultaneously providing persistent identifiers for old loci.

Additional Information

Contact us at the address below to learn more about Illumina's method for CpG identification.

For information about the strategy Illumina uses for identifying SNP alleles and strand, see the technical note "TOP/BOT" Strand and "A/B" Allele. This is available from www.illumina.com on the Downloads page under the Support tab.

For information about Illumina DNA methylation analysis products, download the DNA Methylation Analysis Data Sheet from www.illumina.com/literature.

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

© 2010 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, GoldenGate Indexing, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, CSPro, GenomeStudio, Genetic Energy, HiSeq, and HiScan are registered trademarks or trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners. Pub. No. 270-2007-006 Current as of 1 February 2008

illumina®