TECHNICAL NOTE

CpG loci identification

Identify and track CpG loci for uniform reporting of methylation data generated using Infinium™ Methylation assays

illumina®

Introduction

There are more than 28 million CpG loci, also referred to as CpG dinucleotides or CpG sites, in the human genome. Due to the increasing interest in epigenetics research, particularly DNA methylation, there is a 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, there is no unique and unambiguous identifier to be used for clear reference to specific CpG loci in databases and other communications.

In addition to unique CpG loci identifiers, 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 refer to the cytosine on the forward strand, reverse strand, or both strands without ambiguity.

Rather than using the evolving public databases to provide accurate CpG loci identifiers and strand orientation, Illumina has developed a method to designate CpG loci consistently 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 refSNP IDs (rs#) in the SNP database (dbSNP). This number is based on sequence information only and is unaffected by genome version. The standardized Illumina nomenclature also parallels the top and bottom (TOP/BOT) strand nomenclature commonly used for SNP strand designation (Table 1).

This technical note describes the Illumina 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.

CG identifier designation

The flanking sequences around the CpG dinucleotide are used to generate unique CpG cluster IDs (cg#), similar in concept to NCBI 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 (for example, N) in this flanking sequence are included. A unique "CpG cluster number" (cg#) is assigned to each unique 122-base CpG locus (Table 1).

The requirement for unique sequences differentiates the strategy of CpG cluster assignment from the concept of NCBI 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.

Table 1: CG locus designation in the Illumina CG database

Cluster CG#	Chromosome	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

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.

Strand designation

In addition to 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 (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 (ie, 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, then the sequence is designated BOT.

For more information on identifying SNP alleles and strand, see our technical note on TOP/BOT strand and A/B allele.

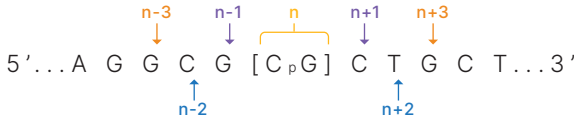

Figure 1: Sequence walking for CG sites—The CpG dinucleotide is defined as position 'n' and pairs of bases are sequentially analyzed outward.

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.

Summary

To refer to CpG loci in any species without ambiguity, Illumina has developed a consistent CpG loci database to ensure uniformity in the reporting of methylation data. The advantage of the Illumina 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 older loci.

Learn more

Infinium Methylation Assay, illumina.com/science/technology/microarray/infinium-methylation-assay

Identifying SNP alleles and strand, illumina.com/documents/products/technotes/technote_topbot.pdf

illumina®

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel
techsupport@illumina.com | www.illumina.com

M-GL-00921 v1.0