TECHNICAL NOTE

# "TOP/BOT" Strand and "A/B" Allele

A guide to Illumina's method for determining Strand and Allele for the GoldenGate® and Infinium™ Assays.

## INTRODUCTION

To address DNA strand designation and orientation for both human and non-human species, Illumina has developed a consistent and simple method to ensure uniformity in the reporting of genotype calls.

The method that Illumina has developed uses the top (TOP) and bottom (BOT) designations based on the polymorphism itself, or the contextual surrounding sequence. This document provides a description of the TOP/BOT method, as well as the generalized nomenclature of "Allele A" and "Allele B" within the Illumina genotyping system. Beginning in mid-2005, dbSNP also adopted this TOP/BOT nomenclature and has included this designation for all SNP entries.

## BACKGROUND

Historically, strand designation and orientation have presented many challenges to researchers in the field of genetics when it comes to consistency and comparing results across platforms and organizations. This confusion is compounded by the re-assembly of the genome during its working phases and extends currently to the updates within and across public Single Nucleotide Polymorphism (SNP) databases.

Rather than referencing the evolving public databases to provide accurate SNP strand and orientation, Illumina has developed a method to consistently designate SNPs based on the actual or contextual sequence of each individual SNP. The advantage of this method is that it will consistently designate the same SNP orientation and allele calls even if public SNP databases and genome assemblies change. This will enable researchers worldwide to easily correlate the genotype calls made today to research that may have been completed several years ago. Researchers can be confident that the same genotype calls are being made across time.

Additionally, although the human genome has been annotated and extensively SNP genotyped, the study of non-human species is growing rapidly. Much of this work is being done on species for which SNP databases are in their initial stages or do not yet exist. Many researchers that study both humans and non-human species also rely on proprietary SNP sequences that may not yet have been incorporated into public databases or for which assembly and orientation have not been established. There is a need, especially in non-human species, for a consistent and simple method for ensuring uniformity in the reporting of genotype calls to the research community.

## DESIGNATION OF STRAND AND ALLELE BASED ON THE ACTUAL POLYMORPHISM

The simplest case of determining strand and allele designations occurs when one of the possible variations of the SNP is an adenine (A), and the remaining variation is either a cytosine (C) or guanine (G). In this instance, the sequence for this SNP is designated TOP, and the A nucleotide is designated Allele A. Therefore, the C or G is Allele B.

Similar to the rules of reverse complementarity, when

TABLE 1: STRAND AND ALLELE DESIGNATIONS FOR UNAMBIGUOUS SNPS

| SNP Name[1] | Sequence[2] | Strand Designation | Allele A | Allele B |
|---|---|---|---|---|
| rs363040 | ...AGGAGGCTAG[G/T]CTCGCAGAGC... | BOT | T | G |
| rs536477 | ...GAGATTTAGG[A/G]AAAGATGTGA... | TOP | A | G |
| rs684517 | ...ACCAGGTACT[C/T]TGAACTTTAC... | BOT | T | C |
| rs2034107 | ...CATCTCCCCC[A/C]AAATCAGTTT... | TOP | A | C |

[1]SNP name and sequence are from dbSNP version 126
[2]Sequence shortened to 10 base pairs on each side of SNP as noted by the ellipses (...) for illustrative purposes. The unambiguous pairings used to determine Strand and Allele are noted in red.

one of the possible variations of the SNP is a thymine (T), and the remaining variation is either a C or a G, the sequence for this SNP is designated BOT and the T nucleotide is designated Allele A. The C or the G nucleotide is Allele B. See *Table 1* for more examples.

For these unambiguous situations, a condensed way of remembering these cases is that the Illumina method aims to designate the A as TOP and Allele A, and the T as Allele A on BOT.

If the SNP does not fit into either of these categories, then the surrounding sequence is used for Strand and Allele determination.

### DESIGNATION OF STRAND AND ALLELE BASED ON THE SURROUNDING SEQUENCE

If the SNP is an [A/T] or a [C/G], then the above rules do not apply. For the [A/T] SNP, the presence of an A would indicate TOP and the presence of a T would indicate BOT. But, both the A and the T would be designated as Allele A resulting in ambiguity. It would be similarly confusing to attempt to designate Strand and Allele for the [C/G] SNP. These [A/T] and [C/G] pairings are considered ambiguous for the purpose of determining Strand and Allele based on the SNP alone. Illumina employs a 'sequence walking' technique to designate Strand and Allele for [A/T] and [C/G] SNPs.

For this sequence walking method, the actual SNP is considered to be position 'n'. The sequences immediately before and after the SNP are 'n-1' and 'n+1', respectively (*see Figure 1*). Similarly, two base pairs before the SNP is 'n-2' and two base pairs after the SNP is 'n+2', etc.

FIGURE 1: SEQUENCE-WALKING TECHNIQUE FOR AMBIGUOUS SNPS


Using this method, sequence walking continues until an unambiguous* pairing is present. In the instance of *Figure 1*, this occurs at the n-1|n+1 pairing, as the nucleotides in these positions are C and T, respectively.

To designate Strand, when the A or T in the first unambiguous pair is on the 5' side of the SNP, then the sequence is designated TOP. When the A or T in the first unambiguous pair is on the 3' side of the SNP, then the sequence is designated BOT.

To designate Allele for an [A/T] SNP, when the Strand is TOP then Allele A = A and Allele B = T. When the Strand is BOT, then Allele A = T and Allele B =A.

To designate Allele for a [C/G] SNP, when the Strand is TOP then Allele A = C and Allele B = G. When the Strand is BOT then Allele A = G and Allele B = C.

Examples of Strand and Allele designation for [A/T] and [G/C] SNPs are shown in *Table 2*.

TABLE 2: STRAND AND ALLELE DESIGNATIONS FOR AMBIGUOUS SNPS

| SNP Name[1] | Sequence[2] | Strand Designation | Allele A | Allele B |
|---|---|---|---|---|
| rs1535632 | ...ACGGGGACA**G**[A/T]**T**ATGTTAACT... | BOT | T | A |
| rs363334 | ...ATGAGTG**A**AT[C/G]AA**G**CACTATT... | TOP | C | G |
| rs7101540 | ...AAATTCAGA**T**[A/T]**C**AGAATCTTT... | TOP | A | T |
| rs7113791 | ...GATGGACAG**G**[A/T]**T**GACCTCTAG... | BOT | T | A |
| rs778833 | ...GGTTAAAT**G**T[C/G]A**A**GGTGAGCT... | BOT | G | C |
| rs4933195 | ...ATGCTAAT**A**A[A/T]A**C**ATTAAAGT... | TOP | A | T |
| rs903997 | ...ATGAGA**A**AGT[C/G]TGA**G**AGTGCA... | TOP | C | G |
| rs1942968 | ...CTACATGA**C**T[C/G]T**T**TATGTTAC... | BOT | G | C |

[1]SNP name and sequence are from dbSNP version 126
[2]Sequence shortened to 10 base pairs on each side of SNP as noted by the ellipses (...) for illustrative purposes. The unambiguous pairings used to determine Strand and Allele are noted in red.

*An unambiguous pairing refers to a pair of nucleotides that are of the following combinations: A/G, A/C, T/C, or T/G. The pair is considered to be the nucleotides that are equidistant on either side of the SNP (e.g., n-1/n+1 or n-5/n+5).

TABLE 3: CONSISTENT STRAND AND ALLELE DESIGNATION

| SNP Name | Sequence[1] | Strand Designation | Allele A | Allele B |
|---|---|---|---|---|
| SNP1_Assembly1 | ....GGACCCGCAA[G/A]GAGGGCGCGG... | TOP | A | G |
| SNP1_Assembly2 | ...CCGCGCCCTC[C/T]TTGCGGGTCC... | BOT | T | C |
| | | | | |
| SNP2_Assembly1 | ...GGTAGCCTGA[A/T]ACCCCCAAGA... | TOP | A | T |
| SNP2_Assembly2 | ...TCTTGGGGGT[A/T]TCAGGCTACC... | BOT | T | A |

[1]Sequences are shortened to ten base pairs on either side of SNP as noted by the ellipses for illustrative purposes.
The unambiguous pairings used to determine Strand and Allele are noted in red.

## CONSISTENT STRAND AND ALLELE DESIGNATION

*Table* 3 demonstrates the utility of Illumina's TOP/BOT strand and A/B allele designation method. SNP1 was present in two assembly versions of this model genome, however the submitted sequences for assembly 1 and 2 are actually reverse-complements of each other. Situations such as this are not uncommon, especially in the beginning genome assembly stages where the sequence is evolving from small, unoriented contigs into large, assembled chromosomes.

Illumina's method maintains appropriate strand orientation and accompanying allele designation regardless of submitted sequence, which will also be carried through to the generation of final reports and nucleotide calls. SNP2 also demonstrates the consistency of orientation for an ambiguous SNP.

## ADDITIONAL INFORMATION

Contact us at the address below to learn more about determining strand and allele for GoldenGate and Infinium products.

**Illumina Technical Support**
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. 370-2006-018 27Jun06

illumina®
making sense out of life