A method for identifying genomic regions with high risk of consensus base call error in Next Generation Sequencing

Julio Avelar-Barragan, Ting Wang, Harry Lopez, Paul Gontarz, David Sayer
Thermo Fisher Scientific, West Hills, CA, USA

Abstract

Purpose: Next Generation Sequencing offers superior HLA genotyping resolution compared to other molecular typing methods. However, certain regions of the HLA genes, e.g., homopolymers, microsatellites, GC rich regions and regions susceptible to sequence mismapping, remain at high risk of base call error. We have developed an approach for assessing sequence quality whereby we calculate the percentage of the 2nd most frequent base that contributes to a consensus base call at every sequence position analyzed for all loci. Identification of poor-quality regions enable focused analysis strategies that minimize the risk of genotyping errors.

Results: Across 11 HLA loci, we observed that 100% of exonic bases fell outside a base call range previously determined to result in errors, indicating that they were called with high fidelity. Beyond exons, we also identified regions that were challenging for Illumina's sequencing chemistry by quantifying bases within the error range. Among 144 samples, we identified 792,567 non-coding base calls, with 14.3% falling within the 15-30% PSB range. Subsequently, this insight informed our analysis algorithm, leading to improved confidence in the final HLA typing results of the assay.

Introduction

Next-generation sequencing **(NGS)** for HLA genotyping is experiencing increased adoption, driving continuous advancements in both technology and sample diversity. A significant development in this field is the expansion of genomic regions used for typing, now encompassing potentially full-length genes to enhance HLA typing resolution. However, these expanded regions present new challenges in deriving reliable and accurate typing information, particularly in the absence of predicate devices for comparison. For instance, difficult to sequence regions, such as homopolymers, microsatellites, and GC-rich regions, can lead to typing errors. Additionally, random sequencing errors, misalignment of sequencing reads, and inherent allele biases can make typing of highly polymorphic heterozygous loci challenging. As the field progresses, it is crucial to establish methods for distinguishing and validating genomic regions that consistently yield accurate HLA typing results. Here, we present a method to identify base positions with alternative base call percentages falling outside the expected range of homozygosity or heterozygosity. Using this metric, sequences with high background and increased risk of consensus base call error can be pinpointed to minimize genotyping mistakes and improve assay or software development.

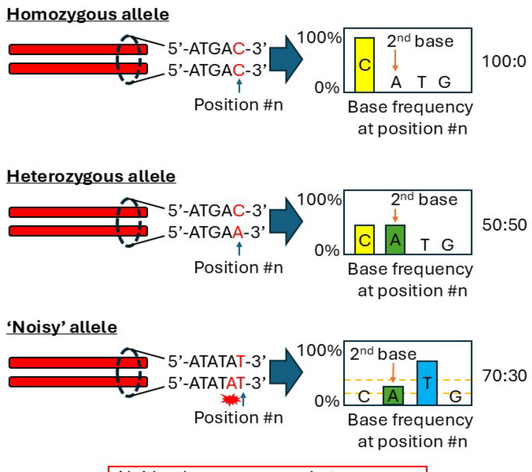
Materials and methods

Sample Preparation

One hundred and forty-four genomic DNA samples were enriched for 11 HLA loci using an NGS target enrichment assay. The loci enriched were HLA- A, B, C, DPA1, DPB1, DQA1, DQB1, DRB1, and DRB3/4/5. After enrichment, sample libraries were prepared for sequencing on Illumina's MiSeq platform.

Data Analysis

The resulting sequencing reads underwent analysis with One Lambda, Inc.'s TypeStream Visual Analysis Software™ to calculate the percentage of a second base call **(PSB)** at each base position of every locus. In the absence of nonspecific regions, homozygous alleles should have a PSB of 0%, with no alternative base calls, while heterozygous ones should have a PSB of 50%. Regions which are challenging to sequence or align can cause PSB to deviate from 0% or 50% and may result in unreliable base calling. To identify regions at high risk of consensus base call error, we calculated the PSB at each nucleotide position for every locus across 144 samples.



Results

Using the percentage of second base to identify genomic regions at high risk for consensus base call errors

After sequencing, TSV was used to map reads to a reference database, provide typing information, and calculate the PSB for each base position per loci. Visualization of the PSB revealed that both class I and II exons were well characterized, with 0% of nucleotides falling within a PSB range of 15-30%. By comparison, class II HLA loci contained substantially more nucleotides falling within a high risk PSB range in noncoding genomic regions (introns and UTRs). The maximum total percentage of bases falling within a PSB range of 15-30% for class I non-coding regions was 0.26%. For class II non-coding regions, this was a max of 22.85%, specifically within DRB3/4/5 introns. Inspecting the raw sequences revealed regions containing insertions, deletions, homopolymer microsatellites, and GC-rich regions.

Figure 1. A bar plot displaying the total percentage of nucleotides within a specific locus (y-axis) for a given PSB range (x-axis). A PSB range of 15-30% is at an increased risk for base call error and is colored in red.

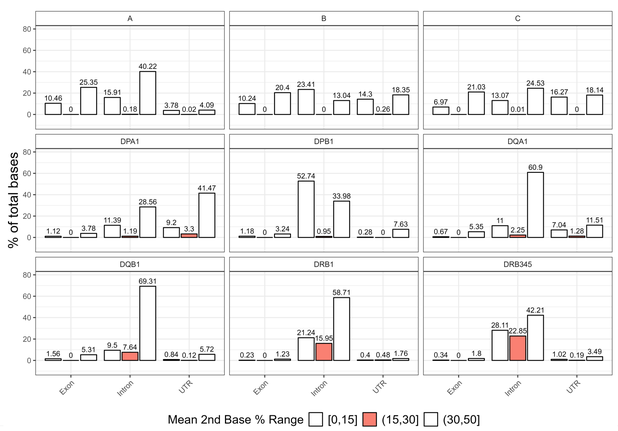
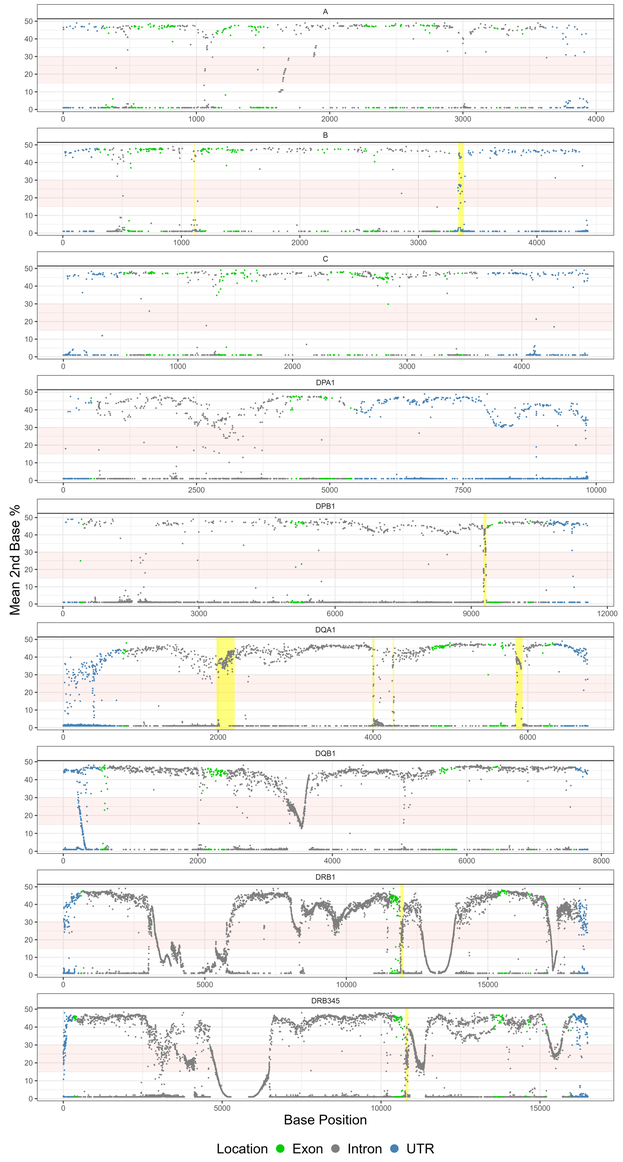


Figure 2. A scatter plot displaying the mean percentage of second base (PSB) for 144 samples (y-axis) against the nucleotide position along specified HLA loci (x-axis). Data points are color-coded by genomic region: blue for untranslated regions (UTRs), gray for introns, and green for exons. The red shaded area indicates a PSB range of 15-30%. Yellow highlighted regions denote examples of areas with increased risk of base call errors.



Conclusions

We developed a method in which we quantify the percentage of bases positionally to identify regions at risk for base call errors. This is centered on the following principals:

- Homozygous bases should not have any alternative base calls, while heterozygous ones should be 50:50 ratio.
- Regions which are challenging to sequence or align reliably can cause the percentage of the second most frequent base to deviate from 0 or 50%.
- We determined that regions with a second base percentage range of 15-30% should be avoided for informing HLA typing.
- Assessing the percentage of second base can be done in the absence of predicate devices.

Using the PSB metric, error prone regions can be pinpointed to minimize genotyping mistakes and improve assay or software development. Future applications include novel genotype validation and assay quality control.

Acknowledgements

We would like to thank Chris Ventura, JJ Chen, and Alexandre Vlassov for their valuable guidance and mentorship in the creation of this poster.