Title
Impact of three Illumina library construction methods on GC bias and HLA genotype calling

Permalink
https://escholarship.org/uc/item/43n7s4wq

Journal
Human Immunology, 76(2-3)

ISSN
0198-8859

Authors
Lan, James H
Yin, Yuxin
Reed, Elaine F
et al.

Publication Date
2015-03-01

DOI
10.1016/j.humimm.2014.12.016

Peer reviewed

Published in final edited form as:
Hum Immunol. 2015 March ; 76(2-3): 166–175. doi:10.1016/j.humimm.2014.12.016.

Impact of Three Illumina Library Construction Methods on GC Bias and HLA Genotype Calling

James H Lan, MD[1,2,*], **Yuxin Yin, PhD**[1,*], **Elaine F Reed, PhD**[1], **Kevin Moua, MS**[1], **Kimberly Thomas, PhD**[1], and **Qiuheng Zhang, PhD**[1,†]

[1]UCLA Immunogenetics Center, Department of Pathology & Laboratory Medicine, Los Angeles, CA, USA

[2]University of British Columbia, Clinician Investigator Program, Vancouver, BC, Canada

Abstract

Next-generation sequencing (NGS) is increasingly recognized for its ability to overcome allele ambiguity and deliver high-resolution typing in the human leukocyte antigen (HLA) system. Using this technology, non-uniform read distribution can impede the reliability of variant detection, which renders high-confidence genotype calling particularly difficult to achieve in the polymorphic HLA complex. Recently, library construction has been implicated as the dominant factor in instigating coverage bias. To study the impact of this phenomenon on HLA genotyping, we performed long-range PCR on 12 samples to amplify HLA-A, -B, -C, -DRB1, and -DQB1, and compared the relative contribution of three Illumina library construction methods (TruSeq Nano, Nextera, Nextera XT) in generating downstream bias. Here, we show high GC% to be a good predictor of low sequencing depth. Compared to standard TruSeq Nano, GC bias was more prominent in transposase-based protocols, particularly Nextera XT, likely through a combination of transposase insertion bias being coupled with a high number of PCR enrichment cycles. Importantly, our findings demonstrate non-uniform read depth can have a direct and negative impact on the robustness of HLA genotyping, which has clinical implications for users when choosing a library construction strategy that aims to balance cost and throughput with data quality.

Keywords

HLA genotyping; NGS; transposase; Illumina Nextera; GC Bias

1. Introduction

Human leukocyte antigen (HLA) genotyping is important in a number of clinical applications, including donor-recipient matching in hematopoietic stem cell transplantation (HSCT), identification of anti-donor HLA antibodies in solid organ transplantation, disease

[†]Corresponding author: Qiuheng Zhang, PhD, Assistant Professor, Pathology & Laboratory Medicine, UCLA Immunogenetics Center, 1000 Veteran Avenue, Room 1-520, 90024, (Tel): (+1) 310-206-0708, (Fax): (+1) 310-206-3216.
[*]Both authors contributed equally to this work

Disclosures
The authors have no conflicts of interest to declare.

association studies involving autoimmunity, and “personalized” risk assessment of hypersensitivity drug reactions [1–5]. Over the years, HLA typing has evolved from traditional serology-based techniques to the current molecular gold-standard in Sanger sequence-based typing (SBT) [6]. Despite this important progress, the Sanger method has reached a bottleneck both in throughput and its frequent complication of allele ambiguity. Ambiguity at the allele level can arise from incomplete exon sequencing, as well as the inability of SBT to set phase for linked polymorphisms. Recently, a number of groups have demonstrated the feasibility of using next generation sequencing (NGS), or “massive parallel” sequencing technologies, to overcome these limitations [7–11]. By exploiting its unique feature of massive clonal sequencing, NGS has the capacity to deliver unambiguous HLA typing at high-throughput and low cost.

While NGS is a much welcomed addition to our current arsenal of sequencing methods, the practicality of performing full-length HLA sequencing is highly dependent on the ability to expedite or automate library construction, the first step of all second-generation sequencing workflows [12]. This cumbersome process begins with DNA fragmentation (either mechanical or enzymatic), followed by end-repair, adaptor ligation, size selection (gel or bead method), and a library enrichment step using limited-cycle PCR. The entire procedure can take up to 9 hours when using standard library preparation kits, with a substantial component of hands-on time. Evidently, with the increasing efficiency of all bench-top platforms, the rate-limiting step in NGS data output is no longer the actual DNA sequencing time [12]; rather, the upfront library preparation imposes significant limitations not only on practical typing throughput, but also the ability of laboratories to satisfy sample turn-around-time.

To overcome this barrier, a number of novel chemistries have been developed to ameliorate the bottleneck of library construction. Among these, the Nextera Library Construction Kit (Illumina, San Diego, CA) is regarded as one of the most convenient and user-friendly approaches. By adapting the “cut and paste” mechanism of transposases to library construction, Nextera is able to combine DNA fragmentation and adaptor ligation into a single reaction, thereby shortening library construction time from 9 hours to only 90 minutes for up to 96 samples. Furthermore, Nextera XT, a similar kit which utilizes the same “tagmentation” method but requires only 1 ng DNA input, appears to be perfectly suited to the sequencing of bone marrow registry donor DNA samples, which typically contain DNA isolated from buccal swabs with concentrations that are far lower than the 100–200 ng required by conventional methods.

Despite this development, one recent report describes a higher degree of coverage bias being observed when using transposase-based library construction protocols [12]. Coverage (read depth) refers to the number of times a given nucleotide is represented in the aligned reads [13]. It is one of the most important data quality metrics in NGS, as it has a direct relationship with the sensitivity and specificity of variant detection [14]. Importantly, low read depth can confound allelic variant and genotype calling. The magnitude of this effect could be further amplified in the extremely polymorphic HLA system, where more than 10,000 alleles have been defined thus far (IMGT/HLA Database 3.15.0).

Herein, we sequenced and performed coverage analysis at 5 classical HLA gene loci (HLA-A, -B, -C, -DRB1, -DQB1). The objectives of this study were to evaluate the variable impact of standard TruSeq Nano and two transposase-based library preparation methods (Nextera and Nextera XT) in generating downstream coverage bias. Additionally, we sought to determine whether this bias could affect the accuracy of genotype calling at these genes.

2. Materials and methods

2.1 Genomic DNA (gDNA) Samples

Samples used in this study were selected from 12 healthy bone marrow volunteers representative of HLA alleles routinely encountered in our clinical laboratory. gDNA was isolated by the EZ1 Advanced XL system (QIAGEN, Valencia, CA) and then sequenced under high resolution by SBT to derive allele level G group reference typing at HLA-A, -B, -C, -DRB1, and -DQB1 (Table 1). Research approval for performing NGS sequencing on these samples was granted by the UCLA Institutional Review Board (IRB#14-000516).

2.2 PCR Amplification, Pooling, Clean-Up

Long-range PCR primers provided by One Lambda (Canoga Park, CA) amplified gDNA from promoter to 3′-UTR at HLA-A, -B, -C, -DRB1, and -DQB1. In brief, each PCR reaction consisted of: 25 ng gDNA, 4 μl polymerase buffer, 1.6 μl dNTP mixture, 2 μl primer mix, and 0.8 μl *LA Taq* DNA polymerase (TaKaRa Bio Inc., Japan), making a final volume of 20 μl. Each HLA locus was amplified in a single reaction, except -DRB1, which required 2 sets of overlapping primers to capture the entire length of the gene. GeneAmp PCR system 9700 (Life Technologies, Carlsbad, CA) thermal cycler conditions for class I genes were: 94°C for 2 min, followed by 30 cycles of (98°C for 10 sec, 67°C for 3 min). For class II, the optimized conditions were: 94°C for 2 min, followed by 30 cycles of (98°C for 10 sec, 70°C for 3 min). PCR products were confirmed on 1% agarose gel (Sigma-Aldrich, St. Louis, MO), and then mixed with Agencourt AMPure XP beads (Beckman Coulter, Brea, CA) in 0.6× PCR reaction volume to undergo purification. Qubit fluorometer 2.0 (Life technologies, Grand Island, NY) was used to quantitate and normalize amplicon concentrations, accounting for the different amplicon fragment lengths; this was followed by equimolar pooling of all HLA loci. Aliquots from each PCR pool were subjected to three Illumina library construction protocols to allow unbiased downstream coverage comparisons. Fig. 1 provides an overview of the main workflow and technical differences between these methods.

2.3 Library Construction

2.3.1 TruSeq Nano DNA Sample Preparation Kit (Standard Method)—Aiming for an insert size of 550 bp, 200 ng of pooled PCR products were sheared by sonication using Covaris M220 (Covaris, Woburn, MA) with duty factor 20%, peak incident power 50 W, and 200 cycles per burst for 60 sec. Amplicon fragments underwent further purification with Agencourt AMPure XP beads in 1.6× reaction volume. After end repair, library templates proceeded to double size selection: large fragments were removed using diluted AMPure XP beads in 1.6× reaction volume; small fragments were removed by mixing 30 μl of undiluted AMPure XP beads with 250 μl of sample supernatant containing the DNA of interest.

Following 3′ adenylation, Illumina adapter indices were ligated to the purified fragments, such that pooled PCR products from each clinical sample were labeled with a unique adaptor index. 25 μl of each library sample was then mixed with 5 μl PCR Primer Cocktail and 20 μl Enhanced PCR Mix to carry out target library enrichment with a limited-cycle PCR: 95 °C for 3 min, followed by 8 cycles of (98°C for 20 sec, 60°C for 15 sec, 72°C for 30 sec), and 72°C for 5 min. Enriched libraries were then purified with AMPure XP beads in 1:1 ratio.

2.3.2 Nextera DNA Sample Preparation Kit—

Simultaneous amplicon fragmentation and adaptor sequence ligation were performed according to manufacturer protocol. Briefly, 50 ng input DNA (in 20 μl) was mixed with tagmentation buffer (25 μl) and enzyme (5 μl), then incubated at 55°C for 5 min. The tagmented DNA was purified using the Zymo DNA Clean & Concentrator kit (Zymo Research, Irvine, CA). A limited-cycle PCR was carried out to enrich and perform dual indexing on the tagmented DNA: 72°C for 3 min, 98°C for 30 sec, and 5 cycles of (98°C for 10 sec, 63°C for 30 sec, 72°C for 3 min). Indexed libraries were purified using AMPure XP beads in the quantity of 0.5× reaction volume. Library validation and quantification were performed on the Agilent Technologies 2200 Tapestation (Agilent Technologies, Santa Clara, CA).

2.3.3 Nextera XT DNA Sample Preparation Kit—

This protocol followed the same workflow as Nextera with the exception of a few notable differences. First, only 1 ng of input DNA was required for the tagmentation reaction. Second, a higher number of PCR cycles was utilized for DNA enrichment: 72°C for 3 min, 95°C for 30 sec, and 12 cycles of (95°C for 10 sec, 55°C for 30 sec, 72°C for 30 sec), and 72°C for 10 min. Finally, instead of using Zymo DNA Clean, PCR clean-up was performed by mixing amplified libraries with AMPure beads in 0.5× library volume. To maintain consistency between protocols, library normalization and pooling were performed in the same fashion as in TruSeq Nano and Nextera, as described in 2.4.

2.4 Library Normalization, Pooling, and Sequencing on MiSeq

Sequence-ready libraries were validated and quantitated on the High Sensitivity D1000 ScreenTape (Agilent Technologies, Santa Clara, CA) to allow for library normalization and equimolar pooling of all study samples. Pooled libraries were diluted and loaded at 12 pM on a MiSeq flow cell, with 15% phiX spiked in. Three separate paired-end sequencing runs were performed overall (one for each protocol), using MiSeq Reagent Kit v2, 500 cycles (Illumina, San Diego, CA).

2.5 HLA Coverage Analysis and Genotyping

Raw sequence outputs were imported as FASTQ files into NGSengine (Gen Dx 1.3.0, Utrecht, Netherlands) for read alignment and genotype calling (using IMGT/HLA Database 3.15.0 as reference). A read length of 15 bp or greater was prerequisite for read alignment. The default minimum coverage threshold used to assign genotypes was set at 20×. The *Alignment View* function of NGSengine allows full-length coverage visualization of the assigned alleles. Due to the lack of complete genomic reference sequence for a number of DRB1 alleles, DRB1 sequence outputs were imported as FASTQ into Omixon Target 1.8.0

(Omixon, Budapest, Hungary), to permit exon-only coverage analysis. Depth bias was calculated by taking the ratio of the maximum to minimum coverage identified for each allele as indicated on the coverage plot.

2.6 GC Content Calculation

GC content was calculated by importing reference allele sequences from IMGT/HLA Database 3.15.0 into an interactive web-based *DNA Base Composition Analysis Tool* (*J. Zheng, Queen's University, Canada*) [33]. GC% was calculated in 100 bp sliding windows, shifting one nucleotide at a time, then normalized to 60% to facilitate identification of GC-rich territories.

2.7 Statistical Analysis

Statistical significance of differences in coverage bias was determined by applying the Kruskal-Wallis non-parametric method to compare the median max/min coverage ratio (depth bias) derived between the three groups. Multiple comparisons were evaluated using a Wilcoxon Rank Sum test with a Bonferroni correction; p-values were two-sided. A p-value < 0.01 was considered significant. Data were analyzed using STATA (StataCorp. 2013. Stata Statistical Software: Release 13, College Station, TX).

3. Results

3.1 Comparison of Library Fragment Size between Covaris and Transposase-Treated Samples

Covaris and transpose-based methods generated library fragments with different lengths. Library size assessment of a representative sample prepared using the three different protocols is shown in Supplementary Fig. 1a. Covaris shearing together with the use of SPRI (solid phase reversible immobilization) beads for size selection produced the anticipated median library size of approximately 700 bp (550 bp fragment insert plus 135 bp dual indices) while Nextera and Nextera XT generated slightly larger libraries (900–1000 bp). This disparity was due to the use of double size selection to remove both small and large fragments in TruSeq Nano, while only single size selection was performed to remove small fragments in Nextera and Nextera XT. Notwithstanding this difference, after downstream sequencing and trimming, the resultant median insert size of the aligned reads was similar between all three protocols (Supplementary Fig. 1b). Of note, transposase-treated samples exhibited a broader size range (Nextera: 279 – 468 bp, Nextera XT: 283 – 401 bp) compared to that generated by TruSeq Nano (366 – 394 bp).

3.2 Delineating Whole-Gene Coverage Distribution and Bias in the HLA Complex

Different alleles within each HLA gene exhibited the same coverage pattern with little variations (Supplementary Fig. 2a – 2d). Most of the class I alleles demonstrated coverage depression in the first 200–1400 bp of the gene; this similarity is likely the result of gene duplication and rearrangement in evolution, which has led to shared sequence homology and base composition between these loci [15]. In contrast to class I, coverage loss for DQB1 alleles occurred at a different position, located between the first 1800–2800 bp of the gene (Supplementary Fig. 2d).

3.3 Comparative Analysis of Coverage Bias between TruSeq Nano, Nextera and Nextera XT Library Construction Methods

Next, we performed side-by-side comparisons of coverage bias produced by different library preparation methods. A representative example is depicted in Fig. 2a. As shown, TruSeq Nano generated fairly even depth across all HLA loci, indicating little bias in this method by gross inspection. In contrast, both transposase-based protocols generated significant read loss in the susceptible regions that were defined above (Fig. 2a, boxes). This phenomenon was especially evident in samples sequenced using Nextera XT, where the degree of bias was large enough to result in gaps for certain alleles (Fig 2a, arrows). To further quantify the magnitude of this effect, we calculated the maximum to minimum coverage ratio of each allele in all of the study samples to derive boxplots in Fig. 2b. As shown, the median max/min coverage ratio at each HLA locus differed significantly between all three library preparation protocols (except between TruSeq Nano and Nextera in HLA-A). This implies that the strategy of library construction alone can generate varying gradations of downstream bias. In the case of Nextera XT, extreme bias was observed in all sequenced samples, with max/min coverage ratios typically in the 25–30 fold range. In addition to prominent inter-protocol bias differences, the bias variability within each library preparation method was also distinctly different from each other. TruSeq Nano generated the smallest variability, while progressively larger variances were detected in Nextera and Nextera XT (Fig 2b, whiskers).

3.4 Characterizing HLA Regions with Coverage Bias

To further elucidate the etiology of bias in the HLA complex, we proceeded to correlate GC content with regions of coverage depression, as other groups using NGS to sequence bacterial genomes have reported this association [12, 16]. First, the mean GC% of all alleles sequenced at each locus was calculated in 100 bp sliding windows, shifting one nucleotide position at a time, to construct locus-specific GC% plots. Next, using GC% > 60 as the cut-off to define high GC content, we obtained normalized GC content plots in Fig. 3, where areas under the curve with normalized GC% >1.0 represent GC-rich territories. In Fig. 4, we overlaid coverage plots from Fig. 2a with their respective normalized GC% plots – this demonstrated a correlation where GC abundance was a good predictor of coverage loss at both class I and II loci. The same overlaying coverage plots for HLA-A and -B are shown in Supplementary Fig. 3.

3.5 Impact of Coverage Bias on HLA Genotyping Accuracy

The practical significance of coverage bias in NGS HLA typing extends beyond throughput and cost reduction. As demonstrated in Fig. 2a, many of the GC-rich regions in HLA encompass clinically important exons which encode polymorphic antigen-recognition sites on HLA molecules. Coverage loss in these regions increases the risk of allele mis-assignment, particularly if multiple closely-related alleles harbor polymorphisms within these poorly represented sites. To test this hypothesis, we next sought to determine whether different library preparation methods could influence the accuracy of HLA genotype calling. As shown in Fig. 5, both transposase-based methods, which have a tendency towards greater GC bias, indeed produced more typing errors compared to TruSeq Nano. Table 2 provides a list of the alleles mis-assigned by NGS according to the library construction protocol utilized. During troubleshooting we manipulated the minimum depth threshold (to 5×, 10×, 50×) used by Gen Dx to assign genotypes – this did not improve the accuracy of allele calling in the mistyped samples. Instead, coverage bias leading to read imbalance between two heterozygous alleles appeared to be the main culprit behind the majority of the mis-assigned alleles. As an example, Figure 6a illustrates the contribution of coverage bias to the mistyping of C*03:04 in sample 063 prepared by Nextera and Nextera XT. Reference typing of sample 063 shows two heterozygous alleles C*03:04 and C*01:02, each carrying a sequence motif of TAA and CCT at positions 353, 355, 361, respectively. Using TruSeq Nano, the most bias-free method of the three, an inherent read imbalance was observed for the two motifs at baseline (TAA: 35%; CCT: 65%) – at this degree the imbalance did not impact the algorithm's ability to align reads at these positions. When using transposase-based methods, however, the CCT reads became over-represented (Nextera: 83%; Nextera XT: 79%). This coverage imbalance resulted in the software's preferential usage of the overabundant CCT reads to create the hybrid mistyped allele C*03:17, which is identical to C*03:04 in their exon 2–3 coding sequence except for the TAA → CCT substitution in exon 3. In this study, alleles mis-assigned through this mechanism occurred when the less abundant alleles were present in <22% of the total reads in a heterozygous sample. Using the current software, the overall depth of the SNP motifs did not necessarily correlate with read balance, as shown in Figure 6b. In this example, although Nextera achieved a higher overall depth at SNP positions 354–357 when compared with Nextera XT (171 versus 106 read depth), it still exhibited a greater degree of read imbalance between the two alleles compared to its counterpart (Nextera: 79%/21%, Nextera XT: 65%/35%); this difference led to the selective mistyping of B*27:05:02 as B*27:14 in Nextera but not in Nextera XT. The same mechanism of error can also occur at the exon level. Supplementary Fig. 4 uses IMGT/HLA's *Sequence Alignment Tool* to show that the mistyped B*40:184 allele in Nextera sample 063 was in fact a hybrid of the exon 2 sequence of Allele 2 (B*40:02:01) and the exon 3 sequence of Allele 1 (B*39:01:01:01).

In contrast to coverage bias, protocol-independent errors accounted for the minority of allele mis-assignment. Software-related mistyping, defined as allele mis-assignment represented by an adequate global coverage (> 100) and free of depth bias, accounted for two of the errors, which have been resolved with the updated Gen Dx 1.4.0 software. Similarly, primer-related allele imbalance led to the mistyping of DQB1 in sample 046 in all three methods - these primers have since been optimized to resolve this issue.

3.6 HLA-DRB1 Coverage Analysis

A majority of DRB1 alleles lack complete reference sequences in the IMGT/HLA Database 3.15.0. Moreover, in those with complete reference sequences, DRB1 alleles can vary greatly in their whole-gene length, from 11–18 kb in our sequenced samples. Together, these factors preclude the use of a generic GC% plot to represent the base composition of all DRB1 alleles. To bypass this issue, we performed coverage analysis on exons 2 and 3 of DRB1 to focus on the most clinically relevant, allele-defining sequences. Unlike other HLA loci, DRB1 exon-only GC% plot displays even base composition with no identifiable GC-rich exons (Fig. 7). Thus, we would predict little bias in DRB1. Surprisingly, we observed a consistent tendency for coverage loss at the terminal end of exon 2 in a majority of our DRB1 reads (Supplementary Fig. 5). Evidently, another factor unrelated to GC content was the source of this phenomenon. We examined the exon 2 sequence of DRB1 and found a unique pattern of intronic $(GT)_x(GA)_y$ repeats ranging from 42 – 90 bp at the DRB1 exon 2-intron 2 junction, which correlated with the low sequence coverage of this region (Supplementary Fig. 6). We also performed a microsatellite search within all the alleles sequenced in this study to determine if a similar association could be found. Using Tandem Repeats Finder [17] and imputing a period a size of 10, the algorithm explored a given allele's DNA sequence for the presence of tandem repeats with a pattern size ranging between 1–10 nucleotides. Overall, we found only one tandem repeat $(TTTA)_z$ present within the intron 1 of certain DRB1 alleles, in addition to the $(GT)_x(GA)_y$ motif we previously identified. Coverage analysis of this region revealed a diminished depth akin to the pattern we observed at the terminal end of DRB1 exon 2 (Supplementary Fig. 7).

3.7 Effect of Modifying Library Enrichment Conditions on GC Bias

Having established the association of depth bias in GC-rich regions, particularly in transposase-treated samples, we performed additional experiments to determine if simple library protocol modifications, such as PCR cycle reduction or the addition of a PCR denaturant (betaine), could further improve coverage uniformity. Here, we exposed a Nextera XT-treated study sample to two experimental conditions: i) reduced enrichment PCR cycles (6×), and ii) addition of 1M betaine to the PCR reaction (12× PCR cycles). The output coverage profiles are shown in Supplementary Fig. 8. In both instances, our adaptations were unable to contribute improvements over the existing commercial Illumina protocol.

4. Discussion

Accurate single nucleotide polymorphism (SNP) determination is critical to the success of HLA genotyping by NGS [18]. Using this methodology, short fragment reads are first assembled by alignment to reference sequences; allele-defining SNPs are then identified and used by software to derive the most likely genotype. Uneven sequence coverage hampers this effort by introducing uncertainty into variant calling, which increases the complication of genotype mis-assignment. Recently, library construction has been implicated as the dominant factor in generating downstream bias [16, 19]. From the inception, non-uniform DNA shearing can introduce fragment length bias, which may lead to preferential removal of certain reads during subsequent size-selection. Further, the method of barcode/adaptor tagging is increasingly recognized to influence both the quality and quantity of mapped reads [20]. Finally, target enrichment by PCR can favor the amplification of some library fragments over others, creating additional bias. Together, these factors underscore the importance of understanding the linkage between library construction and output data quality, as this phenomenon may affect final read assembly and the accuracy of genotype calling.

In line with previous NGS reports, we found high GC content (>60%) to be a good predictor of low read depth in the HLA complex [12, 16, 21, 22]. This effect was observed in all three library preparation protocols, but particularly striking when using the transposase-based strategy. Several groups suggest enrichment PCR in library construction to be the principal offender of GC bias [16, 23, 24]. While we note the same finding in our study, this factor alone is unlikely to explain the differential magnitude of bias we observed between TruSeq Nano and transposase-based methods. For one, the number of PCR enrichment cycles in TruSeq Nano was actually higher than the number of cycles utilized in Nextera, which would lead one to predict more GC bias in the conventional protocol. Given our unexpected results, it is more probable that the use of the transposase itself introduced another distinct mechanism of GC bias unrelated to PCR amplification. Indeed, previous mechanistic studies demonstrated transposon-mediated DNA integration to be a non-random process [25–27]. Further, when Adey *et al* compared mechanical versus transposase-based DNA fragmentation, they also observed a more pronounced signature of GC bias in transposase-based sequencing of human and *Drosophila* genomes [12].

GC bias was most pronounced when sequencing libraries were prepared using Nextera XT. This was likely the synergistic effect of transposase-mediated bias being coupled with a high number of PCR enrichment cycles. This combination led to unusually low reads, even regional gaps, in important allele-defining exons containing high GC%. Additionally, Nextera XT library samples exhibited large variances in their coverage bias, which is especially undesirable in clinical HLA typing, where a high standard of precision and reproducibility is demanded. We attempted to improve GC bias by manipulating Nextera XT's library PCR cycle and introducing betaine as a PCR denaturant. Overall, these modifications did not yield enhancements over the existing commercial protocol. Successful application of Nextera XT to clinical genotyping, therefore, will likely require a trial of various PCR additives/enzymes under different conditions or a multi-prong approach to rescue GC-rich target sequences.

We also found a tendency towards depth loss at the exon 2 terminus of DRB1, which appeared to be GC independent. This bias was traced to a long stretch of intronic dinucleotide $(GT)_x(GA)_y$ repeats which flank the 3′ terminus of exon 2. This tandem repeat motif was previously identified in the DRB region of humans and other primates [28]. Additional microsatellite search using Tandem Repeats Finder [17] identified a separate $(TTTA)_z$ repeat within the intron 1 of certain DRB1 alleles which also supports this coverage bias relationship. Genome regions containing highly repetitive DNA have long posed sequencing difficulties even for the traditional Sanger method [29, 30]. In NGS, this refractory motif can predispose to sequencing errors, which may result in excessive read removal in subsequent QC filtering. Alternatively, short reads with redundant sequences face an increased risk of mal-alignment during data assembly, which could contribute to the generation of mapping artifacts [31]. Given our limited dataset, the robustness of this observation needs to be further evaluated with a larger pool of DRB1 alleles.

In this study, tagmentation generated greater coverage non-uniformity compared to mechanical DNA shearing by sonication. Importantly, we showed this depth bias had a direct and negative effect on genotype calling. Algorithm and primer-related issues were responsible for only a minority of the errors observed in this study. In contrast, allele coverage bias was the principal cause of genotyping errors found in transposase-treated samples. Sequence alignment of mis-assigned alleles with their reference counterparts demonstrated key nucleotide changes in positions that exhibit coverage bias (exon 2 and beginning of exon 3 in class I genes). In a heterozygous combination, reads from one allele may be lost disproportionally compared to reads from the other allele. When a sufficient imbalance threshold is reached (<22% of total reads in this study), the software may favor using reads from the dominant allele while ignoring sequences from the minor allele (treated as background artifacts). As demonstrated in our study, this mechanism can generate hybrid consensus sequences which end in genotyping errors.

Currently, most commercial NGS HLA software incorporate various quality metrics such as allele coverage and balance to inform users on the reliability and confidence of genotype calling. The majority of these indicators provide allele health information on the global level, which may not necessarily identify regional read imbalance at polymorphic SNP sites as a potential quality issue. Besides incorporating this data into the quality display of future NGS software, reducing coverage bias as well as considering the degree of allele imbalance at key exon/SNP positions may prove important in averting this form of allele mis-assignment. Additionally, application of new bioinformatics algorithms to NGS read assembly has been show to partially compensate for the adverse effects of sequencing bias and tandem repeats [32].

In conclusion, library construction is not a random event. Variations in this step of NGS methodologies can impart significant differences in downstream bias, particular in GC-rich regions, which we demonstrate can carry through to affect final genotype calling. The magnitude of this phenomenon is greatest in transposase-based library preparation protocols compared to the conventional approach, which uses sonication to generate DNA fragments. These findings need to be considered and accounted for when adapting NGS technology to perform HLA sequencing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

JHL is supported by funding from the University of British Columbia Clinician Investigator Program. The authors give thanks to One Lambda for supplying the HLA PCR primers.

References

1. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, et al. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood. 2007; 110:4576. [PubMed: 17785583]
2. Flomenberg N, Baxter-Lowe LA, Confer D, Fernandez-Vina M, Filipovich A, Horowitz M, et al. Impact of HLA class I and class II high-resolution matching on outcomes of unrelated donor bone marrow transplantation: HLA-C mismatching is associated with a strong adverse effect on transplantation outcome. Blood. 2004; 104:1923. [PubMed: 15191952]
3. de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet. 2006; 38:1166. [PubMed: 16998491]

4. Fernando MM, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, et al. Defining the role of the MHC in autoimmunity: a review and pooled analysis. PLoS Genet. 2008; 4:e1000024. [PubMed: 18437207]

5. Pavlos R, Mallal S, Phillips E. HLA and pharmacogenetics of drug hypersensitivity. Pharmacogenomics. 2012; 13:1285. [PubMed: 22920398]

6. De Santis D, Dinauer D, Duke J, Erlich HA, Holcomb CL, Lind C, et al. 16(th) IHIW: review of HLA typing by NGS. Int J Immunogenet. 2013; 40:72. [PubMed: 23302098]

7. Shiina T, Suzuki S, Ozaki Y, Taira H, Kikkawa E, Shigenari A, et al. Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. Tissue Antigens. 2012; 80:305. [PubMed: 22861646]

8. Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, Slavich L, et al. Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum Immunol. 2010; 71:1033. [PubMed: 20603174]

9. Holcomb CL, Höglund B, Anderson MW, Blake LA, Böhme I, Egholm M, et al. A multi-site study using high-resolution HLA genotyping by next generation sequencing. Tissue Antigens. 2011; 77:206. [PubMed: 21299525]

10. Lange V, Böhme I, Hofmann J, Lang K, Sauter J, Schöne B, et al. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics. 2014; 15:63. [PubMed: 24460756]

11. Wang C, Krishnakumar S, Wilhelmy J, Babrzadeh F, Stepanyan L, Su LF, et al. High-throughput, high-fidelity HLA genotyping with deep sequencing. Proc Natl Acad Sci U S A. 2012; 109:8676. [PubMed: 22589303]

12. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 2010; 11:R119. [PubMed: 21143862]

13. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014; 15:121. [PubMed: 24434847]

14. Koboldt DC, Ding L, Mardis ER, Wilson RK. Challenges of sequencing human genomes. Brief Bioinform. 2010; 11:484. [PubMed: 20519329]

15. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet. 2009; 54:15. [PubMed: 19158813]

16. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011; 12:R18. [PubMed: 21338519]

17. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27:573. [PubMed: 9862982]

18. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011; 12:443. [PubMed: 21587300]

19. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, et al. Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. BMC Genomics. 2012; 13:1. [PubMed: 22214261]

20. Bystrykh LV. Generalized DNA barcode design based on Hamming codes. PLoS One. 2012; 7:e36852. [PubMed: 22615825]

21. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 2008; 36:e105. [PubMed: 18660515]

22. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456:53. [PubMed: 18987734]

23. Frey UH, Bachmann HS, Peters J, Siffert W. PCR-amplification of GC-rich regions: ‘slowdown PCR’. Nat Protoc. 2008; 3:1312. [PubMed: 18714299]

24. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods. 2009; 6:291. [PubMed: 19287394]

25. Ason B, Reznikoff WS. DNA sequence bias during Tn5 transposition. J Mol Biol. 2004; 335:1213. [PubMed: 14729338]

26. Reznikoff WS. Transposon Tn5. Annu Rev Genet. 2008; 42:269. [PubMed: 18680433]

27. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. Insertion site preference of Mu, Tn5, and Tn7 transposons. Mob DNA. 2012; 3:3. [PubMed: 22313799]

28. Doxiadis GG, de Groot N, Claas FH, Doxiadis II, van Rood JJ, Bontrop RE. A highly divergent microsatellite facilitating fast and accurate DRB haplotyping in humans and rhesus macaques. Proc Natl Acad Sci U S A. 2007; 104:8907. [PubMed: 17502594]

29. Kieleczawa J. Fundamentals of sequencing of difficult templates–an overview. J Biomol Tech. 2006; 17:207. [PubMed: 16870712]

30. Zhao X, Haqqi T, Yadav SP. Sequencing telomeric DNA template with short tandem repeats using dye terminator cycle sequencing. J Biomol Tech. 2000; 11:111. [PubMed: 19499047]

31. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012; 13:36.

33. Zheng, J. DNA base composition analysis tool. Queen's University; Canada: Accessed Feb 15, 2014. <http://molbiol-tools.ca/DNA_composition.htm>

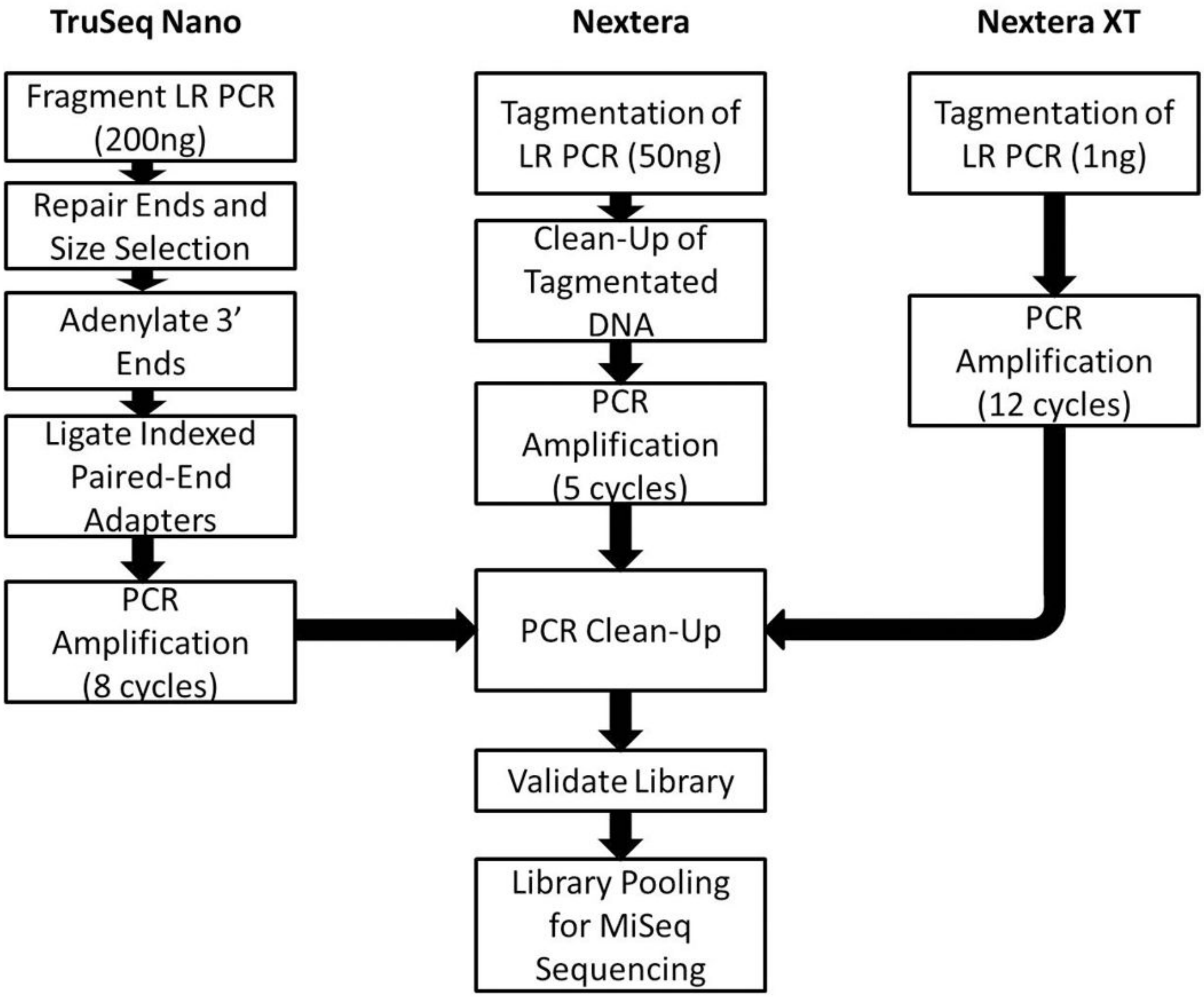

TruSeq Nano
Nextera
Nextera XT
Fragment LR PCR (200ng)
Repair Ends and Size Selection
Adenylate 3' Ends
Ligate Indexed Paired-End Adapters
PCR Amplification (8 cycles)
Tagmentation of LR PCR (50ng)
Clean-Up of Tagmentated DNA
PCR Amplification (5 cycles)
Tagmentation of LR PCR (1ng)
PCR Amplification (12 cycles)
PCR Clean-Up
Validate Library
Library Pooling for MiSeq Sequencing

Fig. 1.

Comparison of three Illumina library construction protocols. Covaris M220 was used to fragment DNA in TruSeq Nano.

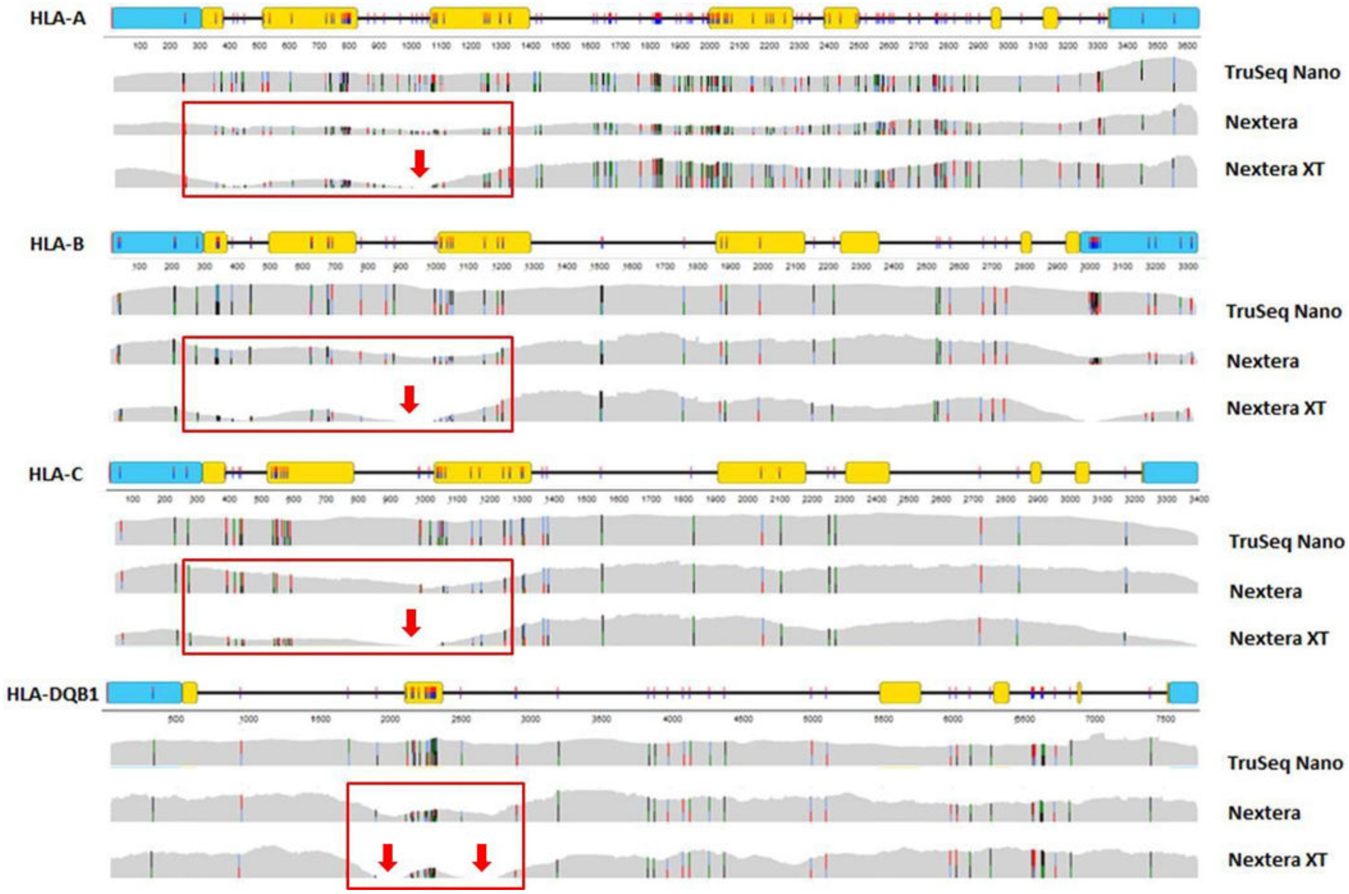


Fig. 2a.

Impact of library construction on coverage bias. Coverage data taken from a representative sample is shown: depth bias in class I genes occur in the same gene region, in contrast to that observed in HLA-DQB1. Transposase-based protocols generated larger depth bias compared to TruSeq Nano (red boxes). The magnitude of bias was greatest in Nextera XT, which produced regional gaps (arrows). Blue box, UTR; yellow box, exon.

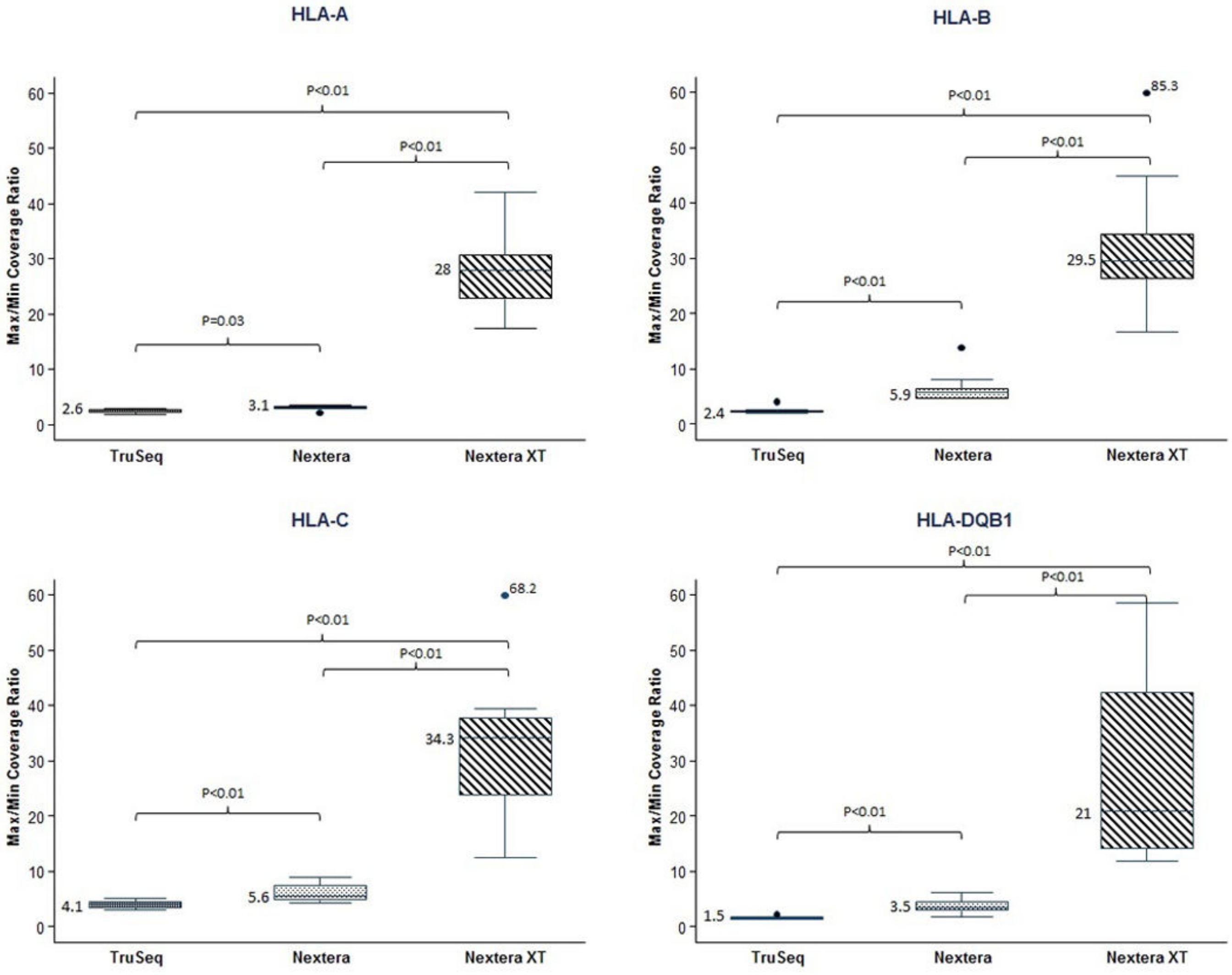


Fig. 2b.

Quantifying bias between protocols. The magnitude of bias was statistically significant between all three protocols ($p < 0.01$), except between TruSeq Nano and Nextera at HLA-A ($p < 0.05$).

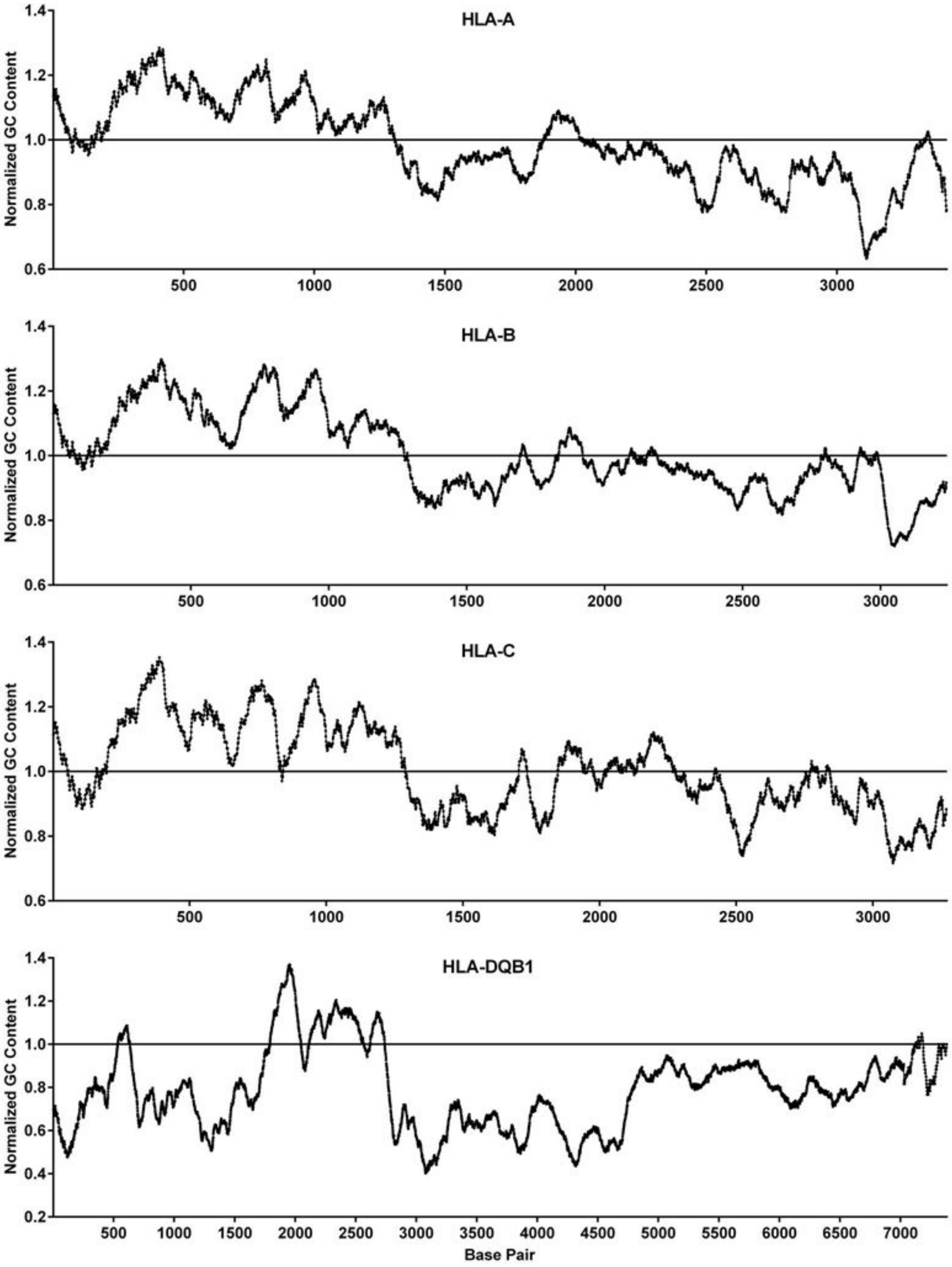


Fig. 3.

Normalized GC content plot. Each GC% plot represents the mean GC content of all alleles sequenced at each locus calculated in 100 bp sliding windows, shifting one nucleotide position at a time and normalized to 60%. Areas under the curve with normalized GC% > 1.0 denote GC-rich regions.

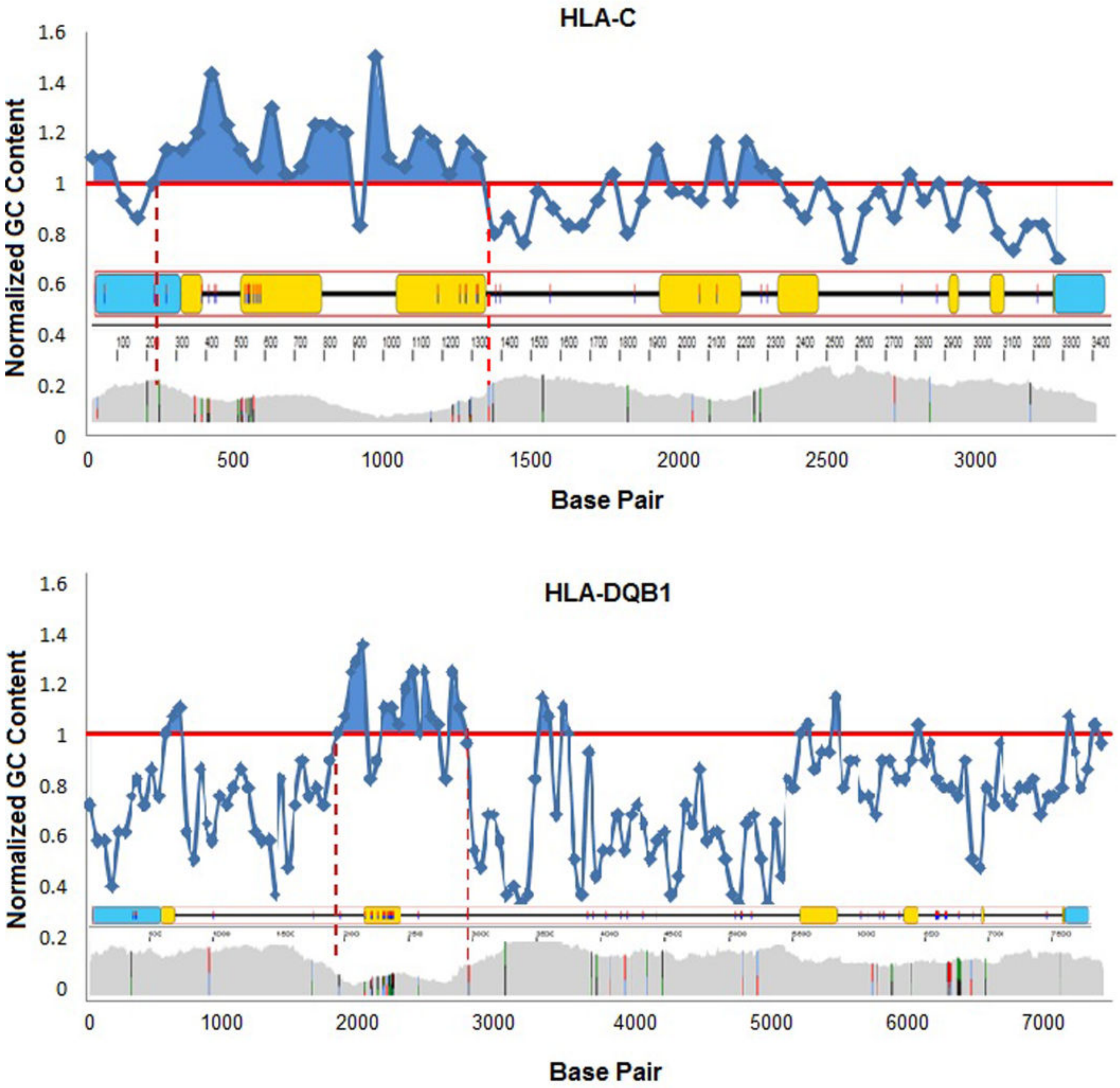


Fig. 4.

GC-rich territories correlate with regions of low depth coverage. Shaded areas represent GC content > 60% (only HLA-C, -DQB1 shown here). Blue box, UTR; yellow box, exon.

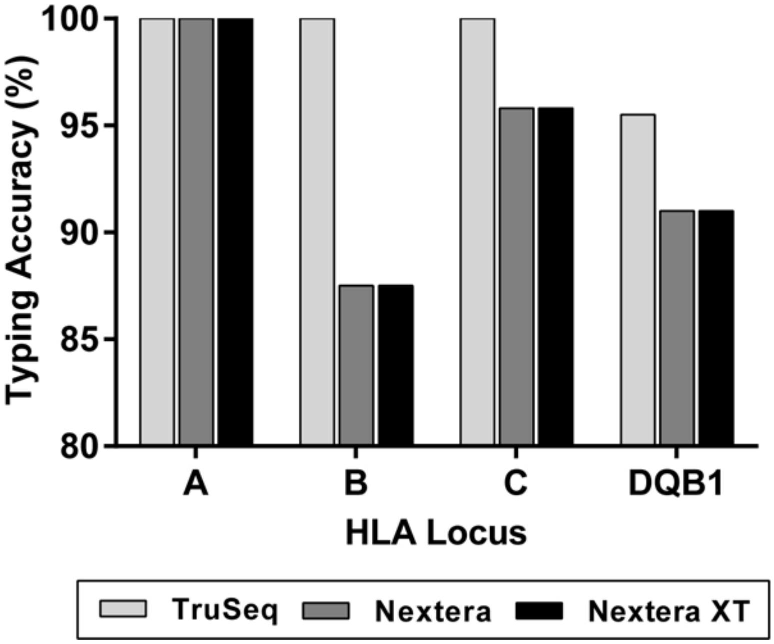
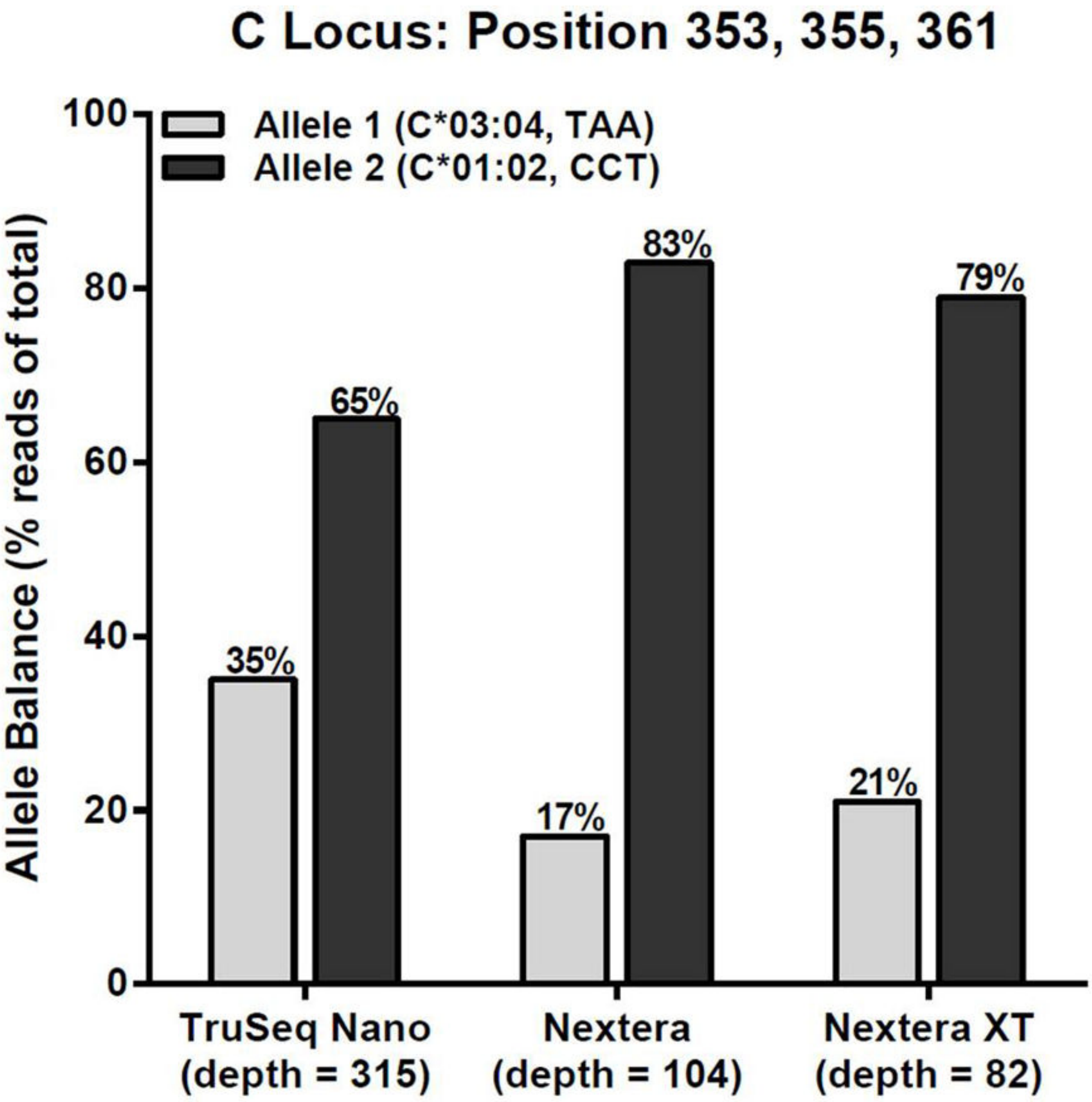


Fig. 5.

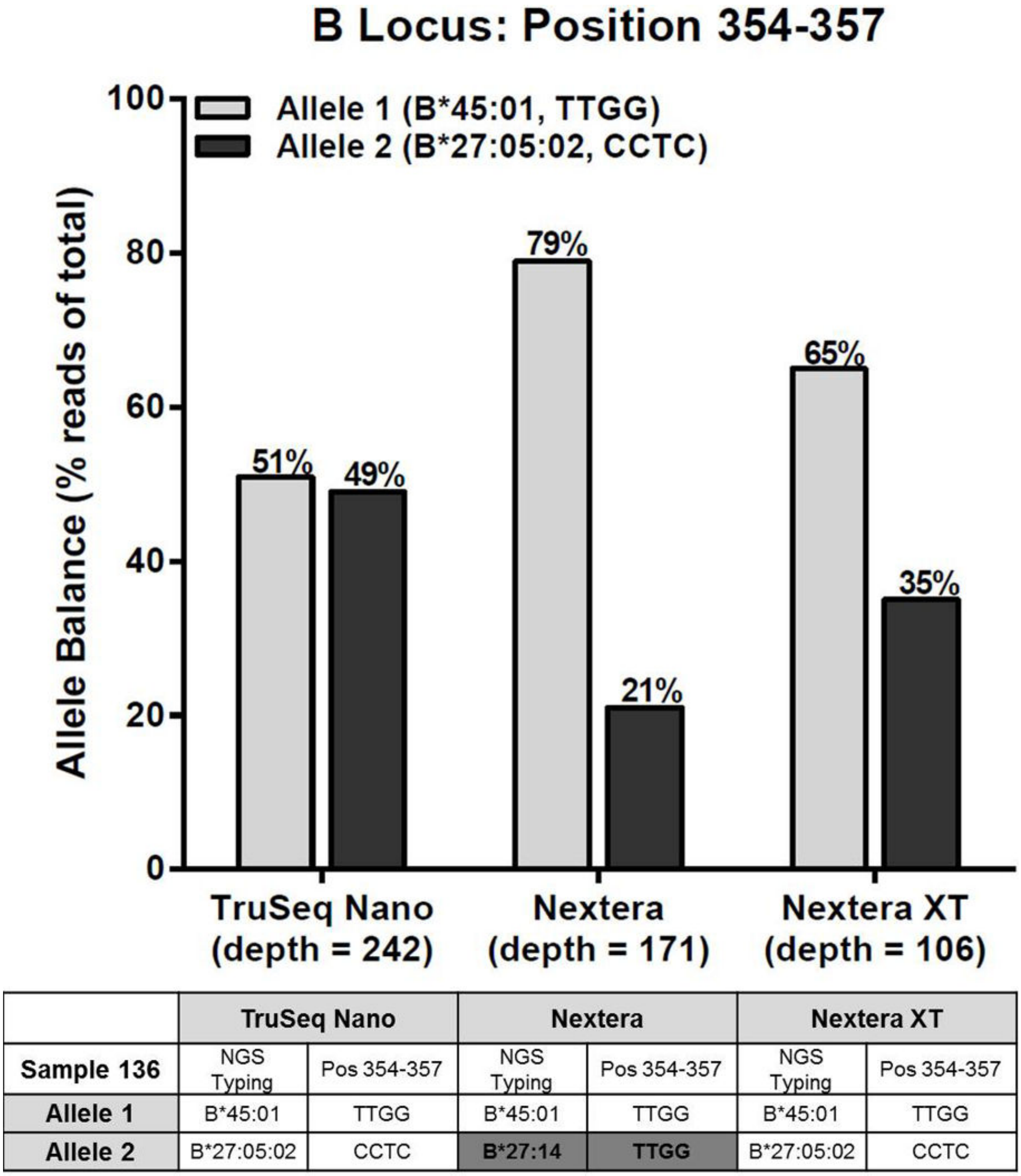
Accuracy of NGS-derived genotype calling at each HLA locus. More typing errors were observed when using transposase-based protocols.



	TruSeq Nano		Nextera		Nextera XT	
Sample 063	NGS Typing	Pos 353, 355, 361	NGS Typing	Pos 353, 355, 361	NGS Typing	Pos 353, 355, 361
Allele 1	C*03:04	TAA	**C*03:17**	**CCT**	**C*03:17**	**CCT**
Allele 2	C*01:02	CCT	C*01:02	CCT	C*01:02	CCT

Fig. 6a.

Allele mis-assignment due to regional read imbalance. In sample 063, allele 1 (C*03:04) was mistyped as C*03:17 in Nextera and Nextera XT. Sequence alignment of C*03:04 and C*03:17 revealed a change in SNP motif TAA (C*03:04) → CCT (C*03:17) at positions 353, 355, 361. Interrogation of read balance at these positions showed an overabundance of CCT reads contributed by allele 2 (C*01:02) relative to the TAA reads of allele 1 (C*03:04) in Nextera and Nextera XT. The over-represented CCT reads were mis-mapped into the construction of the consensus sequence of allele 1 (C*03:04), generating the mistyped hybrid C*03:17 allele.



	TruSeq Nano		Nextera		Nextera XT	
Sample 136	NGS Typing	Pos 354-357	NGS Typing	Pos 354-357	NGS Typing	Pos 354-357
Allele 1	B*45:01	TTGG	B*45:01	TTGG	B*45:01	TTGG
Allele 2	B*27:05:02	CCTC	**B*27:14**	**TTGG**	B*27:05:02	CCTC

Fig. 6b.

Typing discrepancy between transposase-treated samples. The same phenomenon of regional read imbalance as described in Fig. 6a likely contributed to the mistyping of sample 136 in Nextera. This error did not occur in Nextera XT, as the imbalance of TTGG versus CCTC reads was not as skewed as that observed in Nextera.

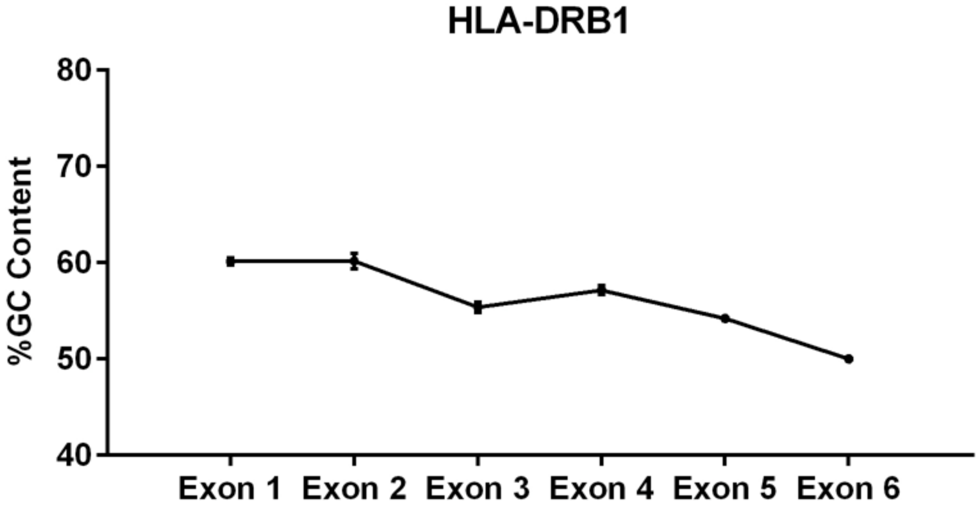


Fig. 7.

DRB1 exon-only GC% plot. The mean GC% of all DRB1 alleles sequenced in this study was calculated for each exon (error bars, standard deviation). GC% was fairly even across all DRB1 exons.

Table 1

Reference HLA genotypes of samples used in this study obtained via Sanger sequencing.

Sample ID	HLA-A	HLA-B	HLA-C	HLA-DRB1	HLA-DQB1
220	24:02 32:01	39:06 40:02	02:02 07:02:01G	14:06 15:01	03:01 06:02
906	02:01 24:02	39:01 40:02	07:02:01G 15:02	01:01 03:01	05:01 02:01
949	24:02 33:03	15:25 46:01	01:02 04:03	09:01 12:02	03:01 03:03
047	23:01:01G 68:03	35:01:01G 45:01	07:02:01G 16:01	04:07:01G 11:02	03:02 03:19
233	02:01 25:01	15:01 51:01	03:03:01G 15:02	11:01 12:01:01G	03:01 03:01
063	02:01 24:02	40:01 56:01	01:02 03:04	04:01 08:01	03:02 04:02
133	03:01 24:02	15:01 35:01:01G	03:03:01G 04:01:01G	01:01 08:01	05:01 04:02
135	23:01:01G 29:02	44:03 45:01	06:02:01G 16:01	07:01 11:01	02:02 06:02
136	02:01 26:01	27:05:02G 45:01	02:02 16:01	04:08 13:03	03:01 03:01
137	01:01 02:01	08:01 15:01	01:02 07:01:01G	01:01 03:01	05:01 02:01
288	02:01 02:02	15:17 41:01	07:01:01G 17:01:01G	04:05 13:02	06:04 02:02
289	02:01 68:02	14:02 35:02	04:01:01G 08:02	01:02 11:04	05:01 03:01

G designates a group of alleles that share the same nucleotide sequence across exons encoding the antigen recognition sites of HLA molecules.

Table 2

NGS typing errors and their proposed etiology. Coverage bias accounted for the majority of typing errors observed in this study. Samples 063 and 136 are used as representative examples to demonstrate the adverse effect of coverage bias on allele mis-assignment in FigS. 6a and 6b and Supplementary Fig. 4.

Method	HLA-locus	Sample ID	Reference by SBT	Mis-assigned by NGS	Cause of Typing Error
TruSeq Nano	DQB1	047	03:02	03:01:01	Primer-related allele imbalance
Nextera	B	136	27:05:02G	27:14	Coverage bias
		906	39:01/40:02	40:184	Coverage bias
		063	56:01/40:01	40:129	Coverage bias
	C	063	03:04	03:17	Coverage bias
	DQB1	047	03:02	03:01:01	Primer-related allele imbalance
		288	06:04	06:09:01	Software
Nextera XT	B	220	39:06	39:01:01	Coverage bias
			40:02	40:06:01	Coverage bias
		906	39:01	39:02:02	Coverage bias
	C	063	03:04	03:17	Coverage bias
	DQB1	047	03:02	03:01:01	Primer-related allele imbalance
		288	06:04	06:09:01	Software

SBT, Sanger sequence-based typing.