NEXT-GENERATION SEQUENCING AND HLA-TYPING:
METHOD VALIDATION USING 1000 GENOMES PUBLIC SAMPLES

Endre Major, Krisztina Rigo, Attila Berces*, Szilveszter Juhos
Omixon Biocomputing – *www.omixon.com*

omixon
Targeted NGS
Data Analysis

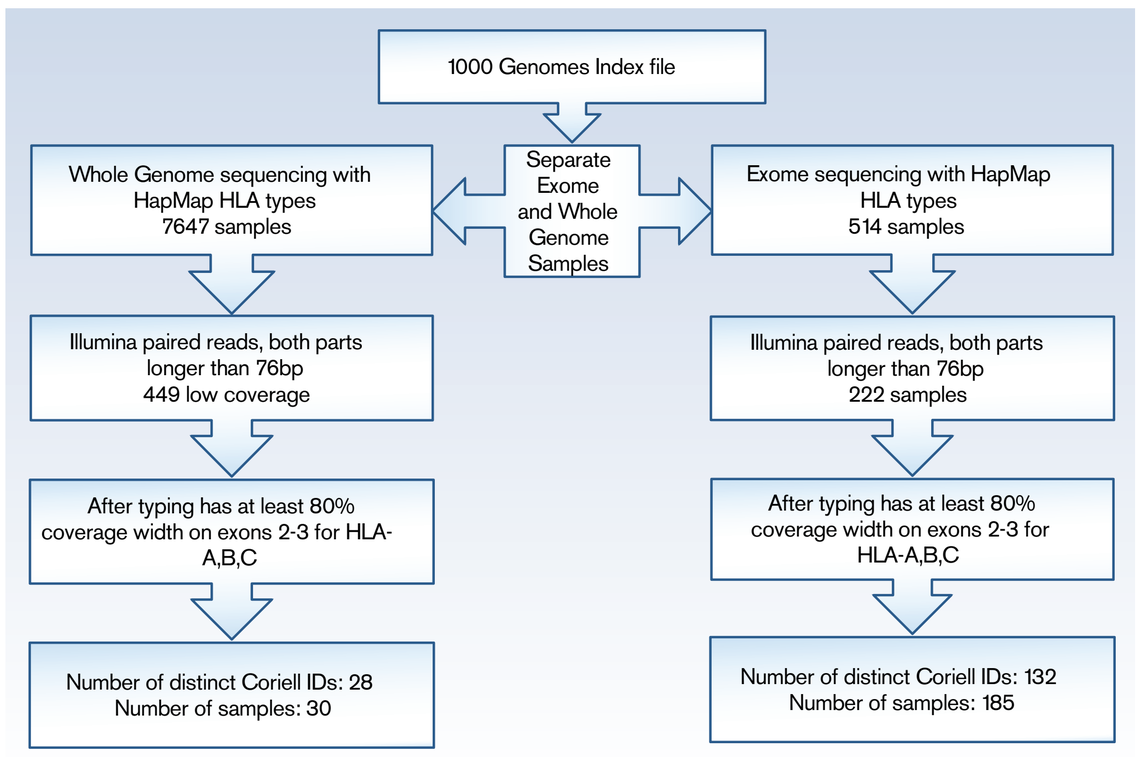
Introduction

Since their first commercialisation, next generation sequencing (NGS) technologies have spread rapidly through molecular genetic laboratories, taking over more and more tasks and overshadowing traditional Sanger sequencing due to their comparatively low cost and fast sample turnaround times. The driving force of these large-scale sequencing studies is hunting for mutations that can be associated to phenotypes, traits and diseases (Kilpinen and Barrett, 2013). The human leukocyte antigen (HLA) genes are also showing significant correlation to the susceptibility for, or the protection from certain diseases. Still, using NGS for HLA typing is somehow lagging behind. The MHC region containing the HLA genes is the most polymorphic part of the human genome and this makes the usual reference-based alignment of NGS reads highly unreliable. In most cases sequence-based typing (SBT) means Sanger sequencing, or relatively long Roche-454 reads. Illumina data with shorter read lengths is rarely used for HLA typing, and Illumina reads of other studies, like whole-genome or whole-exome studies were never used for this purpose. Large-scale sequencing efforts, like the 1000 Genomes (KG) project, are fundamentally aimed at population genetics and were not intended to precisely genotype individuals. However, having the results of the HapMap project, which was using SNP array genotyping techniques, it is possible to compare HLA types obtained using SNP arrays and the corresponding types calculated from the public KG short-read data.

In this poster, we are going to test the feasibility of accurate HLA typing from whole-genome and whole-exome paired Illumina data, concentrating on MHC-I genes HLA-A, HLA-B and HLA-C. We are also presenting current limitations of the HLA typing procedure, and provide simple measures for quality control to achieve maximal accuracy from 1000 Genomes data.

Successful HLA typing

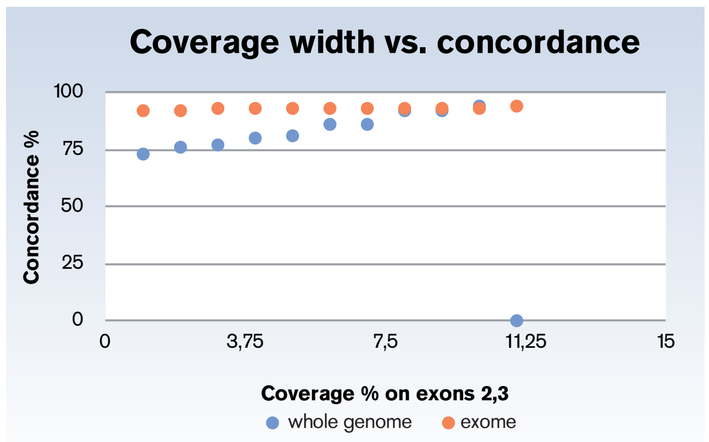
Typical view of a successful result of HLA typing in the genome browser: all the exons have coverage, and the few reads covering the hypervariable regions evenly makes it possible to find the right candidate.



Workflow of sample filtering and typing. The original index file was filtered to find all samples with available HLA types (HapMap samples) and sequenced by Illumina paired technology.

Methods

Paired Illumina reads were selected from the public repository of the 1000 Genomes project. Erlich et al. (2011) conducted a similar validation for these samples using 454 technology, and even corrected some types that were originally determined by sequence specific oligonucleotide (SSO) hybridization. Because having data for these exons is crucial for correct typing, the only samples that passed our quality control (QC) were those where at least 80% of exons 2 and 3 were covered by at least 10% of the average coverage. To speed up typing further, we have pre-filtered all the sets to have only those reads that are mappable to the IMGT/HLA database (Robinson et al., 2013). The filter selected those pairs where at least one of the reads can be aligned with no more than two mismatches and one soft-clip to some allele in the database, and where the orientation of the paired, mapped reads is forward-reverse. Pairs that can be mapped only with indels as mismatches besides nucleotide mismatches were also discarded. Our searching algorithm conceptually is similar to previous studies (Wang et al., 2012) but after alignment we are looking for reads that are able to align to HLA sequences with no or little mismatches. We are than analysing coverage data, depth of and evenness of coverage and percent of exons covered. Using only the CDS nucleotide part of IMGT/HLA the result of a typing experiment is a list of allele pairs, allowing typing to at least four digit but usually to six digit precision.



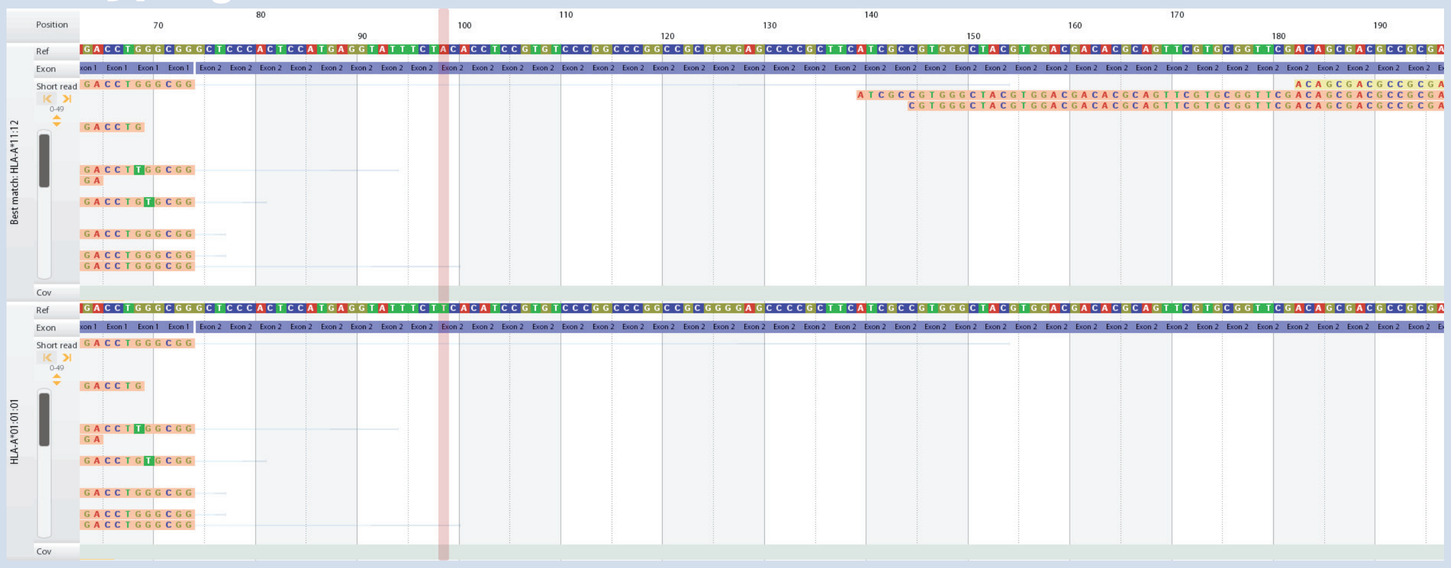
Whole genome samples are generally low coverage, additionally they are not covering exons 2 and 3 evenly. Generally, one can expect good concordance only if more than 80% of these exons are covered by reads.

Results

For the relatively short and highly variable HLA genes coverage depth can be misleading: even if the number of reads covering the gene is high on average, there can be important regions not covered by any reads. For highly concordant samples we would recommend 95% or higher coverage width on exons 2 and 3 of HLA-A, B and C.

The overall concordance for QC-passed samples at 80% coverage width is 86% (±15%) for whole-genome and 93% (±11%) for whole-exome samples. If we increase the coverage width to 95% these values increased to 92% (±11%) and 94% (±10%). These values indicate that reliable HLA typing is possible even from sequencing data that was not originally designed for this purpose.

Mistyping



Mistyping occurs when reads are missing from crucial regions. In this case HLA-A*01:01:01 is the correct type, but no reads are covering the locations of the SNPs differentiating it from HLA-A*11:12

References:

1, 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA - A map of human genome variation from population-scale sequencing. Nature, 2010; 467:1061-1073.
2, Erlich RL, Jia X, Anderson S, Banks E, Gao X, Carrington M, Gupta N, DePristo MA, Henn MR, Lennon NJ, de Bakker PI - Next-generation sequencing for HLA typing of class I loci. BMC Genomics, 2011; 12:42.
3, International HapMap Consortium - The International HapMap Project. Nature, 2003; 426:789-796.
4, Kilpinen H, Barrett JC - How next-generation sequencing is transforming complex disease genetics. Trends Genet, 2013; 29:23-30.
5, MHC Consortium - Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. Nature, 1999; 401:921-923.
6, Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SGE - The IMGT/HLA Database. Nucleic Acids Res, 2013; 41:D1222-1227
7, Wang C, Krishnakumar S, Wilhelmy J, Babrzadeh F, Stepanyan L, Su LF, Levinson D, Fernandez-Viña MA, Davis RW, Davis MM, Mindrinos M - High-throughput, high-fidelity HLA genotyping with deep sequencing. Proc Natl Acad Sci USA, 2012; 109:8676-8681.

Contact:

Email: attila.berces@omixon.com

Address: Petzval J. utca 56, Budapest, H-1119, Hungary