kTypeR – accurate and efficient alignment-free HLA genotyping using nanopore sequencing

Steffen Klasberg[1], Kathrin Putke[1], Markus Fuhrmann, Vineeth Surendranath[1], Alexander H Schmidt[1,2], Vinzenz Lange[1], Gerhard Schöfl[1]

[1]DKMS Life Science Lab, St. Petersburger Str. 2, 01069 Dresden, Germany;
[2]DKMS, Kressbach 1, 72072 Tübingen, Germany

Introduction

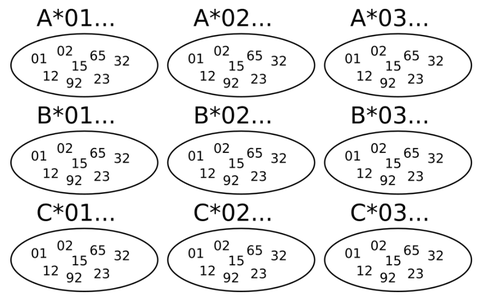
Attempts to leverage nanopore sequence data for HLA genotyping are rapidly gaining traction in the HLA community as the technology promises easy and cost-effective library preparation and reads that cover the full extent of HLA Class I and Class II genes. Current typing strategies using nanopore sequences all rely on classical alignment-based methods. Due to the high per-read error-rates of nanopore reads, these algorithms typically perform considerably less robustly than when applied to short-reads. Here, we present ktypeR, which uses a fundamentally different strategy for genotyping. KtypeR compares all k-mers of each individual read to the pre-processed IPD-IMGT/HLA database and infers the most likely true alleles by majority vote.

K-merize sequences

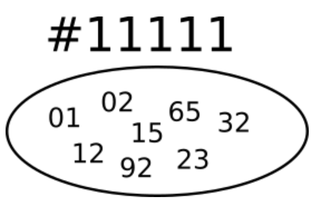
Each k-mer (we use k = 31) in a sequence is inferred in a sliding window. The nucleotide sequence is hashed to a 64 bit integer. The reverse complement of each sequence is considered identical, as we do not know the direction of a read. The smallest of both hashes represents the k-mer.

Prepare the database

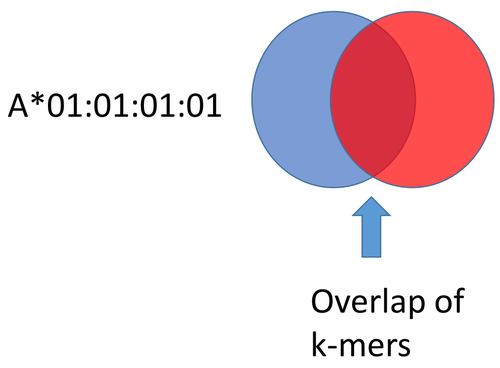
The IPD-IMGT/HLA database contains all known HLA allele. As some alleles are only known partially, we impute the missing sequence using the most similar complete allele. All 31-mers are inferred, hashed and the unique k-mer content of each allele is collected for later comparisons.



Genotype reads



K-merize, hash and collect unique k-mers of each read.



Infer the overlap of the read k-mers and the k-mers of each allele.

Allele	overlap
A*01:01:01:01	348
A*01:01:01:03	344
A*01:01:01:05	339
A*03:01:01:01	301

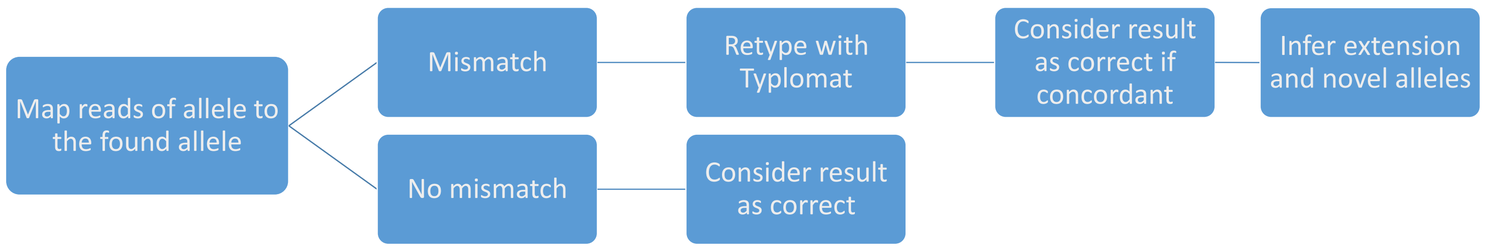
Per Read: Take alleles within a threshold **t** as potential results. Here, A*01:01:01:01 and A*01:01:01:03 with a threshold of max(overlap) – 4.

Allele	count
A*01:01:01:03	3
A*32:01:01:01	2
A*01:01:01:01	1
A*01:01:01:02	1

Per Locus: Keep the two best mutually exclusive alleles as result. Here, A*01:01:01:03 and A*32:01:01:01.

Post-processing

Check if the genotyping results are plausible and reliable by mapping the reads supporting each inferred allele to their reference. An allele is considered correct if no mismatch is found. If a mismatch occurs, perform a second round of genotyping using mapping and our in-house genotyping tool *Typlomat.* An allele is considered correct if both genotypings are concordant. Novel and extended alleles can be validated by inspecting the mapping in a last step.



Results

We developed kTypeR using three sequencing runs (Minion, R10 pore) with a total of 384 samples, each containing the six major HLA loci. A maximum of 243 samples are multiplexed using one barcode per sample. kTypeR is multi-threaded and can process 10,000 reads in 5 minutes on 8 cores. Post-processing, if necessary, takes 1 to several minutes per sample, depending on the mismatch status. We achieve an outstanding accuracy of > 98 % alleles correctly genotyped in the 4th field using the R10 pore. On the same data set, the graph alignment genotyper, HLA*LA [Dilthey et al. 2019], and a commercial solution perform considerably worse, with a 4-field accuracy of ~ 93 % and ~ 78 % alleles correctly genotyped, respectively. A first glance at data from the new R10.3 pore promise a 4-field accuracy of up to 100 %. As a conclusion, kTypeR shows great potential to outperform existing classical tools in speed and accuracy and demonstrated the feasibility of using nanopore data for mid- and high-throughput HLA genotyping at full resolution.

DKMS Life Science Lab

GitHub ***Download from GitHub:*** *https://github.com/DKMS-LSL*

DKMS Life Science Lab

St. Petersburger Str. 2 • 01069 Dresden • www.dkms-lab.de