Layered Genotyping for Efficient High-Throughput High-Accuracy Typing with Nanopore Data

Steffen Klasberg[1], Kathrin Putke[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

Although long-read single-cell sequencing technologies, like Oxford Nanopore Technologies' (ONT) nanopore sequencing, continue to mature, they are not yet truly challenging short-read NGS in the sequence-based HLA genotyping space. Despite their read length advantages, the read quality often does not suffice for conventional, mapping-based approaches to genotyping. Here, we show that substantial improvements in long-read genotyping quality and efficacy can be achieved by a layered approach to genotyping, where we leverage alignment-free *k*-mer-based probabilistic classification and statistical learning in addition to exact sequence alignment.

Workflow

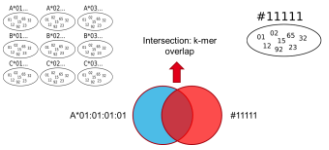

Figure 1: Principle of `kTypeR` Each allele in the IPD-IMGT/HLA database is assigned a set of unique *k*-mers as well as each read in a sample. The intersection of *k*-mer sets between each read and each allele results in candidate alleles for each sample and each locus.

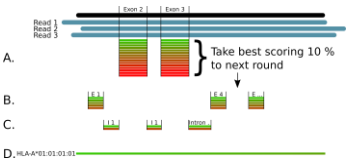

Figure 2: `Typlomat` The allele-specific reads are mapped against the resulting candidate allele. This mapping is hierarchically scored against the IPD-IMGT/HLA database.

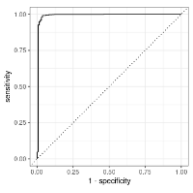

Figure 3: ROC curve of the random forest model of `bonsai` predicting genotyping accuracy. Different statistics of `kTypeR` and `Typlomat` serve as an input to this model.

Figure 4: Screenshot of our app for manual inspection. Parameters for a decision are, e.g. the number of reads supporting an allele, `bonsai` scores, mismatches to the reference sequence or the mapping of allele-specific reads to its reference.

Conclusion

Our genotyping pipeline was developed and tested with most nanopore chemistries and pores, i.e. with R9.4, R10, R10.3 and R10.4. The best results can be obtained using the latest R10.4 chemistry. Our results show that ONT's nanopore sequencing can be used for high-accuracy genotyping of HLA and other relevant genes to the highest resolution. The prediction of correctness of results, either single alleles or zygosity of loci, reduces the need for manual inspection to a minimum. The models are also expected to improve as mora data will accrue over time. We suggest that routine application of nanopore sequence data for high-throughput HLA genotyping will become feasible once the workflow is stabilized to yield constant and comparable results.

Methods

Our layered approach to HLA genotyping uses three progressively applied procedures. A first layer is `kTypeR`, a *k*-mer based genotyping algorithm. In a first step, each read of a multiplexed sample is allocated to its locus of origin, after which all reads of a locus are processed to find the most reliable candidate alleles. During this step sets of allele-specific reads are collected which serve as an input to a second layer. `Typlomat` utilizes iterative mapping of allele-specific reads to the sequences of potential candidate alleles. The final mapping, represented as a PWM, is the target of an hierarchical scoring against all possible candidate alleles. Output data from both, `kTypeR` and `Typlomat`, serve as input to the final layer, the evaluation of results using `bonsai`. This evaluation is carried out using two different random forest models, the first for the prediction of correctness of each found allele, the second for the prediction of correctly assigned homozygosity. Finally, each allele with a predicted correctness below a defined threshold, as well as loci with more then two potential results, need to be inspected manually.

Results

A test run with 18 multiplexed loci (7 HLA class I, 8 HLA class II, 2 MIC, 3 bloodgroup, CCR5) of 42 samples results in an accuracy of 100 % correctly assigned alleles. Six cases of false-positive homozygous loci were found, i.e. the partner allele was missing.

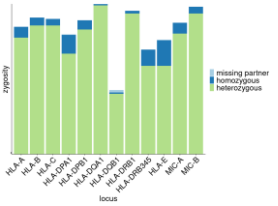

Figure 5: Zygosity per locus. Six loci were falsely claimed homozygous (4 HLA-DQB1, 1 HLA-DPA1, 1 HLA-E).

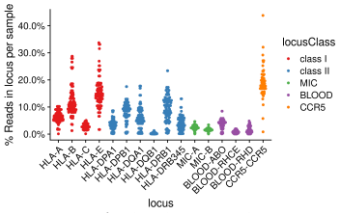

Figure 6: Fraction of reads per locus per sample. .

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