Graph-guided-assembly and typing of the HLA alleles from RNA-seq data

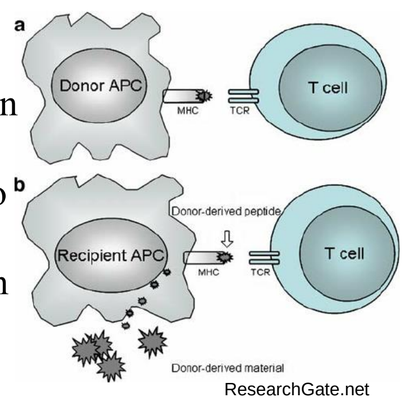
Shayna Mallett, Computer Science
Mentor: Dr. Heewook Lee
School of Computing and Augmented Intelligence
Biodesign's Center for Biocomputing, Security, and Society

Abstract

The HLA, Human Leukocyte Antigens, are encoded by a polymorphic set of genes where even a single base change can impact the function of the body's immune response to foreign antigens [1]. Although many methods exist to type these alleles using whole-genome sequencing (WGS), few can use RNA sequencing (RNA-seq) to show the functional expression of the alleles with its inconsistency in coverage, and none of these allow for novel allele discovery. We present an approach using partially ordered graphs to project sequenced data onto the known alleles allowing for accurate and efficient typing of the HLA genes with flexibility for discovering new alleles and tolerance for poor sequence quality. This graph-guided approach to assembling and typing the HLA genes from RNA-seq has applications throughout precision medicine, facilitating the prevention and treatment of autoimmune diseases where allele expression can change. It is also a necessary step for determining donors for organ transplants with the least likelihood of rejection. This novel approach of combining database matching with partially ordered graphs for assembling genetic sequences of RNA-seq data could be applied towards typing other alleles.

Significance of the HLA genes

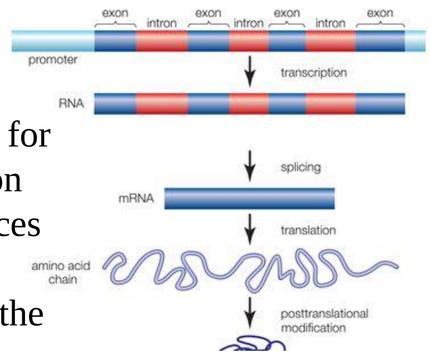
HLA genes encode the MHC for function of the adaptive immune system in self-recognition. The pathogen driven selection on these genes creates a bias towards diversity in pathogen recognition and consequently polymorphism in the sequence [3]. Matching alleles between organ donor and recipient reduces the risk of rejection [1]. Allele type can also indicate a predisposition to autoimmune diseases and expression has been shown to change in some cancer cells [4].



ResearchGate.net

Why RNA-seq?

RNA-seq captures changes in the HLA alleles that can occur from transcribing the DNA sequence into RNA. However, it has uneven coverage of genes capturing only the transcripts expressed at the time of sequencing, and much higher error rates resulting in noisy data. Paired end reads are used for the connecting information between different sequences that it provides when two distinct reads come from the same molecule.



Britanicca.com

Methods and Results

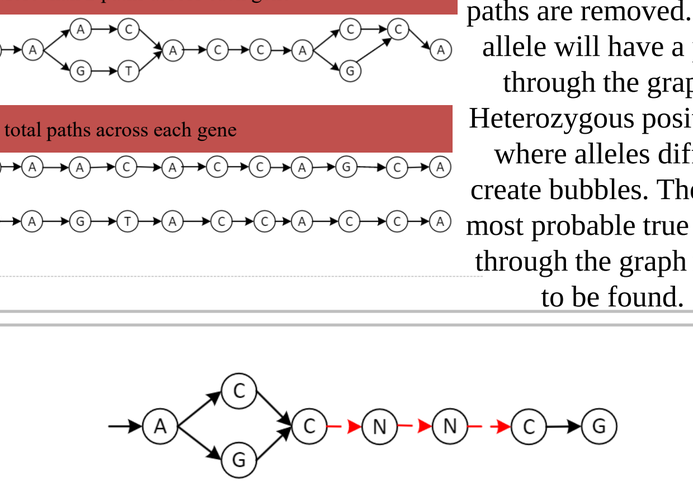


Figure 1. Database sequences of the known HLA genes are used to construct a graph representing all possible alleles. The exon junction sites for each allele are used in splice aware alignment with STAR of RNA-seq data to a reference genome then the panel of HLA genes. Finally, the aligned reads are projected onto the graph and the most probable pair of alleles for each gene is found.

Figure 2. A graph is constructed representing every sequence in the MSA as a path. Reads are aligned and unused paths are removed. Each allele will have a path through the graph. Heterozygous positions, where alleles differ, create bubbles. The two most probable true paths through the graph need to be found.

Figure 3. Portions of the graph that lack coverage due to changing expression of mRNA are padded to allow the sequenced portions to connect while leaving the unsequenced sections ambiguous for allele calling.

Column Noise Processing

$$argmax_{G^i} \prod_{n \in N^i} \prod_{r \in R_n} \left[\frac{P(r \mid A_1^i) + P(r \mid A_2^i)}{2}\right]$$

$$P(r \mid A^i) \begin{cases} (1-\epsilon_r) & Match : n = A^i \\ (\frac{\epsilon_r}{3}) & Mismatch : n! = A^i \\ \frac{1}{4} & N \end{cases}$$

Figure 4. To find the likelihood of the observed data given a column genotype(G_i), the joint probability of every read ($r \in R_n$) in each observed node($n \in N^i$) is found. Each read is equally likely to come from either haplotype (A_1^i, A_2^i) and the average probability of the read given a haplotype, which is a single base in a column is found from the sequencing and alignment error probability (ϵ_r).

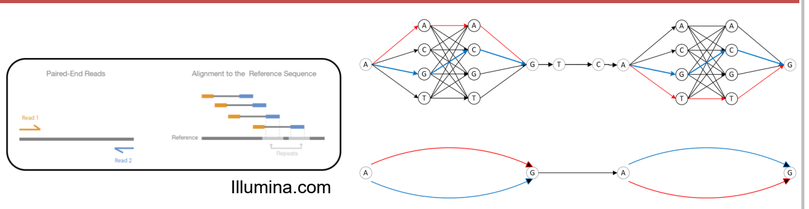
Bubble Processing

$$argmax_{G_b} \prod_{r \in R_b} \prod_{i}^{|b|} \left[\frac{P(r^i \mid A_{b1}^i) + P(r^i \mid A_{b2}^i)}{2}\right]$$

$$P(r^i | A_b^i) \begin{cases} (1-\epsilon_{r^i}) & Match : r^i = A^i \\ (\frac{\epsilon_{r^i}}{3}) & Mismatch : r^i! = A^i \\ \frac{1}{4} & N \end{cases}$$

Figure 5. To find the likelihood of the observed data given a bubble genotype (G_b), the joint probability of every position (i) on every observed read across the bubble ($r \in R_b$) is found. Each read is equally likely to come from either haplotype (A_1^i, A_2^i) and the average probability of the read given a haplotype, is found from the sequencing and alignment error probability (ϵ_r).

Bubble Merging

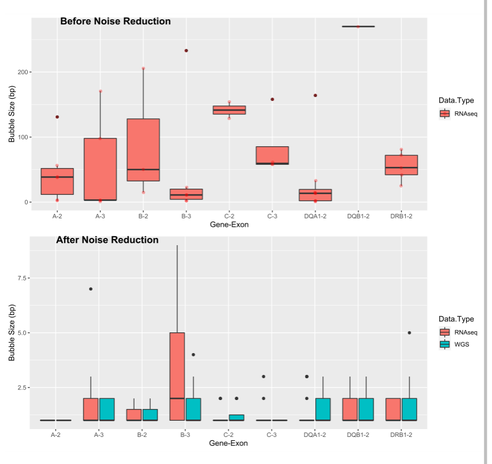


Illumina.com

Figure 6. A single contiguous read won't be able to span multiple bubbles. Paired end reads come from different ends of the same molecule, and therefore the same haplotype. Paths between bubbles are combined with the highest amount of phasing.

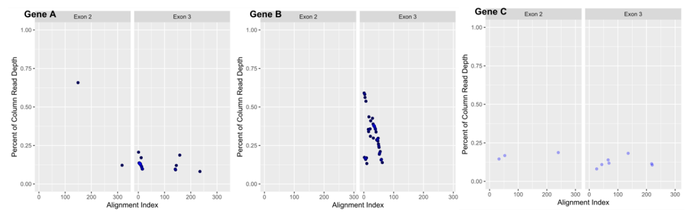
Noise Reduction

Figure 7. Before noise reduction there are intractable bubble sizes with a maximum size of 270bp. After noise reduction the bubble sizes are similar to the true alleles with a maximum size of 9bp.



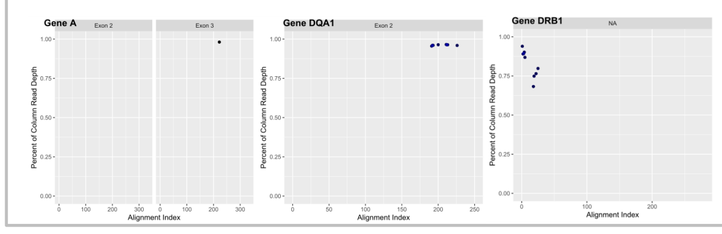
Added Heterozygosity

Figure 8. True allele is homozygous but the algorithm decides on two bases. These can be removed in further processing or prevented by addressing alignment and projection errors.



Missed Heterozygosity

Figure 9. True allele is heterozygous but the algorithm decides on a single most probable base. This type of error can't be fixed once removed but could be prevented in preprocessing Class II read filtering errors.



Future Work

Testing

Real data

Using real data with known HLA alleles we can test the accuracy of the algorithm against a variety of qualities of data

Simulated data

Simulating mRNA reads from a full genome we can further test the accuracy of the algorithm since the genome sequence is known.

We can also test the algorithm's ability to call novel alleles by removing the true allele of a sample from the database

References

(1) Gao, X.; Nelson, G. W.; Karacki, P.; Martin, M. P.; Phair, J.; Kaslow, R.; Goedert, J. J.; Buchbinder, S.; Hoots, K.; Vlahov, D.; O'Brien, S. J.; Carrington, M. New England Journal of Medicine 2001, 344, Publisher: Massachusetts Medical Society eprint: https://doi.org/10.1056/NEJM200105313442203, 1668–1675.
(2) Bravo-Egana, V.; Sanders, H.; Chitnis, N. Human Immunology 2021, 82, 478–487.
(3) Prugnolle, F.; Manica, A.; Charpentier, M.; Gu´egan, J. F.; Guernier, V.; Balloux, F. Current Biology 2005, 15, Publisher: Elsevier, 1022–1027.
(4) McGranahan, N. et al. Cell 2017, 171, Publisher: Elsevier, 1259–1271.e11.
(5) Lee, H.; Kingsford, C. Genome Biology 2018, 19, 16.
(6) Stark, R.; Grzelak, M.; Hadfield, J. Nature Reviews Genetics 2019, 20, Number: 11 Publisher: Nature Publishing Group, 631–656.
(7) Orenbuch, R.; Filip, I.; Comito, D.; Shaman, J.; Pe'er, I.; Rabadan, R. Bioinformatics 2020, 36, 33–40.
(8) Boegel, S.; L¨ower, M.; Sch¨afer, M.; Bukur, T.; de Graaf, J.; Boisgu´erin, V.; T¨ureci, O.; Diken, M.; Castle, J. C.; Sahin, U. Genome Medicine 2012, 4, 102.
(9) Buchkovich, M. L.; Brown, C. C.; Robasky, K.; Chai, S.; Westfall, S.; Vincent, B. G.; Weimer, E. T.; Powers, J. G. Genome Medicine 2017, 9, 86.
(10) Yamamoto, F.; Suzuki, S.; Mizutani, A.; Shigenari, A.; Ito, S.; Kametani, Y.; Kato, S.; Fernandez-Vi˜na, M.; Murata, M.; Morishima, S.; Morishima, Y.; Tanaka M.; Kulski, J. K.; Bahram, S.; Shiina, T. Frontiers in Immunology 2020, 11.
(11) HLA Nomenclature @ hla.alleles.org.
(12) Auton, A. et al. Nature 2015, 526, Number: 7571 Publisher: Nature Publishing Group, 68–74.
(13) Szolek, A.; Schubert, B.; Mohr, C.; Sturm, M.; Feldhahn, M.; Kohlbacher, O. Bioinformatics 2014, 30, 3310–3316.
(14) Grinnemo, K.-H.; Sylv´en, C.; Hovatta, O.; Dellgren, G.; Corbascio, M. Cell and tissue research 2008, 331, 67–78. 13

Acknowledgements

Many thanks to Dr. Lee for his guidance on the project

FURI

Ira A. Fulton Schools of Engineering
Arizona State University