High Throughput HLA typing in the Cloud: How to make it rain HLA Genotypes

Ariella Sasson, Efi Melista, & Ming-shan Chien

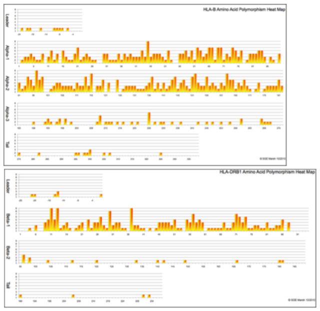
Abstract

The human leukocyte antigens (HLAs), part of the major histocompatibility complex (MHC), contain the most polymorphic genes in the human genome. The classical HLA class I and II genes define the specificity of adaptive immune responses. Genetic variation at the HLA genes is associated with susceptibility to autoimmune and infectious diseases and plays a major role in transplantation medicine and immunology. In addition to the "traditional" use of HLA genes in matching donors and recipients in transplantation, numerous HLA alleles have been associated with disease and implicated in tumor immunogenicity. Understanding the exact molecular mechanisms by which the HLA molecules are involved in the pathophysiology of these diseases remains challenging. Up until recently, HLA genotyping was performed using low resolution techniques, serology and molecular-based methods; the introduction of high-resolution methodology, such as next-generation sequencing (NGS), revolutionized the characterization of the MHC and generated an explosion of new HLA alleles. However, the highly complex genomic nature of the MHC rejects the assumptions made by traditional NGS bioinformatic algorithms. Consequently, the challenge of this MHC/HLA bioinformatic problem has impeded the exploration of large scale NGS disease cohort data in relation to this region. De novo assembly of MHC is ideal for precisely describing this region but is also problematic. A set of bioinformatic algorithms specifically tuned to the strengths and limitations of NGS data and the polymorphic nature of the MHC, we can genotype an individual's HLA genes effectively and easily from whole-exome or whole-genome sequencing. Using these algorithms, we have developed a solution for high throughput HLA typing solution using AWS batch that is time and cost-effective and will allow us to type historical NGS data for this valuable information. HLA genotyping is a critical piece required to benefit fully from personalized medicine approaches and further our understanding of immune response.

Background: Complexity of the MHC

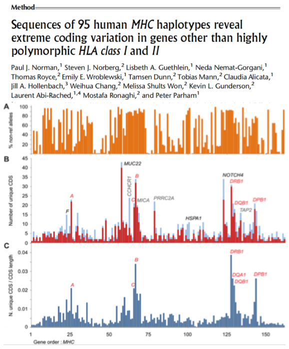
The MHC is the most diverse area of the Human Genome. Due to the diversity, this region breaks the traditional assumptions which underpin NGS bioinformatic pipelines. While exemplified by the HLA genes, this diversity is not limited to the HLA genes. Currently, there are more than 25000 HLA alleles in the IMGT/HLA database and the number is growing exponentially.

HLA Polymorphism (HLA-B and DRB1)



http://hla.alleles.org/alleles/heat_maps.html

Polymorphism outside the MHC



Method

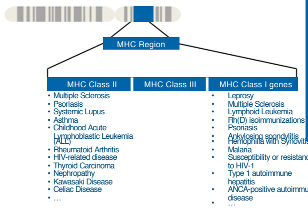
Sequences of 95 human *MHC* haplotypes reveal extreme coding variation in genes other than highly polymorphic *HLA class I* and *II*

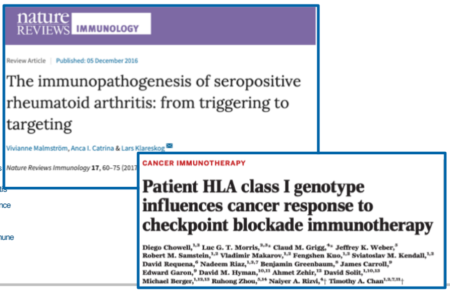
Paul J. Norman,[1] Steven J. Norberg,[2] Lisbeth A. Guethlein,[1] Neda Nemat-Gorgani,[1] Thomas Royce,[2] Emily E. Wroblewski,[1] Tamsen Dunn,[2] Tobias Mann,[2] Claudia Alicata,[1] Jill A. Hollenbach,[3] Weihua Chang,[2] Melissa Shults Won,[2] Kevin L. Gunderson,[2] Laurent Abi-Rached,[1,4] Mostafa Ronaghi,[2] and Peter Parham[1]

Why is this interesting?

"In studying the genetics of human disease, the major histocompatibility complex (MHC) region is arguably the most important part of the genome"
-Norman *et al. 2017*

The MHC and specifically the HLA genes have been associated with many diseases. There is evidence that the HLA genotypes can effect response potentially allowing for a biomarker to help identify patients that will respond positively to treatment.





nature REVIEWS IMMUNOLOGY

Review Article | Published: 05 December 2016

The immunopathogenesis of seropositive rheumatoid arthritis: from triggering to targeting

Vivianne Malmström, Anca I. Catrina & Lars Klareskog

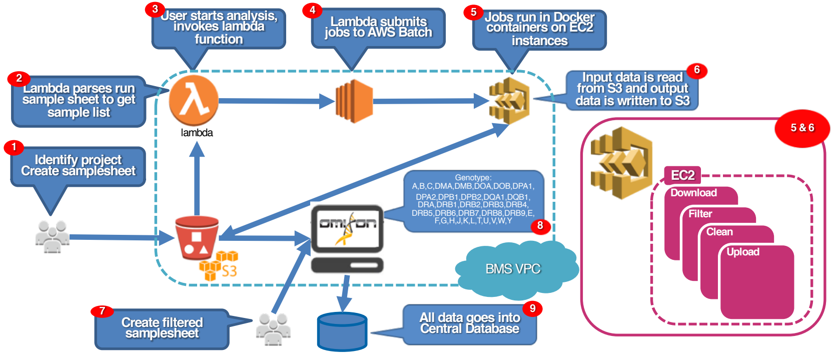
Nature Reviews Immunology **17**, 60–75 (2017

CANCER IMMUNOTHERAPY

Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy

Diego Chowell,[1,2] Luc G. T. Morris,[2,3*] Claud M. Grigg,[4*] Jeffrey K. Weber,[5] Robert M. Samstein,[1,2] Vladimir Makarov,[1,2] Fengshen Kuo,[1,2] Sviatoslav M. Kendall,[1,2] David Requena,[6] Nadeem Riaz,[1,2,7] Benjamin Greenbaum,[8] James Carroll,[9] Edward Garon,[9] David M. Hyman,[10,11] Ahmet Zehir,[12] David Solit,[1,10,13] Michael Berger,[1,12,13] Ruhong Zhou,[5,14] Naiyer A. Rizvi,[4†] Timothy A. Chan[1,2,7,11†]

What We Built – AWS batch


1 Identify project Create samplesheet
2 Lambda parses run sample sheet to get sample list
3 User starts analysis, invokes lambda function
4 Lambda submits jobs to AWS Batch
5 Jobs run in Docker containers on EC2 instances
6 Input data is read from S3 and output data is written to S3
lambda
S3
Genotype: A,B,C,DMA,DMB,DOA,DOB,DPA1, DPA2,DPB1,DPB2,DQA1,DQB1, DRA,DRB1,DRB2,DRB3,DRB4, DRB5,DRB6,DRB7,DRB8,DRB9,E, F,G,H,J,K,L,T,U,V,W,Y
8
BMS VPC
7 Create filtered samplesheet
9 All data goes into Central Database
5 & 6
EC2
Download
Filter
Clean
Upload

Caveats and Performance

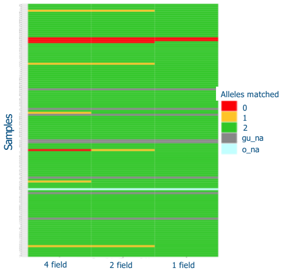
Highest accuracy (4 fields) can be achieved using targeted panels for high depth and completeness of the gene with the longest reads possible. WGS can achieve similar results, but since we don't always have access to samples, we have no other option but to sacrifice on result resolution (2/3 fields).

Caveats Exome/RNASeq

- Possible genotype ambiguity from only considering exons
- Lower certainty in gene assembly
- Off-the-shelf software not compatible with HLA-type changes from tumor mutations

DRB1 Orthogonal Validation

- Samples of interest included Omixon HLA genotyping from DNA WES Tumor & Normal & RNASeq in addition to orthogonal HLA genotyping via NGS targeted sequencing
- 129 patients → 9 samples removed (either one source missing)
- 120 samples:
  - 113 samples (94.2%) perfectly matched at 3 fields
  - 4 samples (3.3%): systematic mismatch of 1 allele (Omixon, 14:54 ~ 14:01)
  - 3 samples (2.5%): From WGS data, we figured that the Tumor and normal samples have been swapped and are not from the same person.

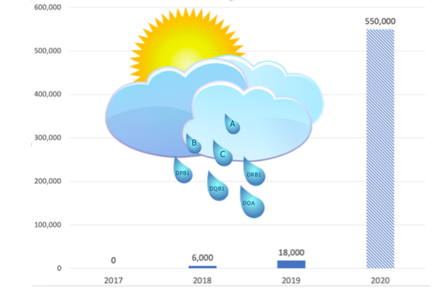


Correctly called 236 of the 240 possible DRB1 alleles (98.3% alleles, 96.6% samples).

What's Next

Implement & optimize a fully automated pipeline

Process data!!!!



Acknowledgements

BMS
- Natallia Kalinava
- Ming-shan Chien
- Steven Vasquez Grinnell
- Sujaya Srinivasan
- Simon Papillon-Cavanagh
- Ryan Golhar
- Carlos Rios
- Stefan Kirov
- Isaac Neuhaus
- Joseph Szustakowski

Omixon
- Peter Meintjes
- Efi Melista
- Nandor Varga

WHO ARE YOU WORKING FOR?

Bristol-Myers Squibb