PhD thesis

Hongzhi Cao

MHC region and its related disease study
Method for decipher the polymorphism and fine-mapping disease causal mutations of MHC region

Academic advisor:
Anders Krogh
University of Copenhagen, Denmark

Jun Wang
University of Copenhagen, Denmark
BGI-Shenzhen, China

The PhD School of Science, Faculty of Science, University of Copenhagen
Submitted: 23/03/2015
Dissertation for the degree of philosophiae doctor (PhD)

Department of Biology, University of Copenhagen
Copenhagen, Denmark

And

BGI-Shenzhen
Shenzhen, China
March 2015

Author: Hongzhi Cao

Title: MHC region and its related disease study
Subtitle: Method for decipher the polymorphism and fine-mapping disease causal mutations of MHC region

Academic advisors:
Anders Krogh
University of Copenhagen, Denmark
Jun Wang
University of Copenhagen, Denmark
BGI-Shenzhen, China
The PhD School of Science, Faculty of Science, University of Copenhagen
Submitted: March 2015
PREFACE

This PhD project started in 2011 as collaboration between the Department of Biology, University of Copenhagen and BGI-Shenzhen. The work presented here has been performed at both institutions by supervision of Prof. Anders Krogh and Prof. Jun Wang.
ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who has helped me during my time as a PhD student. First of all, I would like to thank my supervisors Prof. Anders Krogh and Prof Jun Wang for giving me this opportunity and introducing me to the research field within human genomics and genetics, exploring the mystery of the most complex genomic region – Major histocompatibility complex (MHC), and making the time of my PhD exciting and interesting, in addition to all help, guidance and motivation.

Thanks to my colleagues, friends at BGI and University of Copenhagen for all the help, guidance, collaborations and making me feel very easy and pleasure to work and study here.

I wish to thank all my co-authors, without your helps, comments and guidance with data analysis, interpretation of results and writing, I won’t make it so smooth.

Last, but not least, thanks to the endless support from my family, my gentle and considerate wife Yijie, the kindness and great parents of my wife’s and mine. You are my source of power and happiness. Mostly, to my little princess Little September/小九. Your appearance is the happiest thing of my life and makes my life much more wonderful.
ABSTRACT

The major histocompatibility complex (MHC) is one of the most gene dense regions in the human genome and many disorders, including primary immune deficiencies, autoimmune conditions, infections, cancers and mental disorder have been found to be associated with this region. However, due to a high degree of polymorphisms within the MHC, the detection of variants in this region, and diagnostic HLA typing, still lacks a coherent, standardized, cost effective and high coverage protocol of clinical quality and reliability. And owing to marked linkage disequilibrium of MHC region, genes/mutations involved in the above diseases have not, with very few exceptions, been identified. Currently popular next-generation sequencing (NGS) technology, comprising massively parallel single-molecule sequencing, can generate billions of bases from amplified single DNA molecules within several days, holding great promise for achieving cost-effective and high-throughput and high accurate genotyping. Based on the NGS platform, by using delicately designed bioinformatics methods, we systematically developed new tools/methods, including high throughput HLA genotyping method – RCHSBT, targeted sequencing based SNV detection as well as HLA gene typing and large structural variation detection using optical mapping technic, to provide comprehensive and accurate information of the MHC region and apply them into disease causal mutation’s fine-mapping.
**Sammendrag**

Table of Contents

1 Introduction .................................................................................................................................................. 10
  1.1 Major histocompatibility complex ........................................................................................................ 10
    1.1.1 Function of MHC region gene in human immune system ................................................................. 11
    1.1.2 MHC related diseases ....................................................................................................................... 11
  1.2 High throughput sequencing technology ............................................................................................. 18
    1.2.1 Roche 454 ........................................................................................................................................ 19
    1.2.2 Illumina Solexa/Hiseq/Miseq ........................................................................................................... 19
    1.2.3 Life Technologies Ion Torrent PGM/Proton ..................................................................................... 20
    1.2.4 PacBio ............................................................................................................................................. 20
    1.2.6 Oxford Nanopore ............................................................................................................................. 20
    1.2.7 Others ............................................................................................................................................ 21
  2 Aims ............................................................................................................................................................. 22
    2.1 Current problem for MHC region related study ....................................................................................... 22
    2.2 Specific aims of this project ................................................................................................................... 22
  3 Results ....................................................................................................................................................... 23
    3.1 Novel method for MHC region study ..................................................................................................... 23
      3.1.1 Paper I: High throughput HLA typing Method ............................................................................... 23
      3.1.2 Paper II: Integrated tools to study whole MHC region .................................................................... 25
    3.2 Construction Chinese MHC database .................................................................................................... 27
    3.3 Fine-mapping disease related mutation ................................................................................................. 30
    3.4 Paper III: Novel method and technology in genome study ................................................................. 30
  4 Conclusions ............................................................................................................................................... 32
  5 Future Perspective .................................................................................................................................... 33
  6 References ................................................................................................................................................ 34
  7 Appendix .................................................................................................................................................. 42
ABBREVIATIONS

MHC  Major Histocompatibility Complex
HLA  Human Leukocyte Antigen
BAC  Bacterial Artificial Chromosome
APC  Antigen Presenting Cell
IST  Immunosuppression
UNOS United Network for Organ Sharing
DGF  Delayed Graft Function
PID  Primary Immune Deficiency
IgAD IgA Deficiency
RA  Rheumatoid Arthritis
SLE  Systemic Lupus Erythematos
GD  Grave’s Disease
AIDS Acquired Immune Deficiency Syndrome
HIV  Human Immunodeficiency Virus
HBV  Hepatitis B Virus
HCV  Hepatitis C Virus
HGP  Human Genome Project
NGS  Next Generation Sequencing
SBS  Sequence By Synthesis
CG  Complete Genomics
DNB  DNA Nano Ball
SBH  Sequencing By Hybridization
SMRT Single Molecule Real Time sequencing
ZMW  Zero-Mode Waveguide
SBT  Sequencing Based Typing
SNP Single Nuclear Polymorphism
LD  Linkage Disequilibrium
SV  Structural Variation
GWAS Genome Wide Association Study
HSCT  Hematopoietic Stem Cell Transplantation
PE  Paired End
CNV  Copy Number Variation
KIR  Killer cell Immunoglobulin-like Receptor
TRA  T cell Receptor Alpha
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRB</td>
<td>T cell Receptor Beta</td>
</tr>
<tr>
<td>IGH</td>
<td>Immunoglobulin Heavy</td>
</tr>
<tr>
<td>IGL</td>
<td>Immunoglobulin Light</td>
</tr>
<tr>
<td>MAM</td>
<td>Maximum Allowable Mismatch</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Major histocompatibility complex

The major histocompatibility complex (MHC) is a group of genes that found on the cell surface, and they control a major part of the immune system in all vertebrates and help the immune system recognize foreign materials. In humans, the MHC is also called as human leukocyte antigen (HLA). The genomic region encoding MHC locates on the short arm of chromosome 6 on human and its essential for the human immune system [1]. It is also the densest gene region of the human genome, with over 200 identified loci and more than 120 expressed genes. Most of the genes in this region play key roles in human immune system. The typical MHC region is a 3.6M region, which further divided into three major regions according to the genes’ function (Fig. 1). The class I and class II genes play essential role in antigen presentation and main related with adaptive immunity, including the well-known HLA-A,-B,-C in class I region and DRB and DQB gene in class II region. Gene in class III region mainly related with innate immunity [2].

![Diagram of MHC regions on chromosome 6](image)

**Figure 1.** Location and organization of genes for MHC [2].

Due to its importance, MHC was the first multi-megabase region of the human genome to be completely sequenced [1]. In 2004, the sequences of the two common and disease associated MHC haplotypes, the PGF (HLA-A*03:01-B*07:02-C*07:02-DRB1*15:01-DQB1*06:02) and COX (HLA-A*01:01-B*08:01-C*07:02-DRB1*03:01-DQB1*02:01, also named as 8.1 haplotype) in Caucasian have been provided by Sanger sequencing the bacterial artificial chromosome (BAC) using cell lines from consanguineous individual [3]. Subsequently, sequence of another disease
associated MHC haplotype QBL (HLA-A*26:01-B*18:01-C*05:01-DRB1*03:01-DQB1*02:01) was provided by employing the same strategy [4]. To the end of 2008, sequence of eight different MHC haplotypes [5] were provided by the MHC haplotype project in total and the haplotype sequence of PGF has been introduced into the reference sequence of human genome (since NCBI35), and becomes the largest single haplotype sequence of the human genome so far. Pairwise comparison of these eight MHC haplotypes has revealed dramatic variants, showing the great polymorphism of this region and the great diverse of differences between different haplotypes.

1.1.1 Function of MHC region gene in human immune system

As illustrated in the review paper [2], a man dies after he transplants a heart; a woman is crippled because of her rheumatoid arthritis; a child becomes a coma after she gets cerebral malaria; another child dies in an infection as he has immunodeficiency; an elderly man gets serious hepatic cirrhosis because of iron overload. Though these five different clinical situations are different, it all involves in human MHC system.

The importance of MHC to human immune system main due to two classic gene clusters within the region: HLA class I and class II genes. The class I molecules are expressed on the surface of most human cells although the expression level were different between each tissue. While constitutive expression of class II molecules are normally only expressed in a specific group of immune cells called antigen presenting cells (APC) that includes dendritic cells, macrophages, B cells, activated T cells and thymic epithelial cells [6]. One of the most important functions of class I and class II MHC molecules are to present short, pathogen-derived peptides to T cells to initiate the human adaptive immune responses. Furthermore, class III region comprises important gene involved in innate immunity which including the complement genes and genes encoding for some of the inflammatory cytokines [1].

The key to a healthy immune system is its remarkable ability to distinguish between “self” and “non-self”. Among these process, MHC molecules play a fundamental role. Processing endogenous protein and loading the generated short peptides onto class I molecules are happening all the way in body’s cells to mark as “self” and avoid tricking immune reaction main by cytotoxic CD8+ T cell. Meanwhile, the processed exogenous proteins, presented as small peptides, will load onto class II molecules to mark as “non-self”, hence tricking the immune reaction by CD4+ T cell and B cell [7].

1.1.2 MHC related diseases

Due to the importance of genes in this region with immune system, many disorders were discovered to relate with MHC (Fig. 2). According to their mechanism, they can be classified into following groups:
Figure 2: Partial of diseases that found to be related with MHC region [8].

1.1.2.1 Rejection after transplantation

1.1.2.1.1 Graft and Patient Survival
Despite significant advances in immunosuppression (IST) and post-operative patient management
after transplantation, the 5-year survival rates are still significantly low in lung transplant recipients (51%), and better rates are observed in heart (73% for male, 69% female); liver ~78% and Kidney ~89% (http://www.srtr.org/annual_reports). Graft survival records from the united network for organ sharing (UNOS) registry assessing the fraction of graft failures attributed to immunological or non-immunological factors shows that 10-year graft survival rates for living HLA-mismatched donors to be 66%. Analyses of the overall 10-year graft failure rate for cadaver donors (60%) showed that 18% of failures were due to HLA factors, as observed through mismatched living donor grafts [9].

1.1.2.1.2 Delayed Graft Function

Delayed Graft Function (DGF) is a commonly observed adverse post-transplant that impacts graft survival [10]. Risk factors for DGF include non-immunologic factors such as donor-age, cold ischemia time, and recipient ancestry. Immunologic factors have been shown to be associated with DGF including amino-acid differences at over 60 polymorphic sites of HLA-A in a prospective study involve ~700 kidney transplant recipients showed that combinations of amino acid mismatches at crucial HLA-A sites were related with DGF [11]. In European ancestry recipients, amino acid difference at position 62, 95 or 163, which known to play important function as located in the antigen recognition site were found to relate with increased DGF risk. Additionally, decreased risk for DGF was associated with mismatches at HLA-A sites (149, 184, 193 or 246), indicating that evolutionary features of HLA-A variants separating HLA-A families and lineages among D-R pairs may be intertwined with the strength of allogenicity underpinning DGF. While these findings indicate that amino acid variants at important function site in the antigen recognition site of the HLA-A molecule have great relationship with DGF, there may also be additional common and rare variants with similar protective and deleterious influences in MHC region.

1.1.2.2 Primary immune deficiency

Primary immune deficiency (PID), a kind of disorder that part of the body’s immune system is missing or impaired. Most primary immune deficiencies are genetic disorders. Some of selected diseases include:

1.1.2.2.1 IgA deficiency

Selective IgA deficiency (IgAD) is the most common form of primary immunodeficiency disorders in the Caucasian population. It is defined as serum IgA levels <= 0.07 g/l with normal serum levels of IgG and IgM in individuals >= 4 years old [12]. It affects approximately 1 in 600 of Caucasians [13]. However, its prevalence varies among different ethnical populations ranging from 1:143 in Saudi Arabia [14] to 1:23,255 in Japan [15]. Though intense researches, the genetic basis of IgAD remains elusive. However, it is well known that IgAD is strongly associated with MHC, especially the 8.1 haplotype [16].

1.1.2.2.2 Complement deficiency

Complement deficiency is a disorder caused by one of proteins involving the complement system is absent or does not work properly. There are a group of genes in the MHC class III region relate with human complement system. The most common complement deficiency was in C2. Other deficiencies were also found in C1q, C1r, C1s, C4 and also have been reported to link with immune disease [17].

1.1.2.2.3 MHC Class I deficiency
Major histocompatibility complex class I deficiency is a disorder with severe expression reduce of MHC class I genes in the cell surface. It is a rare disease with high biological heterogeneity and has a variable clinical phenotype. Mutations in gene TAP1 or TAP2 which coding for the transporter associated with antigen presentation in the MHC region have been found in patients with this disorder [18].

1.1.2.3 Autoimmune condition
So far, autoimmune diseases comprise the largest number of diseases that found to be related with MHC region (Table 1). It characterize as the body’s immune responses directly against one’s own tissues. This may be restricted to specific organs or involve a particular tissue in different places. Some represented autoimmune diseases are given below.

1.1.2.3.1 Psoriasis
Psoriasis is a common autoimmune and hyper proliferative skin disease, characterized by thick, silvery-scale patches. The prevalence of psoriasis is 0.1-0.3% in Asian and 2-5% in Caucasian [19]. Psoriasis has a strong hereditary component and the major determinant that defined before is PSORS1, which located in the MHC region. It could explains 35%-50% of psoriasis’ heritability [20].

1.1.2.3.2 Rheumatoid arthritis
Rheumatoid arthritis (RA) is a chronic inflammatory disorder, mainly affects the joints of feet and hands. The prevalence of RA is approximately 1% worldwide [21]. Genetic studies reflect this disease is influenced by both genetic and environment factors [22]. The strongest genetic risk factor for RA is HLA-DRB1 gene that located in MHC region. Risk alleles that identified in MHC region include DR4, DR1, and B27.

1.1.2.3.3 System lupus erythematos
Systemic lupus erythematos (SLE) is a chronic inflammatory and classical autoimmune disease in which the body’s immune system attacks its healthy tissue by mistake. It characterized by a diverse array of autoantibodies, immune complex deposition and complement activation, and influenced by both genetic and environmental factors. Most of the cases were women and frequently starting at childbearing age. The MHC class II gene DRB1 was reported to relate with SLE (Table 1).

1.1.2.3.4 Graves disease
Grave’s disease (GD) is an autoimmune disease and characterized by lymphocytic infiltration of the thyroid gland and the production of thyrotropin receptor autoantibodies (TRAbs), leading to produce excess thyroxin, hence to hyperthyroidism. The incidence of GD shows a geographical variation as well as the seasonal trends, suggesting that environmental triggers such as infections might be an important contributor to the GD pathogenesis [23]. However, twin and sibling studies have reported an increased risk ratio for GD [24], indicating genetic plays an important rule on the development of this disease. The most genetic associated loci is the HLA-DRB1 (DRB1*03,DRB1*11,DRB1*01) gene that is locating in MHC region.
Table 1 Autoimmune disease related loci within MHC

<table>
<thead>
<tr>
<th>Disease</th>
<th>Related variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ankylosing spondylitis</td>
<td>B<em>27, DQAI</em>01:03, A<em>02:01, B</em>13:01</td>
</tr>
<tr>
<td>Autoimmune hepatitis</td>
<td>rs2187668, DRB1<em>03:01/DRB1</em>0301, DRB1*0401</td>
</tr>
<tr>
<td>Autoimmune hepatitis type-1</td>
<td>DRB1<em>0301, DQAI</em>0101/DRB1*1301</td>
</tr>
<tr>
<td>Behçet's disease</td>
<td>B<em>51/rs116799036, rs12525170, rs114854070, C</em>1602</td>
</tr>
<tr>
<td>Celiac disease</td>
<td>DQA1<em>05, DQB1</em>02/rs2647044</td>
</tr>
<tr>
<td>Crohn disease</td>
<td>A<em>33, B</em>45, C<em>3, A26, DRB1</em>03, DRB1<em>08/DRB1</em>07, DQB1*0303</td>
</tr>
<tr>
<td>Graves disease</td>
<td>DRB1<em>03, DRB1</em>11, DRB1<em>01/DRB1</em>09:01, DRB1<em>12:02, DPB1</em>05:01, DQB1*03:02,</td>
</tr>
<tr>
<td></td>
<td>DRB1<em>15:01, DRB1</em>16:02 / DRB1<em>0301, DRB1</em>0802, DRB1<em>1403, DRB1</em>0701, DRB1*1302</td>
</tr>
<tr>
<td>IgA nephropathy</td>
<td>DRB1<em>03, DRB1</em>10/DRB1*04/rs660895, rs1794275, rs2523946</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>DRB1<em>15:01, DRB1</em>03:01, DRB1<em>13:03, DRB1</em>04:04, DRB1*04:01,</td>
</tr>
<tr>
<td></td>
<td>DRB1<em>14:01, A</em>02:01, rs9277489, rs2516489, B*37</td>
</tr>
<tr>
<td>Myasthenia Gravis</td>
<td>DQA1<em>01, DQB1</em>0502/DRB1<em>04, DRB1</em>03, DQB1<em>02, DQB1</em>03 / DQB1*05:02</td>
</tr>
<tr>
<td>Narcolepsy</td>
<td>DRB1<em>1501, DQB1</em>0602/rs2858884/DQB1<em>06:01, DQB1</em>03:02,</td>
</tr>
<tr>
<td></td>
<td>rs3117242, DPB1*05:01</td>
</tr>
<tr>
<td>Primary biliary cirrhosis</td>
<td>DRB1<em>0801, DRB1</em>11, DRB1<em>13/DRB1</em>0405, DRB1<em>0803/DRB1</em>08,</td>
</tr>
<tr>
<td></td>
<td>DRB1<em>11, DRB1</em>14/rs2856683, rs9275312, rs9275390, rs7775228,</td>
</tr>
<tr>
<td></td>
<td>rs2395148, rs9277535, rs806156, rs9357512, rs315363, rs9277565,</td>
</tr>
<tr>
<td></td>
<td>rs2281389, rs660895, rs9501626</td>
</tr>
<tr>
<td>Psoriasis</td>
<td>C<em>06:02, C</em>12:03, B(AA 67,9), A(AA 95), DQA1(AA 53)</td>
</tr>
<tr>
<td>Rheumatic heart disease</td>
<td>B<em>51, C</em>04, DRB1<em>01, DQB1</em>08/DRB1<em>07/B</em>13, DRB1<em>01, DRB1</em>04,</td>
</tr>
<tr>
<td></td>
<td>DRB1<em>07, DQB1</em>02, DRB1*13</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>DRB1(AA 9), B(AA 9)/DRB1(AA 11,71,74), B(AA 9), DPB1(AA 9)</td>
</tr>
<tr>
<td>Systemic lupus</td>
<td>rs3131379, rs1270942/rs9271366, rs9275328/ DRB1*15:01,</td>
</tr>
<tr>
<td>erythematous</td>
<td>DRB1<em>13:02, DRB1</em>14:03</td>
</tr>
<tr>
<td>Systemic sclerosis</td>
<td>rs6457617/rs9275224, rs6457617, rs9275245, rs3130573 / DRB1*15:02,</td>
</tr>
<tr>
<td></td>
<td>DRB5*01:02</td>
</tr>
<tr>
<td>Ulcerative colitis</td>
<td>rs9268480, rs660895/rs2395185/DRB1(AA 11)/DRB1<em>13, DRB1</em>08</td>
</tr>
<tr>
<td>Vitiligo</td>
<td>rs11966200, rs9468925/rs7758128/A<em>33:01, B</em>44:03, DRB1*07:01 / rs9468925</td>
</tr>
</tbody>
</table>

1.1.2.4 Infection disease
Infection disease is one of the leading reasons for human morbidity and mortality, especially for children [25]. Genes involved in the immune response locate within regions that experience the most selective pressure. It is generally assumed that resistance to infection exerts selection pressure fuels the generation of MHC variation and, since its discovery, MHC has stood out as the leading candidates for infection disease susceptibility. However, currently, only several infection diseases have been clearly identified to be associated with this region (Table 2). Some of the selected infection diseases were given below.
<table>
<thead>
<tr>
<th>Disease</th>
<th>Relevant variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARS</td>
<td>A*0201</td>
</tr>
<tr>
<td>HIV-1 control</td>
<td>AA_B_97</td>
</tr>
<tr>
<td>HIV-TB</td>
<td>B*4006</td>
</tr>
<tr>
<td>HCV - sustained response to therapy</td>
<td>B*04</td>
</tr>
<tr>
<td>Hepatitis C virus-associated hypertrophic cardiomyopathy</td>
<td>DPB1*0901</td>
</tr>
<tr>
<td>Mycosis fungoides</td>
<td>DQB1*03</td>
</tr>
<tr>
<td>Malaria</td>
<td>DRB1*04</td>
</tr>
<tr>
<td>Leprosy</td>
<td>DRB1*15</td>
</tr>
<tr>
<td>Allergic bronchopulmonary aspergillosis</td>
<td>DRB1<em>1501, DRB1</em>1503</td>
</tr>
<tr>
<td>Melioidosis</td>
<td>DRB1*1602</td>
</tr>
<tr>
<td>AIDS</td>
<td>rs2395029</td>
</tr>
<tr>
<td>Dengue shock syndrome</td>
<td>rs3132468</td>
</tr>
<tr>
<td>Hepatitis B vaccine response</td>
<td>rs3135363(C)</td>
</tr>
<tr>
<td>Kaposi's sarcoma</td>
<td>rs6902982(G)</td>
</tr>
<tr>
<td>HCV-induced liver cirrhosis</td>
<td>rs910049(A)</td>
</tr>
<tr>
<td>Hepatitis B virus-related hepatocellular carcinoma</td>
<td>rs9275319(G)</td>
</tr>
<tr>
<td>HPV</td>
<td>rs9357152(G)</td>
</tr>
</tbody>
</table>

1.1.2.4.1 Acquired immune deficiency syndrome
Acquired immune deficiency syndrome (AIDS) is the final as well as the most serious stage after infected with human immunodeficiency virus (HIV). The HIV infection and HLA genes’ information of MHC region are one of the most convincing correlations of human infection and offer the opportunity to develop novel interventions and therapies. Gao et al. [26] reported specific HLA-B*35-Px subtypes as being responsible for the correlation between HLA-B*35 and rapid progression to AIDS, and they also have confirmed the known protective effect of the HLA-B*27 and B*57 subtypes against progression to AIDS.

1.1.2.4.2 Hepatitis virus infection
Chronic hepatitis virus infection is an important healthy issue, especially in Asian region like China that cause huge public health concern. Hepatitis B and C virus infections are estimated to account for 70% of the global burden of liver disease [27]. The clinical outcomes after infection were different, from clearance of infection to hepatocellular carcinoma. Many researches proved that genetic information play important role in determining the clinical phenotypes. Most of these researches focused on the relationship between HLA genes of MHC region and infection and HLA class II heterozygote advantage has been demonstrated for clearance of hepatitis B virus (HBV) infection [28].

1.1.2.5 Cancer
Human immune system and HLA genes in MHC region plays important role in the development of cancer but the detailed mechanism still poorly understood. Cells of cancer tissue could express some genes that the normal pairs do not, and the short peptides derives from these specific expressed genes will load to MHC class I molecules. These peptides may sever as the target for the immune system. One of the first cancer that reported to relate with MHC genes was Hodgkin’s lymphoma [29]. Many tumors could modulate MHC class I genes’ expression to escape the immune system’ recognition and attack. Cancers that have been showed to related with MHC
region so far were given below.

Table 3: Cancer related loci within MHC

<table>
<thead>
<tr>
<th>Disease</th>
<th>Related variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdominal aortic aneurysm</td>
<td>DQA1*0102</td>
</tr>
<tr>
<td>Acute lymphoblastic leukemia</td>
<td>DQB1<em>0201, DQB1</em>0501, DQB1<em>0301/DRB1</em>15/rs2227956, rs1043618, rs1061581</td>
</tr>
<tr>
<td>Basal cell carcinoma</td>
<td>DRB1<em>07/DRB1</em>01</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>DRB1<em>1301/A</em>30,A<em>31, A</em>24 /DQA1<em>0301,DRB1</em>1303, DQA1<em>0505, DRB1</em>1301/DQB1*02</td>
</tr>
<tr>
<td>Cervical cancer</td>
<td>DRB1<em>1501/rs4282438/DRB1</em>13,DRB1<em>3,DRB1</em>09,DRB1*1201</td>
</tr>
<tr>
<td>Chronic lymphocytic leukemia</td>
<td>A<em>02:01,DRB4</em>01:01</td>
</tr>
<tr>
<td>Diffuse large B cell Lymphoma</td>
<td>DRB1<em>01,C</em>03</td>
</tr>
<tr>
<td>Esophageal squamous cell carcinoma</td>
<td>DQB1<em>05, DQB1</em>06/rs10484561, rs6457327/DPB1*03:01, rs2647012 /DRB1(AA 13)/rs17203612, rs3130437</td>
</tr>
<tr>
<td>Follicular lymphoma</td>
<td>DRB1<em>07, DRB1</em>12/ DQB1<em>0502, DQB1</em>0602/rs2596542, rs9275319</td>
</tr>
<tr>
<td>Hepatocellular carcinoma</td>
<td>rs6903608, rs2281389, DQA1*02:01</td>
</tr>
<tr>
<td>Hodgkin lymphoma</td>
<td>DRB1<em>01, C</em>03</td>
</tr>
<tr>
<td>Large B-cell lymphoma</td>
<td>DQA1*03</td>
</tr>
<tr>
<td>Lung adenocarcinoma</td>
<td>A<em>0201, A</em>2601, B<em>1518, B</em>3802, DRB1<em>0401, DRB1</em>0402, DRB1<em>1201, DRB1</em>1302/rs1052486, rs3117582</td>
</tr>
<tr>
<td>Nasopharyngeal carcinoma</td>
<td>rs2517713, rs2975042, rs29232, rs3129055 , rs9258122/A<em>31, A</em>33, A<em>19, DRB1</em>01, DQB1<em>05, DQB1</em>02/ B<em>46, A</em>24/rs3869062</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>A*02</td>
</tr>
<tr>
<td>Papillary thyroid carcinoma</td>
<td>B*51:01</td>
</tr>
<tr>
<td>Testicular germ cell tumors</td>
<td>DRB1<em>0405, DQB1</em>0401/ B*41</td>
</tr>
</tbody>
</table>

1.1.2.6 Neurological disorder

Immune system abnormal during brain development can cause serious effects on neuronal function and there are more and more evidences show that immune dysregulation is associated with neurodevelopmental disorders [30]. Other neurological disorders, including Alzheimer’s disease [31], Parkinson diseases [32], Schizophrenia [33] and Autism spectrum disorder (ASD) [30] were reported to related with MHC region, which indicated the importance of immune system with these disorders.

1.1.2.7 Others

Remarkably, MHC is also found to be associated with some other conditions, such as smoking behavior [34] and drag sensitivity which includes abacavir induced hypersensitivity with HLA-B*57:01 [35] and carbamazepine sensitivity with B*15:02 and A*31:01 [36]. It has also been reported that MHC is related with mate choice [37].
1.2 High throughput sequencing technology

DNA sequencing is a process of reading and determining the precise order of nucleotides in a DNA molecule. Sanger sequencing method has dominated the sequencing filed for decades since it first developed in 1977 and has paved the road for our understanding of the human genome as well as many other species. Sequencing technology has exploded since the completion of the Human Genome Project (HGP). The rapid development of sequencing methods, especially the arises of next generation sequencing (NGS) methods (Fig. 3) which enable to produce thousands or millions of sequences parallel has greatly accelerated the medical and biological research and discovery. The most advantage of NGS method is its high-throughput, thus lead to produce an enormous volume of data cheaply.

**Figure 3: Development of sequencing methods**

Currently, the most popular NGS technology are Illumina’s Hiseq, Roche’s 454, Life’s Ion Torrent/Proton, PacBio’s SMRT sequencing and Complete Genome’s CG. Detailed review of the mechanism and comparison of these technologies have been published before [38-43]. Along with the emergence of massively parallel sequencing technology, the cost of sequencing a single human genome is dramatically decreased (Fig. 4) and there has been a drive to improve. The newest sequencers, named as ‘third generation’ or ‘single molecular’ sequencing method [44], are able to sequencing DNA directly. Although the throughputs are not comparable to current NGS methods, they have great potential and advantage in clinical usage currently, and may dominate the sequence filed in the future after persistent technical improvement.
1.2.1 Roche 454

Roche 454 DNA sequencer was the first next-generation sequencing method that has been commercialized (Fig. 3). The method was developed by 454 life Sciences and was acquired by Roche in 2007. The technology is derived from the technological convergence of pyrosequencing and emulsion PCR [46]. DNA was first sheared into small fragments and ligated with adapter. Then the adapter-ligated NDA fragments were fixed to beads in a water-in-oil emulsion and amplified by PCR. Finally, these DNA-bound beads were placed to a fiber chip and placed into the system for sequencing.

1.2.2 Illumina Solexa/Hiseq/Miseq

This sequencing technology is based on sequencing by synthesis (SBS) and was first released in 2006 by Solexa [47]. It employs a modified dNTPs that contains a terminator to blocks further synthesis that similar to the Sanger sequencing. The whole sequence process constitutes by three major steps: 1) Sample preparation. DNA is sheared to an appropriate size, i.e., 500bp (according to the sequencing strategy). Then, the ends of fragmented DNA are polished and ligated with unique adapters. 2) Cluster Generation by Bridge PCR. The constructed DNA library is loaded to the flow cell. Using a specific designed process, DNA fragment ligate with the oligonucleotides of the flow cell surface and amplified. 3) Sequencing by synthesis. Each sequencing cycle consists of loading a single fluorescent nucleotide and imaging by a high-resolution camera.

In 2007, Illumina acquired Solexa and have provided a series of next-generation sequencing machines based on the same theory which including Genome Analyzer IIx, Hiseq, Miseq and so on. As the development of this technology, the output of raw sequencing data increased from 1G (GA) at the first begin to near 600G per run (Hiseq2000) and read length increased from single-end 35bp to current pair-end 300bp (Miseq).
1.2.3 Life Technologies Ion Torrent PGM/Proton
Ion Torrent semiconductor sequencing [48] also uses sequencing by synthesis technology, but has moved away from optics altogether. Library preparation involves coating beads in individual library fragments and each bead is then deposited in its own well of a semiconductor plate. The wells are flooded sequentially with each of the four DNA base pairs. If one is incorporated into the growing strand, the hydrogen ions released causes a change in pH and is recorded by the semiconductor chip at the bottom of each well. Homopolymer runs are detected as larger spikes in pH change. Ion Torrent runs are faster due to the absent of optics and lack of modified bases which requiring chemical alteration. Currently, Ion Torrent provides two different sequencing systems: Proton and PGM. The Ion Proton system enables the highest throughput with up to 10 Gb reads in 2-4 hours with read length up to 200bp. While for Ion PGM, although the throughput is only ~2G, the read length could be as long as 400bp.

1.2.3 Complete Genomics
By brought diverse technologies together, Complete Genomics (CG) created a comprehensive platform [49-50] for large-scale human genomics studies. Like other NGS method, the whole sequencing process can also be divide into three major process: 1) Library Construction. Besides shearing DNA into an appreciated size, CG inserts four adaptors into each DNA fragment during the library construction. 2) DNA library amplification and loading to patterned surface. Unlike other sequencing methods which DNA amplification is performed on surfaces or in emulsions, CG’s amplification process performed in solution and produces amplified DNA clusters which named as DNA nano-balls (DNBs). These produced DNBs are loaded on the patterned surface. 3) Sequencing by hybridization (SBH). Pools of probes labeled with four different dyes were hybridized and ligated to the template to produce high accuracy reads.

1.2.4 PacBio
The PacBio sequencing system provided by Pacific Biosciences is one of current next-generation sequencing technologies which uses a single molecule real time sequencing (SMRT) method [51]. This method is also SBS technology that based on real-time imaging of fluorescently marked nucleotides when synthesizing along the DNA template molecules. The DNA sequencing is performed on a chip that contains many zero-mode waveguide (ZMW). Inside each ZMW, a single active DNA polymerase with a single molecule DNA template is immobilized to the bottom through which light can penetrate and create a visualization chamber that allows monitoring of the activity of the DNA polymerase at a single molecule level. The signal from a phospho-linked nucleotide incorporated by the DNA polymerase is detected as the DNA synthesis proceeds which results in the DNA sequencing in real time. As a result, the read length is not uniform but has an approximately lognormal distribution. Meanwhile, sequencing error could be reduced after the process of circular consensus sequencing and assembly.

1.2.6 Oxford Nanopore
The basic idea of Nanopore based sequencing method is threading the DNA molecule through a small hole and measuring the physical property which is related with each nucleotide that passed the hole. It has great potential to become a direct, fast and inexpensive sequencing technology.
Oxford Nanopore was founded in 2005. It translate the academic nanopore research into a commercial, electronics-based sensing technology [52] which could be used into analysis of single molecules, including DNA, RNA and proteins.

1.2.7 Others
Besides these commercial sequencing methods showed above, there were some other high throughput sequencing methods that appeared and used before. Such as Applied Biosystems SOLiD system [53] which using 2-based encoding and ligation sequencing rather than SBS and Helios Biosciences’ Helicos Genetic Analysis System [54], which was the first commercial NGS method to use the principle of single molecule fluorescent sequencing. However, these technologies were gradually withdrawn from the sequencing market because of technical or commercial problems.
2 Aims

2.1 Current problem for MHC region related study

1) Choosing a reliable and user-friendly genotyping method is one of the most important issues for researches or studies which mainly rely on HLA typing information. Sanger sequencing, distinguished for its long read length and accuracy, has dominated high-quality sequencing for decades since it adapted as the gold standard of genotyping. However, Sanger sequencing based typing (SBT) is labor intensive and low throughput, hindering it being used as a routine method, especially in large-scale genotyping projects.

2) Despite the impressive cumulative list of MHC associations of different disorders during recent decades [8], there has been a surprising lacking of progress in identifying causal genes and variants behind these disease associations. This is likely to be due to the complex allelic and genetic structure of this region with an extended linkage disequilibrium (LD) and a high degree of polymorphisms [55].

3) Effective imputation methods and large-scale reference panels are available to determine the high polymorphism HLA genes in silico and fine-map diseases associations within classical HLA genes [56–58]. However, adequately sized reference panels are still not available for many ethnic groups. Moreover, current available reference panels lack MHC-wide ascertainment of variation, and may thus not allow fine mapping of the causal mutation or even miss parts of disease associations. In addition, it has been shown that MHC haplotypes, polymorphisms and recombination rates are highly divergent among individuals and populations [59].

4) Structural variation (SV) have been proven to be an important source of human genetic diversity and disease susceptibility [60-63]. Base pair differences arising from SVs occur on a significantly higher order (>100 fold) than point mutations [64-65], and data from the 1000 Genomes Project show population-specific patterns of SV prevalence [66]. The MHC is one of the most polymorphism regions on genome, there exists abundant SVs between different MHC haplotypes [5], [67]. However, due to its complexity, current methods could hardly reveal this important information.

2.2 Specific aims of this project

Take the advantage of next generation sequencing methods together with advanced bioinformatics analysis to:

1) Develop novel, accurate and cost-effective HLA typing method for routinely usage, especially in large-scale genotype projects.

2) Develop novel tools that could efficiently obtain information of the whole MHC region which could be used to solve the problem that hard to identify the disease causal mutation due to the high polymorphism and linkage dis-equilibrium in genome wide association study (GWAS) for complex disease study in MHC region.

3) Construct a MHC reference panel database that contains comprehensive variants information, including SNPs, Indels and HLA typing, for Han Chinese population, which could serve as a basis for studies on a variety of MHC associated disorders.

4) Adopt new technology in revealing SVs of MHC region as well as in whole genome.
3 Results

3.1 Novel method for MHC region study

3.1.1 Paper I: High throughput HLA typing Method

Sanger based PCR-SBT is the golden standard for genotyping but has been underused because of its low-throughput, labor intensive and high cost of sequencing. NGS methods provide high throughput and inexpensive sequencing, holding great promise for applying to PCR-SBT. However, the short read length is a major challenge.

HLA gene typing is one of the most important applications of genotyping as high-resolution donor-recipient HLA matching is essential for the success of unrelated donor hematopoietic stem cell transplantation (HSCT). However the extremely polymorphic and complex of this region poses a great challenge for obtaining high accurate and resolution result.

Given this situation, we developed a novel method RCHSBT (Fig. 5), a NGS based PCR-SBT assays that takes the advantage of massively parallel high-throughput and paired-end sequencing, which provides accurate HLA gene typing result using assembled haploid sequences. Utilizing this method, amplicons with length as long as Sanger sequencing reads can be readily genotyped.

Figure 5. Outline of RCHSBT procedures.

In order to develop a high throughput, high accurate and low cost HLA typing methods, we adapted steps as described below:

1) First, each amplicon was amplified with Index PCR. Then, amplicons from different sample were pooled together and send for NGS sequencing. By using the index information within each sequenced read, it enables us to group the sequenced reads into each sample’ corresponding amplicon.
Figure 6. Amplicon information of each target gene. (A) For HLA-A, -B and -C, exon 1 to exon 7 were amplified by four reactions. For DQB1, exon 2 and exon 3 were amplified together. For DRB1, only part of exon 2 was amplified. Q stands for DQB1, R stands for DRB1; (B) Agarose gel showing amplicons from PCR. The amplicons ranged between 250 bp and 1,000 bp in length. Amplicon Q2 and Q3 were loaded in the same lane.

2) After align the sequenced reads to corresponding reference, we adopted a Maximum Allowable Mismatches (MAM) cutoff that used for filtering out the aligned low-quality reads to get accurate variants of each amplicon (Table 4).

3) Detected heterozygous variants were phased into haplotype by using reads’ pair-end information.

4) Re-construction of each amplicon’s two haplotype sequences. Sequenced reads were classified into two groups according to previous phased heterozygous markers and assembled into consensus sequences.
5) HLA gene typing by using haplotype sequence information.

All in all, the key to provide accurate HLA gene typing result is by taking the advantage of advanced bioinformatics analysis method to construct each genotyped site’s two-haplotype sequences using the pair-end information from NGS sequencing data. It fully takes the advantage of high throughput NGS sequencing technology which enable ultra-high sequence depth and high polymorphic of HLA genes to phase the haplotype sequence.

### 3.1.2 Paper II: Integrated tools to study whole MHC region

Although previous studies have broaden the database of variations in MHC regions, detecting the variations in human MHC region remains a challenge due to the relatively high cost for whole genome sequencing or PCR-based Sanger sequencing. High throughput NGS technologies have decreased the cost for genome sequencing dramatically. But, sequencing a genome with high coverage (~30-fold, about 100 Gb in total) would still costs several thousands dollars. In addition to the cost for sequencing, analyzing and storage of these vast amounts sequencing data would place a substantial burden on a research center’s bioinformatics infrastructure. For researchers who only interesting in MHC region or HLA gene’s typing information only, targeted sequencing stands out as a promising approach.

Target capture sequencing mainly used to enrich and sequence the interested region, such as exons of genes or the whole exome of entire human genome [68-69]. This strategy could be a cost-effective solution for analyzing MHC region as well. However, It was thought to be difficult to enrichment of large and high variable genomic region, as there exists ubiquitous repetitive sequence and diverse between different haplotypes. In 2011, Pröll et al [70] used microarray capture together with 454 sequencing to analyze the 3.5 million base pair MHC region of an acute myeloid leukemia patient who underwent unrelated HLA-matched allogeneic stem cell transplantation. It proved the feasibility of capture-based method in studying genomic region as complex as MHC. However, the low capture efficiency hinders it from routine usage.
Given this situation, taking the usage of the eight different MHC haplotype sequences in the hg19, together with NimbleGen, we sophisticated design probes that could efficiently capture the whole MHC region (Fig. 7).

**Figure 7. Variant and probe density of MHC.** The figure in middle indicates the variant density along MHC. In the figure on the bottom, a short black vertical line stands for a probe. Additionally, we align all the probes to the eight haplotypes of MHC and find all the regions with higher depth have more probes than other regions.

Based on the captured data, not only we could provide comprehensive variants information of the covered region, but also, can provide the typing result of the 26 most polymorphism HLA genes that recorded in the IMGT/HLA database. These enable researchers to study the relationship between genetic variants with relevant traits [8] at different levels and fine-map the disease causal variant (Fig. 8).
Figure 8. Three-dimensional ribbon models for the HLA-DR, HLA-B and HLA-DP proteins that related with seropositive Rheumatoid arthritis [56].

3.2 Construction Chinese MHC database

It has been shown that MHC haplotypes, polymorphisms and recombination rates are highly divergent among individuals and populations [59], [71]. However, our knowledge of sequence features and diseases/traits associations in MHC region was mainly based on European population. To better understand the genomic feature and to fine mapping disease associated loci in MHC region, constructing a reference database based on each studied population is essential. Using our developed MHC capture array [72], we set out to deep sequence the entire MHC region which close to 5 Mb, from upstream of HLA-A to downstream of HLA-DP, in 9,906 normal Chinese donors in order to provide a population wide database (Table 5), which including SNP (Fig. 9), Indel, HLA gene type (Fig. 10) and MHC haplotype, that will serve as a basis for studies on a variety of MHC associated disorders.

Table 5 Basic data statistics. There are 9,906 samples across MHC 5M region. For example the number 1.35 in the formula “1.46±0.38” represents the mean value while the 0.38 represents the standard deviation. The average depth of MHC is about 55X and the coverage of MHC is about 94.67%. It's adequate for us to do a credible result.

<table>
<thead>
<tr>
<th>Capture sequencing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>#of individuals</td>
</tr>
<tr>
<td>Raw bases(G)</td>
</tr>
<tr>
<td>Mapped bases(G)</td>
</tr>
<tr>
<td>Mapped bases on target(G)</td>
</tr>
<tr>
<td>Capture specificity (%)</td>
</tr>
<tr>
<td>Average sequence depth(fold)</td>
</tr>
<tr>
<td>Fraction of target covered &gt;= 1X (%)</td>
</tr>
<tr>
<td>Fraction of gene covered &gt;= 1X (%)</td>
</tr>
<tr>
<td>Fraction of target covered &gt;= 5X (%)</td>
</tr>
<tr>
<td>Fraction of gene covered &gt;= 5X (%)</td>
</tr>
<tr>
<td>Fraction of target covered &gt;= 10X (%)</td>
</tr>
<tr>
<td>Fraction of gene covered &gt;= 10X (%)</td>
</tr>
</tbody>
</table>
Figure 9. Heterozygosity and functional catalog of common SNPs in the MHC region. (a) The fraction of variants and their contribution to heterozygosity at different allele frequencies. 27% variants are relatively low frequent (MAF<0.1), but 85% of heterozygous sites were due to common variants (MAF>0.05). (b) The number of synonymous SNPs and nonsynonymous SNPs as the increase of minor allele frequency. The nsSNPs have greater diversity among individuals than sSNPs, especially in low frequency variants.
Figure 10. Allelic frequencies at five classic HLA sites. Allele frequency for HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQB1 in Chinese population were given in the figure. Only the top 20 alleles at each locus were showed in diagram. The common alleles (Frequency >5%) were showed in crimson pillars, the low-frequency alleles were showed in light blue pillars and the rare alleles were showed in green pillars.
3.3 Fine-mapping disease related mutation

One of the most important usages of the constructed database is fine mapping of disease causing mutations in the MHC region. For example, rs2281388 (located 5.1 kb downstream of HLA-DPB1) were previously reported to be associated with Graves’ disease in the Han Chinese population [73] but were in insufficient LD ($r^2<0.8$) with currently known nsSNP variants. However, using our database, rs2281388 shows a strong LD ($r^2=0.91$) with the nearby nsSNP rs11551421, which alters a Valine to a Methionine in the fourth exon (HLA-DPB1:NM_002121:exon4:c.G700A:p.V215M) of HLA-DPB1. Rs2281388 was an almost perfect predictor of the DPB1*0501 allele (tagging LD=0.96), suggesting that nsSNP rs11551421 in the DPB1*0501 allele is a potential or even causal functional variant underlying Graves’ disease (Fig. 11).

![Figure 11. Fine map disease causal variants using constructed database.](image)

The SNP rs2281388 was reported to be related with Grave disease in a previous study, based on our constructed database, this SNP were in high LD ($r^2>0.8$) with a functional non-synonymous SNP in HLA-DPB1.

3.4 Paper III: Novel method and technology in genome study

As the development of NGS sequencing technology, it’s much easier and faster to obtain information of a single [46-47], [53-54], [74] or population’s genome [66], [75-76] information than before [77-78]. The well-developed bioinformatics [79–85] could provide each genome’s variants accurately. However, due to the disadvantage of all currently NGS technology – short read length, most of the analyses were focus on the small variants including SNP, Indel as well as copy number variation (CNV) which main takes the usage of depth information. It has little power to detect and analysis the large SVs of human genome, especially for complex region like MHC, T cell receptor alpha/beta (TRA/TRB), Killer cell immunoglobulin-like receptor (KIR), and immunoglobulin heavy/light locus (IGH/IGL) region.

Optical mapping and new genome mapping technologies based on nicking enzymes provide low resolution but long range genomic information and has been successfully used for assisting the genome de novo assembly [86] and for detecting large-scale structural variants and rearrangements.
[67] that cannot be detected using the paired end information provide by current NGS sequencing methods. Thus, taking the advantage of a high throughput optical map technique – Irys platform that based on Nano-channel, we provided the comprehensive large SVs information in the human genome. Moreover, it also gave out valuable information for the most complex genomic regions (Fig. 12).

**Figure 12. Consensus genome maps compared to hg19 in the MHC region** [87]. The green bars represent the hg19 in silico motif map; the blue bars represent consensus genome maps. Large SVs can be seen in the RCCX, HLA-D and HLA-A regions. The Cox and PGF genome maps are shown below for the HLA-A region. HLA: human leukocyte antigen; RCCX: RP-C4-CYP21-TNX module.
4 Conclusions

The demand of developing novel and advanced methods along with emerged new technology is the eternal motivation for genomic study as well as other fields. NGS methods, which is well known for its high throughput, and hence, low sequencing cost, has great potential utility in the sequence field ever since their emergence. However, the relatively shorter read length and higher sequence errors, compare to Sanger sequencing, are two major challenges before using them as a routine method.

Due to the importance of human MHC to immune system, it's one of the most intensive studied genomic regions. However, due to the limitation of previous methods, there was a lot of room for improvement. SBT was the golden standard for HLA gene typing. Sanger based sequencing method, well known for its long read length and accuracy, has dominated this fields for decades. The High throughput NGS sequencing technologies have great potential of usage in genome analyze, especially in genotyping which has been dominated by Sanger sequencing method. However, the two major challenges NGS method need to be solved before it could be applied in genotyping.

Based on sophisticate designed bioinformatics analysis method and making the usage of specific feature of HLA genes – extremely high polymorphism, we achieved nearly the same accuracy as Sanger sequencing based method for HLA genotyping. Situation was the same for the entail MHC and genome region analysis. Results of the studies showed above fully demonstrate the great advantage of new technology and advanced bioinformatics analysis method in genomic related analysis and applications.
5 Future Perspective

Being able to obtain accurate genotyping information of HLA gene, MHC as well as the entail genome was just the beginning. The ultimate goal for genomic research is to reveal the relationship between genotype and phenotype, then utilize this information in personalized medicine, including disease diagnosis, organ transplantation guidance, disease risk prediction, particularly, for complex diseases. To reach this aim, a lot of work still needed to do in the future which includes, but not limited to, novel technology development, new bioinformatics method conception and new biological mechanism discovery.
6 References


35


7 Appendix


Paper I:

A Short-Read Multiplex Sequencing Method for Reliable, Cost-Effective and High-Throughput Genotyping in Large-Scale Studies

By Hongzhi Cao\textsuperscript{1,2}, Yu Wang\textsuperscript{1,3}, Wei Zhang\textsuperscript{1}, Xianghua Chai\textsuperscript{1}, Xiandong Zhang\textsuperscript{1}, Shiping Chen\textsuperscript{1}, Fan Yang\textsuperscript{1}, Caifen Zhang\textsuperscript{1}, Yulai Guo\textsuperscript{1}, Ying Liu\textsuperscript{1}, Zhoubiao Tang\textsuperscript{1}, Caifen Chen\textsuperscript{1}, Yaxin Xue\textsuperscript{1}, Hefu Zhen\textsuperscript{1}, Yinyin Xu\textsuperscript{1}, Bin Rao\textsuperscript{1}, Tao Liu\textsuperscript{1}, Meiru Zhao\textsuperscript{1}, Wenwei Zhang\textsuperscript{1}, Yinxin Li\textsuperscript{1}, Xiuqing Zhang\textsuperscript{1}, Laurent C. A. M. Tellier\textsuperscript{1,2}, Anders Krogh\textsuperscript{2}, Karsten Kristiansen\textsuperscript{2}, Jun Wang\textsuperscript{1,2,4,5}, and Jian Li\textsuperscript{1,2}

\textsuperscript{1}BGI-Shenzhen, Shenzhen, China;
\textsuperscript{2}Department of Biology, University of Copenhagen, Copenhagen, Denmark;
\textsuperscript{3}School of Bioscience and Biotechnology, South China University of Technology, Guangzhou, China;
\textsuperscript{4}King Abdulaziz University, Jeddah, Saudi Arabia;
\textsuperscript{5}The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

Publication details
Published in Human mutation 35(2):270 (2014)
A Short-Read Multiplex Sequencing Method for Reliable, Cost-Effective and High-Throughput Genotyping in Large-Scale Studies

Hongzhi Cao,†,‡,§†, ¹,² Yu Wang,¹,³,§† Wei Zhang,¹ Xianghua Chai,¹ Xiadong Zhang,¹ Shiping Chen,¹ Fan Yang,¹ Caifen Zhang,¹ Yulai Guo,¹ Ying Liu,¹ Zhoubiao Tang,¹ Caifen Chen,¹ Yaxin Xue,¹ Hefu Zhen,¹ Yinyin Xu,¹ Bin Rao,¹ Tao Liu,¹ Meiru Zhao,¹ Wenwei Zhang,¹ Yingrui Li,¹ Xiuxing Zhang,¹ Laurent C. A. M. Tellier,¹,² Anders Krough,² Karsten Kristiansen,² Jun Wang,¹,²,4,5 and Jian Li¹,²,†

¹BGI-Shenzhen, Shenzhen, China; ²Department of Biology, University of Copenhagen, Copenhagen, Denmark; ³School of Bioscience and Biotechnology, South China University of Technology, Guangzhou, China; ⁴King Abdulaziz University, Jeddah, Saudi Arabia; ⁵The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

Communicated by Ian N. M. Day
Received 17 April 2013; accepted revised manuscript 29 August 2013.
Published online 6 September 2013 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22439

ABSTRACT: Accurate genotyping is important for genetic testing. Sanger sequencing-based typing is the gold standard for genotyping, but it has been underused, due to its high cost and low throughput. In contrast, short-read sequencing provides inexpensive and high-throughput sequencing, holding great promise for reaching the goal of cost-effective and high-throughput genotyping. However, the short-read length and the paucity of appropriate genotyping methods, pose a major challenge. Here, we present RCHSBT—reliable, cost-effective and high-throughput sequence based typing pipeline—which takes short-sequence reads as input, but uses a unique variant calling, haploid sequence assembling algorithm, can accurately genotype with greater effective length per amplicon than even Sanger sequencing reads. The RCHSBT method was tested for the human MHC loci HLA-A, HLA-B, HLA-C, HLA-DQB1, and HLA-DRB1, upon 96 samples using Illumina PE 150 reads. Amplicons as long as 950 bp were readily genotyped, achieving 100% typing concordance between RCHSBT-called genotypes and genotypes previously called by Sanger sequence. Genotyping throughput was increased over 10 times, and cost was reduced over five times, for RCHSBT as compared with Sanger sequence genotyping. We thus demonstrate RCHSBT to be a genotyping method comparable to Sanger sequencing-based typing in quality, while being more cost-effective, and higher throughput.

KEY WORDS: short-read next-generation sequencing; high-throughput; genotyping; HLA

Introduction

Genetic testing plays an important role in physician decision making [Carrington and O’Brien, 2003; Napolitano et al., 2005], and population genetic screening for common inherited single-gene disorders is an economical and effective way to reduce the incidence of disease [Bell et al., 2011; Castellani et al., 2009; Liao et al., 2005]. One of the most common types of genetic testing is genotyping. A variety of genotyping tests are available, including Restriction Fragment Length Polymorphism, PCR-sequence-specific oligonucleotides [Ng et al., 1993], PCR-sequence-specific primers [Natio et al., 2001], reverse dot-blot [Gravitt et al., 1998], and PCR-sequencing based typing (PCR-SBT) [Bettinotti et al., 1997]. Among these methods, PCR-SBT based on Sanger sequencing offers the highest resolution, and is the accepted genotyping reference method [Danzer et al., 2007, 2008].

Sanger sequencing, well known for its accuracy and long-read length (800–1000bp), has dominated high-quality sequencing for about 30 years since its introduction as the gold standard of genotyping. But high cost and low throughput has made it underused, particularly for large-scale genotyping projects. Currently popular small-size next-generation sequencing (NGS) technology, comprising massively parallel single-molecule sequencing, can generate billions of bases from amplified single DNA molecules within days, holding great promise for achieving cost-effective and high-throughput genotyping. However, short-read length and error-prone reading, are significant weaknesses of the method [Metzker, 2009; Veloveryting et al., 2009].

454 sequencing—a sequencing technique developed by Roche Nimblegen with the advantage of longer reads—was the first NGS used for genotyping. Several groups have reported successful use of 454 sequencing for HLA genotyping [Bentley et al., 2009; Danzer et al., 2013; Gabriel et al., 2009; Galan et al., 2010; Holcomb et al., 2011; Lank et al., 2010]. Similar in strategy, they combine
multiplex identifier sequence (MIDs) tagged PCR with direct parallel deep sequencing of the pooled amplicons. Genotypes are called by comparing sorted sequencing data to allele sequences in the gene database. Deep sequencing minimizes the importance of random sequencing errors, but this is still not much of an improvement over Sanger SBT in cost and throughput. At the same time, the amplicons involved are restricted to having a length within the maximum-read length of the sequencing machine, which increases the complexity and number of routes to failure in the method.

Few reports about the success or failure of short-read NGS based genotyping methods have been published, though many short-read NGS platforms, such as Illumina sequencing, have lower sequencing costs and much higher sequencing throughput compared with 454 sequencing [Bamshad et al., 2011]. The short sequence-read length, and the lack of a coherent genotyping methodology, is a major challenge. Currently, most short-read NGS methods have a read length of no more than 150 bp (sequence with paired-end read, 150 × 2), so applying these methods to genotyping with the strategy mentioned above can only involve amplicons shorter than 300 bp, otherwise many ambiguous genotypes will be obtained. For target DNA fragments longer than 300 bp, multiple iterations of PCR will be needed to control the amplicon length, increasing the complexity and the cost of the test.

No previously reported NGS-based genotyping method has so far showed significant improvement in cost and throughput, when compared with Sanger SBT.

It is for this reason that we have developed RCHSBT: an NGS-based genotyping method that uses a massively parallel paired-end (PE) sequencing, effective variant phasing, and haploid sequence assembling pipeline. To rigorously assess its performance, we tested it for HLA-A (MIM #142800), HLA-B (MIM #142830), HLA-C (MIM #142840), HLA-DQB1 (MIM #604305), and HLA-DRB1 (MIM #142857), upon 96 previously diagnosed samples using Illumina PE 150 reads. We demonstrate RCHSBT to be a genotyping method comparable to Sanger sequencing based typing in quality, while being more cost effective, and higher throughput.

**Material and Methods**

** Samples**

A total of 96 purified DNA samples with known HLA types from volunteers were selected. Written consent forms were collected from all volunteers who were involved in this study. These samples had been extensively characterized by a variety of HLA testing methods, including Sanger SBT. The DNA sample information is listed in Supp. Table S1.

RCHSBT follows four steps (Fig. 1A and B):

1. Amplification, shearing, and sequencing

15 HLA Class I and Class II exonic regions were amplified by tag PCR (Supp. Fig. S1a). In this assay, both forward and reverse PCR primers respectively contain an identifier 14 PCR amplifications were performed for each of the 96 samples (Supp. Fig. S1b). PCR products of these 96 samples were then pooled
together. The pooled amplicons were fragmented, recovered and used for library construction. The library was sequenced with PE 150 bp reads on Illumina Miseq (See Supp. Methods for detail).

(2) Grouping

DNA sequence reads of low quality were filtered out. The cleaned PE reads were grouped to individual samples, based on the MIDs, and grouped to the specific loci, by genomic primer sequence.

(3) Mapping and haplод sequence assembling

The MIDs and genomic primer sequence were trimmed before grouped PE reads were aligned locally to the reference sequence (Supp. Table S2) by BWA. Variants were called by using SAMtools. The aligned PE reads with a number of mismatches larger than its maximum allowable mismatches (MAM) cutoff (see Supp. Methods for detail) was filtered. After filtering, the variants were called again by SAMtools. The number of aligned PE reads supporting linkage of any two particular, heterozygous SNPs in the amplicon region was investigated (Fig. 2A,1). All the reliable blocks were integrated to phase the heterozygous SNPs of the amplicon. The alignments were sorted into two parts (heterozygous alleles) or one part (homozygous allele), according to each given pair of complete SNP haploids. Consensus sequences were assembled for each part by using SAMTools in haplод mode. Pairs of haplod sequence were used for the genotyping which ensued.

(4) Genotyping

Genotypes were obtained by comparing the thus determined haplод sequences to allele sequences already described in the gene database (IMGT, http://www.ebi.ac.uk/imgt/hla/). Utilizing RCHSBT, amplicons with the length of Sanger sequencing reads were readily genotyped by short PE sequencing.

Results

PCR Amplification and Miseq Sequencing

A total of 96 purified DNA samples with known HLA type were amplified with a unique set of PCR primers at five loci (HLA-A, HLA-B, HLA-C, HLA-DQB1, and HLA-DRB1), using tag PCR. The amplicons ranged from 290 to 950 bp in length. For most PCR, only a single band was observed following amplification and gel electrophoresis. Multiple bands, apparently weak bands, or no bands were observed by gel electrophoresis when there was nonspecific amplification, poor amplification, or amplification failure. All amplicons of the 96 samples were pooled, and randomly sheared.

Sequencing libraries with fragments including library adapters between 400 and 700 bp in length were recovered by gel slicing. The length range of the recovered DNA fragments was confirmed by use of DNA bioanalyzer 2100. Sequencing was performed on Illumina Miseq with PE 150 bp reads, in a single run. A total of 1.89 Gb of read data, 6.30 M PE reads was generated. 97.08% and 85.23% of the sequence fragments had an average quality value over Q20 and Q30, respectively. The average percentage of GC content was 59.94%.

Sequencing Data Grouping and Variants Calling

Of the original 1.89 Gb of raw data, 1.86 Gb of high quality sequence data remained for analysis after all low quality PE reads were purged (See Supp. Methods for detail). For the cleaned PE reads, 70.45% contained MIDs and genomic primer sequence in at least one end, and could be grouped to specific amplicons of specific samples. 0.59% had a MID/genomic primer mis-pairing, likely caused by template switching during PCR cycles [Odelberg et al., 1995]. The remaining 28.96% reads, without MIDs or genomic primer sequences, could not be readily utilized.

The poor amplification and amplification failure of some PCR output can be confirmed by grouped reads. Groups with apparently low or entirely absent reads will be excluded from the following analysis. Grouped reads were aligned to reference (Supp. Table S2) at HLA-A, HLA-B, HLA-C, HLA-DQB1, and HLA-DRB1 by BWA [Li and Durbin, 2009; Li et al., 2008]. As the alignment parameters set were not strict (See Supp. Methods for detail), reads with many sequencing errors could still be mapped to the reference. 93.05% of all the grouped PE reads, with both ends intact, were successfully mapped to the reference sequence, under the initial alignment parameters. The unaligned and single-end aligned reads were purged from mapping.

Variants were called by SAMtools [Li et al., 2009], in diploid mode (See Supp. Methods for detail). The ratios of each amplicon’s heterozygous variants’ two alleles were analyzed. 94.16% had a ratio above 1:6 (the limit of detection for Sanger sequencing) on average. We evaluated variant calling accuracy of RCHSBT at this step by comparing consistency of variants on all 96 samples detected by RCHSBT and Sanger SBT (Supp. Table S3). The rate of false positive and false negative variant calling was 6.21% and 0%, respectively—loci with allele amplification failure were not included in this statistic.

Based on the high accuracy of the initial variant calling, we used the MAM strategy to individually filter aligned PE reads (See Supp. Methods for detail). After this filtering, 18.38% of the alignments with mismatch counts larger than MAM were removed from variant calling.

The rate of heterozygous variants with allele ratios above 1:6 was increased to 98.87%, and the rate of false positive calling reduced to 0.48%, but the rate of false negative calling was increased to 0.08%, as the result of this filtering. This increase in the rate of heterozygous variants with an allele ratio above 1:6, and this concurrent reduction in false positives, mainly came from amplicons of HLA-C (particularly C4). Before MAM filtering, 21.10% of heterozygous variants from amplicons of HLA-C had allele ratios below 1:6, and the rate of false positive variant calling was 26.57%. After MAM filtering, the ratio of heterozygous variants from amplicons of HLA-C with an allele ratio below 1:6 was reduced to 0.22% and the rate of false positive variant calling was reduced to 0.18% (Supp. Table S3).

Depth of each site, for each amplicon of each of the 96 samples, was plotted in Supp. Figure S2. Average depth was 789× for all amplicons. For the exonic region, 99.9% have depths above 100×, and 86.67% above 700×. For all grouped PE reads, because they have at least one end containing MIDs and genomic primer sequences, all amplicons appeared to have the highest depth at the two ends, except for HLA-DRB1, which had highest depth in the middle. This was because HLA-DRB1 was the only amplicon shorter than 300 bp, a length that was within the sequence length of the Miseq PE 150 read.

Variant Phasing and Haplod Sequence Assembly

In order to further improve variant calling accuracy, we chose to refine sequence reads aligned to the reference sequence. We constructed a series of blocks for every amplicon, using the PE reads. Each block reflected the linkage pattern of two particular heterozygous variants. We integrated all of those blocks to phase heterozygous variants of the amplicons (Fig. 2A).

HUMAN MUTATION, Vol. 00, No. 00, 1–6, 2013
Figure 2. Schematic representation of variant phasing and haploid sequence assembly. A: Schematic of variant phasing: A,I: Record linkage relationships between bases located at heterozygous variant sites, according to paired-end reads. A,II: Blocking: All types of linkage relationships between heterozygous variants (blocks) are determined, as supported by PE reads. A,III: Phasing: Haploid variants are obtained, by integrating all the blocks to phase. B: Haploid sequence assembly: B,I: Before sorting, reads from both alleles are mixed. B,II: Sorting: According to the haploid variants, reads from the same alleles are sorted and clustered. B,III: Haploid sequences of the alleles are assembled according to their assigned read cluster.

For amplicons shorter than 770 bp, 99.87% were completely phased. 0.09% had one phasing breakpoint, 0.04% had two phasing breakpoints. For amplicons longer than 770 bp, few amplicons were completely phased, due to there not being enough long DNA fragments sequenced to support the linkage of heterozygous variants. We analyzed the length distribution of the DNA fragments, using the aligned PE reads (Supp. Fig. S3). Over 85.02% of the recovered fragments were within a length range of 150–400 bp, and only 2.52% had a length over 400 bp. The incomplete haploid pairs within each amplicon were cross matched with each other to construct complete haploid pairs. With pairs of haploid variants, we sorted the aligned sequence reads of every amplicon into two parts, and constructed consensus haploid sequences for each portion of the sorted sequence reads (Fig. 2B).

Variants were called by SAMtools for each haploid sequence, running in the haploid mode (See Supp. Methods for detail). We evaluated the improvement of variant calling accuracy by variant phasing and haploid sequence assembly (Supp. Table S3). Through variant phasing and haploid sequence assembly, the rate of false positive and false negative variant calling was reduced to 0.41% and 0.02%, respectively. Pairs of haploid variants were used for the ensuing genotyping.

Genotyping

The resulting haploid variants of each exon were compared with the local HLA variants database (Supp. Table S4), in order to determine exonal genotypes. For each individual sample, the genotypes of the most frequently studied exons (See Supp. Methods for detail) of the given HLA gene were defined as the candidate HLA allele genotype, and used for method comparison.

The genotyping of other, less frequently studied exons were used to filter the candidate genotypes. After filtering, in cases where there were more than two possible HLA allele genotypes, two of the probable allele types were randomly paired. The genotypes, as defined by these paired alleles, were then confirmed by a logical algorithm. By logical deduction, if there was any ambiguity in the choice of genotype, information was inferred from intron variants to resolve the ambiguity.

Of the 480 genotypes from all 96 samples considered, assignment was possible in 97.50% of the cases, and of those 97.50%, 100% of assignments were concordant with the original Sanger genotypes attached to the samples. Notably, among the assigned genotypes, there were several InDel genotypes, such as C*04:82, which has nine bases inserted in exon 5, compared with the reference. The assigned genotypes for HLA-A, HLA-B, HLA-C, and HLA-DQB1 of all samples were 100% unambiguous (Table 1). Though one of the assigned genotypes for HLA-B of sample L002 was still ambiguous after logical judgment (B*15:02:01/15:01:01:01, B*15:25:01/15:15), the ambiguity was resolved by inference, using heterozygous variant information of intron 2 from the same gene. Because we only amplified parts of exon 2 of HLA-DRB1 for genotyping, the unambiguous assignment rate of HLA-DRB1 was only 48.96%.
We performed an analysis to discover the reason why some alleles were not assigned genotypes. For the 12 unassigned alleles, half of them were caused by nonspecific amplification, 1/3 of them were caused by allele amplification failure, and the rest were caused by insufficient data, resulting from poor PCR amplification efficiency.

### Discussion

Using RCHSBT, which takes short sequencing reads as input, but uses a unique variant calling, haploid sequence assembling algorithm, we can accurately genotype with greater effective length per amplicon than even Sanger sequencing reads. In the study, amplicons as long as 950 bp—over sixfold the length of the sequence-read length were readily genotyped, using Illumina PE 150 sequence reads. According to the principle of RCHSBT, the longest sequence assembled can be two times the length of the longest PE sequencing insert. As the maximum DNA length in Illumina sequencing flow-cells is ca. 800 bp, the longest amplicon that can be assembled by RCHSBT is 1,600 bp with a single amplification. This high maximum length could greatly increase the flexibility of primer design, and consequently reduce the complexity of the assay.

Besides comparatively short-read length, a high rate of sequencing error is another major difficulty for NGS-based genotyping when compared with Sanger, especially for clinical genotyping, where resultant diagnosis errors are poorly tolerated. This is why the reliability of RCHSBT, reliant on more accurate methods for variant calling, variant phasing, and haploid assembly, is important. In RCHSBT, the MAM strategy effectively filters out both reads with excess sequencing errors, and reads which are mapped improperly. This filtering facilitates accurate variant calling.

In the pilot study, 18.38% of aligned reads with mismatch counts larger than their MAM were removed from variant calling, by MAM filtering. Using each sample’s known genotype corresponding variants as a reference, overall rates of false positive variant calling were reduced from 6.21% to 0.48%, while overall rates of false negative variant calling increased from 0% to 0.08%—this latter increase due to MAM filtering removing reads with both an extraordinarily high error rate, but also utile information content, particularly impacting variant calling in low depth regions (Supp. Table S3). However, the majority of false negatives can be reduced by using the haploid assembly process, hence the only moderate increase in FN. In the study, the overall rate of false negative variant calling was reduced from 0.08% to 0.02%, after application of the haploid assembly process. Besides this, the overall rate of false positive variant calling was reduced to 0.41% (Supp. Table S3).

Beyond this—and without considering nonspecific amplification, amplification failure, and poor amplification factors—the final variant calling was 100% consistent with the previous Sanger SBT results, as previously described. Combined with deep sequencing (789× on average), the analysis pipeline provides extreme reliability, effectively comparable to the best-of-class, as assigned genotypes were 100% concordant with the Sanger results.

The key point of RCHSBT is to construct the haploid sequence accurately. The haploid sequence construction process is independent of allele database and allele frequency information, so, it can be used for the genotyping of genes without allele database and allele frequency information, opening a wider application than previous reported methods. Consequently, RCHSBT can be run independently of allele database and frequency information, in order to genotype new and rare alleles.

As other NGS-based genotyping methods mainly adopt the 454 sequencing platform, the short-read NGS based RCHSBT has a substantive advantage in terms of sequencing costs and sequencing throughput. RCHSBT capacity to genotype long amplicons (up to 1.6 kb for Illumina) with short reads enables genotyping with fewer amplicons than current 454 based genotyping methods, reducing cost and improving throughput. A multisite study, using high-resolution HLA genotyping by 454 GS FIX, reported that they could genotype 20 samples for 10 HLA genes (17 exons involved) in a single sequencing run, with reagent costs approximately equivalent to Sanger SBT [Holcomb et al., 2011]. In our pilot study, we genotyped 96 samples on 5 HLA genes (24 exons involved) in a single Illumina Miseq sequencing run, enabling us to cut HLA typing costs by 80%, and to improve HLA typing throughput by 10× compared with Sanger SBT.

Compared with other short-read NGS based genotyping methods, RCHSBT constitutes an immense improvement. In a previously reported Illumina genotyping study [Wang et al., 2012], encoded sequencing adaptors were used to identify the source of sequence reads for each sample, requiring library construction for each sample. In this study, both primer MIDs and encoded sequencing adaptors were used to trace the source of sequence reads for each sample. Amplicons with different primer MIDs were pooled together when constructing library, which greatly reduced the number of libraries constructed. Although the initial synthesis of MID primers may be expensive, the cost of primer synthesis for each test will be comparatively lower in large-scale projects using RCHSBT, and the number of primers synthesized can be adjusted according to the scale of the project. The previous study reported genotyping four HLA genes (25 exons in total) in 180 samples, for a single lane of Illumina Hiseq 2000. With the same rate of output (30G), we estimate that at least 1,500 samples, genotyping on five HLA genes, can be processed at the same time in a single Hiseq 2000 lane using RCHSBT, based on the result of our pilot study. In unpublished data, we have randomly picked half of the 1.89 Gb raw data of the pilot study for genotyping, run RCHSBT on that data, and the results were comparable to results obtained by running RCHSBT on the whole set of raw data. This implies potential for further improvement of cost and throughput of genotyping.

Although the performance of RCHSBT was obtained from testing of HLA genes, similar result can be predicted when testing nonpolymorphic genes. As few variants will be called for nonpolymorphic genes, variants phasing and haploid sequence assembly will have little effect on genotyping accuracy. But as each sequence reads contain few SNPs, their initial mapping are accurate, and deep sequencing help remove the influence of sequencing error, people can still genotype long amplicons accurately by RCHSBT using short sequence reads.

### Table 1. Genotyping Consistence Statistic Result of HLA Genes for 96 Samples

<table>
<thead>
<tr>
<th></th>
<th>HLA-A</th>
<th>HLA-B</th>
<th>HLA-C</th>
<th>HLA-DQB1</th>
<th>HLA-DRB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sample with unambiguous genotype assignment</td>
<td>94</td>
<td>92</td>
<td>93</td>
<td>96</td>
<td>47</td>
</tr>
<tr>
<td>Number of sample with ambiguous genotype assignment</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>46</td>
</tr>
<tr>
<td>Number of sample with no genotype assignment</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Consistence of assigned genotype with Sanger result (%)</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
RCHSBT can meet the requirements of both medium-size genotyping projects, and large-scale genotyping projects. Combined with one of the very rapid NGS platforms, such as Illumina Miseq, hundreds of samples can be genotyped in under a week, at levels suitable for clinical genotyping. When reporting time is not very important, people can choose high-throughput NGS platforms, such as Illumina HiSeq. 2000. Tens of thousands of samples can be genotyped in a single HiSeq 2000 run, at costs lower than Miseq, making this method very suitable for large-scale genotyping projects, such as HLA genotyping for global scale bone marrow registries, gene screening for common monogenic disease like thalassemia, and the like.

Although the advantages of RCHSBT were obvious, the limitations must be considered. The longer the amplicon, the less percentage of raw data can be used. PE reads having no MIDs and genomic primer sequence in at least one end are useless, as they cannot be grouped to specific samples for the following analysis. In unpublished data, we tested amplicons with a length of 1.4 kb using RCHSBT, and we found only about 40% of the raw data were useful. Also, as we found the longer the amplicon, the lower the depth in the middle of amplicon was, much more data are needed for genotyping long amplicons using RCHSBT. Another limitation is about SNP phasing. For SNPs within the same amplicons, we can phase them well, and the maximum length of amplicon can be two times the length of the longest PE sequencing insert. For SNPs within different amplicons, we can phase them well only in the case of the two amplicons are overlapped and they have common SNPs between the two SNPs we are going to phase.

In this assay, most of our time and labor was spent on PCR amplification. We know that the key to scaling PCR-based sequence enrichment involves automation, miniaturization, and multiplexing of reactions. There are commercially available platforms which enable miniaturized PCR by using microfluidics (Fluidigm). The combination of RCHSBT and microfluidic PCR can further improve genotyping throughput and cost, making it highly suitable for studies involving large numbers of amplifications, or large numbers of genes for each sample.

Acknowledgments

The authors thank the members of BGI-Shenzhen genome sequencing team for help with sequencing.

Disclosure statement: The authors declare no conflict of interest.

References

An integrated Tool to study MHC Region: Accurate SNV Detection and HLA Genes Typing in Human MHC Region Using Targeted High-Throughput Sequencing.

By Hongzhi Cao, Jinghua Wu, Yu Wang, Hui Jiang, Tao Zhang, Xiao Liu, Yinyin Xu, Dequan Liang, Peng Gao, Yepeng Sun, Benjamin Gifford, Mark D’Ascenzo, Xiaomin Liu, Laurent C. A. M. Tellier, Fang Yang, Xin Tong, Dan Chen, Jing Zheng, Weiyang Li, Todd Richmond, Xun Xu, Jun Wang, Yingrui Li

1Science and Technology Department, BGI-Shenzhen, Shenzhen, China, 2Roche NimbleGen, Inc., Madison, Wisconsin, United States of America, 3Department of Biology, University of Copenhagen, Copenhagen, Denmark, 4King Abdulaziz University, Jeddah, Saudi Arabia, 5The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

Publication details
Published in Plos One 8(7):e69388 (2013)
An Integrated Tool to Study MHC Region: Accurate SNV Detection and HLA Genes Typing in Human MHC Region Using Targeted High-Throughput Sequencing

Hongzhi Cao1,3, Jinghua Wu1,2, Yu Wang1,3, Hui Jiang1,3, Tao Zhang1,2, Xiao Liu1,3, Yinyin Xu1, Dequan Liang1, Peng Gao1, Yepeng Sun1, Benjamin Gifford2, Mark D’Ascenzo2, Xiaomin Liu1, Laurent C. A. M. Tellier1,3, Fang Yang1, Xin Tong1, Dan Chen1, Jing Zheng1, Weiyang Li1, Todd Richmond2, Xin Xu1, Jun Wang1,3,4,5, Yingrui Li1*

1 Science and Technology Department, BGI-Shenzhen, Shenzhen, China, 2 Roche NimbleGen, Inc., Madison, Wisconsin, United States of America, 3 Department of Biology, University of Copenhagen, Copenhagen, Denmark, 4 King Abdulaziz University, Jeddah, Saudi Arabia, 5 The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

Abstract

The major histocompatibility complex (MHC) is one of the most variable and gene-dense regions of the human genome. Most studies of the MHC, and associated regions, focus on minor variants and HLA typing, many of which have been demonstrated to be associated with human disease susceptibility and metabolic pathways. However, the detection of variants in the MHC region, and diagnostic HLA typing, still lacks a coherent, standardized, cost effective and high coverage protocol of clinical quality and reliability. In this paper, we presented such a method for the accurate detection of minor variants and HLA types in the human MHC region, using high-throughput, high-coverage sequencing of target regions. A probe set was designed to template upon the 8 annotated human MHC haplotypes, and to encompass the 5 megabases (Mb) of the extended MHC region. We deployed our probes upon three, genetically diverse human samples for probe set evaluation, and sequencing data show that ~97% of the MHC region, and over 99% of the genes in MHC region, are covered with sufficient depth and good evenness. 98% of genotypes called by this capture sequencing prove consistent with established HapMap genotypes. We have concurrently developed a one-step pipeline for calling any HLA type referenced in the IMGT/HLA database from this target capture sequencing data, which shows over 96% typing accuracy when deployed at 4 digital resolution. This cost-effective and highly accurate approach for variant detection and HLA typing in the MHC region may lend further insight into immune-mediated diseases studies, and may find clinical utility in transplantation medicine research. This one-step pipeline is released for general evaluation and use by the scientific community.

Introduction

The MHC region, one of the most gene-dense regions of the human genome, is located on the short arm of human chromosome 6. It covers over 200 genes, 128 of which are predicted to be expressed [1]. Most MHC genes play a fundamental role in immunity, and show a close relationship with immune-mediated diseases [2,3]. Over 30 years, numerous studies have demonstrated association between some MHC alleles and some disease susceptibilities. Today, more than 100 diseases, - with relation to infection, inflammation, autoimmunity, drug sensitivity, and transplantation medicine -have been reported to be associated with proteins coded by the genes in MHC region [4].

In addition to its high gene density, the MHC region is also one of the most complex regions in the human genome, due to the extremely high density of polymorphism and linkage disequilibrium (LD). This inherent complexity has made identification of the underlying, causative variants contributing to disease phenotypes by genome-wide association study (GWAS) a challenge. The single nucleotide variations (SNV) which distinguish Pgf and Cox, the two most significant MHC haplotypes in the European population, are at variant densities of ~3.4 × 10^{-3}, higher than the estimated average heterozygosity between any other two haplo-
types in the human genome [5]. Recognizing the importance of fully informative polymorphism and haplotype maps of the MHC region, as pertaining to MHC-related-diseases, the MHC Haplotype Consortium has conducted the MHC Haplotype Project between 2000 and 2006, and provided the sequence and annotations of eight different HLA-homozygous-typing haplotypes (PGF, COX, QBL, APD, DBB, MANN, MCF and SSTO) [6].

The Single Nucleotide Polymorphism (SNP) count in dbSNP for the MHC region has increased rapidly from 69,076 SNPs in database version 130 (Apr 30, 2009) to 169,279 in database version 135 (Oct 13, 2011), due in large part to the progress of the 1000 genome project. Although next generation sequencing (NGS) has made the cost of human genome sequencing decrease dramatically, the detection and analysis of all variants in the human MHC region for a large human cohort by whole genome sequencing (WGS) is still a challenge for researchers. For researchers who focus on genomic variations in MHC region, and for researchers who use genomic information to do HLA typing, targeted region sequencing stands out as a promising and neotypal approach.

The human leukocyte antigen (HLA) genes, the most studied genes in the MHC region, encode cell-surface proteins responsible for antigen peptide presentation in adaptive immune response. Considering the key role of HLA genes in immunology, HLA typing is widely used for matching donors and recipients in organ and hematopoietic stem cell transplantation in the clinical context [7,8], particularly for variants at the four gene loci HLA-A, -B, -C and -DRB1. High resolution HLA matching is required for unrelated donor hematopoietic cell transplants [9]. Although precise HLA matching improves overall transplant survival and reduces the risk of rejection, graft-versus-host disease (GVHD) remains a significant and potentially life-threatening complication after hematopoietic cell transplantation (HCT), even when using HLA-matched transplants. Besides the five classic HLA genes (HLA-A, -B, -C, -DRB1 and -DQB1), additional genes in the MHC region, encoding unidentified transplantation antigens, are proposed to be responsible for GVHD [10]. Proll et al. found 3,025 small variations between recipients and donors undergoing unrelated HLA-matched allogeneic stem cell transplantation, and proposed that various differences causing non-synonymous amino acid changes may lead to GVHD [11]. Therefore, choosing an HLA matched unrelated donor with the smallest number of differences to the recipient may help to reduce risk of rejection and increase odds of transplantee survival.

Target region capture sequencing was developed mainly to enrich and sequence specific regions of particular interest, such as specific exons or entire exomes [12,13]. Target Region Sequencing (TRS) may be a cost-effective solution for studying MHC region; however, successful enrichments of large, contiguous, highly variable genomic regions has previously been thought to be problematic, because of ubiquitous repetitive regions and otherwise high diversity between haplotypes. In this study, a probe set was carefully designed for the target regions, using sequences of eight MHC haplotypes, to capture DNA fragments from the human MHC region, and then processed via high throughput sequencing. Using this probe set, we are able to sequence ~97% of the defined MHC region with high SNV calling accuracy by this capture sequencing technology. We have also developed a HLA typing pipeline to type all of the genes in the IMGT/HLA database (http://www.ebi.ac.uk/imgt/hla,Release 3.9.0) with high accuracy, using this capture sequencing data or whole genome sequencing data. This toolset will improve and extend the usage of both capture sequencing and whole genome sequencing data in related studies which rely on HLA typing.

Results

Evaluation of Target Capture Sequencing, in Terms of Coverage

We generate sequence libraries from genomic DNA (gDNA) of the three samples mentioned in the method. After filtering reads with low sequence quality or sequencing adaptor, the purged data are mapped to the human genome reference sequence hg19, and more than 60% of the mapped reads are proved to align to the MHC region (Table 1). Although the probes themselves cover only 72.8% of the defined MHC region (3,620,871 bp of 4,970,558 bp), the alignment result proves that 97% of the defined MHC region is covered by at least one read, and over 95% by at least four reads (Table 1).

Several hundred samples have been captured and sequenced using this method, and the coverage of the MHC region is over 96% for all samples (data not shown). We investigated the 3–4% of uncovered regions and found that more than 99% of the uncovered bases were located in a long repeating region with length >2000 base per repeat (data not shown), which was difficult to solve with capture method due to NGS reads typically being too short to span the breadth of the long repeat region. The MHC region depth distribution showed a similar Poisson distribution for all three samples (Figure 1), indicating an even enrichment of the MHC region.

For further evaluation of the coverage of the genes in the MHC region, RefSeq genes are downloaded from the UCSC table browser (Hg19) [14], of which 213 genes and 380 transcripts are annotated in the defined MHC region. The average coverage of the longest transcripts of the 213 genes in the MHC region was 99.69% (Table S1). As to the coverage of the exons, the average coverage of all exons was 99.94% and 210 of 213 genes were 100% covered for all exons for all three samples (Table S1). This high coverage of the genes in MHC region makes it possible to detect all variations therein.

Evaluation of Variation Detection Accuracy

Use of the capture sequencing method for mutation discovery is critically dependent on the accurate detection of polymorphisms and genotypes. Based on the target capture sequencing data, using GATK tools (v1.43) for YH, NA18532 and NA18555, respectively, we detected 19,002, 19,007 and 20,022 SNPs, and 2,332, 2,077, and 2,513 insertions and deletions (InDels).

When analyzing the distribution of variations, we found that most of those SNPs were located in HLA-A, -B, -C, -DR, -DP and -DQ loci (Figure 2), leading to high degree of polymorphism within those genes. An average 3% of the SNPs identified in our study were determined to be novel, defined by their not being present in dbSNP132, or not being extant in the 1000 Genomes project (1000 Genomes Project Consortium, http://www.1000genomes.org, released on 23 November 2010). Using the ANNOVAR bioinformatics toolset (http://www.openbioinformatics.org/annovar) to functionally annotate the genetic variants in MHC region for the three case samples, we showed that about 2% of all detected variants were within exonic regions, 16% within intronic regions, and 70% within regions defined as intergenic (Table S2). Further investigation showed that most of the non-synonymous variants resulting in changes to amino acid sequences - and hence probable functional differences in the protein - were also located in HLA-A, -B, -C, -DR, -DP and -DQ genes, putatively resulting in different peptide-binding interactions with T cell receptors.

In order to evaluate the accuracy of SNP detection, we compared our genotype result with the Illumina 2.5 M BeadChip
genotype result for YH, and with HapMap genotype result for NA18555 and NA18532. There were 6,099 sites on the Illumina 2.5 M BeadChip which are within the MHC region, and 13,920 and 14,036 genotypes in the HapMap database for NA18532 and NA18555 respectively. 6,066 (99.5%), 13,775 (99.0%) and 13,895 (99.0%) of these genotypes are identified by target capture sequencing data with an accuracy of 98.83%, 97.82% and 98.26%, respectively, for the three study samples (Table S3, S4 and S5). The rate of divergence (about 2%) of our genotype results from the HapMap genotypes is calculated to be higher than the estimated average error rate of HapMap genotypes of 0.5% [15].

Two-thirds of the divergent loci show missing data at one allele in the HapMap genotype, which indicating that these inconsistent sites might result from failure of amplification, caused by the high background heterozygosity of the MHC region. Due to this high degree of divergence between any given pair of MHC haplotypes, biased enrichment data may likewise lead to an erroneous result. In order to evaluate the bias of our reads on two different MHC haplotypes, we therefore count the support read number for each allele at the heterozygous locus in the BeadChip for YH, and HapMap data for NA18532 and NA18555. This count shows that most of the sites had equal support reads for the non-reference allele and reference allele, by which we infer that our probes display good enrichment balance for the two haplotypes (Figure 3).

In order to validate the authenticity of the novel SNPs and InDels, two already proved fosmids covering the HLA-B gene locus were generated, based on our previous project using YH gDNA. Both of the selected fosmids covered the MHC region of chr6:31,317,939 – 31,344,876 (hg19), but they each are templated on different haplotypes. Using the sequence data from fosmids, we validated the accuracy of 379 of the 382 SNPs (99.21%) (including 14 of the 15 novel SNPs) and 35 of the 37 InDels (94.59%) (data not shown).

### Description of HLA Typing Method

Since the target capture sequencing data can be expected to cover more than 99% of genes in MHC region, it has the potential to give all MHC HLA gene types. Here, based on the above cited TRS data, we have developed an HLA typing pipeline which enables typing all of the HLA genes in the IMGT/HLA database until the date of publishing. The pipeline takes the aligned BAM files as input, and outputs the most reliable HLA types for each gene (Figure 4).

A brief summary of the pipeline was as follow:

1. **Read preparation**: Paired reads with either end mapped to target gene region, and unmapped paired reads with high quality, are categorized by gene, and extracted from the BAM. (2) **Quality control and error correction**: The collected reads of each gene are aligned to the IMGT/HLA database, which contains the sequence of all currently known types, using Basic Local Alignment Search Tool (BLAST, version 2.2.24). Reads with a maximum of two discordant (SNP/InDel) to the most similar reference sequence in the database are kept and corrected to the closest reference sequence, while maintaining and adding to a record of the discordant SNP/InDel variants. (3) **Haplotype sequence assemble**: For each exon, reads are randomly assembled, with a minimum overlap region length of 10 bp for reads with no difference allowed in the overlap region, so as to construct haplotypes. Artificial haplotypes that do not exist in IMGT/HLA database are filtered out, and remaining haplotypes are given quality scores as defined below (see method). (4) **Typing according to assembled haplotype sequence**: The final haplotype of the sample is determined by rank of quality score (See method).
Evaluation of HLA Typing Accuracy

Firstly, we test our HLA typing pipeline on simulated data. Based on the sequence of eight known haplotypes, sequenced by Sanger’s MHC haplotype project in 2008 with a foreknown HLA type result, we simulated regional sequence data for a diploid sample MHC, and ran the HLA typing pipeline upon it. For each sample, we typed 26 HLA genes, as recorded in IMGT/HLA database. The simulated data showed the accuracy of our method at 2 and 4 digital resolutions, and at sequencing depths ranging from $20^6$ to $100^6$ with sequencing error rates ranging from 0–2% (Table S6). The results of this typing were then validated against the HLA type given by the MHC haplotype project.

We then used this pipeline on three real human genome samples selected for this study. For each sample, we typed the same 26 HLA genes recorded in the IMGT/HLA database. In order to estimate the accuracy of the HLA typing result, we typed the nine HLA genes with the highest diversity in the IMGT/HLA database using the gold standard: Polymerase Chain Reaction (PCR) based Sanger sequencing, at 4 digital resolution. When our pipeline was compared with this standard, 100% of the 54 HLA alleles typed by our method were found consistent with those of the Sanger standard at 4 digital resolution (Table S7).

In a second test project, 190 human samples which had already been typed by traditional SSO or PCR based Sanger sequencing method (the same gold standard) were captured and sequenced using our pipeline, comparing in particular on five genes (HLA-A, -B, -C, -DRB1 and -DQB1). The comparison showed that the concordance of our method with the traditional gold standard was 96.4% at 4 digital resolution, and 98.8% at 2 digital resolution (data not shown). This result indicates the high performance and reliability of our targeted region pipeline.

Discussion

In our study, we have developed an efficient approach for targeted region sequencing of the MHC region of the human genome, focusing on detection of minor variants and HLA types, in a high-coverage, high-reliability and low-cost framework. Beyond the 3.4 Mb of the standard MHC region, we designed probes for a continuous 5 Mb genomic region, to include further information from the extended high linkage disequilibrium area of the broader MHC region.

Our data presents 97% coverage for the broader MHC region and 99.69% for the transcripts of the crucial MHC genes, in all three testing samples. Compared with the method of Proll [11], our study showed a significantly higher coverage of the MHC region. Moreover, the accuracy of SNVs identified by our target capture sequencing data was in particularly high concordance with genotype data established by the most reliable methodology for this purpose, indicating the high degree of reliability of our approach.

To overcome the challenge of accurately genotyping the extremely high density of polymorphisms and linkage disequilibria in the human MHC region, our method optimized the standard target capture sequencing pipeline. Unlike traditional probes, which are designed only based on one of the haplotypes of the human reference sequence, the probe design for the MHC region in this study is based on all eight major MHC haplotype sequences. This enables capture of less common haplotypes in this region, ultimately reducing capture bias and failure. We also
shear DNA fragments to ∼500 bp, increasing the coverage of the MHC region compared to the shorter fragments length around ∼200 bp (Table S8) which is more typically used in conventional exon capture libraries. This is particularly useful in the MHC because probes must be annealed to unique regions, but some repetitive regions, which are not covered by the probes, can be enriched by flanking to unique genome regions close to them, by using a longer and more unique genome fragment.

Based on this high-coverage, low-bias capture data, we achieve exceptionally accurate SNP and InDel results, and exceptionally accurate HLA gene type results in MHC region, at a level which before could only be achieved by a more specialized and expensive experimental method, or based on a large database. The traditionally utilized HLA typing methods - such as sequence specific oligonucleotide (SSO), sequence specific primer (SSP) [16], PCR based capillary sequencing and next generation sequencing method [17] - only detect variants in two or three exons, and it is very possible for new alleles in other regions of the gene to be missed when using these traditional methods. Another forward-looking method has been reported for the imputation classical HLA alleles and corresponding amino acid based on the SNP information in the MHC region [18], but it demands a huge well studied reference panel. Our data covers essentially all regions of known HLA genes, and has the potential for more accuracy and higher resolution in typing, as well the potential for the detection of novel types (a not infrequent occurrence in this highly polymorphic region), and an important clinical clue in the event of tissue rejection and GVHD. High coverage data allows us to detect all variants and type all HLA genes in the MHC region in one-step with low cost. Some researchers suggest that the transplant barrier is defined not only by classical HLA genes, but also by non-HLA genetic variation within the MHC [10], but many previously dominant methods have been unable to gather data of sufficient breadth to verify or refute this contention.

One limitation of our target region method, as well as most other methods, is the inability to fully cover long repeated regions, and present 100% coverage of MHC region. This difficulty may later be overcome by shearing gDNA to longer fragments (5000 bp, or longer), capturing using probes of this size, and sequencing on the third generation sequencing platforms. A solution to this difficulty is important and may open the possibility of phasing long MHC haplotype blocks at low cost, ultimately reducing the risk of transplantation rejection and GVHD for those less able to afford a more expensive solution.

**Conclusion**

The MHC region harbors the highest density of immunity-crucial genes in the human genome, and is arguably the region

---

Figure 3. Capture bias evaluation using heterozygous genotypes in the Beadchip for three samples. X-axis denotes the support reads number of the reference allele and Y-axis denotes the support reads number of the non-reference allele. doi:10.1371/journal.pone.0069388.g003
most central to immunogenetics and tissue transplantation. The ability to accurately call variants and HLA types is very important to immunology related studies in general, and to clinical application in particular. In this paper, we outline our development of a novel, low cost pipeline for the identification of variants and typing of HLA alleles in the MHC region in a one step protocol with high accuracy. This pipeline has great potential for use in the study of immune disease and transplantation medicine.

Materials and Methods

Design of MHC Capture Probes

The classical MHC region is defined as a ~3.4 Mb region stretching between the gene C6orf40 (ZFP57) and the gene HCG24 [19]. In this study, due to extreme linkage disequilibrium of the region, and due to the presence of MHC-relevant genes nearby [19], we extended the MHC region to the telomeric gene GPX5 and the centromeric gene ZBTB9 in the probes design, corresponding to the genomic region from chr6:28,477,797 to chr6:33,448,354 in the human reference genome (NCBI release GRCh37, UCSC release hg19), encompassing a total of 4,970,558 bp.

Unlike previous studies of the MHC region, we also include the other seven MHC haplotypes aside from PGF - namely COX, QBL, APD, DBB, MANN, MCF and SSTO - in our probes design so as to improve the coverage of the MHC region and minimize the effect of the regionally high polymorphism density. A maximum allowance of greater than 5 close matches is set to get optimal capture efficiency of such high polymorphic regions. In total, 3,620,871 bp (about 72.8% of the 5 Mb MHC region) was covered by the capture probes for PGF, and the coverage rose to 4,150,331 bp (83.5%) when probes extended by 100 bp at both end. The design name is 110729_HG19_MHC_L2R_D03_EZ and design file can be downloaded from Roche NimbleGen website (http://www.nimblegen.com/products/seqcap/ez/designs/index.html) in the product “human MHC design”. Five

![Figure 4. The workflow of HLA typing method. doi:10.1371/journal.pone.0069388.g004](http://www.nimblegen.com/products/seqcap/ez/designs/index.html)
capture arrays were initially produced for testing the uniformity of coverage, and the regions with extremely low or high sequencing depth were rebalanced by adjusting the number of probes - probe count for regions with extremely low sequencing depth was increased, and for regions with extremely high sequencing depth was decreased. Target region sequencing of the MHC was conducted for three samples using the set of rebalanced probes.

**Samples Collection**

Three samples were used to evaluate the performance of the MHC region capture strategy. One was YH, which had previously been whole genome deep sequenced in 2008 [20], and genotyped using Illumina 2.5 M BeadChip. The other two samples were cell line DNA, NA18532 and NA18555, which were purchased from Coriell Institute, and which had previously been analyzed using high-throughput SNP genotyping in the HapMap project [21].

Nine HLA allele (HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQB1, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-G) types for these samples were already detected using the gold standard sequence based typing method, in which the exonic regions of HLA genes were initially amplified by PCR, and the amplicons were subsequently sequenced using Sanger method.

**Whole Genome Shotgun Library Construction**

High molecular weight gDNA was extracted using DNeasy Blood & Tissue Kits (QIAGEN, 69581). Shotgun libraries were generated from 3 micrograms (\(\mu\)g) of genomic DNA followed by the manufacturer’s instruments (Illumina). gDNA in Tris-EDTA was sheared into 400–600 bp fragments using a Covaris S2 (Covaris). The overhangs were then converted to blunt ends, using T4 Polynucleotide Kinase, T4 DNA polymerase and Klenow (Covaris). The fragments were then A-tailed using Klenow (3‘ to 5‘ exo-). Next, Illumina sequencing adapters with a single “T” base overhang were linked to the A-tailed sample using T4 DNA Ligase. Finally, the fragments with adapters were enriched via four cycles of PCR.

**MHC Region Capture and High Throughput Sequencing**

One \(\mu\)g of prepared sample library was hybridized to the capture probes following the manufacturer’s protocols (Roche NimbleGen). The hybridization mixture was incubated for 70 h at 65°C. After the hybridization was finished, the target fragments were captured with M-280 Streptavidin Dynabeads (Invitrogen), and then the captured sample was washed twice at 47°C and three more times at room temperature using the manufacturer’s buffers. The captured target fragments were amplified using Platinum® Pfx DNA Polymerase (Invitrogen) at the following condition: 94°C for 2 min, followed by 15 cycles of 94°C for 15 s, 58°C for 30 s, 72°C for 50 s and a final extension of 72°C for 5 min. The PCR products were purified and sequenced with standard 2×91 paired-end reads on the Illumina HiSeq2000 sequencer following manufacturer’s instructions.

**Sequencing Read Mapping and Variant Calling**

Low quality and adapter contaminated reads were filtered out to get the clean data and then they were aligned to the reference assembly (UCSC Hg19) with haplotype sequences removed using Burrows-Wheeler Aligner (BWA, version 0.5.9) with the parameters -o 1 -e 63 -i 15 -L -k 2 -l 31 -q 10 -I [22]. PCR duplicates were removed by Samtools (version 0.1.17) [23]. Base qualities were recalibrated and reads were realigned around potential insertions and/or deletions using The Genome Analysis Toolkit (GATK) software (version 1.43). SNPs and InDel were also detected using GATK for the regions with at least 4× sequencing depth as described previously [24,25].

**HLA Genes Haplotype Sequences Assemble**

As showed in Figure 4, De novo assembling method was used to construct the haplotype sequence for each exon by using overlapped reads. The assembled sequence was randomly paired if the whole exon was broken by low depth region. We filtered those sequences that did not exist in the IMGT/HLA database. As for the remaining haplotype sequences, we scored them according to the following formula:

\[
Score = C^2 \times N / \log(R)
\]

C: Coverage of this assembled sequence to its corresponding exon;
N: Number of reads used to construct the assembled sequence;
R: Reliability of this assembled sequence. It was calculated using the following formula:

\[
R = \frac{\sum (X_i - \bar{X})^2 + \bar{X}}{\log(df)}
\]

Xi: The depth of each base of the assembled sequence;
X: The mean depth of the assembled sequence;
df: The degrees of freedom of the assembled sequence. df = N - 1, N equals the length of assembled sequence.

**Typing According to Assembled Haplotype Sequence**

We combined the assembled sequences (called haplotypes) of all exons together to produce a group of types, and then scored them using the following formula:

\[
TScore = N \times S
\]

N: The number of reads that support the haplotype.
S: The score of the haplotype.

Based on the assumption that number of reads mapped to the correct haplotype is larger than that mapped to the incorrect haplotype, we first select the two types with the highest TScore, which can explain the highest number of the sequence reads for all exons as candidate types, and then we divide the sequence reads into two corresponding groups. We judge whether a HLA gene would be a homozygous or heterozygous, depending on the ratio of unique supporting reads number of the sub-optimal type, compared to the supporting reads number of the best type. The threshold of a heterozygote is 0.1. If the ratio is smaller then 0.1, the gene is defined as a homozygote, and the type with the largest score is chosen as the dominant type. Otherwise, the typed gene is defined as heterozygote, and the two types with the highest score are chosen as the final type result. Finally, we assess the reliability of the final type by comparing it with the closed type using AScore:

\[
AScore = \frac{(S0 - S1)^2}{(S0 - S1)^2 + (S0 - S2)^2}
\]
S0: The full score of the all assembled haplotypes of the typed gene;
S1: The score of the closed type;
S2: The score of the final type.

On the other hand, all the discordant information recorded in the process of "error correction" was judged using variant calling methods to check whether there is a novel SNP or InDel comparing to an extant type. If so, the relationship of the novel variant to the given type was also described.

Data Simulation and HLA Typing

Basing on the sequence of eight MHC haplotypes given by the MHC haplotype project in 2008, we randomly simulated one diploid sample’s MHC region sequence data using wgsim (v 0.2.3) provided by the Samtools package. First, we randomly selected two from the eight haploids, at various sequencing depths and various sequencing error rates using parameters –e and –N. The insertion size was set to 500 bp, which was the same as our target capture sequencing data, and the SD was 10. For each of the given conditions, we simulated 36 human MHC region data sets. We then worked out each simulated sample’s 26 HLA gene’s type result, using the method we describe above. The statistical accuracy result is given in Table S6.

Data Access

The raw target capture sequencing data of MHC region for sample YH, NA18526 and NA18555 has been deposited at the NCBI Sequence Read Archive (SRA) under accession no. SRA065346. The HLA typing pipeline is implemented in Perl script and can be freely downloaded at http://soap.genomics.org.cn/SOAP-HLA.html.

Supporting Information

Table S1 Coverage of whole gene body and exome of the MHC genes by capture sequencing data.

Table S2 Distribution of variations within different genomics functional regions for three samples.

Table S3 Comparison of MHC capture SNPs and Illumina 2.5 M genotyping alleles for sample YH. We classified both the MHC capture alleles and the alleles that were called by genotyping into three categories: (1) Hom ref (homozygotes where both alleles are identical to the reference); (2) Hom mut (homozygotes where both alleles differ from the reference); (3) Het ref (heterozygotes where only one allele is identical to the reference); The number of MHC capture sequencing sites that are consistent with genotyping at both alleles, at one allele, or that are inconsistent at both alleles were categorized as 2, 1, and 0, respectively.

Table S4 Comparison of MHC capture genotypes and HapMap genotypes for sample NA18532.

Table S5 Comparison of MHC capture genotypes and HapMap genotypes for sample NA18555.

Table S6 The accuracy of HLA typing method using simulated data at different sequencing depth with different sequencing error rate.

Table S7 HLA alleles typed by PCR based Sanger sequence method and target capture sequence method.

Table S8 Coverage of the MHC region using 200 bp insert size library.

Author Contributions

Conceived and designed the experiments: XX. J. Wang YL. Performed the experiments: J. Wu HJ Xiao Liu JZ WL. Analyzed the data: HC J. Wu YW TZ YX DL PG Xiaomin Liu FY XT DC. Wrote the paper: J. Wu. Designed and produced the probe: HC BG MD’A TR. Developed HLA typing pipeline: HC TZ. Assisted in writing manuscript: YW TZ. Revised the manuscript: HC HJ Xiao Liu YS LCAMT.

References

Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology

By Hongzhi Cao\textsuperscript{1,3,4}, Alex R Hastie\textsuperscript{2}, Dandan Cao\textsuperscript{1,3}, Ernest T Lam\textsuperscript{2}, Yuhui Sun\textsuperscript{1,5}, Haodong Huang\textsuperscript{1,5}, Xiao Liu\textsuperscript{1,4}, Liya Lin\textsuperscript{1,5}, Warren Andrews\textsuperscript{2}, Saki Chan\textsuperscript{2}, Shujia Huang\textsuperscript{1,3}, Xin Tong\textsuperscript{1}, Michael Requa\textsuperscript{2}, Thomas Anantharaman\textsuperscript{2}, Anders Krogh\textsuperscript{4}, Huanming Yang\textsuperscript{1,3}, Han Cao\textsuperscript{2} and Xun Xu\textsuperscript{1,3}

\textsuperscript{1}BGI-Shenzhen, Shenzhen 518083, China.
\textsuperscript{2}BioNano Genomics, San Diego, California 92121, USA.
\textsuperscript{3}Shenzhen Key Laboratory of Transomics Biotechnologies, Shenzhen 518083, China.
\textsuperscript{4}Department of Biology, University of Copenhagen, Copenhagen 2200, Denmark.
\textsuperscript{5}School of Bioscience and Biotechnology, South China University of Technology, Guangzhou 511400, China.

Publication details
Published in GigaScience 3:34 (2014)
Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology

Hongzhi Cao1,3,4†, Alex R Hastie2†, Dandan Cao1,3†, Ernest T Lam2†, Yuhui Sun1,5, Haodong Huang1,5, Xiao Liu1,4, Liya Lin1,5, Warren Andrews2, Saki Chan2, Shujia Huang1,5, Xin Tong1, Michael Requa2, Thomas Anantharaman2, Anders Krogh4, Huanming Yang1,3, Han Cao† and Xun Xu1,3*

Abstract

Background: Structural variants (SVs) are less common than single nucleotide polymorphisms and indels in the population, but collectively account for a significant fraction of genetic polymorphism and diseases. Base pair differences arising from SVs are on a much higher order (>100 fold) than point mutations; however, none of the current detection methods are comprehensive, and currently available methodologies are incapable of providing sufficient resolution and unambiguous information across complex regions in the human genome. To address these challenges, we applied a high-throughput, cost-effective genome mapping technology to comprehensively discover genome-wide SVs and characterize complex regions of the YH genome using long single molecules (>150 kb) in a global fashion.

Results: Utilizing nanochannel-based genome mapping technology, we obtained 708 insertions/deletions and 17 inversions larger than 1 kb. Excluding the 59 SVs (54 insertions/deletions, 5 inversions) that overlap with N-base gaps in the reference assembly hg19, 666 non-gap SVs remained, and 396 of them (60%) were verified by paired-end data from whole-genome sequencing-based re-sequencing or de novo assembly sequence from fosmid data. Of the remaining 270 SVs, 260 are insertions and 213 overlap known SVs in the Database of Genomic Variants. Overall, 609 out of 666 (90%) variants were supported by experimental orthogonal methods or historical evidence in public databases. At the same time, genome mapping also provides valuable information for complex regions with haplotypes in a straightforward fashion. In addition, with long single-molecule labeling patterns, exogenous viral sequences were mapped on a whole-genome scale, and sample heterogeneity was analyzed at a new level.

Conclusion: Our study highlights genome mapping technology as a comprehensive and cost-effective method for detecting structural variation and studying complex regions in the human genome, as well as deciphering viral integration into the host genome.

Keywords: Genome mapping, Structural variation, Repeat units, Epstein-Barr virus (EBV) integration

* Correspondence: han@bionanogenomics.com; xuxun@genomics.cn
† Equal contributors
2BioNano Genomics, San Diego, California 92121, USA
1BGI-Shenzhen, Shenzhen 518083, China
Full list of author information is available at the end of the article

© 2014 Cao et al; licensee BioMed Central. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
Background

A structural variant (SV) is generally defined as a region of DNA 1 kb and larger in size that is different with respect to another DNA sample [1]; examples include inversions, translocations, deletions, duplications and insertions. Deletions and duplications are also referred to as copy number variants (CNVs). SVs have proven to be an important source of human genetic diversity and disease susceptibility [2-6]. Base pair differences arising from SVs occur on a significantly higher order (>100 fold) than point mutations [7,8], and data from the 1000 Genomes Project show population-specific patterns of SV prevalence [9,10]. Also, recent studies have firmly established that SVs are associated with a number of human diseases ranging from sporadic syndromes and Mendelian diseases to common complex traits, particularly neurodevelopmental disorders [11-13]. Chromosomal aneuploidies, such as trisomy 21 and monosomy X have long been known to be the cause of Down’s and Turner syndromes, respectively. A microdeletion at 15q11.2q12 has been identified as causal for Prader-Willi syndrome [14], and many submicroscopic SV syndromes have been revealed since then [15]. In addition, rare, large de novo CNVs were identified to be enriched in autism spectrum disorder (ASD) cases [16], and other SVs were described as contributing factors for other complex traits including cancer, schizophrenia, epilepsy, Parkinson’s disease and immune diseases, such as psoriasis (reviewed in [11] and [12]). With the increasing recognition of the important role of genomic aberrations in disease and the need for improved molecular diagnostics, comprehensive characterization of these genomic SVs is vital for, not only differentiating pathogenic events from benign ones, but also for rapid and full-scale clinical diagnosis.

While a variety of experimental and computational approaches exist for SV detection, each has its distinct biases and limitations. Hybridization-based approaches [17-19] are subject to amplification, cloning and hybridization biases, incomplete coverage, and low dynamic range due to hybridization saturation. Moreover, detection of CNV events by these methods provides no positional context, which is critical to deciphering their functional significance. More recently, high-throughput next generation sequencing (NGS) technologies have been heavily applied to genome analysis based on alignment/mapping [20-22] or de novo sequence assembly (SA) [23]. Mapping methods include paired-end mapping (PEM) [20], split-read mapping (SR) [21] and read depth analysis (RD) [22]. These techniques can be powerful, but are tenuous and biased towards deletions owing to typical NGS short inserts and short reads [24,25]. De novo assembly methods are more versatile and can detect a larger range of SV types and sizes (0 ~ 25 kb) by pair-wise genome comparison [23-25]. All such NGS-based approaches lack power for comprehensiveness and are heavily biased against repeats and duplications because of short-read mapping ambiguity and assembly collapse [9,10,26]. David C. Schwartz’s group promoted optical mapping [27] as an alternative to detect SVs along the genome with restriction mapping profiles of stretched DNA, highlighting the use of long single-molecule DNA maps in genome analysis. However, since the DNA is immobilized on glass surfaces and stretched, the technique suffers from low throughput and non-uniform DNA stretching, resulting in imprecise DNA length measurement and high error rate, hindering its utility and adoption [24,27-29]. Thus, an effective method to help detect comprehensive SVs and reveal complex genomic regions is needed.

The nanochannel-based genome mapping technology, commercialized as the “Irys” platform, automatically images fluorescently labeled DNA molecules in a massively parallel nanochannel array, and was introduced as an advanced technology [30] compared to other restriction mapping methods because of high-throughput data collection and its robust and highly uniform linearization of DNA in nanochannels. This technology has previously been described and used to map the 4.7-Mb highly variable human major histocompatibility complex (MHC) region [31], as well as for de novo assembly of a 2.1-Mb region in the highly complex Aegilops tauschii genome [32], lending great promise for use in complete genome sequence analysis. Here, we apply this rapid and high-throughput genome mapping method to discern genome wide SVs, as well as explore complex regions based on the YH (first Asian genome) [33] cell line. The workflow for mapping a human genome on Irys requires no library construction; instead, whole genomic DNA is labeled, stained and directly loaded into nanochannels for imaging. With the current throughput, one can collect enough data for de novo assembly of a human genome in less than three days. Additionally, comprehensive SV detection can be accomplished with genome mapping alone, without the addition of orthogonal technologies or multiple library preparations. Utilizing genome mapping, we identified 725 SVs including insertions/deletions, inversions, as well as SVs involved in N-base gap regions that are difficult to assess by current methods. For 50% of these SVs, we detected a signal of variation by re-sequencing and an additional 10% by fosmid sequence-based de novo assembly whereas the remainder had no signal by sequencing, hinting at the intractability of detection by sequencing. Detailed analyses showed most of the undetected SVs (80%, 213 out of 270) could be found overlapped in the Database of Genomic Variant (DGV) database indicating their reliability. Genome mapping also provides valuable haplotype information on complex regions, such as MHC, killer cell Immunoglobulin-like receptor (KIR), T cell receptor alpha/beta (TRA/TRB) and
immunoglobulin light/heavy locus (IGH/IGL), which can help determine these hyper-variable regions’ sequences and downstream functional analyses. In addition, with long molecule labeling patterns, we were able to accurately map the exogenous virus’ sequence that integrated into the human genome, which is useful for the study of the mechanism of how virus sequence integration leads to serious diseases like cancer.

**Data description**

High-molecular weight DNA was extracted from the YH cell line, and high-quality DNA was labeled and run on the Irys system. After excluding DNA molecules smaller than 100 kb for analysis, we obtained 303 Gb of data giving 95× depth for the YH genome (Table 1). For subsequent analyses, only molecules larger than 150 kb (223 Gb, ~70X) were used. De novo assembly resulted in a set of consensus maps with an N50 of 1.03 Mb. We performed “stitching” of neighboring genome maps that were fragmented by fragile sites associated with nick sites immediately adjacent to each other. After fragile site stitching, the N50 improved to 2.87 Mb, and the assembly covered 93.0% of the non-N base portion of the human genome reference assembly hg19. Structural variation was classified as a significant discrepancy between the consensus maps and the hg19 in silico map. Further analyses were performed for highly repetitive regions, complex regions and Epstein-Barr virus (EBV) integration. Supporting data is available from the GigaScience database, GigaDB [34-36].

**Analyses**

**Generation of single-molecule sequence motif maps**

Genome maps were generated for the YH cell line by purifying high-molecular weight DNA in a gel plug and labeling at single-strand nicks created by the Nt.BspQI nicking endonuclease. Molecules were then linearized in nanochannel arrays etched in silicon wafers for imaging [31,32]. From these images, a set of label locations on each DNA molecule defined an individual single-molecule map. Single molecules have, on average, one label every 9 kb and were up to 1 Mb in length. A total of 932,855 molecules larger than 150 kb were collected for a total length of 223 Gb (~70-fold average depth) (Table 1). Molecules can be aligned to a reference to estimate the error rates in the single molecules. Here, we estimated the missing label rate is 10%, and the extra label rate is 17%. Most of the error associated with these reference differences are averaged out in the consensus de novo assembly. Distinct genetic features intractable to sequencing technologies, such as long arrays of tandem repeats were observed in the raw single molecules (Additional file 1: Figure S1).

**De novo assembly of genome maps from single-molecule data**

Single molecules were assembled de novo into consensus genome maps using an implementation of the overlap-layout-consensus paradigm [37]. An overlap graph was constructed by an initial pairwise comparison of all molecules >150 kb, by pattern matching using commercial software from BioNano Genomics. Thresholds for the alignments were based on a p-value appropriate for the genome size (thresholds can be adjusted for different genome sizes and degrees of complexity) to prevent spurious edges. This graph was used to generate a draft consensus map set that was improved by alignment of single molecules and recalculation of the relative label positions. Next, the consensus maps were extended by aligning overhanging molecules to the consensus maps and calculating a consensus in the extended regions. Finally, the consensus maps were compared and merged where patterns matched (Figure 1). The result of this de novo assembly is a genome map set entirely independent of known reference or external data. In this case, YH was assembled with an N50 of 1.03 Mb in 3,565 maps and an N50 of 2.87 Mb in 1,634 maps after stitching fragile sites (Additional file 1: Figure S2 and Additional file 1: Table S1). These genome maps define motif positions that occur on every 9 kb on average, and these label site positions have a resolution of 1.45 kb. The standard deviation for interval measurements between two labels varies with length. For example, for a 10 kb interval, the standard deviation (SD) is 502 bp, and for a 100 kb interval, it is 1.2 kb. Consensus genome maps were aligned to an in silico Nt.BspQI sequence motif map of hg19. Ninety-nine percent of the genome maps could align to hg19 and they overlap 93% of the non-gap portion of hg19.

**Structural variation analysis**

Using the genome map assembly as input, we performed structural variation detection (Figure 1), and the genome maps were compared to hg19. Strings of intervals between labels/nick motifs were compared and when they diverged, an outlier p-value was calculated and SVs were

---

### Table 1 Molecule collection statistics under different length thresholds

<table>
<thead>
<tr>
<th>Length cutoff (kb)</th>
<th>No. Molecules</th>
<th>Total length (Gb)</th>
<th>Estimated depth (X)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1,568,969</td>
<td>303</td>
<td>95</td>
</tr>
<tr>
<td>120</td>
<td>1,313,852</td>
<td>275</td>
<td>86</td>
</tr>
<tr>
<td>150</td>
<td>932,855</td>
<td>223</td>
<td>70</td>
</tr>
<tr>
<td>180</td>
<td>659,149</td>
<td>179</td>
<td>56</td>
</tr>
<tr>
<td>200</td>
<td>523,146</td>
<td>153</td>
<td>48</td>
</tr>
<tr>
<td>300</td>
<td>170,432</td>
<td>68</td>
<td>21</td>
</tr>
<tr>
<td>500</td>
<td>22,944</td>
<td>14</td>
<td>4</td>
</tr>
</tbody>
</table>

*Estimated depth based on 3.2 Gb genome size.
called at significant differences (See Methods for details), generating a list of 725 SVs including 59 that overlapped with N-base gaps in hg19 (Additional file 2, Spreadsheet 3). Based on the standard deviation of interval measurements, 1.5 kb is the smallest insertion or deletion that can be confidently measured for an interval of about 10 kb if there is no pattern change. However, if label patterns deviate from the reference, SVs with a net size difference less than 1.5 kb can be detected. Additional file 1: Figure S1 shows three mapping examples (one deletion, one insertion, and one inversion) of gap region SVs. We present these 59 events separately although technically, in those cases, genome mapping detected structural differences between the genome maps and reference regions. For the remaining 666 SVs, 654 of them were insertions/deletions (Figure 2) while 12 were inversions (Additional file 2, Spreadsheet 1 & 2). Out of the 654 insertions/deletions, 503 were defined as insertions and 151 were deletions, demonstrating an enrichment of insertions for this individual with respect to the hg19 reference (Figure 2). Of the 59 SV events that span N-gap regions, 5 of them were inversions. Of the remaining 54 events, 51 were estimated to be shorter than indicated and 3 longer. These gap-region related SVs indicate a specific structure of gap regions of the YH genome compared to the hg19 reference.

In order to validate our SVs, we first cross-referenced them with the public SV database DGV (http://dgv.tcag.ca/dgv/app/home) [38]. For each query SV, we required 50% overlap with records in DGV. We found that the

![Figure 1](https://example.com/figure1.png)

**Figure 1** Flowchart of consensus genome map assembly and structural variant discovery using genome mapping data.

![Figure 2](https://example.com/figure2.png)

**Figure 2** Size distribution of total detected large insertions (green) and deletions (purple) using genome mapping. The comparative histogram bars in red and blue respectively represent deletions and insertions supported by NGS. NGS: next-generation sequencing.
majority of the SVs (583 out of 666; 87.5%) could be found (Additional file 2, Spreadsheet 1 & 2), confirming their reliability. Next, we applied the NGS discordant paired-end mapping and read depth-based methods, as well as fosmid-based de novo assembly (See Methods for detail), and as a result, detected an SV signal in 396 (60%, Figure 2) out of 666 SVs by at least one of the two methods (Figure 2, Additional file 2, Spreadsheet 1 & 2). For the remaining 270 SVs, 79% (213 out of 270, Additional file 2, Spreadsheet 1 & 2) were found in the DGV database. Overall, 91% (609 out of 666, Additional file 2, Spreadsheet 1 & 2) of SVs had supporting evidence by retrospectively applied sequencing-based methods or database entries.

We wanted to determine if SVs revealed by genome mapping, but without an NGS supported signal, had unique properties. We firstly investigated the distribution of NGS-supported SVs and NGS-unsupported SVs in repeat-rich and segmental duplication regions. However, we did not find significant differences between them (data not shown) which was in concordance with previous findings [27]. We also compared the distribution of insertions and deletions of different SV categories and found that SV events that were not supported by sequencing evidence were 97% (260 out of 268) insertions; in contrast, the SVs that were supported by sequencing evidence were only 61% (243 out of 396, Figure 2, Additional file 2, Spreadsheet 1) insertions showing insertion enrichment (p = 2.2e-16 Chi-squared test, Figure 2) in SVs without sequencing evidence. In addition, we further investigated the novel 57 SVs without either sequencing evidence or database supporting evidence. We found that the genes they covered had important functions, such as ion binding, enzyme activating and so forth, indicating their important role in cellular biochemical activities. Some of the genes like ELMO1, HECW1, SLC30A8, SLC16A12, JAM3 are reported to be associated with diseases like diabetic nephropathy, lateral sclerosis, diabetes mellitus and cataracts [39], providing valuable foundation for clinical application (Additional file 2, Spreadsheet 1 & 2).

Highly repetitive regions of the human genome

Highly repetitive regions of the human genome are known to be nearly intractable by NGS because short reads are often collapsed, and these regions are often refractory to cloning. We have searched for and analyzed one class of simple tandem repeats (unit size ranging from 2-13 kb) in long molecules derived from the genomes of YH (male) and CEPH-NA12878 (female). The frequencies of these repeating units from both genomes were plotted in comparison with hg19 (Figure 3). We found repeat units across the entire spectrum of sizes in YH and NA12878 while there were only sporadic peaks in hg19, implying an under representation of copy number variation as described in the current reference assembly. Furthermore, we have found a very large peak of approximately 2.5-kb repeats in YH (male, 691 copies) but not in NA19878 (female, 36 copies; Figure 3). This was further supported by additional genome mapping in other males and females demonstrating a consistent and significant quantity of male-specific repeats of 2.5 kb (unpublished). As an example, Additional file 1: Figure S3 shows a raw image of an intact long molecule of 630 kb with two tracts of at least 53 copies and at least 21 copies of 2.5-kb tandem repeats (each 2.5-kb unit has one nick label site, creating the evenly spaced pattern) physically linked by another label-absent putative tandem repeat spanning over 435 kb, and Additional file 1: Figure S4 shows convincing mapping information. Unambiguously elucidating the absolute value and architecture of such complex repeat regions is not possible with other short fragment or hybridization-based methods.

Complex region analysis using genome mapping

Besides SV detection, genome mapping data also provide abundant information about other complex regions in the genome. For complex regions that are functionally important, an accurate reference map is critical for precise sequence assembly and integration for functional analysis [40-43]. We analyzed the structure of some complex regions of the human genome. They include MHC also called Human leukocyte antigen (HLA), KIR, IGL/IGH, as well as TRA/TRB [44-48]. In the highly variable HLA-A and –C loci, the YH genome shared one haplotype with the previously typed PGF genome (used in hg19) and also revealed an Asian/YH-specific variant on maps 209 and 153 (Additional file 1: Figure S5), respectively. In the variant haplotype (Map ID 153), there is a large insertion at the HLA-A locus while at the HLA-D and RCCX loci, YH had an Asian/YH-specific insertion and a deletion. In addition to the MHC region, we also detected Asian/YH-specific structural differences in KIR (Additional file 1: Figure S6), IGH/IGL (Additional file 1: Figure S7), and TRA/TRB (Additional file 1: Figure S8), compared to the reference genome.

External sequence integration detection using genome mapping

External viral sequence integration detection is important for the study of diseases such as cancer, but current high-throughput methods are limited in discovering integration break points [49-51]. Although fiber fluorescence in situ hybridization (FISH) was used to discriminate between integration and episomal forms of virus utilizing long dynamic DNA molecules [52], this method was laborious, low-resolution and low-throughput. Thus, long, intact high-resolution single-molecule data provided
by genome mapping allows for rapid and effective analysis of which part of the virus sequence has been integrated into the host genome and its localization. We detected EBV integration into the genome of the cell line sample.

The EBV virus map was assembled de novo during the whole genome de novo assembly of the YH cell line genome. We mapped the de novo EBV map to in silico maps from public databases to determine the strain that was represented in the cell line. We found that the YH strain was most closely related, although not identical, to strain B95-8 (GenBank: V01555.2). To detect EBV integration, portions of the aligned molecules extending beyond the EBV map were extracted and aligned with hg19 to determine potential integration sites (Additional file 1: Figure S9). There are 1,340 EBV integration events across the genome (Figure 4). We found that the frequency of EBV integration mapping was significantly lower than the average coverage depth (~70X), implying the DNA sample derived from a clonal cell population is potentially more diverse than previously thought, and that this method could reveal the heterogeneity of a very complex sample population at the single-molecule level. Also, the integrated portion of the EBV genome sequence was detected with a larger fraction towards the tail (Additional file 1: Figure S10). Besides integration events, we also found EBV episome molecules whose single-molecule map could be mapped to the EBV genome, free of flanking human genomic regions.

Discussion

Structural variants are increasingly frequently shown to play important roles in human health. However, available technologies, such as array-CGH, SNP array and NGS are incapable of cataloguing them in a comprehensive and unbiased manner. Genome mapping, a technology successfully applied to the assembly of complex regions of a plant genome and characterization of structural variation and haplotype differences in the human MHC region, has been adopted to capture the genome-wide structure of a human individual in the current study. Evidence for over 600 SVs in this individual has been provided. Despite the difficulty of SV detection by sequencing methods, the majority of genome map-detected SVs were retrospectively found to have signals consistent with the presence of an SV, validating genome mapping for SV discovery. Approximately 75% of the SVs discovered by genome mapping...
were insertions; this interesting phenomenon may be a method bias or a genuine representation of the additional content in this genome of Asian descent that is not present in hg19, which was compiled based on genomic materials presumably derived from mostly non-Asians. Analysis of additional genomes is necessary for comparison. Insertion detection is refractory to many existing methodologies [24,25], so to some extent, genome mapping revealed its distinct potential to address this challenge. Furthermore, functional annotation results of the detected SVs show that 30% of them (Additional file 2, Spreadsheet 1 & 2) affect exonic regions of relevant genes which may cause severe effects on gene function. Gene ontology (GO) analysis demonstrates that these SVs are associated with genes that contribute to important biological processes (Additional file 2, Spreadsheet 1 & 2 and Additional file 1: Figure S11), reflecting that the SVs detected here are likely to affect a large number of genes and may have a significant impact on human health. Genome mapping provides us with an effective way to study the impact of genome-wide SV on human conditions. Some N-base gaps are estimated to have longer or shorter length or more complex structurally compared to hg19, demonstrating that genome mapping is useful for improving the human and other large genome assemblies. We also present a genome-wide analysis of short tandem repeats in individual human genomes and structural information and differences for some of the most complex regions in the YH genome. Independent computational analysis has been performed to discern exogenous viral insertions, as well as exogenous episomes. All of these provide invaluable insights into the capacity of genome mapping as a promising new strategy for research and clinical application.

The basis for the genome mapping technology that enables us to effectively address shortcomings of existing methodologies is the use of motif maps derived from extremely long DNA molecules hundreds of kb in length. Using these motif maps, we are able to also access challenging loci where existing technologies fail. Firstly, global structural variations were easily and quickly detected. Secondly, evidence for a deletion bias which is commonly observed with both arrays and NGS technology generates sequence-specific motif-labeled DNA molecules and analyzes these motif maps using an overlap-layout-consensus algorithm, subsequent performance and resolution largely depends on motif density (any individual event endpoints can only be resolved to the nearest restriction sites). For example, the EBV integration analysis in this study was more powerful in the high-density regions (Additional file 1: Figure S10). Hence, higher density labeling methods to increase the information density that may promote even higher accuracy and unbiased analysis of genomes are currently being further developed. When data from genome mapping is combined with another source of information, one can achieve even higher resolution for each event. In addition, reducing random errors like extra restriction sites, missing restriction sites and size measurement is important for subsequent analysis. Finally, improvements to the SV detection algorithm will provide further discovery potential, and balanced reciprocal translocations can be identified in genome maps generated from cancer model genomes (personal communication, Michael Rossi).

The throughput and speed of a technology remains one of the most important factors for routine use in clinical screening as well as scientific research. At the time of manuscript submission, genome mapping of a human individual could be accomplished with fewer than three nanochannel array chips in a few days. It is anticipated that a single nanochannel chip would cover a human size genome in less than one day within 6 months, facilitating new studies aimed at unlocking the inaccessible portions of the genome. In this way, genome mapping has an advantage over the use of multiple orthogonal methods that are often used to detect global SVs. Thus, it is now feasible to conduct large population-based comprehensive SV studies efficiently on a single platform.

Methods

High-molecular weight DNA extraction

High-molecular weight (HMW) DNA extraction was performed as recommended for the CHEF Mammalian Genomic DNA Plug Kit (BioRad #170-3591). Briefly, cells from the YH or NA12878 cell lines were washed with 2x with PBS and resuspended in cell resuspension buffer, after which 7.5 x 10⁵ cells were embedded in each gel plug. Plugs were incubated with lysis buffer and protease K for four hours at 50°C. The plugs were
washed and then solubilized with GELase (Epicentre). The purified DNA was subjected to four hours of drop dialysis (Millipore, #VCWP04700) and quantified using Nanodrop 1000 (Thermal Fisher Scientific) and/or the Quant-IT dsDNA Assay Kit (Invitrogen/Molecular Probes).

**DNA labeling**

DNA was labeled according to commercial protocols using the IrysPrep Reagent Kit (BioNano Genomics, Inc). Specifically, 300 ng of purified genomic DNA was nicked with 7 U nicking endonuclease Nt.BspQI (New England BioLabs, NEB) at 37°C for two hours in NEB Buffer 3. The nicked DNA was labeled with a fluorescent-dUTP nucleotide analog using Taq polymerase (NEB) for one hour at 72°C. After labeling, the nicks were ligated with Taq ligase (NEB) in the presence of dNTPs. The backbone of fluorescently labeled DNA was stained with YOYO-1 (Invitrogen).

**Data collection**

The DNA was loaded onto the nanochannel array of BioNano Genomics IrysChip by electrophoresis of DNA. Linearized DNA molecules were then imaged automatically followed by repeated cycles of DNA loading using the BioNano Genomics Irys system.

The DNA molecules backbones (YOYO-1 stained) and locations of fluorescent labels along each molecule were detected using the in-house software package, IrysView. The set of label locations of each DNA molecule defines an individual single-molecule map.

**De novo genome map assembly**

Single-molecule maps were assembled de novo into consensus maps using software tools developed at BioNano Genomics. Briefly, the assembler is a custom implementation of the overlap-layout-consensus paradigm with a maximum likelihood model. An overlap graph was generated based on pairwise comparison of all molecules as input. Redundant and spurious edges were removed. The assembler outputs the longest path in the graph and consensus maps were derived. Consensus maps are further refined by mapping single-molecule maps to the consensus maps and label positions are recalculated. Refined consensus maps are extended by mapping single molecules to the ends of the consensus and calculating label positions beyond the initial maps. After merging of overlapping maps, a final set of consensus maps was generated and used for subsequent analysis. Furthermore, we applied a “stitching” procedure to join neighboring genome maps. Two adjacent genome maps would be joined together if the junction a) was within 50 kb apart, b) contained at most 5 labels, c) contained, or was within 50 kb from, a fragile site, and d) also contained no more than 5 unaligned end labels. If these criteria were satisfied, the two genome maps would be joined together with the intervening label patterns taken from the reference in silico map.

**Structural variation detection**

Alignments between consensus genome maps and the hg19 in silico sequence motif map were obtained using a dynamic programming approach where the scoring function was the likelihood of a pair of intervals being similar [53]. Likelihood is calculated based on a noise model which takes into account fixed sizing error, sizing error which scales linearly with the interval size, misaligned sites (false positives and false negatives), and optical resolution. Within an alignment, an interval or range of intervals whose cumulative likelihood for matching the reference map is worse than 0.01 percent chance is classified as an outlier region. If such a region occurs between highly scoring regions (p-value of 10e-6), an insertion or deletion call is made in the outlier region, depending on the relative size of the region on the query and reference maps. Inversions are defined if adjacent match-groups between the genome map and reference are in reverse relative orientation.

**Signals refined by re-sequencing and de novo assembly based methods**

In order to demonstrate the capacity of genome mapping for the detection of large SVs, we tested the candidate SVs using whole-genome paired-end 100 bp sequencing (WGS) data with insert sizes of 500 bp and fosmid sequence-based de novo assembly result. SVs were tested based on the expectation that authentic SVs would be supported by abnormally mapped read pairs, and that deletions with respect to the reference should have lower mapped read depth than average [20,22,23]. We performed single-end/(paired-end + single-end) reads ratio (sp ratio) calculations at the whole-genome level to assign an appropriate threshold for abnormal regions as well as depth coverage. We set sp ratio and depth cutoff thresholds based on the whole genome data to define SV signals. Insertions with aberrant sp ratio and deletions with either sp ratio or abnormal depth were defined to be a supported candidate.

We also utilized fosmid-based de novo assembly data to search for signals supporting candidate SVs. We used contigs and scaffolds assembled from short reads to check for linearity between a given assembly and hg19 using LASTZ [54]. WGS-based and fosmid-based SV validation showed inconsistency and/or lack of saturation as each supported unique variants (Additional file 1: Figure S2) [24].
EBV integration detection
Single-molecule maps were aligned with a map generated in silico based on the EBV reference sequence (strain B95-8; GenBank: V01555.2). Portions of the aligned molecules extending beyond the EBV map were extracted and aligned with hg19 to determine potential integration sites.

Availability of supporting data
The data sets supporting the results of this article is available in the GigaScience GigaDB, repository [55]. See the individual GigaDB entries for the YH Bionano data [35] and YH fosmid validation data [36], which is also available in the SRA [PRJEB7886].

Additional files

Additional file 1: Figure S1. Comparison of consensus genome maps and hg19 reference across gap regions. Figure S2. Consensus genome map coverage of human reference assembly (hg19). Figure S3. Examples of repetitive sequence detected in intact single molecules by genome mapping. Figure S4. Consensus genome map compared to hg19 in a long tandem repeat region. Figure S5. Consensus genome maps compared to hg19 in the MHC region. Figure S6. Consensus genome maps compared to hg19 in the XIR region. Figure S7. Consensus genome map compared to hg19 in the IGH and IGL regions. Figure S8. Consensus genome maps compared to hg19 in the TRA and TRB region. Figure S9. Single-molecule alignment to EBV in silico motif map (strain B95-8) showing evidence of strain variation and heterogeneous integration. Figure S10. Distribution of integrated portions of the EBV genome. Figure S11. GO annotations of genes within called SVs. Table S1. Summary of consensus genome map assembly. Additional file 2: Spreadsheet 1. Summary of insertions and deletions detected in BioNano genome maps. Spreadsheet 2: Summary of insertions detected in BioNano genome maps. Spreadsheet 3: Validation of insertions and deletions detected in BioNano genome maps.

Abbreviations
Array-CGH: Array-based comparative genomic hybridization; AS: De novo sequence assembly; ASD: Autism spectrum disorder; BCR: B cell receptor; CNV: Copy number variant; DGV: Database of genomic variants; EBV: Epstein-Barr virus; FISH: Fluorescence in situ hybridization; GO: Gene ontology; HLA: Human leukocyte antigen; HMW: High-molecular weight; IGH: Immunoglobulin heavy locus; IGL: Immunoglobulin light locus; KIR: Killer cell immunoglobulin-like receptor; LRC: Leukocyte Receptor Complex; MHC: Major histocompatibility complex; NGS: Next-generation sequencing; PCR: Polymerase chain reaction; PEM: Pair-end mapping; RD: Read depth; SNP: Single nucleotide polymorphism; SR: Split read; SV: Structural variation; TCR: T cell receptor; TRA: T cell receptor alpha locus; TRB: T cell receptor beta locus; WGS: Whole-genome sequencing; YH: YanHuang.

Competing interests
The authors have the following interests: Alex R. Hastie, Ernest T. Lam, Warren Andrews, Saki Chan, Michael Requa, Thomas Anantharaman and Han Cao were employees of BioNano Genomics at the time of the study, and they owned company stock options.

Authors’ contributions
XX, HC and HZC managed the project. HC, HZC, ARH, DC, ETL and SC designed the analyses. HC, ARH, ETL, W A, MR and TA produced genome mapping data. HZC, ARH, DC, ETL, YS, HH, XL, LY, WA, SC, SH, XT, MR, TA, AK and HY performed the data analyses. HZC, ARH and DC most of the writing with contributions from all authors. All authors read and approved the final manuscript.

Acknowledgements
We wish to recognize BGI-Shenzhen’s sequencing platform for generating the data in this study. We thank the faculty and staff at BGI-Shenzhen who contributed to this project. We also thank everyone at BioNano Genomics who contributed to the development of the Irys system and the team’s helpful discussion and analysis. This work was supported by the State Key Development Program for Basic Research of China—973 Program (NO.2011CB809002); the National High Technology Research and Development Program of China - 863 Program (NO.2012AA02A102); the Shenzhen Municipal Government of China (NO.JC201105260191A); Shenzhen Key Laboratory of Transomics Biotechnologies (NO.CX201108250096A).

Author details
1BGI-Shenzhen, Shenzhen 518083, China. 2BioNano Genomics, San Diego, California 92121, USA. 3Shenzhen Key Laboratory of Transomics Biotechnologies, Shenzhen 518083, China. 4Department of Biology, University of Copenhagen, Copenhagen 2200, Denmark. 5School of Bioscience and Biotechnology, South China University of Technology, Guangzhou 511400, China.

Received: 21 May 2014 Accepted: 28 November 2014
Published: 30 December 2014

References


doi:10.1186/2047-217X-3-34
Cite this article as: Cao et al: Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. GigaScience 2014 3:34.