

# SCIENTIFIC REPORTS



OPEN

## Haplotyping the human leukocyte antigen system from single chromosomes

Received: 04 February 2016

Accepted: 29 June 2016

Published: 27 July 2016

Nicholas M. Murphy<sup>1,2</sup>, Matthew Burton<sup>3</sup>, David R. Powell<sup>4</sup>, Fernando J. Rossello<sup>5,6</sup>, Don Cooper<sup>7</sup>, Abha Chopra<sup>7</sup>, Ming Je Hsieh<sup>1</sup>, David C. Sayer<sup>8</sup>, Lavinia Gordon<sup>9</sup>, Mark D Pertile<sup>10,11</sup>, Brian D. Tait<sup>12</sup>, Helen R. Irving<sup>1</sup> & Colin W. Pouton<sup>1</sup>

We describe a method for determining the parental HLA haplotypes of a single individual without recourse to conventional segregation genetics. Blood samples were cultured to identify and sort chromosome 6 by bivariate flow cytometry. Single chromosome 6 amplification products were confirmed with a single nucleotide polymorphism (SNP) array and verified by deep sequencing to enable assignment of both alleles at the HLA loci, defining the two haplotypes. This study exemplifies a rapid and efficient method of haplotyping that can be applied to any chromosome pair, or indeed all chromosome pairs, using a single sorting operation. The method represents a cost-effective approach to complete phasing of SNPs, which will facilitate a deeper understanding of the links between SNPs, gene regulation and protein function.

Genetic haplotyping and genome phasing are increasingly prominent components of both sequencing technologies and clinical genotyping/phenotyping. Chromosomal phasing or haplotype phasing refers to the physical linkage of chromosomal polymorphisms unencumbered by the ambiguity of polyploidy. The effect of phase and the quantifiable effects that the phasing of inter-chromosomal single nucleotide polymorphisms (SNPs), copy number variations (CNVs), insertions/deletions (INDELs) and structural rearrangements have on the expression of adjacent co-inherited genes is a challenging but increasingly crucial component of contemporary medicine. We sought to adapt published methods using flow cytometry to sort chromosomes in suspension. This allowed us to acquire single chromosomes and to rapidly determine the haplotype phase of genes which are of critical importance to immune function in bone-marrow transplantation<sup>1</sup>.

Located on the short arm of human chromosome 6, the major histocompatibility complex (MHC) has been extensively haplotype sequenced due to its relevance to transplant matching and disease diagnosis<sup>2</sup>. The highly polymorphic human leukocyte antigen (HLA) system genes are central to the recognition of self from non-self. For this reason, the matching of HLA loci sequences is a cornerstone of donor/recipient matching in organ and haematopoietic stem cell transplantation (HSCT)<sup>3</sup>.

During antigen presentation, the MHC Class I genes (*HLA-A*, *HLA-B* and *HLA-C*) present cytosolic peptides and MHC Class II genes (*HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1*) present extracellular peptides. Matching alleles of MHC Class I and II is standard, and there is also growing interest in selecting killer immunoglobulin-like receptor (KIR) alleles<sup>4–6</sup>. While donor/recipient mismatches are tolerated in solid organ transplants, matching in HSCT is stringent, due to increasing severity of graft versus host disease with rising numbers of HLA mismatches. HLA genes are co-dominant and are inherited as conserved ancestral

<sup>1</sup>Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, VIC, Australia. <sup>2</sup>Preimplantation Genetic Diagnosis, Melbourne IVF, Melbourne, VIC, Australia. <sup>3</sup>Flow Cytometry & Imaging, Murdoch Children's Research Institute, Parkville, VIC, Australia. <sup>4</sup>Bioinformatics Platform, Monash University, Clayton, VIC, Australia. <sup>5</sup>Department of Anatomy and Developmental Biology, Monash University, Clayton, VIC, Australia. <sup>6</sup>Australian Regenerative Medicine Institute, Monash University, Clayton, VIC, Australia. <sup>7</sup>Institute for Immunology and Infectious Diseases, Murdoch University, Murdoch, WA, Australia. <sup>8</sup>Conexio Genomics, Fremantle, WA, Australia. <sup>9</sup>Australian Genome Research Facility, The Walter and Eliza Hall Institute of Medical Research, Royal Parade, Parkville, VIC, Australia. <sup>10</sup>Victorian Clinical Genetics Services, Parkville, VIC, Australia. <sup>11</sup>Department of Paediatrics, University of Melbourne, Parkville, VIC, Australia. <sup>12</sup>Transplant Services, Australian Red Cross Blood Service, Parkville, VIC, Australia. Correspondence and requests for materials should be addressed to H.R.I. (email: helen.irving@monash.edu) or C.W.P. (email: colin.pouton@monash.edu)

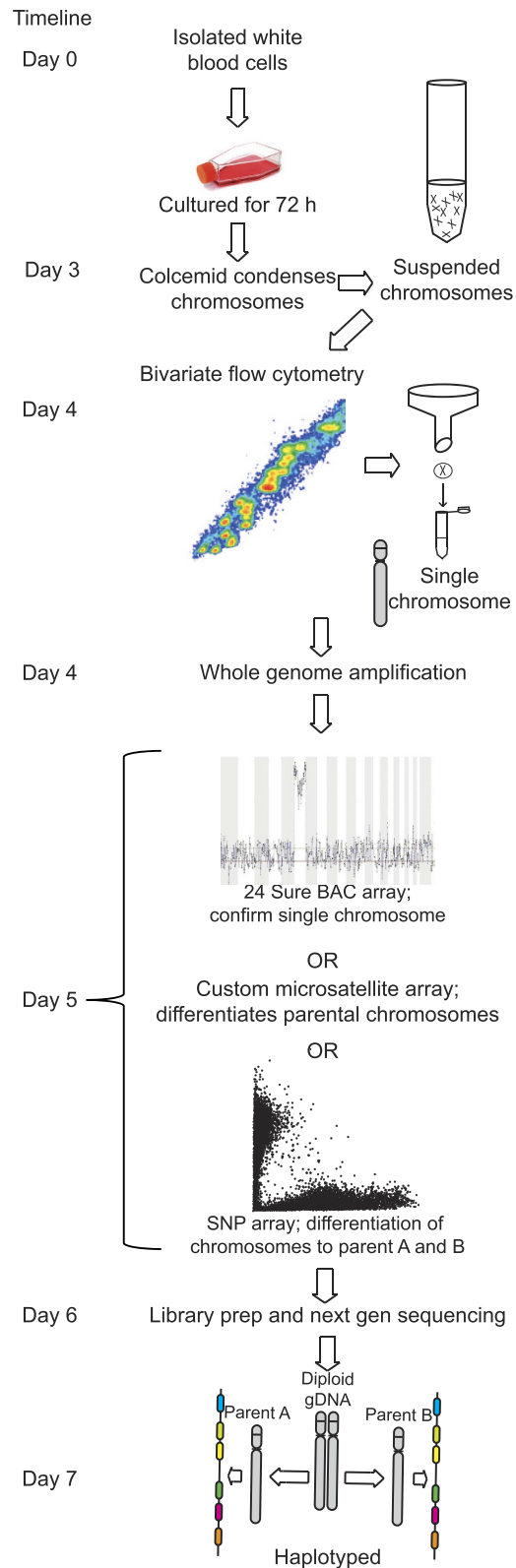
haplotypes from each parent<sup>7</sup>. At present the definitive method of generating haplotype data on an individual is by family pedigree<sup>8</sup>. The demand for a high degree of matching has led to the establishment of HSCT donor registries worldwide which contain HLA data on some 16 million potential donors. Despite the technical challenge of HLA haplotyping, there is evidence that HLA haplotype matching decreases acute Graft-versus-host-disease (GvHD) without modifying risk of relapse, possibly due to non-HLA alleles on the haplotype, and can increase survival in unrelated transplants<sup>9–11</sup>. It is not feasible to establish which donors are haplotype-matched with a particular patient, therefore current recruitment and HLA typing of donors on to HSCT registries does not include definition of HLA haplotypes. Resolving the sequence of each HLA allele directly from diploid genomic DNA requires a significant sequencing campaign and presents a complex computational problem. To date, the most successful HLA-typing approaches employ targeted fragment enrichment for allele calling but do not result in haplotype phasing<sup>12</sup>. There is a need to develop straightforward processes whereby the phase of alleles linked in series can be established to improve clinical outcomes. The ability to resolve phase may have relevance beyond HLA typing; HLA typing may also be used in healthcare areas besides transplantation, specifically in drug hypersensitivity<sup>13</sup> but also more widely in disease association genomics<sup>14</sup>. The value of higher resolution typing will be able to be assessed with greater precision as further typing is done.

Physical separating parentally inherited homologous chromosomes simplifies sequencing in the laboratory, facilitates analysis and clinical decision-making<sup>14</sup>. Amplification products can undergo standard high throughput sequencing of each of the two chromosomes inherited by a single individual, or alternatively, can be genotyped directly using SNP arrays. The method can be used to haplotype any individual, avoiding the need for segregation studies through extensive family genotyping. The method described below can be applied to any chromosome pair, and we have been successful in sorting a single parental copy of all chromosomes in a single FACS experiment (data not shown). Due to the critical need of haplotyping in HSCT, we chose to exemplify the method for this manuscript using chromosome 6 (Fig. 1).

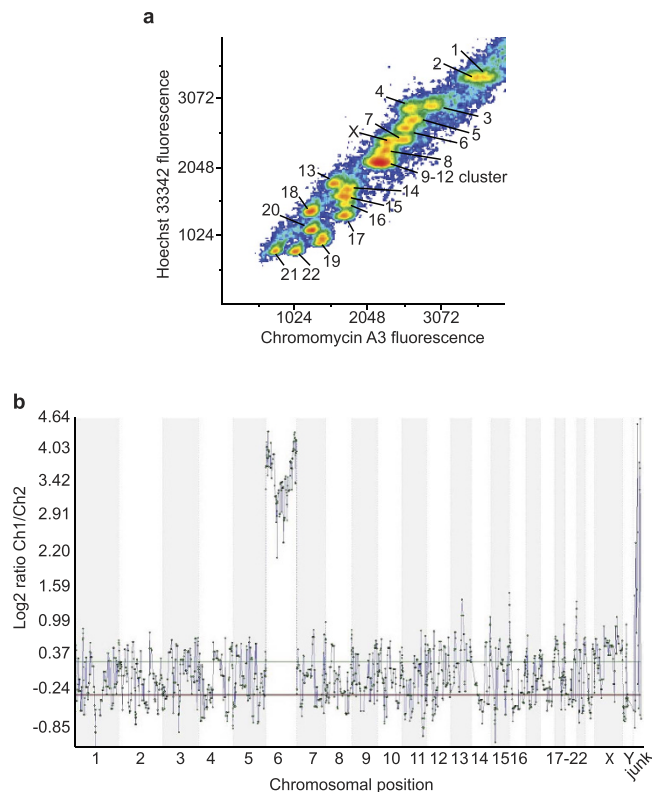
## Results

We haplotyped the most relevant HLA alleles by single chromosome isolation using bivariate flow cytometry followed by whole chromosome amplification. A prospective bone-marrow donor was HLA typed at the Victorian Transplantation and Immunogenetics Service in 2008 using genomic DNA sequence based typing<sup>15,16</sup>. The donor is heterozygous at all loci with the exception of HLA-A, thus haplotype phasing from the blood of this donor provided a convincing validation of the method. The following alleles at the HLA loci were assigned: *HLA-A*\*01:01; *HLA-B*\*08:01,27EKN (\*27:05/\*27:13); *HLA-C*\*01:02,07CVAG (\*07:01/\*07:06/\*07:18/\*07:52); *HLA-DRB1*\*01:01,\*03:01; *HLA-DRB3*\*01:01; *HLA-DQB1*\*02:01,\*05:01; *HLA-DPB1*\*01:01,04BDKS (\*04:02/\*06:02). The chromosome suspension was prepared directly from blood. Blood cells were stimulated with phytohemagglutinin, arrested in metaphase, lysed, and the chromosomes were stabilised in a polyamine buffer. The addition of DNA base-specific intercalating dyes enabled bivariate identification of chromosomal homolog clusters with a flow cytometer. The resolution of the bivariate flow karyotype was tested one day after cell lysis and a high quality flow karyotype was obtained with each chromosome cluster clearly resolved except for the chromosome 9–12 cluster (Fig. 2a), typical for human bivariate flow karyotypes<sup>17</sup>. Chromosomes in suspension were stored for up to 90 days at 4 °C prior to single chromosome acquisition. The older preparations had decreased cluster resolution with overlap in the periphery of neighbouring chromosomal clusters compared to one day old preparations, but nevertheless were resolvable. Single signals of the bivariate flow karyotype chromosome 6 cluster were sorted into separate PCR tubes. In order to confirm resolution of chromosomes, 48 copies of single chromosome 6 were acquired into separate PCR tubes. Thirty six samples were amplified with the PicoPLEX whole genome amplification Kit (Rubicon Genomics, USA) and 24 samples were randomly selected for amplification. Of these, five samples were found to contain only fragments of chromosome 6. Samples were then run on 24 Sure (+) BAC microarray (BlueGnome, Cambridge) to select samples containing DNA from flow-sorted then amplified chromosome 6 samples. Five of the 24 samples contained amplified DNA from chromosome 6 without any non-chromosome 6 DNA (Fig. 2b). To determine their individual identity from one or the other parent, these samples containing chromosome 6 DNA were used for microsatellite testing to compare individual chromosome 6 samples against diploid genomic DNA. Microsatellite maker testing at the polycystic kidney and hepatic disease 1 (PKHD1) locus (6p12.2) was undertaken using a custom multiplex mix for the gene and flanking microsatellites. The donor genomic DNA contained three loci with heterozygote peaks that were suitable for determining flow-sorted-then-amplified samples containing single chromosome 6 (i.e. from one parent). Four of the flow-sorted-then-amplified chromosome 6 samples had homozygous peaks of the same size that matched one of the two dual peaks identified from diploid genomic DNA. An additional sample had homozygote peaks at sizes matching the other peaks identified with the genomic DNA samples, i.e. peaks that were entirely distinct from those identified in the other flow-sorted then amplified samples. This indicated to us that one sample was a flow-sorted and amplified chromosome 6 inherited from one parent and four samples were flow-sorted and amplified chromosome 6 inherited from the donors other parent.

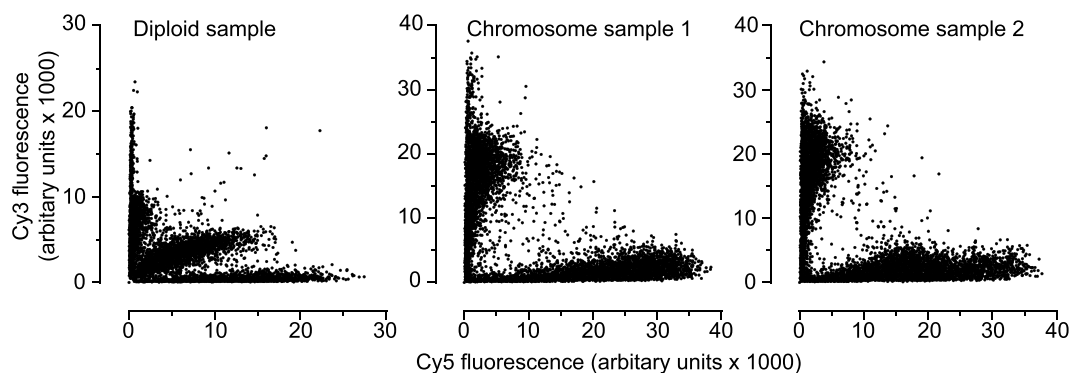
SNP genotyping arrays were performed using amplification products to confirm the parental chromosome source (Fig. 3). SNPs from the diploid genomic DNA of the donor were compared to SNP calls from the whole chromosome amplification products on a SNP array. Two examples identified as single chromosome 6 samples were confirmed as being of separate parental origin, and these samples matched the donor genomic DNA with microsatellite markers. The SNP calling software designates each polymorphic base as either A or B, thereby calling a SNP as one of AA or BB (homozygotes) or AB (heterozygote). In the diploid genomic DNA sample, the 33655 SNPs probed were called as 20.0% (6755) heterozygotes AB (central cluster) with 31.2% annotated as AA (vertical cluster) and 48.6% BB (horizontal cluster) (Fig. 3). Using an arbitrary raw fluorescence units (rfu) cut-off value of 5000, sample #1 called the respective A and B SNPs uniquely to its haplotype 97.5% (2572/2627 and 2584/2665) and sample #2 assigning 97.3% (2159/2217 and 2931/3012) uniquely (Fig. 3). False heterozygotes



**Figure 1. Molecular haplotyping of HLA: overview of the workflow.** A patient's white blood cells are purified and cultured. Addition of colcemid arrests cell division at metaphase, condensing chromosomes sufficiently to survive vigorous hypotonic cell lysis. Bivariate flow cytometry allows visual recognition of chromosome 6 which is then flow sorted into single PCR tubes. A whole genome amplification kit optimised for single cells yields sufficient material to confirm the presence of chromosome 6 via BAC and/or SNP array to discover the parental source, quality, and detection of contamination. Library generation and paired-end next-generation sequencing at 110X depth, yields sufficient coverage for HLA allele calling and haplotype phasing.

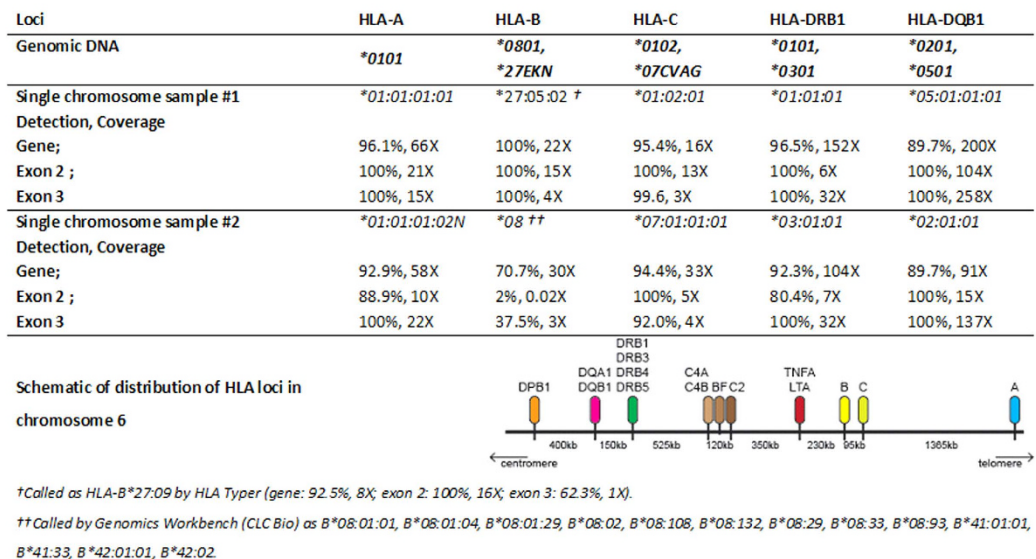


**Figure 2. Separation of chromosomes and confirmation of isolated chromosome 6 samples.** (a) Bivariate flow plot of single chromosome acquisition sorting. The X-axis indicates the fluorescence emitted by Chromomycin A3 specific binding to guanine-cytosine (C-G) regions of DNA, the Y-axis indicates the fluorescence emitted by Hoechst 33258 binding to adenine-thymine (A-T) regions of DNA. (b) 24 Sure (+) BAC microarray (BlueGnome, Cambridge) of chromosome 6 signal hybridised against male reference genomic DNA indicating the presence of chromosomal DNA from chromosome 6. Background noise for the remaining chromosomes indicates a lack of DNA from chromosomes 1–5, 7–22, X or Y.



**Figure 3. Cy3 and Cy5 fluorescence values obtained from Human Core Exome SNP array (Illumina) for 200 ng diploid gDNA (left), flow-sorted-then-amplified single chromosome 6 sample #1 (centre), and flow-sorted-then-amplified single chromosome 6 sample #2 (right).**

were called at 1.48% (498/33656 SNPs) for sample #1, and 1.03% (346/33656) for sample #2, enabling SNP haplotypes to be determined for 98.5% of SNPs for sample #1 and 99.0% for sample #2. Comparatively, the Theta allele calls were made in 99.9% of the genomic DNA variants, with Sample #1 at 76.8% (25835/33655) and Sample #2 at 64.8% (21803/33655) using GenomeStudio basecalling. Of the Theta allele SNPs called as heterozygote (6650 SNPs) from the genomic DNA, 42.2% (2771 SNPs) were differential homozygote calls (i.e. AA/BB, or BB/AA) on the amplified single chromosomes, with 8.9% (599 SNPs) corresponding to bases on both amplified samples and 50.1% missing due to allele dropout. Theta allele SNPs called as homozygote on the genomic DNA, 18.9% (5100/26999 SNPs) unavailable for analysis due to non-calling of one or both amplified samples, were matched in 52.0% (14029/26999) SNPs across the two amplified samples and discordant in 29.1% of calls (Supplementary



**Figure 4. Allele typing of transplant-associated loci using sequence specific typing (genomic DNA) compared to single chromosome amplification and sequencing of the isolated parental chromosome 6 samples.** Alleles were confirmed by sequencing. The loci used are within exon 2 and 3 of HLA Class I genes (*HLA-A*, *HLA-B* and *HLA-C*) and HLA Class II genes (*HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1*). The percentage values are the percentage of the gene/feature covered by the aligned sequence data. Therefore, it is possible for exon 2 and 3 to be 100% covered, but the gene to only be 89.7% covered, due to poor coverage in another part of the gene.

Fig. S1. It is foreseeable that optimisation of the amount of amplified single chromosome product loaded on the SNP array can markedly improve the basecalls. However, the results indicated to us that the full length of each flow-sorted-then-amplified chromosome of was of distinct parental origin and was amplified and not lost during sorting.

The two samples of single flow-sorted-then-amplified chromosome 6 that were identified as being from each parent were used to generate libraries for sequencing. To analyse the HLA region, we applied software specifically designed for HLA typing, namely Omixon HLA Typer. After sequencing (average of 44X coverage per library), sample #1 successfully called at low-resolution *HLA-A\*01*, *DQB1\*05*, *HLA-DRB1\*01* and for sample #2, *HLA-A\*01*, *HLA-C\*07*, *HLA-DQB1\*02*, *HLA-DRB1\*03* corresponded with standard alleles determined by sequence based typing. To improve sensitivity on alleles unresolved in the previous analysis, depth of coverage was increased to 110X, on average, per library. As a result, sample #1 yielded 6/6 calls corresponding to sequence based typing while sample #2 yielded 5/6 of the alleles. Accuracy of sequencing reads was determined by aligning against conventional sequence based typing of the six transplant matching loci. There was a high degree of variability of depth of coverage over the key HLA regions used in transplant matching and across the haplotype (Fig. 4). For instance, the exon 2 region of MHC class II genes (*HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1*) and exon 2 and exon 3 of MHC class I genes (*HLA-A*, *HLA-B* and *HLA-C*) are used in PCR-based matching and the gene coverage for these regions ranged from 70 to 100% (Fig. 4). The coverage of the *HLA-B* locus in exons 2 and 3 was particularly low for single chromosome sample #2, which was possibly a consequence of the combination of a degree of degradation of sample #2 that was also reflected in the SNP array calls, and shorter fragment lengths of amplifying *HLA-B* for the single-cell amplification kit<sup>18</sup>. Omixon HLA Typer failed to assign the *HLA-B* locus on single chromosome sample #2, presumably due to low coverage at exon 2. We therefore used CLC Genomics Workbench to call *HLA-B* alleles for both sample #1 and 2. It called sample #1 as an alternative *HLA-B\*27* allele *HLA-B\*27:05:02* (*HLA-B\*27:09* for Omixon HLA Typer) and sample #2 as the ambiguous group: *B\*08:01:01*, *B\*08:01:04*, *B\*08:01:29*, *B\*08:02*, *B\*08:108*, *B\*08:132*, *B\*08:29*, *B\*08:33*, *B\*08:93*, *B\*41:01:01*, *B\*41:02:01*, *B\*41:33*, *B\*42:01:01* or *B\*42:02*, the *B\*08* being concordant with the allele assignment via sequence based typing. Pooling sequencing runs (154X depth of coverage) did not appreciably improve allele calling, possibly indicating allele dropout at key exons during amplification. Haplotypes were assembled using allele calls from HLA Typer (Table 1). Although *HLA-B\*08* did not have sufficient depth of coverage to be successfully called with Omixon HLA Typer, CLC Genomics Workbench successfully resolved the allele as *HLA-B\*08*, which was less discriminating than the assignment made by sequence based typing. Single chromosome sample #1 was identified as matching the A1-B8-DR3 haplotype<sup>19</sup>. As previously mentioned, the relatively low sequencing depth of coverage at the key exons of *HLA-B\*08* and *HLA-B\*27*, is possibly due to dropout during amplification and requires further exploration. To validate the phasing we used HLA\*IMP:02, to impute alleles directly from the HumanCoreExome array. Allele assignment of the genomic DNA was predominantly accurate, though due to the low density of the HumanCoreExome over the MHC region, *HLA-A* and *HLA-DPB1* were unable to be imputed. *HLA-B\*08:01*, (*HLA-B\*27:05* called falsely as *HLA-B\*07:05*), *HLA-C\*01:02*, *\*0701*,

Locus	Sample #1	Sample #2
<i>HLA-F</i>	*01:01:03:01 (64X)	*01:01:03:01 (64X)
<i>HLA-V</i>	*01:01:01:01 (20X)	*01:01:01:01 (32X)
<i>HLA-G</i>	*01:01:02:01 (105X)	*01:01:02:01 (80X)
<i>HLA-H</i>	*02:01:01:01 (88X)	*02:01:01:01 (69X)
<i>HLA-K</i>	*01:01:01:01 (114X)	*01:01:01:01 (63X)
<i>HLA-A</i>	*01:01:01:01 (64X)	*01:01:01:02N (58X)
<i>HLA-J</i>	*01:01:01:01 (86X)	*01:01:01:02 (71X)
<i>HLA-L</i>	*01:01:02 (109X)	*01:01:02 (47X)
<i>HLA-E</i>	*01:01:01:01 (219X)	*01:01:01:01 (218X)
<i>HLA-C</i>	*01:02:01 (16X)	*07:01:01:01 (32X)
<i>HLA-B</i>	*27:05:02 (22X)	*08 (30X)
<i>MICA</i>	*001 (112X)	*008:01:01 (94X)
<i>MICB</i>	*005:02:02 (106X)	*005:02:02 (115X)
<i>HLA-DRA</i>	*01:01:02 (60X)	*01:02:02 (83X)
<i>HLA-DRB9</i>	*09:01 (33X)	*01:01 (30X)
<i>HLA-DRB2</i>	—	*01:01 (46X)
<i>HLA-DRB3</i>	—	*01:01:02:01 (109X)
<i>HLA-DRB6</i>	*06:01 (2X)	—
<i>HLA-DRB1</i>	*01:01:01 (152X)	*03:01:01 (104X)
<i>HLA-DQA1</i>	*01:01:01 (52X)	*05:01:01:02 (108X)
<i>HLA-DQB1</i>	*05:01:01:02 (200X)	*02:01:01 (91X)
<i>HLA-DOB</i>	*01:01:03:01 (169X)	*01:01:01:02 (191X)
<i>HLA-TAP2</i>	*01:01:02 (136X)	*02:01:02:01 (126X)
<i>HLA-TAP1</i>	*03:01 (202X)	*01:02N (112X)
<i>HLA-DMB</i>	*01:01:01:01 (87X)	*01:01:01:02 (81X)
<i>HLA-DMA</i>	*01:01:01:01 (33X)	*01:01:01:02 (42X)
<i>HLA-DOA</i>	*01:01:02:03 (85X)	*01:01:05 (81X)
<i>HLA-DPA1</i>	*01:03:01:05 (80X)	*02:01:02 (78X)
<i>HLA-DPB1</i>	*04:02:01:02 (306X)	*01:01:01:01 (35X)

**Table 1. Allele typing of the individual generated by HLA Typer (Omixon).** Numbers in brackets indicate the depth at which the mean coverage was assigned per locus.

*HLA-DRB1*\*01:01,\*03:01, *HLA-DQA1*\*01:01,\*05:01, *HLA-DQB1*\*05:01\*02:01 were correctly identified. Sample 1 was phased as *HLA-C*\*01:02, *HLA-DRB1*\*01:01, *HLA-DQA1*\*01:01, *HLA-DQB1*\*05:01 with Sample 2 only resolving the *HLA-DRB1*\*03:01 allele.

## Discussion

To fully validate the haplotypes, a family study would provide additional depth to the study<sup>8</sup>. The data presented here indicating highly disparate SNP calls and HLA allele imputation gives confidence that the two chromosomes selected for high throughput sequencing were of separate parental origin, but also emphasises the limitations of SNP arrays for precise phasing. Chromosome separation offers genuine phasing and there is no possibility of recombination between the two chromosomes during the amplification steps. A further advantage of chromosome separation is that phasing is chromosome wide and therefore haplotypes do not need to be imputed or rely on typing parental samples. One of the haplotypes identified in this study was the most common Caucasian haplotype A1-B8-DR3-DQ2<sup>19,20</sup>, which simplified analysis. With the allele calling of the sequence-based typing contiguous to the high throughput sequencing allele calls, the combined results indicate proof of successful haplotype phasing of HLA. A turnaround time of approximately one week from blood extraction to phased sequenced data is feasible using this approach (Fig. 1) and highly desirable for urgent cases in transplantation laboratories.

The method described addresses a crucial problem in clinical HSCT when using unrelated donors, which is the ability to discriminate between allele-matched or haplotype-matched donor-recipient pairs. Despite evidence that haplotype matching gives superior clinical outcomes there is currently no method that can routinely provide haplotype matching<sup>21</sup>. An added benefit of the approach described in this study is the ability to unambiguously define extended haplotypes which include polymorphic non-HLA genes and future directions of the research could include measurement of linkage disequilibrium between the observed SNP array calls and the HLA alleles assigned. Conversely, the high degree of redundancy in the number of unused reads in order to elucidate loci such as *HLA-B* indicates that approaches aimed to either specifically amplify the MHC or limit the amplification of the non-MHC sequences would increase the economy of the sequencing depth of coverage for the desired regions. Alternatively, amplification of single chromosomes using long-range amplification followed by locus-specific PCR and sequencing would allow alleles to be individually called and haplotypes generated in the context of transplant HLA-matching<sup>22</sup>. The approach applied here has been previously proposed as an effective means to produce a fully phased genome<sup>1</sup>. Phasing by dilution to ultra-low copy numbers has been shown as an effective

method for genome phasing<sup>23</sup>. However, due to the depth of sequencing required to resolve alleles indicated in this study, phasing by dilution may not be well-suited to directly haplotyping HLA.

In conclusion, this study has shown that complete HLA haplotype phasing can be achieved by amplifying single chromosomes and performing deep high-throughput sequencing. This study attempted to address the difficulty of assigning HLA haplotypes in the absence of family studies by applying an established but under-utilised technique, which if developed further, would be of immense benefit to the broader landscape of genetics. Implications that flow from this study extend far beyond the field of clinical transplantation and include the ability to phase multiple SNPs across the genome. This would provide insight into SNPs associated in a range of diseases where SNP phase has not been determined. Reporting of haplotype information will provide clinicians, patients and medical researchers with a better understanding of cis and trans-interactions in functional polymorphisms and their impact on disease susceptibility and progression, leading in turn to more informed treatment options.

## Methods

**Human blood samples.** All experimental protocols involving human blood samples were approved by the ethics committee of the Australian Bone Marrow Donor Registry (ABMDR: Project 2011/03). Methods for handling of blood samples were carried out in accordance with the approved guidelines. Informed consent was obtained from all subjects by the ABMDR.

**Chromosome suspension.** The blood sample (50 mL) was collected in lithium heparin tubes from an individual registered with the Australian Bone Marrow Donor Registry (ABMDR). To generate a chromosome suspension directly from blood, white cells were separated by a density gradient centrifugation method using Histopaque 1077 (Sigma USA) and centrifugation at  $1000 \times g$  for 10 min. White blood cells were then aspirated with a Pasteur pipette and washed twice by resuspending in 50 mL phosphate buffered saline (PBS) and centrifuged at  $250 \times g$  for 5 min and once in 50 mL of RPMI 1640 (Life Technologies, USA). The pellet was then suspended in 40 mL RPMI 1640 with 1 mL phytohemagglutinin (Gibco, USA) and 100  $\mu$ L of fetal calf serum in a T75 flask and cultured at 37 °C, 5% CO<sub>2</sub> for 72 hours. 400  $\mu$ L of 10 g/mL colcemid (Life Technologies, USA) was added and after 6 hours, the cells were harvested by washing twice with PBS and centrifuging at  $250 \times g$  for 5 mins. The pellet was gently suspended in 15 mL of 0.075 M KCl solution at 37 °C for 20 mins followed by centrifugation at  $250 \times g$  for 5 mins. The pellet was gently suspended in freshly prepared 3 mL of ice-cold polyamine buffer (0.5 mM spermidine (Sigma), 0.2 mM spermine (Sigma), 5 mM EGTA, 20 mM EDTA, 150 mM Tris, 0.25% Triton-X 100, 30 mM DTT and 10 mM MgSO<sub>4</sub>)<sup>24</sup>. After 5 mins the cell suspension was vigorously vortexed for 15 seconds, releasing chromosomes into suspension. The suspension was then stored at 4 °C for subsequent sorting by bivariate flow cytometry.

**Chromosome sorting.** In order to identify the chromosome clusters, 16 hours before flow cytometry, 10  $\mu$ L of 1 mg/mL DNA stain Hoechst 33258 and 32  $\mu$ L of 2.5  $\mu$ g/mL of Chromomycin A3 (that preferentially bind adenine-thymine and guanine-cytosine rich DNA sequences respectively), were added to 2 ml of chromosome suspension. One hour prior to sorting 10  $\mu$ L sodium citrate (1 M) and 50  $\mu$ L of sodium sulphite (500 mM) was added to facilitate chromosome condensing and stability, prior to filtering through a 20  $\mu$ m filter (Partec) to eliminate cellular debris. The flow karyotype was generated on a MoFlo cell sorter (Beckman Coulter) using a 70  $\mu$ m nozzle and a sheath pressure of 56 psi. Chromomycin A3 was excited with an Argon Ion laser tuned to 457.9 nm with power set to 300 mW. Hoechst 33258 was excited with a solid state laser emitting at 355 nm with power set at 300 mW. Chromomycin A3 fluorescence was collected using a 490LP filter and Hoechst 33258 fluorescence with a 400SP and 457.9 nm notch laser blocking filter. A sorting gate was placed over the chromosome 6 cluster and a single chromosome 6 was sorted into each PCR tube.

**Single chromosome amplification.** DNA amplification of the single chromosomes was performed using the single cell whole genome amplification PicoPLEX Kit (Rubicon Genomics, USA) as per manufacturer's instructions. Half of the amplification product was then purified using a GeneJet PCR purification kit (Thermo Scientific, USA) for library preparation, and the non-purified product was retained for SNP array testing.

**BAC Array.** 24 Sure (+) BAC microarray (BlueGnome, Illumina, USA) was used for single chromosome 6 screening, described by Tobler *et al.*<sup>25</sup>. Briefly, DNA samples were labelled with random primers, hybridised to a BAC microarray with Cy3, scanned on InnoScan 710 (Innopsys) and reference compared against a euploid male labelled with Cy5. Raw fluorescence was measured by Mapix (Innopsys) and signal strength normalised into the presence of chromosomal DNA by Bluefuse Multi (BlueGnome, Illumina, USA).

**SNP Array.** 200 ng of the amplified product was used on the HumanCoreExome BeadChip array. An arbitrary cut-off value for SNP calls was set at  $\geq 5000$  relative fluorescent units (rfu), i.e. 'AA' if fluorescence  $> 5000$  rfu for 'A' base and  $< 5000$  rfu for 'B' base, 'AB' if fluorescence  $> 5000$  rfu for 'A' and 'B' base, 'BB' if fluorescence  $< 5000$  rfu for 'A' base and  $> 5000$  rfu at 'B' base.

**Sequencing.** Amplified products selected for library preparation were quantified via TapeStation (Agilent, USA) sequencing using Quant-IT PicoGreen (Invitrogen, USA) against a lambda DNA standard. Libraries of 1 ng of dsDNA product were used with the Nextera XT DNA Sample Preparation kit (Illumina, USA), size selected at 444 bp (from 214 – 775 bp) and 392 bp (from 167–737 bp) for each sample and sequenced using 100 bp paired-end on a HiSeq2500 (Illumina, USA). Sequencing on a single lane yielded 96,839,980 and 103,733,615 paired reads for the two libraries.

**Software.** HLA typer (Omixon) was used to assemble paired-end reads against reference HLA-alleles<sup>26</sup>. Alleles were called for each loci as the allele having the highest mean depth of coverage across each gene region. The maximum fragment size was fixed at 600 bp, multi-mapping, *i. e.* reads that mapped to multiple genes were discarded, as were rare alleles. For the *HLA-B* gene, sequencing reads were also aligned with CLC bio genomics workbench (CLC Genomics Workbench 7.3 (<http://www.clcbio.com>)). In house software (HLA Allelecaller) used these alignments combined with the reference data in the international ImMunoGeneTics information system (IMGT) database to assign all alleles concordant with the data from exons one to four inclusive<sup>27</sup>. HLA\*IMP:02 was used to perform allele imputation from HumanCoreExome BeadChip array data<sup>28</sup>. The genotyping results from the HumanCoreExome-12v1-0\_B were exported as PLINK input report files from Illumina GenomeStudio (v2010.4.0.301128). The ped and map files were read into PLINK (v1.07)<sup>29</sup>. The `-make-bed` command was used to create a binary file. This file was read into PSEQ<sup>30</sup> and subsequently exported as a VCF. The VCF file was edited to VCF format v4.0 and read into the HLA\*IMP front end. The file underwent conversion to oxford HLA standard format, and missing data was removed (with a threshold of 0.2). The resulting file was loaded into the web interface of HLA\*IMP and all loci except HLA A and HLA DPB1 were imputed.

## References

1. Yang, H., Chen, X. & Wong, W. H. Completely phased genome sequencing through chromosome sorting. *Proc. Natl. Acad. Sci. USA* **108**, 12–17 (2011).
2. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nature Rev. Genet.* **12**, 215–223 (2011).
3. Warren, E. H. *et al.* Effect of MHC and non-MHC donor/recipient genetic disparity on the outcome of allogeneic HCT. *Blood* **120**, 2796–2806 (2012).
4. Beksac, M. & Dalva, K. Role of killer immunoglobulin-like receptor and ligand matching in donor selection. *Bone Marrow Res.* **2012**, 271695, doi: 10.1155/2012/271695 (2012).
5. Ruggeri, L. *et al.* Role of natural killer cell alloreactivity in HLA-mismatched hematopoietic stem cell transplantation. *Blood* **94**, 333–339 (1999).
6. Sheldon, S. & Poulton, K. In *Methods In Molecular Biology™* Vol. 333 (eds P. Hornick & M. Rose) Ch. 7, 157–174 (Humana Press, 2006).
7. Dorak, M. T. *et al.* Conserved extended haplotypes of the major histocompatibility complex: further characterization. *Gene Immunity* **7**, 450–467 (2006).
8. Handyside, A. H. *et al.* Karyomapping: a universal method for genome wide analysis of genetic disease based on mapping crossovers between parental haplotypes. *J. Med. Genet.* **47**, 651–658 (2010).
9. Joris, M. M. *et al.* The impact of frequent HLA haplotypes in high linkage disequilibrium on donor search and clinical outcome after unrelated haematopoietic SC. *Bone Marrow Transplantation* **48**, 483–490 (2013).
10. Petersdorf, E., Malkki, M., Gooley, T., Martin, P. & Guo, Z. MHC haplotype matching for unrelated hematopoietic cell transplantation. *PLoS Med.* **4**, e8 (2007).
11. Tay, G. K. *et al.* Matching for MHC haplotypes results in improved survival following unrelated bone marrow transplantation. *Bone Marrow Transplantation* **15**, 381–385 (1995).
12. Wittig, M. *et al.* Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nuc. Acids Res.* **43**, e70 (2015).
13. Phillips, E. J. & Mallal, S. A. Pharmacogenetics of drug hypersensitivity. *Pharmacogenomics. Pharmacogenomics* **11**, 973–987 (2010).
14. Crawford, D. C. & Nickerson, D. A. Definition and clinical importance of haplotypes. *Annu. Rev. Med.* **56**, 303–320 (2005).
15. Hoppe, B. & Salama, A. In *Methods in Molecular Medicine* (ed M. Beksac) Ch. 6, 71–79 (Humana Press, 2007).
16. Smith, L. K. In *Methods in Molecular Biology* Vol. 882 (eds F. T. Christiansen & B. D. Tait) Ch. 5, 67–86 (Humana Press, 2012).
17. Lebo, R. V. *et al.* High-Resolution Chromosome Sorting and DNA Spot-Blot Analysis Assign McArdle's Syndrome to Chromosome 11. *Science* **225**, 57–59 (1984).
18. Treff, N. R., Su, J., Tao, X., Northrop, L. E. & Scott, R. T. Single-cell whole-genome amplification technique impacts the accuracy of SNP microarray-based genotyping and copy number analyses. *Molecular Human Reproduction* **17**, 335–343 (2011).
19. Horton, R. *et al.* Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project. *Immunogenetics* **60**, 1–18 (2008).
20. Larsen, C. E. *et al.* Dominant Sequences of Human Major Histocompatibility Complex Conserved Extended Haplotypes from HLA-DQA2 to DAXX. *PLoS Genetics* **10**, e1004637 (2014).
21. Witt, C. *et al.* Unrelated donors selected prospectively by block-matching have superior bone marrow transplant outcome. *Human Immunol.* **61**, 85–91 (2000).
22. Garcia, M. A. A., Yebra, B. G., Flores, A. L. L. & Guerra, E. G. The Major Histocompatibility Complex in transplantation. *J. Transplantation* **2012**, 842141 (2012).
23. Peters, B. A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190–195 (2012).
24. Carter, N. P. In *Cell Biology* Ch. 19, 133–135 (Elsevier Science, 2006).
25. Tobler, K. J. *et al.* Two different microarray technologies for preimplantation genetic diagnosis and screening, due to reciprocal translocation imbalances, demonstrate equivalent ploidy and clinical pregnancy rates. *J. Assisted Reprod. Genet.* **31**, 843–850 (2014).
26. Major, E., Rigó, K., Hague, T., Bérces, A. & Juhos, S. HLA typing from 1000 genomes whole genome and whole exome Illumina data. *PLoS One* **8**, e78410 (2013).
27. Robinson, J. *et al.* The IPD and IMGT/HLA database: allele variant databases. *Nuc. Acids Res.* **43**, D423–D431 (2015).
28. Dilthey, A. *et al.* Multi-population classical HLA type imputation. *PLoS Comp. Biol.* **9**, e1002877 (2013).
29. Purcell, S. *et al.* PLINK: A Tool set for whole-genome association and population-based linkage analyses. *Am. J. Human Genet.* **81**, 559–575 (2007).
30. PLINK/SEQ. PLINK/SEQ: A library for the analysis of genetic variation data. v0.10, released 14-July-2014. Date accessed 12/10/2015. <https://atgu.mgh.harvard.edu/plinkseq/>.

## Author Contributions

N.M.M. designed and performed experiments, performed bioinformatics, analysed data and wrote the paper; M.B. performed flow cytometry and chromosome acquisition; M.J.H. performed experiments, D.R.P., F.J.R., D.C., A.C. and D.C.S., gave technical support and conceptual advice, performed bioinformatics and experimental design on sequencing data L.G. and M.D.P. designed and performed experiments and performed bioinformatics on SNP arrays, B.D.T., gave technical support and conceptual advice, designed experiments and wrote the paper, H.R.I. and C.W.P. supervised the project, designed experiments, analysed data and wrote the paper.



## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Murphy, N. M. *et al.* Haplotyping the human leukocyte antigen system from single chromosomes. *Sci. Rep.* **6**, 30381; doi: 10.1038/srep30381 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016