Identifying novel transcriptional regulatory elements of *HLA-A* alleles through the evaluation of the 5' un-translated region sequences

by

Saiyuri Singh

(214500828)

Submitted in fulfillment of the requirements for the degree of Master of Medical Science (Virology) in the School of Laboratory Medicine and Medical Sciences, University of KwaZulu-Natal

Declaration

I, Miss Saiyuri Singh, declare as follows:

1. That the work described in this thesis has not been submitted to UKZN or other tertiary institution for purposes of obtaining an academic qualification, whether by myself or any other party.
2. That my contribution to the project was as follows:
   - All literature reviews and writing were done by me
   - All laboratory procedures were performed by me independently namely:
     - Western blot
     - Nuclear and cytoplasmic extraction
     - Agarose gel electrophoresis
     - Electrophoretic mobility shift assay and optimization
     - Chromatin immunoprecipitation assay and optimization
     - Cell culture of Raji and THP-1 cell lines
     - Polymerase chain reaction
     - Real time polymerase chain reaction, as well as
     - DNA and RNA extraction
   - I performed sequence analysis using online predictive software for transcription factor binding (AliBaba2.1, CTCFBSDB2.0)
   - All unpaired t-tests were done by me using GraphPad Prism 8
3. That the contributions of others to the project were as follows:
   a. Dr. Veron Ramsuran (supervisor): came up with the project and the idea of using chromatin immunoprecipitation. Discovered the putative CTCF binding site and conducted sequence analysis.
   b. Dr. Smita Kulkarni: allowed me to use her lab in Texas, conducted training on sequence analysis and suggested the use of the electrophoretic mobility shift assay to get initial results.
   c. Prof. Thumbi Ndung'u: supplied FRESH samples
   d. Dr. Mary Carrington: provided sequences and expression data that formed the basis for this project
   e. Mr Hoang Vinh Nguyen: assisted in the optimization of protocols and generation of ideas for the write-up and conclusions
   f. Dr. Ravesh Singh: assisted in the optimization of protocols
   g. Miss Mishka Danielle Muthen: assisted in the optimization of protocols and generation of ideas for the write-up and conclusions
   h. Mrs. Kimone Fisher: assisted in the optimization of protocols and generation of ideas for the write-up and conclusions

Signed.  Date: _____18/01/2020_____

Candidate

As the candidate's Supervisor, in addition to the above, I confirm and agree to the submission of this dissertation

Signed  Date: _____18/01/2020_____

Supervisor

Dedication

To Dr. Suvira Ramlall

For being my guardian angel through this Masters journey, I could not have come this far without your love, support and guidance

To Mishka Danielle Muthen, Kimone Fisher and Lisa Naidoo

For sticking together, working as a team and supporting each other through every hurdle on our academic journey, thank you for teaching me all that you did

Acknowledgements

I'd like to acknowledge the following people and organizations:

Mr. Mahomed Ahmed for kindly giving me THP-1 cells and teaching me how to culture them

Dr. Gila Lustig for generously giving me Raji cells

Prof. Faizal Bux for allowing the use of the sonicator and molecular lab at the Institute for Water and Wastewater Technology at the Durban University of Technology

Dr. Ravesh Singh for allowing the use of his lab for cell culture and real time polymerase chain reaction

Dr. Richard John Lessells, Dr. Sinaye Ngcapu and Dr. Eduan Wilkinson for being my thesis advisors for my Masters

The National Research Foundation (NRF) for funding me during my Masters Research

The Sub-Saharan Network for TB/HIV Research Excellence (SANTHE) for funding my project as well as my training in Texas

Presentations

Oral Presentations:

- Singh S, Muthen M.D, Fisher K.L, Nguyen H.V, Kulkarni S, Ramsuran V. Identifying *HLA-A* regulatory factors. School of Laboratory Medicine and Medical Science Research Day 2018, University of KwaZulu-Natal, Westville Campus. 29 August 2018.
- Singh S, Muthen M.D, Fisher K.L, Nguyen H.V, Kulkarni S, Ramsuran V. Identifying *HLA-A* regulatory factors. Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE) Annual Research Day 2019, Westville Country Club. 3-4 June 2019
- Singh S, Muthen M.D, Fisher K.L, Nguyen H.V, Kulkarni S, Ramsuran V. Effect of a polymorphic CTCF binding site on *HLA-A* mRNA expression. College of Health Sciences Research Symposium 2019, Nelson R Mandela School of Medicine. 1 November 2019.

Poster Presentations

- Singh S, Muthen M.D, Fisher K.L, Nguyen H.V, Kulkarni S, Ramsuran V. Identifying *HLA-A* regulatory factors. Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE) Annual Meeting 2019, Nairobi, Kenya. 3-5 October 2019.
- Singh S, Muthen M.D, Fisher K.L, Nguyen H.V, Kulkarni S, Ramsuran V. Identifying *HLA-A* regulatory factors. 9th Annual Infectious Diseases in Africa Symposium 2019, River Club, Cape Town. 7-14 October 2019.

Table of contents

List of figures and tables

Figures:

Tables:

Abstract

Sub-Saharan Africa holds approximately half the population living with human immunodeficiency virus (HIV) in the world (~19.6 million), of which around 7.2 million cases are found in South Africa. Although antiretroviral therapy can suppress viral loads to below detectable levels in most cases, drug resistance is a growing problem. Therefore, identifying novel treatment strategies are warranted against HIV. The strongest human genetic associations with HIV disease have been found within the human leukocyte antigen (HLA) region. The expression levels of various *HLA* genes have been associated with HIV disease outcomes. Increased *HLA-A* mRNA expression results in poor HIV outcomes due to the inhibition of natural killer (NK) cells since high mRNA expression of *HLA-A* results in high protein expression of HLA-E which serves as an inhibitory receptor for NK cells. Identifying factors that regulate the expression of *HLA-A* has the potential to serve as an avenue for HIV drug target sites. DNA methylation has previously been identified as one of the factors responsible for *HLA-A* expression regulation. In this study, we aimed to identify additional regulatory mechanisms for the *HLA-A* gene. The identification of a putative CCCTC-binding factor (CTCF) binding site upstream of *HLA-A* suggested that CTCF may play a role in regulation of *HLA-A*. Sequence alignments about 2 kilobases (2KB) upstream of the transcriptional start site (TSS) were analysed for polymorphisms that associate with *HLA-A* expression. Six *HLA-A* promoter variants (*rs9260084, rs9260086, rs9260092, rs9260101, rs9260116* and *rs41560714*) were observed to significantly associate with *HLA-A* mRNA expression. However, only one single nucleotide polymorphism (SNP), *rs9260084* (-993G>A), was predicted to disrupt a CTCF binding site. Despite the predicted disrupted binding site, using a chromatin immunoprecipitation (ChIP) assay, we did not detect any difference in CTCF binding across the -993 G>A variants. Additional transcriptional regulators, Nuclear Factor 1 (NF1), Ras related protein (RAP1) and glucocorticoid receptor (GR), were predicted to have differential binding to -993G>A, -226G>A and -885C>G, respectively. The results provided here serve as a basis for further studies exploring the role *HLA-A* promoter variants have in regulating *HLA-A* expression. These variants may serve as potential target sites for future therapeutic intervention against HIV.

Chapter 1

Introduction

Since the first diagnosis of HIV in humans in the 1980's, over 70 million people globally have been infected with the virus. Approximately half that number have died from acquired immunodeficiency syndrome (AIDS) related causes [1]. At the end of 2017, an estimated 36.9 million people were living with HIV on a global scale [1]. Half of these infected individuals (approximately 19.6 million) reside within Sub-Saharan Africa [2].

South Africa has one of the highest burdens of HIV infection worldwide with approximately 7.7 million people living with HIV by 2019 [3]. South Africa hosts the largest HIV treatment programme globally with 80% of antiretroviral therapy (ART) treatment being funded by the government [4]. The number of new infections each year in South Africa remains high despite the fact that 5 million people were on treatment in mid-2019 [3].

HIV is transmitted via infected body fluids such as blood during blood transfusions, via the sharing of needles during intravenous drug use, via semen during sexual intercourse and from mother to child during breastfeeding [5, 6]. Following entry into the body, HIV attaches itself to receptors known as CD4 and C-C chemokine receptor type 5 (CCR5) or C-X-C chemokine receptor type 4 (CXCR4) most commonly on CD4+ T lymphocytes. HIV is an intracellular virus and causes cell lysis depleting the CD4+ cell count of the infected individual. CD4+ T lymphocytes play a crucial role in combatting infections. A low CD4+ T cell count is therefore associated with an increased risk of acquiring opportunistic infections [7]. Once an HIV infected individual is exposed to opportunistic infections and their CD4+ T cell count decreases, AIDS sets in. In 1996 ART became a standard of care treatment to control and reduce HIV viral load levels in the blood [8]. Treatment, in the form of ART has since been made available worldwide. ART, when adhered to, can suppress HIV viral loads but cannot cure the disease due to the establishment of viral reservoirs within immune-privileged sites such as the lymph node [9]. HIV is also able to lay dormant within a cell, a concept known as latency [10, 11].

Due to non-adherence the emergence of HIV drug resistant strains has become a concern regarding HIV treatment. Successful drug resistance testing on 697 HIV positive individuals from all age groups across South Africa showed that 22.8% of individuals who self-reported not taking ART and who tested negative for ART were infected with HIV drug resistant mutants [12]. In addition, of those who reported having taken ART and tested positive for ART, 55.7% had drug resistant HIV mutants [12]. It is therefore imperative that researchers investigate alternate avenues – such as host genetics – when updating HIV treatment regimens. Host genetics has shown a lot of promise in terms of HIV treatment and cure. Two people have been cured to date with the use of host genetics. A mutation

called Δ32 in the CCR5 gene results in loss of CCR5 expression on the cell surface [13]. If a person has two copies of Δ32 they are naturally resistant to HIV strains which are CCR5-tropic [13]. Two patients who had a form of white blood cell cancer – termed The London Patient and The Berlin Patient – underwent bone marrow transplants with donor bone marrow which had the double Δ32 mutation. Since the donor marrow was naturally resistant to CCR5-tropic HIV, ART was stopped in both patients and various body fluids were routinely tested for HIV [13, 14]. The Berlin Patient, who was deemed cured of HIV in 2008 [13], is still HIV free 11 years later and The London Patient was reported to be HIV free in March, 2019, 18 months after stopping ART [14].

Host genetics involves elucidating the impact of individual genes on disease outcome. Host genes have been used as targets for drugs against HIV. It is already known that HIV uses CCR5 as a receptor to gain access to CD4+ T cells. It is also known that the loss of CCR5 on the cell surface will prevent viral entry into the cell. CCR5 has therefore been used as a drug target for a class of ART known as CCR5 inhibitors [15]. The first CCR5 antagonist known as Maraviroc functions by binding to CCR5 therefore barring HIV from binding CCR5 [15, 16]. Other host proteins naturally inhibit HIV. For example, tetherin (also known as BST-2, HM1.24 or CD317) has been shown to inhibit the release of HIV virions from infected cells [17, 18], apolipoprotein B editing complex 3 (APOBEC3) cytidine deaminase has the ability to induce lethal mutations in the HIV genome [19] and tripartite motif 5-alpha (TRIM5α) proteins can block incoming retroviral capsids as they fuse with the host cell membrane [20]. Studies have identified a group of host genes as having the strongest association with HIV. These genes are the human leukocyte antigen (*HLA*) genes [21].

HLA genes are located within the major histocompatibility complex (MHC) region which is found on the short arm of chromosome 6 in humans (6p21.3) [22]. There are two distinct *HLA* classes known as *HLA* Class I and Class II. *HLA* class I genes are divided in classical and non-classical. The classical genes are *HLA-A, HLA-B* and *HLA-C*. The HLA class I molecules are expressed on the surface of all nucleated cells in humans. This differs from class II because class II is found only on professional antigen presenting cells (APC) e.g. dentritic cells, and on B cells. Class I genes have traditionally been shown to play a role in influencing the human immune response via the binding and presentation of antigenic peptides to various immune cells [23, 24]. Although the major function of HLA is presenting antigens to T cells, some HLA class I molecules serve as ligands for NK cells such as HLA-E which is a ligand for the inhibitory NK cell receptor NKG2A [25]. The relationship between HLA molecules and the immune response has led to various *HLA* genes associating with a diverse range of diseases including HIV [26, 27].

Individual *HLA* alleles are associated with increased or decreased risk of acquiring diseases based on the specific peptide that is presented to the immune system. [28]. These diseases include viral [25, 29-37], bacterial [38, 39], parasitic[39] and autoimmune [29, 40]. Different groups of *HLA* alleles confer

increased susceptibility towards or protection against HIV in different populations. An example is a study in which various *HLA* alleles were examined for HIV associations in an Argentinian population. The *HLA-B*18* and *B*39* groups of alleles were found to occur more frequently in HIV positive subjects (p=0.058 and 0.008 respectively, odds ratio=3.84 and 11.96 respectively) suggesting that those groups are associated with increased risk of acquiring HIV. The *B*44* and *B*55* groups were not found at all in HIV positive subjects (p=0.013 and 0.056 respectively, odds ratio=0 for both) which suggests that those groups may have a protective effect against HIV [41]. In Cameroon it was found that a pairing of *B*44* with *A*32* was significantly associated with protection from HIV (p=0.03) [42].

HLA Class I genes have exhibited strong associations with HIV outcomes [21]. *HLA* class I homozygosity was associated with accelerated HIV progression through narrow responses of cytotoxic T lymphocyte (CTL) [21, 43, 44]. *HLA-B*57, B*27, B*35-Px, B*51, Bw4, B*58:01, B*13* and *B*81:01* are some of the allele groups that associate with slow HIV disease progression [21]. In 2010, a genome wide association study (GWAS) reported that in Caucasians, the top hit SNP, *rs9264942* ($p=2.8 \times 10^{-35}$, OR 2.9), was found in the *HLA* region [45] which means this SNP had the strongest association with HIV compared to every other SNP analysed. Several other GWAS have identified SNPs within *HLA* loci as top hits in association with HIV [21]. These GWAS support the numerous other studies linking HIV disease outcomes to various *HLA* genes and gene expression.

A recent avenue of research has begun to associate *HLA* genes with diseases independent of peptide binding and presentation. These studies have demonstrated the expression levels of *HLA* genes contribute to various diseases [25, 31, 46]. Studies have shown that *HLA-A* and HLA-C expression levels vary across alleles [31, 46, 47]. HLA-C cell surface expression was measured using the monoclonal antibody DT9 and the expression levels correlated significantly with HLA-C allotypes after variance was analysed ($p=5 \times 10^{-21}$) [31]. In African Americans, HLA-C*03:02 showed the lowest expression with a mean fluorescence intensity (MFI) of approximately 90 and HLA-C*14:02 showed the highest expression with a MFI of just over 300 [31]. *HLA-A* mRNA expression was measured using specific *HLA-A* specific primers in European American (EA) individuals [46]. The expression levels correlated significantly with *HLA-A* lineages and were continuously distributed ($R=0.6$, $p=5 \times 10^{-25}$) [46]. The lowest expressed allele was *HLA-A*03* with an average expression of approximately 0.25 as measured by the $2^{-\Delta\Delta Ct}$ method [46]. The highest expressed allele was *HLA-A*24* with an average expression of approximately 1.0 [46]. Variation of *HLA-A* and *HLA-C* expression levels have shown to associate with progression of many diseases including Crohn disease where in a case control study, high HLA-C expression was associated with increased risk of developing Chron disease ($p=3 \times 10^{-7}$, OR=1.35) [24, 31], and HIV where increased HLA-C

expression is associated with protection and increased *HLA-A* expression is associated with increased progression to AIDS [25, 31].

Elevated *HLA-C* expression levels have been associated with a decrease in HIV disease progression [31, 49]. This association was determined based on a positive correlation between elevated *HLA-C* expression levels and CTL responses [31]. When *HLA-C* expression levels were elevated, the peptides they presented to CTL elicited a greater immune response compared to peptides presented by low expressing alleles [31]. There is a vast repertoire of peptides that bind HLA molecules and only a few of these bind strongly enough to elicit an immune response. An increase in expression of HLA molecules by as little as two-fold is advantageous as more molecules are present therefore increasing the likelihood of peptide binding that is strong enough to generate an immune response [31].

Many of the mechanisms responsible for *HLA-C* regulation have been determined. One mechanism is the downregulation of *HLA-C* by the microRNA mIR-148a [24] which accounts for approximately 9% of *HLA-C* expression variation. A polymorphism in the 3' untranslated region (3'UTR) of *HLA-C* marks upregulation or downregulation of *MIR148A* which in turn causes downregulation or upregulation of *HLA-C* respectively [24]. When there is high expression of mIR-148a, mIR-148a binds to *HLA-C*, directly inhibiting transcription. Another regulatory mechanism is the binding of a transcription factor, Oct-1, to a polymorphic region within the *HLA-C* promoter [50]. This SNP is termed rs2395471 and is an A/G variation which accounts for up to 36% of *HLA-C* variation [50]. Oct-1 binds strongly to alleles containing the A variant which in turn displays significantly higher expression levels compared to the G variant to which Oct-1 binds weakly [50]. These regulatory mechanisms are useful in raising the expression of *HLA-C* in order to combat HIV and similar studies are underway regarding the other Class I genes such as *HLA-A*.

A study published in 2018 showed that increased expression levels of *HLA-A* results in poor control of HIV and subsequently leads to worse disease outcomes. This study examined *HLA-A* mRNA expression in approximately 9,700 individuals from 22 cohorts across three ethnic backgrounds (Africans, Caucasians and Hispanics) [25]. The authors reported higher mRNA expression of *HLA-A* was associated with higher viral load due to receptor-mediated inhibition of natural killer cells [25]. A signal peptide found on the leader sequence of *HLA-A, -B* and *-C* is responsible for the stabilization of HLA-E on the cell surface [25, 51].

HLA-E serves as a ligand for CD94/NKG2A which is an inhibitory receptor found on NK cells [25]. High HLA-E expression is responsible for the inhibition of NK cells thereby dampening NK cell response to HIV. When there is methionine at residue -21 on the signal peptide (-21M) HLA-E expression is promoted and stabilized. *HLA-A* and *-C* are fixed for -21M however *HLA-B* contains a polymorphism which results in either methionine or threonine (-21T) at that position [25]. The study found that homozygotes for *HLA-B* -21M had a positive correlation with high *HLA-A* expression and

high HLA-E expression and therefore had higher HIV viral loads compared to homozygotes for -21T who had lower expression levels of both *HLA-A* and HLA-E and therefore lower HIV viral loads [25]. These findings highlight the crucial role *HLA-A* mRNA expression plays in HIV disease outcomes. Although *HLA-A* expression variation is only one of the factors that contribute to disease progression, it is imperative to determine regulatory mechanisms for *HLA-A* mRNA expression in order to better understand how the regulation of *HLA-A* impacts the risk of acquiring HIV as well as the progression to AIDS.

Studies have identified factors regulating *HLA-A* expression at the post-transcriptional level [52, 53]. One study looked at the role of interferon gamma (IFN-γ) in the regulation of HLA-A [52]. Increased IFN-γ stimulation coupled with chromosome maintenance region 1 (CRM-1) promotes the expression of HLA-A on the cell surface [52]. Another study looked at the function of the differential usage of polyadenylation signals (PAS) in regulating HLA-A cell surface expression [53]. The binding of a protein known as syncrip to the long 3'UTR resulted in lower expression compared to when syncrip was knocked down [53]. These two studies examined at HLA-A expression on the cell surface however it is mRNA expression which has been implicated in HIV disease. There has only been one study which explored regulation of *HLA-A* mRNA. Ramsuran et al 2015, showed that DNA methylation is one of the mechanisms responsible for regulating the expression levels of *HLA-A* at the allelic level [46].

As mentioned previously, *HLA-A* mRNA expression varies across alleles. For example, *HLA-A*24* displays very high expression levels compared to *HLA-A*03* which displays one of the lowest expression levels [46]. DNA methylation is able to lower the expression levels by preventing the binding of transcription factors to the *HLA-A* promoter. The promoter of *HLA-A*03* is heavily methylated thereby preventing transcription factor binding and reducing expression whereas *HLA-A*24* contains much fewer methylation sites, allowing transcription factors to bind and enhance expression. Methylation does not account for the full level of variation seen within *HLA-A* mRNA, therefore, in this study we propose *HLA-A* expression levels are regulated by multiple factors.

Using online predictive software, a binding site was found approximately 1kb upstream of the *HLA-A* TSS that potentially binds CTCF [46]. CTCF is an 82-kDa protein made up of 11 zinc-fingers and is highly conserved across species [54]. CTCF is found ubiquitously within the human genome. CTCF plays many roles within the human genome including long range chromatin interactions, gene activation, gene repression and insulator functions. CTCF was found to play a role in regulating all of the classical *HLA* Class II genes [55]. The binding of CTCF in conjunction with the Class II transactivator (CIITA) and a subunit of the X1 box factor, regulatory factor X5 (RFX5), is responsible for the upregulation of *HLA* Class II [56]. Loss of any of these factors resulted in almost complete loss of Class II expression [56]. CTCF also ties in with DNA methylation, which is one of the

regulators of *HLA-A* expression. The presence of methylation has been shown to prevent CTCF binding onto a CTCF binding site [48].

CTCF playing such a crucial role in Class II expression coupled with the putative CTCF binding site within the *HLA-A* promoter identified by Ramsuran et al. (2015) and the fact that CTCF cannot bind in the presence of DNA methylation leads to the hypothesis that CTCF has a role in Class I expression. This study therefore explores predicted CTCF binding to variants upstream of the *HLA-A* TSS that associate with expression and determine the role of CTCF in the regulation of *HLA-A*.

Aim

To discover variants within the promoter region of *HLA-A* loci which play a role in the regulation of *HLA-A* expression

Objectives

- To examine the 2kb region upstream of the TSS of *HLA-A* to identify variants
- To associate variable SNPs with expression using an unpaired t-test
- To determine CTCF binding sites across variable SNPs using online prediction software, electrophoretic mobility shift assays and chromatin immunoprecipitation
- To determine additional potential transcription factor binding sites using online predictive software

Hypothesis

CTCF regulates *HLA-A* expression at the putative binding site found by Ramsuran et al., 2015.

Rationale

There is a gap in knowledge regarding the factors that regulate HLA-A *expression levels. In this study we examined* HLA-A *regulators by screening the promoter region for variants which may disrupt potential transcription factor binding sites thereby regulating expression.*

Chapter 2

Effect of a polymorphic CTCF binding site on *HLA-A* mRNA expression

Saiyuri Singh[1,2], Mishka Danielle Muthen[1,2], Kimone Leigh Fisher[1,2,3], Hoang Vinh Nguyen[4], Smita Kulkarni[4,5,6], Ravesh Singh[1,7], Thumbi Ndung'u[2,8], Mary Carrington[5,6] and Veron Ramsuran[1,2,3,4,5,6]

Affiliations: [1]School of Laboratory Medicine and Medical Science, University of KwaZulu-Natal, Durban, South Africa. [2]Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), Africa Health Research Institute (AHRI), Durban, South Africa. [3]Centre for AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa. [4] Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX 78227, USA. [5] Ragon Institute of MGH, MIT and Harvard, Cambridge, MA 02139, USA. [6] Cancer and Inflammation Program, Leidos Biomedical Research Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 21702. [7]National Health Laboratory Services, Inkosi Albert Luthuli Hospital, Durban, South Africa. [8]HIV Pathogenesis Programme, Doris Duke Medical Research Institute, Durban, South Africa

Abstract

Many associations with HIV disease have been found within the human leukocyte antigen (HLA) region. Increased *HLA-A* mRNA expression results in poor HIV outcomes due to the inhibition of natural killer (NK) cells. Identifying *HLA-A* regulatory factors is crucial to identify additional HIV drug target sites. One of the factors responsible for *HLA-A* mRNA expression regulation is DNA methylation. However gene regulation is multifactorial and additional regulatory mechanisms have yet to be identified. CCCTC-binding factor (CTCF) plays a crucial role in the regulation of Class II, however the role CTCF plays on *HLA-A* has yet to be explored. A predicted CTCF binding site was previously found upstream of the *HLA-A* transcription start site (TSS). *HLA-A* sequence alignments upstream of the TSS were analysed for polymorphisms which mark expression. An *HLA-A* promoter variant, -993G>A, which significantly marks expression (p=0.0002) and to which CTCF was predicted to differentially bind was identified. An electrophoretic mobility shift assay (EMSA) and chromatin immunoprecipitation were used to confirm CTCF binding. We observed that CTCF did not bind differentially and does not regulate *HLA-A* expression at position -993. Additional transcription factors (NF1, Rap1 and GR) were predicted to bind differentially to -993G>A -226G>A and -885C>G. These data provide a basis for further study surrounding the role of transcription factors regulating *HLA-A* expression.

Background

HLA genes are located within the major histocompatibility complex (MHC) region on the short arm of chromosome 6. The classical *HLA* Class I genes are made up of *HLA-A, HLA-B* and *HLA-C*. These genes have traditionally been shown to have a role in influencing the human immune response via the

presentation of antigenic peptides to various immune cells [1, 2]. Previously, specific HLA alleles were associated with risk of acquiring certain diseases as well as with the outcome of certain infections [3]. These diseases include viral [4-13], bacterial [14, 15], parasitic and autoimmune [4, 16]. Different groups of *HLA* alleles confer increased susceptibility towards or protection against HIV in different populations. An example is a study in which various *HLA* alleles were examined for HIV associations in an Argentinian population. The *HLA-B*18* and *B*39* groups of alleles were found to occur more frequently in HIV positive subjects (p=0.058 and 0.008 respectively, odds ratio=3.84 and 11.96 respectively) suggesting that those groups are associated with increased risk of acquiring HIV. The *B*44* and *B*55* groups were not found at all in HIV positive subjects (p=0.013 and 0.056 respectively, odds ratio=0 for both) which suggests that those groups may have a protective effect against HIV [17].

A recent avenue of research has begun to associate *HLA* genes with diseases independent of peptide binding and presentation. These studies have demonstrated that the expression levels of *HLA* genes contribute to various diseases [6, 8, 18]. Studies have shown that *HLA-A* and *HLA-C* expression levels vary across alleles [6, 18, 19]. HLA-C cell surface expression was measured using the monoclonal antibody DT9 and the expression levels correlated significantly with HLA-C allotypes after variance was analysed ($p=5 \times 10^{-21}$) [6]. In African Americans, HLA-C*03:02 showed the lowest expression with a mean fluorescence intensity (MFI) of approximately 90 and HLA-C*14:02 showed the highest expression with a MFI of just over 300 [6]. *HLA-A* mRNA expression was measured using specific *HLA-A* specific primers in European American (EA) individuals [18]. The expression levels correlated significantly with *HLA-A* lineages and were continuously distributed ($R=0.6$, $p=5 \times 10^{-25}$) [18]. The lowest expressed allele was *HLA-A*03* with an average expression of approximately 0.25 as measured by the $2^{-\Delta\Delta Ct}$ method [18]. The highest expressed allele was *HLA-A*24* with an average expression of approximately 1.0 [18]. Variation of *HLA-A* and *HLA-C* expression levels have shown to associate with progression of many diseases including Crohn disease where in a case control study, high HLA-C expression was associated with increased risk of developing Chron disease ($p=3 \times 10^{-7}$, OR=1.35) [2, 6], and HIV where increased HLA-C expression is associated with protection and increased *HLA-A* expression is associated with increased progression to AIDS [6, 8].

A study published in 2018 has shown that an increase in the expression of the *HLA-A* gene is associated with poor control of HIV and subsequently leads to worse disease outcomes. This effect was measured in 9,763 individuals from 22 cohorts across three ethnic backgrounds (Africans, Caucasians and Hispanics) [8]. Higher expression of *HLA-A* was associated with higher viral load due to receptor-mediated inhibition of natural killer cells [8]. A signal peptide found on the leader sequence of *HLA-A, -B* and *-C* is responsible for the stabilization of HLA-E on the cell surface [8, 20]. HLA-E serves as a ligand for an inhibitory receptor found on NK cells known as CD94/NKG2A

[8]. High HLA-E expression is responsible for the inhibition of NK cells thereby dampening NK cell response to HIV. If there is a methionine at residue -21 on the signal peptide (-21M), HLA-E expression is promoted and stabilized. *HLA-A* and -*C* are fixed for -21M however *HLA-B* contains a polymorphism which results in either methionine or threonine (-21T) at that position [21]. The study found that homozygotes for *HLA-B* -21M had a positive correlation with high *HLA-A* expression and high HLA-E expression and therefore had higher HIV viral loads compared to homozygotes for -21T who had lower expression levels of both *HLA-A* and HLA-E and therefore lower HIV viral loads [8]. These findings highlight the crucial role *HLA-A* mRNA expression plays in HIV disease outcomes. It is therefore imperative to determine regulatory mechanisms for *HLA-A* mRNA expression in order to better understand how regulation of *HLA-A* impacts HIV disease outcomes.

Previous studies reported varying polyadenylation signals (PAS) and differential interferon gamma (IFN-γ) stimulation as factors regulating *HLA-A* post-transcriptionally [22, 23]. These factors regulate the expression of HLA-A on the cell surface and not at the mRNA level. The association between HIV and *HLA-A* is only observed at the mRNA level [8]. There is therefore a need to investigate factors that regulate *HLA-A* at the mRNA level for possible therapeutic target sites. A previous study, identified DNA methylation as one of the factors contributing to *HLA-A* mRNA regulation [18]. However, DNA methylation may account for only a portion of the variation observed in *HLA-A* allelic expression levels.

Using online predictive software, a binding site was found approximately 1kb upstream of the *HLA-A* transcription start site that potentially binds CTCF [18]. CTCF is an 82-kDa protein made up of 11 zinc-fingers and is highly conserved across species [24]. CTCF is found ubiquitously within the human genome and plays many roles within the human genome including long range chromatin interactions, gene activation, gene repression and insulator functions. CTCF binding is dependent on DNA methylation, which is one of the regulators of *HLA-A* expression. The presence of methylation within a CTCF binding site has been shown to prevent CTCF binding [25]. CTCF was found to play a role in regulating all of the classical *HLA* Class II genes [26] and is therefore a strong candidate for the possible regulation of *HLA-A*.

In this study we explore the effects of the putative CTCF binding site on *HLA-A* expression as well as determine which variants within the promoter region are responsible for the regulation of *HLA-A* expression.

Methods

Ethics

Ethical approval for this study was obtained from the Biomedical Research Ethics Committee (BREC). The BREC approval number is BE217/18.

Samples

Cell lines: Cell lines were used as positive and negative controls for the chromatin immunoprecipitation assay. Raji cells, a B cell line, was grown in order to be used for the detection of Histone H3K and normal IgG. THP-1 cells, a monocytic cell lines, were also used for the detection of Histone H3K and normal IgG. Both THP-1 and Raji cell lines used for controls were cultured using 90% RPMI-1640 and 10% fetal bovine serum (FBS) with antibiotic supplementation.

Peripheral blood mononuclear cells (PBMC): 20 HIV negative samples from the FRESH (Females Rising through Education, Support and Health) cohort were used for chromatin immunoprecipitation. This is a cohort of 300, initially non-infected women, aged 18 to 23. The samples were chosen based on whether they were homozygous for specific variants (either A/A or G/G).

Sequence analysis and expression data

Sequences from homozygous samples and expression data from 216 healthy European American (EA) individuals from the Research Donor Program at the Frederick National Laboratory for Cancer Research [18] were analysed visually using BioEdit. These sequences and expression data were both provided by the Carrington lab. Each variant that was found within the sequence alignment was plotted the corresponding *HLA-A* allele using Microsoft Excel. The variants were compared against *HLA-A* mRNA expression on GraphPad Prism 8 using an unpaired t-test to identify which variants significantly marked expression. (*HLA-A* genotypic data and expression are available within Ramsuran et al., 2015, Ramsuran et al., 2017 and Ramsuran et al., 2018.)

Determining putative CTCF binding sites

Online predictive software, CTCFBSDB 2.0 (http://insulatordb.uthsc.edu), was used to predict CTCF binding to variants which mark *HLA-A* expression. CTCFBSDB 2.0 contains data from sources that have determined tens of thousands of CTCF binding sites in seven species (human, macaque, mouse, rat, dog, opossum and chicken) using various methods such as chromatin immunoprecipitation (ChIP), ChIP-sequencing, ChIP-exo (a modification of ChIP-seq) and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [27]. The sequence surrounding each variant with approximately 100bp flanking sequence on either side of the variant was inserted into the "scan" option of the database. Putative binding motifs are then generated along with a binding score. If the score is above 3 or below -3, CTCF has a greater likelihood of binding to the motif.

Electrophoretic mobility shift assay (EMSA)

The electrophoretic mobility shift assay was performed to determine whether CTCF binds to the binding motif which was predicted using CTCFBSDB 2.0.

Biotinylated target DNA, on which the CTCF may bind, was obtained from Integrated DNA Technologies. This DNA contained the sequence of the predicted CTCF binding site along with either the A variant or the G variant, some flanking sequence and biotin probes attached at the end.

Nuclear and cytoplasmic extractions were performed on Jurkat cells using the NE-PER Nuclear and Cytoplasmic Extraction Kit (ThermoFisher Scientific) as per the manufacturer's protocol. The nuclear fraction of jurkat cells was extracted in order to obtain potential transcription factors that may bind. Cells were washed by suspending the cell pellet in phosphate buffered saline (PBS) and pelleted by centrifugation. The supernatant was removed and discarded. Ice-cold CER I and protease inhibitors were added to the cell pellet which was then vortexed and placed on ice. Ice-cold CER II was then added to the tube which was again vortexed and incubated on ice. The supernatant containing the cytoplasmic extract was then transferred to a pre-chilled tube and stored at -80°C. The insoluble fraction, which contains the nuclei, was resuspended in ice-cold NER containing protease inhibitors and nuclear purification was performed. The purified nuclear extract was transferred to a new pre-chilled tube and stored at -80°C until further use. The purity of the extracts was determined by performing a Western Blot analysis using an antibody against α-tubulin for the cytoplasmic extract and an antibody against KU80 for the nuclear extract.

An electrophoretic mobility shift (EMSA) was performed to determine whether any transcription factors bind to the putative CTCF binding site and whether these transcription factors bind differentially based on the variant. The assay was performed using the Lightshift Chemiluminescence EMSA kit (ThermoFisher Scientific) according to the manufacturer's guidelines. Briefly, a 6% native polyacrylamide gel was set to pre-run for one hour at 100V. Binding reactions were performed by adding the biotinylated DNA fragments to ultrapure water, 10X Binding Buffer, 1µg/µL Poly (dI•dC) and the Jurkat nuclear extract. The mixture was incubated for 20 minutes at room temperature. Loading dye was added to the binding reactions which were then loaded and run on the 6% native polyacrylamide gel. Gel contents were transferred to a nylon membrane electrophoretically. Transferred DNA was fixed onto the membrane via UV crosslinking. The membrane then underwent blocking with the addition of stabilized streptavidin-horseradish peroxidase following washing. A working solution made up of equal parts of luminol and stable peroxide solution was then added to the membrane which was incubated for five minutes without shaking. The membrane was then viewed using the ChemiDoc™ Touch Imaging System (Bio-Rad).

Chromatin immunoprecipitation (ChIP) assay

The chromatin immunoprecipitation assay was run to determine whether CTCF was found on the putative CTCF binding site in live, human cells. ChIP was performed on Raji cells (as the positive and negative controls), THP-1 cells (as the second positive and negative controls) and 20 healthy

donor PBMC samples, ten of which were homozygous for A alleles and ten of which were homozygous for G alleles using the SimpleChIP Enzymatic Chromatin IP kit (Magnetic Beads) (Cell Signalling Technology) according to the manufacturer's guidelines. Raji cells, THP-1 cells and PBMCs were treated with formaldehyde at a final concentration of 1%. Glycine was added to stop the cross-linking reaction. Cells were washed with ice-cold phosphate-buffered saline, pelleted by centrifugation at 4°C and lysed using buffer A. The resulting nuclei were resuspended in ice-cold buffer B and digested with micrococcal nuclease. The enzyme reaction was stopped by the addition of EDTA. Nuclei were lysed by sonication on ice using a QSONICA sonicator (Lasec). The sheared chromatin was collected by centrifugation and a total of 2µg anti-CTCF monoclonal antibody (MERCK) was added to the test samples. For the positive control and negative control, 10µl of antibody against histone H3 (Cell Signaling Technology) and 1µl normal IgG control antibody was added respectively.

After overnight incubation at 4°C with rotation, ChIP-grade protein G magnetic beads were added and samples were incubated further. Cross-linking was reversed by incubation with the addition of proteinase K and NaCl. Purified DNA was eluted and real-time PCR was performed with the SimpleChIP qPCR Master Mix (Cell Signalling Technology) and primers (https://primer3plus.com) specific for the CTCF binding region upstream of *HLA-A* (Integrated DNA Technologies) (Table S2). Primers for insulin like growth factor (IGF) [28] were also used to confirm the specificity of the CTCF antibody. The amount of immunoprecipitated DNA in each sample was calculated using the percentage of input method.

Exploring online predictive software for binding of additional transcription factors

The sequence of variants marking *HLA-A* expression were inserted into AliBaba2.1 (http://gene-regulation.com/pub/programs/alibaba2) in order to predict whether additional transcription factors bind to the variants. This tool uses matrices that are constructed using a database called TRANSFAC 3.5 public. Matrices are constructed for specific sequences starting with a database of known transcription factor binding sites and leading to the identification of new putative binding sites [29].

Determining linkage disequilibrium

Variants marking *HLA-A* expression were analysed for linkage disequilibrium using LDmatrix (https://ldlink.nci.nih.gov/?tab=ldmatrix) in order to determine similarities between variants.

Results

Analysis of 2KB sequences upstream of the TSS from *HLA-A* homozygous alleles revealed 34 variable SNPs across alleles (Fig. S1). Each of these SNPs were compared with corresponding *HLA-A* mRNA expression levels, which were generated from 216 healthy Caucasian donors. *HLA-A* mRNA expression levels from donors containing either the wild type or mutated variant were compared using an unpaired t-test ('wild type' refers to the variant found on *HLA-A*01*, the reference sequence, and 'mutant' refers to the variant not found on *HLA-A*01*) (Table. S1). Six SNPs, *rs41560714_T>G* (-55T>G), *rs9260116_A>G* (-226A>G), *rs9260101_G>A* (-772G>A), *rs9260092_C>G* (-885C>G), *rs9260086_G>C* (-960G>C) and *rs9260084_G>A* (-993G>A), were found to significantly associate with *HLA-A* mRNA expression levels (Fig.1). Four SNPs (-772, -885, -960 and -993) marked expression with equal significance when comparing homozygous mutant variants to homozygous wild type which indicates a dominance effect (p=0.0002) (Fig. 1C, D, E, F). The genotypes found at higher levels were -772A/A, -885G/G, -960C/C and -993A/A. These genotypes were 1.2385 fold higher. The -55 SNP also significantly marked expression of homozygotes (p=0.0191), donors with TT were significantly higher by 1.1511 fold compared to GG genotype. Within the -226 SNP only one donor possesses the AA genotype, however we observed an expression difference between the high expressing heterozygotes, AG, and low expressing homozygotes, GG (p=0.0405). The fold change in the average expression of AG vs GG was 1.4657.

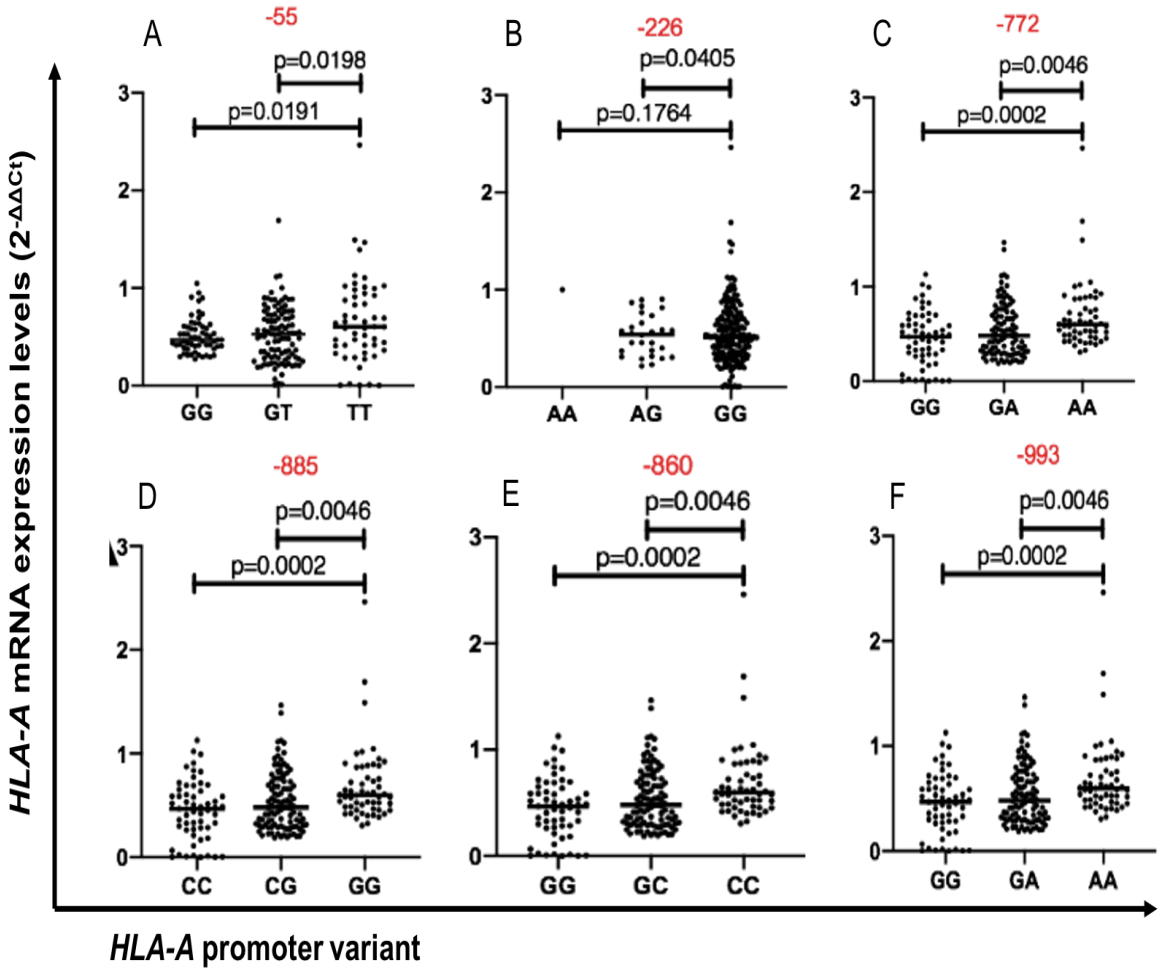


Figure 1. Unpaired t-test six *HLA-A* promoter SNPs that significantly marked expression. Unpaired t-tests were performed on 34 SNPs found upstream of the *HLA-A* TSS using expression data from 216 healthy European American individuals from the Research Donor Program at the Frederick National Laboratory for Cancer Research. Only six SNPs, located at positions -55, -226, -772, -885, -860 and -993 (labelled in red) located relative to the TSS, significantly marked *HLA-A* expression. HLA-A mRNA expression was measured using a real-time PCR assay. The SNPs were identified from 2KB alignments of common homozygous *HLA-A* alleles

Following the association of six SNPs with *HLA-A* mRNA expression, we examined putative binding of CTCF, based on a previous study by Ramsuran et al. (2015), using the *in silico* prediction tool (CTCFBSDB2.0, http://insulatordb.uthsc.edu). The sequences surrounding and including the -55, -226, -772, -885, -960 and -993 variants were inserted into the software with default settings and putative binding sites were observed (Fig. 2 A-F). No predicted CTCF binding was observed for SNPs -55, -226, -772, -885 or -960 (Fig. 2 A, B, C, D and E). However, the -993 variant showed a disrupted CTCF binding site when -993A is present and predicted CTCF binding to -993G with a binding score of 12.6319 (Fig 2F). A binding score above 3 is suggestive of a CTCF binding match.

Figure 2. CCCTC-binding factor, CTCF, is predicted to bind to -993G but not to -993A or any other of the five SNPs. Using online predictive software for CTCF binding called CTCFBSDB 2.0, all six SNPs which significantly mark expression were analysed for CTCF binding. The red letters indicate the sequence upstream of the *HLA-A* TSS. The arrows indicate the position of each mutation. The blue box represents a predicted CTCF binding site (GGCAGTGCA) and the numbers (-55, -226, -772, -885, -960 and -993) indicate the position of the mutation relative to the *HLA-A* TSS. There was differential binding of CTCF to -993G and not -993A, F. There was no predicted CTCF binding to any of the other SNPS (-55, -226, -772, -885 and -960) A, B, C, D, and E.

Since CTCF was predicted to bind -993A and -993G differentially, we identified which *HLA-A* alleles possess the -993A or G alleles. A sequence alignment of common homozygous *HLA-A* alleles revealed which alleles contain -993G and -993A (Fig. 3). *A*01, A*03, A*11, A*29, A*30, A*31, A*32, A*33* and *A*78* possess -993G whereas *A*02, A*23, A*24, A*25, A*26, A*34, A*66,* and *A*68* possess -993A.

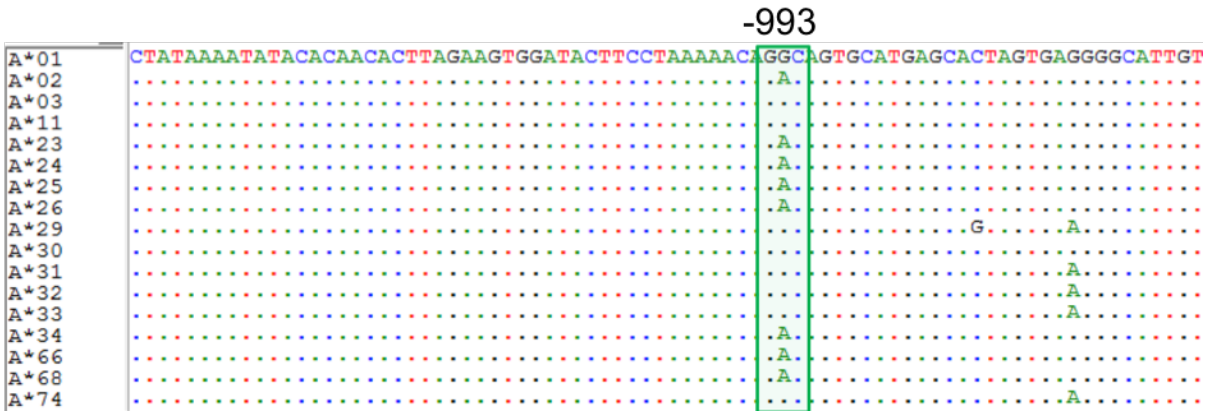


Figure 3. Sequence alignment of *HLA-A* alleles indicating the mutation within the putative CTCF binding site, -993G>A (boxed in green). *HLA-A* sequences ~2kb upstream of the TSS from common homozygous *HLA-A* alleles. *HLA-A*01* was used as a reference sequence against which all other sequences were aligned therefore all bases are represented by coloured letters. Below the reference sequence each dot indicates that the same base pair is present and the letters indicate that a different base i.e. a mutation is present. -993G>A, boxed in green, showed differential CTCF binding based on -993G>A.

Since only position -993G>A showed predicted CTCF binding, we examined presence of differential protein binding at this position. An EMSA was employed to detect differences in protein binding. Biotinylated oligonucleotides containing either -993G or -993A were run on a polyacrylamide gel along with the nuclear extract of Jurkat cells to determine whether any protein found within the nuclear extract bound to the fragment. Binding is seen by an upwards shift of the fluorescent bands on the gel. A control containing no protein extract was run for each oligonucleotide to determine where unbound DNA would lie i.e. determining a base from which an upwards shift can be seen. A second control containing excess unbiotinylated DNA was run to show that excess unlabelled DNA will bind the protein and not fluoresce thereby confounding results. This upwards shift was observed for the fragment possessing -993G but not for the fragment containing -993A (Fig. 4). These results indicate that a protein binds to -993G but not to -993A.

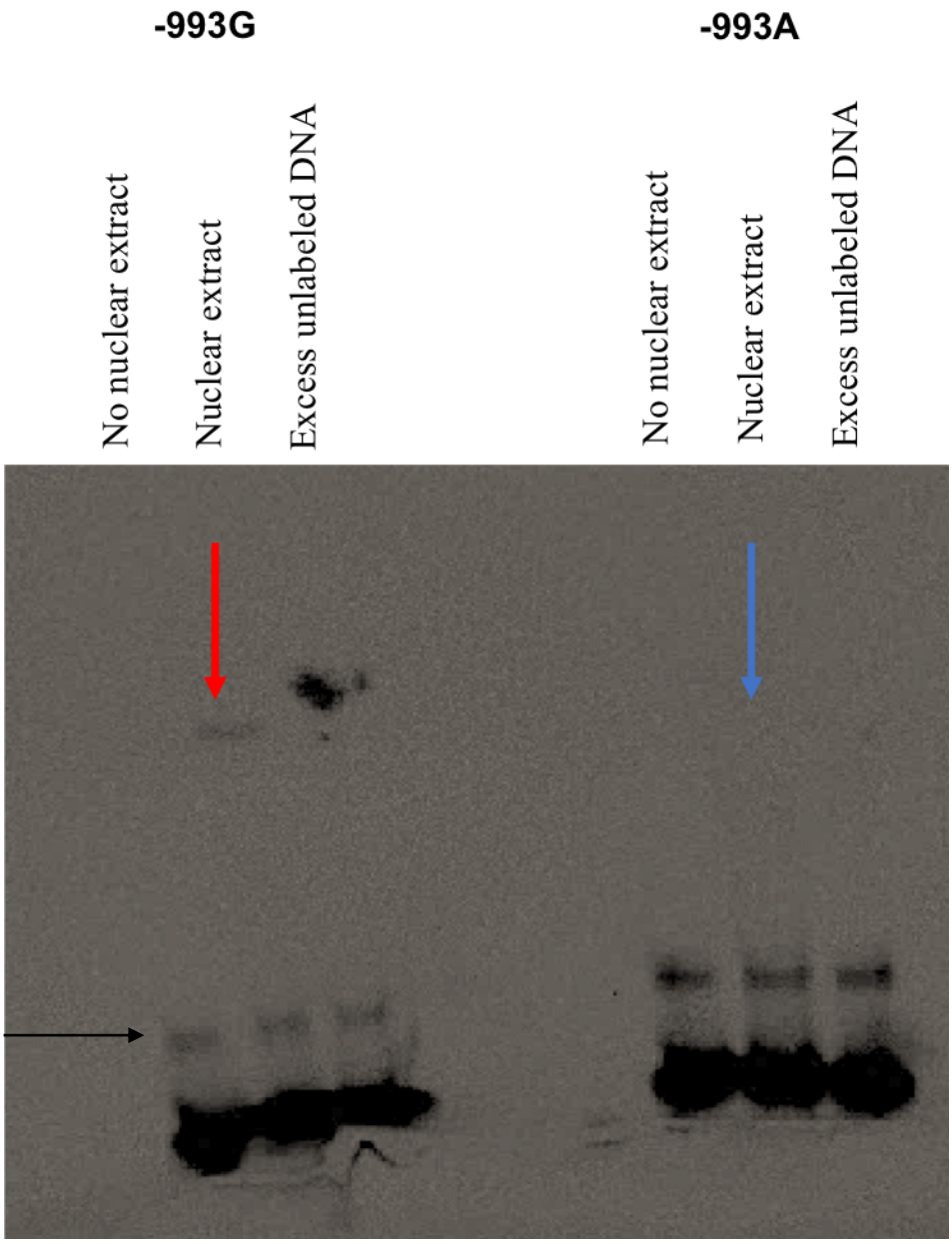


Figure 4. EMSA showing protein binding to -993G but not to -993A. An EMSA was performed using biotinylated oligonucleotides possessing -993G and -993A and the nuclear extract from Jurkat cells. The first control for this assay was a lane containing no nuclear extract to indicate the position of unbound bands (black arrow). The second control was a lane in which excess non-biotinylated fragments were added to show that excess unlabelled fragments bind the extract and confound results by not fluorescing. There was an upwards shift seen in the presence -993G (red arrow) indicating that a protein found in the Jurkat nuclear extract binds to the fragment. There is no visible shift in the presence of -993A (blue arrow).

In order to confirm the EMSA results as well as to identify the unknown protein as CTCF, chromatin immunoprecipitation was performed using a CTCF specific antibody. The assay was run on 20

healthy donor PBMC samples, 10 homozygous for -993G and 10 homozygous for -993A. The qPCR results were calculated using the percentage of input method [30] and plotted using GraphPad Prism 8 software. An antibody against Histone H3 was used as a positive control for the experiment with primers for exon 3 of the human RPL3 gene. Amplification of RPL3 was seen on average around cycle 20 of the qPCR reaction indicating that the immunoprecipitation was successful. Only 19 PBMC samples were reported. One sample was excluded due to differing results when the qPCR was repeated. The results indicated no significant difference between the binding of CTCF to -993G and -993A (p=0.2773) (Fig. 5).

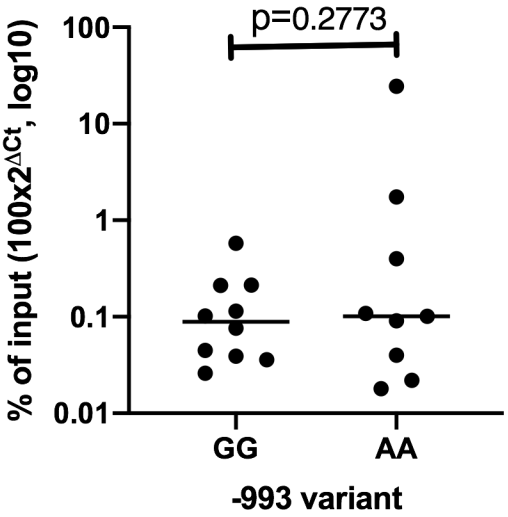


Figure 5. ChIP results confirm CTCF binding to both -993G and -993A with a similar affinity. 20 healthy donor PBMC samples were immunoprecipitated. CTCF binding was calculated using the percentage of input method. One sample was excluded due to non-conformity between technical replicates. Each dot represents one sample. The amount of CTCF that binds to -993G was similar to the amount that binds to -993A (p=0.2773).

Due to the lack of CTCF differential binding, we aimed to identify additional predicted transcriptional factor binding using online predictive software called AliBaba2.1 for the six variants associated with *HLA-A* expression. Sequences surrounding and including the -55, -226, -772, -885, -960 and -993 variants were inserted into the prediction tool. There was no predicted transcription factor binding to either variant present at -55 (Fig. 6A). Ras-related protein 1 (Rap1) showed binding to -226G but not to -226A. Pancreas transcription factor 1 beta (PTF1B) was shown to bind -772G>A independent of the variant present (Fig. 6C). -885 had differential binding of glucocorticoid receptor (GR) which plays a role in transcription when bound to glucocorticoids (Fig. 6D). GR was predicted to bind to -885G but not to -885C. There was no predicted transcription factor binding to -960 (Fig. 6E). Nuclear

Factor 1 (NF1) is predicted to bind differentially to the exact CTCF binding site containing the SNP at position -993 (Fig. 6F). NF1 was predicted to bind to -993G but not to -993A.

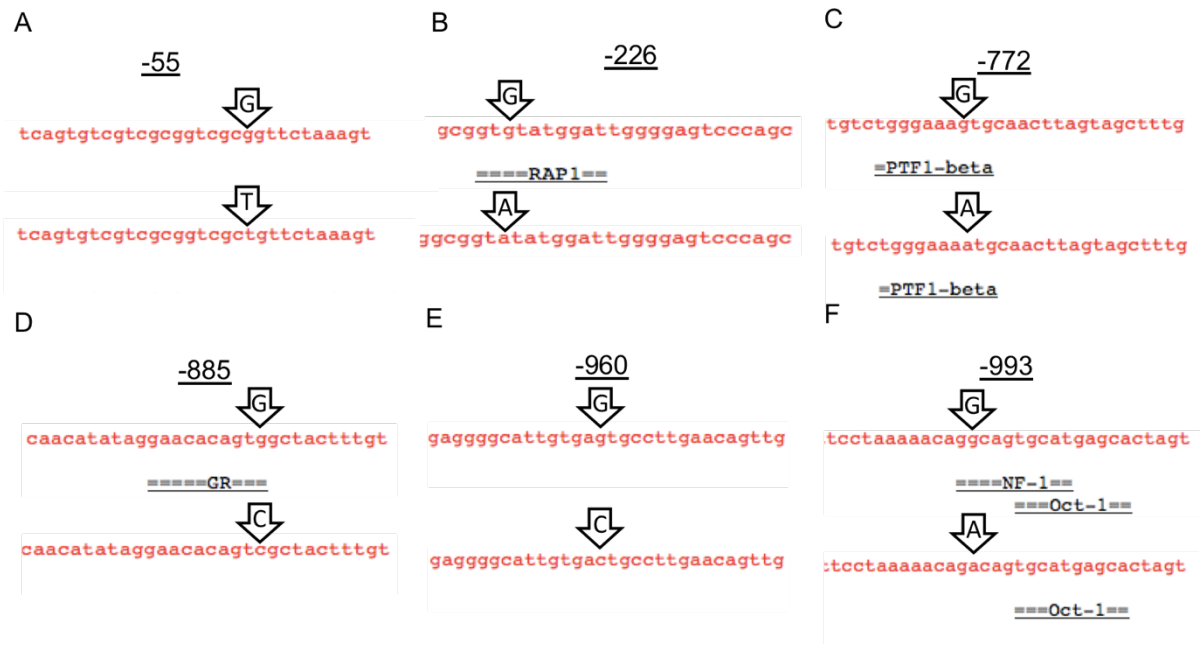


Figure 6. Transcription factors predicted to bind to the six SNPs (AliBaba2.1). Using an online predictive tool for transcription factor binding called AliBaba2.1, transcription factors were predicted to bind to the top four mutations. The red letters indicate the sequence upstream of the *HLA-A* TSS. The arrows indicate the position of each mutation. "=" indicates the predicted transcription factor binding site and the numbers (-55, -226, -772, -885, -960 and -993) indicate the position of the mutation relative to the *HLA-A* TSS. A, there was no predicted transcription factor binding to -55. B, Rap1 showed binding to -226G but not to -226A. C, there was predicted binding of PTF1B to -772 regardless of the variant. D, GR was predicted to differentially bind to -885. E, there was no predicted transcription factor binding to -960. F, -993 showed differential binding of NF1 based on the variant.

Due to the four of the six SNPs displaying similar patterns and marking expression with the same significance, LD analysis was performed to determine if any of the SNPs were in LD with one another. The rs numbers of the six SNPs were inserted into LDMatrix and a matrix and correlation score were generated for individual populations (African, Ad Mixed American, East Asian, European and South Asian) as well as for all populations combined. Of the six SNPs, LD information was only available for five. Information on -993 has yet to be found on LD databases. The three SNPs at positions -960, -885 and -772 all display strong LD with one another in all populations available on LD Link (r^2=1) (Fig. 7). In each individual population (African, Ad Mixed American, East Asian, European and South Asian) strong LD was consistently observed between -960, -885 and -772 (r^2=1.)

(Fig. 7). -226 correlated very weakly with -772, -885 and -960 (r^2=0.042). -55 also had a weak correlation with -772, -885 and -960 (r^2=0.114). This means that -55 and -226 are not in LD or are in very weak LD with -772, -885 and -960. -55 and -226 are not in LD with one another since the correlation score for these two SNPs was also very low (r^2=0.038).

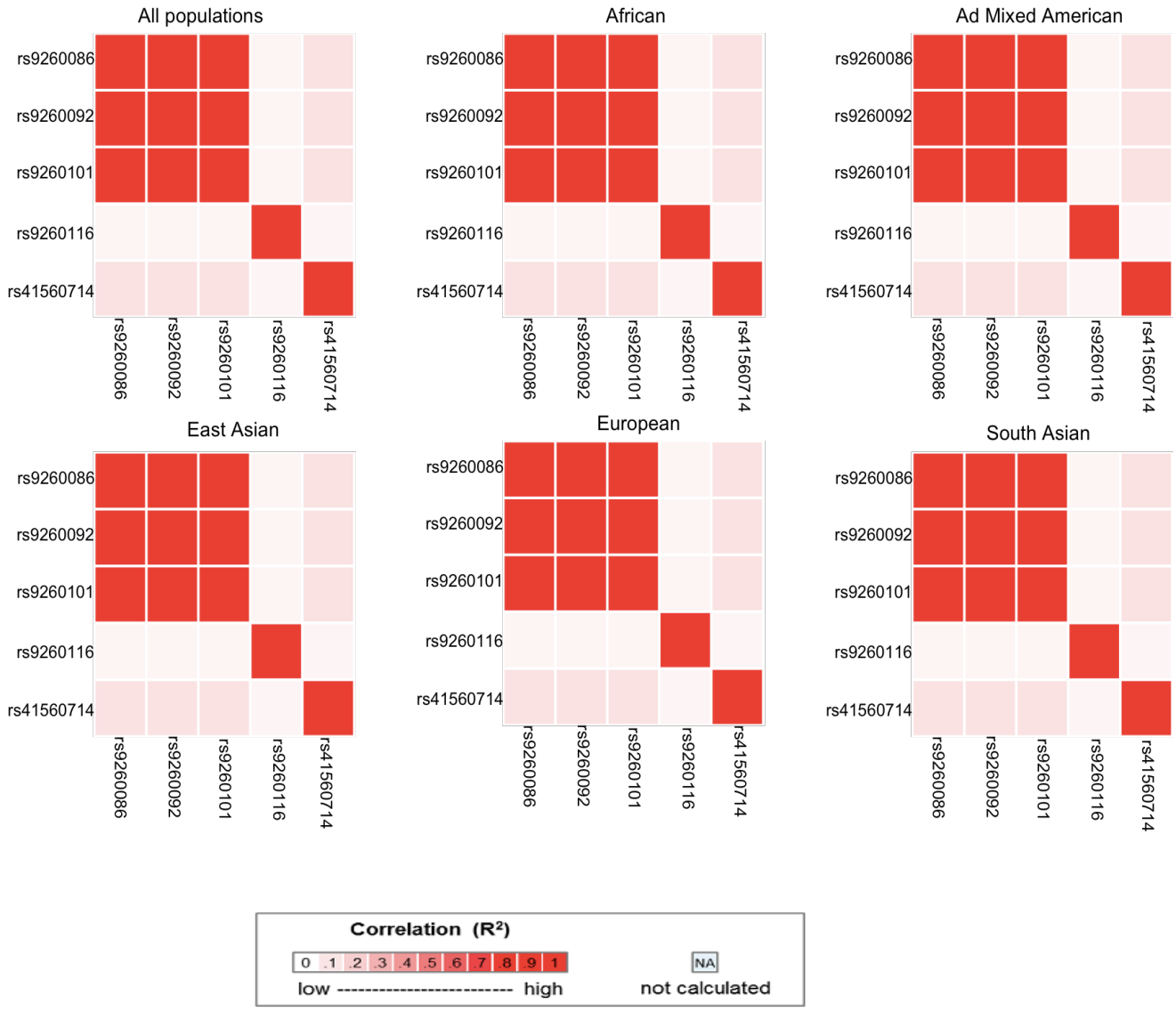


Figure 7. Linkage disequilibrium matrix plot for five out of six SNPs. The top four SNPs were analysed for linkage disequilibrium (LD) using LDmatrix. There was no LD information for -993. -960 *(rs9260086)*, -885 (*rs9260092*) and -772 (*rs9260101*) displayed strong LD with each other (r^2=1) in All populations as well as in African, Ad Mixed American, East Asian, European and South Asian populations individually. -226 (*rs9260116)* displayed very weak LD with -772, -885 and -960 (r^2=0.042). -55 (*rs41560714*) displayed weak LD with -772, -885 and -960 (r^2=0.114). -225 and -55 displayed very weak LD with one another (r^2=0.038).

Discussion

This study aimed to discover variants within the promoter region of *HLA-A* loci which play a role in the regulation of *HLA-A* expression. The objectives of this study were carried out in order to better understand how the regulation of *HLA-A* impacts the risk of acquiring HIV as well as progression to AIDS.

The 2kb region upstream of the *HLA-A* TSS revealed 34 variants with the potential to mark *HLA-A* expression. It is unsurprising that so many variants were found since *HLA* genes are 20-fold more diverse than the rest of the genome [31]. Following analysis using an unpaired t-test, only six out of the 34 variants (-55, -226, -772, -885, -960 and -993) were found to significantly mark expression. Out of the six significant variants, four (-772, -885, -960 and -993) showed identical patterns to one another. Analysis of LD between the variants revealed that -772, -885 and -960 were in perfect LD with one another. Visually -993 appears to be in perfect LD with -772, -885 and -960 however there was no LD information on -993. The lack of LD information on -993 was because this variant did not appear on any of the chips used to determine LD.

CTCFBSDB 2.0 revealed that CTCF may bind differentially to -993G>A. EMSA results seemed to confirm this finding since there was a shift seen for -993G but not -993A. Our ChIP results demonstrated that CTCF binds both -993G and -993A with a similar affinity. CTCF therefore does not regulate *HLA-A* mRNA expression at position -993. EMSA showed differing results to ChIP in terms of CTCF binding to -993G>A. Although EMSA results have been shown to correlate with ChIP results in terms of CTCF in some literature [32, 33] other literature revealed that the nature of CTCF binding observed in the EMSA differed from the CTCF binding that was seen when confirmed using ChIP [34, 35]. This may also be due to alternative (transcription) factors affecting CTCF binding. It could also be the presence of methylation sites within putative CTCF binding sites [34] since the presence of methylation has been shown to prevent CTCF binding onto a CTCF binding site [25].

Of the six SNPs shown to associate with *HLA-A* mRNA expression levels, two SNPs (-55G>T and -226G>A) were located within the core promoter region of the *HLA-A* gene. Despite -55G>T not showing any predicted transcription factor binding, this SNP lies 2 base pairs outside of the TATA box. The TATA box is where multiple transcription factors such as TATA binding protein and TFIID bind to regulate expression [36]. It is possible the mutation at position -55 might affect the transcription factor binding on the TATA box and therefore regulate the mRNA expression of *HLA-A*. However, this is yet to be confirmed and future experiments are required.

The second SNP located within the *HLA-A* core promoter, -226G>A, is predicted to differentially bind Rad1. Rad1 binding is predicted when G is present, but binding is disrupted when A is present. The Rad1 protein has been shown to play a major role in regulation of gene expression in yeast [37-

40]. In humans, Rad1 acts as a checkpoint protein [41, 42]. It forms part of a complex known as 9-1-1 which is activated to stop the cell cycle when DNA is damaged or DNA replication is incomplete [41, 42]. Further work is needed in the area to determine the contribution Rad1 has on *HLA-A* mRNA expression.

NF1 showed differential binding to the CTCF binding site at -993. NF1 acts as a transcriptional activator in both humans and viruses. It has been shown that the level of NF1 binding is associated with the level of transcription [43]. When NF1 binds strongly to a gene the transcription levels of that gene are much higher than genes to which NF1 binds weakly [43]. Further analysis on the interaction between NF1 binding -993G>A needs to be performed as NF1 is a possible regulator of *HLA-A* expression. Another transcription factor, GR, showed differential binding to -885C>G. GR is a ligand-activated nuclear receptor [44] which means that it needs to bind to steroid ligands in order to function as a transcription factor. GR modulates transcription of genes in two ways: via the binding of receptor dimers to specific palindromic sequences called glucocorticoid response elements (GREs) and indirectly by interacting with other transcription factors e.g. Nuclear Factor kappa beta (NF-κβ) [45] and activator protein 1 (AP-1) [44, 46, 47]. The binding of GR to -885C>G could be a regulatory mechanism for *HLA-A* expression. PTF1B was found to bind to -772G>A regardless of the variant however this transcription factor is involved in pancreatic function and is not relevant to *HLA-A* regulation.

Limitations of this study include the relatively small sample size. It is possible that differences in CTCF binding can be seen in a much larger sample group. The type of sample used may also be a limitation. Bulk PBMCs were used for the ChIP assay. It is possible differential binding of CTCF can be seen across individual cell types. There is also the possibility of CTCF regulating *HLA-A* in different ethnic groups since different ethnic groups have varying susceptibility to diseases [48-50] and therefore may have variation in *HLA-A* expression. *HLA-A* expression was measured in Caucasians, Hispanics and Africans previously and there was no significant difference in expression data across those three ethnic groups [8] disproving the previous statement.

Further study needs to be conducted on the mechanisms of *HLA-A* regulation. The differential binding of NF1 to -993G>A and GR to -885C>G need to be confirmed for possible effects on *HLA-A* regulation since the variants are highly likely to be in perfect LD with one another. Studies should also consider looking at *HLA-A* expression across different cell types to determine whether regulatory mechanisms vary across the different cells that make up PBMCs. In conclusion, although CTCF does not regulate *HLA-A* expression by differential binding to -993G>A, the possible binding of NF1 or GR to a SNP in perfect LD with -993G>A serve as a basis for further study regarding *HLA-A* expression regulation in order to better understand the implications *HLA-A* regulation has on HIV.

References

1. Moss, P., et al., *Extensive conservation of alpha and beta chains of the human T-cell antigen receptor recognizing HLA-A2 and influenza A matrix peptide.* Proceedings of the National Academy of Sciences, 1991. **88**(20): p. 8987-8990.
2. Kulkarni, S., et al., *Genetic interplay between HLA-C and MIR148A in HIV control and Crohn disease.* Proc Natl Acad Sci U S A, 2013. **110**(51): p. 20705-10.
3. Risch, N., *Assessing the role of HLA-linked and unlinked determinants of disease.* Am J Hum Genet, 1987. **40**(1): p. 1-14.
4. Gough, S.C.L. and M.J. Simmonds, *The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action.* Current genomics, 2007. **8**(7): p. 453-465.
5. Apps, R., et al., *Relative expression levels of the HLA class-I proteins in normal and HIV-infected cells.* J Immunol, 2015. **194**(8): p. 3594-600.
6. Apps, R., et al., *Influence of HLA-C expression level on HIV control.* Science, 2013. **340**(6128): p. 87-91.
7. Thomas, R., et al., *A novel variant marking HLA-DP expression levels predicts recovery from hepatitis B virus infection.* Journal of virology, 2012. **86**(12): p. 6979-6985.
8. Ramsuran, V., et al., *Elevated HLA-A expression impairs HIV control through inhibition of NKG2A-expressing cells.* Science, 2018. **359**(6371): p. 86-90.
9. Borghans, J.A., et al., *HLA alleles associated with slow progression to AIDS truly prefer to present HIV-1 p24.* PLoS One, 2007. **2**(9): p. e920.
10. Weiskopf, D., et al., *Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8+ T cells.* Proc Natl Acad Sci U S A, 2013. **110**(22): p. E2046-53.
11. Rao, X., et al., *HLA Preferences for Conserved Epitopes: A Potential Mechanism for Hepatitis C Clearance.* Front Immunol, 2015. **6**: p. 552.
12. Malavige, G.N., et al., *HLA class I and class II associations in dengue viral infections in a Sri Lankan population.* PLoS One, 2011. **6**(6): p. e20581.
13. Thio, C.L., et al., *Racial differences in HLA class II associations with hepatitis C virus outcomes.* J Infect Dis, 2001. **184**(1): p. 16-21.
14. Ebringer, A. and C. Wilson, *HLA molecules, bacteria and autoimmunity.* J Med Microbiol, 2000. **49**(4): p. 305-11.
15. Singh, R.K. *HLA and Skin Diseases*. 2014 [cited 2019 29/11/19]; Available from: https://www.slideshare.net/RKSKUSHWAHA/hla-and-skin-disorders.
16. Matzaraki, V., et al., *The MHC locus and genetic susceptibility to autoimmune and infectious diseases.* Genome Biol, 2017. **18**(1): p. 76.
17. de Sorrentino, A.H., et al., *HLA class I alleles associated with susceptibility or resistance to human immunodeficiency virus type 1 infection among a population in Chaco Province, Argentina.* The Journal of infectious diseases, 2000. **182**(5): p. 1523-1526.

18. Ramsuran, V., et al., *Epigenetic regulation of differential HLA-A allelic expression levels.* Hum Mol Genet, 2015. **24**(15): p. 4268-75.

19. Ramsuran, V., et al., *Sequence and Phylogenetic Analysis of the Untranslated Promoter Regions for HLA Class I Genes.* J Immunol, 2017. **198**(6): p. 2320-2329.

20. Braud, V.M., et al., *HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C.* Nature, 1998. **391**(6669): p. 795-9.

21. Horowitz, A., et al., *Class I HLA haplotypes form two schools that educate NK cells in different ways.* Sci Immunol, 2016. **1**(3).

22. Kulkarni, S., et al., *Posttranscriptional Regulation of HLA-A Protein Expression by Alternative Polyadenylation Signals Involving the RNA-Binding Protein Syncrip.* J Immunol, 2017. **199**(11): p. 3892-3899.

23. Browne, S.K., et al., *Differential IFN-γ stimulation of HLA-A gene expression through CRM-1-dependent nuclear RNA export.* The Journal of Immunology, 2006. **177**(12): p. 8612-8619.

24. Filippova, G.N., et al., *An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes.* Mol Cell Biol, 1996. **16**(6): p. 2802-13.

25. Bell, A.C. and G. Felsenfeld, *Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene.* Nature, 2000. **405**(6785): p. 482.

26. Majumder, P. and J.M. Boss, *CTCF controls expression and chromatin architecture of the human major histocompatibility complex class II locus.* Molecular and cellular biology, 2010. **30**(17): p. 4211-4223.

27. Ziebarth, J.D., A. Bhattacharya, and Y. Cui, *CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization.* Nucleic acids research, 2012. **41**(D1): p. D188-D194.

28. Kim, T.H., et al., *Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.* Cell, 2007. **128**(6): p. 1231-45.

29. Grabe, N., *AliBaba2: context specific identification of transcription factor binding sites.* In silico biology, 2002. **2**(1): p. S1-S15.

30. Lin, X., L. Tirichine, and C. Bowler, *Protocol: Chromatin immunoprecipitation (ChIP) methodology to investigate histone modifications in two model diatom species.* Plant methods, 2012. **8**(1): p. 48-48.

31. Trowsdale, J. and J.C. Knight, *Major histocompatibility complex genomics and human disease.* Annu Rev Genomics Hum Genet, 2013. **14**: p. 301-23.

32. Li, W., et al., *Identification of critical base pairs required for CTCF binding in motif M1 and M2.* Protein & cell, 2017. **8**(7): p. 544-549.

33. Tang, M., et al., *Restraint of angiogenesis by zinc finger transcription factor CTCF-dependent chromatin insulation.* Proceedings of the National Academy of Sciences, 2011. **108**(37): p. 15231-15236.

34. Kotova, E.S., et al., *Binding of Protein Factor CTCF within Chicken Genome Alpha-Globin Locus.* Acta naturae, 2016. **8**(1): p. 90-97.

35. Pugacheva, E.M., et al., *Comparative analyses of CTCF and BORIS occupancies uncover two distinct classes of CTCF binding genomic regions.* Genome biology, 2015. **16**(1): p. 161.

36. Watson, J.D., *Molecular Biology of the Gene*. 1987: Benjamin/Cummings Publishing Company.

37. Goto, G.H., et al., *Binding of Multiple Rap1 Proteins Stimulates Chromosome Breakage Induction during DNA Replication.* PLoS Genet, 2015. **11**(8): p. e1005283.

38. Wu, A.C.K., et al., *Repression of Divergent Noncoding Transcription by a Sequence-Specific Transcription Factor.* Mol Cell, 2018. **72**(6): p. 942-954.e7.

39. Wu, A.C.K. and F.J. Van Werven, *Transcribe this way: Rap1 confers promoter directionality by repressing divergent transcription.* Transcription, 2019. **10**(3): p. 164-170.

40. Challal, D., et al., *General Regulatory Factors Control the Fidelity of Transcription by Restricting Non-coding and Ectopic Initiation.* Mol Cell, 2018. **72**(6): p. 955-969.e7.

41. Bao, S., et al., *Disruption of the Rad9/Rad1/Hus1 (9–1–1) complex leads to checkpoint signaling and replication defects.* Oncogene, 2004. **23**(33): p. 5586.

42. Parrilla-Castellar, E.R., S.J. Arlander, and L. Karnitz, *Dial 9–1–1 for DNA damage: the Rad9–Hus1–Rad1 (9–1–1) clamp complex.* DNA repair, 2004. **3**(8-9): p. 1009-1014.

43. Gronostajski, R.M., et al., *Stimulation of transcription in vitro by binding sites for nuclear factor I.* Nucleic acids research, 1988. **16**(5): p. 2087-2098.

44. Muzikar, K.A., N.G. Nickols, and P.B. Dervan, *Repression of DNA-binding dependent glucocorticoid receptor-mediated gene expression.* Proceedings of the National Academy of Sciences, 2009. **106**(39): p. 16598-16603.

45. McKay, L.I. and J.A. Cidlowski, *Cross-talk between nuclear factor-kappa B and the steroid hormone receptors: mechanisms of mutual antagonism.* Mol Endocrinol, 1998. **12**(1): p. 45-56.

46. Heck, S., et al., *A distinct modulating domain in glucocorticoid receptor monomers in the repression of activity of the transcription factor AP-1.* The EMBO journal, 1994. **13**(17): p. 4087-4095.

47. De Bosscher, K., W. Vanden Berghe, and G. Haegeman, *The interplay between the glucocorticoid receptor and nuclear factor-kappaB or activator protein-1: molecular mechanisms for gene repression.* Endocr Rev, 2003. **24**(4): p. 488-522.

48. Lau, C., G. Yin, and M. Mok, *Ethnic and geographical differences in systemic lupus erythematosus: an overview.* Lupus, 2006. **15**(11): p. 715-719.

49. Ackerman, M.J., et al. *Ethnic differences in cardiac potassium channel variants: implications for genetic susceptibility to sudden cardiac death and genetic testing for congenital long QT syndrome*. in *Mayo clinic proceedings*. 2003. Elsevier.

50. URAYAMA, K.Y. and A. MANABE, *Genomic evaluations of childhood acute lymphoblastic leukemia susceptibility across race/ethnicities.* 臨床血液, 2014. **55**(10): p. 2242-2248.

Supplementary material

Table S1: Positions of SNPs found upstream of the *HLA-A* TSS and their rsID's

Position	Variant	rsID
-55	T>G	*rs41560714*
-142	A>T	*rs9260119*
-155	C>T	*rs9260118*
-226	A>G	*rs9260116*
-407	G>A	*rs2734903*
-412	T>C	*rs9260112*
-414	A>C	*rs9260111*
-617	C>T	*rs9260106*
-716	T>A	*rs1230326099*
-772	A>G	*rs9260101*
-789	A>G	*rs9260100*
-808	G>T	*rs9260099*
-809	A>C	*rs9260098*
-812	A>G	*rs9260097*
-822	C>T	*rs9260096*
-833	A>T	*rs9260095*
-860	C>G	*rs9260094*
-878	C>T	*rs9260093*
-885	C>G	*rs9260092*
-892	A>G	*rs9260091*
-906	C>T	*rs9260090*
-923	T>C	*rs9260089*
-940	A>C	*rs9260088*
-956	C>A	*rs9260087*
-960	G>C	*rs9260086*
-993	G>A	*rs9260084*
-1084	C>A	*rs9260083*
-1180	G>T	*rs9260081*
-1263	C>T	*rs9260079*
-1300	A>G	*rs9260078*
-1379	G>A	*rs9260077*
-1396	C>T	*rs78179089*
-1647	T>C	*rs9260069*
-1684	T>G	*rs9260067*

Table S2. 34 *HLA-A* SNPs upstream of the TSS. 34 *HLA-A* SNPs found upstream of the TSS according to *HLA-A* allele. Each letter indicates a single nucleotide and the numbers indicate the position of the SNP relative to the TSS.

	-1684	-1647	-1396	-1379	-1300	-1263	-1180	-1084	-993	-960	-956	-940	-923	-906	-892	-885	-878	-860	-833	-822	-812	-809	-808	-789	-772	-716	-617	-415	-413	-407	-226	-155	-142	-55
A*01	T	T	C	G	A	T	G	A	G	C	C	C	C	T	A	G	T	G	A	C	G	A	G	T	G	T	T	A	T	G	G	C	A	T
A*02	G	T	C	A	G	C	T	A	A	G	A	C	C	T	A	C	T	C	T	T	A	C	T	G	A	C	T	T	C	A	G	T	T	G
A*03	T	T	C	G	A	T	G	A	G	C	C	C	C	T	A	G	T	G	A	C	G	A	G	T	G	T	T	A	T	G	G	C	A	T
A*11	T	T	C	G	A	T	G	A	G	C	C	C	C	T	A	G	T	G	A	C	G	A	G	T	G	T	T	A	T	G	G	C	A	T
A*23	G	C	T	A	G	T	T	A	A	G	A	C	C	T	A	C	T	C	T	T	A	C	T	G	A	C	T	T	C	G	G	T	T	T
A*24	G	C	T	A	G	T	T	A	A	G	A	C	C	T	A	C	T	C	T	T	A	C	T	G	A	C	T	T	C	G	G	T	T	T
A*25	G	C	C	A	G	C	T	A	A	G	A	C	C	T	G	C	C	C	T	T	A	C	T	G	A	C	C	T	C	A	A	T	T	G
A*26	G	C	C	A	G	C	T	A	A	G	A	C	C	T	G	C	C	C	T	T	A	C	T	G	A	C	C	T	C	A	A	T	T	G
A*29	G	C	T	A	G	T	T	C	G	C	A	A	T	C	A	G	T	C	T	T	A	C	T	G	G	C	T	T	C	G	G	T	T	G
A*30	T	T	C	G	A	T	G	A	G	C	C	C	C	T	A	G	T	G	A	C	G	A	G	T	G	T	T	A	T	G	G	C	A	T
A*31	G	C	T	A	G	T	T	C	G	C	A	A	T	C	A	G	T	C	T	T	A	C	T	G	G	C	T	T	C	G	G	T	T	G
A*32	G	C	T	A	G	T	T	C	G	C	A	A	T	C	A	G	T	C	T	T	A	C	T	G	G	C	T	T	C	G	G	T	T	G
A*33	G	C	T	A	G	T	T	C	G	C	A	A	T	C	A	G	T	C	T	T	A	C	T	G	G	C	T	T	C	G	G	T	T	G
A*34	G	T	C	A	G	C	T	A	A	G	A	C	C	T	G	C	C	C	T	T	A	C	T	G	A	C	C	T	C	A	G	T	T	G
A*66	G	C	C	A	G	C	T	A	A	G	A	C	C	T	G	C	C	C	T	T	A	C	T	G	A	C	C	T	C	A	A	T	T	G
A*68	G	T	C	A	G	C	T	A	A	G	A	C	C	T	G	C	C	C	T	T	A	C	T	G	A	C	C	T	C	A	G	T	T	G
A*74	G	C	T	A	G	T	T	C	G	C	A	A	T	C	A	G	T	C	T	T	A	C	T	G	G	C	T	T	C	G	G	T	T	G

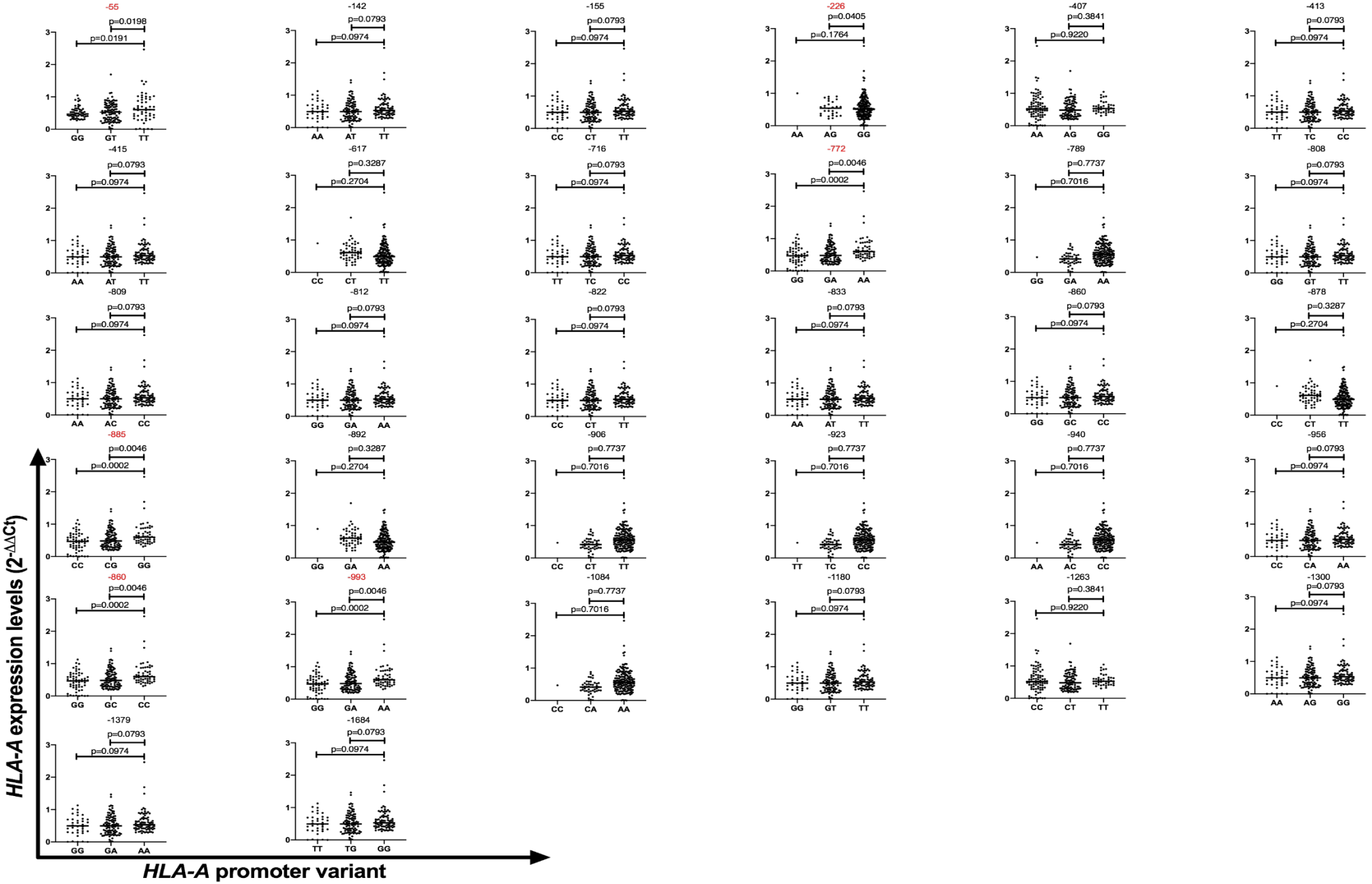
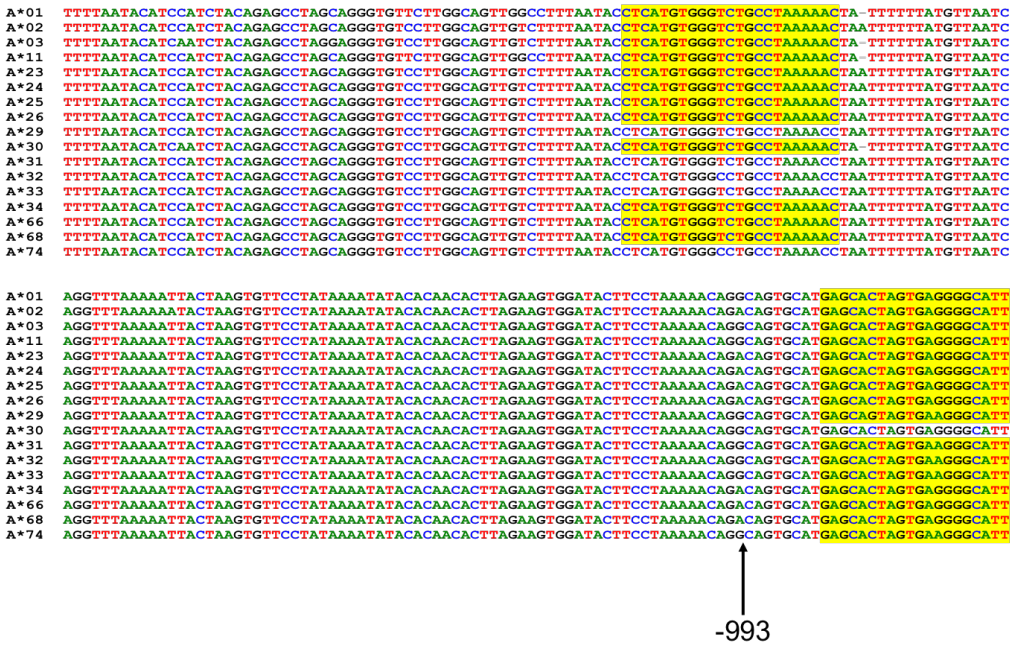

-55
p=0.0198
p=0.0191
GG GT TT
-142
p=0.0793
p=0.0974
AA AT TT
-155
p=0.0793
p=0.0974
CC CT TT
-226
p=0.0405
p=0.1764
AA AG GG
-407
p=0.3841
p=0.9220
AA AG GG
-413
p=0.0793
p=0.0974
TT TC CC
-415
p=0.0793
p=0.0974
AA AT TT
-617
p=0.3287
p=0.2704
CC CT TT
-716
p=0.0793
p=0.0974
TT TC CC
-772
p=0.0046
p=0.0002
GG GA AA
-789
p=0.7737
p=0.7016
GG GA AA
-808
p=0.0793
p=0.0974
GG GT TT
-809
p=0.0793
p=0.0974
AA AC CC
-812
p=0.0793
p=0.0974
GG GA AA
-822
p=0.0793
p=0.0974
CC CT TT
-833
p=0.0793
p=0.0974
AA AT TT
-860
p=0.0793
p=0.0974
GG GC CC
-878
p=0.3287
p=0.2704
CC CT TT
-885
p=0.0046
p=0.0002
CC CG GG
-892
p=0.3287
p=0.2704
GG GA AA
-906
p=0.7737
p=0.7016
CC CT TT
-923
p=0.7737
p=0.7016
TT TC CC
-940
p=0.7737
p=0.7016
AA AC CC
-956
p=0.0793
p=0.0974
CC CA AA
-860
p=0.0046
p=0.0002
GG GC CC
-993
p=0.0046
p=0.0002
GG GA AA
-1084
p=0.7737
p=0.7016
CC CA AA
-1180
p=0.0793
p=0.0974
GG GT TT
-1263
p=0.3841
p=0.9220
CC CT TT
-1300
p=0.0793
p=0.0974
AA AG GG
-1379
p=0.0793
p=0.0974
GG GA AA
-1684
p=0.0793
p=0.0974
TT TG GG
HLA-A expression levels (2-ΔΔCt)
HLA-A promoter variant

Figure S1. Unpaired t-test of wild type vs mutant variants of 34 *HLA-A* SNPs. Unpaired t-tests were performed on 34 SNPs found upstream of the *HLA-A* TSS using expression data from 216 healthy European American individuals from the Research Donor Program at the Frederick National Laboratory for Cancer Research. Only six SNPs, located at positions -55, -226, -772, -885, -860 and -993 (labelled in red) located relative to the TSS, significantly marked *HLA-A* expression. HLA-A mRNA expression was measured using a real-time PCR assay. The SNPs were identified from 2KB alignments of common homozygous *HLA-A* alleles.

Table S3: Table of primers used for sequencing and Real Time PCR

Task	Primer name	Primer sequence
Sequencing	HLA-A_PRO_F1	TATCCCCTCATATGCTCAAGTG
	HLA-A_PRO_F2	GGACTCACACAGAAACTCAGAGCTA
	HLA-A_PRO_F3	CCAGGCGTGGCTCTCA
	HLA-A_PRO_R1	CTCCCACTCCTTACCTGTCCA
	HLA-A_PRO_R2	GGAAGTATCCACTTCTAAGTGTTGTG
	HLA-A_PRO_R3	CAGGCACTTGAGCATATGAGG
PCR	HLA-A F1	CTCATGTGGGTCTGCCTAAAAAC
	HLA-A R1	AATGCCCCTCACTAGTGCTC
	IGF/H19 F1	GTGGCTCCCATGAGTGTTCT
	IGF/H19 R1	AGTTGTGGAATCGGAAGTGG



Primer pair 1

	Sequence (5'->3')	Length	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CTCATGTGGGTCTGCCTAAAAAC	23	59.81	47.83	4.00	0.00
Reverse primer	AATGCCCCTCACTAGTGCTC	20	59.46	55.00	8.00	2.00

Figure S2. Sequence alignment highlighting the primer positions for -993G>A and primer BLAST showing that primers do not bind elsewhere in the genome. *HLA-A* primers specific for -993G>A bound to *HLA-A* alleles and not to any other site in the human genome. Six alleles (A*29, A*30, A*31, A*32, A*33 and A*74) were excluded when choosing samples for ChIP due to mutations within the primer binding site.

Chapter 3

Synthesis

HIV remains one of the largest disease burdens in Sub-Saharan Africa. In South Africa the HIV positive population continues to grow each year [57] despite the fact that free treatment is widely available. Two of the reasons for the continual increase in the HIV positive population is poor adherence to medication and the emergence of drug resistant HIV strains. Finding additional mechanisms of controlling the disease is therefore of utmost importance.

Host genetic factors have been shown to correlate with multiple diseases. Studies have also shown that high expression of HLA-C was associated with lower HIV viral loads [31]. In contrast, high mRNA expression of *HLA-A* was shown to correlate with high HIV viral loads and poor disease outcome [25]. This correlation is due to the fact that high mRNA expression of *HLA-A* stabilizes the expression of HLA-E on the cell surface [25]. HLA-E serves as a ligand for an inhibitory NK cell receptor. NK cells cytotoxic activity is therefore inhibited by the interaction [25]. Sites of variation in the *HLA-A* gene therefore have the potential to act as HIV drug target sites provided that they regulate mRNA expression.

Our study investigated the role of CTCF in *HLA-A* expression regulation since CTCF has a variety of functions in the human genome and plays a large role in the regulation of Class II. The variant, -993G>A, which was found within a putative CTCF binding site showed promise in terms of becoming a novel HIV drug target site. Following sequence analysis -993G/A, along with five other mutations (-960G>C, -885C>G, -772G>A, -226G>A and -55T>G), was shown to significantly mark the expression of *HLA-A.* Predicted CTCF binding determined using online predictive software showed differential CTCF binding only to -993G>A, making -993G>A the target of further analysis. However the ChIP assay showed that CTCF bound both -993A and -993G almost equally and CTCF does not regulate *HLA-A* expression at position -993.

To determine whether CTCF plays a role at all in *HLA-A* expression an experiment may be performed where CTCF is knocked down from *HLA-A* and expression is measured before and after the knockdown. Any change in expression would mean that CTCF does play a role in the regulation of *HLA-A*. This process can also be carried out for the other Class I genes such as *HLA-B* and *-C* to determine if CTCF regulates Class I as a whole the way it regulates Class II.

No transcription factors were predicted to bind to position -55 however this SNP may be regulating transcription in other ways. -55 lies 2bp outside the TATA box of *HLA-A* [47]. The TATA box is a core promoter region which is responsible for transcription initiation [62]. Some of these transcription factors include the general transcription factor family such as TFIID of which TATA binding protein is a subunit [62]. TFIID and TATA binding protein along with other factors make up the transcription preinitiation complex which is what causes transcription to occur [62]. The close proximity of -55 to

this transcription hub could mean that the variant present regulates transcription therefore further study should be done around the -55 SNP to determine its role in transcription regulation.

-226 showed differential binding of Rad1. Rad1 bound to G but not to A. Rad1 plays an important role in regulation of gene expression in yeast [63-66] however the role of this transcription factor is quite different in humans. In humans Rad1 does not regulate transcription, it instead acts as a platform for the 9-1-1 complex which functions in halting the cell cycle when DNA is damaged or DNA replication is incomplete [67, 68].

NF1 showed predicted differential binding to the CTCF binding site of -993G>A. NF1 traditionally acts as a transcriptional activator in both humans and viruses. There isn't much evidence of NF1 acting as a repressor however it has been shown that the level of NF1 binding is associated with the level of transcription [69]. When NF1 binds strongly to a gene the transcription levels of that gene are much higher than genes to which NF1 binds weakly [69]. Further analysis on the interaction between NF1 binding -993G/A needs to be performed as NF1 is a possible regulator of *HLA-A* expression.

Another transcription factor, GR, showed differential binding to -885C>G. GR is a ligand-activated nuclear receptor [70] which means that it needs to bind to steroid ligands such as cortisol and dexamethasone in order to perform its transcription factor activities. GR modulates transcription of genes in two ways: via the binding of receptor dimers to specific palindromic sequences called glucocorticoid response elements (GREs) and indirectly by interacting with other transcription factors e.g. Nuclear Factor kappa beta (NF-κβ) [71] and activator protein 1 (AP-1) [70, 72, 73]. Since GR is found on genes responsible for immune modulation, there have been highly effective drugs that target GR in order to reduce inflammation [70] making the binding of GR to -885C>G a good candidate for further study surrounding *HLA-A* expression regulation. PTF1B was found to bind to -772G>A regardless of the variant however this transcription factor is involved in pancreatic function and is not relevant to *HLA-A* regulation.

Both NF1 and GR could be working in conjunction with one another through LD. LD was measured between the top four SNPs using LDmatrix. There is currently no LD information available for -993G>A however the other three SNPs were all in perfect LD with one another. Visually, -993G>A seems to be in perfect LD with the other three SNPs as well. -993 and -885 appear to be in perfect LD where whenever there is a G in -993 there is a G in -885. Whenever there is an A in -993 there is a C in -885. NF1 binds when there is a G present and binding is lost when there is an A present. GR binds when there is a G present and binding is lost when there is a C present. The perfect LD that may be present between -993G>A and -885C>G therefore warrants further study of both NF1 and GR binding for their potential roles in *HLA-A* expression regulation.

A possible limitation of the study is the small sample size. Using a much larger sample size for ChIP may yield different results. Another limitation is that the study was not performed across different ethnicities. People of different ethnic groups have a large variation in SNPs found within the human genome [58]. Different ethnicities also have varying susceptibility to diseases [59-61]. It is possible that variation in *HLA-A* expression and regulation may be witnessed across ethnicities and that CTCF could show differential binding in certain ethnicities. This may not be true because *HLA-A* expression was measured in Caucasians, Hispanics and Africans previously and there was no significant difference in expression data across those three ethnic groups [25]. A third limitation could be the cell type used. Due to sample availability, bulk PBMCs were used in the ChIP assay. There is no published data available on the regulation of *HLA-A* across the different cell types that make up PBMCs. There may be variation in *HLA-A* expression across cell types and therefore different modes of regulation across various mononuclear cells. It would have been useful to have measured expression in specific cell types that make up bulk PBMCs and determine the pattern of CTCF binding on those cells.

Regulating *HLA-A* expression has been a topic of interest since disease associations were made with *HLA-A* expression. Apart from HIV, the outcomes of other diseases such as autoimmune vitiligo [48] and various cancers [74, 75] are affected by *HLA-A* expression. High expression of *HLA-A*, especially *HLA-A*02:01*, is associated with increased risk of developing autoimmune vitiligo due to the tendency of *HLA-A*02:01* to present peptides of melanocytes that initiate an autoimmune response mediated by T cells [48, 76]. In colorectal cancer, patients with a lower expression of *HLA-A* tend to have better disease outcomes compared to patients in which *HLA-A* is not down-regulated [74]. The mechanism behind this is hypothesized to be due to a lack of NK cell inhibition therefore allowing NK cells to clear cancer cells which try to escape into circulation [74]. These findings support the need for *HLA-A* expression regulation because similar to HIV, the outcomes of autoimmune vitiligo and colorectal cancer will be improved by lowering the expression of *HLA-A*.

There are, however, findings that do not support the need to downregulate *HLA-A*. In prostate cancer, downregulation of *HLA-A* is linked with metastasis. Patients diagnosed with benign prostatic hyperplasia expressed *HLA-A* at high levels whereas those who had metastatic prostate carcinomas had either diminished or no *HLA-A* expression [75]. High expression of *HLA-A* is also required to regulate the activation of NK cells [25]. If *HLA-A* expression is diminished, HLA-E will not be sufficiently expressed on the cell surface leading to a decrease in the inhibition of NK cells. These sensitive mechanisms are in place to regulate NK cell responses and could have catastrophic effects if altered. There is still much that is unknown about the genetics of the immune response. Very few studies have looked at *HLA-A* expression as a whole in relation to diseases. Lowering the mRNA expression of *HLA-A* will lower the cell surface expression of the HLA-A protein ultimately reducing the amount of HLA-A restricted peptides that are presented to immune cells. Using mechanisms

which regulate the expression of *HLA-A* as a drug target for HIV could lead to off-target effects that would expose patients to other issues such as increased susceptibility to viral, bacterial or parasitic infections, loss of regulation or over-regulation of other areas of the immune response and even a diminished response to certain cancers e.g. prostatic carcinoma. All these above mentioned issues need to be taken into consideration and investigated before a site of *HLA-A* expression regulation can be used as a drug target site to improve HIV disease outcomes.

In summary, high expression of *HLA-A* has been linked with poor disease outcomes in those living with HIV. Regulating the expression of *HLA-A* will aid in slowing down the rate of disease progression to AIDS and will also aid in lowering viral loads. Apart from DNA methylation it remains unknown what the additional mechanisms of *HLA-A* mRNA expression regulation are, therefore stressing the need to find what additional regulatory mechanisms may exist. A variant, -993G>A, within a putative CTCF binding site approximately 1kb upstream of the *HLA-A* TSS marked expression. Although CTCF binding to -993G>A was confirmed, the lack of differential binding excluded CTCF as a regulator of *HLA-A* expression at that site. It was found that NF1 is predicted to bind differentially to -993G>A and GR is predicted to bind differentially to -885C>G, making these two variants and transcription factors good targets for further study on *HLA-A* regulatory factors especially since they appear to be in perfect LD with one another. Using *HLA-A* expression regulatory sites as an HIV drug target is warranted since this treatment could be used to prevent autoimmune vitiligo and improve the prognosis of colorectal cancer however it would worsen the outcomes of prostatic carcinomas and possibly worsen the outcomes of other cancers as well as viral, bacterial and parasitic diseases since *HLA-A* is after all an immune gene. Additional research and careful consideration of the effects of altering *HLA-A* expression needs to be carried out when using this gene to develop a drug.

In conclusion, this study found that CTCF does not regulate *HLA-A* expression at position -993. We did discover putative differential binding of NF and GR, both of which play a role in transcription regulation. Confirmation of the nature of the binding of these transcription factors could prove useful in understanding the mechanisms of *HLA-A* regulation. This study serves as a basis for further investigation into the regulatory mechanisms surrounding *HLA-A* in order to better understand the association between *HLA-A* expression and HIV.

References

1. WHO. *HIV/AIDS Global situation and trends*. 2017 [cited 2019 12/04]; Available from: https://www.who.int/gho/hiv/en/.
2. AVERT. *HIV AND AIDS IN EAST AND SOUTHERN AFRICA REGIONAL OVERVIEW*. 2017 [cited 2019 12/04]; Available from: https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/overview#footnote1_euodcq4.
3. UNAIDS. *AIDSinfo*. 2019 [cited 2019 30/11]; Available from: http://aidsinfo.unaids.org.
4. UNAIDS. *South Africa*. 2017 [cited 2019 12/04]; Available from: http://www.unaids.org/en/regionscountries/countries/southafrica.
5. Jaffe, H.W., D.J. Bregman, and R.M. Selik, *Acquired immune deficiency syndrome in the United States: the first 1,000 cases.* Journal of Infectious Diseases, 1983. **148**(2): p. 339-345.
6. Levy, J.A., *HIV pathogenesis: 25 years of progress and persistent challenges.* Aids, 2009. **23**(2): p. 147-160.
7. Damtie, D., et al., *Common opportunistic infections and their CD4 cell correlates among HIV-infected patients attending at antiretroviral therapy clinic of Gondar University Hospital, Northwest Ethiopia.* BMC research notes, 2013. **6**: p. 534-534.
8. Lange, J. and J. Ananworanich, *The discovery and development of antiretroviral agents.* Antivir Ther, 2014. **19**(Suppl 3): p. 5-14.
9. Wong, J.K. and S.A. Yukl, *Tissue reservoirs of HIV.* Current opinion in HIV and AIDS, 2016. **11**(4): p. 362.
10. Siliciano, R.F. and W.C. Greene, *HIV latency.* Cold Spring Harbor perspectives in medicine, 2011. **1**(1): p. a007096.
11. Eisele, E. and R.F. Siliciano, *Redefining the viral reservoirs that prevent HIV-1 eradication.* Immunity, 2012. **37**(3): p. 377-388.
12. Moyo, S., et al., *HIV drug resistance among virally unsuppressed respondents in the 5th South African National HIV Prevalence, Incidence, Behaviour and Communication Survey, 2017.* 2019.
13. Allers, K., et al., *Evidence for the cure of HIV infection by CCR5Delta32/Delta32 stem cell transplantation.* Blood, 2011. **117**(10): p. 2791-9.
14. Warren, M. *Second patient free of HIV after stem-cell therapy*. 2019 [cited 2019 30/11]; Available from: https://www.nature.com/articles/d41586-019-00798-3??utm_medium=affiliate&utm_source=commission_junction&utm_campaign=3_nsn6445_deeplink_PID7267413&utm_content=deeplink.
15. Rao, P.K.S., *CCR5 inhibitors: Emerging promising HIV therapeutic strategy.* Indian journal of sexually transmitted diseases and AIDS, 2009. **30**(1): p. 1-9.
16. Dorr, P., et al., *Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-*

human immunodeficiency virus type 1 activity. Antimicrob Agents Chemother, 2005. **49**(11): p. 4721-32.

17. Neil, S.J., T. Zang, and P.D. Bieniasz, *Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu.* Nature, 2008. **451**(7177): p. 425-30.
18. Van Damme, N., et al., *The interferon-induced protein BST-2 restricts HIV-1 release and is downregulated from the cell surface by the viral Vpu protein.* Cell Host Microbe, 2008. **3**(4): p. 245-52.
19. Sheehy, A.M., et al., *Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein.* Nature, 2002. **418**(6898): p. 646-50.
20. Stremlau, M., et al., *The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys.* Nature, 2004. **427**(6977): p. 848-53.
21. Martin, M.P. and M. Carrington, *Immunogenetics of HIV disease.* Immunological reviews, 2013. **254**(1): p. 245-264.
22. Marsh, S.G., P. Parham, and L.D. Barber, *The HLA factsbook*. 1999: Elsevier.
23. Moss, P., et al., *Extensive conservation of alpha and beta chains of the human T-cell antigen receptor recognizing HLA-A2 and influenza A matrix peptide.* Proceedings of the National Academy of Sciences, 1991. **88**(20): p. 8987-8990.
24. Kulkarni, S., et al., *Genetic interplay between HLA-C and MIR148A in HIV control and Crohn disease.* Proc Natl Acad Sci U S A, 2013. **110**(51): p. 20705-10.
25. Ramsuran, V., et al., *Elevated HLA-A expression impairs HIV control through inhibition of NKG2A-expressing cells.* Science, 2018. **359**(6371): p. 86-90.
26. Shiina, T., H. Inoko, and J. Kulski, *An update of the HLA genomic region, locus information and disease associations: 2004.* HLA, 2004. **64**(6): p. 631-649.
27. Thomson, G., *A review of theoretical aspects of HLA and disease associations.* Theoretical Population Biology, 1981. **20**(2): p. 168-208.
28. Risch, N., *Assessing the role of HLA-linked and unlinked determinants of disease.* Am J Hum Genet, 1987. **40**(1): p. 1-14.
29. Gough, S.C.L. and M.J. Simmonds, *The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action.* Current genomics, 2007. **8**(7): p. 453-465.
30. Apps, R., et al., *Relative expression levels of the HLA class-I proteins in normal and HIV-infected cells.* J Immunol, 2015. **194**(8): p. 3594-600.
31. Apps, R., et al., *Influence of HLA-C expression level on HIV control.* Science, 2013. **340**(6128): p. 87-91.
32. Thomas, R., et al., *A novel variant marking HLA-DP expression levels predicts recovery from hepatitis B virus infection.* Journal of virology, 2012. **86**(12): p. 6979-6985.
33. Borghans, J.A., et al., *HLA alleles associated with slow progression to AIDS truly prefer to present HIV-1 p24.* PLoS One, 2007. **2**(9): p. e920.
34. Weiskopf, D., et al., *Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8+ T cells.* Proc Natl Acad Sci U S A, 2013. **110**(22): p. E2046-53.

35. Rao, X., et al., *HLA Preferences for Conserved Epitopes: A Potential Mechanism for Hepatitis C Clearance.* Front Immunol, 2015. **6**: p. 552.

36. Malavige, G.N., et al., *HLA class I and class II associations in dengue viral infections in a Sri Lankan population.* PLoS One, 2011. **6**(6): p. e20581.

37. Thio, C.L., et al., *Racial differences in HLA class II associations with hepatitis C virus outcomes.* J Infect Dis, 2001. **184**(1): p. 16-21.

38. Ebringer, A. and C. Wilson, *HLA molecules, bacteria and autoimmunity.* J Med Microbiol, 2000. **49**(4): p. 305-11.

39. Singh, R.K. *HLA and Skin Diseases*. 2014 [cited 2019 29/11/19]; Available from: https://www.slideshare.net/RKSKUSHWAHA/hla-and-skin-disorders.

40. Matzaraki, V., et al., *The MHC locus and genetic susceptibility to autoimmune and infectious diseases.* Genome Biol, 2017. **18**(1): p. 76.

41. de Sorrentino, A.H., et al., *HLA class I alleles associated with susceptibility or resistance to human immunodeficiency virus type 1 infection among a population in Chaco Province, Argentina.* The Journal of infectious diseases, 2000. **182**(5): p. 1523-1526.

42. Mekue, L.M., et al., *HLA A* 32 is associated to HIV acquisition while B* 44 and B* 53 are associated with protection against HIV acquisition in perinatally exposed infants.* BMC pediatrics, 2019. **19**(1): p. 249.

43. Carrington, M., et al., *HLA and HIV-1: heterozygote advantage and B* 35-Cw* 04 disadvantage.* Science, 1999. **283**(5408): p. 1748-1752.

44. O'Connor, S.L., et al., *MHC Heterozygote Advantage in Simian Immunodeficiency Virus–Infected Mauritian Cynomolgus Macaques.* Science translational medicine, 2010. **2**(22): p. 22ra18-22ra18.

45. Pereyra, F., et al., *The major genetic determinants of HIV-1 control affect HLA class I peptide presentation.* Science, 2010. **330**(6010): p. 1551-7.

46. Ramsuran, V., et al., *Epigenetic regulation of differential HLA-A allelic expression levels.* Hum Mol Genet, 2015. **24**(15): p. 4268-75.

47. Ramsuran, V., et al., *Sequence and Phylogenetic Analysis of the Untranslated Promoter Regions for HLA Class I Genes.* J Immunol, 2017. **198**(6): p. 2320-2329.

48. Hayashi, M., et al., *Autoimmune vitiligo is associated with gain-of-function by a transcriptional regulator that elevates expression of HLA-A*02:01 in vivo.* Proc Natl Acad Sci U S A, 2016. **113**(5): p. 1357-62.

49. Thomas, R., et al., *HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C.* Nature genetics, 2009. **41**(12): p. 1290-1294.

50. Vince, N., et al., *HLA-C Level Is Regulated by a Polymorphic Oct1 Binding Site in the HLA-C Promoter Region.* Am J Hum Genet, 2016. **99**(6): p. 1353-1358.

51. Braud, V.M., et al., *HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C.* Nature, 1998. **391**(6669): p. 795-9.

52. Browne, S.K., et al., *Differential IFN-γ stimulation of HLA-A gene expression through CRM-1-dependent nuclear RNA export.* The Journal of Immunology, 2006. **177**(12): p. 8612-8619.

53. Kulkarni, S., et al., *Posttranscriptional Regulation of HLA-A Protein Expression by Alternative Polyadenylation Signals Involving the RNA-Binding Protein Syncrip.* J Immunol, 2017. **199**(11): p. 3892-3899.

54. Filippova, G.N., et al., *An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes.* Mol Cell Biol, 1996. **16**(6): p. 2802-13.

55. Majumder, P. and J.M. Boss, *CTCF controls expression and chromatin architecture of the human major histocompatibility complex class II locus.* Molecular and cellular biology, 2010. **30**(17): p. 4211-4223.

56. Majumder, P., et al., *The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions.* J Exp Med, 2008. **205**(4): p. 785-98.

57. statssa.gov. *Mid-year population estimates 2018*. 2018 [cited 2019 21/10]; Available from: https://www.statssa.gov.za/publications/P0302/P03022018.pdf.

58. Huang, T., Y. Shu, and Y.-D. Cai, *Genetic differences among ethnic groups.* BMC genomics, 2015. **16**: p. 1093-1093.

59. Lau, C., G. Yin, and M. Mok, *Ethnic and geographical differences in systemic lupus erythematosus: an overview.* Lupus, 2006. **15**(11): p. 715-719.

60. Ackerman, M.J., et al. *Ethnic differences in cardiac potassium channel variants: implications for genetic susceptibility to sudden cardiac death and genetic testing for congenital long QT syndrome*. in *Mayo clinic proceedings*. 2003. Elsevier.

61. URAYAMA, K.Y. and A. MANABE, *Genomic evaluations of childhood acute lymphoblastic leukemia susceptibility across race/ethnicities.* 臨床血液, 2014. **55**(10): p. 2242-2248.

62. Watson, J.D., *Molecular Biology of the Gene*. 1987: Benjamin/Cummings Publishing Company.

63. Challal, D., et al., *General Regulatory Factors Control the Fidelity of Transcription by Restricting Non-coding and Ectopic Initiation.* Mol Cell, 2018. **72**(6): p. 955-969.e7.

64. Wu, A.C.K., et al., *Repression of Divergent Noncoding Transcription by a Sequence-Specific Transcription Factor.* Mol Cell, 2018. **72**(6): p. 942-954.e7.

65. Wu, A.C.K. and F.J. Van Werven, *Transcribe this way: Rap1 confers promoter directionality by repressing divergent transcription.* Transcription, 2019. **10**(3): p. 164-170.

66. Goto, G.H., et al., *Binding of Multiple Rap1 Proteins Stimulates Chromosome Breakage Induction during DNA Replication.* PLoS Genet, 2015. **11**(8): p. e1005283.

67. Bao, S., et al., *Disruption of the Rad9/Rad1/Hus1 (9–1–1) complex leads to checkpoint signaling and replication defects.* Oncogene, 2004. **23**(33): p. 5586.

68. Parrilla-Castellar, E.R., S.J. Arlander, and L. Karnitz, *Dial 9–1–1 for DNA damage: the Rad9–Hus1–Rad1 (9–1–1) clamp complex.* DNA repair, 2004. **3**(8-9): p. 1009-1014.

69. Gronostajski, R.M., et al., *Stimulation of transcription in vitro by binding sites for nuclear factor I.* Nucleic acids research, 1988. **16**(5): p. 2087-2098.

70. Muzikar, K.A., N.G. Nickols, and P.B. Dervan, *Repression of DNA-binding dependent glucocorticoid receptor-mediated gene expression.* Proceedings of the National Academy of Sciences, 2009. **106**(39): p. 16598-16603.

71. McKay, L.I. and J.A. Cidlowski, *Cross-talk between nuclear factor-kappa B and the steroid hormone receptors: mechanisms of mutual antagonism.* Mol Endocrinol, 1998. **12**(1): p. 45-56.

72. Heck, S., et al., *A distinct modulating domain in glucocorticoid receptor monomers in the repression of activity of the transcription factor AP-1.* The EMBO journal, 1994. **13**(17): p. 4087-4095.

73. De Bosscher, K., W. Vanden Berghe, and G. Haegeman, *The interplay between the glucocorticoid receptor and nuclear factor-kappaB or activator protein-1: molecular mechanisms for gene repression.* Endocr Rev, 2003. **24**(4): p. 488-522.

74. Menon, A.G., et al., *Down-regulation of HLA-A expression correlates with a better prognosis in colorectal cancer patients.* Laboratory investigation, 2002. **82**(12): p. 1725.

75. Lu, Q.L., et al., *Decreased HLA-A expression in prostate cancer is associated with normal allele dosage in the majority of cases.* J Pathol, 2000. **190**(2): p. 169-76.

76. Salazar-Onfray, F., et al., *Synthetic peptides derived from the melanocyte-stimulating hormone receptor MC1R can stimulate HLA-A2-restricted cytotoxic T lymphocytes that recognize naturally processed peptides on human melanoma cells.* Cancer research, 1997. **57**(19): p. 4348-4355.