Topics in cancer genomics

Sachet Ashok Shukla
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd
Part of the Bioinformatics Commons

Recommended Citation
Topics in cancer genomics

by

Sachet Ashok Shukla

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Co-majors: Bioinformatics and Computational Biology;
Statistics

Program of Study Committee:
Philip M. Dixon, Co-major Professor
Julie A. Dickerson, Co-major Professor
Alicia L. Carriquiry
Mark S. Kaiser
Dan Nettleton

Iowa State University
Ames, Iowa
2014
Copyright © Sachet Ashok Shukla, 2014. All rights reserved.
DEDICATION

To my family, especially the rising stars of the Shukla clan - Shivansh, Shivika, Rishabh and Shivali
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>xv</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xvi</td>
</tr>
<tr>
<td>CHAPTER 1. GENERAL INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 HLA typing and related applications</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 HLA genes and their function</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2 HLA diversity and nomenclature</td>
<td>2</td>
</tr>
<tr>
<td>1.1.3 Whole exome sequencing</td>
<td>3</td>
</tr>
<tr>
<td>1.1.4 Motivation</td>
<td>3</td>
</tr>
<tr>
<td>1.2 microRNAs and gene expression regulation</td>
<td>4</td>
</tr>
<tr>
<td>1.2.1 Motivation</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Dissertation organization</td>
<td>5</td>
</tr>
<tr>
<td>1.4 Contribution</td>
<td>5</td>
</tr>
<tr>
<td>CHAPTER 2. DETECTION OF SOMATIC CHANGES OF THE CLASS I HLA LOCIS USING EXOME SEQUENCING</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Methods</td>
<td>7</td>
</tr>
<tr>
<td>2.2.1 Developing POLYSOLVER</td>
<td>7</td>
</tr>
<tr>
<td>2.2.2 Alignment</td>
<td>8</td>
</tr>
<tr>
<td>2.2.3 First stage allele inference</td>
<td>11</td>
</tr>
<tr>
<td>2.2.4 Posterior probability calculations</td>
<td>12</td>
</tr>
</tbody>
</table>
2.2.5 Second stage allele inference ........................................... 15
2.2.6 Detection of somatic mutations in HLA genes using POLYSOLVER . 15
2.2.7 Power calculation for RNA-Seq validation ................................. 17
2.3 Results ................................................................................. 18
  2.3.1 Training set ..................................................................... 18
  2.3.2 Validation of POLYSOLVER .............................................. 18
  2.3.3 Comparison to ATHLATES .............................................. 19
  2.3.4 Comparison with HLA mutations identified using standard methods . 20
  2.3.5 Mutational spectrum of HLA class I genes and their functional effects . 22
  2.3.6 Patterns of HLA mutations across tumor types .......................... 26
2.4 Discussion ............................................................................. 27

CHAPTER 3. SYSTEMATIC IDENTIFICATION OF PERSONAL TUMOR-
SPECIFIC NEOANTIGENS IN CHRONIC LYMPHOCYTIC LEUKEMIA 29
3.1 Introduction ........................................................................... 29
3.2 Methods ................................................................................. 31
  3.2.1 A pipeline for the systematic identification of tumor neoantigens . . . . . . . 31
  3.2.2 Analysis of CLL gene expression ........................................ 34
3.3 Results .................................................................................. 34
  3.3.1 Frequency of missense and frameshift mutations in 91 CLL samples . . . . 34
  3.3.2 Somatic missense mutations were predicted to generate neopeptides that
       bind personal HLA class I alleles ........................................... 35
  3.3.3 More than half of predicted HLA-binding neopeptides showed direct bind-
       ing to HLA proteins in vitro .................................................. 36
  3.3.4 80% of mRNAs encoding predicted neoantigens are expressed in CLL . 37
  3.3.5 Large numbers of candidate neoantigens were predicted across diverse
       cancers .............................................................................. 37
3.4 Discussion ............................................................................. 38
CHAPTER 4. MICRORNA REGULATORY NETWORK INFERENCE IDENTIFIES MIR-34A AS A NOVEL REGULATOR OF TGF-β SIGNALING IN GLIOBLASTOMA

4.1 Introduction .............................................. 47
4.2 Methods .................................................. 50
  4.2.1 Application of CLR to the TCGA data set ..... 50
  4.2.2 Filters ................................................ 52
  4.2.3 Identification of subtype-specific edges ...... 54
4.3 Results ..................................................... 55
  4.3.1 Copy number filter .................................. 55
  4.3.2 Direct targets filter ................................ 56
  4.3.3 Subtype expression differences drive the identification of edges in the network ..... 57
  4.3.4 Discriminatory miRs between the PN and MS subtypes .... 57
  4.3.5 miR-34a ............................................. 59
  4.3.6 mir-34a and the transcription factor hypothesis .... 61
  4.3.7 Identification of subtype-specific edges ..... 61
4.4 Discussion .................................................. 62
  4.4.1 Shortcomings of the CLR algorithm ............. 64
  4.4.2 Integrated model .................................... 65
  4.4.3 Results ............................................. 66

CHAPTER 5. SUMMARY ............................................. 75

  5.1 HLA typing and mutation detection .................. 75
  5.2 Identification of personalized neoantigens in cancer patients ........ 76
  5.3 microRNA network inference .......................... 76

BIBLIOGRAPHY .................................................. 78
LIST OF TABLES

Table 2.1 Test set: 8 CLL samples .......................................................... 18

3.1 Distribution of mutation classes, summed neoORF sizes and number of predicted binding peptides across 13 cancers .................................................. 38

3.2 HLA-peptide binding affinities of known functionally derived immunogenic mutated epitopes across human cancers using NetMHCpan. Epitopes from missense mutations (NSCLC: non-small cell lung cancer; MEL: melanoma; CLL: chronic lymphocytic leukemia; RCC: clear cell renal carcinoma; BLCA: bladder cancer; NR: not reported;). Yellow: IC50 < 150 nM, green: IC50 150–500 nM and grey: IC50 > 500 nM. We discovered that these epitopes fell into 4 groups. Group 1: Mutated and wildtype peptides with IC50<150 nM; Group 2: mutated and wildtype peptides with IC50 150–500 nM; Group 3: mutated peptide with IC50 < 150 and wildtype peptide with IC50 > 500nM and Group 4: mutated and wildtype peptides with IC50 > 500nM. ........................................... 41

3.3 HLA-peptide binding affinities of known functionally derived immunogenic mutated epitopes across human cancers using NetMHCpan: Minor alleles. Epitopes from minor histocompatibility antigens (MM: multiple myeloma; HM: hematological malignancy; B-ALL: B cell acute lymphocytic leukemia). ................................................................. 42

4.1 Overlap of mir-34a targets with subtype signature genes ....... 59

4.2 Integrating copy number and direct targets filters with the network defined by the 8 miRs ................................................................. 60
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3</td>
<td>Survival analysis results</td>
<td>60</td>
</tr>
<tr>
<td>4.4</td>
<td>Transcription Factor Analysis</td>
<td>62</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 2.1  Relationship between GC%, informative site density and coverage in a typical HLA gene ............................... 8
Figure 2.2  Example alignment of reads to the HLA gene sequences in the reference assembly ........................................... 9
Figure 2.3  Schematic of the POLYSOLVER algorithm. .................................................. 10
Figure 2.4  Specificity of tag libraries of different lengths .................................................. 12
Figure 2.5  Framework for detection of somatic changes in HLA genes using POLYSOLVER. Mutation detection algorithms Mutect and Strelka were incorporated for calling point mutations and indels, following MHC-I typing of the germline by POLYSOLVER. .................................................. 16
Figure 2.6  Application of POLYSOLVER of 133 HapMap samples. ................................. 19
Figure 2.7  Comparison of POLYSOLVER and ATHLATES. POLYSOLVER has good accuracy even for low coverage samples. Both algorithms were applied to the 133 HapMap set and the results compared by library size. .................................................. 20
Figure 2.8  Comparison of somatic mutations identified by TCGA across cancers using standard approaches to those identified by the POLYSOLVER framework (n=2688). .................................................. 21
Figure 2.9  Comparison of HLA mutations identified by standard methods and POLYSOLVER by tumor type (SKCM = Melanoma, LUSC = Lung squamous cell carcinoma, LUAD = Lung adenocarcinoma, BLCA = Bladder, HNSC = Head and neck, COAD = Colon adenocarcinoma, READ = Rectum adenocarcinoma, UCEC = Uterine, GBM = Glioblastoma multiforme, OV Ovarian, BRCA = Breast, CLL = Chronic lymphocytic leukemia). ...

Figure 2.10  Validation of mutations at the transcriptomic level. Mutations identified by POLYSOLVER using DNA sequencing were validated at the transcriptome level using RNA sequencing of the same samples. ...

Figure 2.11  Distribution of HLA mutations across exons. ...

Figure 2.12  Comparison of spectrum of mutations in non-HLA genes and HLA genes. The ratio of number of mutations of a particular types to the number of silent mutations is compared between the non-HLA and HLA genes for all mutation types. ...

Figure 2.13  HLA mutations in contact residues. Contact residues in the HLA molecules are in actual physical contact with the peptide. ...

Figure 2.14  Pan-cancer distribution of HLA mutations across functional domains/exons. Distribution of HLA mutations across functional domains and tumor types. Potential loss-of-function events were found to be distributed across the length of the protein, denoted by the blue (out of frame indels) and red (nonsense mutations) dots across the top. The histogram immediately below summarizes the number of events identified at each positions. The vertical panel on the left shows the expected versus observed number of mutations for each disease type. The central panel shows patterns of mutations detected in each tumor type. There were several positions with recurrent events; the lowermost panel depicts recurrent positions with frequency $>= 3$. The disease and allele group of the detected mutations are also indicated in this panel. ...
Figure 3.1 Schematic representation of a strategy to systematically discover tumor neoantigens. Tumor specific mutations in cancer samples are detected using whole-exome (WES) or whole-genome sequencing (WGS) and identified through the application of mutation calling algorithms (such as Mutect) (Cibulskis et al. (2013)). Next, candidate neoepitopes can be predicted using well-validated algorithms (NetMHCpan) and their identification can be refined by experimental validation for peptide-HLA binding and by confirmation of gene expression at the RNA level. These candidate neoantigens can be subsequently tested for their ability to stimulate tumor-specific T cell responses.

Figure 3.2 Frequency of classes of point mutations that have the potential to generate neoantigens in CLL. Analysis of WES and WGS data generated from 91 CLL cases reveals that (A) missense mutations are the most frequent class of the somatic alterations with the potential to generate neo-epitopes, while (B) frameshifts and (C) splice-site mutations constitute rare events.

Figure 3.3 The numbers and affinity distributions of peptides predicted from 31 CLL cases with available HLA typing. Patients expressing the 8 most common HLA -A, -B alleles in the Caucasian population are marked in grey.
Figure 3.4 Application of the NetMHCpan prediction algorithm to functionally-defined neoepitopes and CLL cases. (A) The amino acid sequences of 33 functionally identified cancer neoepitopes reported in literature (table 3.2, Table 3.3) were tested by NetMHCpan, and the resulting predicted binding (IC50) to their known restricting HLA allele are presented, sorted on the basis of predicted binding affinity. (B) Distribution of the number of predicted peptides with HLA binding affinity < 150 nM (black) and 150–500 nM (grey) across 31 CLL patients with available HLA typing information (Table 3.3). (C) Peptides with predicted binding (IC50 < 500 nM by NetMHCpan) from 4 patients were synthesized and tested for HLA -A and -B allele binding using a competitive MHC I allele-binding assay. The percent of predicted peptides with evidence of experimental binding (IC50 < 500 nM) are indicated. (D) The distribution of gene expression for all somatically mutated genes (n=347) from 26 CLL patients, and for the subset of gene mutations encoding neoepitopes with predicted HLA binding scores of IC50 < 500 nM (n=180). No-low: genes within the lowest quartile expression; medium: genes within the 2 middle quartiles of expression; and high: genes within the highest quartile of expression.
Figure 3.5  Estimation of tumor neoantigen load across cancers. (A) Box plots comparing overall somatic mutation rates detected across cancers by massively parallel sequencing. LUSC: Lung squamous cell carcinoma, LUAD: Lung adenocarcinoma, ESO AD: Esophageal adenocarcinoma, DLBCL: Diffused large B-cell lymphoma, GBM: Glioblastoma, Papillary RCC: Papillary renal cell carcinoma, Clear Cell RCC: Clear cell renal carcinoma, CLL: Chronic lymphocytic leukemia, AML: Acute myeloid leukemia. The distributions (shown by box plot) of (B) the number of missense, frameshift and splice-site mutations per case across 13 cancers, (C) the summed neoORF length generated per sample and of the predicted neopeptides with IC50 < 150 nM (D) and with < 500 nM (E) generated from missense and frameshift mutations. For all panels, the left and right ends of the boxes represent the 25th and 75th percentile values, respectively, while the segment in the middle is the median. The left and right extremes of the bars extend to the minimum and maximum values.

Figure 4.1  CLR characteristics. The blue line shows the distribution of measures of association (mutual information and correlation) between all miR:mRNA pairs. The red line shows the distribution of measures of association values for all CLR-identified miR:mRNA pairs.
Figure 4.2  CLR network grouped by connectivity among miRs and mRNAs in regions of copy number aberration. Size of the node represents the number of edges between the miR and mRNAs in each group. There are 9 sub-networks represented. Edges in the network where both participating miR and mRNA are in regions of copy number alterations account for less than 3.9% of the total connectivity (green circles). Edges where either the participating miR or the mRNA is in region of gain or loss, but not both, account for 34.1% of the network (orange circles). Approximately 62% of the network involves miRs and mRNAs without evidence of genomic alterations in glioblastoma (blue circle).

Figure 4.3  Network of CLR edges where both miR (diamonds) and mRNA (circles) are in regions of copy number gain (purple nodes) or loss (green nodes).

Figure 4.4  Putative direct-target subnetworks among CLR-identified edges in glioblastoma. miRs are represented as green diamonds and genes as circles; classical signature genes (gray), neural signature genes (dark green), proneural signature genes (red), mesenchymal signature genes (blue), genes not in any of the molecular subtype signatures (pink). CL, classical; MS, mesenchymal; NL, neural; PN, proneural.

Figure 4.5  mir-34a:PDGFRA edge. Proneural samples have high levels of PDGFRA and low levels of mir-34a and the reverse trend is apparent for the mesenchymal samples.

Figure 4.6  Heatmap of correlations between 8 most discriminative miRs with proneural (PN) and mesenchymal (MS) signature genes in the CLR network. Each cell in the heatmap is the correlation between the corresponding miR and signature gene across samples.
Figure 4.7  mir-34a blocks tumorigenesis in vivo. Nude mice were transplanted with premalignant proneural-like cells and transduced with 6 pre-miRs corresponding to 8 discriminant miRs or the GFP control (n = 10 per group). miR-34a and miR-142 potently impaired tumorigenesis when compared with GFP control.

Figure 4.8  Prognostic value of mir-34a within the proneural subtype. KaplanMeier survival analysis of patients in the TCGA data set (n = 271). Low expression of miR-34a stratifies a subgroup of patients with proneural glioblastoma with a significant survival advantage compared with others in TCGA cohort (log-rank $P = 1.12 \times 10^{-5}$).

Figure 4.9  mir-34a model of action in GBM

Figure 4.10  Subtype-specific edge example: mir-195:SPRY4

Figure 4.11  Distribution of p-values by CLR. The complex correlation structure introduced by CLR results in a significantly skewed distribution.

Figure 4.12  Distribution of p-values by integrated modeling. The distribution is well-behaved and has a shape similar to the theoretically expected mixture of a uniform and a beta.

Figure 4.13  Top 1% of the inferred regulatory network (176/17539 edges)
First and foremost, I would like to thank my major professor Philip Dixon for his patient support during my rather unconventional road toward the Ph.D. I am deeply grateful to him for taking me on as his student and for going the extra mile to compensate for my physical absence at ISU on numerous occasions. I would also like to thank Cathy Wu and Gad Getz who have been my mentors for the HLA work and have provided critical biological and computational insights. I am grateful to Lynda Chin for introducing me to high throughput cancer genomics and providing guidance on the network inference project. Mark Kaiser, Alicia Carriquiry, Julie Dickerson, Dan Nettleton and Volker Brendel - thank you for holding me to a high standard and not letting me off the hook easily! Finally, I would like to express my profound gratitude to my mother for her unwavering faith in me, and to Mahima whose steadfast commitment to our family underlies all of my endeavors.
ABSTRACT

Large-scale projects such as the The Cancer Genome Atlas (TCGA) have generated extensive exome libraries across several disease types and populations. Detection of somatic changes in HLA genes by whole-exome sequencing (WES) has been complicated by the highly polymorphic nature of these loci. We developed a method POLYSOLVER (POLYmorphic loci reSOLVER) for accurate inference of class I HLA-A, -B and -C alleles from WES data, and achieved 97% accuracy at protein level resolution when this was applied to 133 HapMap samples of known HLA type. By applying POLYSOLVER in conjunction with somatic change detection tools to 2688 tumor/normal pairs TCGA that were previously analyzed by conventional approaches, we re-discovered 37 of 56 (66%) HLA mutations, while further identifying 23 new events. An analysis of WES data from a larger set of 3768 tumor/normal pairs by POLYSOLVER revealed 131 class I mutations with an enrichment for potentially loss-of-function events. 3% of samples had at least one HLA event with 95 of 131 mutations in the T cell interacting and peptide binding domains. Recurrent hotspot sites of missense, nonsense and splice site mutations were discovered that suggest positive selection, and support immune evasion as an important pathway in cancer.

Exome sequencing has also revealed a large number of shared and personal somatic mutations across human cancers. In principle, any genetic alteration affecting a protein-coding region has the potential to generate mutated peptides that are presented by surface HLA class I proteins that might be recognized by cytotoxic T cells. Utilizing POLYSOLVER in conjunction with knowledge of mutations in other genetic loci inferred from exome data, we developed a pipeline for the prediction and validation of such neoantigens derived from individual tumors and presented by patient-specific alleles of the HLA proteins. We applied our computational pipeline to 91 chronic lymphocytic leukemias (CLL) that had undergone whole-exome sequencing. We predicted ∼22 mutated HLA-binding peptides per leukemia (derived from ∼16 missense
mutations), and experimentally confirmed HLA binding for \( \sim 55\% \) of such peptides. Finally, we computationally predicted HLA-binding peptides with missense or frameshift mutations for several cancer types and predicted dozens to thousands of neoantigens per individual tumor, suggesting that neoantigens are frequent in most tumors. The neoantigen prediction pipeline can also elucidate the neoantigens unique to a particular cancer patient and help in the design of personalized immune vaccines.

MicroRNAs (miRs) are a class of non-coding small RNAs that regulate gene expression by promoting mRNA degradation or by inhibiting mRNA translation. Context Likelihood of Relatedness (CLR) is genetic network reconstruction method that considers the local network context in assessing the significance of connections while also allowing for detection of non-linear associations. Leveraging TCGA multidimensional data in glioblastoma, we inferred the putative regulatory network between microRNA and mRNA using the CLR algorithm. Interrogation of the network in context of defined molecular subtypes identified 8 microRNAs with a strong discriminatory potential between proneural and mesenchymal subtypes. Integrative in silico analyses, a functional genetic screen, and experimental validation identified miR-34a as a tumor suppressor in proneural subtype glioblastoma. Mechanistically, in addition to its direct regulation of platelet-derived growth factor receptor-alpha (PDGFRA), promoter enrichment analysis of CLR–inferred mRNA nodes established miR-34a as a novel regulator of a SMAD4 transcriptional network. Clinically, miR-34a expression level is shown to be prognostic, where miR-34a low-expressing glioblastomas exhibited better overall survival. This work illustrates the potential of comprehensive multidimensional cancer genomic data combined with computational and experimental models to enable mechanistic exploration of relationships among different genetic elements across the genome space in cancer.
CHAPTER 1. GENERAL INTRODUCTION

Cancer is the second leading cause of death in the U.S. with 574,743 deaths recorded in 2011 (http://www.cdc.gov/nchs/fastats/lcod.htm). Recent technological advances have enabled comprehensive characterization of genomic, expression, epigenetic and structural changes accompanying this disease. This work addresses two exigent issues in current cancer research. In the first part we explore the possibility of high fidelity typing of the highly polymorphic HLA gene loci using whole exome data and expound on two applications: detection of somatic changes in HLA genes and identification of personal tumor-specific neoantigens. The second part elucidates an integrative framework that utilizes mRNA and microRNA expression in conjunction with copy number and methylation data to infer the regulatory circuits of microRNAs.

1.1 HLA typing and related applications

The Human Leukocyte Antigen (HLA) locus in the human genome comprises a set of highly polymorphic genes that play an important role in the immune response. Exact determination of HLA gene variants in an individual is known as HLA typing and has numerous applications including identification of compatible organ donors, understanding autoimmunity and immune biology, and design of personalized cancer vaccines. HLA typing is typically a focused effort informed by directed experimental protocols. Whole exome sequencing (WES, capture sequencing), on the other hand, is a widely used technique for high-throughput sequencing of the coding regions of genes across the genome. Although the use of WES as a research and clinical tool is expanding exponentially, the non-specificity and relatively low-fidelity of WES compared to directed experimental protocols makes it challenging to use this strategy for HLA
typing. This work describes a computational method that enables high precision HLA typing using WES data.

1.1.1 HLA genes and their function

The HLA genes are important effectors of the immune response in humans and are found in all nucleated cells of the body. They bind aberrant endogenously processed peptides and present them for recognition by the T-cells thereby inducing an immune response. The adaptive immune response in humans depends upon the recognition of antigens presented either externally to the system or those that correspond to endogenous foreign peptides. The latter can arise when invading pathogens like viruses have successfully penetrated host cells through direct infection or internalization by phagocytosis and their proteins are present internally in the host cells. Host proteins that have acquired somatic changes through errors in DNA replication or in the evolution of diseases such as cancer may also give rise to intracellular antigens. All such aberrant proteins are processed by the cellular machinery into fragments that are 8-10 amino acids long and subsequently presented on the cell surface bound to specialized proteins collectively referred to as the ‘major histocompatibility complex’ (MHC). The destruction of such cells is critically mediated by T cells that can recognize and respond to the MHC:peptide complexes. There are two classes of MHC genes - I and II - that bind peptides corresponding to proteins found in the cytosol and intracellular membrane-bound vesicles respectively. The class I and II MHC genes also elicit responses from different T cell components. MHC genes are also called human leukocyte antigen (HLA) genes in humans and the most well characterized MHC class I genes are HLA-A, HLA-B and HLA-C. HLA genes are located on chromosome 6 of the genome and have been implicated in several important roles such as graft-vs-host disease in transplants and autoimmune disorders in addition to their pivotal role in cell-mediated immune responses.

1.1.2 HLA diversity and nomenclature

HLA genes are inherited at birth and do not undergo widespread somatic changes in the life of the organism. There is widespread heterogeneity at these loci across the population however
and frequencies of HLA alleles vary by ethnicity. The IMGT database (http://www.ebi.ac.uk/ipd/imgt/hla/) is a repository of all known HLA allele sequences and has 2129 HLA-A, 2796 HLA-B and 1672 HLA-C recorded alleles (v3.10).

HLA allele nomenclature follows a convention where the name of the corresponding gene is suffixed by successive sets of digits that correspond to decreasing functional relevance. The first 2 sets of digits are the most important; the 1st set corresponds to the serological activity of the allele and the second set to the protein sequence. For example, consider allele HLA-A*02:01:01:01. All alleles that have the prefix HLA-A*02 (ex. HLA-A*02:01:01:01, HLA-A*02:03:01) will elicit similar immune responses. All alleles with the same first 2 sets of digits (ex. HLA-A*02:01:01:01, HLA-A*02:01:02) will also give rise to the exact same peptide sequence, i.e they differ from each other only by synonymous changes.

1.1.3 Whole exome sequencing

Whole exome sequencing (WES, also referred to as capture sequencing), is a ubiquitous technique that is used routinely in cancer analysis for characterization of somatic changes. The technique is designed to sequence most of the exons, which constitute the protein-coding regions in the genome, across the known set of genes. We have used Agilent’s SureSelect Target Enrichment Workflow which comprises shearing genomic DNA followed by capture of exonic regions with 120-mer biotinylated cRNA (complementary RNA) baits. The baits overlap in sequence within an exon and comprise about 45 MB in total. The targeted exome region is 33 MB in size. The capture step is followed by sequencing on Illumina machines generating 76 base pair long paired-end reads. The target coverage is 100x across the exome, with 80% of the targets having > 20x coverage.

1.1.4 Motivation

Large-scale projects such as the TCGA (The Cancer Genome Atlas) have already generated extensive exome libraries across several disease types and populations. Several issues with capture sequencing including low coverage and difficulty of alignment in the HLA region have prevented inference of HLA typing from capture data thus far. We describe here a high fidelity
method to harness existing exome data for HLA typing focusing on HLA-A, HLA-B and HLA-C genes. Knowledge of HLA alleles at a given locus would enable more sensitive mutation detection in these genes. Combined with the knowledge of mutations in other genetic loci it may further help elucidate potential immune escape mechanisms in cancer. Lastly, peptide binding predictions may help elucidate the neoantigens unique to a particular cancer patient and help in the design of personalized immune vaccines.

1.2 microRNAs and gene expression regulation

MicroRNAs (miRs) are a class of non-coding small RNAs produced by RNA polymerase II as hairpins of longer precursor RNAs that are subsequently processed to approximately 22-nucleotide-long fragments by RNase III enzymes, Drosha and Dicer. Mature miRs regulate gene expression by promoting mRNA degradation or by inhibiting mRNA translation. The connection between miRs and cancer was first implicated by their genomic alteration and dysregulated expression in various human tumors. Multiple miRs have since been identified to promote or suppress oncogenesis in various tumors, presumably by modulating gene expression in the oncogenic and tumor suppressor networks.

1.2.1 Motivation

Several global analyses of miR:gene relationships have relied either on sequence-based interaction based methods or on general correlation coefficient methods. Correlation based methods are restricted to detection of linear trends only. It would be revealing instead to utilize a more general measure of association that is also sensitive to non-linear patterns. Moreover it would be informative to study the significance of each miR:gene association in the context of the local network defined by all possible edges involving the miR and gene molecules under consideration. Such an approach would help to filter out those miRs and genes in the reconstructed miR-mRNA network that have spurious similarities with large numbers of other gene/miR species.
1.3 Dissertation organization

This dissertation comprises 3 main chapters, each corresponding to a journal article. Chapter 2 describes a novel algorithm for HLA typing and detection of somatic changes in HLA genes using whole exome data. Chapter 3 outlines a neoantigen discovery pipeline that implicitly uses the HLA typing algorithm developed in Chapter 2 and exemplifies the utility of this pipeline in identifying neoantigens unique to individual patients as well as in estimating the neoantigen load across different types of cancers. Chapter 4 describes a framework for inference of regulatory connections of microRNAs based on integration of high-throughput expression, copy number and methylation data with sequence based prediction databases.

1.4 Contribution

The chapters have been modified from the original journal articles to emphasize the aspects that have been my original contributions. Chapter 2 is a computational paper in which I have contributed to the conception and development of the POLYSOLVER algorithm and the corresponding mutation detection framework, RNA-Seq validation, pan-cancer analysis of mutations and assessment of functional impact of the discovered mutational spectrum. In Chapter 3, I was responsible for development of the neoantigen discovery pipeline from exome data, pan-cancer analysis and gene expression analysis. In Chapter 4, I have contributed to the conception and development of the overall methodology, strategy for identification of most discriminatory miRs between different subtypes of the disease, gene set enrichment, direct target prediction, promoter analysis and development of the new miR:mRNA regulatory network discovery approach that addresses the statistical issues discovered during application of the Context Likelihood of Relatedness algorithm.
CHAPTER 2. DETECTION OF SOMATIC CHANGES OF THE CLASS I HLA LOCI USING EXOME SEQUENCING

Modified from a paper in preparation

Sachet A. Shukla, Mohini Rajasagi, Philip M. Dixon, Grace Tiao, Kristian Cibulskis, Adam Kiezun, Carrie Sougnez, Catherine J. Wu*, Gad Getz*.

* denotes equal contribution

2.1 Introduction

Recent large-scale whole-exome sequencing (WES) efforts have revealed the existence and relative high frequency of somatic changes in HLA class I genes in head and neck cancer, lung cancer, and diffuse large B cell lymphoma (Stransky et al. (2011), TCGA (2012a), Lawrence et al. (2014)). The HLA locus, located on chromosome 6, is among the most polymorphic regions of the human genome, with thousands of documented alleles for each gene. Critical mediators of the cytotoxic T cell response, these class I alleles serve to present intracellular peptides on the cell surface by their encoded proteins in a form that can be recognized by the T cell receptor. The findings of recurrent somatic mutations in HLA genes have strongly suggested immune evasion as a positively selected pathway in the development and progression of certain cancers.

Each individual expresses six major MHC class I alleles, with 3 alleles each inherited from maternal and paternal genes. Conventionally, determination of HLA type is performed using serology- and/or PCR-based methods that are labor-intensive and time-consuming. Although
HLA typing information theoretically should be directly extractable from whole-exome sequencing (WES) data, a number of challenges to the precise and accurate detection of somatic HLA mutations are evident. First, the highly polymorphic nature of this locus is expected to lead to suboptimal alignment to canonical reference sequences which would generally be a misrepresentation of the true underlying individual haplotypes. Second, the HLA genes are GC-rich and therefore apt to suffer from poor read retrieval due to lowered capture efficiency and potential errors in sequencing, and subsequent difficulties in alignment. Overall, our hypothesis was thus that conventional mutation detection methods underestimate somatic events within this biologically important and complex locus.

To this end, we developed the algorithm POLYSOLVER (POLYmorphic loci reSOLVER), which enables high precision HLA typing while using relatively low coverage WES data. This was achieved through optimization of alignment followed by a novel model-based approach for accurate inference of HLA class I alleles. By comparing cancer WES data from 3768 samples against HLA typing inferred by POLYSOLVER from matched normal DNA, we improved the sensitivity and precision for detecting somatic changes in HLA genes compared to conventional mutation calling approaches. In principle, a similar approach may be applied to other highly polymorphic regions of the genome, such as the class II loci, MIC genes and TAP genes. Our approach is readily applicable to the growing body of WES datasets across human diseases that rely on or are affected by antigen recognition.

2.2 Methods

2.2.1 Developing POLYSOLVER

As we expected with any region that is GC-rich, we confirmed that the HLA gene loci within WES data from a well-characterized set of 8 Caucasian patients with chronic lymphocytic leukemia (CLL) (Wang et al. (2011)) demonstrated relatively poor capture efficiency and less efficient sequencing and alignment. As shown in Suppl Figure 1, we observed that the informative sites were preferentially enriched in regions of high GC content (p=5.7e-06) with each of the three class I genes displaying a decreasing GC% trend in the direction of transcript-
tion (Figure 2.1). We also noted a marked inverse correlation of GC content with coverage in each of HLA-A, B and C genes (corr=-0.51, p=4.6e-13) (median coverage at the HLA loci in the 8 cases was 59 vs 89 median coverage in non-HLA sites). Naive alignment of the capture reads to the canonical human genome in the HLA region was error-prone with only about 27% of the reads aligning without mismatches or indels (Figure 2.2).

Figure 2.1 Relationship between GC%, informative site density and coverage in a typical HLA gene

These observations motivated us to seek an alternate alignment strategy which would mitigate errors in alignment to hyperpolymorphic regions. We reasoned that optimal retrieval of reads derived from HLA genes followed by accurate alignment to a library of all known HLA alleles would serve as a solid foundation for subsequent computational inference. We therefore developed a two-stage approach to infer HLA in which the first stage was to precisely perform HLA alignments, while the second was to infer individual alleles by computing and ranking posterior probability-derived scores (Figure 2.3).

2.2.2 Alignment

For the first stage, we sought to achieve high precision in HLA alignment by generating a full-length genomic reference library of all known HLA alleles (6597 unique entries based on the IMGT database (Robinson et al. (2013))). Sequence data and its alignments to a reference
Figure 2.2  Example alignment of reads to the HLA gene sequences in the reference assembly

genome are usually stored as a BAM file, which is a binary version of the corresponding SAM file (Sequence Alignment/Map format (Li et al. (2009))). To ensure maximal retrieval of true HLA reads while restricting inclusion of reads that might result in spurious alignments to the HLA loci, we performed a pre-selection step in which all reads mapping within 1 Kb of the HLA genes were extracted from the processed BAM files. We moreover selected all reads perfectly matching at least one bait in either orientation of a secondary library of unique 38-mer baits generated from the HLA sequence library. Only the best-scoring alignments for each read were preserved and used in the subsequent inference steps.

2.2.2.1 Polymorphic allele library creation

We constructed a reference library of known HLA alleles (6597 unique entries) based on the IMGT database (v3.10; http://www.ebi.ac.uk/ipd/imgt/hla/). Missing exons for incompletely sequenced allele cDNAs were deduced by multiple sequence alignment of the missing allele with all available cDNA sequences. Missing introns were inferred by alignment of the cDNA sequence with the nearest full-length genomic sequence.
The final library comprised full-length genomic sequences of 2129 HLA-A, 2796 HLA-B and 1672 HLA-C alleles.

### 2.2.2.2 HLA read extraction and alignment from WES data

To maximally retrieve true HLA reads, we first extracted all reads mapping within 1 Kb of the HLA genes from the processed BAM files. We then performed a further extraction by using a secondary 38-mer HLA sequence library and 'fished' for any reads perfectly matching at least one tag in either orientation. We then realigned the pooled reads to the genomic reference allele library using Novocrafts Novoalign software (http://www.novocraft.com). All best-scoring allele matches for each read pair were preserved and used for subsequent allele inference. Reads selected by this process are subjected to following post-processing filters prior to inference of alleles: (i). reads with insertions or deletions are discarded (ii). both mates must match the allele in question.
2.2.2.3 Choice of tag library length

The choice of 38 mers for creation of HLA tag library was based on maximizing the specificity of the library for HLA genes while maintaining 100% sensitivity in the context of downstream read filtering. To assess the specificity of tag libraries of different lengths derived from the HLA genes, we created a sequence set of 21,000 non-polymorphic genes and recorded the fraction of non-polymorphic genes that matched at least one tag from the library. Specificity was defined as 1 minus this fraction (Figure 2.4). To account for the possibility of sequencing errors or genuine somatic mutations, we decided to preserve reads that had no more than one misalignment or indel reference to the allele reference. For 100% sensitivity, this implies that the tag library length be no more than half the length of the sequencing reads. If a misalignment happens at or close to the edges of a read, there would still be multiple tags of length=(read length/2) in the middle of the read which would guarantee its retrieval. In the extreme case, if a misalignment happens in the middle base (if the read length is odd) or in one of the middle two bases (if the read length is even), there would still be one tag of length=(read length/2) which would ensure it’s retrieval. Since we had 76 base pair reads, we decided to use the 38 mer tag library, which yielded a 23.3% specificity. The low specificity value is not an issue since non-specific reads are discarded in the post-alignment filtering prior to the posterior probability computations.

2.2.3 First stage allele inference

We devised a posterior probability derived score, which was a measure of how likely it was to see the observed set of HLA reads assuming a true underlying HLA allele. The two alleles for each HLA gene were inferred in two different stages. Posterior probability derived scores were calculated for each reference allele in the first stage. The likelihoods of all aligned read pairs to a given allele were computed based on their respective alignments, quality scores and insert size probabilities based on the empirical insert size distribution of all read pairs in the sequencing run. Population-based allele probabilities were used as priors in the model. The alleles that maximized the posterior probability derived scores were identified as one of the two alleles at each locus in the sample.
2.2.4 Posterior probability calculations

The posterior probability calculations for alleles corresponding to each HLA gene are performed separately. We calculate a posterior probability derived score for each allele as described below:

Let

\[ M \equiv \# \text{ alleles corresponding to the HLA gene} \]
\[ N \equiv \# \text{ reads aligning to at least one allele} \]
\[ N_m \equiv \# \text{ reads aligning to allele } a_m \]
\[ N_T \equiv \# \text{ reads in the sequencing run} \]
\[ f_m \equiv \text{population based prior probability of allele } m. \]

These frequencies were derived from HLA typing studies conducted in various ethnic groups. For each ethnic group, the population based frequency for an allele was estimated as the weighted average of the document frequencies.
in HLA typing studies for that ethnic group, where the weights were the number of individual in each study.

\( r_{k1} \equiv \) first segment of read pair \( r_k \). Each sequenced read pair comprises two segments (also called mates) that correspond to two ends of the same DNA fragment.

\( r_{k2} \equiv \) second segment of read pair \( r_k \)

\( d_k \equiv \) insert length of read pair \( r_k \)

\( l_{k1} \equiv \) length of first segment of read pair \( r_k \)

\( l_{k2} \equiv \) length of second segment of read pair \( r_k \)

\( q_i \equiv \) quality of sequenced base \( i \). The quality of each base is determined by the instrument at the time of sequencing and are finally stored in the BAM file.

\( e_i \equiv \) probability that the sequenced base \( i \) is an error

\[
e_i = 10^{-\frac{q_i}{10}} \quad (2.1)
\]

We can use the quality scores of the alignment to build a model for the error generating process. Let us say that a given read \( r_k \) does in fact derive from an allele \( a_m \) and their sequence relationship allowing for miscalls in the sequencing process is accurately captured in the alignment. Let \( Y_{Ai}, Y_{Ci}, Y_{Gi} \) and \( Y_{Ti} \) denote random variables corresponding to observing bases A, C, G and T respectively at position \( i \) in read \( r_k \) in its alignment to allele \( a_m \). Then

\[ Y_{Ai}, Y_{Ci}, Y_{Gi}, Y_{Ti} \sim Multinomial(n = 1; \alpha_{Ai}, \alpha_{Ci}, \alpha_{Gi}, \alpha_{Ti}) \quad (2.2) \]

where

\[ \alpha_{Bi} = 1 - e_i \text{ if reference base at position } i \text{ in } a_m \text{ is } B \]

\[ = \frac{e_i}{3} \text{ otherwise} \]

Let \( D \) denote a random variable for the observed insert length of a paired read in the sequencing run based on alignment to the complete genome. For a given read pair \( r_k \), the
empirical insert size distribution can be used to calculate the probability of observing the
insert length $d_k$ as

$$P(D = d_k) = \frac{\sum_{l=1}^{N_T} I(d_l = d_k)}{N_T} \quad (2.3)$$

Assuming positional independence of quality scores, and independence of generated reads
and their insert sizes, the probability of observing $r_k$ given allele $a_m$ is then

$$P(r_k|a_m) = \prod_{i=1}^{l_{k1}} \alpha_i \prod_{j=1}^{l_{k2}} \alpha_j \cdot P(D = d_k) \quad (2.4)$$

The posterior probability of allele $a_m$ is

$$P(a_m|r_1, r_2, \ldots, r_N) = \frac{\prod_{k=1}^{N} P(r_k|a_m) \cdot f_m}{\prod_{k=1}^{N} P(r_k)} \propto \prod_{k=1}^{N} P(r_k|a_m) \cdot f_m \quad (2.5)$$

where

$$P(r_k) = \sum_{m=1}^{M} P(r_k|a_m) f_m \quad (2.6)$$

We define a posterior probability derived score for allele $a_m$ based on the log transformation
of the above equation as

$$L_m = \sum_{k=1}^{N} \sum_{i=1}^{l_{k1}} \log \alpha_i + \sum_{k=1}^{N} \sum_{j=1}^{l_{k2}} \log \alpha_j + \sum_{k=1}^{N} \log P(D = d_k) + \log(f_m) \quad (2.7)$$

Read pairs that do not align to an allele are assumed to have zero probability of deriving
from that allele. The actual HLA alleles corresponding to the individual should have relatively
higher numbers of good alignments but direct score maximization would instead be biased
in favor of alleles with fewer mapped reads. This is accounted for by shifting the frame of
reference to $s_k$, where $s_k$ is the theoretical least score achievable for read pair $r_k'$ with perfect
base qualities and segment lengths equal to those of $r_k$. As we only preserve the best-scoring
alignments for each read pair, the contribution of a given read pair would be equal for all
matched alleles, and the read pair’s score contributions would maintain allele equality upon
shift of reference. Since 93 is the maximum achievable base quality under Illumina 1.8+ format, $s_k$ is computed as

$$s_k = (l_{k1} + l_{k2}) \cdot \log 10^{-9.3} \approx -23 \cdot (l_{k1} + l_{k2}) \quad (2.8)$$

The final posterior probability derived score is now given by

$$L_m = \sum_{k=1}^{N} l_{k1} \log \alpha_i + \sum_{k=1}^{N} l_{k2} \log \alpha_j + \sum_{k=1}^{N} \log P(D = d_k) + \log(f_m) - Ns_k \quad (2.9)$$

2.2.5 Second stage allele inference

Regarding the inference of the second class I HLA allele, a significant complicating factor stems from the fact that any HLA gene in a given individual may be either a homozygous or heterozygous allele. Since alleles corresponding to the same allele group or protein sequence tend to have highly similar DNA sequences, we observed that selecting the top two alleles when the posterior probabilities were simply sorted in order was incorrectly biased in favor of declaring homozygous winners. On the other hand, a complete depletion of reads mapping to the winning allele at the time of inference of the first allele followed by a recalculation of the posterior probabilities yielded only heterozygous calls. To balance between these two extremes such that we could reliably distinguish between homozygous and heterozygous second alleles, we developed a heuristic for inference of the second allele that accounted for its sequence similarity with the first inferred allele. This was accomplished by recalculating scores for the second allele by first identifying reads that were also aligned to the first allele followed by halving their alignment-based score contribution, while leaving the contribution of reads that were exclusively aligned to the second allele intact. As in the first stage, the alleles with maximal posterior probability derived scores for each gene were identified as the second alleles.

2.2.6 Detection of somatic mutations in HLA genes using POLYSOLVER

A common strategy for reliably identifying somatic mutations from WES data is to compare matched tumor and germline (normal, unaffected tissue) capture sequencing data from the same
individual while correcting for private SNPs within that same genome (Cibulskis et al. (2013)). We considered the possibility that certain true somatic HLA mutations may be missed because of lack of proper alignment in this region, which might be overcome by applying POLYSOLVER. As represented schematically in Figure 2.5, we tested this idea by first applying POLYSOLVER to germline WES data to infer the HLA type of the individual. This germline HLA sequence then served as a synthetic reference against which pre-selected HLA reads from matched tumor were compared, and mutations called by Mutect. We used Strelka (Saunders et al. (2012)) for detection of insertions and deletions.

![Figure 2.5](image)

**Figure 2.5** Framework for detection of somatic changes in HLA genes using POLYSOLVER. Mutation detection algorithms Mutect and Strelka were incorporated for calling point mutations and indels, following MHC-I typing of the germline by POLYSOLVER.

We first inferred the HLA type of the individual using the germline data. Reads that likely derived from the HLA genes were pre-selected as described before from both tumor and germline experiments. These two sets of reads were aligned separately to the HLA allele library. This effectively simulated a mini paired tumor/normal HLA capture experiment for the individual with the HLA allele library serving as a synthetic reference 'HLA genome' and each allele functioning as a 'chromosome'. The approach remains unaffected by the extremely high degree of homology across the alleles since all the best alignments for each read are reported. Since a significant portion of the reads mapped to multiple alleles, it was necessary to modify the tumor and germline HLA genome-aligned BAM files such that the `NotPrimaryAlignment` bit flag was
turned off and the mapping quality was changed to a non-zero value for all reads. These steps were required to accurately run mutation detection tools such as Mutect which would otherwise identify all alignments except one as secondary alignments, and therefore ignore them. Noisy alignments where both mates did not align to the same reference allele as well as those where at least one mate had more than one mutation, insertion or deletion event compared to the reference allele were discarded prior to mutation calling. The MuTect algorithm (Cibulskis et al. (2013)) was then applied to these BAMs to identify mutations and only those that were reported in the inferred alleles for the individual were preserved. The Strelka software (Saunders et al, 2012) was used for detection of somatic insertions and deletions in HLA genes. The detected somatic changes were screened against a panel of normal variants created based on 133 normal HapMap samples to reduce the rate of false positives. The remaining somatic events were annotated for the gene compartment (intron, exon, splice site) and protein change.

2.2.7 Power calculation for RNA-Seq validation

We sought orthogonal validation of detected mutations that were identified exclusively by POLYSOLVER, TCGA or commonly by both methods by probing available RNA-Seq data for the corresponding samples from the CGHub resource. A mutation was considered validated in RNA-Seq if there were at least 2 reads supporting the mutation. To determine the power, we first calculated the allelic fraction of the mutation based on the exome data ($p$). Then, given the total number of reads aligning at the position in the RNA-Seq data ($N$), power was calculated as the probability that we would detect at least 2 reads bearing the alternate allele in the RNA-Seq data under the binomial distribution $Binom(N, p)$ i.e.

$$Power = \sum_{k=2}^{N} \binom{N}{k} p^k (1 - p)^{N-k}$$  \hspace{1cm} (2.10)

Sites that had less than 80% power were considered under-powered and removed from the RNA-Seq validation pool.
2.3 Results

2.3.1 Training set

WES from cases of chronic lymphocytic leukemia (CLL) patients has been previously reported (Wang et al. (2011)). We used a subset of 8 CLL WES cases within this cohort for which conventional PCR-based HLA typing information was also available to develop and train this algorithm. By using the final developed method, we correctly identified 6 of 6 alleles in 7 cases and 5 of 6 alleles in 1 case. In total, 47 of 48 (97.9%) alleles were identified correctly at the protein level (Table 2.1). The misidentified allele had a population frequency of less than 0.05% and was therefore missed by the algorithm. The data set had one homozygous case that was correctly called.

Table 2.1 Test set: 8 CLL samples

<table>
<thead>
<tr>
<th>Sample</th>
<th>HLA type</th>
<th># correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLL-100</td>
<td>HLA-A01:01 HLA-A03:02 HLA-B08:01 HLA-B44:03 HLA-C05:01 HLA-C07:02</td>
<td>6</td>
</tr>
<tr>
<td>CLL-024</td>
<td>HLA-A03:01 HLA-A01:02 HLA-B14:02 HLA-B38:01 HLA-C08:02 HLA-C12:03</td>
<td>5</td>
</tr>
<tr>
<td>CLL-050</td>
<td>HLA-A02:01 HLA-A11:01 HLA-B44:02 HLA-B51:01 HLA-C05:01 HLA-C15:02</td>
<td>6</td>
</tr>
<tr>
<td>JP</td>
<td>HLA-A11:01 HLA-A24:02 HLA-B51:01 HLA-B15:01 HLA-C03:03 HLA-C15:02</td>
<td>6</td>
</tr>
<tr>
<td>CLL-023</td>
<td>HLA-A02:01 HLA-A03:01 HLA-B35:01 HLA-B57:01 HLA-C04:01 HLA-C06:02</td>
<td>6</td>
</tr>
<tr>
<td>CLL-014</td>
<td>HLA-A02:01 HLA-A32:01 HLA-B44:02 HLA-B15:01 HLA-C05:01 HLA-C07:04</td>
<td>6</td>
</tr>
<tr>
<td>CLL-095</td>
<td>HLA-A02:01 HLA-A02:01 HLA-B38:01 HLA-B51:01 HLA-C05:01 HLA-C12:03</td>
<td>6</td>
</tr>
<tr>
<td>CLL-084</td>
<td>HLA-A01:01 HLA-A02:01 HLA-B08:01 HLA-B44:02 HLA-C07:01 HLA-C05:01</td>
<td>6</td>
</tr>
</tbody>
</table>

Protein level accuracy 47 (97.9%)
Allele group accuracy 47 (97.9%)

2.3.2 Validation of POLYSOLVER

We applied POLYSOLVER to WES data from a set of 133 HapMap samples comprising 15 Caucasians, 42 Blacks, 41 Chinese and 35 Japanese individuals with known HLA types (Erlich et al. (2012); http://www.1000genomes.org)). As shown in Figure 2.6, 774 of 798 (97%) alleles were identified correctly at the protein level resolution while allele groups were correctly typed in
787 of 798 (98.7%) instances. There were no significant differences in performance by ethnicity (94.4% - 98.8%), HLA gene (95.9% - 98.1%) or allele inference stage (96.7% - 97.2%). There were 42 homozygous alleles in this set, all of which were correctly identified.

2.3.3 Comparison to ATHLATES

POLYSOLVER was compared to ATHLATES (accurate typing of human leukocyte antigen through exome sequencing) (Liu et al. (2013)), a program which infers HLA type from exome data using an assembly-based approach. As stated in the original description of this algorithm, ATHLATES requires higher coverage data than POLYSOLVER, with a minimum requirement of 175 million pair-end reads and \( \geq 10x \) coverage of at least 96% of the exons. By contrast, the bulk of TCGA data has been reported to be much lower, with a median (range) of reads/sample. Likewise, the median number reads of our validation set of 133 HapMap samples was 55 million reads (range 37 - 149 million). Indeed, ATHLATES failed to call 75% (597/798) of the alleles in the HapMap samples because of this relatively low coverage although its performance improved with increasing library size. Of the ones that were called, it achieved a 68% (136/201) accuracy at the 4-digit level. In contrast, POLYSOLVER consistently had >96% accuracy (96.5% - 100%) across the range of reads (Figure 2.7).
2.3.4 Comparison with HLA mutations identified using standard methods

The Cancer Genome Consortium (TCGA) has undertaken the comprehensive characterization of several different cancer types across multiple data modalities. In order to test the validity of this approach, we assembled a data set of 2688 cases of matched tumor and germline DNA that had previously been analyzed using standard methods. 2437 of these samples were curated from the TCGA and represented 10 diverse solid tumors (TCGA (2012a), Ding et al. (2008), Stransky et al. (2011), TCGA (2012b), TCGA (2008), TCGA (2011), TCGA (2012c)), while 251 cases were separately collected from previously reported datasets of CLL cases (Wang et al. (2011)) and melanoma (Hodis et al. (2012)). Mutation lists for lung squamous, lung adenocarcinoma, bladder, uterine, head and neck, colon and rectum, glioblastoma, ovarian and breast were taken from the TCGA project (available through the Sage Bionetworks’ Synapse resource (http://www.synapse.org/#!SYNAPSE:syn1729383). The list of mutations for melanoma and chronic lymphocytic leukemia were obtained from the corresponding project resources respectively (Wang et al, 2011; Hodis et al, 2012). There were 56 HLA gene (HLA-A, B and C)
mutations that were identified in these 2688 samples by standard approaches. These samples were then analyzed by first typing with POLYSOLVER and then using Mutect and Strelka as described previously. This approach could identify only 37/56 mutations that were identified by standard methods, and detected 23 additional mutations (Figure 2.8, Figure 2.9). Visualization by IGV suggested that most of the mutations that were identified exclusively in SYNAPSE may be false positives. On the other hand, mutations that were identified exclusively by the POLYSOLVER+mutation detection approach seem to have good sequence based support.

![Figure 2.8](image_url) Comparison of somatic mutations identified by TCGA across cancers using standard approaches to those identified by the POLYSOLVER framework (n=2688).

We reasoned that presence of detected mutations in individual-matched transcriptomic data (RNA-Seq) would provide independent confirmation of these events. We focused exclusively on mutations that had RNA-Seq data available and also were adequately powered for mutation detection (Figure 2.10). Most of the current RNA-Seq alignment methods including MapSplice have poor recall for detecting insertions and deletions (Engstrom et al, 2013).

All of the 11 missense mutations that were identified by POLYSOLVER and standard approaches were validated at the RNA-Seq level. The allelic fraction of the mutations in the expression data varied from 8% to as high as 98%. In case of 7 missense mutations that were unique to POLYSOLVER, the validation rate was 71% (5/7) with allele fraction in the 6-38% range. None of the 5 mutations that were uniquely identified by TCGA were validated at the
expression level. The rate of RNA-Seq detection of nonsense and non-stop mutations was 75% (9/12 and 3/4) for events that were identified by both methods and those that were detected by exclusively by POLYSOLVER, while neither of the two events identified by TCGA alone had any mutation-bearing reads in RNA-Seq data. A similar pattern was observed for splice site mutations with similar validation rates for events in the common and POLYSOLVER exclusive categories (2/3 and 3/4 respectively), whereas the only such event that was unique to the TCGA resource could not be confirmed. These results suggest that our framework is both more sensitive and specific in detection mutations in these highly polymorphic loci than conventional methods.

### 2.3.5 Mutational spectrum of HLA class I genes and their functional effects

We then applied the POLYSOLVER driven mutation detection framework to an extended set of 3768 samples by including an additional 1080 TCGA samples to our original collection of 2688 tumor/normal pairs. In total, 131 somatic HLA mutations were detected in 114 of 3768 samples.
Figure 2.10 Validation of mutations at the transcriptomic level. Mutations identified by POLYSOLVER using DNA sequencing were validated at the transcriptome level using RNA sequencing of the same samples.

(3%) of individuals. Events in HLA-A and HLA-B comprised 43.5% and 39.7% of the total number of mutations respectively, whereas HLA-C only had 16.8% of the events (Figure 2.11).

The identified mutations were not restricted to a gene loci, mutation class or spectrum (Figure 2.11, Figure 2.12). The highest frequency of mutations occurred in exon 4 (35.1%) that encodes the alpha-3 domain of the HLA protein that binds to the CD8 co-receptor of T cells. Abrogation of this function could lead to a loss of T cell recognition and thereby lead to a loss of immune response. A mutation in alpha 3 domain in residues 223–229 can prevent low avidity T cells from binding to MHC/peptide complex (Pittel et al, 2003). About 13% of exon 4 mutation (6/46) were found in this region in our data set were found in this region.

About 37.4% of mutations were found in exons 2 and 3 that code for the alpha1 (13%) and alpha2 (24.4%) peptide binding domains. These domains bind the peptide that is usually 9 amino acids long and the MHC:peptide complex is then transported to the cell surface for recognition by cytotoxic T cells. Several residues on the HLA molecule are known to interact with specific positions of the binding peptide (Brusic et al. (2002)). There are two major anchor grooves in the HLA molecule that bind to positions 2 and 9 respectively of the peptide and mutation in either one could lead to a profound effect on the biochemical stability of the MHC:peptide complex. A secondary anchor groove that interacts primarily
with the sixth amino acid of the peptide lies between the two primary anchor grooves. 25% of mutations (12/49) in the peptide binding domains were in residues that come in contact with the peptide and 9 of these were in positions that comprised one of the two primary anchor grooves. Interestingly, most of these mutations (7/9) were in the second primary groove (Figure 2.13). In addition, 5 of the 12 HLA contact residue mutations can potentially interact with the secondary anchor groove.

A significant number of mutations also occur in exon 1 (11.5%) which is translated to create the leader sequence that helps in correct localization of the MHC molecule. An abundance of truncating and frameshifting events were detected in this exon which could lead to a complete deletion of the protein.

9.9% of mutations were in exon 5 which encodes for the transmembrane region. Mutations in this domain might result in an inability of the MHC molecule to anchor itself on the cell membrane or incorrect localization.
The least number of mutations were recorded in the cytoplasmic tail encoded by exons 6 and 7 (4.6%). Surprisingly, 2 non-stop mutations were also recorded, which may cause decrease of transcript levels by mRNA degradation due caused by a missing polyA tail.

We compared the ratio of the number of somatic changes to the number of silent mutations in HLA and non-HLA genes (Figure 2.12). This ratio is around 2.5 for both categories, which suggests that most missense mutations in HLA genes accrue as a result of systemic dysregulation of DNA repair and other error correcting mechanisms that accompany malignancy and are therefore non-causative passenger mutations. However this ratio is significantly enriched for nonsense, splice site and frameshift indels in HLA genes compared to the non-HLA genes. While nonsense mutations profoundly affect protein function due to truncation, frameshift indels can also potentially have a significant effect the expression levels of the alleles by activating mRNA degradation pathways. Splice site mutations may similarly affect mRNA or protein structure as a result of aberrant splicing or intronic retention. These results suggest that HLA mutations are significantly enriched for loss-of-function events.
Figure 2.13  HLA mutations in contact residues. Contact residues in the HLA molecules are in actual physical contact with the peptide.

2.3.6 Patterns of HLA mutations across tumor types

We observed a wide disparity in the frequency and nature of HLA mutations across the 12 tumor types in our data set (Figure 2.14). Colon, head and neck, melanoma and lung squamous tumors had the highest frequency of HLA mutations (mutations/sample) while breast, glioblastoma and ovarian had the lowest. Although melanoma had a high rate of HLA mutations, there were in fact fewer observed mutations that expected based on the average mutation rate in these samples derived from non-HLA genes. Colon cancer was the only tumor types where there was clear evidence of mutations in more than one HLA allele in the same sample which might indicate increased significance of HLA mutations in this disease. The contrast between the HLA mutation spectrum between colon and rectal is also striking given how similar they are otherwise. While there seem to be certain preferred mutational hotspots, somatic changes, particularly loss-of-function events like nonsense and frameshifting events were found in all functional domains.

We also observed several positions with recurrent events, which might be indicative of their biological relevance as important driver events in the development or progression of the disease. The ”hotspot” mutation positions comprised events across HLA genes, tumor and mutation types, and we did not detect position-dependent tumor or gene specific effects with a few notable exceptions. Residue 7, along with residue 209, was the most frequently mutated position with
a preponderance of truncating mutations that were all exclusively in head and neck samples. Residue 209 is present at the 5’ end of the TCR binding domain and is characterized by insertions in primarily colon cancer patients. There were three recurrent splice site mutation positions spanning the junctions between exons 1 and 2, 3 and 4 and 4 and 5. Also notably bladder was the only tumor type with two recurrent non-stop mutations.

2.4 Discussion

The application of the described framework for detection of mutations in HLA-A, B and C genes across thousands of characterized samples yielded many more mutations that were missed by standard mutation detection techniques. A preliminary visualization suggests that disclosed example method is both more sensitive and specific compared to existing techniques. The improved performance can be attributed to (i) using the correct reference allele sequences, (ii) preselection of reads corresponding to HLA genes, (iii) using a precise alignment algorithm, and (iv) judicious post-filtering of the alignments prior to mutation calling. Since immune evasion is an important mechanism for the development and progression of cancer, the acquisition of somatic changes in HLA regions may be an important driver or abetting events in at least some of the patients. This hypothesis is supported by the fact that recurrent HLA mutations, as well as mutations in exons 2 and 3 responsible for coding the peptide binding domains that are primarily responsible for presentation of the aberrant peptides to the immune system are regularly found.

A framework for inference of constituent alleles and more sensitive and specific detection of somatic changes in polymorphic loci such as HLA by first inferring the correct reference sequences for these genes using relatively low coverage exome or other types of sequencing data has been described in this work. The framework can harness multiple different somatic change algorithms and leverage their individual strengths. This example methodology will assist in gleaning knowledge from the vast repositories of sequence data that have been generated thus far. Newer sequencing technologies are enabling increased coverage and longer read lengths, which would help further in resolving polymorphic loci into their correct constituent alleles and lead to increasingly better performance of the framework.
Figure 2.14  Pan-cancer distribution of HLA mutations across functional domains/exons. Distribution of HLA mutations across functional domains and tumor types. Potential loss-of-function events were found to be distributed across the length of the protein, denoted by the blue (out of frame indels) and red (nonsense mutations) dots across the top. The histogram immediately below summarizes the number of events identified at each position. The vertical panel on the left shows the expected versus observed number of mutations for each disease type. The central panel shows patterns of mutations detected in each tumor type. There were several positions with recurrent events; the lowermost panel depicts recurrent positions with frequency $\geq 3$. The disease and allele group of the detected mutations are also indicated in this panel.
CHAPTER 3. SYSTEMATIC IDENTIFICATION OF PERSONAL TUMOR-SPECIFIC NEOANTIGENS IN CHRONIC LYMPHOCYTIC LEUKEMIA

Modified from a paper submitted to Blood 2014

Mohini Rajasagi, Sachet A. Shukla, Edward F. Fritsch, Derin B. Keskin, David DeLuca, Ellese Carmona, Wandi Zhang, Carrie Sougnez, Kristian Cibulskis, John Sidney, Kristen Stevenson, Jerome Ritz, Donna Neuberg, Vladimir Brusic, Stacey Gabriel, Eric S. Lander, Gad Getz, Nir Hacohen, and Catherine J. Wu

3.1 Introduction

Recent progress in the development of potent vaccine adjuvants, clinically effective vaccine delivery systems and agents that overcome tumor-induced immunosuppression has been striking, and strengthens the possibility that long-awaited effective therapeutic cancer vaccines are feasible (Hodi et al. (2010), Topalian et al. (2012), Taneja (2012), Mellman et al. (2011)). However, the problem of antigen selection for cancer vaccines has remained an unsolved challenge. A review of past cancer vaccine efforts has revealed a general lack of efficacy that may stem from their focus on overexpressed or selectively expressed tumor-associated native antigens as vaccine targets (Boon et al. (2006), Rosenberg et al. (2004)). Mounting an effective response against these native immunogens requires overcoming the challenging hurdles of breaking cen-
tral and peripheral tolerance, while risking the generation of autoimmunity (Mellman et al. (2011)). The rare examples of successful cancer vaccines in humans have targeted foreign pathogen-associated antigens (van der Burg and Melief (2011)), idiotype vaccines derived from patient-specific rearranged immunoglobulins (Baskar et al. (2004)), or a vaccine based on a mutated growth factor receptor (Nedergaard et al. (2012)). These contrasting studies point to the importance of selecting immunogens that are distinct from self, where central and peripheral tolerance are not an issue and the risk of autoimmunity is lower.

A primary hallmark of tumorigenesis is the accumulation of mutations in cancer cells. These mutations are found both in genes that drive cancer, and those that do not (passenger mutations) (Garraway and Lander (2013)). Regardless of driver status, these mutations provide a potential opportunity to specifically target tumor cells through the creation of tumor specific novel immunogenic peptides (neoantigens). These neoantigens are generated from peptides encoded by gene alterations that are present in tumor but not normal tissue, and therefore represent highly promising vaccine immunogens (Hacohen et al. (2013), Heemskerk et al. (2013)). Seminal studies have suggested the immunotherapeutic potential of neoantigens and have shown that: (a) mice and humans can mount T cell responses against mutated antigens (Sensi and Anichini (2006), Parmiani et al. (2007)); (b) mice can be tumor-protected by immunization with a single mutated peptide present in the tumor (Mandelboim et al. (1995)); and (c) memory cytotoxic T lymphocyte (CTL) responses to mutated antigens are generated in patients who have unexpected long-term survival or have undergone effective immunotherapy (Lennerz et al. (2005), Zhou et al. (2005)). However, neoantigens also are almost exclusively personal, found uniquely in the tumor of each individual patient, and therefore have not been used for immunotherapy due to technical difficulties in their identification and testing (Sensi and Anichini (2006)).

Two recent technologies have now overcome this limitation. First, massively-parallel sequencing has now enabled not only the analysis of putative driver genes and pathways across various cancers (Garraway and Lander (2013), Vogelstein et al. (2013)), but also the comprehensive identification of the tens to thousands of somatic protein-coding mutations which may create epitopes that can be recognized immunologically in an individual- and tumor-specific
fashion. Second, refinements in class I HLA prediction algorithms have enabled the reliable prediction of peptide binding for a broad range of class I HLA alleles (Lin et al. (2008), Zhang et al. (2011)).

Herein, we report that putative neoantigens identified though sequential application of massively parallel sequencing followed by HLA binding prediction are immunogenic in humans and can target malignant cells in a tumor-specific fashion. Initially, we focused on chronic lymphocytic leukemia (CLL), a common adult B cell malignancy that remains largely incurable but is potentially immune-responsive on the basis of reports of spontaneous regression and susceptibility to the graft-versus-leukemia effect (Dores et al. (2007), Zenz et al. (2010), Del Giudice et al. (2009)). Based on leukemia-specific gene mutations that we identified in patients with CLL (Landau et al. (2013), Wang et al. (2011)), we predicted candidate leukemia neoantigens and then monitored neoantigen-specific memory T cell responses in patients who had undergone allo-HSCT (Burkhardt et al. (2013)). Our approach demonstrates the feasibility of identifying immunogenic neoantigens and provides a basis for designing a truly personalized immunotherapeutic vaccine in humans.

3.2 Methods

3.2.1 A pipeline for the systematic identification of tumor neoantigens

We leveraged recent advances in sequencing technologies and peptide epitope prediction to generate a two-step pipeline to systematically discover candidate tumor-specific HLA-bound neoantigens. As depicted in Figure 3.1, we began with DNA sequencing of tumors (by either whole-exome (WES) or whole-genome sequencing (WGS)) in parallel with matched normal DNA, followed by comprehensive identification of tumor specific non-synonymous somatic mutations (Lawrence et al. (2013), Cibulskis et al. (2013)). Next, we used the well-validated prediction algorithm, NetMHCpan, to predict candidate tumor specific mutated peptides with the potential to bind personal class I HLA proteins and hence be presented to CD8+ T cells (Zhang et al. (2011)). Finally, to identify a smaller number of epitopes for deeper investigation, we selected a subset of candidate peptide antigens based on experimental validation of their
binding to HLA and expression of cognate mRNAs in autologous leukemia cells.

Figure 3.1  Schematic representation of a strategy to systematically discover tumor neoantigens. Tumor specific mutations in cancer samples are detected using whole-exome (WES) or whole-genome sequencing (WGS) and identified through the application of mutation calling algorithms (such as Mutect) (Cibulskis et al. (2013)). Next, candidate neoepitopes can be predicted using well-validated algorithms (NetMHC-pan) and their identification can be refined by experimental validation for peptide-HLA binding and by confirmation of gene expression at the RNA level. These candidate neoantigens can be subsequently tested for their ability to stimulate tumor-specific T cell responses.

3.2.1.1 Whole exome capture sequencing data for CLL and other cancers

The list of somatic mutations detected in the 91 CLL patients used in this study has been previously reported (Wang et al. (2011)). The list for melanoma was obtained from dbGaP database (phs000452.v1.p1) and for the 11 other cancers, through TCGA (available through the Sage Bionetworks’ Synapse resource (https://www.synapse.org/#!Synapse:syn1729383)). We genotyped the HLA-A, HLA-B and
3.2.1.2 **HLA typing from exome capture data**

We developed the POLYSOLVER (POLYmorphic loci reSOLVER) algorithm that would enable HLA typing of HLA -A, -B and C alleles of individuals using capture data (please see Chapter 2 for details). We constructed a dedicated sequence library consisting of all known HLA alleles (6597 unique entries), based on the IMGT database. From this resource, a secondary library of 38-mers was generated, and putative reads emanating from the HLA locus were extracted from total sequence reads based on perfect matches against it. The extracted reads were then aligned to the IMGT-based HLA sequence library using the Novoalign software (http://novocraft.com), and HLA alleles were inferred through a two-stage calculation. In the first stage, population-based frequencies were used as priors for each allele and the posterior probability derived scores were calculated based on quality and insert size distributions of aligned reads. Alleles with the highest scores for each of HLA-A, B and C genes were identified as the first set of alleles. A heuristic weighting strategy of the computed scores in conjunction with the first set of winners were then used to identify the second set of alleles.

3.2.1.3 **Pipeline for prediction of peptides derived from gene mutations with binding to personal HLA alleles**

MHC-binding affinity was predicted across all possible 9-mer and 10-mer peptides generated from each somatic mutation and the corresponding wildtype peptide using NetMHCpan (Version 2.4). These tiled peptides were analyzed for their binding affinities (IC50 nM) to each class I allele in the patients HLA profile. An IC50 value of less than 150 nM was considered a predicted strong binder, an IC50 of 150–500 nM was considered a predicted intermediate to weak binder, while an IC50 > 500 nM was considered a non-binder. Experimental confirmation of predicted peptides binding to HLA molecules (IC50 < 500 nM) was performed using a competitive MHC class I allele-binding assay and has been described in detail elsewhere (Cai et al. (2012), Sidney et al. (2001)).
3.2.2  Analysis of CLL gene expression

CTL responses against an epitope would only be useful if the gene encoding the epitope is expressed in the target cells. Of the 31 patient samples sequenced and typed for HLA, 26 were subjected to genome-wide expression profiling (Brown et al. (2012)). We reanalyzed previously reported microarray data (NCI Gene Expression Omnibus accession GSE37168). Affymetrix CEL files were processed using the affy package in R. The Robust Multichip Analysis (RMA) algorithm models the observed intensities as a mixture of exponentially distributed signal and normally distributed noise and was used for background correction. This was followed by quantile normalization across arrays to facilitate comparison of gene expression under different conditions. The individual probe-level was finally summarized using the median polish approach to get robust probeset-level values (Bolstad et al. (2003), Irizarry et al. (2003)). Gene-level values were obtained by selecting the probe with the maximal average expression for each gene. Batch effects in the data were removed by using the Combat program (Johnson et al. (2007)).

3.3  Results

3.3.1  Frequency of missense and frameshift mutations in 91 CLL samples

We applied this pipeline to a large dataset of sequenced CLL samples (Wang et al. (2011)). From 91 cases that were sequenced by either WES or WGS, a total of 1838 non-synonymous mutations were discovered in protein-coding regions, corresponding to a mean somatic mutation rate of 0.72 (0.36 s.d.) per megabase (range, 0.08 to 2.70), and a mean of 20 non-synonymous mutations per patient (range, 2 to 76). We identified 3 general classes of mutations that would be expected to generate amino acid changes and hence potentially be recognized immunologically as non-self. The most abundant class consisted of missense mutations that cause single amino acid changes, representing 90% of somatic mutations per CLL. Of the 91 samples, 99% harbored missense mutations and 69% had between 10-25 missense mutations (Figure 3.2A). The other two classes of mutations, frameshifts and splice-site mutations (mutations at exon-intron junctions) have the potential to generate longer stretches of novel amino acid sequences entirely specific to the tumor (neo-open reading frames, or neoORFs), with a higher number
of neoantigen peptides per alteration (compared to missense mutations). However, consistent with data from other cancer types (Lawrence et al. (2013)), neoORF-generating mutations were \( \sim 10 \) fold less abundant than missense mutations in CLL (Figure 3.2B-C). Given the prevalence of missense mutations, we focused our initial studies on the analysis of neoantigens generated from this class of mutations.

### 3.3.2 Somatic missense mutations were predicted to generate neopeptides that bind personal HLA class I alleles

T cell recognition of peptide epitopes by the T cell receptor (TCR) requires the display of peptides bound within the binding groove of HLA molecules on the surface of antigen-presenting cells. Recent comparative studies across the >30 available class I prediction algorithms have shown NetMHCpan to consistently perform with high sensitivity and specificity across HLA alleles (Zhang et al. (2011)). To assess the predictive capacity of NetMHCpan, we tested whether the algorithm would correctly predict binding for 33 mutated epitopes that were already identified in the literature on the basis of their functional activity (i.e. ability to stimulate antitumor cytolytic T cell responses) or were characterized as immunogenic minor histocompatibility antigens (Table 3.2,3.3). Among all tiled 9-mer and 10-mer possibilities, NetMHCpan identified all 33 functionally validated mutated epitopes as the best binding peptide among the possible choices for the given mutation. We observed that the median predicted binding affinity (IC50) to the reported HLA restricting elements for each of the 33 mutated epitopes was 32 nM (range, 3-11,192 nM). By setting the predicted IC50 cut-offs to 150 and 500 nM, we could capture 82 and 91% of the functionally validated peptides, respectively (Figure 3.2A) (Sette et al. (1994)).

On the basis of its high degree of sensitivity and specificity, we then applied NetMHCpan prospectively to the 31 of 91 CLL cases for which HLA typing information was available (Wang et al. (2011)). We considered peptides with IC50 < 150 nM as strong binders, IC50 150-500 nM as intermediate to weak binders, and IC50 > 500 nM as non-binders, respectively (31). For 91 CLL cases, we found a median of 10 strong binding peptides (range, 2-40) and 12 intermediate to weak binding peptides (range, 2-41) per case. Hence, in total, a median of 22 (range, 6-81) peptides per case was predicted with IC50 < 500 nM (Fig. 3.2B, table 3.3).
Figure 3.2  Frequency of classes of point mutations that have the potential to generate neoantigens in CLL. Analysis of WES and WGS data generated from 91 CLL cases reveals that (A) missense mutations are the most frequent class of the somatic alterations with the potential to generate neo-epitopes, while (B) frameshifts and (C) splice-site mutations constitute rare events.

3.3.3 More than half of predicted HLA-binding neopeptides showed direct binding to HLA proteins in vitro

To experimentally validate the predicted IC50 nM scores, we performed a competitive MHC I allele binding assay (Sidney et al. (2001)) and focused on class I-A and B alleles. To this end, we synthesized 112 mutated 9- or 10-mer peptides, with predicted IC50 scores of less than 500 nM, identified from 4 CLL cases (Pt 1-4). The experimental results correlated with the binding predictions. Experimental binding (defined as IC50 < 500 nM) was confirmed in 76.5% and 36% of peptides predicted with IC50 of < 150 nM or 150-500 nM, respectively (Figure 3.2C). In total, ~55% (61/112) of predicted peptides were experimentally validated as binders to personal HLA alleles.
3.3.4 80% of mRNAs encoding predicted neoantigens are expressed in CLL

We classified the expression level of 347 genes (that harbor mutations in CLL samples) as low/absent (lowest quartile), medium (middle two quartiles) or high (highest quartile) expression. As shown in Figure 3.2D, 80% of the 347 mutated genes (or 79% of the 180 mutations with predicted HLA-binding) were expressed at medium or high expression levels.

3.3.5 Large numbers of candidate neoantigens were predicted across diverse cancers

The overall somatic mutation rate of CLL is similar to other blood malignancies, but low in comparison to solid malignancies (Lawrence et al. (2013)) (Figure 3.5A). To examine how tumor type and mutation rate could impact the abundance and quality of candidate neoantigens, we applied our pipeline to publicly available WES data from 13 malignancies including high (melanoma (MEL), lung squamous (LUSC) and adeno (LUAD) carcinoma, head and neck cancer (HNSC), bladder cancer, colon and rectum adenocarcinoma), medium (glioblastoma (GBM), ovarian, clear cell renal carcinoma (clear cell RCC), breast cancer) and low (CLL, acute myeloid leukemia (AML)) mutation rate cancers. To perform this analysis, we further implemented a recently developed algorithm that enables inference of HLA typing from WES data (see POLYSOLVER, Chapter 2).

The mutation rate in these solid malignancies was an order of magnitude higher than for CLL and was associated with an increased median number of missense mutations. For example, melanoma displayed a median of 300 (range, 34-4276) missense mutations per case, and while RCC had 41 (range, 10-101), respectively. Frameshift and splice-site mutations in melanoma and RCC were increased by only 2-3 fold in frequency as compared to CLL and summed neoORF length per sample were increased only moderately (by 5-13 fold). Overall, the median number of predicted neopeptides with IC50 < 500 nM generated from missense and frameshift events per sample was proportional to the mutation rate; this was approximately 20- and 4-fold higher for melanoma (488; range, 18-5811) and RCC (80; range, 6-407)), respectively, compared to CLL (24; range 2-124). With a more stringent threshold of IC50 < 150 nM, the corresponding
Table 3.1  Distribution of mutation classes, summed neoORF sizes and number of predicted binding peptides across 13 cancers

<table>
<thead>
<tr>
<th>Cancer type</th>
<th># of mutations/sample median (range)</th>
<th>Summed NeoORF length/sSample</th>
<th># of predicted peptides median (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Missense</td>
<td>Frame shift</td>
<td>Splice site</td>
</tr>
<tr>
<td>MEL</td>
<td>300 (34-4270)</td>
<td>2 (0-16)</td>
<td>4 (0-101)</td>
</tr>
<tr>
<td>LUSC</td>
<td>212 (0-2397)</td>
<td>3 (0-28)</td>
<td>5 (0-37)</td>
</tr>
<tr>
<td>LUAD</td>
<td>172.5 (6-8971)</td>
<td>7 (0-61)</td>
<td>5 (0-127)</td>
</tr>
<tr>
<td>BLCA</td>
<td>161.5 (28-1194)</td>
<td>8 (0-22)</td>
<td>4 (0-22)</td>
</tr>
<tr>
<td>HNSC</td>
<td>95 (2-1400)</td>
<td>5 (0-106)</td>
<td>2 (0-29)</td>
</tr>
<tr>
<td>COAD</td>
<td>93 (32-5902)</td>
<td>4 (1-182)</td>
<td>0 (0-96)</td>
</tr>
<tr>
<td>READ</td>
<td>72.5 (37-1837)</td>
<td>2 (0-31)</td>
<td>0 (0-2)</td>
</tr>
<tr>
<td>GBM</td>
<td>47 (0-169)</td>
<td>2 (0-16)</td>
<td>1 (0-5)</td>
</tr>
<tr>
<td>OV</td>
<td>42 (9-149)</td>
<td>1 (0-7)</td>
<td>1 (0-6)</td>
</tr>
<tr>
<td>RCC</td>
<td>41 (10-101)</td>
<td>5 (0-22)</td>
<td>1 (0-8)</td>
</tr>
<tr>
<td>BRCA</td>
<td>25 (1-300)</td>
<td>2 (0-54)</td>
<td>1 (0-8)</td>
</tr>
<tr>
<td>CLL</td>
<td>16 (0-75)</td>
<td>1 (0-9)</td>
<td>1 (0-6)</td>
</tr>
<tr>
<td>AML</td>
<td>7 (0-20)</td>
<td>1 (0-2)</td>
<td>0 (0-3)</td>
</tr>
</tbody>
</table>

Numbers of predicted neopeptides per sample were 212, 35 and 10 for melanoma, RCC and CLL, respectively (Figure 3.5, Table 3.1).

3.4 Discussion

Each individual tumor harbors a broad spectrum of shared and personal genetic alterations that continue to evolve in response to the environment and often lead to resistance to therapy (Garraway and Lander (2013)). Given the uniqueness and plasticity of tumors, an optimal therapy may need to be customized based on the exact mutations present in each tumor and may need to target multiple nodes to avoid resistance (Ene and Fine (2011), Longo (2012), Greaves and Maley (2012), Nazarian et al. (2010)). The vast repertoire of human CTLs has the potential to create such a therapy that targets multiple, personalized tumor antigens.
This opportunity motivated us to develop a pipeline that systematically identifies CTL target antigens harboring tumor-specific mutations. Leveraging massively parallel sequencing and algorithms that effectively predict HLA-binding peptides, we predicted tumor neoantigens in a variety of low and high mutation rate cancers, and experimentally identified long-lived CTLs that target leukemia neoantigens in CLL patients. Our results, together with recent studies in humans (Robbins et al. (2006)) and mice (Castle et al. (2012)), support the existence of protective immunity targeting tumor neoantigens, and our pipeline provides a method for selecting neoantigens for future personalized tumor vaccines.

We applied our pipeline to a unique group of CLL patients who developed clinically evident durable remission associated with anti-tumor immune responses following allogeneic-HSCT. These graft-versus-leukemia responses have typically been attributed to allo-reactive immune responses targeting hematopoietic cells (Bleakley and Riddell (2004)). However, we now find that the GvL response is also associated with CTLs that recognize personal leukemia neoantigens. These results are consistent with prior studies by our group and others demonstrating the existence of GvL-associated CTLs with specificity for tumor- rather than allo-antigens (Burkhardt et al. (2013), Nishida et al. (2009), Zhang et al. (2010)). Further supporting the importance of neoantigen-reactive CTLs in cancer surveillance, an elegant study of a long-term melanoma survivor found that CTLs targeting neoantigens are significantly more abundant and sustained than those against non-mutated overexpressed tumor antigens (Lennerz et al. (2005)). In agreement with the melanoma study, we found that neoantigen-specific T cell responses in CLL patients were long-lived (on the order of several years) memory T cells (based on their rapid stimulation kinetics in vitro) and associated with continuous disease remission (Lu et al. (2013)). We thus hypothesize that neoantigen-reactive CTLs play an active role in controlling leukemia in transplanted CLL patients.

More generally, we estimated the abundance of neoantigens across many tumors and found \( \sim 1.5 \) HLA-binding peptides with IC50\(<500\mathrm{nM} \) per point mutation and \( \sim 4 \) binding peptides per frameshift mutation. As expected, the rate of predicted HLA binding peptides mirrored the somatic mutation rate per tumor type (Figure 3.5). We used two approaches to study the relationship between predicted binding affinity and immunogenic neoantigens that induce CTLs.
A retrospective application of our pipeline to published immunogenic tumor neoantigens (i.e. in which reactive CTLs were observed in patients) demonstrated that the vast majority (91%) of functional neoantigens are predicted to bind HLA with IC50<500nM (with ~70% of wild type counterpart epitopes predicted to bind at a similar affinity) (Table 3.2, Table 3.3). This test using a gold standard set of neoantigens confirms that our pipeline largely classifies true positives correctly. A prospective prediction of neoeptopes followed by functional validation showed that 6% (3/48) of our predicted epitopes were associated with neoantigen-specific T cell responses in patients – comparable to the rate of 4.8% found recently for melanoma (Robbins et al. (2006)). The low proportion does not necessarily imply low prediction accuracy for the algorithm. Rather, number of true neoantigens is greatly underestimated by our experiments because: (i) allo-HSCT is a general cellular therapy likely to induce only a small number of neoantigen-specific T cell memory clones; and (ii) our T cell expansion methods are not sensitive enough to detect nave T cells that represent a much larger part of the repertoire but with much lower precursor frequencies. Beyond the issue of prediction accuracy, we have also not yet measured the frequency of CTLs that target neoORFs a class of neoantigens that we expect to be more specific (for lack of a wild type counterpart) and immunogenic (as a result of bypassing thymic tolerance). Furthermore, we have not considered peptides binding to the less common HLA alleles because peptide binding predictions are not yet accurate for these alleles. Future improvements in predictive algorithms, and eventually development of direct physical methods for detecting HLA-binding peptides using mass spectrometry, will make it possible to more effectively select neoantigens presented by tumor HLA proteins. Considering the limitations described here, we conclude that there will be many more neoantigens per tumor.

Given the large number of candidate neoantigens, are there additional considerations that could help select the most useful antigens for targeting tumors? First, is it important to have high RNA or protein expression for CTLs to detect HLA-peptide complexes? While no empirical studies have tested this question in human trials, it appears that very low levels (i.e. even a single peptide-HLA protein) may be sufficient for a cell to be targeted by a CTL (Sykulev et al. (1996)), suggesting that high expression may not be required for inclusion of a neoantigen in a vaccine. Second, while theory suggests that essential cancer genes (drivers) are less likely to
Table 3.2  HLA-peptide binding affinities of known functionally derived immunogenic mutated epitopes across human cancers using NetMHCpan. Epitopes from missense mutations (NSCLC: non-small cell lung cancer; MEL: melanoma; CLL: chronic lymphocytic leukemia; RCC: clear cell renal carcinoma; BLCA: bladder cancer; NR: not reported;). Yellow: IC50 < 150 nM, green: IC50 150–500 nM and grey: IC50 > 500 nM. We discovered that these epitopes fell into 4 groups. Group 1: Mutated and wildtype peptides with IC50<150 nM; Group 2: mutated and wildtype peptides with IC50 150–500 nM; Group 3: mutated peptide with IC50 < 150 and wildtype peptide with IC50 > 500 nM and Group 4: mutated and wildtype peptides with IC50 > 500nM.

<table>
<thead>
<tr>
<th>Group</th>
<th>Cenno</th>
<th>Disease</th>
<th>Clinical Response</th>
<th>T cell response</th>
<th>HLA Allele</th>
<th>Mutated</th>
<th>Wildtype</th>
<th>WMUT IC50</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ME-1</td>
<td>NSCLC</td>
<td>Yes</td>
<td>Yes</td>
<td>A*02:01</td>
<td>FLDEFMMDV</td>
<td>3</td>
<td>FLDEFMMDV</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RELMM2</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*01:01</td>
<td>LTDDRLTFCY</td>
<td>3</td>
<td>LTDDRLTFCY</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRX5</td>
<td>ME</td>
<td>No</td>
<td>Yes</td>
<td>A*20:01</td>
<td>LLDDISLVS</td>
<td>5</td>
<td>LLDDISLVS</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MATN5</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*11:01</td>
<td>KTLFDPQGK</td>
<td>5</td>
<td>KTLFDPQGK</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DDX9T</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*08:01</td>
<td>EAIQRTRF</td>
<td>10</td>
<td>EAIQRTRF</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RAP4</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>B*07:02</td>
<td>KAPHRAGAY</td>
<td>10</td>
<td>KAPHRAGAY</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAS7</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*23:01</td>
<td>SLADEARVYL</td>
<td>12</td>
<td>SLADEARVYL</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SIRT2</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*23:01</td>
<td>KPSSEVTUK</td>
<td>14</td>
<td>KPSSEVTUK</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EP2</td>
<td>NSCLC</td>
<td>No</td>
<td>Yes</td>
<td>A*18:02</td>
<td>ETVBSEESNV</td>
<td>16</td>
<td>ETVBSEESNV</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GARSH</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*02:01</td>
<td>GVEGLTTTV</td>
<td>11</td>
<td>GVEGLTTTV</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HSP 70</td>
<td>RCC</td>
<td>No</td>
<td>Yes</td>
<td>A*22:01</td>
<td>SLFEGIOYTF</td>
<td>23</td>
<td>SLFEGIOYTF</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ACTN1</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*22:01</td>
<td>MASKNYKLY</td>
<td>29</td>
<td>MASKNYKLY</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CDK12</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*11:01</td>
<td>OLKDLFTK</td>
<td>33</td>
<td>OLKDLFTK</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KIAA1440</td>
<td>RCC</td>
<td>Yes</td>
<td>Yes</td>
<td>A*11:01</td>
<td>GTACEDLY</td>
<td>33</td>
<td>GTACEDLY</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAU3</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*22:01</td>
<td>IKNAMMNK</td>
<td>34</td>
<td>IKNAMMNK</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PPFR3</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*11:01</td>
<td>YTDRCDDYIV</td>
<td>49</td>
<td>YTDRCDDYIV</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MUM-2</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>B*14:02</td>
<td>SELFRSGIDSY</td>
<td>184</td>
<td>SELFRSGIDSY</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KAAP229</td>
<td>BLCA</td>
<td>Yes</td>
<td>Yes</td>
<td>B*44:02</td>
<td>ACPPGTV</td>
<td>200</td>
<td>ACPPGTV</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GRIMM</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*23:01</td>
<td>TLWILLOT FK</td>
<td>282</td>
<td>TLWILLOT FK</td>
<td>179</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CSNKA1</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*02:01</td>
<td>GLFGDITAI</td>
<td>5</td>
<td>GLFGDITAI</td>
<td>1312</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DORF</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*22:01</td>
<td>ILDKLVHL</td>
<td>32</td>
<td>ILDKLVHL</td>
<td>7900</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CTNNB1</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*24:02</td>
<td>DYYDSSHF</td>
<td>41</td>
<td>DYYDSSHF</td>
<td>10748</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SRCP</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*13:01</td>
<td>KIDLAVACQ</td>
<td>48</td>
<td>KIDLAVACQ</td>
<td>14976</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OSS</td>
<td>MEL</td>
<td>No</td>
<td>Yes</td>
<td>B*14:03</td>
<td>KEELDILL</td>
<td>60</td>
<td>KEELDILL</td>
<td>1161</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MYH2</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*02:01</td>
<td>KINNPYK</td>
<td>141</td>
<td>KINNPYK</td>
<td>460</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MARY-1</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>A*02:01</td>
<td>FLEGIEVRKTV</td>
<td>1115</td>
<td>FLEGIEVRKTV</td>
<td>404</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NP2C</td>
<td>MEL</td>
<td>Yes</td>
<td>Yes</td>
<td>B*39:02</td>
<td>AGQITKEV</td>
<td>7914</td>
<td>AGQITKEV</td>
<td>670</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CEB4</td>
<td>MEL</td>
<td>No</td>
<td>Yes</td>
<td>A*02:01</td>
<td>ADDPHGSHFV</td>
<td>1192</td>
<td>ADDPHGSHFV</td>
<td>28222</td>
</tr>
</tbody>
</table>
Table 3.3  HLA-peptide binding affinities of known functionally derived immunogenic mutated epitopes across human cancers using NetMHCpan: Minor alleles. Epitopes from minor histocompatibility antigens (MM: multiple myeloma; HM: hematological malignancy; B-ALL: B cell acute lymphocytic leukemia).

<table>
<thead>
<tr>
<th>Group</th>
<th>Gene</th>
<th>Disease</th>
<th>Clinical Response</th>
<th>T cell response</th>
<th>Mutated Allele</th>
<th>Wildtype Allele</th>
<th>WT/MUT IC50</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ECOF-1</td>
<td>MM</td>
<td>Yes</td>
<td>Yes</td>
<td>B'07.02</td>
<td>RPAAIRPPLAL</td>
<td>3</td>
<td>0.7</td>
</tr>
<tr>
<td>1</td>
<td>KIAA0223</td>
<td>HM</td>
<td>Yes</td>
<td>NR</td>
<td>A'02:01</td>
<td>VLDDLLEEA</td>
<td>17</td>
<td>140</td>
</tr>
<tr>
<td>1</td>
<td>BCL2A1</td>
<td>HM</td>
<td>NR</td>
<td>Yes</td>
<td>A'24:02</td>
<td>DLYQVYGLQI</td>
<td>22</td>
<td>34</td>
</tr>
<tr>
<td>1</td>
<td>BCL2A1</td>
<td>HM</td>
<td>NR</td>
<td>Yes</td>
<td>A'24:02</td>
<td>KFEDDIINW</td>
<td>36</td>
<td>27</td>
</tr>
<tr>
<td>1</td>
<td>HB-1</td>
<td>B-ALL</td>
<td>NR</td>
<td>Yes</td>
<td>B'44:03</td>
<td>EERKGLSHYW</td>
<td>81</td>
<td>67</td>
</tr>
</tbody>
</table>

Develop resistance under immune pressure, it is not clear that the three neoantigens (mutated FND3CB, ALMS1, C6orf89) found in our study to induce long-lived CTLs play any functional role in the tumor. Third, the clonality of the mutations may be important. We found that the three CLL neoantigens were clonal or near-clonal (data not shown (Carter et al. (2012))). We thus suggest that CTLs associated with clinical responses can target tumor-specific mutations that are present in the bulk of the cancer mass (whether recurrent or not). However, it is clear that only large clinical trials can identify neoantigen properties that are important for targeting tumors.

It is becoming clear that targeting multiple personal tumor epitopes is likely a useful therapeutic approach. In a recent study of the B16 murine melanoma model, HLA-binding neoantigens were predicted based on tumor mutations. Mice that were immunized with the corresponding mutated peptides developed CTLs specific to the mutated but not wild type peptide and controlled disease therapeutically and prophylactically (Castle et al. (2012)). In addition, recent studies in mice showing immune editing and escape to dominant CTL responses support the notion of targeting multiple epitopes to avoid resistance (DuPage et al. (2012), Matsushita et al. (2012)). A prospective clinical trial should now be possible to design based on our pipeline for selecting neoantigens with the goal of inducing personalized and specific anti-tumor immune
responses.
Figure 3.3  The numbers and affinity distributions of peptides predicted from 31 CLL cases with available HLA typing. Patients expressing the 8 most common HLA -A, -B alleles in the Caucasian population are marked in grey.
Figure 3.4 Application of the NetMHCpan prediction algorithm to functionally-defined neoepitopes and CLL cases. (A) The amino acid sequences of 33 functionally identified cancer neoepitopes reported in literature (table 3.2, Table 3.3) were tested by NetMHCpan, and the resulting predicted binding (IC50) to their known restricting HLA allele are presented, sorted on the basis of predicted binding affinity. (B) Distribution of the number of predicted peptides with HLA binding affinity < 150 nM (black) and 150–500 nM (grey) across 31 CLL patients with available HLA typing information (Table 3.3). (C) Peptides with predicted binding (IC50 < 500 nM by NetMHCpan) from 4 patients were synthesized and tested for HLA -A and -B allele binding using a competitive MHC I allele-binding assay. The percent of predicted peptides with evidence of experimental binding (IC50 < 500 nM) are indicated. (D) The distribution of gene expression for all somatically mutated genes (n=347) from 26 CLL patients, and for the subset of gene mutations encoding neoepitopes with predicted HLA binding scores of IC50 < 500 nM (n=180). No-low: genes within the lowest quartile expression; medium: genes within the 2 middle quartiles of expression; and high: genes within the highest quartile of expression.
Figure 3.5  Estimation of tumor neoantigen load across cancers. (A) Box plots comparing overall somatic mutation rates detected across cancers by massively parallel sequencing. LUSC: Lung squamous cell carcinoma, LUAD: Lung adenocarcinoma, ESO AD: Esophageal adenocarcinoma, DLBCL: Diffused large B-cell lymphoma, GBM: Glioblastoma, Papillary RCC: Papillary renal cell carcinoma, Clear Cell RCC: Clear cell renal carcinoma, CLL: Chronic lymphocytic leukemia, AML: Acute myeloid leukemia. The distributions (shown by box plot) of (B) the number of missense, frameshift and splice-site mutations per case across 13 cancers, (C) the summed neoORF length generated per sample and of the predicted neopeptides with IC50 $< 150$ nM (D) and with $< 500$ nM (E) generated from missense and frameshift mutations. For all panels, the left and right ends of the boxes represent the 25th and 75th percentile values, respectively, while the segment in the middle is the median. The left and right extremes of the bars extend to the minimum and maximum values.
CHAPTER 4. MICRORNA REGULATORY NETWORK INFERENCE IDENTIFIES MIR-34A AS A NOVEL REGULATOR OF TGF-β SIGNALING IN GLIOBLASTOMA

Modified from a paper published in Cancer Discovery 2012


* denotes equal contribution

4.1 Introduction

Glioblastoma is the most common primary brain tumor in adults. Patients with newly diagnosed glioblastoma have a median survival of 12 months with generally poor responses to chemoradiotherapy (Furnari et al. (2007)). Recent genome-wide profiling studies have shown extensive genetic heterogeneity among glioblastoma samples with distinct molecular subtypes; that these transcriptomic subtypes reflect distinct underlying biology is supported by observed differences in clinical outcome of the patients (Verhaak et al. (2010)), enrichment of different genomic and epigenetic alterations within each subtype, and differential activation of major signaling pathways (Brennan et al. (2009), Noushmehr et al. (2010)).

Noncoding RNAs have emerged as an important class of regulatory molecules in both normal
MicroRNA (or miR) is a class of noncoding small RNAs produced by RNA polymerase II as hairpins of longer precursor RNAs that are subsequently processed to approximately 22-nt-long fragments by RNase III enzymes, Drosha and Dicer. Mature miRs regulate gene expression by promoting mRNA degradation or by inhibiting mRNA translation (Bartel (2004), Krol et al. (2010)). The connection between miRs and cancer was first implicated by their genomic alteration and dysregulated expression in various human tumors (Calin et al. (2004)). Multiple miRs have since been identified to promote or suppress oncogenesis in various tumors, presumably by modulating gene expression in the oncogenic and tumor suppressor networks. In addition, recent studies have proposed new mechanisms of miR-mRNA regulation such as modulation of mRNA with competitive miR-binding sites (sponge interactions) or mRNAs that affect constituents of the miR regulatory machinery (nonsponge interactions; Sumazin et al. (2011), Karreth et al. (2011), Tay et al. (2011)).

Global views of the relationship between miR and mRNA expression have been reported. For instance, Su and colleagues (Su et al. (2011)) used integrative genomics and genetic techniques to characterize the roles of mouse miRs within the mouse liver miR:mRNA network; Dong and colleagues (Dong et al. (2010)) deciphered the pathway connecting mutations under the glioblastoma miR:mRNA expression network; Mestdagh and colleagues (Mestdagh et al. (2011)) established the miR body map online resource to dissect miR function through integrative genomics; Grigoryev and colleagues (Grigoryev et al. (2011)) presented the genome-wide miR regulation of T-lymphocyte activation through the mapped miR, mRNA, and protein networks; and Sharbati and colleagues (Sharbati et al. (2011)) studied macrophage infection via an integrated miR:mRNA network. In these global integrative miR:mRNA network analyses, either general correlation coefficient methods or putative miR target prediction methods have been used to construct or map miR:mRNA connections. These approaches preferentially quantify linear dependencies between pairwise variables.

Recognizing that the functional relationship between 2 variables in cancer is not necessarily linear, we explored a mutual information-based approach in this work that scores miR-mRNA interaction strength on the basis of relevant expression contexts. This mutual information-based approach has been used recently by Sumazin and colleagues (Sumazin et al. (2011)) to
uncover a novel class of modulators of miR:mRNA interactions. Here, we applied the Context Likelihood of Relatedness (CLR; Faith et al. (2007)) network modeling algorithm to generate pairwise measures of associations on the basis of mutual information through calculation of the entropy. At the heart of the CLR algorithm is a unique statistical background correction test which uses the full set of mutual information values to estimate a significance value for each miR:mRNA pair under a given observed network context (see Supplementary Information). The algorithm evaluates the mutual information value of a miR:mRNA pair against the background mutual information distribution of all mRNAs in the data set with the miR, as well as with the distribution of mutual information values of all miRs with the mRNA under consideration. A combined z-score summarizing these 2 comparisons is generated, and the list of all such pairwise z-scores is subsequently used to generate P values by comparing with the normal distribution. A stringent false discovery rate of 5% is finally applied to identify putative miR:mRNA regulatory edges. Interactions whose mutual information values are outliers in the right tails of the pertinent context background distributions of mutual information scores have the greatest likelihood of being identified as significant. This background correction method allows the CLR algorithm to filter out those edges between miRs and mRNA that have spurious similarities with large numbers of other genemiR species.

Network inference approaches in general have a tendency to have high false positive rates which can be partly explained by the noisy nature of the input data. The CLR approach is no exception and is also not well-suited to inferences of directionality and differentiating between direct and indirect interactions. We try to compensate for these limitations by using different complementary approaches and filters for assessing the biological significance and relative importance of edges, miRs, transcription factors etc to arrive at a manageable list of candidates for experimental validation.

The Cancer Genome Atlas (TCGA) has characterized the genomes of glioblastoma on multiple dimensions including coding and noncoding RNAs (TCGA (2008)). We applied the CLR modeling algorithm to this multidimensional data set to infer putative regulatory relationships (edges) between miRs and mRNAs in glioblastoma. Specifically, we were interested in directional miR:mRNA interactions where the miR downregulates mRNA expression either directly
through binding or indirectly through intermediary effectors. Against this global network, we explored the functional relationship between miRs and mRNAs in gliomagenesis.

4.2 Methods

4.2.1 Application of CLR to the TCGA data set

A total of 290 GBM samples that had matched mRNA and miR were used in the analysis. Because the CLR algorithm relies on variability in the dataset to estimate significant edges, we first filtered the 534 miR and 17814 mRNA species profiled on the corresponding platforms based on the variability they showed in the 290 GBM samples. We chose to apply CLR on those miR and mRNA transcripts, which had variance across the 290 samples, higher than the median variance across all species. This yielded 267 miR and 8907 transcripts among which we looked to infer regulatory relationships. Similarly for the subtype specific analysis, we required that the miR and mRNA transcripts, have variance across their respective subtypes, higher than the median variance across all species in that subtype.

4.2.1.1 The CLR Method

For the analysis of the matched miR:mRNA expression profiling data we utilized a previously described method, namely the CLR (Context Likelihood of Relatedness) algorithm (Faith et al. (2007)). The CLR algorithm builds upon the relevance network strategies, by applying a background correction step. After computing the mutual information between regulators and their potential target genes, CLR calculates the statistical likelihood of each mutual information value within its network context. The algorithm compares the mutual information between a miR/gene pair to the background distribution of mutual information scores for all possible miR/gene pairs that include either the miR or its target. After this background correction, the most probable interactions are those whose mutual information scores stand significantly above the background distribution of mutual information scores. This step removes many of the false correlations in the network by eliminating promiscuous cases in which one transcription factor weakly covaries with a large numbers of genes or one gene weakly covaries with many transcript-
tion factors. Such promiscuity arises when the assayed conditions are inadequately or unevenly sampled, or when microarray normalization fails to remove false background correlations due to inter-lab variations in methodology. The mutual information for two discrete random variables $X$ and $Y$ is defined as:

$$MI(X, Y) = \sum_{i,j} P(x_i, y_j) \log \left( \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \right)$$ (4.1)

where $P(x_i)$ is the probability that $X = x_i$. For genes, $X$ and $Y$ represent a transcription factor and its potential target gene, and $x_i$ and $y_i$ represent particular expression levels.

CLR uses the matrix of mutual information (MI) values between all probe sets on the array. The CLR algorithm estimates a likelihood of the mutual information score for a particular pair of genes, $i$ and $j$ by comparing the MI value for that pair of genes to a background distribution of MI values. The background distribution is constructed from two sets of MI values: $MI_i$, the set of all the mutual information values for gene $i$ (in row or column $i$), and $MI_j$, the set of all the mutual information values for gene $j$ (in row or column $j$). Because of the sparseness of biological regulatory networks, most MI scores in each row of the mutual matrix represent random background mutual information. We approximate this background MI as a joint normal distribution with $MI_i$ and $MI_j$ as independent variables and compute the z-score for each of those distributions

$$Z_i = \frac{MI(\text{i}^{th} \text{ entry in i}^{th} \text{ row}) - \text{mean}(\text{i}^{th} \text{ row})}{\sigma(\text{i}^{th} \text{ row})}$$ (4.2)

$$Z_j = \frac{MI(\text{j}^{th} \text{ entry in i}^{th} \text{ column}) - \text{mean}(\text{j}^{th} \text{ column})}{\sigma(\text{j}^{th} \text{ column})}$$ (4.3)

The new metric for the relationship between gene $i$ and gene $j$ is then

$$Z_{ij} = \frac{Z_i + Z_j}{\sqrt{2}}$$ (4.4)

where $Z_i$ and $Z_j$ are the z-scores of the marginal distributions (Whitlock (2005)). The new metric is then used as a measure of the relationships among the genes. P-values corresponding to the entire set of z-scores are then calculated by comparing to the normal distribution; a
stringent FDR of 5% is finally applied to the computed p-values which defines the inferred network. As can be seen from Figure 4.1, CLR identifies several connections in the main body of the network that would have been missed by the relevance network framework.

![Mutual information](image1.png) ![Correlation](image2.png)

Figure 4.1 CLR characteristics. The blue line shows the distribution of measures of association (mutual information and correlation) between all miR:mRNA pairs. The red line shows the distribution of measures of association values for all CLR-identified miR:mRNA pairs.

### 4.2.2 Filters

Three different filters were implemented in order to prioritize candidates for experimental validation.

#### 4.2.2.1 Copy number filter

We hypothesized that edges in the network where one or both nodes were in regions of copy number alteration had evidence of additional mechanism(s) of dysregulation and were therefore had greater biological relevance. Segmented copy number data were downloaded from the TCGA data portal (http://tcga-data.nci.nih.gov/tcga) and were processed using the CN Tools (http://www.bioconductor.org/packages/2.6/bioc/html/CNTools.html) package of the Bio-conductor project to derive a matrix with genes as rows and samples as columns by assign-
ing each gene the corresponding segment value for each sample. The SGOL (Segment Gain or Loss) score was then calculated for each gene by summarizing the positive or negative values for a given gene across samples using the cghMCR package (http://www.bioconductor.org/packages/2.6/bioc/html/cghMCR.html). Genes with large positive and small negative SGOL scores have high amplitude and/or recurrence of alterations across samples. The threshold used for a single sample was >0.3 for gain and <-0.3 for loss. Thresholds of SGOL > 28 and SGOL <-32, corresponding to the 90th percentiles of gain and loss SGOL analyses respectively, were set to filter genes based on copy number data.

### 4.2.2.2 Gene expression affected by copy number filter

We further hypothesized that genes that displayed expression changes that was indeed driven by DNA gain or loss were better experimental targets. Gene weight analysis can be used to evaluate the influence of copy number alterations on gene expression (Carrasco et al. (2006)). In this work, gene weight analysis was applied to find genes or miRs with high correlation between their expression and copy number levels. For a particular probe gene/miR, expression values of all samples were divided into two groups according to copy number status. For example, expression in tumors with copy number alterations (CNA) compared to without CNA. The Gene weight score was calculated by the following formula:

\[
GW = \frac{m_T - m_R}{\sigma_T + \sigma_R}
\]

(4.5)

where \(m_T, m_R\) and \(\sigma_T, \sigma_R\) denote the mean and standard deviations for expression values for two groups T and R respectively. Statistical significance was determined by permuting sample labels for expression data. A low p-value indicates a strong correlation between the expression and copy number level of the gene or miR under consideration.

### 4.2.2.3 Direct targets filter

We were interested in prioritizing miR:mRNA edges where the possibility of a direct molecular interaction between the two molecules was high. Based on the assumption that miRs downregulate their target genes by binding to their 3' UTR regions, we defined a subset of
mRNA nodes as putative direct targets of their miR nodes based on the following parameters: i) a significant negative correlation (Pearson correlation coefficient $\leq -0.3$) between expression of miR and mRNA; and ii) sequence-based prediction of interaction in all three databases, namely Pictar, TargetScan and Miranda (Chen and Rajewsky (2006), Grimson et al. (2007), Krek et al. (2005), Lewis et al. (2005), Betel et al. (2008)).

4.2.3 Identification of subtype-specific edges

To enumerate all miR:mRNA associations that are subtype-specific, we next repeated CLR modeling within each molecular subtype as defined by TCGA (classical, neural, proneural, mesenchymal) on the assumption that subtype-specific relationships would have a high likelihood of being buried in the noise of a global network. We performed a permutation test as follows:

We first apply the CLR algorithm to samples belonging to the 4 different subtypes separately. Significances of miR:mRNA associations that are identified exclusively in one subtype are assessed by three permutation tests as follows. Let us denote the 4 subtypes by A, B, C and D, with A being the subtype in which the uniqueness of a given correlation is being evaluated. 

(i). Compute the difference in miR:mRNA correlations between subtype A and one of the other subtypes, say subtype B.

(ii). Permute the sample labels between subtypes A and B and compute the difference in correlations in the new set. Repeat this for a 1000 permutations to get the background distribution.

(iii). The $p$-value is assigned to the original difference calculated in step (i) by comparing it against the background distribution.

(iv). Repeat steps (i)-(iii) for subtypes C and D, i.e. compare A to C and compare A to D separately.

(v). If the three $p$-values computed in steps (i)-(iv) are all less than 0.01, the correlation is considered unique to subtype A.
4.3 Results

Application of CLR identified a total of 26,297 edges between 254 miR hubs and 6,152 mRNA nodes.

4.3.1 Copy number filter

Integration with copy number profiles revealed that a third (34.1%) of the miR nodes or mRNA nodes linked to CLR-inferred edges resided in regions of copy number aberration, and less than 3.9% of all edges had both miR and mRNA localized in regions of genomic alterations (Figures 4.2 and 4.3). That these edges are subjected to additional levels of deregulation hints at their biological relevance in GBM.

Figure 4.2 CLR network grouped by connectivity among miRs and mRNAs in regions of copy number aberration. Size of the node represents the number of edges between the miR and mRNAs in each group. There are 9 subnetworks represented. Edges in the network where both participating miR and mRNA are in regions of copy number alterations account for less than 3.9% of the total connectivity (green circles). Edges where either the participating miR or the mRNA is in region of gain or loss, but not both, account for 34.1% of the network (orange circles). Approximately 62% of the network involves miRs and mRNAs without evidence of genomic alterations in glioblastoma (blue circle).
4.3.2 Direct targets filter

Application of the direct targets filter identified only 45 (0.17%) of the 26,297 CLR-inferred edges as putatively direct (Figure 4.4, Supp Table 3), suggesting that the predominant relationship between miR and mRNAs may be indirect, possibly through modulation of intermediary molecules such as transcription factors.
4.3.3 Subtype expression differences drive the identification of edges in the network

Four predominant subtypes—classical (CL), neural (NE), proneural (PN) and mesenchymal (MS)—have been identified in GBM based on transcriptomic analysis (Verhaak et al. (2010)) and these have been shown to have biological and clinical significance. As a first step to mine the computed network, we asked whether the CLR-inferred global network captures the salient transcriptomic features of the four molecular subtypes of GBM. Here, we looked for differential expression of the miR nodes and mRNA nodes between any two molecular subtypes. Interestingly, the variability among molecular subtypes appeared to be the predominant drivers of relationships defined by CLR. We find that in 67% of CLR edges, the mRNA changes levels between two subtypes in one direction while the miR that it is associated with changes levels in the opposite direction ($p < 0.001$). Figure 4.5 shows one such identified edge between mir-34a and PDGFRA. It is clear from the figure that PN samples have high levels of PDGFRA and low levels of mir-34a and the reverse trend is apparent for the MS samples. While the difference in expression between the subtype signature genes is not surprising, it is striking that the CLR-identified miRs associated with these genes should show a reciprocal and opposite change of expression along with their mRNA nodes. In particular, the most significant difference was observed between proneural and mesenchymal subtypes where nearly half of the edges in the global network (or 12,673 edges, 48%) reflect the differences between these two subtypes. This observation suggested that miR regulation of mRNA may play an important role in defining the molecular signatures of these two subtypes. We postulate that these strong reciprocal differences in mRNA and miR expression levels between two subtypes (predominantly between PN and MS) might be the underlying effect that is driving the identification of several CLR edges in the global analysis.

4.3.4 Discriminatory miRs between the PN and MS subtypes

To explore further the role of miRs in defining the molecular signatures of subtypes, we looked for CLR-inferred edges among the 685 genes, comprising genes overexpressed in the
four subtypes, which was used by Verhaak et al (Verhaak et al. (2010)). Of these 685 genes, 506 of them (73%) have edges to a miR node in the global CLR network. Conversely, of the 2,984 inferred edges to these 506 subtype-classifier genes, a disproportionate number (70%) of the edges are part of the PN and MS subtype signatures (e.g. 328 to classical, 560 to neural, 858 to PN and 1,238 to MS signature genes), suggesting that specific miR nodes may contribute to gene signatures underlying these two molecular subtypes.

To identify a core network of PN and MS differences, we looked for miR hubs that had significant number of connections to either PN and MS gene signatures, which were defined as genes that are highly expressed in the respective groups. We scored each miR hub with:

\[
\text{score}_{\text{mirna}}(i) = \frac{d_1 - d_2}{(L_{\text{mes}} + L_{\text{pro}})/2 + L_{\text{hub}}} \tag{4.6}
\]

where \(d_1\) is the number of MS signatures genes which are negatively correlated with the miR expression and the number of PN signature genes which are positively correlated with the miR expression; \(d_2\) is the number of MS signatures genes which are positively correlated with the miR and the number of PN signature genes which are negatively correlated with the miR. \(L_{\text{mes}}\) is the size of the MS signature, \(L_{\text{pro}}\) is the size of the PN signature and \(L_{\text{hub}}\) is the number of connections identified by CLR for the miR. This score would be high in two cases; miRs whose edges were to genes that showed negative correlation with MS genes and positive correlation with PN genes or vice versa. The score obtained for each miR was Z-transformed and a false positive rate of 0.001 was applied to identify 8 miR nodes with the greatest discriminatory potential between the PN and MS subtypes. Table 4.1 shows the overlap of subtype signature genes with the CLR targets of mir-34a, which is one of the 8 identified discriminatory miRs. Figure 4.6 shows a heatmap of correlations of the PN and MS signature genes that are associated with these 8 miRs by CLR. These correlations follow a definite pattern 5 miRs have strong negative correlations with PN genes and positive correlations with MS genes; the other 3 miRs show an opposite trend.
4.3.5 miR-34a

We wanted to pick a miR for thorough experimental follow-up. The 8 identified discriminatory miRs were enlisted for functional genomic screens in vitro and in vivo. One such result is highlighted in Figure 4.7 which shows that only mir-34a was able to completely block tumorigenesis in vivo. In parallel to the functional genetic screens, we also explored the biological relevance of these miRs to human GBM through integrative analyses. Based on the thesis that biologically important genetic elements will be subjected to genomic alterations as an alternative mechanism of deregulation, we integrated copy number information on these 8 miRs and their 2,699 mRNA nodes to look for genomic evidence of relevance for either the miRs or their mRNA nodes. As detailed in Table 4.2, three of these 8 miRs (miR-34a, miR-181c, miR-181d) were resident in regions of genomic alteration, 2 with corresponding expression alterations. Similarly, 185 of the mRNA nodes connected to these 8 miRs were found to be targeted for copy number alterations, 46 (∼25%) of which showed corresponding expression alterations. Correspondence of expression and copy number change was evaluated using Gene Weight Analysis. When applying the rules for putative direct-target definition based on expression correlation and sequence-level prediction by three commonly utilized algorithms (Pictar, TargetScan, Miranda), we found that miR-34a is the only miR node that resides in a region of frequent loss and harbors a putative direct CLR-edge, PDGFRA that is in a region of genomic gain. Furthermore, we asked whether expression of any of the miRs tracks with prognosis in

<table>
<thead>
<tr>
<th></th>
<th>All (342)</th>
<th>Negative (92)</th>
<th>Positive (250)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical (162)</td>
<td>18</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>Neural (129)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Proneural (178)</td>
<td>29</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>Mesenchymal (216)</td>
<td>44</td>
<td>0</td>
<td>44</td>
</tr>
</tbody>
</table>

Table 4.1 Overlap of mir-34a targets with subtype signature genes
Table 4.2 Integrating copy number and direct targets filters with the network defined by the 8 miRs

<table>
<thead>
<tr>
<th>miRNA nodes</th>
<th>All mRNA nodes</th>
<th>mRNA nodes with gain</th>
<th>mRNA nodes with gain and correlated expression</th>
<th>mRNA nodes with loss</th>
<th>mRNA nodes with loss and correlated expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) miRNA nodes with gain</td>
<td>399 edges (2 miRNA-304 mRNA)</td>
<td>34 edges (2 miRNA-28 mRNA)</td>
<td>0 edges (2 miRNA-7 miRNA)</td>
<td>7 edges (2 miRNA-6 miRNA)</td>
<td></td>
</tr>
<tr>
<td>(B) miRNA nodes with gain correlated with expression</td>
<td>399 edges (2 miRNA-304 mRNA)</td>
<td>34 edges (2 miRNA-28 mRNA)</td>
<td>0 edges (2 miRNA-7 miRNA)</td>
<td>7 edges (2 miRNA-6 miRNA)</td>
<td></td>
</tr>
<tr>
<td>(C) miRNA nodes with loss</td>
<td>342 edges (1 miRNA-342 mRNA)</td>
<td>21 edges (1 miRNA-21 mRNA)</td>
<td>6 edges (1 miRNA-6 mRNA)</td>
<td>16 edges (1 miRNA-16 mRNA)</td>
<td></td>
</tr>
<tr>
<td>(D) miRNA nodes with loss correlated with expression</td>
<td>0 edges (0 miRNA-0 mRNA)</td>
<td>0 edges (0 miRNA-0 mRNA)</td>
<td>0 edges (0 miRNA-0 mRNA)</td>
<td>0 edges (0 miRNA-0 mRNA)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3 Survival analysis results

<table>
<thead>
<tr>
<th>miRs</th>
<th>log2-exp HazardRatio</th>
<th>log2-exp Wald P</th>
<th>log2-exp FDR</th>
<th>log2-exp C Index</th>
<th>dichotomized HazardRatio</th>
<th>dichotomized Wald P</th>
<th>dichotomized FDR</th>
<th>dichotomized C Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-34a</td>
<td>1.19</td>
<td>0.02036</td>
<td>0.016</td>
<td>0.545</td>
<td>2.2</td>
<td>2.20E-05</td>
<td>0.0047</td>
<td></td>
</tr>
<tr>
<td>miR-181d</td>
<td>0.81</td>
<td>0.03649</td>
<td>0.2</td>
<td>0.461</td>
<td>0.63</td>
<td>0.00214</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>miR-9</td>
<td>0.89</td>
<td>0.02883</td>
<td>0.2</td>
<td>0.463</td>
<td>0.65</td>
<td>0.00265</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>miR-181c</td>
<td>0.81</td>
<td>0.03271</td>
<td>0.2</td>
<td>0.464</td>
<td>0.69</td>
<td>0.00552</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>miR-22</td>
<td>1.17</td>
<td>0.06752</td>
<td>0.27</td>
<td>0.522</td>
<td>1.6</td>
<td>0.0245</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>miR-142-3p</td>
<td>1.12</td>
<td>0.1167</td>
<td>0.35</td>
<td>0.321</td>
<td>1.4</td>
<td>0.028</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>miR-142-5p</td>
<td>1.081</td>
<td>0.4428</td>
<td>0.89</td>
<td>0.514</td>
<td>1.3</td>
<td>0.091</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>miR-223</td>
<td>0.9965</td>
<td>0.9552</td>
<td>0.96</td>
<td>0.499</td>
<td>1.3</td>
<td>0.277</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

all GBM or within a specific GBM subtype. As summarized in Table 4.3, when dichotomized, miR-34a expression was the only miR with significant prognostic value (Bonferroni-adjusted p value = 0.0047) with a Hazard Ratio of 2.2. It has already been established in literature that PN samples have a better prognosis than the other subtypes. Importantly, miR-34a expression does not simply identify PN subtype tumors which are known to have overall better survival (Verhaak et al. (2010)); instead, high miR-34a expression defines a subset of PN GBMs with poor prognosis similar to that of the other subtypes (Figure 4.8). In other words, miR-34a is a prognostic biomarker for aggressive PN GBM. Collectively, the functional genetic data and integrative analyses singled out miR-34a as a strong candidate for further validation.
4.3.6 mir-34a and the transcription factor hypothesis

Consistent with the observation in the global CLR network, less than 1% of the miR-34a subnetwork (three of the 342 nodes or 0.88%) were computationally predicted to be direct, suggesting that the majority of the miR-34a putative regulatory relationships was not modulated by direct miR-mediated RNA degradation/translation impairment. We thus postulate that indirect mechanisms such as transcriptional regulation mediated through intermediate regulators, e.g. transcription factors, may be at play. To identify such intermediates in the network, we performed an in silico enrichment analysis for transcription factor (TF) binding sites. Transcription factor binding site motifs from the TRANSFAC database (release 10.1, Matys et al. (2006)) were identified in genomic regions comprising 8kb upstream and 2kb downstream of the transcription start site of genes based on coincidental prediction by the CisGenome (conservation cutoff = 50, likelihood ratio = 500, third order background Markov model) and MotifScanner (prior probability of 1 motif copy = 0.9) programs (Aerts et al. (2003), Aerts et al. (2005), Ji et al. (2006)). Two sets of genes corresponding to mir-34a CLR targets that define either the PN or MS signatures were compared for differential over-representation of transcription factor binding sites using this approach. For each TF the proportion of genes with at least one corresponding binding site were compared between the two groups. 34 vertebral transcription factors were thus identified (p.val < 0.05). This list was further screened for transcription factors that were overexpressed and had a statistically significant negative correlation with mir-34a (corr < -0.3, p-val < 0.05) in the PN group. We ended up with 3 potential TF candidates (Table 4.4). PBX was not expressed in our cell systems; Myc and Smad4 were experimentally validated as actual intermediaries through which mir-34a effects its subtype specific action. In parallel, the mir-34a:PDGRA edge was also validated as a direct interaction. Figure 4.9 summarizes our current mode of action of mir-34a in GBM.

4.3.7 Identification of subtype-specific edges

Several subtype-specific edges were identified by the subtype-specific procedure described in the methods section. For example, the miR-195:SPRY4 regulatory relationship was not
observed in the global network but was specific to the proneural subtype (Figure 4.10). Correspondingly, we found that miR-195 was capable of modulating SPRY4 expression in E6/E7T immortalized human astrocytes, with functional impact on proliferation, but failed to do so in two established GBM cells, suggesting that miR regulation of its mRNA targets can be highly context-specific.

### 4.4 Discussion

Glioblastoma is a heterogeneous disease characterized by distinct molecular subtypes underlying different biologic behaviors and response to therapies. Leveraging the multidimensional TCGA data set, reverse-engineering with the CLR algorithm has provided an inferred map of the putative miR mRNA regulatory network in glioblastoma. Integrating this network model with molecular subtype denition and functional genomic screen, as well as in silico sequence-based target prediction and promoter analysis, we prioritized miR-34a for downstream mechanistic studies. These studies uncovered a novel regulatory network emanating from miR-34a, which acts as a tumor suppressor in proneural-like glioblastoma, in part, through direct action on PDGFRA as recently shown by Silber and colleagues (Silber et al. (2012)) in addition to commandeering of the SMAD4 transcriptomic network to regulate ID1 and ID3 levels. That the majority of mRNAs computed to link to a miR node appears to be regulated indirectly through transcription factors, such as SMAD4 (this study) and MYC (Christoffersen et al. (2010), Choi et al. (2011)) likely serves as an amplier of an effect of miR on the global transcriptome. This finding thus provides a rationale for an alternative approach to inhibit transcription factor activity through modulating its upstream miR regulatory node. In sum, this work illustrates not only the power of comprehensive cancer genomic data sets such as that of TCGA and

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>Overrepresentation in proneural (p-value)</th>
<th>Overexpression in proneural (p-value)</th>
<th>Correlation with mir-34a in proneural</th>
<th>Correlation with mir-34a in proneural (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYC</td>
<td>0.0257</td>
<td>0.0146</td>
<td>-0.33</td>
<td>0.00442</td>
</tr>
<tr>
<td>PBX1</td>
<td>5.40E-03</td>
<td>6.02E-07</td>
<td>-0.42</td>
<td>0.00029</td>
</tr>
<tr>
<td>SMAD4</td>
<td>0.04133</td>
<td>2.30E-05</td>
<td>-0.41</td>
<td>0.00037</td>
</tr>
</tbody>
</table>
the importance of mining and interpreting such data sets in the context of cancer biology but also the value of computational and experimental models in enabling an understanding of the underlying complexity of the disease.

Although miR-34a has been implicated in multiple cancer types (Hermeking (2010)), its roles in the regulation of SMAD4-ID1-ID3 have not been previously suggested. While the relationship between PDGFRA and miR-34a is predicted by sequence-based algorithms, the relationship between miR-34a and SMAD4 or its downstream ID1 and ID3 are not; therefore, the hypothesis that they could be mediators of miR-34a activity came only through unbiased analysis for transcription factor binding site enrichment in the promoters of CLR-inferred mRNA nodes linked to miR-34a. This reinforces the power of global system biology approaches in generating unanticipated hypotheses.

Finally, our findings in this study have potential clinical application, as miR-34a expression level is shown in 2 independent cohorts of glioblastoma to stratify patients into good and poor prognosis subgroups with significant difference in overall survival. Furthermore, within the TCGA cohort, we found that miR-34a carries significant overlap in prognostic significance with glioma-CpG island methylator phenotype (G-CIMP) status (Noushmehr et al. (2010); data not shown), suggesting a possible mechanistic relationship between miR-34a and G-CIMP, although elucidation of the molecular basis for this relationship will require further studies.

Previous studies have reported miR-34a expression level as a prognostic parameter. For instance, in pancreatic ductal adenocarcinoma, miR-34a loss (i.e., low to no expression) is associated with a decreased survival probability (Jamieson et al. (2012)); in other words, miR-34a-expressing pancreatic ductal adenocarcinoma has a relatively better survival. In breast cancer, although high miR-34a expression is correlated with poor prognosis factors including positive nodal status, high tumor grade, estrogen receptor negativity, HER2 positivity, and high proliferation rate, after adjusting for these known prognostic parameters in multivariate analysis, high miR-34a expression is in fact associated with a lower risk of recurrence or death from breast cancer (Peurala et al. (2011)), indicating that high levels of miR-34a are a good prognostic factor. In contrast to these previous studies, our analyses of 2 independent cohorts of glioblastoma showed that miR-34a low-expressing glioblastomas have better outcome with
longer overall survival. In other words, glioblastoma tumors driven by inactivation of miR-34a are less aggressive than glioblastomas that evolve through deregulation of other genetic elements. The differences in prognostic significance of miR-34a loss in different tumor types likely reflect the modulatory effects of preexisting genetic alterations and the specific susceptibility of different cell types to the aberrant activation of any given pathway. This is not dissimilar to the case of another glioma gene, IDH1, whose specific point mutation affecting a key residue in the protein (R132) has been shown to be oncogenic. Interestingly, gliomas carrying this mutation in IDH1 as well as analogous mutations affecting IDH2 have a significantly better prognosis (Lu et al. (2012), Parsons et al. (2008), Yan et al. (2009), Koivunen et al. (2012)). On the other hand, in cytogenetically normal acute myelogenous leukemia carrying NPM1 mutations, IDH1 mutations at the same residue are a poor prognostic factor as patients with IDH1 mutations do worse (Marcucci et al. (2010), Paschka et al. (2010)). In summary, we illustrate here that computational network modeling of the complex interrelationships among diverse genetic elements can generate a logical framework in which to explore and understand the genetics and biology of cancers, and when integrated with disease knowledge and clinical annotation can lead to discovery of new pathogenetic insights in addition to potential prognostic biomarkers or therapeutic targets. In this regard, we believe that the results from this study should motivate future efforts to explore the therapeutic implication of miR-34a reconstitution. The potent tumor-suppressive activity in our preclinical models would suggest possible therapeutic benefit of miR-34a reconstitution by tumor-targeted delivery in low miR-34a-expressing glioblastoma. In view of its mechanism of action through PDGFRA, MYC, and SMAD4, one may further speculate that reconstitution of miR-34a could represent an attractive strategy to deliver combination therapy against multiple bona fide cancer gene targets.

4.4.1 Shortcomings of the CLR algorithm

The CLR algorithm introduces a very complex correlation structure between the computed Z-scores since the value in each cell is a function of the mutual information distributions along the corresponding row and column. This is evident in the observed p-value distribution as shown in Figure 4.11. It is clear that the assumption of independence is strongly violated. A
lesser concern is that the assumption of normality of the mutual information values is also not true.

We developed a novel approach for miR regulatory network inference that better deals the issues seen in the CLR algorithm. The algorithm models gene expression as a function of methylation and copy number status of the gene. The reduced model of each gene is compared against the full model that includes additional terms capturing miR expression. MiRs that result in a significant improvement in the goodness of fit of the full model are identified as putative regulators of the gene. The resulting p-value distributions are well-behaved and amenable to robust statistical inference. Three sequence prediction algorithms - Miranda, Pictar and Targetscan - are subsequently used to prioritize miR:gene associations for experimental validation. The application of the approach to other cancer types should help delineate disease-specific miRNA:mRNA regulatory networks.

4.4.2 Integrated model

The following equations describe the approach to evaluate the regulatory miR connections of one gene.

Let

\[ i \equiv \text{individual} \]

\[ k \equiv \text{miR} \]

\[ c_i \equiv \text{copy number of gene in individual } i \]

\[ m_{ij} \equiv \text{beta value of methylation probe } j \text{ of the gene in individual } i \]

For all miRs \( r_k \), the following two models are computed:

\[ l_1 \equiv y_i = \beta_0 + \beta_1 c_i + \sum_j \beta_j \log \left( \frac{m_{ij}}{1-m_{ij}} \right) + \epsilon \]

\[ l_2 \equiv y_i = \beta_0 + \beta_1 c_i + \sum_j \beta_j \log \left( \frac{m_{ij}}{1-m_{ij}} \right) + r_{ik} + \epsilon \]  

\[ \epsilon \sim N(0, \sigma^2) \]  

(4.7)

The improvement of fit to the model of expression of the gene resulting from including
expression of miR \( k \) in addition to its copy number and methylation levels can be assessed by the following F-statistic:

\[
F_k = \frac{SSE_{l1} - SSE_{l2k}}{\frac{df_{l1} - df_{l2k}}{SSE_{l2k}}} \sim F(df_{l1} - df_{l2k}, df_{l2k}) \quad (4.8)
\]

The process is repeated for all genes. The distribution of the resulting set of p-values approximate the theoretically expected mixture of uniform and beta (Figure 4.12). miR:mRNA edges significant at 5% FDR define the final network.

### 4.4.3 Results

The integrated model was applied to a TCGA repository of 482 matched samples with expression data across 470 miRs and 12042 genes. 17539 miR:mRNA edges were discovered at the 5% level involving 223 miRs and 3328 genes (Figure 4.13).
Figure 4.4 Putative direct-target subnetworks among CLR-identified edges in glioblastoma. miRs are represented as green diamonds and genes as circles; classical signature genes (gray), neural signature genes (dark green), proneural signature genes (red), mesenchymal signature genes (blue), genes not in any of the molecular subtype signatures (pink). CL, classical; MS, mesenchymal; NL, neural; PN, proneural.
Figure 4.5 mir-34a:PDGFRA edge. Proneural samples have high levels of PDGFRA and low levels of mir-34a and the reverse trend is apparent for the mesenchymal samples.

Figure 4.6 Heatmap of correlations between 8 most discriminative miRs with proneural (PN) and mesenchymal (MS) signature genes in the CLR network. Each cell in the heatmap is the correlation between the corresponding miR and signature gene across samples.
Figure 4.7  

mir-34a blocks tumorigenesis in vivo. Nude mice were transplanted with premalignant proneural-like cells and transduced with 6 pre-miRs corresponding to 8 discriminant miRs or the GFP control (n = 10 per group). miR-34a and miR-142 potently impaired tumorigenesis when compared with GFP control.
Figure 4.8  Prognostic value of mir-34a within the proneural subtype. KaplanMeier survival analysis of patients in the TCGA data set (n = 271). Low expression of miR-34a stratifies a subgroup of patients with proneural glioblastoma with a significant survival advantage compared with others in TCGA cohort (log-rank $P = 1.12 \times 10^{-5}$).
Figure 4.9  mir-34a model of action in GBM
Figure 4.10  Subtype-specific edge example: mir-195:SPRY4
Figure 4.11  Distribution of p-values by CLR. The complex correlation structure introduced by CLR results in a significantly skewed distribution.

Figure 4.12  Distribution of p-values by integrated modeling. The distribution is well-behaved and has a shape similar to the theoretically expected mixture of a uniform and a beta.
Figure 4.13  Top 1% of the inferred regulatory network (176/17539 edges)
CHAPTER 5. SUMMARY

5.1 HLA typing and mutation detection

Recent large scale whole exome sequencing (WES) efforts have identified potentially functional somatic changes in HLA type I genes in several tumor types (Stransky et al. (2011), TCGA (2012a), Ding et al. (2008)). These genes have also been implicated as possible drivers of disease by virtue of recurrence in head and neck, lung squamous and diffuse large B-cell lymphoma (Lawrence et al. (2014)). However the HLA loci are highly polymorphic and it is likely that a considerable number of identified mutations are incorrect due to the reliance of traditional approaches on the canonical reference genome for alignment. Prior determination of the exact complement of alleles in such genomic regions of high variability should be useful in increasing the sensitivity and specificity of standard somatic change detection algorithms at these sites. We developed a method POLYSOLVER (POLYmorphic loci reSOLVER) comprising alignment optimization followed by a novel model-based algorithm for accurate inference of HLA class I alleles from WES data. We additionally presented a framework for sensitive and precise detection of mutations in class I genes which is enabled by prior prediction of correct reference alleles by application of POLYSOLVER on germline data. Application of POLYSOLVER on a set of 133 Hap Map samples with known HLA typing achieved 97% accuracy at the protein level resolution. The use of POLYSOLVER in conjunction with somatic change detection tools (Mutect, Strelka) on 2688 TCGA samples confirmed 30% of the events that were already discovered using conventional approaches, while also identifying about 50% more events. Moreover events that were exclusively identified by standard methods were found to be mostly false positives. The described methodology can easily be extended to other polymorphic loci such as Type II HLA genes, MIC2A, MIC2B and TAP genes.
5.2 Identification of personalized neoantigens in cancer patients

Genome sequencing has revealed a large number of shared and personal somatic mutations across human cancers. In principle, any genetic alteration affecting a protein-coding region has the potential to generate mutated peptides that are presented by surface HLA Class I proteins that might be recognized by cytotoxic T cells. To test this possibility, we developed a pipeline for the prediction and validation of such neoantigens derived from individual tumors and presented by patient-specific alleles of the HLA proteins. We applied our computational pipeline to 91 chronic lymphocytic leukemias (CLL) that had undergone whole-exome sequencing (WES). We predicted 22 mutated HLA-binding peptides per leukemia (derived from ∼16 missense mutations), and experimentally confirmed HLA binding for 55% of such peptides. Finally, we computationally predicted HLA-binding peptides with missense or frameshift mutations for several cancer types and predicted dozens to thousands of neoantigens per individual tumor, suggesting that neoantigens are frequent in most tumors.

5.3 microRNA network inference

Leveraging The Cancer Genome Atlas (TCGA) multidimensional data in glioblastoma, we inferred the putative regulatory network between microRNA and mRNA using the Context Likelihood of Relatedness modeling algorithm. Interrogation of the network in context of defined molecular subtypes identified 8 microRNAs with a strong discriminatory potential between proneural and mesenchymal subtypes. Integrative in silico analyses, a functional genetic screen, and experimental validation identified miR-34a as a tumor suppressor in proneural subtype glioblastoma. Mechanistically, in addition to its direct regulation of platelet-derived growth factor receptor-alpha (PDGFRα), promoter enrichment analysis of context likelihood of relatedness-inferred mRNA nodes established miR-34a as a novel regulator of a SMAD4 transcriptional network. Clinically, miR-34a expression level is shown to be prognostic, where miR-34a low-expressing glioblastomas exhibited better overall survival. This work illustrates the potential of comprehensive multidimensional cancer genomic data combined with computational and experimental models in enabling mechanistic exploration of relationships among
different genetic elements across the genome space in cancer.

We illustrate here that network modeling of complex multidimensional cancer genomic data can generate a framework in which to explore the biology of cancers, leading to discovery of new pathogenetic insights as well as potential prognostic biomarkers. Specifically in glioblastoma, within the context of the global network, promoter enrichment analysis of network edges uncovered a novel regulation of TGF-β signaling via a Smad4 transcriptomic network by miR-34a.


Brennan, C., Momota, H., Hambardzumyan, D., Ozawa, T., Tandon, A., Pedraza, A., and


The microrna body map: dissecting microrna function through integrative genomics. *Nucleic Acids Res.*


