Published in final edited form as:
Biol Blood Marrow Transplant. 2012 January ; 18(1 Suppl): S151–S160. doi:10.1016/j.bbmt.2011.11.011.

Applications of Next Generation Sequencing to Blood and Marrow Transplantation

Michael Chapman, MD[1], **Edus H. Warren III, MD, PhD**[2], and **Catherine J. Wu, MD**[3]

[1]Department of Haematology, Cambridge University, Cambridge UK

[2]Program in Immunology, Fred Hutchinson Cancer Research Center, Seattle, WA

[3]Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston MA

Abstract

Since the advent of next-generation sequencing (NGS) in 2005, there has been an explosion of published studies employing the technology to tackle previously intractable questions in many disparate biological fields. This has been coupled with technology development that has occurred at a remarkable pace. This review discusses the potential impact of this new technology on the field of blood and marrow stem cell transplantation. Hematologic malignancies have been among the forefront of those cancers whose genomes have been the subject of NGS. Hence, these studies have opened novel areas of biology that can be exploited for prognostic, diagnostic, and therapeutic means. Because of the unprecedented depth, resolution and accuracy achievable by NGS, this technology is well-suited for providing detailed information on the diversity of receptors that govern antigen recognition; this approach has the potential to contribute important insights into understanding the biologic effects of transplantation. Finally, the ability to perform comprehensive tumor sequencing provides a systematic approach to the discovery of genetic alterations that can encode peptides with restricted tumor expression, and hence serve as potential target antigens of GvL responses. Altogether, this increasingly affordable technology will undoubtedly impact the future practice and care of patients with hematologic malignancies.

Keywords

genome analysis; exome sequencing; transplantation; TCR

I. Next generation sequencing for understanding hematologic malignancies

This section describes some of the main analytical considerations in NGS, reviews some of the key findings from sequencing hematologic malignancies, and finally looks to the future use of the technology and the challenges and the opportunities that that will bring.

Correspondence should be addressed to: Catherine J. Wu, MD, Dana-Farber Cancer Institute, Harvard Institutes of Medicine, Rm 416B, 77, Avenue Louis Pasteur, Boston MA 02115, cwu@partners.org.

Dr. Warren serves on the Scientific Advisory Board of Adaptive TCR, Inc., but has no financial interest in the company and receives no financial compensation of any kind from the company Drs. Wu and Chapman have no conflicts of interest to disclose

Analytical considerations

Generating sequence—A detailed description of the different NGS technologies is beyond the scope of this discussion; these are reviewed elsewhere (1). Briefly, NGS methods such as 454, SOLiD, and Illumina all involve a process of fragmenting DNA, ligating adapters, and immobilizing the fragments via the adapters to create libraries. The libraries then undergo a process of amplification, generating multiple copies of each DNA fragment. The immobilized, amplified DNA is then sequenced in parallel by a fluorescence- or chemiluminescence-based method, yielding billions of short sequence reads.

If there is no prior selection of the DNA, this is known as whole genome shotgun (WGS) sequencing. Alternatively, by initially hybridizing the DNA fragments to target-specific baits, selected areas of DNA can be captured for sequencing, whilst excluding the rest (2). Frequently, the baits are designed to the coding portion of the genome, resulting in whole exome (WE) sequencing. If only coding mutations are of interest, WE sequencing offers significant cost advantages over WGS sequencing as the same coverage can be achieved with far fewer reads.

Unlike genome-wide association studies (GWAS), the aim in cancer genome sequencing is generally to detect somatic mutations, i.e. those mutations that are unique to the tumor. This necessitates sequencing both tumor and normal tissue from the same patient. This is because the rate of private single nucleotide polymorphisms (SNPs) - variations in the DNA sequence that are unique to the individual and not annotated in databases - is of the order of 1 SNP per 10000 bases. The choice of the normal tissue is governed by the tumor sequenced and the normal tissue should not be contaminated by tumor cells. For example, when sequencing the acute myeloid leukemia (AML) genome, it was necessary to use skin to provide normal tissue (3), whereas in sequencing the multiple myeloma genome, peripheral blood was used (4). However, in the latter case, even though cases of plasma cell leukemia had been excluded, three samples had to be omitted from the analysis because of the presence of low levels of circulating malignant plasma cells.

Sequencing data processing—Having generated the short sequence reads, these are aligned to the human reference genome. Whilst simple in principle, there are a number of difficulties in practise. Given the number of reads and the size of the genome, highly efficient algorithms have to be employed to perform the mapping. However, these algorithms are less accurate than ‘traditional’ alignment algorithms, such as BLAST (5). Compared to Sanger sequencing reads, the reads from NGS are short and can often map in more than one place, especially with repetitive or low complexity DNA. Sequence variations resulting from either sequencing errors or from somatic mutations or SNPs often cause mis-alignment. The algorithms struggle particularly to accurately align reads containing small insertions or deletions (indels). For this reason, it is often necessary to perform local re-alignments in areas containing indels with more accurate but slower algorithms (4).

Sequencing coverage is the average number of reads covering any particular base. The greater the coverage, the greater the chance of accurately calling sequence variations. For cancer genome sequencing, 30X coverage is typical, although accurate mutation calling can be made with coverage of 20X (Illumina technical note: Calling Sequencing SNPs). It is important to remember that the coverage is not uniform across the genome. GC-rich regions are covered poorly, partly because of difficulties with aligning the sequence reads. With hybrid capture techniques in WE sequencing, not all exons may be targeted by the baits or the baits may fail to reliably capture their targets. This may mean that the ability to detect mutations in some genes is markedly impaired (4).

Mutation detection—Point mutations are relatively straightforward to identify by comparison of the reads with the reference sequence they have been mapped to. Poorly mapped reads will give rise to factitious sequence variations and should be excluded prior to mutation calling. Statistical arguments are applied to define the likelihood that the variant is both real and absent from the normal tissue from the same patient. Only mutations scoring above a given threshold are accepted. As a final step, false positive mutations arising from known artefacts of NGS are excluded. A broadly similar approach can be applied to small indels (i.e. indels smaller than the length of a single read), although these are more challenging. Larger indels and translocations are detected with algorithms that detect reads in which the ends are separated by an unexpected length of sequence or indeed are on different chromosomes. With all these approaches, there is a balance to be struck between sensitivity and specificity. In any sequencing project, validation of identified mutations by an alternative sequencing method is usually undertaken and it is becoming clear that the various mutation-calling algorithms have improved markedly in a short period of time.

Identifying driver mutations—One of the most critical steps in the analytical process is the ability to distinguish between the minority of mutations that afford some growth or survival advantage to the tumor (driver mutations), from the majority of mutations that occurred randomly and became fixed within the tumor clones (passenger mutations). This can be done statistically, arguing that across a number of patients with the same tumor, the driver mutations should occur more frequently than would be expected by chance. Such calculations do not simply consider the frequency of the mutations, but also the size of the mutated genes, their base composition, the nature of the mutated bases, and the level to which the genes are expressed in the tumor. Using this framework, there are two broad approaches for defining driver mutations. One is to sequence a small number of samples in a discovery set, define a list of mutated genes, then perform directed sequencing of these genes in a much larger validation set to detect the driver mutations. This is a relatively economical approach, but there is a risk of missing less frequently mutated genes in the discovery set. The alternative model, adopted by The Cancer Genome Atlas (http://cancergenome.nih.gov) and others, is to sequence a large number of samples up-front. The driver mutations are defined in the discovery set and the problem of lower frequency mutations is overcome to some extent, but it is currently a more expensive approach.

Genome analysis of hematologic malignancies

Acute Myeloid Leukemia (AML)—The first cancer genome to be sequenced by NGS was that of a cytogenetically normal AML by a group at the Genome Institute at Washington University (3). This represented an extraordinary technical feat with the technology available at the time and was an important proof of principle. They identified around 31,600 novel somatic mutations and then focused on those that affected genic coding sequences. Following validation by re-sequencing, they were left with 8 true novel somatic coding mutations, in addition to two well described AML mutations, i.e. insertions into the *FLT3* (*FLT3*-ITD) and *NPM1* genes. The same group followed this work with the WGS sequencing of a second AML genome and the directed sequencing of mutated genes in a further 188 AML samples (6). They found a mutation affecting the isocitrate dehydrogenase gene at residue 132 (*IDH1* R132) in 9% of AML samples, exclusively in cases with intermediate risk cytogenetics. Whilst these mutations had not previously been identified in AML, they were known to occur commonly in glioma. Subsequent studies have demonstrated an association between mutations in *IDH1* and *NPM* mutations, with *IDH1* associated with a worse prognosis in *NPM* mutated/*IDH1*-ITD negative AML (7). More recently, the Washington University group adopted a similar NGS approach to identify mutations in the DNA methyltransferase, *DNMT3A*, in 22% of cases of AML (8). Similar to

IDH1, this was associated with intermediate-risk cytogenetics, but predicted a poor prognosis.

Multiple Myeloma (MM)—As discussed previously, an alternative approach to that described in the sequencing of AML is to sequence several tumors up-front. This was the approach that was taken in a large, multi-center study at the Broad Institute to sequence the MM genome, in what was one of the first studies of its kind (4). This study used a combination of WGS and WE sequencing to explore the genomes of 38 MM cases. 10 genes were mutated more frequently than would be expected by chance, including previously described mutations in *NRAS*, *KRAS*, and *TP53*. A quarter of the samples were affected by mutations in one of two genes, *DIS3* and *FAM46C*, not previously implicated in cancer and known or predicted to be involved in RNA processing and/or translation initiation. A single, known activating mutation in the *BRAF* gene (G469A), previously described in melanoma, was found which prompted the genotyping for known *BRAF* mutations in a large separate cohort of MM patients. 4% of samples were affected by these mutations. A highly effective *BRAF* inhibitor, PLX4032, is already under investigation in a Phase 3 clinical trial in melanoma and these results suggest that trials of PLX4032 in MM in targeted individuals would be promising. Having multiple samples in the initial sequencing cohort enabled the application of network analyses to look for mutations in multiple genes targeting the same pathway. By this means, this group was able to confirm and extend the observations of mutations affecting the NF-3B pathway in MM (9)(10) and identify novel mutations predicted to affect histone methylation. Finally, the presence of non-coding mutations clustering in regulatory regions of the genome in a statistically recurrent manner was demonstrated. Over a quarter of samples had mutations in the promoter or first intron of the putative tumor suppressor *BCL7a*. One potential weakness of this study, currently being addressed by further discovery phase sequencing, was that the size was too small and the patient group was too heterogeneous to draw any conclusions about the prognostic significance of any of the mutations. However, it provided an important demonstration of the power of examining multiple genomes by NGS in the discovery set.

Chronic lymphocytic leukemia (CLL)—A collaboration between multiple centers in Spain and the Cancer Genome Project (CGP) at the Wellcome Trust Sanger Institute (WTSI) undertook WGS sequencing of four cases of CLL (11). The mutation rates were relatively low, with only 45 genes affected by coding mutations across all the samples. These genes informed further sequencing in a much larger series. Four genes, *NOTCH1*, *MYD88*, *XPO1*, and *KLHL6* were recurrently mutated with an apparently non-random distribution. The expression of these mutated genes was examined in relationship with an established predictive biologic marker, the degree of somatic hypermutation of the immunoglobulin heavy chain variable region [IGHV]. *NOTCH1* and *XPO1* appeared to be associated with the more aggressive unmutated IGHV status whereas *MYD88* and *KLHL6* appeared to be associated with mutated IGHV status. The *NOTCH1* mutations were frequent (12%), contained premature stop codons predicted to result in activation and stabilization of the protein, and predicted for poor overall survival, although it was not clear whether or not this is independent of the associated unmutated IGHV phenotype.

A second CLL sequencing project, employing 91 tumors in its discovery set, has recently been accepted for publication (*Wang et al, in press*). This group, from the Dana-Farber Cancer Institute together with the Broad Institute, described 9 genes mutated at significant frequency, among them *NOTCH1* and *MYD88*. They confirmed statistically significant associations of these two genes with unmutated and mutated IGHV status, respectively. In addition, they demonstrated significant associations with trisomy 12 and heterozygous 13q deletion, respectively. Perhaps more important, however, was their observation of frequent mutations affecting the gene *SF3B1*. After *TP53*, this was the most frequently mutated gene,

with 15% of CLL affected. There was a strong association with del(11q) and, in a multivariate Cox analysis, *SF3B1* mutation was predictive of poor prognosis, establishing it as an independent prognostic marker. *SF3B1* is a component of the catalytic core of the spliceosome and these investigators were able to demonstrate that *SF3B1* mutation was associated with aberrant splicing in CLL. It is likely that these mutations are associated with widespread changes in the transcriptome, echoing the large-scale transcriptional changes predicted to occur as a result of the frequent *DIS3* and *FAM46C* mutations in MM.

Myelodysplastic Syndrome (MDS)—Recently, frequent mutations in genes in the RNA splicing machinery have been also detected in myelodysplastic syndrome by two independent groups (12)(13). Papaemmanuil *et al*, representing the CGP group at WTSI performed WE sequencing in 9 samples from patients with MDS (12). They identified 46 mutations affecting protein coding across these cases. Intriguingly, 6 out of 9 samples exhibited mutations in *SF3B1*, the same splicing gene found to be mutated in 15% of CLL (*Wang et al, in press*). Targeted sequencing in a much larger cohort revealed that *SF3B1* was mutated in 20% of cases of MDS. These mutations were associated with ringed sideroblasts and a benign clinical course. Many of the mutations were recurrent and there was considerable overlap with the mutations seen in CLL, including the commonest mutation, *K700E*. This strengthens the idea that these mutations alter rather than abrogate function of the spliceosome. Finally, to understand whether any part of the transcriptome is particularly affected by these mutations, Papaemmanuil *et al.* performed gene expression profiling (12). Of note, *SF3B1* mutation was associated with downregulation of several pathways relating to mitochondrial function, including genes involved in the mitochondrial ribosome and the electron transport chain.

In an alternate approach, Yoshida et al. (13), representing the ICGC, examined 29 patients with MDS by whole exome sequencing of paired tumor/control DNA, and subsequently, 582 subjects with myeloid neoplasm by high-throughput mutation screening of pooled DNA. These investigators also identified mutations on multiple components of spliceosome that function in the recognition of 3′-splice site during pre-mRNA splicing, including *SF3B1* and *U2AF35*. Expression of a mutant form of *U2AF35* in HeLa cells induced the increased expression of genes that function in nonsense-mediated mRNA decay (NMD) pathway, indicating activation of cellular responses to abnormally spliced RNA. Furthermore, both exon array and RNA sequencing analysis confirmed the increased expression of non-exon region of genome in the cells transfected with the mutated *U2AF35*. Intriguingly, expression of mutated *U2AF35* suppressed cell proliferation both *in vitro* and *in vivo*, suggesting that the oncogenic effects of these spliceosome mutations are mediated through mechanisms other than cell proliferation.

Hairy Cell Leukemia (HCL)—WE sequencing of a patient with HCL revealed five nonsynonymous coding mutations (14). One of these was a known activating *BRAF* mutation (V600E), well-described in melanoma and latterly in MM (see above). Strikingly, Sanger sequencing of *BRAF* in a further 47 cases of HCL revealed 100% presence of the V600E mutation, suggesting that it is an obligate driver of the disease and that HCL might well be highly susceptible to PLX4032.

Challenges and opportunities of cancer genome sequencing for the future

There are a number of factors that may influence the future of NGS in cancer. One is the concept of statistical power. Genes whose true mutation frequency is 10% should be identified as significant at 80% power in approximately 75 samples. However, for genes with a true mutation frequency of 5%, around 200 samples are required for the same power. For a true frequency of 3%, several hundred samples are required. Compounding this is the

concept of the Winner's Curse. Originally applied to bidding for oil drilling rights in the Gulf of Mexico (where the winner of the bid frequently overpays) and later to the interpretation of GWAS (where polymorphism rates determined in the discovery sets are frequently found to be large overestimates following validation), it is equally applicable to NGS. Many of the observed frequencies from these initial NGS studies may therefore overestimate the true frequencies. Taken together with the fact that there have been not yet been any discovery sets in the hematologic malignancies adequately powered to detect mutations less frequent than about 10%, it is entirely conceivable that we are missing the vast majority of driver mutations, which may affect large numbers of genes at low frequency. This has implications for understanding the interplay of these mutations and for identifying true independent prognostic markers. Sequencing of thousands of tumors may be required to address these issues.

Fortunately, the cost of NGS has fallen at an astonishing rate. The introduction of the so-called third generation of sequencing machines, capable of true single molecule sequencing, will drive down costs and increase productivity further (15), and it may be the case that the computational infrastructure becomes the rate-limiting step to progress. Other novel technologies that are either operational or in the pipeline are RNA sequencing (RNAseq) and single cell sequencing. The former will allow us to probe the transcriptome in an unbiased fashion, examining non-coding and coding RNA, allelic expression of mutations, and splicing patterns. The latter will enable us to ask questions about the clonal composition and evolution of tumors, revealing potential novel Achilles' heels for targeted therapy as well as novel markers of minimal residual disease.

Perspectives

NGS is a truly revolutionary technology that in a short time has identified novel biology in the hematologic malignancies that would unlikely have been discovered by traditional hypothesis-driven research. These new findings offer the potential for exploring targeted therapeutics and some, such as BRAF mutations in MM and HCL, suggest that existing therapies for other conditions should be tested for novel indications. Another promising avenue is to use the novel mutations in risk stratification to better employ current treatment modalities. For example, mutations in intermediate risk AML could in theory be used to define groups of patients who may benefit from early marrow or stem cell transplantation. The principal barrier to this approach is an incomplete understanding of how these mutations interact with one another and with more 'traditional' diagnostic information, such as cytogenetics. It may require much larger sequencing efforts to understand this and to define single mutations or combinations of mutations with independent prognostic significance. However, it is hoped that the staggering rate of technology development in this field will bring these answers sooner rather than later.

II. Next generation DNA sequencing for probing lymphocyte repertoires and tracking lymphoid malignancies

The generation of antigenic specificity and diversity of lymphoid cells

B- and T-lymphocytes are distinguished from all other somatic cells by the fact that much of their biology – indeed, their primary function – is directed by DNA sequence information that is not encoded within the germline. The antigenic specificity of B- and T-cells is in large part determined by the amino acid sequence in the complementarity-determining regions (CDRs) of their antigen receptors. The CDR1 and CDR2 regions in both B- and T-cell antigen receptors expressed by antigen-naïve lymphocytes are encoded in the germline, but the sequence that encodes the CDR3 region – arguably the most critical determinant of

antigenic specificity – is generated during lymphocyte development by recombination between noncontiguous gene segments in the B- and T-cell receptor loci.

The CDR3 regions in the β and δ chains of αβ and γδ T cell receptors (TCRs) and in the heavy chain of B-cell receptors (BCRs) are formed by recombination between noncontiguous variable (V), diversity (D), and joining (J) gene segments in the TCRβ, TCRδ, and IgH loci, while the CDR3 regions in the α and γ chains of αβ and γδ TCRs and in BCR light chains are formed by recombination between analogous sets of variable and joining gene segments in the TCRα, TCRγ, κ light chain, and λ light chain loci. The existence of multiple variable, diversity, and joining gene segments in the four T-cell and three B-cell antigen receptor loci allows for a large number of distinct TCR and BCR CDR3 sequences to be encoded. CDR3 sequence diversity in both TCRs and BCRs is further increased by template-independent addition and deletion of nucleotides at the junctions between the different classes of gene segments. Somatic hypermutation of previously rearranged B-cell receptor genes, which is not limited to the CDR3 region, can also occur in B cells after initial antigen encounter, further increasing diversity within the immunoglobulin repertoire. The adaptive immune system utilizes this ingenious multi-component strategy to generate an extremely diverse repertoire of B- and T-cell antigen receptors that can collectively recognize the universe of potential pathogens.

Deep sequencing of the CDR3 region to probe repertoire diversity

The enormous magnitude of the CDR3 sequence diversity that can be created with this strategy is far too great to permit comprehensive exploration and definition using conventional capillary-based DNA sequencing. The advent of next-generation sequencing over the last 6 years (16), however, has enabled analysis of the B- and T-cell receptor CDR3 region sequence repertoires realized in any given individual with unprecedented depth, resolution, and accuracy (17)(18)(19)(20).

Because the CDR3 region in the vast majority of successfully rearranged BCR and TCR genes comprises no more than 60 nucleotides, encoding no more than 20 amino acids, the CDR3 region sequence repertoire is ideally suited to comprehensive definition by current sequencing platforms that can generate on the order of 2×10^8 (or more) sequence reads of length ≥60 nucleotides in a single sequencing run. The capacity for generating such extremely large sequence datasets has made it necessary to enjoin the efforts of computational biologists to develop analytical strategies to deal with the deluge of sequence data, and has thus provided the rationale for the nascent field of computational immunology. The development of these powerful computational and molecular strategies for probing the BCR and TCR CDR3 sequence repertoires expressed in lymphocyte populations of virtually any degree of complexity has, in turn, made it possible to address biological questions that were never before amenable to direct experimental analysis.

Normal adult human B and T cell repertoire diversity – parameters and questions

In one example, high-throughput sequencing of rearranged BCR and TCR genes from peripheral blood lymphocytes of healthy adults has begun to define the critical characteristics of the B- and T-cell repertoires that are established and maintained in adult life. The number of unique IgH, TCRα, and TCRβ rearrangements that can be detected in the blood of a single individual, for example, provide a basis on which to estimate the total number of unique B- and αβ T-cell receptors that are present in the individual at any one time. The number of distinct IgH rearrangements present in the peripheral blood B cell pool, for example, appears to be at least ~2×10^6 (17). If comparable diversity exists within the Igκ and Igλ repertoires, as is thought to be likely, this would suggest that the diversity of B-cell antigen receptors expressed in the peripheral blood B cell repertoire is potentially very

high. Analogous studies of TCRβ rearrangements in peripheral blood αβ T cells (18)(19)(20)(21) have likewise established lower bounds for the diversity of TCRβ chains expressed in the naïve and memory CD4+ and CD8+ compartments of healthy adults, and indicate that the total number of unique TCRβ chains in the peripheral blood is at least 3–4 × 10^6 (19). Although published data on the diversity of TCRα chains expressed in peripheral blood αβ T cells are not as extensive as for TCRβ, the available data suggest that TCRα diversity may be comparable to that of TCRβ (20). Global comparison of the TCRβ repertoires expressed in peripheral blood $CD8^+$ T cells from different individuals of diverse geographic and ethnic origin and sharing few or no MHC class I alleles has shown that the overlap between the repertoires of any two individuals is far higher than expected and also seemingly independent of HLA type (22). Whether the TCRα repertoires, and, more importantly, the repertoires of TCR αβ heterodimers, expressed in different individuals exhibit similar overlap has not yet been determined.

An important challenge currently facing the new discipline of computational immunology is to define how, and to what extent, effective clinical immunity against pathogens relates to and depends on the enormous diversity that characterizes lymphocyte antigen receptor repertoires. Is there, for example, a minimum or threshold level of repertoire diversity required for effective immunity against the wide spectrum of viral, bacterial, fungal, and protozoan pathogens that we typically encounter in our lives? To what extent is the diversity that exists in the entire B- and T-lymphocyte repertoires found in the repertoires of the numerous functional subsets of B- and T-cells, and is the diversity within any specific subset particularly important? For example, is receptor diversity in the regulatory T cell repertoire – which, on the basis of published data, appears to be comparable to that in the effector T cell repertoire (20)– related in any systematic way to autoimmune disease, or to the occurrence of chronic GVHD after allogeneic HCT? Studies addressing these questions are currently in progress.

Monitoring rare T cell populations

High throughput sequencing has recently been used to study diversity within the small minority of T cells in peripheral blood that express γδ rather than αβ T-cell receptors (23). Commitment of T cells to the αβ or γδ lineages takes place in the thymus, but the factors that influence this lineage decision have not been well defined (24)(25)(26). In contrast to αβ T cells, which primarily recognize peptide antigens presented by class I or class II MHC molecules, γδ T cells recognize a variety of unconventional ligands without the participation of class I or class II MHC (27). Although it has been estimated that the potential diversity of antigen receptors that can be expressed in γδ T cells exceeds that of αβ T cells or B cells (27), the antigen receptors expressed by peripheral blood γδ T cells are overwhelmingly dominated by specific subsets characterized by a very limited range of antigenic specificity. Indeed, deep sequencing of the rearranged TCRγ genes expressed in peripheral blood γδ T cells from three healthy adults revealed that >45% of the TCRγ CDR3 sequences from the three individuals were identical to a previously described sequence found in a shared γδ TCR that is specifically reactive with nonpeptide prenyl pyrophosphate antigens (28).

Deep sequencing of both the TCRβ and TCRγ loci in αβ and γδ T cells in peripheral blood of healthy adults has also provided valuable insights into the process of αβ versus γδ lineage commitment in the thymus that have important implications for the monitoring of T-lymphoid malignancies using tumor-specific TCR rearrangements (23). Although the vast majority of αβ and γδ T cells in peripheral blood carry rearranged TCRγ genes, suggesting that the TCRγ locus rearranges prior to αβ/γδ lineage commitment, a miniscule fraction (<4%) of γδ T cells appear to have rearranged TCRβ loci, suggesting that rearrangements of TCRβ occur only in T cells that have committed to the αβ lineage. These results therefore

suggest that the TCRγγ locus should be the focus of molecular strategies for monitoring T-lymphoid malignancies.

Monitoring of malignant clones

The enormous capacity of current high-throughput sequencing platforms and the profound sequencing depth that they provide are increasingly being exploited to monitor the lymphocyte repertoires in patients with B- and T-lymphoid malignancies and to identify the malignant clone[s] at the time of diagnosis and to track them during and after therapy. Serial monitoring of the peripheral blood B cell compartment in patients with CLL, follicular NHL, and posttransplant lymphoproliferative disease (PTLD) using IgH CDR3 sequencing, for example, can uniquely identify the malignant B cell clones that drive the disease and follow their suppression as therapy is administered, as well as their subsequent reappearance, before clinical relapse is detectable (17)(29). Moreover, high-throughput sequencing can provide accurate and precise assessments of the volume of disease in a given sample, such as peripheral blood, bone marrow, or lymph node.

Perspectives

Recent studies suggest that sequencing methods can reproducibly detect the presence of a T cell clone with a specific TCR rearrangement in a background of 100,000 others (30), which is comparable to the sensitivity of PCR-based methods. In contrast to PCR-based methods, however, high-throughput sequencing provides comprehensive information about the entire lymphocyte repertoire in a patient, not just the malignant clone[s], and will therefore provide insights into the disease process that would not be obtainable with a PCR-based approach. For this reason, it is anticipated that high-throughput sequencing will soon replace PCR for monitoring disease burden in patients with lymphoid malignancies, particularly if the cost of such sequencing continues to fall as rapidly as it has over the last several years (http://www.synthesis.cc/assets_c/2011/06/carlson_cost%20per_base_june_2011.html). Indeed, it is tempting to speculate that serial monitoring of lymphocyte repertoires with high-throughput sequencing will soon become a standard feature of the care of patients with lymphoid malignancies, autoimmune diseases, and immunodeficiency, as well as patients undergoing allogeneic hematopoietic cell transplantation or other forms of immunotherapy.

III. Genome analysis to discover targets of GvL responses

The Holy Grail, still: Separating GvL from GvHD

Several lines of evidence have definitively established the critical role played by donor-derived T cells following allogeneic HSCT (allo-HSCT) in generating curative responses. These include the observations of improved relapse-free survival following allogeneic compared to autologous HSCT, increased disease relapse following HSCT when using T cell-depleted grafts, and examples of leukemia regression observed following infusion of donor T cells (31). Donor-derived T cells can mediate GvL effects through two general mechanisms. Firstly, engraftment of donor cells restores normal immune function, thus overcoming tumor- or treatment-induced host immune defects and restoring immunosurveillance of malignant cells. Secondly, donor-derived T cells may recognize host antigens and eliminate cells bearing these antigens. Recognition of the beneficial effects of GVL has led to major changes in the landscape of allo-HSCT. In particular, they have provided the underlying rationale for developing less intensive preparative regimens, that have broadened the availability of allo-HSCT as a therapeutic option to older patients and to those individuals with co-morbidities that otherwise would not have been able to withstand more intensive chemotherapy or radiation. These regimens typically do not completely eliminate host hematopoiesis, and hence rely on GvL for their efficacy.

Unfortunately, these desired immunologic effects come at a cost. Too often, GvL responses following allo-HSCT arise in the setting of acute or chronic GvHD. While the immunologic targeting of tumor cells is beneficial, similar targeting of normal recipient tissues remains a major cause of morbidity and mortality following HSCT. Thus a major goal of allotransplantation remains separation of GvL from GvHD, so that curative responses can be generated with minimal toxicity.

Genetic basis of GvHD vs GvL targets

One approach to distinguish GvL from GvHD effects is to define differences in their target antigen specificities. Numerous studies have provided evidence that GvHD arises from immunologic recognition of polypeptides that are encoded by genetic polymorphisms existing throughout the human genome (32), that differ between donor and recipient. Transplantation of mature T cells during allogeneic HSCT results in the transfer of large numbers of cells capable of recognizing these alloantigens. The various mechanisms by which genetic polymorphisms (minor histocompatibility antigens, or ‘mHA’) can give rise to allo-antigens include: amino acid substitutions that create antigenic peptides, creation of alternate transcripts, modification of proteasomal processing, post-translational modifications, or gene deletions. The clinical significance of a mHA is highly dependent on the tissues and cell types that express the target antigen. Targeting allo-antigens that are broadly expressed in normal recipient tissues (hematopoietic and non-hematopoietic) results in GvHD. When these allo-antigens are also expressed on leukemia cells, targeting these antigens contributes to GvL. When mHA are only expressed in hematopoietic tissues, donor T cells targeting these antigens result in the elimination of recipient hematopoiesis and conversion to full donor hematopoiesis. Finally, when leukemia cells also express these mHA, targeting hematopoietic allo-antigens can result in GvL without concomitant GvHD.

Alternatively, GvL responses can also be observed if immune responses are directed against antigens that are expressed solely on the tumor cell. A handful of antigens with leukemia-restricted expression are already known. These include leukemia-specific antigens (epitopes arising from chromosomal rearrangements such as BCR-ABL), virally encoded antigens (latent EBV epitopes), over-expressed self-antigens (proteinase-3, WT-1), cancer-testis antigens (NY-ESO-1) or mutated/modified self-antigens (31). While recipients with leukemia may have become tolerant to these antigens, normal donors remain capable of developing effective immune responses after transplantation. Overall, tumor-associated antigens have been challenging to identify since conventional methods for identifying tumor or transplantation antigens are laborious, typically requiring isolation of patient T cell clones followed by determination of their peptide-HLA target using expression cloning. In recent years, antibody responses to several GvL-associated antigens have been identified, many of which appear to represent tumor-associated rather than mHA, and that have the potential to elicit donor T cells immunity (33)(34). Nonetheless, relatively little is known about the nature of tumor specific antigens. If *bona fide* GvL target antigens were identified, novel immunotherapy approaches, perhaps through vaccination or adoptive T cell therapy, could be implemented to generate and maintain tumor control and eradication through development of tumor-specific responses.

Neoantigens as potential GvL targets

Tumor neoantigens have been previously proposed to be an immunologically important class of tumor-specific antigens (35), but this hypothesis could not be rigorously tested until now because of technical barriers to their identification. Just like mHA, they can potentially arise as a result of genetic changes, but this time, as a reslt of tumor-driven mutation rather than from polymorphism: somatic mutations that give rise to frameshift insertions or deletions, gene fusions, alternative splicing. The potential effectiveness of targeting mutated

antigens, or neoepitopes, in the immune control of tumors has been appreciated in seminal studies showing that: (a) mice and humans often mount T cell responses to mutated antigens (35)(36); (b) mice can be protected from a tumor by immunization with a single mutated peptide that is present in the tumor (37); (c) spontaneous or vaccine-mediated long-term melanoma survivors mount strong memory cytotoxic T cell (CTL) responses to mutated antigens (38)(39)(40); and finally (d) that patients with follicular lymphoma show molecular remission when immunized with patient-specific mutated immunoglobulin proteins that are present in autologous tumor cells (41)(42). However, as targets for vaccination, they have rarely been used in vaccines due to the technical difficulties in identifying them (35).

Comprehensive discovery of tumor neoantigens

The obstacles for discovering personal tumor neo-antigens have recently been potentially surmounted with the advent of next-generation sequencing technology. Recent large-scale traditional sequencing efforts have demonstrated that an average tumor may have tens to hundreds of protein-coding changes (43)(44). Such mutated proteins have the potential to: (a) uniquely mark a tumor for recognition and destruction by the immune system (35), thus reducing the risk for autoimmunity; and (b) avoid central and peripheral T cell tolerance, allowing the antigen to be recognized by more effective, high avidity T cells receptors. In this instance, ‘passenger’ mutations, that may not have been of interest from the standpoint of oncogenesis, do have potential relevance for eliciting immunity. From a probability standpoint, higher mutation rates may generate more chances to generate epitopes. A recent *in silico* analysis of sequences derived from tumor and normal cells in the same patients suggest that these somatic mutations provide ~10 novel neoepitopes that can bind HLA-A*0201 for tumor vaccine development (45).

Perspectives

Fusing genomic data with immunologic studies will enable the evaluation of the immunologic effects of personal tumor neoantigens in a way that has not been historically possible using conventional methodologies of T cell antigen discovery. Recent studies in cancer genome sequencing have increasingly suggested that greater diversity of genetic changes in tumors exist than previously anticipated, and include point mutations, gene fusion and alternative splicing events, the vast majority of which appear to be private to an individual tumor.

These results have interesting immunologic implications. Based on the older studies described above and on the difficult challenge that overexpressed antigens are likely to be expressed in some normal tissue, we anticipate that personal tumor neoantigens will play an increasingly important role in the development of highly focused and potent cancer vaccines. Our tool kit for generating effective vaccines has vastly increased in recent years, and range from the development of novel vaccine delivery methods, of more potent adjuvants, and of highly active checkpoint blockade inhibitors (*Brusic and Wu, in press*). In this context, the importance of defining the truly tumor-specific antigens is heightened, to ensure that focused potent immune responses that could lead to effective destruction of tumor cells (without autoimmunity) can be implemented. The reality of integrating whole tumor sequencing into the clinical therapeutic setting is increasingly feasible as the costs of genome sequencing drop. Deeper investigation in this area can address the many as yet unanswered questions related to tumro noantigens include: Which and what fraction of tumor neoantigens are detected by T cells? How frequent are neoantigen-specific memory and effector T cells in circulation and in the tumor? How much avidity do T cells have for these antigens? And are neoantigen-specific T cells functional? Addressing these questions will inform any potential applications of tumor neoantigens in vaccines.

Acknowledgments

M.C. acknowledges support from Leukaemia and Lymphoma Research. E.H.W. is supported by a Clinical Scientist Award in Translational Research from the Burroughs Wellcome Fund (1007475) and NIH grants P30 CA015704-37, R43 DK089783, and R56 AI081860. C.J.W is supported by a Clinical Investigator award from the Damon-Runyon Cancer Research Foundation (CI-38-07) and acknowledges support from the Leukemia and Lymphoma Society Translational Research Program, the Blavatnik Family Foundation, and from the NIH (NCI -1R01CA155010-01A1).

References

1. Morozova O, Marra Ma. Applications of next-generation sequencing technologies in functional genomics. Genomics. 2008; 92:255–64. [PubMed: 18703132]
2. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology. 2009; 27:182–9.
3. Ley TJ, Mardis ER, Ding L, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature. 2008; 456:66–72. [PubMed: 18987736]
4. Chapman, Ma; Lawrence, MS.; Keats, JJ., et al. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011; 471:467–72. [PubMed: 21430775]
5. Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. Nature Biotechnology. 2009; 27:455–7.
6. Mardis ER, Ding L, Dooling DJ, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. NEJM. 2009; 361:1058–66. [PubMed: 19657110]
7. Paschka P, Schlenk RF, Gaidzik VI, et al. IDH1 and IDH2 mutations are frequent genetic alterations in acute myeloid leukemia and confer adverse prognosis in cytogenetically normal acute myeloid leukemia with NPM1 mutation without FLT3 internal tandem duplication. Journal of Clinical Nncology. 2010; 28:3636–43.
8. Ley TJ, Ding L, Walter M, et al. DNMT3A Mutations in Acute Myeloid Leukemia. NEJM. 2010; 363:2424–2433. [PubMed: 21067377]
9. Annunziata CM, Davis RE, Demchenko Y, et al. Frequent engagement of the classical and alternative NF-κB pathways by diverse genetic abnormalities in multiple myeloma. Cancer Cell. 2007; 12:115–130. [PubMed: 17692804]
10. Keats JJ, Fonseca R, Chesi M, et al. Promiscuous Mutations Activate the Non-Canonical NF-kB Pathway in Multiple Myeloma. Cancer Cell. 2007; 12:131–144. [PubMed: 17692805]
11. Puente XS, Pinyol M, Quesada V, et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature. 2011:3–7.
12. Papaemmanuil E, Cazzola M, Boultwood J, et al. Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts. NEJM. 2011; 365:1384–1395. [PubMed: 21995386]
13. Yoshida K, Sanada M, Shiraishi Y, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. Nature. 2011; 478:64–69. [PubMed: 21909114]
14. Tiacci E, Trifonov V, Schiavoni G, et al. BRAF Mutations in Hairy-Cell Leukemia. NEJM. 2011; 364:2305–2315. [PubMed: 21663470]
15. Munroe DJ, Harris TJR. Third-generation sequencing fireworks at Marco Island. Nature Biotechnology. 2010; 28:426–8.
16. Mardis ER. A decade's perspective on DNA sequencing technology. Nature. 2011; 470:198–203. [PubMed: 21307932]
17. Boyd SD, Marshall EL, Merker JD, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Trans Med. 2009; 1:1–16.
18. Freeman JD, Warren RL, Webb JR, Nelson BH, Freeman JD, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Research. 2009:1817–1824. [PubMed: 19541912]
19. Robins HS, Campregher PV, Srivastava SK, et al. Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. Blood. 2009; 114:4099–107. [PubMed: 19706884]

20. Wang C, Sanders CM, Yang Q, et al. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. PNAS. 2010; 107:1518–23. [PubMed: 20080641]
21. Warren RL, Freeman JD, Zeng T, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Research. 2011; 21:790–797. [PubMed: 21349924]
22. Robins HS, Srivastava SK, Campregher PV, et al. Overlap and effective size of the human CD8+ T cell receptor repertoire. Sci Transl Med. 2010; 2:47ra64.
23. Sherwood AM, Desmarais C, Livingston RJ, et al. Deep Sequencing of the Human TCRγ and TCRβ Repertoires Suggests that TCR Rearranges After and T Cell Commitment. Sci Transl Med. 2011; 3:90ra61–90ra61.
24. Kreslavsky T, Gleimer M, Garbe A, vonBoehmer H. αβ versus γδ fate choice: counting the T-cell lineages at the branch point. Immunol Rev. 2010; 238:169–81. [PubMed: 20969592]
25. Kreslavsky T, von Boehmer H. γδ TCR ligands and lineage commitment. Seminars in Immunology. 2010; 22:214–21. [PubMed: 20447836]
26. Ciofani M, Zúñiga-Pflücker JC. Determining γδ versus αβ T cell development. Nature Rev Immunology. 2010; 10:657–63. [PubMed: 20725107]
27. Carding SR, Egan PJ. γδ T cells: functional plasticity and heterogeneity. Nature Rev Immunology. 2002; 2:336–45. [PubMed: 12033739]
28. Wang H, Fang Z, Morita CT. Vgamma2Vdelta2 T Cell Receptor recognition of prenyl pyrophosphates is dependent on all CDRs. Journal of Immunology. 2010; 184:6209–22.
29. Kalos M, Levine BL, Porter DL, et al. T Cells with Chimeric Antigen Receptors Have Potent Antitumor Effects and Can Establish Memory in Patients with Advanced Leukemia. Sci Transl Med. 2011; 3:95ra73–95ra73.
30. Robins H, Desmarais C, Matthis J, et al. Ultra-sensitive detection of rare T cell clones. Journal of Immunological Methods. 2011:6–11.
31. Wu C, Ritz J. Induction of tumor immunity following allogeneic stem cell transplantation. Adv Immunol. 2006; 90:133–73. [PubMed: 16730263]
32. Mullally A, Ritz J. Beyond HLA: the significance of genomic variation for allogeneic hematopoietic stem cell transplantation. Blood. 2007; 109:1355–62. [PubMed: 17008540]
33. Biernacki AM, Marina O, Zhang W, et al. Efficacious immune therapy in chronic myelogenous leukemia (CML) recognizes antigens that are expressed on CML progenitor cells. Cancer Research. 2010; 70:906–15. [PubMed: 20103624]
34. Zhang W, Choi J, Zeng W, et al. Graft-versus-leukemia antigen CML66 elicits coordinated B-cell and T-cell immunity after donor lymphocyte infusion. Clinical Cancer Research. 2010; 16:2729–39. [PubMed: 20460482]
35. Sensi M, Anichini A. Unique tumor antigens: evidence for immune control of genome integrity and immunogenic targets for T cell-mediated patient-specific immunotherapy. Clinical Cancer Research. 2006; 12:5023–32. [PubMed: 16951217]
36. Parmiani G, Filippo A, De Novellino L, Castelli C. Unique human tumor antigens: immunobiology and use in clinical trials. Journal of Immunology. 2007; 178:1975–9.
37. Mandelboim O, Vadai E, Fridkin M, et al. Regression of established murine carcinoma metastases following vacination with tumour-associated antigen peptides. Nature Medicine. 1995; 1:1179–83.
38. Zhou J, Shen X, Huang J, Hodes RJ, Rosenberg SA, Robbins PF. Telomere Length of Transferred Lymphocytes Correlates with In Vivo Persistence and Tumor Regression in Melanoma Patients Receiving Cell Transfer Therapy. The Journal of Immunology. 2005; 175:7046–7052. [PubMed: 16272366]
39. Lennerz V, Fatho M, Gentilini C, et al. The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. PNAS. 2005; 102:16013–8. [PubMed: 16247014]
40. Huang J, El-Gamil M, Dudley ME, Li YF, Rosenberg Sa, Robbins PF. T cells associated with tumor regression recognize frameshifted products of the CDKN2A tumor suppressor gene locus and a mutated HLA class I gene product. Journal of Immunology. 2004; 172:6057–64.

41. Baskar S, Kobrin CB, Kwak LW. Autologous lymphoma vaccines induce human T cell responses against multiple, unique epitopes. Journal of Clinical Investigation. 2004; 113:1498–1510. [PubMed: 15146248]

42. Timmerman JM, Czerwinski DK, Davis Ta, et al. Idiotype-pulsed dendritic cell vaccination for B-cell lymphoma: clinical and immune responses in 35 patients. Blood. 2002; 99:1517–26. [PubMed: 11861263]

43. Sjöblom T, Jones S, Wood LD, et al. The consensus coding sequences of human breast and colorectal cancers. Science. 2006; 314:268–74. [PubMed: 16959974]

44. Thomas RK, Baker AC, Debiasi RM, et al. High-throughput oncogene mutation profiling in human cancer. Nature Genetics. 2007; 39:347–51. [PubMed: 17293865]

45. Segal NH, Parsons DW, Peggs KS, et al. Epitope landscape in breast and colorectal cancer. Cancer Research. 2008; 68:889–92. [PubMed: 18245491]