Published in final edited form as:

Semin Hematol. 2014 October ; 51(4): 250–258. doi:10.1053/j.seminhematol.2014.08.003.

Sequencing the AML Genome, Transcriptome, and Epigenome

Elaine R. Mardis, Ph.D

Robert E. and Louise F. Dunn Distinguished Professor of Medicine, Professor of Genetics and Molecular Microbiology, Co-director, The Genome Institute, Washington University School of Medicine, St. Louis, MO USA

Abstract

Leukemia is a disease that develops as a result of changes in the genomes of hematopoietic cells, a fact first appreciated by microscopic examination of the bone marrow cell chromosomes of affected patients. These studies revealed that specific subtypes of leukemia diagnosis correlated with specific chromosomal abnormalities, such as the t(15;17) of acute promyelocytic leukemia¹ and the t(9;22) of chronic myeloid leukemia². Over time, our genomic characterization of hematologic malignancies has moved beyond the resolution of the microscope to that of individual nucleotides in the analysis of whole genome sequencing data using state-of-the-art massively parallel sequencing (MPS) instruments and algorithmic analyses of the resulting data. In addition to studying the genomic sequence alterations that occur in patient's genomes, these same instruments can decode the methylation landscape of the leukemia genome and the resulting RNA expression landscape of the leukemia transcriptome. Broad correlative analyses can then integrate these three data types to better inform researchers and clinicians about the biology of individual acute myeloid leukemia (AML) cases, facilitating improvements in care and prognosis.

Overview of Massively Parallel Sequencing

Historically, DNA sequencing by the Sanger dideoxynucleotide method³ required two decoupled steps; 1) the DNA polymerase-catalyzed sequencing reaction during which fluorescent labels are incorporated into the ladder of synthesized strands and 2) the electrophoretic separation and detection of the fluorescently labeled sequencing reaction fragments. While this method was used successfully for small and large sequencing projects, including sequencing the Human Reference Genome⁴, the introduction of massively parallel, or “next generation” sequencing (NGS) platforms over the past decade has radically changed the nature of DNA sequencing^{5,6}. There are fundamental differences in this type of sequencing that inherently enable massive data generation in a truncated timeframe, at significantly lower cost than by Sanger-based methods. In particular, a next-generation

© 2014 Elsevier Inc. All rights reserved.

Correspondence to: T: 314-286-1805 F: 314-286-1810 emardis@wustl.edu Post: 4444 Forest Park Ave., Box 8501, St. Louis MO 63108 USA.

Conflict of interest statement: No conflicts exist.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

platform sequencing library requires only a few discrete molecular biology reactions to generate, starting from either fragmented high molecular weight genomic DNA, cDNA (RNA converted to DNA by reverse transcriptase) or PCR products. Following library generation, the fragments are polymerase amplified on a solid surface, either a non-porous bead or flat microfluidic channel. The amplification of discrete fragments enables the signals generated during the sequencing reaction to have sufficient intensity for detection. Data generation takes place after priming of all DNA fragments in the amplified population, using a step-wise process to extend from each primer end with an incorporated nucleotide, remove the excess nucleotides, and detect the identity of that incorporated nucleotide at each fragment location. The en-masse data generation mode enabled by these methods invokes the term “massively parallel”, and the generated data volumes can be sufficient to produce coverage of a complete human genome. Unlike Sanger sequencing, which produces data from a population of fragments for each reaction, massively parallel platforms generate data from individually amplified library fragments. As a result, the data are digital, maintaining a 1:1 relationship in terms of the fragment population to the copy number present in the input DNA source. Downstream analyses can invoke this digital nature in exact quantitation of DNA copy number, RNA expression, the proportions of mutation-containing cells in a population, and other characterizations.

During the multiple rounds of sequence data generation steps, various sources of noise emerge, ultimately limiting the read length that can be obtained at the point that signal and noise are not well differentiated and the quality of the called nucleotide is questionable. As manufacturers of these platforms have improved their chemistry, molecular biology, and imaging sensitivity and decreased their cycle times, the read lengths have improved. At present, however, read lengths are significantly shorter than from Sanger-based sequencing data, making read alignment to a reference genome necessary for interpretation of the resulting data. However, the improving quality of the Human Reference Genome and of algorithms that accurately place sequencing reads from a massively parallel experiment onto the genome sequence fueled a revolution in our ability to characterize genomes at many levels, as will be described. In the sequencing of cancer-derived DNA, the typical goal is to determine changes that are unique to the tumor genome, or “somatic”. Here, the scope and digital nature of MPS enables a vast number of somatic variant discoveries when comparing data from tumor to matched normal tissue isolates from patient samples. The following sections explore this comparison across the spectrum from whole genome to exome to gene panel sequencing assays. In addition to surveying the tumor-specific alterations of DNA, the RNA landscape also can be studied as can the methylation landscape. RNA and methylation patterns are perhaps most informative when compared to a matched normal tissue from the same patient. However, a comparator or matched normal tissue is often not possible to obtain or the cell type not identified/known, so interpretation may be limited to comparing results either to large numbers of surveyed normal tissues or to reference data sets, when they exist.

Challenges of Next-Generation Sequencing using Clinical AML Samples

In general, the challenges of next-generation assays to characterize DNA derived from bone marrow or blood cells are highly unique to the hematologic malignancies and are quite

different from those of solid tissue malignancies. For example, bone marrow samples often provide abundant tumor DNA/RNA sources for assay, with low amounts of contaminating normal cells. By contrast, some tissues (e.g., myeloid sarcomas) will require flow sorting to enrich the tumor cell population prior to nucleic acid isolation, a process that often produces a limited number of cells and hence a limited source of nucleic acid for the NGS assays. Another challenge is due to the comparator normal tissue, often obtained from skin taken at the marrow biopsy site. In many cases, skin biopsies from AML patients with high circulating white cell counts (or leukemias with monocytic differentiation) will present a contaminating source of tumor nucleic acids, due to the capillary beds present in skin. Furthermore, limited amounts of skin are available for nucleic acid isolation, and substituting buccal swabs or mouthwash samples contributes a microbiome component to the DNA/RNA isolation that will make the sequencing less efficient in terms of obtaining coverage on the human genome. As mentioned above, for RNA and methylation analyses, a true comparator normal tissue may either be not known or is controversial. In this case, comparisons of expression or methylation signatures across multiple affected tissues are required to build a case that any given expressed gene(s) or methylation change is contributing to disease onset or progression.

Whole Genome Sequencing

Figure 1a outlines a process for whole genome sequencing (WGS) data generation that is applicable to generate the coverage needed from a MPS platform. WGS is widely considered to be the gold standard for cancer sequencing because it provides data that can be interpreted across the broad spectrum of somatic variation, from single nucleotide substitution mutations to large structural rearrangements. All of these types of variants have been demonstrated to create “drivers” of cancer development and hence a WGS approach will provide the most complete information about somatic alterations, but these are also the most expensive and difficult data to analyze⁷.

There are multiple considerations when planning a WGS study to compare matched tumor and normal genomes, that include having a fundamental understanding of 1) the purity of tumor cells being used for the sequencing library, 2) the relative amount of chromosomal amplification or loss present in the tumor genome, and 3) the types of variants that one wants to detect from the resulting data. All three factors contribute to the amount of genome “coverage” to be generated, where coverage is defined as the depth of whole genome sampling required to discover variants. In particular, if the tumor purity is below 100%, additional coverage will be required to compensate for the normal cell genome contributions to the library. If there are large-scale chromosomal arm or whole chromosome amplifications, additional sequencing coverage will be needed to bring diploid regions of the genome up to sufficient coverage levels to detect variants. Typically, ploidy is gauged either by signal strength-based analysis of the tumor DNA after hybridization to a high-density single nucleotide polymorphism (SNP) microarray platform or by cytogenetic examination of the tumor chromosomes.

Since most tumor WGS projects seek to identify single nucleotide substitution and focused insertion/deletion mutations, a minimum coverage of 30-fold for the tumor genome (with

caveats listed above) is required. Although copy number and structural variant alterations can be detected from 30-fold coverage, the false positive rate (FPR) of discovery and the false negative rate (FNR) of missed events of these types both decrease with increasing coverage. Routinely, tumor DNA WGS libraries are sequenced at 50–60-fold coverage to enhance discovery of more complex alterations and the confidence of mutation detection increases commensurately. Lower tumor purity can be compensated by increasing coverage, as can chromosomal ploidy differences. In particular, one helpful gauge of achieving appropriate coverage for mutation detection is by comparing the variants identified by massively parallel data (aligned to the genome and with variants identified genome-wide) to the SNP calls at these loci as obtained from the array data proposed above to gauge chromosomal amplification. In particular, as one approaches 98% concordance between SNP array calls and single nucleotide variant calls from MPS data, one can be assured of sufficient coverage for mutation detection.

Library construction for optimal WGS coverage requires an approach that fulfills the following criteria: 1) sufficient diversity in the library fragment population to obtain the desired coverage without exhausting uniqueness, 2) input DNA quantities sufficient to provide the desired coverage without exhausting uniqueness, 3) tight size fractions that enable structural variant detection with precision and low FPR. In situations with low input and/or fragmented DNA (e.g. from formalin-preserved tissue isolates), a method to provide library fragment diversity and minimize duplication rates must be utilized.

The primary challenge to maintaining a diverse fragment population during library construction is due to the use of polymerase chain reaction (PCR) at various steps, illustrated in Figure 1b. In particular, PCR can introduce amplification biases (often referred to as “jackpotting”) that result in certain fragments being preferentially amplified and not others. Often, this preferential amplification is a function of A+T content and of DNA fragment size. Furthermore, if a limited amount of DNA (below 50 ng) is used for library construction, the low input portends a relative lack of diversity that is exacerbated by PCR amplification steps. In general, since limited material is common in clinical samples, at the Washington University Genome Institute, we have carefully titrated the input amounts of several components in library construction such as the synthetic adapters, reduced the number of DNA loss steps by use of a universal buffer for several of the enzymatic treatments, and established a process for “titrating” to find the minimum number of PCR amplification cycles before a high degree of amplification bias develops⁸. For WGS projects, we also commonly select 2–4 size ranges during the final library preparatory steps, either by polyacrylamide gel sizing and excision relative to a co-electrophoresed size standard or by using an automated sizing device such as the Caliper XT (Perkin Elmer) or the Blue Pippin (SAGE Biosciences). These approaches can isolate discrete size fractions of +/-50 bp that are amenable to somatic structural variant detection algorithms (i.e. BreakDancer⁹). In sequencing these different size fractions, we produce equal amounts of data from 2–4 of the collected fractions to ensure library fragment diversity and hence, complete genome coverage.

Once genome coverage is accumulated to the desired depth for both tumor and normal WGS, aligned to the Human Reference Genome and checked for satisfactory correlation to

array-called SNPs, the analytical process to compare tumor to normal can begin. Although a thorough review of somatic variant detection algorithms and pipelines is not the focus of this review, it is the case that there is no perfect algorithm for any given type of somatic variant detection. In particular, because of the situation discussed above, when tumor cells are present in the skin biopsy used for the comparator normal genome, somatic variant detection in hematopoietic malignancies poses unique challenges. Due to the large numbers of algorithms available to identify point mutations, focused insertion/deletions and structural/copy number variants, a common approach is to identify the top 2–3 algorithms for each variant class (often these will use different algorithms to identify variants) and to combine their output at the first pass to ensure comprehensive discovery. Since this approach also assures a significant false positive rate, using a secondary validation step to cull false positives is necessary. We have found the best approach to whole genome variant validation is to have a custom set of capture probes synthesized for each genome set of variant sites, enabling hybrid capture from the whole genome libraries of tumor and normal so these sites can be sequenced, aligned and analyzed a second time. The following sections outline the hybrid capture approach and its fundamentals.

A secondary advantage of generating data from specific capture probes is the high coverage depth that results at the subsequently validated variant sites in the genome. In particular, our group^{10,11} and others^{12–14} have developed statistical approaches to estimating the genomic heterogeneity in a tumor sample based on sites of high coverage depth. In particular, since the majority of these sites are heterozygous, the read population will consist of reads carrying either the wild-type or mutant allele. Since each read originates from one fragment in the library, the one-to-one relationship predicts that the number of variant-carrying reads correlates to the variant allele fraction (VAF) in the tumor cell population. By this logic, the oldest variants in the population should be present in every cell and hence should represent 50% VAF. Newer variants are present in fewer cells as calculated by their VAF. In each case, VAF must be adjusted for tumor cell content of the sequenced cells and for variants that are found in copy number altered regions of the genome. A secondary clustering method then examines the VAF calculated for all variants and predicts the founder clone (50% VAF) and the additional subclones present by virtue of shared VAFs.

Exome Sequencing by Hybrid Capture

While whole genome sequencing offers a comprehensive ability to discover across the broad range of variant types, it is an expensive option and is certainly the most difficult approach from an analytical sense primarily due to the size and complexity of the human genome. While initially this was the only option for human genome sequencing, in 2008, several publications described the concept of “exome” sequencing^{15,16}, where the exome is defined as the ~1.5% of the human genome that contains annotated genes. These methods essentially involve solution-phase hybridization between biotinylated synthetic DNA or RNA probes that correspond to specific exon sequences from the genome, and the whole genome library fragments discussed earlier. Here, by virtue of sequence similarity, library fragments that contained exons or portions of exons would hybridize the sequence-specific probes in solution. Upon completion of the hybridization step, the library-probe hybrids were selectively precipitated from solution by the addition of streptavidin-coated magnetic beads

that bound the biotin conjugates on the probes, and were then pulled to the reaction tube bottom by the application of a magnet. Reversing the hybridization by addition of heat permitted the release of the captured fragments into solution. These selected library fragments could then be amplified by PCR, quantitated and sequenced. While the resulting sequencing data are aligned to the Human Genome Reference as the first step in analysis, the focused areas for variant discovery are specified by the territory these probes are designed to capture (defined by a BED format file). As such, the process of identifying exonic variants is more rapid, although fewer types of variants can be identified. There exist multiple commercial vendors of exome capture reagents that may offer significant additional coverage of regions around the exons of known genes, as well as including 5' and 3' untranslated regions, some known non-coding RNAs, and known promoter regions, to cite a few examples. A typical exome capture experiment aims for around 100-fold coverage depth on average, although coverage at any given locus can vary quite a bit from that average. Typically, exons with high G+C or A+T content are not captured adequately, and there can be overall variability from one hybrid capture to another, leading to false negatives when either tumor or normal or both aren't adequately covered to enable detection of variant bases.

Exome sequencing to compare tumor and normal has been done extensively across many tumor types, due to its low cost and ease of analysis compared to WGS. However, structural variants are not detected by this method as a general rule, and copy number analysis is inherently difficult because differential probe performance in hybridization can masquerade as reporting copy number differences. The cost to produce exome capture data is around 1/10th that of the whole genome sequencing process, when a commercial exome probe set is used. This low cost reflects increases in per run sequencing throughput, permitting libraries that carry a DNA barcode identifier to be pooled in an equimolar pool of 4–8 libraries, hybridized to the capture probe set and then sequenced together. Once sequencing is completed, specific software uses the information from the DNA barcode to separate reads that originated from different libraries into “bins”. Each bin of reads then is aligned to the Human Reference Genome, permitting each sample to be evaluated individually for variants.

As discussed for WGS, the depth of exome sequencing data can be exploited to calculate VAFs and these values subsequently can be clustered to estimate genomic heterogeneity in the sample, where higher depth will provide more confident VAF calculation. This can be limited in tumor types such as AML that have very low numbers of mutations in genes¹⁷ because there insufficient numbers of variants to generate confident clustering. Sequencing depth can be important for other reasons, especially as it relates to the capability to detect pathogenic mutations that may be present in only a small proportion of cells and may indicate an “Achilles Heel” that could prove responsive to one or more targeted therapies, or indeed may predict the tumor's capability to acquire resistance if a specific therapy is applied. As these types of mutations are better characterized, a set of probes can be included to enhance coverage at these sites which will maximize detection sensitivity. In addition, for areas where commercial exome kits are known to routinely under-perform in capture, or where a novel, non-exonic region merits variant detection, additional synthetic probes corresponding to these regions can be synthesized and “spiked” into the exome capture reagent to provide sequencing read coverage.

Targeted Sequencing by Hybrid Capture or Multiplex PCR

Even exome sequencing can provide a large amount of data for analysis, and often one knows exactly the genes that one wishes to target, hence hybrid capture can be used with a defined probe set. Typically, this is referred to as a targeted “panel” of genes and follows the same basic steps as exome capture. Probes can be designed with the help of a commercial manufacturer who can synthesize and biotinylate them, providing the resulting probes as an equimolar mixture or as individual probes to mix and match. Alternatively, capture probes can be produced by first designing PCR primers to amplify the regions of interest, and then biotinylating the PCR products by supplying biotinylated nucleotides into the PCR¹⁸. There is a lower limit to targeted capture; typically around 200kb of target space is minimally efficient. Lower than this, the yield of sequenced regions of interest relative to off-target reads becomes quite low and the efficiency of sequencing is compromised.

However, smaller gene numbers or only sequencing specific mutational “hotspots” are often desired, and another approach can be utilized at this sub-200kb sequencing scale. This approach, called multiplex PCR, attempts to design specific amplification primers for each site of interest such that the primers have approximately matched median annealing temperatures (T_m), but do not have sequence similarity across the primer set. In a multiplex PCR, all primers are combined with the genomic DNA of interest into a PCR and amplification occurs en masse by temperature cycling. At the completion of multiplex PCR, the resulting fragments are made into a library and sequenced by MPS. Subsequent alignment and variant analysis is as described for exome capture. There are several challenges in multiplex PCR including 1) the need to optimize primer pairs so that amplification is as robust as possible at all sites and primer-primer interactions are minimized, and 2) the need to design amplicons of approximately the same size range, so biases in amplification are minimized and sequencing coverage is maximized across the length of each product. The first challenge is made difficult when one wishes to study gene families or if exons across certain genes have high G+C or A+T content. The second challenge requires a relatively uniform PCR product size as determined by the sequencing read length, such that genes with long exons will have multiple PCR products to represent the entire exon sequence. Another challenge is in the detection of PCR biasing (jackpotting), discussed earlier. Essentially, it is impossible to use conventional NGS de-duplication algorithms to detect and remove duplicates of PCR products because they begin and end at the same sequences by definition--those provided by the primer pairs themselves. Hence, a secondary orthogonal validation scheme is a must, especially if high depth coverage is planned. As these challenges are realized, often what results will be multiple PCR primer sets that each have specific annealing and extension temperatures optimized for the loci being targeted. Prior to sequencing, the resulting PCR products are pooled to produce a broader target region of data than possible from a single multiplex reaction.

Transcriptome Sequencing

As interest in RNA assays increases, due to the realization that sequencing RNA from tumors provides a broad variety of data germane to our understanding of cancer biology, one also identifies many attendant challenges of “RNA-seq”. RNA is transcribed from the

genome across a broad range of sizes and while not all functions are known for all RNAs produced in the cell, there are emerging lines of evidence that non-coding RNAs of several classes are integral to cancer biology^{19–23}. Furthermore, RNA-based discovery includes cataloguing of expression levels, alternative splicing, mutation expression, fusion detection and others, and hence depends on the use of many different algorithms. As such, an RNA-seq analytical pipeline can be quite lengthy and complicated to assemble and to validate.

RNA is a labile molecule, so obtaining high quality RNA from clinical samples can often be difficult, especially due to the common pathology practice of formalin fixation and paraffin embedding wherein formalin crosslinks chemically with the nucleic acid backbone resulting in strand breaks and hence degradation. Similarly, paraffin at 65° F or higher temperatures will degrade RNA by hydrolysis during the embedding process.

Another difficult aspect of RNA sequencing from clinical samples is that limited quantities are typical, especially when derived from samples with low tumor content since laser capture microdissection or flow sorting must be used to enhance the tumor cell content into the isolation. Unlike DNA, increasing coverage on RNA sequencing data cannot compensate for low tumor cell content from a bulk isolate, so tumor cell enrichment steps must be performed prior to RNA isolation.

However, RNA amplification is straightforward and has high fidelity to the original expression levels of genes in the tissue, so input as low as 50 pg of total RNA can result in a workable yield of amplified total RNA. If polyA selection is desired as a means to solely focus sequencing on gene-encoding RNAs, the use of low input samples is obviated. Namely, polyA+ RNA constitutes about 2% of the total RNA population; hence, isolation of polyA+ RNA from low input samples will not yield a workable RNA amount for downstream library construction and sequencing.

Recently, we demonstrated that very low amounts (50 pg) of total RNA from fresh frozen or FFPE tissue can be sequenced to obtain improved on-gene coverage by use of an intermediate hybrid capture step that precedes sequencing²⁴. Such an approach can also be utilized in cases where a fusion gene transcript is suspect, or for diagnostic purposes (i.e. known fusion drivers in hematologic malignancies such as PML-RAR α , BCR-ABL, and others). In these cases, including probes into the hybrid capture reagent that tile along the exons of the fusion transcript loci, coupled with focused analysis to identify the fusions captured will yield fusion transcripts if present. Several groups now are exploring single cell RNA sequencing to examine the inter-cell differences in gene expression between cells from the same tissue^{25,26}. Careful analyses of these data to exclude results that are stochastic noise are required for single cell experiments²⁷.

As with DNA sequencing, the analysis of RNA-seq data initiates with alignment to the Human Reference Genome, but requires specific algorithms for this specific purpose and these may be fine-tuned further to permit specific types of RNA-based discovery^{28,29}. For example, detection of alternative splicing or of fusion transcripts is relatively difficult, yet is aided by specific algorithms that identify split read alignments, effectively permitting the sequencing read to be split across a gene or genomic segment³⁰.

Studying RNA by sequencing is informative for many reasons. First, over-expression of specific RNAs may be detected in tumors that are absent correlative signals in the DNA (such as a regional amplification, for example). Since over-expressed genes can be drivers of cancer development (HER2/neu, EGFR, and others), it is critical to collect these data because over-expression is not always simply a consequence of chromosomal amplification. Second, studying the expression of mutated genes identified from DNA analysis in the RNA-seq data can be informative regarding which mutated genes are also expressed in the tumor cells. On average around one-half of the genes mutated in a cancer genome are not expressed in the corresponding RNA population, either by selective silencing of the mutated allele, or due to lack of gene expression in its entirety. Third, although detecting structural variations in DNA has a high FPR, the combined evidence from DNA (by end read mapping) coupled with RNA fusion detection (from NGS) can confirm predicted fusions^{31–33}. Ultimately, as transcription factor and enhancer/repressor binding sites in the genome are better characterized, mutations in these regions may be correlated to RNA expression for the affected genes, representing yet another type of correlative analysis³⁴. Taken together, the integration of DNA and RNA sequencing data is providing very powerful insights into tumor biology.

Methylome and Epigenome Sequencing

Another type of DNA-based characterization that is becoming increasingly common is whole genome or targeted methylation sequencing (“methyl-seq”)^{35–38}. The addition of methyl groups to DNA is one regulatory mechanism utilized by the cell’s transcription machinery to indicate which genes are silenced and which are active. Unmethylated cytosines can be readily converted by treatment of the DNA with sodium bisulfite³⁹, providing a strategy to quantify methylated vs. unmethylated cytosine residues. These studies provide an integral data set with which to characterize cancer genomes. There are nuances to the types of cytosine methylation found on DNA, including differential methylation with 5-hydroxy methyl groups, formyl methyl groups, and perhaps others yet to be described. Efforts to chemically mark these different types of methyl groups, followed by sequencing are being developed as commercial kits^{40,41}.

The major challenges in methylation sequencing are two-fold; 1) the ideal depth of coverage remains somewhat elusive in terms of single CpG resolution⁴², especially for detecting minor methylation marks such as 5-hydroxymethylC residues, and 2) bioinformatic analysis of the resulting data to identify unmethylated and methylated nucleotides remains challenging in the context of a high FPR/FNR (where the latter is due to suboptimal coverage). One approach to address the cost of generating higher coverage is to down-sample the methylated genome by virtue of reduced representation bisulfite sequencing^{43,44}. More recently, commercial vendors have produced probes for targeted hybrid capture at known methylation sites. These are utilized following whole genome library construction and bisulfite conversion to select regions for methylation sequencing and allow more focused analysis.

As with transcriptome analysis, discussed earlier, the comparative changes in methylation between tumor and normal cell genomes are most highly informative but do require a

matched differentiation stage cell type of origin in the hematologic malignancies, to inform the comparison. Integration of the methylation changes identified in the tumor with coincident changes in tumor-specific gene expression can identify those methylation changes that truly impact the gene expression profile⁴⁵.

Methylation patterns are but one of the many regulatory aspects of cancer genomics that merit NGS-based comparison to normal progenitor cells⁴⁶. Indeed, cancer genomics discovery efforts are continuing to identify alterations to histones that package DNA and make it available (or not) for transcription. Like methylation, DNA accessibility data can be correlated to transcriptional activity, as can the presence or absence of bound transcription factors. While methylome studies are possible with the DNA isolated from banked tissues and blood samples, many of the aforementioned methods require cell lines or non-transformed cells growing actively in culture (when possible) to provide the input quantities of DNA required for the assays, or to perform the preparatory steps required to isolate protein-bound DNA. Examples of different methods for studying general DNA accessibility include; 1) genome-wide assays of DNaseI hypersensitivity to map chromatin accessibility which provides an unbiased general characterization of open chromatin⁴⁷, 2) formaldehyde treatment to produce DNA-protein crosslinks in nucleosome-bound chromatin regions, followed by isolation of non-crosslinked ("open chromatin") DNA for sequencing (FAIRE-seq)⁴⁸, 3) ATAC-seq, a transposase-based method that directly inserts sequencing adaptors for NGS into native chromatin⁴⁹.

More specific isolation of protein-bound DNA may be achieved by the combination of chromatin immunoprecipitation (ChIP) and high throughput sequencing (ChIP-seq). These assays require highly specific antibodies to the protein(s) of interest. Data from ChIP-seq experiments can provide insight into how, for example, mutations in epigenetic regulators affect epigenetic marks across the genome. ChIP approaches already have been informative toward understanding mechanisms of leukemogenesis in specific hematologic malignancies bearing mutations in *ASXL1*. In this study, *ASXL1* loss-of-function mutations resulted in a marked reduction of tri-methylated histone H3 lysine 27 (H3K27me3) occupancy by inhibiting polycomb repressive complex 2 (PRC2) recruitment to specific oncogenic target loci⁵⁰. Assays for epigenetic marks across the genome have an uncertain future in the clinic, but they are clearly informing our understanding of AML biology.

Conclusions

The development of methods to sequence genomes, transcriptomes and epigenomes, either in their entirety or in part, coupled with advanced analytical and integration approaches, has transformed our understanding of the genomic origins of leukemia. Although there remains much to be learned about basic discovery genomics in leukemia, there also are efforts to introduce these approaches into clinical care of leukemia patients, ultimately impacting their therapeutic options and outcomes in the disease. This chapter has presented basic concepts regarding the underlying molecular biology of general sequencing approaches for studying DNA, RNA and epigenetics. Equally important and challenging are the analytical approaches required to make sense of the sequencing data emerging from these massively

parallel instruments, and to integrate across data types to construct a more nuanced understanding of the genomic landscape and how it shapes tumor biology.

Acknowledgments

Supported in part by NIH/NHGRI 5U54HG003079

References

1. Rowley JD, et al. Further evidence for a non-random chromosomal abnormality in acute promyelocytic leukemia. *Int J Cancer*. 1977; 20:869–872. [PubMed: 271143]
2. Rowley JD. Chromosome abnormalities in the acute phase of CML. *Virchows Arch B Cell Pathol*. 1978; 29:57–63. [PubMed: 103312]
3. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977; 74:5463–5467. [PubMed: 271968]
4. Finishing the euchromatic sequence of the human genome. *Nature*. 2004; 431:931–945. nature03001 [pii]. 10.1038/nature03001 [PubMed: 15496913]
5. Mardis ER. A decade's perspective on DNA sequencing technology. *Nature*. 2011; 470:198–203. nature09796 [pii]. 10.1038/nature09796 [PubMed: 21307932]
6. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. *Cell*. 2013; 155:27–38. S0092-8674(13)01141-0 [pii]. 10.1016/j.cell.2013.09.006 [PubMed: 24074859]
7. Mardis ER. The \$1,000 genome, the \$100,000 analysis? *Genome Med*. 2010; 2:84. gm205 [pii]. 10.1186/gm205 [PubMed: 21114804]
8. Mardis, E.; McCombie, WR. *Molecular Cloning: A Laboratory Manual*. 4. Green, M.; Sambrook, J., editors. Cold Spring Harbor Laboratory Press; 2012. p. 735-892.
9. Chen K, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*. 2009; 6:677–681. nmeth.1363 [pii]. 10.1038/nmeth.1363 [PubMed: 19668202]
10. Ding L, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*. 2012; 481:506–510. nature10738 [pii]. 10.1038/nature10738 [PubMed: 22237025]
11. Walter MJ, et al. Clonal architecture of secondary acute myeloid leukemia. *N Engl J Med*. 2012; 366:1090–1098. 10.1056/NEJMoa1106968 [PubMed: 22417201]
12. Roth A, et al. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods*. 2014; 11:396–398. nmeth.2883 [pii]. 10.1038/nmeth.2883 [PubMed: 24633410]
13. Attolini CS, et al. A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proc Natl Acad Sci U S A*. 2010; 107:17604–17609. 1009117107 [pii]. 10.1073/pnas.1009117107 [PubMed: 20864632]
14. Nik-Zainal S, et al. The life history of 21 breast cancers. *Cell*. 2012; 149:994–1007. S0092-8674(12)00527-2 [pii]. 10.1016/j.cell.2012.04.023 [PubMed: 22608083]
15. Gnirke A, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*. 2009; 27:182–189. nbt.1523 [pii]. 10.1038/nbt.1523 [PubMed: 19182786]
16. Hodges E, et al. Genome-wide in situ exon capture for selective resequencing. *Nat Genet*. 2007; 39:1522–1527. ng.2007.42 [pii]. 10.1038/ng.2007.42 [PubMed: 17982454]
17. Welch JS, et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell*. 2012; 150:264–278. S0092-8674(12)00777-5 [pii]. 10.1016/j.cell.2012.06.023 [PubMed: 22817890]
18. Duncavage EJ, et al. Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. *J Mol Diagn*. 2011; 13:325–333. S1525-1578(11)00024-9 [pii]. 10.1016/j.jmoldx.2011.01.006 [PubMed: 21497292]
19. Di Leva G, Garofalo M, Croce CM. MicroRNAs in cancer. *Annu Rev Pathol*. 2014; 9:287–314. 10.1146/annurev-pathol-012513-104715 [PubMed: 24079833]

20. Tili E, Michaille JJ, Croce CM. MicroRNAs play a central role in molecular dysfunctions linking inflammation with cancer. *Immunol Rev.* 2013; 253:167–184.10.1111/imr.12050 [PubMed: 23550646]
21. Deng G, Sui G. Noncoding RNA in oncogenesis: a new era of identifying key players. *Int J Mol Sci.* 2013; 14:18319–18349. ijms140918319 [pii]. 10.3390/ijms140918319 [PubMed: 24013378]
22. Han BW, Chen YQ. Potential pathological and functional links between long noncoding RNAs and hematopoiesis. *Sci Signal.* 2013; 6:re5. scisignal.2004099 [pii]. 10.1126/scisignal.2004099 [PubMed: 23962981]
23. Young RS, Ponting CP. Identification and function of long non-coding RNAs. *Essays Biochem.* 2013; 54:113–126. bse0540113 [pii]. 10.1042/bse0540113 [PubMed: 23829531]
24. Cabanski CR, et al. cDNA Hybrid Capture Improves Transcriptome Analysis on Low-Input and Archived Samples. *J Mol Diagn.* 2014 S1525–1578(14)00072-5 [pii]. 10.1016/j.jmoldx.2014.03.004
25. Deng Q, Ramskold D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science.* 2014; 343:193–196. 343/6167/193 [pii]. 10.1126/science.1245316 [PubMed: 24408435]
26. Ramskold D, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol.* 2012; 30:777–782. nbt.2282 [pii]. 10.1038/nbt.2282 [PubMed: 22820318]
27. Brennecke P, et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods.* 2013; 10:1093–1095. nmeth.2645 [pii]. 10.1038/nmeth.2645 [PubMed: 24056876]
28. Mutz KO, Heikenbrinker A, Lonne M, Walter JG, Stahl F. Transcriptome analysis using next-generation sequencing. *Curr Opin Biotechnol.* 2013; 24:22–30. S0958-1669(12)00131-0 [pii]. 10.1016/j.copbio.2012.09.004 [PubMed: 23020966]
29. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013; 29:15–21. bts635 [pii]. 10.1093/bioinformatics/bts635 [PubMed: 23104886]
30. Feng H, Qin Z, Zhang X. Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. *Cancer Lett.* 2013; 340:179–191. S0304-3835(12)00657-X [pii]. 10.1016/j.canlet.2012.11.010 [PubMed: 23196057]
31. Ge H, et al. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics.* 2011; 27:1922–1928. btr310 [pii]. 10.1093/bioinformatics/btr310 [PubMed: 21593131]
32. Benelli M, et al. Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics.* 2012; 28:3232–3239. bts617 [pii]. 10.1093/bioinformatics/bts617 [PubMed: 23093608]
33. Supper J, et al. Detecting and visualizing gene fusions. *Methods.* 2013; 59:S24–28. S1046-2023(12)00253-8 [pii]. 10.1016/j.ymeth.2012.09.013 [PubMed: 23036331]
34. Khurana E, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science.* 2013; 342:1235587. 342/6154/1235587 [pii]. 10.1126/science.1235587 [PubMed: 24092746]
35. Lister R, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature.* 2009; 462:315–322. nature08514 [pii]. 10.1038/nature08514 [PubMed: 19829295]
36. Sandoval J, Esteller M. Cancer epigenomics: beyond genomics. *Curr Opin Genet Dev.* 2012; 22:50–55. S0959-437X(12)00019-6 [pii]. 10.1016/j.gde.2012.02.008 [PubMed: 22402447]
37. Krueger F, Kreck B, Franke A, Andrews SR. DNA methylome analysis using short bisulfite sequencing data. *Nat Methods.* 2012; 9:145–151. nmeth.1828 [pii]. 10.1038/nmeth.1828 [PubMed: 22290186]
38. Gao F, et al. Clustering of Cancer Cell Lines Using A Promoter-Targeted Liquid Hybridization Capture-Based Bisulfite Sequencing Approach. *Technol Cancer Res Treat.* 201410.7785/tcrt.2012.500416
39. Thomassin H, Oakeley EJ, Grange T. Identification of 5-methylcytosine in complex genomes. *Methods.* 1999; 19:465–475. S1046-2023(99)90883-6 [pii]. 10.1006/meth.1999.0883 [PubMed: 10579942]

40. Booth MJ, et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*. 2012; 336:934–937. science.1220671 [pii]. 10.1126/science.1220671 [PubMed: 22539555]
41. Booth MJ, et al. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat Protoc*. 2013; 8:1841–1851. nprot.2013.115 [pii]. 10.1038/nprot.2013.115 [PubMed: 24008380]
42. Harris RA, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol*. 2010; 28:1097–1105. nbt.1682 [pii]. 10.1038/nbt.1682 [PubMed: 20852635]
43. Gu H, et al. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc*. 2011; 6:468–481. nprot.2010.190 [pii]. 10.1038/nprot.2010.190 [PubMed: 21412275]
44. Gertz J, et al. Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLoS Genet*. 2011; 7:e1002228. PGENETICS-D-11-00600 [pii]. 10.1371/journal.pgen.1002228 [PubMed: 21852959]
45. Liu Z, et al. Integrated Analysis of DNA Methylation and RNA Transcriptome during In Vitro Differentiation of Human Pluripotent Stem Cells into Retinal Pigment Epithelial Cells. *PLoS One*. 2014; 9:e91416. PONE-D-13-38313 [pii]. 10.1371/journal.pone.0091416 [PubMed: 24638073]
46. Ryan RJ, Bernstein BE. Molecular biology. Genetic events that shape the cancer epigenome. *Science*. 2012; 336:1513–1514. 336/6088/1513 [pii]. 10.1126/science.1223730 [PubMed: 22723401]
47. John S, et al. Genome-scale mapping of DNase I hypersensitivity. *Curr Protoc Mol Biol*. 2013; Chapter 27(Unit 21):27.10.1002/0471142727.mb2127s103 [PubMed: 23821440]
48. Simon JM, Giresi PG, Davis IJ, Lieb JD. A detailed protocol for formaldehyde-assisted isolation of regulatory elements (FAIRE). *Curr Protoc Mol Biol*. 2013; Chapter 21(Unit21): 26.10.1002/0471142727.mb2126s102 [PubMed: 23547014]
49. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. 2013; 10:1213–1218. nmeth.2688 [pii]. 10.1038/nmeth.2688 [PubMed: 24097267]
50. Abdel-Wahab O, et al. ASXL1 mutations promote myeloid transformation through loss of PRC2-mediated gene repression. *Cancer Cell*. 2012; 22:180–193. S1535-6108(12)00303-0 [pii]. 10.1016/j.ccr.2012.06.032 [PubMed: 22897849]

Figure 1. Overview of the whole genome sequencing (WGS) and analysis workflow for paired tumor and normal samples

(a) Sequencing data representing paired end reads from next generation sequencing libraries of DNA isolated from tumor and normal comparator cells are first aligned separately to the Human Reference Genome. Different types of variants are detected with different algorithms, depending upon the variant types. Variants are detected in the tumor and normal genomes, and then compared to one another, enabling a determination of truly somatic variants (i.e. tumor-unique). Finally, as shown on the right of the figure, a Circos plot diagram can be used to comprehensively represent the somatic variant landscape of each tumor. (b) A detailed series of steps is followed to make a massively parallel sequencing library, beginning with a genomic DNA sample. High molecular weight DNA extracted from human tissues can be fragmented by sonication using high frequency sound waves. Following the fragmentation, the ends are repaired to blunt them using a combined enzymatic fill-in and removal of overhanging ends. The blunted ends are phosphorylated, enabling the addition of adenosine at the ends, and providing a hybridization and ligation point for sequencing platform-specific adapters. The adapters are synthetic DNAs that

correspond to covalently linked adapters on the support surface used for an enzymatic amplification step prior to sequencing. Several library construction steps, marked by blue stars, require the polymerase chain reaction (PCR) to amplify the products prior to the next process step, which can introduce bias in representation of library content.