Quantitative Biology 2020, 8(2): 95–108
https://doi.org/10.1007/s40484-020-0203-8

REVIEW

Applications of probability and statistics in cancer genomics

Xiaotu Ma*, Sasi Arunachalam, Yanling Liu

Department of Computational Biology, and Cancer Biology Program, Comprehensive Cancer Center, St Jude Children's Research Hospital, Memphis, TN 38105, USA
* Correspondence: xiaotu.ma@stjude.org

Received November 26, 2019; Revised December 3, 2019; Accepted December 4, 2019

***Background*: The past decade has witnessed a rapid progress in our understanding of the genetics of cancer and its progression. Probabilistic and statistical modeling played a pivotal role in the discovery of general patterns from cancer genomics datasets and continue to be of central importance for personalized medicine.**

***Results*: In this review we introduce cancer genomics from a probabilistic and statistical perspective. We start from (1) functional classification of genes into oncogenes and tumor suppressor genes, then (2) demonstrate the importance of comprehensive analysis of different mutation types for individual cancer genomes, followed by (3) tumor purity analysis, which in turn leads to (4) the concept of ploidy and clonality, that is next connected to (5) tumor evolution under treatment pressure, which yields insights into cancer drug resistance. We also discuss future challenges including the non-coding genomic regions, integrative analysis of genomics and epigenomics, as well as early cancer detection.**

***Conclusion*: We believe probabilistic and statistical modeling will continue to play important roles for novel discoveries in the field of cancer genomics and personalized medicine.**

Keywords: cancer genomics; sequence analysis; probability and statistics

Author summary: With the rapid technology development and extensive research efforts in past decade, genomics approaches are playing an increasingly important role in human health problems such as cancer. A significant challenge in this endeavor is the analytical complexity associated with its big data nature that requires talents from scientists with quantitative background. In this review we aim to provide an introduction of genomic analyses by accounting for essential biological concepts, as well as exciting new frontiers for novel discoveries.

INTRODUCTION

It has long been appreciated that somatic genetic aberrations characterize cancer cells. For example, microscopic observation with chromatin staining reveals dramatic differences between cancerous and normal cells [1]. Such karyotyping techniques have led to the discovery of Philadelphia chromosome in leukemia [2], for which current well-known targeted therapies have been developed (imatinib, also known as Gleevec). With the revolutionary Sanger sequencing technique [3], many genes responsible for cancer, including *TP53*, *CDKN2A*, and *RB1* [4], were discovered. Due to the low throughput nature of Sanger sequencing, it is essentially impossible to perform a systematic study of cancers at the genome level with this technique. Recent technological advancements, including microarray technology in the 2000s and next generation sequencing (NGS) technology in the 2010s, have enabled the systematic delineation of genetic defects in tumors from a large number of patients [5,6]. As a result, our understanding of genetics in most cancer types, including both childhood [7,8] and adult [9–11] malignancies, has exponentially expanded in the past decade. Indeed, the research successes of NGS technology have prompted enthusiasm for its eventual use for clinical applications [12], including novel diagnostic

applications such as liquid biopsy [13] which can enable early detection [14]. However, the complex nature of the cancer biology and the associated genome-wide "big data" involves numerous quantitative challenges in order to correctly analyze and interpret the data. In this review we introduce cancer genomics from a statistical perspective, hoping it can help scientists with a quantitative background to contribute on cancer genomics problems.

MUTATION

Life is encoded by genetic information found in DNA or RNA. Such information needs to be handed down to the next generation during reproduction and to be copied from mother cells to daughter cells during development in multi-cellular organisms such as humans. As with the phrase "nothing is perfect", the copying of genetic information (through polymerases) is error-prone, albeit at a low rate of ~10^{-9} per cellular division [15], due to effective molecular mechanisms repairing the errors [16]. In this sense, every individual is genetically a "mosaic" [17] from the time of the first cell division post fertilization, because ~2 mutations are expected between the mother and daughter cells. Collectively at least 7×10^{13} mutations could have happened by assuming 3.7×10^{13} cells in a human body [18]. Although most mutations are harmless, some mutations can result in non-cancerous health problems such as Huntington disease [19], as well as cancers to be discussed in this paper. Harmful mutations can be either somatic (acquired post fertilization) or inherited (called germline mutation) from parents — the latter results in elevated predisposition for certain diseases [20].

Linear DNA sequences are composed of chemically basic compounds (called bases or base pairs) including adenine (A), cytosine (C), guanine (G), and thymine (T), which are modeled as text strings in this paper. Genetic errors can be roughly categorized into single nucleotide variants (SNV; similar to single letter misspellings in text editing, Fig. 1A), insertions and deletions (indels; similar to inserted letters or missing letters, Fig. 1B), structural variations (SV; similar to swapping of paragraphs, Fig. 1C) and copy number variations (CNV; addition or loss of large paragraphs, Fig. 1D). The enormous number of possible mutations renders the cancer genomics studies computationally intensive. For example, by denoting the length of human genome as $N=3.1\times10^9$ base pairs, the number of possible SNVs would be $3N\approx10^{10}$ since each base can be converted to three other possible bases, the number of possible small indels (for simplicity, less than 20 bp) would be $2\times N\times\sum_{i=1\ldots20}4^i\approx10^{22}$, while the number of possible structural variations would be $N^2\approx10^{19}$.

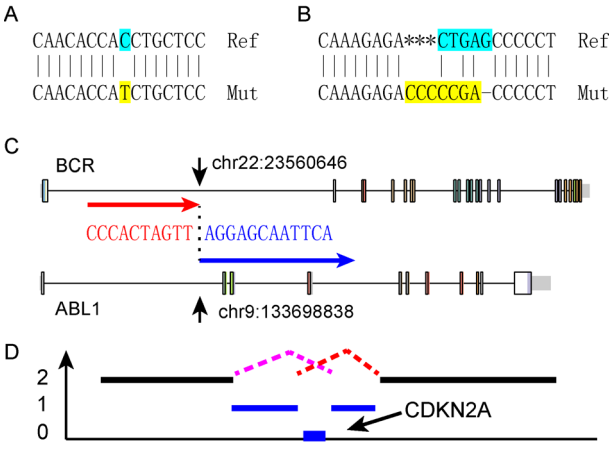


Figure 1. Mutation types. (A) Single nucleotide variant (SNV), where reference allele C (cyan) was mutated to mutant allele T (yellow). Vertical dashes indicate match. (B) Small insertion/deletion (indel). Insertions are indicated by asterisks (*) while deletions are indicated by dash (-). (C) Structural variant (SV). Illustrated is an example BCR-ABL fusion (Philadelphia chromosome), where DNA part of BCR (red arrow) is fused to ABL1 (blue arrow) by somatic mutation. The chimeric DNA sequence from tumor demonstrate both BCR (red) and ABL1 (blue) components. Exact break points (hg19) are indicated by black arrows. Boxes in BCR and ABL1 indicate exon/intron structure. (D) Copy number variant (CNV). Shown is a cartoon of chromosome 9p with CDKN2A homozygous deletion, caused by two distinct structural variants (purple and red dashed lines connecting the breakpoints). As a result, this region demonstrates copy number states of 2, 1, and 0 (*y*-axis).

TUMOR SUPPRESSOR GENE AND ONCOGENE

Genes such as *TP53*, *CDKN2A*, and *RB1* are known to functionally drive cancer development [4]. These genes are broadly categorized into tumor suppressor genes (TSG) and oncogenes, based on their functional role in tumorigenesis. These concepts have been concisely defined by Bert Vogelstein as follows [21]: a tumor suppressor gene is a gene that, when inactivated by mutation, increases the selective growth advantage of the cell in which it resides; while an oncogene is a gene that, when activated by mutation, increases the selective growth advantage of the cell in which it resides. A critical component of this definition is natural selection as stressed by "growth advantage", because cancer in effect is the outgrowth of a subpopulation of cells. Here we informatically illustrate this concept by using the central dogma of molecular biology, which indicates that DNA information is transferred to RNA and then to protein. As shown in Fig. 2A, a eukaryotic gene can be disrupted by numerous mutation classes, such as promoter loss,

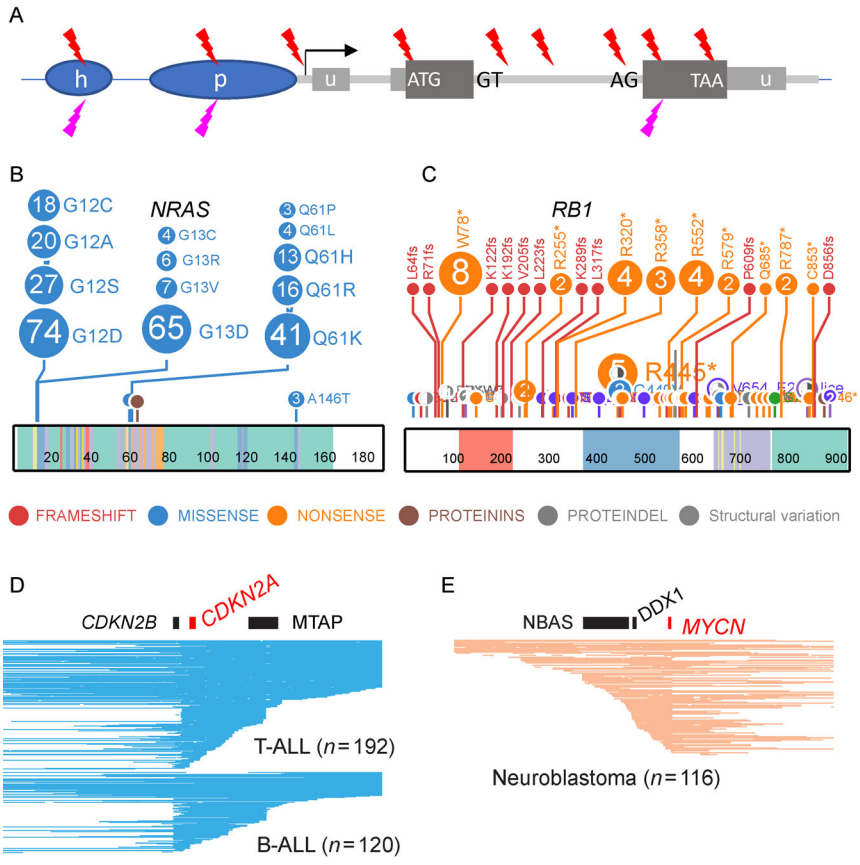


Figure 2. Cancer driver genes. (A) From central dogma perspective, a gene can be disrupted by many different mutations throughout the gene body (thunder bolts), including enhancer (h), promoter (p), untranslated region (u), translation start site (ATG), coding exons, splicing donor (GT) and acceptor (AG), translation stop site (TAA/TAG/TGA). Tumor suppressor genes typically have a diverse pattern of mutation types that mostly result in loss of function. On the other hand, function of a gene can be enhanced by much limited number of mutations (purple thunder bolts), mostly in coding exons, or enhancer (h) and promoter (p) to increase the expression strength, and this pattern (gain of function) is associated with oncogenes. Also shown are example oncogene and tumor suppressor gene, including *NRAS* (B), *RB1* (C), *CDKN2A* (D), and *MYCN* (E), where the number of tumor specimens harboring different mutations is illustrated in circles. For example, 74 specimens have missense mutations changing 12th amino acid (glycine, G) of NRAS to amino acid (glutamine, D). In *RB1*, most mutations either lead to frameshift or early translation stop codon (denoted by *) and thereby disrupting *RB1*. Copy number loss is indicated by blue lines, while copy number gain is indicated by red lines, with number of samples indicated. Data obtained (accessed Nov 1, 2019) from https://pecan.stjude.cloud/.

mutation in protein-coding exons, disruption of critical transcriptional and translational signals including translation start codon (ATG), splicing donor site (GT) and splicing acceptor site (AG) mutations, mutations leading to translation stop codons (TAG/TAA/TGA), or simply deletion of the whole gene or selected exons. On the other hand, a gene can be activated, or enhanced, by very limited number of mutations, such as stronger promoter by new DNA binding motifs of a trans-regulatory transcription factor, mutations in certain amino acids leadings to enzymatic activation, or extension of protein coding exons. As a result of natural selection, if a gene can act as tumor driver, it will exhibit elevated mutational frequency above the background mutational frequency in a large cohort of tumor samples — a fact utilized by many driver gene discovery algorithms [22–24]. A candidate driver can then be classified as TSG or oncogene based on its mutational patterns. For example, NRAS has clustered mutations in amino acids 12 and 13 (both glycine) and 61 (glutamine) (Fig. 2B), thereby demonstrating limited means of activation, and can be classified as an oncogene. By contrast, *RB1* has mostly mutations leading to stop codons and frameshift mutations roughly evenly distributed along the gene body (Fig. 2C), thereby demonstrating diverse means of gene disruption, and can be classified as tumor suppressor gene. Because a DNA sequence has three possible open reading frames, an indel can generally lead to a frameshift unless it has length that is a multiple of 3 and indels are therefore frequently found in tumor suppressor genes. Conversely, some genes

can be frequently disrupted by copy number loss (such as *CDKN2A*, Fig. 2D) or copy number gain (*MYCN*, Fig. 2E).

On the other hand, a protein-coding gene may have a variety of distinct functions. As a result, it can be challenging to classify a gene to be TSG or oncogene by the mutational patterns, such as *TP53* [25]. In this case, functional studies are needed for the correct classification. Mutations not deemed to be cancer drivers, such as silent mutations, can be tentatively categorized as passenger mutations.

The above observations highlight the need for comprehensive analysis of the cancer genome, which leads to considerations regarding the desired NGS platform (*e.g.*, whole genome versus whole exome sequencing). Although whole-exome sequencing can save 98% of the sequencing cost by focusing on ~2% [26] of the gene coding regions, it is not powered to detect copy number and structural variations. For example, our comprehensive analysis of 1,699 pediatric cancers indicated that 62% of driver events in pediatric cancers are due to copy number or structural variations, highlighting the need of comprehensive analysis of the tumor genome, preferably by whole genome sequencing [7].

The copy number variation profiles of *CDKN2A* and *MYCN* are interesting examples for illustrating the complexity of driver gene discovery. For example, the gene *MTAP* is so close to *CDKN2A* that it is frequently co-deleted with *CDKN2A*, and any frequency enrichment-based statistical detection methods [22–24] would detect *CDKN2A* and *MTAP* simultaneously. This is also true for *NBAS* in the vicinity of *MYCN*. Additional biological considerations are clearly needed to rule out these genes as cancer drivers. Similar complexity exists for other mutation types. For example, earlier studies [27] reported *CSMD3* as cancer drivers based on its over-represented mutation frequency. However, later studies found that the mutation frequency of *CSMD3* is not significant after its length was accounted for [28]. Additional confounding factors for driver gene discovery include transcription-coupled repair [28], tumor level hypermutation due to DNA repair deficiency [29], and local (*i.e.*, in certain chromosomal regions of a given tumor) hypermutation known as kataegis [30].

POWER CONSIDERATIONS

Because cellular growth is controlled by a plethora of molecular pathways, each involving multiple genes, it is rare that a single gene would dominate the causal list. Rather, multiple driver genes are typically associated with a given cancer type. In addition, because different signaling pathways and transcriptional programs are active in different tissues/organs, cancer driver genes typically demonstrate a strong cancer-type dependency. For example, our pan-cancer analysis of childhood malignancies indicated that *CDKN2A* loss is enriched in pediatric T-lineage and B-lineage leukemias, while *NOTCH1* mutations are found almost exclusively in T-lineage leukemias [7]. Typically, many genes are drivers in a given cancer type, and most of the genes have very low mutational frequencies (Fig. 3A).

Clearly, when designing a study cohort, power calculations must be performed to understand what cohort

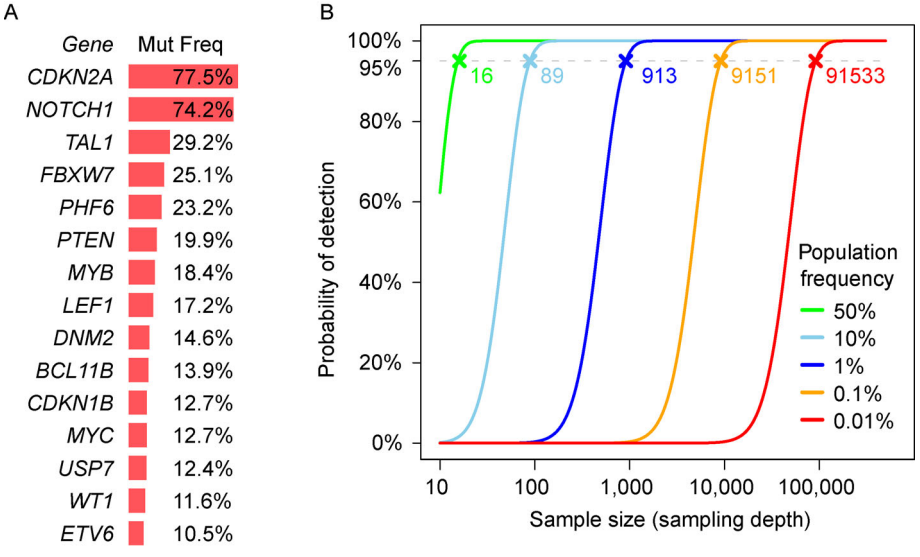


Figure 3. Mutation frequency of drivers and power consideration for gene discovery. (A) Long tail nature of top 15 cancer driver genes in pediatric T-lineage leukemia, ordered by mutation frequency. (B) Binomial probability of detecting events (> 5 recurrences) with respect to sample size (or sampling depth) at population frequencies ranging from 50% to 0.01%. Minimum sample sizes required to achieve 95% probability of detection are indicated with "x".

size is needed to discover driver genes with a predefined mutation frequency. Another similar numeric consideration in cancer genomics is the design of sequencing depth. Although sequencing costs have dramatically decreased in recent years, a balance between sequencing depth, targeted mutant allele fraction (further discussed later) to detect, platform (whole genome, whole exome or gene panel, transcriptome), and cohort size still need to be carefully considered. By assuming a predefined population frequency of the subject of interest (a driver gene, or a given mutation), a binomial distribution can be used to calculate the chance of discovery, as shown in Fig. 3B. For example, a cohort size of 16 would ensure 95% chance of detecting an event (minimum recurrence of 5) when the population frequency is 50%. However, for events with lower population frequency, such as 0.01%, a cohort size of 91,533 is needed to ensure 95% chance of detection. This calculation also applies to determination of sequencing depth as discussed later.

TUMOR PURITY AND TUMOR-IN-NORMAL CONTAMINATION

To discover cancer driver (as well as passenger) mutations from a patient's tumor, DNA sequencing data from a given tumor must be compared with that from a normal tissue sample of the same patient, typically known as tumor-normal matched sequencing. This is because every individual generally inherits 4–5 million polymorphisms (*i.e.*, DNA differences at the population level) from parents, and a significant proportion of these polymorphisms have a low population frequency so that it is difficult to comprehensively catalogue them in current polymorphism databases such as dbSNP [31]. Due to the low somatic mutation rate (mostly <5 mutations per million base pairs per tumor) [23], undocumented inherited polymorphisms can easily mask the signal of true somatic mutations if a normal control sample is not sequenced. Matched tumor-normal paired sequencing is therefore the most effective study design to delineate the somatic mutational landscape for cancer patients.

However, the complex nature of cancer biology renders the analysis of paired sequencing data highly challenging. For example, biopsy methods have no guarantee of acquiring pure cancer cells from the tumor (Fig. 4A). As a result, it is not uncommon to see a tumor sample containing a considerable fraction of normal cells, which results in low tumor purity (also known as cancer cell fraction, CCF). This issue is especially concerning when studying relapsed leukemias, due to the intensive monitoring of disease status that can result in early detection of tumor regrowth. In general, a tumor with reduced tumor purity can still be reasonably studied by increased sequencing depth.

On the other hand, it can also be challenging to obtain a normal sample lacking cancer cells (Fig. 4A). For

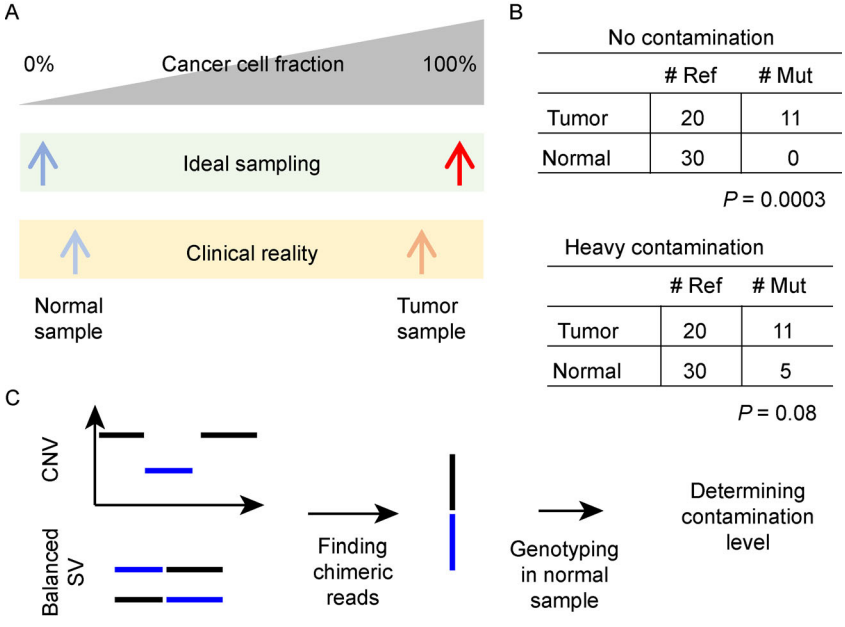


Figure 4. Tumor purity and tumor-in-normal contamination. (A) Statistical model. A biopsy sample can have tumor purity ranging from 0% to 100%. An ideal scenario is to have 0% CCF in the normal sample and 100% CCF in the tumor sample. However, in practice the normal sample can have small fraction of cancer cells and the tumor sample may contain normal cells as well. (B) Statistical detection of true somatic mutation can be compromised by tumor-in-normal-contamination. (C) Assessing potential tumor-in-normal contamination by genotyping chimeric reads (where DNA sequence from one locus (black lines) is fused to another locus (blue lines)) in normal sample ascertained from copy number variation or structural variation.

leukemia, a common practice is to use a biopsy sample acquired at the end of induction treatment (day 28/29), known as a remission sample, as a normal control. However, complete remission is not always achievable, such as in refractory cases. The tumor burden of remission samples might be significant even if there is morphologically complete remission. The proposal of using a skin biopsy as a normal control cannot guarantee a lack of cancer cells because of blood vessels in the skin. For solid tumors, the frequently used tumor-adjacent tissues might contain cancer cells due to localized infiltration [32], while distant tissues might contain metastatic cancer cells. Further, blood from solid tumor patients might contain circulating tumor cells [33], especially in patients with metastatic tumors, rendering this option also sub-optimal. Nail biopsy has been proposed recently as a highly promising normal control although its clinical utility remains to be established [34]. Clearly, detection power can be severely compromised in a tumor-normal matched sequencing design when the tumor-in-normal contamination level is substantial (Fig. 4B), and analytical caveat must be noted to avoid under-diagnosis. However, this is a particularly challenging task to achieve, as robust estimation of tumor-in-normal contamination in normal samples relies on detected somatic mutations, and our ability of detecting somatic markers is reduced in contamination scenarios — a logical deadlock.

There are a few helpful practices to investigate potential tumor-in-normal contamination. First, most tumors harbor clonal copy number alterations, such as *CDKN2A* deletions in T-lineage ALL, as mentioned earlier. Such copy number alterations are usually associated with structural variations, and the presence of chimeric reads spanning the breakpoint in normal samples would conclusively indicate tumor-in-normal contamination as such alterations are virtually non-existent in the normal population (Fig. 4C). Clearly, this strategy does not apply to tumors without clonal copy number alterations, such as cytogenetically normal acute myeloid leukemia (AML) [35]. In such cases, it is frequent that the tumor is driven by chimeric fusions — in this case balanced translocations. Therefore, tumor-only analysis of chimeric fusions followed by detection of the corresponding chimeric fusion in the matched normal sample is often effective in detecting contamination (Fig. 4C). When these routes fail, interaction with the laboratory is needed to rule out mistakes in sample handling. Previously unidentified mechanisms, such as epigenetics, might be considered as well.

TUMOR PURITY, PLOIDY AND MUTANT ALLELE FRACTION

Due to the low somatic mutation rate [23], most mutations are heterozygous, *i.e.*, only one of the two parental alleles being mutated. The mutant allele fraction (MAF), defined as the number of mutant reads divided by the total number of reads at a mutated locus, of such (clonal, discussed later) mutations are theoretically one half of tumor purity (denoted as π): $MAF = \pi/2$. However, both point mutations (*e.g.*, SNV and indel) and interval mutations (*e.g.*, CNV) can happen in a given tumor, and it is frequent that these different types of mutations can collectively affect the same genomic regions (Fig. 5A). Because multiple copy number alteration events can repeatedly affect a given genomic region, a complex relationship between MAF, tumor purity π, and copy number status can be illustrated with mathematical formulation (Fig. 5B, C). For example, in the case of copy neutral loss of heterozygosity (CN-LOH), $MAF = \pi$, while a non-linear relationship exists for 1-copy loss (Fig. 5D). Such complex dependency has a profound effect in mutation detection and clonal evolution analysis detailed later.

Generally, mutant allele fraction (MAF) is a good indicator of timing of the corresponding events. To illustrate this point, our analysis of published [36] somatic mutation data in a melanoma cell line (COLO829) on chr1q (with all four copies from the same parent) revealed three distinct groups of mutations (Fig. 6A, B): Early group (that with allele fraction 100%); Middle group (that with allele fraction 50%); Late group (that with allele fraction 25%). This data clearly indicated three distinct periods of mutational events in this tumor (Fig. 6C): 1) generation of a single copy of chr1q with "Early" mutations, followed by doubling of this chromosome; 2) generation of "Middle" mutations on each of the two copies, followed by another doubling; 3) generation of "Late" mutations. As a result, the number of mutations in each group can be used to estimate the relative duration of each period as following: Early:Middle:Late = 492:205.5:15.5 = 32:13:1.

INTRA-TUMOR HETEROGENEITY

Multiple driver mutations can be acquired in a given tumor, frequently at different time points and therefore in different population of cancer cells. Within such a tumor, genetically homogeneous cell population can be defined as a "clone". Most tumors are multiclonal, and clones with late mutations are termed "subclonal" to clones with early mutations, which can lead to a hierarchical cellular lineage graph defined by mutations and their timing (Fig. 7A). Due to difficulty in ascertaining the timing of mutations, operationally mutations are generally classified into clonal and subclonal based on mutant allele frequencies. For example, clone 3 is subclonal to clone 2. It should be noted that such an operational definition may

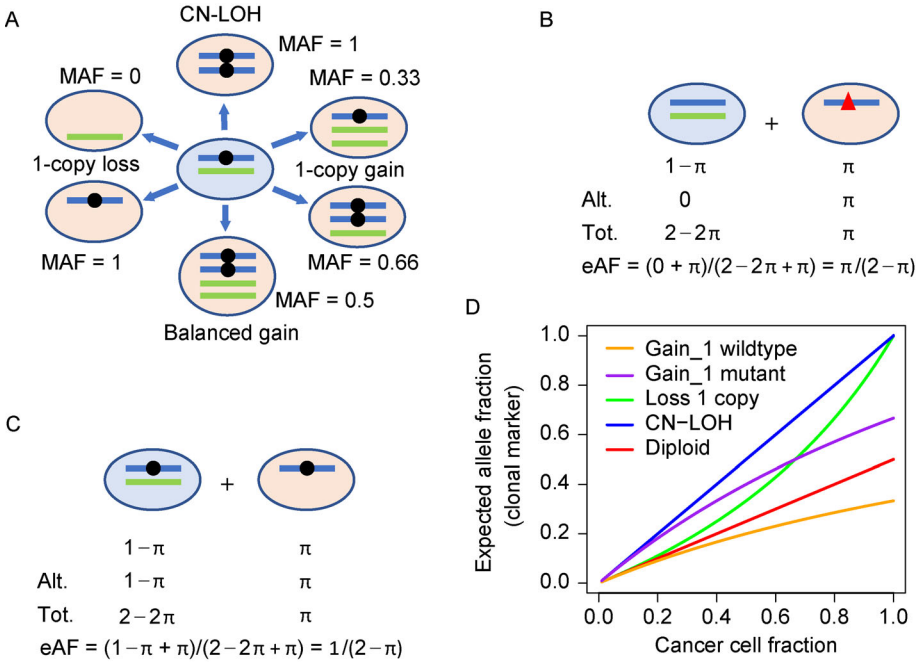


Figure 5. A mathematical model of ploidy, clonality, and tumor purity. (A) Example scenarios of copy gain/loss (differentiating paternal (blue) or maternal (green)). Mutant allele fraction (MAF) of the marker (solid dot) is illustrated for each scenario. (B) Example showing calculation of expected allele fraction (eAF) of somatic mutation (red star) from a tumor with purity π. (C) Example showing calculation of expected allele fraction of germline polymorphism (black dot) from a tumor with purity π. (D) Expected allele fraction of somatic mutation as a function of tumor purity and selected local ploidy. Gain_1 wildtype: the mutation allele has 1 copy while the wildtype has two copies; Gain_1 mutant: the mutant allele has two copies while the wildtype has 1 copy; Loss 1 copy: only 1 mutant copy exists; CN-LOH: only 2 mutant copies exist; Diploid: 1 mutant and 1 wildtype copy.

sometime generate misleading results, because a constant growth rate is implicitly assumed for all subclones. For example, clone 4 might be misclassified as subclonal to clone 2 by allele frequency analysis. Single cell profiling of mutations can be used to resolve such difficulties. In addition, the "subclonal" relationships inferred from allele frequency analysis may not necessarily imply the hierarchical structure as illustrated in Fig. 7A. For example, the Early, Middle, and Late mutations in Fig. 6 are present in every cancer cell of COLO829 cell line. Their allele frequency difference is a result of mutational timing and copy number changes instead of relative cellular compositions as illustrated in Fig. 7A.

Inference of the clonal structure is of great interest in understanding mechanisms governing tumorigenesis. This is typically achieved through clustering analysis of MAFs (Fig. 6B). Notably, statistical models such as binomial mixture modeling have turned out to be highly effective for such clustering analyses, especially for higher dimension data [29]. Because of the high selection pressure exerted by therapy (*e.g.*, surgery and chemo), clonal composition of the diagnosis tumor generally undergoes dramatic changes. Therefore, comparison of clonal composition between time points offers us great potential to learn important insights on the efficacy of treatment and how cancer cells evolve under selection pressure. Indeed, comparison of whole exome sequencing datasets between relapse and diagnosis tumors from twenty relapsed ALL patients [29] has led to common rise and fall clonal evolution patterns of ALL relapse (Fig. 7B). First, predominant subclones at diagnosis are generally cleared, indicating high chemotherapeutic efficacy. Second, the relapsed tumors were generally seeded from a minor subclone (often associated with RAS mutations) detectable at diagnosis, indicating a potentially shared molecular mechanism of cancer cells to survive treatment. Third, relapsed tumors were found to share mutations in *NT5C2*, indicating a common molecular mechanism of drug resistance. This observation has been reinforced by our recent analysis of serial samples from relapsed ALL [37].

In addition to temporal analysis of clonal evolution, spatial intra-tumor heterogeneity is expected in solid tumors. Indeed, our analysis of multiple biopsies from osteosarcoma patients revealed clear evidence on cross-seeding between different biopsy sites [38]. Collectively, pilot spatiotemporal analyses of tumors have led to exciting new insights on the evolution of cancer cells,

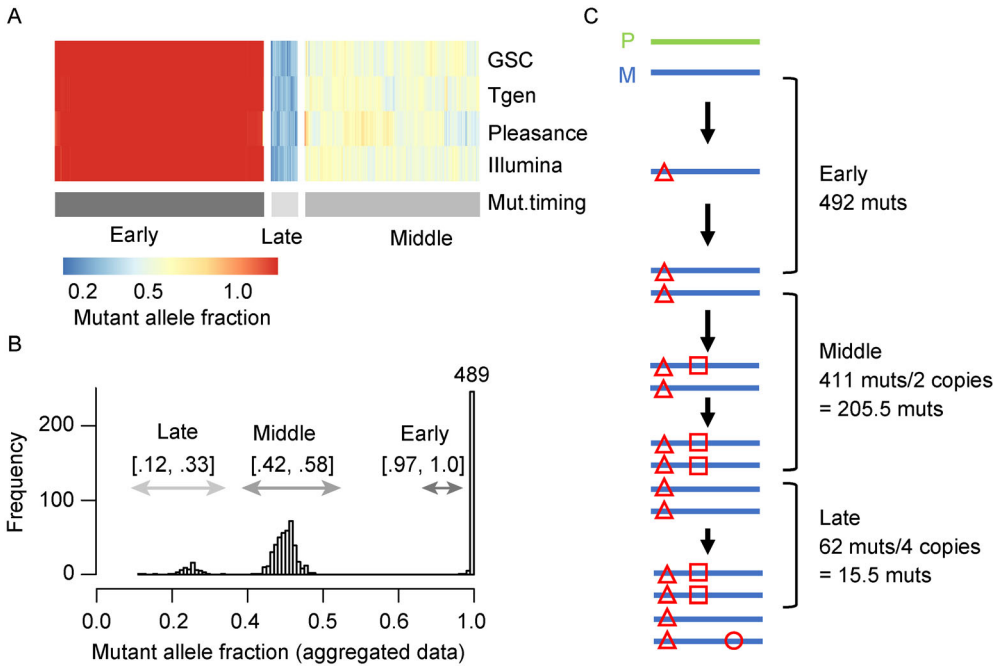


Figure 6. Inferring mutational timing in melanoma cell line COLO829. In this cell line, it is known that chr1q (from 150M to 249M) has four copies from the same parental origin due to loss of heterozygosity and re-duplication. (A) Mutant allele fraction of somatic mutations in this region formed three clusters consistently supported by four sequencing efforts (GSC, Tgen, Pleasance, and Illumina): Early, Middle, and Late, which are defined by aggregating read counts from all four sequencing efforts (B). (C) Proposed model of mutational timing during tumorigenesis of COLO829. One of the two parental alleles of 1q was lost (for simplicity, assuming paternal allele was lost), and the remaining allele acquired 492 mutations (triangle) before its first duplication. Afterwards, 411 mutations (square) were acquired in one of the two copies of 1q before the second duplication event, following which additional 62 mutations (circle) were acquired in one of the four copies of 1q before diagnosis.

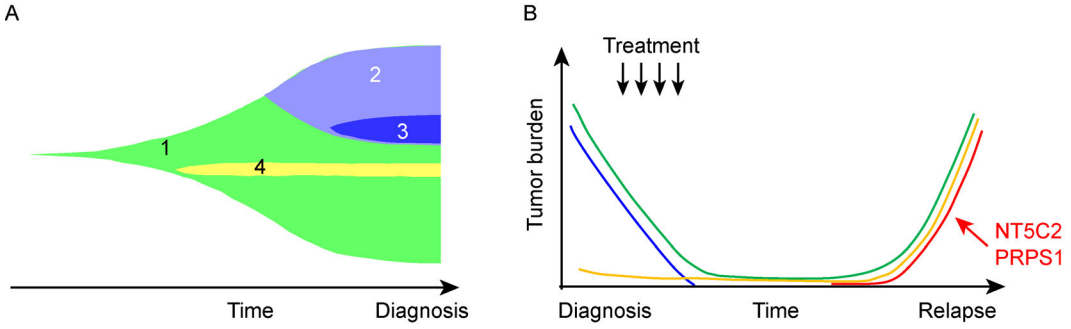


Figure 7. Intra-tumor Heterogeneity. (A) Modeling clonal evolution. The population size of a given subclone is a function of its growth rate and its age. For example, subclone #4 has a small diagnosis fraction due to lower growth rate than subclone #2 and #3. (B) Common clonal evolution patterns in relapsed ALL. Upon diagnosis, intensive treatment will be applied so that the overall tumor burden will reduce. Predominant diagnosis subclone (blue) was eradicated by treatment; while a minor subclone (orange) that also carries the ancestral mutations (green) can survive treatment and seed relapse. New mutations (red), such as *NT5C2* and *PRPS1*, can be acquired and lead to drug resistance.

sometimes with consideration of the therapeutic pressure. Further comprehensive spatiotemporal analyses are needed to develop better treatment strategies for improved cure rates.

FUTURE DIRECTIONS

Early detection of cancers

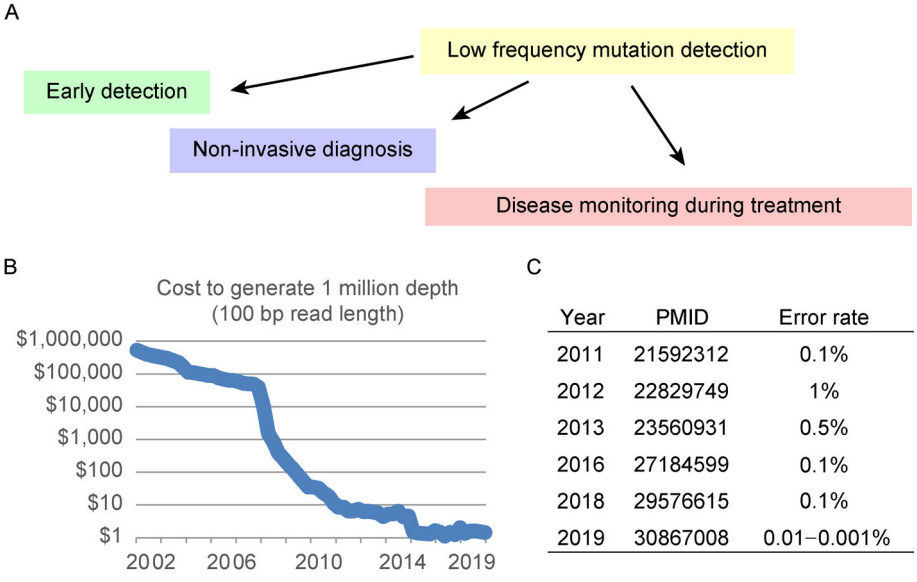
With the dramatic increase of our knowledge on genomic aberrations in many cancer types, there is great interest in the early detection of cancers, either before symptom or during treatment to monitor relapse (Fig. 8A). Indeed, our temporal study of drug resistance mutations in *PRPS1/NT5C2* indicated the feasibility of early detection of relapsed leukemia [37,39], which may inform alternative treatment options when drug resistance mutations are acquired. Detecting low frequency mutations has a profound impact for adult solid tumor patients, for which tumors might be diagnosed non-invasively, by using bodily fluids such as peripheral blood, which generally contains circulating tumor cells or cell free DNA derived from cancer cells, albeit at a very low frequency. As discussed earlier (Fig. 3), ultra-high depth sequencing would be needed to interrogate low frequency mutations. For example, at least 91,533 reads need to be sequenced to achieve 95% chance to detect a mutation (with at least 5 recurrences) with population frequency 0.01%. Fortunately, the dramatic cost reduction in next generation sequencing technology has rendered this endeavor economically feasible (Fig. 8B).

However, a significant challenge in detecting low frequency mutations is associated with the high error rate in conventional NGS technology (*e.g.*, Illumina's HiSeq/NovaSeq sequencers). It has been reported [40–43] in the past decade that the error rate of conventional NGS technology is between 1% and 0.1% (Fig. 8C), so that the limit of detection is commonly set to above 1% [44]. With this observation, we recently conducted a carefully designed experiment to evaluate computational error suppression approaches and discovered [45] that the error rate of conventional NGS technology can be suppressed to between 0.01% and 0.001% (Fig. 8C), two-order of magnitude lower than generally reported. This breakthrough has opened entirely new opportunities for early detection of cancers, although its practical utility needs to be carefully evaluated in real clinical and biological settings, such as the recent findings on ubiquitously existing clonal hematopoiesis in aged healthy populations [46].

Despite these exciting breakthroughs, there are still numerous challenges for its real applications. For example, the PCR errors appear to dominate the NGS accuracy. Consensus sequencing methods, such as UMI-sequencing or duplex sequencing [40], are proposed to suppress such errors. However, the associated barcode-clashing problem [40] and barcoding uniformity problem [47] remains to be a significant numerical challenge that needs careful studies.

Integrative analyses of genomics and epigenomics datasets

Epigenomic regulation, such as DNA methylation, has



Year	PMID	Error rate
2011	21592312	0.1%
2012	22829749	1%
2013	23560931	0.5%
2016	27184599	0.1%
2018	29576615	0.1%
2019	30867008	0.01–0.001%

Figure 8. Detecting low frequency mutations. (A) Potential applications of low frequency mutation detection. (B) Sequencing cost analysis. Data (accessed Nov 1, 2019) from www.genome.gov/sequencingcostsdata. (C) Reported error rate of next generation sequencing technology in past decade. Pubmed ID of corresponding literatures listed.

been extensively studied in cancers as a biomarker for defining disease subtypes [48], and aberrant promoter methylation of tumor suppressor genes, such as *CDKN2A* [49]. Consistently, some of our work has clearly implied functional roles of aberrant epigenetic regulation in pediatric cancers. For example, analysis of allele specific expression of somatic point mutations in cancer driver genes has revealed preferential expression of hotspot oncogenic mutations [7], indicating addiction of cancer cells to these mutant proteins. A convincing example is *WT1* p.D447N, which was found to demonstrate strong allele specific expression across three leukemia samples, indicating a common functional mechanism of *WT1* mutation and epigenetic dysregulation in leukemia.

A more profound example indicating the potential functional involvement of epigenetics in tumorigenesis is found in a neuroblastoma sample. In our analysis of 136 pediatric neuroblastoma samples [7], *MYCN* amplification and ATRX disruption demonstrated a trend towards mutual exclusivity (Fig. 9A, $P = 0.12$ and statistical significance was not reached), which is consistent with functional studies [50]. However, there was one neuroblastoma case clearly harboring both *MYCN* high amplification and an *ATRX* frameshift mutation (T1582fs). To resolve this discrepancy, we noticed that this tumor sample was from a female patient with two X chromosomes intact in the tumor (*ATRX* is on chromosome X). Fortunately, the transcriptome sequencing data was also available. Upon examination of *ATRX* expression, we found that although the wildtype *ATRX* allele is expressed at a high level, the allele with frameshift mutation T1582fs is not expressed at all. Therefore, it is highly likely that an epigenetic mechanism (here X-inactivation, where one copy of X chromosome is inactivated in females) modulated the effect of *MYCN* and *ATRX* co-mutation (Fig. 9B). Under this hypothesis, the *ATRX* mutation T1582fs would not carry a pathogenic function for this patient because it is in the inactivated X chromosome and therefore the *MYCN* mutant cases should be updated from 32 to 33 and *ATRX* mutant cases is still 15, in turn statistical significance is reached ($P = 0.02$, Fig. 9B). Obviously additional experiments are warranted to validate this hypothesis. It should be noted that this mutation is classified as harmful in general as it will disrupt the normal function of *ATRX*.

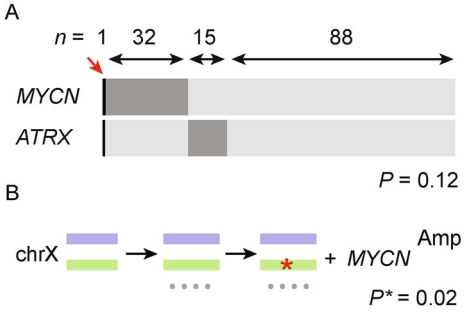


Figure 9. Integrating cancer genomics with epigenomics data. (A) Mutation status of *MYCN* and *ATRX*, two primary driver genes for neuroblastoma, from 136 cases. Dark gray: *MYCN* or *ATRX* mutant; black: *MYCN* and ATRX double-mutant (in which ATRX has mutation T1582fs). *MYCN* and *ATRX* mutant samples has a trend of mutual exclusivity, though statistical significance is not reached ($P = 0.12$, two-sided Fisher's exact test). (B) The *MYCN* and *ATRX* double-mutant sample is from a female patient with both alleles of chrX intact in the tumor (blue and green represent the two parental alleles, respectively). Since only the wildtype *ATRX* allele is expressed, it is hypothesized that the T1582fs mutation happened on the allele with X-inactivation (gray solid dots), and therefore carries no function, so that it is still biologically compatible with *MYCN* high amplification. As a result, the updated *P* value of 0.02 has reached statistical significance for the mutual exclusivity.

Collectively, these examples highlight the challenges and opportunities in integrative statistical analysis of genomics and epigenomics data to gain further mechanistic understandings of the cancer development as well as to correctly interpretation of functional consequences of mutations in real clinical settings.

Non-coding genome

Although extensive efforts have been invested in the protein-coding regions, which consists ~2% of human genome [26], our understanding of non-coding mutations is still in its infancy, leaving ~98% of somatic mutations poorly understood. Nevertheless, convincing results on functional non-coding mutations have been discovered. In pediatric cancers, active promoters/enhancers can be brought to a developmentally silenced gene, and the resulting aberrant high expression of corresponding gene can drive the tumorigenesis. Examples include the translocation between immunoglobulin heavy chain locus and *EPOR* [51], *DUX4* [52] in pediatric B-lineage leukemia. In T-lineage leukemia, enhancer activation by small indels that created a novel transcription factor binding motif has been described for gene *TAL1* [53]. Moreover, mutations at *TERT* promoter have been implied in melanoma [54]. Despite these incredible progresses on regulatory mutations, there is still lack of systematic informatics method to detect and associate the non-coding mutations to genes in disease-relevant fashion.

In addition to the highlight challenging tasks in understanding transcription regulatory mutations, there are also potentially splicing regulatory mutations and translation regulatory mutations. For the former, intronic mutations might create novel splicing acceptors so that

the coding sequences are dramatically affected, such as the STAG2 intronic mutation [7]. Since splicing machinery scans the pre-mRNA from 5′ to 3′, only novel splicing acceptors can theoretically affect the host protein. Clearly, transcriptome sequencing data is needed to confirm the *in-silico* prediction of splicing aberrations. For the latter, a mutation in the 5′ untranslated region might create a novel translation start codon and dramatically affect the protein translation of the canonical open reading frame [55]. However, proteomics data will be needed to confirm the functional impact of such mutations. Nevertheless, we believe innovative mathematical modeling will prove effective in discovering novel non-coding mechanisms driving tumorigenesis.

Etiology

With the large number of tumors being profiled for somatic mutations, a natural next question is on the cause of these events. A default hypothesis to answer this question is that the somatic mutations happen at random, although it remains to be precisely defined for "random". Interestingly, nonnegative matrix factorization decomposition [56] of large collections of somatic mutation data has linked mutations to different mutagenesis processes [57] including DNA methylation and smoking etc. Striking evidence on mutagenic effect from aristolochic acids and their derivatives have been discovered in liver cancers [58]. Recently, therapy-induced mutations were found to drive ALL relapse [37]. However, it should be noted that the association of mutation signatures to certain mutagens from tumor sequencing data are not necessarily causal relationships, as argued by Brash [59]. As a result, intensive efforts [60] have been investigated on establishing the causal relationships between mutagens and actual mutational data through controlled experiments.

On the other hand, signatures of other mutation types, such as indels and SVs, are not well understood yet. For indels, a key challenge can be traced back to the accuracy of the mutation calling algorithms. For example, a study of complex indels by Ding and colleagues has indicated that nearly all complex indels were overlooked (81.1%) or misannotated (17.6%) [61]. Without a precise allele annotation, the signature analysis, which depends on the actual nucleotide sequences around the mutations, will be error prone. This argument also applies to SVs, though the effect might be slightly less. For example, site-specific recombination mechanism has been proposed for *E2A-PBX1* positive childhood B-ALL because of clustered genomic breakpoints in intron 14 of *E2A* gene [62]. However, the general genomic pattern of disease-type defining driver fusions has not been well established.

Clearly, etiology of the different mutation types might have temporal dependence, such as smoking cessation in lung cancer patients. In this regard, it is possible that mutational signature analysis coupled with clonality measurements could result in unprecedented insights on the tumorigenesis process. However, the accuracy of clonality analysis is critically dependent on the standard deviation of mutant allele fraction estimate: $\sqrt{p \times (1-p)/N}$, where N is the sequencing depth and p is the underlying population frequency. As a result, variants discovered by whole genome sequencing (mostly 30) would have standard deviation of ~0.1, while that discovered by whole exome sequencing (mostly 100) would have standard deviation of ~0.05, and such statistical uncertainties can easily mess up the clustering algorithm. This observation, coupled with the variability in tumor purity, renders the modeling extremely challenging. Nevertheless, such analysis is at least feasible to certain special scenarios, such as the analysis of subtype-defining fusion genes, as well as in the diagnosis-relapse tumor sequencing settings, where the clonality can be relatively confidently ascertained. Alternative technologies, such as single cell sequencing, might render this proposal broadly possible to large cohorts, though associated sequencing cost would be still prohibitive at present.

CONCLUSIONS

The ability to profile all mutations in tumors at an economically feasible cost has rendered the concept of personalized and precision medicine a near-reality. However, acquisition of genomic data is just a first step toward this ambitious goal. The accurate analysis and correct interpretation of the massive amount of data, at personalized level, and even at spatial-temporal level for given individuals in the foreseeable future, is the key. To bring this technology to individual patients, multiple expertise, such as from quantitative scientists, are greatly needed to accelerate the development of methods and tools to better analyze, understand, and interpret genomic and epigenomics data to facilitate improved cancer outcomes.

ACKNOWLEDGEMENTS

X.M. is partly supported by The Innovation in Cancer Informatics (ICI) Fund. The authors are grateful to the editorial support by Makeda Porter-Carr.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Xiaotu Ma, Sasi Arunachalam and Yanling Liu declare that they have no conflict of interests.

This article is a review article and does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Nowell, P. C. (2007) Discovery of the Philadelphia chromosome: a personal perspective. J. Clin. Invest., 117, 2033–2035
2. Nowell, P. H. D. (1960) A minute chromosome in human chronic granulocytic leukemia. Science, 132, 1497
3. Sanger, F. and Coulson, A. R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol., 94, 441–448
4. Weinberg, R. A. (1991) Tumor suppressor genes. Science, 254, 1138–1146
5. Downing, J. R., Wilson, R. K., Zhang, J., Mardis, E. R., Pui, C. H., Ding, L., Ley, T. J. and Evans, W. E. (2012) The Pediatric Cancer Genome Project. Nat. Genet., 44, 619–622
6. Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 455, 1061–1068
7. Ma, X., Liu, Y., Liu, Y., Alexandrov, L. B., Edmonson, M. N., Gawad, C., Zhou, X., Li, Y., Rusch, M. C., Easton, J., *et al.* (2018) Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature, 555, 371–376
8. Gröbner, S. N., Worst, B. C., Weischenfeldt, J., Buchhalter, I., Kleinheinz, K., Rudneva, V. A., Johann, P. D., Balasubramanian, G. P., Segura-Wang, M., Brabetz, S., *et al.* (2018) The landscape of genomic alterations across childhood cancers. Nature, 555, 321–327
9. Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S. and Getz, G. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. Nature, 505, 495–501
10. Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D. M., Niu, B., McLellan, M. D., Uzunangelov, V., *et al.* (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell, 158, 929–944
11. Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., Lawrence, M. S., Zhang, C. Z., Wala, J., Mermel, C. H., *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. Nat. Genet., 45, 1134–1140
12. Rusch, M., Nakitandwe, J., Shurtleff, S., Newman, S., Zhang, Z., Edmonson, M. N., Parker, M., Jiao, Y., Ma, X., Liu, Y., *et al.* (2018) Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. Nat. Commun., 9, 3962
13. Crowley, E., Di Nicolantonio, F., Loupakis, F. and Bardelli, A. (2013) Liquid biopsy: monitoring cancer-genetics in the blood. Nat. Rev. Clin. Oncol., 10, 472–484
14. Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A. A., Wong, F., Mattox, A., *et al.* (2018) Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science, 359, 926–930
15. Tomasetti, C., Vogelstein, B. and Parmigiani, G. (2013) Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc. Natl. Acad. Sci. USA, 110, 1999–2004
16. Kunkel, T. A. and Erie, D. A. (2015) Eukaryotic mismatch repair in relation to DNA replication. Annu. Rev. Genet., 49, 291–313
17. Forsberg, L. A., Gisselsson, D. and Dumanski, J. P. (2017) Mosaicism in health and disease—clones picking up speed. Nat. Rev. Genet., 18, 128–142
18. Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., *et al.* (2013) An estimation of the number of cells in the human body. Ann. Hum. Biol., 40, 463–471
19. Testa, C. M. and Jankovic, J. (2019) Huntington disease: A quarter century of progress since the gene discovery. J. Neurol. Sci., 396, 52–68
20. Zhang, J., Walsh, M. F., Wu, G., Edmonson, M. N., Gruber, T. A., Easton, J., Hedges, D., Ma, X., Zhou, X., Yergeau, D. A., *et al.* (2015) Germline mutations in predisposition genes in pediatric cancer. N. Engl. J. Med., 373, 2336–2346
21. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. Jr and Kinzler, K. W. (2013) Cancer genome landscapes. Science, 339, 1546–1558
22. Pounds, S., Cheng, C., Li, S., Liu, Z., Zhang, J. and Mullighan, C. (2013) A genomic random interval model for statistical analysis of genomic lesion data. Bioinformatics, 29, 2088–2095
23. Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature, 499, 214–218
24. Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D., Mardis, E. R., *et al.* (2012) MuSiC: identifying mutational significance in cancer genomes. Genome Res., 22, 1589–1598
25. Soussi, T. and Wiman, K. G. (2015) TP53: an oncogene in disguise. Cell Death Differ., 22, 1239–1249
26. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. Nature, 431, 931–945
27. Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. Nature, 489, 519–525
28. Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature, 499, 214–218
29. Ma, X., Edmonson, M., Yergeau, D., Muzny, D. M., Hampton, O. A., Rusch, M., Song, G., Easton, J., Harvey, R. C., Wheeler, D. A., *et al.* (2015) Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. Nat. Commun., 6, 6604
30. Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., *et al.* (2012) Mutational processes molding the genomes of 21 breast cancers. Cell, 149, 979–993

31. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res., 29, 308–311
32. Griffith, M., Miller, C. A., Griffith, O. L., Krysiak, K., Skidmore, Z. L., Ramu, A., Walker, J. R., Dang, H. X., Trani, L., Larson, D. E., *et al.* (2015) Optimizing cancer genome sequencing and analysis. Cell Syst., 1, 210–223
33. Sundling, K. E. and Lowe, A. C. (2019) Circulating tumor cells: overview and opportunities in cytology. Adv. Anat. Pathol., 26, 56–63
34. Kakadia, P. M., Van de Water, N., Browett, P. J. and Bohlander, S. K. (2018) Efficient identification of somatic mutations in acute myeloid leukaemia using whole exome sequencing of fingernail derived DNA as germline control. Sci. Rep., 8, 13751
35. Mrózek, K., Heerema, N. A. and Bloomfield, C. D. (2004) Cytogenetics in acute leukemia. Blood Rev., 18, 115–136
36. Craig, D. W., Nasser, S., Corbett, R., Chan, S. K., Murray, L., Legendre, C., Tembe, W., Adkins, J., Kim, N., Wong, S., *et al.* (2016) A somatic reference standard for cancer genome sequencing. Sci. Rep., 6, 24607
37. Li, B., Brady, S. W., Ma, X., Shen, S., Zhang, Y., Li, Y., Szlachta, K., Dong, L., Liu, Y., Yang, F., *et al.* (2019) Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. Blood, 135, 41–55
38. Brady, S. W., Ma, X., Bahrami, A., Satas, G., Wu, G., Newman, S., Rusch, M., Putnam, D. K., Mulder, H. L., Yergeau, D. A., *et al.* (2019) The clonal evolution of metastatic osteosarcoma as shaped by cisplatin treatment. Mol. Cancer Res., 17, 895–906
39. Li, B., Li, H., Bai, Y., Kirschner-Schwabe, R., Yang, J. J., Chen, Y., Lu, G., Tzoneva, G., Ma, X., Wu, T., *et al.* (2015) Negative feedback-defective *PRPS1* mutants drive thiopurine resistance in relapsed childhood ALL. Nat. Med., 21, 563–571
40. Salk, J. J., Schmitt, M. W. and Loeb, L. A. (2018) Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat. Rev. Genet., 19, 269–285
41. Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet., 17, 333–351
42. Mardis, E. R. (2013) Next-generation sequencing platforms. Annu. Rev. Anal. Chem. (Palo Alto, Calif.), 6, 287–303
43. Glenn, T. C. (2011) Field guide to next-generation DNA sequencers. Mol. Ecol. Resour., 11, 759–769
44. Cheng, D. T., Mitchell, T. N., Zehir, A., Shah, R. H., Benayed, R., Syed, A., Chandramohan, R., Liu, Z. Y., Won, H. H., Scott, S. N., *et al.* (2015) Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J. Mol. Diagn., 17, 251–264
45. Ma, X., Shao, Y., Tian, L., Flasch, D. A., Mulder, H. L., Edmonson, M. N., Liu, Y., Chen, X., Newman, S., Nakitandwe, J., *et al.* (2019) Analysis of error profiles in deep next-generation sequencing data. Genome Biol., 20, 50
46. Young, A. L., Challen, G. A., Birmann, B. M. and Druley, T. E. (2016) Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat. Commun., 7, 12484
47. Ulz, P., Heitzer, E., Geigl, J. B. and Speicher, M. R. (2017) Patient monitoring through liquid biopsies using circulating tumor DNA. Int. J. Cancer, 141, 887–896
48. Figueroa, M. E., Lugthart, S., Li, Y., Erpelinck-Verschueren, C., Deng, X., Christos, P. J., Schifano, E., Booth, J., van Putten, W., Skrabanek, L., *et al.* (2010) DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. Cancer Cell, 17, 13–27
49. Ma, X., Wang, Y. W., Zhang, M. Q. and Gazdar, A. F. (2013) DNA methylation data analysis and its application to cancer research. Epigenomics, 5, 301–316
50. Zeineldin, M., Federico, S., Chen, X., Xu, B., Stewart, E., Naranjo, A., Hogarty, M.D., Dyer, M.A. (2020) *MYCN* amplification and *ATRX* mutations are incompatible in neuroblastoma. Nat. Commun., 11, 913
51. Iacobucci, I., Li, Y., Roberts, K. G., Dobson, S. M., Kim, J. C., Payne-Turner, D., Harvey, R. C., Valentine, M., McCastlain, K., Easton, J., *et al.* (2016) Truncating erythropoietin receptor rearrangements in acute lymphoblastic leukemia. Cancer Cell, 29, 186–200
52. Zhang, J., McCastlain, K., Yoshihara, H., Xu, B., Chang, Y., Churchman, M. L., Wu, G., Li, Y., Wei, L., Iacobucci, I., *et al.* (2016) Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. Nat. Genet., 48, 1481–1489
53. Mansour, M. R., Abraham, B. J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A. D., Etchin, J., Lawton, L., Sallan, S. E., Silverman, L. B., *et al.* (2014) Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science, 346, 1373–1377
54. Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L. and Garraway, L. A. (2013) Highly recurrent *TERT* promoter mutations in human melanoma. Science, 339, 957–959
55. Zhang, H., Si, X., Ji, X., Fan, R., Liu, J., Chen, K., Wang, D. and Gao, C. (2018) Genome editing of upstream open reading frames enables translational control in plants. Nat. Biotechnol., 36, 894–898
56. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. and Stratton, M. R. (2013) Deciphering signatures of mutational processes operative in human cancer. Cell Reports, 3, 246–259
57. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A. L., *et al.* (2013) Signatures of mutational processes in human cancer. Nature, 500, 415–421
58. Ng, A. W. T., Poon, S. L., Huang, M. N., Lim, J. Q., Boot, A., Yu, W., Suzuki, Y., Thangaraju, S., Ng, C. C. Y., Tan, P., *et al.* (2017) Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. Sci. Transl. Med., 9, eaan6446
59. Brash, D. E. (2015) UV signature mutations. Photochem. Photobiol., 91, 15–26
60. Petljak, M., Alexandrov, L.B., Brammeld, J.S., Price, S., Wedge, D.C., Grossmann, S., Dawson, K.J., Ju, Y.S., Iorio, F., Tubio, J.M.

C., *et al.* (2019) Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. Cell, 176, 1282–1294

61. Ye, K., Wang, J., Jayasinghe, R., Lameijer, E. W., McMichael, J. F., Ning, J., McLellan, M. D., Xie, M., Cao, S., Yellapantula, V., *et al.* (2016) Systematic discovery of complex insertions and deletions in human cancers. Nat. Med., 22, 97–104
62. Wiemels, J. L., Leonard, B. C., Wang, Y., Segal, M. R., Hunger, S. P., Smith, M. T., Crouse, V., Ma, X., Buffler, P. A. and Pine, S. R. (2002) Site-specific translocation and evidence of postnatal origin of the t(1;19) E2A-PBX1 fusion in childhood acute lymphoblastic leukemia. Proc. Natl. Acad. Sci. USA, 99, 15101–15106

