CellPress

Opinion

Early Detection of Cancer in Blood Using Single-Cell Analysis: A Proposal

Alexander Krasnitz,[1] Jude Kendall,[1] Joan Alexander,[1] Dan Levy,[1] and Michael Wigler[1,*]

Here, we explore the potential of single-cell genomic analysis in blood for early detection of cancer; we consider a method that screens the presence of recurrent patterns of copy number (CN) alterations using sparse single-cell sequencing. We argue for feasibility, based on *in silico* analysis of existing single-cell data and cancer CN profiles. Sampling procedures from existing diploid single cells can render data for a cell with any given profile. Sampling from multiple published tumor profiles can interrogate cancer clonality via an algorithm that tests the multiplicity of close pairwise similarities among single-cell cancer genomes. The majority of common solid cancers would be detectable in this manner. As any early detection method must be verifiable and actionable, we describe how further analysis of suspect cells can aid in determining risk and anatomic origin. Future affordability rests on currently available procedures for tumor cell enrichment and inexpensive methods for single-cell analysis.

Advantages of Single-Cell Analysis for Early Detection of Cancer

Cancer kills by spreading to distant sites. At the time of the first clinical presentation, metastasis has often already occurred. Were it otherwise, most cancers would be curable by surgery. It follows that there may be a window of time when detection of cancer and its timely extirpation might result in a cure. Some – perhaps most – tumors spread to distant sites via the blood. The evidence of this is that, in patients with metastatic disease, cancer cells are found in blood [1] and bone marrow [2]. Even for nonmetastatic malignancies, cancer cells may be present in the bone marrow [3]. Therefore, a periodic screen for cancer cells in the blood might detect disease before symptomatic clinical presentation and at a stage before malignant cells have successfully colonized elsewhere. To be useful such a test must have high **sensitivity** (see Glossary) and **specificity**.

We argue that such screening should be based on a nearly universal signal in the genomic DNA of cancer cells: almost all cancers stem from a clonal population of cells each bearing a shared profile of **DNA CN variation (CNV)**. This profile is acquired somatically and through clonal expansion and is distinct from the germline CNV profile of the patient. Several considerations distinguish screens based on enriched cell populations from screens based on **cell-free DNA (cfDNA)** in the circulation, and there are three potential advantages of analyzing cells over free DNA. First, both methods are plagued by a high background from the normal genome. However, cancer cells can be enriched from the blood [4–8] thus dramatically enhancing the signal of the cancer over the normal fraction. Second, once identified, suspected cancer cells can be exhaustively analyzed, either singly or in pools. For example, subsequent deeper

Trends

Metastasis is the most lethal aspect of cancer. To preempt metastatic spread, tumors must be detected early.

There is ample experimental evidence that large-scale DNA copy number (CN) variation is ubiquitous in multiple tumor types.

Clonal expansion is a hallmark of cancer; clonal expansion of cells with massively altered CN profiles is not observed outside cancer and should be targeted for early detection. Single-cell genomic analysis may be the best method to ascertain multiple genomic alterations occurring within the same cell.

Massively parallel CN profiling of individual cells by sparse sequencing is accurate and affordable.

Powerful microfluidic technologies exist to isolate small numbers (in the thousands) of candidate circulating tumor cells from blood samples.

Similarities among single-cell CN profiles may be exploited to identify clonal subpopulations. Transcriptional, epigenetic, or immunohistochemical examination of analytes pooled from suspicious malignant clonal cells found in blood may offer an opportunity to determine the tissue of origin.

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

*Correspondence: wigler@cshl.edu (M. Wigler).

 http://dx.doi.org/10.1016/j.molmed.2017.05.005

sequencing of single or pooled cancer cell DNA can confirm clonality, enable risk assessment, identify coding mutations, and help predict anatomic origin for subsequent image-based screening. The third advantage of cell-based screens is that they can start with a predictable, and an almost universal, cancer signature: CNV. Although cfDNA methods can also be based on CNV, the technical obstacles are enormous and the sensitivity is dubious [9]. To date, most cfDNAs in the blood of cancer patients have been tested through deep sequencing of coding regions from known cancer-related genes [10]. This approach, however, is problematic; numerous target genes remain unknown and many mutations of functional consequence might not reside within coding domains. As a consequence, such a method would have low sensitivity. Moreover, an observed somatic coding variant in a target gene will typically be uninterpretable. Many sequence variants might arise by somatic mutation in tissues and be represented in the blood yet not be harbingers of disease [11,12]. Finally, it takes more than one gene mutation to generate a malignant cell [13], but observing many pathogenic mutations in the circulation does not allow us to infer whether these mutations co-occur within the same cell. As a result, a gene-focused method would be likely to have low specificity. A method with low specificity would be a huge burden to patients and the medical system alike.

Powerful methods can now be applied to enrich cancer cells in the blood based on selecting for the cancer component or by filtering out blood elements based on surface markers and size [5,14]. Methods for analysis of single-cell DNA [15,16], RNA [17,18], and protein composition [19] are also increasing in power and decreasing in cost. From DNA analysis we can infer whether any atypical cells derive from a cancerous clonal population and assess the malignant potential [20–22]. The tissue of origin might be inferred from **RNA expression profiles** [23] or from **DNA methylation patterns** [24,25]. Once we identify the tissue of origin, more conventional imaging tools could be used to verify the location of malignant cells.

Thus, practical questions emerge: (i) what algorithms allow the confident detection of sparse cancer cells in a much larger population of cells; and (ii) are there enough neoplastic cells in the circulation to be detected? In this Opinion article, we illustrate and discuss a prospective protocol to test this (Figure 1, Key Figure). Indeed, most prevalent cancer types display extensive CN alterations (Figure 2A and Table S1 in supplemental information online) [26]. These alterations comprise **CN profiles** that must be shared by cancer cells from a given patient. From single-cell sequence analysis, and using CN data from 3852 cases of solid tumors described in The Cancer Genome Atlas (TCGA) [26], we have concluded that characteristic tumor profiles might be detectable in patients carrying a low load of tumor cells in the blood.

Performance of Single-Cell-Based Early Detection: A Virtual Clinical Assessment

To conceptualize a single-cell-based method of early detection of cancer, we envision the following procedure. Following a standard draw of 10 ml of blood, hematopoietic cell components could be depleted using surface antigens [5] or tumor-like epithelial cells directly selected for epithelial cell adhesion molecule (EpCAM) expression [14], or both (Figure 1). As we indicate below, very low **sequence coverage** (ultrasparse whole-genome sequencing) [15] on single cells enables sufficient DNA CN profiling to accomplish the goal of detecting a small subpopulation of **clonally related (CR)** tumor cells.

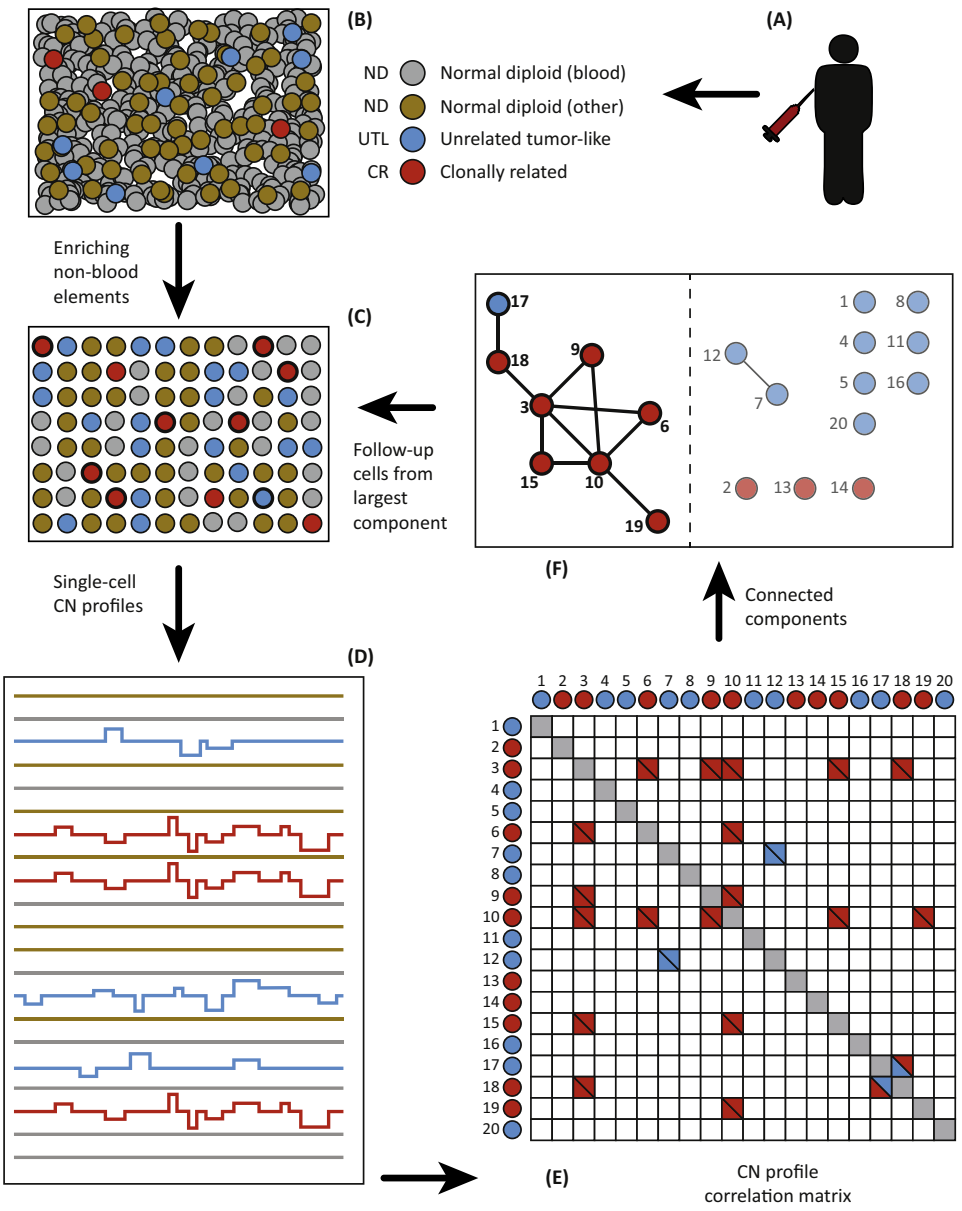
If the only challenge were to detect low numbers of a CR subpopulation in an overwhelming population of **normal diploid (ND) cells**, the algorithmic task would be straightforward. However, troublesome subpopulations of **unrelated tumor-like (UTL) cells** can also be expected to be present. These are likely to result from chromosome degradation in the dying normal cells, each with an abnormal and unique CN profile, ubiquitously observed in single-cell

Glossary

Bins: here, sequence intervals into which a genome is partitioned.
Cell-free DNA (cfDNA): DNA shed into the bloodstream as a result of cell disintegration.
Clonally related (CR) subpopulation: cells originating from a tumor clone.
Copy number (CN) profile: the CN of a genomic DNA sequence as a function of genomic position.
Counts per bin: sequence read counts in each of the bins into which the genome is partitioned.
DNA copy number (CN) variation (CNV): large-scale CN losses and gains.
DNA methylation patterns: the tissue-specific degree of DNA methylation as a function of genomic position.
Edge: a pairwise relationship between two vertices in a graph; if vertices are graphically represented by points, edges are represented by intervals connecting pairs of points.
Extreme-value theory: a branch of statistics dealing with extremely rare observations.
False-negative rate: the fraction of patients with at least one clonal tumor cell per milliliter of blood that is missed by the test; equals one minus the sensitivity.
False-positive rate: the fraction of tumor-free patients among those in whom the test indicates the presence of a tumor clone in the blood; equals one minus the specificity.
Graph: here, a set of objects called vertices and a (0,1)-valued function on pairs of vertices called adjacency; a pair of vertices with an adjacency of one are said to be connected by an edge.
Integer copy number (CN) profile: an approximation of an observed CN profile by an integer function of the genomic position.
Largest (connected) component: the largest subset of vertices in a graph such that there is a path of edges between any two vertices in the subset.
Modal: here, a value of DNA CN observed in a greater portion of the genome.
Normal diploid (ND) cell: a cell with a DNA CN of two throughout the genome, with the possible exception of the X and Y chromosomes.

Key Figure

Schematic of Single-Cell Detection Procedure in Blood



RNA expression profiles: numerical data comprising RNA expression levels for each gene in a given biological entity.
Sensitivity: the ratio of the number of patients testing positive for CTCs to the number of patients with CTCs being tested.
Sequence coverage: the total length of sequence reads divided by the length of the genome being sequenced.
Sequence read: a portion of an individual DNA molecule in which the nucleotide sequence is determined by a sequencing apparatus.
Shared copy number (CN) profile: a CN profile common to several cells.
Specificity: the ratio of the number of true positives to the number of all positive test results.
Stochastic copy number (CN) noise: variation in the CN profile due to random effects inherent to DNA preparation and sequencing procedures.
Unrelated tumor-like (UTL) cells: cells with significantly altered CN but not originating from a tumor clone.
Vertex: an object in a graph; vertices in a graph are customarily illustrated as points.

Figure 1. (A,B) Blood is drawn from a subject. (C) Blood elements are removed and cells of epithelial origin can be enriched. (D) Single-cell barcoded DNA libraries are prepared and sequenced, leading to single-cell copy number (CN) profiles. (E) Pairwise correlation analysis is performed. (F) A 'graph' is constructed where 'vertices' represent single-cell profiles or 'bin vectors' and two vectors are connected by an 'edge' if they are highly correlated. The largest 'connected component' is identified (F) and resides on the left-hand side of the broken vertical line. This identifies clonally correlated cells, indicated in bold. In the illustrated example, the largest connected component of the CN profiles is shown and predominantly stems from the clonally related (CR) tumor cells (3, 6, 9, 10, 15, and 19, in red). Not all tumor cells might be identified this way (2, 13, and 14). In this example, one unrelated tumor-like (UTL) cell (17) is fortuitously linked to the largest connected component. Two UTL cells (7 and 12) also form a small connected component. From the barcodes taken from

(See figure legend on the bottom of the next page.)

analyses. Conceptualizing a numerical method, we have assumed that some number of such UTL cells would be present. They are troublesome because they will not appear as normal and thus might be confused with tumor cells. However, we have also assumed that UTL cells might share CNVs with the tumor subpopulation only by chance. We argue that, in the absence of an overwhelming number of UTL cells, it is possible to effectively assess the presence of a small population with a **shared CN profile**.

A detection procedure would be considered successful if it is both sensitive (low **false-negative detection rate**) and highly specific (very low **false-positive detection rate**). Success would critically depend on many factors. First, sensitivity would depend on the degree of CNV in the CR cells, their count, and their proportion in the population. Second, specificity would depend on the number and proportion of UTL cells. Last, success would depend on operational factors such as the accuracy of determination of the CN profiles of individual cells and the computational protocol used to detect similar CN profiles in the presence of noise (the UTL profiles).

Modeling Single-Cell Sequence Data from Cells in Blood

To properly model a sample of cells that are subjected to an early detection procedure, we must specify for each category of cells in the sample the expected CN properties and the expected cell count. For tumor cells, published **integer CN profiles** for 3852 tumors representing eleven common solid tumor types can be used in this analysis [26,27]. We can assume that a sample obtained from a patient with one of these tumors can contain CR cells, all sharing a CN profile with the tumor. We can further assume the CR cell count to be 10 cells per 10 ml, well within the range of published circulating tumor cell (CTC) counts [6,8,28]. To model CR cells, one tumor profile is selected for each clone. The 3852 published profiles are also used to model UTL cells: random chromosomes from genomes of archived tumor profiles are sampled. Modeling UTL cells as random tumor cells adds a burden to the task but is a conservative procedure to follow. In the absence of experimental data on the number of UTL cells with substantial CNV, we can allow this number to vary within a broad range of 10–200 UTL cells per specimen. With these assumptions one might prefer to err on the side of caution, as large numbers of cancer-like cells in a specimen can pose a challenge for the achievement of specificity of early detection. Finally, ND cells can be assumed to have strictly diploid genomes and the specimen might be expected to contain 10^3–10^5 such cells. We anticipate that ND cells do not affect the specificity of early detection even if present in large numbers.

Next, sparse single-cell **sequence reads** can be simulated. The methodology is outlined here (Figure 1) and described in full detail in the supplemental information online. To model a read set with a specified coverage and CN profile, an empirical read set can be selected from an in-house (not yet published) collection of 1306 single-cell genomes of ND cells with an average of 2×10^6 reads per genome. First, the needed number of 1306 read sets is selected and no read set is chosen twice. By not selecting the same read set twice, spurious correlations can be avoided, albeit limiting the simulation of sequence data to no more than that number of cells. The read set is then resampled with replacement to achieve a sparser read set that would correspond to a cell with the desired CN profile and coverage. The mean sampling rate per read can be set to reach an average overall coverage of 1.25×10^5 reads. This process might at first seem overly complicated, but it captures, nevertheless, the **stochastic noise** that is inherent to single-cell read data.

the largest component, one can return to select cells in (C) (in bold) for deeper analysis; this might confirm cell clonal relations, potentially assess prognosis and therapeutic options, and determine the anatomic origin of the primary tumor.

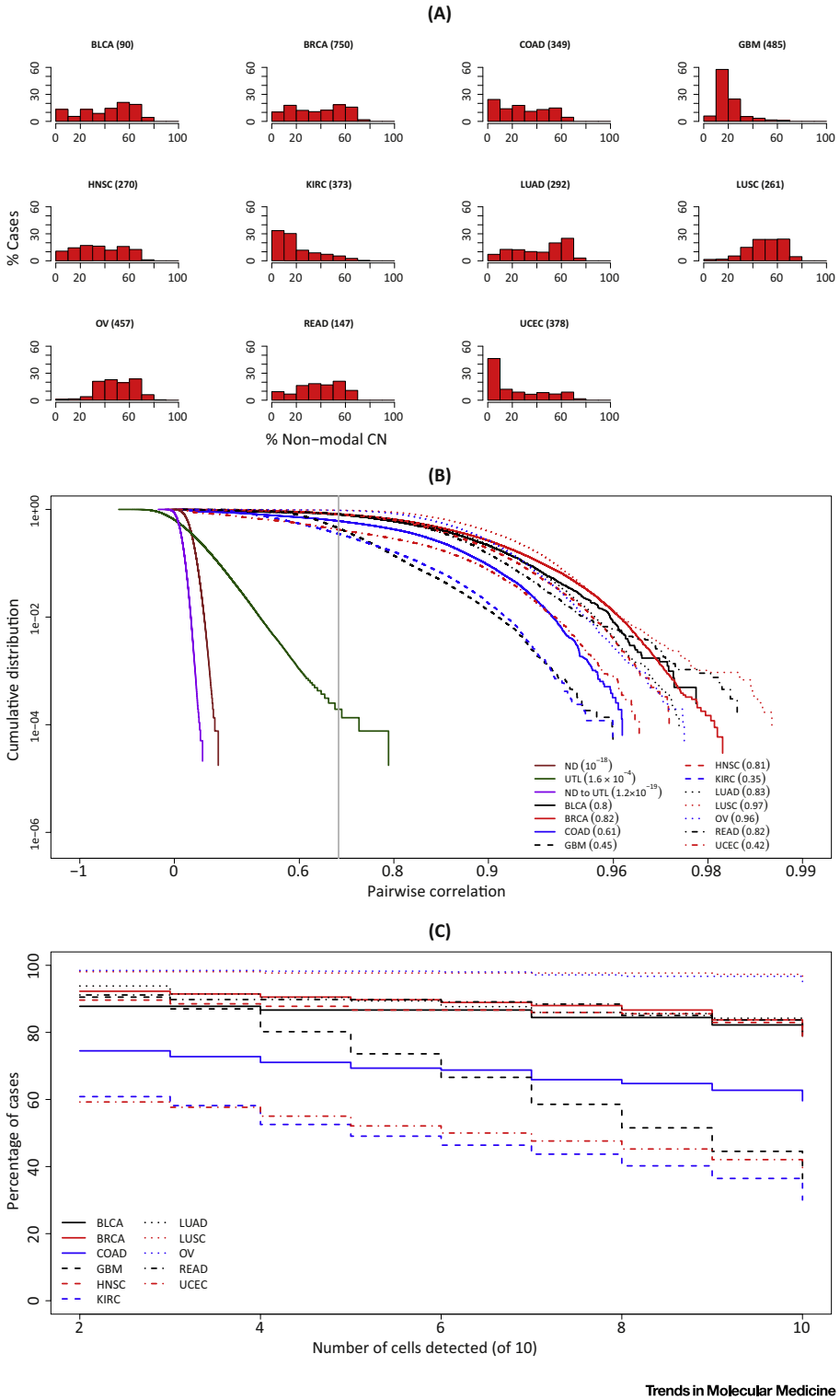


Figure 2. Detection of Clonal Cells in Circulation. (A) Ubiquity of copy number (CN) variations across multiple cancer types; each histogram shows the percentage of patient cases (vertical axis) with a given percentage of the genome with an abnormal non-modal CN value (horizontal axis). The cancer types represented are (in alphabetical order) bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), glioblastoma multiforme (GBM),

(See figure legend on the bottom of the next page.)

Table 1. Estimates of False-Positive Rate[a]

S	10 UTL cells	20 UTL cells	50 UTL cells	100 UTL cells	200 UTL cells
2	0.0086	0.028	0.15	0.43	0.84
3	3.00×10^{-4}	0.0018	0.028	0.13	0.51
4	1.00×10^{-4}	1.00×10^{-4}	0.0036	0.044	0.3
5	0	1.00×10^{-4}	4.00×10^{-4}	0.012	0.16
6	0	0	2.00×10^{-4}	0.0031	0.08
7	0	0	0	6.00×10^{-4}	0.034
8	0	0	0	3.00×10^{-4}	0.013
9	0	0	0	0	0.0049
10	0	0	0	0	0.0012
12	0	0	0	0	1.00×10^{-4}

[a]Assuming that 10, 20, 50, 100, or 200 UTL cells are present in the specimen (columns), the frequency of observing a connected component of at least $S = 2, 3, \ldots, 10$, or 12 cells is estimated (rows; cf. 'Methods' in the supplemental information online).

Clonal Detection Procedure

To derive CN profiles from simulated read data, a tested and published pipeline for sparse reads from cancer and diploid cells can be used [15,29]: the genome is partitioned into contiguous **bins** and the set of sequence reads is converted into read **counts per bin**. The output of the pipeline is thus a bin vector comprising counts whose values represent CN estimates in those bins. From two cells, each with bin vectors, a Pearson correlation can be calculated (Figure 1D, E). Two CR cells would then exhibit high Pearson correlation and UTL cells would have close to zero correlation. The distributions of pairwise correlation coefficients within the three populations (ND, UTL, and CR) are shown in Figure 2B.

A simple heuristic is then introduced to detect the presence of a clonal population with a shared CN profile (Figure 1F). Individual cells in the specimen are considered **vertices** in a **graph** and a pair of vertices is connected by an **edge** if the pairwise Pearson correlation of the bin vectors exceeds a given empirical test threshold. In this analysis a correlation threshold of 0.7 is chosen because correlations of this magnitude are observed frequently among CR cells but never for ND cells (Figure 2B). Pairwise correlations of this magnitude might occur among the UTL cells, but at a frequency of approximately one per 10^4 pairs (Figure 2B). The largest component of the graph comprises the largest set of vertices and their connections such that a path of vertices and edges connects any two vertices (Figure 1F). The vertices of the largest component are

head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian cystadenocarcinoma (OV), rectal adenocarcinoma (READ), and uterine corpus endometrial carcinoma (UCEC). The number of cases for each type is listed parenthetically and is based on data published by The Cancer Genome Atlas (TCGA) [26]. Thus, for example, 60% of GBMs (~500 cases) have CN alterations in >10–20% of the genome. (B) The graph presents the cumulative probability distribution (CDF) for the correlation of pairs of single-cell CN profiles. The CDF is plotted for pairs of normal diploid (ND) cells (unbroken brown line), for pairs of clonally unrelated tumor-like (UTL) cells (unbroken green line), and for mixed pairs of a ND cell and an UTL cell (unbroken magenta line). For comparison, the CDF for pairs of profiles from two cells simulated from the same cancer are plotted and the analysis of each of the 11 tumor types considered is shown, as described in the inset legend. The gray vertical line indicates the discriminatory threshold of 0.7 correlation. The values of the CDF at this threshold are indicated in the inset for each curve in parentheses. These values are taken from the simulations of the 11 tumor types and for UTL cells and from estimates based on extreme-value theory for ND cells and for the mixed ND–UTL pairs. (C) To assess the sensitivity of detection, for each tumor type the cumulative percentage of patients is plotted against the number of clonal cells detected (of ten assumed to be present in the specimen). For example, for BRCA (unbroken red line) a sensitivity of close to 90% is achieved if the number of clonal cells detected is at least six. For KIRC (broken blue line), the sensitivity is approximately 50% for that threshold.

selected to represent a clone of cells with shared CN profiles and the number of vertices in the connected component can provide an estimate of the number of clonal cells in the specimen. Tumor cells can be considered 'detected' if the number of estimated clonal cells in the specimen exceeds a set value *S* as described below (e.g., eight detected cells are shown in Figure 1F).

Sensitivity and Specificity of Clonal Detection

With this procedure the sensitivity and specificity of clone detection can be assessed. We consider these separately. For sensitivity we can ask how often a large connected component can be detected for ten CR cells taken from each of the 3852 tumors, each with published integer CN profiles [26]. Then, for specificity, we can ask how often ND or UTL cells – present in given numbers – result in a large connected component and therefore represent a false-positive signal.

For sensitivity we can assume that ten CR cells from the tumor are present in the specimen. Our summary analysis is depicted in Figure 2C, where for each of the 11 tumor types the fraction of cases (on the y-axis) with a detected clonal population of size *S* (on the x-axis) is depicted. Thus, the graph shows that this detection method would be highly sensitive in the majority of tumor types considered (Figure 2C). For example, a sensitivity of approximately 90% could be achieved in breast cancer if the size *S* of the **largest connected component** is set to be at least 6 (Figure 2B). For ovarian cancer sensitivity could approach 100%, and this is relevant as this disease is most often diagnosed at a late stage, portending an adverse outcome.

To evaluate the specificity of detection, we can measure the frequency of false detection within a population containing only ND and UTL subpopulations. The criterion for a false positive would be defined by the presence of a connected component of highly correlated cells within the specimen. This false-positive rate would therefore depend crucially on the frequency of high correlations in cell pairs within and between subpopulations. Pearson correlations of CN profiles of ND cells – either among themselves or with CN profiles of UTL cells – are highly unlikely to exceed the threshold value of 0.7. None of the empirical correlations of this kind are found to exceed this threshold. Furthermore, using methods of **extreme-value theory** [30–32]

Box 1. Clinician's Corner

- Metastasis is by far the most lethal aspect of cancer. It is likely that cancer mortality would be significantly reduced if most tumors were detected early enough to preempt metastatic spread.
- Early detection of tumors cannot be accomplished consistently by methods in present clinical practice as none of these methods is simultaneously sensitive, specific, and applicable across multiple tumor types. For example, MRI or CT is able to detect a tumor at many anatomic locations but not before the tumor is macroscopically large. Similarly, blood-based tests for antigens such as prostate-specific antigen (PSA) focus on a single tumor type and lack specificity.
- There is mounting evidence that tumor cells enter the circulation and are disseminated from the primary site to distant organs early in disease. It is therefore likely that such cells are present in the circulation even before the disease is detectable by imaging or physical examination.
- Here we argue that, for multiple tumor types, a combination of molecular techniques can be used to detect CTCs, even if the latter are as rare as one per milliliter of blood.
- The procedure we envision involves three critical steps: (i) candidate CTCs are separated from blood cells using surface proteins that discriminate the two populations; (ii) the genome of each candidate CTC is examined individually for aberrations in DNA CN; and (iii) CTCs are identified from cell candidates based on a shared pattern of aberrations.
- The entire detection procedure might be accomplished with existing molecular technologies at a cost of approximately US$1000 per test.
- Once CTCs are identified, a more stringent molecular analysis can be used and may include a combination of deep sequencing and analysis of methylation or gene expression patterns. This information might be used to confirm malignancy risk and identify the putative tissue of origin, thus guiding possible diagnosis and treatment.

the probability of exceeding this threshold with at least one member of the pair being ND can be expressed in terms of the generalized Pareto distribution and estimated to be below 10^{-18}. Consequently, even with approximately 5×10^9 pairwise correlations in a specimen of 10^5 cells, the ND cell subpopulation might present no challenge to the specificity of detection and could be safely dropped from the determination of the false-positive rate.

By contrast, correlations among CN profiles of UTL cells might exceed the threshold of 0.7 with a low but non-negligible empirical probability (e.g., 1.6×10^{-4}; Figure 2B). If present in sufficient numbers, these cells might occasionally form sizable connected components as defined above, leading to the false discovery of a clonal population. The corresponding false-positive rate can be determined empirically by simulating multiple random sets of 10, 20, 50, 100, or 200 UTL cells and computing the size of the largest connected component for each set respectively (Table 1). In Table 1 the corresponding empirical false-positive rate is listed for each of the five set sizes and for a range of threshold values of *S*. For example, in a specimen with 100 UTL cells, false discovery of a connected component comprising six or more cells would rarely occur (approximately one per 3000 cases). UTLs might occasionally contribute to the largest connected component (Figure 1E,F) but could be eliminated on follow up (Figure 1C–F).

Follow-Up Analysis from Positive Findings

As discussed above, false-positive rates can be low but are not 'zero'; consequently, any level of false positives is unwanted. Moreover, a true positive must be actionable. Therefore, detection of a clonal pattern in blood would require us to do more. Certainly, a repeat test starting with larger volumes of blood would be warranted. Once the positive cells are identified, they can be further analyzed, either singly or in pools, provided one uses methods that enable the reexamination of single-cell nucleic acids (Figure 1C–F). This is presently out of reach with our repertoire of molecular tools but may be possible using either addressable arrays of single-cell nucleic acids or addressable nucleic acid libraries.

Deeper DNA sequencing of candidate cell nucleic acid can enable the virtual elimination of false positives, as UTL cells with correlated profiles (low coverage and resolution) might rarely exhibit coincident individual CN events (higher coverage and resolution), whereas clonal tumor cells would. Moreover, details of CN profiles (overall ploidy, number and location of CN events, etc.) combined with knowledge of tissue of origin are likely to yield (ultimately) a good assessment of the risk of malignancy and the source of the neoplasm [20–22]. Using pooled nucleic acid from tumor cells that are confirmed, we might reach sufficient sequence depth to observe therapeutically actionable mutation patterns. Finally, strong clues to the tissue of origin can in turn be discovered from methylation patterns, chromatin structure, and gene expression data [24,25]. Comprehensive databases with this type of information are currently being compiled and will be most useful [33,34]. Defining an anatomic origin of neoplasia will be critical for diagnostics, where, once the locations of probable primary sites are identified, subsequent scanning/imaging at high resolution could validate any preliminary blood-based findings.

Concluding Remarks

We have outlined a sparse genomic sequencing method for the detection of cancer from single cells in blood (Figure 1). It comprises components that include: (i) enriching atypical cells from blood (Figure 1A–C); (ii) separating individual cells (Figure 1C); (iii) performing inexpensive CN profiling (Figure 1D); and (iv) detecting clonal CN profiles computationally (Figure 1D–F). The technical components (i)–(iii) are presently feasible and we posit that, based on simulation analysis under certain conditions, (iv) could be used to detect clonal cancer cells with low false-positive rates. No false positive is without harm and a true positive will require actionable information, which might be attained by tracing back and performing deep-sequencing analysis

Outstanding Questions

What is the specific abundance of malignant clonal cells in circulation, before the emergence of macroscopic disease, by tumor type?

Are non-clonal aneuploid cells in circulation rare enough to keep false-detection rates low?

How frequently are non-malignant clonal cell populations bearing substantial CNV found in blood?

What categories of patients are most likely to benefit from, and should be a priority target for, single-cell-based early detection procedures?

of candidate single-cell nucleic acids either singly or in pools (Figure 1C–F). In this scenario one might confirm risk of disease, reduce false-positive findings to negligible levels, discover critical mutations of possibly therapeutic significance, or determine the anatomic origin of the primary site of clonality. This information would then be actionable.

From another angle, we have set the sequence coverage for genomic analysis in a way that the procedure should be affordable. At approximately 125 000 reads per single cell (given the least expensive high-throughput sequence platform presently available), we anticipate that the cost basis for the sequencing alone might be constrained to approximately US$1 per cell. Detecting ten tumor cells in the presence of 1000 diploid cells might cost approximately US $1000 for the sequencing component. At this sequence depth, false-positive rates may be acceptable if no more than 100 non-clonal cells with abnormal genomes are present per enriched blood specimen. A more expensive test, with higher coverage, would reduce false-positive rates. The follow-up analysis after a positive detection would undoubtedly be more expensive.

Of note, other conditions might arise to confound detection, such as the presence of benign populations in the blood carrying CN alterations (see Outstanding Questions and Box 1). We cannot presently answer this and many other critical questions, such as how often, in what numbers, and at what stage tumors release malignant cells into the blood. One central question is whether this detection method in blood reveals malignant neoplasms sufficiently early so that appropriate interventions can be made and the incidence of metastasis reduced for any cancer type. The procedure might most benefit certain patients with increased risk; for example, the elderly, those with genetic predisposition, or those with suspected lesions for whom invasive surgical biopsy is not the best option. Of relevance, these ideas are not restricted to the detection of cancer signatures in blood but might also be of value in detecting low levels of nonspecific cancer signatures in any biological specimen.

Acknowledgments

The authors are grateful to David Donoho for informative discussions in the early phases of this work. This work was supported by the National Cancer Institute (A.K. and M.W., award U01CA188590), the National Institutes of Health (M.W., award 5R01CA181595-03), the Simons Foundation (A.K., D.L., and M.W.), and the Breast Cancer Research Foundation (M.W.).

Supplemental Information

Supplemental information associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.molmed.2017.05.005.

References

1. Ignatiadis, M. *et al.* (2015) Circulating tumor cells and circulating tumor DNA: challenges and opportunities on the path to clinical utility. *Clin. Cancer Res.* 21, 4786–4800
2. Shiozawa, Y. *et al.* (2015) Bone marrow as a metastatic niche for disseminated tumor cells from solid tumors. *Bonekey Rep.* 4, 689
3. Demeulemeester, J. *et al.* (2016) Tracing the origin of disseminated tumor cells in breast cancer using single-cell sequencing. *Genome Biol.* 17, 250
4. Pantel, K. (2016) Blood-based analysis of circulating cell-free DNA and tumor cells for early cancer detection. *PLoS Med.* 13, e1002205
5. Sarioglu, A.F. *et al.* (2015) A microfluidic device for label-free, physical capture of circulating tumor cell clusters. *Nat. Methods* 12, 685–691
6. Theil, G. *et al.* (2016) The use of a new cellcollector to isolate circulating tumor cells from the blood of patients with different stages of prostate cancer and clinical outcomes – a proof-of-concept study. *PLoS One* 11, e0158354
7. van der Toom, E.E. *et al.* (2016) Technical challenges in the isolation and analysis of circulating tumor cells. *Oncotarget* 7, 62754–62766
8. Zhao, L. *et al.* (2016) Enhanced and differential capture of circulating tumor cells from lung cancer patients by microfluidic assays using aptamer cocktail. *Small* 12, 1072–1081
9. Heitzer, E. *et al.* (2013) Establishment of tumor-specific copy number alterations from plasma DNA of patients with cancer. *Int. J. Cancer* 133, 346–356
10. Bettegowda, C. *et al.* (2014) Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* 6, 224ra24
11. Martincorena, I. *et al.* (2015) Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* 348, 880–886
12. van der Putten, L.J.M. *et al.* (2017) Molecular profiles of benign and (pre) malignant endometrial lesions. *Carcinogenesis* 38, 329–335

13. Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell* 144, 646–674
14. Swennenhuis, J.F. *et al.* (2016) Improving the CellSearch® system. *Expert Rev. Mol. Diagn.* 16, 1291–1305
15. Baslan, T. *et al.* (2015) Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. *Genome Res.* 25, 714–724
16. Gawad, C. *et al.* (2016) Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* 17, 175–188
17. Kolodziejczyk, A.A. *et al.* (2015) The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58, 610–620
18. Trombetta, J.J. *et al.* (2014) Preparation of single-cell RNA-seq libraries for next generation sequencing. *Curr. Protoc. Mol. Biol.* 107, 4.22. 1-4.22.17
19. Spitzer, M.H. and Nolan, G.P. (2016) Mass cytometry: single cells, many features. *Cell* 165, 780–791
20. Blegen, H. *et al.* (2003) DNA amplifications and aneuploidy, high proliferative activity and impaired cell cycle control characterize breast carcinomas with poor prognosis. *Anal. Cell. Pathol.* 25, 103–114
21. Hicks, J. *et al.* (2006) Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* 16, 1465–1479
22. Hieronymus, H. *et al.* (2014) Copy number alteration burden predicts prostate cancer relapse. *Proc. Natl. Acad. Sci. U. S. A.* 111, 11139–11144
23. Stancel, G.A. *et al.* (2012) Identification of tissue of origin in body fluid specimens using a gene expression microarray assay. *Cancer Cytopathol.* 120, 62–70
24. Moran, S. *et al.* (2016) Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol.* 17, 1386–1395
25. Sun, K. *et al.* (2015) Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci. U. S. A.* 112, E5503–E5512
26. Zack, T.I. *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* 45, 1134–1140
27. Carter, S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* 30, 413–421
28. Gao, W. *et al.* (2016) Analysis of circulating tumor cells from lung cancer patients with multiple biomarkers using high-performance size-based microfluidic chip. *Oncotarget* 8, 12917–12928
29. Kendall, J. and Krasnitz, A. (2014) Computational methods for DNA copy-number analysis of tumors. *Methods Mol. Biol.* 1176, 243–259
30. Coles, S. (2001) *An Introduction to Statistical Modeling of Extreme Values Springer Series in Statistics,* Springer
31. Gilleland, E. and Katz, R.W. (2016) extRemes 2.0: an extreme value analysis package in R. *J. Stat. Softw.* 72, 1–39
32. Gilleland, E. *et al.* (2013) A software review for extreme value analysis. *Extremes* 16, 103–119
33. Bujold, D. *et al.* (2016) The Internationa Human Epigenome Consortium Data Portal. *Cell Syst.* 3, 496–499 e2
34. Roadmap Epigenomics Consortium *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330