Studying heterogeneity in a cancer population using primary tumor sectioning and single cell PCR

by Travis Zack

April 16, 2010
Introduction
Cancer is a broad term for somatic tissue overcoming its natural programming as a small, often specialized, part in a multi-cellular organism and transforming into an entity capable of survival and proliferation independent of its traditional context within the organism [1]. In this sense, cancer can be thought of as a genome that has evolved from multicellular roots into a rapidly evolving parasite (of unknown organization) living within the ecosystem defined by the host. This traditional view of cancer as a disease has slowly changed into one that takes the stochastic nature of its development into account when attempting to treat it. Acknowledging that cancer in different patients may have acquired different dependencies, a large amount of current research is being dedicated to detecting differences between patients on a gene by gene basis in an effort to find patterns of dependency that can be targeted [2, 3]. The goal of such research is to identify, based on the cancers genotype, which subset of patients will respond well to a particular treatment [4]. However, this “personalized medicine” approach has had mixed results [5]. Though we have been fairly successful at identifying the best gene to target in a particular individual and designing a treatment that will cause remission, invariably a drug-resistant malignancy emerges quickly, and often growth is quicker and more aggressive than it was pre-treatment. This suggests that, though we may have identified the genetic dependencies of the vast majority of the cancer cells, the heterogeneity of the population has allowed it to survive the selective pressure of drug treatment and emerge more virulent. Additionally, the exact organization of this heterogenous collection of parasitic cells has been subject to much recent debate [6, 7, 8]. Up until recently, this collection of cells was modeled as a colony of single cell organisms under intense environmental and intraspecies selection for certain “cancer” defining traits, such as evasion of apoptosis and unlimited proliferation potential. However, recently many prominent cancer biologists have suggested a “cancer stem cell” theory to explain observations such as cancer’s incredible resilience to drugs and inability to proliferate when transplanted in low quantities into mouse models [7]. In this model, a cancerous tumor is organized much like a multicellular organ with very rare, undifferentiated progenitor cells having a low proliferation rate and unlimited replicating potential which “drive” the tumor and differentiated cells with very fast proliferation, but finite duplication, that make up the bulk of the tumor mass [8]. However, direct genomic evidence for either case is lacking as most current primary tumor sequencing involves bulk characterization of the cancer genome, so almost all heterogeneity is lost. By using a single cell, transcriptome sequencing technique, we will attempt to quantifying the extent of heterogeneity within a human primary tumor, and use this information to answer three fundamental biological questions. Thus our stated Aims are:

Aim 1: Design a Protocol to sequence individual cancer cells in a tumor population. (Section 2)
Aim 2: Determine whether a static population analysis of primary tumor cells support the Cancer Stem Cell/Cancer Tissue theory, or the single-celled, survival of the fittest model of cancer evolution. (Section 3.2)
Aim 3: Use the phylogenetic associations between mutations in different known cancer-related genes to determine a temporal sequence of emergent driver mutations. (Section 3.3)
Aim 4: Determine whether local environment creates spatial heterogeneity in driver mutation selection. (Section 3.4)

1. Aim 1: Experimental Procedure

When looking at cancer from a population analysis standpoint, the first question to tackle is how do we measure heterogeneity in a cancer population quantitatively? There are two main technical papers that we will build on to design a protocol to address this question [9, 10]. Kumaresan et al.
(2008) describes an emulsion based technique that allows one to isolate single cells with DNA coated beads, which will allow us to barcode individual cells. Emulsion PCR has been a useful method to amplify very small amounts of nucleic acids, all the way down to a single molecule [11] as the starting material. In addition, we will make use of a recent Nature Methods [10] paper to remove cells from tissue sections while maintaining spatial information about the local environment where they are from. By combining the high-throughput of the first paper with the tissue handling techniques of the second paper, our experimental section should accomplish our goal.

1.1 Bead design and production
The first step will be emersing Strepavidin-coated beads into a pool of DNA biotinylated on the 5’ end. The Biotinylated DNA (bDNA) will be synthesized from scratch and will contain four regions as shown in Figure 1A. Sequentially from the biotinylated 5’ end, there is a short buffer region to allow the PCR enzymes to reach the end of the sequences unhindered by the bead surface, followed by a universal primer region (UP), an unique barcode region 9-bp long, followed by a poly-T region for mRNA reverse transcription. Though we will only make 2,000 unique barcodes, we want redundancy in the uniqueness of barcode so even if one base is mutated during PCR, the cDNA it is attached to will not be mis-called.

To construct unique DNA-coated beads, Streptavidin beads will be placed in wells containing the bDNA, with each well containing bDNA with a unique barcode. This will require approximately 2000 wells, but should only need to be done once at the beginning of the experiment. Afterwards, a small amount of beads can be removed from each well, pooled, and washed for each local sequencing event.

1.2 Primary tumor sectioning and local population collection
Primary tumor is to be obtained from clinical collaboration. This tumor will be subjected to fresh frozen sectioning using a cryostat microtome according to [10] with a average section width of about 15 μm. Once sectioned and placed on a slide, clumps of 300-500 cells can be removed from multiple areas using a laser capture microdissection (LCM). Removal of these cells using requires destruction of the rest of the slide, but adjacent slides of the sectioned tumor can be used to recover local environment information. Once all the local regions of interest are recovered, slides can be subjected to ImmunoHistoChemical (IHC) staining, followed by 3D tumor reconstruction to provide local environment information for Aim 4. The captured regions of interest can be removed from the slides one by one, placed in individual tubes, and fixed to await sequencing.

1.3 Microfluidics engineering
The experimental design in [9] is diagrammed in Fig 2 below. Cells and DNA coated beads are added in dilute concentrations to a solution containing all the components required for your particular PCR reaction (Primers, buffer, MgSO₄, and high fidelity polymerase, etc.). This solution is flowed through a injector junction, which creates droplets of water with very uniform volumes in a stream of emulsion oil. By adjusting the concentration of the cells and beads, you can insure each droplet contains at most one cell, though may contain multiple beads (not optimal). The bead design will depend on the target nucleic acids. For our initial studies, we are going to use poly-T mRNA reverse transcription. However, if this method proves too non-specific for our single cell technique, we could alter the protocol to target specific mRNA.

1.4 Experimental Flow
Once the individual components have been prepared, you can mix cells from a local population with a solution of barcoded beads. The concentration of beads should be much greater than concentration of cells, but the final ratio may have to be experimentally determined to balance efficiency of cell collection and sequencing. You would like most of your individual cells to be trapped with only one bead, but a small number will be trapped with more than one. The solution, which should contain all
the necessary elements for Reverse Transcription, is sent the microfluidic device, which creates water droplets (Fig. 2, [9]). These water droplets can then sent through a modified flow cytometer on their way to a PCR tube. The flow cytometer must be modified in such a way that the hydrodynamic focusing is based on the viscosity and density of the emulsion oil, not water (as is typical). It will sort droplets by determining the number of beads in each droplet, and removing droplets with more than two beads. This step is necessary for population genetics studies because sampling more than once from the same cell may lead to underestimating genetic variation in your population, as well as needless sequencing.

Once your local population is separated into single cell/single bead droplets and collected in a single PCR tube, the tube can be heated to allow cell permeabilization. This should be followed by a single round of PCR, which will create cDNA covalently-linked to your beads. The droplets can then be combined, and the beads are extracted and washed. The cDNA from the beads is eluted by following the protocol outlined from Invitrogen (under “Streptavidin-coated Dynabeads”). This cDNA is added to a PCR solution and is again sent through emulsion droplet formation, collected in eppendorf tubes, and subjected to multiple rounds of high fidelity PCR using the Poly-T primers and the Universal Primer Region (Fig. 1A) [12]. This final separation is to control for mutations occurring during PCR (see Analysis). All of the cDNA from this entire local population is then mixed and sent for pyrosequencing. A bulk sample of normal somatic tissue from the same patient should be sequenced at the same time to accomplish Aims 2 and 3.

2. Analysis

2.1 Minimizing error Our biggest concern will be separating mutations derived from errors during PCR amplification and real variability in the primary tumor. However, this technique allows us to separate these relatively easily, as long as we assume all mRNA from a single cell is derived from a unique genetic sequence (that is, we ignore functional copy number variation). This is, of course, a big assumption given the frequency and importance of copy number variation, but unavoidable. Because mRNA from each cell contains a unique barcode, we will throw out a mutation if it occurred in less than 80% (cutoff arbitrary and will have to be optimized) of cDNA that has the same barcode (referred to as error cDNA). This should remove your two main sources of noise. Specifically:

1. If there are errors during Reverse Transcription, this error will not be the same on different mRNA on the same bead, so removing a mutation that isn’t ubiquitously associated with a specific barcode removes these mutation events.

2. If there are errors during PCR amplification, these errors would not be identical from droplet to droplet, again ensuring that if a large fraction of reads from the same barcode come up with a different base pair at this location, this event probably occurred during PCR. Though there are approximately 10x more unique barcodes than there are cells per local population, there may be cases where two beads with the same barcode end up with two different cells. This process will create a false negative, as you will lose all differences between these two cells, and take the union as a truncated datapoint.

2.2 Aim 2: “Cancer Stem Cell” model vs. “Yeast culture” model

Using the mutational information of a sample of the cancer population, we can attempt to differentiate between these two models based on mutation analysis, both within a local population and more globally. We will do this using software such as Phylip to create the most parsimonious, rooted phylogenetic tree (the root being the normal tissue sequence) using non-ubiquitous mutations as defining traits. This approach will allow us to validate the cancer stem cell theory: if we find that the Pruffer
sequence of these trees infer a statistically higher rate of multifurcation than expected from typical single cell selection models, we may be able to claim this cancer grows through cancer stem cells. That is, we would see more “star shaped” phylogenetic trees as opposed to the bifurcation events that are usually associated with symmetric cell division [13]. This would point to persistent, underlying progenitor or stem cells. This analysis could be done on both local and global populations, depending on the amount of variation we are able to catalog.

2.3 Aim 3: Driver mutation emergence
Simply by sequencing many different cells separately, we can tease out the order which genes known to be cancer related were mutated. By sequencing the normal, somatic tissue, we know the root of our trees created in Aim 2. We can then look along these trees to see which mutations that are in known cancer-related genes are shared by the most branches. If a driver mutation is present in an outgroup of a second driver mutation, than this mutation occurred first. This has been a very hot topic of research, but up until recently, has only been done using very broad brush strokes, limiting the amount of temporal information that can be obtained [14, 15]. By using this technique you could theoretically determine the exact order of driver mutations in the primary tumor. Unfortunately, this would be more of an observation than true actionable data, as information about the progression of a single tumor can’t be extrapolated to general rules about tumor formation. However, this information may provide detailed clues and insight into drivers of cancer at each stage of its development. Additionally, there may be enough mutational events in between different driver mutations to provide rough estimates about length of evolutionary time it took to develop each mutation.

2.4 Aim 4: Determining the microenvironment’s effect on selection
Lastly, one of the most interesting questions may be how a cancer cell’s microenvironment may alter the selective advantages of particular driver mutations. Obviously, a cancer cell’s “environment” consists of many variables, but two important ones are nutrient supply and the tumor/normal boundary. Using the 3D reconstruction of the tumor, we can determine both the proximity of the tumor/normal boundary and the amount of relative nutrient availability based on location and diameter of nearby blood vessels. Assuming that cancer cells don’t move around that much in a tumor mass (which may be a poor assumption depending on stage of the tumor) we would expect local populations to be more closely related the closer they are to one another. However, if there are similar selective pressures on cells in similar environments, than populations on the surface of the tumor may be more genetically similar to one another than to a population in the center, even if they were collected on opposite sides.

3. Cost analysis, Short-comings and Future directions

3.1 Projected Cost
Though not very labor intensive, this will be a very expensive project due to the large amounts of DNA sequencing and synthesis required. Of course, as with any project, it is difficult to accurately predict overall cost, especially with sequencing rapidly becoming less expensive. However we will try to calculate a upper estimate. Assuming it takes $10,000 to sequence the human genome to 30-40X average coverage, and assuming we get the entire exome from every cell we measure, if we measure 300 cells in each local population, and 100 different local populations the total sequencing would be equivalent to 300 human genomes, or $3,000,000. For our synthesized DNA, the maximum length of each oligonucleotide would be around 40bp. If the synthesis of each of the 2,000 different barcoded DNA was done entirely independently, and we used a typical commercial vender number (Invitrogen)
of $0.38/base it would cost around $30,000. The cost of the microfluidics fabrication would probably be minimal, but we could add as much as another $50,000 if we had to buy a new Flow Cytometer for modification and probably at least another $100,000 for trial and error during experimentation before we are ready for sequencing of a real primary tumor. So overall price tag comes in around $3.35 million. However, if sequencing drops to the $1,000/genome threshold, that number would drop under $1 million.

3.2 Short-comings and Future directions

Besides the biggest short-coming listed in the first part of this section (i.e. the price tag), this study could be improved in many ways. A more targeted approach towards cancer relevant genes might provide more insight into driver mutations without the exhorbitant price tag. This protocol is very capable of such an approach and would only require a couple simple extra steps during bead production. On the opposite side of the spectrum, a look at whole genome sequencing would perhaps provide more population and evolution information without having to sequence so many individual cells, as you would be able to detect much more variation in the non-coding region from cell to cell. This may be more difficult as there currently isn’t a technique to obtain the full sequence of a human genome using a single cell. However, if this technology becomes available, this would be a great application. An important point/ flaw of this study is that with all this sequencing, we are still looking at a what amounts to a single cancer event; we are simply improving the resolution at which we observe such an event. In that since, it is important to realize that as cancer is inherently a random and unpredictable series of events, more datapoints (read tumors) are required in order to justify conclusions about general cancer biology. Single cell transcriptome sequencing of several other patient tumors would provide more evidence for what we will observe. Additionally, single cell transcriptome analysis from multiple tumors resected from the same individual would delve further into the differences between pre- and post-metastatic cancer than sequencing studies already done on the topic.

4. Conclusions

Transcriptome analysis of a tumor on the single cell level will provide a fundamental understanding about its evolution, heterogeneity, organization, and regiospecific dependencies. Though it is a large project, most of the infrastructure and technology is available for such a study, and the information it would provide would transverse many areas of cancer biology and would be immediately clinically relevant. For example, when patients are treated with oncogene specific drugs, tumors expressing that oncogene initially respond very well, decreasing in size and going into remission. However, in most cases a drug-resistant tumor will re-emerge very quickly, and people have mapped the drug resistance to either specific mutations on the targeted gene, or oncogenic mutations downstream in the same pathway [16]. If we can gain an understanding about how often coding mutations in a drug-targeted oncogene appear before the drug is even introduced, we may have a better understanding on the drug combinations required to suppress the emergence of drug resistant tumors. Furthermore, if we find some spatial bias to certain oncogenes, it may suggest un-discovered dependancies that go untreated in current cancer treatments, allowing some cells to survive. This novel technique builds upon previous single cell nucleic acid research to provide the data necessary for a population analysis of a primary tumor. By determining the heterogeneity of a cancer population, we will not only have a better understanding of cancer biology, but we may be able to provide more effective personalized medicine through informed combinational therapy.
References

Figure 1: Emulsion based PCR method to isolate single cells and beads together in water droplet in an oil bath [9]. A. shows how the microfluidics device is set up to accomplish this. B. How their PCR was set up; ours is set up slightly differently.
Figure 2: Basic bead design. Each bead will have a huge number of poly-T primers coating it, followed by a unique barcode, a universal primer sequence, and a sequence of random DNA to offset the universal primer from the bead.