Amplification and thrifty single molecule sequencing of recurrent somatic structural variations

Anand D Patel, Richard Schwab, Yu-Tsueng Liu, et al.

Genome Res. published online December 4, 2013
Access the most recent version at doi:10.1101/gr.161497.113

Amplification and thrifty single molecule sequencing of recurrent somatic structural variations

Anand Patel[1,*], Richard Schwab[2], Yu-Tsueng Liu[2,*], and Vineet Bafna[1,3,*]

[1]Bioinformatics and Systems Biology Program, University of California San Diego, 9500 Gilman Dr., La Jolla 92093

[2]Moores Cancer Center, 3855 Health Sciences Dr., La Jolla 92093

[3]Department of Computer Science, University of California San Diego, 9500 Gilman Dr., La Jolla 92093

December 2, 2013

Abstract

Deletion of tumor suppressor genes as well as other genomic rearrangements pervade cancer genomes across numerous different types of solid tumor and hematologic malignancies. However, even for a specific rearrangement, the breakpoints may vary between individuals, such as the recurrent CDKN2A deletion. Characterizing the exact breakpoint for structural variants (SVs), has utility as patient specific tumor biomarkers. We propose AmBre (Amplification of Breakpoints), a method to target SV breakpoints occurring in samples composed of heterogeneous tumor and germline DNA. Additionally, AmBre validates SVs called by whole exome/genome sequencing and hybridization arrays. AmBre involves a PCR-based approach to amplify the DNA segment containing a SV's breakpoint and then confirms breakpoints using sequencing by Pacific Biosciences RS. To amplify breakpoints with PCR, primers tiling specified target regions are carefully selected with a simulated annealing algorithm to minimize off-target amplification and maximize efficiency at capturing all possible breakpoints within the target regions. To confirm correct amplification and obtain breakpoints, PCR amplicons are combined without barcoding and long-read sequenced simultaneously using a single SMRT cell. Our algorithm efficiently separates reads based on breakpoints. Each read group supporting the same breakpoint corresponds with an amplicon and a consensus amplicon sequence is called. AmBre was used to discover *CDKN2A* deletion breakpoints in cancer cell lines: A549, CEM, Detroit562, MOLT4, MCF7 and T98G. Also, we successfully assayed *RUNX1-RUNX1T1* reciprocal translocations by finding both breakpoints in the Kasumi-1 cell line. AmBre successfully targets SVs where DNA harboring the breakpoints are present in 1:1000 mixtures.

Introduction

Cancer develops through a series of genetic mutations, with tumor cells acquiring pernicious mutations that eventually lead to metastatic disease. The DNA mutations contributing to oncogenesis are not limited to point mutations, but include large chromosomal rearrangements, duplications, and deletions. It has been suggested that recurring mutations are the likely drivers for cancer, and might be viable

*to whom correspondence should be addressed. A Patel adp002@ucsd.edu, YT Liu ytliu@ucsd.edu, and V Bafna vbafna@cs.ucsd.edu

biomarkers for disease detection and prognosis. For instance, a translocation occurs between chromosome 21 and 8 that fuses *RUNX1* and *RUNX1T1* genes in 12% of acute myeloid leukemia (AML) cases (Xiao et al. 2001). The fusion results in a chimeric oncoprotein. The chimeric protein contributes to initial leukemia cell growth mostly through transcriptional repression of wild-type *RUNX1* targets (Downing 1999). Alternatively, the loss of DNA may also contribute to cancer progression. For example, many human cancers frequently delete chromosome 9p21-22 locus containing *MTAP*, *CDKN2A*, and *CDKN2B* genes. The locus encodes INK4 proteins ($p15^{INK4B}$, $p16^{INK4A}$) that inhibit cyclin-dependent kinases, CDK4 and CDK6, and $p14^{ARF}$, which inactivates MDM2 and thereby regulating TP53. Thus, expression of these proteins is responsible for G1 cell-cycle arrest and independently signaling apoptosis (Wessely 2010; Kim et al. 2012). Homozygous deletions frequent the 9p21-22 locus, in particular *CDKN2A*, which encodes both $p16^{INK4A}$ and $p14^{ARF}$, as the single event diminishes expression of multiple proteins each with unique tumor suppressor activity.

In a clinical setting, driver DNA lesions can be used to (a) detect tumor DNA in individuals and (b) monitor tumor burden during or after treatment. Michor et al. (2005) and Bartley et al. (2010) demonstrated how identification of *BCR-ABL1* gene fusion at the DNA level in leukemia patients leads to a more sensitive test for measuring tumor burden than current *BCR-ABL1* mRNA tests. Measuring changes in tumor burden during therapeutic treatment is critical for checking therapy effectiveness and deciding to continue treatment. Their approach focuses on the frequent translocation of *BCR-ABL1* in leukemia and has not been applied to solid tumors. In a more recent study, circulatory biomarkers were assessed in their ability to monitor metastatic breast cancer (Dawson et al. 2013). The researchers applied a variety of sequencing methods to identify point mutations in *PIK3CA* and *TP53* and other somatic structural variations for use as circulatory tumor DNA markers. They found that circulatory tumor DNA had the highest correlation with tumor burden and greater dynamic range than current standard of care CA 15-3 biomarker and circulatory tumor cell counting.

These studies all focused on tumor burden monitoring *after* the specific lesion had been fully characterized. While monitoring is easy for point mutations and structural variants with known breakpoints, it is very difficult when the breakpoint of the structural variation is not known. At the same time, large variants are potentially much more specific for tumor detection and monitoring, and a test that could identify them reliably would have higher sensitivity for monitoring tumor burden. Reliable and sensitive identification of breakpoints in tumor DNA could also serve as a diagnostic for early detection.

Whole genome sequencing experiments (analyzed with appropriate tools like BreakDancer (Chen et al. 2009), Pindel (Ye et al. 2009), and SVDetect (Zeitouni et al. 2010)) have the potential to identify point mutations and structural variations in individual samples. However, clinical tumor samples are a mixture of tumor cells and normal cells and require ultra-deep sequencing to analyze tumor DNA.

Therefore, current approaches apply ultra-deep sequencing after targeted amplification of select genes (Harismendy et al. 2011). Unfortunately, these methods are unable to reliably identify structural variation with uncertain breakpoints. Alternatively, DNA hybridization microarrays (SNP-arrays), which are still widely used in clinics, are capable of calling copy number variation, from which deletions and gene amplifications can be inferred. However, the technology is only reliable with homogeneous samples and only reports low resolution boundaries estimates (Greenman et al. 2010), insufficient for performing tumor burden monitoring assays. Thus, a challenge remains how to detect DNA markers, specifically, somatic structural variations, in a complex patient sample containing a mixture of tumor DNA and germline DNA. This is particularly challenging when the exact breakpoints are needed for quantitative DNA assays.

To identify unknown DNA breakpoints associated with known translocations and deletions, we describe a pipeline *AmBre (Amplification of Breakpoints)*, which builds upon the PAMP approach (Liu and Carson 2007). PAMP is a PCR assay, developed to selectively amplify the tumor DNA sequence

containing a structural variation. To illustrate how PAMP works, consider a deletion on chr9 (*CDKN2A* locus) with unknown breakpoints located around the *CDKN2A* gene. Illustrated in Figure 1, a tiling of evenly spaced forward (blue arrows) primers and reverse primers (red arrows) are selected around the *CDKN2A* gene. The spacing between primers is approximately 1kb apart. The innermost forward and reverse primers are distantly spaced such that they will not amplify sequence from germline DNA.

All tiling primers are used in a single multiplex PCR. Any *CDKN2A* deletion in the tumor DNA will lead to a forward and reverse primer being proximally located (< 2kb) on the tumor DNA, resulting in a targeted DNA amplification of the tumor DNA harboring the deletion, but not germline DNA. This strategy takes advantage of polymerases having a limited amplifying length and genomic rearrangements within tumor DNA resulting in novel adjacencies of germline DNA sequences for selective and sensitive amplification of tumor DNA over germline DNA.

Although it has potential, PAMP has challenges. In the multiplexed reaction, all primers must be evenly spaced so as to amplify any deletion in the region and primer pairs cannot dimerize. In a large (say, 100kb) region, this implies we need to find a design of 100 applicable primers from a large candidate set of over 5000 potential primers. An exhaustive search of all candidate primer combinations is infeasible (5000 candidate primers and 50 to 100 primers desired would result in searching $\sum_{50 \leq i \leq 100} \binom{5000}{i} \approx 10^{211}$ combinations). Bashir et al. formulated PAMP primer tiling as a computational problem and defined a cost associated with each subset of candidate primers (Bashir et al. 2007). Furthermore, the authors showed simulated annealing (Kirkpatrick 1984) could efficiently find low cost PAMP primer designs for contiguous breakpoint regions. Even with these improvements, PAMP is limited to recurrent structural variations where breakpoints appear in short breakpoint regions (< 40kb), as a large number of primers in a single reaction inevitably leads to loss of sensitivity with off-target DNA synthesis and increased spurious primer-primer interactions. Finally, PAMP detects the amplified product and identifies breakpoints via DNA hybridization arrays (Bashir et al. 2009) which had the additional challenge of designing probes that match the primer designs.

Overview of AmBre and Results. AmBre resolves these issues with a three phase approach (Figure 2). The first (*AmBre-design*) involves a revised computational approach to designing multiplex primers on discontiguous DNA regions ignoring regions known to not contain breakpoints. This requires some changes to the optimization function and results in a more flexible design with better performance on sparse regions. The output of this phase is a collection of primers that can be mixed in a single multiple primer reaction.

In the second, experiment phase (*AmBre-amplify*), long range PCR amplifies target amplicons, which reduces the number of primers required in a single reaction. For example, PAMP, using their proposed traditional PCR, would require 600 primers to cover a 600kb region, with over $180,000$ putative interactions. In contrast, to cover the same region, AmBre would need < 100 primers with only $5,000$ possible interactions, which improves reliable amplification from proposed designs. In AmBre, the amplified products are sequenced using the Pacific Biosciences RS (PacBio) platform (English et al. 2012). Our analysis allows us to mix the amplicons prior to sequencing, with computational separation of breakpoints in the third phase.

The final, computational phase (*AmBre-analyze*) involves a customized analysis of sequenced reads to identify DNA breakpoints for each tumor genome. The analysis involves clustering of split mapped reads followed by error correction, and sequence reconstruction around the breakpoint regions. We demonstrated that AmBre can successfully detect targeted structural variations (potential tumor DNA biomarkers) by identifying *CDKN2A* deletion breakpoints in the cancer cell lines A549, CEM, Detroit562, MCF7, MOLT4, and T98G. AmBre resolved breakpoints for MCF7 and T98G, which had not been previously discovered by other studies. Furthermore, AmBre easily extends to identify translocations and inversions, which is demonstrated here with *RUNX1-RUNX1T1* translocation in the cancer cell line

Kasumi-1.

Results

Designing primers

The input to AmBre-design is a collection of genomic intervals for the forward region, denoted by F, a collection of genomic intervals for the reverse region (R), and parameter d. The output is a collection of forward primers in F and reverse primers located in R spaced apart by approximately d. AmBre-design has the following steps:

- Candidate primer generation from target breakpoint regions, where oligonucleotides are selected according to thermodynamic properties. Primers with significant self-dimerization are eliminated. Primer pairs that are likely to dimerize or cause off-target amplifications are marked as incompatible (Methods).
- The list of candidate primers and incompatible primer pairs are used to design an optimal set of primers based on considerations outlined below.

Denote a primer design, P as a subset of candidate primers numbered according to the order of genomic start locations $l_1, l_2, l_3, \ldots l_n$. Let set E denote incompatible primer pairs. We associate a cost $C(P)$ with each design, and seek to find designs with minimum cost. Our formulation of cost differs from Bashir et al. to accommodate sparser primer designs and targeting discontiguous regions (see S1). The parameter d is set to be half the maximum feasible PCR amplicon size. Thus, for long-range polymerases used here, we use $d = 6500$, corresponding to a desirable amplicon size ≤ 13kb. The cost of the design is a sum of incompatibility costs for each pair, and coverage costs.

For the coverage, let $\Delta_i(P) = l_{i+1} - l_i$ denote the gap between adjacent pairs. If $\Delta_i(P) > d$, we run the risk of the product being too long to be amplified. On the other hand, if $\Delta_i(P) << d$, we have a design with extra primers that greatly decrease the efficiency of the reaction. Let parameter ρ, with $0 < \rho \leq 1$, describe a target density $1 + \rho$ of primers every d bp, corresponding to a primer every $\frac{d}{1+\rho} \simeq (1-\rho)d$ bp. Ideally, the distance between adjacent primers is bounded by $(1-\rho)d \leq \Delta_i(P) \leq d$. A design is penalized if the distances violate these constraints. Formally,

$$C(P) = \sum_{(i,j)\in E} w_p + \sum_i \max\{\Delta_i(P) - d, 0, (1-\rho)d - \Delta_i(P)\} \quad (1)$$

Experiments revealed that even a single incompatible pair severely diminishes the multiple primer reaction (Bashir et al. 2007). Therefore, we set $w_p = \infty$ for our designs. We empirically choose $\rho = 0.2$. Similar to Bashir et al., simulated annealing is used to find low cost primer designs by applying our cost function (Figure 3 and Methods). The algorithm explores the large space of all primer designs by initiating a random primer subset and improving the primer subset with iterative addition or removals of primers. Since the algorithm involves randomization and has parameters governing convergence to low cost designs, simulated annealing is repeated multiple times under different rates of convergence. The lowest cost primer design from all simulated annealing runs is used as the final primer tiling design (Figure 3).

Design results: To test Ambre-design, we analyzed cell-line copy number data to identify a large clustering of deletions in the *CDKN2A* region (Greenman et al. 2010). We identified a 380kb region surrounding the *CDKN2A* gene, 230kb upstream and 150kb region downstream of *CDKN2A* that captures

breakpoints in 55 of the 109 *CDKN2A* deletion cell-lines considered. We chose $d = 6500$, as 13kb products can be reliably amplified with LongAmp Taq DNA polymerase (New England Biolabs, NEB).

The candidate primer generation and primer filtering stages resulted in 5181 candidate primers. As shown in Figure 3a, the candidate primers are uniformly spread across breakpoint regions suggesting good tiling primer designs may exist. The simulated annealing algorithm is repeated for 12 different rates of convergence with the fastest convergence rate having a 10 minute average runtime and slowest convergence rate having a 864 minute average runtime (Figure S2). When $d = 6500$, the lowest cost solution (AMBRE-68) requires only 68 primers with 99.99% in silico capture of simple *CDKN2A* deletions that may occur in the 380kb breakpoint region (Figure 3b).

Sequencing amplified sequences harboring SVs

Sequencing the AmBre-amplify DNA confirms capture of *CDKN2A* deletions. We used PacBio RS technology due to its long reads, ideal for structural variation calling, and throughput, appropriate for medium sized experiments. Using computation, we correct for the high inherent error in PacBio sequencing.

Furthermore, if different samples do not share breakpoints (for example, all amplicons are of different sizes and amplify from different primer pairs within the design), the samples can be mixed and sequenced on a single run without additional barcoding. We employed this strategy with *CDKN2A* deleted samples on a single SMRT cell and relying on computation to deconvolute the breakpoints.

Define a *breakpoint* as a pair of disjoint coordinates a and b on a reference, and a non-template sequence s (of length ℓ) such that the sample sequence brings a and b together, separated only by the insertion of s. The objective of Ambre-analyze is to take as input a collection of PacBio sample sequences aligned to the reference genome and output a collection of breakpoints along with the sequence around each breakpoint. The code for this tool is stand-alone and can be used in the analysis of PacBio reads for SV detection. Ambre-analyze works by (a) alignment trimming (defined below), (b) breakpoint clustering of fragments, and (c) consensus sequence generation around each breakpoint (Figure 2, see Methods).

Alignment trimming: Denote a local alignment (Chaisson and Tesler 2012) as a pair of intervals from the fragment and reference that can be aligned with a small number of edits. A split mapped fragment F supports a breakpoint (a, b, s) with two local-alignments (denoted as $(F_a, G_a), (F_b, G_b)$). In the ideal case, G_a ends at a, and G_b begins at b, while the fragment segment between F_a and F_b is exactly the inserted sequence s (Methods). However, in real data, a fragment can span multiple breakpoints, sequence errors can result in spurious incorrect alignments, and the alignments output by standard tools like BLASR will have inaccurate boundaries. Specifically, inaccurate boundaries might result in overlapping consecutive segments F_a, F_b. AmBre-analyze resolves these errors by choosing the optimal alignment segments covering the fragment F. For a fragment F, the input is a chain of local alignments $\mathcal{F} = (F_a, G_a), (F_b, G_b), \ldots$. The output is a subset $\mathcal{F}' = (F'_a, G'_a), \ldots$ of $\mathcal{F}$, with alignment boundaries trimmed so (1) none of the fragment segments $F'_a, F'_b, \ldots$ overlap, (2) the number of distinct alignments is minimized, and (3) most of fragment F is covered. The second and third objectives reinforce the notion that a typical fragment covers a small number of breakpoints and is mostly well-aligned except for non-template insertion sequence. The first objective helps to narrow down the breakpoint coordinates. To clarify, consider a trimmed reference interval G'_a that ends at x and a consecutive interval G'_b beginning at y, while the gap between corresponding fragment segments is L. Then, we expect that $a > x$, $b < y$, and

$$L \simeq \ell + (a - x) + (y - b)$$

Thus, the fragment constrains the location of the breakpoint (a, b) to lie in a small region between x, y. In the next section, we will use information from multiple fragments to further narrow the breakpoint location. Given these three distinct objectives, the alignment trimming algorithm works by combining them into a single objective function, and uses a dynamic programming approach to identify the optimal trimming (Methods).

Fragment clustering: Consider a two dimensional representation of the genomic space with F and R being the vertical and horizontal axes, respectively. In this representation, a true breakpoint (a, b) is represented by a point, and each split-mapped read (x, y, L) is represented by a triangle of possible breakpoints (a, b) that satisfy $(a - x) + (y - b) \leq L$ (Methods). Multiple reads supporting the same breakpoint represent multiple triangles whose intersection reduces the uncertainty in breakpoint determination. Furthermore, if reads from multiple AmBre-amplify experiments are combined, the split-mapped reads will cluster according to overlap, revealing breakpoints for each experiment sample. We develop a fast, customized method to recover the aggregated read clusters for each breakpoint (Methods). The method took 2.5 min seconds on a single desktop core to analyze all local alignments from $52,000$ reads from a single PacBio SMRT cell experiment.

Consensus sequence determination: Predicted amplicon sequences are generated from the breakpoint estimates. In turn, these templates are supplied as reference sequences into PacBio's SMRT Analysis Resequencing Protocol. The analysis protocol calls consensus amplicon sequences by correcting the predicted templates.

Identifying *CDKN2A* deletion given DNA break clustering

AmBre exploits the fact that variable breakpoints aggregate along fragile regions of the chromosome by designing primers around the fragile regions. We used this idea to produce a single design for five cancer cell lines: A549, CEM, Detroit562, MCF7, and T98G. Breakpoints were estimated by copy number changes for four cancer cell-lines (A549, CEM, MCF7, and T98G) from SNP-array data (Greenman et al. 2010) (Table 1 and Figure S3) and the breakpoint was given for a fifth cell line (Detroit562) from prior studies. The error in breakpoint estimation for SNP-array data is roughly 10kb. Thus, to generate cluster target regions, each breakpoint estimate was expanded to be a 10kb interval and overlapping intervals were merged. This created four regions (F) upstream of *CDKN2A* and three downstream regions (R), and the target regions were used as input for AmBre-design ($d = 6500$bp). AmBre-design output a high quality 16 primer design (AMBRE-16) with primers spaced apart by approximately 6kb to cover the 100kb input region. The design was used by AmBre-amplify on DNA samples from each cell line. The experiment successfully amplified DNA from each cell line (Figure S4), where each line produced a unique sized amplicon even though each reaction uses the same set of 16 primers.

PCR products were mixed together for simultaneous preparation and sequencing on a single SMRT cell. The sequence data was the input to AmBre-analyze. The tool BLASR (Chaisson and Tesler 2012) identified 52k alignable fragments. After clustering in AmBre-analyze, we retrieved deep coverage of every breakpoint (although with six clusters instead of five; see below), with A549 having the lowest coverage of 400 fragments and CEM having the highest coverage of $18,000$ fragments (Figure 4). The difference in coverage is due to different amplicon sizes, where shorter amplicons are easier to load onto a PacBio SMRT cell than longer amplicons. Newer PacBio instrumentation is expected to normalize for this sequencing bias (Mason and Elemento 2012).

AmBre-analyze generated consensus sequence for each cell line. A549, CEM, and Detroit562 breakpoints (Figure S5-6) are concordant with previous studies (Kitagawa et al. 2002; Sasaki et al. 2003; Bashir et al. 2009). The A549 harbors a complex structural variation where in addition to a large

DNA segmental loss including *CDKN2A*, there is a 325bp internal inversion occurring at the deletion breakpoint junction. AmBre-analyze resolved the complex event as two separate breakpoints. The A549 amplicon template was created by ordering the reference segments corresponding to the two breakpoints. After template refinement, the A549 amplicon sequence matched the sequence found by Bashir et al. (2009).

To our knowledge, the nucleotide sequence for MCF7 and T98G had not been previously characterized in spite of previous efforts, including whole genome sequencing of the MCF7 cell line. The ease of the discovery in our experiment attests to the value of a targeted approach to SV detection. Both MCF7 and T98G sequences were confirmed using Sanger sequencing. Interestingly, the SNP-array estimate for MCF7 breakpoint is 15kb away from the AmBre detected breakpoint. The difference may be due to SINE and LINE repeats that mark the region of the upstream MCF7 breakpoint, a fact confirmed by the Sanger reads (Figure S5). Repetitive sequences are known to confound structural variation analysis and possibly explains why previous genome sequencing studies of MCF7, have not annotated the *CDKN2A* deletion breakpoints (Hampton et al. 2009; 2011).

We analyzed the physical properties of DNA around the breakpoints of *CDKN2A* deletions using the BreakSeq pipeline (Lam et al. 2009). All five deletion events were predicted to result from non-homologous end joining (NHEJ). According to Lam et al. (2009), a characteristic of NHEJ is lower DNA duplex stability near the breakpoints of a structural variation. They assessed DNA duplex stability based on predictions of helix stability (average dissociation free energy of overlapping dinucleotides) and DNA flexibility (average twist angle of overlapping dinucleotides). We found no strong association to lower DNA duplex stability in *CDKN2A* deletion breakpoints, albeit we are analyzing much fewer structural variations (Figure S7). Alternatively, Kitagawa et al. (2002) suggested that the *CDKN2A* deletion in CEM is due to illegitimate V(D)J recombination, which is evidenced by V(D)J recombination motifs discovered near the deletion breakpoints.

Characterizing *CDKN2A* deletion assuming no DNA break clustering

Also, AmBre applies to contiguous break regions. We developed a 68 primer design to capture *CDKN2A* deletions with breaks in a 380kb region (AMBRE-68, Figure 3).

In AmBre-amplify experimentation, we observed that the high amount of multiplexing, and larger amplicon lengths ($>$ 4kb) reduce amplification efficiency. Using all AMBRE-68 primers in a single reaction resulted in amplification of only the 2.2kb A549 *CDKN2A* deletion loss (data not shown). To mitigate this effect, sub-sampling of primers from a design and performing multiple reactions per sample using different primer sets improved amplification results. To test whether the AMBRE-68 primers selected were viable at some level of subsampling, we sampled the nearest forward and reverse primer in AMBRE-68 to each *CDKN2A* break in cell lines: A549, CEM, Detroit562, MCF7, MOLT4, T98G. This resulted in a nine primer subset, which again captures the *CDKN2A* deletion in each cell line. Of these cell lines, five lines resulted in amplicons ranging in lengths from 2.2kb to 7.5kb (Figure 5). The Detroit562 breakpoints did not fall within the target breakpoint region given to AmBre-design and the expected amplicon size using the closest AMBRE-68 primers is 16kb. Thus, Detroit562 did not amplify with the nine primer subset. For each remaining cell line, the observed amplicon length matched the spacing between *CDKN2A* breakpoints and nearest primers in AMBRE-68 design. Thus, a universal primer design divided into multiple primer subset experiments can be used to identify SVs.

Characterizing *RUNX1-RUNX1T1* translocations

AmBre also captures more complex rearrangements like interchromosomal translocations. This was demonstrated with an experiment characterizing *RUNX1-RUNX1T1* gene fusion, the result of a translo-

cation between chr21 and chr8. In the tumor genome, breakpoint ends lie within a 30kb region $chr21:36,205,000-36,235,000$ in the *RUNX1* intron, and a 55kb region $chr8:93,030,000-93,085,000$ in *RUNX1T1*, and the derivative chromosome 8 (Der8) encodes a fusion oncoprotein. In some cases, the translocation is balanced and also generates a fusion of *RUNX1T1-RUNX1* on a derivative chromosome 21 (Der21). To capture the translocation producing Der8, we used AmBre to design 10 reverse primers in the *RUNX1* region and 18 forward primers in the *RUNX1T1* region with ~ 3kb primer spacing. Similarly, to capture Der21 breakpoints, 10 forward and 19 reverse primers were designed in the *RUNX1* and *RUNX1T1* regions, respectively. Recall, a ~ 3kb primer spacing supposes the maximum product size is approximately 6kb. The primer designs were tested on Kasumi-1, which carries the balanced translocation with both Der8 and Der21 breakpoints characterized (Xiao et al. 2001). AmBre spaced the primers in the two regions unaware of the true Kasumi-1 breakpoints and we assayed the Der8 and Der21 chromosomes in two independent reactions using the respective 28 and 29 primers. The primers closest to the breakpoints produces a 3.5kb and 2.7kb amplicon from Der8 and Der21, respectively (Figure 6). Both reactions resulted in a strong signal and virtually no background noise, despite there being close to 30 primers in each reaction.

Furthermore, we investigated subsampling of primers and efficacy in generating longer amplicons. For each primer design, we divided the forward and reverse primers based on index parity when sorted by chromosome position. Thus, there are four primer sets: forward odd (FO), forward even (FE), reverse odd (RO), and reverse even (RE), with primers spaced by approximately 6kb. The forward and reverse primer sets make four combinations: FO $\cup$ RO, FO $\cup$ RE, FE $\cup$ RO, and FE $\cup$ RE, primers for capturing target breakpoints. These combinations can be treated as four new primer designs, each with a maximum product size of 12kb, but half as many primers. This gives us the opportunity to assess amplification efficiency across different amplicon lengths and primer density per reaction using the same DNA template. In the original 28 primer design, the Kasumi-1 breakpoints for Der8 were generated by the sixth forward and ninth reverse primer. Thus, trying the 14 primer designs FE $\cup$ RO, FO $\cup$ RO, and FO $\cup$ RE produces 3.5kb, 6.8kb, and 10.1kb amplicons (Figure 6). Similarly, the 29 primer design for Der21 was subsampled into three reactions. Each reaction resulted in a strong signal band at the expected amplicon size and all six amplicon were confirmed to span the Der8 and Der21 breakpoint via Sanger sequencing (Figure S8). From each reaction, a general trend of better amplification for shorter amplicon lengths is observed. However, there was no significant difference in amplification efficiency between using all primers and half the primers to generate the shortest amplicons. Longer amplicons had strong signal, but weaker false products were visible. This effect is not seen with the shorter amplicons and false products may be more prevalent in reactions with greater number of primers and longer amplicons.

Dealing with tumor heterogeneity

The AmBre assay, unlike other methods, can target DNA with a SV in the context of high background of germline DNA. This feature is important for sensitive detection of tumor DNA and establishing a patient specific tumor DNA marker for monitoring tumor burden. We successfully amplified a 2.2kb *CDKN2A* deletion sequence from A549 and a 3.6kb deletion sequence from MCF7 starting with A549 and MCF7 genomic DNA mixed with HEK genomic DNA (Figure S9). Each reaction starts with a heterogeneous mixture of approximately $400ng$ with tumor to wild-type gDNA mixture ratios of 1:1, 1:10, 1:100, 1:1000. In a realistic application for AmBre, each reaction contains numerous primers where only 2 primers are responsible for amplification. In the experiment, each reaction contains 16 primers sampled from AMBRE-68 around *CDKN2A* deletion breakpoints for each cell line. In the heterogeneity experiment of A549, strong amplification is observed for each mixture ratio whereas for MCF7 there is clearly a reduction of amplification efficiency as the fraction of starting cancer cell line gDNA decreases (Figure S9). Amplification of longer amplicons with AmBre in the complex gDNA sample is also possible,

however with reduced sensitivity (Figure S10). The sensitivity for the AmBre assay is largely dependent on expected amplicon length. *CDKN2A* deletion breakpoints corresponding to a smaller amplicon in a particular AmBre primer design are more easily amplified.

Discussion

AmBre addresses the challenge of highly sensitive SV targeting in complex DNA mixtures. This is accomplished with a careful design of tiling primers that enables amplification of DNA harboring the SV if present in the mixture and a specialized PacBio analysis pipeline to confirm SV breakpoints. AmBre was used to discover breakpoints associated with *CDKN2A* deletion in cancer cell-lines MCF7 and T98G. In addition, we demonstrated amplification occurs even in a complex DNA mixture where 1 in every 1000 DNA molecules contain the *CDKN2A* deletion. These features of AmBre are clinically important. A SV breakpoint specific to a cancer patient could serve as a personalized biomarker, where a quantitative PCR assay could accurately measure the patient's tumor burden (Michor et al. 2005; Bartley et al. 2010). With advancements in microfluidics and droplet PCR, quantifying $1-3$ copies of tumor DNA in a complex sample is possible (Hatch et al. 2011).

If the problem is to simply observe a SV, there are numerous high-throughput methods; SNP hybridization arrays (SNP-array), whole exome sequencing (WES), and whole genome sequencing (WGS). However, these methods are not ideal for a clinical application in tumor burden monitoring. SNP-arrays and WES give copy number read outs of DNA, which hint at the presence of SVs and a low resolution estimate of corresponding breakpoints. Without a high accuracy breakpoint estimate, a quantitative PCR assay specific to tumor DNA cannot be designed. WGS is capable of breakpoint calling, but would require an exorbitant amount of deep sequencing to capture SVs occurring in a low fraction of DNA. Harismendy et al. (2011) reported the extent of this sequencing challenge, where more than 1500X coverage of cancer mutational hotspots (71.1kb region) was necessary to capture single nucleotide variants (SNVs) occurring with prevalence greater than 5% in the sample.

Therefore, a targeted approach for mutation detection is preferred to a high throughput untargeted mutation discovery for clinical practice. A high throughput method captures numerous SVs and SNVs where follow-up functional analysis is required for each mutation to determine its potential as a cancer driver or passenger mutation. Alternatively, there are numerous targetable SVs known to drive cancer progression, and are being used in clinical laboratories to confirm cancer diagnosis and guide therapy. The most notable example, CML patients with the *BCR-ABL1* translocation are treated with tyrosine kinase inhibitors. The patient's response to therapy can be reliably tracked by measuring tumor DNA containing *BCR-ABL1* gene fusion from blood samples (Michor et al. 2005; Bartley et al. 2010). Unfortunately, such success in tumor burden monitoring has not been observed for patients with solid tumors.

In this work, we present AmBre's application to capture *RUNX1-RUNX1T1* translocations in AML cases and *CDKN2A* deletions, which are prevalent in many types of cancer. Using the accompanying software, this approach can be easily extended to target other SVs, like *BCR-ABL1* in chronic myeloid leukemia, *EML4-ALK* in lung cancer, and *TMPRSS2-ERG* in prostate cancer. For *EML4-ALK* and *TMPRSS2-ERG*, DNA breaks within introns and rearrangement of the chromosome fuse the genes together, similar to *RUNX1-RUNX1T1* gene fusion. The remaining challenge for AmBre is a limited targetable breakpoint region. We presented a design capturing breakpoints falling within a 100kb and proposed a multiple primer subset strategy for encompassing a 380kb breakpoint region. Further development is necessary to capture SVs with breakpoints appearing in a greater than 1mb range. AmBre is a first step to sensitive tumor DNA monitoring test for solid tumors. Extending the approach with improvements of applying multiple primer designs to target the same SV or the use of microfluidic devices may lead to an ultra-sensitive assay capable of minimally invasive early cancer detection.

Methods

AmBre: Primer generation and filtering

Primer3 2.3.0 (Rozen et al. 2000) was used with long-range PCR specific parameters to identify 31bp candidate AmBre primers that were capable of amplification under the same thermocycling conditions. To minimize the chance of off-target amplification, candidate primers were aligned to the reference human assembly (GRCh37) using BLAT (Kent 2002). Define an *end-aligning* match as an exact match of length > 18 between the $3'$ end of a primer and an off-target location. Primers with greater than 10 end-alignments were removed as having a high chance for off-target amplification. Second, pairs of primers that have compatible end-alignments within a $2d$ long off-target region were marked as incompatible. Finally, each pair (including a self-pair) was tested for dimerization using MultiPlx (Kaplinski et al. 2005). Primers with self-dimerization (maximum binding energy ΔG less than -8.0 kcal/mol for any region) were removed and pairs with high binding affinity (maximum binding energy ΔG less than -4.0 kcal/mol for primer-primer 3end binding or -8.0 kcal/mol for any region of primers) were marked as incompatible. The remaining candidate primers and incompatibilities formed the input to AmBre primer selection.

AmBre: Primer selection with simulated annealing

A final AmBre primer design was selected from a filtered list of candidate primers (P_U) and primer-primer compatibilities. To compute an optimal primer design, a low cost P according to $C(P)$, we applied a simulated annealing (Kirkpatrick 1984) procedure. We computed an initial design P using a random subset of 6 primers. Define the neighboring design of P, $N(P)$, as either the removal of a single primer from P, or the addition of a single primer $p \notin P$ to P followed by removal of all primers $p' \in P$ s.t. $(p, p') \in E$. The simulated annealing procedure described in Algo 1 was used to compute low cost designs.

Algorithm 1 Simulated Annealing Algorithm

procedure SIMULATEDANNEALING(P_U, C)
  $P \leftarrow Random(P_U, 6)$ ▷ Initialize random primer set P with size 6
  for $t = T_1, T_2, T_3, \ldots$ **do** ▷ Iterate until design is stable
    $l \leftarrow Random(P_U, 1)$
    if $C(N_l(P)) < C(P)$ or $Random[0,1] < e^{-\frac{C(N_l(P)) - C(P)}{t}}$ **then**
  ▷ Move to neighboring design if improves or with probability proportional to extra cost and iteration
      $P \leftarrow N_l(P)$
    end if
  end for
  return P
end procedure

The temperature schedule, $T_1, T_2, T_3, \ldots$, linearly decreases depending on intercept and slope parameters m and b. Parameters tested for T were combinations of $m = 1, 0.1, 0.01, 0.001$ and $b = 10^4, 10^5, 10^6$. The maximum number of iterations ran was determined by the temperature schedule, $2b + \frac{b}{m}$, and constrained to be at least 10^6 and at most 10^8 iterations. Each parameter set was repeated 3 times. The lowest cost primer design of all runs was used as the final design. Figure S2 demonstrates convergence to design minima under different parameters of T for a target CDKN2A breakpoint region of length 380kb.

AmBre-analyze: PacBio sequence analysis

Alignment trimming: BLASR computed local alignments between the PacBio reads and human reference assembly were provided as input to alignment trimming. An alignment pair $(F_a, G_a), (F_b, G_b)$ with $a << b$ between a fragment F and reference G imply a breakpoint. The goal of alignment trimming is to trim the ends of each alignment for each fragment F, so that (a) each segment of F participates in a single alignment; and, (b) F is maximally covered.

We first remove local alignments encompassed by other alignments (e.g., 4 in Figure 7). We sort remaining alignments by their location on the fragment, so that alignment i starts before alignment j if and only if $i < j$. Let $b_s(i)$ and $b_e(i)$ denote the fragment breaks before the beginning and after the end of alignment i.

We represent alignments on a grid with alignments as rows and fragment positions as columns (Figure 7). An alignment is a series of breaks on the fragment (i.e. $(1, b_1)$ to $(1, b_5)$ in Figure 7). Alignments are chained together to cover a portion of F exactly once. To chain adjacent alignments, for each alignment j with an alignment i that terminates before j starts, add a jump from $(i, b_e(i))$ to $(j, b_s(j))$ (for instance $(1, b_e(1))$ to $(3, b_s(3))$). Also, for each alignment j overlapping an earlier alignment i on the fragment, add a jump from $(i, b_s(j))$ to $(j, b_e(i))$ (for instance $(2, b_e(3))$ to $(3, b_s(2))$) if i spans $b_s(j)$ and j spans $b_e(i)$. By this process, any alignment chain covers positions exactly once.

$$w[(i,u),(j,v)] = \begin{cases} \text{Aln}[i,u,v] & \text{if } i = j \\ \frac{1}{2}\left(\text{Aln}[i,u,v] + \text{Aln}[j,u,v]\right) + J(u,v) & \text{if } i \neq j \text{ and } i,j \text{ overlap from } u \text{ to } v \\ J(u,v) & \text{Otherwise.} \end{cases}$$

An alignment chain is scored by summing local alignment scores ($\text{Aln}[i,u,v]$ for alignment i for fragment coordinates u to v) and penalizing for jumps between alignments ($J(u,v)$ for alignment u to v). A high scoring alignment chain corresponds to trimmed alignments that aligns well and covers most of the fragment. The score of a chain is computed using dynamic programming. Let $S(j,v)$ denote the score of the best chain ending at (j,v). Then,

$$S(j,v) = \max_{(i,u)}\{S(i,u) + w[(i,u),(j,v)]\} \tag{2}$$

In the recursion, (i,u) is the start of alignment j, start of a jump to (j,v) (i.e. if $(j,v) = (3, b_e(2))$ then (i,u) could be $(2, b_s(3))$), or previous position on alignment j where a jump ends (i.e. if $(j,v) = (2, b_e(2))$ then $(i,u) = (2, b_e(1))$). By not computing the score for each alignment and fragment position on the grid, the optimal trimmed alignment chain is quickly found.

Along the maximum scoring chain, each jump, $(F'_a, G'_a), (F'_b, G'_b)$, represents a breakpoint estimate $(a, b, F'_j - F'_i)$. For example, the jump from 1 to 3 correspond with breakpoint estimate $(x_1, y_2, 6)$.

In this formulation, two alignments that overlap may contribute to a high score since the overlap segment is scored as the average of both alignment scores. Above, for a breakpoint estimate from overlapping alignments, we use boundaries around the overlap and do not resolve a tighter breakpoint within the overlap segment. Finding a tighter breakpoint estimate would require computing S for all breaks within overlap intervals, which is inefficient for thousands of fragments. In any case, the conservative breakpoint estimates are improved with downstream clustering and refinement steps.

Breakpoint clustering: Breakpoint estimates from all fragments supporting the same breakpoint are aggregated into groups using a sweep line algorithm. Sindi et al. (2009), applied a similar geometric approach to efficiently identify structural variations using discordant paired end reads.

For a breakpoint estimate (x,y,L), the true breakpoint junctions (a,b) in reference G lies between $x \ldots x+L$ and $y-L \ldots y$, respectively, subject to $a - x + y - b < L$. Here, we assume L, a spacing

length on F, is a reasonable estimate for breakpoint uncertainty on G and the effect of sequencing deletion errors at the breakpoint junction is minimal. On a $G \times G$ plane, each breakpoint estimate x, y, and L with the above constraints defines a triangle which contains the true breakpoint (a, b) (Figure 7 and Figure 4).

A line sweeps the plane and tracks when breakpoint triangles overlap along the sweep line. Here, a cluster is a collection of triangles where each triangle overlaps one or more triangles in the cluster. The consensus breakpoint (a, b) for the cluster is the mode of (x, y) estimates (see Figure 4).

Accounting for reverse orientation alignments: With a slight modification, we can account for alignments in the reverse complement orientation to capture structural variations with inversions and bidirectional PacBio reads. PacBio reads DNA amplicons in both directions, in particular, read in the forward direction produces an alignment chain $(F_x, G_x), (F_y, G_y)$ and in the reverse direction $(H_y, RC(G_y)), (H_x, RC(G_x))$ where RC reverse complements the sequence G. This is resolved by relabeling reverse complement alignments by a $-$, such that H supports the breakpoint $(-y, -x)$.

The relabeling applies naturally to the sequence analysis pipeline. Alignment-trimming relies only on projections on sequenced fragments and therefore does not change. Each DNA amplicon containing a breakpoint is associated with two breakpoint estimates, (x, y) generated from forward reading and $(-y, -x)$ from reverse reading.

In addition, the constraints of $-y$,$-x$,L in relation to $-a$, $-b$ remain the same, therefore both forward and reverse direction breakpoint estimates have the same triangle orientation on the $G \times G$ plane. All forward and reverse breakpoints are simultaneously recovered with the sweep line algorithm.

Using reverse complement alignments, breakpoints associated with inversions, like A549, are captured. In this case, a breakpoint corresponds with $(-x, y)$ and $(-y, x)$ or $(x, -y)$ and $(y, -x)$.

Breakpoint reconstruction: In the final step, predicted amplicon templates for each cluster are created by joining reference sequence $G(6500-a, a)$ and $G(b, b+6500)$. The PacBio SMRT Analysis 1.4 pipeline for Resequencing is performed to refine the amplicon template predictions using all fragments generated from the SMRT cell (Figure S6). The Resequencing protocol involves running BLASR for mapping followed by Quiver for consensus sequence calling. The protocol accurately recovered the sequence around breakpoints; the consensus amplicon sequence starting at aligned $25 - a$ and ending at $b + 25$ matched either sequencing from previous studies or independent Sanger sequencing chromatogram (Figure 5). For clusters with $L > 0$, adding L "N" nucleotides at the breakpoint junction of the predicted amplicon template had no effect on PacBio Resequencing protocol. In both cases, the correct amplicon breakpoint junction sequence was found.

Experimental Methods

A549, CEM, Detroit562, and T98G cells were thawed from Moore's Cancer Biorepository. MCF7, HeLa, and HEK (293T) cells were collected from Rosenfeld Lab. Standard DNAzol protocol was used for DNA extraction and DNA was quantified with NanoDrop 2000 spectrophotometer. DNA products are visualized on 1% agarose gels with EtBr. Gel images are either color value inverted or color curve adjusted uniformly across the image for visual enhancement. All PCRs were performed on a BioRad iCycler instrument.

All PCR experiments used the following thermocycling conditions; initial denaturation at $95°C$ for 3 min, 10 cycles at $94°C$ for 20 sec, $64°C$ for 30 sec, $66°C$ for 15 min, 28 cycles at $94°C$ for 5 sec, $64°C$ for 30 sec, $66°C$ for 15 min + 20 sec for each cycle, final extension at $64°C$ for 45 min, and $4°$ hold.

AMBRE-16 experiment

See supplemental materials for primer sequences. Standard protocol for NEB Crimson LongAmp Taq is used for $50\mu l$ PCR reactions with the following changes. The same mix of 16 primers was used in each reaction where each primer is present with final concentration of $0.2\mu M$. Starting genomic DNA for each cell line reaction is 10ng. QIAquick PCR purification kit was used to clean up PCR samples. Samples were quantified and $2\mu g$ of A549 reaction sample was mixed with $1\mu g$ of each remaining cell line reaction sample and submitted for PacBio sequencing at the UCSD BioGem Core facility. Loading of DNA samples onto a PacBio SMRT cell is biased towards sequencing smaller amplicons and increasing the amount of A549 reaction sample containing an 11kb DNA fragment was necessary to sufficiently sequence the A549 DNA fragment.

AMBRE-68 experiment

See supplemental materials for primer sequences. Standard protocol for NEB Crimson LongAmp Taq is used for $50\mu l$ PCR reactions with the following changes. The same mix of 9 primer was used in each reaction where each primer is present with final concentration of $0.4\mu M$. Starting genomic DNA for each cell line reaction is 20ng.

RUNX1-RUNX1T1 experiment

See supplemental materials for primer sequences. Standard protocol for NEB Crimson LongAmp Taq is used for $25\mu l$ PCR reactions with the following changes. All primer at 0.4 μM PCR experiments were under the conditions; initial denaturation at $95°C$ for 1 min, 10 cycles at $94°C$ for 20 sec, $63°C$ for 30 sec, $68°C$ for 2 min, 28 cycles at $94°C$ for 5 sec, $61°C$ for 30 sec, $66°C$ for 2 min $+$ 5 sec for each cycle, final extension at $64°C$ for 30 min, and $4°$ hold. Subsampling experiments used the same primer concentration and thermocycling conditions except extension times for the first phase is 7 min and the second phase is 7 min with 10 sec increase per cycle.

Tumor:Wild-type genomic DNA heterogeneity experiment

See supplemental materials for primer sequences. Standard protocol for NEB Crimson LongAmp Taq is used for $50\mu l$ PCR reactions with the following changes. Each primer has final concentration $0.4\mu M$. Each reaction contains $\approx 400ng$ gDNA, with the following tumor to normal DNA ratios: $200ng : 200ng$, $40ng : 400ng$, $4ng : 400ng$, $0.4ng : 400ng$. Normal DNA is derived from HEK cells.

MCF7 and T98G PCR validation

Primer pair sequences were generated using Primer3 2.3.0 given short genomic sequence around the MCF7 and T98G breakpoints as determined by PacBio sequencing and analysis. See supplemental materials for primer sequences. Standard protocol for NEB Standard Taq is used for $50\mu l$ PCR reactions starting with 250ng of genomic DNA.

Data Access

The sequencing data have been deposited at the NCBI Sequence Read Archive (SRA; `http://www.ncbi.nlm.nih.gov/sra`) under accession number SRX353044. The AmBre software is available at `http://bix.ucsd.edu/AmBre`.

Acknowledgements

The work was funded by NIH RO1-HG004962. Also, we thank the Rosenfeld laboratory for supplying MCF7 and HEK cell lines.

Disclosure Declaration

Figure Legends

1 Figure

PAMP tiling design for capture of *CDKN2A* deletions. *CDKN2A* upstream and downstream breakpoint regions are defined on a germline genome, blue and red lines, respectively. Tiled forward primers (blue arrows) and reverse primers (red arrows) are spaced $\approx 1kb$ apart (width of hashed boxes) (not to scale with reference). Overlap of blue box and red box on tumor DNA represents a forward and reverse primer pair are less than $2kb$ apart and will lead to amplification of tumor DNA harboring *CDKN2A* deletion breakpoints.

2 Figure

AmBre pipeline with primer designing and PacBio long fragment sequence analysis.

3 Figure

Designing AMBRE-68 **A)** Candidate primers are uniformly distributed in *CDKN2A* locus suggesting good primer designs are possible. AmBre-design is tasked to capture *CDKN2A* deletion upstream and downstream breakpoints in regions $chr9:21,730,000-21,965,000$ and $chr9:21,975,000-22,129,000$ (GRCh37 coordinates), respectively. **B)** Final low cost 68-primer design to capture *CDKN2A* deletions in 380kb breakpoint region. The solution has a 97.6% and a 99.7% coverage of breakpoint regions. The fraction of break pairs captured by the design (resulting in amplicon length <13kb) is 99.99%.

4 Figure

Aggregates of breakpoints from each PacBio fragments after sweep line clustering. Target amplicons are strongly supported by fragments and breakpoints are well separated. Only breakpoints with $L<1$kb are displayed for inset boxes. The height of each cluster corresponds with number of fragments supporting the breakpoint(depth of breakpoint coverage).

5 Figure

Subsampling of 9 primers from the complete AMBRE-68 tiling design results in clean amplification of *CDKN2A* loss DNA fragments in six cell lines. From left to right, lanes contain 1kb Plus GeneRuler DNA ladder, PCR products from samples A549 (2.2kb), CEM (5.8kb), MCF7 (3.6kb), MOLT4 (6.8kb), T98G (7.5kb), HEK, and water. The expected lengths of each amplicon according to AMBRE-68 design are listed in parentheses. HEK cells (no *CDKN2A* deletion) and H_2O are negative controls.

6 Figure

Characterizing *RUNX1-RUNX1T1* balanced translocation in Kasumi-1. Lanes 1,2,4,6 and 8 contain 1kb Plus GeneRuler DNA ladder, PCR products from Kasumi-1 Der8 with all 28 primers (3.5kb), 14 primer FE ∪ RO (3.5kb), 14 primer FO ∪ RO (6.8kb), 14 primer FO ∪ RE (10.1kb). Lanes 3,5,7 and 9 contain matching water controls, which show no contamination. Lanes 10,12,14, and 16 contain PCR products from Kasumi-1 Der21 with all 29 primers (2.7kb), 15 primer FO ∪ RO (2.7kb), 15 primer FE ∪ RO (6.1kb), and 14 FE ∪ RE (8.1kb). Gel was loaded with $2\mu l$ for lanes 2,3,4,5,10,11,12 and 13, and $4\mu l$ for remaining volumes. Reactions with shorter amplicons amplified extremely well and lesser volumes were used for visualization on the gel. The expected amplicon lengths according to the Der8 and Der21 design are listed in parenthesis.

7 Figure

A) Fragment-segmentation example for local alignments $1, 2, 3$, and 4 along a PacBio fragment. **B)** Triangle representation of adjacent alignments $1, 2$, and 3 on $G \times G$ plane.

Figures

Figure 1

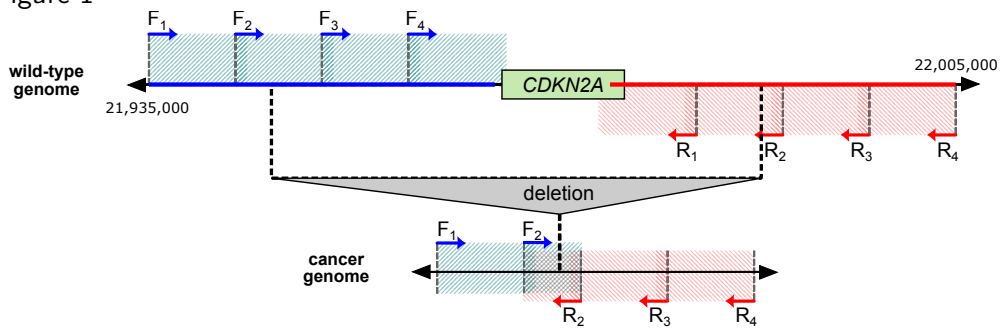


Figure 2
AmBre pipeline

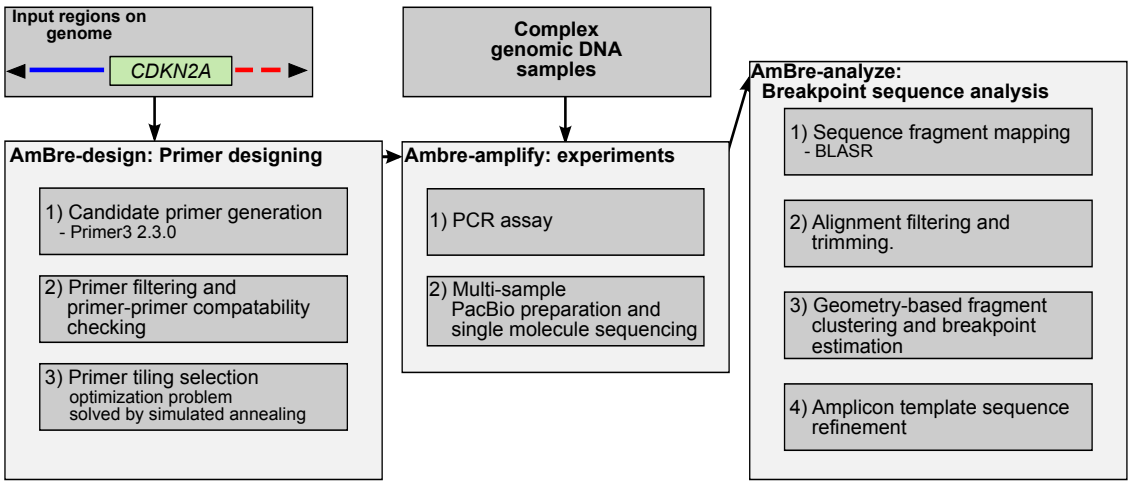


Figure 3

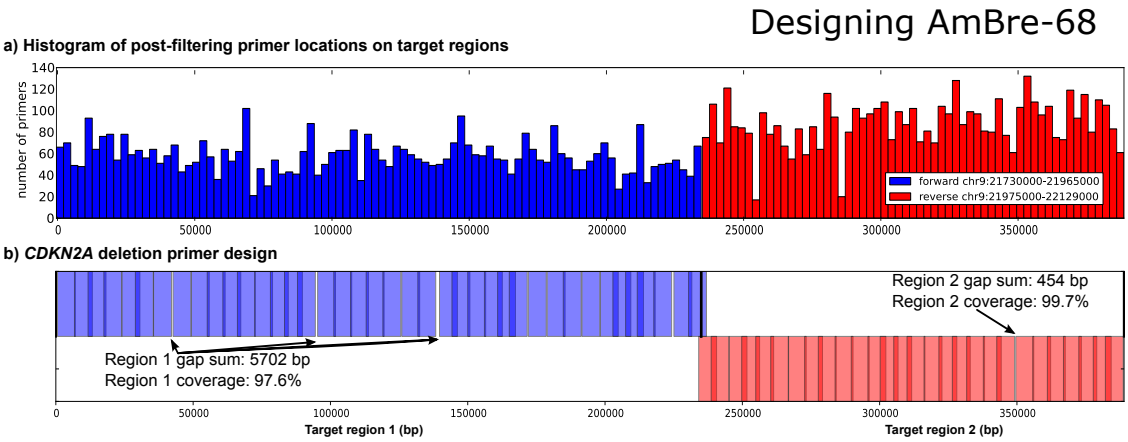


Figure 4

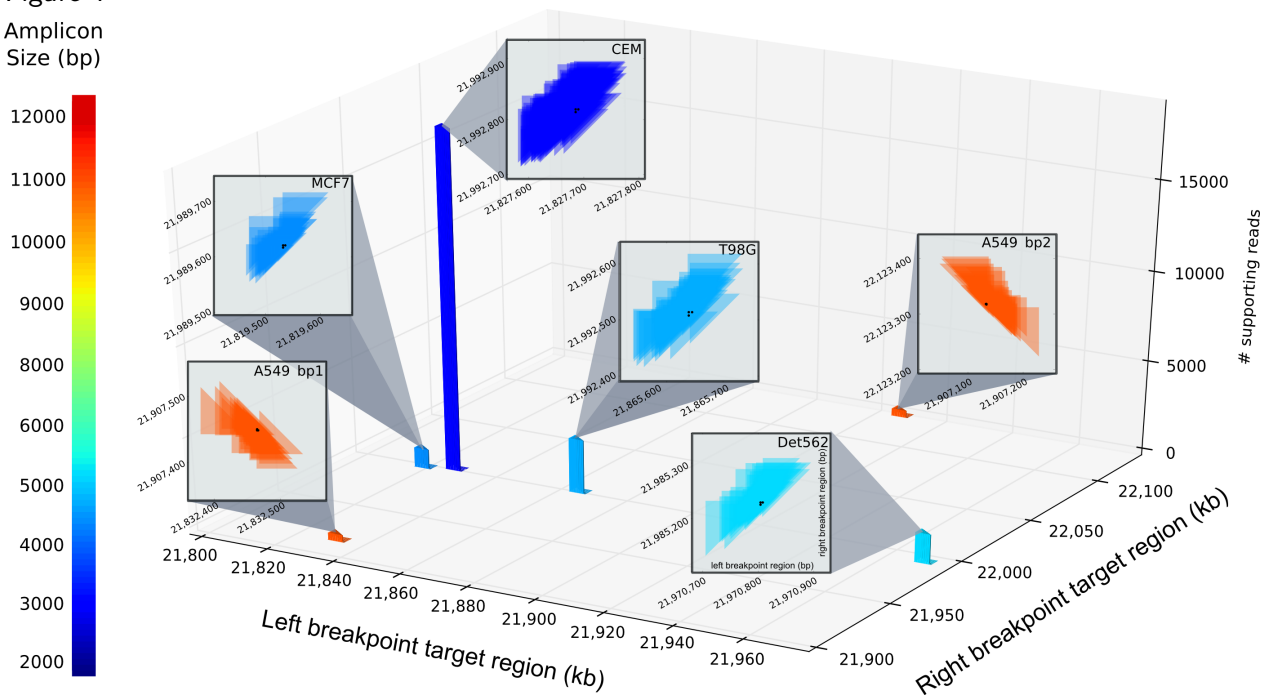


Figure 5

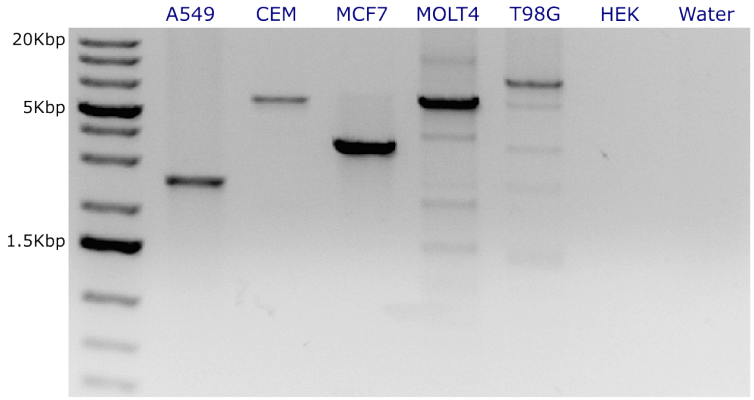


Figure 6

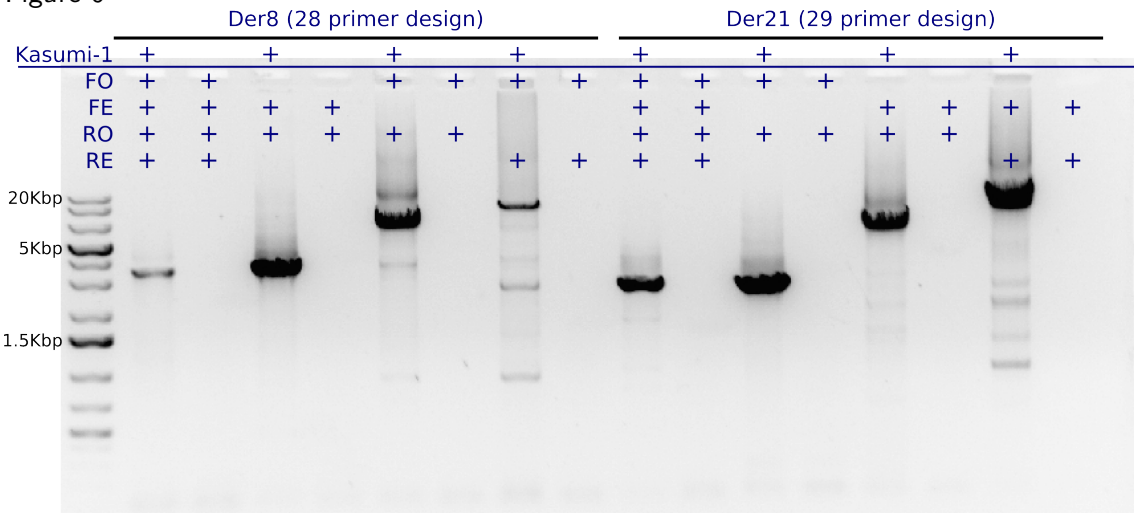
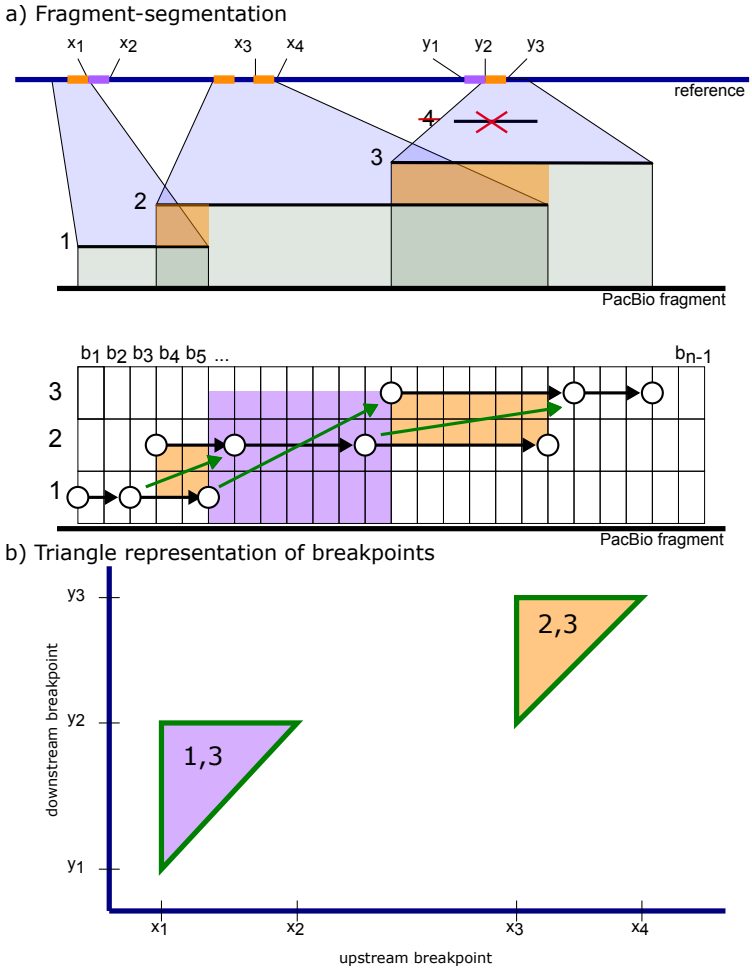


Figure 7



Tables

1 Table

Cell Line	Type	Estimated Breaks	Our Breaks	Estimated Deletion Size	True Deletion Size	Difference in Breaks
A549	lung adenocarcinoma	21833542 - 22121634	21832459 - 22123318	288092	290859	1083 - 1684
CEM	lymphoblastic leukemia	21828110 - 21992808	21828685 - 21996997	168887	164123	575 - 4189
Detroit562*	pharynx carcinoma		21970804 - 21985229		14425	
MCF7	breast carcinoma	21834611 - 21989073	21819532 - 21989621	154462	170089	**15079** - 548
T98G	glioblastoma	21868909 - 21991923	21865639 - 21992514	123014	126875	3270 - 591

Caption Five cell-lines with CDKN2A deletion breakpoints in GRCh37. Estimated breakpoints are according to CGP (Greenman et al. 2010). CGP coordinates were converted from NCBI36 to GRCh37 using UCSC liftover (Hinrichs et al. 2006). The break coordinates for Detroit562 were identical to Bashir et al. (2009) and the cell-line was not examined by CGP.

References

Bartley P, Ross D, Latham S, Martin-Harris M, Budgen B, Wilczek V, Branford S, Hughes T, Morley A 2010. Sensitive detection and quantification of minimal residual disease in chronic myeloid leukaemia using nested quantitative PCR for BCR-ABL DNA. *International Journal of Laboratory Hematology*, **32**: e222–e228.

Bashir A, Liu Y, Raphael B, Carson D, Bafna V 2007. Optimization of primer design for the detection of variable genomic lesions in cancer. *Bioinformatics*, **23**: 2807–2815.

Bashir A, Lu Q, Carson D, Raphael B, Liu Y, Bafna V 2009. Optimizing PCR assays for DNA based cancer diagnostics. In *Research in Computational Molecular Biology*, (pp. 220–235). Springer.

Chaisson M. J, Tesler G 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, **13**: 238.

Chen K, Wallis J. W, McLellan M. D, Larson D. E, Kalicki J. M, Pohl C. S, McGrath S. D, Wendl M. C, Zhang Q, Locke D. P, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, **6**: 677–681.

Dawson S.-J, Tsui D. W, Murtaza M, Biggs H, Rueda O. M, Chin S.-F, Dunning M. J, Gale D, Forshew T, Mahler-Araujo B, et al. 2013. Analysis of Circulating Tumor DNA to Monitor Metastatic Breast Cancer. *New England Journal of Medicine*.

Downing J. R 1999. The AML1-ETO chimaeric transcription factor in acute myeloid leukaemia: biology and clinical significance. *British Journal of Haematology*, **106**: 296–308.

English A. C, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny D. M, Reid J. G, Worley K. C, et al. 2012. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PloS ONE*, **7**: e47768.

Greenman C, Bignell G, Butler A, Edkins S, Hinton J, Beare D, Swamy S, Santarius T, Chen L, Widaa S, et al. 2010. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, **11**: 164–175.

Hampton O, Den Hollander P, Miller C, Delgado D, Li J, Coarfa C, Harris R, Richards S, Scherer S, Muzny D, et al. 2009. A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Research*, **19**: 167–177.

Hampton O, Miller C, Koriabine M, Li J, Den Hollander P, Carbone L, Nefedov M, Ten Hallers B, Lee A, De Jong P, et al. 2011. Long-range massively parallel mate pair sequencing detects distinct mutations and similar patterns of structural mutability in two breast cancer cell lines. *Cancer genetics*, **204**: 447–457.

Harismendy O, Schwab R, Bao L, Olson J, Rozenzhak S, Kotsopoulos S, Pond S, Crain B, Chee M, Messer K, et al. 2011. Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. *Genome Biology*, **12**: R124.

Hatch A. C, Fisher J. S, Tovar A. R, Hsieh A. T, Lin R, Pentoney S. L, Yang D. L, Lee A. P 2011. 1-Million droplet array with wide-field fluorescence imaging for digital PCR. *Lab on a chip*, **11**: 3838–3845.

Hinrichs A, Karolchik D, Baertsch R, Barber G, Bejerano G, Clawson H, Diekhans M, Furey T, Harte R, Hsu F, et al. 2006. The UCSC genome browser database: update 2006. *Nucleic Acids Research*, **34**: D590–D598.

Kaplinski L, Andreson R, Puurand T, Remm M 2005. MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics*, **21**: 1701–1702.

Kent W 2002. BLATthe BLAST-like alignment tool. *Genome Research*, **12**: 656–664.

Kim J, Deluna A, Mungrue I, Vu C, Pouldar D, Civelek M, Orozco L, Wu J, Wang X, Charugundla S, et al. 2012. The Effect of 9p21. Coronary Artery Disease Locus Neighboring Genes on Atherosclerosis in Mice. *Circulation*.

Kirkpatrick S 1984. Optimization by simulated annealing: Quantitative studies. *Journal of statistical physics*, **34**: 975–986.

Kitagawa Y, Inoue K, Sasaki S, Hayashi Y, Matsuo Y, Lieber M. R, Mizoguchi H, Yokota J, Kohno T 2002. Prevalent involvement of illegitimate V(D)J recombination in chromosome 9p21 deletions in lymphoid leukemia. *Journal of Biological Chemistry*, **277**: 46289–46297.

Lam H, Mu X, Stütz A, Tanzer A, Cayting P, Snyder M, Kim P, Korbel J, Gerstein M 2009. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature biotechnology*, **28**: 47–55.

Liu Y, Carson D 2007. A novel approach for determining cancer genomic breakpoints in the presence of normal DNA. *PLoS One*, **2**: e380.

Mason C, Elemento O 2012. Faster sequencers, larger datasets, new challenges. *Genome Biology*, **13**: 314.

Michor F, Hughes T, Iwasa Y, Branford S, Shah N, Sawyers C, Nowak M 2005. Dynamics of chronic myeloid leukaemia. *Nature*, **435**: 1267–1270.

Rozen S, Skaletsky H, et al. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*, **132**: 365–386.

Sasaki S, Kitagawa Y, Sekido Y, Minna J. D, Kuwano H, Yokota J, Kohno T 2003. Molecular processes of chromosome 9p21 deletions in human cancers. *Oncogene*, **22**: 3792–3798.

Sindi S, Helman E, Bashir A, Raphael B 2009. A geometric approach for classification and comparison of structural variants. *Bioinformatics*, **25**: i222–i230.

Wessely R 2010. Atherosclerosis and Cell Cycle: Put the Brakes On!: Critical Role for Cyclin-Dependent Kinase Inhibitors? *Journal of the American College of Cardiology*, **55**: 2269–2271.

Xiao Z, Greaves M, Buffler P, Smith M, Segal M, Dicks B, Wiencke J, Wiemels J 2001. Molecular characterization of genomic AMLI-ETO fusions in childhood leukemia. *Leukemia*, **15**: 1906–1913.

Ye K, Schulz M. H, Long Q, Apweiler R, Ning Z 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**: 2865–2871.

Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-Né P, Nicolas A, Delattre O, Barillot E 2010. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, **26**: 1895–1896.