Amplification and thrifty single molecule sequencing of recurrent somatic structural variations

Anand Patel, Richard Schwab, Yu-Tsueng Liu, and Vineet Bafna

December 2, 2013

Supplemental Materials

1 Comparison to Bashir et al. cost function

The cost function, $C(P)$ differs from Bashir et al. (2007), $B(P)$. Unlike $C(P)$, $B(P)$ penalizes for loss of break region coverage independent from limiting too many primers in a design. More specifically, $B(P)$ applies a cost weight to the sum of too large primer spacings and limits the number of primers by adding the cost $w_G \frac{|P|d}{T}$. If $\frac{|P|d}{T}$ is greater than primer density parameter ρ, then $B(P) = \inf$ by setting $w_G = \inf$, otherwise $w_G = 0$. While $B(P)$ works well for large regions requiring numerous primers, it does not work well for smaller regions with sparser selection of primers.

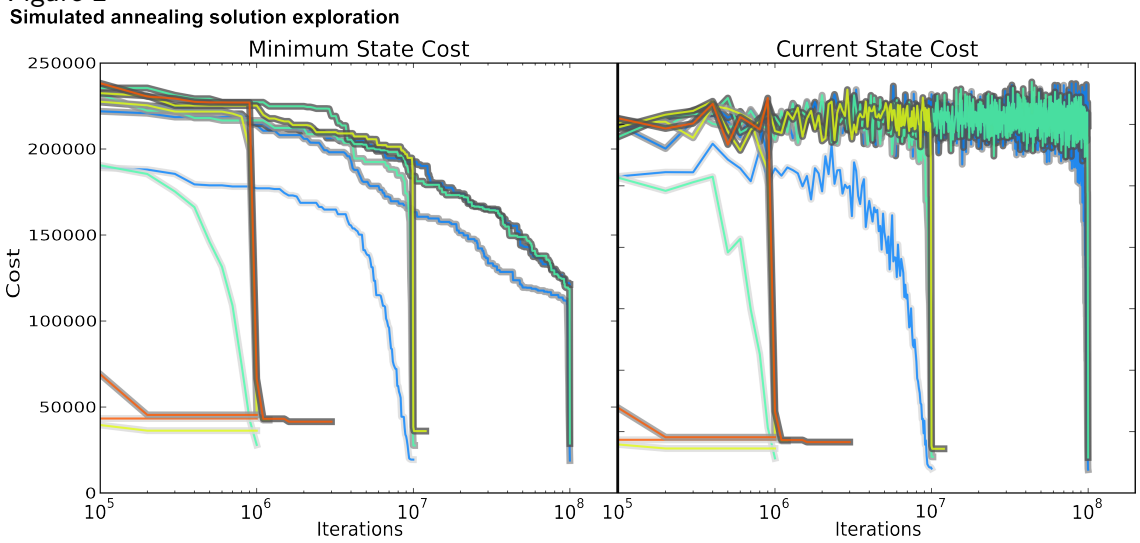
Consider a one-sided case where the input is one region of length T in F and R is an empty set. For $1 \leq \rho < 1 + \frac{d}{T}$ and a primer set (P) with $\lceil \frac{T}{d} \rceil$ primers, the cost of adding any compatible primer to P without primer removal for $B(N_l(P))$ is $\inf$ whereas for $C(N_l(P))$ is definite. This becomes an issue when requiring sparse primer selection. An extreme example with any primer density $1 \leq \rho < 1.5$ where $T = 2d$, the simulated annealing procedure using B would limit to only primer subsets of size 0, 1, and 2. However, setting $\rho = 1.5$ would have a definite cost for primer subsets of size P. Using B, how to set ρ with sparse selection of primers in multiple regions is also not trivial. The reformulated cost function C, has the additional property where $C_{\rho_i}(N_l(P)) > C_{\rho_j} N_l(P)$ for $\rho_i < \rho_j$, therefore changing ρ has a reasonable effect on the new cost of primer subsets.

Also, under sparse primer selection, B does not necessarily evenly space primers. Consider again the above one-sided case. Let $N = \frac{\rho T}{d}$ (number of primers desired) and $N_e = N - \lfloor \frac{T}{d} \rfloor$ (number of extra primers than necessary to cover T) and suppose each position of T has a fully compatible candidate primer. The ideal primer subset (P_I) would select a primer every $\frac{T}{N}$ positions of T. Consider a contrived primer subset (P_C) where a primer is selected every d positions of T and $\frac{d}{N_e+1}$ positions of T, which places primers to cover T entirely and places all extra primers within the first covered segment of T. Note the cost $B(P_I)$ equals the cost $B(P_C)$ even though P_I is clearly better. The cost $C(P_I)$ is zero and less than cost $C(P_C)$ if there are extra primers to be placed, $N_e > 0$. This scenario is unlikely to occur with a large region and low primer density, however may occur with sparse primer selection and high primer density. Nonetheless, C reports sensible costs for both situations.

It follows from above, low-cost solutions with unevenly spaced primers can be generated by setting a primer density threshold for $B(P)$ too high. For discontiguous target regions, in particular, some target regions may receive more extra primers than should be placed in that region. Our cost function penalizes for irregular primer spacing, thus each target region will receive the appropriate number of primers.

2 Figure for simulated annealing convergence

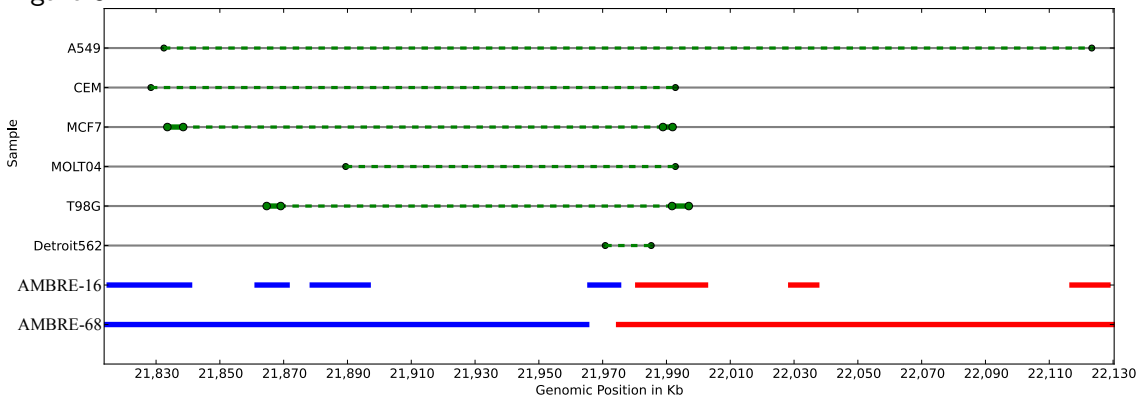
Figure 2



Caption Designing AMBRE-68 Simulated annealing using different convergence rates, is used to select good primer designs with lowest cost. The convergence rate that finds the lowest cost primer design will depend on the input given to AmBre-design.

3 AmBre designs with breakpoint estimates from CGP

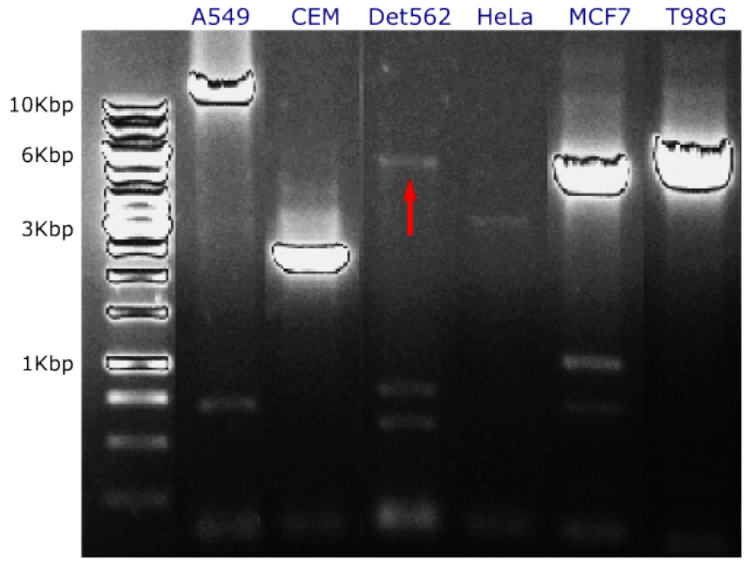
Figure 3



Caption Breakpoint estimates for A549, CEM, MCF7, and T98G from CGP. Last two rows are AMBRE-16 and AMBRE-68 input target regions.

4 Figure for AMBRE-16 amplifications

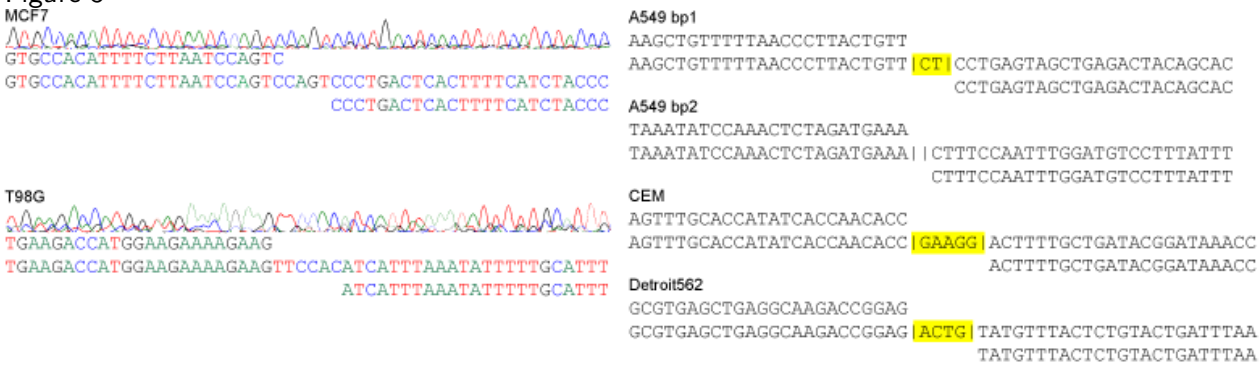
Figure 4



Caption PCR products of AMBRE-16 on cell-lines: A549 (lane 2), CEM (lane 3), Detroit562(lane 4), HeLa(lane 5), MCF7 (lane 6), and T98G (lane 7). $4\mu l$ of $1kb$ GeneRuler in lane 1. Lanes are reactions starting with $10ng$ cell-line genomic DNA. HeLa cells (no *CDKN2A* deletion reported by CGP) and H_2O are negative controls. Arrow denotes weak Detroit562 band; another PCR had stronger amplification and was used for subsequent sequencing.

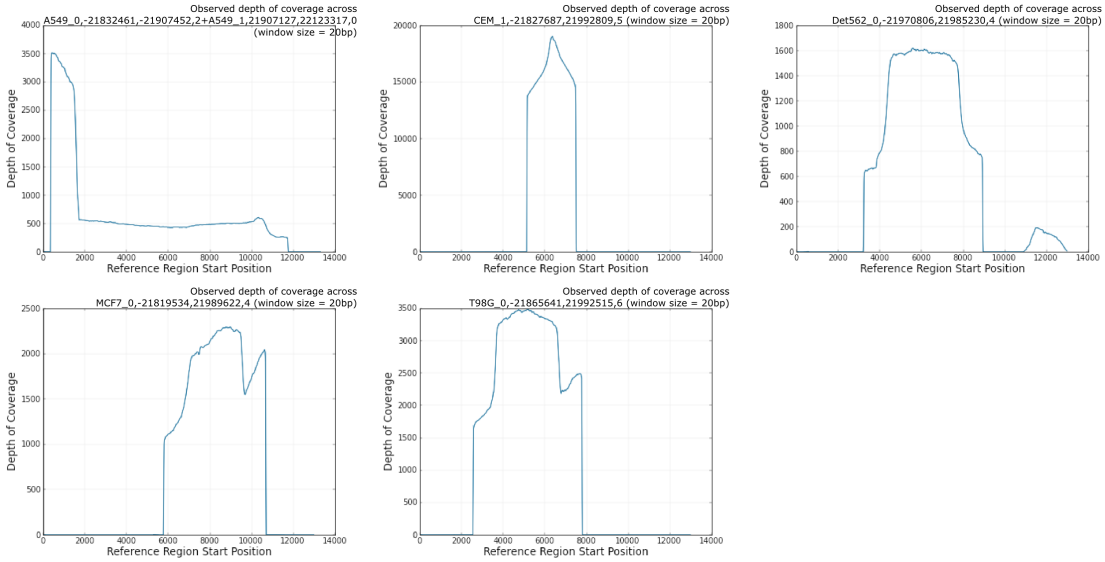
5 Breakpoint sequences

Figure 5



Caption Breakpoint sequences for A549, CEM, Detroit562, MCF7, and T98G with orthogonal validation chromatogram of MCF7 and T98G. AmBre-analyze captures both breakpoints and non-templated insert sequence (highlighted in yellow).

6 PacBio coverage of refined amplicon sequences



Caption BLASR remapping to consensus amplicon templates. The GC content for A549, CEM, Detroit562, MCF7, and T98G amplicons is 0.40, 0.42, 0.42, 0.38, and 0.38, respectively.

Primer sequences for AMBRE-16 and AMBRE-68

AMBRE-16 primer sequences	
A2-21815419-T-r0.2-3	TTTCTCTCTTAGATTGGAATAATTGGTGGAAC
A2-21818824-T-r0.2-3	TACGTTGTCATTAGCTATAATCACCATGCAG
A2-21826327-T-r0.2-3	TTAATAAGCCTTCTAGTCTGGAAGATTCCAC
A2-21833470-T-r0.2-3	TCACTTCCTTCTGGTTATAGAGACAGAATTG
A2-21861703-T-r0.2-3	ATGGCAATAAGTGATTATCAGAACAATGCTC
A2-21867239-T-r0.2-3	GTTACTCTTGTCTTATTCTCAACAGCAGAGG
A2-21879976-T-r0.2-3	TTTCAGTCATGGAAAATCTAAGGATTATGTG
A2-21885329-T-r0.2-3	TCTTAAGAGGTTGGGCAGGATTACTATAACC
A2-21891029-T-r0.2-3	TAGGAACCGTAGTTTGAGAACAACTGTTCTAG
A2-21967542-T-r0.2-3	CCCCTAGAGTTTCTATTCATCATTTTAACCG
A2-21987671-F-r0.2-3	TATTATGTGACCCTTTGTATGAATTGGAAAAG
A2-21993806-F-r0.2-3	ACACACACAGTAGGAAAGGTGTATTTCAAGC
A2-22001902-F-r0.2-3	ACAAGGACTTGACTGGAAGATAGAAGACCTAG
A2-22036684-F-r0.2-3	AATGGACAGATGACTCCTAACTTTGACATTAG
A2-22121823-F-r0.2-3	AGTTATAGAATCAATCCAGGCATCCAAAAAC
A2-22128237-F-r0.2-3	GTAGCAACAGCAGTAGCAGTATAACAGCAAC

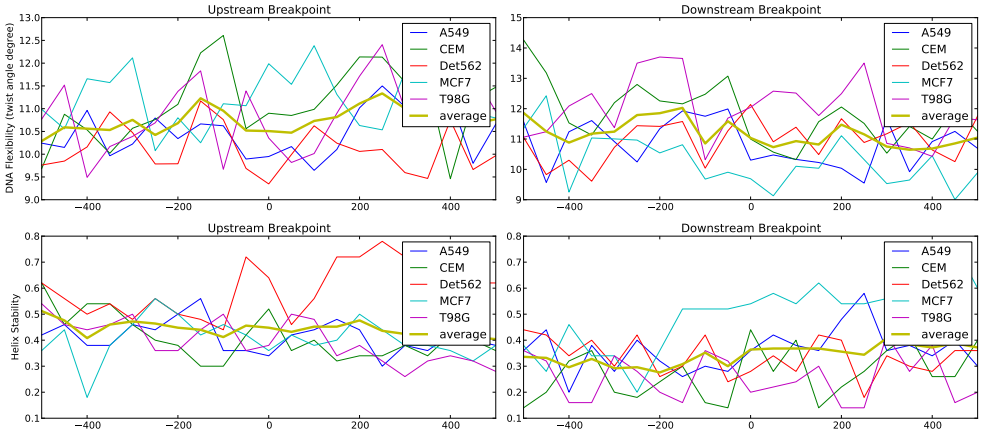
9 primer sequences used from AMBRE-68 design	
chr9-21817993-T	TACTCATCACGGGTTAACAATTTCTTCTCTC
chr9-21825345-T	GATGTCTTTCTTGACGTAAAATTTCATTTCC
chr9-21831236-T	AATCCTGATGATTGTAGGAAATCAGTACACC
chr9-21861845-T	TTTATTTAAAGTGTATGTTTCCTGCGTCCTC
chr9-21960424-T	CACTTTCTGACTGCACTTTCTTGAAGTATTC
chr9-21991669-F	CCACGTTAGTTCATCACTATCAACTACCATG
chr9-21991669-F	CCACGTTAGTTCATCACTATCAACTACCATG
chr9-21996284-F	GTAAGAGTGATCGCTAAATCTCACTTTTTCC
chr9-22124303-F	TACATTTTGTCTACCCATCCCTTTCTTAATG

MCF7 and T98G validation primer sequences

Primer sequences used for MCF7 validation	
forward	CAAGAGGCCCTGGTGTGT
reverse	GGGGTTAGTGGACTCGAGAC

Primer sequences used for T98G validation	
forward	CCAGCAAGGCAAAGAACTGA
reverse	TTAGCCACTGTGACCGGTAA

7 DNA helix stability around breakpoints



Caption Using code from BreakSeq pipeline, DNA flexibility for the 6 breaks around proposed non-homologous end joining DNA breaks showed no significant deviation.

8 *RUNX1-RUNX1T1* translocation Sanger confirmation

Subsampled positive PCRs from Kasumi-1 were sent to GENEWIZ for sequencing after Qiagen PCR clean-up. Each sample sequenced with a forward primer upstream and a reverse primer downstream of the expected Kasumi-1 breakpoint. Samples A,B and C are shortest to longest amplicons from Der8. Similarly, samples D,E and F are amplicons from Der21.

```
BLASTN 2.2.28+
Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and
Webb Miller (2000), "A greedy algorithm for aligning DNA
sequences", J Comput Biol 2000; 7(1-2):203-14.

Query= KASUMI Derivative chromosome 8 AML-ETO

Length=240
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value
lcl|8236  A-a1-1_kas_T_A08.ab1                                         422   2e-122
lcl|8237  B-a1-1_kas_T_B08.ab1                                         444   5e-129
lcl|8238  C-a1-1_kas_T_C08.ab1                                         346   1e-99
lcl|8239  A-a1-1_kas_F_D08.ab1                                         255   2e-72
lcl|8240  B-a1-1_kas_F_E08.ab1                                         298   4e-85

ALIGNMENTS
Query  1    TGCATTTTTCCCCAGAGGGCCCTCTCAGCTTTTTCTGATCGAGACTTTCTGGCCATGCCTAGACTTGGGAGACTGTGGGACTTGATAACA  90
8236   282  ..........................................................................................  193
8237   273  ..........................................................................................  184
8238   277  ......................N....AAAAAAN.N.....................................................  188
8240   21                                                                          .....N........N  35

BREAKPOINT            >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
Query  91   GACCCTGATTCTTCTTTATGAGTGAAAAGCTTGAGAACACTTTCCTGTATATTTAGACATTTATTGCTTTCATAATTAATTTATCTGACA  180
8236   192  ..........................................................................................  103
8237   183  ..........................................................................................  94
8238   187  .....C..............................C......N...................N.......................  98
8239   51                .....................................................................     128
8240   36   ..........................................................................................  125

Query  181  AATAATATCAAGAAGGTTGTGTTTCTGAAAACAACCTTATAGATCTTTAGCAAGCAACAT  240
8236   102  ..............................................                 55
8237   93   ..........................................................     34
8238   97   ................N.......A...........AA.........N.N.N......     40
8239   129  ..........................................................     188
8240   126  ..........................................................     185

Query= KASUMI Derivative chromosome 21 ETO-AML
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value
lcl|19454  D-a2-2_kas_T_G08.ab1                                        438   2e-127
lcl|19455  E-a2-2_kas_T_H08.ab1                                        438   2e-127
lcl|19456  F-a2-2_kas_T_A09.ab1                                        438   2e-127
lcl|19457  D-a2-2_kas_F_B09.ab1                                        285   3e-81
lcl|19458  E-a2-2_kas_F_C09.ab1                                        285   3e-81
lcl|19459  F-a2-2_kas_F_D09.ab1                                        292   2e-83

ALIGNMENTS
Query  1    AGTTTCACTCTCTTTGTCTTAAAAGTAGAAACAGTTATTCTGCCTTGCTAACTTTCACAGGATGGCAATAGTGATTGTAGATATGCTGTG  90
19454  313  ..........................................................................................  224
19455  312  ..........................................................................................  223
19456  310  ..........................................................................................  221
19457  20                                                                                  .......  26
19458  20                                                                                  .......  26
19459  16                                                                               ...........  26

Query  91   TACCTGTATAAAGAGGTTATTAAACCTCTTGGCAGTCAGAGGGGAAAAAAGATGTCTATGCGATTGTAACGGTCTGGCTGTATATACAGT  180
19454  223  ..........................................................T...............................  134
```

```
19455  222  ..........................................................T.............................  133
19456  220  ..........................................................T.............................  131
19457  27   ..........................................................T.............................  116
19458  27   ..........................................................T.............................  116
19459  27   ..........................................................T.............................  116

Query  181  TTACTCTGCCCCTGCTGTTTTACAACTTAAGAAAACAGTTGTTATATGGAGAAGCACCCT  240
19454  133  ............................................................  74
19455  132  ............................................................  73
19456  130  ............................................................  71
19457  117  ............................................................  176
19458  117  ............................................................  176
19459  117  ............................................................  176
```

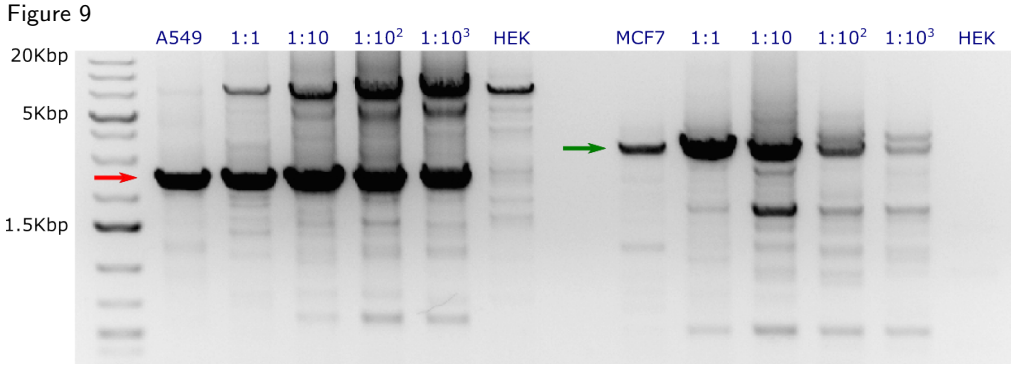
Primer sequences for Kasumi-1 breakpoints

Der8 primer sequences for 3.5Kbp	
d2.1-6-T	AGTACACTAGAGCACCATAAGAATACAATCC
d2.1-9-F	CAATACTCTCGCTACTCAAAGCTTGTTTC
Der8 primer sequences for 6.8Kbp	
d2.1-6-T	AGTACACTAGAGCACCATAAGAATACAATCC
d2.1-10-F	CCAGTAGGAAGACAGTCATGTGAAGATG
Der8 primer sequences for 10.1Kbp	
d2.1-5-T	CAAGCCAAAATACCACAATCATCCTAAGAC
d2.1-10-F	CCAGTAGGAAGACAGTCATGTGAAGATG
Der8 primer sequences validation	
kas-a1.1-T	TGGCGAGAATCAAACCAAACATTG
kas-a1.1-F	TTTTCTGATCGAGACTTTCTGGCC

Der21 primer sequences for 2.7Kbp	
d1.1-9-T	TTTGCCCTTCTATAATAGACAGTCTTCGAG
d1.1-13-F	CTTCCTCAGGACTTGTTTGCTTTAATGATTC
Der21 primer sequences for 6.1Kbp	
d1.1-8-T	AAGAGCAAAGAGTGACACATCTTTTCATC
d1.1-13-F	CTTCCTCAGGACTTGTTTGCTTTAATGATTC
Der21 primer sequences for 8.5Kbp	
d1.1-8-T	AAGAGCAAAGAGTGACACATCTTTTCATC
d1.1-14-F	CAGTCTTACTTTGTTGTCTTTTATCTTCAGTCC
Der21 primer sequences validation	
kas-a2.2-T	TTCCACGCATTTAGTTTTTCCCCA
kas-a2.2-F	CTGCCTTGCTAACTTTCACAGGAT

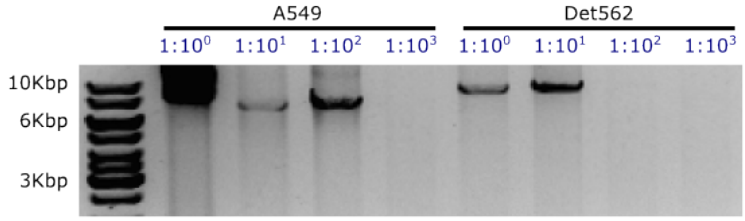
9 Amplification in complex gDNA samples with AMBRE-68 on A549-HEK and MCF7-HEK

Figure 9



Caption Successful A549 (red arrow) and MCF7 (green arrow) *CDKN2A* deletion amplification with heterogeneity ratios $1:1, 1:10, 1:100, 1:1000$ (lanes 3-6 for A549 and lanes 10-13 for MCF7) and 16 primers starting with $400ng$ of gDNA. Lane 1 contains 1kb Plus Gene Ruler DNA ladder. Lanes 2 and 9 are A549 and MCF7 positive control reactions starting with 20ng of homogenous gDNA. Lanes 7,14 are negative control reactions with wild-type DNA and lanes 8,15 are water negative control reactions with corresponding 16 primer mixes.

10 Amplification in complex gDNA samples with longer PCR products and lower multiplexing



Caption A549 and Detroit562 with 6 primers and heterogeneity ratios.

Starting with a total $250ng$ of DNA material, we demonstrate amplification of a 7.6kbp CDKN2A deletion sequence from A549 and a 9.7kbp deletion sequence from Detroit562 with 6 primers and tumor to wildtype DNA mixtures of 1:1,1:10, and 1:100 (Fig. 10).

Standard protocol for NEB Crimson LongAmp Taq is used for $25\mu l$ PCR reactions with the following changes. Each primer has final concentration $1.1\mu M$. Each reaction contains $\approx 250ng$ of DNA, with the following tumor to normal DNA ratios: $125ng : 125ng$, $25ng : 250ng$, $2.5ng : 250ng$, $0.25ng : 250ng$. Normal DNA is derived from HEK cells.

References

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.