Supplemental Information for Michael F. Berger *et al*., Integrative Analysis of the Melanoma Transcriptome

SUPPLEMENTAL TEXT

Analysis of Gene Expression Levels Using Melanoma RNA-Seq Data

RNA-Seq enables the precise quantification of gene expression levels based on the number of sequence reads originating from each individual transcript (Cloonan et al. 2008; Mortazavi et al. 2008). To convert the observed sequence coverage for each transcript to the underlying gene expression levels, we normalized by the fraction of each transcript that is "alignable" by uniquely-mapping reads. We quantified expression levels of all genes in each sample according to the RPKM measure (reads per kilobase of exon model per million mapped reads) (Mortazavi et al. 2008), reported in **Supplemental Table S4**. As expected, we observed high reproducibility in our expression level estimates, both for a single cDNA library sequenced in multiple Illumina lanes and for multiple cDNA libraries created from a single sample (**Supplemental Figure S7**). To benchmark the expression levels determined from RNA-Seq against microarray-based measurements of gene expression, we performed microarray experiments with RNA from 9 melanoma samples using two Affymetrix platforms: U133A and Human Exon 1.0 ST arrays. RNA-Seq exhibited good correlation with both microarray platforms, with slightly better agreement with human exon arrays ($R = 0.68$ versus $R = 0.63$). Of note, the RNA-Seq data exhibited a much greater dynamic range than all microarray-based estimates, spanning approximately 6 orders of magnitude compared to only 3 on microarrays (**Supplemental Fig. S8**). Additionally, we found RNA-Seq to provide more precise expression level estimates than microarrays for 97% of expressed genes, based on independent replicate experiments (**Supplemental Fig. S9**).

RNA-Seq Data Enables Global Views of Alternative Splice Junctions

Individual sequence reads spanning exon junctions can reveal the relative abundance of different splice isoforms (Pan et al. 2008; Wang et al. 2008). For example, as illustrated in **Supplemental Figure S10**, we detected multiple isoforms of *ADAM15*, which encodes a transmembrane disintegrin protein that is overexpressed and associated with metastatic progression in breast and prostate adenocarcinomas (Kuefer et al. 2006). Alternative splicing of *ADAM15* has been shown to be misregulated in cancer (Ortiz et al. 2004), and particular splice variants have been linked to poorer relapse-free survival of breast cancer patients, potentially due to altered molecular interactions between its cytoplasmic domain and the Src and Brk tyrosine kinases (Zhong et al. 2008).

Recognizing that not all exon junctions in all genes may be known, we created a database of all hypothetical (intragenic) exon junctions based on all annotated exons in the Ensembl database. We then aligned all 51-mer reads to these junction sequences and considered reads mapping to a junction but not to the genome. Exon junctions were confirmed by the presence of at least 2 distinct reads. Of the 204,702 known exon

junctions in Ensembl protein-coding genes, we confirmed 116,153 (57%) in the 10 melanoma samples. We also confirmed 4,313 novel junctions in 2,932 genes. (Taken together with the K-562 RNA-Seq data, we confirmed 5,375 novel junctions in 3,500 genes.) Included among the genes harboring novel intragenic splice junctions are 12 of the 22 genes we identified in melanoma gene fusions: *ANKHD1*, *ARHGEF12*, *C5orf32*, *C9orf127*, *C11orf67*, *CCT3*, *GCN1L1*, *GNA12*, *RECK*, *SCAMP2*, *SHANK2*, and *TLN1*.

Though some of these novel exon junctions may be specific to melanoma, it is likely that most are expressed more generally across multiple human cell types. Of the 4,313 novel junctions in melanoma, 1,555 were detected by at least 1 read in K-562. Part of the remaining discordance may arise from the limited sensitivity to detect isoforms in low- to moderately-expressed genes. Of the 270 novel junctions in melanoma implicated by at least 10 distinct reads, 208 were also detected in K-562.

We next asked how many genes showed evidence for alternative splicing in melanoma. Each individual melanoma sample showed an average of 618 genes simultaneously expressing multiple splice isoforms, as evidenced by overlapping exon junctions (*i.e.*, a single exon joined to 2 downstream exons, or 2 exons joined to the same downstream exon). Considering all melanoma samples together, multiple distinct isoforms were detected for 4,342 genes (including 7 genes involved in fusions), though not necessarily expressed simultaneously in a single sample. The fact that there is significant overlap between gene fusions and alternatively-spliced genes is not surprising, as the sensitivity for detecting gene fusions and detecting alternative splicing depends on the observed sequence coverage at a locus. Other RNA-Seq studies interrogating many different tissue types have estimated that as many as 95 percent of human genes are differentially spliced in a tissue-specific manner (Pan et al. 2008; Wang et al. 2008).

We considered the possibility that the 4,342 alternatively-spliced genes detected here are enriched for genes important for melanoma genesis and progression and examined the overlap with independent gene sets. Of 379 curated cancer-related genes (Futreal et al. 2004), 116 were detected in our melanoma dataset ($P = 0.0004$). Of 1,788 genes with somatic mutations in the Cosmic database (Forbes et al. 2008), 488 were detected in our melanoma dataset ($P = 0.000037$). Finally, of 71 genes significantly amplified or deleted and differentially expressed in melanoma (Lin et al. 2008), 34 were alternatively spliced here ($P = 0.0000005$). As above, the significance of overlap may arise in part due to the underlying expression levels of these genes, though this relationship is not obvious.

Allele-Specific Gene Expression in the Melanoma Transcriptome

It is occasionally the case that the two alleles of a particular gene exhibit differential expression due to alternate *cis* regulatory elements, DNA methylation, or DNA copy number variation. Genotyping assays interrogating common germline SNPs have revealed allele-specific mRNA expression in a large fraction of genes (15-20%) in primary cancer cells (Milani et al. 2009). In the case of somatic (heterozygous) mutations, the mutant allele may be preferentially expressed in the context of a tumor.

We therefore examined the allelic ratios determined by RNA-Seq of all novel, validated base mutations to determine if somatic mutations exhibited preferential expression.

On average, these somatic mutations exhibited an overall derived allele frequency of 0.72, notably higher than the expected 0.50 value for balanced allelic expression. This discrepancy arises for two reasons. First, 8 of the somatic mutations exhibit monoallelic expression due to loss of heterozygosity in the underlying genomic DNA (confirmed by genotyping). Removing these cases, the average derived allele frequency decreases to 0.62. Second, there is an ascertainment bias in those variants predicted by RNA-Seq, such that sites with greater derived allele frequency are more likely to be detected. To compensate for this bias, we examined the novel variants that were called by RNA-Seq but proved to be present in the germline. We observed no significant difference in the derived allele frequency between novel somatic and germline variants. We noticed a slightly higher frequency for non-silent somatic mutations compared to silent somatic mutations, but this did not reach statistical significance ($P = 0.19$; Mann Whitney U-test).

Despite the lack of enrichment in base mutation expression across the melanoma data set as a whole, we nevertheless observed several interesting individual cases of apparent allele-specific expression. Using annotated SNPs in the coding regions of the 22 genes implicated in melanoma gene fusions in this study, we examined the allelic ratios for all heterozygous positions with at least 20x sequence coverage. Notably, we found evidence for allele-specific expression of the gene *SLC12A7* in the M000921 and 501Mel transcriptomes (**Supplemental Fig. S11**). *SLC12A7,* which is also involved in a gene fusion event in this same cell line, has been shown previously to undergo allele-specific expression in 13-38% of primary leukemic samples (Milani et al. 2009). Here, the Illumina reads overlapping SNP rs1058508 (reference = T) exhibit allele counts of C[31] T[19] in M000921 and C[20] T[7] in 501 Mel. SNP 6.0 microarray experiments suggest that M000921 is a high ploidy tumor (data not shown), implying that the imbalanced allelic expression may be a direct result of unequal copy number at the *SLC12A7* locus. However, 501 Mel appears to be diploid at this locus, suggesting that expression may be influenced by cis-acting regulatory elements or epigenetic marks. These data illustrate a systematic means to interrogate allele-specific gene expression in RNA-Seq data, while offering additional evidence in support of genomic alterations of *SLC12A7* as possible contributors to melanoma biology.

SUPPLEMENTAL METHODS

Estimation of Gene Expression Levels

To obtain an accurate measure of transcript abundance, we took into account the fact that not every transcript is completely “alignable” by uniquely-mapping reads. We generated a set of synthetic overlapping 51-mer reads corresponding to every position in the Ensembl transcripts with zero mismatches, and we aligned these back to the reference transcriptome and genome using BWA, subject to the same parameters and filtering criteria as in the original alignment. This revealed the fraction of each transcript that is alignable, which we then used as a normalization factor to convert the observed sequence

coverage for each transcript to the underlying gene expression levels. Expression levels were converted to the RPKM measure (reads per kilobase of exon model per million mapped reads) (Mortazavi et al. 2008).

Gene Expression Microarrays

Affymetrix U133A microarray experiments were performed and analyzed in a previous study (Lin et al. 2008). Affymetrix Human Exon 1.0 ST were performed (www.genome-explorations.com) and processed using the aroma.affymetrix package in Bioconductor. RMA (Robust Multichip Average) was applied in combination with quantile normalization to generate gene centric expression values, using CDF files based on remapping of probes to the human genome. This resulted in expression values for 18,632 genes.

Quantifying Alternative Splicing

We created a database of all hypothetical (intragenic) exon junctions based on all annotated exons in the Ensembl database. Sequences extended 41 bases upstream and downstream of each junction. This produced a total of 2,230,710 hypothetical junctions, 204,702 of which exist in Ensembl transcripts. All 51-mer Illumina reads were aligned to the database of junction sequences using the ImperfectLookupTable (ILT) tool of the Arachne genome assembly suite (Jaffe et al. 2003). Exon junctions were confirmed by the presence of at least 2 distinct reads harboring no more than 4 mismatches with the junction sequence and no fewer than 10 mismatches with their best placements in the human genome. In all, 120,466 junctions in 12,462 genes were confirmed in the 10 melanoma samples combined, including 4,313 novel junctions in 2,932 genes.

To identify genes exhibiting alternative splicing, we considered cases in which at least 2 confirmed exon junctions within a single gene overlapped with each other, producing distinct transcripts. A single exon could be joined to 2 downstream exons, 2 exons could be joined to the same downstream exons, or 2 junctions could involve alternating exons. 2,498 genes exhibited multiple isoforms simultaneously expressed in a single melanoma sample; 4,342 genes demonstrated evidence of multiple isoforms considering data from all 10 melanoma samples combined. (16,663 genes in Ensembl contain at least 3 exons and are theoretically capable of exhibiting alternative splicing.)

Comparing Precision of Gene Expression Methods

In general, estimates of gene expression will be less accurate for genes expressed at lower levels. To determine the precision in expression levels estimates from RNA-Seq and microarrays, we examined replicate experiments for RNA from K-562. 18 independent microarray experiments using U133A Affymetrix arrays (GEO accession GSE1922) were analyzed, and 2 independent cDNA libraries were sequenced in separate Illumina lanes. Each library produced approximately 2.8 million read-pairs aligning uniquely to an annotated gene. The coefficient of variation (CV) was empirically determined for each gene. The observed CV from RNA-Seq closely agreed with a theoretical estimate based

on a Poisson model (CV = $1/\sqrt{n}$ for n reads mapping to a given gene). Genes were called "present" in Affymetrix if they were present in at least 9 of 18 microarrays. RNA-Seq gave more precise measurements of expression level for approximately 6,700 genes, including 97% of all present genes (**Supplemental Fig. S9**).

REFERENCES

Cloonan, N., A.R. Forrest, G. Kolle, B.B. Gardiner, G.J. Faulkner, M.K. Brown, D.F. Taylor, A.L. Steptoe, S. Wani, G. Bethel et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5:** 613-619.

Forbes, S.A., G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J.W. Teague, P.A. Futreal, and M.R. Stratton. 2008. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10:** Unit 10 11.

Futreal, P.A., L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M.R. Stratton. 2004. A census of human cancer genes. *Nat Rev Cancer* **4:** 177-183.

Jaffe, D.B., J. Butler, S. Gnerre, E. Mauceli, K. Lindblad-Toh, J.P. Mesirov, M.C. Zody, and E.S. Lander. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* **13:** 91-96.

Kuefer, R., K.C. Day, C.G. Kleer, M.S. Sabel, M.D. Hofer, S. Varambally, C.S. Zorn, A.M. Chinnaiyan, M.A. Rubin, and M.L. Day. 2006. ADAM15 disintegrin is associated with aggressive prostate and breast cancer disease. *Neoplasia* **8:** 319-329.

Lin, W.M., A.C. Baker, R. Beroukhim, W. Winckler, W. Feng, J.M. Marmion, E. Laine, H. Greulich, H. Tseng, C. Gates et al. 2008. Modeling genomic diversity and tumor dependency in malignant melanoma. *Cancer Res* **68:** 664-673.

Mortazavi, A., B.A. Williams, K. McCue, L. Schaeffer, and B. Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621-628.

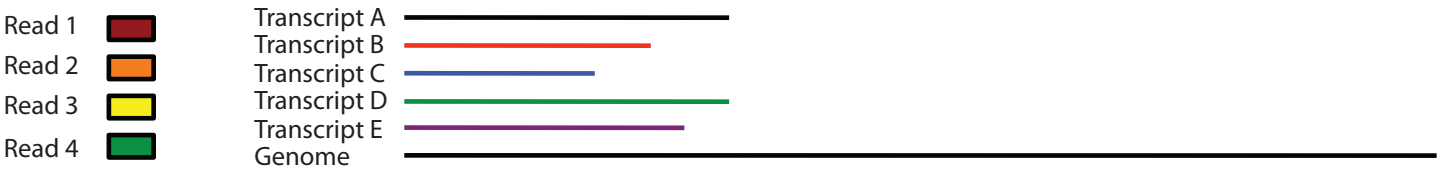
Ortiz, R.M., I. Karkkainen, and A.P. Huovila. 2004. Aberrant alternative exon use and increased copy number of human metalloprotease-disintegrin ADAM15 gene in breast cancer cells. *Genes Chromosomes Cancer* **41:** 366-378.

Pan, Q., O. Shai, L.J. Lee, B.J. Frey, and B.J. Blencowe. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40:** 1413-1415.

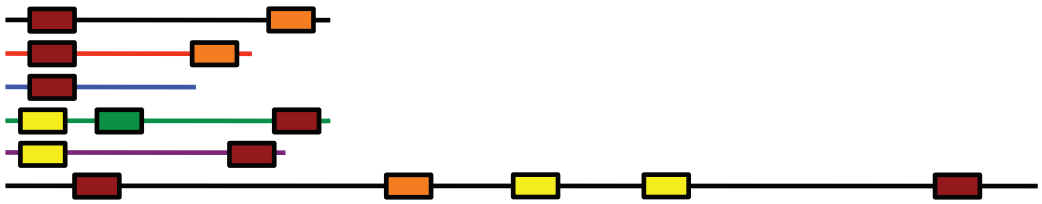
Wang, E.T., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, and C.B. Burge. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470-476.

Zhong, J.L., Z. Poghosyan, C.J. Pennington, X. Scott, M.M. Handsley, A. Warn, J. Gavrilovic, K. Honert, A. Kruger, P.N. Span et al. 2008. Distinct functions of natural ADAM-15 cytoplasmic domain variants in human mammary carcinoma. *Mol Cancer Res* **6:** 383-394.

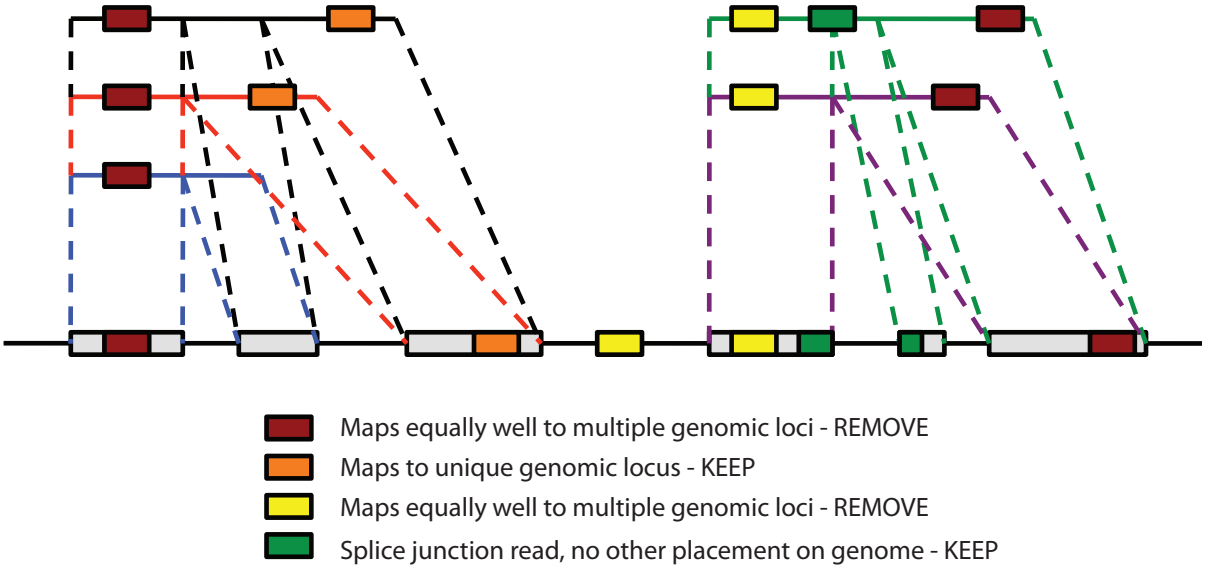
Supplementary Figure S1



(1) Align reads to reference file (transcriptome and genome), retaining all best placements per read.



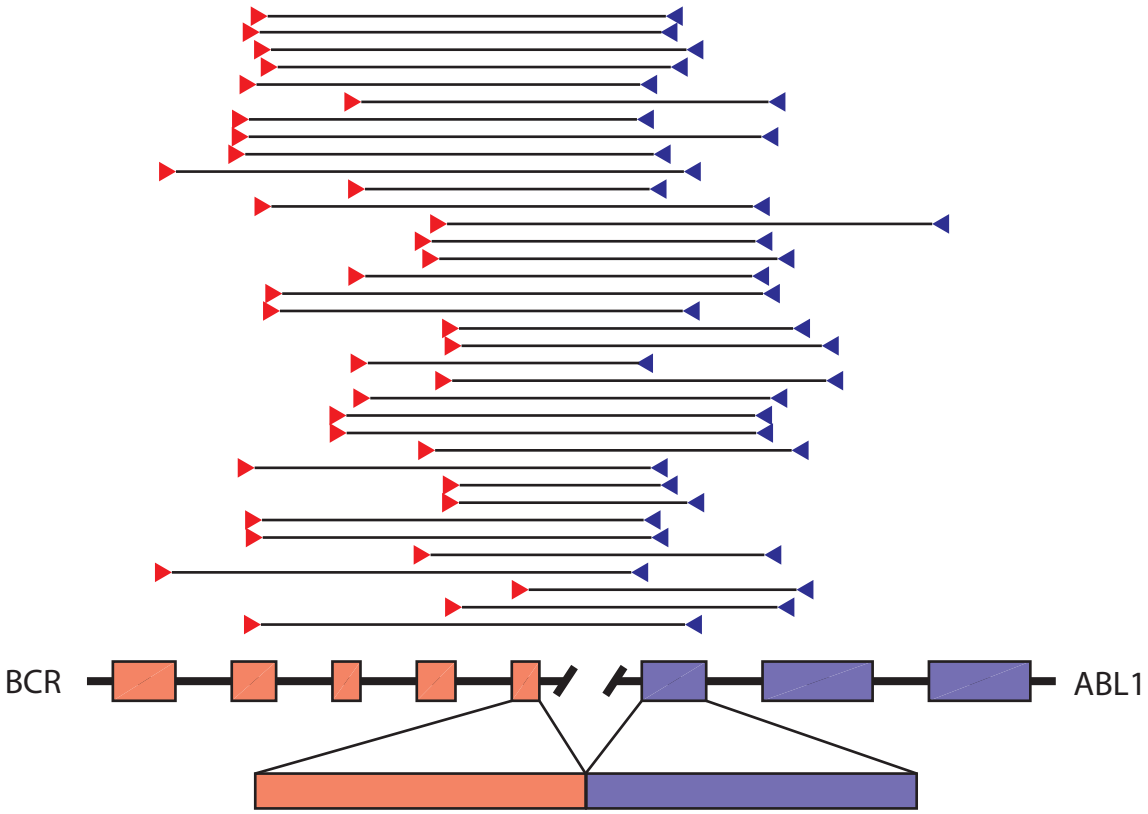
(2) Map transcript placements to genomic coordinates, preserving introns. Filter reads with ambiguous placement on genome.



(3) Isolate mate pairs where both reads independently map to a unique genomic locus (not shown).

Supplementary Figure S1: Schematic of multi-tiered alignment strategy. Reads are aligned to a reference sequence composed of all (redundant) transcripts and the human genome using BWA (Lin and Durbin, *Bioinformatics*, 2009). Reads aligning to transcripts are mapped to their genomic coordinates; reads spanning splice junctions are split. Mate pairs for which both reads map to a unique genomic locus are kept for further consideration. Pairs mapping to different loci are examined for candidate gene fusions (Methods). Pairs mapping to a single locus are examined for sequence mutations, gene expression levels, alternative splicing, and allele-specific expression.

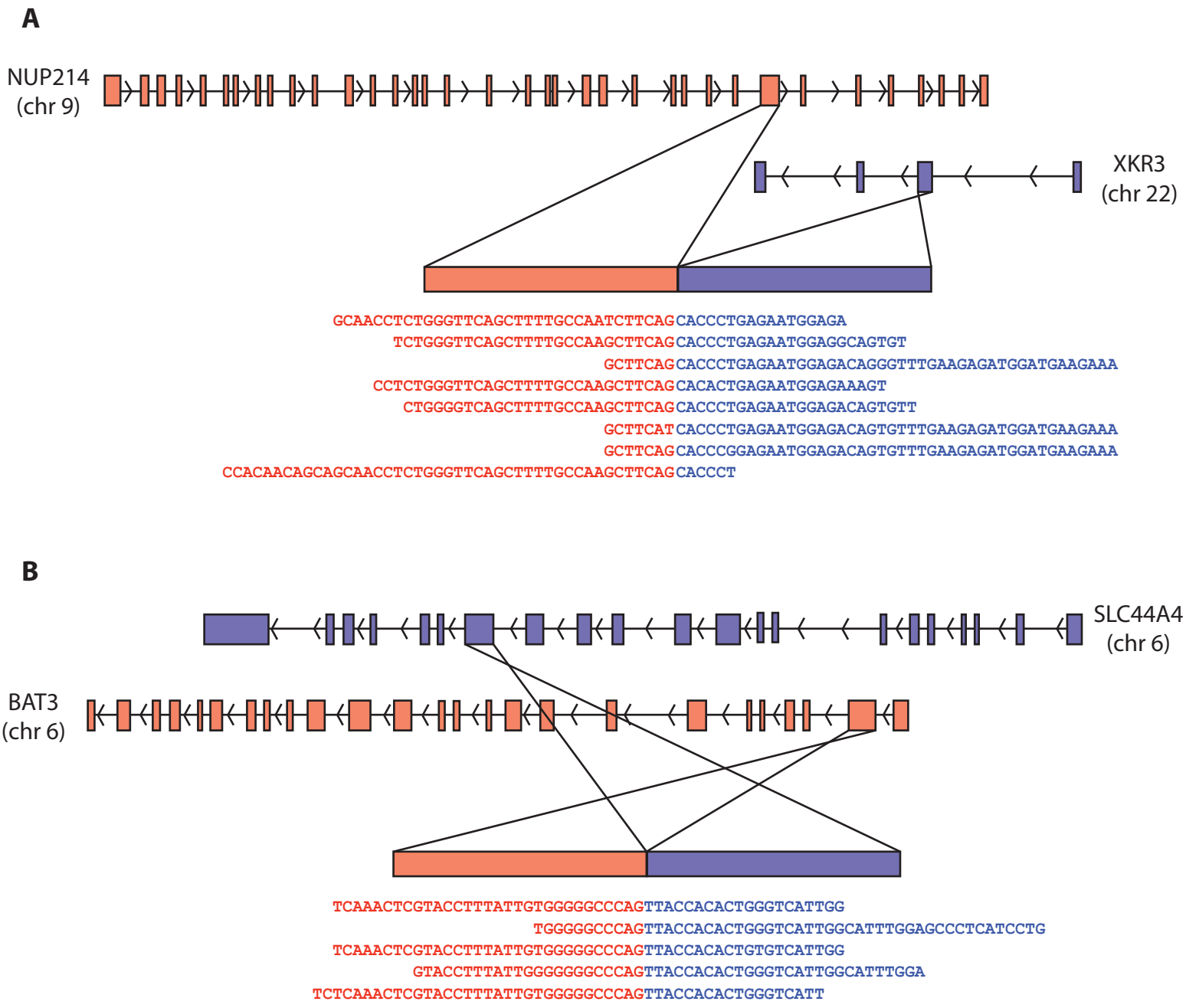
Supplementary Figure S2



AGCAGAGTTCAAAAGCCCTTCAGAGGCCAGTAGCATCTGACTTTGAGCCCC
CTCAGCCACTGGATTTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGC
CATCGTCCACTCAGCCACTGGATTTAAGCAGAGTTCAAAAGCCCTTCAGCG
ATTTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGA
GTCATCGTCCACTCAGCCACTGGATTTAAGCAGAGTTCAAAAGCCCTTCAG
GTCCACTCAGCCACTGGATTTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCA
AGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAGCCTC
CATCGTCCACTCAGCCACTGGATTTAAGCAGAGTTCACAAGCCCTTCAGCG
AGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAGCCTC
AGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGCGCCTC
TTTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAG
CATCGTCCACTCAGCCACTGGATTTAAGCAGAGTTCAAAAGCCCTTCAGCG
GTCATCGTCCACTCAGCCACTGGATTTAAGCAGAGTTCAAAAGCCCTTCAG
CTGGATTTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACT
CATCGTCCACTCAGCCACTGGATTTAAGCAGAGTTCAAAAGCCCTTCAGCG
AAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAGCCT
AAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAGCCT
GCAGAGTGCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAGCCTCA
ATTTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGA
GCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAGCCTCA
CCACTAGATTTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTG
AGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGAGCCTC
ATTTAAGCAGAGTTCAAAAGCCCTTCAGCGGCCAGTAGCATCTGACTTTGA

Supplementary Figure S2: BCR-ABL1 gene fusion in the K-562 leukemia cell line. A chimeric transcript joining exon 14 of BCR on chromosome 22 and exon 2 of ABL1 on chromosome 9 is implicated by 37 distinct read-pairs (top) and 23 individual fusion-spanning reads (bottom).

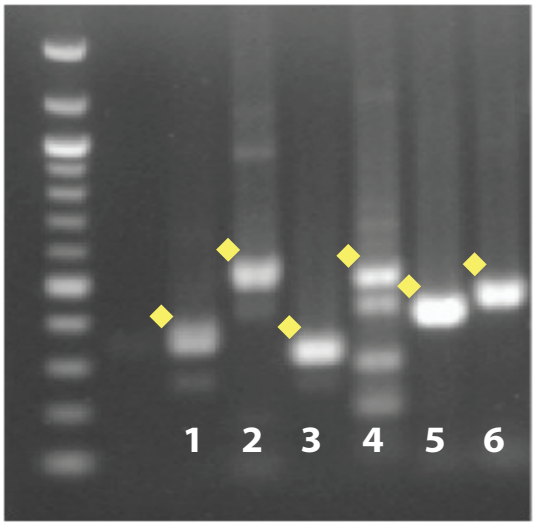
Supplementary Figure S3



Supplementary Figure S3: Novel gene fusions in K-562 leukemia cell line. (A) NUP214-XKR3 gene fusion. 44 read-pairs (not shown) and 8 individual fusion-spanning reads (bottom) support a chimeric transcript involving NUP214 and XKR3. (B) BAT3-SLC44A4 gene fusion. 22 read-pairs (not shown) and 5 individual fusion-spanning reads (bottom) support a chimeric transcript involving BAT3 and SLC44A4.

Supplementary Figure S4

A



Gene or Gene Fusion	Expected Product
(1) BAT3 (5′ region)	368 bp
(2) BAT3 - SLC44A4	541 bp
(3) BCR (5′ region)	317 bp
(4) BCR - ABL1	507 bp
(5) NUP214 (5′ region)	400 bp
(6) NUP214 - XKR3	446 bp

B

BAT3 - SLC44A4 (sequence across fusion point)

CTTCGAGANCCCCAAATACTATCGGGGAAACGGAAGTGGCCGTCGGTGGCAGANACCTGTCNGCCATGGAGCCTAATGATAGTA
CCAGTACCGCTGTGGAGGAGCCTGACAGCTTGGAGGTGTTGGTGAAGACCTTGGACTCTCAAACTCGTACCTTTATTGTGGGGG
CCCAGTTACCACACTGGGTCATTGGCATTTGGAGCCCTCATCCTGACCCTTGTGCAGATAGCCCGGGTCATCTTGGAGTATATTGA
CCACAAGCTCAGAGGAGTGCAGAACCCTGTAGCCCGCTGCATCATGTGCTGTTTCAAGTGCTGCCTCTGGTGTCTGGAAAAATTT
ATCAAGTTCCTAAACCGCAATGCATACATCATGATCGCCATCTACGGGAAGAATTTCTGTGTCTCAGCCAAAAATGCGTTCATGCTA
CTCATGCGAAACATTGTCAGGGTGGTC

NUP214 - XKR3 (sequence across fusion point)

GTTTGGCCAAAGCAACGCTCCTGCTTTTGGGCAGAGTCCTGGCTTTGGACAGGGAGGCTCTGTCTTTGGTGGTACCTCAGCTGC
CACCNNNANCNNCAGCAACCTCTGGGTTCAGCTTTTGCCAAGCTTCAGCACCCTGAGAATGGAGACAGTGTTTGAAGAGATGG
ATGAAGAAAGCACAGGAGGAGTTTCATCTTCGAAAGAAGAAATAGTCCTTGGCCAGAGACTCCATCTAAGCTTTCCTTTTAGCATT
ATCTTCTCAACTGTTCTCTACTGTGGTGAGGTTGCCTTTGGTTTATACATGTTTGAAATTTATCGAAAAGCTAATGACACATTCTGGA
TGTCATTTGCC

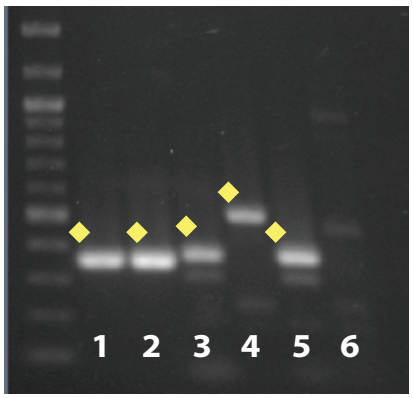
Supplementary Figure S4: RT-PCR validation of novel K-562 gene fusions. (A) RT-PCR verifying presence of fusion transcripts in K-562 mRNA. Primers were designed to amplify regions upstream of the fusion point (lanes 1, 3, 5) and spanning the fusion point (lanes 2, 4, 6) for 3 gene fusions discovered from RNA-seq: BCR-ABL1 (known), BAT3-SLC44A4 (novel), and NUP214-XKR3 (novel). All three chimeric transcripts produced bands the appropriate size. (B) Bands were excised from the gel and subjected to Sanger sequencing. Sequencing results confirmed the identities of the novel fusion products and mapped the fusion points to the same base positions as RNA-seq. Exon 2 of BAT3 is fused to exon 15 of SLC44A4, and exon 29 of NUP214 is fused to exon 2 of XKR3.

Supplementary Figure S5

A

SCAMP2 - WDR72
(sample M010403)

RT-PCR in:
M010403 (lanes 1,3,4)
501 Mel (lanes 2,5,6)



Gene or Gene Fusion	Expected Product
(1) Actin	368 bp
(2) Actin	368 bp
(3) SCAMP2 (5′ region)	376 bp
(4) SCAMP2 - WDR72	519 bp
(5) SCAMP2 (5′ region)	376 bp
(6) SCAMP2 - WDR72	N/A

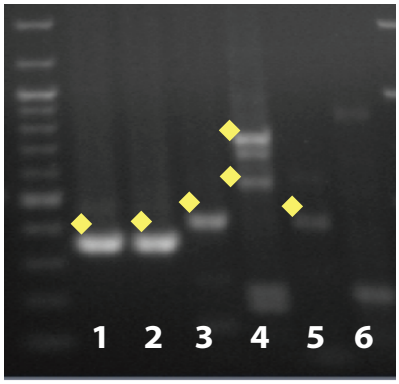
SCAMP2 - WDR72 (sequence across fusion point)

TCTGTGCCGTGCTCTCAGTCTTCCTCCTGCAGCGGGAACAGAGCCGTCCCTTTGTTATGGGCTACATGAATGAAAGGAAAGAGCCTTTTTA
CAAGGTACTTTTCTCTGGAGAAGTCTCAGGAAGAATTACTTTGTGGCACATCCCTGATGTTCCTGTATCCAAGTTTGATGGTTCTCCTAGAG
AGATACCAGTAACTGCCACCTGGACTCTTCAAGATAATTTTGATAAGCATGATACTATGTCACAAAGTATTATTGACTATTTCTCTGGGCTTAA
AGATGGGGCAGGAACTGCTGTAGTCACTTCATCAGAGTATATTCCAAGTCTTGATAAACTAATATGTGGCTGTGAAGATGGGACAATTATCAT
TACCCAGGCTTTGAATGCTGCCAAAGCAAGACTTCTGGAAGGTGGTTCTTTAGTAAAAGATTCTCCCCCTCATAAAGTTCTTA

B

RECK - ALX3
(sample M000921)

RT-PCR in:
M000921 (lanes 1,3,4)
M990802 (lanes 2,5,6)



Gene or Gene Fusion	Expected Product
(1) Actin	368 bp
(2) Actin	368 bp
(3) RECK (5′ region)	434 bp
(4) RECK - ALX3	572 bp AND 761 bp
(5) RECK (5′ region)	434 bp
(6) RECK - ALX3	N/A

RECK - ALX3 (sequence across fusion point)

Top Band

TTCCATCTGGAGATCCCTGTCTTCCGTACTTTTGTGTTCAAGGTTGCAAACTGGGAGAAGCTTCTGATTTCATTGTCCGTCAAGGGA
CACTAATCCAGGTGCCATCATCTGCAGGGGAAGTTGGTTGTTATAAAATCTGTTCATGTGGACAAAGTGGACTCTTAGAAAACTGTA
TGGAAATGCACTGTATAGACCTCCAGAAGTCTTGTATTGTTGGAGGAAAAAGAAAAAGTCTGGTTCCAGAANNGANANNAAGTN
NNAAAGCGCGAGCGTTATGGGAAGATCCAGGAGGGGNGGAACCCCTTCACGGCTGCCTATAACATCTCTGTGCTGCCCCGTACTGA

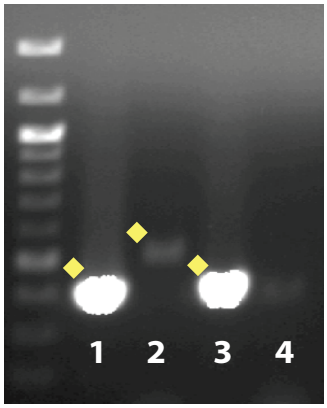
Bottom Band

GAGATGTGGAAAGCAATAGCTTGTTCACTGCAGATTAAACCTTGTCATAGTAAATCTCGGGGAAGTATTATTTGCAAATCAGATTGTG
TGGAGATTCTTAAAAAATGTGGAGACCAGAACAAATTCCCTGAAGACCACACAGCTGAAAGTATTTGTGAGCTTCTGTCACCTACAG
ATGATCTGAAGAATTGTATACCTTTGGATACATACCTCAGGCCAAGTACTTTAGGTAACATTGTAGAAGAAGTGACTCATCCCTGTAAC
CCAAATCCTTGCCCTGCCAATGAGCTCTGTGAAGTAAACCGAAAAGGATGTCCATCTGGAGATNCNNGTCTTCCGTACTTTTGTGTTC
AAGGTCTGGTTCCAGAACCGCAGAGCCAAGTGGCGAAAGCGCGAGCGTTATGGGAAGATCCAGGAGGGGCGGAACCCCTTCACGG
CTGCCTATAACATCTCTGTGCTGCCCCGTACTGACAGCCACCCTCAGCTGCAGAACTCCCTGNGGGCCA

C

C9orf127 - TLN1
(sample M000921)

RT-PCR in:
M010403 (lanes 1,2)
501 Mel (lanes 3,4)



Gene or Gene Fusion	Expected Product
(1) TLN1 (3′ region)	389 bp
(2) C9orf127 - TLN1	506 bp
(3) TLN1 (3′ region)	389 bp
(4) C9orf127 - TLN1	N/A

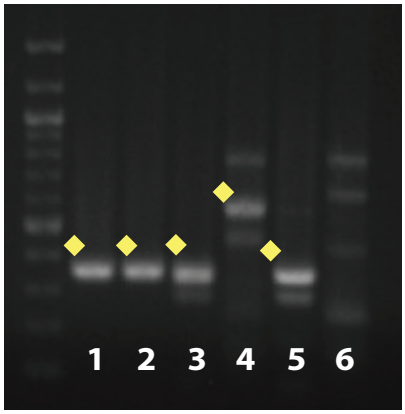
C9orf127 - TLN1 (sequence across fusion point)

ACTGCCCNGACCAAAGTGTGACTGTGTATTTCCGGTCCGGGGCACCCCCTGTCATCAATCCCCTGCATACACACTTCCCAGGGGACACAGC
TGTGCCTGGGGTTTTCTCACTGACCCTCAGCTGGACACTGCCCAACCGCACCTCAGGCATCTTTAACGTCAGCAGCCCCTTACCTGGGGAC
TGGTTCTTGGCTGCCCACCTTCCCCAGGCCCACGGCCACATCTCTGTCAAGGATGCGCTAATGCAGCTCGCCAAAGCTGTGGCAAGTGCTG
CAGCTGCCCTGGTCCTCAAGGCCAAGAGTGTGGCCCAGCGGACAGAGGACTCGGGACTTCAGACCCAAGTTATTGCTGCAGCAACACAG
TGTGCCCTATCCACTTCCCAACTAGTGGCCTGTACTAAGGTGGTGGCACCTACAATCAGCTCACCTGTCTGCCAAGAGCAACTGAA

D

KCTD2 - ARHGEF12
(sample M000216)

RT-PCR in:
M000216 (lanes 1,3,4)
501 Mel (lanes 2,5,6)



Gene or Gene Fusion	Expected Product
(1) Actin	368 bp
(2) Actin	368 bp
(3) KCTD2 (5′ region)	345 bp
(4) KCTD2 - ARHGEF12	569 bp
(5) KCTD2 (5′ region)	345 bp
(6) KCTD2 - ARHGEF12	N/A

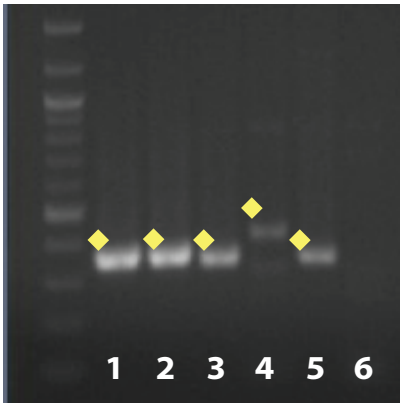
KCTD2 - ARHGEF12 (sequence across fusion point)

TTTGGTCCTATCCTCAACTACCTCCGCCACGGGAAACTCATCATCACTAAGGAGTTGGCAGAAGAAGGTGTGCTGGAGGAAGCGGAGTTTT
ACAACATCGCGTCCCTTGTGCGGCTGGTTAAGGAAAGGATACGGGACAATGAGAACAGAACTTCACAAGGCCCCGTGAAGCACGTGTACA
GAGTCCTGCAGTGTCAGGAAGAAGAGCTCACGCAGATGGTGTCCACGATGTCCGACGGCTGGAAATTCGAACAGGAAACTATTTGATCCT
TGATGGCTATGACCCAGTGCAGGAGAGTTCCACAGATGAGGAGGTTGCTTCCTCACTTACCCTGCAGCCCATGACAGGCATCCCTGCTGTG
GAATCCACCCACCAGCAGCAACATTCTCCTCAGAATACTCACTCCGATGGGGCAATTTCACCATTCACCCCCGAATTTCTGGTCCAGCAGCG
CTGGGGAGCTATGGAGTATTCCTGTTTTGAGATCCAGAGT

E

GCN1L1 - PLA2G1B
(sample M980409)

RT-PCR in:
M980409 (lanes 1,3,4)
501 Mel (lanes 2,5,6)



Gene or Gene Fusion	Expected Product
(1) Actin	368 bp
(2) Actin	368 bp
(3) GCN1L1 (5′ region)	344 bp
(4) GCN1L1 - PLA2G1B	419 bp
(5) GCN1L1 (5′ region)	344 bp
(6) GCN1L1 - PLA2G1B	N/A

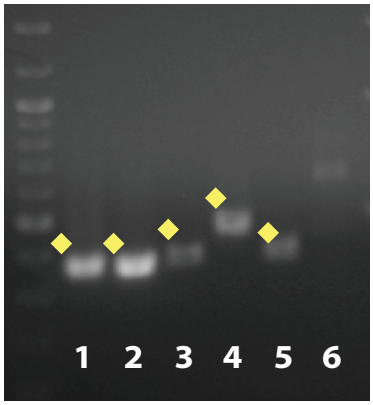
GCN1L1 - PLA2G1B (sequence across fusion point)

ACACCGCCCTGCGCGCGGGCCAGCGGGTTATCTCCATGTACGCTGAGACAGCCATCGCCCTGCTGCTGCCCCAGCTAGAGCAAGGCCTCTT
TGATGACCTTTGGAGAATCAGTGGCCGCCGCCGACAGCGGCATCAGCCCTCGGGCCGTGTGGCAGTTCCGCAAAATGATCAAGTGCGTGAT
CCCGGGGAGTGACCCCTTCTTGGAATACAACAACTACGGCTGCTACTGTGGCTTGGGGGGCTCAGNNNNNCCCGTGGATGAACTGGACNA
GTGCTGCCNGACACATGACAACTGCTATG

F

ANKHD1 - C5orf32
(sample M990802)

RT-PCR in:
M000216 (lanes 1,3,4)
501 Mel (lanes 2,5,6)



Gene or Gene Fusion	Expected Product
(1) Actin	368 bp
(2) Actin	368 bp
(3) C5orf32 (3′ region)	380 bp
(4) ANKHD1 - C5orf32	453 bp
(5) C5orf32 (3′ region)	380 bp
(6) ANKHD1 - C5orf32	N/A

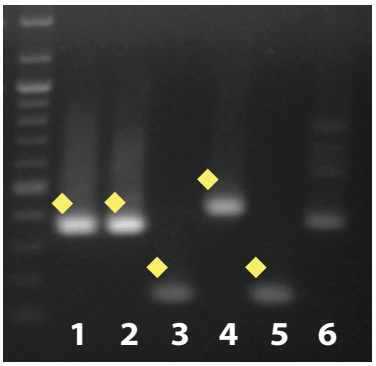
ANKHD1 - C5orf32 (sequence across fusion point)

TGTGGAGGTGNNNNTNTGGCAGTGACGAGGACGAAGTGTCCGAGNNTATGTNGTAGAAGNCCAAAGAAGANATGAGCTAGGACCATC
CNCNNNNCTCNCANCCTNCTGGACGNCTNNCTGTTGCT

G

RB1 - ITM2B
(sample M990802)

RT-PCR in:
M990802 (lanes 1,3,4)
501 Mel (lanes 2,5,6)



Gene or Gene Fusion	Expected Product
(1) Actin	368 bp
(2) Actin	368 bp
(3) RB1 (5′ region)	156 bp
(4) RB1 - ITM2B	421 bp
(5) RB1 (5′ region)	156 bp
(6) RB1 - ITM2B	N/A

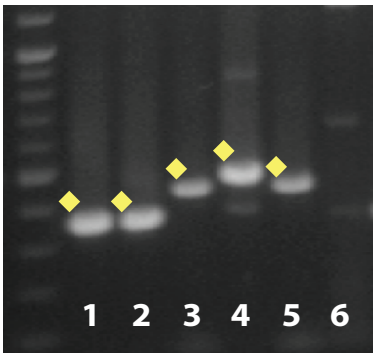
RB1 - ITM2B (sequence across fusion point)

AGANATNTNAAATAAANNNNNNCANNCNAGANTTNNNNNTCTCGTCNNNTTGAGTTTGAAGAAACAGAAGAACCTGATTTTACTGCATTA
TGTCAGAAATTAAAGATACCAGATCATGTCAGAGAGAGAGCTTGGTTAACTTGGGAGAAAGTTTCATCTGTGGATGGAGTATTGCCAGATGA
CGTGTACTACTGTGGAATAAAGTACATCAAAGATGATGTCATCTTAAATGAGCCCTCTNNAGNNNNNNNNCTGCTCTCTACCAGACAATTGA
AGA

H

SLC12A7 - C11orf67
(sample 501 Mel)

RT-PCR in:
501 Mel (lanes 1,3,4)
M000921 (lanes 2,5,6)



Gene or Gene Fusion	Expected Product
(1) Actin	368 bp
(2) Actin	368 bp
(3) SLC12A7 (5′ region)	429 bp
(4) SLC12A7 - C11orf67	464 bp
(5) SLC12A7 (5′ region)	429 bp
(6) SLC12A7 - C11orf67	N/A

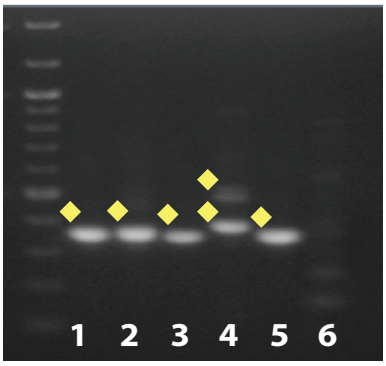
SLC12A7 - C11orf67 (sequence across fusion point)

TGCTGCACAACATGCGTGTGTACGGCACGTGCACGCTCNNCTCATGGCCCTGGTGGTCTTCGTGGGCGTCAAGTATGTCAACAAGCTGGCG
CTGGTCTTCCTGGCCTGCGTCGTGCTGTCCATCCTGGCCATCTATGCCGGCGTCATCAAGTCTGCCTTCGACCCCCCGGACATCCCCATTCTCC
TGGTGTGCAGCCTGCAGATGTGAAGGAAGTTGTTGAGAAGGGTGTACAGACTCTTGTGATTGGCCGAGGGATGAGTGAGGCCTTGAAGGTG
CCTTCATCAACTGTGGAGTACCTCAAGAAACATGGCATTGATGTGCA

I

CCT3 - C1orf61
(sample 501 Mel)

RT-PCR in:
501 Mel (lanes 1,3,4)
M990802 (lanes 2,5,6)



Gene or Gene Fusion	Expected Product
(1) Actin	368 bp
(2) Actin	368 bp
(3) CCT3 (5′ region)	353 bp
(4) CCT3 - C1orf61	381 bp and 515 bp
(5) CCT3 (5′ region)	353 bp
(6) CCT3 - C1orf61	N/A

CCT3 - C1orf61 (sequence across fusion point)

Top Band

CCCAACAGTGGNGATCAGTGCTTACCGCNNAGCATTGGATGATATGATCAGCACCCTAAAGAAAATAAGCCTTCATCTCCTCGTCCTCC
AGGCAGCACAAGCCATTGTGGAATCTCCACCAGGTGTACAGAACGGTGCC

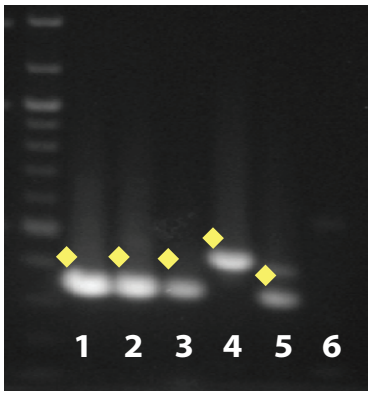
Bottom Band

CAGGATGAAGAGGTTGGAGATGGGACCACATCAGTAATTATTCTTGCAGGGGAAATGCTGTCTGTAGCTGAGCACTTCCTGGAGCAGC
AGATGCACCCAACAGTGGTGATCANNNNNACCNCAAGGCATTGGATGATATGATCAGCACCCTAAAGAAAATAAGACCAAAAAGAAG
AATGTACTTCATCTGGTTGGGCTGGATTCCCTCTGATAAGCCTTCCCAGTTGACTGAAAGATGAGNCTAGGCTCTAGCAAGTTGAAGTC
AAACCAGCTCCTTCAAGAAGCTTTGAGCA

J

GNA12 - SHANK2
(sample 501 Mel)

RT-PCR in:
501 Mel (lanes 1,3,4)
M990802 (lanes 2,5,6)



Gene or Gene Fusion	Expected Product
(1) Actin	368 bp
(2) Actin	368 bp
(3) GNA12 (5′ region)	357 bp
(4) GNA12 - SHANK2	448 bp
(5) GNA12 (5′ region)	357 bp
(6) GNA12 - SHANK2	N/A

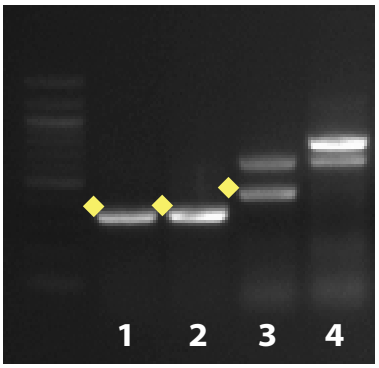
GNA12 - SHANK2 (sequence across fusion point)

GATGTTCCTGATGGCCTTCGAGAACAAGGCGGGGCTGCCTGTGGAGCCGGCCACCTTCCAGCTGTACGTCCCGGCCCTGAGCGCACTCTGG
AGGGATTCTGGCATCAGGGAGNNNCAGCCGGAGAAGCGAGTTTCAGCTGGTGGCCATAATAGCAGGCAACTTTGAGCTGGCAGAATACAT
CAAGAACCACAAGGAAACAGACATTGTGCCCTTCCGAGAGGNCCCGGCGTACTCCAAC

K

PARP1 - MIXL1
(sample 501 Mel)

RT-PCR in:
501 Mel (lanes 1,3)
M000921 (lanes 2,4)



Gene or Gene Fusion	Expected Product
(1) Actin	368 bp
(2) Actin	368 bp
(4) PARP1 - MIXL1	439 bp
(6) PARP1 - MIXL1	N/A

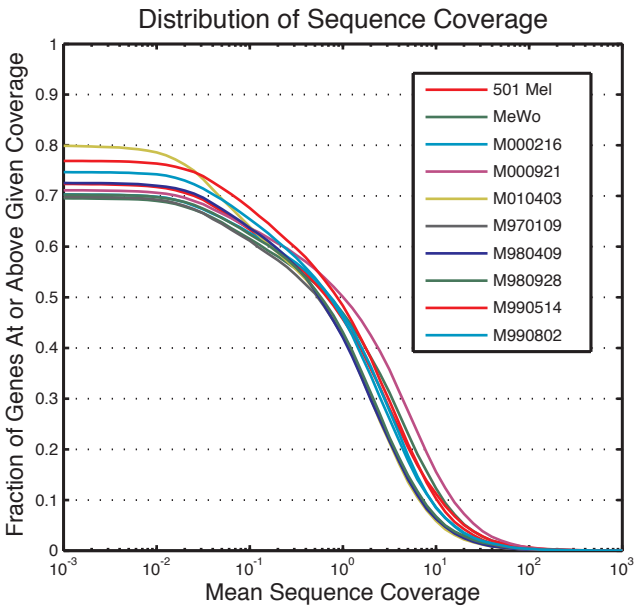
PARP1 - MIXL1 (sequence across fusion point)

AAGAGCAGCTCCCAGGAGTCAAGAGTGAAGGAAAGAGAAAAGGCGATGAGGTGGATGGAGTGGATGAAGTGGCGAAGAAGAA
ATCTAAAAAAGAAAAAGACAAGGATAGTAAGCTTGAAAAAGCCCTAAAGGTATGGTTCCAGAACAGGCGTGCCAAGTCTCGGCG
TCAGAATGGGAAATCCTTCCTACCTTTGGCTGGGCCGGATATTATCCTCAACCACTGTGCTCCTGGAACTGAAACGAAATGTCTGA
AGCCCCACCCGCCTCTTGAGGTAGATGTGAACTGCCTGCCCGTAACCAAACTCGGTTGGAGGGGGCATCTCTGACTCTAGCTCCC
AAGGTCAGAATTTTGTAACCCGTTCACCAAAAAAAGAAANACGGTTCCCCTCTCTCTG

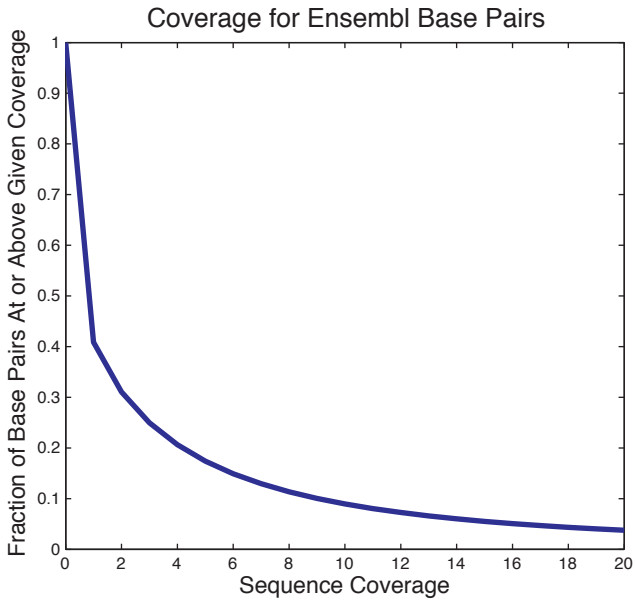
Supplementary Figure S5: RT-PCR validation of novel melanoma gene fusions. RT-PCR and sequencing reactions were performed as in Supplementary Figure S4, with additional cell lines not harboring the corresponding gene fusion used as negative controls. (A) SCAMP2-WDR72; (B) RECK-ALX3; (C) C9orf127-TLN1; (D) KCTD2-ARHGEF12; (E) GCN1L1-PLA2G1B; (F) ANKHD1-C5orf32; (G) RB1-ITM2B; (H) SLC12A7-C11orf67; (I) CCR3-C1orf61; (J) GNA12-SHANK2; (K) PARP1-MIXL1.

Supplementary Figure S6

A



B

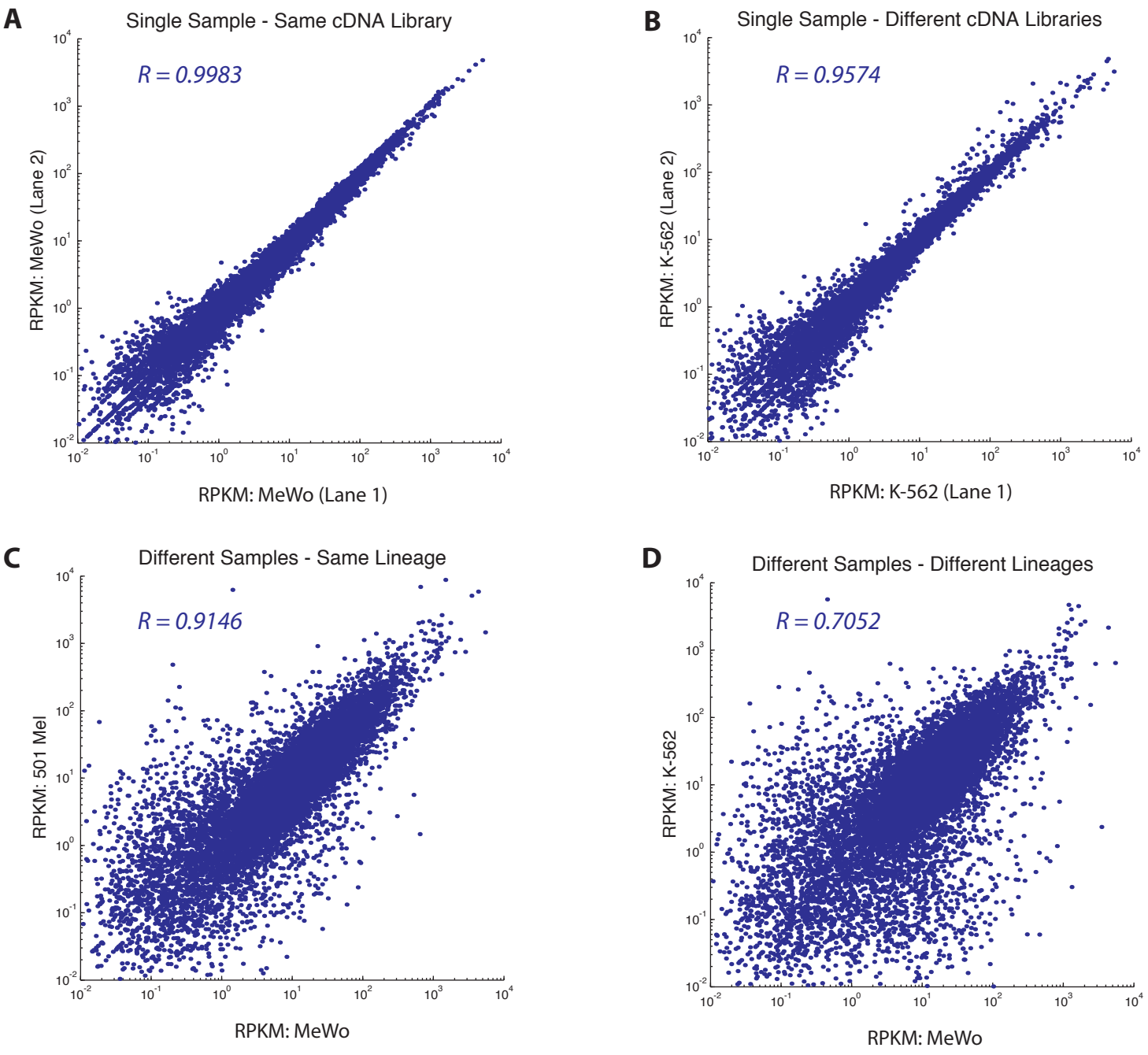


MeWo (1 Illumina lane)

	Percent of Genes at Threshold (avg)	Percent of Bases at Threshold
0.1x	61.5%	N/A
1x	46.9%	40.9%
10x	12.3%	9.0%
20x	5.2%	3.8%
100x	0.4%	0.3%

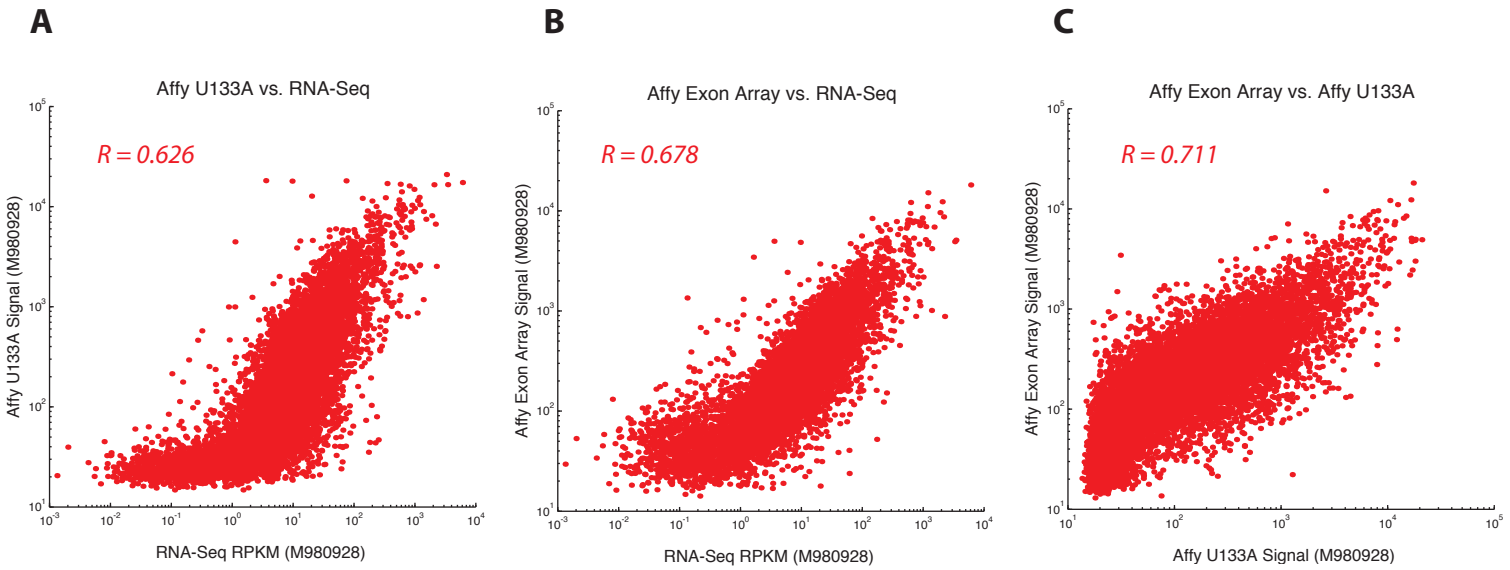
Supplementary Figure S6: Sequence Coverage Per Lane of RNA-seq. (A) Distribution of sequence coverage for all protein-coding genes in Ensembl. Between 70 and 80% of genes are expressed at detectable levels; between 40 and 50% of genes are expressed at greater than 1x mean coverage; and between 5 and 15% of genes are expressed at greater than 10x mean coverage. (B) Nucleotide-level view of coverage for all bases contained in an Ensembl transcript, for one Illumina lane of MeWo. Each base is covered by an integral number of reads.

Supplementary Figure S7



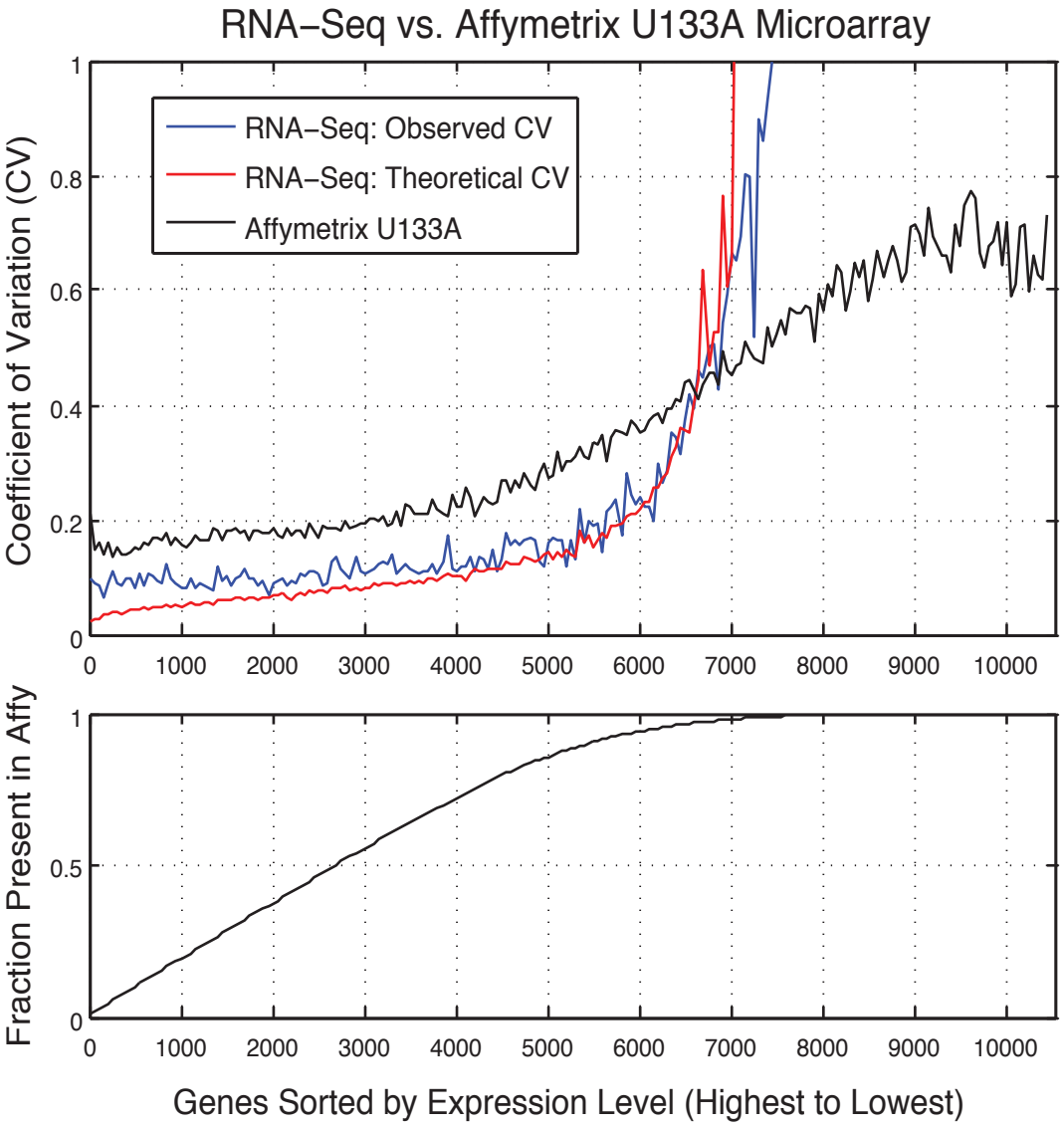
Supplementary Figure S7: Reproducibility of Gene Expression in RNA-seq. Gene expression levels were estimated for 17.343 protein-coding genes in Ensembl (at least 50% alignable) according to the observed mean sequence coverage over each gene. Gene expression levels are quantified according to the RPKM measure (reads per kilobase per million reads mapped). (A) Single cDNA library (MeWo) sequenced in two Illumina lanes. (B) Two independent cDNA libraries from the same sample (K-562) sequenced in two Illumina lanes. (C) Different samples of the same lineage (melanoma cell lines MeWo and 501 Mel). (D) Different samples of different lineages (melanoma cell line MeWo and leukemia cell line K-562).

Supplementary Figure S8



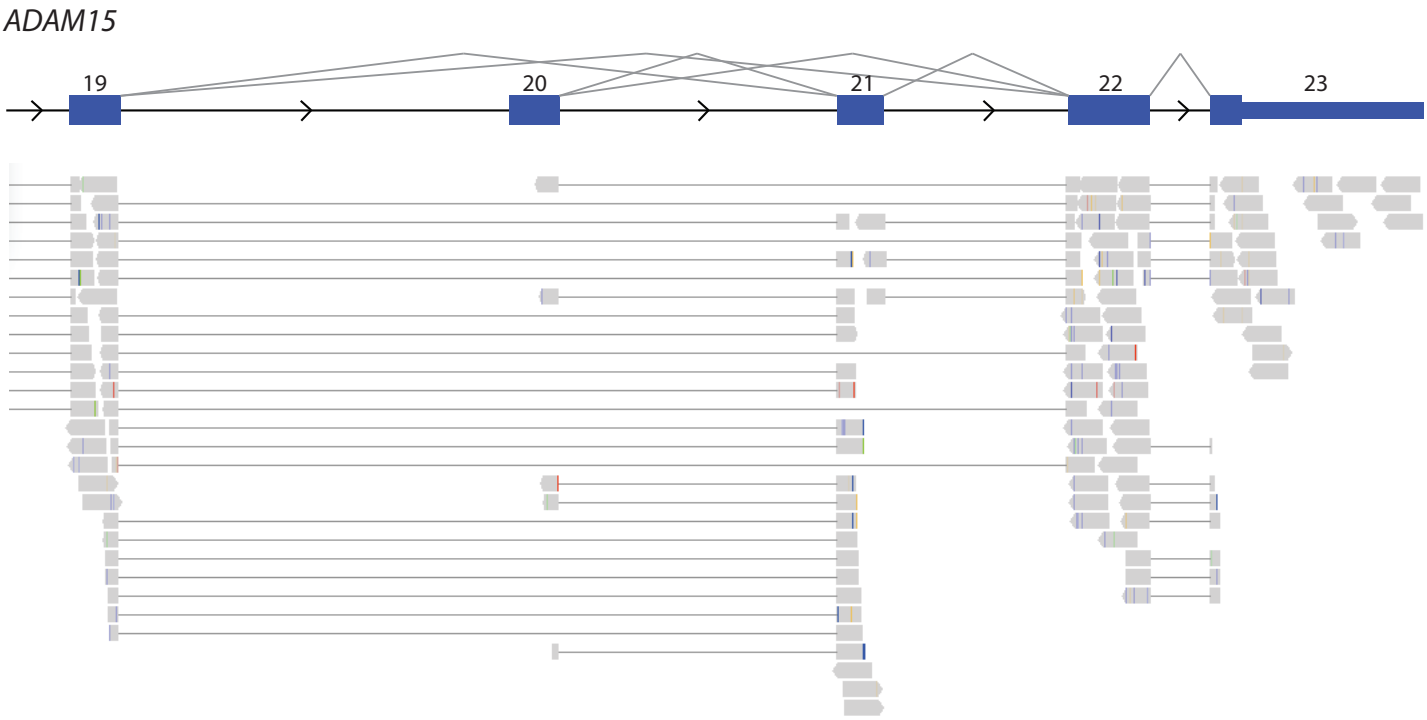
Supplementary Figure S8: Comparison of Gene Expression Platforms. Affymetrix U133A and Human Exon microarrays were run for 9 melanoma samples (including M980928, above). 10,602 genes were common to all 3 platforms. (A) Affy U133A vs. RNA-seq. (B) Affy Human Exon array vs. RNA-seq. (C) Affy Human Exon array vs. Affy U133A.

Supplementary Figure S9



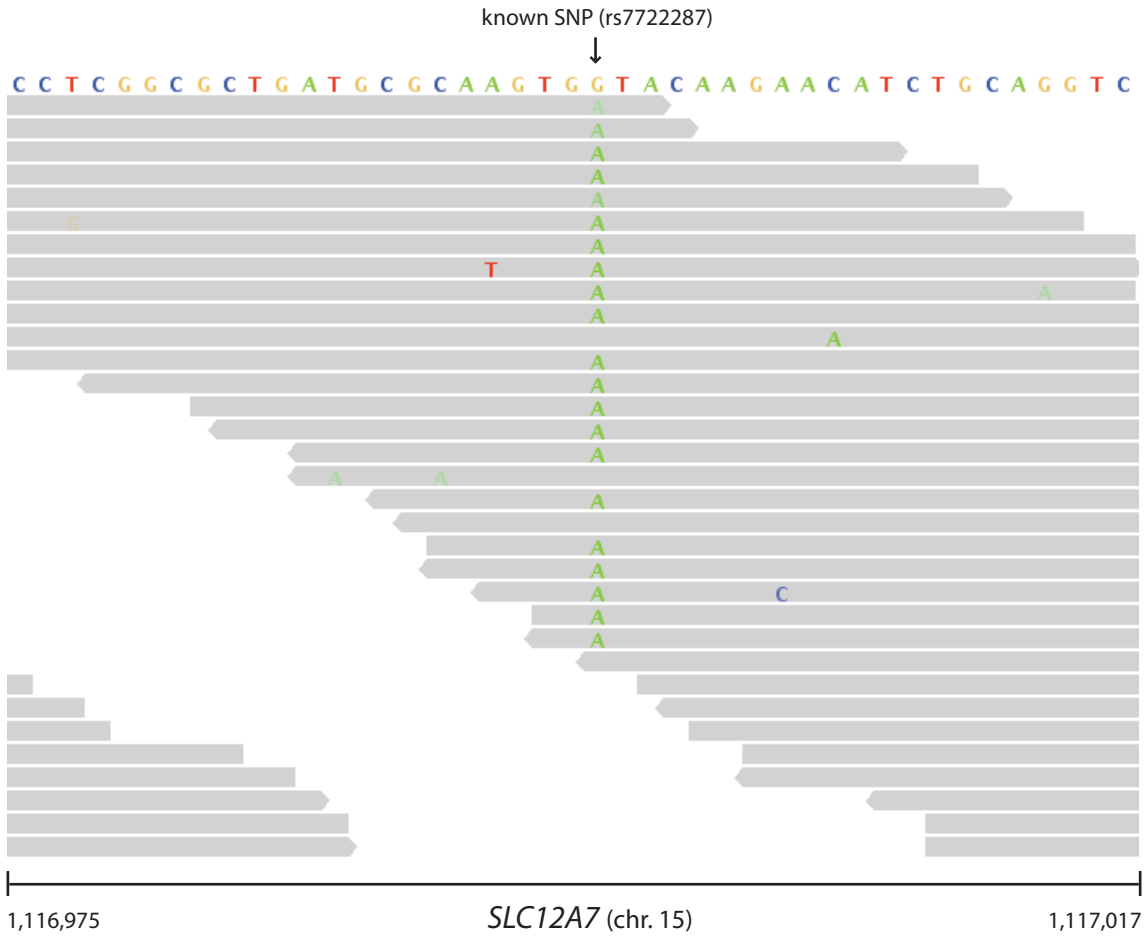
Supplementary Figure S9: Comparing Precision Between RNA-seq and Microarrays. *Top*: Coefficient of variation for replicate K-562 datasets (2 RNA-seq libraries and 18 Affymetrix U133A microarrays). Genes are sorted from highest expression to lowest expression in each platform, and CVs are averaged in bins of 50 genes. Theoretic CV estimates for RNA-seq are based on a Poisson model of the number of reads mapping to each gene, showing good agreement with the empirical CV. RNA-seq estimates are more precise than Affymetrix for the 6,700 most highly-expressed genes. *Bottom*: Cumulative distribution of expressed genes (i.e., called "present" in at least 9 of 18 replicate Affymetrix arrays). The 6,700 genes called more precisely by RNA-seq account for 97% of all expressed genes.

Supplementary Figure S10



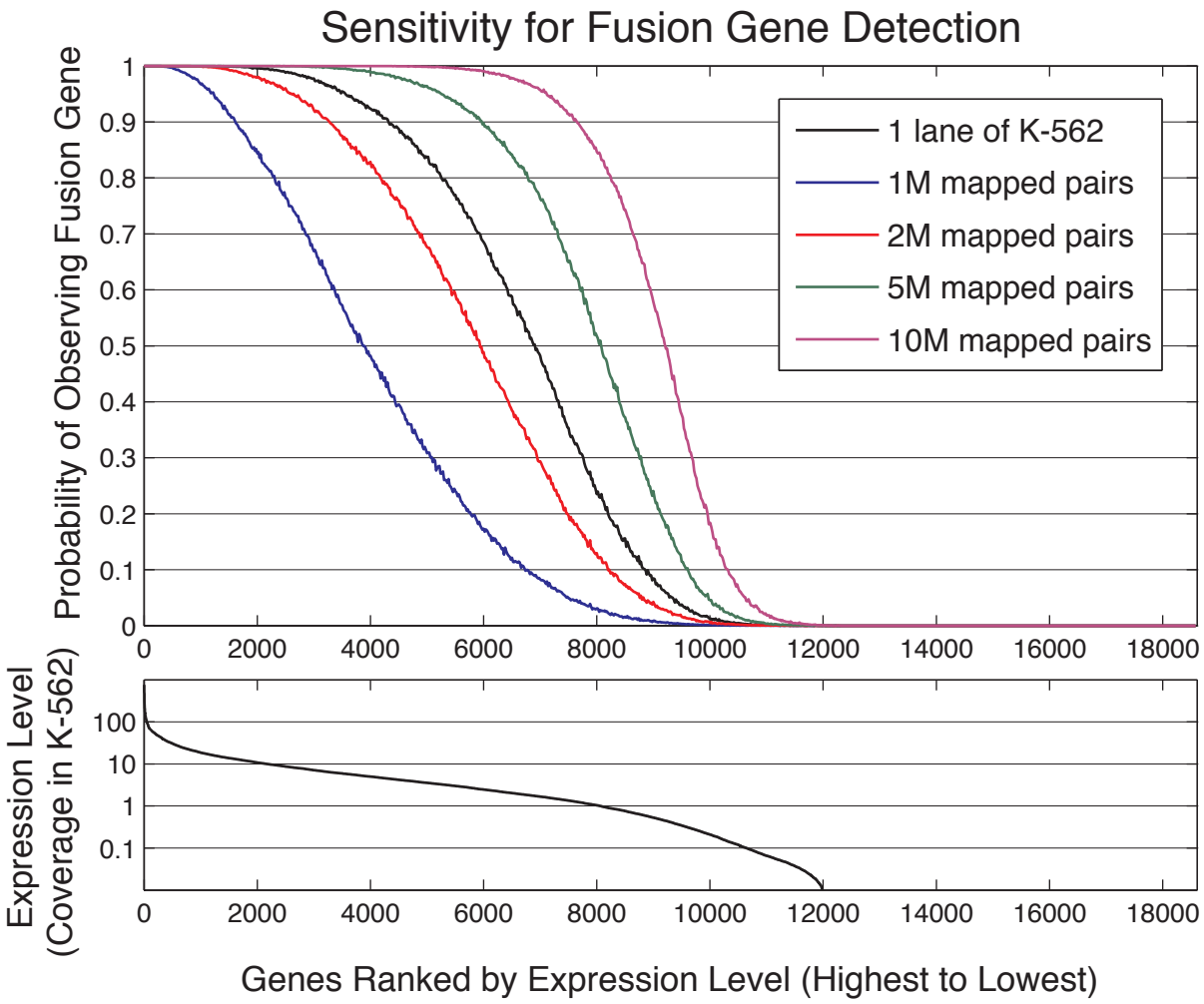
Supplementary Figure S10: Alternatively spliced transcripts are implicated by junction-spanning reads. Iluumina 51-mer reads mapping to the 3′ end of *ADAM15* in melanoma sample M000921 are shown as gray boxes (arrowheads denote directionality of reads; colored bars indicate single base mismatches with respect to the reference human genome hg18). Individual reads spanning exon-exon junctions are fractured and connected by thin lines across the intervening intronic sequence. Multiple *ADAM15* isoforms co-occur in this sample, as evidenced by overlapping junctions. Alternatively-spliced transcripts joining exons 19 and 22, and exons 20 and 22, have been linked to poorer relapse-free survival of node-negative breast cancer patients (Zhong et al. 2008). This figure was generated using the Integrative Genomics Viewer (IGV): http://www.broadinstitute.org/igv.

Supplementary Figure S11



Supplementary Figure S11: Allele-specific expression in melanoma. SLC12A7 exhibits allele-specific expression in melanoma short-term culture M000921, as seen in the allelic ratios at heterozygous SNP rs7722287. The reference base, G, is observed 4 times, and the alternate base, A, is observed 21 times. This figure was generated using the Integrative Genomics Viewer (IGV): http://www.broadinstitute.org/igv/.

Supplementary Figure S12



Supplementary Figure S12: Sensitivity for gene fusion detection as a function of gene expression level. In order to assess the utility of paired-end RNA-seq as a general method for gene fusion discovery, we considered each gene in Ensembl (18,615 total) to harbor a single hypothetical fusion point. Using the observed number of read pairs mapping to each gene in the K-562 sample, the length of each transcript, and the average cDNA fragment length amplified in an Illumina cluster, we calculated the probability of detecting this fusion point in each gene. We modeled the number of read-pairs mapping to each transcript as a Poisson distribution, the mapped position of read-pairs as a uniform distribution, and the cDNA fragment length as a Gaussian distribution (428 ± 105, as observed for K-562), and we postulated the breakpoint to occur in only 1 of 2 copies of the gene. Genes are sorted in descending order by observed sequence coverage, and the probability of discovering a hypothetical fusion point in each gene is shown for bins of 20 genes. The probability of detecting a gene fusion is greatest for genes expressed at high levels and decreases to zero for genes that are not expressed. (One lane of K-562 contained 2.9 million distinct pairs that mapped uniquely to annotated genes in Ensembl.)