MMBGX Supplementary Material

Ernest Turro, Alex Lewin, Anna Rose, Margaret J. Dallman and Sylvia Richardson

August 25, 2009

FIGURE S1. Plot of mean background probe log intensities by GC category and associated central 95% confidence interval on a whole-transcript GeneChip. On the real scale, the variability of non-specific hybridisation within GC categories is very large.

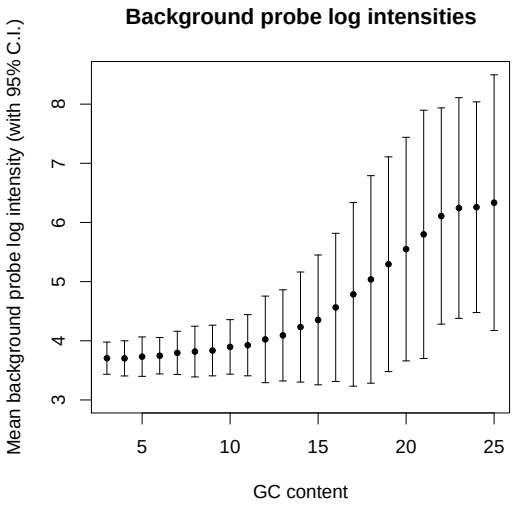


FIGURE S2. Plots showing the means and variances of log PM intensities of single-matching probes against the means and variances of PM intensities of multi-match probes. Each point represent one of the 33 arrays in the Affymetrix human Gene 1.0 ST array tissue mixture data set. Both the means and variances are consistently higher for multi-match probes, supporting the assumption that the contribution to the signal from each transcript is additive.

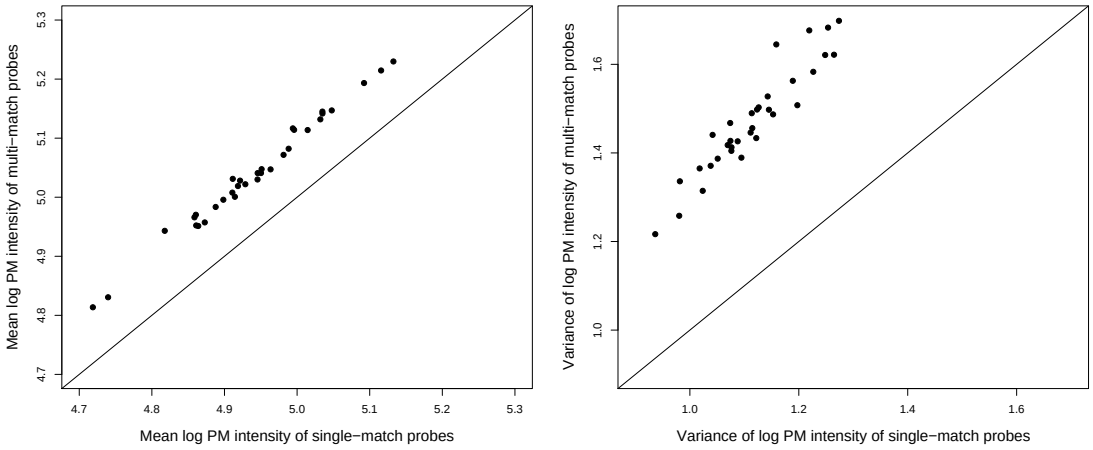


FIGURE S3. Speed-up from parallel computations. A human Gene array analysis was carried out with MMBGX three times using 1, 2, 3, 4, 5, 6, 7 and 8 cores on an 8-core machine. The average speed was recorded and a speedup graph plotted. The base of 2 hours 45 minutes was reduced to only 22 minutes using 8 cores, a speedup of 7.36.

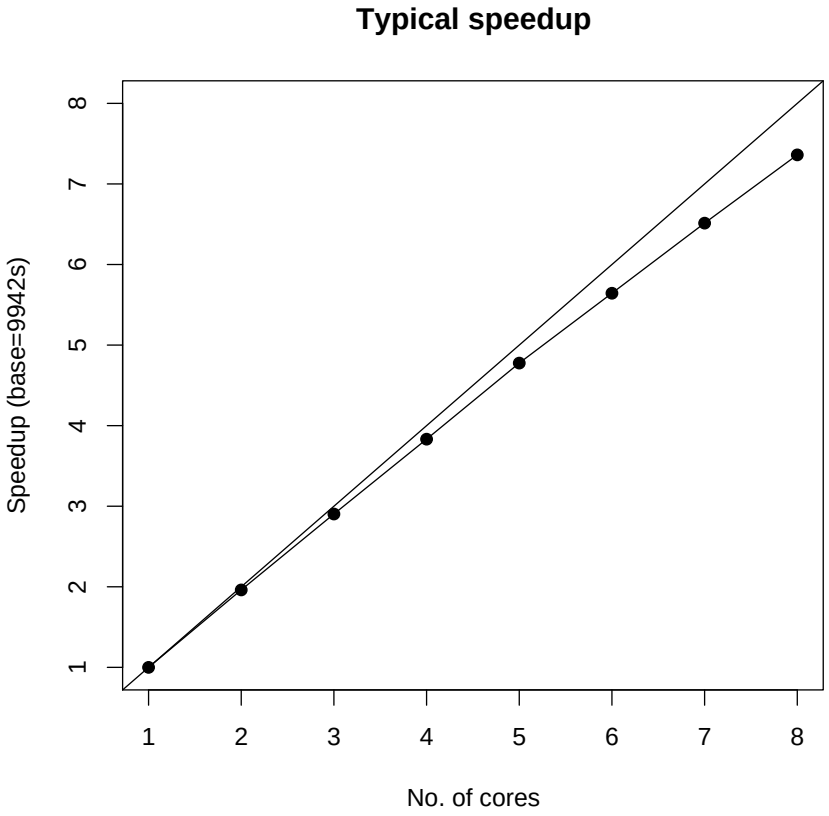


FIGURE S4. We ran MMBGX on a human Gene 1.0 ST array discarding (masking) probes that match to multiple gene-probesets. The two scatterplots show that the standard errors of the log expression measure for probesets without multi-match probes are not affected by the masking while the errors for multi-mapping probesets with at least one single-match probe increase as a result of the masking. The histogram of standard errors for multi-mapping probesets with no single-match probes is mostly within the range of single-mapping probesets (standard error < 1.5), indicating that the expression of genes that are completely eliminated by masking can be well estimated by MMBGX.

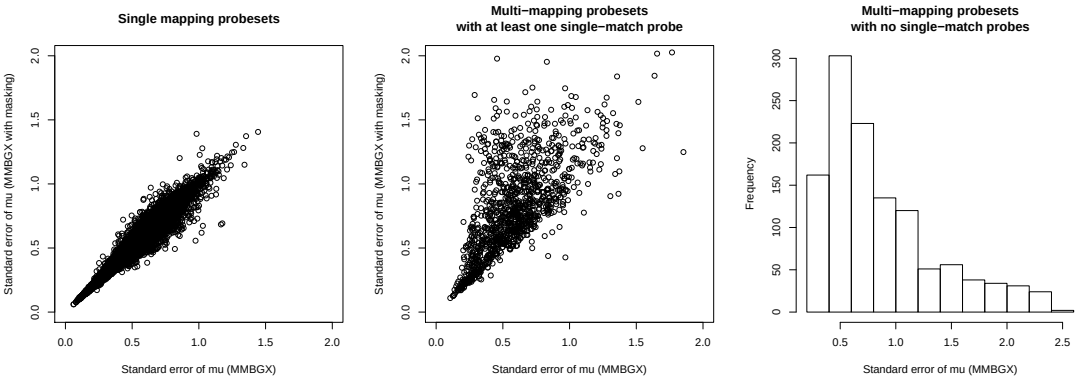


FIGURE S5. Full representation of the 2309 multi-mapping probesets on the human Gene 1.0 ST array. Red dots represent probesets while black dots represent probes. Small groups of two or three probesets are near the periphery of the figure, while more complicated structures appear in the centre.

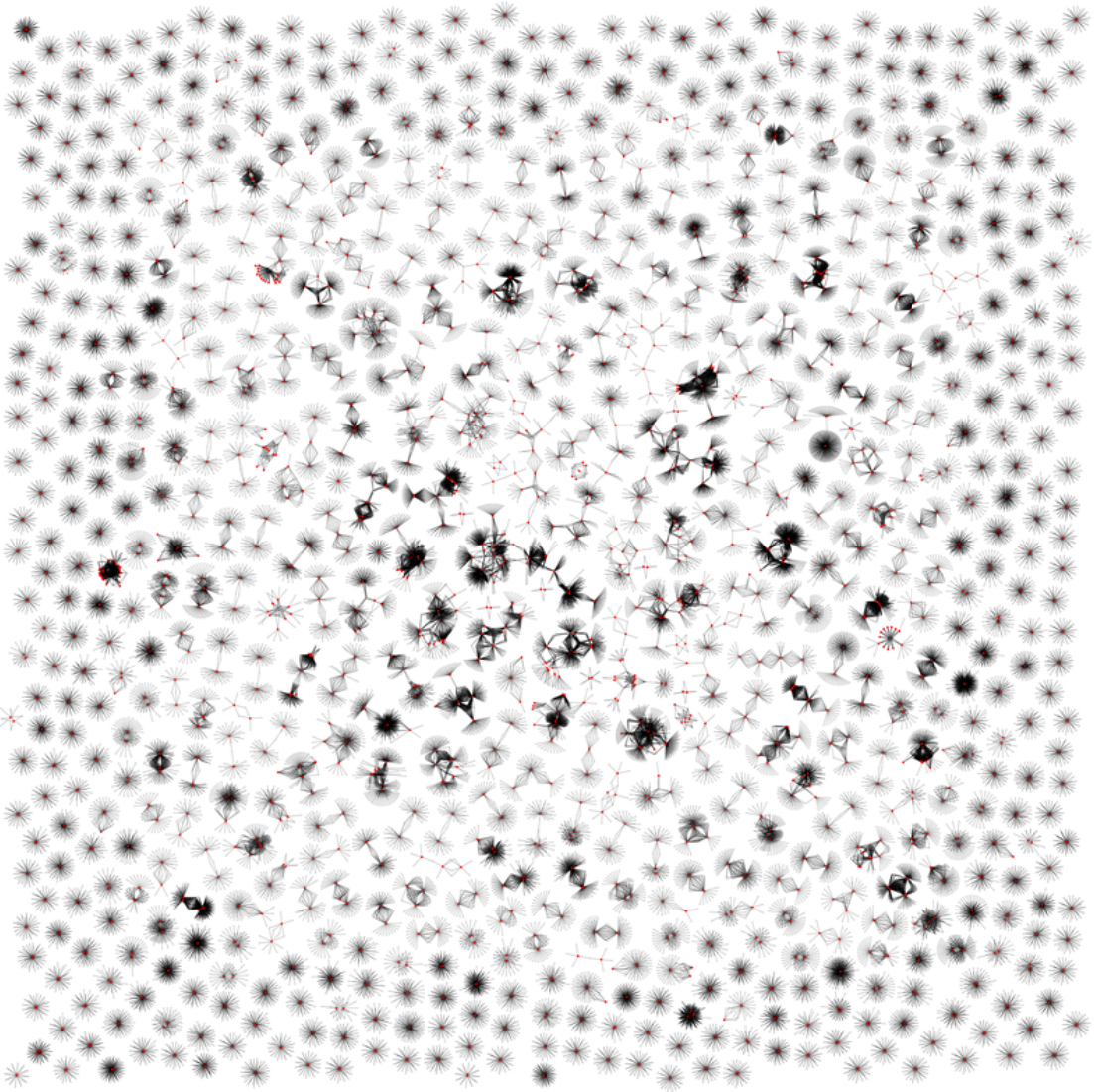


FIGURE S6. Histogram of the proportion of multi-match probes in each transcript-targeting probeset on the human Exon 1.0 ST array. About half the probesets are made up of 90% or less multi-match probes.

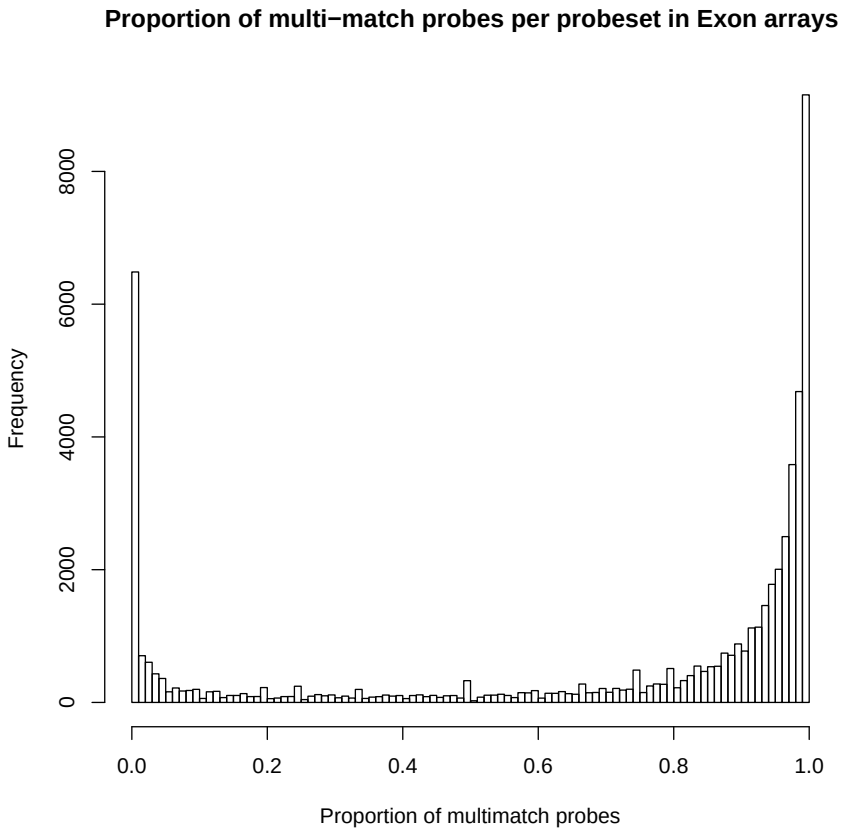
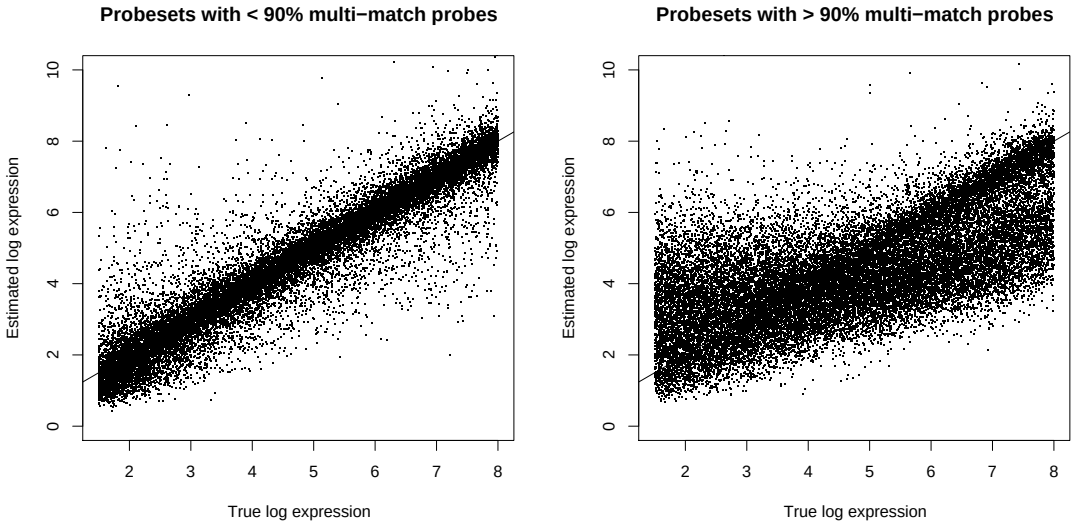


FIGURE S7. Scatterplots showing the performance of MMBGX on simulated Exon array data. MMBGX recovers the signal very well for probesets with less than 90% multi-match probes. There is increased variability and some shrinkage when the proportion of multi-match probes exceeds 90%, but the error in the estimates increases commensurately: 95% central credible intervals from the posterior distributions contain the true simulated value about 95% of the time (not shown).



ALGORITHM S1. A fully Bayesian model of α_c, the mean of the prior on $\log(\sigma^2_{gc})$, exhibited slow mixing of the MCMC sampler, making reliable inference computationally inefficient. We therefore developed an Empirical Bayes algorithm that estimates α_c very accurately prior to making inference on the other parameters. Before MCMC sampling, the algorithm proceeds once as follows:

1. For each k, c, r
   (a) Set $\widehat{\gamma_{kcr}}$ to the empirical mean of the logarithm of background probes in affinity category k, array r and condition c.
   (b) Set $\widehat{\delta^2_{kcr}}$ to the empirical variance of the logarithm of background probes in affinity category k, array r and condition c.
2. For each j, c, r
   (a) Sample $\log \widehat{H_{jcr}}$ from $N(\widehat{\gamma_{k(j)cr}}, \widehat{\delta^2_{k(j)cr}})$
   (b) If $PM_{jcr} - \widehat{H_{jcr}} > 1$, then $\widehat{S_{jcr}} := PM_{jcr} - \widehat{H_{jcr}}$. Otherwise, return to step a.
3. For each g, c
   (a) $\widehat{\mu_{gc}} := (|R(c)| \cdot |J(g)|)^{-1} \cdot \sum_{R(c)} \sum_{J(g)} \log \frac{\widehat{S_{jcr}}}{|G(j)|}$
   (b) $\widehat{\sigma^2_{gc}} := (|R(c)| \cdot |J(g)| - 1)^{-1} \cdot \sum_{R(c)} \sum_{J(g)} \left[\log \frac{\widehat{S_{jcr}}}{|G(j)|}\right]^2 - (|R(c)| \cdot |J(g)|/(|R(c)| \cdot |J(g)| - 1)) \cdot \widehat{\mu_{gc}}^2$
4. For each c
   (a) $\alpha_c := G^{-1} \sum_g \log \widehat{\sigma^2_{gc}}$
   (b) $\widehat{\beta^2_c} := (G-1)^{-1} \cdot \sum_g \log \left[\widehat{\sigma^2_{gc}}\right]^2 - (G/(G-1)) \cdot \alpha_c^2$,

where $J(g)$ indexes the set of probes matching transcript g, $R(c)$ indexes the set of replicates in condition c, and G is the total number of transcripts. In step 2a, we capture the uncertainty in the true intensity of non-specific hybridisation by sampling from the GC-specific empirical distribution, $N(\widehat{\gamma_{k(j)cr}}, \widehat{\delta^2_{k(j)cr}})$. For a given GC category, the lower the value of PM_{jcr}, the more likely it is that the true H_{jcr} is located near the leftmost support of its prior distribution. The algorithm captures this property through resampling, essentially truncating $\widehat{H_{jcr}}$ to a maximum of PM_{jcr}.

All parameters other than α_c are estimated using MCMC samplers. In order to speed up convergence, we set the starting values for the samplers to the estimates $\widehat{S_{jcr}}$, $\widehat{H_{jcr}}$, $\widehat{\mu_{gc}}$, $\widehat{\sigma^2_{gc}}$ and $\widehat{\beta^2_c}$.

NOTE S1. Primers for PCR were designed in exons flanking spliced exons for Cd97, B4galt5, Clec5a, Rac1, Csf2rb2 and Slc23a2 with the following primer sequences (5'-3'):

mCD97-F	gagttacacctgcgtctgtaacc
mCD97-R2	ttccagccttgacgacagatgc
mCD97-R3	aagtagggaacggtggctcttg
mB4galt5-F1	catgacgtcgatcacatacctg
mB4galt5-R1	cttcctcagcagagcgtacctg
mClec5a-F1	agaaagagatcagatccctgaatc
mClec5a-R1	agtgtggatccttgtgttgcac
mRac1-F1	ctatgggacacagctggacaag
mRac1-R1	catccctaagatcaagcttcgtc
mCsf2rb2-F1	tcagtgtcctgtgagctcagtg
mCsf2rb2-F2	aagcattgaaggttctgtggc
mSlc23a2-F1	tcactacattgctgcagacaac
mSlc23a2-R1	cttctatcagtgaggacatgatga

NOTE S2. Despite the gene-level focus of the Gene arrays, Affymetrix has assigned a small number of their probesets to alternative isoforms of the same gene. In these cases, the probe signal captures the expression of multiple alternative isoforms and should be split appropriately.

In the human arrays, a small number of probesets interrogate assemblies other than the reference genome. For example, in the human Gene 1.0 ST arrays, there are probesets that interrogate alternative haplotype assemblies of the chromosome 6 MHC region, the chromosome 22 region containing the CYP2D6 gene or the chromosome 5 region containing the SMN1 gene. In these cases, probes target different versions of the same transcript, and hence their signals should not be split. We therefore discard the probesets mapping to alternative haplotypes and focus on probing only the reference genome.

TABLE S1. MMBGX and Gardina workflow results for genes that tested negative for differential splicing by RT-PCR in Gardina *et al.*.

Gene	Probe-set ID	Forward	Reverse	Forward Primer	Reverse Primer	MMBGX result	MMBGX remarks	Gardina result
BTNL3	2844898	Ex3	Ex6	GGAGGTATCCAGTTACTCTG	TCCGTGCTTTCTTCTCCAGT	Incompatible		Positive
CABIN1	3939720	Ex4	Ex6	CTTGATCTGCAGAAACATGAC	ATACCAGAGGTTGACATCTG	Incompatible		Positive
ENOSF1	3795922	Ex2	Ex4	GGTGTGGAATTACCTTCACT	AGGCTTTCCCTCCTGCTTG	Negative		Positive
FAM44B	2887648	Ex1	Ex4	CATCGTGGAGCAGCTCAAG	GCACTCTTATGTAACCGAATC	Positive		Positive
FAT		Ex26	Ex27	GTCTGAAGTGCAGTCTCTGA	GCTCATCAATCACCTCATAG	Negative	Only 1 unfiltered transcript.	Positive
FAT	2797442	Ex6	Ex7	CCTTTACTGTGATGGAAAGTG	GACTTCTGTTCTGCATCAAG			Positive
FN1		Ex32	Ex34	TTAGTGTCTATGCTCAGAATC	CGATATCCAGTGAGCTGAAC	Negative		Positive
FN1		Ex39	Ex41	ATTTATGTCATTGCCCTGAAG	GAAATGATGTACTCAGAAGTG			Positive
GBA	2437271	Ex1	Ex3	CTGCCTTCAGAGTCTTACTG	GAAGTAGAAGCAATCCCTGTGA	Positive		Positive
LRP8		Ex4	Ex7	CAGCCACAAGTGTGTACCT	TCTTGAGGTCAGTGCAGATG	Negative	One case incompatible, one negative.	Positive
LRP8	2413218	Ex14	Ex19	CAGACATGAAGAGGTGCTAC	CATCATCTTCAAGGCTTAATG			Positive
NME2		Ex5	Ex9	TGGAGACTTCTGCATACAAG	GTCTTTCAGGTCAATGTAGTG	Incompatible		Positive
PFKP		Ex15	Ex17a-R	AAATCGGCTGGACAGATGTC	GCTGACAGCTCCAGGAGT		Two cases incompatible, one negative. All transcripts at very low levels in normal tissue, some up in tumour.	Positive
PFKP		Ex16	Ex19	TGCTGATCATCGGTGGATTC	GTCAGGTGCTCCACGTTG	Negative		Positive
PFKP	3232406	Ex18-Ext	Ex19	CTTGCTGCACTCACTCTTG	GTCAGGTGCTCCACGTTG			Positive
PTK9L	2676011	Ex7	Ex9	TCTACAAGCACACCCATGAG	GACGCTGCAAGTGCGTTC	Incompatible		Positive
RAD51	3590094	Ex3	Ex8	GGGAATTAGTGAAGCCAAAG	GCTACATTATCCAGGACATC	Negative		Positive
SRFS2		Ex2	Ex5	GGAGCGGTGTCCTCTTAAG	GCATTACTCAACTGCTACAC	Incompatible		Positive
STK25	2607278	Ex7	Ex9	CTGAAGCACAAGTTCATCAC	CCTCGGCTGCCTCTTGAC	Incompatible		Positive
FIP1L1	2727236	Ex4	Ex7	CTCCATCTGGAATTGAAGATG	ATGCTTCCAGGTGCATCAAG	Incompatible		Positive
PLAU	3252046	Ex3	Ex5	GCAGCAATGAACTTCATCAAG	GCTGAAGAGCATCAGATCTG	Incompatible		Positive
PTK7	2907697, 2907705	Ex6	Ex7b	CCTGGAAGCCACACTTCAC	TCCCAGTAGTGAACTGATTC		All transcripts at very low levels in normal tissue, some up in tumour.	Positive
PTK7		Ex6	Ex11	CCTGGAAGCCACACTTCAC	TGGAGGCAATGCAAGTGTAG	Negative		Positive
PTK7	2907697, 2907705	Ex7	Ex14	AGACAGGATGTCAACATCAC	CAATGGTCTGGATCATCTTG			Negative
PTK7		Ex15	Ex18	ACAAGTGATAAGATGCACTTC	GTTGGACAGGTGCTCCATG			Negative
RAN		Ex1	Ex3	GGCGTTTGAATTGCGCTTC	CAAATTCACCAGTCAAATGAC	Incompatible		Positive
TENS1	3049621	Ex12	Ex17	CCATGCAGCCTGTGTACAC	TCGCGTAAAGGCTGCCATC	Incompatible		Positive
TNS		Ex16	Ex16b	GTGTCTGTGGACTATAACAC	TTCTGGCTTGGCAACTTGAC	Incompatible		Positive
CASP4	3389310	Ex2	Ex5	TATGGCAGGACAAATGCTTC	GGCAGATGGTCAAACTCTGT	Incompatible		Positive
CDH11	3694667	Ex12	Ex14	CAGAGGCCTACATTCTGAAC	TCAAAGGCTTCTGTGTCTTC	Positive	Both transcripts up in tumour, but by slightly different amounts.	Positive
COL11A1	2425849, 2425850	Ex1_1	Ex2	ACAGTACTCTCAGCTTGTTG	AGCAGTATCTGAGCCTTTAG			Positive
COL11A1	2425849, 2425850	Ex1_2	Ex2	GGAGCCGTGGTCCTCTAG	AGCAGTATCTGAGCCTTTAG	Negative	Two cases incompatible, one negative. All transcripts at very low levels in normal tissue, some up in tumour.	Positive
COL11A1		Ex5	Ex7	GCATTATAGTCCAGACTGTGA	CTGTCTGGTAACTTTCCATTG			Positive
CST2	3901398	Ex1	Ex3	GAGGGTGGCATCTATGATG	TCTTGACACCTGGAATTCAC	Incompatible		Positive
KIAA1199	3604205	Ex21	Ex23	GCACTTTCCGAAAGTTTGTG	GCAGGTTACTGGTCTTGTAG	Incompatible		Positive
MUC4	2712246	Ex23	Ex25	ATCCAGCTCTTGCTCAGTG	ACTCCTCCGTGGAACGTG	Incompatible		Positive
NACA	3458111, 3458123	Ex2	Ex7	TGCTACAGAGCAGGAGTTG	TTTACTGACTGGTTCTTCATC	Incompatible		Positive
NME1	3726945	Ex1	Ex3	GTGCAAGTGCTGCGAACCA	GTTTGATCGCAATGAAGGTAC	Negative		Positive
PSD	3304306	Ex14	Ex16	AGAGACGGAGCTCAAGAATG	CTGCTCTTCAGCCTCCTTG	Incompatible		Positive
VEGF		Ex4	Ex8	CACCAAGGCCAGCACATAG	CTGTATCGATCGTTCTGTATC	Negative		Negative
NCAM1		Ex10	Ex14_2	AGTCCATGTACCTTGAAGTG	CTTACGGCGTACGTTGTTTC	Negative	One case incompatible, one negative.	Negative
NCAM1		Ex13	Ex19	GCATTCCAAGTGGTATGATG	TCCTGCTTGATCAGGTTCAC			Negative
SIAHBP1		Ex1	Ex6	TGGCGACGGCGACCATAG	GTAGACGCGGCACATGATG	Negative		Negative
MST1R		Ex10	Ex12	CCAACCTAGTTCCACTGAAG	ACCAAGGAGCGTGCTCTGT	Negative		Negative
FGFR3		Ex6	Ex10	CGTGGAGTTCCACTGCAAG	TGGTGTGTTGGAGCTCATG	Negative		Negative

TABLE S2. COSIE results for genes the 12 positively validated for which gels were provided and the 36 negatively validated genes in Gardina *et al.*. Transcript clusters or probesets which were filtered by the method are indicated with "NA".

Gene	TcId	PsId	P-value	Result (FDR <.2; alpha=0.001025257)	Result (FDR<.3); alpha=0.01245986)
ACTN1	3569814	3569830	5.40E-05	TRUE	TRUE
ATP2B4	2375706	2375764	2.85E-03	FALSE	TRUE
VCL	3252071	3252128	1.97E-03	FALSE	TRUE
CALD1	3025545	3025632	1.67E-04	TRUE	TRUE
SLC3A2	3333711	NA			
COL6A3	2605321	2605386	1.54E-05	TRUE	TRUE
CTTN	3338552	3338589	9.81E-02	FALSE	FALSE
FN1	2598261	2598321	1.37E-02	FALSE	FALSE
TPM1	3597338	3597382	4.25E-03	FALSE	TRUE
CD44	3326635	3326711	5.11E-02	FALSE	FALSE
ITGB4	3735151	3735208	NA		
RAC1	2989050	2989068	NA		
BTNL3	2844888	2844890	2.88E-02	FALSE	FALSE
CABIN1	3939707	NA			
ENOSF1	3795866	3795922	8.05E-03	FALSE	TRUE
FAM44B	2887633	NA			
FAT	2797393	2797410	2.51E-03	FALSE	TRUE
FN1	2598261	2598328	9.28E-03	FALSE	TRUE
GBA	2437205	2437232	1.03E-02	FALSE	TRUE
LRP8	2413203	NA			
NME2	3726960	NA			
PFKP	3232349	3232391	2.34E-03	FALSE	TRUE
PTK9L	2676009	NA			
RAD51	3590086	3590096	2.72E-03	FALSE	TRUE
SFRS2	3771800	NA			
STK25	2607262	NA			
FIP1L1	2727226	2727235	1.09E-02	FALSE	TRUE
PLAU	3252036	3252054	1.13E-03	FALSE	TRUE
PTK7	2907671	NA			
RAN	3438027	3438031	5.66E-04	TRUE	TRUE
TENS1	3049522	3049621	1.41E-03	FALSE	TRUE
TNS	2599153	2599212	9.33E-04	TRUE	TRUE
CASP4	3389273	3389298	1.51E-01	FALSE	FALSE
CDH11	3694657	3694727	5.49E-03	FALSE	TRUE
COL11A1	2425756	2425837	1.13E-03	FALSE	TRUE
CST2	3901387	NA			
KIAA1199	3604147	NA			
MUC4	2425756	2425837	1.13E-03	FALSE	TRUE
NACA	3458097	NA			
NME1	3726934	3726942	1.65E-01	FALSE	FALSE
PSD	3304301	NA			
VEGF	2908179	2908180	4.32E-04	TRUE	TRUE
NCAM1	3349293	3349364	1.10E-02	FALSE	TRUE
SIAHBP1	3157817	NA			
MST1R	2674919	2674958	3.62E-03	FALSE	TRUE
FGFR3	2715016	NA			

TABLE S3. FIRMA results for genes the 12 positively validated for which gels were provided and the 36 negatively validated genes in Gardina *et al.*. Transcript clusters or probesets which were filtered by the method are indicated with "NA".

Gene	TcId	PsId	P-value	Result (FDR <.2; alpha=4.4774e-05)	Result (FDR<.3); alpha= 0.003178101)
ACTN1	3569814	3569830	1.58E-04	FALSE	TRUE
ATP2B4	2375706	2375766	2.55E-03	FALSE	TRUE
VCL	3252071	3252128	6.43E-06	TRUE	TRUE
CALD1	3025545	3025632	1.44E-04	FALSE	TRUE
SLC3A2	3333711	3333717	8.88E-01	FALSE	FALSE
COL6A3	2605321	2605386	6.28E-03	FALSE	FALSE
CTTN	3338552	3338589	3.83E-03	FALSE	FALSE
FN1	2598261	2598321	7.47E-03	FALSE	FALSE
TPM1	3597338	3597382	1.20E-02	FALSE	FALSE
CD44	3326635	3326714	5.52E-02	FALSE	FALSE
ITGB4	3735151	3735208	NA		
RAC1	2989050	2989068	3.07E-02	FALSE	FALSE
BTNL3	2844888	2844898	2.87E-02	FALSE	FALSE
CABIN1	3939707	3939739	6.02E-03	FALSE	FALSE
ENOSF1	3795866	3795922	3.45E-02	FALSE	FALSE
FAM44B	2887633	2887648	1.23E-03	FALSE	TRUE
FAT	2797393	2797449	4.15E-02	FALSE	FALSE
FN1	2598261	2598330	8.82E-04	FALSE	TRUE
GBA	2437205	2437223	1.04E-01	FALSE	FALSE
LRP8	2413203	2413218	3.04E-04	FALSE	TRUE
NME2	3726960	3726971	3.68E-02	FALSE	FALSE
PFKP	3232349	3232365	5.78E-03	FALSE	FALSE
PTK9L	2676009	2676018	3.10E-03	FALSE	TRUE
RAD51	3590086	3590096	4.51E-03	FALSE	FALSE
SFRS2	3771800	3771810	6.05E-03	FALSE	FALSE
STK25	2607262	2607278	3.49E-03	FALSE	FALSE
FIP1L1	2727226	2727235	2.70E-03	FALSE	TRUE
PLAU	3252036	3252051	2.16E-04	FALSE	TRUE
PTK7	2907671	2907707	1.94E-04	FALSE	TRUE
RAN	3438027	3438033	9.65E-04	FALSE	TRUE
TENS1	3049522	3049643	1.87E-03	FALSE	TRUE
TNS	2599153	2599225	6.07E-03	FALSE	FALSE
CASP4	3389273	3389317	2.93E-02	FALSE	FALSE
CDH11	3694657	3694703	3.98E-04	FALSE	TRUE
COL11A1	2425756	2425850	6.15E-06	TRUE	TRUE
CST2	3901387	3901398	2.19E-04	FALSE	TRUE
KIAA1199	3604147	3604213	1.09E-04	FALSE	TRUE
MUC4	2425756	2425850	6.15E-06	TRUE	TRUE
NACA	3458097	3458098	1.93E-02	FALSE	FALSE
NME1	3726934	3726957	7.38E-03	FALSE	FALSE
PSD	3304301	3304306	1.78E-03	FALSE	TRUE
VEGF	2908179	2908192	1.29E-03	FALSE	TRUE
NCAM1	3349293	3349296	3.76E-04	FALSE	TRUE
SIAHBP1	3157817	3157827	5.22E-03	FALSE	FALSE
MST1R	2674919	2674946	4.32E-02	FALSE	FALSE
FGFR3	2715016	2715045	4.72E-02	FALSE	FALSE