Appendix

Materials and Methods

For more details on data collection, curation, and assessment of significance see the appendix of the original meta-analysis publication [12].

Gene Mapping
Where possible, the platform-specific annotation files for each global gene expression platform were obtained and the sequence features were mapped to an NCBI Entrez Gene ID. Gene mapping was performed as previously described [12], with the exception that the Clone/Gene ID Converter tool [37], instead of the DAVID Resource [A1], was used to map sequence features to Entrez Gene IDs. Also, for the sequence features that could not be mapped by the Clone/Gene ID Converter tool, a Bioperl script [A2] was implemented to fetch the corresponding nucleotide sequence from NCBI Genbank [A3] and aligned with BLASTN (maximum E-value of 10^{-5}, minimum sequence similarity of 90%) to a database made up of all human transcript sequences from NCBI RefSeq [A4]. Sequences that hit multiple genes were omitted from further analysis.

Total Gene Lists
Of the 25 studies included in this meta-analysis, 22 compared cancer versus normal tissue with a microarray platform, while one utilized Serial Analysis of Gene Expression (SAGE). In the 22 studies, we were able to obtain annotation files for 13 of them. From the annotation files of these 13 studies, we were able to map 88.6% of the features to an Entrez Gene ID. Therefore, for each of the nine microarray cancer versus normal studies in which the annotation files could not be obtained, this same fraction of genes was randomly chosen from the superset of the gene lists of the thirteen studies. Finally, to obtain a total gene list for the SAGE study, all gene names in the tag to gene mapping data from SAGE Genie [38] were mapped to Entrez Gene IDs. Table A1 summarizes the mapping success rate for the three comparisons that we performed.

Gene Ontology Analysis
We performed Gene Ontology [A5] analysis on the 573 multi-study genes in the cancer versus normal comparison with GOStat [A6], which utilizes Fisher's Exact test. To correct for false discovery, the Benjamini & Hochberg correction was applied. Significance was set at $P < .001$.

Results

Consistently Reported Differentially Expressed Genes
In the cancer versus normal comparison, there were 573 multi-study genes. Tables 4 and 5 contain differentially expressed genes reported in at least five studies, while Tables A2 and A3 contain the consistently reported up- and down-regulated genes, respectively, in three or four cancer versus normal studies.

In the adenoma versus normal comparison, there were 39 multi-study genes (Tables A4 and A5). In the simulations, an average of 10.64 (95% CI, 10.61 to 10.68) genes was observed with an overlap of two, while the actual data contained 37. For an overlap of three, an average of 0.07 (95% CI, 0.067 to 0.073) of a gene was observed in the simulations, while two genes were observed with an overlap of three in the real data.

Finally, in the cancer versus adenoma comparison, significance in overlap was not observed. The studies included in this comparison are listed in Table A6.

Significantly Over-represented Gene Ontology Terms
A gene ontology analysis of multi-study genes from the cancer versus normal comparison identified 24 significantly over-represented terms (Table A7). Multi-study genes tended to encode products that remained in the cytoplasm, specifically ribosomal complexes, or were exported into the extracellular space. Significant biological processes included responses to stress and cell cycle control. Significant molecular functions involved binding (RNA and unfolded proteins) and translation.

References

A1. Dennis G Jr, Sherman BT, Hosack DA, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 2003;4:P3.

A2. Stajich JE, Block D, Boulez K, et al. The Bioperl toolkit: Perl modules for the life sciences. Genome Res 2002;12:1611-1618.

A3. Wheeler DL, Barrett T, Benson DA, et al. Database resources for the National Center for Biotechnology Information. Nucleic Acids Res 2007;35:D5-D12.

A4. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2007;35:D61-D65.

A5. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25-29.

A6. Beissbarth T, Speed TP. GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 2004;20:1464-1465.