Supplementary Methods

Uniqueome Generation

Genome sequences were downloaded from http://genome.ucsc.edu/. A reference genome was constructed from regular chromosome files. For mammalian genomes, “random” and haplotype variant files were not used; mitochondria was used. For the remaining genomes, all available sequences were used. We systematically evaluated the uniqueness of genome derived N-mers defined under different combinations of length and stringency criteria. To do this, an all-against-all alignment was performed independently for each chosen word size and criteria. Alignments were performed with the Imagenix Sequence Alignment System (ISAS; http://www.imagenix.com/).

Unlike aligners that use the Burrows Wheeler Transformation, such as BWA, Bowtie, and SOAP2 (Langmead*, et al.*, 2009; Li*, et al.*, 2009; Li*, et al.*, 2009), ISAS does not compress the data to be searched. The entire reference, along with a full index based on a non-injective hashing function are stored in an uncompressed form in memory. While this makes the memory requirement of the software several times larger (eg. almost 16 GBytes for a 3 GBase human reference), the speed of the alignment makes “all against all” for the large mammalian genomes feasible. For any given K-mer in the reference, the index is used to accelerate an exhaustive comparison with all other K-mers in the reference. In un-gapped mode, a comparison simply counts the number of corresponding base positions which are not identical. In gapped mode, each comparison computes the Levenshtein distance (Levenshtein, 1966). Because the gapped mode comparison is computationally more expensive, a second hashing function is used, which adds a further acceleration.

Alignments were run in the non-heuristic mode (filter=0). Turning on filtering speeds up the alignments, but could miss a small proportion of tags. All base-space alignments allow the mismatches to be placed anywhere within the tag. For color-space, the 25mer (mode=2) and 50mer (mode=5) alignments allow the errors to be placed anywhere within the tag. For all other lengths, the global setting was used, meaning that the mismatches are limited to 2 within the first 25nt. For example, using the global=60,6 command allows matching to occur at 60mers with up to 2 mismatches in the first 25nt, and up to 6 mismatches in total.

Generating Uniqueome BED files

To display results, ISAS matching results were converted into BED files for viewing with the Galaxy (Giardine*, et al.*, 2005), the UCSC genome browser (Kuhn*, et al.*, 2009), or IGV (http://www.broadinstitute.org/igv). BED files are strand specific, and record a value of 1 where a sequence starting at a given genomic coordinate on that strand is unique within the genome. For example, in the file named hg19_uniqueome.unique_starts.color-space.25.2.positive.BED.gz, the second line "chr10<tab>64762,tab.64767<tab>1<newline>" indicates that on chr10, tags of length 25 on the positive strand that start between 64762 and 64766 are all unique in the genome. All coordinates are 0-based (the first nucleotide in the chromosome is a 0), and the BED files are the standard half-open format (left closed, right open). Negative strand BED files are created by offsetting the positive strand files by the tag length. Filenames are in the format [genome]_uniqueome.unique_starts.[space].[length].[mismatches].[strand].BED.gz. To facilitate the loading of large files into genome browsers, bigBed files are also available (*.bb). All files can be downloaded from http://grimmond.imb.uq.edu.au/uniqueome/ or http://www.imagenix.com/uniqueome/.

Generating Uniqueome Coverage Plots

Whilst the BED files provide the starting positions of unique sequence in the genome, sequencing data is often displayed as coverage plots (such as wiggle files; (Cloonan*, et al.*, 2008; Mortazavi*, et al.*, 2008)). It is useful then to understand how the uniqueome contributes to the peaks and troughs often seen with massively parallel sequencing data. Whilst a tag starting at a given coordinate may not be uniquely mappable, that coordinate may still be covered by unique tags starting at neighboring coordinates. Coverage values are expressed as rounded integer percentiles of full coverage. eg. A value of 100 indicates that 100% of overlapping N-mers are unique and contribute to coverage of that coordinate. Similarly a value of 50 indicates that 50% of overlapping N-mers are unique. To generate these coverage plots, the BED files generated above were converted to coverage "wiggle" plots, and then every position was divided by the word size, multiplied by 100, and rounded to the nearest integer. These coverage plots are formatted as BED files as above. Filenames are in the format [genome]_uniqueome.coverage.[space].[length].[mismatches].Wig.gz. To facilitate the loading of large files into genome browsers, bigWig files are also available (*.bw).

Comparison of multiple short read aligners

To compare problematic alignment regions in different short-read aligners, we used mouse embryonic stem cell sequencing data (SRA accession number SRX019275) previously reported (Guttman, *et al.*, 2010). All reads were trimmed to 50nt, and aligned to the mm9 reference (excluding "random" assemblies) using BWA (Li, *et al.*, 2009), Bowtie (Langmead, *et al.*, 2009), and SOAP2 (Li, *et al.*, 2009) in nucleotide-space. Tags were also converted to color-space and aligned with mapreads (http://www.solidsoftwaretools.com) and SHRiMP (Rumble, *et al.*, 2009). In all cases, only tags that were placed uniquely in the genome allowing 2 mismatches in base-space, and 5 mismatches in color-space were used for visualization in the UCSC genome browser.

Correlation of RNAseq data with microarray data

In order to compare the impact of uniqueome normalization versus multi-mapping tag rescue or straight RPKM values, we used mouse embryoid body (EB) sequencing data (SRA accession number SRA000306) and EB microarray data (GEO accession number GSE10518) previously reported in Cloonan *et al* (2008). Sequencing data was mapped using mapreads v2.4.1 as a part of RNA-MATEv1.1 (Cloonan, *et al.*, 2009), mapping only 35mers and allowing up to 3 color-space mismatches. No tag positions were masked, and valid-adjacent errors were counted as 2 mismatches. Multi-mapping rescue was set to false. Each replicate was matched and processed independently. Genomic matching tags were assigned to RefSeq genes using Galaxy (Giardine, *et al.*, 2005), according to the strategy outlined in the user manual for RNA-MATEv1.1 (http://grimmond.imb.uq.edu.au/RNA-MATE/). RPKMs were calculated based on either the full length of each RefSeq gene (Raw tag counts), or the unique length of each RefSeq gene (Uniqueome normalized tag counts) determined using the strategy outlined in Supplementary File 1. To generate "non-unique tag rescue counts", RNA-MATEv1.1 was run as above, but multi-mapping rescue was set to true. RPKMs were calculated as described above. RPKMs for all sequencing replicates were then averaged.

Illumina Sentrix Bead Array (single color) data was quantile normalized using the Limma package (Smyth, *et al.*, 2003) in R. Expression was considered to be gene-centric (rather than probe-centric), using RefSeq probe annotations. For correlation, we used only RefSeq genes where the microarray detection score = 1, and where the sequencing RPKM (of either rescued or uniqueome corrected data) was ≥ 10 (n = 3580). Pearson correlations (r) were calculated on Log_2 normalized data. No effort was made to correct for alternative splice variants, or the genomic location of the microarray probes.

References

Cloonan, N., *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing, *Nat Meth*, **5**, 613-619.

Cloonan, N., *et al.* (2009) RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data, *Bioinformatics*, **25**, 2615-2616.

Giardine, B., *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis, *Genome Res*, **15**, 1451-1455.

Guttman, M., *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs, *Nat Biotechnol*, **28**, 503-510.

Kuhn, R.M., *et al.* (2009) The UCSC Genome Browser Database: update 2009, *Nucleic Acids Res*, **37**, D755-761.

Langmead, B., *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol*, **10**, R25.

Levenshtein, V.I. (1966) Binary Codes Capable of Correcting Deletions, Insertions, and Reversals, *Soviet Physics Doklady* **10**, 707-710.

Li, H., *et al.* (2009) Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, **25**, 1754-1760.

Li, R., *et al.* (2009) SOAP2: an improved ultrafast tool for short read alignment, *Bioinformatics*, **25**, 1966-1967.

Mortazavi, A., *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nat Meth*, **5**, 621-628.

Rumble, S.M., *et al.* (2009) SHRiMP: accurate mapping of short color-space reads, *PLoS Comput Biol*, **5**, e1000386.

Smyth, G.K., *et al.* (2003) Normalization of cDNA microarray data, *Methods*, **31**, 265-273.