Short-read sequencing reveals dynamic range of transcript abundance in CHO cells

Faraaz N. K. Yusufi[1], Nitya M. Jacob[3], Bhanu Chandra Mulukutla[3], Anne Kantardjieff[3], Kathryn Johnson[3], Bernard Liat Wen Loo[1], Song Hui Chuah[1], Peter Morin Nissom[1], Miranda Yap[1,2], Dong-Yup Lee[1,2], Wei-Shou Hu[3*]

[1] Bioprocessing Technology Institute, A*STAR, 20 Biopolis Way, #06-01 Centros, Singapore 138668
[2] Department of Chemical & Biomolecular Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119620
[3] Department of Chemical Engineering and Materials Science, University of Minnesota, 421 Washington Ave SE, Minneapolis, MN, USA 55414-01232
[*] Contact: acre@cems.umn.edu, Phone: +1-612-626-7630, Fax: +1-612-626-7246

Abstract

Next generation sequencing technologies such as the Solexa/Illumina platform are generating upwards of 3Gbp of short-read sequence data. Such deep-sequencing technology can be utilized to analyze the abundance of transcripts in a cell, as well as to discover previously unknown rare transcripts. A single Solexa sequencing run on a cDNA library produced from a CHO DG44-derived IgG-producing cell line resulted in 55.8 million 46-base-pair reads. This set of sequences was compared with a collection of 80,000 CHO ESTs obtained previously through traditional Sanger sequencing.

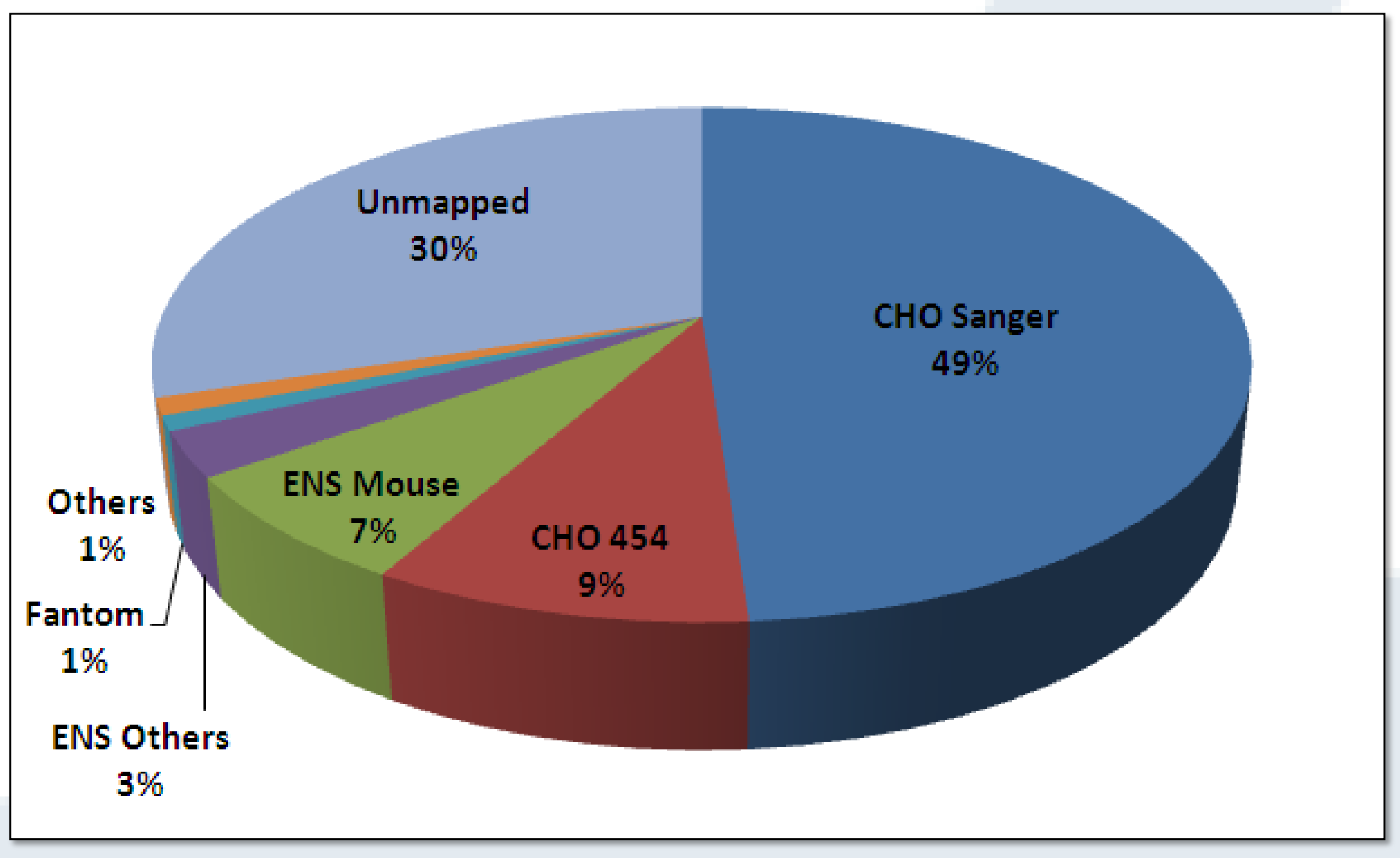
The RMAP algorithm was used to calculate the abundance of each EST in our pre-existing collection by counting the number of Solexa reads that were mapped to each transcript. This mapping provided insight into the wide dynamic range of transcript activity, spanning more than six orders of magnitude. This analysis allowed us to map approximately 70% of Solexa reads to the reference set of CHO ESTs, with the remaining 30% unmapped sequences possibly representing novel CHO transcripts. More than 99% of our EST collection was hit by at least one Solexa read, reinforcing the observation that Solexa provides significant sequencing depth.

The abundance information generated was compared to CHO Affymetrix array results obtained from the same CHO cell line used for sequencing. A Spearman correlation coefficient of 0.78 between the microarray and sequencing data was observed, illustrating good correlation between the two technologies.

Solexa sequencing data was also analyzed in a genomic context in an attempt to identify transcriptional hotspots. By using the Ensembl mouse annotation for every previously known CHO gene, the sequence data was visualized in the context of position along mouse chromosomes. By identifying regions where the unique gene count is low, yet the number of mapped Solexa reads is high, we could locate putative transcriptional hotspots.

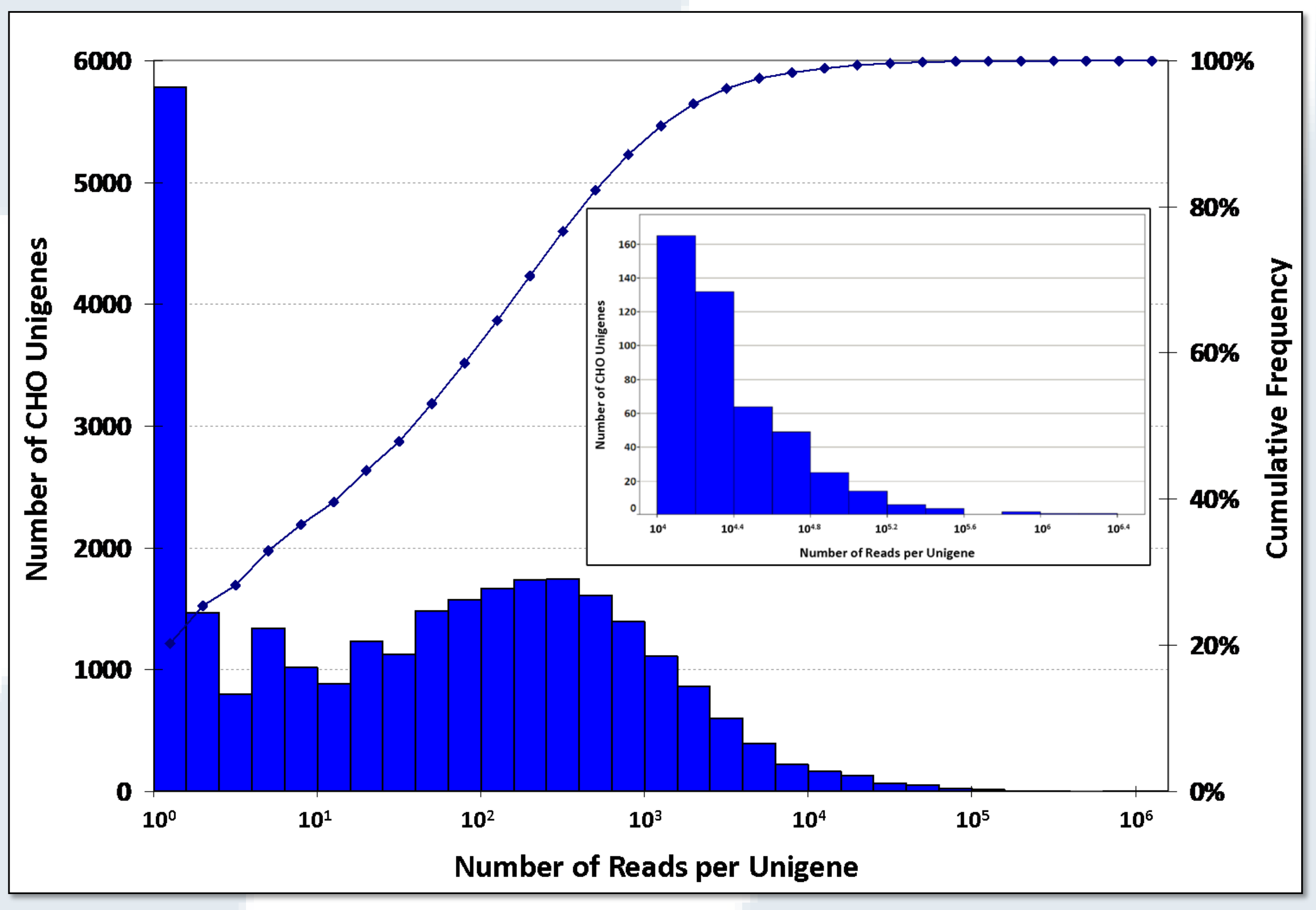
Mapping of Solexa reads

High quality reads were aligned to reference sequences using the RMAP algorithm, allowing up to 2 mismatches in the search. The reference sequences included our current collection of CHO ESTs obtained from Sanger and 454 sequencing, Ensemble v48 cDNA collections for mouse, human, rat and squirrel, the FANTOM3 database of mouse cDNA's, the Repbase Update database of eukaryotic repetitive sequences and the NONCODE collection of noncoding RNA's. Of the 55.8 million Solexa reads approximately 70% were mapped to the reference set with the remaining 30% unmapped as shown below.



Summary of Solexa mapping to reference sequence sets

To characterize the transcriptome dynamics in CHO, we examined the mapping frequency of Solexa reads to our set of 28,914 CHO unigenes. More than 99% of the unigenes were mapped to by at least one Solexa read, with the remaining fraction of unigenes sequenced from either tissue-derived cDNA or from chemically-treated cDNA. The following figure shows the frequency distribution of mapped Solexa reads to CHO unigenes, with the bins plotted in logarithm scale. The inset shows an expanded view of unigenes with a mapping frequency greater than 10^4. The cumulative frequency is shown on the right y-axis.



Frequency distribution of mapped Solexa reads to CHO unigenes

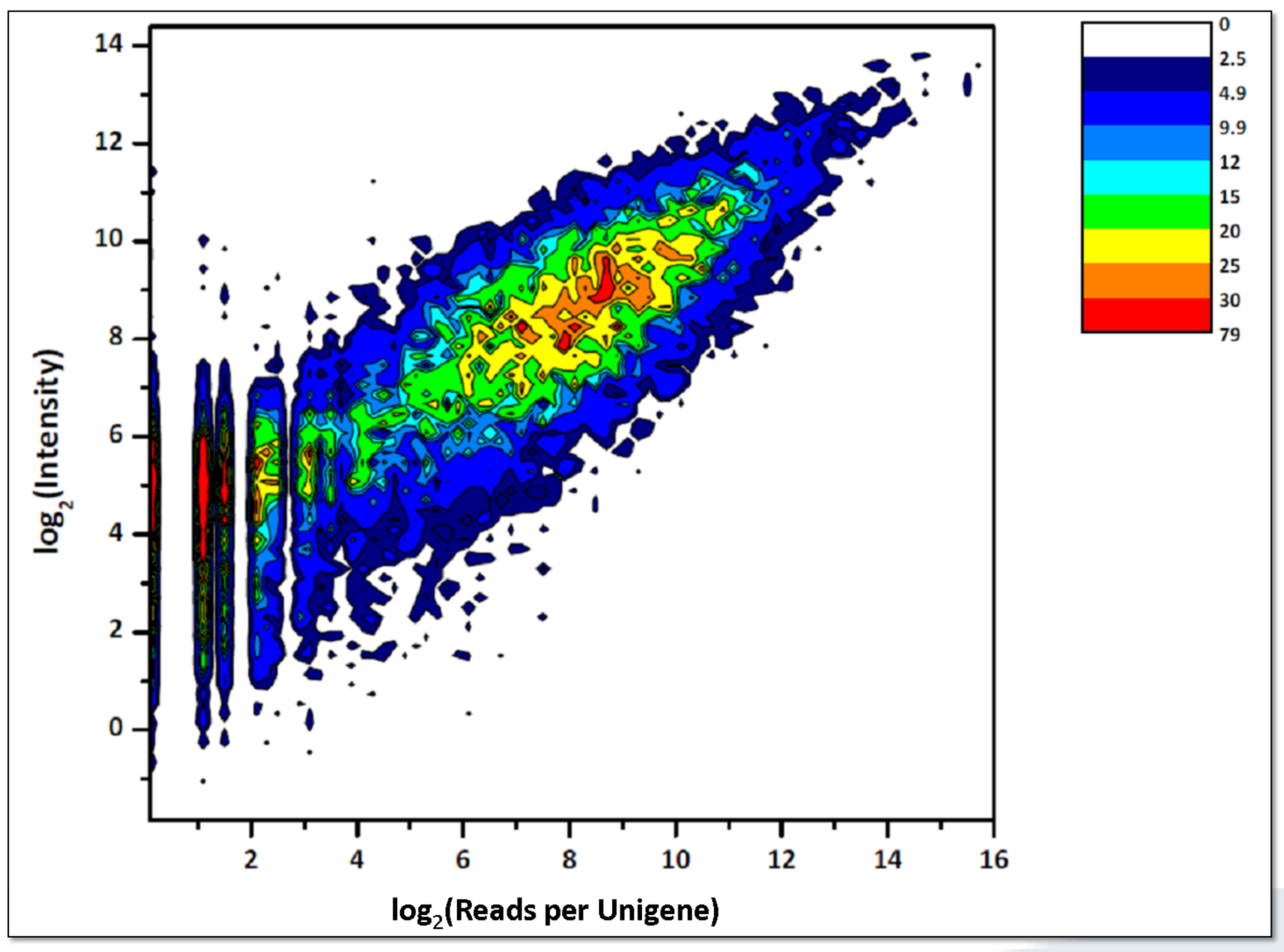
The number of reads mapped to a unigene is indicative of its transcript abundance level. For our repertoire this number was found to vary over a wide dynamic range, from one to more than 1.6 million hits. The most abundant transcripts are listed in the table below.

Most abundant CHO unigenes identified through mapping of Solexa reads

Gene Description	Solexa Hits	EST Length (bp)
Recombinant immunoglobulin heavy chain C gene segment	1,644,460	1,964
Recombinant immunoglobulin Kappa light chain C gene segment	753,593	962
Glyceraldehyde-3-phosphate dehydrogenase	670,311	1,301
cDNA sequence BC005624	297,190	1,290
Heat shock protein 5	247,121	2,529
Eukaryotic translation elongation factor 1 alpha 1	226,264	1,834
Ferritin heavy chain 1	184,459	934
Cricetulus griseus Csr1	181,157	1,148
Pyruvate kinase, muscle	162,868	865
Actin, beta, cytoplasmic	151,084	1,026
Enolase 1, alpha non-neuron	132,826	1,762
Calreticulin	132,136	1,917

Comparison with microarray data

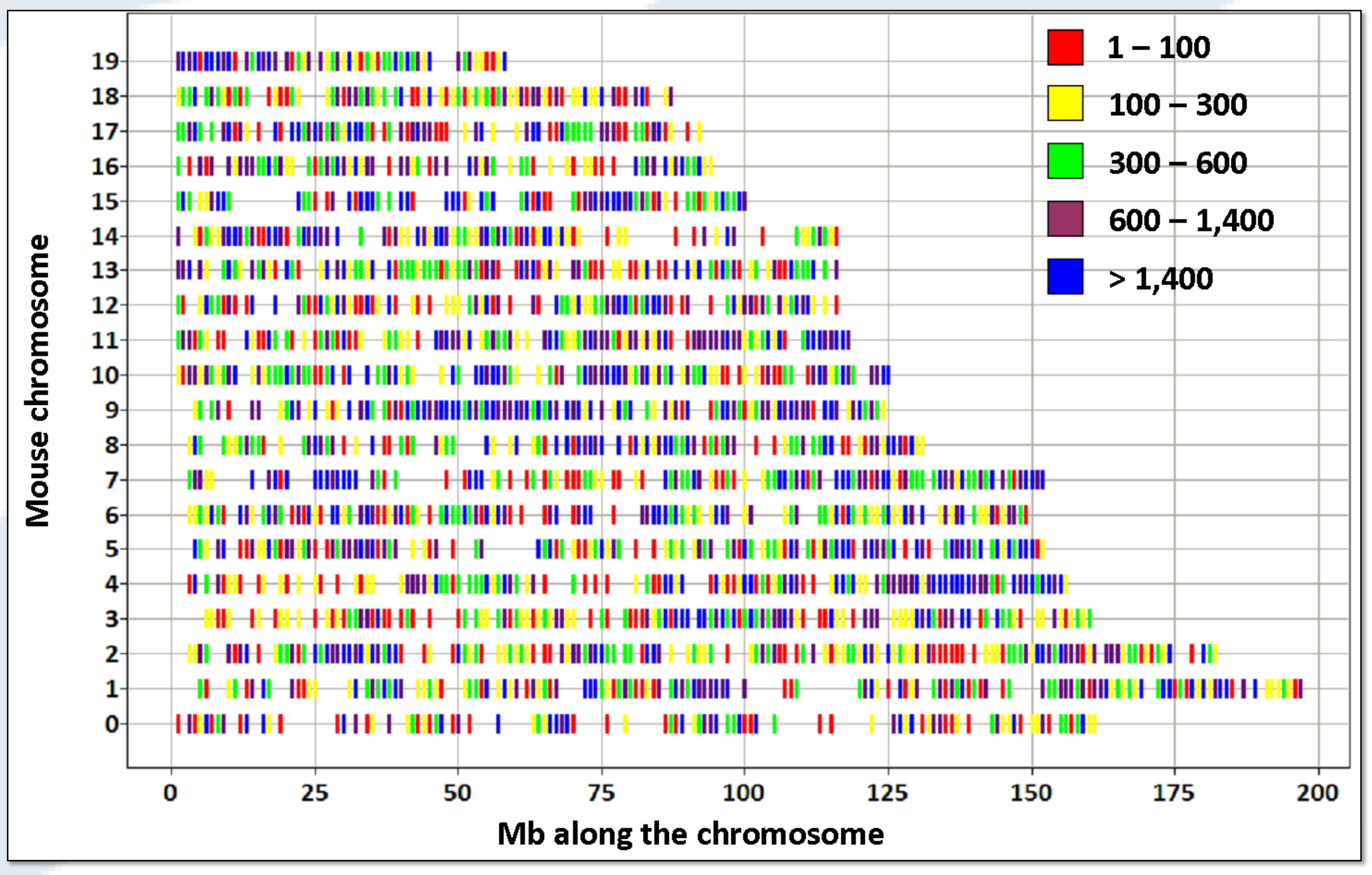
To assess the performance of Solexa sequencing for transcriptome analysis, the same RNA sample was used for sequencing and hybridization onto CHO Affymetrix arrays. Overall, a good correlation was seen between the abundance levels assessed by the two methods, with a Spearman correlation coefficient of 0.78. The number of reads mapped to each unigene is plotted against the intensity of the corresponding probe sets on the Affymetrix CHO microarray in the figure below. The contour plot shows the log2-transformed count of Solexa reads mapped to CHO unigenes versus the log2-transformed Affymetrix intensity of corresponding CHO unigenes.



Correlation of reads mapped per CHO unigene and Affymetrix signal intensities

Identifying transcriptional hotspots

The abundance of Solexa reads which mapped to mouse orthologs was examined in a genomic context in order to identify highly transcribed regions. Read counts were summed over 1 Mb bins along mouse chromosomes and subsequently divided by the number of unique genes within each bin. These normalized abundances are plotted along mouse genomic coordinates in the figure below. Color thresholds were set based on the distribution of normalized mapped read frequencies to mouse orthologs. Each color threshold captures 20% of the normalized abundance distribution. Clusters of highly expressed transcripts as well as low abundance regions on each chromosome can be seen.



Normalized abundance of Solexa reads mapped to orthologous mouse genes

Summary and Conclusions

While high throughput sequencing technologies have only recently become widely available, they have already altered the landscape of transcriptome characterization. We have employed the Illumina Solexa GAII platform to characterize the transcriptome of an antibody-producing CHO)cell line. More than 55 million sequence reads were generated and mapped to an existing set of CHO unigenes derived from expressed sequence tags (ESTs), as well as several public sequence databases.

The extensive depth of sequencing granted insight into the wide dynamic range of different transcripts, spanning more than six orders of magnitude. A large fraction of sequences unearthed have not been previously reported or seen in our existing sequence data set. With the depth that ultra high-throughput sequencing methods can reach, one can expect that the entire transcriptome of this industrially important organism will be decoded in the near future.

References

Mortazavi, A., et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods 5(7): 621-8.

Smith, A. D., et al. (2008). "Using quality scores and longer reads improves accuracy of Solexa read mapping." BMC Bioinformatics 9: 128.

Kantardjieff, A., et al. (2009). "Developing genomic platforms for Chinese hamster ovary cells." Biotechnology Advances.

21st Meeting of the European Society for Animal Cell Technology, ESACT, June 2009, Dublin, Ireland