Supplement A:
Estimation of Number of Unique Transcripts from Limited Sage Data

Anisimov SV, Tarasov KV, Tweedie D, Stern MD, Wobus AM, Boheler KR: SAGE Identification of Gene Transcripts with Profiles Unique to Pluripotent Mouse R1 Embryonic Stem Cells

We start from the premise that there are N total unique transcripts, each of which is expressed with an expression level p_i normalized so that

$$\sum_{i=1}^{N} p_i = 1$$

i.e. p_i is the probability that an arbitrary sampled transcript (tag) will be i. If m tags have been sampled, the chance that transcript i has *not* been "hit" yet is $(1-p_i)^m$, so the chance that it *has* been detected is $1-(1-p_i)^m$. The expected number of unique transcripts that have been detected is then:

$$U(m) = \sum_{i=1}^{N}[1-(1-p_i)^m] = N-\sum_{i=1}^{N}(1-p_i)^m \approx N-\sum_{i=1}^{N} e^{-mp_i}$$

where, in the third expression, we have approximated the m^{th} power of a number close to one by an exponential function, which is quite accurate for the large m we deal with here.

If we define the density of expression levels by the function $\rho(p)$, then the sum can be expressed as:

$$U(m) = N-\int \rho(p) e^{-mp} dp$$

The true density function ρ is "hairy": a series of unit delta-functions located at the values p_i, but it can be approximated by a histogram H_i. Then we get:

$$U(m) = N-\sum_{j=1}^{L} H_j e^{-mq_j} = \sum_{j=1}^{L} H_j(1-e^{-mq_j})$$

where H_j is the true number of unique transcripts with true expression level in the j^{th} histogram

bucket whose expression level is q_j. Of course, we don't know H, what we know is the *measured* histogram h_j describing how many unique transcripts have been found after examining $m_0 = 137832$ total tags with *estimated* transcription level j/m_0 where j is the number of copies of the tag found. At the low-expression end, a fraction of unique transcripts will be missing. But we know, as described above, exactly what that fraction is, so we can correct for it to estimate H, thereby obtaining:

$$U(m) = \sum_{j=1}^{L} h_j \frac{1 - e^{-mq_j}}{1 - e^{-m_0 q_j}}$$

This function plateaus at the estimated value N^* given by

$$N^* = \sum_{j=1}^{L} \frac{h_j}{1 - e^{-m_0 q_j}}$$

which gives $N^* = 63722$.

The function $U(m)$ is plotted as the solid line in figure 1. As may be seen, it is not a good fit – in particular, the slope at the rightmost point is substantially less than that of the data, giving little confidence that this function can be used to extrapolate to the plateau, N^*. Part of the problem is that we have neglected sequencing error. In order to account for sequencing error, we assume that a fraction r of sampled tags contain sequencing errors that cause them to be counted as a spurious sequence chosen at random from among the $N_{max} = 4^{10}$ possible sequences. Repeating the above analysis, and assuming that $N_{max} >> N$ and that the number of "hits" on a particular sequence by spurious tags is Poisson-distributed, we obtain the following expressions:

$$u(m,r) = \sum_{j=1}^{L} \frac{\left(1 - e^{-m\left(\frac{r}{N_{\max}} + \frac{k_j(1-r)}{m_0}\right)}\right)\left(h_j e^{\frac{m_0 r}{N_{\max}}} - \frac{\left(\frac{m_0 r}{N_{\max}}\right)^{k_j} N_{\max}}{k_j!}\right)}{e^{\frac{m_0 r}{N_{\max}}} - e^{k_j(r-1)}} + \left(1 - e^{-\frac{mr}{N_{\max}}}\right) N_{\max}$$

$$N^*(r) = \sum_{j=1}^{l} \frac{h_j e^{\frac{m_0 r}{N_{\max}}} - \frac{\left(\frac{m_0 r}{N_{\max}}\right)^{k_j} N_{\max}}{k_j!}}{e^{\frac{m_0 r}{N_{\max}}} - e^{k_j(r-1)}}$$

where k_j is the number of tag copies corresponding to the j^{th} bin of the empirical histogram, *i.e.* $q_j=k_j/m_0$.

$N^*(r)$, the estimated number of unique tags in the presence of sequencing error, is plotted in figure 2. The function $U(m,r)$ is plotted as the dashed lines in figure 1 for r=0.049, the error rate estimated from the measured fidelity of our sequencing, and r=0.161, the error rate that would give the best fit to the data. It is apparent that the model could only fit the data if the rate of sequencing errors were more than triple what we believe it to be.

The likely cause of this discrepancy is that, in the above analysis, we have assumed that the number of copies of a tag detected is a valid estimate of the expression level of its transcript. Because of sampling error, this will be inaccurate in the low-expression bins. Most seriously, the one-copy bin may contain a (possibly small) fraction of the tags corresponding to transcripts whose expression level is substantially less than $1/m_0$. This problem cannot be corrected in any purely mathematical way. Since the observed number of unique transcripts is still rising rapidly at the one-copy level, we have no estimate of the transcription rate of the least-expressed gene. We are at liberty to assume the existence of arbitrarily large numbers of arbitrarily low-expression transcripts, most of which are undetected after sampling 137832 tags. The seriousness of this problem can be illustrated, in a contrived way, by assuming that all the transcripts in the one-copy bin actually have a true transcription level k_1 less than the nominal value of 1. Figure 3 plots the number of unique transcripts estimated with this assumption, as a function of k_1. If the average expression level of all the transcripts in the one-copy bin were actually only 0.1 (fractional expression $1/(10\ m_0) = 7.25\times10^{-7}$, then N^* would rise to the implausibly high value 268,000. The number of unique transcripts estimated above must therefore be regarded as lower limits; a more accurate estimate cannot be performed until enough tags have been sequenced to identify the rising limb of the histogram and estimate the minimum transcription level of expressed genes.

Figure 1 of Supplement A

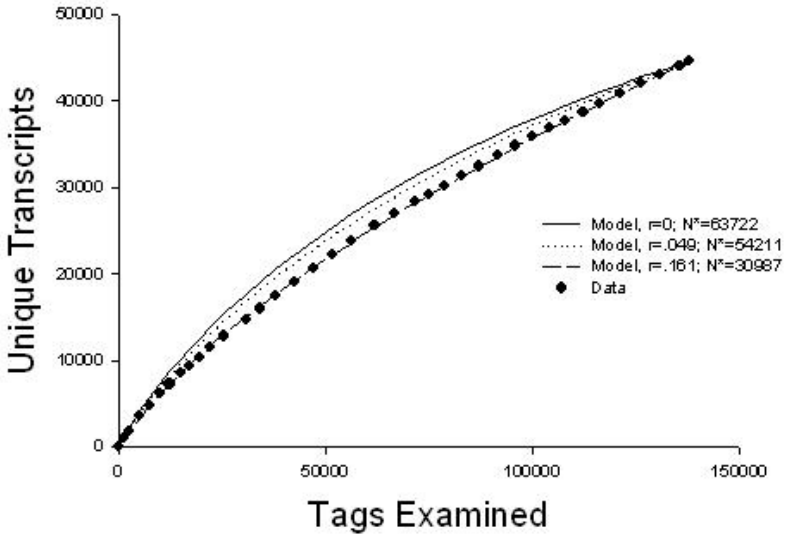

Figure 2 to Supplement A

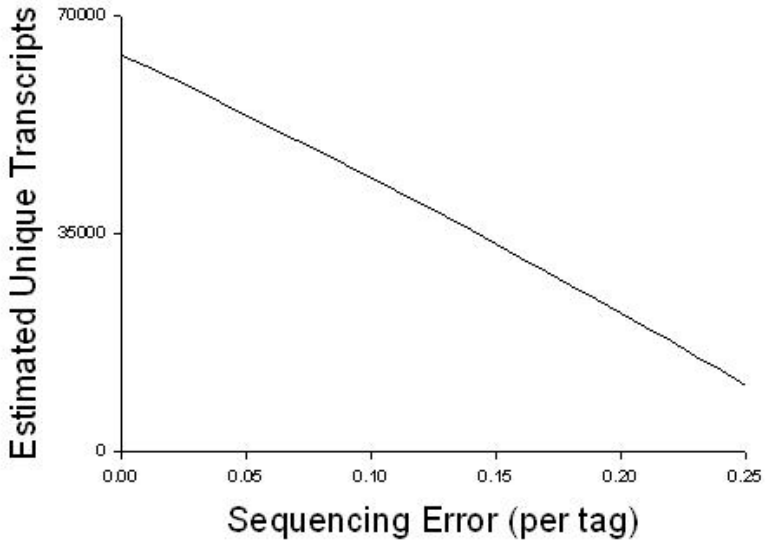

Figure 3 to Supplement A

Supplement B:
100 Most Abundant Tags in the R1 ES cell SAGE library

Anisimov SV, Tarasov KV, Tweedie D, Stern MD, Wobus AM, Boheler KR: SAGE Identification of Gene Transcripts with Profiles Unique to Pluripotent Mouse R1 Embryonic Stem Cells

[a] Tags sorted according to abundance within individual functional categories.
[b] 3'-RACE data generated in this study.
Similarity search was performed with GenBank Non-Redundant database (release 120.0) and EST database (release 120.0).

SUPPLEMENT B. 100 Most Abundant Tags in R1 ES Cells

N	Tag Sequence [a]	Count (N)	Abundance (%)	GenBank Matches	Accession number
				Protein Synthesis	
1.	TGACCCCGGG	874	0.6341	Ubiquitin/60S ribosomal fusion protein	AF118402
2.	GGATTTGGCT	648	0.4701	EST, similar to acidic ribosomal phosphoprotein P1	W99992
3.	TAAAGAGGCC	625	0.4534	Ribosomal protein S26 (Rps26)	U67770
4.	AGGAAGGCGG	583	0.4229	EST, similar to ribosomal protein L36	W87139
5.	GATGACACCA	555	0.4026	Ribosomal protein S28	U11248
6.	GGGAAGGCGG	537	0.3896	EST, similar to Human 40S ribosomal protein S3A	W36575
7.	CAAGGTGACA	458	0.3322	Ribosomal protein S2 (Llrep3)	U89418
8.	GCAGAGTGCG	429	0.3112	Ribosomal protein S6	Y00348
9.	GGGTGGCCCA	408	0.296	EST, similar to Rat 40S ribosomal protein S13	BF227078
10.	GGCTTTGGTC	395	0.2865	Acidic ribosomal phosphoprotein P1	U29402
11.	CACCACCGTT	391	0.2836	Ribosomal protein L7a (Surf3)	M21460
12.	AGAGCGAAGT	380	0.2756	Ribosomal protein L41	U93862
13.	GATTCCGTGA	359	0.2604	EST, similar to Human 60S ribosomal protein L37 (Rpl37)	W99095
14.	AAAACAGTGG	344	0.2495	Ribosomal protein L37a	X73331
15.	CGCCGCCGGC	321	0.2328	EST, similar to Rat 60S ribosomal protein L35 (Rpl35)	W82556
16.	CTAGTCTTTG	311	0.2256	S29 ribosomal protein	L31609
17.	ATACTGAAGC	302	0.2191	60S ribosomal protein	U28917
18.	TGGTGACAAA	298	0.2162	EST, similar to Rat ribosomal protein L4	BB223769
19.	GTGGGCGTGT	279	0.2024	Ribosomal protein S15 (Insulinoma, rig)	M33330
20.	GTTGCTGAGA	264	0.1915	Ribosomal protein L10 (QM)	X75312
21.	CCCGTGTGCT	262	0.19	EST, similar to Rat 40S ribosomal protein S9	W91659
22.	AATCCTGTGG	241	0.1748	Ribosomal protein L8 (Rpl8)	U67771
23.	TCTGTGCACC	239	0.1733	Ribosomal protein S11	U93864
24.	TGGCCCAAAT	239	0.1733	S16 ribosomal protein	M11409
25.	GCCAAGGGTC	237	0.1719	Ribosomal protein L29	AF236069
26.	AACAATTTGG	229	0.1661	60S ribosomal Protein L9	AF260271
27.	CTGAACATCT	222	0.161	Acidic ribosomal phosophoprotein PO	X15267
28.	ATCCGAAAGA	218	0.1581	Ribosomal protein L18 (Rpl18)	L04128
29.	CCCCAGCCAG	214	0.1552	Ribosomal protein S3	X76772
30.	ATTCTCCAGT	203	0.1472	EST, similar to Human 60S ribosomal protein L17	W98916
31.	CAGAACCCAC	199	0.1443	Ribosomal protein Ke-3 (S13 homologue)	M76763
32.	AGGTCGGGTG	192	0.1393	EST, similar to 60S ribosomal protein L13A (Rpl13a)	W85200
33.	GTGAAACTAA	192	0.1393	EST, similar to ribosomal protein S4	BE336059
34.	AAGGCAAAGA	188	0.1363	EST, similar to Yeast probable 60S ribosomal protein L14EA	W33609
35.	TATGTCAAGC	185	0.1342	Ribosomal protein S12	X15962
36.	CTAATAAAGC	173	0.1255	Monoclonal nonspecific suppressor factor β (fau)	D26610
37.	CAGTCTCTCA	168	0.1218	EST, similar to Rat 40S ribosomal protein S10	W90870
38.	CCAGAACAGA	165	0.1197	EST, similar to ribosomal protein L30	W97761
39.	CTGTAGGTGA	163	0.1182	EST, similar to Human 40S ribosomal protein S23	W77106
40.	TGTAGTGTAA	161	0.1168	Ribosomal protein S8	X73829
41.	GGCTTCGGTC	155	0.1124	EST, similar to acidic ribosomal phosphoprotein P1	W98417
42.	TGGATCAGTC	155	0.1124	Ribosomal protein L19	M62952
43.	GTGACCACGG	145	0.1052	18S ribosomal protein	K01364
44.	CCTTTGAGAT	136	0.0986	Ribosomal protein S5	U78085
45.	GAGGAGAAGA	134	0.0972	J1 protein (yeast ribosomal protein L3 homologue)	Y00225
46.	CTGCTATCCG	132	0.0957	Ribosomal protein L5	X83590
47.	CCTACCAAGA	130	0.0943	EST, similar to Human 40S ribosomal protein S20	W99238
				Cytoplasmic Proteins	
48.	CCCTGGGTTC	669	0.4853	Ferritin light chain (Ftl2)	J04716
49.	AGGCAGACAG	561	0.407	Elongation factor 1-α (EF 1-α, eEF-Tu)	X13661
50.	GGCAAGCCCC	430	0.3119	Csa19	U12403
51.	TGGGCAAAGC	404	0.2931	EST, similar to Human elongation factor 1-γ	W97412
52.	GCCAAGTGGA	328	0.2379	Elongation factor 2 (ef-2)	M76131
53.	GTGAGCCCAT	278	0.2016	EST, similar to heat-shock protein hsp84	W97396
54.	TCAGGCTGCC	247	0.1792	Ferritin heavy chain (fth)	M24509
55.	CGCTGGTTCC	219	0.1588	EST, similar to nonamer binding protein	W99121
56.	GATGTGGCTG	193	0.14	Elongation factor 1-β homolog	AF029844
57.	CTCGAGTCTC	150	0.1088	Poly(A) binding protein	X65553
58.	ATTGTACCAG	134	0.0972	Cellular retinoic acid-binding protein (Crabp)	M31552
59.	GAGGCTTTGC	131	0.095	Heat-shock protein hsp86	M36830

SUPPLEMENT B. 100 Most Abundant Tags in R1 ES Cells (Continued)

				Cytoplasmic Proteins (Continued)	
60.	TATCCCACGC	118	0.0856	EST, similar to calgizzarin	BE135639
61.	CCCTACTTCA	116	0.0841	H19	X58196
				Enzymes	
62.	GCCTCCAAGG	825	0.5985	Glyceraldehyde-3-phosphate dehydrogenase	M32599
63.	GCCCGGGAAT	299	0.2169	Hexokinase	J05277
64.	CCTACTAACC	237	0.1719	Aldolase A (Alda)	Y00516
65.	CACCACCACA	209	0.1516	mGpi1	AB008920
66.	TTCCAGCTGC	189	0.1371	Phosphoglycerate mutase 1 (Pgam1)	AF283667
67.	CAATAGAGAC	145	0.1052	Pyruvate kinase M	D38379
68.	GCAATCTGAT	141	0.1022	X chromosome-linked phosphoglycerate kinase (pgk-1)	M15668
69.	CCTCAGCCTG	139	0.1008	Cathepsin D (Ctsd)	X52886
70.	GCCTGTGGCC	125	0.0906	EST, similar to protein disulfide isomerase (P4hb)	BB541528
				Cytoskeleton Proteins	
71.	TGGCTCGGTC	364	0.264	Cytoplasmic γ-actin	X13055
72.	GCAGGCACTC	345	0.2503	β-tubulin (isotype Mβ 5)	X04663
73.	CGCCTGCTAG	318	0.2307	Pro-α-2(I) collagen (Col1a2)	X58251
74.	GAAGCAGGAC	244	0.177	Cofilin	D00472
75.	GGGGAAATCG	241	0.1748	Testis-specific thymosin β-10	Z48496
76.	CCCTGAGTCC	192	0.1393	Cytoskeletal β-actin and A-X actin	X03672
77.	GGCTGGGGGC	187	0.1356	Profilin (Pfn)	X14425
78.	ATGTCTCAAA	159	0.1153	α-tubulin isotype M-α-2	M13446
79.	TGCTACCCTT	115	0.0834	Vascular smooth muscle α-actin	X13297
				Energetic Metabolism Proteins	
80.	GCTGCCCTCC	1294	0.9388	EST, similar to cytochrome C oxidase polypeptide I (COX1)	AV380522
81.	CAGGCCACAC	251	0.1821	ATP synthase β-subunit (β-F1 ATPase)	AF030559
82.	AGCAGTCCCC	187	0.1356	EST, similar to cytochrome C oxidase polypeptide II (COX2)	W98306
83.	ATACTGACAT	128	0.0928	EST, similar to Human cytochrome C oxidase polypeptide III (COX3)	W17837
				Secretory Proteins	
84.	AACGCTGCCA	362	0.2626	Macrophage migration inhibitory factor (Mif, growth factor-induced delayed early response protein, glycosylation-inhibiting factor)	U19825 X66532
85.	GCGGCGGATG	245	0.1777	β-galactoside specific lectin (L14 lectin)	M20692
86.	CAAACTCTCA	239	0.1733	Osteonectin (Sparc, secreted acidic cysteine-rich protein)	W97475
87.	CCAACGCTTT	180	0.1305	EST, similar to fibronectin	X12697
88.	GAACATTGCA	177	0.1284	Osteonectin (Sparc, secreted acidic cysteine-rich protein)	M12414
89.	CATCGCCAGT	117	0.0848	Apolipoprotein E (Apo E)	
				Membrane Proteins	
90.	CAAACACCGT	263	0.1908	Minopontin	X13986
91.	GTCTGCTGAT	230	0.1668	G protein β subunit homologue	D29802
				Repeat Sequences	
92.	GTGGCTCACA	2271	1.6476	B2 repeat sequence	M55334
93.	TGGTTGCTGG	111	0.0805	Alu-type II repeat element	M30379
				Miscellaneous	
94.	TCCACCCAAG	309	0.2241	ETn transposon	X15598
95.	TGGGTTGTCT	302	0.2191	21 kd polypeptide under translational control	X06407
96.	GAGCGTTTTG	225	0.1632	Cyclophilin	X52803
97.	AAGACCCTGG	152	0.1102	pH 34	X57708
				Multiple Matches	
98.	CCCACAAGGT	160	0.116	Putative protein kinase YSK2; Ribosomal protein L27 (Rpl27)	U49949 AF214527
99.	GTGGGCTCAC	135	0.0979	EST, similar to ERCC2; EST, similar to MHC class I H2-D gene	C77489 AA616702
				No Similarity	
100.	TGACCCCCGG	123	0.0892	No matches (3'-RACE [b]: L41 ribosomal-like protein)	-

[a] Tags sorted according to abundance within individual functional categories.

[b] 3'-RACE data generated in this study.

Similarity search was performed with GenBank Non-Redundant database (release 120.0) and EST database (release 120.0).

Supplement C:
Oligonucleotide primers used for RT-PCR

Anisimov SV, Tarasov KV, Tweedie D, Stern MD, Wobus AM, Boheler KR: SAGE Identification of Gene Transcripts with Profiles Unique to Pluripotent Mouse R1 Embryonic Stem Cells

[a] Forward (F) and reverse (R) primers were designed from mouse GenBank database sequences.

SUPPLEMENT C. Oligonucleotide primers.

Transcript	Primer sequence [a]	PCR Product
Alda	F-5'-cccctggccatgcttgcacccaga-3'	224 bp
	R-5'-agaggcctgcagggctcgaccata-3'	
Alad	F-5'-caggagtggccaggtatggcgtaa3'	176 bp
	R-5'-cctcagcagacggacagcctcaat-3'	
Bax	F-5'-cgcgtggttgccctcttcta-3'	223 bp
	R-5'-cggtgaggactccagccaca-3'	
Ctsz	F-5'-tctggaagcagcgctgtgaa-3'	249 bp
	R-5'-cgggtggctgaaggcaagca-3'	
Cdk4	F-5'-cccgtggacatgtggagcgt-3'	172 bp
	R-5'-gggcaaaggctcctcgaggt-3'	
Csa19	F-5'-ccgcagccatgagcagcaaagtct-3'	185 bp
	R-5'-tggggtggacttgagcctgacggt-3'	
Dsn-pending	F-5'-ccccagcgaacatggcctca-3'	195 bp
	R-5'-cgccaacatctcccaccaag-3'	
Ftl2	F-5'-gctgccttggccatggagaagaac-3'	163 bp
	R-5'-ggaggttggtcagatggttgccca-3'	
Fbln2	F-5'-tgctcctgcgctgctggctt-3'	231 bp
	R-5'-aggcacgttgacacagcgga-3'	
Gal	F-5'-ctgcctgagagcaacattgtccgca-3'	149 bp
	R-5'-gcacagtgggcatggtctcaggact-3'	
Galr1	F-5'-gcgatgctggggtgttcctcactat-3'	198 bp
	R-5'-ggaggctaagaagtagcccagccgt-3'	
Galr2	F-5'-agcgcaaggtgacacggatgatcgt-3'	214 bp
	R-5'-cctttgcggaaatgcttggagacca-3'	
	F-5'-gtcgcagctggccaatctgacggtg-3'	226 bp
	R-5'-cgccacgatgacgatcatccgtgtc-3'	
Galr3	F-5'-cgcctgtcgtctggcttcgcactgc-3'	177 bp
	R-5'-ggctggacacgacggagggctcgat-3'	
	F-5'-cagcctctccaagctccaccgaaaa-3'	211 bp
	R-5'-gggcttgggaccaaaagggtctagt-3'	
Ier3	F-5'-gtgcgagatatgcctccctg-3'	214 bp
	R-5'-ccggattcccaagggtcttc-3'	
H19	F-5'-ccctaaccgctcagtccctgggtc-3'	171 bp
	R-5'-tcccctggctcaccaccccagcta-3'	
Hk1	F-5'-ccgcagcgtccagaggaaagagcg-3'	158 bp
	R-5'-agccacccctgtgggcagagtccg-3'	
Klf2	F-5'-ctgctggaggccaagcccaa-3'	287 bp
	R-5'-aggtggtcggacctggagaa-3'	
Tubb5	F-5'-gaagaggaggcctaacggcagagagccct-3'	174 bp
	R-5'-gagcgtgtcccatgtgccaggcaccattt-3'	
Mmp14	F-5'-gcgactgctttactgccagcgt-3'	251 bp
	R-5'-gaccgtcttctgctcagccctc-3'	
Msn	F-5'-acggcacggatctcccagtt-3'	242 bp
	R-5'-tcactgcggtccttggccat-3'	
Nid2	F-5'-gcaccaggcctgcatgcata-3'	273 bp
	R-5'-tggcacggctgatggtccga-3'	
Nme1	F-5'-cgcgtgatgcttggagagac-3'	157 bp
	R-5'-ccagctcctcaggctgaaac-3'	
Pal31	F-5'-cctcccgaagctacctaagt-3	291 bp
	R-5'-ggtgcttcctggtcctctcg-3	
Pf4	F-5'-tgcggtgtttcgaggcctcc-3'	219 bp
	R-5'-atgagctggggaaccgcaca-3'	
Sn	F-5'-cactggctggttgtatgtggttt-3'	205 bp
	R-5'-gtaactgcttgatatagcttgggca-3'	
Tdgf1	F-5'-aaagaacctgccgtacgcga-3'	211 bp
	R-5'-gcagccaggtgccatggagg-3'	
Tnc	F-5'-cgggtccctgaggagcaagc-3'	275 bp
	R-5'-tgcggtctccaaacccagca-3'	

[a] Forward (F) and reverse (R) primers were designed from mouse GenBank database sequences.