Supplemental Materials

Methods

Predicting total reads

We predicted the total number of reads that would be required to obtain the number of MEND reads required for a threshold. Specifically, for each of 765 RNA-Seq samples, we determined the factor required to transform the number of existing MEND reads to the desired number of MEND reads, and then performed the same transformation on the remaining read types (e.g. duplicate, not mapped, etc). This number is an estimate because the fraction of non-duplicate reads, and thus MEND reads, is not consistent across depths. The same library, sequenced more deeply, will have a greater fraction of computationally detected duplicates because the universe of possible non-duplicate reads decreases with every additional read.

Predicting duration of pipelines

Timestamps were recorded before and after each step of each pipeline for each sample. Because both pipelines have steps that are not dependent on input size (e.g. loading reference files info memory), a start-up time for each pipeline was estimated based on the median amount of time required to process a sample containing 1 million total reads. That time was subtracted from the duration of the pipeline, and the remaining adjusted duration was used to calculate the speed of the pipeline in minutes per million total reads. The median speed plus median startup time was used for predicting total pipeline duration for arbitrary read depths.

Tables

Table S1. Sample Information

Sample	Treehouse ID	Project information	Sample ID in project	Age
S1	THR28_0688_S01	SRA: SRP040454	SRR1201253	15
S2	THR14_0298_S01	https://doi.org/10.24370/SD_BHJXBDQK	C15990	1.7
S3	THR17_0392_S01	EGA: EGAD00001000158, https://dx.doi.org/10.1038/nature11327	EGAZ00001000212_81CMBABXX_3	4.5
S4	THR31_0892_S01	dbGap: phs000673.v2.p1, https://dx.doi.org/10.1001/jama.2015.10080	52	<30
S5	THR32_0941_S01	EGA: EGAD00001001927, https://dx.doi.org/10.1016/j.cell.2016.01.015	dkfz_CNS-PNET_15-0069	3

Table S2. Sample read type composition

(Percent MMoM is the fraction of mapped reads that are multi-mapped, i.e.. multi-mapped of mapped)

ID	Total sequences	Not mapped	Duplicates	Non-exonic	MEND	Percent MEND	Percent MMoM
S1	128,535,112	4,835,735	42,042,547	13,092,258	68,564,572	53.3	8.2
S2	90,208,409	3,153,361	28,542,492	4,155,515	54,357,041	60.3	9.1
S3	84,452,915	5,555,879	13,178,433	15,510,051	50,208,552	59.5	6.9
S4	109,466,113	4,779,467	42,401,564	4,410,048	57,875,034	52.9	4.3
S5	109,226,298	1,463,129	34,803,665	5,056,818	67,902,686	62.2	8

Table S3. Sub-sample read type composition

(Percent MMoM is the fraction of mapped reads that are multi-mapped, i.e.. multi-mapped of mapped)

ID	Target MEND Count (M)	Seed	Total sequences	Not mapped	Duplicates	Non exonic	MEND	Percent MEND	Percent MMoM
S1	1	2771	1,191,561	44,931	32,840	144,746	969,044	81.3	8.2
S1	4	5411	5,242,557	197,506	391,682	631,010	4,022,360	76.7	8.2
S1	8	2622	10,907,904	411,416	1,223,836	1,299,434	7,973,218	73.1	8.2
S1	12	9311	16,876,470	634,535	2,357,704	1,994,906	11,889,325	70.4	8.2
S1	16	6933	23,145,211	870,759	3,749,472	2,527,112	15,997,869	69.1	8.2
S1	20	2226	29,709,318	1,118,269	5,389,331	3,227,528	19,974,190	67.2	8.2
S1	24	1120	36,559,290	1,376,837	7,261,258	3,946,870	23,974,325	65.6	8.2
S1	28	6451	43,665,129	1,642,663	9,351,746	4,694,195	27,976,525	64.1	8.2
S1	32	3920	51,002,535	1,919,972	11,649,902	5,459,935	31,972,726	62.7	8.2
S1	36	7519	58,558,309	2,203,450	14,154,168	6,232,705	35,967,986	61.4	8.2
S1	40	4956	66,343,416	2,496,223	16,856,241	7,028,152	39,962,800	60.2	8.2
S1	44	7750	74,358,987	2,797,435	19,761,936	7,834,893	43,964,723	59.1	8.2
S1	48	9314	82,603,751	3,108,502	22,863,202	8,658,070	47,973,977	58.1	8.2
S2	1	7173	1,143,477	39,754	46,370	91,062	966,290	84.5	9.1
S2	4	2711	4,993,927	174,592	473,621	338,840	4,006,874	80.2	9.1
S2	8	1490	10,420,956	364,200	1,399,780	639,665	8,017,312	76.9	9.1
S2	12	3931	16,175,345	566,516	2,629,156	931,813	12,047,861	74.5	9.1
S2	16	3498	22,257,160	779,144	4,126,796	1,444,917	15,906,303	71.5	9.1
S2	20	3732	28,636,088	1,000,462	5,876,649	1,800,789	19,958,188	69.7	9.1
S2	24	4424	35,215,412	1,231,421	7,837,390	2,152,988	23,993,613	68.1	9.1
S2	28	5510	41,978,201	1,466,507	9,992,408	2,510,114	28,009,172	66.7	9.1
S2	32	4233	48,922,115	1,710,060	12,345,431	2,871,763	31,994,861	65.4	9.1
S2	36	1259	56,088,802	1,961,176	14,893,202	3,234,947	35,999,477	64.2	9.1
S2	40	8596	63,469,370	2,217,871	17,648,382	3,602,206	40,000,910	63	9.1
S2	44	6556	71,062,797	2,484,142	20,603,862	3,980,378	43,994,415	61.9	9.1
S2	48	7023	78,868,925	2,756,316	23,764,239	4,359,219	47,989,151	60.8	9.1
S3	1	7649	1,333,110	87,652	12,656	268,050	964,753	72.4	6.8
S3	4	8539	5,468,018	359,684	154,118	1,103,855	3,850,360	70.4	6.9
S3	8	7537	11,170,527	735,105	522,538	2,158,143	7,754,741	69.4	6.9

S3	12	5537	17,089,047	1,123,015	1,060,466	3,314,643	11,590,923	67.8	6.9
S3	16	662	23,225,066	1,528,638	1,745,342	4,500,740	15,450,346	66.5	6.9
S3	20	7248	29,566,535	1,945,301	2,573,368	5,729,301	19,318,564	65.3	6.9
S3	24	6740	36,082,741	2,374,840	3,527,936	6,994,714	23,185,251	64.3	6.9
S3	28	4433	42,807,901	2,815,162	4,614,248	8,281,769	27,096,721	63.3	6.9
S3	32	8154	49,615,635	3,266,617	5,811,608	9,585,236	30,952,174	62.4	6.9
S3	36	4791	56,421,595	3,711,929	7,098,102	10,877,421	34,734,143	61.6	6.9
S3	40	2115	63,293,109	4,164,264	8,471,480	12,179,621	38,477,744	60.8	6.9
S3	44	849	70,219,354	4,617,734	9,939,252	12,952,812	42,709,556	60.8	6.9
S3	50	3541	80,704,867	5,309,035	12,295,582	14,839,989	48,260,261	59.8	6.9
S4	1	5315	1,045,377	45,876	45,752	56,363	897,386	85.8	4.3
S4	4	4596	5,043,401	219,745	561,849	223,190	4,038,617	80.1	4.3
S4	8	2594	10,744,554	468,240	1,738,179	490,852	8,047,283	74.9	4.3
S4	12	3263	16,860,169	736,956	3,329,810	785,977	12,007,426	71.2	4.3
S4	16	2214	23,395,401	1,021,538	5,285,918	1,100,864	15,987,081	68.3	4.3
S4	20	7447	30,319,271	1,323,675	7,582,848	1,435,276	19,977,472	65.9	4.3
S4	24	734	37,446,782	1,635,768	10,132,510	1,777,813	23,900,692	63.8	4.3
S4	28	8899	44,785,147	1,955,663	12,934,337	1,750,624	28,144,523	62.8	4.3
S4	32	3452	52,394,319	2,287,791	15,986,324	2,064,188	32,056,016	61.2	4.3
S4	36	8038	60,348,764	2,635,559	19,320,624	2,393,625	35,998,956	59.7	4.3
S4	40	3320	68,626,163	2,995,634	22,937,514	2,733,306	39,959,708	58.2	4.3
S4	44	3778	77,224,528	3,370,685	26,823,490	3,086,454	43,943,900	56.9	4.3
S4	48	4374	86,145,326	3,759,043	30,981,409	3,456,192	47,948,682	55.7	4.3
S5	1	8803	1,095,917	14,652	36,657	75,806	968,802	88.4	8
S5	4	7362	4,768,333	63,729	388,644	326,465	3,989,495	83.7	7.9
S5	8	4808	9,889,727	132,092	1,163,274	561,700	8,032,662	81.2	7.9
S5	12	3737	15,263,801	203,986	2,194,178	865,464	12,000,172	78.6	8
S5	16	7924	20,893,267	279,963	3,441,558	1,183,454	15,988,293	76.5	8
S5	20	7125	26,762,540	358,672	4,900,294	1,512,722	19,990,853	74.7	8
S5	24	1819	32,823,979	439,686	6,537,826	1,855,955	23,990,511	73.1	8
S5	28	2542	39,075,013	524,456	8,353,716	2,207,201	27,989,640	71.6	8
S5	32	1872	45,508,328	610,341	10,343,082	2,564,317	31,990,588	70.3	8
S5	36	2104	52,114,856	698,827	12,498,577	2,929,734	35,987,718	69.1	7.9
S5	40	753	58,896,397	788,996	14,819,092	3,304,205	39,984,105	67.9	8

S5	44	6788	65,854,950	882,393	17,302,914	3,686,839	43,982,804	66.8	8
S5	48	8931	72,989,829	977,023	19,952,874	4,072,439	47,987,492	65.7	8

Figures

Figure S1

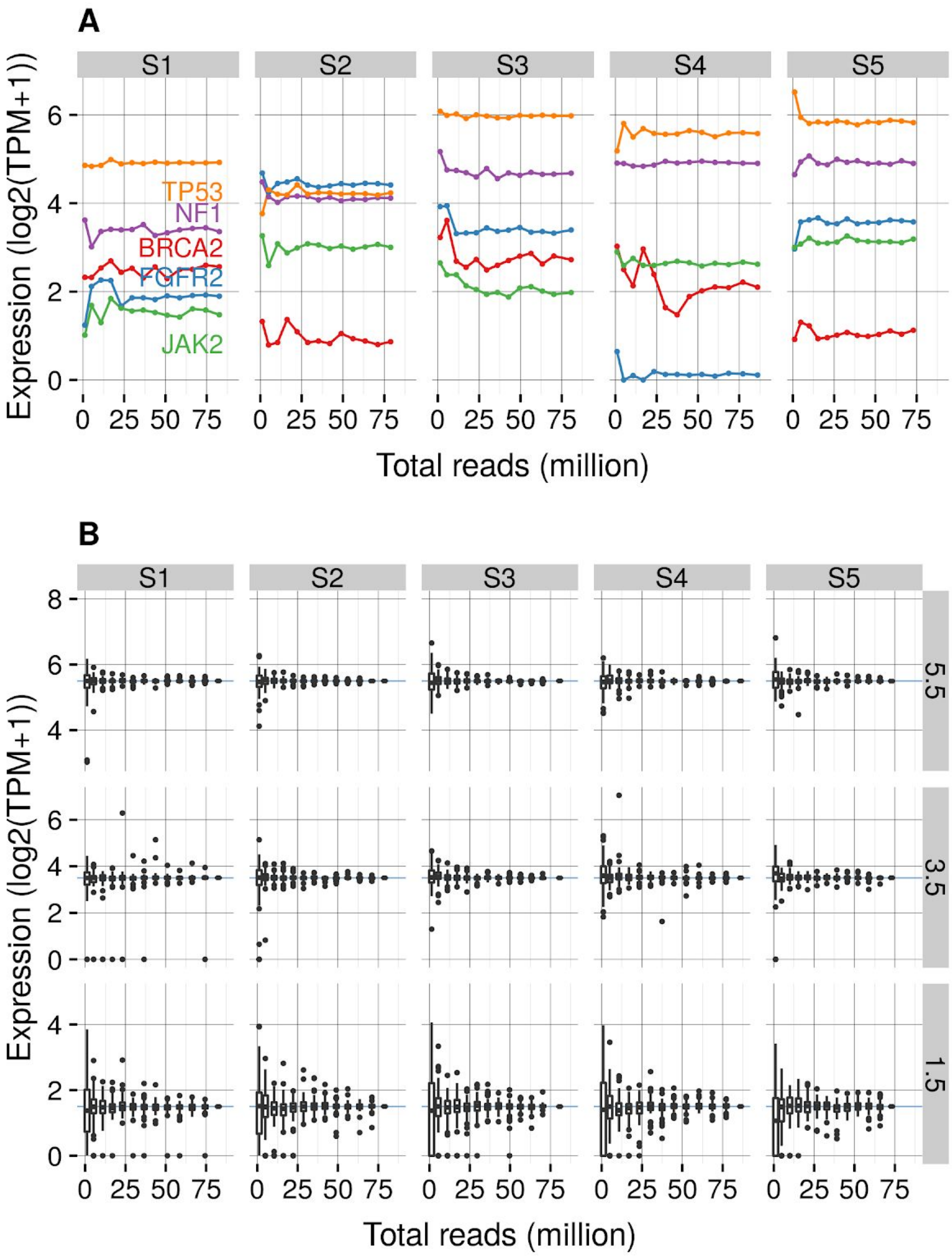

Figure S1. Reproducibility of gene expression measurement increases with depth of sequence in all parent samples.

Caption: Gene expression (y-axis) is plotted against the number of total reads in the measured subsample. A. Each point represents the measurement of a gene in a subsample. Each panel represents a different parent sample. B. Expression of groups of genes with the same expression at higher depths. Each boxplot represents 73 gene measurements. Each panel contains genes that are expressed within 0.02 of 5.5, 3.5 or 1.5 log2(TPM+1) (horizontal blue lines) at the highest depth.

Figure S2

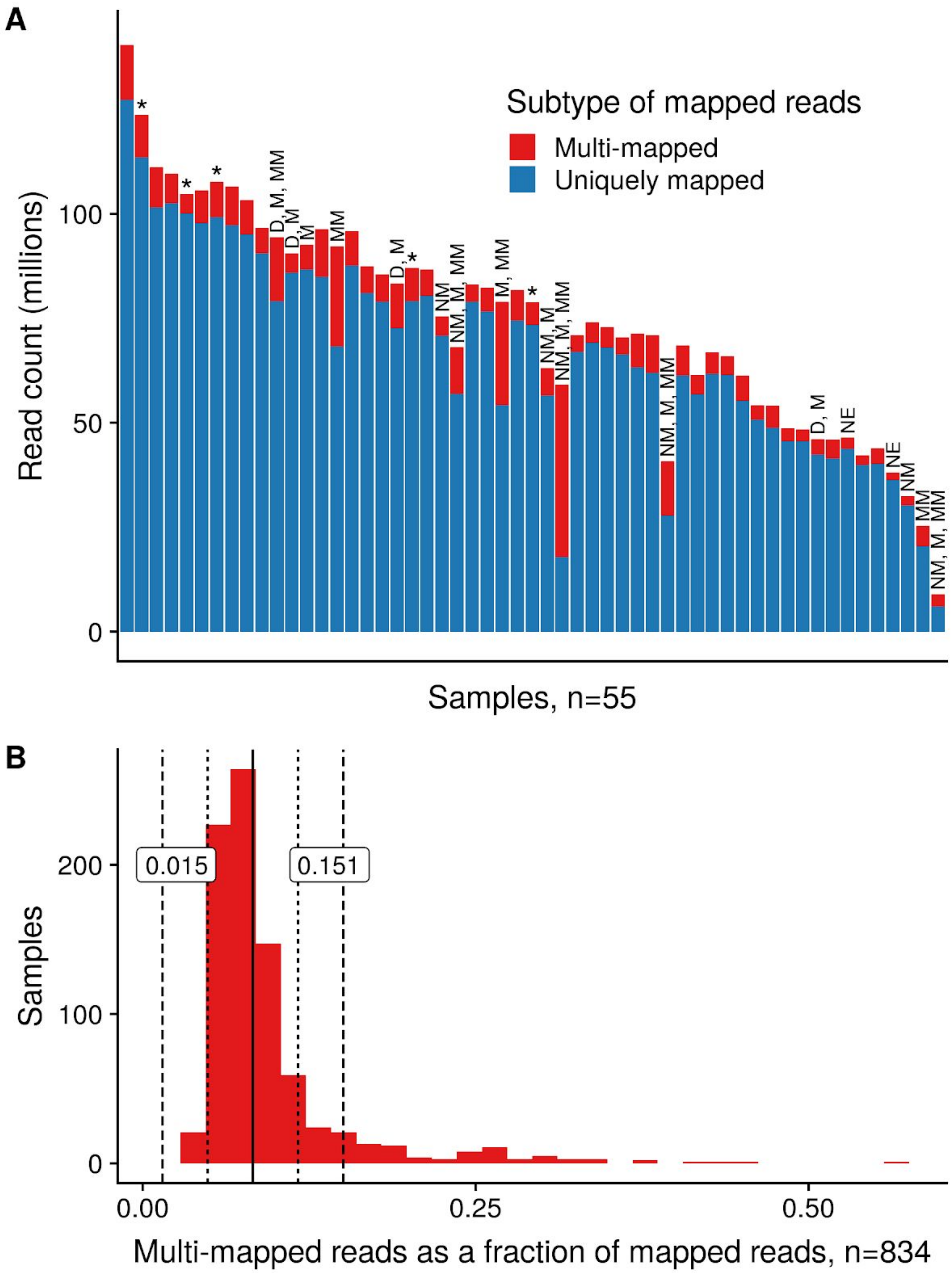
A
Read count (millions)
0
50
100
Samples, n=55
Subtype of mapped reads
Multi-mapped
Uniquely mapped
*
*
*
D, M, MM
D, M
M
MM
D, M
*
NM
NM, M, MM
M, MM
*
NM, M
NM, M, MM
NM, M, MM
D, M
NE
NE
NM
MM
NM, M, MM
B
Samples
0
100
200
0.015
0.151
0.00
0.25
0.50
Multi-mapped reads as a fraction of mapped reads, n=834

Figure S2. The fraction of multi-mapped reads is variable.

Caption: A. Subtypes of mapped reads are present in fifty-five representative RNA-Seq samples, including uniquely mapped (blue) and multi-mapped (red). Asterisks indicate parent samples from Table 1. Samples with text annotations NM (not mapped), D (duplicates), NE (non-exonic), MM (multi-mapped), or M (MEND) have read type ratios outside the limits (Figs 2B, S2B). B. The typical range of multi-mapped reads as a fraction of all mapped reads (μ ± 2σ) is shown for the 834 samples with more than 20 million MEND reads.