Developmental Cell, Volume *23*

Supplemental Information

Silencing of Germline-Expressed Genes by DNA Elimination in Somatic Cells

Jianbin Wang, Makedonka Mitreva, Matthew Berriman, Alicia Thorne, Vincent Magrini, Georgios Koutsovoulos, Sujai Kumar, Mark L. Blaxter, and Richard E. Davis

Inventory of Supplemental Information

Supplementary Information files contain supplemental experimental procedures, 4 figures and 6 tables. Two files (**Table S5** and **S6**) are in separate Excel (.xlsx) sheets. The rest of the figures and tables are combined into this PDF file.

Figures:

Figure S1 is related to **Figure 2**. Representative *A. suum* genomic scaffolds with genome reads, RNA-seq data, and gene models.

Figure S2 is related to **Figure 3**. MEME motif analysis of sequences around breakpoints with telomere addition.

Figure S3 is related to **Figure 5**. Dot plot matrices of differential RNA expression in *A. suum*.

Figure S4 is related to **Table 1**. Base coverage of scaffolds for genome size estimation.

Tables:

Table S1 is related to **Table 1**. *A. suum* genome and cDNA libraries and sequencing.

Table S2 is related to **Figure 2**. *A. suum* repetitive sequences in the germline and somatic genomes.

Table S3 is related to **Figure 3**. *A. suum* somatic DNA breakpoints with telomere addition.

Table S4 is related to **Figure 4**. Loss of *A. suum* duplicated, rearranged loci in somatic cells.

Table S5 is related to **Figure 5**. *A. suum* genes, tissue expression, and eliminated genes.

Table S6 is related to **Figure 5**. Eliminated *A. suum* genes.

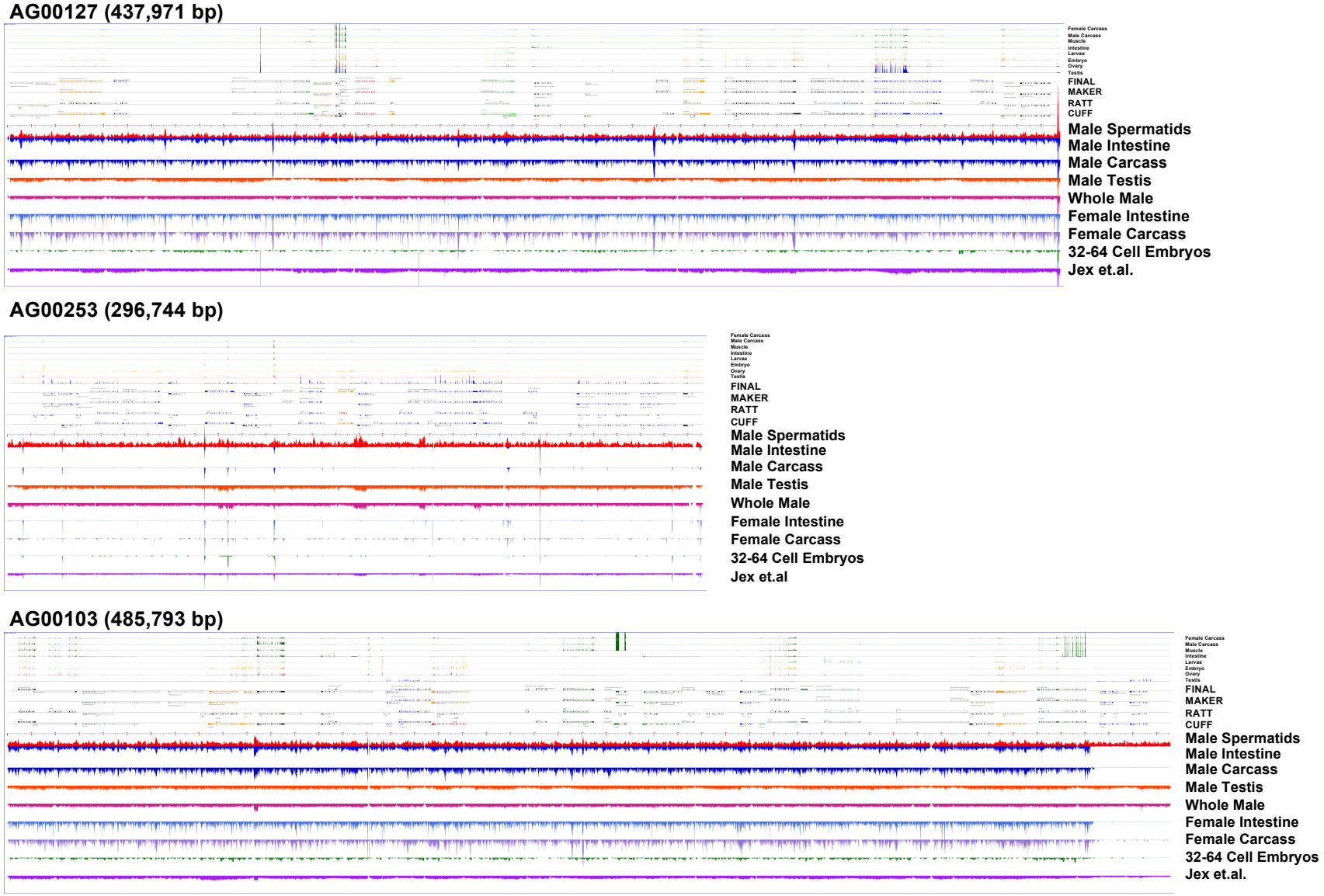

Figure S1 (related to Figure 2). Representative *A. suum* genomic scaffolds with genome reads, RNA-seq data, and gene models. The scaffolds illustrated represent the examples presented in **Fig. 2A**. Note that data for all the scaffolds are available at http://ascaris.nematodegenomes.org/ (see scaffold picture track). Above the coordination line are gene models and RNA-seq data coverage. Four different tracks for the gene models are illustrated. From bottom to top: 1). RNA-seq (tophat/cufflinks); 2). Transferred annotation from the Jax et al. Ascaris assembly (RATT); 3). Gene models from an annotation pipeline (MAKER); and 4). Final integrated gene models. The genes are colored based on their expression using the same scheme as in **Fig. 5**, where blue = testis, red = ovary, orange = embryo, light green = larvae, green = intestine, dark green = other somatic tissues (carcass), and grey is other. Eight normalized RNA-seq data tracks are shown above the gene models. From the bottom to top: 1). Testis; 2). Ovary; 3). Embryo; 4). Larvae; 5). Intestine; 6). Muscle; 7). Male carcass; and 8). Female carcass. Below the coordinate line are genomic reads coverage tracks. Note that the scales for each track are different. Tracks from top to bottom are: 1). Single male spermatids, 2). Single male intestine, 3). Single male other somatic cells, 4). Single male testis, 5). Single whole male (Belgium), 6). Single female intestine, 7). Single female other somatic cells, 8). Mixed population 32-64 cell embryos, and 9). Data from Jex et.al.

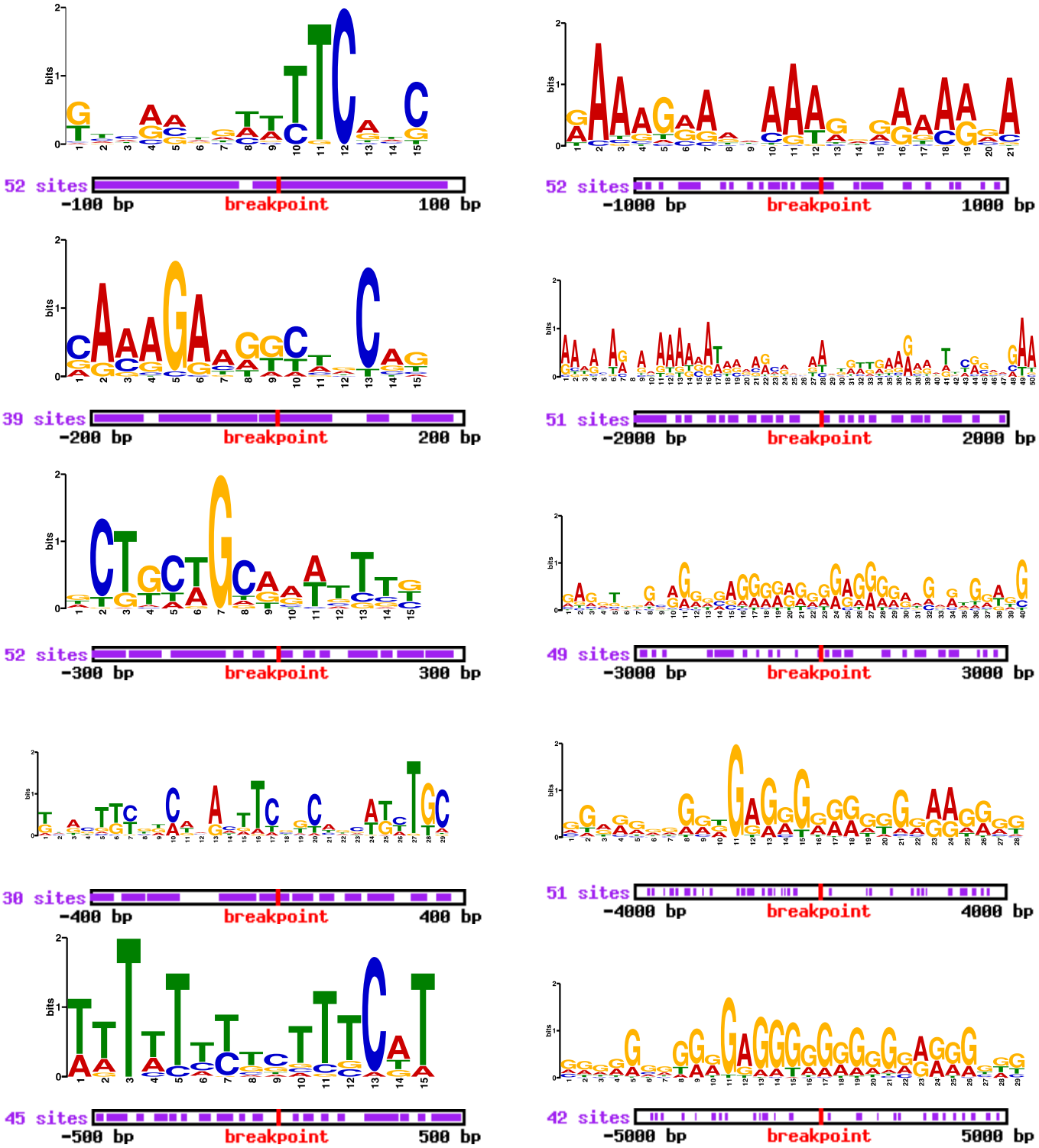

Figure S2 (related to Figure 3). MEME motif analysis of sequences around breakpoints with telomere addition. We used sequences around all the 52 breakpoints flanking 100, 200, 300, 400, 500, 1,000, 2,000, 3,000, 4,000, and 5,000 bp for the MEME motif analysis. Shown are the best motif logos for all these different window sizes. Below each motif logo is the number of sites (out of 52 loci) that have this motif and their positions (purple bars) in the sequence. Note MEME will always give the best motif even if no real motif exists. We see no consistent motif associated with the breakpoints in that: 1). The motif scores are in general not strong; 2). Each window size gives a very different motif; 3). Within each window, the motif sites for different loci are randomly distribution along the window, instead of at a specific location, 4). For over half of the window sizes, not all of the loci contain the motifs; and 5). Using random sequences from the genome or from simulated random sequences give similar results (data not shown).

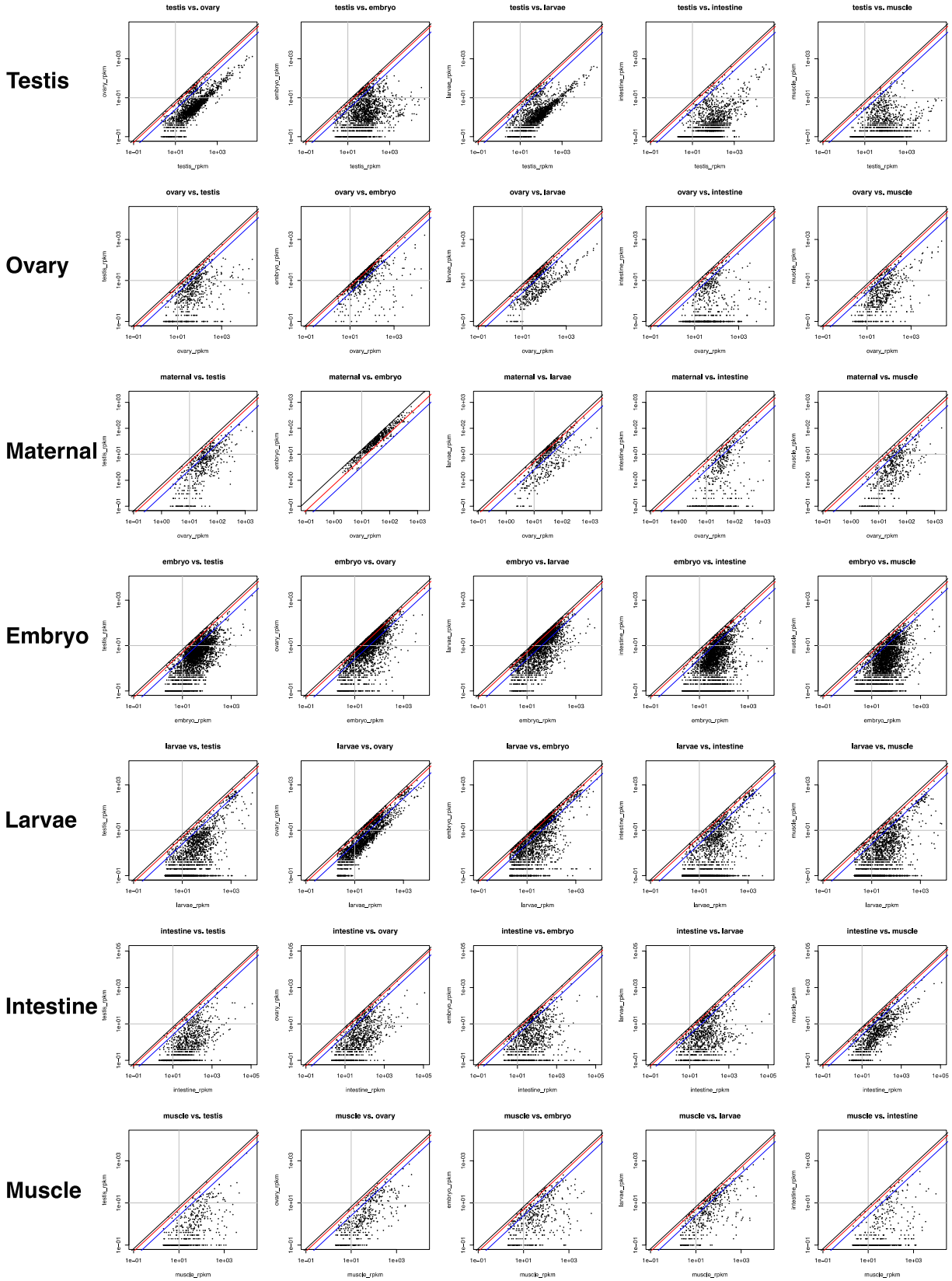

Figure S3 (related to Figure 5). Dot plot matrices of differential RNA expression in *A. suum*. Dot plot illustrating the expression level (RPKM) of groups of enriched genes in one tissue as compared to other tissues. The black, red and blue lines mark 1.5-, 2 and 4-fold difference (2/3-, 1.5-and 3-fold difference for the maternal vs. embryo comparison). Note the x and y axis are in log 10 scale, and y axis value in 0 is shown as 0.1 (1e-01).

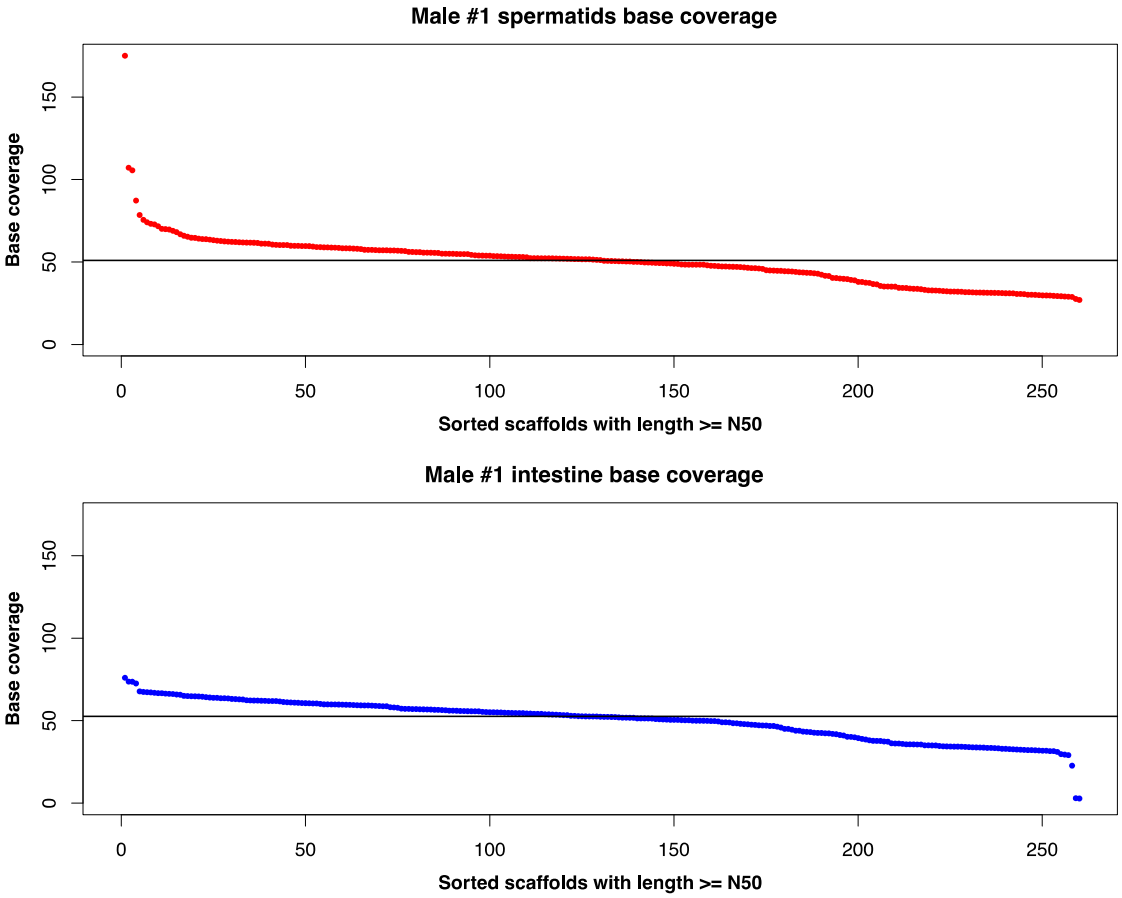

Figure S4 (related to Table 1). Base coverage of scaffolds for genome size estimation. Base coverage for male spermatids and intestine genomic reads for scaffolds with length >= N50. The black line indicates the median coverage value (51.0 x for germline and 52.6 x for somatic).

Table S1 (related to Table 1). *A. suum* genome and cDNA libraries and sequencing

Genome

	DNA source	Insertion size (bp)	Reads (million)	Sequencing type[4]	Genome Coverage (x)
Genomic reads for assembly	Male #1 spermatids	360	134	2 x 150	51.0
	Male #1 testis	5,000	250	2 x 100	80.0
	Male #1 intestine	360	108	2 x 150	52.6
	Male #1 intestine	480	490	2 x 100	182.0
Genomic reads for scaffolding	Male #2 (Belgium)	520	240	2 x 75	54.5
	Male #1 carcass[1]	330	323	2 x 100	95.2
	Female #1 carcass[2]	380	198	2 x 100	52.9
	32-64 cell embryos	5,500	0.43	2 x 700	1.0
Additional Genomic reads for analysis	Female #1 intestine	360	143	1 x 50	23.4
	Jex et.al[3]	170 - 10,000	272	2 x 50, 2 x 100	80.0

cDNA

RNA-seq data for expression analysis	Total Reads (million)	Average Read length (bp)	Total Bases (Gb)
Testis	21.92	84	18.4
Ovary	18.50	91	16.8
Embryo (24 hr - 64 hr)	64.96	40	26.0
Larvae (7 days - L2)	56.57	90	50.9
Intestine	23.20	45	10.7
Muscle (and hypodermis)	51.38	45	23.6
Male carcass[1]	26.83	45	12.1
Female carcass[2]	8.70	45	3.9

1. Male carcass = whole male with testis, spermatids, and intestine removed
2. Female carcass = whole female with ovary and oviduct, uterus with embryos, and intestine removed
3. Jex, A.R., et al., Ascaris suum draft genome. Nature, 2011. 479(7374): p. 529-33.
4. 2 x 150, paired-end read of 150 nt length; 2 x 100, paired-end read of 100 nt length; 1 x 50, single end read of 50 nt length; etc.

Note: for genomic libraries, germline are in red, somatic are in blue, and mixed in purple.

Table S2 (related to Figure 2). *A. suum* repetitive sequences in the germline and somatic genomes

Repeat Sequence Name	Repeat Unit Length (bp)	Germline Base Pair (bp)	Germline Copies[2]	Somatic Base Pair (bp)	Somatic Copies[2]
Chromatin satellites (121bp tandem repeats)[1]	121	29,708,645	245,526	1,630,866	13,478
Telomeric repeats (TTAGGC tandem repeats)	6	1,785,277	297,546	1,061,940	176,990
CATAT tandem repeats	5	381,218	76,244	668,062	133,612
AAAGAG tandem repeats	6	56,528	9,421	134,095	22,349
Ribosomal RNAs (18s-5.8s-26s rRNA)	6,242	2,977,434	477	3,395,648	544
Splice leader RNAs (SL and 5S rRNA)	1,467	391,689	267	384,354	262
PAO LTR retrotransposons	6,346	78,056	12	34,268	5
TAS LTR retrotransposons	7,813	31,352	4	20,323	3
R4 non-LTR retrotransposons	4,687	11,718	3	18,748	4

1. All variants included
2. Error in our estimations for both the germline and somatic copy number is in the range of -13% to +21%

Table S3 (related to Figure 3). *A. suum* somatic DNA breakpoints with telomere addition

ID	Scaffold Name	Scaffold Length	Elimination Start	Elimination End	Eliminated Length	Breakpoint Position	Sperm PE Pairs[1]	Whole Male PE Pairs[1]	Testis MP Pairs[2]	Total Germline PE/MP Pairs[1,2]	Total Somatic PE Pairs[1]	CBR ID[3]	Offset[4] in male (intestine vs. carcass)	Offset[4] in female (intestine vs. carcass)	Offset[4] in intestine (male vs. female)	Offset[4] in other somatic tissue (male vs. female)	Organization in Jex et al. Assembly[5]	PCR confirmed
1	AG00002	1,456,112	1,354,806	1,456,112	101,307	1,354,806	5	48	175	228	79		1500	1700	-2900	-2700	Match	Yes
5	AG00023	857,739	855,800	857,739	1,940	855,800	41	26	209	276	31		200	3800	-3400	200	Match	
6	AG00034	778,791	247,090	299,339	52,250	247,090	18	17	286	321	55		1200	0	1700	500	Match	Yes
8	AG00062	629,399	1	18,153	18,153	18,153	44	149	742	935	131	CBR3	900	0	500	-400	End of Scaffold	
9	AG00066	610,077	585,315	610,077	24,763	585,315	0	30	709	739	108		100	0	0	-100	Match	Yes
10	AG00070	586,243	1	17,637	17,637	17,637	26	66	110	202	174		200	-200	-400	-800	Not Match	
11	AG00074	572,432	485,218	572,432	87,215	485,218	15	54	397	466	37	CBR2	1400	200	1400	200	Match	
12	AG00086	539,107	471,077	523,198	52,122	471,077	19	73	362	454	127		0	0	-200	-200	Match	
15	AG00103	485,793	452,247	485,793	33,547	452,247	32	102	668	802	167		1600	700	1300	400	Match	Yes
16	AG00104	485,063	1	20,356	20,356	20,356	28	108	701	837	79		700	0	-400	-1100	Match	
17	AG00131	427,091	1	18,198	18,198	18,198	18	53	351	422	33		0	0	-1700	-1700	Match	Yes
18	AG00132	426,041	272,023	426,041	154,019	272,023	21	18	621	660	65		500	0	-800	-1300	Match	Yes
19	AG00133	424,953	1	112,987	112,987	112,987	0	10	268	278	50		-100	-300	200	0	Match	
20	AG00134	422,627	1	84,398	84,398	84,398	6	68	664	738	89		0	0	-900	-900	Match	
21	AG00143	404,780	393,904	404,780	10,877	393,904	7	6	365	378	48		400	0	0	-400	Match	
22	AG00150	398,613	372,489	398,613	26,125	372,489	0	3	86	89	66		0	0	0	0	Not Match	
25	AG00173	371,654	1	38,875	38,875	38,875	33	92	848	973	49		100	0	-700	-800	Match	
26	AG00177	367,983	1	2,275	2,275	2,275	10	67	206	283	0		0	0	-200	-200	Not Match	
27	AG00179	366,537	1	154,577	154,577	154,577	47	203	518	768	65		500	0	-700	-1200	Match	Yes
29	AG00192	356,767	1	17,787	17,787	17,787	49	68	514	631	80		500	0	-1300	-1800	Match	
30	AG00193	354,834	353,380	354,834	1,455	353,380	92	234	303	629	75		1000	200	-500	-1300	Match	
33	AG00228	313,013	286,583	313,013	26,431	286,583	46	98	595	739	586		0	0	100	100	Match	
34	AG00236	308,108	294,455	308,108	13,654	294,455	21	117	783	921	75		0	0	-800	-800	Match	
40	AG00270	283,753	213,329	283,753	70,425	213,329	26	70	177	273	89		100	0	-900	-1000	Not Match	
41	AG00275	280,889	243,806	280,889	37,084	243,806	12	94	219	325	57	CBR1	1300	0	200	-1100	Match	
42	AG00277	280,367	147,178	280,367	133,190	147,178	12	46	355	413	43		1000	-1500	-600	-3100	Match	
43	AG00280	278,489	223,303	278,489	55,187	223,303	56	102	1206	1364	20		1000	0	1000	0	Match	
45	AG00308	256,194	1	195,877	195,877	195,877	4	44	360	408	0		100	0	-800	-900	Match	
46	AG00309	254,253	210,533	254,253	43,721	210,533	0	7	573	580	167		0	500	-700	-200	End of Scaffold	
47	AG00315	251,941	1	88,448	88,448	88,448	8	2	585	595	76		0	600	-1300	-700	Match	
48	AG00330	238,052	1	144,040	144,040	144,040	10	142	769	921	55		0	0	0	0	Match	
51	AG00486	160,064	156,443	160,064	3,622	156,443	38	186	725	949	56		0	1900	-800	1100	Match	
52	AG00515	152,148	1	71,702	71,702	71,702	52	186	123	361	56		0	500	-500	0	Not Match	
56	AG00600	127,910	1	34,000	34,000	34,000	27	46	303	376	125		0	0	1100	1100	Match	
57	AG00603	126,784	3,696	126,784	123,089	3,696	25	75	705	805	134		0	500	-1400	-900	Match	
58	AG00617	123,167	94,327	123,167	28,841	94,327	24	73	714	811	17		-200	0	-1400	-1200	Match	
61	AG00637	117,451	37,601	63,833	26,233	37,601	27	60	344	431	53	CBR21	0	0	0	0	Match	
64	AG00639	117,388	98,009	117,388	19,380	98,009	21	83	837	941	96		0	200	-1600	-1400	Match	
66	AG00659	113,593	2,607	113,593	110,987	2,607	4	54	257	315	46		0	500	-1300	-800	Match	
67	AG00720	100,488	1	93,153	93,153	93,153	15	70	525	610	102		0	300	-1800	-1500	Match	
68	AG00736	98,016	86,277	98,016	11,740	86,277	8	27	648	683	324		0	400	-400	0	Match	
69	AG00748	95,392	1	2,844	2,844	2,844	23	140	483	646	67		500	1500	-1700	-700	Match	
77	AG00969	63,375	57,984	63,375	5,392	57,984	31	146	1913	2090	0		0	-100	-500	-600	Match	
78	AG00982	61,535	8,475	61,535	53,061	8,475	34	125	851	1010	28		400	0	-100	-500	Match	
80	AG01041	54,462	50,162	54,462	4,301	50,162	85	463	893	1441	0		0	0	-800	-800	Not Match	
83	AG01089	49,444	7,840	42,652	34,813	42,652	43	154	1214	1411	252		0	-600	-900	-1500	Match	
85	AG01163	42,969	33,687	42,969	9,283	33,687	43	154	1261	1458	252		0	0	-900	-900	Match	
87	AG01217	38,388	5,520	38,388	32,869	5,520	22	22	72	116	130		900	600	-200	-500	Match	Yes
90	AG01352	27,107	1	7,164	7,164	7,164	28	64	1078	1170	75		0	1000	-1200	-200	Match	
91	AG01358	26,537	3,600	26,537	22,938	3,600	39	94	328	461	0		0	100	300	400	Match	
92	AG01368	26,030	11,852	26,030	14,179	11,852	16	45	703	764	210		100	700	-1100	-500	Match	
101	AG01786	6,512	1	4,200	4,200	4,200	3	58	224	285	95		0	0	-1000	-1000	Match	

1. PE pairs are the number of the paired-end read pairs that span the breakpoint. For each PE pair, one read maps upstream of the breakpoint while the other maps downstream (PE correspond to insert sizes of 320-520 bp; see Table S1a)
2. MP pairs are the number of the mate pairs that span the breakpoint. For each MP pair, one read maps upstream of the breakpoint while the other maps downstream (MP correspond to insert sizes of 5000 bp; see Table S1a)
3. CBR = chromosomal breakage region. From Bachmann-Waldmann, C., et al., Chromatin diminution leads to rapid evolutionary changes in the organization of the germ line genomes of the parasitic nematodes A. suum and P. univalens. Molecular and biochemical parasitology, 2004. 134(1): p. 53-64.
4. Offset is the difference in bp between the pair of breakpoints/telomere addition
5. Jex, A.R., et al., Nature, 2011. 479(7374): p. 529-33. Match, the Jex et al. assembly matches our germline assembly; Not Match, the Jex et al. assembly does not match our germline assembly; End of scaffold, the locus is at the end of a scaffold in the Jex. et al assembly.

Table S4 (related to Figure 4). Loss of *A. suum* duplicated, rearranged loci in somatic cells

ID	Scaffold Name	Scaffold Length	Elimination Start	Elimination End	Eliminated Length	DNA Change[1]	Sperm PE Pairs[2]	Whole Male PE Pairs[2]	Testis MP Pairs[3]	Total Germline PE/MP Pairs[2,3]	Total Somatic PE Pairs[2]	Total Germline PE/MP Pairs for Somatic Locus[2,3]	Organization in Jex et al. Assembly[4]	PCR confirmed
2	AG00006	1,173,925	1,144,875	1,173,925	29,051	1,144,875	0	4	235	239	43	813	Match	
3	AG00013	958,964	1	30,936	30,936	30,936	2	56	561	619	78	367	Match	Yes
4	AG00017	921,977	1	23,739	23,739	23,739	36	49	267	352	8	30	Match	
7	AG00034	778,791	247,090	299,339	52,250	299,339	9	1	401	411	0	0	Match	
13	AG00086	539,107	471,077	523,198	52,122	523,198	1	50	411	462	14	446	Not Match	
14	AG00102	491,296	489,600	491,296	1,697	489,600	0	2	46	48	121	324	Match	
23	AG00157	390,349	370,487	390,349	19,863	370,487	4	50	464	518	0	0	Match	
24	AG00172	372,876	362,290	372,876	10,587	362,290	1	4	124	129	67	235	Not Match	Yes
28	AG00191	357,075	1	5,986	5,986	5,986	15	658	1482	2155	80	611	Not Match	Yes
31	AG00197	350,630	291,420	350,630	59,211	291,420	0	5	413	418	61	408	Match	
32	AG00202	342,367	336,594	342,367	5,774	336,594	0	0	23	23	41	197	Not Match	
35	AG00237	306,207	1	10,790	10,790	10,790	0	3	678	681	0	0	Match	
36	AG00243	301,928	294,900	301,928	7,029	294,900	0	0	94	94	NA	NA	Match	
37	AG00246	300,757	4,545	136,187	131,643	4,545	31	85	336	452	102	47	Match	
38	AG00246	300,757	4,545	136,187	131,643	136,187	27	74	1169	1270	644	934	Not Match	Yes
39	AG00248	298,664	1	6,500	6,500	6,500	106	235	454	795	NA	NA	Match	
44	AG00301	263,654	1	18,358	18,358	18,358	23	109	759	891	0	0	Match	
49	AG00338	233,606	70,813	233,606	162,794	70,813	0	0	85	85	157	380	Not Match	
50	AG00441	177,407	174,700	177,407	2,708	174,700	4	30	13	47	34	482	Match	
53	AG00533	145,525	4,711	145,525	140,815	4,711	0	14	31	45	111	1431	Match	
54	AG00554	138,520	1	1,166	1,166	1,166	0	1	0	1	154	206	Match	
55	AG00580	132,901	1	112,604	112,604	112,604	8	78	573	659	37	776	Not Match	
59	AG00631	118,217	1	56,813	56,813	56,813	40	29	352	421	NA	NA	Match	
60	AG00631	118,217	60,493	118,217	57,725	60,493	27	97	350	474	NA	NA	Match	
62	AG00637	117,451	69,480	117,451	47,972	69,480	0	0	312	312	81	108	End of scaffold	Yes
63	AG00637	117,451	37,601	63,833	26,233	63,833	5	17	336	358	106	202	Match	Yes
65	AG00644	116,265	76,431	116,265	39,835	76,431	0	3	196	199	NA	NA	Match	
70	AG00765	92,147	88,507	92,147	3,641	88,507	47	104	347	498	521	450	Match	Yes
71	AG00826	82,593	1	3,973	3,973	3,973	12	47	34	93	278	419	End of scaffold	Yes
72	AG00872	76,200	1	12,358	12,358	12,358	3	29	1367	1399	51	379	Match	
73	AG00872	76,200	19,470	75,200	55,731	19,470	22	91	1312	1425	55	431	Match	
74	AG00872	76,200	19,470	75,200	55,731	75,200	30	65	38	133	83	474	Match	
75	AG00925	69,557	1	800	800	800	16	44	113	173	159	30	Match	
76	AG00937	68,171	63,271	68,171	4,901	63,271	63	163	753	979	205	1332	Match	
79	AG01041	54,462	1	42,393	42,393	42,393	16	37	840	893	11	383	Match	
81	AG01074	51,218	37,644	51,218	13,575	37,644	138	117	518	773	0	0	Match	
82	AG01089	49,444	7,840	42,652	34,813	7,840	16	75	461	552	35	1567	Not Match	
84	AG01140	44,869	1	2,590	2,590	2,590	55	27	360	442	190	4	Match	
86	AG01191	40,705	700	40,705	40,006	700	19	8	13	40	318	128	Match	Yes
88	AG01239	36,528	35,900	36,528	629	35,900	6	22	11	39	71	476	Match	
89	AG01349	27,350	20,664	27,350	6,687	20,664	0	28	32	60	0	123	Not Match	
93	AG01380	25,234	1	13,765	13,765	13,765	22	37	211	270	111	0	Match	
94	AG01381	25,210	1	17,552	17,552	17,552	6	0	739	745	29	88	Not Match	
95	AG01425	22,488	8,875	22,488	13,614	8,875	0	0	489	489	58	482	Match	
96	AG01441	20,997	19,297	20,997	1,701	19,297	10	58	169	237	78	30	Match	
97	AG01487	18,900	972	18,900	17,929	972	2	8	50	60	19	429	Match	
98	AG01606	12,328	1	11,453	11,453	11,453	1	6	17	24	63	88	Match	
99	AG01623	11,662	1	10,391	10,391	10,391	16	20	55	91	123	480	Match	Yes
100	AG01691	9,044	1	3,135	3,135	3,135	0	0	83	83	0	0	Not Match	
102	AG02005	3,085	1	1,960	1,960	1,960	7	28	0	35	224	142	Match	

1. The position in the scaffold where the germline and somatic sequencse diverge due to a presumed rearrangement (see Fig. S1b, part A)
2. PE pairs are the number of the paired-end read pairs that span the breakpoint. For each PE pair, one read maps upstream of the breakpoint while the other maps downstream (PE correspond to insert sizes of 320-520 bp; see Table S1a)
3. MP pairs are the number of the mate pairs that span the breakpoint. For each MP pair, one read maps upstream of the breakpoint while the other maps downstream (MP correspond to insert sizes of 5000 bp; see Table S1a)
4. Jex, A.R., et al., Nature, 2011. 479(7374): p. 529-33. Match, the Jex et al. assembly matches our germline assembly; Not Match, the Jex et al. assembly does not match our germline assembly; End of scaffold, the locus is at the end of a scaffold in the Jex. et al assembly.

Supplemental Experimental Procedures

DNA isolation and genomic library preparation

The tissues or a whole male Ascaris (Belgium) were ground to a fine powder in liquid nitrogen, digested overnight with 0.5% SDS and 150 µg/ml proteinase K in buffer (50 mM Tris-HCl, 100 mM EDTA, and 100 mM NaCl) and the DNA was isolated using phenol-chloroform extraction and ethanol precipitation. DNA from mixed populations of 32-64 cell embryos was isolated using CsCl gradients as described (Davis et al. 1988).

Genomic DNA libraries were constructed from *A. suum* germline and somatic tissues and sequenced (Table S1). Libraries were constructed using standard Illumina protocols and sequenced on the Illumina GAIIx or HiSeq platforms except as noted below. A library was constructed for Sanger sequencing using 32-64 cell embryos DNA fractioned to enrich for 6 kb fragments as previously described (Mitreva et al. 2011). A mate-pair library from testis DNA was prepared using 4.5 to 6.5 kb size-selected fragments (Hydroshear®) ligated to *LoxP*7 adapters and processed as described (Van Nieuwerburgh et al. 2012). A library was constructed from a whole male (Belgium) using a PCR-free protocol (Kozarewa et al. 2009).

Genome assembly assessment and comparison

A. suum cDNA and small RNA data were used to assess the functional coverage of the genomes (Table 1). RNA-seq data from *A. suum* developmental stages and tissues were generated, de-novo assembled, and updated with additional tissue samples and sequencing as described (Wang et al. 2011). The revised cDNA assembly contains 58,085 contigs and 58.1 Mb of sequences (see below). These were mapped to different genome assemblies using BLAT (v. 34) (Kent 2002), and the % of contigs with > 90% sequence mapped and the % of total cDNA bases mapped were calculated. The transcriptomes of small RNAs from *A. suum* germline and different developmental stages of embryos were also used to evaluate the genomes (Wang et al. 2011). For the small RNA mapping, both the genome and cDNAs that mapped to the genome were used as the database, because of the large number of endo-siRNAs in the *A. suum* small RNA populations (Wang et al. 2011).

From the base coverage analysis (see below), while we can clearly distinguish the eliminated sequences from retained ones, we observed low "background" read coverage in the eliminated regions for some somatic tissues (Fig. S1). To assess the purity of each genomic DNA source, we compared the base coverage ratio (R) in eliminated vs. retained regions for each genomic DNA. Using spermatids as the control ($R_{spermatids}$ = 100%), we found an average of 4.5% germline background for the 5 different somatic DNAs ($R_{male\_intestine}$ = 4.9%, $R_{male\_carcass}$ = 1.6%, $R_{female\_intestine}$ = 2.5%, $R_{female\_carcass}$ = 6.5% and $R_{32\text{-}64cell\_embryo}$ = 7.2%), and 85.1% germline for the other 2 germline DNAs (R_{male_testis} = 82.9% and R_{whole_male} = 87.3%). The germline background in the 32-64 cell embryo is expected since ~5% of the cells at this stage are germline. The reproductive tissues encompass the majority of the internal contents of male and female worms and removal of all of this material can be difficult. The germline reads in the somatic samples are likely small amounts of germline contamination.

A genome assembly for *A. suum* was recently published and described as largely a somatic genome (Jex et al. 2011). Our data indicate that this genome is assembled from mixed genomic DNAs (R_{Jex} = 38.7%, approximately 40% germline and 60% somatic). While the Jex et al. (Jex et al. 2011) assembly overall has larger scaffolds, the contig N50 for our assembly is larger (Table 1). In addition, we predict fewer, longer protein coding genes (Table 1), likely due to a reduction in gene fragmentation compared to the previous version (Jex et al. 2011). Finally, our genome libraries were prepared from genomic DNA without any amplification, compared to the whole genome amplification strategy (Jex et al. 2011).

Estimation of genome size and major repetitive sequence copy numbers

A k-mer based genome size estimation was initially performed using data from different genomic libraries. We observed a rather broad range of very high k-mer frequency instead of the peak that would be expected from a perfect dataset. We noted a dominant peak of k-mer singletons, representing sequencing errors, as is normally seen in such analyses. The broad range is likely due to the relatively patchy sequence coverage of the genomes (see Fig S1) as they were constructed from limiting amounts of genomic DNA from the spermatids and intestine of a single male and required extra PCR amplification

cycles during library production. We did not use the whole genome amplification approach to increase starting material, as employed by Jex et al (2011) to minimize the possibilities of mis-assembly. Thus a k-mer based genome size estimation is likely to be rather inaccurate, because of the uncertainty of where the k-mer peak is. We used an alternative method based on the sequence coverage of the scaffold that avoids the k-mer issue. In this method, the size of the *A. suum* haploid genomes (G) in bp is estimated as G = M/X, where M is the span of reads in bp that mapped to the genome assembly and X is the estimated base coverage for non-repetitive regions of the genome. Due to the extreme low proportion of repetitive sequences in the large, assembled scaffolds (Fig. S4), we used the median base coverage for scaffolds with length $\geq$ N50 (n = 260) as the estimated fold coverage. We used this approach to estimate the genome size based on two libraries made and sequenced in parallel from single male spermatids and intestine. Their genome sizes are estimated to be ~334.3 Mb and ~290.7 Mb, respectively (Table 1). This approach was also applied to other libraries estimating the genome size to be ~330.3 Mb for the male testis germline genome, ~338.9 Mb for the whole male (Belgium) which is largely germline, and ~292.2 Mb for the female intestine somatic genome.

To estimate the copy numbers of the 121-bp repetitive sequence in the germline and somatic genome, we mapped all germline or somatic reads to the repetitive unit (all variants included) and their total base pairs mapped (TB) were obtained. For each genome, the copy number (CN) is calculated as CN = TB/(L * X), where L is the length of the repeat unit and X is the estimated base coverage for single copy regions of the genome (same as above). The copy numbers for other notable repetitive elements were calculated the same way (Table S2).

PCR verification of the breakpoints, telomere addition, and other DNA alterations

The PCR strategies to confirm germline DNA loss and telomere addition are illustrated in Figs. 2-4. PCR primers used are shown below.

ID	Primer S1	Primer S2	Primer G1	Primer G2	Primer St	Primer S3
1	CCAAACAATGTTTTTGAAACCA	GGTTAATTTCAGCCATTCAAC	TCAGTTTTCCACAGATCTCCC	GCTCGTCATCATATCAAGCAC	CCTAAGCCTAAGCCTAAGCCTACTG	NA
6	CGGATTCTGATAGTATAAGAGGC	CATTTAAACCAACGATGGGAT	AGACGGCACCCGCTAAATG	CAAAAGTTCACGTTCATGCTTC	AAGCCTAAGCCTAAAACTGGATTCG	NA
9	GCGGCCACTAACACGAATG	TCAACGATCATATCAAGTGGTC	CGATTGAAAGCGAACTCAAC	CAATCATGCCAGATGCTGT	CCTAAGCCTAAGCCTAAGCCTGTG	NA
15	TGACGTCGATCCTCACTGAC	TGGCCTTCAGTTTGCAGTC	ACTACAATTTTCGTGGCTGGC	CCCTGGCGAGGAGAAGAAGTTC	CCTAAGCCTAAGCCTAAGGCGTCCGT	NA
17	CAAGGTGACAAGTCCTGTGA	GTTCGAATATGTCGCAAAGA	GTACCAATTGTAGCGAGATG	CCAGGAGGTGAAACTTACTC	AAGCCTAAGCCTAAGCCTAATTGCATTTAC	NA
18	TTACTCCATAATTTCAGAGCC	ATCTGTGTTTCGTCGCCTTAC	CTGAAGAACACTGAAGTACCC	AGCCACACACTTCAGACGGTG	NA	NA
27	ATTGCAACTGAAAGGACTGTG	AAGTGCTACCCATTTGGTTCG	TGGACGGGCTTTATAGAAGG	TTGACGACGACGATCCAAAA	AAGCCTAAGCCTAAAGCCCCTA	NA
87	GATGCGCAAATAAAACTGGT	CATGTGGTTCAAGGGAAAAT	CCTAAAACAATAGCAGGCG	AAAGGCTCTGCTGCAATATGG	CCTAAGCCTAAGCCTAAGCCTAAAAC	NA
3	GGTTGGGGAAGGGTTAAGAG	GATCGGGAAAGCAATTCAAA	GCACTGAGCCTTGATGATGA	ATTCTTGTCTTCCGGGAACC	NA	TGTGCATTTCGATTTCGGTA
24	CGTGCAGATTACTGACTTCG	CTGGCGAATCTCTTGGTATAG	TCAGTCCTCTTTCGGACATT	GACCTTACCCACTTCACCTC	NA	GCTGGTTCGCCTTGAATA
28	AAGGTGGGATAAGTGCAAAAG	TTGAGCCATAACATGGAAGAGA	GGCGTATCACCATTTTCGTT	CTGTCAGCCGGAACGTTTAC	NA	GGTACCATGCAACCCAATTT
38	CTGAACAGCCCAAGGGAATA	GTTGCAATGAGGGAACATCC	ACTGTCCCGCTGTTTCTGTT	GGTTTTTCGAGCATGTGGAT	NA	GAATGGTGGGGACTGAGCTA
62	TGCAAAGACTTCCGTTCTCA	TTTGCAGTACGTTGCAAAAAG	CTGCGGAAATTGAAAATGGT	GACGAAAGGTGAGCCACAAT	NA	CAAGCGTCCCTTCTCAAATC
63	TTCGAAGTTGCATGTTTTTGGT	CGTCGACCCCAAGTCAACATA	CTTGGCATAATCAACACATACA	CGTTTCTCAGAACGTCTTCG	NA	GTTCCTAGCCAGTCTTCCCAAGTC
70	GAGCTGTATGAGGAGCCATT	TGGAACAATCCAGTGATGTC	GACAGCGGAATTCTACAAGG	GACGTCGTACGTTCCTTCA	NA	GTCGCTGCAACTCTTAATTG
71	CGACCCCGTTCATAATGAGT	ATAAAATGCCGAAACGATGC	CTTTTCCGAGCGTGTTATCC	CACATGACTCATCCGCAAAC	NA	AATTTGTTACGACGCACACG
86	GGTTTTGCTGCTGGAAATGT	TCGGGTATCTGGTGAAAAATG	GGCATCCGATTGTTTCTGTT	AGCGGCAAAACTGCTCTAAA	NA	AATCGGTCACTTTTCGATGG
99	GCAGCACAATACGCTTTCAC	GGTAGCTGCGACGATTTGTT	GGGTAATTGGGTTCGAGTCA	TCTCTCCTCGCTGTGCCTAT	NA	GAAAAGGGGCAAGAACCAAT

Multiple genome comparisons and breakpoint heterogeneity analysis

We mapped genomic reads from different sources (Table S1) to the germline assembly by using either bowtie 2 (Langmead and Salzberg 2012) (for Illumina reads) or bwa (v0.6.1-r104) (Li and Durbin 2010) (for 700 bp Sanger capillary reads). The base coverage uniformity for different tissues was visualized in figures (see Fig. S1 for examples) generated by in-house Perl scripts using the GD module (v2.46) and in a genome browser that uses GMOD/gbrowse (http://www.gbrowse.org/index.html) set of tools. An *A. suum* genome browser with all data can be accessed at http://ascaris.nematodegenomes.org/. The position heterogeneity for the 52 breakpoints with telomere addition was determined to at least 100-bp resolution. Pairwise comparisons of the breakpoint position in different tissues were conducted and bp offset distance of the breakpoints calculated (Fig. 3e).

Motif search around breakpoints

We used the MEME suite (Bailey et al. 2009) to search for potential motifs surrounding the breakpoints with telomere additions. For all 52 loci, we extracted sequences around the breakpoints flanking 100, 200,

300, 400, 500, 1,000, 2,000, 3,000, 4,000 and 5,000 bp. For each set of sequences, we searched for motifs using MEME, with "-minsites 30 -minw 6 -maxw 50" parameters (requiring that at least 30 sites contain the motif, with the minimal motif size 6 bp and the maximum size 50 bp). We also used GLAM2 (Frith et al. 2008) using the default parameters to search for gapped motif sequences. None of the analyses indicate a consistent DNA motif around the breakpoints (Fig. S2). In addition, we observed no consistent association of GC%, repetitive sequence, or other DNA characteristics with the breakpoints. Breakpoints occur in both intergenic regions and within genes.

Gene annotation

For all the predicted genes, NCBI blast (blastx v2.2.23) (Altschul et al. 1997) was used to search against protein databases from nr (e-value cutoff: 1e-10), Swiss-prot (e-value cutoff: 1e-8) and *C. elegans* (e-value cutoff: 1e-5). Blast2go (v2.50) (Gotz et al. 2008) was used to functionally categorize genes (db: go_201204-assocdb-data), and to carry out GO enrichment analysis (Fisher's Test) using eliminated *A. suum* genes as the test-set and all *A. suum* genes as the reference-set (Table S6). Entrez IDs for *C. elegans* genes that are orthologous to *A. suum* genes were used for the IPA analysis (Ingenuity® Systems, www.ingenuity.com) (Table S6). *A. suum* gene families (paralogs) were identified by using tblastx (Altschul et al. 1997) search for all *A. suum* genes against themselves (e-value cutoff: 1e-8), and genes with $\geq$ than 50% match over their entire length were grouped into families.

Supplemental References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389-3402.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37: W202-208.

Davis RE, Davis AH, Carroll SM, Rajkovic A, Rottman FM. 1988. Tandemly repeated exons encode 81-base repeats in multiple, developmentally regulated Schistosoma mansoni transcripts. *Mol Cell Biol* 8: 4745-4755.

Frith MC, Saunders NF, Kobe B, Bailey TL. 2008. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol* 4: e1000071.

Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* 36: 3420-3435.

Jex AR, Liu S, Li B, Young ND, Hall RS, Li Y, Yang L, Zeng N, Xu X, Xiong Z et al. 2011. Ascaris suum draft genome. *Nature* 479: 529-533.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* 12: 656-664.

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature methods* 6: 291-295.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* 9: 357-359.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26: 589-595.

Mitreva M, Jasmer DP, Zarlenga DS, Wang Z, Abubucker S, Martin J, Taylor CM, Yin Y, Fulton L, Minx P et al. 2011. The draft genome of the parasitic nematode Trichinella spiralis. *Nat Genet* 43: 228-235.

Van Nieuwerburgh F, Thompson RC, Ledesma J, Deforce D, Gaasterland T, Ordoukhanian P, Head SR. 2012. Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. *Nucleic Acids Res* 40: e24.

Wang J, Czech B, Crunk A, Wallace A, Mitreva M, Hannon GJ, Davis RE. 2011. Deep small RNA sequencing from the nematode Ascaris reveals conservation, functional diversification, and novel developmental profiles. *Genome Res* 21: 1462-1477.