Table S5. Newbler assembler results for the ten independent runs using randomly selected reads from the total of 13 pyrosequencing runs for DNA extracted from soil from the Park Grass experiment plot 3d at Rothamsted Research. Newbler parameters: Expected depth: 0 (i.e. undefined); Minimum read length: 20; Seed step : 12; Seed length : 16; Seed Count : 1; Minimum overlap length : 40; Minimum overlap identity : 90%; Alignment identity score: 2; Alignment difference score: -3.

Assembler runs:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
runMetrics										
totalNumberOfReads	1257242	2514489	3771729	5028954	6286181	7543423	8800651	10057895	11315094	12572342
totalNumberOfBases	487554794	974670636	1461941647	1949323461	2436671088	2.924E+09	3411376290	3898801814	4386462616	4874169257
Reads Status										
numAlignedReads	57556	185373	352696	554010	788063	1048430	1330404	1634721	1959374	2303182
numAlignedBases	18266373	59672888	114246594	180909939	259220082	346814643	442802556	547016473	658672541	777402244
inferredReadError	5.06%	5.35%	5.43%	5.42%	5.36%	5.31%	5.24%	5.17%	5.11%	5.05%
numberAssembled	31175	104814	200774	319201	459853	620159	798751	994756	1206256	1432866
numberPartial	26381	80500	151773	234651	328066	427926	531532	639829	752966	870034
numberSingleton	1163529	2236204	3260115	4244122	5192433	6112877	7011630	7884093	8738366	9572485
numberRepeat	21	103	268	408	763	1146	1548	1968	2487	3074
numberOutlier	30890	82472	143156	209757	278972	350046	420777	495652	568277	642021
numberTooShort	5246	10396	15643	20815	26094	31269	36413	41597	46742	51862
Total	1257242	2514489	3771729	5028954	6286181	7543423	8800651	10057895	11315094	12572342
largeContigMetrics										
numberOfContigs	3702	9138	17233	27641	40104	53850	68470	84705	101154	118800
numberOfBases	3757449	9786251	17197015	26537742	37785216	50417029	64287965	79618149	95770632	113089732
avgContigSize	1014	1070	997	960	942	936	938	939	946	951
N50ContigSize	1076	1094	998	964	951	949	954	956	965	971
largestContigSize	6361	15448	22645	21204	14380	15875	13912	15752	15426	15425
Q40PlusBases	3208918	8414019	14785259	22817145	32526441	43456627	55545708	68928952	83085358	98268346
Q39MinusBases	548531	1372232	2411756	3720597	5258775	6960402	8742257	10689197	12685274	14821386

Table S5. Newbler assembler results for the ten independent runs using randomly selected reads from the total of 13 pyrosequencing runs for DNA extracted from soil from the Park Grass experiment plot 3d at Rothamsted Research. Newbler parameters: Expected depth: 0 (i.e. undefined); Minimum read length: 20; Seed step : 12; Seed length : 16; Seed Count : 1; Minimum overlap length : 40; Minimum overlap identity : 90%; Alignment identity score: 2; Alignment difference score: -3.

Assembler runs:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
allContigMetrics										
numberOfContigs	7478	21418	41388	65598	94069	124842	157780	192206	228651	266600
numberOfBases	4738744	12993575	23467908	36508692	51965915	69162343	87893864	108053114	129548168	152337226
alignmentDepths										
1	254115	562821	1047133	1689799	2431319	3210275	4011784	4837438	5731987	6632746
2	1203570	3072382	5787534	9165657	13092802	17334645	21793710	26405044	31238812	36348350
3-4	2093933	4747497	8409968	13474365	19616431	26379378	33651362	41544235	49746108	58339992
5-6	1110246	2960407	4631690	6921263	9876485	13396813	17371792	21667238	26454007	31525166
7-8	332199	1556836	2402579	3184272	4289997	5675685	7371741	9243575	11315472	13600152
9-10	94034	788301	1541061	1954878	2376906	2973868	3721430	4618985	5621952	6710813
11-13	36453	424071	1292575	1901405	2252372	2648259	3078596	3676553	4324585	5059861
14-16	10742	116673	520348	1113087	1457370	1641430	1849416	2092694	2379250	2673058
17-19	5515	40973	174771	542959	932733	1128344	1237370	1340651	1458570	1609141
20-22	2440	15876	63147	213932	537004	812516	931154	964149	988136	1061923
23-25	1773	6433	30899	84046	248247	499440	695566	770823	781784	791998
26-28	1125	4689	16783	39095	108385	257508	463531	605929	674295	663610
29-31	893	3213	7901	21079	46572	120284	254402	431273	544012	592207
32-34	471	2465	4487	12294	27656	56103	127353	254512	396612	484982
35-38	291	1395	3732	9901	19892	41132	78045	164154	303232	460025
39-42	200	1135	2169	4596	11067	21353	36487	68246	143552	263660
43-46	50	1150	1570	3449	5774	13134	21329	36671	65517	123733
47-50	132	451	1102	1567	3299	7786	15037	21567	31831	58094
51-55	304	367	942	1408	2421	4825	11218	16476	25135	37157
56-60	90	323	517	1117	1235	3105	5834	11695	17448	24615
61-70	10	368	486	1510	2142	3085	4382	10797	20884	28372
71-80	6	383	282	722	1350	1875	2397	4266	8248	13665

Table S5. Newbler assembler results for the ten independent runs using randomly selected reads from the total of 13 pyrosequencing runs for DNA extracted from soil from the Park Grass experiment plot 3d at Rothamsted Research. Newbler parameters: Expected depth: 0 (i.e. undefined); Minimum read length: 20; Seed step : 12; Seed length : 16; Seed Count : 1; Minimum overlap length : 40; Minimum overlap identity : 90%; Alignment identity score: 2; Alignment difference score: -3.

Assembler runs:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
91-100	0	102	191	245	406	670	1227	1460	1770	2513
101-140	0	15	942	580	508	1155	2609	4075	3786	5084
141-180	0	0	0	1068	699	313	291	892	1964	2207
181-240	0	0	0	41	676	814	406	463	500	921
241-300	0	0	0	0	12	500	981	606	469	400
301-400	0	0	4	0	0	3	33	652	1063	1178
401-500	0	0	20	0	0	7	1	1	15	84
501-600	8	0	5	0	1	5	0	0	12	4
601-700	0	0	1	0	0	1	0	0	1	2
701-850	0	0	0	0	0	3	0	0	5	10
851-1000	0	0	0	0	0	9	0	0	0	14
1001+	0	20	23	26	13	23	29	16	26	38

Read Status – status of the read in the assembly, which can be one of the following:

a. Assembled – the read is fully incorporated into the assembly

b. PartiallyAssembled – only part of the read was included in the assembly, the rest was deemed to have diverged sufficiently to not be included

c. Singleton – the read did not overlap with any other reads in the input

d. Repeat – the read was either:

i. Inferred to be repetitive early in the assembly process. A read can be inferred to be repetitive if >70% of the read's seeds hit to at least 70 other reads. Such reads are excluded from the assembly.

ii. Determined to partially overlap a contig. The portions of such reads that overlap unique contigs are still included in the assembly results.

e. Outlier – the read was identified by the GS De Novo Assembler as problematic, and was excluded from the final contigs (one explanation of these outliers are chimeric sequences, but sequences may be identified as outliers simply as an assembler artifact)

f. TooShort – the trimmed read was too short to be used in the computation (shorter than 50 bases and longer than the value of the minlen parameter, unless 454 Paired End Reads are included in the dataset, in which case, all reads at least “minlen” bases are used).