corsage: Metagenome-enabled error correction of single cell sequencing reads

Andreas Bremges [1]*, Esther Singer [2], Tanja Woyke [2] and Alexander Sczyrba [1]

[1]Center for Biotechnology & Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany
[2]U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

*to whom correspondence should be addressed

1 SINGLE AMPLIFIED GENOMES

*Descibe why and how the SAGs were generated. Cite (Clingenpeel et al., 2014). State average sequencing depth (constant), but different MDA quality levels. Highlight two exceptionally well amplified SAGs (6 & 7), that are closer to isolate-grade genomes. Also highlight the problematic ones mentioned in main text, where the hybrid error correction clearly outperforms SAG-only correction. Table **S1**.*

Table S1. Per-base coverage for the eight *E. coli* SAGs.

SAG	mean	std.dev	min	max	Q1	Q2	Q3
0	289.749	486.983	0	6560	36	112	331
1	297.379	386.285	0	4774	70	169	355
2	272.085	708.577	0	12254	18	57	188
3	260.303	762.898	0	16120	11	49	198
4	284.087	783.65	0	13996	7	43	174
6	266.821	275.48	3	3861	105	179	321
7	278.141	241.48	4	2124	124	195	334
8	299.661	399.265	0	4816	58	168	384

Describe methods: mapping of raw reads with bwa mem (Li, 2013) then samtools to view, sort and depth (Li et al., 2009) options of depth: -q0 -Q0 (with source code modified to remove the hardcoded coverage max). bla bla

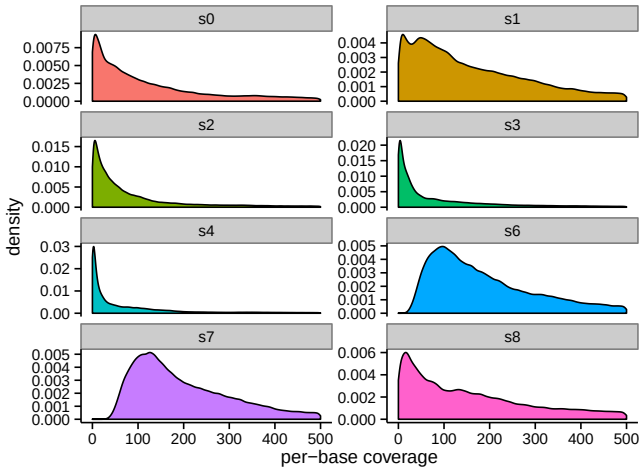

Fig. 1: Per-base coverage for the eight *E. coli* SAGs.

2 MOCK METAGENOME

*Generated at the JGI, initially for internal benchmarking. NCBI accession number or some way to get it via IMG or GOLD (alternatively, might be okay to provide upon request, contact Tanja?). Describe what's in there. Refer to table with molecular weights and mapping statistics (genome coverage), Table **S2**.*

Probably also include a tree containing all 26 genomes, based on their 16S (or – simpler – extracted from iTOL, as all are known!). This might be of interest to see how closely related the stuff in there is. Discuss with others!

3 METRICS PER SAG

*Read-based in Table **S3**, assembly-based in Tables **S4** and **S5**.*

Explain the parameter sweep for -c, the metagenomic coverage threshold. Setting this to 2 already produces much better results, with a higher threshold yielding slightly better results, but more sample-specific. Setting it to 2 should work in all cases, and thus is the default setting. You might get better results if you play around with it, depends on target coverage in the metagenome.

Assembly parameters (k) were tuned to maximize the NG50 value of the BayesHammer + SPAdes assembly. Changing from the default -k 21,33,55 to -k 21,33,55,77 produced better assemblies, thus we picked this combination. --careful reduces the number of misassemblies by (1) less aggressive assemby and (2) polishing of the contigs in a postprocessing step involving read mapping.

REFERENCES

Clingenpeel, S. et al. (2014). Reconstructing each cell's genome within complex microbial communities–dream or reality? *Front Microbiol*, 5:771.

Gurevich, A. et al. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.

Letunic, I. and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–128.

Letunic, I. and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.*, 39(Web Server issue):W475–478.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997*.

Li, H. et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

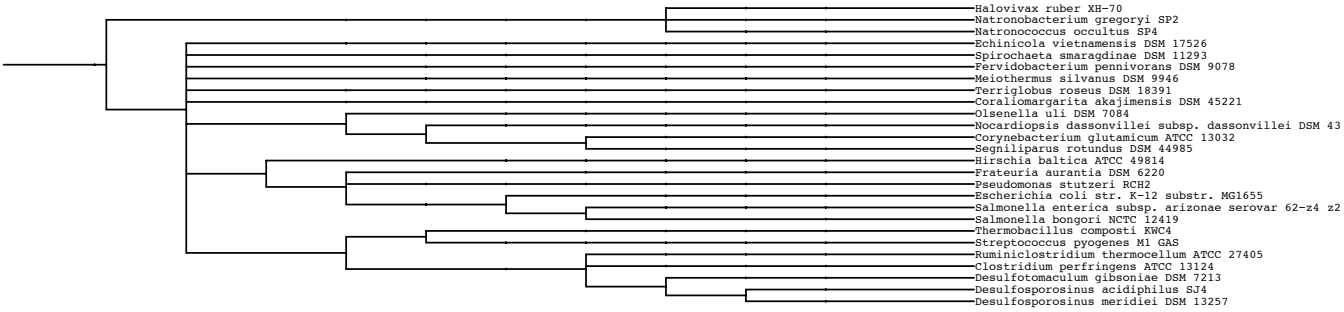

Fig. 2: phyloT/iTOL (Letunic and Bork, 2007, 2011) phylogenetic tree of the 26 members of the mock metagenome.

Table S2. Profile of the mock metagenome.

NCBI Taxonomy ID	Scientific name	ContigLength	MappedReads	AvgCoverage
646529	Desulfosporosinus acidophilus SJ4 : Contig240.3	3897	367804	13410.8450089813
	Desulfosporosinus acidophilus SJ4 : Contig241.2	60447	2309615	5644.16513640048
	Desulfosporosinus acidophilus SJ4 : Contig243.1	4926837	50525496	1520.41947886646
767817	Desulfotomaculum gibsoniae DSM 7213	4855529	24329020	741.835021889479
768704	Desulfosporosinus meridiei DSM 13257	4873567	16066750	487.573607585573
926556	Echinicola vietnamensis DSM 17526	5608040	2160987	57.2232760465332
771875	Fervidobacterium pennivorans DSM 9078	2166381	39566833	2708.24050709455
767434	Frateuria aurantia DSM 6220	3603458	13922996	568.15908080516
797302	Halovivax ruber XH-70	3223876	6060380	274.952728640928
511145	Escherichia coli str. K-12 substr. MG1655	4639675	647555	20.6521661538793
195103	Clostridium perfringens ATCC 13124	3256683	1461840	66.5510195496461
203119	Clostridium thermocellum ATCC 27405	3843301	1542460	59.6166532363715
582402	Hirschia baltica ATCC 49814	3455622	27502745	1182.5057080896
	Hirschia baltica ATCC 49814 plasmid pHbal01	84492	641481	1126.13552762392
583355	Coraliomargarita akajimensis DSM 45221	3750771	11810956	467.028614916773
640132	Segniliparus rotundus DSM 44985	3157527	4886507	225.630138079579
446468	Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111	5767958	5761	0.0573058957780206
	Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111 plasmid pNDAS01	775354	879	0.0682900971685192
526227	Meiothermus silvanus DSM 9946	3249394	27813938	1261.46624232088
	Meiothermus silvanus DSM 9946 plasmid pMESIL01	347854	3461768	1466.53106188228
	Meiothermus silvanus DSM 9946 plasmid pMESIL02	124421	955914	1123.36214947637
633147	Olsenella uli DSM 7084	2051896	7839526	552.985752201866
573413	Spirochaeta smaragdinae DSM 11293	4653970	39431130	1255.9706035922
797304	Natronobacterium gregoryi SP2	3788356	8676937	335.230170554193
694430	Natronococcus occultus SP4 : Contig265	12939	243363	2564.16840559549
	Natronococcus occultus SP4 : Contig266	287963	819298	396.405625028215
	Natronococcus occultus SP4 : Contig267	4013216	11443075	397.25378947956
644801	Pseudomonas stutzeri RCH2 : Contig40.4	2804	62801	3135.38908701855
	Pseudomonas stutzeri RCH2 : Contig44.3	9865	4321	64.2942726811961
	Pseudomonas stutzeri RCH2 : Contig45.2	12763	54681	630.421844393951
	Pseudomonas stutzeri RCH2 : Contig47.1	4575057	5326365	171.651879965649
926566	Terriglobus roseus DSM 18391	5227858	8464573	239.329327805002
717605	Thermobacillus composti KWC4 : Contig54.2	149182	1912248	1900.01226689547
	Thermobacillus composti KWC4 : Contig56.1	4206343	29042078	1016.11924491179
160490	Streptococcus pyogenes M1 GAS	1852441	1502208	120.443499145182
882884	Salmonella enterica subsp. arizonae serovar 62:z4,z23:- str. RSK2980	4600800	1831145	58.7192305685968
218493	Salmonella bongori NCTC 12419	4460105	501312	16.60908140055
196627	Corynebacterium glutamicum ATCC 13032	3309401	1063668	47.6651768703762

Describe methods: mapping of raw reads with bwa mem (Li, 2013) then samtools (Li et al., 2009), input from Esther is needed. Total of 355875608 reads (53381341200bp) sequenced.

Table S3. Performance of error correction

Metric	Program	0	1	2	3	4	6	7	8
Reads	–	9365134	9604918	8811278	8396488	9257066	8609900	8990744	9682468
Perfect	raw	2120932	2179609	1937541	1954244	1872800	2049675	2063874	2183454
	hammer	7656274	8260510	6302861	5970186	6297298	7639715	8068555	8317006
	corsage -c1	8276184	8578469	7861670	7433499	8192894	7736330	8046250	8656282
	corsage -c2	8886436	9188854	8440502	7965810	8829995	8264229	8611867	9272559
	corsage -c3	8965114	9270617	8511335	**8024365**	8904161	8330927	8683349	9347321
	corsage -c4	**8974388**	**9278745**	**8511843**	8023150	**8910290**	**8338543**	**8689922**	**9354172**
	corsage -c5	8943102	9243789	8471748	7993790	8875735	8310857	8654540	9312765
Chimeric	raw	69568	67625	81938	75509	79443	52246	44983	59265
	hammer	72820	70590	87564	80336	84813	53875	46387	61948
	corsage -c1	15129	13376	18281	12859	13501	10913	9954	12303
	corsage -c2	5502	4593	10397	4648	5257	3941	3824	4889
	corsage -c3	**5224**	**4245**	**10181**	**4349**	4802	**3642**	**3555**	**4614**
	corsage -c4	5239	4247	10230	4881	**4777**	3656	3558	4646
	corsage -c5	5318	4350	10321	4703	4896	3698	3611	4731
Better	raw	–	–	–	–	–	–	–	–
	hammer	6743983	7026478	6156387	5669978	6574627	6244274	6645542	7095951
	corsage -c1	6899557	7131162	6610328	6169861	7102763	6311717	6653538	7200704
	corsage -c2	7008163	7236195	6707136	**6260408**	7206731	6403639	6749255	7306961
	corsage -c3	**7017190**	**7247251**	**6715402**	6258028	**7215037**	**6412774**	**6759006**	**7315526**
	corsage -c4	7011329	7242576	6709552	6238573	7211254	6408499	6753669	7310524
	corsage -c5	6987270	7218080	6681337	6211648	7184655	6388594	6729406	7280437
Worse	raw	–	–	–	–	–	–	–	–
	hammer	31644	26951	32315	29096	41584	25270	26959	28960
	corsage -c1	76694	76914	65596	75134	75213	62950	65427	73162
	corsage -c2	25990	25304	20560	27335	27390	20791	21074	24001
	corsage -c3	21353	20867	17696	**23556**	23768	17299	17238	20241
	corsage -c4	**20967**	**20559**	17040	24583	**23721**	**17037**	**16932**	**20007**
	corsage -c5	21386	21512	**16922**	25583	24349	17529	17774	20919
Time	hammer								
	corsage								
RAM	hammer								
	corsage								

A read is said to become *better* (or *worse*) if the best alignment of the corrected sequence has more (or fewer) identical bases to the reference genome than the best alignment of the original sequence. The table gives the number of reads mapped *perfectly*, number of *chimeric* reads (i.e. reads with parts mapped to different places), number of corrected reads becoming *better* and the number of corrected reads becoming *worse* than the original reads. Mapping of raw and corrected reads to reference with BWA-MEM-0.7.12 (Li, 2013), postprocessing with samtools-1.2 (Li et al., 2009) and some custom scripts, all available on GitHub eventually.

Table S4. IDBA-UD assemby statistics

Metric	Program	0	1	2	3	4	6	7	8
NG50	raw	45284	53863	31250	24834	17196	87102	80574	59754
	hammer	40924	51004	29256	24023	17246	**95532**	**87102**	44292
	corsage	**59081**	**73496**	**54946**	**41749**	**31569**	90184	80997	**80997**
# contigs	raw	330	342	514	607	654	201	200	303
	hammer	335	345	530	626	653	**191**	**194**	328
	corsage	**277**	**272**	**409**	**497**	**512**	200	198	**249**
Largest contig	raw	227106	203026	**203098**	141383	102074	**221687**	**232585**	162612
	hammer	203098	157125	197417	141494	107872	**221687**	178322	139398
	corsage	**236473**	**203098**	144213	**141579**	**124628**	221683	221683	**236473**
Total length	raw	4400079	4587934	4400469	4139052	3940625	4640153	**4639167**	4533515
	hammer	4402128	**4591089**	4400334	4144742	3934141	**4641409**	4638005	**4538546**
	corsage	**4408926**	4590112	**4421966**	**4171137**	**3972787**	4639453	4636457	4538141
# misassemblies	raw	16	11	15	32	36	**0**	**0**	10
	hammer	17	11	9	37	37	**0**	**0**	8
	corsage	**9**	**4**	**2**	**13**	**20**	**0**	**0**	**7**
# mismatches per 100 kbp	raw	**4.17**	**3.64**	10.16	17.02	20.41	0.31	0.13	3.16
	hammer	4.88	3.88	12.50	16.35	20.76	**0.15**	**0.11**	**2.90**
	corsage	4.38	3.80	**7.93**	**8.86**	**12.82**	2.30	2.39	3.27
# indels per 100 kbp	raw	0.32	**0.24**	0.69	1.47	1.09	0.13	**0.09**	0.36
	hammer	0.32	0.29	0.79	1.52	1.51	0.11	**0.09**	0.34
	corsage	**0.25**	0.31	**0.53**	**0.90**	**0.85**	**0.09**	**0.09**	**0.18**
Genome fraction (%)	raw	93.656	97.182	93.344	87.892	83.101	98.266	98.245	96.104
	hammer	93.631	97.165	93.304	87.924	82.958	98.211	98.175	96.028
	corsage	**93.928**	**97.431**	**93.988**	**88.827**	**84.053**	**98.271**	**98.252**	**96.265**
# genes	raw	3873	4049	3741	3477	3220	4186	4191	4027
	hammer	3859	4052	3709	3469	3208	4194	4199	4005
	corsage	**3931**	**4126**	**3857**	**3593**	**3347**	**4204**	**4202**	**4088**
NGA50	raw	44890	53863	31250	24224	16220	87102	80574	59589
	hammer	40924	51004	29207	24023	17227	**95444**	**87102**	44292
	corsage	**59081**	**73496**	**54946**	**41749**	**31567**	90184	80997	**80574**
Time	raw								
	hammer								
	corsage								
RAM	raw								
	hammer								
	corsage								

IDBA-UD. Explain metric definitions as outlined by QUAST. State, that everything was computed with QUAST (Gurevich et al., 2013), default settings, contigs greater than 500bp, bla bla. Availability in GitHub.

Table S5. SPAdes assemby statistics

Metric	Program	0	1	2	3	4	6	7	8
NG50	raw	65444	95218	66287	48903	26823	114661	117715	86966
	hammer	**86625**	95218	67436	52817	31448	**120770**	**132608**	105995
	corsage	**86625**	**95517**	**72055**	**53947**	**36236**	112350	**132608**	**112853**
# contigs	raw	447	400	606	718	813	245	233	324
	hammer	302	**275**	474	594	676	**198**	**185**	**250**
	corsage	**288**	279	**418**	**534**	**569**	210	213	263
Largest contig	raw	203603	**224667**	**218793**	**178300**	113773	269308	268816	223154
	hammer	**204882**	203257	**218793**	167410	135551	**312119**	269348	**269318**
	corsage	203394	224320	**218793**	178231	**155221**	268535	**312008**	268327
Total length	raw	**4522153**	**4703061**	**4533214**	**4290463**	**4138591**	**4713277**	**4718163**	**4633297**
	hammer	4443696	4633876	4464269	4233011	4046113	4686582	4689565	4584513
	corsage	4452849	4645907	4471868	4240428	4065827	4698390	4702810	4600454
# misassemblies	raw	15	2	22	45	51	**1**	3	12
	hammer	11	7	19	31	38	**1**	2	7
	corsage	**6**	**3**	**10**	**23**	**22**	**1**	**0**	**6**
# mismatches per 100 kbp	raw	15.30	11.57	34.70	48.21	50.53	2.84	**2.14**	9.72
	hammer	12.70	10.30	30.34	40.41	48.42	**1.27**	2.17	**7.66**
	corsage	**10.41**	**9.32**	**22.69**	**30.86**	**36.47**	5.66	5.21	8.43
# indels per 100 kbp	raw	0.89	1.17	2.26	4.48	4.24	0.31	**0.22**	1.00
	hammer	1.17	1.19	3.16	3.48	4.58	**0.24**	0.35	0.94
	corsage	**0.64**	**0.95**	**2.24**	**3.30**	**3.28**	0.55	0.31	**0.83**
Genome fraction (%)	raw	**94.241**	**97.948**	**94.239**	89.098	84.372	**98.665**	**98.708**	**96.702**
	hammer	94.050	97.527	93.984	89.133	84.220	98.505	98.446	96.459
	corsage	94.223	97.629	94.223	**89.543**	**84.798**	98.580	98.541	96.603
# genes	raw	3876	4117	3782	3532	3281	4217	**4224**	4081
	hammer	3898	4124	3805	3562	3300	4211	4219	4093
	corsage	**3937**	**4133**	**3866**	**3608**	**3390**	**4218**	4220	**4097**
NGA50	raw	65444	95218	66287	41757	26823	114661	117715	86966
	hammer	82361	95218	67436	51035	31446	**120770**	**132608**	99558
	corsage	**86625**	**95435**	**71474**	**52883**	**36219**	112350	**132608**	**105926**
Time	raw								
	hammer								
	corsage								
RAM	raw								
	hammer								
	corsage								

SPAdes --sc --only-assembler -k 21,33,55,77 --careful. Explain metric definitions as outlined by QUAST. State, that everything was computed with QUAST, cite it, default settings, contigs greater than 500bp, bla bla. Availability in GitHub.