Recovery of high-quality assembled genomes via metagenome binning guided with single-cell amplified genomes

Koji Arikawa[1], Keigo Ide[2,3], Masato Kogawa[2], Tatusya Saeki[1], Takuya Yoda[1], Taruho Endoh[1], Ayumi Matsuhashi[1], Haruko Takeyama[2,3,4,5], Masahito Hosokawa[1,4*]

1. bitBiome, Inc., 513 Wasedatsurumaki-cho, Shinjuku-ku, Tokyo 162-0041, Japan.
2. Department of Life Science and Medical Bioscience, Waseda University, 2-2 Wakamatsucho, Shinjuku-ku, Tokyo 162-8480, Japan.
3. Computational Bio Big-Data Open Innovation Laboratory, National Institute of Advanced Industrial Science and Technology, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan.
4. Research Organization for Nano and Life Innovation, Waseda University, 513 Wasedatsurumaki-cho, Shinjuku-ku, Tokyo 162-0041, Japan.
5. Institute for Advanced Research of Biosystem Dynamics, Waseda Research Institute for Science and Engineering, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan.

*Corresponding author: Masahito Hosokawa (masahito.hosokawa@bitbiome.co.jp)

Abstract

Background:

Obtaining high-quality (HQ) reference genomes from microbial communities is crucial for understanding the phylogeny and function of uncultured microbes in complex microbial ecosystems. Despite the improved bioinformatic approaches to generate curated metagenome-assembled genomes (MAGs), existing metagenomic binners often fail to obtain reliable MAGs, and thus, they are nowhere comparable to genomes sequenced from isolates in terms of strain level resolution. Here, we present a single-cell genome-guided metagenome binning (MetaSAG) to reconstruct the strain-resolved genomes from microbial communities at once.

Results:

MetaSAG employs single-cell amplified genomes (SAGs) generated with microfluidic technology as binning guides to recover improved draft genomes with the metagenomic data. To assess the performance of reconstructing genomes from various microbial communities, we compared MetaSAG with four conventional metagenomic binners using a cell mock community, human gut microbiota, and skin microbiota samples. MetaSAG showed precise contig binning and higher recovery rates (>97%) of rRNA and plasmids compared to conventional binners in genome reconstruction from the cell mock community. In human microbiota samples, MetaSAG recovered the largest number of genomes with a total of 103 gut microbial genomes (21 HQ and 65 showed >90% completeness) and 45 skin microbial genomes (10 HQ and 40 showed >90% completeness), respectively. Conventional binners recovered one *Staphylococcus hominis* genome, whereas MetaSAG recovered two *S. hominis* genomes from the identical skin microbiota sample. Single-cell sequencing indicated that these *S. hominis* genomes clearly derived from two distinct strains harboring specifically different plasmids. We found that all conventional *S. hominis* MAGs had

substantial lack or excess of the genome sequences and contamination of other *Staphylococcus* bacteria (*S. epidermidis).*

Conclusions:

MetaSAG enabled us to obtain the strain-resolved genomes in the mock community and human microbiota samples by assigning metagenomic sequences correctly and covering both highly conserved genes such as rRNA genes and unique extrachromosomal elements, including plasmids. MetaSAG will provide HQ genomes that are difficult to obtain with metagenomic analyses alone and will facilitate the understanding of microbial ecosystems by elucidating detailed metabolic pathways and horizontal gene transfer networks through microbial genomes. MetaSAG is available at https://github.com/kojiari/metasag.

Keywords:

Single-cell genomics, Metagenomics, Binning, Software

Background

The accumulation of reference genomes from microbes has provided insight into the ecology and evolution of environmental and host-associated microbiomes. The golden standard for microbial genome sequencing has been to culture specific strains and sequence extracted DNA[1–3]. Recently, metagenomic analysis, which combines direct extraction of genomic DNA from the microbial community with *in silico* reconstruction of each microbial genome sequence from massive sequenced reads, has attracted much attention. A growing number of metagenome-assembled genomes (MAGs) are asserting our understanding of microbial diversities in various environments[4–9].

In a metagenomic approach, genome reconstruction is performed in two steps: (1) assembly of fragmented genome sequences to contigs and (2) binning contigs into lineages as bins. Current state-of-the-art binners rely on nucleotide compositional information such as tetranucleotide frequency, GC content, or sequence coverage[10–12]. However, these tools demonstrate different performances and produce different MAGs including incomplete bins and multi-species composite bins[13]. Composite genomes that aggregate sequences originating from multiple distinct species or strains can yield misleading insights if they are registered as single genomes in the reference database[14]. To solve these problems, several approaches combine and curate the result of multiple binners to generate a large number of high-quality (HQ) genomes[13,15,16]. However, in the real samples, it is difficult to verify the certainty of the binning results because there are numerous microbes without the reference genome and the proportion of microbial species richness among them is unknown.

Single-cell genomics is an alternative approach that enables culture-independent sequencing of microbial genomes[17]. In contrast to metagenomics, single-cell genomics does not require microbial population clonality but instead recovers genome sequences from individual cells. In single-cell genomics, the DNA amplification process often causes amplification biases and incompleteness in genome sequences. Therefore, co-assembly of individual single-cell sequencing data is generally required to compensate for the gaps and

errors in each SAG sequence[18]. However, most SAGs generally have low completeness, and even with co-assembly, produced shortly fragmented contigs, rarely covering their entire genomic area.

Metagenomics assesses the genomes of all microbes present in a sample, whereas single-cell genomics reveals individual genomes. Therefore, it has been suggested that integrating the two can compensate for each of their specific shortcomings[19–21]. However, no efforts have been made to acquire multiple draft genomes of the human microbiota using this hybrid approach. Moreover, its advantages over the conventional metagenomics binning have not been verified. In this study, we developed a novel metagenome binning guided with single-cell amplified genomes (MetaSAG) to recover at once HQ genomes of multiple bacterial strains from the microbial community. We used microfluidic technology-aided approaches to obtain a large number of single-cell amplified genomes (SAGs) for guided binning[22,23]. Mock community and human microbiota samples were tested to compare the binning accuracy and the number of HQ genomes between conventional binners and MetaSAG. We also investigated the integration of single-cell genomes with metagenomes to acquire strain-resolved genomes and to validate the presence of aggregate sequences originating from multiple distinct species in metagenomic bins.

Results

Overview of the single-cell genome-guided metagenome binning (MetaSAG).

Conventional metagenomic phylogenetic classification tools[24,25] and conventional metagenomic binners[10–12] have difficulty in allocating contigs to bins from complex microbial communities in the absence of known microbial genome information as teaching data for classifying closely related species or strains. Our MetaSAG tool uses single-cell amplified genomes, which are newly produced in the same sample, as teaching data for metagenome binning (Fig. 1). The SAGs of uncultured microbes serve as ideal references for metagenome binning from the reference-lack microbial community. These SAGs were obtained using the SAG-gel platform[22,26], which enables obtaining the contamination-less

116 SAGs in a high throughput manner with the aid of microfluidic droplet format. Multispecies
117 SAGs obtained by assembly from each single-cell genome are grouped into individual
118 strains using the ccSAG method[18]. Composite SAGs (CoSAGs) are constructed by re-
119 assembling (co-assembling) single-cell reads (SRs) recognized as identical strains. Based
120 on genome completeness (>50%) and contamination (10%), non-redundant SAGs (nrSAGs)
121 are collected for use as binning references. Besides, metagenomic reads (MRs) are
122 obtained from the same sample and are assembled into metagenomic assembled contigs
123 (MAs). The contigs in nrSAGs are mapped to the contigs in MA to allocate contigs in MAs to
124 single cell-guided bins (sgBins). Finally, the paired nrSAGs and sgBins at the strain level are
125 merged to fill in the gaps for each other and extend the contig length as single-cell-guided
126 MAGs (sgMAGs) or metagenome-guided SAGs (mgSAGs).
127
128 **Evaluation of single-cell genome and metagenome assemblies**
129 To confirm sequence accuracy in nrSAGs and MAs, single-cell genomic and
130 metagenomic sequencings were performed with the same cell mock community containing
131 15 different bacteria (Additional file 1: Table S1). In total, we obtained 48 SRs and one MR
132 with total read lengths of 3.9 Gb and 2.6 Gb, respectively (Additional file 2: Table S2).
133 Following the assembly, 15 nrSAGs were obtained, which covered all species in the
134 mock community. From SAG to CoSAG according to taxonomy identification (Additional file
135 3: Table S3), the average completion rates improved from 33.5% to 66.6%, with low
136 contamination rates of 0.3% and 0.76%, respectively (Additional file 4: Table S4). In 14
137 nrSAGs, approximately ≥98.5% of the total length of each was correctly mapped to
138 reference genomes. In Mock-C00006 (*Lactobacillus delbrueckii*), some contigs (8.5% of the
139 total length) were mapped to other microbial genomes. The original SAGs were obtained
140 from physically isolated single-cells in gel capsules[22]; however free DNA was randomly
141 captured and subsequently amplified simultaneously. The unmapped contigs could have
142 been derived from these free DNA fragments. Alternatively, we confirmed that 1008 contigs
143 of total 1016 MA contigs were mapped to single reference genomes (Additional file 5: Fig.

144 S1). In addition, there were no 16S rRNA gene sequences for *Bacteroides uniformis* and
145 *Escherichia coli* in MA, while all nrSAGs remained individual 16S rRNA sequences
146 (Additional file 5: Fig. S2). Overall, both the nrSAGs and the MA showed a high sequence
147 accuracy as high identity corresponding to reference genomes. Thus, we considered that the
148 metagenomic binning step was crucial for reconstructing each genome from the MA.
149

150 **Comparison of the characteristics of single-cell-guided bins with conventional bins.**

151 We investigated the characteristics of bins collected by MetaSAG and conventional
152 binners (Fig. 2). Three binners, CONCOCT[10], MaxBin2[12], and MetaBAT2[11], were used
153 to construct bins, and DAS_Tool[13] was subsequently used to obtain refined bins. Based
154 on 15 reference genomes (Additional file 1: Table S1), we assessed the taxa of each bin and
155 estimated the total sizes of the contigs incorrectly binned to different bacterial bins and
156 contigs unbinned to any reference genomes (Fig. 2a). The binner with the smallest incorrect
157 binned contig was MetaSAG with 20 kbp, followed by MetaBAT2 with 181 kbp. The binner
158 with the smallest unbinned contig was CONCOCT (1 kbp), while the unbinned contig length
159 was 892 kbp in MetaSAG. Total lengths of contigs unbinned into target sgBins were
160 inversely correlated with the guide SAG completeness (Fig. 2b), suggesting that the SAG
161 completeness strengthens the adequacy of the contig allocation to the bin.

162 We also calculated F1 scores, a harmonic mean of precision and recall, to evaluate
163 the accuracy of bins construction against true reference genomes (Fig. 2c). The precision
164 depends on the less false-positive contig that is incorrectly allocated in the bin. Although the
165 ability to force contigs into bins helps improve completeness, it carries the risk of inclusion of
166 artificial sequences as false-positive contigs and increases contamination rates. MetaSAG
167 demonstrated high-precision bins against all 15 corresponding references. The high-
168 precision bin (F1 score >0.9) for MetaSAG, DAS_Tool, MetaBAT2, MaxBin2, and
169 CONCOCT were 15, 14, 8, 13, and 12, respectively, while all binners except MetaBAT2 had
170 equally high precision values. Alternatively, the recall value depends on the true

171 completeness of the bacterial genome. MetaSAG demonstrated the highest F1 scores
172 among all reference genomes owing to the highest recall value. In this test, SAG qualities
173 were limited to low-quality (LQ) to medium-quality (MQ), which were not the best conditions
174 to guide binning; however, it was still remarkably clear that MetaSAG had the best binning
175 accuracy. Thus, the single-cell guided binning approach of MetaSAG helps accurate and
176 efficient allocation of contigs into different bacterial genomes compared to conventional
177 binners.
178

179 **Effectiveness to merge nrSAGs and sgBins**

180 The merging of paired nrSAGs and sgBins into sgMAG or mgSAG improved several
181 genome assembly quality metrics such as completeness and N50 in multiple microbial
182 communities including human gut microbiota, and human skin microbiota (Fig. 3ab).
183 Although the completeness of either nrSAG or sgBin was low (average: 74.5%), that of
184 sgMAG and mgSAG was much improved (average: 93.6%) (Fig. 3a). N50 metrics of most
185 nrSAGs (average: 48.2 kb) improved after merging nrSAG and sgBin (average: 87.7 kb),
186 except in the case of low completeness of sgBins (Fig. 3b). Low completeness of sgBins
187 occurred often, particularly in skin microbiota (average completeness: 23.1%). This may be
188 due to the inability of metagenomic data to produce qualified MAs owing to some interfering
189 factors, such as human DNA contamination. Thus, to recover sgBins with high
190 completeness, it is necessary to increase the MA mapping rate by improving its assembly
191 accuracy and by increasing SAGs repertoire corresponding to MAs. In addition, rRNA and
192 tRNA gene sequences were often compensated from nrSAGs (Recovery rate of rRNA: 5S:
193 >53.1%, 16S: >94.1%, 23S: >98.5% in nrSAGs; and 5S: >7.5%, 16S: >13.4%, 23S: >14.9%
194 in sgBins)(Fig. 3cb), thus merging of nrSAGs and sgBin is extremely important for
195 incorporating phylogenetic information of draft genomes.
196

197 **Recovery of HQ draft genomes from multiple microbial communities using MetaSAG.**

198 We assessed the quality of all draft genomes according to the Genomic Standards
199 Consortium[9]. From the mock community sample, MetaSAG, DAS_Tool, and MaxBin2
200 constructed MAGs corresponding to 15 reference genomes, whereas MetaBAT2 and
201 CONCOCT constructed more than the expected 15 MAGs, including several LQ MAGs (Fig.
202 4a). Thus, the risk of creating unreliable MAGs must also be deliberated when considering
203 conventional binners. MetaSAG uses nrSAG taxonomy to identify representative species
204 and to extract contigs in MA necessary for binning, such that the risk of producing artificial
205 MAGs that cannot be present in actual samples is diminished. Regarding draft genome
206 quality, MetaSAG produced a total of 13 HQ draft genomes, with better accuracy than other
207 binners (Fig. 4a). For non-chromosomal elements, all plasmid sequences existed in MA;
208 however, these were lost in the plasmid-containing bacterial genomes after binning
209 (Additional file 5: Fig. S3). MetaSAG demonstrated constant and higher plasmid coverage
210 (97.2%) than other binners (50.6-74.5%) in five bacteria.
211 To evaluate the performance of MetaSAG in human gut and skin microbiota, three
212 SR sets (each 96 SR, 100 Mb/SR) and three MRs (each 6 Gb) were obtained, and the
213 assemblies were binned with MetaSAG and other binners. Here, MQ and HQ draft genomes
214 were considered for comparison. MetaSAG was able to construct the largest number of
215 genomes with a total of 103 (21 HQ) and 45 (10 HQ) genomes from the gut and skin,
216 respectively (Fig. 4a and Additional file 6: Table S5). In gut microbiota, none of the HQ
217 genomes were constructed in other binners. Although there were draft genomes that
218 satisfied a completeness >90% and contamination <5% in conventional binners, difficulty in
219 recovery of rRNA and tRNA sequences were clearly indicated (Fig.4 bc). MetaSAG
220 demonstrated consistent high performance in the recovery of rRNA (5S: >42.7%, 16S:
221 >61.2%, and 23S: >66.0%) and tRNA (average: 17.3 ± 2.9) in each microbial sample.
222 MetaSAG used a large number of sequencing reads by incorporating single-cell genomics
223 and metagenomics; however, trends were unchanged, even when the read number used for
224 other binners was equal to that when MetaSAG was used (Additional file 5: Fig. S4).
225

226 **Coverage of MetaSAG-produced draft genomes against bacterial diversity**

227 To determine the extent to which the constructed genome covered all metagenomic
228 sequence fractions, MRs were mapped to their respective genomes and mapping rates were
229 calculated. In MAGs constructed by MaxBin2 and CONCOCT, >90% of MRs were mapped
230 (Fig. 5a). These high mapping rates were considered owing to their algorithm trends of
231 unbinned contig reduction (Fig. 2a). The MR mapping rates in MetaSAG were in the middle
232 of all binners, ranging from 78.9% to 89.5% for gut microbiota and 91.3% to 95.6% for skin
233 microbiota. Regarding the bacterial diversity, MetaSAG detected more bacterial genomes
234 than other binners, with 54 genera in gut microbiota (Fig. 5b) and nine genera in skin
235 microbiota (Fig. 5c). We considered that the metagenomic coverage of MetaSAG could be
236 improved by increasing the number of obtained SAGs from the same samples and the
237 number of detected taxa.

238

239 **Strain-resolved genome analysis based on single-cell genomes toward revealing**
240 **intra-species diversities.**

241 Accurate genomic classification of closely related species and subspecies from the
242 microbial community is important and is required to discuss intra-species diversity.
243 Therefore, to assess separation accuracy of closely related genomes, we assessed the
244 correspondence between MAG and SAG sequences of the same species.

245 In skin microbiota, conventional metagenome binner yielded one *Staphylococcus*
246 *hominis* MAG, whereas MetaSAG yielded two *S. hominis* strain mgSAGs (*S. hominis*
247 BBMGS-S01-101 and *S. hominis* BBMGS-S01-100). We considered that conventional MAG
248 had difficulty in binning contigs to two different strains in the sample. To confirm further
249 details, we calculated the average nucleotide identity (ANI) of the two strain genomes
250 obtained by MetaSAG and other binners against the original SAGs (Fig. 6a) and we
251 confirmed all ANI showed >97% identities. We found that while the presence of two strains is
252 evident at the single-cell level, only MetaSAG was able to output strain-resolved genomes,

253 and the binner produced MAGs which demonstrated increased similarity to only one strain.
254 Notably, we found plasmids in conventional MAGs; however, plasmid assignment to
255 mgSAGs clearly indicated that these two strains had specifically different plasmids (Fig. 6b).
256 Thus, MetaSAG will aid strain-resolved binning and plasmid-host allocation, resulting in an
257 understanding of intra-species diversities and linking mobile gene elements to hosts.
258

259 **Validation of aggregate sequences originating from multiple distinct species.**

260 SAG can be used as a self-check reference to evaluate the appropriateness of
261 conventional MAG binning results, and it may also be possible to remove unsuitable contigs
262 such as aggregate sequences from multiple species. A simple way to detect incorrect
263 sequences in MAG is to map corresponding SAG sequences to MAGs (Fig. 6c). For *S.*
264 *hominis* obtained from human skin microbiota, we screened SAG sequences which were
265 mapped on MAGs obtained with conventional binners. This result clearly indicated that *S.*
266 *hominis* MAGs showed different genome sizes between different binners (1.4 to 3.2 MB)
267 while they showed high completeness (83% to 94%), suggesting substantial lack or excess
268 of the genome sequence, and some sequences from the closed *Staphylococcus* genus (*S.*
269 *epidermidis*) were contaminated in all MAGs (55.6-146.8kb). In particular, we found that the
270 longest contaminated contig (44 Kb) in MAGs of DAS_tool and MaxBin2 showed homology
271 (identity 98.5%) to pSE2 plasmid of *S. epidermidis* (CP066374). The genome sizes of
272 publicly available *S. hominis* isolate genomes are 2.1 to 2.3 Mb and are similar to the draft
273 genome obtained with MetaSAG. BBMGS-S01-101 and BBMGS-S01-100 exhibited some
274 common sequences between *S. hominis* and *S. epidermidis*; however, there were no
275 obvious interspecies aggregate sequences. Using SAGs as references, contigs that have
276 been erroneously removed or included by conventional binners can be correctly assigned,
277 suggesting that even uncultured bacterial genomes can be validated for showing their
278 reliability at the strain-level.
279

Discussion

HQ reference genomes are essential for understanding the phylogeny and function of uncultured microbes in complex microbial ecosystems. In a changing environment, microbes acquire adaptive evolution through repeated genetic mutations and horizontal transfer, etc.[27–30]. To best understand the connections between microbial communities and their habitats is to recover genomes from the communities themselves, rather than referring to genomes of closely related bacteria isolated from different environments.

Despite the cell mock community being a simple sample consisting of 15 different bacteria, the occurrence of false-positive contigs in the conventional MAG suggested the requirement for careful selection of the metagenomic binner depends on the presence of conserved genes and consistency of nucleotide composition. As reported previously[13,15], the tool that utilizes the bin refinement strategy demonstrated high accuracy, which was in agreement with MAG and reference genomes. These tools utilize multiple binners to generate various combinations of bins for reference to each other from single or multiple metagenomics data. Alternatively, MetaSAG generates self-references from the same sample at the single-cell level and they are subsequently guided to bin metagenomic contigs for genome reconstruction. MetaSAG enabled us to obtain the highest qualities in draft genomes, both in the mock and human microbiota samples, by assigning metagenomic sequences in correct bins, as well as by filling the gap in highly common sequences, such as rRNA genes, and linking the host with extrachromosomal elements, such as plasmids. The integration of metagenomics and single-cell genomics has been used to improve genome recovery from environmental bacteria. It was previously reported that metagenomic reads can be used to fill in gaps in SAGs[19], or that SAGs can be used as scaffolds for MAGs[20]. However, the number of constructed genomes was limited, and no tool has been developed to obtain multispecies genomes at once, which is mostly due to the lack of technology that provides good quality SAGs as binning guides. In MetaSAG, the qualities of SAGs obtained by our SAG-gel technology[22] were sufficiently high to prevent false-positive contigs in supervised contig identification. In addition, merging of SAGs with the metagenomic bin

308 aided recovery of rRNA and tRNA sequences, which were frequently lacking in the MAGs
309 obtained by conventional binners. This advantage overcomes the incompleteness of
310 phylogenetic information contained in conventional metagenomic bins, suggesting that this
311 technology can be used to move forward from conventional microbial profiling using 16S
312 rRNA gene amplicon sequencing to metabolic function analysis referring to novel genomes.
313 One of the challenges of MetaSAG is the difficulty in obtaining genome sequences
314 beyond the number of SAGs acquired in advance. To obtain genomes from samples of high
315 microbial diversity or to obtain genomes of rare microbes, it is necessary to obtain a large
316 number of SAGs or to selectively obtain SAGs of the desired taxa. In this study, we recruited
317 the SAGs with the completeness >20% to produce CoSAG with the completeness >50%. To
318 improve the genome number, the approaches are considered to accumulate massive LQ
319 SAGs with low sequencing efforts to produce nrSAG which covers a broad microbial
320 spectrum, or target single-cell genome sequencing with species enrichment
321 techniques[17,31,32]. Another issue with MetaSAG is that it only allows allocation to a single
322 sgBin per contig for binning using nrSAG as a guide. Under this binning condition, if there
323 are multiple bacterial strains with extremely similar sequences, the assignment of MA contig
324 to sgBin may not be fulfilled in any of the strain genomes. Nonetheless, the implementation
325 of contig assignment to multiple sgBin requires careful consideration owing to the complexity
326 of the computational process and the possibility of producing interspecies aggregate
327 sequences. We recommend using mgSAG, where the completeness of the SAG itself is
328 increased and used as primary data, and the metagenome is used as supplementary
329 information. This procedure allows us to obtain strain-resolved genomes and observe
330 differences among strains, taking advantage of the resolution of SAGs.
331 MetaSAG has the ability to control the SAG integration level by adjusting parameters.
332 It is possible to construct representative sequences for each taxonomy rank by appropriately
333 setting single copy marker gene homology, ANI, and tetranucleotide frequency, which are
334 parameters used for SAG integration to CoSAG. These SAGs can be utilized as reference
335 genome sequences against which resulting MAGs are checked for harboring interspecies

aggregate sequences. Verification of the reliability of MAGs is critical because composite genomes that aggregate sequences from several different populations can provide misleading insights when treated and reported as a single genome. By using MetaSAG, if biological samples that are the source of metagenomic data are properly stored and new single-cell data can be obtained, we will be able to increase the accuracy of acquired data curation and MAG by obtaining new single-cell genomes. Besides, MetaSAG can subdivide genomes of individual strains, even for species that cannot be divided into strain levels by metagenomic bins. Single-cell based strain-resolved genome analysis will contribute to our understanding of intraspecies diversity and distribution of non-chromosomal elements[29,33–35].

Conclusion

In conclusion, MetaSAG can integrate SAG and MAG to reconstruct qualified microbial genomes and control their binning resolution based on the numbers and classification of SAGs. Since it can provide reliable HQ genomes from a variety of microbial communities, it will represent a powerful tool to support microbial research that requires reference genome expansion and strain-resolved analysis toward understanding microbial association to the host or environment. Thus, MetaSAG is highly scalable and can be applied to reuse previously acquired metagenomics data and single-cell genomics tools to be developed.

358 **Methods**

359 **Experimental design and sample collection.**

360 Fresh feces were collected by subjects in 15 mL vials containing 3 mL GuSCN
361 solution (TechnoSuruga Laboratory Co., Ltd,) and stored for 2 d maximum, prior to DNA
362 extraction and single-cell encapsulation in droplets.

363 Skin bacterial samples were collected and placed in Dulbecco's phosphate-buffered
364 saline (DPBS) by swabbing the surface of facial skin using sterile cotton applicators (Nissui
365 Pharmaceutical Co., Ltd) pre-moistened with DPBS by subjects and were stored at room
366 temperature for 2 d maximum, prior to DNA extraction and single-cell genome amplification.

367 The mock microbial community (Cell-Mock-001) was obtained from the National
368 Institute of Technology and Evaluation Biological Resource Center, Japan. This mock
369 microbial community was composed of 15 bacterial strains detected in various environments
370 (intestinal, oral, skin, and natural environment) as follows: *Bacteroides uniformis,*
371 *Bifidobacterium pseudocatenulatum, Clostridium clostridioforme, Cutibacterium acnes*
372 *subsp. acnes, Escherichia coli* K-12*, Parabacteroides distasonis, Staphylococcus*
373 *epidermidis, Streptococcus mutans, Acinetobacter radioresistance, Comamonas*
374 *terrigenous, Bacillus subtilis subsp. subtilis, Clostridium butyricum, Corynebacterium*
375 *striatum, Lactobacillus delbrueckii subsp. delbrueckii,* and *Pseudomonas putida*.

376

377 **Single-cell genome sequencing with SAG-gel.**

378 Single-cell genome sequencing was performed with single-cell whole genome
379 amplification (WGA) using the SAG-gel platform according to our previous
380 reports[22,26]. Following homogenization of human feces in GuSCN solution (500 μL), the
381 supernatant was recovered by centrifugation at 2000 ×*g* for 30 s, followed by filtration
382 through 35-μm nylon mesh and centrifugation at 8,000 ×*g* for 5 min. The resulting cell pellets
383 were suspended in PBS, washed twice at 8,000 ×*g* for 5 min. Skin swab samples in DPBS
384 were processed in the same manner except for homogenization.

385 Prior to single-cell encapsulation, cell suspensions were adjusted to 0.1 cells/droplets
386 in 1.5% agarose in PBS to prevent encapsulation of multiple cells in single droplets. Using
387 the droplet generator (On-chip Biotechnologies Co., Ltd.), single microbial cells were
388 encapsulated in droplets and collected in a 1.5-mL tube, which was chilled on ice for 15 min
389 to form the gel matrix. Following solidification, collected droplets were broken with
390 1H,1H,2H,2H-perfluoro-1-octanol (Sigma-Aldrich) to collect beads. Thereafter, the gel beads
391 were washed with 500 μL acetone (Sigma-Aldrich), and the solution was mixed vigorously
392 and centrifuged. The acetone supernatant was removed, 500 μL isopropanol (Sigma-
393 Aldrich) was added, and the solution was mixed vigorously and centrifuged. The isopropanol
394 supernatant was removed, and the gel beads were washed three times with 500 μL DPBS.
395 Thereafter, individual cells in beads were lysed by submerging the gel beads in lysis
396 solutions: first, 50 U/μL Ready-Lyse Lysozyme Solution (Epicentre), 2 U/mL Zymolyase
397 (Zymo research), 22 U/mL lysostaphin (MERCK), and 250 U/mL mutanolysin (MERCK) in
398 DPBS at 37 °C overnight; second, 0.5 mg/mL achromopeptidase (FUJIFILM Wako
399 Chemicals) in PBS at 37 °C for 8 h; and third, 1 mg/mL Proteinase K (Promega) with 0.5%
400 SDS in PBS at 40 °C overnight. At each reagent replacement step, the gel beads were
401 washed three times with DPBS and subsequently resuspended in the next solution.
402 Following lysis, gel beads were washed with DPBS five times and the supernatant was
403 removed. Then, the beads were suspended in Buffer D2 and subjected to multiple
404 displacement amplification (MDA) using a REPLI-g Single Cell Kit (QIAGEN).
405 Following WGA at 30 °C for 2 h, gel beads were washed three times with 500 μL DPBS.
406 Thereafter, beads were stained with 1× SYBR Green (Thermo Fisher Scientific) in DPBS.
407 Following confirmation of DNA amplification by the presence of green fluorescence in the
408 gel, fluorescence-positive beads were sorted into 0.8 μL DPBS in 96-well plates using a
409 FACSMelody cell sorter (BD Bioscience) equipped with a 488-nm excitation laser. Following
410 droplet sorting, 96-well plates proceeded to the second round of WGA or were stored at
411 −30 °C.

412 Following gel bead collection in 96-well plates, second-round MDA was performed
413 with the REPLI-g Single Cell Kit. Buffer D2 (0.6 μL) was added to each well and incubated at
414 65 °C for 10 min. Thereafter, 8.6 μL of MDA mixture was added and incubated at 30 °C for
415 120 min. The MDA reaction was terminated by heating at 65 °C for 3 min. Following second-
416 round amplification, master library plates of SAGs were prepared. For quality control,
417 aliquots of SAGs were transferred to replica plates for DNA yield quantification using a Qubit
418 dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific). For sequencing analysis,
419 sequencing SAG libraries were prepared from the second-round MDA products using
420 QIAseq FX DNA Library Kit (QIAGEN). Ligation adaptors were modified to TruSeq™–
421 Compatible Full-length Adapters UDI (Integrated DNA Technologies). Each SAG library was
422 sequenced using an Illumina HiSeq 2 × 150 bp configuration (Macrogen).
423
424 **16S rDNA sequencing.**
425 To confirm amplification from single-cell genomes and to identify the taxonomy from
426 the mock community sample, 16S rRNA gene fragments V3–V4 were amplified with 341F
427 and 806R primers (Forward, 5′-
428 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3′;
429 reverse, 5′-
430 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3′)
431 and sequenced by Sanger sequencing from SAGs obtained by SAG-gel. Following
432 taxonomy identification with BLAST, every two to four species for mock communities were
433 selected for whole-genome sequencing.
434
435 **Metagenome sequencing.**
436 Total DNA was extracted from mock samples according to the International Human
437 Microbiota Standard protocol Q[36]. The DNeasy Power Soil Pro Kit (QIAGEN) was used for
438 total DNA extraction from fecal and skin swab samples. Metagenomic sequencing libraries

439 were constructed from extracted DNA samples with 10-μL (1/5 volume) reactions of the
440 QIAseq FX DNA Library Kit. Each metagenomic sequencing library was sequenced using an
441 Illumina HiSeq 2 × 150 bp configuration (Macrogen).
442

443 **Pre-processing and assembly of single-cell genomic and metagenomic sequence**
444 **reads**

445 SRs and MRs were individually processed for eliminating LQ reads by using fastp
446 0.20.1[37] with default options or bbduk.sh 38.79[38] (options: qtrim=r, trimq=10,
447 minlength=40, maxns=1, minavgquality=15). Human genome contaminations were removed
448 from SRs and MRs by mapping with bbmap.sh 38.79. SRs were assembled *de novo* using
449 SPAdes 3.14.0 (options for SAG: --sc --careful --disable-rr --disable-gzip-output -t 4 -m 32),
450 and contigs <1000 bp were excluded from subsequent analyses[39]. The MRs were
451 assembled into contigs *de novo* using SPAdes 3.14.0 (options: --meta, -t 12, -m 96).
452

453 **Grouping of the same strain SAGs into CoSAG**

454 SAGs with the completeness >10% in the mock community, 20% in the human
455 microbiota sample, and contamination of <10% were selected with CheckM[40]. ANI was
456 calculated for selected SAGs using fastANI 1.3[41]. The homology of common single-copy
457 marker genes obtained using CheckM v1.1.2 taxonomy workflow (option: -nt --tab_table -t
458 16 domain Bacteria) was calculated by blastn 2.9.0+ with the default option. SAGs with ANI
459 >95%, single-copy marker gene homology >99%, and tetra-nucleotide frequencies
460 correlation >90% were identified in the same strain group. SRs from one SAG were mapped
461 to other SAGs in the same group using MINIMAP2 2.17 (options: -ax sr)[42]. According to
462 the ccSAG procedure[18], potential chimeras that partially aligned were split into aligned and
463 unaligned fragments. The short fragments (<20 bp) were discarded. Clean and chimera-
464 removed reads were obtained using cycles of cross-reference mapping and chimera splitting
465 for each sample in the same group. Quality controlled reads from the same group were co-

466 assembled *de novo* as CoSAG using SPAdes (options: --sc --careful --disable-rr --disable-
467 gzip-output -t 4 -m 32).
468

469 **SAG-guided binning of metagenome contigs**

470 The MAs were individually mapped against the strain-specific nrSAG contig using
471 BWA 0.7.17 with the default option[43]. MA contigs that showed >99% identity (>200bp) to
472 nrSAG contigs were extracted to construct sgBins by assignment based on nrSAG taxa.
473

474 **Merging of nrSAG and sgBin**

475 For the two sets of assemblies, nrSAG and sgBin, CheckM was performed to
476 measure their completeness, and the assembly with higher completeness was defined as
477 the master and that with lower completeness as the slave. Slave assemblies were
478 processed using SeqKit[44], and contigs <10000 bp were removed. Master and slave
479 assemblies were merged using HaploMerger2_20180603[45] to create sgMAGs or
480 mgSAGs. Thereafter, MAGs were reconstructed by using the DAS-tool from MAs that were
481 unclassified as sgBin.
482

483 **Conventional MAG binning**

484 For comparison of MAG quality, multiple binning of metagenomic contigs were
485 conducted with conventional binners including CONCOCT 1.0.0[10], MaxBin 2 2.2.6[12], and
486 MetaBAT 2 2.12.1[11] with default options. To refine the binning results obtained using these
487 three different methods, DAS_Tool 1.1.2[13] was used with default options.
488

489 **Gene prediction, taxonomy identification, and plasmid detection**

490 CDS, rRNAs, and tRNAs were extracted from all SAGs or MAGs by Prokka
491 1.14.6[46] (option: --rawproduct --mincontiglen 200). 16S and 23S rRNA genes with lengths
492 ≥700 and 1100 bp, respectively, were detected. Taxonomy identification was performed

using GTDB-Tk 1.3.0[47] with the default option, using the Release95 database. PlasClass[48] was used for detecting plasmids.

Quality assessment of draft genomes from the mock community

In the mock community sample analysis, ANIs of each draft genome (sgMAG and mgSAG) for the closest reference genome were calculated with fastANI 1.3. The closest taxa with ≥ 99.5% ANI was assigned to each draft genome.

The quality of all obtained SAGs and MAGs were evaluated using QUAST v.5.0.2 (default option)[49], ChcekM v1.1.2 lineage workflow (option: --nt --tab_table -t 16), and identification of 5S, 16S, and 23S rRNA. To assess the accuracy of procured draft genomes in mock community samples, draft genomes were individually mapped to the corresponding taxa reference genome using MINIMAP2 2.17 with default options. The mapping results were converted to the pileup textual format using samtools 1.9[50], and the genomic coverage L for the reference genome was calculated using the following equation.

$$L_i = length(A_i \cap G_g), where\ g = \arg\max_{j \in G}\{ANI(A_i, G_j)\}$$

where A_i represents the ith draft genome. G and G_j represent the set of reference genomes and the jth reference genome of the set, respectively. G_g represents the corresponding reference genome against A_i. When the reference genome is G_g and the draft genome is A_i, precision (P), recall(R), and F value (F_1 score) of the reference genome were calculated using the following equations.

$$P_i = \frac{L_i}{length(A_i)}$$

$$R_i = \frac{L_i}{length(G_g)}$$

$$F_1 score_i = 2\frac{P_i R_i}{P_i + R_i}$$

Declarations

Ethics approval and consent to participate

Studies with human subjects were approved by the School of Science and Engineering at Waseda University (No. 2018-323 and No. 2019-381). The subjects gave their written informed consent prior to sample collection.

Consent for publication

Not applicable.

Availability of data and material

MetaSAG is available from https://github.com/kojiari/metasag. Sequencing data has been deposited in the NCBI database under the BioProject PRJNA692334 (see Additional file 6: Table S5 for details).

Competing interests

MH and HT are shareholders in bitBiome, Inc., which provides single-cell genomics service using SAG-gel workflow as bit-MAP. MH is a founder of bitBiome, Inc. KA, TS, TY, TE, and AM are employed at bitBiome, Inc. KA, KI, MK, HT, and MH are inventors on patent applications submitted by bitBiome, Inc. covering the technique for integration of metagenome and single-cell genome data.

Funding

This work was supported by Tokyo Metropolitan Small and Medium Enterprise Support Center.

Authors' contributions

KA, HT, and MH conceived and designed the experiments. KA, KI, MK, and MH developed MetaSAG. TS, TY, TE, and AM conducted genomics experiments and collected the data. KA

551 and KI conducted bioinformatic analysis of the metagenomic data and single-cell genomic
552 data. KA and MH wrote the manuscript. All authors read and approved the final manuscript.
553

554 **Acknowledgements**
555 The super-computing resource was provided by the Human Genome Center (University of
556 Tokyo). Figure 1 was created with BioRender.com.

557 **References**

558 1. Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, et al. A human gut bacterial
559 genome and culture collection for improved metagenomic analyses. Nat Biotechnol. 2019;37:186–92.

560 2. Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, et al. 1,520 reference genomes from cultivated
561 human gut bacteria enable functional microbiome analyses. Nat Biotechnol. 2019;37:179–85.

562 3. Mukherjee S, Seshadri R, Varghese NJ, Eloe-Fadrosh EA, Meier-Kolthoff JP, Göker M, et al. 1,003
563 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. Nat
564 Biotechnol. 2017;35:676–83.

565 4. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of
566 nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol.
567 2017;2:1533–42.

568 5. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. New insights from uncultivated genomes
569 of the global human gut microbiome. Nature. 2019;568:505–10.

570 6. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic
571 blueprint of the human gut microbiota. Nature. 2019;568:499–504.

572 7. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored
573 Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning
574 Age, Geography, and Lifestyle. Cell. 2019;176:649-662.e20.

575 8. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the
576 tree of life. Nat Microbiol. 2016;1:16048.

577 9. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al.
578 Minimum information about a single amplified genome (MISAG) and a metagenome-assembled
579 genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35:725–31.

580 10. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic
581 contigs by coverage and composition. Nat Methods. 2014;11:1144–6.

582 11. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning
583 algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ.
584 2019;7:e7359.

585 12. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover
586 genomes from multiple metagenomic datasets. Bioinformatics. 2016;32:605–7.

587 13. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes
588 from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol. 2018;3:836–
589 43.

590 14. Shaiber A, Eren AM. Composite Metagenome-Assembled Genomes Reduce the Quality of Public
591 Genome Repositories [Internet]. MBio. 2019. Available from: http://dx.doi.org/10.1128/mBio.00725-19

592 15. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved
593 metagenomic data analysis. Microbiome. 2018;6:158.

594 16. Song W-Z, Thomas T. Binning_refiner: improving genome bins through the combination of
595 different binning programs. Bioinformatics. 2017;33:1873–5.

596 17. Woyke T, Doud DFR, Schulz F. The trajectory of microbial single-cell sequencing. Nat Methods.
597 2017;14:1045–54.

598 18. Kogawa M, Hosokawa M, Nishikawa Y, Mori K, Takeyama H. Obtaining high-quality draft
599 genomes from uncultured microbes by cleaning and co-assembly of single-cell amplified genomes.
600 Sci Rep. 2018;8:2059.

601 19. Mende DR, Aylward FO, Eppley JM, Nielsen TN, DeLong EF. Improved Environmental Genomes
602 via Integration of Metagenomic and Single-Cell Assemblies. Front Microbiol. 2016;7:143.

603 20. Roux S, Hawley AK, Torres Beltran M, Scofield M, Schwientek P, Stepanauskas R, et al. Ecology
604 and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-
605 genomics. Elife. 2014;3:e03125.

606 21. Nobu MK, Narihiro T, Rinke C, Kamagata Y, Tringe SG, Woyke T, et al. Microbial dark matter
607 ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. ISME J.
608 2015;9:1710–22.

609 22. Chijiiwa R, Hosokawa M, Kogawa M, Nishikawa Y, Ide K, Sakanashi C, et al. Single-cell genomics
610 of uncultured bacteria reveals dietary fiber responders in the mouse gut microbiota. Microbiome.
611 2020;8:5.

612 23. Hosokawa M, Nishikawa Y, Kogawa M, Takeyama H. Massively parallel whole genome
613 amplification for single-cell sequencing using droplet microfluidics. Sci Rep. 2017;7:5199.

614 24. Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, et al. Phylogenetic
615 classification of short environmental DNA fragments. Nucleic Acids Res. Oxford University Press
616 (OUP); 2008;36:2230–9.

617 25. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. Genome Biol.
618 2008;9:R151.

619 26. Nishikawa Y, Kogawa M, Hosokawa M, Mineta K, Takahashi K, Sakanashi C, et al. Massively
620 parallel single-cell genome sequencing enables high-resolution analysis of soil and marine
621 microbiome [Internet]. bioRxiv. 2020 [cited 2020 Apr 25]. p. 2020.03.05.962001. Available from:
622 https://www.biorxiv.org/content/10.1101/2020.03.05.962001v1.full-text

623 27. Ramiro RS, Durão P, Bank C, Gordo I. Low mutational load and high mutation rate variation in gut
624 commensal bacteria. PLoS Biol. 2020;18:e3000617.

625 28. Van Rossum T, Ferretti P, Maistrenko OM, Bork P. Diversity within species: interpreting strains in
626 microbiomes. Nat Rev Microbiol [Internet]. 2020; Available from: http://dx.doi.org/10.1038/s41579-
627 020-0368-1

628 29. Zlitni S, Bishara A, Moss EL, Tkachenko E, Kang JB, Culver RN, et al. Strain-resolved
629 microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical
630 timescale. Genome Med. 2020;12:50.

631 30. Vatanen T, Plichta DR, Somani J, Münch PC, Arthur TD, Hall AB, et al. Genomic variation and
632 strain-specific functional adaptation in the human gut microbiome during early life. Nat Microbiol.
633 2019;4:470–9.

634 31. Cross KL, Campbell JH, Balachandran M, Campbell AG, Cooper SJ, Griffen A, et al. Targeted
635 isolation and cultivation of uncultivated bacteria by reverse genomics. Nat Biotechnol. 2019;37:1314–
636 21.

637 32. Hatzenpichler R, Krukenberg V, Spietz RL, Jay ZJ. Next-generation physiology approaches to
638 study microbiome function at single cell level. Nat Rev Microbiol. 2020;18:241–56.

639 33. Sorbara MT, Littmann ER, Fontana E, Moody TU, Kohout CE, Gjonbalaj M, et al. Functional and
640 Genomic Variation between Human-Derived Isolates of Lachnospiraceae Reveals Inter- and Intra-
641 Species Diversity. Cell Host Microbe. 2020;28:134-146.e4.

642 34. Yan Y, Nguyen LH, Franzosa EA, Huttenhower C. Strain-level epidemiology of microbial
643 communities and the human microbiome. Genome Med. 2020;12:71.

644 35. Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, Li C, et al. Hybrid metagenomic
645 assembly enables high-resolution analysis of resistance determinants and mobile elements in human
646 microbiomes. Nat Biotechnol. 2019;37:937–44.

647 36. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, et al. Towards standards for
648 human fecal sample processing in metagenomic studies. Nat Biotechnol. 2017;35:1069–76.

649 37. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
650 Bioinformatics. Oxford University Press (OUP); 2018;34:i884–90.

651 38. Bushnell B, Rood J, Singer E. BBMerge - Accurate paired shotgun read merging via overlap.
652 PLoS One. 2017;12:e0185056.

653 39. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new
654 genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol.
655 2012;19:455–77.

656 40. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality
657 of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res.
658 2015;25:1043–55.

659 41. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis
660 of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114.

661 42. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

662 43. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
663 Bioinformatics. 2009;25:1754–60.

664 44. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File
665 Manipulation. PLoS One. 2016;11:e0163962.

666 45. Huang S, Kang M, Xu A. HaploMerger2: rebuilding both haploid sub-assemblies from high-
667 heterozygosity diploid genome assembly. Bioinformatics. 2017;33:2577–9.

668 46. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.

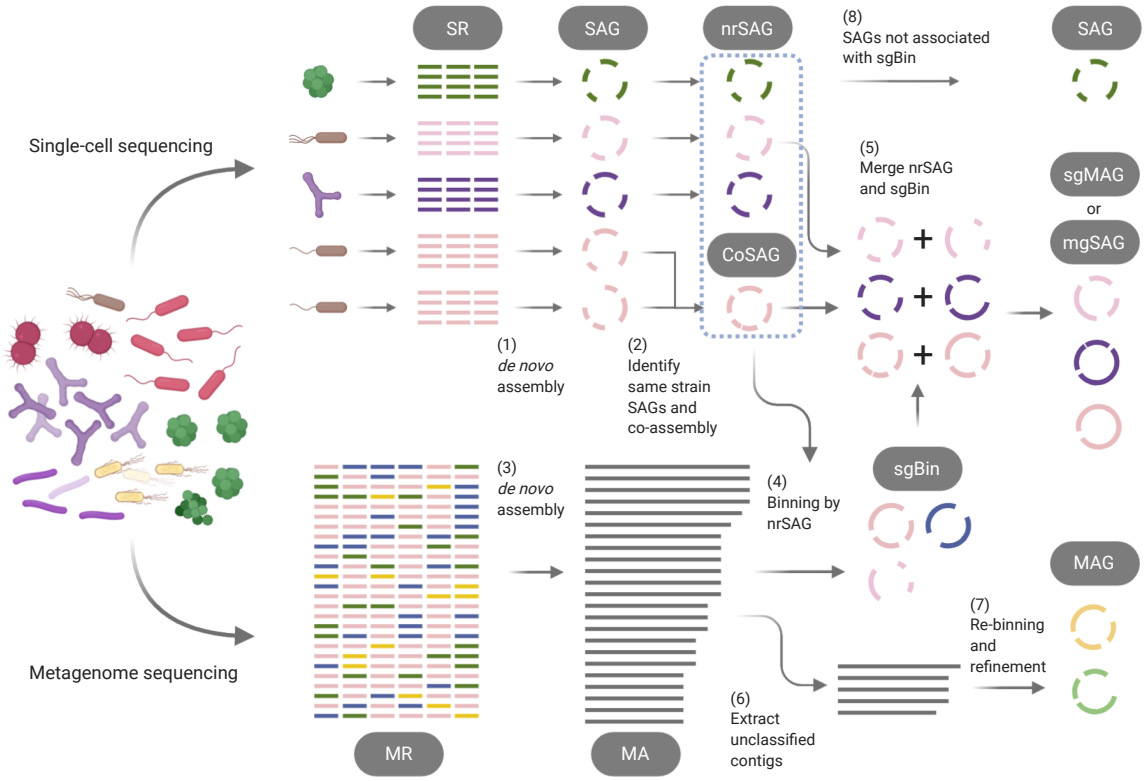
669 47. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with
670 the Genome Taxonomy Database. Bioinformatics [Internet]. 2019; Available from:
671 http://dx.doi.org/10.1093/bioinformatics/btz848

672 48. Pellow D, Mizrahi I, Shamir R. PlasClass improves plasmid sequence classification. PLoS Comput
673 Biol. 2020;16:e1007781.

674 49. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome
675 assemblies. Bioinformatics. 2013;29:1072–5.

676 50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map
677 format and SAMtools. Bioinformatics. 2009;25:2078–9.

678 **Figures**



679

680 **Fig. 1 Overview of metagenome binning guided with single-cell amplified genomes (MetaSAG)**
681 **workflow.** Single-cell sequencing reads (SRs) and metagenomic sequencing reads (MRs) are
682 obtained from the same microbial community. (1) *De novo* assembly of each SR to a single-cell
683 amplified genome (SAG). (2) SAGs of the same strain are identified into the group and co-assembled
684 into composite SAG (CoSAG). (3) *De novo* assembly of MRs to metagenome-assembled contigs
685 (MAs). (4) MA is classified to single-cell genome-guided bin (sgBin) by mapping MA on non-
686 redundant SAG (nrSAG). (5) Paired nrSAGs and sgBins are merged to single-cell genome-guided
687 MAG (sgMAG) or metagenome-guided SAG (mgSAG). (6) Unbinned contigs in MA are extracted and
688 subsequently (7) re-binned and refined by conventional metagenomic binning and refinement tools.
689 (8) Four types of draft genomes (SAG, sgMAG, mgSAG, and MAG) are finally acquired.
690

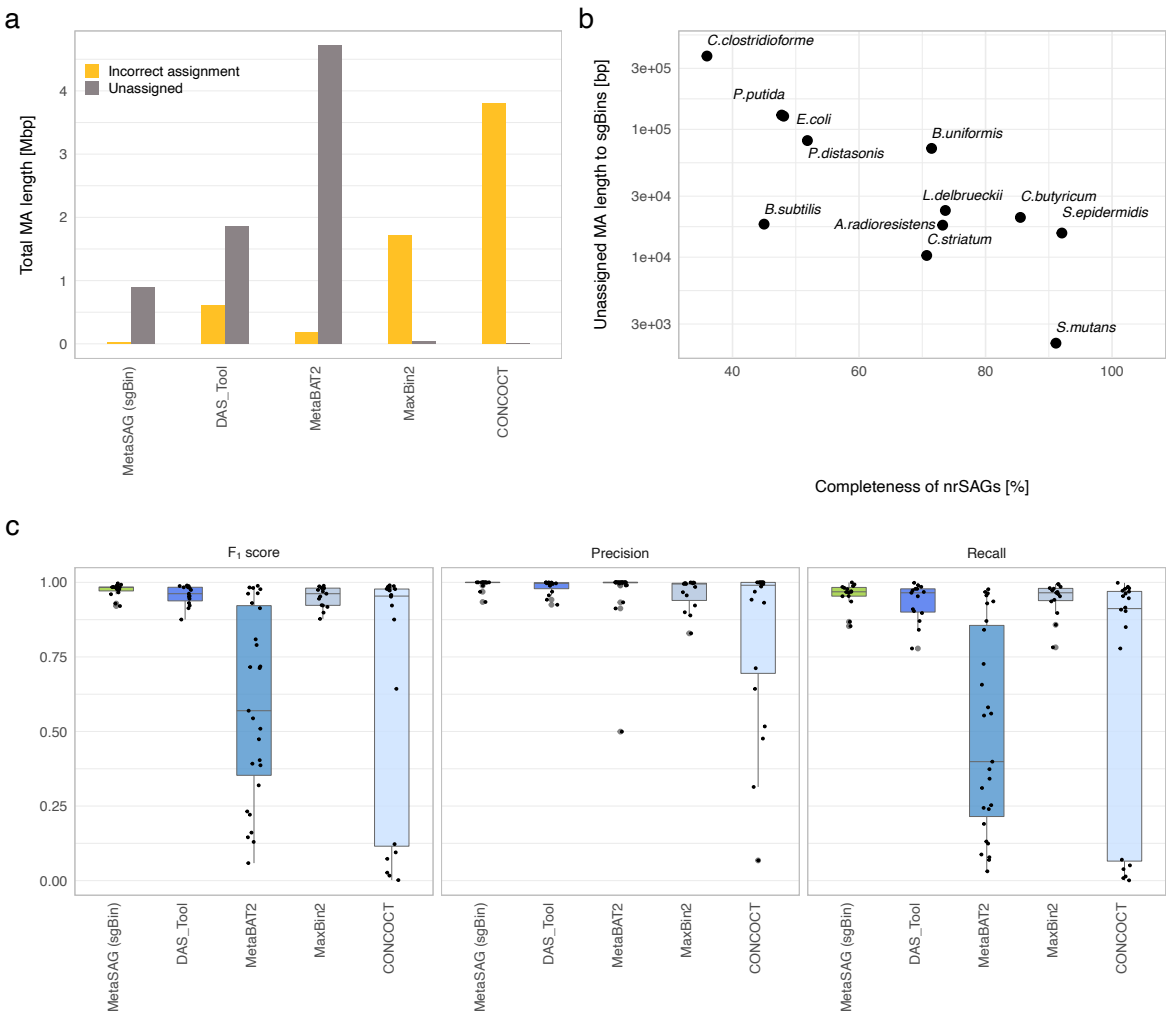


Fig. 2 Precision of single-cell-guided binning of metagenome assembly from a mock microbial community of 15 bacteria. (a) Total length of contigs incorrectly binned to different species and unbinned metagenome assembled contigs (MAs). (b) Correlation between completeness of non-redundant single-cell amplified genomes (nrSAGs) and total length of contigs unbinned to the target single-cell genome-guided bin (sgBin). (c) The plots of F1 scores, precision, and recall of all reported bins (center line, median; box limits, upper and lower quartiles; whiskers, minimum or maximum values between upper and lower quartiles that are extended 1.5 times the interquartile region). Individual values are represented as dots.

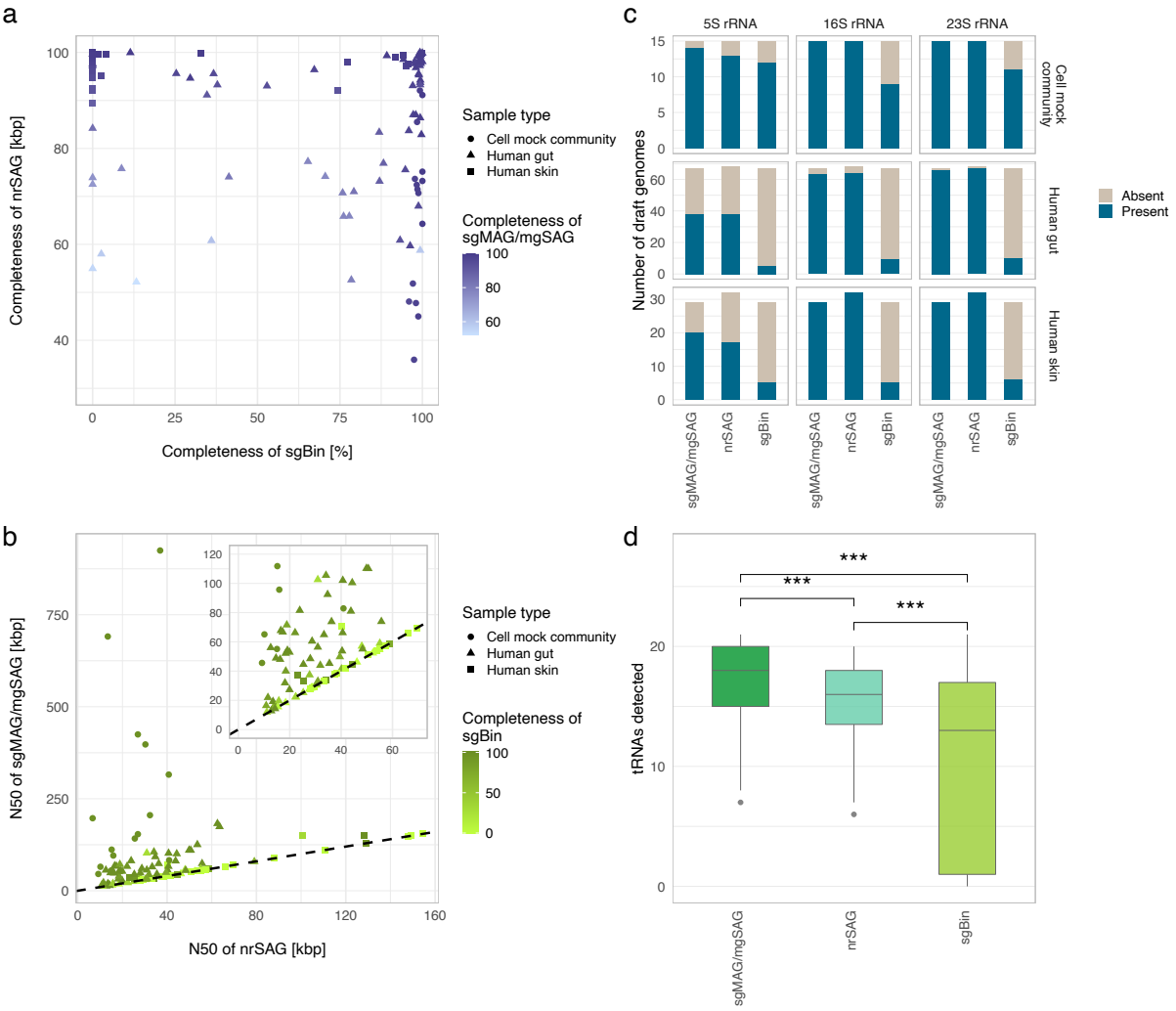


Fig. 3 Assembly quality metrics of single-cell-guided metagenome-assembled genome (sgMAG) and metagenome-guided single-cell amplified genome (mgSAG). All data were collected from a mock community containing 15 bacteria, three human fecal samples, and three human skin swab samples. (a) Scatter plot of completeness of non-redundant single-cell amplified genome (nrSAGs) versus single-cell genome guided bins (sgBins) corresponding to all medium-quality (MQ) and high-quality (HQ) sgMAGs and mgSAGs. (b) Relationship between N50s of nrSAG and of sgMAG or mgSAG. Numbers of rRNA genes (c) and tRNA genes (d) in draft genomes produced in MetaSAG workflow (center line, median; box limits, upper and lower quartiles; whiskers, minimum or maximum values between upper and lower quartiles that are extended 1.5 times the interquartile region, Wilcoxon rank sum test ***: p <0.001).

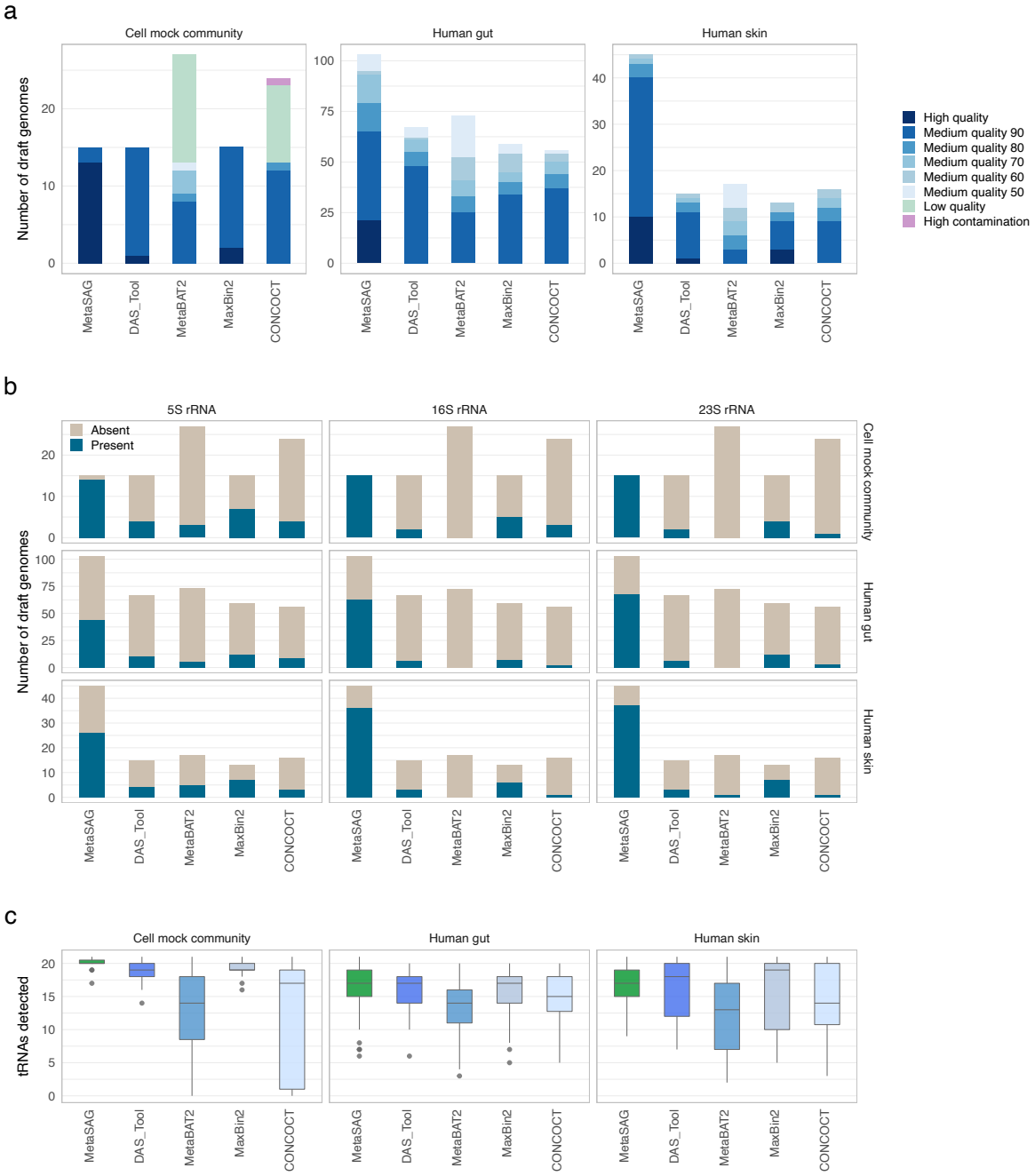


Fig. 4 Draft genomes reconstructed from the cell mock community and human microbiota samples. All data were collected from a mock community containing 15 bacteria, three human fecal samples, and three human skin swab samples. (a) The number of reconstructed genomes per method. Human gut and skin data show medium-quality (MQ) and high-quality (HQ) genomes only. Number of rRNA genes (b) and tRNA genes (c) in draft genomes produced with MetaSAG and other tools (center line, median; box limits, upper and lower quartiles; whiskers, minimum or maximum values between upper and lower quartiles that are extended 1.5 times the interquartile region).

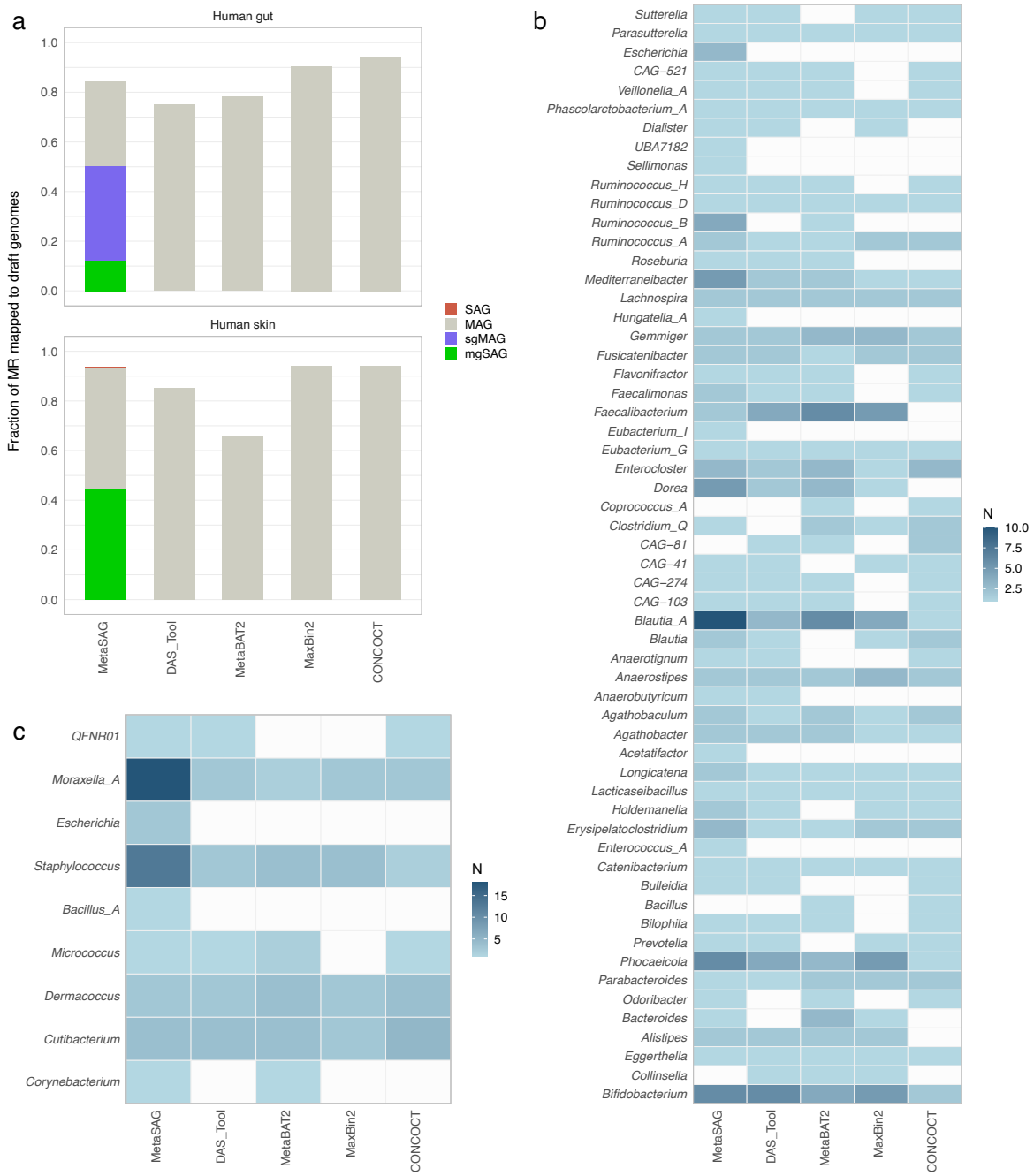


Fig. 5 Diversity of draft genomes reconstructed by MetaSAG.

(a) Fraction of metagenomic reads against draft genomes constructed by each binner. MetaSAG shows a fraction of metagenomic reads against four types of draft genomes. The number of draft genomes acquired from (b) human gut and (c) human skin are collapsed by genus assigned with GTDB-Tk.

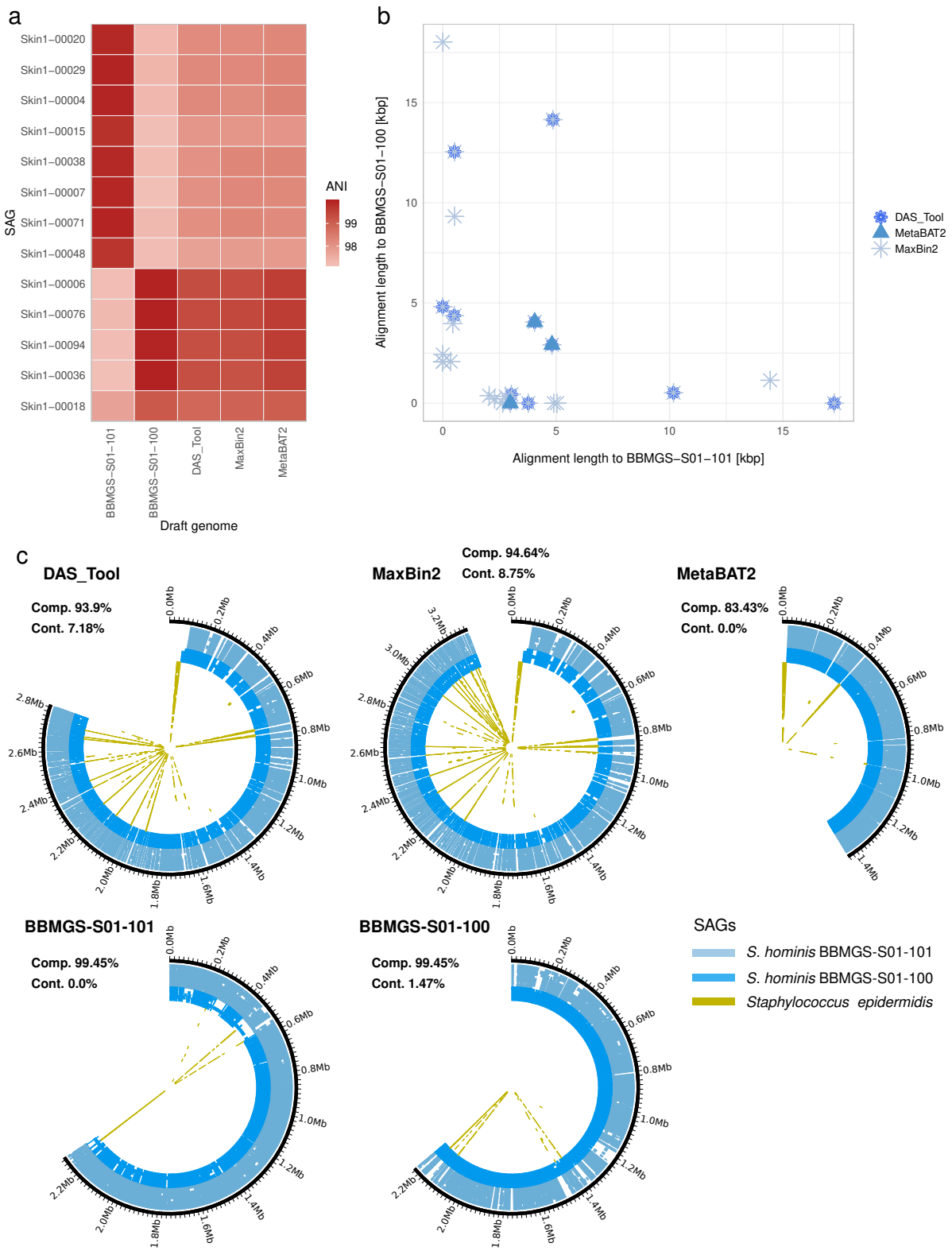


Fig. 6 Strain-resolved analysis of skin microbes for host-plasmid linking and detection of interspecies chimeric sequences.

(a) Mean pairwise genomic similarities between *Staphylococcus hominis* draft genomes obtained with MetaSAG (*S. hominis* BBMGS-S01-101 and *S. hominis* BBMGS-S01-100), and other binners. (b) The

734 scatter plot shows the length of plasmid contigs assigned to BBMGS-S01-101 and BBMGS-S01-100.
735 Different plot symbols show contigs obtained with different binners. (c) Interspecies chimeric
736 sequence detection by alignment of *S. hominis* MAGs derived with conventional binners (yellow), and
737 *S. hominis* BBMGS-S01-101 (light blue) and *S. hominis* BBMGS-S01-100 (blue). The outermost
738 circles show draft genomes and the inner circles show the result of mapping individual SAGs, which
739 belong to the same genus of *Staphylococcus*, to the draft genome.