


ARTICLE

DOI: 10.1038/s41467-017-00342-9

OPEN

Community-like genome in single cells of the sulfur bacterium *Achromatium oxaliferum*

Danny Ionescu¹, Mina Bizic-Ionescu¹, Nicola De Maio², Heribert Cypionka³ & Hans-Peter Grossart ^{1,4}

Polyploid bacteria are common, but the genetic and functional diversity resulting from polyploidy is unknown. Here we use single-cell genomics, metagenomics, single-cell amplicon sequencing, and fluorescence in situ hybridization, to show that individual cells of *Achromatium oxaliferum*, the world's biggest known freshwater bacterium, harbor genetic diversity typical of whole bacterial communities. The cells contain tens of transposable elements, which likely cause the unprecedented diversity that we observe in the sequence and synteny of genes. Given the high within-cell diversity of the usually conserved 16S ribosomal RNA gene, we suggest that gene conversion occurs in multiple, separated genomic hotspots. The ribosomal RNA distribution inside the cells hints to spatially differential gene expression. We also suggest that intracellular gene transfer may lead to extensive gene reshuffling and increased diversity.

¹Leibniz Institute of Freshwater Ecology and Inland Fisheries, Department of Experimental Limnology, Alte Fischerhuetten 2, 16775 Stechlin, Germany.

²Institute for Emerging Infections, Oxford Martin School, University of Oxford, 34 Broad Street, Oxford OX1 3BD, UK. ³Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, 26111 Oldenburg, Germany. ⁴Institute of Biochemistry and Biology, Potsdam University, 14476 Potsdam, Germany. Correspondence and requests for materials should be addressed to D.I. (email: ionescu@igb-berlin.de) or to H.-P.G. (email: hgrossart@igb-berlin.de)

Polyploidy, the condition of having multiple chromosome copies per cell is a frequent phenomenon in eukaryotic organisms¹. Polyploidy is suggested^{2, 3} and in some cases shown⁴ to be advantageous in regulation of gene expression, DNA repair, and supporting large cell sizes. Despite most commonly studied bacteria being haploid¹, polyploid *Archaea* and *Bacteria* (defined as having more than 10 genome copies) are common and can contain up to thousands⁵ of genome copies (hereafter chromosomes). These chromosomes are believed to be nearly identical copies and safeguarded against mutations by gene conversion (asymmetrical homologous recombination resulting in one allele “overwriting” another)⁶.

Polyploidy has been suggested to have a major role in the evolution of eukaryotes by allowing genomic rearrangements and gene duplication^{7, 8} that eventually result in different functionality by similar organisms. In *Bacteria* and *Archaea* the significance of polyploidy has received less attention. Polyploidy can lead to divergence of the coding material allowing the cells to experiment with new gene/protein versions^{5, 9}. Thus, a polyploid bacterium with divergent genome copies would benefit from the genetic diversity of a colony within each single cell⁹. However, observations from the highly polyploid *Epulopiscium* spp. using marker genes and recently from genomic studies of *Candidatus* *Marithrix* sp., suggested that genomic copies within a cell are all extremely similar^{9, 10}, possibly as a consequence of strong gene conversion, within-cell genome population bottlenecks at reproduction, and limited between cell recombination.

Achromatium sp. is the largest known unicellular freshwater bacterium, with several described size classes reaching up to $15 \times 125 \mu\text{m}$ ^{11, 12}. It is a colorless sulfur-oxidizing bacterium typically found at the oxic–anoxic interface in sediments of temperate freshwater lakes¹¹. The cells contain large calcite bodies and sulfur granules^{12, 13}. *Achromatium* was mostly studied in freshwater environments with several species and phylotypes

described¹⁴, but may be found in tidal salt marsh¹² and in mineral springs¹⁵ as well. According to nucleic acid staining^{12, 16}, like other large sulfur bacteria¹⁷, *Achromatium* appears to be polyploid.

Here we study *Achromatium* cells using genomic and metagenomic data from single and pooled “hand-picked” *Achromatium oxaliferum* cells from Lake Stechlin, NE Germany, coupled with 16S ribosomal RNA (rRNA) analysis of 27 single *Achromatium* cells and fluorescence in situ hybridization (FISH). We find extreme intracellular genetic diversity, and suggest that *Achromatium* undergoes intracellular gene duplications, re-assortments, and divergence with reduced or minimal gene convergence, leading to genetic diversity typical for populations rather than single cells. Our data suggests that the cells are equipped with numerous transposases, insertion sequences, and DNA editing factors as the machinery responsible for the intracellular evolution. These processes could explain the highly genetically heterogeneous *Achromatium* population at the level of individual cells.

Results

Evidence of polyploidy. A light micrograph of a dividing *Achromatium* cell from Lake Stechlin overlaid with the parallel DNA staining image (Fig. 1, Supplementary Fig. 1) shows that the individual cells contain multiple DNA spots that are not localized in one single area but rather spread across the cell, mostly in between calcium carbonate bodies. Analysis of several cells showed an average of 199 ± 46 spots. Given that the spots had varying fluorescence intensity, we cannot rule out each spot containing a varying amount of DNA¹⁸, i.e., a different number of chromosomes or chromosomes of varying sizes. Based on previous knowledge on large sulfur bacteria and giant *Firmicutes*¹⁷ these multiple DNA spots confirm the polyploid nature of *Achromatium*.

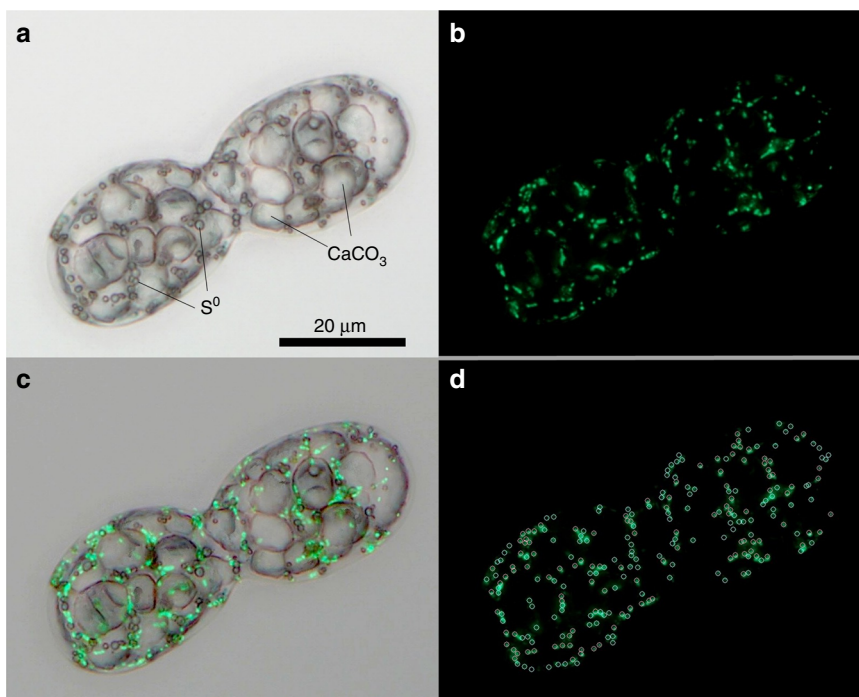


Fig. 1 Dividing cell of *Achromatium* sp. **a** Bright field showing calcite crystals (CaCO_3) and sulfur droplets (S^0). **b** Nucleic acids stained by SybrGreen I in the same cell. **c** Overlay of **a** and **b** showing that sulfur and nucleic acids spots are present in the grooves around the calcites, but not at the same positions. **d** Count of 244 DNA spots using the software tool CountThem. Both bright field and fluorescence images were taken as focus stacks of 22 images covering the full-cell depth and were processed by the stacking program PICOLAY. A similar image stained with the DNA exclusive dye picoGreen, is provided as Supplementary Fig. 1

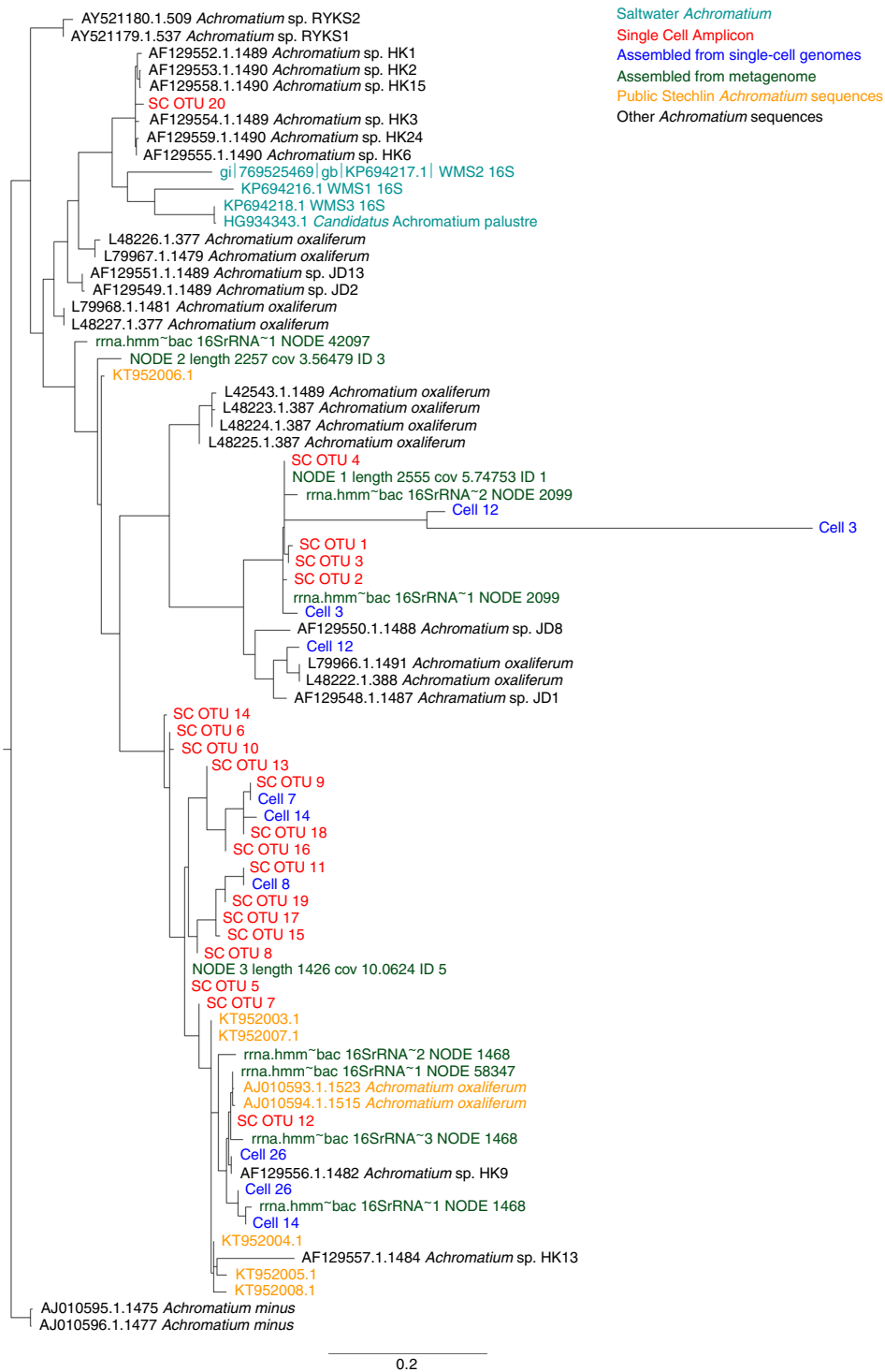


Fig. 2 Maximum likelihood tree of 16S rRNA sequences from *Achromatium*. The sequences were obtained from single *Achromatium* cells from Lake Stechlin, metagenomics data of the same cell population, and reference sequences. A similar tree which includes distance-clustered amplicon sequences is given in Supplementary Fig. 2A. A similar tree created using only full-length sequences to which the shorter ones were added by parsimony is provided as Supplementary Fig. 3

Community-like rRNA diversity in single cells of *Achromatium*. Metagenomic data obtained from sequencing of ca. 10,000 “hand-picked” and well pre-washed *Achromatium oxaliferum* cells from Lake Stechlin were analyzed for the presence of 16S rRNA gene sequences. Most of the 16S rRNA gene reads (>98%) were associated to *Achromatium* sp., suggesting a low level of contamination by the remaining epibiotic bacteria. These

reads assembled into three different full-length 16S rRNA sequences (93–95% similarity; Fig. 2, Supplementary Figs. 2 and 3). Several additional *Achromatium* affiliated partial reads were also identified (>91% identity) in the metagenome assembled data (Fig. 2, Supplementary Figs. 2 and 3).

A section of the 16S rRNA gene was further sequenced from 27 individual *Achromatium* cells. The V1–V4 region was sequenced

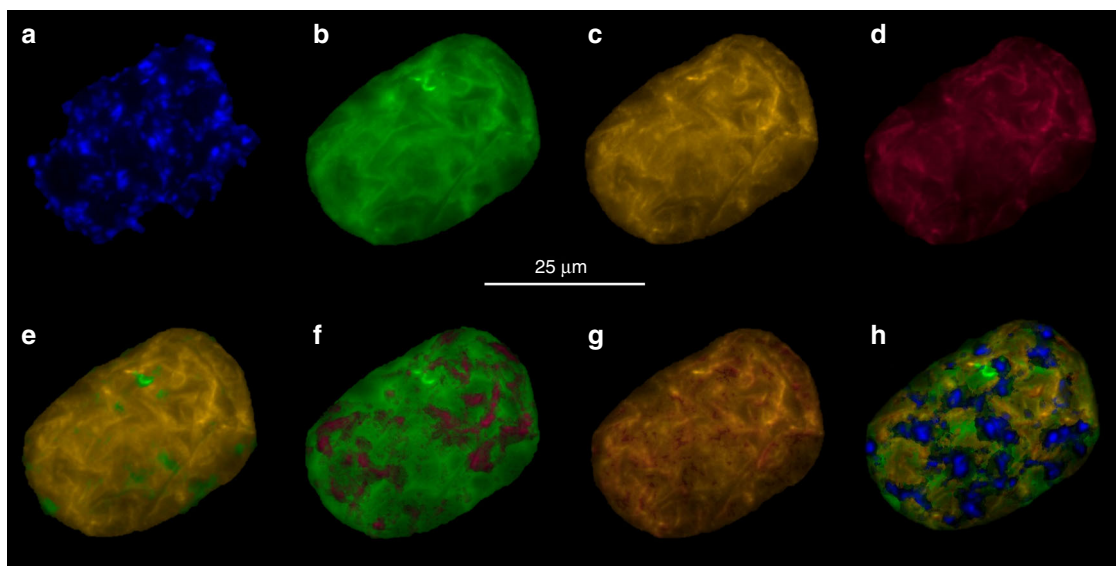


Fig. 3 Fluorescence in situ hybridization. Fluorescence in situ hybridization of a DAPI stained **a** *Achromatium* cell using probes AchroCLII-IV (5'-cgatcgtcgcttgtaggctt-3') **b**, AchroCLIII (5'-cgatcgttgcttgtaggctt-3') **c**, and AchroCLI (5'-ggatcgtcgcttgtaggcca-3') **d**. The used three different probes for *Achromatium* match the clusters in Supplementary Fig. 2A. **e-g** Show overlaid images of 2 probes combinations. An overlay of all probes as well as the DAPI staining is shown in **h**. Results from additional cells are shown in Supplementary Fig. 6

for 5 of these cells, while for the remainder the V5 region was sequenced as a test of contamination prior to single-cell genome sequencing. Both sets were analyzed using the stringent DADA2 R package¹⁹ resulting in 20 and 177 sequence variants, respectively (Fig. 2 and Supplementary Figs. 2–4). Distance based clustering of the same sequences at an identity cutoff of 97% resulted in 1189 and 6909 operational taxonomic units (OTUs), respectively.

We sequenced the genomes of six cells, which resulted in more than 99% *Achromatium* related 16S rRNA gene sequences. From these cells, 10 full or partial 16S rRNA gene sequences were obtained (Fig. 2, Supplementary Fig. 3).

A maximum likelihood phylogenetic analysis of all OTUs overlapping the V1–V4 region alongside previously published *Achromatium* sequences shows no cell-specific clusters (Fig. 2, Supplementary Figs. 2–4). Two earlier sequences reported from Lake Stechlin and assigned to two different species²⁰ appear both as members of a single cluster (cluster 4; Supplementary Fig. 2). This suggests that these sequences belong to members of a single group with elevated genetic diversity. The alignment of the rRNA gene sequences shows that the genetic diversity is concentrated in the hypervariable regions, supporting the idea that this is not the result of random sequencing errors, but of an evolutionary process (Supplementary Fig. 5). This is surprising as the rRNA genes are generally believed to undergo gene conversion resulting in concerted evolution within species^{21, 22}.

Spatially differential expression of rRNA. To confirm the presence and expression of several different rRNA gene alleles in individual cells, two sets of FISH probes were designed: the first based on complete 16S rRNA sequences as assembled from the metagenome, and the second based on the clustering of the V1–V4 amplicons obtained from single cells (Supplementary Fig. 2A). Different cells showed positive signals for 1, 2, or 3 probes of either sets (Fig. 3, Supplementary Figs. 2B and 6), confirming that different *Achromatium* rRNA sequences do not originate from different species but from intracellular diversity. Previous FISH analyses of *Achromatium* from Lake Stechlin have similarly identified cells labeled by one or more probes

designed to target what were thought to be different species¹⁶. It is not unprecedented that bacteria contain multiple and different copies of the rRNA operon with dissimilarities between 16S rRNA genes in one cell up to 11%²³. The FISH signal of the different rRNAs overlapped only partially (Fig. 3, Supplementary Figs. 2B and 6) suggesting that at least in some cases different alleles of the same genes may be expressed in different locations in the cell.

***Achromatium* cells harbor multiple non-identical chromosomes.** The assembly of single-cell genomes ($6 \times \sim 12,000,000$ reads) and metagenomic data ($\sim 96,000,000$ reads) resulted in relatively short contigs ($< 82,000$ nt and $< 56,000$ nt long, respectively). Given the almost absolute presence of *Achromatium* sequences among the 16S rRNA genomic (single cell) and metagenomic reads, the elevated read output would have been sufficient to robustly cover even a large prokaryotic genome (e.g., ~ 2000 -fold coverage of a single 7 Mbp genome). The estimated size of the *Achromatium* genome based on the 6 single cells and the various metagenomic bins ranges between 3.5 and 12 Mbp (Supplementary Data 1). Therefore, we suggest that *Achromatium* does not harbor many identical copies of a single genome (like, e.g., *Epilopiscium*) but rather many diverse chromosomes. This is also supported by the multiple different copies of many of the genes (see below). We hypothesize that our assembled contigs represent the relatively conserved regions in an otherwise highly diverse inter- and intra-cellular (polyploid) chromosomal environment. The genome regions connecting the assembled genes probably vary significantly between chromosomes, resulting in a low sequencing coverage per-variant, insufficient for assembly.

Tetranucleotide binning^{24, 25} coupled to GC distribution and assembly coverage are often used in extracting individual genomes out of metagenomic data and are referred to as binning. The completion and homogeneity of each bin (genome) is evaluated by the percent presence of general or lineage specific sets of “single-copy” marker genes. We have used this technique to evaluate the presence of multiple *Achromatium* species in our samples, each containing a unique genome. Binning of

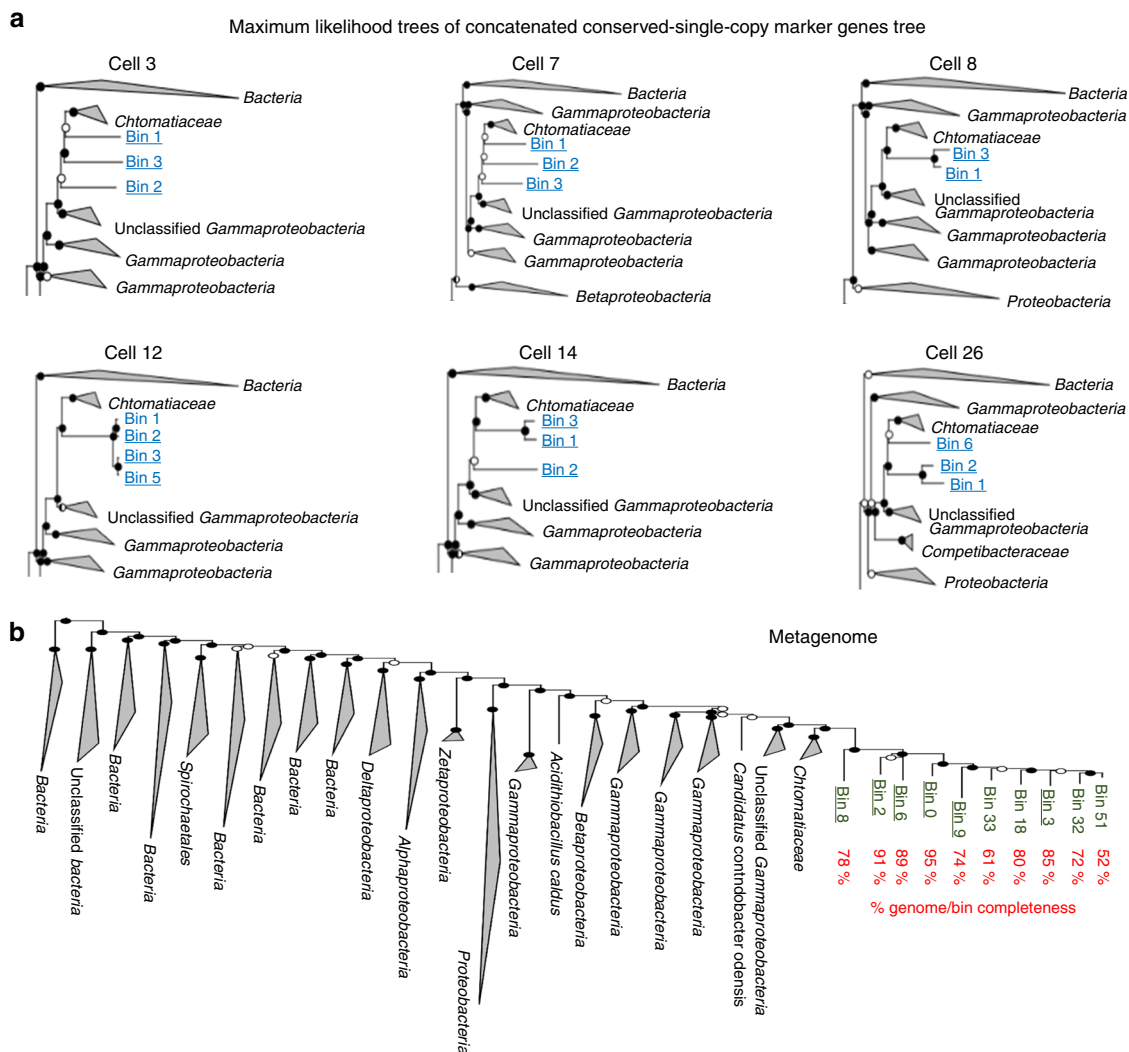


Fig. 4 Maximum likelihood trees calculated from concatenated sequences of conserved single copy marker genes. The genes were found in genes of the 6 sequenced single cells **a** and the different genomic bins of the *Achromatium* metagenome **b**

Table 1 Abundance of various genetic elements in *Achromatium* genomes and metagenomes

	Freshwater/Stechnlin						Saltwater				
	Cell 3	Cell 7	Cell 8	Cell 12	Cell 14	Cell 26	MGM	<i>A. palustre</i>	WMS1	WMS2	WMS3
Insertion sequences	160	145	49	36	383	88	545	27	20	15	85
Transposases (Prokka annotation)	51	47	50	51	93	29	546	19	3	12	27
Origin of replication	1	6	3	1	3	4	19	5	3	5	5

either single-cell genomes or metagenomic data results in multiple bins with a varying degree of completion (<42% for single cells and 52–95% for the metagenome; Supplementary Data 1). Both single cells and metagenomic bins contain multiple copies of some “single copy” marker genes (Supplementary Data 2) with predicted (gene set) duplication levels up to 1.5 times (of 138 genes analyzed; Supplementary Data 1). A similar approach applied to the genomes of four published single *Achromatium* cells from brackish/marine environments (WMS1-3¹⁵ and “*Candidatus Achromatium palustre*”¹² divided the assembled data into 2–4 bins without overlapping marker genes (Supplementary Data 3).

The presence of multiple bins as well as the outcome of phylogenetic trees of concatenated “single copy” marker

genes²⁶ assigned to each bin obtained from the metagenomic or single-cell assemblies could suggest the possible presence of multiple *Achromatium* species whose different genomes can be separated by sequence patterns (Fig. 4). However, several facts rule out this possibility. First, we are confident that single-cell genomes contain only one *Achromatium* cell. Second, the multiple 16S rRNA gene alleles resolved in the metagenome were later found in the single-cell genomes, as well as in the specific amplicon sequences obtained from our single cells. Third, all single-cell genomes contain multiple and different copies of some of the “single copy” marker genes, further supporting our hypothesis of elevated intracellular diversity. This as well confirms that multiple copies of “single copy” marker genes in the metagenomic data must not represent multiple and different



Fig. 5 Maximum likelihood tree of two proteins believed to be single-copy marker genes. The genes encode Arg-tRNA-synthase **a** and ribosomal protein L11 (rplK) **b**. Replicate copies of the genes identified in the same cells are marked by identical shapes. Additional trees of “single copy” marker genes (protein trees) are given in Supplementary Data 4

genomes. Fourth, all metagenomic bins were associated to *Achromatium*, but comparing trees of individual marker genes from different bins showed phylogenetic inconsistencies (i.e., dissimilarity between the trees; Supplementary Data 4). Hence, we hypothesize that the separation into individual genomic/metagenomic bins is a by-product of sequence divergence across chromosomes, and not a consequence of the presence of multiple separate species. Mobile genetic elements such as transposons often use tetranucleotide recognition sites which are disrupted or duplicated upon insertion^{27–29}. The Lake

Stechlin *Achromatium* genomes and metagenome are extremely rich (>90 and >500, respectively) in transposases (Table 1). These are likely responsible for generating diverse tetranucleotide patterns (Fig. 5).

Phylogenetic analysis of “single copy” marker genes shows that individual cells harbor multiple and different copies (e.g., arg-tRNA-synthetase and *rplK* in Fig. 4a, b; for more genes see Supplementary Data 5). The *Achromatium* metagenome further reveals that both inter- and intracellular diversity cover the diversity observed in the overall community.

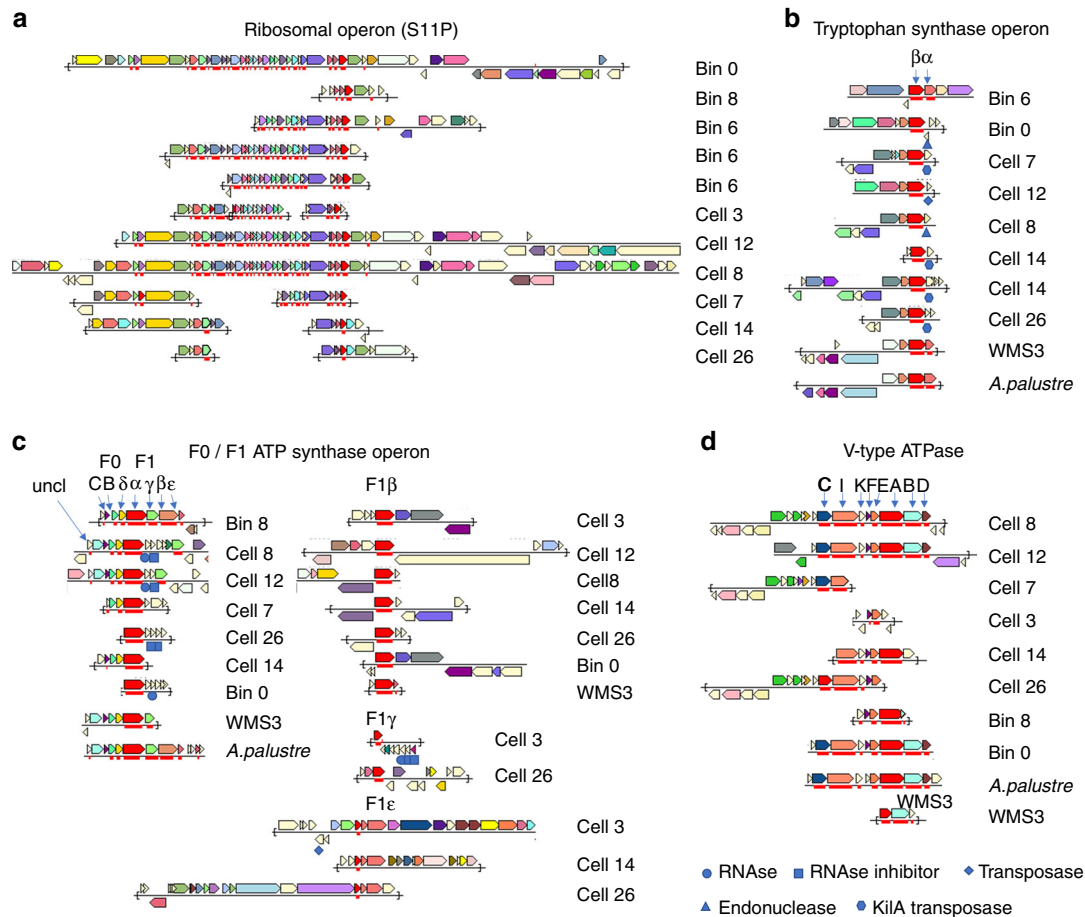


Fig. 6 Gene synteny of four bacterial operons. The operons are: ribosomal **a**, tryptophan synthase **b**, ATP synthase **c**, and V-type ATPase **d**. Additional individual genes from the tryptophan synthase and ATP synthase operons were identified and are not shown here. The data are available on IMG genome IDs 2642422595-9, 2710724216-17 and 2711768587-90

Gene synteny. Gene synteny can be of crucial significance to functionality. While the relocation of individual genes within the genome or between multiple copies of the genome may not have damaging effects, the disruption of operons may be fatal. As such we looked into the gene synteny of the ribosomal protein operons, the tryptophan synthase, the ATP synthase, and the V-type ATPase (Fig. 6), as well as solitary genes such as *recA* (Supplementary Fig. 7).

The gene neighborhood of solitary genes such as *recA* differs between most copies (Supplementary Fig. 7). In contrast, the ribosomal operon gene neighborhood is highly conserved (Fig. 6a). The sequence similarity between copies of ribosomal proteins (e.g., *rplK* Fig. 4b) is higher than in non-operon-based marker genes (Fig. 3b, c, Supplementary Data 6), as previously shown³⁰. This suggests that the conservation pressure differs between different proteins strengthening our hypothesis that the observed phenomenon of intracellular genomic diversity is real and not a methodological artifact. It appears though that not all known operons are fully conserved. The tryptophan synthase and ATP-synthase operons (Fig. 6b, c) have been recovered only in one case as a full set of genes (Fig. 6c). In most other cases the operon was interrupted by known transposable elements, by a gene set consisting of an RNase and RNase inhibitor, or by an endonuclease. Additional single genes belonging to these operons occur individually on other scaffolds; data not shown, available on IMG³¹ with genome IDs 2642422595-9, 2710724216-17, and 2711768587-90. In contrast, the operon of the V-type ATPase seems to be strongly conserved. The effect this has on the

Achromatium functionality is currently not clear. Suzuki et al.³² have shown that for the F0F1 ATP synthase, the separate yet simultaneous expression of some of the genes maintains functionality. Thus, while some operons may be interrupted, they may still function if maintained under similar regulatory factors.

Large genetic diversity within and between *Achromatium* cells. Based on the above results we consider the entire set of assembled sequences associated to *Achromatium* sp. as a “community genome”. A pan-genome comprising all metagenomic sequences attributed to *Achromatium*-associated bins consisted of >3500 proteins, excluding hypothetical genes (Supplementary Data 7). We used two tests to investigate if different proteins are under different evolutionary pressure. In the first test, we calculated for each and across all single-cell genomes the average amino acid distance between copies of proteins that occur more than twice (79–296 proteins, median = 90; Fig. 7a and Supplementary Fig. 8) and for the metagenome the distance between those occurring more than five times (~ 1400 proteins, see Fig. 7b; Supplementary Data 7). In the second test, we calculated from gene copy alignments the average ratio between non-synonymous and synonymous mutations (Ka/Ks) for ~1180 of the above metagenomic proteins which could be annotated using Seed Subsystem³³ (Fig. 7c). Among the single cells, 25% of the proteins have a Dayhoff distance³⁴ smaller than one substitution per site (Fig. 7a). With metagenomic data this number increases to 50% (Fig. 7b) with no correlation between gene copy number and

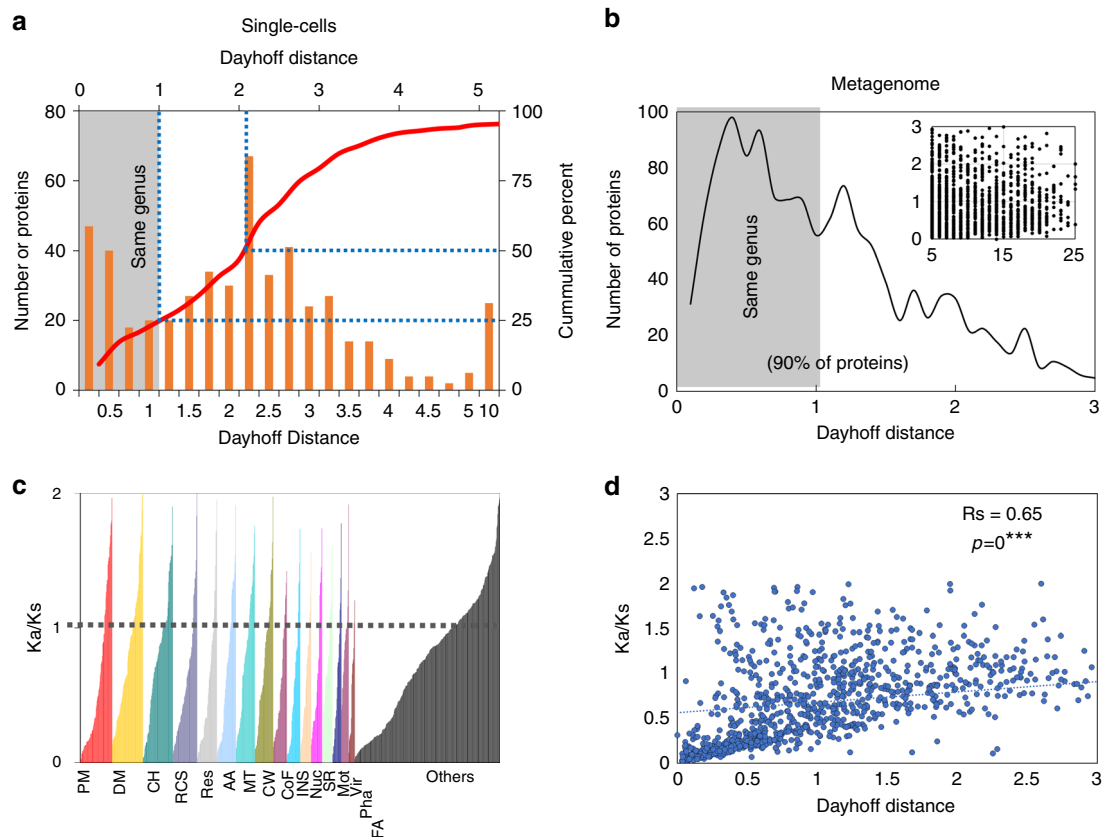


Fig. 7 Distances between multiple copies of the same protein and conservation pressure. **a** Protein numbers ($n = 505$) distribution along the average Dayhoff distance across single cells, between proteins with multiple copies per single cells. The curve shows the cumulative percent of total proteins. A similar distribution across all single cells is shown in Supplementary Fig. 8. **b** Protein numbers ($n = 1400$) distribution along the Dayhoff distance (0-3), showing a decreasing number of proteins as the distance increases. The insert shows the lack of correlation between distance (Y axis) and gene copy number (X axis). **c** Ratio between non-synonymous and synonymous mutations (Ka/Ks) calculated for 1040 proteins grouped based on the first level of the seed subsystem³² hierarchical system. The dashed line at a ratio of 1 marks proteins inferred under conservation pressure (Ka/Ks < 1), stable (Ka/Ks = 1), and evolving (Ka/Ks > 1). The seed groups are labeled as follows: AA amino acids and derivatives; CH carbohydrates; CoF cofactors, vitamins, prosthetic groups, pigments; CW cell wall and capsule; DM DNA metabolism; FA fatty acids, lipids, and isoprenoids; INS iron, nitrogen and sulfur metabolism (small groups combined for plotting purposes); Mot motility and chemotaxis, MT membrane Transport; Nuc nucleosides and nucleotides; Others smaller groups and non-classifiable proteins; Pha phages, prophages, transposable elements, Plasmids; PM protein metabolism; RCS regulation and cell signaling; Res respiration; SR stress response; Vir virulence, disease and defense. **d** Correlation between protein distance and the Ka/Ks ratio (Spearman $R_s = 0.64$)

distance (Fig. 7b, insert). This highlights the high genetic diversity within the community which can be even higher when looking at its individual members. A distance of one substitution per amino acid site has been shown to be at the upper limit of protein distance between species of the same genus. However, a distance of three amino acids substitutions per site would normally suggest a lower phylogenetic relation³⁵. Interestingly, there is not a big difference between overall diversity patterns among single cells and within individuals (Fig. 4a and Supplementary Fig. 8). The salt water *Achromatium* cells/populations have a similar broad diversity among the multiple-copy proteins (Supplementary Fig. 8), despite having single rRNA alleles rather than multiple ones as our freshwater cells have. The Ka/Ks ratio is lower than 0.5 for 410/1180 proteins (lower than 1.0 for 775/1180 proteins), i.e., under evolutionary pressure for conservation. No correlation was found between the Ka/Ks ratio and gene copy number. Hierarchical classification of the proteins based on the Seed Subsystem³³ classification (Level 1) shows no group specific trends. Overall, no specific protein group seems under stronger or weaker conservation pressure. The correlation between the Ka/Ks ratio and the averaged protein distance (Fig. 7d) confirms that this is not a sequencing artifact.

Using freshwater and marine single-cell genomes as well as metagenomic bins (Fig. 4a) we conducted amino acid identity and average nucleotide identity analyses³⁶. Overlaying the results (Supplementary Fig. 9) with the thresholds of Rodriguez-R and Konstantinidis³⁷ places most of freshwater genomes and bins as different species in the same genus (similarly as when compared to the salt water cells). Nevertheless, all the data presented thus far suggest that the genetic diversity of freshwater *Achromatium* from Lake Stechlin does not originate from multiple species. Instead, we propose that each individual *Achromatium* cell harbors genetic diversity at a level typically found between species of the same genus.

Suggested mechanism of genetic diversity generation. Our observations are consistent with multiple chromosomes within cells frequently recombining and undergoing rearrangement, possibly through homologous and non-homologous recombination and transposition. The latter is plausibly responsible for the variability in gene synteny observed in genomic and metagenomic data of *Achromatium* cells from Lake Stechlin. Over 540 different transposases were identified in the metagenome, and up to 90 in single cells, several times more

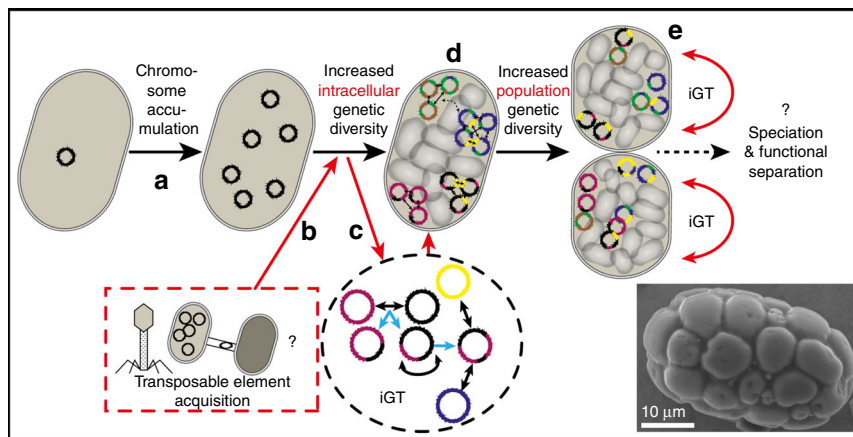


Fig. 8 Proposed model for generation of large intracellular and subsequent population genetic diversity. Genome replication in an ancestral *Achromatium* cell **a** is followed by the acquisition of an unusually high number of transposases via undetermined mechanisms (e.g., phages, horizontal gene transfer); **b**. The genomes in the polyplod cell undergo a process of intracellular gene transfer (iGT) with limited transfer between cell parts **c** due to the multiple calcite bodies which can be seen in the insert. The new chimeric genomes further interact increasing the cellular diversity while maintaining conserved regions such as operons intact **c**. Overall this results in a cell with multiple different chromosomes; **d**, represented by the different colors of each chromosome. Gene conversion does not occur on a cellular level but maintains genome stability at a local scale and thus fixing multiple version of the same gene in different cellular locations. A dividing cell recruits part of the genomic material generating two daughter cells different from each other as well as from the mother cell; **e**. Scale bar in SEM insert is 10 μm

than other *Achromatium* species (Table 1; Supplementary Data 7). This large intra-genus difference is not unprecedented. Sequencing of six *Crocospaera* isolates from two different phenotypic groups found one isolate containing over 1200 transposases. Other isolates from the same phenotype had up to 165 transposases, while isolates from the other phenotype had up to 223³⁸. In contrast to previously studied *Achromatium* genomes^{12, 15}, insertion sequences (IS)—often encoding the transposase genes³⁹ were found in high abundance in our single cell as well as metagenomic data of Lake Stechlin. The number of identified ISs in the similarly annotated genomes of the brackish water *Achromatium*, however, was far lower (Table 1). Increased numbers of IS elements have been associated with environmental stress^{40, 41}, possibly enhancing genetic diversity by mutagenesis.

Transposable elements can explain the intracellular gene shuffling and part of the observed increased divergence. However, an additional mechanism is needed to explain the large genetic diversity in Lake Stechlin (and saltwater) *Achromatium* cells. In fact, a main disadvantage of polyplod, asexual organisms, is that new mutations affecting a few chromosome copies within a cell are thought to have limited phenotypic consequences, and so mutations that would be strongly deleterious if fixed within a cell, can be tolerated for many generations⁴². However, these mutations can go to fixation many generations after their occurrence in a cell, leading to a sudden decrease in fitness (replication load). A way to escape this replication load can be strong gene conversion leading new mutations to either fixation or extinction within cells. This mechanism is not likely to occur in *Achromatium* at the whole cell level, because it would lead to very little within-cell diversity, as observed, e.g., in *Epulopiscium*⁹ and *Haloferax volcanii*⁶. Nevertheless, as visible in Fig. 1 and Supplementary Fig. 1, the *Achromatium* chromosomes (DNA spots) occur in separate clusters. Thus, gene conversion as well as evolutionary selection might take place in separate genetic hotspots (containing clusters of chromosomes) that are minimally or not at all interacting with each other. This would lead to fixation of different neutral and positive mutations in each “compartment”. Cellular rearrangement upon cell division would lead to the propagation of these mutations and the fixation of new ones. The Ka/Ks analysis (Fig. 4b) shows that

most of the proteins analyzed are under evolutionary pressure. This suggests that a gene-conversion-like process might occur, however, on a local (compartment) rather than a global (whole cell) scale.

Another mechanism for polyplod prokaryotes to escape extinction is between-cell recombination (or horizontal gene transfer (HGT)), that could reintroduce beneficial variants lost within cells⁴². HGT can occur via conjugation, viral infection, gene-transfer-agents and the uptake of naked DNA from the environment. We only found the genes required for the latter across the *Achromatium* single-cell genomes and metagenome (e.g., the *comEA* genes and the assembly genes for type IV pili). Thus, while *Achromatium* might have the ability to uptake genetic material from the surrounding environment, HGT between adjacent cells is most likely not the major mechanism leading to the large observed genetic variation.

We propose that following a process of genome replication and accumulation in an ancestral cell of the modern *Achromatium* (Fig. 8a), the acquisition of multiple transposases (Fig. 8b) led to a continuous, intracellular exchange of genomic sequences between individual genomes of the polyplod cell (i.e., intracellular gene transfer (iGT); Fig. 8c). This resulted in a mosaic-like genome (Fig. 8d) consisting of conserved regions (genes and operons, e.g., Fig. 8d, e), which are separated by less conserved spacers. A possible indication of a viral source for these transposases can be found in numerous of the KilA-N domain containing proteins⁴³. The spatial compartmentalization of these genomes alongside possible uptake of DNA from the environment, allows for the formation of multiple stable versions of similar proteins. We further suggest that the genomic diversity of single cells leads to a highly heterogeneous community as dividing cells, which likely split into two genetically different (yet functionally similar) entities (Fig. 8e). Comparing the metagenomic data with the DoriC database⁴⁴, we identified 23 potential origins of replication (Table 1). Interestingly, this number is similar to the number of different copies of the *dnaN* (DNA polymerase III beta subunit; 22 copies) gene that is found adjacent to the OriC and the replication initiator *dnaA*, of which 28 copies have been identified (Supplementary Data 7). Among the single cells <6 potential OriC were identified per genome both in the freshwater and

saltwater cells. It has already been proposed that some bacteria harbor more than one origin of replication on a single chromosome⁴⁵ and this has also been successfully achieved by genome engineering⁴⁵. Thus, it is highly plausible that different genomes (chromosomes) in a single polyploid cell (such as *Achromatium*) harbor different origins of replication. This may allow the cell to regulate the number of different chromosomes present at any given time and subsequently, if different chromosomes harbor different proteins or protein copies with different properties, the cells may be able to tune their expression by enhancing the available number of DNA templates.

Evolutionary and ecological significance of iGT. This proposed mechanism of genome evolution (iGT) has multiple implications for our understanding of microbial evolution and the role of polyploidy. First, as our comparative analysis shows, high intracellular genomic diversity is not a common feature among all members of the *Achromatium* genus (i.e., saltwater vs. freshwater). Nevertheless, a higher than expected diversity among replicate copies of proteins within a cell is observed across the lineage. Interestingly, this is not uniformly depicted by the rRNA marker gene. More genome sequencing from different freshwater and marine environments is needed to understand the full extent of this phenomenon.

It is suggested that one benefit of polyploidy is to allow for the generation of “experimental” versions of functioning proteins (neofunctionalization)^{46, 47}. As long as one (or more) functional copy of a given protein is present, cells can putatively allow divergence to accumulate in alternative copies leading to new specialized functions. The presence of a high number (up to almost 100) of multiple different copies of the same gene in our metagenomic data set suggests that this process might be predominant in *Achromatium*. It remains to be determined which and how many of such gene copies are indeed functional. Interestingly, we also detected multiple copies of group II introns (Supplementary Data 7), a feature that has been shown in the polyploid *Thiomargarita*⁴⁸, hinting towards eukaryotic-like alternative splicing.

Thus, can *Achromatium* cells represent an intermediate evolutionary state between uni- and multicellular life? Multicellular bacteria differ from their eukaryotic counterparts. Often these are chained single cells (e.g., filamentous *Cyanobacteria* and the large sulfur bacterium *Beggiatoa*) each of which can survive alone and give rise to a new filament. In contrast, different cells in a multicellular eukaryotic organism have various functions, a phenomenon which is rare in the prokaryotic domains (exceptions are, for example, heterocysts of N₂-fixing filamentous cyanobacteria⁴⁹). *Achromatium* cells appear to have multiple compartments¹². Inside of those, chromosomes might be independently replicated, and may serve as the basis of functional compartmentalization and multicellularity. Functional compartmentalization has been described in several bacteria⁵⁰ and compartment-specific gene expression has been described in spore-forming *Bacillaceae*⁵¹. At present, no indication exists as to differential expression of genes in different parts of the *Achromatium* cell. Some of the FISH images (Fig. 2b, Supplementary Figs. 2 and 6) suggest that some expression patterns may exist, but this needs further evaluation in more targeted studies. Additionally, Markov and Kasnachev⁴² suggest that, similarly to Lake Stechlin *Achromatium*, the proto-Eukaryote cell may have been a rapidly mutating, chromosome diversifying, polyploid Bacteria/Archaea.

To conclude, we present evidence for increased intracellular genomic diversity in single cells of the freshwater, large sulfur bacterium *Achromatium* from Lake Stechlin. We propose

that intra and inter-cellular gene transfer (iGT and HGT), chromosomal rearrangement, and compartmentalized gene conversion are responsible for this phenomenon. This is supported by the high abundance of transposable elements and the presence of multiple, mostly conserved, versions of the same proteins. Elevated genetic diversity might provide *Achromatium* with exceptional potential for fast adaptation. The extent and functional significance of this phenomenon in permanently or temporarily polyploid bacteria remains to be determined.

Methods

Cell collection. Samples were collected from Lake Stechlin in Germany (53.1520 °N, 13.0233 °E) on several occasions during 2015 and 2016. Surface sediment from the lake's shore at a water depth of ca. 1 m were allowed to settle for a few hours in a glass beaker, after which the surface layer was sieved through a 200 µm hole size mesh onto a glass plate. Single cells were picked under a binocular and transferred to a clean vial by using the different sedimentation of *Achromatium oxaliferum* cells and sand grains when applying a rotational movement. The cells were further transferred until no further coarse contaminants were visible. See details below for single-cell processing.

Achromatium cells are covered in a layer of extracellular polymeric substances to which epibiotic bacteria were attached. To remove this layer and thus minimize foreign DNA contamination, the concentrated cells were washed for 10 min in a 100 mM NaHCO₃ buffer.

Cells for DNA extraction were frozen at -20 °C until further treatment. Samples for fluorescent in situ hybridization were fixed for 14 h at 4 °C in 1% formaldehyde solution (final concentration).

Fluorescence in situ hybridization and DNA staining. DNA staining with SybrGreen I or PicoGreen was performed on non-fixed cells with a mixture of 5 µl of the stock solution in 200 µl Sigma Mowiol plus 5 µl freshly prepared 1 M ascorbic acid in 1× TAE buffer.

For fluorescence in situ hybridization, fixed *Achromatium* cells were washed with distilled water allowing the cells to precipitate prior to water removal. The clean, fixed cells were placed in polylysine covered microscope slides. Fluorescence in situ hybridization was done as previously described⁵² using a 10% and 30% formamide hybridization buffer for the Achro664 and AchroCL probes respectively. Hybridization was carried out for 12–14 h at 46 °C. The probes Achro664I (5'-gcttgctagactacgaag-3'), Achro664II (5'-gcaagctagactacgaag-3') and Achro664III (5'-gctgagctagactacggag-3') were designed using the full-length 16S rRNA sequences obtained from the metagenome. The probes were labeled with 6-FAM, Cy-3, and Cy-5, respectively, and were generated by Biomers (Ulm, Germany). The probes AchroCLI (5'-ggatcgtcgccttgtaggcca-3'), AchroCLIII (5'-cgatcgttgcttgtaggcca-3'), and AchroCLII-IV (5'-cgatcgtcgccttgtaggcca-3') were designed based on the V1–V4 regions of the 16S rRNA gene and start at position 263. The probes were labeled with Cy-3, Cy-5, and 6-FAM, respectively, and were generated by Biomers (Ulm, Germany) and were doubly labeled for enhanced signal.

Images were taken using an Axiovision microscope (Zeiss). An autofluorescence signal was observed in all channels although it required exposure times for image acquisition at least 10 times longer than for the FISH signal. Hybridization experiments using the EUB (I,II,III) probe mix and the non-EUB probe produced, positive and negative signals, respectively.

Scanning electron microscopy. The scanning electron microscopy image was taken using a Jeol JCM-6000 using 15 kV at high-vacuum at a magnification of X1500. The *Achromatium* cells were dried for 3 h 30 °C, placed in an extainer for 15 h, and were sputtered with Au-Pd for 5 min.

Metagenomics. The fragile nature of *Achromatium* was used to extract DNA without additional chemical steps. Circa 10,000 cells were placed in 500 µl water and disrupted on a vortex machine for 30 s after which they were centrifuged in a table top centrifuge (Fresco 17, Thermo Fisher) for 5 min at 17,000×g. The DNA was precipitated with 2 vol. of 29:1 ethanol:sodium acetate (3 M) at -20 °C overnight followed by 30 min centrifugation at 17,000×g at 4 °C. The DNA pellet was washed with ice cold 70% ethanol, centrifuged at 17,000×g for 5 min, dried, and re-suspended in water.

Sequencing was carried out at Molecular Research Laboratories (Mr. DNA), Shallowater, Texas on an Illumina HiSeq resulting in 96,000,000 paired end reads (2 × 151 nucleotides). Metagenome sequencing steps included DNA fragmentation, ligation to sequencing adapters, and purification. Following the amplification and denaturation steps, libraries were pooled and sequenced. DNA (50 ng) from each sample was used to prepare the libraries using Nextera DNA Sample Preparation Kit (Illumina). Library insert size was determined by Experion Automated Electrophoresis Station (Bio-Rad). The insert size of the libraries ranged from 300 to 850 bp (average 500 bp). Pooled library (12 pM) was loaded to a 600 Cycles v3

Reagent cartridge (Illumina) and the sequencing was performed on HiSeq (Illumina).

Raw sequence data was quality trimmed using the Neson Clip tool (<http://www.vicbioinformatics.com/software.neson.clip.shtml>). To remove the non-uniform coverage caused by the genome amplification step the samples were normalized to a coverage of 100-fold using the BBnorm tool (<https://sourceforge.net/projects/bbmap/>). The normalized sequences were assembled using SPAdes⁵³ version 3.7 including the implemented error correction tools. Note that due to SPAdes being specifically designed for assembly of single-cell data, no major differences in assembly were obtained when the normalized and raw data were used.

The assembled data were binned using Metawatt²⁴ version 3.5.2. Bins not associated to *Chromatiaceae* were not used for further analysis. A similar binning approach was applied to four published genomes of *Achromatium*^{12, 15}. Metawatt was also used to extract the sequences of all identified single-copy marker genes as well as partial 16S rRNA sequences.

The combined bins associated with *Achromatium* were analyzed using the Prokka pipeline⁵⁴. Selected bins or bin clusters (as underlined in Fig. 4) were also uploaded to the IMG-ER³¹ system for annotation.

Percent completion of the single cells genomes as well as of the metagenomic bins alongside duplication level were analyzed using Metawatt²³. These data were confirmed by analysis of the same data using CheckM⁵⁵ with the Chromatiaceae marker-genes set, which also provided the estimated genome size.

For the analysis of coding sequences, all copies of individual genes/proteins were extracted and aligned using MUSCLE⁵⁶. Trees showing the diversity within each protein/gene copy were calculated using FastTree⁵⁷ version 2.1. Specific trees were calculated using RAXML⁵⁸. Distance analysis and non-synonymous/synonymous ratio were calculated using the megacc command line version of MEGA⁵⁹ version 7. Distances between trees of proteins were calculated using the TreeDist program of the PHYLIP package.

16S rRNA sequence analysis from the raw metagenomics data were conducted using phyloflash (<https://github.com/HRGV/phyloFlash>).

Single-cell genomics. *Achromatium* cells were collected as mentioned above in two separate occasions. In the first attempt, single cells were placed in a 2 ml tube containing 15 µl of MilliQ water. Five cells were selected for sequencing based on differences in morphology or cells size: a small, medium, and large sized cell as well as two dividing cells. Cells were sequenced at Molecular Research Laboratories (Mr. DNA), Shallowater, Texas, where DNA was extracted by centrifugation. Thanks to the fragile nature of the cells they break when centrifuged allowing the genomic DNA to spill out. Genomic DNA was amplified using the REPLI-g Single Cell kit (Qiagen).

Single-cell genomes were obtained as above; however, the reads were dominated (ca. 90%) by epibionts.

The DNA was then used for 16S rRNA amplicon sequencing on the Illumina MiSeq platform. The 16S rRNA gene V1-3 variable region PCR primers 28f/515r⁶⁰ with barcodes on the forward primer were used in a 28 cycle PCR (five cycles used on PCR products) using the HotStarTaq Plus Master Mix Kit (Qiagen, USA) under the following conditions: 94 °C for 3 min, followed by 28 cycles of 94 °C for 30 s, 53 °C for 40 s, and 72 °C for 1 min, after which a final elongation step at 72 °C for 5 min was performed. After amplification, PCR products were checked by agarose gel electrophoresis to determine the success of amplification and the relative intensity of bands. Multiple samples were pooled (e.g., 100 samples) in equal proportions based on their molecular weight and DNA concentrations. Pooled samples were purified using calibrated Ampure XP beads and the purified PCR product used to prepare the Illumina DNA library. Sequencing was performed following the manufacturer's guidelines.

In a second attempt to obtain clean genomes of single *Achromatium* cells, 50 NaHCO₃ treated single cells were collected into separate 2 ml tube pre-filled with sterile PCR grade water. Each cell was then transferred 3 times into a new similar 2 ml tube. Last, each cell was transferred to a new empty tube in a total volume of ~1 µl water. Of these 25 cells were selected for whole-genome amplification using the Illustra Single Cell GenomiPhi DNA Amplification Kit (GE Healthcare Europe GmbH, Freiburg, Germany) resulting in 22 successful amplification which were sent for 16S rRNA gene sequencing at Molecular Research Laboratories (Mr. DNA), Shallowater, Texas using a PGM machine with primer set 515F-806R. The resulting sequences were analyzed using the DADA2 R¹⁹ package. Six cells with >99% of their 16S rRNA gene sequences attributed to *Achromatium* were selected for single-cell genome sequencing. The sequencing was carried out as described above.

Single-cell genomes were assembled using SPAdes⁵³ (V 3.9) and binned using Metawatt²⁴. Genomic bins associated with *Chromatiaceae* were extracted, the raw reads were mapped to the contained contigs and reassembled. The final scaffolds were annotated using Prokka⁵⁴, RAST⁶¹, and IMG-ER³¹.

Amplicon data analysis. The 16S rRNA amplicon sequences were analyzed in full by three independent pipelines: the proprietary pipeline of the sequencing company based on QIIME⁶², the SILVA NGS pipeline⁶³, and the DADA2 R package¹⁹. For the first, sequences were joined, depleted of barcodes, and sequences < 150 bp or with ambiguous base calls were removed. Sequences were denoised, OTUs generated, and chimeras removed. OTUs were defined by

clustering at 3% divergence (97% similarity). Final OTUs were taxonomically classified using BLASTn against a curated database derived from RDPII and NCBI (www.ncbi.nlm.nih.gov; <http://rdp.cme.msu.edu>). The SILVA pipeline was supplied with a FASTA file containing the assembled reads as provided by the sequencing company.

The analysis using the DADA2 R package¹⁹ was conducted using the raw reads and the default parameters.

The *Achromatium* representative sequences together with all available 16S rRNA *Achromatium* sequences in the database were aligned using the online version of SINA⁶⁴. Phylogenetic trees were calculated using RAXML (v 8.2.8)⁵⁸, the GTR model, and 100 bootstrap analyses.

Data availability. All sequence data are available at the European Nucleotide Archive (ENA) under project number PRJEB14545/ERP016191 (available through the ENA ftp site). Annotated data can be accessed on IMG genome IDs 2642422595-9, 2710724216-17, 2711768587-90 and on MG-RAST under id mgp11828.

Received: 2 September 2016 Accepted: 22 June 2017

Published online: 06 September 2017

References

- Zerulla, K. & Soppa, J. Polyploidy in haloarchaea: advantages for growth and survival. *Front. Microbiol.* **5**, 274 (2014).
- Comai, L. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846 (2005).
- Soppa, J. Polyploidy in archaea and bacteria: about desiccation resistance, giant cell size, long-term survival, enforcement by a eukaryotic host and additional aspects. *J. Mol. Microbiol. Biotechnol.* **24**, 409–419 (2015).
- Slade, D., Lindner, A. B., Paul, G. & Radman, M. Recombination and replication in DNA repair of heavily irradiated *Deinococcus radiodurans*. *Cell* **136**, 1044–1055 (2009).
- Oliverio, A. M. & Katz, L. A. The dynamic nature of genomes across the tree of life. *Genome Biol. Evol.* **6**, 482–488 (2014).
- Lange, C., Zerulla, K., Breuert, S. & Soppa, J. Gene conversion results in the equalization of genome copies in the polyploid haloarchaeon *Haloferax volcanii*. *Mol. Microbiol.* **80**, 666–677 (2011).
- Infante, J. J., Dombek, K. M., Rebordinos, L., Cantoral, J. M. & Young, E. T. Genome-wide amplifications caused by chromosomal rearrangements play a major role in the adaptive evolution of natural yeast. *Genetics* **165**, 1745–1759 (2003).
- Maere, S. et al. Modeling gene and genome duplications in eukaryotes. *Proc. Natl Acad. Sci. USA* **102**, 5454–5459 (2005).
- Mendell, J. E., Clements, K. D., Choat, J. H. & Angert, E. R. Extreme polyploidy in a large bacterium. *Proc. Natl Acad. Sci. USA* **105**, 6730–6734 (2008).
- Salman-Carvalho, V., Fadeev, E., Joye, S. B. & Teske, A. How clonal is clonal? genome plasticity across multicellular segments of a 'candidatus maritrix sp.' filament from sulfidic, briny seafloor sediments in the Gulf of Mexico. *Front. Microbiol.* **7**, 1173 (2016).
- Babenzien, H.-D., Glöckner, F. O. & Head, I. M. in *Bergey's Manual of Systematics of Archaea and Bacteria* 1–8 doi:10.1002/9781118960608.gbm01222 (John Wiley & Sons, Ltd, 2015).
- Salman, V. et al. Calcite-accumulating large sulfur bacteria of the genus *Achromatium* in Sippewissett Salt Marsh. *ISME J.* **9**, 2503–2514 (2015).
- Gray, N. D. in *Inclusions in Prokaryotes* 299–309, doi:10.1007/3-540-33774-1_11 (Springer-Verlag, 2006).
- Gray, N. D. & Head, I. M. in *The Prokaryotes* (eds Rosenberg, E., DeLong, E. F., Lory, S., Stackebrandt, E. & Thompson, F.) 1–14 doi:10.1007/978-3-642-38922-1 (Springer Berlin Heidelberg, 2014).
- Mansor, M., Hamilton, T. L., Fantle, M. S. & Macalady, J. L. Metabolic diversity and ecological niches of *Achromatium* populations revealed with single-cell genomic sequencing. *Front. Microbiol.* **6**, 822 (2015).
- Head, I. M., Gray, N. D., Babenzien, H. D. & Oliver Glöckner, F. Uncultured giant sulfur bacteria of the genus *Achromatium*. *FEMS Microbiol. Ecol.* **33**, 171–180 (2000).
- Angert, E. R. DNA replication and genomic architecture of very large bacteria. *Annu. Rev. Microbiol.* **66**, 197–212 (2012).
- Darzynkiewicz, Z. in *Current Protocols in Cytometry* doi:10.1002/0471142956.cy0702s56 (John Wiley & Sons, Inc., 2011).
- Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
- Glöckner, F. O., Babenzien, H. D., Wulf, J. & Amann, R. Phylogeny and diversity of *Achromatium oxaliferum*. *Syst. Appl. Microbiol.* **22**, 28–38 (1999).
- Stewart, F. J. & Cavanaugh, C. M. Intra-genomic variation and evolution of the internal transcribed spacer of the rRNA operon in bacteria. *J. Mol. Evol.* **65**, 44–67 (2007).

22. Pei, A. et al. Diversity of 23S rRNA genes within individual prokaryotic genomes. *PLoS ONE* **4**, e5437 (2009).
23. Větrovský, T. & Baldrian, P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE* **8**, e57923 (2013).
24. Strous, M., Kraft, B., Bisdorf, R. & Tegetmeyer, H. E. The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front. Microbiol.* **3**, 410 (2012).
25. Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**, 26 (2014).
26. Campbell, J. H. et al. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc. Natl Acad. Sci. USA* **110**, 5540–5545 (2013).
27. Wang, H. G. & Fraser, M. J. TTAA serves as the target site for TFP3 lepidopteran transposon insertions in both nuclear polyhedrosis virus and *Trichoplusia ni* genomes. *Insect. Mol. Biol.* **1**, 109–116 (1993).
28. Fraser, M. J., Cary, L., Boonvisudhi, K. & Wang, H. G. Assay for movement of Lepidopteran transposon IFP2 in insect cells using a baculovirus genome as a target DNA. *Virology* **211**, 397–407 (1995).
29. Chandler, M. et al. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nat. Rev. Microbiol.* **11**, 525–538 (2013).
30. Fu, Y., Deiorio-Haggar, K., Anthony, J. & Meyer, M. M. Most RNAs regulating ribosomal protein biosynthesis in *Escherichia coli* are narrowly distributed to Gammaproteobacteria. *Nucleic Acids Res.* **41**, 3491–3503 (2013).
31. Chen, I.-M. A. et al. IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.* **45**, D507–D516 (2017).
32. Suzuki, T., Ozaki, Y., Sone, N., Feniouk, B. A. & Yoshida, M. The product of uncl gene in F1Fo-ATP synthase operon plays a chaperone-like role to assist c-ring assembly. *Proc. Natl Acad. Sci. USA* **104**, 20776–20781 (2007).
33. Overbeek, R. et al. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res.* **42**, D206–D214 (2014).
34. Dayhoff, M. O., Schwartz, R. N. & Orcutt, B. C. A model of evolutionary change in proteins. *Atlas Protein Seq. Struct* **5**, 345 (1978).
35. Grishin, N. V. From complete genomes to measures of substitution rate variability within and between proteins. *Genome Res.* **10**, 991–1000 (2000).
36. Konstantinidis, K. T. & Tiedje, J. M. Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.* **187**, 6258–6264 (2005).
37. Rodriguez-R, L. M. & Konstantinidis, K. T. Bypassing cultivation to identify bacterial species. *Microbe Mag.* **9**, 111–118 (2014).
38. Bench, S. R. et al. Whole genome comparison of six *Crocospaera watsonii* strains with differing phenotypes. *J. Phycol.* **49**, 786–801 (2013).
39. Mahillon, J. & Chandler, M. Insertion sequences. *Microbiol. Mol. Biol. Rev.* **62**, 725–774 (1998).
40. Touchon, M. & Rocha, E. P. C. Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.* **24**, 969–981 (2007).
41. Foster, P. L. Stress-induced mutagenesis in bacteria. *Crit. Rev. Biochem. Mol. Biol.* **42**, 373–397 (2008).
42. Markov, A. V. & Kaznacheev, I. S. Evolutionary consequences of polyploidy in prokaryotes and the origin of mitosis and meiosis. *Biol. Direct* **11**, 28 (2016).
43. Iyer, L. M., Koonin, E. V. & Aravind, L. Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. *Genome Biol.* **3**, RESEARCH0012 (2002).
44. Gao, F., Luo, H. & Zhang, C.-T. DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. *Nucleic Acids Res.* **41**, D90–D93 (2013).
45. Gao, F. Bacteria may have multiple replication origins. *Front. Microbiol.* **6**, 324 (2015).
46. Roth, C. et al. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J. Exp. Zool. B Mol. Dev. Evol.* **308B**, 58–73 (2007).
47. Andersson, D. I. & Hughes, D. Gene amplification and adaptive evolution in bacteria. *Annu. Rev. Genet.* **43**, 167–195 (2009).
48. Salman, V., Amann, R., Shub, D. a. & Schulz-Vogt, H. N. Multiple self-splicing introns in the 16S rRNA genes of giant sulfur bacteria. *Proc. Natl Acad. Sci. USA* **109**, 4203–4208 (2012).
49. Kumar, K., Mella-Herrera, R. A. & Golden, J. W. Cyanobacterial heterocysts. *Cold Spring Harb. Perspect. Biol.* **2**, a000315 (2010).
50. Cornejo, E., Abreu, N. & Komelli, A. Compartmentalization and organelle formation in bacteria. *Curr. Opin. Cell Biol.* **26**, 132–138 (2014).
51. Hilbert, D. W. & Piggot, P. J. Compartmentalization of gene expression during *Bacillus subtilis* spore formation. *Microbiol. Mol. Biol. Rev.* **68**, 234–262 (2004).
52. Fuchs, B. M., Pernthaler, J. & Amann, R. in *Methods for General and Molecular Microbiology* (eds Reddy, C. A. et al.) 886–896 (SM Press, 2007).
53. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
54. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
55. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
56. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
57. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
58. Stamatakis, A. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
59. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
60. Lane, D. in *Nucleic Acid Techniques in Bacterial Systematics* (eds Stackebrandt, E. & Goodfellow, M.) 115–175 (John Wiley and Sons, 1991).
61. Aziz, R. K. et al. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
62. Caporaso, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
63. Ionescu, D. et al. Microbial and chemical characterization of underwater fresh water springs in the Dead Sea. *PLoS ONE* **7**, e38319 (2012).
64. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).

Acknowledgements

The positions of D.I. and M.B.-I. were financed through the DFG Aquameth (GR 1540/21-1) and BMBF BIBS projects. We thank Dieter Babenzien for the introduction to *Achromatium* ecology and advice on cell collection. Further, we acknowledge Lara Sabelhaus for collecting and cleaning cells, Katrin Attermeyer for graphical assistance, Reingard Rossberg for SEM images, Sina Schorn, Verena Salman-Carvalho, David Walsh, Mark Dopson, Wolfgang Hess and Alicia Muro-Pastor for fruitful discussions and for critically commenting on this manuscript.

Author contributions

D.I.: concept, data generation, data analysis, and wrote the paper. M.B.-I.: data generation, data analysis, and wrote the paper. N.D.M.: data analysis and wrote the paper. H.C.: concept, data generation, and wrote the paper. H.-P.G.: concept, data generation, and wrote the paper.

Additional information

Supplementary Information accompanies this paper at doi:10.1038/s41467-017-00342-9.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons

Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017