Title

ConStrains identifies microbial strains in metagenomic datasets.

Journal

Nature biotechnology, 33(10)

ISSN

1087-0156

Authors

Luo, Chengwei

Knight, Rob

Siljander, Heli

et al.

Publication Date

2015-10-01

DOI

10.1038/nbt.3319

Peer reviewed

HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2016 April 01.

Published in final edited form as:

Nat Biotechnol. 2015 October ; 33(10): 1045–1052. doi:10.1038/nbt.3319.

ConStrains identifies microbial strains in metagenomic datasets

Chengwei Luo[1,2,3], **Rob Knight**[4,5,11], **Heli Siljander**[6,7], **Mikael Knip**[6,7,8,9], **Ramnik J Xavier**[1,2,3,10], and **Dirk Gevers**[1,12]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA [2]Gastrointestinal Unit and Center for the Study of Inflammatory Bowel Disease, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA [3]Center for Computational and Integrative Biology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA [4]Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder, Colorado, USA [5]Howard Hughes Medical Institute, Boulder, Colorado, USA [6]Children's Hospital, University of Helsinki and Helsinki University Hospital, Helsinki, Finland [7]Research Programs Unit, Diabetes and Obesity, University of Helsinki, Helsinki, Finland [8]Folkhälsan Research Center, Helsinki, Finland [9]Department of Pediatrics, Tampere University Hospital, Tampere, Finland [10]Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Abstract

An important fraction of microbial diversity is harbored in strain individuality, so identification of conspecific bacterial strains is imperative for improved understanding of microbial community functions. Limitations in bioinformatics and sequencing technologies have to date precluded strain identification owing to difficulties in phasing short reads to faithfully recover the original strain-level genotypes, which have highly similar sequences. We present ConStrains, an open-source algorithm that identifies conspecific strains from metagenomic sequence data and reconstructs the phylogeny of these strains in microbial communities. The algorithm uses single-nucleotide polymorphism (SNP) patterns in a set of universal genes to infer within-species structures that represent strains. Applying ConStrains to simulated and host-derived data sets provides insights into microbial community dynamics.

Understanding how individual organisms co-exist within a microbial community is crucial to understanding community functions. For example, the study of microbial community

Correspondence should be addressed to R.J.X. (xavier@molbio.mgh.harvard.edu) or D.G. (dirk.gevers@gmail.com).

[11]Current address: Departments of Pediatrics and Computer Science & Engineering, University of California San Diego, La Jolla, California, USA.

[12]Current address: Janssen Human Microbiome Institute, Janssen Research and Development, Cambridge, Massachusetts, USA.

Author Contributions

C.L. and D.G. conceived the project, C.L. designed and implemented the algorithm, C.L., D.G., and R.J.X. designed the experiments and C.L. performed the analysis. M.K., H.S., D.G., and R.J.X. collected and sequenced the samples. C.L., R.K., R.J.X., and D.G. wrote the paper.

Competing Financial Interests

The authors declare no competing financial interests.

dynamics is important in human health, including how to maintain or restore a healthy human microbiome. Metagenomics has revolutionized microbiology by addressing some of these issues in a culture-independent manner. However, state-of-the-art metagenomics approaches are often limited to the species level[1–3] or to partially assembled population consensus genomes[4–6]. Evidence that the unit of microbial action can fall below the species level comes from multiple sources, including culturing[7], single-cell genomics[8], redundant bacterial 16S rRNA gene sequencing[9], internal transcribed spacer sequencing[10], multilocus sequence typing[11], and high-resolution genomic variation[12]. Therefore methods that enable strain resolution from metagenomics datasets are desirable.

Most existing culture-free approaches to identify bacterial strains in communities-have drawbacks that have limited wide adoption. For example, single-cell sequencing requires expensive and laborious efforts in cell sorting and suspension so that analyzing a large community using this approach is not done. Similarly, Hi-C, a sequencing-based approach[13], requires extra steps and budget for cross-linking, library construction, and sequencing. Strain typing methods leveraging strain-level gene copy number variations[14] or strain-level phylogenetic marker SNPs such as canSNPs[15], PathoScope[16], and Sigma[17] rely on the availability of complete reference strain genomes and, with current limitations on these resources, run into challenges when studying the broader diversity found using metagenomic sequencing approaches. An assembly-based approach is dependent on several factors, including genome structure and intra-species divergence. With rare exceptions, assemblers usually fail to produce individual strain assemblies, instead creating either highly fragmented contigs or contigs that only represent population consensus sequences[18,19]; a recent effort in using variation-aware contig graphs for strain identification[20] relies on manual inspection and hence its accuracy is subject to users' experience. In all of these approaches, only a relatively small fraction of strain genomes have been successfully analyzed, and their distribution is usually biased[21]. On the other hand, methods based on single marker genes such as the 16S rRNA gene often lack the resolution to reliably capture intra-specific genomic differences[22].

To overcome this difficulty and increase the utility of metagenome dataset, we developed ConStrains (Conspecific Strains), an algorithm that exploits the polymorphism patterns in a set of universal bacterial and archaeal genes to infer strain-level structures in species populations. Using both *in silico* and previously published host-derived datasets we show that ConStrains recovers intra-specific strain profiles and phylogeny with high accuracy, and captures important features of community dynamics including dominant strain switches and rare strains. The simulated data sets address performance in the context of different within-population diversities, different numbers of strains, the interference from other species within the same community, as well as the scalability of the method using a large *in silico* cohort with 322 samples. Predicted within-species structures as well as the strain genotypes were highly accurate across these simulated datasets. Applying this method to an infant gut development metagenomic data set reveals new insights of strain dynamics with functional relevance. ConStrains is implemented in Python, and the source code is available with this paper (Supplementary Code) and freely available together with full documentation at https://bitbucket.org/luo-chengwei/constrains.

RESULTS

The ConStrains algorithm

Guided by reference species, the ConStrains algorithm compares raw metagenomic reads to reference genomes and identifies patterns in SNPs as the basis for differentiation and quantification of conspecific strains. This approach is fundamentally different from other reference-dependent methods such as Sigma and PathoScope [16,17], because, unlike these methods, using ConStrains can provide reliable predictions for those species with only one genome (complete or draft), as opposed to approaches that rely on availability of a comprehensive reference strain collection. For confident SNP calling, a species requires a minimum of tenfold coverage (Supplementary Fig. 1) within or across all samples considered, which is obtained for all species with a relative abundance of >1% at typical sequencing depths of 5 Gbp. When applied to multiple samples, for example a longitudinal time series or otherwise related samples, strain identities can be traced across the different samples. The algorithm achieves this in two operations: (1) identifying species for which SNPs are detected and quantified, and (2) transforming individual SNPs into SNP profiles representing individual strains.

The first operation is a two-step process. Because the algorithm identifies strains only for those species with sufficient sequencing depth (≥10-fold coverage in at least one sample; Supplementary Fig. 1), the first step uses MetaPhlAn[1] for rapid species composition profiling. For those species with sufficient sequencing depth, a custom database of marker genes is created from the comprehensive PhyloPhlAn marker set[23], against which the raw reads are mapped using Bowtie2[24]. This targeted approach allows for optimized time and computational efficiency. Resulting marker gene alignments are processed with SAMtools[25] to generate a table of coverage by base position from which SNPs are identified. It is important to note that in this process the reference sequences are removed and SNPs are identified *de novo* to minimize reference dependency (Fig. 1a–d and **Online Methods**). We verified that such a SNP selection procedure is sufficiently accurate and uniquely sensitive to disentangle intra-specific diversity (Supplementary Note 1 and Supplementary Fig. 2).

In the second operation, individual SNPs are combined into unique SNP profiles from which strains are identified. Previous methods have approached the challenge of identifying individual organisms from microbial communities using SNPs (for example, oligotyping[26] and minimum entropy decomposition[27]), but were limited to SNPs within the span of a sequence read length. ConStrains overcomes this read length limitation and represents each strain by a barcode-like string of concatenated SNPs spanning hundreds of genes, referred to as the "uniGcode." To derive the strain's uniGcodes within a data set, ConStrains constructs candidate models of strain combinations using a combination of SNP-flow and SNP-type clustering algorithms. Sequentially, the relative abundance of strains in each candidate model across the cohort is estimated using a Metropolis-Hastings Markov Chain Monte-Carlo approach (Fig. 1e–g and **Online Methods**). Finally, to choose the optimal model with the principle of balancing model fitness and complexity, corrected Akaike information criterion (AICc) is employed (Fig. 1h and **Online Methods**). ConStrains repeats these steps for each species with sufficient coverage, then outputs the number of strains and their

respective uniGcodes and relative abundances (Fig. 1i). The uniGcode allows downstream analysis such as cross-sample comparisons and evolutionary studies.

ConStrains identifies strains in large data sets

To validate the performance of ConStrains for strain profiling, we used *in silico* and host-derived data sets. A total of 36 different sets of *k*-strain mixtures were generated using *in silico* genome-based Illumina paired-end read simulation based on ten different *Escherichia coli* strains whose complete genomes are publicly available, representing real-life scenarios of strain admixtures (k = 2–7; Fig. 2a–b and Supplementary Fig. 3a, Supplementary Table 1). These 36 sets of reads were profiled by ConStrains using default settings. Predicted results were compared with the 'true' strain compositions using Jensen-Shannon divergence (JSD; Fig. 2b and Supplementary Fig. 3b). ConStrains successfully predicted the underlying intra-species compositions in all 36 data sets ($P < 1 \times 10^{-5}$; two group *t*-test against random guesses; Fig. 2b), demonstrating a substantial advantage (Supplementary Fig. 4) over reference-base approaches (see Supplementary Note 1 and Supplementary Fig. -5 for details and comparisons). Furthermore, in 34 of the 36 sets of reads (94.44%), the numbers of strains inferred exactly matched the ground truth (Fig. 2a), with the remaining two sets of reads having an additional chimeric strain predicted at an extremely low level (<0.1%). We therefore set the recommended detection limit at 0.1% to reduce such errors computationally. Since this is a relative abundance threshold, one can still target low abundance organisms by increasing sequence depth. In similar simulations with up to 30 *E. coli* strains, ConStrains predicted the strain composition with high confidence when the strain number was less than ten (Fig. 2c), which represents the intra-specific upper bound for more than 95% of metagenomic species (Fig. 2d and Supplementary Note 1). To assess the impact of intra-species recombination on performance, both real sequencing reads from highly recombined *Burkholderia pseudomallei* strains[28] and *in silico*-simulated recombinant strain-derived reads were generated, and no significant adverse impact was identified (Supplementary Note 1). We also further tested the performance in a more realistic metagenomic scenario by embedding *E. coli* strains within communities with various levels of complexity and found our approach remained robust (**Online Methods**, Supplementary Note 2, and Supplementary Table 2). We also found no significant correlation between admixture compositions' alpha diversity and prediction accuracy. These results collectively suggested good algorithm performance (Supplementary Note 1).

We then tested ConStrains using a host-derived metagenomic data set that had previously been analyzed using a manually curated strain identification approach. Using manual strain curation the authors had for the first time described the changes in an infant gut microbiome during the first 24 days of life[4]. All three manually curated *Staphylococcus epidermidis* strains reported in this study were successfully predicted by ConStrains in a fully automated manner, with the predicted relative abundances of each strain over time having highly similar values to the original compositions quantified from the scaffold coverage (JSD avg. = 0.024, s.d. = 0.021; Supplementary Fig. 6). Because the performance of ConStrains' fully automated approach matched well with the manually curated, semi-automated approach described previously[4], but required far less machine and manual resources (ConStrains completed the infant gut data set in 8.5 CPU hours with RAM peak footprint of < 2GB on a

Linux server with Xeon 2.6GHz processors, in contrast to days to weeks of manual curation following assembly), we next applied ConStrains to a very large data set for which a manual effort would not be feasible (for detailed resource usage, see Supplementary Note 5 and Supplementary Table 3).

In the absence of the existence of such a large data set (especially one where true results were known), we used a simulated shotgun data set with intra-specific structure mimicking the natural relative abundance of taxa informed by a recent gut microbiome collection effort for which samples were collected daily over the course of one year[29] (Online Methods and Supplementary Note 3) (Fig. 3a). ConStrains analyzed 91 species with sufficient depth in the 322 *in silico* samples. In total, ConStrains analysed 3.2 terabases of paired-end reads contained 1,361 strains from 320 species, with minimal runtime and infrastructure requirements (Supplementary Note 3). ConStrains achieved high accuracy in individual samples, and also captured crucial information such as dominant strain type changes, for example in *Bacteroides fragilis* (Fig. 3a–c and inset windows 1–3; see Supplementary Table 4 and Supplementary Note 3 for details). This large cohort also enabled us to test factors that might affect the performance of ConStrains, including population complexity, coverage, and relatedness. We found that 10× coverage was necessary for accurate profiling, and that strain relatedness could also affect performance (Supplementary Fig. 7 and Supplementary Note 3). With this thorough benchmarking, we next applied ConStrains to two previously published clinical data sets to illustrate the biological insights strain level analyses can provide.

ConStrains reconstructs strain phylogeny

Lieberman and co-workers previously reported on the genetic variation of *Burkholderia dolosa* in cystic fibrosis patients by combining a selective culturing step with a deep population sequencing approach[30]. We re-analyzed their data set using our ConStrains algorithm and predicted a total of six *B. dolosa* strains in the population with an abundance well above 0.1% (pop-I to pop-VI; Fig. 4a). We compared the uniGcodes from the six strains inferred by ConStrains with the isolate genome sequence by building a phylogenetic tree, and found that all of the colony strains and two population strains (pop-I and pop-II) were closely related (Fig. 4a). Moreover, the combined relative abundance of pop-I and pop-II represented the majority of the population (51.3% and 27.9% for pop-I and pop-II, respectively). This finding corroborated observations based on the colony sequencing approach. In addition, the ConStrains algorithm identified four additional, less abundant strains (pop-III to pop-VI). None of these strains could be inferred by the colony sequencing approach alone, likely because of their low abundance (11.2%, 8.1%, 1.0%, and 0.5%, respectively). To validate these additional predictions, we further examined the polymorphism patterns in these four strains, and compared them against pop-I and pop-II. As shown in Fig. 4b, we found patterns that are unlikely to have resulted from chimeric mixtures of SNPs from pop-I and pop-II ($P < 0.01$, permutation test). This analysis demonstrated that application of ConStrains to cross-sectional datasets, used in addition to a culture-based approach, allows for a comprehensive and efficient discovery of rare strains.

Uncovering strain dynamics in infant gut development

We next analyzed an infant gut development dataset containing 54 samples from 9 subjects collected over the first three years of life (Online methods and Supplementary Fig. 8) to further explore the ability of ConStrains to reveal strain dynamics. ConStrains analysis was run on a total of 75 species that had sufficient sequencing depth for analysis (10×; Fig. 5). Because previously reported strain detection algorithms were limited to studying the population consensus sequences[12], and ConStrains has the capability to untangle intra-species diversity, we first examined the number of strains observed within each species. Nearly all species (94.66%) had more than two strains, with an average of 4.88 strains per subject (±1.54 s.d.; Supplementary Fig. 9). By tracking the uniGcode of each strain in separate individuals, we identified several unique strain-level longitudinal patterns. For instance, the population of *Fecalibacterium prausnitzii* was usually comprised of several strains that maintained a co-dominant profile, in which the strains maintained the same order of abundance; *Ruminococcus gnavus* had highly variable behaviors over time, with different strains dominating the intra-species composition at different time points; in contrast, *Bacteroides ovatus* contained one dominant strain over time keeping other strains relatively rare. *Bifidobacterium bifidum* strains demonstrated comparable dynamic patterns similar to *F. prausnitzii*; moreover, the strains reestablished the same intra-specific diversity even after the abundance of the species dropped below the detection limit (Fig. 5, open boxes). We anticipate that the capability of generating better insights in intra-species dynamics of such health-related species[31–33] will shed light on the role of these organisms in human physiology.

With this goal in mind, we pursued our findings in *Bifidobacterium longum*, an organism linked to human gut health and applied to prevention and treatment of several diseases[33]. We first observed that the phylogeny of *B. longum* strains strongly correlated with their host origins (Fig. 5, phylogenetic treem insert box), which indicated strong individuality of *B. longum* strains. Moreover, in two subjects (4 and 6, Fig. 6a), we observed switches in dominant strain types that were highly correlated with the overall relative abundance of the *B. longum* species. As previous work has shown that a single operon can affect the competitiveness of different *Bacteroides fragilis* strains[34], we evaluated functional differences between different dominant strains. In both subjects, the different strains dominating during consecutive phases (period 2 in subject 4 and period 1 for subject 6; Fig. 6a) carried additional functions that might be crucial to *B. longum*'s successful colonization of the host gut. In particular, presence of the human milk oligosaccharide (HMO) utilization cluster has been shown to result from an adaptation to the human infant gut[35] (Fig. 6b; highlight IV). Some additional functions might underlie formation of a *B. longum* bloom including the presence of the fructose and *L*-fucose utilization gene clusters (Fig. 6b; highlights I and III). Together, these findings might explain why strains with these functions were associated with higher relative abundance of *B. longum* in the infant gut microbiome. We also observed functions specific to strains that were dominant in periods when *B. longum* was less abundant (periods 1 and 3 in subject 4 and period 2 in subject 6; Fig. 6a), most notably that the capsular polysaccharide biosynthesis genes were absent from dominant strains when *B. longum* was more abundant (Fig. 6b; highlight II). Taken together, strain-

level insights provided by ConStrains, combined with functional analyses, could offer candidate targets and hypotheses for future studies.

DISCUSSION

We have shown that the ConStrains algorithm accurately predicts strain-level profiles in large cohorts of metagenomic samples, and that the inferred uniGcodes reconstruct strain phylogeny, within or across cohorts, allowing combined cohort studies. ConStrains is scalable and has minimal resource requirements. In contrast, other approaches[14,16,17] are largely dependent on prior knowledge of reference strain genomes, with sub-species resolution being directly dependent on the number of available reference strains per species. This greatly limits the application of such methods on real metagenomic data, as for most of the human microbiome species only one reference genome is available[14]. Current databases are quickly gaining in intra-species genome representation, but are still far from saturating natural diversity. With just one genome per species, ConStrains can resolve natural diversity occurring within that species, and is therefore agnostic to unknown strains. Future improvements for strain-level analysis include identification of strains in the absence of any reference genomes. It is conceivable that combinaing ConStrains with *de novo* genome assembly from metagenomic data could be an appropriate candidate to overcome this hurdle.

ConStrains is particularly effective for obtaining insights that were previously overlooked using species level findings (Supplementary Note 4 and Supplementary Figs. 10–12), and will thus enable new types of studies. As shown above with the *B. longum* example, combining strain-level profiles with reference genome-based gene coverage analysis can reveal candidate genes for understanding strain-specific beneficial effects and the functions that might contribute to successful colonization in the human gut. ConStrains could also identify strains or genes associated with disease and link variable genomic regions to individual strains, a major challenge in shotgun metagenomics. Strain-level profiles, together with appropriate metadata, could link reference-based or *de novo* assembled genes with individual strains and further interpret unknown strain-specific functions. Our study of the infant gut development cohort captured HMO utilization cluster enrichment shifts in different development periods, which is particularly important because products of the HMO utilization cluster are essential for *B. longum* to exert its probiotic effects[36]. Finally, strain phylogeny could be used across cohorts and add metagenomic means to test fundamental ecological hypotheses, including neutral theory and other adaptive and nonadaptive mechanisms for maintaining sympatric diversity among strains. Although we have applied ConStrains to human microbiome datasets, it can also be applied to environmental samples to test fundamental hypotheses about the role of microbes in the environment that can only be addressed at the strain level.

Online methods

ConStrains algorithm

Extracting target species and informative SNPs—With raw reads from samples S_1, S_2, …, S_n, ConStrains starts with profiling input metagenomes using MetaPhlAn[1] (v1.7) with default settings, with the exception that alignment options are set to “very-sensitive”;

species with average coverage higher than a coverage cutoff (default: 10×) in at least one sample are selected for further strain analysis. For each of the selected species, the corresponding set of the universally conserved genes reported by Segata *et al.*[1] are used as a database, and Bowtie2[24] mapping with default setting is carried out to map reads against those reference genes. Only reads with proper pairing and orientation, no indels, >30 mapping quality, >90 length mapped (overhanging part at gene 5′ and 3′ ends is clipped off before calculation), and at least 95% nucleotide identity with the reference gene are further used. These reads are then piled up onto their respective reference sequences using SAMtools[25], and the reference gene coverage is subsequently calculated on a per-base basis. To filter out genes with spurious mappings due to hypervariable regions or conserved universal motifs, sites with less than default minimum coverage, as well as those that fall outside of the 1.5 interquartile coverage range across the gene length, are masked. Any gene with more than 30% of its length masked is discarded from further analysis. Single nucleotide polymorphism sites (SNPs) are then counted across samples as those unmasked positions where the minor allele has at least two counts or more than 3% in relative abundance.

Strain typing by SNP-flow algorithm—With SNPs extracted, ConStrains first infers the strain composition and their SNP-types using the "SNP-flow" algorithm in per-species per-sample fashion. In this algorithm, all SNP sites are first hierarchically clustered by the Euclidean distance between the frequencies of different alleles defined as

$$d_f(a,b)^2=\sum_{i=1}^{4}(a_i-b_i)^2$$

where *a* and *b* are the frequency vector of the four bases sorted in descending order of the respective SNPs. Clusters that contain less than 5% of the overall SNPs or fewer than ten SNPs are discarded. The centroid of each cluster is selected as representative. These SNP cluster centroids (SCCs) are then ranked in descending order based on their weight quantified as the number of SNPs they represent. Finally, a directed graph is constructed using these SCCs, in which nodes are alleles in these SCCs and each is assigned a "capacity" defined by the allele frequency, and these alleles from neighboring SCCs are connected by edges (Fig. 1e).

In the directed graph constructed in the previous step, nodes are denoted from the same SCC as a layer. With *m* layers in the graph, SNP-flow will explore all possible combinations of paths from the first layer to the last. This way, every such path represents a strain genotype, and its relative abundance, *c*, is defined as the lowest node capacity among all nodes on the path. Once a path is visited, all nodes on this path would subtract their capacity by the path's relative abundance *c* (Fig. 1e). Such a pathfinding and visiting step is repeated until all nodes" capacities are zero, and the visited paths constitute one combination. ConStrains exhausts all possible SNP-type (strains) combinations $\beta = \{\beta_1, \beta_2, \ldots, \beta_k\}$ in each sample with the *i*-th sample's SNP-type $\beta_i = b^i_1 b^i_2 \ldots b^i_h$ where b^i_j is one of the four bases, A, C, G, and T, and the associated strain profile $\alpha_i = (\alpha^i_1, \alpha^i_2, \ldots \alpha^i_h)$ with

$$\sum_{j=1}^{k} \alpha_j^i = 1.$$

For each sample, ConStrains picks the optimal combination that minimizes the fitting error, defined as the discrepancy between expected SNP frequencies and observed frequencies, ε, defined as:

$$\varepsilon = \frac{1}{h} \sum_{i,j} (E_{ij} - o_{ij})^2$$

where E_{ij} is expected frequency of the i-th base at the j-th SNP locale; and similarly, O_{ij} is the observed frequency of the i-th base at the j-th SNP locale in the pileup of aligned reads from the corresponding sample. For instance, C is the second base ($i = 2$), and if we observed two C's and eight A's at the fifth SNP locale ($j = 5$) in the pileup of aligned reads against reference, the frequency of C is 0.2 at that position and thus is referred to as $O_{25} = 0.2$. E_{ij} is inferred using α_i and β_i such that

$$E_{ij} = \sum_{k}^{t} \alpha_k^* b_j^i$$

Inferring strain compositions—To unify these optimal SNP-types into cohort-wide strains, ConStrains next constructs a neighbor-joining tree of the SNP-types from different samples based on sequence percentage identity, and utilizes an internal parameter, Δ_d, defined as the distance between the tree-cutting point and the leaves, to cut the tree. Rather than using a preset value, the algorithm cuts this tree using all possible Δ_d. Each internal node created by such a cut could be viewed as the representative of all the children nodes (SNP-types) on the tree. In doing so, it identifies all possible k clusters defined by the structure of the tree of SNP-types (Fig. 1f), which we refer to as candidate strains.

With the proposed k strains from the previous step, in each sample, we need to find a composition, $\alpha^* = (\alpha^*_1, \alpha^*_2, \ldots, \alpha^*_k)$ with

$$\sum_{i=1}^{k} \alpha_i^* = 1,$$

to minimize the discrepancy between expected SNP frequencies and observed frequencies, ε, as defined previously. This is carried out by a Metropolis-Hasting Monte-Carlo method. ConStrains first generates a number of seeds (default: 1,000) at uniform random on $k-1$ simplex. The top 50 seeds are then selected and each such seed's vicinity on the $k-1$ simplex is iteratively searched. In iteration t, a new point, α^t_{ik}, is picked within the 0.01 radius of the previous point, α^{t-1}_{ik}; and it is accepted as the new point with probability min(1, $q(\alpha^t_{ik}, \alpha^{t-1}_{ik})$), where $q(\alpha^t_{ik}, \alpha^{t-1}_{ik}) = \varepsilon(\alpha^t_{ik})/\varepsilon(\alpha^{t-1}_{ik})$. It repeats the iteration until $|1-q(\alpha^t_{ik},$

$\alpha^{t-1}_{ik})|$is smaller than 0.001 or the maximum number of iterations (10,000) is reached. The composition yielding the lowest ε is selected as optimal α^*_{ik}. ConStrains repeats this step for all samples and all *k,* yielding solutions for each *k*, $\alpha^*_k = (\alpha^*_1, \alpha^*_2, \ldots, \alpha^*_n)$, with corresponding error (Fig. 1g):

$$\varepsilon_k = \sum_{i=1}^{n} \varepsilon_{ik}$$

Selecting the optimal strain model—Corrected Akaike information criterion (AICc) is employed to select optimal *k*. The AICc of each *k* is calculated as:

$$AICc = 2k - 2ln(L) + \frac{2k(k+1)}{n-k-1}.$$

where $L = 1-\varepsilon_k$ denotes the model likelihood. ConStrains selects the *k* with the lowest *AICc* and outputs the associated SNP-types and compositions as final results (Fig. 1h). As noted previously, we suggest filtering strains with less than 0.1% in relative abundance as they present a high probability of being chimeric.

In silico data sets

To simulate *in silico* single species data sets, 62 complete *E. coli* genome sequences were downloaded from NCBI. Ten genomes were selected and their relatedness was shown by a maximum likelihood tree (Supplementary Fig. 3a) constructed from concatenated nucleotide sequences of core genes among the 10 strains using a method similar to Luo *et al.*[19]. 1,000 random compositions were sampled on a Gamma distribution with $k = 1$ and $\theta = 0.5$ for each number of strains ($N = 2$–7). In each set of these 1,000 compositions, Shannon entropy was calculated and based on which these compositions were ranked. The compositions on the 15th, 30th,…, 90th percentiles were picked to form a gradient of intra-specific diversity for each *N*. ART simulator[37] was employed to simulate 100× coverage of 100 bp paired-end Illumina reads using these compositions with default settings for Illumina and library settings as "-m 350 -s 50" (Supplementary Fig. 3a). These samples were further grouped together to simulate single strain series samples (Supplementary Table 1).

These simulated *E. coli* reads were then spiked into *in silico*-constructed metagenomes to measure the impact from other species. Three human microbiome-like metagenomes with low, medium, and high complexity level (referred as LC, MC, and HC, respectively) were simulated based on an aggregated MetaPhlAn[1] profile over all 690 Human Microbiome Project (HMP) samples[38]. *E. coli* and *Shigella* were excluded from the profile, and the rest of the species were ranked based on their average abundance in the HMP cohort. The top 20, 50, and 100 most abundant species were selected for LC, MC, and HC, respectively. The species composition in each *in silico* metagenome was calculated as their relative abundance in the HMP cohort, normalized by their total sum. Genomes of these species were downloaded from NCBI, and a representative strain was selected at random if multiple strains of the same species were present. A total of 100 million 100 bp paired-end Illumina reads were simulated for each set by ART simulator[37] with the same settings as mentioned

previously. Additional data sets for testing the sensitivity and the performance on different numbers of strains and recombined strains were generated in a similar fashion using ART (Supplementary Note 1 for details).

The year-long shotgun metagenome cohort with 322 samples was simulated based on donor A's 16S rRNA amplicon profiles reported in David *et al.*[29]. The operational taxonomy unit (OTU) table was used as a guide for community composition in human microbiomes. To allow simulation at the strain level, however, taxonomy in the OTU table was shifted down by one level. For instance, species composition in the original OTU table was shifted to be the strain composition. NCBI draft and complete genomes were used to match as closely as possible the phylogeny of the original OTUs. Reads were then simulated by ART simulator as previously described. The coverage was set to be 1× per 25 read counts in the 16S OTU table.

Biological data sets

The two infant gut development longitudinal metagenomic data sets used in this study were from a previous study[4] and from our recent effort in tracking nine subjects in a three-year period since birth. For the former set, all metagenomic samples were downloaded from NCBI SRA under accession number SRA052203, and the corresponding assembled *Staphylococcus epidermidis* strains and phage genomes were downloaded from ggKBase as described by Sharon *et al.*[4]. For the latter set, 54 stool samples were collected from nine infant subjects between September 2008 and August 2010 in Finland. Samples were first collected by the subjects' parents and stored in the household freezer before being transferred on dry ice to a laboratory −80 °C freezer. Samples were then shipped to the Broad Institute for DNA extraction, in which QIAamp DNA Stool Mini Kit (Qiagen, Inc., Velencia, CA, USA) was used as described previously[39]. Library construction was carried out following Human Microbiome Project's standard protocol (http://hmpdacc.org/tools_protocols/tools_protocols.php), and 101bp paired-end reads were produced on an Illumina HiSeq 2000 platform. The raw sequences of these samples are available at SRA under BioProject accession number PRJNA269305, and the corresponding sample information is available in Supplementary Table 5.

Prediction accuracy measurement

To measure how close the predicted composition, P, is from the true composition, Q, we applied Jenson-Shannon divergence with minor modifications. Since it is possible that P and Q are of different dimensions, we first padded the one with lower dimension with zeros to match the one with the higher dimension, and then defined a composition M based on sorted P and Q, P' and Q', as:

$$M=\tfrac{1}{2}(P'+Q')$$

Therefore the Jenson-Shannon divergence is:

$$JSD(P\|Q)=\tfrac{1}{2}D(P'\|M)+\tfrac{1}{2}D(Q'\|M).$$

where $D(X\|Y)$ is the Kullback-Leibler divergence defined as:

$$D(X\|Y)=\sum_{i}^{n} ln(\frac{x_i}{y_i})x_i$$

We calculate the SNP typing accuracy as the distance between the inferred SNP tree of strains, T_p, and the true strain tree constructed from concatenated core genes, T_q. First, a distance similar to the symmetric difference introduced by Robinson and Foulds is applied to calculate the distance, d, between these two trees. We then normalize d to the expected basal distance from a random tree with the same leaves. The expected basal distance, d, is the mean distance between T_q and 1,000 randomly generated trees with the same leaves.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Natalia Nedelsky for editorial support. This work was supported in part by the Crohn's and Colitis Foundation of America, the Leona M. and Harry B. Helmsley Charitable Trust, National Institutes of Health (NIH) grants U54 DK102557 (R.J.X.) and R01 DK092405 (R.J.X.)., and the Howard Hughes Medical Institute (R.K.).

References

1. Segata N, et al. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012; 9:811–814. [PubMed: 22688413]
2. Sunagawa S, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013; 10:1196–1199. [PubMed: 24141494]
3. Darling AE, et al. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ. 2014; 2:e243. [PubMed: 24482762]
4. Sharon I, et al. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. 2013; 23:111–120. [PubMed: 22936250]
5. Nielsen HB, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014; 32:822–828. [PubMed: 24997787]
6. Imelfort M, et al. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ. 2014; 2:e603. [PubMed: 25289188]
7. Luo C, et al. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. Proc Natl Acad Sci USA. 2011; 108:7200–7205. [PubMed: 21482770]
8. Kashtan N, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science. 2014; 344:416–420. [PubMed: 24763590]
9. Faith JJ, et al. The long-term stability of the human gut microbiota. Science. 2013; 341:1237439. [PubMed: 23828941]
10. Maslunka C, Gifford B, Tucci J, Gurtler V, Seviour RJ. Insertions or deletions (Indels) in the rrn 16S–23S rRNA gene internal transcribed spacer region (ITS) compromise the typing and identification of strains within the Acinetobacter calcoaceticus-baumannii (Acb) complex and closely related members. PLoS ONE. 2014; 9:e105390. [PubMed: 25141005]

11. Han D, et al. Population structure of clinical Vibrio parahaemolyticus from 17 coastal countries, determined through multilocus sequence analysis. PLoS ONE. 2014; 9:e107371. [PubMed: 25225911]
12. Schloissnig S, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013; 493:45–50. [PubMed: 23222524]
13. Beitel CW, et al. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. PeerJ. 2014; 2:e415. [PubMed: 24918035]
14. Greenblum S, Carr R, Borenstein E. Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species. Cell. 2015; 160:583–594. [PubMed: 25640238]
15. Karlsson E, et al. Eight new genomes and synthetic controls increase the accessibility of rapid melt-MAMA SNP typing of Coxiella burnetii. PLoS ONE. 2014; 9:e85417. [PubMed: 24465554]
16. Hong C, et al. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome. 2014; 2:33. [PubMed: 25225611]
17. Ahn TH, Chai J, Pan C. Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics. 2015; 31:170–177. [PubMed: 25266224]
18. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010; 95:315–327. [PubMed: 20211242]
19. Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT. Individual genome assembly from complex community short-read metagenomic datasets. ISME J. 2012; 6:898–901. [PubMed: 22030673]
20. Nijkamp JF, Pop M, Reinders MJ, de Ridder D. Exploring variation-aware contig graphs for (comparative) metagenomics using MaryGold. Bioinformatics. 2013; 29:2826–2834. [PubMed: 24058058]
21. Lasken RS, McLean JS. Recent advances in genomic DNA sequencing of microbial species from single cells. Nat Rev Genet. 2014; 15:577–584. [PubMed: 25091868]
22. Ivanova N, et al. Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis. Nature. 2003; 423:87–91. [PubMed: 12721630]
23. Segata N, Bornigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun. 2013; 4:2304. [PubMed: 23942190]
24. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9:357–359. [PubMed: 22388286]
25. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]
26. Eren AM, et al. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol. 2013; 4:1111–1119.
27. Eren AM, et al. Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME J. 2014; 9:968–979. [PubMed: 25325381]
28. Nandi T, et al. Burkholderia pseudomallei sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. Genome Res. 2015; 25:129–141. [PubMed: 25236617]
29. David LA, et al. Host lifestyle affects human microbiota on daily timescales. Genome Biol. 2014; 15:R89. [PubMed: 25146375]
30. Lieberman TD, et al. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. Nat Genet. 2014; 46:82–87. [PubMed: 24316980]
31. Sokol H, et al. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proc Natl Acad Sci USA. 2008; 105:16731–16736. [PubMed: 18936492]
32. Crost EH, et al. Utilisation of mucin glycans by the human gut symbiont Ruminococcus gnavus is strain-dependent. PLoS ONE. 2013; 8:e76341. [PubMed: 24204617]
33. Di Gioia D, Aloisio I, Mazzola G, Biavati B. Bifidobacteria: their impact on gut microbiota composition and their applications as probiotics in infants. Appl Microbiol Biotechnol. 2014; 98:563–577. [PubMed: 24287935]

34. Lee SM, et al. Bacterial colonization factors control specificity and stability of the gut microbiota. Nature. 2013; 501:426–429. [PubMed: 23955152]
35. Schell MA, et al. The genome sequence of Bifidobacterium longum reflects its adaptation to the human gastrointestinal tract. Proc Natl Acad Sci USA. 2002; 99:14422–14427. [PubMed: 12381787]
36. Sela DA, et al. The genome sequence of Bifidobacterium longum subsp infantis reveals adaptations for milk utilization within the infant microbiome. Proc Natl Acad Sci USA. 2008; 105:18964–18969. [PubMed: 19033196]
37. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012; 28:593–594. [PubMed: 22199392]
38. Human Microbiome Project. C. Structure, function and diversity of the healthy human microbiome. Nature. 2012; 486:207–214. [PubMed: 22699609]
39. Morgan XC, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 2012; 13:R79. [PubMed: 23013615]

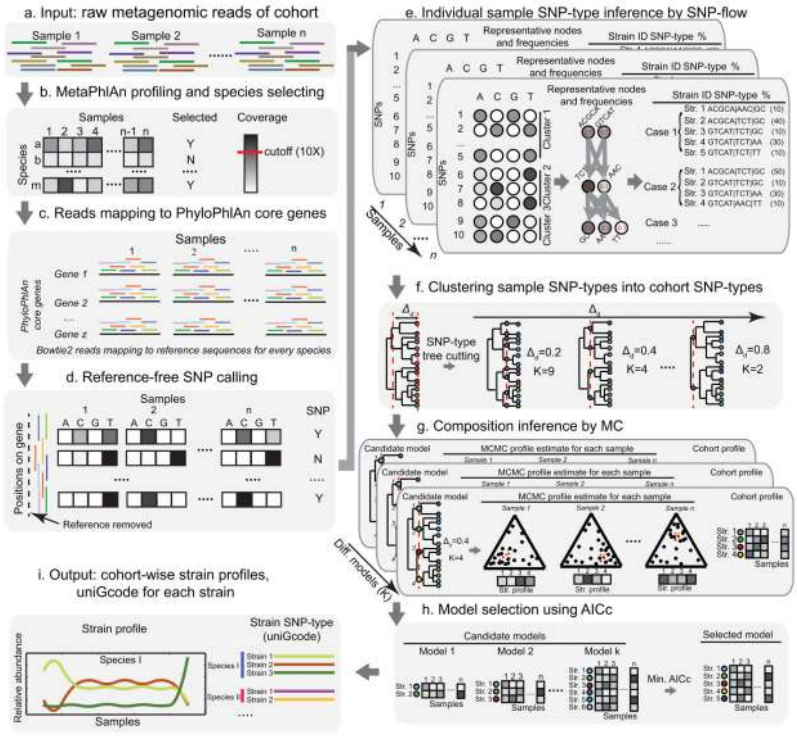

Figure 1.

Overview of the ConStrains algorithm: from raw metagenomic data to strain profiles and uniGcodes. (**a**) ConStrains requires raw metagenomic reads from a single or series of metagenomic samples as input. (**b**) To select species that satisfy a predefined sequencing depth cutoff, the algorithm starts with determining the species composition with MetaPhlAn[1]. (**c**) Next, Bowtie2[24] is used to recruit all reads to a reference database of species-specific marker genes[23]. (**d**) SNPs are called on these recruited reads after quality filtering, removal of reference gene sequence, and reference-free read realignment. (**e**) Resulting SNPs are used by a SNP-flow algorithm to infer all possible SNP-types for each of the samples. (**f**) Such SNP-types across samples are clustered using a tree structure based on their distances to represent candidate strain models; the internal distance cutoff, Δ_d, is varied to exhaust all possible SNP-type clusterings. (**g**) The Metropolis-Hastings Monte-Carlo method is then carried out to infer relative abundances per sample and per species for every candidate strain model. (**h**) These models are then evaluated by corrected Akaike information criterion (AICc) and the model with minimum AICc is selected as the optimal model. (**i**) Finally, the associated strains' relative abundances across samples and their uniGcodes are generated for every species.

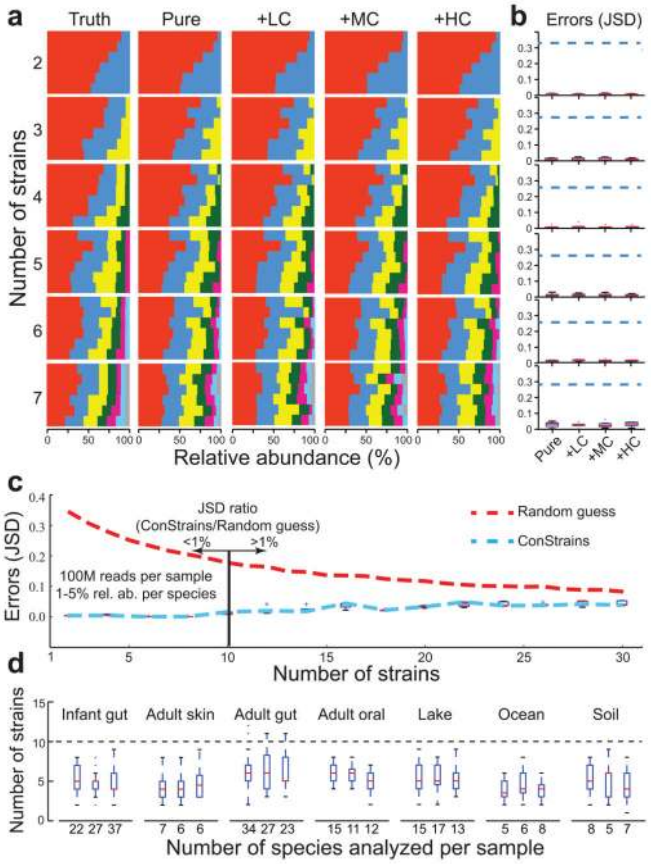

Figure 2. ConStrains correctly predicts the strain composition of *in silico*-simulated data sets

A comparison of true and predicted strain composition profiles of *in silico*-simulated multi-strain mixtures is shown. (**a**) An increasing number of multi-strain mixtures (n = 2–7; rows) was analyzed with ConStrains either containing only the target strains (pure) or in the context of a metagenome of low, medium, and high complexity (+LC, +MC, and +HC, respectively). In each box of barcharts, the colors represent different strains that were mixed in six different ratios (x axis, relative abundance) with a Shannon index (y axis) increasing from top to bottom. In the resulting 144 admixtures, all strains were correctly identified. (**b**) To compare the predictions in abundance for each strain, the Jensen-Shannon Divergence (JSD) between predicted composition and the true composition was determined. Blue dashed lines mark the expected errors from random guesses. The box marks the interquartile range, the red bar marks the interquartile median, whiskers represent the top and the bottom 25% data range, and outliers are marked by crosses. Good performance was obtained for all compositions, with minimal difference in the accuracy of results between pure mixtures and metagenomic mixtures; see also Supplementary Fig. 3b for more detailed graphs. (**c**) Graph showing ConStrains' ability to correctly infer intra-specific structure as a function of the number of strains contained in a sample. Shown is a typical case with the species' relative abundance ranging from 1% to 5% and a sequencing depth of 100 million paired-end reads, though higher abundance or sequencing depth would improve its accuracy. The ConStrains' prediction JSD errors (blue dashed line and boxes) were below 1% of null informative prediction errors (random guess; red dashed line) when the number of strains within a

species was less than ten. (**d**) For comparison, three metagenomic samples were randomly chosen from seven different niches, ranging from adult gut microbiome to a marine planktonic community. More than 95% of the species from these metagenomic samples possessed fewer than ten strains (dashed horizontal line). Dashed lines and whiskers mark the interquartile range; plusses mark the outliers.

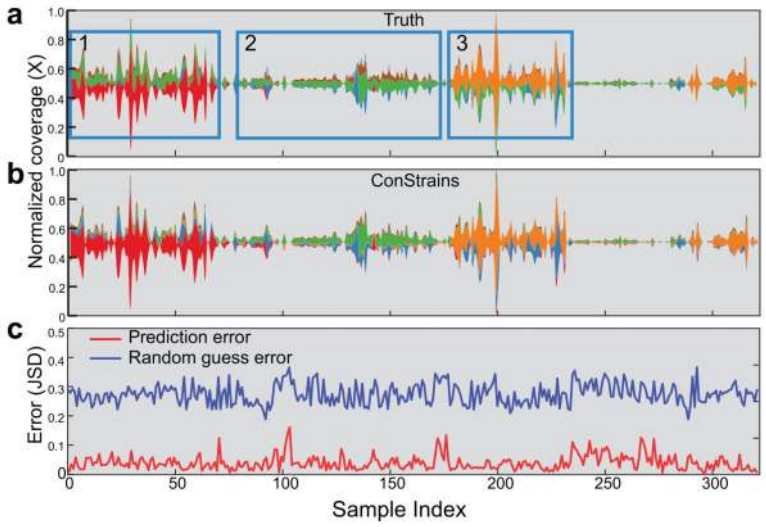

Figure 3.

ConStrains scales to large time series and accurately predicts strain dynamics. In the absence of existing large time series metagenomic data sets, a simulated set with 322 samples was created. Shown are the strain predictions within the *Bacteroides fragilis* species. The (**a**) true and (**b**) ConStrains-predicted relative abundance (*y* axis) of *B. fragilis* strains (stream ribbon width, with different colors representing different strains) in different samples sorted in longitudinal order (*x* axis, sample index) are illustrated. Inset windows 1–3 in **a** indicate periods with different dominant strains. (**c**) Prediction errors (red line) in each sample were measured between the true and predicted profiles using Jenson-Shannon Divergence (*y* axis, JSD). For comparison, random guess error (blue line) is shown to indicate a lower performance boundary. Spikes in error rates above 0.1 JSD are mostly related to time points in which the species average coverage drops below 10×, preventing reliable SNP profiling (Supplementary Fig. 7b).

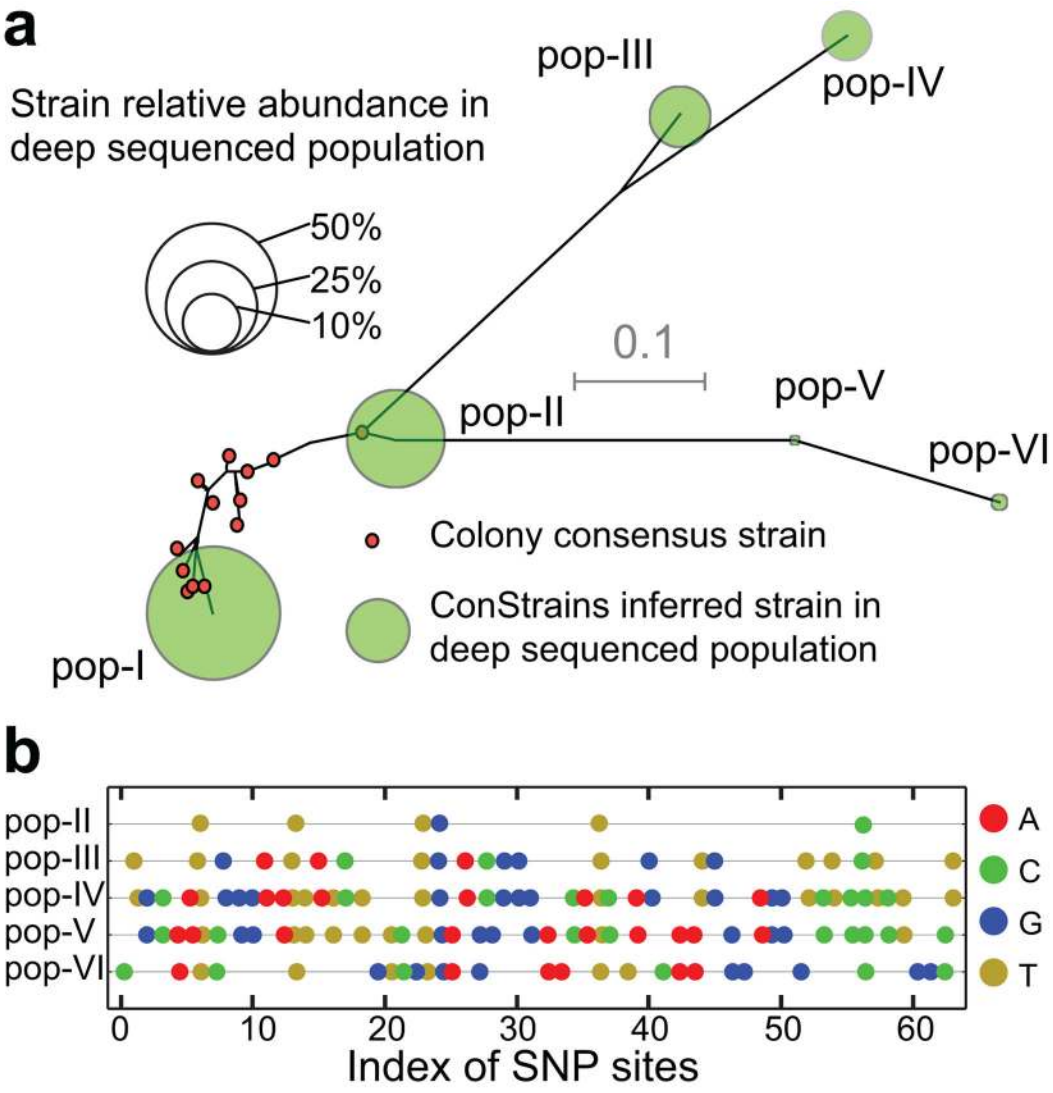

Figure 4. High sensitivity identification of strain phylogeny within a cystic fibrosis *Burkholderia dolosa* population data set

ConStrains was used to re-analyze data from a published study on the genetic variation of *Burkholderia dolosa* populations within cystic fibrosis patients[30]. (**a**) A total of six *B. dolosa* strains (pop-I to pop-VI) were predicted with an abundance of > 0.1% of the species (diameter of green circles proportional to relative abundance). An unrooted neighbor-joining tree on the alignments of the unweighted concatenated SNP profiles was constructed for the predicted strains (green circles) and the corresponding genomic data for the 29 cultured isolates (red circles; gray bar indicates the tree distance scale). These results show that the original study retrieved numerous isolates for the two most dominant strains within the population, but could not isolate the lower abundance strains. Distance between predicted strains and isolates fall within the prediction sensitivity of the ConStrains algorithm (same strain individuals differ with no more than 5% of all SNPs). (**b**) To demonstrate the sensitivity of the algorithm for differentiating strains, the color-coded allelic difference for

each of the predicted strains is shown in reference to the most dominant strain, pop-I. Sites with the same allele as reference (pop-I) were not marked.

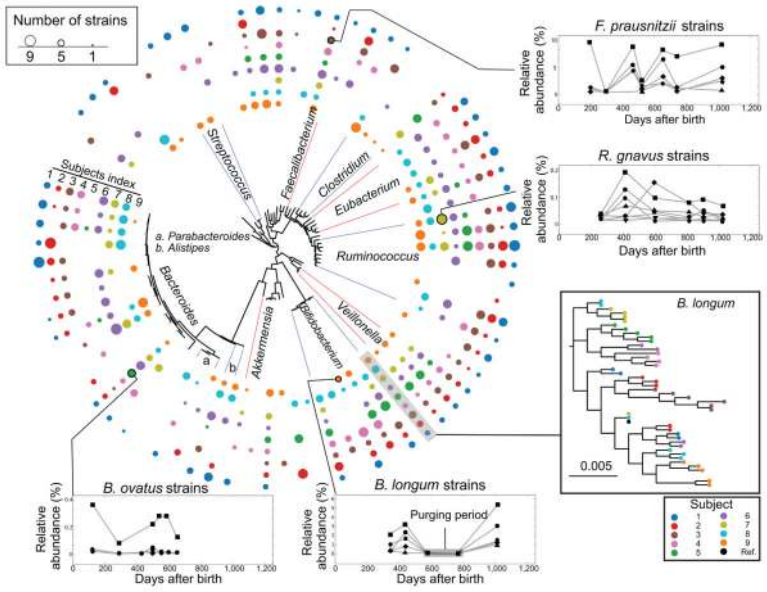

Figure 5.

ConStrains analysis reveals species longitudinal dynamics and functional shifts within an infant gut development cohort. A cohort of nine infants that were sampled throughout the first three years of life, and for which metagenomic data was available for up to nine time points, were analyzed with ConStrains. For a total of 75 species, the depth was sufficient to interpret the underlying strains. The circular tree is constructed using a representative sequence for each species, with the colored outer rings indicating the number of strains observed for each of the nine subjects. Open boxes show the longitudinal dynamics of strains in four selected species; the phylogeny tree insert box shows all strains including the available reference genome of *B. longum*.

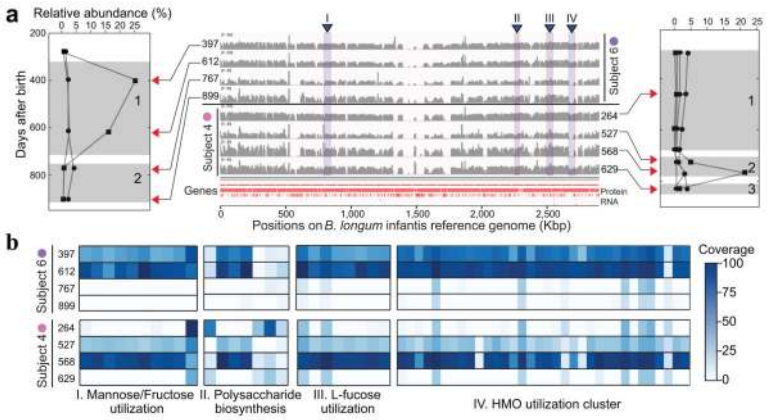

Figure 6. Functional differences in *Bifidobacterium longum* strains at different time points during infant gut microbiome development

(**a**) Two subjects experienced dominant strain switches within the species *B. longum* (flanking panels, periods marked by numbered gray shadows). Each track in the middle panel shows the corresponding sample's coverage over the *B. longum* reference genome. Time points (days after birth) are marked by red triangles. Windows I–IV capture gene content differences before and after dominant strain switches, reflected by the reference genome. (**b**) The four highlighted regions (I–IV in **a**) indicate strain-specific functional cohesion that is also strongly associated with *B. longum* relative abundance in gut microbiome development.