Supplementary Information

Nomenclature of samples. Each sputum sample received a unique identifier, i.e. the two first letters indicate whether the specimen was retrieved from an exocrine pancreatic sufficient (PS) or exocrine pancreatic insufficient (PI) subject. The third letter identifies the age group of the subject (A, child [8-13 years]; B, adolescent and young adult [18 – 23 years]; C, adult [> 28 years]. The fourth letter indicates the sex (M, male; W, female). The following numeral specifies the subject within his/her group. The number after the dot is the number of the serial specimen collected over a one-year period. Example: The specimen PSCW5.2 is the second sputum sampled from the exocrine pancreatic sufficient adult CF female 5.

Methods

Wet lab experimental procedures

Sampling and processing. Sputum was collected within one year on one to four occasions according to the Standard Operating Procedure 530.00 of the CFFT Therapeutics Development Network Coordinating Center [1]. In brief, the CF subjects performed up to 4 cycles of 3-minutes inhalation of 3% hypertonic saline with the Pari Boy S nebulizer (PARI, Starnberg, Germany). Secretions were mobilized by autogenic drainage in order to ensure representative sampling of the whole lung. Expectorated respiratory secretions were flushed with nitrogen, shock-frozen at -80°C and then stored at -80°C or immediately processed.

Samples were collected at regular visits of the CF outpatient clinic or prior to the first dose of a 14-day course of elective intravenous antipseudomonal chemotherapy. Induced sputum was successfully collected from all PI study participants. In contrast, no sputum was recovered from three PS CF children (age group A), one PS CF adolescent (age group B) and one PS CF adult (age group C). These study participants were not successful to produce induced sputum within the 1-year study period.

Control samples. Sterile swabs stabbed for a minute in the agar tube and water used for the set-up of solutions were processed by the same protocol of DNA isolation and DNA library preparation (negative controls). In parallel four respiratory specimens each were collected on the same day from four healthy subjects. Two samples each taken from each subject were split and processed in parallel (technical replicates). Corresponding pairs are identified in Table S5 by ‘name’ and ‘name_F3’.

Fresh or thawed samples were diluted 1:5 with ice-cold 97.5 % phosphate buffered saline/2.5% mercaptoethanol (v/v) and incubated on ice for 2 h under shaking. The suspension was centrifuged (15 min; 3,800 g; 10°C), the pellet was dried for 10 s, dissolved at 4°C in 10 mL bi-distilled water and then incubated on ice for 15 min on a rocker switch. This cycle of centrifugation, drying and incubation in distilled water was repeated twice. The pellet was dissolved in 1 mL 0.1 RDD-buffer (QIAGEN, Hilden) and incubated in two 0.5 mL aliquots with 60 units DNase I for 90 min at 30°C under shaking (350 rpm). The solutions were combined, diluted with 40 mL DNase buffer and centrifuged (15 min; 3,800 g; 10°C). The pellets were washed three times with 10 mL SE-buffer each by centrifugation, then dissolved in 0.5 mL SE-buffer and pelleted again (10 min, 12,000 g, 10°C). Subsequently genomic DNA was purified according to the 'Hard-to-lyse-Bacteria' protocol with the NucleoSpin Tissue kit (Machery-Nagel, Düren). DNA was stored at 4°C in Tris-EDTA buffer. Yield of double-stranded DNA was determined at the Qubit 1.0 fluorimeter with the Qubit dsDNA BR assay kit (Q32850, Agilent technologies). This protocol was found to be an acceptable compromise to obtain some non-stoichiometric amounts of hard-to-lyse mycobacteria and fungi and not to lose all easy-to-lyse mycoplasms.

DNA library preparation. 0.1 - 1 µg of DNA was sheared in a Covaris S2 system. Fragment libraries were prepared at the E120 scale according to the protocols provided by Thermo Life Technologies for SOLiD5500 instruments (generation of libraries: https://tools.lifetechnologies.com/content/sfs /manuals/4460960_5500_FragLibraryPrep_UG.pdf; emulsion PCR: Emulsifier, Amplifier http://tools.lifetechnologies.com/content/sfs/manuals /cms_102275.pdf; Enricher http://tools.lifetechnologies.com/content/sfs/manuals /cms_089261.pdf).

These standard protocols for the generation of fragment libraries for NGS applications generate a bias for GC-rich sequences, as the ligation of the adaptors becomes inefficient for DNA fragments containing more than 65% GC. To compensate for this constraint, we modified the ligation step of the standard protocol (LT/Thermo). The dA tailing reaction was performed in ¼ of the standard volume with Stratec Taq Polymerase instead of the LT- dA tailing enzyme (DNA 9µl; 5x Buffer(LT) 2.5µl, 10mM dATP 0.25µl, Stratec Taq Polymerase 1.25µl; 30 min; 68°C). The incubation conditions of the subsequent ligation were altered to increase life time and performance of the T4 ligase (dA-tailed reaction mix 13 µl; 5x Buffer(LT) 0.75µl; each adaptor (LT,1:20 diluted) 0.5 µl; 10mM dNTP 0.3µl; Quick T4 Ligase (NebNext, NEB) 0.8µl; water 0.1µl; 12 h; 12°C; followed by nick translation (20 min; 72°C)). The subsequent purification and amplification (5 cycles) of the generated fragment library was performed according to LT standard protocols.

Sequencing. The binding of the fragment library to beads was performed according to manufacturer's protocols (EZBead System(LT); E120 scale, P2 post enrichment 17%). Sequencing was performed on a SOLiD 5500XL system (LT) with 75 bp read length and implemented Exact call chemistry (LT). The accuracy of sequencing of the instrument was determined independently to be 99.943 %, i.e. a mean of 57 single nucleotide errors are estimated per 100,000 bp of raw sequence.

In silico analyses

Processing of sequences reads. In total we obtained 2.2 billion color space, quality-trimmed and filtered single-end sequences with an average length of 60 bp. More than 77 million reads (3.5% of the total amounts of reads) were non-human (average of 1.25 million reads (74.6 Mbp) per sample). Raw sequencing data were first processed to remove SOLiD barcode sequences and thereafter trimmed to filter out low quality reads. Sequences with at least 40 bases with a quality score above 20 and a minimum length of 45 bp were selected for the analysis. The trimmed reads were aligned against the human reference genome (NCBI build 37/hg19), first using the ultrafast Bowtie2 [2] and thereafter the unaligned reads were processed using the NovoalignCS (http://www.novocraft.com/) short read aligner.

Non-human reads were then checked for low complexity reads. Low-complexity sequences contain repetitions of nucleotides with low or limited information content, e.g. two- or three-letter repeats. These sequences are prone to cause false positive cross-alignments to human and microbes, so they need to be removed. The grade of complexity was estimated by PRINSEQ (http://prinseq.sourceforge.net/) with the DUST [3] method which calculates the frequency distribution of trinucleotides whereby high scores are attributed to mono-, di- or trinucleotide repeats. A stringent threshold of 5 was necessary to eliminate the low complexity reads.

Non-human sequences were corrected with the software SOLiD Accuracy Enhancer Tool (SAET), which increased the number of mapped reads by 40 - 50% in genomes 1 Kbp-200 Mbp in size (http://solidsoftwaretools.com/gf/project/saet/ and http://bcc.bx.psu.edu/download/saet.2.2/) and reduced the error rate by 3 to 5-fold. Reads were grouped by similarity. If a mismatch was found and it was not supported by high quality reads, the software corrected the low quality read having a 'consensus' sequence.

Reference-based taxonomic classification. A local database of complete microbial reference genomes was created (1,680 bacteria, 610 fungi, 5,804 viruses and 120 archaea) downloaded from the National Centre for Biological Information (NCBI, http://www.ncbi.nlm.nih.gov). Draft or incomplete genomes were not considered because they frequently contain contaminations and implausible sequences.

Non-human corrected reads were aligned to the database using NovoalignCS. Next, the option -r 'None' was used for the identification of matches of viruses, fungi and bacteria at the strain level. The option -r 'All' was used for the remaining reads that aligned to multiple bacterial genomes. These reads were then interrogated with an in-house Perl script whether they could be reassigned to the species level.

Removal of mobile genetic elements. First, a single-sample t-test method was applied to calculate the mean distances among the reads aligned to a specific reference genome. A cutoff p-value of $p < 0.01$ removed most sequences clustered to a specific region. However, in cases of numerous genomic islands in a bacterial genome, the distribution of distances of genome map positions between pairs of sequence reads was interrogated whether it followed a Gaussian distribution centered around '0.25 x genome size' [4].

Normalization. The remaining microbial reads were then normalized by GC content and genome length. The SOLiD technology has a pronounced GC bias in GC-rich regions [5] which affects the quantification of microbial genome abundances. Based on sequencing of a set of bacterial strains of 30% to 71% GC-content on the SOLiD instrument, an empirical algorithm was developed that normalizes each read by its normalized coverage coefficient (dependent variable), based on the GC content of the read and the GC content of the genome to which it has aligned (independent variables) (Chouvarine et al., unpublished). The GC-corrected reads were then normalized by genome length and reported as counts per Mb of reference. Finally, bacterial abundance was normalized to bacterial DNA per human cell present in the metagenomic sample.

Unaligned sequences (from 0.5% to 28.9 % of total amount of sequences per sample) were queried by blastn against the NCBI nt database (downloaded in January 2014) to improve the recovery rate of rare species or incomplete genomes not present in our database. Default values were selected to take the best hit for each sequence match.

Principal component analysis. We performed principal component analysis of bacterial abundances on the genus level of the samples divided into pancreatic sufficient (PS) and pancreatic insufficient (PI) groups. This analysis was performed using R (version 3.2.1). Two different methods were implemented. In the first method the relative bacterial abundances were created by applying the decostand (data,"total") method from the vegan package for R on the abundance count data followed by application of the prcomp function in R for standard PCA. In the second method we used absolute bacterial abundances per human cell to perform the same analysis. In both cases the bacterial data were normalized for GC bias and genome length of the bacterial reference genomes. The size of the concentration ellipses in probability was set to 0.95.

***P. aeruginosa* and *S. aureus* clone analysis.**

a. Assessment of clonal diversity. *P. aeruginosa* and *S. aureus* sequences were aligned with NovoalignCS against the *P. aeruginosa* PAO1 [6] and *S. aureus* Newman [7] reference genomes. Single nucleotide polymorphisms (SNPs) were extracted using Samtools [8]. Reads which differed from the reference sequence by at least one nucleotide substitution were designated as 'mismatch' (*mm*) sequences and reads with 100% sequence identity with the reference sequence were designated as 'matches' (*m*).

To determine the number of clones and clonal variants and their relative abundance, all SNPs were queried which were covered by at least 10 reads at the respective genome position. For each SNP the number of matches n_m and the total number of reads ($n_m + n_{mm}$) covering this genome position were counted.

Each clone X with n_i SNPs and a relative abundance p_X ($0 < p_X \leq 1$) in the population will show a hypergeometric distribution [9] of the ratio $n_m / (n_m + n_{mm})$ at the *i* genome positions. In other words, the relative abundance p_X of the clone is queried *i* –times from $n_{mi} / (n_{mi} + n_{mmi})$.

The ratios $n_{mi} / (n_{mi} + n_{mmi})$ were truncated to absolute percentage and all matches with the same percentage were pooled. Pooled matches were plotted against the truncated ratio $n_m / (n_m + n_{mm})$, i.e. 1% intervals. Clones and clonal variants were differentiated by the number of SNPs that contribute to the truncated ratio $n_m / (n_m + n_{mm})$. The relative abundance p_X of a clone shows up in similar $n_{mi} / (n_{mi} + n_{mmi})$ ratios for many SNPs. In contrast, variants of the same clone are detected if neighbouring $n_{mi} / (n_{mi} + n_{mmi})$ ratios fit the hypergeometric distribution, but are supported by only a few SNPs and thus low numbers of n_{mi} reads.

Considering the accuracy of SOLiD technology of 99.94%, only clones or clonal variants with a relative abundance of at least 0.1% will show reliable signals in a data set of at least 10,000 species-specific reads.

b. **Genotyping of clones.** Clone types have been commonly presented by sequence types derived from multilocus sequence typing (MLST) of housekeeping genes or by hybridization onto multimarker arrays. To extract this information from the metagenome sequences, the reads were mapped onto the MLST or array loci that have previously been used for genotyping in the respective species. In case of *P. aeruginosa* clone type identification, 12 SNPs in seven loci of the core genome were queried that had previously been selected for a multi-marker genotyping device [10]. *S. aureus* clone types were identified by sequence type. Sequence types were downloaded from http://saureus.mlst.net/ and the experimental reads were aligned with NovoalignCS against them. The analysis of *P. aeruginosa* and *S. aureus* clones was performed on 25 samples from 10 subjects and 16 samples from 13 individuals, respectively.

Antimicrobial resistance genes identification. Bacterial sequences were aligned to 'The Comprehensive Antibiotic Resistance Database' (CARD) [11] to define genetic carriage of resistance profiles in the cystic fibrosis lungs.

Uniquely aligned reads carrying a maximum number of 3 SNPs were selected for analysis. *P. aeruginosa* and *S. aureus* sequences aligned against CARD were extracted and aligned (following the same procedure described previously) against the *P.aeruginosa* PAO1 and *S. aureus* Newman reference genomes, respectively. Samtools and SnpEff [12] were used to extract and categorize the effects of the genetic variants on the coding DNA sequences. We detected 132 SNPs present in the aligned *P. aeruginosa* reads, 20 (12.5%) of which were non-synonymous SNPs. Five SNPs were qualified as rare or *de novo* mutations present in less than 20% of the aligned sequences. The *S. aureus* genes contained 221 SNPs and 2 indels of which 30 were non-synonymous SNPs (13 of them rare or *de novo* mutations).

Statistical and phylogenetic analysis. R software (version 3.2.1) was used to perform all statistical analysis. To prepare the Figures the following R packages were used: Figure 1: gridExtra, reshape2, ggplot2, Rmisc, grid; Figure2: gridExtra, reshape2, ggplot2, Rmisc, grid, gtable, scales; Figure5: gridExtra, msir, ggplot2; Figure 7: gridExtra, reshape2, ggplot2, Rmisc; FigureS1: ggplot2; FigureS3: ggplot2. The program MetaPhlAn2 [13, 14] was used for taxonomic classification of normalized sequence data and for the construction of heatmaps of the most abundant species. Phylogenies were constructed with the tool GraPhlAn [15] (Graphical Phylogenetic Analysis).

Access to the metagenome evaluation pipeline. Readers who are interested to use our pipeline for metagenome analysis should contact one of us (PML) by email (pmlmetagenomecf@gmail.com) who will provide the DOI for uploading the scripts.

References

1. CFFT Therapeutics Development Network. Sputum induction using the Nouvag nebulizer with medication cup. 2010.
2. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 2012; **9:** 357–359.
3. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low complexity DNA sequences. *J Comput Biol* 2006; **13:** 1028-1040.
4. Davenport CF, Neugebauer J, Beckmann N, Friedrich B, Kameri B, Kokott S, Paetow M, Siekmann B, Wieding-Drewes M, Wienhöfer M, Wolf S, Tümmler B, Ahlers V,Sprengel F. Genometa--a fast and accurate classifier for short metagenomics shotgun reads. *PLoS One* 2012; **7**: e41224.
5. Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, Hutter B, Jäger N, Kool M, Taylor M, Lichter P, Pfister S, Wolf S, Brors B, Eils R. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PLoS One* 2013; **8**: e66621.

6. Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M, Garber RL, Goltry L, Tolentino E,Westbrock-Wadman S, Yuan Y, Brody LL, Coulter SN, Folger KR, Kas A, Larbig K, Lim R, Smith K, Spencer D, Wong GK, Wu Z, Paulsen IT, Reizer J, Saier MH, Hancock RE, Lory S, Olson MV. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* 2000; **406**: 959-964.
7. Baba T, Bae T, Schneewind O, Takeuchi F, Hiramatsu K. Genome sequence of *Staphylococcus aureus* strain Newman and comparative analysis of staphylococcal genomes: polymorphism and evolution of two major pathogenicity islands. *J Bacteriol* 2008; **190**: 300-310.
8. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078-2079.
9. http://mathworld.wolfram.com/HypergeometricDistribution.html
10. Wiehlmann L, Cramer N, Tümmler B. Habitat-associated skew of clone abundance in the *Pseudomonas aeruginosa* population. *Environ Microbiol Rep* 2015; **7:** 955-960.
11. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJ, Spanogiannopoulos P, Sutherland AD, Tang I,Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother* 2013; **57**: 3348-3357.
12. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 2012; **6**: 80-92.
13. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *Peer J* 2015; **3:** e1029.
14. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 2012; **9**: 811-814.
15. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 2015; 12: 902-903.

Figure Legends

Figure S1. Rank number differences between the frequency of detection and the relative abundance of taxa in the whole data set of metagenomes of CF sputa. All detected taxa were sorted by rank numbers for the total number of reads assigned to the respective taxon and its detection rate in the samples. The figure displays the difference of rank numbers between abundance and frequency of detection for the top 95% of species belonging to the actinobacteria, bacteroidetes, firmicutes, fusobacteria or proteobacteria, respectively. Rank number differences are shown for samples collected from PI (green triangle), PS (blue square) and all patients with CF (orange circle).

Figure S2. Heatmaps of the relatedness of sputum metagenomes of PI (**a**) and PS (**b**) patients based on the abundance of the 25 most frequent bacterial species.

Figure S3. Association of the detection rate of microbial species with the total number of assigned microbial reads in the metagenome sample. The species composition of the individual metagenomes is shown whereby the color of a dot visualizes the detection rate of the respective species in all specimens.

Table Captions

Table S1. Clinical characteristics of the patient cohort.

Table S2. Reads of DNA viruses, bacteria, molds and fungi detected at the species level in the individual sputum metagenomes collected from patients with cystic fibrosis. The nomenclature of samples is explained in the first paragraph of the Supplementary Information (see above).

Table S3. A. Number of species (DNA viruses, bacteria, fungi) detected in sputa collected from pancreatic exocrine sufficient (PS) or insufficient (PI) children (group A, 8-13 years), adolescents and young adults (group B, 18-23 years) and adults (group C, > 28 years) with cystic fibrosis. **B.** Relative abundance in per cent of DNA viruses, bacteria and fungi in the individual CF sputa. **C.** Proportion of anaerobes among the bacteria in the sputum metagenomes. **D.** Number of sequence reads of bacteriophages and their respective bacterial hosts.

Table S4. Normalized abundance and detection rates of microbial species in the sputum metagenomes differentiated by bacteria, DNA viruses and eukaryotic microbes (molds and fungi).

Table S5. Reads and relative abundance of microbial species in the metagenomes retrieved from agar, water and healthy volunteers differentiated by bacteria, DNA viruses and eukaryotic microbes (molds and fungi). The four healthy adult volunteers each provided four respiratory specimens. Two

samples retrieved from each volunteer were split and processed in parallel. These pairs of technical replicates are identified by ‘name’ and ‘name_F3’.

Table S6. Average relative abundance of the dominant phyla, genera and species contributing to 95% of the bacterial communities in the induced sputa collected from PI and PS people with CF. The data are the basis for the taxonomic cladograms shown in Figure 3.

Table S7. Clonal diversity of *S. aureus* and *P. aeruginosa* populations in CF sputum. The number of reads n_m matching with the sequence of the reference genomes *P. aeruginosa* PAO1 and *S. aureus* Newman, respectively are listed in 1% intervals of the truncated ratio n_m / (n_m + n_{mm}). These values are plotted in Figure 7. Only SNPs were considered that were covered by more than ten sequence reads.