Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung

Dana Willner, Matthew R. Haynes, Mike Furlan, Nicole Hanson, Breeann Kirby, Yan Wei Lim, Paul B. Rainey, Robert Schmieder, Merry Youle, Douglas Conrad, Forest Rohwer

ONLINE DATA SUPPLEMENT

Supplementary Methods and Materials

Generation of viral metagenomes

Immediately following transplant surgery at the University of California San Diego Medical Center (UCSD-MC), lung tissue sections were removed from the explanted lungs as described in (1). Lung gross pathology is demonstrated in Fgiure S7. The post-mortem lungs were received from UCSD-MC following autopsy as described in (1). Using a Dowser homogenizer, each dissected lung subsection was sheared and homogenized with 10 ml of Suspension Medium buffer (SM; 1 M NaCl, 10 mM $MgSO_4$, 50 mM Tris-HCl pH 7.4) in a 50 ml Falcon tube, followed by centrifugation at 2000 rpm for 20 minutes to pellet tissue fragments. The supernatant was removed and processed as described in (2).In brief, samples were treated with an equal volume of dithiothreotol, incubated for 30 minutes at 37 ° C, homogenized, and passed through 0.8 micron and 0.45 micron filters to remove intact cells. The viral particles in each filtrate were isolated and concentrated in a cesium chloride gradient (3). The viral fractions were collected and treated with chloroform and DNAse I. Each sample was verified to be free of contaminating cellular material using epifluoresence microscopy and 16S rDNA PCR prior to DNA isolation (3). Viral DNA was extracted using a CTAB/Formamide protocol as described in (3) and amplified using Phi29 polymerase to generate sufficient template DNA for 454 pyrosequencing. The amplified viral DNA was sequenced on the 454 GS-FLX instrument with multiplexing and Titanium chemistry according to the manufacturer's standard protocols.

Initial bioinformatics of viral metagenomic sequences

Multiplexed viral metagenomic libraries were de-convoluted and de-replicated. Taxonomic assignments were made by comparison to the non-redundant databaseat NCBI using BLASTn and tBLASTx (4). Any sequences with significant BLASTn or tBLASTx similarities to eukaryotic DNA or significant BLASTn similarities to bacterial or archaeal DNA (e-value$<10^{-5}$) were removed from the metagenomes prior to further processing.

Viral taxonomic assignment

A custom database was constructed that included all completely sequenced viral genomes available at NCBI (http://www.ncbi.nlm.nih.gov/sites/genome) and all phage genomes in the phage proteomic tree (http://www.phantome.org/). Redundant genome sequences were removed. Viral taxonomy was assigned based on tBLASTx similarities to the database (E-value $<10^{-5}$, >30% similarity, >50% query coverage). Viral community profiles were generated by GAAS using the top option (i.e., using only the top tBLASTx similarity), with no length normalization, requiring a minimum percent identity of 30% and minimum alignment length of 50% of the query, and with an E-value cutoff of 10^{-5} (5).

Metagenomic reads were mapped to individual viral genomes using BLAT and coverage plots generated by Circos for TTV-3 and HPV-49 and by the Integrated Genome Browser for Pf1 (6,7).Accession numbers for the reference sequences used for coverage plots were NC_014081 (TTV-3 complete genome sequence), NC_001591 (HPV-49 complete genome sequence), and NC_001331 (*Pseudomonas* phage Pf1 complete genome sequence).

Diversity of viral metagenomes

Species richness was estimated for each virome using a similarity-independent method based on contig assembly (8,9). Sequences from each virome were repeatedly sub-sampled using a sample size of 1000 sequences to yield 5X coverage and then assembled into contigs by Circonspect (http://biome.sdsu.edu/Circonspect). The resultant contig spectra were combined with average genome size estimates generated by GAAS using tBLASTx for analysis by PHACCS (5,9). Richness estimates for all viromes are based on a logarithmic model, as this model produced the lowest error (9).

To compare the diversity between viromes, the most likely percentage of viral genotypes shared by each virome pair was calculated using MaxiPhi, a method that performs Monte Carlo simulations based on cross-contig spectra (8,10). MaxiPhi results were combined using non-metric multi-dimensional scaling (NM-MDS) with the isoMDS function in the R package MASS. The percent of non-shared genotypes provided the distance metric used to create dissimilarity matrices for NM-MDS. To determine statistical significance empirically for the percentage of shared genotypes, a simulation was conducted. In brief, 100 artifical viral metagenomes were created from the viral database described above using the program Grinder (http://sourceforge.net/projects/biogrinder). Each metagenome had a randomly determined species richness between 1 and 400 viral genotypes to simulate the range of diversity estimated for the viral metagenomes in this study. This set of metagenomes were compared pairwise to each other using MaxiPhi to create an empirical distribution of percent similarity values. Quantiles of this distribution were used to define statistical significance at the 5% and 1% levels.

Contig assembly and phylogenetic analysis

Viral contigs were assembled using the command line version of the CAP3 assembler and taxonomic assignments were made using BLASTn against the non-redundant database at NCBI (11). Open reading frames in the TTV contig were determined using Prodigal (12). The predicted ORF1 was aligned with all other anellovirus ORF1 protein sequences available from UniProt and phylogenetic analysis was performed using PhyML with 100 bootstrap repetitions (13).

Antibiotic resistance genes in viromes

To identify antibiotic resistance genes, virome sequences were compared to the Antibiotic Resistance Database (ARDB) using BLASTx (14). Similarities with E-values of $<10^{-5}$ over at least 80% of the query sequence length were considered significant. Statistically significant differences between antibiotic resistance determinants in different lobes were evaluated using the non-parametric statistical program XIPE (15). Viral contigs were generated using CAP3, translated in all six frames using TranSeq, and aligned to the sequence with the highest BLAST similarity to determine the correct translation frame (11,16). The translated *mexF* contig from the explant lung was aligned to the set of non-redundant MexF sequences available from UniProt using ClustalW2 (17). Phylogenetic analysis of aligned MexF sequences was performed using MrBayes 3.1 with four independent Markov chains over 500,000 generations using the mixed amino acid model (18).

Caveats

Viral DNA extracted from all lobes was amplified using multiple displacement amplification with Phi29 polymerase prior to sequencing. Phi29 has been shown to preferentially amplify small circular DNAs and could artificially inflate the abundance of sequences from small circular genomes such as HPV and TTV (19,20). All of the pyrosequenced metagenomes used in this study were collected and processed in an identical manner, thus all samples were equally subject to any potential biases due to sampling or amplification. HPV or TTV present in any lobes would have been preferentially amplified and detected, further confirming the unique presence of HPV and TTV in distinct anatomical regions.

A second caveat concerns the risk of contamination of lung autopsy samples during excision and transport, as well as contamination with oral microbiota due to aspiration upon death. To prevent contamination, samples were processed upon receipt in the laboratory in a sterile hood using autoclaved reagents and tools. The large lung sections received were resected further so that only interior tissue (i.e. tissue which that had not been previously been exposed to air during autopsy) was processed and used for viral DNA extraction. Although aspiration of oropharyngeal contents can occur upon death, the autopsy report for this patient indicated that the patient had suffocated due to complete blockage of his central airways, including the trachea, by thick, purulent mucus. The tracheal mucus plug that was physically removed from the patient’s trachea at autopsy had completely sealed off the airway, thus making aspiration of oropharyngeal contents into the lungs unlikely.

References

E1. Willner D, Haynes MR, Furlan M, Schmieder R, Lim YW, Rainey PB, Rohwer F, Conrad D. Spatial distribution of microbial communities in the cystic fibrosis lung. *ISME J*. 2011. Available at: http://dx.doi.org/10.1038/ismej.2011.104. Accessed September 10, 2011.

E2. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE*. 2009;4(10):e7370.

E3. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. *Nat Protoc*. 2009;4(4):470-483.

E4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J. Mol. Biol*. 1990;215(3):403-410.

E5. Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux RK, Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, Rohwer F. The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput. Biol*. 2009;5(12):e1000593.

E6. Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: An information aesthetic for comparative genomics. *Genome Research*. 2009;19(9):1639 -1645.

E7. Nicol JW, Helt GA, Blanchard SG, Raja A, Loraine AE. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics*. 2009;25(20):2730 -2731.

E8. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F. The marine viromes of four oceanic regions. *PLoS Biol*. 2006;4(11):e368.

E9. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics*. 2005;6(1):41.

E10. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*.

2010;466(7304):334-338.

E11. Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res*. 1999;9(9):868-877.

E12. Hyatt D, Chen G-L, LoCascio P, Land M, Larimer F, Hauser L. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11(1):119.

E13. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol*. 2003;52(5):696-704.

E14. Liu B, Pop M. ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res*. 2009;37(Database issue):D443-447.

E15. Rodriguez-Brito B, Rohwer F, Edwards RA. An application of statistics to comparative metagenomics. *BMC Bioinformatics*. 2006;7:162.

E16. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. 2000;16(6):276-277.

E17. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007;23(21):2947-2948.

E18. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. 2001;17(8):754-755.

E19. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. *Genome Research*. 2001;11(6):1095 -1099.

E20. Pinard R, de Winter A, Sarkis G, Gerstein M, Tartaro K, Plant R, Egholm M, Rothberg J, Leamon J. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*. 2006;7(1):216.

Table E1. Characteristics of the viral metagenomes. Eukaryotic sequences were detected using BLASTn and tBLASTx comparisons to the non-redundant database. Sequences with similarities to eukaryotic genomes greater than 90% identity and e-values<10^{-5} were removed prior to further processing.

Lung	**Region**	**Number of Sequences**	**Eukaryotic sequences (%)**
Ex-plant	RUL	16,413	17 (0.10)
Ex-plant	RML	10,575	985 (9.31)
Ex-plant	RLL	10,295	2161 (20.99)
Ex-plant	LUL	24,743	52 (0.21)
Ex-plant	Ling	17,754	4463 (25.14)
Ex-plant	LLLA	44,547	16,368 (36.74)
Ex-plant	LLLP	12,584	443 (3.52)
Post-mortem	RUL	219,333	74,638 (34.03)
Post-mortem	RML	186,228	86,789 (46.60)
Post-mortem	RLL	264,916	121,470 (45.85)
Post-mortem	LUL	123,166	11,139 (9.04)
Post-mortem	Ling	140,683	64,989 (46.2)
Post-mortem	LLL	186,517	79,826 (42.8)

Table E2. Dissimilarity matrices for viral metagenomes in the ex-plant (A) and post-mortem (B) lungs. Each entry in the matrix represents the percentage of unshared viral genotypes in each pair of lung viromes as determined by MaxiPhi. Statistical significance was determined empirically using simulated metagenomes as described in methods. A dissimilarity index of 96.67% was the cutoff for significance at the 0.05 level (indicated by *) and an index of 99.21% was the cutoff at the 0.01 level (indicated by **).

A

	RUL	RML	RLL	LUL	Ling	LLLA	LLLP
RUL	0						
RML	83.33	0					
RLL	83.33	100**	0				
LUL	50	100**	100**	0			
LML	83.33	100**	100**	75	0		
LLLA	83.33	99*	64.81	83.33	79.55	0	
LLLP	100**	100	95.46	83.33	96.14	80	0

B

	RUL	RML	RLL	LUL	Ling	LLL
RUL	0.0					
RML	20.0	0.0				
RLL	57.0	0.5	0.0			
LUL	10.0	38.0	54.0	0.0		
Ling	11.0	85.0	81.0	20.0	0.0	
LLL	97.4*	95.1	97*	97*	99.3**	0.0

Figure E1. Medical data for the explant (A-D) and post-mortem (E-H) patients including clinical microbiology and FEV1 % data (Figures 2A and 2E). Antibiotic resistance of the clinical cultures (Figures 2A and 2E): C = ciprofloxacin, F = ceftazidime, I = imipenem, T = tobramycin, and Z = piperacillin. Medical events (Figures 2A and 2E): PE = pulmonary exacerbation, Pnthx = pneumothorax, and Pan = pancreatitis. Included with each medical condition are the antibiotics that were administered, and M=meropenem, Z=Zosyn (pipericillin plus tazobactam), and N=nafcillin. Computed axial tomography (CAT) scans shown for the explant patient are from age 39 (Figure 2B), age 44 (Figure 2C), and subsequent to the lung transplant (Figure 2D). CAT scans for the post-mortem patient are from age 26 (Figure 2F), age 32 (Figure 2G), and two weeks after Figure 2G (Figure 2H).

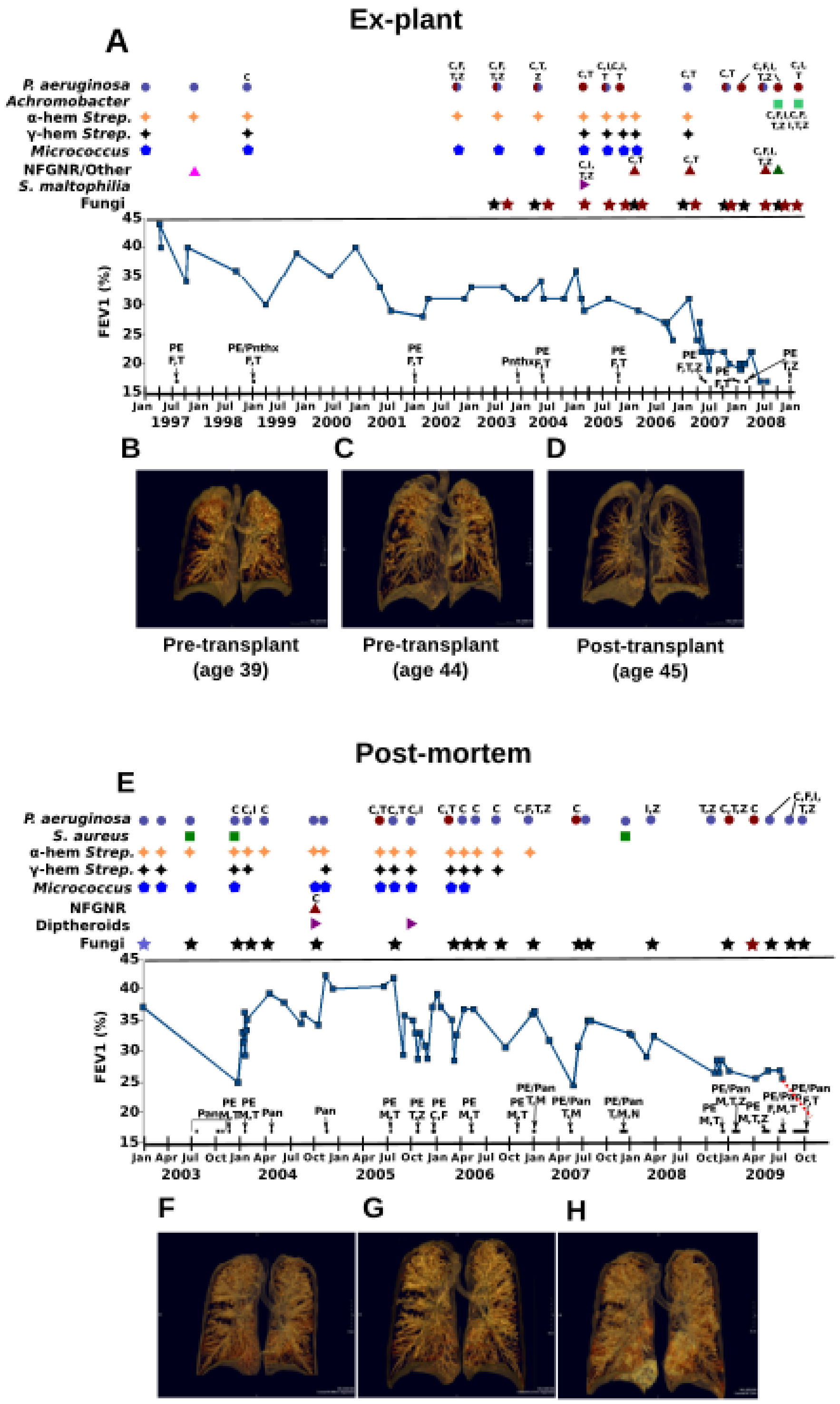

Figure E2. Taxonomic constituency (A,B) of lung viral metagenomes. Viral taxa were identified in the ex-plant (A) and post-mortem (B) lungs using tBLASTx to a custom viral database (e<0.00001, percent similarity > 30%, query coverage > 50%). Each bar respresents the presence of a different virus in the metagenome. Phage are grouped according to the taxonomy of their hosts, and phage of common CF pathogens are indicated in color: *Pseudomonas aeruginosa* in blue, *Burkholderia* spp. in green, *Haemophilus influenzae* in purple, and *Staphylococcus aureus* in red.

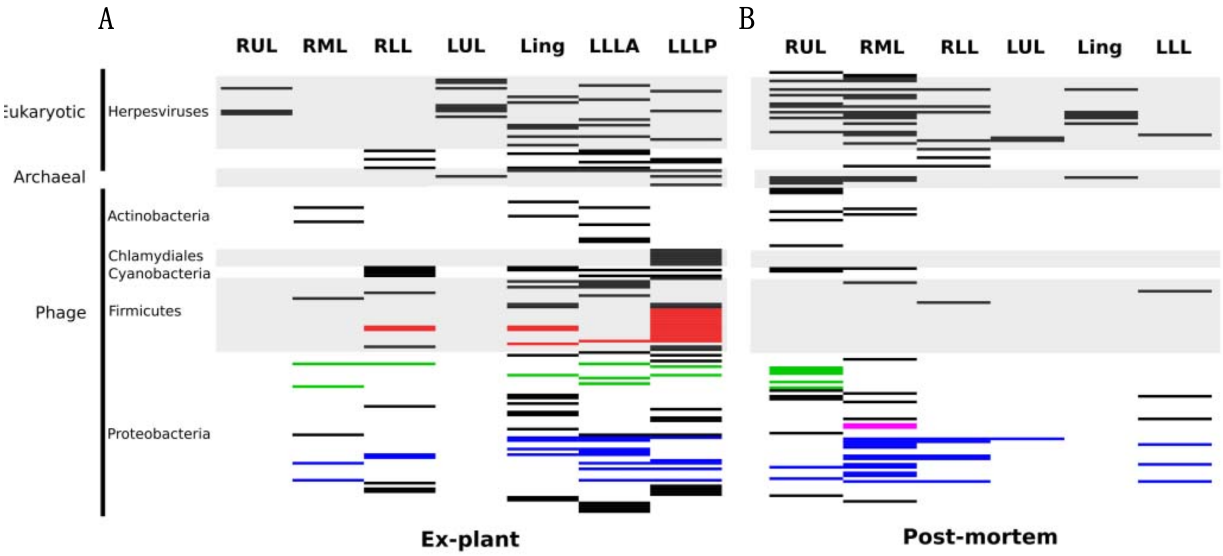

Figure E3. Phylogenetic relationships between predicted ORF1 of post-mortem lung RML Annellovirus contig and ORF1 sequences from known Annellovirus genomes. Numerical labels at branch points represent the percentage of bootstrap trees in which the sequences to the right of the branch point clustered together.

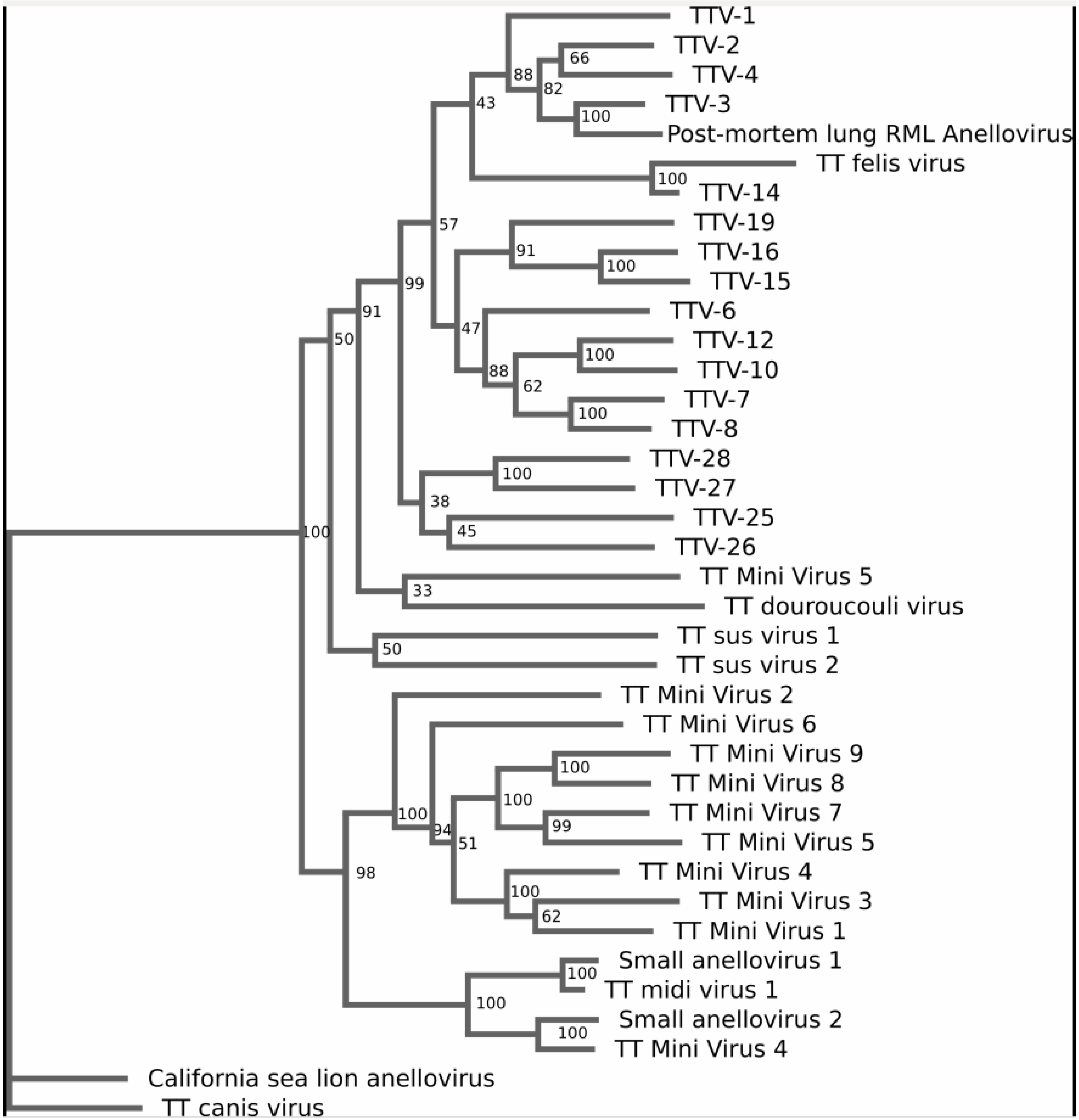

Figure E4. Coverage of phage Pf1 genome by metagenomic sequences from the RML of the post-mortem lung. Each blue bar represents one sequence read. Thin regions of bars represent gaps in the sequence alignment.

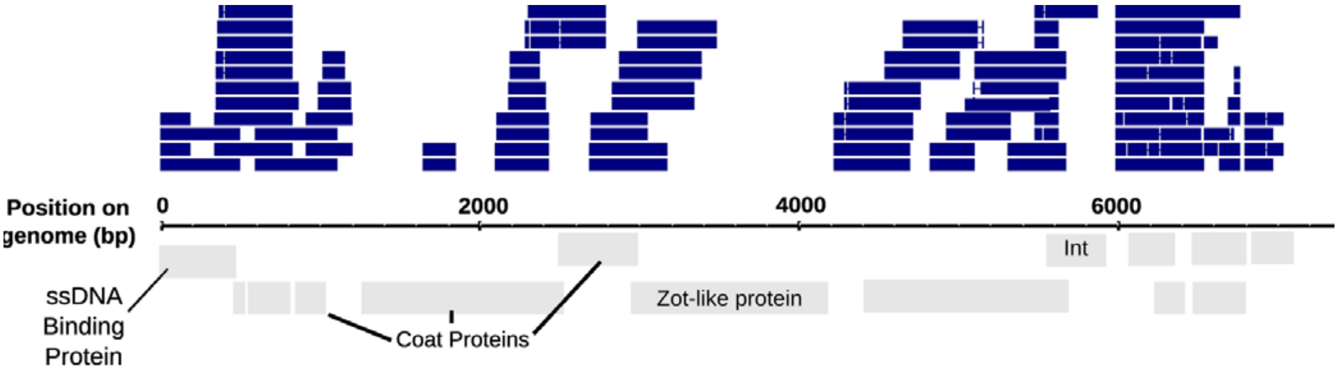

Figure E5. (A,B) Estimated richness of viral genotypes in the ex-plant (A) and post-mortem lungs (B) using contig assemblies of metagenomic sequences as implemented in PHACCs. (C,D) Non-metric multi-dimensional scaling based on the percent of shared viral genotypes bewteen lobes of the ex-plant (C) and post-mortem (D) lungs. The percent of shared genotypes was calculted for each pair of viral metagenomes using MaxiPhi, which conducts Monte Carlo simulations based on cross-contig spectra to determine the relationship between viral communities (reference). The larger the distance between points on the MDS plots, the smaller the percentage of shared viral genotypes.

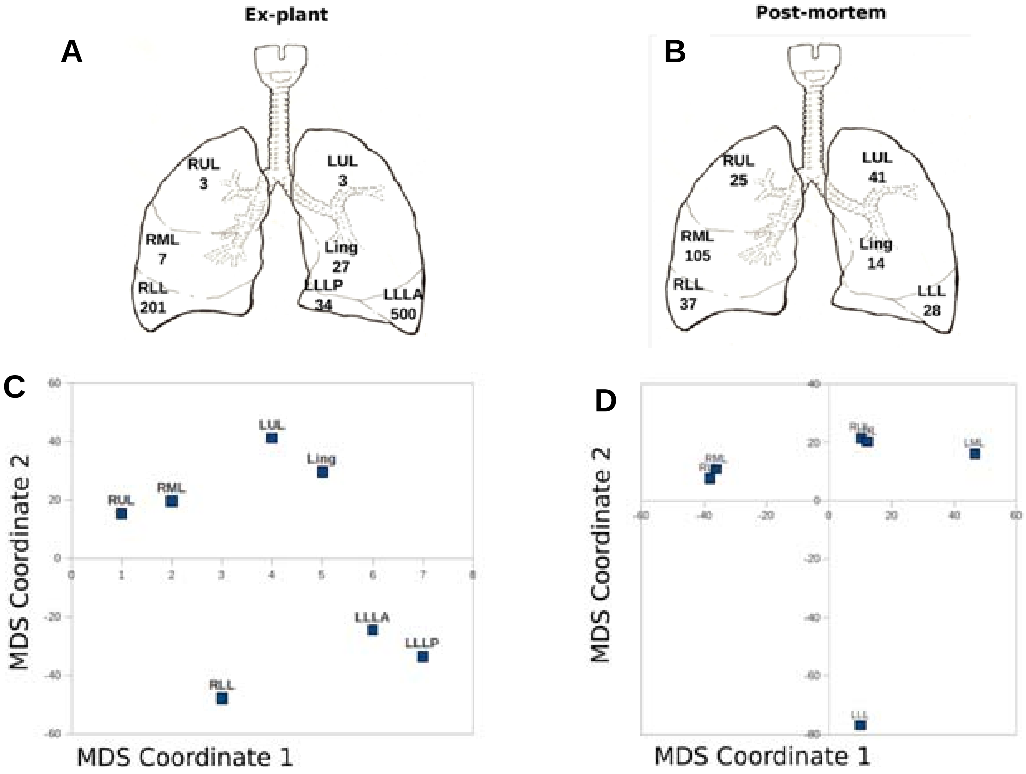

Figure E6. Phylogenetic relationships between the translated mexF sequence from the ex-plant lung RML and non-redundant mexF sequences from UniProt. Accession numbers for UniProt sequences appear in Table S3. Numbers at branch points indicate Bayes values, which represent the proporation of trees in which the sequences to the right of the branch point cluster together. Only non-unity (i.e. not equal to 1.00) Bayes values are shown.

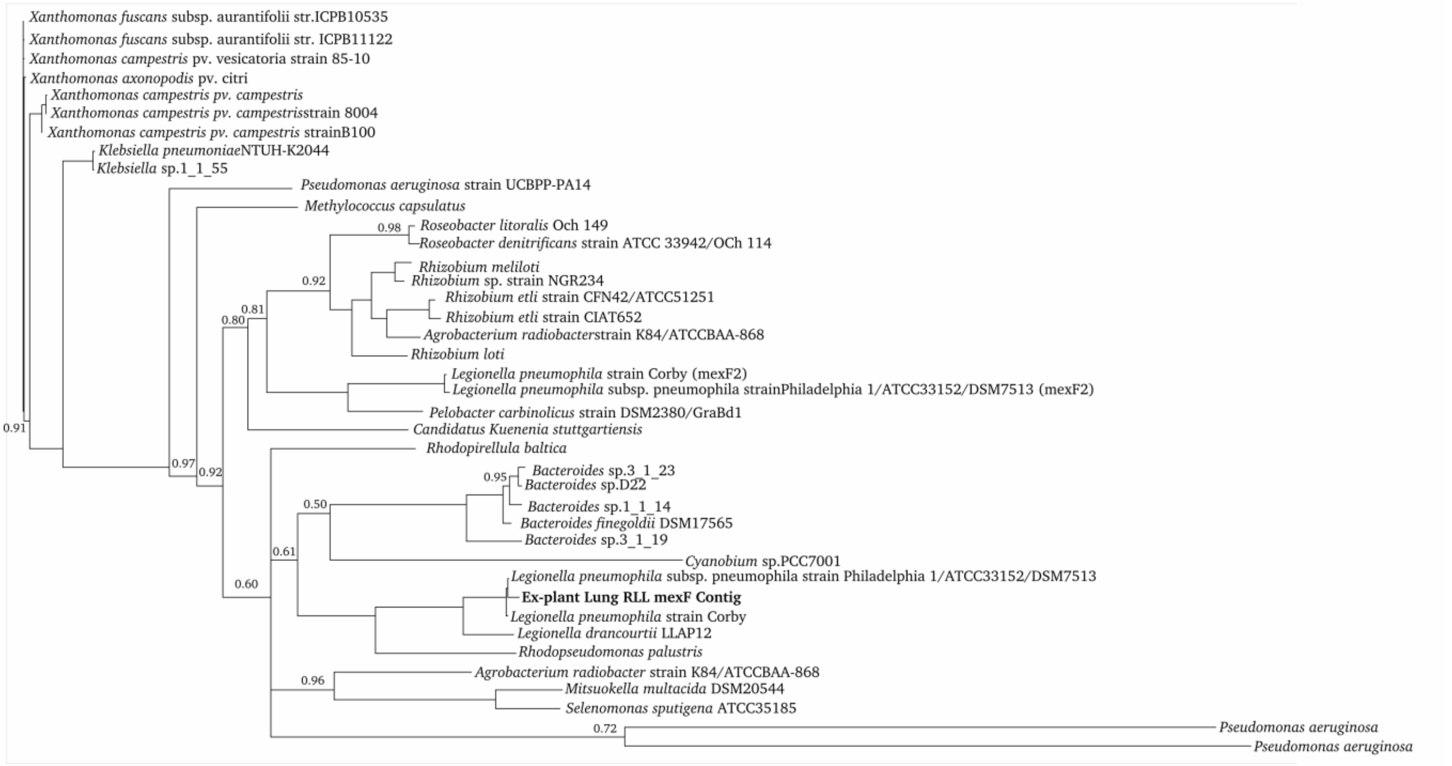
Xanthomonas fuscans subsp. aurantifolii str.ICPB10535
Xanthomonas fuscans subsp. aurantifolii str. ICPB11122
Xanthomonas campestris pv. vesicatoria strain 85-10
Xanthomonas axonopodis pv. citri
Xanthomonas campestris pv. campestris
*Xanthomonas campestris pv. campestris*strain 8004
Xanthomonas campestris pv. campestris strainB100
*Klebsiella pneumoniae*NTUH-K2044
Klebsiella sp.1_1_55
Pseudomonas aeruginosa strain UCBPP-PA14
Methylococcus capsulatus
0.98
Roseobacter litoralis Och 149
Roseobacter denitrificans strain ATCC 33942/OCh 114
Rhizobium meliloti
0.92
Rhizobium sp. strain NGR234
Rhizobium etli strain CFN42/ATCC51251
0.81
0.80
Rhizobium etli strain CIAT652
*Agrobacterium radiobacter*strain K84/ATCCBAA-868
Rhizobium loti
Legionella pneumophila strain Corby (mexF2)
Legionella pneumophila subsp. pneumophila strainPhiladelphia 1/ATCC33152/DSM7513 (mexF2)
Pelobacter carbinolicus strain DSM2380/GraBd1
0.91
Candidatus Kuenenia stuttgartiensis
Rhodopirellula baltica
0.97
0.92
0.95
Bacteroides sp.3_1_23
Bacteroides sp.D22
0.50
Bacteroides sp.1_1_14
Bacteroides finegoldii DSM17565
Bacteroides sp.3_1_19
0.61
Cyanobium sp.PCC7001
Legionella pneumophila subsp. pneumophila strain Philadelphia 1/ATCC33152/DSM7513
0.60
Ex-plant Lung RLL mexF Contig
Legionella pneumophila strain Corby
Legionella drancourtii LLAP12
Rhodopseudomonas palustris
0.96
Agrobacterium radiobacter strain K84/ATCCBAA-868
Mitsuokella multacida DSM20544
Selenomonas sputigena ATCC35185
0.72
Pseudomonas aeruginosa
Pseudomonas aeruginosa

Figure E7. Gross pathology of ex-plant (A) and post-mortem (B) lungs. The ex-plant lungs were photographed intact just after extraction from the patient, and the post-mortem lungs were photographed following dissection at autopsy.