nature biotechnology

OPEN

The draft genome sequence of the ferret (*Mustela putorius furo*) facilitates study of human respiratory disease

Xinxia Peng[1], Jessica Alföldi[2], Kevin Gori[3], Amie J Eisfeld[4], Scott R Tyler[5,6], Jennifer Tisoncik-Go[1], David Brawand[2,22], G Lynn Law[1], Nives Skunca[7,8], Masato Hatta[4], David J Gasper[4], Sara M Kelly[1], Jean Chang[1], Matthew J Thomas[1], Jeremy Johnson[2], Aaron M Berlin[2], Marcia Lara[2,22], Pamela Russell[2,22], Ross Swofford[2], Jason Turner-Maier[2], Sarah Young[2], Thibaut Hourlier[9], Bronwen Aken[9], Steve Searle[9], Xingshen Sun[5], Yaling Yi[5], M Suresh[4], Terrence M Tumpey[10], Adam Siepel[11], Samantha M Wisely[12], Christophe Dessimoz[3,13,14], Yoshihiro Kawaoka[4,15–18], Bruce W Birren[2], Kerstin Lindblad-Toh[2,19], Federica Di Palma[2,22], John F Engelhardt[5,20], Robert E Palermo[1] & Michael G Katze[1,21]

© 2014 Nature America, Inc. All rights reserved.

The domestic ferret (*Mustela putorius furo*) is an important animal model for multiple human respiratory diseases. It is considered the 'gold standard' for modeling human influenza virus infection and transmission[1–4]. Here we describe the 2.41 Gb draft genome assembly of the domestic ferret, constituting 2.28 Gb of sequence plus gaps. We annotated 19,910 protein-coding genes on this assembly using RNA-seq data from 21 ferret tissues. We characterized the ferret host response to two influenza virus infections by RNA-seq analysis of 42 ferret samples from influenza time-course data and showed distinct signatures in ferret trachea and lung tissues specific to 1918 or 2009 human pandemic influenza virus infections. Using microarray data from 16 ferret samples reflecting cystic fibrosis disease progression, we showed that transcriptional changes in the CFTR-knockout ferret lung reflect pathways of early disease that cannot be readily studied in human infants with cystic fibrosis disease.

We performed whole-genome sequencing with DNA from an individual female sable ferret (*M. putorius furo*) and created a genome assembly using ALLPATHS-LG[5]. The draft assembly is 2.41 Gb including gaps, has a contig N50 size of 44.8 kb, a scaffold N50 size of 9.3 Mb and quality metrics comparable to other genomes sequenced using Illumina technology (**Table 1** and **Supplementary Note**). RNA-seq data for annotation were obtained from polyadenylated transcripts using RNA from 24 samples of 21 tissues from male and female ferrets, including both developmental and adult tissues (**Supplementary Table 1**). The genome assembly was annotated using the Ensembl gene annotation system[6] (Ensembl release 70). Protein-coding gene models were annotated by combining alignments of Uniprot[7] mammal and other vertebrate protein sequences and the aforementioned RNA-seq data. The ferret genome can be viewed as unanchored scaffolds along with the Ensembl genome models in both the University of California Santa Cruz and Ensembl genome browser interfaces. We also used the tool liftOver to map the coordinates from the ferret assembly onto the well-finished genome sequence of its phylogenetic neighbor, the domestic dog (*Canis familiaris*; V3.1); this mapping is a useful resource for a genome based on short-read sequencing and facilitates browsing the ferret genome in the surrogate context of dog chromosomes (**Supplementary Fig. 1**).

Using the annotated ferret protein sequences, we constructed a highly resolved phylogenetic tree (**Supplementary Fig. 2**). As expected,

[1]Department of Microbiology, University of Washington, Seattle, Washington, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK. [4]Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin, Madison, Wisconsin, USA. [5]Department of Anatomy and Cell Biology, Carver College of Medicine, University of Iowa, Iowa City, Iowa, USA. [6]Molecular and Cellular Biology Program, Carver College of Medicine, University of Iowa, Iowa City, Iowa, USA. [7]Department of Computer Science, Swiss Federal Institute of Technology (ETH Zurich), Zurich, Switzerland. [8]Swiss Institute of Bioinformatics, Zurich, Switzerland. [9]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [10]Centers for Disease Control and Prevention, Atlanta, Georgia, USA. [11]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, USA. [12]Department of Wildlife Ecology and Conservation, University of Florida, Gainesville, Florida, USA. [13]Department of Genetics, Evolution and Environment, University College London, London, UK. [14]Department of Computer Science, University College London, London, UK. [15]ERATO Infection-Induced Host Responses Project, Japan Science and Technology Agency, Saitama, Japan. [16]Division of Virology, Department of Microbiology and Immunology, Institute of Medical Science, University of Tokyo, Tokyo, Japan. [17]Department of Special Pathogens, International Research Center for Infectious Diseases, Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo, Japan. [18]Laboratory of Bioresponses Regulation, Department of Biological Responses, Institute for Virus Research, Kyoto University, Kyoto, Japan. [19]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden. [20]Center for Gene Therapy, Carver College of Medicine, University of Iowa, Iowa City, Iowa, USA. [21]Washington National Primate Research Center, Seattle, Washington, USA. [22]Present addresses: MRC Functional Genomics Unit, University of Oxford, Oxford, UK (D.B.); Biogen Idec, Cambridge, Massachusetts, USA (M.L.); Division of Biology, California Institute of Technology, Pasadena, California, USA (P.R.); Vertebrate and Health Genomics, The Genome Analysis Center, Norwich, UK (F.D.P.). Correspondence should be addressed to M.G.K. (honey@uw.edu) or R.E.P. (palermor@uw.edu).

Received 20 May; accepted 22 October; published online 17 November 2014; doi:10.1038/nbt.3079

Table 1 Summary details of the ferret genome assembly and associated Ensembl annotation

Mustela putorius furo genome	
Assembly	MusPutFur1.0
Date	June 2011
Total assembly length	2.41 Gb
Total sequence	2.28 Gb
Short read coverage	162 ×
Organization	7,783 unplaced scaffolds
Scaffold N50	9.3 Mb
Ensembl annotation	
Database version	70.1
Date	August 2012
Coding genes	19,910
Noncoding genes	3,614
Pseudogenes	287
Gene transcripts	23,963

Supporting evidence: 1.1×10^9 mRNA-seq reads (Paired end 2 × 100; 220 Gb) from 21 individual tissues including developmental stages, and respiratory tissues from 2009 SOIV infected ferrets.

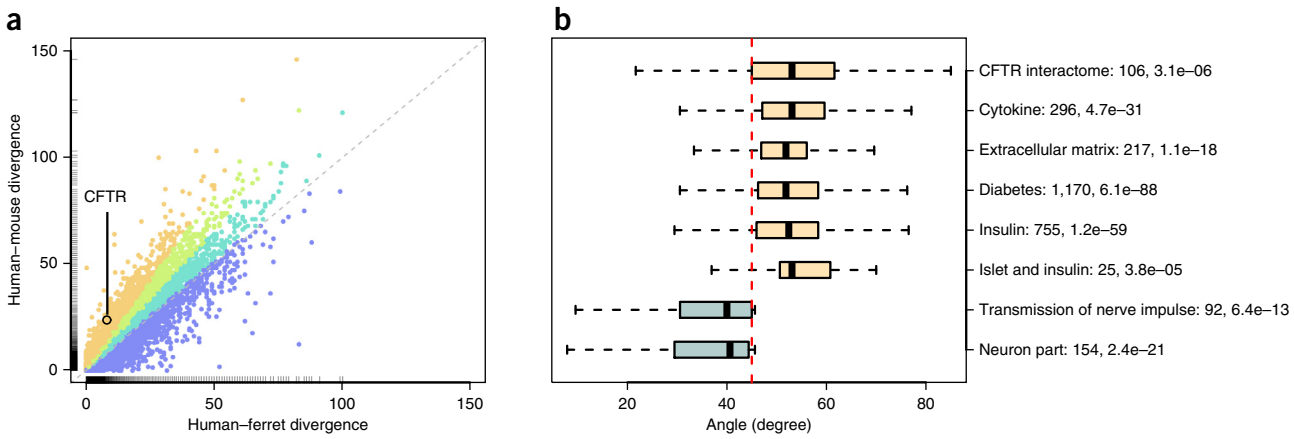
ferret falls within the *Caniformia* suborder of the Carnivores, as represented by the domestic dog, cat, giant panda and walrus, and the support values are high for most clades (**Supplementary Tables 2** and **3** and Methods). Although the clade containing the ferret diverged from a common ancestor before the divergence of the rodent and human/primate lineages, branch lengths in the tree indicate rapid evolution in the rodent clade, which has resulted in less genetic divergence between humans and ferrets than between humans and mice. Indeed, in comparing protein sequences between the species, we found that for 75% of all orthologous triplets, ferret proteins are closer than mouse proteins to human proteins (**Fig. 1a** and **Supplementary Table 4**). For example, the ferret cystic fibrosis transmembrane conductance regulator (CFTR) protein is considerably closer to the human protein than is its mouse counterpart (percentage identities [PAM distance] for ferret to human = 92% [8.1]; mouse to human = 79% [23.3]). Overall, Gene Ontology (GO) terms related to basic cell physiology tend to be enriched among the genes residing in the angular sector, representing the top 25% of genes for which the ferret sequence is closer to human than the mouse ortholog. The enriched GO terms include nucleic acid metabolism, nuclear division, regulation of expression, and protein modification and localization (**Supplementary Fig. 3** and **Supplementary Tables 5** and **6**). Extending this comparison from CFTR to 106 CFTR-interacting proteins, we found that the ferret-to-human protein sequence similarity is significantly greater than the corresponding mouse ortholog (Wilcoxon test P value = 3.1×10^{-6}; **Fig. 1b**). In additional comparisons, we examined gene sets pertinent to cystic fibrosis disease processes including inflammation, lung and pancreatic remodeling, and the regulation of insulin and diabetes; in all cases we found the encoded human proteins to be better conserved in ferret than in mouse (**Fig. 1b** and **Supplementary Fig. 4**). In contrast, proteins encoded by some nervous system–related genes seem to be more divergent from human in ferret than in mouse (**Fig. 1b**). In summary, the overall high sequence similarity between ferret and human proteins shown by these genome-level analyses indicates that many ferret proteins have likely evolved to conserve similar molecular functions as their human protein orthologs.

Next, we investigated whether ferret and human genes exhibit similar tissue expression. We compared the patterns of relative transcript abundance across seven tissues in common between our data set and previously reported human RNA-seq data[8] (**Fig. 1c**). First, we determined the genes with highest relative abundance across all tissues within each species and found that the intersection of these tissue-specific sets between human and ferret was highly significant (chi-squared test P values < 10^{-186}; **Supplementary Table 7**). To refine the sets of tissue-specific genes, we clustered genes with similar expression patterns across the 14 tissue samples (7 from ferret and 7 from human) into 7 disjoint clusters (**Fig. 1c** and **Supplementary Table 8**). This clustering analysis revealed that many ferret and human genes exhibited highly concordant, tissue-specific expression patterns. The assignment of a gene cluster to a specific tissue was evident by its significantly increased expression in that tissue relative to the rest of the tissues of the same species for all comparisons except between skeletal muscle and heart (**Supplementary Table 9**). The similarity between skeletal muscle and heart may be attributed to the presence of striated muscle cells in both tissues. The clusters include transcription factors (TFs) with known tissue specificities in human samples, such as *OLIG2* and *NEUROD2* (brain), *OVOL1* (testis), *MYF6* (heart, skeletal muscle), and lung-specific Iroquois-class homeodomain TFs *IRX2*, *IRX3* and *IRX5*[9]. The same specificity was seen for the ferret tissues. Sequence comparisons showed that ferret TFs in brain, skeletal muscle/heart, lung and kidney gene clusters exhibited even greater similarity to human orthologs than the rest of the ferret genome, suggesting strong conservation of functional regulation (**Supplementary Tables 10** and **11**). Some genes related to immune and inflammatory functions, including TFs associated with Th17 cells (*BATF*, *IRF4*, *AHR*)[10], showed increased expression in ferret and human lungs, which is likely a consequence of the greater proportion of immune cells in this compartment and the possible presence of bronchus-associated lymphoid tissue. The broad similarity between ferret and human tissue-specific gene expression suggests the regulation of gene expression in tissue compartments is also highly conserved.

Ferrets are frequently used as a model for human influenza virus infection, in part due to the similar distribution of viral attachment receptors in the respiratory tract of humans and ferrets[2,11]. We used our genome sequence to profile the transcriptional response of ferrets to pandemic influenza virus. To this end, we infected ferrets with either of two human pandemic influenza viruses—the H1N1 2009 pandemic virus A/CA/04/2009 (CA04) or the reconstructed H1N1 1918 pandemic virus (1918)—and collected samples from both the upper (trachea) and the lower (lung) respiratory tract at 1, 3 and 8 days postinfection (dpi) for transcriptome analysis (**Supplementary Table 12** and **Supplementary Figs. 5** and **6**). To increase the coverage of our transcriptome analysis beyond standard Ensembl annotated genes, including non-polyadenylated transcripts, we performed RNA-seq on total RNA after ribosomal RNA depletion (Methods). To augment standard Ensembl annotation, we predicted additional transcript models using the RNA-seq data collected from both lung and trachea samples from these infected animals and the tissue samples described in the previous paragraph (Online Methods and **Supplementary Table 13**). Additional analyses indicate that the transcripts derived by RNA-seq are enriched with novel protein-coding isoforms and polyadenylated and non-polyadenylated intergenic noncoding RNAs (**Supplementary Figs. 7** and **8**). To make these genomic resources more accessible for gene expression profiling, we also designed and validated two versions of ferret-specific oligonucleotide microarrays: version 1 interrogates 23,582 Ensembl annotated transcripts plus 13,368 intergenic transcripts derived from RNA-seq analysis of ferret mRNAs; version 2 provides broader coverage with probes for an additional 27,288 intergenic transcripts from RNA-seq analysis of ferret total RNAs (**Supplementary Table 14**, **Supplementary Figs. 9–16** and **Supplementary Note**).

As quantified by RNA-seq, host transcriptional changes were much more extensive in infected trachea (9,869 differentially expressed [DE]

© 2014 Nature America, Inc. All rights reserved.



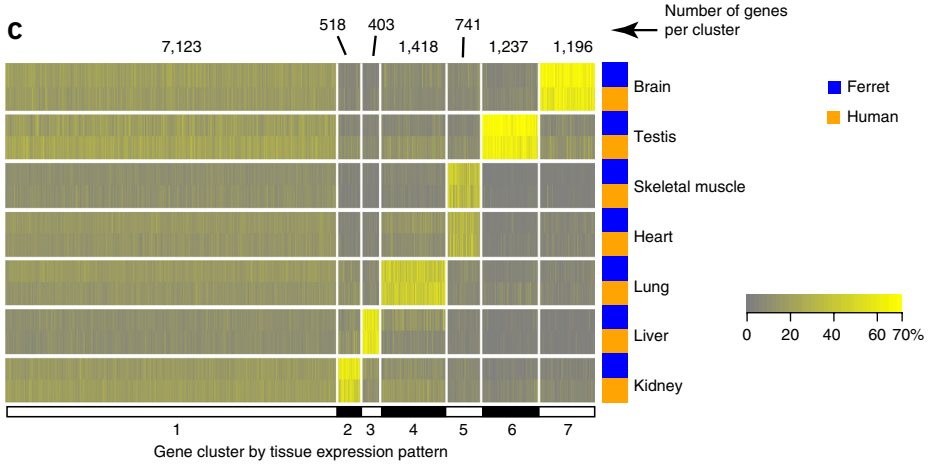
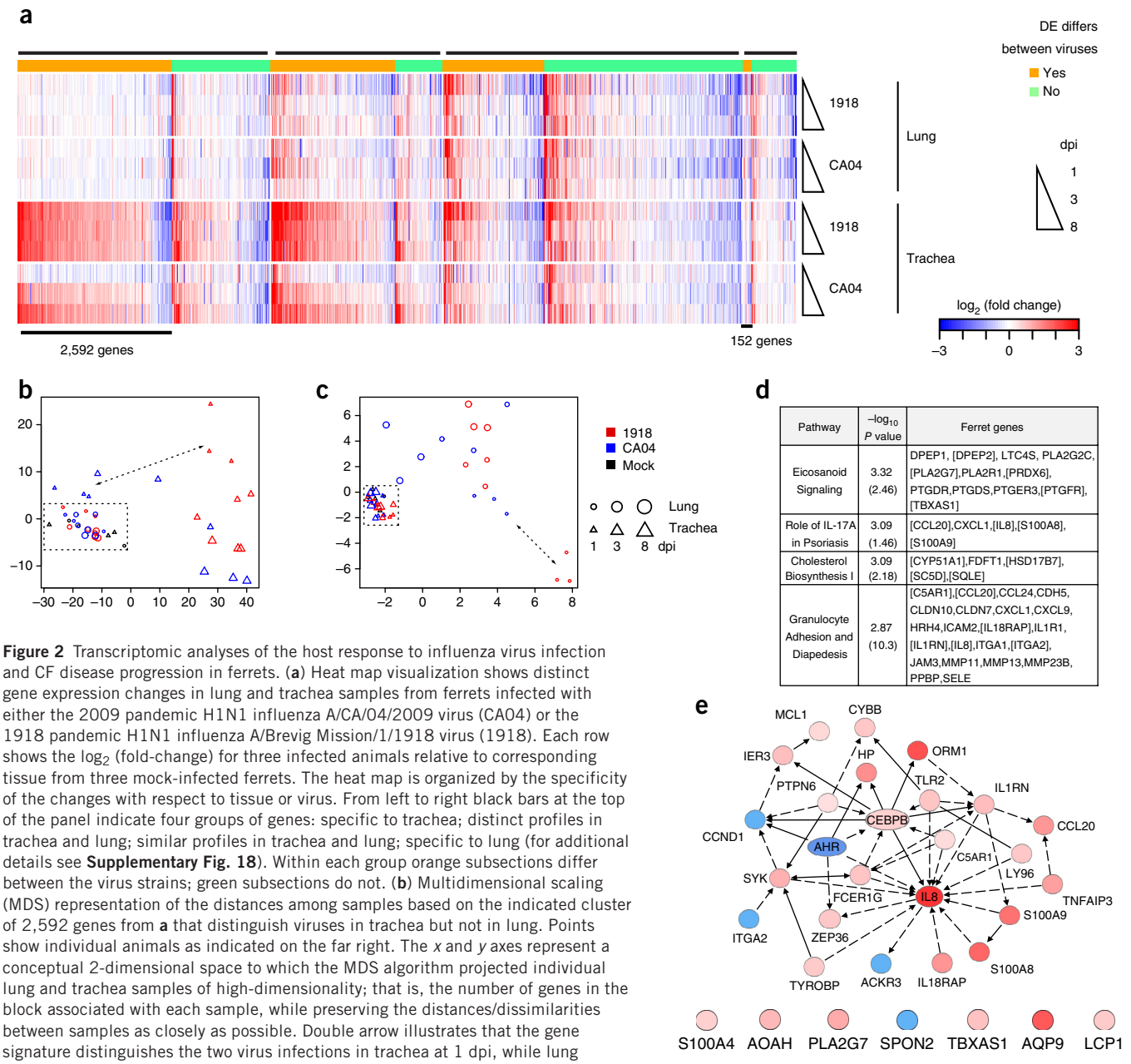


Figure 1 Cross-species comparisons show that ferret protein sequence and tissue-specific expression are similar to that of human. (**a**) Scatter plot of human vs. mouse protein divergence in Point Accepted Mutation (PAM) metric (*y* axis) against the corresponding human vs. ferret protein divergence (*x* axis). Proteins appear above the 45° diagonal (gray dashes) when the ferret sequence is closer to the human sequence than the corresponding mouse sequence. The angle of the line to each protein from the origin is directly related to the ratio of mouse divergence from human sequence and ferret divergence from the human sequence. A greater angle from the origin indicates greater divergence. The quartiles of the distribution of these ratios are displayed in different colors (blue being the least conserved in ferret relative to mouse, and orange-brown being the most conserved). Hatched lines on the axes show the metric distributions for the individual species (**Supplementary Table 4**). (**b**) Boxplots of the angles represented in **a** for proteins in eight selected biological functions. For gene sets related to CF (light yellow), human protein sequence is better conserved in ferret than in mouse. The gene sets highlighted in light yellow are relevant to CFTR function, lung inflammation and remodeling, and insulin secretory defects in CF-related diabetes. For two nervous system–related gene sets (light blue; neuron part (GO:0097458) and transmission of nerve impulse (GO:0019226)), human protein sequence tended to be more conserved in mouse. Next to each function is the number of proteins in the function and the *P* value from one-sided Wilcoxon signed rank test comparing the human-ferret (*x* axis in **a**) vs. human-mouse (*y* axis in **a**) divergence in PAM metric. (**c**) K-means clustering of ferret-human orthologous genes by their tissue expression patterns reveals similarities in tissue specificity. The color scale represents relative abundance across all tissues within each species and is saturated at 70%. Vertical partitions correspond to the seven clusters of genes from the optimal clustering, with numbers of genes per cluster appearing on the top. Horizontal groupings are organized by tissue with ferret and human pairings denoted by the color bar at the side, and highlight the tissue specificity of clusters 2 through 7.

© 2014 Nature America, Inc. All rights reserved.

genes, adjusted *P* value < 0.01) than in infected lung (4,646 DE genes), and the kinetics of the response differed by virus and compartment (**Supplementary Fig. 17**). In the trachea, the 1918 virus induced a pronounced transcriptional response, both in the number of DE genes and in the magnitude of their changes, that commenced at 1 dpi and was largely sustained through day 8; in contrast, infection with the CA04 virus resulted in a gradual escalation of overall transcriptional changes in these same genes, resulting in peak expression by 8 dpi. Different kinetics occurred in the lung, where both viruses induced a similar number of DE genes at 1 dpi, followed by a decline to far fewer DE genes by day 8. A detailed tissue-by-virus comparison revealed distinct transcriptional signatures that differentiated the response to the 1918 and CA04 viruses in the two respiratory tissues (**Fig. 2a**, **Supplementary Fig. 18** and **Supplementary Table 15**). Within the trachea-specific host response, a subset of 2,592 ferret genes distinguished the two viruses, with extensive perturbation at 1 dpi in response to the 1918 virus and minimal alteration in response to CA04 (**Fig. 2b**). This gene set has an over-representation of diverse biological processes such as Apoptosis Signaling, NGF Signaling and Ceramide Signaling (one-sided Fisher exact test *P* values of 4.28×10^{-7}, 1.61×10^{-6} and 3.76×10^{-6}, respectively; **Supplementary Table 16**). Related lipid-receptor signaling systems, such as sphingosine-1-phosphate receptor signaling, can protect the host from influenza virus-induced "cytokine storm" by inhibiting pro-inflammatory responses[12,13]. Similarly, some DE transcripts were exclusively observed within the lung compartment, with a subset of 152 ferret genes that differentiated the two virus infections (**Fig. 2c**). Within this subset, we observed enrichment of Prothrombin Activation Pathway and differential expression of *Il13* and *Il20*, associated with Role of Cytokines in Mediating Communication between Immune Cells (*P* value of 1.53×10^{-2}), which are produced by pulmonary innate lymphoid cells[14] and maturing dendritic cells[15], respectively. In summary, the ferret genomic resources described here enabled a side-by-side comparison of ferret transcriptional responses to two human pandemic influenza viruses. The results revealed that the host response to the two pandemic viruses differs in a tissue compartment–dependent manner.



Pathway	$-\log_{10}$ *P* value	Ferret genes
Eicosanoid Signaling	3.32 (2.46)	DPEP1, [DPEP2], LTC4S, PLA2G2C, [PLA2G7],PLA2R1,[PRDX6], PTGDR,PTGDS,PTGER3,[PTGFR], [TBXAS1]
Role of IL-17A in Psoriasis	3.09 (1.46)	[CCL20],CXCL1,[IL8],[S100A8], [S100A9]
Cholesterol Biosynthesis I	3.09 (2.18)	[CYP51A1],FDFT1,[HSD17B7], [SC5D],[SQLE]
Granulocyte Adhesion and Diapedesis	2.87 (10.3)	[C5AR1],[CCL20],CCL24,CDH5, CLDN10,CLDN7,CXCL1,CXCL9, HRH4,ICAM2,[IL18RAP],IL1R1, [IL1RN],[IL8],ITGA1,[ITGA2], JAM3,MMP11,MMP13,MMP23B, PPBP,SELE

Figure 2 Transcriptomic analyses of the host response to influenza virus infection and CF disease progression in ferrets. (**a**) Heat map visualization shows distinct gene expression changes in lung and trachea samples from ferrets infected with either the 2009 pandemic H1N1 influenza A/CA/04/2009 virus (CA04) or the 1918 pandemic H1N1 influenza A/Brevig Mission/1/1918 virus (1918). Each row shows the $\log_2$ (fold-change) for three infected animals relative to corresponding tissue from three mock-infected ferrets. The heat map is organized by the specificity of the changes with respect to tissue or virus. From left to right black bars at the top of the panel indicate four groups of genes: specific to trachea; distinct profiles in trachea and lung; similar profiles in trachea and lung; specific to lung (for additional details see **Supplementary Fig. 18**). Within each group orange subsections differ between the virus strains; green subsections do not. (**b**) Multidimensional scaling (MDS) representation of the distances among samples based on the indicated cluster of 2,592 genes from **a** that distinguish viruses in trachea but not in lung. Points show individual animals as indicated on the far right. The *x* and *y* axes represent a conceptual 2-dimensional space to which the MDS algorithm projected individual lung and trachea samples of high-dimensionality; that is, the number of genes in the block associated with each sample, while preserving the distances/dissimilarities between samples as closely as possible. Double arrow illustrates that the gene signature distinguishes the two virus infections in trachea at 1 dpi, while lung samples show no separation (dotted rectangle). (**c**) As in **b**, for the indicated cluster of 152 genes that is differentially regulated in lung but not trachea tissues and separates the two virus strains on 1 dpi. (**d**,**e**) Differential transcriptional responses in an experiment comparing lung samples from 15-day-old CF ferrets ($n = 3$) vs. non-CF ferrets ($n = 5$). (**d**) Similar pathways enriched in genes differentially expressed in 15-day-old CF ferret lung samples and CF human bronchial brushings, derived from Ingenuity Pathway Analysis. The values in parentheses are the enrichment *P* values for the corresponding pathways in the genes differentially expressed in CF human bronchial brushing[19]. In brackets are genes that were differentially expressed in both ferret and human CF/non-CF comparisons. (**e**) Network illustration of 32 genes of the function "inflammatory response", which were differentially expressed in the same direction in ferret and human CF data sets (for additional details see **Supplementary Fig. 20**). Red and blue shading reflects the extent of increased or decreased expression, respectively, in CF relative to non-CF individuals. A solid line between two genes indicates direct interaction(s) among them and a dotted line for indirect interaction(s), as documented in the literature.

Genetically engineered cystic fibrosis ferrets model two key components of disease not observed in cystic fibrosis mice, lung disease[16,17] and diabetes[18]. To investigate cystic fibrosis disease progression in the ferret model, we carried out microarray expression analysis on lung specimens from newborn and 15-day-old CFTR knockout (CF) and normal (non-CF) ferrets. In newborn animals, genotypic differences in transcriptomes were limited; 472 DE protein-coding genes were identified using a relaxed threshold (absolute fold-change ≥ 1.4, *P* value ≤ 0.1). Nonetheless, functional analyses of these DE genes showed disturbances in several canonical pathways, including Coagulation System, Primary Immunodeficiency Signaling, Serotonin Receptor Signaling and Signaling in T Helper Cells (**Supplementary Table 17**). Genotype-dependent gene expression differences between 15-day-old animals were much more extensive (1,468 DE protein-coding genes, absolute fold-change ≥ 1.5, false discovery rate ≤ 0.05) and included expression changes in genes from

© 2014 Nature America, Inc. All rights reserved.

pathways involved in Cholesterol Biosynthesis, Eicosanoid Signaling, Granulocyte Adhesion and Diapedesis, and IL8 regulation (**Fig. 2d**,**e** and **Supplementary Table 17**). In a previous study, gene expression in these pathways was also significantly perturbed in human bronchial brushings from adults with cystic fibrosis[19] (**Supplementary Table 17**). Further, in CF ferrets, changes in the expression of most of these genes were highly positively correlated (ANOVA *P* value 4.9×10^{-5}, Pearson correlation coefficient 0.63) with that in human cystic fibrosis samples, consistent with the overall positively correlated changes in expression between day 15 CF ferret and human cystic fibrosis samples (**Supplementary Fig. 19**). The exception of some cholesterol biosynthesis pathway genes may be the result of variation in epithelia sampled in the ferret (intact lung) and human (conducting larger airways), or differences in cystic fibrosis disease status between infants and adults.

Similar expression changes in the CF ferret and human cystic fibrosis data sets were also evident at the level of broader biological functions such as Chronic Inflammatory Disorder, Cell Movement of Phagocytes and Inflammatory Response (**Supplementary Table 18**). As anticipated, *IL8* was one of the most significantly increased inflammatory genes in older CF ferret (14.6-fold upregulated) and human cystic fibrosis (11.7-fold upregulated) samples (**Supplementary Table 17**), consistent with a dominant role of *IL8* in cystic fibrosis lung disease[20]. Indeed, many genes associated with *IL8* regulation, including *CCL20*, *S100A8*, *S100A9*, *IL18RAP*, *IL1RN* and *ITGA2*, were differentially regulated concordantly in CF ferret and human cystic fibrosis lung samples (**Fig. 2e** and **Supplementary Fig. 20**). Although these findings suggest commonalities in the pathways of cystic fibrosis inflammation between the two species, it is worth noting that the dominant bacterial pathogens of the lung are distinctly different between CF ferret and humans with this disease—*Pseudomonas aeruginosa*[21] in humans and enteric pathogens in both young and old CF ferrets[16,17]. Thus, the predominant gene pathways involved in cystic fibrosis inflammatory responses seem to be conserved across ferrets and humans, and to be largely independent of the pathogen's taxa. Of the DE genes for which expression changed in opposite direction between ferret and human data sets, one of the most significant functional pathways included 19 genes associated with Cell Movement (**Supplementary Fig. 20**). This suggests that there are differences in the extent of injury, repair and/or migratory inflammatory cell infiltrates between the ferret and human data sets. Such differences are not surprising given the larger number of DE genes associated with Granulocyte Adhesion and Diapedesis (**Supplementary Table 17**) and inflammation (**Supplementary Table 18**) in the older human cystic fibrosis samples. Despite these differences, the overall positively correlated expression changes, especially the high concordance in key cystic fibrosis-related pathways and functions between 15-day-old CF ferret and adult human cystic fibrosis samples, suggest that many disease changes associated with adult cystic fibrosis in humans may begin in infancy. Thus, the CF ferret represents a tractable model by which to systemically address disease progression–related changes in gene expression at anatomical sites not possible in humans.

Ferrets are extensively used to study human diseases such as influenza virus infection and cystic fibrosis, but the lack of genome sequence information has limited the ability to understand ferret transcriptional responses. Our transcriptomic analyses of the host response to human pandemic influenza virus infection and of cystic fibrosis disease progression in ferrets illustrate how the availability of the ferret genome sequence can enhance the sophistication of ferret respiratory-disease models. The analyses revealed high protein-sequence similarity and shared tissue-expression patterns between ferret and human, suggesting the potential utility of ferret models in a broader set of diseases. The ferret genome will also prove valuable to investigators exploring the conservation genomics of the highly imperiled North American black-footed ferret (*Mustela nigripes*), which is a congener to *M. putorius furo*[22]. The black-footed ferret underwent a population bottleneck in the 1980s leading to greatly diminished genetic diversity, and the resulting congenital defects include reduced immune capacity and anomalies in male fertility[23]. The genomic resources presented here can aid genetic analysis of these defects and the ongoing captive breeding program necessary for the survival of the species[24].

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The *M. putorius furo* genome, GenBank, AEYP00000000; access to assembled contigs and the derived unplaced genomic scaffolds. The assembly can also be found at http://uswest.ensembl.org/Mustela_putorius_furo/Info/Index. Trachea RNA-seq data, SRA, SRP033621; connected to the following BioProjects: genomic reads, PRJNA59869; mRNA-seq reads for tissue survey and lung samples from the influenza model, PRJNA78317. Microarray data sets, GEO: GSE49060 (influenza arrays) and GSE49061 (CF arrays). Agilent array using the developed designs (IDs 048471 and 048472) can be ordered via the Agilent eArray utility (https://earray.chem.agilent.com). Additional community resources related to the paper, including the genomic coordinates for the intergenic and nonpolyadenylated transcripts, and results for the ferret to dog liftOver can be found at http://ucsc.viromics.washington.edu/genomes/ferretGenome.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This project was funded in whole or in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH), Department of Health and Human Services, under Contract Nos. HHSN272200800060C and HHSN272201400005C and Public Health Service Grant P51OD010425 (M.G.K.). For the Broad Institute of MIT and Harvard, this project was funded in whole or in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900018C. D.J.G. also received training grants 9T32OD010423-06 and 5T32RR023916-05 from the NIH Office of the Director. J.F.E., S.R.T., X.S. and Y.Y. (University of Iowa) were supported under the NIH National Institute of Diabetes and Digestive and Kidney Diseases grants R37 DK047967, R24 DK096518 and P30 DK054759, and National Heart, Lung, and Blood Institute grant R01 HL108902. C.D. acknowledges support by SNSF advanced researcher fellowship (#136461). A.S. receives support under NIH National Institute of General Medical Sciences R01 GM102192. For S.S., B.A. and T.H., the Wellcome Trust Sanger Institute is operated by Genome Research Limited, a charity registered in England with number 1021457 and a company registered in England with number 2742969. K.L.-T. also receives support under EURYI ERC. We thank Marshall Farms, New York, for providing three adult sable female ferrets (*M. putorius furo*; 421 days old).

AUTHOR CONTRIBUTIONS

X.P., R.E.P., J.F.E., J.T.-G., A.J.E. and S.M.W. wrote the paper with input from other authors. F.D.P., J.A., B.W.B. and K.L.-T. oversaw genome and transcriptome sequencing, and related computational efforts at the Broad Institute. D.B., P.R. and J.T.-M. performed transcriptome analysis, initial genome annotation and initial expression analysis. J.J. assisted in coordinating samples and sequencing data at the Broad Institute. M.L. and R.S. performed library construction for sequencing. A.M.B. and S.Y. generated the genome assembly. T.M.T. and Y.K. provided influenza infected ferret tissues listed in **Supplementary Table 1**. T.H., B.A. and S.S. were responsible for Ensembl annotation pipeline.

© 2014 Nature America, Inc. All rights reserved.

Y.K. and A.J.E. oversaw the influenza model at the University of Wisconsin. M.H. implemented the ferret infection protocol and generated primary virological data. D.J.G. and M.S. did immunohistochemical staining for influenza A virus antigen. Y.K., A.J.E. and D.J.G. interpreted overall biological outcomes of the influenza model. J.F.E., X.S. and Y.Y. provided normal ferret tissues for genome and transcriptome sequencing, and RNA samples from the CF ferret model. S.M.K. isolated and characterized RNA from normal ferret tissues. J.C. and M.J.T. isolated RNA from samples from the ferret influenza infection model and generated ribosomally depleted RNA. M.J.T. generated total RNA-seq libraries for RNA from ferret trachea and coordinated the generation of trachea total RNA-seq data. G.L.L. oversaw sample handling and coordinated data generation for the ferret influenza model. K.G. and C.D. performed the phylogenetic and ferret-to-human and mouse-to-human comparisons. N.S. generated the GO term enrichments for the differing angular quadrants of **Figure 1a**. S.R.T. evaluated relative conservation of human genes between ferret or mouse for gene sets associated with biomedical models. A.S. advised on comparative genomics analyses. X.P. did the final analysis of all the RNA-seq data, including generating the expanded annotation, evaluating the tissue specificity, and performing the differential expression analysis for the ferret influenza model. J.T.-G. and R.E.P. provided the associated functional interpretation. X.P. also generated the designs for the ferret microarrays. G.L.L. coordinated sample handling and data generation with the ferret microarrays; J.C. generated the microarray data; X.P. performed the statistical comparisons; R.E.P., X.P., J.F.E. and S.R.T. performed the functional interpretation of transcriptional changes between CF ferret and human samples. R.E.P. and M.G.K. coordinated contributions between the collaborating laboratories.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/.

© 2014 Nature America, Inc. All rights reserved.

1. Tripp, R.A. & Tompkins, S.M. Animal models for evaluation of influenza vaccines. *Curr. Top. Microbiol. Immunol.* **333**, 397–412 (2009).
2. van Riel, D. *et al.* Human and avian influenza viruses target different cells in the lower respiratory tract of humans and other mammals. *Am. J. Pathol.* **171**, 1215–1223 (2007).
3. Belser, J.A., Katz, J.M. & Tumpey, T.M. The ferret as a model organism to study influenza A virus infection. *Dis. Model. Mech.* **4**, 575–579 (2011).
4. Schrauwen, E.J. *et al.* Possible increased pathogenicity of pandemic (H1N1) 2009 influenza virus upon reassortment. *Emerg. Infect. Dis.* **17**, 200–208 (2011).
5. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **108**, 1513–1518 (2011).
6. Curwen, V. *et al.* The Ensembl automatic gene annotation system. *Genome Res.* **14**, 942–950 (2004).
7. UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* **41**, D43–D47 (2013).
8. Cabili, M.N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
9. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
10. Ciofani, M. *et al.* A validated regulatory network for Th17 cell specification. *Cell* **151**, 289–303 (2012).
11. van Riel, D. *et al.* H5N1 virus attachment to lower respiratory tract. *Science* **312**, 399 (2006).
12. Walsh, K.B. *et al.* Suppression of cytokine storm with a sphingosine analog provides protection against pathogenic influenza virus. *Proc. Natl. Acad. Sci. USA* **108**, 12018–12023 (2011).
13. Marsolais, D. *et al.* A critical role for the sphingosine analog AAL-R in dampening the cytokine response during influenza virus infection. *Proc. Natl. Acad. Sci. USA* **106**, 1560–1565 (2009).
14. Neill, D.R. *et al.* Nuocytes represent a new innate effector leUK.ocyte that mediates type-2 immunity. *Nature* **464**, 1367–1370 (2010).
15. Wolk, K. *et al.* Maturing dendritic cells are an important source of IL-29 and IL-20 that may cooperatively increase the innate immunity of keratinocytes. *J. Leukoc. Biol.* **83**, 1181–1193 (2008).
16. Sun, X. *et al.* Disease phenotype of a ferret CFTR-knockout model of cystic fibrosis. *J. Clin. Invest.* **120**, 3149–3160 (2010).
17. Sun, X. *et al.* Lung phenotype of juvenile and adult cystic fibrosis transmembrane conductance regulator-knockout ferrets. *Am. J. Respir. Cell Mol. Biol.* **50**, 502–512 (2014).
18. Olivier, A.K. *et al.* Abnormal endocrine pancreas function at birth in cystic fibrosis ferrets. *J. Clin. Invest.* **122**, 3755–3768 (2012).
19. Ogilvie, V. *et al.* Differential global gene expression in cystic fibrosis nasal and bronchial epithelium. *Genomics* **98**, 327–336 (2011).
20. Bonfield, T.L. *et al.* Inflammatory cytokines in cystic fibrosis lungs. *Am. J. Respir. Crit. Care Med.* **152**, 2111–2118 (1995).
21. Ciofu, O., Hansen, C.R. & Hoiby, N. Respiratory bacterial infections in cystic fibrosis. *Curr. Opin. Pulm. Med.* **19**, 251–258 (2013).
22. Ouborg, N.J., Pertoldi, C., Loeschcke, V., Bijlsma, R.K. & Hedrick, P.W. Conservation genetics in transition to conservation genomics. *Trends Genet.* **26**, 177–187 (2010).
23. Wisely, S.M., Buskirk, S.W., Fleming, M.A., McDonald, D.B. & Ostrander, E.A. Genetic diversity and fitness in black-footed ferrets before and during a bottleneck. *J. Hered.* **93**, 231–237 (2002).
24. Wisely, S.M., McDonald, D.B. & Buskirk, S.W. Evaluation of the genetic management of the endangered black-footed ferret (Mustela nigripes). *Zoo Biol.* **22**, 287–298 (2003).

ONLINE METHODS

Animal usage was performed under protocols approved by the Institutional Animal Care and Use Committees (IACUCs) at the University of Wisconsin School of Veterinary Medicine or the University of Iowa. Appropriate biosafety containment was used in the course of infections with the indicated influenza strains.

Genome sequencing and assembly. Three adult sable female ferrets (*M. putorius furo*; 421 days old) obtained from Marshall Farms (via John Engelhardt, Iowa) were sacrificed and specimens sent to the Broad Institute for heterozygosity testing. The individual ID#1420 was selected for sequencing due to its low heterozygosity. The ferret DNA was sequenced to 90× total coverage by Illumina sequencing technology and was composed of 45× coverage using 180 bp fragment libraries, 42× coverage using 3 kb sheared jumping libraries, 2× coverage using 6-14 kb sheared jumping libraries and 1× coverage using ShARC jumping libraries. The reads were assembled into MusPutFur1.0 (accession number AEYP00000000.1) using ALLPATHS-LG[5]. The *M. putorius furo* genome has been reported to have a karyotype of 40 chromosomes[25]. The draft assembly is 2.41 Gb in size and is composed of 2.28 Gb of sequence plus gaps between contigs. The ferret genome assembly has a contig N50 size of 48.8 kb, a scaffold N50 size of 9.3 Mb and quality metrics comparable to other Illumina genomes.

RNA sequencing and assembly. A panel of 24 ferret samples from multiple tissues was RNA sequenced to aid with genome annotation. Developmental (3 staged embryos) and uninfected adult tissues (the individual used for genome sequencing) were obtained from sable ferrets (obtained by the Engelhardt laboratory, Iowa). Two pooled RNA samples were prepared from ferrets that had been infected with strains of 2009 pandemic H1N1; one pool was generated from lung and trachea specimens collected in the laboratory of T. Tumpey (CDC) from two ferrets sacrificed 3 days after infection with A/Mexico/4482/2009 (H1N1). The other RNA pool was generated from materials from laboratory of Y. Kawaoka (Wisconsin), using spleens harvested at days 3 and 6 following infection of two ferrets with A/Wisconsin/WSLH049/2009 (H1N1). All RNAs were extracted at the University of Washington and the RNA-seq libraries were then produced at the Broad Institute by the strand-specific dUTP method from oligo dT polyA-isolated RNA[26]. The libraries were sequenced by Hi-Seq Illumina machines, producing 101 bp reads (3-6 Gb of sequence/tissue). All 24 RNA-seq data sets were assembled via the genome-independent RNA-seq assembler Trinity[27].

Ensembl gene annotation. The genome assembly was annotated by the Ensembl gene annotation system[6] (Ensembl release 70, August 2012). Protein-coding gene models were annotated by combining alignments of Uniprot[7] mammal and other vertebrate protein sequences and RNA-seq data. RNA-seq models were generated from a survey of adult ferret and embryonic ferret tissues, including tissues from an influenza-infected ferret. This pipeline produced 23,963 transcripts arising from 19,910 protein coding genes and 3,614 short noncoding genes. The ferret gene annotation is available on the Ensembl website (http://www.ensembl.org/Mustela_putorius_furo/), including orthologs, gene trees and whole-genome alignments against human, mouse and other mammals. Also included are the tissue-specific mRNA-seq transcript models, indexed BAM files and complete set of splice junctions identified by our pipeline. Further information about the annotation process can be found at http://www.ensembl.org/Mustela_putorius_furo/Info/Annotation, which includes a summary as well as a PDF giving a detailed description of the ferret genebuild.

Comparative genomics. Orthology inference: orthology among the ferret genome and 33 other genomes was inferred using the OMA pipeline[28]. This yielded pairwise orthologs between all species and orthologous groups. The latter were used for 3-way human/mouse/ferret comparisons and species tree inference. The number of genes conserved across mammals and carnivores was computed from hierarchical orthologous groups identified using GETHOGs[29].

Species tree. The 789 orthologous groups covering at least 31 of the 34 species considered were aligned individually and then concatenated. Missing data were represented as "X" characters. Alignment was performed using MAFFT's local-alignment based L-INS-i algorithm. Phylogenetic inference was performed using PhyML, under the JTT, WAG and LG models, and also as a partitioned analysis in RAxML[30,31]. Support values were calculated using bootstrapping (RAxML and PhyML) and approximate Bayesian support (PhyML aBayes).

Scatterplot analysis. To contrast the divergence between human-ferret genes and their human-mouse counterparts, we extracted triplets of orthologs between the three species from all OMA groups computed above. Divergence was computed using two measures: (i) point accepted mutation (PAM) unit estimated by pairwise maximum likelihood distance estimation using Gonnet matrices[32]; and (ii) nucleotide divergence calculated in PhyML from triplet alignments using the general time-reversible model.

Gene ontology (GO) annotations and enrichment analyses. GO terms were assigned to OMA groups by propagating experimental GO annotations (GO evidence codes EXP, IDA, IPI, IMP, IGI, IEP) of any group member to the rest of the group. This procedure assigned 13,509 GO terms, most of them from the Biological Process ontology, to 9,117 OMA groups, resulting in 541,220 GO annotations.

To perform the GSEA analysis[33], we created a list by ordering our data points in **Figure 1b** according to the angle to the *x* axis and determined, for each GO term separately, whether data points are randomly distributed throughout the list or are primarily found at the top or at the bottom. To perform a two-tailed Fisher's exact test, we partitioned the data points in the scatterplot (**Fig. 1b**) into 4 quantiles according its angle to the *x* axis, thereby accounting for the relative evolutionary distance between ferret and mouse. For each GO term, we contrasted each quartile with the other three fourths of the data. We used a two-tailed Fisher's exact test as implemented in R. In both statistical analyses we adjusted the *P* values for multiple testing using the Benjamini & Hochberg method as implemented in R. To organize the enriched GO terms, the terms that passed the enrichment criteria were processed with REVIGO to remove redundant GO terms and cluster semantically similar terms. Very generic GO terms (for example, "macromolecular complex", "organelle part") were excluded, as were singleton terms that were not aggregated into clusters.

Comparison of gene sets specific to models of human health and disease. Increased conservation in gene subsets (**Fig. 1b**) were determined by Wilcoxon signed-rank test in R (version 2.14.1) by the wilcox.test function. Interactions with CFTR (gene subset CFTR interactome in **Fig. 1b**) were obtained from a previously published CFTR IP–mass spec data set[34]. Other five CF-related gene subsets in **Figure 1b** were obtained from http://www.genecards.org[35] using relevancy score cutoffs at the point of greatest Euclidean distance. Two nervous system related GO terms in **Figure 1b** were from the enrichment analysis of proteins with mouse sequences closer to human when tested with Fisher's exact test and Gene Set Enrichment Analysis (false-discovery rate < 0.05) (**Supplementary Tables 5** and **6**).

Expanded custom ferret genome annotation for differential expression analysis. We assembled ferret transcript contigs *de novo* using Trinity[27] from ferret RNA-seq data with default parameters. The RNA-seq data used included mRNA-seq data of the panel of 24 tissue samples for 21 different tissues or tissue mixes (each sample was assembled separately), as well as both mRNA-seq and Total RNA-seq data from the influenza study (all 21 virus infected or control lung samples (each condition and protocol was assembled separately). All assembled transcript contigs were aligned to the ferret reference genome (MusPutFur1.0) using GMAP[36] with default parameters. For those uniquely aligned ferret transcript contigs, their alignments across all mRNA-seq data (transcript contigs from lung Total RNA-seq data were not included here) were merged using cuffmerge (Cufflinks version 2.0.2)[37] to remove redundant alignments and to predict novel genes and transcripts. Predicted transcripts were checked against Ensembl annotation (version 69) to identify: 1) putative intergenic transcripts—those did not overlap with any Ensembl annotated transcripts directly (with class code 'u') and indirectly through any predicted transcripts overlapping Ensembl annotated transcripts, and 2) putative novel isoforms of Ensembl annotated transcripts (with class code 'j', at least one splice junction is shared with a reference transcript). We removed all single exon transcripts or any transcript with the alignment to reference genome shorter than 200 nt to minimize unspliced precursor fragments. For putative

© 2014 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.3079

intergenic transcripts, we also removed a small number of predicted transcripts located within introns of other predicted transcripts. For putative novel isoforms, we removed those predicted transcripts that spanned two or more Ensembl annotated genes to minimize mis-assembled transcripts. Similarly, we predicted intergenic transcripts from lung Total RNA-seq data, which were then filtered against both Ensembl annotation and the newly predicted transcripts from mRNA-seq data. We did not require intergenic transcripts from Total RNA-seq to be spliced, but the length of alignment to reference genome had to be at least 120 nt, which was intended to capture ncRNAs that would be longer than small RNAs like miRNAs. For Total RNA-seq data we did not predict novel isoforms to avoid unspliced transcripts. After filtering we combined all predicted transcripts with Ensembl annotation into one annotation, which was used for all downstream gene quantification and differential expression analysis. This expanded annotation was used for Agilent ferret microarray design.

To investigate if Total RNA-seq captured non-polyadenylated transcripts, we performed both Total RNA-seq and mRNA-seq analysis of 21 ferret lung samples. We reasoned that, for the same gene in the same sample, if Total RNA-seq analysis collected many more short reads than mRNA-seq analysis, that gene likely transcribed non-polyadenylated transcripts, since by polyT priming mRNA-seq analysis selected against non-polyadenylated transcripts. To facilitate the comparison, the raw gene read counts were first preprocessed as follows: i) any gene with less than 50 raw read counts in all 42 RNA-seq measurements were removed to ensure genes to be compared were robustly detected at least once in the samples used here, and ii) all gene raw read counts were scaled by the total read counts of remaining genes in each RNA-seq analysis for each sample. Next, for each gene we counted the number of samples (out of 21 samples in total) in which the scaled read count from Total RNA-seq analysis was much larger (1.5-fold or more) than that from the corresponding mRNA-seq analysis.

Comparative analysis of tissue expression patterns. The panel of ferret tissue RNA-seq data generated for genome annotation was used to quantify ferret gene expressions in each tissue and was processed the same way as described in the influenza study section below; depending on the tissue type, the data was from a single individual or from 2–4 individuals. For human data set, alignment files of RNA-Seq read alignment of 24 human tissues and cell types were downloaded from Human lincRNA Catalog website (http://www.broadinstitute.org/genome_bio/human_lincrnas/?q=home).

These data were derived from specimens collected from 9 individuals, with 1–2 contributing individuals per tissue. Similarly we quantified the expression of human genes (Ensembl 69) in each tissue using HT-seq (http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html). We selected the set of 7 tissues (brain, testis, skeletal muscle, heart, lung, liver and kidney) that were common between two data sets for further comparative analysis. We limited the comparison to 15,597 ferret-human gene pairs that had 1:1 ortholog relationships as defined by Ensembl 69. The normalized read count in counts per million (cpm) for each tissue was obtained using edgeR[38]; in instances when there was more than one tissue donor, the resulting data are averages. To focus on genes that tended to be robustly detected in both ferret and human data sets, we applied two *ad hoc* filters on genes based on the observed abundances. First, before read count normalization we instituted a threshold for genes with a raw read count of at least 100 in at least one of the ferret tissues and at least one of the human tissues (not necessarily the same tissue between ferret and human). Second, to account for the differences in sequencing depth, in the normalized read count we further required each gene to have at least 5 cpm in at least one of the ferret tissues and at least one of the human tissues (again, not necessarily the same tissue between ferret and human). The final working set was 12,636 genes. For each gene within each species, we calculated the relative abundance in each tissue as the ratio between its cpm in the tissue and the sum of cpms across all tissues. To evaluate if orthologous genes tend to exhibit concordant tissue-specific expression across ferret and human, we determined the number of genes with highest relative abundance across all tissues of the same species, as well as the intersection of these tissue-specific sets between human and ferret, and assessed the significance of the intersections using a chi-squared test.

To identify groups of genes with similar expression patterns across tissues and species, we clustered genes using k-means partitioning. We iteratively assessed the number of centers (k) to be used for clustering as following: for a given k, we calculated the difference between the converged, total within-cluster sum of squares vs. that from 500 random data sets generated by random permutation of the actual data matrix; for the series of k tested, $k = 7$ had the maximum difference for the final gene clustering. For the results of the $k = 7$ clustering, we used a Mann-Whitney test to evaluate if the overall expression of a cluster of genes was significantly higher in one tissue relative to the rest tissues of the same species, based on the relative abundances in the normalized count matrix.

Cells and influenza viruses. The 2009 pandemic influenza A/California/04/2009 (H1N1) virus, referred to as CA04, and the 1918 pandemic influenza A/Brevig Mission/1/1918 (H1N1) virus, referred to as 1918, were generated by reverse genetics, as described[39–41]. Madin-Darby canine kidney (MDCK) cells for virus titer measurements were from ATCC, and were grown in Eagle's minimum essential medium (MEM) with 5% newborn bovine calf serum (HyClone, Thermo Fisher Scientific) and penicillin/streptomycin. Cell stocks are periodically restarted from early passage aliquots and routinely monitored for mycoplasma contamination.

Ferret infections. Twenty-one 4- to 8-month-old female ferrets were obtained from Triple F Farms Inc. (Sayre, PA, USA), confirmed serologically negative by hemagglutination inhibition assay for currently circulating influenza viruses and randomly assigned to experimental groups. Individual animals were intramuscularly anesthetized with ketamine and xylazine (5 mg and 0.5 mg per kg of body weight, respectively), followed by intranasal inoculation with 500 µl of phosphate-buffered saline (PBS; $n = 3$) alone, PBS containing 1×10^6 plaque forming units (PFU) of the CA04 virus ($n = 9$), or PBS containing 1×10^6 PFU of the 1918 virus ($n = 9$). On day 1 post-infection (p.i.), 3 animals from each infection group were euthanized, and tracheal and lung tissues were harvested for virological, immunohistochemical staining for influenza A virus antigen and RNA sequencing analysis. Tissues were similarly harvested from 3 additional CA04- or 1918-infected ferrets on days 3 and 8 p.i. Tracheal tissues harvested for each individual analysis were derived from the same general region in each ferret, and lung tissues for all analyses were derived from the same lung lobe. We previously examined pathologic lesions in 1918 virus-infected ferret lung, observing macroscopic pathologic changes by day 3 p.i. that included severe lesions and hemorrhage[40]. Since the primary purpose of the present study was to measure gene expression changes in regions of the lung where macroscopic lesions are known to develop, we carefully selected the lung lobe and lung region based on our previous study and then collected samples from the same region for all animals in the study to be consistent. Sample sizes of 3 animals per condition were in keeping with prior reports for exploratory animal models to characterize influenza infection when models require serial sacrifice. While the sample sizes were not the result of a power analysis for a prespecified effect size, the evaluation of the gene expression differences between conditions is performed with statistical stringencies suitable for exploratory assessment, hypothesis generation and reproducibility by alternate techniques such as qPCR. All procedures with ferrets were approved by the University of Wisconsin School of Veterinary Medicine Animal Care and Use Committee, and were performed in an enhanced biosafety level 3 Agriculture (BSL3-Ag) containment suite. All samples derived from influenza virus-infected ferret tissues and containing infectious virus were manipulated in BSL3-Ag containment. Ensuing analysis of samples from the influenza model was performed without blinding, with the exception of histopathology scoring.

Virus quantification and immunohistochemical staining. Ferret tracheal and lung tissues, frozen at −80 °C at the time of excision, were thawed and homogenized in PBS containing penicillin/streptomycin. Cleared supernatants were titrated on MDCK cells using standard methods. For virus antigen immunohistochemical (IHC) analysis, tissues were preserved by immersion in 10% phosphate-buffered formalin (Sigma-Aldrich). Preserved tissues were paraffin embedded and several 5-µm-thick sections were cut for ferret tracheal and lung tissues. Sections were stained with standard hematoxylin and eosin

© 2014 Nature America, Inc. All rights reserved.

and then processed for IHC staining with an in-house rabbit anti-influenza virus polyclonal antibody (R309) raised against influenza A/WSN/1933 (H1N1) virus[40].

Quantitative reverse transcription (RT-PCR). Quantitative RT-PCR was performed to assess viral mRNA transcripts from infected ferret tracheal and lung samples used for sequencing. Total RNAs were treated with DNase using DNA-free DNase Treatment and Removal Reagents (Ambion, Inc., Austin, TX). cDNAs from total RNAs were generated using the QuantiTect reverse transcription kit (Qiagen Inc.). A custom-designed TaqMan gene expression assay for influenza Matrix (M) sequence (MPCONS2010), with primers that have complete homology with both 1918 and CA04 M sequences, was ordered from Applied Biosystems, Inc. Taqman experiments were performed on the ABI 7500 Real-Time PCR System platform and each sample was run in quadruplicate. Ribosomal RNA (18S) was used as endogenous control to normalize quantification of each target within tissues using Applied Biosystems Sequence Detections Software version 1.3. The relative amount of viral mRNA ($\log_{10}$) is presented in the final results.

RNA extraction and library preparation. Tissues used for RNA sequencing were excised and immediately immersed in RNALater (Ambion) for 24 h at 4 °C, subsequently frozen at −80 °C, and later thawed and homogenized in TRIzol (Life Technologies); RNA was isolated using QIAGEN miRNeasy protocols.

Total RNA from each sample was divided into two pools for whole transcriptome and mRNA library construction. RNA for whole transcriptome analysis was depleted of rRNA using the Epicentre RiboZero Gold protocol (Epicentre) designed for human, mouse and rat samples, but effective in reducing rRNA amounts for ferret total RNA samples. The presence of 18S and 28S rRNA peaks was checked using the Agilent 2100 Bioanalyzer instrument (Agilent). The rRNA depleted RNA was then used to make strand-specific whole transcriptome libraries[26]. Strand-specific mRNA libraries were constructed using the Illumina TruSeq RNA Preparation Kit (Illumina) according to the manufacturer's guide. Both libraries were quality controlled and quantitated using the Agilent 2100 Bioanalyzer instrument and qPCR (Kapa Biosystems).

Transcriptome sequencing, read mapping and differential expression analysis. Constructed libraries were sequenced using Illumina platform with stranded paired end reads, 2 × 100 nt for all mRNA-seq data, 2 × 100 nt for lung total RNA-seq data and 2 × 50 nt for trachea total RNA-seq data. Lung data sets were assembled via the genome-independent RNA-seq assembler Trinity, with each set of three biological replicates assembled into a single transcriptome assembly. However, the quantitative analysis used the ferret genome as a reference, mapping short reads to the ferret genome using RNA-seq aligner STAR[42] with default parameters. The index used for STAR included splicing junctions from the expanded custom annotation constructed below, genome sequences of influenza viruses used in this study, and human and mouse ribosomal RNA sequences. Gene level quantification was based on the Ensembl gene annotations combined with the expanded list of transcribed genomic regions that were identified using the 63 RNA-seq data sets generated from the influenza model (cf. Supplemental Materials). Quantification was performed using HT-seq. The differential expression analysis was performed using edgeR[38]. Clustering and other statistical analyses were performed using R (http://www.r-project.org/).

Influenza model - transcriptomic analysis details. The expression data from both tissues were combined and processed together, using the generalized linear model approach provided by edgeR. Stages in the analysis are outlined in **Supplementary Figure 18**. Genes were differentially regulated vs. mock in any of the infection conditions were partitioned into disjointed clusters reflecting their tissue or virus specificity. Both Ensembl annotated genes and the expanded list of transcribed genomic regions were quantified using mapped RNA-seq reads. Differential analysis using count data was called significant for adjusted P values ≤ 0.01.

Functional analysis of differential gene expression data. Functional analysis was performed using Ingenuity Pathway Analysis (IPA, Ingenuity Systems, Inc). The software tool analyzes the experimental data set in the context of known biological functions and pathways within the Ingenuity Pathways Knowledge Base, a curated repository of biological interactions and functional annotations. Analysis of the data sets used human annotations, based on the Ensembl listing of human-ferret orthologs. The P values associated with functions or pathways were calculated using the right-tailed Fisher's exact test.

Ferret microarray design and performance assessment. We designed two versions of oligonucleotide microarray using Agilent eArray Web portal (https://earray.chem.agilent.com/earray) to profile both Ensembl annotated transcripts and intergenic transcripts derived from ferret RNA-seq data as described above. In both cases, the longest isoform of each locus was selected for probe design with the 'Design with 3′ Bias' checked, and probe length was set to 60 nt. For first version of microarray (design ID: 048471) 36,950 probes were selected to target Ensembl annotated genes and intergenic transcripts uncovered from mRNA-seq data, and is intended to work with conventional experimental protocols using poly(A) priming for cDNA synthesis. The second version (design ID: 048472) has 64,238 probes selected to target Ensembl annotated genes as well as intergenic transcripts uncovered from both mRNA-seq and Total RNA-seq data. It is intended to work with experimental protocols using random priming for cDNA synthesis to capture both poly(A) and non-poly(A) transcripts. The performance of designed microarray was evaluated by comparing the microarray measurements vs. RNA-seq measurements on the same influenza-infected ferret samples.

Microarray measurements and data analysis. Cy3-labeled cRNA probes were prepared using standard approaches as provided by Agilent Technologies. For the ferret array design ID 048471, Agilent kit 5190-2305 yields probes derived from poly-adenylated RNAs in the starting sample. For design ID 048472, labeled probes were generated with the whole transcriptome labeling kit (part number 5190-2943). Hybridizations were performed as per manufacturer instructions and the slides read on an Agilent Technologies model G2565C high-resolution scanner with extended dynamic range. Image files were processed with Agilent Feature Extraction Software, yielding background-corrected fluorescence intensities with flags for those features deemed not significantly different from background. Statistical analyses to determine differentially regulated genes were performed with the Bioconductor package limma, as described above.

Transcriptional profiling of CF and non-CF ferrets. Homozygous CFTR knockout ferrets were generated and reared as described[16]. Non-CF ferrets were either homozygous or heterozygous for functional CFTR genes. Lung samples were collected after sacrifice at birth or at 15 days post partum and flash frozen. For RNA isolation, lung specimens were ground in liquid nitrogen and then immediately suspended in TRIzol (Life Technologies) and RNA isolated with QIAGEN RNeasy protocols. Animals were assigned to groups solely on the basis of age and genotype (i.e., CF or non-CF). At age 15 days, the comparisons used three CF ferrets (1 F, 2M) and 5 non-CF animals (3 F, 2 M); comparisons of newborn animals used four animals of each phenotype (sexes unknown). Experiments were performed under protocols approved by the Institutional Animal Care and Use Committee at the University of Iowa; statistical considerations for sample sizes were described earlier in the context of the influenza infection model. Array measurements for polyadenylated transcripts (design 048471) were performed as described above and statistical comparisons of CF vs. non-CF animals were done as t-tests as implemented in limma; procedures were performed without blinding. Differences between CF vs. non-CF newborn animal were quite limited and were determined without the use of a multiple test correction. The threshold for differential expression was absolute fold-change ≥ 1.4 and un-adjusted P value ≤ 0.1. Expression differences for CF vs. non-CF 15-day-old animals did use a multiple test correction (Benjamini-Hochberg false discovery rate). The threshold for differential expression was absolute fold-change ≥ 1.5 and false discovery rate ≤ 0.05. These analyses were limited to those array probes that interrogated protein-coding genes within the Ensembl annotation for the ferret genome, and functional interpretation used the corresponding human gene symbols based on the Ensembl mapping of ferret-to-human orthologs. See each

© 2014 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.3079

supplemental tables and figures for specifics of filtering criteria for the results of the statistical tests.

Comparative analysis of transcriptional changes in human cystic fibrosis bronchial epithelium. We downloaded the gene expression data on human cystic fibrosis (CF) CF and non-CF bronchial epithelium samples from ArrayExpress (E-MTAB-360)[19]. The Illumina HumanRef-8 Expression BeadChips summary expression data was statistically analyzed using limma with default settings. We filtered out two CF samples ("127 CF BBr", "129 CF BBr") and two non-CF samples ("112 control BBr", "113 control BBr") as potential outliers, due to their relative large deviations from other replicates upon inspecting multidimensional scaling (MDS) plots and the overall expression changes. The final data set included 17 non-CF samples and 10 CF samples. The statistical analysis of differential expression was done at the probe level, and we applied the same multiple test correction (Benjamini-Hochberg false discovery rate) as we did for day 15 ferret CF vs. non-CF comparison. The functional enrichment analysis of differentially expressed genes was performed using Ingenuity Pathway Analysis (IPA), similarly for differentially expressed ferret genes.

To highlight genes concordantly or discordantly differentially expressed in lung samples from day 15 CF ferrets and human CF bronchial epithelium samples, we first identified biological functions enriched in CF/non-CF differentially expressed genes, separately for day 15 ferret CF/non-CF comparison and human CF/non-CF comparison using IPA analysis. From one given function (or a subset of related functions) enriched in both comparisons, we gathered genes differentially expressed in either comparison to identify concordantly or discordantly differentially expressed genes between two comparisons. We applied the following steps to identify genes with concordant expression changes between two comparisons. First, the gene was significantly (fold change ≥ 1.5 and adjusted P value ≤ 0.05) differentially expressed in both comparisons. Second, the gene had the same direction of expression changes in two comparisons. Third, we ranked these genes selected from steps 2 and 3 by their absolute $\log_2$ fold changes within each comparison, i.e., the gene with the largest fold change had a rank of 1 and the gene with the smallest fold change had a rank equal to the total number of these selected genes. This way each gene was assigned with two corresponding ranks, one from the day 15 ferret CF/non-CF comparison and one from the human CF/non-CF comparison. Fourth, we ranked these selected genes by the sum of 1) the difference between their two ranks and 2) the maximum of two ranks. The top genes from this process tended to have both large expression changes in both comparisons and their expression changes tended to be close in magnitude. Similarly we applied the following steps to identify genes with discordant expression changes between two comparisons. First, the gene was significantly (fold change ≥ 1.5 and adjusted P value ≤ 0.05) differentially expressed in both comparisons. Second, the gene had the different direction of expression changes in two comparisons. Third, we ranked these genes selected from steps 2 and 3 by the absolute value of the difference between their two $\log_2$ fold changes, one from the day 15 ferret CF/non-CF comparison and one from the human CF/non-CF comparison. The top gene had the largest difference in fold changes between two comparisons, and with opposite expression changes.

25. Cavagna, P., Menotti, A. & Stanyon, R. Genomic homology of the domestic ferret with cats and humans. *Mamm. Genome* **11**, 866–870 (2000).
26. Levin, J.Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).
27. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
28. Altenhoff, A.M., Schneider, A., Gonnet, G.H. & Dessimoz, C. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.* **39**, D289–D294 (2011).
29. Altenhoff, A.M., Gil, M., Gonnet, G.H. & Dessimoz, C. Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS ONE* **8**, e53786 (2013).
30. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thoUSA.nds of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
31. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
32. Gonnet, G.H., Cohen, M.A. & Benner, S.A. Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443–1445 (1992).
33. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
34. Wang, X. *et al.* Hsp90 cochaperone Aha1 downregulation rescues misfolding of CFTR in cystic fibrosis. *Cell* **127**, 803–815 (2006).
35. Rebhan, M., Chalifa-Caspi, V., Prilusky, J. & Lancet, D. GeneCards: integrating information about genes, proteins and diseases. *Trends Genet.* **13**, 163 (1997).
36. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
37. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
38. McCarthy, D.J., Chen, Y. & Smyth, G.K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
39. Ozawa, M. *et al.* Impact of amino acid mutations in PB2, PB1–F2, and NS1 on the replication and pathogenicity of pandemic (H1N1) 2009 influenza viruses. *J. Virol.* **85**, 4596–4601 (2011).
40. Watanabe, T. *et al.* Viral RNA polymerase complex promotes optimal growth of 1918 virus in the lower respiratory tract of ferrets. *Proc. Natl. Acad. Sci. USA* **106**, 588–592 (2009).
41. Neumann, G. *et al.* Generation of influenza A viruses entirely from cloned cDNAs. *Proc. Natl. Acad. Sci. USA* **96**, 9345–9350 (1999).
42. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

© 2014 Nature America, Inc. All rights reserved.