PRAGMA 28 | Nara-Japan | 9 April 2015

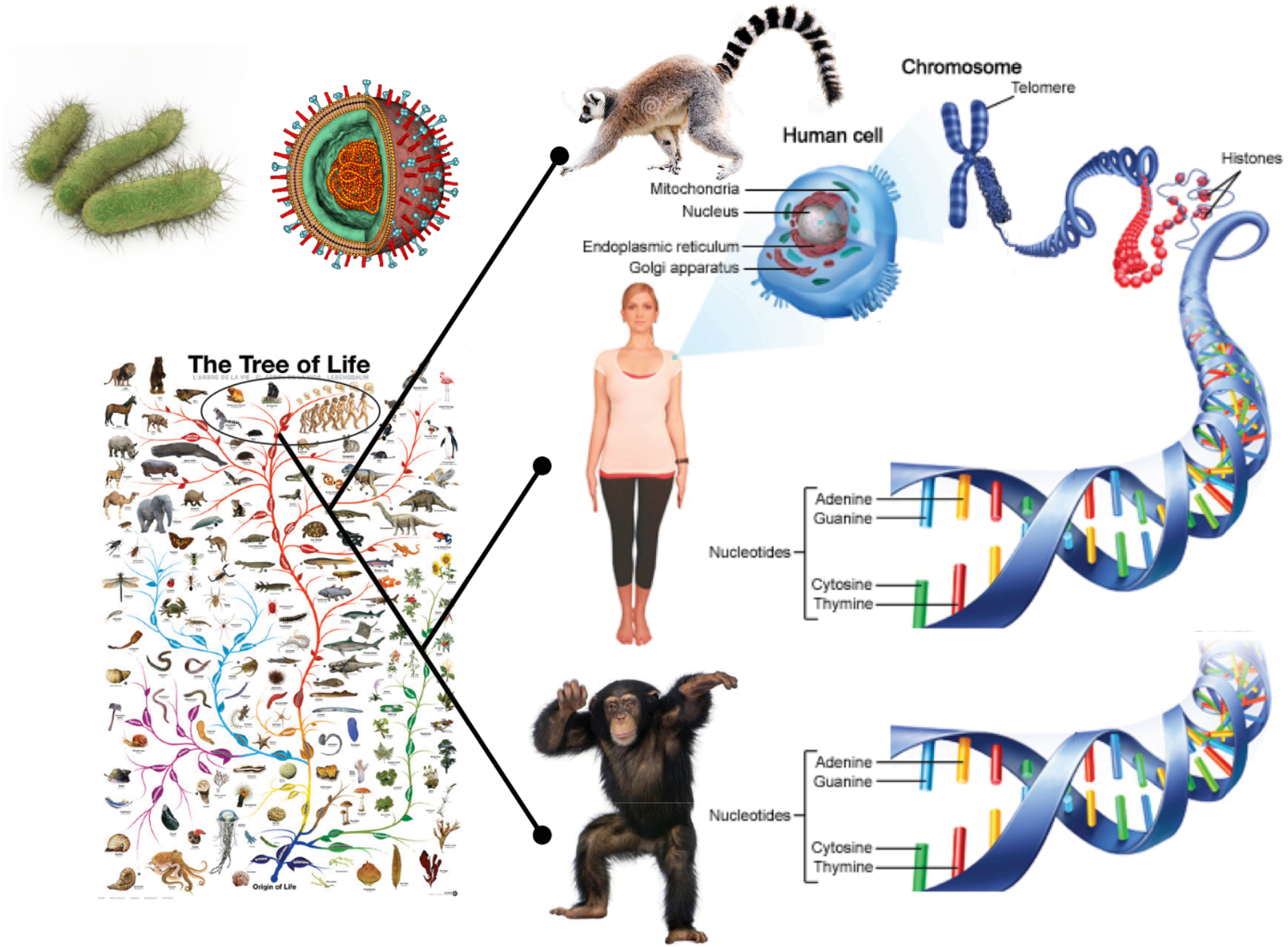
Metagenomic and Phylogenomic Applications in studying Biological & Medical Sciences

Tommy T.-Y. Lam

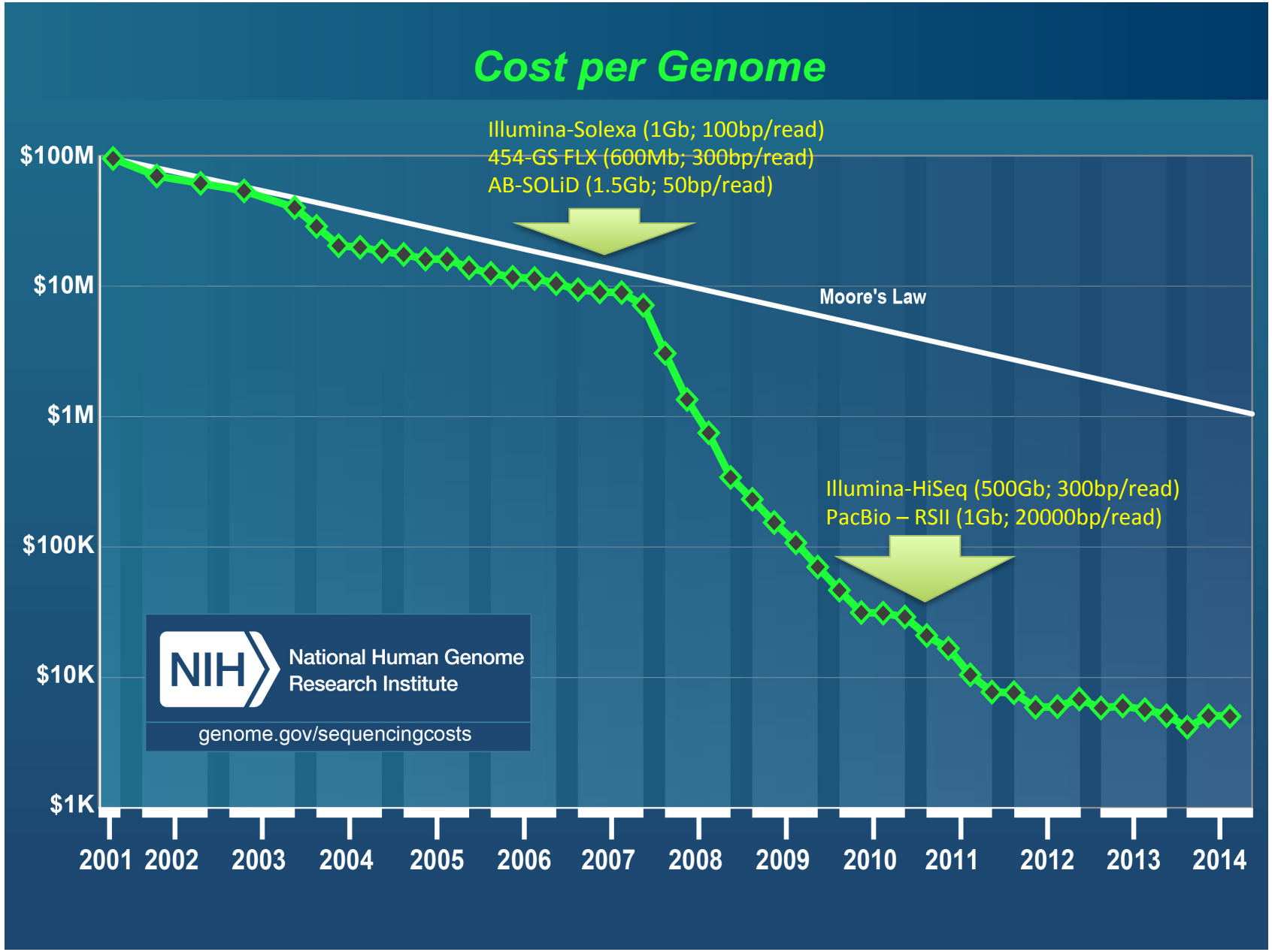
Research assistant professor

SCHOOL OF PUBLIC HEALTH
THE UNIVERSITY OF HONG KONG
香 港 大 學 公 共 衞 生 學 院

Genomic information used in different biological and medical sciences



Explosion of genomic data



Metagenomic studies
of microbiota

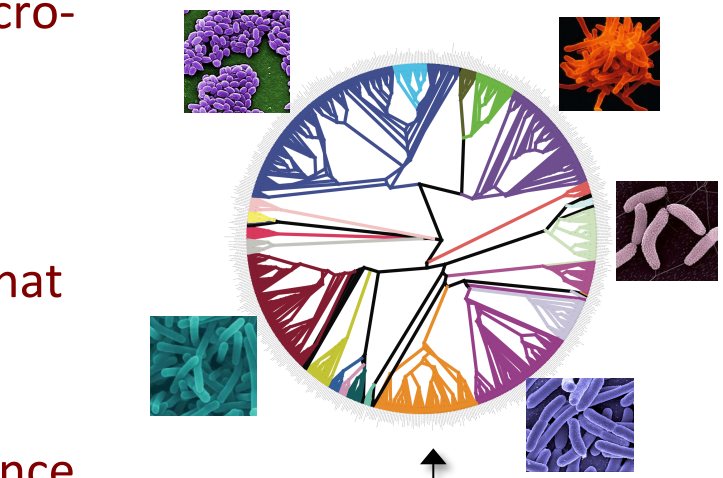
Phylogenomic studies
of organismal evolution

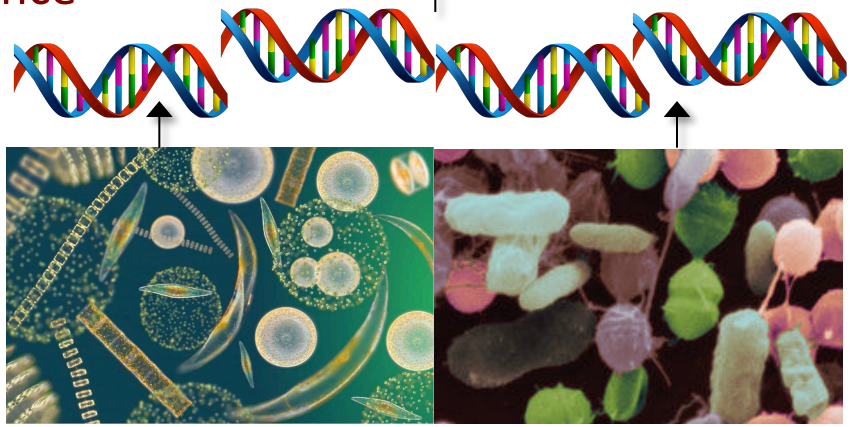
Phylogenomic studies
of pathogen transmission and emergence

Web-based platform ‘Galaxy’

Metagenomics

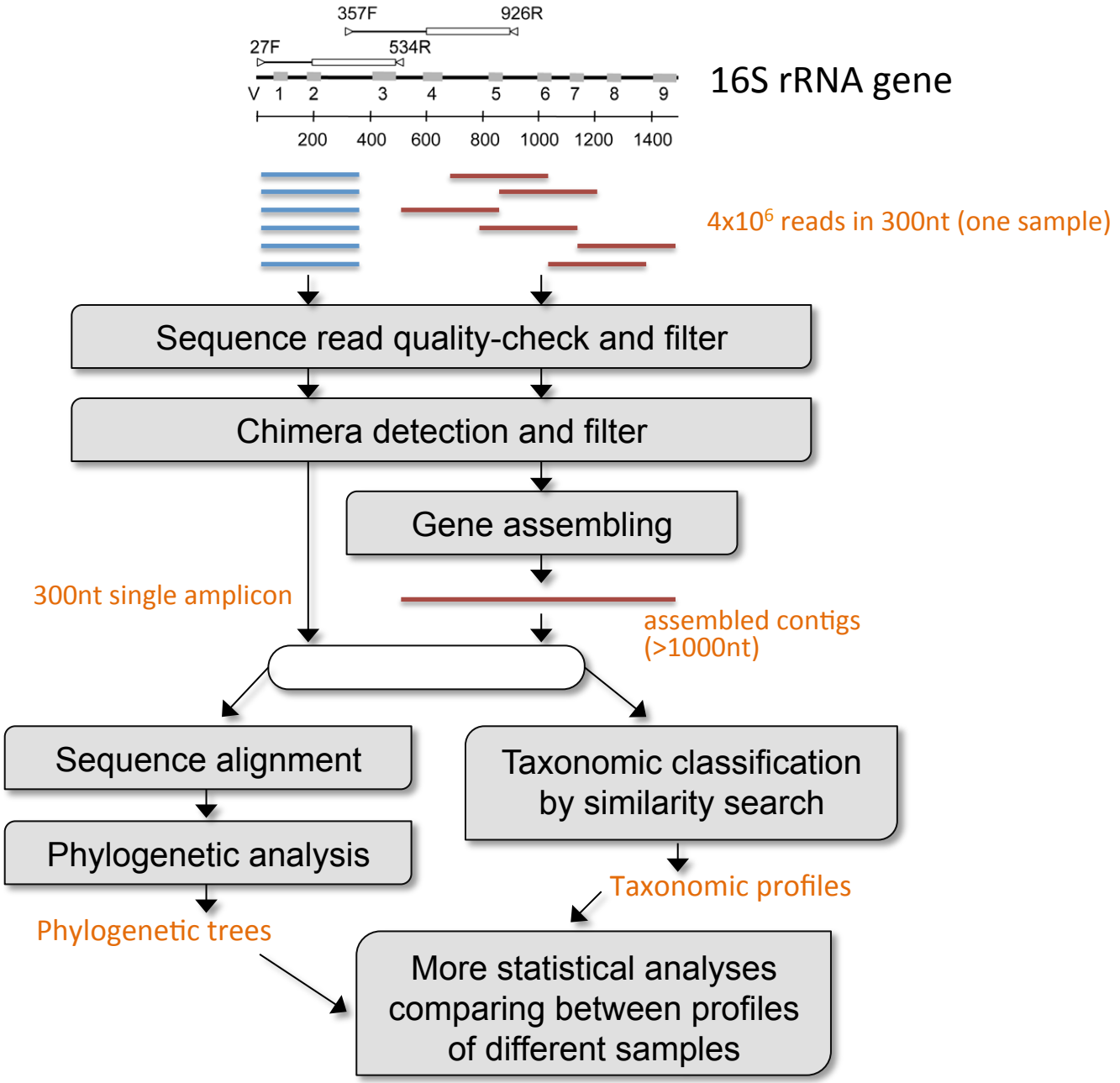
- Diversity in the micro-organismal community ("Microbiota")
- Factors/Diseases that account (are accounted by) the community difference
- Understanding interaction in the ecosystem



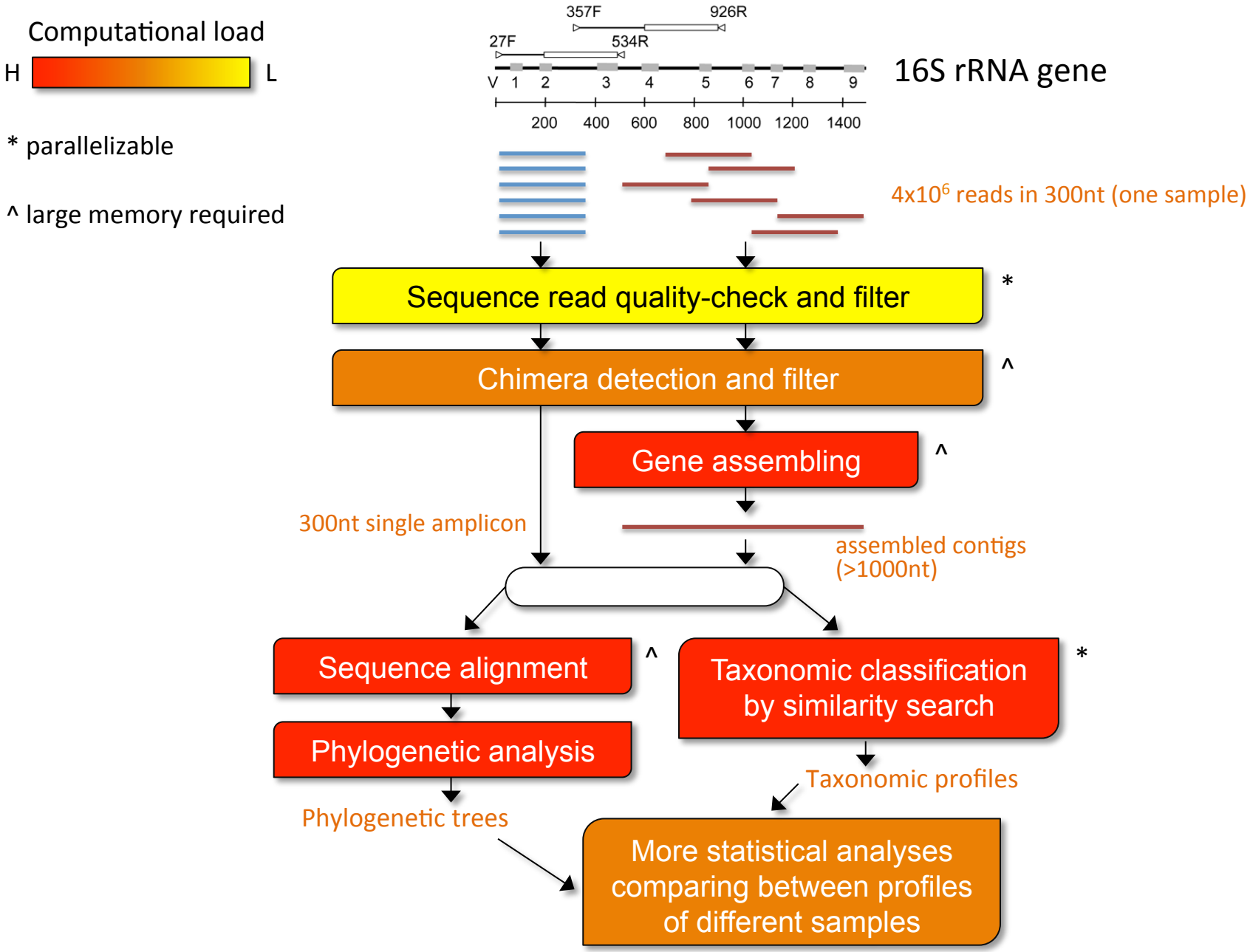


- Discovery of novel important micro-organisms
  - Disease causing agents
  - Health maintaining agents
  - Biochemical reactor with industrial values
  - Key players in ecosystem

Overview of Metagenomic Data Analysis



Overview of Metagenomic Data Analysis



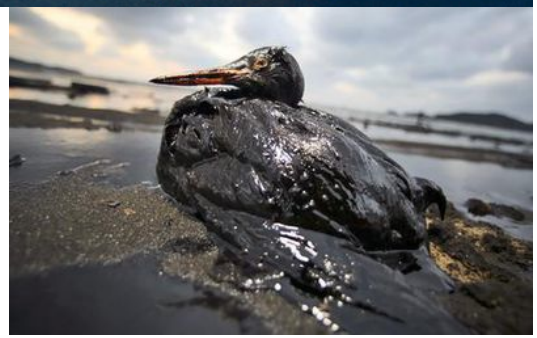
frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 15 March 2013
doi: 10.3389/fmicb.2013.00050

Metagenomic analysis and metabolite profiling of deep-sea sediments from the Gulf of Mexico following the Deepwater Horizon oil spill

Nikole E. Kimes[1†], Amy V. Callaghan[2], Deniz F. Aktas[2,3], Whitney L. Smith[2,3], Jan Sunner[2,3], Bernard T. Golding[4], Marta Drozdowska[4], Terry C. Hazen[5,6,7,8], Joseph M. Suflita[2,3] and Pamela J. Morris[1]*

[1] *Baruch Marine Field Laboratory, Belle W. Baruch Institute for Marine and Coastal Sciences, University of South Carolina, Georgetown, SC, USA*
[2] *Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, USA*
[3] *Institute for Energy and the Environment, University of Oklahoma, Norman, OK, USA*
[4] *School of Chemistry, Newcastle University, Newcastle upon Tyne, UK*
[5] *Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, TN, USA*
[6] *Department of Microbiology, University of Tennessee, Knoxville, TN, USA*
[7] *Department of Earth and Planetary Sciences, University of Tennessee, Knoxville, TN, USA*
[8] *Ecology Department, Lawrence Berkeley National Laboratory, Berkeley, CA, USA*



Marine subsurface environments such as deep-sea sediments, house abundant and diverse microbial communities that are believed to influence large-scale geochemical processes. These processes include the biotransformation and mineralization of numerous petroleum constituents. Thus, microbial communities in the Gulf of Mexico are thought to be responsible for the intrinsic bioremediation of crude oil released by the Deepwater Horizon (DWH) oil spill. While hydrocarbon contamination is known to enrich for aerobic, oil-degrading bacteria in deep-seawater habitats, relatively little is known about the response of communities in deep-sea sediments, where low oxygen levels may hinder such a response. Here, we examined the hypothesis that increased hydrocarbon exposure results in an altered sediment microbial community structure that reflects the prospects for oil biodegradation under the prevailing conditions. We explore this hypothesis using metagenomic analysis and metabolite profiling of deep-sea sediment samples following the DWH oil spill. The presence of aerobic microbial communities and associated functional genes was consistent among all samples, whereas, a greater number of Deltaproteobacteria and anaerobic functional genes were found in sediments closest to the DWH blowout site. Metabolite profiling also revealed a greater number of putative metabolites in sediments surrounding the blowout zone relative to a background site located 127 km away. The mass spectral analysis of the putative metabolites revealed that alkylsuccinates remained below detection levels, but a homologous series of benzylsuccinates (with carbon chain lengths from 5 to 10) could be detected. Our findings suggest that increased exposure to hydrocarbons enriches for Deltaproteobacteria, which are known to be capable of anaerobic hydrocarbon metabolism. We also provide evidence for an active microbial community metabolizing aromatic hydrocarbons in deep-sea sediments of the Gulf of Mexico.

Kimes NE, et al. ***Frontiers in Microbiology***. (2013)

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 15 March 2013
doi: 10.3389/fmicb.2013.00050

Metagenomic analysis and metabolite profiling of deep-sea sediments from the Gulf of Mexico following the Deepwater Horizon oil spill

Nikole E. Kimes[1†], Amy V. Callaghan[2], Deniz F. Aktas[2,3], Whitney L. Smith[2,3], Jan Sunner[2,3], Bernard T. Golding[4], Marta Drozdowska[4], Terry C. Hazen[5,6,7,8], Joseph M. Suflita[2,3] and Pamela J. Morris[1]*

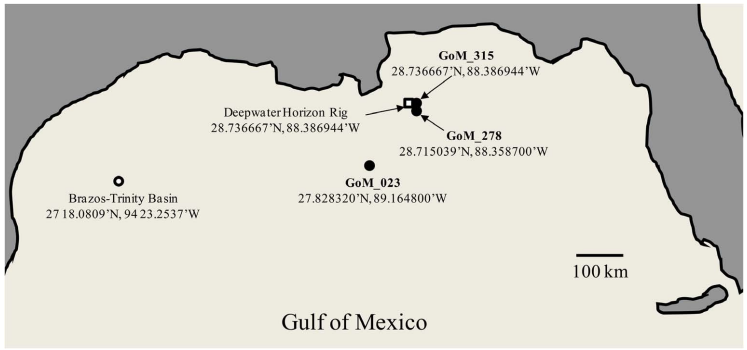


FIGURE 1 | Map of the Gulf of Mexico sampling sites. Open square – DWH rig; filled circle – sampling sites from the current study; open circle – BT Basin sampling site from Biddle et al. (2011). The Peru Margin (PM) sampling sites used for comparison are described in Biddle et al. (2008).

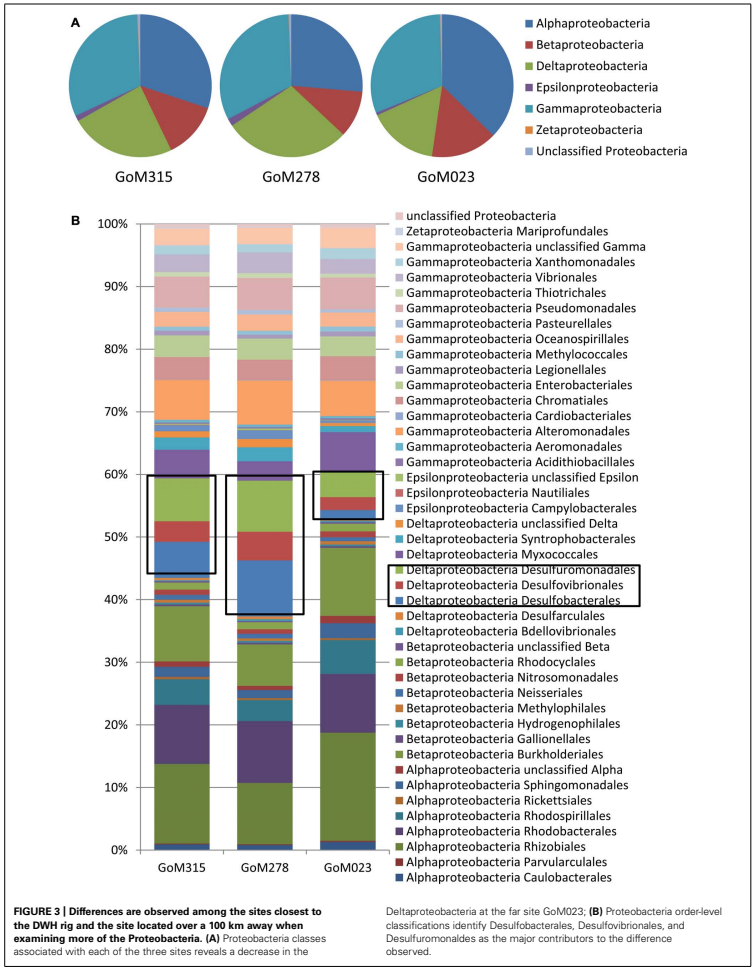


FIGURE 3 | Differences are observed among the sites closest to the DWH rig and the site located over a 100 km away when examining more of the Proteobacteria. (A) Proteobacteria classes associated with each of the three sites reveals a decrease in the Deltaproteobacteria at the far site GoM023; **(B)** Proteobacteria order-level classifications identify Desulfobacterales, Desulfovibrionales, and Desulfuromonaldes as the major contributors to the difference observed.

Oil spill site

- Deltaproteobacteria ↑
- anaerobic functional genes ↑

Deltaproteobacteria are known to be capable of anaerobic hydrocarbon metabolism

Turnbaugh PJ, et al. ***Nature***. (2009) 457:480-4

nature Vol 457|22 January 2009|doi:10.1038/nature07540

LETTERS

A core gut microbiome in obese and lean twins

Peter J. Turnbaugh[1], Micah Hamady[3], Tanya Yatsunenko[1], Brandi L. Cantarel[5], Alexis Duncan[2], Ruth E. Ley[1], Mitchell L. Sogin[6], William J. Jones[7], Bruce A. Roe[8], Jason P. Affourtit[9], Michael Egholm[9], Bernard Henrissat[5], Andrew C. Heath[2], Rob Knight[4] & Jeffrey I. Gordon[1]

The human distal gut harbours a vast ensemble of microbes (the microbiota) that provide important metabolic capabilities, including the ability to extract energy from otherwise indigestible dietary polysaccharides[1–6]. Studies of a few unrelated, healthy adults have revealed substantial diversity in their gut communities, as measured by sequencing 16S rRNA genes[6–8], yet how this diversity relates to function and to the rest of the genes in the collective genomes of the microbiota (the gut microbiome) remains obscure. Studies of lean and obese mice suggest that the gut microbiota affects energy balance by influencing the efficiency of calorie harvest from the diet, and how this harvested energy is used and stored[3–5]. Here we characterize the faecal microbial communities of adult female monozygotic and dizygotic twin pairs concordant for leanness or obesity, and their mothers, to address how host genotype, environmental exposure and host adiposity influence the gut microbiome. Analysis of 154 individuals yielded 9,920 near full-length and 1,937,461 partial bacterial 16S rRNA sequences, plus 2.14 gigabases from their microbiomes. The results reveal that the human gut microbiome is shared among family members, but that each person's gut microbial community varies in the specific bacterial lineages present, with a comparable degree of co-variation between adult monozygotic and dizygotic twin pairs. However, there was a wide array of shared microbial genes among sampled individuals, comprising an extensive, identifiable 'core microbiome' at the gene, rather than at the organismal lineage, level. Obesity is associated with phylum-level changes in the microbiota, reduced bacterial diversity and altered representation of bacterial genes and metabolic pathways. These results demonstrate that a diversity of organismal assemblages can nonetheless yield a core microbiome at a functional level, and that deviations from this core are associated with different physiological states (obese compared with lean).

leanness (BMI = 18.5–24.9 kg m^{-2}) (one twin pair was lean/overweight (overweight defined as BMI $\geq$ 25 and $<$ 30) and six pairs were overweight/obese). They had not taken antibiotics for at least 5.49 $\pm$ 0.09 months. Each participant completed a detailed medical, lifestyle and dietary questionnaire: study enrolees were broadly representative of the overall Missouri population for BMI, parity, education and marital status (see Supplementary Results). Although all were born in Missouri, they currently live throughout the USA: 29% live in the same house, but some live more than 800 km apart. Because faecal samples are readily attainable and representative of interpersonal differences in gut microbial ecology[7], they were collected from each individual and frozen immediately. The collection procedure was repeated again with an average interval between sampling of 57 $\pm$ 4 days.

To characterize the bacterial lineages present in the faecal microbiotas of these 154 individuals, we performed 16S rRNA sequencing, targeting the full-length gene with an ABI 3730xl capillary sequencer. Additionally, we performed multiplex pyrosequencing with a 454 FLX instrument to survey the gene's V2 variable region[13] and its V6 hypervariable region[14] (Supplementary Tables 1–3).

Complementary phylogenetic and taxon-based methods were used to compare 16S rRNA sequences among faecal communities (see Methods). No matter which region of the gene was examined, individuals from the same family (a twin and her co-twin, or twins and their mother) had a more similar bacterial community structure than unrelated individuals (Fig. 1a and Supplementary Fig. 1a, b), and shared significantly more species-level phylotypes (16S rRNA sequences with $\geq$97% identity comprise each phylotype) ($G = 55.2$, $P < 10^{-12}$ (V2); $G = 12.3$, $P < 0.001$ (V6); $G = 11.3$, $P < 0.001$ (full-length)). No significant correlation was seen between the degree of physical separation of family members' current homes and the degree of similarity between their microbial communities

Turnbaugh PJ, et al. ***Nature***. (2009) 457:480-4

nature Vol 457|22 January 2009|doi:10.1038/nature07540

LETTERS

A core gut microbiome in obese and lean twins

Peter J. Turnbaugh[1], Micah Hamady[3], Tanya Yatsunenko[1], Brandi L. Cantarel[5], Alexis Duncan[2], Ruth E. Ley[1], Mitchell L. Sogin[6], William J. Jones[7], Bruce A. Roe[8], Jason P. Affourtit[9], Michael Egholm[9], Bernard Henrissat[5], Andrew C. Heath[2], Rob Knight[4] & Jeffrey I. Gordon[1]

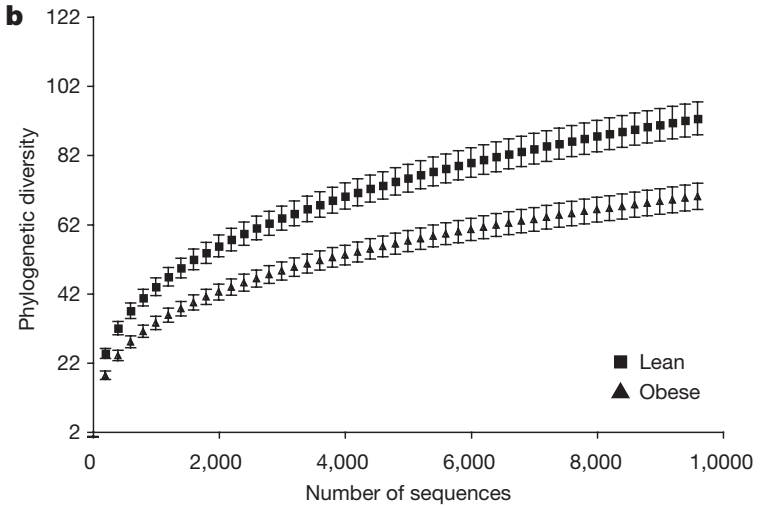


Figure 1 | 16S rRNA gene surveys reveal familial similarity and reduced diversity of the gut microbiota in obese individuals. a, Average unweighted UniFrac distance (a measure of differences in bacterial community structure) between individuals over time (self), twin pairs, twins and their mother, and unrelated individuals (1,000 sequences per V2 data set; Student's *t*-test with Monte Carlo; $*P < 10^{-5}$; $**P < 10^{-14}$; $***P < 10^{-41}$; mean ± s.e.m.). **b**, Phylogenetic diversity curves for the microbiota of lean and obese individuals (based on 1–10,000 sequences per V6 data set; mean ± 95% confidence intervals shown).

Obesity is associated with

- reduced bacterial diversity
- phylum-level changes in microbiota
  - Bacteroidetes ↓
  - Actinobacteria ↑
- altered representation of bacterial genes and metabolic pathways

Ng S.C. et al. ***Journal of Gastroenterology and Hepatology.*** 2013;28:1624-31.

JGH Journal of Gastroenterology and Hepatology JGHF

doi:10.1111/jgh.12306

GASTROENTEROLOGY

Effect of probiotic bacteria on the intestinal microbiota in irritable bowel syndrome

Siew Chien Ng,* Emma F C Lam,* Tommy T Y Lam,‡ Yawen Chan,* Wendy Law,* Pete C H Tse,* Michael A Kamm,†,§ Joseph J Y Sung,* Francis K L Chan* and Justin C Y Wu*

*Institute of Digestive Disease and Department of Medicine and Therapeutics, Li Ka Shing Institute of Health Sciences, Chinese University of Hong Kong, Hong Kong; †St Vincent's Hospital and University of Melbourne, Melbourne, Victoria, Australia; and ‡Department of Zoology, University of Oxford, Oxford, and §Imperial College, London, UK

Key words
IBS, intestinal microbiota, probiotics, pyrosequencing.

Accepted for publication 12 June 2013.

Correspondence
Siew C Ng, Department of Medicine and Therapeutics, Prince of Wales Hospital, Shatin, New Territories, Hong Kong. Email: siewchienng@cuhk.edu.hk

Conflict of interest: There is no conflict of interest to be declared.

Authors' contribution: SC Ng (study concept and design; patient recruitment, and drafting of manuscript), FC Lam (Sample collection and DNA extraction), TT Lam (data analysis, revision of manuscript), WT Law and CH Tse (DNA extraction); MA Kamm, KL Chan and JY Sung (revision of manuscript); CY Wu (study supervision; revision of manuscript).

Abstract

Background and Aim: In irritable bowel syndrome (IBS), the gut microbiota may be altered. Probiotic bacteria appear to be therapeutically effective. We characterized the mucosa-associated microbiota, and determined the clinical and microbiological effects of orally administered probiotic bacteria, in patients with IBS.

Methods: Mucosal microbiota from rectal biopsies of IBS patients and controls were assessed on the V1 and V2 variable regions of the 16S ribosomal RNA gene amplified using 454 pyrosequencing. Clinical symptoms and changes in mucosal microbiota were assessed in IBS patients before and after 4 weeks of treatment with probiotic mix VSL#3.

Results: Ten IBS subjects (eight female; mean age 46 years) were included. At week 4 of probiotic therapy, six patients showed symptom improvement on global symptom assessment compared with baseline ($P = 0.031$). Before therapy, intestinal microbiota of IBS subjects differed significantly from that of healthy controls, with less diversity and evenness than controls ($n = 9$; $P < 0.05$), increased abundance of *Bacteroidetes* ($P = 0.014$) and *Synegitestes* ($P = 0.017$), and reduced abundance of *Actinobacteria* ($P = 0.004$). The classes *Flavobacteria* ($P = 0.028$) and *Epsilonproteobacteria* ($P = 0.017$) were less enriched in IBS. Abundance differences were largely consistent from the phylum to genus level. Probiotic treatment in IBS patients was associated with a significant reduction of the genus *Bacteroides* (all taxonomy levels; $P < 0.05$) to levels similar to that of controls.

Conclusion: In this pilot study, global and deep molecular analysis demonstrates an altered mucosal microbiota composition in IBS. Probiotic leads to detectable changes in the microbiota. These effects of probiotic bacteria may contribute to their therapeutic benefit.

Ng S.C. et al. ***Journal of Gastroenterology and Hepatology.*** 2013;28:1624-31.

JGH Journal of Gastroenterology and Hepatology JGHF

doi:10.1111/jgh.12306

GASTROENTEROLOGY

Effect of probiotic bacteria on the intestinal microbiota in irritable bowel syndrome

Siew Chien Ng,* Emma F C Lam,* Tommy T Y Lam,‡ Yawen Chan,* Wendy Law,* Pete C H Tse,* Michael A Kamm,†,§ Joseph J Y Sung,* Francis K L Chan* and Justin C Y Wu*

*Institute of Digestive Disease and Department of Medicine and Therapeutics, Li Ka Shing Institute of Health Sciences, Chinese University of Hong Kong, Hong Kong; †St Vincent's Hospital and University of Melbourne, Melbourne, Victoria, Australia; and ‡Department of Zoology, University of Oxford, Oxford, and §Imperial College, London, UK

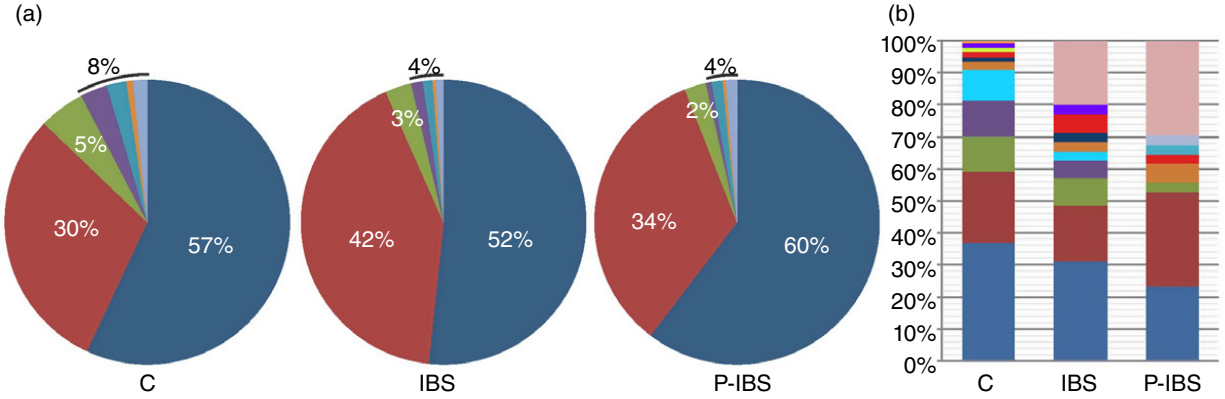


Figure 3 Average abundance of bacterial groups in the mucosal samples from controls (C), IBS patients at baseline (IBS) and week 4 of probiotic therapy (P-IBS). Major phyla are shown in panel (a). ■, *Firmicutes*; ■, *Proteobacteria*; ■, *Actinobacteria**; ■, Unclassified; ■, *Bacteroidetes**; ■, *Fusobacteria*; ■, Others; other minor phyla with trace abundance are shown in panel (b). ■, *Synergistetes**; ■, *Lentisphaerae*; ■, *Gemmatimonadetes*; ■, *Tenericutes*; ■, *Verrucomicrobia*; ■, *Chloroflexi*; ■, *Spirochaetes*; ■, *Nitrospira*; ■, *Deinococcus-Thermus*; ■, *Acidobacteria*; ■, SR1; ■, OD1; ■, *Cyanobacteria**; ■, TM7. Asterisks indicate the phyla that show significant difference ($P < 0.05$, by *t*-test) between controls and IBS patients. Selected bacterial genera are shown in panel (c). Square brackets indicate $P < 0.05$ in the *t*-tests of the two study groups.

IBS is associated with abundance changes of various bacterial lineages

- Bacteroidetes ↑
- Actinobacteria ↓
- Synergistetes ↑
- Cyanobacteria ↓

Probiotic treatment could suppress the increased bacteroidetes, and relieve the disease symptoms

Metagenomic studies
of microbiota

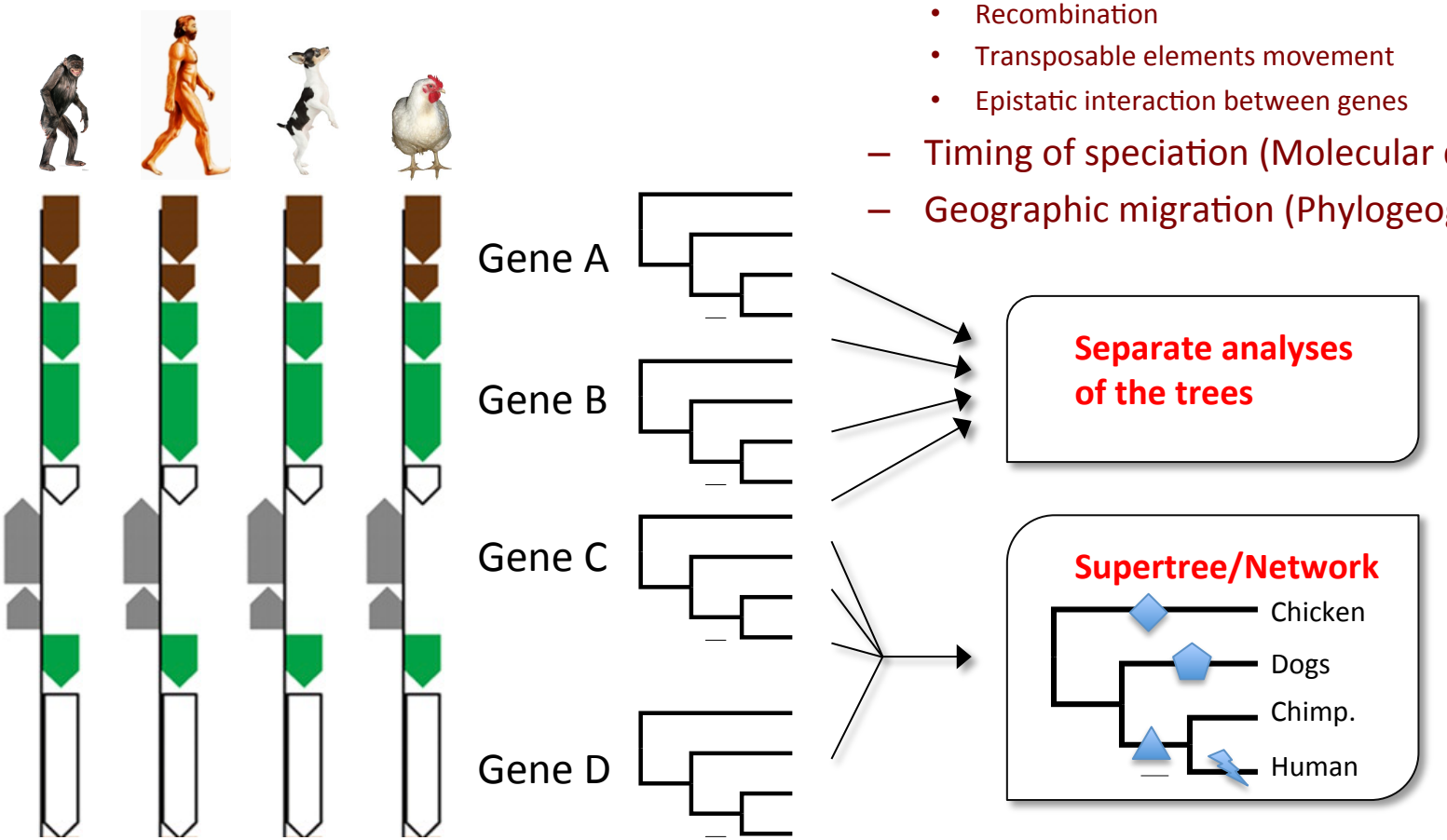
Phylogenomic studies
of organismal evolution

Phylogenomic studies
of pathogen transmission and emergence

Web-based platform ‘Galaxy’

Phylogenomics

- Analyzing the 'whole' genome sequences of the organisms
- Compare their orthologous genes with those of other known species
- Use of phylogenetic tree (or called 'phylogeny') as a analysis tool



- Understand the evolution of the organism, e.g. speciation and adaptation
  - Genes involved in adaptation
  - Molecular mechanisms involved in adaptation
    - Gene/Genome duplication/deletion
    - Selective mutations
    - Recombination
    - Transposable elements movement
    - Epistatic interaction between genes
  - Timing of speciation (Molecular dating)
  - Geographic migration (Phylogeography)

LETTER

D'Hont A. et al. *Nature*. (2012) 488:213-7.

doi:10.1038/nature11241

The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants

Angélique D'Hont[1]*, France Denoeud[2,3,4]*, Jean-Marc Aury[2], Franc-Christophe Baurens[1], Françoise Carreel[1,5], Olivier Garsmeur[1], Benjamin Noel[2], Stéphanie Bocs[1], Gaëtan Droc[1], Mathieu Rouard[6], Corinne Da Silva[2], Kamel Jabbari[2,3,4], Céline Cardi[1], Julie Poulain[2], Marlène Souquet[1], Karine Labadie[2], Cyril Jourda[1], Juliette Lengellé[1], Marguerite Rodier-Goud[1], Adriana Alberti[2], Maria Bernard[2], Margot Correa[2], Saravanaraj Ayyampalayam[7], Michael R. Mckain[7], Jim Leebens-Mack[7], Diane Burgess[8], Mike Freeling[8], Didier Mbéguié-A-Mbéguié[9], Matthieu Chabannes[5], Thomas Wicker[10], Olivier Panaud[11], Jose Barbosa[11], Eva Hribova[12], Pat Heslop-Harrison[13], Rémy Habas[5], Ronan Rivallan[1], Philippe Francois[1], Claire Poiron[1], Andrzej Kilian[14], Dheema Burthia[1], Christophe Jenny[1], Frédéric Bakry[1], Spencer Brown[15], Valentin Guignon[1,6], Gert Kema[16], Miguel Dita[19], Cees Waalwijk[16], Steeve Joseph[1], Anne Dievart[1], Olivier Jaillon[2,3,4], Julie Leclercq[1], Xavier Argout[1], Eric Lyons[17], Ana Almeida[8], Mouna Jeridi[1], Jaroslav Dolezel[12], Nicolas Roux[6], Ange-Marie Risterucci[1], Jean Weissenbach[2,3,4], Manuel Ruiz[1], Jean-Christophe Glaszmann[1], Francis Quétier[18], Nabila Yahiaoui[1] & Patrick Wincker[2,3,4]

Bananas (*Musa* spp.), including dessert and cooking types, are giant perennial monocotyledonous herbs of the order Zingiberales, a sister group to the well-studied Poales, which include cereals. Bananas are vital for food security in many tropical and subtropical countries and the most popular fruit in industrialized countries[1]. The *Musa* domestication process started some 7,000 years ago in Southeast Asia. It involved hybridizations between diverse species and subspecies, fostered by human migrations[2], and selection of diploid and triploid seedless, parthenocarpic hybrids thereafter widely dispersed by vegetative propagation. Half of the current production relies on somaclones derived from a single triploid genotype (Cavendish)[1]. Pests and diseases have gradually become adapted, representing an imminent danger for global banana production[3,4]. Here we describe the draft sequence of the 523-megabase genome of a *Musa acuminata* doubled-haploid genotype, providing a crucial stepping-stone for genetic improvement of banana. We detected three rounds of whole-genome duplications in the *Musa* lineage, independently of those previously described in the Poales lineage and the one we detected in the Arecales lineage. This first monocotyledon high-continuity whole-genome sequence reported outside Poales represents an essential bridge for comparative genome analysis in plants. As such, it clarifies commelinid-monocotyledon phylogenetic relationships, reveals Poaceae-specific features and has led to the discovery of conserved non-coding sequences predating monocotyledon–eudicotyledon divergence.

sequence errors. The assembly consisted of 24,425 contigs and 7,513 scaffolds with a total length of 472.2 Mb, which represented 90% of the estimated DH-Pahang genome size. Ninety per cent of the assembly was in 647 scaffolds, and the N50 (the scaffold size above which 50% of the total length of the sequence assembly can be found) was 1.3 Mb (Supplementary Text and Supplementary Tables 1–3). We anchored 70% of the assembly (332 Mb) along the 11 *Musa* linkage groups of the Pahang genetic map. This corresponded to 258 scaffolds and included 98.0% of the scaffolds larger than 1 Mb and 92% of the annotated genes (Supplementary Text, Supplementary Table 4 and Supplementary Fig. 1).

We identified 36,542 protein-coding gene models in the *Musa* genome (Supplementary Tables 1 and 5). A total of 235 microRNAs from 37 families were identified, including only one of the eight microRNA gene (*MIR*) families found so far solely in Poaceae[8] (Supplementary Tables 6 and 7).

Viral sequences related to the banana streak virus (BSV) dsDNA plant pararetrovirus were found to be integrated in the Pahang genome, with 24 loci spanning 10 chromosomes (Supplementary Text and Supplementary Fig. 2). They belonged to a badnavirus phylogenetic group that differed from the endogenous BSV species (eBSV) found in *M. balbisiana*[9] and most of them formed a new subgroup (Supplementary Fig. 3). Importantly, all of the integrations were highly reorganized and fragmented and thus did not seem to be capable of forming free infectious viral particles, contrary to the eBSV described in *M. balbisiana*[10].

LETTER

D'Hont A. et al. ***Nature***. (2012) 488:213-7.
doi:10.1038/nature11241

The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants

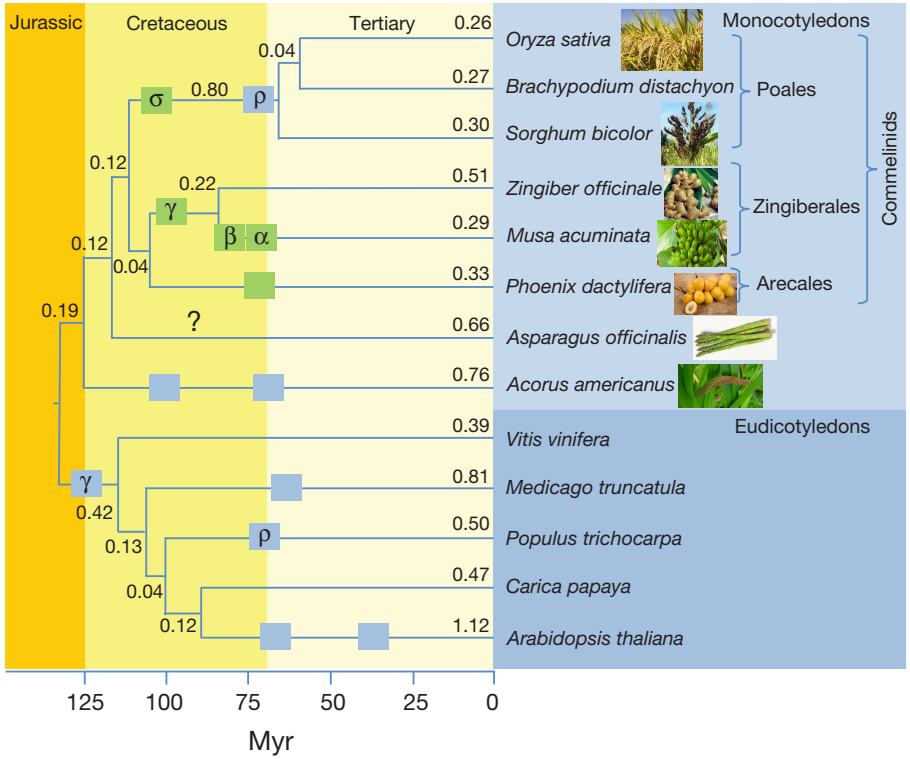


Figure 3 | Timing of whole-genome duplications relative to speciation events within representative monocotyledons and eudicotyledons. Boxes indicate WGD events. Green boxes indicate WGD events analysed in this paper. All nodes have 100% bootstrap support in a maximum likelihood analysis. Branch lengths (synonymous substitution rate) are indicated. The timing of the β WGD event relative to the Musaceae/Zingiberaceae split remains to be clarified.

Li R. et al. ***Nature***. (2010) 463:311-7.

Vol 463|21 January 2010|**doi:10.1038/nature08696** nature

ARTICLES

The sequence and *de novo* assembly of the giant panda genome

Ruiqiang Li[1,2]*, Wei Fan[1]*, Geng Tian[1,3]*, Hongmei Zhu[1]*, Lin He[4,5]*, Jing Cai[3,6]*, Quanfei Huang[1], Qingle Cai[1,7], Bo Li[1], Yinqi Bai[1], Zhihe Zhang[8], Yaping Zhang[6], Wen Wang[6], Jun Li[1], Fuwen Wei[9], Heng Li[10], Min Jian[1], Jianwen Li[1], Zhaolei Zhang[11], Rasmus Nielsen[12], Dawei Li[1], Wanjun Gu[13], Zhentao Yang[1], Zhaoling Xuan[1], Oliver A. Ryder[14], Frederick Chi-Ching Leung[15], Yan Zhou[1], Jianjun Cao[1], Xiao Sun[16], Yonggui Fu[17], Xiaodong Fang[1], Xiaosen Guo[1], Bo Wang[1], Rong Hou[8], Fujun Shen[8], Bo Mu[1], Peixiang Ni[1], Runmao Lin[1], Wubin Qian[1], Guodong Wang[3,6], Chang Yu[1], Wenhui Nie[6], Jinhuan Wang[6], Zhigang Wu[1], Huiqing Liang[1], Jiumeng Min[1,7], Qi Wu[9], Shifeng Cheng[1,7], Jue Ruan[1,3], Mingwei Wang[1], Zhongbin Shi[1], Ming Wen[1], Binghang Liu[1], Xiaoli Ren[1], Huisong Zheng[1], Dong Dong[11], Kathleen Cook[11], Gao Shan[1], Hao Zhang[1], Carolin Kosiol[18], Xueying Xie[13], Zuhong Lu[13], Hancheng Zheng[1], Yingrui Li[1,3], Cynthia C. Steiner[14], Tommy Tsan-Yuk Lam[15], Siyuan Lin[1], Qinghui Zhang[1], Guoqing Li[1], Jing Tian[1], Timing Gong[1], Hongde Liu[16], Dejin Zhang[16], Lin Fang[1], Chen Ye[1], Juanbin Zhang[1], Wenbo Hu[17], Anlong Xu[17], Yuanyuan Ren[1], Guojie Zhang[1,3,6], Michael W. Bruford[19], Qibin Li[1,3], Lijia Ma[1,3], Yiran Guo[1,3], Na An[1], Yujie Hu[1,3], Yang Zheng[1,3], Yongyong Shi[5], Zhiqiang Li[5], Qing Liu[1], Yanling Chen[1], Jing Zhao[1], Ning Qu[1,7], Shancen Zhao[1], Feng Tian[1], Xiaoling Wang[1], Haiyin Wang[1], Lizhi Xu[1], Xiao Liu[1], Tomas Vinar[20], Yajun Wang[21], Tak-Wah Lam[22], Siu-Ming Yiu[22], Shiping Liu[23], Hemin Zhang[24], Desheng Li[24], Yan Huang[24], Xia Wang[1], Guohua Yang[1], Zhi Jiang[1], Junyi Wang[1], Nan Qin[1], Li Li[1], Jingxiang Li[1], Lars Bolund[1], Karsten Kristiansen[1,2], Gane Ka-Shu Wong[1,25], Maynard Olson[26], Xiuqing Zhang[1], Songgang Li[1], Huanming Yang[1], Jian Wang[1] & Jun Wang[1,2]

Using next-generation sequencing technology alone, we have successfully generated and assembled a draft sequence of the giant panda genome. The assembled contigs (2.25 gigabases (Gb)) cover approximately 94% of the whole genome, and the remaining gaps (0.05 Gb) seem to contain carnivore-specific repeats and tandem repeats. Comparisons with the dog and human showed that the panda genome has a lower divergence rate. The assessment of panda genes potentially underlying some of its unique traits indicated that its bamboo diet might be more dependent on its gut microbiome than its own genetic composition. We also identified more than 2.7 million heterozygous single nucleotide polymorphisms in the diploid genome. Our data and analyses provide a foundation for promoting mammalian genetic research, and demonstrate the feasibility for using next-generation sequencing technologies for accurate, cost-effective and rapid *de novo* assembly of large eukaryotic genomes.

Li R. et al. ***Nature***. (2010) 463:311-7.

Vol 463|21 January 2010|doi:10.1038/nature08696 nature

ARTICLES

The sequence and *de novo* assembly of the giant panda genome

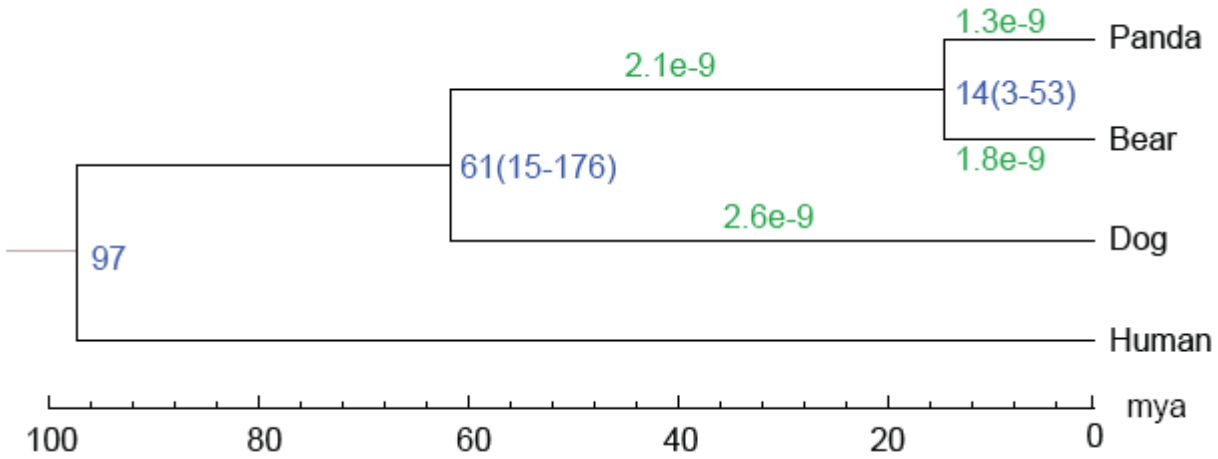
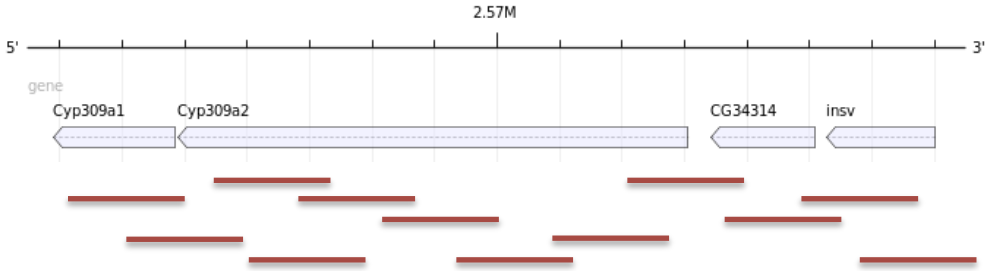


Figure SA6 | Estimation of divergence time and substitution rate. The green numbers on the branches are the estimated substitution rate (substitutions per site per year). The blue numbers on the nodes are the divergence time from present (million years ago, Mya). The calibration time (97 Mya) from human-dog divergence was derived from the TimeTree database (http://www.timetree.org).

- Divergence for giant panda was estimated from the 20 orthologous genes from giant panda, American black bear, dog, human, mouse, and opossum (as an outgroup).
- Giant panda lineage diverged from their common ancestor with bear in about 14 Mya.

Overview of Phylogenomic Data Analysis



Genome

$4x10^8$ reads in 250nt [100Gb]
(HiSeq 2500 specification)

Sequence read quality-check and filter

Chimera detection and filter

Genome assembling

assembled genome
[6Kb - 1Mb for viruses]
[0.1 - 20 Mb for prokaryotes]
[0.01 - 6Gb for eukaryotes]

Gene annotation

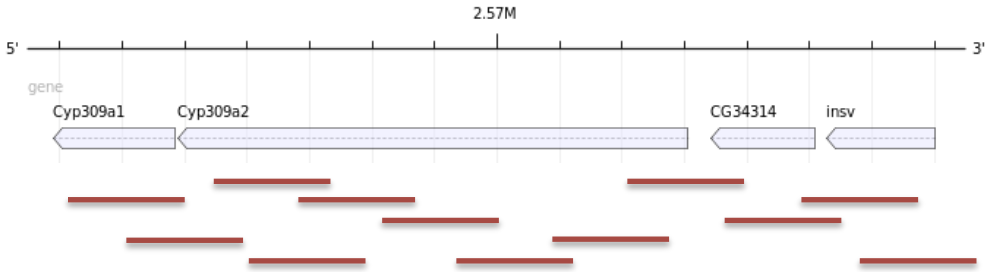
Whole genome alignment (with other species)

Orthologous gene identification and alignment (with other species)

Genome structural and functional analyses

Phylogenetic and evolutionary analyses

Overview of Phylogenomic Data Analysis



Genome

H L

* parallelizable

^ large memory required

$4x10^8$ reads in 250nt [100Gb]
(HiSeq 2500 specification)

Sequence read quality-check and filter *

Chimera detection and filter ^

Genome assembling ^

assembled genome
[6Kb - 1Mb for viruses]
[0.1 - 20 Mb for prokaryotes]
[0.01 - 6Gb for eukaryotes]

Gene annotation *

Whole genome alignment (with other species) ^

Orthologous gene identification and alignment (with other species) *

Genome structural and functional analyses

Phylogenetic and evolutionary analyses *

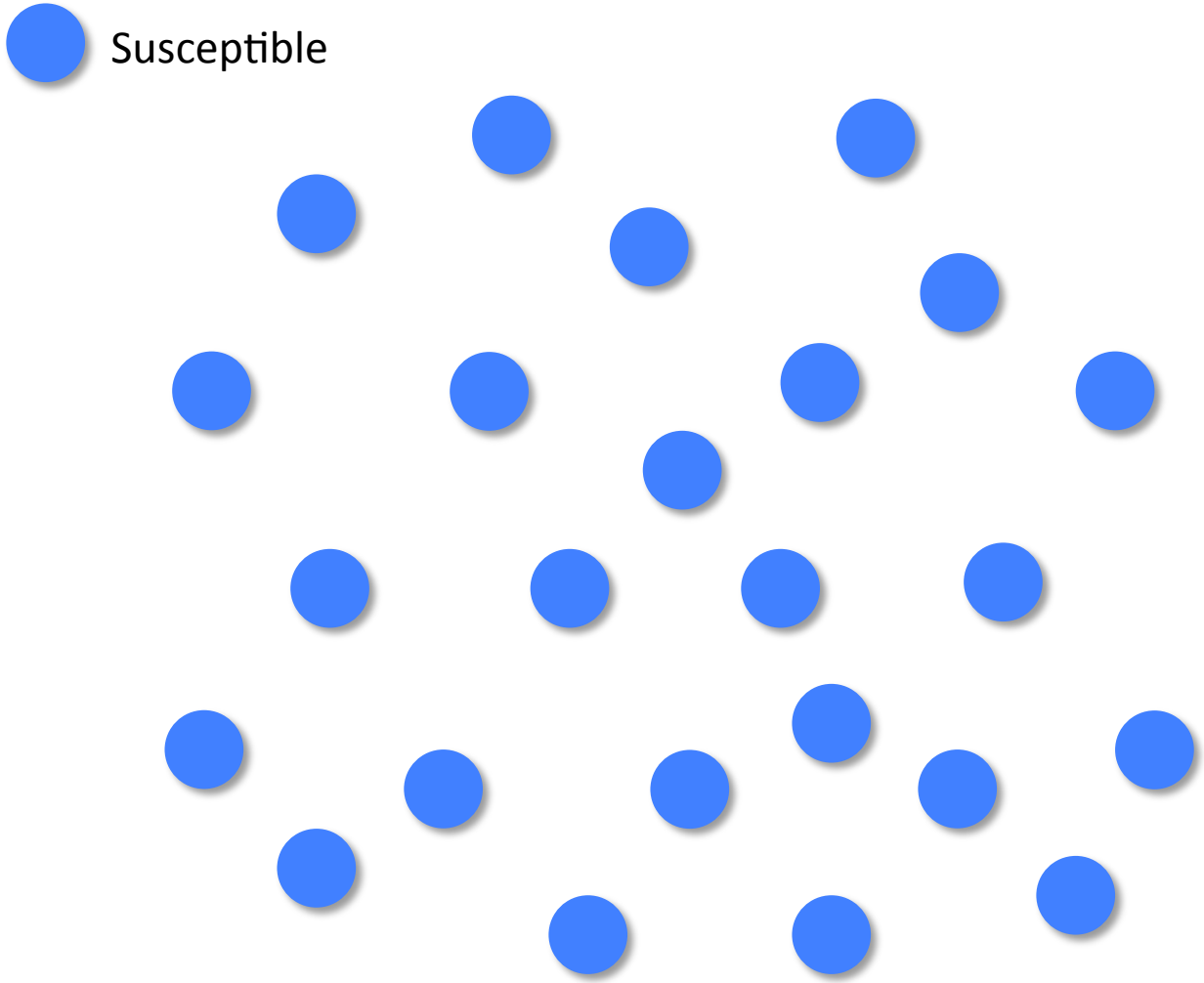
Metagenomic studies
of microbiota

Phylogenomic studies
of organismal evolution

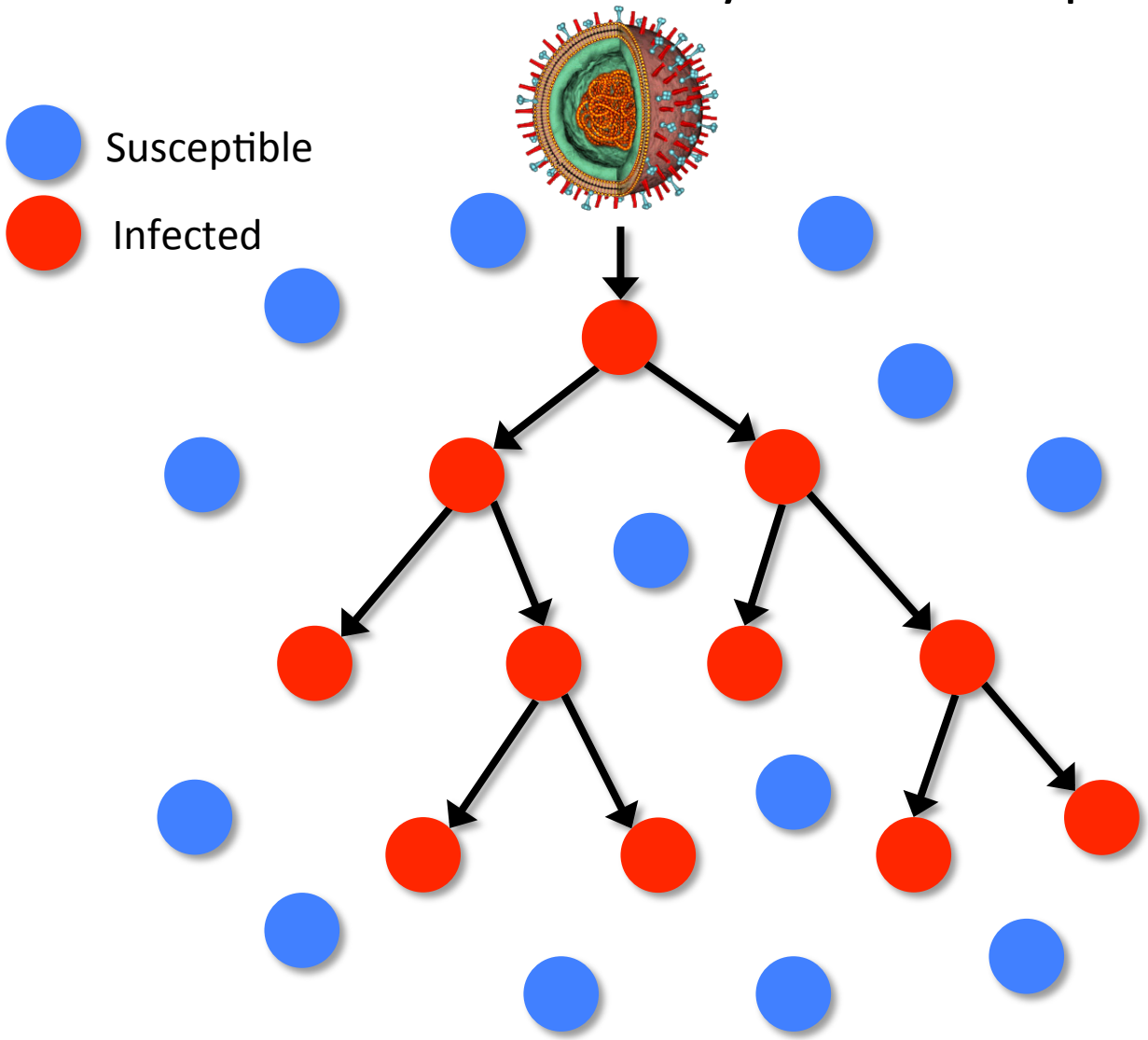
Phylogenomic studies
of pathogen transmission and emergence

Web-based platform ‘Galaxy’

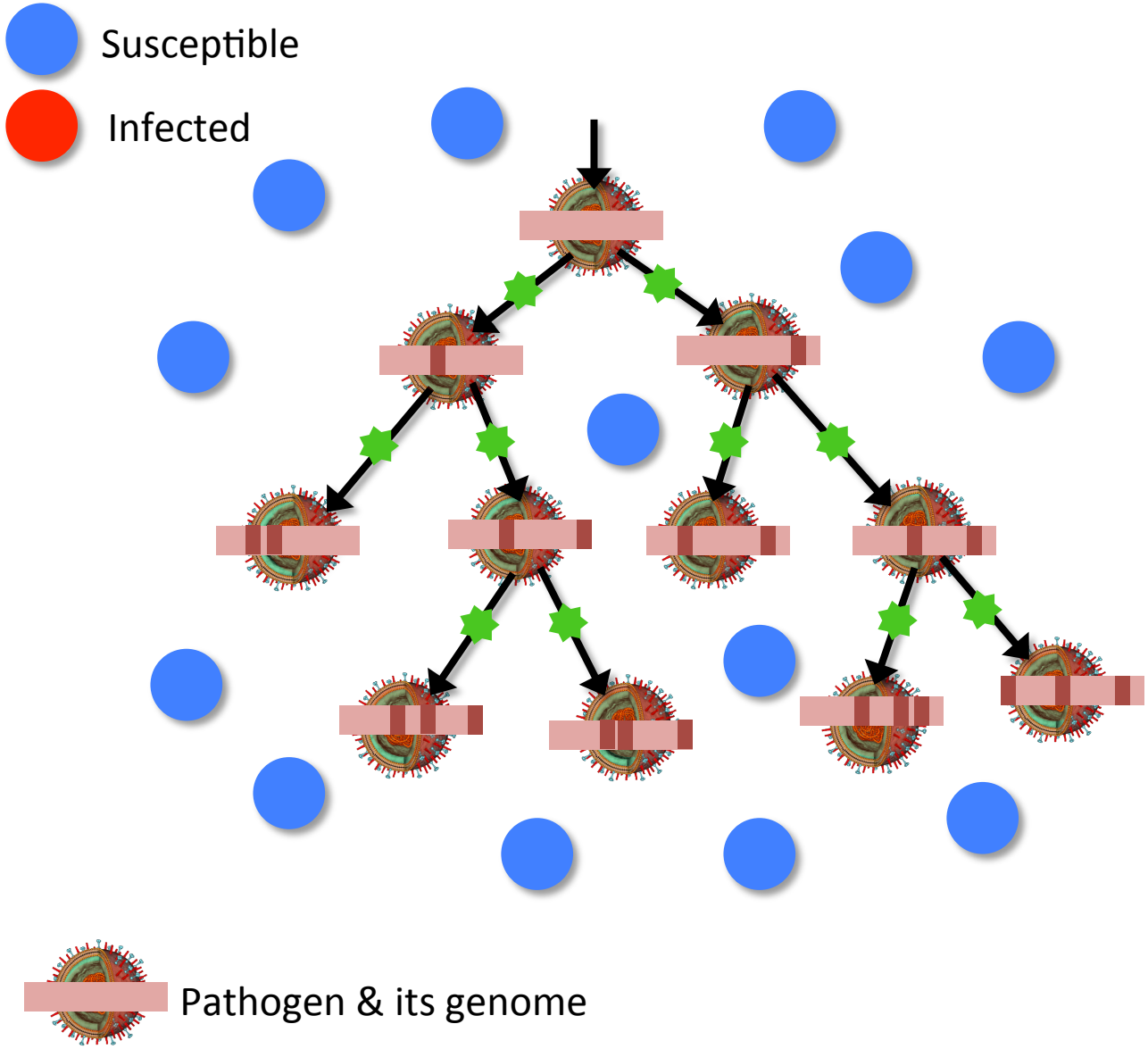
Transmission history is stored in pathogen genomes



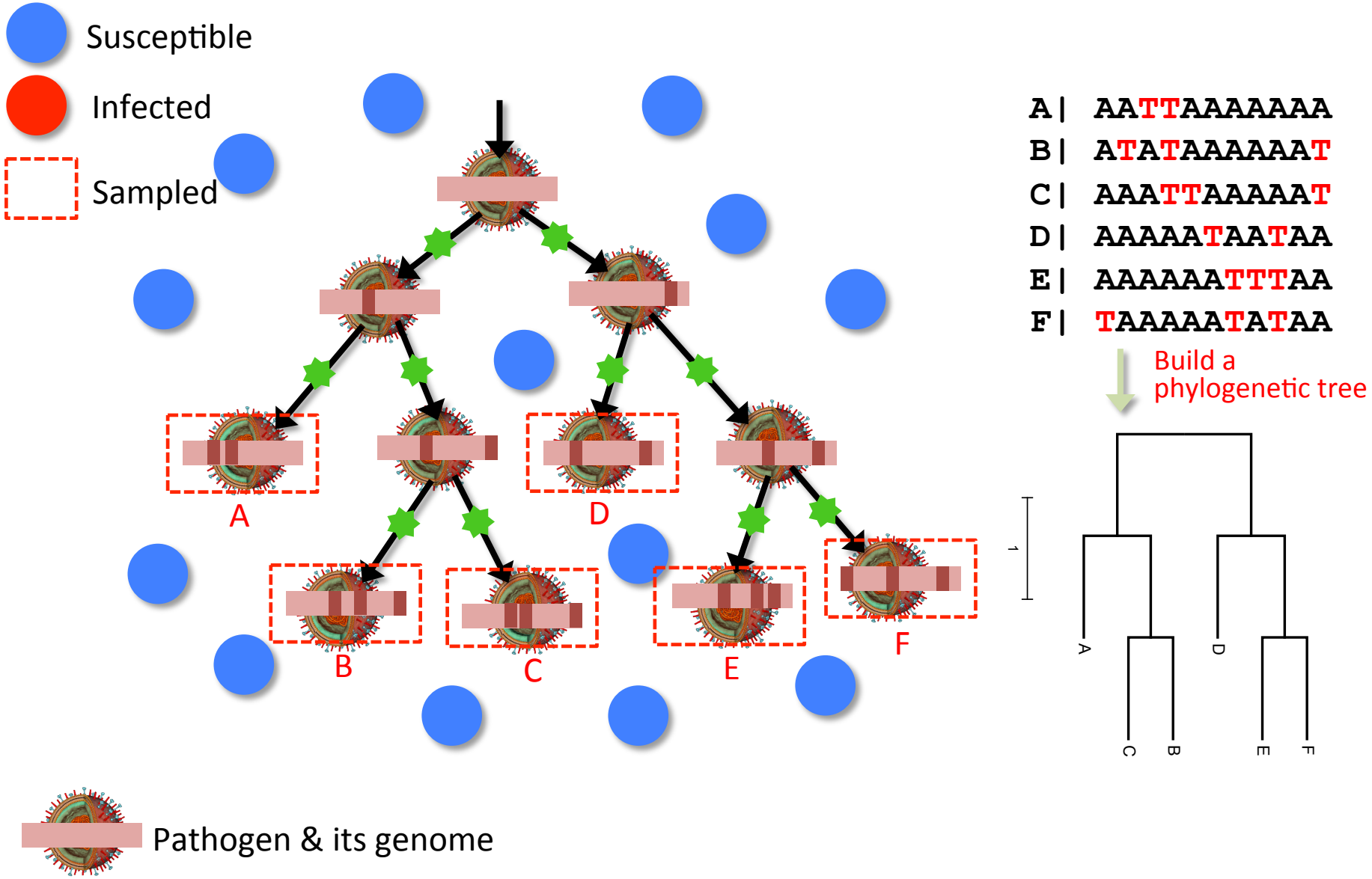
Transmission history is stored in pathogen genomes



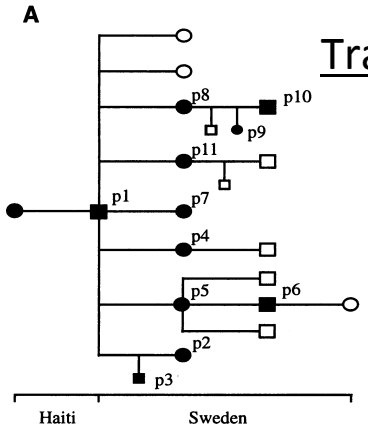
Transmission history is stored in pathogen genomes



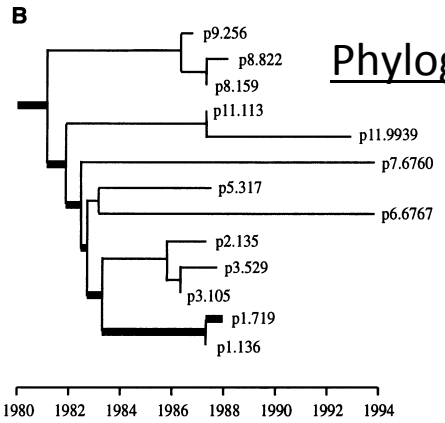
Transmission history is stored in pathogen genomes



Transmission history is stored in pathogen genomes



Transmission history



Phylogenetic tree

FIG. 1. The real time population history of HIV-1 in a Swedish transmission cluster. (*A*) Pattern of the contact tracing (20). Squares denote males and circles females, smaller symbols denote children. Solid and open symbols denote HIV-1 infected and uninfected persons, respectively. (*B*) The true tree, obtained by combining information about when the virus transmissions occurred and when the samples were obtained. Each ramification denotes a transmission event, where the virus population of the donor and recipient continues

Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 10864–10869, October 1996
Evolution

Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis

THOMAS LEITNER*, DAVID ESCANILLA†, CHRISTER FRANZÉN‡, MATHIAS UHLÉN§, AND JAN ALBERT*¶

Experts Express Concerns Over Use Of HIV Fingerprinting To Establish Proof Of HIV Criminal Transmission

By Kieryn Graham and Courtney McQueen No Comment
Published: Feb 15, 2011 4:19 pm

In a recent article in the Lancet Infectious Diseases, experts warned that a forensics technique called HIV phylogenetic analysis, sometimes called HIV fingerprinting, cannot definitively establish whether a specific individual transmitted HIV to another person. The authors also stressed the need for scientists to recognize the limitations of the technique as a basis for proving HIV criminal transmission.

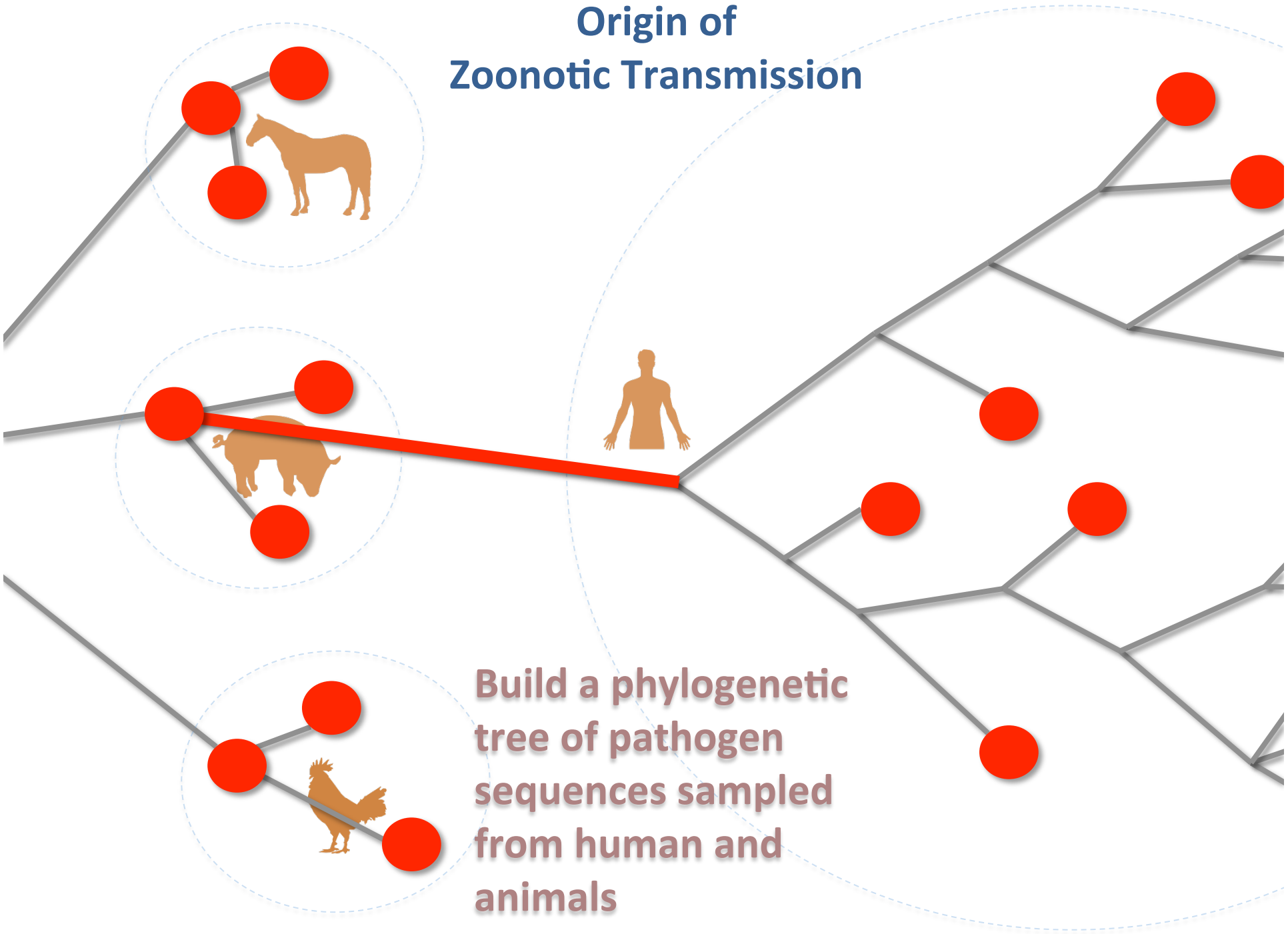
"Phylogenetic analysis is more powerful in its ability to exclude certain scenarios," said Professor Anne-Meike Vandamme of the Rega Institute for Medical Research at Katholieke Universiteit Leuven and coauthor of the Lancet article, in correspondence with The AIDS Beacon. "Phylogenetics can prove that people cannot have infected each other, but it can never prove that people infected each other."

The experts listed several guidelines for scientific experts to follow in order to prevent the misuse of HIV fingerprinting evidence in HIV criminal transmission trials. The recommendations include using a sufficient number of comparison samples from other people in the area who have HIV; not informing experts as to which HIV samples are from the accused and which are from the accuser while the tests are being run; and taking special care to correctly word their findings in light of the limitations of the technique.

HIV Criminal Transmission And Phylogenetic Analysis

Criminal transmission of HIV is defined as the intentional or reckless infection of another person with HIV. In most states, it is illegal for a person with HIV to have unprotected sex without disclosing his or her HIV status.

Which host species did the emerging pathogen originate?


Origin of
Zoonotic Transmission
Build a phylogenetic tree of pathogen sequences sampled from human and animals

ECOLOGY LETTERS

Ecology Letters, (2012) **15**: 24–33 doi: 10.1111/j.1461-0248.2011.01703.x

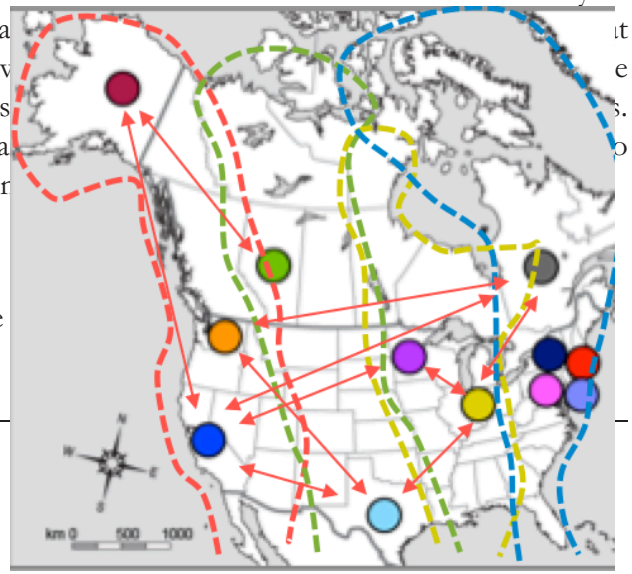
LETTER

Migratory flyway and geographical distance are barriers to the gene flow of influenza virus among North American birds

Tommy Tsan-Yuk Lam,[1] Hon S. Ip,[2] Elodie Ghedin,[3,4] David E. Wentworth,[5] Rebecca A. Halpin,[4] Timothy B. Stockwell,[4] David J. Spiro,[4] Robert J. Dusek,[2] James B. Bortner,[6] Jenny Hoskins,[6] Bradley D. Bales,[7] Dan R. Yparraguirre[8] and Edward C. Holmes[1,9]*

Abstract

Despite the importance of migratory birds in the ecology and evolution of avian influenza virus (AIV), there is a lack of information on the patterns of AIV spread at the intra-continental scale. We applied a variety of statistical phylogeographic techniques to a plethora of viral genome sequence data to determine the strength, pattern and determinants of gene flow in AIV sampled from wild birds in North America. These analyses revealed a clear isolation-by-distance of AIV among sa[illegible]t phylogeographic models incorporating information on the av[illegible]e observed sequence data than those specifying homogeneous[illegible]s. In sum, these data strongly suggest that the intra-continenta[illegible]o major ecological barriers, including spatial distance and aviar



Keywords

Avian influenza, ecological barriers, evolution, flyways, gene

Ecology Letters (2012) **15:** 24–33

ECOLOGY LETTERS

Ecology Letters, (2012) **15**: 24–33 doi: 10.1111/j.1461-0248.2011.01703.x

LETTER

Migratory flyway and geographical distance are barriers to the gene flow of influenza virus among North American birds

Tommy Tsan-Yuk Lam,[1] Hon S. Ip,[2] Elodie Ghedin,[3,4] David E. Wentworth,[5] Rebecca A. Halpin,[4] Timothy B. Stockwell,[4] David J. Spiro,[4] Robert J. Dusek,[2] James B. Bortner,[6] Jenny Hoskins,[6] Bradley D. Bales,[7] Dan R. Yparraguirre[8] and Edward C. Holmes[1,9]*

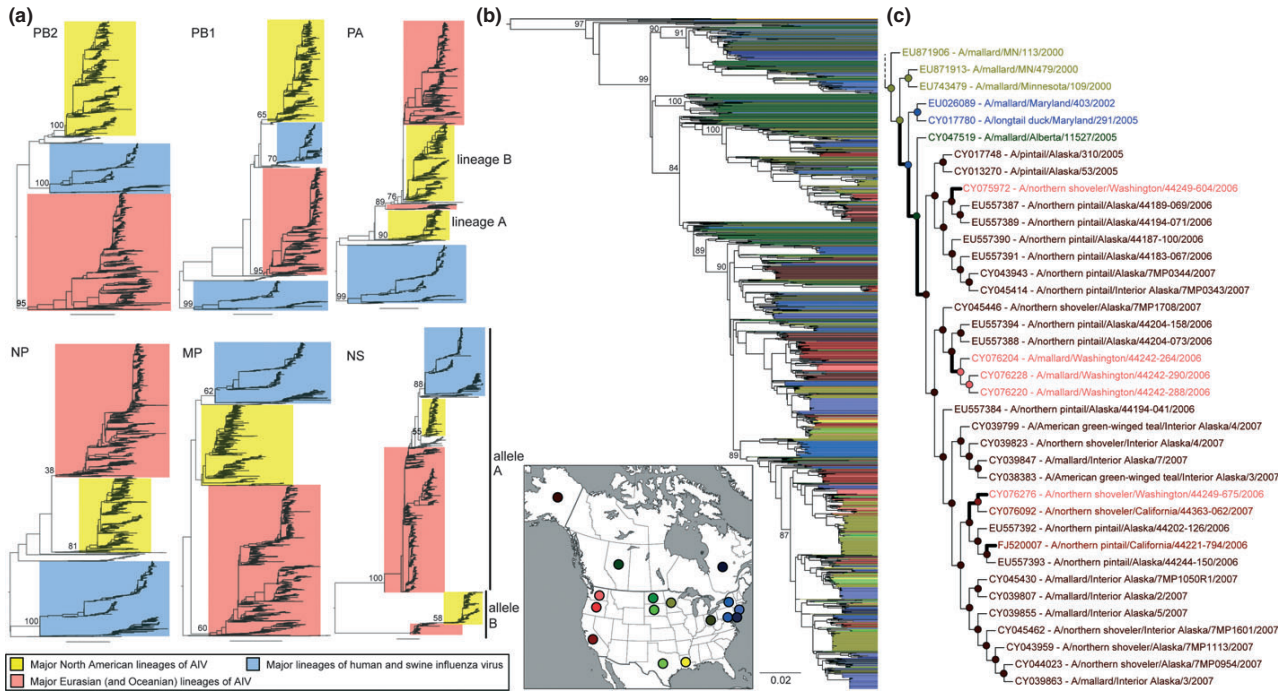


Figure 1 Phylogenetic trees of the internal genome segments of influenza A virus. (a) 'Panoramic' phylogenies of the PB2, PB1, PA, NP, MP and NS segments. Bootstrap support values are shown adjacent to selected nodes, and the scale bar represents 0.1 substitutions/site. Major North American lineages of avian influenza virus (AIV) are highlighted in yellow. (b) Maximum likelihood phylogeny of the major North American wild bird AIV (NA-WB-AIV) lineage inferred from the PB2 gene. Terminal branches are extended and coloured according to the place of sampling as shown in the inset map (14 US states and two Canadian provinces; others shown in grey). The scale bar is 0.02 substitutions/site. (c) A sub-lineage of NA-WB-AIV in the PB2 phylogeny, shown as a cladogram. Geographical states of the ancestral nodes (circles) were estimated, using parsimony, from taxon localities, and their colours are the same as those in panel (a). Changes in geographical state that occurred during evolutionary history are indicated by bold branches.

ECOLOGY LETTERS

Ecology Letters, (2012) **15**: 24–33 doi: 10.1111/j.1461-0248.2011.01703.x

LETTER

Migratory flyway and geographical distance are barriers to the gene flow of influenza virus among North American birds

Tommy Tsan-Yuk Lam,[1] Hon S. Ip,[2] Elodie Ghedin,[3,4] David E. Wentworth,[5] Rebecca A. Halpin,[4] Timothy B. Stockwell,[4] David J. Spiro,[4] Robert J. Dusek,[2] James B. Bortner,[6] Jenny Hoskins,[6] Bradley D. Bales,[7] Dan R. Yparraguirre[8] and Edward C. Holmes[1,9]*

- Transmission of avian influenza virus across avian migratory flyways is more restricted than within a flyway.

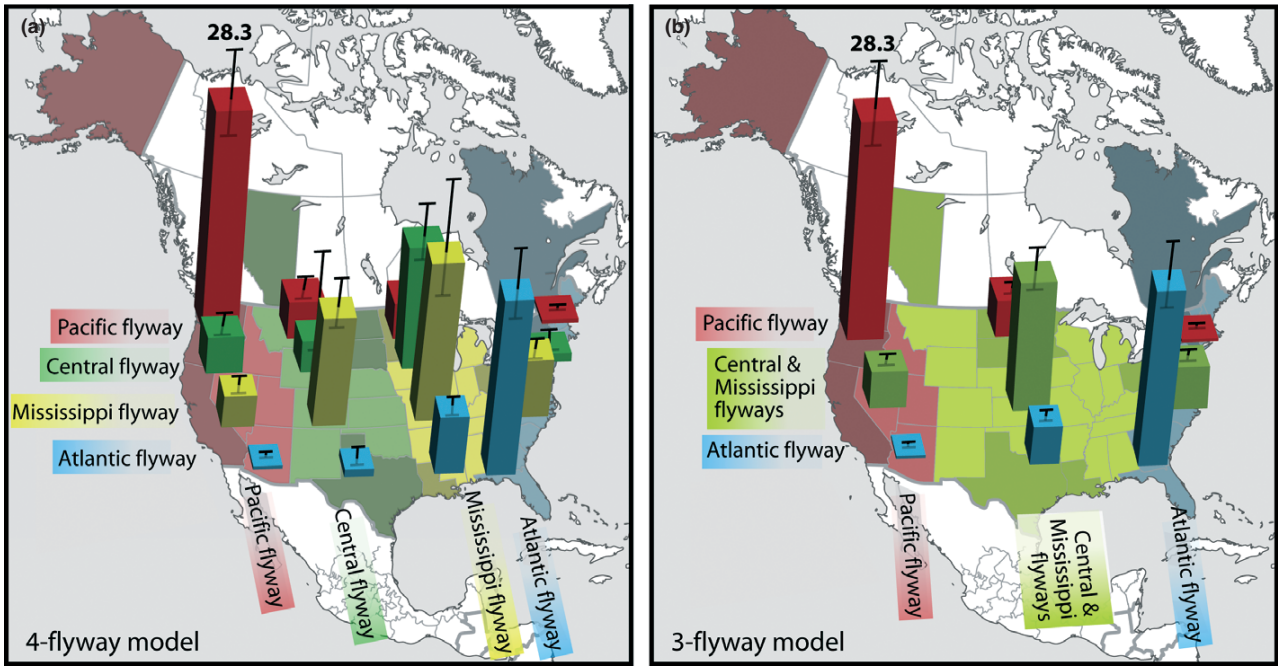


Figure 3 Flyway-specific rates of avian influenza virus (AIV) gene flow. (a) Maximum likelihood estimates (q) of the average level of AIV internal gene flow (excluding the NS gene which has large confidence intervals) within and between the four flyways; Pacific flyway (PF; red), Central flyway (CF; green), Mississippi flyway (MF; yellow) and Atlantic flyway (AF; blue). The 4 × 4 rate matrices were projected onto the map. The colour of the bar and the geographical region where the bar is located denote the flyway where gene flow is measured. For example, the red bar in the yellow terrestrial region represents the extent of gene flow between the PF and MF, while the red bar in the red terrestrial region represents the gene flow within PF. The rate of AIV gene flow within PF is shown next to the bar. Average 95% confidence intervals for the rate estimates are indicated by error bars. (b) The rate matrices (3 × 3) in a 3-flyway model in which CF and MF are merged.

LETTER

Lam T.T.-Y. et al. *Nature*. (2013) 502:241-4.

doi:10.1038/nature12515

The genesis and source of the H7N9 influenza viruses causing human infections in China

Tommy Tsan-Yuk Lam[1,2,3]*, Jia Wang[1,3]*, Yongyi Shen[1,3,4]*, Boping Zhou[2], Lian Duan[2,3], Chung-Lam Cheung[3], Chi Ma[1,3], Samantha J. Lycett[5], Connie Yin-Hung Leung[3], Xinchun Chen[2], Lifeng Li[1,2,3], Wenshan Hong[1], Yujuan Chai[2,3], Linlin Zhou[3], Huyi Liang[1,2,3], Zhihua Ou[1,2,3], Yongmei Liu[1,3], Amber Farooqui[6], David J. Kelvin[6], Leo L. M. Poon[2,3], David K. Smith[1,3], Oliver G. Pybus[7,8], Gabriel M. Leung[1,3], Yuelong Shu[9], Robert G. Webster[10], Richard J. Webby[10], Joseph S. M. Peiris[2,3], Andrew Rambaut[5,11], Huachen Zhu[1,2,3] & Yi Guan[1,2,3]

A novel H7N9 influenza A virus first detected in March 2013 has since caused more than 130 human infections in China, resulting in 40 deaths[1,2]. Preliminary analyses suggest that the virus is a reassortant of H7, N9 and H9N2 avian influenza viruses, and carries some amino acids associated with mammalian receptor binding, raising concerns of a new pandemic[1,3,4]. However, neither the source populations of the H7N9 outbreak lineage nor the conditions for its genesis are fully known[5]. Using a combination of active surveillance, screening of virus archives, and evolutionary analyses, here we show that H7 viruses probably transferred from domestic duck to chicken populations in China on at least two independent occasions. We show that the H7 viruses subsequently reassorted with enzootic H9N2 viruses to generate the H7N9 outbreak lineage, and a related previously unrecognized H7N7 lineage. The H7N9 outbreak lineage has spread over a large geographic region and is prevalent in chickens at live poultry markets, which are thought to be the immediate source of human infections. Whether the H7N9 outbreak lineage has, or will, become enzootic in China and neighbouring regions requires further investigation. The discovery here of a related H7N7 influenza virus in chickens that has the ability to infect mammals experimentally, suggests that H7 viruses may pose threats beyond the current outbreak. The continuing prevalence of H7 viruses in poultry could lead to the generation of highly pathogenic variants and further sporadic human infections, with a continued risk of the virus acquiring human-to-human transmissibility.

duck isolates that were H7N2 and H7N3 viruses. All H9 isolates were H9N2 viruses (80 from LPMs, 5 from farms). At LPMs in Wenzhou, the H7 virus was at its highest prevalence in chickens (10.1%; 46 out of 457), followed by ducks (2.4%; 3 out of 125) and pigeons (1.6%; 3 out of 188). In Rizhao, LPM H7N9 viruses were only found in chickens (0.7%; 8 out of 1,113). Of the chicken isolates, 100% of H7N9, 65.3% of H7N7 and 94.8% of H9N2 viruses were from oropharyngeal swabs (Supplementary Table 1), suggesting that these H7N9 and H7N7 viruses might replicate in the upper respiratory tract of terrestrial poultry, similar to the enzootic H9N2 viruses[6].

These samples were sequenced to investigate the evolutionary history of avian influenza viruses implicated in the current outbreak of H7N9 infections of humans and poultry. Full genome sequences were obtained for 34 H7N7, 4 H7N9 and 19 H9N2 isolates. The H7 and N7/9 genes of 16 mixed H7/H9 infections were sequenced (Supplementary Table 1), as were 3 H7N9 and 3 H7N7 samples that had multiple H9N2-like internal gene segments. The H7 haemagglutinin gene sequences of the H7N9 viruses isolated from chickens in Rizhao formed a tight monophyletic group (Fig. 1a, lineage 'b') with previously reported human and avian viruses from the current H7N9 outbreak. This was most closely related to a group comprising mainly H7N7 viruses obtained from Wenzhou chickens, ducks and pigeons (Fig. 1a, lineage 'c'). All viruses isolated from chickens in these two groups had internal gene complexes that were closely related to those present in co-circulating H9N2 viruses.

To examine the genesis of these H7N9 and H7N7 viruses, we sequenced

LETTER

Lam T.T.-Y. et al. *Nature*. (2013) 502:241-4.

doi:10.1038/nature12515

The genesis and source of the H7N9 influenza viruses causing human infections in China

Tommy Tsan-Yuk Lam[1,2,3]*, Jia Wang[1,3]*, Yongyi Shen[1,3,4]*, Boping Zhou[2], Lian Duan[2,3], Chung-Lam Cheung[3], Chi Ma[1,3], Samantha J. Lycett[5], Connie Yin-Hung Leung[3], Xinchun Chen[2], Lifeng Li[1,2,3], Wenshan Hong[1], Yujuan Chai[2,3], Linlin Zhou[3], Huyi Liang[1,2,3], Zhihua Ou[1,2,3], Yongmei Liu[1,3], Amber Farooqui[6], David J. Kelvin[6], Leo L. M. Poon[2,3], David K. Smith[1,3], Oliver G. Pybus[7,8], Gabriel M. Leung[1,3], Yuelong Shu[9], Robert G. Webster[10], Richard J. Webby[10], Joseph S. M. Peiris[2,3], Andrew Rambaut[5,11], Huachen Zhu[1,2,3] & Yi Guan[1,2,3]

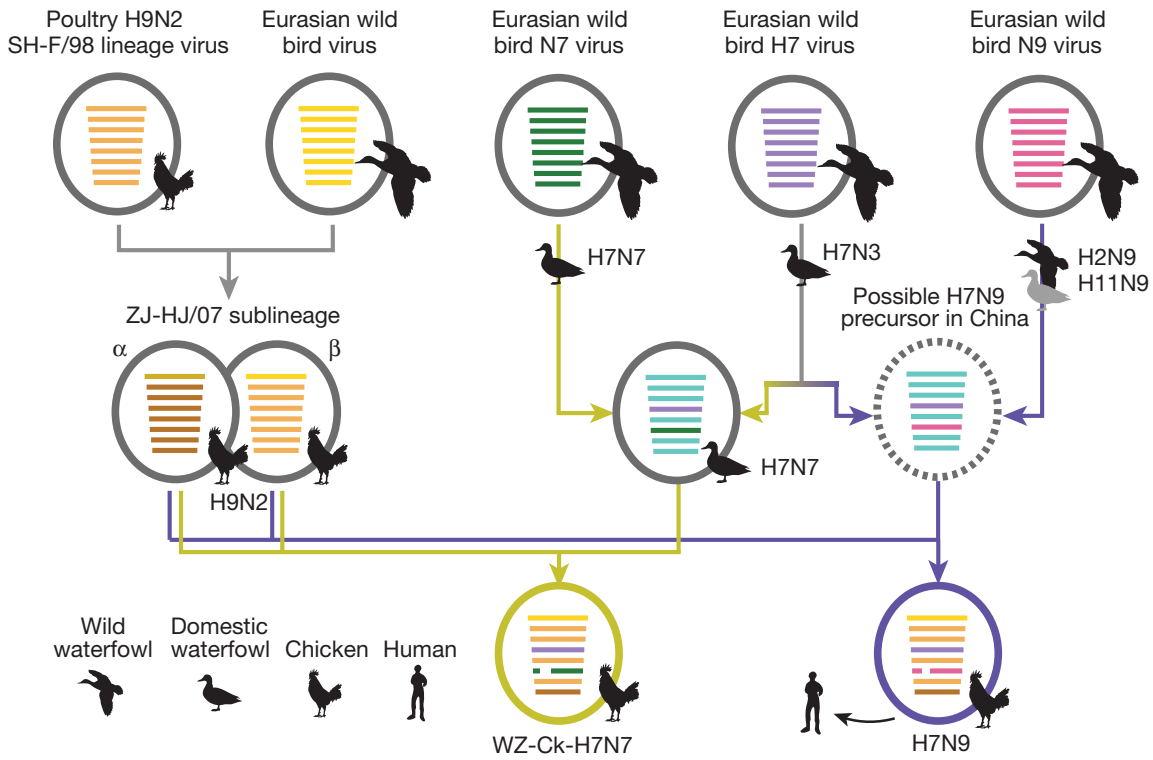


Figure 2 | Evolutionary pathways of the H7N9 and H7N7 viruses. Virus particles are represented by coloured ovals containing horizontal bars that represent the eight gene segments (from top to bottom: PB2, PB1, polymerase acidic, haemagglutinin, nucleoprotein, neuraminidase, matrix and non-structural). Segments in descendent viruses are coloured according to their corresponding source viruses (top) to illustrate gene ancestry through reassortment events. Source viruses for a reassortment are adjacent to arrow tails; arrowheads point to the resulting reassortants. Bars coloured cyan indicate gene segments of the ZJ-5 sub-lineage of wild bird viruses. A broken bar in segment 6 (neuraminidase) indicates a stalk region deletion. The virus indicated by a broken oval represents a hypothetical reassortant.

LETTER

Lam T.T.-Y. et al. *Nature*. (2015) doi:10.1038/nature14348

doi:10.1038/nature14348

Dissemination, divergence and establishment of H7N9 influenza viruses in China

Tommy Tsan-Yuk Lam[1,2,3]*, Boping Zhou[1]*, Jia Wang[2,3]*, Yujuan Chai[2,3]*, Yongyi Shen[2,3]*, Xinchun Chen[1]*, Chi Ma[2,3], Wenshan Hong[2], Yin Chen[4], Yanjun Zhang[4], Lian Duan[1,2,3], Peiwen Chen[1,2], Junfei Jiang[1,3], Yu Zhang[2,3], Lifeng Li[2,3], Leo Lit Man Poon[1,3], Richard J. Webby[5], David K. Smith[2,3], Gabriel M. Leung[3], Joseph S. M. Peiris[1,3], Edward C. Holmes[6], Yi Guan[1,2,3] & Huachen Zhu[1,2,3]

Since 2013 the occurrence of human infections by a novel avian H7N9 influenza virus in China has demonstrated the continuing threat posed by zoonotic pathogens[1,2]. Although the first outbreak wave that was centred on eastern China was seemingly averted, human infections recurred in October 2013 (refs 3–7). It is unclear how the H7N9 virus re-emerged and how it will develop further; potentially it may become a long-term threat to public health. Here we show that H7N9 viruses have spread from eastern to southern China and become persistent in chickens, which has led to the establishment of multiple regionally distinct lineages with different reassortant genotypes. Repeated introductions of viruses from Zhejiang to other provinces and the presence of H7N9 viruses at live poultry markets have fuelled the recurrence of human infections. This rapid expansion of the geographical distribution and genetic diversity of the H7N9 viruses poses a direct challenge to current disease control systems. Our results also suggest that H7N9 viruses have become enzootic in China and may spread beyond the region, following the pattern previously observed with H5N1 and H9N2 influenza viruses[8,9].

The second wave of the H7N9 outbreak that begun in late 2013 has resulted in 318 human cases and over a hundred deaths as of 12 September 2014 (ref. 7), more than twice that of the first wave. Guangdong, which had no reported human infection in the first wave, and Zhejiang have reported the highest numbers of human cases in the second wave[7]. We used influenza surveillance at live poultry markets (LPMs) in Zhejiang, Guangdong, Jiangxi, Jiangsu and Shandong provinces, at specific times or routinely, from October 2013 to July 2014 (Extended Data Tables 1 and 2), and at hospitals in Shenzhen (Guangdong) from December 2013 to April 2014, to trace the evolution and spread of the second wave of the H7N9 outbreak.

Active surveillance in fifteen cities across these five provinces identified 493 H7N9 viruses from oropharyngeal swabs of market chickens, with an average isolation rate of 3.0% (Extended Data Table 1 and Fig. 1). Only five H7N9 viruses were isolated from 2,465 cloacal swabs sampled in chickens in Jiangxi and Guangdong, giving an isolation rate of 0.2%. No H7N9 virus was isolated from domestic ducks during this survey (Extended Data Table 2). These findings highlight that market

LETTER

Lam T.T.-Y. et al. ***Nature***. (2015) doi:10.1038/nature14348

doi:10.1038/nature14348

Dissemination, divergence and establishment of H7N9 influenza viruses in China

Tommy Tsan-Yuk Lam[1,2,3]*, Boping Zhou[1]*, Jia Wang[2,3]*, Yujuan Chai[2,3]*, Yongyi Shen[2,3]*, Xinchun Chen[1]*, Chi Ma[2,3], Wenshan Hong[2], Yin Chen[4], Yanjun Zhang[4], Lian Duan[1,2,3], Peiwen Chen[1,2], Junfei Jiang[1,3], Yu Zhang[2,3], Lifeng Li[2,3], Leo Lit Man Poon[1,3], Richard J. Webby[5], David K. Smith[2,3], Gabriel M. Leung[3], Joseph S. M. Peiris[1,3], Edward C. Holmes[6], Yi Guan[1,2,3] & Huachen Zhu[1,2,3]

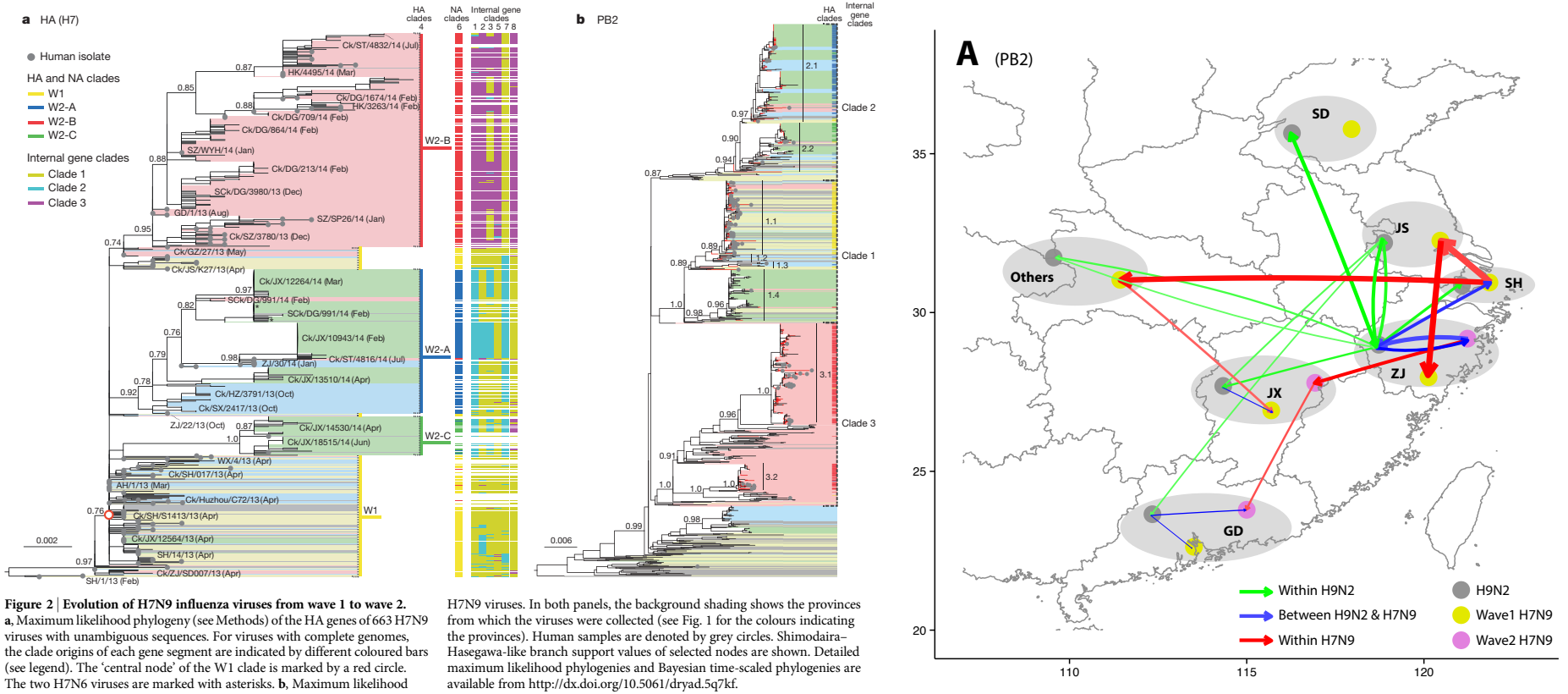


Figure 2 | Evolution of H7N9 influenza viruses from wave 1 to wave 2. **a**, Maximum likelihood phylogeny (see Methods) of the HA genes of 663 H7N9 viruses with unambiguous sequences. For viruses with complete genomes, the clade origins of each gene segment are indicated by different coloured bars (see legend). The 'central node' of the W1 clade is marked by a red circle. The two H7N6 viruses are marked with asterisks. **b**, Maximum likelihood phylogeny of 1,060 ZJ-HJ/07 sub-lineage PB2 genes. Red branches indicate H7N9 viruses. In both panels, the background shading shows the provinces from which the viruses were collected (see Fig. 1 for the colours indicating the provinces). Human samples are denoted by grey circles. Shimodaira–Hasegawa-like branch support values of selected nodes are shown. Detailed maximum likelihood phylogenies and Bayesian time-scaled phylogenies are available from http://dx.doi.org/10.5061/dryad.5q7kf.

Metagenomic studies
of microbiota

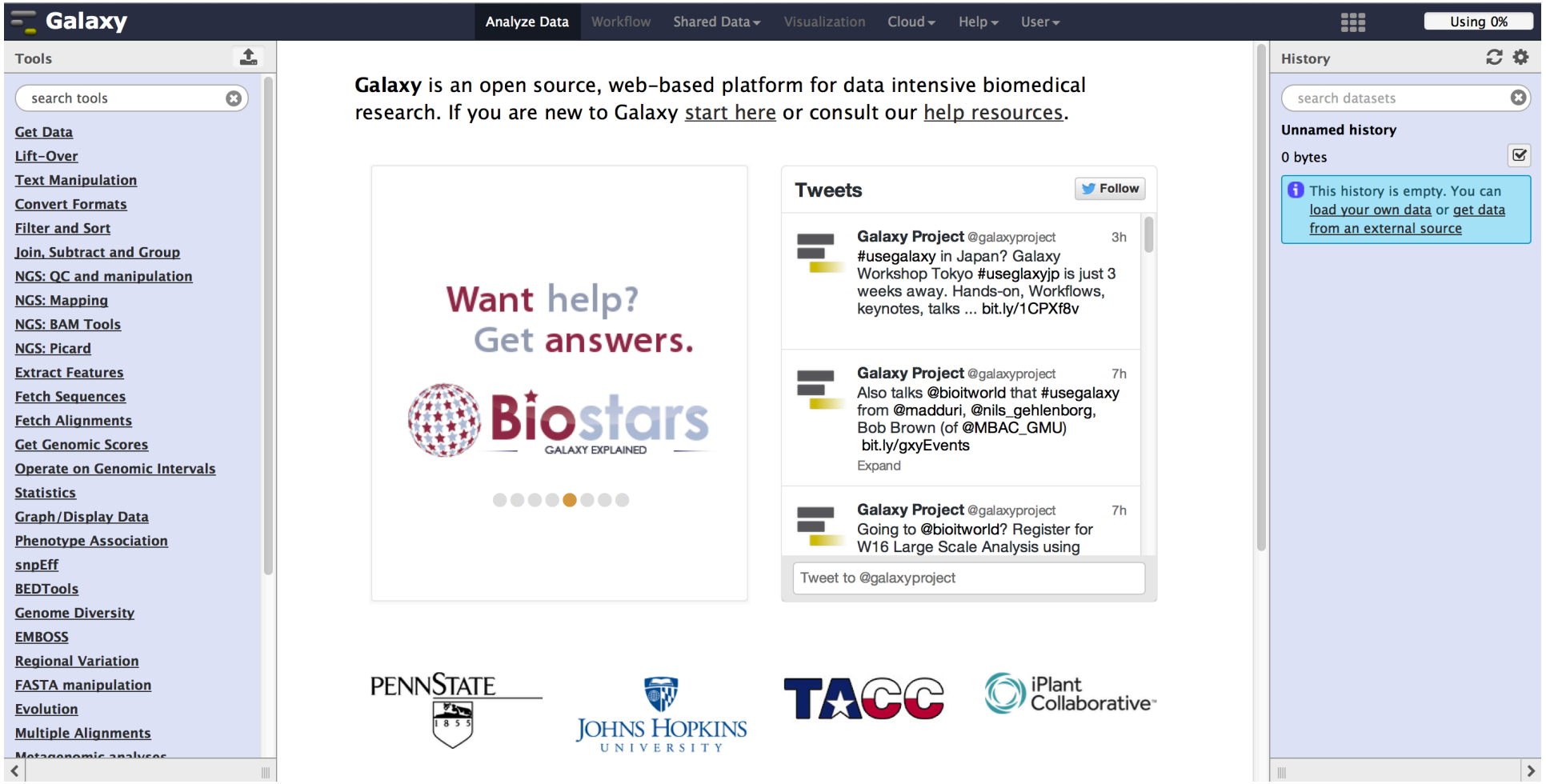
Phylogenomic studies
of organismal evolution

Phylogenomic studies
of pathogen transmission and emergence

Web-based platform ‘Galaxy’

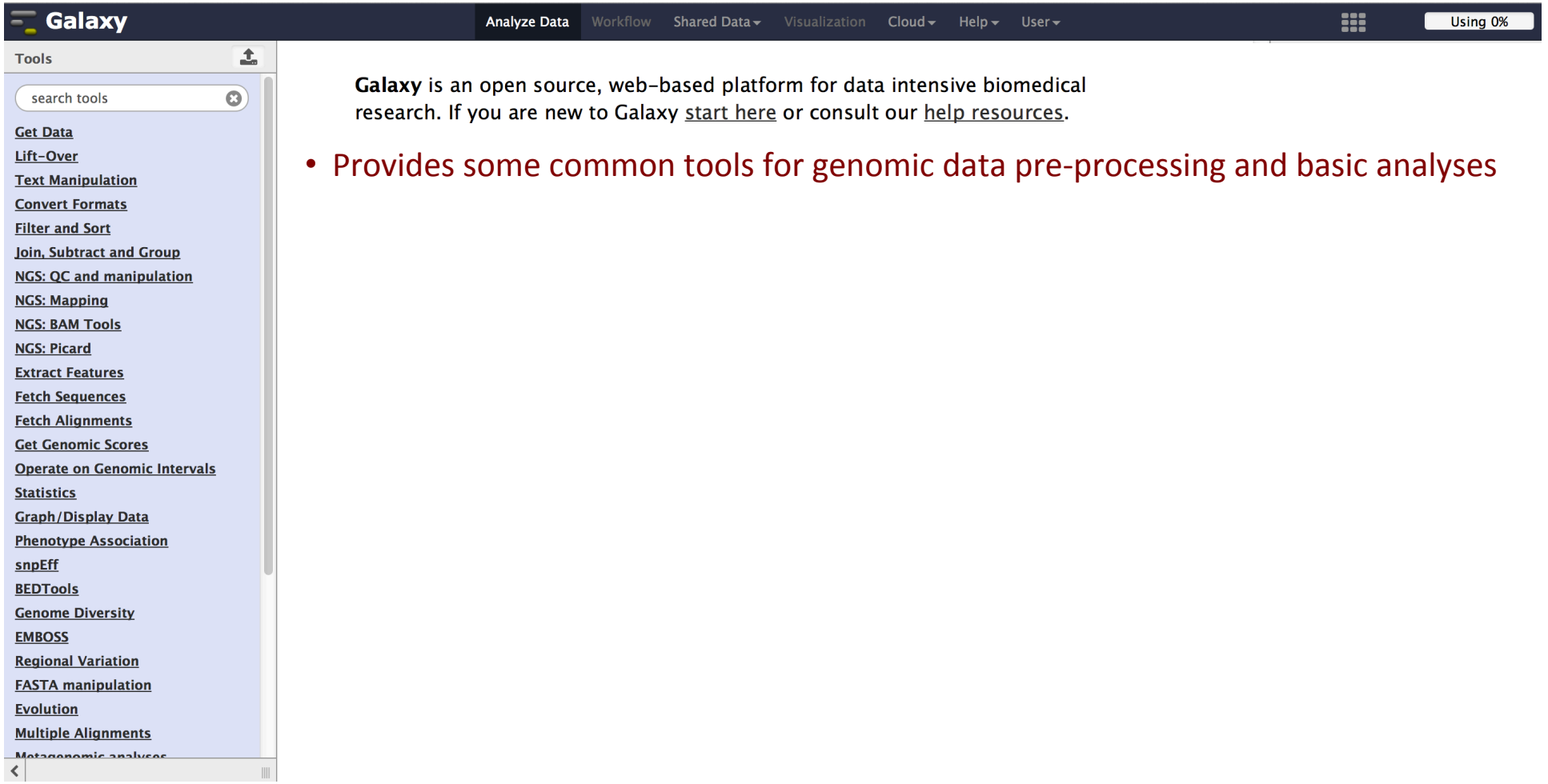
Galaxy

http://galaxyproject.org



Galaxy

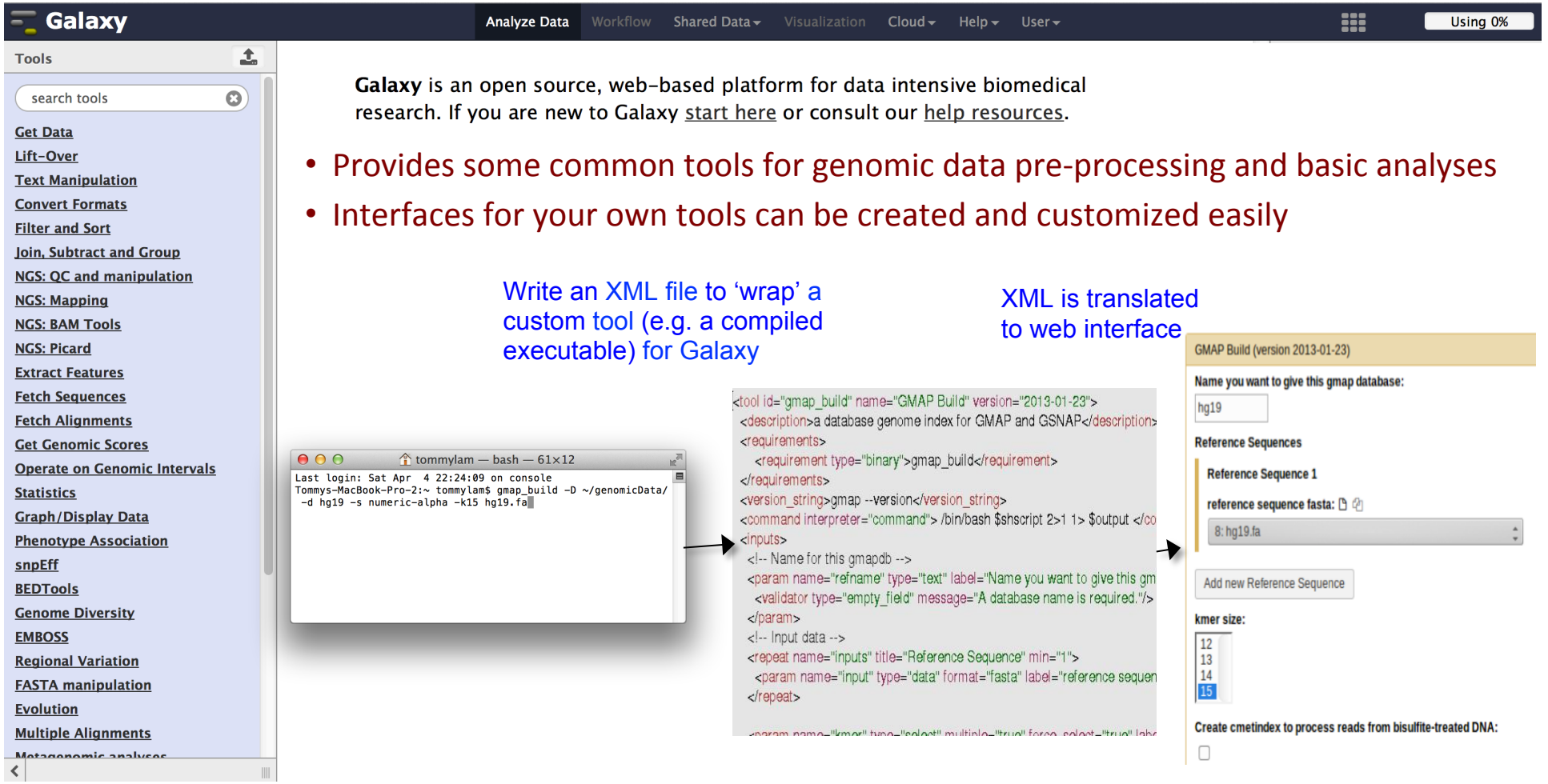
http://galaxyproject.org



- Provides some common tools for genomic data pre-processing and basic analyses

Galaxy

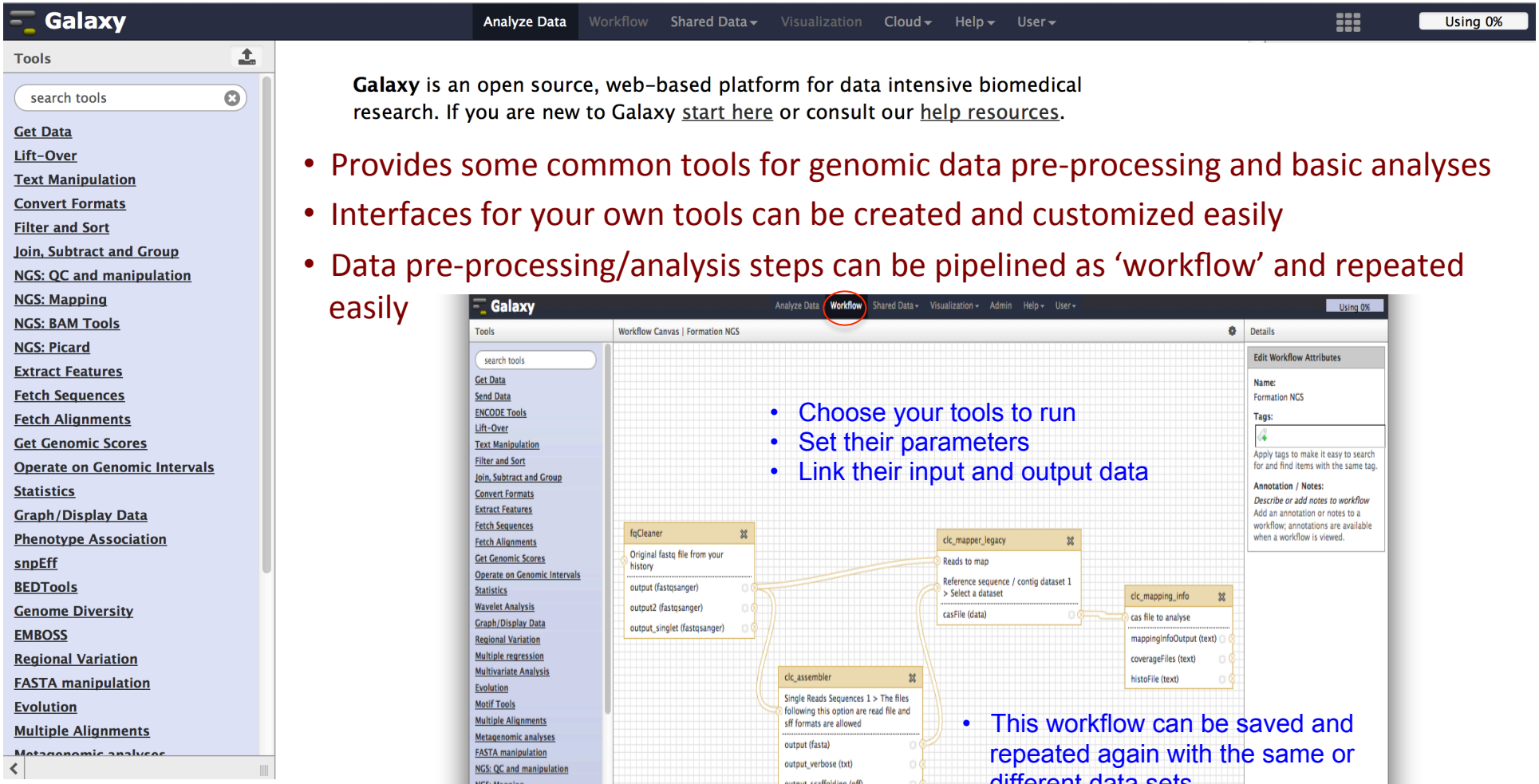
http://galaxyproject.org



- Provides some common tools for genomic data pre-processing and basic analyses
- Interfaces for your own tools can be created and customized easily

Galaxy

http://galaxyproject.org



- Provides some common tools for genomic data pre-processing and basic analyses
- Interfaces for your own tools can be created and customized easily
- Data pre-processing/analysis steps can be pipelined as ‘workflow’ and repeated easily

Galaxy

http://galaxyproject.org

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User | Using 0%

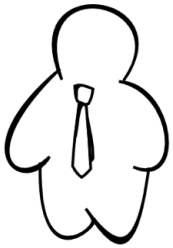
Tools

search tools

- Get Data
- Lift-Over
- Text Manipulation
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
- NGS: QC and manipulation
- NGS: Mapping
- NGS: BAM Tools
- NGS: Picard
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Phenotype Association
- snpEff
- BEDTools
- Genome Diversity
- EMBOSS
- Regional Variation
- FASTA manipulation
- Evolution
- Multiple Alignments
- Metagenomic analyses

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

- Provides some common tools for genomic data pre-processing and basic analyses
- Interfaces for your own tools can be created and customized easily
- Data pre-processing/analysis steps can be pipelined as 'workflow' and repeated easily
- Galaxy can be installed locally on single machine or cloud intra-structure



CloudMan

Contents

1. About Galaxy on the cloud
2. Instantiating a Galaxy instance on the Amazon cloud
3. Detailed steps
4. Galaxy AMIs
5. Determining the size of your cloud cluster
6. Customizing your cloud cluster
7. Notes
8. Presentations
9. Publications

Note: There are several choices for using Galaxy. This page describes installing Galaxy on a *cloud infrastructure* using CloudMan (see below). For other options, see Choices and Cloud.

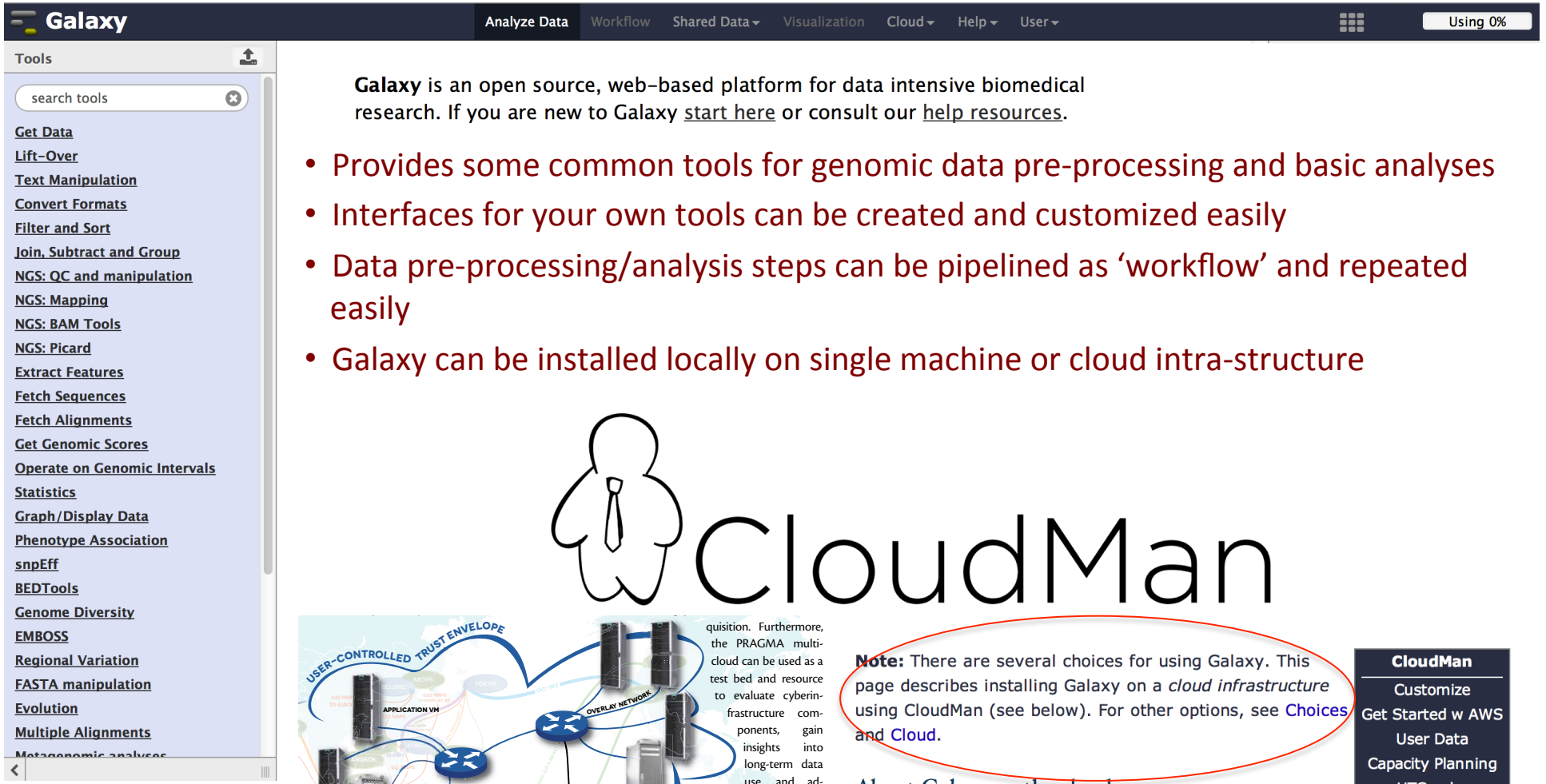
CloudMan
Customize
Get Started w AWS
User Data
Capacity Planning
HTCondor
Hadoop

About Galaxy on the cloud

With sporadic availability of data, individuals and labs may have a need to, over a period of time, process greatly variable amounts of data. Such variability in data volume imposes variable requirements on availability of compute resources used to process given data. Rather than having to purchase and maintain desired compute resources or

Galaxy

http://galaxyproject.org



- Provides some common tools for genomic data pre-processing and basic analyses
- Interfaces for your own tools can be created and customized easily
- Data pre-processing/analysis steps can be pipelined as 'workflow' and repeated easily
- Galaxy can be installed locally on single machine or cloud intra-structure

quisition. Furthermore, the PRAGMA multi-cloud can be used as a test bed and resource to evaluate cyberinfrastructure components, gain insights into long-term data use, and advance science. This year we also established an Experimental Network Testbed (ENT) to allow testing of different SDN approaches, and we standardized how we author virtual clusters and how we use them in a multi-cloud environment.

Note: There are several choices for using Galaxy. This page describes installing Galaxy on a *cloud infrastructure* using CloudMan (see below). For other options, see Choices and Cloud.

CloudMan
Customize
Get Started w AWS
User Data
Capacity Planning
HTCondor
Hadoop

About Galaxy on the cloud

With sporadic availability of data, individuals and labs may have a need to, over a period of time, process greatly variable amounts of data. Such variability in data volume imposes variable requirements on availability of compute resources used to process given data. Rather than having to purchase and maintain desired compute resources or

Credits

Some figures are adapted from the photos/illustrations from the source links below:

- http://www.myvmc.com/uploads/VMC/TreatmentImages/2437_dna_450_v2.jpg
- http://www.activityvillage.co.uk/sites/default/files/images/monkeys_av2.jpg
- http://thumbs.dreamstime.com/z/walking-lemur-8490179.jpg
- http://www.eurographics.ca/uploads/postercartel_product_option.imageEnlarge/2450-0282.jpg
- http://creationrevolution.com/wp-content/uploads/2013/05/Charles-Darwin-tree-of-life-poster.jpg
- http://micro.magnet.fsu.edu/cells/viruses/influenzavirus.html
- http://i.istockimg.com/file_thumbview_approve/49943358/3/stock-photo-49943358-e-coli-bacteria.jpg
- http://advancedgraphics.com/wp-content/uploads/2013/08/1485_Giant%20Panda_50.jpg
- http://www.3plearning.com/wp-content/uploads/2013/08/Screen-Shot-2013-08-05-at-4.15.41-PM-600x241.png
- http://advancedgraphics.com/wp-content/uploads/2013/08/1487_Chimpanzee_28.jpg
- http://www.animalleague.org/assets/images/animal_cutouts/standing.jpg
- http://nobacks.com/wp-content/uploads/2014/11/Chicken-27-280x500.png
- http://jb.asm.org/content/194/1/176/F4.large.jpg
- http://www.geeksnack.com/wp-content/uploads/2014/08/phytoplankton-international-space-station.jpg
- http://static01.nyt.com/images/2013/05/19/magazine/19microbiome3/19microbiome3-videoLarge-v2.jpg
- http://fc00.deviantart.net/fs70/f/2013/136/2/7/commision_for_dna_projecten_bv_version_2_1__logo__by_bastiaandegoede-d65fwi5.png
- http://41.media.tumblr.com/tumblr_ma0cx9aNH21rv4l4do1_1280.jpg
- http://www.oceanrecov.org/assets/components/phpthumbof/cache/b3195fefc66004c30e03da61b960d98d.db635bd86fb20fc073849b0738f5f634.jpg
- http://www.azerb.com/azer428.jpg
- http://www.watersheddiscovery.ca/wp-content/uploads/2012/03/Metagenomics-hotspring-no-text.jpg
- http://upload.wikimedia.org/wikipedia/commons/thumb/5/5f/Desulfovibrio_desulfuricans.jpg/250px-Desulfovibrio_desulfuricans.jpg
- http://www.thermapure.com/wp-content/uploads/2011/05/e_coli.jpg
- http://classconnection.s3.amazonaws.com/70/flashcards/274070/jpg/actinobacteria1307399189425.jpg
- http://media.eol.org/content/2009/11/25/03/29206_580_360.jpg
- http://www.examiner.com/images/blog/EXID7150/images/Firmicutes1USA.jpg
- Lang JM, Darling AE, Eisen JA (2013) Phylogeny of Bacterial and Archaeal Genomes Using Conserved Genes: Supertrees and Supermatrices. PLoS ONE 8(4): e62510. doi:10.1371/journal.pone.0062510
- In the courtesy of Pasteur Paris INDA Presentation Slides

Acknowledgement

- The University of Hong Kong
  - Prof. Yi Guan
  - Prof. Malik Peiris
  - Dr. Maria Zhu
  - Dr. David Smith
  - Prof. Frederick Leung
  - Mr. Wing-Keung Kwan (HKU-CC)
- University of Oxford
  - Prof. Oliver Pybus
- The University of Sydney
  - Prof. Edward Holmes
- Funding source: Royal Society UK