CSC 334: TOPICS IN COMPUTATIONAL BIOLOGY

“Algorithms for Genomic Data”

Fall 2015

Smith College

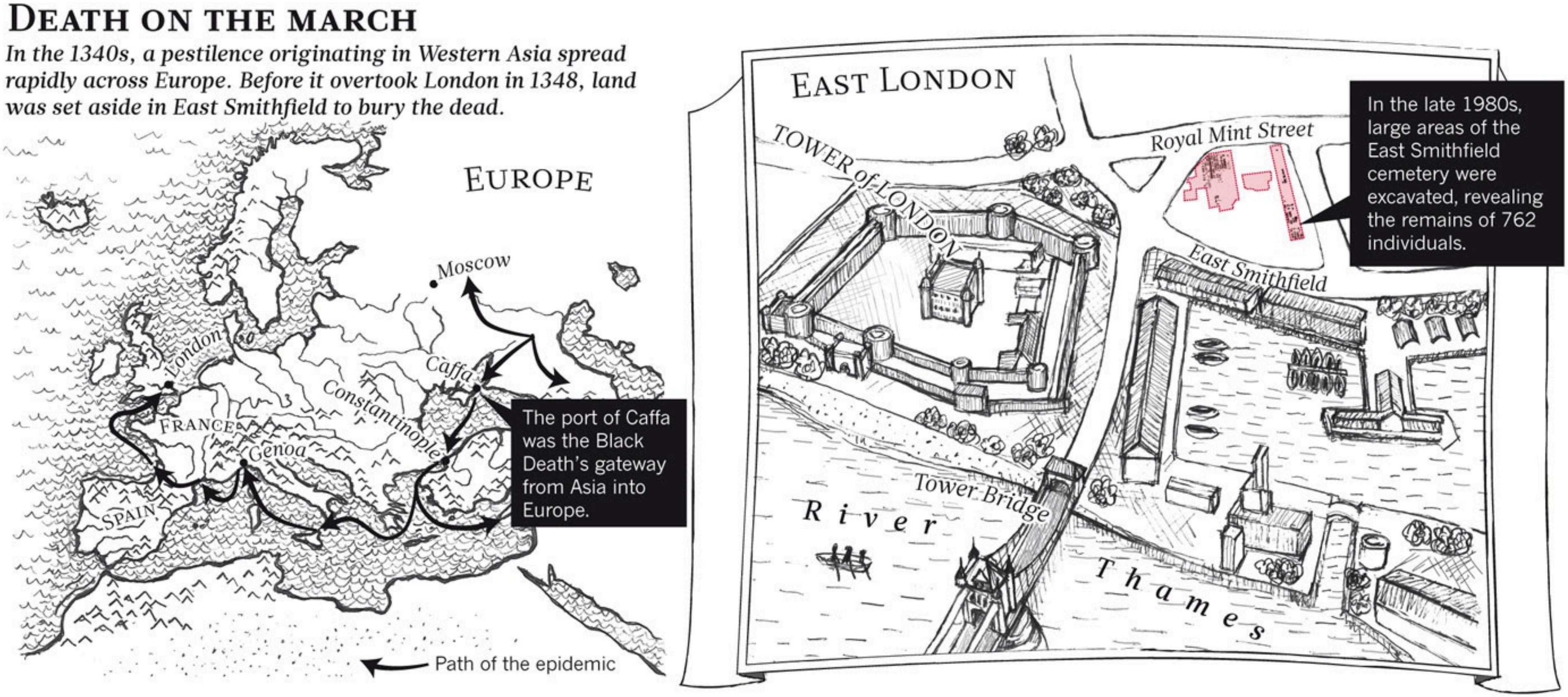
Instructor: Prof. Sara Sheehan

Outline: 10/21

- Ancestral Genome (Black Death) paper

Black Death pandemic

- Peaked in Europe in roughly 1350
- 75-200 million people died
- Spread through the silk road



DEATH ON THE MARCH

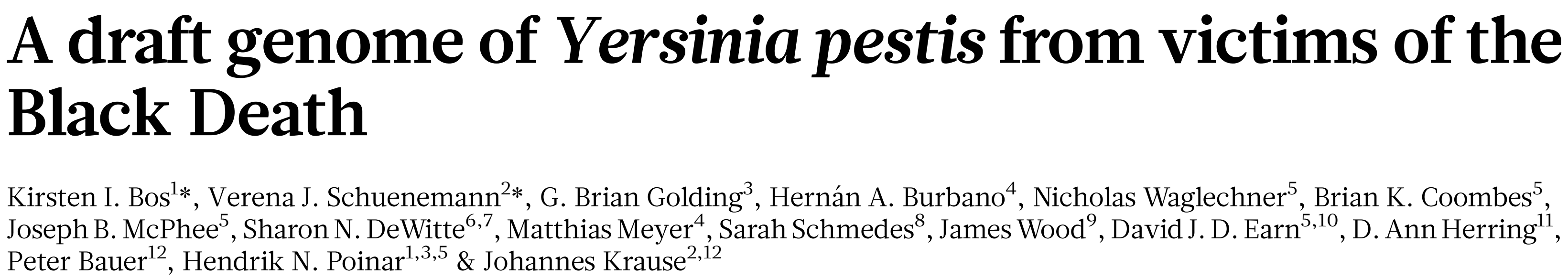
In the 1340s, a pestilence originating in Western Asia spread rapidly across Europe. Before it overtook London in 1348, land was set aside in East Smithfield to bury the dead.

The port of Caffa was the Black Death's gateway from Asia into Europe.

In the late 1980s, large areas of the East Smithfield cemetery were excavated, revealing the remains of 762 individuals.

Plague genome: The Black Death decoded

Black Death agent genomics



A draft genome of *Yersinia pestis* from victims of the Black Death

Kirsten I. Bos[1]*, Verena J. Schuenemann[2]*, G. Brian Golding[3], Hernán A. Burbano[4], Nicholas Waglechner[5], Brian K. Coombes[5], Joseph B. McPhee[5], Sharon N. DeWitte[6,7], Matthias Meyer[4], Sarah Schmedes[8], James Wood[9], David J. D. Earn[5,10], D. Ann Herring[11], Peter Bauer[12], Hendrik N. Poinar[1,3,5] & Johannes Krause[2,12]

> “Comparisons against modern genomes reveal **no unique derived positions** in the medieval organism, indicating that the perceived increased virulence of the disease during the Black Death may not have been due to bacterial phenotype. These findings support the notion that factors other than microbial genetics, such as environment, vector dynamics and host susceptibility, should be at the forefront of epidemiological discussions regarding emerging *Y. pestis* infections.”

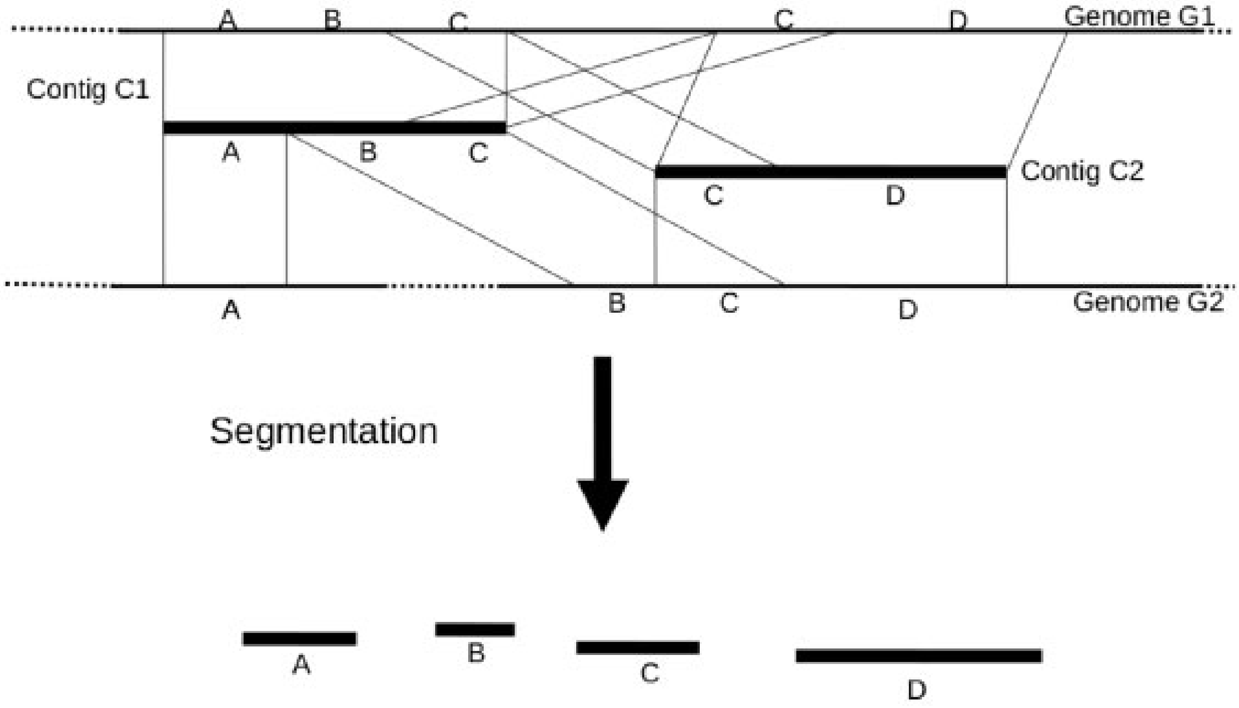


Fig. 1. Illustration of the segmentation procedure to obtain homologous families of markers. For this example, we consider two contigs $C1$ and $C2$ and their alignments on two genomes $G1$ and $G2$. Part C of $C1$ and $C2$ aligns to the same positions in both genomes, including two different positions on $G1$. Parts A and B of $C1$ align at two different positions of $G2$. So the segmentation produces four families, with non-overlapping ancestral markers A, B, C and D. For these four segments, properties (1) and (2) are satisfied, whereas both were violated for $C1$ and $C2$. The family containing segment C contains two ancestral segments, two extant segments from $G1$ and one from $G2$. According to the number of occurrences in other genomes, this family may have a multiplicity >1.

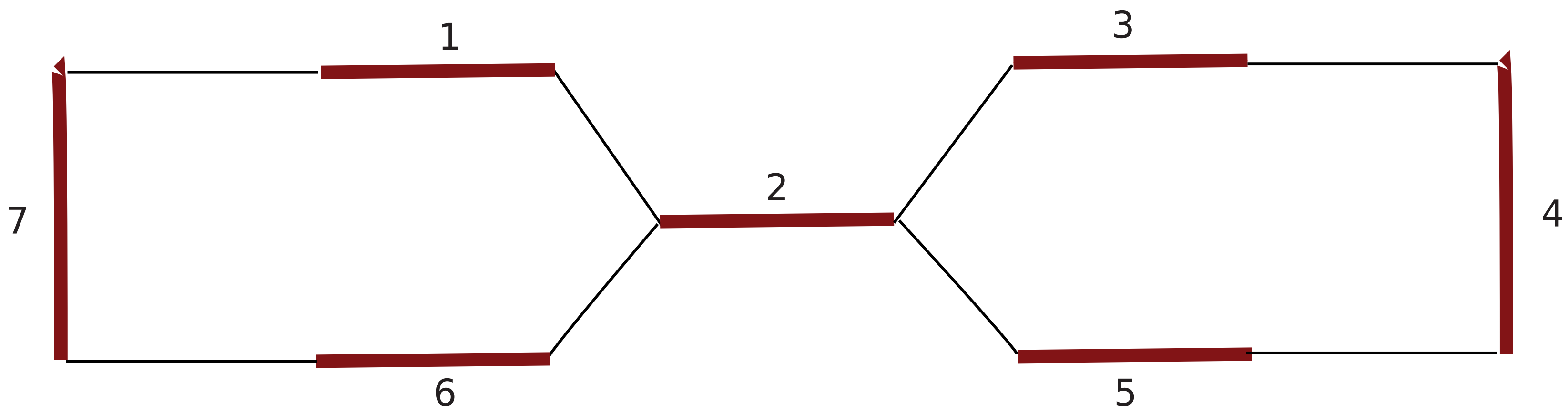


Fig. 2. Illustration of the ambiguity in ordering ancestral markers with multiplicities >1 and of the use of intervals to address it. Here is a toy example where we have markers 1, . . . , 7, drawn with bold segments, and adjacencies between their extremities, drawn with thin lines. Assume every marker has multiplicity 1 except marker 2, which has multiplicity 2. Then every marker extremity has as many adjacencies as its multiplicity predicts. But there are two possible circular orderings of these markers according to these adjacencies: 1,2,3,4,5,2,6,7, or 1,2,5,4,3,2,6,7. Suppose we have in addition repeat spanning intervals 1.2.3 and 5.2.6, then only the first ordering is compatible with them

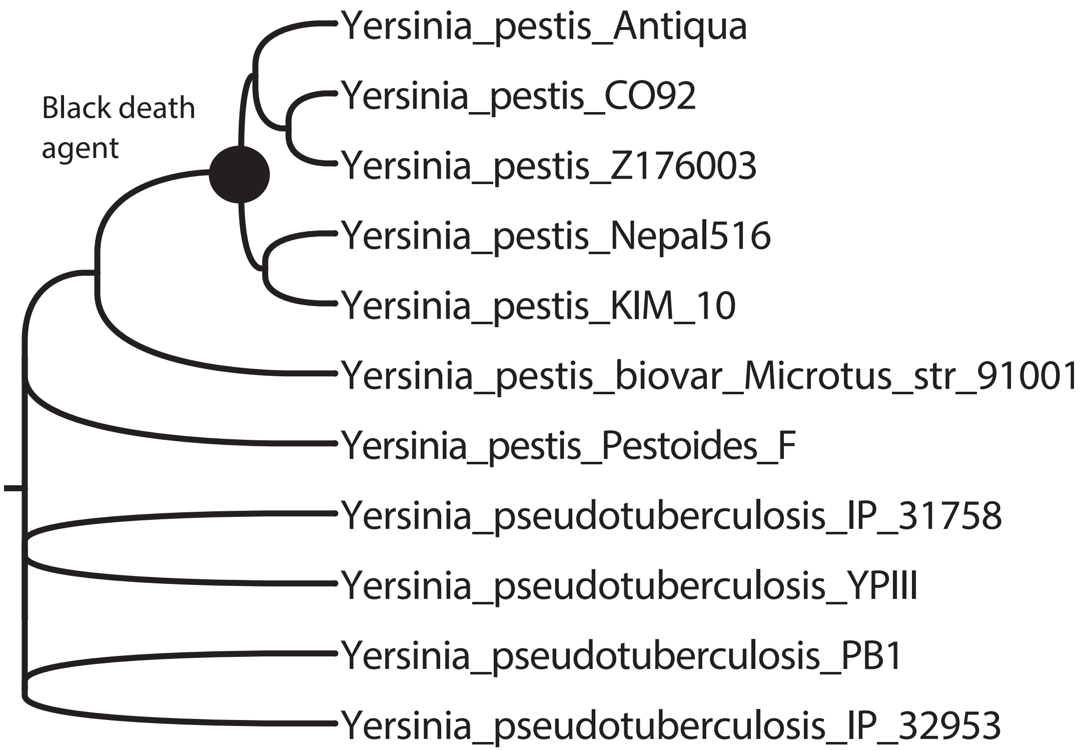
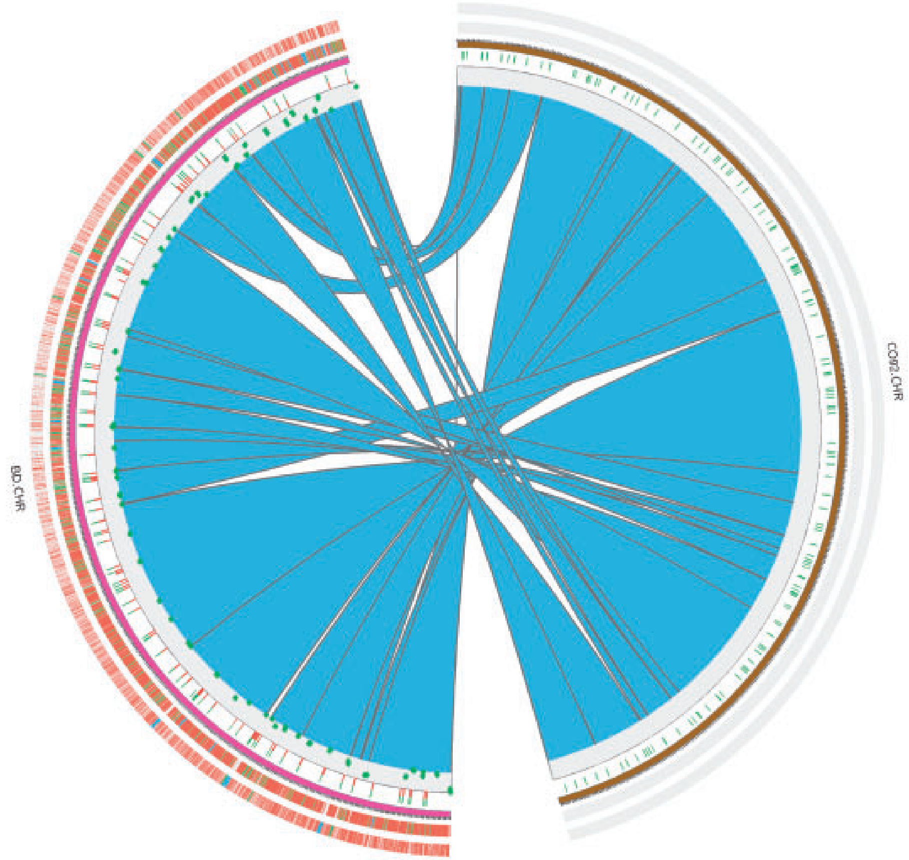

Black death agent
Yersinia_pestis_Antiqua
Yersinia_pestis_CO92
Yersinia_pestis_Z176003
Yersinia_pestis_Nepal516
Yersinia_pestis_KIM_10
Yersinia_pestis_biovar_Microtus_str_91001
Yersinia_pestis_Pestoides_F
Yersinia_pseudotuberculosis_IP_31758
Yersinia_pseudotuberculosis_YPIII
Yersinia_pseudotuberculosis_PB1
Yersinia_pseudotuberculosis_IP_32953

Fig. 3. Phylogeny of the considered genomes from Bos *et al.* (2011)

Black Death agent



Y. pestis CO92

Fig. 4. Comparison of the reconstructed Black Death agent chromosome (left) and of the *Y.pestis CO92* chromosome (right). Outside tracks of the Black Death agent chromosome represents gaps (outer track) and markers (inner track), with red (respectively green, blue) indicating small (resp. mid-length, large) elements. The first two inside tracks represent annotated (green) and inferred (green) insertion sequences. The scattered inside track represents the level of breakpoint reuse in evolutionary scenario between the ancestor and the strains *Y.pestis Antiqua*, *Y.pestis KIM10* and *Y.pestis biovar Microtus str. 91001*. Blue ribbons join colinear chromosome segments. Figure computed using Circos (Krzywinski *et al.*, 2009)

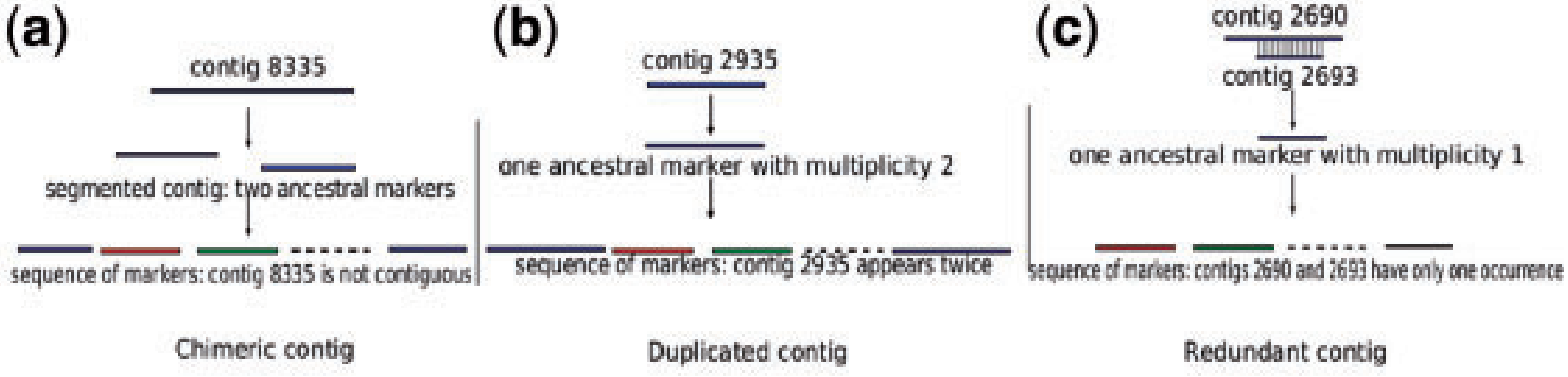

(a)
contig 8335
segmented contig: two ancestral markers
sequence of markers: contig 8335 is not contiguous
Chimeric contig
(b)
contig 2935
one ancestral marker with multiplicity 2
sequence of markers: contig 2935 appears twice
Duplicated contig
(c)
contig 2690
contig 2693
one ancestral marker with multiplicity 1
sequence of markers: contigs 2690 and 2693 have only one occurrence
Redundant contig

Fig. 5. Contig correction: (**a**) the contig is cut during the segmentation procedure, but not joined during the marker ordering; (**b**) the contig is found to have two occurrences in the marker ordering; (**c**) two contigs contain the same DNA sequence and this sequence is predicted to have only one occurrence in the marker ordering