

# **Филогения**

**Алгоритмы в биоинформатике**

**Антон Елисеев**

**[eliseevantoncoo@gmail.com](mailto:eliseevantoncoo@gmail.com)**

# Что было на прошлой лекции

- Перестановки внутри X хромосомы человека и мыши
- Хотспоты и Random Breakage Models
- Перестройки и reversal distance
- Breakpoint Graphs
- synteny block

# Что будет в этой лекции

- Распространение инфекций и эволюция
- Филогенетические деревья
- Формат newick
- UPGMA, WPGMA, Neighbor joining
- Восстановление общего предка

# Распространение инфекций



1346

1347

1348

1349

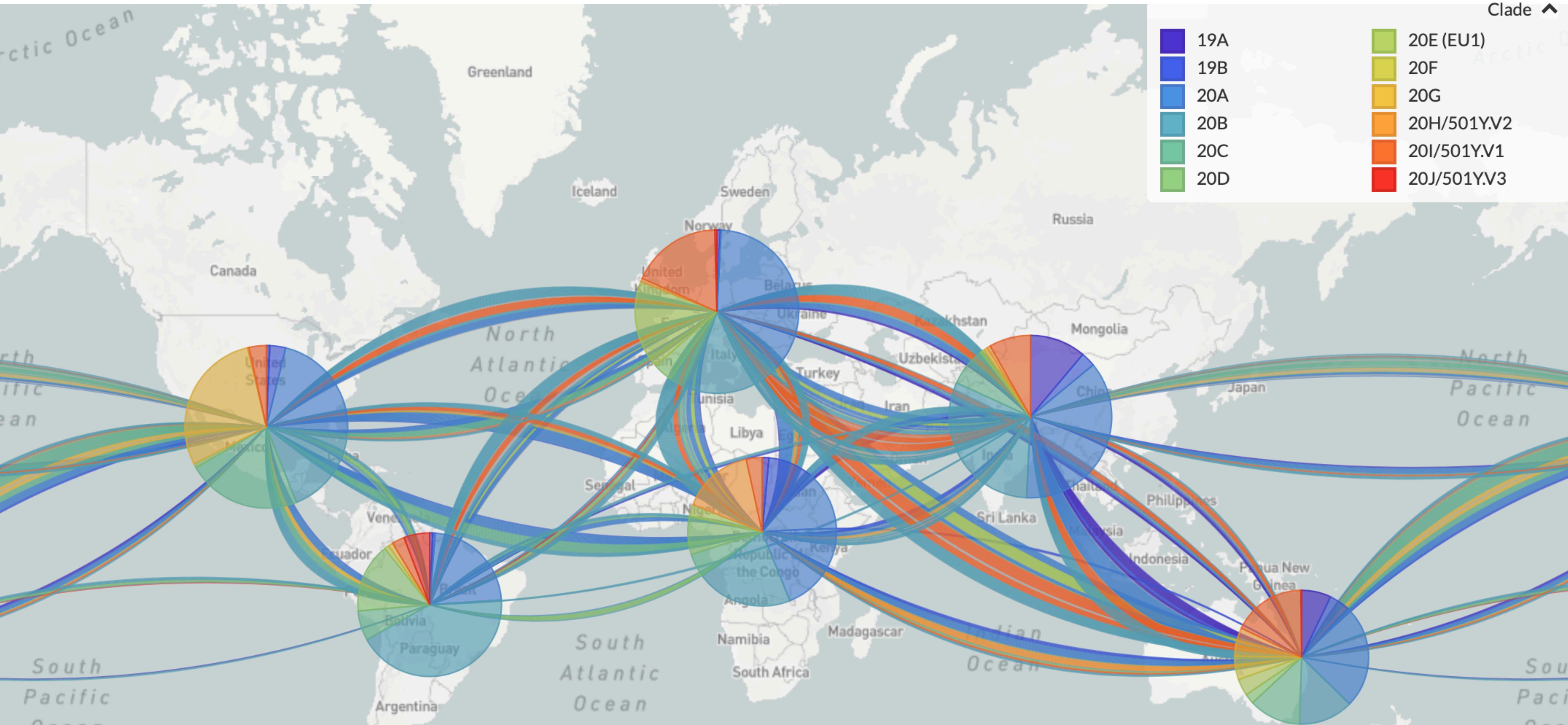
1350

1351

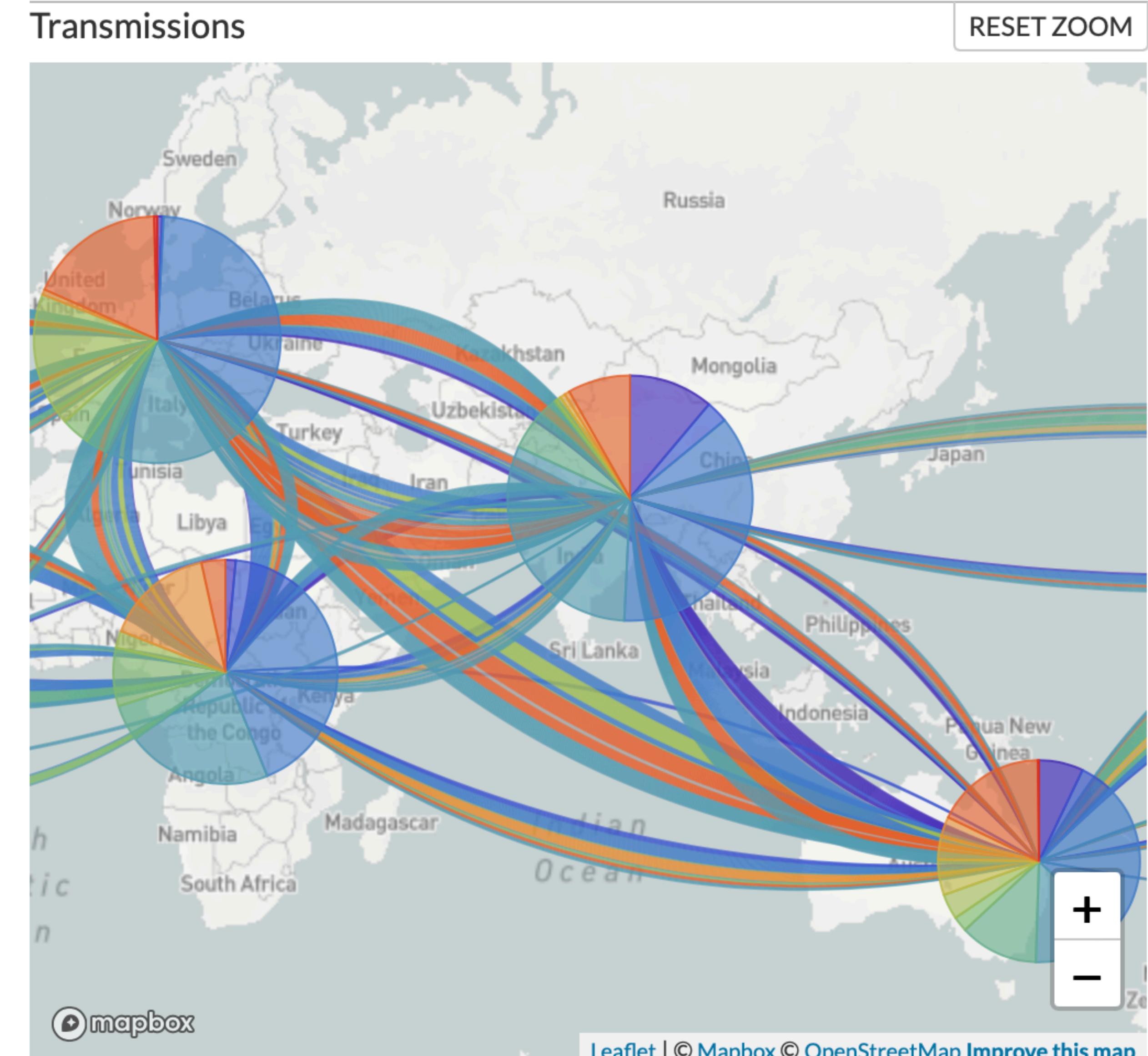
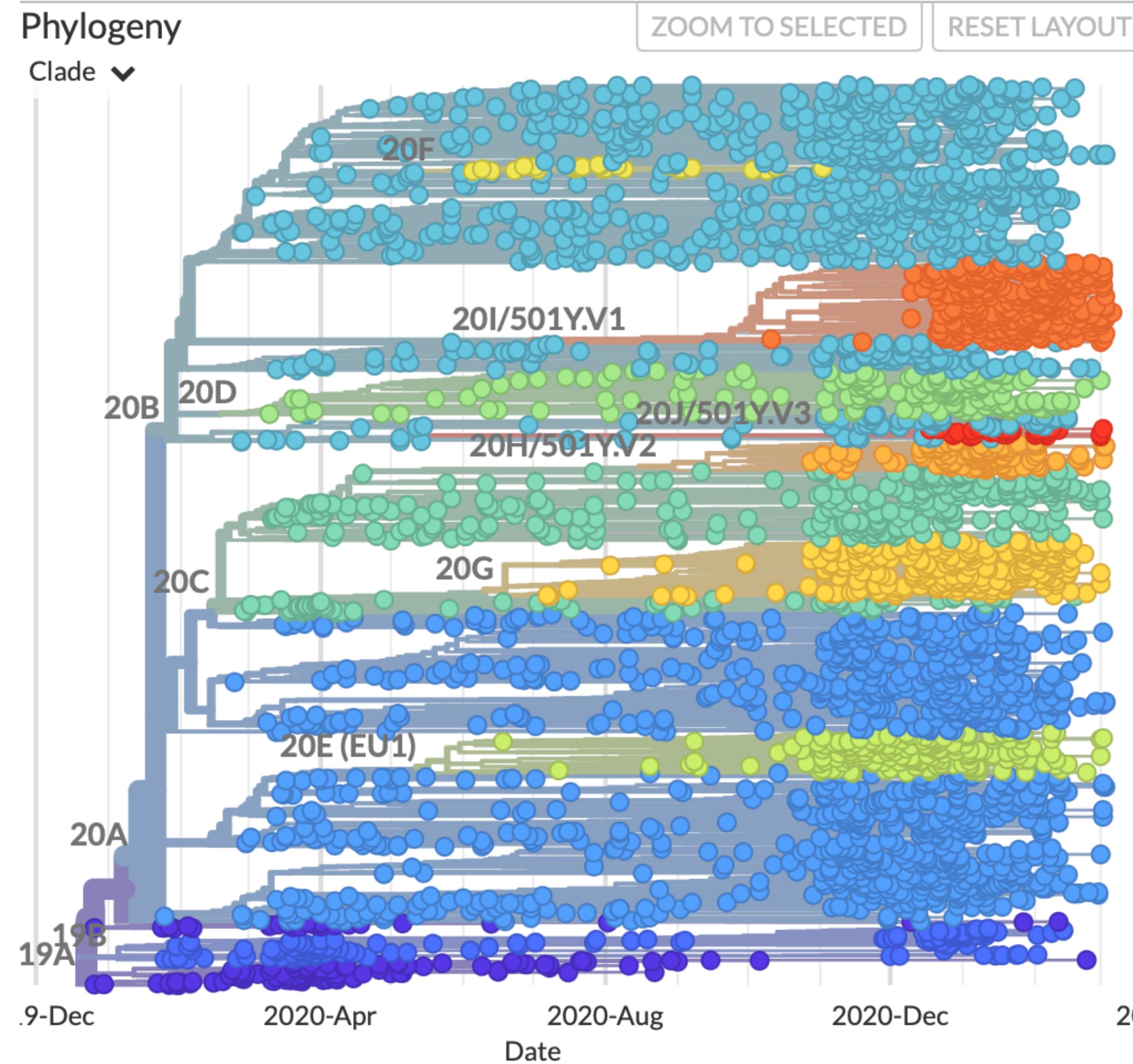
1352

1353

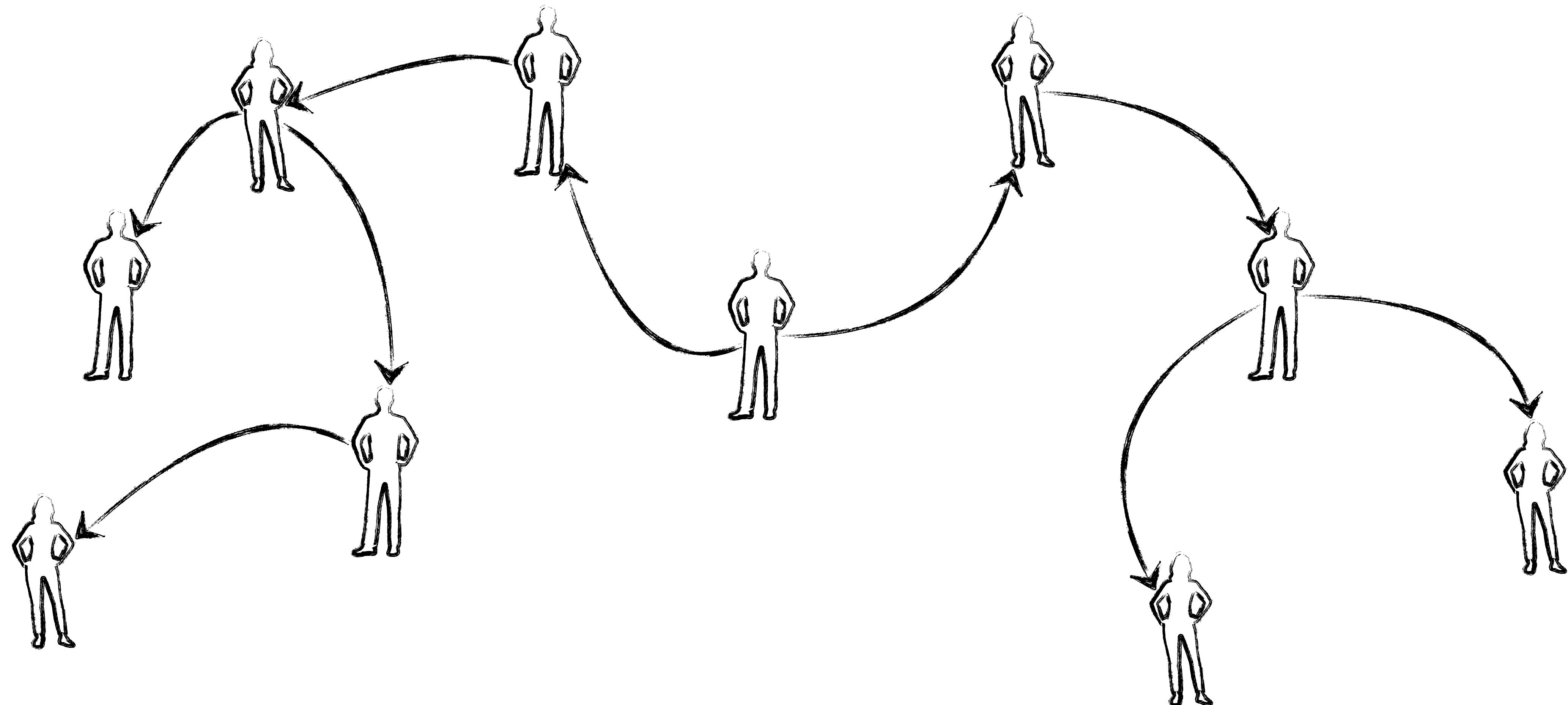
# Распространение инфекций



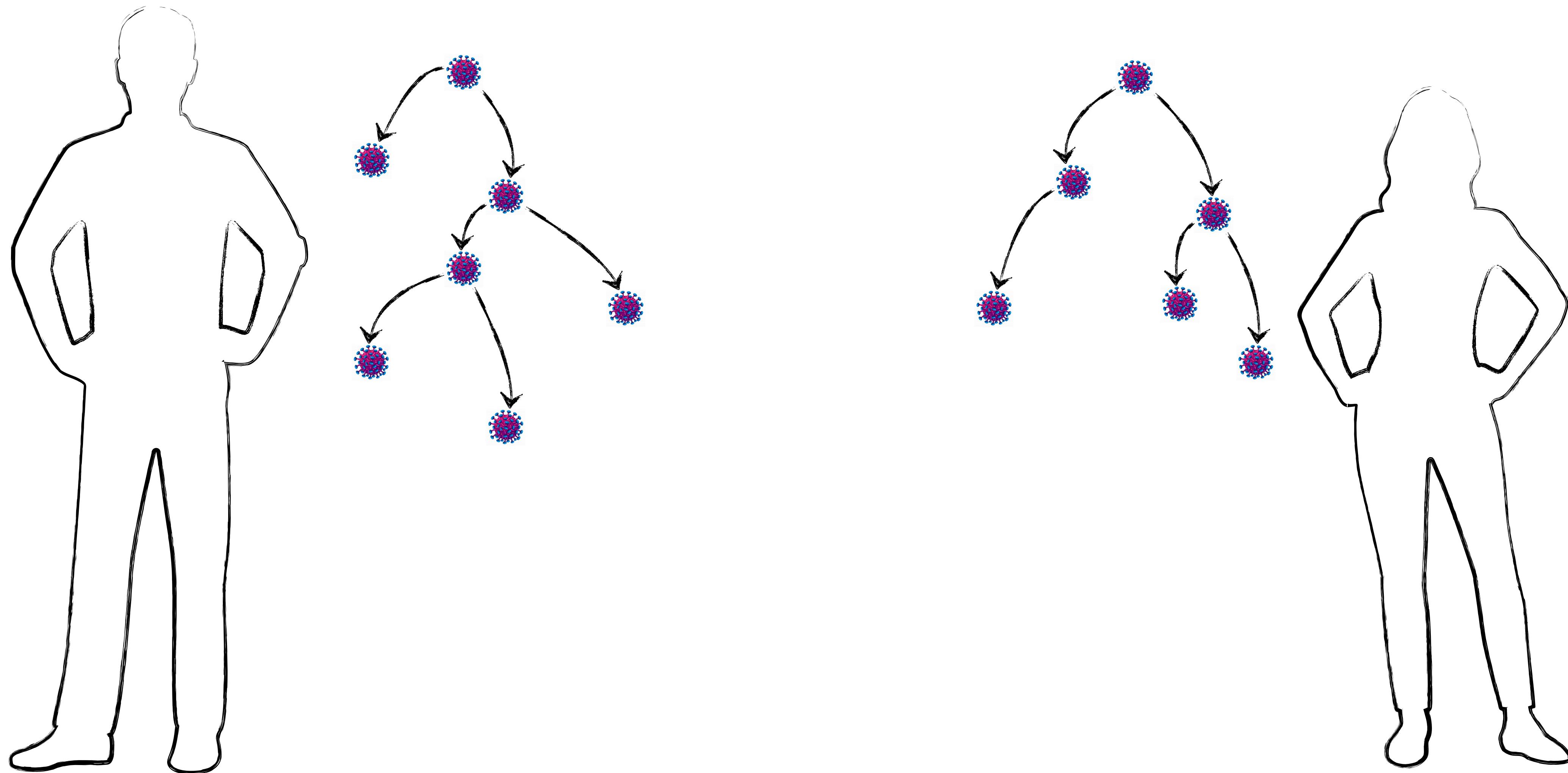
# Распространение инфекций



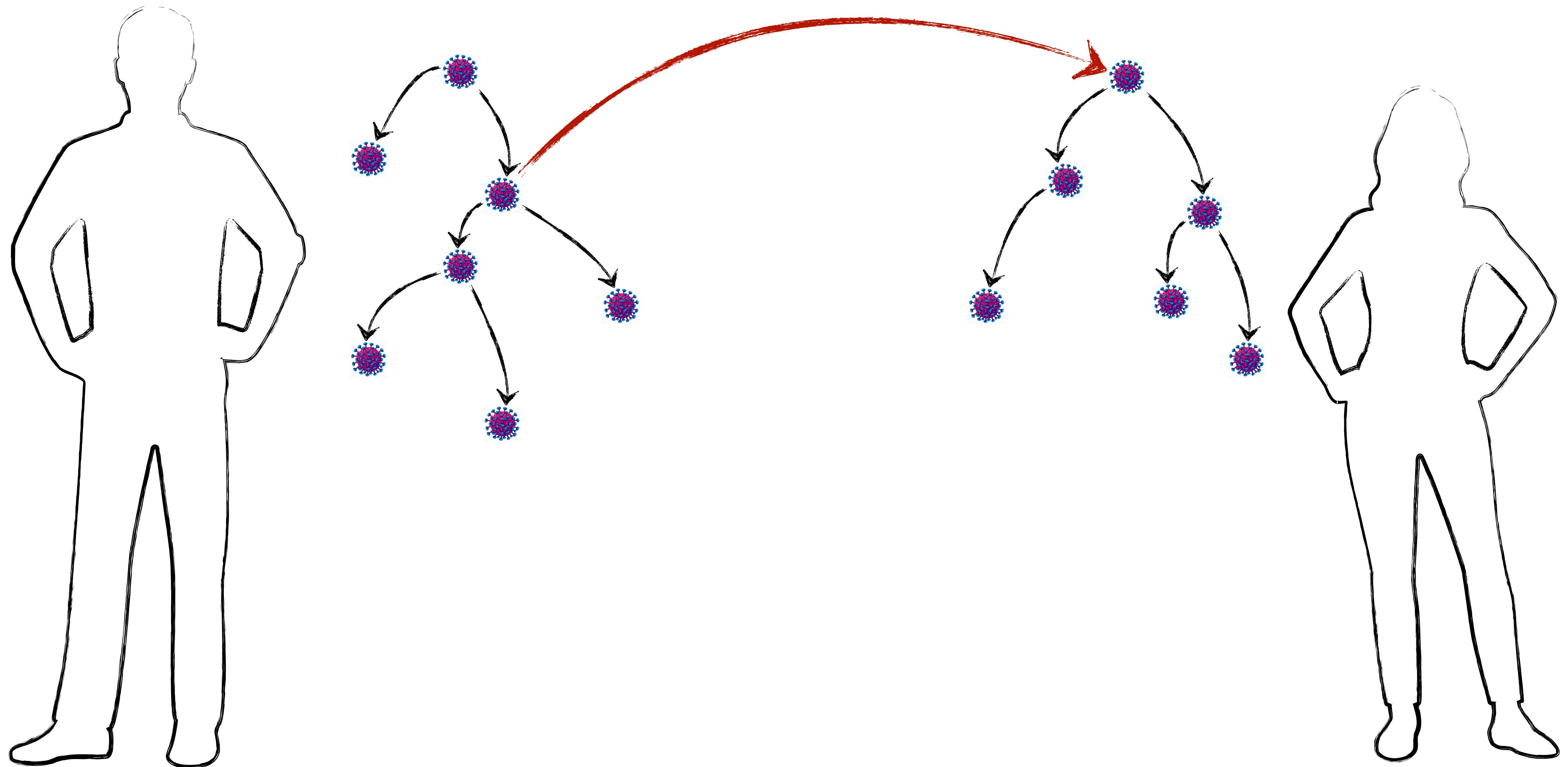
# Распространение инфекций



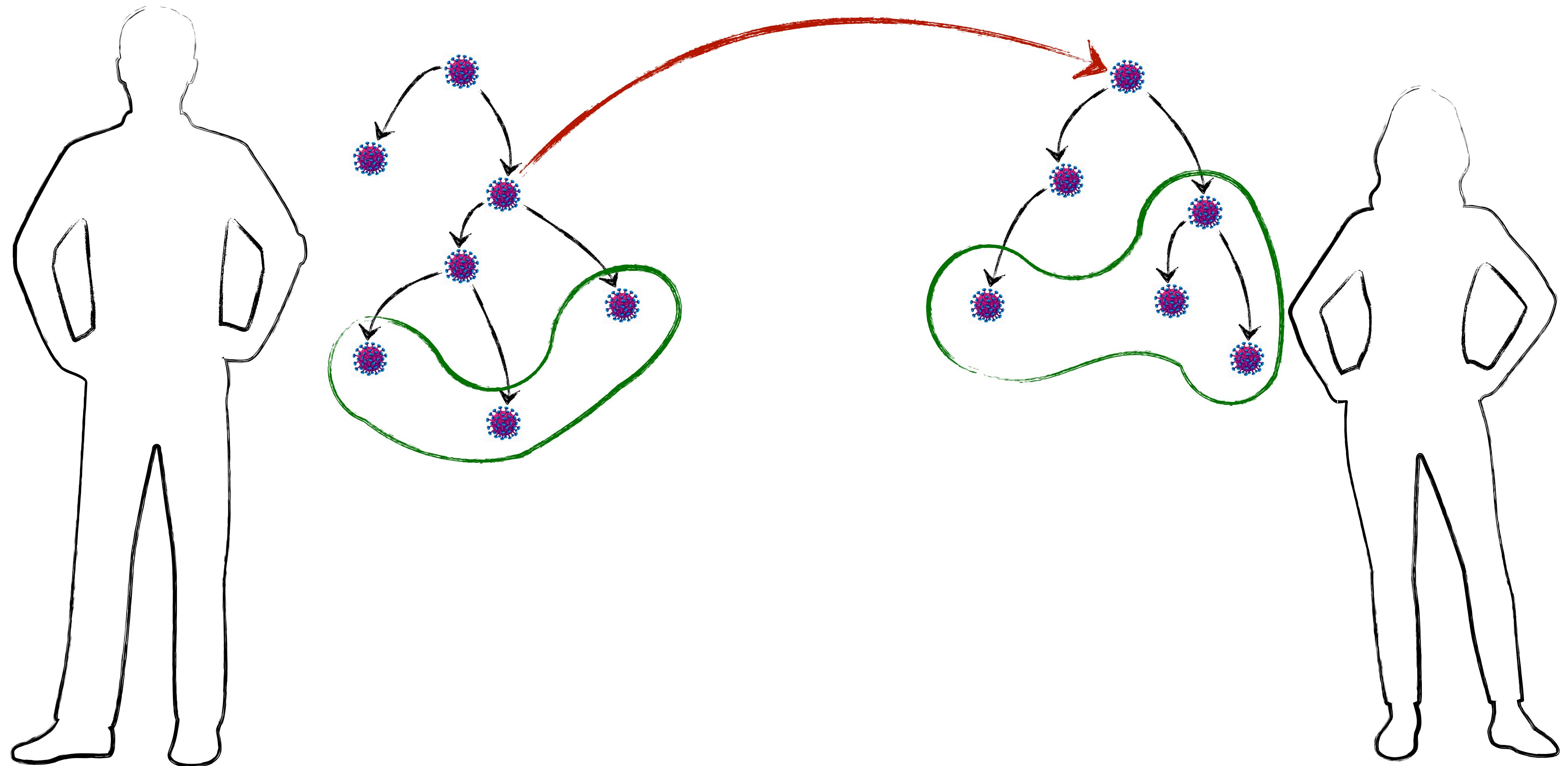
# Распространение инфекций



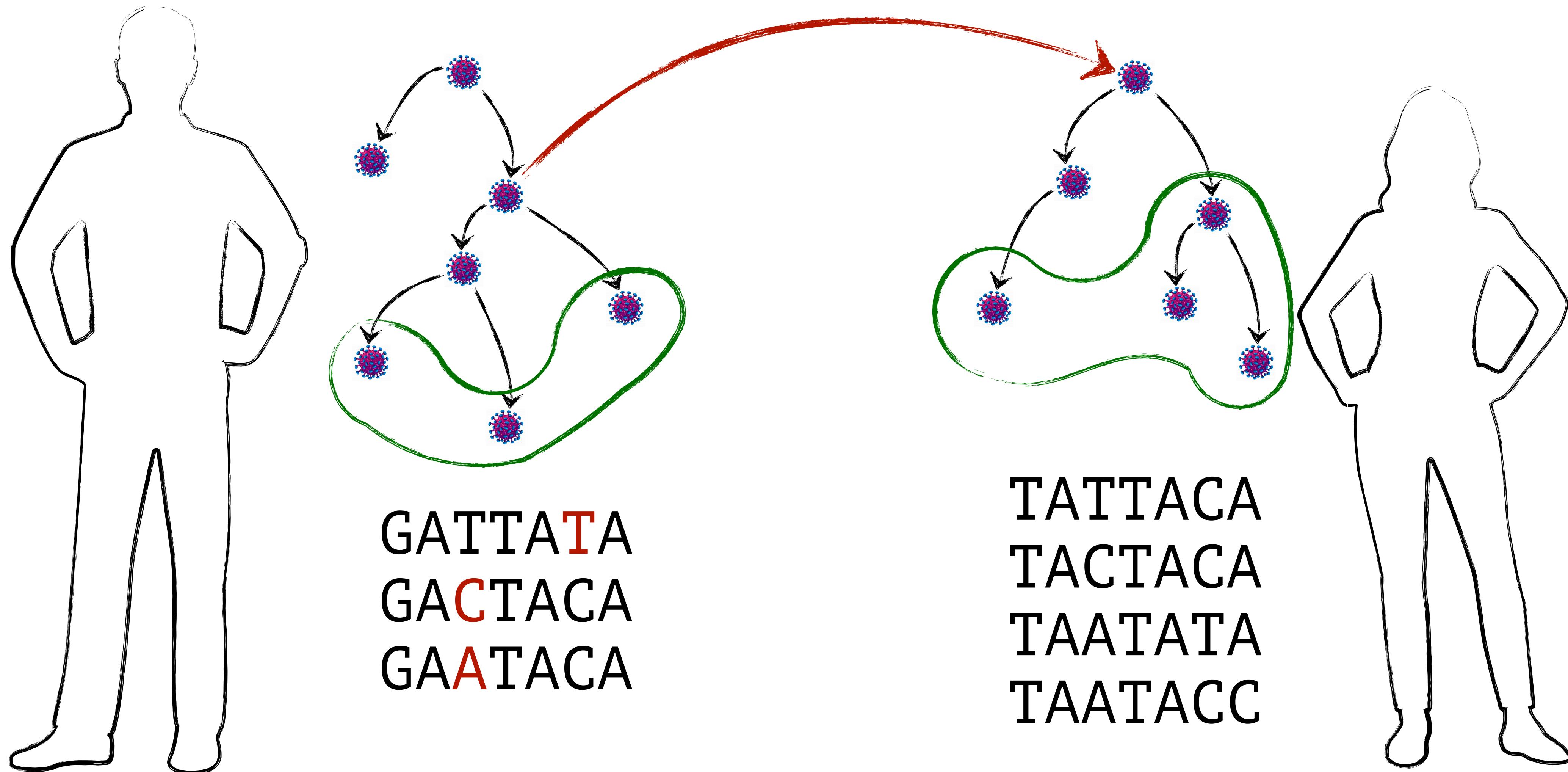
# Распространение инфекций



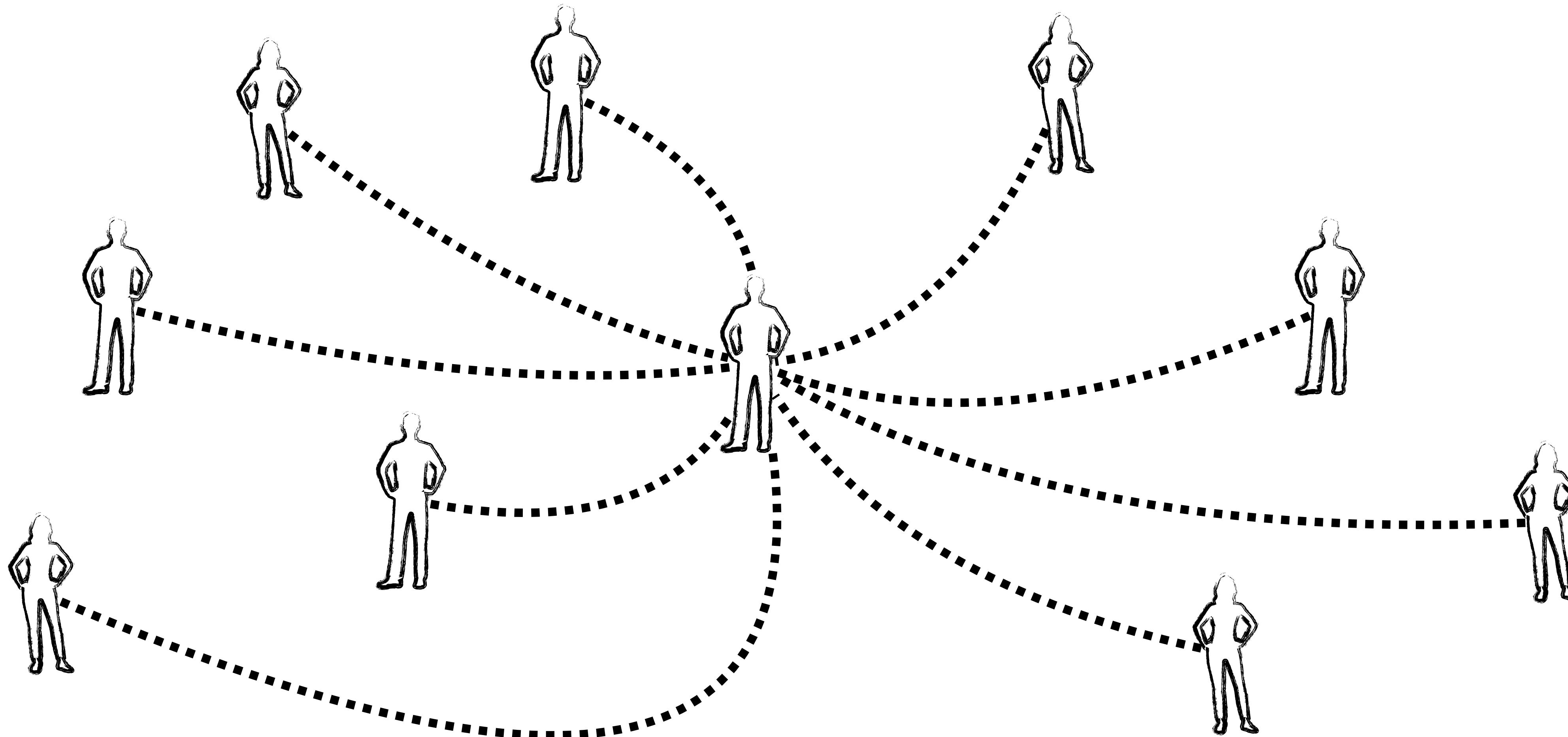
# Распространение инфекций



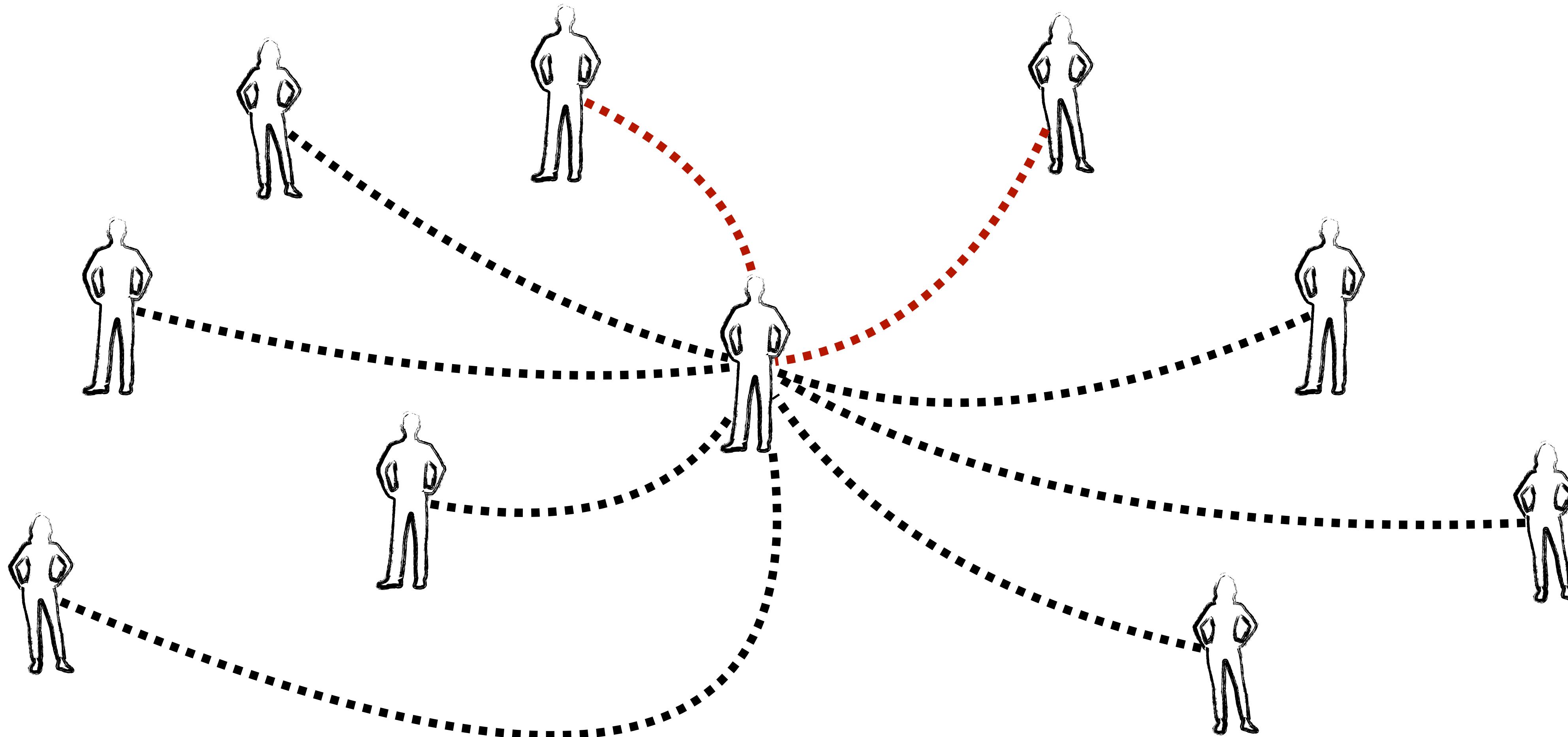
# Распространение инфекций



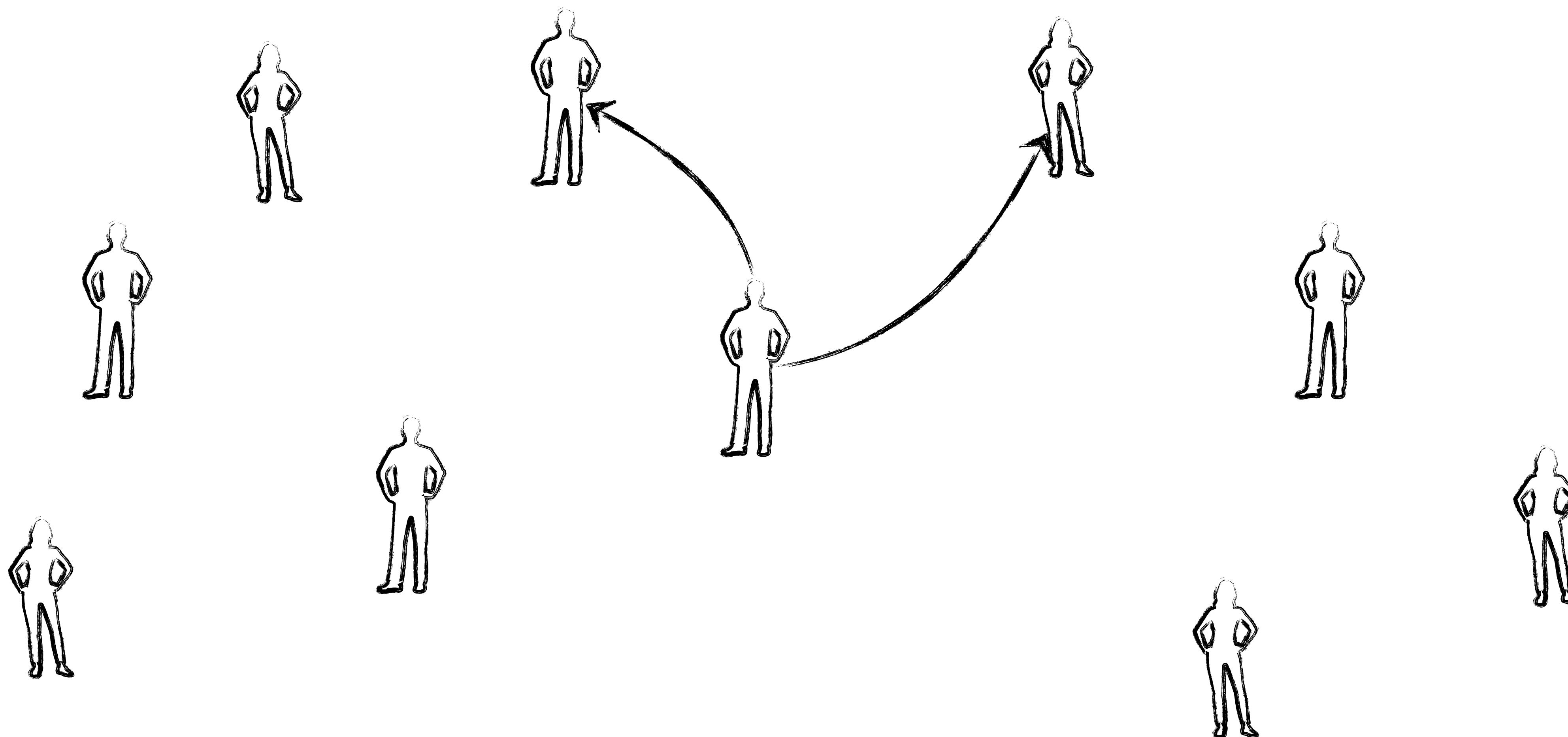
# Распространение инфекций



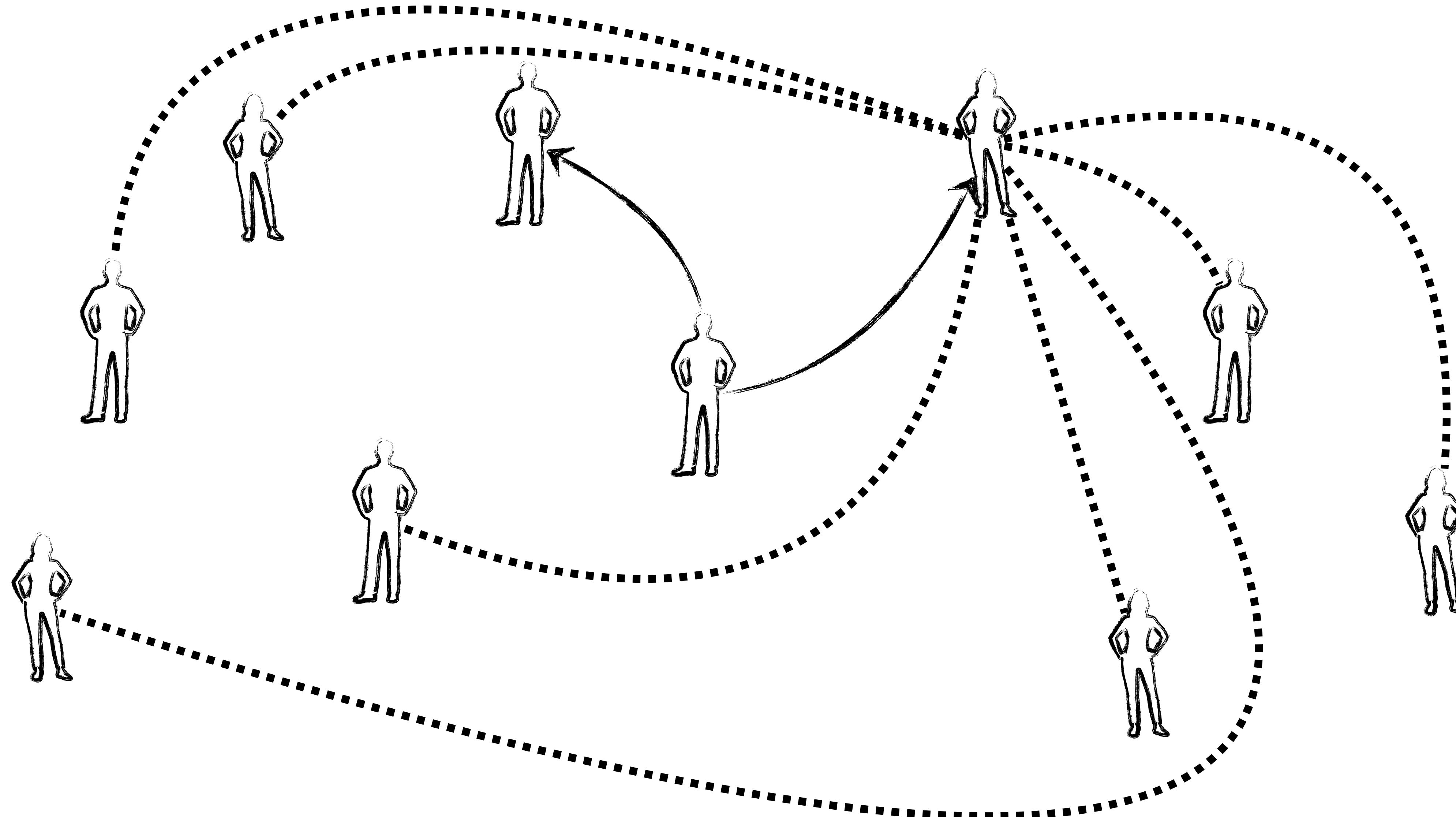
# Распространение инфекций



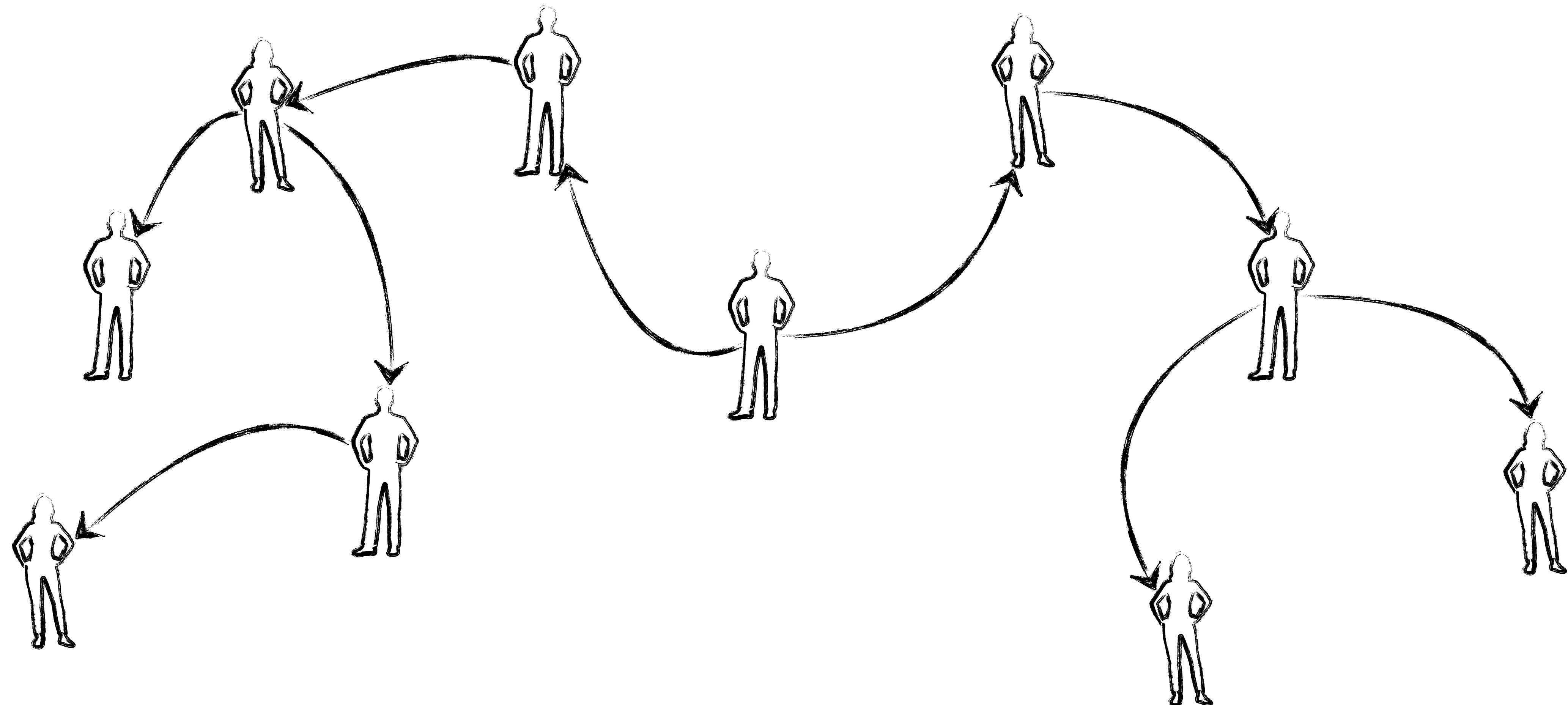
# Распространение инфекций



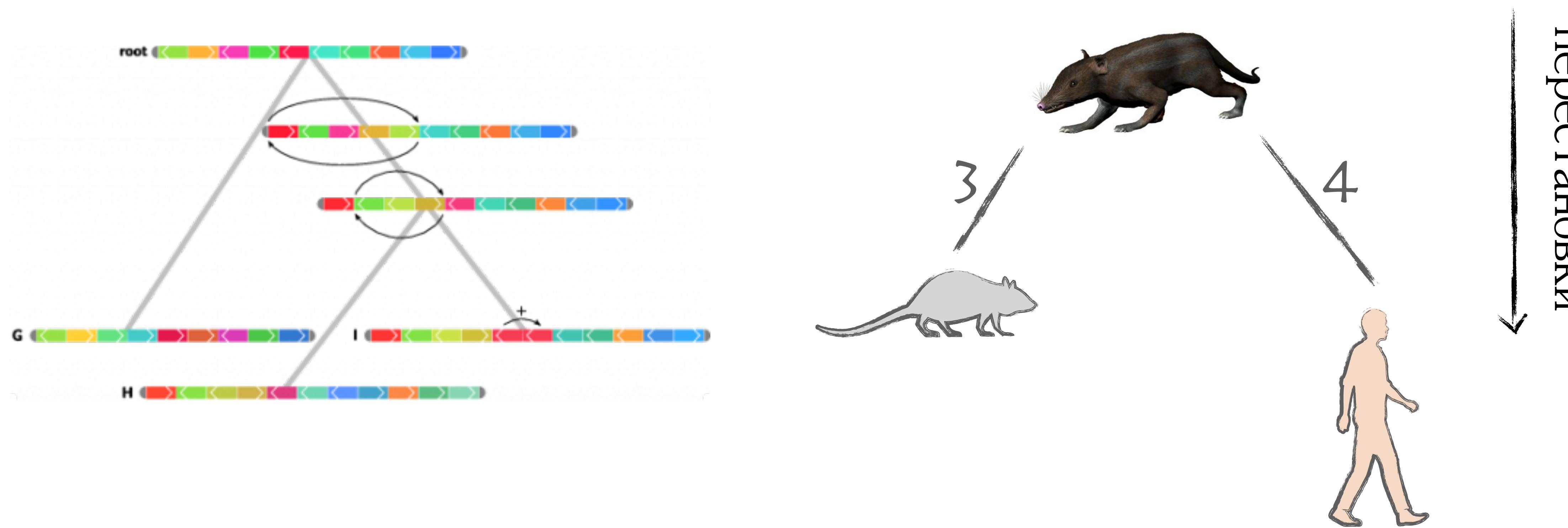
# Распространение инфекций



# Распространение инфекций



# Восстановление эволюционной картины

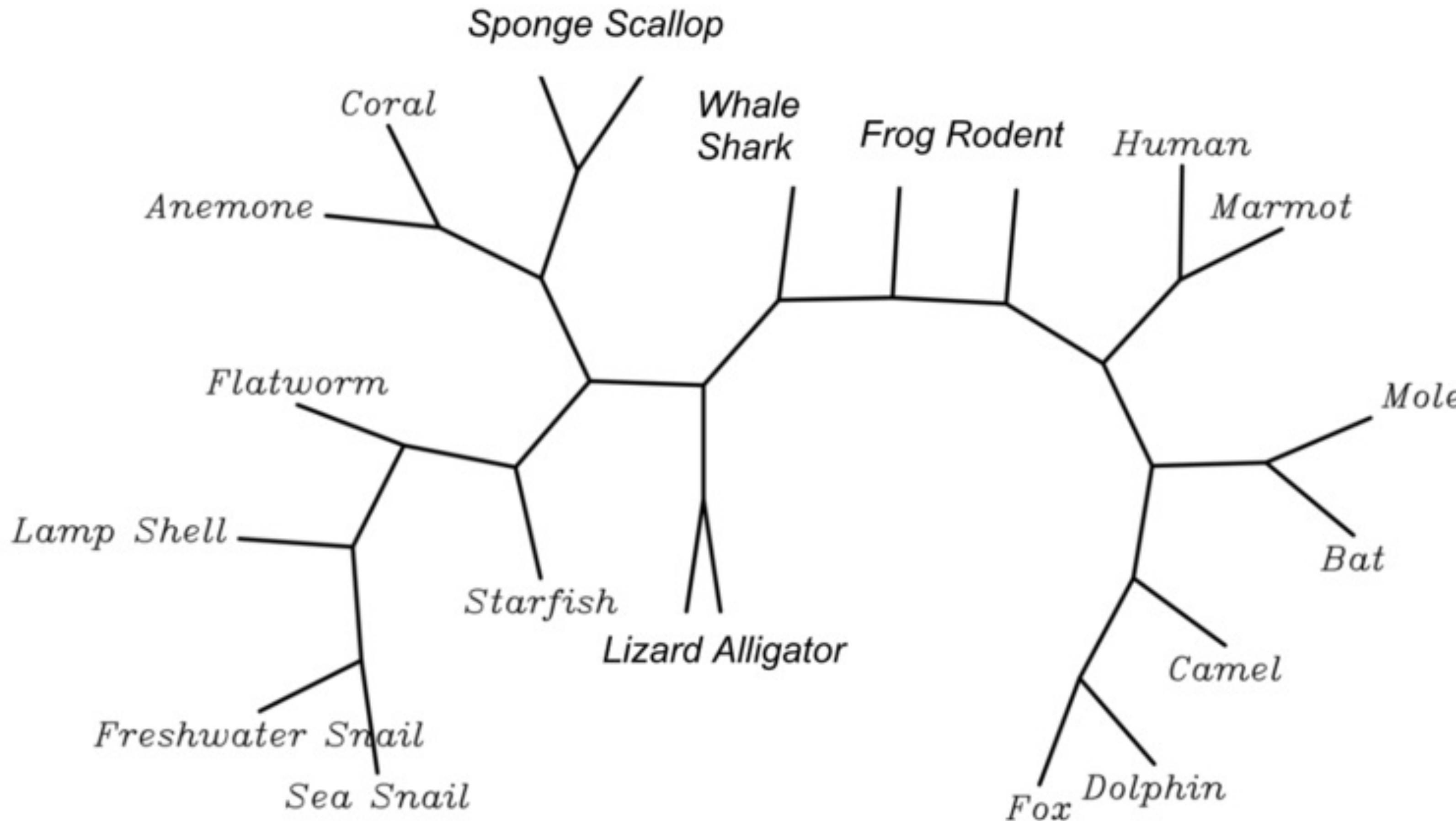


# Филогенетические деревья

Филогенетическое дерево обладает следующими свойствами:

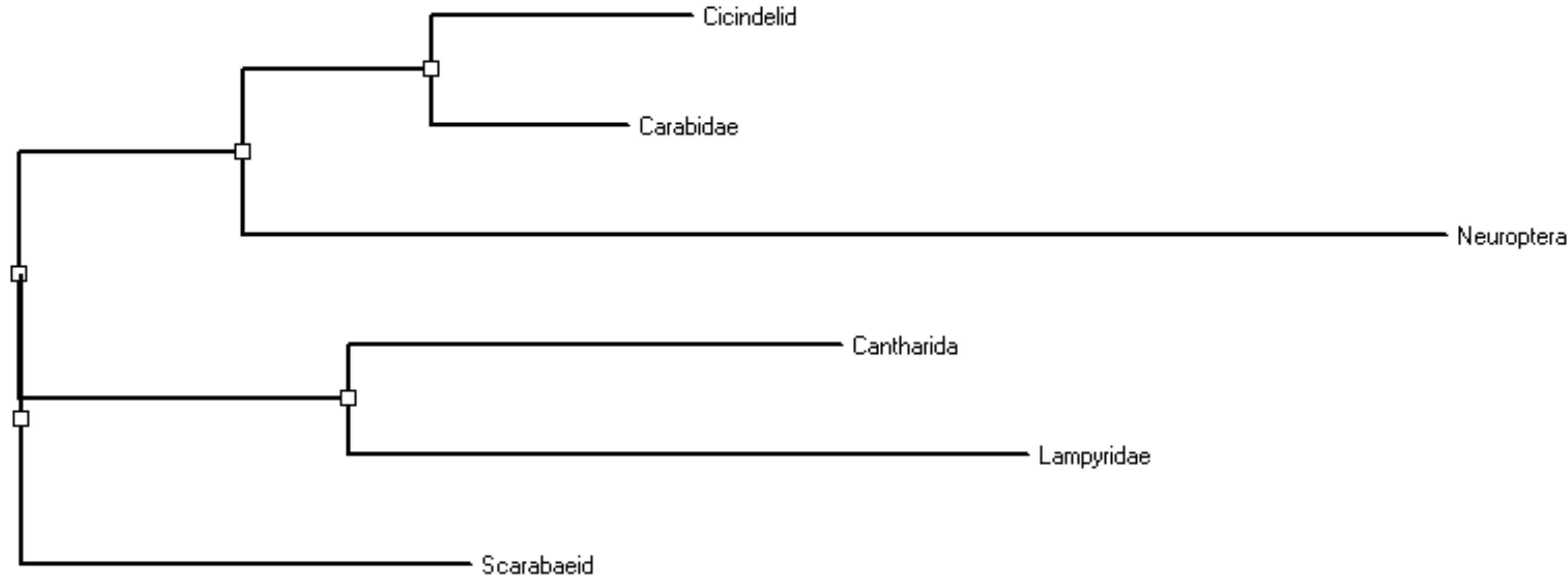
- Это дерево
- $\deg(v_i) \neq 2$
- Листья, как правило, соответствуют наблюдаемым последовательностям
- $\max \deg(v_i) = 3$

# Филогенетические деревья



# Филогенетические деревья

Филогенетическое дерево может быть укорененным



# Филогенетические деревья

Множественное выравнивание

-	A	G	G	C	T	A	T	C	A	C	C	T	G
T	A	G	-	CC	T	A	C	C	A	-	-	-	GGG
C	A	G	-	CC	T	A	C	C	A	-	-	-	GG
C	A	G	-	CC	T	A	T	C	A	C	-	G	G
C	A	G	-	C	T	A	T	C	G	C	-	G	G

6			
6	1		
4	4	3	
5	5	4	1

# UPGMA

UPGMA (unweighted pair group method with arithmetic mean)

- Строим матрицу попарных расстояний
- Находим ближайшие последовательности и объединяем, вычеркивая столбец и строку из матрицы
- Пересчитываем расстояния:  $D_{i \cup j, k} = \frac{D_{i,k} + D_{j,k}}{2}$
- Возвращаемся к шагу 2 если осталось более двух последовательностей

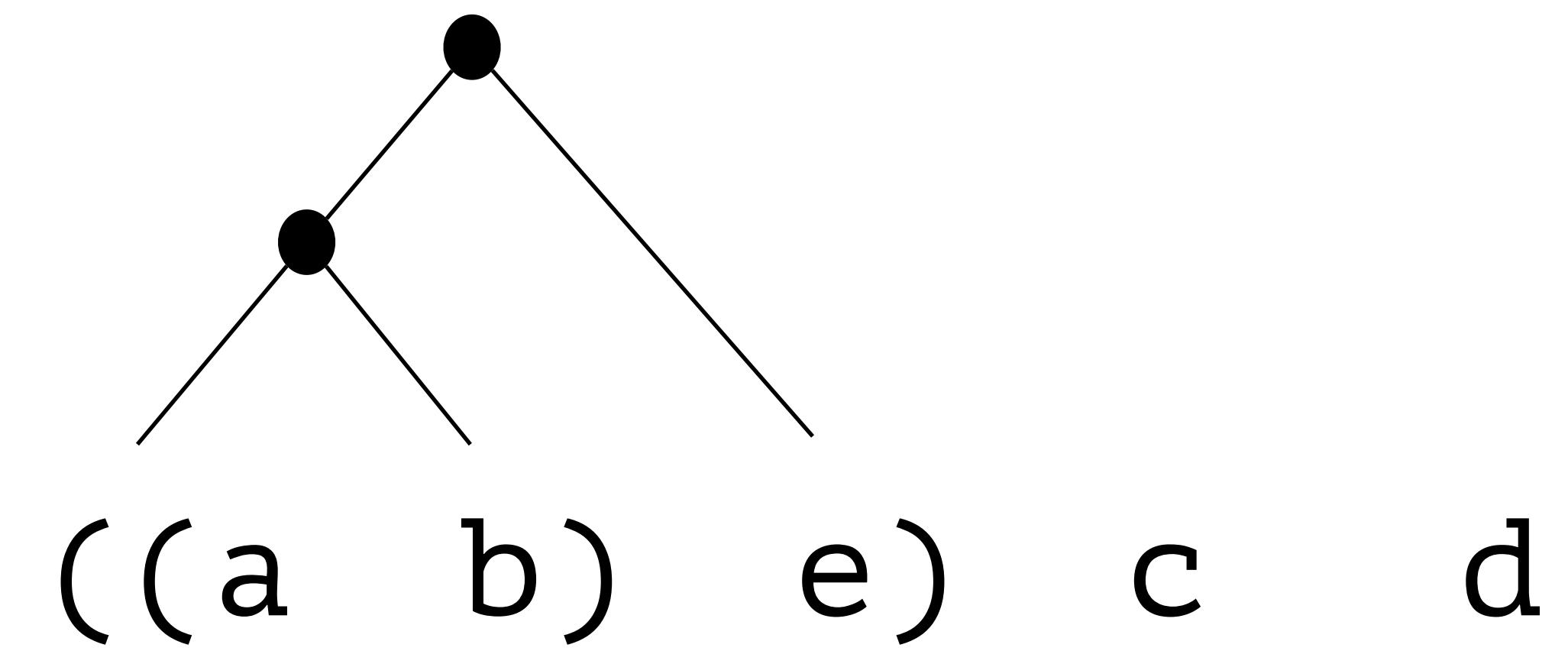
# UPGMA

	a	b	c	d	e
a	0	17	21	31	23
b	17	0	30	34	21
c	21	30	0	28	39
d	31	34	28	0	43
e	23	21	39	43	0



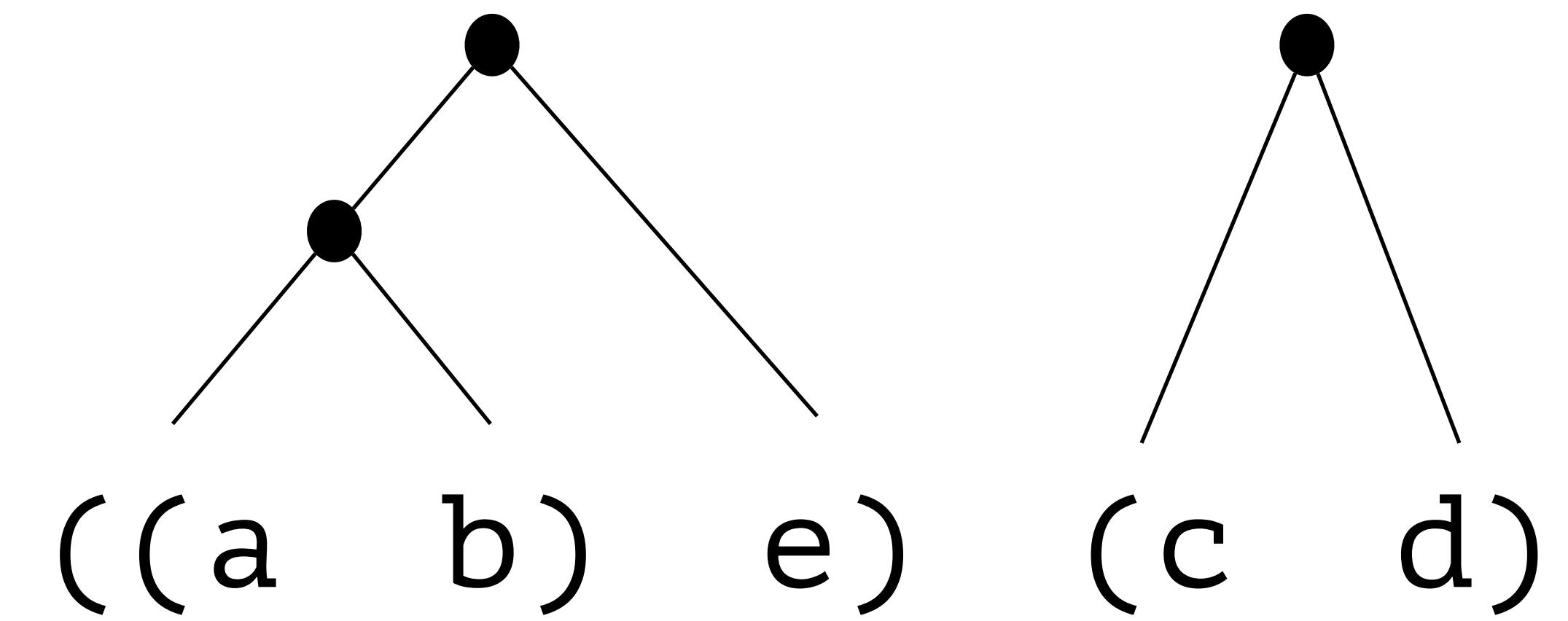
# UPGMA

	(a,b)	c	d	e
(a,b)	0	25.5	32.5	22
c	25.5	0	28	39
d	32.5	28	0	43
e	22	39	43	0



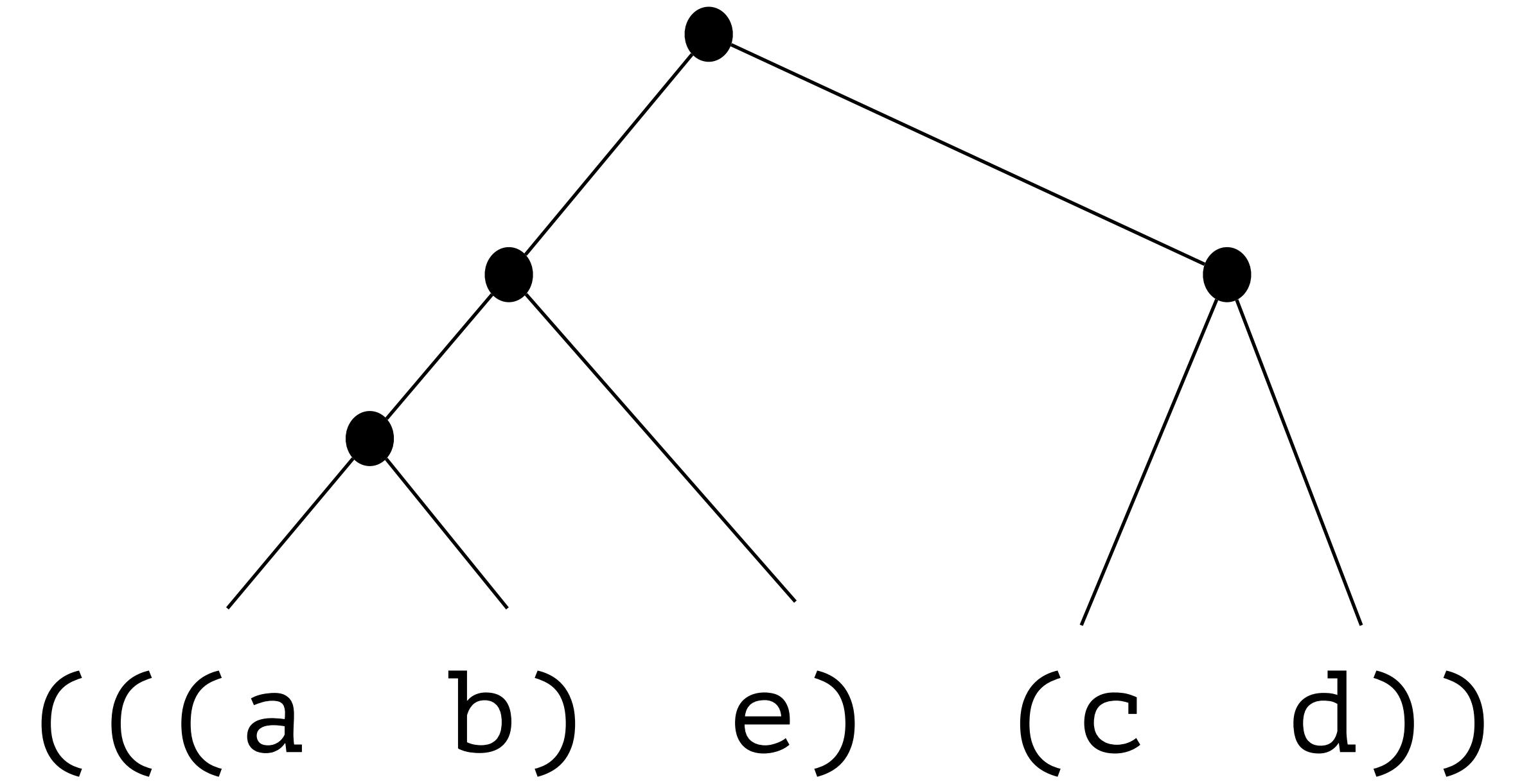
# UPGMA

	<b>((a,b),e)</b>	<b>c</b>	<b>d</b>
<b>((a,b),e)</b>	0	<b>32.25</b>	<b>37.75</b>
<b>c</b>	<b>32.25</b>	0	<b>28</b>
<b>d</b>	<b>37.75</b>	<b>28</b>	0

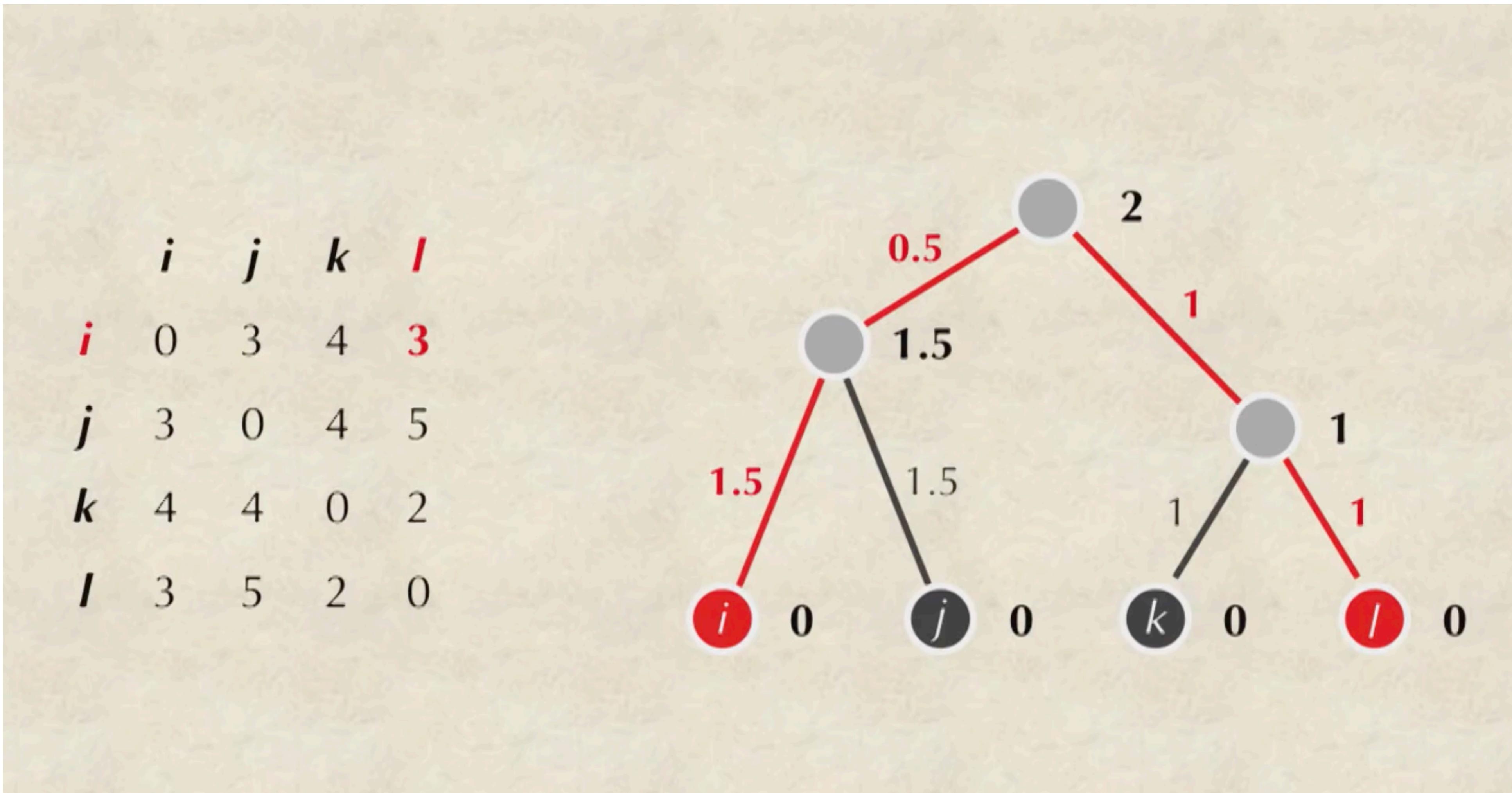


# UPGMA

	((a,b),e)	(c,d)
((a,b),e)	0	35
(c,d)	35	0



# UPGMA



# Neighbor joining

- Матрица попарных расстояний
- По матрице попарных расстояний считается Q-матрица
- Ищется пара ближайших последовательностей по матрице Q
- Они присоединяются к новому узлу, который соединяется с центральным
- Рассчитывается расстояние от каждой из присоединённых последовательностей до нового узла
- Рассчитывается расстояние от каждого из оставшихся последовательностей до нового узла

# Neighbor joining

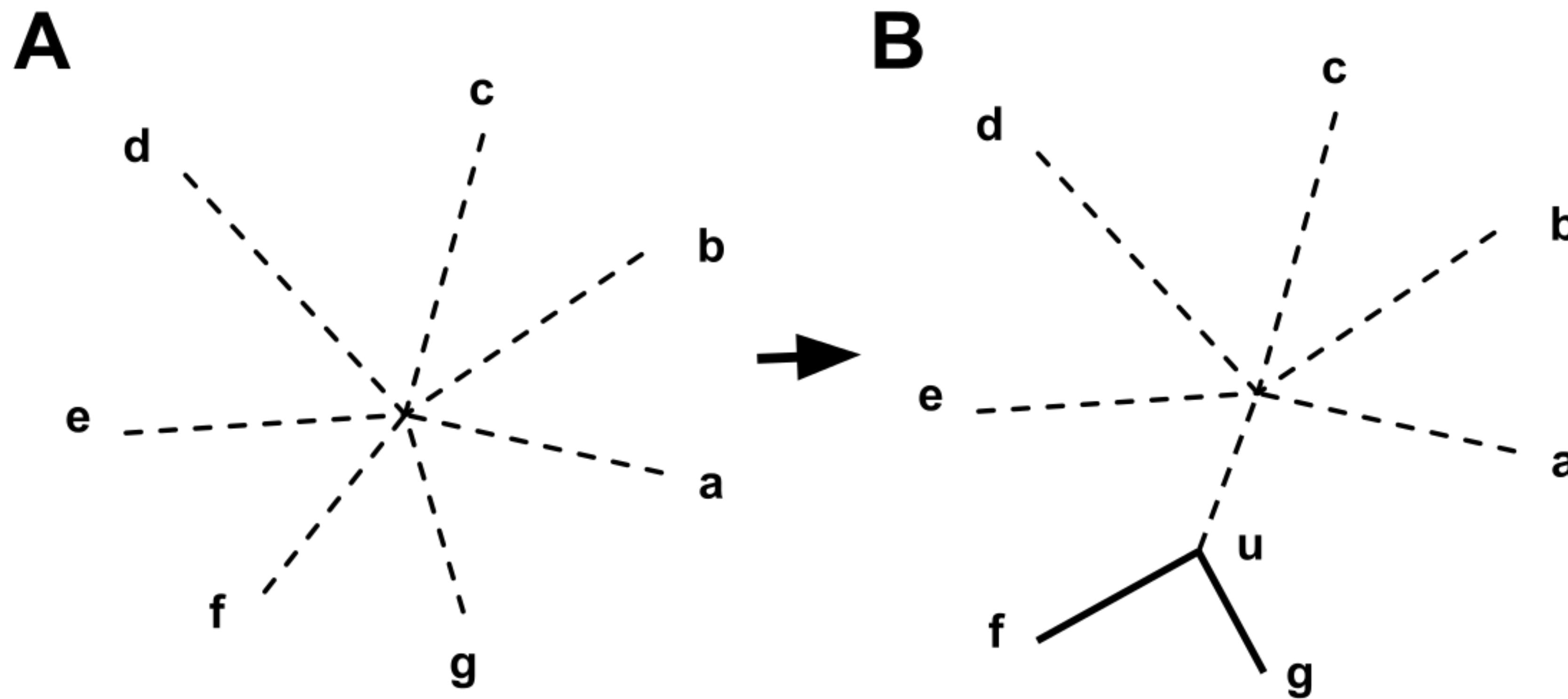
$$\circ D^*(i,j) = (n - 2)d(i,j) - \sum_{k=1}^n d(i,k) - \sum_{k=1}^n d(j,k)$$

**Neighbor-Joining Theorem:** If  $D$  is additive, then the smallest element of  $D^*$  corresponds to neighboring leaves in  $\text{Tree}(D)$ .

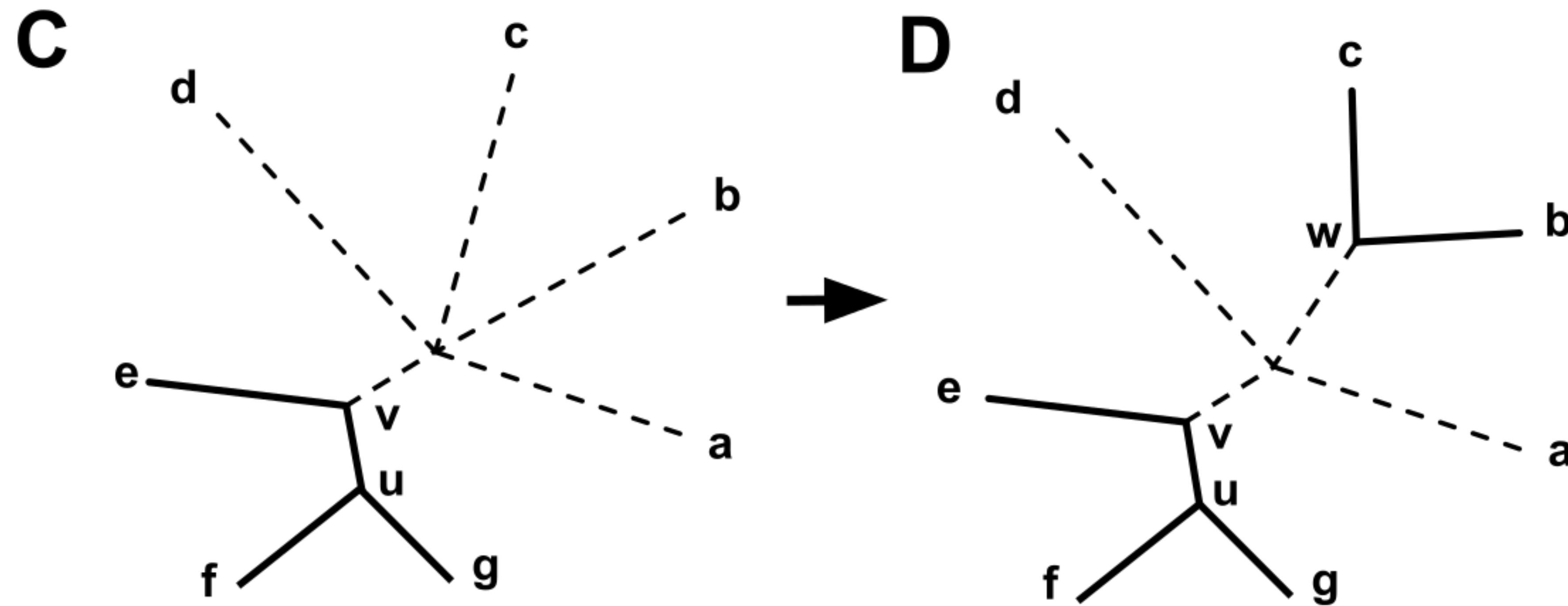
# Neighbor joining

- $$Q(i, j) = (n - 2)d(i, j) - \sum_{k=1}^n d(i, k) - \sum_{k=1}^n d(j, k)$$
- $$\delta(f, u) = \frac{1}{2}d(f, g) + \frac{1}{2(n - 2)} \left( \sum_{k=1}^n d(f, k) - \sum_{k=1}^n d(g, k) \right)$$
$$\delta(g, u) = d(f, g) - \delta(f, u)$$

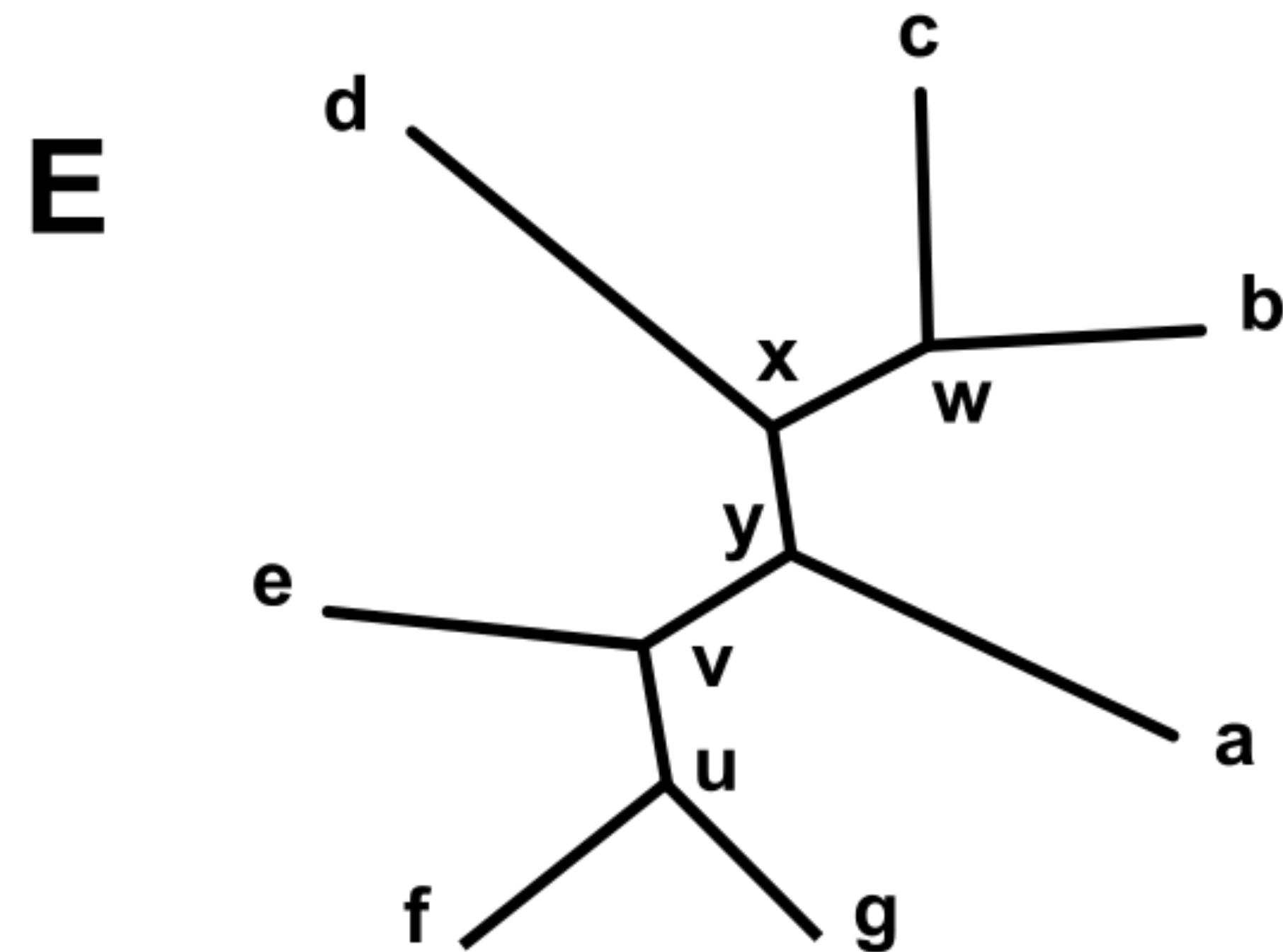
# Neighbor joining



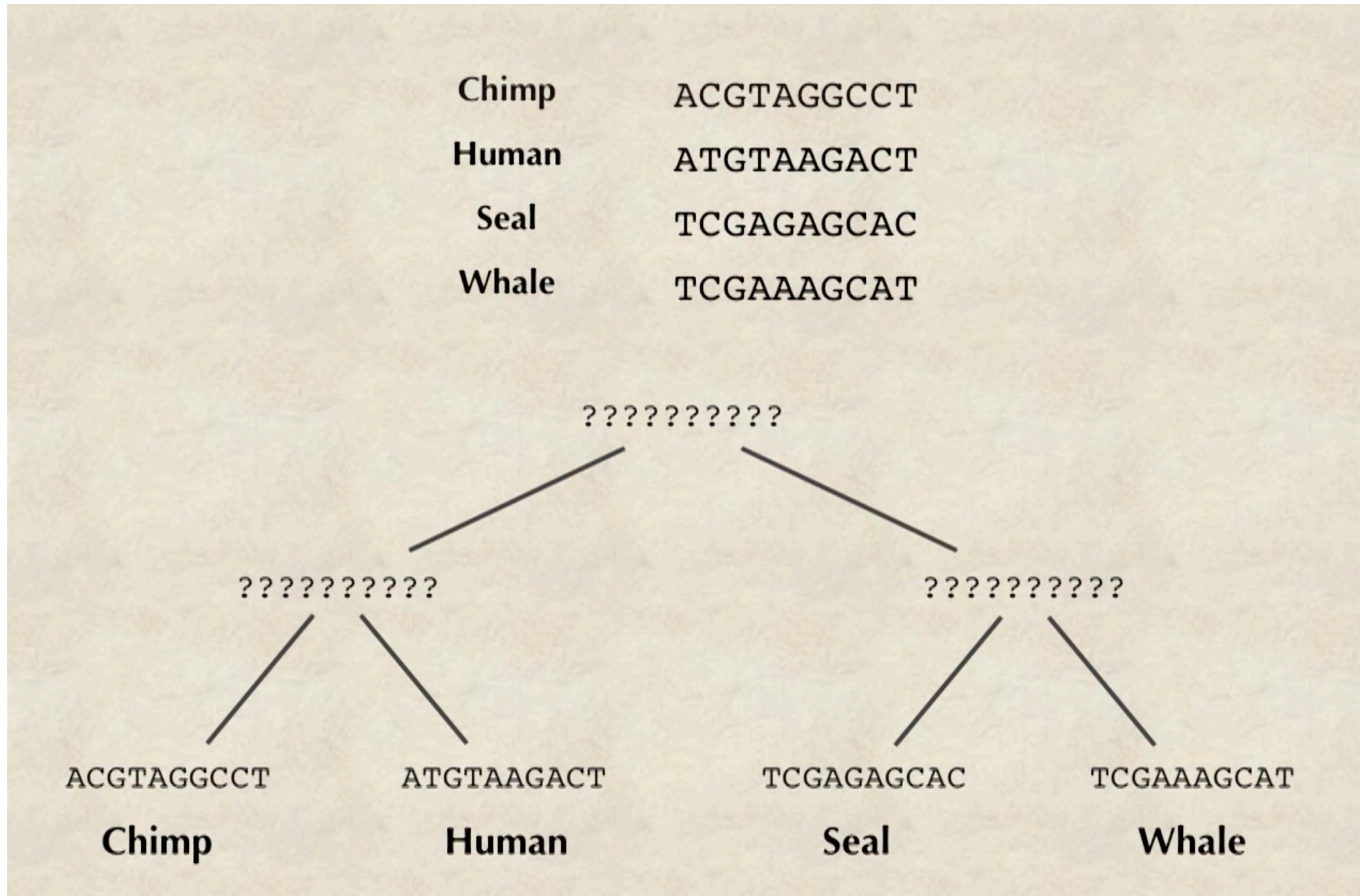
# Neighbor joining



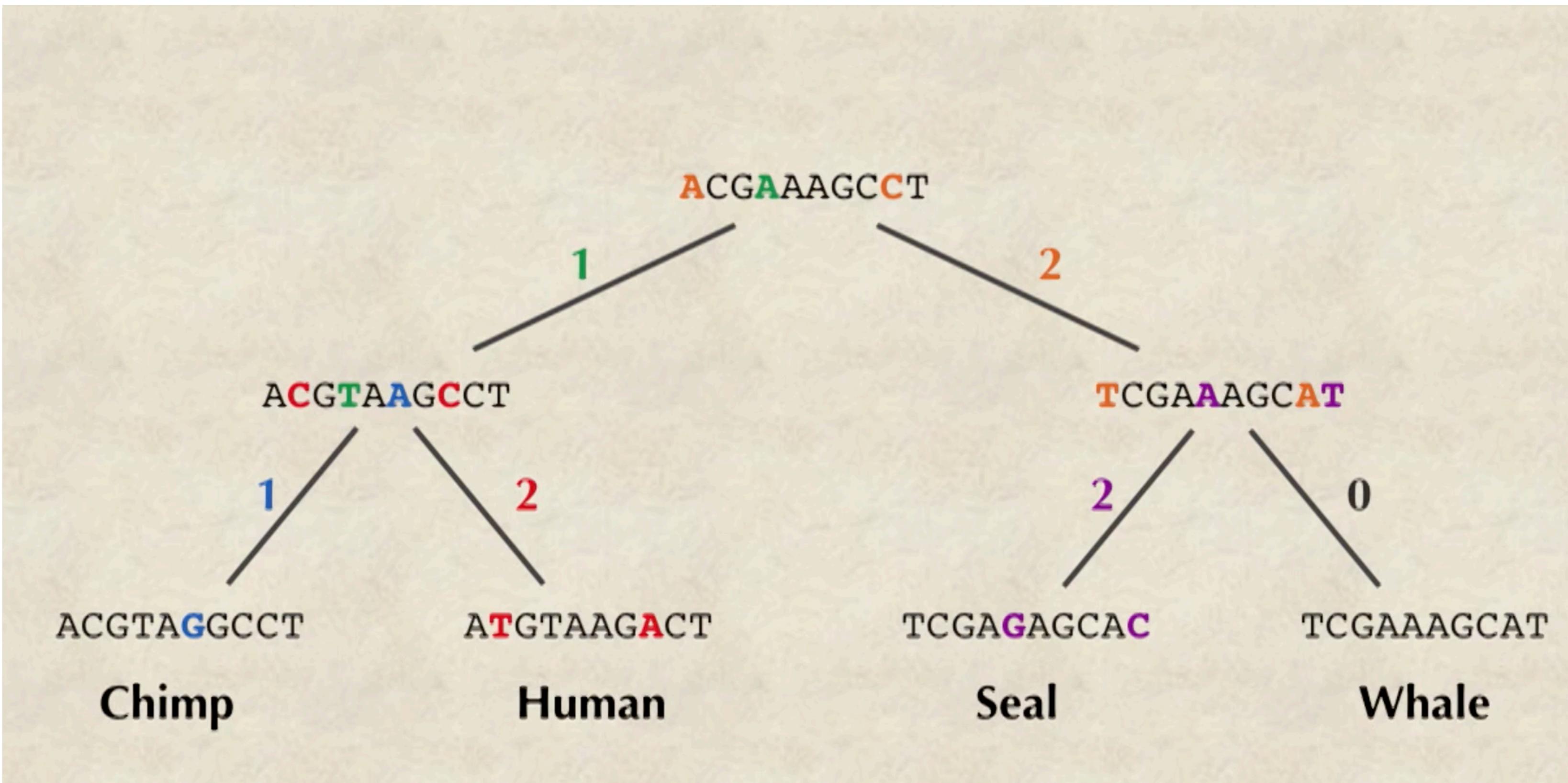
# Neighbor joining



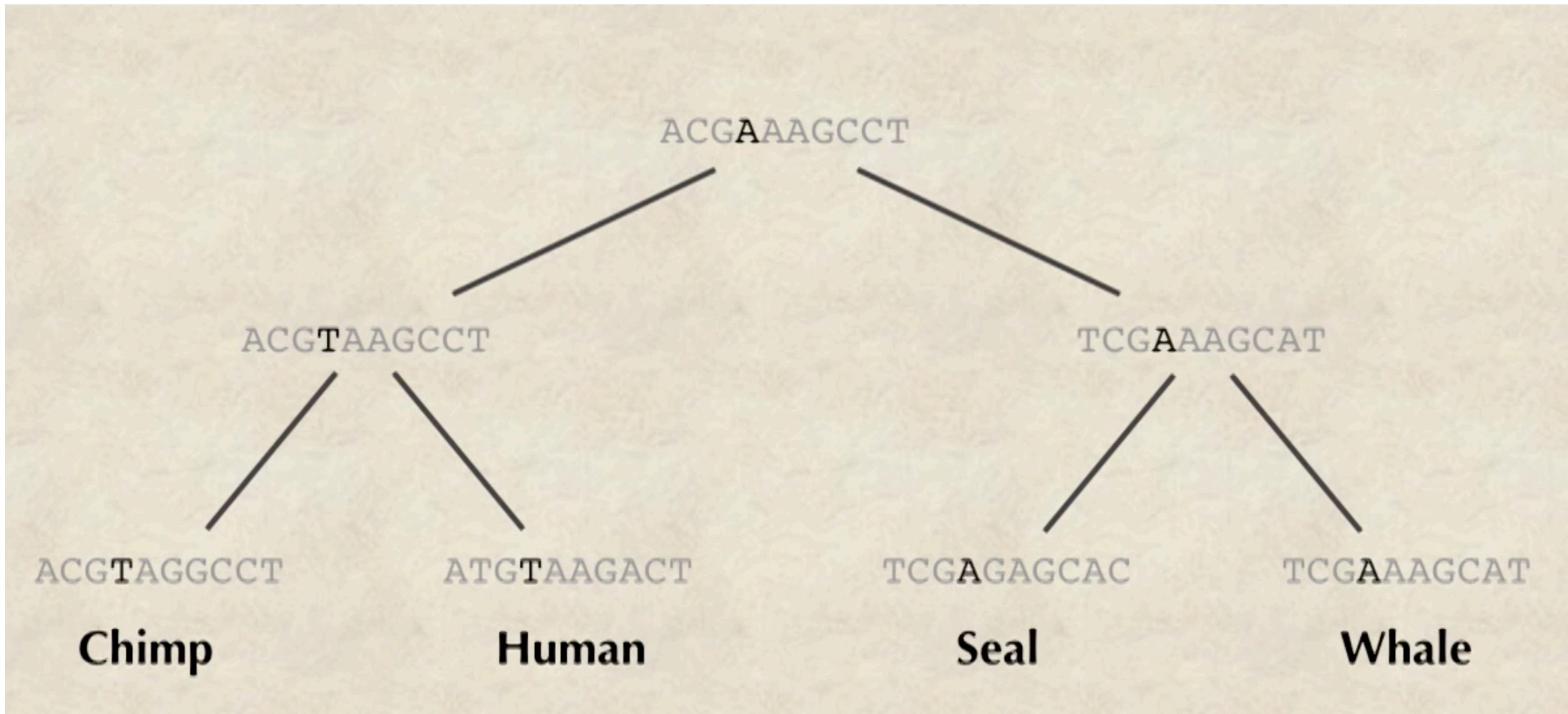
# Нахождение генома предков



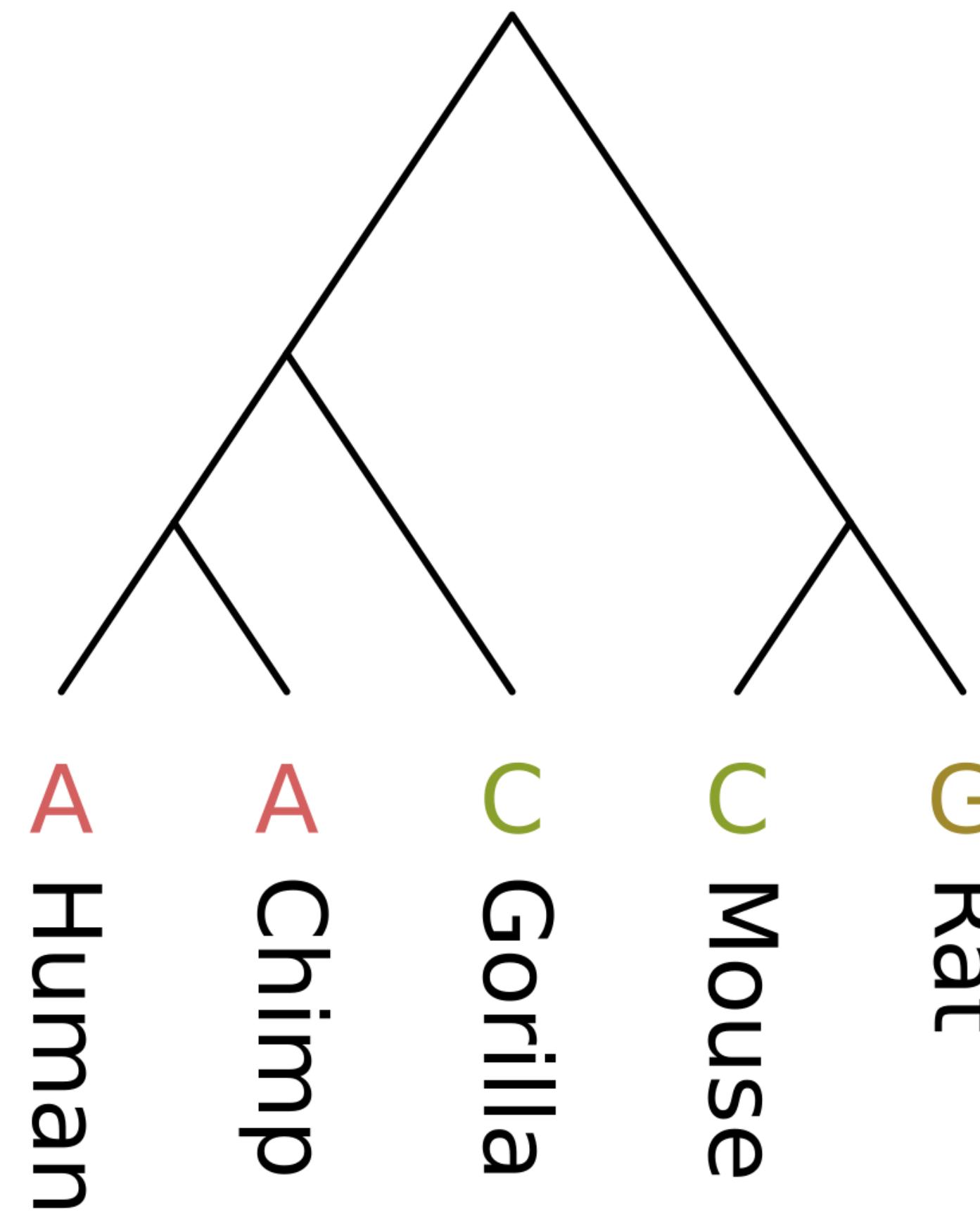
# Нахождение генома предков



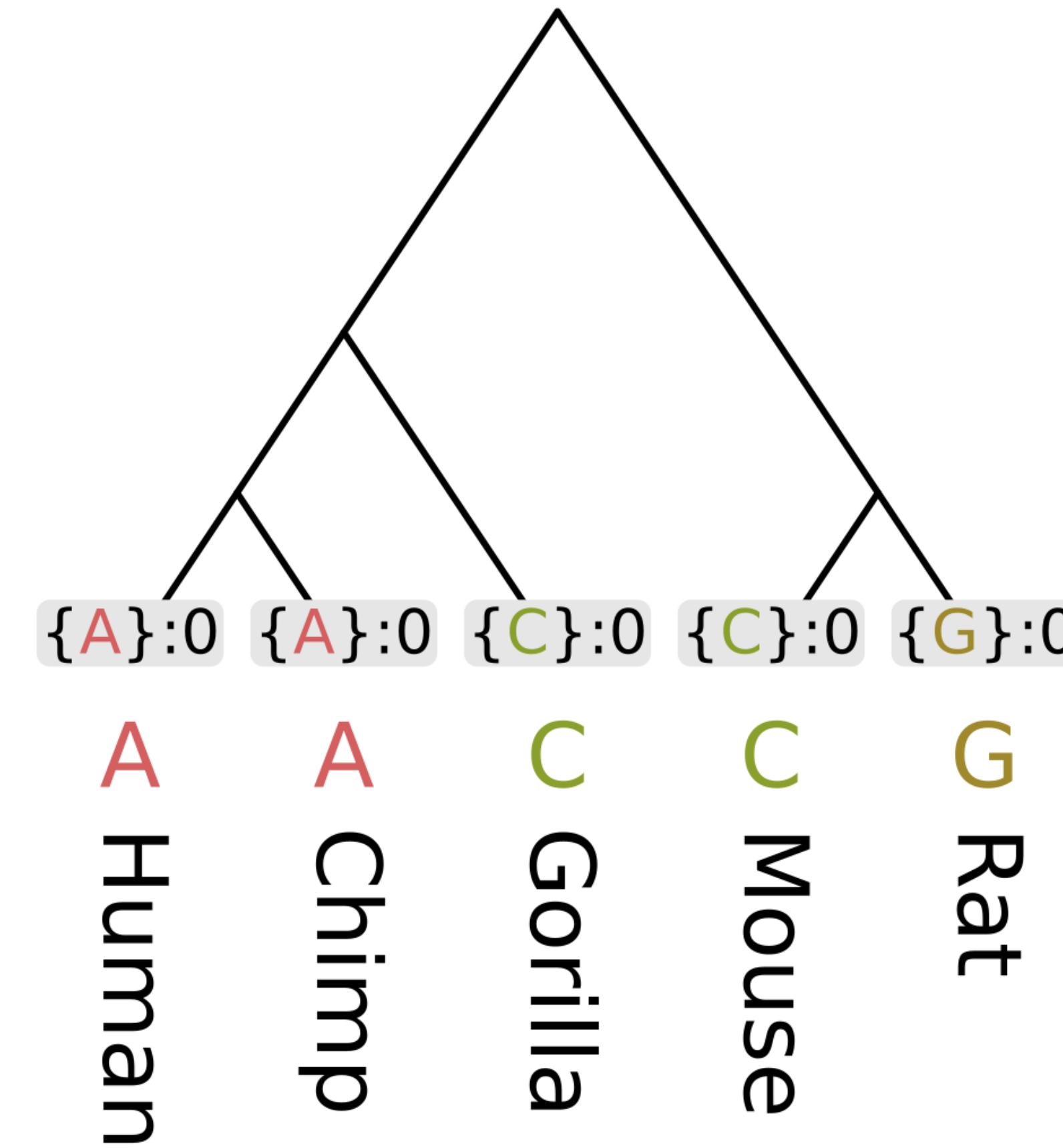
# Нахождение генома предков



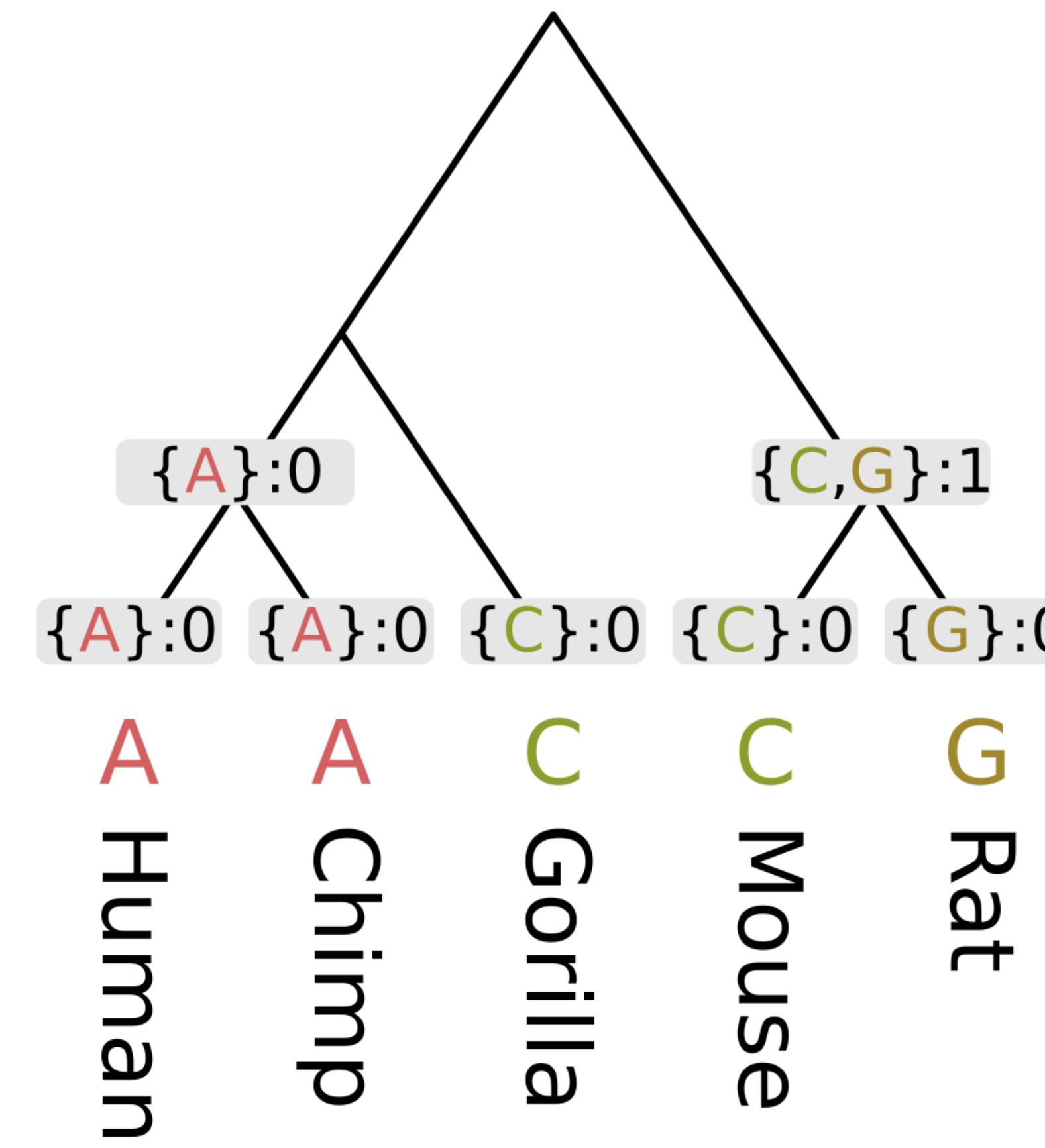
# Нахождение генома предков



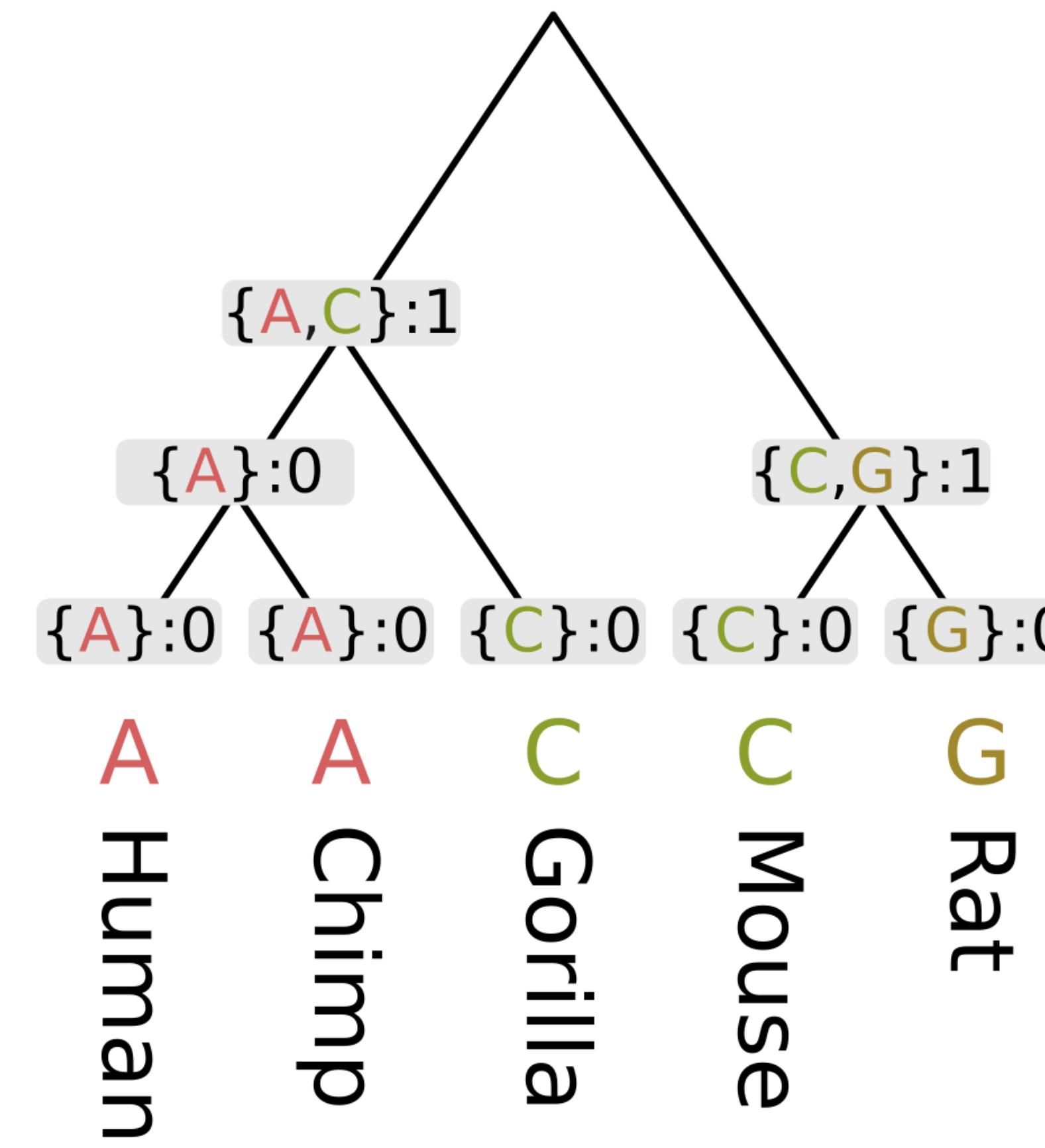
# Нахождение генома предков



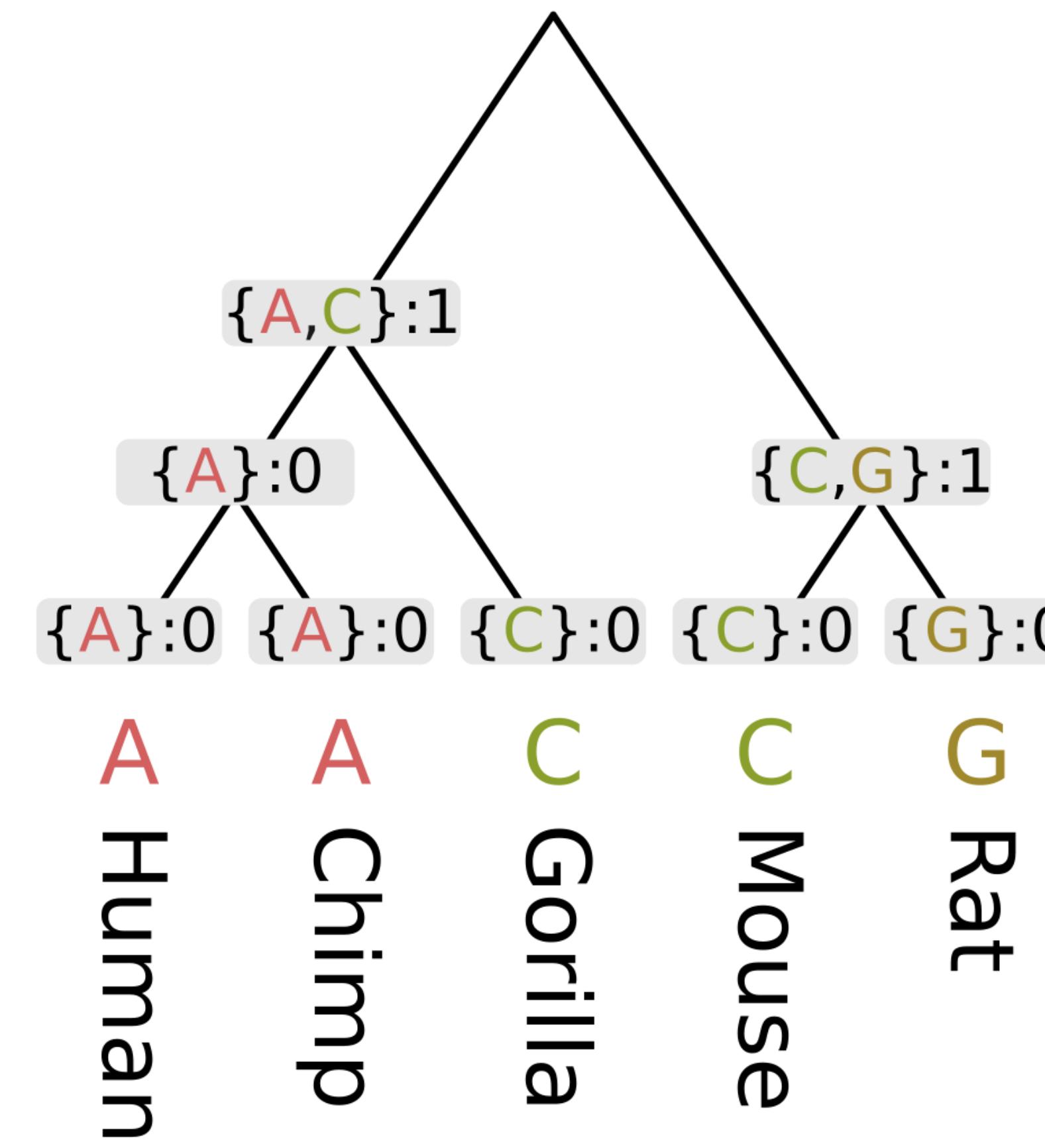
# Нахождение генома предков



# Нахождение генома предков



# Нахождение генома предков



# Резюмируем

- Филогения важна как для эволюционных исследований так и для определения вспышек эпидемий инфекций
- Чтобы найти правдоподобную топологию дерева, можно использовать например метод присоединения соседей
- Для восстановление самих последовательностей используется алгоритм Фитча