Hledání genů skrytými Markovovými modely

Lukáš Ondráček
ondracek.lukas@gmail.com

19. 1. 2018

Zadání

- Cíl: Vytvořit skrytý Markovův model podle anotací jednoho vlákna genomu a najít pomocí něj geny v komplementárním vlákně Viterbiho algoritmem.
- Použitý genom: Yersinia pestis Antiqua (NC_008150.1)
- Implementace: vlastní v Pythonu

1. model

Stavy:

- nekódující: –
- start kodón: <
- kódující: #
- stop kodón: >

Příklad genu:

– – – – – < < < # # # # # # > > > – – – –

	Primární vlákno	Komplementární vlákno
Délka	4 702 289	
Počet genů (délka)	2105 (1 949 511)	2115 (1 921 983)
Počet nalezených úseků	102 (1 875 913)	111 (2 006 087)
Přesně označené geny	0 (0)	0 (0)

2. model

Stavy:

- nekódující: `-`
- start kodón: `<1`, `<2`, `<3` (pro každou pozici jiný)
- kódující: `#1`, `#2`, `#3` (podle pozice v genu modulo 3)
- stop kodón: `>1`, `>2`, `>3`

Příklad genu:

```
- - - - - <1 <2 <3 #1 #2 #3 #1 #2 #3 >1 >2 >3 - - - -
```

	Primární vlákno	Komplementární vlákno
Délka	4 702 289	
Počet genů (délka)	2105 (1 949 511)	2115 (1 921 983)
Počet nalezených úseků	1388 (2 549 327)	1440 (2 595 116)
Přesně označené geny	187 (207 690)	170 (180 012)

3. model

Stavy:

- nekódující: `-`
- start kodón: `<1`, `<2`, `<3`
- kódující: `#1`, `#2`, `#3`
- stop kodón: `>a1`, `>a2`, `>a3` (uprostřed `A`); `>b1`, `>b2`, `>b3` (jinak)

Příklad genu:

```
- - - - - <1 <2 <3 #1 #2 #3 #1 #2 #3 >1b >2b >3b - - - -
```

	Primární vlákno	Komplementární vlákno
Délka	4 702 289	
Počet genů (délka)	2105 (1 949 511)	2115 (1 921 983)
Počet nalezených úseků	1396 (2 561 239)	1474 (2 610 667)
Přesně označené geny	216 (238 995)	208 (225 636)

4. model

Stavy:

- nekódující: `-a` (krátké úseky, < 50); `-b` (dlouhé úseky, ≥ 50)
- start kodón: `<1`, `<2`, `<3`
- kódující: `#1`, `#2`, `#3`
- stop kodón: `>a1`, `>a2`, `>a3`; `>b1`, `>b2`, `>b3`

Příklad genu:

```
-b -b -b <1 <2 <3 #1 #2 #3 #1 #2 #3 >1b >2b >3b -a -a
```

	Primární vlákno	Komplementární vlákno
Délka	4 702 289	
Počet genů (délka)	2105 (1 949 511)	2115 (1 921 983)
Počet nalezených úseků	1524 (2 568 665)	1603 (2 604 498)
Přesně označené geny	278 (309 081)	261 (260 562)

5. model

Stavy:

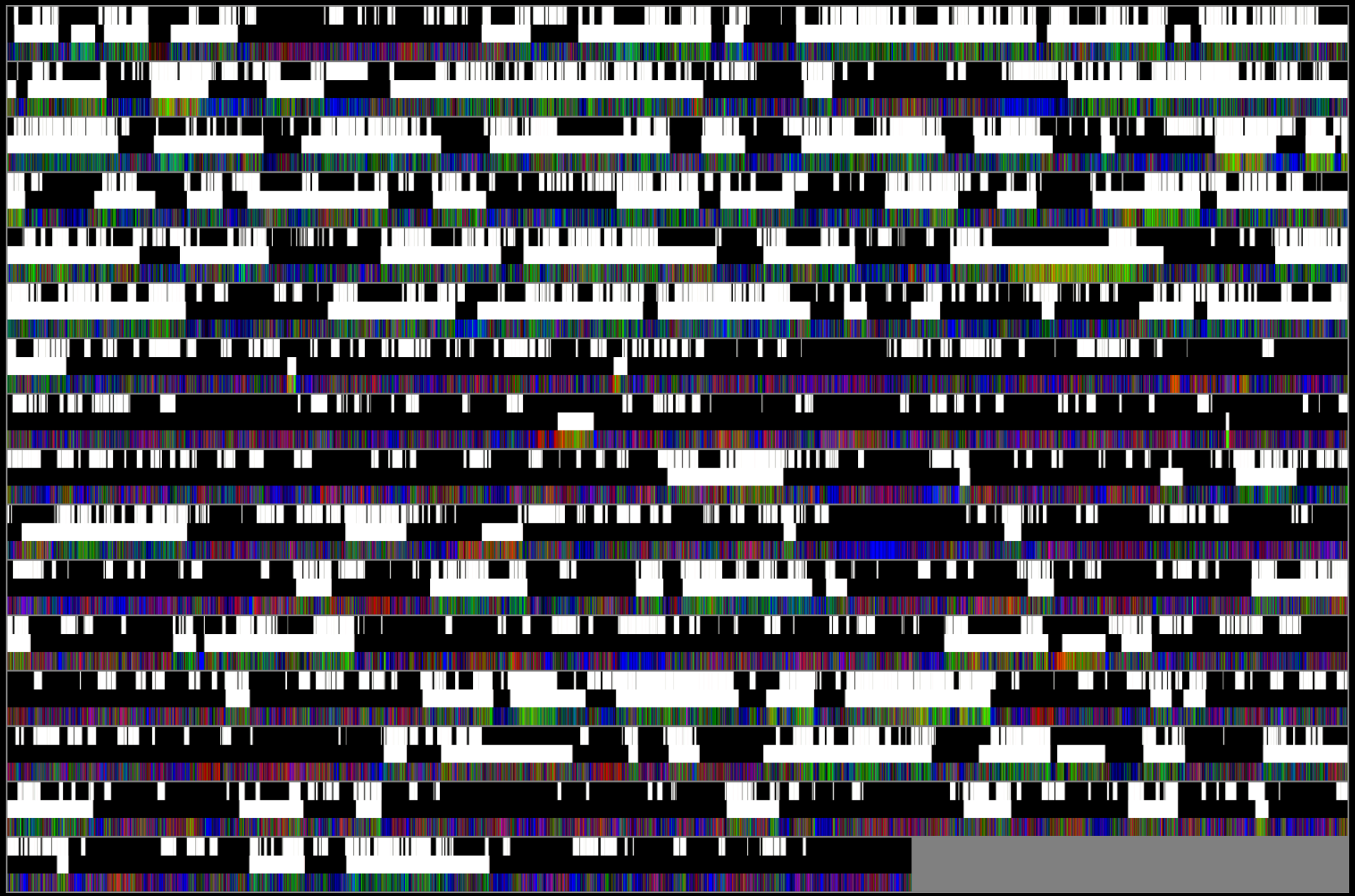
- nekódující: `-a`; `-b`, `-b14`, ..., `-b1` (podle vzdálenosti ke genu)
- start kodón: `<1`, `<2`, `<3`
- kódující: `#1`, `#2`, `#3`
- stop kodón: `>a1`, `>a2`, `>a3`; `>b1`, `>b2`, `>b3`

Příklad genu:

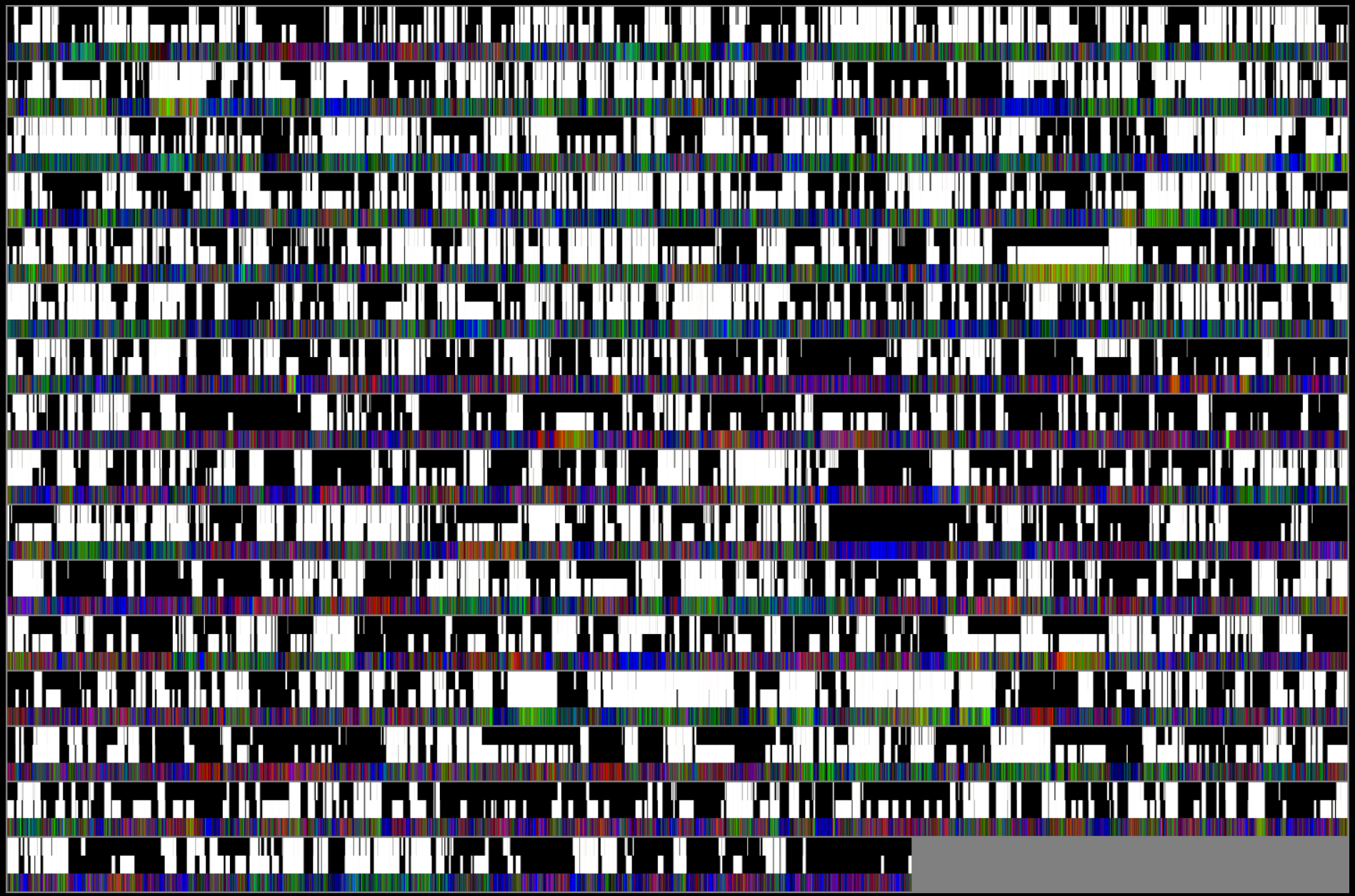
```
-b3 -b2 -b1 <1 <2 <3 #1 #2 #3 #1 #2 #3 >1b >2b >3b -a -
```

	Primární vlákno	Komplementární vlákno
Délka	4 702 289	
Počet genů (délka)	2105 (1 949 511)	2115 (1 921 983)
Počet nalezených úseků	1616 (2 554 513)	1698 (2 595 946)
Přesně označené geny	427 (450 480)	432 (421 668)

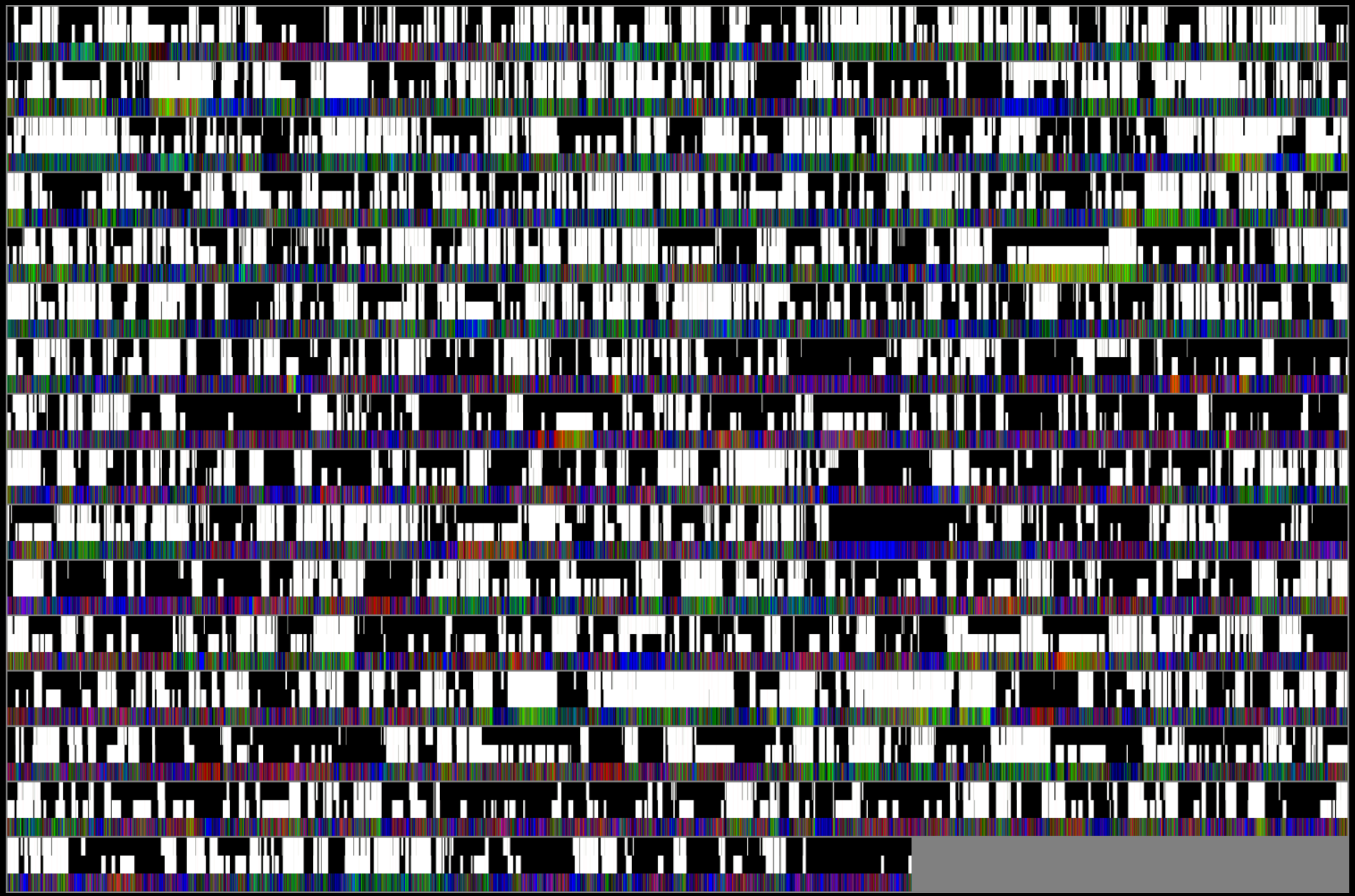
1. model – primární vlákno (0 genů)



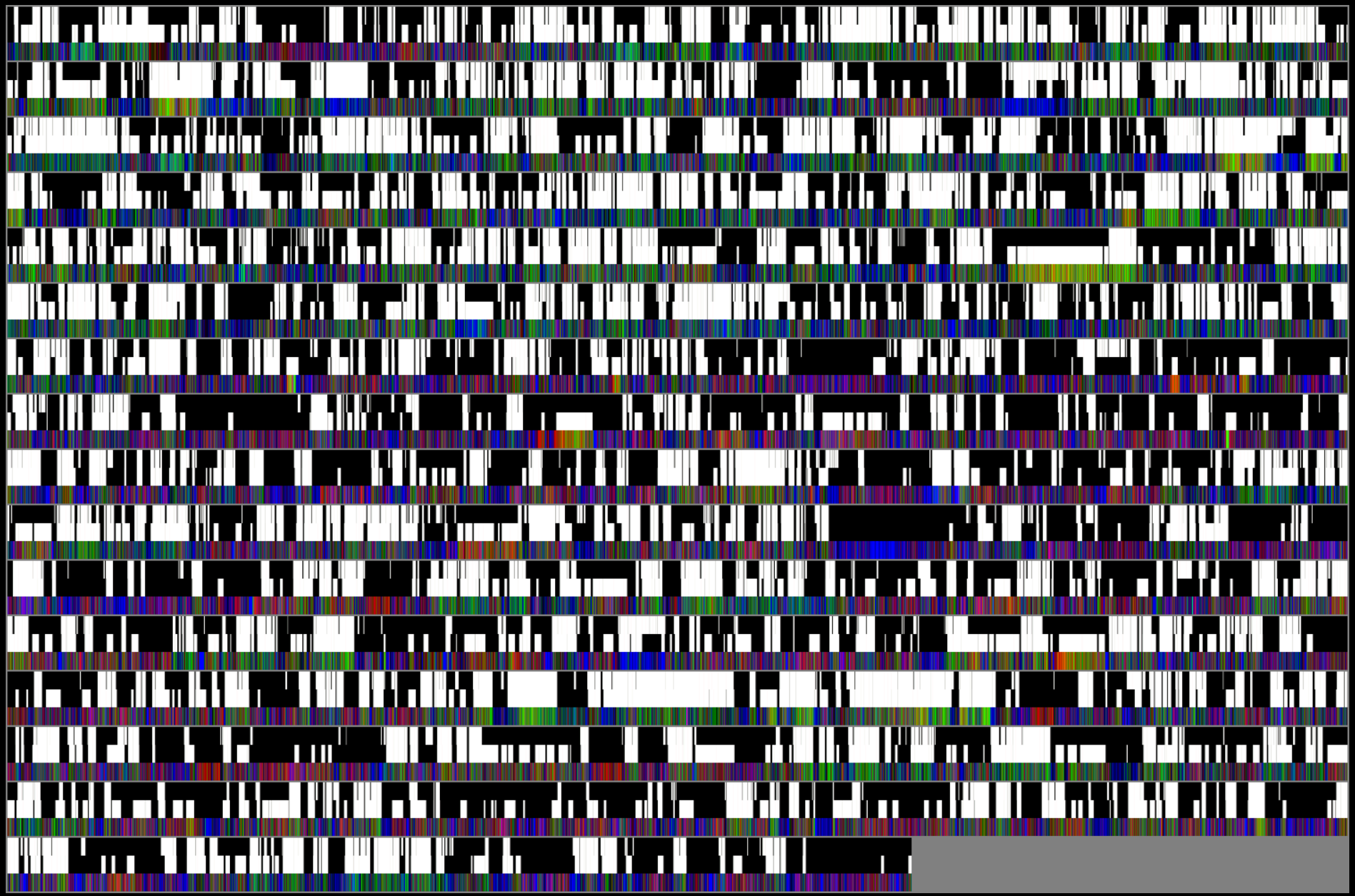
2. model – primární vlákno (187 genů)



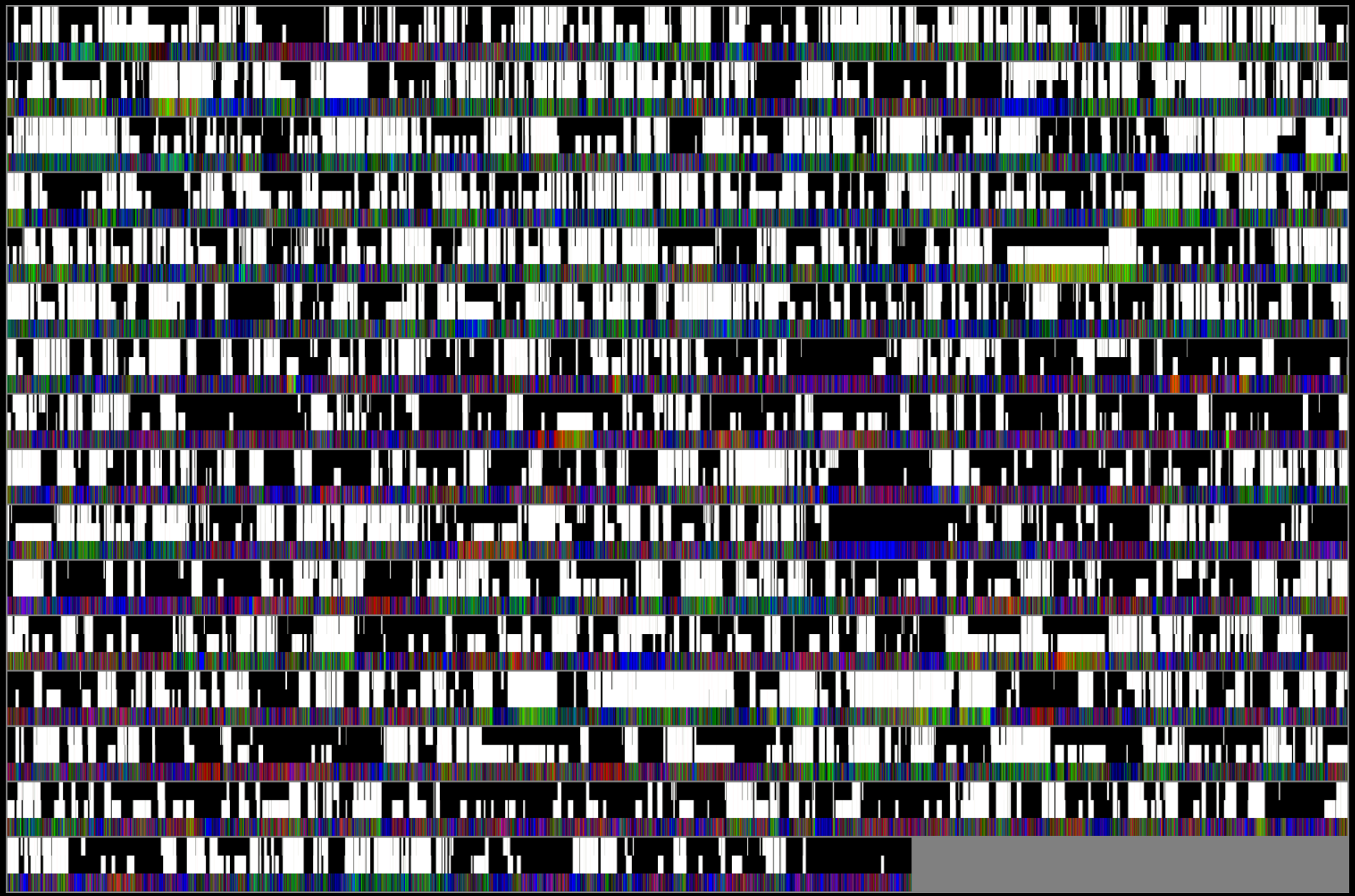
3. model – primární vlákno (216 genů)



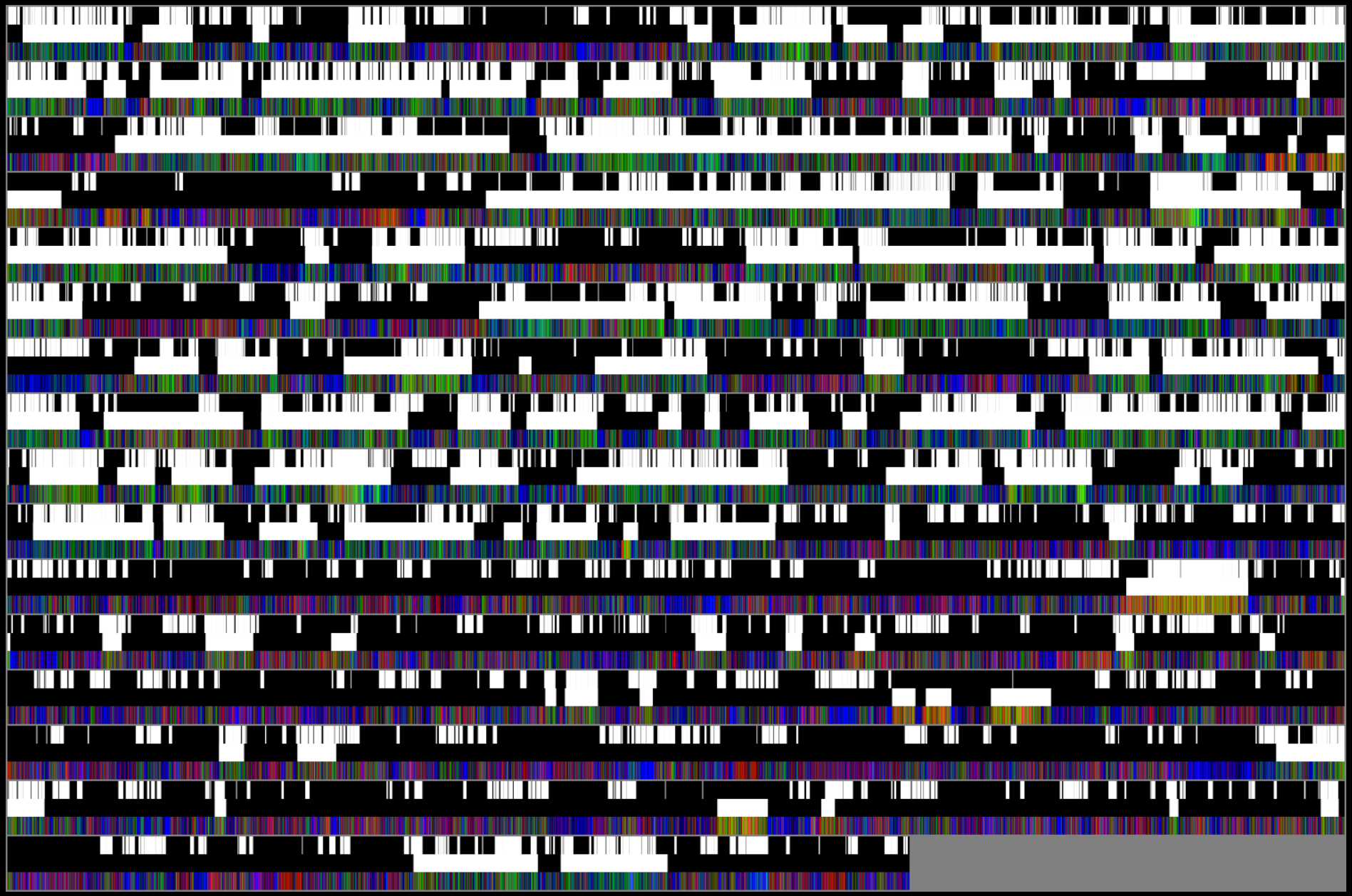
4. model – primární vlákno (278 genů)



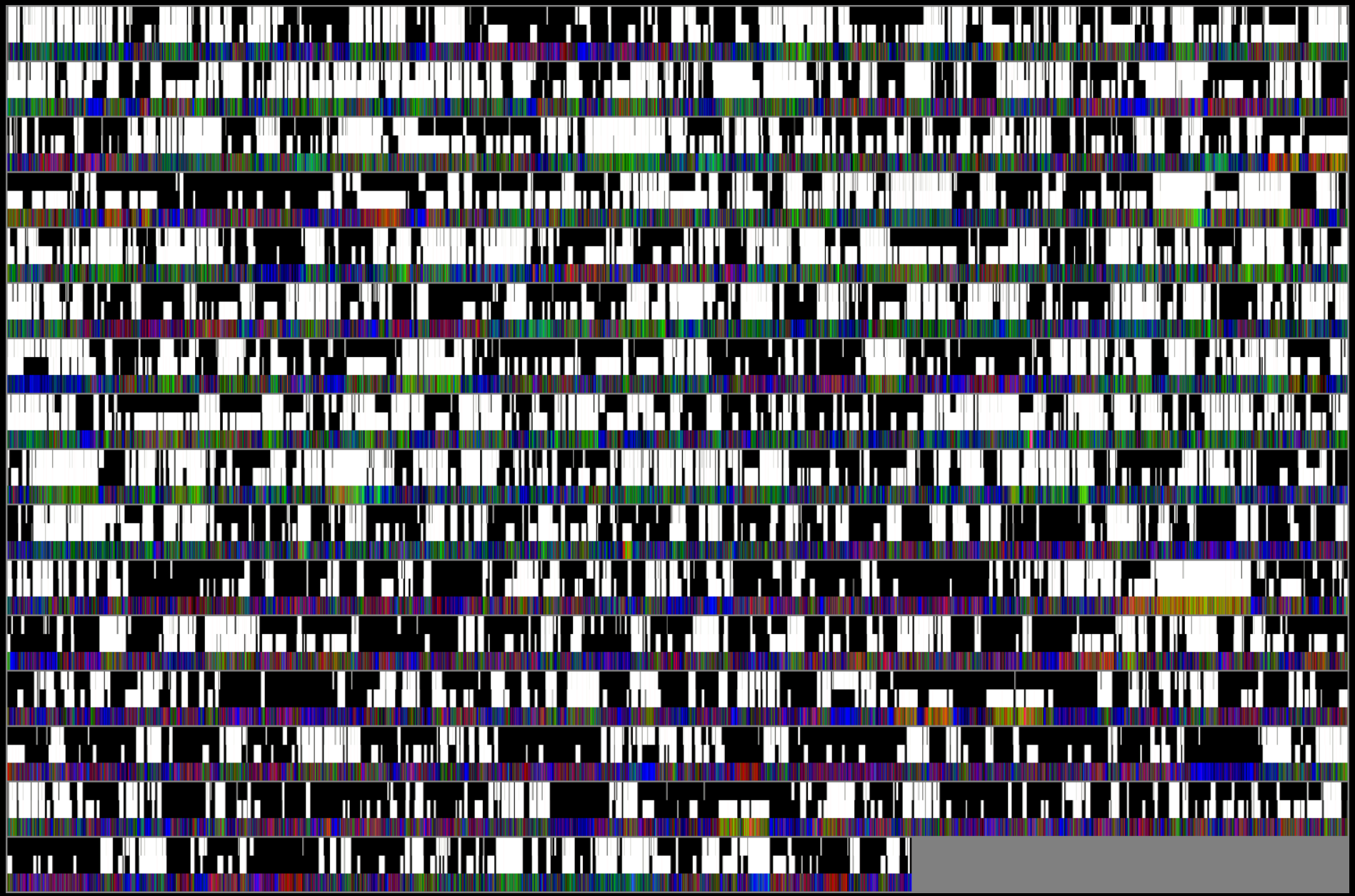
5. model – primární vlákno (427 genů)



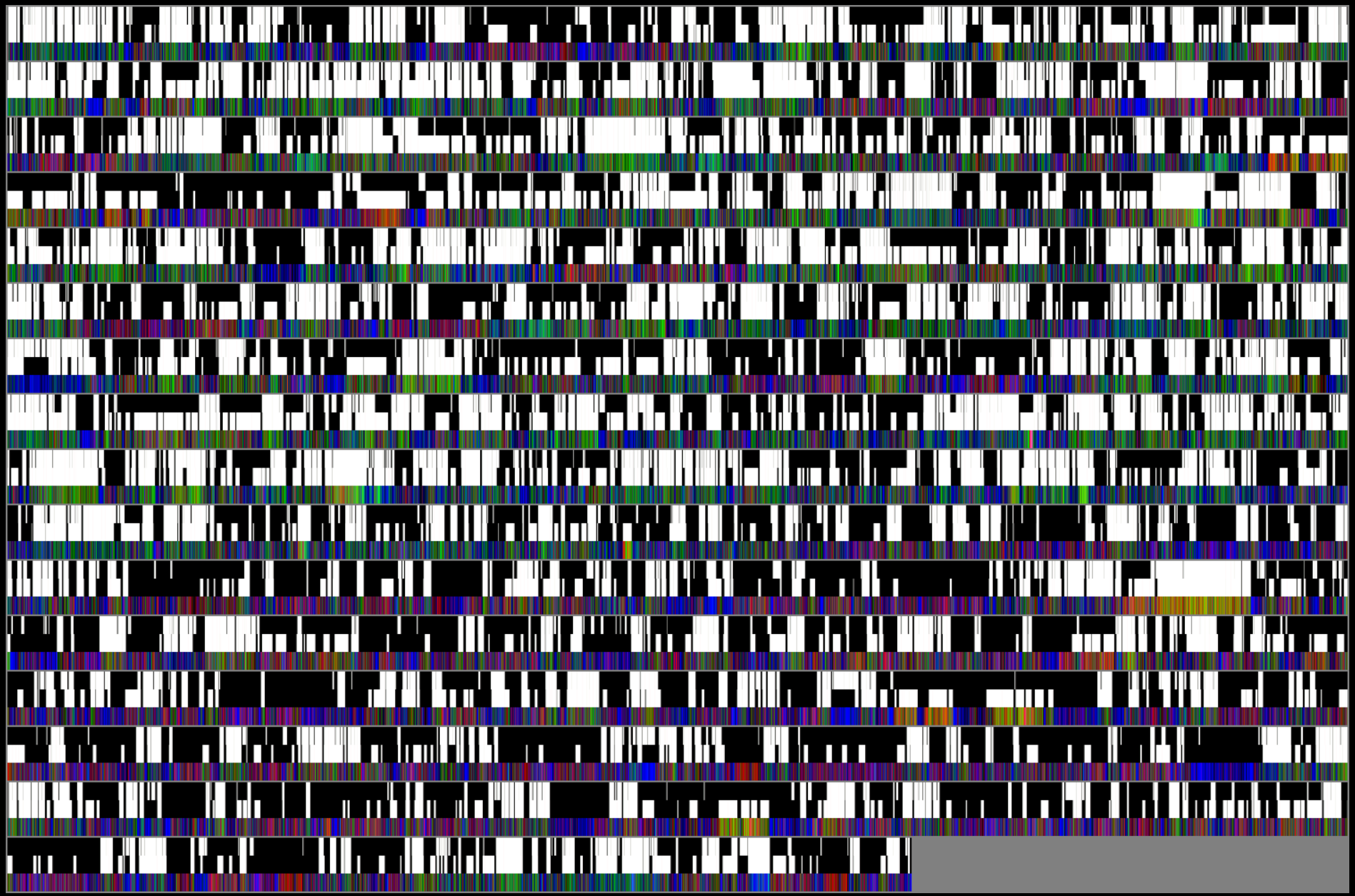
1. model – komplementární vlákno (0 genů)



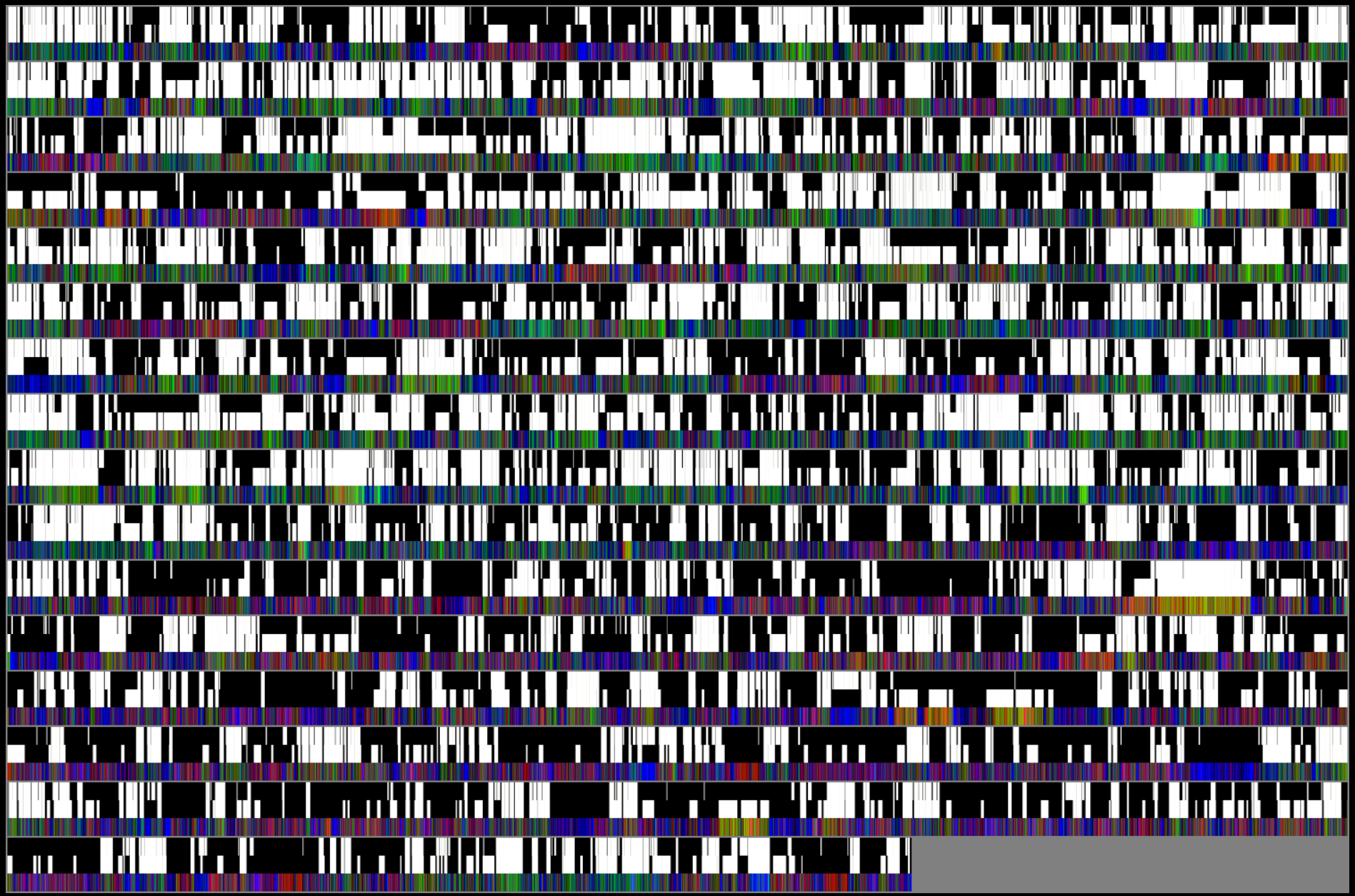
2. model – komplementární vlákno (170 genů)



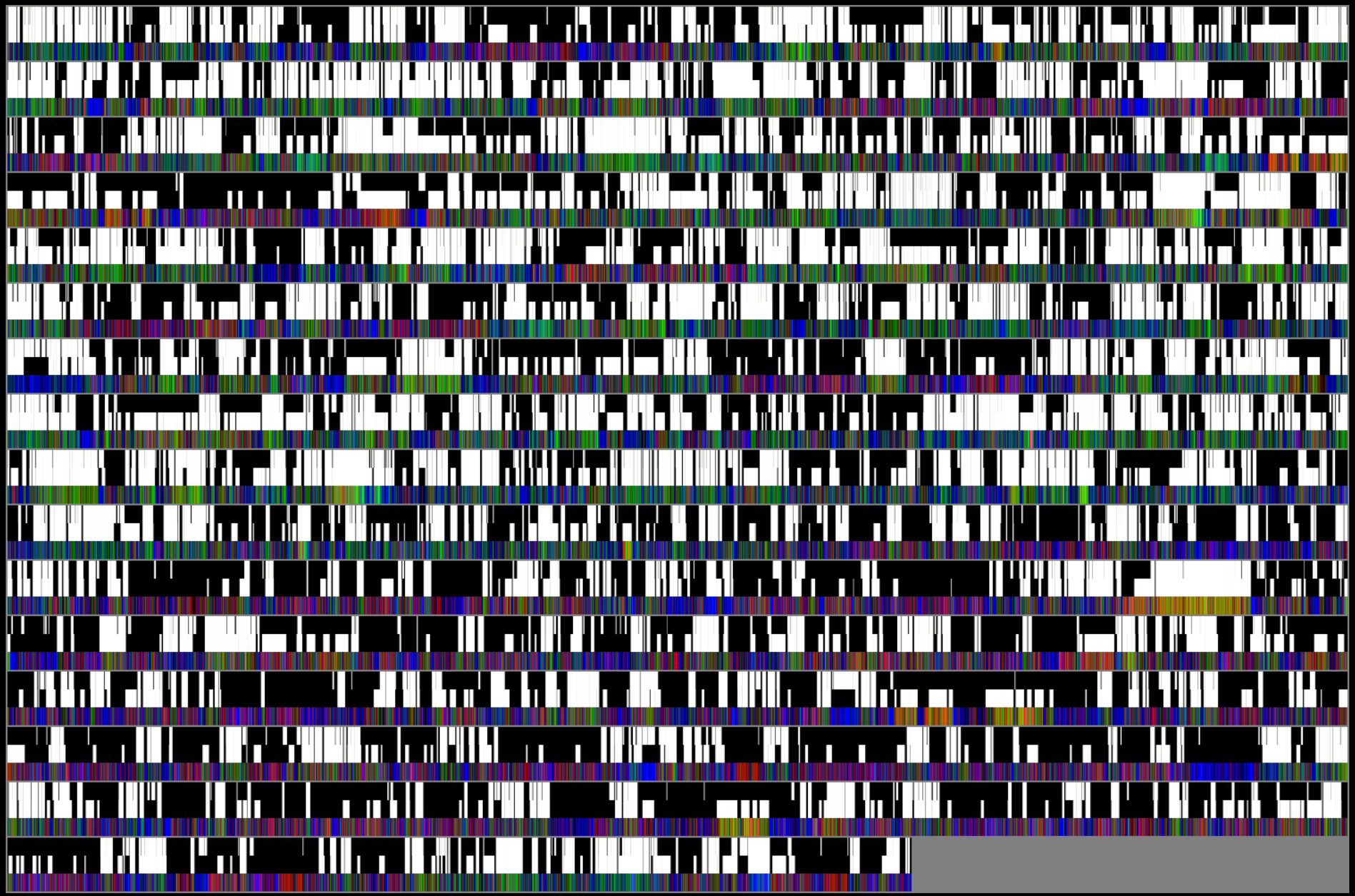
3. model – komplementární vlákno (208 genů)



4. model – komplementární vlákno (261 genů)



5. model – komplementární vlákno (432 genů)



Souhrn

	Primární vlákno	Komplementární vlákno
Délka	4 702 289	
Počet genů (délka)	2105 (1 949 511)	2115 (1 921 983)
Přesně označené geny		
1. model	0 (0)	0 (0)
2. model	187 (207 690)	170 (180 012)
3. model	216 (238 995)	208 (225 636)
4. model	278 (309 081)	261 (260 562)
5. model	427 (450 480)	432 (421 668)