Compression-based Machine Learning for DNA analysis

Diogo Pratas *et al.*

pratas@ua.pt

http://sweet.ua.pt/pratas

IEETA
UNIVERSITY OF AVEIRO

DATA SCIENCE PORTUGAL

DS PORTUGAL

What is compression ?

The process of reducing storage

What is relative compression ?

The process of reducing storage using exclusively a reference

REFERENCE: THIS IS MORE THAN A WORD

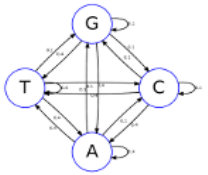
TARGETS:

THAT IS MORE THAN A WORD
THESE ARE WORDS
THOSE ARE SWORDS
ARE YOU READING THIS ?

REFERENCE → **MODEL OF THE REFERENCE** → **TARGET**

$$\log P(x^n|k) = \sum_{i=1}^{n-1} \log P(x_i|k, x^{i-1}) + \log P(x_n|k, x^{n-1})$$

$$\log p_{k,n} = \gamma \log p_{k,n-1} + \log P(x_n|k, x^{n-1}),$$



$$p_{k,n} = p_{k,n-1}^{\gamma} P(x_n|k, x^{n-1})$$

$$w_{k,n} = \frac{p_{k,n}}{\sum_{k \in \mathcal{K}} p_{k,n}}.$$

Genomic (DNA) sequences are large codified messages, from an alphabet of four symbols $\Theta = \{A, C, G, T\}$, describing most of the structure of all known living organisms. A huge amount of genomic data has been generated, with diverse characteristics, rendering the data deluge phenomenon a serious problem in most genomics centers. As such, most of the data are discarded (when possible), while other are compressed using general purpose algorithms, often attaining modest data reduction results.

Several specific algorithms have been proposed for the compression of genomic sequences, such as [1–7], but only some of them have been made available as usable and reliable compression tools, and those have been developed to some specific purpose or data characteristic [8–11].

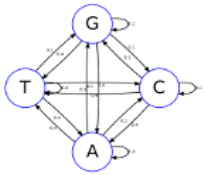
The dramatic increase of sequenced genomes, given the reduced sequencing costs, and the high redundancy characteristics, led to the development of genomic reference sequence compression. Several compressors of this type have been proposed [12–16], although most of them seem to be less efficient in handling sequences with higher rates of mutation.

MODEL OF THE REFERENCE

HOW WELL DOES IT LEARN ABOUT THE REFERENCE ?

$$\log P(x^n|k) = \sum_{i=1}^{n-1} \log P(x_i|k, x^{i-1}) + \log P(x_n|k, x^{n-1})$$

$$\log p_{k,n} = \gamma \log p_{k,n-1} + \log P(x_n|k, x^{n-1}),$$



$$p_{k,n} = p_{k,n-1}^{\gamma} P(x_n|k, x^{n-1})$$

$$w_{k,n} = \frac{p_{k,n}}{\sum_{k \in \mathcal{K}} p_{k,n}}.$$

- COMPUTATIONAL TIME
- COMPUTATIONAL MEMORY
- COMPRESSION RATIO

WE NEED EFFICIENT MODELS TO REPRESENT THE REFERENCE

MARKOV MODEL (FINITE-CONTEXT MODEL)

→ **A model that provides probability estimates that depend on the recent past of the sequence**

Portuguese Football League winners 2002-2016:

9-Porto
5-Benfica
1-Sporting

Who is going to win in 2017?

2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
S	P	P	B	P	P	P	P	B	P	P	P	B	B	B	?

CONTEXT K=1

	P	B	S	TOTAL
P	6	3	0	9
B	2	3	0	5
S	1	0	0	1
TOTAL	10	6	0	

2017?

Pr(P|B)=2/5

Pr(B|B)=3/5

Pr(S|B)=0

(CONTEXT=1)

2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
S	P	P	B	P	P	P	P	B	P	P	P	B	B	B	?

CONTEXT K=2

	P	B	S	Total
PP	3	3	0	6
PB	1	1	0	2
PS	0	0	0	0
BB	0	**1**	0	**1**
BP	1	0	0	1
BS	0	0	0	0
SS	0	0	0	0
SP	1	0	0	1
SB	0	0	0	0
Total	6	5	0	

2017?

Pr(P|BB)=0

Pr(B|BB)=1

Pr(S|BB)=0

(CONTEXT=2)

Which context Markov model is *better*? Context 1, 2, 3, 4, 5, ...?

Which context Markov model is *better*? Context 1, 2, 3, 4, 5, ...?

→ It depends on the nature and size of the trainning data

Therefore, we consider multiple Markov models using different contexts

Approaches:

1. Competition
2. Cooperation

Approaches:

1. Competition

2. Cooperation

1. COMPETITION

What is the best model for the next 10 years?

2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016

... SPPBPPPPBPPPBBB

PAST

2017 2018 2019 2020 2021 2022 2023 2024 2025 2026

BSPBBPBPPP CONTEXT 1
BPPPPPPBBP CONTEXT 2
BBSBBPBBPS CONTEXT 3
BBBBBBBBBB CONTEXT 5

BBBBBBBBBB REALITY

D. Pratas, A. J. Pinho. Compressing the human genome using exclusively Markov models. Proc. of the 5th Int. Conf. on Practical Applications of Computational Biology & Bioinformatics, PACBB 2011, Salamanca, Spain, vol. 93, p. 213-220, April 2011.

1. COMPETITION

What is the best model for the next 10 years?

2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016

... SPPBPPPPBPPPBBB

PAST

2017 2018 2019 2020 2021 2022 2023 2024 2025 2026

BSPBBPBPPP CONTEXT 1

BPPPPPPBBP CONTEXT 2

BBSBBPBBPS CONTEXT 3

BBBBBBBBBB CONTEXT 5

BBBBBBBBBB REALITY

A. J. Pinho, D. Pratas. MFCompress: a compression tool for FASTA and multi-FASTA data. Bioinformatics, vol. 30, no. 1, p. 117-118, January 2014


CLÁSSICO DA LUZ ESGOTADO E O DO RESTELO PARA LÁ CAMINHA
BENFICA LEVADO AO COLO


#colinho
BTV


GIOVANNI GALLI
ADVINCULA
BENFICA JOGOU COM 1 A MAIS
EM 12 JOGOS

Approaches:

1. Competition
2. Cooperation

2. COOPERATION

MARKOV MODEL CONTEXT **1**

MARKOV MODEL CONTEXT **2**

MARKOV MODEL CONTEXT **3**

MARKOV MODEL CONTEXT **5**

MIXTURE (Maestro)

EACH MODEL IS WEIGHTED ACCORDING TO ITS PERFORMANCE ALONG THE TIME

A. J. Pinho, D. Pratas, P. J. S. G. Ferreira. Bacteria DNA sequence compression using a mixture of finite-context models. Proc. of the IEEE Workshop on Statistical Signal Processing, SSP 2011, Nice, France, p. 125-128, June 2011

Generally cooperation works better than competition…

specially in DNA sequences

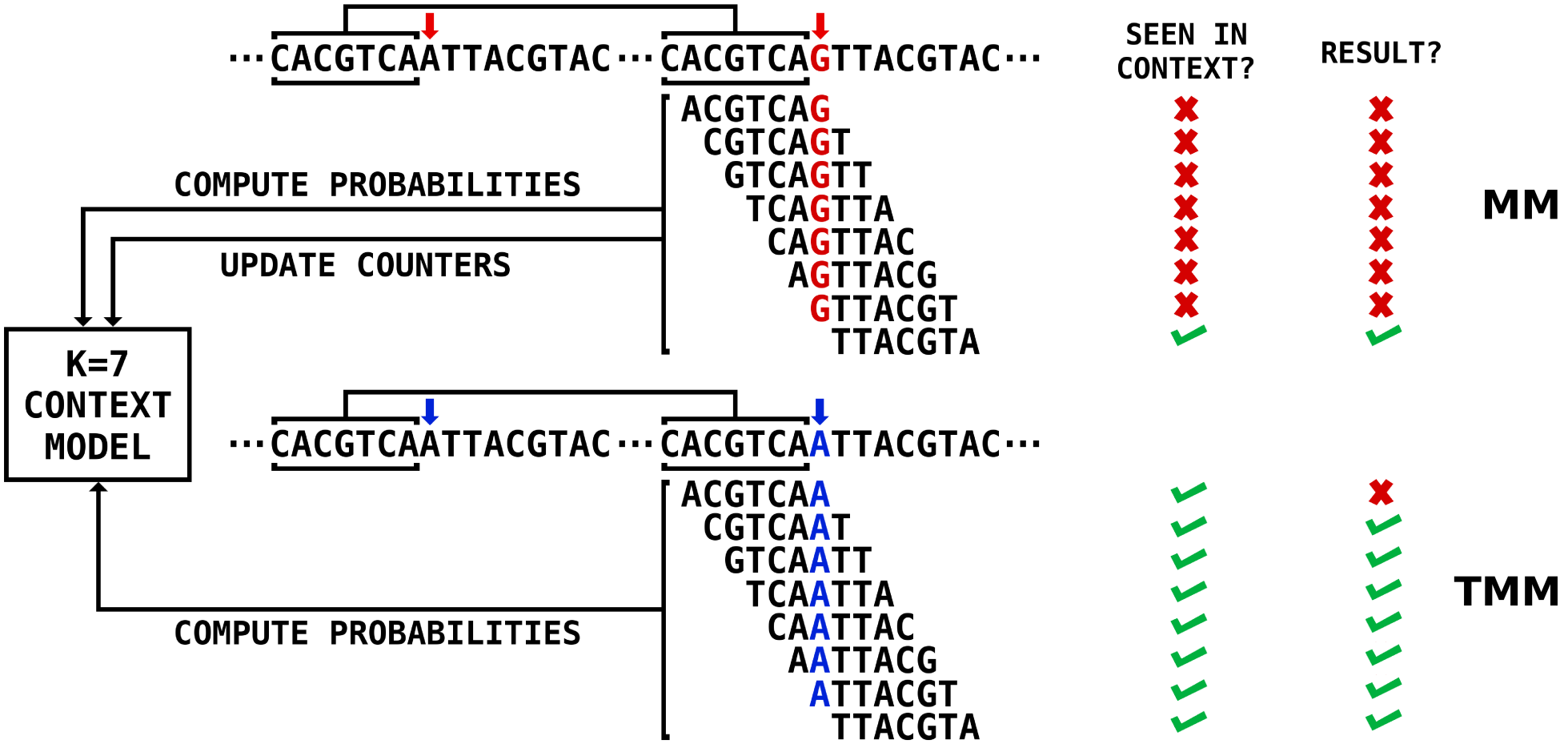
Tolerant Markov Model (TMM)

→ a probabilistic-algorithmic Markov model

It assigns probabilities according to a conditioning context that considers the last symbol, from the sequence to occur, as the most probable, given the occurrences stored in the memory, such as those from y, instead of the true occurring symbol.

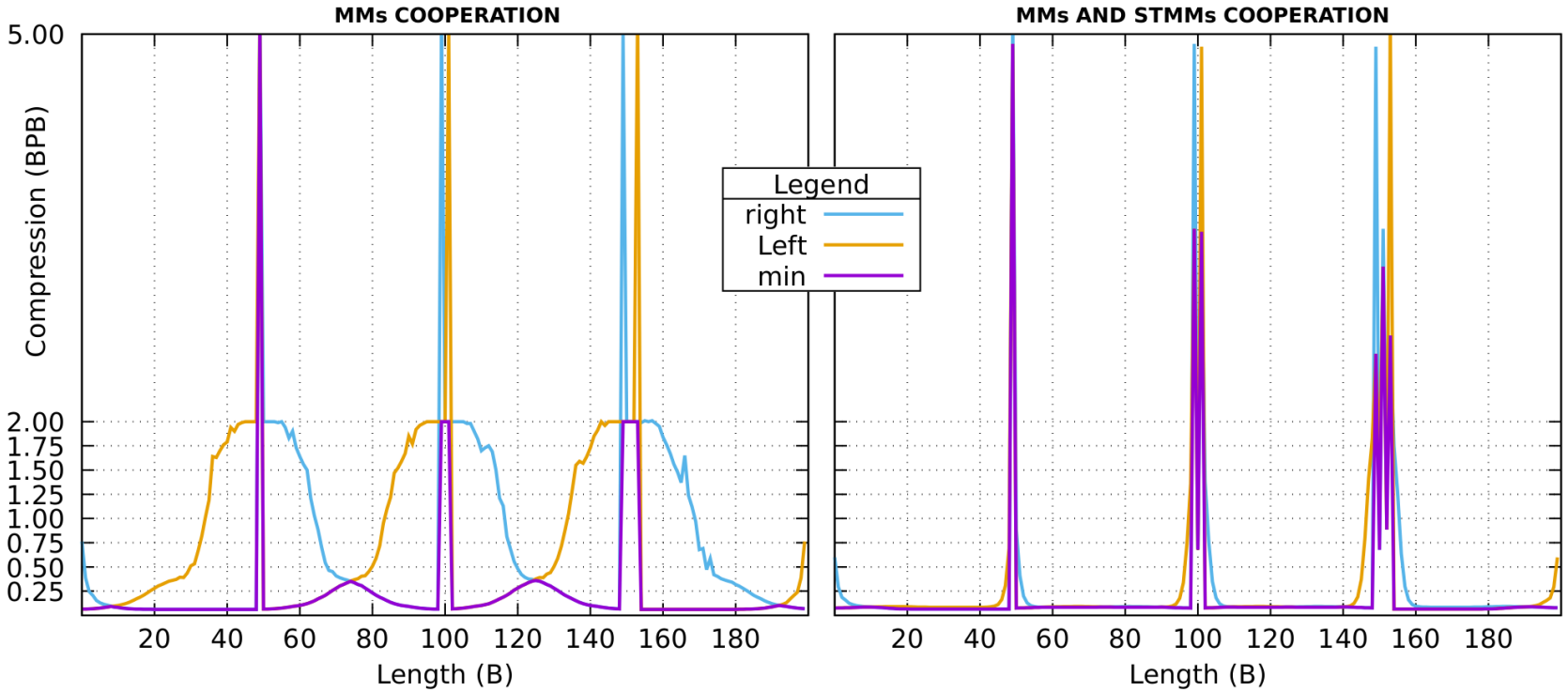
D. Pratas, A. J. Pinho, P. J. S. G. Ferreira. Efficient compression of genomic sequences. Proc. of the Data Compression Conference, DCC-2016, Snowbird, UT, p. 231-240, March 2016.

Tolerant Markov Model (TMM)



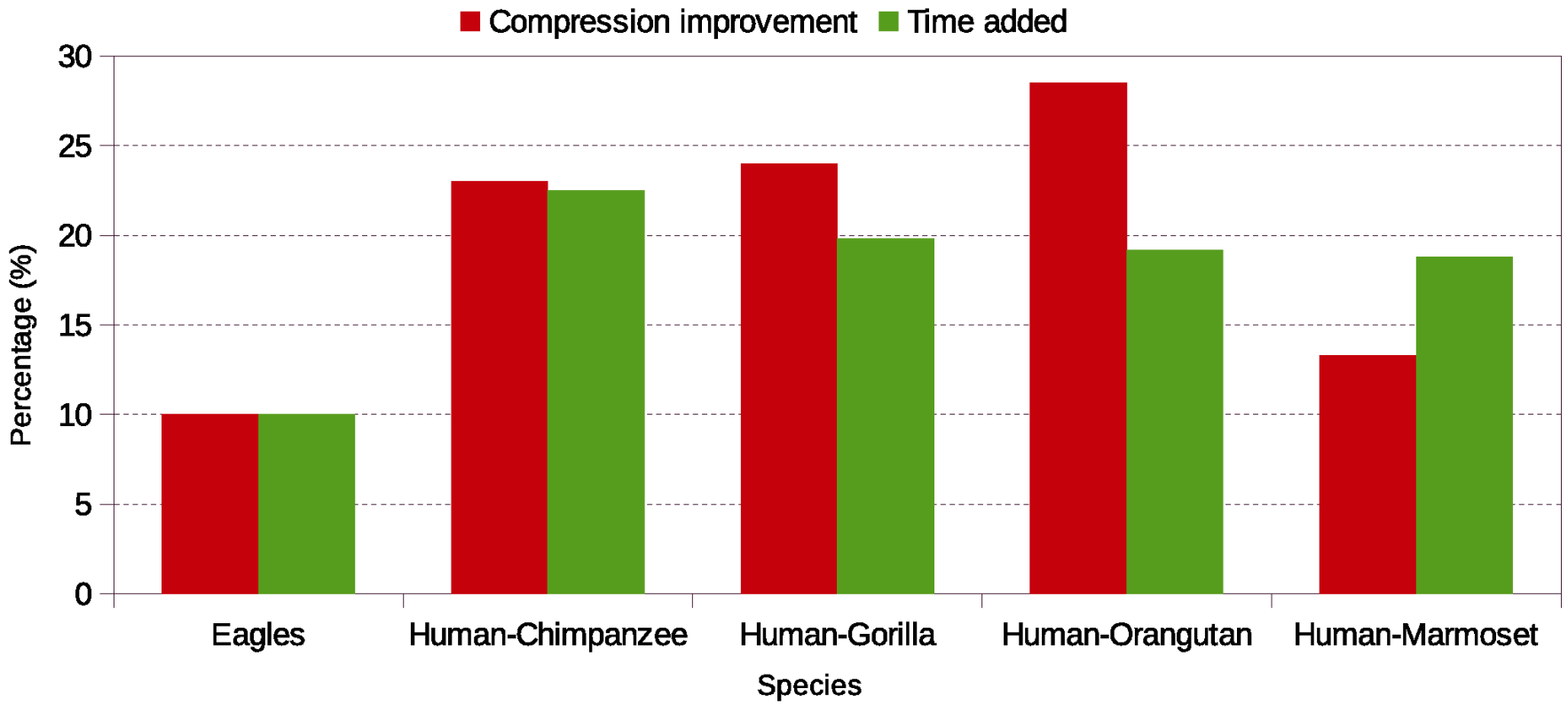
D. Pratas, A. J. Pinho, P. J. S. G. Ferreira. Efficient compression of genomic sequences. Proc. of the Data Compression Conference, DCC-2016, Snowbird, UT, p. 231-240, March 2016.

Tolerant Markov Model (TMM)



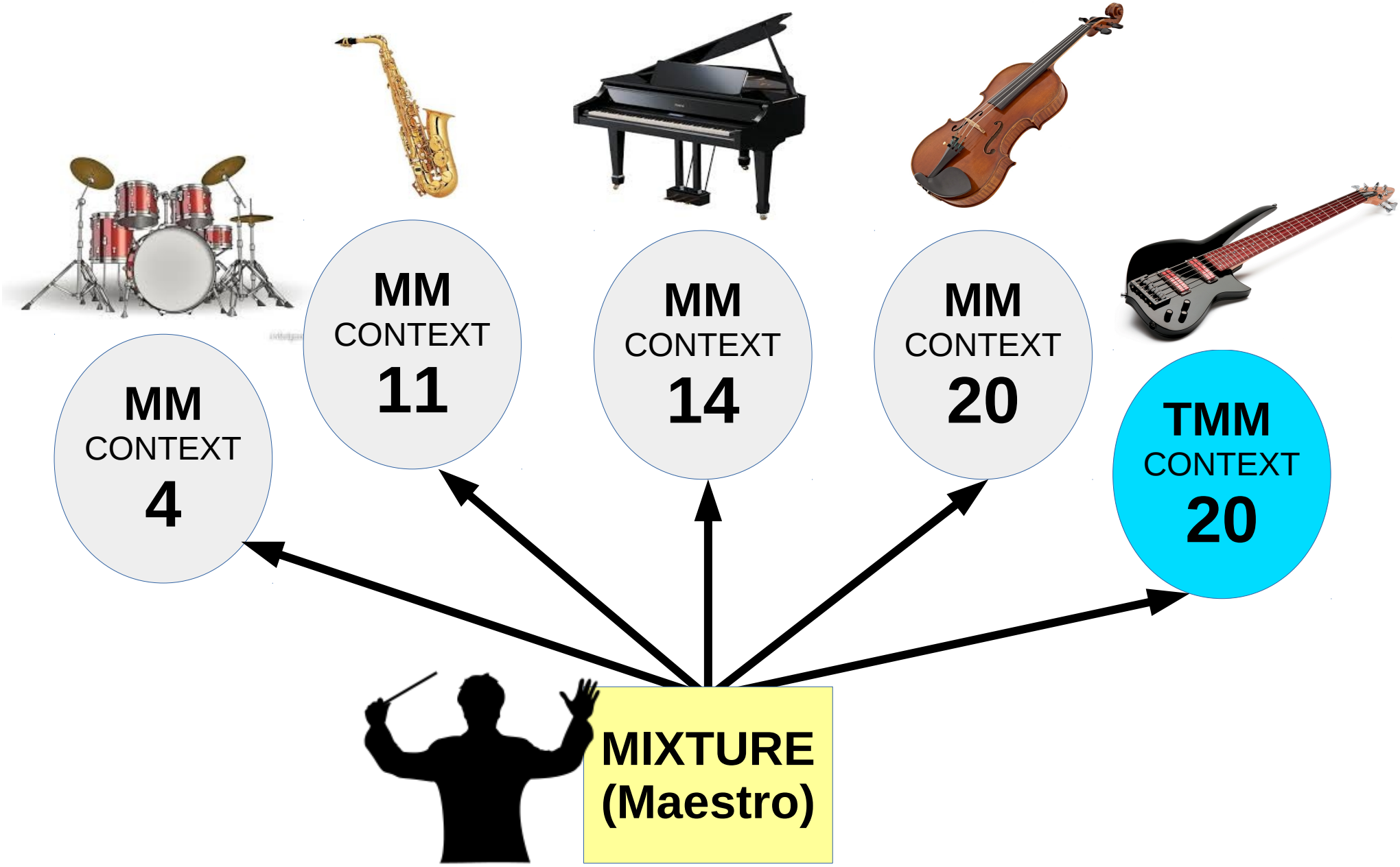
D. Pratas, M. Hosseini, A. J. Pinho. Substitutional tolerant Markov models for relative compression of DNA sequences. PACBB, 2017.

Tolerant Markov Model (TMM)



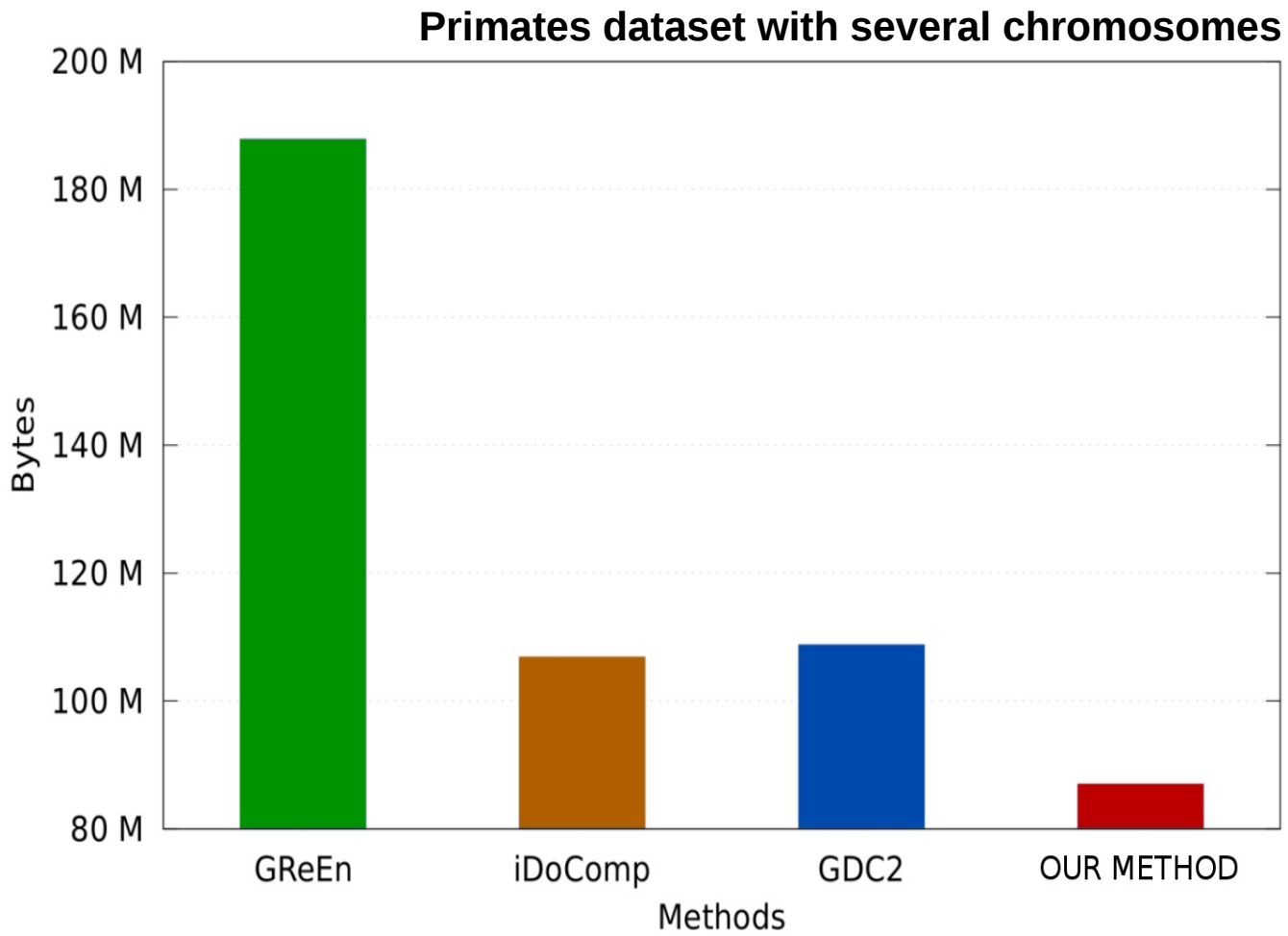
D. Pratas, M. Hosseini, A. J. Pinho. Substitutional tolerant Markov models for relative compression of DNA sequences. PACBB, 2017.

DNA relative compressor



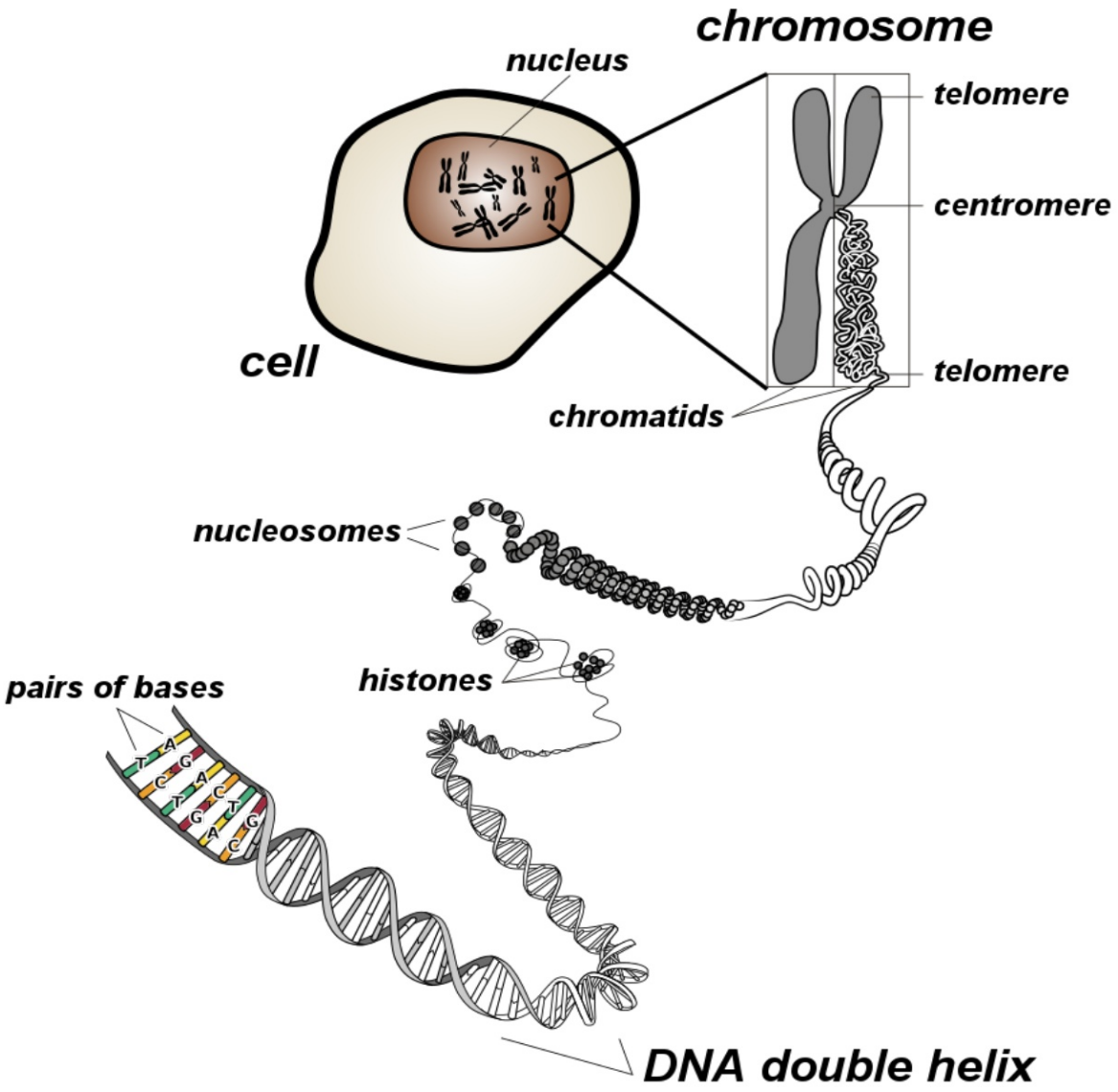
We use this compressor to measure relative similarity

DNA relative compressor



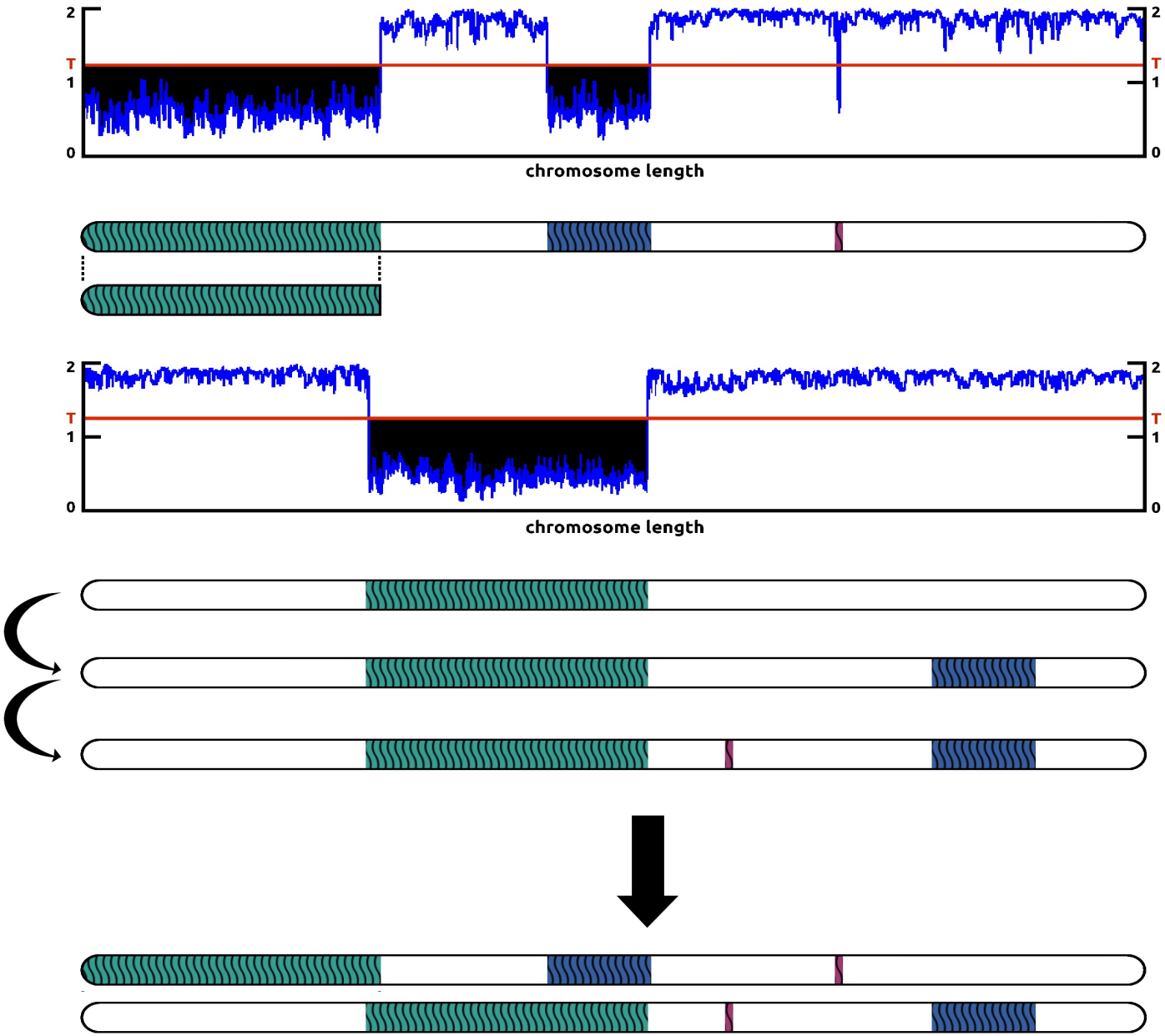
D. Pratas, A. J. Pinho, P. J. S. G. Ferreira. Efficient compression of genomic sequences. Proc. of the Data Compression Conference, DCC-2016, Snowbird, UT, p. 231-240, March 2016.

DNA sequences



DNA sequences

Similar regions detection

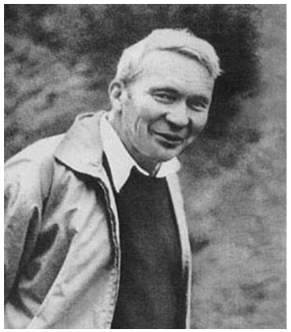


Rearrangements detection

Chromosome

5

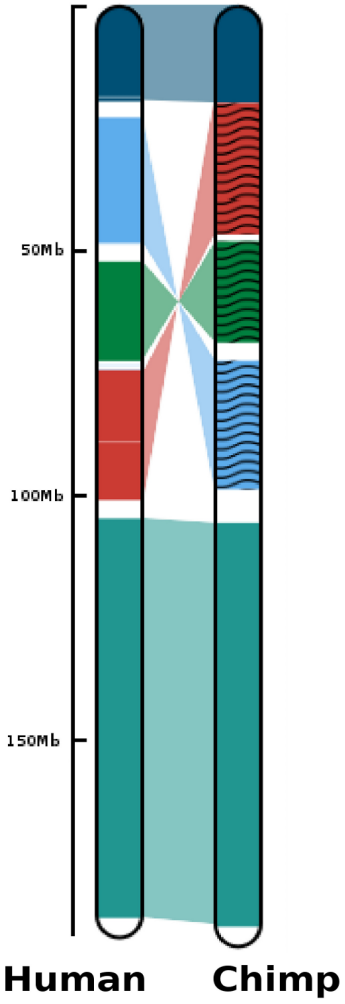
Human



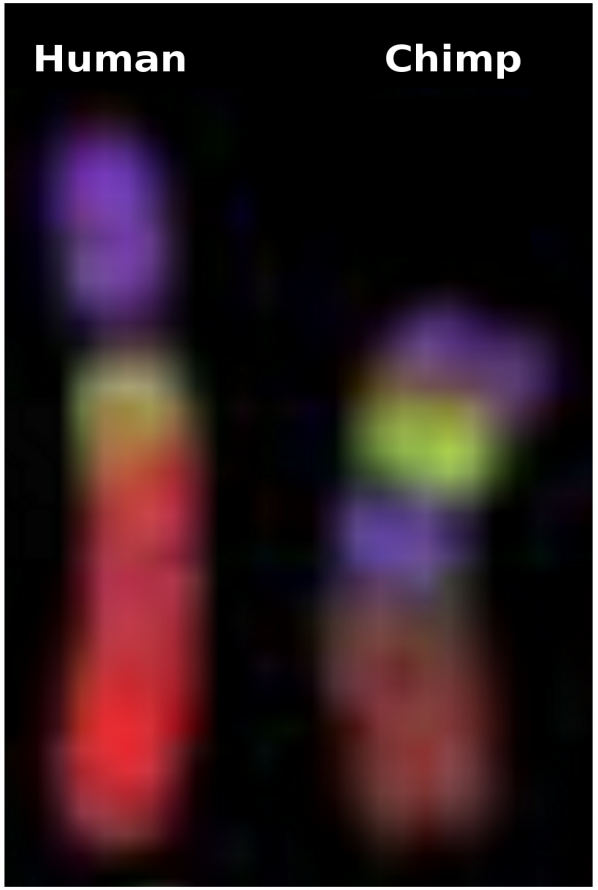
Vs

Chimp

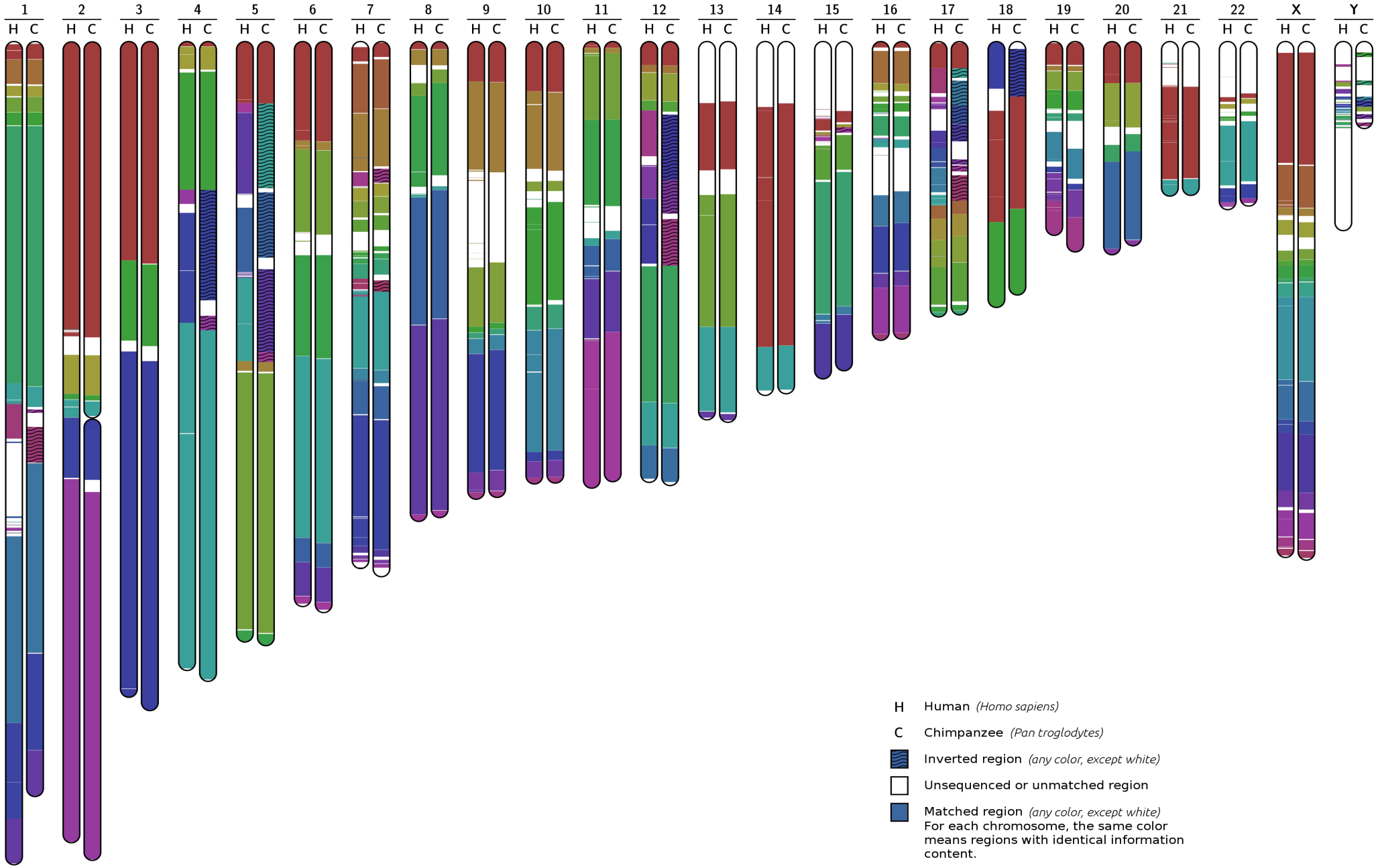
Computational approach



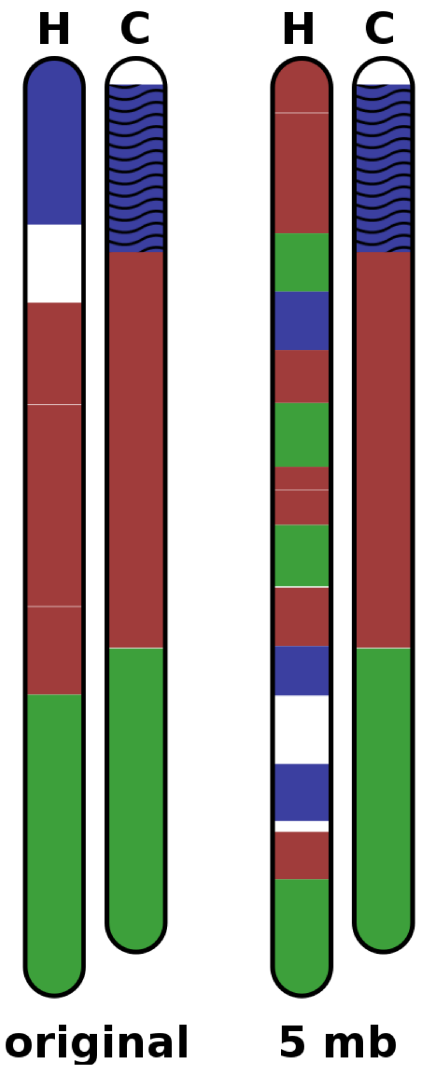
FISH (biological approach)



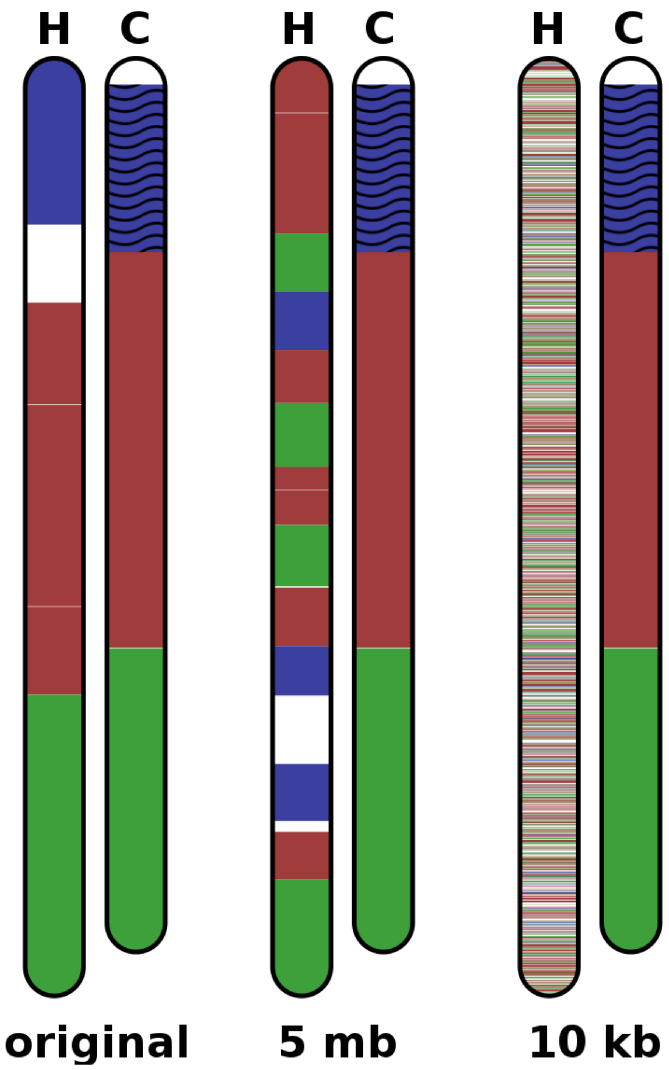
D. Pratas, R. Silva, A. J. Pinho, P. J. S. G. Ferreira. An alignment-free method to find and visualise rearrangements between pairs of DNA sequences. Scientific Reports, vol. 5, p. 10203, May 2015.



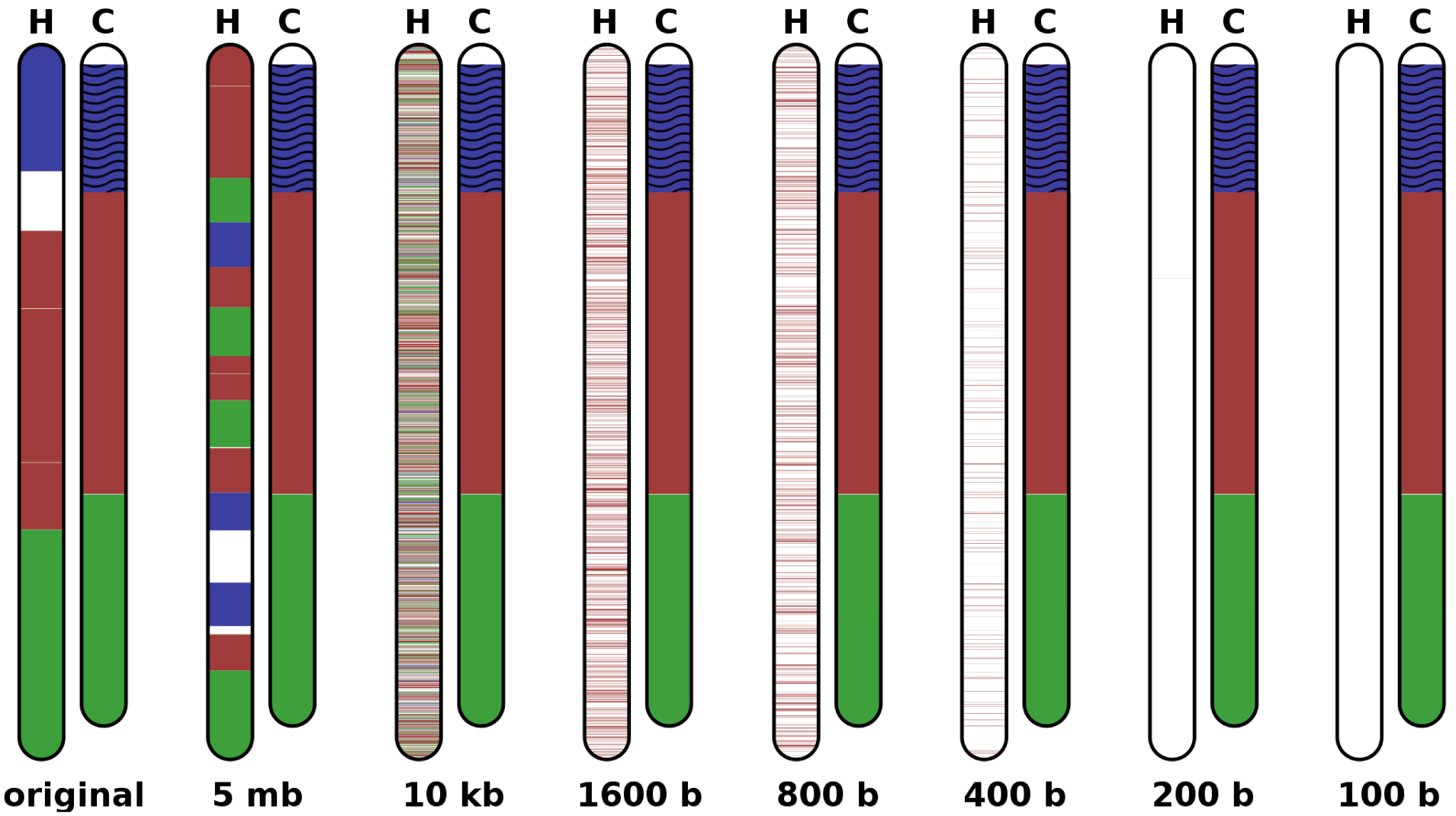
D. Pratas, R. Silva, A. J. Pinho, P. J. S. G. Ferreira. An alignment-free method to find and visualise rearrangements between pairs of DNA sequences. Scientific Reports, vol. 5, p. 10203, May 2015.



D. Pratas, R. Silva, A. J. Pinho, P. J. S. G. Ferreira. An alignment-free method to find and visualise rearrangements between pairs of DNA sequences. Scientific Reports, vol. 5, p. 10203, May 2015.



D. Pratas, R. Silva, A. J. Pinho, P. J. S. G. Ferreira. An alignment-free method to find and visualise rearrangements between pairs of DNA sequences. Scientific Reports, vol. 5, p. 10203, May 2015.



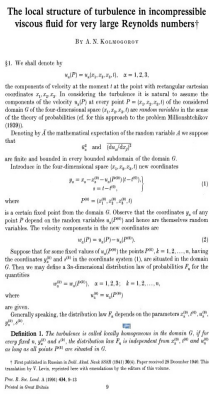
D. Pratas, R. Silva, A. J. Pinho, P. J. S. G. Ferreira. An alignment-free method to find and visualise rearrangements between pairs of DNA sequences. Scientific Reports, vol. 5, p. 10203, May 2015.

DNA extraction kit

FASTQ Reads

```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
@SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
BBCBBBBBB@@BAB?BBBBCBC>BBBAA8>BBBAA@
```

DNA Sequence



The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers†

By A. N. Kolmogorov

§1. We shall denote by

$$u_\alpha(P) = u_\alpha(x_1, x_2, x_3, t), \quad \alpha = 1, 2, 3,$$

the components of velocity at the moment t at the point with rectangular cartesian coordinates x_1, x_2, x_3. By considering the turbulence it is natural to assume the components of the velocity $u_\alpha(P)$ at every point $P = (x_1, x_2, x_3, t)$ of the considered domain G of the four-dimensional space (x_1, x_2, x_3, t) are random variables in the sense of the theory of probabilities (cf. for this approach to the problem Millionshchikov (1939)).

Denoting by $\bar{A}$ the mathematical expectation of the random variable A we suppose that

$$u_\alpha^2 \quad \text{and} \quad \overline{(\delta u_\alpha^r / \delta x_\beta)^2}$$

are finite and bounded in every bounded subdomain of the domain G.

Introduce in the four-dimensional space (x_1, x_2, x_3, t) new coordinates

$$\left.\begin{aligned} y_\alpha &= x_\alpha - x_\alpha^{(0)} - u_\alpha(P^{(0)})(t - t^{(0)}), \\ \tau &= t - t^{(0)}, \end{aligned}\right\} \tag{1}$$

where

$$P^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, t)$$

is a certain fixed point from the domain G. Observe that the coordinates y_α of any point P depend on the random variables $u_\alpha(P^{(0)})$ and hence are themselves random variables. The velocity components in the new coordinates are

$$v_\alpha(P) = u_\alpha(P) - u_\alpha(P^{(0)}). \tag{2}$$

Suppose that for some fixed values of α, β the points $P^{(k)}$, $k = 1, 2, \ldots, n$, having the coordinates $y_\alpha^{(k)}$ and $\tau^{(k)}$ in the coordinate system (1), are situated in the domain G. Then we may define a $3n$-dimensional distribution law of probabilities F_n for the quantities

$$v_\alpha^{(k)} = v_\alpha(P^{(k)}), \quad \alpha = 1, 2, 3; \quad k = 1, 2, \ldots, n,$$

where

$$v_\alpha^{(k)} = v_\alpha(P^{(k)})$$

are given.

Generally speaking, the distribution law F_n depends on the parameters $x_\alpha^{(0)}, t^{(0)}, u_\alpha^{(0)}, y_\alpha^{(k)}, \tau^{(k)}$.

Definition 1. *The turbulence is called locally homogeneous in the domain G, if for every fixed u, $y_\alpha^{(k)}$ and $\tau^{(k)}$, the distribution law F_n is independent from $x_\alpha^{(0)}$, $t^{(0)}$ and $u_\alpha^{(0)}$ as long as all points $P^{(k)}$ are situated in G.*

† First published in Russian in *Dokl. Akad. Nauk SSSR* (1941) **30**(4). Paper received 28 December 1940. This translation by V. Levin, reprinted here with emendations by the editors of this volume.

Proc. R. Soc. Lond. A (1991) **434**, 9–13
Printed in Great Britain

9

Assemble reads

The local structure of turbulence in incompressible viscous fluid for large Reynolds numbers†

By A. N. Kolmogorov

§1. We shall denote by

$$u_\alpha(P) = u_\alpha(x_1, x_2, x_3, t), \quad \alpha = 1, 2, 3,$$

the components of velocity at the moment t at the point with rectangular cartesian coordinates x_1, x_2, x_3. In considering the turbulence it is natural to assume the components of the velocity $u_\alpha(P)$ at every point $P = (x_1, x_2, x_3, t)$ of the considered domain G of the four-dimensional space (x_1, x_2, x_3, t) are *random variables* in the sense of the theory of probabilities (cf. for this approach to the problem Millionshtchikov (1939)).

Denoting by $\bar{A}$ the mathematical expectation of the random variable A we suppose that

$$\bar{u}_\alpha^2 \quad \text{and} \quad \overline{(\mathrm{d}u_\alpha/\mathrm{d}x_\beta)^2}$$

are finite and bounded in every bounded subdomain of the domain G.

Introduce in the four-dimensional space (x_1, x_2, x_3, t) new coordinates

$$\left.\begin{aligned} y_\alpha &= x_\alpha - x_\alpha^{(0)} - u_\alpha(P^{(0)})(t - t^{(0)}), \\ s &= t - t^{(0)}, \end{aligned}\right\} \tag{1}$$

where $P^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, t)$

is a certain fixed point from the domain G. Observe that the coordinates y_α of any point P depend on the random variables $u_\alpha(P^{(0)})$ and hence are themselves random variables. The velocity components in the new coordinates are

$$w_\alpha(P) = u_\alpha(P) - u_\alpha(P^{(0)}). \tag{2}$$

Suppose that for some fixed values of $u_\alpha(P^{(0)})$ the points $P^{(k)}, k = 1, 2, \ldots, n$, having the coordinates $y_\alpha^{(k)}$ and $s^{(k)}$ in the coordinate system (1), are situated in the domain G. Then we may define a $3n$-dimensional distribution law of probabilities F_n for the quantities

$$w_\alpha^{(k)} = w_\alpha(P^{(k)}), \quad \alpha = 1, 2, 3; \quad k = 1, 2, \ldots, n,$$

where $u_\alpha^{(0)} = u_\alpha(P^{(0)})$

are given.

Generally speaking, the distribution law F_n depends on the parameters $x_\alpha^{(0)}, t^{(0)}, u_\alpha^{(0)}, y_\alpha^{(k)}, s^{(k)}$.

Definition 1. *The turbulence is called locally homogeneous in the domain G, if for every fixed n, $y_\alpha^{(k)}$ and $s^{(k)}$, the distribution law F_n is independent from $x_\alpha^{(0)}$, $t^{(0)}$ and $u_\alpha^{(0)}$ as long as all points $P^{(k)}$ are situated in G.*

† First published in Russian in *Dokl. Akad. Nauk SSSR* (1941) **30**(4). Paper received 28 December 1940. This translation by V. Levin, reprinted here with emendations by the editors of this volume.

The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers†

By A. N. Kolmogorov

§1. We shall denote by

$$u_\alpha(P) = u_\alpha(x_1, x_2, x_3, t), \quad \alpha = 1, 2, 3,$$

the components of velocity at the moment t at the point with rectangular cartesian coordinates x_1, x_2, x_3. In considering the turbulence it is natural to assume the components of the velocity $u_\alpha(P)$ at every point $P = (x_1, x_2, x_3, t)$ of the considered domain G of the four-dimensional space (x_1, x_2, x_3, t) are *random variables* in the sense of the theory of probabilities (cf. for this approach to the problem Millionshtchikov (1939)).

Denoting by $\bar{A}$ the mathematical expectation of the random variable A we suppose that

$$\bar{u}_\alpha^2 \quad \text{and} \quad \overline{(\mathrm{d}u_\alpha/\mathrm{d}x_\beta)^2}$$

are finite and bounded in every bounded subdomain of the domain

Introduce in the four-dimensional space (x_1, x_2, x_3, t) new coordinates

$$\left.\begin{aligned} y_\alpha &= x_\alpha - x_\alpha^{(0)} - u_\alpha(P^{(0)})(t - t^{(0)}), \\ s &= t - t^{(0)}, \end{aligned}\right\} \tag{1}$$

where $P^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, t)$

is a certain fixed point from the domain G. Observe that the coordinates y_α of any point P depend on the random variables $u_\alpha(P^{(0)})$ and hence are themselves random variables. The velocity components in the new coordinates are

$$w_\alpha(P) = u_\alpha(P) - u_\alpha(P^{(0)}). \tag{2}$$

Suppose that for some fixed values of $u_\alpha(P^{(0)})$ the points $P^{(k)}, k = 1, 2, \ldots, n$, having the coordinates $y_\alpha^{(k)}$ and $s^{(k)}$ in the coordinate system (1), are situated in the domain G. Then we may define a $3n$-dimensional distribution law of probabilities F_n for the quantities

$$w_\alpha^{(k)} = w_\alpha(P^{(k)}), \quad \alpha = 1, 2, 3; \quad k = 1, 2, \ldots, n,$$

where $u_\alpha^{(0)} =$

are given.

Generally speaking, the distribution law F_n depends on the parameters $x_\alpha^{(0)}, t^{(0)}, u_\alpha^{(0)}, y_\alpha^{(k)}, s^{(k)}$.

Definition 1. *The turbulence is called locally homogeneous in the domain G, if for every fixed n, $y_\alpha^{(k)}$ and $s^{(k)}$, the distribution law F_n is independent from $x_\alpha^{(0)}$, $t^{(0)}$ and $u_\alpha^{(0)}$ as long as all points $P^{(k)}$ are situated in G.*

† First published in Russian in *Dokl. Akad. Nauk SSSR* (1941) **30**(4). Paper received 28 December 1940. This translation by V. Levin, reprinted here with emendations by the editors of this volume.

Proc. R. Soc. Lond. A (1991) **434**, 9–13
Printed in Great Britain 13

The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers†

By A. N. Kolmogorov

§1. We shall denote by

$$u_\alpha(P) = u_\alpha(x_1, x_2, x_3, t), \quad \alpha = 1, 2, 3,$$

the components of velocity at the moment t at the point with rectangular cartesian coordinates x_1, x_2, x_3. In considering the turbulence it is natural to assume the components of the velocity $u_\alpha(P)$ at every point $P = (x_1, x_2, x_3, t)$ of the considered domain G of the four-dimensional space (x_1, x_2, x_3, t) are *random variables* in the sense of the theory of probabilities (cf. for this approach to the problem Millionshtchikov (1939)).

Denoting by $\bar{A}$ the mathematical expectation of the random variable A we suppose that

$$\bar{u}_\alpha^2 \quad \text{and} \quad \overline{(\mathrm{d}u_\alpha/\mathrm{d}x_\beta)^2}$$

are finite and bounded in every bounded subdomain of the domain

Introduce in the four-dimensional space (x_1, x_2, x_3, t) new coordinates

$$\left.\begin{aligned} y_\alpha &= x_\alpha - x_\alpha^{(0)} - u_\alpha(P^{(0)})(t - t^{(0)}), \\ s &= t - t^{(0)}, \end{aligned}\right\} \tag{1}$$

where

$$P^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, t)$$

is a certain fixed point from the domain G. Observe that the coordinates y_α of any point P depend on the random variables $u_\alpha(P^{(0)})$ and hence are themselves random variables. The velocity components in the new coordinates are

$$w_\alpha(P) = u_\alpha(P) - u_\alpha(P^{(0)}). \tag{2}$$

Suppose that for some fixed values of $u_\alpha(P^{(0)})$ the points $P^{(k)}, k = 1, 2, \ldots, n$, having the coordinates $y_\alpha^{(k)}$ and $s^{(k)}$ in the coordinate system (1), are situated in the domain G. Then we may define a $3n$-dimensional distribution law of probabilities F_n for the quantities

$$w_\alpha^{(k)} = w_\alpha(P^{(k)}), \quad \alpha = 1, 2, 3; \quad k = 1, 2, \ldots, n,$$

where

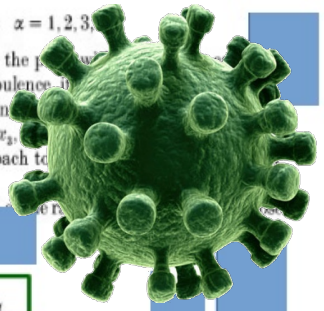
$$u_\alpha^{(0)} =$$

are given.

Generally speaking, the distribution law F_n depends on the parameters $x_\alpha^{(0)}, t^{(0)}, u_\alpha^{(0)}, y_\alpha^{(k)}, s^{(k)}$.

Definition 1. *The turbulence is called locally homogeneous in the domain G, if for every fixed n, $y_\alpha^{(k)}$ and $s^{(k)}$, the distribution law F_n is independent from $x_\alpha^{(0)}$, $t^{(0)}$ and $u_\alpha^{(0)}$ as long as all points $P^{(k)}$ are situated in G.*

† First published in Russian in *Dokl. Akad. Nauk SSSR* (1941) **30**(4). Paper received 28 December 1940. This translation by V. Levin, reprinted here with emendations by the editors of this volume.



The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers†

By A. N. Kolmogorov

§1. We shall denote by

$$u_\alpha(P) = u_\alpha(x_1, x_2, x_3, t), \quad \alpha = 1, 2, 3,$$

the components of velocity at the moment t at the point with rectangular cartesian coordinates x_1, x_2, x_3. In considering the turbulence it is natural to assume the components of the velocity $u_\alpha(P)$ at every point $P = (x_1, x_2, x_3, t)$ of the considered domain G of the four-dimensional space (x_1, x_2, x_3, t) are *random variables* in the sense of the theory of probabilities (cf. for this approach to the problem Millionshtchikov (1939)).

Denoting by $\bar{A}$ the mathematical expectation of the random variable A we suppose that

$$\bar{u}_\alpha^2 \quad \text{and} \quad \overline{(\mathrm{d}u_\alpha/\mathrm{d}x_\beta)^2}$$

are finite and bounded in every bounded subdomain of the domain

Introduce in the four-dimensional space (x_1, x_2, x_3, t) new coordinates

$$\left.\begin{aligned} y_\alpha &= x_\alpha - x_\alpha^{(0)} - u_\alpha(P^{(0)})(t - t^{(0)}), \\ s &= t - t^{(0)}, \end{aligned}\right\} \tag{1}$$

where

$$P^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, t)$$

is a certain fixed point from the domain G. Observe that the coordinates y_α of any point P depend on the random variables $u_\alpha(P^{(0)})$ and hence are themselves random variables. The velocity components in the new coordinates are

$$w_\alpha(P) = u_\alpha(P) - u_\alpha(P^{(0)}). \tag{2}$$

Suppose that for some fixed values of $u_\alpha(P^{(0)})$ the points $P^{(k)}, k = 1, 2, \ldots, n$, having the coordinates $y_\alpha^{(k)}$ and $s^{(k)}$ in the coordinate system (1), are situated in the domain G. Then we may define a $3n$-dimensional distribution law of probabilities F_n for the quantities

$$w_\alpha^{(k)} = w_\alpha(P^{(k)}), \quad \alpha = 1, 2, 3; \quad k = 1, 2, \ldots, n,$$

where

$$u_\alpha^{(0)} =$$

are given.

Generally speaking, the distribution law F_n depends on the parameters $x_\alpha^{(0)}, t^{(0)}, u_\alpha^{(0)}, y_\alpha^{(k)}, s^{(k)}$.

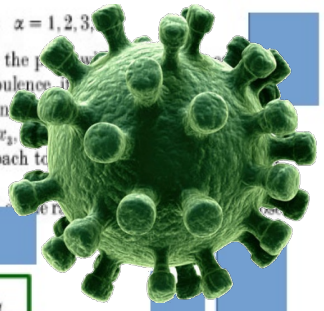
Definition 1. *The turbulence is called locally homogeneous in the domain G, if for every fixed n, $y_\alpha^{(k)}$ and $s^{(k)}$, the distribution law F_n is independent from $x_\alpha^{(0)}$, $t^{(0)}$ and $u_\alpha^{(0)}$ as long as all points $P^{(k)}$ are situated in G.*

† First published in Russian in *Dokl. Akad. Nauk SSSR* (1941) **30**(4). Paper received 28 December 1940. This translation by V. Levin, reprinted here with emendations by the editors of this volume.

Proc. R. Soc. Lond. A (1991) **434**, 9–13
Printed in Great Britain

Ancient DNA

Much more complex...



The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers†

By A. N. Kolmogorov

§1. We shall denote by

$$u_\alpha(P) = u_\alpha(x_1, x_2, x_3, t), \quad \alpha = 1, 2, 3,$$

the components of velocity at the moment t at the point with rectangular cartesian coordinates x_1, x_2, x_3. In considering the turbulence it is natural to assume the components of the velocity $u_\alpha(P)$ at every point $P = (x_1, x_2, x_3, t)$ of the considered domain G of the four-dimensional space (x_1, x_2, x_3, t) are *random variables* in the sense of the theory of probabilities (cf. for this approach to the problem Millionshtchikov (1939)).

Denoting by $\bar{A}$ the mathematical expectation of the random variable A we suppose that

$$\bar{u}_\alpha^2 \quad \text{and} \quad \overline{(\mathrm{d}u_\alpha/\mathrm{d}x_\beta)^2}$$

are finite and bounded in every bounded subdomain of the domain

Introduce in the four-dimensional space (x_1, x_2, x_3, t) new coordinates

$$\left.\begin{aligned} y_\alpha &= x_\alpha - x_\alpha^{(0)} - u_\alpha(P^{(0)})(t - t^{(0)}), \\ s &= t - t^{(0)}, \end{aligned}\right\} \tag{1}$$

where

$$P^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, t)$$

is a certain fixed point from the domain G. Observe that the coordinates y_α of any point P depend on the random variables $u_\alpha(P^{(0)})$ and hence are themselves random variables. The velocity components in the new coordinates are

$$w_\alpha(P) = u_\alpha(P) - u_\alpha(P^{(0)}). \tag{2}$$

Suppose that for some fixed values of $u_\alpha(P^{(0)})$ the points $P^{(k)}, k = 1, 2, \ldots, n$, having the coordinates $y_\alpha^{(k)}$ and $s^{(k)}$ in the coordinate system (1), are situated in the domain G. Then we may define a $3n$-dimensional distribution law of probabilities F_n for the quantities

$$w_\alpha^{(k)} = w_\alpha(P^{(k)}), \quad \alpha = 1, 2, 3; \quad k = 1, 2, \ldots, n,$$

where

$$u_\alpha^{(0)} =$$

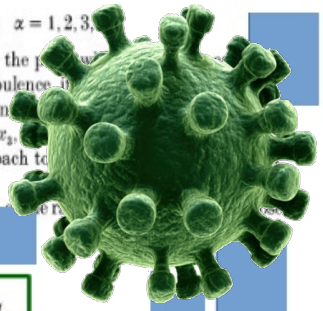
are given.

Generally speaking, the distribution law F_n depends on the parameters $x_\alpha^{(0)}, t^{(0)}, u_\alpha^{(0)}, y_\alpha^{(k)}, s^{(k)}$.

Definition 1. *The turbulence is called locally homogeneous in the domain G, if for every fixed n, $y_\alpha^{(k)}$ and $s^{(k)}$, the distribution law F_n is independent from $x_\alpha^{(0)}$, $t^{(0)}$ and $u_\alpha^{(0)}$ as long as all points $P^{(k)}$ are situated in G.*

† First published in Russian in *Dokl. Akad. Nauk SSSR* (1941) **30**(4). Paper received 28 December 1940. This translation by V. Levin, reprinted here with emendations by the editors of this volume.

Proc. R. Soc. Lond. A (1991) **434**, 9–13
Printed in Great Britain



The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers†

By A. N. Kolmogorov

§1. We shall denote by

$$u_\alpha(P) = u_\alpha(x_1, x_2, x_3, t), \quad \alpha = 1, 2, 3,$$

the components of velocity at the moment t at the point with rectangular cartesian coordinates x_1, x_2, x_3. In considering the turbulence it is natural to assume the components of the velocity $u_\alpha(P)$ at every point $P = (x_1, x_2, x_3, t)$ of the considered domain G of the four-dimensional space (x_1, x_2, x_3, t) are *random variables* in the sense of the theory of probabilities (cf. for this approach to the problem Millionshtchikov (1939)).

Denoting by $\bar{A}$ the mathematical expectation of the random variable A we suppose that

$$\bar{u}_\alpha^2 \quad \text{and} \quad \overline{(\mathrm{d}u_\alpha/\mathrm{d}x_\beta)^2}$$

are finite and bounded in every bounded subdomain of the domain

Introduce in the four-dimensional space (x_1, x_2, x_3, t) new coordinates

$$\left.\begin{aligned} y_\alpha &= x_\alpha - x_\alpha^{(0)} - u_\alpha(P^{(0)})(t - t^{(0)}), \\ s &= t - t^{(0)}, \end{aligned}\right\} \tag{1}$$

where $P^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, t)$

is a certain fixed point from the domain G. Observe that the coordinates y_α of any point P depend on the random variables $u_\alpha(P^{(0)})$ and hence are themselves random variables. The velocity components in the new coordinates are

$$w_\alpha(P) = u_\alpha(P) - u_\alpha(P^{(0)}). \tag{2}$$

Suppose that for some fixed values of $u_\alpha(P^{(0)})$ the points $P^{(k)}, k = 1, 2, \ldots, n$, having the coordinates $y_\alpha^{(k)}$ and $s^{(k)}$ in the coordinate system (1), are situated in the domain G. Then we may define a $3n$-dimensional distribution law of probabilities F_n for the quantities

$$w_\alpha^{(k)} = w_\alpha(P^{(k)}), \quad \alpha = 1, 2, 3; \quad k = 1, 2, \ldots, n,$$

where $u_\alpha^{(0)} =$

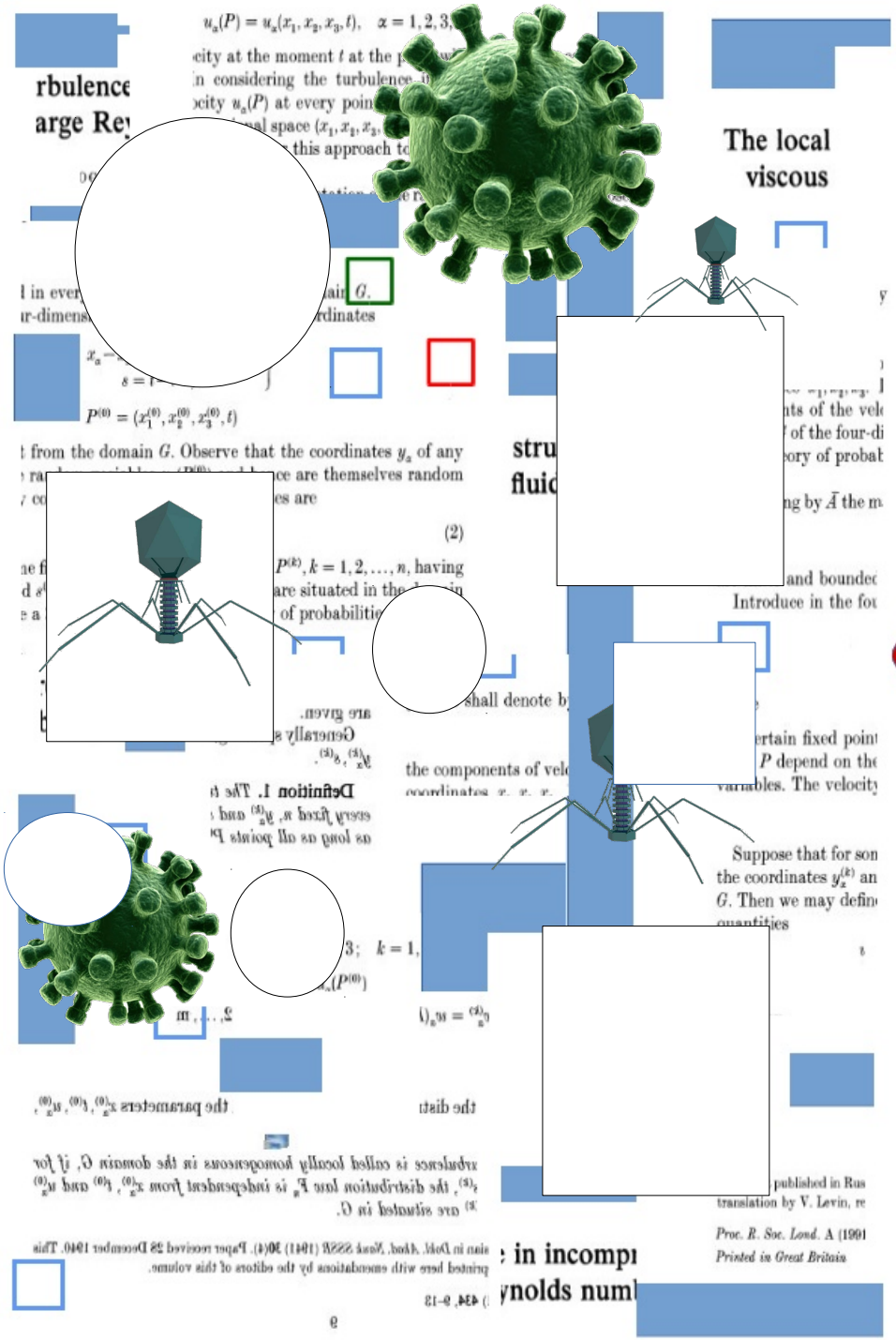
are given.

Generally speaking, the distribution law F_n depends on the parameters $x_\alpha^{(0)}, t^{(0)}, u_\alpha^{(0)}, y_\alpha^{(k)}, s^{(k)}$.

Definition 1. *The turbulence is called locally homogeneous in the domain G, if for every fixed n, $y_\alpha^{(k)}$ and $s^{(k)}$, the distribution law F_n is independent from $x_\alpha^{(0)}$, $t^{(0)}$ and $u_\alpha^{(0)}$ as long as all points $P^{(k)}$ are situated in G.*

† First published in Russian in *Dokl. Akad. Nauk SSSR* (1941) **30**(4). Paper received 28 December 1940. This translation by V. Levin, reprinted here with emendations by the editors of this volume.

Proc. R. Soc. Lond. A (1991) **434**, 9–13
Printed in Great Britain



The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers†

By A. N. Kolmogorov

§1. We shall denote by

$$u_\alpha(P) = u_\alpha(x_1, x_2, x_3, t), \quad \alpha = 1, 2, 3,$$

the components of velocity at the moment t at the point with rectangular cartesian coordinates x_1, x_2, x_3. In considering the turbulence it is natural to assume the components of the velocity $u_\alpha(P)$ at every point $P = (x_1, x_2, x_3, t)$ of the considered domain G of the four-dimensional space (x_1, x_2, x_3, t) are *random variables* in the sense of the theory of probabilities (cf. for this approach to the problem Millionshtchikov (1939)).

Denoting by $\bar{A}$ the mathematical expectation of the random variable A we suppose that

$$\bar{u}_\alpha^2 \quad \text{and} \quad \overline{(\mathrm{d}u_\alpha/\mathrm{d}x_\beta)^2}$$

are finite and bounded in every bounded subdomain of the domain

Introduce in the four-dimensional space (x_1, x_2, x_3, t) new coordinates

$$\left.\begin{aligned} y_\alpha &= x_\alpha - x_\alpha^{(0)} - u_\alpha(P^{(0)})(t - t^{(0)}), \\ s &= t - t^{(0)}, \end{aligned}\right\} \tag{1}$$

where

$$P^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, t)$$

is a certain fixed point from the domain G. Observe that the coordinates y_α of any point P depend on the random variables $u_\alpha(P^{(0)})$ and hence are themselves random variables. The velocity components in the new coordinates are

$$w_\alpha(P) = u_\alpha(P) - u_\alpha(P^{(0)}). \tag{2}$$

Suppose that for some fixed values of $u_\alpha(P^{(0)})$ the points $P^{(k)}, k = 1, 2, \ldots, n$, having the coordinates $y_\alpha^{(k)}$ and $s^{(k)}$ in the coordinate system (1), are situated in the domain G. Then we may define a $3n$-dimensional distribution law of probabilities F_n for the quantities

$$w_\alpha^{(k)} = w_\alpha(P^{(k)}), \quad \alpha = 1, 2, 3; \quad k = 1, 2, \ldots, n,$$

where

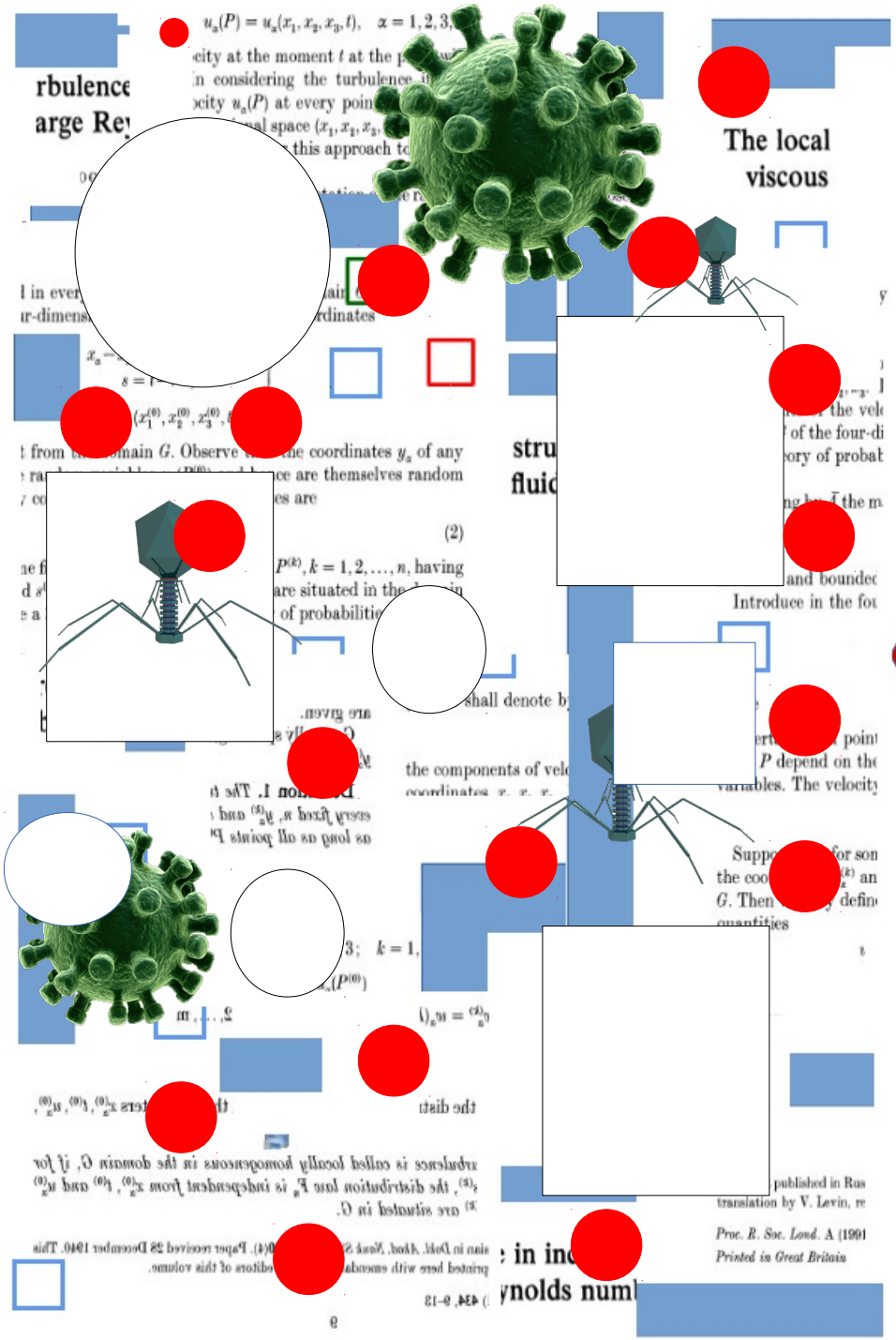
$$u_\alpha^{(0)} =$$

are given.

Generally speaking, the distribution law F_n depends on the parameters $x_\alpha^{(0)}, t^{(0)}, u_\alpha^{(0)}, y_\alpha^{(k)}, s^{(k)}$.

Definition 1. *The turbulence is called locally homogeneous in the domain G, if for every fixed n, $y_\alpha^{(k)}$ and $s^{(k)}$, the distribution law F_n is independent from $x_\alpha^{(0)}$, $t^{(0)}$ and $u_\alpha^{(0)}$ as long as all points $P^{(k)}$ are situated in G.*

† First published in Russian in *Dokl. Akad. Nauk SSSR* (1941) **30**(4). Paper received 28 December 1940. This translation by V. Levin, reprinted here with emendations by the editors of this volume.



The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers†

By A. N. Kolmogorov

§1. We shall denote by

$$u_\alpha(P) = u_\alpha(x_1, x_2, x_3, t), \quad \alpha = 1, 2, 3,$$

the components of velocity at the moment t at the point with rectangular cartesian coordinates x_1, x_2, x_3. In considering the turbulence it is natural to assume the components of the velocity $u_\alpha(P)$ at every point $P = (x_1, x_2, x_3, t)$ of the considered domain G of the four-dimensional space (x_1, x_2, x_3, t) are *random variables* in the sense of the theory of probabilities (cf. for this approach to the problem Millionshtchikov (1939)).

Denoting by $\bar{A}$ the mathematical expectation of the random variable A we suppose that

$$\bar{u}_\alpha^2 \quad \text{and} \quad \overline{(\mathrm{d}u_\alpha/\mathrm{d}x_\beta)^2}$$

are finite and bounded in every bounded subdomain of the domain

Introduce in the four-dimensional space (x_1, x_2, x_3, t) new coordinates

$$\left.\begin{aligned} y_\alpha &= x_\alpha - x_\alpha^{(0)} - u_\alpha(P^{(0)})(t - t^{(0)}), \\ s &= t - t^{(0)}, \end{aligned}\right\} \tag{1}$$

where

$$P^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, t)$$

is a certain fixed point from the domain G. Observe that the coordinates y_α of any point P depend on the random variables $u_\alpha(P^{(0)})$ and hence are themselves random variables. The velocity components in the new coordinates are

$$w_\alpha(P) = u_\alpha(P) - u_\alpha(P^{(0)}). \tag{2}$$

Suppose that for some fixed values of $u_\alpha(P^{(0)})$ the points $P^{(k)}, k = 1, 2, \ldots, n$, having the coordinates $y_\alpha^{(k)}$ and $s^{(k)}$ in the coordinate system (1), are situated in the domain G. Then we may define a $3n$-dimensional distribution law of probabilities F_n for the quantities

$$w_\alpha^{(k)} = w_\alpha(P^{(k)}), \quad \alpha = 1, 2, 3; \quad k = 1, 2, \ldots, n,$$

where

$$u_\alpha^{(0)} =$$

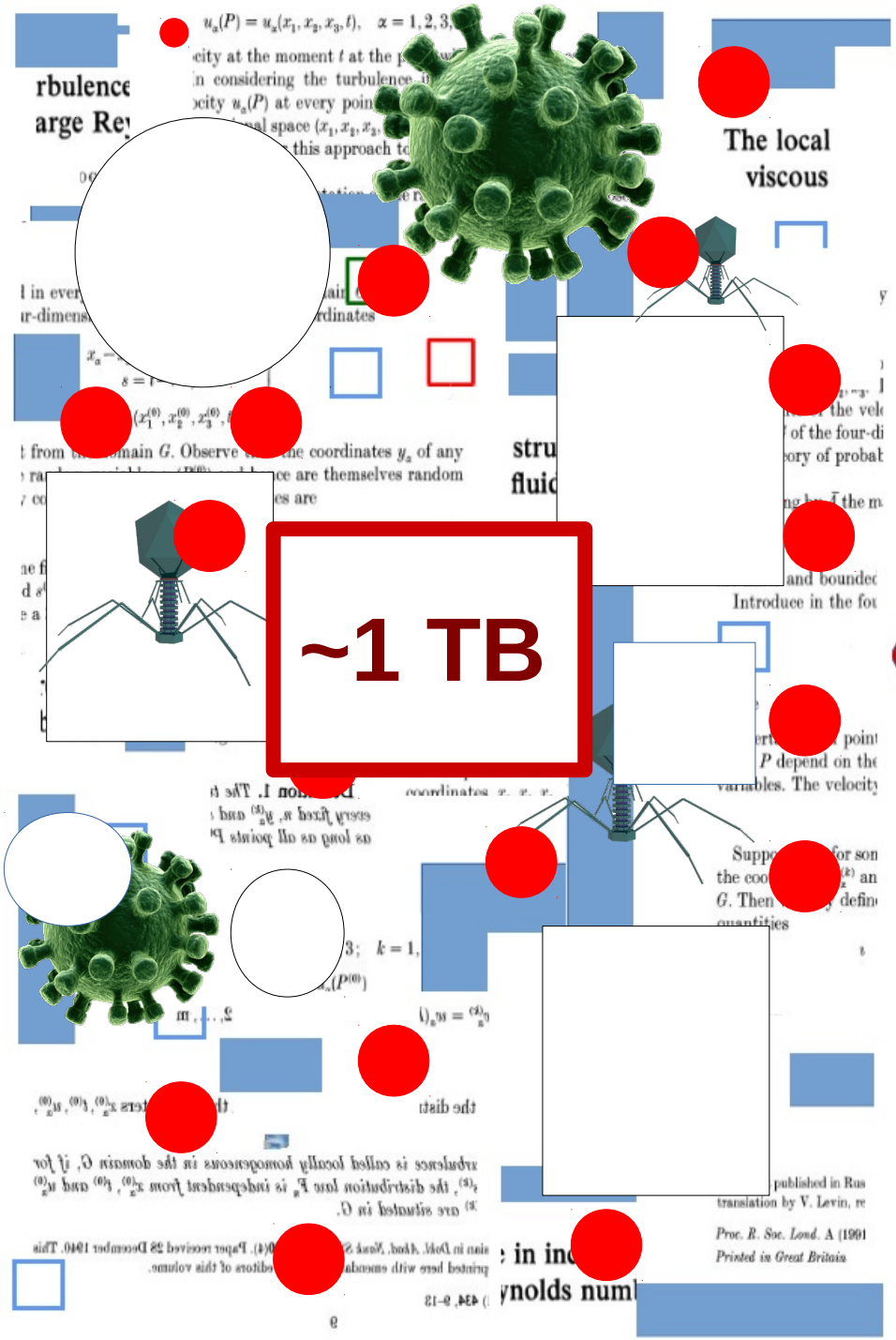
are given.

Generally speaking, the distribution law F_n depends on the parameters $x_\alpha^{(0)}, t^{(0)}, u_\alpha^{(0)}, y_\alpha^{(k)}, s^{(k)}$.

Definition 1. *The turbulence is called locally homogeneous in the domain G, if for every fixed n, $y_\alpha^{(k)}$ and $s^{(k)}$, the distribution law F_n is independent from $x_\alpha^{(0)}$, $t^{(0)}$ and $u_\alpha^{(0)}$ as long as all points $P^{(k)}$ are situated in G.*

† First published in Russian in *Dokl. Akad. Nauk SSSR* (1941) **30**(4). Paper received 28 December 1940. This translation by V. Levin, reprinted here with emendations by the editors of this volume.

Proc. R. Soc. Lond. A (1991) **434**, 9–13
Printed in Great Britain



The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers†

By A. N. Kolmogorov

§1. We shall denote by

$$u_\alpha(P) = u_\alpha(x_1, x_2, x_3, t), \quad \alpha = 1, 2, 3,$$

the components of velocity at the moment t at the point with rectangular cartesian coordinates x_1, x_2, x_3. In considering the turbulence it is natural to assume the components of the velocity $u_\alpha(P)$ at every point $P = (x_1, x_2, x_3, t)$ of the considered domain G of the four-dimensional space (x_1, x_2, x_3, t) are *random variables* in the sense of the theory of probabilities (cf. for this approach to the problem Millionshtchikov (1939)).

Denoting by $\bar{A}$ the mathematical expectation of the random variable A we suppose that

$$\bar{u}_\alpha^2 \quad \text{and} \quad \overline{(\mathrm{d}u_\alpha/\mathrm{d}x_\beta)^2}$$

are finite and bounded in every bounded subdomain of the domain

Introduce in the four-dimensional space (x_1, x_2, x_3, t) new coordinates

$$\left.\begin{aligned} y_\alpha &= x_\alpha - x_\alpha^{(0)} - u_\alpha(P^{(0)})(t - t^{(0)}), \\ s &= t - t^{(0)}, \end{aligned}\right\} \tag{1}$$

where

$$P^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, t)$$

is a certain fixed point from the domain G. Observe that the coordinates y_α of any point P depend on the random variables $u_\alpha(P^{(0)})$ and hence are themselves random variables. The velocity components in the new coordinates are

$$w_\alpha(P) = u_\alpha(P) - u_\alpha(P^{(0)}). \tag{2}$$

Suppose that for some fixed values of $u_\alpha(P^{(0)})$ the points $P^{(k)}, k = 1, 2, \ldots, n$, having the coordinates $y_\alpha^{(k)}$ and $s^{(k)}$ in the coordinate system (1), are situated in the domain G. Then we may define a $3n$-dimensional distribution law of probabilities F_n for the quantities

$$w_\alpha^{(k)} = w_\alpha(P^{(k)}), \quad \alpha = 1, 2, 3; \quad k = 1, 2, \ldots, n,$$

where

$$u_\alpha^{(0)} =$$

are given.

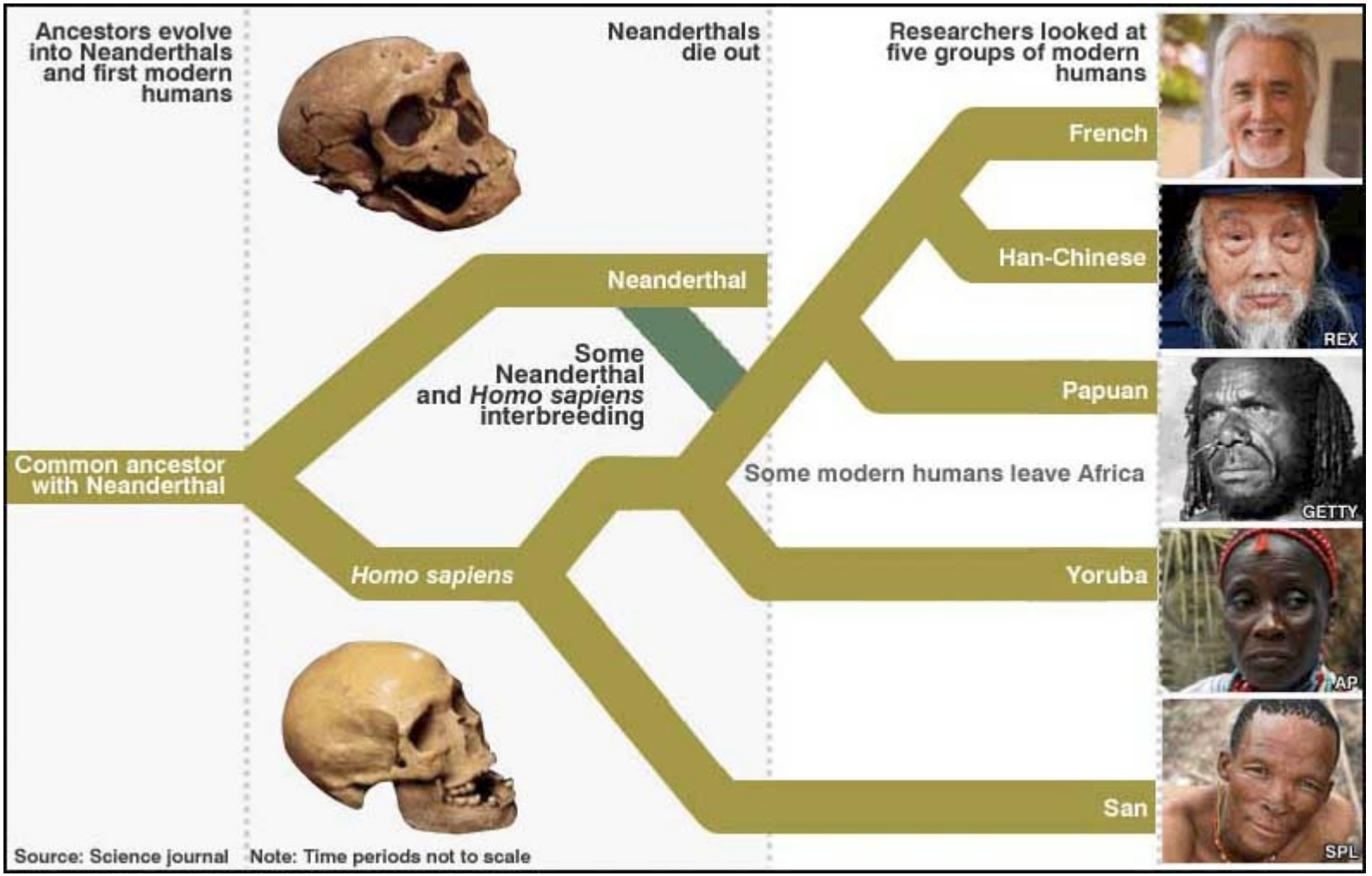
Generally speaking, the distribution law F_n depends on the parameters $x_\alpha^{(0)}, t^{(0)}, u_\alpha^{(0)}, y_\alpha^{(k)}, s^{(k)}$.

Definition 1. *The turbulence is called locally homogeneous in the domain G, if for every fixed n, $y_\alpha^{(k)}$ and $s^{(k)}$, the distribution law F_n is independent from $x_\alpha^{(0)}$, $t^{(0)}$ and $u_\alpha^{(0)}$ as long as all points $P^{(k)}$ are situated in G.*

† First published in Russian in *Dokl. Akad. Nauk SSSR* (1941) **30**(4). Paper received 28 December 1940. This translation by V. Levin, reprinted here with emendations by the editors of this volume.

Proc. R. Soc. Lond. A (1991) **434**, 9–13
Printed in Great Britain

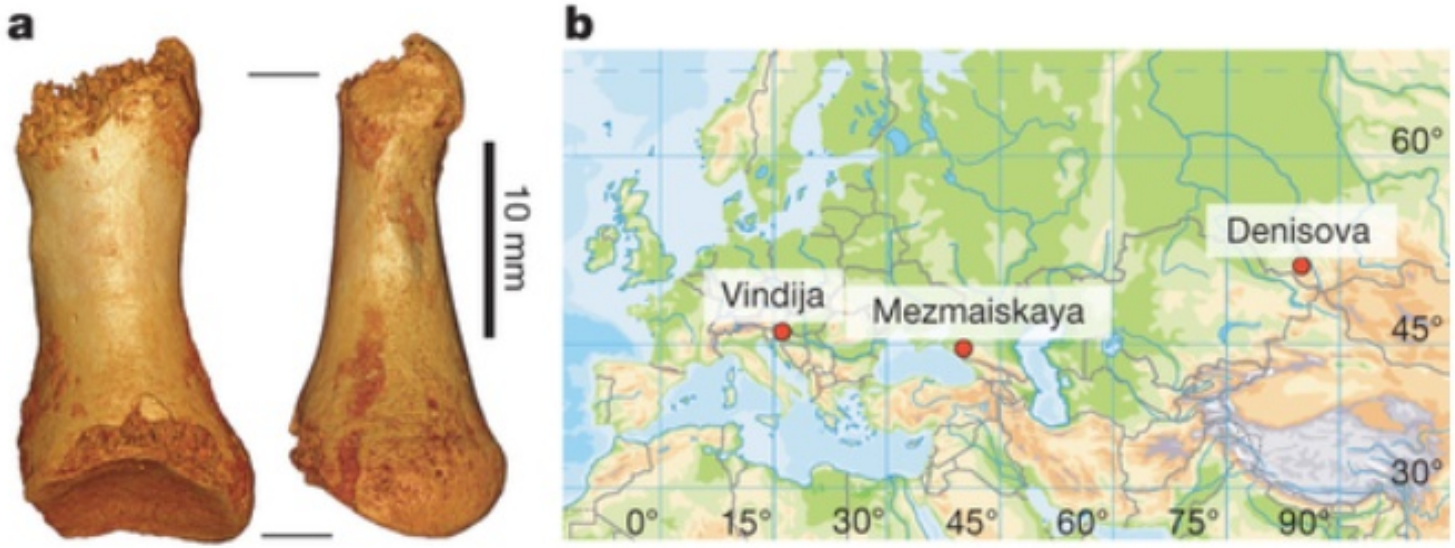
Timeline



Neanderthals approximate period: ~400,000 to ~40,000 years ago; Unknown reason of extinction.

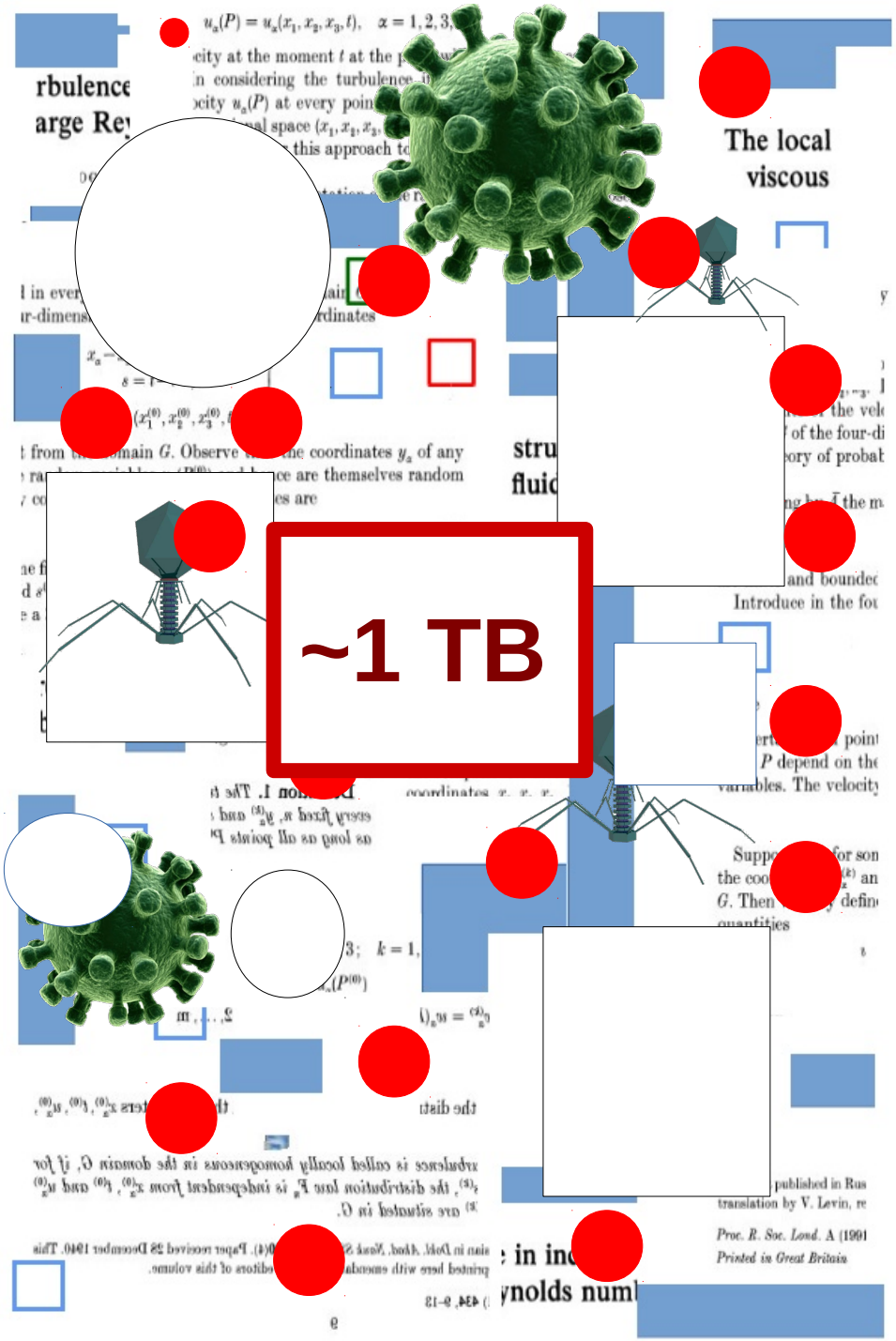
Neanderthal (~50,000 years ago)

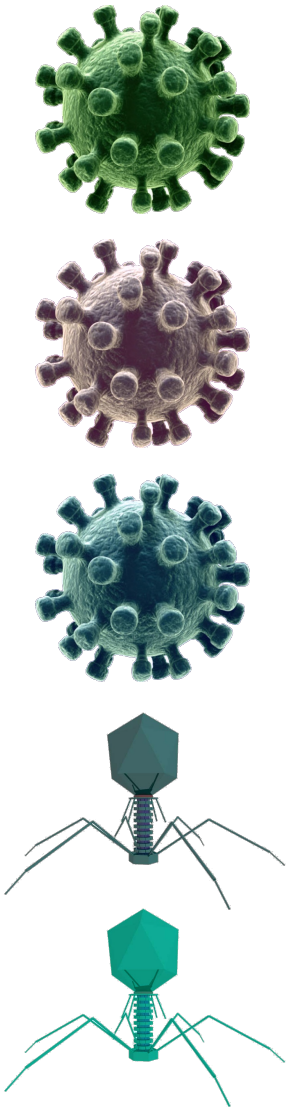
Figure 1: Toe phalanx and location of Neanderthal samples for which genome-wide data are available.



a, The toe phalanx found in the east gallery of Denisova Cave in 2010. Dorsal view (left image), left view (right image). Total length of the bone is 26 mm. **b**, Map of Eurasia showing the location of Vindija Cave, Mezmaiskaya Cave and D..

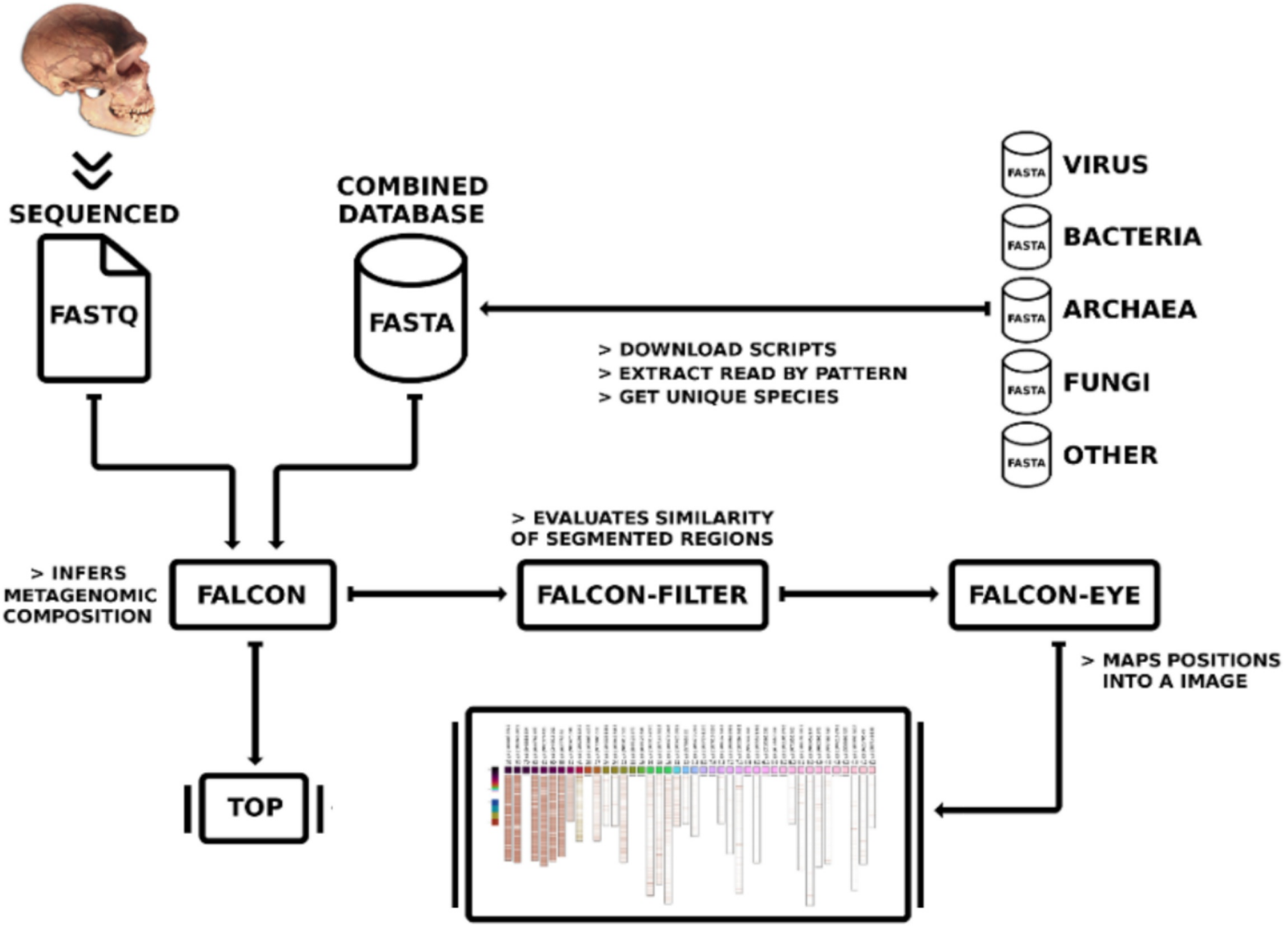
Prüfer, K. *et al* (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature, 505(7481), 43-49.





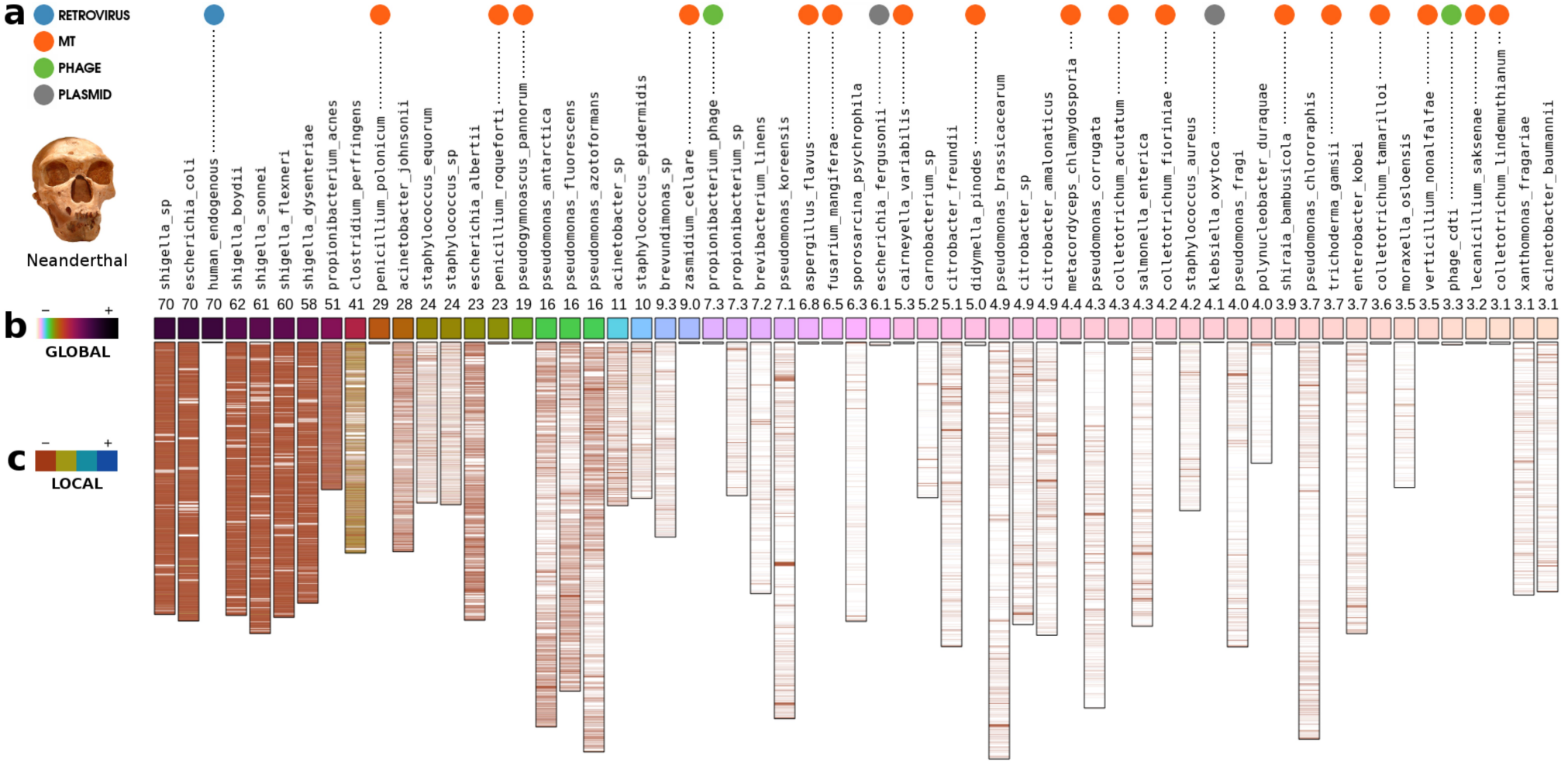
All known microorganisms

Metagenomic sample composition pipeline

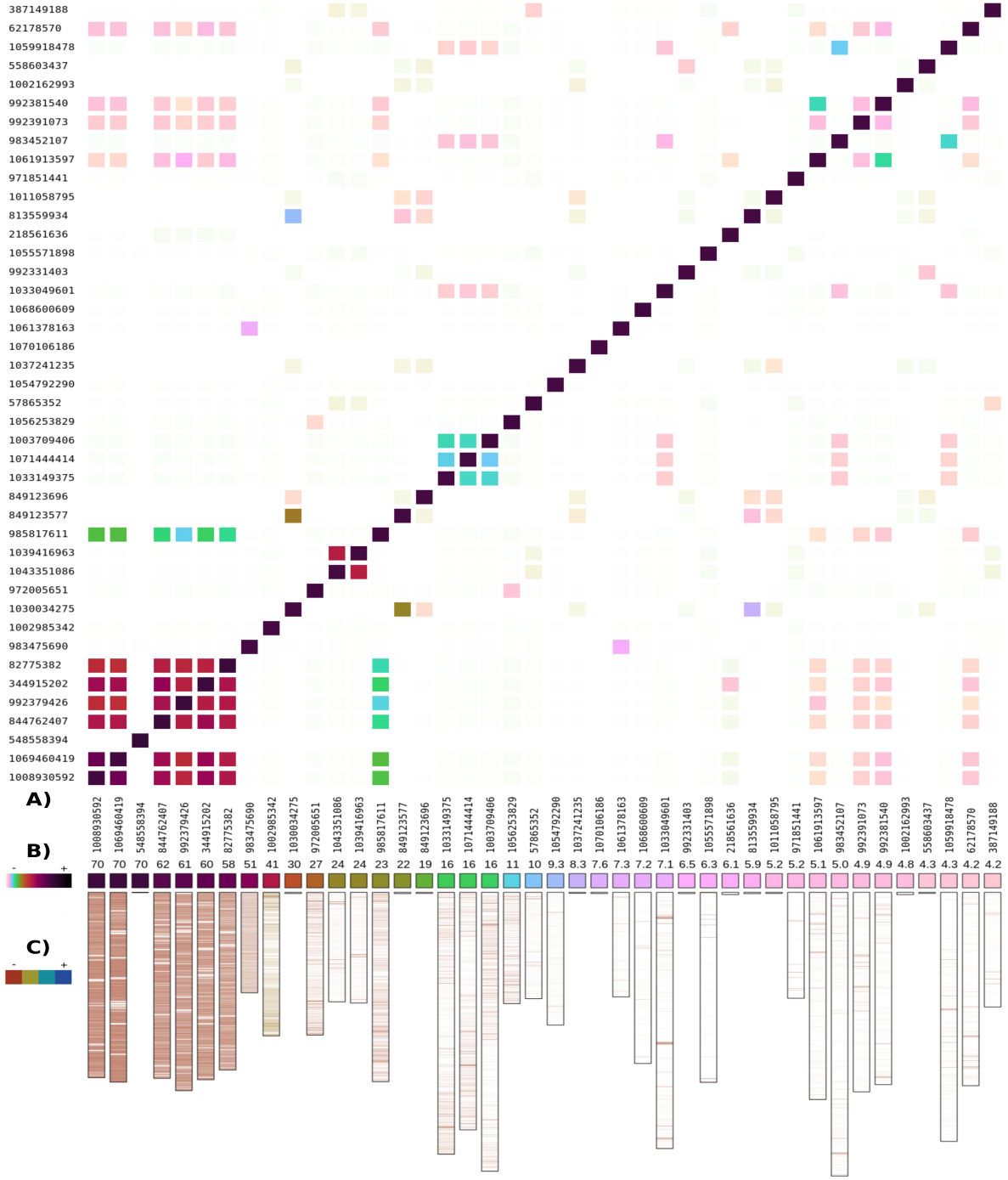


D. Pratas, R. Silva, J. M. O. S. Rodrigues, A. J. Pinho, P. J. S. G. Ferreira. FALCON: a compression method to infer metagenomic composition of ancient DNA. (Under submission).

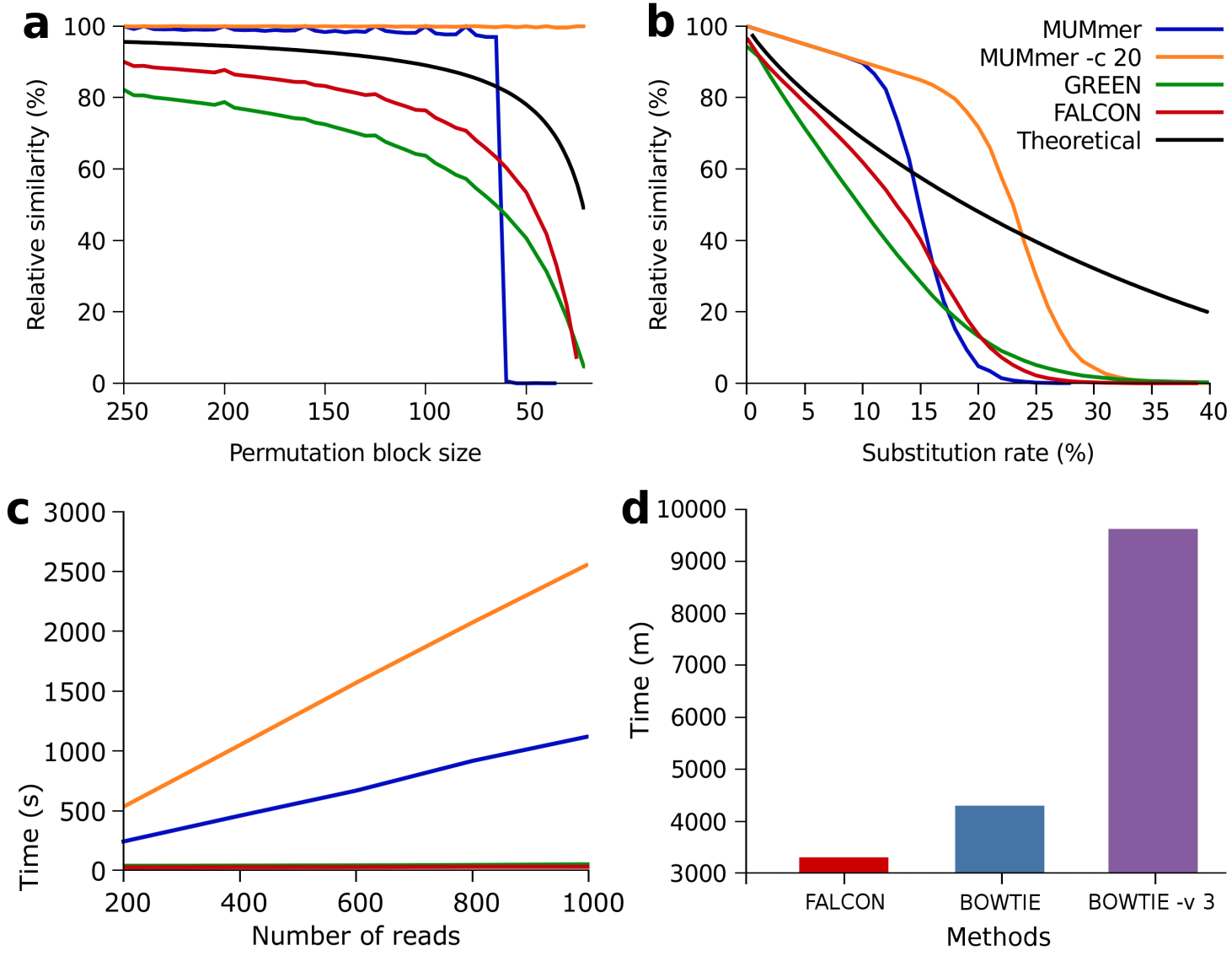
Neanderthal metagenomic sample composition



D. Pratas, R. Silva, J. M. O. S. Rodrigues, A. J. Pinho, P. J. S. G. Ferreira. FALCON: a compression method to infer metagenomic composition of ancient DNA. (Under submission).

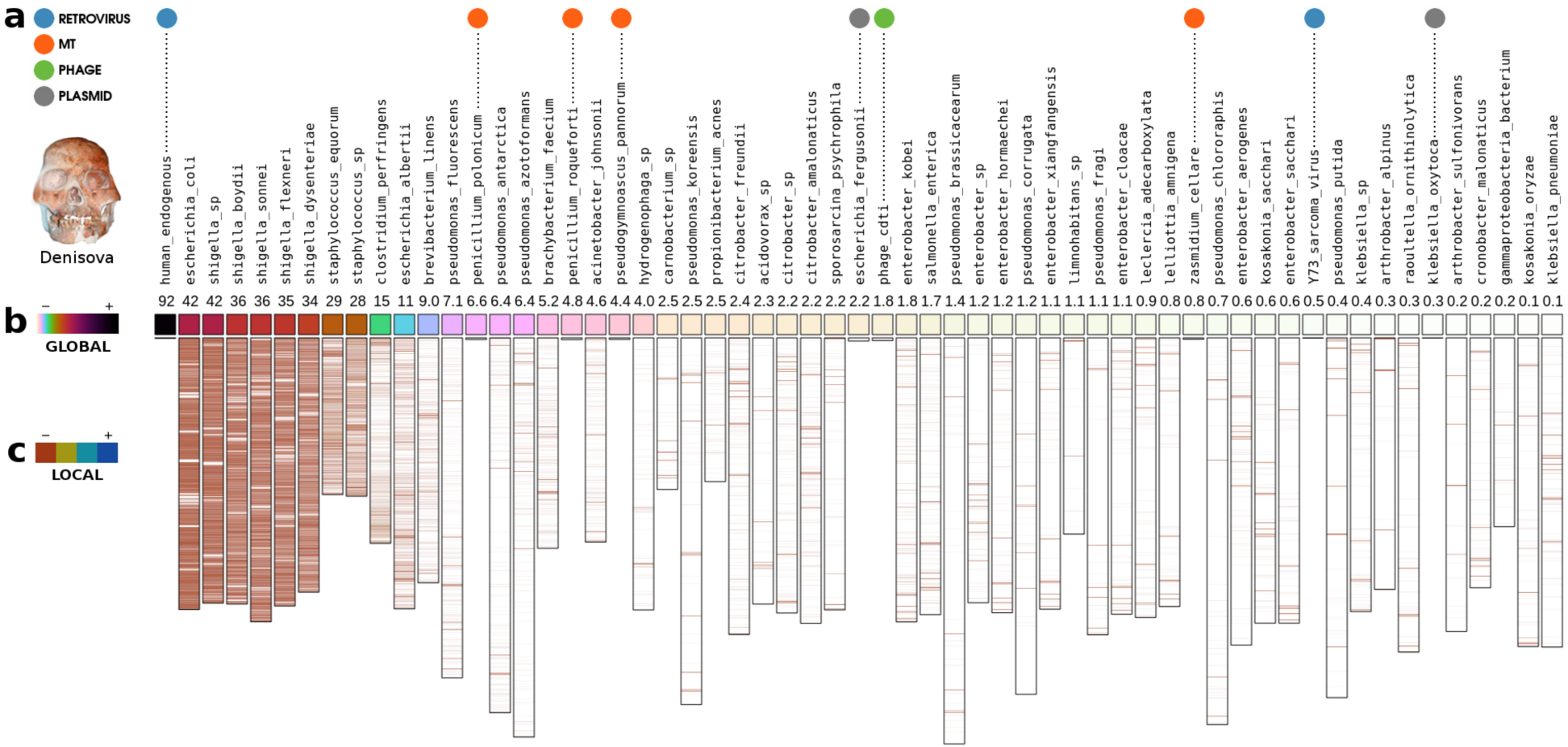


D. Pratas, R. Silva, A. J. Pinho, P. J. S. G. Ferreira. Unsupervised measurement of relative similarity between non-assembled genomic sequences. (submitted).



D. Pratas, R. Silva, J. M. O. S. Rodrigues, A. J. Pinho, P. J. S. G. Ferreira. FALCON: a compression method to infer metagenomic composition of ancient DNA. (Under submission).

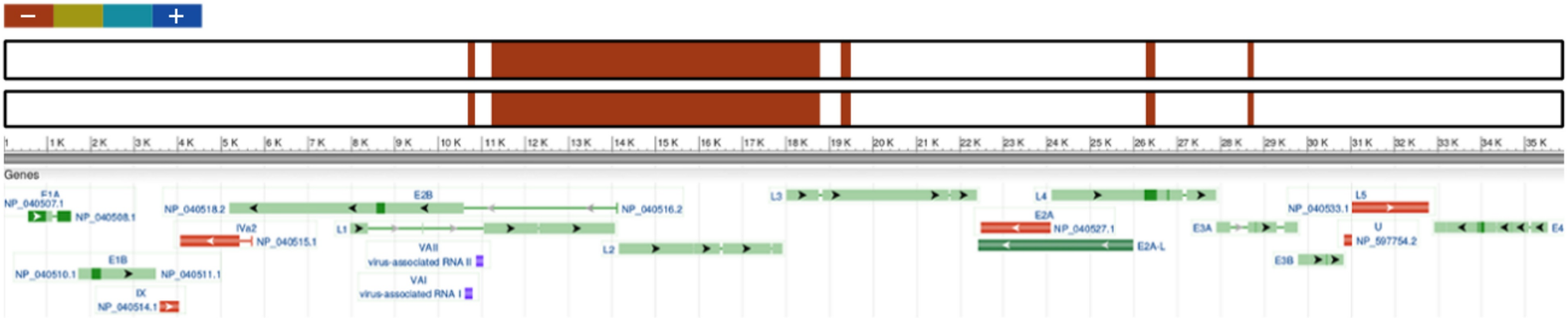
Denisova metagenomic sample composition



D. Pratas, R. Silva, J. M. O. S. Rodrigues, A. J. Pinho, P. J. S. G. Ferreira. FALCON: a compression method to infer metagenomic composition of ancient DNA. (Under submission).

Denisova metagenomic sample composition

Local similarities of Human Adenovirus with an ancient virus:



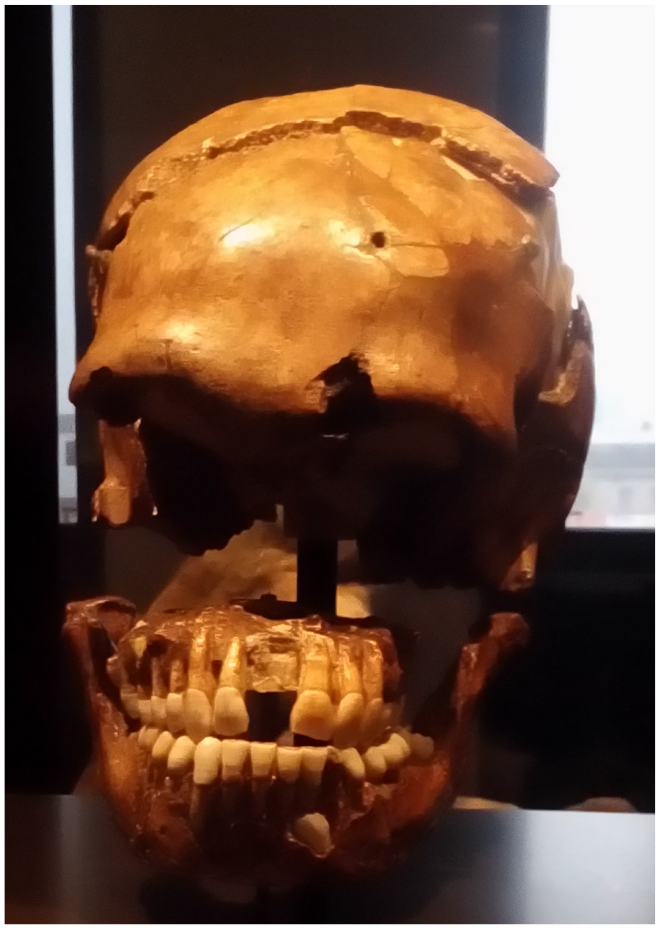
Associated with respiratory infections

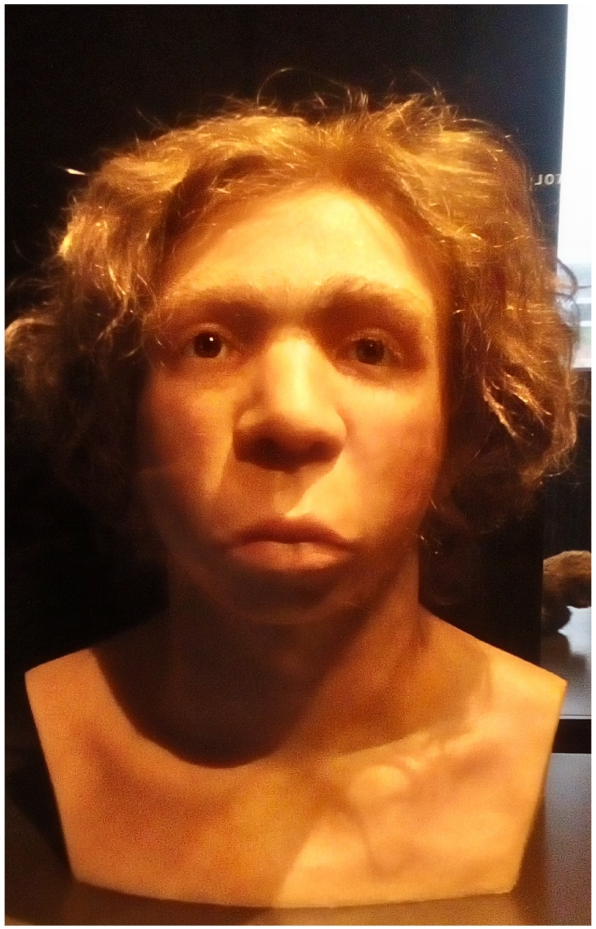
D. Pratas, R. Silva, J. M. O. S. Rodrigues, A. J. Pinho, P. J. S. G. Ferreira. FALCON: a compression method to infer metagenomic composition of ancient DNA. (Under submission).

Ancient genomes hide a key for efficient therapeutics

What makes us human?

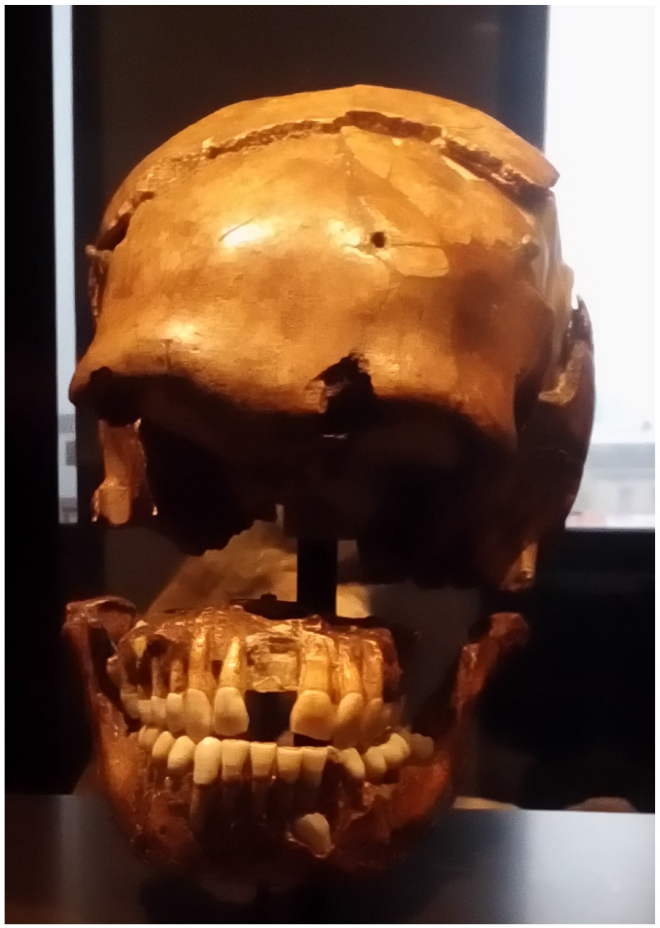
Differences between Humans and Neanderthals

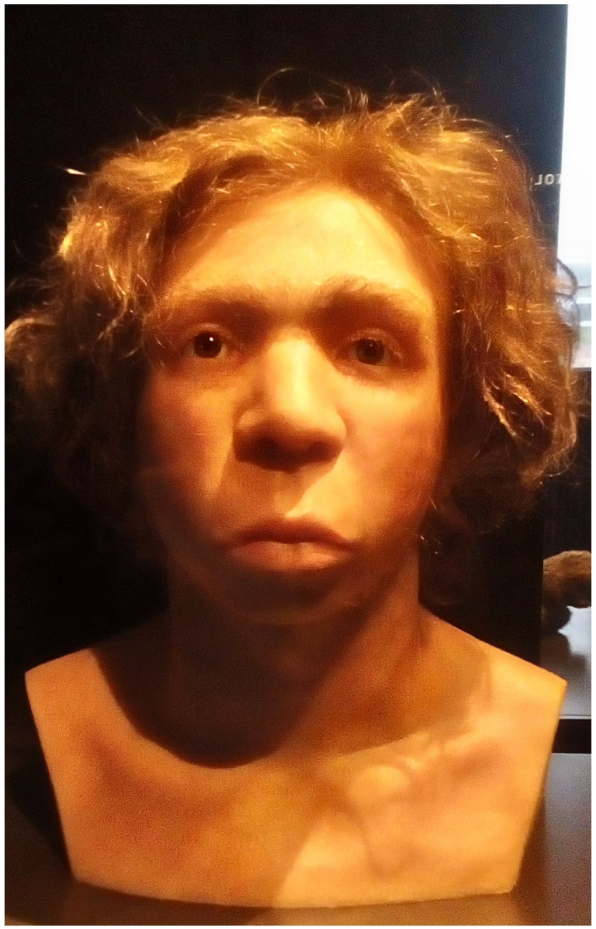




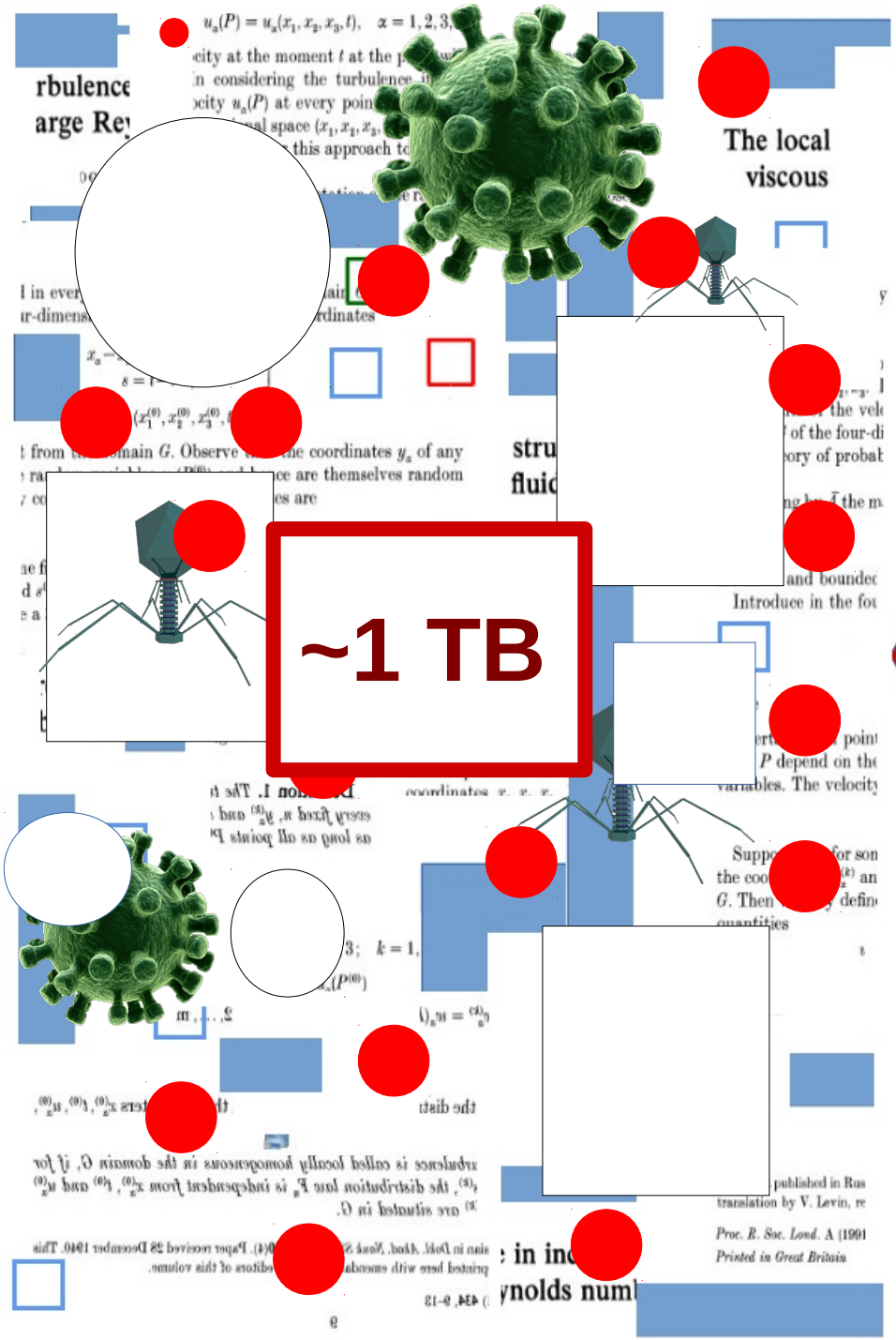
45,000 years old Neanderthal in the Neues museum at Berlin. Photo and edition by Pratas *et al*, May, 2017.

Differences between Humans and Neanderthals





45,000 years old Neanderthal in the Neues museum at Berlin. Photo and edition by Pratas *et al*, May, 2017.



The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers†

By A. N. Kolmogorov

§1. We shall denote by

$$u_\alpha(P) = u_\alpha(x_1, x_2, x_3, t), \quad \alpha = 1, 2, 3,$$

the components of velocity at the moment t at the point with rectangular cartesian coordinates x_1, x_2, x_3. In considering the turbulence it is natural to assume the components of the velocity $u_\alpha(P)$ at every point $P = (x_1, x_2, x_3, t)$ of the considered domain G of the four-dimensional space (x_1, x_2, x_3, t) are *random variables* in the sense of the theory of probabilities (cf. for this approach to the problem Millionshtchikov (1939)).

Denoting by $\bar{A}$ the mathematical expectation of the random variable A we suppose that

$$\bar{u}_\alpha^2 \quad \text{and} \quad \overline{(\mathrm{d}u_\alpha/\mathrm{d}x_\beta)^2}$$

are finite and bounded in every bounded subdomain of the domain

Introduce in the four-dimensional space (x_1, x_2, x_3, t) new coordinates

$$\left.\begin{aligned} y_\alpha &= x_\alpha - x_\alpha^{(0)} - u_\alpha(P^{(0)})(t - t^{(0)}), \\ s &= t - t^{(0)}, \end{aligned}\right\} \tag{1}$$

where

$$P^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, t)$$

is a certain fixed point from the domain G. Observe that the coordinates y_α of any point P depend on the random variables $u_\alpha(P^{(0)})$ and hence are themselves random variables. The velocity components in the new coordinates are

$$w_\alpha(P) = u_\alpha(P) - u_\alpha(P^{(0)}). \tag{2}$$

Suppose that for some fixed values of $u_\alpha(P^{(0)})$ the points $P^{(k)}, k = 1, 2, \ldots, n$, having the coordinates $y_\alpha^{(k)}$ and $s^{(k)}$ in the coordinate system (1), are situated in the domain G. Then we may define a $3n$-dimensional distribution law of probabilities F_n for the quantities

$$w_\alpha^{(k)} = w_\alpha(P^{(k)}), \quad \alpha = 1, 2, 3; \quad k = 1, 2, \ldots, n,$$

where

$$u_\alpha^{(0)} =$$

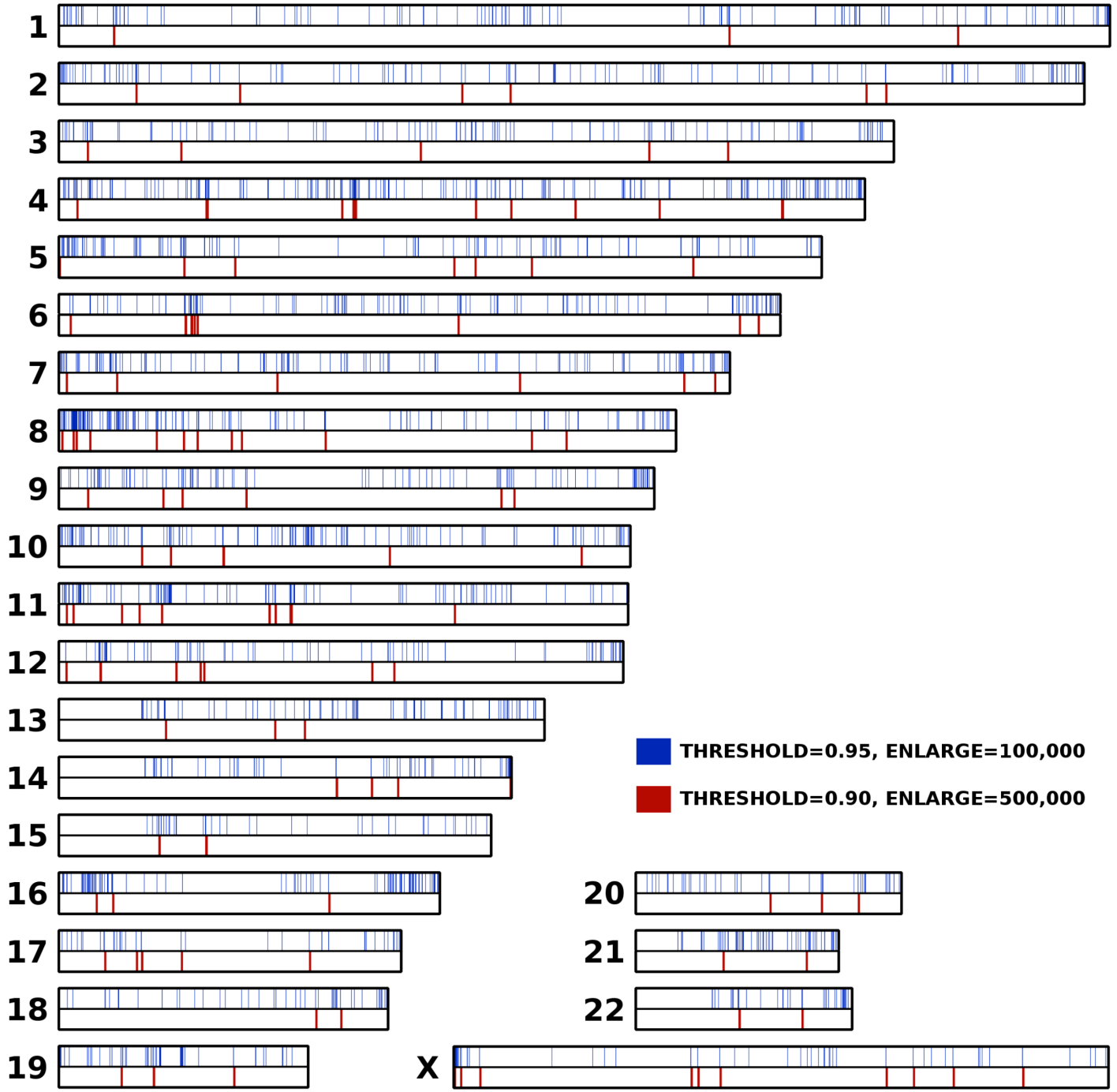
are given.

Generally speaking, the distribution law F_n depends on the parameters $x_\alpha^{(0)}, t^{(0)}, u_\alpha^{(0)}, y_\alpha^{(k)}, s^{(k)}$.

Definition 1. *The turbulence is called locally homogeneous in the domain G, if for every fixed n, $y_\alpha^{(k)}$ and $s^{(k)}$, the distribution law F_n is independent from $x_\alpha^{(0)}$, $t^{(0)}$ and $u_\alpha^{(0)}$ as long as all points $P^{(k)}$ are situated in G.*

† First published in Russian in *Dokl. Akad. Nauk SSSR* (1941) **30**(4). Paper received 28 December 1940. This translation by V. Levin, reprinted here with emendations by the editors of this volume.

Proc. R. Soc. Lond. A (1991) **434**, 9–13
Printed in Great Britain



D. Pratas, M. Hosseini, R. Silva, A. J. Pinho, P. J. S. G. Ferreira. Visualization of distinct DNA regions of the modern human relatively to a Neanderthal genome. IBPRIA, May 2017.

Modern-human genes associated with:

brain (neurotransmiters and synapses)

D. Pratas, M. Hosseini, R. Silva, A. J. Pinho, P. J. S. G. Ferreira. Visualization of distinct DNA regions of the modern human relatively to a Neanderthal genome. IBPRIA, May 2017.

Modern-human genes associated with:

brain (neurotransmiters and synapses)

hearing

D. Pratas, M. Hosseini, R. Silva, A. J. Pinho, P. J. S. G. Ferreira. Visualization of distinct DNA regions of the modern human relatively to a Neanderthal genome. IBPRIA, May 2017.

Modern-human genes associated with:

brain (neurotransmiters and synapses)

hearing

blood

D. Pratas, M. Hosseini, R. Silva, A. J. Pinho, P. J. S. G. Ferreira. Visualization of distinct DNA regions of the modern human relatively to a Neanderthal genome. IBPRIA, May 2017.

Modern-human genes associated with:

brain (neurotransmiters and synapses)

hearing

blood

fertility

D. Pratas, M. Hosseini, R. Silva, A. J. Pinho, P. J. S. G. Ferreira. Visualization of distinct DNA regions of the modern human relatively to a Neanderthal genome. IBPRIA, May 2017.

Modern-human genes associated with:

brain (neurotransmiters and synapses)

hearing

blood

fertility

immune system

D. Pratas, M. Hosseini, R. Silva, A. J. Pinho, P. J. S. G. Ferreira. Visualization of distinct DNA regions of the modern human relatively to a Neanderthal genome. IBPRIA, May 2017.

Modern-human genes associated with:

brain (neurotransmiters and synapses)

hearing

blood

fertility

immune system

and others with unknown function...

D. Pratas, M. Hosseini, R. Silva, A. J. Pinho, P. J. S. G. Ferreira. Visualization of distinct DNA regions of the modern human relatively to a Neanderthal genome. IBPRIA, May 2017.

Next challenges:

- Visualization of distinct DNA regions of the modern human relatively to a Denisova genome

- Classification of ancient or contamination microorganisms present in a Neanderthal and Denisova genomes

- Reconstruction of ancient microorganisms

- Study the co-evolution between host and pathogen

Thanks for the attention!

Work supported by FCT funds

pratas@ua.pt