Possible traces of resonant signaling in the genome

Ivan Savelyev [1], Max Myakishev-Rempel [1,2]

1: Localized Therapeutics, San Diego, CA, USA

2: DNA Resonance Research Foundation, San Diego, CA, USA

Abbreviations:

HIDERs - Homologous If Decoded Elements, Repetitive.

Ninety-seven years ago, Alexander Gurwitsch proposed the existence of a morphogenetic field that is created by the body and is responsible for developing and maintaining the shape of the body (Gurwitsch, 1922). He and others demonstrated that biological organisms influence the development of each other at short distances and that some of this influence is blocked by optical filters, suggesting that morphogenic field is of electromagnetic nature (Gurwitsch, 1988; Volodyaev and Beloussov, 2015). In 1968, Frohlich predicted that cell and organelle membranes in the presence of constant flux of energy produce coherent waves in the millimeter wave region thus creating a coherent state and enabling electric wave signaling in living organisms (Frohlich, 1988). In 1973, Miller and Web further proposed that it is DNA that produces the morphogenic field and that the genomic code is directly sending and receiving the information from the morphogenic field (Miller and Webb, 1973). The experiments verifying the existence of biological fields involve two samples such as cell culture aliquots in sealed quartz cuvettes separated by optical filters. When one of the aliquots is perturbed, the second one may catch a signal that is transferred non-chemically and is blocked by light impermeable filters. Such effects are often referred to as "non-chemical cell-cell communication" and are reviewed in refs (Cifra et al., 2011; Scholkmann et al., 2013; Trushin, 2004; Xu et al., 2017). Burlakov experimentally demonstrated that optical distortion by quartz retroreflectors of the field produced by fish embryos causes developmental abnormalities, thus confirming that the field is morphogenic and electromagnetic (Burkov et al., 2008; Burlakov et al., 2012).

Although the existence of the field and its morphogenic and electromagnetic nature have been demonstrated, the involvement of DNA in its generation proposed in 1973 by Muller and Webb have not been proven yet. There have been proposed many models for oscillations in DNA that involve the movement of groups of atoms in DNA (referred here as mechanical oscillations) (Scott, 1985; Volkov and Kosevich, 1987). Spectroscopic detection of coherent mechanical oscillations in DNA was reported at THz range (Sajadi et al., 2011). We proposed that in addition to mechanical oscillations in DNA, there are oscillations of delocalized electron clouds in the base stack (Polesskaya et al., 2018) and of delocalized proton clouds of the hydrogen bonds in the base stack (Savelyev et al., 2019). Moreover, we suggested that these oscillations occur in DNA sequence-dependent manner and provide the primary medium for the formation of the morphogenic field. We suggested that since electron and proton clouds are located inside the base stack and are light, they don't cause significant movements of the atoms and therefore don't cause significant movement of water and this way avoid thermal dissipation of energy. We suggested that therefore, the electron and proton cloud oscillations are a more likely medium for the morphogenic field than the mechanical oscillations of DNA which should cause dissipation of energy into the movement of the surrounding water (Polesskaya et al., 2018; Savelyev et al., 2019).

We suggested that electroacoustic resonances between similar DNA sequences form the basis of signaling within the genome and coordinates the function of the cell. We also suggested possible mechanisms by which these oscillations channeled by the microtubules from one nucleus to another forming an oscillation network of the body. This way, we transformed an idea of a diffuse morphogenic field into the model of the morphogenic field traveling between the nuclei via tunnels formed by microtubules. This also explained how nature may avoid the dissipation of the electroacoustic signals in largely amorphous tissues (Savelyev et al., 2019). We

further implicated genomic repeats as primary candidate sequences to serve as resonators. We suggested that the fact that the 300 base pair-long Alu repeat occurs 1.1 million times in each of our cells makes it the best candidate for serving as a resonator by the mere number of copies improving the quality of oscillations and reducing the dissipation of the signal. We also suggested that the primary function of genomic repeats such as telomeric, centromeric, simple repeats and transposable elements is to support the resonant signaling in the genomes of complex organisms. We suggested (Savelyev et al., 2019) that this resonant signaling system is deliberately supported by the cells via the flux of ATP and other biochemical energy in accordance with Frohlich models (Fröhlich, 1968). We suggested that similar DNA sequences resonate with each other forming a resonating network within the nucleus, between the nuclei and across all nuclei. In this process some of the repetitive sequences are energized by chemical processes, their oscillations are transmitted along the base stack and cause the oscillations in similar sequences. This way conformational changes in chromatin in one place lead to conformational changes in chromatin of similar DNA sequences allowing for resonant signaling within the nucleus and across the organism. We suggested that this process is deliberate, developed by evolution for higher organisms and that the cell spends ATP and other types of chemical energy on supporting this resonant genomic signaling. This way chromatin immediately and mechanistically is mediating the interaction between the electromagnetic resonant signaling and molecular signaling in a DNA sequence-specific manner. In this signaling, resonance properties of DNA sequences provide specificity and ATP energy allows amplification of electromagnetic resonant signals and conversion of them to molecular signals. For example, oscillations in some Alu sequences might be induced by ATP-dependent chromatin remodeling factors, these oscillations may be transmitted via the base stack to the second group of Alu elements, via electromagnetic resonance, the Alu elements of the second group would begin resonant oscillation, this oscillation would be amplified by ATP-dependent chromatin remodeling factors bound to them, causing chromatin opening and transcription of nearby genes. This mechanism would explain why Alu elements are enriched in gene promoters.

Although there are experimental demonstrations of morphogenic field effects, the involvement of DNA in its formation is yet to be proven. The prediction of frequencies of oscillations in DNA is not trivial since DNA could support a number of modes of oscillation including various modes of oscillation of mechanical, electron and proton clouds. Since DNA is wrapped around nucleosomes, chromatin state should also be considered. We suggest that sequence-specific oscillations in DNA could spread over an extremely wide range of frequencies.

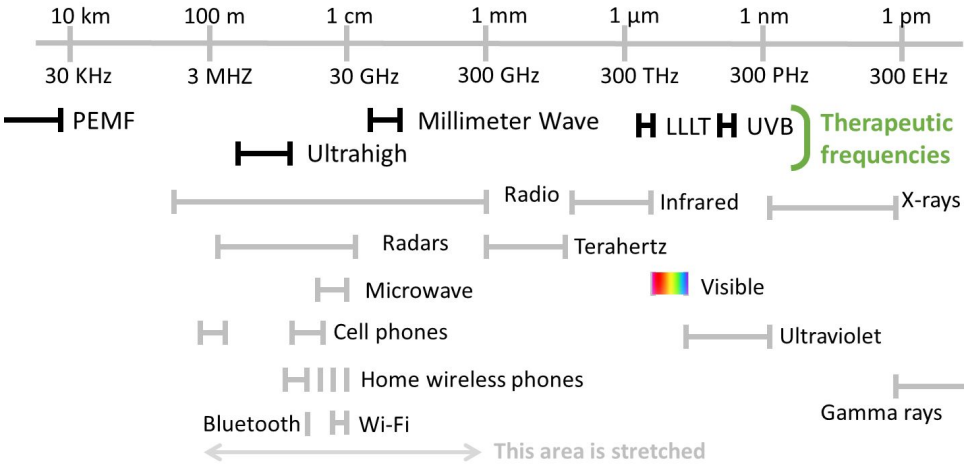


Fig. [Spectrum] Frequency ranges used for therapy. (LLLT – low-level light therapy, PEMF - pulsed electromagnetic field).

Some insight might be obtained from electromagnetic frequencies used in physical therapy. Especially informative would be those frequencies, which produce effects at extremely low power suggesting that they tap onto electromagnetic resonant signaling. Such frequencies are shown in Fig. [Spectrum]. Specifically, the

following therapeutic ranges of electromagnetic frequencies are exhibit significant effects at low power and thus are likely to be tapping on existing signaling pathways: pulsed electromagnetic field therapy (Binder et al., 1984), ultra high-frequency therapy (Lushnikov et al., 2004), millimeter wave therapy (Usichenko et al., 2003), low-level light therapy (Bjordal et al., 2003), and UVB (Lowe et al., 1991). We suggest that these frequencies are good candidate frequencies for resonant oscillations in DNA. Since the frequency depends on the mass of the oscillator, shorter DNA repeats should oscillate at higher frequencies than the longer ones. Based primarily on these assumptions, we propose the following approximate prediction of resonance frequencies of the genomic repeats, **Table [Wavelengths]**. Note that the natural wavelength of the oscillator can be much larger than its size. Recently, for radioelectronics, nanomechanical magnetoelectric (ME) antennas have been developed, which resonate with wavelengths 1000 times larger that their size (Nan et al., 2017; Shi et al., 2016). An additional conversion factor which could allow DNA to resonate at higher frequencies, that the therapeutic electromagnetic waves shown onto the biological tissue, might induce oscillations in the tissue which could spread acoustically. Thus an electromagnetic wavelength in the air might be converted to acoustic in body tissue, thus shortening the wavelength approximately 200,000 times. Although the predictions in Table [Wavelengths] are preliminary and need to be tested experimentally, they may help understanding possible mechanisms underlying the possible mechanistic connection between electromagnetic therapies and the proposed resonant genomic signaling.

Table [Wavelengths]: A very approximate prediction of resonance wavelengths of genomic repeats

Repeat unit length		Periodic	Type	wavelength	PEMF	UHF	MWT	LLLT	UVB
				light	37km	0.3m	7mm	800nm	300nm
				sound	186m	1.5um	30nm	4nm	1.5nm
1 bp	0.3 nm	y	simple						
2 bp	0.7 nm	y	simple						
3 bp	1.0 nm	y	simple						
4 bp	1.3 nm	y	simple						
6 bp	2.0 nm	y	telomeric						
171 bp	57 nm	y	centromeric						
260 bp	86 nm	n	MIR						
300 bp	100 nm	n	Alu						
1000 bp	332 nm	n	Mariner						
6000 bp	1992 nm	n	LINE1						

(UHF - ultra high frequency, MWT - millimeter wave therapy)

Since so far, there is no published evidence for the resonant genomic signaling, we attempted searching for its traces in the genome computationally. Since we believe that the majority of repetitive sequences in the genome are involved in meaningful resonant signaling, we hypothesized that some of the unique (non-repetitive) sequences in the genome might have evolved to resonate with the genomic repeats. Accordingly, we hypothesized that it is not necessary for the unique sequence to be identical to the repeat, that for resonance, it might need to be only superficially similar to the sequence of the repeat: for example, it is possible that some oscillations involve primarily the electron clouds of the aromatic rings (Savelyev et al., 2019). This way only purine-pyrimidine structure of the resonating sequences should be similar and their primary sequences could be different. This simplification of the sequence from the primary sequence to the purine-pyrimidine sequence is further called "purine code". Similarly, for the oscillations which involve primarily the proton clouds of the delocalized protons of the hydrogen bonds in basepairs, only the patterns of these bonds should be similar and the primary sequence could be different. This simplification of the sequence from primary to strong/weak (3 bonds /2 bonds per base pair) is further called "strong code". The recoding rules used here are listed in **Table [Codes]**.

Table [Codes]. Our recoding rules

Purine code	A G	→	R - purines
	C T	→	Y - pyrimidines
Strong code	G C	→	S - strong
	A T	→	W - weak
Amino code	A C	→	M - amino
	G T	→	K - keto
Thymine code	T	→	T - thymine
	A G C	→	V - not thymine

We can not chemically rationalize the classification of nucleotides traditionally called "amino" and "keto" but we used it as "amino" code because it produced statistical results too, see below. We have also noticed that T differs very much chemically from A, G, and C, since it contains neither keto nor amino group and therefore introduced "thymine" code. These rules are derived from the IUPAC nucleotide classification which in turn, is based on the chemical structure of the bases. Therefore we attempted the search for sequences which are unique (non-repetitive), but after recoding (simplification) become similar to genomic repeats or each other. We will refer to them as HIDERs (Homologous upon recoding). Since four recoding schemes were used in this study, Table [Codes], four types of HIDERs will be discussed: purine, strong, amine and thymine HIDERs in accordance with the recoding rules used to find them.

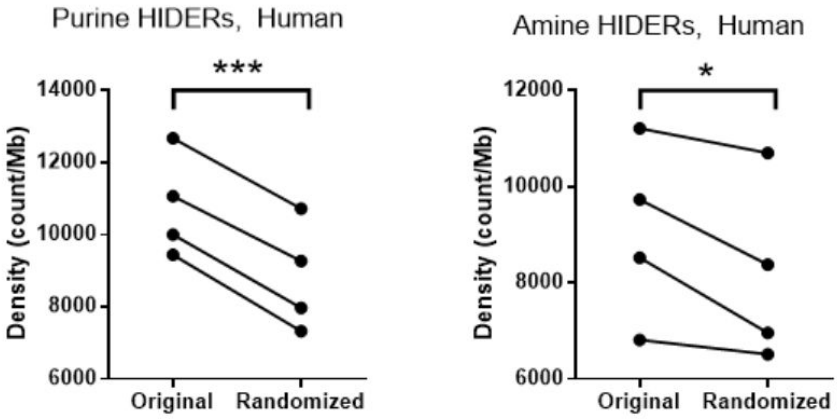
On the sequence level HIDERs are unique (nonrepetitive) sequences which are homologous to other sequences after recoding. On the physical level, we expect that these are mostly genomic repeats are engaged in resonant signaling and that HIDERs have evolved to take part in this signaling since they are partly similar to some of the genomic repeats or to each other.

Methods

We utilized the genomic data from the UCSC genome browser (https://genome.ucsc.edu). The repeats were masked using Repeat Masker (http://repeatmasker.org/) followed by a heuristic removal of repeats Ugene 1.32.0 (http://ugene.net/). Recoding was done as described in **Table [Codes]** using custom C++ programs, provided in the Supplement. HIDERs were detected by searching for similar fragments in the recoded sequences using Ugene. Thus obtained annotations were summarized in Google Sheets (https://www.google.com/sheets/). For statistics, random genomic fragments of predetermined size were picked, analyzed as above and the significance was determined using the t-test. As controls, randomized sequences were used. To reproduce the overall sequence structure and variations of nucleotide densities, randomization was done only on unmasked parts of the sequence 20 nucleotides at a time, the sequence was randomized in 20 bp bins using a custom C++ program.

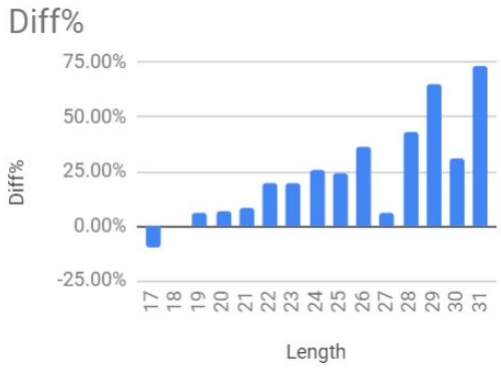
Results

Four 90 kb pieces were selected at random from the Human genome. The repeats were masked, the unmasked sequence contained no repeats. The sequence was recoded to the purine code as in Table [Codes]. The recoded sequence was searched for any homologies (HIDERs) longer than 19 bases and thousands of HIDERs were found. The sequence was randomized and the search for HIDERs was repeated. The counts are shown on Fig. [Counts].



The original sequence was found to contain 22% more HIDERs than randomized, P<0.001, suggesting that they are functional and are enriched in the process of evolution.

Other genomes were analyzed.



In Arabidopsis, Purine HIDERs demonstrated a strong increase of HIDERs' density with HIDERs' length. This also suggests that longer HIDERs are functional and preferentially selected for in the process of evolution.

Какие данные есть: Картинка соотношение между частотами количеством хоресов, хоресов в исходной последовательности и в уравномизированном (уравновешенном) виде по всем видам кодировок, по нескольким биологических видам. Можно посмотреть на распределение по длинам, зависимость длин, кодировок и видов не совпадает, не все так очевидно. по типу: периодические и непериодические повторы.

Обсудить, что какие-либо повторы могли образоваться естественным образом, но тот факт что в разных видах с разными частотами, подсказывает что это не из-за частотообразования. По одной кодировки еще могли образовываться с каким-либо предпочтением, но чтобы по всей – это больше похоже на резонанс.

Был факт. Теломерная последовательность объединена. Просто чередование пуринов-пиримидинов тоже объединено. Объяснить это тяжело. Особый вопрос –почему объединено? Есть какая-то выборка против.

Здесь мы упоминаем, что поле А может образоваться из-за полиаденилирования, что это тоже естественный химический процесс. Пурины и пиримидины могут образовываться друг из друга. Какие синонимические замены бывают? Таблица кодировки. Что на что заменяется? Эти частоты более

разрешены в геноме. Хотя малая часть в геноме занимается кодировкой белков, поэтому не должны влиять сильно.

Выборка против Стоп-кадонов. Мы не подбирали последовательности, а взяли первые попавшиеся. На них провели анализ. Обсудить параметры, почему именно такие параметры? Сначала репит маскером, потом ю-геном и от этого практически ничего не зависит. Разными способами показать, что не зависит. Параметры не подбирали специальным образом, чтобы усилить сигнал. Единственный параметр, который важен – это после какой частоты. И мы его четко показываем. Все остальные параметры не критичны. Это не доказательство того, что резонанс есть, возможное доказательство.

Эксперименты, которые могли бы быть сделаны (in vivo, in vitro), позволят показать что эти последовательности резонируют по какому-то определенному коду. Если код (структуру) нарушить, то что-то изменяется в функционале. Методы генной инженерии могут применяться для того, чтобы менять структуры и показывать зависимость функции от структуры. Важно упомянуть , что помимо хоресов, базары играют основную роль. хоресы являются дополнением к базарам. Отношение между хоресами и базарами радикально меняется в разных видах. Каким-то образом у млекопитающих произошел взрыв базаров. До этого базары были запрещены, у млекопитающих они выскочили. Надо упомянуть, что теломеры у разных видов разные. Хотелось бы сравнить частоты встречи теломер человека, млекопитающих, насекомых (5,6,7). Мое предположение – 6 у млекопитающих представлено ярче, хотя может быть и наоборот. Можно взять 2 или более видов с разной длиной теломеров, показать, что есть зависимость частот хоресов, резонирующих с теломерами. Зависит от формы теломера. Лучше 3: насекомые с правильной теламерой (7), млекопитающие (6) и растения (5). Показать влияние распределения теломерных хоресов.

Физическая терапия при помощи волн полезна, но не получила распространение, потому что не понимают принципа работы. Геномный резонанс мог бы объяснить посему свет и мм терапия полезны. Если удастся разгадать морфогенное поле, то мы в дамках.

Я совершено не понимаю, почему амины так названы и так срабатывают. Как срабатывает тиминовый код – это понятно. Присутствие тимина - это 1; отсутствие – 0. Есть несколько вариантов:

1)Важны оба;

2)Важен только тимин

3)Важен только нетимин

Сравнение:

1)тимин, нетимин, нетимин, нетимин (1000) -

2) тимин, нетимин, тимин, нетимин (1010) – тимин обязан быть на своем месте.

3) Важны другие места – главное чтобы тимин там отсутствовал.

Может быть важны оба варианта.

Пример. Парковочные места: вам важны только не занятые места. Расчёска расчёсывает даже тогда, когда у нее не хватает зубьев. Если у нее будет не хватать дырок – она не будет расчесывать.

Принцип перевеса – одна из двух букв важнее, чем вторая. Одна из двух частей кода важнее, чем вторая.

Binder, A., Parr, G., Hazleman, B., Fitton-Jackson, S., 1984. Pulsed electromagnetic field therapy of persistent rotator cuff tendinitis. A double-blind controlled assessment. Lancet 1, 695–698.

Bjordal, J.M., Couppé, C., Chow, R.T., Tunér, J., Ljunggren, E.A., 2003. A systematic review of low level laser therapy with location-specific doses for pain from chronic joint disorders. Aust. J. Physiother. 49, 107–116.
Burkov, V.D., Burlakov, A.B., Perminov, S.V., Kapranov, Y.S., Kufal, G.E., 2008. [Correction of Long Range Interaction Between Biological Objects Using Corner-Cube Reflectors]. Biomedical Radioelectronics 41–48.
Burlakov, A.B., Kapranov, Y.S., Kufal, G.E., Perminov, S.V., 2012. [About possible influence on biological object electromagnetic fields]. Weak and ultraweak fields and radiation in biology and medicine-Proceedings of IV International Congress 111–112.
Cifra, M., Fields, J.Z., Farhadi, A., 2011. Electromagnetic cellular interactions. Prog. Biophys. Mol. Biol. 105, 223–246.
Frohlich, H., 1988. Theoretical physics and biology. Biological Coherence and Response to External Stimuli. Berlin: Springer-Verlag 1–24.
Fröhlich, H., 1968. Long-range coherence and energy storage in biological systems. Int. J. Quantum Chem. 2, 641–649.
Gurwitsch, A., 1922. Über den Begriff des Embryonalen feldes. Wilhelm Roux Arch. Entwickl. Mech. Org. 51, 383–415.
Gurwitsch, A.A., 1988. A historical review of the problem of mitogenetic radiation. Experientia 44, 545–550.
Lowe, N.J., Prystowsky, J.H., Bourget, T., Edelstein, J., Nychay, S., Armstrong, R., 1991. Acitretin plus UVB therapy for psoriasis. Comparisons with placebo plus UVB and acitretin alone. J. Am. Acad. Dermatol. 24, 591–594.
Lushnikov, K.V., Shumilina, Y.V., Yakushina, V.S., Gapeev, A.B., Sadovnikov, V.B., Chemeris, N.K., 2004. Effects of low-intensity ultrahigh frequency electromagnetic radiation on inflammatory processes. Bull. Exp. Biol. Med. 137, 364–366.
Miller, R.A., Webb, B., 1973. Embryonic Holography: An Application of the Holographic Concept of Reality. DNA Decipher Journal 2.
Nan, T., Lin, H., Gao, Y., Matyushov, A., Yu, G., Chen, H., Sun, N., Wei, S., Wang, Z., Li, M., Wang, X., Belkessam, A., Guo, R., Chen, B., Zhou, J., Qian, Z., Hui, Y., Rinaldi, M., McConney, M.E., Howe, B.M., Hu, Z., Jones, J.G., Brown, G.J., Sun, N.X., 2017. Acoustically actuated ultra-compact NEMS magnetoelectric antennas. Nat. Commun. 8, 296.
Polesskaya, O., Guschin, V., Kondratev, N., Garanina, I., Nazarenko, O., Zyryanova, N., Tovmash, A., Mara, A., Shapiro, T., Erdyneeva, E., Others, 2018. On possible role of DNA electrodynamics in chromatin regulation. Prog. Biophys. Mol. Biol. 30, 1e5.
Sajadi, M., Furse, K.E., Zhang, X.-X., Dehmel, L., Kovalenko, S.A., Corcelli, S.A., Ernsting, N.P., 2011. Detection of DNA--Ligand Binding Oscillations by Stokes-Shift Measurements. Angew. Chem. Int. Ed. 50, 9501–9505.
Savelyev, I.V., Zyryanova, N.V., Polesskaya, O.O., Myakishev-Rempel, M., 2019. On The Existence of The DNA Resonance Code and Its Possible Mechanistic Connection to The Neural Code. Neuroquantology 17. https://doi.org/10.14704/nq.2019.17.2.1973
Scholkmann, F., Fels, D., Cifra, M., 2013. Non-chemical and non-contact cell-to-cell communication: a short review. Am. J. Transl. Res. 5, 586–593.
Scott, A.C., 1985. Soliton oscillations in DNA. Phys. Rev. A Gen. Phys. 31, 3518–3519.
Shi, Y., Choi, M., Li, Z., Kim, G., Foo, Z., Kim, H., Wentzloff, D., Blaauw, D., 2016. 26.7 A 10mm3 syringe-implantable near-field radio system on glass substrate, in: 2016 IEEE International Solid-State Circuits Conference (ISSCC). pp. 448–449.
Trushin, M.V., 2004. Distant non-chemical communication in various biological systems. Riv. Biol. 97, 409–442.
Usichenko, T.I., Ivashkivsky, O.I., Gizhko, V.V., 2003. Treatment of rheumatoid arthritis with electromagnetic millimeter waves applied to acupuncture points--a randomized double blind clinical study. Acupunct. Electrother. Res. 28, 11–18.
Volkov, S.N., Kosevich, A.M., 1987. Conformation oscillations of DNA. Mol. Biol. 21, 797–806.
Volodyaev, I., Beloussov, L.V., 2015. Revisiting the mitogenetic effect of ultra-weak photon emission. Front. Physiol. 6, 241.

Xu, J., Yang, F., Han, D., Xu, S., 2017. Wireless Communication in Biosystems. arXiv [physics.bio-ph].

SUPPLEMENT
Custom programs zip file.
Detailed methods and results

SUPPLEMENT

Search strategy.

The genomic sequences were searched for repeating sections. For the search, the UGENE program's Find Repeats algorithm was used. The algorithm allows customizing the minimum length of a repeating element.

The search for repeats took place in the original genomic sequences and degenerate, 4 types of recoding were considered

Purine (AG)/ Pyrimidine(CT)
Strong(CG) / Weak (AT)
Keto (GT) / Amino (AC)
Thymine (T and non-T(ACG))

To create degenerate sequences in a text editor, pairs of letters, for example, G is replaced with A and C is replaced with T (used Notepad ++)

In the genomic sequences, a two-stage search for duplicate elements was performed. At the first stage, the sequence was uploaded into the online Repeatmasker service (http://repeatmasker.org/cgi-bin/WEBRepeatMasker). At the second stage, the file with the masked N repetitions was checked by the Find Repeats algorithm built into UGENE.
The result was a sequence in which for some minimum length (17, 19 bases) and higher lengths there are no repeating elements.

The n-masked sequence was randomized in a special program (Juan) so that only portions of the sequence between N. were randomized.

The original and randomized sequences were transformed into degenerate codes.

In degenerate sequences, duplicate elements were searched. 100% of homologous degenerate regions have nonhomologous non-degenerate analogs.

For each repeating part of the sequence are known:
Repeating Sequence length
Coordinates of the beginning and end of the repeating sequence (index)

Repeating Sequence Code (ACGT)

4 species were considered: human, mouse, Drosophila melanogaster, Arabidopsis. Sequences of 90000 bases long were studied. For each species, 4 original sequences and 4 randomized were studied.

The search algorithm was tested for sensitivity to a type of degenerate code. For example, Purine code may look like AT, GT, AC, GC. The algorithm showed the independence of the results from the transformation method.

Results

Human Genome

4 original and 4 randomized sequences were investigated, each sequence was selected at arbitrary position and was 90Kb long. The selection was done only once.
1> hg38_dna range = chr1: 100000000-100090000
2> hg38_dna range = chr1: 100090001-100180000
3> hg38_dna range = chr1: 100180001-100270000
4> hg38_dna range = chr1: 100270001-100360000

Nucleotide statistics

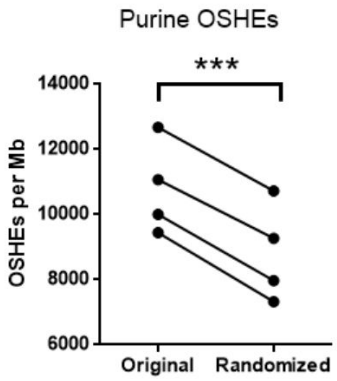
A:	A:	A:	A:
14 590	14 552	12 271	12 005
16.20%	16.20%	13.60%	13.30%
C:	C:	C:	C:
7 408	7 109	6 579	7 006
8.20%	7.90%	7.30%	7.80%
G:	G:	G:	G:
8 323	6 986	6 566	7 640
9.20%	7.80%	7.30%	8.50%
N:	N:	N:	N:
43 489	47 335	52 827	50 065
48.30%	52.60%	58.70%	55.60%
T:	T:	T:	T:
16 190	14 018	11 757	13 284
18.00%	15.60%	13.10%	14.80%

Repeat search algorithm settings: repeat length longer than 18 bases, 100% identity. The search for repetitions was performed in direct sequence. One repeat count is equivalent to a couple of repeating sequences.

Purine code (A (AG) - purines T (CT) - pyrimidines).
The conversion was done:
A to A
G to A
C to T
T to T
The letters in the right column were selected arbitrarily, so the homology can be tested using the U-gene program.

Seq	Orig	RND	Diff %
1	1141	965	15.42506573
2	996	834	16.26506024
3	850	659	22.47058824
4	900	717	20.33333333
	ttest	0.000090606403 75	

With high confidence, there is a difference between the number of purine repeats in the original and randomized sequences.



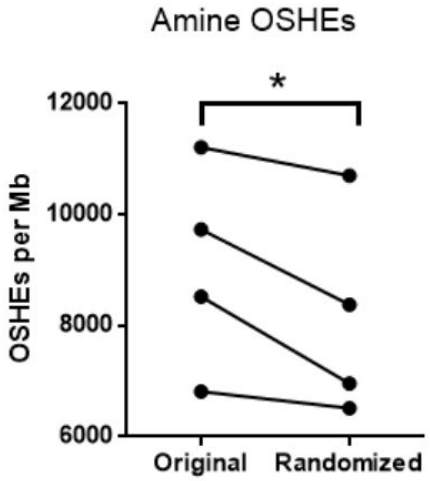
Strong code (A (AT) weak C (GC) strong)

Seq	Orig	RND	Diff %
1	2714	2618	3.54%
2	2368	2368	0.00%
3	1695	1600	5.60%
4	1736	1718	1.04%
	ttest	0.07	

The difference between the number of repeats in strong recoding between original and randomized sequences is significantly lower than for purine recoding.

Amine code (A (AC) amines G (GT) keto)

seq	Orig	RND	DIff%
1	1009	963	4.56%
2	876	754	13.93%
3	613	586	4.40%
4	767	626	18.38%
	ttest	0.03	



thymine code (T(T) ACG(G)):

Seq	Orig	RND	Diff%
1	2394	2391	0.13%
2	2342	2257	3.63%
3	1950	1962	-0.62%
4	2006	2006	0.00%
	ttest	0.4556750027	

The difference between the number of repeats in amine recoding between original and randomized sequences is lower than for purine recoding and higher than for strong recoding.

Repetition density data for various recodings.

The repetition density is equal to the ratio of the number of repetitions divided by the number of bases in the sequence. The percentage value expresses the probability of detecting a repeat for 1 base.

	Orig	RND	Diff %
Purine	2.54%	2.14%	15.43%
	2.21%	1.85%	16.27%
	1.89%	1.46%	22.47%
	2.00%	1.59%	20.33%
Strong	6.03%	5.82%	3.54%
	5.26%	5.26%	0.00%
	3.77%	3.56%	5.60%
	3.86%	3.82%	1.04%
Amin	2.24%	2.14%	4.56%
	1.95%	1.68%	13.93%
	1.36%	1.30%	4.40%
	1.70%	1.39%	18.38%

Results for mouse genome.

4 original and 4 randomized sequences were investigated.

1> mm10_dna range = chr3: 32500000-32590000
2> mm10_dna range = chr3: 32590001-32680000
3> mm10_dna range = chr3: 32680001-32770000
4> mm10_dna range = chr3: 32770001-32860000

Nucleotide statistics

A:	A:	A:	A:
14 906	13 385	13 603	9 367
16.60%	14.90%	15.10%	10.40%
C:	C:	C:	C:
10 592	10 248	10 556	7 807
11.80%	11.40%	11.70%	8.70%

G:	G:	G:	G:
11 522	9 810	11 458	8 469
12.80%	10.90%	12.70%	9.40%
N:	N:	N:	N:
36 651	43 714	39 807	53 214
40.70%	48.60%	44.20%	59.10%
T:	T:	T:	T:
16 330	12 843	14 573	11 143
18.10%	14.30%	16.20%	12.40%

Repeat search algorithm settings: repeat length longer than 16 bases, 100% identity. The search for repetitions was performed in direct sequence. One repeat count is equivalent to a couple of repeating sequences.

Purine code (A (AG) - purines T (CT) - pyrimidines)

Seq	Orig	RND	Diff%
1	4130	3853	6.71%
2	3351	3100	7.49%
3	3949	3557	9.93%
4	2304	2141	7.07%
	ttest	0.01052341811	

Compared to purine repeats in the human genome in the mouse genome, the difference between the original and random sequences is lower.

Strong code (T (AT) weak C (GC) strong)

Seq	Orig	RND	Diff%
1	4758	4406	7.40%
2	3547	3269	7.84%
3	4180	3815	8.73%
4	2491	2234	10.32%
	ttest	0.001341065586	

Compared to repeats in severe recoding in the human genome and in the mouse genome, the difference between the original and random sequences is higher.

Amine code (A (AC) amines G (GT) keto)

Seq	Orig	Rnd	Diff%
1	4127	3875	6.11%
2	3116	3011	3.37%
3	3631	3615	0.44%
4	2212	2103	4.93%
	ttest	0.09015002833	

thymine code (T(T) ACG(G)):

Seq	Orig	RND	Diff%
1	5126	5098	0.55%
2	4079	4022	1.40%
3	4754	4760	-0.13%
4	3143	3333	-6.05%
	ttest	0.6519794767	

Compared to amin repeats in the human genome in the mouse genome, the difference between the original and random sequences is lower.

Repetition density data for various recodings.

The repetition density is equal to the ratio of the number of repetitions divided by the number of bases in the sequence. The percentage value expresses the probability of detecting a repeat for 1 base.

	Orig	RND	Diff%
Purine	9.18%	8.56%	6.71%
	7.45%	6.89%	7.49%
	8.78%	7.90%	9.93%
	5.12%	4.76%	7.07%
Strong	10.57%	9.79%	7.40%
	7.88%	7.26%	7.84%
	9.29%	8.48%	8.73%

	5.54%	4.96%	10.32%
Amin	9.17%	8.61%	6.11%
	6.92%	6.69%	3.37%
	8.07%	8.03%	0.44%
	4.92%	4.67%	4.93%

Repeat density data for the human genome and mouse genome is not comparable because in human genome data repeat length is longer than 18 and in mouse genome data is longer than 16.

Results for Drosophila melanogaster genome

4 original and 4 randomized sequences were investigated.

dm6_dna range = chr2L: 200000-290000
dm6_dna range = chr2L: 400000-490000
dm6_dna range = chr2L: 800000-890000
dm6_dna range = chr2L: 1200000-1290000

Nucleotide statistics

A:	A:	A:	A:
24 185	23 248	23 314	21 882
26.90%	25.80%	25.90%	24.30%
C:	C:	C:	C:
19 153	19 507	21 042	17 586
21.30%	21.70%	23.40%	19.50%
G:	G:	G:	G:
19 347	19 430	20 693	17 448
21.50%	21.60%	23.00%	19.40%
N:	N:	N:	N:
3 145	4 031	2 459	10 382

3.50%	4.50%	2.70%	11.50%
T:	T:	T:	T:
24 171	23 785	22 493	22 703
26.90%	26.40%	25.00%	25.20%

Repeat search algorithm settings: repeat length longer than 16 bases, 100% identity. The search for repetitions was performed in direct sequence. One repeat count is equivalent to a couple of repeating sequences.

Purine code (A (AG) - purines T (CT) - pyrimidines)

seq	Orig	RND	Diff%
1	9441	9520	-0.84%
2	9334	9294	0.43%
3	9562	9668	-1.11%
4	8425	8138	3.41%
	t-test	0.7185432213	

Strong code (T (AT) weak G (GC) strong)

Seq	Orig	RND	Diff%
1	9416	9424	-0.08%
2	9340	9250	0.96%
3	9559	9605	-0.48%
4	8370	8236	1.60%
	ttest	0.3845699634	

Amine code (C (AC) amines G (GT) keto)

seq	orig	RND	Diff%
1	9633	9340	3.14%
2	9370	9150	2.40%
3	9444	9607	-1.70%

4	8312	8298	0.17%
	ttest	0.4429646252	

thymine code (T(T) ACG(G)):

Seq	Orig	RND	Diff%
1	5065	4989	1.50%
2	5118	4990	2.50%
3	5688	5569	2.09%
4	4324	4311	0.30%
	ttest	0.04931176011	

Repetition density data for various recodings.

The repetition density is equal to the ratio of the number of repetitions divided by the number of bases in the sequence. The percentage value expresses the probability of detecting a repeat for 1 base.

	Orig	RND	Diff%
Purine	20.98%	21.16%	-0.84%
	20.74%	20.65%	0.43%
	21.25%	21.48%	-1.11%
	18.72%	18.08%	3.41%
Strong	20.92%	20.94%	-0.08%
	20.76%	20.56%	0.96%
	21.24%	21.34%	-0.48%
	18.60%	18.30%	1.60%
Amino	21.41%	20.76%	3.04%
	20.82%	20.33%	2.35%
	20.99%	21.35%	-1.73%
	18.47%	18.44%	0.17%

The repeat density data for the data from the mouse genome and the Drosophila genome are comparable with each other (the same repeat length). When comparing the data, it can be seen that Drosophila has a higher density of degenerate repeats.
In Drosophila, the difference between the numbers of repeats in degenerate sequences (original vs randomized) is significantly less than for the mouse genome data.

Results for Arabidopsis thaliana genome

4 original and 4 randomized sequences were investigated.

> hub_329263_araTha1_dna range = chr3: 400000-490000
> hub_329263_araTha1_dna range = chr3: 600000-690000
> hub_329263_araTha1_dna range = chr3: 1490000-1580000
> hub_329263_araTha1_dna range = chr3: 1800000-1890000

A:	A:	A:	A:
27 498	27 757	27 352	28 409
30.60%	30.80%	30.40%	31.60%
C:	C:	C:	C:
16 978	16 126	16 869	16 569
18.90%	17.90%	18.70%	18.40%
G:	G:	G:	G:
16 503	16 491	16 180	16 201
18.30%	18.30%	18.00%	18.00%
N:	N:	N:	N:
708	1 363	2 293	522
0.80%	1.50%	2.50%	0.60%
T:	T:	T:	T:
28 297	28 263	27 311	28 297
31.40%	31.40%	30.30%	31.40%

Repeat search algorithm settings: repeat length longer than 16 bases, 100% identity. The search for repetitions was performed in direct sequence. One repeat count is equivalent to a couple of repeating sequences.

Comparing the data for the human genomes of the mouse and Drosophila, we found different statistical patterns. Exploring the Arabidopsis genome, we looked at additional data sets to refine the data.

For Arabidopsis, the number of repeats from 17 bases, the number of repeats from 19 bases, and the number of repeats of 17 and 18 bases in length were counted separately. Calculated the ratio between the indicators of 17+ and 19+ length repetitions

Purine code (A (AG) - purines T (CT) - pyrimidines)

	17+		
seq	Orig	RND	**diff%**
1	9994	9952	0.42%
2	9875	9793	0.83%

3	9587	9551	0.38%
4	10069	9919	1.49%
		ttest	0.05982588058

	19+		
seq	Orig	RND	**diff%**
1	3952	3717	5.95%
2	4087	3640	10.94%
3	4007	3559	11.18%
4	4087	3654	10.59%
	ttest	0.00489185544	

Seq	**the ratio between 19+ and 17+ diff%**
1	14.15
2	13.17
3	29.77
4	7.11

	17&18		
Seq	Orig	RND	diff%
1	6042	6235	-3.19%
2	5788	6153	-6.31%
3	5580	5992	-7.38%
4	5982	6265	-4.73%
	ttest	0.00736995623	

It can be seen that for repeats of shorter length (17,18) in randomized sequences of repeating elements there is significantly more than in original ones. The inverse relationship for longer repeats (more than 18). This suggests the uneven significance of purine repeats of different lengths in the Arabidopsis genome.

Strong code (T (AT) weak C (GC) strong)

	17+		
Seq	Orig	RND	diff%
1	10974	10605	3.36%
2	10811	10620	1.77%
3	10811	10620	1.77%
4	10687	10544	1.34%
	ttest	0.02064090793	

	19+		
Seq	Orig	RND	diff%
1	6177	5829	5.63%
2	6338	6056	4.45%
3	5935	5734	3.39%
4	6533	6158	5.74%
	ttest	0.004426852539	

Seq	the ratio between 19+ and 17+ diff%
1	1.675483797
2	2.518427039
3	1.916936974
4	4.289813202

	17&18		
Seq	Orig	RND	diff%
1	4797	4776	0.44%
2	4473	4564	-2.03%
3	4876	4886	-0.21%
4	4154	4386	-5.58%
	ttest	0.261315775	

The pattern of the greater difference in the number of repetitions between the original and randomized sequences for large repetition lengths is preserved for strong recoding

Amine code (A (AC) amines T(GT) keto)

	17+		
seq	Orig	RND	diff%
1	9998	9785	2.13%
2	9765	9760	0.05%
3	9676	9593	0.86%
4	10141	10018	1.21%
	ttest	0.09170330686	

	19+		
seq	Orig	RND	diff%
1	3806	3459	9.12%
2	3725	3581	3.87%
3	3604	3397	5.74%
4	3818	3645	4.53%
	ttest	0.01679889455	

seq	ratio between 19+ and 17+ diff%
1	4.279511717
2	75.49852349
3	6.69581322
4	3.735819205

	17&18		
seq	Orig	RND	diff%
1	6192	6326	-2.16%
2	6040	6179	-2.30%
3	6072	6196	-2.04%
4	6323	6373	-0.79%
	ttest	0.01265586913	

thymine code (T(T) ACG(G)):

Seq	Orig	RND	Diff%

1	7158	7261	-1.44%
2	6854	6836	0.26%
3	7089	6969	1.69%
4	7123	7034	1.25%
	ttest	0.5755986781	

The pattern in the distribution of repeat lengths in amino recoding is similar to purine. The difference between the distribution of relatively short (17, 18) and long (19+) repeats between the original and randomized sequences is noticeable.

Comparison of Interspecies results

General information about repeating elements.

The data on repeat density in degenerate sequences (purine, strong, and amine recoding) are calculated. Compared with the data on the number of masked N and unmasked elements for each sequence.

The percentage represents repetition density. The repetition density is equal to the probability of detecting repetition on 1 basis.

The ratio of non-repeating sequence elements to N-masked in non-degenerate sequences is calculated.

Human	19+	Orig1	RND1	Orig2	RND2	Orig3	RND3	Orig4	RND4
	P	4.91%	4.15%	4.67%	3.91%	4.57%	3.55%	4.51%	3.59%
	S	11.67%	11.26%	11.10%	11.10%	9.12%	8.61%	8.69%	8.60%
	A	4.34%	4.14%	4.11%	3.53%	3.30%	3.15%	3.84%	3.14%
	Base count	46512	46512	42666	42666	37174	37174	39936	39936
	Base to N ratio	1.069511831	1.069511831	0.9013626281	0.9013626281	0.7036931872	0.7036931872	0.7976830121	0.7976830121
Mouse	17+	Orig1	RND1	Orig2	RND2	Orig3	RND3	Orig4	RND4
	P	15.48%	14.44%	14.48%	13.39%	15.73%	14.17%	12.53%	11.64%
	S	17.84%	16.52%	15.33%	14.12%	16.66%	15.20%	13.54%	12.15%
	A	15.47%	14.53%	13.46%	13.01%	14.47%	14.40%	12.03%	11.43%
	Base count	53350	53350	46287	46287	50194	50194	36787	36787
	Base to N ratio	1.455621948	1.455621948	1.058859862	1.058859862	1.260934007	1.260934007	0.6913030406	0.6913030406
Drosophila	17+	Orig 1	RND1	Orig2	RND2	Orig3	RND3	Orig4	RND4
	P	21.74%	21.92%	21.71%	21.62%	21.85%	22.09%	21.16%	20.44%

	S	21.74%	21.70%	21.73%	21.52%	21.84%	21.94%	21.03%	20.69%
	A	22.18%	21.51%	21.80%	21.29%	21.58%	21.95%	20.88%	20.84%
	Base count	86856	86856	85970	85970	87542	87542	79619	79619
	Base to N ratio	27.61717011	27.61717011	21.32721409	21.32721409	35.60065067	35.60065067	7.668946253	7.668946253
Arabidopsis	17+	Orig1	RND1	Orig2	RND2	Orig3	RND3	Orig4	RND4
	P	22.20%	22.12%	21.94%	21.76%	21.30%	21.22%	22.38%	22.04%
	S	24.38%	23.56%	24.02%	23.60%	24.02%	23.60%	23.74%	23.44%
	A	22.22%	21.74%	21.70%	21.68%	21.50%	21.32%	22.54%	22.26%
	Base Count	89292	89292	88637	88637	87707	89478	89478	89478
	Base to N ratio	126.1186441	126.1186441	65.03081438	65.03081438	38.24989097	38.24989097	171.4137931	171.4137931

Below is a table of the relationship between the masked and unmasked portions of the sequences for all species studied. The ratio for the human genome is taken as 100%

Repeat density	
Human	100.00%
Mouse	82.03%
Drosophila	5.12%
Arabidopsis	1.17%

There are significant differences in the density of repeating sequences in the non-degenerate form of the genomes of various species. The inverse relationship is shown for the density of repeating sequences in degenerate encodings.
For different types of recoding, there is heterogeneity in the distribution of repeating sequences. We assume that this heterogeneity is associated with their participation in life.
We noticed a heterogeneity in the distribution of repeating sequences within a single recoding for different types of organisms, which suggests a possible evolutionary component in the organization of a degenerate code.
Of particular importance is the traced heterogeneity in the length distribution of the repeating elements.
The following is a study of the relationship between the number of repeating elements for different lengths.

We took 2 sequences: one original and one randomized for each type. Spent the grouping of repeating sequences by length. 17 and 18 base pairs, 19-23 base pairs, 24+ (more than 23 bases). For all recodings, the number of repeating sequences in the original and randomized sequences was compared.

Purine recoding:

		Purine / Pyrimidine		
	Length	**Repeat (couple) number**		
		Orig	RND	Diff%
Drosophila	17-18	11704	12302	-5.11%
	19-23	6870	6486	5.59%
	24+	311	255	18.01%
Mouse	17-18	5504	5334	3.09%
	19-23	2647	2303	13.00%
	24+	113	73	35.40%
Human	19-23	2184	1878	14.01%
	24+	101	55	45.54%
Arabidopsis	17-18	6042	6235	-3.19%
	19-23	1856	1777	4.26%
	24+	120	81.5	32.08%
The content of repeats 19-23 and 24+ in different types of organisms				
		Orig	RND	
	Drosophila	95.67%	96.22%	
		4.33%	3.78%	
	Mouse	95.91%	96.93%	
		4.09%	3.07%	
	Human	95.58%	97.15%	
		4.42%	2.85%	
	Arabidopsis	93.93%	95.61%	
		6.07%	4.39%	

Strong recoding:

		Strong/Weak		
	Length	**Repeat (couple) number**		
		Orig	RND	Diff%
Drosophila	17-18	10662	11002	-3.19%
	19-23	7533	7271	3.48%
	24+	638	576	9.72%
Mouse	17-18	5432	5194	4.38%
	19-23	3800	3388	10.84%
	24+	284	230	19.01%
Human	19-23	4662	4414	5.32%
	24+	766	822	-7.31%
Arabidopsis	17-18	4797	4776	0.44%
	19-23	5521	5192	5.96%
	24+	656	637	2.90%
	The content of repeats 19-23 and 24+ in different types of organisms			
	Orig	Rnd		
Drosophila	92.19%	92.66%		
	7.81%	7.34%		
Mouse	93.05%	93.64%		
	6.95%	6.36%		
Human	85.89%	84.30%		
	14.11%	15.70%		
Arabidopsis	89.38%	89.07%		
	10.62%	10.93%		

Amin recoding:

		Amine/Keto		
	Length	**Repeat (couple) number**		
		Orig	RND	Diff%
Drosophila	17-18	12466	12178	2.31%

	19-23	6546	6420	1.92%
	24+	254	262	-3.15%
Mouse	17-18	5600	5332	4.79%
	19-23	2568	2324	9.50%
	24+	86	94	-9.30%
Human	19-23	1924	1880	2.29%
	24+	94	46	51.06%
Arabidopsis	17-18	6192	6326	-2.16%
	19-23	3627	3323	8.38%
	24+	179	136	24.02%
	The content of repeats 19-23 and 24+ in different types of organisms			
	Orig	Rnd		
Drosophila	96.26%	96.08%		
	3.74%	3.92%		
Mouse	96.76%	96.11%		
	3.24%	3.89%		
Human	95.34%	97.61%		
	4.66%	2.39%		
Arabidopsis	95.30%	96.07%		
	4.70%	3.93%		

Выводы для распределения по длинам.