

Гены развития и домашнего хозяйства в геноме *Drosophila melanogaster*

Жимулев И.^{1*}, Ватолина Т.¹, Левицкий В.², Цуканов А.²

¹ Институт молекулярной и клеточной биологии СО РАН, Новосибирск, Россия

² Институт цитологии и генетики СО РАН, Новосибирск, Россия

* zhimulev@mcb.nsc.ru

Ключевые слова: политенные хромосомы; гены домашнего хозяйства; гены развития; мотивы промоторов; *Drosophila*; диски и междиски; bands; interbands; белки CHRIZ/CHROMATOR

Мотивация и цель: Проблему генов «домашнего хозяйства» и генов развития мы рассматривали еще в 1975 г. Для этого использовали политенные хромосомы *Drosophila melanogaster*. Эти хромосомы постоянно находятся в состоянии интерфазы, что позволило наблюдать гены как в активном, так и в неактивном состоянии.

Политенные хромосомы имеют поперечный рисунок дисков и междисков, а также гигантские размеры и хорошо различимые два типа дисков: черные и серые. О генетической организации этих хромосомных структур было мало известно. В начале 1970-х гг. внимание авторов привлекло то обстоятельство, что для постоянной активности генов материал хромосомы также должен быть постоянно открытым для транскрипции (или деконденсированным). Этим свойством обладают только междиски, а неактивные гены должны быть расположены в плотно конденсированных дисках, поэтому все гены генома были гипотетически разделены на две группы генов: развития и домашнего хозяйства, локализованные соответственно в междисках – гены домашнего хозяйства, и в дисках – гены развития [1].

Результаты: Лишь много лет спустя мы разработали методы, позволяющие идентифицировать материал дисков и междисков хромосом на молекулярной (физической) карте ДНК, в результате чего можно определить, какие гены расположены в той или иной структуре хромосомы. Был предложен метод локализации транспозонов в политенных хромосомах под электронным микроскопом (ЭМ) и последующего выявления сайта встраивания последовательности нуклеотидов на физической карте (например, транспозон P-lArB, содержащий концевые последовательности P-элемента, а также части lac-оперона *E. coli* и части генов дрозофилы *ADH*, *rosy* и плазмиды Bluescript). Из материала транспозона в хромосоме образовывался новый диск, что можно было увидеть под ЭМ, если внедрение происходило в междиск [2].

Затем были использованы данные, полученные в ходе реализации программы modENCODE (модельный организм ENCYclopedia of DNA Elements) [3], и, таким образом, картировали белки, расположенные в междисках близко к месту встраивания транспозона [2]. Использовали также данные трех типов классификаций хроматина в полном геноме дрозофилы [4, 6]. Для классификаций использовали модификации гистонов и белки хроматина (H3K36Me, RNAPolII, Мус, инсуляторные белки BEAF-32, Mod (mdg4), Su(Hw), Zw5, CTCF, CP190),

ассоциированные с промоторами генов белки Z4 и Chriz, а также белки, локализованные в ориджинах репликации (ORC2 и ORC6). Эти и другие белки активного хроматина были использованы для построения тепловых карт колокализации данных белков [4–6]. Далее мы сосредоточились на распределении и анализе типов хроматина, которые были идентифицированы с помощью 4НММ (скрытая марковская модель). Эта модель учитывает данные связывания белков «открытого хроматина» и других белков для клеточных линий Кс, S2, BG3 и Clon8. В результате сформировали четыре основных типа хроматина, названных нами аквамарином, лазуритом, малахитом и рубином [2, 7].

Согласно этой модели и данным последующих исследований, установлен белковый состав всех четырех состояний. Состояние хроматина «аквамарин» (13 % генома), локализованного в междисках политенных хромосом, имеет наиболее обширный перечень белков и модификаций гистонов. Самым распространенным состоянием хроматина в геноме является рубин (черные диски) (48 % генома). Проверка локализации этих типов хроматина ясно показала, что аквамарин связан с междисками со значительным обогащением междискового белка CHRO (CHRIZ) [2, 72] и 5'-UTR генов. Лазуритовый хроматин соответствует телам генов и морфологически соответствовал серым дискам, соседствующим с междисками. Сделан вывод о том, что ген «домашнего хозяйства» занимает две структуры политенных хромосом: промотор находится в междиске [2, 7], а тело гена – в сером диске ниже.

В результате анализа баз данных по данным RNA-seq обнаружено, что количество транскриптов в хроматине состояния «аквамарин» в 22–27 раз превышает таковое в хроматине состояния «рубин». Путем подсчета транскриптов в различных типах клеток дрозофилы, специфичных для той или иной стадии развития, в тканях или органах личинок и взрослых мух, культурах клеток и т. п. было показано, что транскрипционно активные гены в основном расположены в междисках (аквамариновом хроматине).

Для более глубокого понимания биологических функций белков, кодируемых генами, локализующимися в четырех различных состояниях хроматина в геноме дрозофилы, мы использовали биоинформатический ресурс Gene Ontology-DAVID [8] для анализа больших чисел генов. Установлено, что гены, промоторы которых связаны с хроматином «аквамарин», участвуют в метаболических процессах, таких как протеолиз, окислительно-восстановительный процесс, в различных процессах биосинтеза ДНК, РНК, белков, липидов и пептидов и т. д. Гены развития участвуют в восприятии запаха, вкуса, химических раздражителей, транспорте ионов, развитии кутикулы и метаболическом процессе хитина, размножении многоклеточных организмов, реакции на бактерии, поведение, нейропептидный сигнальный путь, реакция на углекислый газ, циркадный ритм и т. д.

В результате дальнейшего биоинформатического анализа мы установили, что геном содержит два типа генов: из 13 574 белок-кодирующих генов дрозофилы (FlyBase R 5.57) 6562 гена, 5'-области которых (от –300 до +200 п. н.) лежат в аквамарине, и 3162 гена развития – в рубине.

Используя инструмент STREME, мы провели поиск мотивов *de novo* в каждом из состояний хроматина и обнаружили четыре мотива, характерных только для генов домашнего хозяйства. Было обнаружено, что они содержат четыре специфических мотива, тогда как гены развития – только один ТАТА бокс. Около 30 % генов домашнего хозяйства не несут в себе ни одного из вышеупомянутых мотивов.

Выявлены особенности расположения генов этих групп в геноме: они различаются по длинам генов, интронов, экзонов и межгенных спейсеров, а также по количеству транскриптов, продуцируемых этими генами, и количеству экзонов в генах этих двух групп.

Гены развития имеют гораздо более длинные интроны, чем интроны генов домашнего хозяйства (в 4.5 раза больше). Гены домашнего хозяйства распределены значительно плотнее в геноме: межгенные спейсеры в пять раз короче, чем в кластерах генов развития.

Выводы: Предложена модель организации генома, позволяющая выделить и охарактеризовать две группы генов – домашнего хозяйства и развития, которые различаются по многим признакам: локализацией в структурах хромосом, различиями состояний хроматина, степени укладки материала хромосомы и наборами белков и модификаций гистонов в местах их локаций, организацией процесса репликации (ранняя и поздняя), обилием сайтов инициации репликации (ORC). Значительно различаются структура самих генов, распределение мотивов в промоторах генов, размеры межгенных спейсеров, длина и числа интронов и экзонов, числа транскриптов с одного гена, участие в процессах инсуляции хроматина различных белков. Таким образом, фактически в пределах одного генома существуют два генома, каждый со своей организацией.

Финансирование: Работа поддержана грантом РФФИ (№ 24-14-00133).

Developmental and housekeeping genes in the genome of *Drosophila melanogaster*

Zhimulev I.^{1*}, Vatolina T.¹, Levitsky V.², Tsukanov A.²

¹ Institute of Molecular and Cellular Biology, SB RAS, Novosibirsk, Russia

² Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

* zhimulev@mcb.nsc.ru

Key words: polytene chromosomes; housekeeping genes; developmental genes; promoter motifs; *Drosophila*; bands; interbands; CHRIZ/CHROMATOR proteins

Motivation and Aim: We considered the problem of “housekeeping” genes and developmental genes back in 1975. For this purpose we used polytene chromosomes of *Drosophila melanogaster*. These chromosomes are constantly active in interphase, which allowed us to observe genes in both active and inactive states.

Polytene chromosomes have a transverse pattern of bands and interbands, as well as giant size and well-distinguishable two types of bands: black and gray. Little was known about the genetic organization of these chromosomal structures. In the early 1970s, the authors' attention was drawn to the fact that for constant gene activity, the chromosome material must also be permanently open for transcription (or decondensed). Only interbands have this property, however inactive genes should be located in densely condensed bands, so all genes of the genome were hypothetically divided into two groups: developmental and housekeeping, localized respectively in interbands – housekeeping genes, and in bands – developmental genes [1].

Results: It was not until many years later that we developed methods to identify the bands and interbands material of chromosomes on a molecular (physical) map of DNA, which

could then determine which genes are located in a particular chromosome structure. A method was proposed to localize transposons in polytene chromosomes under an electron microscope (EM) and then identify the nucleotide sequence insertion site on the physical map (e. g., the P-lArB transposon containing the P-element end sequences, as well as parts of the lac operon of *E. coli* and parts of the *Drosophila ADH*, *rosy*, and Bluescript plasmid genes. The transposon material formed a new band in the chromosome, which can be seen under EM if the insertion took place in the interband [2].

Then, we used the data from the modENCODE program (model organism ENCyclopedia of DNA Elements) [3] and thus mapped proteins located in interbands close to the transposon insertion site [2]. We also used data from three classifications of chromatin in the full *Drosophila* genome [4, 6]. These are histone modifications H3K36Me, RNAPolII, Myc, insulator proteins BEAF-32, Mod (mdg4), Su(Hw), Zw5, CTCF, CP190, gene promoter-associated proteins Z4 and Chriz, and proteins localized in replication origins (ORC2 and ORC6). These and other active chromatin proteins were placed on a heat map [4–6]. Next, we focused on the distribution and analysis of chromatin types that were identified using 4HMM (hidden Markov model). This model takes into account the binding data of “open chromatin” proteins and other proteins for Kc, S2, BG3 and Clone8 cell lines. As a result, four main types of chromatin were formed, which we named aquamarine, lazurite, malachite, and ruby [2, 7].

According to this model and the data of subsequent studies, the protein composition of all four states has been established. The aquamarine chromatin state (13 % of the genome), localized in the interbands of polytene chromosomes, has the most extensive list of proteins and histone modifications. The most common chromatin state in the genome is ruby (black disks) (48 % of the genome. Inspection of the localization of these chromatin types clearly showed that aquamarine is associated with interbands, with significant enrichment of the interband protein CHRO (CHRIZ) ([2, 7] and 5'-ends of genes; lazurite chromatin corresponds to gene bodies and morphologically corresponded to gray bands adjacent to interbands. It is concluded that the housekeeping gene occupies two polytene chromosome structures: the promoter is located in the interband [2, 7] and the gene body is located in the gray band below.

Analysis of the RNA-seq databases revealed that the number of transcripts in the aquamarine state chromatin is 22–27 times higher than that in the ruby one. By counting transcripts in different types of *Drosophila* cells specific for this or that developmental stage, in tissues or organs of larvae and adult flies, cell cultures, etc., it was shown that transcriptionally active genes are mainly located in interbands (aquamarine chromatin). To better understand the biological functions of proteins encoded by genes localized in four different chromatin states in the *Drosophila* genome, we used the Gene Ontology-DAVID bioinformatics resource [8] for large gene number analysis. Genes whose promoters are associated with aquamarine chromatin were found to be involved in metabolic processes such as proteolysis, redox, various biosynthesis processes, DNA, RNA, proteins, lipids and peptides, etc. Developmental genes are involved in odor perception, taste, chemical stimuli, ion transport, cuticle development and chitin metabolic process, reproduction of multicellular organisms, response to bacteria, behavior, neuropeptide signaling pathway, response to carbon dioxide, circadian rhythm, etc.

Through further bioinformatics analysis, we found that the genome contains two types of genes: of the 13,574 protein-coding genes in *Drosophila* (FlyBase R 5.57),

6,562 genes whose 5'-regions (–300 to +200 bp) lie in aquamarine, and 3,162 developmental genes lie in ruby.

Using the STREME tool, we searched for de novo motifs in each of the chromatin states and found four motifs unique to housekeeping genes. Genes were found to contain four specific motifs, whereas developmental genes contain only one TATA box. About 30 % of housekeeping genes do not carry any of the above motifs.

The arrangement of genes from these groups in the genome was found to differ in the lengths of genes, introns, exons and intergenic spacers, as well as the number of transcripts produced by these genes and the number of exons in the genes of these two groups.

Developmental genes have much longer introns than the introns of housekeeping genes (4.5 times longer). Housekeeping genes are distributed much more densely in the genome: intergenic spacers are five times shorter in storage than in developmental gene clusters; their median value is 3769.5 bp in developmental stage genes and only 706 bp in housekeeping genes.

Conclusion: Genome organization model has been proposed to identify and characterize two groups of genes: housekeeping and developmental. They differ in many ways: localization in chromosome structures, differences in chromatin states, degree of stacking of chromosome material and sets of proteins and histone modifications at their localization sites, organization of the replication process (early and late), and abundance of replication initiation sites (ORCs). The structure of genes themselves, the distribution of motifs in gene promoters, the size of intergenic spacers, the length and number of introns and exons, the number of transcripts from one gene, and the participation of various proteins in chromatin insulation processes differ significantly. Thus, there are actually two genomes within one genome, each with its own organization.

Funding: The study is supported by RFS grant (No. 24-14-00133).

Список литературы/References

1. Zhimulev I.F., Belyaeva E.S. Proposals to the problem of structural and functional organization of polytene chromosomes. *Theor Appl Genet.* 1975;45(8):335-340
2. Zhimulev I., Vatolina T., Levitsky V., Tsukanov A. Developmental and housekeeping genes: Two types of genetic organization in the *Drosophila* genome. *Int J Mol Sci.* 2024;25(7):4068
3. The modENCODE Consortium et al. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science.* 2010;330(6012):1787-1797. doi 10.1126/science.1198374
4. Jamrich M., Greenleaf A.L., Bautz E.K. Localization of RNA polymerase in polytene chromosomes of *Drosophila melanogaster*. *Proc Natl Acad Sci USA.* 1977;74(5):2079-2083
5. Filion G.J. et al. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell.* 2010;143(2):212-224. doi 10.1016/j.cell.2010.09.009
6. Kharchenko P.V. et al. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature.* 2011;471(7339):480-485. doi 10.1038/nature09725
7. Zhimulev I.F. et al. Genetic organization of polytene chromosome bands and interbands in *Drosophila melanogaster*. *PLoS One.* 2014;9:e101631. doi 10.1371/journal.pone.0101631
8. Chintapalli V.R., Wang J., Dow J.A. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nat. Genet.* 2007;39(6):715-720. doi 10.1038/ng2049