УДК 573.2

# КОМПЬЮТЕРНЫЙ АНАЛИЗ И ФУНКЦИОНАЛЬНАЯ АННОТАЦИЯ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ AP2/ERF В ГЕНОМЕ *ARABIDOPSIS THALIANA* L.

© 2014 г. **О.А. Черных[1], В.Г. Левицкий[1, 2], Н.А. Омельянчук[1], В.В. Миронова[1, 2]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: kviki@bionet.nsc.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

Поступила в редакцию 12 сентября 2014 г. Принята к публикации 1 октября 2014 г.

У растений этилен участвует как в регуляции процессов развития, так и в ответе на стрессовые воздействия. Сигнал с рецепторов этилена активирует гены одного из самых больших семейств транскрипционных факторов, APETALA2/ETHYLENE RESPONSE FACTORs (ERFs). Сайты связывания ERF транскрипционных факторов содержат специфический GCCGCC-мотив и называются GCC-боксами. В настоящей работе две компьютерные программы для предсказания сайтов связывания транскрипционных факторов (oPWM и SiteGA) применены для анализа последовательностей экспериментально подтвержденных GCC-боксов. Распознаны GCC-боксы и исследовано их распределение в геноме *Arabidopsis thaliana* L. Проведены функциональная аннотация распознанных GCC-боксов и анализ их роли в ответе на фитогормон этилен.

**Ключевые слова:** этилен, транскрипционные факторы, сайты связывания, арабидопсис.

## ВВЕДЕНИЕ

Фитогормон этилен является единственным газообразным гормоном растений. Он участвует в контроле большого числа важнейших процессов в жизни растительного организма, начиная от роста и развития до ответов на стрессы биотического и абиотического происхождения (Shakeel *et al.*, 2013). За последние годы были идентифицированы многие ключевые компоненты сигнального пути от рецепторов этилена до связывания транскрипционных факторов (ТФ) с промоторами генов-мишеней (Fujimoto *et al.*, 2000).

Действующая модель пути передачи этиленового сигнала включает в себя следующие семейства генов: *ETR1-like*/*ETR2-like*, *CTR1, EIN2, EIN3/EIL1* и *ERF* (Ethylene Responsive Elements) (Shakeel *et al.*, 2013). Передача сигнала этилена осуществляется по механизму негативной регуляции, т. е. в отсутствие этилена все компоненты пути находятся в активированном состоянии, блокируя реакцию на этилен на транскрипционном уровне. Рецепторы этилена представлены двумя подсемействами (ETR1-like и ETR2-like) и расположены на мембране эндоплазматического ретикулума (Bleecker, Kende, 2000). Рецепторы находятся в комплексе с серин/треониновой протеинкиназой CTR1. При связывании этилена CTR1 становится неактивной, что приводит к дефосфорилированию и протеолитическому расщеплению следующего компонента пути EIN2 (Ju, Chang, 2012; Qiao *et al.*, 2012). EIN2 служит главным позитивным регулятором, от которого сигнал этилена передается к ТФ EIN3/EIL1 и ERF, локализованным в ядре.

Транскрипционные факторы активируются каскадно, т. е. один за другим. EIN3/EIL1 активирует ТФ ERF, который, в свою очередь, связывается со специфической последовательностью, называемой GCC-боксом и имеющей консенсус GCCGCC (Ohme-Takagi, Shinshi, 1995; Solano *et al.*, 1998). Консенсус GCCGCC

консервативен для многих видов растений и широко используется для распознавания ERF мишеней (Shinshi *et al.*, 1995; Fujimoto *et al.*, 2000; Choudhury *et al.*, 2008).

К настоящему моменту в промоторах отдельных генов различных видов растений исследованы GCC-боксы, например, в работах Solano с соавт. (1998) и Nakano с соавт. (2006) изучено влияние нуклеотидных замен в их последовательности. GCC-боксы находят в промоторах некоторых связанных с ответом на патогенные воздействия генов, многие из которых являются этилен-зависимыми (Stepanova, Ecker, 2000). Тем не менее только для небольшого числа потенциальных генов-мишеней экспериментально показана функциональность GCC-боксов в их промоторах в реакции на этилен.

В нашей работе с помощью компьютерных программ oPWM и SiteGA, разработанных для предсказания сайтов связывания ТФ, проведен поиск GCC-боксов в геноме *Arabidopsis thaliana* и изучены особенности их распределения. Мы также проанализировали связь между наличием в промоторах GCC-боксов и этилен-чувствительностью генов, используя данные полногеномных микрочип-экспериментов, в которых было исследовано изменение экспрессии генов в ответ на воздействие этилена.

## МАТЕРИАЛЫ И МЕТОДЫ

### Создание выборок GCC-боксов

На основе данных из публикаций нами были составлены обучающая выборка (1) и позитивная выборка (2). Обучающая выборка (1) состоит из 24 последовательностей, найденных в промоторах генов семи видов растений (табл. 1), и содержит нуклеотидные последовательности GCC-боксов, подтвержденных в экспериментах, исследовавших этилен-чувствительность генов. Позитивная выборка (2) составлена на основе данных из публикаций и содержит промоторы 54 генов *A. thaliana*, экспрессировавшихся в ответ на воздействие этилена. Отбор генов произведен по нескольким условиям: эффект этилена на экспрессию генов проверен методом количественного ПЦР (qRT-PCR), в публикации имеются точные данные по увеличению или уменьшению экспрессии генов под действием этилена. Экспрессия генов должна быть установлена в течение 3–24 ч после обработки этиленом. Гены позитивной выборки (2) не повторяют состав обучающей выборки (1).

### Распознавание потенциальных GCC-боксов

Выборка 24 экспериментально подтвержденных GGC-боксов собрана по литературным данным. Выравнивание сайтов было произведено с помощью разработанного нами подхода на основе генетического алгоритма (Mironova *et al.*, 2014). Для распознавания использованы разработанные нами ранее методы oPWM и SiteGA (Levitsky *et al.*, 2007).

Метод oPWM – модификация канонического метода весовых матриц, это позволяет рассматривать динуклеотидный контекст и в полной мере привлекать фланкирующие районы. Отличительная черта метода SiteGA состоит в учете зависимостей между удаленными позициями сайтов, что не предусмотрено в методе весовых матриц.

Лого, моно- и динуклеотидное лого на рис. 1, *а*, *б* представляют консервативные контекстные характеристики GCC-боксов, выявляемые с помощью метода oPWM. Согласно методике SiteGA, для распознавания GCC-боксов нами было отобрано $N = 60$ локально-позиционированных динуклеотидов (ЛПД), каждый из которых представлен позициями начала и конца участка и типом динуклеотида.

Для расчета наиболее значимых зависимостей между ЛПД мы провели следующий анализ. Для $60 \times (60 - 1)/2$ ЛПД были отобраны корреляции пар, значимость которых задана условиями $p < 0{,}05$, $p < 0{,}01$ или $p < 0{,}001$. Затем для каждого из этих условий рассчитана плотность ЛПД (рис. 1, *в*). Для одной корреляции вклад каждой позиции учитывался как $0{,}5/L_1$ и $0{,}5/L_2$ ($L_1$ и $L_2$ означают длины участков двух ЛПД). В данной работе проанализированы сайты, распознанные обоими методами.

### Анализ данных микрочип-эксперимента

Выборки данных микрочип-экспериментов взяты из базы NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/). Для GCC-боксов всего было проанализировано шесть экспериментов

**Таблица 1**

Обучающая выборка нуклеотидных последовательностей экспериментально подтвержденных GCC-боксов

| Ген | Сайт | Вид | Статья |
|---|---|---|---|
| *HLS1* | TTAACGCAGACATAGCCGCCATTTTCAACTTCTCACTC | *Arabidopsis thaliana* | Fujimoto *et al.*, 2000 |
| *PDF1.2a* | ATTCAGATTAACCAGCCGCCCATGTGAACGATGTAGCA | | Zarei *et al.*, 2011 |
| *Chn48* | ATAAAAAGGTAAGAGCCGCCACATAATATATGTAACCT | *Nicotiana tabacum* | Shinshi *et al.*, 1995 |
| *ACS3* | CTATTACATAGTAAGCCGCCCACCGTATCTCAAAATAG | | Zhang *et al.*, 2009 |
| *Cel5* | GTCACATTTTTATCGTCGCCGTGAATTGTGGTATAGTA | *Lycopersicon esculentum* | Tournier *et al.*, 2003 |
| *prb-1b* | CAAGTATGACTAATGGCGGCTCTTATCTCACGTGATG | *Nicotiana tabacum* | Sessa *et al.*, 1995 |
| *PR-5* | GGCCTTTACATTTAGCCGCCTAGCTCTATCTTTACCAA | *Nicotiana sylvestris* | Sato *et al.*, 1996 |
| *gln2* | GCCTCCTCATTAGAGCCGCCACTAAAATAAGACCGATC | *Nicotiana tabacum* | Grimmig *et al.*, 2003 |
| *ATHCHIB* | TTGATCACGAACCCGCCGCCTCATATTCATAATTAAAG | *Arabidopsis thaliana* | Samac *et al.*, 1990 |
| *CH5B* | TTCACGCTTGGGAAGCCGCCGGGGTGGGCCCGCAGAAA | *Phaseolus vulgaris* | Solano *et al.*, 1998 |
| *CHN50* | GGATGAAGCTAAAAGCCGCCAGTCTCACTAAGAAAAAT | *Nicotiana tabacum* | Nakano *et al.*, 2006 |
| *Osmotin* | TCTATGTGCGAAAAGCCGCCATACTCCTATATAAACCA | | |
| *RD29B* | AGAAACAAATGTATGTCGGCCAACAAGTTAATTTGGGT | *Arabidopsis thaliana* | Cheng *et al.*, 2013 |
| *ELI3-2* | CGGATTATGTCAACACCGCCATGGAACGGCTTGCAAAG | | |
| *GEA6* | GAGAGAAGAATTACACCGACGATTTCACCATGAAGAGA | | |
| *LEA4* | TATCTTGTTCTCTCGCCGACCAAGACTTGCTATAAATA | | |
| *COR15B* | GAAAAAAAAAGCAGGTCGGCCATGAAATTGTGGCTACA | | |
| *COR47-DRE1* | TCTTATTTCTTGAAGTCGGTAGATGAATATCATGATAT | | |
| *HSP101-DRE1* | CTTTAATTTATACAAGCCTCCTTTTTTGTACATCTATTT | | |
| *HSP70-DRE* | CTGAATTTTGACTTGCCGACTCCCCTGCTTGCTACTTT | | |
| *PDF1.2b* | AGTCAGATTAACCAGCCGCCCATGCAAAGCCAAAGCAG | *Arabidopsis thaliana* | Wang *et al.*, 2013 |
| *AT2G37130* | ACTTTCTTAATTATGGCGGCTGTAAATACATGTACAAT | | |
| *PMT1* | TATTATATCGAGTTGCGCCCTCCACTCCTCGTGTCCAA | *Nicotiana tabacum* | Sears *et al.*, 2014 |
| *AtHAK5* | TAAAAGTTTCAACAGCCGGCAATACGTGTTTGAGACGC | *Arabidopsis thaliana* | Son *et al.*, 2012 |
| *ACS* | AACACGTCATTGTTGCCGCCAACACTGAAGCTTCCTAT | *Musa acuminata* | Choudhury *et al.*, 2008 |
| *ACO* | GAGACCGATGGAAGCAGCCAAACTTGGTCCCCGATC | | |
| *BERF1* | TCCTCCATCACTGTGCCGCCCGTGCTCGCCTCTCCCGG | *Hordeum vulgare* | Osnato *et al.*, 2010 |

GCC-бокс подчеркнут. Выборка использована для применения методов распознавания SiteGA и oPWM.

(табл. 2). Из данных микрочип-экспериментов были взяты значения экспрессии генов при контрольной обработке растения воздухом и обработке этиленом. При наличии нескольких реплик эксперимента было рассчитано среднее значение экспрессии. Для расчета значимости был проведен t-тест. Ген считался значимо регулируемым этиленом, если (1) отношение Ф уровней экспрессии обработанного и контрольного образцов было не менее/более заданного порогового значения (Ф = 1,5/0,667 для активации/репрессии) и (2) отличие средних уровней экспрессии для обработанных и контрольных реплик было значимым по t-тесту ($p < 0,05$). Долю генов, содержащих предсказанные сайты и увеличивших или уменьшивших свою экспрессию под воздействием этилена, сравнили с долей всех генов в геномной выборке, имею-

щих такую же реакцию на этилен (табл. 3). Был проанализирован район [–300; +1] от старта инициации транскрипции (коровый промотор). Всего в анализ вошло 21 098 генов, для которых были аннотированы старты транскрипции (база ENSEMBL) и имелись данные экспериментов с микрочипами. Ассоциация между наличием GCC-бокса в коровом промоторе и этилен-зависимой экспрессией признавалась значимой, если была выявлена, как минимум, в двух экспериментах. Этот порог был выбран по биномиальному распределению (см. ниже).

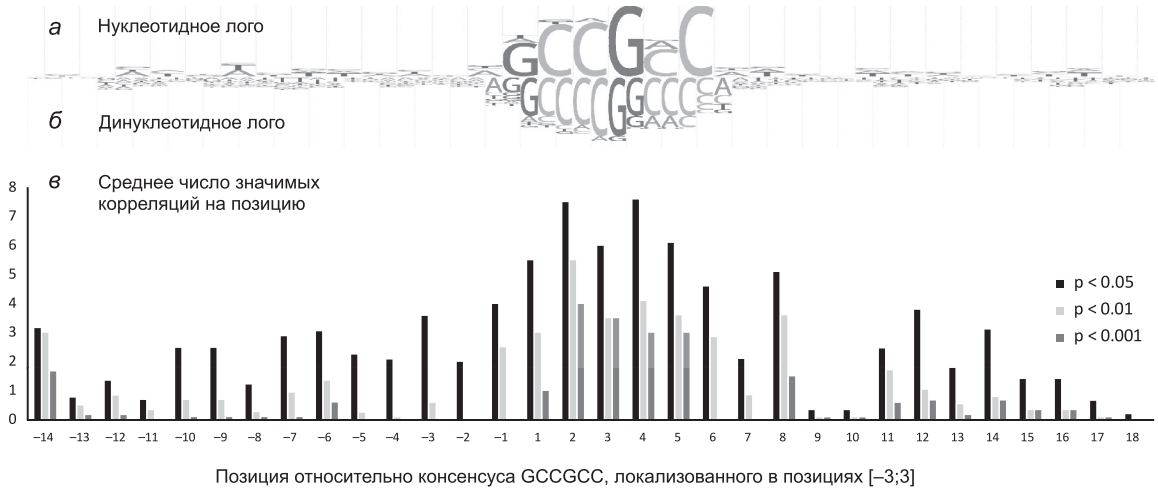


**Рис. 1.** Контекстные особенности GCC-сайтов, в которых функциональность GCC-боксов подтверждена экспериментально, выявленные методом oPWM (Levitsky *et al.*, 2007), осуществленным для моно- (*а*) и динуклеотидной (*б*) позиционно-весовой матрицы (PWM). Лого для PWMs рассчитаны в соответствии с Kulakovskiy и соавт. (2010) (*в*). Метод SiteGA (Levitsky *et al.*, 2007) выявил значимые корреляции между частотами локально-позиционированных динуклеотидов. По оси ординат показано среднее число значимых корреляций на позицию относительно консенсуса GCCGCC (ось абсцисс).

**Таблица 2**

Эксперименты с микрочипами, использованные для анализа ассоциаций наличия в промоторных районах генов GCC-боксов с этилен-чувствительной экспрессией этих генов

| GEO ID | Условия обработки | Ткань | Ссылка |
|---|---|---|---|
| GSE14247 | 10 ppm этилен 4 ч | Трехнедельные растения | Qiao *et al.*, 2012 |
| GSE39384 | 10 μM АЦК 1 ч | Семидневные проростки | Goda *et al.*, 2008 |
| GSE39384 | 10 μM АЦК 30 мин | | |
| GSE39384 | 10 μM АЦК 3 ч | | |
| GDS3505 | 10 ppm этилен 4 ч | Трехдневные проростки | Stepanova *et al.*, 2007 |
| GSE5174 | 10 ppm 4 ч | Этиолированные растения, выращенные в темноте | Olmedo *et al.*, 2006 |

Данные из базы NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/). Во всех экспериментах использованы данные по дикому фенотипу. Для проведения анализа взяты данные по изменению экспрессии генов при обработке растения воздухом (контроль) и этиленом. АЦК – 1-амино-циклопропан-1-карбоновая кислота – предшественник этилена.

**Таблица 3**

Схема для статистического анализа данных микрочип экспериментов с использованием углового преобразования Фишера для сравнения долей

| Число генов | Наличие предсказанного GGC-бокса | |
|---|---|---|
| | да | нет |
| Увеличивших/уменьшивших уровень экспрессии | *a* | *b* |
| Не изменивших уровень экспрессии | *c* | *d* |

Чтобы подтвердить статистическую значимость различий, посчитанных для выборок генов, мы использовали отношения *a/c* и *b/d* (табл. 3) между: 1) числом этилен-зависимых генов, значимо увеличивших или уменьшивших свою экспрессию в микрочиповом эксперименте (*a* и *b*) и 2) числом генов, не изменивших свою экспрессию в полногеномной выборке (*c* и *d*). Первая пропорция $p_1$ была рассчитана для выборки генов, в которых присутствовал потенциальный GCC-бокс. Вторая пропорция $p_2$ была рассчитана для генов, не содержащих этилен-чувствительные элементы.

Поскольку математическое ожидание значений *a* и *c* в некоторых случаях не превышало 10, для оценки значимости нами использован тест Фишера для сравнения долей (угловое преобразование Фишера для сравнения долей):

$$p_1 = 2 * arcsin\left(\sqrt{\frac{a}{a+c}}\right), p_2 = 2 * arcsin\left(\sqrt{\frac{b}{b+d}}\right).$$

Так как расчет отношений проводился для шести микрочип-экспериментов, для учета множественного сравнения мы применили биномиальное распределение, а именно рассчитали минимально необходимое количество микрочип-экспериментов для оценки найденной ассоциации как значимой ($p < 0{,}05$).

В результате расчет биномиального распределения

$$P(k) = \sum_{k}^{6} \frac{6!}{(6-k)!\,k!} 0.05^k 0.95^{6-k}$$

при $k = 3$ дает $P(3) = 0{,}0083 < 0{,}01$, тогда как $P(2) = 0{,}0394 < 0{,}05$. Следовательно, при $k = 2$, $k = 3$ имеем статистическую значимость, т. е. значимость изменения экспрессии в двух или более микрочип-экспериментах из шести можно рассматривать как неслучайное событие. Функциональная аннотация GCC-боксов проведена в системе AgriGO (Du *et al.*, 2010) с использованием инструмента SEA (Singular enrichment analysis).

## РЕЗУЛЬТАТЫ

### Распознавание GCC-боксов

Для распознавания GCC-боксов в геноме *A. thaliana* нами были созданы обучающая выборка (1), содержащая нуклеотидные последовательности экспериментально подтвержденных GCC-боксов, и позитивная выборка (2) этилен-чувствительных генов. Для полногеномного распознавания GCC-боксов мы применили две компьютерные программы, oPWM и SiteGA, основанные соответственно на методе весовых матриц и анализе локально-позиционированных динуклеотидов (Levitsky *et al.*, 2007). С помощью обеих программ исследованы нуклеотидные последовательности из обучающей выборки (1) (см. табл. 1). В результате анализа найдены особенности нуклеотидного контекста как в самом GCC-боксе, так и на его флангах (рис. 1).

Для распознавания был применен объединенный метод SiteGA&PWM, предполагающий распознавание потенциального сайта одновременно SiteGA и oPWM. С использованием позитивной выборки (2) были выбраны пороги обоих методов (0.934 SiteGA, 0.687 oPWM). При этом условии нами распознаются 15 из 27 сайтов обучающей выборки (1) (ошибка 1-го рода = 44 %). Как ошибку 2-го рода можно использовать вероятность распознавания сайтов для полного генома $1{,}9 \times 10^{-5}$.

### Распределение GCC-боксов в геноме *A. thaliana*

Найденные контекстные особенности (рис. 1) были использованы для распознавания GCC-боксов во всем геноме *A. thaliana*. Всего GCC-боксы были распознаны в промоторах [–2000; +1] 941 гена, что составляет 3,5 % генома *A. thaliana*. Для анализа особенностей распределения GCC-боксов мы рассчитали их относительную плотность распределения в различных районах генома: 5′- и 3′НТР, транскриптах, экзонах, интронах и промоторных районах (рис. 2, *а*).

Неожиданным оказалось, что повышенная плотность GCC-боксов найдена не только в промоторных районах, но и в экзонах, большую часть которых составляют кодирующие. Наибольшая плотность GCC-боксов обнаружена в экзонах, высокая плотность также была характерна для 5′НТР. Для исследования функциональной важности найденных особенностей мы проанализировали плотность распределения GCC-боксов в различных областях (5′НТР, экзонах, интронах, промоторных районах и 3′НТР) генов из позитивной выборки (2), составленной из нуклеотидных последовательностей 54 генов, для которых реакция на этилен была показана экспериментально.

Для этих генов, как и в целом геноме, повышенная плотность GCC-боксов была найдена в экзонах, но в отличие от полногеномного распределения в выборке этилен-чувствительных генов значимое обогащение было найдено еще и в промоторных областях [–1000; +1]. GCC-боксы в 5′НТР генов позитивной выборки не найдены. Далее мы проанализировали более

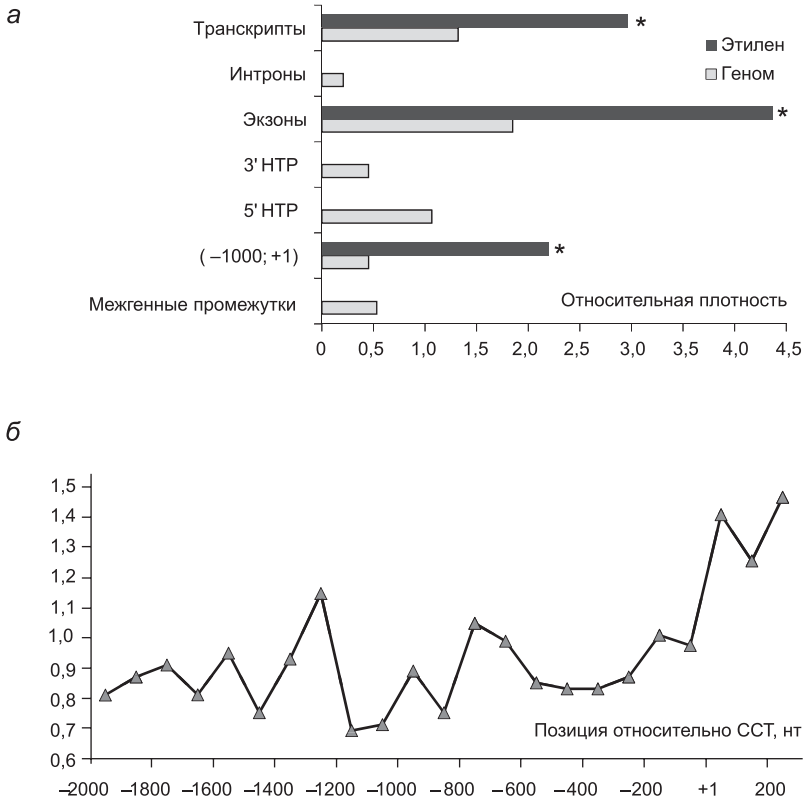


**Рис. 2.** Распределение GCC-боксов в геноме *A. thaliana*: *а* – плотность распределения GCC-боксов в разных областях генома для полного генома и позитивной выборки. Каждая плотность нормализована на среднюю плотность мотива в полном геноме. * отмечена значимость различий между плотностью в полном геноме и позитивной выборке по *t*-тесту ($p$ <0,05); *б* – относительная плотность распределения GCC-боксов вдоль регуляторных областей [–2000; +200] в полном геноме.

детально плотность распределения GCC-боксов вдоль регуляторных областей генов *A. thaliana*. GCC-боксы оказались локализованы вдоль промоторов неравномерно – плотность распределения была выше вокруг старта инициации транскрипции (рис. 2, *б*). Из проведенного анализа можно сделать следующие выводы. Значимое обогащение предсказанных методами PWM и SiteGA сайтов связывания ERF факторов в регуляторных районах этилен-чувствительных генов свидетельствует об адекватности распознавания GCC-боксов комбинированием этих двух методов.

Распознанные в настоящей работе GCC-боксы оказались неравномерно распределены в геноме *A. thaliana*, их повышенная плотность в кодирующих участках генома отмечена не только при анализе генома в целом, но и для выборки этилен-чувствительных генов, у которых плотность распределения GCC-боксов в экзонах оказалась при этом даже значимо выше, чем в целом по геному (рис. 2, *а*).

Обогащение GCC-бокса в экзонах может быть объяснено высокой встречаемостью тринуклеотида GCC в кодирующих участках, так как он является кодоном для одной из самых распространенных аминокислот – аланина. Альтернативное объяснение этого результата: GCC-бокс может быть дуоном – кодоном, который кроме своего прямого назначения составляет часть сайта связывания ТФ (Stergachis *et al.*, 2013). Однако нами было показано, что доля генов, изменивших свою экспрессию в ответ на этилен и несущих GCC-боксы в экзонах, не превышает доли генов, изменивших свою экспрессию в ответ на этилен в целом по геному. Таким образом, функциональность GCC-боксов в экзонах генов не подтверждена.

### Функциональная аннотация GCC-боксов

Функциональная аннотация GCC-боксов состояла из двух задач: (1) функциональная аннотация генов, содержащих в промоторах GCC-боксы, по геномной онтологии (ГО) и (2) поиск ассоциаций между наличием GCC-бокса в промоторе и этилен-чувствительностью генов. В анализе был использован список генов с предсказанными GCC-боксами в районах [–500; +1]. Выбор данного участка обоснован особенностями распределения GCC-бокса в 5′-регуляторных областях этилен-чувствительных генов и в среднем по геному, а именно: плотность распределения GCC-боксов в среднем по геному выше вокруг старта инициации транскрипции (рис. 2, *б*), при этом в 5′НТР этилен-чувствительных генов сайты не были обнаружены (рис. 2, *а*).

### Функциональная аннотация генов, содержащих GCC-боксы

Для исследования процессов, связанных с экспрессией генов, содержащих потенциальные GCC-боксы, мы провели функциональную аннотацию генов, в промоторах которых они были распознаны. Число генов с GCC-боксами только в прямой ориентации – 2 029. Число генов, содержащих GCC-боксы в обратной ориентации, – 1 941. Число генов с GCC-боксами в обеих цепях – 3 971. Аннотация проводилась в системе AgriGO (Du *et al.*, 2010). В результате выявлен ряд терминов, значимо обогащенных для трех выборок генов, содержащих GCC-боксы в: (1) прямой (GCC+); (2) обратной (GCC–); или (3) любой ориентации (GCC+/–) относительной цепи транскрипции.

На рис. 3 представлена блок-схема терминов ГО, относящихся к функциям генов, содержащих GCC-боксы. Значимо обогащенные термины обозначены светло- и темно-серым ($p < 0{,}05$ и $p < 10^{-4}$ по тесту Бенджамини соответственно). Термины ГО, значимо обогащенные для генов GCC+ и GCC–, отличались. GCC+ оказались ассоциированы с метаболизмом азотных соединений, а GCC– с организацией клеточных компонентов. Однако, когда мы объединили гены, содержащие GCC-боксы в обеих ориентациях, были найдены дополнительные значимо обогащенные термины, в том числе связанные с постэмбриональным развитием ($p < 0{,}007$ по тесту Бенджамини).

### Анализ взаимосвязи между наличием в промоторе GCC-бокса и этилен-чувствительностью гена

Для проверки функциональности потенциального GCC-бокса в промоторе и его связи с этилен-зависимой экспрессией генов нами

был проведен анализ данных шести микрочип-экспериментов, в которых было исследовано воздействие этилена на проростки *A. thaliana* (табл. 2). Для отдельного полногеномного эксперимента сравнили доли генов, увеличивших или уменьшивших свою экспрессию под воздействием этилена в среднем по геному с таковой для выборки генов, содержащих предсказанные GCC-боксы (табл. 4). Связь между наличием потенциальных GCC-боксов

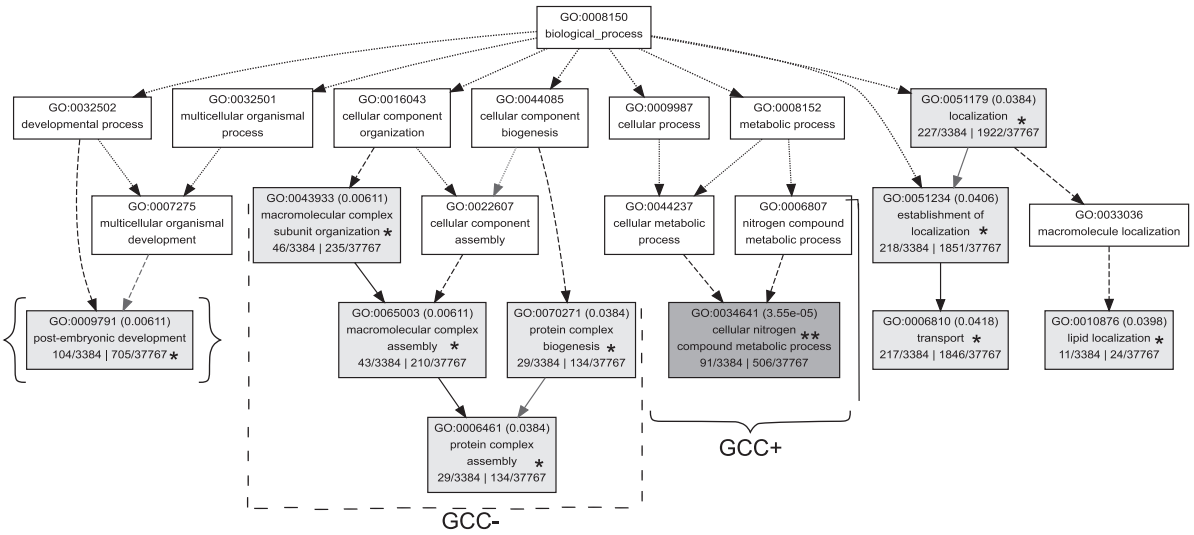


**Рис. 3.** Блок-схема терминов GO (Gene Ontology), значимо обогащенных для генов, содержащих GCC-боксы.

Фигурными скобками отмечены термины, обогащенные для GCC+ (сайт в прямой ориентации), штриховой линией – термины для GCC– (сайт в обратной ориентации). В двойных фигурных скобках – термины, связанные с развитием. Значимость обогащения по тесту Бенджамини: $p < 0{,}05$* (светло-серые блоки), $p < 10^{-4}$** (темно-серые). В блоках указано число распознанных генов (*x, y*) в формате *x*/3 384, *y*/37 767, где 3 384 – число аннотированных генов, содержащих GCC-бокс из 3 971 гена *A. thaliana* и поданных нами на анализ в систему AgriGO, а 37 767 – число генов, составляющих, по последним данным, полный геном *A. thaliana*.

**Таблица 4**

Ассоциация наличия GCC-боксов в промоторных районах генов ([–300; +1] относительно сайтов старта транскрипции) с этилен-зависимой активацией экспрессии генов

| GEO ID | Цепь | Число генов с GCC-боксами, увеличившими экспрессию (*a*) | Число генов с GCC-боксами (*c*) | Отношения *a/c* \| *b/d*, % | *p*-value |
|---|---|---|---|---|---|
| GSE14247 | +/– | 141 | 631 | 22,3 \| 14,3 | $9{,}0 \times 10^{-8}$ |
| GSE14247 | + | 83 | 372 | 22,3 \| 14,3 | $3{,}1 \times 10^{-5}$ |
| GSE14247 | – | 59 | 267 | 22,1 \| 14,3 | $4{,}6 \times 10^{-4}$ |
| GDS3505 | +/– | 17 | 631 | 2,7 \| 1,7 | 0,04 |
| GSE3938 3 ч | – | 7 | 267 | 2,6 \| 0,9 | $1{,}4 \times 10^{-2}$ |

Представлены значимые результаты для прямой (+), обратной (–) и обеих (+/–) цепей, полученные при анализе данных шести микрочип-экспериментов (см. табл. 2): *a* – число генов, содержащих в промоторе хотя бы один GCC-бокс и увеличивших свою экспрессию в ответ на воздействие этилена; *c* – число генов, содержащих GCC-бокс в промоторе, среди всех проанализированных генов генома (D = 21 098). Также указаны отношения *a/c* и *b/d*, с помощью которых были высчитаны значимости (*p*-value). Объяснение расчетов с величинами *a* – *d* см. в табл. 3.

и этилен-зависимой экспрессией признавали значимой, если значимые различия в экспрессии генов обнаруживались более чем в двух микрочип-экспериментах, что соответствует $p < 0,05$ по биномиальному распределению.

В результате проведенного анализа показано, что наличие GCC-боксов в промоторном районе [–300; +1] значимо ассоциировано с активацией экспрессии генов в ответ на воздействие этилена на растения (табл. 4). Случаев подавления экспрессии для генов, содержащих GCC-боксы, не выявлено. Мы также проанализировали, влияет ли ориентация GCC-бокса в промоторе гена на его этилен-чувствительность. Активация генов в ответ на воздействие этилена оказалась характерной для обратной ориентации (GCC–). Оба результата были выявлены в двух из шести микрочип-экспериментов (значимость $p < 0,05$ по биномиальному критерию). Для прямой ориентации (GCC+) была найдена одна ассоциация с активацией экспрессии в ответ на воздействие этилена, что по биномиальному распределению не значимо. Для сравнительной оценки функциональной значимости GCC-боксов в различных их ориентациях необходимо привлечь результаты дополнительных микрочип-экспериментов по оценке изменения экспрессии генов в ответ на воздействие этилена. Известно, что семейство ТФ ERF, специфически связывающихся с GCC-боксами, очень велико (125 генов у *A. thaliana*) (Nakano *et al.*, 2006) и лишь некоторые представители этого семейства изменяют свою экспрессию в ответ на этилен. Выявленные ассоциации свидетельствуют в пользу адекватности примененных нами методов распознавания GCC-боксов, так как с их использованием получены значимые результаты как по функциональной аннотации, так и по анализу данных микрочип-экспериментов.

## ЗАКЛЮЧЕНИЕ

Нами проведен биоинформатический анализ последовательностей GCC-боксов, которые являются сайтами связывания ТФ семейства ERF. Наличие GCC-бокса в промоторе гена исследователи ассоциируют с их этилен-чувствительностью (Fujimoto *et al.*, 2000; Stepanova *et al.*, 2007). Первые GCC-боксы были найдены в генах, значимо изменяющих свою экспрессию в ответ на этилен и связанных с ответом на стрессовые воздействия (Sessa *et al.,* 1995; Sato *et al.*, 1996). Однако, так как ранее полногеномный анализ GCC-боксов не проводился, достоверность этой ассоциации была неизвестна. Нами впервые осуществлен анализ распределения GCC-боксов в геноме *A. thaliana*. В результате показано, что GCC-боксы действительно обогащены в проксимальных районах этилен-чувствительных генов и ассоциированы с активирующим эффектом этилена. Установлено, что этилен не только увеличивает, но и уменьшает экспрессию ряда генов (Hess *et al.*, 2011; Cheng *et al.*, 2013; Mase *et al.*, 2013). Вероятно, на характер изменения экспрессии генов влияет локализация GCC-боксов в прямой или обратной цепи относительно старта инициации транскрипции. Также возможно влияние других ТФ в пути передачи этиленового сигнала, например, EIN3/EIL1 (Solano *et al.*, 1998; Alonso *et al.*, 2003), сайты связывания которых изучены недостаточно. Детальное прояснение этого вопроса и роли ориентации сайтов относительно старта инициации транскрипции для проксимального района является целью наших ближайших исследований.

Предсказанные нами GCC-боксы оказались значимо обогащены не только в промоторах, но и в кодирующих частях этилен-чувствительных генов. Это может свидетельствовать либо о том, что GCC-бокс может быть частью дуона (Stergachis *et al.*, 2013), либо в ряде случаев о том, что регулируемые ERF-гены могут экспрессироваться с альтернативных стартов транскрипции, а значит, сайт располагается в 5′-области для некоторых изоформ, кодируемых геном. Однако нами было показано, что предсказанные сайты в экзонах не имеют функционального значения. Изучение этого вопроса требует более детального анализа с привлечением дополнительных экспериментальных данных.

Также нами выявлены некоторые закономерности влияния ориентации GCC-боксов относительно старта инициации транскрипции на функцию гена. Группы генов, имеющих GCC-боксы в обратной ориентации, были значимо ассоциированы с увеличением экспрессии в ответ на этилен. Для GCC-боксов в прямой ориентации нами не было показано значимой

связи с этилен-чувствительностью. Интересно, что гены, содержащие GCC-боксы в определенной ориентации, оказались специфически обогащены другими терминами генной онтологии помимо постэмбрионального развития. Это свидетельствует о том, что GCC-боксы могут быть функциональны для ТФ ERF, которые не участвуют в генной сети передачи этилена, но участвуют в других важных для метаболизма клетки процессах.

## БЛАГОДАРНОСТИ

Выражаем благодарность И.В. Мироновой за помощь в составлении обучающей выборки и И.В. Медведевой за экстракцию данных из базы ENSEMBL. Работа поддержана грантом РФФИ-12-04-33112-мол-а-вед и бюджетным проектом ИЦиГ СО РАН VI.61.1.2.

## ЛИТЕРАТУРА

Alonso J.M., Stepanova A., Solano R. *et al.* Five Components of the Ethylene-Response Pathway Identified in a Screen for Weak Ethylene-Insensitive Mutants in Arabidopsis // Proc. Natl AS USA. 2003. V. 100. No. 5. P. 2992–2997.

Bleecker A.B., Kende H. Ethylene: A Gaseous Signal Molecule in Plants // Ann. Rev. Cell Developmental Biology. 2000. V. 16. No. 1. P. 1–18.

Cheng M., Liao P., Kuo W., Lin T. The Arabidopsis ETHYLENE RESPONSE FACTOR1 Regulates Abiotic Stress-Responsive Gene Expression by Binding to Different Cis-Acting Elements in Response to Different Stress Signals // Plant Physiology. 2013. V. 162. No. 3. P. 1566–1582.

Choudhury S.R., Roy S. *et al.* Characterization of Differential Ripening Pattern in Association with Ethylene Biosynthesis in the Fruits of Five Naturally Occurring Banana Cultivars and Detection of a GCC-Box-Specific DNA-Binding Protein // Plant Cell Reports. 2008. V. 27. No. 7. P. 1235–1249.

Du Z., Zhou X., Ling Y. *et al.* agriGO: A GO Analysis Toolkit for the Agricultural Community // Nucleic Acids Research. 2010. V. 38. P. 64–70.

Fujimoto S., Ohta M. *et al.* Arabidopsis Ethylene-Responsive Element Binding Factors Act as Transcriptional Activators or Repressors of GCC Box-Mediated Gene Expression // The Plant Cell. 2000. V. 12. No. 3. P. 393–404.

Goda H., Sasaki E., Akiyama K. *et al.* The AtGenExpress Hormone and Chemical Treatment Data Set: Experimental Design, Data Evaluation, Model Data Analysis and Data Access // Plant J.: For Cell Molecular Biology. 2008. V. 55. No. 3. P. 526–542.

Grimmig B., Gonzalez-Perez M. *et al.* Ozone-Induced Gene Expression Occurs via Ethylene-Dependent and Independent Signalling // Plant Molecular Biology. 2003. V. 51. No. 4. P. 599–607.

Hess N., Klode M. *et al.* The Hypoxia Responsive Transcription Factor Genes ERF71/HRE2 and ERF73/HRE1 of Arabidopsis Are Differentially Regulated by Ethylene // Physiologia Plantarum. 2011. V. 143. No. 1. P. 41–49.

Ju C., Chang C. Advances in Ethylene Signalling: Protein Complexes at the Endoplasmic Reticulum Membrane // AoB Plants. 2012. No. 1.

Kulakovskiy I.V., Boeva V.A., Favorov A.V., Makeev V.J. Deep and Wide Digging for Binding Motifs in ChIP-Seq Data // Bioinformatics. 2010. V. 26. No. 20. P. 2622–2623.

Levitsky V.G., Ignatieva E.V., Ananko E.A. *et al.* Effective Transcription Factor Binding Site Prediction Using a Combination of Optimization, a Genetic Algorithm and Discriminant Analysis to Capture Distant Interactions // BMC Bioinformatics. 2007. V. 8. No. 1. P. 481.

Mase K., Ishihama N., Mori H. *et al.* Ethylene-Responsive AP2/ERF Transcription Factor MACD1 Participates in Phytotoxin-Triggered Programmed Cell Death // Molecular Plant-Microbe Interactions. 2013. V. 26. No. 8. P. 868–879.

Mironova V., Omelyanchuck N., Levitsky V. Computational analysis of Auxin Responsive Elements in *Arabidopsis thaliana* Genome // BMC Genomics review. 2014. In print.

Nakano T., Suzuki K., Fujimura T., Shinshi H. Genome-Wide Analysis of the ERF Gene Family in Arabidopsis and Rice // Plant Physiology. 2006. V. 140. No. 2. P. 411–432.

Ohme-Takagi M., Shinshi H. Ethylene-Inducible DNA Binding Proteins That Interact with an Ethylene-Responsive Element // Plant Cell. 1995. V. 7. No. 2. P. 173–182.

Olmedo G., Guo H., Gregory B. *et al.* ETHYLENE-INSENSITIVE5 Encodes a 5′→3′ Exoribonuclease Required for Regulation of the EIN3-Targeting F-Box Proteins EBF1/2 // Proc. Natl AS USA. 2006. V. 103. No. 36. P. 13286–13293.

Osnato M., Stile M.R., Wang Y. *et al.* Cross Talk between the KNOX and Ethylene Pathways Is Mediated by Intron-Binding Transcription Factors in Barley // Plant Physiology. 2010. V. 154. No. 4. P. 1616–1632.

Qiao H., Shen Z., Huang S. *et al.* Processing and Subcellular Trafficking of ER-Tethered EIN2 Control Response to Ethylene Gas // Science. 2012. V. 338. No. 6105. P. 390–393.

Samac D. A., Hironaka C.M., Yallaly P.E., Shah D.M. Isolation and Characterization of the Genes Encoding Basic and Acidic Chitinase in Arabidopsis Thaliana // Plant Physiology. 1990. V. 93. No. 3. P. 907–914.

Sato F., Kitajima S. *et al.* Ethylene-Induced Gene Expression of Osmotin-like Protein, a Neutral Isoform of Tobacco PR-5, Is Mediated by the AGCCGCC Cis-Sequence // Plant Cell Physiology. 1996. V. 37. No. 3. P. 249–255.

Sears M., Zhang H., Rushton P. *et al.* NtERF32: A Non-NIC2 Locus AP2/ERF Transcription Factor Required in Jasmonate-Inducible Nicotine Biosynthesis in Tobacco // Plant Molecular Biology. 2014. V. 84. No. 1-2. P. 49–66.

Sessa G., Meller Y., Fluhr R. A GCC Element and a G-Box Motif Participate in Ethylene-Induced Expression of the PRB-1b Gene // Plant Molecular Biology. 1995. V. 28. No. 1. P. 145–153.

Shakeel S.N., Wang X., Binder B.M., Schaller G.E. Mechanisms of Signal Transduction by Ethylene: Overlapping

and Non-Overlapping Signalling Roles in a Receptor Family // AoB Plants. 2013. V. 5. No. 1.

Shinshi H., Usami S., Ohme-Takagi M. Identific ation of an Ethylene-Responsive Region in the Promoter of a Tobacco Class I Chitinase Gene // Plant Molecular Biology. 1995. V. 27. No. 5. P. 923–932.

Solano R., Stepanova A., Chao Q., Ecker J.R. Nuclear Events in Ethylene Signaling: A Transcriptional Cascade Mediated by ETHYLENE-INSENSITIVE3 and ETHYLENE-RESPONSE-FACTOR1 // Genes Development. 1998. V. 12. No. 23. P. 3703–3714.

Son G.H., Wan J., Kim H. *et al.* Ethylene-Responsive Element-Binding Factor 5, ERF5, Is Involved in Chitin-Induced Innate Immunity Response // Molecular Plant-Microbe Interactions. 2012. V. 25. No. 1. P. 48–60.

Stepanova A.N., Ecker J.R. Ethylene Signaling: From Mutants to Molecules // Current Opinion Plant Biology. 2000. V. 3. No. 5. P. 353–360.

Stepanova A.N., Yun J., Likhacheva A.V., Alonso J.M. Multilevel Interactions between Ethylene and Auxin in Arabidopsis Roots // Plant Cell. 2007. V. 19. No. 7. P. 2169–2185.

Stergachis A., Haugen E. *et al.* Exonic Transcription Factor Binding Directs Codon Choice and Affects Protein Evolution // Science. 2013. V. 342. No. 6164. P. 1367–1372.

Tournier B., Sanchez-Ballesta M. *et al.* New Members of the Tomato ERF Family Show Specific Expression Pattern and Diverse DNA-Binding Capacity to the GCC Box Element // FEBS Letters. 2003. V. 550. No. 1-3. P. 149–154.

Wang P., Du Y., Zhao X. *et al.* The MPK6-ERF6-ROS-Responsive Cis-Acting Element7/GCC Box Complex Modulates Oxidative Gene Transcription and the Oxidative Response in Arabidopsis // Plant Physiology. 2013. V. 161. No. 3. P. 1392–1408.

Zarei A., Körbes A.P., Younessi P. *et al.* Two GCC Boxes and AP2/ERF-Domain Transcription Factor ORA59 in Jasmonate/ethylene-Mediated Activation of the PDF1.2 Promoter in Arabidopsis // Plant Molecular Biology. 2011. V. 75. No. 4-5. P. 321–231.

Zhang Z., Zhang H., Quan R., Wang X.C., Huang R. Transcriptional Regulation of the Ethylene Response Factor LeERF2 in the Expression of Ethylene Biosynthesis Genes Controls Ethylene Production in Tomato and Tobacco // Plant Physiology. 2009. V. 150. No. 1. P. 365–377.

# COMPUTATIONAL ANALYSIS AND FUNCTIONAL ANNOTATION OF AP2/ERF TRANSCRIPTION FACTOR BINDING SITES IN *ARABIDOPSIS THALIANA* L. GENOME

**O.A. Chernykh[1], V.G. Levitsky[1, 2], N.A. Omelyanchuk[1], V.V. Mironova[1, 2]**

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: kviki@bionet.nsc.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia

## Summary

The plant hormone ethylene regulates both developmental processes and various stress responses in plants. Ethylene perception in plants is followed by activation of some transcription factors from the large family of APETALA2/ETHYLENE response factors (ERFs). ERF TF binding sites contain a specific GCCGCC motif, called GCC-box. In this study, we applied TF binding site recognition tools oPWM and SiteGA for sequence analysis of experimentally proven GCC-boxes. We carried out GCC box recognition and tested its distribution in the *Arabidopsis thaliana* L. genome. Functional annotation and microarray data analysis of the genes possessing predicted GCC-boxes elucidated their role in ethylene response.

**Key words:** ethylene, transcription factor, binding site, *Arabidopsis thaliana*.