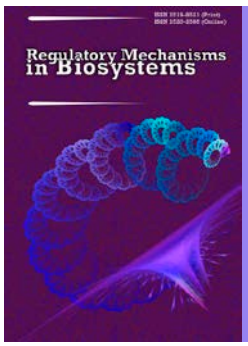

Regulatory Mechanisms in Biosystems

ISSN 2519-8521 (Print)
ISSN 2520-2588 (Online)
Regul. Mech. Biosyst., 8(3), 343–348
doi: 10.15421/021753

Grouping and clustering of maize Lancaster germplasm inbreds according to the results of SNP-analysis

K. V. Derkach*, T. M. Satarova**, V. V. Borysova*, V. Y. Cherchel*, B. V. Dzyubeckiy*

**Institute of Grain Crops of National Academy of Agrarian Science of Ukraine, Dnipro, Ukraine*
***Oles Honchar Dnipro National University, Dnipro, Ukraine*

Article info

Received 19.06.2017
Received in revised form 28.07.2017
Accepted 03.08.2017

Institute of Grain Crops of NAAS of Ukraine, V. Vernadsky Str., 14, Dnipro, 49067, Ukraine.

Oles Honchar Dnipro National University, Gagarin Ave., 72, Dnipro, 49010, Ukraine. Tel.: +38-056-236-26-18. E-mail: satarova2008@ukr.net, kvderkach@gmail.com

Derkach, K. V., Satarova, T. M., Borysova, V. V., Cherchel, V. Y., & Dzyubeckiy, B. V. (2017). Grouping and clustering of maize Lancaster germplasm inbreds according to the results of SNP-analysis. Regulatory Mechanisms in Biosystems, 8(3), 343–348. doi: 10.15421/021753

The objective of this article is the grouping and clustering of maize inbred lines based on the results of SNP-genotyping for the verification of a separate cluster of Lancaster germplasm inbred lines. As material for the study, we used 91 maize (*Zea mays* L.) inbred lines, including 31 Lancaster germplasm lines and 60 inbred lines of other germplasms (23 Iodent inbreds, 15 Reid inbreds, 7 Lacon inbreds, 12 Mix inbreds and 3 exotic inbreds). The majority of the given inbred lines are included in the Dnipro breeding programme. The SNP-genotyping of these inbred lines was conducted using BDI-III panel of 384 SNP-markers developed by BioDiagnostics, Inc. (USA) on the base of Illumina VeraCode Bead Plate. The SNP-markers of this panel are biallelic and are located on all 10 maize chromosomes. Their range of conductivity was >0.6. The SNP-analysis was made in completely automated regime on Illumina BeadStation equipment at BioDiagnostics, Inc. (USA). A principal component analysis was applied to group a general set of 91 inbreds according to allelic states of SNP-markers and to identify a cluster of Lancaster inbreds. The clustering and determining hierarchy in 31 Lancaster germplasm inbreds used quantitative cluster analysis. The share of monomorphic markers in the studied set of 91 inbred lines equaled 0.7%, and the share of dimorphic markers equaled 99.3%. Minor allele frequency (MAF) > 0.2 was observed for 80.6% of dimorphic markers, the average index of shift of gene diversity equaled 0.2984, PIC on average reached 0.3144. The index of gene diversity of markers varied from 0.1701 to 0.1901, pairwise genetic distances between inbred lines ranged from 0.0316–0.8000, the frequencies of major alleles of SNP-markers were within 0.5085–0.9821, and the frequencies of minor alleles were within 0.0179–0.4915. The average homozygosity of inbred lines was 98.8%. The principal component analysis of SNP-distances confirmed the isolation of the Lancaster group within the general set of analyzed inbred lines. Two-dimensional component analysis showed that the first principal component (PCA1) accounted for 36.0% of total variation and divided the investigated set of 91 inbred lines into two fractions, while all the inbred lines which are considered Lancaster based on pedigree information were included in one of the fractions. The second principal component (PCA2), which accounted for 12.1% of total variation, separated most of the Lancaster germplasm inbred lines from the others in this fraction, although the overlapping of the locations of Lancaster and non-Lancaster inbred lines was observed. Qualitative cluster analysis of 31 Lancaster germplasm inbred lines allowed to identify two clusters: the first one includes 23 inbred lines of Ukrainian selection and the well known Mo17 inbred line (77.4% of total number of analysed lines) inbred line, and the second cluster included 6 inbred lines of Ukrainian selection and the well known Oh43 inbred line (22.6% of total number of analysed lines) inbred line. The isolation of two clusters within Lancaster germplasm indicates the genetic diversity in this plasm. The evaluation of genome similarities through allelic states of SNP-markers can successfully be used for classification and systematization of the gene pool of maize genetic resources.

Keywords: *Zea mays*; markers of single nucleotide polymorphism of DNA; principal component analysis; cluster analysis

Групування та кластеризація ліній кукурудзи зародкової плазми Ланкастер за результатами SNP-аналізу

К. В. Деркач*, Т. М. Сатарова**, В. В. Борисова*, В. Ю. Черчель*, Б. В. Дзюбецький*

**Інститут зернових культур НААН України, Дніпро, Україна*
***Дніпровський національний університет імені Олеся Гончара, Дніпро, Україна*

Охарактеризовано репрезентативність маркерів однонуклеотидного поліморфізму ДНК панелі BDI-III з 384 SNP-маркерів для 91 лінії кукурудзи Дніпровської селекційної програми різних зародкових плазм. Проведено групування досліджуваного масиву ліній за алельним станом SNP-маркерів та ідентифікацію кластера плазми Ланкастер серед загальної добірки ліній шляхом проведення двовимірного

принципового компонентного аналізу. За результатами якісного кластерного аналізу методом повного зв'язку відмічено генетичне різноманіття всередині плазми Ланкастер за варіюванням алельного стану SNP-маркерів. Серед проаналізованих ліній зародкової плазми Ланкастер виділено два підкластери. Перший підкластер об'єднує 23 лінії української селекції та широковідому лінію Mo17 (77,4%), другий – 6 ліній української селекції та широковідому лінію Oh43 (22,6%).

Ключові слова: *Zea mays*; маркери однонуклеотидного поліморфізму ДНК; принциповий компонентний аналіз; кластерний аналіз

Вступ

Внутрішньовидова систематизація генетичних ресурсів культурних рослин – актуальна проблема, особливо для монотипних родів, таких як кукурудза (*Zea mays* L.). За рахунок багаторічної селекції усередині одного виду існують сотні, тисячі і навіть мільйони локальних популяцій, ліній, гібридів, сортів. Упорядкування цієї маси представників того чи іншого виду культурних рослин залежно від їх генетичної спорідненості та геномна характеристика – необхідні завдання молекулярної генетики та біотехнології.

Сучасна селекція кукурудзи будується на основі гетерозисних моделей і полягає у використанні лінійного матеріалу різних типів зародкової плазми. Типи зародкової плазми – окремі групи, які об'єднують значну кількість споріднених генотипів кукурудзи, тобто тих, які мають спільне походження, як правило, походять від одного вихідного сорту. Для України селекційно найперспективніша група ліній кукурудзи зародкової плазми Ланкастер. Вирощуються також лінії зародкових плазм Айодент, Лакон, Рейд. Лінії плазми Ланкастер відрізняються високою комбінаційною здатністю, інтенсивним стартовим ростом, середньою стійкістю до хвороб і, головне, значною посухо- та жаростійкістю, що актуально в умовах глобального потепління (Dzjubec'kij et al., 2012; Derkach et al., 2016).

Сучасні генотипи зародкової плазми Ланкастер походять від вільнозапиленого сорту Lancaster Sure Crop, який створила родина Hershey з округу Ланкастер, штат Пенсильванія (США). Родина Hershey у 1860–1910 рр. працювала з місцевим напівкременистим сортом, який характеризувався малими тонкими качанами лавандового (бузкового) кольору. Спочатку відбирали за середньою довжиною качана, добре дозрілим насінням, неураженістю пліснявою. Постійно відбирали качани з кременистими зернами, хоча схрещування проводили з пізніми зубоподібними сортами. Пізніше під час відбору почали звертати увагу на міцніші корені та більші качани. Відібраний із цього типу зародкової плазми сорт кукурудзи названо Sure Crop через його ранньостиглість і посухостійкість, що забезпечувало сталий впевнений врожай. Окремо селекцією вільнозапиленого сорту Lancaster займалася родина Richey біля Ла Саля, штат Іллінойс, із 1888 по 1920 рр., створивши сорт Richey Lancaster. Насіння сорту Lancaster Sure Crop потрапило до них через мігрантів із Міннесоти, які привезли його з Пенсильванії. Селекцію вели на більшу довжину качана, більшу масу насіння, вищу врожайність та кращу схожість насіння. Сучасні лінії плазми Ланкастер переважно походять від двох ліній першого циклу самозапилення сорту Ланкастер – Oh40B (від сорту Richey Lancaster) та C103 (від сорту Lancaster Sure Crop), хоча практичного значення набули лінії другого циклу самозапилення Mo17 (від лінії C103) та Oh43 (від лінії Oh40B) (Bennetzen and Hake, 2009).

Історично склалося, що приналежність ліній кукурудзи до певного типу зародкової плазми, тобто внутрішньовидова класифікація, яка покликана бути філогенетичною, ґрунтується на даних педігрі (родоводів), тобто взаємовідносини ліній оцінюються залежно від їх походження. Така класифікація недосконала, бо родоводи мають такі недоліки як суб'єктивність, пропуски в записах, неможливість відновлення тих їх частин, які належать до XIX – першої половини XX сторіччя, міксування неспоріднених генотипів у генотипі нащадка. Внутрішньовидова класифікація кукурудзи, основана на оцінюванні, наприклад, комбінаційної здатності (поділ зразків не на зародкові плазми, а на гетерозисні групи), також не досконала, оскільки прояв ознаки комбінаційної здатності залежить не тільки від генотипу, а і від факторів довкілля. Поділ генофонду культурної рослини по групах адаптації до мінливих умов довкілля зустрічається з тим, що генотипи однієї групи адаптації не обов'язково гомологічні за походженням (Dzjubec'kij et al., 2012). Разом із цим, використання різних типів молекулярно-генетичних маркерів, зокрема, маркерів однонуклеотидного поліморфізму ДНК (single nucleotide polymorphism), тобто SNP-маркерів, розглядається як потенційно ефективний інструмент внутрішньовидової класифікації та систематизації генофонду культурних рослин, що враховують спорідненість і варіювання на рівні геномів, на відміну від методів, які базуються на порівнянні за фенотипічними ознаками (Elshire et al., 2013; Wu et al., 2016; Zhang et al., 2016; Mikić et al., 2017).

У зв'язку із цим, мета нашого дослідження – групування та кластеризація ліній кукурудзи за результатами SNP-генотипування для верифікації окремого кластера ліній зародкової плазми Ланкастер.

Матеріал і методи досліджень

Матеріалом дослідження виступали 91 лінія кукурудзи (*Zea mays* L.), зокрема, 31 лінія зародкової плазми Ланкастер і 60 ліній інших зародкових плазм (23 лінії плазми Айодент, 15 ліній плазми Рейд, 7 ліній плазми Лакон, 12 ліній плазми Мікс та 3 лінії екзотичної плазми). Переважна більшість ліній входить до Дніпровської селекційної програми.

SNP-генотипування цих ліній проводили за панеллю BDI-III з 384 SNP-маркерів, розробленою фірмою BioDiagnostics, Inc. (США) на основі Illumina VeraCode Bead Plate. SNP-маркери панелі BDI-III біалельні, розташовані на всіх 10 хромосомах кукурудзи, мають ранг конструктивності > 0,6 (Venkatramana et al., 2012). SNP-аналіз виконано у повністю автоматизованому режимі на обладнанні Illumina BeadStation на базі фірми BioDiagnostics, Inc. (США).

Оцінювання репрезентативності SNP-маркерів для набору з 91 лінії кукурудзи проводили за Lu et al. (2009) за такими показниками: частота пропущених даних, частота мономорфних та диморфних маркерів, частота мажорних та мінорних алелів, показник зсуву генного різноманіття маркера, показник генного різноманіття лінії, індекс інформативності, гомозиготність і гетерозиготність зразків.

Частоту пропущених даних за маркером *i* (%, missing data points) розраховували як процентне відношення кількості зразків, за якими не вдалося визначити алельний стан маркера *i*, до загальної кількості зразків, для яких проводили визначення алельного стану маркера *i*. Частоту мономорфних маркерів у даному наборі зразків (%) розраховували як процентне відношення кількості біалельних маркерів, які в даному наборі зразків виявили мономорфний стан, до загальної кількості біалельних маркерів, за якими проаналізовано дану сукупність зразків. Частоту диморфних маркерів у даному наборі зразків (%) розраховували як відношення кількості біалельних маркерів, які в даному наборі селекційних зразків виявили диморфний стан, до загальної кількості біалельних маркерів, за якими проаналізовано набір зразків. Частоту мажорного алеля за маркером *i* в даному наборі зразків розраховували як відношення кількості зразків, в яких виявлено мажорний алельний стан за маркером *i*, до загальної кількості зразків, проаналізованих за маркером *i*. Даний показник визначається в частках одиниці і завжди більший за 0,5. Частоту мінорного алеля за маркером *i* в даному наборі зразків (minor allele frequency, MAF) розраховували як відношення кількості зразків, у яких виявлено мінорний алельний стан за маркером *i*, до загальної кількості зразків,

проаналізованих за маркером *i*. Даний показник визначається в частках одиниці і завжди менший за 0,5.

Показник зсуву генного різноманіття для маркера *i* у даному наборі зразків розраховували за формулою:

$$1 - [(p_i^+)^2 + (p_i^-)^2],$$

де p_i^+ та p_i^- – частоти альтернативного стану маркера *i* для набору зразків, проаналізованих за даним маркером. Показник визначали в частках одиниці (в межах 0,0–0,5).

Показник генного різноманіття зразка *k* (gene diversity) визначали за формулою:

$$(n_A \bullet n_A - 1 + n_T \bullet n_T - 1 + n_Г \bullet n_Г - 1 + n_Ц \bullet n_Ц - 1)/2N \bullet (N - 1),$$

де n_A, n_T, $n_Г$ і $n_Ц$ – кількість маркерів, які в маркерному сайті містять, відповідно, аденін, тимін, гуанін та цитозин, N – загальна кількість маркерів, за якими досліджено зразок *k*. Показник розраховували в частках одиниці. Його значення можуть коливатися від 0 до 0,5 і збільшуються у разі посилення генного різноманіття лінії.

Індекс інформативності маркера *i* (polymorphism information content value, PIC) розраховували за формулою:

$$1-[(p_i^+)^2 + (p_i^-)^2] - 1 \bullet 2 \bullet (p_i^+)^2 \bullet (p_i^-)^2,$$

де p_i^+ та p_i^- – частоти альтернативного стану маркера *i* для зразків, проаналізованих за даним маркером. PIC визначається в частках одиниці і змінюється від 0 до 0,375.

Гомозиготність зразка *k* (%) розраховували як процентне відношення кількості маркерів, що виявили гомозиготний алельний стан у зразка *k*, до загальної кількості маркерів, за якими проаналізовано зразок *k*. Гетерозиготність зразка *k* (%) розраховували як процентне відношення кількості маркерів, що виявили гетерозиготний алельний стан у зразка *k*, до загальної кількості маркерів, за якими проаналізовано зразок.

Завдання групування досліджуваного масиву ліній за алельним станом SNP-маркерів та ідентифікації кластера ліній плазми Ланкастер серед загальної сукупності ліній виконували шляхом проведення принципового компонентного аналізу. Цей аналіз (PCA, метод головних компонент) дозволяє провести зменшення кількості об'єктів аналізу за рахунок нових основних змінних величин, досягти скорочення розмірності опису, здійснити візуалізацію даних та виділити значиму інформацію. Основи принципового компонентного аналізу розроблено в працях Abdi and Williams (2010).

Для встановлення ієрархії ліній у проаналізованому масиві проведено якісний кластерний аналіз 31 лінії кукурудзи плазми Ланкастер. Ми застосували один з ієрархічних агломеративних методів кластерного аналізу – метод повного зв'язку (complete-linkage clustering, далекі сусіди) (Brereton, 2003; Sivolap et al., 2011).

Дані в таблиці наведено у вигляді X ± m, де X – середнє значення показника, m – довірчий інтервал (P = 0,05).

Результати

Аналіз алельного стану SNP-маркерів у наборі з 91 лінії кукурудзи дозволив визначити тип нуклеотиду в маркерних сайтах у 97,8% ліній. Частка мономорфних маркерів у досліджуваному наборі ліній склала 0,7%, диморфних – 99,3%. Частотою мінорного алеля (MAF) > 0,2 характеризувалися 80,6% диморфних маркерів. Середній показник зсуву генного різноманіття маркера *i* дорівнював 0,2984, а PIC використаних маркерів у даному наборі ліній у середньому сягав 0,3144 за потенційно можливого діапазону значень 0–0,3750. Наведені показники свідчать про виконання вимог, які висуваються до інформативних SNP-маркерів, і правомочність використання інформації, отриманої під час SNP-аналізу.

Основні характеристики набору з 91 лінії кукурудзи, визначені за результатами SNP-генотипування за панеллю BDI-III, такі: середня гомозиготність досліджених ліній − 98,8%, частоти мажорних алелів SNP-маркерів коливалися в межах 0,5085–0,9821, а мінорних алелів – 0,0179–0,4915. Показник генного різноманіття ліній перебував у межах 0,1701–0,1901 (за потенційно можливого розмаху від 0 до 0,5, показник збільшується у разі посилення генного різноманіття лінії). Значення PIC використаних маркерів варіювало в максимально можливих межах. Попарні генетичні дистанції між дослідженими лініями перебували в діапазоні 0,0316–0,8000.

Результати принципового компонентного аналізу генетичних SNP-дистанцій між усіма лініями проаналізованого набору наведено на рисунку 1. Двовимірний компонентний аналіз показав, що перший принциповий компонент (PCA1) пояснює 36,0% загального варіювання та поділяє досліджуваний масив з 91 лінії на дві фракції: А і Б. Уздовж осі OX фракція А візуалізується в інтервалі –1,0 – –0,2, а фракція Б – в інтервалі –0,2 – +0,8. При цьому всі лінії, які за педігрі вважаються Ланкастер (білі кружечки на рис. 1), потрапляють у фракцію Б, розташовану на верхньому (правому) кінці осі OX.

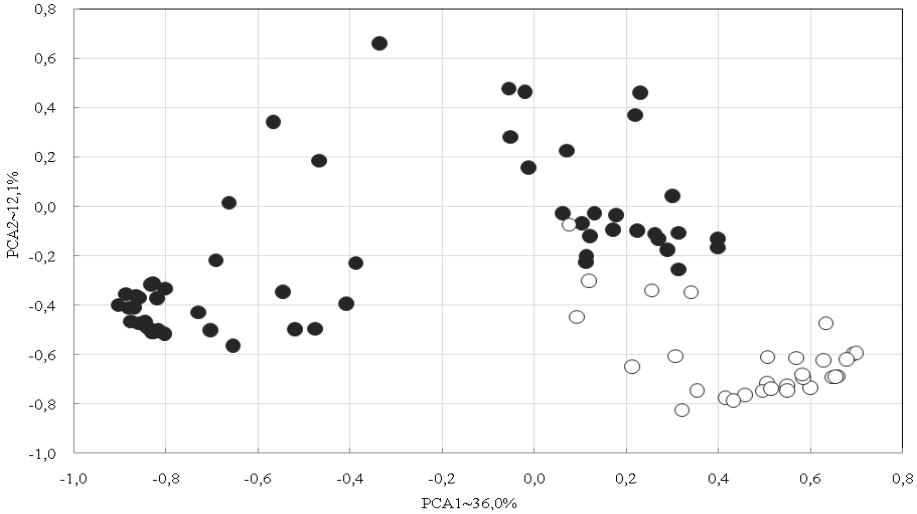

Рис. 1. Двовимірна діаграма принципового компонентного аналізу (PCA) даних SNP-генотипування 91 лінії кукурудзи: окремі лінії представлені кружками; білим кольором вказано лінії, які за педігрі належать до плазми Ланкастер (L), а чорним – лінії решти зародкових плазм (NL); масив досліджуваних ліній за першим принциповим компонентом PCA1 розпадається на дві фракції (А та Б)

Другий принциповий компонент (PCA2), який пояснює 12,1% загального варіювання, відокремлює більшість ліній Ланкастер (нижній кінець осі OY) від решти ліній фракції Б, хоча має місце деяке перекриття зон розташування L- та NL-ліній. Таким чином, принциповий компонентний аналіз попарних SNP-дистанцій підкреслює відокремленість групи ліній Ланкастер у загальному масиві проаналізованих ліній. Для з'ясування генетичних взаємовідносин ми провели кластеризацію

ліній плазми Ланкастер за розміром генетичних SNP-дистанцій. За Sivolap et al. (2011), «кластер» – група об'єктів, які мають спільні властивості, а характеристики «кластера» – внутрішня однорідність та зовнішня ізольованість. Кластерна модель ієрархізації об'єктів передбачає, що об'єкти з подібними властивостями належать до одного кластера. Найчастіше вона має вигляд дендрограми. На рисунку 2 показано дендрограму генетичних взаємовідносин ліній кукурудзи плазми Ланкастер за результатами SNP-аналізу, побудовану методом повного зв'язку.

На дендрограмі (рис. 2) виділяються два кластери, перший з яких об'єднує 23 лінії української селекції плазми Ланкастер та відому лінію Mo17 (77,4% загальної кількості проаналізованих ліній), а другий – 6 ліній української селекції плазми Ланкастер та відому лінію Oh43 (22,6%). Виділення двох кластерів вказує на різноманіття генетичного матеріалу усередині даної плазми. Дві з трьох фенотипічно подібних ліній ДК267, ДК212 та ДК6080, відібрані з вихідної популяції за участю лінії Oh43, під час кластеризації за SNP-даними увійшли до кластера, спільного з Oh43 (ДК267, ДК212), а лінія ДК6080 потрапила до кластера з Mo17. Цей цікавий факт може свідчити про те, що серед предків вихідних генотипів лінії ДК6080 окрім Oh43 могла бути присутньою також лінія Mo17, і комплекс саме її алельних варіантів маркерів перейшов у спадковість до ДК6080. Лінії ДК633266 та ДК298, у родоводі яких присутні лінії Oh43 та Mo17, входять до одного кластера з лінією Mo17, тобто саме від цієї лінії вони отримали більшу частку специфічних алельних варіантів маркерів, ніж від лінії Oh43.

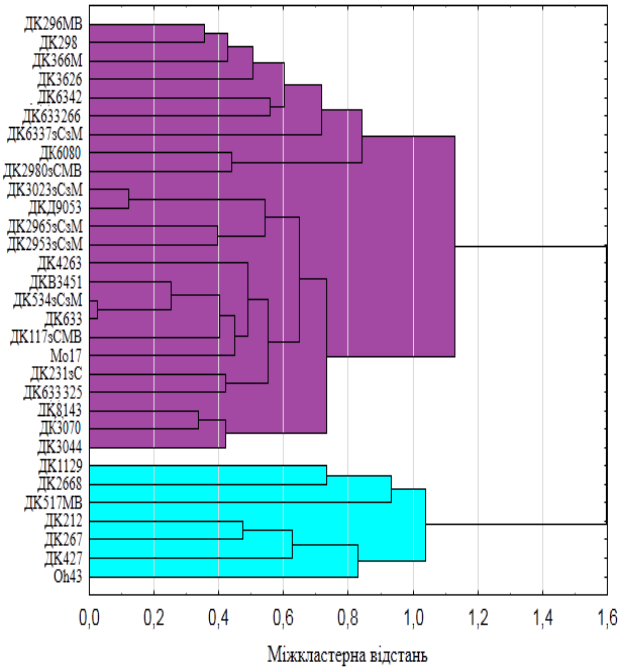

Рис. 2. Дендрограма генетичних взаємовідносин ліній кукурудзи плазми Ланкастер за результатами SNP-аналізу, побудована методом повного зв'язку

У таблиці наведено оцінки варіювання основних показників, що характеризують однонуклеотидний поліморфізм ліній плазми Ланкастер порівняно з аналогічною інформацією для групи неланкастерівських ліній.

Загалом частоти мажорних алелів SNP-маркерів коливалися для L-ліній в межах 0,5238–0,9688, а для NL-ліній – у ширших межах 0,5085–0,9821 (табл.). Середня частота мажорного алеля в групі L-ліній достовірно перевищувала аналогічний показник у NL-ліній і була ближчою до медіани потенційно можливого інтервалу значень частот мажорних алелів (0,7500), тоді як у NL-ліній цей показник зсунутий у бік менших частот. За коефіцієнтом варіації цієї ознаки достовірних відмінностей між двома групами ліній не зафіксовано.

Показник генного різноманіття у ліній плазми Ланкастер коливався в межах 0,1701–0,1873, а у NL-ліній – у межах 0,1725–0,1901 (за потенційно можливого розмаху від 0 до 0,5, показник збільшується за посилення генного різноманіття лінії). Тобто цей показник в групах L- та NL-ліній коливався в близькому діапазоні, відповідно 0,0172 та 0,0176, але його значення в групі L-ліній достовірно менші (в середньому на рівні 0,1774), порівняно з групою NL-ліній (в середньому 0,1808). Ті самі закономірності простежено відносно середнього значення показника зсуву генного різноманіття. Коефіцієнти варіації двох груп для показника генного різноманіття були низькими та достовірно не відрізнялися. Для показника зсуву генного різноманіття коефіцієнт варіації групи L-ліній суттєво (в 1,72 раза) перевищував аналогічний показник групи NL-ліній.

Таблиця

Варіювання основних показників однонуклеотидного поліморфізму ДНК у ліній кукурудзи

Характеристика		Група ліній	
		L	NL
Кількість досліджених ліній, шт.		32	59
Частота мажорного алеля, частка одиниці	$Lim_{potential}$	0,5000–1,000	
	Lim	0,5238–0,9688	0,5085–0,9821
	X ± m	0,7678 ± 0,0159	0,6977 ± 0,0143
	V, %	17,0 ± 4,3	16,8 ± 3,1
Показник генного різноманіття ліній, частка одиниці	$Lim_{potential}$	0–0,5000	
	Lim	0,1701–0,1873	0,1725–0,1901
	X ± m	0,1774 ± 0,0015	0,1808 ± 0,0009
	V, %	2,4 ± 0,6	2,1 ± 0,4
Показник зсуву генного різноманіття, частка одиниці	$Lim_{potential}$	0–0,5000	
	Lim	0–0,5000	0–0,5000
	X ± m	0,3045 ± 0,0180	0,3866 ± 0,0133
	V, %	51,0 ± 12,8	29,6 ± 5,5
PIC, частка одиниці	$Lim_{potential}$	0–0,3750	
	Lim	0–0,3750	0–0,3750
	X ± m	0,2462 ± 0,0132	0,3053 ± 0,0091
	V, %	46,4 ± 11,6	25,6 ± 4,7
Попарні генетичні дистанції, частка одиниці	$Lim_{potential}$	0–1,0000	
	Lim	0,0035–0,5333	0,0316–0,8000
	X ± m	0,3377 ± 0,0099	0,4229 ± 0,0061
	V, %	33,5 ± 8,4	30,5 ± 5,6

Значення PIC використаних маркерів в обох групах ліній варіювало в максимально можливих межах. Вищі значення PIC виявлено у більш поліморфній групі NL-ліній. Коефіцієнт варіації ознаки у групі L-ліній також суттєво (в 1,81 раза), перевищив цей показник для NL-ліній.

У групі ліній неланкастерівських плазм діапазон генетичних дистанцій був більшим (0,7684), ніж у L-ліній (0,5298), ліміти ширші (0,0316–0,8000 проти 0,0035–0,5333), а середнє значення достовірно вищим (0,4229 ± 0,0061 проти 0,3377 ± 0,0099). Максимальна генетична дистанція у групі NL-ліній зафіксована між лініями B73 та ДКД2725СВзМ, а у групі L-ліній – між Oh43 та ДК8143. Коефіцієнти варіації даної ознаки у двох груп достовірно не розрізнялися. Генетична дистанція між двома групами, L-ліній та NL-ліній, склала 0,4742.

Порівняння варіювання за основними показниками однонуклеотидного поліморфізму ДНК ліній кукурудзи очікувано свідчить про більше генетичне різноманіття групи ліній, де зібрані представники декількох неланкастерівських плазм, але виявляє також певний рівень різноманіття групи ліній Ланкастер.

Оскільки сучасні лінії української селекції створені на базі популяцій різної складності із залученням у тому числі компонентів, які не належать до груп Mo17 та Oh43, оцінювання їх подібності між собою за SNP-маркерами дозволить обрати доцільну стратегію для вибору альтернативних генотипів із метою їх використання у синтезі високогетерозисних гібридів.

Таким чином, оцінювання подібності геномів за визначеними, константними сайтами чи локусами може слугувати інформативним джерелом для класифікації та систематизації генофонду зразків культурних рослин.

Обговорення

У праці Semagn et al. (2012) з 450 інбредних ліній, досліджених за 1065 SNP-маркерами, і праці Wen et al. (2012), які досліджували 350 генотипів кукурудзи раси Tuxpeño, а також значну кількість ліній із різноманітних селекційних програм США за 1536 SNP-маркерами, частота пропущених даних склала 10%. Xu et al. (2017) за спеціально підібраною панеллю з 55 000 SNP-маркерів проаналізували 593 лінії кукурудзи помірного та тропічного поясів і повідомляють про частоту пропущених даних на рівні 0,3–5,2% (у середньому 1,83%). Romay et al. (2013) показали, що кількість пропущених даних можна зменшити під час повторного сиквенування, а також висловлюють припущення про залежність кількості пропущених даних від генотипу лінії, що сиквенується. Так, генотипи, близькі до референсного геному, яким є геном лінії B73, дають найменшу кількість пропущених даних (до 20%), інші – до 30%, а лінія SA24 після 25-разового генотипування мала лише 16% пропущених даних. У нашому дослідженні частотою пропущених даних >20% характеризувалися лише шість маркерів. Вони виключені з аналізу та подальших розрахунків, але інформацію про їх алельний стан використано для ліній, де тип нуклеотиду було встановлено.

Середня гетерозиготність проаналізованих зразків склала 1,2%, у середньому 0,0–0,4 гетерозиготних SNP-сайтів на одну лінію, що свідчить про високу (на рівні 98,8%) гомозиготність досліджених ліній. У нашому дослідженні гетерозиготні SNP-локуси виключали з аналізів і розрахунків. Wen et al. (2012) показали, що гетерозиготність за SNP-маркерами коливається від 1% у ліній CIMMYT до 36% у місцевих американських популяцій.

У праці Semagn et al. (2012) за даними SNP-генотипування спорідненість на рівні 5–50% мали 79% ліній, а 94% попарних генетичних дистанцій вкладалися в діапазон 0,3–0,4, оскільки мали вузьку генетичну базу з CIMMYT-програм для Східної та Південної Африки. Більшість ліній автори вважають унікальними для цієї селекційної програми. Xu et al. (2017) показав, що у кукурудзи помірного та тропічного поясів спорідненість 59% досліджених ліній перебувала на рівні 0–0,5%, а 25% ліній – на рівні 5–20%, тобто даний набір складений зі значно віддалених між собою в генетичному відношенні, різноманітних і навіть унікальних зразків. У нашому наборі досліджених ліній, переважна більшість яких входить до Дніпровської селекційної програми, визначена за результатами SNP-генотипування спорідненість ліній у діапазоні 0–5% не зареєстрована, а найбільша кількість пар ліній (65,2%) має значення спорідненості в інтервалі 40–60%.

Показник генного різноманіття ліній Дніпровської селекційної програми перебував у діапазоні 0,1701–0,1901. Це значно менше, ніж для популяцій CIMMYT (0,2669) і ліній із програми US-GEM для розширення зародкової плазми США (0,3891), які вміщують значні колекції, у тому числі тропічних, азійських, південноамериканських генотипів (Wen et al., 2012). Низькі значення показника генного різноманіття ліній та його вузький діапазон ми пов'язуємо зі специфічним набором ліній, що аналізували (ліній Дніпровської селекційної програми), генезис яких проходив у специфічних і суворих відносно посухи та спеки умовах степової зони України.

Таким чином, загальні характеристики генотипування, проведеного нами за панеллю з 384 SNP-маркерів для 91 лінії кукурудзи, більшість яких створена за Дніпровською селекційною програмою, відповідають установленим вимогам і перебувають на рівні, отриманому іншими авторами під час генотипування шляхом часткового сиквенування геномів ліній кукурудзи різного походження за SNP-панелями.

Wu et al. (2016), вивчивши 544 лінії за 362008 SNP-маркерами за допомогою принципового компонентного аналізу, встановили, що перші дві принципові компоненти пояснюють 32,8% (19,1% та 13,7%) загального варіювання для всього масиву ліній кукурудзи, 31,0% (17,6% та 13,4%) – для групи низинних тропічних ліній, 30,1% (18,1% та 12,0%) для субтропічної групи ліній середньої висоти та 30,8% (18,2% та 12,6%) – для тропічної гірської групи ліній. У нашому дослідженні перша та друга принципові компоненти пояснюють разом 48,1% загального варіювання, що на 15,3% вище, ніж у праці (Wu et al., 2016). Semang et al. (2012) визначили 237 алелів за 236 SNP-маркерами, які найкраще відділяли три групи ліній між собою. Перша та друга принципові компоненти під час аналізу за цими алелями пояснювали 99,8% (93,6% та 6,2%) загального варіювання масиву з 450 ліній кукурудзи. Під час дослідження розподілу 770 помірних і тропічних / субтропічних ліній за 1034 SNP-маркерами встановлено, що перша та друга принципові компоненти пояснюють лише 6,1% та 2,8% загального варіювання (Lu, 2009). Метод принципового компонентного аналізу дозволив ефективно розділити 346 досліджених ліній на чотири підгрупи: дві гетерозисні групи (Iowa Stiff Stalk Synthetic та Non-Stiff Stalk), група тропічної / субтропічної кукурудзи та змішана група (Li, 2017).

Mikić et al. (2017), проаналізувавши 96 ліній кукурудзи помірного поясу Південно-Східної Європи за 5 SSR-маркерами, виділили 6 кластерів, серед яких кластери, що містили лінії BSSS, Ланкастер і Айодент. Автори довели значне різноманіття ліній плазми Ланкастер і перспективність її використання в селекційному процесі, а також для асоціативного картування цільових ознак. Smith et al. (2015) на основі генотипування 380 ліній Техаської селекційної програми за 766 SNP-маркерами здійснили їх поділ на групи, які відповідали групам за педігрі BSSS, NSS, Айодент та тропічній групі. Разом із цим, Semagn et al. (2012) не підтверджують зв'язок між результатами генотипування за 1065 SNP-маркерами 450 ліній кукурудзи та поділ на гетерозисні групи за комбінаційною здатністю, оціненою як у діалельних, так і в тестерних схрещуваннях. Автори пояснюють це впливом генотипу батьківських компонентів, а також тестера на фенотипічний прояв комбінаційної здатності і її не збіжність з результатами геномного аналізу. Під час аналізу генетичної структури 367 ліній кукурудзи, поширених у Китаї (Wu, 2014), за 56110 SNP-маркерами виділено дві великі групи: до першої з них увійшли лінії місцевої зародкової плазми, до другої – інтродукованої плазми, а також п'ять підгруп, що відповідали різним гетерозисним групам (Reid Yellow Dent, Lancaster Sure Crop, P-группа, Tang Sipingtou та Tem-tropical група, всередині яких також спостерігали генетичну гетерогенність, причому найменша гетерогенність відзначена всередині P-групи, а найбільша – всередині Tem-tropical групи). Генетична різнорідність відмічена також серед 59 ліній INERA, проаналізованих за 1057 SNP-маркерами (Dao, 2014), і серед 156 ліній Північно-Каролінського Державного університету (Nelson, 2015), і у 284 ліній селекції Університету Міннесоти (Schaefer, 2013).

Висновки

Загальні характеристики генотипування, проведеного за панеллю з 384 маркерів однонуклеотидного поліморфізму ДНК для 91 лінії кукурудзи Дніпровської селекційної програми, відповідають установленим вимогам і перебувають на рівні, отриманому іншими авторами у процесі SNP-аналізу ліній кукурудзи різного походження. Результати принципового компонентного аналізу підтверджують наявність серед проаналізованого масиву окремої групи ліній, які за педігрі належать до зародкової плазми Ланкастер. Результати якісного кластерного аналізу методом повного зв'язку свідчать про варіювання всередині плазми Ланкастер, де виділено лінії, близькі як до Mo17, так і до Oh43.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews, Computational Statistics.

Bennetzen, J. L., & Hake, E. S. A. (2009). Handbook of maize. Genetic and genomics. New York. Springer Science.
Brereton, R. G. (2003). Chemometrics: Data analysis for the laboratory and chemical plant. Wiley, Chichester.
Dao, A., Sanou, J., Mitchell, S. E., Gracen, V., & Danquah, E. Y. (2014). Genetic diversity among INERA maize inbred lines with single nucleotide polymorphism (SNP) markers and their relationship with CIMMYT, IITA, and temperate lines. BMC Genetics, 15, 127–140.
Derkach, K. V., Abraimova, O. E., & Satarova, T. M. (2016). Reguljacija morfogenezu *in vitro* u linij kukurudzi grupi Lankaster [Regulation of *in vitro* morphogenesis in maize inbreds of the Lancaster group]. Visnyk of Dnipropetrovsk Universitet. Biology, Ecology, 24(2), 253–257 (in Ukrainian).
Dzjubec'kij, B. V., Bodenko, N. A., Fed'ko, M. M., & Gusak, J. V. (2012). Stvorennja seredn'opiznih gibridiv kukurudzi na bazi plazmi Lankaster (C_{103}) [Creation of medium-late hybrids of corn based on Lancaster germplasm (C_{103})]. Bjuleten' Institutu Sil's'kogo Gospodarstva Stepovoji Zony NAAN Ukrayiny, 3, 8–11 (in Ukrainian).
Elshire, R. J., Acharya, C. B., Mitchell, S. E., Flint-Garcia, S. A., McMullen, M. D., Holland, J. B., Buckler, E. S., & Gardner, C. A. (2013). Comprehensive genotyping of the USA national maize inbreds seed bank. Genome Biology, 14(6), R55.
Li, X., Jian, Y., Xie, C., Wu, J., Xu, Y., & Zou, C. (2017). Fast diffusion of domesticated maize to temperate zones. Scientific Reports, 7, 2077–2089.
Lu, Y., Yan, J., Guimarães, C. T., Taba, S., Hao, Z., Gao, S., Chen, S., Li, J., Zhang, S., Vivek, B. S., Magorokosho, C., Mugo, S., Makumbi, D., Parentoni, S. N., Shah, T., Rong, T., Crouch, J. H., & Xu, Y. (2009). Molecular characterization of global maize breeding germplasm based in genome-wide single nucleotide polymorphisms. Theoretical and Applied Genetics, 120(1), 93–115.
Mikić, S., Kondić-špika, A., Brbaklić, L., Stanisavljević, D., Ćeran, M., Trkulja, D., & Mitrović, B. (2017). Molecular and phenotypic characterisation of diverse temperate maize inbred lines in Southeast Europe. Zemdirbyste-Agriculture, 104(1), 31–40.
Nelson, P. T., Krakowsky, M. D., Coles, N. D., Holland, J. B., Bubeck, D. M., Smith, J. S. C., & Goodman, M. M. (2016). Genetic characterization of the North Carolina State University Maize Lines. Crop Science, 56, 259–275.
Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., Elshire, R. J., Acharya, C. B., Mitchell, S. E., Flint-Garcia, S. A., McMullen, M. D., Holland, J. B., Buckler, E. S., & Gardner, C. A. (2013). Comprehensive genotyping of the USA national maize inbreds seed bank. Genome Biology, 14(6), R55.
Schaefer, C. M., & Bernardo, R. (2013). Population structure and single nucleotide polymorphism diversity of historical Minnesota maize inbreds. Crop Science, 53(4), 1529–1536.
Semagn, K., Magorokosho, C., Vivek, B. S., Makumbi, D., Beyene, Y., Mugo, S., Prasanna, B. M., & Warburton, M. L. (2012). Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers. BMC Genomics, 13, 113–124.
Sivolap, J. M., Kozhuhova, N. J., & Kalendar', R. N. (2011). Variabel'nost' i specifichnost' genomov sel'skohozjajstvennyh rastenij [Variability and specificity of genomes of agricultural plants]. Astroprint, Odessa (in Russian).
Smith, S. D., Murray, S. C., & Heffner, E. (2015). Molecular analysis of genetic diversity in a Texas maize (*Zea mays* L.) breeding program. Maydica, 60, 1–8.
Venkatramana, P., Carlson, C., Blackstad, M., Bialozynski, R., Schultz, Q., & Kaufman, B. (2010). Development and characterization of single nucleotide polymorphism (SNP) panel for marker assisted backcrossing in corn. Seed Technology, 32(2), 153–154.
Wen, W., Franco J., Chavez-Tovar, V. H., Yan, J., & Taba, S. (2012). Genetic characterization of a core set of a tropical maize race Tuxpeño for further use in maize improvement. PLoS One, 7(3), e32626.
Wu, X., Li, Y., Shi, Y., Song, Y., Wang, T., Huang, Y., & Li, Y. (2014). Fine genetic characterization of elite maize germplasm using high-throughput SNP genotyping. Theoretical and Applied Genetics, 127, 621–631.
Wu, Y., San Vicente, F., Huang, K., Dhliwayo, T., Costich, D. E., Semagn, K., & Babu, R. (2016). Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. Theoretical and Applied Genetics, 129, 753–765.
Xu, C., Ren, Y., Jian, Y., Guo, Z., Zhang, Y., Xie, C., Fu, J., Wang, H., Wang, G., Xu, Y., Li, P., & Zou, C. (2017). Development of a maize 55 K SNP array with improved genome coverage for molecular breeding. Molecular Breeding, 37(3), 20–34.
Zhang, X., Zhang, H., Li, L., Lan, H., Ren, Z., Liu, D., Wu, L., Liu, H., Jaqueth, J., Li, B., Pan, G., & Gao, S. (2016). Characterizing the population structure and genetic diversity of maize breeding germplasm in Southwest China using genome-wide SNP markers. BMC Genomics, 17(1), 697–704.