ПРИМЕНЕНИЕ МЕТОДОВ МОЛЕКУЛЯРНОЙ ГЕНЕТИКИ ДЛЯ ОЦЕНКИ ГЕНЕТИЧЕСКОЙ ГЕТЕРОГЕННОСТИ ПОПУЛЯЦИЙ ОСНОВНЫХ ЛЕСООБРАЗУЮЩИХ ВИДОВ

УДК 575.113.1/575.133

ПРЕДВАРИТЕЛЬНЫЕ РЕЗУЛЬТАТЫ ПОЛНОГЕНОМНОГО *DE NOVO* СЕКВЕНИРОВАНИЯ ЛИСТВЕННИЦЫ СИБИРСКОЙ (*LARIX SIBIRICA* LEDEB.) И СОСНЫ КЕДРОВОЙ СИБИРСКОЙ (*PINUS SIBIRICA* DU TOUR)

© 2014 г. К. В. Крутовский[1, 2, 3, 4], Н. В. Орешкова[1, 5], Ю. А. Путинцева[1], А. А. Ибе[1, 6], К. О. Дейч[1, 6], Е. А. Шилкина[1, 6]

[1] *Сибирский федеральный университет*
Научно-образовательный центр геномных исследований
660036, Красноярск, Академгородок, 50а/2
[2] *Институт общей генетики им. Н. И. Вавилова РАН*
119991, ГСП-1, Москва, ул. Губкина, 3
[3] *Геттингенский университет*
Германия, 37077, Геттинген, ул. Бюсгенвег, 2
[4] *Техасский агро-механический университет*
США, Техас, 77843, Колледж Стейшн
[5] *Институт леса им. В. Н. Сукачева СО РАН*
660036, Красноярск, Академгородок, 50/28
[6] *Филиал Российского центра защиты леса*
Центр защиты леса Красноярского края
660036, Красноярск, Академгородок, 50а/2
E-mail: kkrutovsky@gmail.com, oreshkova@ksc.krasn.ru, yuliya-putintseva@rambler.ru, aaibis@mail.ru, kse-zhdanova@yandex.ru, helenbeauty74@mail.ru

Поступила в редакцию 20.06.2014 г.

В результате секвенирования ядерной ДНК сибирской лиственницы (*Larix sibirica* Ledeb.) и сосны кедровой сибирской (*Pinus sibirica* Du Tour) с помощью высокопроизводительного секвенатора Illumina HiSeq2000 и последующей первичной обработки данных для лиственницы получено 2 906 977 265 высококачественных парноконцевых нуклеотидных сиквенсов (чтений) генома и просеквенировано (прочитано) 576 млрд пар нуклеотидных оснований (п. н. о.), что соответствует 48-кратной длине генома лиственницы (48X покрытие), равного 12.03 млрд п. н. о., а для кедра получено 3 427 566 813 чтений генома (679 млрд п. н. о.), что соответствует 29-кратной длине (29X покрытие) генома кедра (23.6 млрд п. н. о.). Этих данных пока недостаточно для полной сборки и аннотации данных геномов, но полученные нуклеотидные сиквенсы уже позволили обнаружить и разработать эффективные высокополиморфные молекулярно-генетические маркеры, микросателлитные локусы, необходимые для популяционно-генетических исследований и идентификации происхождения древесины. Также продолжаются исследования однонуклеотидных полиморфизмов (так называемых «снипов»). Кроме того, геномные исследования российских бореальных лесов и связанных с ними фитопатогенов позволят обнаружить биомаркеры для решения важных научных и хозяйственных задач по сохранению лесных генетических ресурсов и селекции более устойчивых к заболеваниям и неблагоприятным факторам среды пород с ускоренным ростом и улучшенной древесиной.

Ключевые слова: *геном, de novo секвенирование, лиственница сибирская (Larix sibirica Ledeb.), сосна кедровая сибирская (Pinus sibirica Du Tour).*

ВВЕДЕНИЕ

Лиственница сибирская (*Larix sibirica* Ledeb.) и сосна кедровая сибирская (*Pinus sibirica* Du Tour) – важнейшие виды бореальных лесов Сибири, имеющие огромное экономическое, экологическое и эстетическое значение и играющие важнейшую биосферную роль в регуляции глобального климата. Однако изучение лиственницы, сосны и других важных пород хвойных, таких как сосна кедровая сибирская (кедр), тормозится практически полным отсутствием данных об их геноме и генах, контролирующих важные адаптивные и селекционные признаки.

Данные исследования позволят разработать высокоинформативные молекулярно-генетические маркеры, которые могут быть эффективно использованы для определения происхождения древесины, изучения и мониторинга генетической изменчивости хвойных лесов, их адаптации к изменению климата и для создания селекционных и природоохранных программ. Эти задачи заявлены как приоритетные для лесного хозяйства в Комплексной программе развития биотехнологий в Российской Федерации на период до 2020 г., утвержденной правительством Российской Федерации 24 апреля 2012 г.

Основной целью исследований является проведение полногеномного секвенирования, сборки, сравнительного анализа и аннотирования геномов основных лесообразующих пород Российской Федерации – лиственницы сибирской и кедра сибирского.

МАТЕРИАЛЫ И МЕТОДЫ

Исследовали референсные деревья *L. sibirica* (окр. с. Черное озеро, Республика Хакасия) и *P. sibirica* (Ермаковский р-н Красноярского края). Для выделения ДНК у лиственницы сибирской использовали хвою, гаплоидный каллус, у кедра – гаплоидный эндосперм (мегагаметофит) семени.

Этапы пробоподготовки для парноконцевых ДНК библиотек (paired-endlibrary) и непосредственного секвенирования проведены согласно требованию компании Illumina (www.bu.edu/iscf/files/2011/05/TruSeq_DNA_SamplePrep_Guide_15005180_A.pdf). Секвенирование образцов проводили на высокопроизводительном секвенаторе ДНК нового поколения HiSeq 2000 компании Illumina.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Первичная обработка результатов секвенирования. По результатам секвенирования проведена первичная биоинформатическая обработка данных. Всего получено 3 619 538 913 парноконцевых чтений (нуклеотидных последовательностей) генома лиственницы сибирской и 3 997 217 664 чтений генома кедра. Оценка качества полученных нуклеотидных последовательностей (сиквенсов) проводилась при помощи программы FASTQC (www.bioinformatics.babraham.ac.uk/projects/fastqc). Последовательности праймеров и адаптеров удалялись из сиквенсов с помощью программы Trimmomatic (www.usadellab.org/cms/?page=trimmomatic).

Для всех полученных данных использовались следующие параметры: среднее качество внутри скользящего окна шириной в 4 нуклеотида должно быть не ниже $Q = 20$, исключались из анализа сиквенсы, длина которых меньше 20. После удаления праймеров и адаптеров сиквенсы вновь проверялись при помощи программы FASTQC. В случае необходимости уточнялся список адаптеров и праймеров для обрезки и расчеты производились вновь.

В результате первичной обработки данных получено 2 906 977 265 высококачественных парноконцевых чтений генома лиственницы сибирской с $Q > 20$ (т. е. вероятность ошибки секвенирования не более 1 %) или 576 млрд п. н. о., что обеспечивает 48-кратное покрытие генома (12.03 млрд п. н. о.) и 3 427 566 813 чтений генома (679 млрд п. н. о.) кедра, что составляет 29-кратное покрытие соответствующего генома (23.6 млрд п. н. о.).

Сборка хлоропластного генома лиственницы сибирской. Из общего числа полученных нуклеотидных последовательностей произведена выборка сиквенсов хлоропластного генома на основе опубликованных в Genbank геномов хлоропластов близкородст-

Таблица 1. Предварительные результаты сборки геномов

Вид, геном	Количество контигов	Общая длина всей сборки	Длина наиболее длинного контига	N50	N90
		п.н.о.			
Larix sibirica:					
хлоропластный	3	125 489	64 254	64 254	57 603
ядерный	56 192	19 712 820	57 603	337	219
митохондриальный	142	126 882	3655	1081	505
Pinus sibirica					
Ядерный	7 737 815	2 830 700 997	81 100	363	231

венных видов *Larix occidentalis* (FJ899578.1, длина генома 119 680 п. н. о.), *Larix decidua* (NC_016058, длина генома 122 474 п. н. о.).

Отбор последовательностей, составляющих хлоропластный геном лиственницы сибирской, производили при помощи программы для картирования коротких сиквенсов на геномы средних размеров bowtie2 (bowtie-bio.sourceforge.net/bowtie2/index.shtml). Данная программа основана на алгоритме построения FM-индекса, основанном на преобразовании Барроуза-Уилера (Burrows, Wheeler, 1994).

В результате картирования на хлоропластные геномы отобраны высокогомологичные последовательности, очевидно относящиеся к хлоропластному геному лиственницы сибирской. Для сборки отобранных последовательностей использовали несколько программ-ассемблеров (SPAdes, Velvet, ABySS). Наиболее качественная сборка получена при помощи SPAdes (табл. 1).

Сборка митохондриального генома лиственницы сибирской. Для сборки митохондриального генома использована стратегия, аналогичная сборке хлоропластного генома. Однако поиск в Genbank показал, что на 01.09.2013 г. сведения о полном митохондриальном геноме близкородственных видов, таких, например, как лиственница западная или лиственница европейская, отсутствуют. Кроме того, на указанный момент в банке генетических данных не имелось сведений о полных митохондриальных геномах других представителей семейства Pinaceae.

В качестве матрицы для картирования митохондриальных сиквенсов было решено использовать еще официально не опубликованную рабочую сборку митохондриального генома *Pinus taeda* (сосна ладанная), доступную для загрузки по адресу loblolly.ucdavis.edu/bipod/ftp/Genome_Data/genome/pinerefseq/Pita/mito. Анализ картирования с помощью программы bowtie2 показал, что менее 0,01 % сиквенсов тотальной ДНК лиственницы сибирской картируются на митохондриальный геном ладанной сосны. Тем не менее была предпринята попытка использовать картировавшиеся сиквенсы для сборки митохондриального генома при помощи программы SPAdes (табл. 1).

Полученные сборки митохондриального генома лиственницы сибирской требуют дальнейшей проверки и подтверждения через получение дополнительных сиквенсов и улучшение сборки, поскольку в генетических базах данных в настоящее время содержится крайне мало нуклеотидных сиквенсов митохондриальных геномов не только близкородственных видов, но и вообще хвойных. Для увеличения числа митохондриальных сиквенсов мы планируем дальнейшее секвенирование митохондриального генома лиственницы, используя образцы, обогащенные митохондриальной ДНК.

Предварительные результаты сборки генома. Исходя из анализа литературы по данным полногеномного секвенирования и сравнения полученных результатов с опубликованными данными аналогичных проектов по ели обыкновенной (Nystedt et al., 2013) и ели белой (Birol et al., 2013) сделан вывод о том, что 48-кратное покрытие генома лиственницы сибирской и 29-кратное покрытие генома кедра недостаточны для сборки их полных геномов. Необходимость большего числа сиквенсов для полной сборки связана с высоким содержанием высоко-

Таблица 2. Микросателлитные локусы, обнаруженные в нуклеотидных сиквенсах и контигах генома лиственницы сибирской

Тип повтора	Количество локусов	Средняя длина локуса, п. н. о.	Встречаемость на 1 млн п. н. о.
Динуклеотидный	563	20.62	28.57
Тринуклеотидный	140	31.84	7.10
Тетрануклеотидный	7	53.86	0.36
Пентануклеотидный	3	60.00	0.15
Гексануклеотидный	10	59.90	0.51

повторяющейся (75–80 %) ДНК (Nystedt et al., 2013; Birol et al., 2013; Neale et al., 2014; Wegrzyn et al., 2014; Zimin et al., 2014) и огромными размерами геномов исследуемых видов: 12.03 млрд п. н. о. у лиственницы сибирской и 23.6 млрд п. н. о. у кедра, что почти в 4 и 7 раз соответственно больше генома человека (3.2 млрд п. н. о.).

Полученные предварительные данные (см. табл. 1) не являются полными, но их уже можно использовать для поиска и обнаружения микросателлитных (SSRs) локусов.

Идентификация микросателлитных (SSRs) и однонуклеотидных (SNPs) локусов. Для идентификации микросателлитных локусов использована программа SciRoKo (kofler.or.at/bioinformatics/SciRoKo/index.html) в режиме поиска идеальных повторов (PerfectRepeatsMode) в полученных предварительных сборках геномов лиственницы и кедра. В табл. 2 приведена общая статистика по найденным мотивам.

Всего в полученных контигах генома лиственницы сибирской найдено 723 микросателлитных локуса, соответствующих заданным параметрам.

ЗАКЛЮЧЕНИЕ

Для лиственницы сибирской и кедра сибирского получены данные высокого качества, обеспечивающие 48- и 29-кратное покрытие их геномов соответственно. Этих сведений еще недостаточно для полногеномной сборки, но уже удалось получить черновые сборки хлоропластных, митохондриальных и ядерных геномов для поиска микросателлитных локусов.

Работа выполнена при частичной финансовой поддержке гранта РФФИ № 14-04-01462а и гранта Правительства РФ «Геномные исследования основных бореальных лесообразующих хвойных видов и их наиболее опасных патогенов в Российской Федерации» (договор № 14.Y26.31.0004), выделенного в рамках конкурсной программы государственной поддержки научных исследований, проводимых под руководством ведущих ученых в российских образовательных учреждениях высшего профессионального образования, научных учреждениях государственных академий наук и государственных научных центрах Российской Федерации.

СПИСОК ЛИТЕРАТУРЫ

Birol I., Raymond A., Jackman S. D. et al. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data // Bioinformatics. 2013. V. 29. N. 12. P. 1492–1497.

Burrows M., Wheeler D. A block sorting lossless data compression algorithm // Technical Report 124, Digital Equipment Corporation, 1994.

Neale D. B., Wegrzyn J. L., Stevens K. A. et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies // Genome Biology. 2014. V. 15. N. 3. R59 DOI: 10.1186/gb-2014-15-3-r59

Nystedt B., Street N. R., Wetterbom A. et al. The Norway spruce genome sequence and conifer genome evolution // Nature. 2013. V. 497. N. 7451. P. 579–584.

Wegrzyn J. L., Liechty J. D., Stevens K. A. et al. Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation // Genetics. 2014. V. 196. N. 3. P. 891–909.

Zimin A., Stevens K. A., Crepeau M. W. et al. Sequencing and assembly of the 22-Gb loblolly pine genome // Genetics. 2014. V. 196. N. 3. P. 875–890.

Preliminary Results of *De Novo* Whole Genome Sequencing of the Siberian Larch (*Larix sibirica* Ledeb.) and the Siberian Stone Pine (*Pinus sibirica* Du Tour)

K. V. Krutovsky[1, 2, 3, 4], N. V. Oreshkova[1, 5], Yu. A. Putintseva[1], A. A. Ibe[1, 6], K. O. Deych[1, 6], E. A. Shilkina[1, 6]

[1] *Siberian Federal University*
Genome Research and Education Centre
Akademgorodok, 50a/2, Krasnoyarsk 660036 Russian Federation
[2] *N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences*
Gubkin str., 3, Moscow, 119333 Russian Federation
[3] *University of Göttingen*
Büsgenweg, 2, Göttingen, D-37077 Germany
[4] *Texas A&M University*
HFSB 305, 2138 TAMU, College Station, Texas, 77843 USA
[5] *V. N. Sukachev Institute of Forest, Russian Academy of Sciences, Siberian Branch*
Akademgorodok, 50/28, Krasnoyarsk, 660036 Russian Federation
[6] *Branch of the Russian Centre for Forest Protection*
Centre for Forest Protection of Krasnoyarsk Territory
Akademgorodok, 50a/2, Krasnoyarsk, 660036 Russian Federation
E-mail: kkrutovsky@gmail.com, oreshkova@ksc.krasn.ru, yuliya-putintseva@rambler.ru, aaibis@mail.ru, kse-zhdanova@yandex.ru, helenbeauty74@mail.ru

The Illumina HiSeq2000 DNA sequencing generated 2 906 977 265 high quality paired-end nucleotide sequences (reads) and 576 Gbp for Siberian larch (*Larix sibirica* Ledeb.) that corresponds to 48X coverage of the larch genome (12.03 Gbp), and 3 427 566 813 reads and 679 Gbp for Siberian stone pine (*Pinus sibirica* Du Tour) that corresponds to 29X coverage of its genome (23.6 Gbp). These data are not enough to assemble and annotate whole genomes, but the obtained nucleotide sequences have allowed us to discover and develop effective highly polymorphic molecular genetic markers, such as microsatellite loci that are required for population genetic studies and identification of the timber origin. Sequence data can be used also to discover single nucleotide polymorphisms (SNPs). Genomic studies of Russian boreal forests and major phytopathogens associated with them will also allow us to identify biomarkers that can be used for solving important scientific and economic problems related to the conservation of forest genetic resources and breeding more resilient and fast growing trees with improved timber and resistance to diseases and adverse environmental factors.

Keywords: *genome, de novo sequencing, Siberian larch (Larix sibirica Ledeb.), Siberian stone pine (Pinus sibirica Du Tour).*