Uluslararası Fen Araştırmalarında Yenilikçi Yaklaşımlar Dergisi
International Journal of Innovative Approaches in Science Research 2022, Vol. 6 (1), 12-22
https://doi.org/10.29329/ijiasr.2022.433.2

Original article

Bacillus Boroniphilus Genom Dizisinin Tamamlanması Çalışmaları ve Gap Bölgelerindeki Transpozaz Kodlayan Gen Dizilerinin Belirlenmesi

Genome Sequence Completion Studies of *Bacillus Boroniphilus* and Transposase Encoding Sequences in Gap Regions

Merve Sezer Kürkçü [a], Bekir Çöl [b, *], Hilal Özdağ [c], Yeşim Doğan [d] & Zeynep Özkeserli [c]

[a] Centers of Research Laboratory, Muğla Sıtkı Koçman University, Muğla, Turkey
[b] Department of Biology, Faculty of Science, Muğla Sıtkı Koçman University, Muğla, Turkey
[c] Department of Biotechnology Institute, Faculty of Science, Ankara University, Ankara, Turkey
[d] Department of Anthropology, Faculty of Science, Ankara University, Ankara, Turkey

Özet

Yeni nesil sekanslama (next-generation sequencing, NGS) teknolojisinin geliştirilmesi ile canlılardaki genomik bilginin açığa çıkartılması önemli derecede hız kazanmıştır. Bu gelişme ile genomları belirlenen organizmaların sayısındaki artış dikkat çekmektedir. Bu sekanslama yönteminde temel yaklaşım, tüm genomun kısa dizilere bölünmesi ve fragmentler halinde okunması ve ardından bunlar için geliştirilen yazılım programları ile bu fragmanların birleştirilmesidir (Contig Assembly). Ancak, bu kısa dizileri okuma yaparak sekanslama (short-read sequencing) platformlarının bazı kısıtlamalarının olduğu görülmüştür. Bu yeni sekanslanan genom dizilerinin çoğunun tamamlanmamış, genetik içeriği temsil eden taslaklar halinde kaldığı bilinmektedir.

Bu çalışmada, yüksek bor konsantrasyonu içeren ortamlarda yaşayabilen Bacillus boroniphilus bakterisinin genom belirlenmesi çalışmaları sırasında, belirlenemeyen sekans dizilerinin hangi DNA dizileri olduğunun açıklığa kavuşturulması amaçlanmıştır. Genom birleştirme işlemleri sırasında görülen gap (boşluk) bölgelerinin sekans dizilerinin belirlenebilmesi için yüzlerce primer dizayn edilerek PZR işlemleri sonucu elde edilen ürünlerin sekans analizlerinin sonuçlarının değerlendirilmesi ile neredeyse hepsinin transpozaz olduğu görülmüştür. Gap bölgeleri hakkında elde edilen bu bilgiler yeni nesil genom sekanslama çalışmalarındaki verimliliğin arttırılması için oldukça önemlidir.

Anahtar Kelimeler: Transpozaz, Yeni Nesil Genom Sekanslama, Gap (Boşluk), Bacillus Boroniphilus, Shotgun Sekanslama.

Abstract

The introduction and development of next-generation sequencing (NGS) has had a significant impact on revealing genomic information in living things. With this development, it attracts attention in the number of organisms whose genomes are identified in a wide range. The basis of this sequencing method is that the entire genome is read into fragments by dividing them into short sequences and then combined with software programs developed for them. However, short-read sequencing platforms have some

* **Corresponding author:**

Bekir Çöl, Department of Biology, Faculty of Science, Muğla Sıtkı Koçman University, Muğla, Turkey.
Email: bcol@mu.edu.tr

limitations. Most of these newly sequenced genome sequences were found to remain incomplete, in drafts representing genetic content.

In this study, it is aimed to clarify which DNA sequences are undetectable sequences during the genome determination studies of Bacillus boroniphilus bacteria which can live in environments with high boron concentration. In order to determine the sequence sequences of the gap regions seen during genome assembly processes, hundreds of primers were designed and the results of the sequence analyzes of the products obtained as a result of the PCR processes were found to be transposases. This information about gap regions is very important for increasing the efficiency of next generation genome sequencing studies.

Keywords: Transposase, Next Generation Sequencing, Gap, Bacillus Boroniphilus Shotgun Sequencing.

Received: 01 February 2022 * **Accepted:** 16 March 2022 * **DOI:** https://doi.org/10.29329/ijiasr.2022.433.2

GİRİŞ

Yeni Nesil Dizileme (NGS) teknolojileri olarak adlandırılan ikinci nesil dizileme platformlarının ortaya çıkışı, açık veritabanlarındaki genomik verilerin bilim dünyasına kazandırılmasında önemli derecede artış sağlamıştır (Chain vd., 2009).

Illumina HiSeq, IonTorrent PGM, Roche 454 FLX ve ABI SOLiD gibi NGS platformları, Sanger sekanslama tekniği (Sanger vd., 1977) ile karşılaştırıldığında, bu platformların daha yüksek bir sekanslama kapasitesine sahip oldukları ve bu avantajlarından dolayı önemli ölçüde daha fazla veri ürettikleri görülmüştür (Liu vd., 2012). Ancak bu sistemlerde karşılaşılan başlıca sorun, genomun kısa uzunluktaki okumalardan dolayı daha az tamamlanması ve parçalar halinde kalmasıdır (Liu vd., 2012).

Bakteriyel tüm genom sekanslama çalışmaları, yeni keşfedilen organizmaların genomlarının belirlenmesi, bilinen organizmaların genom dizilerinin tamamlanması ve farklı ya da benzer canlı türlerinin genom dizilerinin karşılaştırılması ile genetik düzeyde ne kadar ilişkili olduklarının anlaşılması için önemli bilgiler sunmaktadır. Bakterilerde genom belirleme çalışmaları incelendiğinde Shotgun sekanslama tekniğinin, genom diziliminde kullanılan yöntemler arasında öncül teknolojilerden biri olduğu görülmektedir.

Tüm genom shotgun sekanslama yöntemi ilk olarak 1979 yılında Staden tarafından öne sürülmüştür. Ardından bu yöntem ile ilk sekanslanan genom 1981'de yayınlanan karnıbahar mozaik virüsü olmuştur (Gardner vd., 1981). Teknik geliştirilirken, rastgele parçalanan DNA fragmentlerinin sıralı bir şekilde dizilebilmesi ve programların bunları birbirlerine ekleyerek dizilimi tamamlayabilmesi için DNA fragmentleri uzunluğu hakkında bir çok görüş öne sürülmüş ve kullanılmıştır (Edwards ve Caskey, 1991; Edwards vd., 1990; Roach vd., 1995).

Geliştirilen teknik, 1995 yılında The Institute for Genomic Research (TIGR) tarafından *Haemophilus influenzae* genomu (Fleischmann vd., 1995), 2000 yılında Celera Genomics tarafından ilk

olarak *Drosophila melanogaster* (meyve sineği) ve ardından insan genomunu dizilemek için kullanılmıştır (Adams vd., 2000). Shotgun sekanslama için elde edilen genomik DNA'nın tek bir hücre tipinden gelmesinin önemli olduğu belirtilmiştir (Lasken, 2012; Ishoey vd., 2008).

Shotgun sekanslamada, dizilenecek olan DNA rastgele olarak çok fazla sayıda küçük parçalara bölünmektedir (kırılır, kesilir). Daha sonra tüm genom dizilimini sıralı halde elde etmek için bilgisayar programları kullanılarak bu küçük DNA parçalarının birbirleri ile örtüşen uçlarını eşleştirilmesi ile ana iskelet parçaları ("scaffold kontigleri") oluşturulmaktadır. Bu contigler birbirlerine eklenerek sürekli bir sıra halinde bir araya getirilmektedirler. Bu şekilde tüm genom dizisinin sıralı bir şekilde dizilmesi sağlanmaktadır (Staden, 1979; Anderson, 1981). Tüm genomun assembly edilmesi (bir araya getirilmesi) için çeşitli programlar kullanılmaktadır. Birçok çalışma için yararlı olmasına rağmen, taslak genomların bitmemiş ve parçalanmış yapısından dolayı bu genomlarda, karşılaştırmalı genomik ve yapısal genomik analizlerinin verimli bir şekilde yapılması mümkün değildir (Ricker vd., 2012). Bunun yanı sıra, dizi, kapsam dışına (örneğin; kontig ya da scaffold ucuna) yerleştirilmişse ya da yanlış biraraya getirilmişse bazı genler kaçırılabilir (Klassen ve Currie, 2012). Bu yüzden çeşitli yazılımlar geliştirilmiştir. Klasik olarak kullanılan Phrap (http://www.phrap.org/phredphrap/) ve CAP3 (Huang ve Madan, 1999) gibi sekans assembler programlarının yerini De Bruijn graph algoritmasına dayanan (Pevzner vd., 2001; Compeau vd., 2011), Velvet (Zerbino ve Birney, 2008), ABySS (Simpson vd., 2009), Ray (Boisvert vd., 2010), SPAdes (Bankevich vd., 2012), SOAPdenovo (Luo vd., 2012) ve Newbler (Genivaldo vd., 2013) gibi programlar almıştır.

Newbler assembler programı, 454 Life Sciences firmasının geliştirmiş olduğu ve 454 sekanslama cihazları ile birlikte dağıtımı yapılan, doğru contigleri oluşturmak ve pirosekanslamada yapılan hataları da göz önünde bulundurarak geliştirilen bir yazılım programdır (Genivaldo vd., 2013). Bu program ile okumalardan elde edilen nükleotit dizilerinin birbirleri ile contigleri oluşturularak dizilerin ard arda eklenmesi ile scaffoldlar (iskele) oluşturulur ve bu scaffoldların birleştirilmesi ile genom dizisinin sıralı bir şekilde dizilmesi sağlanmaktadır. Ancak bu dizi oluşturulurken hem oluşturulan scaffoldlar arasında hem de scaffoldlar içerisinde dizinin devamlılık gösteremediği boşluk (gap) bölgeleri oluşmaktadır. Genom tamamlama contig ya da scaffoldların bilinmeyen bölgelerinin (gap bölgeleri) tamamlanmış sekanslara dönüştürülmesi ile gerçekleştirilir. Bir taslak genomu (draft genome), bir veya birkaç contig ve scaffold (iskele) dizisinden oluşabilir (Land vd., 2015). Burada hedef, genomun boşluk bölgelerinin tamamlanması ile tamamı dizilenmiş ("closed") ve eksiksiz elde edilmesidir (Mardis vd., 2002; Maiti ve Bouvagnet, 2001).

Özellikle yeni keşfedilmiş ve fenotipik olarak da ilginç özelliklere sahip mikroorganizmaların tüm genom dizilerinin bilinmesi, ardından gelecek olan çalışmalar için temel oluşturmakta ve yeni çalışmalara yol açmaktadır. *Bacillus boroniphilus* bakterisi de hem çok çalışılmamış bir bakteri olup hem de bu ekstrem özelliklere sahip bakteriler arasında yer almaktadır.

Bacillus boroniphilus bor açısından ekstremofilik bir bakteridir. Bu bakteri 450 mM borik asit içeren besiyeri ortamında üreme göstermektedir. *Bacillus boroniphilus* bakterisi hem yaşayabilmesi için bora ihtiyaç duymakta hem de çoğu canlının yaşayamayacağı bor konsantrasyonunda yaşamını devam ettirebilmektedir (Iftikhar vd., 2007). Doğada bor ile doğrudan temas halinde olan ve başka canlıların yaşayabilmesinin imkansız olduğu yüksek bor konsantrasyonlarında yaşayabildiği görülen canlılar oldukça ilgi çekici olmuştur. Bu canlılar ile yapılan çalışmalar sonucunda borun biyokimyasal ve moleküler rollerinin araştırılmasına olanak sağlayacağı düşünülmüştür. *Bacillus boroniphilus* bakterinin genom belirleme çalışmaları 2013 yılında tamamlanmış olup literatür dünyasına kazandırılmıştır (Çöl vd., 2014). Bu bakterinin genom dizisinin belirlenmesinde Çöl ve arkadaşları (2013) tarafından shotgun dizileme yöntemi (shotgun sequencing) kullanılmıştır.

Bacillus boroniphilus bakterisinin full genom sekansı belirlenirken elde edilen scaffold kontiglerinin bazı bölgelerinde birbirleri ile örtüşen kısımlarının olmadığı görülmüştür. Çalışmada, shotgun ve paired-end verileri kullanılarak yapılan birleştirmeler sonucu elde edilen "build"'de iki çeşit gap (boşluk) bölgesine rastlanmıştır. Bunlardan bir tanesi scaffoldların içerisinde bulunan gap'ler, diğeri ise scaffold arası gap'lerdir. Bu çalışmada *Bacillus boroniphilus*'un genomunun tamamlanması için bu gap (boşluk) bölgelerindeki sekans dizileri belirlenmiştir.

MALZEMELER ve YÖNTEMLER

Full genom dizisinin eksiksiz tamamlanması, genomdaki bilgi kaybını azaltmak ve ilgili organizmanın genomik özelliklerinin daha eksiksiz gösterilmesini sağlamaktadır. Draft (tamamlanmamış, taslak) genomlardaki aralık bölgelerinin kapatılmasında genel olarak izlenilen yol şu şekildedir; "1. Birbirini izleyen contiglerin iki ucundan primerlerin dizayn edilmesi", "2. PZR amplifikasyonu", "3. Sanger sekanslama", "4. Bölgesel assembly", "5. Manuel tamamlama". Bu çalışmada da PZR amplifikasyonu yöntemi kullanılmıştır.

Roche 454 Sekanslama platformu kullanılarak gerçekleştirilen shotgun dizileme deneyleri ile *Bacillus boroniphilus*'un *de novo* genom sekansı belirleme çalışmaları gerçekleştirilmiştir. Çöl ve ark.'larının gerçekleştirdiği bu çalışmada 84,872,624 bazlık okuma elde edilmiştir. Genomun 5' ucundan 3' ucuna doğru olarak sıralanabilmesi için 4 farklı eşleşebilen sonlu genom kütüphaneleri hazırlanmıştır. Tam plaka okuma sonucu 194.092.510 total baz sekanslanmıştır. Birleştirme analizleri sonucunda 3 tanesi büyük olmak üzere toplam 13 iskelet (scaffold) elde edilmiştir (Çöl vd., 2014).

Primer dizaynı

Scaffold kontiglerinin birleştirme çalışmalarında "build"lerdeki gap bölgelerinin sekans dizilerinin belirlenebilmesi için toplamda 298 (149 çift) adet primer tasarlanmıştır (Çizelge 1).

Çizelge 1. Scaffold kontiglerinin birleştirme işlemleri sonucunda elde edilen "build"lerdeki gap bölgelerinin sekans dizilerinin belirlenebilmesi için dizayn edilen primerlerin listesi.

Contig	**Position**	**Primer name**	**Primer #**	**seq**
contig00002	6121-6143	Sc01_C002_F1	col_ripurt.5	aatgtatacaaactagctaaata
contig00003	283-304	Sc01_C003_R1	col_ripurt.6	aaccggtgcgtatcttatatta
contig00003	8899-8917	Sc01_C003_F1	col_ripurt.7	cacatcctgttagtcggct
contig00004	441-461	Sc01_C004_R1	col_ripurt.8	caatcatcccggtaatagaat
contig00012	23980-24004	Sc01_C012_F1	col_ripurt.25	cggaaatattaaactaatgacaaga
contig00013	370-392	Sc01_C013_R1	col_ripurt.26	ttaaaacggtaagacaacaattt
contig00014	21321-21341	Sc01_C014_F1	col_ripurt.29	tggtgcgtgtacagaaataag
contig00015	379-403	Sc01_C015_R1	col_ripurt.30	aattgtcattgtttagacacagttt
contig00017	26674-26698	Sc01_C017_F1	col_ripurt.35	gtttatcccatctttatcctattct
contig00018	395-414	Sc01_C018_R1	col_ripurt.36	agattcagtaattgaattgt
contig00018	19223-19243	Sc01_C018_F1	col_ripurt.37	tgccatgttagcactactcac
contig00019	433-451	Sc01_C019_R1	col_ripurt.38	tcaggaaaacatcccgata
contig00021	5895-5912	Sc01_C021_F1	col_ripurt.43	ccgcccaggttcttatta
contig00022	476-496	Sc01_C022_R1	col_ripurt.44	taaggattatccgcacataga
contig00024	7857-7881	Sc01_C024_F1	col_ripurt.49	ctttggagtgaaataaaaggtagta
contig00025	369-392	Sc01_C025_R1	col_ripurt.50	gaaatctttgataacccagtagaa
contig00028	41957-41979	Sc01_C028_F1	col_ripurt.57	tcatattatctaaaatacttgac
contig00029	370-390	Sc01_C029_R1	col_ripurt.58	tggttgatgcagctttctatt
contig00039	57172-57194	Sc01_C039_F1	col_ripurt.79	cgcatctctaaataataccacag
contig00040	331-355	Sc01_C040_R1	col_ripurt.80	cattataaatagtatttataggacg
contig00042	23615-23636	Sc01_C042_F1	col_ripurt.85	gcatagccatcatttacacaga
contig00043	274-292	Sc01_C043_R1	col_ripurt.86	tgccatatgcactagcgtt
contig00051	23423-23442	Sc01_C051_F1	col_ripurt.104	aaatcgagtcgtaccacaca
contig00052	434-455	Sc01_C052_R1	col_ripurt.105	aagatttataggagtaattgta
contig00053	5694-5713	Sc02_C053_F1	col_ripurt.108	caatcccattctcaaacgac
contig00054	267-289	Sc02_C054_R1	col_ripurt.109	ggcagatattgtttctacgtaag
contig00062	7008-7028	Sc02_C062_F1	col_ripurt.126	ggaaatattttaaaagataac
contig00063	105-128	Sc02_C063_R1	col_ripurt.127	ccaaagatttatcaaaatcattgc
contig00068	14834-14856	Sc02_C068_F1	col_ripurt.138	tcatatatgaaccacgatgaaat
contig00069	332-354	Sc02_C069_R1	col_ripurt.139	tgaaggctgtcaagtaaattatt
contig00074	81476-81494	Sc02_C074_F1	col_ripurt.150	gaaggtcatgacatttaca
contig00075	126-149	Sc02_C075_R1	col_ripurt.151	tcctacttatctgagttaacggac
contig00079	4222-4245	Sc02_C079_F1	col_ripurt.160	ggataaggtaaacagagtaaacga
contig00080	469-488	Sc02_C080_R1	col_ripurt.161	caccgcattaattcgataca
contig00084	151144-151158	Sc02_C084_F1	col_ripurt.170	ggaggaataggggga
contig00085	493-512	Sc02_C085_R1	col_ripurt.171	ctgaacctagtactacttcg
contig00088	77485-77504	Sc03_C088_F1	col_ripurt.178	aagcgcaatgataatgacac
contig00089	483-506	Sc03_C089_R1	col_ripurt.179	tcccaaactttacagtcaaattat
contig00098	24116-24133	Sc04_C098_F1	col_ripurt.198	agcgataaagagagagtt
contig00099	300-319	Sc04_C099_R1	col_ripurt.199	gctgaatgcagcatagtggt

contig00106	10972-10995	Sc04_C106_F1	col_ripurt.214	gttaagatccatagttacgtagtt
contig00107	381-399	Sc04_C107_R1	col_ripurt.215	cgggttgcaattcacatac
contig00107	9835-9858	Sc04_C107_F1	col_ripurt.216	gcttattccaggatattaaattgt
contig00108	433-457	Sc04_C108_R1	col_ripurt.217	cgactataaaaccagtaacctacaa
contig00110	27299-27317	Sc04_C110_F1	col_ripurt.222	tgactggcttgctgagttc
contig00111	397-419	Sc04_C111_R1	col_ripurt.223	caaaccattttcataggaaagac
contig00115	5034-5053	Sc04_C115_F1	col_ripurt.232	gttaaatccttatggaacta
contig00116	448-466	Sc04_C116_R1	col_ripurt.233	acaaacatttgagctaata
contig00122	26520-26543	Sc05_C122_F1	col_ripurt.246	aggagcaaattaaactcataaaga
contig00123	373-395	Sc05_C123_R1	col_ripurt.247	cccaagattgtctatatgacgaa
contig00126	2113-2135	Sc05_C126_F1	col_ripurt.254	ccaagctgcttgtattacttcac
contig00127	513-533	Sc05_C127_R1	col_ripurt.255	aataataacaatgcaaatact
contig00143	15593-15616	Sc07_C143_F1	col_ripurt.288	aagtacaacatatattaaaagtaa
contig00144	538-559	Sc07_C144_R1	col_ripurt.289	ggcaacgtcactctaactgata
contig00144	64479-64502	Sc07_C144_F1	col_ripurt.290	caaaccagtcgtaaataatacaaa
contig00145	389-413	Sc07_C145_R1	col_ripurt.291	attatcttcttctggaaatctttct
contig00146	8582-8601	Sc08_C146_F1	col_ripurt.294	cctccagcatcatatgtgtt
contig00147	609-623	Sc08_C147_R1	col_ripurt.295	ggcgaactcggcatt
contig00147	7805-7820	Sc08_C147_F1	col_ripurt.296	tcgccatcgtatagac
contig00148	467-491	Sc08_C148_R1	col_ripurt.297	aatcagagaataaccctaaatcatt

DNA izolasyonu ve PZR amplifikasyonu

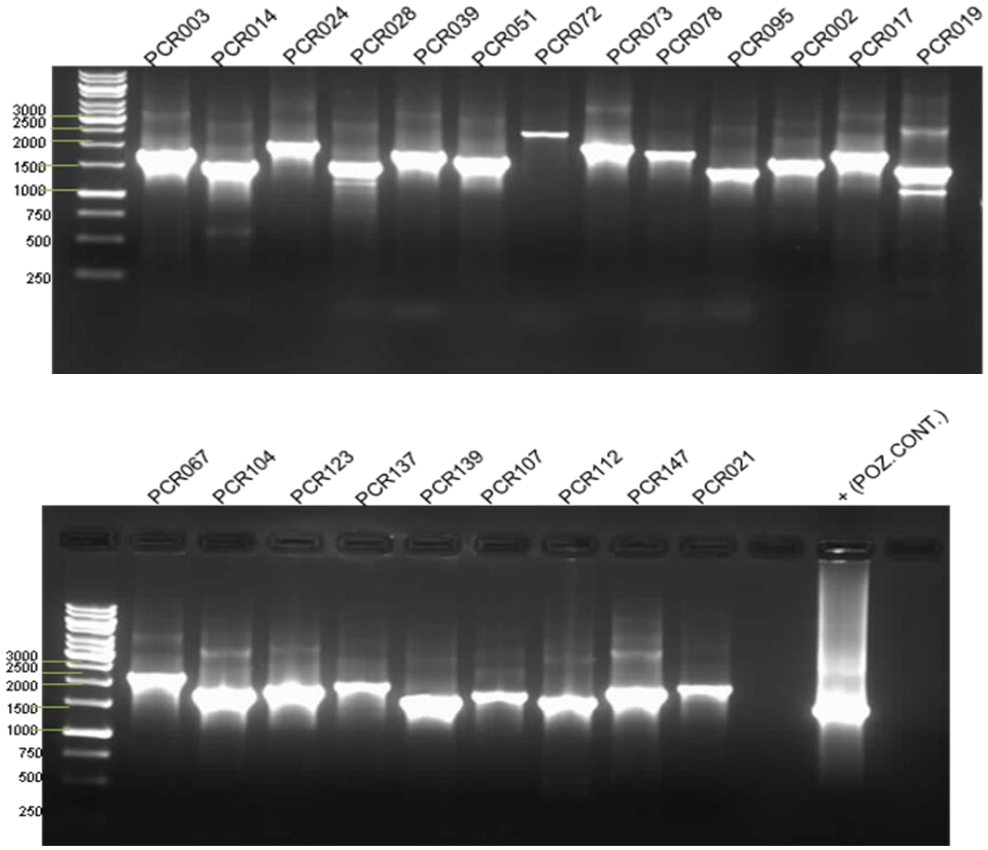
Kalıp DNA olarak B. boroniphilus genomik DNA'sı kit (Zymo, D4068) kullanılarak izole edilmiştir. Uygun primer çiftleri için birkaç farklı Tm derecesi kullanılarak en uygun PZR koşullarına ulaşılmıştır ve PZR reaksiyonları gerçekleştirilmiştir. Elde edilen PZR ürünleri jel elektroforez sisteminde yürütülerek görüntülenmiştir.

Sekanslama ve Biyoinformatik analizler

Elde edilen PZR ürünleri Sanger Sekanlama Tekniği ile sekanslanmış ve ürünlerin DNA dizileri belirlenmiştir. Bu aşama hizmet alımı ile gerçekleştirilmiştir. Elde edilen nükleotid sekans sonuçları BioEdit biyoinformatik yazılım paketiyle analiz edilmiştir. Gelen sonuçlardaki QV≥20 değerlerinin dikkate alınması ve BioEdit programıyla okunan dizilerin piklerinin göz önünde bulundurulması ile sekans kalitesi değerlendirilmiştir. Elde edilen okumalar Bioedit programının contig analizi (CAP) ile birleştirilerek contig gen dizileri elde edilenler bu diziler ile elde edilemeyenler sekans dizileri ile BlastN ve BlastX veritabanları kullanılarak analiz edilmiştir (https://blast.ncbi.nlm.nih.gov/ Blast.cgi). PZR ürünlerinin Blast analizleri ile tanımlanmaları gerçekleştirildikten sonra NCBI Blast veritabanından GenBank numaraları alınarak veritabanına yüklenmiştir (https://submit.ncbi.nlm.nih.gov/).

BULGULAR

Uygun Tm derecelerine karar verilen örnekler PZR ile çoğaltılarak 70 adet PZR reaksiyonu gerçekleştirilmiş ve elde edilen ürünler agaroz jellerde yürütülmüştür (Şekil 1). PZR ürünlerinden seçilen 30 ürün dizilenerek (Çizelge 2) BlastN ve BlastX (https://blast.ncbi.nlm.nih.gov/Blast.cgi) veritabanları ile biyoinformatik analizleri gerçekleştirilmiştir. Sekans sonucu elde edilen bilgiler değerlendirildiğinde "gap" bölgelerinin çoğunluğunun "transpozaz" olduğu tespit edilmiştir (Çizelge 3).



Şekil 1. PZR ürünlerinin %0,8lik agaroz jel görüntülerine örnek

Çizelge 2. Sekansa gönderilen PZR ürünlerinin listesi

	Sl No	**Scaffold**	**ContigA**	**ContigB**	**expected gap**	**primer1 Name**	**primer2 Name**
1	PZR002	Sc001	contig00002	contig00003	1400	Sc01_C002_F1	Sc01_C003_R1
2	PZR003	Sc001	contig00003	contig00004	1251	Sc01_C003_F1	Sc01_C004_R1
3	PZR014	Sc001	contig00014	contig00015	1266	Sc01_C014_F1	Sc01_C015_R1
4	PZR017	Sc001	contig00017	contig00018	1149	Sc01_C017_F1	Sc01_C018_R1
5	PZR018	Sc001	contig00018	contig00019	1301	Sc01_C018_F1	Sc01_C019_R1
6	PZR021	Sc001	contig00021	contig00022	861	Sc01_C021_F1	Sc01_C022_R1
7	PZR024	Sc001	contig00024	contig00025	987	Sc01_C024_F1	Sc01_C025_R1
8	PZR028	Sc001	contig00028	contig00029	1047	Sc01_C028_F1	Sc01_C029_R1
9	PZR039	Sc001	contig00039	contig00040	966	Sc01_C039_F1	Sc01_C040_R1
10	PZR042	Sc001	contig00042	contig00043	899	Sc01_C042_F1	Sc01_C043_R1
11	PZR051	Sc001	contig00051	contig00052	788	Sc01_C051_F1	Sc01_C052_R1
12	PZR052	Sc002	contig00053	contig00054	2118	Sc02_C053_F1	Sc02_C054_R1
13	PZR061	Sc002	contig00062	contig00063	1495	Sc02_C062_F1	Sc02_C063_R1
14	PZR067	Sc002	contig00068	contig00069	20	Sc02_C068_F1	Sc02_C069_R1
15	PZR073	Sc002	contig00074	contig00075	575	Sc02_C074_F1	Sc02_C075_R1
16	PZR078	Sc002	contig00079	contig00080	1868	Sc02_C079_F1	Sc02_C080_R1
17	PZR083	Sc002	contig00084	contig00085	577	Sc02_C084_F1	Sc02_C085_R1
18	PZR086	Sc003	contig00088	contig00089	1225	Sc03_C088_F1	Sc03_C089_R1
19	PZR095	Sc004	contig00098	contig00099	483	Sc04_C098_F1	Sc04_C099_R1
20	PZR103	Sc004	contig00106	contig00107	1855	Sc04_C106_F1	Sc04_C107_R1
21	PZR104	Sc004	contig00107	contig00108	971	Sc04_C107_F1	Sc04_C108_R1
22	PZR107	Sc004	contig00110	contig00111	5378	Sc04_C110_F1	Sc04_C111_R1
23	PZR112	Sc004	contig00115	contig00116	1873	Sc04_C115_F1	Sc04_C116_R1
24	PZR118	Sc005	contig00122	contig00123	1351	Sc05_C122_F1	Sc05_C123_R1
25	PZR122	Sc005	contig00126	contig00127	657	Sc05_C126_F1	Sc05_C127_R1
26	PZR137	Sc007	contig00143	contig00144	5457	Sc07_C143_F1	Sc07_C144_R1
27	PZR138	Sc007	contig00144	contig00145	1108	Sc07_C144_F1	Sc07_C145_R1
28	PZR139	Sc008	contig00146	contig00147	477	Sc08_C146_F1	Sc08_C147_R1
29	PZR147	Sc008	contig00149	contig00150	1137	Sc08_C147_F1	Sc08_C148_R1
30	PZR012	Sc001	contig00012	contig00013	928	Sc01_C012_F1	Sc01_C013_R1

Çizelge 3. Sekansa gönderilen PZR ürünlerinin Blast analizleri ile tanımlanmaları

Sl No	**F-R/Contiq Sequence**	**GenBank Number**
PZR002_F	IS1380 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370944
PZR002_R	IS1380 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370945
PZR003_contig	IS1380 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370931
PZR014_F	IS701 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370946
PZR014_R	transposase IS4 family protein [*Bacillus boroniphilus strain DSM 17376*]	MN370947
PZR017_contig	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370933
PZR018_F	IS1380 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370948
PZR018_R	IS1380 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370949
PZR021_R	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370973
PZR024_contig	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370934
PZR028_contig	IS701 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370935
PZR039_F	hypothetical protein [*Bacillus boroniphilus strain DSM 17376*]	MN370942
PZR039_R	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370950
PZR042_F	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370951
PZR042_R	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370952
PZR051_contig	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370936
PZR061_F	IS1380 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370955
PZR061_R	IS1380 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370956
PZR073_contig	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370937
PZR078_contig	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370938
PZR083_F	hypothetical protein G3A_22275 [*Bacillus boroniphilus strain DSM 17376*]	MN370943
PZR083_R	hypothetical protein G3A_22275 *[Bacillus boroniphilus strain DSM 17376*]	MN370957
PZR086_F	IS1380 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370958
PZR086_R	IS1380 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370959
PZR095_R	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370960
PZR103_F	MULTISPECIES: IS66 family insertion sequence element accessory protein TnpB [*Bacillus boroniphilus strain DSM 17376*]	MN370961
PZR103_R	IS66 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370962
PZR104_contig	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370939
PZR112_contig	group II intron reverse transcriptase/maturase [*Bacillus boroniphilus strain DSM 17376*]	MN370940
PZR118_F	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370963
PZR118_R	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370964
PZR122_F	IS1380 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370965
PZR122_R	IS1380 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370966
PZR137_F	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370967
PZR137_R	IS4 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370968
PZR138_F	IS1380 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370969
PZR138_R	IS1380 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370970
PZR139_F	tranposase [*Bacillus boroniphilus strain DSM 17376*]	MN370971
PZR139_R	transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370972
PZR147_contig	IS3 family transposase [*Bacillus boroniphilus strain DSM 17376*]	MN370941
PZR012_contig	HAD family hydrolase [*Bacillus boroniphilus strain DSM 17376*]	MN370932

SONUÇ

Yeni nesil dizileme teknolojileri, genomların daha hızlı ve uygun şekilde dizilenmelerine olanak sağlamaktadır ancak, özellikle büyük genomlara sahip organizmaların tüm genom dizisinin açığa çıkartılmasında önemli biyoinformatik zorluklar yaşanmaktadır. Newbler assembler ve mapper (gsAssembler, gsMapper), özellikle Roche/454 Life Science sekanslama teknolojisindeki okumalarla çalışmak için geliştirilmiştir. Bu programlar, genom dizileme çalışmalarındaki verilerle başa çıkmak için en uygun programlardan biridir. Newbler birçok büyük ve küçük genomun assembly edilmesi (bir araya getirilmesi) için kullanılmıştır (bakteriler, Atlantik morina balığı, domates vb.). Son zamanlarda bu sisteme çoklu sekanslama teknolojileri (multiple sequencing technologies) de eklenerek, program hibrid assembly (birleştirilmesi) programlarından biri haline getirilmiştir. 2013 yılında Genom Biyolojisi ve Teknolojisindeki Gelişmelerde (Advances in Genome Biology and Technology (AGBT)), Roche, Newbler programını, hibrid 454 ve Illumina veri seti ile birlikte insan genomunu geliştirmek için kullandığını duyurmuştur. Ancak, sekans kalitesinin düşük olması, tekrarlayan diziler ve kısa okuma uzunlukları, de novo genom kurulumunu zorlaştırmaktadır. Bunlar, çoğu zaman sekans içerisinde boşluk "gap" bölgelerine ve bilinmeyen bölgelerde tespit edilemeyen nükleotid dizilerine sebep olmaktadır. Bu boşlukların bazıları, işlenmemiş datalardaki, açığa çıkartılmamış bilgilerin tekrardan işlenmesi ile kapatılabilir.

Bacillus boroniphilus bakterisinin genom tamamlama çalışmaları sırasında belirlenemeyen sekans dizilerinin açıklığa kavuşturulması için yapılan bu çalışmada, gap bölgelerini tamamlamak için yapılan PZR işlemleri sonucunda bu bölgelerin neredeyse hepsinin transpozaz oldukları görülmüştür.

Transpozonlar genom içerisinde tekrarlayan dizilerdir. Genel olarak bakıldığında tekrarlayan diziler, genomun farklı bölgelerinde çoklu kopyalardan meydana gelen dizilerdir. Bu konu hakkında literatür incelendiğinde, bu farklı tekrar dizilerinden gelen sekans okumaları birbirine çok benzer olduğu için assembly (birleştirme) programlarının bunlar arasında ayrım yapmakta zorlandığı, gördüğü benzer diziyi ekleme yaparak diziyi birleştirmeye devam ettiği bu yüzden scaffoldlarda gap (boşluk) bölgelerine neden olduğu görülmüştür. Ayrıca, bu yüzden bu dizilerin miktarı ve dağılımının dizilerin biraraya getirilmesini büyük ölçüde etkilediği görülmüştür. Bu durum, genomdaki birbirine uzak olan bölgelerin biraraya getirildiği yanlış assemblylere (birleştirmelere) ve takrarlayan dizilerin kopya ve tekrarlarının yanlış tahminlerine yol açabilmektedir.

Tekrarlanan dizilerin assembly (bir araya getirilme) problemini çözebilmek için, okumaların, tekrar dizilerinin çevresindeki sekansları da kapsayacak şekilde uzun olması ve tekrarlayan dizi ile çevresindeki dizilerin birlikte okunması gerekmektedir. Bu nedenle eğer tekrar dizisi yüksek olan bir genomla çalışıldığı biliniyorsa uzun okuma yapabilen bir teknoloji ile assembly (bir araya getirilmesi) yapılması uygun olabilir.

Sonuç olarak elde edilen bilgiler ışığında Newbler Assembler programının algoritmasının, full genom sekans çalışmalarında elde edilen dizilerin sıralamasında yetersiz görüldüğü tartışılmıştır. Genom bitirme çalışmaları hem yoğun optimizasyon ve tekrarlar gerektiren çalışmalar olup hem de oldukça uzun, masraflı ve ince çalışmalardır. Yapılan literatür taramalarında shotgun dizi sekanslama çalışmalarında görülen gap bölgelerinin genellikle tekrarlayan dizilerden meydana geldiği ve bu sekans dizilerinin gap bölgelerine entegrelerinin daha zor olduğu belirtilmiştir. Transpozaz dizilerinin de tekrarlayan DNA dizileri olması karşımıza çıkan bu sonucu açıklamaktadır. Ayrıca genom dizileri ortaya çıkartılırken karşılaşılan transpozaz dizilerinin anlaşılmasının da bakterilerdeki genom çeşitliliğini anlamada yardımcı olacağı düşünülmektedir. Yeni nesil genom sekanslama çalışmalarındaki verimliliğin arttırılması için, gap bölgeleri hakkında elde edilen bu tür bilgiler kullanılmalıdır.

Acknowledgement

Bu çalışma Bor Araştırma Enstitüsü (BOREN Ç0238) tarafından desteklenmiştir.

REFERENCES

1. Iftikhar A., Akira Y, Fujiwara T. 2007 “A novel highly boron tolerant bacterium, *Bacillus boroniphilus* sp. nov., isolated from soil, that requires boron for its growth. Extremophiles.” 2007 Mar;11(2):217-24. Epub 2006 Oct 27.
2. Çöl, B., Ozkeserli, Z., Kumar, D., Ozdag, H., & Alakoç, Y. D. (2014). “Genome Sequence of the Boron-Tolerant and -Requiring Bacterium *Bacillus boroniphilus*.” Genome announcements, 2(1), e00935-13. doi:10.1128/genomeA.00935-13
3. Lasken R. S. “Genomic sequencing of uncultured microorganisms from single cells” (2012) Nat Rev Microbiol, 10, pp. 631-640. doi:10.1038/nrmicro2857
4. Ishoey T., Woyke T., Stepanauskas R., Novotny M., Lasken R. S. (2008) “Genomic sequencing of single microbial cells from environmental samples” Curr Opin Microbiol, 11, pp. 198-204.
5. Staden, R. (1979). "A strategy of DNA sequencing employing computer programs". Nucleic Acids Research. 6 (70): 2601–10. doi:10.1093/nar/6.7.2601. PMC 327874. PMID 46119
6. Anderson, S. (1981). "Shotgun DNA sequencing using cloned DNase I-generated fragments". Nucleic Acids Research. 9 (13): 3015–27. doi:10.1093/nar/9.13.3015. PMC 327328. PMID 6269069.
7. Gardner, Richard C.; Howarth, Alan J.; Hahn, Peter; Brown-Luedi, Marianne; Shepherd, Robert J.; Messing, Joachim (1981). "The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing". Nucleic Acids Research. 9 (12): 2871–2888. doi:10.1093/nar/9.12.2871. ISSN 0305-1048. PMC 326899. PMID 6269062.
8. Edwards, A., Caskey, T. (1991). "Closure strategies for random DNA sequencing". Methods: A Companion to Methods in Enzymology. 3 (1): 41–47. doi:10.1016/S1046-2023(05)80162-8

9. Edwards, A., Voss, H., Rice, P., Civitello, A., Stegemann, J., Schwager, C.; Zimmerman, J.; Erfle, H.; Caskey, T.; Ansorge, W. (1990). "Automated DNA sequencing of the human HPRT locus". Genomics. 6 (4): 593–608. doi:10.1016/0888-7543(90)90493-E. PMID 2341149.

10. Roach, JC; Boysen, C; Wang, K; Hood, L (1995). "Pairwise end sequencing: a unified approach to genomic mapping and sequencing". Genomics. 26 (2): 345–353. doi:10.1016/0888-7543(95)80219-C. PMID 7601461

11. Fleischmann, RD; et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd". Science. 269 (5223): 496–512. Bibcode:1995Sci...269..496F. doi:10.1126/science.7542800. PMID 7542800 .

12. Adams, MD; et al. (2000). "The genome sequence of Drosophila melanogaster" (PDF). Science. 287 (5461): 2185–95. Bibcode:2000Sci...287.2185.. CiteSeerX 10.1.1.549.8639. doi:10.1126/science.287.5461.2185. PMID 10731132.

13. Maiti A. K., Bouvagnet P. (2001) "Assembling and gap filling of unordered genome sequences through gene checking" Genome Biology volume 2(9), DOI: 10.1186/gb-2001-2-9-preprint0008.

14. Mark J. Chaisson and Pavel A. P. (2008). "Short read fragment assembly of bacterial genomes". Genome Res. 2008. 18: 324-330.

15. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al.(2006) "Genome sequencing in microfabricated high-density picolitre reactors." Nature 437:376–380.

16. Ronaghi, M., Mathias, U., Nyren, P.(1998) DNA sequencing: A sequencing method based on real-time pyrophosphate. Science 281:363–365.

17. Chain P. S. G., Grafham D V, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, et al. (2009) "Genomics. Genome project standards in a new era of sequencing." Science 326:236-237.

18. Sanger F, Nicklen S and Coulson AR (1977) "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci USA 74:5463-5467.

19. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L and Law M (2012) "Comparison of next-generation sequencing systems." J Biomed Biotechnol 2012:251364.

20. Huang X, Madan A (1999) "CAP3: A DNA sequence assembly program." Genome Res 9:868-877.

21. Pevzner P. A, Tang H., Waterman M.S. (2001) "An Eulerian path approach to DNA fragment assembly." Proc Natl Acad Sci U S A 98:9748-9753.

22. "Compeau P.E.C., Pevzner P.A., Tesler G. (2011) "How to apply de Bruijn graphs to genome assembly." Nat Biotechnol 29:987-991.

23. Zerbino D.R., Birney E. (2008) "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs." Genome Res 18:821-829.

24. Simpson J.T, Wong K., Jackman S.D., Schein J.E., Jones S.J.M., Birol I. (2009) "ABySS: A parallel assembler for short read sequence data." Genome Res 19:1117-23.

25. Boisvert S., Laviolette F.,nCorbeil J. (2010) "Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies." J Comput Biol 17:1519-1533.

26. Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D., et al. (2012) "SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing." J Comput Biol 19:455-477.

27. Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q., Liu Y., et al. (2012) "SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler." Gigascience 1:18.

28. Mardis E., McPherson J., Martienssen R., Wilson R.K., McCombie W.R. (2002) "What is finished, and why does it matter." Genome Res 12:669-671.

29. Land M., Hauser L., Jun S.R, Nookaew I., Leuze M.R., Ahn T.H., Karpinets T., Lund O., Kora G., Wassenaar T., et al. (2015) "Insights from 20 years of bacterial genome sequencing." Funct Integr Genomics 15:141-161.

30. Ricker N., Qian H. Fulthorpe R.R. (2012) "The limitations of draft assemblies for understanding prokaryotic adaptation and evolution." Genomics 100:167-175.

31. Klassen J.L., Currie C.R. (2012) "Gene fragmentation in bacterial draft genomes: Extent, consequences and mitigation." BMC Genomics 13:14.

32. Genivaldo, G.Z., Silva, Bas E., Dutilh, D., Matthews, K., Elkins, R., Schmieder, Elizabeth A., Dinsdale, R. A. E. (2013). "Combining de novo and reference-guided assembly with scaffold_builder". Source Code Biomed Central. 8 (23): 23.