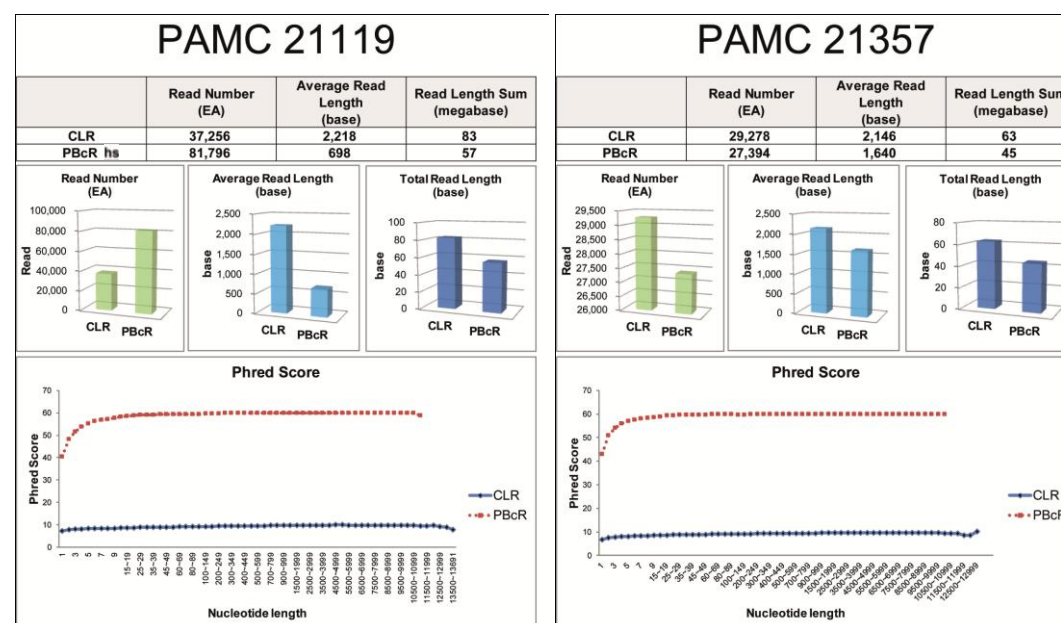
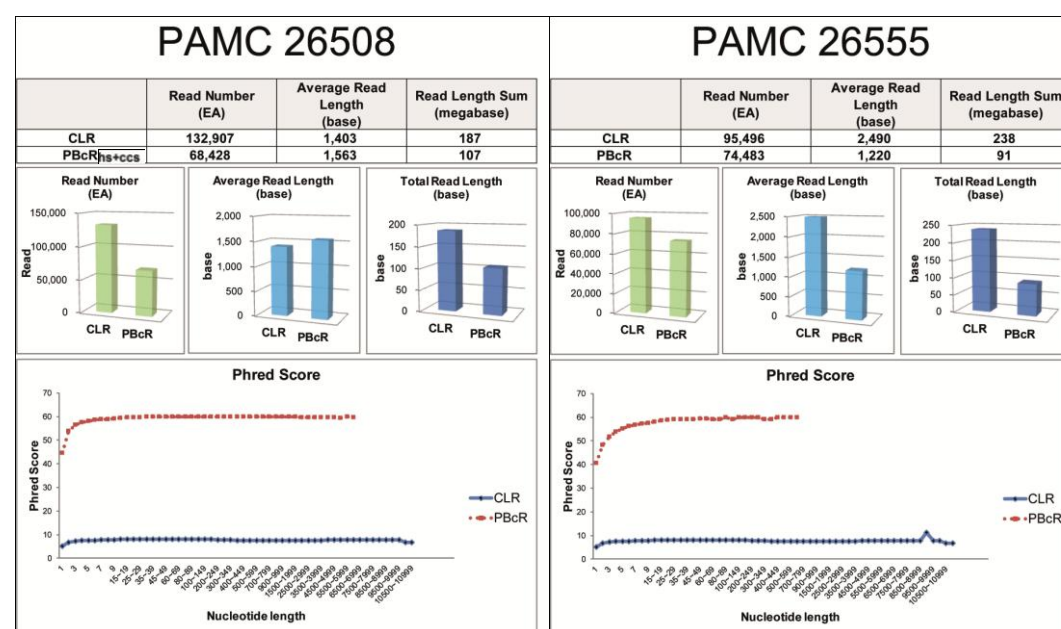
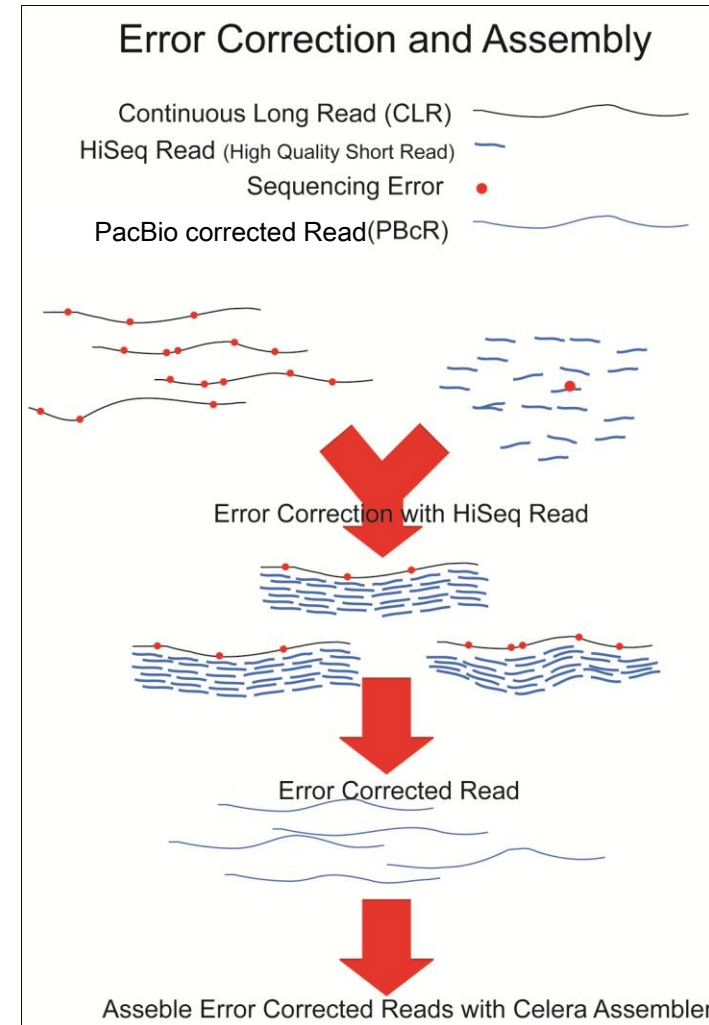


Sujin Kim¹, Seung-Chul Shin², Hyun Park² & Jong-Eun Lee¹
¹DNA Link, Inc. Seoul Korea / ²Korea Polar Research Institute, Incheon Korea

Abstract

In the last year, high-throughput sequencing technologies have progressed from proof-of-concept to production quality. Although each technology is able to produce vast quantities of sequence information, in every case the underlying chemistry limits reads to very short lengths. We present an examining *de novo* assembly comparison with bacterial genome assembly varying genome size (from 3.1Mb to 7.6Mb) and different G+C contents (from 43% to 71%), respectively. We analyzed Solexa reads, 454 reads and Pacbio RS reads from *Streptomyces* sp. (Genome size, 7.6 Mb; G+C content, 71%), *Psychrobacter* sp. (Genome size, 3.5 Mb; G+C content, 43%), *Salinibacterium* sp. (Genome size, 3.1 Mb; G+C content, 61%) and *Frigoribacterium* sp. (Genome size, 3.3 Mb; G+C content, 63%). We assembly each bacterial genome using Celera assembler 7.0 with and without PacBio RS reads. We found out that the assemble result with Pacbio RS reads have less contigs and scaffolds, and better N50 values.

Methods



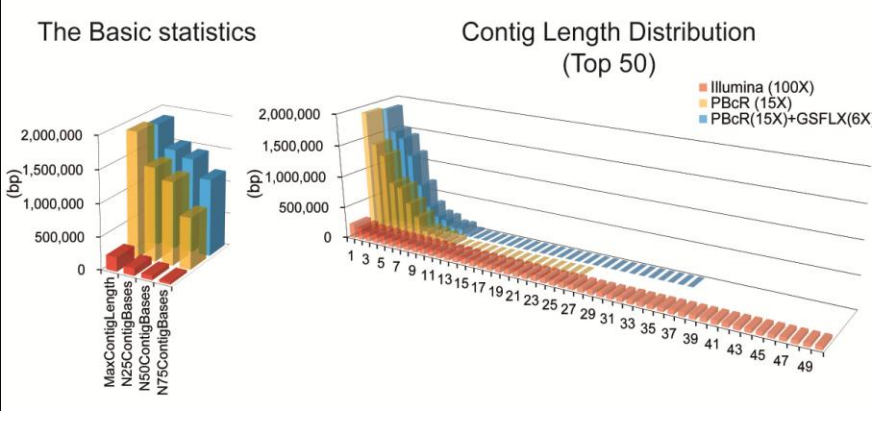
PAMC26508 : *Streptomyces* sp.
GC content : 70.90%
Genome Size : 7.6Mb

Assembler : Celera Assembler
Sequencing technologies :
Illumina : 500bp paired-end library
GS-FLX : 8kb paired-end library
Pacbio : Continuous Long Read

Assembly results

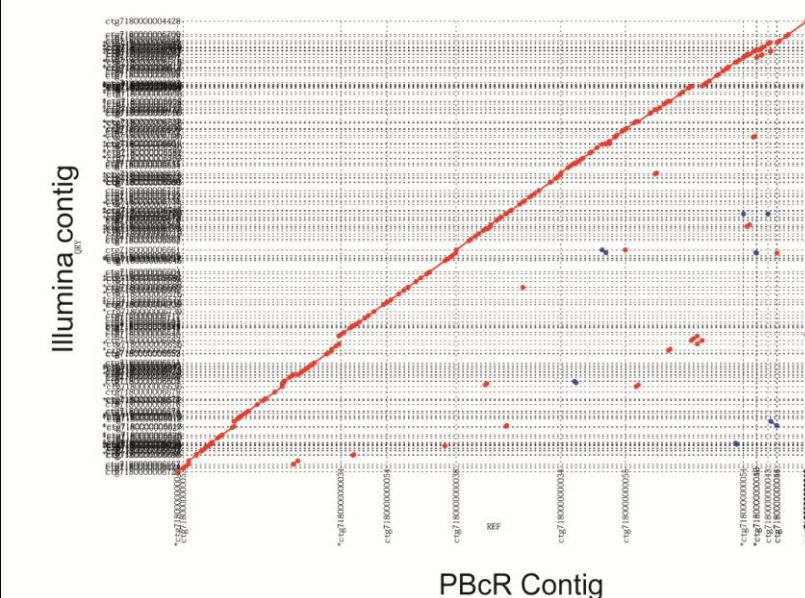
	Illumina (100X)	PbCR (15X)	PbCR(15X)+GSFLX(6X)
Scaffolds			
TotalScaffolds (EA)	80	26	31
TotalContigsInScaffolds (EA)	185	26	36
TotalBasesInScaffolds (bp)	7,595,626	7,669,881	7,784,673
MaxBasesInScaffolds (bp)	449,494	1,915,364	1,871,049
N50ScaffoldsBases (bp)	202,976	1,268,506	1,434,415
Contigs			
TotalContigsInScaffolds (EA)	185	26	36
MaxContigLength (bp)	221,220	1,915,364	1,869,455
N50ContigBases (bp)	68,326	1,268,506	1,434,415
TotalBigContigs (10kb+)	141	11	11
BigContigLength (bp)	7,372,844	7,637,132	7,723,210

The improvement of de novo assemblies with Pacbio RS



Assembly comparison

Program : mummer 3.0



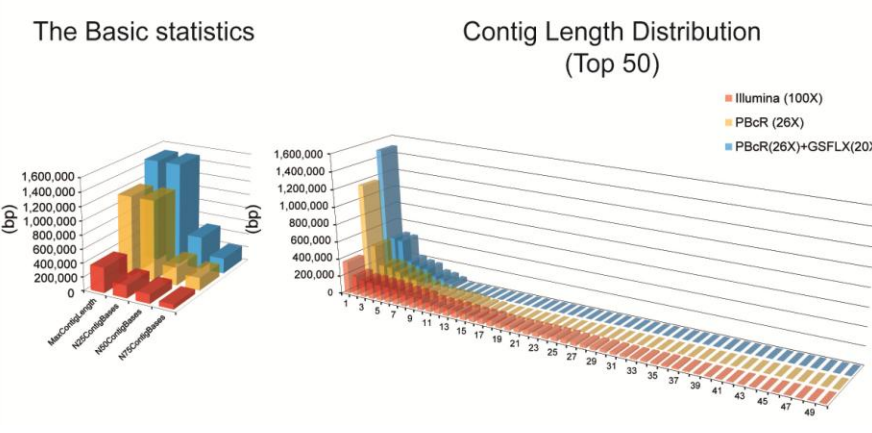
PAMC26555 : *Frigoribacterium* sp.
GC content : 62.87%
Genome Size : 3.3Mb

Assembler : Celera Assembler
Sequencing technologies :
Illumina : 500bp paired-end library
GS-FLX : 8kb paired-end library
Pacbio : Continuous Long Read

Assembly results

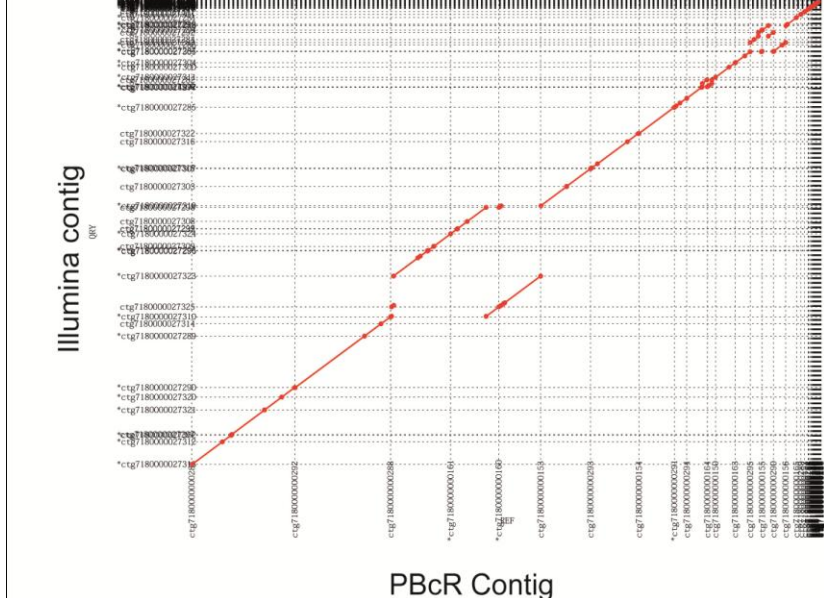
	Illumina (100X)	PbCR (26X)	PbCR(26X)+GSFLX(20X)
Scaffolds			
TotalScaffolds (EA)	75	60	80
TotalContigsInScaffolds (EA)	93	60	96
TotalBasesInScaffolds (bp)	3,310,520	3,324,735	3,432,448
MaxBasesInScaffolds (bp)	453,832	1,164,979	2,443,060
N50ScaffoldsBases (bp)	177,721	263,283	2,443,060
Contigs			
TotalContigsInScaffolds (EA)	93	60	96
MaxContigLength (bp)	366,603	1,164,979	1,492,435
N50ContigBases (bp)	136,524	263,283	469,317
TotalBigContigs (10kb+)	38	16	9
BigContigLength (bp)	3,203,637	3,246,689	3,246,608

The improvement of de novo assemblies with Pacbio RS



Assembly comparison

Program : mummer 3.0



Results

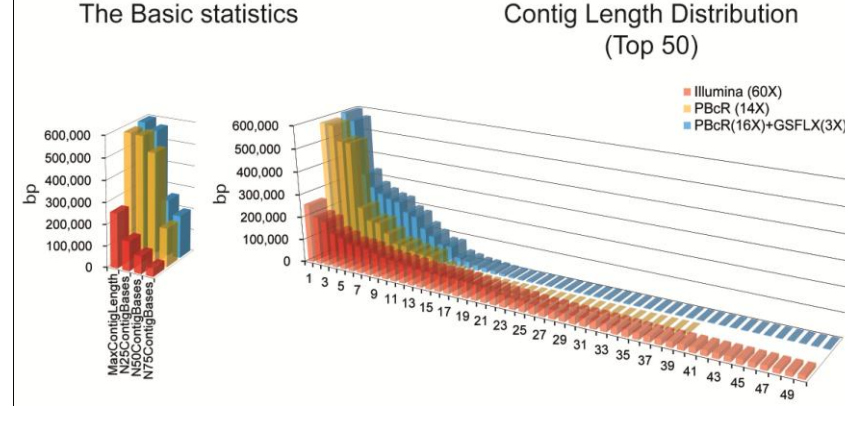
PAMC21119 : *Psychrobacter* sp.
GC content : 43.34%
Genome Size : 3.5 Mb

Assembler : Celera Assembler
Sequencing technologies :
Illumina : 500bp paired-end library
GS-FLX : 8kb paired-end library
Pacbio : Continuous Long Read

Assembly results

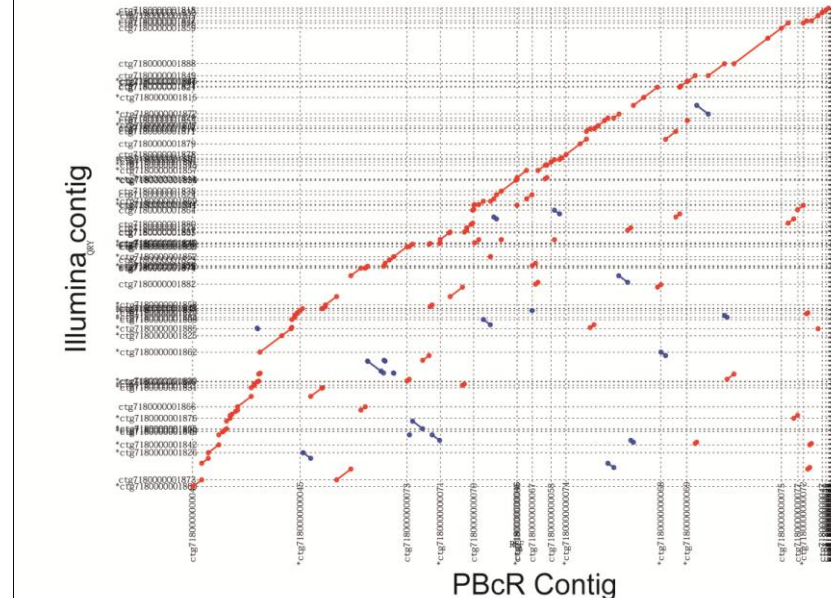
	Illumina (60X)	PbCR (14X)	PbCR(16X)+GSFLX(3X)
Scaffolds			
TotalScaffolds (EA)	67	39	60
TotalContigsInScaffolds (EA)	76	39	67
TotalBasesInScaffolds (bp)	3,443,158	3,502,463	3,591,719
MaxBasesInScaffolds (bp)	352,708	582,027	598,054
N50ScaffoldsBases (bp)	93,139	511,916	258,230
Contigs			
TotalContigsInScaffolds (EA)	76	39	67
MaxContigLength (bp)	285,324	582,027	598,054
N50ContigBases (bp)	82,863	511,916	258,230
TotalBigContigs (10kb+)	60	16	19
BigContigLength (bp)	3,372,293	3,439,229	3,441,479

The improvement of de novo assemblies with Pacbio RS



Assembly comparison

Program : mummer 3.0



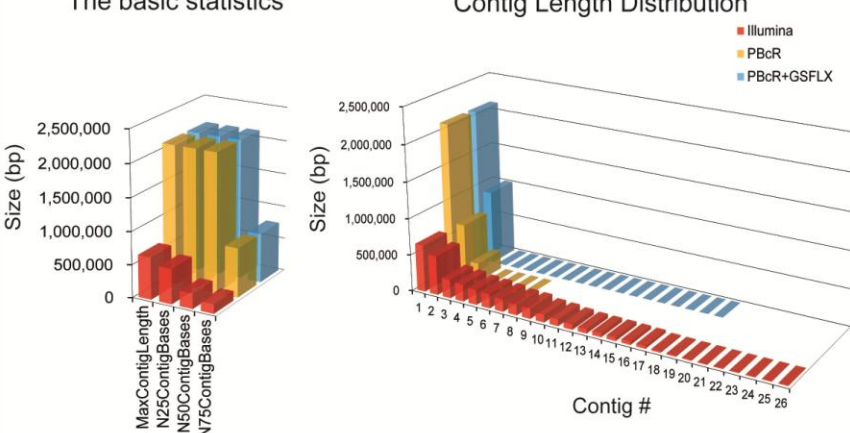
PAMC21357 : *Salinibacterium* sp.
GC content : 60.2 %
Genome Size : 3.1 Mb

Assembler : Celera Assembler
Sequencing technologies :
Illumina : 500bp paired-end library
GS-FLX : 8kb paired-end library
Pacbio : Continuous Long Read

Assembly results

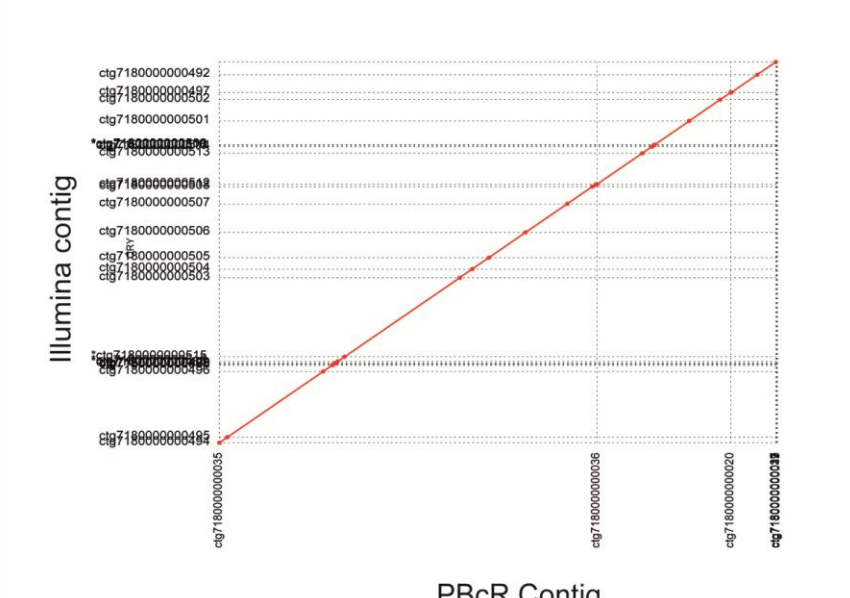
	Illumina (100X)	PbCR (16X)	PbCR(16X)+GSFLX(8X)
Scaffolds			
TotalScaffolds (EA)	9	7	9
TotalContigsInScaffolds (EA)	26	7	9
TotalBasesInScaffolds (bp)	3,102,360	3,119,401	3,120,661
MaxBasesInScaffolds (bp)	751,436	2,110,572	2,112,607
N50ScaffoldsBases (bp)	631,463	2,110,572	2,112,607
Contigs			
TotalContigsInScaffolds (EA)	26	7	9
MaxContigLength (bp)	643,647	2,110,572	2,112,607
N50ContigBases (bp)	233,362	2,110,572	2,112,607
TotalBigContigs (10kb+)	20	3	3
BigContigLength (bp)	3,082,467	3,111,472	3,110,513

The improvement of de novo assemblies with Pacbio RS



Assembly comparison

Program : mummer 3.0



Summary

General Infomation

Strain No.	Scientific name	GC content (%)	Genome Size (Mb)	HiSeq (Coverage)	PbCR (Coverage)	GSFLX (Coverage)	CDS (Coverage)
PAMC 26508	<i>Streptomyces</i> sp.	70.09	7.6	100X	15X	6X	27X
PAMC 26555	<i>Frigoribacterium</i> sp.	62.87	3.3	100X	26X	20X	
PAMC 21119	<i>Psychrobacter</i> sp.	43.34	3.5	60X	14X	3X	
PAMC 21357	<i>Salinibacterium</i> sp.	60.2	3.1	100X	16X	8X	

Assembly Summary

Strain No.	Scientific name	PAMC 26508	PAMC 26555	PAMC 21119	PAMC 21357
Contigs (EA)	HiSeq	185	93	76	26
	PbCR	26	60	39	7
Max Contig Bases Bases (bp)	PbCR+GSFLX	36	96	67	9
	HiSeq	221,220	366,603	255,324	643,647
N50 Contig Bases Bases (bp)	PbCR	1,915,364	1,164,979	582,027	2,110,572
	PbCR+GSFLX	1,869,455	1,492,435	598,064	2,112,607
Total Big Contigs (10kb <)	HiSeq	68,326	136,524	82,863	233,362
	PbCR	1,268,506	263,283	511,916	2,110,572
Big Contig Length	PbCR+GSFLX	1,434,415	469,317	258,230	2,112,607
	HiSeq	141	38	60	20
	PbCR	16	16	19	3
	PbCR+GSFLX	11	9	9	3
	HiSeq	7,372,844	3,203,637	3,372,293	3,082,467
	PbCR	7,637,132	3,246,689	3,439,229	3,111,472
	PbCR+GSFLX	7,723,210	3,246,808	3,441,479	3,110,513