Novogene

Technical Note:

Enhanced Amplicon Library Preparation with Phased Primers for Increased Complexity

Introduction

16S/18S/ITS Amplicon sequencing is a widely adopted method for identifying species and determining their composition using metabarcoding-specific primers that target particular short sequences within a metagenomic sample. Amplicon libraries inherently lack sequence complexity because they focus on specific genes or regions, resulting in low sequence diversity. This is a major concern for Illumina sequencing platforms, which require high-diversity sequences for optimal performance and low-diversity libraries often suffer from reduced sequencing quality. At Novogene, we address this issue by using phased barcodes to increase the complexity of amplicon libraries, thereby improving sequencing quality on the NovaSeq sequencing platforms.

This technical note highlights how amplicon libraries constructed with phased barcodes are sufficiently complex, thus reducing the need to add PhiX to the sequencing reaction. This technical note also explains how raw sequencing data obtained from this library preparation strategy can be effectively demultiplexed and merged.

Challenges of Low Diversity Libraries

Libraries with high nucleotide polymorphisms diversity contain equal proportions of all four nucleotides (A, C, G, and T) in each sequencing cycle. On the other hand, low-diversity libraries have imbalanced nucleotides proportion within a sequencing cycle. An amplicon library, being base-unbalanced due to in-line barcodes and other molecular identities like amplification primers, will typically have regions of low diversity within an otherwise diverse library.

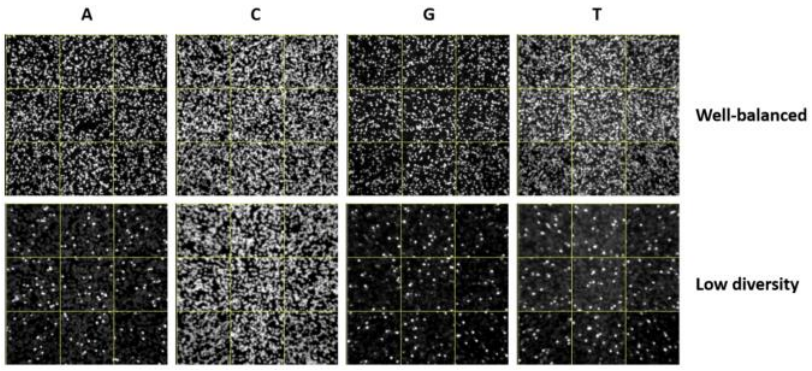

Figure 1: The thumbnail images within a cycle illustrate the relative proportion of nucleotides for a well-balanced library (top panel) and a low-diversity library (bottom panel) for a single tile of the MiSeq sequencing cycle. In the low-diversity library, the signal densities of clusters in channel C are relatively high, while the signal densities in channels A, G, and T are significantly lower.

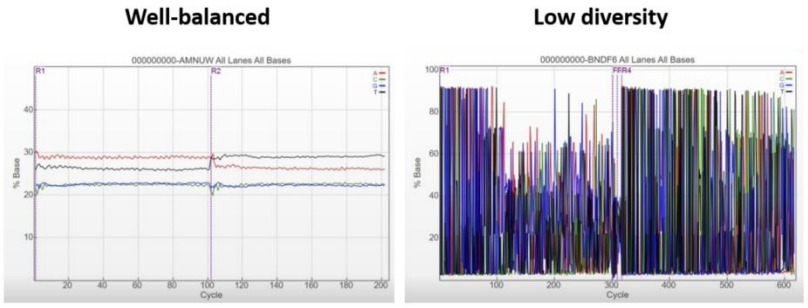

Figure 2. Sequencing results analyzed using Sequencing Analysis Viewer (SAV) software revealed that for well-balanced libraries, the signals for all four nucleotides each accounted for approximately 25% of the total (left). In contrast, poorly balanced libraries exhibited wide variations in the proportions of each nucleotide (right).

Unbalanced libraries lead to uneven fluorescence signals, which adversely impact base calling, quality value calculations, and data output. The NovaSeq platforms use dual-channel (Red-Green 2-Channel) SBS technology, identifying bases in each cycle with two fluorescent dyes and two images. In dual-channel platforms, base balance primarily affects sequencing quality by influencing the following aspects:

1. Cluster Passing Filter
2. Phasing/Pre-phasing
3. Color matrix correction

Cluster Passing Filter

In sequencing-by-synthesis (SBS) platforms those from Illumina, clusters of identical DNA molecules are generated on a flow cell. Each cluster emits a fluorescent signal as nucleotides are added during sequencing. High-diversity libraries contain a wide range of sequences, resulting in distinct and easily separable clusters. In contrast, low-diversity libraries consist of very similar or identical sequences, causing overlapping fluorescent signals from different clusters. This reduced cluster separation makes it difficult for the sequencer to differentiate between clusters, leading to higher error rates.

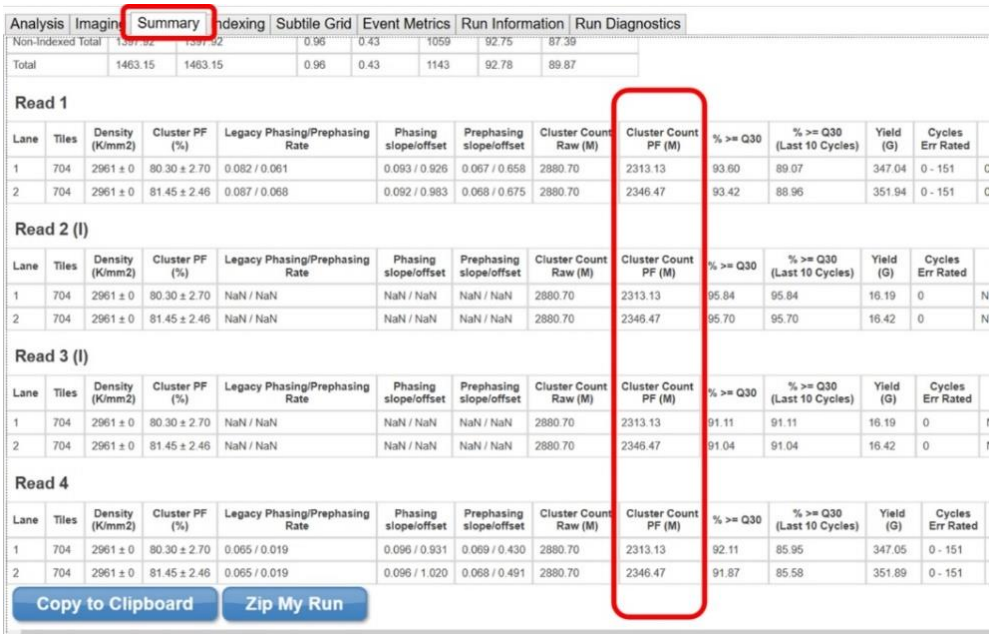

Figure 3. Poor base balance may cause signal quality of some clusters to fail quality filter, making these clusters considered lost clusters, thus reducing the total number of Cluster Count passing filter (PF).

Phasing/Pre-Phasing

Phasing occurs when the synthesis of the new DNA strand lags, causing some nucleotides to be incorporated in later cycles and the shifting of the detected sequence. In contrast, pre-phasing happens when the DNA strand synthesis progresses ahead of schedule, causing nucleotides to be incorporated earlier than expected, making the signal from a cluster appear prematurely. Both phasing and pre-phasing can significantly reduce the accuracy of sequencing low-diversity libraries. They increase noise and error rates, misinterpret signals, and decrease base-calling accuracy, which can result in index hopping and cross-contamination.

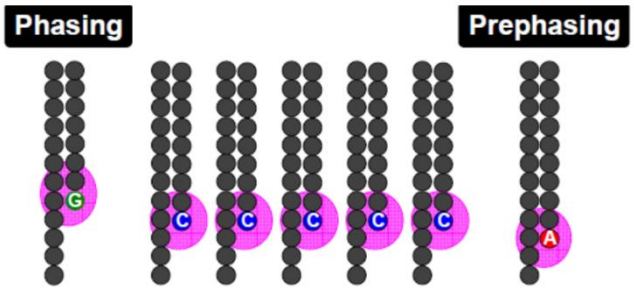

Figure 4. Poor base balance in a library can lead to less efficient amplification of certain bases compared to others in some cycles, resulting in different signal detection efficiencies and inhomogeneous base signals. This increases Phasing and Pre-phasing offsets. When the phasing or pre-phasing percentage in a cluster exceeds 15-20%, it can cause issues in the base calling process, leading to the cluster being filtered out.

Color Matrix Correction

The color matrix converts detected fluorescent signals into base identifiers. In low diversity libraries, imbalances in base composition can cause deviations in signal strength ratios, resulting in an inaccurate color matrix. During sequencing, the color matrix correction maps these signals to specific bases. When the color matrix is inaccurate due to base balance issues, it can impair signal resolution and accurate base identification, leading to erroneous sequencing results and reduced quality and reliability of the data.

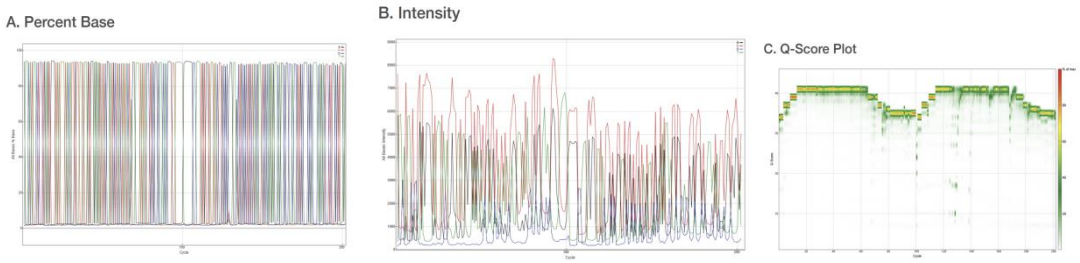

Figure 5. Sequence Analysis Viewer (SAV) data from low-diversity samples, such as amplicon samples, show that nucleotide distribution affects cluster identification and, consequently, the color matrix correction process.

Incorporating Phased Primers to Address Low Diversity Amplicon Libraries

To increase sequence diversity during a sequencing run, a common method is to spike in a high-diversity library, such as PhiX control, into the low-diversity library. While this approach improves overall diversity, it only partially mitigates the negative effects on

sequencing quality. Novogene uses a more effective method which is the incorporation of phased primers to enhance apparent diversity of the library.

Phased primers create small variations at both ends of the sequences, resulting in a staggered alignment of sequencing reads. This staggered alignment helps reduce systematic bias and synchronization issues, enhances error correction, and improves overall sequencing accuracy. Libraries prepared with phased primers exhibit a higher degree of nucleotide diversity, which increases sequencing quality. This increased nucleotide diversity also ensures that phased libraries maintain a very homogeneous read quality distribution, with only a slight decline as sequencing progresses through read 1.

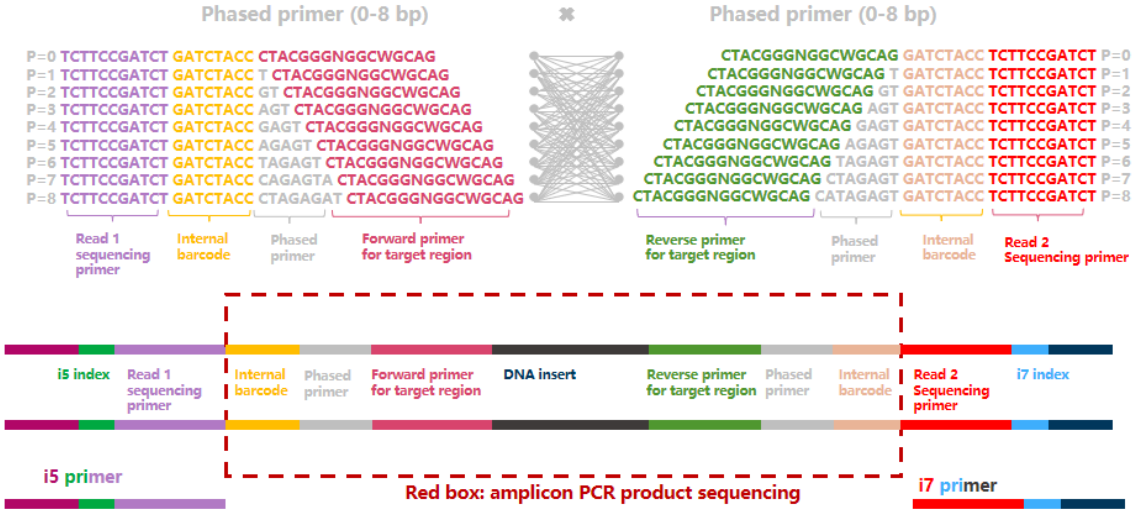

Figure 6. During the PCR amplification stage, 0-8 bases are added to the original 8 bp barcode. The number of bases added is determined individually for each sample and varies between samples. These additional bases are specifically positioned between the internal barcode and the forward/reverse primer.

Adopting PE256 Sequencing Strategy

With the addition of phased primers, the overall length of the amplicon will increase. To determine the maximum tag length achievable under different sequencing strategies, we use the TagLengthCutoff equation:

TagLengthCutoff (tags length) =2×minReadLengthNeed−overlap value−primer-barcode

Where:

- minReadLengthNeed: length required by the sequencing strategy
- LossPercentage: data loss rate

Using this equation, we can see that the PE256 sequencing strategy, compared to the previous PE250 strategy, will result in a lower loss rate despite increased amplicon size due to phased primers. Illumina's NovaSeq sequencing system supports the PE256 sequencing strategy.

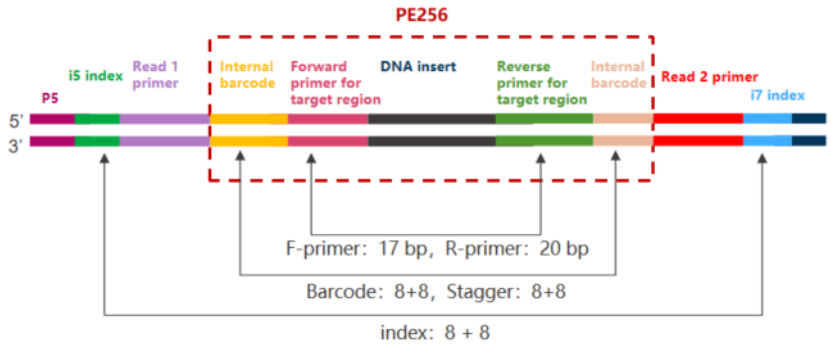

Figure 7. According to the kit manual, the maximum supported number of cycles is 538, which theoretically allows for a sequencing strategy of up to PE258.

The NovaSeq 6000 sequencing system by Illumina is designed for high throughput and flexibility, making it suitable for a variety of applications, from whole-genome sequencing to targeted sequencing, offering a total of 538 cycles. The number of cycles in a sequencing run determines the read length, with each cycle adding a single nucleotide to the growing DNA strand. For example, a run described as "2x250 + 38" means 250 cycles for read 1, 250 cycles for read 2, and an additional 38 cycles for indexing or other purposes.

Appendix: NovaSeq 6000 v1.5 Reagent kits

Flow cell	Reagent kit	Catalog no.
S4	NovaSeq 6000 S4 Reagent Kit v1.5 (300 cycles)	20028312
	NovaSeq 6000 S4 Reagent Kit v1.5 (200 cycles)	20028313
	NovaSeq 6000 S4 Reagent Kit v1.5 (35 cycles)	20044417
S2	NovaSeq 6000 S2 Reagent Kit v1.5 (300 cycles)	20028314
	NovaSeq 6000 S2 Reagent Kit v1.5 (200 cycles)	20028315
	NovaSeq 6000 S2 Reagent Kit v1.5 (100 cycles)	20028316
S1	NovaSeq 6000 S1 Reagent Kit v1.5 (300 cycles)	20028317
	NovaSeq 6000 S1 Reagent Kit v1.5 (200 cycles)	20028318
	NovaSeq 6000 S1 Reagent Kit v1.5 (100 cycles)	20028319
SP	NovaSeq 6000 SP Reagent Kit v1.5 (500 cycles)	20028402
	NovaSeq 6000 SP Reagent Kit v1.5 (300 cycles)	20028400
	NovaSeq 6000 SP Reagent Kit v1.5 (200 cycles)	20040719
	NovaSeq 6000 SP Reagent Kit v1.5 (100 cycles)	20028401
Xp	NovaSeq Xp 4-Lane Kit v1.5	20043131
	NovaSeq Xp 2-Lane Kit v1.5	20043130

Figure 8. NovaSeq 6000 v1.5 Reagent Kit - The NovaSeq 6000 v1.5 Reagent Kit supports up to 500 cycles with an additional 38 cycles of reagents.

Data demultiplexing from phased barcoded amplicon libraries

Sequencing mixed amplicon libraries requires two rounds of demultiplexing. The first round demultiplexes the mixed library from the sequenced lane using the library indexes. The second round demultiplexes each amplicon from the mixed library based on the barcode and phased primer. This step requires demultiplexing both read 1 and read 2 for each amplicon using the corresponding barcode, phased nucleotide, and amplification primer. Novogene's amplicon sequencing service provides raw demultiplexed data and raw sequenced data with the barcodes, phased nucleotides, and amplification primers removed. Each data release includes a detailed list of the barcodes and phased nucleotide combinations.

FKDN240190840-1A	CCATAGCTC,GCCATGAAGGATCA	GCGGTAATTCCAGCTCCAA,AATCCRAGAATTTCACCTCT
FKDN240190839-1A	GCCTATACGTGTC,CCAGGAATAGTG	GCGGTAATTCCAGCTCCAA,AATCCRAGAATTTCACCTCT
FKDN240190855-1A	AGGACCTA,CGTATCACGAGCGAGT	GCGGTAATTCCAGCTCCAA,AATCCRAGAATTTCACCTCT
FKDN240190854-1A	AGGCTACACTCTCT,TACCACTGC	GCGGTAATTCCAGCTCCAA,AATCCRAGAATTTCACCTCT
FKDN240190853-1A	CCATATCG,CTAGAGACAGAGTAGT	GCGGTAATTCCAGCTCCAA,AATCCRAGAATTTCACCTCT
FKDN240190900-1A	TCGCCTAATGTTAG,CACTAAGGC	GCGGTAATTCCAGCTCCAA,AATCCRAGAATTTCACCTCT
FKDN240190898-1A	AGCCACTTCGCGAG,CTAGCGAAC	GCGGTAATTCCAGCTCCAA,AATCCRAGAATTTCACCTCT
EKDO240002721-1A	TACGCATCGT,AGACCTTCTCTAC	GCGGTAATTCCAGCTCCAA,AATCCRAGAATTTCACCTCT

Figure 9. A detailed list of the amplification primer sequences, the 8 bp barcodes, and any phased nucleotides added for each corresponding sample will be provided to customers. This will be included alongside the raw sequenced data and the processed data where primers, barcodes, and phased nucleotides have been removed, all within the data release.

Reads Merging

Reads merging is conducted after demultiplexing to improve sequencing data accuracy and quality. This process merges the forward and reverse reads from paired-end sequencing into a single, continuous sequence. The steps involved include quality filtration, trimming, alignment, reads merging, and error correction.

Phased nucleotides, along with barcodes and amplification primers, are trimmed before alignment. The resulting insert sizes are expected to be consistent across samples. The addition of phased nucleotides adds variation in the overlap length, but it remains within the recommended minimum overlap of 10 bp or less. This ensures high-quality sequences without the need to adjust the current analysis pipeline.

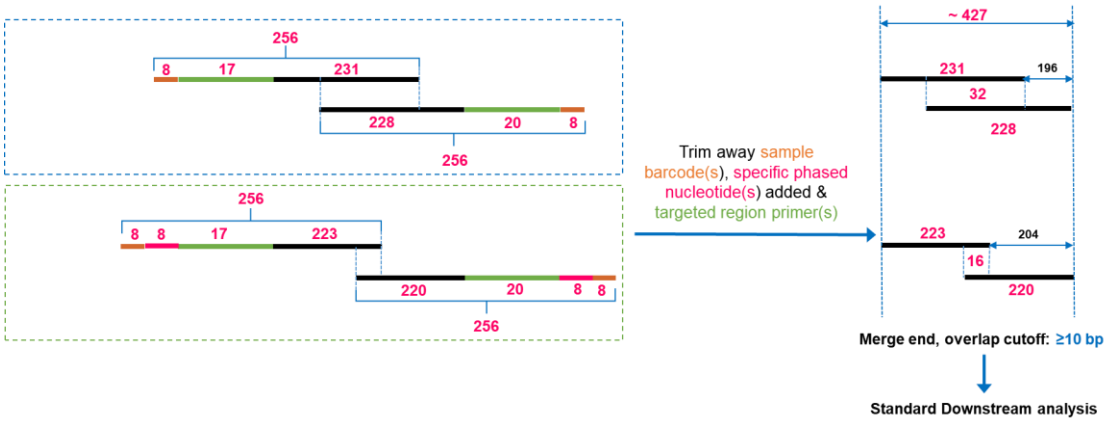

Figure 10. A schematic to illustrate the trimming process before alignment and the reads merging process. The resulting insert size remains constant, but although the overlap region varies between samples, it stays within the recommended overlap cutoff parameters of ≤ 10 bp.

Conclusion

Novogene has adopted an improved amplicon library preparation strategy to tackle compatibility issues when sequencing low complexity libraries like 16S/18S/ITS amplicon libraries on Novaseq sequencers. This strategy increases library complexity with addition of phased primers, which enhances overall sequencing quality, an increase in usable reads for reads merging and downstream analysis, and a decrease in reads trimmed as a result of poor quality. Novogene has also found a reduced need for PhiX spike-ins to increase complexity in each sequencing run.

Novogene's amplicon sequencing data release includes raw sequenced data, data without phased primers and barcodes, merged raw tags, and clean tags after chimera removal. A detailed document outlining the phased nucleotides, barcodes, and amplification primer sequences is also provided. This improved amplicon library preparation strategy with phased primers enables the production of high-quality sequencing data without requiring users to modify their existing analysis pipeline.

References:

1. "General Instrumentation Reference Material List." Illumina, knowledge.illumina.com/instrumentation/general/instrumentation-general-reference_material-list/000006529.
2. General Instrumentation Reference Material List." Illumina, knowledge.illumina.com/instrumentation/general/instrumentation-general-reference_material-list/000001543.
3. How Many Cycles of SBS Chemistry Are in My Kit?" Illumina, support.illumina.com/content/illumina-support/apac/zh_CN/bulletins/2016/10/chinese_how-many-cycles-of-sbs-chemistry-are-in-my-kit-.html?langsel=/us/.
4. "HiSeq Low-Diversity Sequencing." Illumina, www.illumina.com/documents/products/technotes/technote-hiseq-low-diversity.pdf.
5. "Increase Complexity of Amplicon Libraries Using Phased Primers." SciLifeLab, ngisweden.scilifelab.se/2021/01/tech-note-increase-complexity-of-amplicon-libraries-using-phased-primers/.
6. "Introduction to Genomics." Biomedical Hub, biomedicalhub.github.io/genomics/01-part1-introduction.html.
7. "PhiX Control v3: Optimizing Cluster Density and Sequencing Performance." Illumina, www.illumina.com/content/dam/illumina-support/documents/products/technotes/technote_phixcontrolv3.pdf.

www.novogene.com/amea-en/

marketing_amea@novogeneait.sg

July 31, 2024