Webinar Q&A: How library preparation affects sequencing accuracy

Q: How would you limit DNA damage in the typical DNA extraction/library prep?

A: Damage can be limited by performing acoustic shearing of genomic DNA in 1X TE (10 mM Tris pH 8, 1 mM EDTA), rather than less buffered solutions. Treating the genomic DNA with the NEBNext FFPE DNA Repair Mix after shearing, but before library preparation, will also reduce DNA damage prior to sequencing. More information about DNA damage and repair mixes can be found on our website.

Additional reference: Costello, M., Pugh, T. J., Fennell, T. J., Stewart, C., Lichtenstein, L., Meldrim, J. C., et al. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research*, 41(6), e67–e67.

Q: Can you talk a little more about PreCR repair? How does it work?

A: The PreCR Repair Mix and the NEBNext FFPE DNA Repair Mix are enzyme cocktails that recognize and repair DNA damage. Both DNA repair mixes contain DNA repair enzymes that recognize and remove damaged bases; then a DNA polymerase and ligase will replace the excised base and seal the resulting nick. Multiple repair enzymes recognize specific types of DNA damage typically resulting from oxidation, hydrolysis, UV irradiation and shearing. The ligase present in the mixture will not ligate blunt DNA ends, nor nicks near a mismatch, thereby avoiding ligation chimeras. We have also shown that DNA repair reduces false positive mutations significantly, but does not change variant frequency (indicating that DNA repair does not introduce new mutations). Both the PreCR Repair Mix and the NEBNext FFPE DNA Repair Mix repair the same spectrum of DNA damage, but the NEBNext FFPE DNA Repair Mix has been optimized and validated for use in next-generation sequencing workflows.

Q: Would you recommend DNA damage repair before Illumina sequencing as a normal part of a protocol?

A: For routine sequencing (fresh samples, high input and high sequencing coverage), DNA damage repair is probably not necessary. However, for low input samples, archived clinical samples (especially DNA extracted from FFPE samples), or for detecting low-frequency variation in heterogeneous populations, then DNA damage repair can help increase library yields, sequencing quality and reduce false positive mutation calls.

Q: With Illumina sequencing and NEBNext, what percentage of errors (on average) originate from sample prep? How does the number of PCR-mediated mutations introduced compare to the number of mutations introduced during the actual sequencing?

A: For most applications, Illumina sequencing errors from single-pass reads will be more abundant than the errors originating from DNA damage or PCR. However, for examining low-frequency variation (for example, tumor heterogeneity), errors originating from DNA damage or PCR can look very similar to low-frequency variation, making it more difficult to identify true allelic variation.

Q: How do you correct for substitution mutations in your library? What would you want to consider if you are looking for somatic mutations in your sample?

A: Higher coverage can help distinguish substitution errors from true somatic variation, so obtaining greater sequencing depth may help. True somatic variation is more likely to appear in multiple reads, while sequencing errors and polymerase substitutions are more likely to be randomly distributed. For removing substitution and sequencing errors completely, different library preparation methods are required. PCR-free library preparation methods are available, as well as methods that utilize Unique Molecular Identifiers (see references below).

For detecting somatic variation, it is also recommended to be aware of DNA damage that can be introduced during library preparation (especially during shearing). To reduce oxidative damage during acoustic shearing, sonication in a buffered solution (1X TE, pH 8) is recommended. In addition, DNA repair after shearing, but before library preparation, will repair oxidative and other types of DNA damage. Additionally, calculating the GIV score can be used for quality control analysis.

(1) Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W., & Vogelstein, B. (2011). Detection and quantification of rare mutations with massively parallel sequencing. PNAS U.S.A., 108(23), 9530-9535.

(2) Schmitt, M. W., Kennedy, S. R., Salk, J. J., Fox, E. J., Hiatt, J. B., & Loeb, L. A. (2012). Detection of ultra-rare mutations by next-generation sequencing. PNAS U.S.A., 109(36), 14508–14513.

Q: In other words, if you are limited by the errors originating from library prep, is it worth improving sequencing accuracy at this point?

A: Yes, there is still a need to improve sequencing accuracy to allow greater sensitivity for detecting rare mutations. Technology development to improve sequencing accuracy remains an active area of research.

Q: Could you explain the GIV score a bit more?

A: DNA damage only affects one nucleotide of a pair, leading to an imbalance of one particular type of mutation. Thus, DNA damage results in a systematic and global excess of particular mutations (for example G to T) and a specific signature: an excess of one mutation in the first paired read (R1), and a corresponding excess of the reverse complement of the mutation in the second paired read (R2). To estimate the extent of DNA damage in an Illumina data set, we compute a global imbalance value (GIV) score. A GIV score greater than 1 indicates DNA damage, while undamaged DNA will have a GIV score of 1.

More specifically, the analysis strategy consists of deconvoluting both the origin and orientation of variants and computing a global imbalance value (GIV). The GIV score is determined using the following equation:

GIV_{G_T} = ((C1v + C2v) / (C1 + C2)) / ((C1v_RC + C2v_RC)/(C1_RC +C2_RC))

continued on page 2

With C1v = Number of G to T variants in R1 (read 1); C1 = Total number of G in R1; C2v = Number of C to A variants in R2 (read 2); C2 = Total number of C in R2; C1v_RC = Number of C to A variants in R1; C1_RC = Total number of C in R1; C2v_RC = Number of G to T variants in R2; C2_RC = Total number of G in R2. For more information see Chen, L., et. al., Science (2017), 355(6326):752-756.

The source code to calculate the GIV score is available at here.

Q: Can you use a GIV score to remove low frequency variants caused by DNA damage? If so does it work with downstream analysis.

A: The current algorithm will not support the removal of variants caused by DNA damage. The GIV score is a global measure of the imbalance in the dataset, and is indicative of the extent of DNA damage, but cannot pinpoint which particular mutations originate from damage.

Q: So the best route to minimize damage is to use a high fidelity polymerase and DNA repair enzymes?

A: Yes, using the NEBNext FFPE DNA Repair mix to repair damage, and a high-fidelity DNA polymerase (like the NEBNext Ultra II Q5 Master Mix), will reduce sequencing errors due to DNA damage and PCR polymerase mistakes. Reducing the number of PCR cycles will also help.

Q: What are the sequencing accuracy (e.g. SNP calling) implications of cross over events during PCR?

A: The implications of template-switching in PCR is most applicable to NGS assays that rely on amplifying and sequencing very similar targets. Amplicon sequencing applications such as microbial identification by 16S rRNA, HLA genotyping, or viral population studies can be affected by PCR-mediated recombination artifacts. PCR-based multiplex target enrichment strategies could also be affected. Calling individual SNPs in distinct genomic sequences probably won't be affected.

Q: Why did you only do the tests with 16 cycles and not 30–40 cycles as most PCRs?

A: We used 16 cycles of PCR in our assays to ensure that the amplification is still in the linear range (meaning that the replication efficiency per cycle is constant throughout amplification). We could then determine the number of template doublings that occurred based on the input DNA amount and PCR yield, and then normalize the raw error rate (errors counted after PCR) to the number of doubling events for each polymerase. This allows us to determine the errors per replication event, which makes it easier to compare different polymerases (regardless of the replication efficiency of each enzyme) and compare our results to previous studies.

Q: You used PacBio sequencing to analyze the types of error. How would one analyze errors from the PacBio process?

A: To determine the background error rate for our PacBio-based fidelity assay, we sequenced plasmid DNA (that had been repaired using PreCR to remove background DNA damage). Based on the typical sequencing output and the *in vivo* error rate of plasmid replication, we expect plasmid DNA to have an undetectable error rate in our assay. Any errors detected in the plasmid libraries were attributed to PacBio sequencing and library preparation. The background error rate for the fidelity assay was determined to be 9.6×10^{-8} substitutions per base, but higher for insertions/deletions (3.1×10^{-6} per base). Note, our assay is for single-stranded consensus reads (top and bottom strand separately). The standard PacBio sequence analysis tools analyze duplex reads (top and bottom strand combined), which is more accurate. For example, duplex reads will compensate for sequencing artifacts resulting from DNA damage.

Q: Is there a preferential genomic content for *Taq* errors?

A: *Taq* DNA polymerase prefers to make mistakes at A's and T's, so A/T rich templates would produce more errors than G/C rich templates. In contrast, the high-fidelity polymerases (related to Family B) generally make more mistakes at G's and C's, so G/C rich templates will be relatively more error-prone than A/T rich templates (though the overall mutation rate will be reduced compared to *Taq*).

Q: Why is Q5 able to maintain an extremely low error rate while thermal cycling seems to introduce significant DNA damage?

A: Q5 (and archaeal Family B DNA polymerases in general) are unable to replicate past deoxyuridine (dU) in a template strand. As cytosine deamination (conversion to dU) was found to be the major mutagenic DNA damage observed during thermocycling, the polymerase is preferentially replicating undamaged templates, and this may partly account for the observed low error rate.

Q: Is there a way to reduce or avoid primer dimers during amplification? And what method can be used to remove primer dimers?

A: Some primer sets seem to be more prone to forming dimers than others; however it is unclear what causes some primers to be more susceptible to dimerization. One possible experiment is to try to design new primer pairs using online primer design or other sequence analysis tools, and testing several primer sets to see if any produce less primer dimers. For next-generation sequencing library prep, adaptor dimers can be removed by using a size selection protocol with SPRISelect® or AMPure® beads after PCR.

Q: How many recombination event/kb can we expect?

A: We measured the rate of recombination to be on average 1×10^{-4} per base for *Taq* polymerase, which corresponds to once every 10 kb replicated. However, as PCR errors get copied in later cycles, the actual number of recombinants in the final product will be much higher (depending on the number of cycles and replication efficiency of the target). There is also evidence in the literature that certain sequence contexts can also promote recombination, so this rate will likely be sequence-dependent.

continued on page 3

Q: Is there a preferential genomic content for cross-overs mediated by *Taq* polymerase?

A: In our study, we identified specific inverted repeat elements (in the lacZ gene) that promote template-switching (but between the top- and bottom- strand), suggesting that template-switching can be induced by specific sequence contexts. For the recombination assay described in this webinar (template-switching between different strands), the templates were designed to be random sequences and not prone to secondary structure, and we could not see any obvious sequence bias for recombination events.

Q: Is there a way to reduce chimeric events (for long range PCRs in particular)? Is there any way to minimize PCR mediated recombination in your samples?

A: We want to further study this issue. Reducing the number of cycles is the easiest way to reduce the number of recombinants in the final PCR product.

Q: Was cytosine deamination in any way, shape, or form affected by methylation or genomic CpG content?

A: We did not examine the effect of methylation or CpG content on cytosine deamination.

Q: For a diverse multiplex PCR library, how much of a factor do differences in priming efficiency affect final biases?

A: For multiplex amplification reactions, with multiple primer sets amplifying multiple targets, differences in priming efficiency has been shown to introduce bias. Primer pairs that are less effective at priming may result in that amplicon being artificially underrepresented in the final product. Amplification bias can also affect the final representation of each target. Amplification biases can be reduced by increasing the template concentration and performing fewer PCR cycles.

Q: Does the rate of heating and cooling during the PCR reaction affect DNA damage?

A: We did not study DNA damage for different temperature cycling protocols, however, it is likely that longer denaturation times increase damage. Slower ramp times may increase the amount of time at elevated temperatures, and may influence the amount of DNA damage. More likely, the total number of PCR cycles will have a larger effect.

Q: Do you believe most errors in the PCR experiments come from damage to the bases in the template or growing strand or from damage to the nucleotide pool? Does the fact that thermal cycling itself causes damage suggest it is damage in the template? Maybe different types of damage in the template vs the nucleotide pool?

A: Thermal cycling can damage dNTP pools as well as DNA templates. In our study, cytosine deamination was the major cause of mutagenic DNA damage to template DNA. In a separate unpublished study, we found that cytosine deamination was also the major damage to dNTP pools.

Q: Do you think repair of the amplified libraries (to correct for DNA damage arising from thermocycling) should be added to library prep workflows?

A: For applications where DNA repair is recommended (low frequency variant-calling, archived clinical samples, or low input libraries), repairing genomic DNA prior to library preparation is recommended (rather than repairing after library preparation and PCR). Unrepaired DNA damage (especially cytosine deamination and guanine oxidation) will be captured as mutations during PCR, and the resulting mutations won't be recognized as damaged bases subject to DNA repair. Regarding a possible second DNA repair step after PCR and thermocycling, we have not looked at whether DNA damage from thermocycling affects variant-calling. My guess is that PCR thermocycling damage will have less of an effect on variant-calling than damage from shearing, as thermocycling damage will be distributed among the copies of each sequence, whereas mutagenic damage from shearing will be replicated in all subsequent copies.

Q: Would lowering the ramping rate (i.e. 2°C /sec) prevent amplification bias?

A: We haven't explicitly studied the effect of ramp rate on amplification bias so it's difficult to answer. We do know that adjusting the extension temperature can have a significant impact on bias so it's possible that adjusting the time spent getting to temperature could also impact overall bias profiles.

Q: Would the PCR errors, for example the ones caused by heating and cooling during thermocycling, affect Sanger sequencing as it does NGS?

A: DNA damage during thermocycling likely does not affect Sanger sequences as much as next-generation sequencing. If the PCR product is directly sequenced, then the sequencing read will be the consensus of the entire population, and all of the individual, randomly-distributed errors (from substitutions or damage) would not appear in the sequencing read. If the PCR product was cloned into a vector and sequenced from bacterial colonies, then the bacterial repair pathways would correct DNA damage before sequencing.

New England Biolabs, Inc., 240 County Road, Ipswich, MA 01938-2723 Telephone: (978) 927-5054 Toll Free: (USA Orders) 1-800-632-5227 (USA Tech) 1-800-632-7799 Fax: (978) 921-1350 email: info@neb.com