PCR Amplification during Library Preparation Confounds Variant Discovery in Whole Genome Sequencing

Maxwell A. Sherman[1], Tiziana Sanavia[1], Daniel Kwon[1], Rachel Rodin[2], Christopher Walsh[2], Peter Park[1]

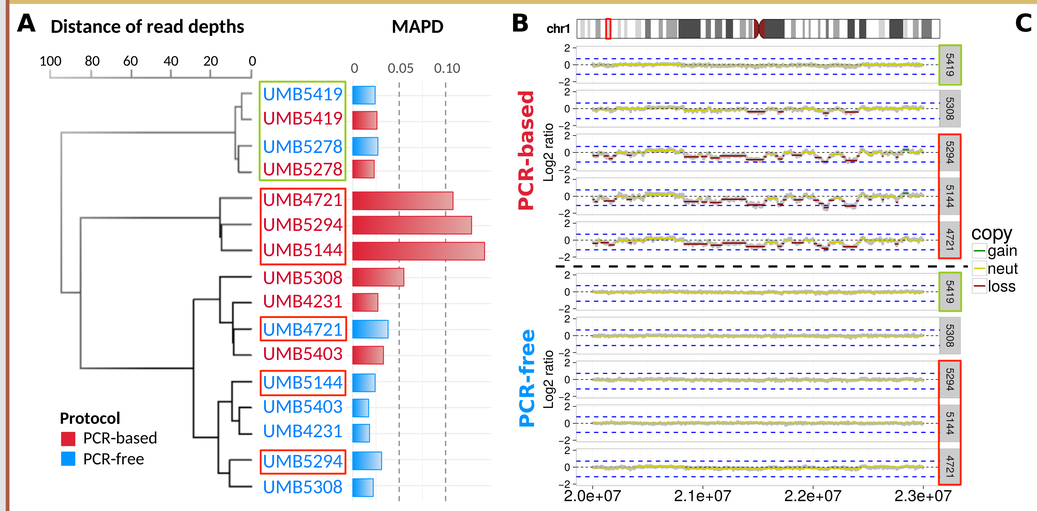
[1]Harvard Medical School, Department of Biomedical Informatics
[2]Boston Childen's Hospital, Departments of Neurology and Pediatrics

Introduction

The vast majority of publicly available whole genome sequencing data were prepared using PCR amplification during library preparation. Here we directly compare **PCR-based** to **PCR-free** libraries from the same samples and find that PCR-based preparations have the potential to:

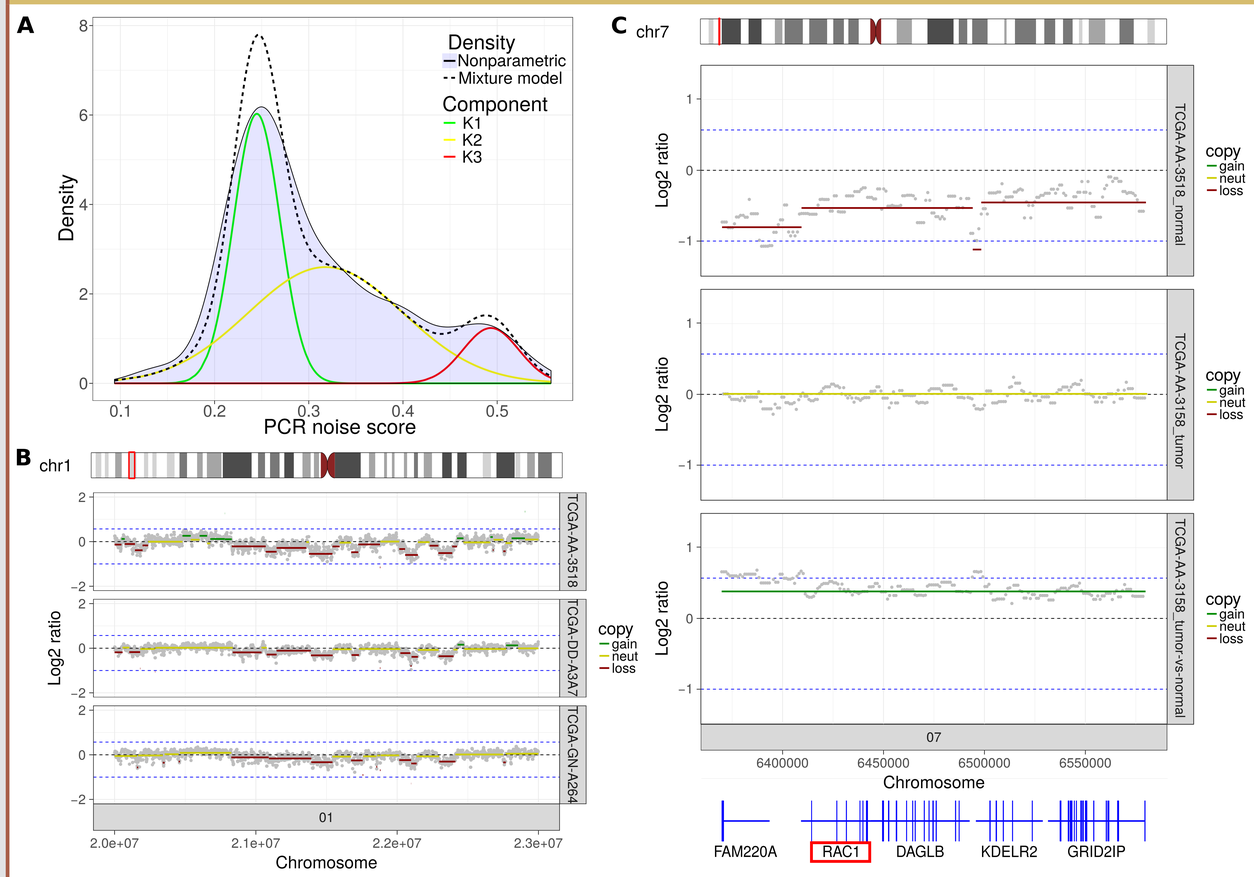
1. Induce systematic non-biological read depth variation, possibly leading to false-positive copy number alteration calls.
   - Analysis of 697 high coverage WGS normal tissue samples from TCGA revealed the PCR noise pattern in at least 10% samples.
   - The noise pattern enriches read-depth CNV calls across thousands of genomic loci, covering hundreds of reported disease-associated genes.
2. Mask thousands of true-positive SNVs while introducing thousands of false-positive SNVs.
   - SNVs called only in PCR-based libraries are characterized by a distinct mutational signature.

PCR Amplification Creates an Artificial Read Depth Pattern



A. Libraries display PCR-specific read depth patterns, with high noise PCR-based samples clustering together.
B. Example 3 MB region on chr1 illustrating the PCR read depth noise pattern.
C. High noise PCR-based libraries share >50% of CNV calls between libraries. Corresponding PCR-free replicates share <9% of CNV calls.

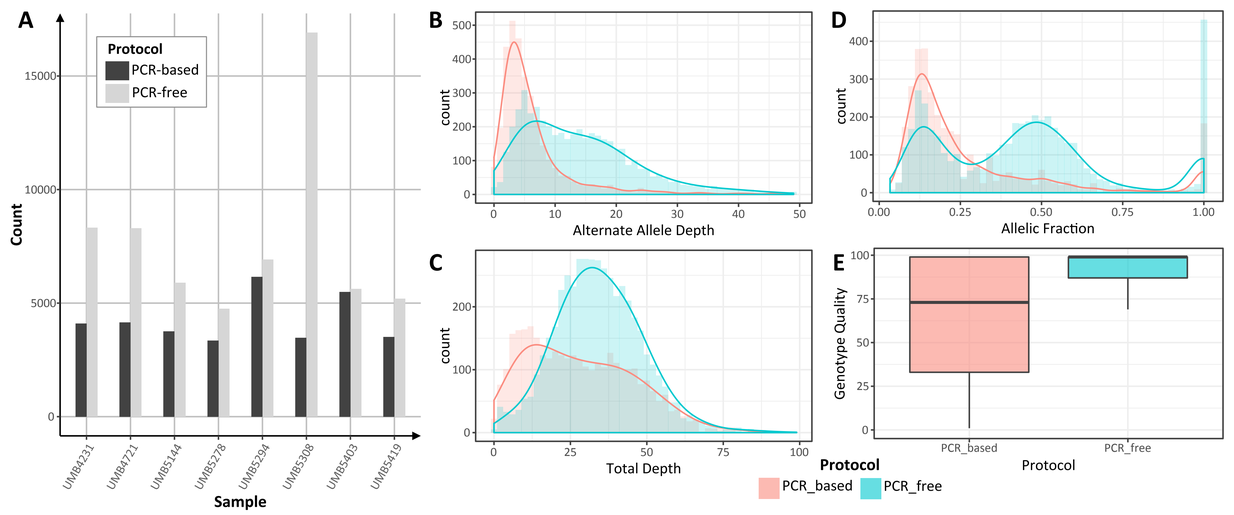
PCR Read Depth Noise Pattern is Present in TCGA Samples



A. Three-component Gaussian mixture model used to detect the PCR noise pattern in 697 normal TCGA tissue samples.
- 73 (10.5%) of samples belong to the most extreme component, indicating presence of the PCR read depth signature.

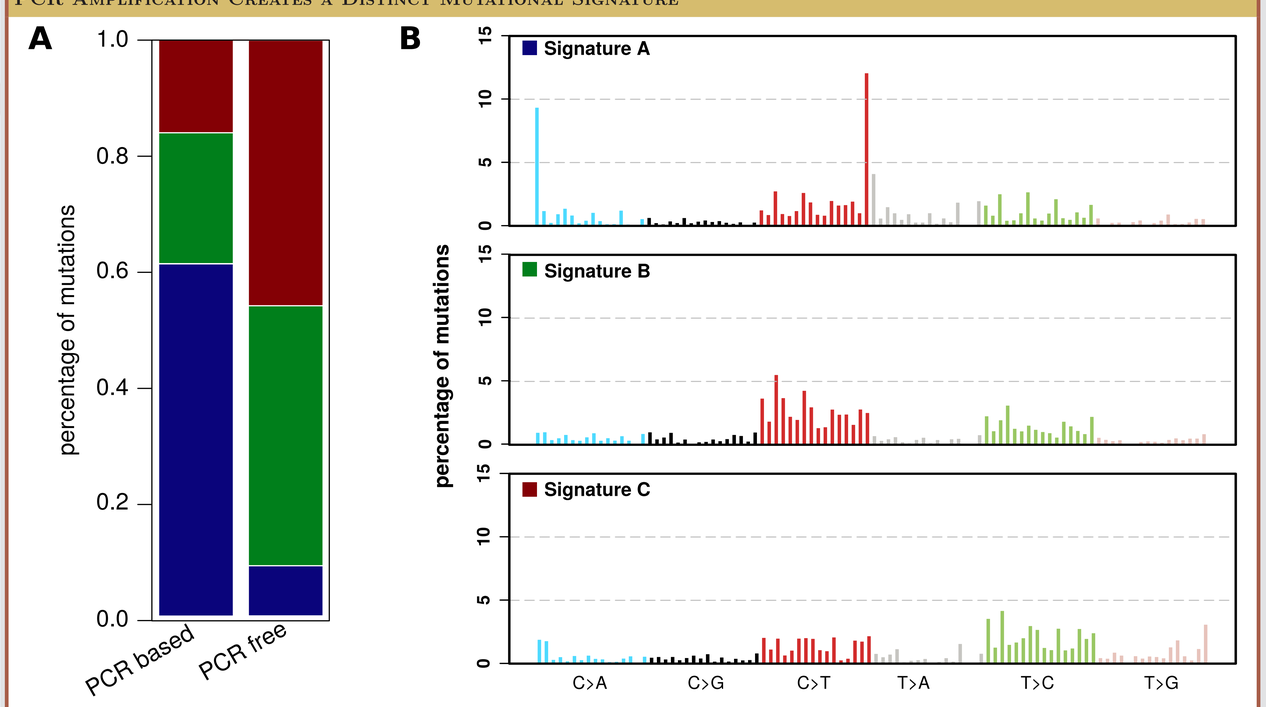
B. The 3 MB example region of three representative high noise TCGA samples exhibits a similar pattern as the samples in the above panel.
C. The PCR noise pattern can induce false-positive CNV calls when performing tumor-normal read depth comparison calling.
- 7389 genes (32% of human genes) are enriched in high noise samples compared to noise-free samples.
- Includes 55% of genes in COSMIC, 58% of genes in SFARI, 49% of genes in AutWorks, and 44% of genes for intellectual disability in the HPO.

PCR Amplification Masks True-Positive and Induces False-Positive SNV Calls



A. SNV calls private to PCR-based and PCR-free libraries from the same tissue. One average 7742 ± 3938 calls are unique to PCR-free replicates and 4251 ± 1028 calls are unique to PCR-based replicates.
B. Alternate allele depth of private SNV calls reveals an excess of PCR-based calls with few alternate reads.
C. Total depth at the private SNV call sites shows PCR-based calls have skewed coverage compared to PCR-free calls.
D. Allele fraction of private SNV calls reveals an excess of PCR-based calls with very low allele fraction, indicative of false-positive calls.
E. Overall genotype quality is lower in PCR-based calls than PCR-free calls ($p < 3e\text{-}16$, Wilcoxon rank-sum test).
- Combined evidence from **B-E** suggests that PCR amplification potentially misses true-positive calls and induces false-positive calls.
- We hypothesize this is due to amplification imbalance and PCR polymerase error during the amplification process.

PCR Amplification Creates a Distinct Mutational Signature



A. Three mutational signatures contribute to private SNV calls. PCR-based calls are dominated by Signature A while PCR-free calls are mostly Signature B and C.
B. Signatures A, B and C.
- **Signature A** is dominated by AC>AA and TC>TT mutations; this feature is distinct from all COSMIC signatures.
- **Signature B** is similar to COSMIC signature 1, which is common in cancer samples.
- **Signature C** is similar to COSMIC signature 5, which is common to most samples.

Conclusions

1. PCR Amplification during library preparation can introduce an artificial read depth variation.
   - The signature is found in at least 10% of TCGA WGS normal tissue samples.
   - The signature enriches for thousands of genes, many of which are associated with disease states including cancer, Autism, and intellectual disability.
2. PCR masks thousands of true-positive SNVs and induces thousands of false-positive SNVs.
   - Masking of true variants may occur due to amplification imbalance.
   - We suspect the introduction of false-positive SNVs occurs due to PCR polymerase error.
3. PCR amplification exhibits a unique mutational signature.
4. WGS data should be checked for the presence of the PCR amplification read depth signature prior to performing read-depth based CNV calling.
5. Care should be taken in curating SNV calls from PCR amplified data.

Methods

- **Aligner**: BWA-mem to GRCh37d5
- **CNV caller**: BICseq2 / CNVnator
- **SNV caller**: GATK HaplotypeCaller

Acknowledgements

We thank Alison Barton and Carl Vitzhum for help aligning samples.

Joe Luquette and Giorgio Melloni for valuable guidance in the analysis of SNVs.

Research funded by a NIH (1U01MH106883) and the Harvard Ludwig Center.