Aird et al. Genome Biology 2010, 11(Suppl 1):P3
http://genomebiology.com/2010/11/S1/P3

Genome Biology

POSTER PRESENTATION Open Access

Analyzing and minimizing bias in Illumina sequencing libraries

Daniel Aird*, Wei-Shen Chen, Michael Ross, Kristen Connolly, Jim Meldrim, Carsten Russ, Sheila Fisher, David Jaffe, Chad Nusbaum, Andreas Gnirke

From Beyond the Genome: The true gene count, human evolution and disease genomics
Boston, MA, USA. 11-13 October 2010

Although Illumina shot-gun reads cover most genomes almost completely, sequences with extreme base compositions are often underrepresented or missing. Bias can potentially be introduced at any step during the library construction in the lab, on the Illumina instrument, in data processing or at the sequence analysis stage. Here we set out to evaluate sources of bias and ameliorate the effects.

To dissect the library construction process, we developed a panel of qPCR assays for loci ranging from 6% to 90% GC that work well in a pool of three microbial DNA samples of different base composition: *Plasmodium falciparum* (19% GC), *Escherichia coli* (51% GC) and *Rhodobacter sphaeroides* (69% GC). We also developed qPCR assays for loci in the human genome that represent four categories of underrepresented sequence motifs as well as GC-rich promoters known to be underrepresented or missing in 'whole' genome sequencing data sets.

We tracked the relative abundance of these loci throughout the standard Illumina library protocol and saw no significant introduction of bias in the initial steps including shearing, end repair, adaptor ligation and size selection. However, GC-rich and extremely GC-poor sequences were depleted during the subsequent PCR-enrichment step. Using qPCR as a readout, we tested different PCR enzymes, the addition of betaine and/or DMSO, and thermocycling profile variations. The choice of PCR instrument itself and the ramp rate had a significant effect on the GC profile of the PCR product, especially when using the recommended amplification conditions (Phusion HF and 10s denaturation per cycle).

Our optimized conditions produce PCR-amplified libraries that display little systematic bias between 15% and 80% GC that resulted during sample preparation. We saw significantly improved representation of challenging human sequence motifs both in the PCR-amplified library (qPCR assay) and in the final Illumina reads. Our conditions are also more reliable and robust because they minimize the effect of PCR instrument and ramp rate. These conditions are currently being implemented in the Sequencing Platform at the Broad Institute. Finally, we still observe some bias in the sequencing readout, which is introduced by steps subsequent to sample preparation, including cluster generation and sequencing. These sources of bias are the object of ongoing investigations.

Published: 11 October 2010

doi:10.1186/gb-2010-11-S1-P3
Cite this article as: Aird *et al.*: Analyzing and minimizing bias in Illumina sequencing libraries. *Genome Biology* 2010 **11**(Suppl 1):P3.

Broad Institute, Genome Sequencing, Cambridge, MA 02141, USA