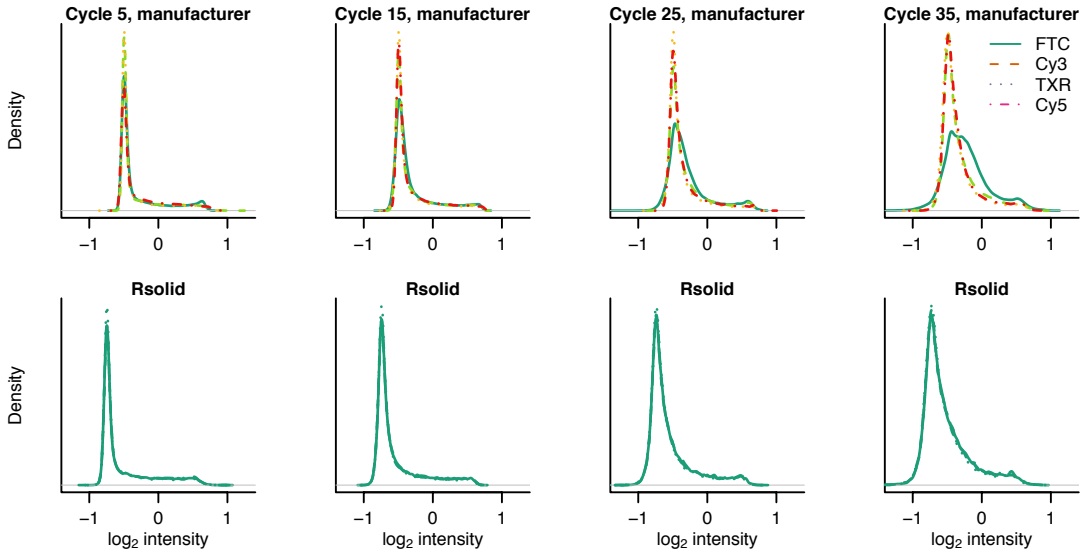
Intensity normalization improves color calling in SOLiD sequencing

Hao Wu, Rafael A. Irizarry and Héctor Corrada Bravo
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD USA 21205

Supplementary Figures



Supplementary Figure 1. Fluorescence intensity measurement distributions for each color in four sequencing cycles. Before normalization, the distribution of intensities in the FTC channel are skewed towards a higher range compared to the other channels in later cycles, resulting in the FTC bias seen in (**Fig. 1a**).

Supplementary Tables

Supplementary Table 1. Mapping statistics for two *E. coli* genomic DNA samples before and after quantile normalization. Sample 1 are 50bp reads, sample 2 are 36bp reads. Observe there is an increase in the total number of mapped reads. Most striking is the increase in perfect matches, indicating a higher accuracy in the color calls after normalization. These samples were processed on two different labs, with independent library preparations and sequencing machines.

		Before QN	After QN	% change
Sample 1	Total Mapped Reads	660850	710226	7.47
1826966	0 mismatches	246542	281590	14.22
total reads	1 mismatch	169708	180460	6.34
	2 mismatches	134467	138811	3.23
	3 mismatches	110133	109365	-0.70
Sample 2	Total Mapped Reads	14090775	14985313	6.35
30296640	0 mismatches	5490005	6202116	12.97
total reads	1 mismatch	3511552	3679413	4.78
	2 mismatches	2794532	2829559	1.25
	3 mismatches	2294686	2274225	-0.89

Supplementary Table 2. Mapping statistics for two *H. sapiens* genomic DNA 36-bp samples from the One Thousand Genomes Project. Each sample was processed in different sequencing centers, thus independent library preparation and sequencing machines. In general, accuracy improved more in Sample 2. Anecdotally, although biases towards the end of reads are found in both samples, the color balance in early cycles of Sample 2 correspond closer to the color balance expected from the human genome. That is, our method is more successful in removing bias in later cycles when the quality of the earlier cycles is good. In other words, good library preparation leads to better bias-correction.

		Before QN	After QN	% change
Sample 1	Total Mapped Reads	493683	515566	4.43
1392626	0 mismatches	165546	177487	7.21
total reads	1 mismatch	125067	131213	4.91
	2 mismatches	109794	112039	2.04
	3 mismatches	93276	94827	1.66
Sample 2	Total Mapped Reads	700605	732433	4.54
2085077	0 mismatches	188330	205331	9.03
total reads	1 mismatch	178920	191088	6.80
	2 mismatches	173863	178733	2.80
	3 mismatches	159492	157281	-1.39

Supplementary Table 3. Effect of normalization on mapped reads from *E. coli* sample. Each row corresponds to reads in each mapping strata (0 mismatches, 1 mismatch, ...) before normalization, columns indicate the percentage of reads in each mapping strata after normalization. For instance, all perfectly mapped reads remain mapped after normalization (89.5% still perfectly mapped) while 25.9% of reads mapped with one mismatch become perfect matches and 7.8% of unmapped reads before normalization can be mapped after normalization.

before/after	0	1	2	3	>=4
0	89.5	9.1	1.1	0.2	0.0
1	25.9	58.5	12.8	2.4	0.5
2	8.1	26.6	47.7	13.6	4.0
3	2.9	10.7	26.6	41.1	18.7
>=4	0.2	0.8	2.1	4.7	92.2

Supplementary Table 4. Effect of normalization on accuracy in *E. coli* sample. More reads map uniquely after normalization, there are fewer errors per mapped read, and, more importantly, there is a ~6.4% reduction in the number of valid adjacent errors. The latter results in a lower rate of false-positive SNP calls (Fig. 1d of main text).

	Before Normalization	After Normalization	% change
Uniquely mapped reads	12969144	13794868	+6.36
Colors w/ errors per mapped read	1.05	0.99	-4.92
Valid adjacent color errors per mapped read	0.027	0.025	-6.42

Supplementary Methods

Quantile Normalization*:* The Rsolid normalization procedure assumes that the distribution f_{jc} of intensity measurements for cycle j and color channel c is the mixture of two distributions:

$$f_{jc} = (1-p_c) f_{0j} + p_c f_{1j},$$

where p_c is the proportion of dinucleotides corresponding to color c in the sample being sequenced, while f_{0j} and f_{1j} are background and signal distributions for the corresponding sequencing cycle. The mixture model of signal and background distributions is motivated by Fig. 1b in the main text, and is the principle behind existing model-based base-calling methods for second-generation sequencing data[1]. Accordingly, the main assumption in our model is that the signal and background intensity distributions are the same for all color channels, while the different mixture proportions make the resulting distributions different.

Our quantile normalization method follows four steps:

- ***Estimate sample color proportions****:* for this step we use the fact that early sequencing cycles are usually good and use these to compute sample color proportions p_c as the proportion of times color channel c has highest intensity over all measurements in sequencing cycles 3-5. If a practical estimate of dinucleotide proportions can be derived from sequence analysis, our method can use that estimate instead.
- ***Estimate background and signal distributions for each cycle****:* in each cycle j, we take the highest intensity for each read as coming from the signal distribution f_{1j}, and the other three measurements as coming from the background distribution f_{0j}.
- ***Compute reference distribution for each channel and cycle****:* for each channel c and cycle j we sample from the estimated signal and background distributions f_{1j}

and f_{0j} according to estimated proportion p_c, this results in a reference distribution f_{jc}.

- ***Quantile normalize****:* we use a standard quantile normalization procedure[2] to normalize the observed intensities in each channel to the reference distributions obtained in step 3.

This gives a new set of intensities from which color-calls are made by selecting the highest of the four intensities for each read and cycle. This procedure is computationally efficient. The most expensive step is sorting the observed intensity measurements in each channel and cycle, and the reference distributions created for normalization. In our tests, we are able to process a complete flow cell of data (~300 million reads) in two hours using ten commodity computing cluster nodes. Intensity measurements for each panel are normalized together, since intensity distributions are different across panels. For data from recently released Solid systems (3plus system), which uses smaller beads and therefore produces more intensity data, our algorithm scales as expected and we are able to process a complete flow cell of data in 4 hours. Note that our software scales at the same rate as a sorting algorithm. Also note that the amount of time consumed by our software is trivial compared to the computational requirements of mapping these data.

Materials*:* Software implementing the quantile normalization procedure is available online at http://rafalab.jhsph.org/Rsolid. The *H. sapiens* samples from the 1000 genomes project are available online from the Short Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra), with accession numbers SRR010750 (Sample 1) and SRR011318 (Sample 2). The *H. sapiens* samples were mapped to the UCSC hg18 reference using Corona Lite software (http://solidsoftwaretools.com/gf/project/corona/), allowing 3 or fewer mismatches and no trimming. The *E. coli* samples were mapped to the DH10B, NCBI Reference Sequence NC_010473 using SOCS[3]. SNP calls in Fig. 1c were made using Corona for both mapping and SNP calling.

Acknowledgements: This work was partially supported by NIH grants P41HG004059 and R01HG005220. We thank Sarah Wheelan and Srinivasan Yegnasubramanian and Gil McVean and Paul Flicek for assistance in obtaining the 1000 genomes project data.

References

1. Bravo, H.C. and R.A. Irizarry, *Model-Based Quality Assessment and Base-Calling for Second-Generation Sequencing Data.* Biometrics, 2009.
2. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.
3. Ondov, B.D., et al., *Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications.* Bioinformatics, 2008. **24**(23): p. 2776-7.