Data release 1 - let7a

Quantum Biosystems

January 23, 2014

1 Summary

This document describes the basecalling of a sample containing the following DNA `TGAGGTAGTAGGTTGTATAGTT` sample. This data release shows shows the capability of our platform to sequence single molecules of DNA with almost no substitution errors, and a limited number of homopolymer errors. It marks the beginning of our data release programme where we invite researchers to engage in the validation and development of our platform.

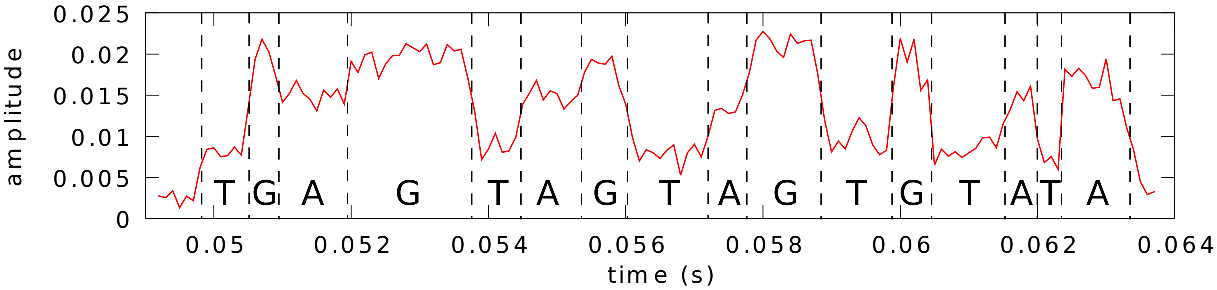

Figure 1: A 16-base read with homopolymers collapsed.

2 Basecalling

Base calling has been performed using a simple hidden Markov model (HMM) as shown in Figure 1. Each base is represented with one state, plus an additional state for the baseline read level. Transition probabilities were uniformly set to 0.001 between all states to give a most likely dwell time of around 10-20 samples. State emissions were modelled with single Gaussians over the sample amplitude.

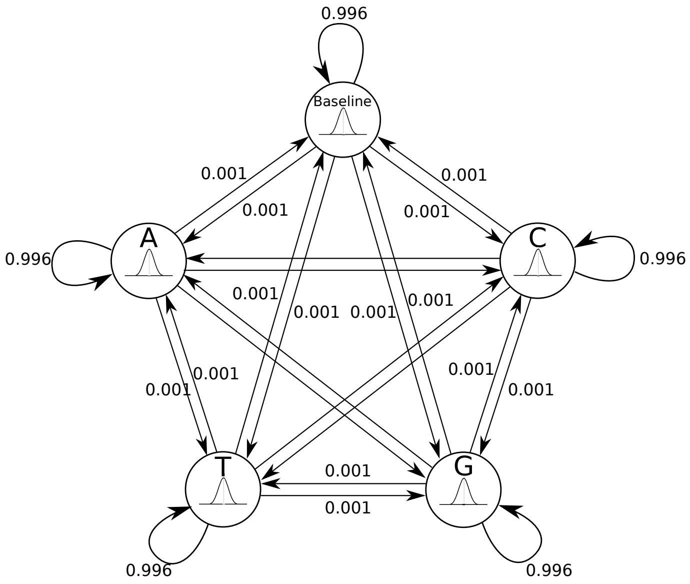

Figure 2: Hidden Markov model used in this analysis.

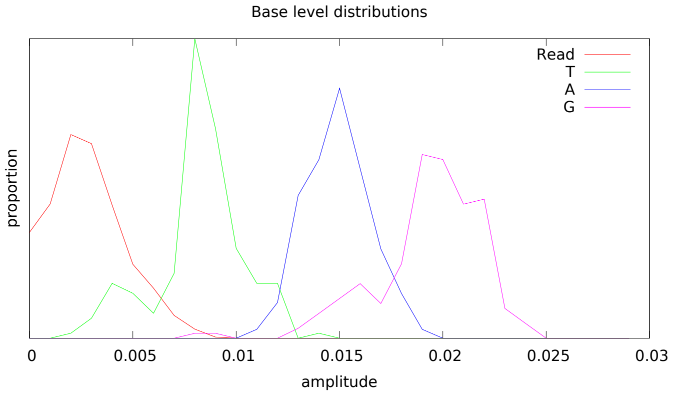

Figure 3: The amplitude of samples assigned to bases and the baseline read level.

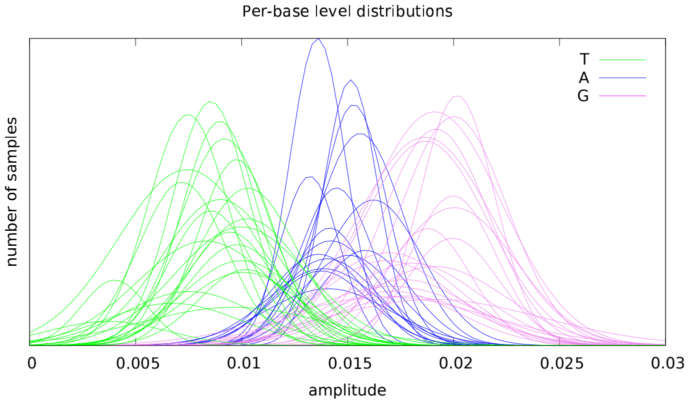

Figure 4: Distribution of sample amplitudes for individual bases in the reads.

Emission means were estimated from the distribution of samples in the entire data set under the assumption that it is a mixture of four Gaussians. However the standard deviations of the Gaussians were fixed to 0.001 for the read level, A, and T, and to 0.0015 for G based on observations of earlier data. The C emission was set to an arbitrary high value as no examples are present in the data. Figure 2 shows the distribution of samples in the data as classified by the HMM. Each component has been normalised to aid visualisation, as the majority of samples are at the read baseline.

Where the histogram of Figure 2 illustrates the variation amongst all samples of each base, Figure 3 shows how much of this variation is due to differences between each example of each base. Each base is shown as a Gaussian distribution with its samples' mean and standard deviation.

3 Dwell time

Figure 4 gives the duration in samples that each base dwells in the read head. These values are determined from the longer reads in the data so that the homopolymers TT and GG can be assigned based on the reads' alignment to the reference sequence. The filled bars in the histograms show the proportion believed to represent homopolymer subsequences.

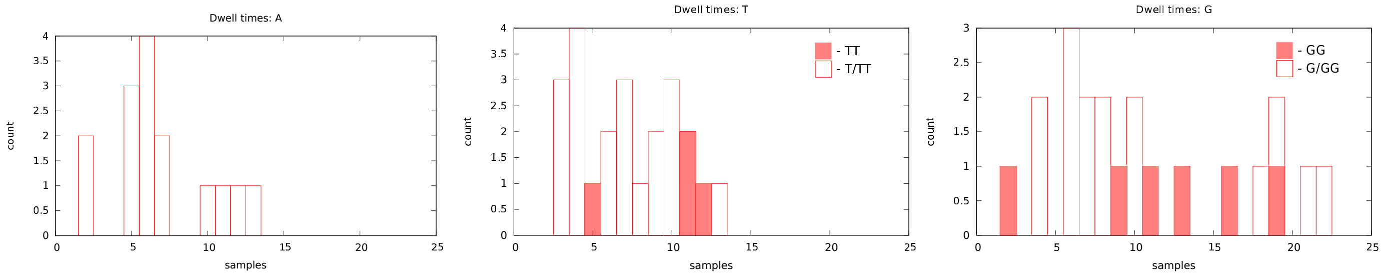

Figure 5: Dwell times of each base with homopolymers manually annotated.

4 Read length distribution

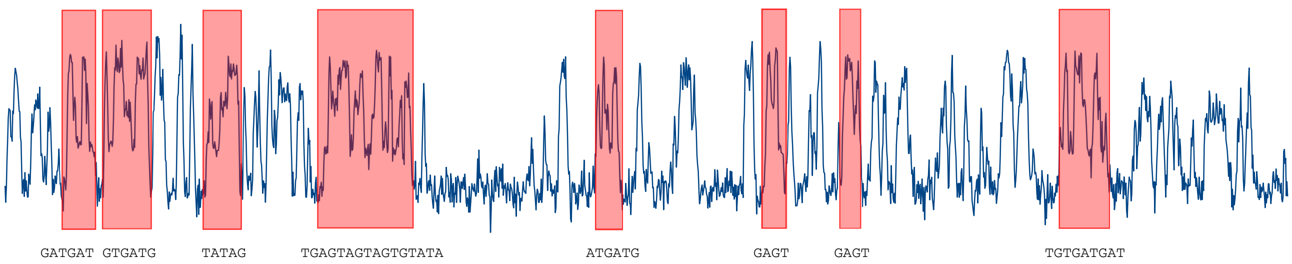

Figure 6: Data set with reads of four or more bases highlighted.

After collapsing homopolymer subsequences, the reference sequence contains 18 bases. This data set contains 23 reads, of which the longest is 16 bases (again treating homopolmers as single bases). Events in the signal that most likely represent a read of single base were ignored in these counts.

The average read length was 3.95, for a total of 91 bases over 0.2 seconds. The 8 reads of 4 or more bases are shown in Figure 5.

5 Future releases

Figure 6 shows the HMM's base call of the longest read in the data set. Interested readers are encouraged to further explore the raw data for this preliminary release at `http://www.quantumbiosystems.com/data`

It is our intent to continue regular data releases as the technology and product develops. In the coming weeks, experiments on DNA with full base coverage, slower movement rates, reduced noise, and higher sampling rates will produce new data sets that we will release to the community.