➢ Assignment 2

BASE-CALLING FOR ROCHE 454 LIFE SCIENCES

The Roche 454 platform starts by constructing an adaptor flanked single strand DNA sequence library. The sequence fragments are bound to beads and amplified on the beads by emulsion PCR in order to increase the downstream signal intensity. Ideally, during this process a single template is attached to each bead leading to uniform clusters on each bead. The beads are then deposited onto an array of picoliterscale wells such that each well contains a single bead. After these preparatory steps, the actual sequencing begins using the pyrosequencing method .In every sequencing cycle, a single species of nucleotides is introduced. In wells where the nucleotides are incorporated, this results in the release of pyrophosphate which eventually leads to a burst of light. The light is detected using a CCD sensor and software detects wells containing template DNA. This step includes image analysis and base-calling. For a more detailed description we refer to the original paper .

A number of sources of errors have been described [9]. Firstly, there is a risk of mixed clusters, caused by the binding of different DNA fragments to a single bead. In such a case, it will be impossible to detect a clear signal and the acquired data from the wells containing such beads has to be excluded. Secondly, in every cycle there is a slight chance of incomplete synthesis of the complementary DNA strand which leads to phasing. Similarly if the reagent of a previous cycle was not perfectly removed, it is possible that multiple different bases are incorporated, resulting in pre-phasing .The main source of errors is, however, due to thresholding. Thresholds are needed to determine whether a base was incorporated or not. The thresholds necessary to determine the lenghts of homopolymers are even more delicate. Homopolymers are consecutive runs of the same base. Since all bases of a homopolymer are included in a single cycle, the length of the homopolymer has to be inferred from the signal intensity. Incorrect prediction of homopolymer lengths leads to insertions and deletions which are by far the most frequent errors associated with the pyrosequencing technology .

In the original 454 paper, wells containing templates are identified by detecting the key sequence ‘TCAG’ at the start of the sequence .The number of incorporated bases is determined from the intensity of the emitted light. It is shown that the intensity is linear with the lengths of the homopolymer, thus allowing for easy classification. A prior on the homopolymer lengths of An external file that holds a picture, illustration, etc.

Object name is bbq077i2.jpg is used. In order to compensate for an incomplete extension rate of 0.1–0.3% and a carry forward rate (pre-phasing) of 1–2% a detailed physical model is proposed. If, frequent ambiguous intensity levels are detected for a given read, that read is filtered out as a low quality read. This allows to exclude wells containing multiple templates. Finally, a Phred like quality score is assigned to every called base. This quality score corresponds to the log-probability that the base was not an overcall, that is, the predicted homopolymer length was not too long.

In Pyrobayes, Quinlan et al. proposed to improve the above procedure by adapting an empirical prior on the homopolymer lengths and by using a classifier based on an empirical measure of the signal intensity. This challenges the validity of a simple linear classifier. As they illustrated in their report, using this more empirical approach does not reduce the total error rate. However, Pyrobayes clearly outperforms the native base-caller in substitution error rate and in the accuracy of the Phred quality scores. Thus, they argued that Pyrobayes is superior in the context of single nucleotide polymorphism (SNP) prediction.