Massively Parallel DNA Sequencing using Single Molecule Array Technology

Anthony J. Cox[1]

Keywords: Whole genome resequencing, single molecule array.

1 Introduction

Solexa is developing a unique technology that has the potential to transform the economics of DNA sequencing by allowing the sequence of millions of individual DNA molecules to be rapidly determined in parallel. Our approach obviates the need for sorting, cloning and amplification of genomic DNA samples and so lab preparation and reagent overheads are also drastically reduced. The applications of a re−sequencing technology range from SNP determination to transcriptomics.

As well as outlining the basic ideas behind Solexa's sequencing platform, this presentation will describe the high throughput bioinformatics pipeline that we are developing to process the large volumes of image and sequence data that our platform will generate.

2 Sequencing Technology

Genomic DNA is first purified and sheared and the resulting fragments are then immobilized onto a surface as primed single strands at a density of around 100 million molecules per square centimetre. These molecules are then sequenced with a base−by−base sequencing strategy employing proprietary polymerases and modified fluorescently labelled nucleotides. Sensitive optical methods allow the outcome of sequencing reactions to be observed simultaneously at a resolution of millions of individual DNA molecules, thus enabling their sequences to be determined in a massively parallel fashion. We aim to obtain at least 25 bases of sequence from each molecule imaged.

3 Bioinformatics Pipeline

The primary output of the sequencing process consist of a large number of images, each similar to Figure 1. The first task of our bioinformatics pipeline is to convert these data into a set of DNA sequences. This requires rapid processing of the images in real time. The resulting sequences are then aligned to the reference human genome sequence, allowing for sequencing errors and naturally occurring differences. We have developed software that has made it feasible for the large number of inexact alignments required to be performed on a Linux cluster with a modest number of nodes. Lastly, sequencing errors are filtered out to leave behind the genuine variation between the reference sequence and the genome being sequenced, in a final stage analogous to the consensus generation stage of shotgun sequence assembly.

[1] on behalf of Solexa Limited, Chesterford Research Park, Little Chesterford, Essex CB10 1XL, United Kingdom. anthony.cox@solexa.com

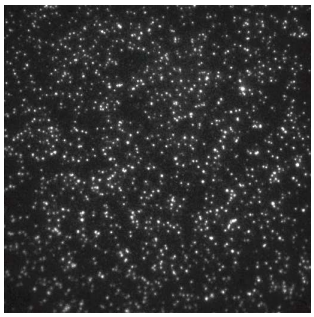

Figure 1: Actual Single Molecule Array image. Each spot is a single DNA molecule with labelled nucleotide attached.

4 Post Sequencing

We have secured UK government funding and set up academic collaborations for the co−development of systems and methods for storage and analysis of genome variation data derived from single molecule array sequencing.