Chapter 2
DNA Hash Pooling and its Applications

Dennis Shasha
New York University, USA

Martyn Amos
Manchester Metropolitan University, UK

ABSTRACT
In this article we describe a new technique for the comparison of populations of DNA strands. Comparison is vital to the study of ecological systems, at both the micro and macro scales. Existing methods make use of DNA sequencing and cloning, which can prove costly and time consuming, even with current sequencing techniques. Our overall objective is to address issues such as whole genome detection, fragment detection and sample similarity. Because our method is similar in spirit to hashing in computer science, we call it DNA hash pooling. To illustrate this method, we describe protocols using pairs of restriction enzymes. The in silico empirical results we present reflect a sensitivity to experimental error. Our method, performed as a filtering step prior to sequencing, may reduce the amount of sequencing required (generally by a factor of 10 or more). Even as sequencing becomes cheaper, an order of magnitude remains important.

INTRODUCTION
Biologists often examine large and diverse populations of organisms (for example, molecules, microbes or plants). This is particularly the case in fields such as microbial ecology, which studies the interactions between living microorganisms (such as algae, or bacteria) and their environment. One of the most significant and challenging problems in these areas of biology is to compare the species in different samples. This task is often made even more difficult by the fact that many “wild” organisms resist laboratory cultivation (and, thus, have unknown phenotypes and their genomes are unknown), or may be present in a population in relatively low numbers.

The study of metagenomics has emerged in recent years (Zhou, 2003; Handelsman, 2004; Tringe et al., 2005; Huson et al., 2007) to perform what has been described as “environmental
forensics”, including the quantification of relative abundances of known species, and the estimation of the number of “unknown” species in a given environment (Huson et al., 2007). The potential impact of this new field is huge, with applications ranging from medicine to agriculture and biotechnology. Furthermore, the insights gained will be of significant assistance in advancing our understanding of biodiversity in both new and familiar environments, such as frozen Antarctic lakes and the human gut (Handelsman, 2004).

The rest of this article is organized as follows: We provide a brief overview of metagenomics, describing current techniques and practices. We discuss several shortcomings of existing methods, motivating the work presented in the following section. Here, we present our new method for population analysis, along with the results of in silico experiments. We conclude with a discussion of the implications of this work, and suggest possible future lines of enquiry.

AN OVERVIEW OF METAGENOMICS

The term metagenomics was first used in print by Handelsman et al. (Handelsman et al., 1998) to describe the study of genetic information recovered from environmental samples. In line with the other “omics” (e.g., proteomics, metabolomics), the emphasis lies on sets of products and/or genes, the suffixes “-ome” and “-omics” being interpreted to imply totality or collectivity (Lederberg & McCray, 2001).

According to Schloss and Handelsman, “metagenomics is the culture-independent analysis of a mixture of microbial genomes (termed the metagenome) using an approach based either on expression or sequencing” (Schloss & Handelsman, 2005). Only a tiny fraction (estimated at below 1%) of micro-organisms found in nature are amenable to isolation and culture for further study (Ammann et al., 1995). This means that we need alternative (“culture-independent”) strategies for studying the vast majority of microbes found in the wild, a significant number of which will be unknown to science (Ward et al., 1990).

The possible implications of metagenomics are many and wide-ranging. As Handelsman argues, Microbiology has long relied on diverse methods for analysis, and metagenomics can provide the tools to balance the abundance of knowledge attained from culturing with an understanding of the uncultured majority of microbial life. Myriad environments on Earth have not been studied with culture-independent methods ... and they invite further analysis. Metagenomics may further our understanding of many of the exotic and familiar habitats that are attracting the attention of microbial ecologists, including deep sea thermal vents, acidic hot springs, permafrost, temperate, desert, and cold soils, Antarctic frozen lakes, eukaryotic host organs, the human mouth and gut, termite and caterpillar guts, plant rhizospheres and phyllospheres, and fungi in lichen symbioses. With improved methods for analysis ...and attraction of diverse scientists to identify new problems and solve old ones, metagenomics will expand and continue to enrich our understanding of microorganisms. (Handelsman, 2004)

Methods for Metagenomics

Early efforts to characterise environmental samples used ribosomal RNA (rRNA). This class of RNA plays a major role in protein synthesis, and rRNA makes up at least 80% of the RNA molecules in a typical eukaryotic cell. The gene encoding one particular molecule - 16S rRNA - is non-coding - rather than “becoming” (or being translated into) a protein, the RNA molecule resulting from its transcription “assists” in protein synthesis. The gene is highly conserved, meaning that it has been preserved throughout the process of speciation through evolution (the implication being that it performs a fundamental role in basic