

by many fragments which overlap. A covered region where there is overlap between consecutive fragments is known as a contig, and it is important to realize that not every covered region is a contig. If one fragment ends just before another fragment begins, we consider the ensemble to consist of separate contigs, although there is no exposed region in between them.

2 The Model

The model most commonly used in the literature requires only that fragments overlap by at least one site in order to form a contig, which is a gross simplification of scientific practice, where many nucleotides are required to overlap to ensure that they really do represent the same area of the genome. We also assume that there are F fragments, each of length L , and the genome is of length G . In reality, L is variable due to the fact that restriction enzymes cut at a specific pattern, and not necessarily after a certain number of nucleotides. Given that there are N fragments of length L , the probability, p , that any given site is the beginning of a fragment is F/G and the coverage a is given by FL/G . We alter this model slightly by assuming that a cut of size L is made independently at each site with probability p , so that the expected number of fragments is Gp (ignoring edge effects).

The genome will thus be represented as a sequence of Y 's and N 's, where Y signifies that a fragment begins at that site, and N means that no fragment begins at that site. For example, if $L = 3$, then a single fragment with no overlapping fragments can be represented by YNN . Hence, the sequence $NNYNNYYNNNNYN$ has three contigs composed of four fragments, and four exposed sites forming two exposed regions. In order to make the sequence more easily readable, exponential notation will be employed; thus the above sequence may be written as $N^2YN^2Y^2N^4YN$.

Previous research in the mathematics of reconstruction (see, e.g. W. J. Ewens and G. R. Grant, *Statistical Methods in Bioinformatics*, Springer Verlag, New York, 2001) has concentrated on the evaluation of the mean number of contigs, the mean contig size, and the mean proportion of the genome covered by contigs: The mean number of contigs is Fe^{-a} , the mean contig size is approximately $L \left(\frac{e^a - 1}{a} \right)$ and mean proportion of the genome covered is found to be $1 - e^{-a}$. These results naturally lead to questions about the probability distributions of these quantities. This paper examines the distributions of (i) number of contigs, (ii) the contig size, (iii) the number of exposed regions,

(iv) the size of the exposed regions, (v) and the proportion of the genome covered. The Stein-Chen method is employed to derive Poisson approximations for (i) and (iii), and a compound Poisson approximation for (v) – with the compounding distributions derived in (ii) and (iv) playing a key role.

Extensions to the case of anchoring and Markov generation of the sequence of cuts will be provided. Much of the work presented is new; in other cases, there is significant overlap with the research of Arratia, Lander, Waterman, Schbath and others.