Coverage statistics for sequence census methods

Steven N. Evans, Valerie Hower and Lior Pachter

May 3, 2010

Abstract

Background: We study the statistical properties of fragment coverage in genome sequencing experiments. In an extension of the classic Lander-Waterman model, we consider the effect of the length distribution of fragments. We also introduce the notion of the *shape* of a coverage function, which can be used to detect abberations in coverage. The probability theory underlying these problems is essential for constructing models of current high-throughput sequencing experiments, where both sample preparation protocols and sequencing technology particulars can affect fragment length distributions.

Results: We show that regardless of fragment length distribution and under the mild assumption that fragment start sites are Poisson distributed, the fragments produced in a sequencing experiment can be viewed as resulting from a two-dimensional spatial Poisson process. We then study the jump skeleton of the the coverage function, and show that the induced trees are Galton-Watson trees whose parameters can be computed.

Conclusions: Our results extend standard analyses of shotgun sequencing that focus on coverage statistics at individual sites, and provide a null model for detecting deviations from random coverage in high-throughput sequence census based experiments. By focusing on fragments, we are also led to a new approach for visualizing sequencing data that should be of independent interest.

1 Introduction

The classic "Lander-Waterman model" [15] provides statistical estimates for the read coverage in a whole genome shotgun (WGS) sequencing experiment via the Poisson approximation to the Binomial distribution. Although originally intended for estimating the extent of coverage when mapping by fingerprinting random clones, the Lander-Waterman model has served as an essential tool for estimating sequencing requirements for modern WGS experiments [17]. Although it makes a number of simplifying assumptions (e.g. fixed fragment length and uniform fragment selection) that are violated in actual experiments, extensions and generalizations [19, 18] have continued to be developed and applied in a variety of settings.

The advent of "high-throughput sequencing", which refers to massively parallel sequencing technologies has greatly increased the scope and applicability of sequencing experiments. With the increasing scope of experiments, new statistical questions about coverage statistics have emerged. In particular, in the context of *sequence census methods*, it has become important to understand the *shape* of coverage functions, rather than just coverage statistics at individual sites.

Sequence census methods [20] are experiments designed to assess the content of a mixture of molecules via the creation of DNA fragments whose abundances can be used to infer those of the original molecules. The DNA fragments are identified by sequencing, and the desired abundances inferred by solution of an inverse problem. An example of a sequence census method is ChIP-Seq. In this experiment, the goal is to determine the locations in the genome where a specific protein binds. An antibody to the protein is used to "pull down" fragments of DNA that are bound via a process called chromatin immunoprecipitation (abbreviated by ChIP). These fragments form the "mixture of molecules" and after purifying the DNA, the fragments are determined by sequencing. The resulting sequences are compared to the genome, leading to a *coverage function* that records, at each site, the number of sequenced fragments that contained it. As with many sequence census methods, "noise" in the experiment leads to random sequenced fragments that may not correspond to bound DNA, and therefore it is necessary to identify regions of the coverage function that deviate from what is expected according to a suitable null model.

The purpose of this paper is not to develop methods for the analysis of ChIP-Seq (or any other sequence census method), but rather to present a null model for the shape of a coverage function that is of general utility. That is, we propose a definition for the shape of a fragment coverage function, and describe a random instance assuming that fragments are selected at random from a genome, with lengths of fragments given by a known distribution. The distinction between our work and previous statistical studies of sequencing experiments, is that we go beyond the description of coverage at a single location, to a description of the change in coverage along a genome.

2 The shape of a fragment coverage function

We begin by explaining what we mean by a *coverage function.* Given a genome modeled as a string of fixed length N, a coverage function is a function $f : \{1, \dots, N\} \longrightarrow \mathbb{Z}_{\geq 0}$. The interpretation of this function, is that $f(i)$ is the number of sequenced fragments obtained from a sequencing experiment that cover position i in the genome. It is important to note that N is typically large; for example, the human genome consists of approximately 2.8 billion bases. Because N is very large, we replace the finite set $\{1, \dots, N\}$ with $\mathbb{R}$, and re-define a coverage function to be a function $f : \mathbb{R} \longrightarrow \mathbb{Z}_{\geq 0}$. This helps to simplify our analysis.

We next introduce an object that describes a sequence coverage function's shape. Our approach is motivated by recent applications of topology including persistent homology [2, 21] and the use of critical points in shape analysis [1, 5, 6]. For a given coverage function $f : \mathbb{R} \longrightarrow \mathbb{Z}_{\geq 0}$, we will define a rooted tree, which is a particular type of directed graph with all the directed edges pointing away from the root. This tree T_f is based on the *upper-excursion sets of* f: $U_h := \{(x, f(x)) | f(x) \geq h\}$, $h \in \mathbb{Z}_{\geq 0}$ and keeps track of how the sets U_h evolve as h decreases. Long paths in T_f represent features of the coverage function that persist through many values of h.

Specifically, for each $h \in \mathbb{Z}_{\geq 0}$, let C_h denote the set of connected components of the upper-excursion set U_h. We define the rooted tree $T_f = (V, E)$ as follows

- Vertices in V correspond to the connected components in the collection $\{C_h\}_{h \in \mathbb{Z}_{\geq 0}}$
- $(i, j) \in E$ provided their corresponding connected components $c_i \in C_{h_i}$ and $c_j \in C_{h_j}$ with $h_i < h_j$ satisfy $h_i = h_j - 1$ and $c_j \subset c_i$.

Note that the root of T_f corresponds to the single connected component in C_0. The tree T_f is very similar to a contour tree [1, §4.1], which is built using level sets of a function, and a join tree [3].

Indeed, suppose we ignore every vertex that is adjacent to only one vertex with greater height. Then, the remaining vertices of T_f correspond to (equivalence classes of) local extrema of f. Each local maximum of f yields the birth of a new connected component as we sweep down through $h \in \mathbb{Z}_{\geq 0}$ while a local minimum of f merges connected components. Since we do not require f to have distinct critical values (as is frequently assumed), the vertices in T_f can have arbitrary degrees, as is depicted in Figure 1C.

In the sequel, we will use the following equivalent characterization that can be found in [7, §2.3]. Given a coverage function $f : \mathbb{R} \longrightarrow \mathbb{Z}_{\geq 0}$ with $f(a) = f(b) = 0$ and $f(x) > 0$ for $x \in (a, b)$, we form an integer-valued sequence $x_0, \ldots, x_{2n}$ that records the changes in height of f on the interval $[a, b]$. The sequence $x_0, \ldots, x_{2n}$ consists of the y values that f travels through from $x_0 := f(a) = 0$ to $x_{2n} := f(b) = 0$ and satisfies

$$x_0 = x_{2n} = 0,$$
$$x_i > 0 \text{ for } 0 < i < 2n,$$
$$|x_i - x_{i-1}| = 1 \text{ for } 1 \leq i \leq 2n.$$

Such a sequence is called a *lattice path excursion away from* 0. Next, we define an equivalence relation on the set $\{0, 1, \ldots, 2n\}$ by setting

$$i \equiv j \Longleftrightarrow x_i = x_j = \min_{i \leq k \leq j} x_k.$$

The equivalence classes under this relation are in 1 : 1 correspondence with the connected components in the upper-excursion sets of $f|_{[a,b]}$. One equivalence class is $\{0, 2n\}$, and if $\{i_1, \ldots, i_p\}$ is an equivalence class with $0 < i_1 < i_2 < \ldots < i_p$ then $x_{i_1 - 1} = x_{i_1} - 1$, whereas $x_{i_q - 1} = x_{i_q} + 1$ for $2 \leq q \leq p$. Conversely, any index i with $x_{i-1} = x_i - 1$ is the minimal element of an equivalence class. We use the minimal element of each equivalence class as its representative. Thus, we can view the vertices of $T_{f|_{[a,b]}}$ as the set $\{0\} \cup \{i | x_{i-1} = x_i - 1\}$. Two indices $i_1 < i_2$ are adjacent in $T_{f|_{[a,b]}}$ provided $x_{i_2} = x_{i_1} + 1$ and $x_k \geq x_{i_1}$ for $i_1 \leq k \leq i_2$. Figure 1 gives an example of a coverage function together with its lattice path excursion $(0, 1, 2, 3, 4, 3, 2, 3, 4, 5, 4, 3, 2, 3, 2, 1, 0)$ and rooted tree. The minimal elements of each equivalence class in Figure 1B are depicted with red squares.

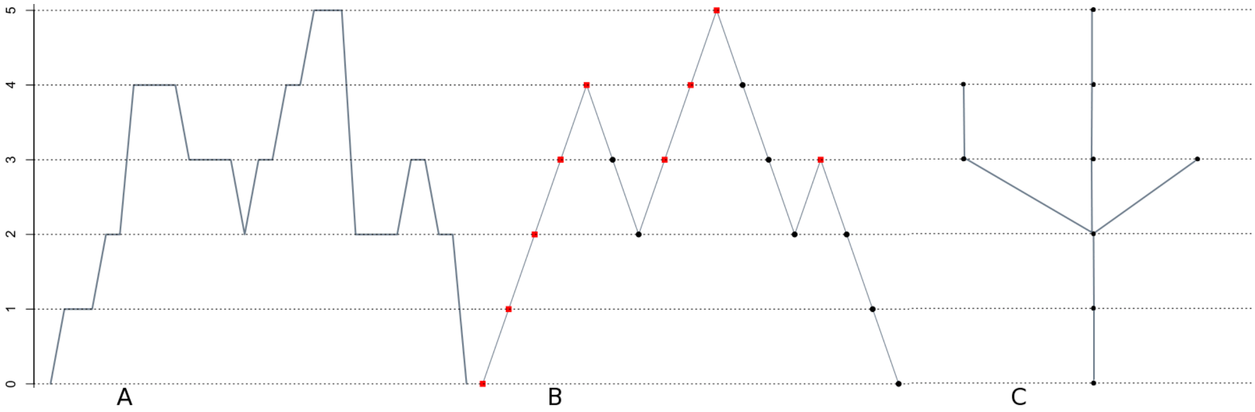

Figure 1: A coverage function (A) with its lattice path excursion (B) and rooted tree (C).

3 Planar Poisson processes from sequencing experiments

In order to model random coverage along the genome, we use a Poisson process to give random starting locations to the fragments. Specifically, suppose that we have a stationary Poisson point process on $\mathbb{R}$ with intensity ρ. At each point of the Poisson point process we lay down an interval that has that point as its left end-point. The lengths of the successive intervals are independent and identically distributed with common distribution μ. We will use the notation X for a coverage function built from this process and X_t for the height at a point t.

Let $t_1, t_2, \cdots$ be the left-end points and $l_1, l_2, \cdots$ be the corresponding lengths of intervals. The interval given by (t_i, l_i) will cover a nucleotide t_0 provided $t_i \leq t_0$ and $t_i + l_i \geq t_0$. We can view this pictorially by plotting points $\{(t_j, l_j)\}$ in the plane. Then X_{t_0}—the number of intervals covering t_0—is the number of points in the triangular region below. We now recall the definition of a two-dimensional Poisson process and refer the reader to [10, §6.13] or [4, §2.4] for the details. Suppose Γ is a locally finite measure on the Borel σ-algebra $\mathscr{B}(\mathbb{R}^2)$. A random countable subset Π of $\mathbb{R}^2$ is called a *non-homogeneous Poisson process with mean measure* Γ if, for all Borel subsets A, the random variables $N(A) := \#(A \cap \Pi)$ satisfy:

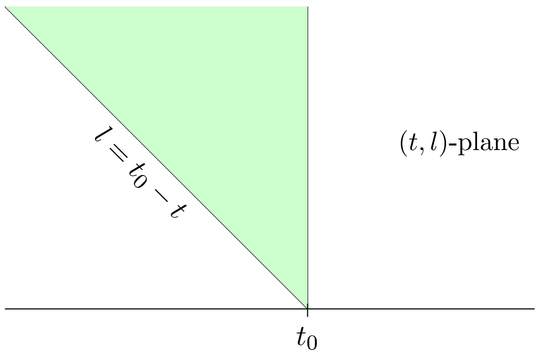

Figure 2: A two dimensional view of a sequencing experiment.

1. $N(A)$ has the Poisson distribution with parameter $\Gamma(A)$, and
2. If $A_1, \cdots, A_k$ are disjoint Borel subsets of $\mathbb{R}^2$, then $N(A_1), \cdots, N(A_k)$ are independent random variables.

The following theorem is a consequence of [14, Proposition 12.3].

Theorem 3.0.1. *The collection $\{(t_i, l_i)\}$ of points obtained as described above is a non-homogeneous Poisson process with mean measure $\rho\, m \otimes \mu$. Here m is Lebesgue measure on $\mathbb{R}$.*

We compute the expected value $\mathbb{E}[X_t] = \rho\, m \otimes \mu(\text{wedge})$:

$$\begin{aligned} \rho\, m \otimes \mu(\text{wedge}) \quad &= \rho \int_{-\infty}^{t} \int_{t-u}^{\infty} \mu(dv)du \\ &= \rho \int_{-\infty}^{t} \mu((t-u, \infty))du \\ &= \rho \int_{0}^{\infty} \mu((s, \infty))ds. \end{aligned}$$

3.1 Fragment lengths have the exponential distribution

We treat the simplest case first, namely the case where the distribution μ of fragment lengths is exponential with rate λ. Then, we have $\mu((s,\infty)) = \mathbb{P}\{l > s\} = e^{-\lambda s}$, and

$$\mathbb{E}(X_t) = \rho \int_0^\infty e^{-\lambda s} ds = \frac{\rho}{\lambda}.$$

Claim 1. *The process X is a stationary, time-homogeneous Markov process.*

Proof. It is clear that X is stationary because of the manner in which it is constructed from a Poisson process on $\mathbb{R}^2$ that has a distribution which is invariant under translations in the t direction; that is, the random set $\{(t_i, l_i)\}$ has the same distribution as $\{(t_i + t, l_i)\}$ for any fixed $t \in \mathbb{R}$. Since μ is exponential, it is memoryless, meaning for any interval length l with an exponential distribution

$$\mathbb{P}\{l > a + b | l > a\} = \mathbb{P}\{l > b\}.$$

This means that probability that an interval covers t_2 knowing that it covers t_1 is the same as the probability that an interval starting at t_1 covers t_2. Thus, the probability that $X_{t_2} = k$ given X_t for $t \leq t_1$ only depends on the value of X_{t_1}. Indeed, in terms of time, $\mathbb{P}\{X_{t_2} = k | X_{t_1} = k'\}$ depends only on $t_2 - t_1$. □

More specifically, X is a birth-and-death process with birth rate $\beta(k) = \rho$ in all states k and death rate $\delta(k) = k\lambda$ in state $k \geq 1$. Note that as the exponential distribution is the only distribution with the memoryless property, we lose the Markov property when μ is not exponential.

To build the tree of §2, we are interested in the jumps of the coverage function $f(t) = X_t$. We hence consider the jump chain of X— a discrete-time Markov chain with transition matrix

$$P(i,j) = \begin{cases} 1, & \text{if } i = 0 \text{ and } j = 1, \\ \frac{\rho}{\rho + i\lambda}, & \text{if } i \geq 1 \text{ and } j = i + 1, \\ \frac{i\lambda}{\rho + i\lambda}, & \text{if } i \geq 1 \text{ and } j = i - 1, \\ 0, & \text{otherwise.} \end{cases}$$

Suppose now we have a lattice path excursion starting at 0. Given a vertex v of the associated tree at height k, we are interested in the number of offspring (at height $k+1$) of this vertex. Suppose i_0 is the minimal equivalence class representative for vertex v, and suppose $[i_0] = \{i_0, i_1, \cdots, i_n\}$ with $i_0 < i_1 < \cdots < i_n$. Then, we have $x_{i_r} = k$ for $0 \leq r \leq n$, $x_{i_r+1} = k+1$ for $0 \leq r \leq n-1$, $x_{i_n+1} = k-1$, and $x_t > k$ for $i_0 < t < i_n$ with $t \neq$ some i_r. From the Markov property, for $0 \leq j \leq n$, $\mathbb{P}\{x_{i_j+1} = k+1 | x_{i_j} = k\} = \frac{\rho}{\rho+\lambda k}$ and $\mathbb{P}\{x_{i_j+1} = k-1 | x_{i_j} = k\} = \frac{\lambda k}{\rho+\lambda k}$. The resulting tree is a Galton-Watson tree with generation-dependent offspring distributions (see [8, 9, 12, 13] for more on Galton-Watson trees). Indeed, we have

$$\mathbb{P}\{\text{a vertex at height } k \text{ has } n \text{ offspring}\} = \left(\frac{\rho}{\rho + \lambda k}\right)^n \frac{\lambda k}{\rho + \lambda k},$$

which is the probability of n failures before the first success in a sequence of independent Bernoulli trials where the probability of success equals $\frac{\lambda k}{\rho+\lambda k}$.

3.2 Fragment lengths have a general distribution

Suppose that we have a general distribution μ for the fragment lengths. We observe X at some fixed "time" – which might as well be 0 because of stationarity, and ask for the conditional probability given X_0 that the next jump of X will be upwards. We know from the above that if μ is exponential with rate λ, then conditional on $X_0 = k$ this is $\rho/(\rho + k\lambda)$.

Let T denote the time until the next segment comes along. This random variable has an exponential distribution with rate ρ and is independent of X_0 [4, §2.1]. If we condition on $X_0 = k$, the two-dimensional Poisson point process must have k points in the region

$$A := \{(t, l) : -\infty < t \leq 0,\ -t < l < \infty\}.$$

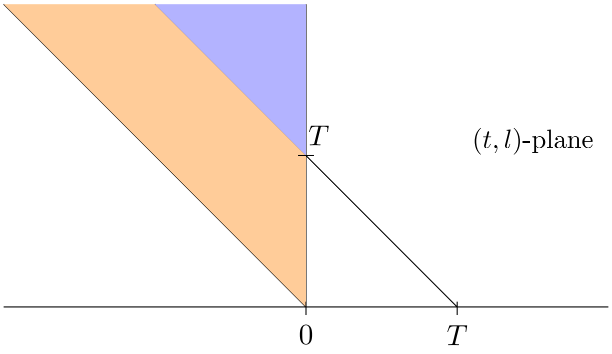

Figure 3: A wedge from the planar Poisson process.

Conditionally, these k points in A have the same distribution as k points chosen at random in A according to the probability measure

$$\frac{\rho\, m \otimes \mu(B)}{\rho\, m \otimes \mu(A)} \quad \text{for} \quad B \subset A$$

However, in order that the next jump after 0 is upwards, the two-dimensional Poisson point process must have no points in the orange region

$$\{(t, l) : -\infty < t \leq 0,\ -t < l < T - t\}$$

as these segments end before time T. This leaves the k points lying in the blue region

$$B_T := \{(t, l) : -\infty < t \leq 0,\ T - t \leq l < \infty\},$$

which occurs with probability $\left(\frac{\rho \int_T^\infty \mu((u,\infty))\, du}{\rho \int_0^\infty \mu((u,\infty))\, du}\right)^k$. Thus, conditional on $X_0 = k$, the probability that the next jump will be upwards is

$$\int_0^\infty \left(\frac{\int_t^\infty \mu((u, \infty))\, du}{\int_0^\infty \mu((u, \infty))\, du}\right)^k \rho e^{-\rho t}\, dt.$$

Write $p(k)$ for this quantity. A reasonable approximation to the jump skeleton Z of X is to take it be a discrete-time Markov chain on the nonnegative integers with transition probabilities

$$P(i,j) = \begin{cases} 1, & \text{if } i = 0 \text{ and } j = 1, \\ p(i), & \text{if } i \geq 1 \text{ and } j = i+1, \\ 1 - p(i), & \text{if } i \geq 1 \text{ and } j = i-1, \\ 0, & \text{otherwise.} \end{cases}$$

The resulting tree is then a Galton-Watson tree with generation dependent offspring distributions, where

$$\mathbb{P}\{\text{a vertex at height } k \text{ has } n \text{ offspring}\} = p(k)^n(1-p(k)).$$

Example 3.2.1. *Suppose μ is the point mass at L (that is, all segment lengths are L). Then*

$$\mu((u,\infty)) = \begin{cases} 1, & u < L \\ 0, & u \geq L \end{cases},$$

and

$$\int_t^\infty \mu((u,\infty))du = \begin{cases} \int_t^L du = L - t, & t < L \\ 0, & t \geq L. \end{cases}$$

This gives

$$\begin{aligned} p(k) \quad &= \int_0^L \frac{(L-t)^k}{L^k} \rho e^{-\rho t} dt \\ &= \int_0^1 w^k \rho e^{-\rho(L-Lw)} L dw \\ &= \theta e^{-\theta} \int_0^1 w^k e^{\theta w} dw \quad \textit{for} \quad k \geq 1, \end{aligned}$$

where $\theta := \rho L = \mathbb{E}[X_0]$. We integrate by parts and find that $p(k) = \theta e^{-\theta} q(k)$ where

$$\begin{aligned} q(k) \quad &= \frac{w^k e^{\theta w}}{\theta}\Bigg|_{w=0}^{w=1} - \frac{k}{\theta}\int_0^1 w^{k-1} e^{\theta w} dw \\ &= \frac{e^\theta}{\theta} - \frac{k}{\theta} q(k-1) \quad \textit{for} \quad k \geq 2, \end{aligned}$$

which yields the recursion

$$p(k) = 1 - \frac{k}{\theta} p(k-1), \quad k \geq 2, \quad \textit{with} \quad p(1) = 1 - \frac{1}{\theta} + \frac{e^{-\theta}}{\theta}.$$

Solving explicitly, we obtain

$$p(k) = k! \left(\sum_{j=0}^k \frac{(-1)^{k-j}}{j!\theta^{k-j}} + \frac{(-1)^{k-1}e^{-\theta}}{\theta^k} \right) \quad \textit{for} \quad k \geq 1.$$

4 Discussion

Our observation that randomly sequenced fragments from a genome form a planar Poisson process in $(position, length)$ coorindates has implications beyond the coverage function analysis performed in this paper. For example we have found that the visualization of sequencing data in this novel form is useful for quickly identifying instances of sequencing bias by eye, as it is easy to "see" deviations from the Poisson process. An example is shown in Figure 4 where fragments from an Illumina sequencing experiment are compared with an idealized simulation (where the fragments are placed uniformly at random). Specifically, paired-end reads from an RNA-Seq experiment conducted on a GAII sequencer were mapped back to the genome and fragments inferred from the read end locations. Bias in the sequencing is immediately visible, likely due to non-uniform PCR amplification [11] and other effects. We hope that others will find this approach to visualizing fragment data of use.

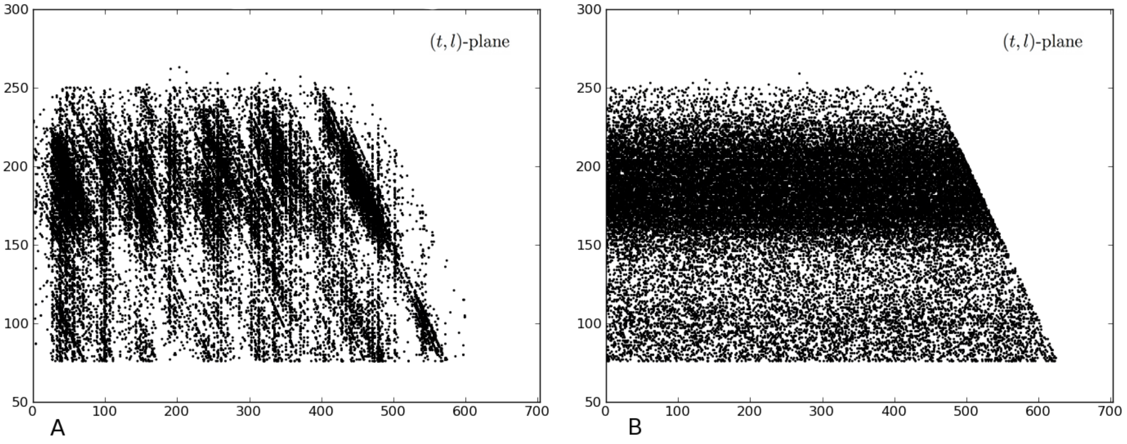

Figure 4: (A) Fragments from a sequencing experiment shown in the (t, l) plane. (B) The spatial Poisson process resulting from fragments with the same length distribution as (A) but with position sampled uniformly at random.

The "shape" we have proposed for coverage functions was motivated by persistence ideas from topological data analysis (TDA). In the context of TDA, our setting is very simple (1-dimensional), however unlike what is typically done in TDA, we have provided a detailed probabilistic analysis that can be used to construct a null hypothesis for coverage-based test statistics. For example, we envision computing test statistics [16] based on the trees constructed from coverage functions and comparing those to the statistics expected from the Galton-Watson trees. It should be interesting to perform similar analyses with high-dimensional generalizations for which we believe many of our ideas can be translated. There are also biological applications, for example in the analysis of pooled experiments where fragments may be sequenced from different genomes simultaneously.

Indeed, we believe that the study of sequence coverage functions that we have initiated may be of use in the analysis of many sequence census methods. The number of proposed protocols has exploded in the past two years, as a result of dramatic drops in the price of sequencing. For example, in January 2010, the company Illumina announced a new sequencer, the HiSeq 2000, that they claim "changes the trajectory of sequencing" and can be used to sequence 25Gb per day. Although technologies such as the HiSeq 2000 were motivated by human genome sequencing a surprising development has been the fact that the majority of sequencing is in fact being used for sequence census experiments [20]. The vast amounts of sequence being produced in the context of complex sequencing protocols, means that a detailed probabilistic understanding of random sequencing is likely to become increasingly important in the coming years.

5 Acknowledgements

SNE is supported in part by NSF grant DMS-0907630 and VH is funded by NSF fellowship DMS-0902723. We thank Adam Roberts for his help in making Figure 4.

6 Author Contributions

LP proposed the problem of understanding the random behaviour of coverage functions in the context of sequence census methods. VH investigated the jump skeleton based on ideas from topological data analysis. SE developed the probability theory and identified the relevance of Theorem 3.0.1. SNE, VH and LP worked together on all aspects of the paper and wrote the manuscript.

References

[1] S. Biasotti, D. Giorgi, M. Spagnuolo, and B. Falcidieno. Reeb graphs for shape analysis and applications. *Theoretical Computer Science*, 392(1-3):5 – 22, 2008. Computational Algebraic Geometry and Applications.

[2] Gunnar Carlsson. Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):255–308, 2009.

[3] Hamish Carr, Jack Snoeyink, and Ulrike Axen. Computing contour trees in all dimensions. *Comput. Geom.*, 24(2):75–94, 2003. Special issue on the Fourth CGC Workshop on Computational Geometry (Baltimore, MD, 1999).

[4] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes.* Springer Series in Statistics. Springer-Verlag, New York, 1988.

[5] Mark de Berg and Marc van Kreveld. Trekking in the Alps without freezing or getting tired. *Algorithmica*, 18(3):306–323, 1997. First European Symposium on Algorithms (Bad Honnef, 1993).

[6] Herbert Edelsbrunner, John Harer, and Afra Zomorodian. Hierarchical Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete Comput. Geom.*, 30(1):87–107, 2003. ACM Symposium on Computational Geometry (Medford, MA, 2001).

[7] Steven N. Evans. *Probability and real trees*, volume 1920 of *Lecture Notes in Mathematics.* Springer, Berlin, 2008. Lectures from the 35th Summer School on Probability Theory held in Saint-Flour, July 6–23, 2005.

[8] Dean H. Fearn. Galton-Watson processes with generation dependence. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. IV: Biology and health*, pages 159–172, Berkeley, Calif., 1972. Univ. California Press.

[9] I. J. Good. The joint distribution for the sizes of the generations in a cascade process. *Proc. Cambridge Philos. Soc.*, 51:240–242, 1955.

[10] Geoffrey R. Grimmett and David R. Stirzaker. *Probability and random processes*. Oxford University Press, New York, third edition, 2001.

[11] K Hansen, SE Brenner, and S Dudoit. Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 2010.

[12] Theodore E. Harris. *The theory of branching processes*. Dover Phoenix Editions. Dover Publications Inc., Mineola, NY, 2002. Corrected reprint of the 1963 original [Springer, Berlin; MR0163361 (29 #664)].

[13] Peter Jagers. Galton-Watson processes in varying environments. *J. Appl. Probability*, 11:174–178, 1974.

[14] Olav Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

[15] ES Lander and MS Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2:231–239, 1988.

[16] FA Matsen. A geometric approach to tree shape statistics. *Systematic Biology*, 4:652–661, 2006.

[17] JL Weber and EW Myers. Human whole-genome shotgun sequencing. *Genome Research*, 7:401–409, 1997.

[18] MC Wendl. A general coverage theory for shotgun DNA sequencing. *Journal of Computational Biology*, 13:1177–1196, 2006.

[19] MC Wendl and W Brad Barbazuk. Extension of Lander-Waterman theory for sequencing filtered DNA libraries. *BMC Bioinformatics*, 6:245, 2005.

[20] B Wold and RM Myers. Sequence census methods for functional genomics. *Nature Methods*, 5:19–21, 2008.

[21] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274, 2005.