The Annals of Applied Statistics
2016, Vol. 10, No. 2, 726–755
DOI: 10.1214/15-AOAS892

SCAN STATISTICS ON POISSON RANDOM FIELDS WITH APPLICATIONS IN GENOMICS

BY NANCY R. ZHANG[2,3], BENJAMIN YAKIR[1], LI C. XIA[2] AND DAVID SIEGMUND[1]

University of Pennsylvania, Hebrew University of Jerusalem, Stanford University School of Medicine and Stanford University

The detection of local genomic signals using high-throughput DNA sequencing data can be cast as a problem of scanning a Poisson random field for local changes in the rate of the process. We propose a likelihood-based framework for such scans, and derive formulas for false positive rate control and power calculations. The framework can also accommodate modified processes that involve overdispersion. As a specific, detailed example, we consider the detection of insertions and deletions by paired-end DNA-sequencing. We propose several statistics for this problem, compare their power under current experimental designs, and illustrate their application on an Illumina Platinum Genomes data set.

1. Introduction. Developments during the last three decades in biology, especially in genetics and fMRI analysis, have motivated theoretical and applied research on the detection of local signals in large fields of data. These data are obtained simultaneously from markers distributed across an entire genome or from scans of the entire brain; for early examples see Lander and Botstein (1989), Karlin, Dembo and Kawabata (1990), Feingold, Brown and Siegmund (1993), Worsley et al. (1992) and Siegmund and Worsley (1995). Over most of the region of observation the data in such studies are noise, while signals appear as local "peaks." The random field is often conveniently assumed to be Gaussian based on approximations using the central limit theorem. Control for the multiple comparisons involved in searching the field for local signals is achieved by using the theory of maxima of Gaussian fields to obtain a significance threshold that controls the overall false positive rate. This approach requires that the normal distribution provide an adequate approximation in the extreme tail of the distribution, which suggests that one be skeptical of the accuracy of the resulting thresholds, especially in many cases where Poisson-like data are involved and the Poisson rate is not large. For asymptotic analyses in various concrete problems without using a

Received April 2014; revised September 2015.
[1]Supported in part by NSF Grant DMS 1043204.
[2]Supported in part by NIH R01 HG006137-01.
[3]Supported in part by the Sloan Foundation.
Key words and phrases. Scan statistics, Poisson processes, change-point detection, next-generation sequencing, structural variation.

Gaussian assumption, see Rabinowitz and Siegmund (1997), Tang and Siegmund (2001), Peng and Siegmund (2005), Chan and Zhang (2007) and Siegmund, Yakir and Zhang (2011).

This paper is motivated by several problems arising from high-throughput DNA sequencing data, where the data can be modeled by a Poisson process, possibly nonhomogeneous, or in some cases a mixture of Poisson processes to deal with overdispersion. The signal of interest involves a local change in a functional of the process. In addition to our primary problem of detecting structural variations using DNA sequencing data, similar scans for local signals arise in the detection of transcription factor binding sites in chromatin immuno-precipitation followed by sequencing [ChIP-Seq, cf. Schwartzman et al. (2013)] and in detection of alternative transcription start and end sites in RNA sequencing. See also Song et al. (2013) for an application to word counts in DNA sequence analysis.

The applications mentioned in the preceding paragraph can be cast as a problem of scanning a Poisson field for clusters of a prespecified configuration. In particular, for structural variant detection, the signal is in the form of a local mixture. We derive approximations for the false positive rates of likelihood-ratio and score-based scan statistics in reasonably general settings. We also study the power of these statistics as a function of scientifically interpretable parameters. Although we use likelihood-based models, the usual regularity conditions of likelihood theory are not satisfied, since the parameters defining the location and the expected height of a peak are not separately identifiable, that is, we cannot speak of the location of a peak unless the expected height is different from the expected value of the background. We also study settings where the problem is irregular in a second sense, that the likelihood function is not differentiable in the parameter defining the location of a signal.

Before concentrating on the detailed models for structural variant detection, we study a simplified model that may have more general interest and that is relatively much easier to analyze. This model can be viewed either as a marked Poisson process, as a compound Poisson process or, alternatively, as a two-dimensional Poisson process where one coordinate of the process involves location in time and/or space, and the other coordinate involves the value of the observation available at that location.

For an example outside of genomics, see Braun et al. (2008) who discuss the use of high-energy neutrino telescopes to search for point sources of galactic and extragalactic neutrinos. Here, atmospheric neutrinos have one energy distribution, and against this random background one would like to identify the existence of localized clusters of neutrinos that have a different distribution of energy. A plausible model is a marked Poisson field, where in certain small regions determined by the point spread function of a distant point source the distribution of the marks associated with a fraction of the points is different from the distribution of the background [Braun et al. (2008)]. Scanning the field for these distinguished point sources has been called a "look elsewhere" test by Gross and Vitells (2010), who

use a Gaussian random field and the well-developed theory of maxima of random fields [e.g., Adler and Taylor (2007)] to obtain approximate p-values.

Segmentation of financial time series using a similar model is discussed by Toth, Lillo and Farmer (2010). See also the series of papers by Kulldorff and collaborators [Kulldorff and Nagarwalla (1995), Kulldorff (1997, 1999)] for applications to detection of clustered disease outbreaks.

We introduce the motivating genomic applications in Section 2. In Section 3, we give a general framework for scans of Poisson random fields, first illustrating it on a simple mixture model (Section 3.1), and then on a more detailed and realistic model for the problem of structural variant detection (Section 3.3). The simple mixture model is more transparent and also leads to insights that help us understand more complex settings. The procedures for p-value approximation for scan statistics on Poisson random fields are sketched in the Supplementary Materials [Zhang et al. (2016)]. In Section 4 we give approximations for the simple mixture model, examine their accuracy by numerical experiments, and analyze power under different scanning designs. In Section 5, we return to explore in more detail the more realistic models for structural variant detection formulated in Section 3.3. We conclude with a discussion in Section 7.

The theory and methods described in this paper are at the core of `SWAN`, a comprehensive statistical pipeline for genomic structural variant detection. `SWAN` is an open source R library available at https://bitbucket.org/charade/swan/wiki/Home. The accuracy of SWAN is compared to mainstream structural variant detection methods on a in silico spike-in data set in Section 5.

2. Motivating examples from sequencing experiments. High-throughput short read sequencing, often referred to as "next-generation sequencing," provides data for quantifying DNA, RNA, protein binding and many other genome-wide features in biology. A good overview of the technology and its applications can be found in three articles in the November 2009 issue of Nature Methods: Flicek and Birney (2009), Medvedev, Stanciu and Brudno (2009) and Pepke, Wold and Mortazavi (2009). As our main example, we consider DNA sequencing, which is described by Medvedev, Stanciu and Brudno (2009). The DNA sequencing pipeline is briefly summarized in Figure 1: In step (1), double-stranded DNA is extracted from the sample of interest and fragmented. In step (2), the fragments are usually amplified, then fragments within a certain size range are selected and made into a DNA library. In step (3), a fixed number of bases, referred to as *reads*, is read by a sequencer from one or both ends of each fragment. When both ends of the fragment are read, the data are referred to as *paired ends*, since the reads are paired, one coming from each end of the double-stranded DNA molecule. Since sequencing is unidirectional, one read of each pair comes from the plus strand of each double stranded fragment, with the other read of the pair coming from the minus strand. Finally, in step (4), the reads are mapped to a reference genome template; Flicek and Birney (2009) give a good overview of this step. Our data include the mapping

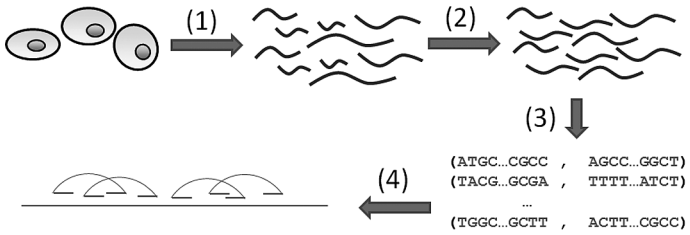

FIG. 1. *Outline of a DNA sequencing experiment.*

location and strand orientation of the reads as well as other attributes that describe the mapping, for example, whether the entire read or only a partial read has been mapped.

For a fragment with both ends mapped, we define the *mapped insert length* to be the number of bases between the start position of the minus-strand read and the start position of the plus-strand read. If the sequenced genome is identical to the reference genome in the region spanned by the reads, then the mapped insert length is simply the length of the fragment from which the read pair is derived minus the length of one read. Important fixed quantities in our models, which are chosen by the experimenter, are the length of each read, R, and the distribution of the insert lengths, which we characterize by a distribution function F with mean δ and standard deviation σ.

2.1. *Detection of copy number variation.* Copy number variation refers to the deletion or duplication of a chromosomal segment of DNA. DNA sequencing has been used to detect copy number variation because the density of reads mapped to a genome interval depends on the relative quantity of that piece of DNA in the sequenced sample [Campbell et al. (2008); Chiang et al. (2009); Abyzov et al. (2011); Shen and Zhang (2012)]. A simple model assumes that the start positions of the mapped reads follow a nonhomogeneous Poisson process $N(t)$ of intensity $\rho(t)$, where t is position along the genome. That is, for $s < t$, $N(t) - N(s)$ is the number of reads that map to the region $(s, t]$ on the reference genome. We will call $N(t)$ the *coverage process*. The function $\rho(t)$ depends not only on the copy number but also on features that are local to the neighborhood of t. For example, it has been shown that the percentage of bases that are G or C dramatically influence the baseline coverage. Assuming that a reliable estimate of $\rho(t)$ is available, we can model a deletion of $(t_1, t_2]$ as a local drop of the intensity function to, say, $\exp(\beta)\rho(t)$, where $\beta < 0$. The log-likelihood ratio for the process N_t, with and without the deletion, is

$$\beta[N_{t_2} - N_{t_1-R}] - \big[\exp(\beta) - 1\big] \int_{t_1-R}^{t_2} \rho(t)\, dt. \tag{2.1}$$

Since the boundaries of the deletion are unknown, a scan statistic would involve maximization of (2.1) over an appropriate range of $t_1 < t_2$, and perhaps also over a

reasonable range for β. This model can obviously also be used to detect increases in DNA copy number.

2.2. *Detection of structural variants*. Structural variants are insertions, deletions, inversions and translocations of segments of DNA in the genome. Deletions result in a loss of copy number, and insertions of DNA in the form of tandem duplications represent copy number gain. Thus, structural variations may cause copy number variation. However, a translocation, which is the movement of a DNA segment from one position in the genome to another, does not result in a change of copy number, although it represents a deletion at the original site and an insertion at the new site.

Structural variants are always parameterized with respect to the reference genome to which reads are mapped. For example, deletion of $[s, t]$ refers to deletion in the target genome of the DNA sequence starting at s and ending at t in the reference genome. Paired-end DNA sequencing allows the detection and sometimes the precise positioning of structural variants. Figure 2 shows how deletions and insertions produce telltale signatures in the mapping of paired-end reads. Figure 2(a) shows a deletion of bases s to t in the reference genome. The deleted region is labeled B and is spanned by regions A and C. In the absence of structural change, the mapped insert length is random with a known distribution F. Fragments that span the deletion point, that is, those that start in A and end in C in the target, produce read pairs that map farther apart in the reference than expected

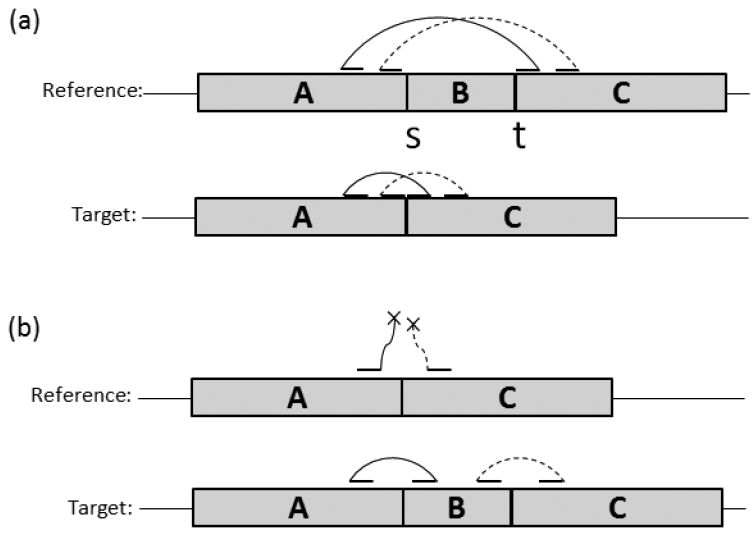

FIG. 2. *Mapping of paired-end reads in region of deletion* (a) *and insertion* (b). *Arcs join reads within a pair. In* (a), *the deleted region, labeled* B, *spans bases* s *to* t *in the reference genome, and is flanked by regions A and C. In* (b), *the inserted segment B is flanked by regions A and C in the target genome, with A and C joined together in the reference genome.*

under F. Now consider Figure 2(b), where an insertion B, spanned by A and C, starts at position t in the reference genome. A read that overlaps with B would fail to map whenever B is a foreign sequence with no homolog in the target genome, or it would map far from its mate if B is a "domestic" insertion from a distant location of the reference. Read pairs where one read maps successfully and the other fails to map, maps in the same orientation or maps too far from the first are called *hanging pairs*. Deletions can also produce hanging pairs whenever one read of a given pair straddles the boundary between A and C in the target genome. Similarly, fragments that span insertions produce read pairs that map closer to each other than expected under F. In Sections 3.1 and 3.3, we describe models and statistics that exploit these patterns to detect structural variants.

2.3. *Detection of transcription factor binding sites.* Chromatin immunoprecipitation (ChIP) is a technique for isolating from a DNA sample only those DNA fragments bound to a protein of interest. Sequencing reads from the ends of the DNA fragments derived from ChIP then mapping these reads to a reference template allows us to detect the binding locations of the protein in the genome of the sample. One expects to see an increase in the density of mapped reads near the binding site. Under the assumption that the binding site is short compared to R, it is natural to assume that the "shape" of the peak centered on the site is roughly triangular. Schwartzman et al. (2013) consider statistics of the form

$$Z_t = \int g_w(t-s)\,dN_s, \tag{2.2}$$

where N_s is the Poisson counting process with jumps at the centers of mapped reads, and g is a symmetric kernel. Schwartzman suggested the function $g_w(s) = (1-|s|/w)^+/w$ as a plausible "matched filter." An alternative kernel is a Gaussian probability density function with standard deviation w. The scale parameter w indicating the width of the signal may be known or unknown. Under this setup, the log-likelihood ratio for testing the intensity function $\rho(s)$ against the alternative of a peak at t of the form $\exp(\beta g_w(t-s))$ equals

$$\ell(t,w,\beta) = \beta Z_t - \int \big[\exp\big(\beta g_w(t-s)\big) - 1\big]\rho(s)\,ds. \tag{2.3}$$

Since the location t is unknown, one might consider one of several statistics maximized over candidate values of t. The simplest would be the score statistic, $\partial\ell/\partial\beta|_{\beta=0} = Z_t - \int g_w(t-s)\rho(s)\,ds$. An alternative would be $\max_\beta \ell(t,w,\beta)$, where the maximum is over (in addition to t) some appropriate range of values of $\beta > 0$ and perhaps also w. Below we shall see a model for detecting insertions and deletions that also involves convolution of a smooth function with a Poisson process.

2.4. *Modeling overdispersion.* It has often been found that the coverage process N_t is overdispersed. The overdispersion can be handled by using a negative binomial process or, equivalently, a gamma mixture of Poisson processes. An alternative that may be more suitable for our present purposes is to assume that a small fraction of DNA fragments, hence points in the Poisson process, occur with random multiplicity. For example, assume that the size of a jump in the process is k with binomial probability having parameters j, α and conditioned to be at least 1. In this case, after multiplication of the rate function $\rho(t)$ by $[1-(1-\alpha)^j]/(j\alpha)$ to maintain the expected null intensity, the log-likelihood becomes

$$\beta[N_{t_2}-N_{t_1-R}]-\big[\Omega(t_2)-\Omega(t_1-R)\big]\big\{\big[1+\alpha\big(\exp(\beta)-1\big)\big]^j-1\big\}/(j\alpha). \tag{2.4}$$

Its null variance is inflated by the factor $1+(j-1)\alpha$, and it reduces to the Poisson case for either $j=1$ or $\alpha\to 0$.

3. Models and scan statistics.

3.1. *A simple mixture model.* Consider first a simplified model for the detection of insertions and deletions using the mapped insert lengths in paired-end sequencing. We consider for now only those pairs where both reads are unambiguously in opposite orientation. For read pair i, let x_i^+ and x_i^- be the mapped positions of the plus- and minus-strand reads, respectively. For a reference template of length m, $(x_i^+, x_i^-) \in \{1,\dots,m-R+1\}^2$. The mapped insert length, which we denote by y_i for read pair i, is defined by $y_i \equiv x_i^- - x_i^+$.

If there are no structural variants, y_i has distribution $F_0(dy)$ with density f_0, mean δ and standard deviation σ. As described in Section 2.2, deletions cause an increase in mapped insert length, and small insertions cause a decrease. We introduce a parameter w, where the sign of w is positive for deletions and negative for insertions, while $|w|$ is the number of bases in the deleted or inserted segment. Apart from the length of the variant, another important parameter influencing the strength of the signal is purity, denoted by r. For example, heterozygous variants, which are carried by 50% of the genomes in a diploid individual, have purity 0.5. In cancer tissue, where the cell population is usually genetically heterogeneous, the purity of a given mutation is a continuous fraction. Both w and r are usually unknown, although it is informative to study statistics defined by particular values of these parameters. As illustrated in Figure 2, read pairs derived from fragments containing the deletions/insertions have *mapped* insert lengths coming from the mixture distribution $F_1(dy) = (1-r)F_0(dy) + rF_0(dy-w)$.

To detect insertions and deletions, we consider as a toy model the two-dimensional Poisson random field

$$N(dt,dy) = \sum_{i=1}^{n} I\big(x_i^+ \in dt, y_i \in dy\big).$$

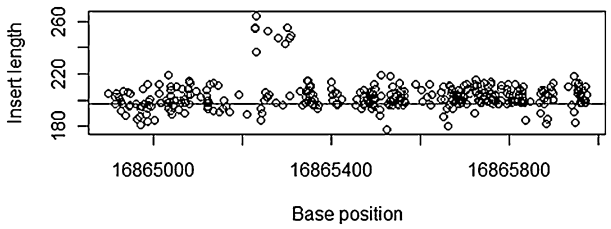

FIG. 3. *Region on chromosome* 22 *from NA*12878 *sequenced by the* 1000 *Genomes Project. This region contains a putative heterozygous deletion of* ∼50 *bases. The mean insert size is shown by the horizontal line.*

We assume for N the null intensity function $\lambda(dt, dy) = \rho(t)\,dt\,F_0(dy)$, where $\rho(t)$ is the rate with which plus-strand reads map to genome position t. Alternatively, we can think of this process as a marked (or compound) Poisson process with rate $\rho(t)$ and marks that follow F_0 or F_1.

This logic prompts the construction of an alternative intensity function

$$\lambda_1(dt, dy) = \begin{cases} \rho(t)\,dt\,F_1(dy), & t \in [s-\delta, s); \\ \rho(t)\,dt\,F_0(dy), & \text{otherwise}, \end{cases} \tag{3.1}$$

for a structural variant starting at s. The log-likelihood ratio of λ_1 versus λ_0 is

$$\begin{aligned} \ell = &\int \log\bigl(\lambda_1(dt, dy)/\lambda_0(dt, dy)\bigr) N(dt, dy) \\ &- \int \bigl[\lambda_1(dt, dy) - \lambda_0(dt, dy)\bigr]. \end{aligned} \tag{3.2}$$

The log-likelihood is indexed by the location parameter s and the parameters w and r in the specification of F_1. A scan of the genome for large values of the log-likelihood, varying s and possibly also r and w, can be used to detect insertions and/or deletions.

Figure 3 shows an example of a region on chromosome 22 containing a signal from the sequencing of NA12878 by the 1000 Genomes Project. Note the cluster of points with an elevated mean at base position 16,865,200–16,865,300. Since the mean increased by around 50 bases in approximately half of the read pairs in this 100 base window, we suspect that there is a heterozygous deletion of roughly 50 bases starting at ∼16,865,300.

Compared to the models introduced below in Section 3.3, the scan statistic suggested here has a simple, general structure due to the assumption that the rate function for the two-dimensional process is a product of separate one-dimensional rate functions in t and y. This structure leads to relatively simple theoretical properties that may be of general interest for problems involving mixtures in compound Poisson processes. Despite its simplicity, this toy model leads to insights in our problem.

3.2. *General framework and notation.* Assume that the observed data are a counting process $\{N(dz) : z \in \Omega\}$, that has a null intensity function $\lambda_0(z)$ on the domain Ω. For example, in the single-read sequencing setup of Sections 2.1 and 2.3, $N(z)$ is the coverage process, with z being a one-dimensional index for genome location. In the mixture model proposed in Section 3.1, $z = (t, y)$, $\Omega = [0, m] \times \Re$, and $N(z)$ counts the number of read pairs with the plus-strand read mapping to a given location and mapped insert length within a given range. The signal of interest in all cases is a local change in intensity, which is represented by an alternative intensity function $\lambda_1(z)$ that relies on one or more parameter(s), collectively denoted by τ. For example, in Section 3.1, τ can be the single parameter s for genome location but can also be the vector (s, r, w) which includes the proportion and length of the variant. We introduce the representation

$$\lambda_1(dz) = e^{\beta k_\tau(z)} \lambda_0(dz), \tag{3.3}$$

where we call $k_\tau(z)$ the kernel function. The parameter β plays a technical role in false positive rate calculations. The alternative of interest is often $\beta = 1$. The log-likelihood ratio (3.2) can be written as

$$\ell_\tau = \beta \int k_\tau(z) N(dz) - \psi_\tau(\beta), \tag{3.4}$$

where

$$\psi_\tau(\beta) = \int \{\exp[\beta k_\tau(z)] - 1\} \lambda_0(dz). \tag{3.5}$$

The scan statistics (2.1) and (2.3) for detecting copy number variants and peaks in ChIP-Seq data are also of this form.

Almost no restrictions are placed on the kernel function, except that the integral (3.5) exists for β in some neighborhood of 0 and that, when appropriate, ψ_τ can be differentiated with respect to the scan parameters under the integral.

The simple model in Section 3.1 has a kernel that is separable in t and y,

$$k_\tau(t, y) = 1\{s - \delta \le t \le s\} g(y; w, r), \tag{3.6}$$

where $g(y; w, r) = \log[1 - r + r f_0(y - w)/f_0(y)]$. The separability allows simple p-value approximations and more direct analysis of power.

In general, we will consider as raw material for scan statistics the random fields (3.4), indexed by the unknown parameters τ. Note that the random field is in general not differentiable in the parameter s for location, but can be differentiable in the other parameters such as r, w in (3.6).

3.3. *More realistic models for structural variants.* The mixture model suggested above neglects several features of the problem of detecting structural variants by paired-end reads. Here we propose alternative models that take into account more specific features of paired-end sequencing, albeit at the cost of a more complicated analysis of the false positive error rate.

Let n be the number of read pairs where at least one read within the pair is successfully mapped to the template. As before, let x_i^+ and x_i^- be the leftmost base positions of the plus- and minus-strand reads, respectively, for pair i. Successfully mapped reads have positions in $\{1, \ldots, m - R + 1\}$. Some reads will fail to map to the reference template, in which case we assign its position the value ∞. Reads may fail to map due to sequencing or mapping error, or due to inclusion of a segment of DNA that does not have a match in the reference. Read pairs where the plus (minus) strand fails to map are called *hanging plus- (minus-) strand pairs*. See Figure 2(b) for an illustration of hanging pairs produced by an insertion. Let p be the probability of a hanging read due to experimental error. A conservative estimate of p can be obtained from $n^{-1}\sum_i [I(x_i^+ = \infty) + I(x_i^- = \infty)]$. Since hanging reads caused by true structural variants are only a very small proportion of the overall number of hanging reads, we expect this conservative estimate to be very accurate. Thus, we assume that p is known.

In Section 2.2 we discussed hanging pairs in a broader context that also includes read pairs that are mapped too far apart or in reverse orientation. The models and statistics we introduce below easily adapt to the broader definition, but the notation will be much simpler under the narrow definition. The important thing is that, given whatever the definition may be for a hanging pair, p should be empirically estimated by n_H/n, where n_H is the total number of read pairs that satisfy this definition.

Let $\kappa(t)$ be the rate with which reads (either plus- or minus-strand) map to position t. For mathematical convenience we embed the discrete mapping positions into the continuous interval $[0, m]$ and let

$$N(du, dv) = \sum_{i=1}^{n} I\big(x_i^+ \in du, x_i^- \in dv\big), \qquad u, v \in [0, m]. \tag{3.7}$$

Then, in the notation of Section 3.2, N is an inhomogeneous Poisson process with $z = (u, v)$, $\Omega = ([0, m] \cup \infty)^2$, and intensity function

$$\lambda_0(u, v) = \begin{cases} (1 - p)\kappa(u)\kappa(v) f(v - u), & u, v \in [0, m]; \\ \dfrac{1}{2} p\kappa(u) \displaystyle\int_u^m \kappa(x) f(x - u)\, dx, & u \in [0, m], v = \infty; \\ \dfrac{1}{2} p\kappa(v) \displaystyle\int_0^v \kappa(x) f(v - x)\, dx, & u = \infty, v \in [0, m]. \end{cases} \tag{3.8}$$

The integrals in the second and third lines account for the possible different insert lengths, which are unobserved because of the hanging read. We assume that hanging reads are equally likely to be a hanging plus strand or a hanging minus strand. Note that the marginal intensity for a read to map to t is $\kappa(t)$. If we assume constant κ, then $\lambda_0(u, v)$ simplifies to $(1 - p)\kappa^2 f(v - u)$ for properly mapped read pairs, and $p\kappa^2/2$ for hanging pairs.

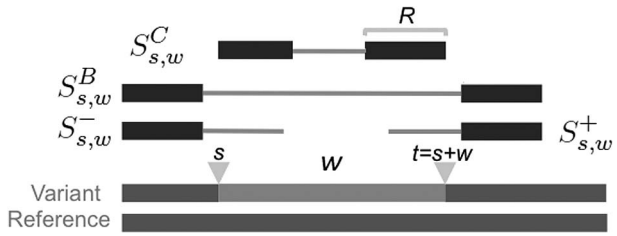

FIG. 4. *Examples of the four types of informative read pairs in the neighborhood of a deletion at* $[s, s+w)$: $S^C_{s,w}$ *have at least one read that covers the window*; $S^B_{s,w}$ *bracket the window*; $S^+_{s,w}$ *have hanging plus-strand reads, with the minus strand mapping to the right of the window*; *and* $S^-_{s,w}$ *have hanging minus-strand reads, with the plus-strand read mapping to the left of the window.*

Now consider testing the alternative hypothesis that a proportion r of the genomes in the sample contain a deletion of width w beginning at reference location s. In reference to the window $[s, s+w)$, the sample space Ω of possible paired-end mappings can be partitioned into the following nonoverlapping sets:

$$S^C = S^C_{s,w} = \{(u,v) : s - R < u < s + w \text{ or } s - R < v < s + w\};$$

$$S^B = S^B_{s,w} = \{(u,v) : u \le s - R \text{ and } v > s + w\};$$

$$S^+ = S^+_{s,w} = \{(u,v) : u = \infty \text{ and } v > s + w\};$$

$$S^- = S^-_{s,w} = \{(u,v) : u \le s - R \text{ and } v = \infty\};$$

$$S^0 = S^0_{s,w} = \Omega \setminus (S^C_{s,w} \cup S^B_{s,w} \cup S^+_{s,w} \cup S^-_{s,w}).$$

Figure 4 shows examples of the first four categories of read pairs: $S^C_{s,w}$ is the set of pairs where at least one read intersects the window $[s, s+w)$; $S^B_{s,w}$ is the set of pairs that *b*racket the window; $S^+_{s,w}$ is the set of hanging plus-strand pairs where the minus-strand read maps to the right of the window; $S^-_{s,w}$ is the set of hanging minus-strand pairs where the plus-strand read maps to the left of the window. Finally, $S^0_{s,w}$ contains all of the remaining pairs, which are uninformative about whether there is a deletion of $[s, s+w)$. To simplify notation, we will sometimes suppress the suffix s, w.

We use the notation from Section 3.1, where the parameter set of the model is $\tau = (s, w, r)$, where s is the location of the putative variant, w is the width, and r is the proportion of genomes in the sample that carry the variant. We call those genomes that carry the variant the *carriers*, and those that do not carry the variant the *noncarriers*.

Let $\lambda_1(u, v)$ be the rate function under the alternative of a deletion with parameters τ. To derive λ_1, we consider the probability under the alternative of read pairs belonging to each of the above sets separately. The deletion should not affect the rate with which pairs map to S^0. Pairs in S^C can only come from the noncarrier

genomes, with probability $1-r$, and thus

$$\lambda_1(u,v) = \lambda_0(u,v)[1-r], \qquad (u,v) \in S^C. \tag{3.9}$$

A pair in S^B can be generated in two ways: It can be from a noncarrier chromosome, with rate $(1-r)\lambda_0(u,v)$, or it can be from a fragment containing the deletion from the carrier chromosome, with rate $r(1-p)\kappa(u)\kappa(v)f(v-u-w)$. Thus,

$$\lambda_1(u,v) = \lambda_0(u,v)\big[1-r+rf(v-u-w)/f(v-u)\big], \qquad (u,v) \in S^B. \tag{3.10}$$

Now consider the hanging minus-strand pairs. A pair mapping to $(u,v) \in S^-$ can be from a noncarrier chromosome, with rate $(1-r)\lambda_0(u,v)$, or it can be from a carrier chromosome. In the latter case, there are two explanations for the minus-strand read failing to map: It can be due to sequencing error or it can be due to the read overlapping the deletion point. Thus, for $(u,v) \in S^-$,

$$\begin{aligned}\lambda_1(u,v) &= (1-r)\lambda_0(u,v) + r\left[\lambda_0(u,v) + (1-p)\kappa(u)\int_{s-R}^{s} f(t-u)\kappa(t)\,dt\right] \\ &= \lambda_0(u,v)\left[1 + \frac{2r(1-p)}{p\int_u^m \kappa(x)f(x-u)\,dx}\int_{s-R}^{s} f(t-u)\kappa(t)\,dt\right].\end{aligned} \tag{3.11}$$

With similar reasoning, we have for $(u,v) \in S^+$

$$\lambda_1(u,v) = \lambda_0(u,v)\left[1 + \frac{2r(1-p)}{p\int_0^v \kappa(x)f(v-x)\,dx}\int_{s+w-R}^{s+w} f(v-t)\kappa(t)\,dt\right]. \tag{3.12}$$

Now we see that the alternative rate function can be written in the form of (3.3) with $\beta = 1$, $k_\tau = 0$ for $(u,v) \in S^0$, and k_τ equal to the log of the term in square brackets in (3.9)–(3.12) for (u,v) belonging to, respectively, S^C, S^B, S^+ and S^-. The log-likelihood scan statistic thus evaluates to

$$\ell_\tau = \beta\big[Z_\tau^C + Z_\tau^B + Z_\tau^+ + Z_\tau^-\big] - \psi_\tau(\beta), \tag{3.13}$$

where Z_τ^C, Z_τ^B, Z_τ^+ and Z_τ^- are the sum of the kernel k_τ over the sets S^C, S^B, S^+ and S^-, respectively. We call Z_τ^C, Z_τ^B, Z_τ^+ and Z_τ^- signature-specific scores or, simply, scores, since they summarize the evidence for a deletion from, respectively, the coverage process, the bracketing pairs, the hanging plus-strand pairs and the hanging minus-strand pairs. If κ were assumed constant, the scores for the hanging pairs simplify significantly to

$$Z_\tau^+ = \sum_{i:(x_i^+,x_i^-)\in S^+} \log\left\{1 + \frac{2r(1-p)}{p}\big[F\big(x_i^- - s + w - R\big) - F\big(x_i^- - s + w\big)\big]\right\},$$

$$Z_\tau^- = \sum_{i:(x_i^+,x_i^-)\in S^-} \log\left\{1 + \frac{2r(1-p)}{p}\big[F\big(s - x_i^+\big) - F\big(s - R - x_i^+\big)\big]\right\}.$$

From these simplified versions, we see that the hanging pairs scores are weighted counts of the hanging pair of the given type in the region before the start of the deletion (for Z^-) or after the end of the deletion (for Z^+), where the weights depend on the insert length distribution F.

The above reasoning can be easily modified to handle insertions. For testing the alternative of an insertion of width w between template positions s and $s+1$ in a proportion r of the chromosomes, we redefine the sets

$$S^C_{s,w} = \{(u,v) : s - R < u \le s \text{ or } s - R < v \le s\};$$
$$S^B_{s,w} = \{(u,v) : u \le s - R \text{ and } v > s\};$$
$$S^+_{s,w} = \{(u,v) : u = \infty \text{ and } v > s\};$$
$$S^-_{s,w} = \{(u,v) : u \le s - R \text{ and } v = \infty\}.$$

Then $\lambda_1(u,v)$ remains the same as (3.9) for S^C, and the same as (3.10) with $-w$ replaced by $+w$ for S^B. For the hanging minus-strand pairs,

$$\lambda_1(u,v) = \lambda_0(u,v)\left[1 + \frac{2r(1-p)}{p\int_u^m \kappa(x)f(x-u)\,dx}\int_{s-R}^{s+w} f(t-u)\kappa(t)\,dt\right], \tag{3.14}$$

and, for the hanging plus-strand pairs,

$$\lambda_1(u,v) = \lambda_0(u,v)\left[1 + \frac{2r(1-p)}{p\int_0^v \kappa(x)f(v-x)\,dx}\int_{s-w-R}^{s} f(v-t)\kappa(t)\,dt\right]. \tag{3.15}$$

REMARK 1. There is an important difference between insertions and deletions for the hanging read statistic. For insertions both Z^+ and Z^- should give a peak at the point of the insertion in the reference genome, hence they can be combined by addition. For deletions of the interval $(s, s+w)$, Z^- should give a peak at s, while Z^+ should give a peak at $s+w$. These two statistics will reinforce each other if w is small enough for the two peaks to overlap. Since w is unknown, alignment can be accomplished by maximizing the sum of the two statistics over a range of w values. To compensate for the increased number of multiple comparisons, one must use a higher significance threshold. As we shall see, the hanging read statistics are most useful for detecting short variants, where the bracketing pairs statistics have little power. The ideal range depends on the true value of w and on other unknown parameters. For simplicity in what follows we ignore this possibility and consider in our power studies the probability that the higher of the two peaks exceeds the appropriate threshold.

REMARK 2. Our focus in this paper is detection of insertions and small deletions, which are the hardest to detect by currently available methods. Coverage based statistics (e.g., Z^C) have low sensitivity for these types of variants, and so we ignore them in our analysis. In the Supplementary Materials, we include some

discussions about more complex types of variants, such as Tandem Duplications, Translocations and Inversions. These structural changes can be viewed as specific compositions of insertions and deletions. Whereas the scans we describe will likely reject the null and report a signal when the region contains such a complicated structural variant, they are not designed to identify the variant type correctly, nor are they optimized in terms of sensitivity. A further complicating scenario is when an individual carries two different structural variants at the same locus, one on each homologous chromosome copy. With so many possibilities, we recommend that instead of formulating every possible alternative, a practical strategy is to scan only for insertions and deletions, then apply a second stage classification of the signals by a more detailed modeling of the reads in the rejected regions.

REMARK 3. Like the simplified model of Z^B proposed in Section 3.1, it is also possible to develop a simplified model for the "hanging read" statistics, Z^+ and Z^-. If we assume that there is no variability in the insert lengths, that is, $\sigma = 0$, then for a mapped positive strand read beginning in the interval $[s - \delta, s - \delta + R]$ the corresponding negative strand will not map (a) whenever there is a deletion beginning at s or (b) with probability p, even if there is no deletion. Hence, a simple detection statistic would be obtained by counting the number of reads beginning in each interval of length R, the other end of which does not map, and claiming a detection of a deletion at s whenever the sum of positive strand reads that begin $[s - \delta, s - \delta + R]$ and negative strand reads that begin $[s + \delta - R, s + \delta]$ is too large to be determined by chance. An appropriate modification would serve to detect insertions. Some numerical experimentation suggests that this simplified test is less powerful than the more detailed likelihood-based procedure described above, although it might be more robust. Numerical examples are given in Section 5.1.

It is not a priori clear whether one should try to combine the scores S^C, S^B, S^+ and S^- into a single statistic, as in ℓ_τ, or treat them separately, for example, by a scan with only Z^B to target relatively long intervals and $Z^+ + Z^-$ to target short intervals, then use a Bonferroni bound to correct for using two different statistics. In Section 4.5 we will explore the sensitivity of the various types of scans.

4. Analysis of scan statistic from Section 3.1. We first consider scan statistics derived from (3.4) with kernel (3.6) corresponding to the simple mixture model in Section 3.1. We will derive their p-value approximations and study their power. Our p-value approximation approach relies on a measure transformation technique that shifts the distribution toward the desired alternative within the scan window. Some details of the method are given in the Supplementary Materials. See Siegmund, Yakir and Zhang (2010), Yakir (2013) and the references cited there for a more comprehensive illustration of this method and its applications to other problems.

In the notation introduced earlier, let

$$Z(t,w,r)=\int_y\int_{t-\delta}^{t} g(y;w,r)N(ds,dy).$$

The log-likelihood is given by [cf. (3.4)]

$$\ell(t,w,r)=\beta Z(t,w,r)-\Omega_\delta(t)\psi(\beta;w,r),$$

where $\Omega_\delta(t)=\int_{t-\delta}^{t}\rho(s)\,ds$ and $\psi(\beta;w,r)=\int\{\exp[\beta g(y;w,r)]-1\}\,dF_0(y)$.

The expected value of $\ell(t;w,r)$ can be expressed as $\Omega_\delta(t)J(\beta,w,r)$, where $J(\beta,w,r)=[\beta D_\beta\psi-\psi]$, with D_β denoting differentiation with respect to β. The parameter J is the Kullback–Leibler information.

We consider detection statistics of the form

$$\max Z(t;w,r) \tag{4.1}$$

and

$$\max \ell(t;w,r). \tag{4.2}$$

The maximum can extend over (t,w) or over (t,w,r) in some suitable range. We assume that t changes by discrete amounts $\Delta>0$. We start by considering arbitrary fixed values of w and r, and then explore the effect on power by considering a range of values for w, say $[w_0,w_1]$. We also consider maximization over a range of values of r, but our power calculations show that this maximization does not give a clear boost in sensitivity.

REMARK 4. The statistic (4.1) is essentially the scan statistic studied by Chan and Zhang (2007). Chan and Zhang, however, study specific "scoring" functions g that do not depend on unknown parameters. The rate $\rho(t)$ is also held constant, and thus under a fixed window size (4.1) is equivalent to (4.2). They do not consider a general maximum likelihood analysis of alternatives to the null model, and their calculations are equivalent to using what we have called the formal alternative with the value β defined by (4.3) below.

4.1. *Approximate p-values assuming homogeneity of Poisson rate.* Let $\rho(t)=\rho$ for all t, so $\Omega_\delta(t)=\rho\delta$ is independent of t. Assume β is chosen so that

$$\mathbb{E}_\tau\big[\ell(t;w,r)\big]=\rho\delta J(\beta;w,r)=x_0. \tag{4.3}$$

Then for large x_0 and $\rho\delta$, $\mathbb{P}_0\{\max_{1\le t\le m}\ell(t;w,r)\ge x_0\}\approx$

$$1-\exp\Big\{-me^{-x_0}\rho\big[\xi(\beta)-\xi(0)\big](2\pi\rho\delta)^{-1/2}\sigma(\beta)^{-1} \times \nu\Big(\frac{2\rho^{1/2}[\xi(\beta)-\xi(0)]}{[\sigma^2(\beta)+\sigma^2(0)]^{1/2}}\Big)\Big\}, \tag{4.4}$$

where $\xi(\beta)$ and $\sigma^2(\beta)$ are parameters of the local random walk, as explained in the Supplementary Materials; ν is the function defined in Siegmund (1985) and given approximately for purposes of numerical evaluation in Siegmund and Yakir [(2007), page 112]; and where we have for simplicity assumed that F_0 is a nonlattice distribution.

REMARK 5. The function $\nu(y)$ is always between 0 and 1 and approximately equals 1 for small values of $y > 0$. Although using the precise value of ν improves the quality of the approximation, in what follows we occasionally take ν identically equal to one. This simplifies some calculations, and numerical experimentation indicates that it rarely affects the power by more than a few percent.

The approximations when we also maximize over w or over w and r are more complicated. Consider, for example, the event

$$R = \Big\{ \max_{t, w_0 \le w \le w_1} \big[\ell(t; w, r) - x_{w,r} \big] \ge 0 \Big\}. \tag{4.5}$$

Let $J(\beta_w, w, r) = \mathbb{E}_\tau[\ell(t; w, r)]/(\rho\delta)$, with β_w chosen so that $\rho\delta J(\beta, w, r) = x_{w,r}$. Then $\mathbb{P}_0(R) \approx$

$$1 - \exp\Big\{ -m\rho \int_{w_0}^{w_1} \frac{\exp(-x_{w,r})[\xi(\beta_w) - \xi(0)]\nu\{\cdot\}[\Sigma(w)]^{1/2}}{2\pi(\rho\delta)^{1/2}\sigma(\beta_w)}\, dw \Big\}, \tag{4.6}$$

where the function ν has the same argument as in the preceding approximation, and where $\Sigma(w) = \mathbb{E}_\tau\{-D_w^2[\ell(t, w, r) - x_{w,r}]\}$.

If one were to also maximize over both w and r, the integral then becomes two-dimensional, $\Sigma = \Sigma(w, r)$ is the determinant of the expectation of the negative Hessian, and there is one more factor of $1/(2\pi)^{1/2}$. A similar result holds if there are more parameters. It appears that for the examples of this paper there are significant edge effects when one maximizes over r, so the first order asymptotic approximation given here may not be adequate, with resulting implications for the power.

REMARK 6. When we fix the values of the parameters w, r, the statistics $\max_t Z(t; w, r)$ and $\max_t \ell(\beta, w, r)$ are equivalent in the sense that a suitable threshold for one is a known linear function of the corresponding threshold for the other. This is not so if we maximize with respect to w or w, r; see Table 2 in Section 5.1.

4.2. *Approximation accuracy.* We have performed a small Monte Carlo experiment to evaluate the accuracy of (4.4). Results are given in the Supplementary Materials, where we find that the approximation is slightly conservative but fairly close to the Monte Carlo value, even when the approximation $\nu = 1$ is used (see Remark 5).

It is interesting to compare our approximations to results obtained using the theory of maxima of Gaussian random fields. Consider, for example, the scenario $m = 2000$, $\rho = 0.5$, $\delta = 200$, $w = 2$, $r = 0.2$, $\sigma = 1$ (row 8 of Supplementary Table 1). Our p-value for $x = 5$ is 0.056 (using $\nu = 1$) and 0.048 (using computed value of ν), whereas the Monte Carlo value is 0.044. The approximation in Siegmund and Yakir [(2007), page 112], which is known to be very accurate when the field is in fact Gaussian, would suggest that the p-value for this scenario is less than 0.01. The discrepancy becomes larger with larger m, since this pushes the threshold farther out into the tail of the distribution, where a Gaussian approximation can be extremely anti-conservative. For example, in the first row of Table 2 in Section 5.1 a 0.05 significance threshold based on the approximations of this paper would suggest a p-value of 3×10^{-8} if a Gaussian approximation is used. If we use the 0.05 threshold suggested by Gaussian theory, the actual false positive rate would be virtually one.

4.3. *Nonhomogeneous processes.* In the case that the underlying Poisson process is nonhomogeneous with intensity $\rho(t)$, the approximations given above apply with only slight modifications. Consider the case of fixed r. Since the measure transformation used to derive the approximations decomposes the boundary crossing probability into a sum of m terms, each depending on t, the expressions given in the exponents in the approximations (4.4) and (4.6) change to a sum of m terms, instead of a single expression multiplied by m. For the tth term, the definition of $\beta = \beta_{t,w}$ depends on both t and w, since $\rho\delta$ in the definition of the cumulant generating function and throughout the approximation is replaced by $\Omega_\delta(t)$. In addition, the factor $\rho[\xi(\beta_{t,w}) - \xi(0)]$ becomes $\rho(t)\xi(\beta_{t,w}) - \rho(t+\delta)\xi(0)$.

4.4. *Piecewise smooth processes.* We digress here to consider briefly the model (2.3) for ChIP-Seq data, where g_w is twice continuously differentiable. Assume that $Z_{t,w} = \int g_w(t-s)\,dN_s$. Let $\tau = (t, w)$ and consider $\mathbb{P}\{\max_{t,w} Z(t, w) \geq x_1\}$, where the max extends over $t_0 < t < t_1$ and $w_0 < w < w_1$. Then $\ell(t, w) = \beta Z(t, w) - \psi(\beta; t, w)$, where $\psi(\beta; t, w) = \int\{\exp[\beta g_w(t-s)] - 1\}\rho(s)\,ds$, $J(\beta; t, w) = -\int\{\exp[\beta g_w(t-s)] - 1 - \beta g_w(t-s)\}\rho(s)\,ds$, and $x_{w,t} = \beta x_1 - \psi(\beta; t, w)$. Putting $\tau = (t, w)$ and setting β to satisfy $\mathbb{E}_\tau[Z(t, w)] = x_1$, we find that the probability of interest is

$$\begin{aligned} &\approx \int_{t_0}^{t_1}\int_{w_0}^{w_1} \exp\bigl[-J(\beta; t, w)\bigr] \\ &\qquad \times \left[\frac{\mathbb{E}_\tau\{-D^2[\ell(t, w) - x_{w,t}]\}}{\beta^2 \operatorname{Var}_\tau \ell_t}\right]^{1/2} dw\,dt/(2\pi)^{3/2}. \end{aligned} \tag{4.7}$$

In the case $\rho(t) = \rho$ for all t, the integrand is a constant function of t, except near the end points, so the integral with respect to t can be simplified to multiplication by $t_1 - t_0$. See Siegmund and Worsley (1995) for justification and examples in the case of Gaussian processes. A version of this approximation is used for the statistics Z^+, Z^- for hanging reads, as shown in Section 3.3.

4.5. *Marginal power.* The statistics of the preceding section are all of the form $\max_\tau Y_\tau$. To study their power, suppose $\mathbb{P}\{Y_\tau \geq x\}$ is maximized at $\tau = \tau_0$ under the alternative. It seems reasonable to define the local power of the detection scheme to

$$\mathbb{P}\{Y_{\tau_0} \geq x\} + \mathbb{P}\Big\{Y_{\tau_0} < x, \max_\tau Y_\tau \geq x\Big\}. \tag{4.8}$$

Since the second term is usually small compared to the first, we define $\mathbb{P}\{Y_{\tau_0} \geq x\}$ to be the marginal power. In this section we consider again a homogeneous process and use the marginal power, evaluated by means of a normal approximation, to compare different scan designs.

In our numerical study we assume that $\delta = 200$, $m = 1{,}000{,}000$, $\rho = 0.5$ and maximize over $[0.5 < w \leq 5]$. We compare the marginal power of four different scans: (1) $Z = \max_{t,w} Z(t; w, r_0)$, (2) $\ell = \max_{t,w} \ell(t; w, r_0)$, where $r_0 = 0.1$, (3) $\ell_2 = \max_{t,w,r} \ell(t; w, r)$, where the maximum over r is restricted to the range $0.03 \leq r \leq 0.2$, and (4)

$$\ell(w_0, w_1; 0.1) = \max\Big[\max_t \ell(t, w_0, 0.1)/b_0, \max_t \ell(t, w_1, 0.1)/b_1\Big],$$

where $w_0 = 1.0$, $w_1 = 3.5$.

We assumed the standard deviation σ of the null insert size distribution F_0 is one. When σ is not 1, a shift of size w in our computations corresponds to an actual change of size σw.

For all statistics the significance level based on the approximations given above with $\nu = 1$ is about 0.05. For $\ell(2; 0.1)$, Z and ℓ, we obtained the thresholds 11.4, 11.54 and 12.05, respectively. For ℓ_2, the situation is more complicated, since the tail probability is largest at the largest values of r. Hence, as an overall approximation for the significance level, we have added, to the approximations involving the max over w and $r_0 \leq r \leq r_1$ given above, an edge correction involving max over w at r_1. The edge correction produced the threshold 12.87. For $\ell(w_0, w_1; 0.1)$ we used a Bonferroni bound to combine the two statistics, where $b_0 = 12.34$ and $b_1 = 11.9$ were chosen so that the individual statistics had 0.025 significance level. The column headed "Opt" gives the power for the statistic $\max_t \ell(t; w, r)$ for the indicated values of w, r and the 0.05 threshold (which depends on w, r and is omitted). Since this problem is statistically irregular, we do not know whether using the true parameters to define the log-likelihood ratio actually achieves maximum power. It does, however, seem a reasonable measure of the power that might be achieved with complete knowledge of the parameters. The statistic ℓ, which uses a nominal fixed value $r_0 = 0.1$ and is adaptive with respect to w, does remarkably well.

From these numbers it appears that when r is not too far from the assumed value, r_0, the statistic Z is slightly more powerful than ℓ, but it can be considerably less powerful when r is quite different from r_0. The statistic ℓ_2 seems less powerful than ℓ, even when the actual value of r is not close to the assumed value r_0. It

TABLE 1
Parameters are $\rho = 0.5$, $r = 0.1$, $\delta = 200$, $m = 1{,}000{,}000$; max *over* $w \in [0.5, 5]$, $r \in [0.03, 0.3]$

r_1	w_1	"Opt"	$\ell(2, 0.1)$	$Z(0.1)$	$\ell(1, 3.5; 0.1)$	ℓ	ℓ_2
0.1	2.5	0.53	0.52	0.52	0.47	0.50	0.45
0.1	3.0	0.82	0.80	0.81	0.80	0.80	0.78
0.1	2.25	0.32	0.32	0.31	0.21	0.29	0.24
0.3	1.4	0.54	0.43	0.38	0.47	0.50	0.46
0.3	2.0	0.96	0.96	0.95	0.95	0.96	0.96
0.5	1.0	0.63	0.31	0.26	0.48	0.59	0.54
0.5	1.5	0.99	0.96	0.95	0.98	0.98	0.98
0.03	4.0	0.55	0.37	0.43	0.46	0.50	0.47
0.03	4.5	0.70	0.51	0.58	0.61	0.66	0.64
0.02	5.0	0.64	0.38	0.48	0.50	0.58	0.55

is possible that the performance of ℓ_2 has been adversely affected by our *ad hoc* method of controlling the significance level. The statistic $\ell(1, 3.5; 0.1)$ is much simpler than ℓ and seems to be only slightly, but consistently less powerful over the range of parameters considered here.

The rate parameter ρ of the driving Poisson process is effectively the sample size, hence an important determinant of the power. Smaller values of ρ lead to lower significance thresholds but, even so, to less power.

5. Analysis of scan statistics from Section 3.3. We now consider the more detailed model to detect insertions and deletions proposed in Section 3.3. The log-likelihood ratio statistic under this model is a sum of the signature-specific scores. In practice, each score can be used on its own as a scan statistic or the scores can be summed in various combinations. Our power comparisons below show that the different scores achieve power in different regions of the parameter space. Although summing them improves power under very special conditions, overall it often results in a loss of power compared to applying each score individually and then adjusting the p-value by the Bonferroni inequality. Hence, we discuss p-value control only for the individual score Z_t^B and the sum of the hanging read scores, $Z^+ + Z^-$. The parameters w, r are set to fixed values in the case of deletions, where, to align peaks, we maximize over a range of w as discussed above. It would be possible to maximize these statistics with respect to the unknown parameters w, r, but some numerical experimentation shows that it is possible—and much simpler—to select robust values.

Consider first the score Z^B for detecting deletions using bracketing pairs. Here, the parameter τ is the triple (s, w, r). The kernel function corresponding to the alternative is $k_\tau(z) = \log[1 - r + rf(v - u + w)/f(v - u)]I(z \in S_{s,w}^B)$. It will

be convenient to put $g(x) = g(x; w, r) = \log[1 - r + rf(x + w)/f(x)]$, so the cumulant generating function of Z_τ^B is given by

$$\psi_\tau(\beta) = (1-p)\int_{u<s-R}\kappa(u)\int_{v>s+w}\kappa(v)f(v-u) \times \{\exp[\beta g(v-u)] - 1\}\,dv\,du, \tag{5.1}$$

which for a homogeneous process simplifies to $(1-p)\kappa^2\psi_0(\beta)$, where

$$\psi_0(\beta) = \int_{w+R}^{\infty}(x-w-R)f(x)\{\exp[\beta g(x)] - 1\}\,dx. \tag{5.2}$$

A similar analysis applies to insertions, in which case w is the negative of the insert size, the range of integration for v in (5.1) changes to $v > s$, and the range of integration in (5.2) changes to (R, ∞). Given the cumulant generating function, the false positive rate for a scan using Z_τ^B can be obtained along the lines of the results for the mixture model. In particular, for fixed w, r we have the approximation (4.4) with $\delta = 1$, since δ is incorporated into the definition of ψ_0.

If we model increased dispersion as suggested in (2.4), (5.2) becomes

$$\psi_0(\beta) = \int_{w+R}^{\infty}(x-w-R)f(x)\{[1+\alpha(\exp[\beta g(x)] - 1)]^j - 1\}\,dx/j\alpha. \tag{5.3}$$

For the mean of the local random walk the integral becomes

$$\int_{w+R}^{\infty} g(x)f(x)\{\exp[\beta g(x)][1+\alpha(\exp[\beta g(x)] - 1)]^{(j-1)} - 1\}\,dx.$$

Numerical examples for scans using Z_τ^B, with $R = 36$, $p = 0.03$, $\delta = 200$, $\sigma = 10$, indicate that the statistics behave similarly to those discussed for the toy mixture model. One distinction worth noting between insertions and deletions in this model is that while power increases with the length of a deletion, it can decrease for insertions when w becomes a substantial fraction of the insert length, since an insert must span the insertion in the target genome for the read pair to be informative.

Now consider the scores Z^+ and Z^-, or their sum, which uses hanging reads for detection. For example, the kernel function for Z^- is $k_\tau(z) = \log(1 + 2rp^{-1}(1-p)\{F(s-u) - F(s-R-u)\})I(z \in S_{s,w}^-)$. Since these kernels are continuous and vanish at $\pm\infty$, we can use a version of the approximations (4.7), modified as described in Remark 1.

5.1. *Power comparison.* We now examine the power of the scans based on the bracketing pair score Z^B and the hanging pair score $Z^H = Z^+ + Z^-$. The marginal power is computed as described in Section 4.5.

The sequencing and library preparation parameters that influence power are the length of the read, R, the mean δ and standard deviation σ of the insert length distribution, and the sequencing coverage (that is, the average value of $R\kappa^2$). Power of

the hanging reads score also depends heavily on the value of p, the probability of a mapping error leading to a hanging read under the null hypothesis. Together, these parameters determine the null distribution. Table 2 in the Supplementary Materials shows the value of R and the estimated values of δ, σ and p for a few typical publicly available data sets. Earlier sequencing data sets had read length 36; the read length is now 100 and expected to increase further. Insert sizes usually range between 200 and 1000. Increased insert lengths come at a cost of an increased standard deviation, and thus the overall effect on the power of the trend toward longer inserts is not so clear. The rate of hanging reads also varies widely between data sets, and is usually between 0.001 and 0.05. We will analyze power under two settings that we found in the empirical data: $R = 36$, $p = 0.03$, $\delta = 200$, $\sigma = 10$ and $R = 100$, $p = 0.033$, $\delta = 220$, $\sigma = 63$. We also consider a few examples with longer insert lengths and smaller p.

Although the values of R, p, δ and σ more or less fall within standard ranges, depth of coverage can vary widely across studies, depending on the goals of the experiment. "Low-coverage" usually refers to cases where each genomic position is covered by an average of 10 reads or less, and "high coverage" to cases where each genomic position is covered by an average of 40 reads or more. In some studies, for example, in the sequencing of virus populations, extremely high coverage in the hundreds or thousands is desired. These are so-called "deep sequencing" experiments, where the mutations of interest are sometimes present at very low frequencies ($r < 1\%$) in the sample.

We will examine two scenarios for coverage. For the first, $\kappa^2 = 0.27$, which represents coverage of $\sim 30\times$ when read length is 100 bases. This scenario is a common setting in current sequencing experiments. For the second, $\kappa^2 = 5$ and we study only the $R = 100$ setting, which represents deep sequencing with an average coverage of 500. The latter setting is common in the sequencing of bacterial and viral genomes, as well as in targeted sequencing experiments. In both cases, we let $m = 1{,}000{,}000$. Larger values of m are likely to occur in practice, but do not seem to yield additional insights.

We found through simulations and numerical studies that, as for the toy mixture model, the power of the scores is not particularly sensitive to the *assumed* values of r and w used to define the scores. Thus, for simplicity, we set $w = 30$ base pairs in Z^H for insertion, and $|w| = 30$ base pairs for Z^B. The assumed value of r is set to 0.1 in all statistics.

For the most part, power increases with the true size of the signal (w) and its true frequency in the sample (r), both of which are properties of the alternative distribution. These are chosen so that at least one of Z^B and Z^H has moderate power. We expect to find that Z^H has relatively more power than Z^B to detect short variants, and relatively less to detect longer variants.

Consider first the case of insertions under moderate coverage ($\kappa^2 = 0.27$). Marginal power for varying values of (r, w) is given in Table 2. Observe that when $R = 36$, Z^B has better power than Z^H when the insertion is large, provided that it

TABLE 2
Marginal power: insertions

R, δ, σ, p	r	w	Hanging reads	Bracketing pairs
36, 200, 10, 0.03	0.5	10	0.94	0.00
36, 200, 10, 0.03	0.5	20	0.98	0.74
36, 200, 10, 0.03	0.5	100	1.00	1.00
36, 200, 10, 0.03	0.2	50	0.66	0.71
36, 200, 10, 0.03	0.1	100	0.11	0.80
36, 200, 10, 0.03	0.1	150	0.11	0.51
100, 220, 63, 0.033	0.5	10	1.00	0.00
100, 220, 63, 0.033	0.5	100	1.00	0.00
100, 220, 63, 0.033	0.1	100	0.46	0.00
100, 220, 63, 0.033	0.1	200	0.91	0.00
100, 220, 63, 0.01	0.1	200	0.79	0.00
100, 220, 63, 0.033	0.2	10	0.70	0.00
100, 220, 63, 0.01	0.2	10	0.94	0.00
100, 220, 63, 0.033	0.2	100	0.97	0.00
100, 400, 63, 0.033	0.5	100	1.00	0.72
100, 400, 63, 0.033	0.3	200	1.00	0.21

is still substantially smaller than the value of δ. The statistic Z^H has better power than Z^B to detect short insertions at high frequency. When $R = 100$, Z^B has no power in the situations studied here, except for the case where the mean insert length was 400.

Table 3 shows the power for detecting deletions of varying (r, w) when coverage is moderate. The power of Z^B for deletions increases monotonically with the size of the deletion. In comparison to Z^H, Z^B has better power when deletion size is large and when r is small. As in the case of insertions, Z^H has better power for longer reads, and it is preferable to Z^B when the true value of r is large. The power of Z^H does not depend on the size of the deleted region. Note that the larger value of σ that appears to be a concomitant of the larger values of R and δ can lead to a loss of power for Z^B.

We can also consider the effect on the significance threshold of the model of excess variability. Assume that each fragment is produced in a binomial number of copies, conditioned to be at least one, with parameters j, α. For example, let $j = 5$ and $\alpha = 0.03$, which leads to a roughly 15% increase in the null variance of the statistics. For the bracketing pair statistic, the true false positive rate would increase from 0.05 to about 0.25; and it would be about the same for the hanging read statistic. Hence, a small amount of excess variability, if not properly accounted for, can lead to a substantial increase in the false positive errors.

Finally, we consider the case of detecting low-frequency mutations using deep sequencing ($R = 100$, $\delta = 400$, $\sigma = 63$, $p = 0.033$, $\kappa^2 = 5$). We consider only deletions and examine the setting where the length of the deletion is large and the

TABLE 3
Marginal power: detecting deletions with $\kappa^2 = 0.27$, $m = 10^6$

R, δ, σ, p	r	w	Hanging reads	Bracketing pairs
36, 200, 10, 0.03	0.5	10	0.71	0.00
36, 200, 10, 0.03	0.5	20	0.71	0.84
36, 200, 10, 0.03	0.5	100	0.71	1.00
36, 200, 10, 0.03	0.1	100	0.00	0.99
36, 200, 10, 0.03	0.1	150	0.00	0.99
36, 200, 10, 0.03	0.05	150	0.00	0.96
36, 200, 10, 0.03	0.01	150	0.00	0.64
36, 200, 10, 0.03	0.01	250	0.00	0.75
100, 220, 63, 0.033	0.5	10	0.99	0.00
100, 220, 63, 0.033	0.3	10	0.80	0.00
100, 220, 63, 0.033	0.3	100	0.80	0.40
100, 220, 63, 0.033	0.3	150	0.80	0.95
100, 220, 63, 0.033	0.2	150	0.36	0.75
100, 400, 63, 0.033	0.2	100	0.36	0.35
100, 400, 63, 0.01	0.2	100	0.81	0.36
100, 400, 63, 0.01	0.2	150	0.81	0.94

frequency is small. Table 4 shows the marginal power for varying (r, w). Compared to Table 3, we see that Z^B is more competitive against Z^H in the high depth, low r, large w scenario.

6. Structural variant detection.

6.1. *Comparisons to mainstream algorithms on spike-in data.* Simulated and real data sets were prepared to evaluate the effectiveness of the likelihood-based approach to structural variant detection described in Section 3.3. The simulation imitates a real sequencing experiment by taking a one megabase region of the human reference genome hg19, adding deletions and insertions in silico, fragmenting

TABLE 4
Marginal power: detecting deletions with $R = 100$, $\delta = 400$, $\sigma = 63$, $p = 0.033$, $\kappa^2 = 5$

r	w	Hanging reads	Bracketing pairs
0.10	5	1.00	0.00
0.07	50	0.99	0.02
0.07	100	0.99	0.96
0.05	150	0.80	1.00
0.02	200	0.02	0.93
0.01	250	0.00	0.71

the resulting sequence and mapping the fragment ends back to hg19 (more details in the supplement). This way, the type, position, length and mixture proportion of the variants can be controlled. On the spike-in data set, the accuracy of a given algorithm can be measured by precision and sensitivity, defined as

$$\text{Precision} = \frac{\#\text{True Positives}}{\#\text{ Positives}} \times 100\%; \qquad \text{Sensitivity} = \frac{\#\text{ True Positives}}{\#\text{True}} \times 100\%.$$

Precision is the true discovery rate, with the positive and true positive counts constrained to signals of a certain type and size. For each program and each type of signal, we also computed the combined accuracy, which is the harmonic mean of precision and sensitivity. Our likelihood-based approach, as implemented under the default setting in the software SWAN, is compared to four existing algorithms: BreakDancer [Chen et al. (2009)], CNVnator [Abyzov et al. (2011)], Delly [Rausch et al. (2012)] and Pindel [Ye et al. (2009)]. Among the many existing algorithms for calling structural variants, these four algorithms were shown by the 1000 Genomes Project [The Genomes Project Consortium (2015)] to be competitive and have become mainstream.

First, we compared the accuracy of SWAN to existing approaches for detecting homozygous variants. As shown in the left panel of Figure 5, SWAN is the only program capable of detecting medium to large insertions (100 bp to 10 kbp), while also maintaining high precision and sensitivity for small insertions and deletions of all sizes. BreakDancer can find medium to large deletions, but breaks down for small deletions. CNVnator is sensitive for large deletions, as expected since it relies only on changes in coverage. Delly uses both hanging reads and insert size, which allows it to detect deletions as small as 50 bp, although its accuracy is low for small deletions. Finally, we found Pindel to be more accurate than the other existing methods, as reported by the authors [Ye et al. (2009)]. Like SWAN, Pindel can detect different size deletions while also capturing some very small insertions with size around half of the read length. However, unlike SWAN, Pindel has no power for medium to large insertions (size $>$ read length). We attribute the wide spectrum accuracy of SWAN to the fact that it combines multiple signature-specific scores. When one signal is missing or weak, SWAN relies on other signals.

Next, we examined the performance of SWAN when only a half or a quarter of the DNA in the sample contains a mutation. The half (50%) setting corresponds to the heterozygous case in diploid samples, while the quarter (25%) setting corresponds to more difficult cases such as tumor samples where the variant is a somatic mutation present in only a sub-clone of cells. The results are summarized in the right panel of Figure 5. For most of the tested variant types, including all sizes of insertions and medium to large deletions, SWAN maintains $>$90% in accuracy, precision and sensitivity. This accuracy holds for both half and quarter mixtures. For small deletions (100 bp, roughly 3 times the insert size standard deviation), we see a drop in the sensitivity of SWAN to approximately 50% for the quarter mixture. For very small deletions (50 bp), we see a drop in sensitivity for the 50%

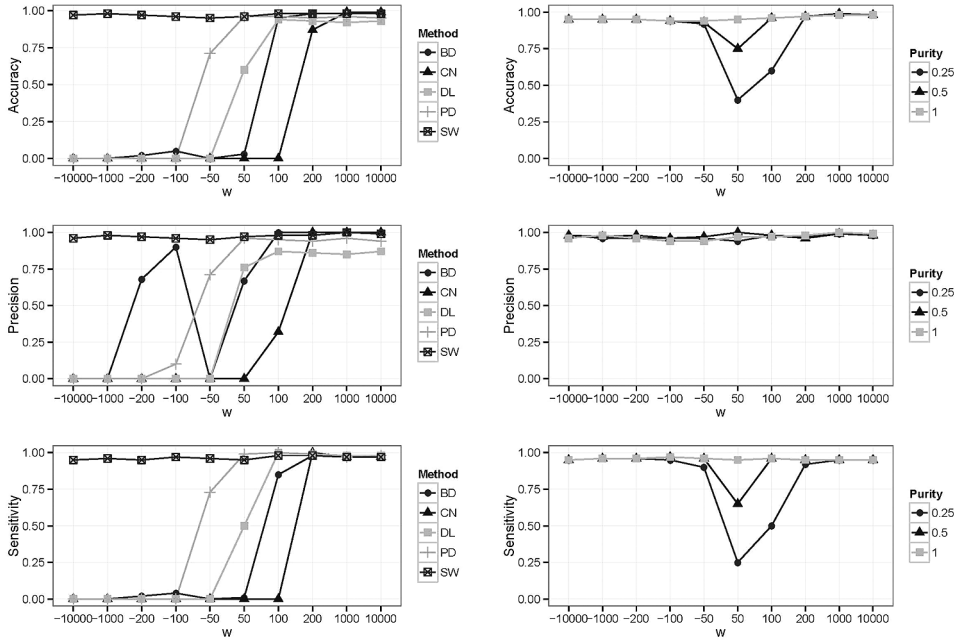

FIG. 5. *The plots on the left show the F-1 accuracy, precision and sensitivity of five programs (BD = Breakdancer, CN = CNVnator, DL = Delly, PD = Pindel and SW = SWAN) on spike-in data. The variant size is coded as negative for insertions and positive for deletions. The plots on the right show the F-1 accuracy, precision and sensitivity of SWAN for variants at varying purity levels.*

mixture and an additional drop at the 25% setting. Therefore, in this low-noise setting, SWAN is sensitive for all except very small deletions.

6.2. *Analysis of platinum genomes trio data.* We also analyzed a family trio comprised of NA12877 (father), NA12878 (mother) and NA12882 (son) that was sequenced as part of the Platinum Genomes project at Illumina [Eberle et al. (2014)] on the HiSeq 2000 system. The BAM files were download from the EMBL-EBI website with study accession PRJEB3381 and sample accessions ERS189473, ERS189474 and ERS189490. These data, whose summary statistics are described in Table 2 of the Supplementary Materials, are of very high quality by most sequencing metrics. The insert size distributions of the three samples are unimodal and normal shaped. For simplicity we restrict our analysis to the insert size score.

Since the majority of our insertion/deletion discoveries are less than one kb while existing validated structural variants are mainly of larger sizes [The Genomes Project Consortium (2015)], we will use the familial relationship in the trio data to assess the accuracy of SWAN detections. In total, SWAN reported 8874 events comprised of 8813 deletions and 61 insertions in the son (NA12882), 9235 events comprised of 9150 deletions and 85 insertions in the father (NA12877)

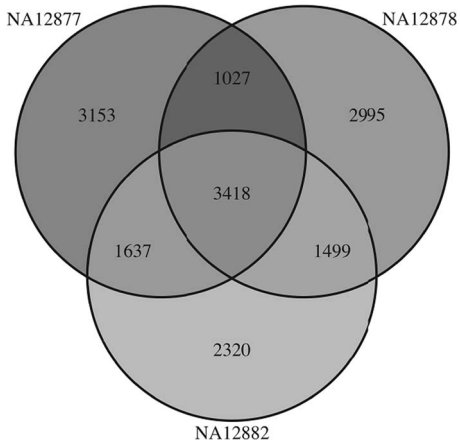

FIG. 6. *The total and overlapping number of detections made by SWAN on the samples NA*12877, *NA*12878 *and NA*12882. *NA*12877 *and NA*12878 *are parents to NA*12882.

and 8939 events comprised of 8859 deletions and 80 insertions in the mother (NA12878); see Figure 6. Among the deletions, the frequency of deletions diminishes rapidly as the length increases; see Figure 1 in the Supplementary Materials. While the size distribution of structural variants (SVs) is not well characterized for smaller deletions and insertions, results from Chaisson et al. (2015) show a similar rapid decrease (for events >1 kb). Our findings support their claim and also imply that an exponential rate of decrease remains true for smaller deletions. This observation agrees with Li et al. (2011), where the authors used de novo assembly to find small SVs and reported an exponential rate of decrease for SVs less than 1 kb in length. Results from the 1000 Genome Project [The Genomes Project Consortium (2015)] have a similar exponential trend in deletion size.

Another interesting aspect is the sharing of SV events between different pairs of the trio. As shown in the Venn Diagram of Figure 6, 57% and 55.4% of the 8874 events detected in the child overlap with those made in the mother and the father, respectively. (The Illumina provided VCF files show overlapping SNP calls are 77% and 78% for the same data.) In total, 77% of the detections in the child have overlap in one or both parents. In contrast, of the detections made in the mother, about 49% overlap with those made in the father, and of the detections made in the father, 48% overlap with calls made in the mother. From familial overlap analysis, it is impossible to estimate sensitivity and specificity, since a SWAN call in the child which is not shared in the parents may be a false positive in the child, a false negative in one or both parents, or a true de novo SV in the child. Nevertheless, the fact that we have a significantly higher overlap between the child and at least one parent, compared to the overlap between parents, lends some support for our results.

7. Summary and discussion. We studied scan statistics for Poisson-type data, with emphasis on several statistics that are useful for detecting local genomic signals in next-generation sequencing experiments, in particular, for structural variant detection by paired-end whole genome sequencing. Despite their different formulations, analytic significance approximations for these statistics can be obtained through a general framework that involves embedding the statistics into an exponential family (3.3) and applying the measure transformation technique described in Siegmund, Yakir and Zhang (2011); see also Yakir (2013).

We analyzed in detail a mixture model which may be viewed as a simplified version of the model for bracketing read pairs (3.10) but may also be of broader interest. We developed approximations for the significance level and power which reveal a complex picture regarding the dependence of power on the choice of scanning parameter(s), the assumed homogeneity of the process and the values of nuisance parameters. The key observations are summarized in Section 4.5.

For structural variant detection using paired-end sequencing, we formulated a model that incorporates three different features of the data: Read coverage, mapped insert length and hanging read pairs. While the bracketing pairs statistics have increasing power to detect longer deletions, their power to detect insertions first increases, then decreases with the length of the insertion. The power of the hanging read statistics to detect deletions does not depend on the length of the deletion, while their power to detect insertions increases with the length of the insertion and approaches an asymptote typically less than one.

In the empirical data that we have examined, fragments with larger mean insert lengths also have substantially larger standard deviations. Also, such libraries tend to exhibit skewness and sometimes even multimodality. A consequence of the contemporary move to increase read and insert length is that, relatively speaking, the hanging read statistics gain power, but the bracketing pairs statistics can lose substantial power.

Our analyses in Section 5 assume constant, known read coverage κ, although read coverage fluctuates along the genome. If we allow κ to vary, the thresholds would change with genomic position. A compromise might be to segment the genome into blocks of approximately homogeneous read coverage, then scan each block separately. The global p-value would then be computed from the sum of the block-wise p-values. In implementing this approach, one may want to ignore regions of low coverage, since a substantial amount of power is inevitably lost in those regions and cannot be recovered by adjustment of the significance threshold. It is also straightforward to adapt our threshold analysis to the case where the regional read coverage is estimated, but, unless the region is very short (on the order of a few hundreds of base pairs), accounting for variability in the estimate of the background rate has negligible impact.

An open question is whether or how different statistics should be combined to improve detection accuracy. We found in our power analysis that summing the

scores, as in the log-likelihood, is rarely better than applying each score individually. The reason is that for most alternative settings there is one score that is substantially more powerful than the others, and incorporating the others by simple addition contributes mainly noise. Thus, it appears to be better to apply each score individually and then combine detections using a Bonferroni correction.

While we have been focusing mainly on control of the family-wise error rate, in genomic studies the false discovery rate (FDR) is often an appealing mode of multiple testing control. The boundary crossing probabilities can be converted into the expected number of false discoveries under the null and used for FDR control as described in Siegmund, Zhang and Yakir (2011).

Acknowledgments. We would like to thank the Editor and Associate Editor who handled this paper for their attention to this paper and constructive comments.

SUPPLEMENTARY MATERIAL

Supplement to "Scan statistics on Poisson random fields with applications in genomics" (DOI: 10.1214/15-AOAS892SUPP; .pdf). Detailed comments on complex structural variants, moment calculations, expectation and covariance structure of the local field, details of p-value calculations and Monte Carlo accuracy evaluations of Section 4.2. Also contains details of spike-in experiment and summaries of real sequencing data sets.

REFERENCES

ABYZOV, A., URBAN, A. E., SNYDER, M. and GERSTEIN, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. **21** 974–984.

ADLER, R. J. and TAYLOR, J. E. (2007). *Random Fields and Geometry*. Springer, New York. MR2319516

BRAUN, J., DUMM, J., DE PALMA, F., FINLEY, C., KARLE, A. and MONTARULI, T. (2008). Methods for point source analysis in high energy neutrino telescopes. *Astroparticle Physics* **29** 299–305.

CAMPBELL, P. J., STEPHENS, P. J., PLEASANCE, E. D., O'MEARA, S., LI, H., SANTARIUS, T., STEBBINGS, L. A., LEROY, C., EDKINS, S., HARDY, C., TEAGUE, J. W., MENZIES, A., GOODHEAD, I., TURNER, D. J., CLEE, C. M., QUAIL, M. A., COX, A., BROWN, C., DURBIN, R., HURLES, M. E., EDWARDS, P. A. W., BIGNELL, G. R., STRATTON, M. R. and FUTREAL, P. A. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics* **40** 722–729.

CHAISSON, M. J. P., HUDDLESTON, J., DENNIS, M. Y., SUDMANT, P. H., MALIG, M., HORMOZDIARI, F., ANTONACCI, F., SURTI, U., SANDSTROM, R., BOITANO, M., LANDOLIN, J. M., STAMATOYANNOPOULOS, J. A., HUNKAPILLER, M. W., KORLACH, J. and EICHLER, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517** 608–611.

CHAN, H. P. and ZHANG, N. R. (2007). Scan statistics with weighted observations. *J. Amer. Statist. Assoc*. **102** 595–602. MR2370856

CHEN, K., WALLIS, J. W., MCLELLAN, M. D., LARSON, D. E., KALICKI, J. M., POHL, C. S., MCGRATH, S. D., WENDL, M. C., ZHANG, Q., LOCKE, D. P., SHI, X., FULTON, R. S., LEY, T. J., WILSON, R. K., DING, L. and MARDIS, E. R. (2009). BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6** 677–681.

CHIANG, D. Y., GETZ, G., JAFFE, D. B., O'KELLY, M. J. T., ZHAO, X., CARTER, S. L., RUSS, C., NUSBAUM, C., MEYERSON, M. and LANDER, E. S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6** 99–103.

EBERLE, M. A., KALLBERG, M., CHUANG, H.-Y., TEDDER, P., HUMPHRAY, S., BENTLEY, D. and MARGULIES, E. H. (2014). Platinum Genomes: A systematic assessment of variant accuracy using a large family pedigree. In *ASHG 2013 Annual Meeting*.

FEINGOLD, E., BROWN, P. O. and SIEGMUND, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am. J. Hum. Genet.* **53** 234–251.

FLICEK, P. and BIRNEY, E. (2009). Sense from sequence reads: Methods for alignment and assembly. *Nat. Methods* **6** S6–S12.

THE GENOMES PROJECT CONSORTIUM (2015). A global reference for human genetic variation. *Nature* **526** 68–74.

GROSS, E. and VITELLS, O. (2010). Trial factors for the look elsewhere effect in high energy physics. *The European Physical Journal C* **70** 525–530.

KARLIN, S., DEMBO, A. and KAWABATA, T. (1990). Statistical composition of high-scoring segments from molecular sequences. *Ann. Statist.* **18** 571–581. MR1056327

KULLDORFF, M. (1997). A spatial scan statistic. *Comm. Statist. Theory Methods* **26** 1481–1496. MR1456844

KULLDORFF, M. (1999). Spatial scan statistics: Models, calculations, and applications. In *Scan Statistics and Applications* (J. Glaz and N. Balakrishnan, eds.) 303–322. Birkhäuser, Boston, MA. MR1697758

KULLDORFF, M. and NAGARWALLA, N. (1995). Spatial disease clusters: Detection and inference. *Stat. Med.* **14** 799–810.

LANDER, E. S. and BOTSTEIN, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121** 185–199.

LI, Y. et al. (2011). Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat. Biotech.* **29** 723–730.

MEDVEDEV, P., STANCIU, M. and BRUDNO, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **6** S13–S20.

PENG, J. and SIEGMUND, D. (2005). The admixture model in linkage analysis. *J. Statist. Plann. Inference* **130** 317–324. MR2128010

PEPKE, S., WOLD, B. and MORTAZAVI, A. (2009). Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* **6** S22–S32.

RABINOWITZ, D. and SIEGMUND, D. (1997). The approximate distribution of the maximum of a smoothed Poisson random field. *Statist. Sinica* **7** 167–180. MR1441152

RAUSCH, T., ZICHNER, T., SCHLATTL, A., STÜTZ, A. M., BENES, V. and KORBEL, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28** i333–i339.

SCHWARTZMAN, A., JAFFE, A., GAVRILOV, Y. and MEYER, C. A. (2013). Multiple testing of local maxima for detection of peaks in ChIP-Seq data. *Ann. Appl. Stat.* **7** 471–494. MR3086427

SHEN, J. J. and ZHANG, N. R. (2012). Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. *Ann. Appl. Stat.* **6** 476–496. MR2976479

SIEGMUND, D. (1985). *Sequential Analysis*: *Tests and Confidence Intervals*. Springer, New York. MR0799155

SIEGMUND, D. O. and WORSLEY, K. J. (1995). Testing for a signal with unknown location and scale in a stationary Gaussian random field. *Ann. Statist.* **23** 608–639. MR1332585

SIEGMUND, D. and YAKIR, B. (2007). *The Statistics of Gene Mapping*. Springer, New York. MR2301277

SIEGMUND, D., YAKIR, B. and ZHANG, N. (2010). Tail approximations for maxima of random fields by likelihood ratio transformations. *Sequential Anal.* **29** 245–262. MR2747524

SIEGMUND, D., YAKIR, B. and ZHANG, N. R. (2011). Detecting simultaneous variant intervals in aligned sequences. *Ann. Appl. Stat.* **5** 645–668. MR2840169

SIEGMUND, D. O., ZHANG, N. R. and YAKIR, B. (2011). False discovery rate for scanning statistics. *Biometrika* **98** 979–985. MR2860337

SONG, K., REN, J., ZHAI, Z., LIU, X., DENG, M. and SUN, F. (2013). Alignment-free sequence comparison based on next-generation sequencing reads. *J. Comput. Biol.* **20** 64–79. MR3021670

TANG, H. K. and SIEGMUND, D. (2001). Mapping quantitative trait loci in oligogenic models. *Biostatistics* **2** 147–162.

TOTH, B., LILLO, F. and FARMER, J. D. (2010). Segmentation algorithm for non-stationary compound Poisson processes. *Eur. Phys. J. B* **78** 235–243.

WORSLEY, K. J., EVANS, A. C., MARRETT, S. and NEELIN, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *J. Cereb. Blood Flow Metab.* **12** 900–918.

YAKIR, B. (2013). *Extremes in Random Fields*: *A Theory and Its Applications*. Wiley, Chichester. MR3241226

YE, K., SCHULZ, M. H., LONG, Q., APWEILER, R. and NING, Z. (2009). Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25** 2865–2871.

ZHANG, N. R., YAKIR, B., XIA, L. C. and SIEGMUND, D. (2016). Supplement to "Scan statistics on Poisson random fields with applications in genomics." DOI:10.1214/15-AOAS892SUPP.

N. R. ZHANG
DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19004
USA
E-MAIL: nzh@wharton.upenn.edu

L. C. XIA
DEPARTMENT OF MEDICINE
STANFORD UNIVERSITY SCHOOL OF MEDICINE
69 CAMPUS DRIVE
STANFORD, CALIFORNIA 94304
USA
E-MAIL: li.xia@stanford.edu

B. YAKIR
DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY OF JERUSALEM
JERUSALEM 91905
ISRAEL
E-MAIL: msby@mscc.huji.ac.il

D. SIEGMUND
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL
390 SERRA MALL
STANFORD, CALIFORNIA 94305-4065
USA
E-MAIL: dos@stat.stanford.edu