A note on optimal sampling strategy for structural variant detection using optical mapping

Weiwei Li[a], Jan Hannig[a], and Corbin D. Jones[b]

[a]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; [b]Department of Biology and Integrative Program for Biological & Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

ABSTRACT

Structural variants compose the majority of human genetic variation, but are difficult to accurately assess using current genomic sequencing technologies. Optical mapping technologies, which measure the size of chromosomal fragments between labeled markers, offer an alternative approach. As these technologies mature toward becoming clinical tools, there is a need to develop an approach for determining the optimal strategy for sampling biological material in order to detect a structural variant at some threshold. Here we develop an optimization approach using a simple, yet realistic, model of the genomic mapping process using a hypergeometric distribution and probabilistic concentration inequalities. Our approach is both computationally and analytically tractable and includes a novel approach to getting tail bounds of hypergeometric distribution. We show that if a genomic mapping technology can sample most of the chromosomal fragments within a sample, comparatively little biological material is needed to detect a variant at high confidence.

ARTICLE HISTORY

Received 7 October 2019
Accepted 23 January 2020

KEYWORDS

Structural variant detection; hypergeometric distribution; concentration inequality

1. Introduction

Structural variants (SV), insertions, deletions, trans-locations, copy number variants, are by far the most common types of human genetic variation (Chaisson, Wilson, and Eichler 2015). They have been linked to large number of heritable disorders (Hurles, Dermitzakis, and Tyler-Smith 2008). Technologies to assay the presence or absence of these variants have steadily improved in ease and resolution (Audano et al. 2019; Huddleston and Eichler 2016). Whole genome shotgun (WGS) DNA sequencing can detect small variants (less than 10 bp) readily and can detect some classes of large SV. This approach, however, is inferential and often struggles to capture copy number variation in gene families or to correctly estimate the size of insertions. An alternative approach, genomic mapping (such as the technology of BioNano Genomics), addresses the deficiencies of WGS by providing linkage and size information from ordered fragments of chromosomes spanning tens to hundreds of kilobases. In contrast to WGS, genomic mapping approaches directly observe SV, rather than inferring the existence of a SV from patterns of mismatch in WGS data. In the near future, these genome

CONTACT Weiwei Li weiweili@live.unc.edu Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

mapping technologies are expected to be used for clinical diagnosis of SV known to be associated with genetic disorders.

In a clinical setting, the cells or tissues needed for analysis may be hard to obtain, which poses several important statistical questions: what is the minimum amount of starting material necessary to have some confidence of detecting a target fragment? What is the optimal sampling strategy for the primary and derived material throughout the process? How best to model the technical errors—such as failure to digest at a site—during the processing of the data as these errors can lead to false positives and negatives? As is often the case, answering these questions motivated an exploration and expansion of the statistical machinery used to model this biological process.

Our contributions are twofold. From the algorithmic perspective, we explored, both theoretically and empirically, the connection between hypergeometric distribution and binomial distribution. We showed that under certain conditions, the tail bounds of binomial distribution can be used to control that of hypergeometric distribution. From the clinical and experimental perspective, we built an extensible model for estimating the amount of material needed for optical mapping of a genome. As these technologies move into clinical practice–such as diagnostics for chromosome abnormalities—there is critical need to be able to determine if enough genomic material is available for applying this assay.

The rest of this paper is organized as follows: in Section 2, we describe the statistical modeling of the sampling problem and our sampling strategy, in Section 3, we introduce the implementation details of our sampling algorithm, in Section 4 we present our numerical results on synthetic data sets, in Section 5 we summarize the conclusions of our paper. Proofs are relegated to the Appendix.

2. Statistical model

In this section, we abstract our sampling procedure into an "urn sampling" model. As DNA is processed through the optical mapping procedure, we imagine the material passing through a series of urns. Assume we have 46 different types of long sequences (i.e., chromosomes), each type has n copies (i.e., n cells), so we have $46n$ long sequences in total. We assume only one type of long sequences contains the target sequence, or

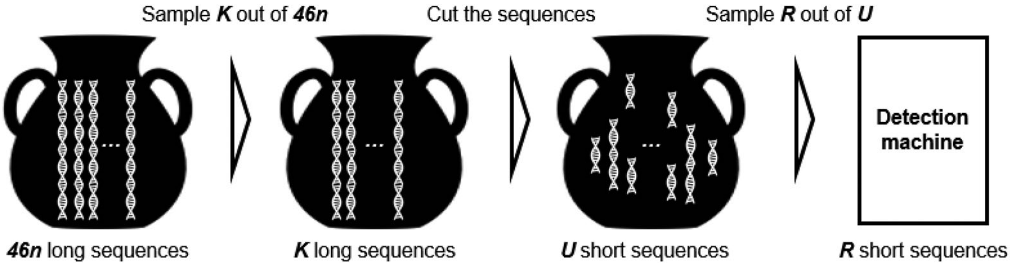

Figure 1. Three urn demonstration of the algorithm: the first urn contains raw biological materials. The second urn contains materials sampled from the first urn. The third urn contains materials from the second urn that are cut into shorter segments. Content of the third urn is sampled and assayed in the detection machine.

Table 1. Non random quantities.

Notation	Definition
n	Number of cells in the first urn (copies of each type of long sequences)
K	Number of sequences sampled from the first urn
R	Number of sequences sampled from third urn
L	Approximated length of long sequence
l	Approximated length of short sequence
T	Threshold on detectability of target sequences
f	Length of fragment of interest
c	Approximated ratio between lengths of long and short sequences
Q	Minimum number of target sequences we want in the detection machine
p	Minimum confidence in achieving the goal

the fragment of interest. The basic idea of our sampling model is shown in Figure 1. The notations introduced below are summarized in Table 1.

The first urn contains our original biological material, total of $46n$ long sequences out of which n of them contain the target sequence. At the first stage, we sample K sequences without replacement from the first urn, and put them in the second urn. The second urn will therefore contain a random number X of target sequences. All of the K long sequences in second urn are cut (a.k.a. nicked and labeled) at random locations according to a Poisson process and placed into the third urn. The third urn will therefore contain a random number of U sequences out of which W are target sequences. The content of the third urn models the biological material prepared for assay in a detection machine. Finally, we sample R smaller sequences without replacement out of the third urn and put them into a detection machine. There will be a random number Y of target sequences processed by the detection machine, and the goal is to assure that for some pre-specified values Q and p, we have the probability of $Y \geq Q$ is at least p. Throughout the experiment, the variables (n, K, R) are in our control and we will find the conditions on them to achieve our goal. Throughout this paper, we call the long sequence in the second urn which contains the fragment of interest as "target sequence."

Next we state the following biological assumptions:

(1) The length of target sequence is f.
(2) The lengths of long sequences in the first urn are approximately L, here $L \gg max(f, T)$.
(3) Short sequences in the third urn have lengths approximately l, and $c \approx \frac{L}{l}$.

We proceed by describing the probabilistic parts of our model. The distributions and their expectations are summarized in Table 2. There are X target sequences in the second urn. It is straightforward to see $X \sim H(46n, n, K)$, a hypergeometric distribution with $46n$ samples, n samples of interest and K as sampling size.

Let U_i $(i = 1, 2, ..., K)$ denotes the number of cuts on ith long sequence in the second urn. Combine with the third assumption above, we assume that U_i follows a Poisson distribution with mean c. Note that U_i cuts divide the sequence into $(U_i + 1)$ shorter sub-sequences. Consequently, $U = \sum_{i=1}^{K}(U_i + 1)$ is the total number of short sequences in the third urn, and $(U - K)$ follows Poisson distribution with mean cK.

Write W as the number of the sequences in the third urn that contain the target sequence. The distribution of W is more complicated than that of X. Assuming $X > 0$,

Table 2. Random quantities and their expectations.

Notation	Distribution	Expectation
X	$H(46n, n, K)$	$\frac{K}{46}$
U_i	$\mathrm{Poi}(c)$	c
W	$\sum_{i=1}^{X} \mathrm{Ber}(q_i(U_i))$	$\frac{K(2e^{ct_1t_3} - e^{ct_2t_3})}{46e^c}$
$Y \vert U, W$	$H(U, W, R)$	$\frac{WR}{U}$

we have at least 1 target sequence contained in the second urn. We have $W = \sum_{i=1}^{X} B_i$, where fix X, $\{B_i\}_{i=1}^{X}$ are independent Bernoulli random variables. Condition on $\{U_i\}_{i=1}^{K}$, the probability of success q_i of random variable B_i satisfies

$$q_i(U_i)\begin{cases} \geq 2(t_1t_3)^{U_i} - (t_2t_3)^{U_i} & \text{if } T \geq f \\ = t_3^{U_i} & \text{otherwise} \end{cases} \tag{1}$$

respectively, with $t_1 = \frac{L-T}{L-f}, t_2 = \frac{L-2T+f}{L-f}$, and $t_3 = 1 - \frac{f}{L}$, respectively. The proof is found in Appendix A.

Finally, when we condition on U and W, the number of target sequences in the detection machine Y follows a hypergeometric distribution with parameters (U, W, R).

2.1. Analytical calculations

In this section, we present the analytical results of our statistical modeling. Our goal is to set the sampling parameters K and R so that we can guarantee

$$P(Y \geq Q) \geq p, \text{ for pre-specified } Q \text{ and } p \tag{2}$$

Now we consider R_{low}, such that with pre-fixed quantities p_0, U and W

$$P(Y \geq Q | U, W, R \geq R_{\text{low}}) \geq p_0 \tag{3}$$

Note here $Y|U, W \sim H(U, W, R)$. We will find R_{low} as a function of U, W, p_0 from tail bounds on hypergeometric distribution.

In this paper, we use the concentration inequality in Lemma 2 on binomial distribution to control the tail bounds of hypergeometric distribution. Specifically, we will use the following theorem:

Theorem 2.1. *Let $h \sim H(A, B, C)$ be a hypergeometric random variable, $B_a \sim Bin\left(C, \frac{B}{A}\right)$ and $B_b \sim Bin\left(A - C, \frac{B}{A}\right)$ be two binomial random variables. Then under conditions on A, B, C and x listed in Appendix B,*

$$P(h \leq x) \leq P(B_a \leq x) \tag{4}$$

$$P(h \leq x) \leq P(B_b \geq B - x) \tag{5}$$

The proof is in Appendix B. Numerical results presented in Section 4 suggest that for large C (5) is a better bound, in the remaining cases we will use (4).

Usually, one would want to fix (A, B, C) and calculate the tail bounds with different x. In this case only Property 1 is needed to ensure the validity of Theorem 2.1. The remaining properties proved in Appendix B ensure the validity of Theorem 2.1 for the other cases needed in Algorithm 1 when (A, B, C) are changing. In the subsequent calculations, we assume the conditions for Theorem 2.1 are met. In particular, we will use large deviation bounds from Lemma 2 on the two binomial distributions: $\mathrm{Bin}\left(R, \frac{W}{U}\right)$ and $\mathrm{Bin}\left(U - R, \frac{W}{U}\right)$ to find R_{low} in (3).

Write $R_{\mathrm{low}} = R_{\mathrm{low}}(U, W, p_0)$. Note that U and W are typically unknown. Therefore, R_{low} itself is still a random quantity and we need to further find a upper bound for R_{low} depending on n and K, this is denoted by $\hat{R}_{\mathrm{low}}$. With large probability, sampling $\hat{R}_{\mathrm{low}}$ sequences in the third urn is enough to guarantee sampling no less than R_{low} samples.

It is fairly straightforward to see R_{low} increases with W and decreases with U. Now we fix Q and p_0, and write U_{up} and W_{low} as the probabilistic upper/lower bounds for U and W, respectively. From (4) and (5) we can find $\hat{R}_{\mathrm{low}}$ directly from tail bounds on $\mathrm{Bin}\left(R, \frac{W_{\mathrm{low}}}{U_{\mathrm{up}}}\right)$ and $\mathrm{Bin}\left(U_{\mathrm{up}} - R, \frac{W_{\mathrm{low}}}{U_{\mathrm{up}}}\right)$. In particular, the steps needed to determine $\hat{R}_{\mathrm{low}}$ for a given K and n are summarized here:

(1) Use Lemma 2 on binomial distributions $\mathrm{Bin}\left(K, \frac{1}{46}\right)$ and $\mathrm{Bin}\left(46n - K, \frac{1}{46}\right)$ to find lower bound X_{low} of X. Here X_{low} depends only on n, K and p_1 so that: $P(X \geq X_{\mathrm{low}}) \geq p_1$.

(2) Set $X := X_{\mathrm{low}}$ from step 1. Note that W is the summation of X_{low} independent Bernoulli trials. Hence from Lemma 2 we can find lower bound W_{low} of W depending only on n, K, L, f, T, c, p_1, p_2 so that:$P(W \geq W_{\mathrm{low}} | X \geq X_{\mathrm{low}}) \geq p_2$. Consequently $P(W \geq W_{\mathrm{low}}) \geq p_1 p_2$.

(3) Use inequality from Lemma 1 to find U_{up} and U_{low} depending only on c, K, p_3 so that: $P(U \geq U_{\mathrm{low}}) \geq p_3$ and $P(U \leq U_{\mathrm{up}}) \geq p_3$.

(4) Use Lemma 2 on binomial distributions $\mathrm{Bin}\left(R, \frac{W_{\mathrm{low}}}{U_{\mathrm{up}}}\right)$ and $\mathrm{Bin}\left(U_{\mathrm{up}} - R, \frac{W_{\mathrm{low}}}{U_{\mathrm{up}}}\right)$ to find $\hat{R}_{\mathrm{low}}$ so that:

$$\begin{aligned} P(\hat{R}_{\mathrm{low}} \geq R_{\mathrm{low}}) &\geq P(U \leq U_{\mathrm{up}}, W \geq W_{\mathrm{low}}) \\ &\geq P(U \leq U_{\mathrm{up}}) + P(W \geq W_{\mathrm{low}}) - 1 \\ &= p_3 + p_1 p_2 - 1 \end{aligned}$$

Note that we need to ensure the needed sample size R is not larger than the available number of short sequences U. To this end, both $\hat{R}_{\mathrm{low}}$ and U_{low} are deterministic functions of given constants and we can add numerical constraint on $\hat{R}_{\mathrm{low}}$ to force it smaller than U_{low}. A key observation from our numerical result is, as K gets larger, U_{up} and U_{low} will be more concentrated around the mean $cK + K$, while R_{low} will be much smaller than U_{low}. Therefore, we need to find a lower bound K_{min} on K to ensure $U_{\mathrm{low}} \geq \hat{R}_{\mathrm{low}}$.

Finally, given that we choose K and $\hat{R}_{\mathrm{low}}$ as our sampling sizes at two stages, respectively. The following relations are true:

$$\begin{aligned} P(Y \geq Q) &\geq P(Y \geq Q, R \geq R_{\mathrm{low}}, U \geq R) \\ &\geq p_0 \cdot P(\hat{R}_{\mathrm{low}} \geq R_{\mathrm{low}}, U \geq \hat{R}_{\mathrm{low}}) \\ &\geq p_0 \cdot \left[P(\hat{R}_{\mathrm{low}} \geq R_{\mathrm{low}}) + P(U \geq \hat{R}_{\mathrm{low}}) - 1\right] \\ &\geq p_0(2p_3 + p_1 p_2 - 2) \end{aligned} \tag{6}$$

It suffices to set the desired probability p equal to the right-hand-side of (6). The exact selection of $\{p_i\}_{i=0}^3$ can be found in Section 3. We will also show in Section 3.1 that the range of K is $[K_{\min}, 45n]$. While not every K in this range is feasible, a straightforward monotone analysis shows that as long as K is larger than a certain threshold, the solution $\hat{R}_{\text{low}}$ always exists.

2.2. Optimal sampling strategy

In this section, we discuss how to use the formulas derived in Section 2.1 to find the optimal values of n and K for any given p and Q. Specifically, assume there is a user-specified cost function $f(n, K)$ over number of samples n and the sampling size from first urn. In this paper we assume $f(\cdot,\cdot)$ is a monotone increasing function of both n and K.

The proposed procedure is summarized here:

(1) Solve for $\{p_i\}_{i=0}^3$ such that $p = p_0(2p_3 + p_1p_2 - 2)$.
(2) For fixed n, we calculate $K_{\min}$.
(3) For any fixed n and K such that $K \geq K_{\min}$, we calculate $\hat{R}_{\text{low}}$.
(4) Return: $(n, K, \hat{R}_{\text{low}})$.

The implementation details are discussed in Section 3. In reality the amount of biological materials is limited, hence there is an upper bound on n and there are only finite number of $(n, K, \hat{R}_{\text{low}})$ to consider. We do not need to consider any $R > \hat{R}_{\text{low}}$ as that would lead to sub-optimal design. However, for fixed n, we do need to consider $K > K_{\min}$, because larger K might lead to smaller $\hat{R}_{\text{low}}$ and a more efficient solution.

Assume we have a cost function $C(K, R)$ that increases with K and R. We only have finitely many $(n, K, \hat{R}_{\text{low}})$ to consider and a brute force search among all the possible triples will yield the optimal $(n, K, \hat{R}_{\text{low}})$ minimizing the cost function.

Due to technology limits, we may have certain constraints on sampling percentages: for example, we can only sample 80% in the first stage, and 50% from the second stage. We can still use a brute force search only considering the cases that do satisfy these extra constraints.

3. Sampling algorithm

In this section, we discuss the implementation details of optimal sampling strategy in Section 2.

The following quantities should be specified/calculated beforehand:

(1) Specify the values of L, f, T, p, Q, n, c according to the particular application.
(2) Select $p_0 = \sqrt{p}, 3p_3 - 2 = \sqrt{p}$ and $p_1 = p_2 := \sqrt{p_3}$ so that the right-hand-side of (6) becomes p.
(3) Compute: $t_1 = \frac{L-T}{L-f}, t_2 = \frac{L-2T+f}{L-f}, t_3 = 1 - \frac{f}{L}$ and set $Q_1 = \frac{2e^{ct_1t_3} - e^{ct_2t_3}}{e^c}, v = Q_1 - Q_1^2$. Here Q_1 and v are the expected value and variance of Bernoulli $\text{Ber}(q_i(U_i))$ random variable.

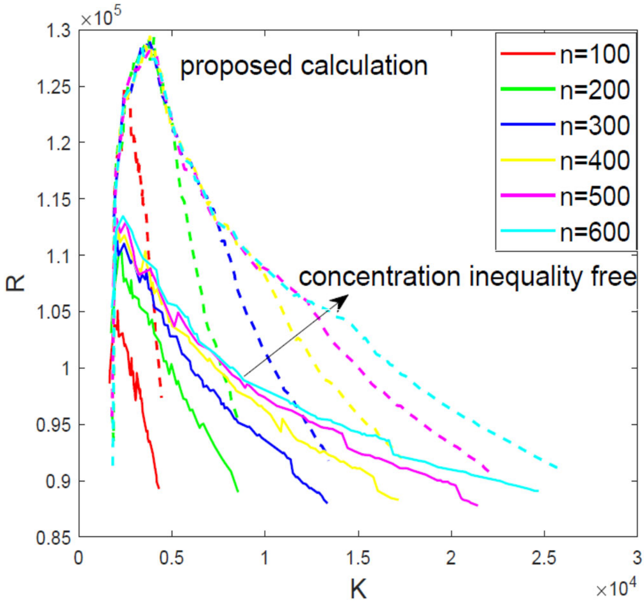

Figure 2. Approximation results: plot of K vs $\hat{R}_{\text{low}}$ using Algorithm 1 for n ranges from $n = 100$ to 600, different colors correspond to different n. Curves at the bottom correspond to concentration inequality free results, while dotted curves at the top correspond to results calculated from our algorithm.

3.1. Lower bound on K

We need to find the lower bound $K_{\min}$ of K such that with large probability we have at least Q target sequences in the third urn. Equivalently, we want $R \geq Q$. To this end, we assume the cutting process in urn 2 does not break any target sequences and we take everything out from urn 3. Therefore, we only need to make sure X is larger than Q with high probability. In Section 4, we solved both (4) and (5) to get different lower bounds for K, similarly with different lower bounds on K we will have different lower bounds for downstream quantities like X, U, etc.

3.2. Lower bound on R

Algorithm 1 can be used to calculate $\hat{R}_{\text{low}}$ with pre-fixed n and K. Please note, that we use tail bounds of binomial distribution to approximate that of hypergeometric distribution in step 1, 2, and step 4. Here, steps 1 and 2 only require property 1 and 2 in Appendix B. Step 4 additionally needs Property 3 and 4, because we need the relations in (4) and (5) to be true with both $W \geq W_{\text{low}}$ and $U \leq U_{\text{up}}$. For each fixed n, the range of K is relatively small, thus for each input n we can simply try all the possible K and calculate the corresponding smallest R (denoted by R_{low}) that achieves our goal. To make our algorithm more efficient, we can first find the smallest K that can give us a lower tail that is larger than Q (any smaller K will not be feasible, see our supporting code for details), call this $K_{\min}$. For each K from $K_{\min}$ to $45n$, we use Algorithm 1 to find R_{low}.

Algorithm 1. Computing $\hat{R}_{\text{low}}$ from fixed n and K

```
Apply Lemma 2 to $B_a \sim \text{Bin}\left(K, \frac{1}{46}\right)$ and $B_b \sim \text{Bin}\left(46n - K, \frac{1}{46}\right)$. Solve the following system:
    $-\log(c(1-p_1)) = Kh\left(\frac{t}{K} + 1 - \frac{1}{46}, 1 - \frac{1}{46}\right)$
    $c = \max\left\{2, \sqrt{4\pi Kh\left(\frac{t}{K} + 1 - \frac{1}{46}, 1 - \frac{1}{46}\right)}\right\}$
  and set $X_{\text{low}_1} = \frac{K}{46} - t$. Similarly we can solve for $X_{\text{low}_2}$. Set $X_{\text{low}} := \max(X_{\text{low}_1}, X_{\text{low}_2})$.
Fix $X$ to be $X_{\text{low}}$. Solve the following system
    $-\log(c(1-p_2)) = X_{\text{low}}h\left(\frac{t}{X_{\text{low}}} + 1 - Q_1, 1 - Q_1\right)$
    $c = \max\left\{2, \sqrt{4\pi X_{\text{low}}h\left(\frac{t}{X_{\text{low}}} + 1 - Q_1, 1 - Q_1\right)}\right\}$
  and set $W := Q_1 * X_{\text{low}} - t$.
Calculate the $p_3$ lower bound $U_{\text{low}}$ and upper bound $U_{\text{up}}$ for $U$ from Lemma 1.
Apply Lemma 2 to $B_c \sim \text{Bin}\left(R, \frac{W_{\text{low}}}{U_{\text{up}}}\right)$ and $B_d \sim \text{Bin}\left(U_{\text{up}} - R, \frac{W_{\text{low}}}{U_{\text{up}}}\right)$, and solve for $R_{\text{low}_1}$ from the following system
    $r\frac{W_{\text{low}}}{U_{\text{up}}} - t = Q$
    $-\log(c(1-p_0)) = rh\left(\frac{t}{r} + 1 - \frac{W_{\text{low}}}{U_{\text{up}}}, 1 - \frac{W_{\text{low}}}{U_{\text{up}}}\right)$
    $c = \max\left\{2, \sqrt{4\pi rh\left(\frac{t}{r} + 1 - \frac{W_{\text{low}}}{U_{\text{up}}}, 1 - \frac{W_{\text{low}}}{U_{\text{up}}}\right)}\right\}$
  set $R_{\text{low}_1} := r$. Similarly we can solve for $R_{\text{low}_2}$.
Set $\hat{R}_{\text{low}} = \min\{R_{\text{low}_1}, R_{low_2}\}$ and output $(n, K, \hat{R}_{\text{low}})$.
```

4. Numerical results

For our numerical results, the calculations were based on biologically reasonable parameters: $L = 250{,}000{,}000$, $f = 50{,}000$, $T = 75{,}000$, $c = 60$, $p = 0.95$, and $Q = 20$.

In Figure 2, we plot our original calculation results from Algorithm 1 together with the results without using any concentration inequalities (we get the tail points by the inverse of cumulative distribution functions, which is applicable for relatively small n); both of them have the similar patterns. From original calculation results we can find two "kinks" for each fixed n. This is because when K is small, we will need to sample almost everything from the second stage, which will force us to choose the correspond

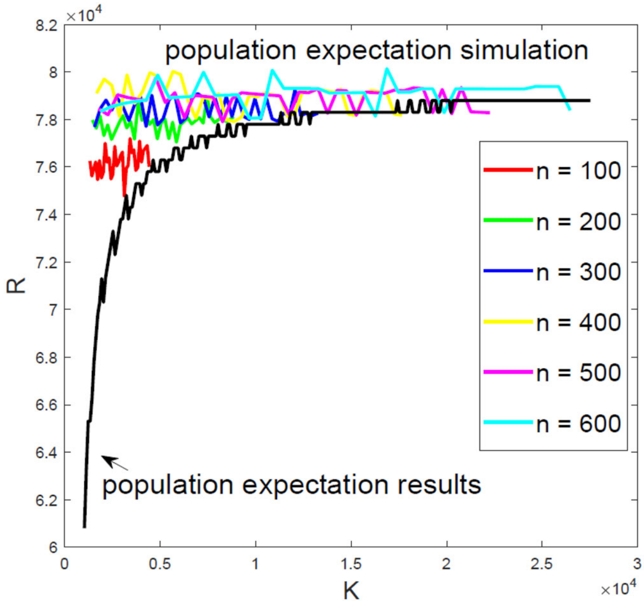

Figure 3. Simulation results and population expectation results: n ranges from 100 to 600.

Table 3. Minimization of cost function.

Constraint 1 (%)	Constraint 2 (%)	n	K	R	$\frac{K}{46n}$ (%)	$\frac{R}{U_{\text{low}}}$ (%)
100	50	100	3652	111,850	79.39	49.96
80	20	300	8716	106,220	63.16	19.91
50%	100%	100	1812	91322	39.39%	82.03%
50%	50%	200	4136	126630	44.96%	49.96%
20%	80%	500	2572	124370	11.18%	78.8%

$\text{Bin}\left(U_{\text{up}} - R, \frac{W_{\text{low}}}{U_{\text{up}}}\right)$ for Y as the binomial bounds. Then as K gets larger but not big enough, we will use $\text{Bin}\left(R, \frac{W_{\text{low}}}{U_{\text{up}}}\right)$ for both stages. Finally, K will get close to $45n$ which again forces to use $\text{Bin}\left(U_{\text{up}} - R, \frac{W_{\text{low}}}{U_{\text{up}}}\right)$ at the first sampling stage. The performance of our algorithm is slightly more conservative than the concentration inequality free approach in the sense that we ask for more samples. However, each lower bound of our algorithm can be solved efficiently using numerical method, while using inverse cdf function is generally slow for large n.

In Figure 3 we plot the simulation results together with population expectation results. Here, the simulation means of each n and fixed K we create large amount of X, W, and U. Then for each simulation trial, we use a brute force search to find the smallest R that can gives us (2). Note this simulation is an "averaging" approach while our algorithm is more like a tolerance interval approach, thus they are not comparable and we put them into two separate figures. The population expectation results means we replace W and U directly by their expectations, and again brute force search for the smallest R. From Figure 3 we can see as K gets larger, these two results will be very close, which implies for large K, we can approximately use expectations of U and W to conduct the calculation.

Table 3 provides examples the minimization results based on a linear cost function, $C(K, R) = aK + bR$, under various constraints. In particular we use $a = 60$, $b = 1$ and various sampling percentage constraints on both sampling stages. Under all constraints, the algorithm tends to sample as many as possible in the second stage.

We have also applied our algorithm to other choices of Q. The lessons learned are similar to what we have shown here. In the supporting materials we provide MATLAB code that can be used to calculate optimal sampling strategy with different parameters.

5. Conclusions

In this article, we have developed an optimization approach for estimating the amount of material needed for genomic mapping based on a simple, yet realistic, model of the process that uses a novel result regarding the tail bounds of the hypergeometric distribution. Our approach is both computationally and analytically tractable and we show that *if* a genomic mapping technology can sample most of the chromosomal fragments within a sample, comparatively little biological material is needed to detect a variant at high confidence.

Funding

Jan Hannig's research was supported in part by the National Science Foundation under Grant No. DMS-1512945 and IIS-1633074 and DMS-1916115. Corbin D. Jones' research was supported by DEB-1457690, University Cancer Research Fund, 1-R01-HG007407-01 (J.M. Ramsey PI).

ORCID

Weiwei Li http://orcid.org/0000-0002-7402-6874
Jan Hannig http://orcid.org/0000-0002-4164-0173
Corbin D. Jones http://orcid.org/0000-0001-7281-7130

References

Audano, P. A., A. Sulovari, T. A. Graves-Lindsay, S. Cantsilieris, M. Sorensen, A. E. Welch, M. L. Dougherty, B. J. Nelson, A. Shah, S. K. Dutcher, et al. 2019. Characterizing the major structural variant alleles of the human genome. *Cell* 176 (3):663–75. doi:10.1016/j.cell.2018.12.019.

Chaisson, M. J., R. K. Wilson, and E. E. Eichler. 2015. Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics* 16 (11):627–40. doi:10.1038/nrg3933.

Huddleston, J., and E. E. Eichler. 2016. An incomplete understanding of human genetic variation. *Genetics* 202 (4):1251–4. doi:10.1534/genetics.115.180539.

Hurles, M. E., E. T. Dermitzakis, and C. Tyler-Smith. 2008. The functional impact of structural variation in humans. *Trends in Genetics* 24 (5):238–45. doi:10.1016/j.tig.2008.03.001.

Short, M. 2013. Improved inequalities for the Poisson and binomial distribution and upper tail quantile functions. *ISRN Probability and Statistics* 2013:1–6. doi:10.1155/2013/412958.

Appendix A: Proof of Equation (1)

Proof. There are X copies of the target fragments in the second urn. Some of the fragments of interest might not survive during the cutting process, therefore we have $W \le X$. Define $\{A_i\}_{i=1}^{X}$

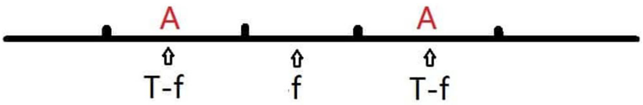

Figure A1. Demonstration of cutting: DNA sequence with target fragment. The *f* zone and at least one of the *A* zones should have no cuts to provide a valid target sequence.

as the event that the ith target fragment survives (i.e., being intact after cutting procedure) and is placed in the third urn. Given that we have U_i cuts on the ith target sequence, the locations of these U_i cuts are then uniformly distributed, therefore $p(A_i) = (1 - f/L)^{U_i}$.

Next, in order for the target fragment to be usable by the detector, it has to be longer than T. If $T \leq f$, the sequences that contain the target fragment are always longer than T, then $q_i(U_i) = p(A_i)$. Otherwise we estimate q_i from a lower bound using the probability of an event $A_i \cup E_i$, where E_i is the event of not having cuts within $T - f$ on either one or the other side of the target sequence (see Figure A1). Recall that $t_1 = \frac{L-T}{L-f}, t_2 = \frac{L-2T+f}{L-f}, t_3 = 1 - \frac{f}{L}$. Then by inclusion and exclusion $p(E_i \mid A_i) = 2(t_1)^{U_i} - (t_2)^{U_i}$ and consequently

$$q_i(U_i) \geq p(A_i)p(E_i|A_i) = 2(t_1 t_3)^{U_i} - (t_2 t_3)^{U_i}$$

□

Appendix B: Hypergeometric distribution and binomial bounds

In this section, we discuss the conditions needed for Theorem 2.1. We use the same notations as in Theorem 2.1. For fixed positive integer x, consider the following two inequalities

$$P(h = x) \leq P(B_1 = x) \tag{B1}$$

$$P(h = x) \leq P(B_2 = B - x) \tag{B2}$$

Note that if (B1) is true for all $x' \leq x_0$, then (4) is true for $x = x_0$, similarly for (B2). We write $r = \frac{B}{A}$ and expand the above two inequalities as

$$\frac{\binom{A-C}{B-x}}{\binom{A}{B}} \leq r^x(1-r)^{C-x} \tag{B3}$$

and

$$\frac{\binom{C}{x}}{\binom{A}{B}} \leq r^{B-x}(1-r)^{A-B-C+x} \tag{B4}$$

respectively. From now on we use (B3) and (B4) to derive the conditions needed for Theorem 2.1.

Property 1. If (B3) and (B4) are true for some fixed A, B, C and $x = x_0 \leq \frac{BC}{A}$. Then Theorem 2.1 is true for $x = x_0$.

Proof. We use backward mathematical induction on x to prove (B3) and (B4) are true for any $x \leq x_0$. Assume (B3) and (B4) are true for some $x = x_0 \leq \frac{BC}{A}$. Then for $x := x_0 - 1$, it suffices to have the following two inequalities

$$\frac{r}{1-r}\frac{A - C - B + x_0}{B - x_0 + 1} \leq 1, \text{ and } \frac{1-r}{r}\frac{x_0}{C - x_0 + 1} \leq 1$$

which only require $x_0 \leq 1 - r + \frac{BC}{A}$ and $x_0 \leq r + \frac{BC}{A}$, obviously true. Therefore, (B3) and (B4) are true for any $x \leq x_0$, this means Theorem 2.1 is true for $x = x_0$. □

Property 2. Assume (B3) and (B4) are true for some fixed A, B, $C = C_0$, and $x = x_0 \le \frac{BC_0}{A}$. Then Theorem 2.1 is true for any $C \in (C_0, A - B)$.

The proof of Property 2 is almost the same as that of Property 1. We present the proof sketch here and the details are omitted: one can fix $x = x_0$ and do forward mathematical induction on C_0 to show that (B3) and (B4) are true for any $C \in (C_0, A - B)$. The property is proved by using Property 1 and the fact that $x_0 \le \frac{BC_0}{A} < \frac{BC}{A}$.

Property 3. Assume the following inequalities are true for some constants A_{up} and B_{low}

$$\begin{gathered}2A \ge 2B + C, \quad Q \le g\frac{BC}{A}, \quad A \ge \frac{3}{(3-2g)}B + \frac{2}{(3-2g)}C + \frac{3}{(3-2g)}\frac{3A}{C} \\ x \ge 5, \quad \psi(A_{\text{up}}) \ge 0, \quad (B - Q)C \ge B(2Q + 1), \quad A \ge 3B + CG(B_{\text{low}}) \ge 0\end{gathered}$$

where $g = \frac{xA}{BC}$ and

$$\begin{aligned}\psi(A) &= (C-1)\log(A) + (C-x)\log(A-B-1) - C\log(A-1) - (C-x-1)\log(A-B) \\ &\quad + \log(A-C) - \log(A-B-C+x) \\ G(B) &= (x-1)\log(B+1) + (C-x)\log(A-B-1) - x\log B - (C-x-1)\log(A-B) \\ &\quad + \log(1+B-x) - \log(A-B-C+x)\end{aligned}$$

For fixed B, x and C, if (B3) is true for $A = A_{\text{up}}$, then (4) is true for any $A \in (\max\{B, C\}, A_{\text{up}})$; for fixed A, x and C, if (B3) is true for $B = B_{\text{low}}$, then (5) is true for any $B \in (B_{\text{low}}, A)$.

Proof. Using Property 1, we only need to show (B3) is true accordingly. Again we use (backward) mathematical induction on A. Given $\frac{\binom{A-C}{B-x}}{\binom{A}{B}} \le r^x(1-r)^{C-x}$. We need $\frac{\binom{A-1-C}{B-x}}{\binom{A-1}{B}} \le \left(\frac{B}{A-1}\right)^x\left(1-\frac{B}{A-1}\right)^{C-x}$. It suffices to show

$$\frac{\binom{A-1-C}{B-x}}{\binom{A-1}{B}} \le \left(\frac{B}{A-1}\right)^x\left(1-\frac{B}{A-1}\right)^{C-x}\left(\frac{A}{B}\right)^x\left(1-\frac{B}{A}\right)^{x-C}\frac{\binom{A-C}{B-x}}{\binom{A}{B}}$$

the inequality above is equivalent to $\psi(A) \ge 0$. Take first order derivative of $\psi(A)$ with respect to A we have:

$$\psi'(A) = -\frac{C}{A-1} + \frac{C-1}{A} + \frac{1}{A-C} - \frac{C-x-1}{A-B} + \frac{C-x}{A-B-1} - \frac{1}{A-B-C+x}$$

If $\psi'(A) \le 0$ for $A \le A_{\text{up}}$, the result is proved by using the monotonicity of $\psi(A)$ and the assumption that $\psi(A_{\text{up}}) \ge 0$. Now we will prove $\psi'(A) \le 0$. It suffices to show:

$$\begin{aligned}&-B^3C^2 + B^3C - B^2C^3 + B^2C^2x - B^2Cx + B^2C - BC^3 + BC^2x + BC^2 - BCx \\ &+A(3B^2C^2 - 3B^2C + 2BC^3 - 2BC^2x + 2BCx - 2BC + C^2x - Cx^2) \\ &+A^2(-3BC^2 + 3BC - C^2x + Cx^2 - 2Cx + x^2 + x) + A^3(2Cx - x^2 - x) \le 0\end{aligned} \tag{B5}$$

The fist line of (B5) is obviously negative by noting the following facts

$$B^2C^2x \le B^2C^3, \quad B^2C \le B^2Cx, \quad B^3C \le B^3C^2, \quad BC^2x + BC^2 \le BC^3 + BCx$$

For the rest lines, we use the following relations:

$$Ax^2 + Ax + ACx^2 \le 3AB^2C + 2BC^2x, \quad C^2x + 2BCx \le 2ACx, \quad -x^2 - x \le 0$$

where the last inequality follows by assumption $2A \ge 2B + C$. For the rest parts, we want $2xA^2 + 3AB + 2BC^2 + 3B^2C \le 3ABC$. It suffices to show

$$(3-2g)A \geq 3B + 2C + \frac{3A}{C}$$

which is equivalent to

$$A \geq \frac{3}{(3-2g)}B + \frac{2}{(3-2g)}C + \frac{3}{(3-g)}\frac{3A}{C}$$

this follows directly from the assumptions. Thus, the first part of Property 3 is proved.

Now we prove the second part. From mathematical reduction on B, we want $\frac{\binom{A-C}{B+1-x}}{\binom{A}{B+1}} \leq \left(\frac{B+1}{A}\right)^x\left(1-\frac{B+1}{A}\right)^{C-x}$. It suffices to have

$$\frac{\binom{A-C}{B+1-x}}{\binom{A}{B+1}} \leq \left(\frac{B+1}{A}\right)^x\left(1-\frac{B+1}{A}\right)^{C-x} \cdot \left(\frac{A}{B}\right)^x\left(1-\frac{B}{A}\right)^{x-C}\frac{\binom{A-C}{B-x}}{\binom{A}{B}}$$

which is equivalent to $G(B) \geq 0$. Similarly as before, we want this function increases with $B \geq B_{\text{low}}$, from which we only need to check $G(B_{\text{low}}) \geq 0$ and this follows from our assumption. Consider the first order derivative of $G(B)$:

$$G'(B) = \frac{1}{1+B-x} + \frac{C-x-1}{A-B} - \frac{C-x}{A-B-1} + \frac{x-1}{1+B} - \frac{x}{B} + \frac{1}{A-B-C+x}$$

Then $G'(B) \leq 0$ requires

$$\begin{aligned}
&-B^3(C-1)(C-2x) + B^2C^2(x-2) - BC(x-1) + BC^2(x-1) - 3B^2x + B(x-1)x \\
&-BC(x-1)x + B^2x^2 + B^2C(2+2x-x^2) + A^3(1-x)x \\
&+A^2\left[(x-1)x + 3B(x-1)x + C(x-1)x + (1-x)x^2\right] \\
&+A(-3B^2(x-1)x - C(x-1)x - 2BC(x-1)x + 2B(x-1)^2x + (x-1)x^2) \leq 0
\end{aligned} \tag{B6}$$

We can expand the first line of (B6) and write it as:

$$\begin{aligned}
&-B^3C^2 + B^3(2x+1)C - 2xB^3 + B^2C^2x - 2B^2C^2 - BCx + BC + BC^2x - BC^2 - 3B^2x \\
&+ Bx^2 - Bx - BCx^2 + BCx + B^2x^2
\end{aligned}$$

we want to show the above line is non positive. Note that

$$\begin{aligned}
&-BCx + BC \leq 0, \quad BC^2x - 2B^2C^2 \leq 0, \quad -BC^2 \leq 0 \\
&-Bx \leq 0, \quad -3B^2x + Bx^2 \leq 0, \quad -BCx^2 + BCx \leq 0, \quad B^2x^2 - 2xB^3 \leq 0
\end{aligned}$$

Finally, we only need $-B^3C^2 + B^3(2x+1)C + B^2C^2x \leq 0$, which follows from our assumption: $(B-x)C \geq B(2x+1)$.

From $x \geq 5$ we immediately get: $2+2x-x^2 \leq 0$, hence $B^2C(2+2x-x^2) \leq 0$. For the second and third terms at the second line of (B6) we show:

$$A(1-x)x + (x-1)x + 3B(x-1)x + C(x-1)x + (1-x)x^2 \leq 0$$

it suffices to have $A - 3B - C \geq 0$, which is our assumption. It is fairly straightforward to prove the last line of (B6) is non negative, hence we omit it here. □

Property 4. Assume the following inequalities are true for constants A_{up} and B_{low}

$$\psi(A_{\text{up}}) \geq 0, \quad Ax \geq B + x + 2Bx, \quad G(B_{\text{low}}) \geq 0$$

where

$$\begin{aligned}\psi(A) &= (A-B-C+x-1)\log(A-1-B) + (A-C)\log(A) - (A-C-1)\log(A-1) \\ &\quad - (A-B-C+x)\log(A-B) + \log(A-B) - \log(A) \\ G(B) &= (B-x)\log(B+1) + (A-B-C-1+x)log(A-1-B) - (B-x)\log(B) \\ &\quad - (A-B-C+x-1)\log(A-B)\end{aligned}$$

For fixed B, x and C, if (B4) is true for $A = A_{\mathrm{up}}$, then (5) is true for any $A \in (\max\{B, C\}, A_{\mathrm{up}})$; for fixed A, x, and C, if (B4) is true for $B = B_{\mathrm{low}}$, then (5) is true for any $B \in (B_{\mathrm{low}}, A)$.

Proof. Using Property 1, we only need to show (B4) is true accordingly. Same as before we use (backward) mathematical induction on A. For $A = A_{\mathrm{up}}$ we want:

$$\frac{\binom{C}{x}}{\binom{A-1}{B}} \le \frac{\binom{C}{x}}{\binom{A}{B}} \left(\frac{B}{A-1}\right)^{B-x} \left(\frac{A-1-B}{A-1}\right)^{A-B-C+x-1} r^{x-B}(1-r)^{-A+B+C-x}$$

which is equivalent to $\psi(A_{\mathrm{up}}) \ge 0$. Similarly as the proof of Property 3, it suffices to show

$$\begin{aligned}\psi'(A) = \frac{C+1-A}{A-1} + \frac{A-C-1}{A} + \frac{A-B-C+x-1}{A-B-1} - \frac{A-B-C+x-1}{A-B} \\ + \log\frac{A(A-B-1)}{(A-1)(A-B)} \le 0\end{aligned}$$

for any $A \le A_{\mathrm{up}}$ that satisfies the assumptions. It suffices to have $B + B^2 + BC + B^2C + A(-B - 2BC - x) + A^2x \le 0$, this only needs $B + 1 \le A$, which is obviously true according to our assumptions. Similarly, for the second part we need for $B = B_{\mathrm{low}}$:

$$\frac{\binom{C}{x}}{\binom{A}{B+1}} \le \frac{\binom{C}{x}}{\binom{A}{B}} \left(\frac{B+1}{A}\right)^{B+1-x} \left(\frac{A-1-B}{A}\right)^{A-B-C+x-1} r^{x-B}(1-r)^{-A+B+C-x}$$

and it suffices to have $G(B) \ge 0$ for any $B \ge B_{\mathrm{low}}$ that satisfies the assumptions. Again we prove the monotonicity of $G(B)$:

$$\begin{aligned}G'(B) = \frac{-B-C+A+x-1}{A-B} - \frac{-B-C+A+x-1}{-B+A-1} - \log(-B+A-1) + \log(A-B) \\ - \frac{B-x}{B} + \frac{B-x}{B+1} - \log(B) + \log(B+1) \ge 0\end{aligned}$$

it suffices to show $B + B^2 + BC + B^2C - BA - Ax - 2BAx + A^2x \ge 0$, which can be proved by using our assumption $Ax \ge B + x + 2Bx$. Thus the second part is proved. □

Appendix C: Lemmas

To make this paper self-contained, in this section we list the lemmas from Short (2013) that were slightly modified to be applicable in our calculation. Detailed proof can be found in relevant references.

Lemma 1. *(Bounds on Poisson distribution.) Let $U \sim Pos(m)$. Then for $p \in (0, 1)$,*

$$\mathbb{P}\left[U \le m + \Phi^{-1}(p)\sqrt{m} + \frac{\Phi^{-1}(p)^2}{6}\right] \ge p \tag{C1}$$

$$\mathbb{P}\left[U \geq m - \sqrt{-2m\log\left(1-p\right)}\right] \geq p \tag{C2}$$

here $\Phi^{-1}(\cdot)$ is the inverse cdf function of standard normal distribution.

The following inequality is tighter than Chernoff's bound.

Lemma 2. *(Large deviation bound on binomial distribution.) Let $X \sim \mathrm{Bin}(n,p)$, and $h(a,b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}$, $a,b \in (0,1)$. Then for a fixed $t \in (0,np)$,*

$$\mathbb{P}[X \geq np - t] \geq 1 - \frac{\mathrm{e}^{-nh(1-p+\frac{t}{n},1-p)}}{\max\left\{2, \sqrt{4\pi nh\left(1-p+\frac{t}{n},1-p\right)}\right\}}$$