Deforesting Coverage Data

Jack Fraser-Govil

November 14, 2023

1 INTRODUCTION & MOTIVATION

Coverage data belies a wealth of data – we have already seen that a hidden set of biases can be seen when considering the coverage-frequency distribution.

In addition, Coverage can also be used to detect the presence of large-scale copy-number variations and other forms of duplication or deletion (which we shall generically call 'gains' and 'losses'). In the case of losses it is easy to see why this would be the case, it is perhaps less obvious why it happens in the case of gains.

Provided the read length is smaller than the scale of the duplications we are interested in, it is generally not possible to infer which copy of a motif a read originated from, and so the alignment does not record this as an insertion/duplication of bases – it simply assigns them to the original copy in the reference, leading to a spurious over-coverage of the reference, as shown in Figure 1.

Only reads which bridge the gap between consecutive copies would permit us to properly infer the presence of duplication – however, if the number of duplicates is greater than two, then there are multiple such bridge points, and the problem is highly degenerate and we must rely on the coverage to infer the number of duplications.

The problem with this kind of inference is that Coverage is already inherently stochastic and highly noisy - coverage plots covering any significant portion of the genome are difficult to interpret due to a large degree of shot noise which – as we have seen – has a significantly higher dispersion than Poisson shot noise, with the noise changing on (approximately) a per-base resolution.

Plots of the base coverage are - to be charitable - hard to interpret, reminiscent of the notorious 'Lyman-Alpha forest' in astrophysical spectroscopy. Whilst large-scale trends can (just about) be picked out by eye, we might desire a more rigorous and sophisticated methodology.

The aim of this work is to generate a computational tool which can peer through the 'Forest', and see the underlying pattern.

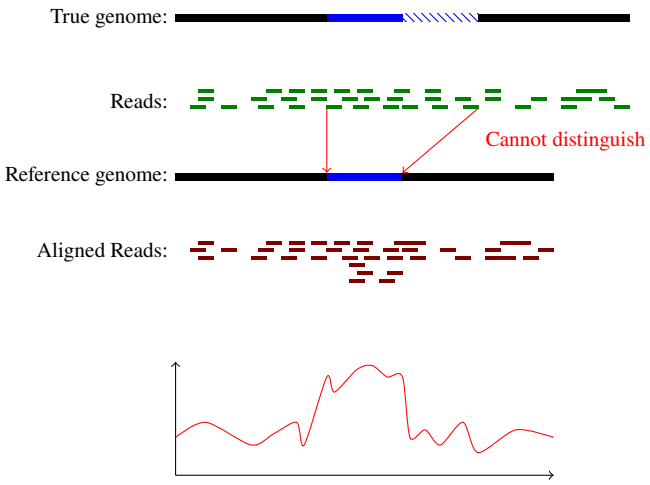

Figure 1. A depiction of duplication leading to a higher coverage. The blue hatched region is duplicated with respect to the reference, but since the reads are much shorter than the duplicated region, they cannot be properly assigned as duplicates, instead leading to a higher coverage rate.

2 MATHEMATICAL THEORY

2.1 Naive Smoothing

The most obvious solution is to simply pass a smoothing kernel over the data, and use that to infer the underlying mean function that the data is oscillating around. The Nadarya-Watson (cite?) method uses a kernel function $K_\lambda(x,y) = \frac{1}{\lambda}k\left(\frac{|x-y|}{\lambda}\right)$, and the mean estimate of a set of data (x_i, y_i) is given by:

$$\hat{m}_\lambda(x) = \frac{\sum_i^n K_\lambda(x,x_i)y_i}{\sum_i K_\lambda(x,x_i)} \tag{1}$$

The kernel should be sufficiently decaying that $K_\lambda(x,y) = 0$ when $|x-y| \gg \lambda$. Since we know that x_i is always an integer, and it spans the entire set between 0 and C (the chromosome size), we can also therefore infer that $|x - x_i|$ is an integer (presuming we are limiting ourselves to prediction on the integers – prediction on the reals is meaningless in this case). We can therefore trivially precompute most of this function - especially if we define some boundary over which we can assume the kernel is equal to zero, which we write as some multiple t of the kernel radius – i.e. $k(t) \approx 0$.

Smoothing in this fashion serves as a generic way to extract a smooth curve from the extremely noisy data present (ii.e. the 'forest') - however it fails us on a number of grounds:

- The smoothing is not biologically motivated (no underlying mechanism present) – it is merely a fancy way of drawing a line through the data, without regard to any other information we have about the system – for example, the knowledge that coverage can never be negative, which should biases the distribution to be lopsided.
- This reduces the coverage to a smooth curve. This is undesirable because:

  – It gives us no indication about where the gains and losses begin or end (the quantity of real interest to us) – we would need to do some more statistical post-processing in order to infer the edges of the transitions. The smoothing only helps human eyes see things, it does not particularly help in our numerical affairs.

  – The action of the kernel is (by definition) to smooth out sharp transitions. However – when a transition occurs we *expect* it to be a sharp transition. The kernel method necessarily smooths out these edges – it is difficult to distinguish between 'noise' and 'genuine transitions'.

2.2 Poisson Fitting

A more scientific (and less 'let's just draw a line through the data') method would exploit the fact that we know the data should (more or less) follow a Poisson distribution. The advantage (other than just aesthetic) of this over the mean-smoothing approach is that the Poisson distribution (and indeed, any realistic coverage distribution) is asymmetric around the peak value – the Poisson covers the semi-infinite interval $[0, \infty]$. Therefore, although we should expect the sample mean of a Poisson to be equal to the Poisson Parameter λ, deviations away from this are asymmetric in a way that the Kernel smoothing mechanism might have trouble with.

From previous work we know that the distribution is non-Poissonic on a global scale (i.e. the entire chromosome/genome). However, we might hypothesise that the distribution is obeyed *locally*, but with variations in the mean of the distribution over the spatial extent of the chromosome. In this case, the probability distribution of the coverage at index i is:

$$p(k,i) = \frac{[\lambda(i)]^k \, e^{-\lambda(i)}}{k!} \tag{2}$$

The deforestation exercise therefore reduces to inferring the behaviour of λ, given the observed distribution.

We assume that we have some way of parameterising the function $\lambda(i) = \lambda(i, \vec{\theta})$, and so we wish to infer the parameters θ which maximise the following Likelihood:

$$\begin{aligned} \mathcal{L}(\vec{\theta}, \{i, y_i\}) &= \sum_i \left[y_i \ln(\lambda(i|\vec{\theta})) - \lambda(i|\vec{\theta}) - \ln(y_i!) \right] + \text{Prior}(\vec{\theta}) \\ &= \text{const} + \sum_i \left[y_i \ln(\lambda(i|\vec{\theta})) - \lambda(i|\vec{\theta}) \right] + \text{Prior}(\vec{\theta}) \end{aligned} \tag{3}$$

This has gradient:

$$\frac{\partial \mathcal{L}}{\partial \vec{\theta}} = \sum_i \left(\frac{y_i}{\lambda(i|\vec{\theta})} - 1 \right) \frac{\partial \lambda(i|\vec{\theta})}{\partial \vec{\theta}} \tag{4}$$

Hence a basic gradient descent method can easily find the most likely distribution – modulo some prior which enforces smoothness.

This, however, has a similar problem to the Naive fitting method: when the Prior is weak, it is just as noisy as the data. When the prior is strong, it simply smooths over all but the largest deviations, and does not have sharp transition edges which would allow us to identify the position of gains and losses.

2.2.1 A Note On Priors

The most obvious Prior for imposing the kind of smoothness we are looking for is something on the form:

$$\log \text{Prior}(\lambda(\theta)) = -\sum_i \frac{(\lambda_i - \lambda_{i-1})^2}{\ell^2} \tag{5}$$

This works, insofar as it imposes a penalty whenever subsequent values of λ are far apart from each other, and smaller when they are close together – but has the problem of 'fine tuning': the prior has a 'strength' as an external parameter, which then controls the 'spikiness' of the output curve. By selecting this value appropriately, the user is able to either make the output ust as much of a forest as the input data, or perfectly smooth: and anything in between.

The user is then left to fine tune the dials of this parameter until they get something they would find acceptable. The problem with this is that it invariably leaves the user projecting their own expectations on the input – you only get out what you put in.

In such a case, a Bayesian statistician would say that we have formulated our Prior poorly – it has a free parameter (the strength), over which we have actually no prior knowledge. In fact, on further examination the strength (formulated as a length scale over which the binding happens) isn't even the quantity we are wishing to constrain since – as we have already noted – we don't even want a smooth output curve.

We must therefore ask what it is we wish our smoothing-prior to do since it evidently isn't 'smoothing' – since we are looking for a highly discontinuous output function! We want our prior to be such that it causes the model to disregard small-scale oscillations in the coverage as merely being noise. Smoothing length scales had this as a side effect – but with the unfortunate side effect that they also smoothed over what we (probably) believe to be sharp, discontinuous edges.

2.3 Attempt 3: Harmonic Fitting

The problem with the prior attempts is that they attempt to fit some continuous function to the data, in the hope of elucidating some distinct upward and downward steps in the data, which represent 'gains' and 'losses' – in so doing they are messy and noisy, and only marginally easier to interpret than the raw data.

We should like a method which actually assigns discrete values, and hence makes it easier to mechanistically identify regions of gains and losses in the genome. We should also like the Priors which we place on the models to be relevant to the quantities which we are actually studying: rather than attempting to smooth out noise via some arcane length scale, we should instead place strict boundaries on what we do and do not consider relevant.

We propose the following solution, which takes inspiration from the 'harmonics' observed in musical instruments: when you play an 'A' on a piano, you do not merely produce a pure tone at 440Hz, but a superposition of *harmonics* of that tone: at 880Hz, 1320HZ and so on.

So it is with our hypothesis: we are asserting that there is a single, mean coverage depth but that, due to the effects of normal heterogeneity and subsequent gains and losses, parts of the genome are being amplified or suppressed. This results in 'harmonics' of the mean coverage depth: integer multiples of the 'fundamental harmonic'

Hence, we should only fit a single frequency, but with the knowledge that each datapoint should then be fit to a integer harmonic of that frequency.

The model probability of a datapoint occuring due to the q^{th} harmonic of some fundamental frequency ν is:

$$p(k|\nu) = \left(\frac{[q\nu]^k \, e^{-q\nu}}{k!} \right) \tag{6}$$

Therefore the likelihood of a given harmonic $q \in \mathbb{N}$ (q must be a non-negative integer to satisfy the harmonic constraint) is:

$$\begin{aligned} \mathcal{H}(q) &= p(q|k,\nu) \\ &= \frac{p(k|q,\nu)p(q)}{p(k)} \\ &= \frac{[q\nu]^k \, e^{-q\nu}}{k!} \times \text{Prior}(q) \end{aligned} \tag{7}$$

The most likely value of q can therefore be found by a simple integer search: this is technically an infinite task, but we can restrict ourselves to a finite subset via our prior. This is a statement of Bayesian maximum Likelihood: we search for the most likely value of q, which is determined both by the data and our knowledge of what a reasonable

value of q is. The restriction of q to the natural numbers (which we define to include 0) is the 'harmonic' assumption: everything should exist in multiples of the fundamental frequency.

We expect that most sequences in the DNA will have $q = 2$, i.e. they occur only once on each copy of the chromosome. A deletion (or a heterozygous sequence) on one chromosome will give $q = 1$, and a deletion on both chromosomes give $q = 0$. An amplification likewise gives $q = 3$ if the sequence is duplicated on one only chromosome and so on.

It is worth noting that we are detecting the absolute multiplicity, rather than the relative: Motifs which have multiplicity in the reference (such as in the centromere or telomere) will show up as high-q values, and low q values will appear where mis-alignment has occurred. Therefore motifs which are already repetitive but which have suffered gains or losses will move from one q to another – we will only be able to detect these regions by comparing to a healthy sample.

2.3.1 Improving the model

We have formulated a basic probability assignment model using harmonics but if we were to apply this directly to the data we would find a number of problems since, at the moment, it applies only on a resolution of an individual base. A base with a coverage of 2 might be assigned a $q = 0$ (or not, as we shall see), whilst every other adjacent base has a $q = 10$ – we need to be able to account for noise.

The simplest way to do this is to discretise the data and find the most likely value of q for a block of bases of some length L, assign that a q and then move onto the next block. This has the problem that any transitions which occur in the middle of the block might get missed, and so the 'position' of the transition is dependent on the resolution of the model which is less than desirable.

We shall find a way around this by a proper formulation of our Prior.

We also note that in a region where a deletion has genuinely occurred there might still be 'peaks' in the coverage due to spurious alignment or contamination from a small population of cells without the total deletion which ruins our 'pure harmonic' assumption. This is a problem since even in the case of a 'block' of datapoints over a total deletion is total, the probability model will never be able to pull down a region containing non-zero data to a value of $q = 0$: a prior cannot force something impossible to become possible!

We must therefore refine our probability model slightly.

2.3.2 Error-Prone Probability Model

There are two possible sources of additional noise which we handle slightly differently. The first source of noise is an inherent deviation from the underlying Poisson assumption - this follows from our previous work that the distribution is highly non-Poisson for a variety of reasons: we term this *process noise*, and it is a fundamental part of the underlying biology. The second source is *experimental noise* - contamination from different cell populations, misalignments and so on.

The process noise we model by assuming that the underlying Probability model is, instead of being a pure-Poisson, marginalised against a Gamma distribution with mean μ and variance σ^2 (in the common parameterisation this is $(\alpha, \beta) = \left(\frac{\mu^2}{\sigma^2}, \frac{\mu}{\beta^2}\right)$):

$$\begin{aligned} p(k|\mu,\sigma) &= \int_0^\infty p(k|\lambda)p(\lambda|\mu,\sigma)\mathrm{d}\lambda \\ &= NB\left(k; \frac{\mu^2}{\sigma^2}, \frac{\mu}{\mu+\sigma^2}\right) \\ &\left(NB(k,r,p) = \frac{\Gamma(k+r)}{k!\Gamma(r)}(1-p)^k p^r\right) \end{aligned} \tag{8}$$

Where $NB(k; n, r)$ is the usual Negative Binomial distribution, written in terms of Real (rather than integer) r, and $\Gamma(x)$ is the usual Gamma function. This resulting distribution has mean $\mu = q\nu$ (recall: this was the mean value of the coverage depth) but variance $\mu + \sigma^2$ - and so we see that our distribution is Poisson-esque, albeit with a higher variance.

The Harmonic assumption is therefore embedded in this internal function: $\mu_i = q\nu$, where q is an integer. We assume that σ^2 is a global constant.

However, although this has the effect of broadening the tails of the distribution (and hence making the inference much more accepting of deviations away from the expected mean), it does not help the $q = 0$ problem, since $NB(k, 0, p) > 0$ only if $k = 0$. In order for that, we must sum over the probability that k was assigned erroneously:

The probability that $p(k_{\mathrm{obs}}|q)$ is therefore found from:

$$p(k_{\mathrm{obs}}|k,\nu,\sigma^2) = \left(\sum_k NB\left(k; \frac{q^2\nu^2}{\sigma^2}, \frac{q\nu}{q\nu+\sigma^2}\right) \times p(k_{\mathrm{obs}}|k)\right) \tag{9}$$

For simplicity's sake, it is probably easiest to use the L_1 kernel:

$$p(k_{\mathrm{obs}}|k) = \mathcal{N}\exp(-\gamma|k_{\mathrm{obs}} - k|) \tag{10}$$

2.3.3 The Prior

In order to properly formulate a Prior, we must first give some thought about exactly what question we are trying to answer. We are trying to ascertain the 'harmonic' at each index of the genome, given the observed coverage – but we want the model to disregard small-scale oscillations in the coverage as merely being noise. Smoothing length scales had this as a side effect – but with the unfortunate side effect that they also smoothed over what we (probably) believe to be sharp, discontinuous edges.

Let us therefore use this as a strict prior: there should be no oscillations on a scale shorter than some enforced length scale L - we assert that any gains or losses which span L bases or fewer are merely noise. Any determination if a change in harmonic has occurred must therefore consider *at least* L bases.

A standard Bayesian hypothesis can be formulated at the index i to determine if there has been a transition by considering the Hypothesis "the harmonic on the domain $[q, q + L]$ is equal to q_i", and testing all reasonable values of q_i (we suggest between 0 and 10 as being reasonable values). The odds-Likelihood of a transition is then:

$$p(\text{transition } q_{i-1} \to q) = \frac{p(D_i \to D_{i+L}|q)}{p(D_i \to D_{i+L}|q_{i-1})} \times \mathrm{Prior}(q|q_{i-1}) \tag{11}$$

It simply suffices to find if any of these terms is greater than 1 and, if any, which is the largest: this is the assigned value of q. The prior then allows us to penalise 'marginal' jumps by only assigned jumps

when the evidence reaches a threshold:

$$\text{Prior}(q|q_{i-1}) = \begin{cases} 0 < \alpha \leq 1 & \text{if } q = q_{i-1} \\ 1 & \text{else} \end{cases} \tag{12}$$

The case where $\alpha \approx 1$ means we accept a transition when the evidence is very marginal, $\alpha \ll 1$ means we require an overwhelming amount of evidence. Given some of the other post-processing that we will be doing (to locate transition edges more precisely), and provided L is sufficiently large that most spurious jumps will be eliminated, we use $\alpha = 0.5$ as a good first port of call.

Should we find that $q \neq q_{i-1}$, we have therefore been able to robustly identify that there is a transition in the region $[i, i+L]$ - but it is not true that the transition edge is at i – if $\alpha = 1$ the transition is probably somewhere in the middle of the region, whilst as $\alpha \to 0$ it gets closer to i. To robustly identify the edge of the transition we need to do a second hypothesis test – that the transition $q_{i-1} \to q$ occurs at an index j, the Likelihood of which is computed as:

$$\mathfrak{p}(\text{transition at } j|i, \text{data}, D) = p(D_i \to D_{j-1}|q_{-1}) \times p(D_j \to D_{j+L}|q) \tag{13}$$

I.e., we sucessively compute the probability that every point left of j is at the original $q = q_{i-1}$, and every point to the right is at the new value of q: the one with the highest value of $\mathfrak{p}$ is where we assign j.

2.3.4 *Initial Frequency Detection*

The most important free variable in our model is ν, the fundamental frequency of the model.

If we erroneously assign ν, the model produce results which veer towards the nonsensical since the majority of the data will lie off-harmonic, and so the model will place transitions at more or less spurious points. We can mitigate some of this problem by a gradient descent approach – however we note that due to the discontinuous nature of the assignments that this is only useful if our initial guess is good.

We can also trivially see that there is a high degree of degeneracy in the assignment of ν since $\nu \to \nu/10$, $q \to 10q$ produces identical results – albeit at the cost of reducing the power of our Harmonic approach, since it relies on the fact that the harmonics are highly distinct – our prior in this case is that the majority of the genome should have $q = 2$, being a normal non-heterogenous DNA sequence.

We propose that the following algorithm would identify the best possible starting value for ν:

- Start with $\nu = 1$ (or other small value)
- Assign qs across the dataset without using

2.4 Attempt 4: Harmonic Network

The Harmonic fitting method shows a vast improvement over the previous iteration - however it still suffers from a number of drawbacks, mostly associated with the somewhat dubious means by which the Bayesian tests were 'bootstrapped' from each other (technically each successive probability should be fully conditioned on the previous set, rather than just stuffing that all into a simplistic prior).

The Harmonic Network method is an attempt to a) make the assignment robust and mathematically sound and b) devise a method which is conceptually easier to understand and follow.

For this reason, we therefore find it much more convenient to think of the assignment of the harmonics to the genome as a *network*, such that finding the optimal assignment is equivalent to finding the shortest path through the network.

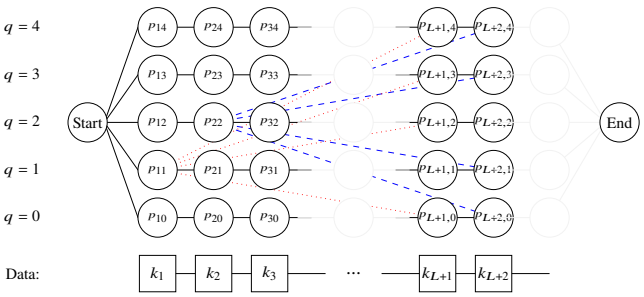

Figure 2. An example of a harmonic network – only two 'jump vertices' are shown (in red and blue) for clarity. In the full network, every node p_{iq} is connected to $p_{i+1,q}$ and $p_{i+L,k\neq q}$. The graph is directional - vertices can only be traversed from left to right.

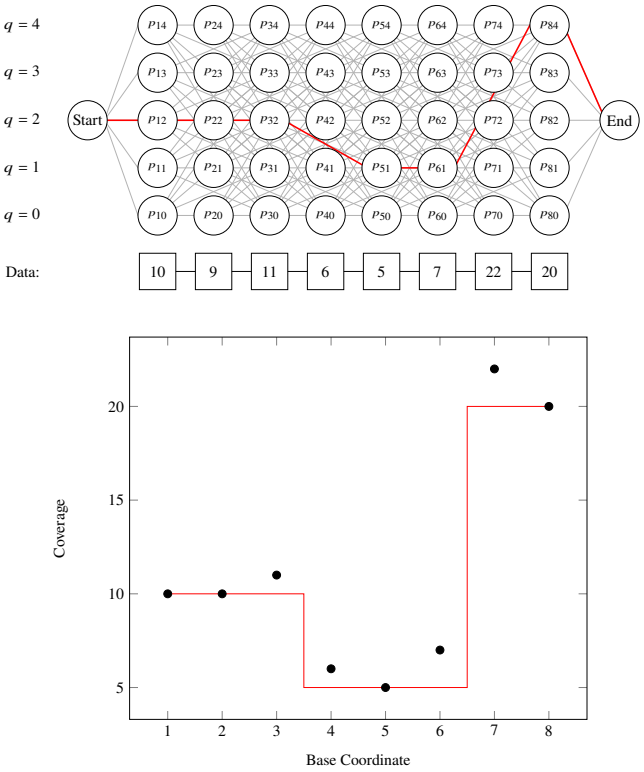

Figure 3. (Top panel) A demonstration of an optimal path through a network with $L = 2$ and $\nu = 5$, given some example coverage data. (Bottom panel) a projection of this path back onto the coverage distribution. For aesthetic reasons we have placed the transitions at half-integers – in practice non-integer values of base index are meaningless.