Supplementary Note - GangSTR Model

In order to estimate maximum likelihood genotypes at each repeat from Illumina sequencing data, GangSTR utilizes a model that combines information from multiple aspects of paired end reads. This note describes the full GangSTR model.

1 Likelihood Model

1.1 Symbols and Definitions

1.1.1 Model Representation

Our goal is to find the underlying genotype of a diploid individual at a known Tandem Repeat (TR) locus. For an individual that carries A and B copies of the TR motif on each allele we represent the genotype as $\langle A, B\rangle$. Likelihood is represented by L and defined by Equation (1), where $\overrightarrow{Data}$ corresponds to alignment information. We can interpret L as the probability of observing $\overrightarrow{Data}$ considering an underlying genotype $\langle A, B\rangle$.

$$L = P(\overrightarrow{Data}; \langle A, B\rangle) \tag{1}$$

GangSTR computes this probability for different values of A and B according to an optimization algorithm (see section 2) to find the genotype that maximizes the log-likelihood $\mathcal{L}$.

$$\begin{aligned} \langle \tilde{A}, \tilde{B}\rangle &= \arg\max \mathcal{L} \\ &= \arg\max logP(\overrightarrow{Data}; \langle A, B\rangle) \end{aligned} \tag{2}$$

1.1.2 Insert Size Distribution

We model the insert size δ with a limited Gaussian random variable Δ with the following distribution:

$$f_\Delta(\delta) = \frac{1}{C\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(\delta-\mu)^2} \qquad ; r \leq \delta \leq \infty \tag{3}$$

In this equation μ is average insert size, σ is insert size standard deviation, C is a normalization constant to account for limited range of δ, and r is the read length.

Integration of this probability density function arises several times throughout the rest of this document. We compute these integrals using a helper Gaussian distribution X:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad ; -\infty \leq x \leq \infty \tag{4}$$

and it's cumulative density function (CDF):

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx \tag{5}$$

1) For $r \leq a, b \leq \infty$:

$$\begin{aligned} \int_a^b f_\Delta(\delta)d\delta &= F_\Delta(b) - F_\Delta(a) \\ &= \frac{1}{C}\{F_X(b) - F_X(a)\} \end{aligned}$$

2) For $r \leq a, b \leq \infty$:

$$\begin{aligned}\int_a^b (\delta-\mu) f_\Delta(\delta) d\delta &= \frac{1}{C\sqrt{2\pi}\sigma}\int_a^b (\delta-\mu) e^{-\frac{1}{2\sigma^2}(\delta-\mu)^2} d(\delta-\mu) \\ &= -\frac{\sigma}{C\sqrt{2\pi}}\{e^{-\frac{1}{2\sigma^2}(b-\mu)^2} - e^{-\frac{1}{2\sigma^2}(a-\mu)^2}\} \\ &= -\frac{\sigma^2}{C}\{f_X(b) - f_X(a)\}\end{aligned}$$

1.2 Computation of Log Likelihood

The log likelihood for genotype $\langle A, B\rangle$ is calculated using the information extracted from the alignment file. This information consists of several read pairs r_i, that we assume to be independent from each other. We separate the calculation into a term for information extracted from read-pairs ($\mathcal{L}_P$, see section 1.2.1), and a term corresponding to the number of observed fully repetitive reads ($\mathcal{L}_N$, see section 1.2.2):

$$\mathcal{L} = logP(\overrightarrow{Data}; \langle A, B\rangle) = \overbrace{log\prod_i P(r_i; \langle A, B\rangle)}^{\mathcal{L}_P} + \overbrace{logP(|FRR|; \langle A, B\rangle)}^{\mathcal{L}_N} \tag{6}$$

1.2.1 Read pair Term

The first term in (6) is calculated by extracting evidence from every informative read pair. Each read pair belongs to one of the classes of informative reads.

$$\begin{aligned}\mathcal{L}_C &= log\prod_i P(r_i; \langle A, B\rangle) \\ &= \sum_i logP(r_i; \langle A, B\rangle) \\ &= \sum_i log\sum_j P(r_i, c^j; \langle A, B\rangle) \qquad (7)\end{aligned}$$

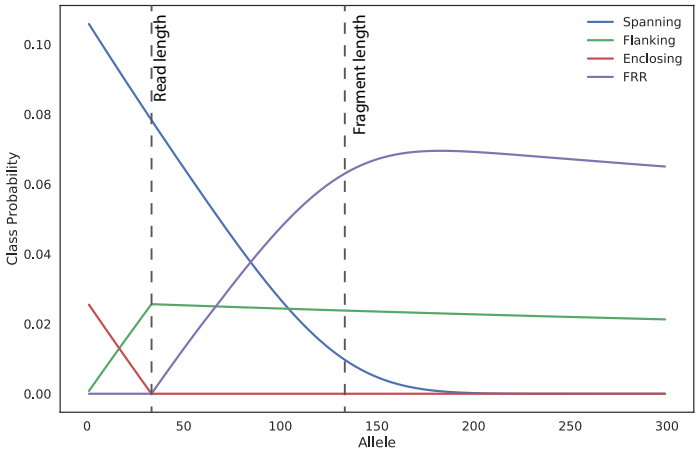
Step (7) shows expansion of probability over all possible read classes, c^j. Our definition is such that $P(r_i, c^j; \langle A, B\rangle)$ is only non-zero when r_i belongs to class c^j. We denote the specific class that r_i belongs to by c_i:

$$\begin{aligned}&= \sum_i logP(r_i, c_i; \langle A, B\rangle) \\ &= \sum_i log \underbrace{P(r_i|c_i; \langle A, B\rangle)}_{\text{Read Probability}} \underbrace{P(c_i; \langle A, B\rangle)}_{\text{Class Probability}} \qquad (8)\end{aligned}$$

Step (8) shows the definition of Read and Class Probability. Read Probability is described by the distribution of read-pairs in each class of reads (see section 1.4). Class probability corresponds to the abundance of reads from each class with respect to other classes (see section 1.3).

$$\begin{aligned}&= \sum_i log\frac{1}{2}\Big\{P(r_i|C(r_i); A)P(C(r_i); A) \\ &+ P(r_i|C(r_i); B)P(C(r_i); B)\Big\} \qquad (9)\end{aligned}$$

Finally, step (9) separates the contribution of each allele A and B assuming reads originate from both chromosomes with equal probability.



Supplementary Figure 1: Class probabilities as a function of repeat allele length. Results were calculated using 3bp long repeat unit, read length=100bp, and fragment length=400bp. Blue=spanning reads, green=flanking reads, red=enclosing reads, and purple=FRR reads.

1.2.2 Repetitive Read Count Term

The second term in (6) assigns a likelihood to the total number of observed fully repetitive reads (both on-target and off-target). We use a Poisson distribution with parameter λ to model the number of reads covering a region of the genome. Large TR expansions cause the FRR count to inflate beyond this average. Our model assigns a likelihood to the observed number of FRRs. Both on-target (reads aligned to the TR locus) and off-target (read pairs that are mapped elsewhere) FRRs are included in this term.

$$\begin{aligned}\mathcal{L}_N &= logP(|FRR|; \langle A, B \rangle) \\ &= log\frac{e^{-\lambda}\lambda^{|FRR|}}{|FRR|!} \\ &= -\lambda + |FRR| \cdot log\lambda - log\big(|FRR|!\big)\end{aligned}$$

We use the sterling approximation to calculate $\log(|FRR|!)$ for large $|FRR|$ values:

$$log\big(n!\big) \approx \left(n + \frac{1}{2}\right) \cdot log(n) - n + \frac{1}{2}log(2\pi) \tag{10}$$

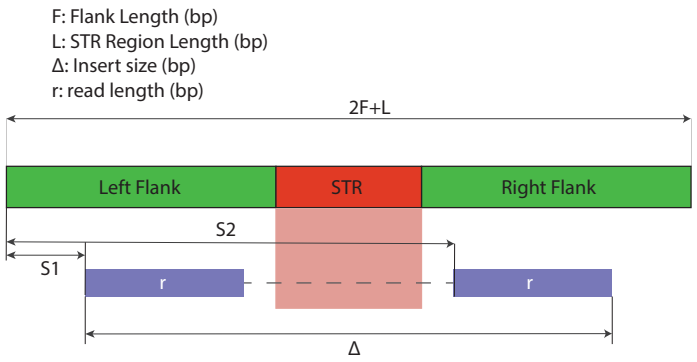
The Poisson parameter λ models the expected number of observed FRR reads, which is linearly related to the size of alleles A and B. Assuming uniform coverage C_v, read length r and motif length m we can calculate λ using (11). The unit step function $u(.)$ ensures alleles shorter than read length have 0 expected FRR reads.

$$\lambda = u(A - \frac{r}{m}) \cdot \frac{C_v(A \cdot m - r)}{2r} + u(B - \frac{r}{m}) \cdot \frac{C_v(B \cdot m - r)}{2r} \tag{11}$$

1.3 Class Probability

Class probability describes the probability of a read pair belonging to any specific class, considering uniform coverage [2]. For any value of underlying allele A, this probability can give an intuition on the relative abundance of different classes of reads **(Supplementary Figure 1)**.

We use **Supplementary Figure 2** as a general diagram for informative read pairs. In this figure, S_1 and S_2 are the distances of the beginning of each read in the read pair from the origin, which is F base pairs before the TR locus. The value F corresponds to the size of the flanking region that was used for



Supplementary Figure 2: Depiction of notation for describing GangSTR read classes. Green reperesents non-repetitive sequence, red represents the repeat region, and blue represents sequence reads.

extracting informative reads surrounding a TR locus of size L base pairs, or A copies of the motif of length m.

To calculate the class probability for each class of informative reads, we find the probability of S_1 and S_2 taking different ranges of values. We use the insert size distribution described in section 1.1.2, and assume the fragments are uniformly distributed (uniform coverage). The following sections describe the process for each class of informative reads.

1.3.1 Class Probability of Enclosing Reads

Without loss of generality, we assume the first mate in the pair is enclosing. The calculation is similar for the other mate (Equation (12)). Assuming uniformity of coverage, we use a uniform distribution to find the probability of a TR region being enclosed in a read.

$$\begin{aligned} P(c_i = E; A) &= P(S_1 < F, S_1 + r > F + A \cdot m) \\ &= P(F + A \cdot m - r < S_1 < F) \\ &= \frac{(F) - (F + A \cdot m - r)}{2F + A \cdot m - 2r} \\ &= \frac{(r - A \cdot m)}{2F + A \cdot m - 2r} \end{aligned} \tag{12}$$

1.3.2 Class Probability of Spanning Reads

A read pair is classified as spanning if it's two mates are mapped in the flanking region before and after the TR locus.

$$
\begin{aligned}
P(c_i = S; A) &= P(S_1 < x, S_2 > x + A \cdot m - r) \\
&= \int_{2r}^{2F+A\cdot m} P(S_1 < x, S_2 > x + A \cdot m - r | \Delta = \delta) f_\Delta(\delta) d\delta \\
&= \int_{2r}^{2F+A\cdot m} P(S_1 < x, S_1 + \Delta - r > x + A \cdot m - r | \Delta = \delta) f_\Delta(\delta) d\delta \\
&= \int_{2r}^{2F+A\cdot m} P(x + A \cdot m - \Delta < S_1 < x) | \Delta = \delta) f_\Delta(\delta) d\delta \qquad (13) \\
&= \int_{2r}^{2F+A\cdot m} \frac{(x) - (x + A \cdot m - \delta)}{2F + A \cdot m - 2r} u(\delta - A \cdot m) f_\Delta(\delta) d\delta \qquad (14) \\
&= \int_{max\{2r, A\cdot m\}}^{2F+A\cdot m} \frac{\delta - A \cdot m}{2F + A \cdot m - 2r} f_\Delta(\delta) d\delta
\end{aligned}
$$

Step function $u(.)$ is introduced in (14) to satisfy the condition in (13), $x + A \cdot m - \Delta < x$, which simplifies to $\Delta > A \cdot m$. This condition is then imposed in the integral limit. We continue the calculation using the helper integrals from Section 1.1.2.

$$
\begin{aligned}
P(c_i = S; A) &= \int_{max\{2r, A\cdot m\}}^{2F+A\cdot m} \frac{(\delta - \mu) + \mu - A \cdot m}{2F + A \cdot m - 2r} f_\Delta(\delta) d\delta \\
&= \frac{1}{2F + A \cdot m - 2r} \Big\{ (\mu - A \cdot m) \int_{max\{2r, A\cdot m\}}^{2F+A\cdot m} f_\Delta(\delta) d\delta \\
&+ \int_{max\{2r, A\cdot m\}}^{2F+A\cdot m} (\delta - \mu) f_\Delta(\delta) d\delta \Big\} \\
&= \frac{\mu - A \cdot m}{C(2F + A \cdot m - 2r)} \Big[F_X\big(2F + A \cdot m\big) - F_X\big(max\{2r, A \cdot m\}\big) \Big] \\
&- \frac{\sigma^2}{C(2F + A \cdot m - 2r)} \Big[f_X\big(2F + A \cdot m\big) - f_X\big(max\{2r, A \cdot m\}\big) \Big] \\
&= \frac{1}{C(2F + A \cdot m - 2r)} \Big\{ (\mu - A \cdot m) \Big[F_X\big(2F + A \cdot m\big) - F_X\big(max\{2r, A \cdot m\}\big) \Big] \\
&- \sigma^2 \Big[f_X\big(2F + A \cdot m\big) - f_X\big(max\{2r, A \cdot m\}\big) \Big] \Big\}
\end{aligned}
$$

1.3.3 Class Probability of Flanking Reads

Without loss of generality, we assume the first mate in the pair is flanking. The calculation is similar for the other mate. Assuming the uniformity of coverage, we use a uniform distribution to find the probability

of observing a flanking read.

$$\begin{aligned} P(c_i = F; A) &= P\big(S_1 < F, S_1 + r < F + A \cdot m, S_1 + r > F\big) \\ &= P\big(F - r < S_1 < min\{F, F + A \cdot m - r\}\big) \\ &= \frac{min\{F + A \cdot m - r, F\} - (F - r)}{2F + A \cdot m - 2r} \\ &= \frac{F + min\{A \cdot m - r, 0\} - F + r}{2F + A \cdot m - 2r} \\ &= \frac{min\{A \cdot m, r\}}{2F + A \cdot m - 2r} \end{aligned} \tag{15}$$

1.3.4 Class Probability of FRRs

$$\begin{aligned} P(c_i = FRR; A) &= P(S_1 \leq x, x \leq S_2 \leq x + A \cdot m - r) \\ &= \int_{2r}^{2F+A \cdot m} P(S_1 \leq x, x \leq S_1 + \Delta - r \leq x + A \cdot m - r | \Delta = \delta) f_\Delta(\delta) d\delta \\ &= \int_{2r}^{2F+A \cdot m} P(S_1 \leq x, S_1 \leq x + A \cdot m - \delta, S_1 \geq x + r - \delta) f_\Delta(\delta) d\delta \end{aligned} \tag{16}$$

We combine the inequalities describing S_1 in (16) to derive conditions that need to hold for this integral to have non-zero value.

- $\left.\begin{matrix} S_1 \geq x + r - \delta \\ S_1 \leq x + A \cdot m - \delta \end{matrix}\right\} \Rightarrow x + A \cdot m - \delta \geq x + r - \delta \Rightarrow A \cdot m \geq r$

  $\Rightarrow$ This condition is the clear condition underlying presence of FRR reads. Smaller TR lengths have 0 probability of having an FRR read.

- $\left.\begin{matrix} S_1 \geq x + r - \delta \\ S_1 \leq x \end{matrix}\right\} \Rightarrow x \geq x + r - \delta \Rightarrow \delta \geq r$

  $\Rightarrow$ The lower limit of the integral is $\delta \geq 2r$, hence this condition is satisfied for the range of possible δ values.

Since there are two upper bounds for S_1 in (16), we need to consider two different scenarios:

- $x \leq x + A \cdot m - \delta \Rightarrow \delta \leq A \cdot m$
  Therefore, for $2r \leq \delta \leq A \cdot m$; $A \cdot m \geq 2r$, integrand is simplified to:
  $\Rightarrow P(S_1 \leq x, S_1 \leq x + A \cdot m - \delta, S_1 \geq x + r - \delta) = P(x + r - \delta \leq S_1 \leq x)$
  For $A \cdot m < 2r$, this part has no contribution.

- $x > x + A \cdot m - \delta \Rightarrow \delta > A \cdot m$
  Similarly, for $A \cdot m \leq \delta \leq 2F + A \cdot m$, integrand is simplified to:
  $\Rightarrow P(S_1 \leq x, S_1 \leq x + A \cdot m - \delta, S_1 \geq x + r - \delta) = P(x + r - \delta \leq S_1 \leq x + A \cdot m - \delta)$

Continuing integration for $A \cdot m \geq 2r$:

$$
\begin{aligned}
P(c_i = FRR; A) &= \int_{2r}^{A \cdot m} P(x + r - \delta \leq S_1 \leq x) f_\Delta(\delta) d\delta \\
&+ \int_{A \cdot m}^{2F + A \cdot m} P(x + r - \delta \leq S_1 \leq x + A \cdot m - \delta) f_\Delta(\delta) d\delta \\
&= \int_{2r}^{A \cdot m} \frac{(x) - (x + r - \delta)}{2F + A \cdot m - 2r} f_\Delta(\delta) d\delta \\
&+ \int_{A \cdot m}^{2F + A \cdot m} \frac{(x + A \cdot m - \delta) - (x + r - \delta)}{2F + A \cdot m - 2r} f_\Delta(\delta) d\delta \\
&= \int_{2r}^{A \cdot m} \frac{(\delta - \mu) + (\mu - r)}{2F + A \cdot m - 2r} f_\Delta(\delta) d\delta \\
&+ \int_{A \cdot m}^{2F + A \cdot m} \frac{A \cdot m - r}{2F + A \cdot m - 2r} f_\Delta(\delta) d\delta \\
&= \frac{1}{C(2F + A \cdot m - 2r)} \Bigg\{ - \sigma^2 \big[f_X(A \cdot m) - f_X(2r) \big] \\
&+ (\mu - r) \big[F_X(A \cdot m) - F_X(2r) \big] \\
&+ (A \cdot m - r) \big[F_X(2F + A \cdot m) - F_X(A \cdot m) \big] \Bigg\} \qquad ; A \cdot m \geq 2r
\end{aligned}
$$

The result is similar for $A \cdot m < 2r$, except the first two terms are zero in this case:

$$
P(c_i = FRR; A) = \frac{A \cdot m - r}{C(2F + A \cdot m - 2r)} \Big\{ F_X(2F + A \cdot m) - F_X(A \cdot m) \Big\} \quad ; A \cdot m < 2r \tag{17}
$$

1.4 Read Probability

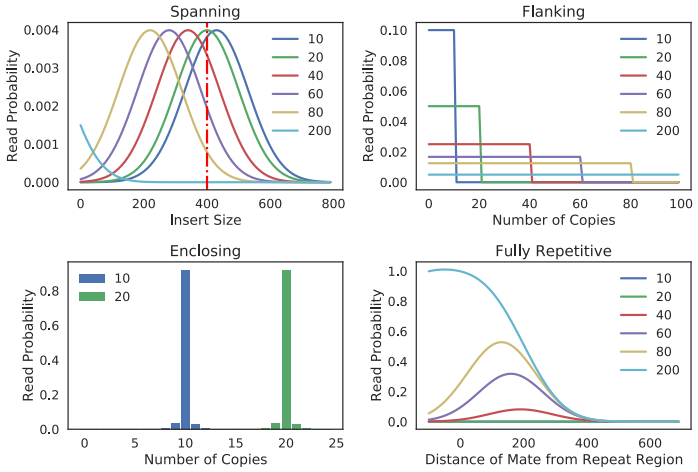
For each class of informative reads, the read probability describes the distribution of the informative characteristic of the class, given an underlying allele A **(Supplementary Figure 3)**. The details of read probability for each class of informative reads is presented in the following sections.

1.4.1 Enclosing Reads

Enclosing reads contain the whole repeating region, as well as flanking regions before and after. Therefore, the number of copies can be directly extracted after performing the local realignment step (see section 3).

The HipSTR stutter model [5] explains the distribution of the number of repeat copies in enclosing reads. Equation (18) shows the probability of a read with r_i copies having an error of length δ copies compared to the underlying true number of copies A. In this model, u and d correspond to the probability of stutter adding or removing copies of the motif, and ρ_s is the parameter of the geometric distribution that governs the number of stutter deviations from true number of copies A.

$$
P(r_i - A = \delta | c_i = E; A) = \begin{cases} 1 - u - d & \delta = 0 \\ u \rho_s \left(1 - \rho_s\right)^{\delta - 1} & \delta > 0 \\ d \rho_s \left(1 - \rho_s\right)^{-\delta - 1} & \delta < 0 \end{cases} \tag{18}
$$



Supplementary Figure 3: Read probabilities for each read class. Panels give read probabilities for spanning (top left), flanking (top right), enclosing (bottom left), and fully repetitive (bottom right) reads. In each plot, color indicates the number of repeats in an allele, the x-axis gives the data value modeled for each read type, and the y-axis gives the probability distribution for each data type based on the allele size. For enclosing reads, alleles with 40 or more repeats are not shown since those cannot be spanned by short (100bp) reads.

1.4.2 Flanking Reads

Flanking reads with n copies of the motif imply that one of the alleles has at least n copies of the motif. We use a uniform distribution (similar to [4]) to model the distribution of reads in the flanking class:

$$P(r_i = n | c_i = F; A) = \begin{cases} \frac{1}{A} & n \leq A \\ 0 & n > A \end{cases} \tag{19}$$

1.4.3 Spanning Reads

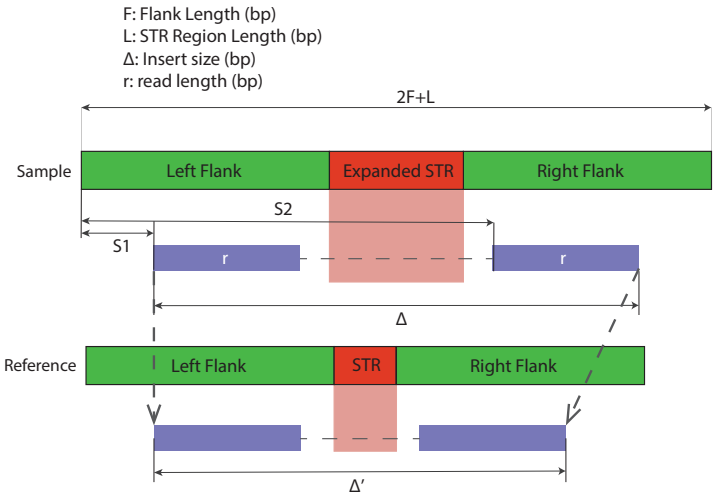
Fragments that completely span the TR region can create spanning read pairs. Spanning read pairs consist of two mates that are mapped to the flanking region before and after the TR. During alignment, spanning reads originating from an expanded TR allele experience a decrease in the insert size (**Supplementary Figure 4**). Therefore, the distribution of insert sizes for spanning reads is similar to the insert size distribution in section 1.1.2, with a decrease in average insert size by an amount equal to the size of expansion. If the reference has R copies of an m base pair motif, we can describe the class probability of spanning reads with the following Gaussian distribution:

$$P(r_i | c_i = S) \sim N(\mu - (A - R) \cdot m, \sigma) \tag{20}$$

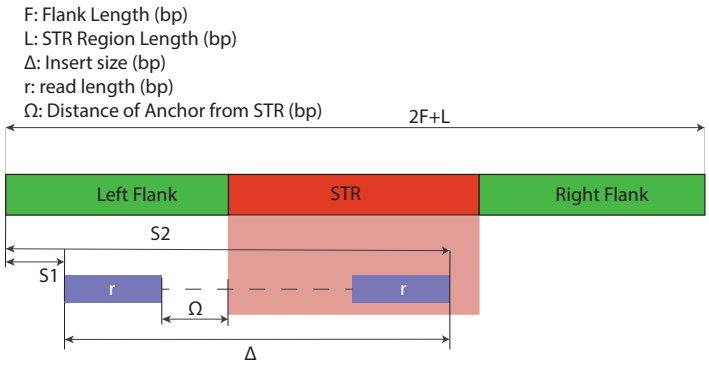
1.4.4 Fully Repetitive Reads (FRRs)

In section 1.2.2 we discussed how we use the number of FRR reads (both extracted from target locus and off-target regions) to create the repetitive read count term in the likelihood model. Here we discuss another informative aspect of FRR reads (**Supplementary Figure 5**).

Anchored FRRs are read pairs that contain one read completely consisting of repeats, while the other mate pair is mapped to the flanking region before or after the TR locus. Using the insert size distribution



Supplementary Figure 4: Modeling spanning reads at repeat expansions. Reads mapping to expanded TRs have an apparent decrease in fragment length distributions.



Supplementary Figure 5: Modeling fully repetitive reads. The distance of the non-repetitive mate to the STR is modeled as a function of repeat length.

(see 1.1.2), we model the distance of the anchor read from the repeat region (shown by Ω) to obtain read probability of this class of reads. We use the notation from section 1.1.2 to derive the read probability of anchored FRR reads.

$$P(r_i|c_i = FRR; A) = P(\Omega + 2r < \Delta < \Omega + r + L) \tag{21}$$
$$= F_\Delta(\Omega + r + L) - F_\Delta(\Omega + 2r) \tag{22}$$
$$= \frac{1}{C}[F_X(\Omega + r + L) - F_X(\Omega + 2r)] \tag{23}$$

On the other hand, fragments that originate from within the repeating region generate FRR read pairs (both mates repetitive). These read pairs do not have an anchor, and are most likely aligned to one of the off-target regions associated wiht the locus. These read pairs contribute to both FRR count term (adding two FRR reads, see section 1.2.2) and read pair term (FRR class probability computed for $\Omega = -r$, see section 1.2.1).

2 Optimization

In order to find the maximum likelihood diploid genotype, we implement optimization in multiple steps. Each step performs search in a different range of alleles. This modular approach to grid search helps prevent

reporting sub-optimal genotypes as a result of the optimization algorithm finding local maxima. First, any enclosing allele a_i with support of 2 or more reads is added to the list of potential alleles. In the second step, each potential enclosing allele, a_i, is used to perform 1-dimensional optimization of the likelihood function to find allele b_i, were $< a_i, b_i >$ minimizes the likelihood function. Next, multiple rounds of 2-dimensional optimization are performed to find $< c_i, d_i >$ genotypes that minimize the likelihood function. In each round the optiomizer uses a different initial point ($< short, short >$, $< short, long >$, etc.), which helps prevent reporting local opima. Any potential allele from each step, a_i, b_i, c_i, d_i, is added to the list of potential alleles. In the final step we compare the likelihood from any combination of alleles in this list, to find the maximum likelihood genotype. All 1 and 2-dimensional optimization is performed using COBYLA algorithm [3] implemented in the NLopt library [1]. This algorithm is a derivative-free optimization method that has a faster running time compared to exhaustive grid search.

2.1 Confidence Interval (Bootstrap)

After extracting all of the informative reads and performing optimization, we use the bootstrap method to calculate confidence interval for the estimated genotypes. The number of bootstrap samples, N_b, is set by the user. In each bootstrap round, GangSTR resamples the set of informative reads (with replacement) to create a bootstrap sample. This sample undergoes the optimization procedure described in section 2 to calculate a pair of bootstrap estimates. GangSTR records all bootstrap estimates in separate lists for shorter and longer alleles. These lists are then sorted and used to find the confidence interval at the desired level of significance.

3 Local Realignment

Large repeat expansions create large insertions that are not present in the reference. In order to perform a more accurate local realignment, we follow a method similar to that used in Tredparse [4]. Flanking regions before and after the TR locus are used to create artificial reference sequences with different numbers of repeat copies. Each read is realigned to these artificial reference sequences, starting from a reference with only one copy of the motif and growing. The reference with the highest realignment score is used to derive the number of repeat copies in the read. The realignment is performed using an efficient implementation of Smith-Waterman algorithm [6].

References

[1] Steven G Johnson. *The NLopt nonlinear-optimization package*. 2014. URL: http://ab-initio.mit.edu/nlopt.

[2] E. S. Lander and M. S. Waterman. "Genomic mapping by fingerprinting random clones: a mathematical analysis". In: *Genomics* 2.3 (Apr. 1988), pp. 231–239.

[3] Michael JD Powell. "A direct search optimization method that models the objective and constraint functions by linear interpolation". In: *Advances in optimization and numerical analysis*. Springer, 1994, pp. 51–67.

[4] H. Tang et al. "Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes". In: *Am. J. Hum. Genet.* 101.5 (Nov. 2017), pp. 700–715.

[5] T. Willems et al. "Genome-wide profiling of heritable and de novo STR variations". In: *Nat. Methods* 14.6 (June 2017), pp. 590–592.

[6] Mengyao Zhao et al. "SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications". In: *PloS one* 8.12 (2013), e82138.