APPENDIX A

DOUBLE-BARREL SHOTGUN ALGEBRA

I define the length of an insert to be 1, without loss of generality. Insert length is assumed to be constant. For this appendix, in a departure from the notation of the main text, I assume a constant sequence read length of length f, with f [0,0.5]. If f>0.5, the fragments overlap at the center of the insert. To maintain as much simplicity as possible, I will consider only the limiting case as $T \rightarrow 0$. I will explicitly calculate only the probability of obtaining a single scaffold (i.e., all clones form one scaffold). However, the equations developed along the way can be adapted with a healthy dose of algebra to give distributions for the lengths of scaffolds as well as other variables of interest. I assume stepwise addition of characterized inserts to a project, in order to aid my descriptions. I will refer to these inserts (i.e., "clones") as C_1, C_2, or C_3, with the subscript designating the order of addition to the project. Adding all the clones simultaneously does not alter the results. I assume a circular target with $G>n$. A less restrictive assumption on target size can be made with little loss of accuracy.[1]

To aid my descriptions I will refer to each sequence read by whether or not it is the right or left read from an insert, with a subscript for which of the inserts the sequence read is obtained from. Thus the left read from C_1 is designated L_1; the right read from C_3 is R_3.

I will use the symbol $\cap$ to designate overlap. I will use S to indicate the state of a project with a single scaffold. Therefore $P(S|L_1 \cap L_2)$ means "the probability of a project having a single scaffold given that the left end of clone 1 overlaps the left end of clone 2." $P(S|(L_1 \cap L_2)\wedge\sim(L_1 \cap R_3))$ means "the probability of a project having a single scaffold given that the left end of clone 1 overlaps the left end of clone 2 but not the right end of clone 3."

---

[1] If G is smaller than this, it becomes possible for islands to wrap around on themselves. This alters the number and character of the possible topologies. It will tend to increase the probability of a single island, possibly quite significantly. If the target is linear, slight "edge effects" alter the calculations. These effects tend to be minor, particularly if $G>>I$. Therefore, the calculations presented here as exact for really big circular targets are actually better approximations to real linear targets than they are to real circular targets.

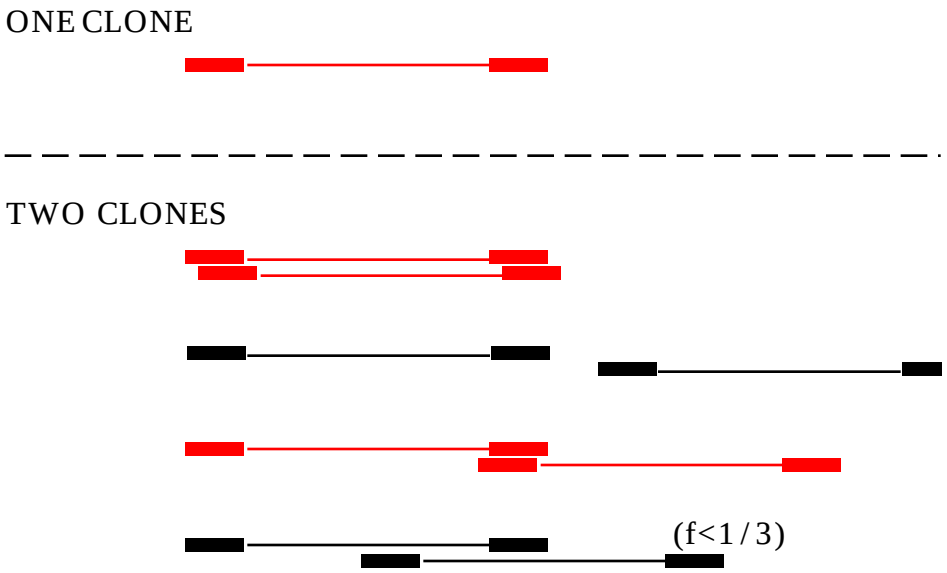

Figure A.1. Topologies for one- and two-clone double-barrel configurations. Single-scaffold topologies are highlighted in red. In some instances, sequence islands will span regions depicted here as SMGs.

To aid in the descriptions that follow, cartoon sketches of possible clone topologies are provided in Figure A.1 and Figure A.2.

A.1 ONE CLONE

One clone, by definition, will always form one scaffold.

A.2 TWO CLONES

There are two cases to consider.

1.15.1 CASE 1 $\frac{1}{2} \geq f \geq \frac{1}{3}$

The second clone will intersect the first clone with probability 2/*G*. Note that the average number of clones in a scaffold will be:

$$\frac{1\left(\frac{2}{G}\right) + 2\left(1 - \frac{2}{G}\right)}{2\left(\frac{2}{G}\right) + 1\left(1 - \frac{2}{G}\right)} = \frac{2 - \frac{2}{G}}{1 + \frac{2}{G}}$$

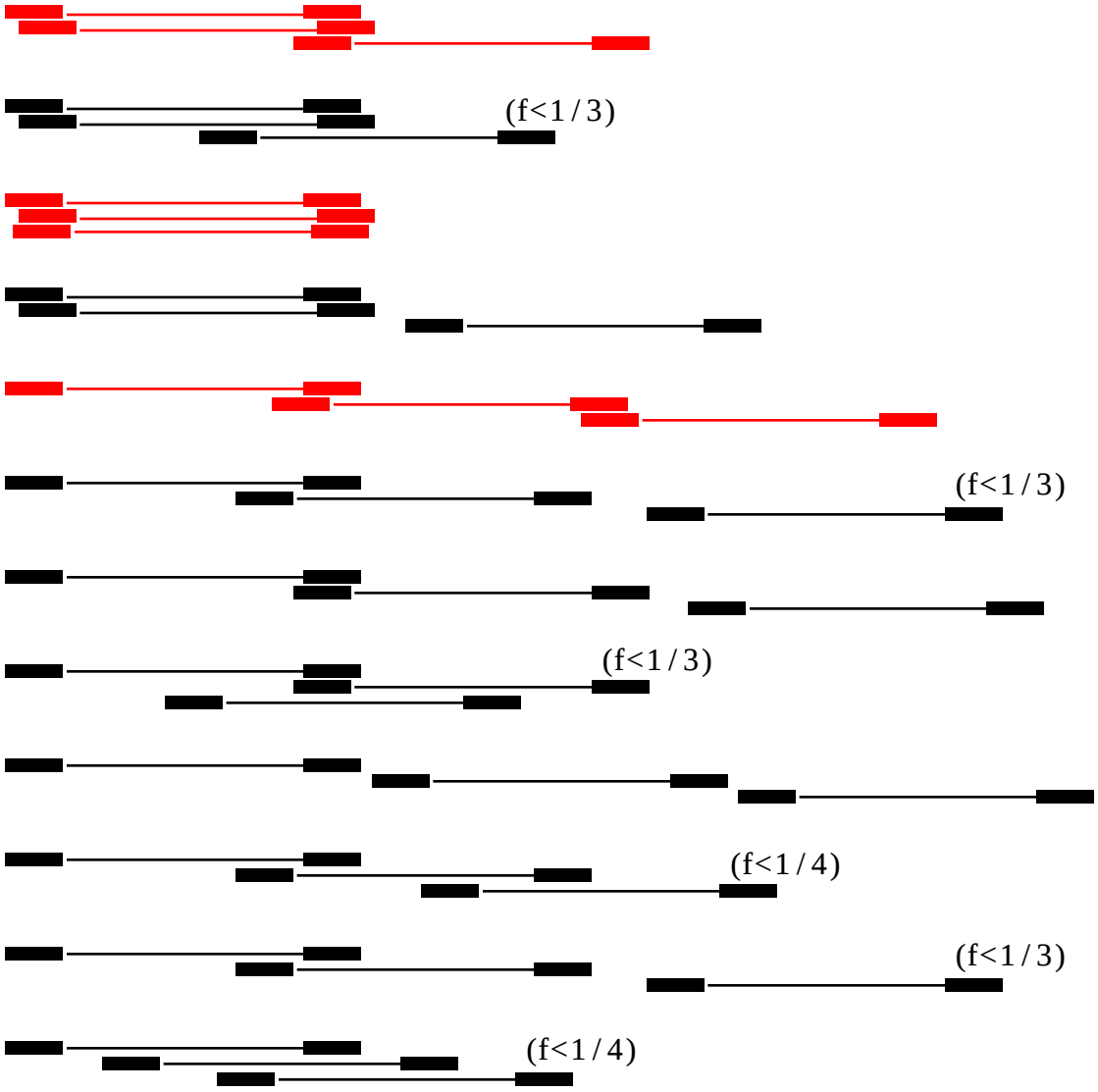

Figure A.2. Topologies for three-clone double-barrel configurations. The number of possible topologies rises rapidly with the number of clones. Single-scaffold topologies are highlighted in red. The first and third configurations shown here represent any topology in which their overlapping sequence ends form a sequence island, even if all three ends do not mutually intersect. In some instances, sequence islands will span regions depicted here as SMGs.

1.15.2 CASE 2 $f < \frac{1}{3}$

The second clone will intersect the first clone with probability $6f/G$. Note that the average number of clones in a scaffold will be:

$$\frac{1\left(\frac{6f}{G}\right) + 2\left(1 - \frac{6f}{G}\right)}{2\left(\frac{6f}{G}\right) + 1\left(1 - \frac{6f}{G}\right)} = \frac{2 - \frac{6f}{G}}{1 + \frac{6f}{G}}$$

A.2 THREE CLONES

The cases with one and two clones were relatively simple. The case with three clones is also simple, in that no complex mathematics is needed, but some care must be taken to ensure that all the topologies are properly accounted for. There are three cases to consider.

1.15.1 CASE 1 $\frac{1}{2} \geq f \geq \frac{1}{3}$

When $f \geq \frac{1}{3}$, a fragment cannot fall in the unsequenced gap between the two characterized ends of a clone without overlapping at least one of the ends. This limits the number of topologies that must be considered.

There are two possible configurations for the first two clones. Either they overlap, or they do not.

configuration 1 $C_2 \cap C_1$

This configuration will occur with probability $2/G$. All possible extents of overlap are equally likely. Fixing the leftmost end of C_1 at 1, then the leftmost end of C_2 will vary from 0 to 2. The expected probability of C_3 overlapping either or both of C_1 and C_2 can be calculated as

$$\frac{\frac{1}{G}\int_0^1 (3 - x)\,dx + \frac{1}{G}\int_1^2 (1 + x)\,dx}{2} = \frac{5}{2G}$$

For example, when the leftmost end of C_2 is at the origin, the leftmost end of C_3 can be anywhere from -1 to 2 and still overlap C_1 or C_2, giving it a probability of $3/G$ of joining the scaffold. Due to symmetry, the two integrals contribute equally.

configuration 2 $\sim(C_2 \cap C_1)$

This configuration will occur with probability $1-\frac{2}{G}$. All possible non-overlapping states are equally likely. Fixing the rightmost end of C_1 at 0, and letting x represent the rightmost end of C_2, the probability of C_3 overlapping both of C_1 and C_2 is:

$$\frac{\frac{1}{G}\int_0^1 (1-x)\,dx + \frac{1}{G}\int_1^{G-3} 0\,dx + \frac{1}{G}\int_{G-3}^{G-2} \left(x-(G-3)\right)dx}{G-2} = \frac{1}{G(G-2)}$$

Again, due to symmetry, the first and last integrals give the same result.

combining the probabilities for the two configurations

The probability of obtaining a single scaffold given $\frac{1}{2} \geq f \geq \frac{1}{3}$ is the sum of the probabilities from the last two sub-sections:

$$P(S) = \left(1-\frac{2}{G}\right)\frac{1}{G(G-2)} + \frac{2}{G}\frac{5}{2G} = \frac{6}{G^2}$$

1.15.2 CASE 2 $\frac{1}{3} \geq f \geq \frac{1}{4}$

There are four possible topologies for the first two clones.

configuration 1 $L_1 \cap L_2$

This configuration will occur with probability $2f/G$. Explicitly writing out the probability of the third clone intersecting the scaffold formed by the first two clones:

$P(S)=$

$P(S|R_3 \cap L_1)$

$+ P[S|(R_3 \cap R_1) \wedge \sim(R_3 \cap L_1)]$

$+ P[S|(R_3 \cap L_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1)]$

$+ P[S|(R_3 \cap R_2) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1) \wedge \sim(R_3 \cap L_2)]$

$+ P[S|(L_3 \cap L_1) \wedge \sim(R_3 \cap L_1) \wedge \sim(R_3 \cap R_1) \wedge \sim(R_3 \cap L_2) \wedge \sim(R_3 \cap R_2)]$

$+ P[S|(L_3 \cap R_1) \wedge \sim (R_3 \cap L_1) \wedge \sim (R_3 \cap R_1) \wedge \sim (R_3 \cap L_2) \wedge \sim (R_3 \cap R_2) \wedge \sim (L_3 \cap L_1)]$

$+ P[S|(L_3 \cap L_2) \wedge \sim (R_3 \cap L_1) \wedge \sim (R_3 \cap R_1) \wedge \sim (R_3 \cap L_2) \wedge \sim (R_3 \cap R_2) \wedge \sim (L_3 \cap L_1) \wedge \sim (L_3 \cap R_1)]$

$+ P[S|(L_3 \cap R_2) \wedge \sim (R_3 \cap L_1) \wedge \sim (R_3 \cap R_1) \wedge \sim (R_3 \cap L_2) \wedge \sim (R_3 \cap R_2) \wedge \sim (L_3 \cap L_1) \wedge \sim (L_3 \cap R_1) \wedge \sim (L_3 \cap L_2)] =$

$2f/G +$

$+ 2f/G$

$+ \frac{-1 + 8f - 14f^2}{4fG}$

$+ \frac{-1 + 8f - 14f^2}{4fG}$

$+ 0$

$+ \frac{8f - 8f^2 - 1}{4fG}$

$+ 0$

$+ \frac{-1 + 8f - 14f^2}{4fG} =$

$$\frac{8f - 8f^2 - 1}{4fG} + 3\left(\frac{-1 + 8f - 14f^2}{4fG}\right) + \frac{16f^2}{4fG} =$$

$$\frac{-2 + 16f - 17f^2}{2fG}$$

The following identities help in the above calculation:

1. $P(S|R_3 \cap L_1) = 2f/G$

2. R_3 cannot hit L_1 if R_3 hits R_1 so $P[S|(R_3 \cap R_1) \wedge \sim (R_3 \cap L_1)] = P(S|R_3 \cap L_1) = 2f/G$

3. $P[S|(R_3 \cap L_2)\wedge\sim(R_3 \cap L_1)\wedge\sim(R_3 \cap R_1)]=$

$$\frac{\frac{1}{G}\left(\int_0^f (f-x)\,dx+\int_f^{4f-1}(x-f)\,dx+\int_{4f-1}^{2f}(x-f)-(x-(1-2f))\,dx\right)}{2f}=$$

$$\frac{\frac{1}{G}\left(\int_0^f (f-x)\,dx+(x-f)+\int_{1-2f}^{2f}(1-3f)\,dx\right)}{2f}=$$

$$\frac{\frac{1}{G}\left(\left[fx-\frac{1}{2}x^2\right]_0^f+\left[\frac{1}{2}x^2-fx\right]_f^{1-2f}+(1-3f)(4f-1)\right)}{2f}=$$

$$\frac{\frac{1}{G}\left(\frac{f^2}{2}+\frac{1-6f+9f^2}{2}+\left(-12f^2+7f-1\right)\right)}{2f}=$$

$$\frac{\frac{1}{G}\left(\frac{-1+8f-14f^2}{2}\right)}{2f}=$$

$$\frac{-1+8f-14f^2}{4fG}$$

4. If R_3 hits R_2 it cannot hit L_2 so:

$P[S|(R_3 \cap R_2)\wedge\sim(R_3 \cap L_1)\wedge\sim(R_3 \cap R_1)\wedge\sim(R_3 \cap L_2)]=P[S|(R_3 \cap R_2)\wedge\sim(R_3 \cap L_1)\wedge\sim(R_3 \cap R_1)]=$

$$\frac{-1+8f-14f^2}{4fG}$$

and this probability is in turn equal to $P[S|(R_3 \cap L_2)\wedge\sim(R_3 \cap L_1)\wedge\sim(R_3 \cap R_1)]$.

5. If L_3 hits L_1 then R_3 hits R_1, so:

$P[S|(L_3 \cap L_1)\wedge\sim(R_3 \cap L_1)\wedge\sim(R_3 \cap R_1)\wedge\sim(R_3 \cap L_2)\wedge\sim(R_3 \cap R_2)]=0$

6. If L_3 hits R_1 then L_3 cannot hit L_1 and R_3 cannot hit L_2, L_1, or R_1 so:

$P[S|(L_3 \cap R_1)\wedge\sim(R_3 \cap L_1)\wedge\sim(R_3 \cap R_1)\wedge\sim(R_3 \cap L_2)\wedge\sim(R_3 \cap R_2)\wedge\sim(L_3 \cap L_1)]=$

$P[S|(L_3 \cap R_1) \wedge \sim (R_3 \cap R_2)] =$

$$\frac{\frac{1}{G}\left(\int_0^{2f} 2f\,dx + \int_{1-2f}^{2f} (1-x)\,dx\right)}{2f} =$$

$$\frac{\frac{1}{G}\left(\left(2f - 4f^2\right) + \left(2f - \frac{1}{2}\right)\right)}{2f} =$$

$$\frac{\frac{1}{G}\left(4f - 4f^2 - \frac{1}{2}\right)}{2f} =$$

$$\frac{8f - 8f^2 - 1}{4fG}$$

7. If L_3 hits L_2 then R_3 hits R_2 so:

$P[S|(L_3 \cap L_2) \wedge \sim (R_3 \cap L_1) \wedge \sim (R_3 \cap R_1) \wedge \sim (R_3 \cap L_2) \wedge \sim (R_3 \cap R_2) \wedge \sim (L_3 \cap L_1) \wedge \sim (L_3 \cap R_1)] = 0$

8. If L_3 hits R_2 then: R_3 cannot hit L_1, L_2, or R_2; L_3 cannot hit L_2; L_3 hits L_1 if and only if R_3 hits R_1. Therefore:

$P[S|(L_3 \cap R_2) \wedge \sim (R_3 \cap L_1) \wedge \sim (R_3 \cap R_1) \wedge \sim (R_3 \cap L_2) \wedge \sim (R_3 \cap R_2) \wedge \sim (L_3 \cap L_1) \wedge \sim (L_3 \cap R_1) \wedge \sim (L_3 \cap L_2)] =$

$P[S|(L_3 \cap R_2) \wedge \sim (R_3 \cap R_1) \wedge \sim (L_3 \cap R_1)] =$

$$\frac{-1 + 8f - 14f^2}{4fG}$$

This last equality is symmetrical to the case in identity 4.

configuration 2 $\sim(C_1 \cap C_2)$

Since C_1 and C_2 do not overlap, in order for there to be one scaffold after the addition of C_3, it must link C_1 and C_2. I consider only the case where C_1 is to the left of C_2. This case is symmetrical with the case where C_1 lies to the right of C_2. The probability of one or the other of these cases occurring is $1 - \frac{2}{G}$. However, with probability $1 - \frac{4}{G}$, the first two clones will be greater than a clone length apart, with zero chance that the third clone will link them into one scaffold. I will condition the following probabilities on (i) C_1 lies to the left of C_2,

and (ii) C_1 and C_2 are not separated by more than one clone length. With these conditions, the probability of C_3 hitting both C_1 and C_2 is:

$P[S|(R_3 \cap R_1)\wedge(R_3 \cap L_2)]+P[S|(L_3 \cap R_1)\wedge(L_3 \cap L_2)]+P[S|(L_3 \cap R_1)\wedge(R_3 \cap L_2)]=$

$$\frac{-1+8f-6f^2}{2G}$$

The following identities help in the above calculation:

1. $P[S|(R_3 \cap R_1)\wedge(R_3 \cap L_2)]= P[S|(L_3 \cap R_1)\wedge(L_3 \cap L_2)]= \int_0^f (f-x)\,dx + \int_f^1 0 = \frac{f^2}{2G}$

2. $P[S|(L_3 \cap R_1)\wedge(R_3 \cap L_2)]= \int_0^{1-2f} (x+4f-1)dx + \int_{1-2f}^1 (1-x)\,dx = \frac{-1+8f-8f^2}{2G}$

configuration 3 $R_1 \cap L_2$

This configuration will occur with probability $2f$. Note that this is symmetrical with the configuration $L_1 \cap R_2$, which will also occur with probability $2f$, so I will not explicitly address $L_1 \cap R_2$. Explicitly writing out the probability of the third clone intersecting the scaffold formed by the first two clones:

$P(S)=$

$P(S|R_3 \cap L_1)$

$+ P(S|R_3 \cap R_1)$

$+ P[S|(R_3 \cap L_2)\wedge\sim(R_3 \cap L_1)\wedge\sim(R_3 \cap R_1)]$

$+ P[S|(R_3 \cap R_2)\wedge\sim(R_3 \cap R_1)]$

$+ P[S|(L_3 \cap R_1)\wedge\sim(R_3 \cap L_2)\wedge\sim(R_3 \cap R_2)]$

$+ P[S|(L_3 \cap R_2)\wedge\sim(L_3 \cap R_1)]=$

$$\frac{-1+8f-7f^2}{fG}$$

The following identities help in the above calculation:

1. $P(S|R_3 \cap L_1)=P(S|R_3 \cap R_1)=2f/G$

2. $P[S|(R_3 \cap L_2) \wedge \sim (R_3 \cap L_1) \wedge \sim (R_3 \cap R_1)] = \frac{-1 + 8f - 14f^2}{4fG}$

3. $P[S|(R_3 \cap R_2) \wedge \sim (R_3 \cap R_1)] = P[S|(L_3 \cap R_2) \wedge \sim (L_3 \cap R_1)] =$

$$\frac{\int_0^{4f-1} (x + 1 - 2f)dx + \int_{4f-1}^{2f} 2f\,dx}{2fG} = \frac{-1 + 8f - 8f^2}{4fG}$$

4. $P[S|(L_3 \cap R_1) \wedge \sim (R_3 \cap L_2) \wedge \sim (R_3 \cap R_2)] = P[S|(L_3 \cap R_1) \wedge \sim (R_3 \cap L_2) \wedge \sim (L_3 \cap L_2)] =$

$$\frac{\frac{1}{G}\left(\int_0^f (f - x)dx + \int_f^{1-2f} (x - f)dx + \int_{1-2f}^{2f} (1 - 3f)dx\right)}{2f} =$$

$$\frac{-1 + 8f - 14f^2}{4fG}$$

configuration 4 $(C_1 \cap C_2) \wedge \sim ((R_1 \ L_1) \cap (R_2 \ L_2))$

This is the case where the clones overlap, but still form two scaffolds because none of their characterized ends overlap. The clones are interleaved. Again, this can occur in two symmetrical states, one with the left end of C_1 to the left of the left end of C_2, and one with C_1 to the right. Each of these states will occur with probability $\frac{1-3f}{G}$. I will condition on the state with C_1 to the left in the following. Explicitly writing out the probability of the third clone intersecting the scaffold formed by the first two clones:

$P(S) =$

$P[S|(R_3 \cap L_1) \wedge (R_3 \cap L_2)]$

$+ P[S|(R_3 \cap R_1) \wedge (R_3 \cap L_2)]$

$+ P[S|(R_3 \cap R_1) \wedge (R_3 \cap R_2)]$

$+ P[S|(L_3 \cap R_1) \wedge (L_3 \cap R_2)] =$

$$\frac{10f - 2}{G}$$

The following identity helps in the above calculation:

$P[S|(R_3 \cap L_1) \wedge R_3 \cap L_2)] = P[S|(R_3 \cap R_1) \wedge (R_3 \cap L_2)] = P[S|(R_3 \cap R_1) \wedge (R_3 \cap R_2)] =$

$$P[S|(L_3 \cap R_1) \wedge (L_3 \cap R_2)] = \frac{\int_0^{1-3f} (f-x)dx}{(1-3f)G} = \frac{5f-1}{2G}$$

combining the probabilities for the four configurations

With each possible configuration represented by a roman numeral, we have:

$$P(S)=P(i)P(S|i)+P(ii)P(S|ii)+P(iii)P(S|iii)+P(iv)P(S|iv)=$$

$$\left(\frac{2f}{G}\right)\left(\frac{-2+16f-17f^2}{2fG}\right)+2\left(\frac{1}{G}\right)\left(\frac{-1+8f-6f^2}{2G}\right)$$

$$+2\left(\frac{2f}{G}\right)\left(\frac{-1+8f-7f^2}{fG}\right)+2\left(\frac{1-3f}{G}\right)\left(\frac{-2+10f}{G}\right)=$$

$$\frac{-11+88f-111f^2}{G^2}$$

1.15.3 CASE 3 $\frac{1}{4} \geq f \geq 0$

There are four possible topologies for the first two clones. These are the same topologies considered in Case 2. However, the probabilities must be computed differently, as some of the final topologies that were previously impossible are now possible. Each of the configurations is as described in Case 2, so I will omit explanations where they are identical to Case 2.

configuration 1 $L_1 \cap L_2$

$P(S)=$

$P(S|R_3 \cap L_1)$

$+ P(S|R_3 \cap R_1)$

$+ P[S|(R_3 \cap L_2) \wedge \sim(R_3 \cap L_1)]$

$+ P[S|(R_3 \cap R_2) \wedge \sim(R_3 \cap R_1)]$

$+ P(S|L_3 \cap R_1)$

$+ P[S|(L_3 \cap R_2) \wedge \sim(L_3 \cap R_1)]=$

$2f/G+2f/G+f/2G+f/2G+2f/G+f/2G=15f/2G$

configuration 2 $\sim(C_1 \cap C_2)$

$P(S)=P[S|(R_3 \cap R_1)\wedge(R_3 \cap L_2)]+P[S|(L_3 \cap R_1)\wedge(L_3 \cap L_2)]+P[S|(L_3 \cap R_1)\wedge(R_3 \cap L_2)]=\frac{5f^2}{G}$

The following identities help in the above calculation:

1. $P[S|(R_3 \cap R_1)\wedge(R_3 \cap L_2)]=P[S|(L_3 \cap R_1)\wedge(L_3 \cap L_2)]= \int_0^f (f-x)\,dx + \int_f^1 0 = \frac{f^2}{2G}$

2. $P[S|(L_3 \cap R_1)\wedge(R_3 \cap L_2)]= \int_0^{1-4f} 0\,dx + \int_{1-4f}^{1-2f} (x+4f-1)\,dx + \int_{1-2f}^{1} (1-x)\,dx = \frac{4f^2}{G}$

configuration 3 $R_1 \cap L_2$

$P(S)=P(S|R_3 \cap L_1)$

$+ P(S|R_3 \cap R_1)$

$+ P[S|(R_3 \cap L_2)\wedge\sim(R_3 \cap R_1)]$

$+ P(S|R_3 \cap R_2)$

$+ P[S|(L_3 \cap R_1)\wedge\sim(L_3 \cap L_2)]$

$+ P(S|L_3 \cap R_2)=$

$\frac{9f}{G}$

configuration 4 $(C_1 \cap C_2)\wedge\sim((R_1 \ \ L_1)\cap(R_2 \ \ L_2))$

$P(S)=$

$P[S|(R_3 \cap L_1)\wedge(R_3 \cap L_2)]$

$+ P[S|(R_3 \cap R_1)\wedge(R_3 \cap L_2)]$

$+ P[S|(R_3 \cap R_1)\wedge(R_3 \cap R_2)]$

$+ P[S|(L_3 \cap R_1)\wedge(L_3 \cap R_2)]=$

$\frac{2f^2}{(1-3f)G}$

The following identity helps in the above calculation:

$P[S|(R_3 \cap L_1)\wedge R_3 \cap L_2)]=P[S|(R_3 \cap R_1)\wedge(R_3 \cap L_2)]=P[S|(R_3 \cap R_1)\wedge(R_3 \cap R_2)]=$

$$P[S|(L_3 \cap R_1) \wedge (L_3 \cap R_2)] = \frac{\int_0^f (f - x)\,dx + \int_f^{1-3f} 0}{(1-3f)G} = \frac{f^2}{2(1-3f)G}$$

combining the probabilities for the four configurations

$P(S)=P(i)P(S|i)+P(ii)P(S|ii)+P(iii)P(S|iii)+P(iv)P(S|iv)=$

$$\frac{65 f^2}{G}$$

A.3 MORE THAN THREE CLONES

The algebra quickly gets more difficult. However, a few simple statements can be made.

As *f* becomes a smaller fraction of the clone length, more topologies become possible. This is the same as holding *f* constant and increasing the clone length, which is what is done in actual practice when longer inserts are used to build clone libraries. This is a partial explanation for why long clones are better. The more topologies the clones have available to them in order to form a single scaffold, the more likely they are to form a single scaffold. The probability of clones forming a particular topology never drops as the ratio of clone length to *f* increases. This probability is illustrated in Figure A.3 for the case of three inserts.

It is interesting to note that this curve is not smooth at the ratio of three. I predict that this is the only point at which this curve will not be smooth, regardless of the number of inserts (note that the curve is not defined for ratios less than two). Below a ratio of three, the opposite ends of the insert are present in each other's potential region of overlap. This results in a concave curve. Above a ratio of three, I predict that the curve will be both smooth and convex.

The number of topologies obviously does not rise continuously as *f* diminishes. Rather the number of topologies takes discrete jumps at each $f \in \{1/x | x \text{ is an integer} \leq n\}$. The maximum number of topologies is reached when f=1/*n*, so there is no further gain in the number of topologies by increasing the clone length beyond *nf*. Since, in practice, *n* is universally greater than 100, and *f* for sequencing projects is at least 400 bp, this implies no topology gains for subclones longer than 40 kb. Since 10 kb is the maximum routine length for sequencing templates, this limit will not be reached in practice.

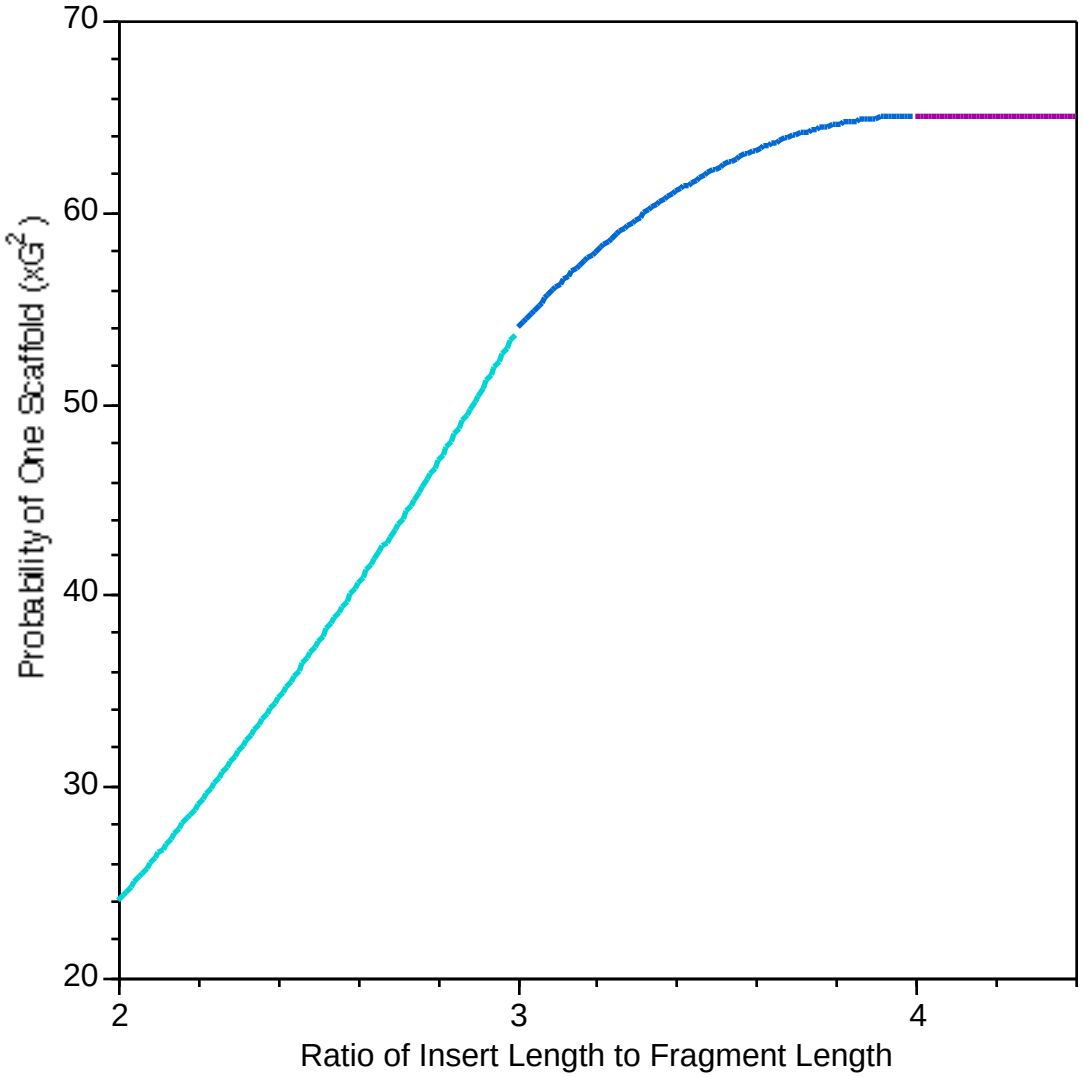

Figure A.3. For the case of three inserts, the probability of obtaining one scaffold versus the ratio of insert length to fragment length. The probability increases as the insert length is increased (and the fragment length remains constant). The character of the increase changes with each discrete increase in the ratio of insert length to read length, eventually reaching a plateau.

Evaluating the number of topologies is not a simple task. Nevertheless, it is undoubtedly feasible, given sufficient effort. Lee Newberg (1993 and 1996) has evaluated the number of possible topologies for traditional random subcloning maps.[2] By his counting, the number of topologies for n=2 is 2, for n=3 is 10, and for n=4 is 94. This number rises exponentially, with 1.3×10^{10} topologies at n=10 and 3.2×10^{27} topologies at n=20. The number of pairwise topologies will rise even faster. Furthermore, the smaller the ratio of f to clone length (up to the $1/n$ limit), the faster the rise.

The number of topologies affects primarily the probability of the clones forming a single scaffold. This will also tend to increase the length of the average scaffold. Another factor influencing the length of the average scaffold will be the clone length. With the number of topologies held constant (i.e., with $1/x - 1 < f < 1/x$), the average scaffold length will rise proportionally with the clone length. This rise will continue even after the limit of $f < 1/n$ is reached.

Therefore there is always a monotonically increasing rise in average scaffold length as insert length is increased. I postulate that this curve will be convex, indicating that the most benefit from increasing insert length will occur with shorter inserts (with the exception of insert:fragment ratios below three). I have not proved this postulate.

Despite the increase in scaffold length with respect to insert length, this increase is not always useful. Consider the limiting case of a single clone (n=1). The single scaffold formed will grow linearly in length with respect to insert length. If the insert length equals the target length, this scaffold will be complete by definition. This increase in scaffold length is however of little use to the researcher: as the completeness (i.e., resolution) of the map decreases, the density of SMGs diminishes. Increases in scaffold length accompanied by increases in SMG density are what is needed at the lab bench. Such an increase cannot be gained by longer clones alone; the longer clones must be accompanied by an increase in available topologies. Thus, although there is a theoretical gain in scaffold length above a clone:read ratio of n, there is no further practical gain. Again, this last point can be considered trivia, as the clone:read ratio will never reach n in practice.[3]

---

[2] Newberg uses a slightly different definition of topology. He treats the clones as distinguishable. However, his results are easily adaptable.

Much work remains to be done on this difficult problem and it may be that approaches that remain hidden today will ultimately provide more insight for a mathematical model for pairwise end sequencing.

---

[3] One might imagine that it could happen during a BAC end sequencing project, where the clone:read ratio is approximately 100000:400 = 250. However, such sequences are poorer quality sequences that will not usually be used for constructing finished sequence. They would rarely be used as part of a project undertaken strictly as outlined in this dissertation. Additionally, although 250 clones might be used for analyzing a small target, 250 is significantly less than the actual *n* that would be used for a human-genome-sized project (see Venter et al., 1996; Siegel et al., 1999).