Supplemental Information

Table 1S: Primers Used for Polymerase Chain Reaction (PCR).

Primer Name	Sequence (5'→3')	Target
kps Reg1 KO#1	GTTACAACCCATTGATTTAGCATAAATAAATTATAGTGGGTTCGGGTTTGT TGTGTAGGCTGGAGCTGCTTC	*kps* Region I, 5' / pKD4
kps Reg1 KO#2	TGGAAATGATTTTTTGGCTACTTAAAATTCAAAAGATATTGACTTGAAATAT GGGAATTAGCCATGGTCC	*kps* Region I, 3' / pKD4
kps Reg1 #1	ATGTTCCCGGTGGTCAACATGCTTCCAGCACTCCTT	*kps* Region I, 5'
kps Reg1 #2	CCTCTTTGCACGATAAAAGGATTTTCTTG	*kps* Region I, 3'
BP 1F	GTACCGCGCTTAAACGTTCAG	BAR-PATH
BP 1R	AATTctgaacgtttaagcgcggtacAGC	BAR-PATH
BP 2F	GTACCGCGCTTAAATAGCCTG	BAR-PATH
BP 2R	AATTcaggctatttaagcgcggtacAGC	BAR-PATH
BP 3F	GTACCGCGCTTAAAAGTCTCG	BAR-PATH
BP 3R	AATTcgagacttttaagcgcggtacAGC	BAR-PATH

BP 4F	GTACCGCGCTTAATAACGTGG	BAR-PATH
BP 4R	AATTccacgttattaagcgcggtacAGC	BAR-PATH
BP 5F	GTACCGCGCTTAAACTGGTAG	BAR-PATH
BP 5R	AATTctaccagtttaagcgcggtacAGC	BAR-PATH
BP 6F	GTACCGCGCTTAAGCATGTTG	BAR-PATH
BP 6R	AATTcaacatgcttaagcgcggtacAGC	BAR-PATH
BP 7F	GTACCGCGCTTAATGTAACCG	BAR-PATH
BP 7R	AATTcggttacattaagcgcggtacAGC	BAR-PATH
BP 8F	GTACCGCGCTTAAAATCTCGG	BAR-PATH
BP 8R	AATTccgagattttaagcgcggtacAGC	BAR-PATH
BP 9F	GTACCGCGCTTAATAGGCAAG	BAR-PATH
BP 9R	AATTcttgcctattaagcgcggtacAGC	BAR-PATH
BP 10F	GTACCGCGCTTAACAATCGTG	BAR-PATH
BP 10R	AATTcacgattgttaagcgcggtacAGC	BAR-PATH
BP 11F	GTACCGCGCTTAATCAAGACG	BAR-PATH

BP 11R	AATTcgtcttgattaagcgcggtacAGC	BAR-PATH
BP 12F	GTACCGCGCTTAACTAGTAGG	BAR-PATH
BP 12R	AATTcctactagttaagcgcggtacAGC	BAR-PATH
BP 13F	CTTGCGGCGTATTACGTTCAG	BAR-PATH
BP 13R	AATTctgaacgtaatacgccgcaagAGC	BAR-PATH
BP 14F	CTTGCGGCGTATTATAGCCTG	BAR-PATH
BP 14R	AATTcaggctataatacgccgcaagAGC	BAR-PATH
BP 15F	CTTGCGGCGTATTAAGTCTCG	BAR-PATH
BP 15R	AATTcgagacttaatacgccgcaagAGC	BAR-PATH
BP 16F	CTTGCGGCGTATTTAACGTGG	BAR-PATH
BP 16R	AATTccacgttaaatacgccgcaagAGC	BAR-PATH
BP 17F	CTTGCGGCGTATTACTGGTAG	BAR-PATH
BP 17R	AATTCTACCAGTAATACGCCGCAAGAGC	BAR-PATH
BP 18F	CTTGCGGCGTATTGCATGTTG	BAR-PATH
BP 18R	AATTCAACATGCAATACGCCGCAAGAGC	BAR-PATH

BP 19F	CTTGCGGCGTATTTGTAACCG	BAR-PATH
BP 19R	AATTCGGTTACAAATACGCCGCAAGAGC	BAR-PATH
BP 20F	CTTGCGGCGTATTAATCTCGG	BAR-PATH
BP 20R	AATTCCGAGATTAATACGCCGCAAGAGC	BAR-PATH
BP 21F	CTTGCGGCGTATTTAGGCAAG	BAR-PATH
BP 21R	AATTCTTGCGGCGTATTTAGGCAAGAGC	BAR-PATH
BP 22F	CTTGCGGCGTATTCAATCGTG	BAR-PATH
BP 22R	AATTCACGATTGAATACGCCGCAAGAGC	BAR-PATH
BP 23F	CTTGCGGCGTATTTCAAGACG	BAR-PATH
BP 23R	AATTCGTCTTGAAATACGCCGCAAGAGC	BAR-PATH
BP 24F	CTTGCGGCGTATTCTAGTAGG	BAR-PATH
BP 24R	AATTCCTACTAGAATACGCCGCAAGAGC	BAR-PATH
BP-2A	Tgattaagatgaattcatgggaattagccatggtcc	BAR-PATH
BP-2B	Tgattaagatgaattcgtgacacaggaacacttaacggctgac	BAR-PATH
BP-2C	tgattaagatgaattccgcactgagaagcccttagagcctc	BAR-PATH

BP-8K	gcttcaaaagcgctctgaagttcctatac	BAR-PATH
BP-8C	Cgtgccgatcaacgtctcattttcg	BAR-PATH
BP-1	gaaccgtaggccggataaggcgtttacgccgcatccggcacatagttaacagctcgtgtaggctggagctgc ttc	BAR-PATH
BP-5	/5Phos/ctacttcttcgcctctgcaaccactttgctacccacgccgcggttattgtattcc	BAR-PATH
BP-6	ggaatacaataaccgcggcgtgggtagcaaagtggttgcagaggcgaagaagtaggct	BAR-PATH

3

4

Mathematical model of tag diversity subsequent to a population bottleneck

As discussed in the main text, IBC formation is clonal and hypothesized to be the main contribution to bacterial persistence in the bladder. Assuming that bacteria carrying different tags are equivalent both in terms of IBC formation and in detection efficiency, the number of tags expected to be detected due solely to IBCs can be modeled as a multinomial distribution. Because tag detection does not differentiate between whether one or multiple IBCs contained that tag, we are interested simply in the number of different tags (multinomial outcomes) detected at all, and not in their relative abundance. In short, the problem of how many tags remain following a population bottleneck is identical to asking how many distinct bar codes are chosen when they are randomly sampled with replacement, the number of samples taken representing the population bottleneck. From the point of view of tags detected, we can calculate other populations from other bacterial niches by including additional samplings (multinomial trials), with one additional sample per bacterial clone.

This problem has been previously examined in (Ma, 2001), where formulas for the number of "live" terms in a complete multinomial expansion were presented - this corresponds precisely to our problem of calculating how many tags are expected to be detected in a given experiment. In the notation of (Ma, 2001), n is the number tags, representing mutually exclusive categories that result from a sampling (i.e. multinomial choices); k is the number of bacterial clones (either IBCs or clones included in other niches), or the number of multinomial trials; m is the number of tags detected, or those that have been sampled at least once in an experiment (live terms). A closed formula for the complete multinomial distribution is well known: $a_{k,p}^{i_1,i_2,\ldots,i_n} = n!/k!(n-k_1)!(n-k_2)!\ldots i_1^{k_1} i_2^{k_2} \ldots i_n^{k_n}$ However, as noted by (Ma, 2001), this is computationally intractable for values of n and k that we are interested in. (Ma, 2001) developed a closed formula for calculating individual terms of the multinomial expansion, but use of this requires iterating over all combinations of live terms, which is equivalent to finding all combinations of a set of k integers that sum to n. To translate from this into the distribution of total number of live terms is also computationally intractable for reasonable n.

Therefore, we have extended the reasoning used by (Ma, 2001) to directly calculate the number of multinomial samplings that have a given number of live terms. Our calculation has polynomial complexity because it does not explicitly calculate each term of the multinomial; instead, it calculates sums of distinct sets of the multinomial expansion, which are precisely the values that we are interested in.

We do a direct calculation of the number of ways that k samples from a multinomial with n possibilities, all with equal probability, results in exactly m outcomes occuring at least once. We let T_m be the number of possibilities that have exactly m live terms. For $m = 1$, this is trivial; for every sample, there is only 1 possibility, and thus there is $1^k = 1$ combination. However, there are n choices for which outcome is the only successful one; thus, there are $n \times 1 = n$ total ways that exactly one outcome is successful in all k samples: $T_1 = \binom{n}{1}1^k = n$.

For $m = 2$, each of the k samples has two possibilities, resulting in 2^k total combinations. However, 2^k counts all the possibilities of no more than 2 successful outcomes; this includes outcomes where only one of the two is successful (all k choices result in the same outcome) and thus these must be subtracted. There are $\binom{2}{1} = 2$ possibilities for the single successful outcome, thus we have $2^k - \binom{2}{1}1^k$ total outcomes where exactly two possibilities are successful. There are $\binom{n}{2}$ ways to pick which two outcomes are successful, giving us a total of $T_2 = \binom{n}{2}[2^k - \binom{2}{1}1^k]$ ways to have exactly two successful outcomes after k trials.

A similar reasoning is used for $m = 3$. Given the 3 outcomes, there are 3^k possibilities that have no more than those 3 outcomes; we must now subtract the possibilities that only 2 or 1 of the 3 desired outcomes is successful. When subtracting the possibilities that only 2 of the desired 3 outcomes is successful, we must only count those outcomes where exactly 2 of the outcomes is successful and not include the subset where only 1 of the 2 outcomes is successful. As above, the possibilities for exactly 2 successful outcomes is $2^k - \binom{2}{1}1^k$, and there are $\binom{3}{2}$ combinations of two outcomes, giving $\binom{3}{2}[2^k - \binom{2}{1}1^k]$. The correction for outcomes with only one successful outcome is $\binom{3}{1}[1^k]$. Again, there are $\binom{n}{3}$ ways of choosing a subset of three successful outcomes, giving a total of $T_3 = \binom{n}{3}[3^k - \binom{3}{2}[2^k - \binom{2}{1}1^k] - \binom{3}{1}[1^k]]$.

We now introduce a recursive definition for the term $c_{m,k} = m^k - \Sigma_{i=1}^{m-1}\binom{m}{i}c_{i,k}$ for $m > 1$, and $c_{1,k} = 1^k = 1$. The term $c_{m,k}$ represents the number of ways to get exactly m successful outcomes in k trials. Note that $c_{m,k}$

does not depend on the total number of possible outcomes. Then we can simplify the above expressions to:

$T_1 = \binom{n}{1} c_{1,k}$

$T_2 = \binom{n}{2} c_{2,k}$

$T_3 = \binom{n}{3} c_{3,k}$

In general, we have

$T_m = \binom{n}{m} c_{m,k}$

which is now, as expected, dependent on the total number of possible outcomes.

References

Ma, N. 2001. ”Complete multinomial expansions”. Applied Mathematics and Computation 124(3):365-70.

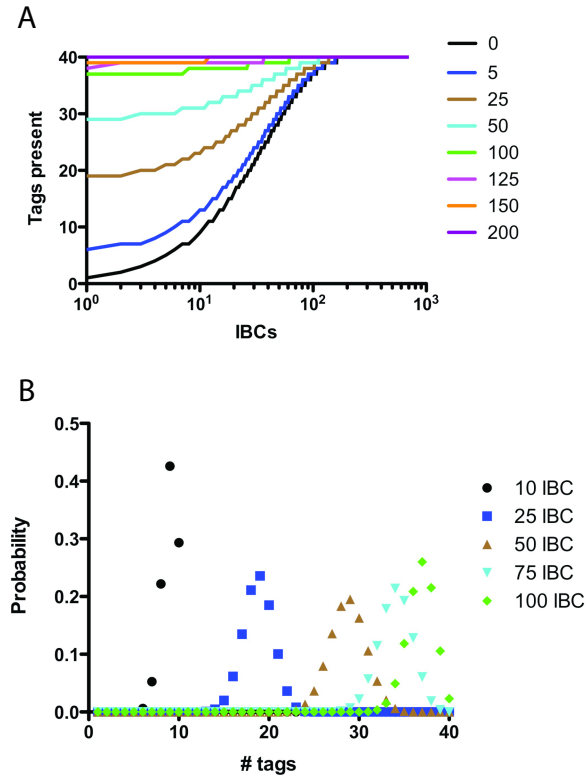

Figure 1. Predictive stochastic selection models of tag diversity in relationship to clonal intra- and extra-cellular communities. (A): Median number of tags (total = 40) expected based on a multinomial expansion with increasing contributions of tags from theoretical, clonal extracellular populations (see supplemental material). (B): The probability distributions for the likelihood of detecting a specific number of tags based on the amount of IBCs formed.