Article type: Focus Article

The Limits of Multiplexing

Dan Shen[1,*] , D.P. Dittmer [2,*] and J. S. Marron [3]

Keywords

Multiplexing, Genomics, nanostring$^{\text{TM}}$, DASL$^{\text{TM}}$, Probability

Abstract

We were motivated by three novel technologies, which exemplify a new design paradigm in high throughput genomics: nanostring$^{\text{TM}}$, *DNA-mediated Annealing, Selection, extension, and Ligation* (DASL)$^{\text{TM}}$ and multiplex real-time *quantitative polymerase chain reaction* (QPCR). All three are solution hybridization based, and all three employ on 10-1000 DNA sequence probes in a small volume, each probe specific for a particular sequence in a different human gene. Nanostring$^{\text{TM}}$ uses 50-mer, DASL and multiplex QPCR use $\sim$20-mer probes. Assuming a 1 nM probe concentration in a 1 μl volume, there are 10^{-9} x 10^{-9} x 6.23 x 10^{23} or 6.23 x 10^{5} molecules of each probe present in the reaction compared to 10-1,000 target molecules. Excess probe drives the sensitivity of the reaction. We are interested in the limits of multiplexing, i.e. the probability that in such a design a particular probe would bind to any other, sequence-related probe rather than the intended, specific target. If this were to happen with appreciable frequency, this would result in much reduced sensitivity and potential failure of this design. We established upper and lower bounds for the probability that in a multiplex assay at least one probe would bind to another sequence-related probe rather than its cognate target. These bounds are reassuring, because for reasonable degrees of multiplexing (10^3 probes) the probability for such an event is practically negligible. As the degree of multiplexing increases to $\sim 10^6$ probes, our theoretical boundaries gain practical importance and establish a principal upper limit for the use of highly multiplexed solution-based assays vis-$\grave{a}$-vis solid-support anchored designs.

Recently solution-based multiplex hybridization based methods have been developed and used for *messenger ribonucleic acid* (mRNA) profiling experiments that were previously the purview of solid-state, anchored methods the so-called microarrays or chips. By most practical accounts their performance seems

* Correspond to: danshen@usf.edu; ddittmer@med.unc.edu

[1] Interdisciplinary Data Sciences Consortium, Department of Mathematics and Statistics, University of South Florida.

[2] Department of Microbiology and Immunology and Center for AIDS Research, Comprehensive Cancer Center, University of North Carolina at Chapel Hill.

[3] Department of Statistics and Operations Research, Comprehensive Cancer Center, University of North Carolina at Chapel Hill.

equal, but practical experiments represent examples with a bias towards reporting positive data. They are not exhaustive and do not represent a general solution. Because massive multiplexing involves 1000 - 10^6 probes, individual experimental validation is no longer feasible.

At the time the reaction volume of *polymerase chain reaction* (PCR) and hybridization assays has been reduced due to nanotechnology. Conventional PCR instruments now can use the 1536-well format e.g. the Roche system with 1 μl. Newer microfluidics-based machines perform up to 20,000 individual reactions on the same chip e.g. Fluidigm system with 0.85 - 10 nanoliter. Recently, a picoliter device has been described by White et al. (2011). As the engineering pushes the technical boundaries of miniaturization, it becomes important to define the statistical boundaries of experimental designs.

Problem

We were concerned that the multiplex design introduces the potential for cross-hybridization among probe molecules, resulting in a loss of sensitivity or detection failure. This possibility is not present in solid-state, anchored designs, since in this case there are no free probes available in solution only the target molecules. Traditionally cross-hybridization refers to a scenario where the probe binds to a second, related but not the intended target. In solution-based multiplex designs there exists in addition the possibility that the probe cross-hybridizes to another probe rather than any target at all. Since every probe has to be in excess over any potential target in order to ”drive” the hybridization reaction to completion, cross-hybridization to an unrelated probe would be favored and would prevent detection of the cognate target. Here we determined theoretical bounds for this cross-hybridization problem.

Figure 1 lays out the problem. A microarray can be seen as a set S of probes in a special orientation. Each probe of length d (solid arrow) is physically attached to the support surface. Two probes never touch each other. These are then hybridized to a mixture of target mRNAs. The probes are in molar excess compared to the target. Each probe is an oligonucleotide of length d, i.e. a d-mer. The d-mer is made up of the four bases A, T, C, G. We assume that A and T have the same frequency, as do C and G. Furthermore. we parameterize the CG-ratio (i.e. the relative frequency of A or T) as p, where $0 \leq p \leq 1$. Each d-mer binds to the target mRNA with perfect complementarity and we assume 100 percent efficiency. We also assume that a given d-mer does not bind non-specifically to non-target mRNAs. These assumptions are supposed to hold in maximum efficient instruments or assays. In a solid-state anchored microarray, this is the only interaction that can take place (Figure 1 panel A). There is no binding of d-mer probes to each other.

The situation is different in a solution-hybridization based multiplex design, e.g. a bead array (Figure 1 panel B). In a bead array probes are coupled to individual beads and in principle any two beads with complementary probes (blue and red arrows in Figure 1) can hybridize to each other. Other design rely on free oligonuclotides/ probes with no

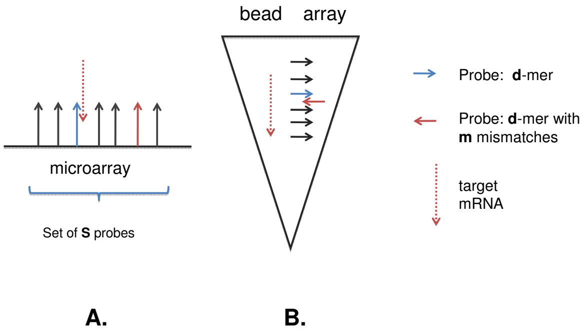

Figure 1: Conceptual illustrations of binding possibility for (A) solid-state anchored micro array, (B) multiplex solution array. Solid arrows indicate probes, the dotted arrow the correct target. The blue arrow refers to a probe or oligonucleotide of length d, which is desired, perfectly complementary to the target mRNA (dashed arrow). The red arrow refers to an oligonucleotide of length d, which is similar to the target mRNA and thus can bind the blue probe except for m mismatches. The blue-red interactions are possible due to sequence complementarity.

beads attached. In addition to each d-mer binding is cognate target mRNA, each d-mer can also bind to any other d-mer in the probe set. If the d-mer binds to another d-mer rather than the target mRNA, the assay fails.

It is important to keep in mind that for solution hybridization based designs the concentration of probe is orders of magnitude large than the target mRNA that is being detected. Otherwise, the assay would not be quantitative. Since hybridization efficiency is a function of probe concentration, unwanted probe d-mer to probe d-mer hybridizations poses a novel problem for solution-based multiplex approaches.

In a complete set that contains all possible d-mers, there exists for each d-mer one and only one perfectly complementary d-mer. For example, for a d-mer of sequence $ACTG$ the perfect complement would be $TGAC$. In a complete set S, all d-mer probes would hybridize to their complementary d-mer probe and none would bind to the target mRNA. This is avoided in traditional microarray designs, since all d-mers are spatially separated by anchoring them to a solid support matrix, i.e. a slide microarray or ”chip”.

- The size of the complete set, i.e. the number of all possible d-mers is 4^d.
- In praxi, one would never multiplex the complete set, but only a subset of all possible d-mers. This subset S_1 is much smaller than the complete set S.
- In praxi, not only perfectly completely complementary d-mers would hybridize, but also those with m mismatches. The number of mismatches being determined by the stringency of the hybridization reaction. The more mismatches are tolerated by the reaction conditions the more the size of S_1 approaches the size of the complete set S. Suppose $d = 4$, then the size of $S = 64$. Assume we only have

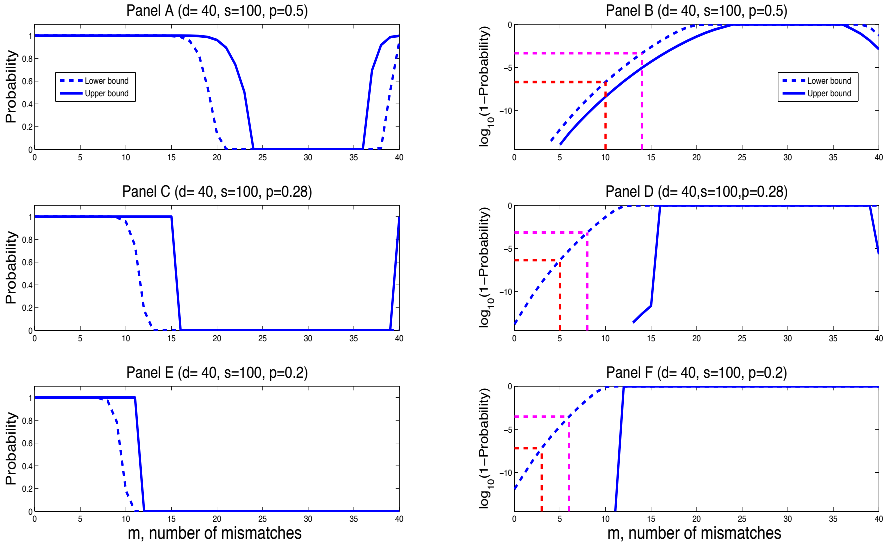

Figure 2: Upper (solid) and lower (dashed) bounds on the probability that there exist no m mismatched d-mers in a subset under the different CG-ratio: Show a absolute probability scale in the left panels. To depict closeness to 1 at higher resolution, the right panels are plotted on the log10(1-probability) scale. Rows compare affect of the CG-ratio p. Show upper and lower bounds are equally very closed. Dashed red and pink lines give specific interesting comparison.

> one probe of length $d = 4$, then size of $S_1 = 1$, e.g. $ACTG$. $S_1/S = 1/64$. Now we allow one mismatch at the end to yield : $ACTg$, $ACTt$, $ACTc$, $ACTa$. The size of $S_1 = 4$ and $S_1/S = 4/64$. If we allow 4 mismatches $S_1 = S$.

This paper gives useful theoretical bounds on how many d-mers can be multiplexed and how these depend on the length, the number of mismatches and the CG-ratio p.

There exists a vibrant literature on the probability of an individual d-mers and the complete set S of all possible sequence permutation, D'yachkov et al. (2005); Bishop et al. (2007); Dyachkov and Voronina (2009). Fewer studies have investigated this problem in the context of subsets S_1 of S and how subset size influences the probability of annealing.

Results

We defined no m mismatched d-mers in a subset S_1 if the number of mismatches between any two d-mers in a subset S_1 doesn't equal to m. The probability of no m mismatched d-mers in a subset S_1 of all possible d-mers is a function of s (the size of the subset S_1), of d (the length of the d-mer), of m (the number of mismatches) and of p (the CG-ratio). Panel A of the Figure 2 show our lower and upper bounds on the probability of no m mismatched d-mers in the subset S_1 for $s = 100$, $d = 40$ and $p = 0.5$. Lower and upper bounds on this probability are close to 1 when m is small, then decrease to 0 for m between 16 and 24 and increase again for m near $d = 40$.

Panels C and E show the lower and upper bounds plot with the same $s = 100$ and $d = 40$ but a different CG-ratio $p = 0.28$ and $p = 0.2$ respectively. The bound curves in Panels C and E have a similar trend as in Panel A. For smaller values of the CG-ratio, the sharp decrease from 1 to 0 happens for smaller m. Also the increase happens for larger m, and does not occur for $p = 0.2$ in Panel C.

Panels of Figure 2 also show that the lower and upper bounds on the probability of no m mismatched d-mers in the subset S_1 is almost equal to 1 when m is small. For example, the vertical scale in the panel A does not effectively distinguish the bounds from 1, for $m < 15$. A better visualization of this practically important range is achieved by applying the log function to 1 minus the bounds as shown in the second column panels of Figure 2.

These transformed plots clearly show the order of magnitude of the difference between the probability and 1. For example, the dashed red and pink lines in the plot show that to have a probability within $0.001 = 10^{-3}$ of 1, we need $m \leq 13, 8$, or 6, for the CG-ratios $p = 0.5, 0.28$, or 0.2 respectively, and to be within 10^{-6}, $m \leq 10, 5$, or 3 respectively.

As mentioned above, the combinatorial problem (Graham, 1995; Riordan, 2012) of finding a general closed form for the probability of no m mismatched d-mers in the subset S_1 is very challenging. This motivated us to instead find closed forms for the lower and upper bounds on the probability of no m mismatched d-mers in the subset S_1.

Without loss of generality, we assume that the CG-ratio $p \leq 0.5$. First, we need to introduce some notation. Given d, m and s, we define N_i^l and N_i^u, $i = 1, \cdots, s-1$, as

- If $i \leq 2^d$, $N_i^l = ip^d$ and $N_i^u = i(1-p)^d$.
- If $2^d \sum_{k=0}^{l-1} \binom{d}{k} < i \leq 2^d \sum_{k=0}^{l} \binom{d}{k}$, where $1 \leq l \leq d$, then $N_i^l = 2^d \sum_{k=0}^{l-1} \binom{d}{k} p^{d-k}(1-p)^k + p^{d-l}(1-p)^l(i - 2^d \sum_{k=0}^{l-1} \binom{d}{k})$ and $N_i^u = 2^d \sum_{k=0}^{l-1} \binom{d}{k} (1-p)^{d-k}p^k + (1-p)^{d-l}p^l(i - 2^d \sum_{k=0}^{l-1} \binom{d}{k})$.

In addition, we need to define M_i^l and M_i^u, $i = 1, \cdots, s-1$, as

- If $i \binom{d}{m} < 2^d$, then $M_i^l = i \binom{d}{m} (1+p)^m p^{d-m}$ and $M_i^u = i \binom{d}{m} (2-p)^m (1-p)^{d-m}$.
- If $2^d \sum_{k=0}^{l-1} \binom{d}{k} < i \binom{d}{m} \leq 2^d \sum_{k=0}^{l} \binom{d}{k}$, where $1 \leq l \leq d-m$, then $M_i^l = 2^d \sum_{k=0}^{l-1} \binom{d}{k} (1+p)^m (1-p)^k p^{d-m-k} + [i \binom{d}{m} - 2^d \sum_{k=0}^{l-1} \binom{d}{k}](1+p)^m (1-p)^l p^{d-m-l}$ and $M_i^u = 2^d \sum_{k=0}^{l-1} \binom{d}{k} (2-p)^m (1-p)^{d-m-k} p^k + [i \binom{d}{m} - 2^d \sum_{k=0}^{l-1} \binom{d}{k}](2-p)^m (1-p)^{d-m-l} p^l$.
- If $2^d \sum_{k=0}^{l-1} \binom{d}{k} < i \binom{d}{m} \leq 2^d \sum_{k=0}^{l} \binom{d}{k}$, where $d-m < l \leq d$, then $M_i^l = 2^d \sum_{k=0}^{d-m-1} \binom{d}{k} (1+p)^m (1-p)^k p^{d-m-k} + 2^d \sum_{k=d-m}^{l-1} \binom{d}{k} (2-p)^{k-d+m} (1+$

$p)^{d-k}(1-p)^{d-m} + [i\binom{d}{m} - 2^d \sum_{k=0}^{l-1}\binom{d}{k}](2-p)^{l-d+m}(1+p)^{d-l}(1-p)^{d-m}$ and $M_i^l = 2^d \sum_{k=0}^{d-m-1}\binom{d}{k}(2-p)^m p^k (1-p)^{d-m-k} + 2^d \sum_{k=d-m}^{l-1}\binom{d}{k}(1+p)^{k-d+m}(2-p)^{d-k}p^{d-m} + [i\binom{d}{m} - 2^d \sum_{k=0}^{l-1}\binom{d}{k}](1+p)^{l-d+m}(2-p)^{d-l}p^{d-m}$.

Note that $N_i^l \leq N_i^u$ and $M_i^l \leq M_i^u$, $i = 1, \cdots, s-1$ and equality holds when the CG-ratio $p = 0.5$.

A convenient notation is:

- The number of possible results for the first d-mer is $M_d = (2p + 2 - 2p)^d = 2^d$.
- Given i d-mers with no m mismatches, the number of possible d-mers that are m-mismatched with the given d-mers is greater than or equal to M_i^l and less than or equal to M_i^u.
- Given two different d-mers that are not m mismatched, the number of possible d-mers H that are m-mismatched with one of them is

  $$H \geq \begin{cases} M_1^l + \binom{d-1}{m} 2^m p^d, & m < d \\ M_1^l + (1+p)^{d-1}p, & m = d \end{cases}.$$

  Let the lower bound be $H^l = M_1^l + \binom{d-1}{m} 2^m p^d$ for $m < d$ and $H^l = M_1^l + (1+p)^{d-1}p$ for $m = d$.

Lower and upper bounds on the probability of no m mismatched d-mers in the subset S_1 are:

- If $M_d - M_{s-1}^u - N_{s-1}^u < 0$, then

  $$P[no\ m\ mismatches] = 0. \tag{1}$$

- If $M_d - M_{s-1}^u - N_{s-1}^u \geq 0$, then

  $$\prod_{i=1}^{s-1} \frac{(M_d - M_i^u - N_i^u)}{(M_d - N_i^l)} \leq P[no\ m\ mismatches] \leq \frac{\left(M_d - M_1^l\right)\prod_{i=2}^{s-1}\left(M_d - \max\left\{H^l, N_i^u\right\}\right)}{\prod_{i=1}^{s-1}\left(M_d - N_i^u\right)}. \tag{2}$$

The derivation of equations (1) and (2) is shown in the supplementary material.

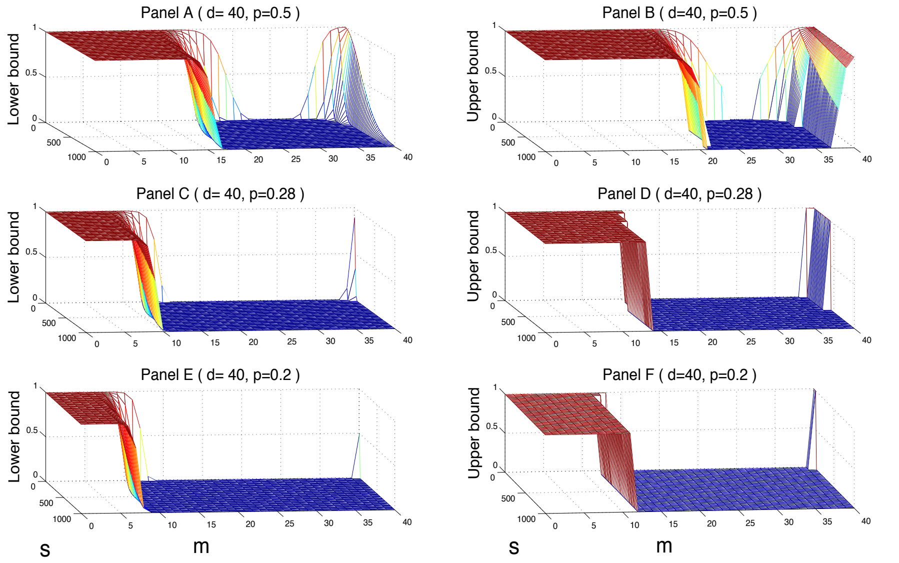

Figure 3: Bounds on the probability of no m mismatched $d = 40$-mers in subsets with different sizes: Extension of Figure 2 to a range of different values of s (the size of the subset). This continues to show large probability for small m (more so for small s), for both lower (Panel A, C and E) and upper (Panel B, D and F) bounds. The bound remain close, indicating good approximation quality, over a range of different CG-ratios, $p = 0.5$ (A and B), 0.28 (C and D) and 0.2 (E and F).

Discussion

Given the CG-ratio p, equations (1) and (2) give bounds on the probability of no m mismatched d-mers in the subset S_1 as functions of s and m. As an example we modeled the bounds for $s = 100$, the size of the subset S_1, in Figure 2. This showed that for probes (d-mers) with 50% ($p = 0.5$) CG- content even if we allow as many as 10 mismatches in a probe of length $d = 40$ the probability that any two probes in this set anneal to each other is 1 in 1,000,000, i.e. very unlikely.

The situation becomes less favorable as the CG- content ratio becomes more skewed. At 20% ($p = 0.2$) CG-content, such as experienced in certain microorganisms (mycoplasma has 24% CG-content) allowing for 8 mismatches yields a chance of in 1 in 1,000 that any two probes would anneal to each other. Because our model is symmetric around 50% CG-content the same reasoning applies to positively skewed ration such as found in *streptomyces* species, which average 72% CG-content. Hence, multiplex assays with > 1000 probes are limited to organisms with balanced CG- content.

Next, we explored how the bounds change when both subset size s and mismatch number m change. We use Figure 3 to illustrate the lower and upper bounds as functions of both the size of the subset s and the number of allowed mismatches m, for the CG-ratios $p = 0.5, 0.28, 0.2$. For practical applications, we want to maximize the size of the subset s, which increases the degree of multiplexing, and we want to minimize the number of allowed mismatches m, which increases specificity. As seen from Figure 3, a practical limit of the degree of multiplexing again depends on the CG-content of the target organism. Up to a set size of $s = 1000$ probes, though it is extremely unlikely that any two probes in a multiplex assay would bind to each other. This assumes large probes of length $d = 40$ or longer as used in the Nanostring$^{\mathrm{TM}}$ assay. The chance of unwanted cross-hybridization increases as the probe length d decreases. At a probe length of $d = 20$, such as used in multiplex PCR and applied to the worst case scenario of a microorganism with heavily skewed CG-content, allowing as little as 2 mismatches per probe may result in cross-hybridization between probes in the probe set. Luckily *homo sapiens* has a balanced CG-content, which allows the use of highly multiplexed assays for clinical applications.

Current solid-state microarrays can achieve a size of $s = 1.8 \times 10^6$ different probes per chip and $m = 1$, since they can detect *single nucleotide polymorphisms* (SNPs). Based on our calculation, we can answer the question: Can a solution-based, multiplex design reach or exceed this performance? For $d = 40$, $s = 1.8 \times 10^6$ and $m = 1$, we have the probability (of no m mismatches) within 10^{-9}, 10^{-3} and $2 * 10^{-2}$ of 1 for the CG-ratios $p = 0.5, 0.28$, or 0.2 respectively. Hence, solution-based SNP arrays based on probe sizes of 40 or longer have comparable performance to solid-state microarrays only for balanced (CG-ratio $p = 0.5$) probes. If the CG-ratio drops, as is known for many microbial genomes, solution-based SNP arrays underperform due to cross-hybridization among probes.

Mismatch and mismatch probability have a concrete biophysical meaning, see Cantor and Schimmel (1980). Every match lowers the free energy ΔG of the probe-target duplex and every mismatch m increases ΔG. Every probe-target duplex has a characteristic melting temperature T, which is a function of ΔG.

We show here that it is extremely unlikely that in a set S of size $s < 1000$ we would encounter any pair of probes of length $d = 40$ with $m < 13, 8$, or 6 (corresponding to the CG-ratio $p = 0.5, 0.28$ or 0.2) mismatches between them.

In sum, the current multiplex assays (e.g. nanostring$^{\mathrm{TM}}$, DASL$^{\mathrm{TM}}$) are expected to work and have a large margin of error built in before they encounter the theoretical boundaries, which we derived here.

As we move into higher and higher modes of multiplexing, it is important to know the principal boundaries of each design. As it is no longer possible to experimentally test all possible failure scenarios or experimentally validate the performance for each and every probe our theoretical understanding needs to improve to near certainty. Otherwise the true potential of highly multiplexed methods cannot be realized.

Acknowledgements. This work was partially supported by the Startup Fund of University of South Florida, and public health service grants CA019014 and AI107810 to DPD.

References

Bishop, M. A., D'yachkov, A. G., Macula, A. J., Renz, T. E. and Rykov, V. V. (2007) Free energy gap and statistical thermodynamic fidelity of dna codes. *Journal of Computational Biology 14*(8), 1088–1104.

Cantor, C. R. and Schimmel, P. R. (1980) *Biophysical chemistry: Part III: the behavior of biological macromolecules*. Macmillan.

D'yachkov, A. G., Vilenkin, P. A., Ismagilov, I. K., Sarbaev, R. S., Macula, A., Torney, D. and White, S. (2005) On dna codes. *Problems of Information Transmission 41*(4), 349–367.

Dyachkov, A. G. and Voronina, A. N. (2009) Dna codes for additive stem similarity. *Problems of Information Transmission 45*(2), 124–144.

Graham, R. L. (1995) *Handbook of combinatorics*, Volume 1. Elsevier.

Riordan, J. (2012) *Introduction to combinatorial analysis*. Courier Corporation.

White, A. K., VanInsberghe, M., Petriv, I., Hamidi, M., Sikorski, D., Marra, M. A., Piret, J., Aparicio, S. and Hansen, C. L. (2011) High-throughput microfluidic single-cell rt-qpcr. *Proceedings of the National Academy of Sciences 108*(34), 13999–14004.