illumina®

Array Differences in Genomic Coverage and Data Quality Impact GWAS Success

David Delano, Michael Eberle, Luana Galver, and Carsten Rosenow*

Introduction

Recent successes in genome-wide association studies (GWAS) are evidenced by a steady increase in the number of scientific publications in the literature. This is likely to be a direct reflection of the level of efficiency with which investigators are now able to associate clinical phenotypes with SNPs. These successes have come despite complex challenges. Factors influencing the likelihood of a successful outcome include phenotypic heterogeneity, phenotype prevalence, population stratification, the minor allele frequencies of the genetic elements driving phenotypic variance, and the availability of adequate sample quantities.

In this context, it is not surprising that the choice of genotyping platform draws significant attention and discussion. The selection strategy and quantity of markers, as well as the robustness of data generated by genotyping arrays, are known to affect study power and the likelihood of GWAS success. Illumina has designed its arrays with the intent of enabling investigators with a combination of high data quality and optimized genomic coverage. In many cases, this combination has provided the fastest path to significant associations and publication. This document demonstrates multiple measures of coverage and data quality and describes how each contributes to the performance of Illumina's genotyping platform.

Analysis Methods and Results

Genomic Coverage

The fundamental requirement that a genotyping array must fulfill in a genome-wide association study is the ability to interrogate the locus or loci associated with the trait of interest. The method of interrogation may be direct—by including the genetic variants responsible for the observed phenotypic differences—or indirect—by including SNPs that are in sufficient linkage disequilibrium with the causative variants to be used as surrogates for them[1].

While practical restrictions prevent the inclusion of every potential causative variant on a single array, the data from the International HapMap Project provide a means to successfully implement the indirect approach. The tag SNP strategy applied in selecting the content of Illumina's genotyping arrays was based on this approach[2]. The level of genomic coverage resulting from this strategy has been evaluated and described by publicly-funded studies[3–5]. Their findings demonstrate that the coverage Illumina's strategy produces is unparalleled in the industry.

The simplest and most commonly used measure to compare arrays is the global coverage that each provides. Global coverage is defined as the fraction of common SNPs (minor allele frequency ≥ 0.05) across the genome that are tagged by, or are in linkage disequilibrium with (typically with a threshold of $r^2 > 0.8$), the SNPs on the array. The relative levels of global coverage across the tested arrays determined by University of Michigan researchers, M. Li and colleagues, are shown in Figure 1. The Illumina Human1M BeadChip offers the highest global coverage across each population tested, with 93% of common SNPs in the CEU population tagged at $r^2 \geq 0.8$ (Figure 1). While less than the Human1M, the coverage of the lower density HumanHap550 and HumanHap650Y BeadChips still offer a level of coverage in the CEU population higher than any competing array. (The HumanHap550 BeadChip has been replaced by the Human610-Quad BeadChip, and the HumanHap650Y and Human1M BeadChips have been replaced by the Human1M-Duo BeadChip, with full backwards compatibility.)

Since the content on the arrays was based in part on the Phase II HapMap reference data set, it is a valid concern that coverage measures may be overestimated in the HapMap populations. This issue has been addressed by a number of studies and was found not to impact the validity of the previous conclusions[2,3]. Most recently, Bhangale et al. used resequencing data across multiple genes generated from both HapMap and non-HapMap CEU individuals to assess the coverage of arrays that had been previously calculated based on Phase II HapMap data. While the coverage levels calculated by this new method were reduced across all tested platforms relative to those determined based on the HapMap data, the Illumina Human1M and HumanHap650Y BeadChips maintained their markedly superior performance.

Efficiency

While the superior coverage delivered by the Illumina Human1M BeadChip was expected given the higher number of included tag SNPs, the coverage of the HumanHap650Y and HumanHap550 BeadChips was notably higher than that of competing arrays, despite including 29% and 41% fewer SNPs, respectively[5] (Figure 1). These data highlight the efficiency of coverage provided by the tag SNP strategy. This is a critical factor to study design because the Bonferroni multiple-testing correction factor increases linearly with the number of SNPs assayed (0.05 × number of SNPs). Using fewer SNPs and more efficient content can lead to increased association scores.

Gene Coverage

While global coverage assessments are promising indicators of an array's utility in a GWAS, the success of a study ultimately depends on whether the array includes the causative allele or a SNP that tags it. Since a substantial portion of causative variants are expected to lie within the coding and regulatory regions of genes, it is important to ensure that SNPs located within these regions are consistently tagged by an array. Figure 2 shows the number of genes covered by each array tested as a function of coverage percentage across three HapMap populations. Consistent with the results from the global coverage analysis, the Illumina Human1M BeadChip delivers greater genic coverage in all three HapMap populations than competing arrays. Furthermore,

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA

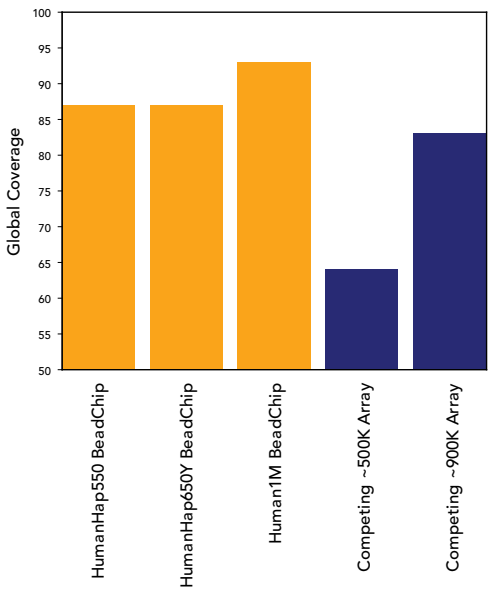
the HumanHap550 and HumanHap650Y BeadChips exhibit genic coverage higher than competing arrays in the CEU population. Thus, the tag SNP strategy provides an unparalleled advantage at the level of individual gene coverage, which directly impacts the likelihood of successfully identifying associations.

Call Rate and Coverage Gaps

The theoretical genomic coverage analyses described above are informative and useful for deciding which array to choose for a GWAS. However, the practical accuracy of these calculations depends on the assumption that each of the markers on the array consistently provides usable data. In practice, the actual number of calls depends on a balance set by the investigator between call rate (the total number of markers for which a genotype is generated) and the signal quality threshold used to prevent the inclusion of false calls. Assessing call rate as a function of included false calls is therefore critical to a meaningful comparison between arrays.

Call quality can be inferred indirectly by measuring calls that are not reproducible between technical replicates or can be measured directly by comparing calls to a reference data set, such as HapMap data or parental genotypes. For the analysis shown in Figure 3, technical replicates of whole-genome genotyping data across two sets of trios were generated by an independent service provider using both an Illumina Human1M-Duo BeadChip and a competing array composed of ~900,000 SNPs (two sets of trios for ~2 million and ~1.8 million total calls, respectively). Unreproducible calls were defined as cases where different genotype calls were made at a locus between technical replicates for an array. Figure 3 shows the call rate plotted as a function of the number of unreproducible calls that were included. The number of unreproducible calls that would need to be retained in order to reach any given level of genome coverage is substantially higher for the competing array than for the Illumina Human1M-Duo BeadChip.

Figure 1: Global Genomic Coverage Across Tested Array



Global coverage was calculated for the CEU population as the fraction of total SNPs that are tagged at $r^2 > 0.8$, out of the total number of SNPs included on the array (adapted from Table 1 of Li et al., 2008).

The starting level of coverage provided by an array's content is certainly critical. However, the final call rate after data cleaning, the resulting effective genomic coverage, and the number of false calls included must all be taken into consideration during the design and interpretation of a GWAS.

Miscalls and Data Interpretation

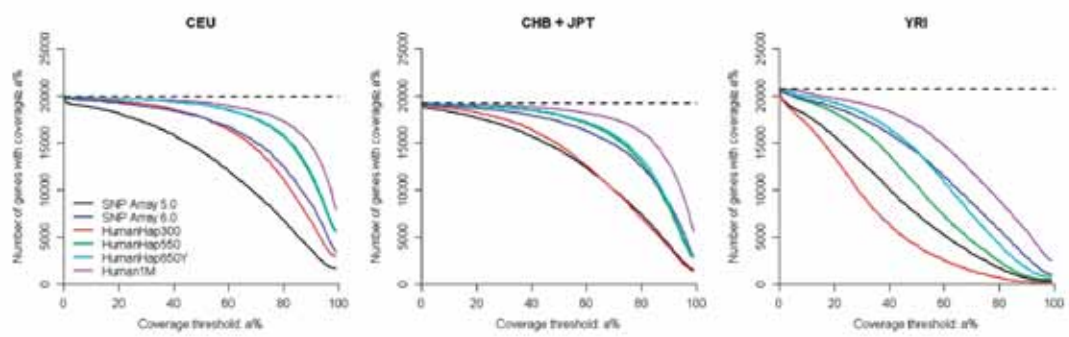
In addition to loss of coverage, discordant calls have the potential to negatively impact GWAS results in other ways. While achieving the highest possible call rate is desirable, including questionable calls at the expense of accuracy can quickly compromise the integrity of the results. The evidence of miscalls at even the highest levels of signal stringency (Figure 3) highlights the importance of a thorough examination of the reliability of data generated by an array platform before designing a study.

A widely used metric to assess genotype call quality is the transmission disequilibrium test (TDT). A TDT can be used to gauge false call rate based on the average frequency with which the minor alleles of all SNPs across a genome-wide array deviate from 0.5, which is expected under the null hypothesis. Data generated from a set of trios using the Illumina HumanHap550, Human610-Quad, and Human1M-Duo BeadChips are shown both directly as the frequency of minor allele transmission and as the chi-squared score that this proportion generates (Table 1). The degree of consistency (< 2.5% deviation) between the observed and expected minor allele transmission frequencies is a clear indication of the accuracy of the calls generated by Illumina BeadChips.

Another commonly used measure of genotype call quality is the frequency of observed inconsistencies of strict Mendelian inheritance. Different genotyping arrays have different rates and profiles of errors, which are important to examine closely when comparing arrays. Figure 5 depicts the quantities of various classes of Mendelian inconsistencies detected in the trio data set described in the call rate analysis. Interestingly, some classes of inconsistencies were overrepresented relative to the overall average. For example, in a substantial number of cases the mother and father were each called as homozygous for the opposite allele (AA × BB) while the son was also called as homozygous (Figure 5, asterisk). This class had the highest incidence for both platforms, though it occurred over three-fold times more often on the competing array than on the Illumina Human1M BeadChip. There were also a substantial number of incidences where the mother and father were each called as homozygotes for the same allele while the son was called either a heterozygote or a homozygote for the other allele (Figure 5, diamond). Across all categories of Mendelian inconsistencies detected, the observed error frequency was markedly higher for the competing ~900K SNP array.

Given the evidence that some false calls are likely retained in the final data set to maintain an adequate call rate (Figure 3), it is useful to quantify any negative repercussions that these calls could have on data interpretation. While loss of power due to dropped calls, limitations in genomic coverage, or other sources are always a concern with regard to type II errors (false negatives), the possibility of including

Figure 2: Number of Covered at Various Coverage Thresholds



Four Illumina BeadChips and two competing arrays were compared by Li and colleagues[5] in terms of the number of genes covered by SNPs on an array. Only gene regions containing ≥ 5 HapMap common SNPs were considered and coverage was evaluated at $r^2 > 0.8$.

Reprinted by permission from Macmillan Publishers Ltd: Eur J Hum Genet, 16:635-43, copyright 2008.

non-random miscalls capable of increasing the frequency of type I errors (false positives) is much more problematic. McCarthy and colleagues recently reviewed several challenges facing investigators running GWAS7. One issue addressed was the potential impact that miscalls included in the final data might have on the interpretation of the results. The authors state:

> "Experience from the first wave of GWA studies has demonstrated that scrupulous attention to detail is required throughout because each stage is fraught with the potential for error and bias. Many of these errors and biases have the potential to generate extreme values for the association test statistic; if uncorrected, these can dominate the tails of the distribution, such that interesting true associations become lost in a sea of spurious signals. Efforts to prevent, detect and eradicate sources of error therefore remain a high priority in GWAS studies, despite continuing improvements in genotyping performance."

It is quite possible that among the false calls detected in the analysis depicted in figure 5 are cases such as those described by McCarthy and colleagues as high-risk for false positive generation. This type of uncertainty in a large experiment requires additional data cleaning exercises as well as more downstream validation. Beyond the genomic coverage, required sample sizes, and call rate offered by a genotyping array, confidence in the accuracy of the results is invaluable to a successful GWAS.

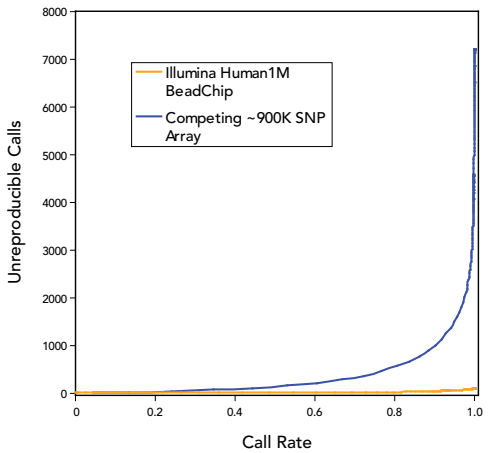
Conclusion

In designing a GWAS, numerous variables have the potential to impact the success or failure of the study. Many of these factors are determined by nature, and the ability to adequately account for them in the study design varies depending on the phenotype and study population. However, the degree of genomic coverage and the accuracy of genotype calls are largely determined by the genotyping array chosen for the study. Given the costs of a GWAS, careful consideration of the relative advantages and disadvantages offered by different platforms is an essential preliminary step in any study.

The data reviewed here illustrate that the design of Illumina's genotyping BeadChips provides clear, substantial advantages over competing arrays as gauged by the critical metrics of genomic coverage, genic coverage, call rate, signal quality, and call accuracy. These factors impact not only the quality and reliability of the data the platform generates, but also the time and resources required to clean and post-process. Ultimately, the speed to publication benefits from careful array choice.

The end result, more publications in less time when using Illumina BeadChips, is evidenced by in-the-field performance. Supplementary table 2 lists a sample of the rapidly growing number of studies (including their time to publication, GWAS odds ratio, and p-value) that substantiate the practical advantages of using the highest quality genotyping arrays[8].

Figure 3: Number of Included Miscalls as a Function of Call Rate



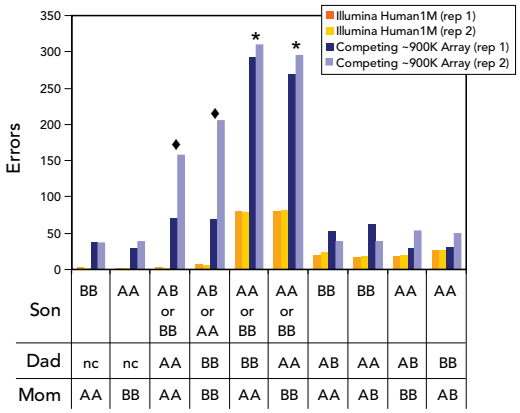
Two different arrays are compared in terms of the unreproducible calls generated at various call rates. Unreproducible calls were defined as cases where the genotype for a given SNP was differentially called between technical replicates on the same platform. Call rate was determined as a function of increasing quality score threshold and the resulting number of included unreproducible calls assessed.

AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAA
AATCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACC
AACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTA
TTAACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGAT
AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAA
AAGATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGA
AACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACA

Figure 1: Transmission Disequilibrum Tests Demonstrate the Accuracy of Illumina Beadchip Genotyping Calls

Array	McNemar x^2		Frequency Minor Allele Transmission	
	Actual	Expected (no errors)	Actual	Expected (no errors)
HumanHap550	314	0–130	0.51	0.5
Human610-Quad	320	0–200	0.512	0.5
Human1M-Duo	503	0–400	0.501	0.5

Figure 5: Mendelian Inconsistencies Detected Across Tri Data



The distribution of Mendelian inconsistencies identified within technical replicates across a set of three triads measured for each of two technical replicates on two array platforms. The number of occurrences for each category and the corresponding calls across mother, father, and son are shown. nc = no call

References

The data used in the analysis depicted in Figure 5 are available from http://www.illumina.com/GWASArray.

The Li et al. article is available for free download from http://www.illumina.com/GWASArray.

1. Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L et al. (2003) Additional SNPs and Linkage-Disequilibrium Analyses Are Necessary for Whole-Genome Association Studies in Humans. Nat Genet 33: 518-521.
2. Eberle MA, Ng PC, Kuhn K, Zhou L, Peiffer DA et al. (2007) Power to Detect Risk Alleles Using Genome-Wide Tag SNP Panels. PLoS Genet 3: 1827-1837.
3. Bhangale TR, Rieder MJ, Nickerson DA (2008) Estimating Coverage and Power for Genetic Association Studies Using Near-Complete Variation Data. Nat Genet 40: 841-843.
4. Li C, Li M, Long JR, Cai Q, Zheng W (2008) Evaluating Cost Efficiency of SNP Chips in Genome-Wide Association Studies. Genet Epidemiol 32: 387-395.
5. Li M, Li C, Guan W (2008) Evaluation of Coverage Variation of SNP Chips for Genome-Wide Association Studies. Eur J Hum Genet 16: 635-643.
6. Schadt EE, Molony C, Chudin E, Hao K, Yang X et al. (2008) Mapping the Genetic Architecture of Gene Expression in Human Liver. PLoS Biol 6: e107.
7. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J et al. (2008) Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges. Nat Rev Genet 9: 356-369.
8. http://www.illumina.com/downloads/GWASArrayWhitePaperSuppT2.pdf

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2010 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, GoldenGate Indexing, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, CSPro, GenomeStudio, Genetic Energy, HiSeq, and HiScan are registered trademarks or trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners. Pub. No. 370-2008-018 Current as of 28 July 2010

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA