Ranking of Genes, SNVs, and Sequence Regions

Ranking elements within various types of biosets for metaanalysis of genetic data.

Introduction

One of the primary applications of the BaseSpace® Correlation Engine is to allow researchers to perform metaanalyses that harness large amounts of genomic, epigenetic, proteomic, and assay data. Such analyses look for potentially novel and interesting results that cannot necessarily be seen by looking at a single existing experiment. These results may be interesting in themselves (eg, associations between different treatment factors, or between a treatment and an existing known pathway or protein family), or they may be used to guide further research and experimentation.

The primary entity in these analyses is the **bioset**. It is a ranked list of elements (genes, probes, proteins, compounds, single-nucleotide variants [SNVs], sequence regions, etc.) that corresponds to a given treatment or condition in an experiment, an assay, or a single patient sample (eg, mutations). For a gene expression experiment, the biosets will consist of gene lists with associated change values and statistical information for each relevant experimental factor. For example, a bioset could consist of a list of Affymetrix probesets, fold-change difference values between treated and control groups, and p-values produced by a t test between treated and control sample probeset intensity values. A sequence-centric bioset from a copy-number analysis experiment, for example, will contain ranked sequence regions with associated gain and loss statistics.

In summary, biosets may contain many of the following columns:

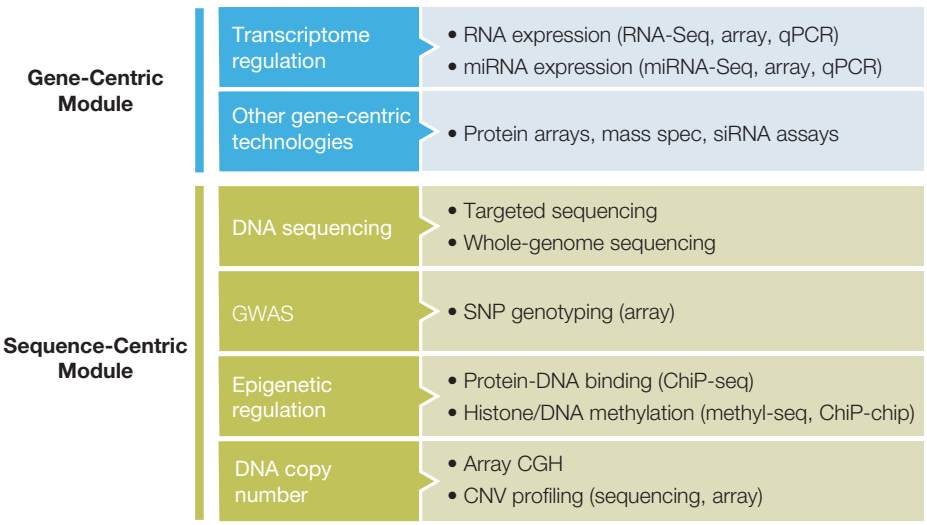
- Identifier of an entity such as a gene, SNV, or sequence region (required)
- Other identifiers of the entity—eg, chromosome, position
- Summary statistics—eg, p-value, fold change, score, rank, odds ratio

Ranking of Elements within a Bioset

Most biosets generated at Illumina include elements that are changing relative to a reference genome (mutations) or due to a treatment (or some other test factor) along with a corresponding **rank** and **directionality**. Typically, the rank will be based on the magnitude of change (eg, fold change); however, other values, including p-values, can be used for this ranking. Directionality is determined from the sign of the statistic: eg, up (+) or down(-) regulation or copy-number gain (+) or loss (-). Some biosets contain signatures from individual samples where the elements may or may not be ranked.

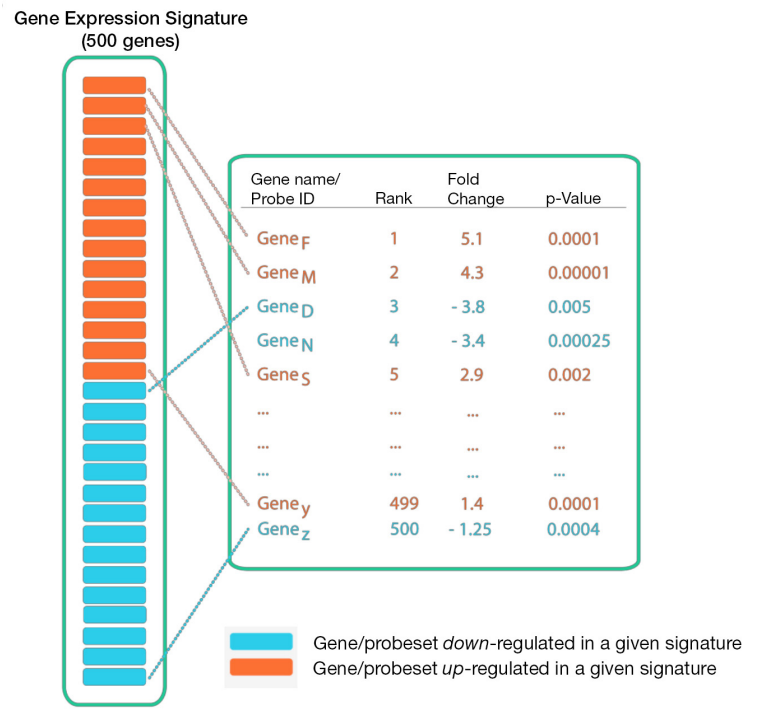
Where applicable, elements within biosets that are generated by Illumina scientists internally are ranked and assigned directionality using predetermined criteria. However, for user-imported biosets, the statistics/values to be used for ranking can be specified during upload.

Figure 1: Overview of Biosets



Biosets within the BaseSpace Correlation Engine comprise gene-centric or sequence-centric signatures from orthogonal types of biological experiments.

Figure 2: Gene Expression Bioset



An example of a gene expression signature, with associated statistics.

The BaseSpace Engine metaanalysis procedures use rank-based comparison of different types of biosets to other biosets, as well as biosets to biogroups (pathways, Gene Ontology [GO] functional groups, sets of genes regulated through a common regulatory motif, and similar biological sets). By looking across different types of data (eg, gene expression, epigenetic regulation, DNA copy-number, methylation, protein abundance, or RNAi targets with assay-associated scores) the ranking system enables comparisons that are independent of the underlying absolute values.

Therefore, in addition to assigning ranks for each element within a bioset, the BaseSpace Engine computes a **normalized rank**, which is essentially a function of the rank and other bioset features (such as data type or platform size). These normalized ranks aim to improve entity comparability across studies, platforms, and even data types.

This technical note describes the BaseSpace Correlation Engine approach to ranking elements within various types of biosets that are generated internally and uploaded into the BaseSpace Knowledge Network. The normalized rank function is also described.

Gene Sets or Biogroups

A biogroup is a collection of genes associated by a specific biological function, pathway, or similar criteria. No numerical information is directly associated with a biogroup. Examples of canonical gene lists represented as biogroups in the BaseSpace Engine include those from the GO Consortium[1] and the Broad Institute's MSigDB.[2]

Ranking Criteria

Genes within a biogroup are unranked, ie, no rank, or normalized rank is assigned.

Directionality

Genes are assigned no direction.

Gene Expression Biosets

These biosets generated from the analysis of microarray data contain gene signatures measuring differential expression between 2 conditions. Typically, this is a list of genes (or probesets), along with associated p-value, fold change, and average expression level in each group (Figure 2).

Ranking Criteria

Probesets within biosets generated internally by Illumina are ranked by magnitude of fold change.[3] When a gene is queried, the probeset with the highest normalized rank in a bioset is used to determine the order or relevance of search results.

Normalized rank is computed based on the rank of the gene in relation to the platform size (Equation 1). For example, a gene that is ranked #1 from a platform comprising 100 genes is considered less significant than a gene that is ranked #1 from a platform comprising 10,000 genes.

$$NormalizedRank = \frac{Rank}{PlatformSize} \times NormalizingConstant \quad (1)$$

The normalizing constant is set to be 22,215 (the size of the Affymetrix Human U133A platform) × 100 (to maintain precision at integer level) = 2,221,500.

When biosets are uploaded without any ranking statistics, all genes are assigned the same rank based on the size of the bioset—ie, number of genes in a bioset ÷ 2. The normalized rank is determined as above using the size of the custom species-specific platform in Equation 1. A species-specific custom platform contains all the known genes from the genome of that species.

Directionality

Directionality of a gene is determined by the sign of the fold change: positive for upregulated genes and negative for downregulated genes.

RNA-Seq Experiments

RNA-Seq experiments represent a newly evolving data type and there are many different ways to approach data analysis. The current BaseSpace Correlation Engine method will be changing as new requirements for handling RNA-Seq data emerge.

Ranking Criteria

Generally, RNA-Seq data enables researchers to generate higher-resolution gene expression readouts compared to traditional microarray experiments. The BaseSpace Engine will enable capture and ranking of RNA-Seq data at the following resolution:

1. Genes
2. Transcripts (eg, RefSeq)
3. Individual exons

Depending on the specific bioset, information from RNA-Seq experiments can contain the above features with associated fold changes, representing differential expression levels. Ranking criteria are then applied in a manner similar to traditional microarray-based experiments. In place of probesets, however, an RNA-Seq bioset may contain exons (with exon-specific statistics), individual transcripts, or genes.

Directionality

Directionality of a gene, transcript, or exon is determined by the sign of the fold change: positive for upregulated genes and negative for downregulated genes.

Body Atlas Biosets

In upcoming releases of the BaseSpace Correlation Engine, Illumina will be generating biosets that will enable researchers to investigate absolute and relative gene expression levels across multiple tissues, cell types, and cell lines. Gene expression experiments from predetermined standard platforms for each type—normal tissues, cell types, and cell lines—are collated and collectively normalized.[4,5] An investigation of batch effects is conducted and, if present, techniques for correction are applied.[6]

Ranking Criteria

In the context of absolute gene expression, genes are ranked by mean tissue, cell type, or cell line signal. In relative gene expression biosets, the mean signal of a gene is compared to a "reference," defined as the median signal among all tissues, cell types, or cell lines. A ranking statistic is computed by taking into account the difference in signal compared to the reference, and the standard deviation.

Directionality

In the context of relative gene expression, the directionality of a gene or transcript is determined by the sign of the fold change: positive for upregulated genes and negative for downregulated genes, relative to the median signal among all tissues for that gene.

SNV Biosets from GWAS and Aggregated Experiments

As seen in Figure 3, SNV biosets from aggregated studies—such as genome-wide association studies (GWAS)—contain significant markers whose genotype is associated with the phenotype under investigation. They also contain accompanying metrics, such as minor allele frequencies and various association test statistics. SNV biosets from individual samples essentially contain the genotype of the individual.

Illumina investigated the distribution of ranks and normalized ranks for gene expression biosets with varying sizes, and from different platforms and study designs. This study aimed to determine a unified ranking scheme for SNVs from genotyping studies, so that they can be compared across SNV studies and also be comparable with gene expression studies. A mapping function based on the sigmoid curve was derived to convert p-values to ranks value with the following aims in mind:

1. The range of ranks for SNV sets is the same as that for gene expression biosets.
2. The relative importance of SNVs within a bioset, as determined by p-value, is preserved by using the sigmoid function. As p-values increase, the SNVs quickly become less significant and saturate for large p-values.
3. Good resolution of rank values will differentiate between SNVs with close p-values.
4. SNVs with the same p-value in different SNV biosets will have the same rank and normalized rank.

Ranking Criteria

For biosets containing SNVs with associated p-values

$$NormalizedRank(p) = round \left(\frac{1}{1 + base^{(-scale \times p)}} \times 2 - 1 \right) \times range \quad (2)$$

$$Rank(p) = \frac{NormalizedRank(p) \times platSize}{range} \quad (3)$$

Figure 3: SNV Bioset from GWAS

An example of an SNV set comprising significant markers from a GWAS, with associated statistics.

Parameter	Default Value
base	10
scale	100
range	2221500 *(for mapping to range of gene expression normalized ranks)*
platSize	36437 *(size of custom human platform)*

Tuning parameters, such as **base** and **scale**, were adjusted so that normalized ranks from SNV sets had similar range of values compared with gene expression biosets of similar sizes. This adjustment also ensured fine resolution of ranks to distinguish one SNV from another for highly significant SNVs.

For biosets containing SNVs without associated p-values or any other ranking columns

Typical examples of these biosets are mutation data from resequencing studies, but users may also upload GWAS data without any associated statistics. In these cases, all rows in the bioset are assigned the same rank based on the size of the bioset as follows:

1. Compute a pseudo p-value p' = totalBasesInBioset ÷ totalBasesInGenome
2. Use Equations 2 and 3 to compute normalized rank and rank with p' in place of p.

For biosets containing SNVs where only some SNVs have associated p-values

Ranks for rows with p-values are assigned as for SNV biosets with associated p-values. The highest p-value of these rows is recorded (*maxP*).

For the remaining rows without p-values:

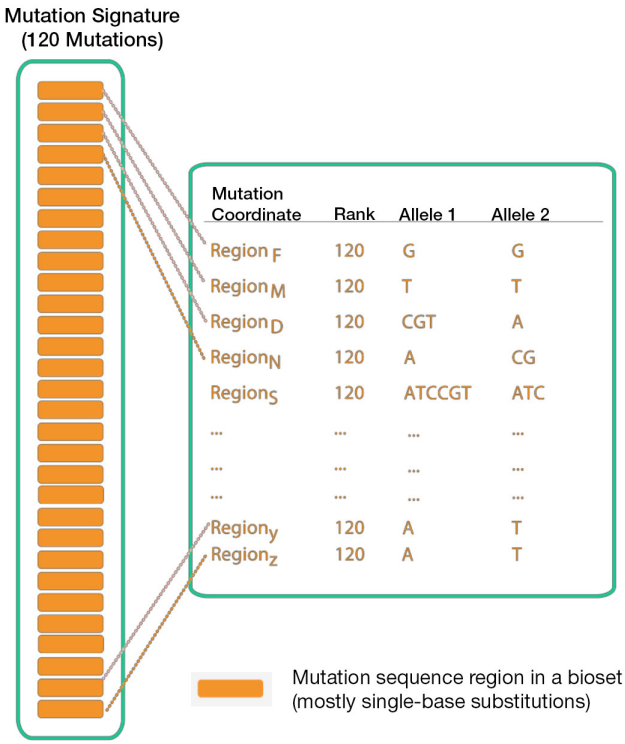
1. Compute a pseudo p-value p'= totalBasesInBioset ÷ totalBasesInGenome
2. Compute the new pseudo p-value p'' = max(p', *maxp*)
3. Use Equations 2 and 3 to compute normalized rank and rank with p'' in place of p.

Parameter	Value
totalBasesInBioset	Sum of bases of all rows
totalBasesInGenome	Species-specific genome size in bp

Directionality

Elements in an SNV bioset are not assigned a direction.

Figure 4: SNV Bioset from Mutation Data



An example of an SNV bioset comprising mutation data from an individual sample, with associated statistics.

Mapping

SNVs are mapped to genes and a gene-centric bioset is also created corresponding to each SNV bioset. Genes carry the same rank and normalized rank as the SNVs to which they map.

Illumina is investigating weighted ranking of SNVs by mutation significance, eg, location of SNV (promoter region, intergenic), and impact on protein sequence (nonsense mutation, missense mutation vs. silent mutation).

SNV Biosets from Mutation and Resequencing Data

Mutation and resequencing biosets from individual samples typically contain the mutation identifier and location, and the alleles. The ranking scheme followed is the same as that described for GWAS data. The typical case is when SNVs have no associated p-values or other ranking columns.

Directionality

Elements in an SNV bioset are not assigned a direction.

Biosets from Experiments Investigating Copy Numbers

Figure 5 shows a bioset generated from an analysis of copy numbers in a cell line.

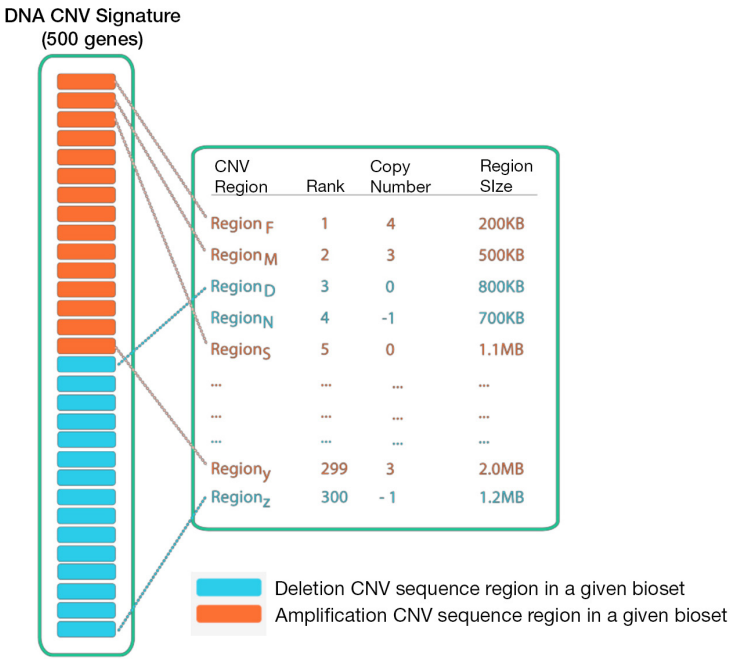
Ranking Criteria

Regions are sorted according to the ranking column (eg, z score) and “raw ranks” are assigned sequentially to each region (Figure 5).

Next, raw ranks are converted to ranks and are assigned to each sequence based on cumulative sequence length and unit region size as follows:

1. For each sequence region, set cumulatedLength to be the sum of sequence lengths of all the regions with better or the same raw rank.
2. Set rank of the region to cumulatedLength ÷ unitRegion, rounded off to the closest integer.
3. Normalized rank is computed as in Equation 4, using the custom platform size and the normalizing constant.

Figure 5: Copy-Number Bioset



An example of a bioset generated from copy-number data, with associated statistics.

Certain constraints are imposed:

1. Smallest rank is 1 (ranks = 0 are raised to 1).
2. When a region is so small that it gets the same rank as the sequences at the previous raw rank, a minimal increment (+1) is added to differentiate them.

Parameter	Value
totalRanks	Species-specific custom platform size
genomeSize	Species-specific genome length
unitRegion	genomeSize/totalRanks

Directionality

Directionality is set to positive for copy-number gain and negative for loss.

Mapping

Sequence regions are mapped to genes, and a gene-centric bioset is created corresponding to each sequence-centric bioset. Genes carry the same rank and normalized rank as the sequence regions to which they map.

Biosets from Epigenetic Data

ChIP-Chip Experiments

Ranking Criteria

ChIP-enriched sequence regions are ranked by a score metric that is the sum of the smoothed reporter levels in the region, each minus the enrichment threshold.[7] Sequence regions are mapped to genes, through the creation of a gene-centric bioset, and the ranks are carried over.

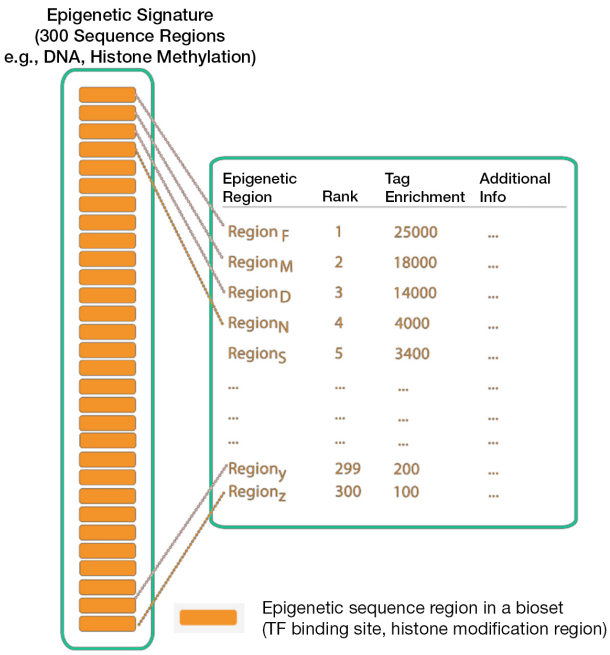
Normalized ranks for these genes are computed in the same way as for gene expression studies (Equation 4). The background platform of the ChIP-chip gene-centric bioset is a custom platform that contains all known genes from the genome of that species. The normalizing constant is set to 2,221,500.

$$NormalizedRank = \frac{Rank}{SpeciesSpecificCustomPlatformSize} \times NormalizingConstant \quad (4)$$

Directionality

Elements in a ChIP-Chip bioset are not assigned a direction.

Figure 6: Epigenetic Data Bioset



An example of an epigenetic signature, with associated statistics.

ChIP-Seq Experiments

Ranking Criteria

ChIP-enriched sequence regions are ranked by the ChIP signal determined by the customized version of QuEST program.[8] Sequence regions are mapped to genes, through the creation of a gene-centric bioset, and the ranks are carried over.

Normalized ranks for these genes are computed as in Equation 4.

Directionality

Elements in a ChIP-Seq bioset are not assigned a direction.

Infinium® Methylation Experiments

Ranking Criteria

Methylated regions are ranked by the absolute value of differential percentage methylation between the test group and the reference group.

Directionality

Directionality is set to positive for positive differential percentage methylation and negative for negative differential percentage methylation.

Mapping

Sequence regions are mapped to genes, and a gene-centric bioset is created corresponding to each sequence-centric bioset. Genes carry the same rank and normalized rank as the sequence regions to which they map.

References

1. Huntley RP, Harris MA, Alam-Faruque Y, Blake JA, Carbon S, et al. (2014) A method for increasing expressivity of Gene Ontology annotations using a compositional approach. BMC Bioinformatics 15: 155.
2. www.broadinstitute.org/gsea/msigdb/index.jsp
3. Shi L, Tong W, Fang H, Scherf U, Han J, et al. (2005) Cross-platform comparability of microarray technology: intraplatform consistency and appropriate data analysis procedures are essential. BMC Bioinformatics 6(Suppl 2): S12.
4. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4(2): 249-264.
5. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics 19(2): 185–193.
6. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8(1): 118–127.
7. Toedling J, Skylar O, Krueger T, Fischer JJ, Sperling S, et al. (2007) Ringo—an R/Bioconductor package for analyzing ChIP-chip readouts. BMC Bioinformatics 8: 221.
8. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, et al. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5(9): 829–834.

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

For Research Use Only. Not for use in diagnostic procedures.

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA