Computational methods for Bioinformatics

ASN file format

Compress DNA string in bits, and then convert the bits into hex.

DNA	Code
A	00
C	01
G	10
T	11

→ → →

Binary	Hexadecimal	Binary	Hexadecimal
0000	0	1000	8
0001	1	1001	9
0010	2	1010	A
0011	3	1011	B
0100	4	1100	C
0101	5	1101	D
0110	6	1110	E
0111	7	1111	F

Code 18.14 The **DecoderDict** function.

```
# create a decoding dictionary
def DecoderDict( ):
    ddct = {}
    ddct['0'] = 'AA'; ddct['1'] = 'AC'; ddct['2'] = 'AG'
    ddct['3'] = 'AT'; ddct['4'] = 'CA'; ddct['5'] = 'CC'
    ddct['6'] = 'CG'; ddct['7'] = 'CT'; ddct['8'] = 'GA'
    ddct['9'] = 'GC'; ddct['A'] = 'GG'; ddct['B'] = 'GT'
    ddct['C'] = 'TA'; ddct['D'] = 'TC'; ddct['E'] = 'TG'
    ddct['F'] = 'TT'
    return ddct
```

Code 18.15 The **DNAFromASN1** function.

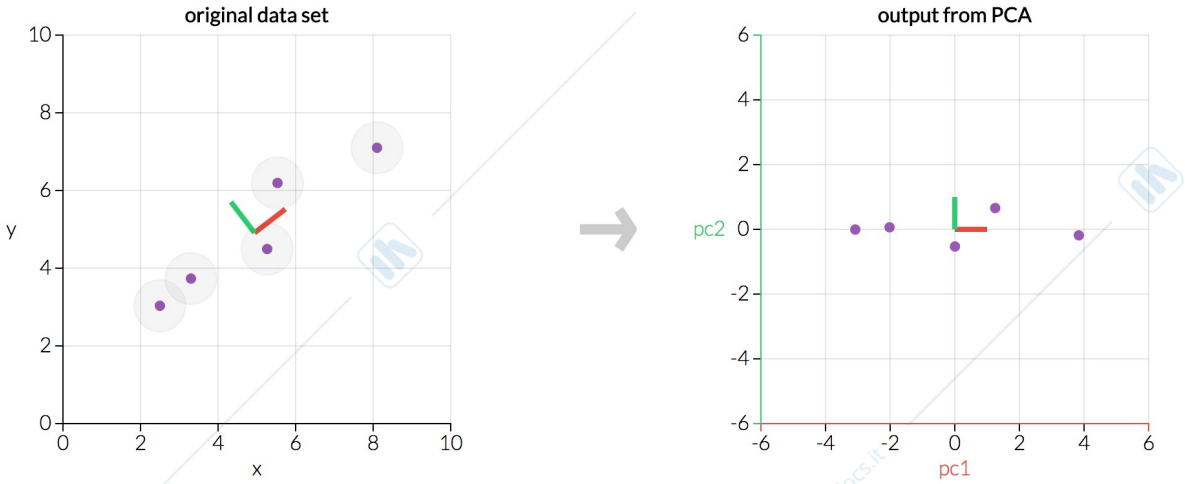
```
def DNAFromASN1( filename, ddct ):
    # read in data
    fp = file( filename )
    a = fp.read()
    fp.close()
    # extract DNA
    loc = a.find( 'ncbi2na' )
    start = a.find( "'",loc )+1
    end = a.find( "'", start+2)
    cpdna = a[start:end] # compressed dna
    cpdna = cpdna.replace( '\n','')
    # decode
    dna = ''
    for i in range( len( cpdna )):
        dna += ddct[ cpdna[i] ]
    return dna

>>> dna = DNAFromASN1( 'c20/nc_006046.asn1', ddct)
>>> dna[:100]
'CCTCTCCTCTCGCGCCGCCAGTGTGCTGGTTAGTATTTCCCCAAACTTTCTTCGAAT
GATACAACAATCACACATGACGTCTACATAGGAGCCCCGGAAG'
```

DNA FILE FROMAT

DNA information is stored in several formats. Two of the most popular are FASTA and Genbank. The FASTA files are very easy to read and this takes only a few lines of code. The Genbank files are considerably more involved and store significantly more information beyond the DNA sequence. They can store identifying information, publication and author information, proteins, identified repeats and much more. Thus, reading these files requires a bit more programming. These programs, however, are not complicated.

PCA Principal Component Analysis

Data generated from experiments may contain several dimensions and be quite complicated. However, the dimensionality of the data may far exceed the complexity of the data. A reduction in dimensionality often allows simpler algorithms to effectively analyze the data. The most common method of data reduction in bioinformatics is principal component analys.



Quite often a single biological experiment can produce a lot of data, but due to time and costs, only a small number of experiments can be run. So their are few data vectors that have a lot of elements. The dimensionality of the data is large, but the dimensionality of the information is not. So, PCA is a very useful tool that reduces the dimensionality of the data without damaging the dimensionality of the information.

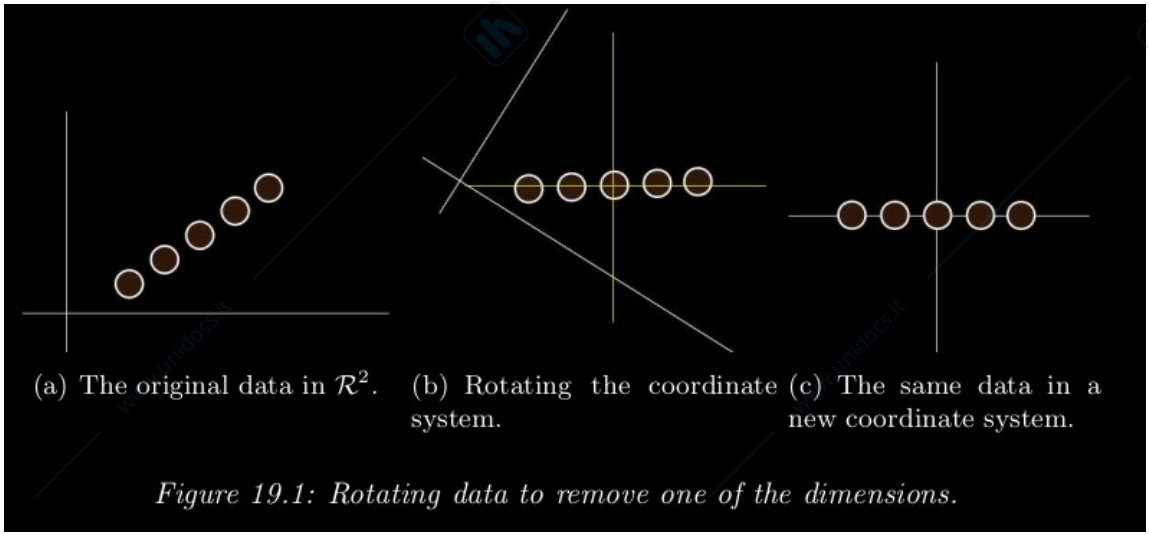


Figure 19.1: Rotating data to remove one of the dimensions.

The logic of PCA (principal component analysis) is to diagonalize the covariance matrix. In doing so, the elements of the data become independent. If there are first order relationships within the data then this new representation will often display these relationships more clearly than the original representation. Diagonalization of the covariance matrix is achieved through mapping the data through a new coordinate system.

The protocol for PCA is,

1. Compute the covariance of the data.
2. Compute the eigenvectors and eigenvalues of the covariance matrix.
3. Determine which eigenvectors to keep.
4. Project the data points into the new space defined by the eigenvectors.

The choice of how many eigenvectors to keep is up to the user and that is based on how sharply the curve bends and how much error the user can allow.

Covariance Matrix

PCA minimizes the covariance within a data set, and this information is contained within a covariance matrix. The purpose of the covariance is that it relates one column to another. Basically, if the data in two columns are positively correlated (when one goes up in value so does the other) then the covariance is positive. If the data in the two columns are negatively correlated then the covariance is negative. If the data in the two columns are independent then the covariance should be zero.

Variance-Covariance Matrix

	1	2	3	4	5
1	σ^2	COV	COV	COV	COV
2	COV	σ^2	COV	COV	COV
3	COV	COV	σ^2	COV	COV
4	COV	COV	COV	σ^2	COV
5	COV	COV	COV	COV	σ^2

Codon Frequencies

There are 64 different codons but only 20 different amino acids which means that many amino acids have multiple associated codons. It is therefore possible that some genomes favor one codon over another in the DNA when producing a gene. If this is true then it is possible to classify genomes according to their codon frequencies. To compute the codon frequencies the number of occurrences of each codon is obtained and these counts are divided by the total number of codons.

Code 20.2 The **CountCodons** function.

```
# codonfreq.py
def CountCodons( dna, codons ):
    N = len( dna )
    counts = np.zeros( 64 )
    for i in range(0, N, 3 ):
        cut = dna[i:i+3]
        if cut in codons:
            ndx = codons.index(cut)
            counts[ndx] += 1
    return counts

>>> cts = cf.CountCodons(dna, codons )
```

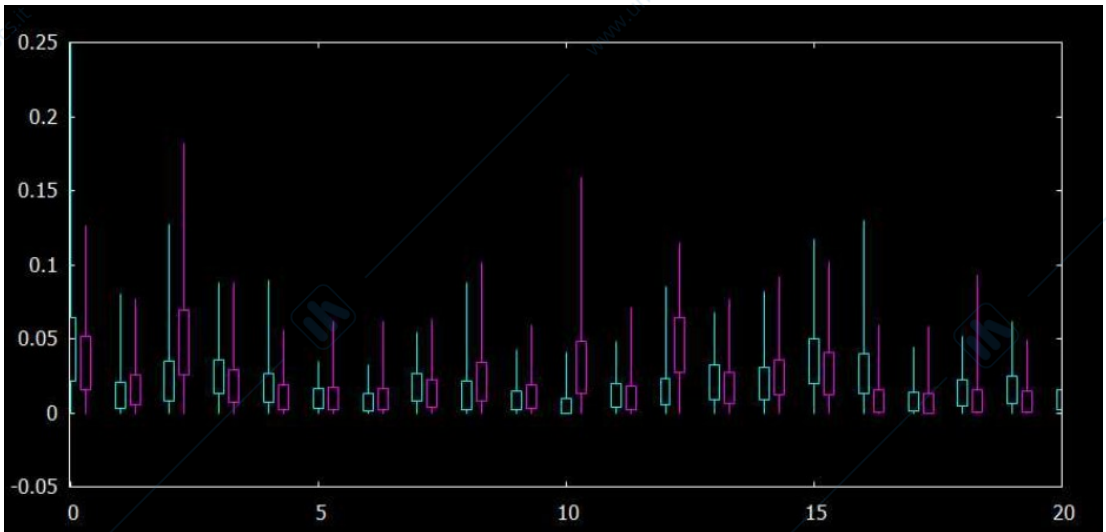


Figure 20.2: The statistics for the first 20 codons for two genomes.

20.3 Comparing Multiple Genomes

The plot in Figure 20.2 shows only a part of the comparison of just two genomes. Comparing multiple genomes requires a different analysis technique. For this task, PCA will be used as proposed by Kanaya *et al.* Kanaya et al., 2001.

The protocol for this experiment is:

1. Gather the names of several bacterial genomes.
2. Compute the codon frequency lists for each genome.
3. Apply PCA to this collection of data.
4. Color code the data points in PCA space for each genome.

The result is shown in Figure 20.3. Each genome is assigned a different color. The data was sufficiently organized in the PCA representation that each genome has its own isolated territory. This indicates that codon frequencies are sufficient for classifying bacterial genomes.

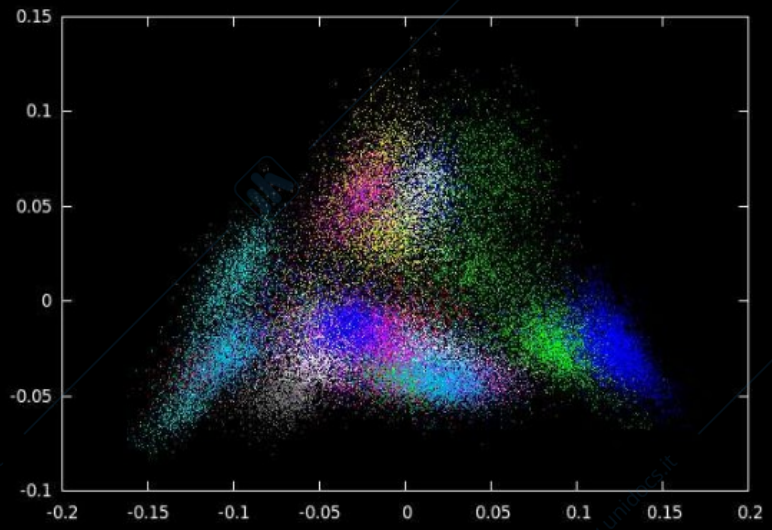


Figure 20.3: PCA mapping for several bacterial genomes.

Some genomes do overlap in this view. However, this only the first two PCA axis and it is always possible that the groups that appear to overlap in this view are not actually overlapping which can be seen in other views.

DNA sequences are complicated structures that have been difficult to decode. A strand of DNA contains coding regions which produce genes and contains non-coding regions which may or may not have functionality. As systems evolve genes were passed on sometimes with small alterations or relocations. Since the non-coding regions are less important in many respects they were often passed on with more alterations. These similarities allow us to infer functionality of a gene by relating it to other genes with known function. The main computational technique for accomplishing this comparison is to align sequences. The purpose of alignment is to demonstrate the similarity of two (or more) sequences. At first this sounds like an easy job. Each sequence has only four bases and it should not be too hard to determine if the sequences are similar.

Indels

This is a serious matter. If the sequences are very long (perhaps thousands of bases) then there are thousands of locations where the indel can occur. Furthermore, the sequences may have several indels and at one location multiple indels may need to be considered. For sequences of significant length it is not possible to consider all possible indels in a brute force computing fashion.

Starting and ending sequence

Given two sequences and the task of global alignment it is still necessary to be concerned with the beginning and ending of the sequences. The sequencing technology tends to have problems calling the very beginning and very end of sequences. Thus, the actual sequence may be longer than necessary.

Sequence length

Another complicating factor is sequence length. Often the alignment algorithms are based on the number of matches. Consider a case in which the sequences are 100 elements long and 90 of them align. Consider a second case in which the sequences are 1000 elements long and 800 of them align. In the second case the score can be higher since many more elements aligned, but the percentage of alignment is greater in the first case. So, some algorithms consider the sequence length when producing an alignment score. Since the strings could be different lengths it is necessary to find the length of the shortest string.

Simple Aligment

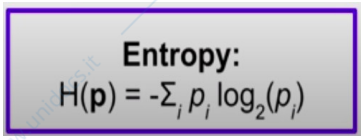
```
RNDKPKFSTARN
||-|||--||-|
RNQKPKWWTATN

++-+++--++-+
```

Table 21.1: The BLOSUM50 matrix.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	-2
R	-2	7	-1	-2	-1	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-1
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	-1	0	-4	-2	-2
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-3
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-3
Q	-1	-1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-1
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-2
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-3
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	-1	-1
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	-1
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	-1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-2
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	0
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	4
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	-2
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	2
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	8
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	-1

BLOUSUM50
Is a matrix for protein sequences alignment. It consider how often an amminoacid change into another. Is also complicated by the fact that some amino acids are more commonly seen than others. There are several versions depending on the mathematical methods employed and the evolutionary time step. The most popular matrices are either PAM or BLOSUM and the number that follows the name indicates the evolutionary time step. The matrices have some differences but are similar enough that only one matrix will be used here.

Entropy of Information



Nella teoria dell'informazione l'entropia di una sorgente di messaggi è l'informazione media contenuta in ogni messaggio emesso.
Negli anni quaranta Shannon riuscì a definire l'equazione con cui calcolare il livello di imprevedibilità di una fonte d'informazione, constatando che la sua formula era praticamente uguale a quella con cui Boltzmann aveva calcolato l'entropia di un sistema termodinamico. Per Shannon il problema era misurare quanta "informazione" contiene un dato messaggio, e quindi, di conseguenza, quanto costa inviarlo, dato un sistema di trasmissione e le difficoltà che un canale di trasmissione (in generale disturbato da "rumore") può trovare. L'entropia così introdotta nella teoria dell'informazione è quindi misura del grado di complessità di un messaggio emesso da una sorgente: come dimostrato nel primo teorema di Shannon, essa fornisce il numero medio minimo di simboli necessari alla codifica del messaggio stesso. In questo contesto l'entropia misura dunque la quantità d'incertezza o informazione presente in un segnale aleatorio, che può essere interpretata anche come la minima complessità descrittiva di una variabile aleatoria, ovvero il limite inferiore della compressione dei dati. L'informazione contenuta in un messaggio è tanto più grande quanto meno probabile era. Un messaggio scontato, che ha un'alta probabilità di essere emesso dalla sorgente contiene poca informazione, mentre un messaggio inaspettato, poco probabile contiene una grande quantità di informazione. L'entropia di una sorgente risponde a domande come: qual è il numero minimo di bit che servono per memorizzare in media un messaggio della sorgente? Quanto sono prevedibili i messaggi emessi dalla sorgente?
L'entropia può essere vista come la casualità contenuta in una stringa, ed è strettamente collegata al numero minimo di bit necessari per rappresentarla senza errori.

Protein Production difficulties due to formation of inclusion body

Protein production at low temperatures not only suppresses the functional activity of the proteins (enzymes), but also prevents the proteins to become insoluble (prevents the formation of inclusion body). Proteins become insoluble regardless of its type, but several factors can cause proteins to become insoluble.

1. High concentration;
concentration of proteins become too high at some localities within the cell.

2. Differences in intracellular reduction-oxidation (redox) environment.

3. Lack of protein modification after translation.

4. Inappropriate interactions with molecular chaperones and/or foldase.

5. Nonspecific disulfide binding.

The first factor described above can be avoided by producing proteins at low temperatures, as production rate of proteins slows down at low temperatures. Use of Rhodococcus is expected to affect factors 1 and 2. Codon usage frequency and the intracellular environment in Rhodococcus change the product

RNA-SEQ

With RNA-Seq, researchers can detect the fine architecture of the transcriptome, such as transcript isoforms, gene fusions, single nucleotide variants, and other features—without the limitation of prior knowledge. Because RNA-Seq does not require predesigned probes, the data sets are unbiased. As a result, RNA-Seq has been shown to detect a higher percentage of differentially expressed genes compared to expression arrays, especially genes with low abundance.

Table 1: Comparison of RNA-Seq technology with expression microarrays

Application	RNA-Seq	Microarray
High run-to-run reproducibility	Yes	Yes
Dynamic range comparable to actual transcript abundances within cells	Yes	No
Able to detect alternative splice sites and novel isoforms	Yes	No
De novo analysis of samples without a reference genome	Yes	No
Re-analyzable data	Yes	No

To understand normal cell development and disease mechanisms, researchers frequently investigate differential expression during development, in specific tissues, or in response to varying conditions. RNA-Seq is currently being used to assess gene expression profiles for the study of a number of complex

diseases, including cardiomyopathy, sudden cardiac death, and Alzheimer's Disease (AD).

CHIP-SEQ
Chromatin immunoprecipitation sequencing (ChIP-Seq) can leverage NGS to efficiently determine the distribution and abundance of DNA-bound protein targets across the genome at base-pair level resolution.

whole-genome bisulfite sequencing (WGBS) - Epigenetics methilation discovery
Aberrant DNA methylation and its impact on gene expression have been implicated in many disease processes, including cancer, neurological disorders, aging, and development. 38,39 High-throughput technologies, such as whole-genome bisulfite sequencing (WGBS), targeted bisulfite sequencing, and methylation microarrays, are powerful tools for investigating the dynamic state of DNA methylation across the genome. WGBS leverages the power of next-generation sequencing to provide a comprehensive view of methylation
patterns across the genome.

expression quantitative trait loci (eQTL)

Glossary

coverage level: The average number of sequenced bases that align to each base of the reference DNA. For example, a whole genome sequenced at 30× coverage means that, on average, each base in the genome was sequenced 30 times.

CpG site: A cytosine-guanine dinucleotide in DNA, separated by a phosphate. The p represents the phosphate group in the DNA backbone, indicating the 5'-3' directionality of the site. The C is always 5' of the 3'. CpG sites are locations where DNA may be methylated. CpG sites are often found in large groups (called CpG islands) in promoters. The methylation status of CpG islands often correlates with gene expression levels.

deep sequencing: Sequencing to high coverage levels. For example, WGS is typically performed to 30×–75× coverage while targeted NGS enables sequencing depths of 5000× or higher.

discovery power: In genomics, the ability to identify novel variants.

epigenetics: The process by which instructions encoded in DNA are turned into RNA and proteins. Gene expression is a process susceptible to variation and evolutionary selection, due to variations in the timing, location, and amount of gene expression.

gene expression: The process by which instructions encoded in DNA are turned into RNA and proteins. Gene expression is a process susceptible to variation and evolutionary selection, due to variations in the timing, location, and amount of gene expression.

genome-wide association study (GWAS): A study that examines genetic variation across many genomes and compares the DNA of people with a phenotypic trait, such as disease, to the DNA of control subjects. The goal of this type of study is to determine the causal variant for a given phenotypic trait, or a variant in linkage disequilibrium with a causal variant. To do this, researchers investigate the association between certain genetic variants (or alleles) and variations of phenotype. It is assumed that by using a large enough population, non-causative variants will disappear into the noise, above which the signal of true causal variants will be detected.

methylation: Occurs when a methyl group (carbon and hydrogen) is added to DNA, often at CpG sites.

next-generation sequencing (NGS): A non-Sanger-based high-throughput DNA sequencing technology. Compared to Sanger sequencing, NGS platforms sequence as many as billions of DNA strands in parallel, yielding substantially more throughput and minimizing the need for the fragment-cloning methods that are often used in Sanger sequencing of genomes.

paired-end reads: A strategy involving sequencing of 2 different regions that are located apart from each other on the same DNA fragment. This strategy provides elevated physical coverage and alleviates several limitations of NGS platforms that arise because of their relatively short read length.

read: In general terms, a sequence "read" refers to the data string of A,T, C, and G bases corresponding to the sample DNA. With Illumina technology, millions of reads are generated in a single sequencing run. In more specific terms, each cluster on the flow cell produces a single sequencing read. For example, 10,000 clusters on the flow cell would produce 10,000 single reads and 20,000 paired-end reads.

reverse transcription PCR (RT-PCR): An application to measure RNA expression levels using qPCR. RNA starting material is reverse transcribed into complementary DNA (cDNA) by the reverse transcriptase enzyme. Expression levels are usually expressed as a relative value, in comparison to the expression of a reference gene.

Sanger sequencing: The sequencing method, also known as capillary electrophoresis sequencing, developed in 1977 by Frederick Sanger. It involves sequencing DNA based on the selective incorporation of chain-terminating dideoxynucleotides by the DNA polymerase enzyme during *in vitro* DNA replication.

sensitivity: In genomics, the ability to detect low-frequency variants, or low-abundance transcripts.

sequencing panel: A subset of genes or genomic regions of interest in a targeted resequencing study. The sequencing panel can be amplified or enriched using sequence-specific probe sets. The sequencing panel, which represents a smaller subset of the whole genome, can be sequenced for a fraction of the time and cost compared to broader sequencing approaches. Targeted resequencing also enables deeper sequencing of the genomic regions of interest.

single nucleotide polymorphism (SNP): A mutation at a single position in a DNA sequence that occurs at >1% frequency in the population. If a SNP occurs within a gene, then the gene is described as having more than one allele. SNPs may or may not lead to variations in the amino acid sequence. Although a particular SNP may not cause a disorder, some SNPs are associated with certain diseases. These associations allow scientists to look for SNPs in order to evaluate an individual's genetic predisposition to develop a disease. In addition, if certain SNPs are known to be associated with a trait, then scientists may examine stretches of DNA near these SNPs in an attempt to identify the gene or genes responsible for the trait.

target region: A specific sequence of the genome, identified as a region of interest, due to possible involvement in or association with biological development, pathogenesis, or other area of study of interest to the investigator. The sequence can be a gene, a gene segment, a gene fusion, a promotor region, part of an intron or exon, or any stretch of sequence of interest to the investigator.

targeted resequencing: An NGS method that involves sequencing a subset of genes or regions of interest rather than sequencing the entire genome. The target region, which represents a smaller subset of the whole genome, can be sequenced for a fraction of the time and cost compared to broader sequencing approaches. Targeted resequencing also enables deeper sequencing of the genomic regions of interest.

transcriptome: Set of all messenger RNA (mRNA) molecules, or "transcripts," produced in one or a population of cells.

whole-genome sequencing (WGS): An NGS method where the entire genome of an organism is sequenced. WGS is ideal for discovery applications, such as identifying causative variants and for novel genome assembly. WGS can detect single nucleotide variants, insertions/deletions, copy number changes, and large structural variants. Due to recent technological innovations, the latest genome sequencers can perform WGS more efficiently than ever before.

Acronyms

ACPA: Anti-citrullinated protein antibodies

AD: Alzheimer's Disease

anti-PD-1: anti-programmed cell death-1

BMI: body mass index

ChIP-Seq: chromatin immunoprecipitation sequencing

CpGs: cytosine-guanine dinucleotides

EWAS: epigenome-wide association studies

eQTL: expression quantitative trait loci

FFPE: formalin-fixed paraffin embedded

GWAS: genome-wide association studies

IPSC: induced pluripotent stem cells

lncRNA: long non-coding RNA

meQTL: methylation quantitative trait loci

miRNA: micro RNA

MWAS: methylome-wide association study

NAFLD: Non-Alcoholic Fatty Liver Disease

NASH: Nonalcoholic Steatohepatitis

NGS: next-generation sequencing

PD: Parkinson's Disease

RA: Rheumatoid Arthritis

RNA-Seq: RNA sequencing

RT-PCR: real time PCR

scRNA-Seq: single-cell RNA-Seq

siRNA: small interfering RNA

SLE: Systemic Lupus Erythematosus

snoRNA: small nucleolar RNA

SNP: single nucleotide polymorphism

TNF: tumor necrosis factor

tRNA: transfer RNA

TWAS: transcriptome-wide association studies

WGBS: whole-genome bisulfite sequencing

WGS: whole-genome sequencing