AbdelRahman Adel
170122 96

Sheet 2

Lecture 1

(1) • Chromosomes have genes inside of them, chromosomes are made out of genes

• Humans have 23 pairs (46) chromosomes in all cells except reproductive cells where they have only 23 chromosomes

(2) • Genome is the whole DNA which is composed of multiple genes & Gene is only a small section of the Genome
• and approximately there are around 20,000 Genes

(3) Central Dogma says

$$DNA \xrightarrow{\text{Transcription}} RNA \xrightarrow{\text{Translation}} Protein$$

(4) Before bioinformatics, processing data took a long time and a lot of effort but after it, it was more efficient to process huge amounts of data and even further efficiency is obtained using parallel processing

(5) Late 90's, early 2000's due to the human genome project

(6) The Human Genome Project was the first large-scale international effort to map and sequence the entire human genome. And it was the most significant project ever since the start of research of the human genome.

(7) Phylogeny: it's the arrangement of species or sequences into groups based on distance and visualizing it is mainly using Phylogenetic trees & Heat maps
↓ more visual

(8) Gene expression Analysis provides dynamic info about genes and their functions and contributes to the personalization & specialization of drugs and disease mechanisms

(9) Microarrays are used to find if there is a specific gene is there or not and it helps with discovering the associated genes with diseases & phenotypes

(10) Omics is the study of something like genomics study of genomes and proteomics (proteome) ... etc & they are significant because they help in discovering biomarkers, diseases pathways and others

(11) Biomarkers indicate exacts genes for diseases unlike other like environmental markers that can indicate more than just the genes and other stuff that can be not related to the disease.

(12) • Disease pathways tells how the disease acts and how it interacts with the human
• Therapeutical " " " the treatment " " " " " "

(13) • Traditional therapy is more based in statistics for example that 90% of diabetics need a specific drug
• Personalized therapy tailors the treatment to an individual only to make the treatment most effective.

Lecture 2

(1) Alignment helps to see how different sequences can be similar or dissimilar to another sequence and it can be used with DNA, RNA & proteins

(2) Global alignment: Aligns the entire sequence → Needleman-Wunsch
Local alignment: Aligns a region with highest density of matches → Smith-Waterman

(3) Pairwise: Only Aligning two sequences to achieve the optimal pairing
Multiple: Aligning Three or more sequences to achieve the optimal pairing

(4) Scoring matrix: two dimensional matrix where the two sequences to be aligned are written along the axises

Scoring scheme: gives match, mismatch, and gap scores that can be used in the scoring matrix

Traceback: after finishing the scoring matrix you trace back to create the sequences.

(5) Inputs: 2 sequences, scoring scheme
Outputs: 2 aligned sequences
Methods: global alignment & local alignment.

(6) Genomic Variation is when at least 2 sequences have a difference between them. A SNP is single Nucleotide Polymorphism where it only looks at a singular base, and an allele has 2 types. Major allele and minor allele and they are the two forms of SNPs

(7) Major alleles, and minor alleles, we identify them statistically if the main SNP is majorly a specified based then it's a major allele else it's a minor allele

(8) Whole Genome Sequence: sequencing ~300,000,000 bases
Whole Exome Sequence: sequencing ~60,000,000 bases
Large Scale Genotyping: sequencing ~1,000,000 bases
↓ less expensive ↑ more expensive

(9) The indicate the start and end of the coding region and without them they can't know the genes of coding so they are not junk DNA

(10) BLAST breaks the query into "words" of fixed length and searches for exact matches while FASTA identifies short runs of identical residues as starting points.

Lecture 3

1. GWAS refers to the study in which hundreds of thousands of SNPs are genotyped across the genome and tested for association with the phenotype of interest
   Input: thousands of SNPs
   Output: Association with phenotype
   Methods: using the p-value & null hypothesis

2. Phenotypes are like features, for example, phenotype of diebetes if you have it or not and these phenotypes can be influenced by genome expressions, so much so that only One SNP can make it happen.

3. Manhattan plot are mainly used for GWAS, and it represembes the P values of each chromosome, on sets a threshold that anything above it is rejected by the null hypothesis

4. We use -log p-value to make it easier to start from zero and increase when the p value decreases. and the coding regions of chromosomes in the x-axis

5. Minor Allele Frequency: if a minor allele fregency under a specific percentage the subject is excluded from the Frequency output
   Genotyping rate: if a SNPs with a genotype rate missing more than 2%. exclude them
   Hardy-Weinberg Equilibrium: if SNP fail HWE then exclude

6. If the missing Rate of a Genotype is low (less than 95% are not missing) (5 more than 2 - 5% missing) the the study of that SNP would only give incorrect results.

7. PED → contains: Family ID, individual ID, paternal ID, maternal ID, sex Phenotype & genotype
   MAP → Contains: chromosome, marker ID, genetic distance, & Physical position
   BIM → Contains: chromosome, variant ID, genetic position, & Physical position
   BED → Binary file requires .fam & .bim
   FAM → contains: family ID, individual ID, paternal ID, maternal ID, sex & phenotype

(8) False positive can occur as typically hundreds of thousands to millions of SNPs are tested simultaneously for association with the phenotype
→ Solution: multiple hypothesis testing is performed & confounders presence.

(9) PLINK
Inputs: PED/MAP or BED/BIM/FAM files
Outputs: Frequency file & bead files
Filters: --geno , --hwe , --maf

Lecture 4:

(1) Population structure → Systematic ancestry differences like geographic proximity or individuals along the same ethnicity

(2) False Positive → when its a positive value but it's actually not positive
False Negative → " " a negative " " " " not negative
↳ Happens when the statistical signal of the marker is not strong enough because of: a) too little evidence or b) the threshold is too low.

(3) PCA creates principle components of the data and uses the first 2 or 3 PCAs (the heaviest and most important PCAs) as XYZ to make it possible to visualize population structure

(4) We can use box plots to show outliers and inliers
or we can use logistical & linear Regression

(5) Additive: for each minor allele you increase the value by 1, Range {0,1,2}
Dominant: if the SNP have a minor allele then it's 1 and 0 if anything else
Recessive: if the SNP is a homozygous minor allele then 1 else wise it's 0

(6) Two sample tests → creating a contingency table using case/control study without using any confounders.

(7) using the contingency table can be used to test of association using discrete test statistic, such as Fisher's exact test & χ^2 test or odds ratio

(8) All of them aim to find the p-value that would reject the null hypothesis with different equations

- Fisher's test: $p = \frac{(a+b)!\,(b+d)!\,(a+c)!\,(c+d)!}{a!\,b!\,c!\,d!\,n!}$
- χ^2 test: $\chi^2 = \sum \frac{(O-E)^2}{E}$ ⇒ then get p from χ^2 chart
- odds Ratio $= \frac{\text{odds}(A)}{\text{odds}(B)}$, $\text{odds}(A) = \frac{P(\text{trait}|A)}{1 - P(\text{trait}|A)}$

(9) Linear models are like $y = \beta_0 + \beta_1 x_G + \beta_2 x_C + \varepsilon$

(y: phenotype, x_G: Genotype, x_C: Covariates, ε: residuals)

(10) logistical models are used when the phenotype is binary not on a scale

(11) with LD we can figure out the correct SNPs if there was any missing snips.

Lecture 5

(1) It's a way to hierarchal clustering using Distance means. Input is a list of sequences.

1. Compute distance between all sequences
2. identify the least distance
3. Make them on the same level
4. Repeat until grouping all sequences

We utilize Euclidean distance and outputs Phylology tree

Lecture 6:-

① RNA polymerase ⟹ Creates an mRNA from DNA
Reverse transcriptase ⟹ creates single strand DNA from mRNA
(RT-PCR) ⇒ from mRNA we create cDNA

② cDNA represents the gene of interest DNA, not the whole DNA and it has a non-coding region to allow coding region to stick to the cDNA using the none coding region

③ Microarray assay:- are arrays that has many genes of the organism with single strand DNA probes tethered to a solid support when the cDNA is put in Microarray it glows a specific color
the Color corresponds to the presence of the gene or not.

④ T-test: how significant is the difference between gene expression in two classes
→ returns p-value
and we can use multiple testing correction of the p-value

⑤ Clustering → unsupervised learning outputs a number of clusters given
Classification → supervised learning that outputs a best fit model to predict other new values

⑥ Regular K-Means is a special case of Fuzzy K-Means when

$$P(\text{label } K \mid x_i, \mu_K) = \begin{cases} 1 & \text{if } x_i \text{ is closest to } \mu_K \\ 0 & \text{other wise} \end{cases}$$

⑦ WGCNA can be used for finding clusters of highly correlated genes, for summarizing such clusters using module eigengene or intra modular hub gene, default method of co-expression similarity $S_{ij} = |cor(x_i, x_j)|$