**UCL DIVISION OF BIOSCIENCES**

***IN COURSE ESSAY SUBMISSION AND FEEDBACK FORM (LEVEL 7)***

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CANDIDATE NO.** | **D** | **V** | **Y** | **M** | **9** |  |  |  | **Submission Date:** |  |
| **Module Code:** | **BIOL0050** | | | | | | | | **Word Count:** |  |
| **Title/Type of Coursework:** | **Assignment 2: Advanced Computational Biology** | | | | | | | | **Turnitin Submission Receipt #** |  |

|  |  |  |
| --- | --- | --- |
| **Have you been diagnosed as having Dyslexia by the UCL Dyslexia Centre?** | **Y** | **N** |

**SELF-ASSESSMENT (FOR STUDENTS TO COMPLETE)**

|  |
| --- |
| **Based on previous feedback, try to identify specific aspects of this work that you would like additional feedback on.** |

**FEEDBACK (FOR MARKERS TO COMPLETE)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Marker Name:** |  | **Date:** |  |
| 1. **What the student did well.** | | | |
|  | | | |
| 1. **What the student did not do well.** | | | |
|  | | | |
| 1. **Action: How might the student improve?** | | | |
|  | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **First Mark (%)** |  | **Second Mark (%)** |  | **Final Mark (%)** |  |

**NOTE:**

All assessed coursework is subjected to review by a Second Marker or the Course Organiser. The mark given here remains provisional until confirmed at the Examination Board.

# Assignment 2: Advanced Computational Biology

All code used to answer the questions are at the end this document.

## Question 1

A)

The median heterozygosity (H) across all SNPs for each of the five populations:

Table 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Population | CEU | JPT | TSI | YRI | PopX |
| Median heterozygosity | 0.299 | 0.279 | 0.307 | 0.304 | 0.311 |

YRI (Nigeria/ Africa) has one of the highest calculated median heterozygosities and CEU (Utah /Americas) one of the lowest this is as expected as the greatest human genetic variation occurs in Africa, where Anatomically Modern Humans (AMH) likely emerged and declines with distance from these areas; indicating modern human spreading across continents.

B)

The median FST, across all SNPs, between every pairing of the five populations = 0.294

The relatively high value for FST suggests that there is low migration between populations, hence the alleles in different populations tend to diverge. These indicate that the subpopulations differ in allele frequencies in comparison to the whole population. It has been established that variation between population groups is greatest when comparing different continental populations. The size of differences can depend on the length of time ago populations split; in humans, the time since splitting is not large, which is why FST is small.

C)

Population X could be a combined panel of CEU+TSI as this may provide more accurate information about groups and populations outside of the four other samples. The average of the two medians is similar to the median heterozygosity of PopX. In addition, the ARG of Popx for a 5 SNP region is similar to the combination of both CEU and TSI.

A close up of text on a whiteboard

Description automatically generatedD)

Time

(present to past)

A close up of text on a whiteboard

Description automatically generated

A close up of text on a whiteboard

Description automatically generated

A close up of text on a whiteboard

Description automatically generated

A close up of a whiteboard

Description automatically generated

Figure 1: ARG(s) of populations CEU; TSI; YRI; JPT respectively

E)

* Is the resulting ARG what you expected? Why or why not?

The collection of all trees in the regions are referred to as the Ancestral Recombination Graph (ARG) for the first five SNPs, of the first two individuals, for each of the five populations. Populations TSI, YRI and JPT have coalescent with recombination. Recombination events have resulted in the regions having many different trees, with which is required to explain the ancestral history. This also indicates that LD may be lower in these regions.

The expected time until each of the coalescent events: Where n = 10 samples

2/ n(n-1) = 0.022 generations

The mean time decreases as n increases because the time until any pair of individuals shares a recent ancestor decreases the more pairs of individuals that are compared.

The expected time until all individuals coalesce: Where n = 5

2(1-1/n) = 1.6 generations

## Question 2

A)

An R function ldsel was used to carry out computational simulations. The model assumes a population undergoing random mating with mutation (mu), recombination (rho) and selection (s) at two biallelic loci.

Median time to fixation= 1965 generations.

With Npop = 1000; from 419 observations where fixation of the A allele occurs.

A screenshot of a cell phone

Description automatically generated

Figure 2: Measures of LD in a simulation of the 2-locus WF model, with mu = 0, s = 0

and rho = 0.0.

The graphs show that the line AB largely follows the A allele line. The ab line shows an opposite pattern to the AB line in the graph. Linkage disequilibrium measuresr2 and |D’| between the two loci over this same time frame shows that r2 (the squared correlation coefficient) has a high value and |D'| is not visible on the graph. r2 = 1 can only arise if the two loci are perfectly correlated, which means that only two haplotypes exist in the population: AB and ab as illustrated on the graph.

B)

The parameter rho reflects the rate of recombination between the two loci. Here, the rate of recombination was increased while keeping the same init and s.

Median time to fixation= 1816 generations

With Npop = 1000; from 200 observations where fixation of the A allele occurs.

For recombination the corresponding measures of I D’ I and r2 have low values indicating that recombination is present. In general: I D’ I and r2 usually decrease with increasing recombination but are affected by loci's allele frequencies. The graph showed fluctuations for I D’ I.

**A screenshot of a cell phone

Description automatically generated**

Figure 3:Measures of LD in a simulation of the 2-locus WF model, with mu = 0, s=0 and rho = 0.01.

C)

The A allelic type at the first locus is undergoing selection, with the parameter s controlling the increase in fitness for haplotypes carrying A.  The effects of selection on LD can be investigated in using simulations.

The Median time to fixation= 496 generations.

For one value of S = 0.02 with Npop = 1000; from 255 observations where fixation of the A allele occurs.

With the initial frequency 0.01 of the selected A allele and from varying different values for selection (s), it appears that the higher the value of s the quicker the points of fixation for the allele frequency of A i.e. haplotypes that carry this mutation increase in frequency, with speed depending on the strength of selection.

A small value for the initial frequency mimics a newly arisen mutation. New mutants move I D’ I towards 1, then decline towards 0. r2 can also increasebut less so, depending on frequency of haplotype on which new mutant arises. As the novel mutant increases in frequency, I D’ I quickly declines to 0 unless where rho is small. As rho is relatively small, I D’ I may remain high for many generations, and meanwhile r2 can also become large. LD reduces over time under random mating and this region around the selected locus becomes narrower due to recombination, but for many generations there can remain an extensive genomic region in high LD with the selected allele.

A screenshot of a cell phone

Description automatically generated

Figure 4:Measures of LD in a simulation of the 2-locus WF model, with mu = 0, s=0.02 and rho = 0.00

D)

Median time to fixation= 2017 generations

For rho=0.3, S = 0.02 with Npop = 1000; from 230 observations where fixation of the A allele occurs.

With increase in rho the |D’I and r2 patterns have lower values indicating that recombination is present.

For an increase in recombination the corresponding measures of I D’ I and r2 should have low values indicating that recombination is present. Also, the graph pattern shows that with increasing values of s there is visibly more crossing over of lines for the haplotypes AB and Ab. In addition, the haplotype ab is lower at higher selection (s) values.

Recombination acts to break down associations (or Linkage Disequilibrium) amongst neighbouring sites (SNPs) over time. Near the end of the simulation the a allele is (nearly) lost, and this has different effects on |D’| and r2. |D0| = 1 as only three of four possible haplotypes are present from the population

E)

The allele A has been subject to recent positive selection as it has been shown to increase rapidly in frequency over a short time and demonstrate a high LD. Thus, regions of high LD (or highly-conserved haplotypes) can be a useful indicator of effects of selection and hence advantageous for the species' survival.

## Code

A screenshot of a cell phone

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Faculty of Life Sciences Level 7 (Postgraduate or Integrated Undergraduate Masters) Essay Marking Guidance**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Marks | 0-39 | 40-49 | 50-59 | 60-69 | 70-85 | 86-100 |
| Fail/Condonable Fail | | Condonable Fail | Pass/2ii | Merit/2i | Distinction/First | Distinction/First |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Knowledge and understanding of field**  *e.g. analysis and synthesis; deploying logical argument supported by evidence; focus on topic; drawing conclusions; ability to communicate ideas or material to diverse audiences* | Demonstrates little knowledge of the field.  Demonstrates significant weaknesses in the knowledge base, and/or simply reproduces knowledge without evidence of understanding.  Shows little or no critical ability.  Poor, inconsistent analysis. Use 0-15 where no answer or makes a single relevant point or a few disconnected sentences/ images. | Demonstrates knowledge of the field and awareness of current evidence and its interpretation, but with some notable weaknesses.  Lacks knowledge and understanding of some key areas.  Offers some appropriate analysis, but with some significant inconsistencies which affect the soundness of argument and/or conclusions.  Demonstrates very limited critical ability. | Demonstrates a sound knowledge and understanding of material within the field and is up-to-date with current developments.  Demonstrates relevant and sound analysis of material presented with some critical evaluation.  Is able to analyse complex questions and make appropriate judgements. May lack clear focus on the question in parts.  Able to communicate arguments, evidence and conclusions to specialist and non-specialist audiences. | Produces work with a clear focus on the question throughout.  Demonstrates a systematic knowledge, understanding and critical awareness of current problems and/or new insights, much of which are at the forefront of the field and informed by it.  Is able to evaluate methodologies critically.  Is able to deal with complex problems both systematically and creatively and make sound judgements in the absence of complete data.  Consistently able to communicate arguments, evidence and conclusions to specialist and non-specialist audiences. | Produces work reflecting excellent understanding of the question.  Displays mastery of a complex and specialised area with notable critical awareness of current problems and/or new insights at forefront of field.  Shows excellent ability to evaluate methodologies critically.  Deals with complex problems systematically and creatively, making excellent judgements.  Consistently able to communicate high-level arguments, evidence and conclusions to diverse audiences. | This work meets and often exceeds the standard for distinction/first class, as described in the 70-85 band, across *all* sub-categories of criteria.  The ideas and analyses presented are of publishable quality (with only very minor amendments) and would be likely to receive that judgement if submitted to a peer-reviewed journal.  Work is of such a quality that the student is clearly highly capable of doctoral research in the discipline and would be prioritized for a PhD studentship or DTP placement. |
| **Research and future enquiry**  *e.g. framing and creating further question or new hypotheses;*  *suggesting appropriate methods for gathering further evidence; awareness of methodological benefits/ limitations;*  *critical analysis of evidence* | Lacks any understanding of how established techniques of research and enquiry are used to create and interpret knowledge in the field.  Little or no skill demonstrated in understanding or selecting techniques applicable to the specific question. | Lacks sufficient understanding of how established techniques of research and enquiry are used to create and interpret knowledge in the field  Demonstrates some skill in selected techniques and/or approaches applicable to the specific question but with significant areas of weakness. | Shows some originality in the application of knowledge, and some understanding of how established techniques of research and enquiry are used to create and interpret knowledge in the discipline  Demonstrates understanding of, and skills in, selected techniques/ approaches applicable to the specific question | Shows originality in the application of knowledge, and a good understanding of how knowledge is created and interpreted in the discipline.  Displays a comprehensive understanding of, and skills in, techniques/approaches applicable to the specific question.  Able to identify gaps in current knowledge and formulate hypotheses or research proposals | Shows originality in application of knowledge, and excellent grasp of how knowledge is created and interpreted in the discipline.  Displays exceptional grasp of a range of techniques applicable to the specific question.  Able to identify gaps in current knowledge and formulate original hypotheses or research proposals |