.
Ниже в сумме 3 задания и 4 вопроса. В каждом задании несколько задач. Перед началом работы создайте копию этого документа. По итогу работы в новом файле напротив каждого Задания укажите задачи, которые сделаны. Напротив каждой задачи укажите ссылку на google spreadsheet и/или notebook (если написан скрипт) с результатами работы. Все скрипты должны быть реализованы на питоне. Если для одного задания в одном документе google spreadsheet/notebook результаты работы нескольких задач, укажите ссылку на документ под Заданием. Любые рассуждения, возникшие по ходу работы над задачами/заданием можно зафиксировать в любом месте, выделив их желтым цветом. Можете сами выбирать, какие задания и какие задачи делать. Например, вы можете сделать Задание 1 полностью и не делать второе. Либо, можете сделать задачи 1-3 Задания 1 и задачи 1-2 Задания 2. Внутри каждого Задания задачи обычно идут по уровню повышения сложности. Не рекомендуется тратить на тестовое больше 6 часов.
Ответы на вопросы по мере компетенций в области. Отвечать только по мере текущих знаний / того, что можете быстро посмотреть в литературе.

Общее введение: все задания касаются вопросов анализа данных целевого секвенирования ДНК. Если вы слабо представляете себе процесс, лучше посмотреть видео, где довольно доступно объясняется базовый принцип технологии.

Задание 1.
Введение: Микросателлит (микросателлитный регион) - регион генома, в котором одна короткая (от 1 до 5 букв) последовательность букв повторяется много раз (от 15 до 35). Микросателлитная нестабильность - часто наблюдаемое явление в опухолевых клетках, являющееся следствием нарушения системы восстановления нарушений ДНК. При этом нарушении ДНК-полимераза при копировании генома ‘проскальзывает’ на микросателлитах, что приводит к уменьшению их длины в геноме клетки. Например, рассмотрим микросателлитный регион **A**, представляющий собой повтор 19 букв T подряд. В опухоли без микросателлитной во всех клетках длина региона **A** будет ровно 19 букв (с маловероятными небольшими отклонения). В опухоли с микросателлитной нестабильностью образуется субпопуляция клеток, в которых регион **A** имеет вариабельную длину - в каких-то клетках он будет длиной 15 букв, где-то длиной 18 букв, где-то 19 букв (за счет того, что проскальзывание ДНК-полимеразы на микросателлитах - стохастический процесс - где-то проскользнет больше, где-то не проскользнет вообще). Если мы проведем секвенирование региона **A** опухоли без микросателлитной нестабильности и получим **N** чтений, покрывающих сайт **A**, то большинство из этих чтений будут содержать последовательность 19 букв T подряд. При этом будут также встречаться и последовательности 18 букв T подряд и 20 - это является следствием ошибок секвенирования. Если мы проведем секвенирование региона A опухоли с микросателлитной нестабильностью, то наблюдаемое разнообразие длин повтора в прочтениях будет значительно шире.
При том, в опухоли с микросателлитной нестабильностью, какие микросателлиты могут иметь вариабельную длину (являются нестабильными), какие-то могут сохранять узкий

спектр длины повтора (являются стабильными) - это также обусловлено стохастичностью процесса проскальзывания ДНК-полимеразы.

<u>Важная дополнительная информация</u>: микросателлит А может также иметь разную длину у разных людей (то есть, в популяции у него вариабельная длина) - у одного человека с рождения в повторе может быть 19 букв Т, у другого - 17, у третьего - на половине хромосом 17 букв, на другой половине 19 букв (за счет того, что, например, от мамы пришло 19 букв, от папы 17 букв).

Условия: было проведено секвенирование 12 образцов опухоли с микросателлитной нестабильностью (образцы MSI-1, MSI-2 и так далее) и 16 образцов опухоли без микросателлитной нестабильности (образцы MSS-1, MSS-2 и так далее). В ходе эксперимента были получены прочтения 33 микросателлитных регионов (регионы STR1, STR2, …, STR33). В ходе предварительной обработки данных для каждого образца был получен спектр наблюдаемых длин каждого повтора. Для каждого повтора в каждом образце получено 20 точек для каждого микросателлитного региона - доля прочтений, содержащих конкретную длину повтора. Результаты сведены в таблицу. По столбцам расположены различные образцы. По строкам - различные длины различных повторов. В ячейках - доля прочтений у конкретного образца, которые содержат конкретную длину конкретного повтора. Например, в данных секвенирования образца MSI-12, повтор STR1 имеет длину 14 букв в 7% прочтений, он же имеет длину 19 букв в 36% прочтений того же образца.

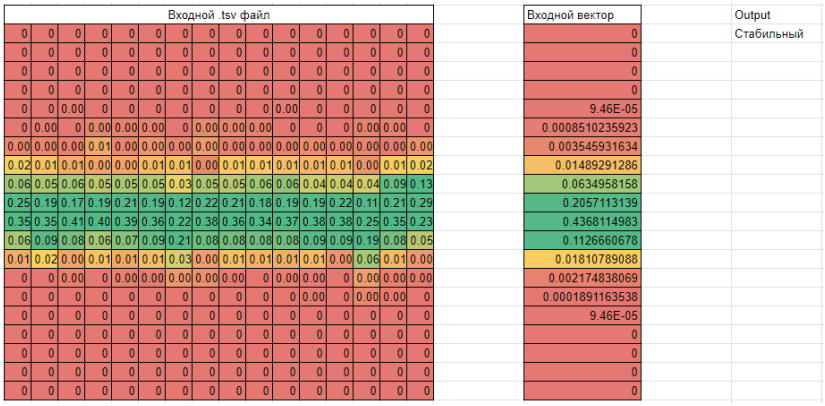
Задачи:

1. Создать новый google spreadsheet файл и скопировать туда содержимое таблицы по ссылке выше.
2. В новой таблицы вручную отметить для каждого образца стабильность/нестабильность микросателлитов STR1, STR2, STR4, STR14 (сделать это в новой строке под названием микросателлита).
3. Написать скрипт, который на вход принимает .tsv файл (20 строк и N колонок, где 20 - разные длины повтора; N - количество образцов <u>без</u> микросателлитной нестабильности) и вектор длиной 20. Скрипт должен выдавать является ли повтор (соответствующий входному вектору) стабильным или нет. См. Ниже пример input/output:

Input/output1:

Входной .tsv файл																Входной вектор	Output
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Нестабильный
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8.58E-05	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8.58E-05	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.001801879102	
0	0	0.00	0	0	0	0	0	0	0	0.00	0	0	0	0	0	0.01656012699	
0	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0	0	0.00	0.00	0	0.0743918658	
0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01922004376	
0.02	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.02	0.0374533442	
0.06	0.05	0.06	0.05	0.05	0.05	0.03	0.05	0.05	0.06	0.06	0.04	0.04	0.04	0.09	0.13	0.07790982024	
0.25	0.19	0.17	0.19	0.21	0.19	0.12	0.22	0.21	0.18	0.19	0.19	0.22	0.11	0.21	0.29	0.1638851946	
0.35	0.35	0.41	0.40	0.39	0.36	0.22	0.38	0.36	0.34	0.37	0.38	0.38	0.25	0.35	0.23	0.3677549444	
0.06	0.09	0.08	0.06	0.07	0.09	0.21	0.08	0.08	0.08	0.08	0.09	0.09	0.19	0.08	0.05	0.09344030203	
0.01	0.02	0.00	0.01	0.01	0.01	0.03	0.00	0.01	0.01	0.01	0.01	0.00	0.06	0.01	0.00	0.01364279892	
0	0	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00	0.00	0	0.00	0.00	0.00	0.001844780986	
0	0	0	0	0	0	0	0	0	0	0	0.00	0	0.00	0.00	0	0.0001716075336	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8.58E-05	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Input/output:



Входной .tsv файл																Входной вектор	Output
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Стабильный
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0.00	0	0	0	0	0	0	0	0.00	0	0	0	0	0	9.46E-05	
0	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0	0	0.00	0.00	0	0.0008510235923	
0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.003545931634	
0.02	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.02	0.01489291286	
0.06	0.05	0.06	0.05	0.05	0.05	0.03	0.05	0.05	0.06	0.06	0.04	0.04	0.04	0.09	0.13	0.0634958158	
0.25	0.19	0.17	0.19	0.21	0.19	0.12	0.22	0.21	0.18	0.19	0.19	0.22	0.11	0.21	0.29	0.2057113139	
0.35	0.35	0.41	0.40	0.39	0.36	0.22	0.38	0.36	0.34	0.37	0.38	0.38	0.25	0.35	0.23	0.4368114983	
0.06	0.09	0.08	0.06	0.07	0.09	0.21	0.08	0.08	0.08	0.08	0.09	0.09	0.19	0.08	0.05	0.1126660678	
0.01	0.02	0.00	0.01	0.01	0.01	0.03	0.00	0.01	0.01	0.01	0.01	0.00	0.06	0.01	0.00	0.01810789088	
0	0	0.00	0	0.00	0.00	0.00	0.00	0.00	0	0.00	0.00	0	0.00	0.00	0.00	0.002174838069	
0	0	0	0	0	0	0	0	0	0	0	0.00	0	0.00	0.00	0	0.0001891163538	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9.46E-05	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

4. Прогнать скрипт для всех повторов всех MSI… образцов. Результаты свести в исходную таблицу (под названием каждого повтора создать новую строку и ввести туда значение работы скрипта).
5. Сделать так, чтобы вместо результата “стабильный/нестабильный” скрипт выводил вероятность того, что входной вектор соответствует нестабильному микросателлиту. Результаты свести в исходную таблицу аналогично пункту 4
6. При условиях, что каждый образец микросателлитной опухоли гетерогенный (содержит как здоровые клетки,в которых нет микросателлитной нестабильности, так и опухолевые клетки, в которых есть микросателлитная нестабильность - это обусловлено процессом забора биологического образца - когда вырезают ткань, в образец может попасть и здоровая и опухолевая ткань), сделать так, чтобы скрипт в дополнении к первичному результату (либо ‘стабильный/нестабильный’ в задаче 3, либо вероятность нестабильности в задаче 5) выводил дополнительно процент опухолевой ткани в образце при анализе каждого отдельного вектора (при допущении, что если микросателлит нестабилен в опухоли, то он нестабилен в каждой опухолевой клетке). Результаты работы скрипта свести в исходную таблицу аналогично пункту 4.

Задание 2.

Условие: Было проведено секвенирование 15 образцов крови для детектирования наследственных генетических вариантов в генах BRCA1, BRCA2, ATM, CHEK2. Суммарно для каждого образца покрыто 40000 точек генома. Покрытие каждой точки для каждого образца составило не менее 100х (то есть, каждая геномная координата из 40000 покрыта не менее одной сотней различных прочтений). Для каждого образца определен набор потенциальных вариантов. Далее путем объединения составлен общий набор потенциальных вариантов всех образцов. Далее для каждого варианта в общем списке для каждого образца подсчитан процент прочтений, которые свидетельствуют о наличии варианта.

Результаты подсчета процентов чтений сведены в таблицу. По колонкам идут разные образцы. По строкам - потенциальные варианты. В столбце count указано, сколько раз этот вариант был реально задетектирован в 1000 других образцов. В ячейках - процент чтений, которые подтверждают наличие конкретного варианта в конкретном образце.

Поскольку везде секвенировалась кровь, в идеальной картине для настоящего варианта должно быть либо 50% прочтений, свидетельствующих о его наличии, либо 100% (соответственно, либо гетерозигота, либо гомозигота - когда у человека вариант есть только одной хромосоме или на обоих хромосомах). В реальности эти значения могут отличаться от 50% и 100% за счет ошибок секвенирования и для гетерозиготы обычно могут варьироваться в пределах от 5% до 95% в зависимости от степени ошибки (но для гомозиготы это значение не может падать ниже 98%).
Основная доля ячеек соответствует артефактам секвенирования - то есть, в действительности варианта нет в образце.
Среди 15 образцов нет родственников. Иначе говоря, если вариант редко детектируется в общей популяции (столбец count), то он должен также редко детектироваться и среди этих 15 образцов.
Каждый вариант должен удовлетворять закону Харди-Вайнберга.То есть, вариант не может всегда обнаруживаться только в гетерозиготе.
Ожидаемое количество вариантов в одном образце - от 20 до 30 (что соответствует средней вариабельности генов BRCA1/2, ATM, CHEK2 в популяции).

Задачи:

1. Создать новый google spreadsheet файл и скопировать туда содержимое таблицы по ссылке выше.
2. В новой таблице вручную отметить варианты, которые по вашему мнению являются настоящими (не артефактами). Сделать это необходимо только для тех вариантов, для которых значение столбца count 10 и менее. Отметить желательно путем выделения ячеек отдельным цветом (фон ячейки). Создать дополнительный столбец, в котором отметить, сколько раз задетектирован каждый проанализированный вариант.
3. Написать скрипт, который делает задачу 2 автоматически. На вход скрипт принимает .tsv файл (N*M, где N - количество вариантов, M - количество образцов). На выход скрипт выдает список задетектированных вариантов (например, “S2-chr11:108175463A>T ; S3 - chr17:41244429C>T ; S11 - chr17:41244429C>T”). Результирующий список вывести на отдельный лист google spreadsheet в два столбца (образец/вариант)
4. В приведенных данных один образец контаминирован другим. Это означает, что в пробирку с ДНК одного образца случайно попало немного ДНК другого образца. Это приводит к подмешиванию вариантов от донора к акцептору. Если, например, у донора есть вариант A, который подтверждается в 50% чтений, а у акцептора нет этого варианта вообще, то в наблюдаемых данных мы увидим вариант A у акцептора, только подтверждаться он будет уже не в 50% чтений, а в значительно меньшем количестве. При том, все варианты от акцептора подмешиваются донору приблизительно в равном количестве. При том от акцептора донору подмешиваются вообще все варианты акцептора. Необходимо найти донора и акцептора в приведенных данных. Результат написать текстом, например

‘контаминация от S1 к S2’. Дополнительно можете расписать на основании каких наблюдений сделан вывод

Задание 3**

Оцените априорную вероятность обнаружения соматического варианта в гене BRCA1 chr17:41209079T>TG (координаты по hg19/GRCH37) при профилировании опухоли молочной железы. Сравнить с вероятностью обнаружения этого же варианта, только с наследственной природой. Молекулярную эпидемиологию рака молочной железы и паттерн мутаций гена можно посмотреть либо в COSMIC, либо в cbioportal (обе базы содержат только соматические мутации, обращайте внимание на версии генома человека). Частоты наследственных вариантов гена BRCA1 в общей популяции можно посмотреть в Gnomad (а также по литературе. Обращайте внимание, что частота отдельного варианта в общей популяции и в популяции пациенток с раком молочной железы - это не одно и то же)

Вопросы:

1. For each of the following sequencing strategy write down limit of detection (i.e. analytical sensitivity, in variant allele frequency) for detection of somatic variants:
    a. Ion Torrent S5, coverage depth of variant site 250x
    b. Ion Torrent S5, coverage depth of variant site 1500x
    c. Ion Torrent S5, coverage depth of variant site 10000x
    d. Illumina MiSeq 150x2 reads, coverage depth of variant site 1500x
    e. Illumina NovaSeq 150x2 reads, coverage depth of variant site 10000x
    f. FASTASeq 100x2 reads, coverage depth of variant site 1500x
2. From the list below check alterations which can be identified based on NGS data (Whole Exome Sequencing (hybridization enrichment, at least 15bp exon padding), tumor-only sequencing (tumor content 60%), FFPE sample, Illumina HiSeq, 2x100bp reads, 300x average depth, at least 95% target regions covered by 50 reads or more):
    a. ALK-EML4 rearrangement
    b. BRCA1 germline missense variant
    c. High Tumor Mutation Burden (at least 10Mut/Mb based on only missense non-hotspot mutations)
    d. Microsatellite Instability
    e. METex14 skipping mutation
    f. EGFR kinase domain amplification
    g. HRD phenotype (based on CNV-derived genomic scars)
    h. HRD phenotype (based on mutational signatures)
    i. TERT promotor mutation
    j. FGFR1 high level amplification (7x or higher)
    k. ERBB2 amplification (2x or higher)
    l. Chromotripsis
    m. BRAF V600E somatic mutation present in 3% of tumor cells or lower
    n. APC somatic mutation present in 3% of tumor cells or lower
    o. 1p/19q co-deletion

3. Targeted sequencing was used to obtain sequences of BRCA1/2 coding regions. Which of the following alignment strategy in which cases (type of biological material analyzed/ enrichment strategy and so on) will be preferable:
    a. Alignment versus human whole genome
    b. Alignment versus human whole exome
    c. Alignment versus GENCODE human BRCA1/2 transcripts
    d. Alignment versus LRG sequences of BRCA1/2 genes
4. Write below SQL statement to retrieve count of somatic mutations identified per each patient (two columns in output: 1 - patientId ; 2 - count of rows from MutationResult table for this patient, where zygosity is 'somatic'). Each row in Barcode table may be associated with several rows from Analysis table, consider only rows from Analysis table where analysisRole is 'Major'.
   Relations: [Patient 1:n Case], [Case 1:n Barcode], [Barcode 1:n Analysis], [Analysis 1:n MutationResult], [Mutation 1:n MutationResult]