Project Title: "Rapid Gene-Drug Interaction Analysis for Personalized NSCLC..."
2024 CYSF Logbook: Isha Goyal.
Grade 10, Renert School.

June 29, 2023: Brainstorming Ideas:

↳ direction: data science, computational analysis — T-test, logistical regression, ANOVA, Tukey HSD.
  ◦ medical dataset?
the best datasets would be from the hospitals / research center
  ⇒ need to contact them on a specific topic ☆ finalize that first ☆
two types of dataset I can look at.
1. The disease / actual medicine (e.g. does obesity influence cancer)
2. The care (stats on ambulances, staffing etc.)
this data may be easier to get → already available?

August 1, 2023: Finding Datasets — all sites bookmarked in google!

◦ WHO: raw excel datasets ⇒ https://www.who.int/
  ↳ but I am not sure what I will do with these datasets
  ↳ some datasets need application/permission
    ☆ good ones (?): global health expenditure dataset [availability]
    tuberculosis data
◦ ~~KAGGLE~~ CDC ⇒ but kinda hard to navigate.
  ↳ use WONDER CDC ☆ (i'm testing it on a cancer set)
  YOU GET DATA FOR FREE! But i'm not sure how to read it XD
◦ KAGGLE ⇒ just make sure to read info to make sure it's legit (ngl... I'd make this a last resort type thing)
  I WILL ☆ WHAT LOOKS GOOD!

August 4, 2023: diverge from data sci → looking @ cool topics!

◦ preimplantation genetic diagnosis: pre fertilization screenings to look at inherited illnesses.
  → PGD (well-established) vs. PGS (not widely supported)

fertilize in a hospital, embryo put back into the woman.

for IVF

PGD: specific gene mutations (select embryos that are unaffected)

PGS: numerical chromosomal abnormalities (↑ likelihood of successful implantation)

so... what about NATURAL BIRTHS? (non-invasive testing)
- biomarkers?

non-invasive pre-natal testing (NIPT)
↳ usually look at aneuploidy disorder
↳ tests small fragments of DNA in plasma
high accuracy for Trisomy 21
but not used for cancers.

I don't think I can do anything about it...

August 9, 2023: Global Health Expenditure database
Explore the database from WHO.

how countries spend on health?
how much $ comes from government, households, donors?
how much comes from financing agreements?
data from 2000 - 2023 ← can monitor how spending changes
↳ how did COVID impact spending?

SO MANY VARIABLES TO CHOOSE FROM (a HUGE dataset)
↳ have to understand each variable
think the dataset is too complicated / NO! WE WONT GIVE UP :)

How to approach this:
↳ have very specific research questions
the goal is not to summarize the entire dataset, but use it to answer certain questions.

what focuses can we have?

- certain countries: mapping changes across different "world"-ed countries, different political systems pre / during / post COVID responses
  - => follow up with a discussion on optimization for future pandemics
- how is $ distributed across different services? what implications does this have for overall performance?
- how have shifts in gov't health expenditure allocation impacted healthcare system resilience, COVID 19 response effectiveness, and subsequent health outcomes during b4 / aftr COVID. [select countries]
  - ↳ follow up w/ external data stuff

future projections

sources of health expenditure

August 11, 2023: what is the project?

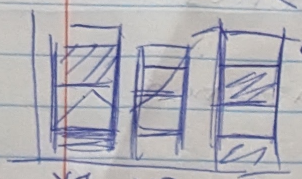
[a context-heavy data analysis]

The Impact of Government Health Expenditure Allocation on System Resilience and COVID-19 Response: A Comparative Study Before, During, and After the Pandemic.

1. Literature Review: summarize existing literature
   - introduce studies that talk about the correlation
   - contextualize the research problem
   - identify gaps
   - refine research question
2. Methodology
   - country selection criteria
   - data sources
   - explain the data limitations.

3. Data Collection & Analysis:
- make graphs for county expenditure

STACKED BAR



- how government health expenditure allocation changed
- how is COVID supplies being funded?

4. CONTEXTUAL INTERPRETATION
- compare all the graphs to political/economic situations → laws and regulations
- what was the COVID response? [lockdown, restriction]
  → reports
- what was the resilience

5. Going Forward: Policy Recommendations
- optimal resource distribution

6. Conclusion / why I am amazing

7. Limitations:
   → no outcome data
   → future steps

(I'm gonna seriously ask mentick now to make this work...)

August 27, 2023 : Country Selection (STEP 2)

we need to select countries based on the following!

- varying economies
- varying political systems
- how they handled the pandemic

ALL AROUND THE WORLD

- early action
- crushing the curve
- best at testing
- quarantining
- economic protection
- public communication
- lenient / laggards

select 4-5 countries.

first hard hit: china, italy

quick response: new zealand, East Asia

lenient: sweden

economic & social protection: SK

the cases i'm looking at: canada, taiwan, india

★ DO BG RESEARCH (ST. 1) B4 FINALIZING ★

September 16, 2023

(congrats doc)

- did some BG research → identified gaps in literature
- reassessed project → no new conclusions

September 24, 2023

Wait.. I dont like my project!

BACK TO THE DRAWING BOARD.

- machine learning? → SO DIFFICULT
- basic data analysis? —

machine learning:

- → medical image classification → classify into diff categories, detect anomalies
- → drug interaction checker // already exists → input list of medications
- → nutrition analysis tool → takes in data and provides diet recommendations

Semptember 25, 2023:

- what NOVEL problem can I solve with machine learning?
  // relation to medicine

September 30, 2023: new topic!

Personalized Medicine: treatment/disease/diagnosis is + other stuff
tailored to patients based on unique genomic data
// not a one drug fits all

- could hard-code or use machine learning to mateh sequencing to predisposition for disease (23 and me) or drug interaction

term: pharmacogenomics!

But:

- does having a human genome suggest drug reaction?
- can this be coded?
- hard-code or machine learning?

example: CYP2C9 Gene: encodes an enzyme that plays crucial role in drug metabolism:
→ several alleles/variants of this gene.

Impacts:
Warfarin → reduced CYP2C9 activity can require lower dosage: bleeding events
NSAIDS (eg. iBUPROfen) → lower enzyme activity may experience slower drug clearance

(1) Reprocess and align against reference genome
(2) Annotate specific genes // find what you want (genome coordinates)
(3) Compare with allele database
(4) Functional Implication.

what I gotta do: FIND A NICHE!

Oct 15 - FINALIZE PROJECT! ⇢ Research on google document

2 ideas to choose between:

HARDER

1. Cancer: often a result of many mutations rather than just one. When user inputs genetic code, can we flag mutations & find probability for disease?

Types of Cancer Genes:
- → Oncogenes: mutated ↑ growth
- → TS genes: slow growth, mistakes, apoptosis
  - usually there to help us! but can mutate
- → DNA repair genes: can't fix mistakes

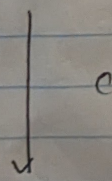
Known changes:
- → BRCA ⇒ mutation: BRCA1, BRCA2
  - women breast & pancreatic cancer!
- → TP53 ⇒ more than 50% of cancers
  - tumor suppress & repair
- → mis-match repair genes: MLH1, MSH2, MSH6, PMS2
- → BCR-ABL Fusion gene: 95% CML, 25% ALL (Leukemia)
- → ALK mutation ⇒ 5% NSCLC (XD)
- → BRAF: melanoma 50%

~~etc...~~ (i mean they kinda?)

niche: doctors don't usually genetic test for cancers!

Probabilities and stuff would be so difficult...

2. Personalized Medicine: how does your genome react to certain drugs and mutation? ★



Molecular profiling: what genes have been mutated, cancer biomarkers → why isn't a drug working?
- prevent, diagnose and treat your disease!

why go off the rack when you have something tailor-made?

so, this is the project! but what are we looking at?

emergency drugs? // emergency care / primary care

- chemotherapy? → side effects → resistance
- emergency drugs?
  - ↳ i feel like the drugs would already be good for all
- primary care? → very broad

project focus: chemotherapy → what cancer?
- → colorectal
- → breast cancer
- → adrenocortical
- → NSCLC

Title: Gene-Drug Interaction Analysis for Personalized NSCLC Chemotherapy Management: A Python-Based Decision Support System.

Nov 3: Talked to Mr. Gordon - lots of insight into the project
- → SNPs
- → papers
- → individuality
- → selling it

NOV 4: DAY 1/3: RENERT PROPOSAL:
1. looking at SNP's: what am I looking for?

NOV 5: Finish Slideshow
→ talk to Gordon ab ethics next week!

PRESENTATION FRI 15TH
(rehearse till then)

NOV. 16: Present to Renert ⇨ GOT IN!

NOV. 24: Meet w/ Gordon?
- Aquired vs. Inherentant resistance: can't measure aquired, only those you were born w/
  - SNP: Single nucleotide base substitutions
  - MND: multiple adjacent substitutions
- ~~At~~ Ancestry.com: not whole genome sequencing, but rather certain polymorphisms → PCR, and analyze
  → thus, I can avoid ethical concerns: it already exists!
- Need to find substitutions for specific drugs. Papers? Banks?

⇨ Gordon is main mentor!
- Need to get ahead in python - research on libraries.

IT REVIEW: Names: P53, PKM2 (Warburg Effect) | once someone has multiple of these, it is more clear that it will progress |
↑ many genes!

Find reviews!

- IS it connected to chemo resistance?
  ↳ maybe I can combine, or do one or another.

Nov. 27, 2023: Meet w/ Gordon

Reflection from presentation:

1. Resistance to drugs: 23&me type of testing can only test INHERITED response. Mutations only occur within the tumor cell ("single-cell genome sequencing")

2. Methodical papers: read same papers on similar topics - how did other scientists approach this? // send to Gordon.

3. From Soares' idea, we can also look into things like the Warburg effect - how easily can can cancer progress. My project can focus on 3 things:
   A) Bad chemotherapies
   B) Good chemotherapies
   C) Cancer progress → look into papers before committing.

Later... technical approach:
↳ what are the technical aspects?
↳ AWS? Artificial Machine Learning?

FOR NEXT WEEK:

① Read methods
② Read Warburg Effect ← what Soares sent!
③ Start looking into genes

Timeline

ONLINE IN-PERSON.

Dec. — Jan. — Feb. — Mar. 15 — Apr. 11

collect all info	program & test.	paper & presentation.

INTELLIGENT LEARNING RENERT

Assaf Gordon <assaf.gordon@renertschool.ca>

Some thoughts about Isha's Science-Fair presentation

2 messages

Assaf Gordon <assaf.gordon@renertschool.ca> Fri, Nov 24, 2023 at 10:29 AM
To: Shahin Jabbari <shahin.jabbari@renertschool.ca>, Iaci Soares <iaci.soares@renertschool.ca>

And by "some thoughts" I mean a looooong diatribe... :)

1. Resistance to drugs:
She mentioned two root-causes: "inherited / predisposition" and "mutations due to treatment".
The "genome sequencing" technology she's mentioning (and/or planning to use, like 23-and-me) is the kind that takes DNA from saliva - it will only detect "inherited / predisposition" kind of SNPs that lead to resistance to drugs.
It can not detect tumor mutations that resulted from chemotherapy or radiation, because these mutations only occur in the actual tumor cells, not in the rest of the cells in your body.
There are more recent advanced technologies that sequence the genome from a single tumor cell (and even from multiple different cells in one tumor) - but these are much harder and more complicated to perform.
Those are called "single-cell genome sequencing" technologies.

Two "review" papers:
Applications of single-cell sequencing in cancer research: progress and perspectives
https://jhoonline.biomedcentral.com/articles/10.1186/s13045-021-01105-2

Understanding tumor ecosystems by single-cell sequencing: promises and limitations
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1593-z

And an example of a breast-cancer single-cell sequencing for breast-cancer research (a shameless plug , since I'm on the paper, but I can also get you in touch with the lead researcher who is a good friend):
https://elifesciences.org/articles/51480

2. Python and "free": she mentions python is "free" and so it can be beneficial - technically this is correct, but there are more nuances there when it comes to what is free and what is not, and it's better to clarify what's going on there, and what will actually be free (I'm not against free and open source software, on the contrary, it's just a bit more complicated).

3. Using "AI" is not "magic solution" - we can't just say "we're going to use AI to do X" and think that "Chat GPT" will read a genome file and give us answers. It's a great idea, but will require lots of work to make it actually work - better set expectations realistically.
There are many examples of papers using "AI" to do something, it will be wise to read some of them, to understand what is meant by "using AI" and how much work the researchers really had to do in order to get results.
Here are few examples (from a quick PubMed search of free papers, not authoritative at all):

Artificial intelligence for multimodal data integration in oncology
https://pubmed.ncbi.nlm.nih.gov/36220072/

An expanded universe of cancer targets
https://pubmed.ncbi.nlm.nih.gov/33667368/

https://pubmed.ncbi.nlm.nih.gov/35879805/
A review of deep learning applications in human genomics using next-generation sequencing data

Nov. 29
→ reading over papers on AI gene sequencing!
(none were very helpful.. I'm gonna do some more tomorrow)

Dec 1. Reading data/ML genomic sequencing paper
- Note: ML isn't possible: taking a python route → look for papers on that
- Most papers use neural networks: DNN, CNN

Deepvariant - graphical variant
NGS - whole genome sequencing
Clairvoyante - allele alternatives!
↖ Clairvoyante is better

Disease variants:
- Deep WAS → identify disease-associated SNPs.

Epigenomics: locations & functions of all chemical tags that mark a genome.

Drug-gene interactions:
- DNN-DTI
- Drug Cell: predict drug response & synergy
- ☆ DeepSynergy: anticancer drugs.

PG. 15 - PYTHON PACKAGES: ⇨ look at those repositories!

For the weekend:
1. Finish paper - look @ libraries
2. Find pytho papers
3. Look @ Warburg Effect
4. Meet Gordon monday - start finding genes! → read papers → how do I
   - find these?
   - what to look for?
   - what format?

future research: WHAT do I need?
→ what part of genes? > read papers
→ what format?

Later... → how can I use python ML properly. || Find papers! (projects, program itself)

conclusion: i am very confused.

Dec 2. Warburg Effect

Energy cells need to function comes from ATP.
Generated in 2 ways
→ glycolysis (Krebs Cycle) ⇨ 2 ATP
→ Oxidative phosphorylation ⇨ 34-35 ATP ✱ what most healthy cells use.
~~but bacteria don't...~~

Krebs cycle can produce lactic acid during anaerobic respiration

- In cancer cells, even w/ O_2, they will perform ~~aerobic~~ anae glycolysis (Warburg Effect). Essentially, they stop ATP via oxidative phosp. & substitute it w/ ↑ glycolysis & glucose uptake.
- Reason is still unknown, but it's suggested that this switch to glycolysis may support cell proliferation
  1. Takes away glucose from T-cells ⇨ doesn't seem to appear in the germline!
  → Though producing less ATP, glycolysis consumes glucose QUICKLY, so surrounding T-cells can't function
  2. Invasiveness
  → ↑ H^+ ions secreted from lactic acid can diffuse into tumorstroma and ↑ invasiveness + metastasis

Other: promotes flux into biosynthetic pathways, allows for signal transduction through ROS (?)

Conc'n: these are TUMOR mutations, not the germline

Dec. 3, 2023: Call w/ Minal Aunty

Owns a lab that does real-time genome sequencing, and integrates machine learning into it. Might give me real patient data
- need to talk to biostatisticians about their work, get ideas
- ask them ab what I am specializing in (NSCLC chemoresistance) + next steps + how to achieve these steps.

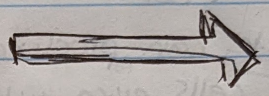
↓ can't

Their panel does cancer PREDISPOSITION, I want to do targeted therapy & chemoresistance. → whole genome sequencing or exome? SNPs?

Dec 7: HTSeq - Python framework for sequencing data
↳ prepping for meeting tmrw!

Dec. 12: Meeting w/ Manil

* Notes in Bio-Aro Meeting Research *



① • I need to get her to send me things
③ • Choose ~~NSCLC~~ or Prostate Cancer - research
② • ~~MAKE A TIMELINE~~

④ After I choose, do LOTS of research (meta-analysis)

Dec. 15 - CYST work

- Finalized timeline

IMPORTANT: 6 JAN ✱: FINISH GENE RESEARCH
JAN 10: CHECKED BY SUAREZ, GORDON, MANIL

✱ Manil said to do lung cancer

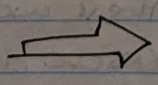
now I gotta do research :(

For rest of time, I was sick :/

Jan 2, 2024: TIME FOR RESEARCH!

Not really sure WHAT I need, so gonna look for now & make a spreadsheet later

✱ need to define drugs I'm using

research next 

GOAL: MAKE A LIST!

BioAro Meeting Research:

1. QUESTION: Genes with chemotherapy resistance: what are some, where can I find them?
    a. Inherited resistance (in germline) or acquired (in tumour cells)? Which will be better/ easier?
        i. For acquired, what control sequence will I use?
    b. How do I read these sequences? What do I need? What don't I need??
2. QUESTION: How complex is programming with Python/ML, given I don't know a lot of programming languages? Can I get it by March?

Research:
Genes for chemoresistance– are they general or broader?
Cisplatin Resistance: p53 sensitizes chemoresistant non-small cell lung cancer via elevation of reactive oxygen species and suppression of EGFR/PI3K/AKT signaling | Cancer Cell International | Full Text (biomedcentral.com)

ZNF300 promotes chemoresistance and aggressive behaviour in non-small-cell lung cancer - PMC (nih.gov)

pKM2 Gene:

Will I be focusing solely on NSCLC, another type of cancer, or do it generally?

What are some papers that use Python ML to genomically sequence? What are some of these programs?

MEETING NOTES:

- Lung cancer is mostly EGFR gene mutation → for that, there could be variants other than EGFR (used as control → for which chemotherapy is good)
    - If the patient has this variant, they will respond well to this drug
    - We have to categorize the variants (sensitive, responsive, and resistant)
- She can generalize the patient data and gene variants found in patients, what treatments were given to them

- Categorize drugs, treatments, and variants they have, and they are sensitive & responsive to them

Maybe not lung cancer?

- They are starting to work with prostate cancer
- How much literature can I get between cancers?

Key search terms: metanalysis, lung cancer, drug resistance, pharmacogenomics analysis**

- Make sure to read this criteria

But the larger the number of samples, the more reliable it is.

She says to check the literature!! – have they written scripts? Can I find the gene sequences?

HOW TO SEARCH:

She can give me them!! – She will give me the <u>well-known</u> **variants**

Go with well-known variants!

ARE HER VARIANTS TUMOR OR GERMLINE?

" For tumours, you need to do FFP tissue biopsy→ When you study, you get the variant in tumours (SOMATIC variants)

When you check your blood, you get GERMLINE variants

Our work in prostate cancer uses FFP tissues and then looks into somatic variants.

You can look at TUMOUR variants "

→ You will see normal and tumour pairs (if they had both tissues comparison– in that tumour you will find the mutation not present in the normal variant, which helps you have a control sequence)

FULL TUMOUR GENOME OR SPECIFIC PARTS?

- People usually just amplify what they need to detect the cancer: DNA is taken from a small part, and amplify
- Either whole exome, or just targeted genomic sequences (what do you want to look for?)

- Targeted genomes will be MUCH EASIER! – identity 70-100 genes (wtf.) and just look for variants for which drugs are sensitive, responsive and resistant.
- Use well-known, "crisp" genes so you have lots of samples for a few genes
- While reading papers, go for TARGETED PANELS specific gene sequences amplified for research.

Stepping point. Literature search.

- DO NOT GO FOR THE ENTIRE SEQUENCE, JUST NEED THE SPECIFIC ALLELE (specific to chromosome position as well). What is the consequence of that mutation?
- She'll send me how to read chromosome position
- Non-synonymous, stop-gain, framework: type of mutation
- RS ID IS IMPORTANT! well-known , published
    - Should be linked to AI to connect to info (eh… k)
- Wtf is an exome vs. genome
- Use ClinVar to find genes :)
- Use metanalysis and pharmacogenomics

Do not fear… she has software that detects mutations & the position, and give you results (which we can tie into a drug response)

Jan. 2nd

SEARCH CISPLATIN

Paper 1: Cisplatin in NSCLC Cancer therapy
"Protein interaction network that indicates highly dysregulated
X TP53, MDM2, and CDKN1A genes -- upregulated in
cisplatin-resistant cells"

Paper 2: Cisplatin & cancer chemotherapy
"Knockdown of CTR1 reduces intracellular accumulation
& uptake of cisplatin"

· Paper 3: chemo-radio resistance
X "↑ CTR1 means more accumulation -- expression of ATP7A
and ATP7B is upregulated → chemoresistance" (ovarian cancer?)
MDR1 pumps it out

Paper 2: miR526b-3p

Paper 4: The Drug-Resistant Mechanism... 5 therapies (platinum therapy)
X → ⇧ hCTR1 increases SENSITIVITY
X → verifies CTR1, ATP7A/7B when upregulated
X → ⇧ MRP2 leads to resistance (small cell lung cancer)
X → ⇧ MRP4 chemor"
→ knocking down HIF2a ↑ A549 cisplatin therapy (idk if relevant)

Paper 5: Drug resistance & combating
X → ⇧ ABCC1 gives chemor"
→ ABCC3?? maybe?

Paper 6: Expression of Multidrug resistance --
X MRP3, MRP (check which gene encodes it) X overexpression
★ check all ABCC/MRP superfamily

Paper 1: ABC transporters: ABCA1, ABCA2, ABCB1 (MDR1), ABCC2, ABCG6
(MRPs): ABCB4, ABCB11, ABCC1-6, ABCC10, ABCC11, ABCG2
only underlined are direct efflux ⇧ verify!
ATP binding cassette

Paper 7: Battling Chemoresistance
→ ABCG2

X other: PHF53 X overexpression sensitizes cells to cisp-1 doc. in NSCLC

GEMCITABINE + DOCETAXEL:

Paper 8: Tumour BRCA1 -- (research study)
→ Gem. resist" X RRM1, RRM2 ⇧
→ ~~BRCA1 ⇧ sensitivity to docs & paclit~~
~~low expression~~

X BRCA1: low exp. ↑ cisplatin sensitivity, high exp. ↑ resistance to pacli. & doce.

(general)

Paper 9: Mapping genetic alterations → look @ references for further research

TP53 mutations ↑ resistance cisplatin WHICH MUTATION? T-T

✓ X · MDM2 & 4 might cause it?

X · ERCC1?

Jan 3rd

SEARCH DOCETAXEL (didn't work)

Paper 10: ~~X~~ PHF23 ↑↓ sensitizes to cisplatin & docetaxel (?)

• Look at MVP/LRP (lung-resistant protein) → what genes are expressed

CP. 11: the molecular mechanism of chemoresistance)

→ Heme oxygenase (HO)-1 ?

X ↑enzyme!

→ ~~Erk~~ ✓ K-ras mutation: res^nt to gefitnib, erlotnib, sunitnib.

↑ enzyme!

X → miR-138, miR-145, miR-489 suppressed chemoresistance

X → EGFR overexpression

activating NF-κB & STAT3

is bad

· cisplatin: EGFR-mediated

? PI3K/Akt & NF-κB pathways

Area for growth: i didn't when taking ~~clinical~~ tho :) ~~trials~~, my project doesn't consider demographics & inc/exc. criteria; the project is very generalized.

→ however, I use the big ones.

Paper 12: Recent Progress in ... OMG.

THE JACKPOT

· EGFR: T790M ✓, C797S ✓

· ALK: G1202R ✓

· ROS1: G2032R ✓ ✱ read whole paper

· RET mutations ? ✓

· MET: D1288, Y1230 ✓

· NTRK: G595 ✓

+ KRAS, ~~G12C~~ ✓, PELP1 ... X

ALSO I only use mutations, not amplification or expression

Paper 13: Drug resistance in Non-Small ...

• NOTCH pathways:

→ CD133 ✓

→ CD44, NANOG ✓, OCT4 ✓, SOX2 ✓, ALDHA1 ✓

platinum taxanes ✓ ↑ ABCB1 ✓, miR-451 X

EGFR inhibito: → ~~T790M~~ ✓, C797S ✓

ALK → C1156Y, L1196M, G1269A, E1512Q

Jan 4: Sequence Hunt!

KEEP IN MIND: targeted panels (1)
RS ID (2)
Clinvar? Genebank? (3)

- Working backwards in paper
- Gene overexpression DOESNT WORK: need mutations itself

Go over previous mutations,
- Ex KRAS ✓
- TP53 ✓
- → Casette-binding
- MVP
- micro-RNAs. ✗ don't work

1) Keap1 deletion

Question: in paper 13, it says CD44, NANOG, etc are resistant if expressed in the lung cells. Can I assume they are expressed if they are?
(2) are they all expressed together?

expression in RNA.
use it! ✓ ☆ just take note to check for expression.

Jan 5, 2024:

Meeting with Manil: How to read genome stuff on Clinvar

Goal plan today:
1) Finish the ~20 gene mutations on spreadsheet ☐ ✗
2) Find more mutations ☐
- Look @ casette-binding ✗
- Look @ MVP/LRP ✗
- Look @ TP53 ✗

useful:
! ASK how to read on BOTH sites
! Ask question above ⇧
! RECORD MEET & TAKE NOTES
! Common sites to find mutation
- cytogenic vs. genomic
- what I need to take note of

Meeting Notes:

Genomic location: 12:

↑ chromosome #

G-T on chromosome 12.

GRCh38 vs. GRCh37

↑ resequencing: new chromosome alignment

- My nucleotide is on chromosome 12, position 25245351

Cytogenetic: chromosome #

→ divided to top (p) & bottom (q) chromosome

12p12.1

(12th chromosome, p-arm, site 12.1)

TITLE: change of amino acid.

c.34G>T

↑ codon 34

↑ base changes G→T

⇨ amino acid (p. Gly12Cys)

glycine → cystine.

- OMIM: mammalian inheritance X
- dbSNP: variant database! + RSID

Bottom: consequence

(if Gly→Cys, it will affect functional consequence: might need pathogenic)

→ you can read the papers to ensure its valid.

We have different understandings of the project...

She wants:
- 100+ variants
- pathogenic
- AI searching database

I could do pathogenic; sensitive; resistant.

I want
- only a dozen
- primarily scanning sequences
- hard-code it in.

NOTE: my project is deviating from chemotherapy & going towards targeted inhibitors. (fine.)

Jan 6, 2024
still doing research :/
Finished 13 I already had, new sheet for expression

"The emerging landscape" - Paper 14
· RAS mutation + targeted therapy
T790M is HUGE!

MET D1288N ✗ can't find
~~EGFR~~ CD74-ROS1 fusion gene! 1st & ALK ★
ALK-resistant: ~~SFM G1202R~~ - 2nd-gen ALK inhibitors.
~~ALK-L1196M~~ ★ already have.

"Rising Resistance" - Paper 15
↳ ~~EGFR T854A~~ ★ ✗ can't find
maybe ~~L858R?~~ ✗ drug sensitive. ~~D761Y~~ ?
~~G595R~~ ‼ 3rd gen might work
Osimertinib

2nd-day resistance (EGFR inhibitor) — can't find fusion gene.
NOT really researched

★ Can do small sample-size with ~15 most common mutations for CYSF
⇨ not a lot of mutations are known [not my fault]
⇨ area to growth: look at expression - resistance
Just make sure what I have is the best it can be
+ CODE IS PRIORITY.

Jan 7, 2024 LAST DAY OF MUTATION SEARCH
Exon-14 skipped is SENSITIVE? ⇨ all papers say another mutation + ex14
↳ can't read it off genome.

Criteria:
- well-researched variant → literature reviews → clinivar
- contributes to NSCLC resistance in either chemo. or targeted therapy
- variant is a result of nucleotide base change.

Exclude novel variants
resistance as a result of over/under expression, OR RNA-transcription (micro-rna's) OR amplification.
inherited resistance: must be present in somatic cells

January 19:
Finished p53 mutation: START CODING!

February 2nd - Dallas helps me download biopython (continues on Monday)
DONE ON MONDAY!!!

February 9th - what does my code need?
- Hard code w/ sequences that are chemoresistance — } BLAST!
- Way to compare your gene to that one
  - ⇨ listing the mutations
  - ⇨ identifying it as problematic
  ] - "normal genome" to compare to

OR: just see if your gene = problem gene.
[output: gene information
mutation
problem
treatment]

IMPACTS!
Paper done by March 7st - Soares edits.
→ intro / problem ①
→ method: ②
↳ inclusion/exclusion ↔ code structure
③ ↳ research methodology • objectives
→ results ~~↳ what genes we~~
- gene search
- code

lymphatic & fistr

C+G C+G EGFR: CTG
CAG ALK: AG
TTT N: G
other 3: T

Feb. 20: Begin Coding

(1) Can't find at how to download variant, so I will just download the gene & check for SNV.

Problem: I am getting 2 letters for EGFR position..

Steps A) get gene
B) find position of mutation
C) make sure it works
repeat the everything

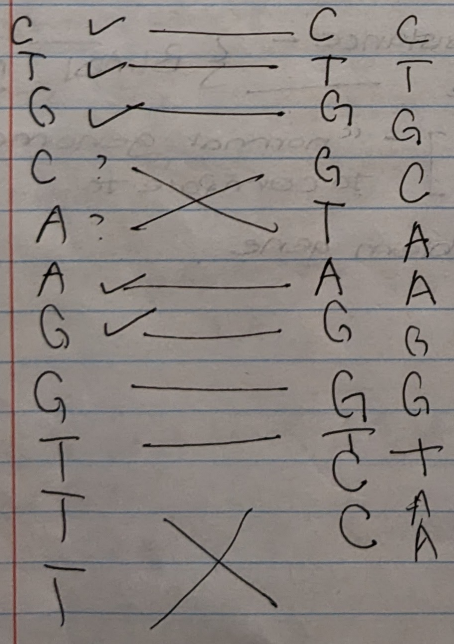
today.

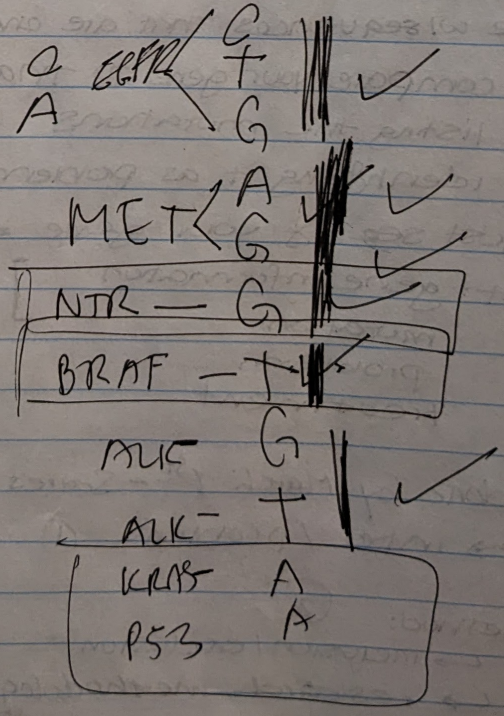
2 sequences.
Calculated the one I don't want)

Update: ~~code~~ now discards the second sequence (I don't need).

NOTE: TAKE CHROMOSOME POSITION

*could not make 3 ALK ~~positive~~ variants work





Feb 21: Align & Detect

Writing this part isn't horrible IF ENTIRE GENE is present

↳ deal w/ insertion/deletion?

11:58AM: wrote some basic code, testing w/ EGFR (made up a random patient genome)

TEST EGFR fna changed first nucleotide from A → T
last from A → G

My sequences are too large..
(LIMITATION)

brain dump:
1) what if insertion/deletion occur? — research that

12:24AM: computer is checking memory.

① do align & detect
② research insert/deletion ⇒ test manually.
③ plan output
④ paper

1:01 - use P53 sequence to test

TESTP53 < 1st nucleotide C→G.
2nd Last nucleotide A→T

NO mutation detected.

1:23. 4th Line: TC → AA (first 2 nucleotides) X

Feb 21: Sequence Alignment

```
from Bio.Align import PairwiseAligner
from Bio import SeqIO

# Define function to align sequences and detect mutations
def align_and_detect_mutations(reference_sequence, patient_sequence):
    aligner = PairwiseAligner()
    aligner.mode = 'global'  # Set the alignment mode to 'global' for starting from the beginning
    alignments = aligner.align(reference_sequence, patient_sequence)

    for alignment in alignments:
        print("Alignment score:", alignment.score)
        print("Aligned sequences:")
        print(alignment)

        # Convert aligned sequences to strings
        aligned_reference = str(alignment.aligned[0])
        aligned_patient = str(alignment.aligned[1])

        # Initialize positions in both sequences
        reference_pos = 0
        patient_pos = 0

        # Iterate over each character in the aligned sequences
        for ref_char, pat_char in zip(aligned_reference, aligned_patient):
            # Check for gaps
            if ref_char == '-':
                print(f"Gap at position {reference_pos + 1} in reference sequence")
            elif pat_char == '-':
                print(f"Gap at position {patient_pos + 1} in patient sequence")
            # Check for mismatches
            elif ref_char != pat_char:
                print(f"Mismatch at position {reference_pos + 1}: Reference {ref_char} - Patient {pat_char}")

            # Update positions
            if ref_char != '-':
                reference_pos += 1
            if pat_char != '-':
                patient_pos += 1

# Open and parse reference sequence (streaming)
with open("P53.fna", "r") as reference_file:
    reference_sequences = SeqIO.parse(reference_file, "fasta")
    reference_record = next(reference_sequences)
    reference_sequence = reference_record.seq
```

Mutation reset

Line 59: first nucleotide: C → G [GAGCACTTT]

mutation reset: won't detect

2:57 PM: P53 copy 1st nucleotide in sequence:

C → G

IT WORKS!!!!

Next step: ① Detect SPECIFIC MUTATION - output → I can test it for all genes

② Make it streamlined for all genes.

③ [Later] User interface.

Feb. 22: Cont

Input: gene to test [

if EGFR: ① plus or m

if ATK... ② detect

1:19: Test P53 Note: minus strand

last nucleotide A → G [

TO-DO:

① Align & Detect X

② "If" specific mutation chemoresistance

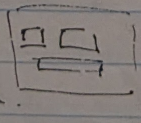
③ Output

④ Cont Paper.

NOTE: (2) @ breakpoint

1:31 - says "no mutations detected" T-T

NOTE: if minus strand, it will show you the COMPLEMENTARY mutation.



3:21PM: trying to make it align.

I might just not align them.

EGFR: T790M - 55181378 (pks)

55019017

C → T

index [162, 360]

OUTPUT

~~Mutation detected at index ~~.

chemoresistant mutation detected @ index ###

Variant Type: SNV. → NAME: ____.

Mutation:

Amino Acid Change:

Molecular Consequence

Drug Resistance

MET - 116672196

BRAF - 142239131

NTRK1 - 156815750

* ADD Chromosome Position.

* code already taken ins.

ALK: 29192774

Feb. 23: Final Day (?) :)

* check BRAF

- removed BRAF.

Feb 24: kept working on paper // ill revist the code later?

Feb. 29: Still writing paper

Flow of code:

Input: Genes to test (EGFR, ALK, MET, NTRK1, KRAS, TP53)

⇒ Import Patient Sequence.

plus ↓

"Align & Detect Mutations" aligns both reference & patient sequence

MINUS ↓

"Compare_sequence" function ~~returns~~ detects nucleotide change ~~& returns mutations & position~~ & compare to chemoresistant

↓ No change: healthy gene

↓ Mutation (non-resistant): prints mutation

↓ yes chemoresistant: print mutation name, type, class, amino acid change, nucleotide change, resistance, molecular consequence.

March 10

- finish slideshow
- start script tmrw

March 12

- finish video
- upload to platforms
- SUBMIT LOGBOOK