

## שחזור עץ פילוגנטי על תאי נאופלזמות מיאלופרוליפרטיביות

שמות חברי הקבוצה:

- גל סזנה 318510633.
- נועה מרגוליס 208635334.
- עדי יפרואימסקי 326045515.
- יואל מרקו 325245280.
- איתן סמסון 214935165.

### רקע:

- המאמר: "Life histories of myeloproliferative neoplasms inferred from phylogenies".
- פרטים: המאמר פורסם בכתב העת Nature, כרך 602, בתאריך 3 בפברואר 2022. המחברים כוללים את ניקולס וויליאמס ואחרים ממוסדות כמו מכון וולקאם סנגר והמחלקה להמטולוגיה באוניברסיטת קיימברידג'.
- תקציר רקע: מחלות מיאלופרוליפרטיביות (MPNs) הן סוג של סרטן דם כרוני הנגרם ממוטציות גנטיות סומטיות בתאי גזע המטופויאטיים (HSCs). המאמר עוסק בזיהוי מוטציות סומטיות ותיעוד ההיסטוריה הגנומית של מושבות מתאי גזע של חולי MPN. זוהו 580,133 מוטציות סומטיות לצורך שחזור אילנות יוחסין המטופויאטיים ובחינת ההיסטוריות והתרחבות של השיבוטים.
- המחקר מצביע על כך שרכישת מוטציות מובילות (driver mutation) מוקדמות והתפתחותן לכל אורך החיים הן הבסיס למחלות מיאלופרוליפרטיביות במבוגרים, דבר שמעלה הזדמנויות להתערבות מוקדמת ומציע מודל חדש להתפתחות סרטן.

## שאלת מחקר ומטרות:

- שאלת מחקר:
  - כיצד מוטציות גנטיות מוקדמות ותהליכי התרחבות שיבוטיים משפיעים על התפתחותן של מחלות מיילופרוליפרטיביות לאורך החיים?
  - מתי בערך בחיים אדם עלול לקבל מוטציית דרייבר?
- מטרות:
  - נרצה לנצל את הידע הנלמד בקורס בנושא בניית עצים פילוגנטיים על מנת לזהות פיצולים (מוטציות) בהתפחות התאים שמובילות לסרטן להבין את התזמון של רכישת מוטציות מובילות (driver mutations) ואת התרחבותן.
  - שיחזור תוצאות המאמר עליו אנחנו מתבססים.
  - הכרת עבודה עם דאטה ביולוגי.

## דאטה:

- [קישור למקור הדאטה סט.](#)
- המידע נלקח מתוך קובץ נתונים גנומיים בפורמט VCF (Variant Call Format), שמטרתו לתאר מוטציות גנטיות באזורים ספציפיים בגנום. במקרה זה, מדובר במוטציות סומטיות שנמצאו בתאי דם שנדגמו מ-12 פרטים. אלו מוטציות שנרכשו במהלך החיים ולא תורשתיות.
- ראו הרחבה על הדאטה ונספח ניתוח הפורמט.

## היפותזה:

מוטציות מניעות שנרכשות בשלבים מוקדמים בחיים מובילות לדינמיקות התרחבות משתנות לפני האבחנה הקלינית. הבנה זו תוכל לשפר את הגילוי המוקדם ולתרום להתאמה אישית של טיפולים.

## שיטות חישוביות:

- שימוש באלגוריתמים של בניית עצים פילוגנטיים (MPBoot).
- בשיעור למדנו שתי שיטות מרכזיות למימוש האלגוריתם הנבדלות זו מזו בקריטריון לבחירת הזוג הכי קרוב של הקודקודים – האחת UPGMA השנייה Neighbor joining.
- מודלים בייזיאניים לחישוב קצבי התפשטות קלונית.
- ניתוח סטטיסטי להשוואת מוטציות מניעות בין קלונים.

## שלבי ביצוע:

1. ניתוח ועיבוד ראשוני של הנתונים ובדיקת איכות.
2. בניית עץ פילוגנטי עבור כל הדאטה המצורף במאמר:
  - השוואת גנים של חולים ברצפים הייעודיים לגנים בריאים ותיעוד היסטוריית המוטציות.
  - נבדוק איזו שיטה של בניית עץ פילו גנטי תניב תוצאות טובות יותר. NJ הוא ללא דרישה לאולטרמטריות.
3. ניתוח התוצאות מהעץ ומסקנות.

## בביבליוגרפיה:

Williams, N. et al. (2022). Life histories of myeloproliferative neoplasms inferred from phylogenies. \*Nature 602, 162–168.

## גיוון הקבוצה:

- 3 סטודנטים לביולוגיה חישובית, סטודנט למדעי המחשב חד חוגי, סטודנטית למדעי המחשב עם חטיבה בביולוגיה.
  - 3 עתודאים, 1 קצין במיל' (תותחנים), 1 קצינה בקבע (חיל הים) – שילוב הזרועות השונים.
  - 2 דתיים, 3 חילוניים.
  - ג'ינג'י אחד.
  - גם בנים וגם בנות.
  - 4 משקפופרים 1 ללא משקפיים.
  - גיוון אתני – ישראל הראשונה עם ישראל השנייה.
- הקבוצה שלנו מגוונת ברקע האקדמי, המקצועי והאישי של חבריה. אנו כוללים סטודנטים ממסלולים שונים כמו ביולוגיה חישובית ומדעי המחשב, לצד עתודאים, קצין במילואים וקצינה בקבע מחילות שונים. בקבוצה יש ייצוג של דתיים וחילוניים, בנים ובנות, וכן גיוון אתני המשקף את החברה הישראלית. אלמנטים אישיים כמו ג'ינג'י אחד ומשקפופרים מוסיפים צבע וגיוון נוסף. המגוון שלנו מאפשר שיתוף פעולה ייחודי וחשיבה רחבה ויצירתית.

## הרחבה על הדאטה

כל המוטציות שנתעסק איתם במאמר הם מוטציות סומטיות ולא מוטציות מתאי נבט, מסוג (סניפ) החלפה של בסיס בודד או מסוג (Indel) הכנסה ומחיקה של בסיס בודד או מספר בסיסים. בתהליך עושים **single-colony whole-genome sequencing** על תאים מהדם או ממוח עצם. כלומר, התאים שנלקחו הופרדו לתאים בודדים והם הופרדו לתרביות שונות, ושם נתנו לכל אחת להתחלק ולהתפתח לקולונה.

כל קולונה מייצגת לנו גנום של הבן אדם בשלב כלשהו בחיים שלו, עם מידע כלשהו של מוטציות שהוא כן סבר ומוטציות שהוא לא סבר. במחקר נבקש להבין איזה מסלול של צבירת מוטציות מוביל לסרטן ומתי בערך אמורה להיות המוטציית דרייבר.

יש בידנו מוטציות שמתאפיינות עם סרטן ולכן נוכל להגיד שקולונה עם מסלול מוטציות שמוביל למוטציה עם סרטן הוא מסלול מוטציות שמוביל לסרטן.

בשביל שני הכיוונים נרצה ליצור עץ פילוגנטי שמכיל את המסלולים של מוטציות מסוימות בגוף.  
אז מה צריך בשביל להפיק את העץ הפילוגנטי?

ראשית, צריך את המידע מהמחקר על איזה מוטציות כל קולונה צברה ואיזה היא לא צברה. יש לנו כבר מקומות שחשודים למוטציה שאנחנו יודעים עליהם. מכל קולונה לקחו כמה תאים ועשו להם **WGS** וקיבלו כמה גנומים מאותה קולונה ובדקו על הגנומים האלו במקומות החשודים במוטציה האם מופיעה אותה מוטציה או לא. ספרו את מספר הפעמים שהופיעה וכתבו "1" אם החליטו שיש, כתבו "0" אם החליטו שאין וכתבו "?" אם לא יכלו להסיק מסקנה חד משמעית. כל המידע הזה מאוגד לתוך קובץ מסוג **VCF** (הסבר על הבנת הקובץ בנספח א'), על הקובץ הזה עושים אנליזות בשביל להוציא את העץ הפילוגנטי.

## נספח א' – ניתוח פורמט הדאטה

הקובץ מתחיל בתיאור של המידע שמוכל בתוכו, הקטע של התיאור של הצגת הנתונים מתחיל ב##. אחריו מגיע טמפלייט של מבנה הצגת שורה. כל שורה היא בדיקה של וריאנט סומטי ספציפי. אחרים יש את השורות של הבדיקות עצמם (הנתונים).

### 1. הצהרת הגרסה של הקובץ

```
##fileformat=VCFv4.1
##INFO=ID=SOMATIC,Number=0,Type=Flag,Description="Variant is Somatic">
##INFO=ID=MUT_TYPE,Number=1,Type=String,Description="Mutation Type">
##INFO=ID=GENE,Number=1,Type=String,Description="Vagrent annotated gene">
##INFO=ID=PROTEIN_CHANGE,Number=1,Type=String,Description="Vagrent annotated protein change">
##INFO=ID=VC,Number=1,Type=String,Description="Vagrent annotated most deleterious variant consequence">
##INFO=ID=IS_IN_EXCLUDED_REGION,Number=1,Type=Integer,Description="In CNA/LOH or sex chromosome region: 0=Not Excluded,1=Excluded">
##FORMAT=ID=GENOTYPE,Number=1,Type=String,Description="Genotype:0=absent,1=present,?=unknown">
##FORMAT=ID=REF_COUNT,Number=1,Type=Integer,Description="Reads presenting a reference allele for this position">
##FORMAT=ID=ALT_COUNT,Number=1,Type=Integer,Description="Reads presenting the specified alternative allele(s) for this position">
```

### 2. פורמט המידע של הרצה של בדיקת קיום מוטציה על קולונות

ג. סוג הוריאנט: somatic או germline

```
##fileformat=VCFv4.1
##INFO=ID=SOMATIC,Number=0,Type=Flag,Description="Variant is Somatic">
##INFO=ID=MUT_TYPE,Number=1,Type=String,Description="Mutation Type">
##INFO=ID=GENE,Number=1,Type=String,Description="Vagrent annotated gene">
##INFO=ID=PROTEIN_CHANGE,Number=1,Type=String,Description="Vagrent annotated protein change">
##INFO=ID=VC,Number=1,Type=String,Description="Vagrent annotated most deleterious variant consequence">
##INFO=ID=IS_IN_EXCLUDED_REGION,Number=1,Type=Integer,Description="In CNA/LOH or sex chromosome region: 0=Not Excluded,1=Excluded">
##FORMAT=ID=GENOTYPE,Number=1,Type=String,Description="Genotype:0=absent,1=present,?=unknown">
##FORMAT=ID=REF_COUNT,Number=1,Type=Integer,Description="Reads presenting a reference allele for this position">
##FORMAT=ID=ALT_COUNT,Number=1,Type=Integer,Description="Reads presenting the specified alternative allele(s) for this position">
```

### ב. סוג המוטציה:

```
##fileformat=VCFv4.1
##INFO=ID=SOMATIC,Number=0,Type=Flag,Description="Variant is Somatic">
##INFO=ID=MUT_TYPE,Number=1,Type=String,Description="Mutation Type">
##INFO=ID=GENE,Number=1,Type=String,Description="Vagrent annotated gene">
##INFO=ID=PROTEIN_CHANGE,Number=1,Type=String,Description="Vagrent annotated protein change">
##INFO=ID=VC,Number=1,Type=String,Description="Vagrent annotated most deleterious variant consequence">
##INFO=ID=IS_IN_EXCLUDED_REGION,Number=1,Type=Integer,Description="In CNA/LOH or sex chromosome region: 0=Not Excluded,1=Excluded">
##FORMAT=ID=GENOTYPE,Number=1,Type=String,Description="Genotype:0=absent,1=present,?=unknown">
##FORMAT=ID=REF_COUNT,Number=1,Type=Integer,Description="Reads presenting a reference allele for this position">
##FORMAT=ID=ALT_COUNT,Number=1,Type=Integer,Description="Reads presenting the specified alternative allele(s) for this position">
```

### i. החלפת בסיס בודד – SNV/SNP

### ii. הוספה ומחיקה – INDEL

### ג. הגן שנבחן בהרצה זו

```
##fileformat=VCFv4.1
##INFO=ID=SOMATIC,Number=0,Type=Flag,Description="Variant is Somatic">
##INFO=ID=MUT_TYPE,Number=1,Type=String,Description="Mutation Type">
##INFO=ID=GENE,Number=1,Type=String,Description="Vagrent annotated gene">
##INFO=ID=PROTEIN_CHANGE,Number=1,Type=String,Description="Vagrent annotated protein change">
##INFO=ID=VC,Number=1,Type=String,Description="Vagrent annotated most deleterious variant consequence">
##INFO=ID=IS_IN_EXCLUDED_REGION,Number=1,Type=Integer,Description="In CNA/LOH or sex chromosome region: 0=Not Excluded,1=Excluded">
##FORMAT=ID=GENOTYPE,Number=1,Type=String,Description="Genotype:0=absent,1=present,?=unknown">
##FORMAT=ID=REF_COUNT,Number=1,Type=Integer,Description="Reads presenting a reference allele for this position">
##FORMAT=ID=ALT_COUNT,Number=1,Type=Integer,Description="Reads presenting the specified alternative allele(s) for this position">
```

### ד. לכל מוטציה, החלובן שבו נעשה המוטציה, השינוי שנעשה ואיפה בדיוק בחלובן זה התרחש

```
##fileformat=VCFv4.1
##INFO=<ID=SOMATIC,Number=0,Type=Flag,Description="Variant is Somatic">
##INFO=<ID=MUT_TYPE,Number=1,Type=String,Description="Mutation Type">
##INFO=<ID=GENE,Number=1,Type=String,Description="Vagrent annotated gene">
##INFO=<ID=PROTEIN_CHANGE,Number=1,Type=String,Description="Vagrent annotated protein change">
##INFO=<ID=VC,Number=1,Type=String,Description="Vagrent annotated most deleterious variant consequence">
##INFO=<ID=IS_IN_EXCLUDED_REGION,Number=1,Type=Integer,Description="In CNA/LOH or sex chromosome region: 0=Not Excluded,1=Excluded">
##FORMAT=<ID=GENOTYPE,Number=1,Type=String,Description="Genotype:0=absent,1=present,?=unknown">
##FORMAT=<ID=REF_COUNT,Number=1,Type=Integer,Description="Reads presenting a reference allele for this position">
##FORMAT=<ID=ALT_COUNT,Number=1,Type=Integer,Description="Reads presenting the specified alternative allele(s) for this position">
```

ה. לכל מוטציה ההשלכות שקרו בגללה, למשל missense, nonsense, 3 prime utr ועוד

```
##fileformat=VCFv4.1
##INFO=<ID=SOMATIC,Number=0,Type=Flag,Description="Variant is Somatic">
##INFO=<ID=MUT_TYPE,Number=1,Type=String,Description="Mutation Type">
##INFO=<ID=GENE,Number=1,Type=String,Description="Vagrent annotated gene">
##INFO=<ID=PROTEIN_CHANGE,Number=1,Type=String,Description="Vagrent annotated protein change">
##INFO=<ID=VC,Number=1,Type=String,Description="Vagrent annotated most deleterious variant consequence">
##INFO=<ID=IS_IN_EXCLUDED_REGION,Number=1,Type=Integer,Description="In CNA/LOH or sex chromosome region: 0=Not Excluded,1=Excluded">
##FORMAT=<ID=GENOTYPE,Number=1,Type=String,Description="Genotype:0=absent,1=present,?=unknown">
##FORMAT=<ID=REF_COUNT,Number=1,Type=Integer,Description="Reads presenting a reference allele for this position">
##FORMAT=<ID=ALT_COUNT,Number=1,Type=Integer,Description="Reads presenting the specified alternative allele(s) for this position">
```

ו. גן CNA/LOH או גן בכרומוזום המין נכתוב 1 אחרת נכתוב 0 (ש-0 מסמן לכלול אותו באנליזות מסוימות ו-1 מסמן לא לכלול אותו)

```
##fileformat=VCFv4.1
##INFO=<ID=SOMATIC,Number=0,Type=Flag,Description="Variant is Somatic">
##INFO=<ID=MUT_TYPE,Number=1,Type=String,Description="Mutation Type">
##INFO=<ID=GENE,Number=1,Type=String,Description="Vagrent annotated gene">
##INFO=<ID=PROTEIN_CHANGE,Number=1,Type=String,Description="Vagrent annotated protein change">
##INFO=<ID=VC,Number=1,Type=String,Description="Vagrent annotated most deleterious variant consequence">
##INFO=<ID=IS_IN_EXCLUDED_REGION,Number=1,Type=Integer,Description="In CNA/LOH or sex chromosome region: 0=Not Excluded,1=Excluded">
##FORMAT=<ID=GENOTYPE,Number=1,Type=String,Description="Genotype:0=absent,1=present,?=unknown">
##FORMAT=<ID=REF_COUNT,Number=1,Type=Integer,Description="Reads presenting a reference allele for this position">
##FORMAT=<ID=ALT_COUNT,Number=1,Type=Integer,Description="Reads presenting the specified alternative allele(s) for this position">
```

3. הפורמט (מיוצג כשלישיה) של הצגת הנתונים עבור בדיקה על קולונה. (כל שלישיה זו קולונה).

א. האם המוטציה מופיעה (1 אם כן 0 אם לא) : מספר הרידים שהייתה בהם מוטציה

: מספר הרידים שלא היה בהם מוטציה

```
##FORMAT=<ID=GENOTYPE,Number=1,Type=String,Description="Genotype:0=absent,1=present,?=unknown">
##FORMAT=<ID=REF_COUNT,Number=1,Type=Integer,Description="Reads presenting a reference allele for this position">
##FORMAT=<ID=ALT_COUNT,Number=1,Type=Integer,Description="Reads presenting the specified alternative allele(s) for this position">
```

#### 4. הכרומוזומים והאורכים שלהם:

```
##contig=<ID=1,assembly=NCBI37,length=249250621,species=Human>  
##contig=<ID=2,assembly=NCBI37,length=243199373,species=Human>  
##contig=<ID=3,assembly=NCBI37,length=198022430,species=Human>  
##contig=<ID=4,assembly=NCBI37,length=191154276,species=Human>  
##contig=<ID=5,assembly=NCBI37,length=180915260,species=Human>  
##contig=<ID=6,assembly=NCBI37,length=171115067,species=Human>  
##contig=<ID=7,assembly=NCBI37,length=159138663,species=Human>  
##contig=<ID=8,assembly=NCBI37,length=146364022,species=Human>  
##contig=<ID=9,assembly=NCBI37,length=141213431,species=Human>  
##contig=<ID=10,assembly=NCBI37,length=135534747,species=Human>  
##contig=<ID=11,assembly=NCBI37,length=135006516,species=Human>  
##contig=<ID=12,assembly=NCBI37,length=133851895,species=Human>  
##contig=<ID=13,assembly=NCBI37,length=115169878,species=Human>  
##contig=<ID=14,assembly=NCBI37,length=107349540,species=Human>  
##contig=<ID=15,assembly=NCBI37,length=102531392,species=Human>  
##contig=<ID=16,assembly=NCBI37,length=90354753,species=Human>  
##contig=<ID=17,assembly=NCBI37,length=81195210,species=Human>  
##contig=<ID=18,assembly=NCBI37,length=78077248,species=Human>  
##contig=<ID=19,assembly=NCBI37,length=59128983,species=Human>  
##contig=<ID=20,assembly=NCBI37,length=63025520,species=Human>  
##contig=<ID=21,assembly=NCBI37,length=48129895,species=Human>  
##contig=<ID=22,assembly=NCBI37,length=51304566,species=Human>  
##contig=<ID=X,assembly=NCBI37,length=155270560,species=Human>  
##contig=<ID=Y,assembly=NCBI37,length=59373566,species=Human>
```



```
##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=##SAMPLE=
```

## 6. הטמפלייט של הרצה על מוטציה:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT			
	PD4781c_lo0005	PD4781c_lo0006	PD4781c_lo0007	PD4781c_lo0008	PD4781c_lo0009	PD4781c_lo0010	PD4781c_lo0011	PD4781c_lo0012	PD4781c_lo0013	PD4781c_lo0014	PD4781c_lo0015
	PD4781c_lo0013	PD4781c_lo0014	PD4781c_lo0015	PD4781c_lo0016	PD4781c_lo0017	PD4781c_lo0018	PD4781c_lo0019	PD4781c_lo0020	PD4781c_lo0021	PD4781c_lo0022	PD4781c_lo0023
	PD4781c_lo0020	PD4781c_lo0021	PD4781c_lo0022	PD4781c_lo0023	PD4781c_lo0024	PD4781c_lo0025	PD4781c_lo0026	PD4781c_lo0027	PD4781c_lo0028	PD4781c_lo0029	PD4781c_lo0030
	PD4781c_lo0027	PD4781c_lo0028	PD4781c_lo0029	PD4781c_lo0030	PD4781c_lo0031	PD4781c_lo0032	PD4781c_lo0033	PD4781c_lo0034	PD4781c_lo0035	PD4781c_lo0036	PD4781c_lo0037
	PD4781c_lo0035	PD4781c_lo0036	PD4781c_lo0037	PD4781c_lo0038	PD4781c_lo0039	PD4781c_lo0040	PD4781c_lo0041	PD4781c_lo0042	PD4781c_lo0043	PD4781c_lo0044	PD4781c_lo0045
	PD4781c_lo0044	PD4781d_lo0001	PD4781d_lo0002	PD4781d_lo0003	PD4781d_lo0005	PD4781d_lo0006	PD4781d_lo0007	PD4781d_lo0008	PD4781d_lo0009	PD4781d_lo0010	PD4781d_lo0011
	PD4781d_lo0009	PD4781d_lo0010	PD4781d_lo0011	PD4781d_lo0012	PD4781d_lo0013	PD4781d_lo0014	PD4781d_lo0015	PD4781d_lo0016	PD4781d_lo0017	PD4781d_lo0018	PD4781d_lo0019
	PD4781d_lo0020	PD4781d_lo0021	PD4781d_lo0022	PD4781d_lo0023	PD4781d_lo0024	PD4781d_lo0025	PD4781d_lo0026	PD4781d_lo0027	PD4781d_lo0028	PD4781d_lo0029	PD4781d_lo0030
	PD4781d_lo0031	PD4781d_lo0032									

- א. מספר הכרומוזום שעליו נמצאת המוטציה
- ב. המיקום של הבסיס שבו מתחילה המוטציה
- ג. הזהות של הבן אדם שעליו אנחנו מריצים את זה
- ד. מה שאמור להיות בלי מוטציה
- ה. מה שאמור להיות אחרי מוטציה
- ו. עד כמה אנחנו בטוחים בתוצאה
- ז. לא בטוח מה זה ה FILTER
- ח. זה החלק של המידע של פורמט המידע שהסברתי בהרחבה ב2
- ט. זה פורמט ההצגה של קולונה שהסברתי בהרחבה ב3
- י. אלו השלשות שמייצגות את הקולונות שכתובות ב5