

אלגוריתמים בביו' חישובית

76558

לקראת ההאקדון

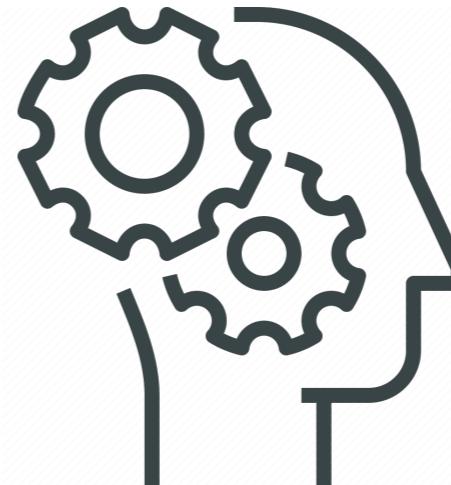
תומי קפלן

12/1/2025

אבני הבסיס לפרויקט מחקר



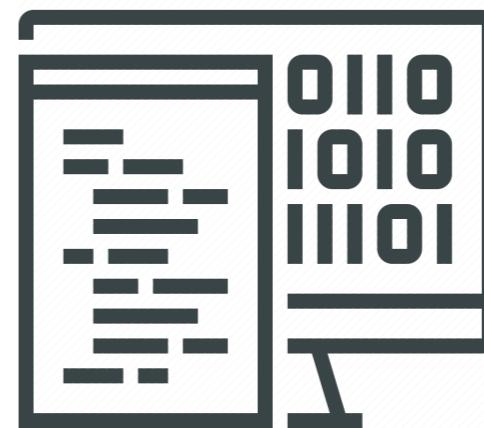
נתונים



שאלה מחקרית



ניתוח סטטיסטי



מודל + למידה

תרגיל 4 - הצעה מחר

דעתה:

גנום של דגי זברה JCSC

מודל:

שימוש בHOMER למציאת

1. שימוש ב-HOMER וseqs-clip של החלבוניות מהמאמר ולמציאת המוטיפים של כל אחד מהם. גנון חלבון. הוא חורג מבחן הסתברותית מההתפלגות הרנדומית של הגנום (0.25 לכל בסיס),

וניתן ליציג מוטיף באמצעות הסתברויות של אות להופיע בעמלה מסוימת, כאשר גודל האות בגרף

פרופורציונלי להסתברות שלה להופיע במוטיף:



לדוגמה במוטיף הנ"ל שתי האותיות הראשונות הן בסבירות מאוד גבוהה G (בodoreות כמעט 1), וקיים סיכוי כמעט זהה ל-G באות

REFERENCES:

g Chan, Mina L. Kojima, Mark E. dscape of Pioneer Factor Activity Reveals nd Genome Activation." *Molecular Cell*.

מיזו לין טין צווען זי

שאלת מחקר:

האם קיים קשר בין אן בו הוא נקשר, לקשירה של שני החלבונים

כלומר נרצה לבדוק האם קיים קשר בין איפונו המוטיף של Nanog באיזוריים שונים בגנום, לבין Pouf, Sox (באותם האיזוריים).

ההשערות שלנו:

H0 - אין הבדל באופי

בין שלושת החלבוניות Nanog, Pouf, Sox

טיפח חזק יותר שלו (ודאות יותר גבוהה
האחרים נקשרים גם הם, נראה מוטיף

חלש יותר שלו.

למה לצפות

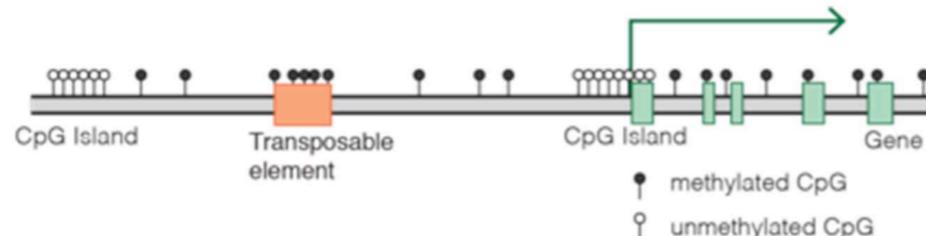
- האקתון, 40% מהציוון בקורס, בקבוצות (3-5).
- מועד: 27/1
- מי שלא יצליח: להגיש פרויקט מחקרי בהיחף מחייב
- תרגיל 4 – מציאת נושא והגשת הצעת מחקר

רעיונות

- כל שאלה שקשורה לבiology או רפואה
- DATA על רצפים או סדרות
- למידה (בולל למידה عمוקה)
- דוגמאות:

השפעת המרחק הגנומי על דפוסי המתילציה בין אתרים CpG עוקבים

Typical mammalian DNA methylation landscape



איור 2: מתילציה אופיינית של DNA ביונקים

dist	0->0	0->1	1->0	1->1
2	749	972	761	3660
13	346	1200	627	4121
5	257	848	671	5451
4	313	635	1227	5244
4	312	1252	533	5468
28	83	741	374	6055

טבלה 1: המדיע בבסיס המודל שבנו, נתוניים אמפיריים על מתילציה של אתרים CpG עוקבים

בדומה לנעשה במודלים האבולוציוניים, אנו מניחים מרקוביות, כלומר, אם אתר מיי-הוקב לו שנמצא במרקח d מمنו, יהיה במרקח M תלואה רק בז' ו- p .

אם כך, עבור מתילציה בגנים, מטריצת הקצב היא מהצורה:

$$R = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}$$

המטריצה היא בהכרח מצורה זו שכן נדרשות לה שתי דרגות חופש. לא נוכל להניח מטריצת קצב, לא יתכן יותר משתי דרגות חופש, שכן הסכום של כל שורה בהכרח?

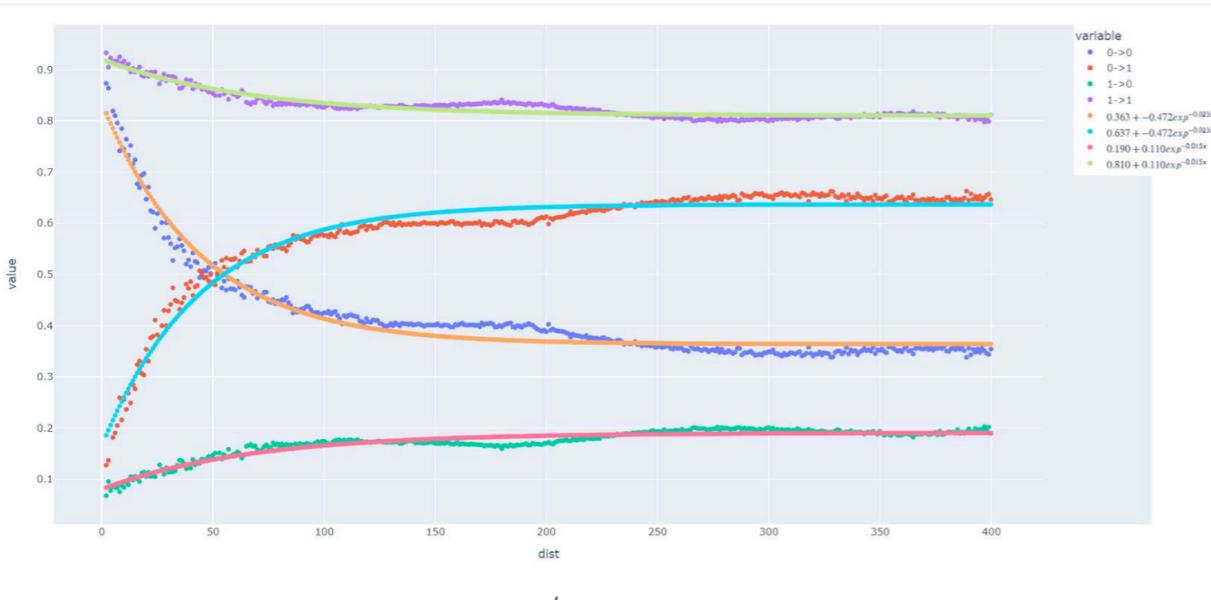
שיטת הנאייה נוספת לכך היא מודל Jukes-Cantor [4]. המודל מניח תדרי בסיס שוים ושיעורי מוטציות שוים. באמצעות הנחה זו, המודל מחשב מרך אבולוציוני על סמך חלוקם של הבדלי נוקלאוטידים בין שני רצפים.

עבור מטריצת קצב מהצורה:

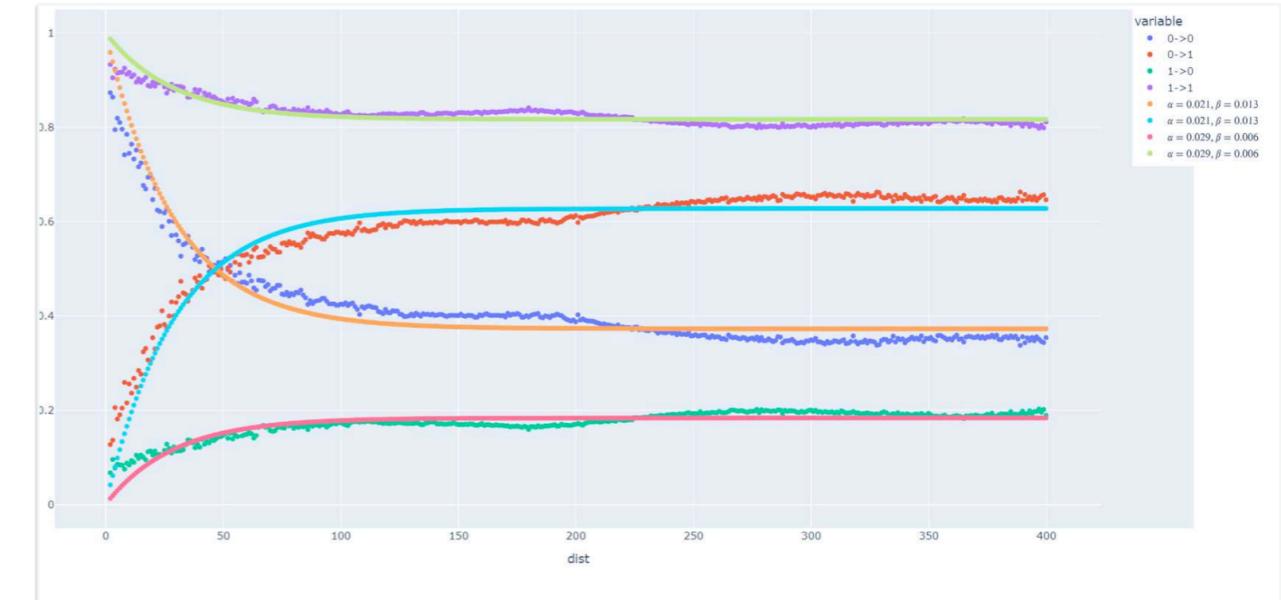
$$R = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

נקבל כי ההסתברויות לעבור מהמצב ה- i למצב ה- j כאשר אורך הענף הוא v , נתון ע"י:

$$P_{ij}(v) = \begin{cases} \frac{1}{4} + \frac{3}{4} e^{-\frac{4v}{3}} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4} e^{-\frac{4v}{3}} & \text{if } i \neq j \end{cases}$$



איור 3: התאמה עבור מטריצת המעלרים שהשכנו מתוך מטריצת הקצב



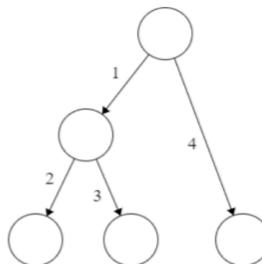
איור 4: התאמה עבור מטריצת המעלרים שהשכנו מתוך מטריצת הקצב

השפעת המרחק הגנומי על דפוסי המתילציה בין אתרי CpG עוקבים

- **סוגי תאים שונים?**
- **בריאים לעומת מותחולים?**
- **מודל חבוי שמתאים לאיזוריים שונים בגנו? למשל,
אי CpG לעומת איזוריים אחרים**
- **מה זה יכול ללמד אותנו על הבiologyה?**

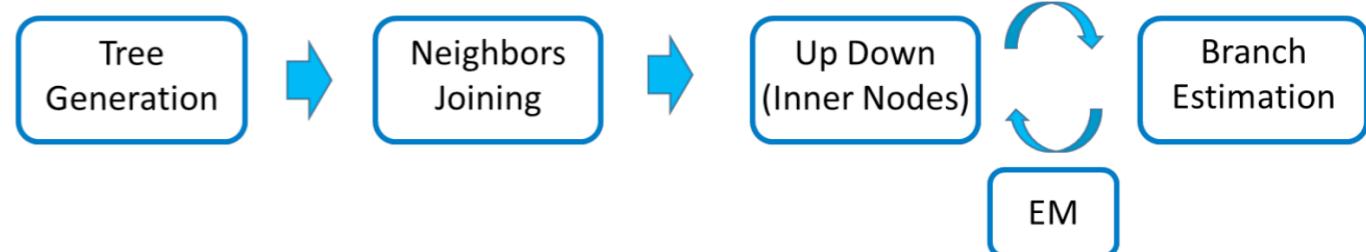
שחזר עצים אבולוציוניים מדטה (סימולציות של רצפים בינארי)

We will start with a small binary tree for ease of implementation and runtime. This can easily be scaled up. The tree we begin with is structured as follows:

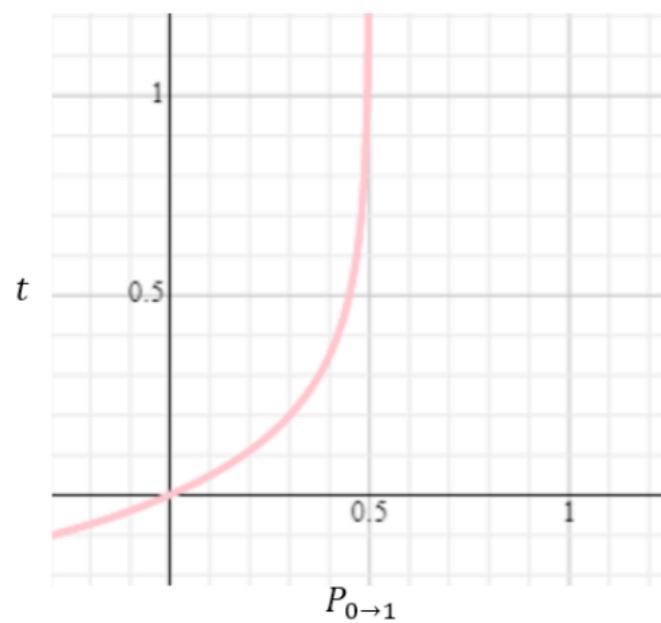


Models and Algorithms

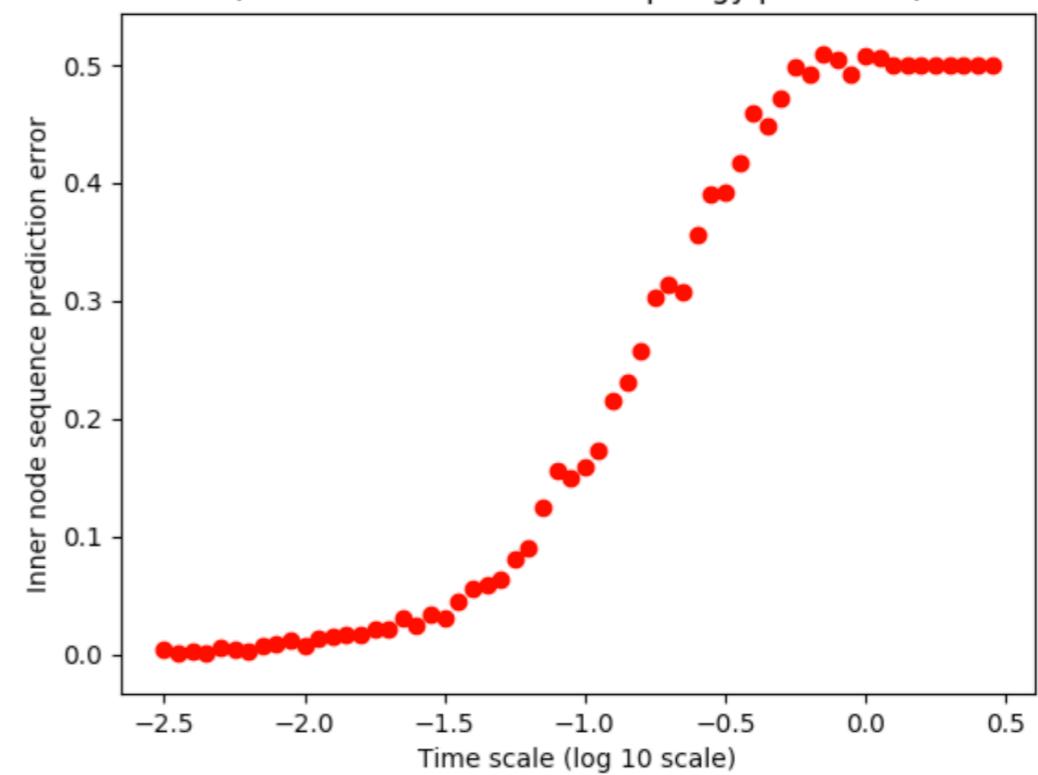
Overview



lengths: Our branch lengths are arbitrarily chosen small integers. This is an arbitrary constraint for ease of implementation, and can be easily changed at a later stage. We will n

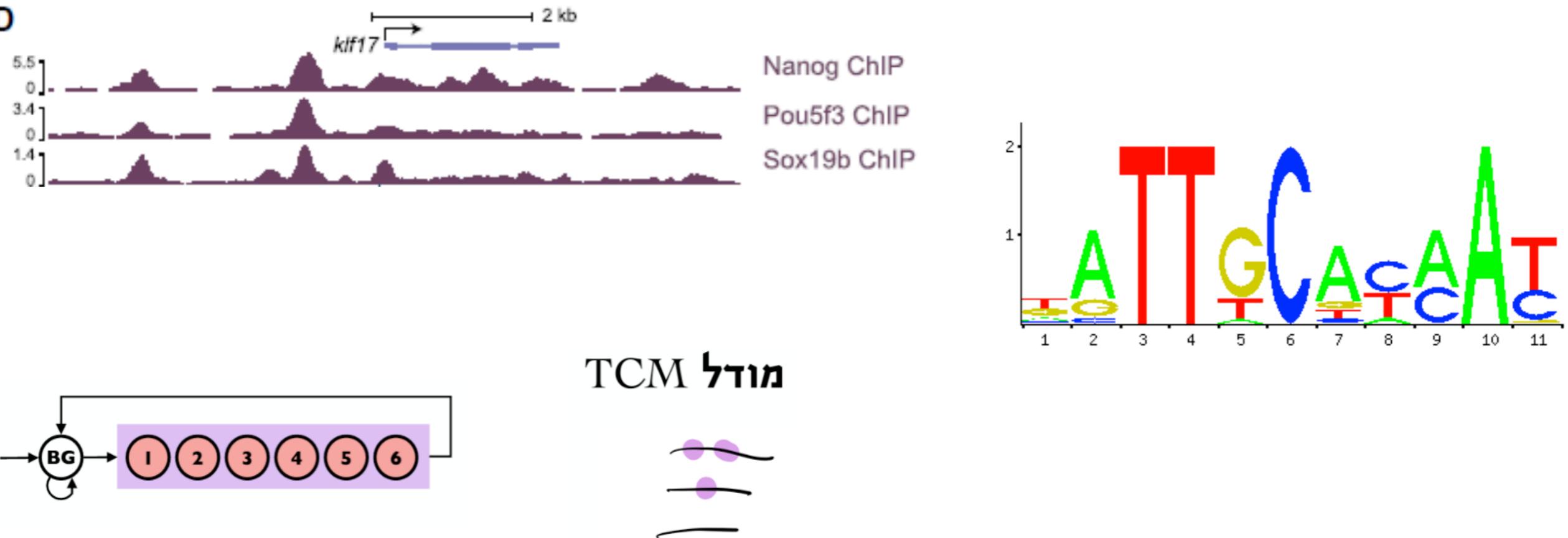


Inner node sequence prediction accuracy vs. Time scales
(for runs with successful topology prediction)



איפיון אתרים קיישור לחלבוניים בדנ"א וקשר לחזק הקישור

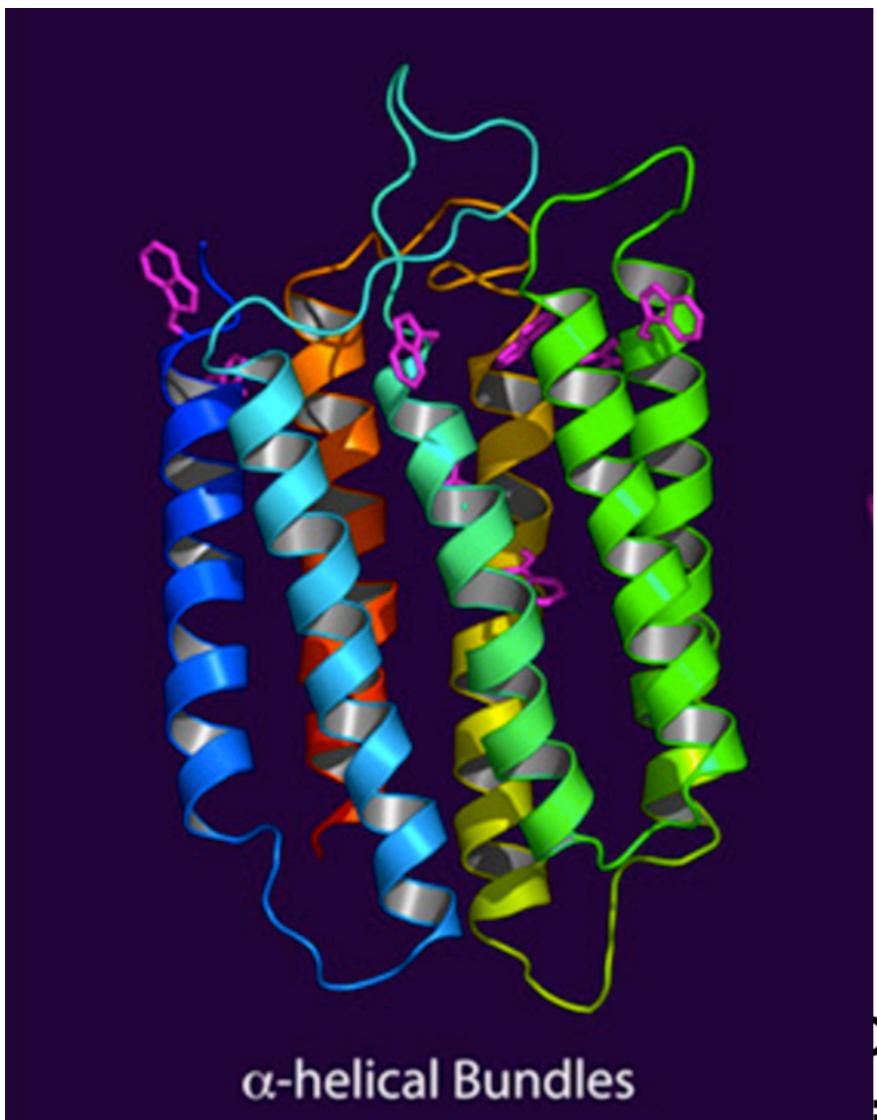
D



השוואה התפלגות ציוני אתר הקישור (לפי צ'יפ-סק)
באייזוריים עם ובלוי המוטיב

זיהוי דומיונים טרנס-מمبرנליים בחלבוניים

Data:



We used the [PDBTM: Protein Data Bank of Transmembrane Proteins](#) as our datasets source.

The [pdbtm_all_seq](#) data set contains ~8500 entries (the data set originally contains ~34000 entries, but only around $\frac{1}{3}$ were not duplicates or copies).

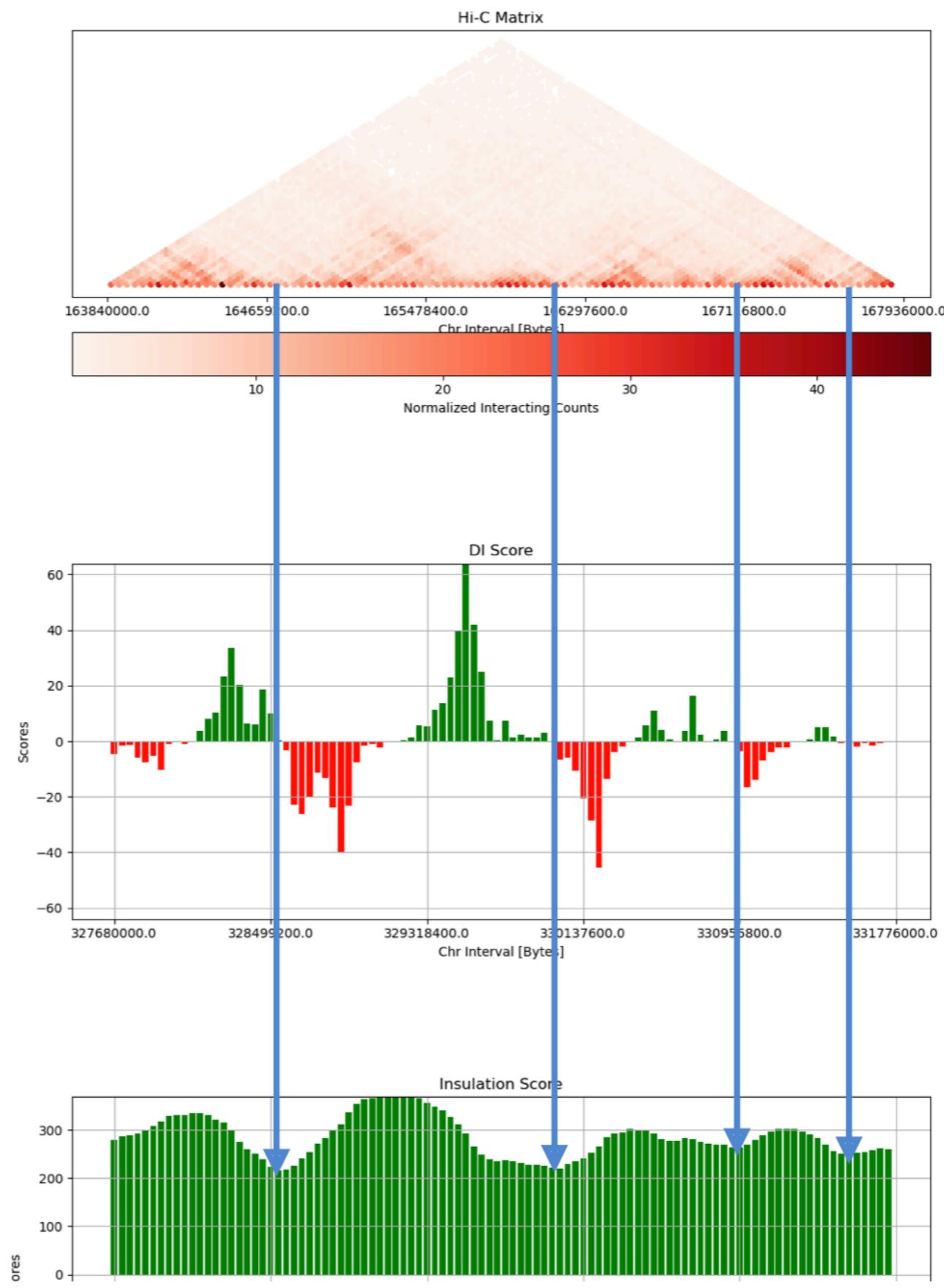
To use the HMM we needed to initialize the start probabilities, transmission matrix, threshold, decoding algorithm and maximum iteration number.

The start probabilities were initialized to an uniform distribution.

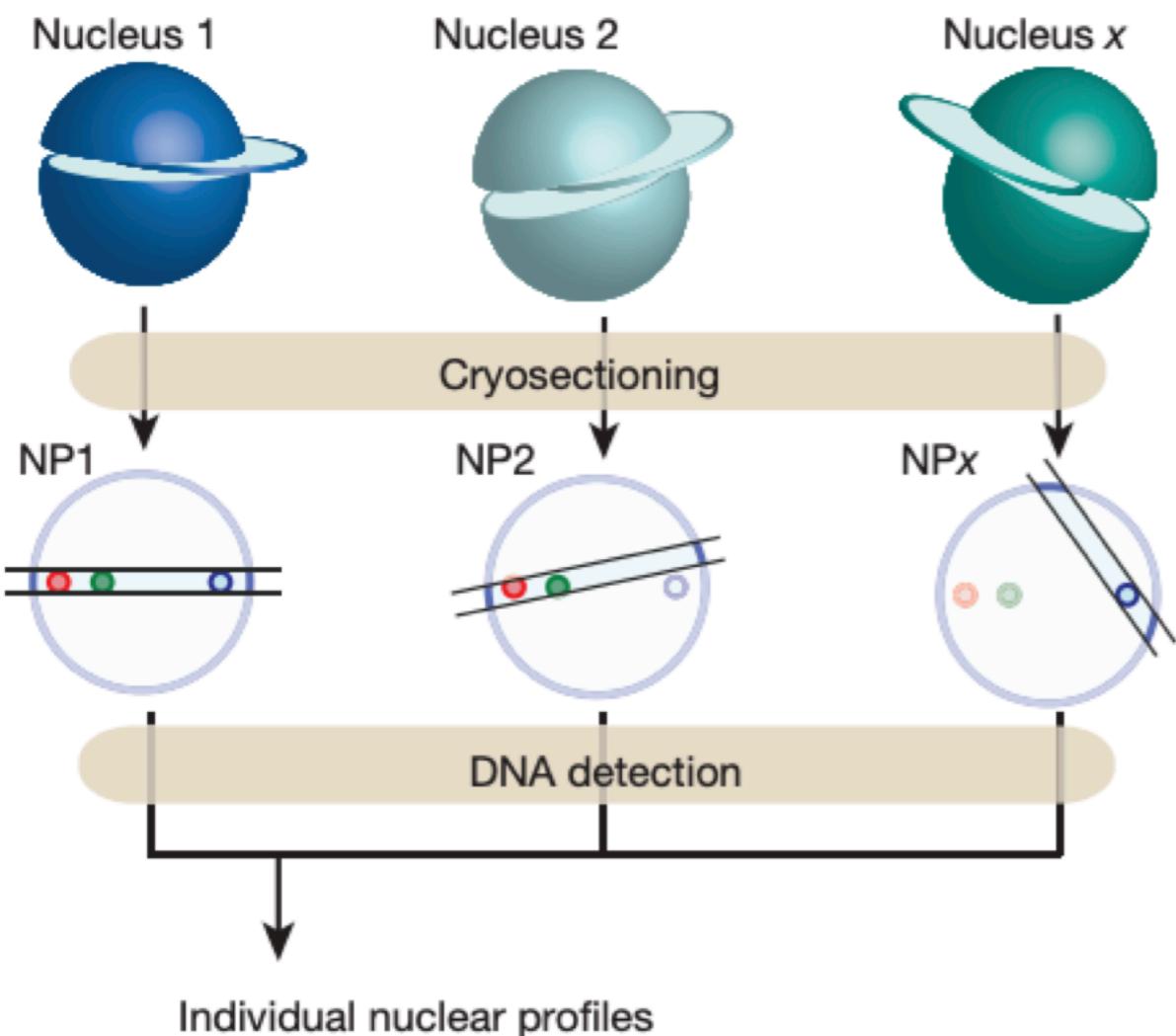
The emissions matrix was initialized to be 2 concatenated identity matrices of size 21x21 forming a 42x21 matrix.

The transition matrix was initialized by using the information from the training data and labels. For each two consecutive amino acids (represented by their one-letter codes) the probability of transitioning from one state to another was calculated.

זיהוי דומיננס תלת מימדיים (טאים) בקייפול הגנים מדאטה Hi-C



זיהוי דומיננס תלת מימדיים (טאדים) מריצוף פרוסות מגרעין התא



	NP1	NP2	NP3	NP4	...	NPx
A	+	+	+	-	...	-
B	+	-	-	+	...	+
C	+	+	-	-	...	-

Computational analysis

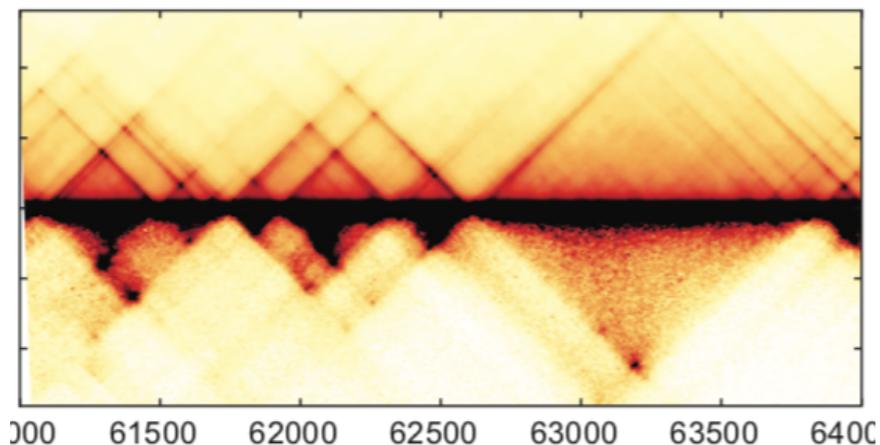
האם עישון בהריון משפיע על בריאות התינוק



עוד רעיונות

- שחרור עז פילוגנטי של תאי מערכת החיסון
 - <https://www.nature.com/articles/s41586-021-03548-6>
 - <https://www.nature.com/articles/s41586-021-04312-6>

Simulated



Real Hi-C

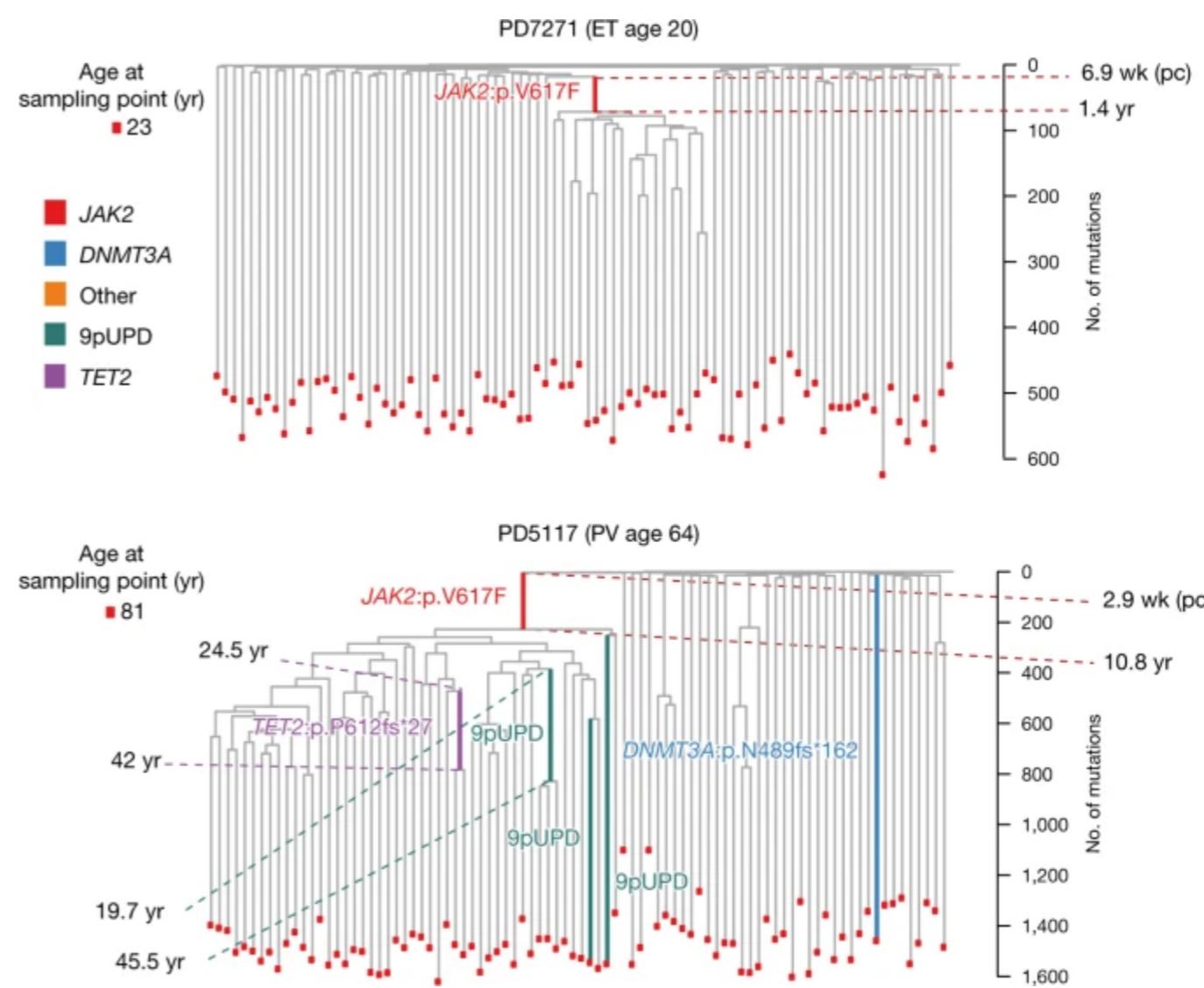
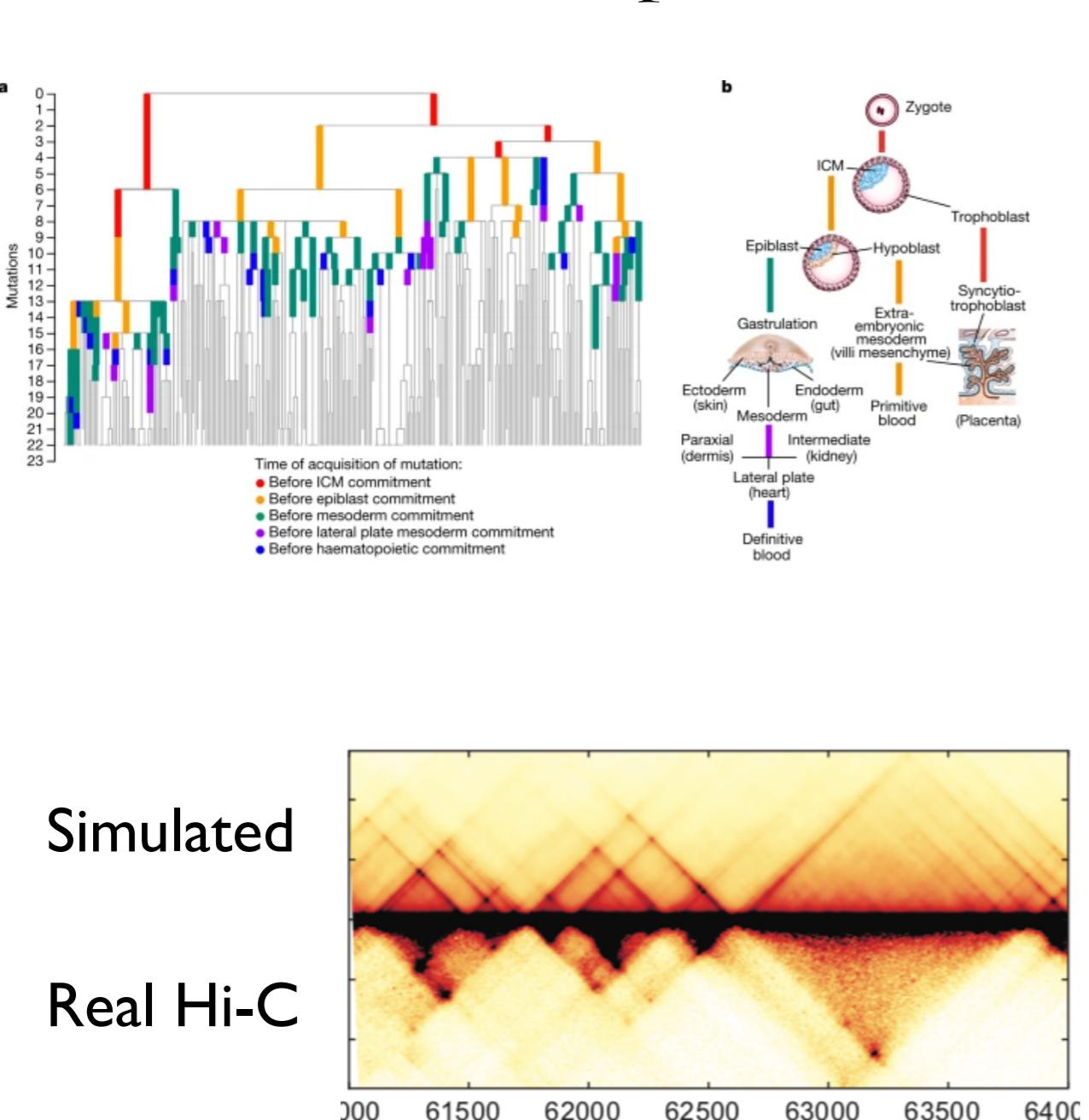
- סימולציות בקייפול הגנים



עוד רעיונות

- שחרור עז פילוגנטי של תאי מערכת החיסון / העיבול

<https://www.nature.com/articles/s41586-021-03548-6>
<https://www.nature.com/articles/s41586-021-04312-6>



- סימולציות בקייפול הגנים

מודל מרקובי מסדר גבוה

- מודל מרקובי מסדר משתנה

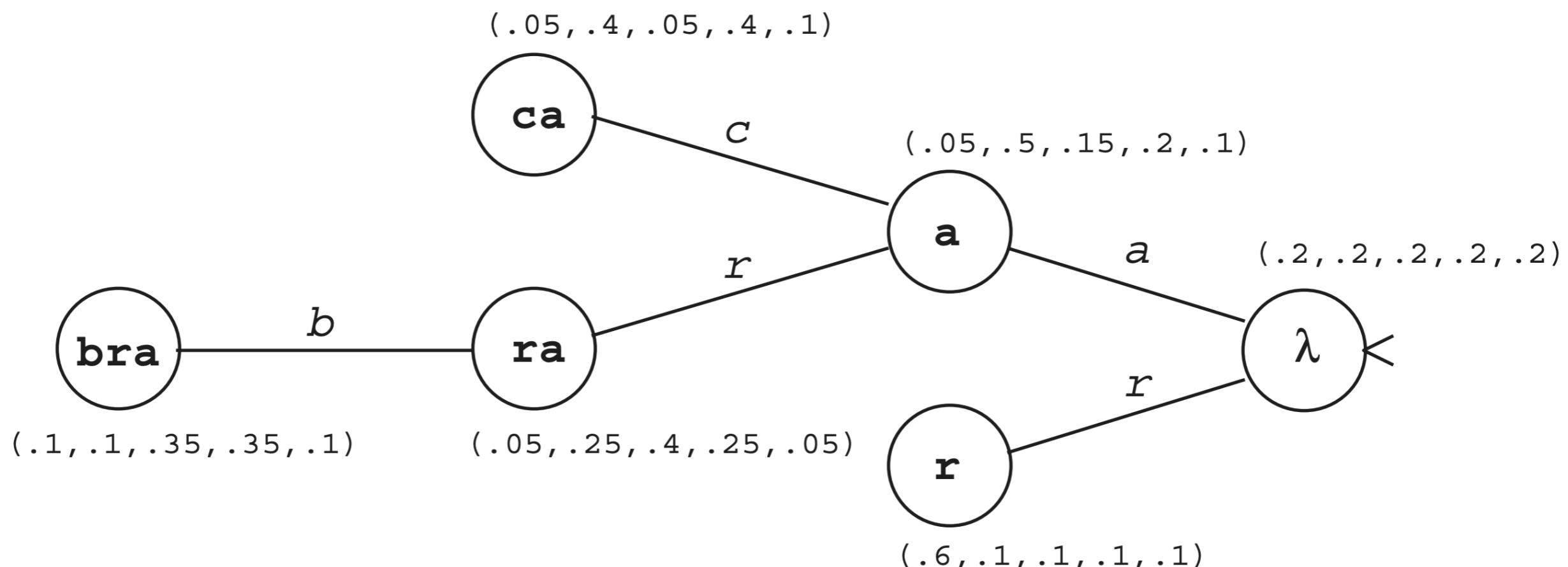
BIOINFORMATICS

Vol. 17 no. 10 2001
Pages 927–934



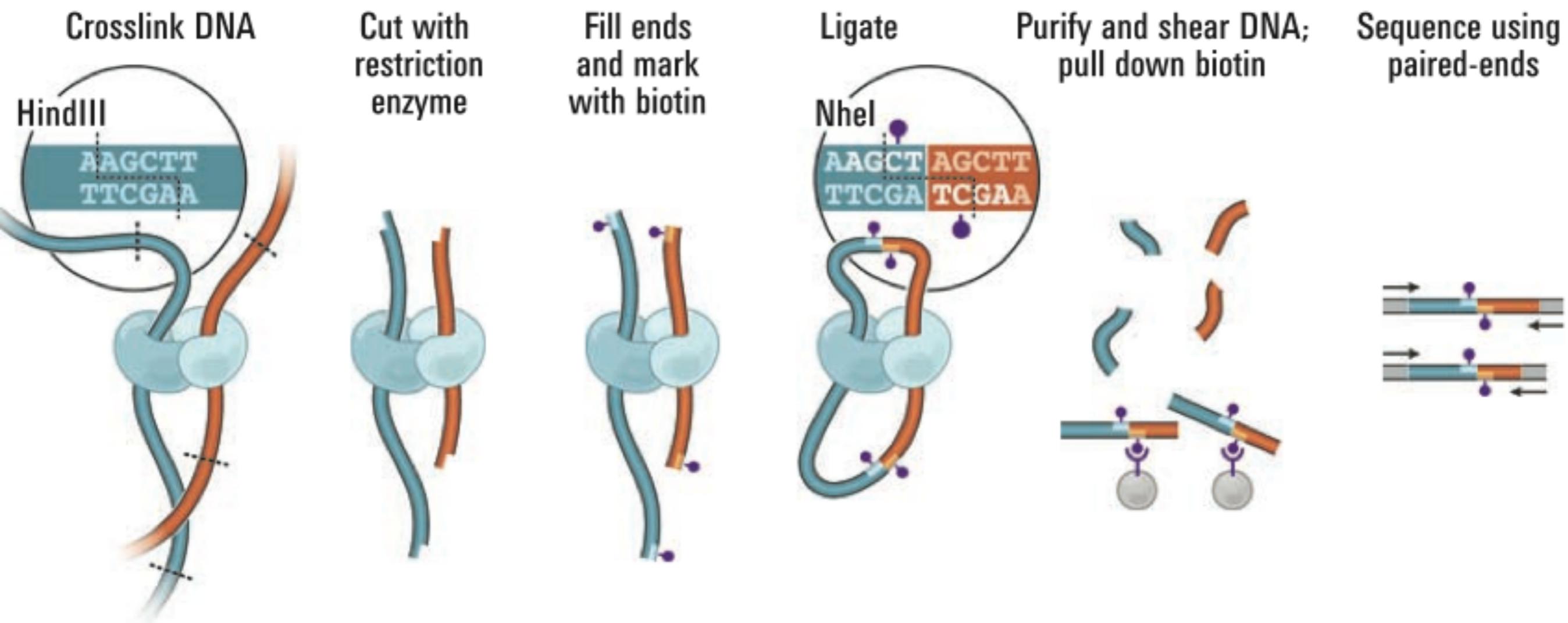
Markovian domain fingerprinting: statistical segmentation of protein sequences

Gill Bejerano^{1,*}, Yevgeny Seldin¹, Hanah Margalit² and
Naftali Tishby^{1,*}



איך נמדד קירבה בין אתרים בגנום?

- ביולוגיה מולקולרית. נחודה, נדביך, ונרצף. Hi-C.

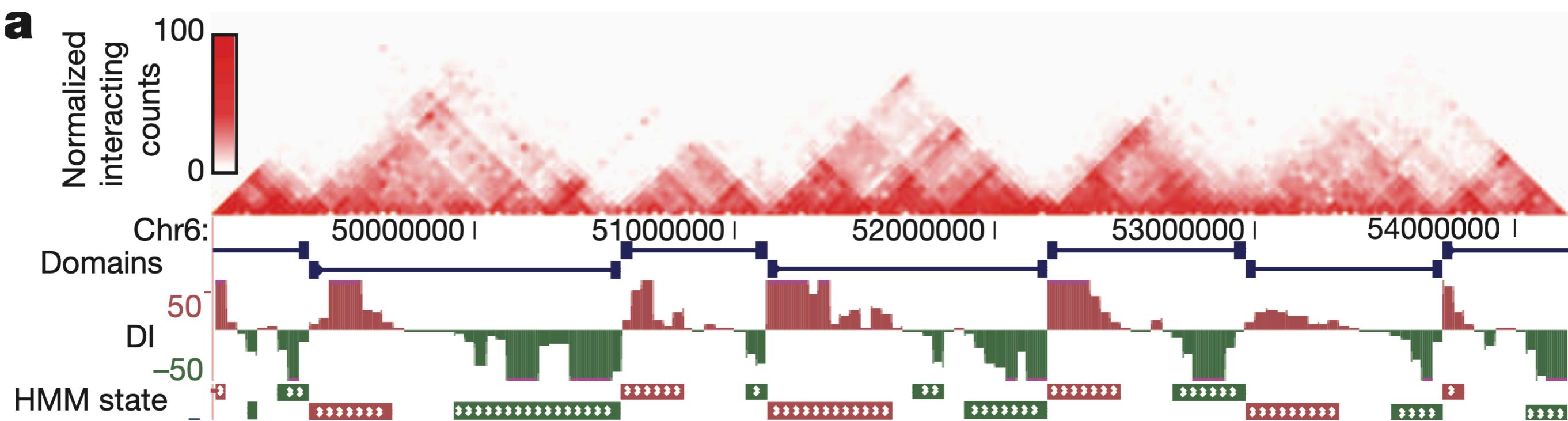


**Comprehensive Mapping of Long-Range
Interactions Reveals Folding Principles
of the Human Genome**

Erez Lieberman-Aiden,^{1,2,3,4*} Nynke L. van Berkum,^{5*} Louise Williams,¹ Maxim Imakaev,² Tobias Ragoczy,^{6,7} Agnes Telling,^{6,7} Ido Amit,¹ Bryan R. Lajoie,⁵ Peter J. Sabo,⁸ Michael O. Dorschner,⁸ Richard Sandstrom,⁸ Bradley Bernstein,^{1,9} M. A. Bender,¹⁰ Mark Groudine,^{6,7} Andreas Gnirke,¹ John Stamatoyannopoulos,⁸ Leonid A. Mirny,^{2,11} Eric S. Lander,^{1,12,13†} Job Dekker^{5†}

איך נמדד קירבה בין אתרים בגנום?

- הגנים מאורגנו בדומיניניס. הקיפול כלל אינו מקרי



Topological domains in mammalian genomes identified by analysis of chromatin interactions

אלגוריתם סטטיסטי לזיהוי דומיניים

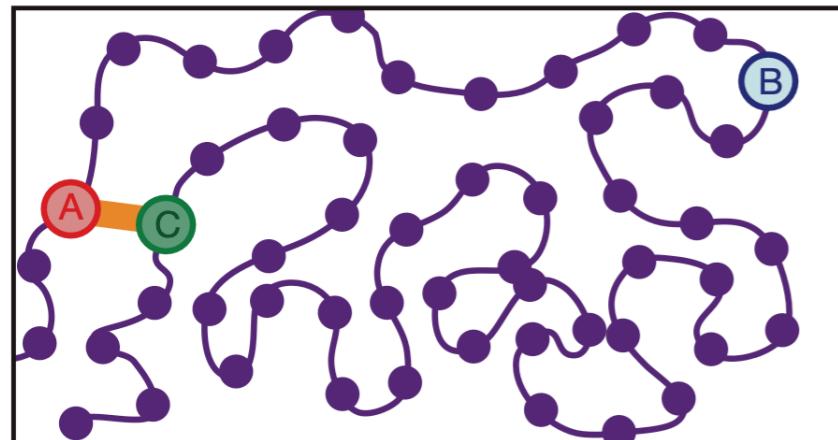
Complex multi-enhancer contacts captured by genome architecture mapping

Robert A. Beagrie^{1,2,3*}, Antonio Scialdone^{4*†}, Markus Schueler¹, Dorothee C. A. Kraemer¹, Mita Chotalia², Sheila Q. Xie^{2†}, Mariano Barbieri^{1,5}, Inês de Santiago^{2†}, Liron-Mark Lavitas^{1,2}, Miguel R. Branco^{2†}, James Fraser⁶, Josée Dostie⁶, Laurence Game⁷, Niall Dillon³, Paul A. W. Edwards⁸, Mario Nicodemi^{4§} & Ana Pombo^{1,2,5,9§}

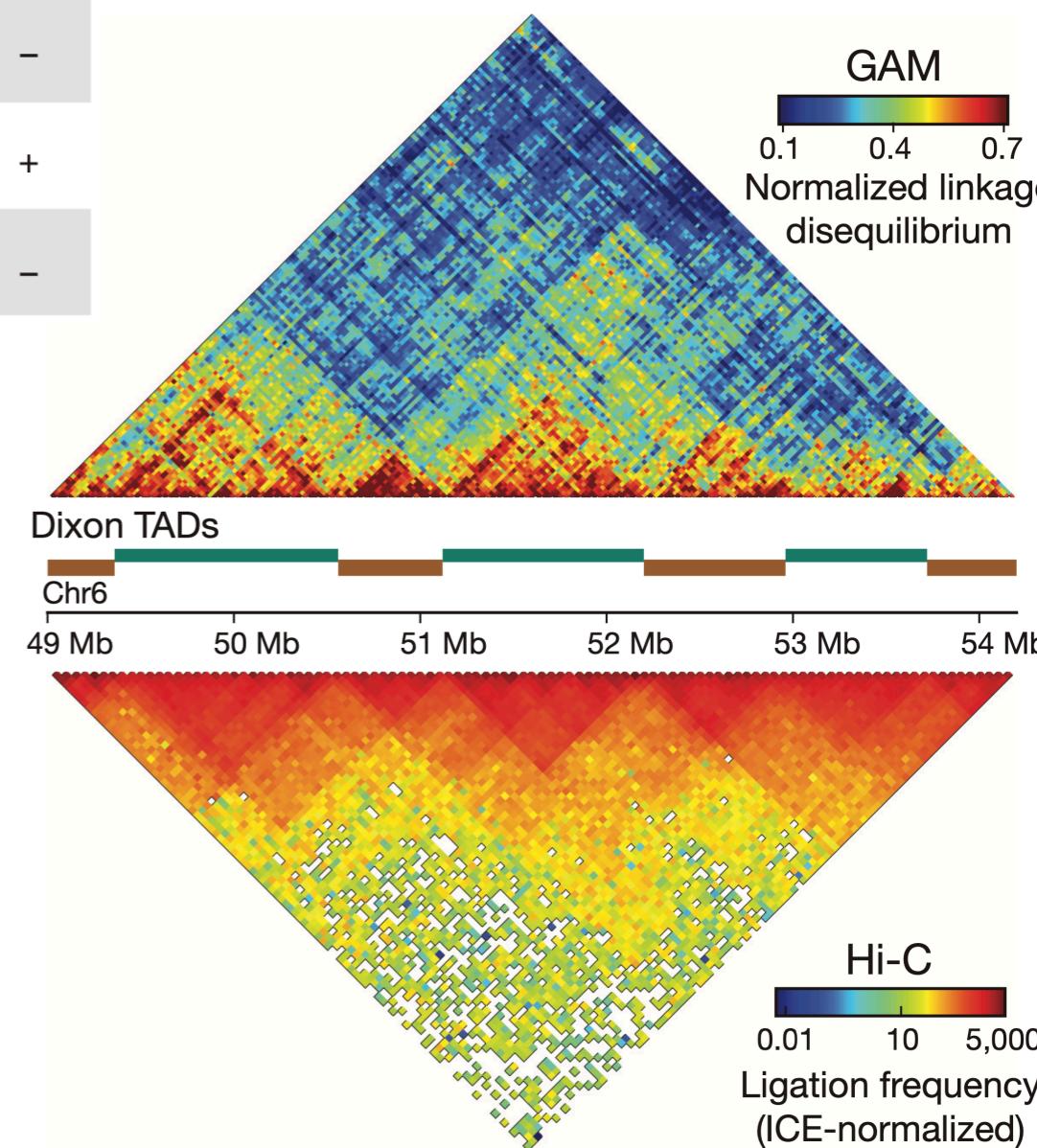
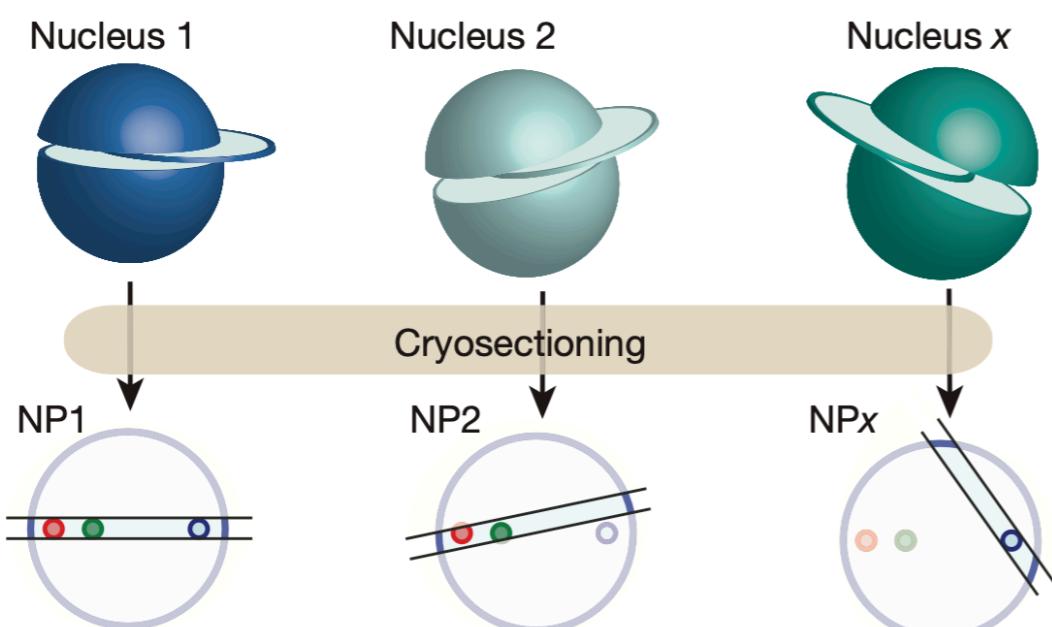
Genomic distance between loci:



Nuclear distance between loci:

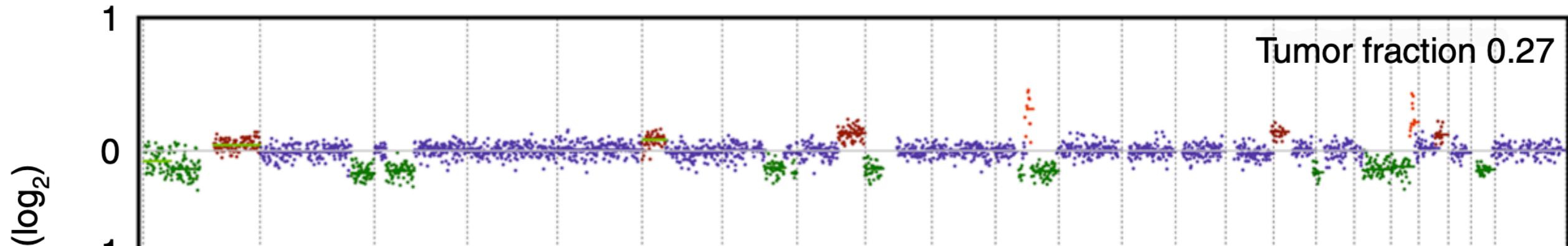


	NP1	NP2	NP3	NP4	...	NPx
A	+	+	+	-	...	-
B	+	-	-	+	...	+
C	+	+	-	-	...	-

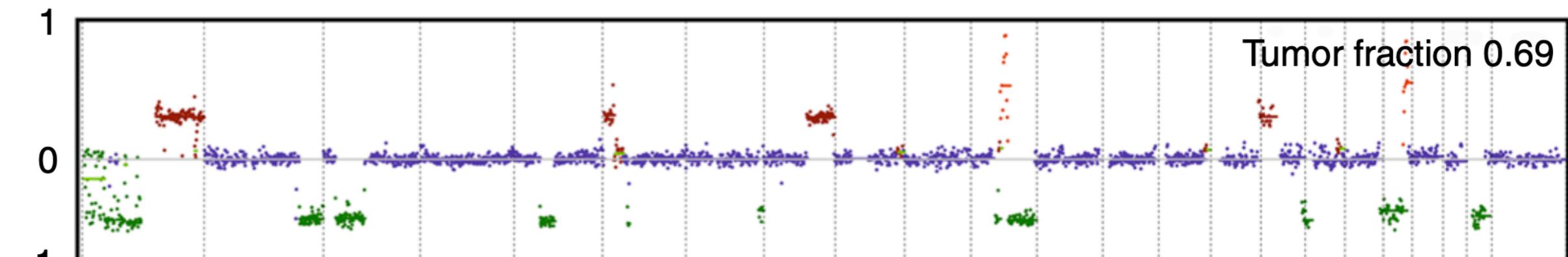


ביופסיה נזולית וسرطان

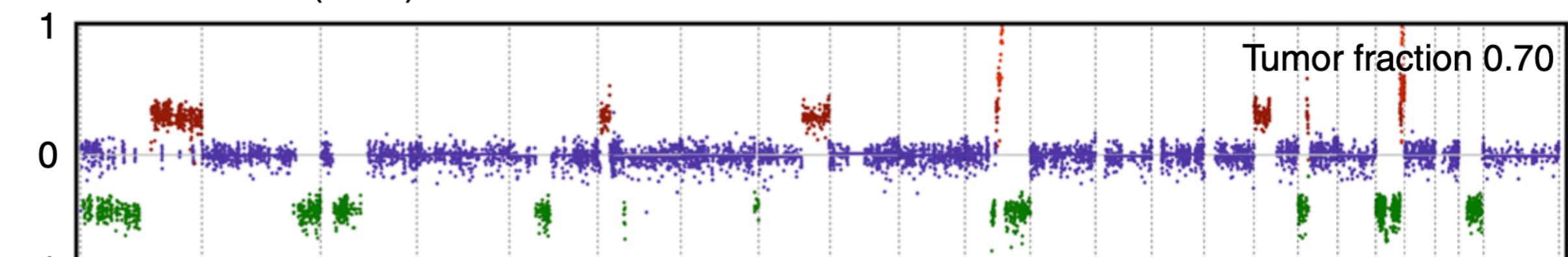
cfDNA - ULP-WGS (0.1 \times)



Tumor - WGS (1 \times)



Tumor - WES (142 \times)



Chromosome

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 X
18 20 22

Gain

Amplification

Copy neutral

Deletion

טנספורט לדzapio דנ'א

>chr2:220282501-220286500

GAGGCTCAGGGTAGCTGCCATAGACATACTGGCAGGCAGGCTTGGCCAGGATCCCTCCGCCTGCCAGGCGTCTCCCTGCCCTCCCTGCCTA
GAGACCCCCACCCTCAAGCCTGGCTGGCTTTGCCTGAGACCCAAACCTCTCGACTCAAGAGAATATTAGGAACAAGGTGGTTAGGGCCTTCCTG
GGAACAGGCCTTGACCCTTAAGAAATGACCCAAAGTCTCTCCTTGACCAAAAAGGGGACCCCTCAAACATAAGGAAGCCTCTTCTGCTGTCCCCT
GACCCCACCTCCCCCACCAGGACGAGGAGATAACCAGGGCTGAAAGAGGCCCTGGGGCTGCAGACATGCTGCTGCCCTGGCGAAGGAT
TGGCAGGCTTGGCCGTACAGGACCCCCGCTGGCTGACTCAGGGCGCAGGCCTCTGCAGGGGAGCTGGCCTCCCGCCCCACGCCACGGCCGCCC
TTCTGGCAGGACAGCAGGGATCTGCAGCTGTCAAGGGAGGGAGGCAGGGCTGATGTCAAGGAGGGATACAAATAGTGCCGACGGCTGGGGCCCTGT
CTCCCCTGCCGCATCCACTCTCCGGCCGCCCTGCCGCCCTCCGTGCCGCCAGCCTGCCCGCCGTACCATGAGCCAGGCCTACT
CGTCCAGCCAGCGCGTGTCCCTACCGCCGCACCTCGGGGGGGGGCTTCCACTCGGCTCCCCGCTGAGTCGCCGTGTTCCGGGGCGGG
TTCTGGCTCTAAGGGCTCCAGCTCGGTGACGTCCCAGGTACCGAGGTGTCGCCACGTGGGGGGGGGGCTGGGTGCTGCCAGGC
CGGCTGGGGACCACCCGACGCCCTCCCTACGGCGCAGGCGAGCTGGACTTCACTGGCCGACGCCGGTGAACCAGGAGTTCTGACCACCGCA
CCAACGAGAAGGTGGAGCTGCAGGAGCTCAATGACCGCTCGCCAACATCGAGAAGGTGCGCTTGGAGCAGCAGAACCGGGCTGCCGA
AGTGAACCGGCTCAAGGGCCCGAGCCGACGCCGAGCTACGGAGGAGCTGCGGGAGCTGCCGCCAGGTGGAGGTGCTCACTAACAG
CGCGCGCGTGCACGTCGAGCGCACAACCTGCTCGACGACCTGCAGGGCTCAAGGCCAAGTGAGGGCCGGACCCAGACTCCTTCTGCCGGC
AGGGCACAGGAGGCTAGGCCTGGGGTCTGGGGTCCCCTGTCAGCACCTGCCTCTCCGGGGGGGGACCCCTCCCTGCCCATGTGGAGAAAGGTC
CTCCACCTGTGTTCAAGGGCCGTGACCTCCAGGTCTCTCCCCCTGCGATCCCCTTGACAGGAGTTCTTGGGGACATAGATCAGGGGTGGA
TATGGGAGAATTAGGGACCCGGTGGACAGCCCCGTTAAAAAGCATTAAAGATGCTGGGGGATATTATGGGGTCAAGGTAGTTGATGGG
AGAGGAAGGGCTGCAGGAGGGCCAGAGGGCAGTGTAGCCAGAGGGAGAAGGGAGGCTGATAGGAGACAGGAAAGCAGGGCAAGGGCCAGACTCCAAG
AACAGCTCTCAGCTCAGCTGTGATGAGGCCCTGGGGAGGTGGGGGGAGGGGGAGCTGGCCCTGGGCCCTGCCGAGACTGTGTTTACAAGGTG
AATGGACAGGCTGGAGAAAAAGGGAGTAGGTGGGGTCACAGCTCTCAGAGAGCTGGGAGGACCTGACTGTAGACTTCACCAGGCTCCAAGAACGAAA
GGGCAGCAAGTGTAGCATATTGTTGGTCCCACCTCTGACAGGCCAAGTGAGCACAGTCACCCCTGCCACCAAGTCATAAAATATTGAGCAGCT
ATATTGGCCAGGCTGGAGCTGGGAACCAGAAACACAGAGGTGGATAAAATAGACACAGTTCTAACCCAGGGAGGTACACAGTCTGGTGGGGACATAG
ACTTCAAGGGTGTGGCTCTGGCAGAGATTGGCCACCTCCTGTGCCCCCTGGGTGGGGCTCTCCACTCCCTGTCTCTGCCCTAACCA
GCAGCCAGGCCCTCCCGCTCTGCTGGACCCACCCCTGGTCAGCCCCCGCCAGTCGTTCCACTGCCAGCTTATCACCGCAACTGTCTTTC
TGTCTGCCCACCCAGGCTGCAGGAGGAGATTCAAGTGAAGGAAGAAGCAGAGAACATTTGGCTGCCAGGTGGATGCAGCTACTCTAGCTCGCA
GCATGGCCTCTGGCTTGCTCTGCCACCTGGTGGCGGTGACCATGTCCTCTCGCTGGCCTCTCCAGGACGTGGATGCAGCTACTCTAGCTCGCA
TTGACCTGGAGCGCAGAATTGAATCTCTCAACGAGGAGATCGCGTTCTAAGAAAGTGCATGAAGAGGTATACCTGGCCCTCTCCTGGGGTCACTG
GGCCATGGGGAAAGCAGCCGGAAAGTGGGGTTGGGTGAGGCTCTGGCTGGGAATAGGGTGTGAGGGTGTGTGGCCCTGAGAGGGGACTGAAGCC
CAGTCATGCCCTACAGGAGATCCGTGAGTTGCAGGCTCAGCTCAGGAACAGCAGGTCCAGGTGGAGATGGACATGTCTAACGACACTCACTGCCGCC
CTCAGGGACATCCGGGCTCAGTATGAGACCATCGCGGCTAAGAACATTCTGAAGCTGAGGAGTGGTACAAGTCGAAGGTGGGTGGCCCTGCCGGGAC
TGGCATTCCGTCCTCTGAATCCCAGCTGGATGTGCTGCCCTGGTACCATCCATGGGAGGAGAGGCCAGAGGCTTCATGCTCCCTGCTCATCCCTA
CCCGTGCCCTGCATCCTCTCATTGGGCCCTTCTCTGCCCTTAGGTGTCAGACCTGACCCAGGCAGCCAACAAGAACACGACGCCCTGCCAG
GCCAAGCAGGAGATGATGGAATACCGACACCAGATCCAGTCCTACACCTGCAGGATTGACGCCCTGAAGGGCACTGTGAGTCCTGCCACCTGCCAGG
CCCTGCCCTCTGTGCAGTTCACACCCTCACTTGTGACCTGGGCCATCATAGATCCTCTGGGCCCTCATCTACTAAATCTACAATAGGGG
TAAAACCAGACAAGTGGATTCCAGTTGGATGCTAAGGAATCAGGGTTCTGGCATCTACCTATGTGGGACTGTGAGGCTGAATGCAATGTCCTTG
TATCTATTATTCTGAGTGTTCACATATAGACTTAATTGAGTTCAACATGGCCTGGACCTGACCATCTGGAGTTGCGCTGCCAGGCCAAAG
CTTCTTGGCTGCTAGTGTCTCTCCCTTGACCTGGTTCCCCCTCTGCCCTGCAGAACGATTCCCTGATGAGGCAGATGCCAGATGGGGAAATTGGAGGAC
CGATTGCCAGTGAGGCCAGTGGCTACCAAGGACAACATTGCGCGCCTGGAGGGAGGAATCCGGCACCTCAAGGATGAGATGCCAGGATCTGCCAG
ACCAGGACCTGCTCAACGTGAAGATGGCCCTGGATGTGGAGATTGCCACCTACCGGAAGCTGCTGGAGGGAGAGGCCGGTGAGGGGCCAGGCAGGA
GCCCGAGTGGAGGTGCGGGGTGCTGGGTGGCCATTCTGTCCCCAGGAGGCTCGAGGATTACTGATTACCTCAACAAGACCTGGAAACAATTTTTT
TTTTGAGATGGAGTTGCTCTGCTGCCAGTCTGGAGTGCAATGGCACCATCTGGCTACTGCAACCTCCGCTCCTGGTTCAAGCAATTCTCCT