

Low level Processing and Visualization of Genome Sequencing Data

GSND 5340Q, BMDA

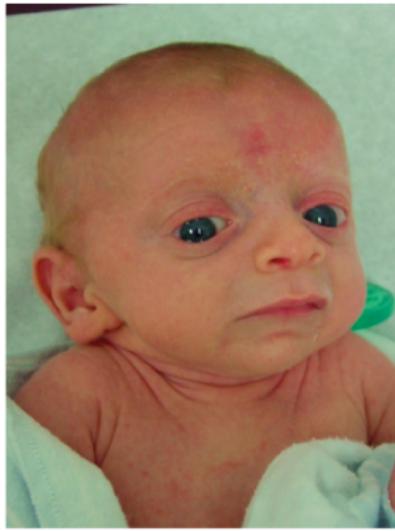
W. Evan Johnson, Ph.D.
Professor, Division of Infectious Disease
Director, Center for Data Science
Rutgers University – New Jersey Medical School

2024-05-06

Section 1

Motivating Example: X-linked Disease

Rare X-linked Disease



Rare X-linked Disease



Uncle #1



Uncle #2

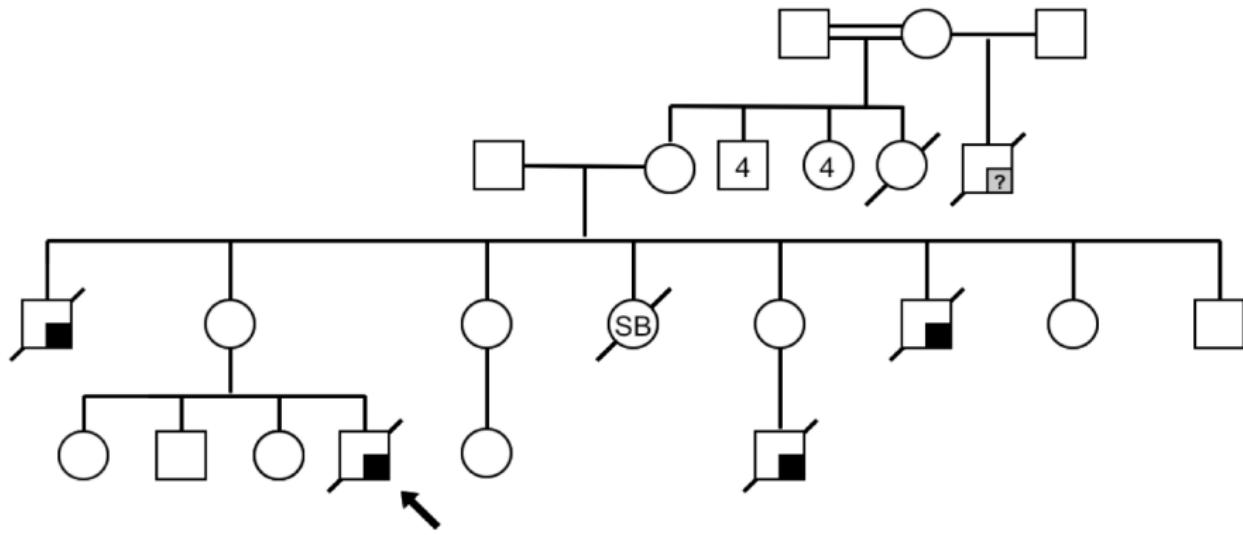


cousin

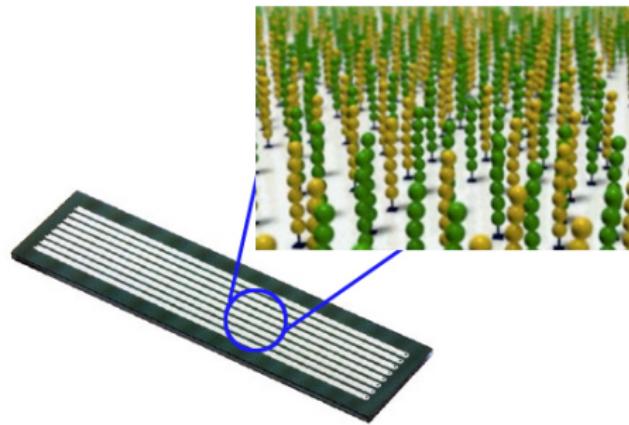


proband

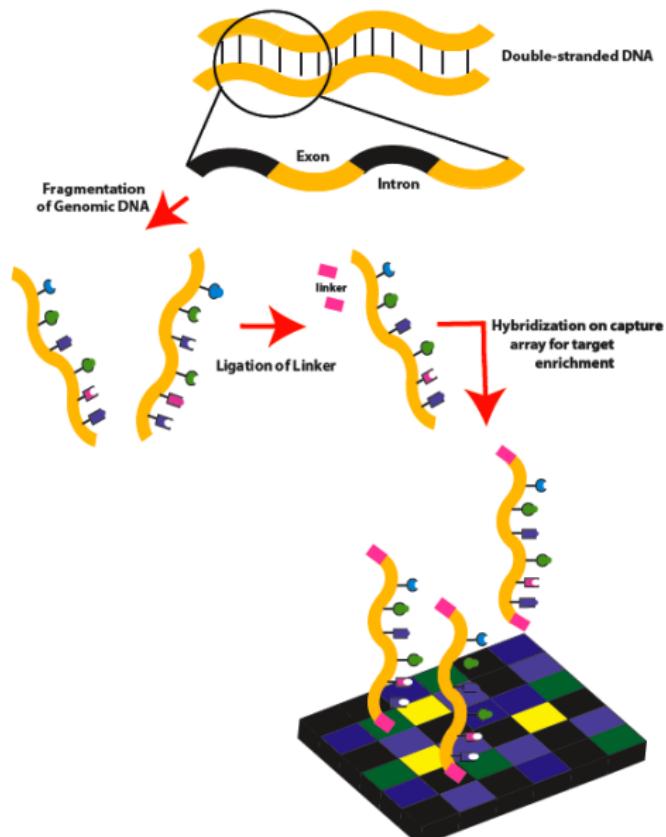
Rare X-linked Disease



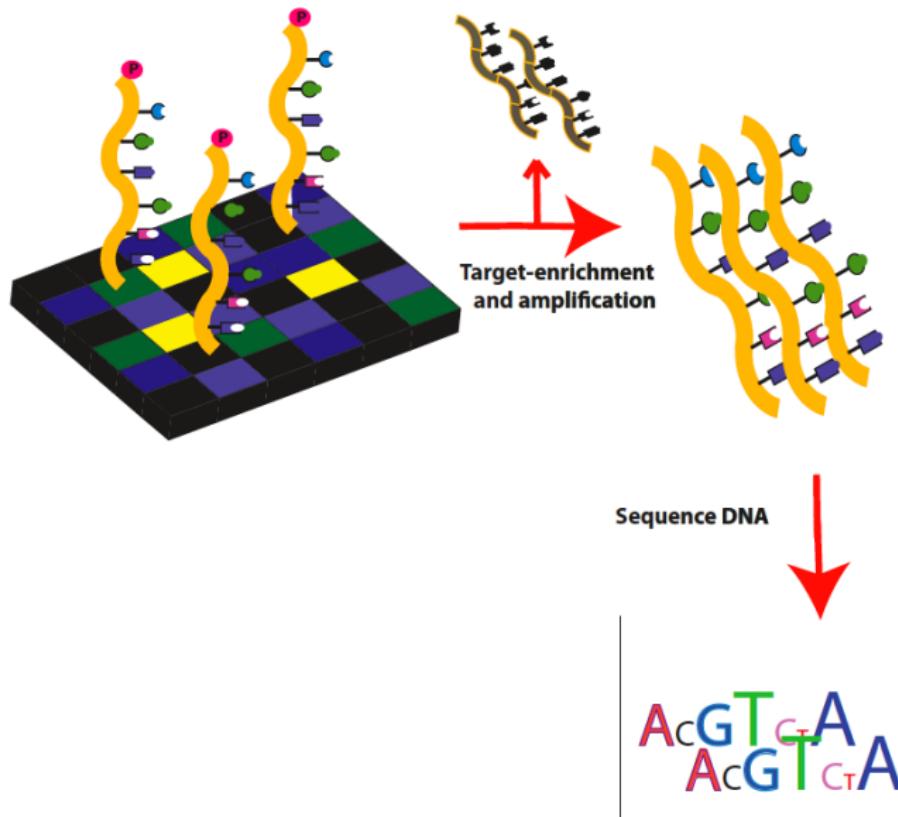
Exon Capture Sequencing



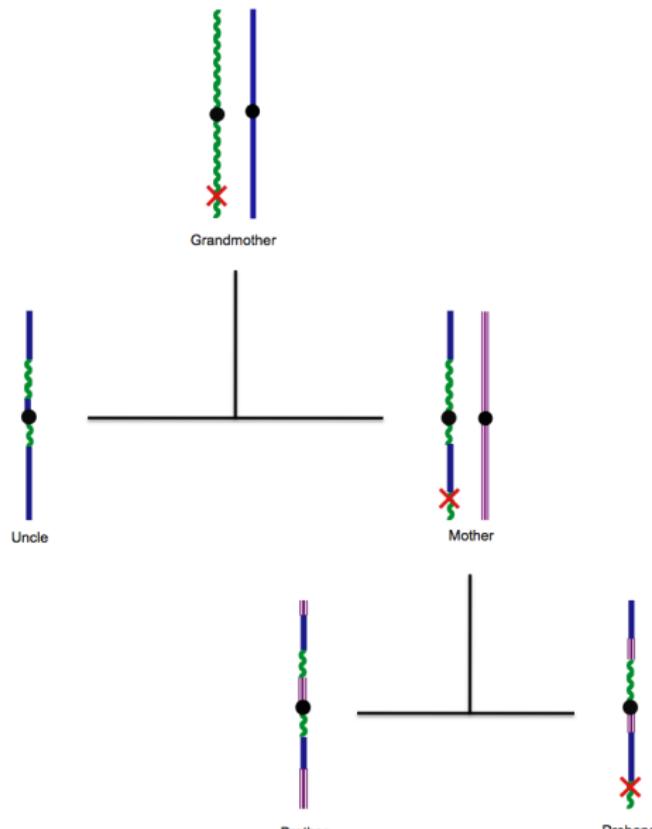
Exome Capture



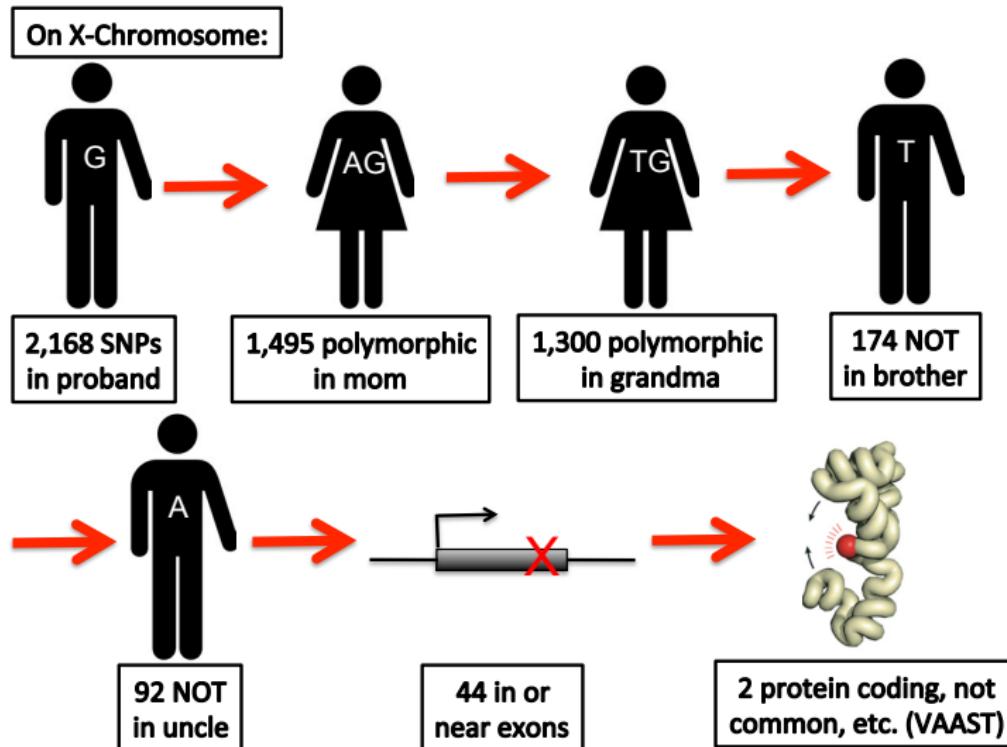
Exome Capture



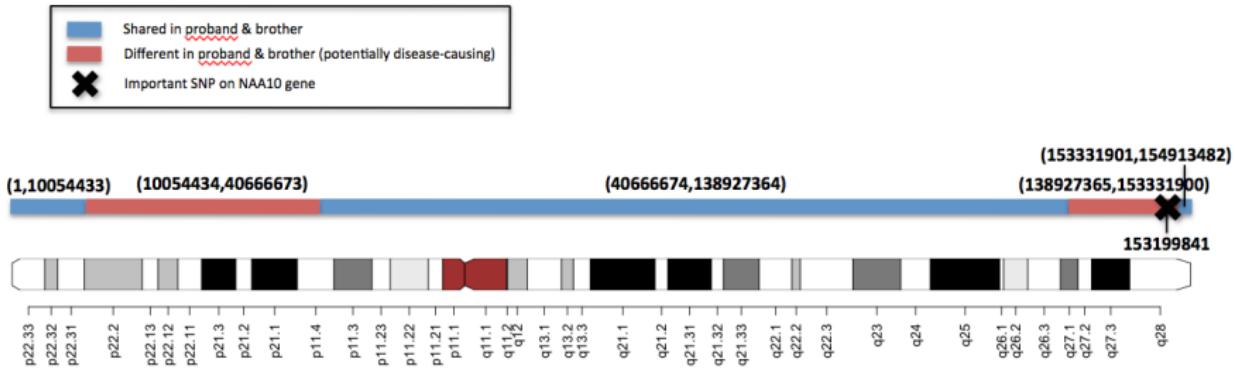
Rare X-linked Disease



Rare X-linked Disease}



Rare X-linked Disease



Rare X-linked Disease

AJHG  **Supports open access**

Submit Log in Register Subscribe Claim

ARTICLE | VOLUME 89, ISSUE 1, P28-43, JULY 15, 2011 [Download Full Issue](#)

PDF [603 KB] Figures Save Share Reprints Request

Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency

Alan F. Rope • Kai Wang ¹⁹ • Rune Evjenth • ... Mark Yandell • Thomas Arnesen • Gholson J. Lyon    Show all authors • Show footnotes

Open Archive • Published: June 23, 2011 • DOI: <https://doi.org/10.1016/j.ajhg.2011.05.017>

 PlumX Metrics

Introduction

We have identified two families with a previously undescribed lethal X-linked disorder of infancy; the disorder comprises a distinct combination of an aged appearance, craniofacial anomalies, hypotonia, global developmental delays, cryptorchidism, and cardiac arrhythmias. Using *X* chromosome exon sequencing and a recently developed probabilistic algorithm aimed at discovering disease-causing variants, we identified in one family a c.109T>C (*p.Ser37Pro*) variant in *NAA10*, a gene encoding the catalytic subunit of the major human N-terminal acetyltransferase (NAT). A parallel effort on a second unrelated family converged on the same variant. The absence of this variant in controls, the amino acid conservation of this region of the protein, the predicted disruptive change, and the co-occurrence in two unrelated families with the same rare

Subjects and Methods

Results

Discussion

Acknowledgments

Supplemental Data

Web Resources

References

Article info



Ad served by Google

[Ad options](#)

[Send feedback](#)

Why this ad? 

N-terminal acetyltransferase (NAA10)

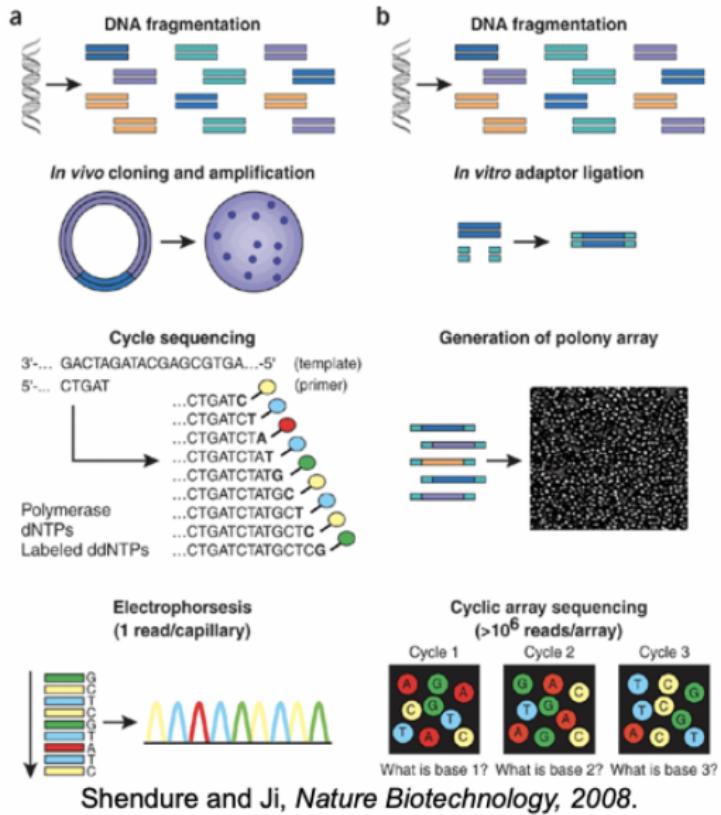
N-terminal acetyltransferase (NAT):

- Common modification (~80-90% of human proteins)
- Depletion from cancer cells linked to cell cycle arrest and apoptosis
(Starheim, *BMC Proc* 2009)
- NAT genes directly implicated as cause of genetic disease
- Mutation demonstrated a significantly impaired biochemical activity *in vitro*
- NAA10 lethal if knocked out of Drosophila

Section 2

Processing Raw Data

Next-Generation Sequencing



Study Design

Whole Exome

- Less expensive
- Nearly complete ascertainment of variation in the coding ~1% of the genome (i.e. exons)
- Will miss functional variants outside of the coding region

Low Coverage Whole Genome

- Less expensive
- Reasonably good ascertainment of shared variation, but not unique variation

Deep Whole Genome

- More expensive
- Capture most of the genetic information
- Sequence the entire genome of each subject

History and Evolution of Illumina Data Output

Illumina sequencers have given output in many different formats:

- Illumina .PRB and .INT files
 - Better access to raw data.
 - Base calling algorithms (Bravo and Irizarry, *Biometrics*, 2010)
 - Mapping algorithms (GNUMAP, NOVO)
 - Confusing formats; Large data files
- Illumina .FASTQ files
- Sanger .FASTQ files

Illumina .INT and .PRB

1	1	125	771	1651.8	2189.6	228.1	549.9	219.0	202.5	48.4	3016.8	127.8	6.1	204
1	1	478	16	1050.0	969.9	149.5	311.7	0.0	0.0	39.3	134.1	0.0	0.0	0.0
1	1	780	553	639.6	980.8	555.2	6412.8	1040.1	4408.7	750.3	638.1	946.2	4351.1	7
1	1	123	685	-116.5	341.5	-14.0	-985.9	231.5	1090.0	88.3	-102.2	240.2	513.6	
1	1	61	934	40.7	87.3	38.5	21.3	16.4	31.7	100.9	68.8	41.4	29.4	40.4
1	1	866	972	2820.5	4698.9	435.8	8502.2	4740.3	4890.2	1491.5	1241.7	2137.5	2505.6	6
40	-40	-40	-40	-40	-40	-40	-40	40	-40	-40	40	-40	-40	-40
28	-28	-40	-40	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
-17	-16	-30	13	-40	26	-26	-40	-21	17	-20	-27	21	-21	-40
-40	40	-40	-40	-40	40	-40	-40	-1	1	-40	-40	-0	-40	-0
-1	-40	-0	-10	-20	-18	16	-35	-1	-8	-2	-13	12	-34	-13
-27	-40	-40	27	40	-40	-40	-40	-40	-40	40	-40	-40	40	-40

.fasta or .fa

```
>chrX
ttgaactcctgacctcaggtatccgcggcctgacccaaaggcgct
cctgcctcagcctcccgagtagctggactacaggtgccaccatgc
.....
caggctaattttgtattttagtagagacgggtttcaccatgttagc
caggatggtctcaatctcctgacctcatgatccgcctgcctcgccccc
>chrY
tgtacacttaaatgggtgaatttatggaatgtgaattataCGTGTGG
CTTGTAAAAAAAAATGATGGAGATGGAGACGTGACTCTAGCGTGAAGGGG
.....
GTGGGGAGAGTAGATCTAGAGTGGAGACACCACTTTAGGAGGTATGATC
cctgccaccatgcctggctaattttgtattttagtagagacagggtt
```

.fastq or .fq (or .fq.gz)

@HWI-EAS240_0001:2:1:1142:17571#0/1

CTCTCTTTCTCCCCANGTCTCCTCATGACCATATCCNTGTTGCCATTGTGTANGGNNNCTTC

+HWI-EAS240 0001;2;1;1142;17571#0/1

@HWI-EAS240 0001;2;1;1142;17571#0/1

CCTCTTTCTCCCANGTCTCTCATGACCATAATCCNTGTTGCCATTGTGTANGNNNNCTT

+HWT=FAS240 0001:2:1:1142:17571#0/1

bababbb` abbbbabbhYB^[[^` ` bbbbbbbbbb` B` ``^` ^` bbbbba^` ^` ^` B00BBN00

@HWT-EAS340_0001:2:1:1143:6453#0/1

CAGGACGCTGCCACTATGNCATGCAAGATGCAAGTCTTANCATATTGAGGATAACANATANNGCC

+HWT_EAS340_0001:3:1:1143:6453#0/1

+HWI-EAS246_0001.2.1.1142.8435#01
[1][so00]\[BL1][1][S]\[X]\-\[soZ\]B11000[0]\-\[ovx\]XPNNNRP11]

CH/UT_EAS243_2001-2-1-1143-10443#2/1

@HWI-EAS240_0001:2:1:1142:19443#0/1

TAGAAAACAGAAAGGTCATTTCTTACTCTTG
AATC TACCGA CGGGCA GAGG GAGG GAGG

+HWI-EAS240_0001:2:1:1142:19443#0/1

bbbbabb^~bbb^bbabbb]bb]~]`]]] bbbbbbYYBYX00000bb bBVY^HB000BVBOJO

.FASTQ Comparison

Quality scores (PHRED)

Quality scores (Phred)

- Sanger Phred: Range=(0,40), $P = 1 - 10^{-(ASCII-33)/10}$
- Solexa: Range= (-5,40), $P = \frac{10^{(ASCII-64)/10}}{1+10^{(ASCII-64)/10}}$
- Illumina 1.3: (0,40), $P = \frac{10^{(ASCII-64)/10}}{1+10^{(ASCII-64)/10}}$
- Illumina 1.5: Range=(2,40), $P = 1 - 10^{-(ASCII-64)/10}$
- Illumina 1.8: Same as Sanger except Range=(0,41)

.FASTQ Comparison

> Sanger	> Solexa	> Illumina1.3	> Illumina1.5
ASCII	Quality	ASCII	Quality
!	33 0.0000	;	59 0.2403
"	34 0.2057	<	60 0.2847
#	35 0.3690	=	61 0.3339
\$	36 0.4988	>	62 0.3869
%	37 0.6019	?	63 0.4427
&	38 0.6838	@	64 0.5000
'	39 0.7488	A	65 0.5573
(40 0.8005	B	66 0.6131
)	41 0.8415	C	67 0.6661
*	42 0.8741	D	68 0.7153
+	43 0.9000	E	69 0.7597
,	44 0.9206	F	70 0.7992
-	45 0.9369	G	71 0.8337
.	46 0.9499	H	72 0.8632
/	47 0.9602	I	73 0.8882
0	48 0.9684	J	74 0.9091
1	49 0.9749	K	75 0.9264
2	50 0.9800	L	76 0.9406
		M	77 0.9523
		N	78 0.9617
		O	79 0.9693
		P	80 0.9755
		Q	81 0.9804
		R	82 0.9844
		S	83 0.9876
		T	84 0.9900
		U	85 0.9921

Session info

```
sessionInfo()
```

```
## R version 4.4.0 (2024-04-24)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sonoma 14.2.1
##
## Matrix products: default
## BLAS:    /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK:  /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/Denver
## tzcode source: internal
##
## attached base packages:
## [1] stats      graphics   grDevices utils      datasets   methods    base
##
## loaded via a namespace (and not attached):
## [1] compiler_4.4.0  fastmap_1.1.1   cli_3.6.2       tools_4.4.0
## [5] htmltools_0.5.8.1 rstudioapi_0.16.0 yaml_2.3.8     rmarkdown_2.26
## [9] knitr_1.46     xfun_0.43      digest_0.6.35   rlang_1.1.3
## [13] evaluate_0.23
```