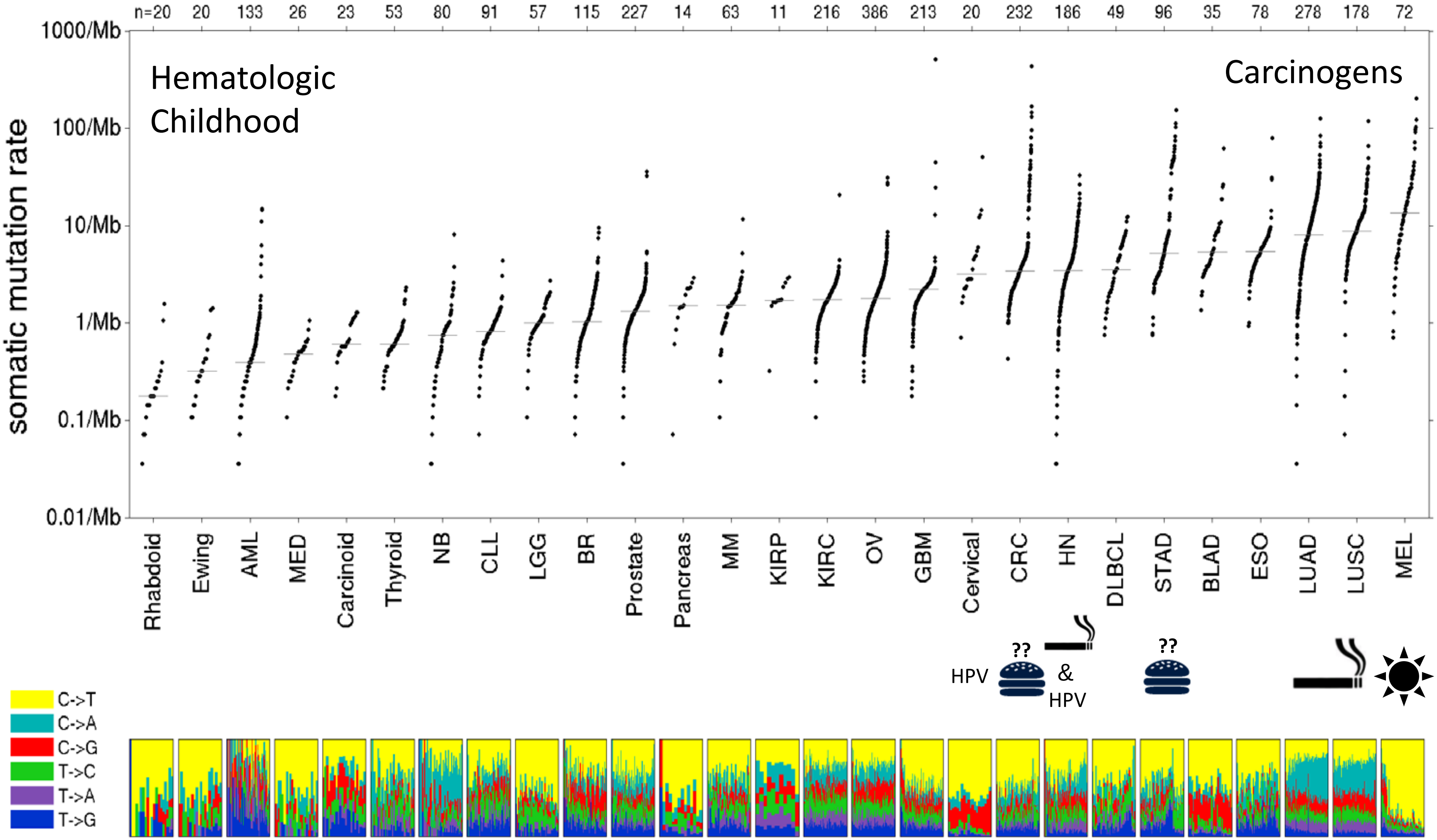
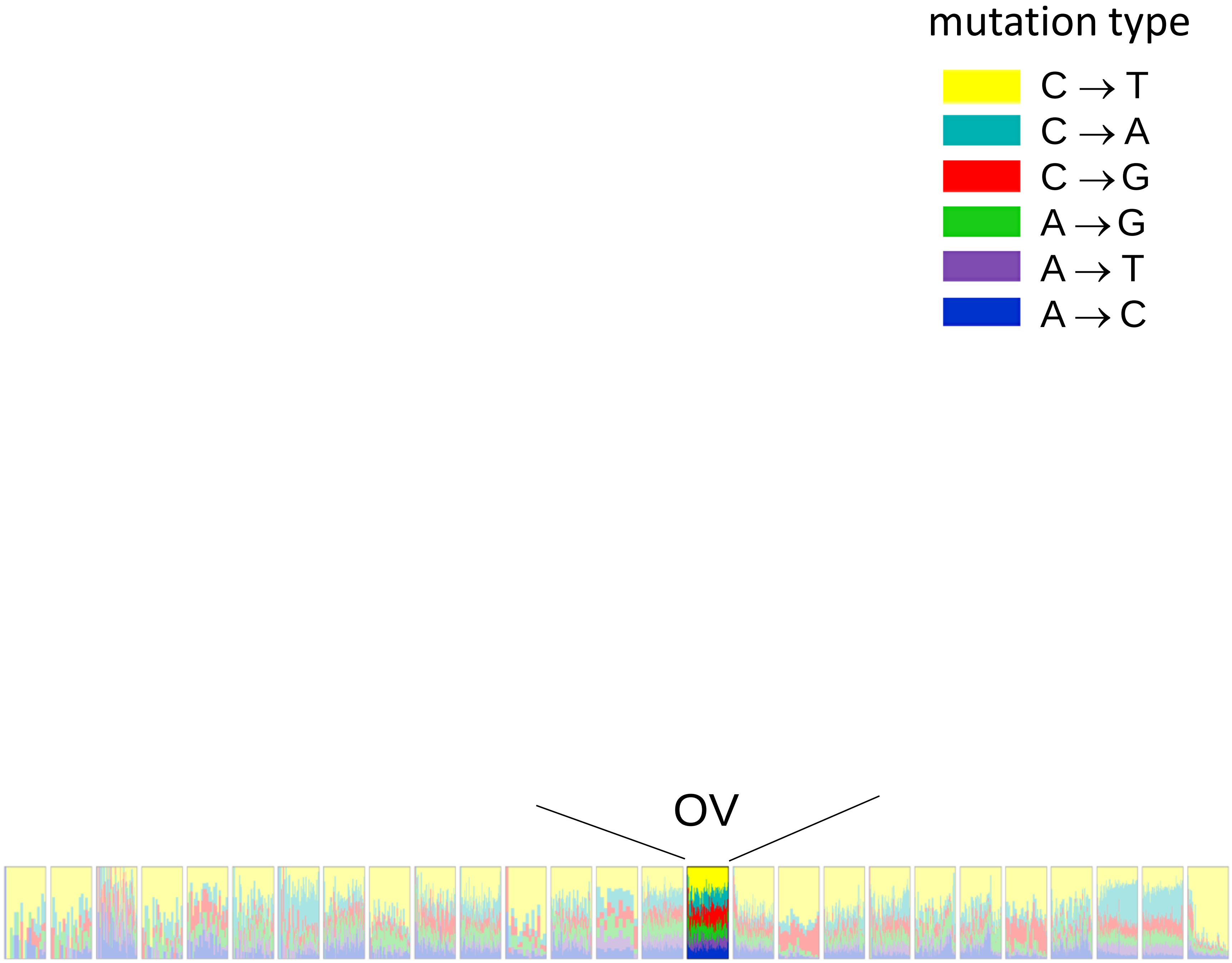
The spectra of somatic mutations across many tumor types

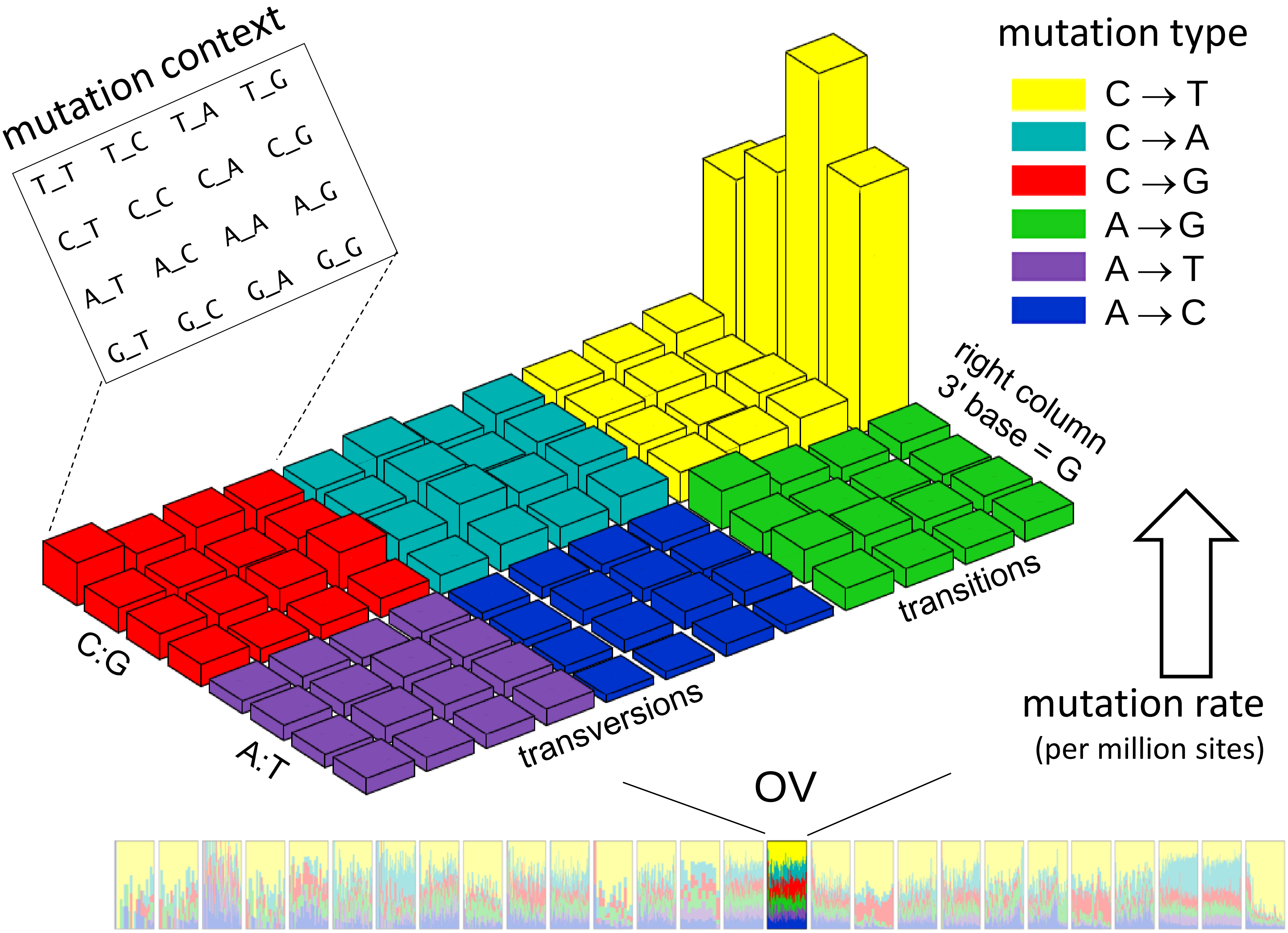
Mike Lawrence
Broad Institute of Harvard and MIT

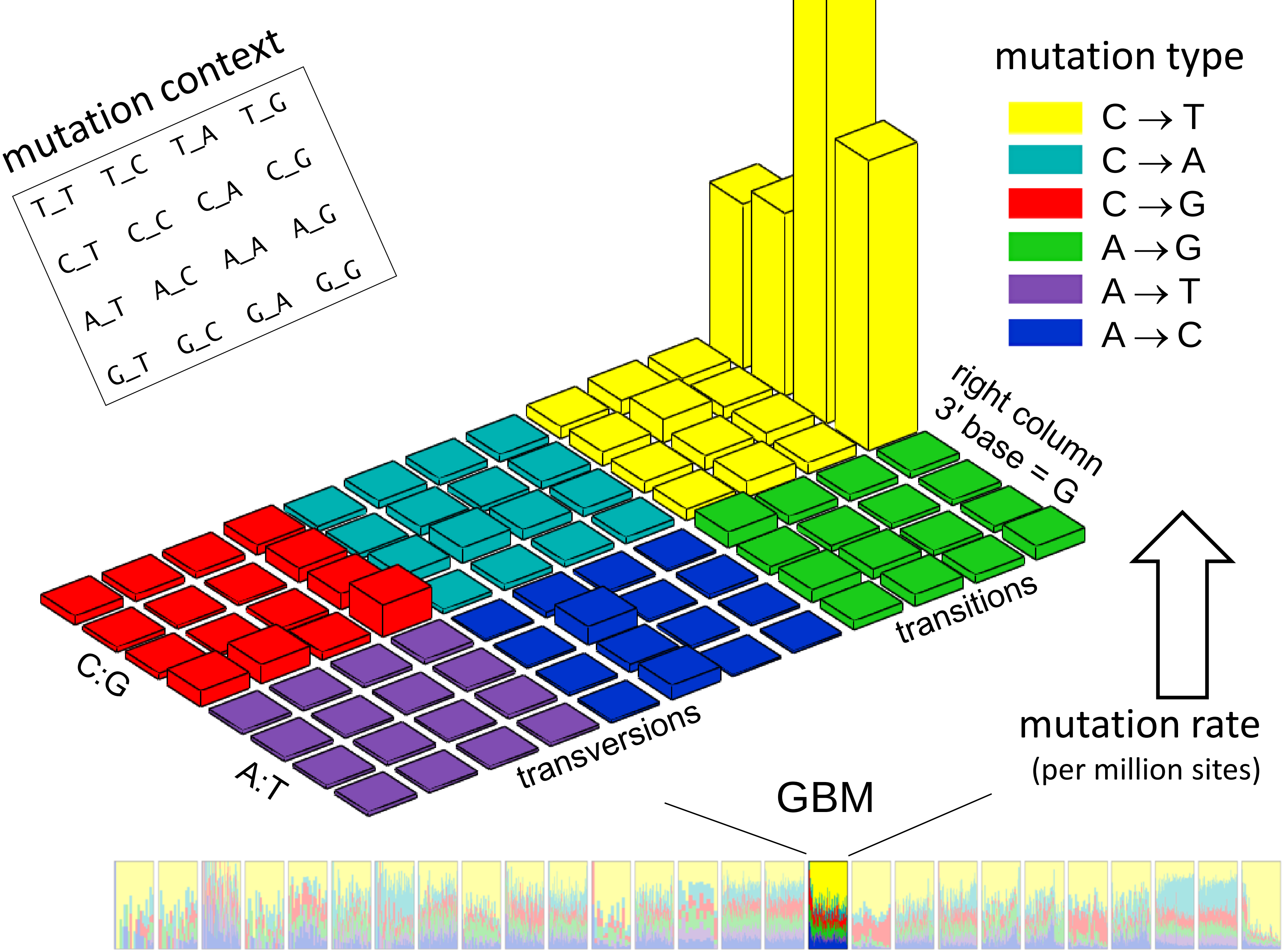
1st Annual TCGA Scientific Symposium
November 17, 2011

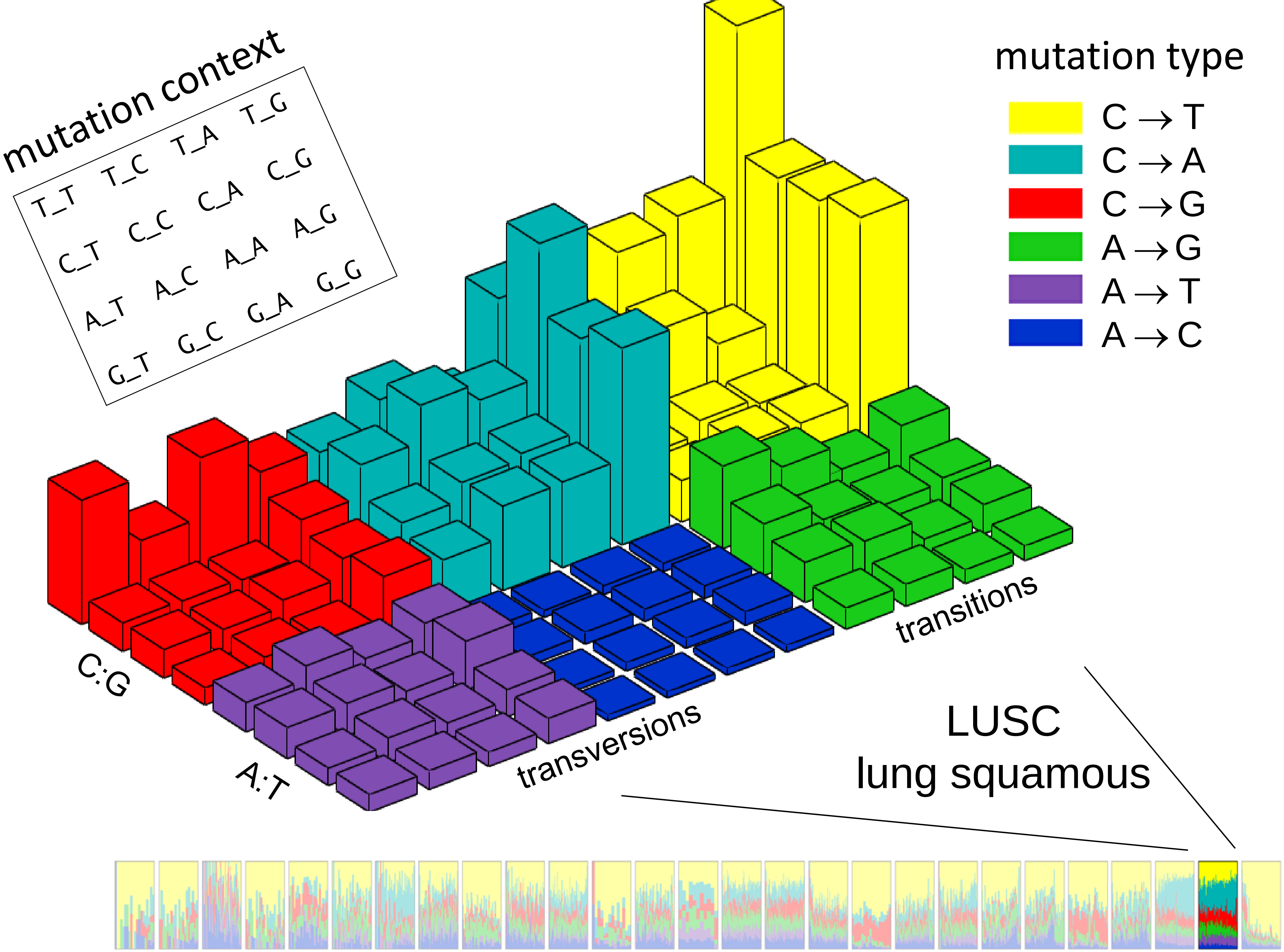
mutation rates across cancer

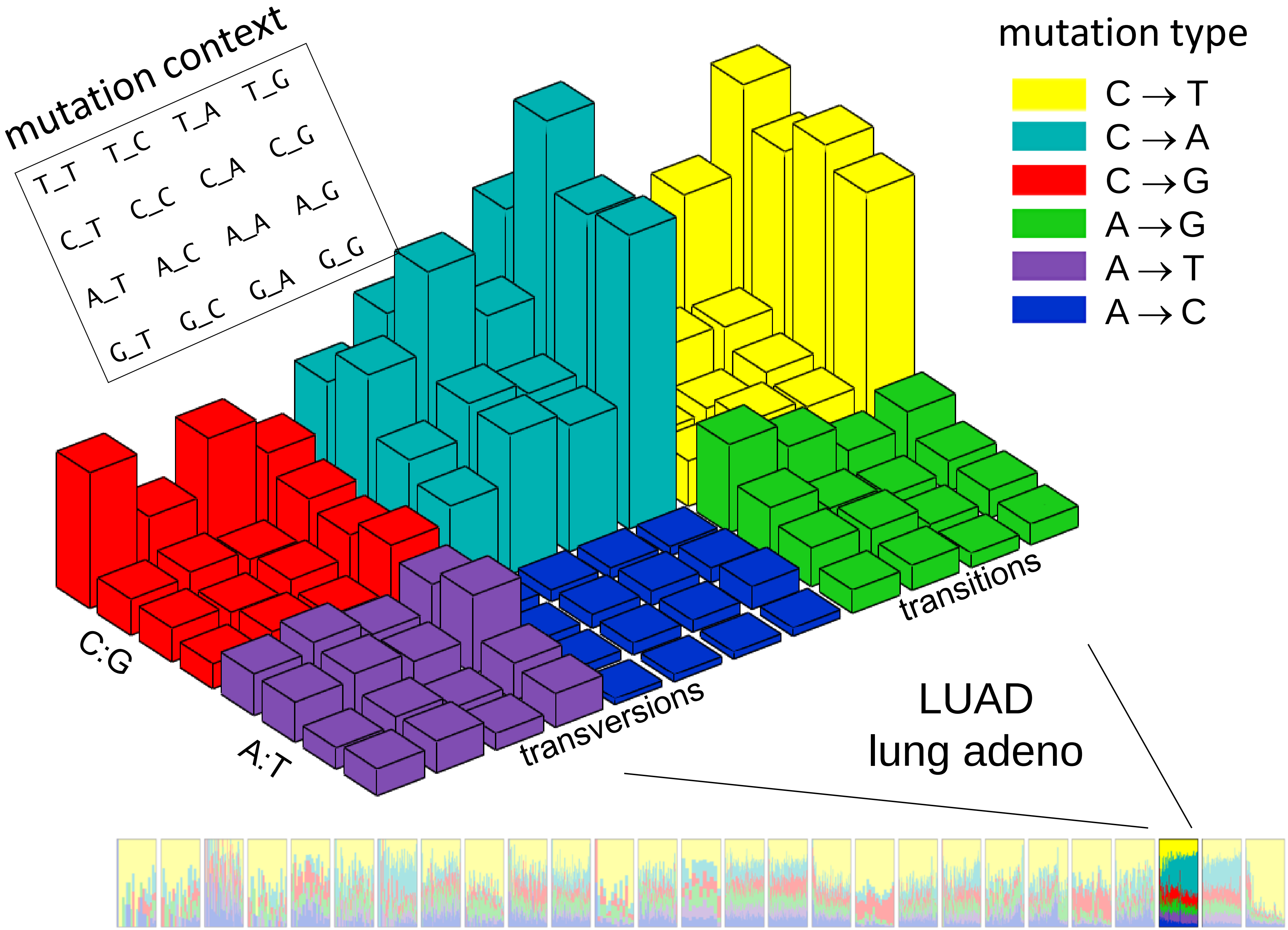


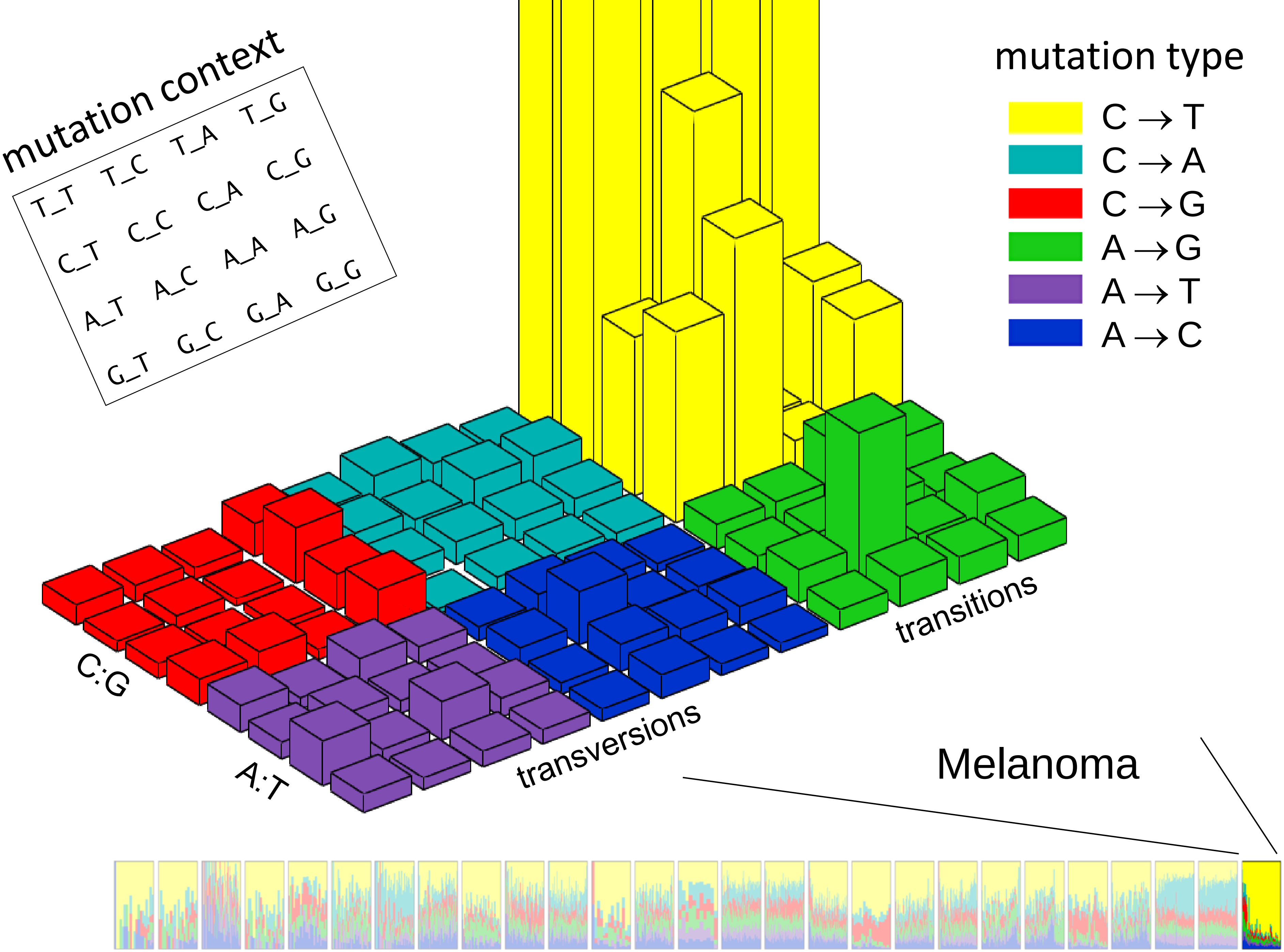

mutation type
🟨 C → T
🟦 C → A
🟥 C → G
🟩 A → G
🟪 A → T
🟦 A → C
OV

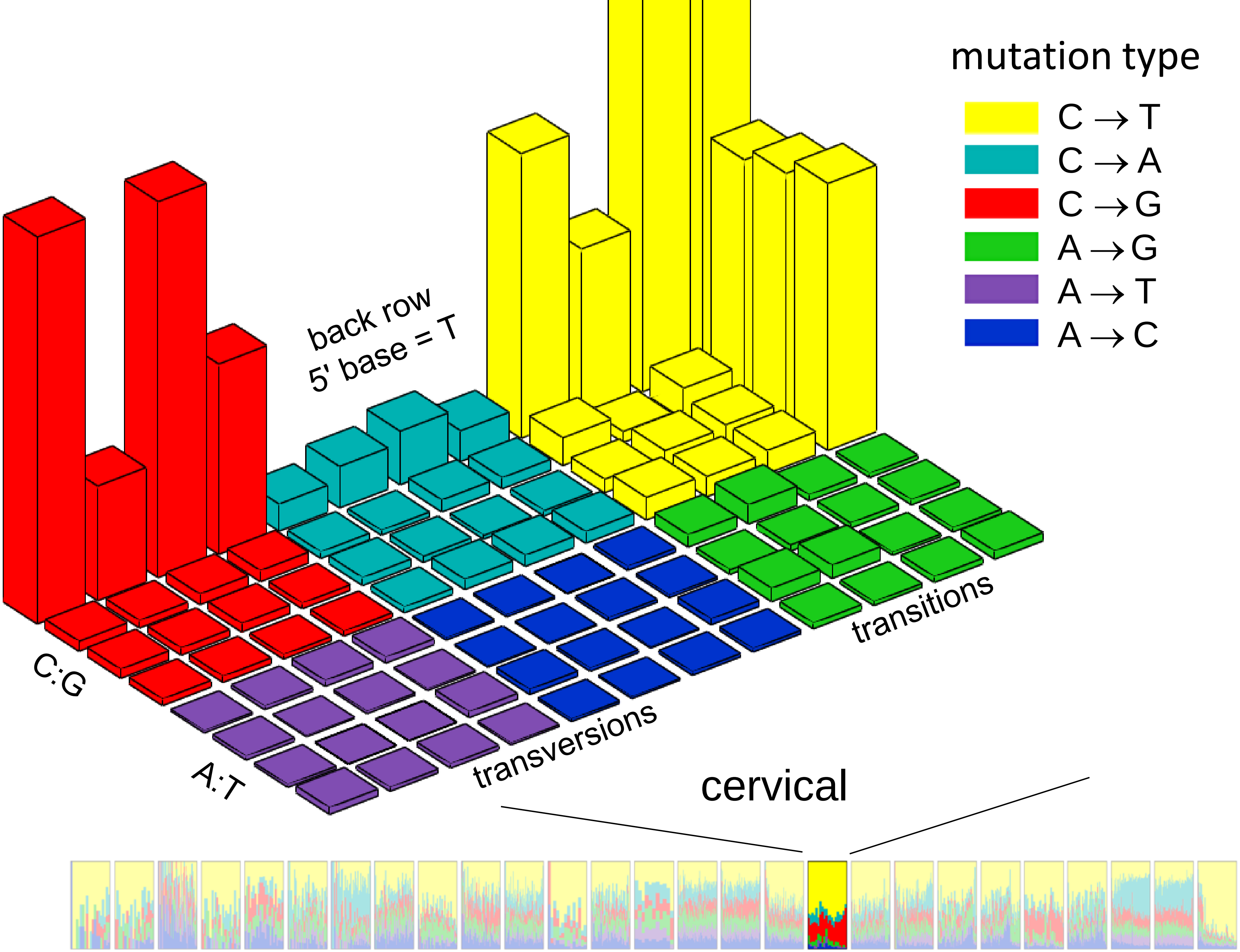

mutation context
T_T T_C T_A T_G
C_T C_C C_A C_G
A_T A_C A_A A_G
G_T G_C G_A G_G
mutation type
C → T
C → A
C → G
A → G
A → T
A → C
right column
3' base = G
C:G
A:T
transversions
transitions
mutation rate
(per million sites)
OV

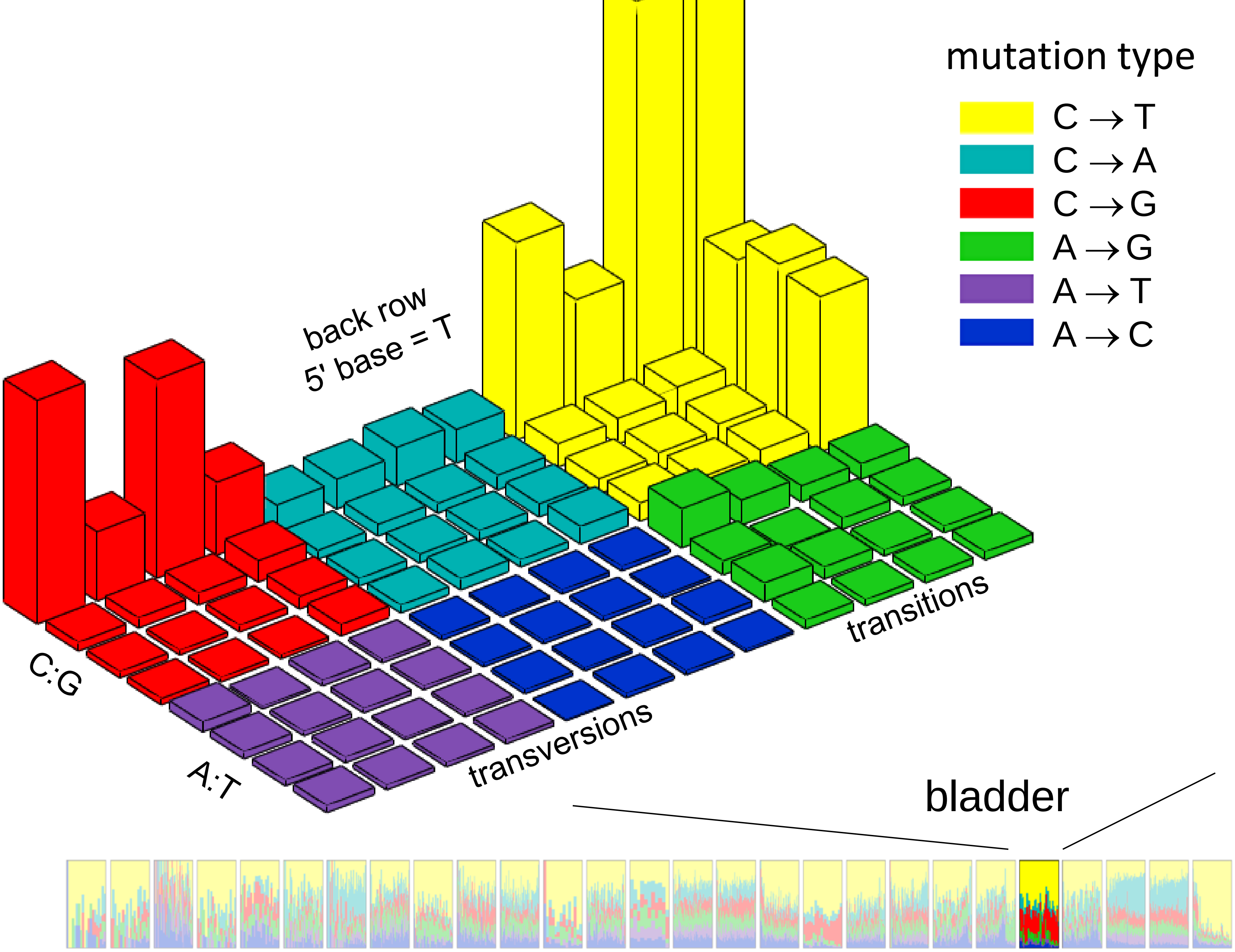

mutation context
T_T T_C T_A T_G
C_T C_C C_A C_G
A_T A_C A_A A_G
G_T G_C G_A G_G
mutation type
C → T
C → A
C → G
A → G
A → T
A → C
right column
3' base = G
C:G
A:T
transversions
transitions
mutation rate
(per million sites)
GBM

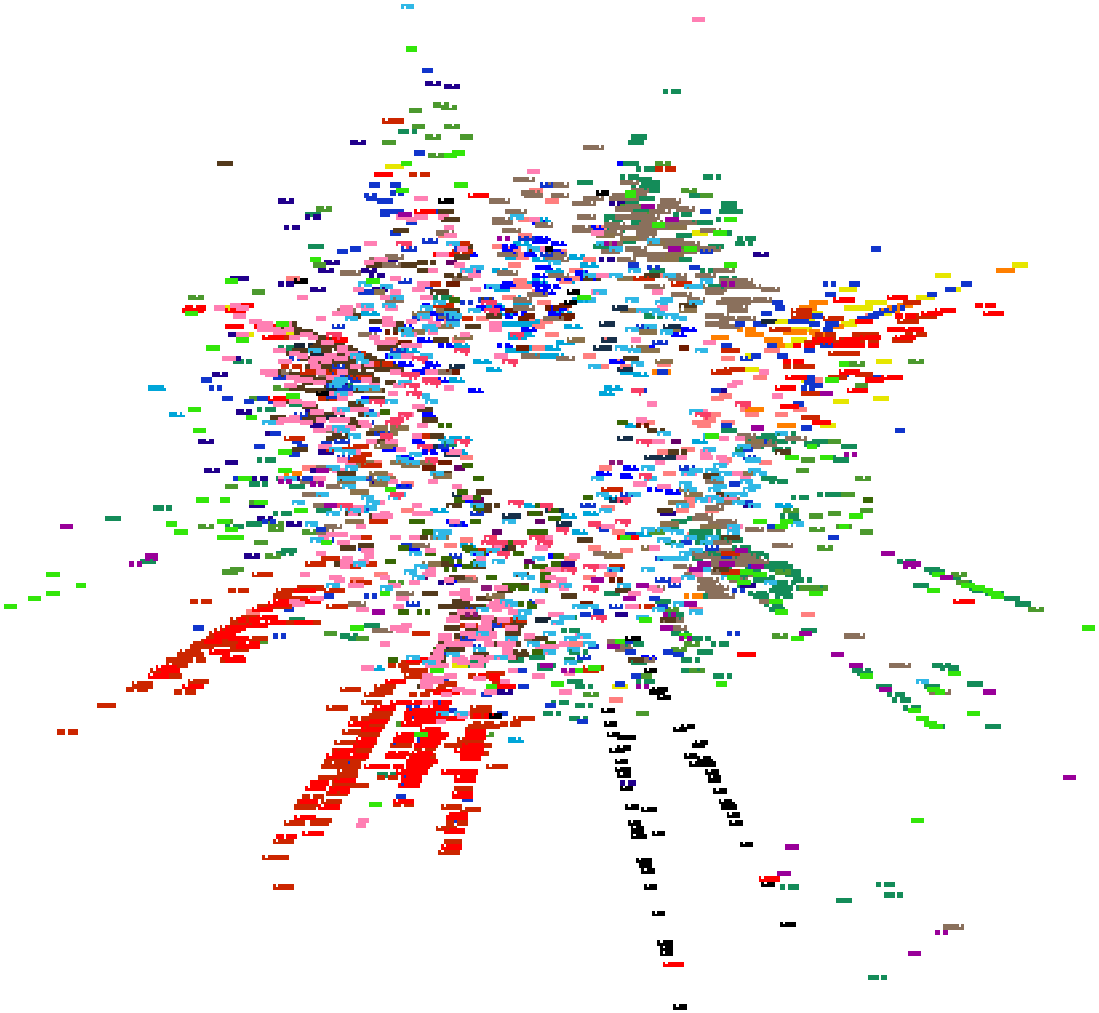

mutation context
T_T T_C T_A T_G
C_T C_C C_A C_G
A_T A_C A_A A_G
G_T G_C G_A G_G
mutation type
🟨 C → T
🟦 C → A
🟥 C → G
🟩 A → G
🟪 A → T
🟦 A → C
C:G
A:T
transversions
transitions
LUSC
lung squamous

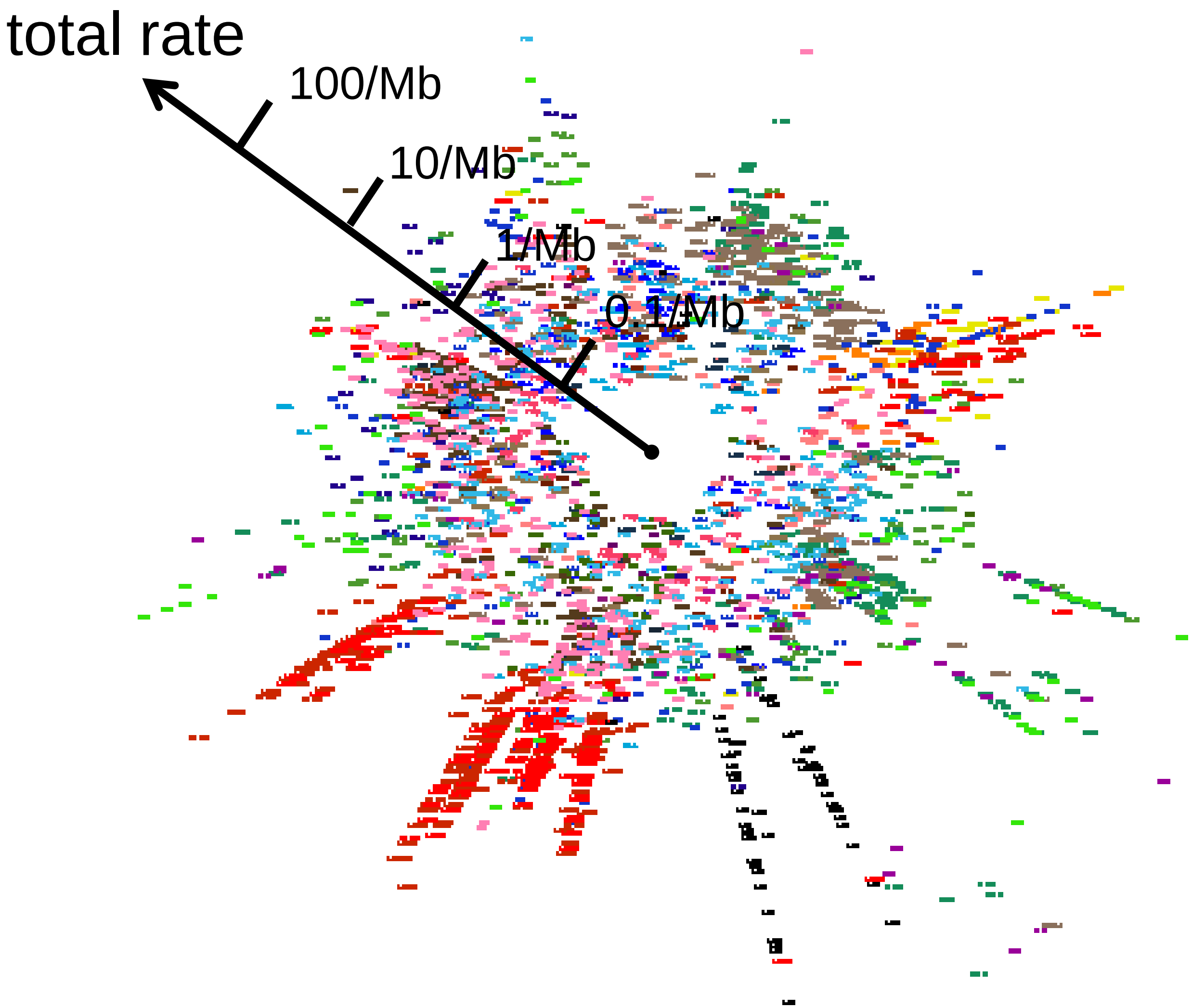

mutation context
T_T T_C T_A T_G
C_T C_C C_A C_G
A_T A_C A_A A_G
G_T G_C G_A G_G
mutation type
C → T
C → A
C → G
A → G
A → T
A → C
C:G
A:T
transversions
transitions
LUAD
lung adeno

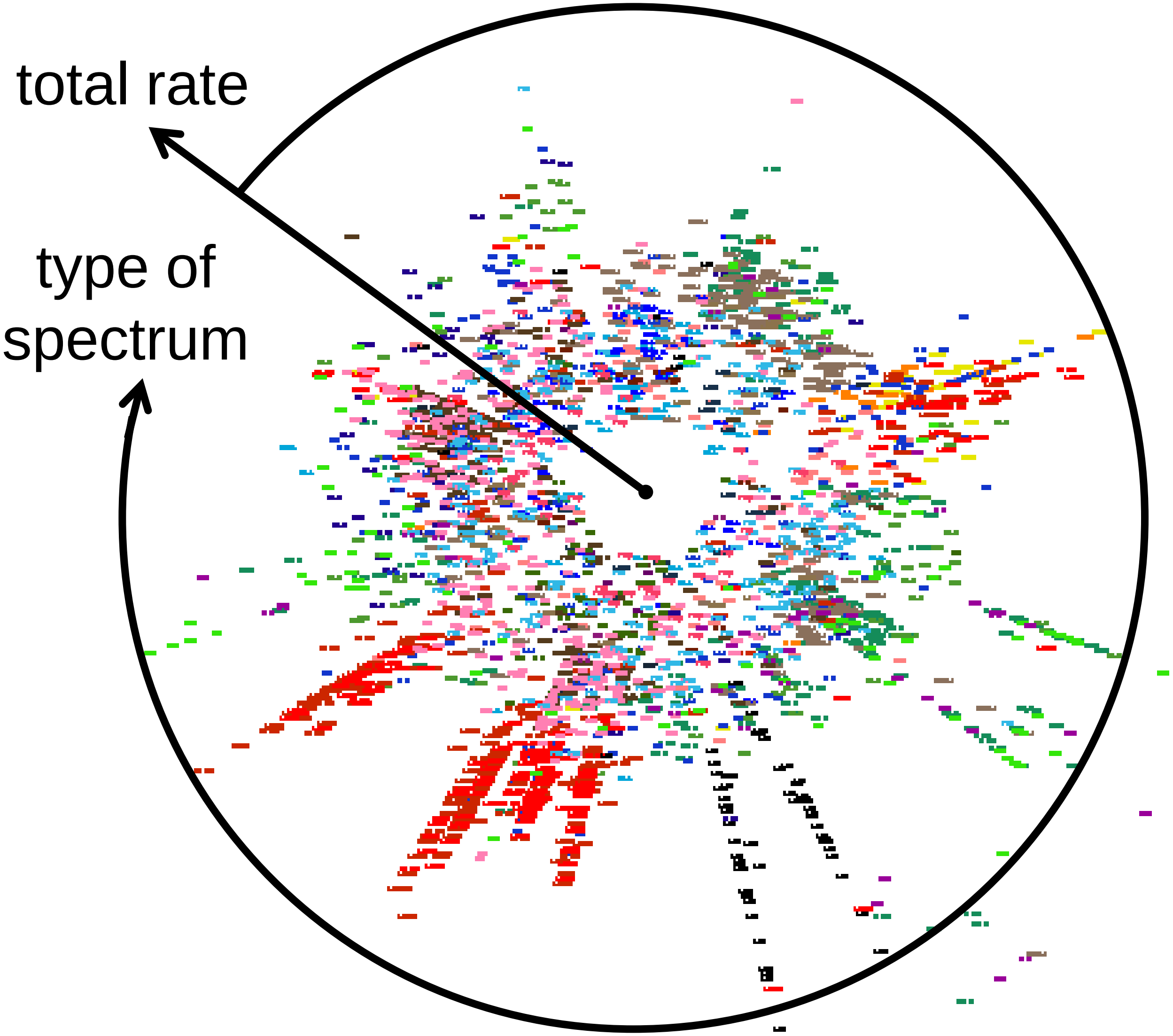

mutation context
T_T T_C T_A T_G
C_T C_C C_A C_G
A_T A_C A_A A_G
G_T G_C G_A G_G
mutation type
C → T
C → A
C → G
A → G
A → T
A → C
C:G
A:T
transversions
transitions
Melanoma

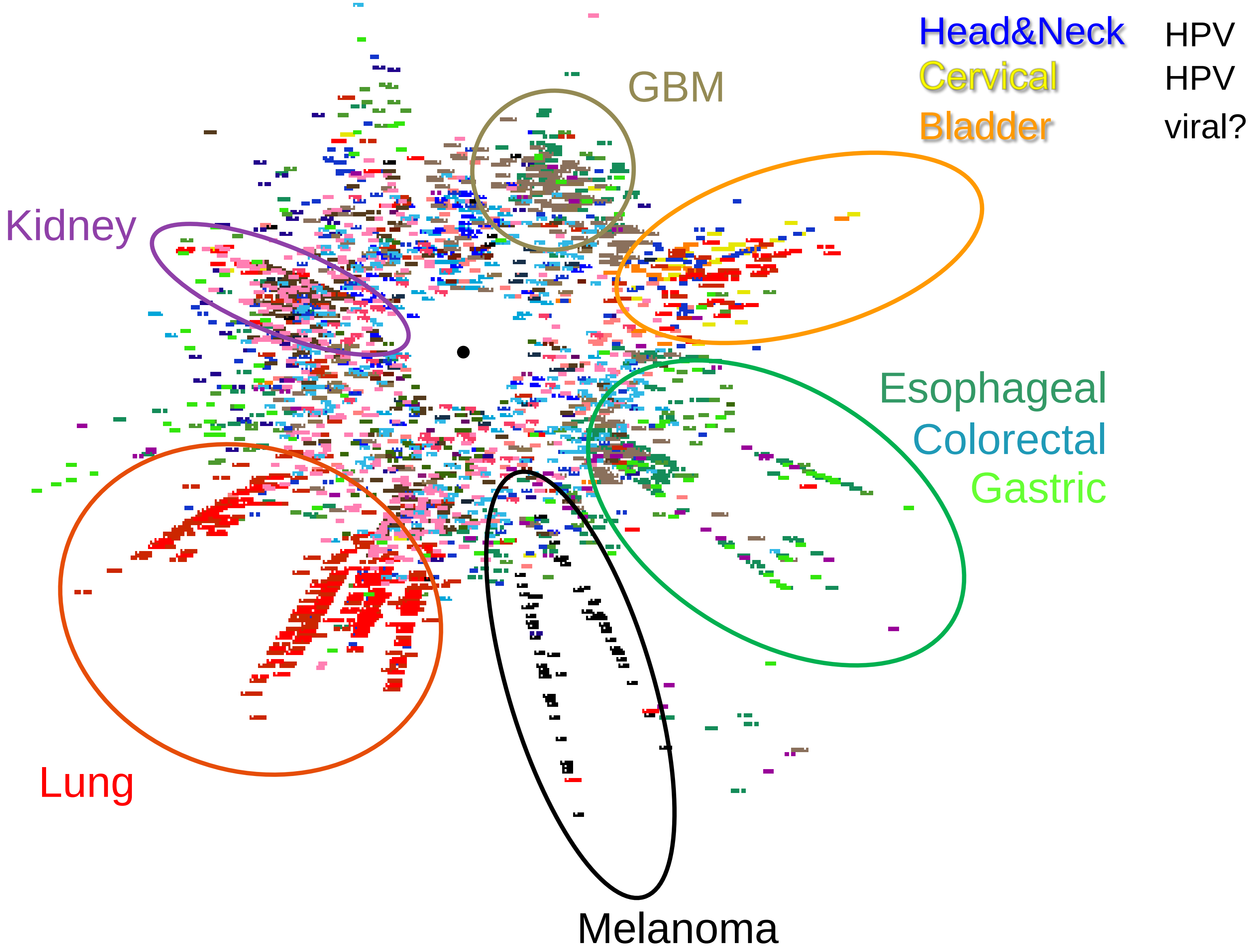

mutation type
C → T
C → A
C → G
A → G
A → T
A → C
back row
5' base = T
C:G
A:T
transversions
transitions
cervical

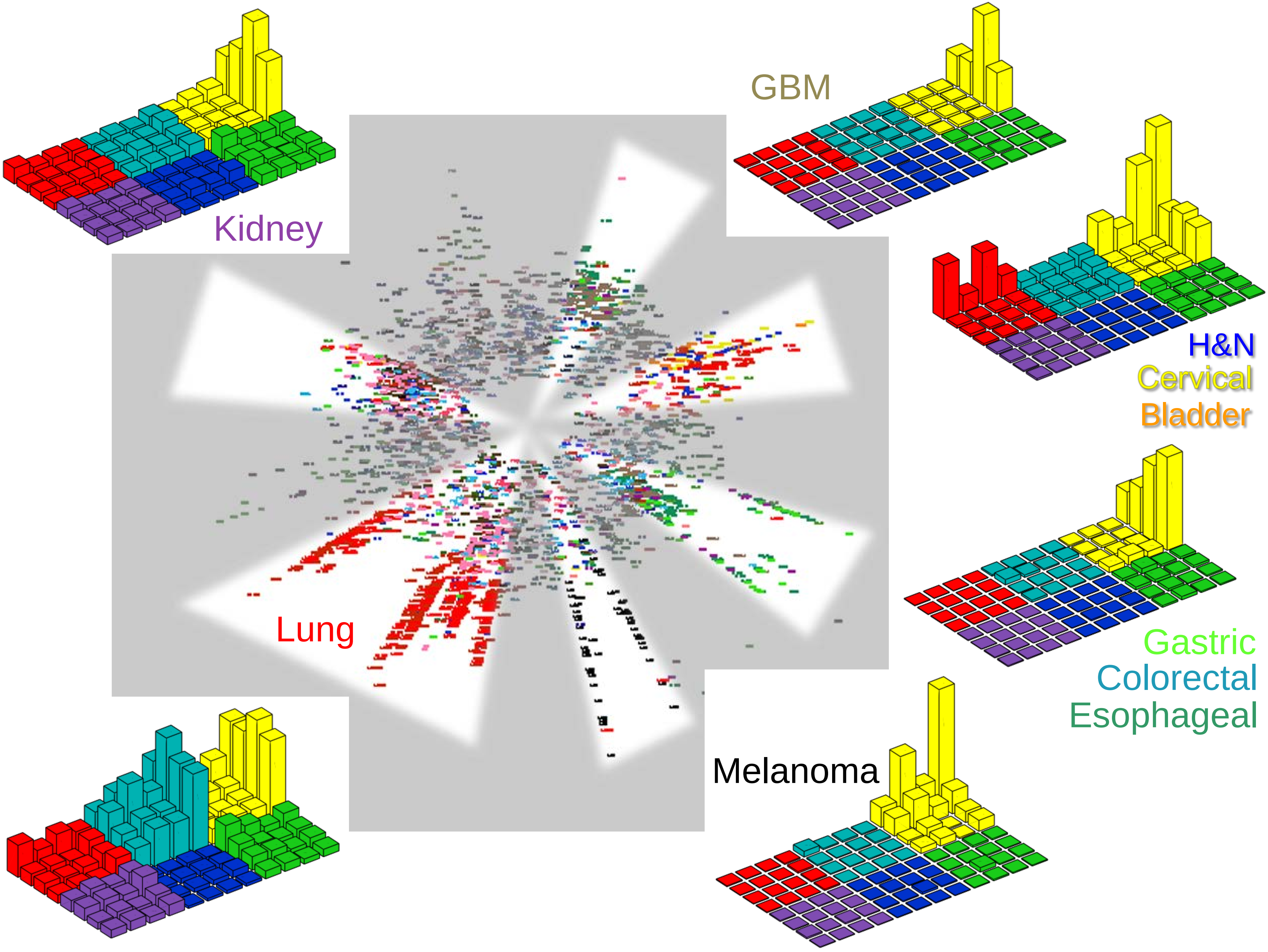

mutation type
C → T
C → A
C → G
A → G
A → T
A → C
back row
5' base = T
C:G
A:T
transversions
transitions
bladder



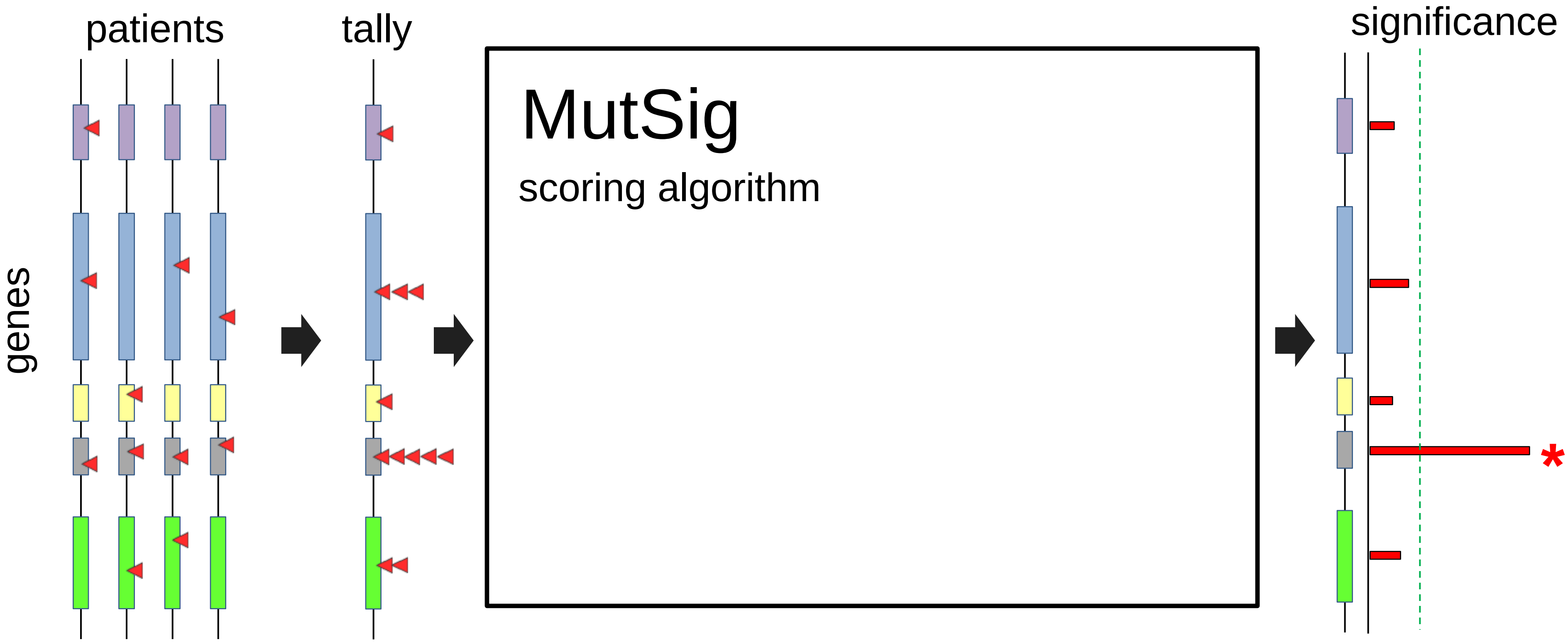

total rate
100/Mb
10/Mb
1/Mb
0.1/Mb


total rate
type of spectrum


Head&Neck HPV
Cervical HPV
Bladder viral?
GBM
Kidney
Esophageal
Colorectal
Gastric
Lung
Melanoma


Kidney
GBM
H&N
Cervical
Bladder
Gastric
Colorectal
Esophageal
Melanoma
Lung

finding significantly mutated genes



patients

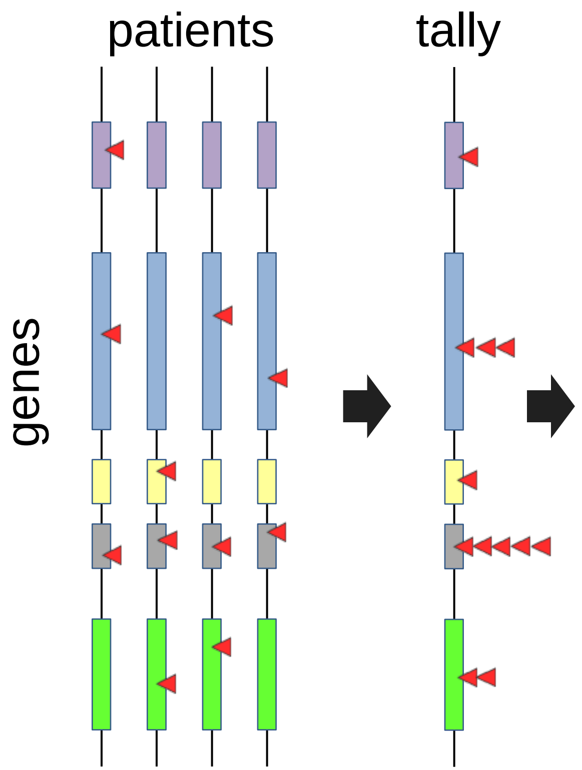
genes

tally

MutSig

scoring algorithm

significance



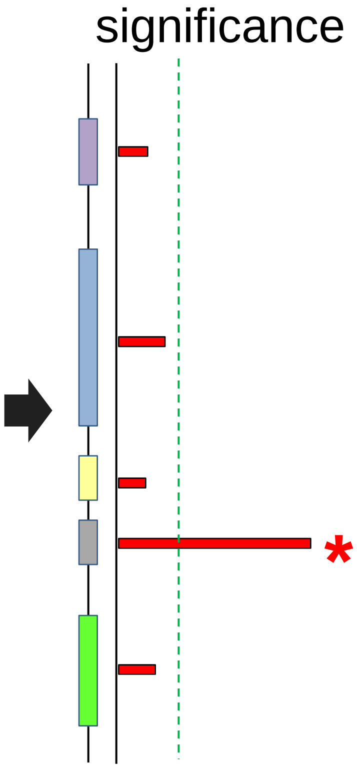
MutSig

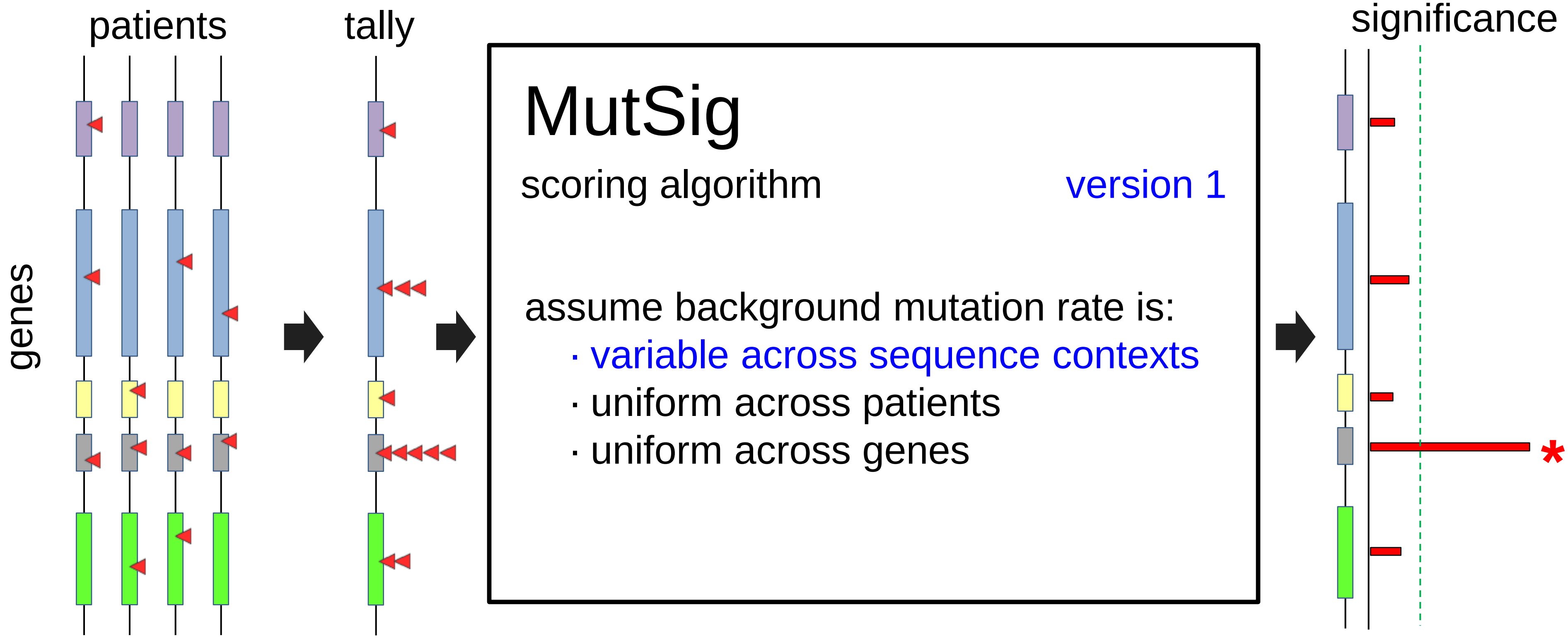
scoring algorithm

version 0

assume background mutation rate is:

- uniform across sequence contexts
- uniform across patients
- uniform across genes





MutSig

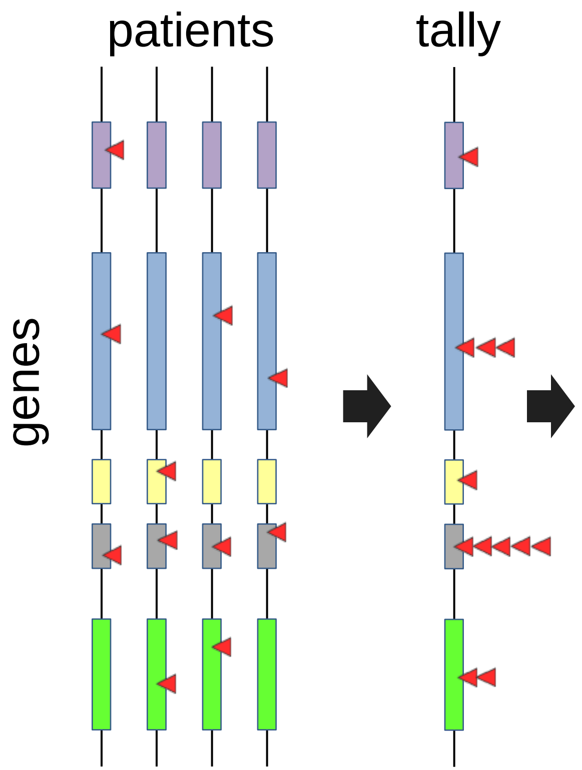
scoring algorithm version 1

assume background mutation rate is:

- · variable across sequence contexts
- · uniform across patients
- · uniform across genes

▼ C→T (UV-induced)

▼ A→T



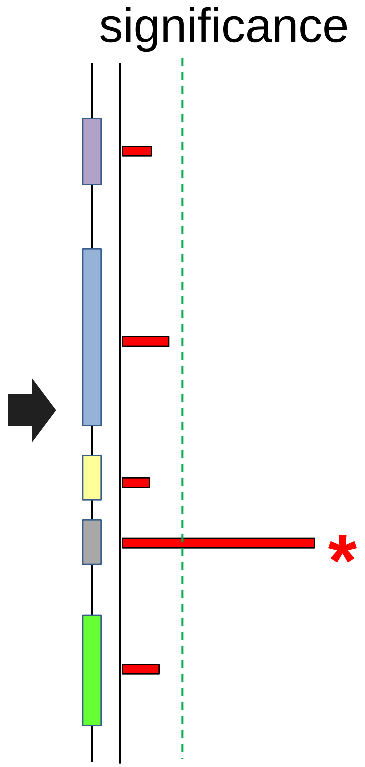
MutSig

scoring algorithm

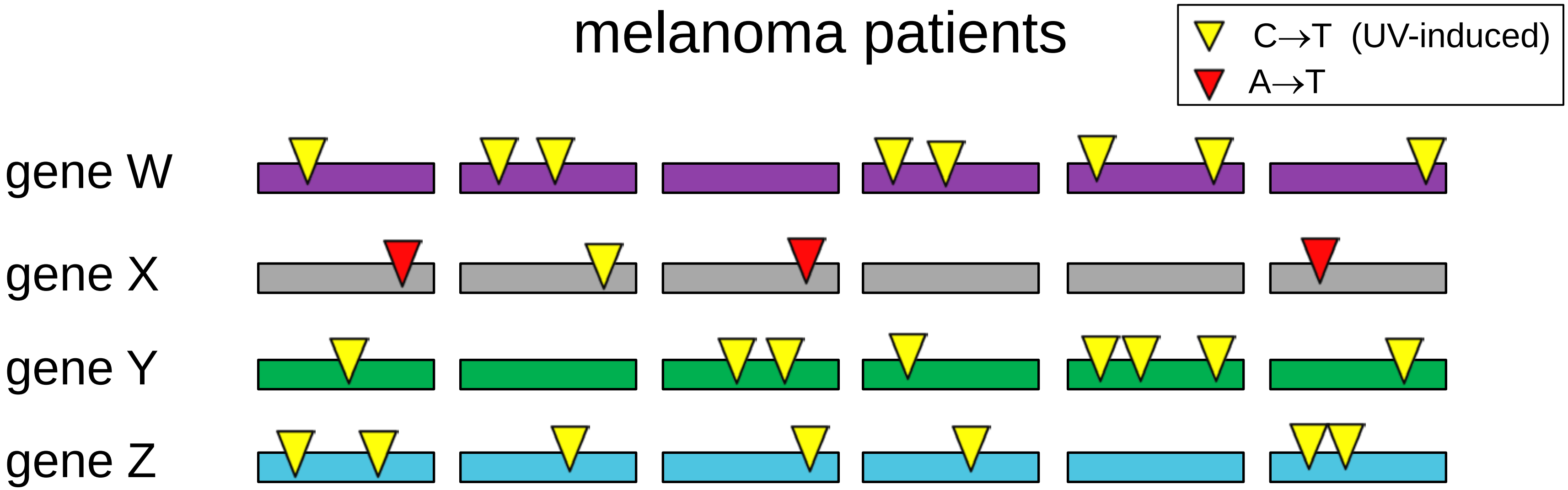
version 1

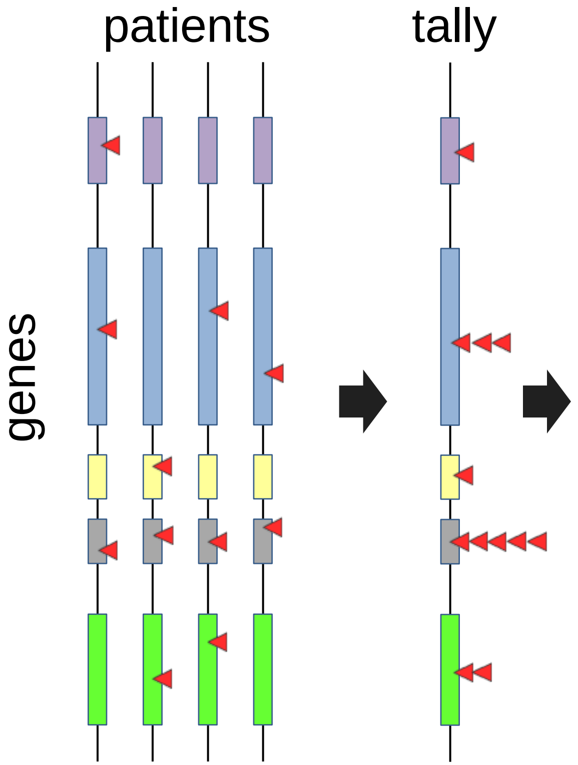
assume background mutation rate is:

- variable across sequence contexts
- uniform across patients
- uniform across genes



melanoma patients



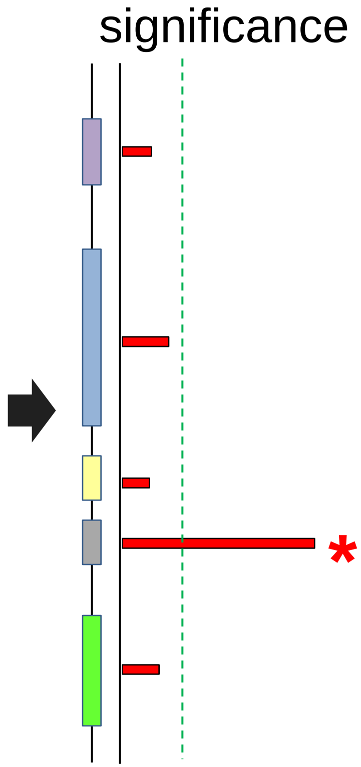


MutSig

scoring algorithm version 2

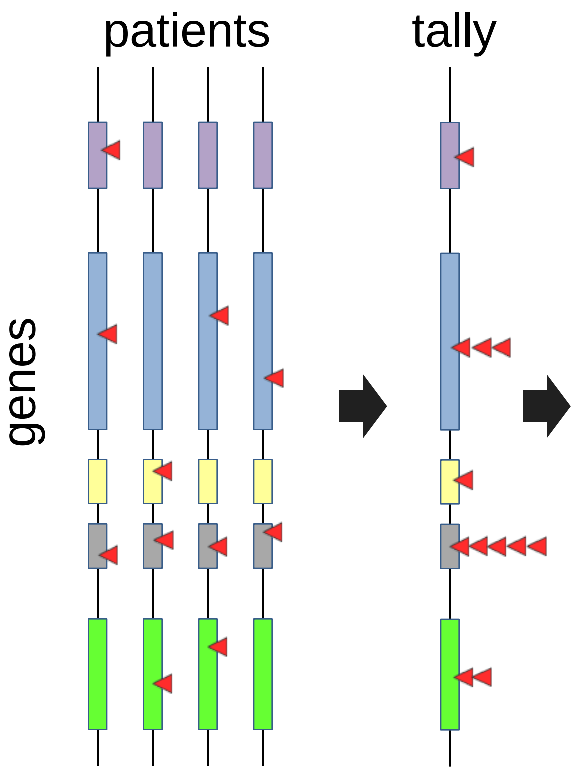
assume background mutation rate is:

- variable across sequence contexts
- variable across patients
- uniform across genes



patient 1
low mutation rate

patient 2
high mutation rate

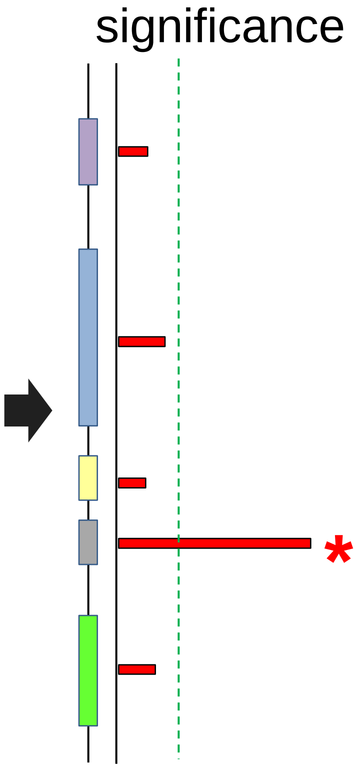


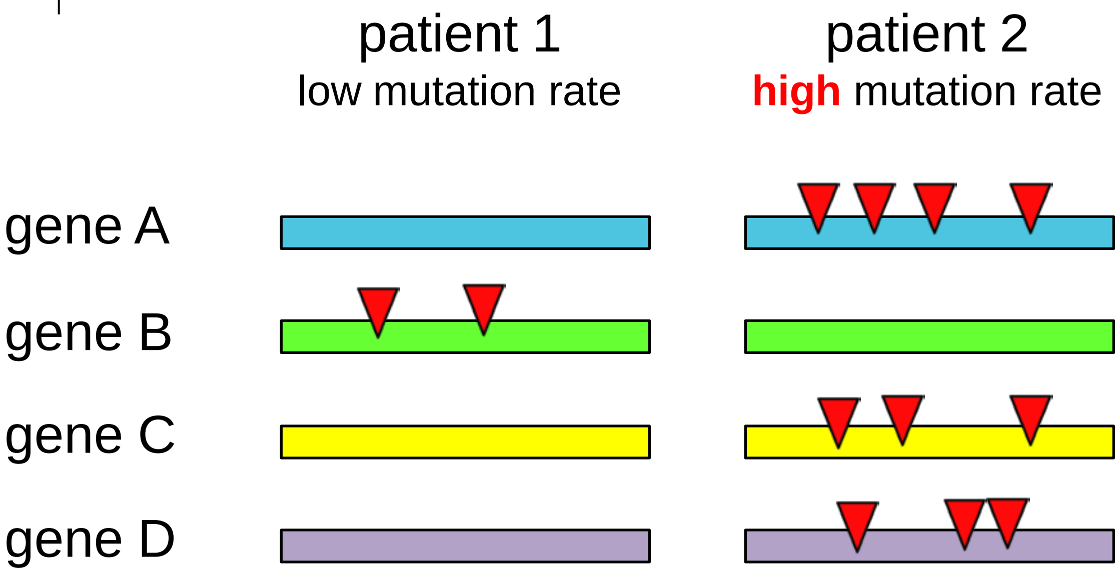
MutSig

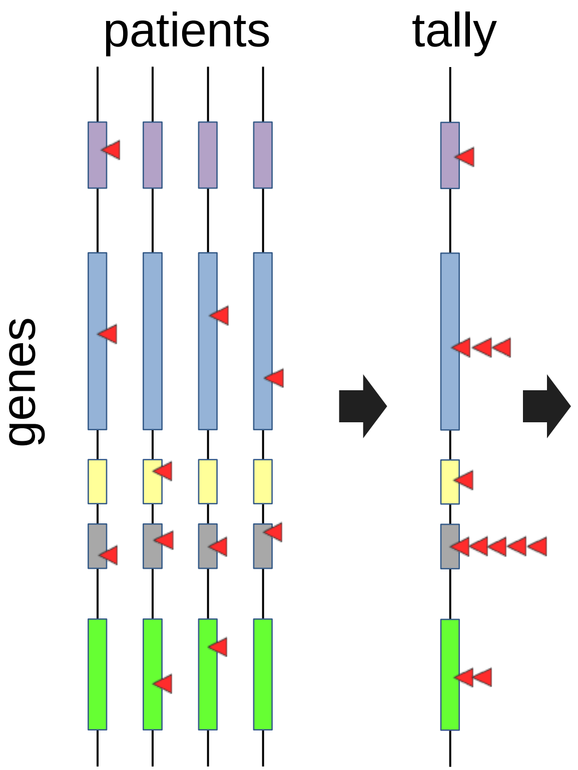
scoring algorithm version 2

assume background mutation rate is:

- · variable across sequence contexts
- · variable across patients
- · uniform across genes







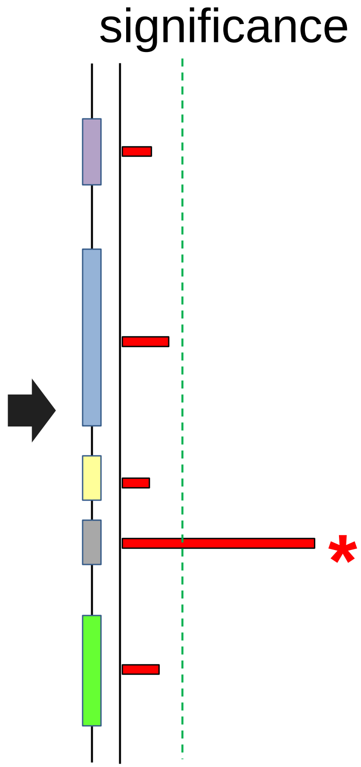
MutSig

scoring algorithm

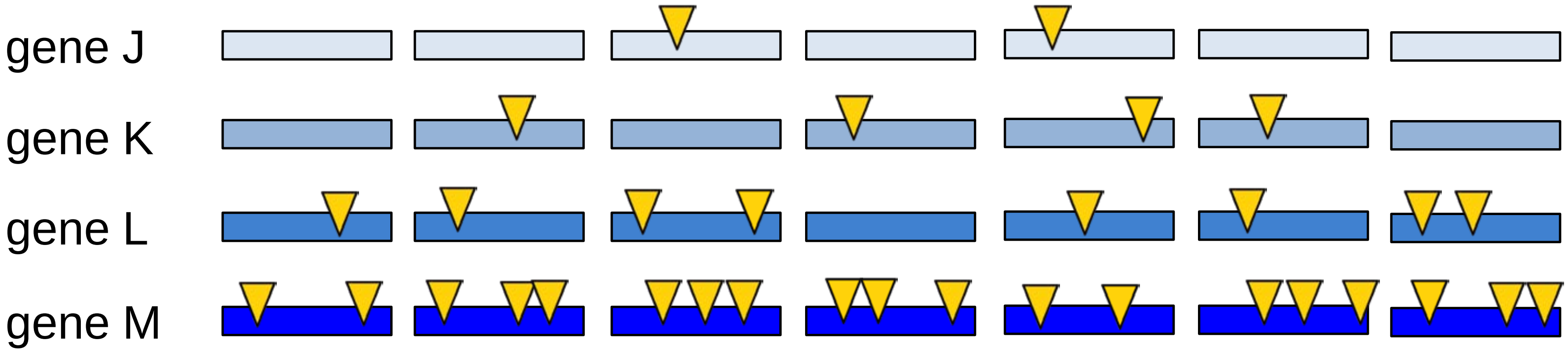
version 2

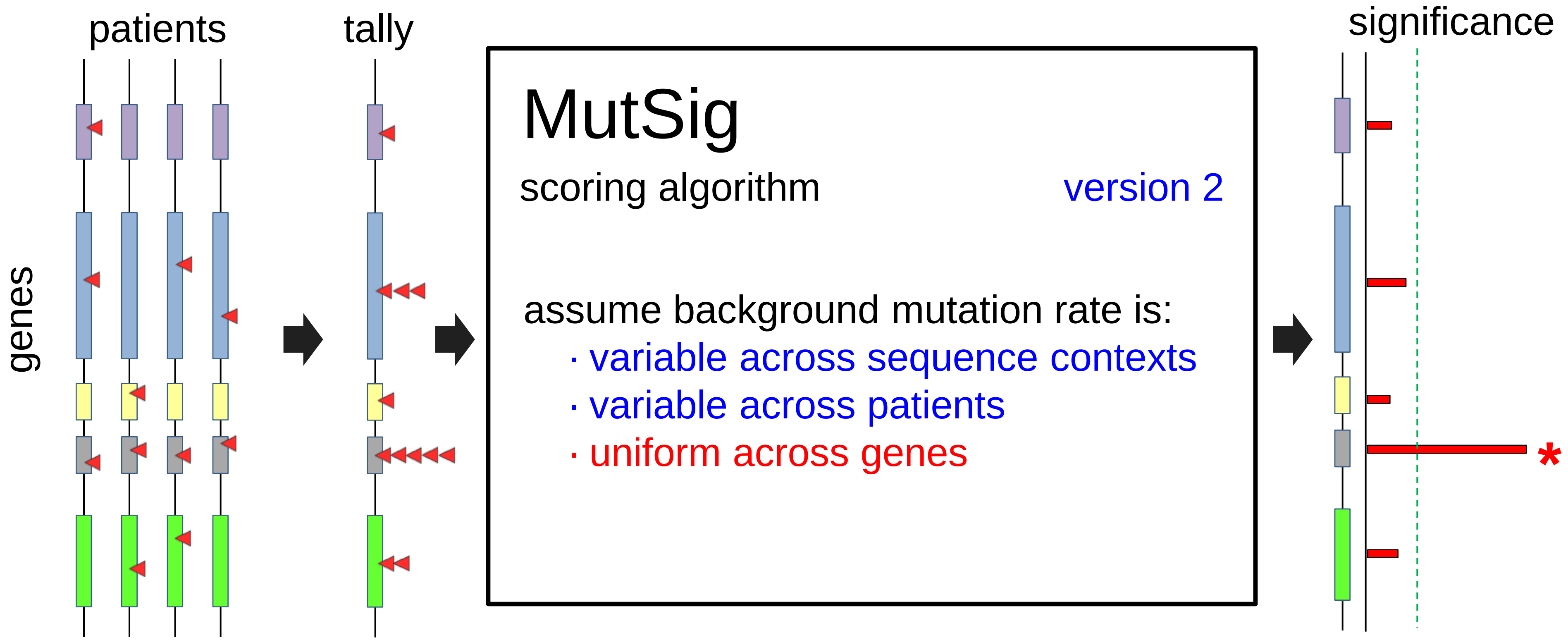
assume background mutation rate is:

- · variable across sequence contexts
- · variable across patients
- · uniform across genes



Problem: mutation rate is heterogeneous across genes





MutSig

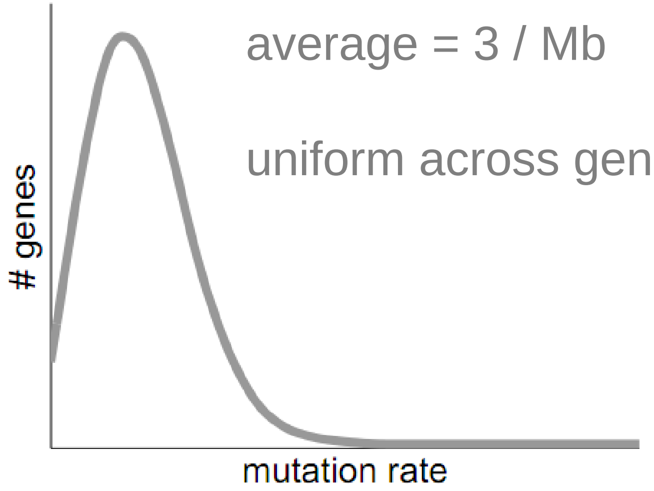
scoring algorithm

version 2

assume background mutation rate is:

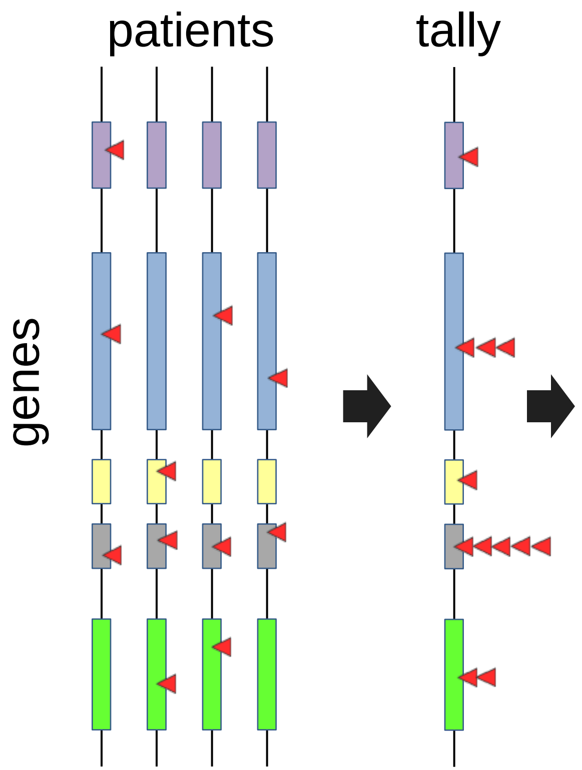
- variable across sequence contexts
- variable across patients
- uniform across genes

Problem: mutation rate is heterogeneous across genes



average = 3 / Mb

uniform across genes



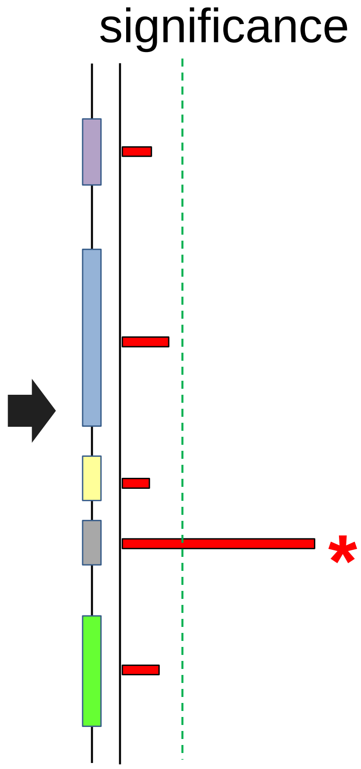
MutSig

scoring algorithm

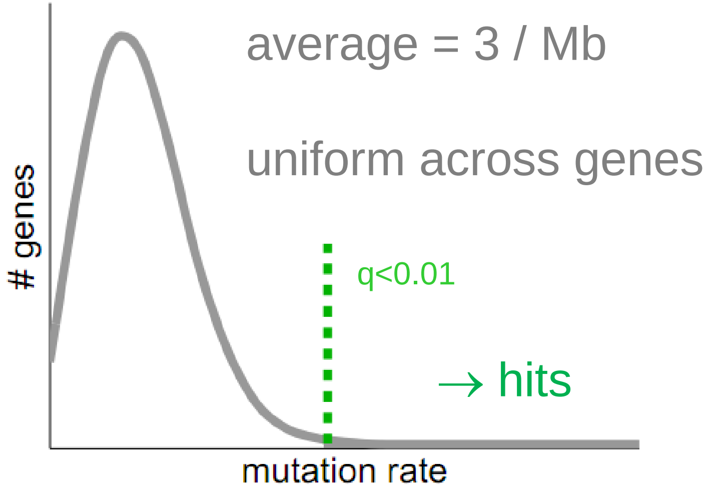
version 2

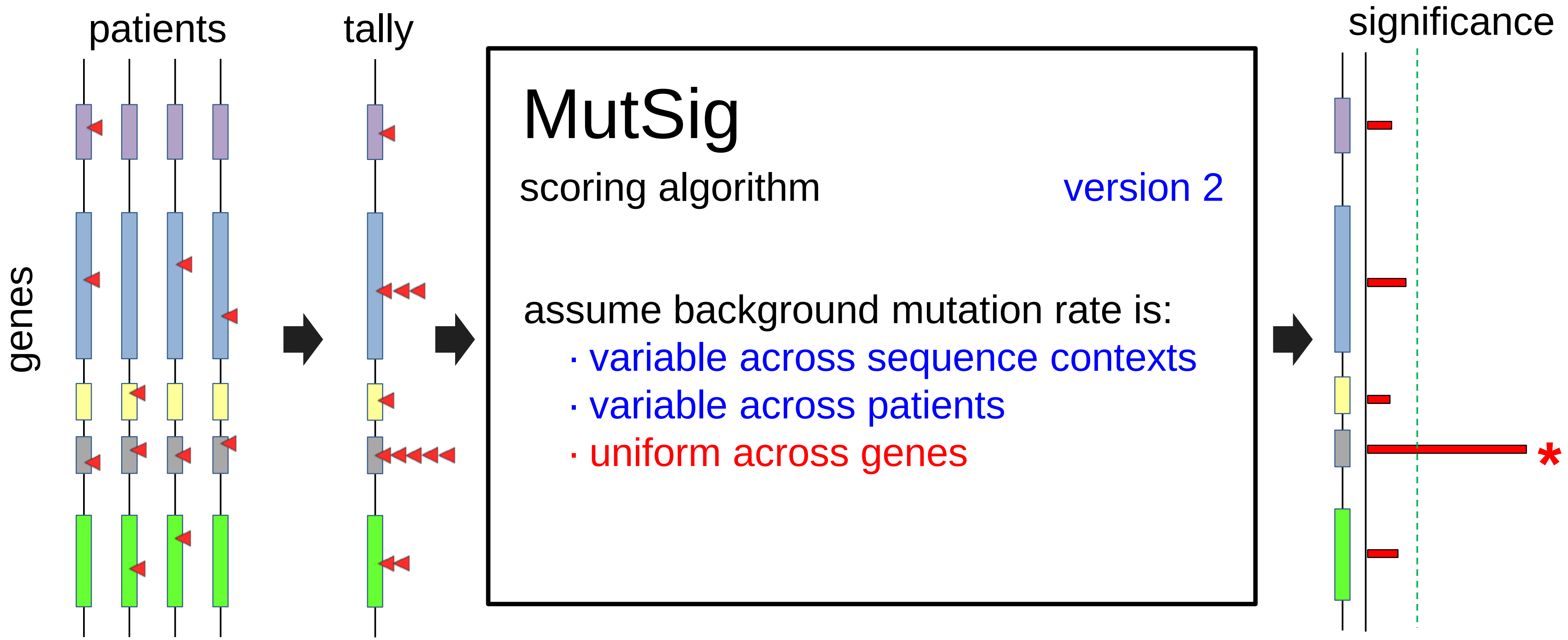
assume background mutation rate is:

- variable across sequence contexts
- variable across patients
- uniform across genes



Problem: mutation rate is heterogeneous across genes





MutSig

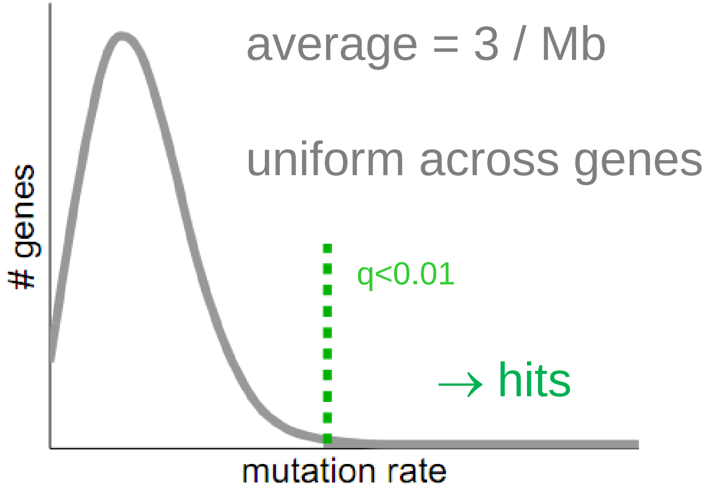
scoring algorithm

version 2

assume background mutation rate is:

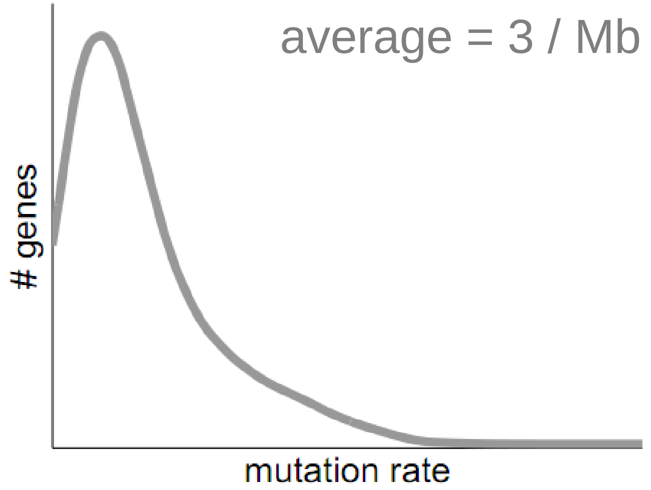
- variable across sequence contexts
- variable across patients
- uniform across genes

Problem: mutation rate is heterogeneous across genes

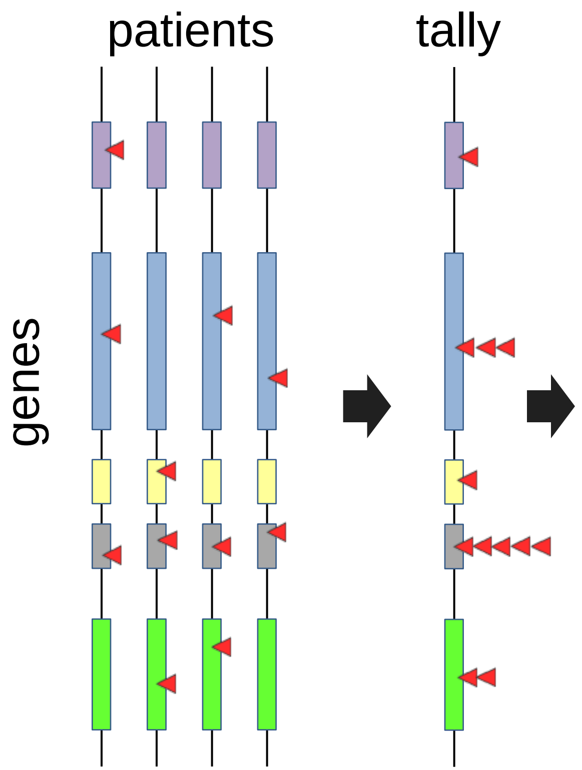


average = 3 / Mb

uniform across genes



average = 3 / Mb

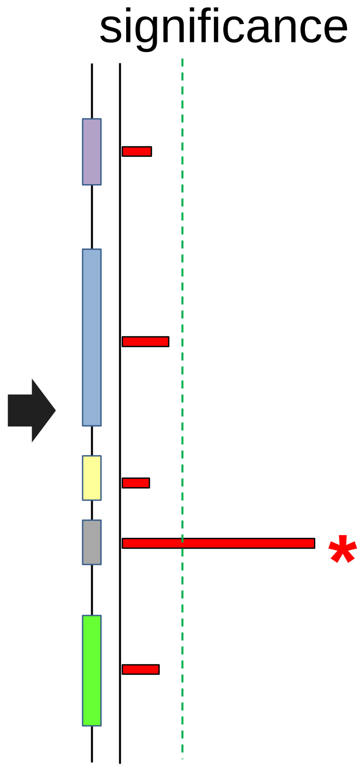


MutSig

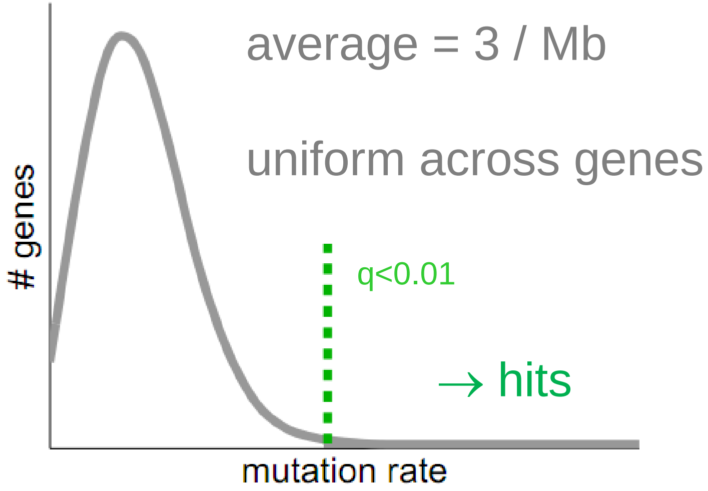
scoring algorithm version 2

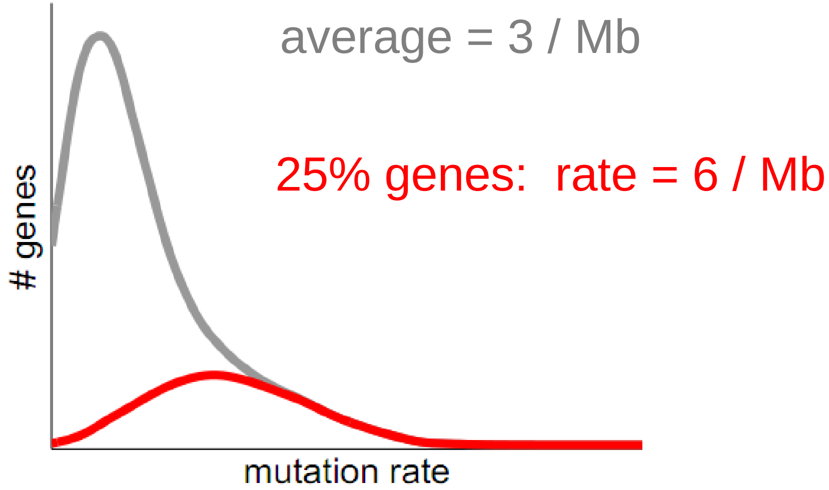
assume background mutation rate is:

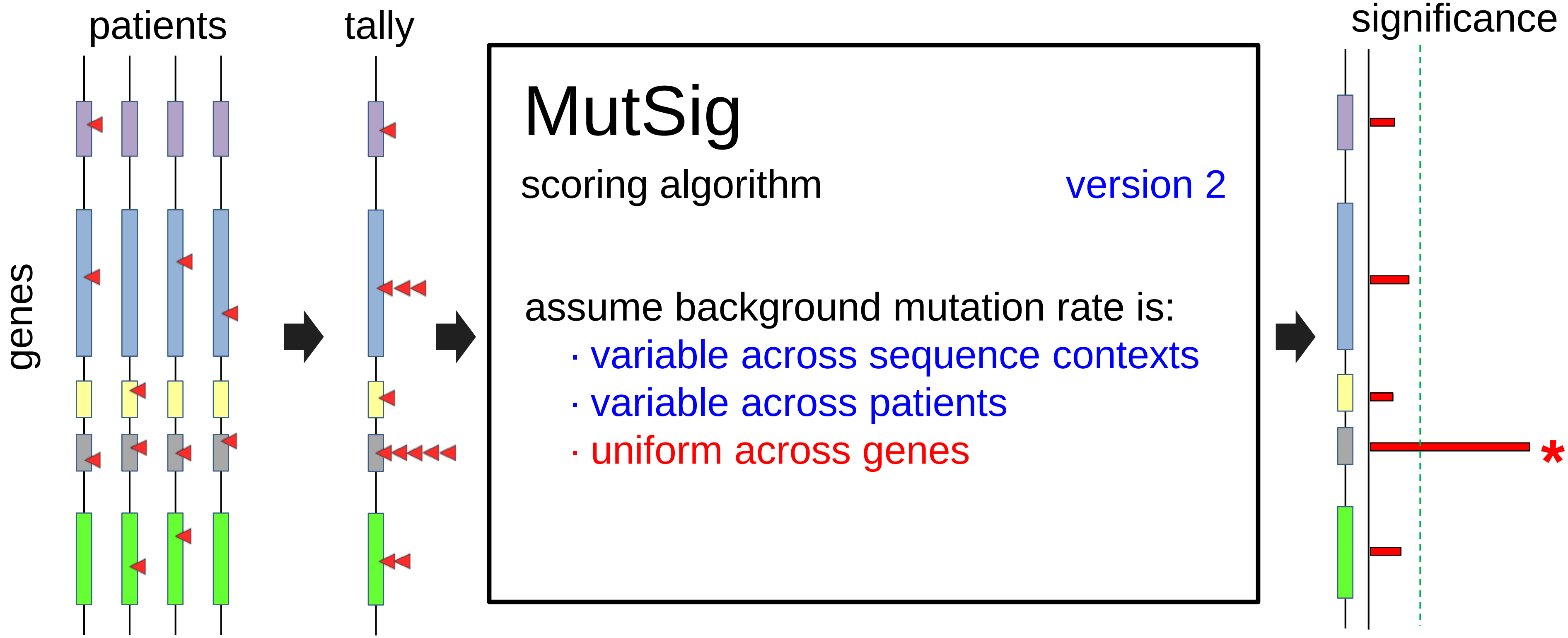
- variable across sequence contexts
- variable across patients
- uniform across genes



Problem: mutation rate is heterogeneous across genes







MutSig

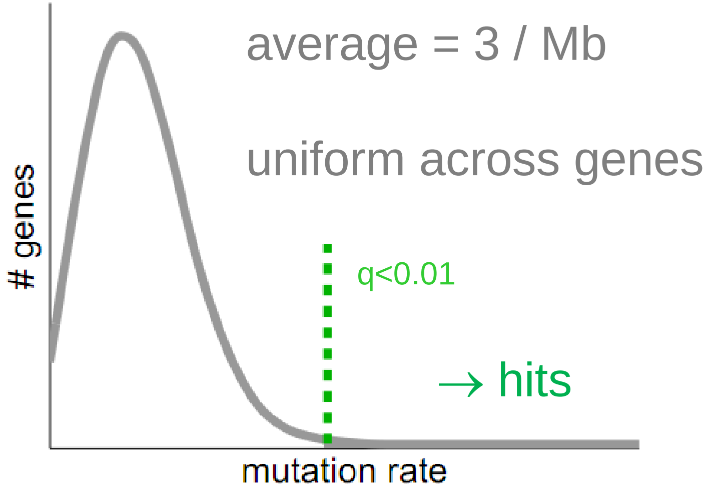
scoring algorithm

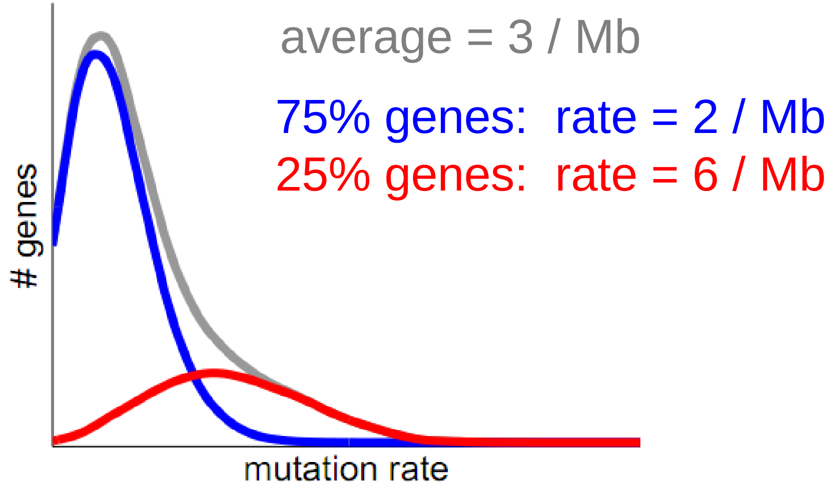
version 2

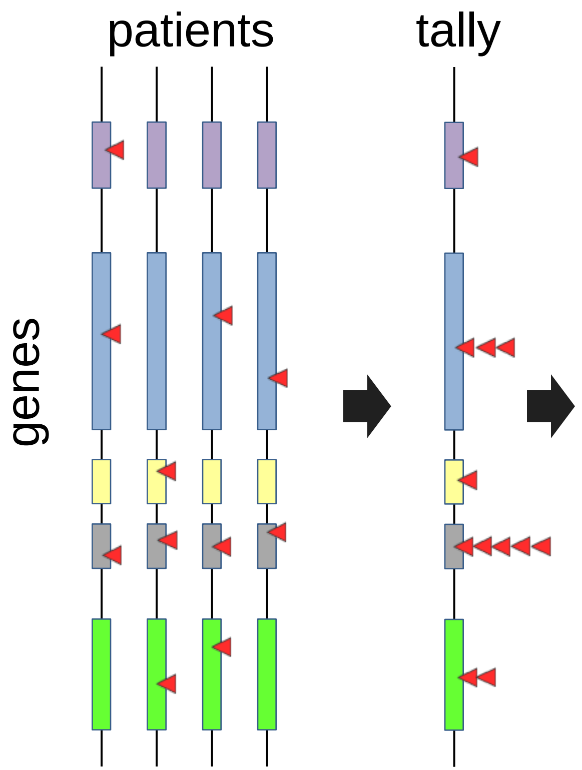
assume background mutation rate is:

- · variable across sequence contexts
- · variable across patients
- · uniform across genes

Problem: mutation rate is heterogeneous across genes







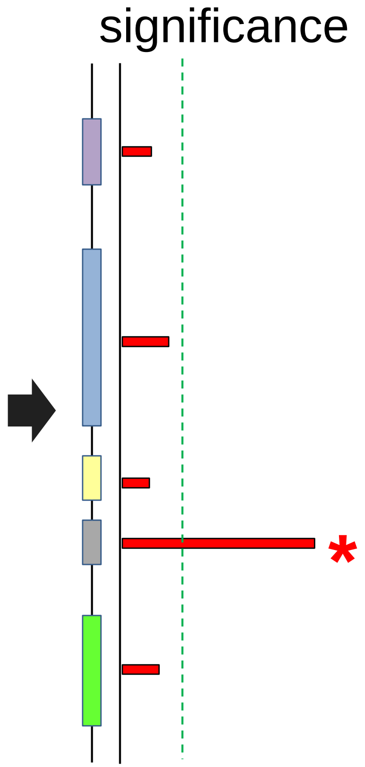
MutSig

scoring algorithm

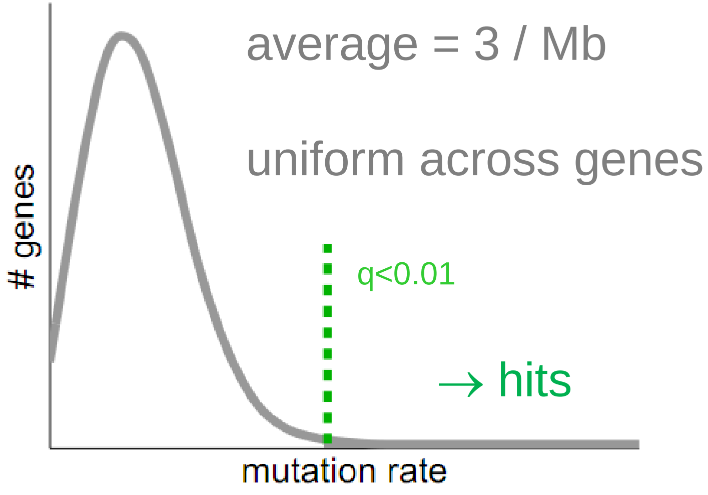
version 2

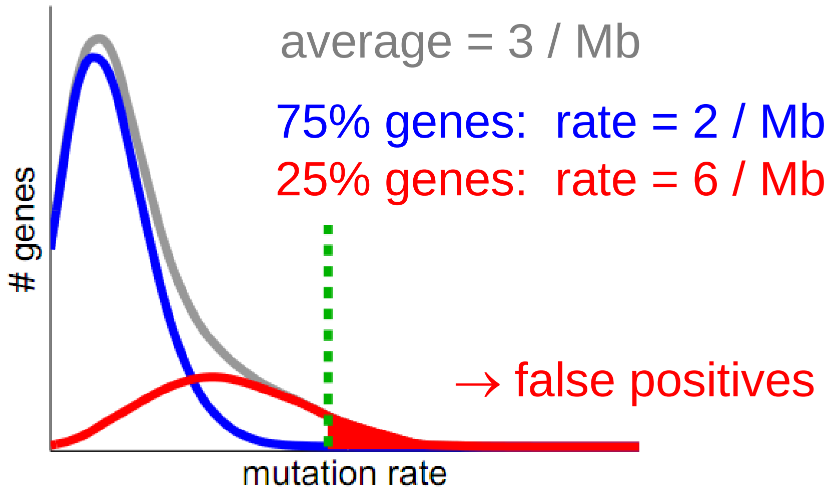
assume background mutation rate is:

- variable across sequence contexts
- variable across patients
- uniform across genes



Problem: mutation rate is heterogeneous across genes





Lung cancer

457 patients

180 lung squamous cell carcinoma
277 lung adenocarcinoma

average mutation rate = 10 / Mb

MutSig results version 0
(assuming uniform background mutation rate across genes)

all of these genes are extremely significant ($q<10^{-7}$)

total of
843 genes significantly mutated
(q<0.01)

* known lung cancer genes

#1	*	**TP53**
#2	*	**KRAS**
#7		OR4A15
#13	*	**KEAP1**
#14		OR8H2
#15	*	**STK11**
#17		OR2T4
#25		OR2T3
#31		OR2T6
#48		CSMD3
#49		OR5D16
#55		RYR2
#100		CSMD1
#139	*	**PIK3CA**
#158		RYR3
#159		MUC16
#161		OR2T33
#169	*	**NFE2L2**
#172		OR10G8
#180		OR2L8
#198		MUC17
#217		TTN

Bryan Hernandez
Peter Hammerman
Marcin Imielinski
Matthew Meyerson

Lung cancer

457 patients

180 lung squamous cell carcinoma
277 lung adenocarcinoma

average mutation rate = 10 / Mb

MutSig results version 0

(assuming uniform background mutation rate across genes)

all of these genes are extremely significant ($q<10^{-7}$)

total of 843 genes significantly mutated ($q<0.01$)

* known lung cancer genes

#1	*	**TP53**
#2	*	**KRAS**
#7		OR4A15
#13	*	**KEAP1**
#14		OR8H2
#15	*	**STK11**
#17		OR2T4
#25		OR2T3
#31		OR2T6
#48		CSMD3
#49		OR5D16
#55		RYR2
#100		CSMD1
#139	*	**PIK3CA**
#158		RYR3
#159		MUC16
#161		OR2T33
#169	*	**NFE2L2**
#172		OR10G8
#180		OR2L8
#198		MUC17
#217		TTN

"fishy" genes

olfactory receptors

(146 with $q<0.01$)

"cub and sushi" proteins

reported to be tumor suppressors but significantly mutated in almost every tumor type (including TCGA ovarian)

ryanodine receptors

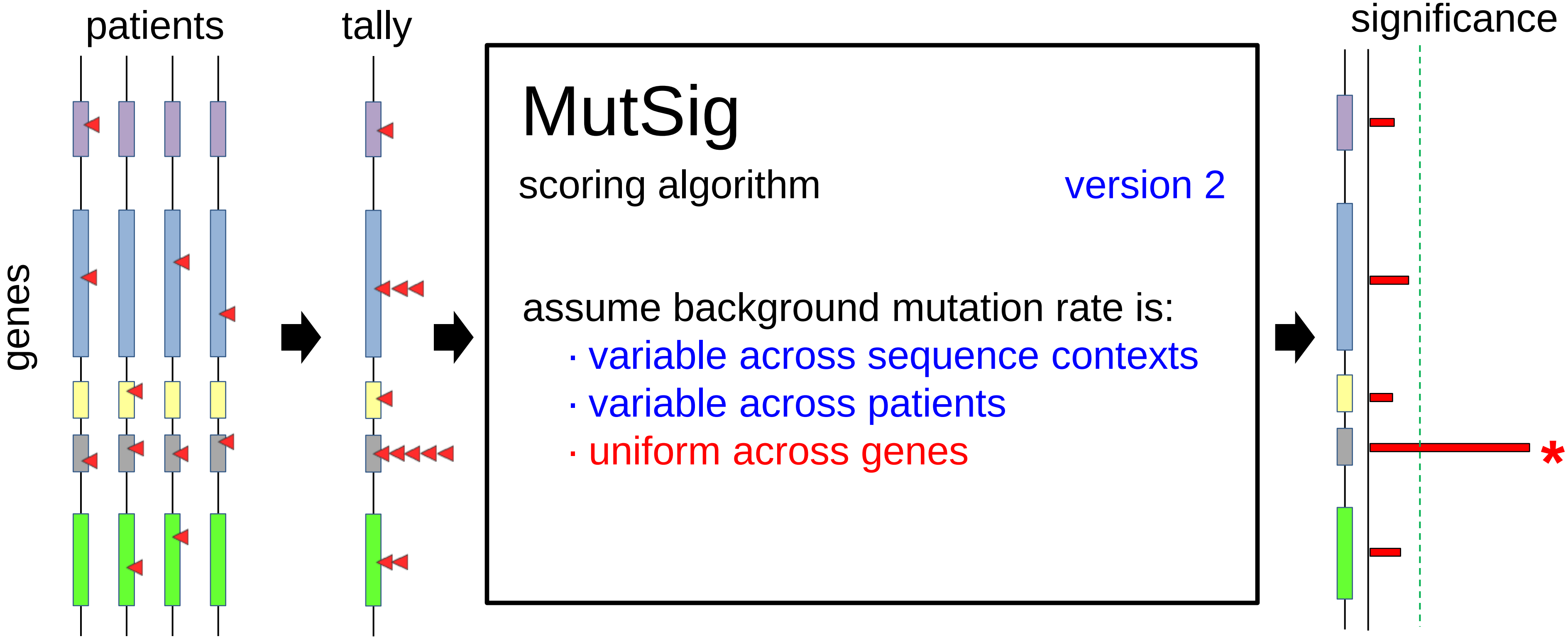
cardiac calcium channels

mucins

gel-forming proteins

titin

largest human protein
100x bigger than p53
34,350 amino acids
100 Kb coding sequence



MutSig

scoring algorithm

version 2

assume background mutation rate is:

- variable across sequence contexts
- variable across patients
- uniform across genes

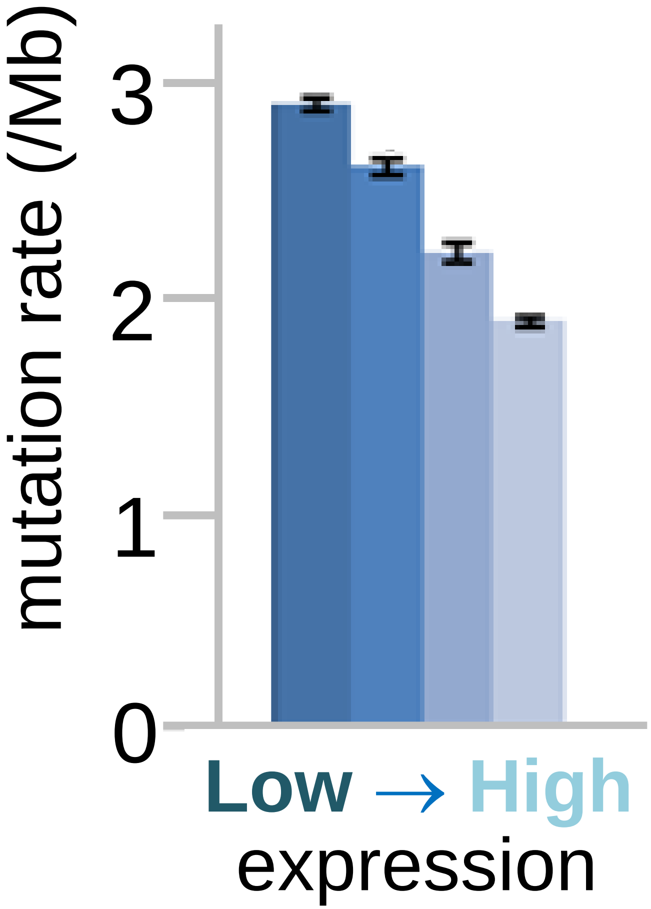
Problem: mutation rate is actually heterogeneous across genes

Challenge: predict gene-specific background mutation rates

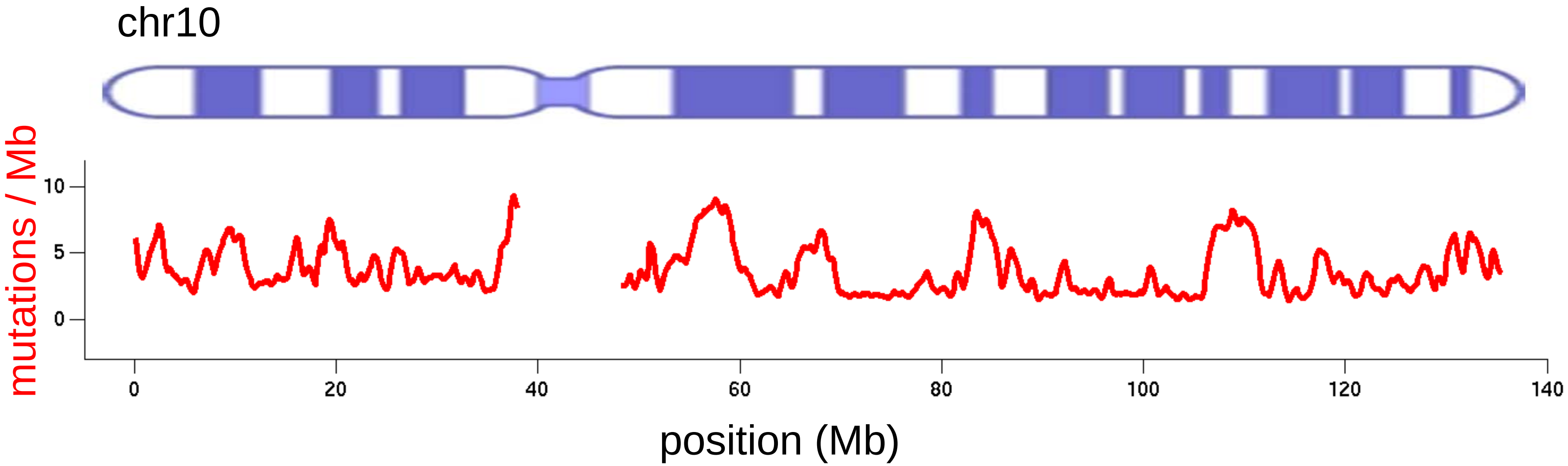
We eventually want to learn the background mutation rate of every gene
(and all possible mutations at all basepairs!)

As we sequence more and more samples, we get closer to this goal.

Highly expressed genes have lower mutation rates



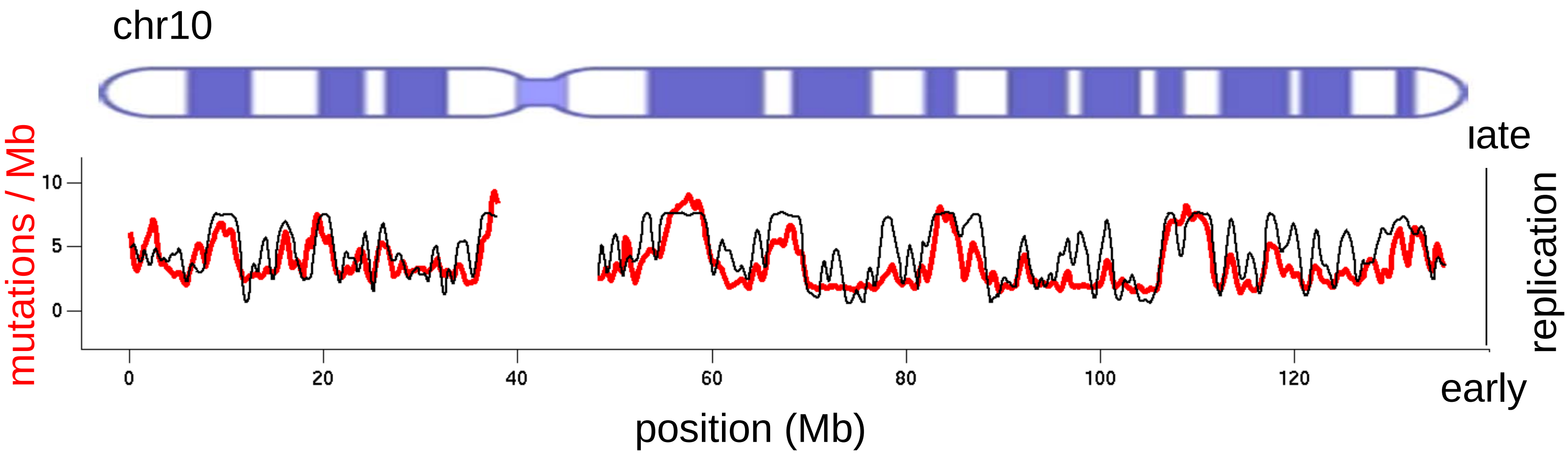
Chapman et al. *Nature* (2011)



background mutation rate
varies ten-fold or more
across the genome

shown:
noncoding mutation rate
from TCGA lung cancer dataset

Early-replicating genes have lower mutation rates



background mutation rate varies ten-fold or more across the genome

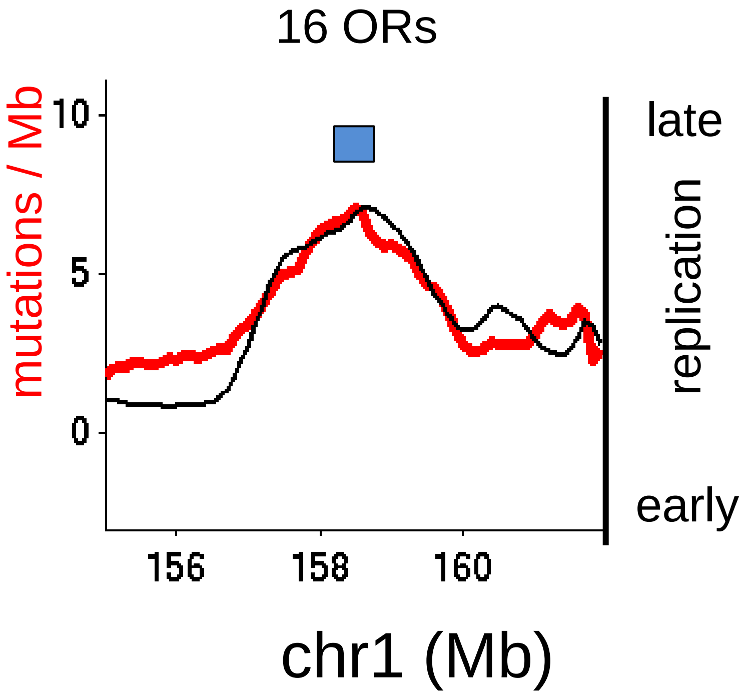
highly correlated

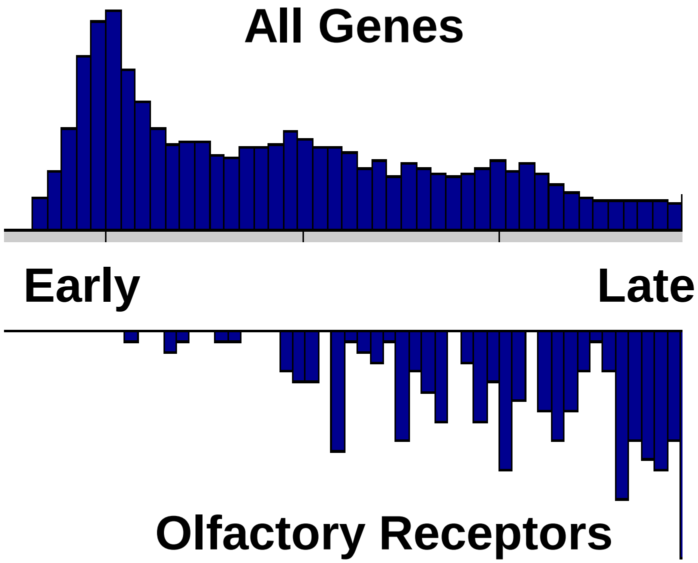
replication time also varies greatly across the genome

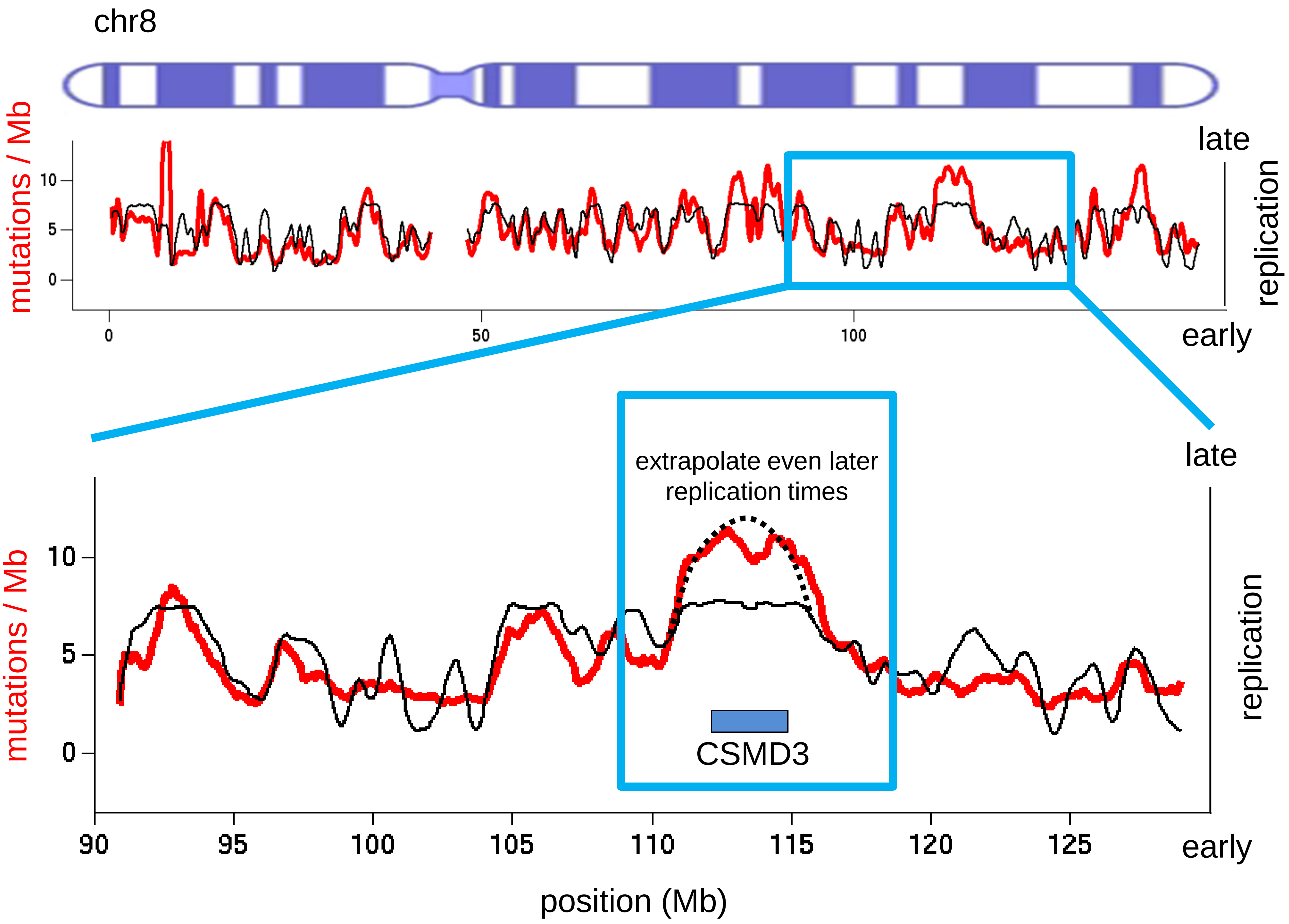
shown:
noncoding mutation rate
from TCGA lung cancer dataset

Sunyaev Lab (Harvard/BWH)
Stamatoyannopoulos et al. (2009) *Nat. Gen.*
shown: replication time measurements from
Chen et al. (2010) *Genome Research* 20:447

Late replication explains most olfactory receptors

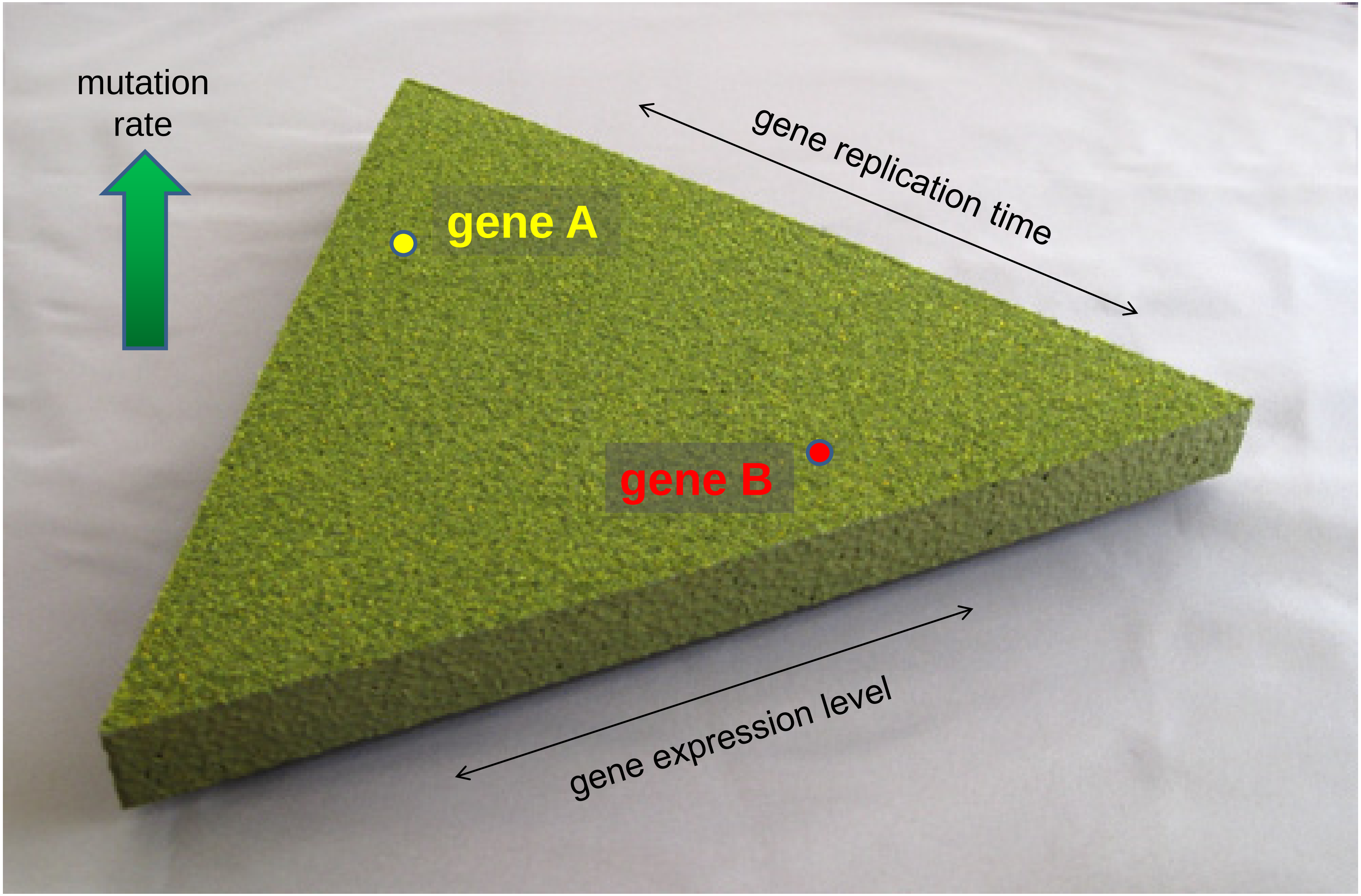




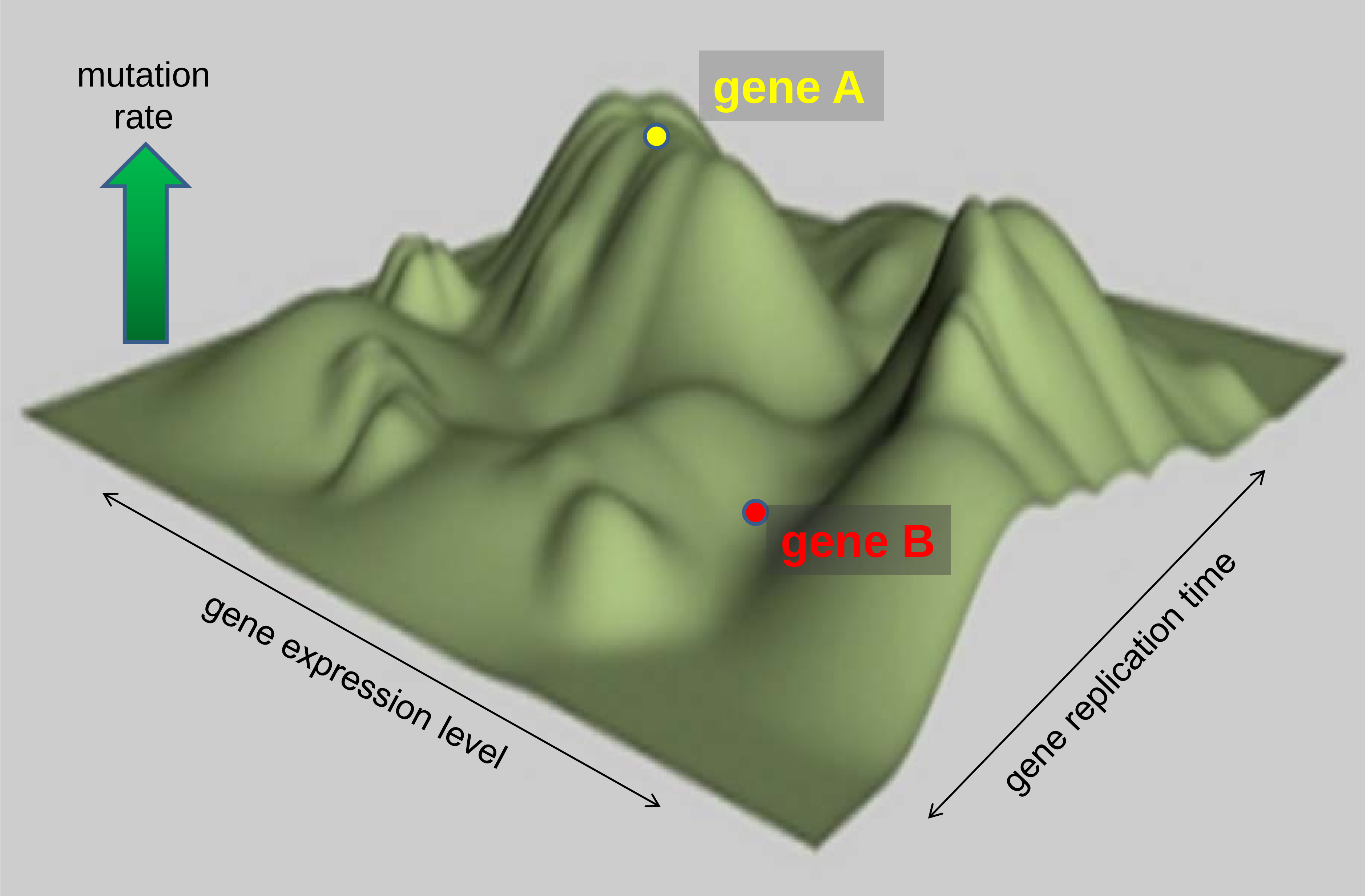


extrapolate even later replication times

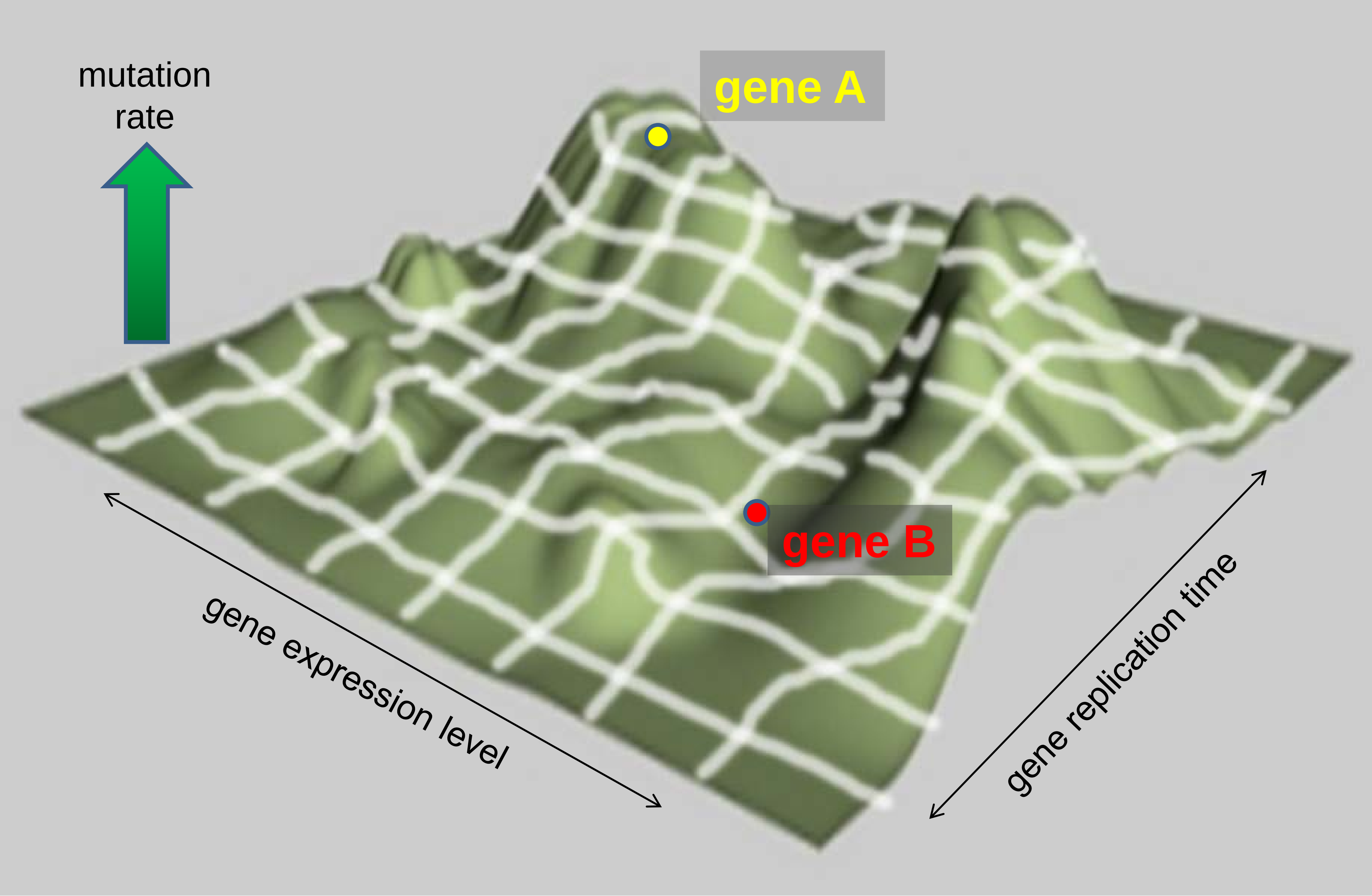
initial model assumed a flat mutational landscape



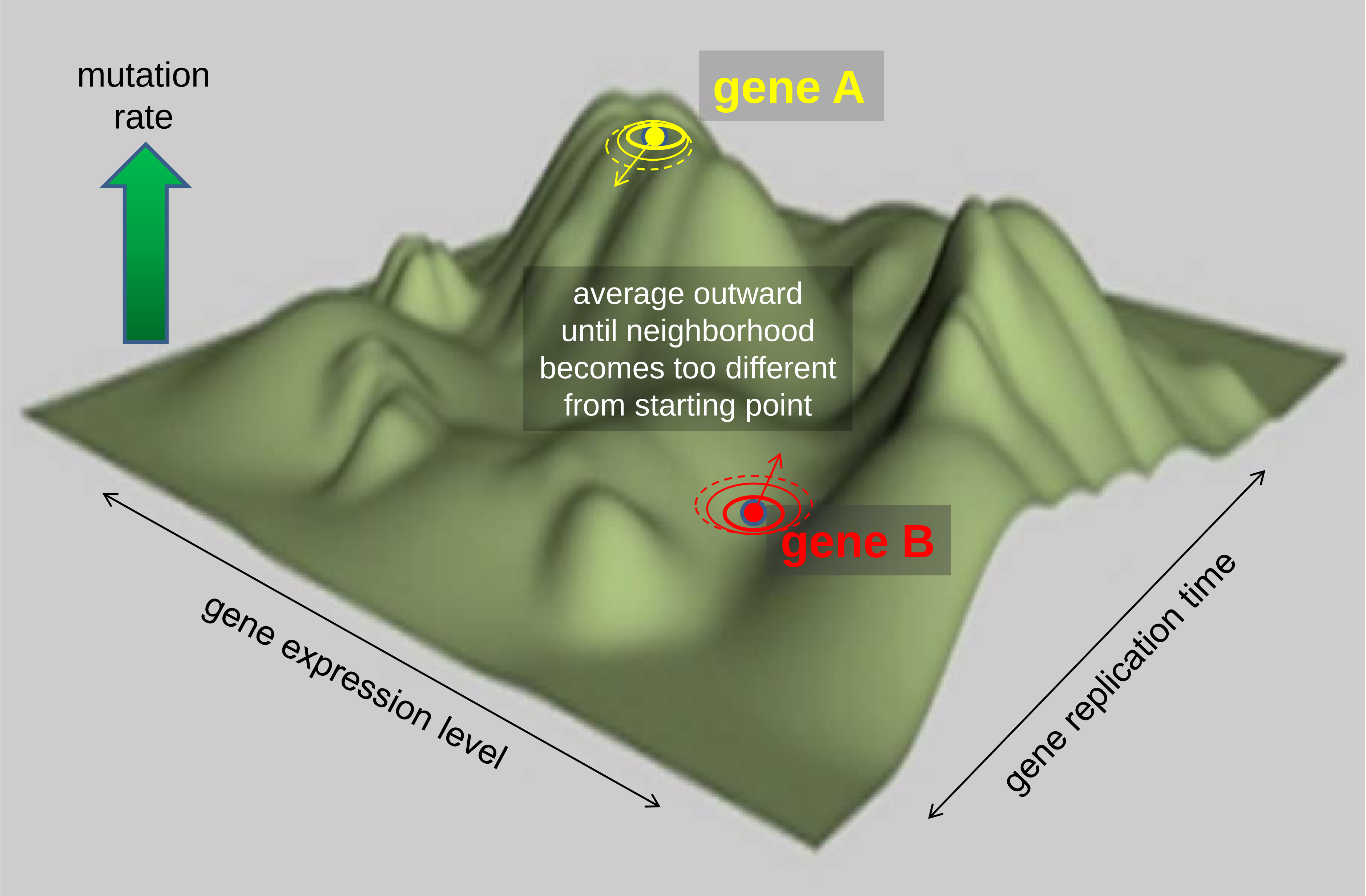
landscape is actually *not* flat



improve estimate by binning together similar genes...



...or by local regression


mutation rate
gene A
average outward until neighborhood becomes too different from starting point
gene B
gene expression level
gene replication time

Lung cancer

* known lung cancer genes

"fishy" genes

MutSig v0
assuming uniform
bkgd mutation rate
across all genes

$q<10^{-7}$

#1	*	**TP53**
#2	*	**KRAS**
#7		OR4A15
#13	*	**KEAP1**
#14		OR8H2
#15	*	**STK11**
#17		OR2T4
#25		OR2T3
#31		OR2T6
#48		CSMD3
#49		OR5D16
#55		RYR2
#100		CSMD1
#139	*	**PIK3CA**
#158		RYR3
#159		MUC16
#161		OR2T33
#169	*	**NFE2L2**
#172		OR10G8
#180		OR2L8
#198		MUC17
#217		TTN

843 genes
significantly mutated
($q<0.01$)

Lung cancer

* known lung cancer genes

"fishy" genes

MutSig v0

assuming uniform
bkgd mutation rate
across all genes

#1	*	TP53	$q<10^{-7}$
#2	*	KRAS	
#7		OR4A15	
#13	*	KEAP1	
#14		OR8H2	
#15	*	STK11	
#17		OR2T4	
#25		OR2T3	
#31		OR2T6	
#48		CSMD3	
#49		OR5D16	
#55		RYR2	
#100		CSMD1	
#139	*	PIK3CA	
#158		RYR3	
#159		MUC16	
#161		OR2T33	
#169	*	NFE2L2	
#172		OR10G8	
#180		OR2L8	
#198		MUC17	
#217		TTN	

843 genes
significantly mutated
($q<0.01$)

improved MutSig

using gene-specific
background mutation rates

*	STK11	#1	$q<10^{-5}$
*	NFE2L2	#4	
*	TP53	#7	
*	KRAS	#8	
*	KEAP1	#11	
*	PIK3CA	#12	

52 genes
significantly mutated
($q<0.01$)

. . . .

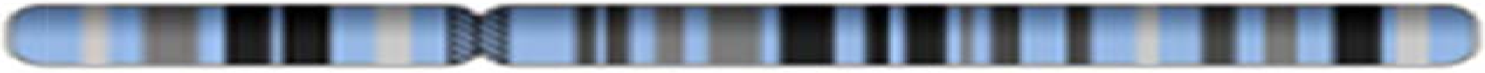
*	OR8H2	#181	q~0.2
	OR5T2	#276	
	OR10J3	#334	
	CSMD3	#388	
	MUC17	#2614	q=1
	RYR2	#2898	
	CSMD1	#4482	
	TTN	#4825	
	MUC16	#5650	
	RYR3	#11496	

*most significant
olfactory receptor

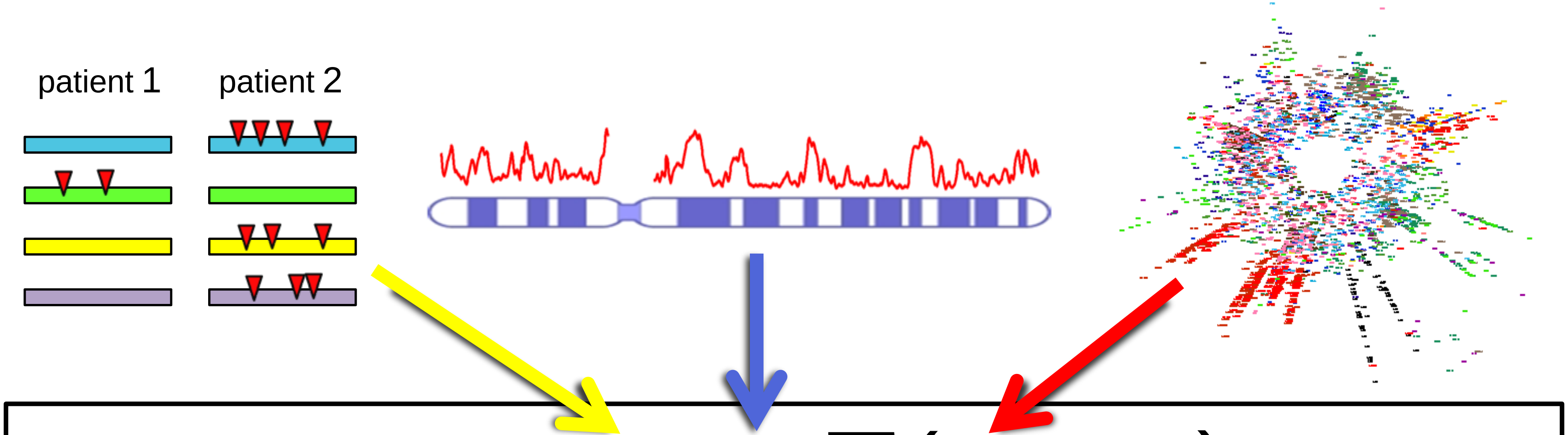
Correcting for variation in mutation rate

	Before	After
Lung Squamous	261 (50 OR)	18 (0 OR)
Lung Adeno	511 (93 OR)	33 (1 OR)
Melanoma	177 (7 OR)	61 (0 OR)
Prostate	3 (0 OR)	3 (0 OR)
DLBCL	32 (1 OR)	15 (0 OR)

Ultimate solution: Learn the background rate



putting it all together



$$\mu_o \cdot F_{p,s,c} \;=\; \mu_o \cdot F_p \cdot F_g \sum_k \left(w_{p,k} \cdot v_{k,c} \right) \;=\; \mu_{p,s,c}$$

overall mutation rate across entire dataset

relative mutation rate in
gene *g*
patient *p*
category *c*

relative mutation rate of patient *p*

relative mutation rate of gene *g*

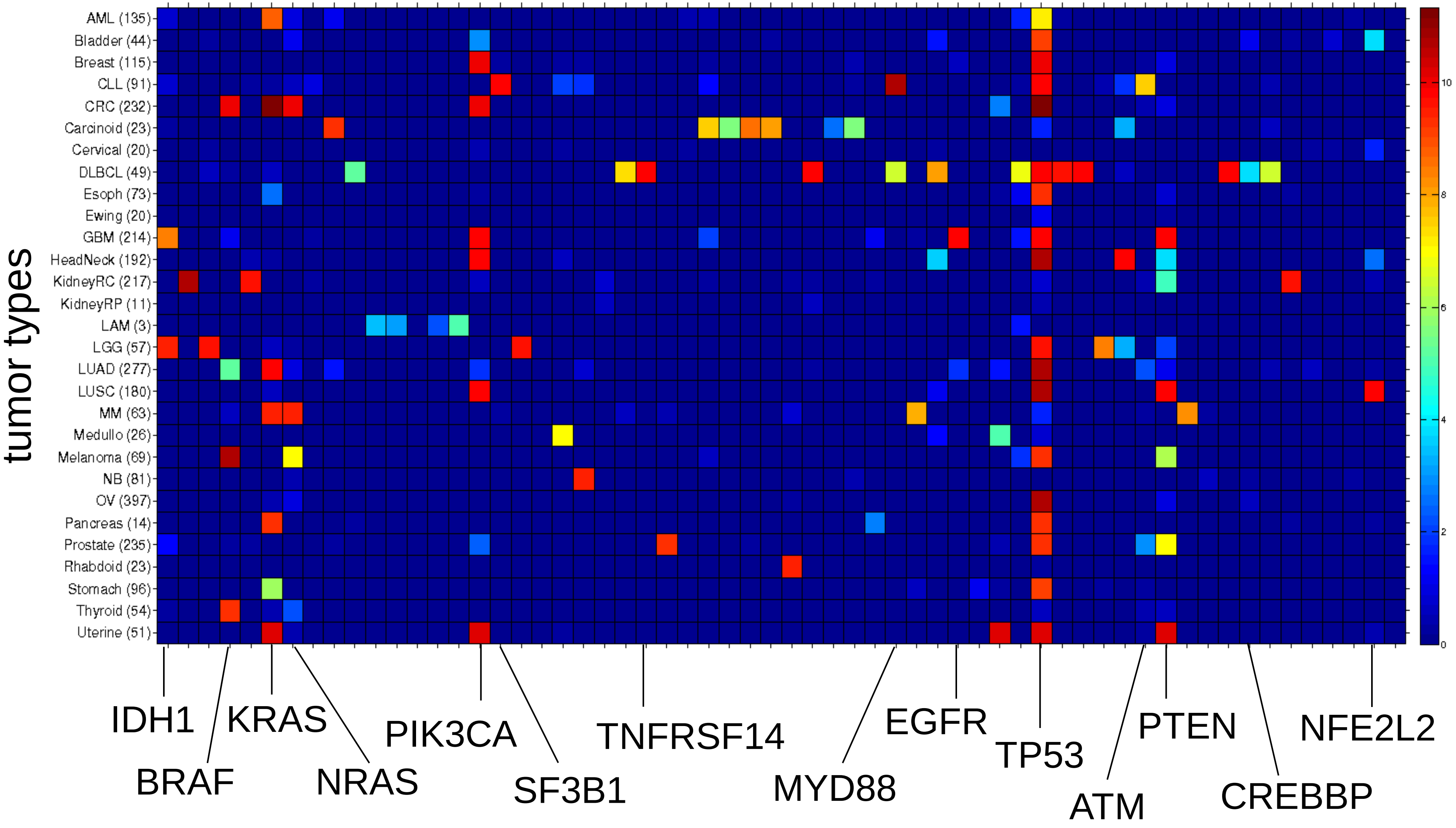
sum across *k* "factors" (i.e. mutational processes)

weight of factor *k* in patient *p*

contribution of factor *k* to mutation category *c*

mutation rate in
gene *g*
patient *p*
category *c*

significantly mutated genes across tumor types



Acknowledgements

MutSig Team

Petar Stojanov
Bryan Hernandez
Marcin Imielinski
Peter Hammerman
Gregory Kryukov
Eran Hodis
Chip Stewart

Analysis Team

Kristian Cibulskis
Andrey Sivachenko
Scott Carter
Gordon Saksena
Yotam Drier
Alex Ramos
Aaron McKenna
Rui Jing
Lihua Zou
David DeLuca
Elena Helman
Jaegil Kim
Cheng-Zhong Zhang
Sylvan Baca
Trevor Pugh
Nam Pho
Andrew Cherniak
Alex Kostic
Peter Chen
Nikhil Wagle

(cont.)

Mike Berger
Mike Chapman
Jens Lohr
Derek Chiang
Craig Mermel
Nicolas Stransky
Roel Verhaak
Barbara Weir
Shantanu Banerji

Firehose

Doug Voet
Mike Noble
Pei Lin
Dan DiCara
Lee Lichtenstein
Robert Zupko
Peter Carr

NHGRI NCI

THE CANCER GENOME ATLAS

Gaddy Getz

Stacey Gabriel **Matthew Meyerson**

Levi Garraway **Eric Lander**

Lynda Chin **Todd Golub**

Broad Institute of Harvard and MIT

Dana Farber Cancer Institute

Shamil Sunyaev
Paz Polak
Leonid Mirny

Rameen Beroukhim
Steve Schumacher
Barbara Tabak

Cathy Wu
Jennifer Brown
Lili Wang
Youzhong Wan
Dan-Avi Landau
Aviv Regev
Nathalie Pochet

Adam Bass
Austin Dulak

Project Management

Carrie Sougnez
Erica Shefler
Elizabeth Nickerson
Daniel Auclair
Marisa Cortes
Kristin Thompson

Jim Robinson
Helga Thorvaldsdottir
Marc-Danie Nazaire
Jill Mesirov

Broad Institute Sequencing Program and Platform

Tim Fennell
Sara Chauvin
Lauren Ambrogio
Sheila Fisher
Joshua Levin
Xian Adiconis
Andreas Gnirke
David Jaffe
Toby Bloom
Chad Nusbaum

Broad Institute Biological Genetic Analysis Platform

Robb Onofrio
Brendan Blumenstiel
Huy Nguyen
Mellisa Parkin
Wendy Winckler

Broad Institute Biological Samples Platform

Kristin Ardlie

Mark DePristo
Eric Banks
Kiran Garimella