ゲノムの大規模データを解析する

— 転写制御領域の解読と設計 —

矢田 哲士

九州工業大学大学院情報工学研究院

ytetsu@bio.kyutech.ac.jp

研究グループ

鈴木 穣　　(東大・新領域)

入江 拓磨　(東大・新領域)

谷口 丈晃　(三菱総研)

Our data (Irie *et al.* in prep.)

ヒトプロモーター (~1,100-nt)

EF1a1，GAPDH，DDX5

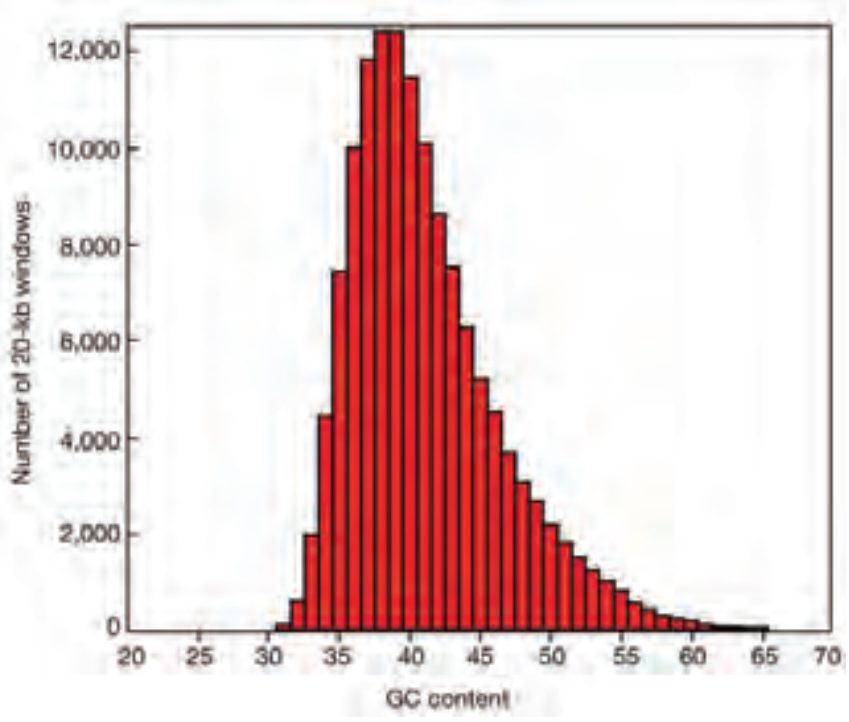
39,000，28,000，16,000 muts / wt

突然変異率: 1.64, 1.59, 1.82 %

置換 (92%)，**挿入** (1%)，**欠失** (7%) **を導入**

HEK293 **での転写強度を測定**

なぜ EF1a1，GAPDH，DDX5？



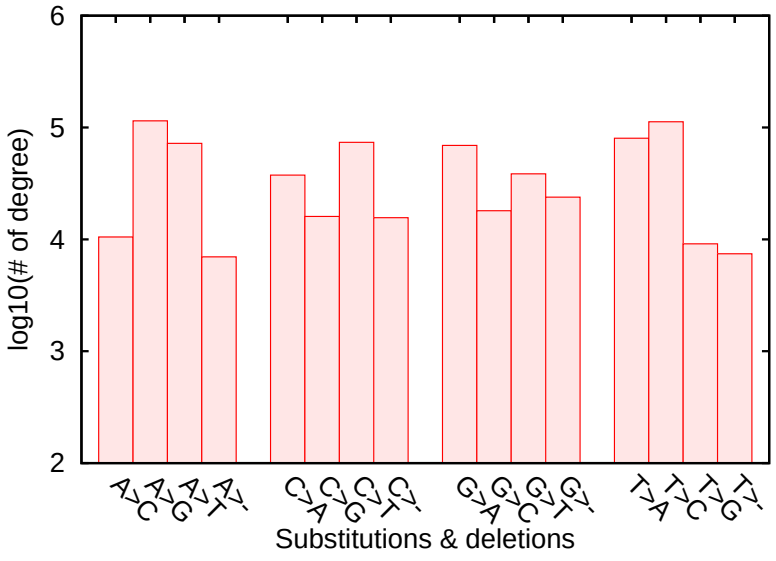
(IHGC 2001)

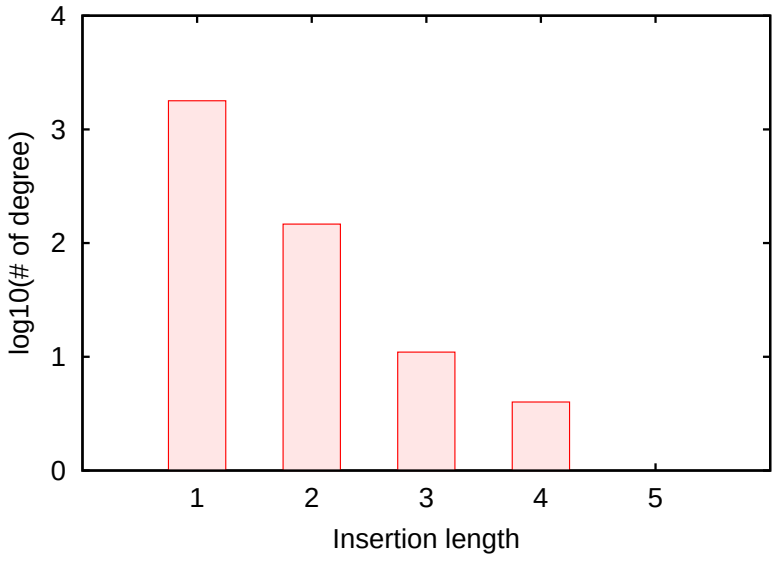
GC **含量**: 59.3, 64.9, 63.3 %

TATA box

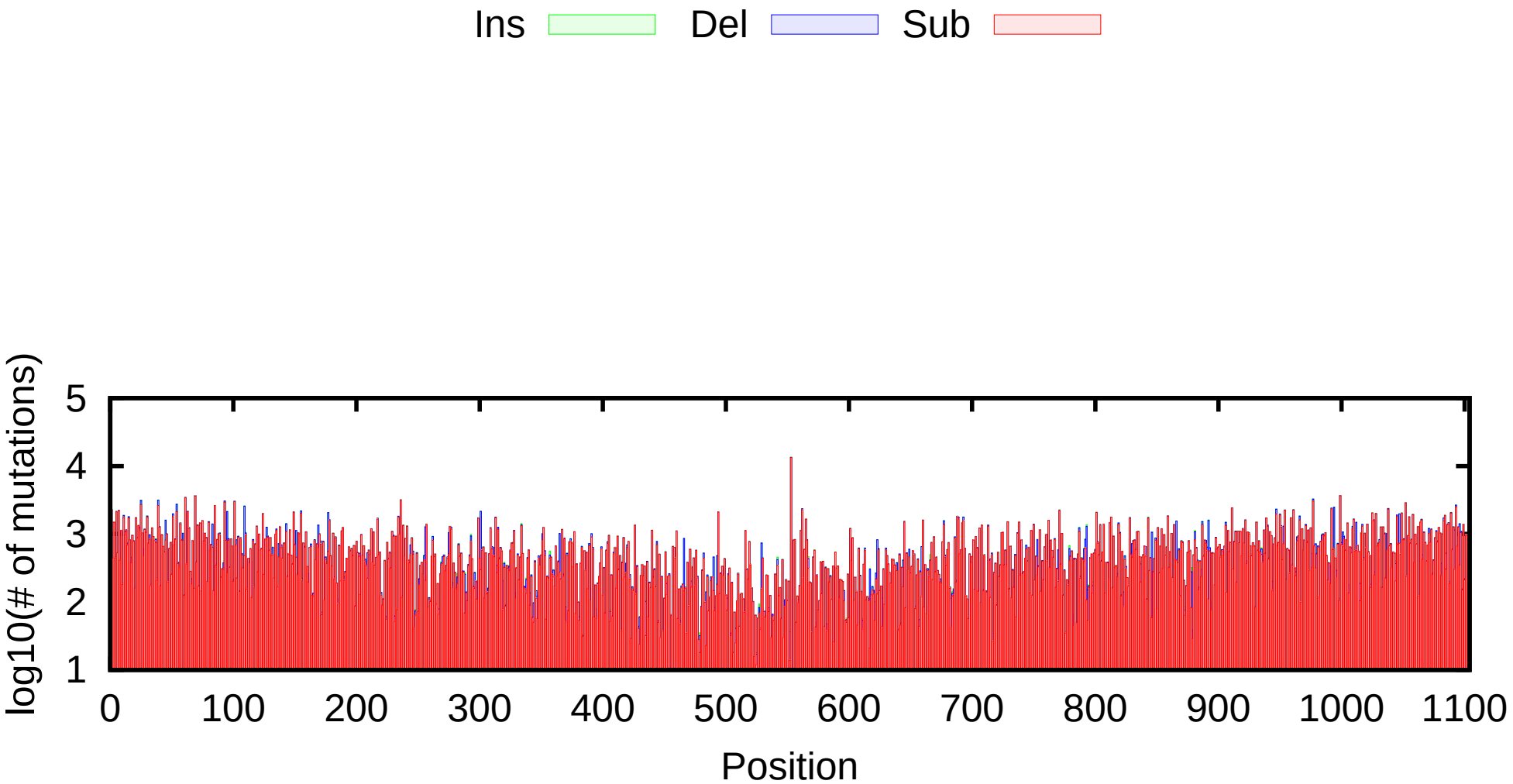
HEK293 **で強い転写活性**

導入された変異のスペクトル (EF1a1プロモーター)

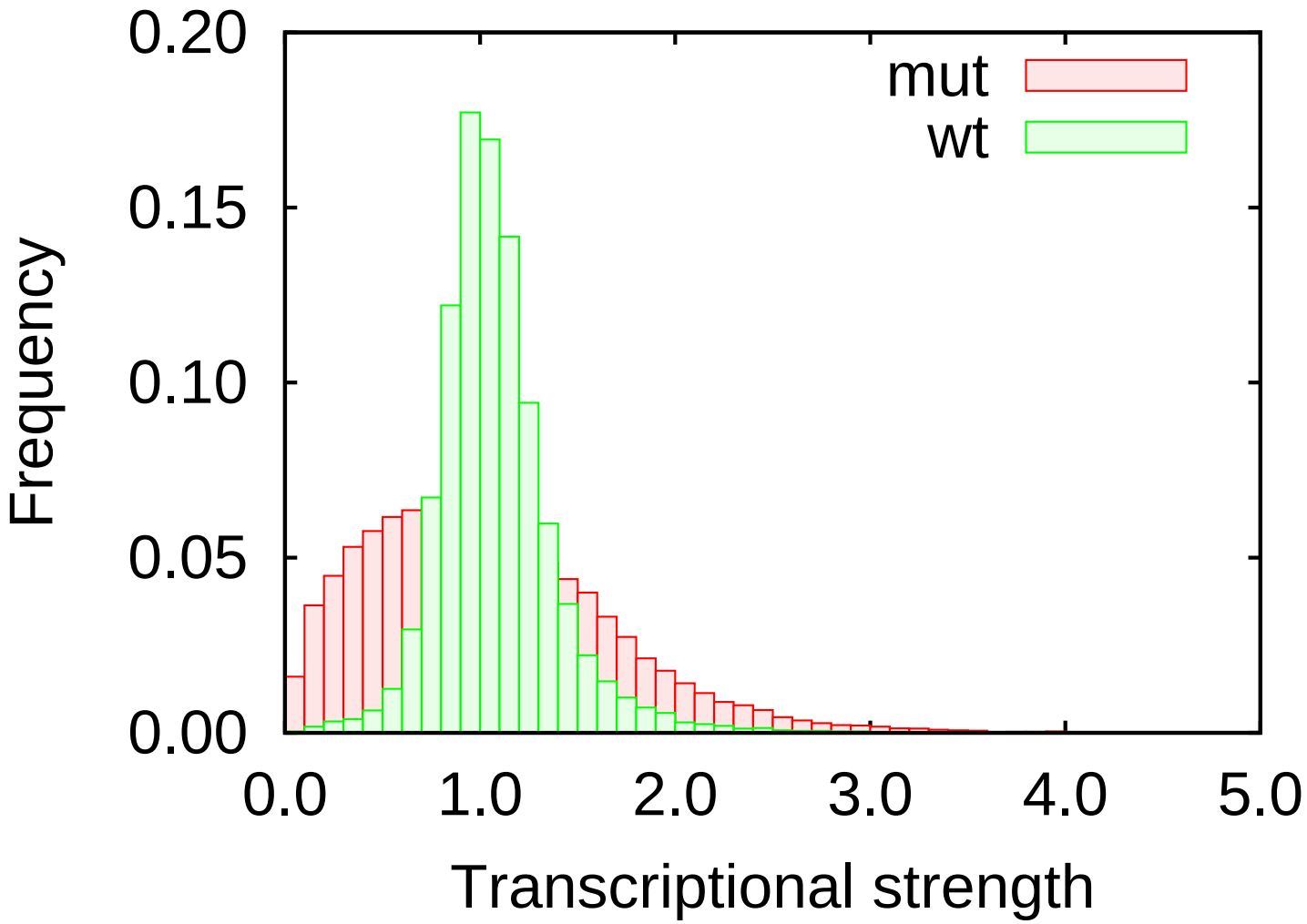




導入された変異の位置的な分布 (EF1a1プロモーター)



転写強度のダイナミックレンジ (EF1a1プロモーター)



Quantitative sequence-activity modeling (QSAM) (Jonsson 1993)

$$\log Y = B + \sum_{b,i} A_{bi} X_{bi}$$

A_{bi} 位置 i の塩基 b の転写への寄与
X_{bi} 位置 i の塩基が b ならば 1，そうでなければ 0
B 転写のベースライン

説明変数の選択

LASSO (Tibshirani 1996)

Least-square linear regression problem with regularization by the l_1-norm

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left[\frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda P(\beta) \right]$$

where

$$P(\beta) = \|\beta\|_{l_1} = \sum_{j=1}^{p} |\beta_j|$$

BOLASSO (BOotstrap LASSO)

各変数が選択される頻度から、各々の回帰への寄与を推定 (Bach 2008)

選択頻度の高い変数から順に、回帰に関連のあるものを決定 (Rohart 2011)

帰無仮説：i 番目以降の変数は関連がない

帰無仮説が棄却される限り i を増やす

回帰モデルの導出

粗視モデル → 構造化 → 微視モデル

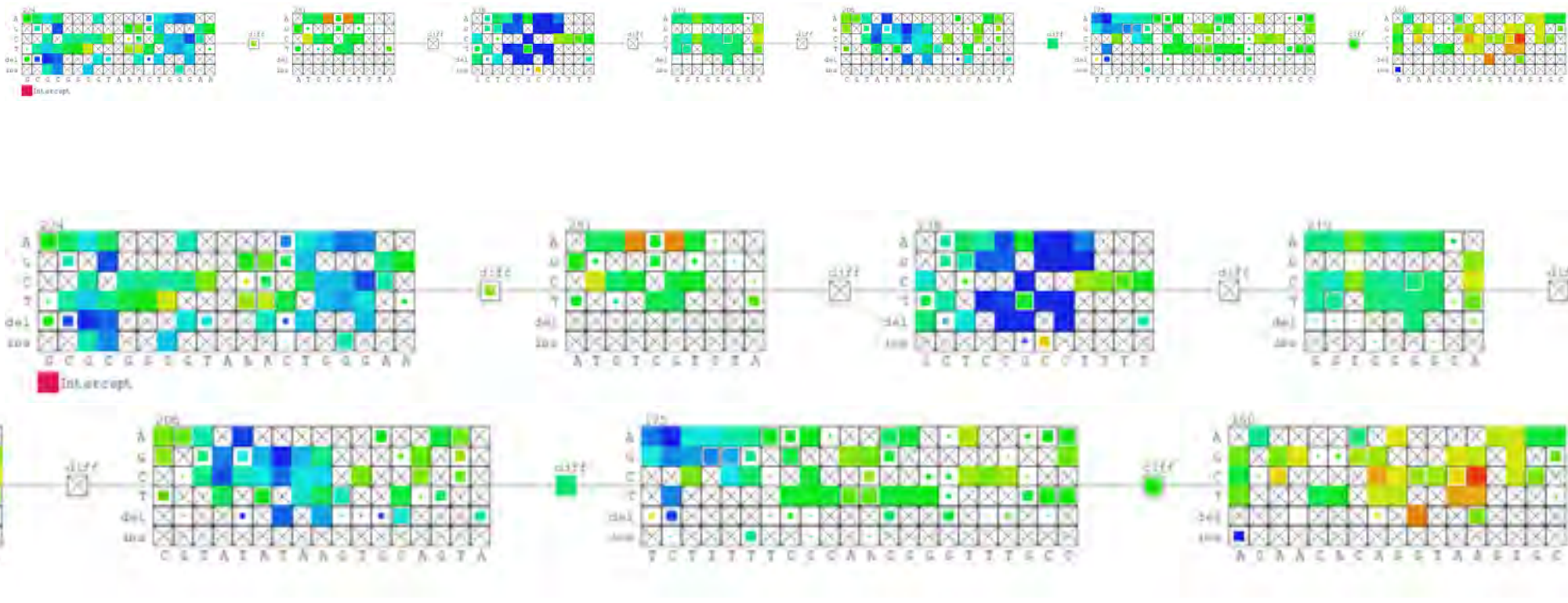
TFBS

各位置における塩基，欠失，挿入配列長の
転写強度への寄与

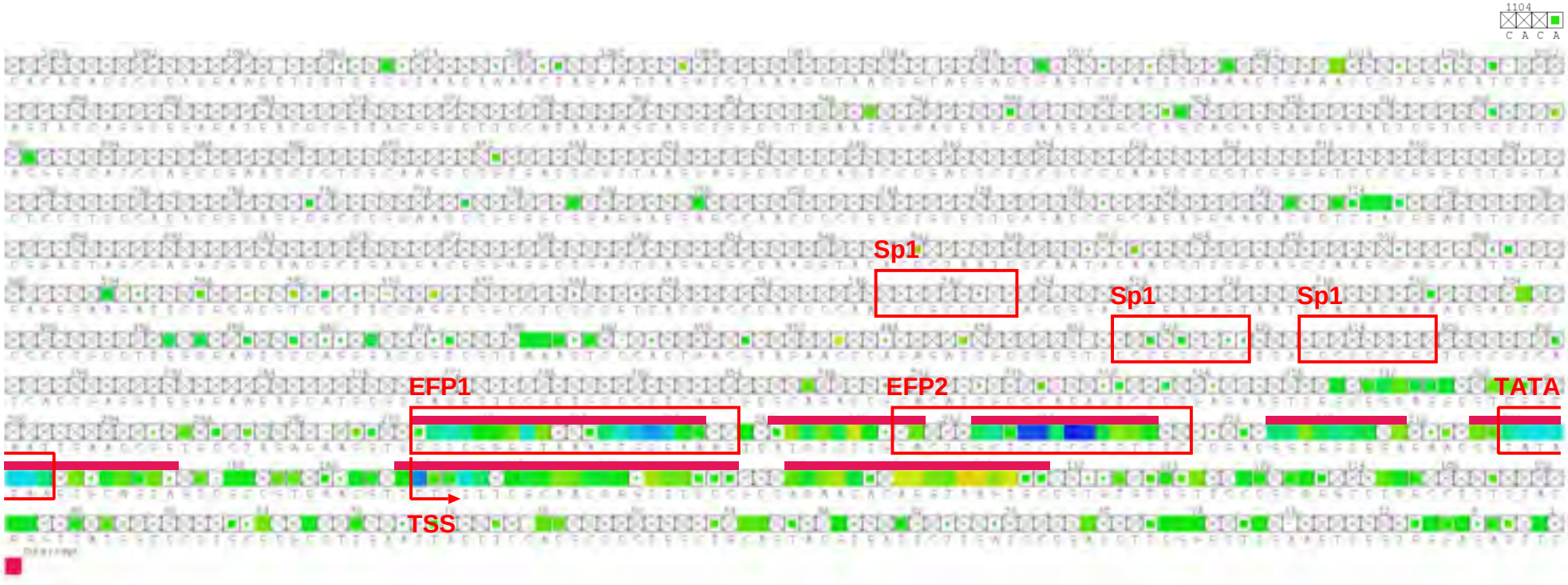
スペーサー

野生型のスペーサー配列長との差 (絶対値) の
転写強度への寄与

転写強度の回帰モデル (EF1a1プロモーター)



アノテーションと比べる (EF1a1プロモーター)

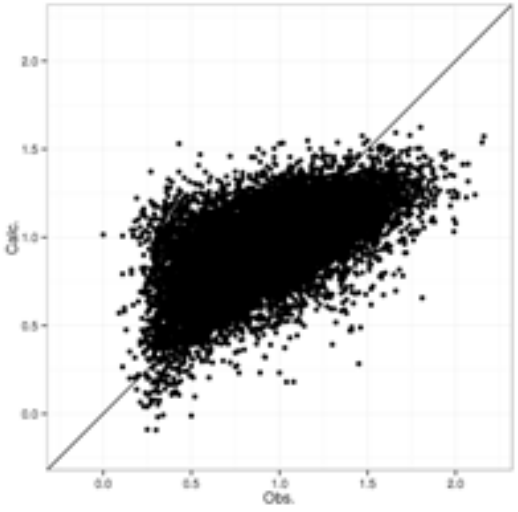


転写強度の回帰モデルの性能

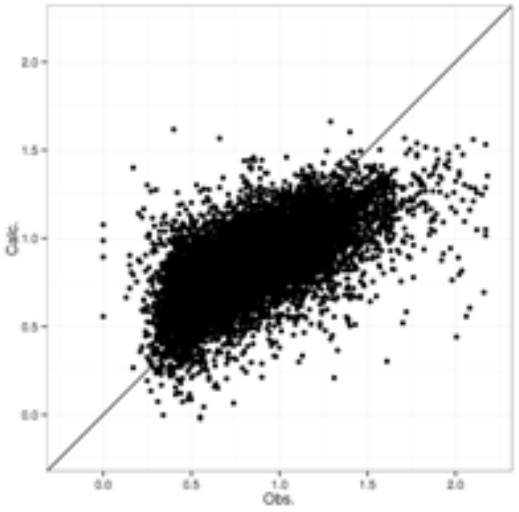
Promoter	Model	
	# of param.	$R^{\dagger}$
EF1a1	314	0.655
GAPDH	254	0.649
DDX5	116	0.606

† 10 **分割クロス検定**

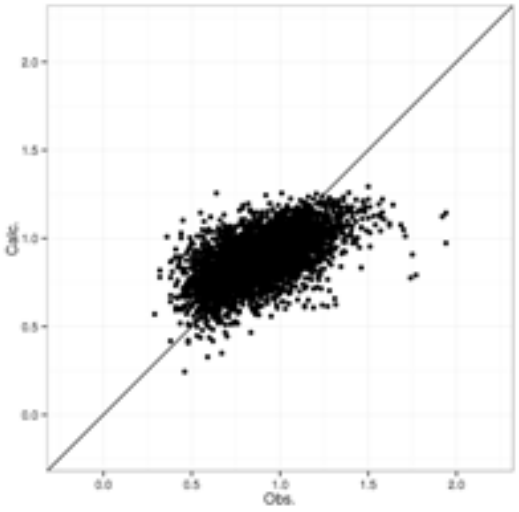
EF1a1



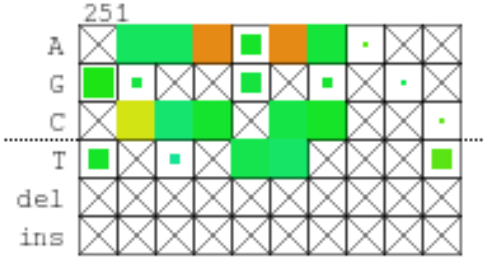
GAPDH



DDX5



転写強度を高める
(EF1a1プロモーター)



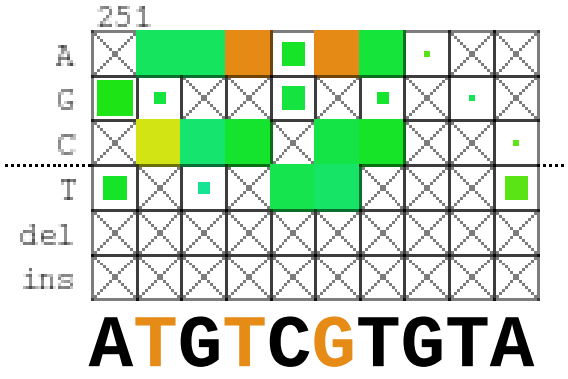
ATGTCGTGTA

ACGACATGTA

TFBS	Str.	Score		Seq.
		Core	Matrix	
V$CPHX_01	−	0.985	0.650	agTGATGtcgtgta
V$GRE_C	+	0.952	0.789	gtgatgtcgtGTACTg
V$SP100_04	+	0.928	0.886	gaTGTCGtgtactgg
V$RHOX11_01	+	0.904	0.745	aaagtGATGTcgtgtac

TFBS	Str.	Score		Seq.
		Core	Matrix	
V$CREB1_Q6	−	1.000	0.952	agTGACGacatg
V$IRX2_01	+	1.000	0.915	tgacgACATGtactggc
V$IRX2_01	−	1.000	0.907	gtgacgaCATGTactgg
V$GRE_C	+	0.952	0.852	gtgacgacatGTACTg

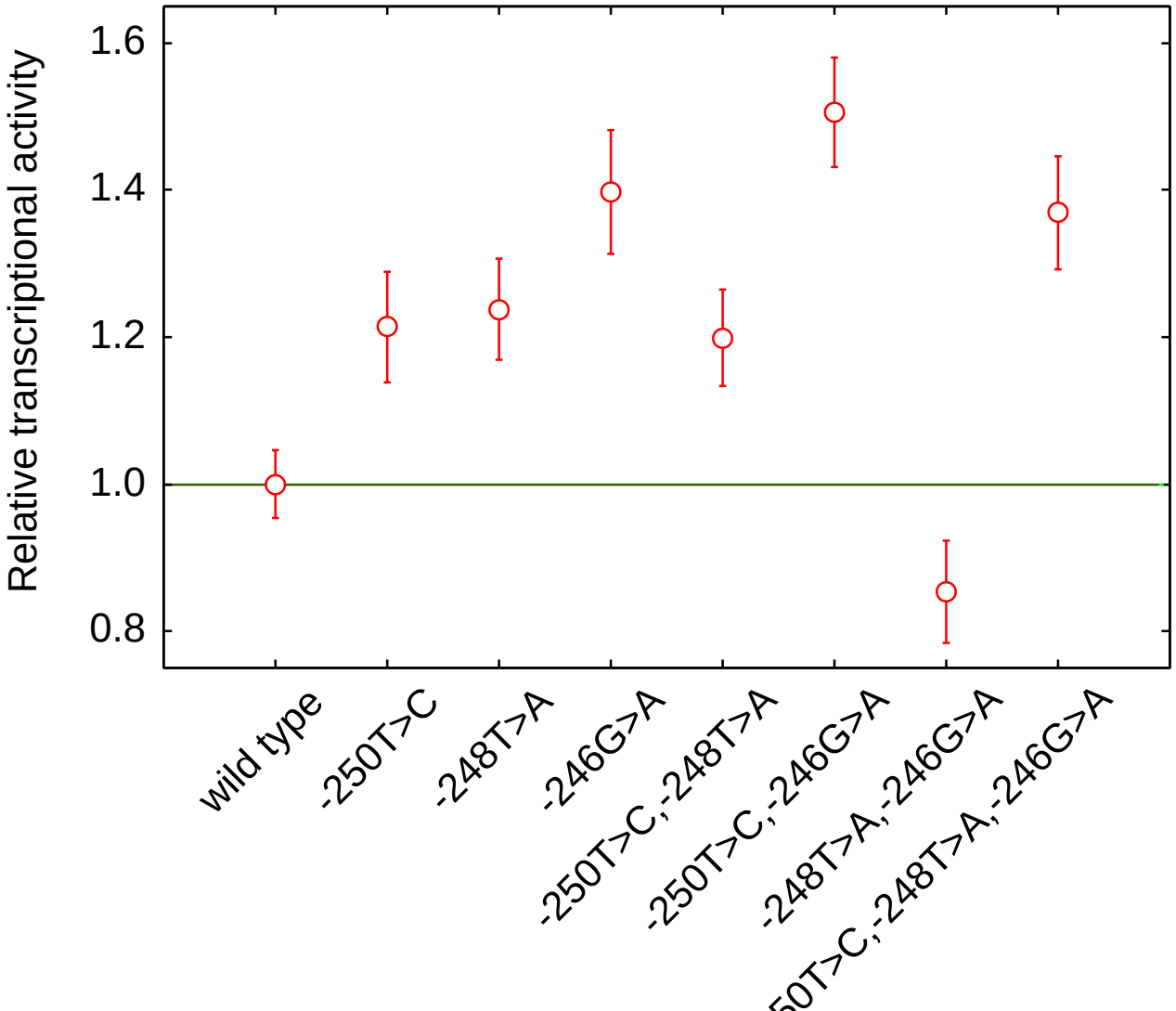
野生型配列を改変する (EF1a1プロモーター)



0. *wt*
1. **−250**T>C
2. **−248**T>A
3. **−246**G>A
4. **−250**T>C, **−248**T>A
5. **−250**T>C, **−246**G>A
6. **−248**T>A, **−246**G>A
7. **−250**T>C, **−248**T>A, **−246**G>A

Luciferase reporter assay

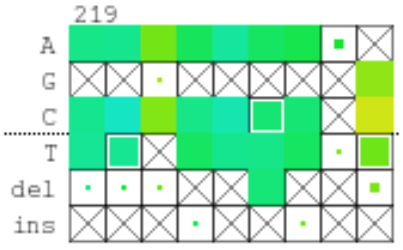
転写強度は高められたか？
(EF1a1プロモーター)



実験結果を説明できるか？ (EF1a1プロモーター)

Obs.	Calc.	#	TFBS	Score	Promoter
1.00	1.05	—	V$SP100_04	0.89	*wt*
1.21	1.18	907	V$SP100_04	0.95	**−250**T>C
1.24	1.28	330	V$CREB1_Q6	0.98	**−248**T>A
1.40	1,28	389	V$CREB1_Q6	0.93	**−246**G>A
1.20	1.42	9	V$CREB1_Q6	0.94	**−250**T>C , **−248**T>A
1.51	1.41	7	V$CREB1_Q6	1.00	**−250**T>C , **−246**G>A
0.85	1.51	1	V$IRX2_01	0.92	**−248**T>A , **−246**G>A
1.37	1.65	0	V$CREB1_Q6	0.95	**−250**T>C , **−248**T>A , **−246**G>A

転写強度を高める（2）
(EF1a1プロモーター)

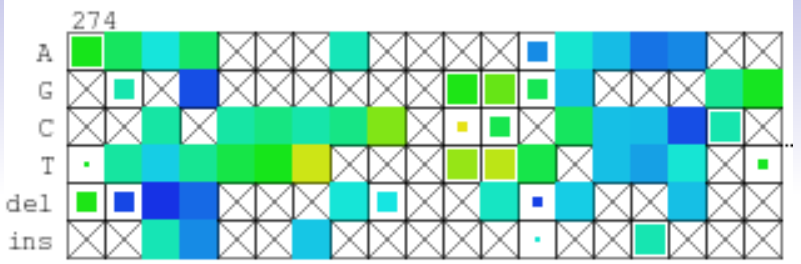


GGTGGGGGA

GGTGGGGGC

TFBS	Str.	Score		Seq.
		Core	Matrix	
V$IK_Q5	+	1.000	0.894	tggGGGAGaa
V$MUSCLEINI_B	–	1.000	0.877	ttcccgaGGGTGggggag
V$GRE_C	–	1.000	0.801	gAGAACcgtatataag
V$TATA_01	–	0.936	0.884	gagaaccgTATATaa
V$HELIOSA_02	+	0.853	0.873	tggGGGAGaac

TFBS	Str.	Score		Seq.
		Core	Matrix	
V$EGR1_Q6	+	1.000	0.953	gtGGGGGcga
V$MUSCLEINI_B	–	1.000	0.890	ttcccgaGGGTGggggcg
V$TATA_01	–	0.936	0.885	gcgaaccgTATATaa
V$MYB_05	+	0.881	0.855	gggcgaaCCGTAtataa



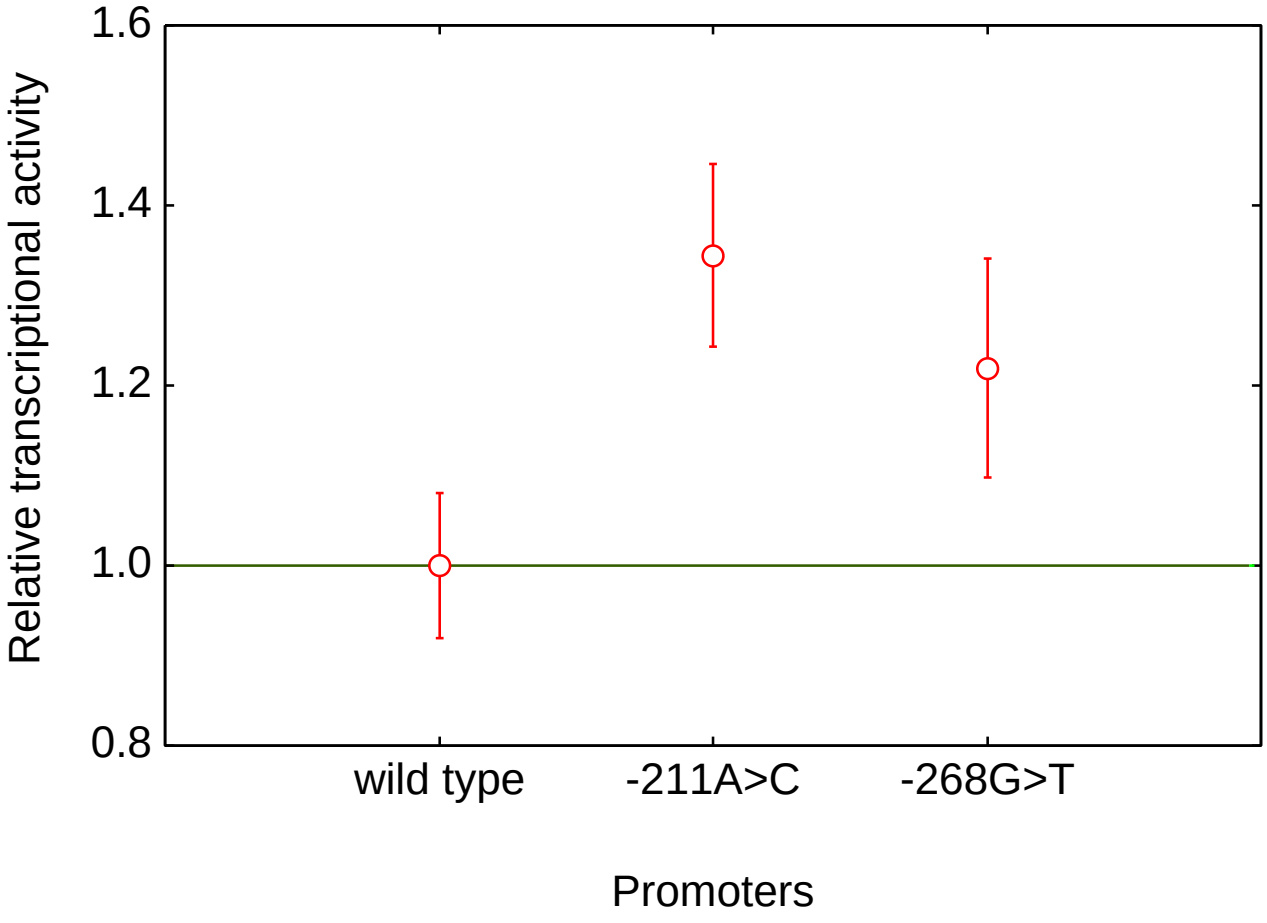
GCGCGGGGTAAACTGGGAA

GCGCGGTGTAAACTGGGAA

TFBS	Str.	Score		Seq.
		Core	Matrix	
V$CHCH_01	+	1.000	0.989	CGGGGt
V$E2F_Q6_01	–	1.000	0.838	aggTGGCGcggg
V$ZFP161_04	–	1.000	0.748	gaaggtgGCGCGgg
V$HOXD12_01	+	1.000	0.715	gcgcgggGTAAActggg

TFBS	Str.	Score		Seq.
		Core	Matrix	
V$TBX5_01	+	1.000	0.886	cgcGGTGTaaac
V$E2F_Q6_01	–	1.000	0.838	aggTGGCGcggt
V$ZFP161_04	–	1.000	0.751	gaaggtgGCGCGgt
V$HOXD12_01	+	1.000	0.716	gcgcggtGTAAActggg
V$RHOX11_01	+	0.910	0.905	ggcgcGGTGTaaactgg

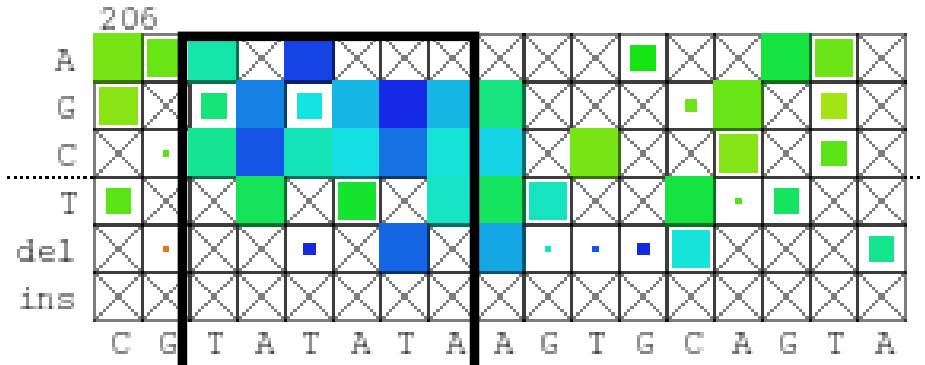
転写強度は高められたか？（２）
(EF1a1プロモーター)



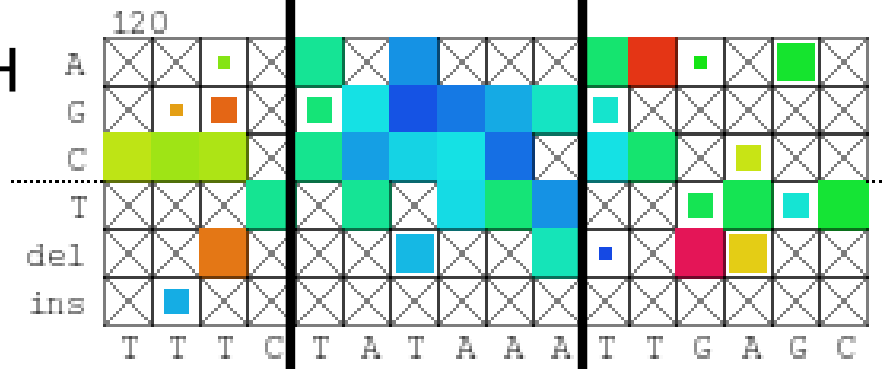
TFBSは交換できるか？

TATA boxes

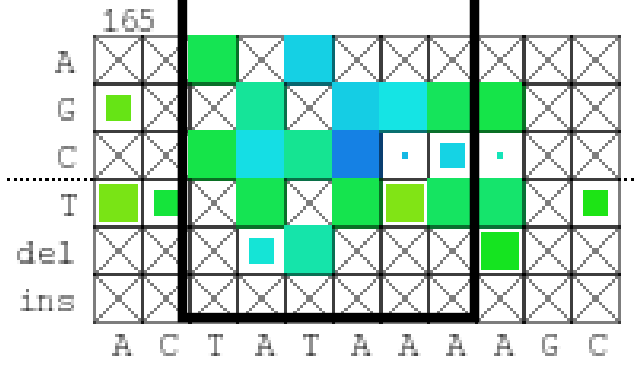
EF1a1



GAPDH



DDX5



まとめ

大規模な変異型プロモーターの塩基配列と転写強度を同時に測定

ヒト GC-rich プロモーターの転写強度を推定する回帰モデルを導出

回帰モデルに基づいたプロモーター配列の改変に成功

次のステップ

GC-rich	AT-rich	
√	—	TATA-containing
—	—	TATA-less

ゲノムワイド，組織（培養）細胞ワイド

メチル化，ヒストン修飾，
クロマチン構造