次世代シークエンシングによるSTRタイピングのための機械学習

2021/07/17

京都大学(院)医学研究科 統計遺伝学分野

山田　亮

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Forensic STR allele extraction using a machine learning paradigm

Yao-Yuan Liu[a], David Welch[b], Ryan England[a,c], Janet Stacey[c], SallyAnn Harbison[c,*]

[a] Forensic Science Program, School of Chemical Sciences, University of Auckland, 38 Princes Street, Auckland 1010, New Zealand
[b] School of Computer Science, University of Auckland, 38 Princes Street, Auckland 1010, New Zealand
[c] Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland 1142, New Zealand

ARTICLE INFO

Keywords:
STR extraction
Machine learning
Massively parallel sequencing
Bioinformatics

ABSTRACT

We present a machine learning approach to short tandem repeat (STR) sequence detection and extraction from massively parallel sequencing data called Fragsifier. Using this approach, STRs are detected on each read by first locating the longest repeat stretches followed by locus prediction using k-mers in a machine learning sequence model. This is followed by reference flanking sequence alignment to determine precise STR boundaries. We show that Fragsifier produces genotypes that are concordant with profiles obtained using capillary electrophoresis (CE), and also compared the results with that of STRait Razor and the ForenSeq UAS. The data pre-processing and training of the sequence classifier is readily scripted, allowing the analyst to experiment with different thresholds, datasets and loci of interest, and different machine learning models.

情報があれば、答えが出るはずだ

- 推定
- 予測
- 検定

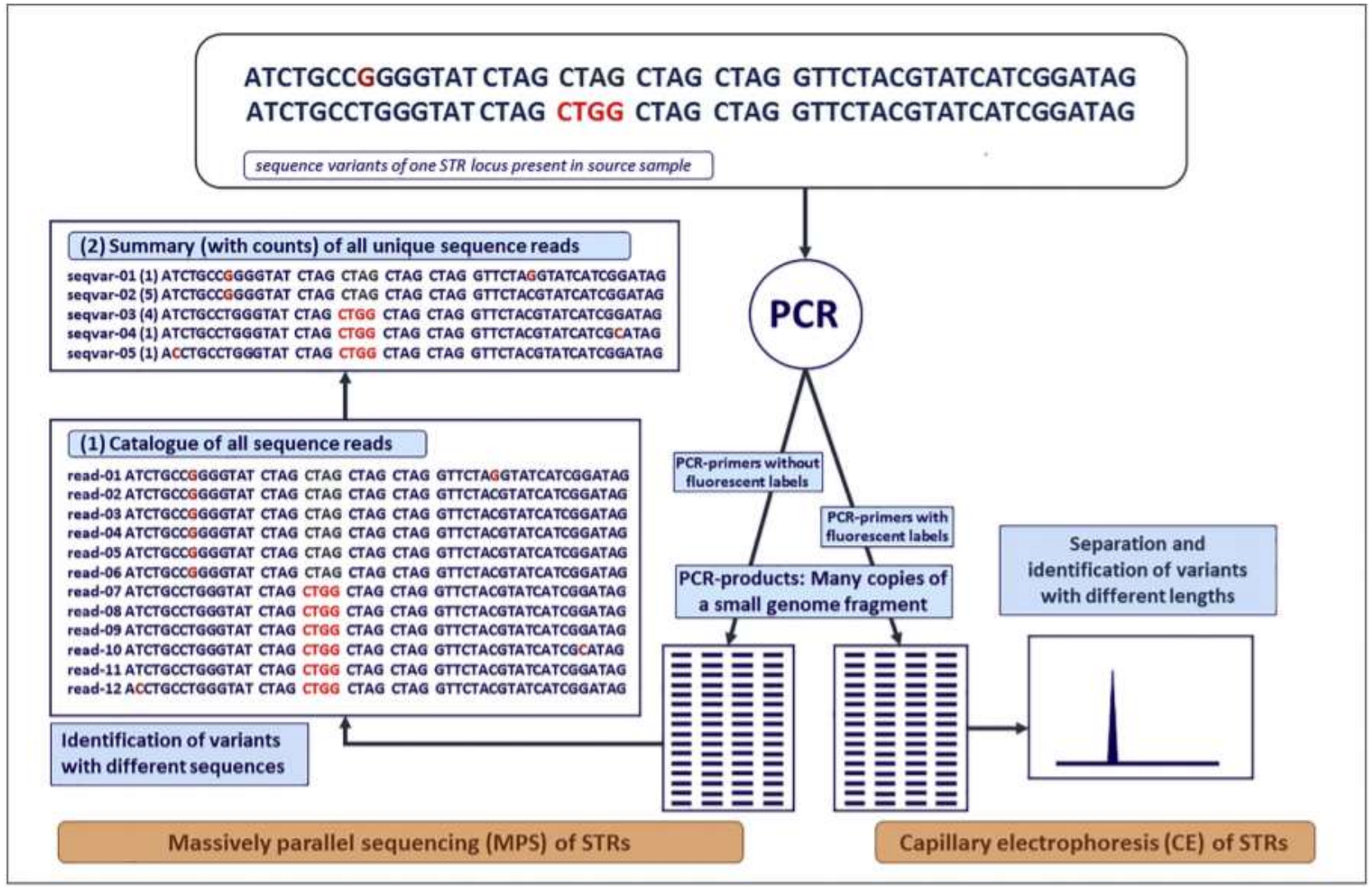
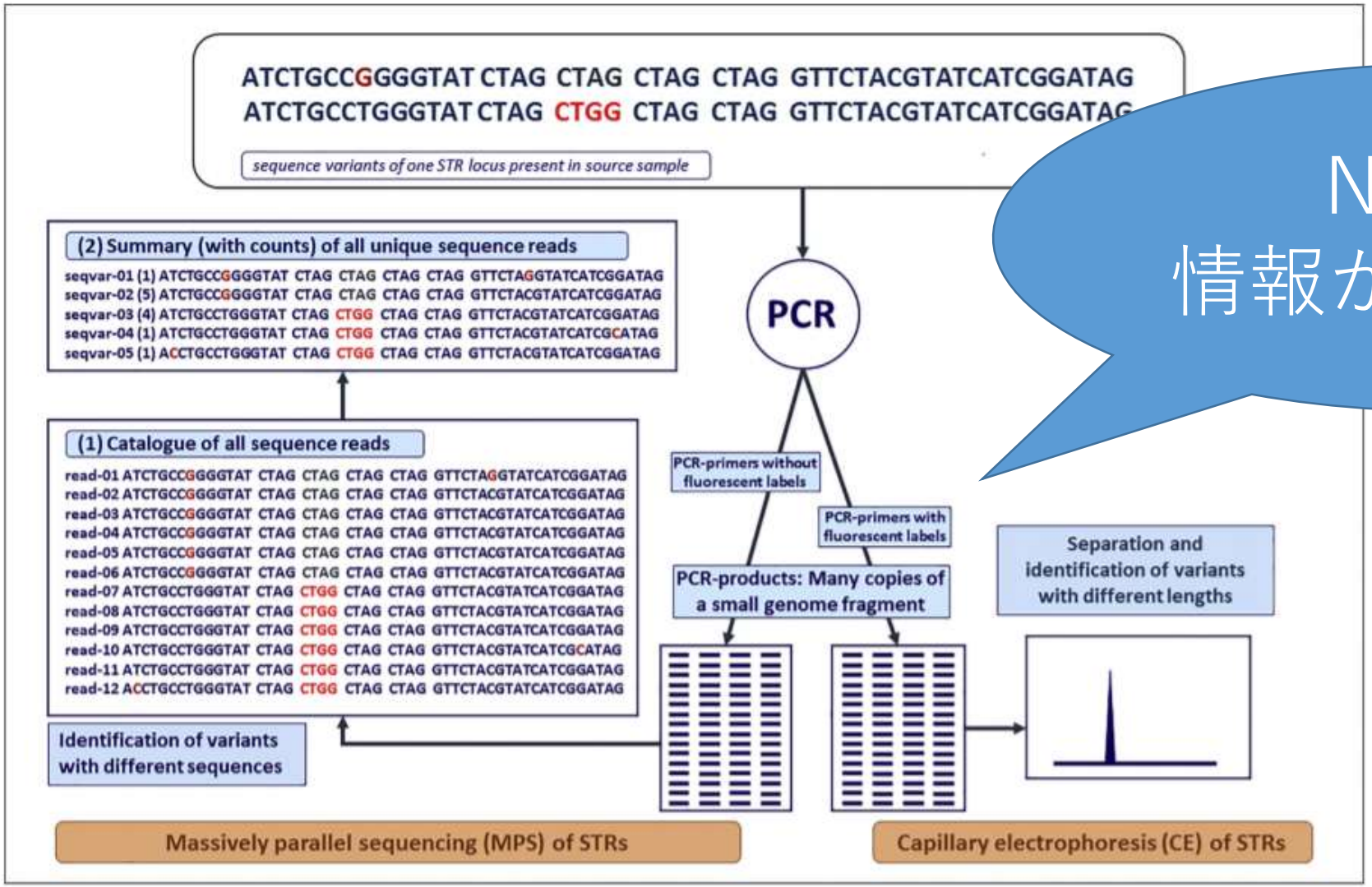


Fig. 1. A simplified explanation of the main differences between CE genotyping of STRs and MPS of STRs.

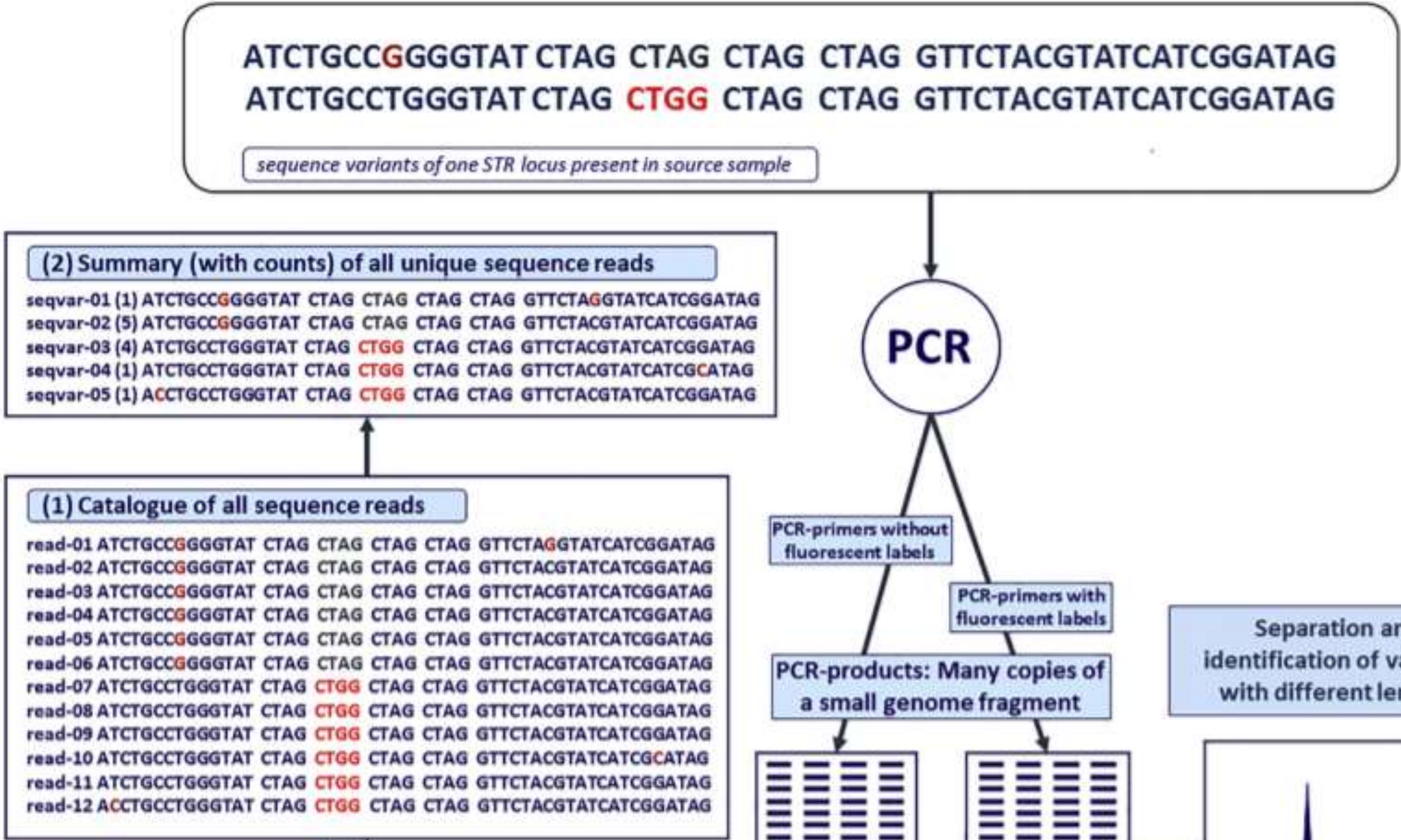
In this theoretical example, a PCR is designed to reveal the genetic variation of a single STR locus. The DNA sample has two different alleles, both four repeats long. One allele contains a CTGG instead of a CTAG repeat unit. The other allele contains a SNP in the 5′flanking region, a T > G mutation 7 bp. prior to the first repeat unit. When using CE, the PCR uses a set of PCR primers of which one primer has a fluorescent label. After PCR, all PCR products will include this fluorochrome, enabling the detecting of the PCR products on a CE platform. As both alleles are 4 repeats long, only a single peak will be visible. For this locus, this DNA sample will be called a homozygote when using CE. When using MPS, the PCR primers are not labeled with a fluorochrome. All PCR products are simply sequenced on a MPS platform, resulting in a long list of sequence reads. For the sake of simplicity, in this example 12 sequence reads are shown (panel 1). Six reads (01 – 06) contain the T > G mutation in the 5′flanking region. Six other reads (07–12) contain the CTGG repeat. Upon further inspection, three more genetic variants are detected. A single read (01) contains a C > G error 7 bp. 3′of the repeat structure. Another read [10] contains a G > C error 16 bp. 3′of the repeat structure. Finally, read 12 shows a T > C error at the second position of the sequence. This full spectrum of sequence variation is summarized in panel 2. A total of five different sequence variants were detected in this sample. Two variants (02 and 03) were seen multiple times and probably reflect the true alleles. A further three variants (01, 04, and 05) are seen only once; their defining SNP likely representing error reads: reads containing PCR or sequence induced sequence errors.



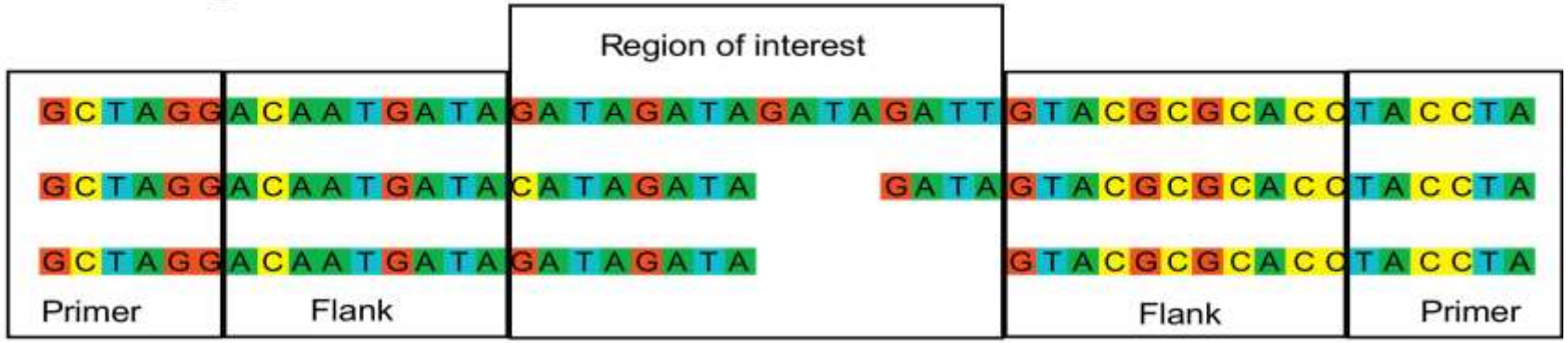
NGSは
情報が細かい！

Fig. 1. A simplified explanation of the main differences between CE genotyping of STRs

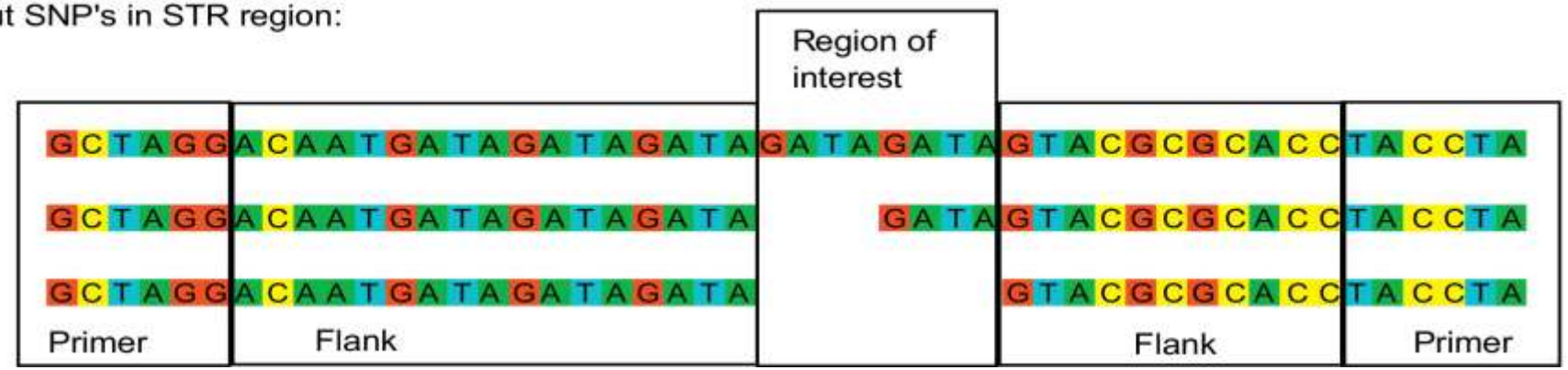
PCR is designed
ingle STR
ent al-
e con-
t unit.
in the
tation 7 bp.
unit. When using CE,
the PCR uses a set of PCR primers of which one primer has a fluorescent label. After PCR, all PCR products will include this fluorochrome, enabling the detecting of the PCR products on a CE platform. As both alleles are 4 repeats long, only a single peak will be visible. For this locus, this DNA sample will be called a homozygote when using CE. When using MPS, the PCR primers are not labeled with a fluorochrome. All PCR products are simply sequenced on a MPS platform, resulting in a long list of sequence reads. For the sake of simplicity, in this example 12 sequence reads are shown (panel 1). Six reads (01 – 06) contain the T > G mutation in the 5′flanking region. Six other reads (07–12) contain the CTGG repeat. Upon further inspection, three more genetic variants are detected. A single read (01) contains a C > G error 7 bp. 3′of the repeat structure. Another read [10] contains a G > C error 16 bp. 3′of the repeat structure. Finally, read 12 shows a T > C error at the second position of the sequence. This full spectrum of sequence variation is summarized in panel 2. A total of five different sequence variants were detected in this sample. Two variants (02 and 03) were seen multiple times and probably reflect the true alleles. A further three variants (01, 04, and 05) are seen only once; their defining SNP likely representing error reads: reads containing PCR or sequence induced sequence errors.


ATCTGCCGGGGTAT CTAG CTAG CTAG CTAG GTTCTACGTATCATCGGATAG
ATCTGCCTGGGTAT CTAG CTGG CTAG CTAG GTTCTACGTATCATCGGATAG
sequence variants of one STR locus present in source sample
(2) Summary (with counts) of all unique sequence reads
seqvar-01 (1) ATCTGCCGGGGTAT CTAG CTAG CTAG CTAG GTTCTAGGTATCATCGGATAG
seqvar-02 (5) ATCTGCCGGGGTAT CTAG CTAG CTAG CTAG GTTCTACGTATCATCGGATAG
seqvar-03 (4) ATCTGCCTGGGTAT CTAG CTGG CTAG CTAG GTTCTACGTATCATCGGATAG
seqvar-04 (1) ATCTGCCTGGGTAT CTAG CTGG CTAG CTAG GTTCTACGTATCATCGCATAG
seqvar-05 (1) ACCTGCCTGGGTAT CTAG CTGG CTAG CTAG GTTCTACGTATCATCGGATAG
(1) Catalogue of all sequence reads
read-01 ATCTGCCGGGGTAT CTAG CTAG CTAG CTAG GTTCTAGGTATCATCGGATAG
read-02 ATCTGCCGGGGTAT CTAG CTAG CTAG CTAG GTTCTACGTATCATCGGATAG
read-03 ATCTGCCGGGGTAT CTAG CTAG CTAG CTAG GTTCTACGTATCATCGGATAG
read-04 ATCTGCCGGGGTAT CTAG CTAG CTAG CTAG GTTCTACGTATCATCGGATAG
read-05 ATCTGCCGGGGTAT CTAG CTAG CTAG CTAG GTTCTACGTATCATCGGATAG
read-06 ATCTGCCGGGGTAT CTAG CTAG CTAG CTAG GTTCTACGTATCATCGGATAG
read-07 ATCTGCCTGGGTAT CTAG CTGG CTAG CTAG GTTCTACGTATCATCGGATAG
read-08 ATCTGCCTGGGTAT CTAG CTGG CTAG CTAG GTTCTACGTATCATCGGATAG
read-09 ATCTGCCTGGGTAT CTAG CTGG CTAG CTAG GTTCTACGTATCATCGGATAG
read-10 ATCTGCCTGGGTAT CTAG CTGG CTAG CTAG GTTCTACGTATCATCGCATAG
read-11 ATCTGCCTGGGTAT CTAG CTGG CTAG CTAG GTTCTACGTATCATCGGATAG
read-12 ACCTGCCTGGGTAT CTAG CTGG CTAG CTAG GTTCTACGTATCATCGGATAG
PCR
PCR-primers without fluorescent labels
PCR-primers with fluorescent labels
PCR-products: Many copies of a small genome fragment
Separation ar
identification of v
with different le

a) With SNP's in STR region:

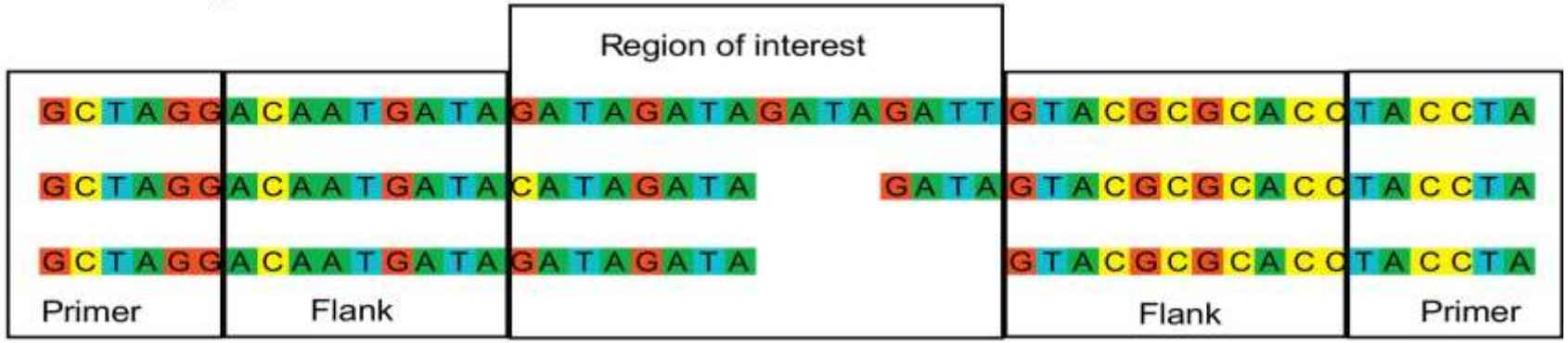


b) Without SNP's in STR region:

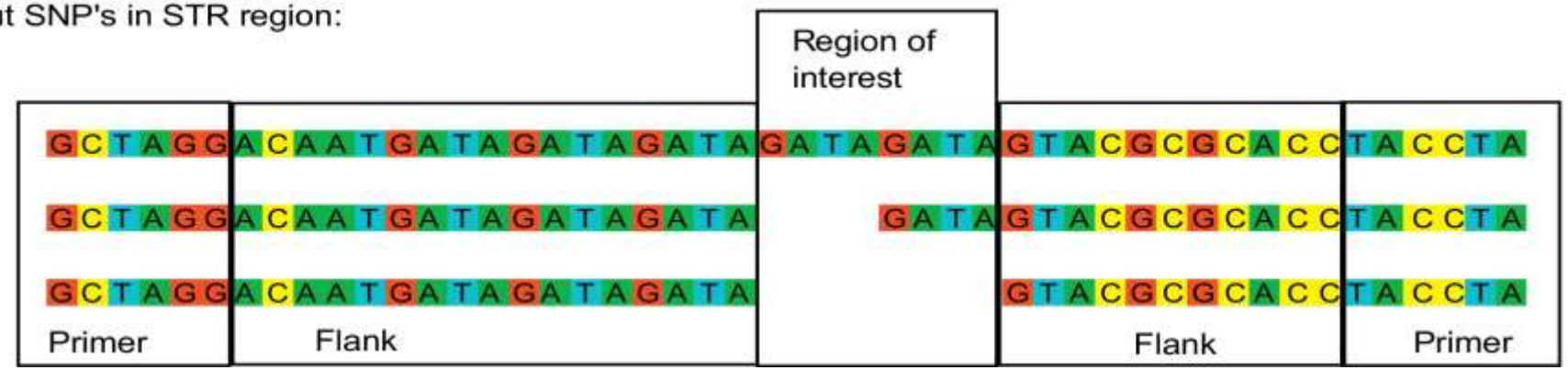


Forensic Science International: Genetics 2014 91-8DOI: (10.1016/j.fsigen.2013.10.012)

a) With SNP's in STR region:



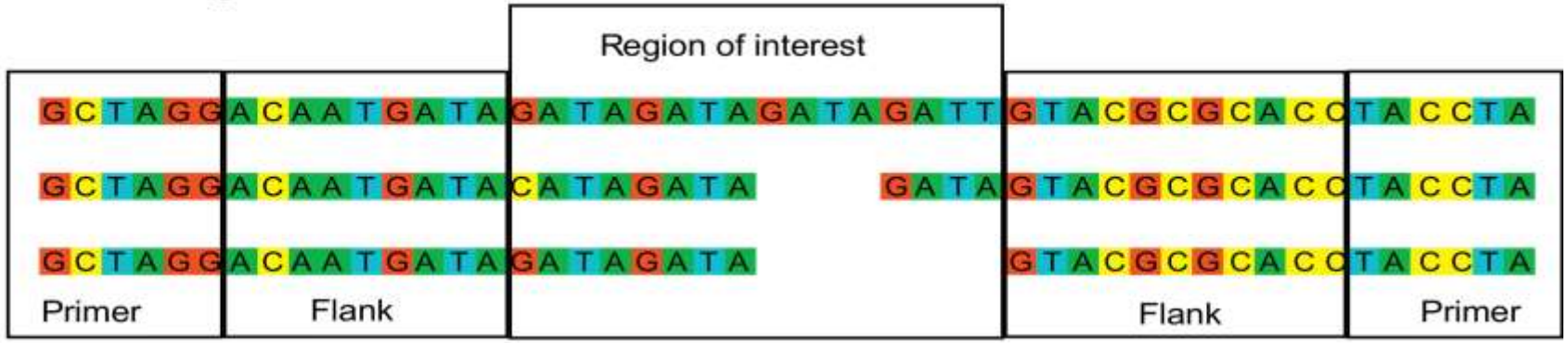
b) Without SNP's in STR region:



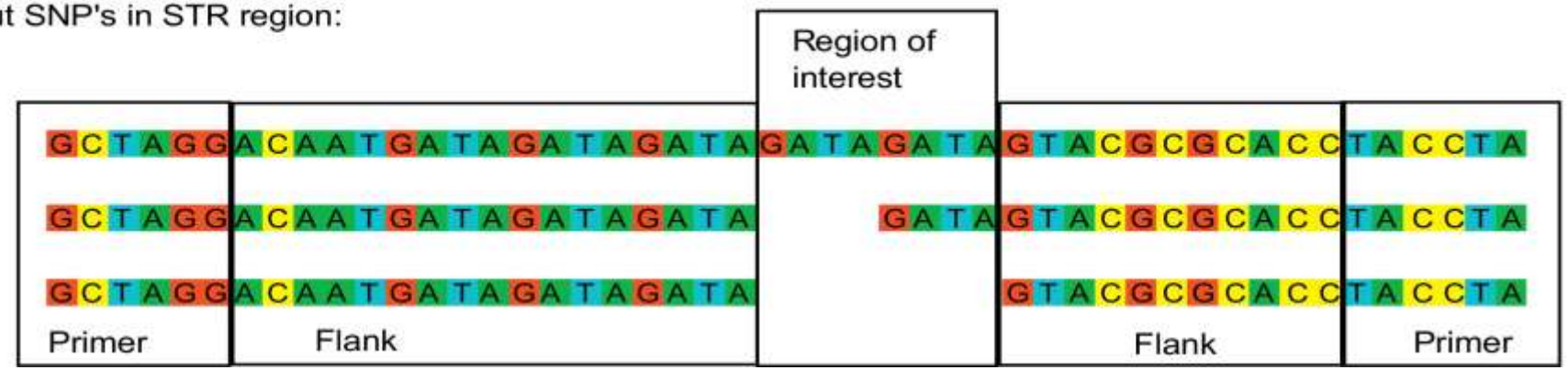
リピートの回数に興味がある

Forensic Science International: Genetics 2014 91-8DOI: (10.1016/j.fsigen.2013.10.012)

a) With SNP's in STR region:

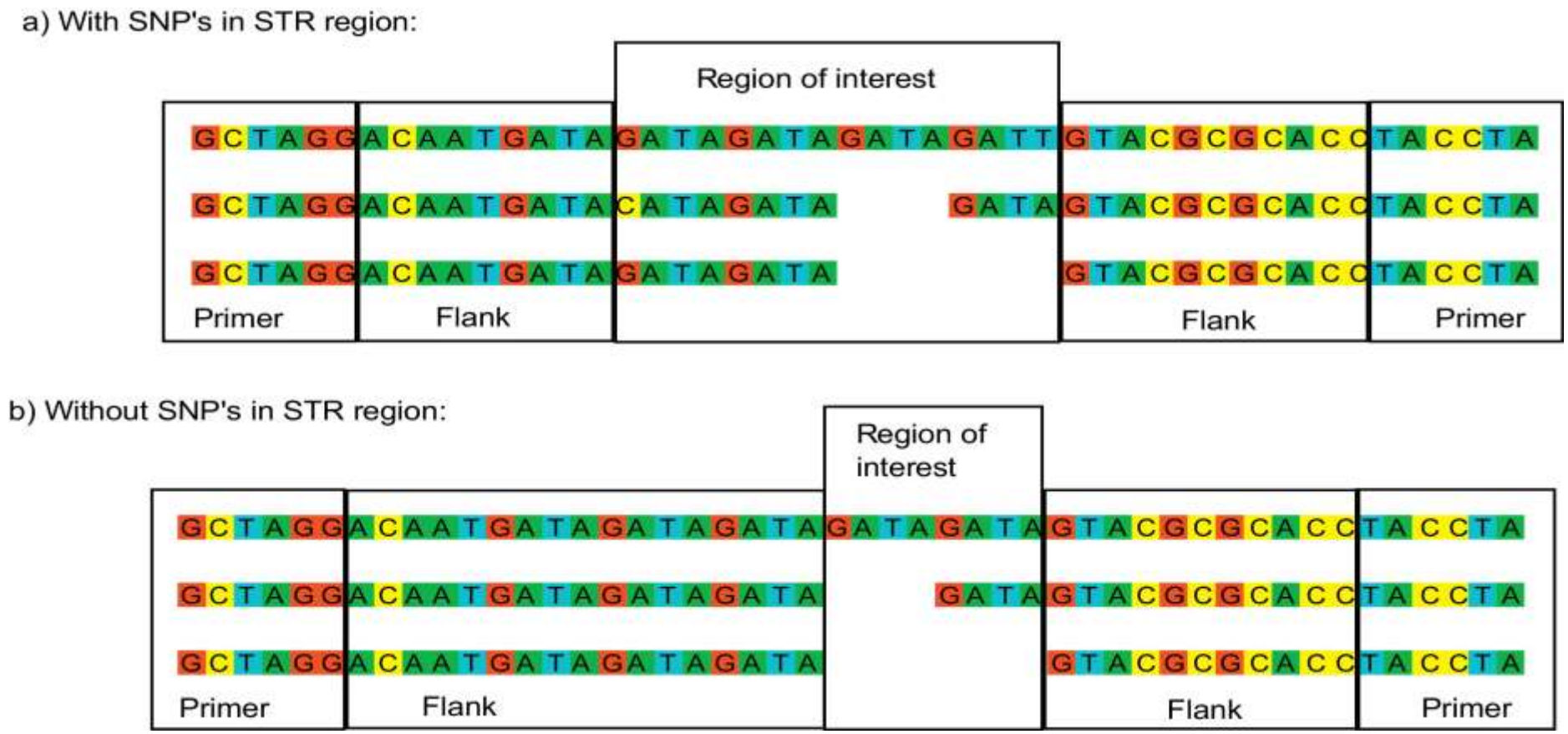


b) Without SNP's in STR region:



リピート回数にはバリエーションが大きい

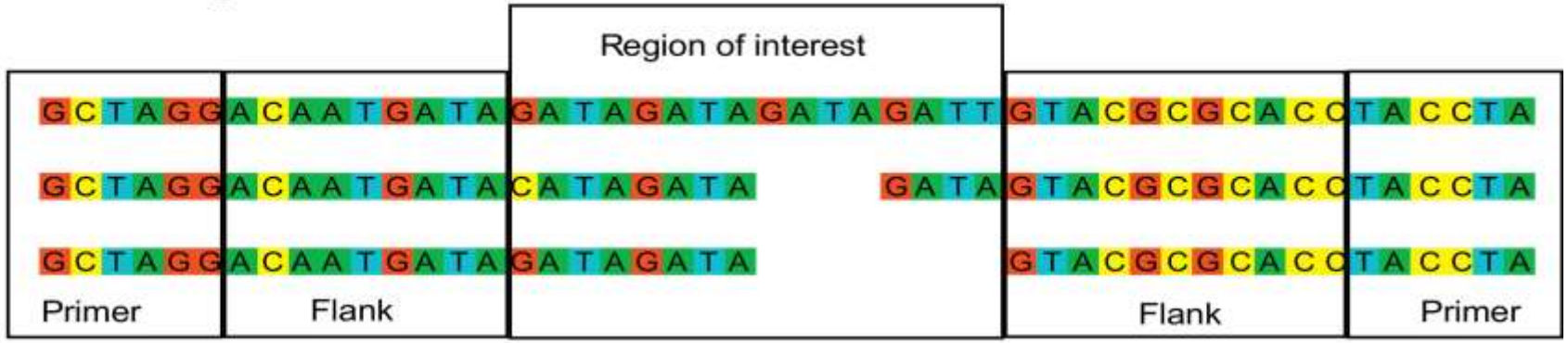
Forensic Science International: Genetics 2014 91-8DOI: (10.1016/j.fsigen.2013.10.012)



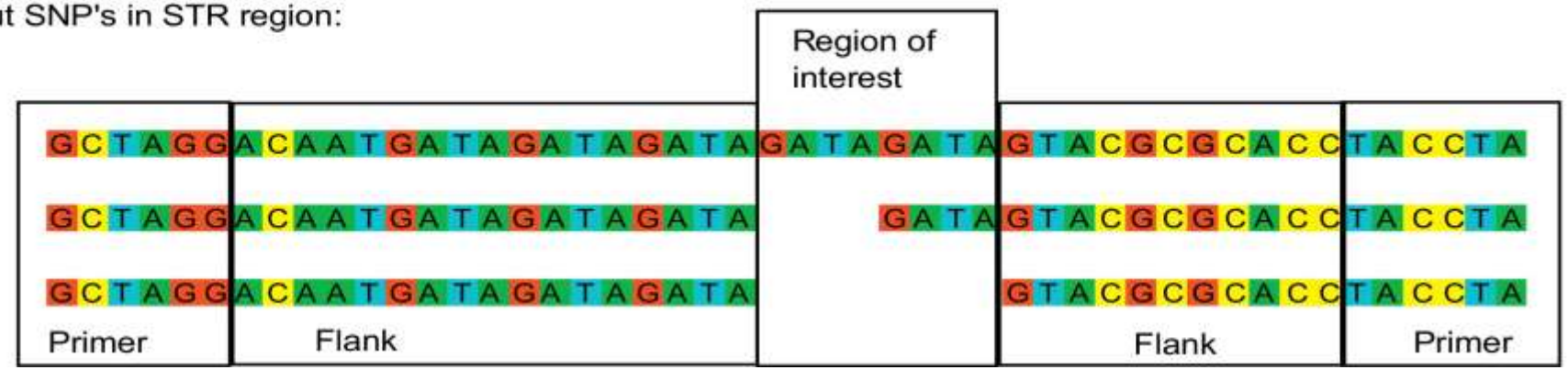
配列比較では、バリエーションの無いところが重要

Forensic Science International: Genetics 2014 91-8DOI: (10.1016/j.fsigen.2013.10.012)

a) With SNP's in STR region:

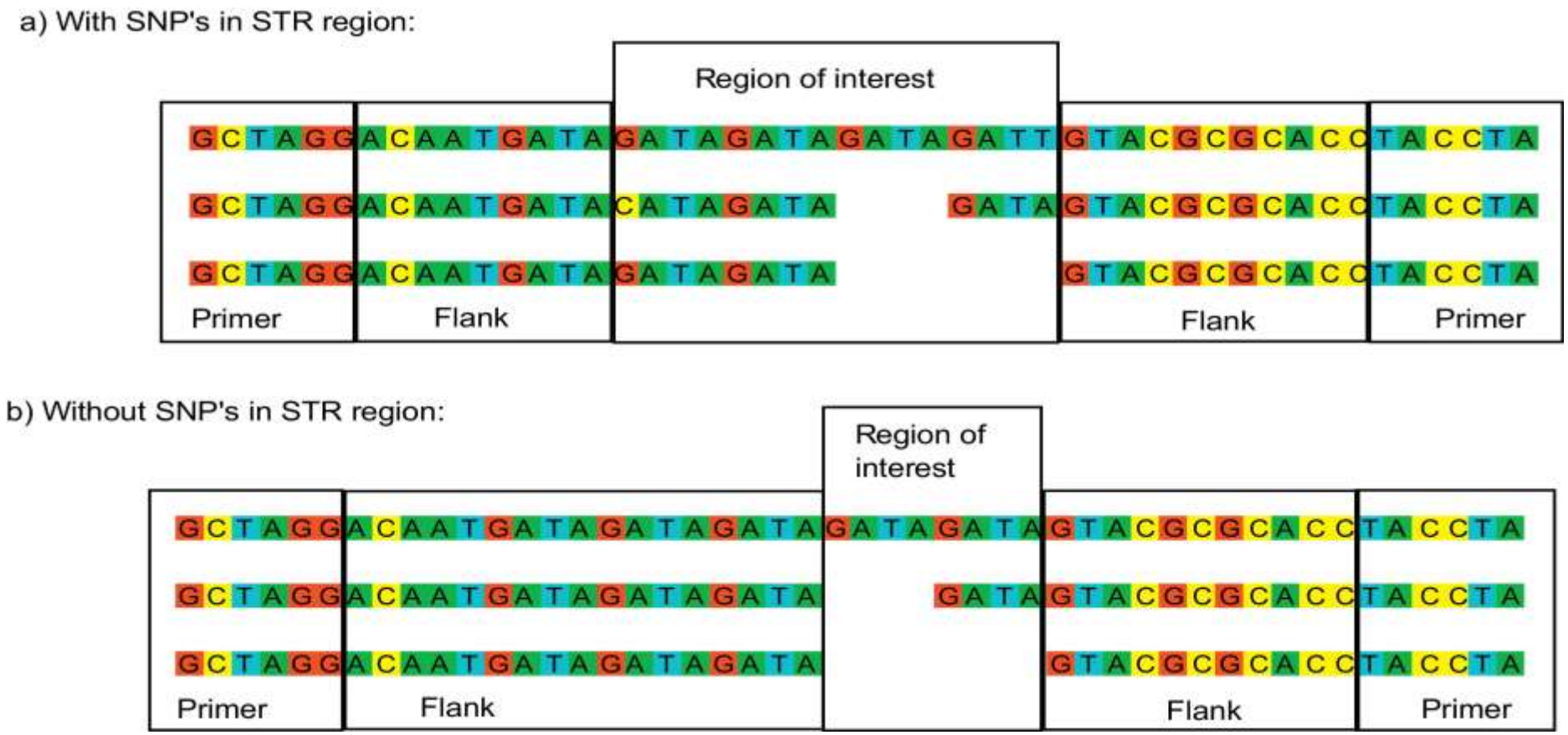


b) Without SNP's in STR region:



リピートの前後Flanking regionsが使いやすい

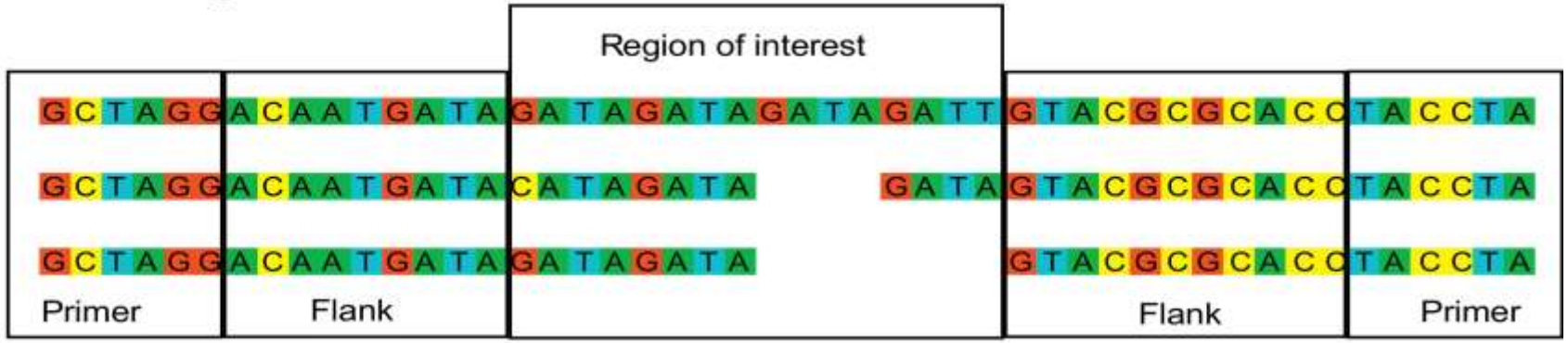
Forensic Science International: Genetics 2014 91-8DOI: (10.1016/j.fsigen.2013.10.012)



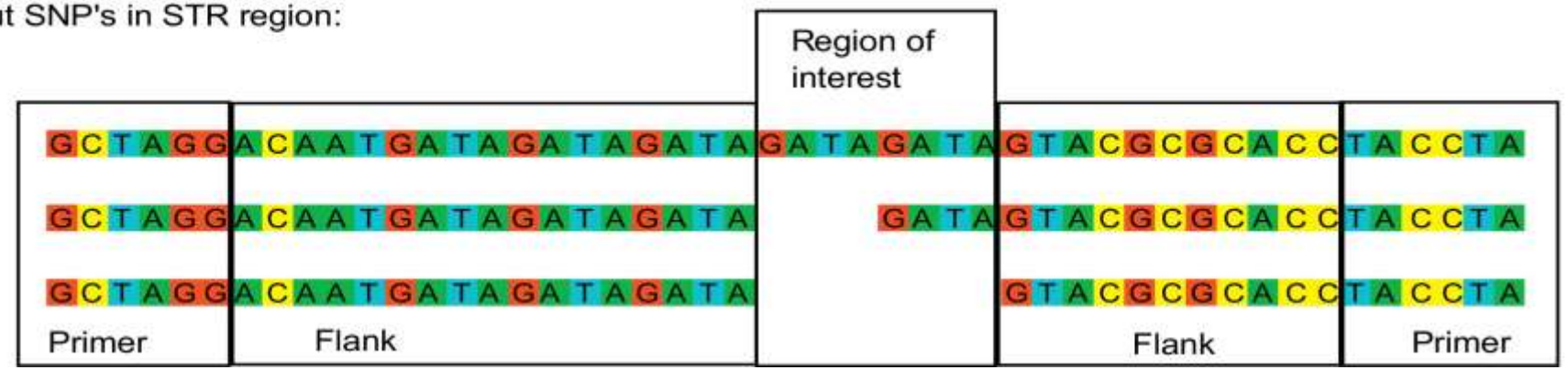
Flanking regionsは使いやすいが、バリエーション(多型)が比較的多い

Forensic Science International: Genetics 2014 91-8DOI: (10.1016/j.fsigen.2013.10.012)

a) With SNP's in STR region:



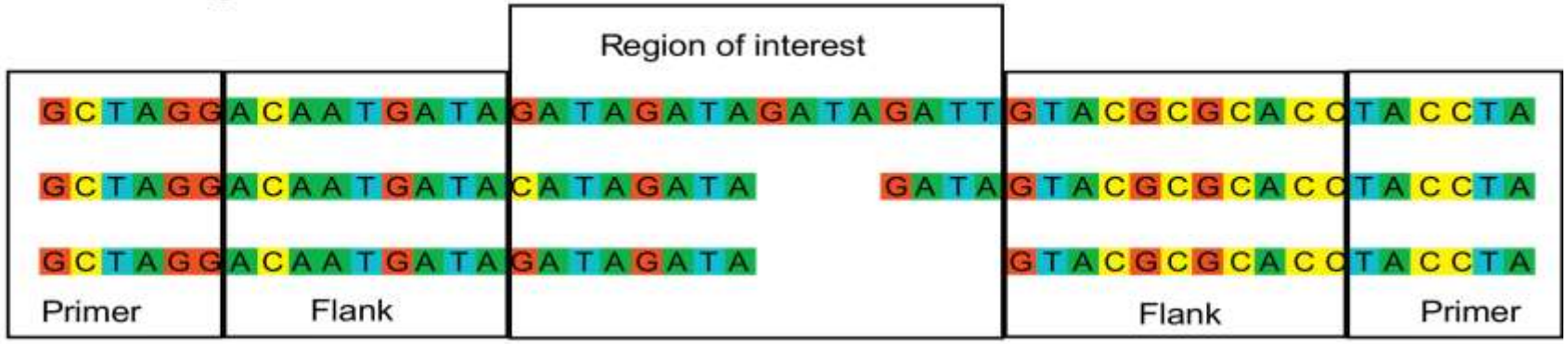
b) Without SNP's in STR region:



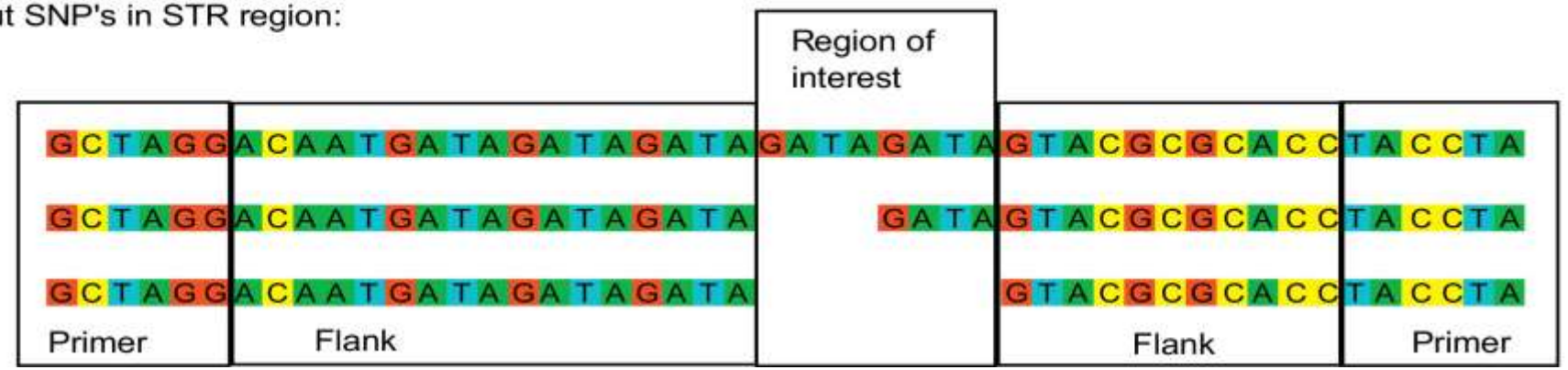
バリエーションが多いと、いわゆる次世代シークエンシングツールが使いにくい

Forensic Science International: Genetics 2014 91-8DOI: (10.1016/j.fsigen.2013.10.012)

a) With SNP's in STR region:

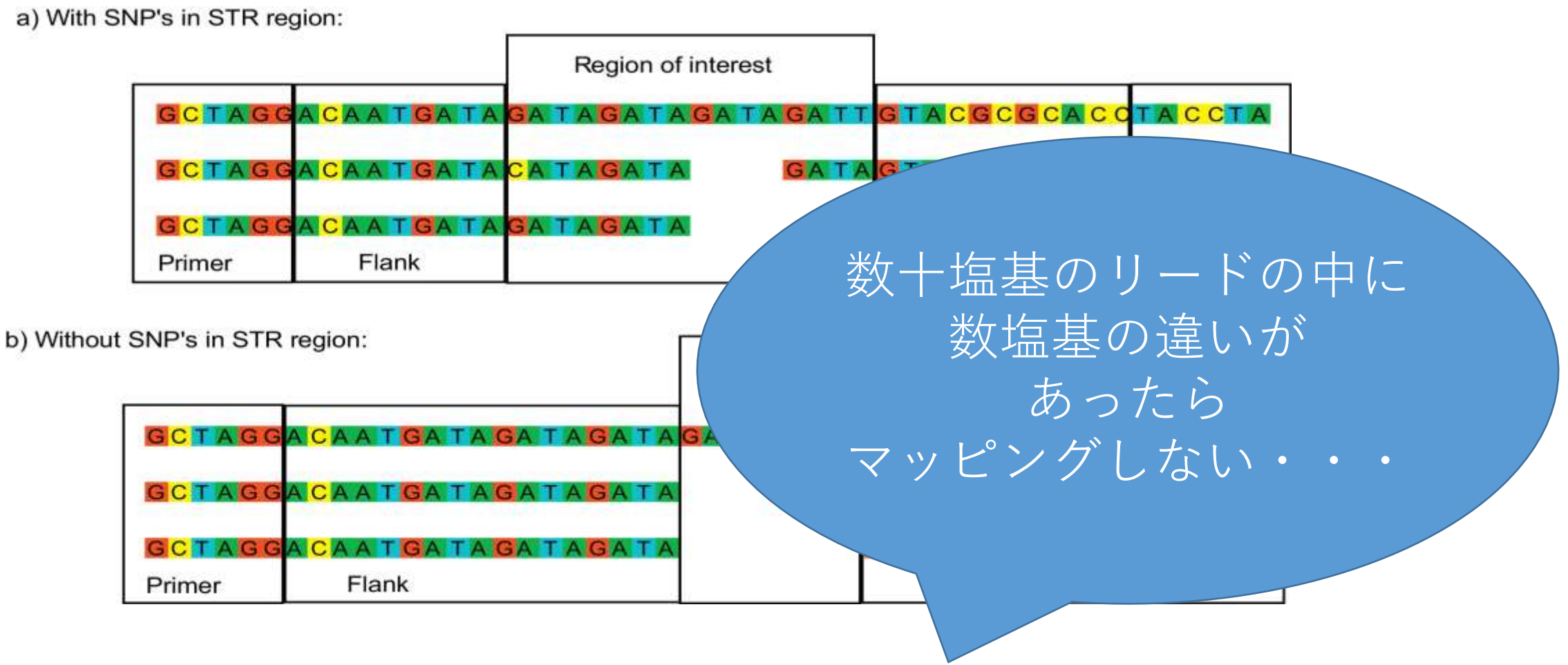


b) Without SNP's in STR region:



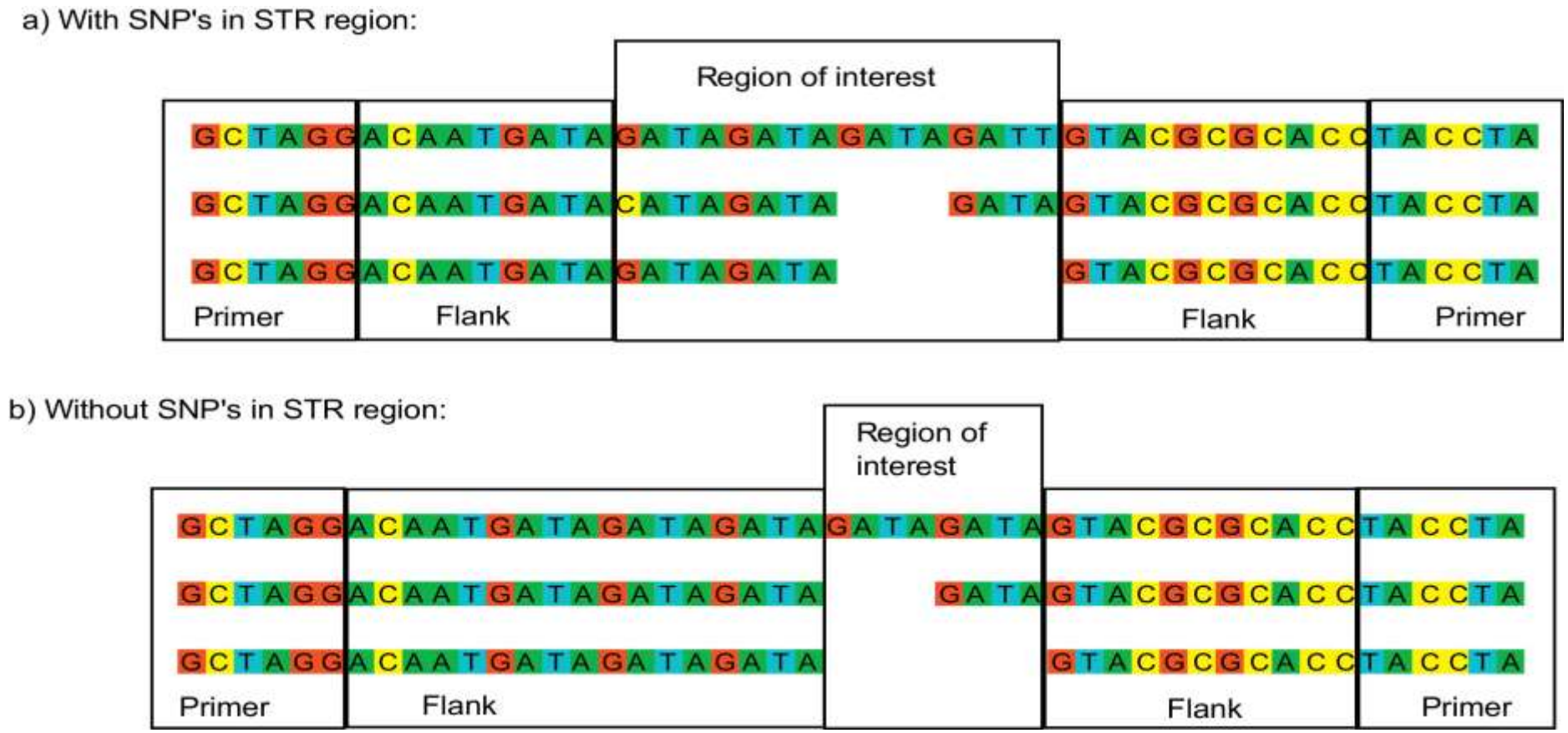
NGS通常ツールは、１塩基違い、２塩基違いを問題にする

Forensic Science International: Genetics 2014 91-8DOI: (10.1016/j.fsigen.2013.10.012)



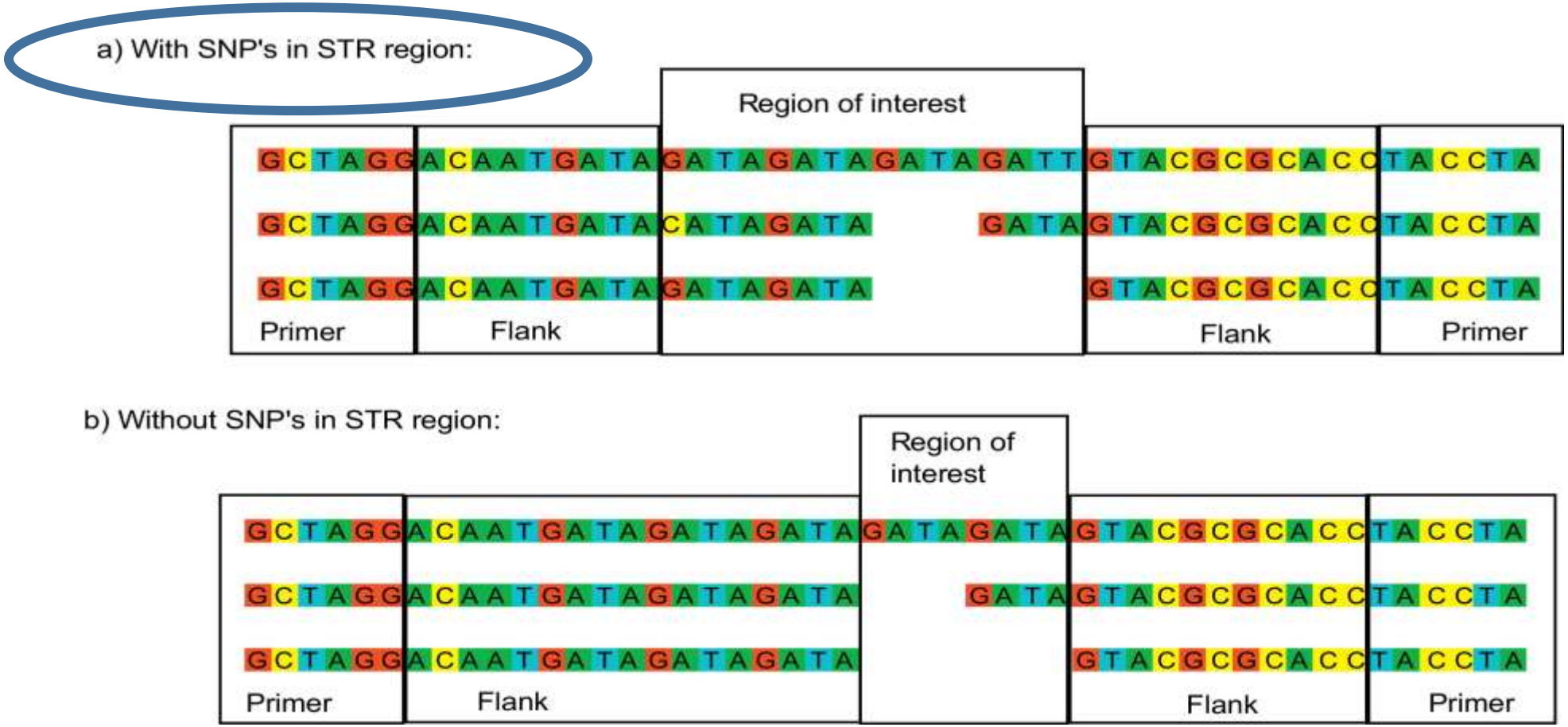
NGS通常ツールは、１塩基違い、２塩基違いを問題にする

Forensic Science International: Genetics 2014 91-8DOI: (10.1016/j.fsigen.2013.10.012)



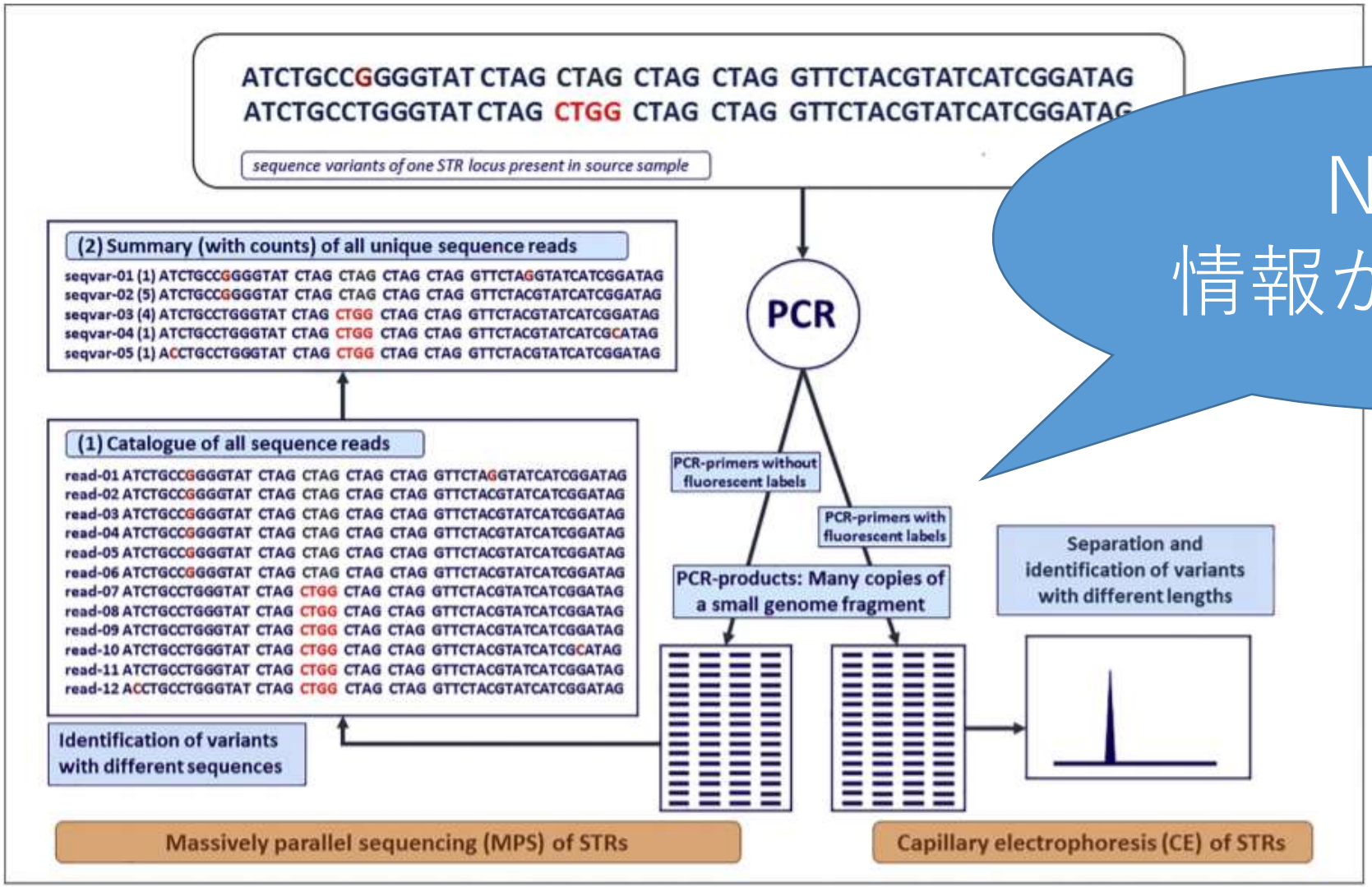
少数箇所の塩基違いは、真の配列違いかもしれないし
シークエンスエラーかもしれない

Forensic Science International: Genetics 2014 91-8DOI: (10.1016/j.fsigen.2013.10.012)



リピート配列にSNPがあると
厳密には「リピート」ではないことになる

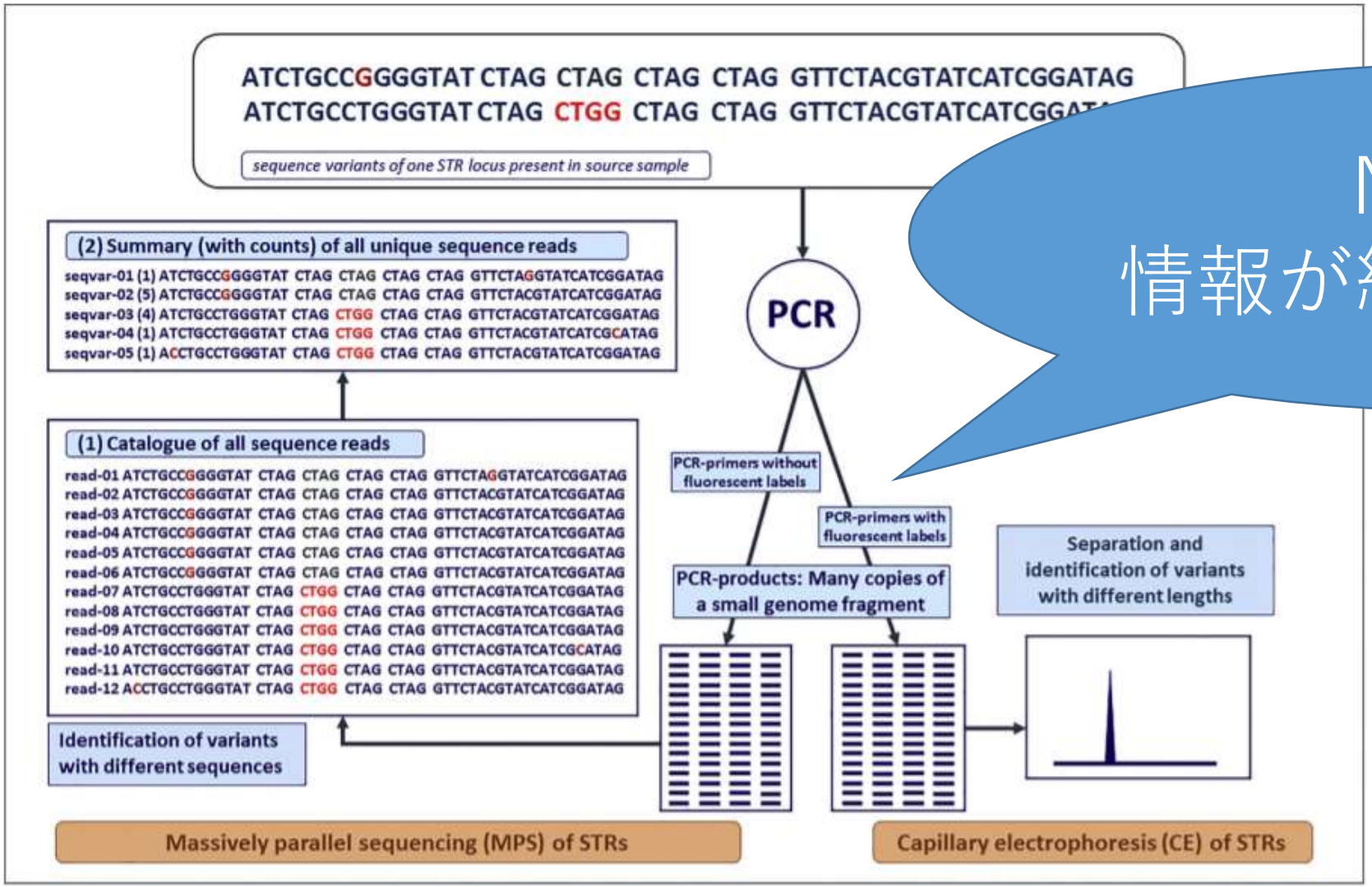
Forensic Science International: Genetics 2014 91-8DOI: (10.1016/j.fsigen.2013.10.012)



NGSは
情報が細かい！

Fig. 1. A simplified explanation of the main differences between CE genotyping of STRs

PCR is designed
ingle STR
ent al-
e con-
t unit.
in the
tation 7 bp.
unit. When using CE,
the PCR uses a set of PCR primers of which one primer has a fluorescent label. After PCR, all PCR products will include this fluorochrome, enabling the detecting of the PCR products on a CE platform. As both alleles are 4 repeats long, only a single peak will be visible. For this locus, this DNA sample will be called a homozygote when using CE. When using MPS, the PCR primers are not labeled with a fluorochrome. All PCR products are simply sequenced on a MPS platform, resulting in a long list of sequence reads. For the sake of simplicity, in this example 12 sequence reads are shown (panel 1). Six reads (01 – 06) contain the T > G mutation in the 5′flanking region. Six other reads (07–12) contain the CTGG repeat. Upon further inspection, three more genetic variants are detected. A single read (01) contains a C > G error 7 bp. 3′of the repeat structure. Another read [10] contains a G > C error 16 bp. 3′of the repeat structure. Finally, read 12 shows a T > C error at the second position of the sequence. This full spectrum of sequence variation is summarized in panel 2. A total of five different sequence variants were detected in this sample. Two variants (02 and 03) were seen multiple times and probably reflect the true alleles. A further three variants (01, 04, and 05) are seen only once; their defining SNP likely representing error reads: reads containing PCR or sequence induced sequence errors.



NGSは
情報が細かすぎる！

Fig. 1. A simplified explanation of the main differences between CE genotyping of STRs

igned

bp.

When using CE, the PCR uses a set of PCR primers of which one primer has a fluorescent label. After PCR, all PCR products will include this fluorochrome, enabling the detecting of the PCR products on a CE platform. As both alleles are 4 repeats long, only a single peak will be visible. For this locus, this DNA sample will be called a homozygote when using CE. When using MPS, the PCR primers are not labeled with a fluorochrome. All PCR products are simply sequenced on a MPS platform, resulting in a long list of sequence reads. For the sake of simplicity, in this example 12 sequence reads are shown (panel 1). Six reads (01 – 06) contain the T > G mutation in the 5′flanking region. Six other reads (07–12) contain the CTGG repeat. Upon further inspection, three more genetic variants are detected. A single read (01) contains a C > G error 7 bp. 3′of the repeat structure. Another read [10] contains a G > C error 16 bp. 3′of the repeat structure. Finally, read 12 shows a T > C error at the second position of the sequence. This full spectrum of sequence variation is summarized in panel 2. A total of five different sequence variants were detected in this sample. Two variants (02 and 03) were seen multiple times and probably reflect the true alleles. A further three variants (01, 04, and 05) are seen only once; their defining SNP likely representing error reads: reads containing PCR or sequence induced sequence errors.

配列情報から
「SNPなしのリピート」を検出することはできる

Forensic Science International: Genetics 44 (2020) 102194

AGTC [AGAT]10 GGA[TCTA]2 [ACT]3 ATGCA

POSSIBLE STRS:
[AGAT]10 GGA[TCTA]2 [ACT]3
[AGAT]10 GGA[TCTA]2
[TCTA]2 [ACT]3
[AGAT]10
[TCTA]2
[ACT]3

Fig. 1. Schematic of possible STRs enumerated from a read containing multiple repeat stretches. As only the possible STRs in rows 1, 2, and 4 contain the identified repeat stretch (when only one repeat stretch is examined) the rest are discarded.

配列情報から
「大雑把なリピート」を検出するのは難しい

Forensic Science International: Genetics 44 (2020) 102194

AGTC [AGAT]10 GGA[TCTA]2 [ACT]3 ATGCA

POSSIBLE STRS:
[AGAT]10 GGA[TCTA]2 [ACT]3
[AGAT]10 GGA[TCTA]2
[TCTA]2 [ACT]3
[AGAT]10
[TCTA]2
[ACT]3

Fig. 1. Schematic of possible STRs enumerated from a read containing multiple repeat stretches. As only the possible STRs in rows 1, 2, and 4 contain the identified repeat stretch (when only one repeat stretch is examined) the rest are discarded.

色々なツールがある

- (できる範囲で)リピート部と、Flanking部に分離する
- Flanking部をレファレンス配列にマッピングする
- その位置を利用して中央部をマッピングする
- その結果「SNPあり・なしに関わらず」リピート数が決まる

- NGSは配列エラーもあるので
  - たくさんのリードから「高質リード」を選んで「真の配列」とする
    - STRの場所とアレルとを決める
  - 「比較的、高質リード」を選んで
    - それぞれのSTRの場所とアレルとを定め
    - その上で、「一番、『らしい』」ものを選ぶ

色々なツールがある

難しい

- (できる範囲で)リピート部と、Flanking部に分離する
- Flanking部をレファレンス配列にマッピングする
- その位置を利用して中央部をマッピングする
- その結果「SNPあり・なしに関わらず」リピート数が決まる

- NGSは配列エラーもあるので
  - たくさんのリードから「高質リード」を選んで「真の配列」とする
    - STRの場所とアレルとを決める
  - 「比較的、高質リード」を選んで
    - それぞれのSTRの場所とアレルとを定め
    - その上で、「一番、『らしい』」ものを選ぶ

今回の紹介論文のツールは

- 難しい「Flanking部をレファレンス配列にマッピングする」ステップを機械学習にやらせてみた

- Flanking配列、中央配列の取り出しには、既存法を使い
- 「比較的、高質の配列」について、STRローカスの予測を機械学習で選択し
- 選ばれたローカスにおけるレファレンス配列との関係から、中央部配列のアレルを決める

学習～教師アリ・教師ナシ

STRタイピングという学習問題の設定方法

- 教師アリ学習するには、「教師」が必要
  - 「教師～正解」ありの情報を使って、『予測モデル』を作る
  - 「正解のない」情報を『予測モデル』に照らして、『正しく予測』したい

- 「絵」の学習は、「この絵は猫」「この絵は猫ではない」という「正解」
- 多変量解析データでは、従属変数の値が「正解」
- STRタイピングでは、何が「正解」か？

STRタイピングという学習問題の設定方法

- 学習するには、「ピース」が必要
- 「絵」の学習は、ピクセル(画素)が「ピース」
- 多変量解析データでは、説明変数が「ピース」
- NGSのShort read(配列情報)は何が「ピース」か？？

ction using a machine learning paradigm

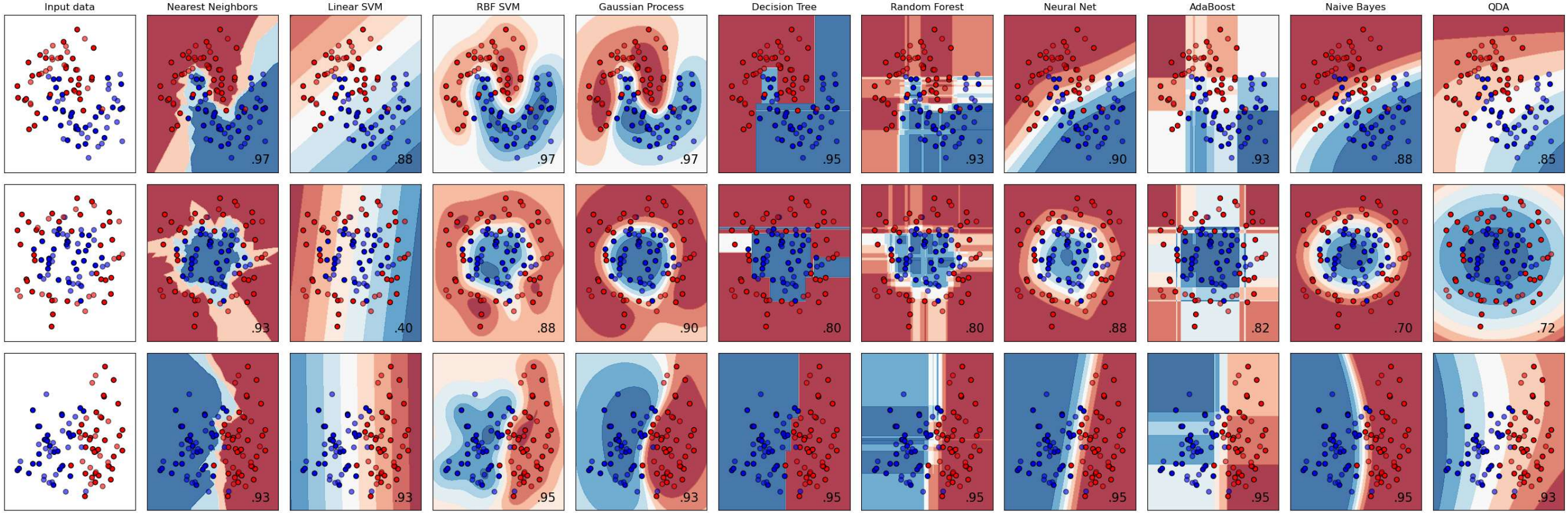
Ryan England[a,c], Janet Stacey[c], SallyAnn Harbison[c,*]

es, University of Auckland, 38 Princes Street, Auckland 1010, New Zealand
38 Princes Street, Auckland 1010, New Zealand
ted, Private Bag 92021, Auckland 1142, New Zealand

ABSTRACT

We present a machine learning approach to short tandem repeat (STR) sequence detection and extraction from massively parallel sequencing data called Fragsifier. Using this approach, STRs are detected on each read by first locating the longest repeat stretches followed by locus prediction using k-mers in a machine learning sequence model. This is followed by reference flanking sequence alignment to determine precise STR boundaries. We show that Fragsifier produces genotypes that are concordant with profiles obtained using capillary electrophoresis (CE), and also compared the results with that of STRait Razor and the ForenSeq UAS. The data pre-processing and training of the sequence classifier is readily scripted, allowing the analyst to experiment with different thresholds, datasets and loci of interest, and different machine learning models.

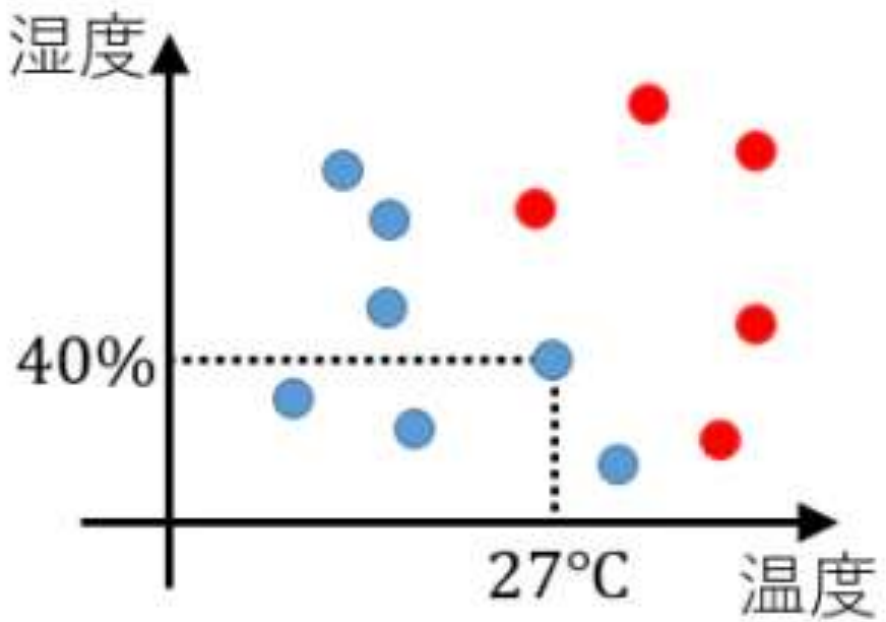
分類：教師アリ


Input data
Nearest Neighbors
Linear SVM
RBF SVM
Gaussian Process
Decision Tree
Random Forest
Neural Net
AdaBoost
Naive Bayes
QDA
.97
.88
.97
.97
.95
.93
.90
.93
.88
.85
.93
.40
.88
.90
.80
.80
.88
.82
.70
.72
.93
.93
.95
.93
.95
.95
.95
.95
.95
.93

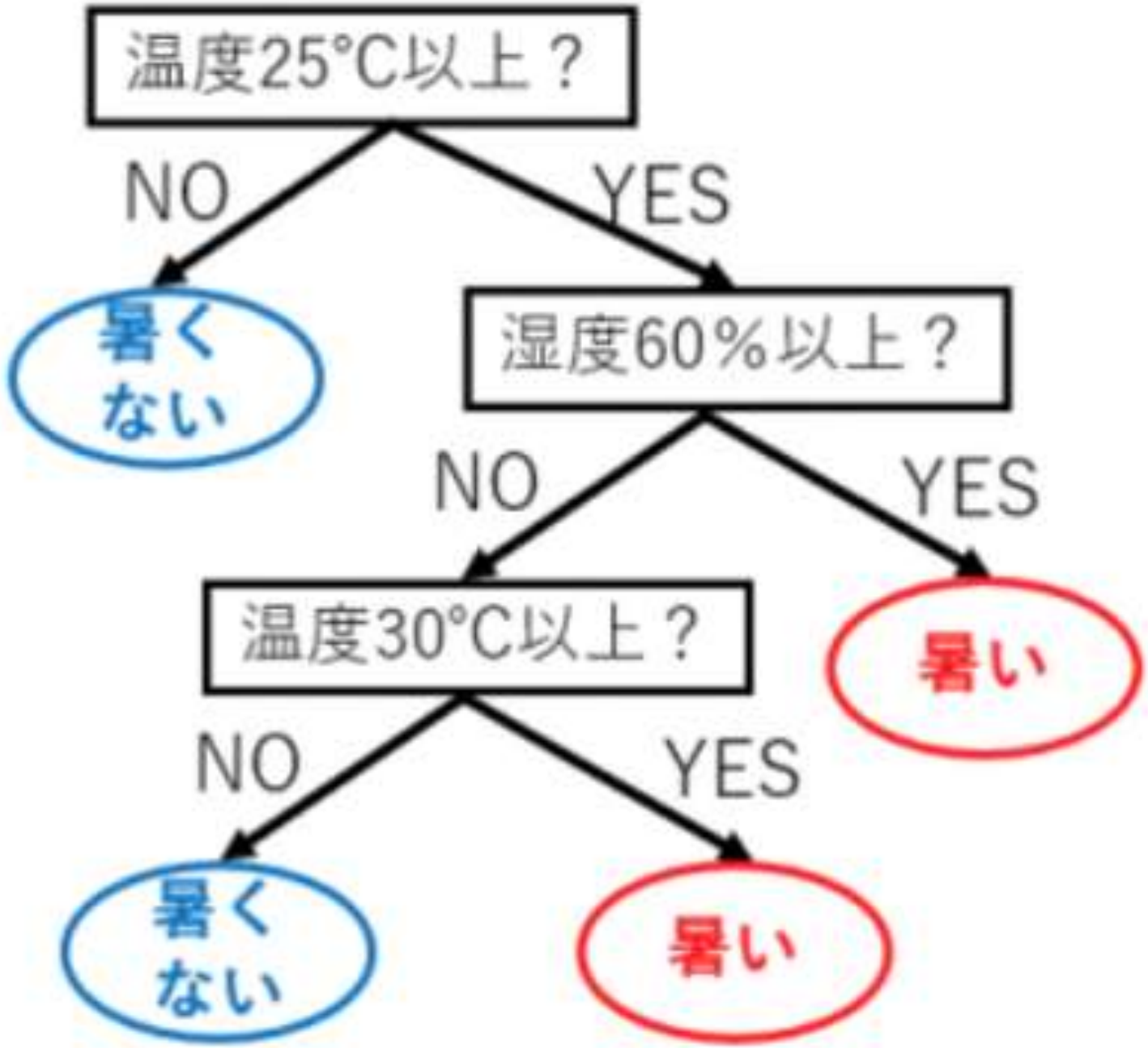
決定木

分類木(Classification Tree)

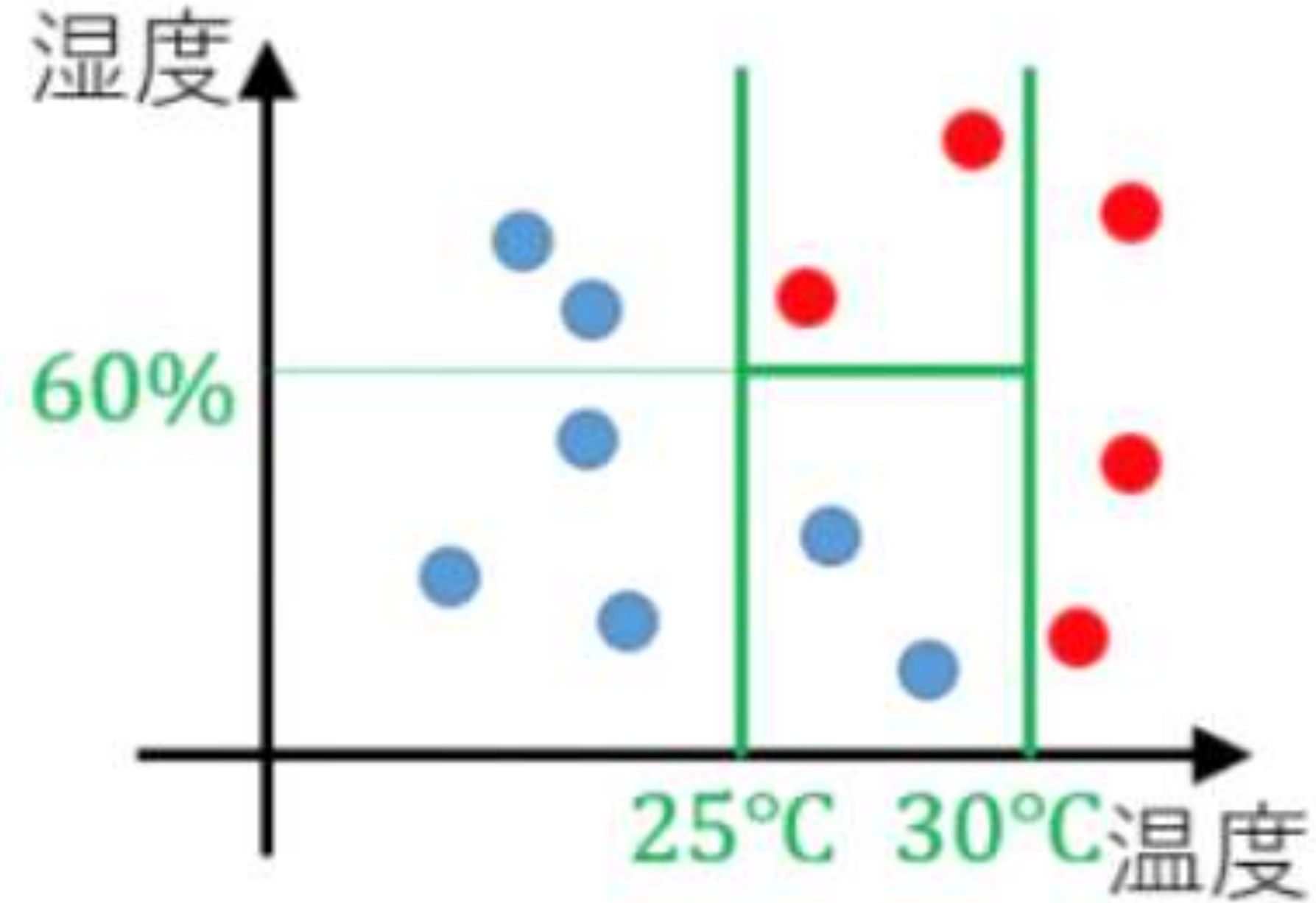
日々の温度と湿度のデータ，その日Aさんが暑いと感じたか暑くないと感じたかかのデータが与えられた状況を考えます．
赤い点が暑いと感じた日，青い点が暑くないと感じた日です．



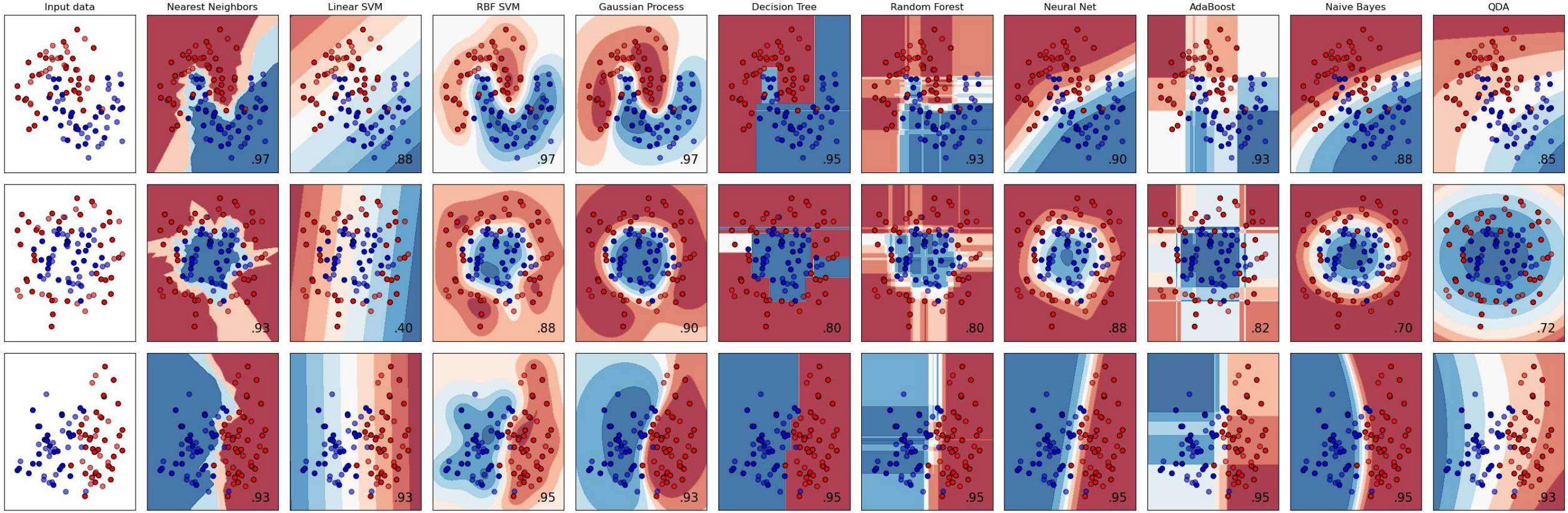
https://qiita.com/3000manJPY/items/ef7495960f472ec14377

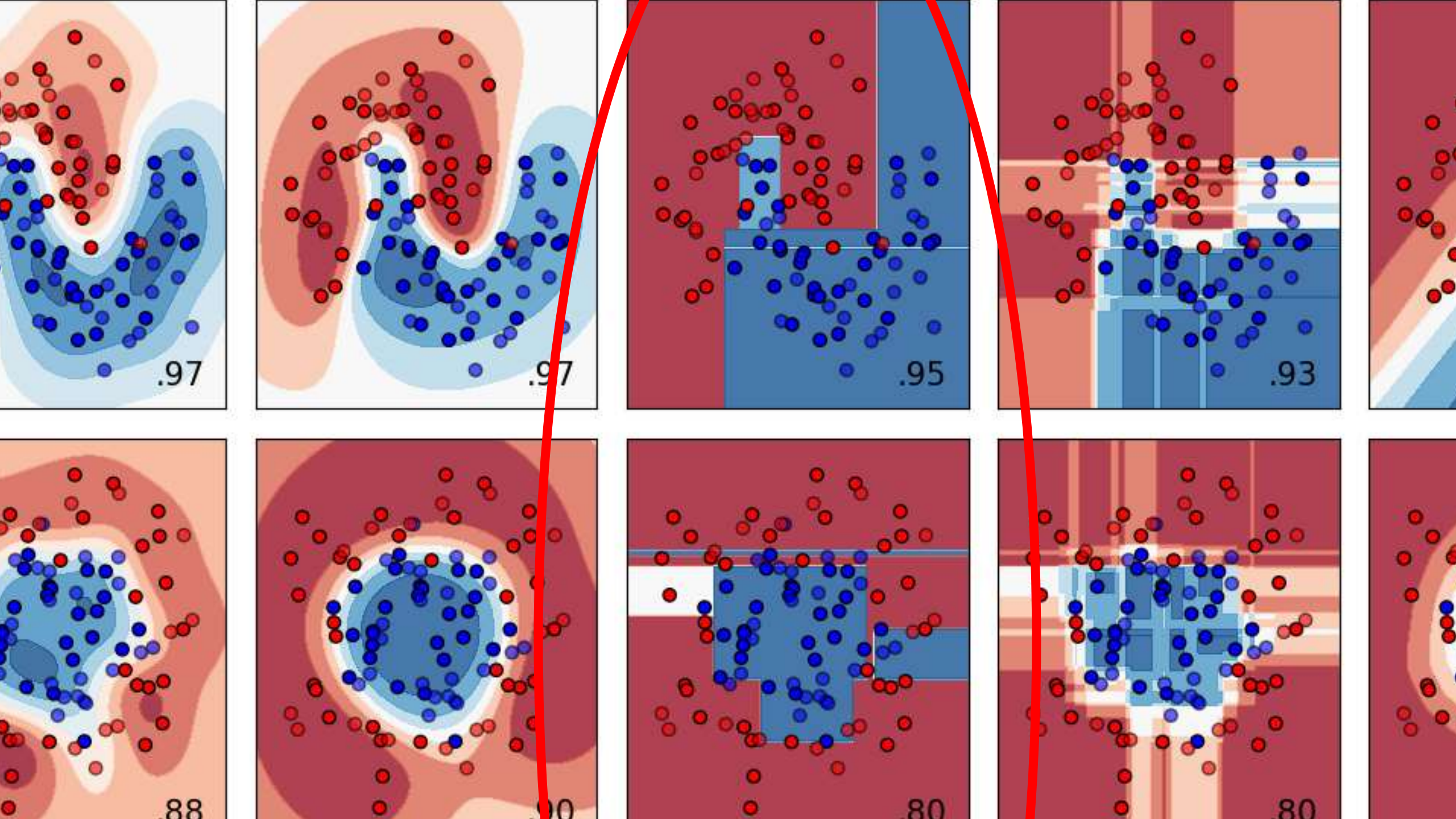


https://qiita.com/3000manJPY/items/ef7495960f472ec14377

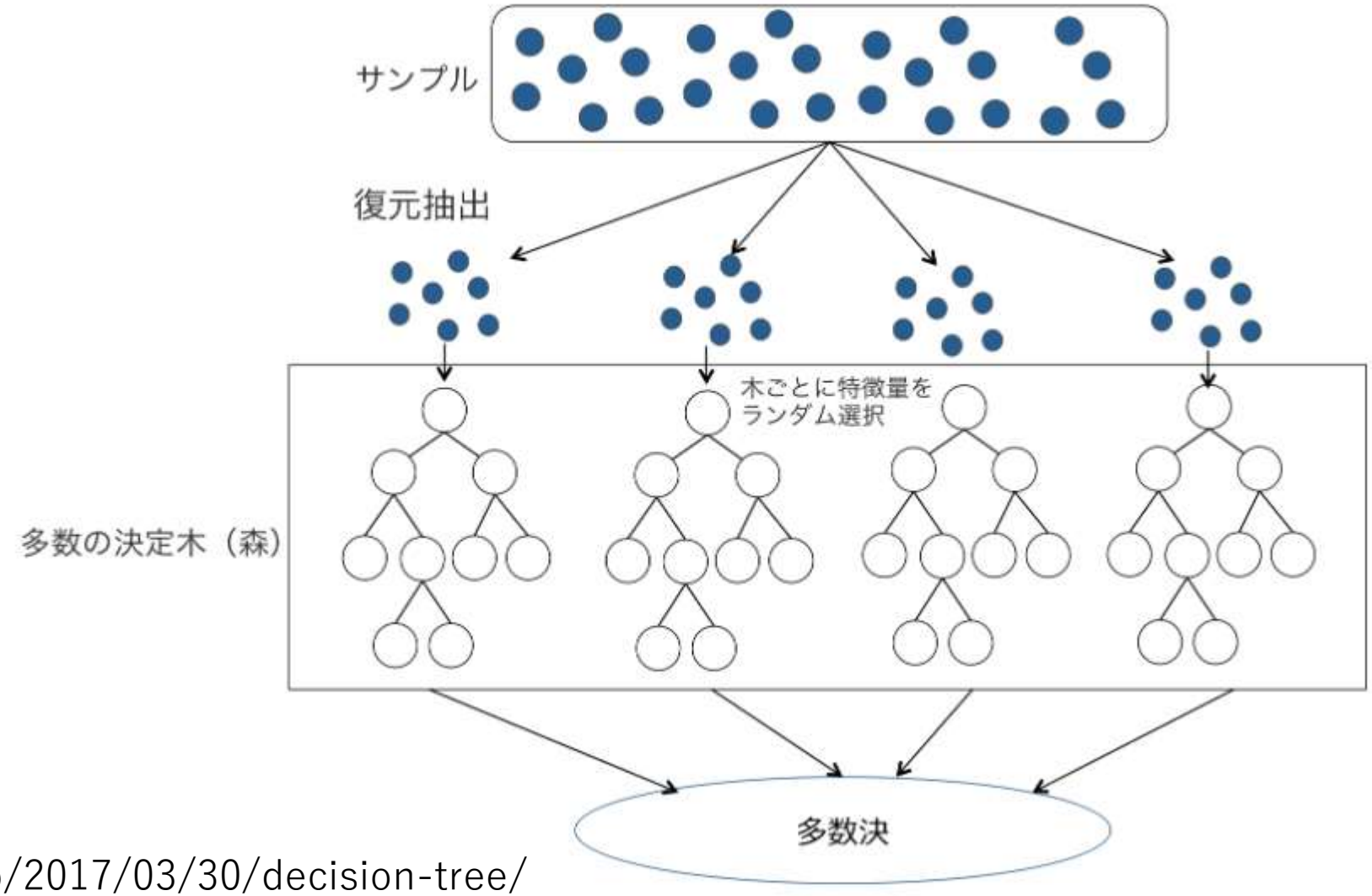


https://qiita.com/3000manJPY/items/ef7495960f472ec14377

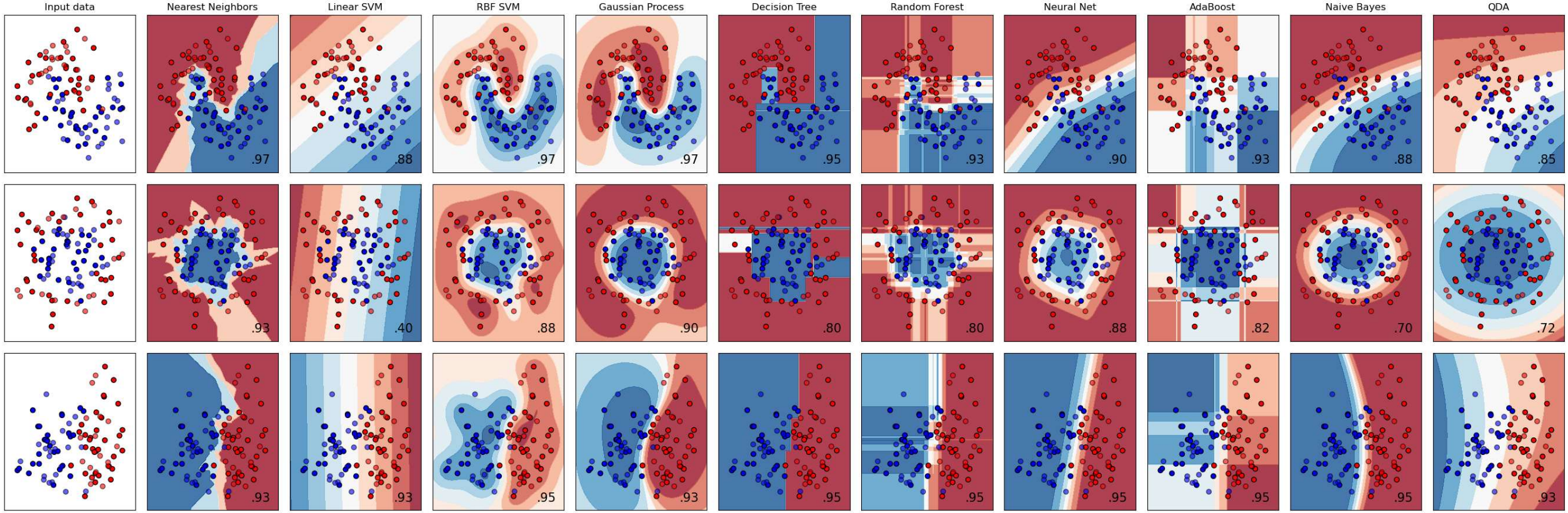

Input data
Nearest Neighbors
Linear SVM
RBF SVM
Gaussian Process
Decision Tree
Random Forest
Neural Net
AdaBoost
Naive Bayes
QDA
.97
.88
.97
.97
.95
.93
.90
.93
.88
.85
.93
.40
.88
.90
.80
.80
.88
.82
.70
.72
.93
.93
.95
.93
.95
.95
.95
.95
.95
.93

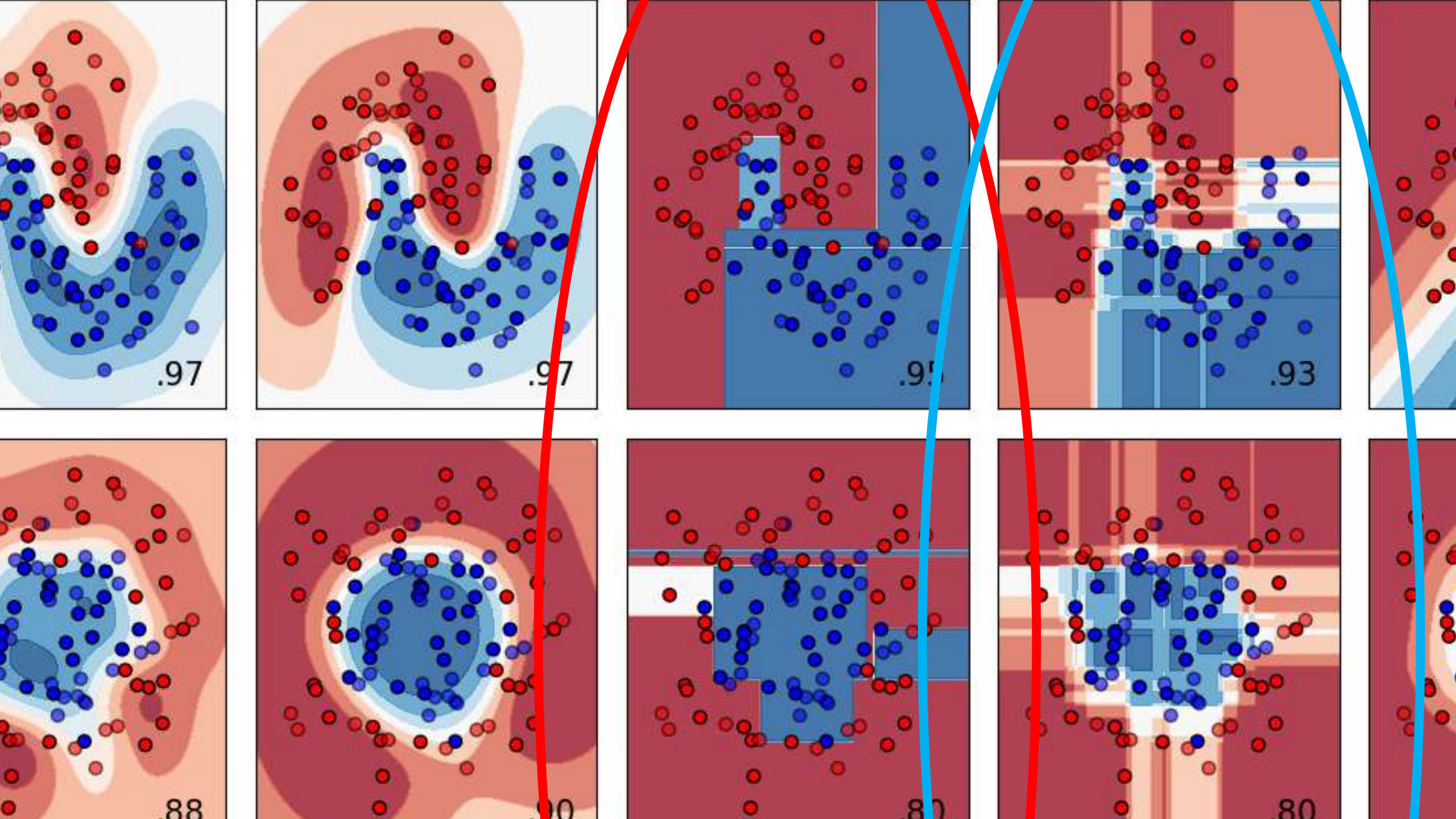

.97
.97
.95
.93
.88
.80
.80

ランダムフォーレスト



https://alphaimpact.jp/2017/03/30/decision-tree/

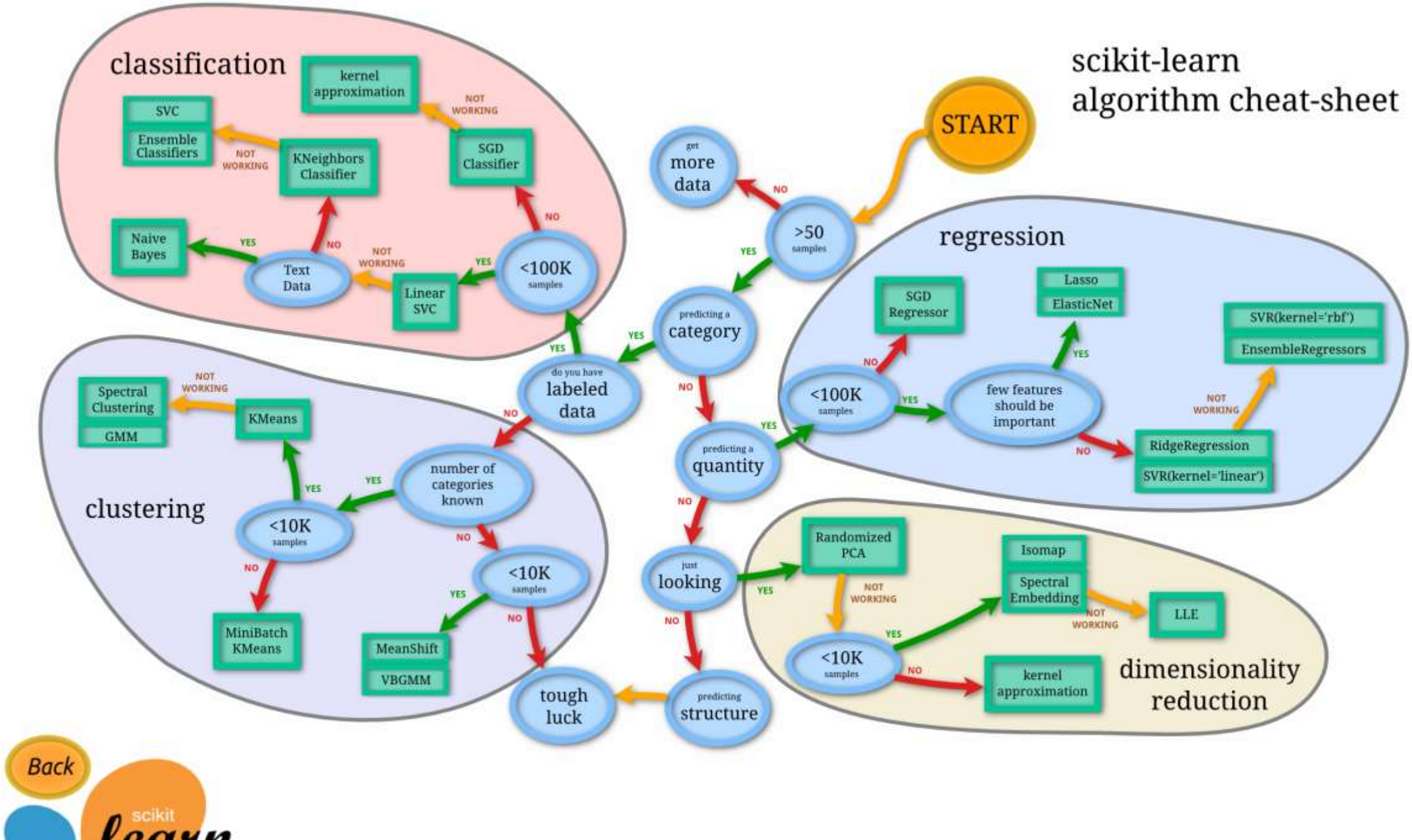

Input data
Nearest Neighbors
Linear SVM
RBF SVM
Gaussian Process
Decision Tree
Random Forest
Neural Net
AdaBoost
Naive Bayes
QDA
.97
.88
.97
.97
.95
.93
.90
.93
.88
.85
.93
.40
.88
.90
.80
.80
.88
.82
.70
.72
.93
.93
.95
.93
.95
.95
.95
.95
.95
.93


.97
.97
.95
.93
.88
.80
.80

決定木とランダムフォーレストの違い

- いくつの木？
  - １　vs.　複数（多数）
- 曖昧さ
  - なし　vs.　あり
- 過剰適合・オーバーフィッティング
  - 深さで調節　vs.　均して調節

機械学習　一般に


scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
NO
Text Data
NOT WORKING
Linear SVC
YES
<100K samples
NO
get more data
NO
>50 samples
YES
YES
predicting a category
YES
do you have labeled data
NO
regression
SGD Regressor
NO
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
<100K samples
YES
few features should be important
YES
NOT WORKING
NO
RidgeRegression
SVR(kernel='linear')
NO
predicting a quantity
YES
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
just looking
YES
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
Back
scikit learn