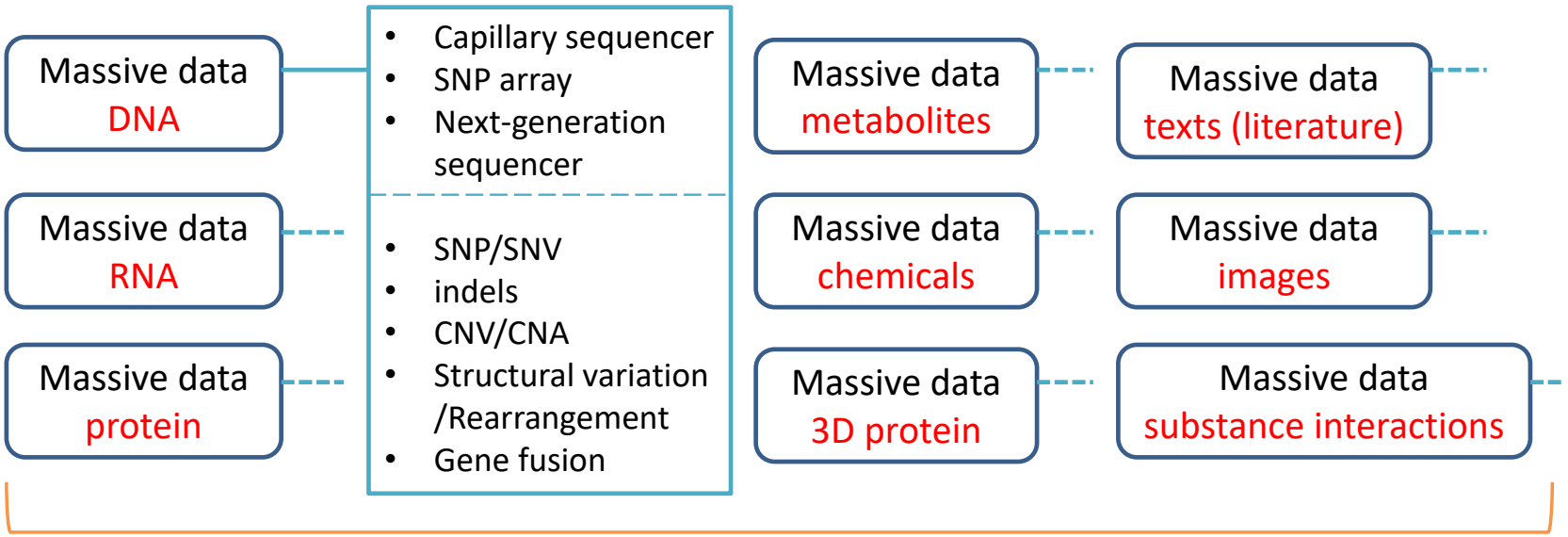
バイオインフォマティクスの概要

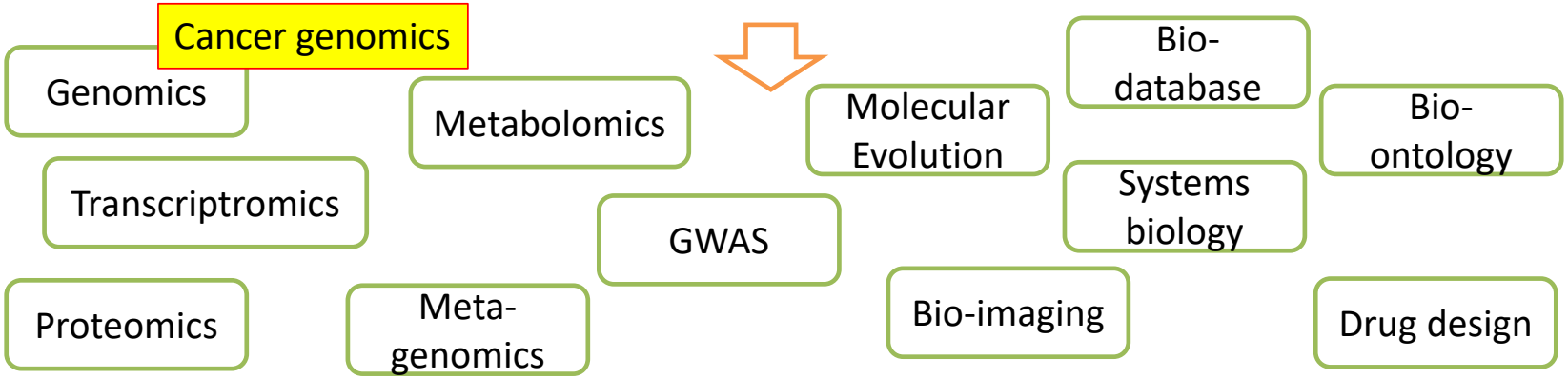
研究所　　生物情報学分野
研究所　バイオインフォマティクス部門
C-CAT　　ゲノムデータ管理室
加藤　護

Outline

1. Rough scope of bioinformatics
   – Based on my limited experiences though...
2. Application examples of bioinformatics
   1. Data analysis in cancer genomics
   2. Software development in cancer genome medicine

Bioinformatics: greedy and cloudy field

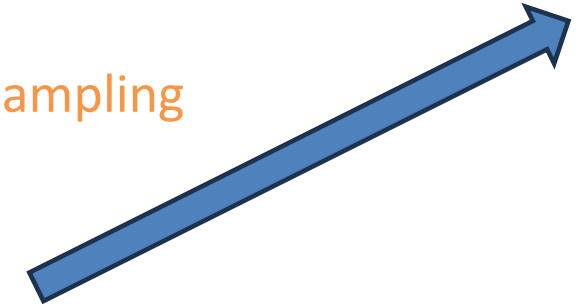




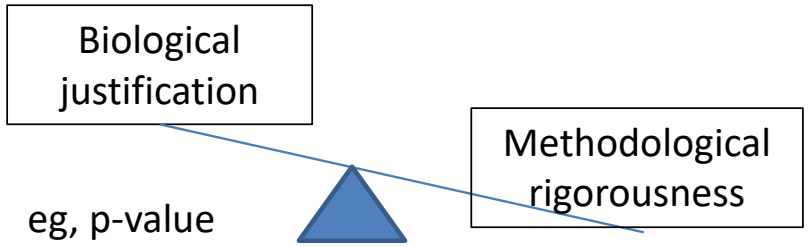
The methodological aspect

The theoretical basis

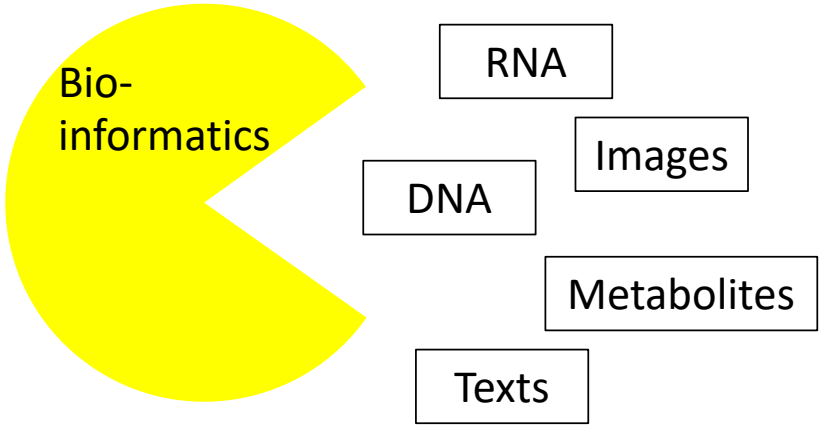
- Mechanics:
  Newton's three laws
- Statistics:
  population and sampling



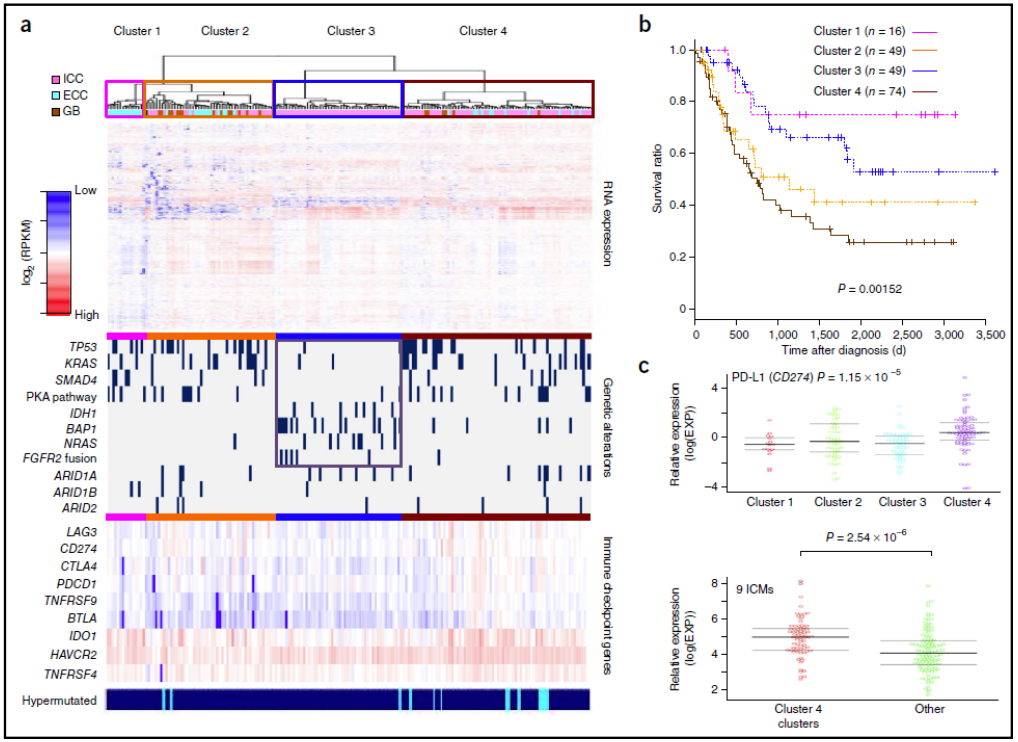
- Bioinformatics:
  NOTHING...?



- Collection of arts?
- The spirit of "pragmatism"
  - ✓Whatever method, as long as useful



Example: RNA analysis in bile-duct cancer



(Nakamura et al, Nat Genet, 2015)

- Bioinformatics support for Shibata group in NCC

NGS data 6 TB data

```
@PERI8:9:45
CCCTCAGCTACGGGGGGGGGGTGGCTTCTTCCTGTTCACCTGGTG
GTGGCGGCTGTGACGCTCCTGCTGCTGCGCAGCCCCAGAACGGC
CGGAGCCATCCCACGCGCTACCGTACCGGCGACATCGATCCAAT
GATACGCGGCTGAGCACA
+
/0(,..0***0000000000%02-..(15030111/322-**%-,(03/24)++-
22/+++230000.+++.2111----%***(**-1,1/*+-(-
**2++*+**/1,0(0..0.4%+++4223+++4*.).***+*024%++2+**+,
...
```

200 columns (samples)

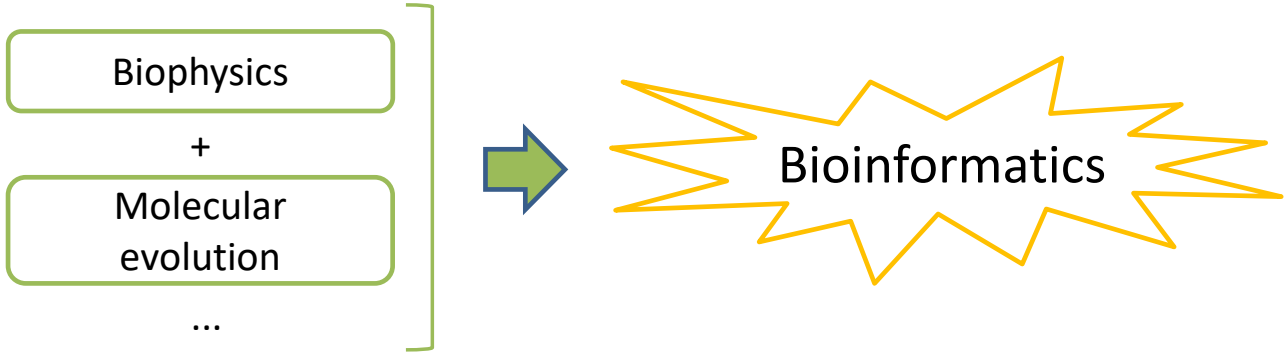
40,000 rows (transcripts)

Gene	BD003T	BD004T	BD005T	BD006T	BD007T
ENST000(	31.35851	81.2562	58.13853	35.76353	40.48326
ENST000(	10.01731	1.137802	32.82091	15.14492	2.095884
ENST000(	0	0	0	0	0
ENST000(	3.982066	1.120111	1.371183	5.04892	2.619011
ENST000(	0	0	0	0	0
ENST000(	0.241376	0.119728	0.021227	0.009749	0
ENST000(	0.061229	0.032396	0.057434	0.039568	0.093569
ENST000(	0	0.581962	0	0	0
ENST000(	146.4966	163.5045	205.3889	162.6099	96.99319
ENST000(	0	0	0	0	0
ENST000(	0	0	0	0	0
ENST000(	8.933542	1.986797	2.840649	4.501061	0.370228
ENST000(	3.923663	0.688758	2.00794	1.949078	0.777532

Clustering analysis

Birth of bioinformatics

- It's in 1970s



Protein sequence **1 letter = 1 amino acid*

1. MKILETPFASGDLSMLVLLPDEVSDLERIEKTINFE...
2. MKILETPFASGDLSMLVLNPDEVSDLERIEKFINFE...
3. MKILETPFSSGDLSMLVLIPDEVSDLERIEKTINFE...

...

How about sorting out sequences obtained so far...?

How different?

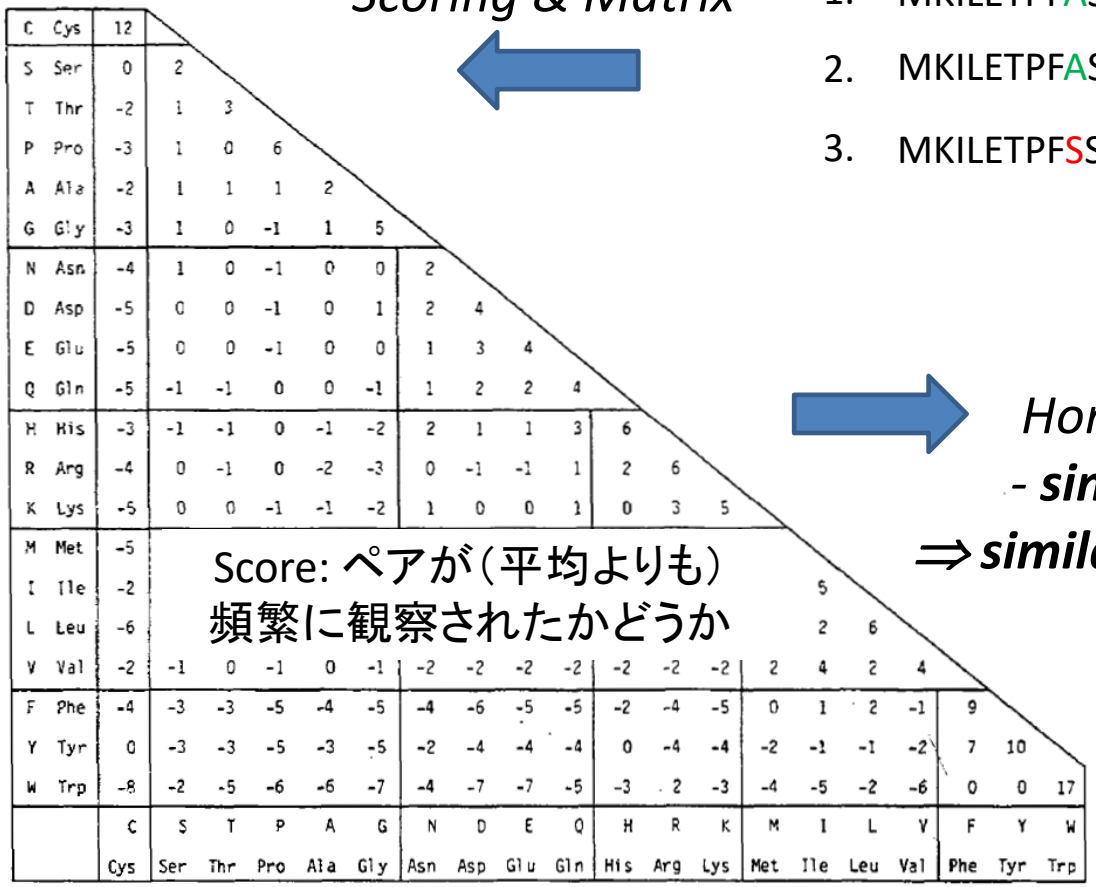
Computer!

Dayhoff matrix

Homology search

Dayhoff matrix & homology search

Scoring & Matrix



C Cys	12																						
S Ser	0	2																					
T Thr	-2	1	3																				
P Pro	-3	1	0	6																			
A Ala	-2	1	1	1	2																		
G Gly	-3	1	0	-1	1	5																	
N Asn	-4	1	0	-1	0	0	2																
D Asp	-5	0	0	-1	0	1	2	4															
E Glu	-5	0	0	-1	0	0	1	3	4														
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4													
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6												
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6											
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5										
M Met	-5													5									
I Ile	-2													2	6								
L Leu	-6																						
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4						
F Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9					
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10				
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17			
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W		
	Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp			

Score: ペアが（平均よりも）
頻繁に観察されたかどうか

(Dayhoff et al, 1978)

1. MKILETPFASGDLSMLVLLPDEVSDLERIEKTINFE...
2. MKILETPFASGDLSMLVLNPDEVSDLERIEKFINFE...
3. MKILETPFSSGDLSMLVLIPDEVSDLERIEKTINFE...

Database

Homology search
- similar <u>sequence</u>
⇒ similar <u>protein function</u>

New. MKILETPGASGDLSMLVLLPDE...

- *Which sequence does this new resemble?*

Homology search

- **DNA**

Matrix

	A	T	G	C
A	5	-4	-4	-4
T	-4	5	-4	-4
G	-4	-4	5	-4
C	-4	-4	-4	5

New: A T G C

Seq1: T T G C

-4 + 5 + 5 + 5 = 11

Seq2: T C G C

-4 - 4 + 5 + 5 = 2

New is more similar to seq1.

- **Protein (amino acid sequence)**

Same idea

Another first technique: “assembly”

- It’s in 1960s
- For amino acid sequencing
  - Here, explained in DNA though
- Simple information processing
  - No biological prediction like in homology search

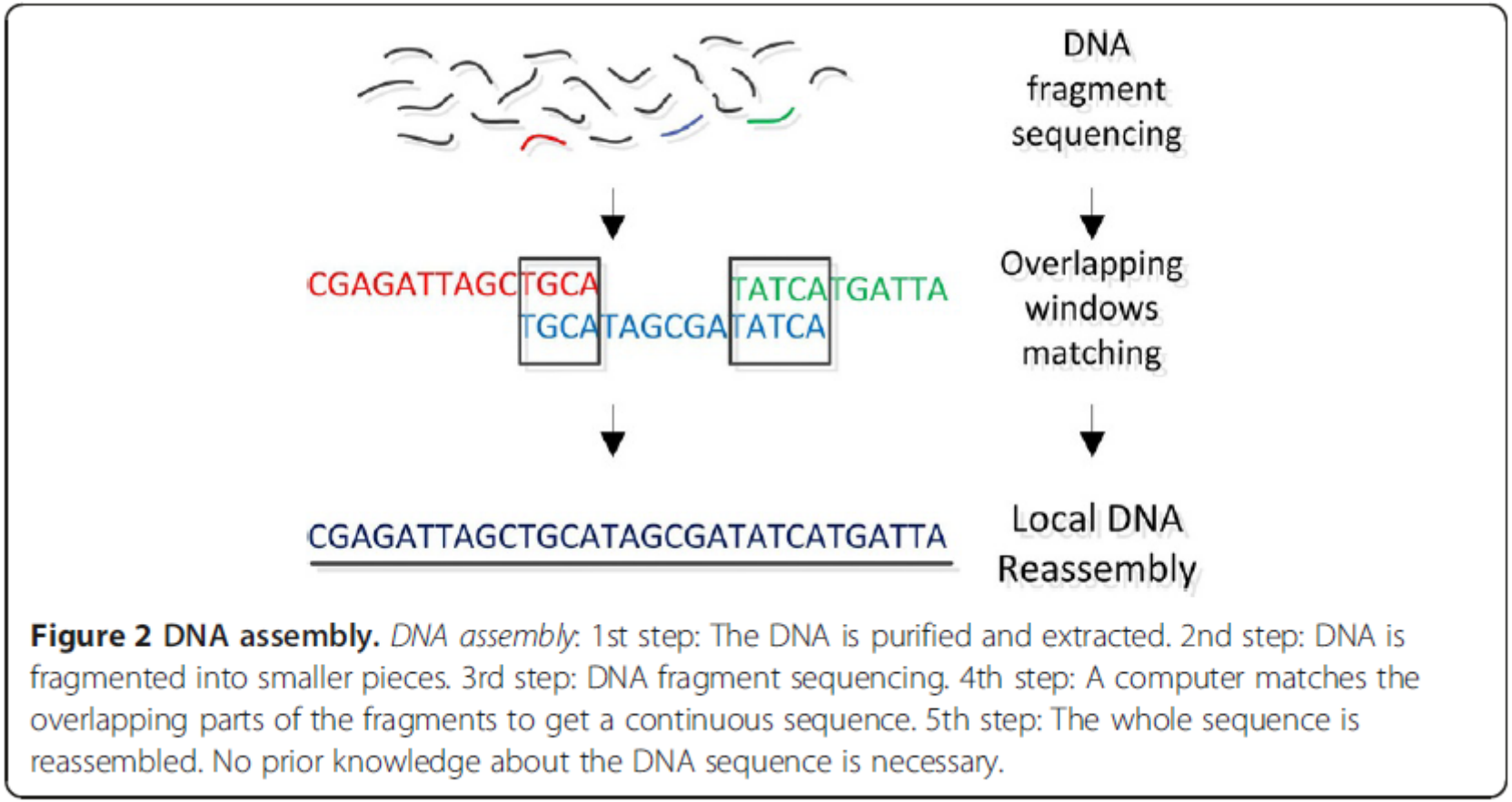
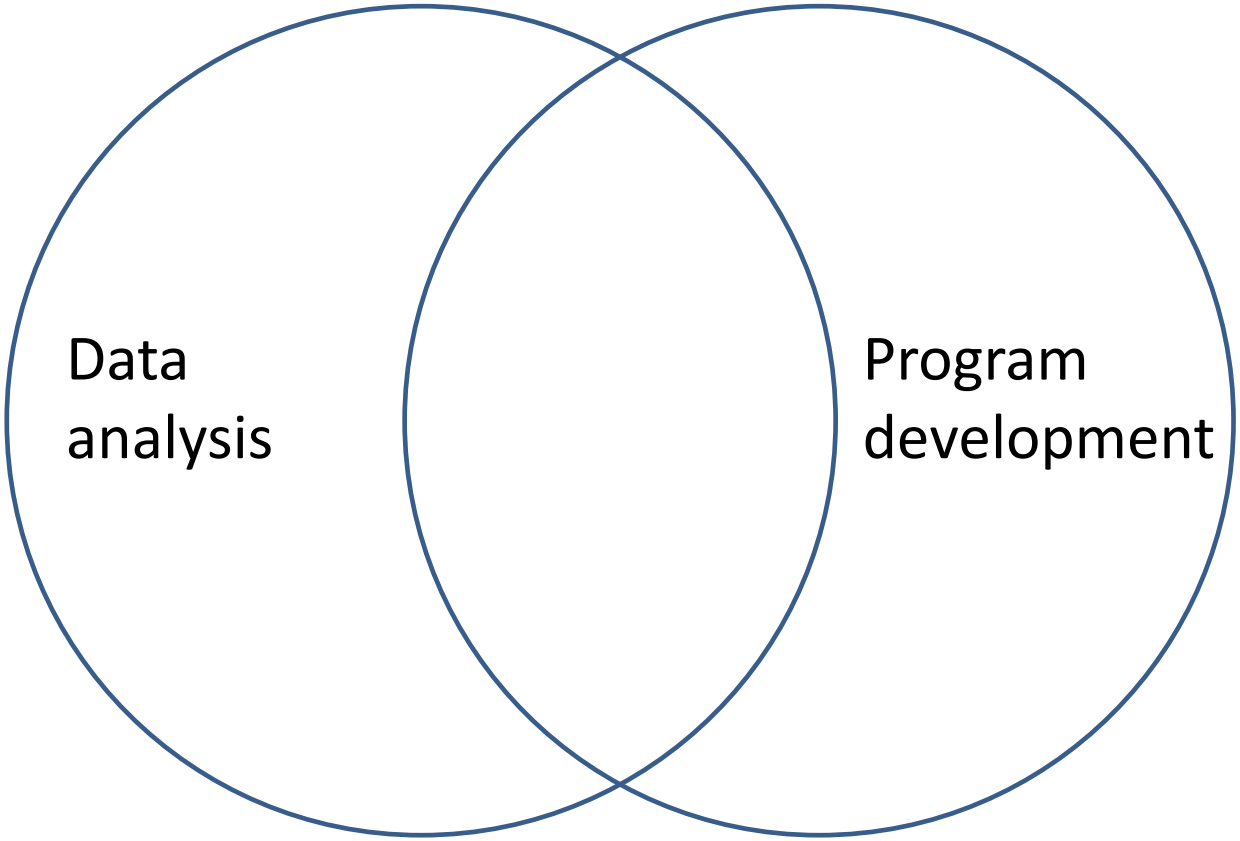


Figure 2 DNA assembly. *DNA assembly*: 1st step: The DNA is purified and extracted. 2nd step: DNA is fragmented into smaller pieces. 3rd step: DNA fragment sequencing. 4th step: A computer matches the overlapping parts of the fragments to get a continuous sequence. 5th step: The whole sequence is reassembled. No prior knowledge about the DNA sequence is necessary.

(Pavlopoulos et al, BioData Mining, 2013)

Two extreme types of studies



Concrete operations: commands and programming

Examples

- Data analysis (of NGS data)

On Linux

```
# make index
bwa index -p human_chrs.fa -a bwtsw human_chrs.fa

# aln
bwa aln -t 6 human_chrs.fa test.fastq         1> test.aln 2> test.aln.err
# samse
bwa samse    human_chrs.fa test.aln test.fastq 1> test.sam 2> test.sam.err

# bwasw
bwa bwasw -t 6 human_chrs.fa test.fastq 1>| tmp.sw.1 2>| tmp.sw.2

# sam -> bam
## with index
samtools view -bS test.sam > test.bam

## no index
samtools faidx    human_chrs.fa (--> .fai) # fasta index
samtools view -bt human_chrs.fa.fai test.sam > test.bam

# bam -> sam
samtools view -h test.bam > test.sam

...
```

- Programming (of catstools)

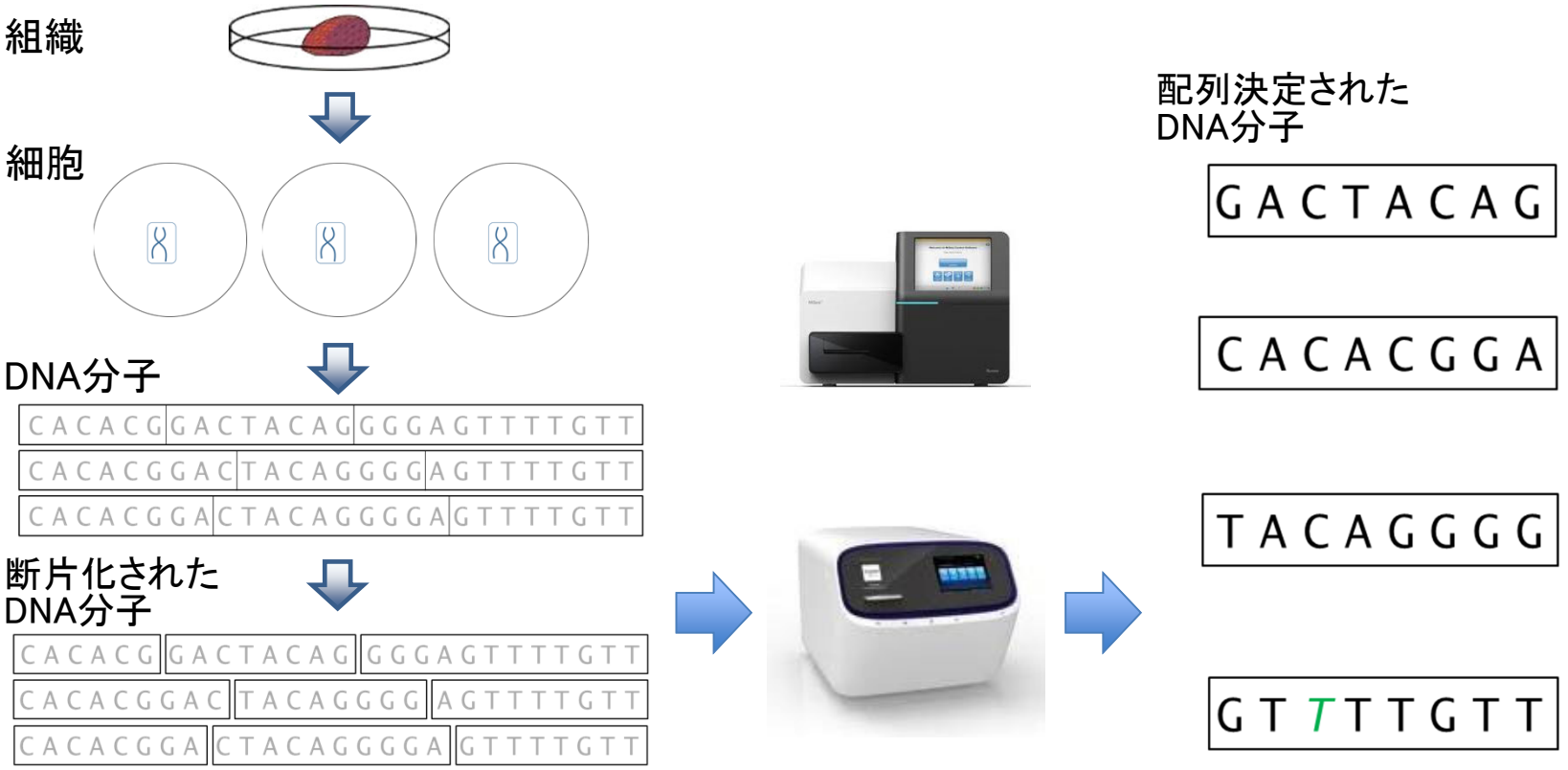
Code Blame 154 lines (121 loc) · 4.71 KB

```python
import logging
import os
from typing import Callable, Dict, List, Optional

from catstools.common import CatsMeta, Variant, filter_variant, iter_variants

logger = logging.getLogger(__name__)


def iter_snv(meta: CatsMeta, filters: List[Callable]):
    """Extracting SNV/insertion/deletion gene mutations"""
    for variant in iter_variants(meta, key='shortVariants'):
        snv = Snv(meta, variant)
        yield from filter_variant(snv, filters)


def filter_somatic(snv: 'Snv'):
    """Provides a filter flag for somatic lineage mutations"""
    return True if snv.origin in ["somatic", "likely somatic"] else False


```

DATA ANALYSIS in cancer genomics

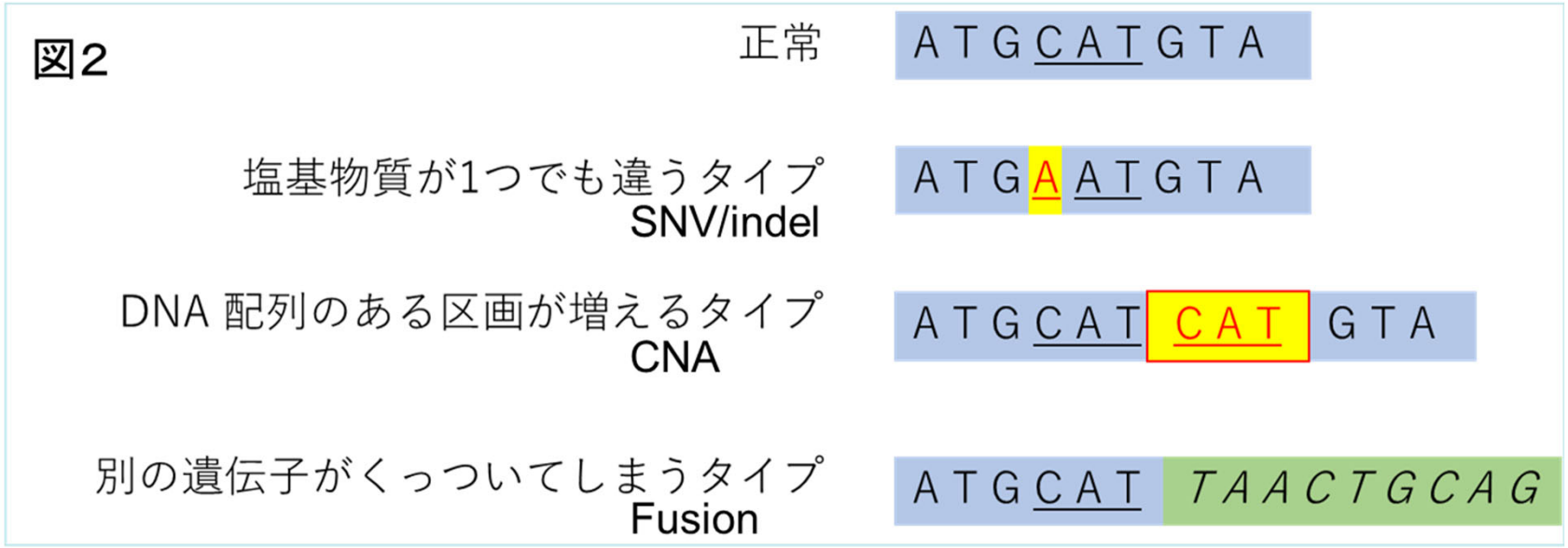
– Examples –

次世代シークエンサー



並列処理技術（DNA を小片に分けて、それぞれを並列処理する）によって、高速にDNA 配列を決定する

ー代表的なDNA変異（異常）の種類ー



（加藤、*アンチ・エイジング医学*、2017）

データ解析項目（バイオインフォマティクス支援項目）

- DNA-seq
  - ▷マッピング
    - ◇BWA-mem
  - ▷変異検出
    - ◇SNV/Indel：cisCall、Mutect2
    - ◇コピー数変化：cisCall、cnvkit
    - ◇構造異常：cisCall、Manta
  - ▷注釈付け
    - ◇ANNOVAR, VEP
    - ◇OncoKB, Cancer Genome Interpreter
- RNA-seq
  - ▷マッピング
    - ◇STAR
  - ▷遺伝子発現量算出
    - ◇FPKM, TPM
  - ▷DEG（differentially expressed genes）解析
- ChIP-seq
  - ▷マッピング
    - ◇Bowtie2
  - ▷ピークコール
    - ◇MACS
- メチル化アレイ
  - ▷ベータ値・M値からの解析

- その他の次世代シークエンス・データ解析
  - ▷TMB（tumor mutation burden）
  - ▷変異シグネチャー解析
    - ◇SigProfiler
  - ▷MSI（microsatellite instability）
    - ◇MSIsensor
  - ▷HRD（homologous recombination deficiency）
    - ◇ShallowHRD, ScarHRD
- 各種データ解析
  - ▷クラスタリング解析
    - ◇hclust, k-means, 系統樹
  - ▷次元圧縮解析
    - ◇PCA, MDS, tSNE, UMAP
  - ▷パスウェイ・GO（gene ontology）解析
    - ◇DAVID
  - ▷GSEA（gene set enrichment analysis）
    - ◇MSigDB
  - ▷生存時間分析
    - ◇カプランマイヤー曲線
    - ◇log-rank 検定
    - ◇Cox 回帰
  - ▷機械学習
    - ◇Random Forest, xgboost
    - ◇LASSO/ridge 回帰、ガウス過程回帰
  - ▷その他データ解析
    - ◇多変量解析、分散分析、各種検定・推定
  - ▷その他応相談

DNA analysis in esophageal squamous cell carcinoma: JCOG0502-A1

Onco-plot

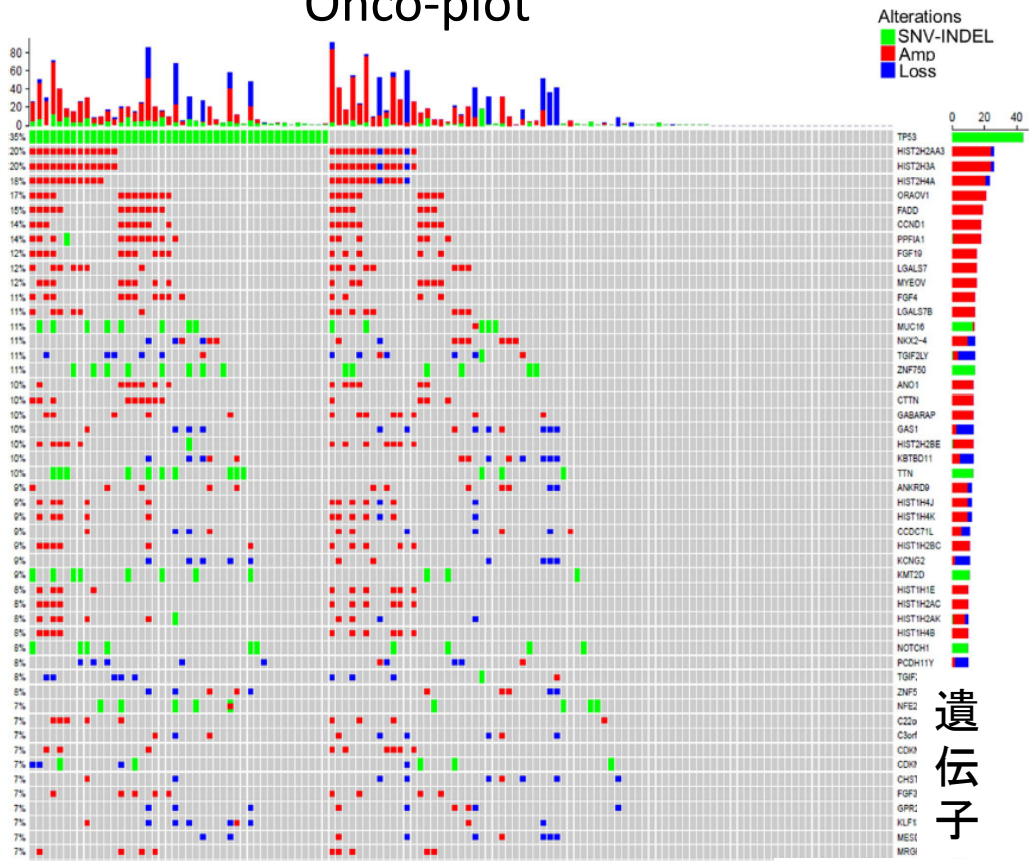


FIGURE 2 Landscape of genomic alterations in 127 T1 esophageal squamous cell carcinoma patients. of genomic alterations per sample. Bottom graph, each genomic alteration for every sample including the total numbe tions for each gene. Left values show the percentage of genomic alterations in every sample for each gene. The genomic alterations are shown in color as follows: green, single nucleotide variant (SNV)-insertion/deletion (INDEL); red, copy number alteration (CNA) amplification (Amp); blue, CNA loss. This figure was generated using the R ComplexHeatmap package[51]

Discovered CNAs along a chromosome

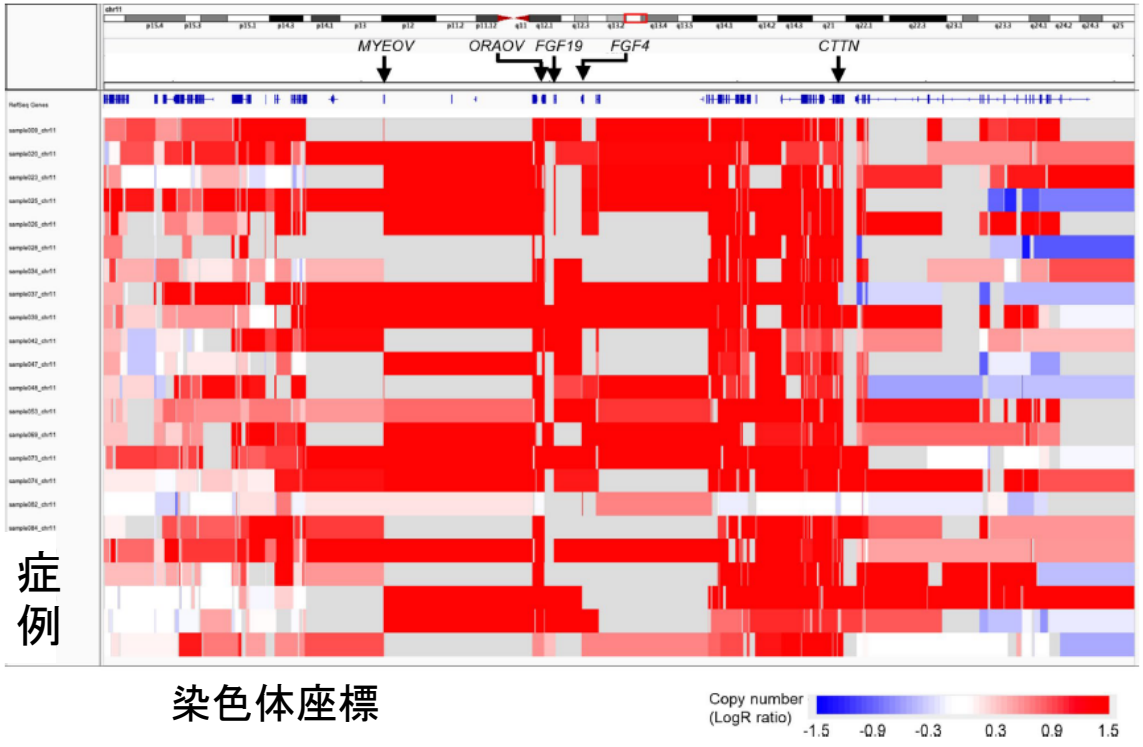


FIGURE 4 Heatmap representation for copy number alterations at chromosome 11q13.3 in 127 patients with T1bN0M0 esophageal squamous cell carcinoma. Red, increased copy number alterations (CNA amplification); blue, decreased copy number alterations (CNA loss). This figure was generated using Integrative Genomic Viewer

(Oshima et al, 2022, *Cancer Science*)

Transcriptome analysis in bile duct cancer

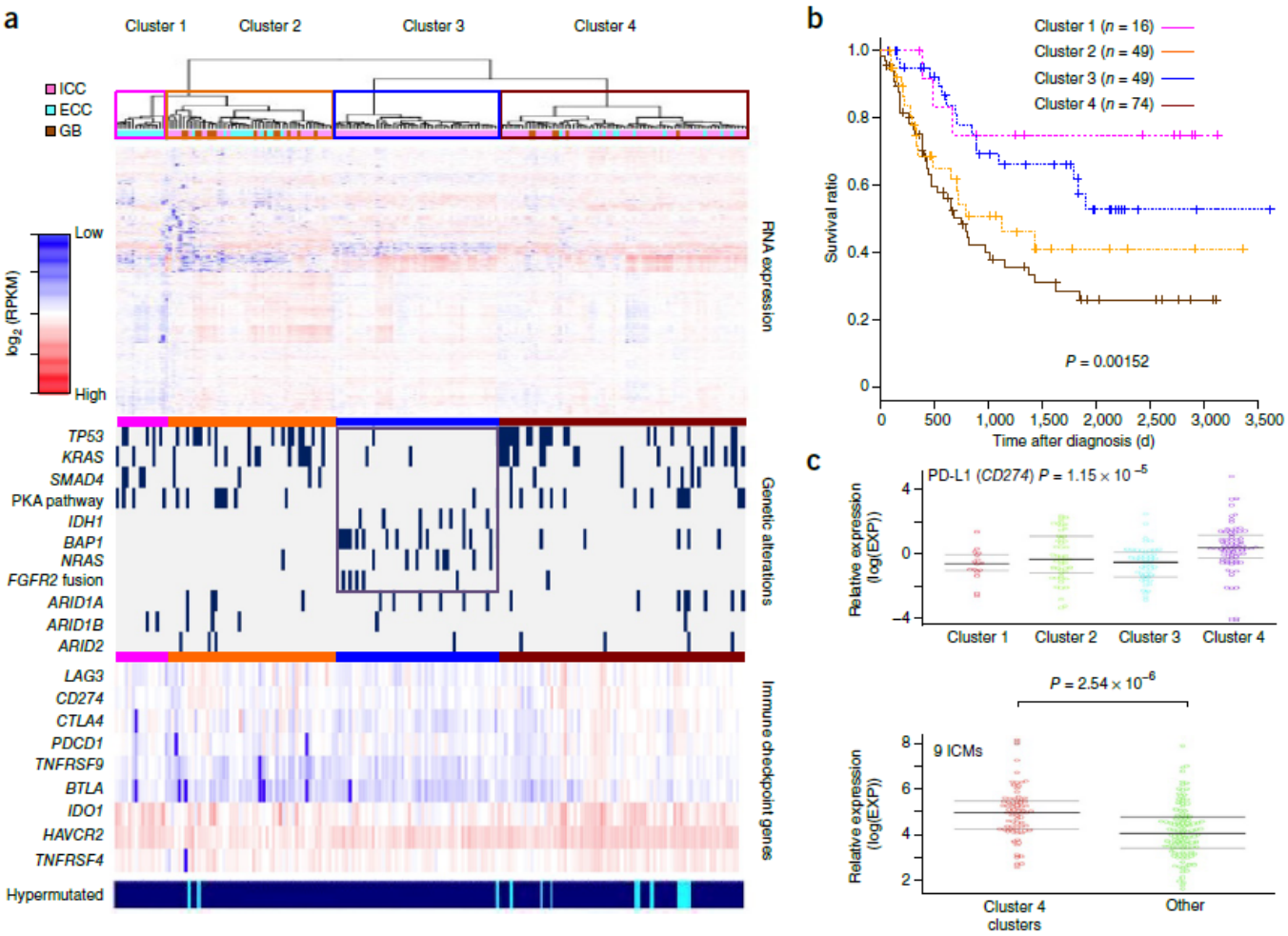


Figure 6 (Nakamura et al, 2015, *Nat Genet*)

Methods

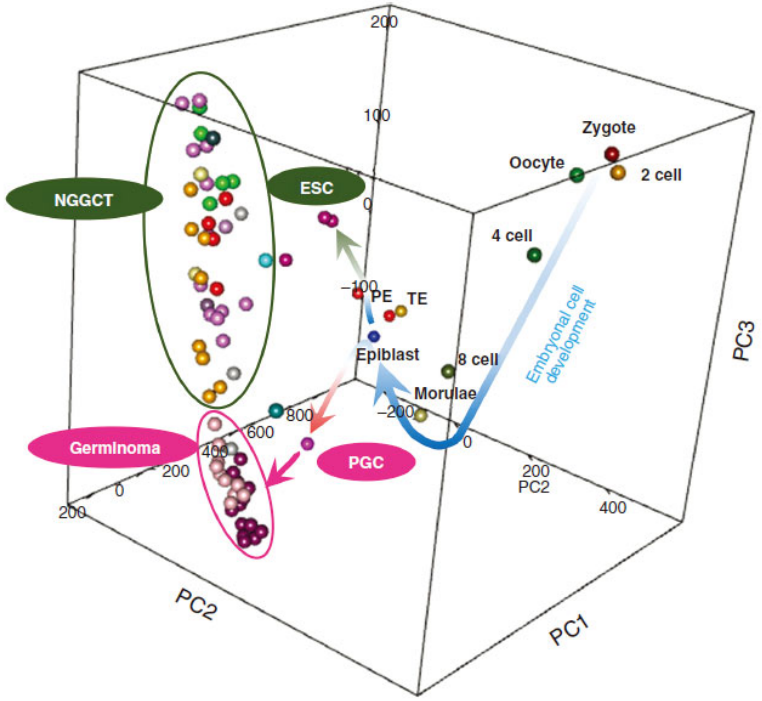
- Clustering analysis by RNA transcriptome
  - RPKM expression data from NGS data
    - RPKM, FPKM, TPM
  - Clustering for log(RPKM+ε)
    - Euclidean distance
    - Ward agglomeration method
- Survival analysis
  - Kaplan-Meier curve
  - Log-rank test
- Wilcox test

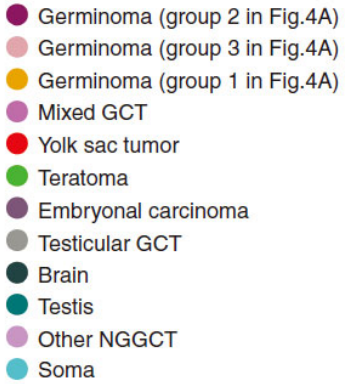
Results

- Cluster 3:
  - DNA alterations of drivers TP53/KRAS/SMAD4 were infrequent
- Cluster 4
  - Worst prognosis
  - Higher immune checkpoint gene expressions
  - Hyper mutation cases

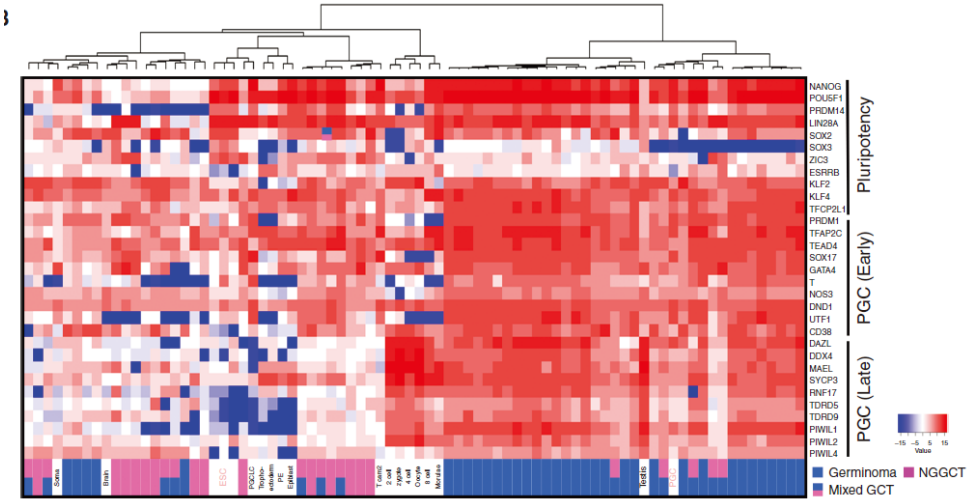
Transcriptome analysis in CNS germ cell tumors

RNA データのPCA（主成分分析）結果を、
右の分類群と組織分類でラベル



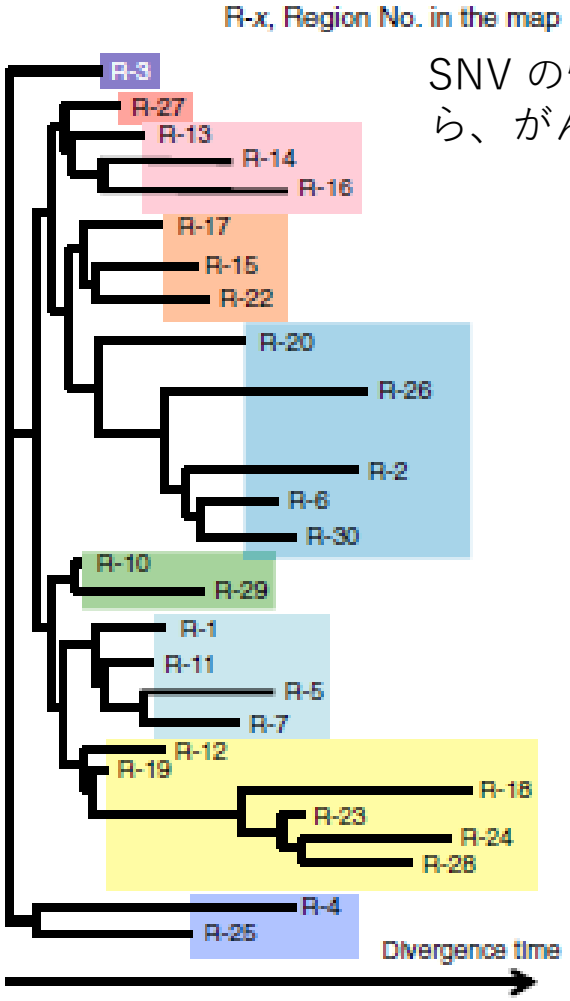


RNA クラスター解析による
germ cell tumor 症例の分類群

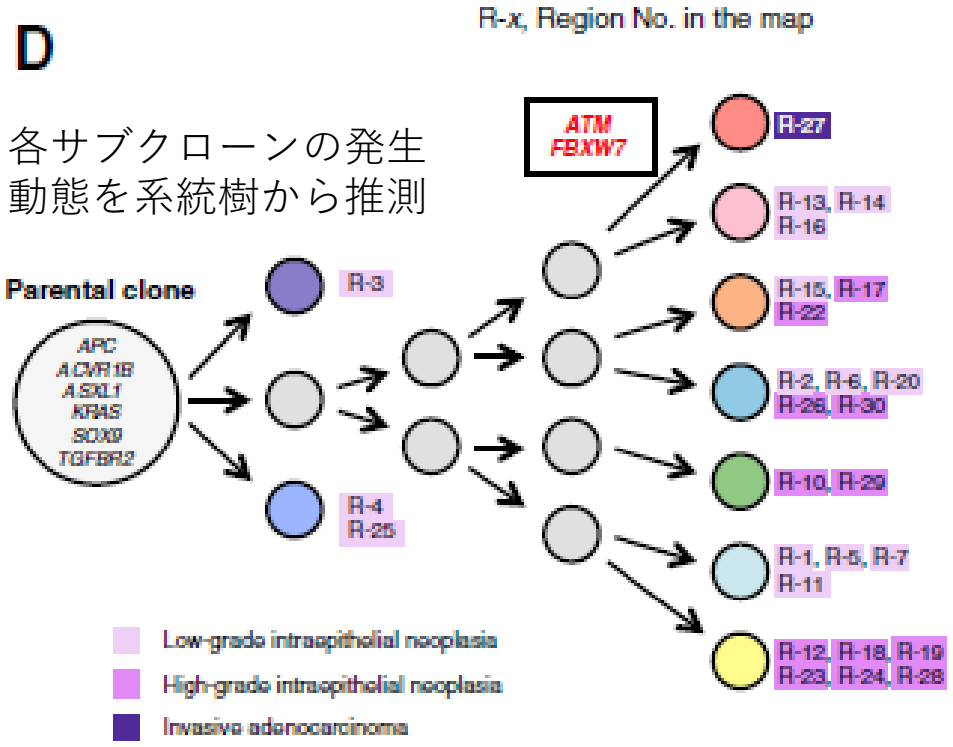


(Takami et al, 2022, *Neuro-Oncology*)

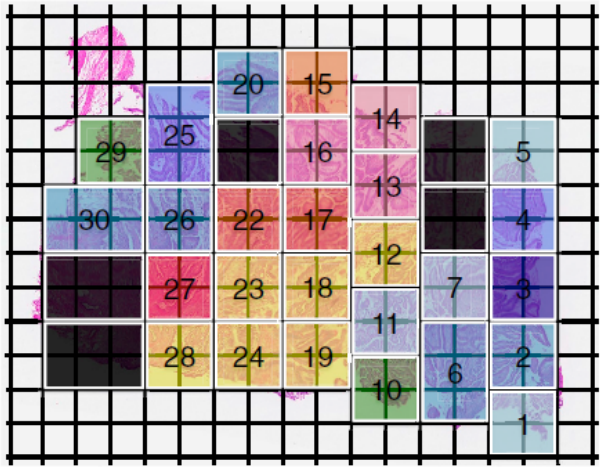
Cancer-cell evolutionary analysis in ampullary carcinoma



SNV の情報に基づいたDNA 配列の類似度から、がん細胞の進化系統樹を推定。



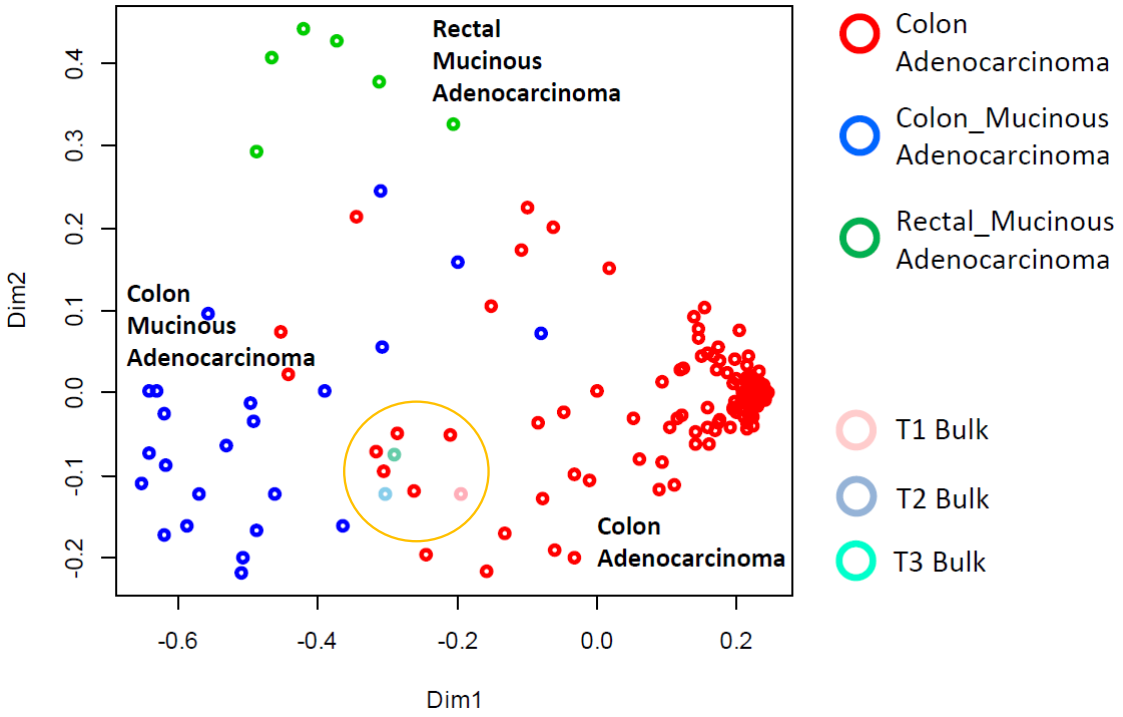
がん細胞サンプリング場所と系統樹はconsistent

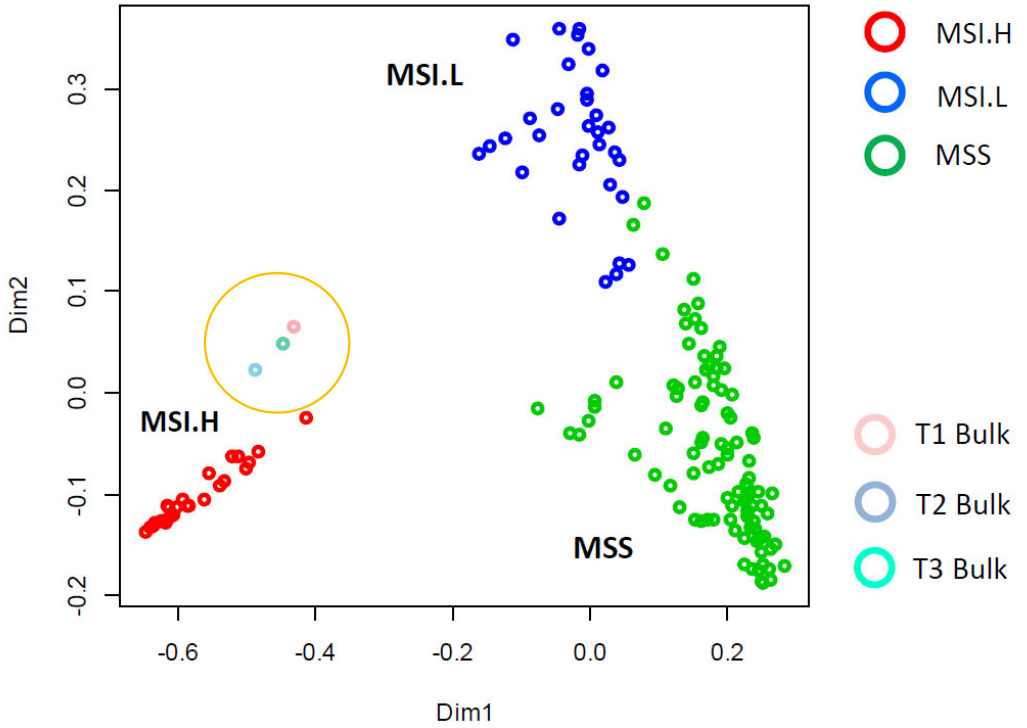


(Yachida et al, 2016, *Cancer Cell*)

Machine learning analysis in colorectal cancer

RNA の発現量から、機械学習によって、ある小腸がんのマウスモデルが、１）ヒトの大腸腺がんに似ていること、２）MSI-H に近いこと、を予測。

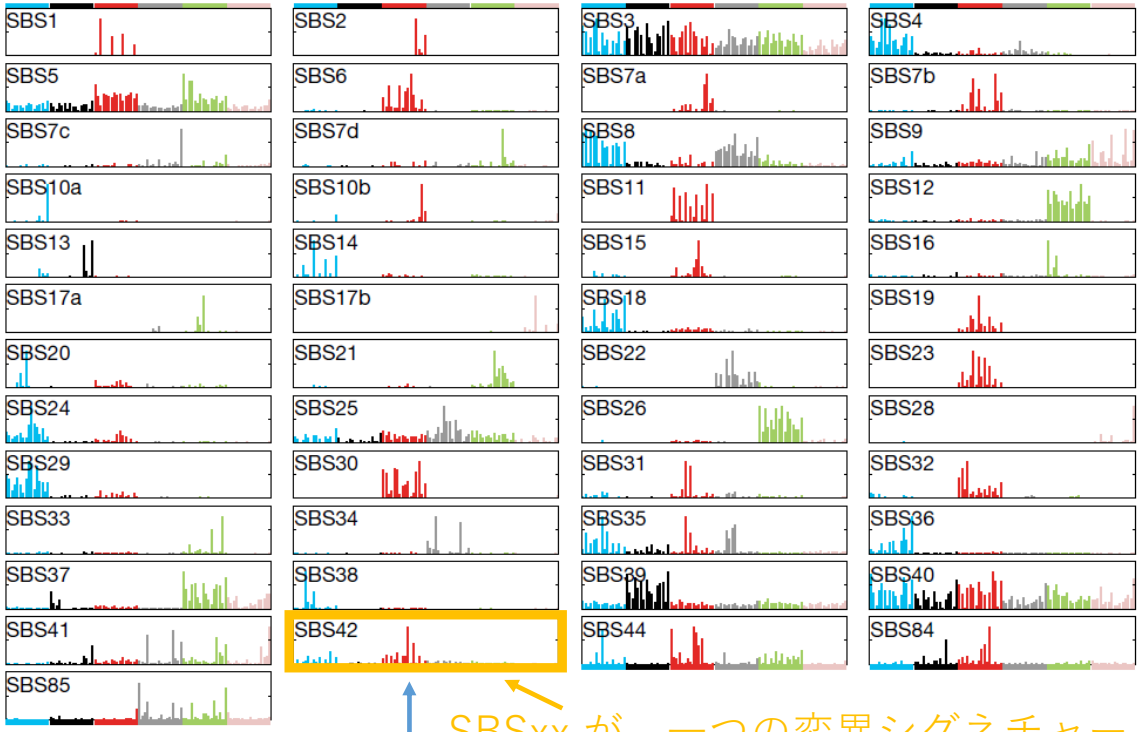




(Ono et al, 2021, *BMC Biology*)

Mutational signature analysis in pan-cancer

変異シグネチャー解析。３連塩基の変異パタンの頻度集計を、NMF という方法で、変異シグネチャーと呼ばれる要素的変異パタン（の頻度）に分解する。

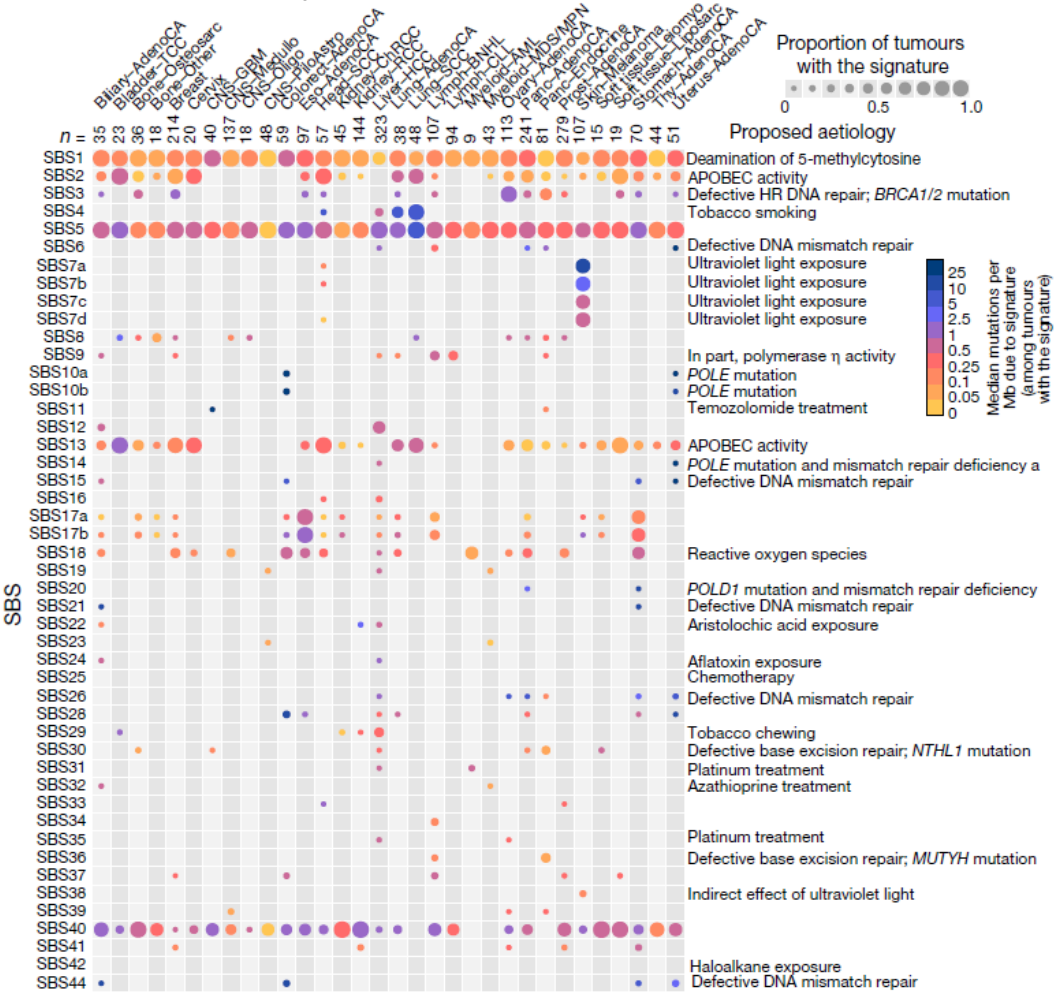


SBSxx が、一つの変異シグネチャー

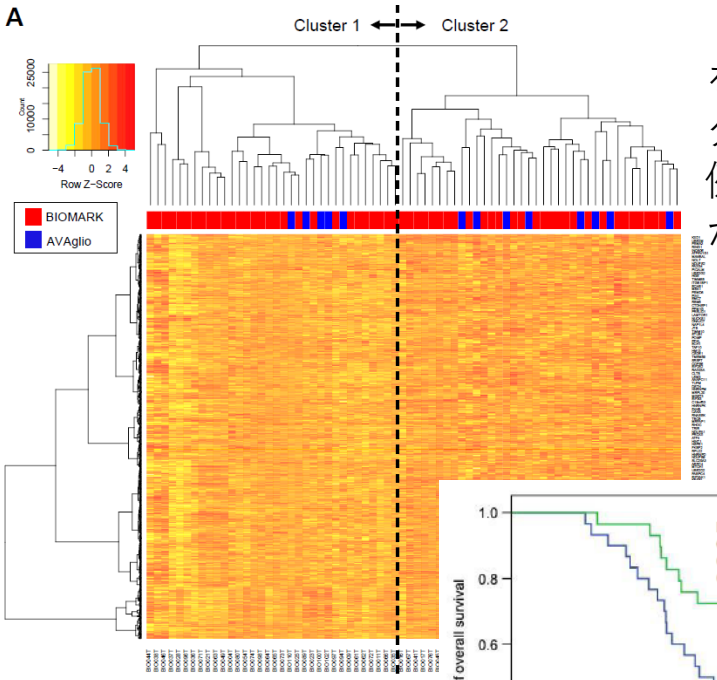
- １縦棒が、３連塩基変異パタンに相当
- 縦軸は、頻度

(Alexandrov et al, 2020, *Nature*)

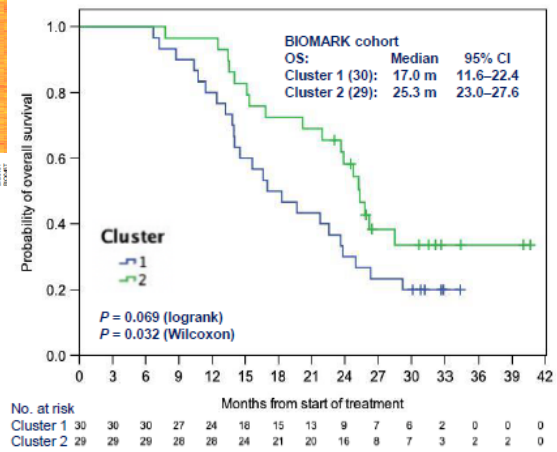
疫学的素因が、変異シグネチャーに対応すると考えられている。



GSEA (gene set enrichment analysis) in glioblastoma



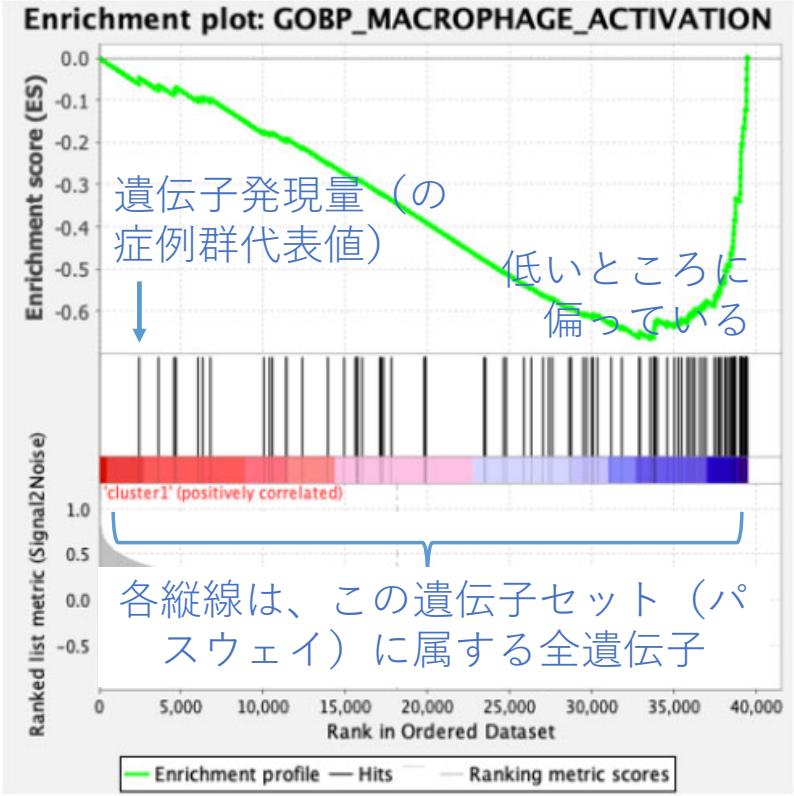
1．RNA 発現量をクラスタリング解析して、症例群が２つに分かれた



２．その２群を生存時間分析したら、OS の違いが確認できた

３．予後が悪い群において、GSEA 解析。

色んな遺伝子セット（パスウェイ）で調べたところ、macrophage activation パスウェイに属する遺伝子発現が低かった。

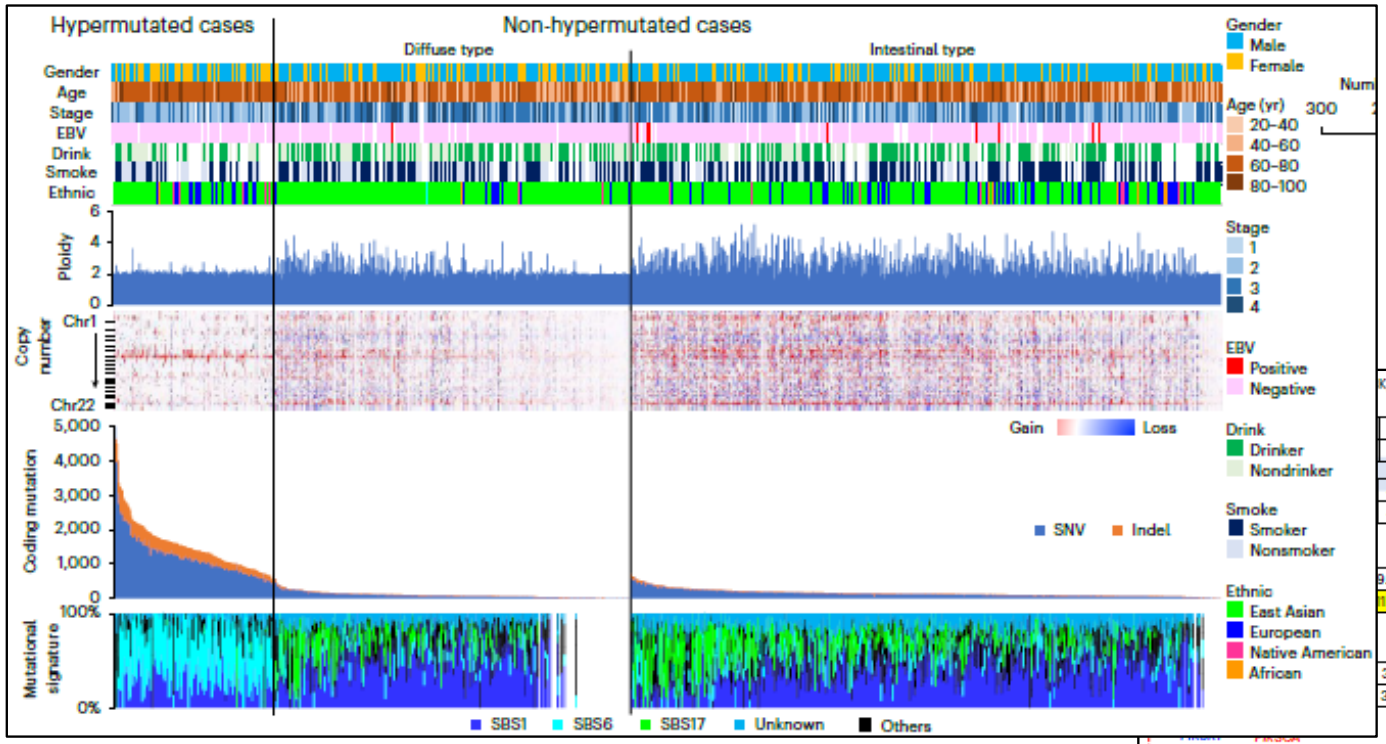


(Nagane et al, 2022, *Cancers*)

Other typical analyses (plots) in gastric cancer

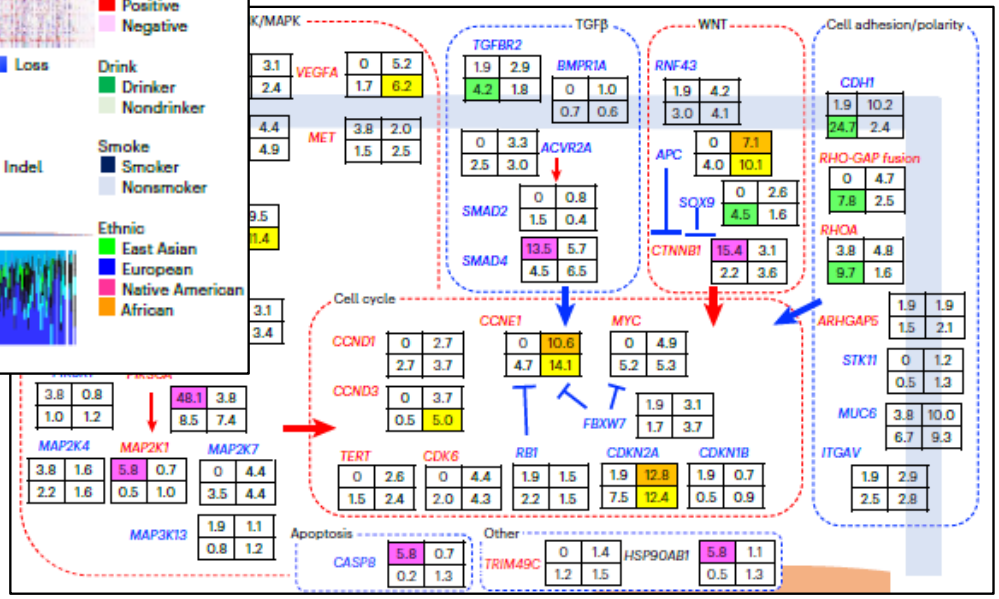
症例

臨床・ゲノムの特徴



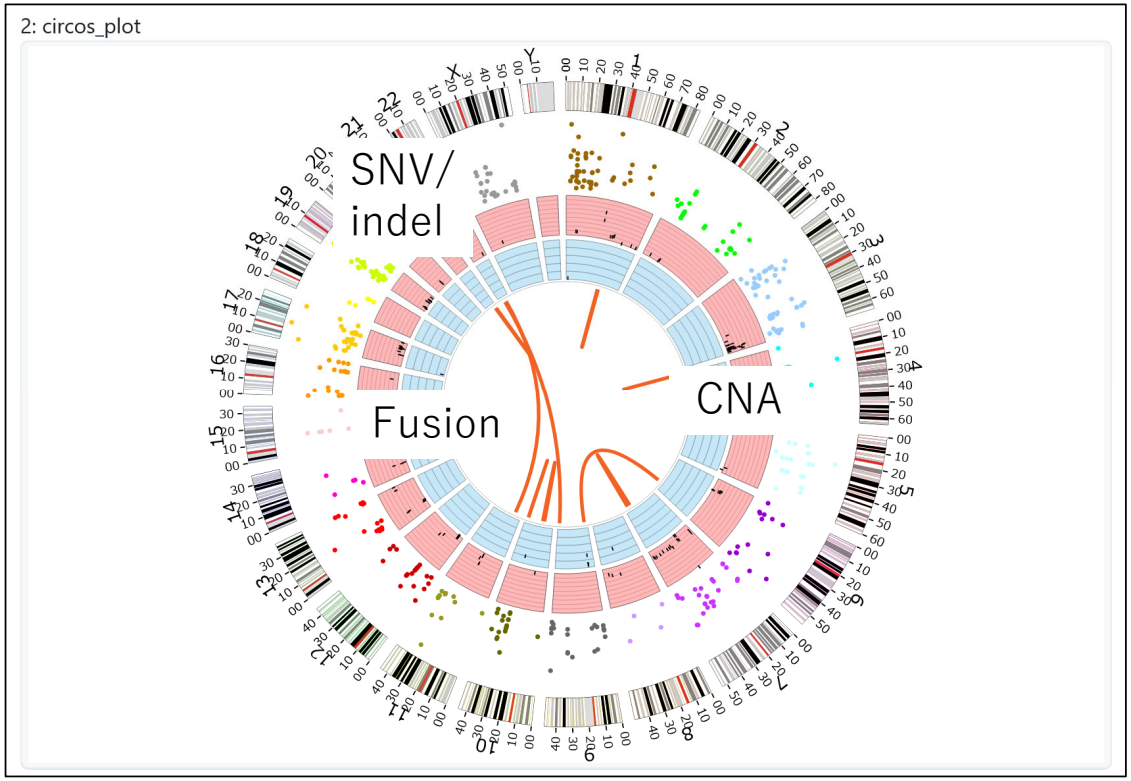
(Totoki et al, 2023, *Nature Genetics*)

遺伝子パスウェイ図
（症例４分類の変異頻度）



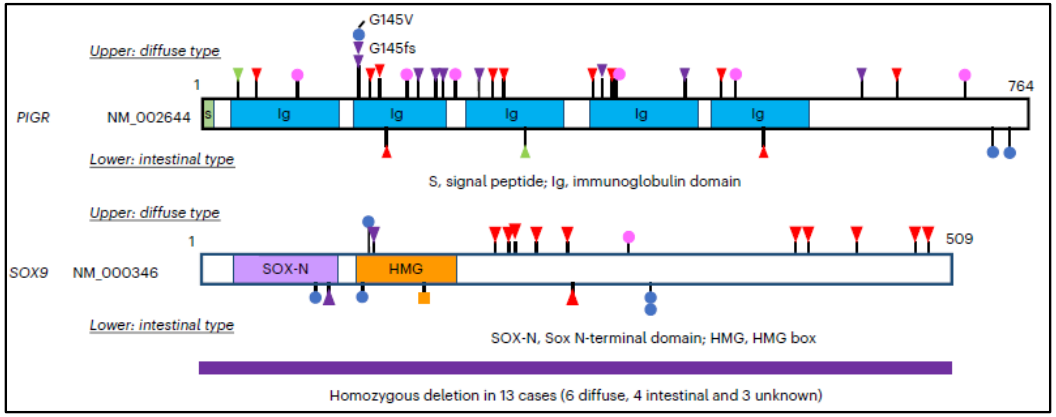
Other typical analyses (plots)

Circos-plot: 全染色体が円上に配置され、
SNV/indel, CNA, fusion がプロットされる



(catstools, https://github.com/ccatg-pub/catstools)

各遺伝子上での変異頻度

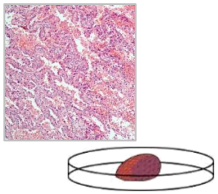


(Totoki et al, 2023, *Nature Genetics*)

PROGRAM DEVELOPMENT
in cancer genome medicine

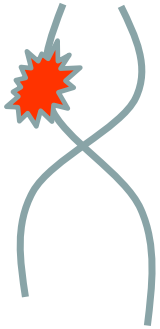
– Examples –

がんゲノム医療（臨床シークエンス）



がん組織

次世代
シークエンサー



EGFR
遺伝子

がん細胞の
ゲノム

遺伝子異常と分子標的薬の表

Table 1 | **Genomic alterations as putative predictive biomarkers for cancer therapy**

Genes	Pathways	Aberration type	Disease examples	Putative or proven drugs
PIK3CA[51,52], *PIK3R1* (REF. 53), *PIK3R2*, *AKT1*, *AKT2* and *AKT3* (REFS 54,55)	Phosphoinositide 3-kinase (PI3K)	Mutation or amplification	Breast, colorectal and endometrial cancer	• PI3K inhibitors • AKT inhibitors
PTEN[56]	PI3K	Deletion	Numerous cancers	• PI3K inhibitors
MTOR[57], *TSC1*[58] and *TSC2* (REF. 59)	mTOR	Mutation	Tuberous sclerosis and Bladder cancer	• mTOR inhibitors
RAS family (*HRAS*, *NRAS*, *KRAS*), *BRAF*[60] and *MEK1*	RAS–MEK	Mutation, rearrangement or amplification	Numerous cancers, including melanoma and prostate cancer	• RAF inhibitors • MEK inhibitors • PI3K inhibitors
Fibroblast growth factor receptor 1 (*FGFR1*), *FGFR2*, *FGFR3*, *FGFR4* (REF. 36)	FGFR	Mutation, amplification or rearrangement	Myeloma, sarcoma and bladder, breast, ovarian, lung, endometrial and myeloid cancers	• FGFR inhibitors • FGFR antibodies
Epidermal growth factor receptor (*EGFR*)	EGFR	Mutation, deletion or amplification	Lung and gastrointestinal cancer	• EGFR inhibitors • EGFR antibodies
ERBB2 (REF. 61)	ERBB2	Amplification or mutation	Breast, bladder, gastric and lung cancer	• ERBB2 inhibitors • ERBB2 antibodies
SMO[62,63] and *PTCH1* (REF. 64)	Hedgehog	Mutation	Basal cell carcinoma	• Hedgehog inhibitor
MET[65]	MET	Amplification or mutation	Bladder, gastric and renal cancer	• MET inhibitors • MET antibodies
JAK1, *JAK2*, *JAK3* (REF. 66), *STAT1*, *STAT3*	JAK–STAT	Mutation or rearrangement	Leukaemia and lymphoma	• JAK–STAT inhibitors • STAT decoys
Discoidin domain-containing receptor 2 (*DDR2*)	RTK	Mutation	Lung cancer	• Some tyrosine kinase inhibitors
Erythropoietin receptor (*EPOR*)	JAK–STAT	Rearrangement	Leukaemia	• JAK–STAT inhibitors
Interleukin-7 receptor (*IL7R*)	JAK–STAT	Mutation	Leukaemia	• JAK–STAT inhibitors
Cyclin-dependent kinases (CDKs;[67] *CDK4*, *CDK6*, *CDK8*), *CDKN2A* and cyclin D1 (*CCND1*)	CDK	Amplification, mutation, deletion or rearrangement	Sarcoma, colorectal cancer, melanoma and lymphoma	• CDK inhibitors
ABL1	ABL	Rearrangement	Leukaemia	• ABL inhibitors
Retinoic acid receptor-α (*RARA*)	RARα	Rearrangement	Leukaemia	• All-trans retinoic acid
Aurora kinase A (*AURKA*)[68]	Aurora kinases	Amplification	Prostate cancer and breast cancer	• Aurora kinase inhibitors
Androgen receptor (*AR*)[69]	Androgen	Mutation, amplification or splice variant	Prostate cancer	• Androgen synthesis inhibitors • Androgen receptor inhibitors
FLT3[70]	FLT3	Mutation or deletion	Leukaemia	• FLT3 inhibitors
MET	MET–HGF	Mutation or amplification	Lung cancer and gastric cancer	• MET inhibitors
Myeloproliferative leukaemia (*MPL*)	THPO, JAK–STAT	Mutation	Myeloproliferative neoplasms	• JAK–STAT inhibitors
MDM2 (REF. 71)	MDM2	Amplification	Sarcoma and adrenal carcinoma	• MDM2 antagonist
KIT[72]	KIT	Mutation	GIST, mastocytosis, leukaemia	• KIT inhibitors
PDGFRA and *PDGFRB*	PDGFR	Deletion, rearrangement or amplification	Haematological cancer, GIST, sarcoma and brain cancer	• PDGFR inhibitors
Anaplastic lymphoma kinase (*ALK*)[9,17,73,74]	ALK	Rearrangement or mutation	Lung cancer and neuroblastoma	• ALK inhibitors
RET	RET	Rearrangement or mutation	Lung cancer and thyroid cancer	• RET inhibitors
ROS1 (REF. 75)	ROS1	Rearrangement	Lung cancer and cholangiocarcinoma	• ROS1 inhibitors
NOTCH1 and *NOTCH2*	Notch	Rearrangement or mutation	Leukaemia and breast cancer	• Notch signalling pathway inhibitors

次世代シークエンサーからのデータ

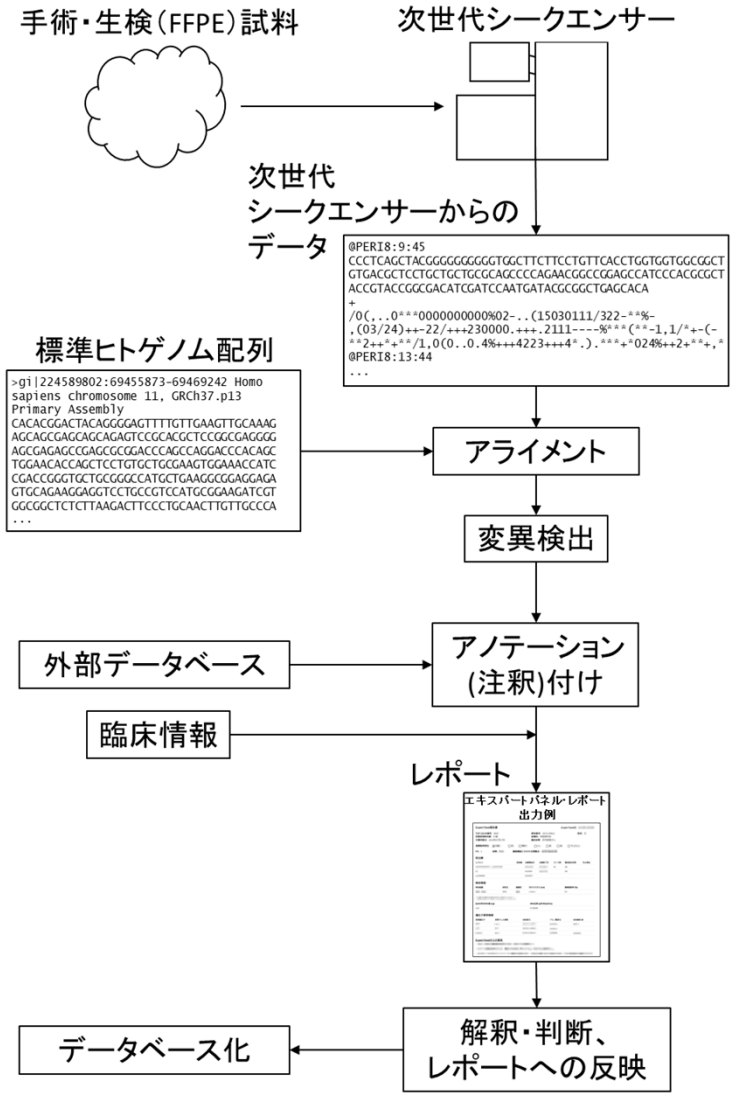
@PERI8:9:45
CCCTCAGCTACGGGGGGGGGGTGGCTTCTTCCTGTTCACCTGGTGGTGGCGGCTGTGACGCTCCTGCT
GCTGCGCAGCCCCAGAACGGCCGGAGCCATCCCACGCGCTACCGTACCGGCGACATCGATCCAATGATA
CGCGGCTGAGCACA
+
/0(,..0***0000000000%02-..(15030111/322-**%-,(03/24)++-22/+++230000.+++.2111----%***(**-1,1/*+-(-
2++*+/1,0(0..0.4%+++4223+++4*.).***+*024%++2+**+,*
@PERI8:13:44
GGAGGACCGTCGCTTGGTGCACCGCGACCTGGCAGCCAGGAACGGTACTGGTGAAAACACCGCAGCAT
GTCAAGATCACAGATTTGGGCTGGCCAAACTCGCGTTGGGTAGCGAAGAGACCGAAGACGTCGCGCCAG
TCG
+
0,.:3683:0+-..+3(+-54707;<89(..69744122+1.44/6:;9;2=:<==
:661:6967+++577%++2++*./.02,45,444/4,13)/413..0)/1)***,*
@PERI8:15:42
GGCTCATCAAGCTCGCTGCTTCCAGGAATGACTGGGAAGGTGGGAAGGAGAGAAGATGCGTGGGTTCTT
CA
+
;(1.9...605344911.452;4<=<A<BB@??=@3;/9/,,0&0&,0654
@PERI8:15:43
TCCCCCCCCCCCCAAATGTTCACTAACTCTGAAACGTGG
ACGCATGGTCTGTA
+
-0221111222(16.445,65569355766><079.00&,440++37,,+,*
...

- Big data with errors

- FastQ in 5GB per case
- We use a cluster machine with 4 X
  - 20-core 2.4 GHz CPUs
  - 128 GB memory
  - 150 TB storage

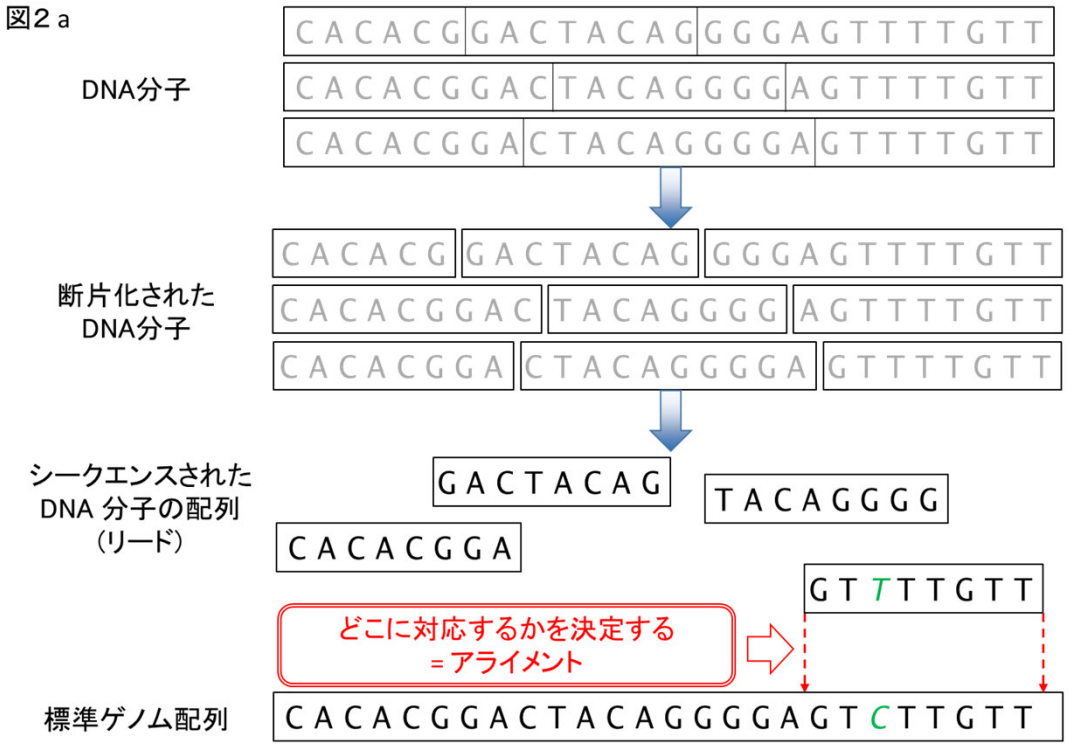
臨床シークエンスにおける バイオインフォマティクス処理の流れ

図1



（加藤、「最新がん個別化医療」、*癌と化学療法*、2016）

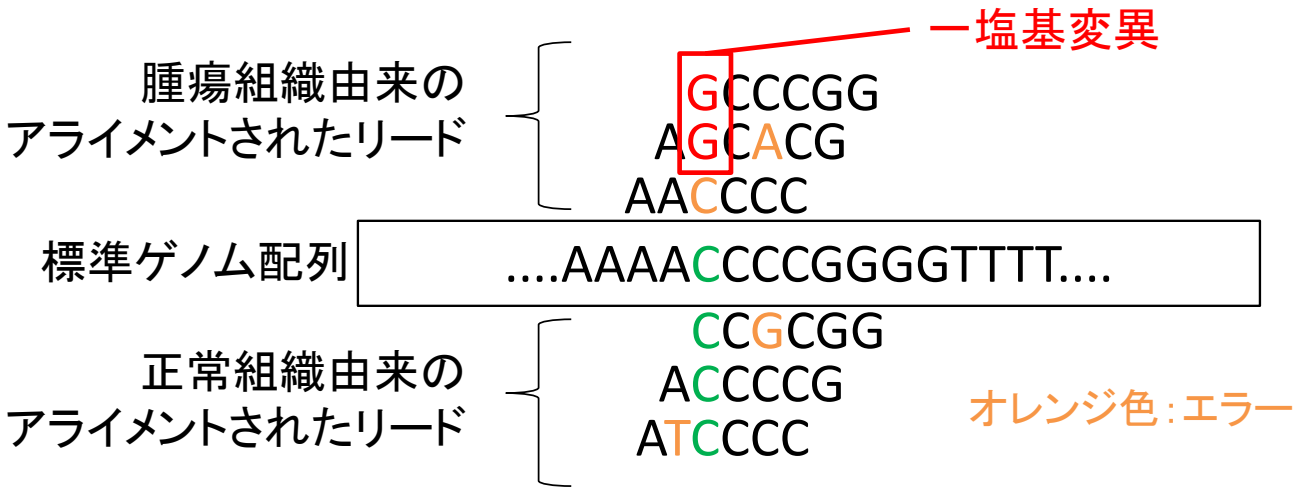
アライメント



アライメントの原理。a. 腫瘍組織から抽出されたDNA 分子は（一般に、ランダムな位置で）断片化され、実験に適した長さに基づいて選択された断片化DNA 分子がシークエンスされる。断片化され位置情報を失った DNA 分子がシークエンスされているので、このままではリードのゲノム上の位置は分からない。そのため、標準ゲノム配列とリード配列のATGC文字の並びを比較して、最も一致する位置を決めるのが、アライメントである。アライメントの最も単純な方法は、最初に、標準ゲノム配列の左端とリードの左端を並べ何文字中何文字一致しているかを見、一致が悪ければ、さらにリードを一文字分ずらして同様に一致度を見、また一致が悪ければ…、を繰り返していく。この方法では効率が悪いので、通常は高速なアルゴリズムを使う。

（加藤、「最新がん個別化医療」、
癌と化学療法、2016）

検出法の原理：点変異 – cisCall –



cisCall では、分割表による統計分析を行っている

	G	Not G
腫瘍由来	2	1
正常組織由来	0	3

仮に数が100倍程度になったとして…

	G	Not G
腫瘍由来	200	100
正常組織由来	10	290

分析数値の例

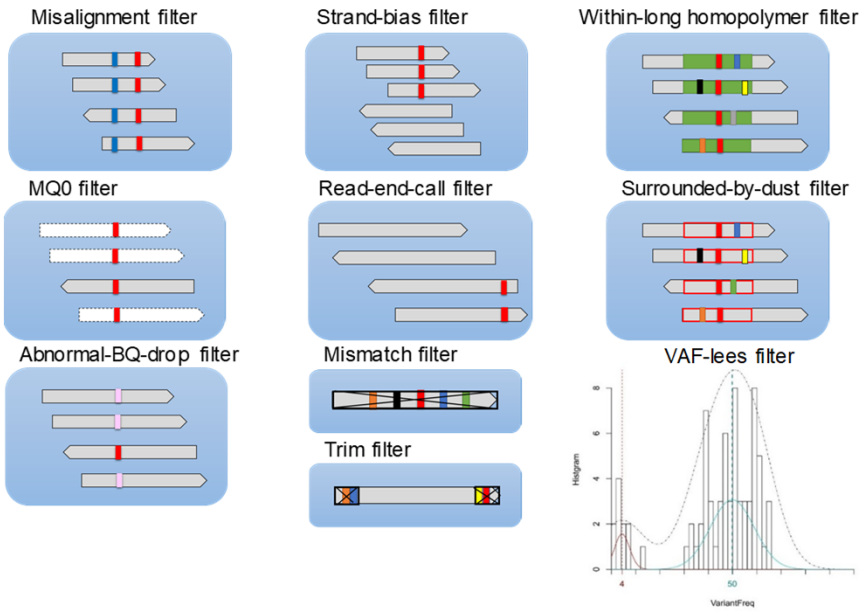
- Fisher's exact test *p* value: 6.8×10^{-68}
- Odds ratio: 57.5（実際には、Prop. Ratio: {200/(200+100)}/{10/(10+290)} = 20

cisMuton Ver5: SNV/indel calling

- Prep filters: mapping-quality and base-quality filters
- Variant extraction: Fisher's exact test
  - *P*-value, proportion ratio (~odds ratio), FC, count

Single-point analysis

- Noise filter set 1
  - Misalignment filter
  - Strand-bias filter
  - Within-long homopolymer filter
  - MQ0 filter
  - Read-end-call filter
  - Surrounded-by-dust filter
  - Abnormal-BQ-drop filter (for Ion indels only)
- Noise filter set 2
  - Second Fisher filter
    - Erroneous-read filter and trimming filter
- Noise filter set 3
  - VAF-lees filter

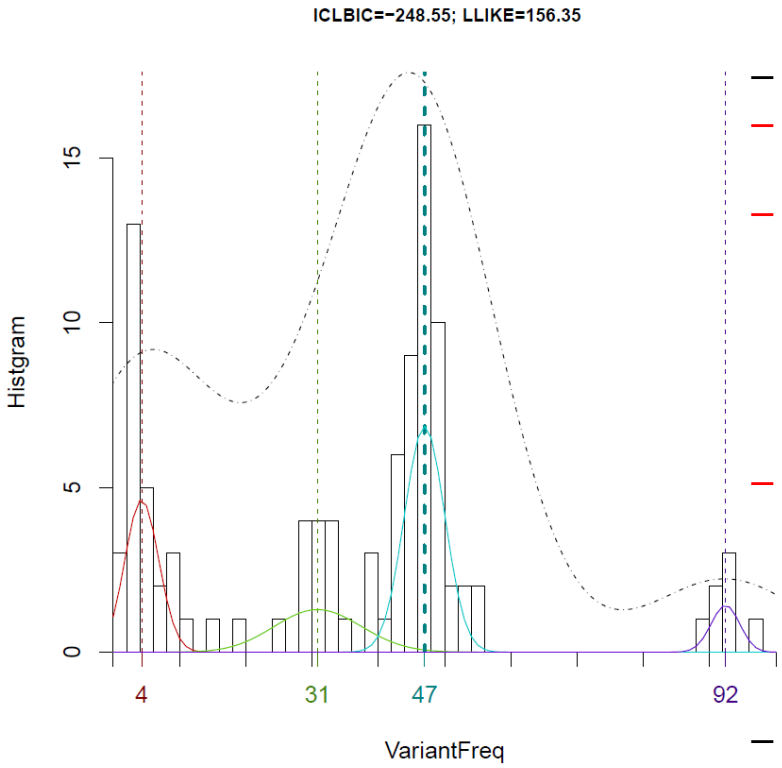


(Kato et al, *Genome Medicine*, 2018)

- Main features

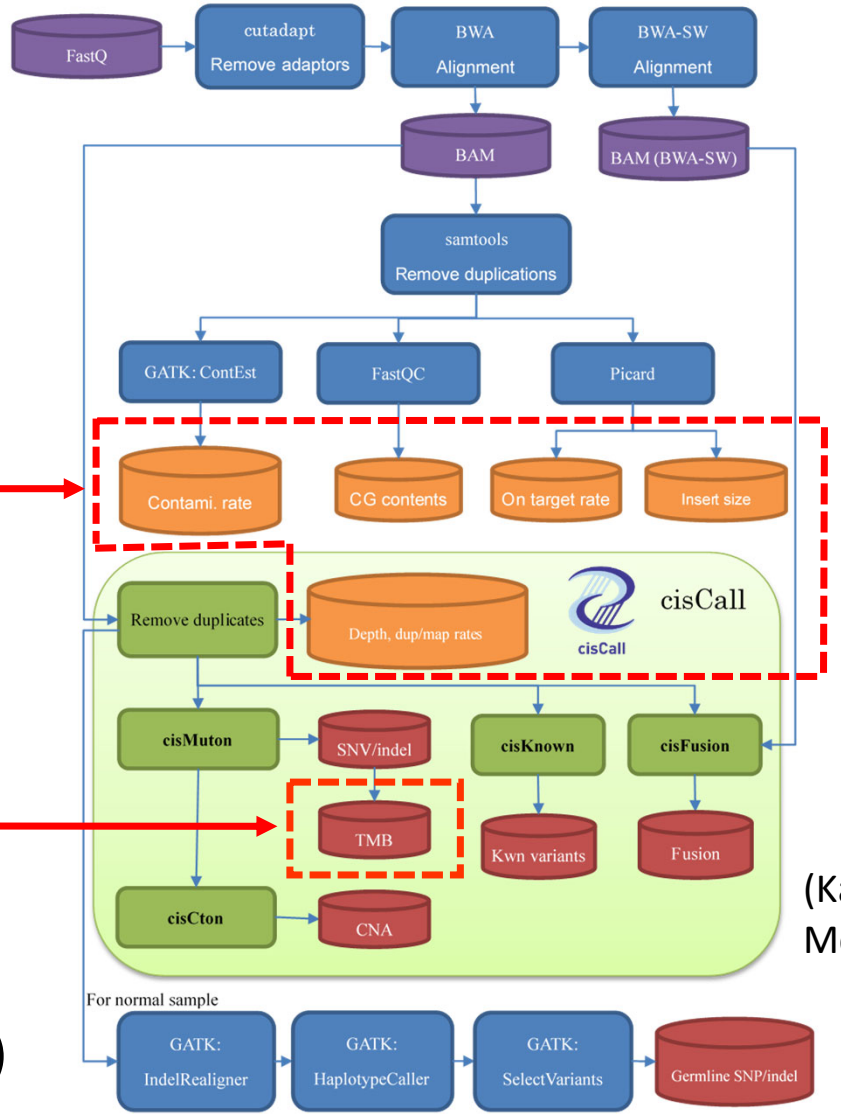
1. Elaborated multi-layer filters
2. Takes every chance to use non-parametric techniques for abrupt FFPE errors
3. Takes every chance to calculate internal control values (e.g., error rates) from observed data for parameter values for flexibility to various clinical settings

フィルターの例



- cisCall のVAF-lees filter
  - 低VAF 値を取る点変異はエラーが多い、という経験則がある
    - 直感的には、signal-to-noise において、signal の強度がnoise の強度に近くなっている状態
  - Signal をnoise (エラー)と分離したい。
  - VAF の分布において、これを**混合ベータ分布**と見て、データから要素ベータ分布に分解する。
  - 平均値が低い要素ベータ分布を、エラー分布と解釈する。
    それ以外の要素ベータ分布を、真の点変異のVAF分布と解釈する。
    - 真の点変異に対する要素ベータ分布の複数性：生物学的には、コピー数変化や腫瘍率の影響によって、点変異VAF 値は複数存在することがある
  - **EM アルゴリズム**によって、分解する。
    - 各データポイントが要素ベータ分布に所属する確率を定義し、その確率的重みを加味したカウントで、ベータ分布の決定パラメータを計算し、決定されたベータ分布（確率密度）を用いてデータポイントの所属確率を再計算して…、と繰り返していくと、真の要素分布が推定できる
  - EM では要素数を与える必要があり、それら複数のケースを評価する必要があるが、その評価は情報量規準で行う。
    - 混合ベータ分布分解用の、**ICM-Bayesian Information Criterion** を使用
      - Akaike's/Bayesian Information Criterion

プログラム・パイプラインの構築



QCs

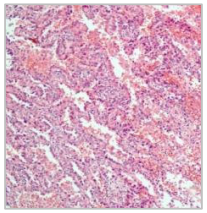
TMB

cisCall+
Pipeline (Ver5)

(Kato, Precision Medicine, NTS, 2018)

がんゲノム医療における変異検出と、ゲノムデータ標準化

Daily-use FFPE tumor samples



Next-generation sequencer (NGS)

```
@PERI8:15:42
GGCTCATCAAGCTCGCTGCTTCCAGGAATGACTGGG
AAGGTGGGAAGGAGAGAAGATGCGTGGGTTCTTCA
+
;(1.9...605344911.452;4<=<A<BB@??=@3;/9/,,0&
0&,065455)+.4831**-%-(*2(-*
@PERI8:15:43
...
```

cisCall

cisCall

Variant call

- Point mutations
- Fusion genes
- Copy number alterations
- Complex known alterations

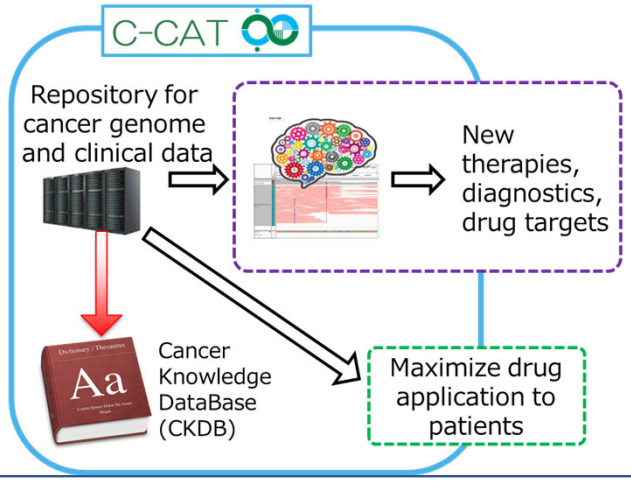
Development of ***cisCall*** (Kato et al, 2018, *Genome Medicine*) for variant calling in NCC Oncopanel

First Medical Device approved by the Japanese government in cancer genome medicine

Development of

- ***cats format***
- ***catstools***

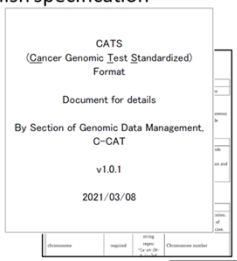
for standardized genomic data format in C-CAT (Kohno et al, 2022, *Cancer Discov*)



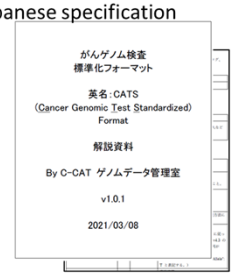
cats format

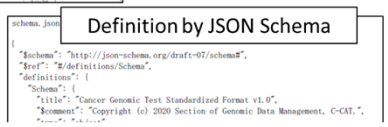
https://www.ncc.go.jp/en/c_cat/section/070/index.html

English specification



Japanese specification

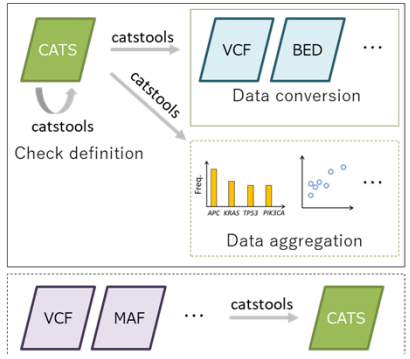




catstools

Tool to manipulate data in CATS format

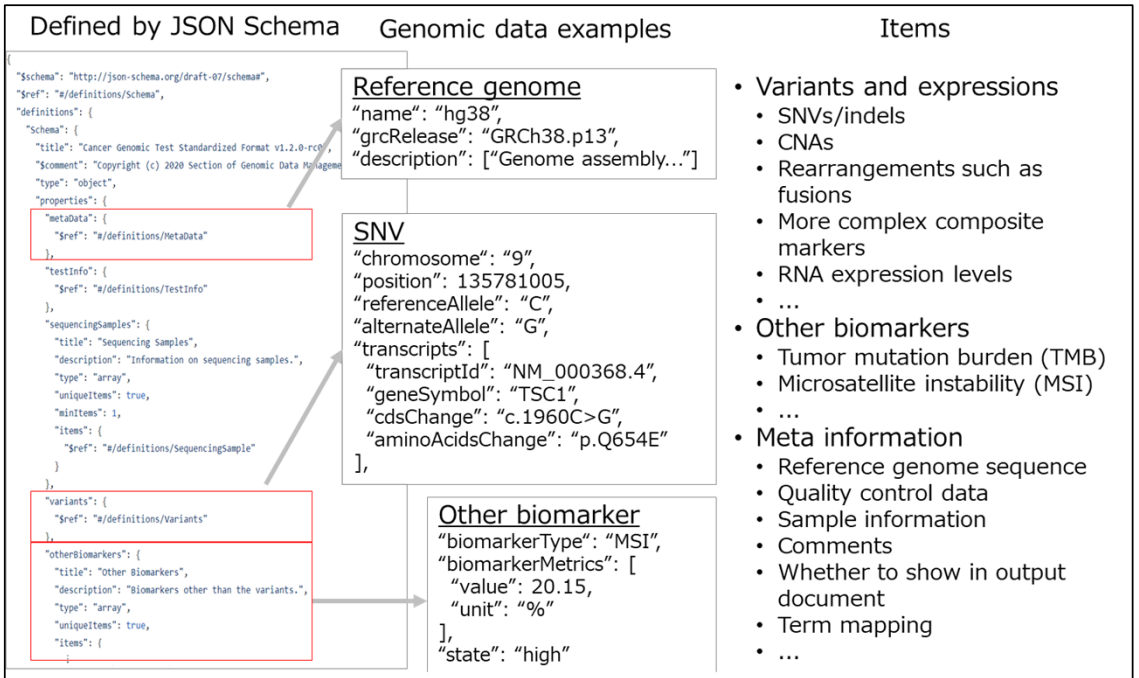
For example, data conversion and aggr



がんゲノム医療におけるゲノムデータの標準化

CATS (CAncer genomic Test Standardized) format

Defined by JSON Schema　Genomic data examples　Items

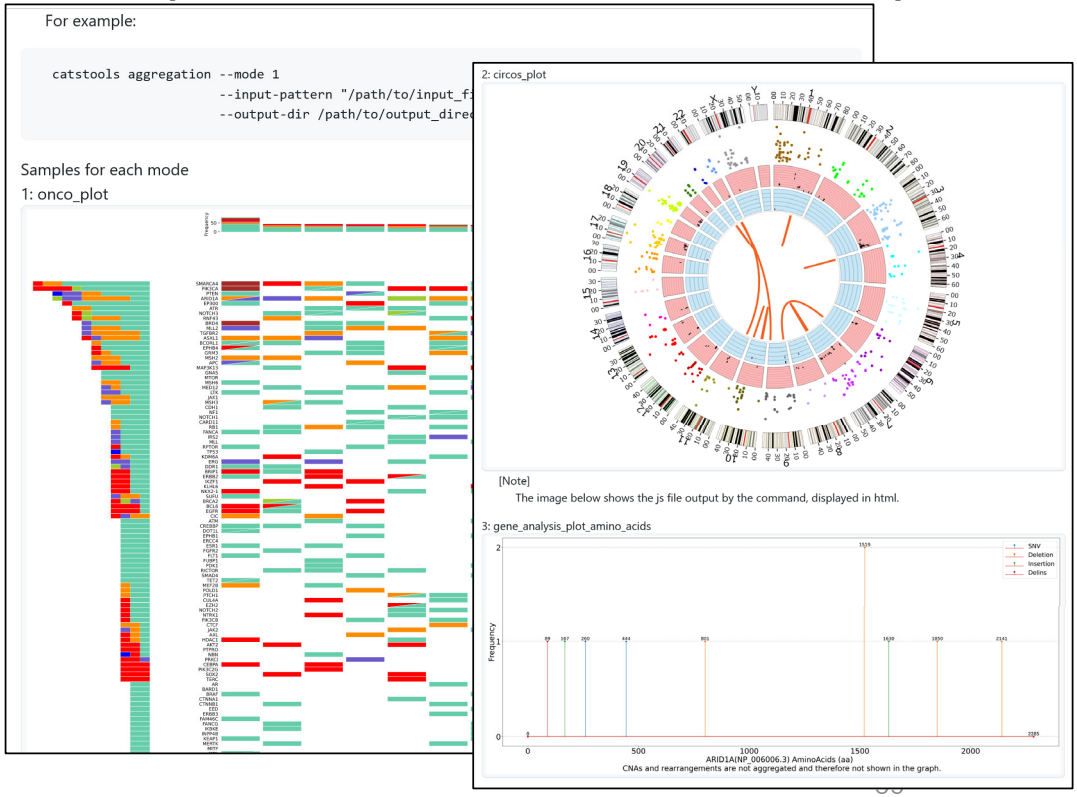


- **Excluding clinical data** in electrical medical record (EMR) in hospitals
  - Because such data are unlikely to be accessible to testing companies or laboratories

(Kohno, Kato, et al, 2022, *Cancer Discov.*)

- **がんゲノム医療を意識した設計**
  - CNA やSV もまとめて
  - 染色体レベルの異常
  - がんゲノム医療に必須のTMB/MSI も
  - ほか、がんゲノム医療に必要な情報

catstools (CATS|TOOLS, not CAT|STOOLS)



- **がんゲノム医療上の操作**
  - 電子カルテ・データ（HL7 FHIR）への接続
  - 簡易版ゲノム医療レポートの作成
  - がんゲノム学データフォーマットとの接続

変異検出プログラムの**全ゲノム・FFPE**シークエンスへの拡張

- FFPE でのがん全ゲノム検査を目指す

- 自動作成されたレポート

シークエンシングレポート

検査会社受付ID:

■サンプル情報

サンプル　検体識別番号:(NM604.T.FF, NM604.N.Fr)

■遺伝子セット情報

遺伝子セット　Gene1147

■体細胞変異一覧

■一塩基変異 / 挿入・欠失情報

No		遺伝子名	腫瘍 アレル頻度[%]	非腫瘍 アレル頻度[%]	変異種類	CDS変化	アミノ酸変化	公共DB情報 COSMIC(登録数)
1	●	BRCA2	17.0(18/106)	0.0(0/57)	frameshift deletion	exon11:c.2196_2202 delGGTCTTG	V733fs*37	-(-)
2	●	CDKN2A	23.2(43/185)	0.0(0/81)	nonsynonymous SNV	exon2:c.242C>T	P81L	13224(24)
3	●	FLNA	9.4(21/223)	0.0(0/64)	nonsynonymous SNV	exon10:c.1486C>T	R496W	1207225(1)
4	●	N4BP2	16.0(20/125)	0.0(0/61)	frameshift deletion	exon4:c.477delC	Q160fs*21	-(-)
5	●	TP53	26.9(54/201)	0.0(0/63)	nonsynonymous SNV	exon5:c.413C>T	A138V	43818(43)
6	-	CARD11	25.0(43/172)	0.0(0/66)	nonsynonymous SNV	exon25:c.3326A>G	E1109G	-(-)
7	-	CD79A	6.0(12/201)	0.0(0/60)	nonsynonymous SNV	exon2:c.279C>G	I93M	-(-)
8	-	CLIP1	6.8(17/251)	0.0(0/59)	splicing	exon3:c.645_657+1 delAGACAGAGTATTGG	.	-(-)
9	-	CSF1R	17.2(35/203)	0.0(0/57)	nonsynonymous SNV	exon14:c.1975G>C	V659L	-(-)
10	-	FANCM	20.9(24/115)	0.0(0/73)	nonsynonymous SNV	exon23:c.6053A>C	K2018T	-(-)
11	-	FHDC1	24.2(45/186)	0.0(0/64)	nonsynonymous SNV	exon12:c.2851G>A	G951S	-(-)
12	-	GNAS	14.4(43/298)	0.0(0/71)	nonsynonymous SNV	exon1:c.1619C>T	P540L	5507676(1)
13	-	KDM5C	23.8(49/206)	0.0(0/59)	nonsynonymous SNV	exon13:c.1839G>C	E613D	-(-)
14	-	NUTM1	21.6(50/232)	0.0(0/56)	nonsynonymous SNV	exon4:c.927G>T	E309D	-(-)
15	-	TSHR	7.8(18/232)	0.0(0/62)	nonsynonymous SNV	exon10:c.899T>G	M300R	-(-)

■遺伝子コピー数異常情報

No	遺伝子名	変異種類	遺伝子コピー数比
-	-	-	-

■遺伝子再構成情報

No	遺伝子名	変異種類	物理位置(染色体:塩基番号)
16	-	APC	long deletion	5:112,826,483-5:112,828,131
17	-	BAIAP2L1	inversion	7:98,364,156-7:98,359,335
18	-	BLM	long deletion	15:90,721,568-15:90,721,684
19	-	CARD11	long deletion	7:2,917,676-7:2,917,966
20	-	CLIP1	tandem duplication	12:122,375,953-12:122,377,177
21	-	CLTC	long deletion	17:59,650,913-17:59,651,433

Summary

- Bioinformatics（生物情報学）は、分野が広く今も拡大しており、境界が曖昧な部分もある。
- 大ざっぱには、生物学に関連した大量の分子データに、コンピュータを適用して研究する分野である。
  - 例えばがんゲノム学では、大量のDNA, RNA データをコンピュータで処理し、新しいがん関連遺伝子の発見や症例群の特徴付けなどを行う。
- 典型的には、データ解析とプログラム開発がある。
- 本講義では、がんゲノム学とがんゲノム医療を主体に紹介した。
  - Bioinformatics は分野が広く、ここでは触れられなかった解析や開発も多々あります。
  - がんゲノム学やがんゲノム医療に絞っても、多々あります。
- がんゲノム学でのデータ解析例の紹介
- がんゲノム医療でのプログラム開発例の紹介

バイオインフォマティクス・
生物統計合同コンサルテーション

◆案内ページ.　所内にもポスターを掲示しています
http://int.res.ncc.go.jp/web/kenkyu_consul/bioinformatics.html

※ 高次元・多層のオミクスデータから分子学的特徴の抽出や意味付けを行うための専門知識と、バイアスや効率を考慮し分子学的特徴と臨床アウトカムとの関連を適切に推論するための方法論を提供します

◆『生物情報学・生物統計学を含む領域横断的なオミックス臨床研究
の 計画作成・実施体制構築準備の手引き』
http://int.res.ncc.go.jp/web/kenkyu_consul/files/tebiki.pdf
(上記の案内ページにリンクがあります)

※生物情報学、生物統計学の視点から、それぞれの役割と研究の各時点で行う典型的な作業・検討事項をまとめています

END