バイオインフォマティクスとは？
What is bioinformatics?

＊日本語で行います

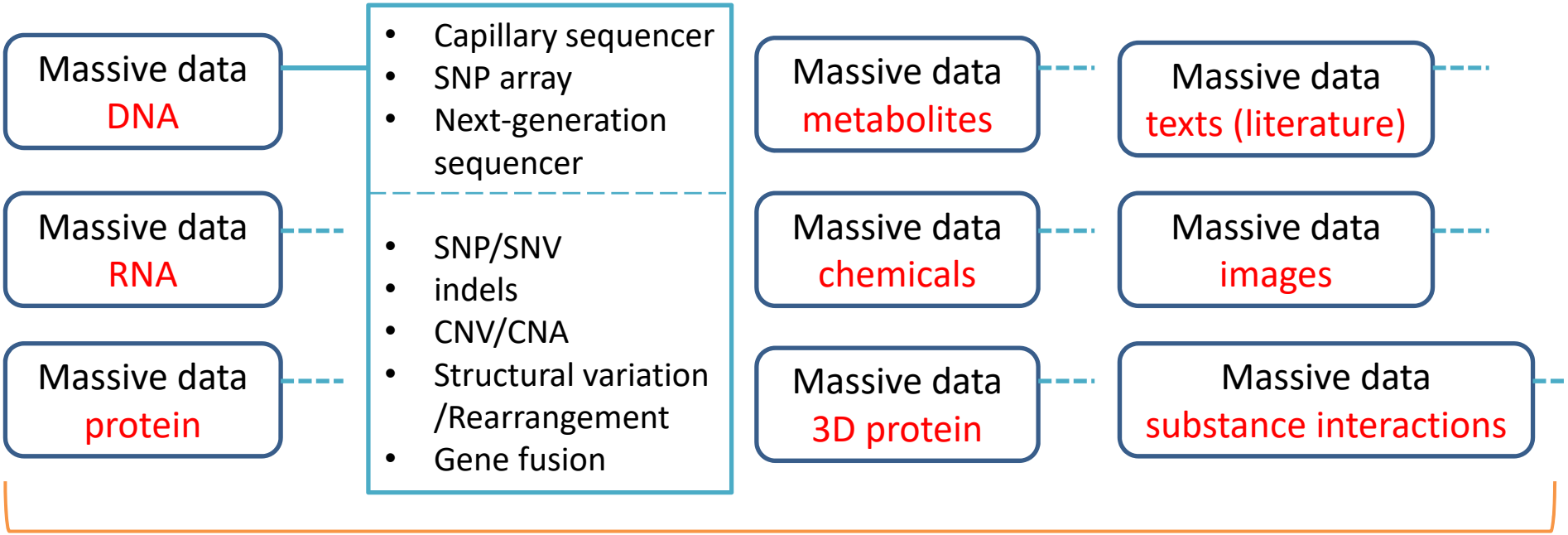
研究所
バイオインフォマティクス部門
加藤 護

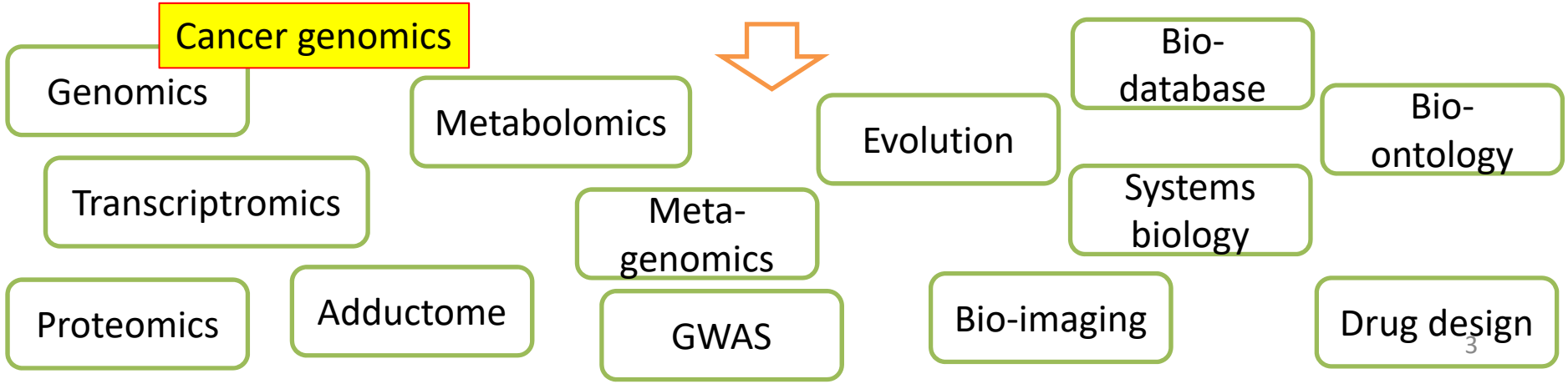
Outline

1. General concepts of bioinformatics
   – Within my limited experiences though...

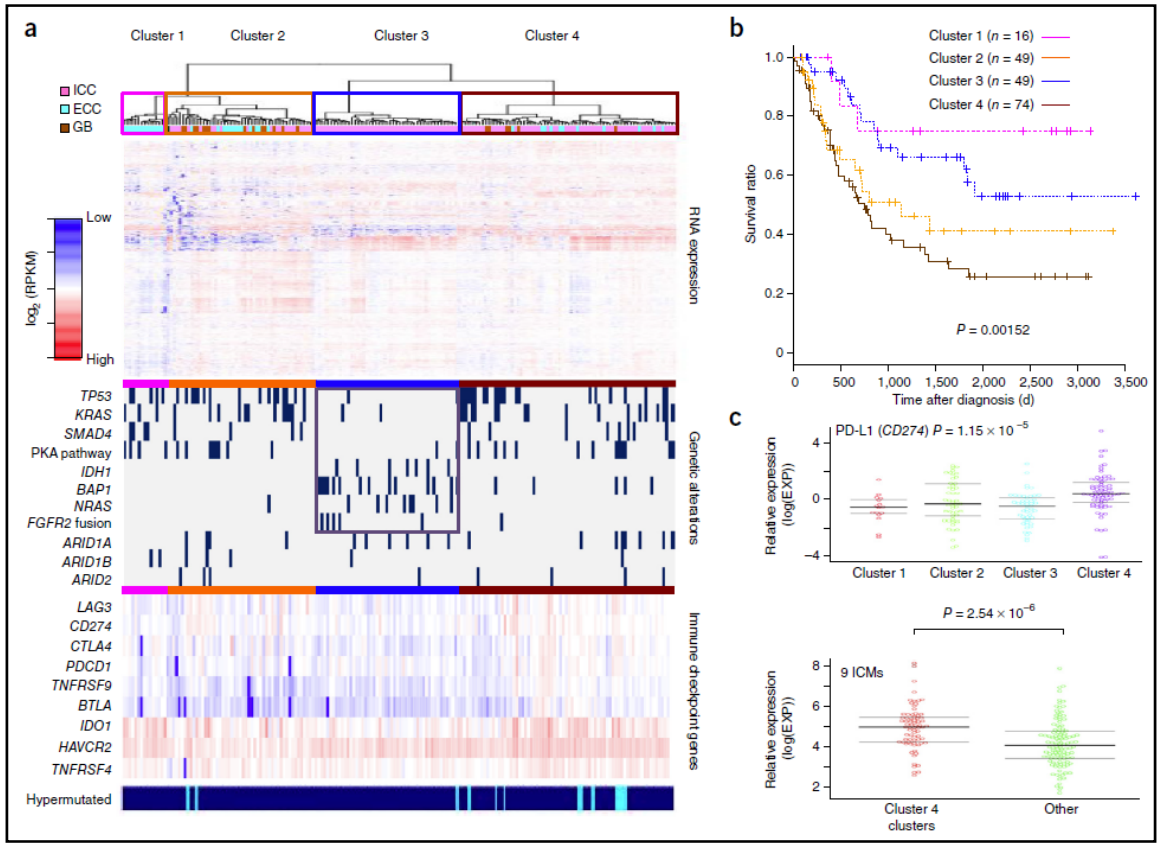
2. Its application to genomic medicine
   – clinical sequencing

Bioinformatics: greedy and cloudy discipline





RNA analysis and the clinical relationships in bile-duct cancer genomics



(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)

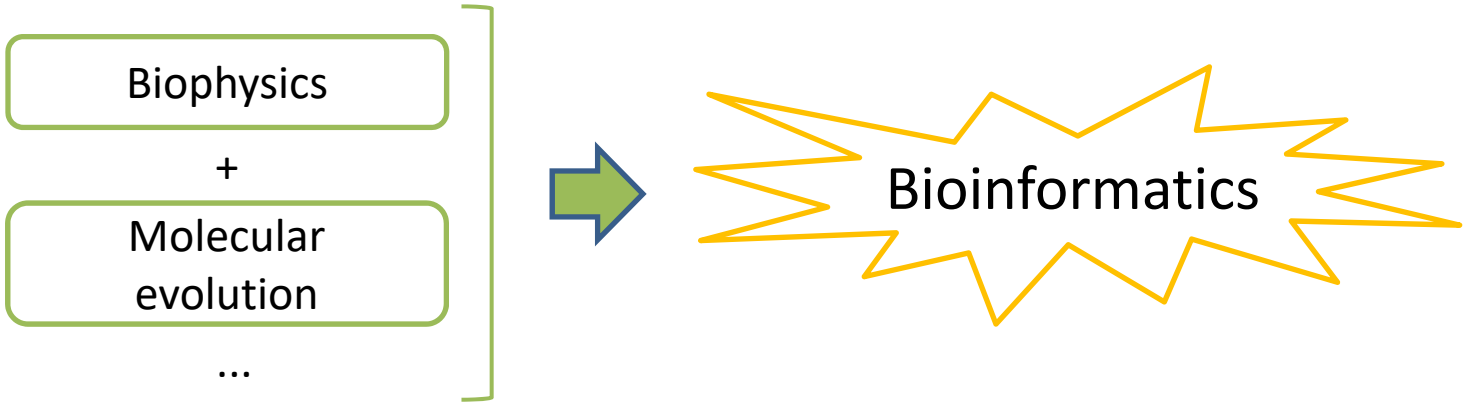
40,000 rows (transcripts)

Gene	BD003T	BD004T	BD005T	BD006T	BD007T
ENST0000	31.35851	81.2562	58.13853	35.76353	40.48326
ENST0000	10.01731	1.137802	32.82091	15.14492	2.095884
ENST0000	0	0	0	0	0
ENST0000	3.982066	1.120111	1.371183	5.04892	2.619011
ENST0000	0	0	0	0	0
ENST0000	0.241376	0.119728	0.021227	0.009749	0
ENST0000	0.061229	0.032396	0.057434	0.039568	0.093569
ENST0000	0	0.581962	0	0	0
ENST0000	146.4966	163.5045	205.3889	162.6099	96.99319
ENST0000	0	0	0	0	0
ENST0000	0	0	0	0	0
ENST0000	8.933542	1.986797	2.840649	4.501061	0.370228
ENST0000	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 different?

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

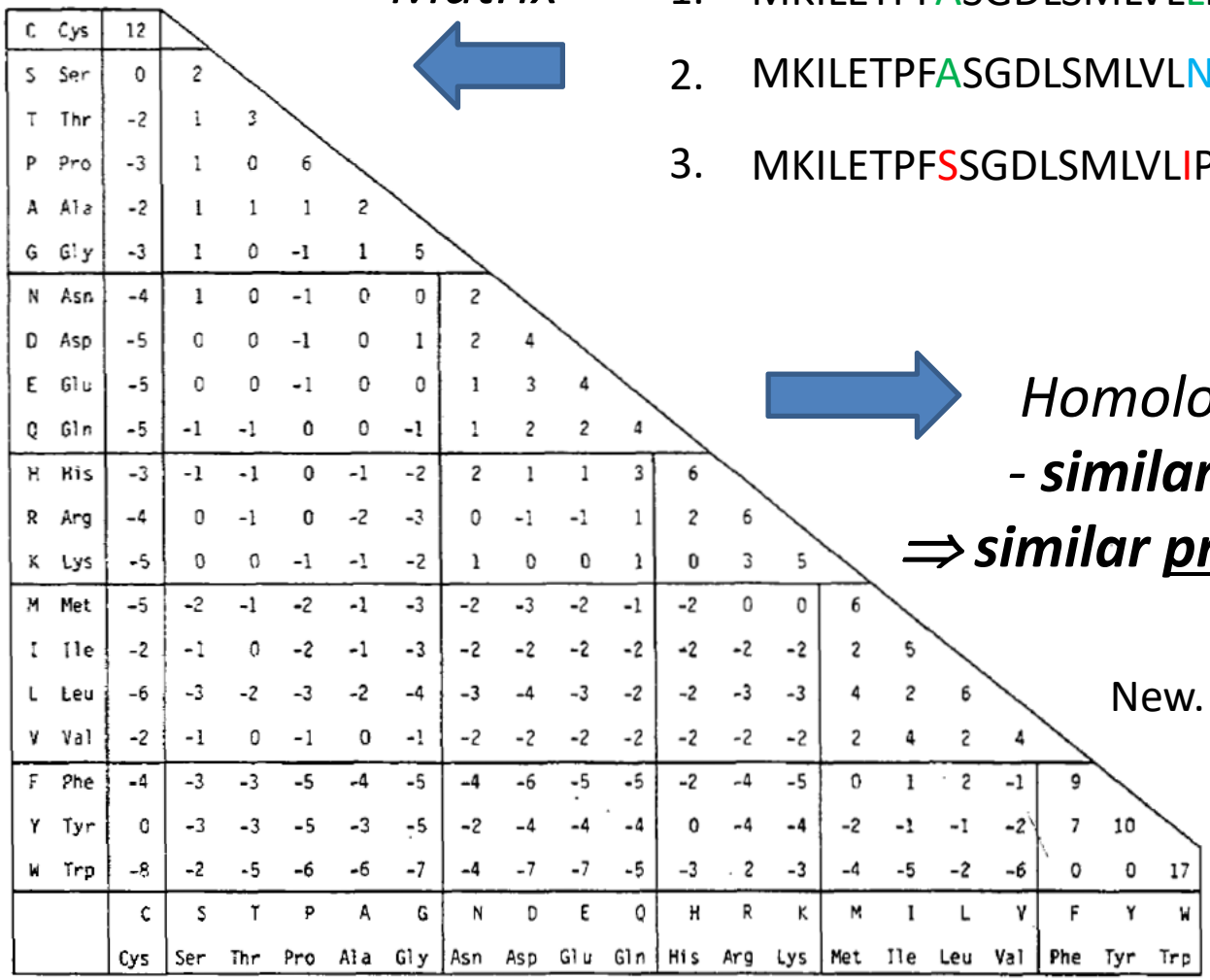
Computer!

Dayhoff matrix

Homology search

Dayhoff matrix & homology search

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	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6					
I	Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5				
L	Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	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

(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

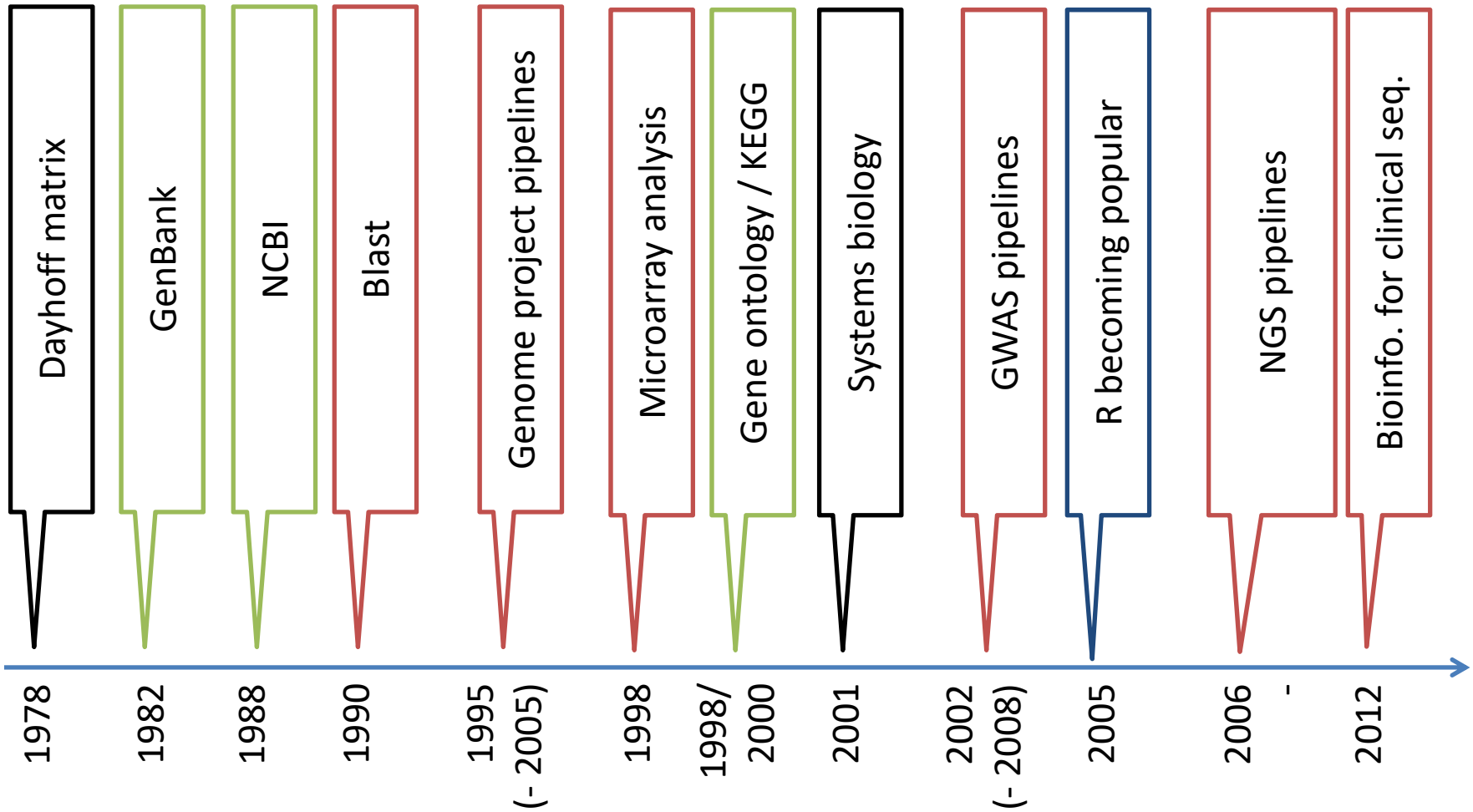
-4 - 4 + 5 + 5 = 2

New is more similar to seq1 even in the protein function. (Say, seq1 function is well-known.)

- **Protein (amino acid sequence)**

Same idea

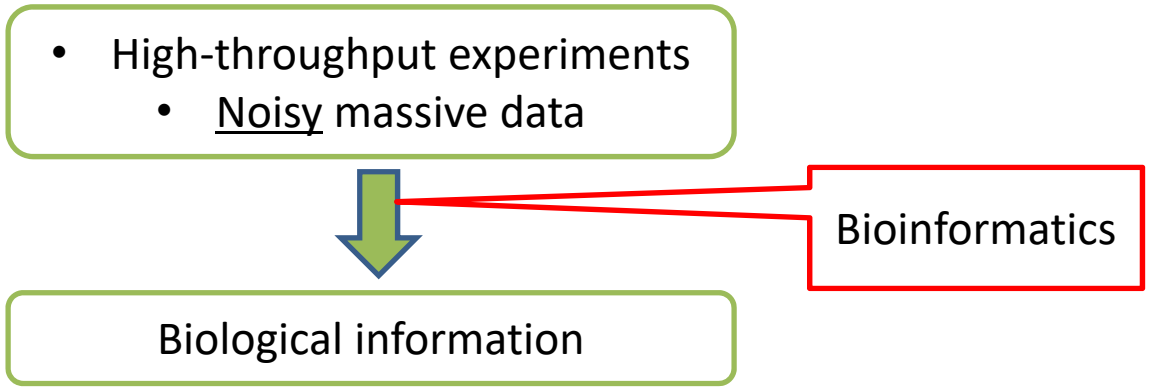
Brief history of bioinformatics



The essence

Signal from Noise

– a needle in a haystack –



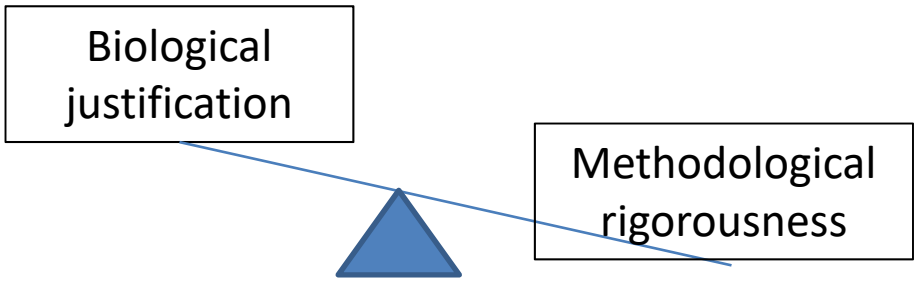
	Data type: phenotypic or molecular?	Data quantity	Emphasis: discovery or proof?
Bioinformatics	Molecular	Great many (>millions)	Discovery

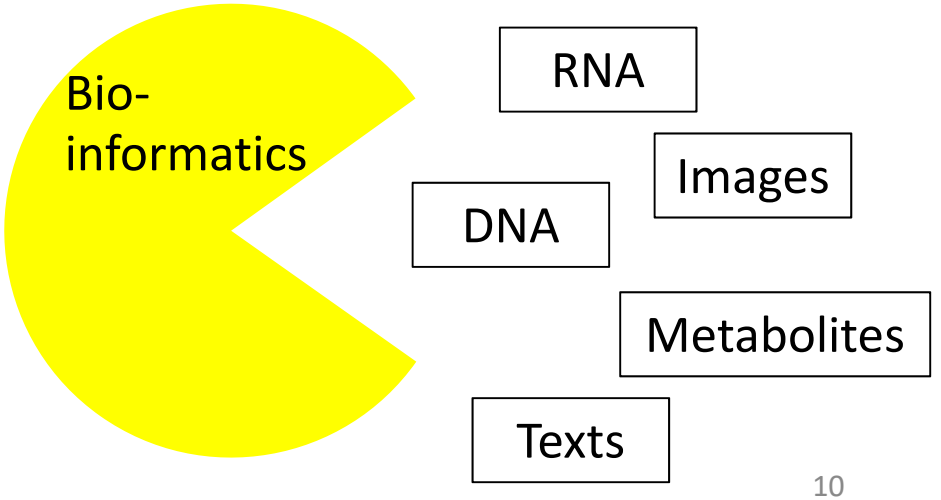
The methodological aspect

The theoretical basis

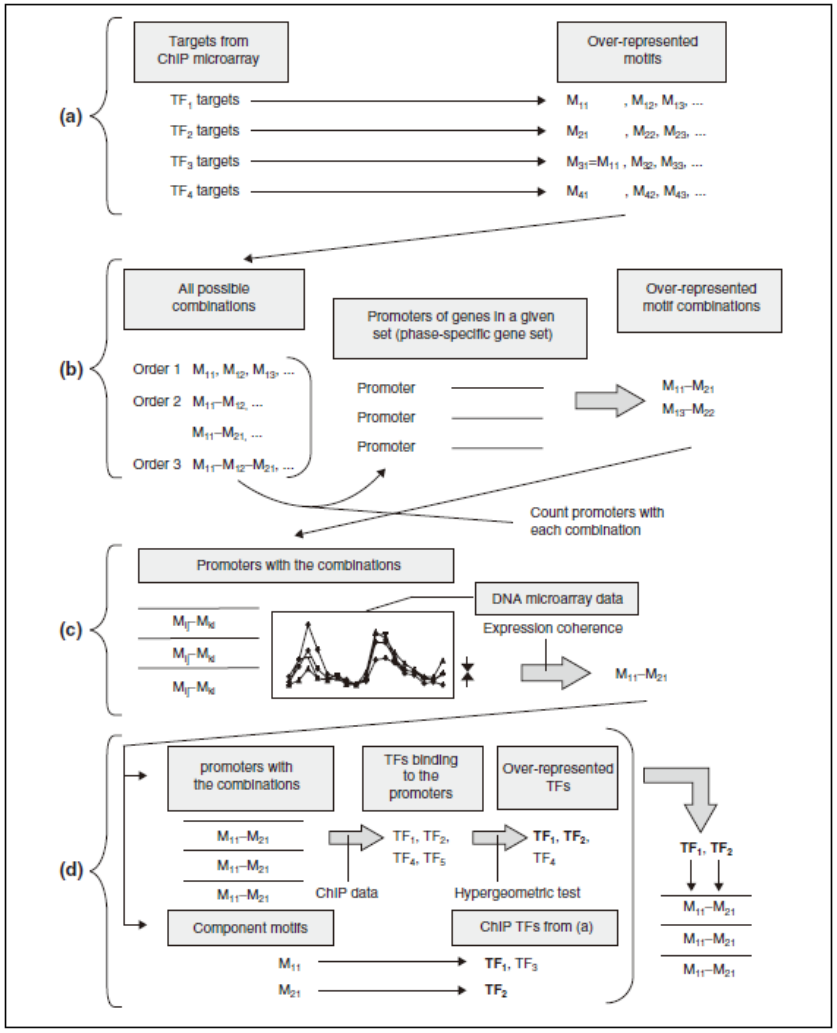
- Mechanics:
  Newton's three laws
- Statistics:
  population and sampling
- Bioinformatics:
  NOTHING…?

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



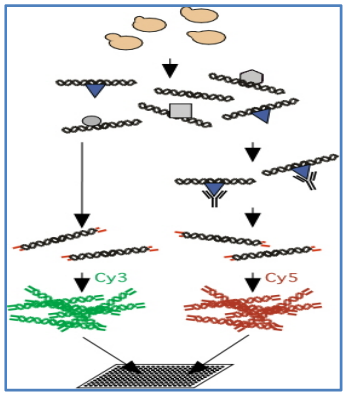


Discovery, relaxing methodological rigorousness



>YOR128C
TCGAAAAAGATCATT...
>YOR119W
CAGTTTATATAAATTT...
> ...

Promoter sequences represented by 200 * 6000 letters



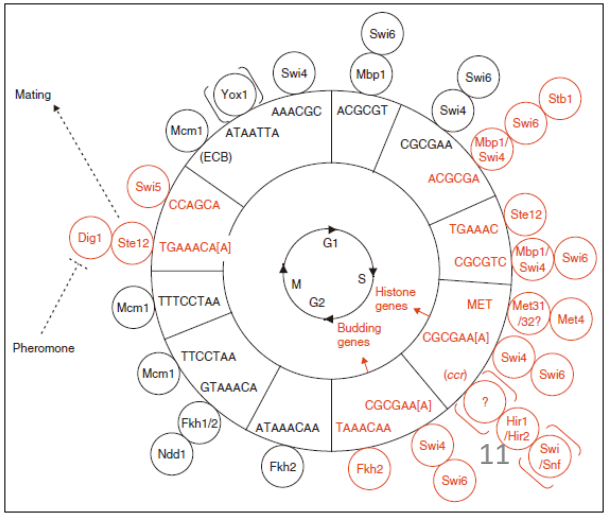
(Lee et al., Cell, 2002)

Binding data represented by 113 * 6000 p-values

In the step (a): search TF-binding motif

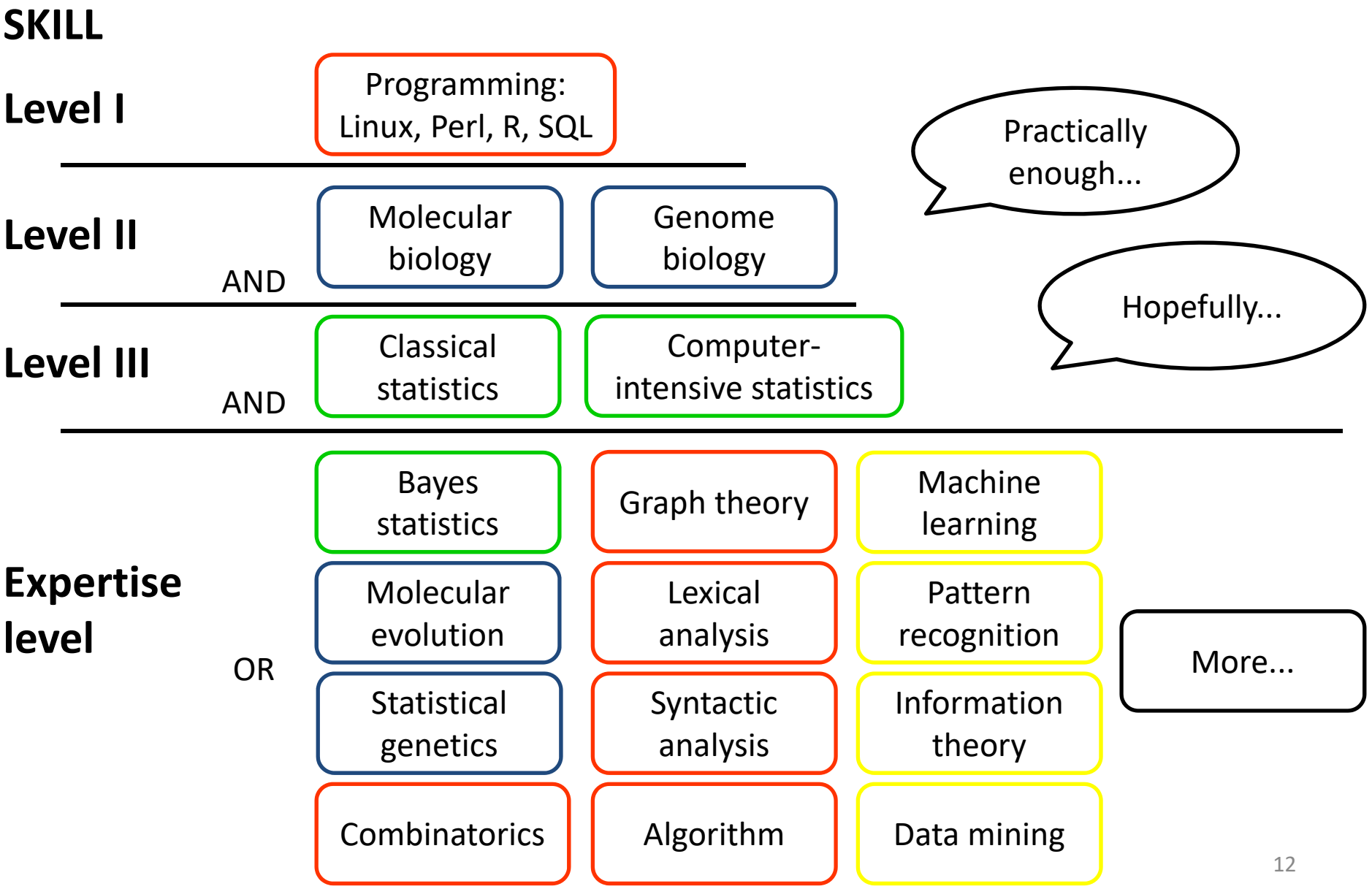
200,000 statistical tests! ⇒ 200,000 p-values

As, simply,
score...



(Kato et al, Genome Biol., 2004)

Therefore, collection of arts



Is level II enough?

Requested tasks

Cutting-edge bioinfo.

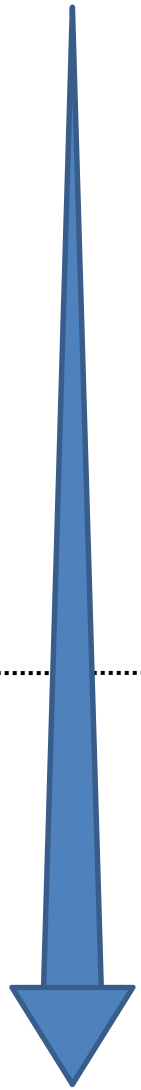
- Routine
  - NGS analysis
    - DNA: SNV/indel/CNA/SV calling
    - RNA: expression calling
  - Microarray analysis
  - General sequence analysis
  - GO/pathway analysis
  - SNP analysis
    - GWAS, haplotype, LD
  - Clustering analysis
  - PCA
- Ad-hoc operation
- Ad-hoc algorithm
- Non-routine data cleaning

- Deep data-interpretation
- Deep analysis

Skill *Experience*

Level II

Level III or more



Routine/Ad-hoc operation?

- NGS routine

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

…
```

- Ad-hoc operation

1. Installation of a public bioinformatics tool
   - Eg, “bwa” on the left
2. Understand the manual
   - Though, sometimes
3. Apply to your own data
   - Eg, command lines on the left
4. Get results

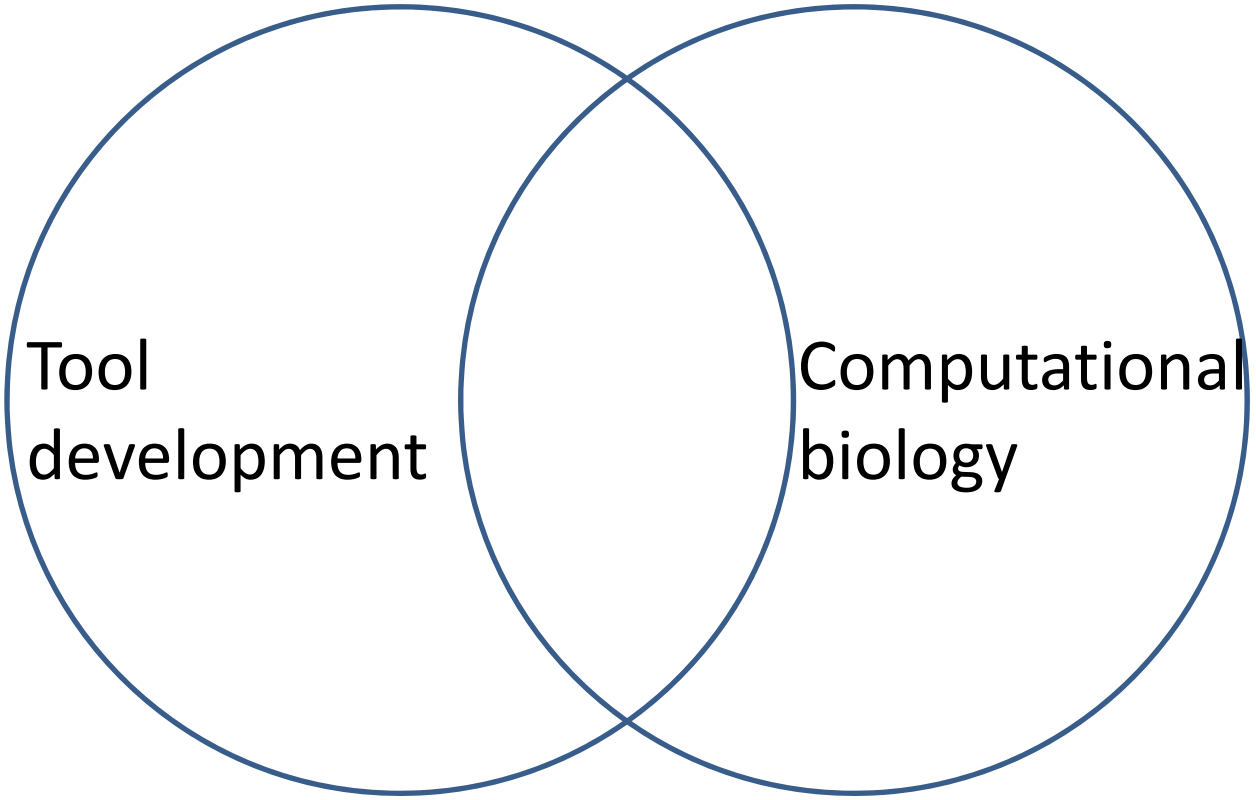
Advanced type of working: division of labor

- When bioinformatics tasks are many and complex, ...

Algorithm design
~= experimental design
by scientists

Implementation (coding)
~= perform experiment
by technical staff / technicians

Two extreme types of studies



Example of tool development

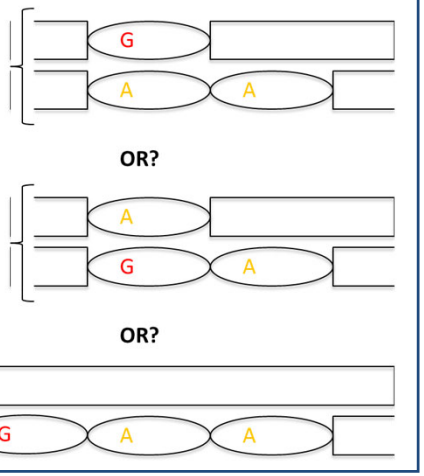
Problem in RETINA data

- What is the configuration of haplotypes(/genotypes)?

RETINA data: total number of bases over homologous chromosomes

Sample name	rs1065852		rs3892097		…
	FAM (G)	Yellow (A)	FAM (C)	Yellow (T)	
NA10860	1	2	1	2	…
NA11992	0	3	0	3	…
NA12878	1	1	1	1	…
NA12239	2	0	2	0	…
…	…	…	…	…	…

rs1065852 for NA10860

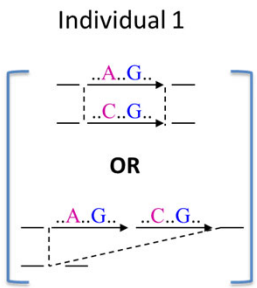


Principle of the algorithms

1. List all possible diplotypes (pairs of haplotypes) that are consistent with the total numbers

Sample	SNVC 1		SNVC 2	
	# of A	# of C	# of G	# of T
Ind. 1	1	1	2	0
Ind. 2	2	1	2	1
Ind. 3	1	0	0	1

SNVC: SNV in a CNV



Sample	Possible diplotype (haplotype / haplotype)
Ind. 1	AG / CG
	AG, CG / --
Ind. 2	…
…	…

","": separator between copy units

"-": deletion

Principle of the algorithms

2. Repeat E- and M-steps to estimate haplotype frequencies
   - Using possible diplotypes obtained at the previous step

Sample	Possible diplotype	Weight (probability that the sample takes this diplotype)	Diplotype frequency under Hardy-Weinberg equilibrium
Ind. 1	haplotype 1 / haplotype 1	$w_{11} \propto F(h1 / h1)$	$F(h1 / h1) = 1 \cdot F(h1) \cdot F(h1)$
	haplotype 1 / haplotype 2	$w_{12} \propto F(h1 / h2)$	$F(h1 / h2) = 2 \cdot F(h1) \cdot F(h2)$
	haplotype 2 / haplotype 3	$w_{13} \propto F(h2 / h3)$	$F(h2 / h3) = 2 \cdot F(h2) \cdot F(h1)$
	…	…	…
Ind. 2	haplotype 1 / haplotype 1	$w_{21} \propto F(h1 / h1)$	$F(h1 / h1) = 1 \cdot F(h1) \cdot F(h1)$
	…	…	…

F(x): frequency of x

Giving arbitrary values to haplotype frequencies → E step: Haplotype frequencies →Diplotype frequencies →Update the weights → Iteration → M step: Number of haplotypes, considering the weights →Haplotype frequencies

Haplotype (14 SNVC sites)	True count	True frequency	Estimated frequency
|AA-AAAAAAAAAAA|	428	0.3639	0.3683
|AA-AAAAAAAABAB|	382	0.3248	0.3235
|BA-AAABAAAAAAB|	244	0.2075	0.2066
|AA-AAAAAAA-AAA|	24	0.0204	0.0194
|----------|	23	0.0196	0.0199
|AA-AAAA---AAAA|	21	0.0179	0.0167
|BA-AAAAAAAAAAB|	18	0.0153	0.0153
|AA-AAAAAAAABAB, AA-AAAAAAAABAB|	16	0.0136	0.0118
|AA-A-AAAAAAAAA|	11	0.0094	0.0079
|AA-AAAAAAAAAAA, AA-AAAAAAAAAAA|	6	0.0051	0.0050
|BAAAAABAAAAAAB|	-	-	0.0009
…	…	…	…

CNV-phaser (Kato et al, Am. J. Hum. Genet., 2008)

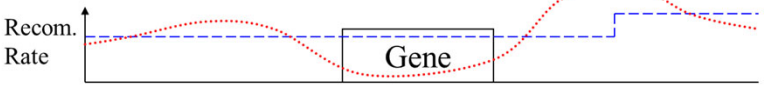
Allelic copy number	True count	True frequency	Estimated frequency
1 (copy)	1130	0.9609	0.9601
0 (copies)	23	0.0196	0.0200
2 (copies)	23	0.0196	0.0200
3 (copies)	-	-	$<10^{-10}$

MOCS-phaser (Kato et al, Bioinformatics, 2008)

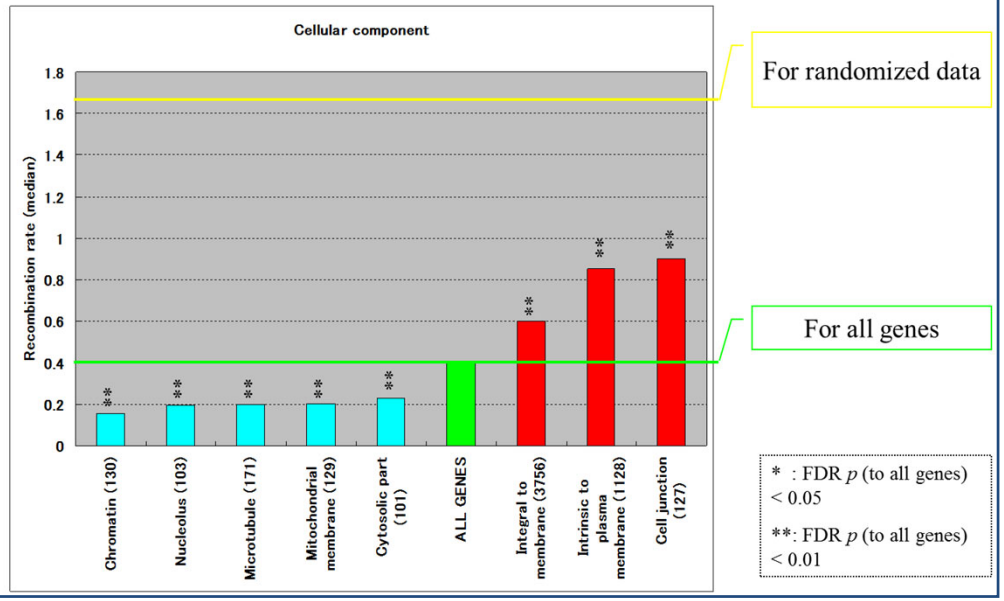
Example of computational biology

Background

- *2003* – The complete sequence of the human genome
  - an "average" sequence
- *2005* – The International HapMap Project, Phase I
  2007 – The International HapMap Project, Phase II
  - Polymorphism in human genomes
  - Focused on single nucleotide polymorphism (SNP)
    - ✓ Catalogued SNP genotypes for 270 individuals in three ethnical populations (Asian, African, European) at 3 million SNP loci
  - Medical application as well as biological investigation
- The first data on human recombination rates at a high resolution
  - New statistical methods (by a HapMap group) to infer recombination rates from large-scale SNP genotyping data
  - kilo-base level high-resolution

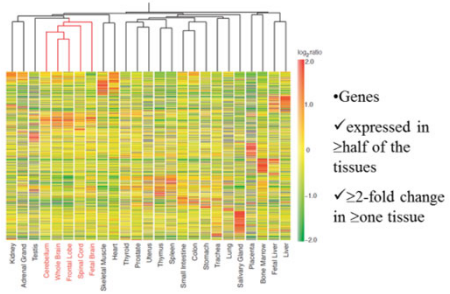


Gene Ontology: *cellar component*

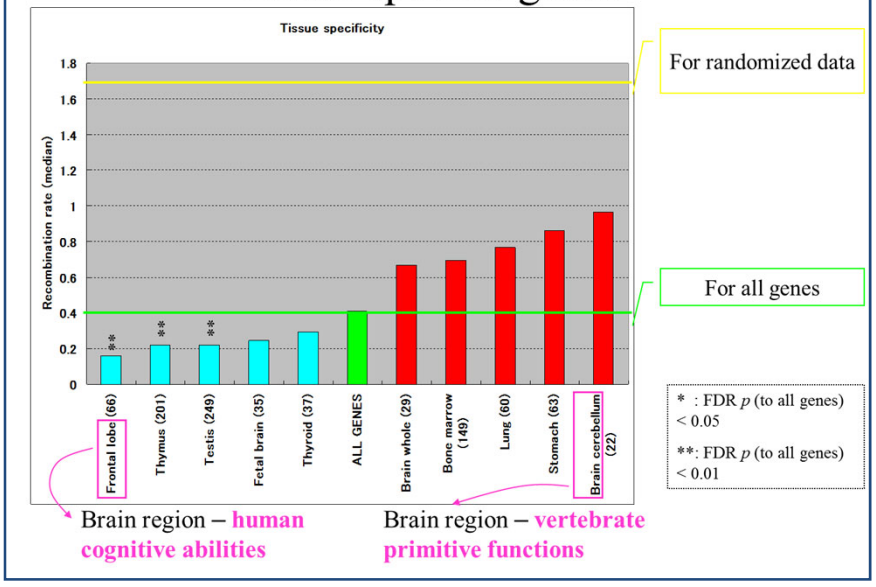


Tissue-specific genes

- Tissue-specific genes
  - We performed microarray experiments.
  - Microarrays with ~30,000 probes for 25 tissue samples
  - Genes highly expressed (FDR $p < 0.001$) in each tissue
  - The clustering analysis confirmed that we successfully measured expression levels.
  - Genes highly expressed in only one tissue → tissue-specific genes

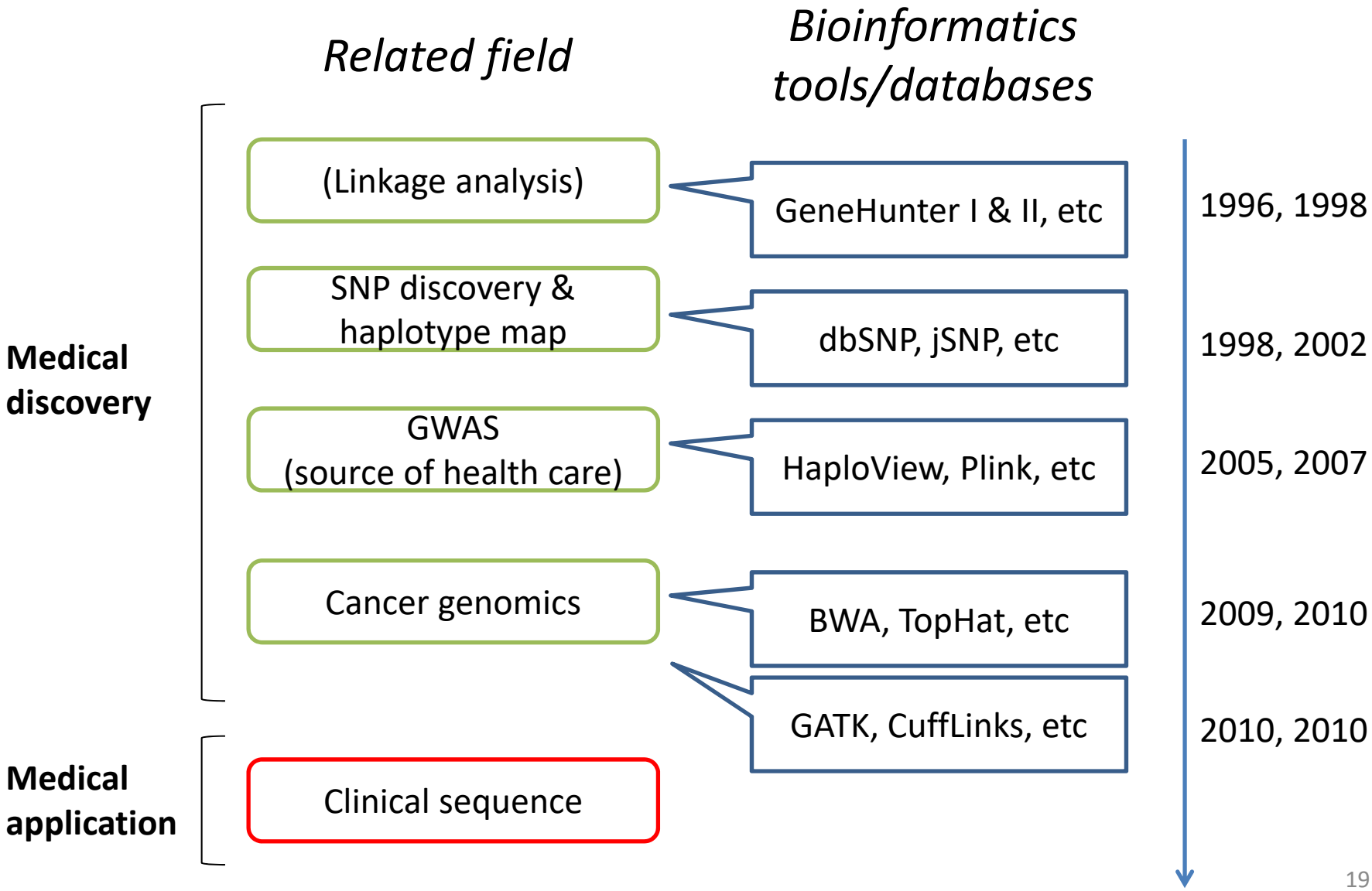


Tissue-specific genes



(Kato et al, Hum. Mol. Genet., 2008)

Genomic medicine & bioinformatics



The latest – clinical sequencing

- Detection of different mutations in just one assay
  - *Multiplex PCR, mass spectrometry, FISH*
    - ✓Mostly, single type and single to several mutations
  - *Next-generation sequencing*
    - ✓Point mutations, fusions, amplifications, deletions
    - ✓All exons of ~100 genes (in our case)
    - ✓Potential for research discovery

Next-generation sequencer

1 G byte / 2 hours

```
@PERI8:15:42
GGCTCATCAAGCTCGCTGCTTCCAGGAATGACTGGGAAGGTGGGAAGGAGAGAAGATGCGTGGGT
TCTTCA
+
;(1.9...605344911.452;4<=<A<BB@??=@3;/9/,,0&0&,065455)+.4831**-%-(*2(-*
@PERI8:15:43
TCCCCCCCCCCCAAATGTTCACTAACTCTGAAACGTGGGTCTCTGACTGGCTCCCGTGCCAGCCAGTC
GACGCATGGTCTGTA
+
-0221111222(16.445,65569355766><079.00&,440++37,,+,**-%-211.411+03,,,,,+,411(..+-+4
...
```

Gene alterations and drugs

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*)[8,37,73,74]	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

Gene alterations and molecularly targeted drugs (Reviewed by Simon et al, 2013)

- If one assay simply detects these, ...

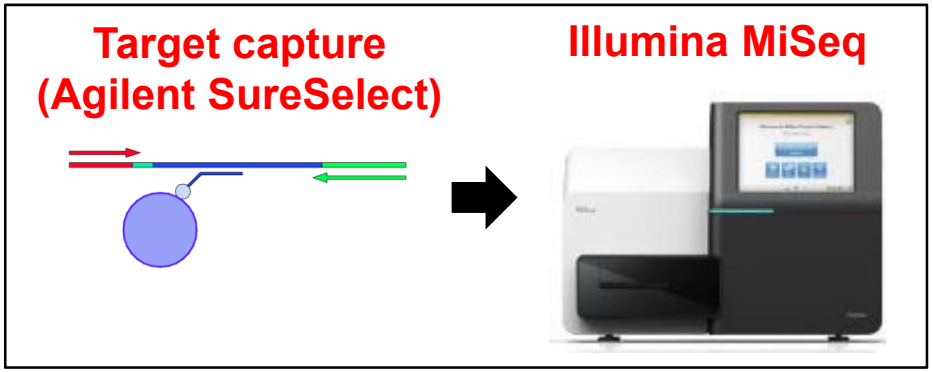
Clinical sequencing for cancer in National Cancer Center, Japan

– Collaboration by doctors and scientists –

薬物治療を検討している患者

- A遺伝子の変異例 → Aタンパク質阻害剤による治療
- B遺伝子の変異例 → Bタンパク質下流分子経路阻害剤による治療
- C遺伝子の増幅例 → 抗Cタンパク質抗体薬による治療

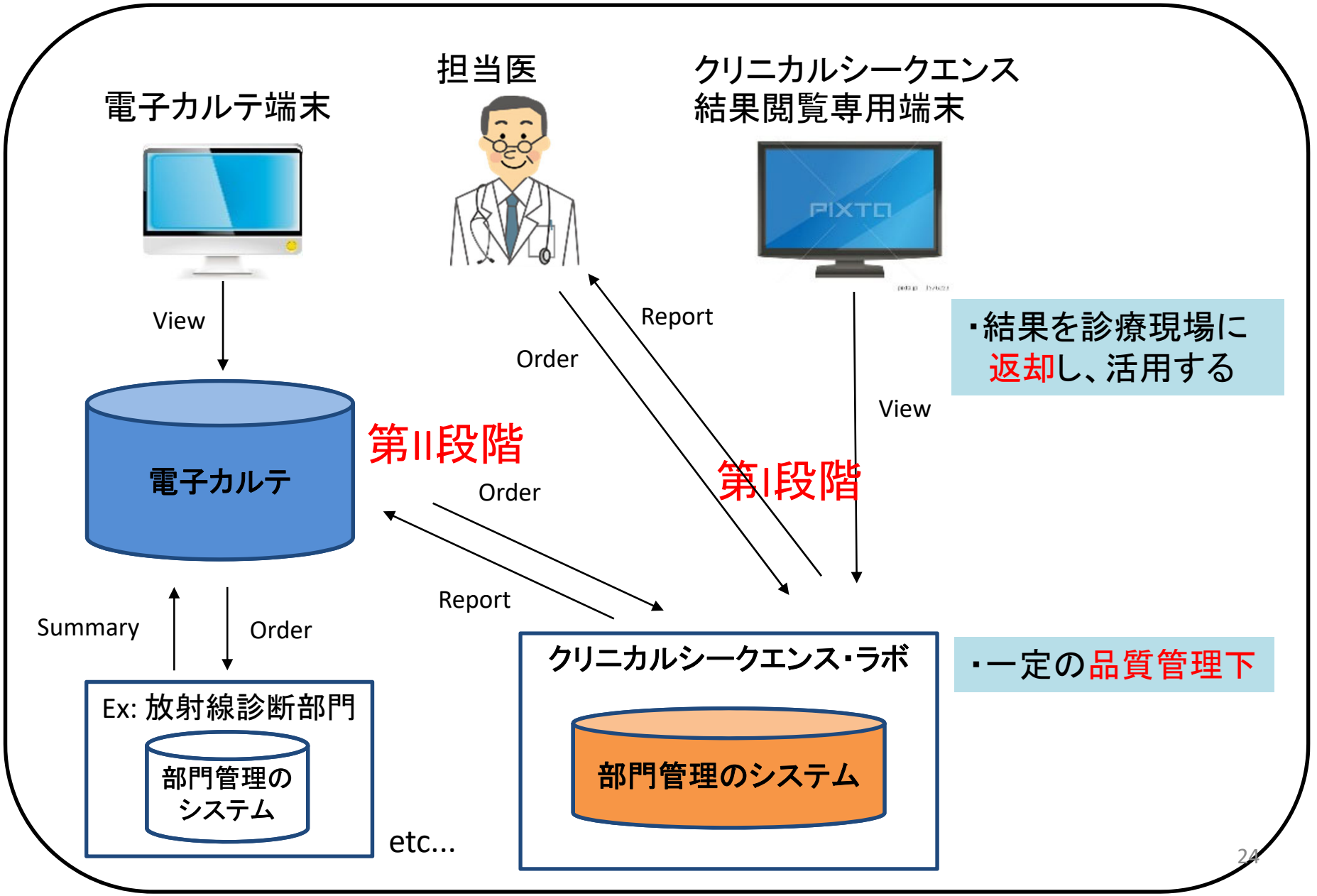
Platform and genes



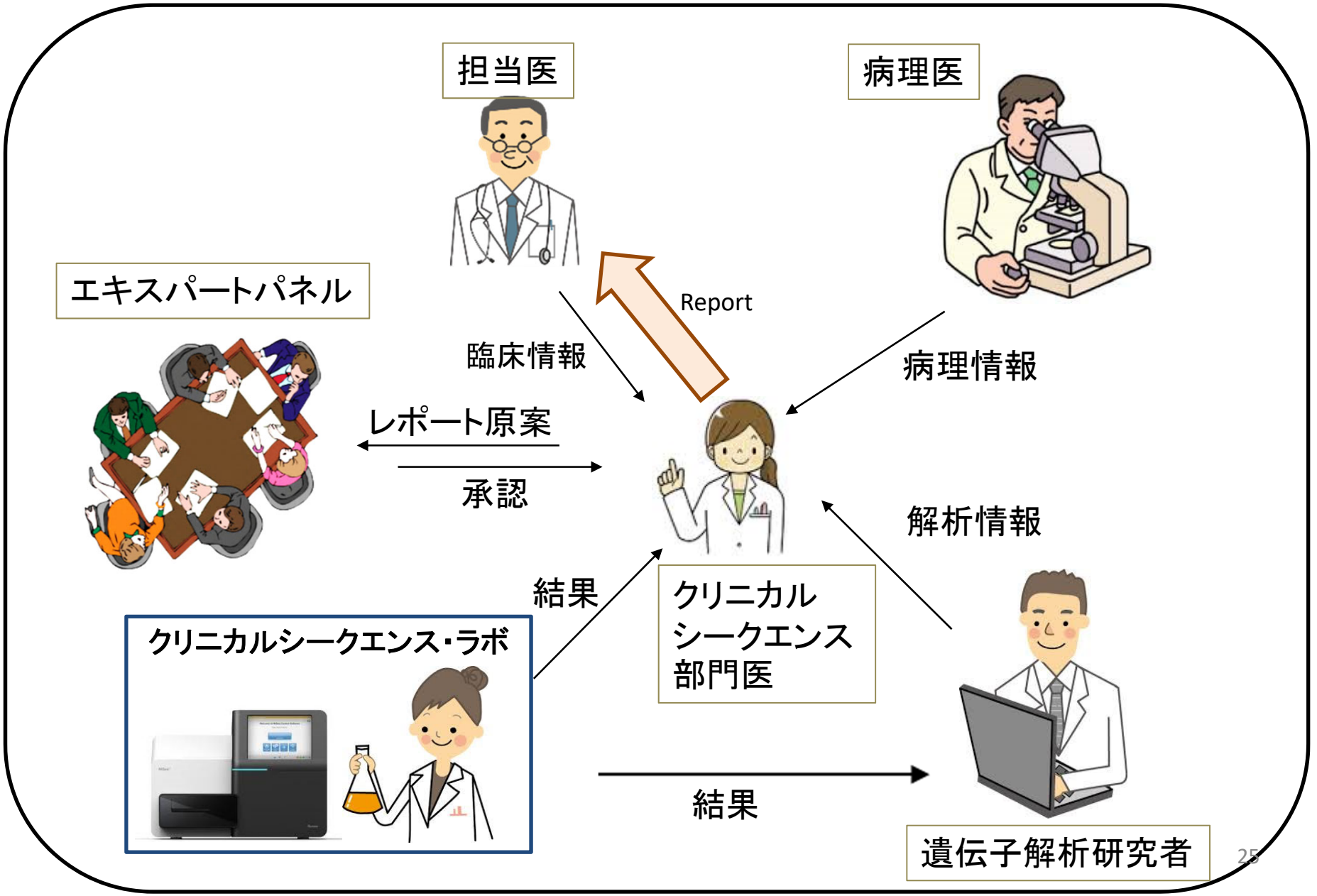
NCC oncopanel v2

Mutation / Amplification					Fusion
ABL1	CREBBP	IGF1R	NOTCH1	ROS1	ALK
AKT1	CTNNB1	IGF2	NOTCH2	SETD2	RET
AKT2	CUL3	IL7R	NOTCH3	SMAD4	ROS1
AKT3	DDR2	JAK1	NRAS	SMARCA4	FGFR2
ALK	EGFR	JAK2	NRG1	SMO	FGFR3
APC	ENO1	JAK3	NT5C2	STAT3	AKT3
ARID1A	EP300	KEAP1	PALB2	STK11	BRAF
ARID2	ERBB2	KIT	PBRM1	TP53	RAF1
ATM	ERBB3	KRAS	PDGFRA	TSC1	NOTCH1
AXIN1	ERBB4	MAP2K1	PDGFRB	VHL	NRG1
BAP1	EZH2	MAP2K4	PIK3CA		
BARD1	FBXW7	MAP3K1	PIK3R1		
BCL2L11	FGFR1	MAP3K4	PTCH1		
BRAF	FGFR2	MDM2	PTEN		
BRCA1	FGFR3	MET	RAC1		
BRCA2	FGFR4	MTOR	RAC2		
CCND1	FLT3	MYC	RAD51C		
CDK4	HRAS	MYCN	RAF1		
CDKN2A	IDH1	NF1	RB1		
CHEK2	IDH2	NFE2L2	RET		

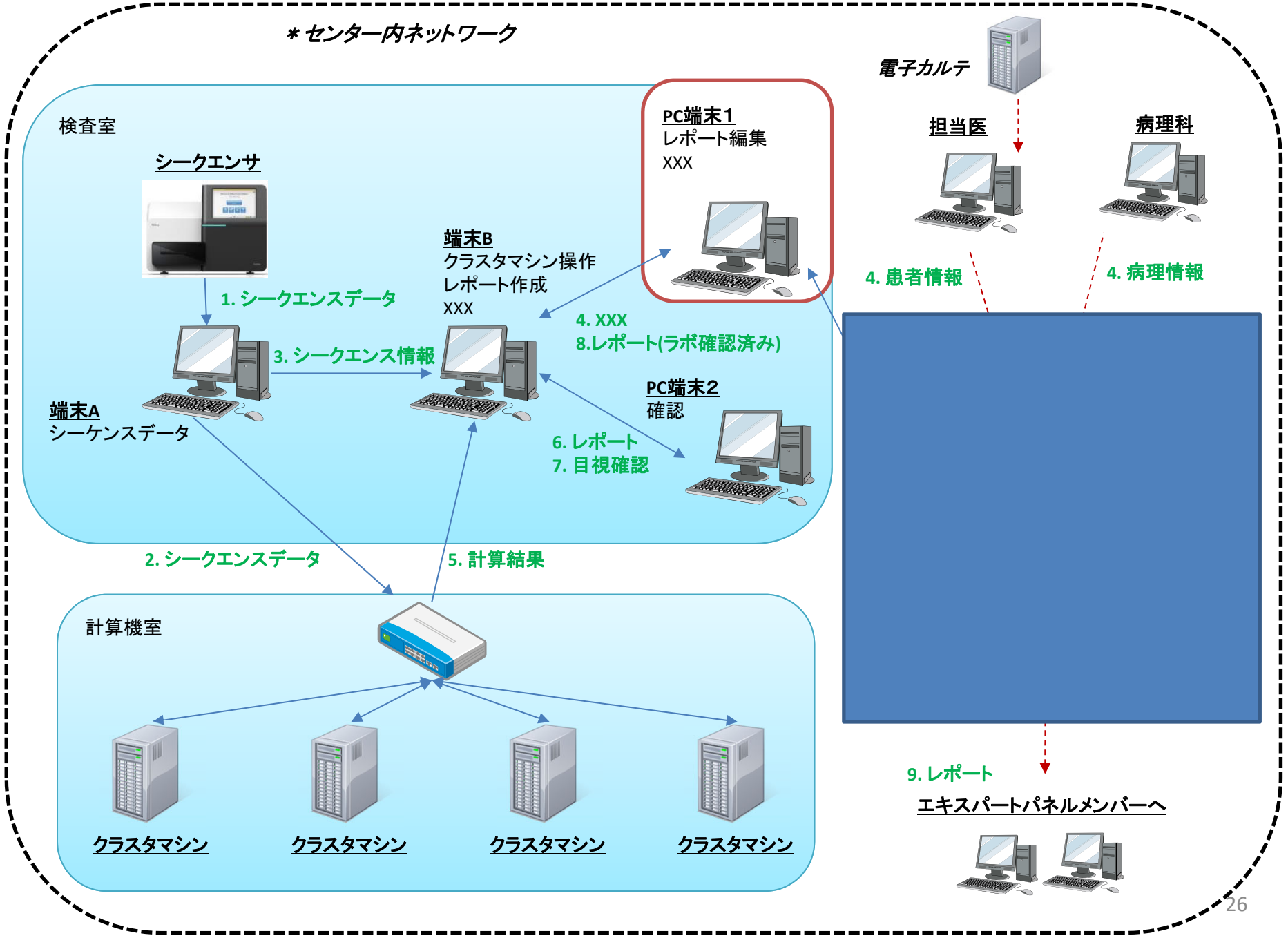
ゲノム医療に向けて：
中央病院内　クリニカルシークエンス・ラボ（仮称）　運用計画（案）



ゲノム医療に向けて：実施体制（暫定案）



■データフロー（暫定案）＊バイオインフォマティクス部門設計



システムモデル Ver. 1

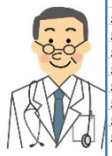
開発した変異検出プログラム（cisCall）

1．計算変異情報

2．シークエンス情報カード

シークエンス情報	試料情報	切り出しの有無
シークエンス情報	試料情報	Qubit測定DNA量（ug）
シークエンス情報	試料情報	qPCR測定DNA量（ug）
シークエンス情報	試料情報	DNA品質（qPCR/Qubit比）
シークエンス情報	測定情報	パネル
シークエンス情報	測定情報	試薬
シークエンス情報	測定情報	使用DNA量（ug）
シークエンス情報	測定情報	シークエンサー
シークエンス情報	測定情報	シークエンサーラン日
シークエンス情報	測定情報	リードデータ名

3．患者情報カード



患者情報	基礎情報	NA	診療科
患者情報	基礎情報	NA	文書同意日
患者情報	基礎情報	NA	性別
患者情報	基礎情報	NA	同意取得時年齢
患者情報	基礎情報	NA	臨床診断
患者情報	基礎情報	NA	遠隔転移部位
患者情報	基礎情報	NA	臓器機能にかかわる問題点
患者情報	基礎情報	NA	PS
患者情報	治療情報	1	レジメン
患者情報	治療情報	1	治療開始日
患者情報	治療情報	1	治療終了日
患者情報	治療情報	1	コース数
患者情報	治療情報	1	放射線
患者情報	治療情報	1	最良効果
患者情報	治療情報	1	中止理由

分類	項目
識別情報	FFPEブロック番号
識別情報	担当病理医
試料情報	検体組織

4．病理情報カード

病理情報	試料情報	腫瘍細胞率（%）
病理情報	試料情報	炎症細胞
病理情報	検査情報	EGFR 変異

開発した
情報統合
プログラム

- 現在これらのパーツを連結し、実際の運用体制を構築中

開発したDB (cisMedi)

シークエンスリポート
出力例

エキスパートパネル・リポート
出力例

エキスパートパネル

確定変異情報

電子カルテ

開発した臨床シークエンス用変異検出プログラム（cisCall）

Sample → Target enrichment → Sequencer → Bioinformatics system → Mutations

- **FFPE**
- **Low tumor purity**

- TargetSeq
- SureSelect
- HaloPlex
- multi-PCR

- Ion PGM
- Ion Proton
- Illumina MiSeq
- Illumina HiSeq

Clinical Sequence Call (*cisCall*) System

- Sequence data
- Read-QC report

cisMuton

- SNV/indel call table
  - Call-QC report

cisFusion

- Fusion call table
  - with breakpoints

cisCton

- CNA call table
  - with visualization

SNV
– single nucleotide variation –
and indel

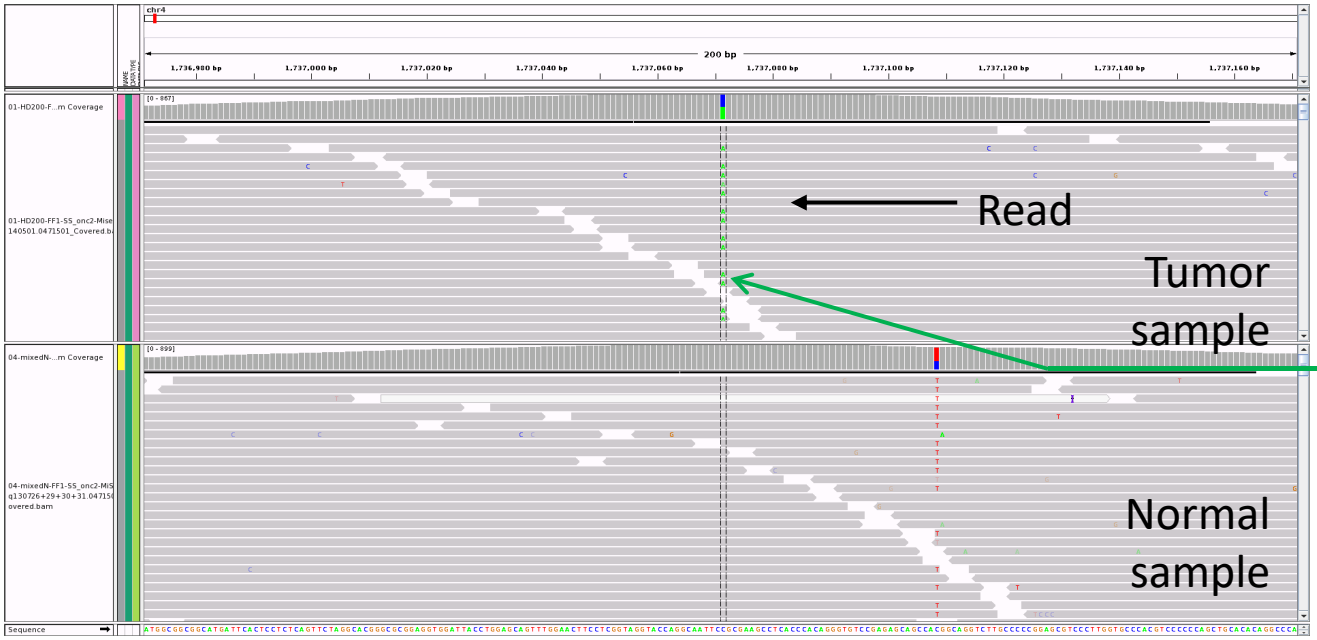
Principle of the detection method

Fragmented DNAs are sequenced by NGS

GCCCGG
AGCCCG
AAGCCC

Variant in tumor sample

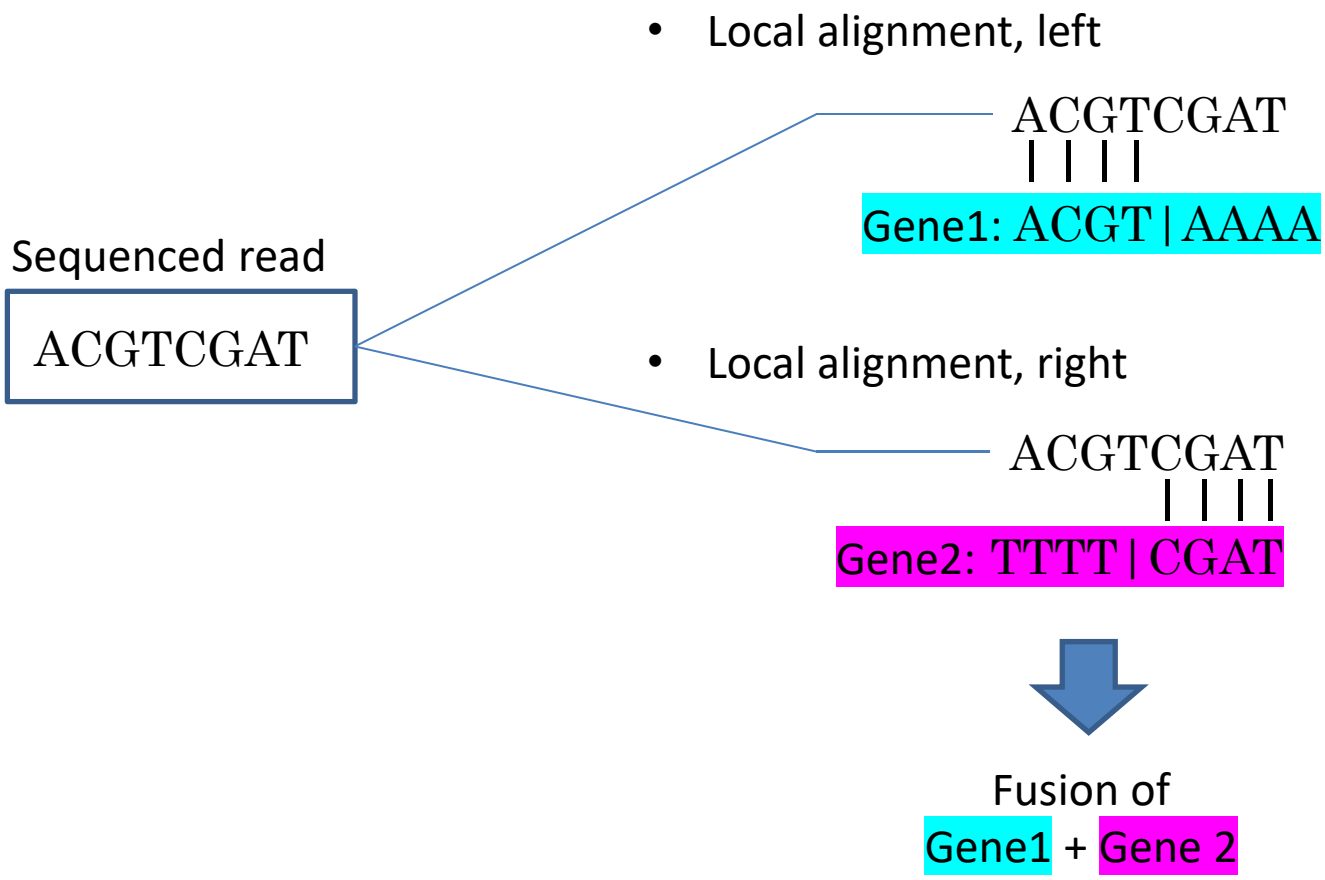
Human genome DNA sequence ....AAAACCCCGGGGTTTT....



- Identify bases <u>present in the tumor but absent in the (unmatched) normal</u>

Fusion genes

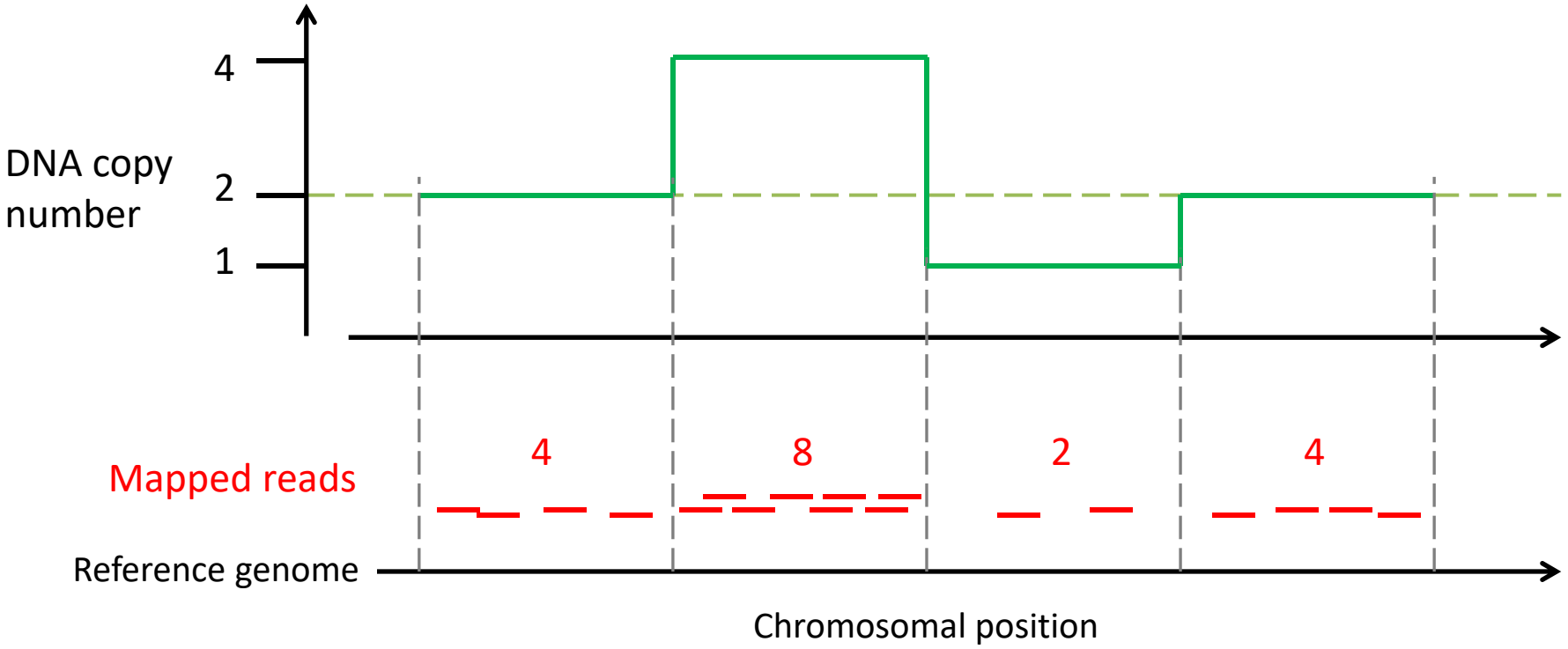
Principle of the detection method



CNA
– copy number alteration –

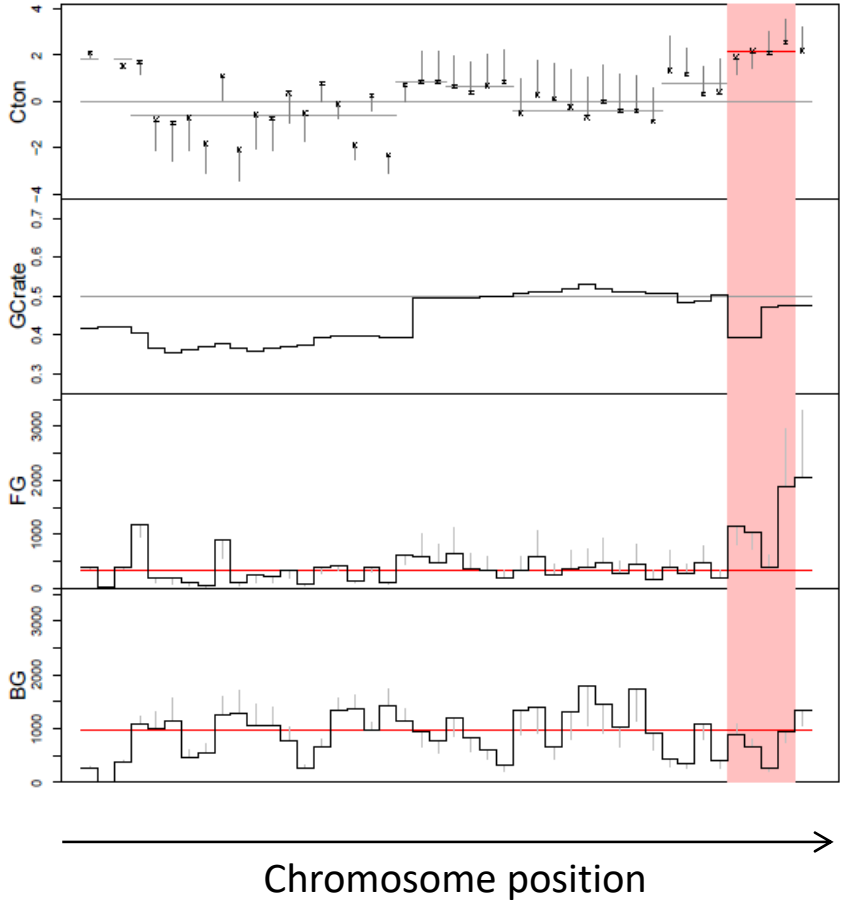
Principle of the detection method

- Mapped sequence reads ↗, copy numbers ↗
- Difference in depth between the tumor and normal

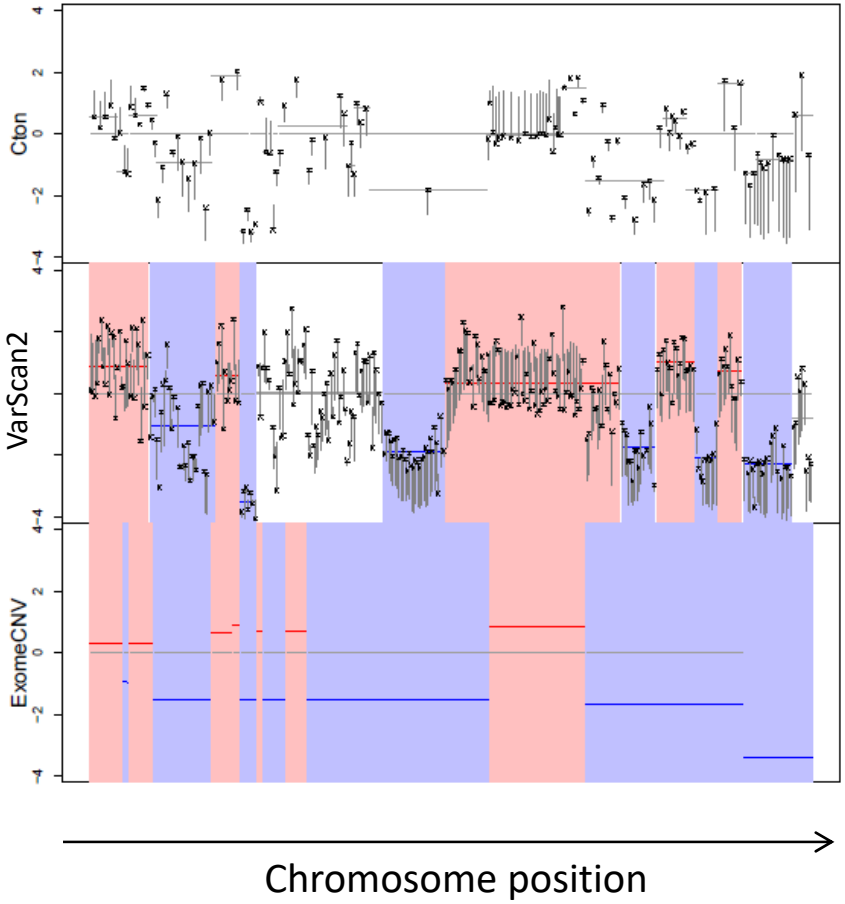


In practice, …
This is the very place of bioinformatics

ポジコン
qPCR 増幅 LogRatio: 3.14



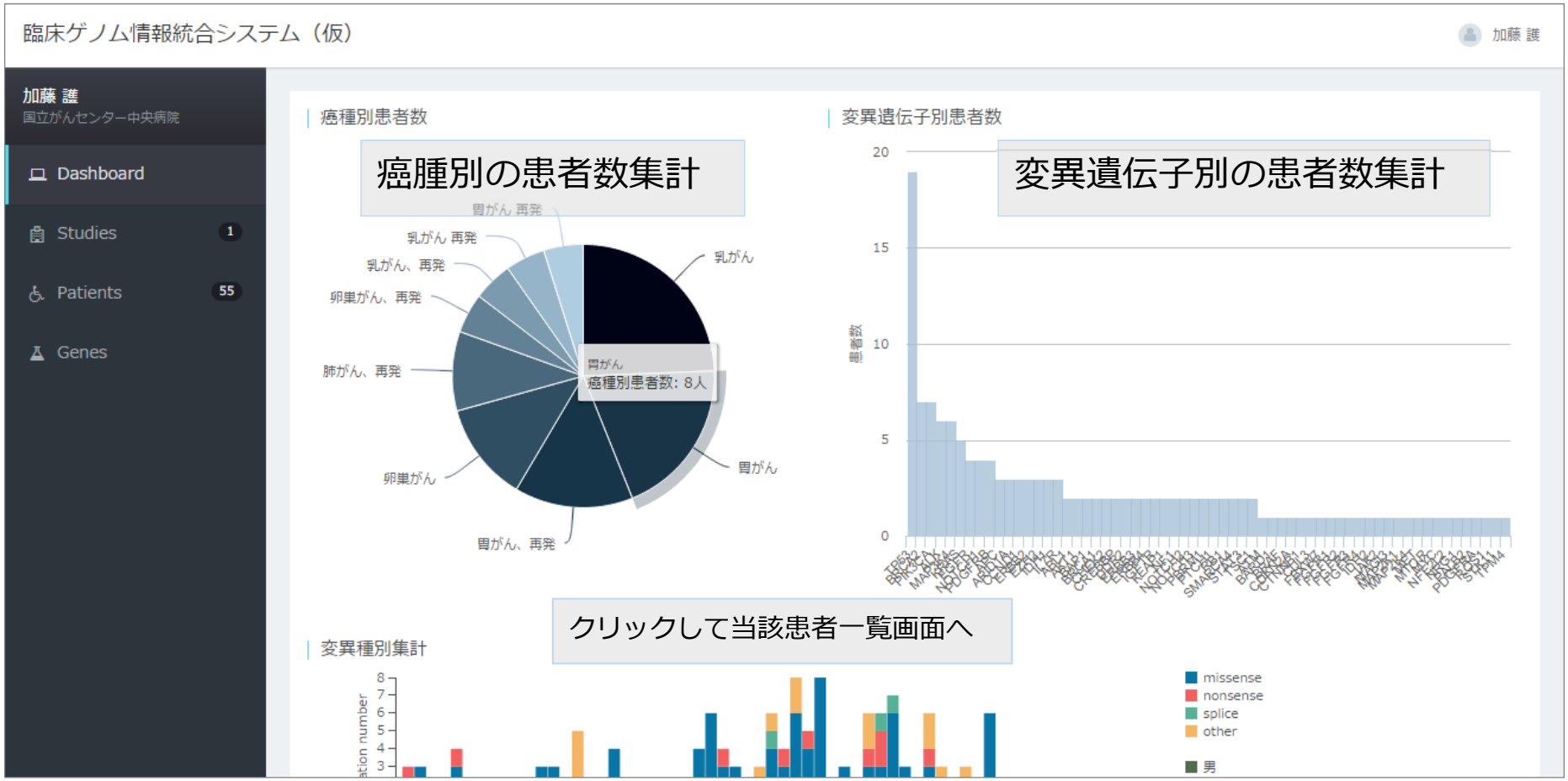
ネガコン
同一検体 FFPE-Frozen



開発したDB (cisMedi):
Database & post-calling system

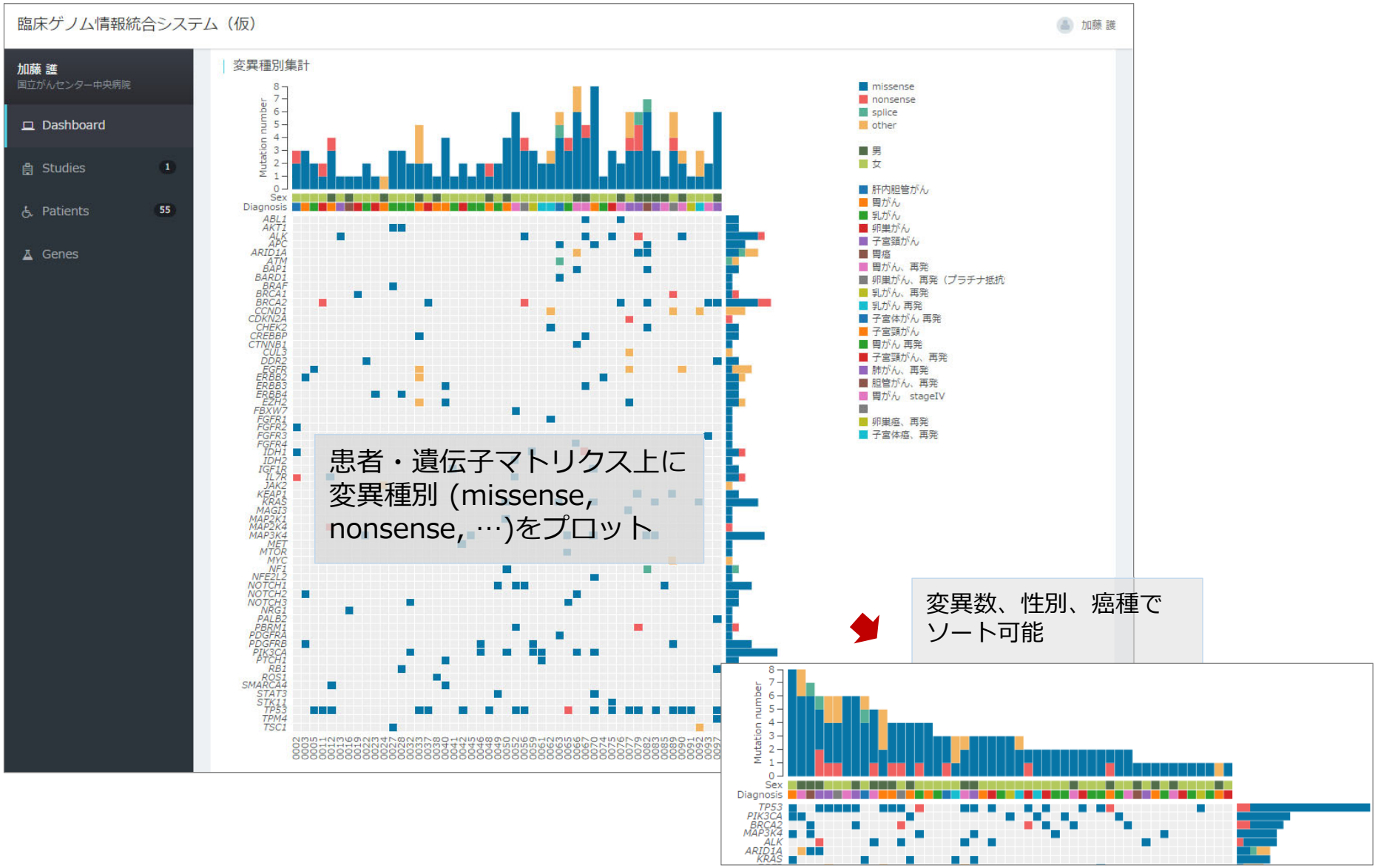
cisMedi system – Clinical Sequence Medical Information system

ダッシュボード画面 #1　集計情報を表示



With Dr. Tsuchihara

ダッシュボード画面 #2


臨床ゲノム情報統合システム（仮）
加藤 護
加藤 護
国立がんセンター中央病院
Dashboard
Studies 1
Patients 55
Genes
変異種別集計
Mutation number
Sex
Diagnosis
missense
nonsense
splice
other
男
女
肝内胆管がん
胃がん
乳がん
卵巣がん
子宮頸がん
胃癌
胃がん、再発
卵巣がん、再発（プラチナ抵抗
乳がん、再発
乳がん 再発
子宮体がん 再発
子宮頸がん
胃がん 再発
子宮頸がん、再発
肺がん、再発
胆管がん、再発
胃がん stageIV
卵巣癌、再発
子宮体癌、再発
患者・遺伝子マトリクス上に
変異種別 (missense,
nonsense, …)をプロット
変異数、性別、癌種で
ソート可能
TP53
PIK3CA
BRCA2
MAP3K4
ALK
ARID1A
KRAS

患者一覧画面

患者個別の情報を表示

患者詳細画面 #1


臨床ゲノム情報統合システム（仮）
加藤 謙
加藤 謙
国立がんセンター中央病院
Dashboard
Studies 1
Patients 55
Genes
匿名化患者番号:
ExpertPanel実施日
1 2013年11月12日
エキスパートパネル・
リポートを出力
臨床情報
Expert Panel報告書を出力
性別: 女
同意取得時年齢:
診療科:
担当医:
検体番号:
文書同意日:
臨床診断:
TNM分類: 再発
遠隔転移部位:
肝
肺
腹膜
リンパ節
骨
脳
その他 ()
PS: 1
保険:
臓器機能にかかわる問題点: なし
前治療
レジメン
放射線
治療開始日
治療終了日
コース数
最良総合効果
中止理由
PD

エキスパートパネル・リポート
出力例

Expert Panel報告書

Expert Panel日:

TOP-GEAR番号:
同意取得時年齢:
文書同意日:

検体番号:
診療科:
臨床診断:

性別: 女

遠隔転移部位: ☑肝臓 ☐肺 ☐腹膜 ☐LN ☐骨 ☐脳 ☐その他 ( )

PS: 1　保険:　臓器機能にかかわる問題点:

前治療

レジメン	放射線	治療開始日	治療終了日	コース数	最良総合効果	中止理由

検体情報

検体組織	採取法	組織型	切片の大きさ (cm)	腫瘍細胞率 (%)
肝臓（原発）	手術	腺癌		50

* 広範な壊死や固定不良は認められない。

Qubit測定DNA量 (ug)	DNA品質 (qPCR/Qubit比)
1.62	0.794499

遺伝子異常情報

変異遺伝子	変異アレル頻度	CDS変化	アミノ酸変化	COSMIC ID
IDH1	33.4	Exon4:c.C394T	p.R132C	28747
IL7R	52.7	Exon5:c.T603A	p.Y201X	
FGFR2	69.4	Exon7:cG870T	p.W290C	1346285

Expert Panelからの意見

・IDH1：既知の機能獲得変異である。対応する治験薬なし。

・IL7R：短縮型変異のため、機能欠失変異と考えられる。対応する治験薬なし。

・FGFR2：COSMICデータベースに複数の登録があり、活性化変異である可能性がある。FGFR阻害剤が候補にあがる

Conclusions

- Bioinformatics – greedy and cloudy discipline
  - ✓ Originally, it's generated from biophysics + molecular evolution
    - ➢ Alignment of protein (amino acid) sequences
    - ➢ Homology search
  - ✓ Expanding into many types of massive biological data
- The essence – signal from noise (a needle in a haystack)
  - ✓ Discovery > proof
    - ➢ Relaxing methodological rigorousness
  - ✓ Pragmatism – take in whatever method if useful
    - ➢ No theoretical basis…?
- Minimully required skills are NOT many
  - ✓ Programming: linux, perl, R, SQL
  - ✓ Biology: molecular biology, genome biology
- Can divide tasks into: algorithm design + program implementation
- Two study types: tool development + computational biology
- Expanding from science to medicine
  - ✓ Bioinformatics for clinical sequencing

END