遺伝子発現解析とデータマイニング(２)

統合薬学教育開発分野
奥野　恭史
okuno@pharm.kyoto-u.ac.jp

Nature 403, 503 - 511 (03 February 2000); doi:10.1038/35000501

Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling

B細胞リンパ腫(DLBCL)の40%は延命、60%は致死

↓

この差が何に起因するのか？は不明

↓

DLBCLのマイクアレイ実験より、発現パターンが2タイプに分類
(GC B-like DLBCLとActivated B-like DLBCL)

↓

実際の患者の延命期間とこれら2タイプが対応している

何がマイニングできるか？

＜遺伝子側クラスタリング＞

1. データセットの中に何種類の遺伝子発現パターンが含まれているか？
2. 遺伝子Xはどの機能カテゴリーに属するか？
3. 機能未知の遺伝子群の発現パターンの中に、すでによく知られた遺伝子の発現パターンと似たものはあるか？

＜細胞・組織（サンプル）側クラスタリング＞

4. 疾患Xのサブタイプを組織の遺伝子発現パターンで認識、発見することができるか？
5. 対象の組織サンプルはどの組織由来か？

何がマイニングできるか？

＜遺伝子相互作用ネットワーク＞

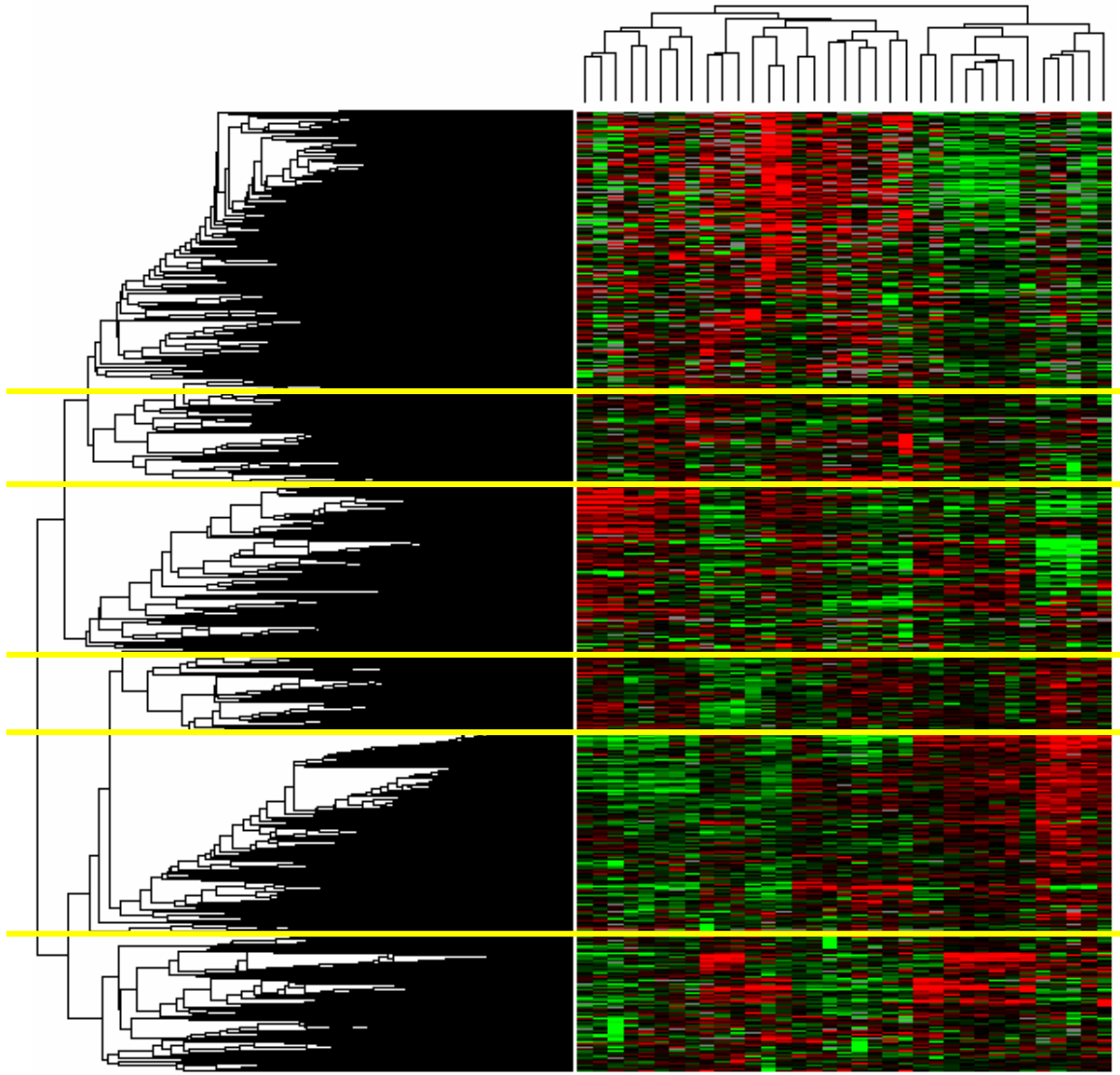
6. 対象の組織サンプルで観測される全ての遺伝子間の相互作用の違いは？
7. 発現パターンが類似した全ての遺伝子ペアを明らかにできるか？

＜遺伝子ハンティング＞

8. 正常と疾患など2群の組織サンプルを最もよく識別できる遺伝子群はどれか？
9. 薬物の影響を受けている遺伝子群は？
10. ある遺伝子Xの発現パターンは他の遺伝子群と比較してどれくらい特異か？
11. 医薬品のターゲットとなる遺伝子は？

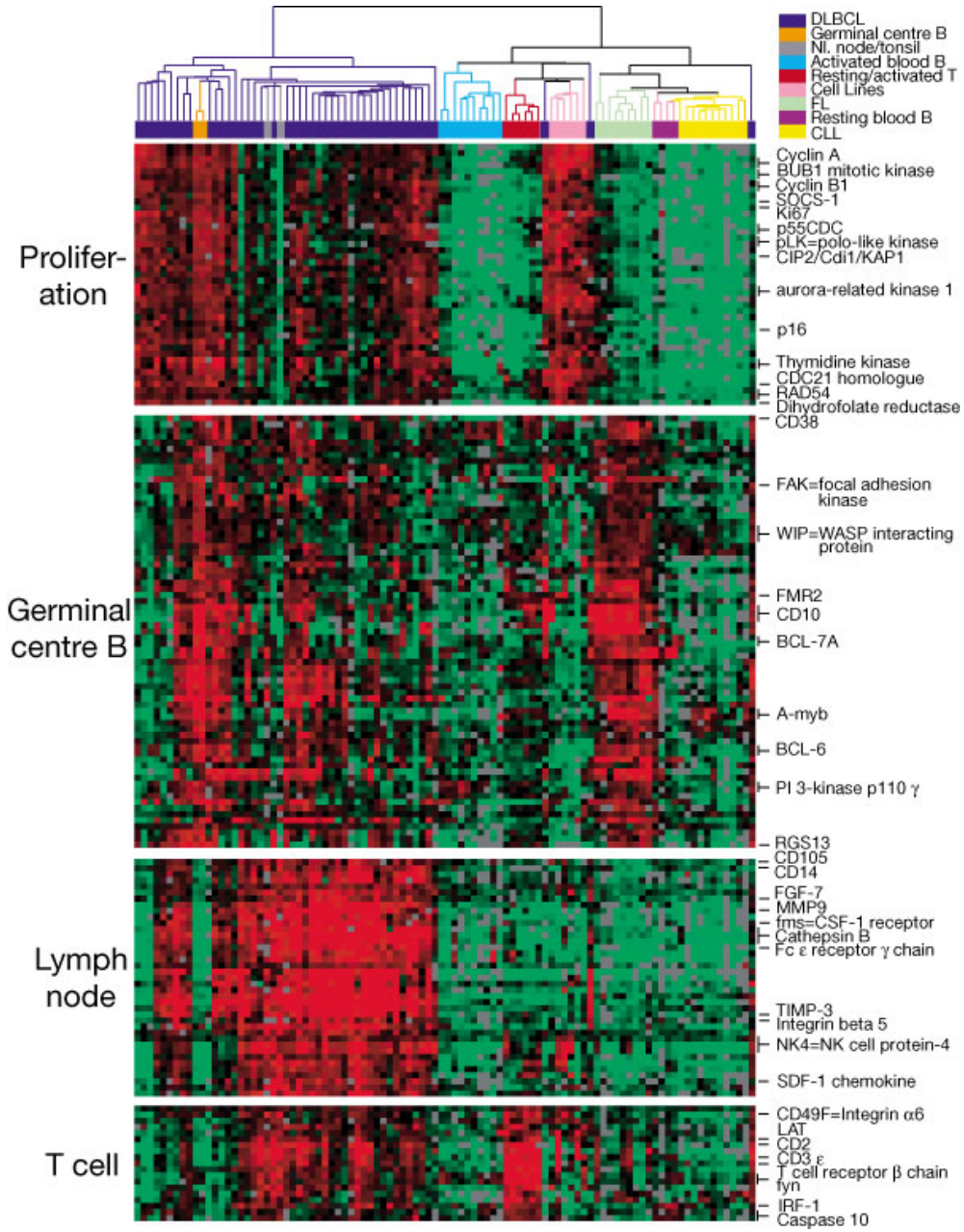
遺伝子側クラスタリング

1.データセット中に何種類の遺伝子発現パターンが含まれているか？



遺伝子側クラスタリング

2. 遺伝子Xはどの機能カテゴリーに属するか？

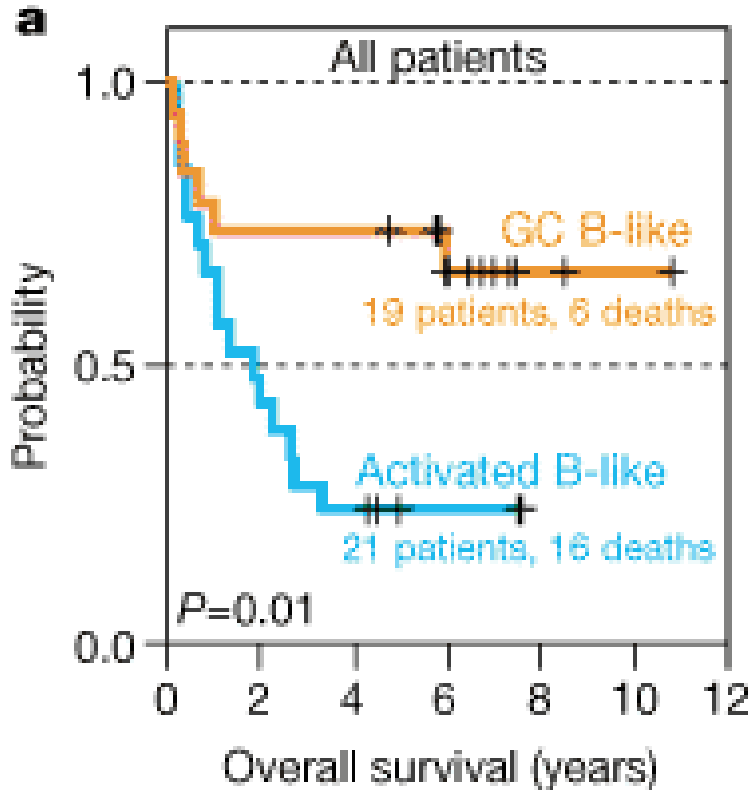


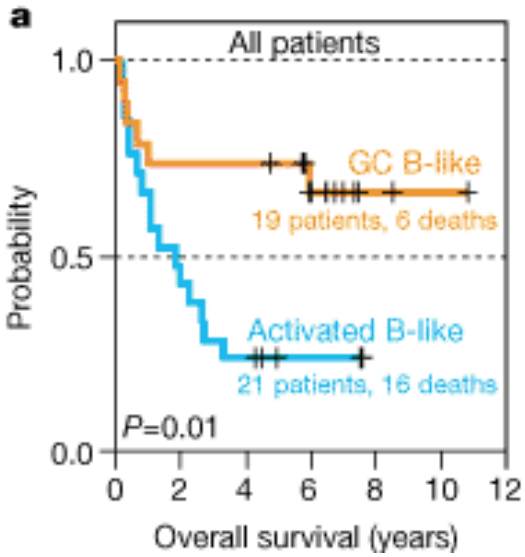
遺伝子側クラスタリング

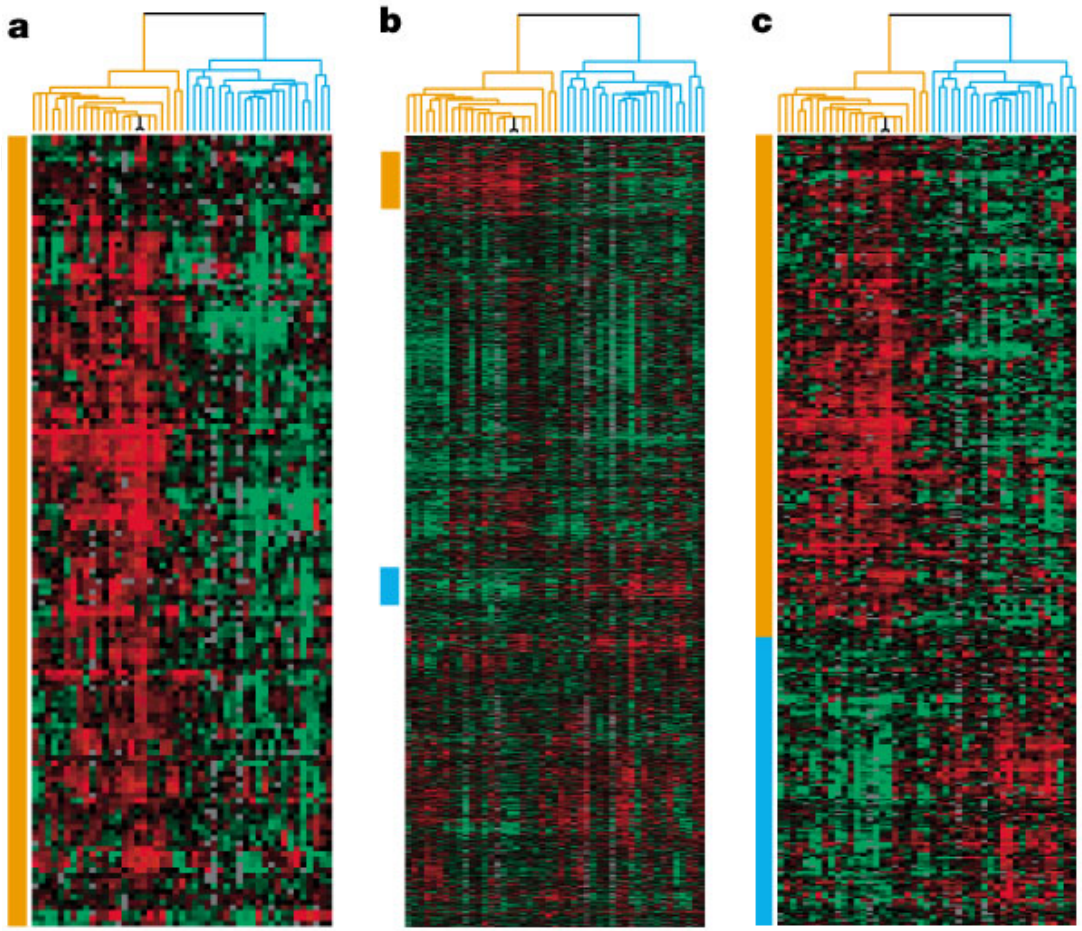
3. 機能未知の遺伝子群の発現パターンの中に、すでによく知られた遺伝子の発現パターンと似たものはあるか？（問題１）

細胞・組織(サンプル)側クラスタリング

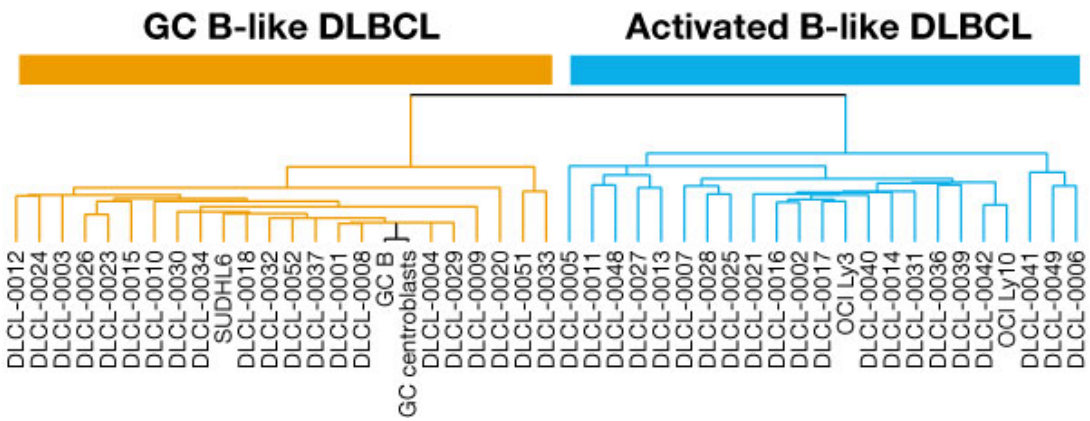
4. 疾患Xのサブタイプを組織の遺伝子発現パターンで認識、発見することができるか？







問題2.
Rで計算し、
DLBCLを分類しなさい



細胞・組織(サンプル)側
クラスタリング

5. 対象の組織サンプルはどの組織由来か？

問題3.
new.txtのアレイデータは、活性型のDLBCLか否か？

遺伝子相互作用ネットワーク

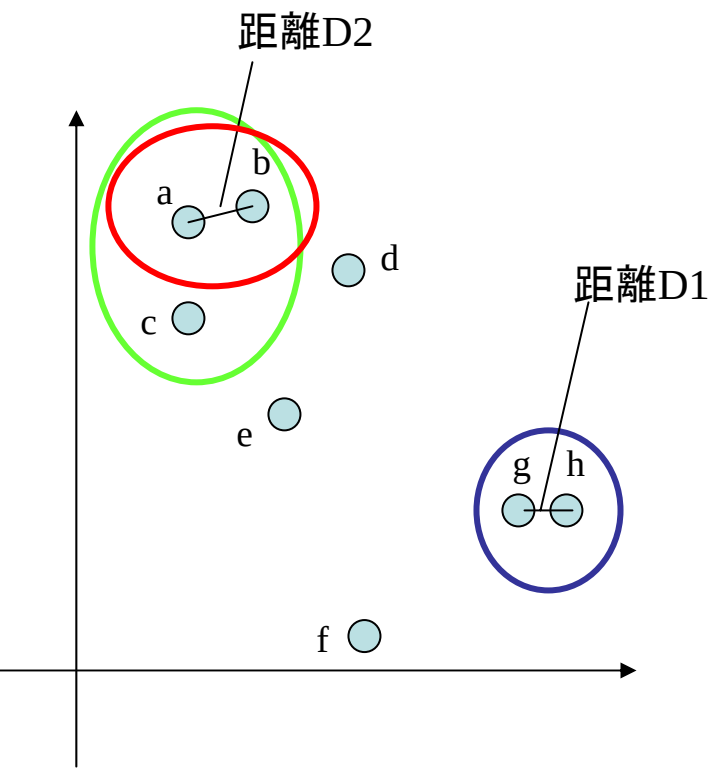
6. 対象の組織サンプルで観測される全ての遺伝子間の相互作用の違いは？
7. 発現パターンが類似した全ての遺伝子ペアを明らかにできるか？

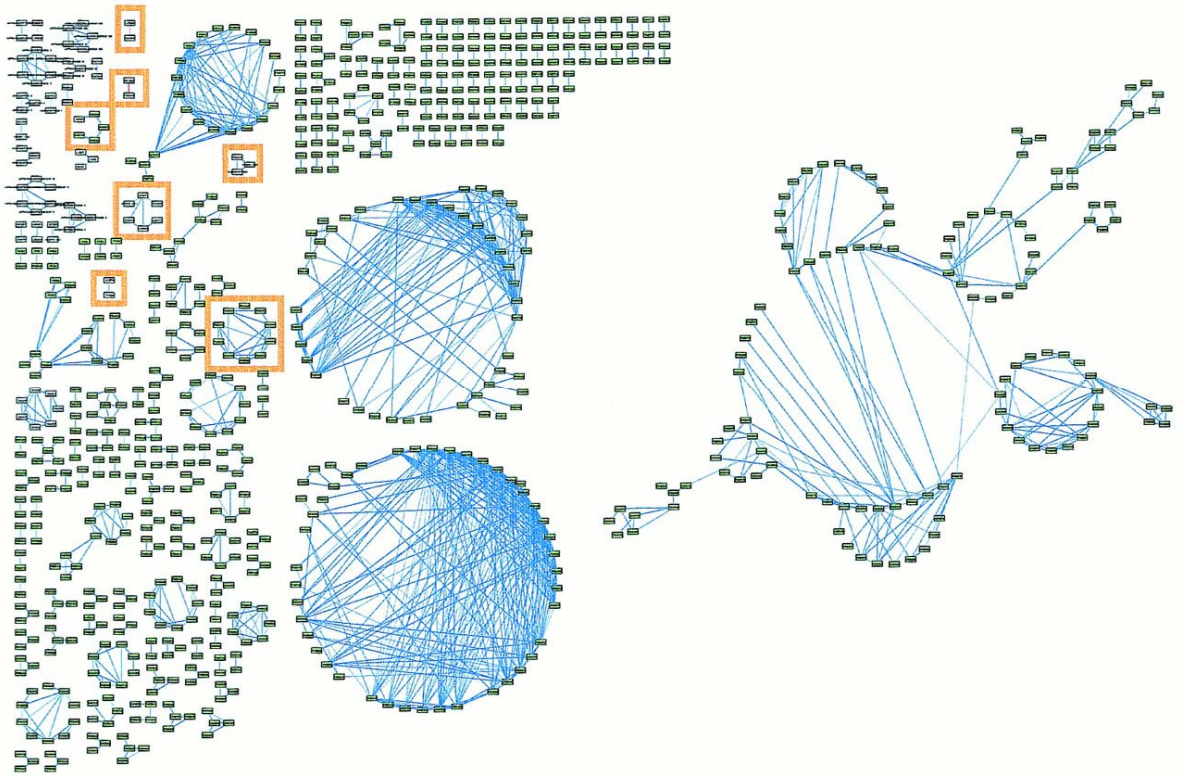
Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks

Atul J. Butte, Pablo Tamayo, Donna Slonim, Todd R. Golub, and Isaac S. Kohane

PNAS 2000 97: 12182-12186

遺伝子相互作用ネットワーク



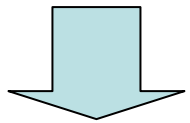


遺伝子ハンティング

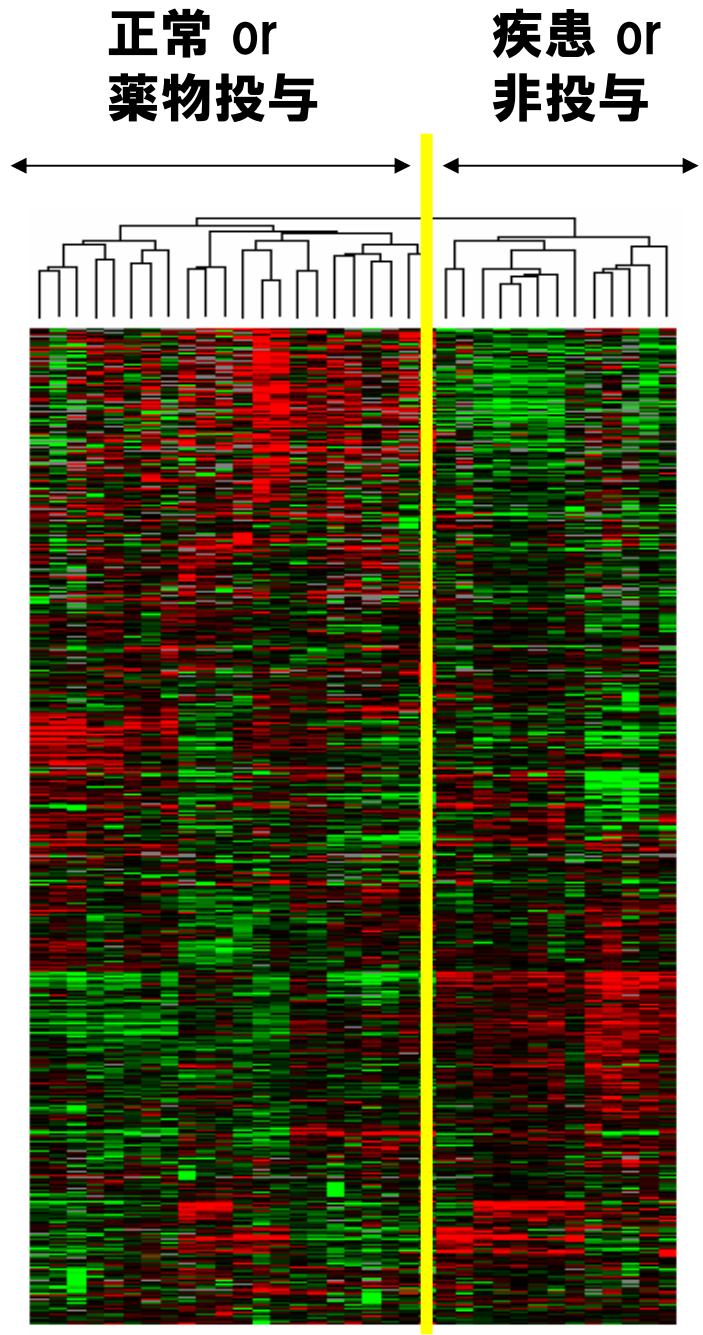
8. 正常と疾患など2群の組織サンプルを最もよく識別できる遺伝子群はどれか？
9. 薬物の影響を受けている遺伝子群は？
10. ある遺伝子Xの発現パターンは他の遺伝子群と比較してどれくらい特異か？
11. 医薬品のターゲットとなる遺伝子は？

8. 正常と疾患など2群の組織サンプルを最もよく識別できる遺伝子群はどれか？

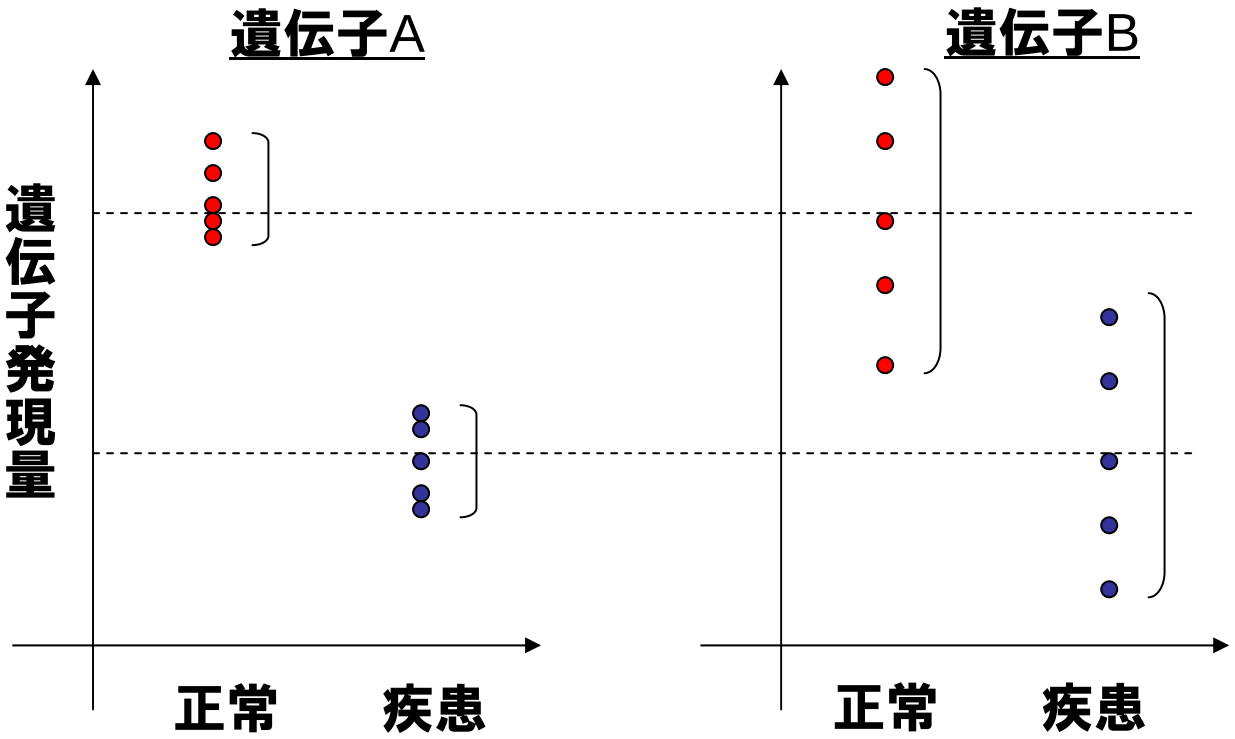
9. 薬物の影響を受けている遺伝子群は？（薬物投与群と非投与群において、発現変化する遺伝子群はどれか？）



2群の発現量の差が大きい遺伝子を探す



2群の発現量の差が大きい遺伝子を探す



・平均値の差は同じ
・分散に違いがある

正常と疾患を区別している遺伝子は？

2集団間に差があるかないかを統計的に調べる

t検定（平均値の差の検定）：p-valueで確率的有意性を評価

Rでt検定（問題４）

```
> x <- grep ("Active", colnames (dat) )
> x
 [1]  2  3  5  6  7 10 11 12 16 17 18 20 24 29
> dat [1,x]
     ActiveDLCL.0051 ActiveDLCL.0033 ActiveDLCL.0026 ActiveDLCL.0052 ActiveDLCL.0001 ActiveDLCL.0012
SMAD6          -0.432          -0.596          -1.086           0.08          -0.376           0.331
     ActiveDLCL.0004 ActiveDLCL.0023 ActiveDLCL.0010 ActiveDLCL.0008 ActiveDLCL.0015 ActiveDLCL.0003
SMAD6          -0.101          -0.528          -0.291          -0.114          -0.013           0.099
     ActiveDLCL.0018 ActiveDLCL.0029
SMAD6           0.099          -0.731
> dat [1,-x]
     DLCL.0027 DLCL.0002 DLCL.0014 DLCL.0011 DLCL.0006 DLCL.0025 DLCL.0031 DLCL.0016 DLCL.0007
SMAD6      -0.6     0.544    -0.009    -0.234     0.247     0.372    -0.157     0.083     0.632
     DLCL.0021 DLCL.0049 DLCL.0048 DLCL.0042 DLCL.0013 DLCL.0005
SMAD6     0.128     0.332     0.651     1.062     0.265    -0.609
> t.test (dat [1,x] , dat [1,-x] )

        Welch Two Sample t-test

data:  dat [1, x]  and dat [1, -x]
t = -2.7978, df = 26.743, p-value = 0.00942
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7659915 -0.1176561
sample estimates:
 mean of x  mean of y
-0.2613571  0.1804667
```

＊Active細胞の列番号を選択

＊Active細胞における1番目の遺伝子の発現量

＊non Active細胞における1番目の遺伝子の発現量

＊t.test関数

＊t検定の結果

問題4.

2群(GCとActivated B-like DLBCL)の発現量の差が大きい遺伝子を探す

複数の遺伝子についてt検定し、top20の遺伝子を決定

```
> dim (dat)
> trst <- 0
> for (i in 1:160) {
+ trst [i] <- t.test (dat [i,x] , dat [i,-x] ) $p.value
+ }
> trst <- as.matrix (trst)
> rownames (trst) <- rownames (dat)
> sorttrst <- trst [order (trst) ]
> sorttrst [1:20]
```

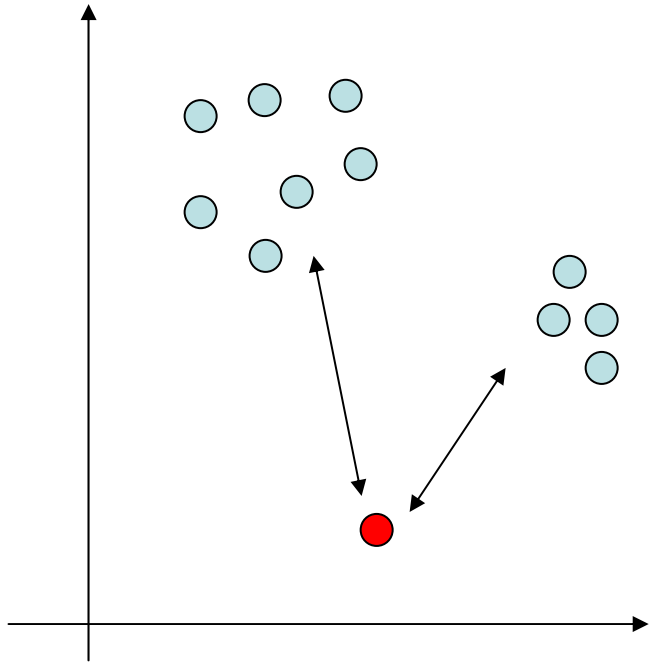
全ての遺伝子(160個)についてt.testを行う

p.valueの結果はtrstに入る

p.valueの値の大小で並べ替え

トップ20遺伝子

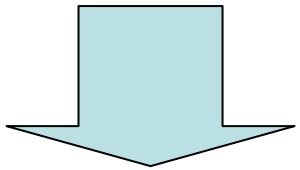
10. ある遺伝子Xの発現パターンは他の遺伝子群と比較してどれくらい特異か？



遺伝子の距離空間において孤立している遺伝子、つまり他の遺伝子から遠い距離にある遺伝子を探せばよい。
(ヒント：距離行列を評価する）

遺伝子ハンティング

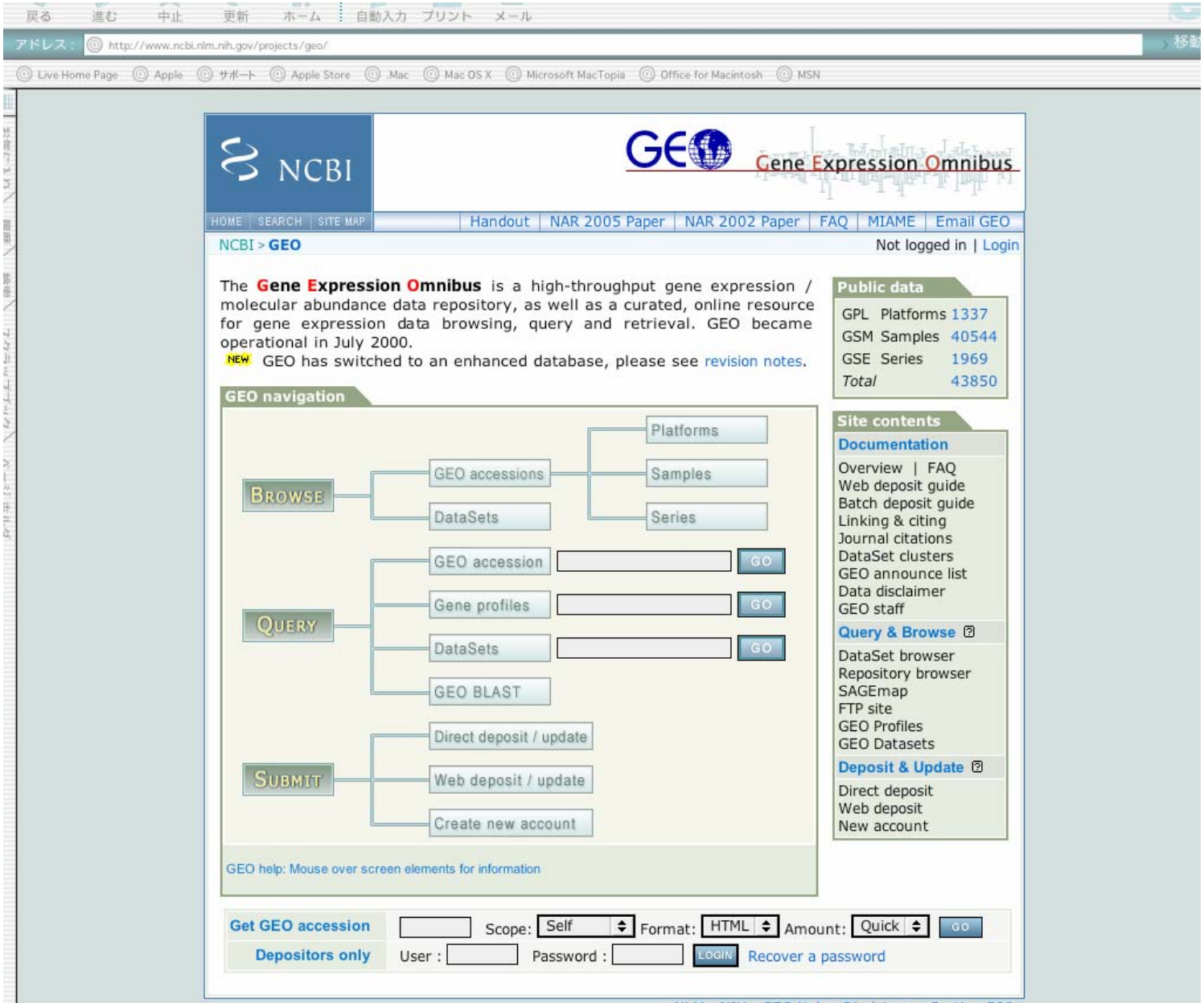
8. 正常と疾患など2群の組織サンプルを最もよく識別できる遺伝子群はどれか？
9. 薬物の影響を受けている遺伝子群は？
10. ある遺伝子Xの発現パターンは他の遺伝子群と比較してどれくらい特異か？（細胞・組織特異的な遺伝子は？）

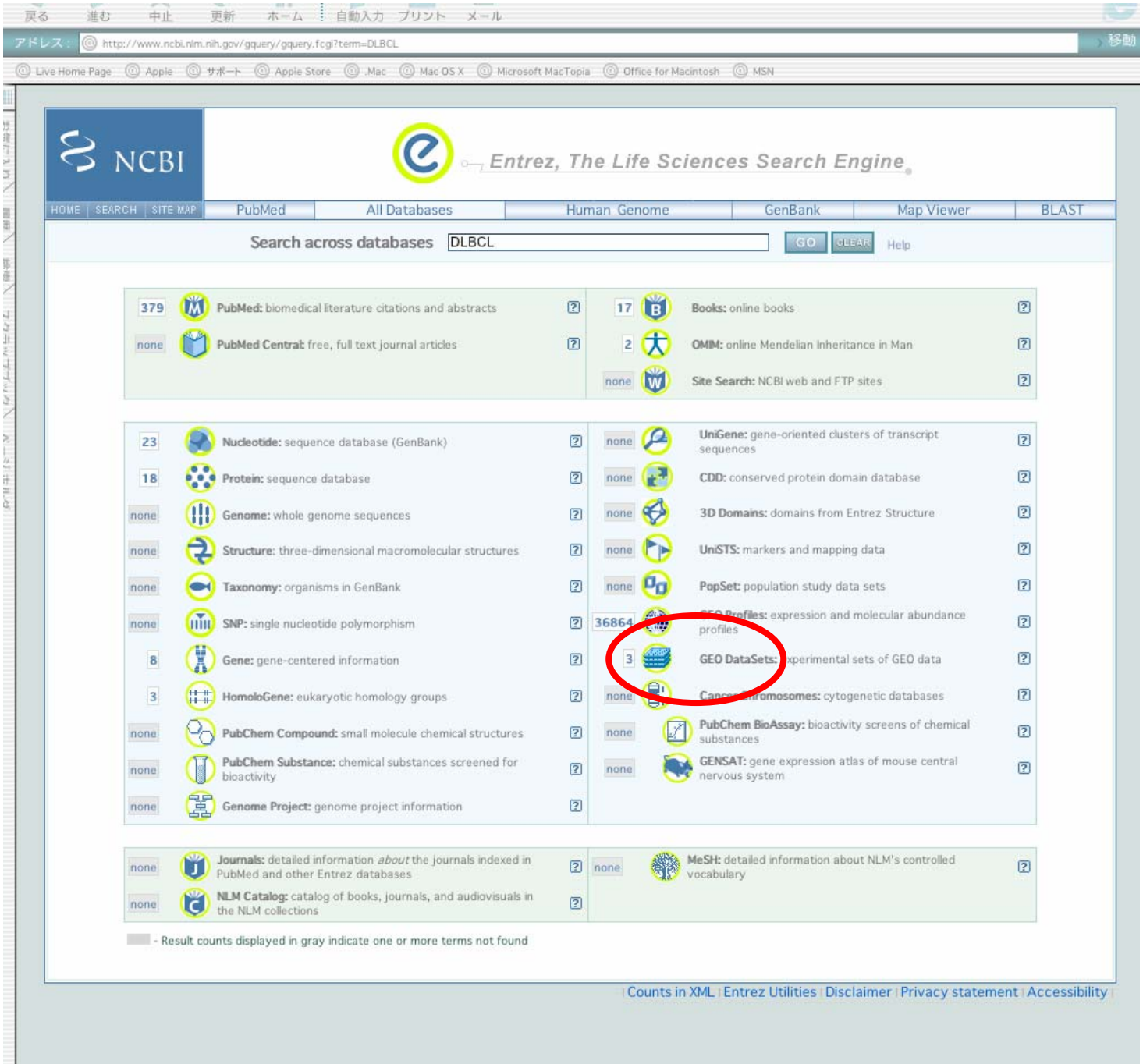


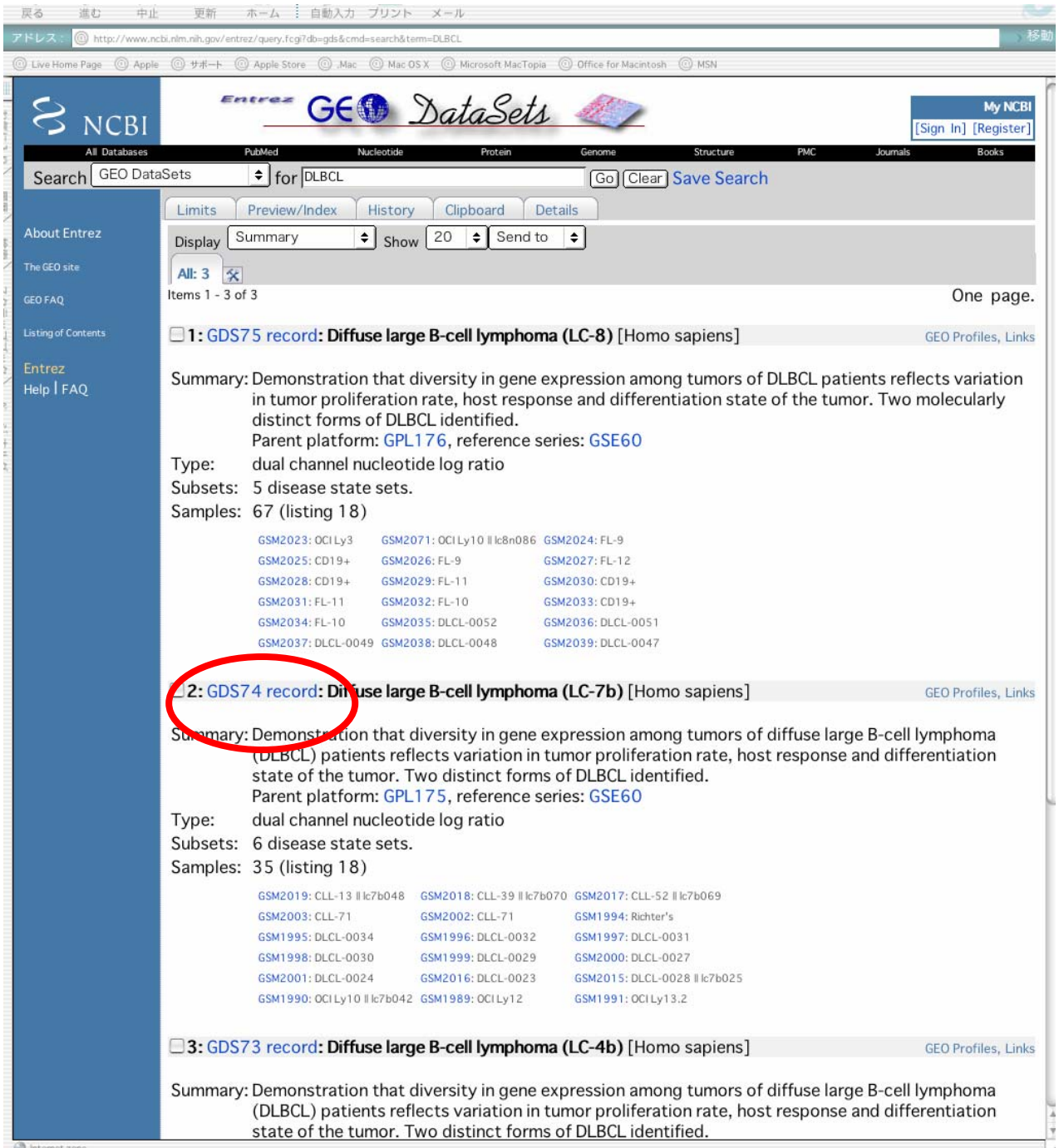
11. 医薬品のターゲットとなる遺伝子は？

マイクロアレイデータベース
NCBI/Gene expression omnibus (GEO)

遺伝子でも何でも良い
例えば、DLBCL


戻る 進む 中止 更新 ホーム 自動入力 プリント メール
アドレス: http://www.ncbi.nlm.nih.gov/projects/geo/
移動
Live Home Page Apple サポート Apple Store .Mac Mac OS X Microsoft MacTopia Office for Macintosh MSN
NCBI
Gene Expression Omnibus
HOME SEARCH SITE MAP
Handout NAR 2005 Paper NAR 2002 Paper FAQ MIAME Email GEO
NCBI > GEO
Not logged in | Login
The Gene Expression Omnibus is a high-throughput gene expression / molecular abundance data repository, as well as a curated, online resource for gene expression data browsing, query and retrieval. GEO became operational in July 2000.
NEW GEO has switched to an enhanced database, please see revision notes.
GEO navigation
BROWSE
GEO accessions
DataSets
Platforms
Samples
Series
QUERY
GEO accession
Gene profiles
DataSets
GEO BLAST
GO
SUBMIT
Direct deposit / update
Web deposit / update
Create new account
GEO help: Mouse over screen elements for information
Public data
GPL Platforms 1337
GSM Samples 40544
GSE Series 1969
Total 43850
Site contents
Documentation
Overview | FAQ
Web deposit guide
Batch deposit guide
Linking & citing
Journal citations
DataSet clusters
GEO announce list
Data disclaimer
GEO staff
Query & Browse
DataSet browser
Repository browser
SAGEmap
FTP site
GEO Profiles
GEO Datasets
Deposit & Update
Direct deposit
Web deposit
New account
Get GEO accession
Scope: Self
Format: HTML
Amount: Quick
GO
Depositors only
User :
Password :
LOGIN
Recover a password
NLM | NIH | GEO Help | Disclaimer | Section 508


戻る 進む 中止 更新 ホーム 自動入力 プリント メール
アドレス: http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi?term=DLBCL
移動
Live Home Page Apple サポート Apple Store Mac Mac OS X Microsoft MacTopia Office for Macintosh MSN
NCBI
Entrez, The Life Sciences Search Engine
HOME SEARCH SITE MAP
PubMed
All Databases
Human Genome
GenBank
Map Viewer
BLAST
Search across databases
DLBCL
GO
CLEAR
Help
379 PubMed: biomedical literature citations and abstracts
none PubMed Central: free, full text journal articles
17 Books: online books
2 OMIM: online Mendelian Inheritance in Man
none Site Search: NCBI web and FTP sites
23 Nucleotide: sequence database (GenBank)
18 Protein: sequence database
none Genome: whole genome sequences
none Structure: three-dimensional macromolecular structures
none Taxonomy: organisms in GenBank
none SNP: single nucleotide polymorphism
8 Gene: gene-centered information
3 HomoloGene: eukaryotic homology groups
none PubChem Compound: small molecule chemical structures
none PubChem Substance: chemical substances screened for bioactivity
none Genome Project: genome project information
none UniGene: gene-oriented clusters of transcript sequences
none CDD: conserved protein domain database
none 3D Domains: domains from Entrez Structure
none UniSTS: markers and mapping data
none PopSet: population study data sets
36864 profiles
3 GEO DataSets: experimental sets of GEO data
none PubChem BioAssay: bioactivity screens of chemical substances
none GENSAT: gene expression atlas of mouse central nervous system
none Journals: detailed information about the journals indexed in PubMed and other Entrez databases
none NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections
none MeSH: detailed information about NLM's controlled vocabulary
- Result counts displayed in gray indicate one or more terms not found
Counts in XML | Entrez Utilities | Disclaimer | Privacy statement | Accessibility


NCBI
Entrez GEO DataSets
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure PMC Journals Books
Search GEO DataSets for DLBCL Go Clear Save Search
Limits Preview/Index History Clipboard Details
About Entrez
The GEO site
GEO FAQ
Listing of Contents
Entrez
Help | FAQ
Display Summary Show 20 Send to
All: 3
Items 1 - 3 of 3
One page.
1: GDS75 record: Diffuse large B-cell lymphoma (LC-8) [Homo sapiens]
GEO Profiles, Links
Summary: Demonstration that diversity in gene expression among tumors of DLBCL patients reflects variation in tumor proliferation rate, host response and differentiation state of the tumor. Two molecularly distinct forms of DLBCL identified.
Parent platform: GPL176, reference series: GSE60
Type: dual channel nucleotide log ratio
Subsets: 5 disease state sets.
Samples: 67 (listing 18)
GSM2023: OCI Ly3 GSM2071: OCI Ly10 ǁ lc8n086 GSM2024: FL-9
GSM2025: CD19+ GSM2026: FL-9 GSM2027: FL-12
GSM2028: CD19+ GSM2029: FL-11 GSM2030: CD19+
GSM2031: FL-11 GSM2032: FL-10 GSM2033: CD19+
GSM2034: FL-10 GSM2035: DLCL-0052 GSM2036: DLCL-0051
GSM2037: DLCL-0049 GSM2038: DLCL-0048 GSM2039: DLCL-0047
2: GDS74 record: Diffuse large B-cell lymphoma (LC-7b) [Homo sapiens]
GEO Profiles, Links
Summary: Demonstration that diversity in gene expression among tumors of diffuse large B-cell lymphoma (DLBCL) patients reflects variation in tumor proliferation rate, host response and differentiation state of the tumor. Two distinct forms of DLBCL identified.
Parent platform: GPL175, reference series: GSE60
Type: dual channel nucleotide log ratio
Subsets: 6 disease state sets.
Samples: 35 (listing 18)
GSM2019: CLL-13 ǁ lc7b048 GSM2018: CLL-39 ǁ lc7b070 GSM2017: CLL-52 ǁ lc7b069
GSM2003: CLL-71 GSM2002: CLL-71 GSM1994: Richter's
GSM1995: DLCL-0034 GSM1996: DLCL-0032 GSM1997: DLCL-0031
GSM1998: DLCL-0030 GSM1999: DLCL-0029 GSM2000: DLCL-0027
GSM2001: DLCL-0024 GSM2016: DLCL-0023 GSM2015: DLCL-0028 ǁ lc7b025
GSM1990: OCI Ly10 ǁ lc7b042 GSM1989: OCI Ly12 GSM1991: OCI Ly13.2
3: GDS73 record: Diffuse large B-cell lymphoma (LC-4b) [Homo sapiens]
GEO Profiles, Links
Summary: Demonstration that diversity in gene expression among tumors of diffuse large B-cell lymphoma (DLBCL) patients reflects variation in tumor proliferation rate, host response and differentiation state of the tumor. Two distinct forms of DLBCL identified.

NCBI

DataSet Record

GEO Gene Expression Omnibus

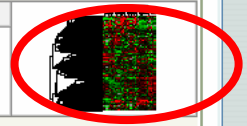
GDS Summary

Accession:	GDS74 ☞ View Expression Profiles (Entrez-GEO)
Title:	Diffuse large B-cell lymphoma (LC-7b)
Summary:	Demonstration that diversity in gene expression among tumors of diffuse large B-cell lymphoma (DLBCL) patients reflects variation in tumor proliferation rate, host response and differentiation state of the tumor. Two distinct forms of DLBCL identified.
Organism:	Homo sapiens
Platform:	GPL175: LC-7b
Experiment type:	dual channel nucleotide
No. of probes:	9216
Value type:	log ratio
Series:	GSE60
PubMed id:	10676951
Series published:	June 27 2002
Last GDS update:	April 04 2003

Subset and Sample Info

Sample selection

check all uncheck all toggle



6 assigned subsets			Two-tailed t-test (A vs B)	
Samples	Type	Description	A	B
(9)	disease state	diffuse large B-cell lymphoma	↔	
(5)	disease state	chronic lymphocytic leukemia	↔	
(3)	disease state	normal lymphoid subset	↔	
(5)	disease state	hematopoetic cell line	↔	
(2)	disease state	follicular lymphoma	↔	
(11)	disease state	activated blood B	↔	
	GDS74 only ranks values subset effects		√	Query A vs. B √

0.050 significance level

35 samples, order: none

GSM2019 : CLL-13 II lc7b048 src1: Lymphopool src2: CLL-13

GSM2018 : CLL-39 II lc7b070 src1: Lymphopool src2: CLL-39

GSM2017 : CLL-52 II lc7b069 src1: Lymphopool src2: CLL-52

GSM2003 : CLL-71 src1: Lymphopool src2: CLL-71

GSM2002 : CLL-71:Richter's

GSM1994 : DLCL-0034

GSM1995 : DLCL-0032

GSM1996 : DLCL-0031

戻る 進む 中止 更新 ホーム 自動入力 プリント メール

アドレス: http://www.ncbi.nlm.nih.gov/projects/geo/gds/analyze/analyze.cgi?ID=GDS74

Live Home Page Apple サポート Apple Store .Mac Mac OS X Microsoft MacTopia Office for Macintosh MSN

NCBI

DataSet Cluster Analysis

HOME SEARCH SITE MAP NCBI

NCBI > GEO > GDS Browser

- Move/resize box to select region of interest
- Double click or hit space bar in box to zoom in cluster region with annotation

Display Options		Clustering	
High expression level:	Red	Distance:	Uncentered Correlation
Low expression level:	Green	Hierarchical:	UPGMA

Change

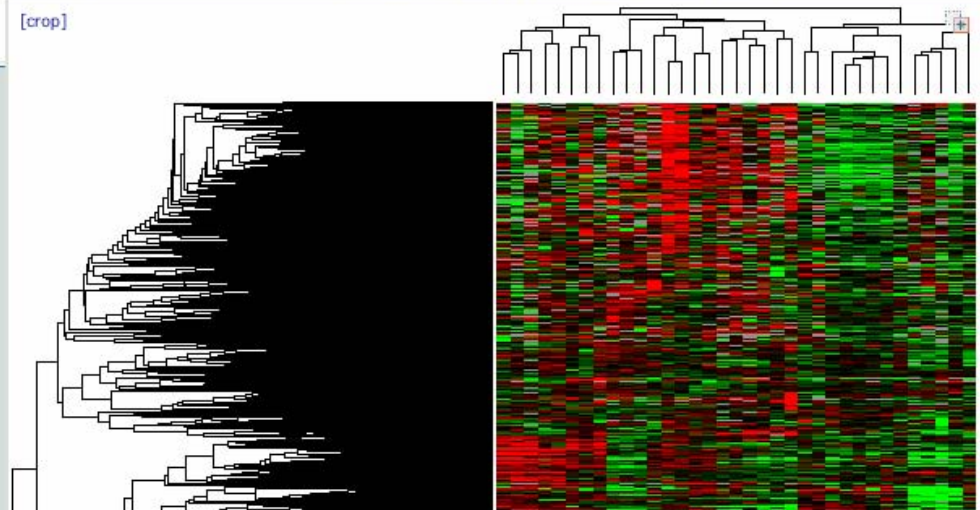
Get selected data | Plot selected gene profiles | Get profiles in Entrez-GEO

Stack selections

Full image : **9216** X **35** spots [Reset]

GDS74 Uncentered Correlation UPGMA - Diffuse large B-cell lymphoma (LC-7b) [Homo sapiens]

[crop]



GDS74 : log expression value of genes vs samples - Diffuse large B-cell lymphoma (LC-7b) [Homo sapiens].

Get profiles in Entrez-GEO Show heat map region Get selected data

戻る 進む 中止 更新 ホーム 自動入力 プリント メール

アドレス: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=geo&WebEnvRq=1&cmd=Select+from+History&query_key=3&WebEnv=0c7iVuEVM5n-RPXRelEaA-59OW95yL518A5jy8hbpD1vax8SE9LLAL@9@EV8I0F 移動

Live Home Page Apple サポート Apple Store .Mac Mac OS X Microsoft MacTopia Office for Macintosh MSN

NCBI

Entrez GEO Profiles

My NCBI
[Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure PMC Journals Books

Search GEO Profiles for GDS74[ACCN] AND ("H12639" OR "AA226782 Go Clear Save Search

Limits Preview/Index History Clipboard Details

About Entrez
The GEO site
GEO FAQ
Listing of Contents
Graph caption
Entrez
Help | FAQ

Display Summary Show 20 Subgroup effect Send to

All: 16

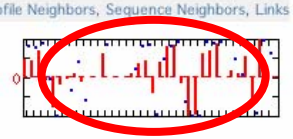
Items 1 - 16 of 16 One page.

1: GDS74 record | GPL175 8298 [Homo sapiens] 35 samples Profile Neighbors, Sequence Neighbors, Links

Annotation: BTK: Bruton agammaglobulinemia tyrosine kinase

Reporter: N75948 IMAGE:295208 (clone)

Experiment: Diffuse large B-cell lymphoma (LC-7b), dual channel nucleotide log ratio

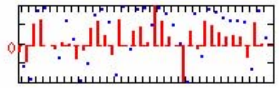


2: GDS74 record | GPL175 6034 [Homo sapiens] 35 samples Profile Neighbors, Sequence Neighbors, Links

Annotation: NFIL3: Nuclear factor, interleukin 3 regulated

Reporter: AA226782 IMAGE:663776 (clone)

Experiment: Diffuse large B-cell lymphoma (LC-7b), dual channel nucleotide log ratio

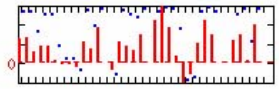


3: GDS74 record | GPL175 9163 [Homo sapiens] 35 samples Profile Neighbors, Links

Annotation: oa54g04.s1 NCI_CGAP_GCB1 Homo sapiens cDNA clone IMAGE:1308822 3' similar to SW:EXTN_TOBAC P13983 EXTENSIN PRECURSOR ;contains element MSR1 repetitive element ;, mRNA sequence

Reporter: AA748376 IMAGE:1308822 (clone)

Experiment: Diffuse large B-cell lymphoma (LC-7b), dual channel nucleotide log ratio

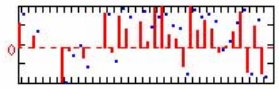


4: GDS74 record | GPL175 8266 [Homo sapiens] 35 samples Profile Neighbors, Sequence Neighbors, Links

Annotation: MAPK11: Mitogen-activated protein kinase 11

Reporter: T71113 IMAGE:84148 (clone)

Experiment: Diffuse large B-cell lymphoma (LC-7b), dual channel nucleotide log ratio

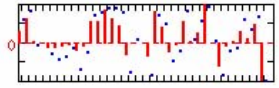


5: GDS74 record | GPL175 8833 [Homo sapiens] 35 samples Profile Neighbors, Sequence Neighbors, Homologs, Links

Annotation: LYN: V-yes-1 Yamaguchi sarcoma viral related oncogene homolog

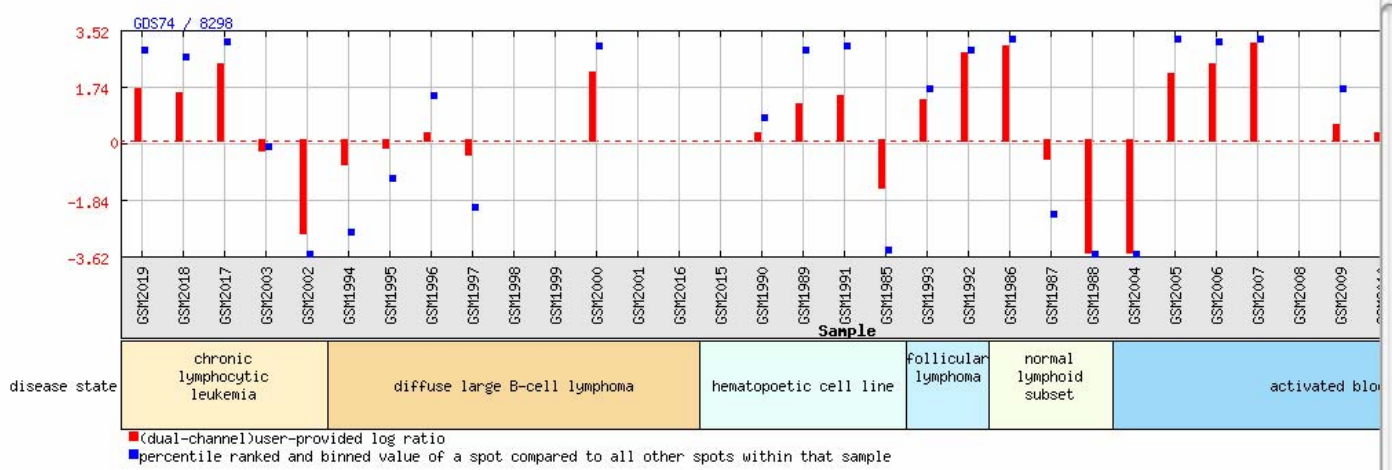
Reporter: R83836 IMAGE:193913 (clone)

Experiment: Diffuse large B-cell lymphoma (LC-7b), dual channel nucleotide log ratio



6: GDS74 record | GPL175 6000 [Homo sapiens] 35 samples Profile Neighbors, Sequence Neighbors, Links

Annotation: SULT1A1: Sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1



Graph caption help

GSM2019:CLL-13 || lc7b048
GSM2018:CLL-39 || lc7b070
GSM2017:CLL-52 || lc7b069
GSM2003:CLL-71
GSM2002:CLL-71;Richter's
GSM1994:DLCL-0034
GSM1995:DLCL-0032
GSM1996:DLCL-0031
GSM1997:DLCL-0030
GSM1998:DLCL-0029
GSM1999:DLCL-0027
GSM2000:DLCL-0024
GSM2001:DLCL-0023
GSM2016:DLCL-0028 || lc7b025
GSM2015:OCI Ly10 || lc7b042
GSM1990:OCI Ly12
GSM1989:OCI Ly13.2
GSM1991:Jurkat
GSM1985:WSU1
GSM1993:FL-5;CD19+
GSM1992:FL-6;CD19+
GSM1986:Tonsil Memory B
GSM1987:Tonsil GC Centroblasts
GSM1988:Tonsil GC B