Gene Expression-Based Precision Medicine

(From Tissue Sample to Biomarker-Based Clinical Trials)

Kao-Tai Tsai, Ph.D.
Celgene Corporation, NJ & JPHCOPH, GSU, GA

Presentation at the 2019 BASS at Charlotte, NC

October 23, 2019

Outline of Presentation

Paradigms of clinical studies - 1

- Classical clinical trial paradigms (as advocated by Peto & others):
  - large population with simple designs to investigate overall efficacy, e.g., the mega trials.
  - mostly for public health perspectives.
- Precision medicine paradigms:
  - the so-called "right treatment to right patients."
  - specifically, targeted treatment for selected patients based on certain clinical/genomics bio-markers.

Precision medicine: the procedures

General procedures for precision medicine treatments

- targeted treatment for selected patients based on certain clinical/genomic bio-markers.
- based on patient's blood/tissue sample, the genomic profile is created.
- markers thought to be related to targeted diseases are identified.
- with these markers and (usually) via statistical modeling, patient selection classifiers are constructed.
- based on the classifiers, companion diagnostic devices are designed to screen patients for clinical trials.

Steps from tissue samples to personalized medicine

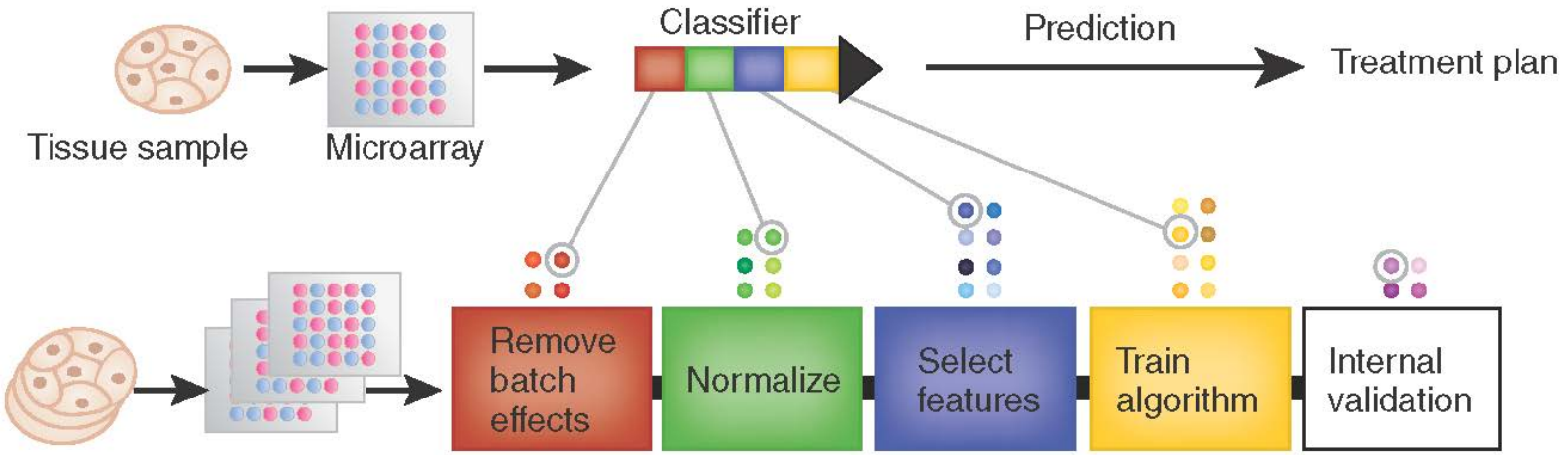
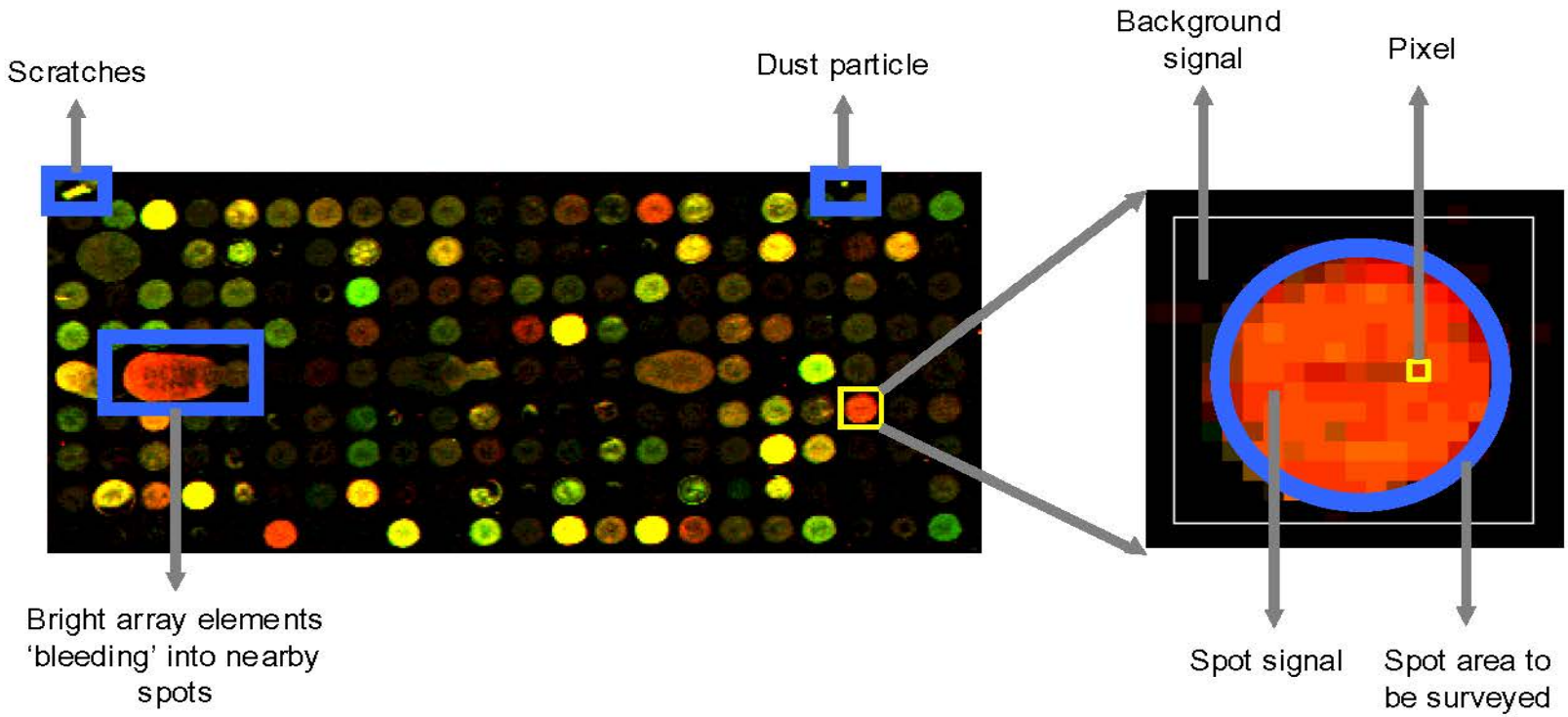


Figure 1: Major steps of gene-based research to personalized medicine

Precision medicine: data quality challenges - 1

General data quality difficulties encountered:

Figure 2: Example of microarray data issues



Precision medicine: data quality challenges - 2

- Example of microarray data issues:
  - The spot area and the background area are shown by a blue circle and a white box, respectively (right panel).
  - Any pixel within the blue circle is treated as a signal from the spot.
  - Pixels outside the blue circle but within the white box is treated as a signal from the background.
  - Images are not perfect, as it is often the case, which leads to many problems with spurious signals from dust particles, scratches, bright arrays, etc. (left panel).

Precision medicine: data quality challenges - 3

- tissue samples may come from various labs and the tissue handling procedures and microarray output quality are not always consistent (e.g., the batch effect).
- microarray signals are weak and biologically interesting results are usually obtained through the analysis of outliers but outliers are difficult to test rigorously.
- data in microarray need to be normalized (before and post normalization can be very different), but multiple normalization methods exist.
- in addition, signal measurement error can be high.

Precision medicine: data quality challenges - 1

Data quality:

- high correlation between genes or group of genes are commonly existent.
- some genes may have zero value (not expressed) for most of the sample with a few small non-zero values (weakly expressed).
- most statistically non-significant genes, such as the house-keeping genes, are not included in the models; however, these genes keep other genes function properly.
- outlying values may be existent, however, it is difficult to test with small sample size of patients.

Precision medicine: analytical challenges - 1

General analytical difficulties encountered:

- the number of patients with available gene profile is usually small (in tens or around hundreds).
- the number of genes in the patient profile is usually quite large (in thousands or tens of thousands).
- high correlation between genes or group of genes are commonly existent.
- the false discovery rate of testing gene effect on outcome is high due to large number (in thousands) of statistical test.
- significant findings in small sample of patients may not be easily generalized to other studies.
- cross-validate the models may be difficult due to unavailable or limited new data from other samples.

Precision medicine: analytical challenges - 2

- statistical modeling tools are mostly mechanical rarely incorporating the specific subject-matter-knowledge.
- *ad hoc* adjustments based on clinical expertise may be either unavailable or highly subjective.
- the genes identified as significant from various statistical modeling methodologies may be quite different.

Precision medicine: regulatory challenges - 1

General regulatory difficulties encountered:

- the FDA made three distinct classes of biomarkers:
  - Standard biomarkers, accepted and approved by the FDA to be used to determine patient eligibility and randomization for the trial. They need to have sufficient evidence from large samples of correct patient classification and sufficient validation.
  - Qualifying biomarkers, markers not yet approved by the FDA but show promise for determining patient eligibility or measuring treatment response.
  - Exploratory biomarkers, markers that are of interest on the basis of promising preliminary data suggesting predictive or prognostic value for cancer treatment.

contents page

Objectives of this presentation

- To describe and contrast some existing approaches used to construct the classifiers and the associated challenges. Generally, there are two major approaches:

  1. clinical decisions plus lots of multiple testing, e.g., Rosenwald et al. (NEJM), etc.
  2. regression-based statistical methodologies such as Lasso, elastic-net, SIS, Significance Analysis of Microarrays, or Tree-based model.

- To describe how the classifiers affect the creation of companion diagnostic devices.
- A few commonly used clinical design based on biomarkers.

contents page

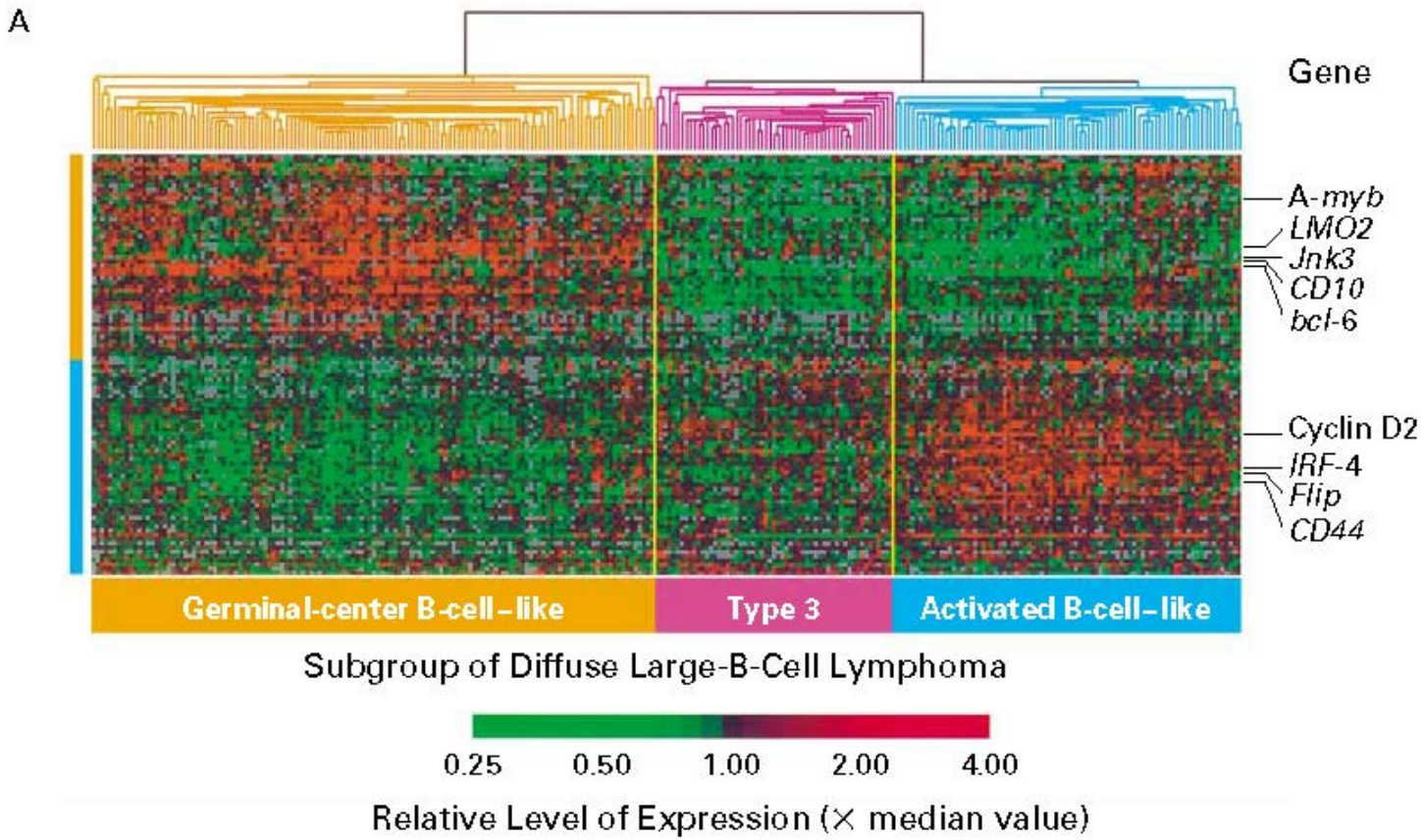
Procedures for classifier construction (clinical) - 1

Case Study of DLBCL:

- Diffuse large-B-cell lymphoma (DLBCL), can be cured by anthracycline-based chemotherapy in only 35 to 40% of patients.
- DLBCL actually comprises several diseases that differ in responsiveness to chemotherapy.
- Study of gene-expression profiles identified two subgroups of DLBCL (GCB & ABC) that had different outcomes after multi-agent chemotherapy.

Procedures for classifier construction (clinical) - 2

Figure 3: Relative Level of Expression



nano page

Procedures for classifier construction (clinical) - 3

Study Procedures:

- Tumor-biopsy specimens and clinical data were obtained retrospectively from 240 patients with untreated DLBCL.
- The microarrays were constructed from 12,196 clones of complementary DNA (i.e., microarray features) and were used to quantitate the expression of messenger RNA in the tumors.
- Data from 670 of 7399 microarray features were significantly associated with a good or a bad outcome.
- To classify the genes that were correlated with outcome, the authors used hierarchical clustering to group them into gene-expression signatures.

Procedures for classifier construction (clinical) - 4

- Among the 162 microarray features associated with a favorable outcome,
  - 15 belonged to the signature that characterizes GCB,
  - 30 were in the lymph-node signature of reactive nonmalignant cells, and
  - 35 were in the MHC class II signature.
- In the proliferation signature, 287 of 1333 microarray features were associated with a poor outcome.
- Since genes within the same gene-expression signature are probably associated with similar biologic aspects of a tumor, the authors combined the genes that were significantly associated with survival $(P < 0.01)$ within each signature.

Procedures for classifier construction (clinical) - 5

- To minimize the number of genes in the outcome predictor, the authors selected 16 genes that were highly variable in expression:
  - 3 germinal-center B-cell genes,
  - 4 MHC class II genes,
  - 6 lymph-node genes, and
  - 3 proliferation genes,

  and averaged the expression values for genes belonging to the same signature.

Procedures for classifier construction (clinical) - 6

TABLE 2. USE OF GENE-EXPRESSION PROFILES TO PREDICT OUTCOME FOR PATIENTS WITH DIFFUSE LARGE-B-CELL LYMPHOMA.*

GENE-EXPRESSION VARIABLE	NO. OF MICROARRAY FEATURES IN SIGNATURE (N=7399)	NO. OF MICROARRAY FEATURES SIGNIFICANTLY ASSOCIATED (P<0.01) WITH A GOOD OUTCOME (N=162)	NO. OF MICROARRAY FEATURES SIGNIFICANTLY ASSOCIATED (P<0.01) WITH A BAD OUTCOME (N=508)	REPRESENTATIVE GENES	GENBANK ACCESSION NO.	P VALUE			RELATIVE RISK OF DEATH AMONG ALL PATIENTS (95% CI)	OUTCOME
						PRELIMINARY GROUP	VALIDATION GROUP	ALL PATIENTS		
Germinal-center B-cell signature	151	15	0	*bcl-6* IMAGE 1334260† IMAGE 814622†	U00115 AA805575 AA236080	<0.001	0.009	<0.001	0.69 (0.59–0.81)	Favorable
MHC class II signature	37	35	0	HLA-DPα HLA-DQα HLA-DRα HLA-DRβ	X00457 X00452 K01171 M20430	<0.001	0.11	<0.001	0.69 (0.57–0.82)	Favorable
Lymph-node signature	357	30	2	α-Actinin Collagen type III α 1 Connective-tissue growth factor Fibronectin KIAA0233 Urokinase plasminogen activator	X15804 X14420 M54995 X02761 D87071 D00244	<0.001	0.04	<0.001	0.72 (0.62–0.85)	Favorable
Proliferation signature	1333	6	287	*c-myc* *E2IG3* *NPM3*	V00568 NM_014366 AF081280	<0.001	0.05	<0.001	1.63 (1.27–2.09)	Poor
Other	5521	76	219	*BMP6*	M60315	0.005	0.08	0.003	1.36 (1.11–1.65)	Poor

*P values indicate the significance of the association between the average value of the gene-expression variable and overall survival. Relative risk indicates the change in risk associated with a change by a factor of 2 in the average value for a given gene-expression signature. CI denotes confidence interval, and MHC major histocompatibility complex.

†IMAGE refers to an Integrated Molecular Analysis of Genomes and Their Expression consortium complementary DNA clone on the Lymphochip microarray.

Procedures for classifier construction (clinical) - 7

Remarks:

- The authors examined each of the 7399 microarray features (i.e., variables) to test for the significance w.r.t. survival and arrived at total of 670 significant variables.
- They showed (via permutation test) that there is a very small probability of false-positive even with such a big number (670) of variables selected.
- Based on clinical judgments, *ad hoc* decisions were made to reduce and aggregate these features into a few signatures so that a simple model can be built for practical clinical implementations.

Procedures for classifier construction (clinical) - 8

- The final "formula" is a Cox survival model:

```
Outcome-predictor-score =
= (0.241 x average of the proliferation signature)
+ (0.310 x value of BMP6)
- (0.290 x average of the GCB signature)
- (0.311 x average of the MHC class II signature)
- (0.249 x average of the lymph-node signature).
```

Procedures for classifier construction (statistical) - 1

- Most common practices are through regressions with distributional assumptions, mathematical derivation, and to find relationship (linear or non-linear) between input variables and outcomes.
- There are two general goals :
  - estimation of the relationship between inputs and outcomes with parsimonious models.
  - prediction of the future outcomes as accurate as possible with any possible models.

Procedures for classifier construction (statistical) - 2

- In the general least squares fitting procedure to estimate $\boldsymbol{\beta}$, one would use the the value of parameters which minimize the RSS of the following equation:

$$RSS_{\text{ls}} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2. \quad (1)$$

- LS fitting work fine for most of cases if sample cases are more than the unknown parameters in Eqn (1).
- In the construction of patient classifiers using molecular profiles, LS fitting does not always work correctly because the sample cases (n) are usually much less than the number of genes (p) as mentioned previously. Hence further assumptions are required to derive a parsimonious model for selection purposes.

Procedures for classifier construction (statistical) - 3

- One usually assumes the "sparsity condition" of the relationship, namely, only a small number of genes are needed in the model (hence most of the β's are zero in Eqn (1)), and uses a penalized regression to estimate the parameters.

In the following, we will discuss a few methods to analyze this kind of data. Generally, these methods involves regression shrinkage to derive a parsimonious model (except for Ridge regression).

Ridge Regression (Hoerl and Kennard (1988)) - 1

- Ridge regression coefficient estimates are the values that minimize

$$RSS_{\text{ridge}} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \qquad (2)$$

  where $\lambda \geq 0$ is a tuning parameter, to be determined separately.
- The quantity $\lambda\sum_{j=1}^{p}\beta_j^2$ is called a shrinkage penalty.
- When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates.
- However, as $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.
- Ridge regression will include all p predictors in the final model.

The Lasso regression (Tibshirani (1996)) - 1

- The lasso coefficients minimize the quantity

$$RSS_{\text{Lasso}} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| \qquad (3)$$

  where $\lambda \geq 0$ is a tuning parameter to be determined separately.
- Lasso does both continuous shrinkage and variable selection simultaneously.
- To perform prediction using the Lasso, one often uses a cross-validation (CV) scheme, to select a reasonable tuning parameter λ minimizing the cross-validated squared error risk.
- Asymptotic results conclude that the estimates based on Lasso are consistent.

The Lasso regression (Tibshirani (1996)) - 2

- Some limitations of Lasso:
  1. In the $p > n$ case, the lasso selects at most n variables before it saturates.
  2. For a group of variables with high pairwise correlations the lasso tends to select only one variable from the group.
  3. For usual $n > p$ situations, if there are high correlations between predictors, it has empirically observed that the prediction performance of the lasso is dominated by ridge regression.

The elastic-Net (Zou and Hastie (2005)) - 1

- For any $\lambda_1 > 0, \lambda_2 > 0$, they defined the "naive elastic net" as

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2. \quad (4)$$

  the estimator $\hat{\beta}$ minimizes equation (4).
- This procedure is also a penalized least squares method. When $\lambda_1 = 0$, it becomes ridge regression, and lasso if $\lambda_2 = 0$.
- However, the naive elastic net does not perform satisfactorily unless it is very close to ridge regression or the lasso in the regression prediction setting.
- To improve the performance of naive elastic net, Zou and Hastie fine-tuned it to a better procedure, elastic net, using the following steps.

The elastic-Net (Zou and Hastie (2005)) - 2

- Given the data (y, X) and penalty parameters (λ_1, λ_2), and define an augmented data set $(\mathbf{y}^*, \mathbf{X}^*)$ by

$$\mathbf{X}^*_{(n+p)\times p} = \sqrt{(1+\lambda_2)}\begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda_2}\mathbf{I}\end{pmatrix}, \quad \text{and} \quad \mathbf{y}^*_{(n+p)} = \begin{pmatrix}\mathbf{y}\\ 0\end{pmatrix}.$$

- Based on the augmented data, if $\hat{\beta}^*$ is the solution of the following lasso-type problem, namely,

$$\hat{\beta}^* = \min_{\beta} |y^* - X^*\beta^*|^2 + \frac{\lambda_1}{\sqrt{(1+\lambda_2)}}|\beta^*|_1,$$

then the elastic net estimator $\hat{\beta}$ can be expressed as

$$\hat{\beta}_{\text{(elastic net)}} = \sqrt{1+\lambda_2}\hat{\beta}^* = (1+\lambda_2)\hat{\beta}_{\text{(naive elastic net)}}.$$

The elastic-Net (Zou and Hastie (2005)) - 3

Remarks:

- Zou & Hastie showed that the naive elastic net problem can be transformed into an equivalent lasso problem using the augmented data.
- Since the sample size in the augmented problem is $n + p$ and $\mathbf{X}^*$ has rank p, the elastic net can potentially select all p predictors in all situations.
- The elastic net can perform an automatic variable selection in a fashion similar to the lasso.
- The elastic net has the "grouping effect" when some covariates are highly correlated.

The Sure Screening Method (Fan ans Li (2008)) - 1

- One of the approaches to handle the high dimensional data is through dimension reduction.
- With dimensions reduced from high to low, the computational burden can be reduced drastically, and accurate estimation can be obtained by using well-developed lower dimensional method.
- The method of sure screening is based on correlation learning to filter out the features that have weak correlation with the response.

The Sure Screening Method (Fan ans Li (2008)) - 2

- Let $W = (w_1, \cdots, w_p)^T$ be a p-vector that is obtained by componentwise regression, i.e.

$$W = X^T y,$$

where the $n \times p$ data matrix X is first standardized columnwise.

- Assume a sparsity model, for any given $\gamma \in (0, 1)$, sort the p componentwise magnitudes of the vector w in a decreasing order and define a submodel

$$M_\gamma = \{1 \leq i \leq p : |w_i| \text{ is the first } [\gamma n] \text{ largest } |w_i|\}, \quad (5)$$

where $[\gamma n]$ denotes the integer part of γn.

The Sure Screening Method (Fan ans Li (2008)) - 3

- This is a straightforward way to shrink the full model $\{1, \cdots, p\}$ down to a submodel M_γ with size $d = [\gamma n] < n$.
- Such correlation learning ranks the importance of features according to their marginal correlation with the response variable and filters out those that have weak marginal correlations with the response variable.
- *Note:* the componentwise regression estimator is a specific case of ridge regression with regularization parameter $\lambda = \infty$, namely, it makes the resulting estimator as little noisy as possible.

The Sure Screening Method (Fan ans Li (2008)) - 4

- **Suggestion for data analysis:** For the problem of ultra-high dimensional variable selection,
  - first to apply a sure screening method such as SIS to reduce the dimensionality from p to a relatively large scale d, say, below sample size n.
  - then use a lower dimensional model selection method such as SCAD, the Dantzig selector, lasso, or adaptive lasso.

Sig. Analysis of Microarrays (Tusher et al. (2001)) - 1

- Compute a test statistic t_j for each gene j and define order statistics $t_{(j)}$ such that $t_{(1)} \geq t_{(2)} \geq \cdots \geq t_{(m)}$.
- Perform B permutations of the responses/covariates $y_1, \cdots, y_n$.
- For each permutation b compute the test statistics $t_{j,b}$ and the corresponding order statistics $t_{(1),b} \geq t_{(2),b}, \geq \cdots, \geq t_{(m),b}$. Note that $t_{(j),b}$ may correspond to a different gene than the observed $t_{(j)}$.
- From the B permutations, estimate the expected value (under the complete null) of the order statistics by

$$\bar{t}_{(j)} = (1/B) \sum_{b}^{B} t_{(j),b}.$$

Sig. Analysis of Microarrays (Tusher et al. (2001)) - 2

- Form a quantile–quantile (Q–Q) plot of the observed $t_{(j)}$ versus the expected $\bar{t}_{(j)}$.
- For a fixed threshold Δ, let

$$
\begin{aligned}
j_0 &= \max\{j : \bar{t}_{(j)} \geq 0\}, \\
j_1 &= \max\{j \leq j_0 : t_{(j)} - \bar{t}_{(j)} \geq \Delta\}, \\
j_2 &= \min\{j > j_0 : t_{(j)} - \bar{t}_{(j)} \leq -\Delta\}.
\end{aligned} \tag{6}
$$

- All genes with $j \leq j_1$ are called significant positive and all genes with $j \geq j_2$ are called significant negative.

Sig. Analysis of Microarrays (Tusher et al. (2001)) - 3

- Define the upper and lower cut points,

$$\begin{aligned} \text{cut}_{up}(\Delta) &= \min\{t_{(j)} : j \leq j_1\} = t_{(j_1)}, \\ \text{cut}_{low}(\Delta) &= \max\{t_{(j)} : j \geq j_2\} = t_{(j_2)}. \end{aligned} \tag{7}$$

  If no such j_1 or j_2 exists, set $\text{cut}_{up}(\Delta) = \infty$ and $\text{cut}_{low}(\Delta) = -\infty$.

- Note:
  - For a given threshold Δ, the expected number of false positives is estimated by computing for each of the B permutations the number of genes with $t_{j,b}$ above $\text{cut}_{up}(\Delta)$ or below $\text{cut}_{low}(\Delta)$, and averaging this number over permutations.
  - A threshold Δ is chosen to control the expected number of false positives under the complete null, at an acceptable nominal level.

Data: TARGET (from NCI website)

	TargetID	Gender	AgeDays	EFSdays	VitalStatus	Osdays	WBCDiag	InitialTherapy	WHOALALClass
1	SJMPAL011914	Male	2613	2813	Alive	2813	29.0	AML	T/M
2	SJMPAL012419	Male	5973	1229	Alive	1229	43.0	AML	B/M
3	SJMPAL012421	Female	4484	1718	Alive	1718	79.4	AML	T/M
4	SJMPAL016342	Male	335	295	Dead	295	4.2		T/M
5	SJMPAL017975	Male	6527	2927	Alive	2927	14.9	AML	B/M
6	SJMPAL019076	Male	2017	NaN	Alive	5297	2.8	ALL	B/M
7	SJMPAL040025	Female	3319	2130	Alive	2130	14.8	ALL	B/M
8	SJMPAL040028	Female	6265	NaN	Alive	3182	43.2	hybrid	B/M
10	SJMPAL040037	Male	5594	963	Alive	963	9.3	ALL	T/M
...									
20	SJMPAL042794	Male	5541	25	Dead	25	289.0	hybrid	T/M
21	SJMPAL042798	Male	1751	443	Dead	443	33.9	AML	NOS (T/B/M)
22	SJMPAL042799	Male	4090	486	Dead	486	335.2	AML	NOS (T/B/M)
23	SJMPAL042801	Male	1963	1169	Alive	1169	32.0	AML	T/M
24	SJMPAL042946	Female	5588	320	Dead	320	22.3	ALL	NOS (T/B)
25	SJMPAL043505	Male	1270	1562	Dead	1562	1.2	AML	T/M
26	SJMPAL043506	Male	2218	NaN	Dead	769	NaN		B/M
28	SJMPAL043508	Female	3505	1806	Dead	1806	9.6	AML	B/M
30	SJMPAL043512	Female	5664	1097	Alive	1097	5.0	ALL	B/M
34	SJMPAL043769	Female	5752	572	Dead	572	32.7	AML	AUL
35	SJMPAL043770	Male	2716	662	Alive	662	62.0	AML	AUL
36	SJMPAL043771	Female	3542	780	Alive	780	70.8	ALL	B/M
37	SJMPAL043772	Male	1796	NaN	Dead	771	0.5		T/M
38	SJMPAL043773	Female	2743	394	Dead	394	35.0	AML	T/M
39	SJMPAL043774	Male	12	156	Dead	156	67.0	ALL	MLL
40	SJMPAL043775	Male	699	838	Alive	838	41.0	ALL	Ph+

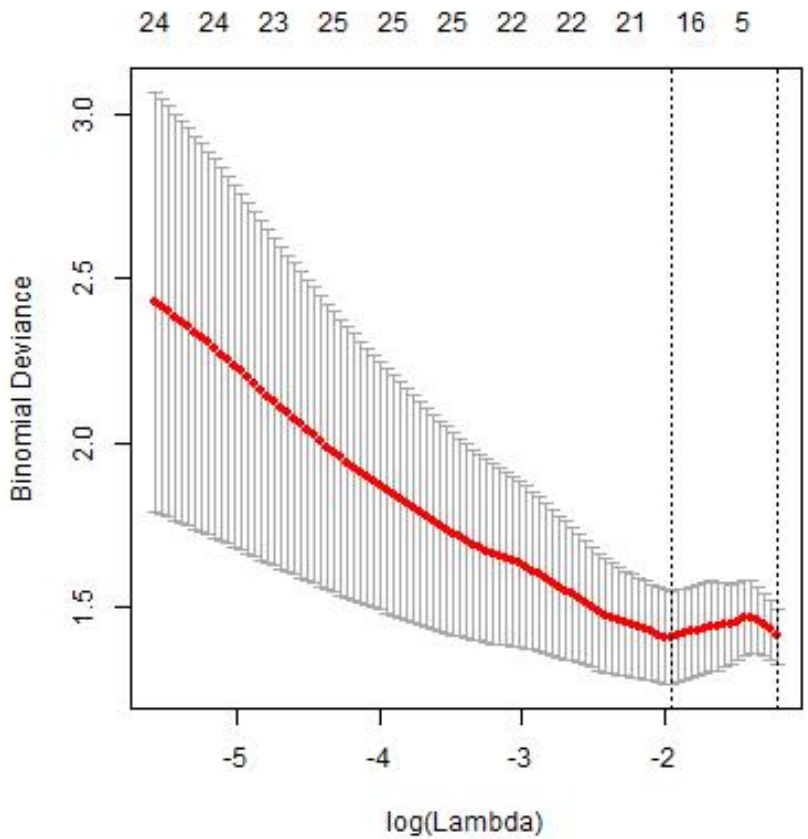
Genes selected (Lasso) - 1

```
Call: glmnet(x = x, y = y, family = c("binomial"), alpha = 1, nlambda = 100)
[1] "fit$nobs" = 33
[1] "Best lambda" = 0.1422173

[1] "Genes selected by Lasso:"
                  [,1]
Gene565   1.417241e+01
Gene3207  1.848242e-01
Gene3224  1.600571e+00
Gene13756 5.226362e-02
Gene31966 2.321998e-01
Gene33435 1.143997e-02
Gene33641 8.407966e-04
Gene35780 2.436470e-03
Gene40857 1.443125e-02
Gene45591 1.740618e-02
```

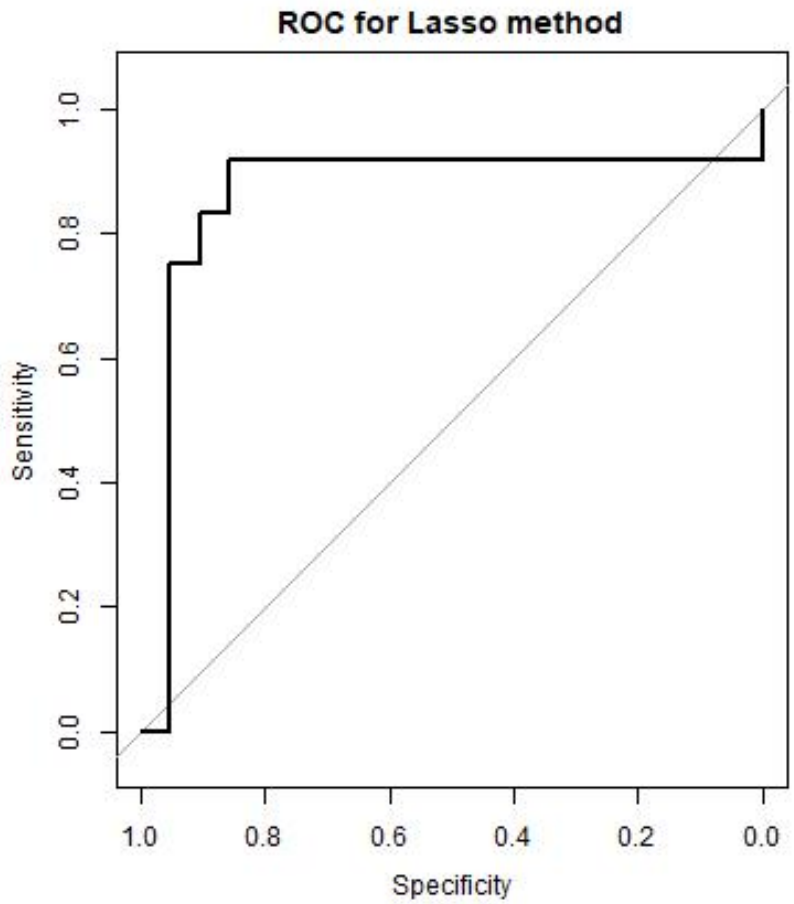
Genes selected (Lasso) - 2

```
[1] "Best lambda" = 0.2161576

[1] "Genes selected by Lasso:"
                  [,1]
Gene31966 0.159299749
Gene33435 0.005406323
Gene40857 0.015507189
Gene45591 0.005127338
```

Cross-validation of Lasso model



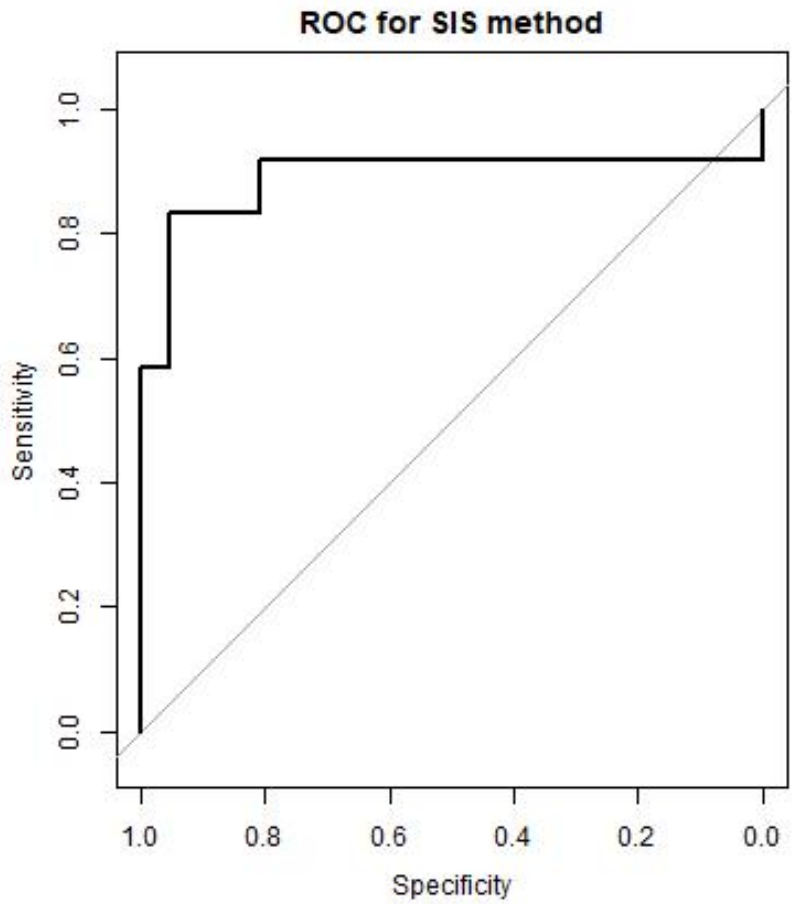
ROC for Lasso prediction



Genes selected (SIS) - 1

```
[1] Selected genes and the estimated coefficients"

Gene31966    1.9192385
Gene33435    0.0965927
```

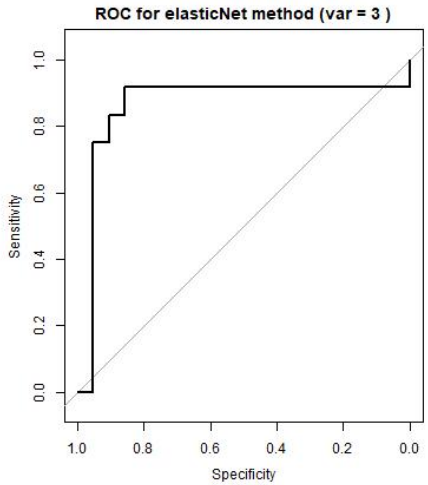
ROC for SIS method

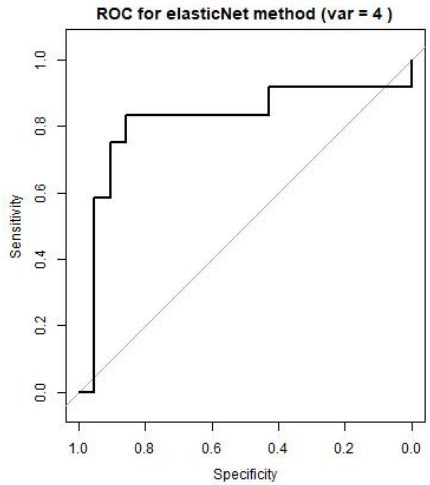


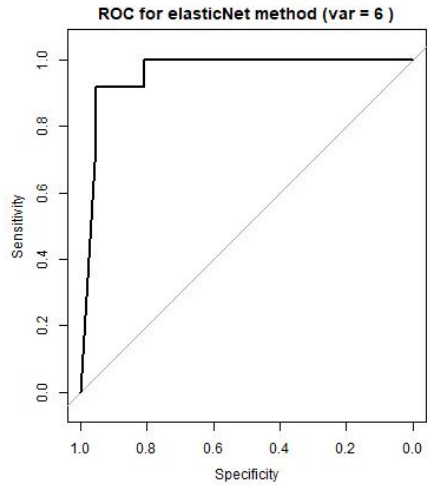
Genes selected (elastic-net) - 1

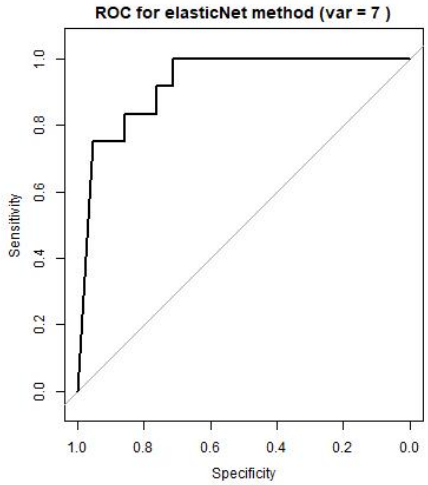
```
Call: enet(x = x, y = y, lambda = 10, max.steps = 50)

Sequence of  moves:
Gene33435 Gene31966 Gene40857 Gene48345 Gene32677
Gene43496 Gene43770 Gene40760 Gene34561 Gene27311
Gene47598 Gene38005 Gene38090 Gene33641 Gene42912
Gene45591 Gene27451 Gene31647 Gene2363  Gene3654
Gene26047 Gene40402 Gene1173  Gene7663  Gene7302
Gene28436 Gene27896 Gene21149 Gene1084  Gene3207
Gene23952 Gene13756 Gene35780 Gene43948 Gene5794
Gene30850 Gene25317 Gene50215 Gene3214  Gene15621
Gene34893 Gene26695 Gene42841 Gene14568 Gene43772
Gene7185  Gene1834  Gene21913 Gene57570 Gene48706
```

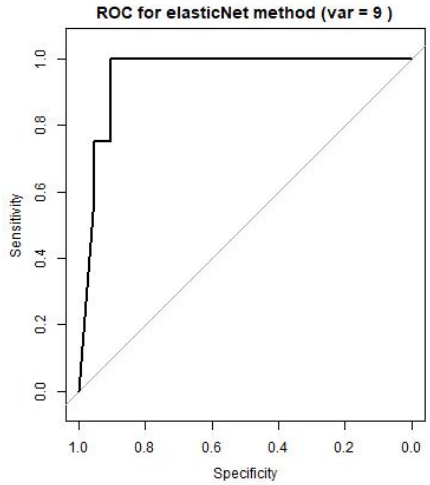
Elastic-Net with various number of terms in the model

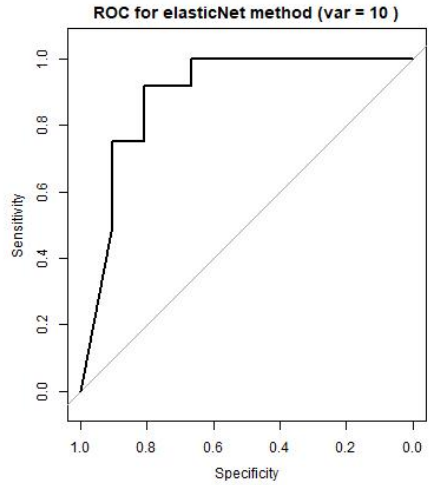








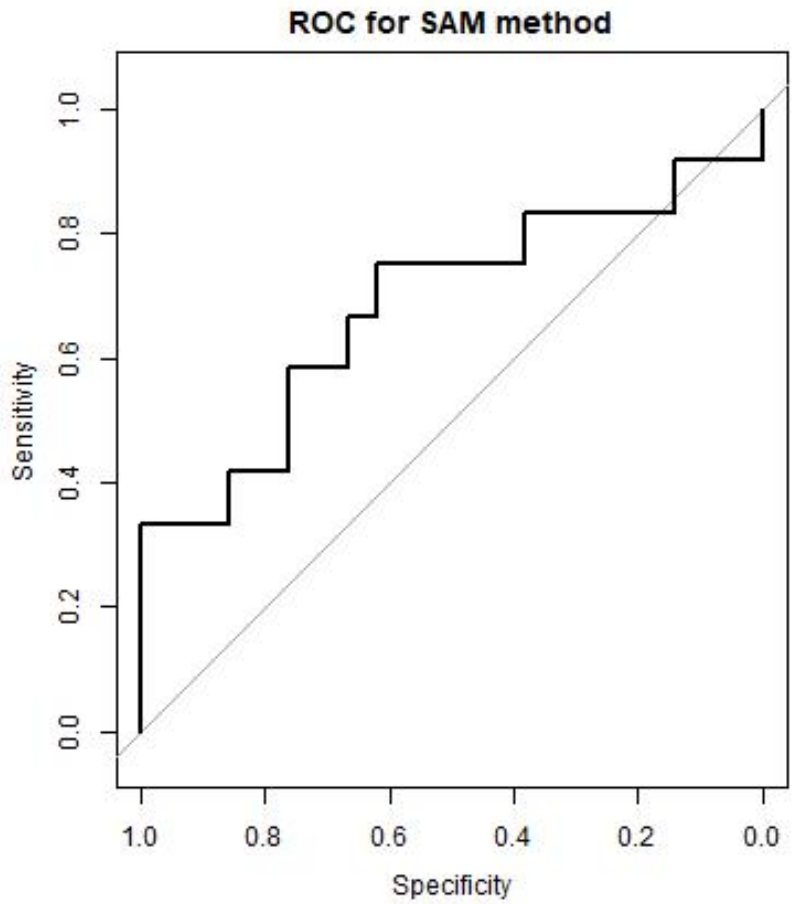




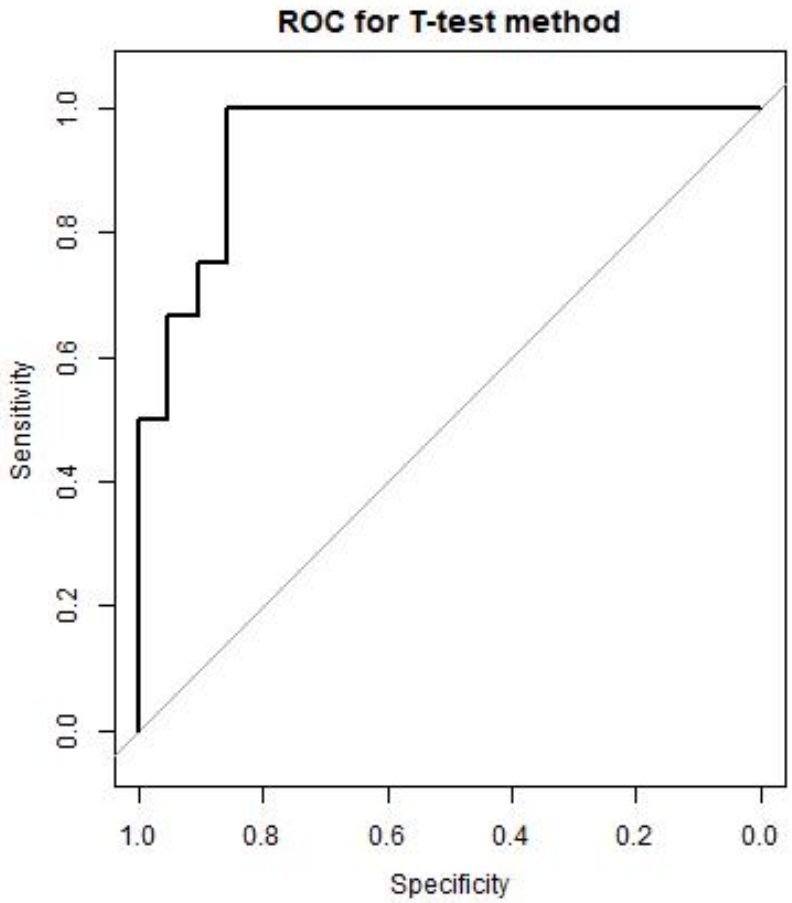
Genes selected (SAM) - 1

```
Genes up
NULL

Genes down
     Gene ID Gene Name Score(d)
[1,] g38824  38824     -4.113
[2,] g26636  26636     -3.996
[3,] g35097  35097     -3.966
[4,] g35323  35323     -3.935
[5,] g28199  28199     -3.822
[6,] g31108  31108     -3.727
[7,] g34538  34538     -3.683
[8,] g2122   2122      -3.679
```

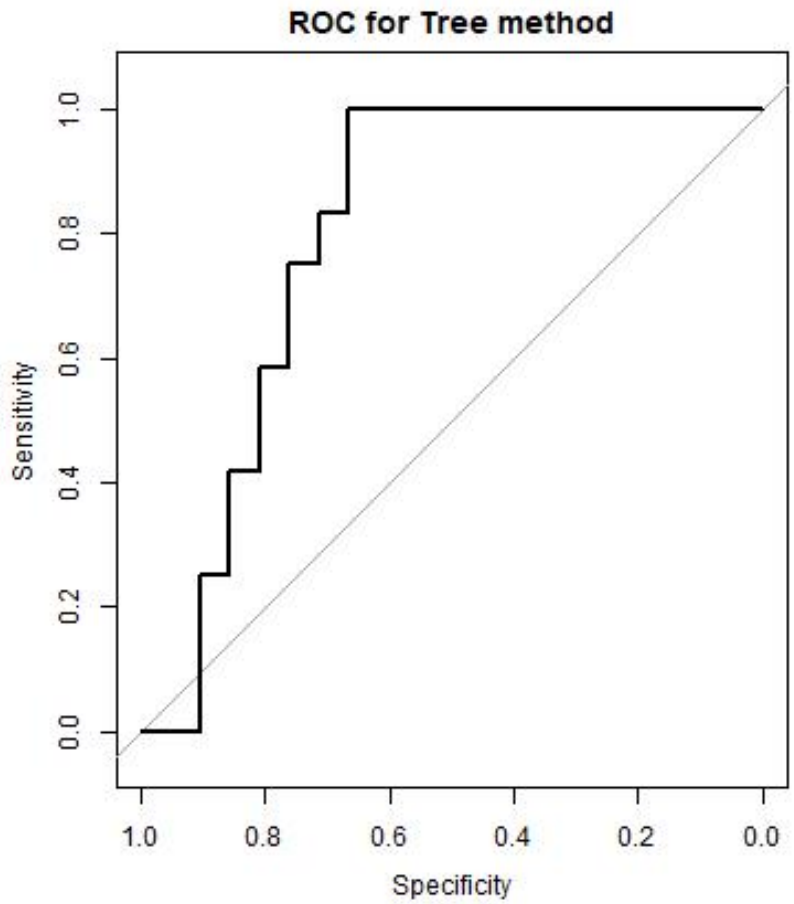
ROC for SAM method



ROC for T-test method



ROC for Tree-based model method



Summary on gene-based predictors

- Several methods exist to construct the gene-based classifier/predictor and they do not always produce the same results.
- Clinical approaches which incorporate *ad hoc* clinical decisions can be quite different from the methods purely based on statistical mechanical optimizations.
- Due to the nature of big-data with large number of genes on correlations and distributions, it is critical to carefully examine the data structure before any attempts of modeling.
- Since many biomarker/gene-based clinical studies depend on the classifier for patient selection, it is utmost important to incorporate both clinical and statistical expertise to produce the classifiers that fit the clinical/scientific objectives.

contents page

IVD companion diagnostic device - 1

Definition: IVD companion diagnostic device (CDx) is an in vitro diagnostic device that provides information that is essential for the safe and effective use of a corresponding therapeutic product.

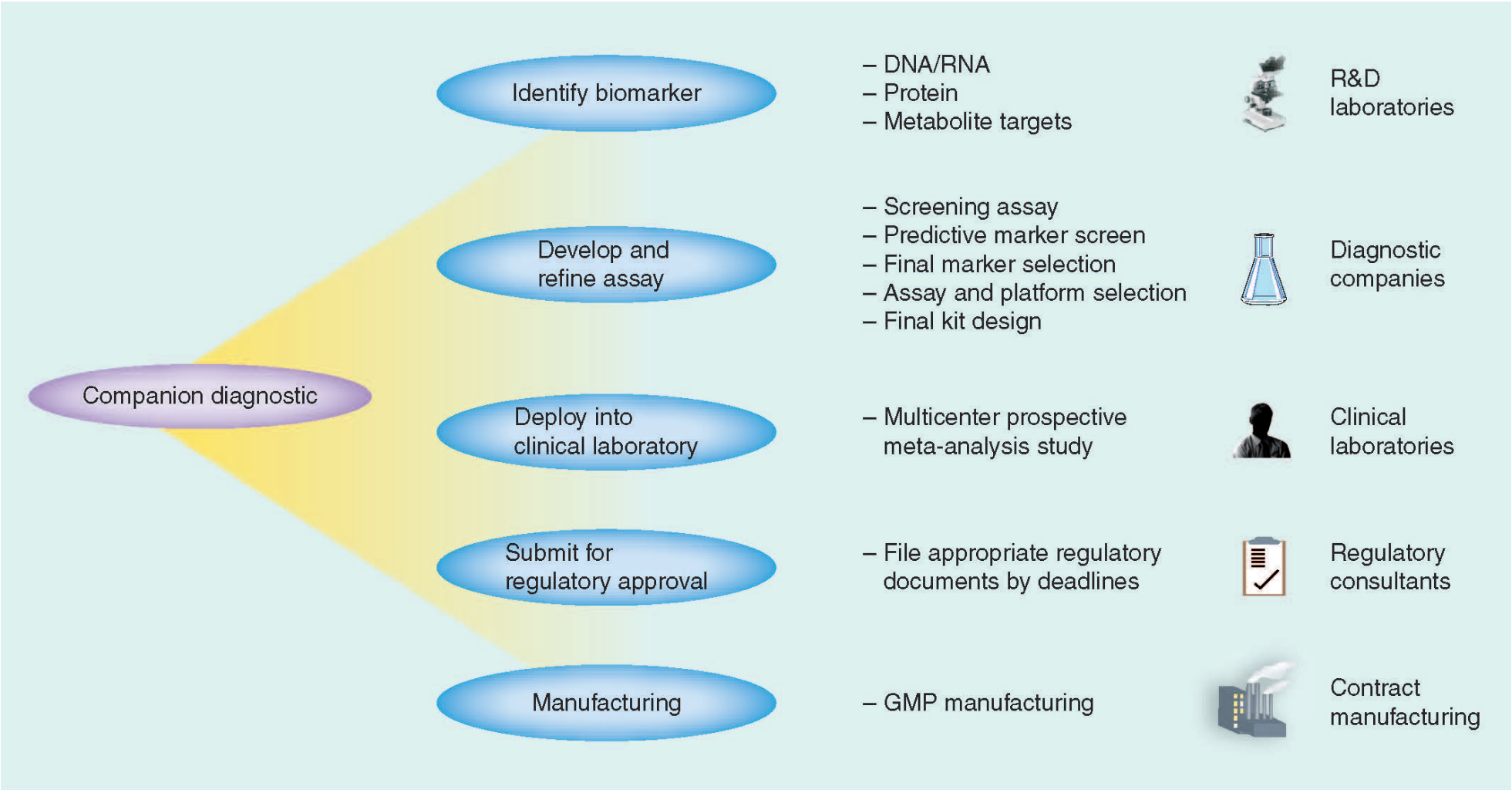
An IVD CDx could be essential for the safe and effective use of a corresponding therapeutic product to:

- Identify patients who are most likely to benefit from the therapeutic product.
- Identify patients likely to be at increased risk for serious adverse reactions as a result of treatment with the therapeutic product.
- Monitor response to treatment with the therapeutic product for the purpose of adjusting treatment (e.g., schedule, dose, discontinuation) to achieve improved safety or effectiveness.

IVD companion diagnostic device - 2

- Identify patients in the population for whom the therapeutic product has been adequately studied, and found safe and effective, i.e., there is insufficient information about the safety and effectiveness of the therapeutic product in any other population.

IVD CDx typical components required in development



Timing of IVD CDx development - 1

- Ideally, a therapeutic product and its corresponding IVD CDx should be developed contemporaneously, with the clinical performance and clinical significance of the IVD CDx established using data from the clinical development program of the corresponding therapeutic product.
- But exceptions do exist ...

Timing of IVD CDx development - 2

Novel Therapeutic Products

- For a novel therapeutic product for which an IVD CDx is essential for the safe and effective use of the product, the IVD CDx should be developed and approved or cleared contemporaneously so that it will be available for use when the therapeutic product is approved.
- FDA generally will not approve the therapeutic product or new therapeutic product indication if the IVD companion diagnostic device is not approved or cleared for that indication.

Timing of IVD CDx development - 3

Approval of a Therapeutic Product without an Approved or Cleared IVD CDx

- FDA may decide that it is appropriate to approve a therapeutic product even though an IVD companion diagnostic device is not approved or cleared contemporaneously if:
  - New Therapeutic Products to Treat Serious or Life-Threatening Conditions
  - Already Approved Therapeutic Products

General Policies:

- If the use of an IVD CDx is essential for the safe and effective use of a therapeutic product, an approved or cleared IVD CDx should be available for use once the therapeutic product is approved.

Example of applications - 1

Example: CC-5013-DLC-002 (ROBUST): patient selection criteria via CDx on GEP.

- ABC type determined using a validated Gene Expression Profile (GEP) assay performed on NanoString's nCounter® Dx Analysis System as assessed by Central Pathology.
- ABC type subjects are reported by the GEP assay as eligible for this inclusion criterion, whereas subjects with an indeterminate, unclassifiable, or GCB type are not eligible.
- The NanoString LST assay determines COO subtype by measuring 15 target genes and 5 housekeeping genes in the RNA from FFPE DLBCL tumor specimens on the nCounter Dx Analysis system.

Example of applications - 2

- A linear computational algorithm that sums the weighted gene expression is utilized to compute a Linear Predictor Score (LPS). The LPS is compared against thresholds that define value ranges for the assignment of an ABC or GCB subtype.

genemap page

Thresholds of LPS value for the assignment of ABC, GCB, or Unclassified subtype to determine eligibility for the ROBUST clinical trial

Subtype Assignment	LPS Cutoffs For Subtype Assignment	ROBUST Trial Eligibility
GCB	$\leqslant$ 1907.8	Ineligible
Unclassified	>1907.8 & < 2433.5	
ABC	$\geqslant$ 2433.5	Eligible

contents page

Biomarker-based clinical trial designs - 1

- Medical treatment for oncology patients is driven by a combination of the expected outcome for the patient (prognosis) and the ability for treatment to improve the expected outcome (prediction).
- Biomarkers aid this process through the estimation of disease-related patient outcomes and by the prediction of patient-specific outcome to treatments.
- Studies had shown that although molecular profiling is expensive, not doing so can be far more expensive and can lead to incorrect conclusions.

Biomarker-based clinical trial designs - 2

- The term 'biomarker' in oncology refers to a broad range of markers, including biochemical markers, cellular markers, cytokine markers, genetic markers, physiological results, radiological measurements, physical signs and pathological assessment.
- In the case of genetic markers, the pharmacogenetic determinants of efficacy and toxicity for many anticancer drugs remain unknown.
- A common approach to understand the genetic determinants of efficacy and toxicity is to look for molecular markers in the tumor itself.

Biomarker-based clinical trial designs - 3

- A prognostic marker is a single trait (or signature of traits) that
  - separates a population with respect to the outcome of interest in the absence of treatment, or regardless of (standard) treatment.
  - It is associated with the disease or the patient and not with a specific therapy.
  - Prognostic marker validation can thus be established using the marker and outcome data from a cohort of uniformly treated patients with adequate follow-up.
- A predictive marker is a single trait (or signature of traits) that
  - separates a population with respect to the outcome of interest in response to a particular treatment.
  - The use of a randomized controlled trial (RCT) as opposed to a cohort or single-arm study is essential for predictive marker validation.

Biomarker-based clinical trial designs - 4

- An important component of biomarker validation relates to biomarker assay issues, including the choice of using a central versus local laboratories for patient selection.
- This choice depends on: (a) the reliability and reproducibility of the assay, and (b) the complexity of the assay.
- Trial designs in the Phase II/III setting for initial marker validation can be classified under enrichment, all-comers or adaptive design categories, elaborated below.
  - Enrichment designs: An enrichment design screens patients for the presence or absence of a biomarker profile and then only includes patients who either have or do not have the profile in the clinical trial.

Biomarker-based clinical trial designs - 5

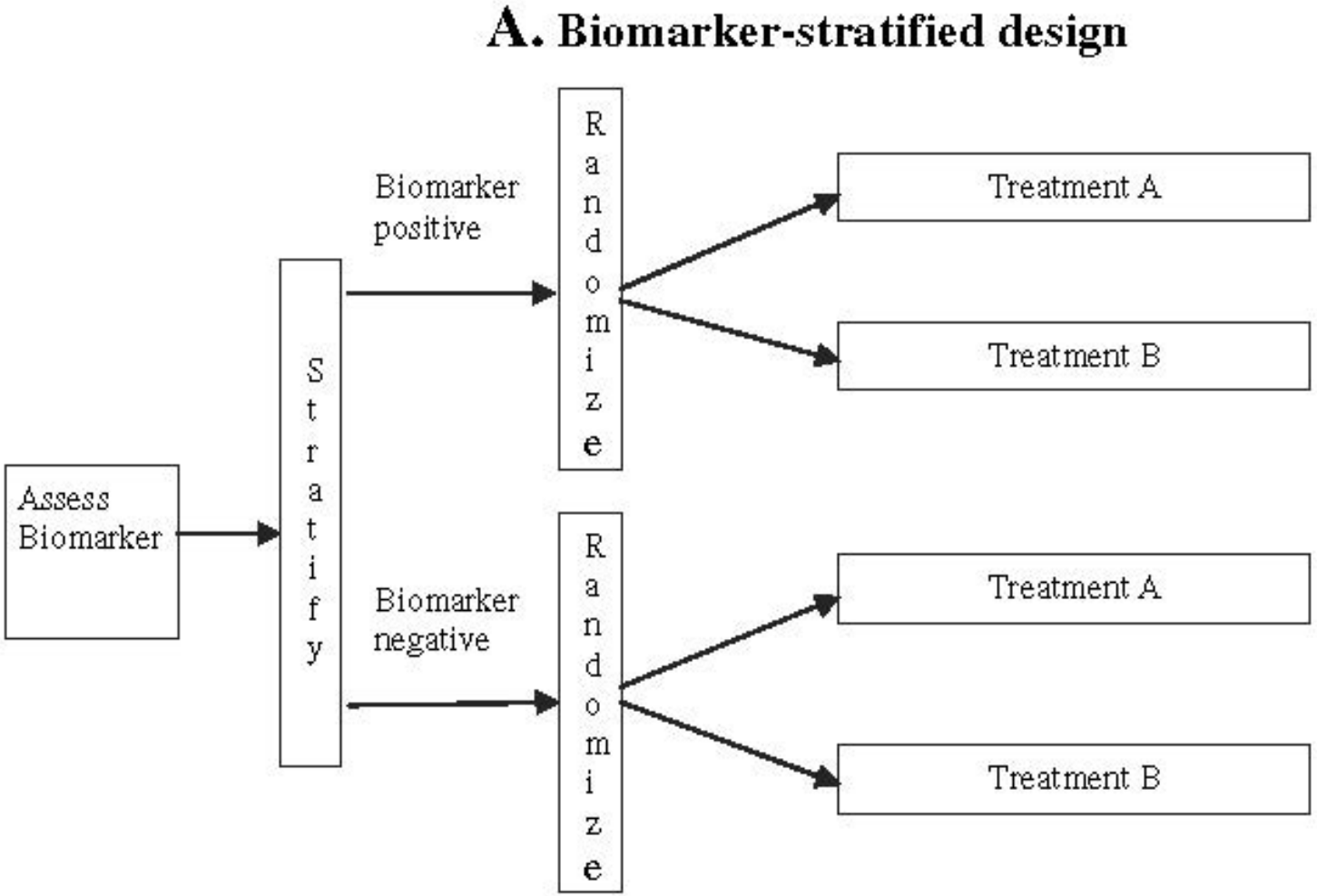
- All-comers (biomarker-stratified) designs: In this design, all patients meeting the eligibility criteria (not include the biomarker status) are entered. The ability to provide adequate tissue may be an eligibility criterion for these designs, but not the specific biomarker result, or the status of a biomarker characteristic.

- Adaptive designs: Adaptive design strategies are a class of randomized Phase II designs by which a variety of marker signatures and drugs can be tested under one umbrella protocol. In these designs, the success of the drug-biomarker subgroup is assessed in an ongoing manner, which allows either the randomization ratio to be altered to place more patients on the most promising arm(s) and/or the underperforming drugs and/or the biomarker subgroups are eliminated midway through the trial.

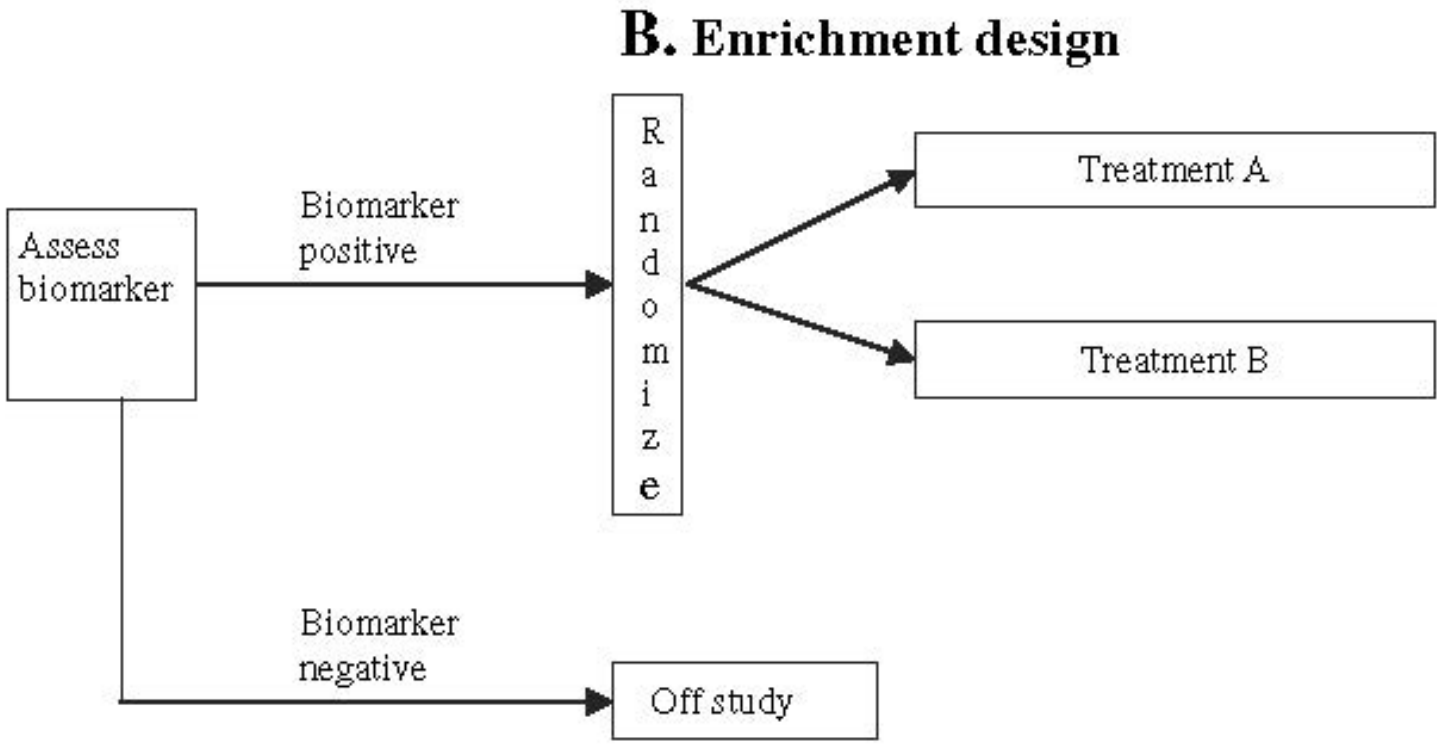
Common Criteria of Biomarker-based design

Criteria for choice of design for marker validation trials			
Criteria	Designs		
	Enrichment	All-comers	Adaptive
Strongly suggest benefit in marker-defined subgroups	Optimal	Not recommended	Appropriate (assess multiple treatments / biomarker subgroups)
Excellent assay reproducibility and validity	Required	Required	Required
Rapid turnaround time	Optimal	Optimal	Optimal

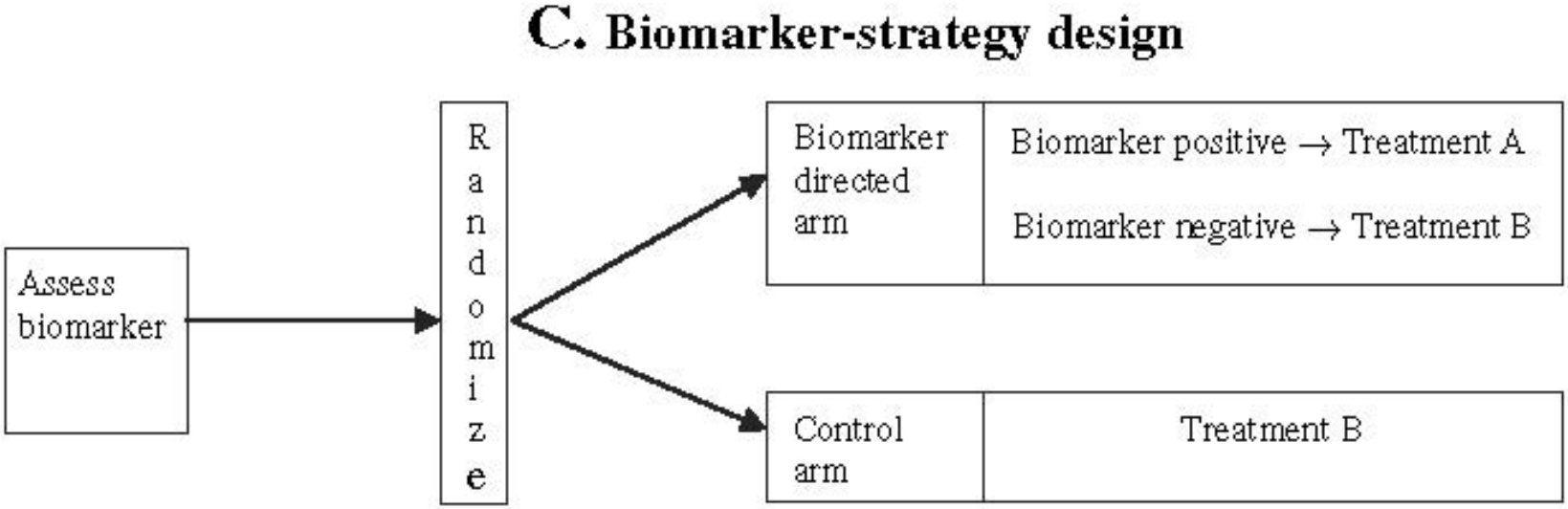
Biomarker stratified design (All comers)



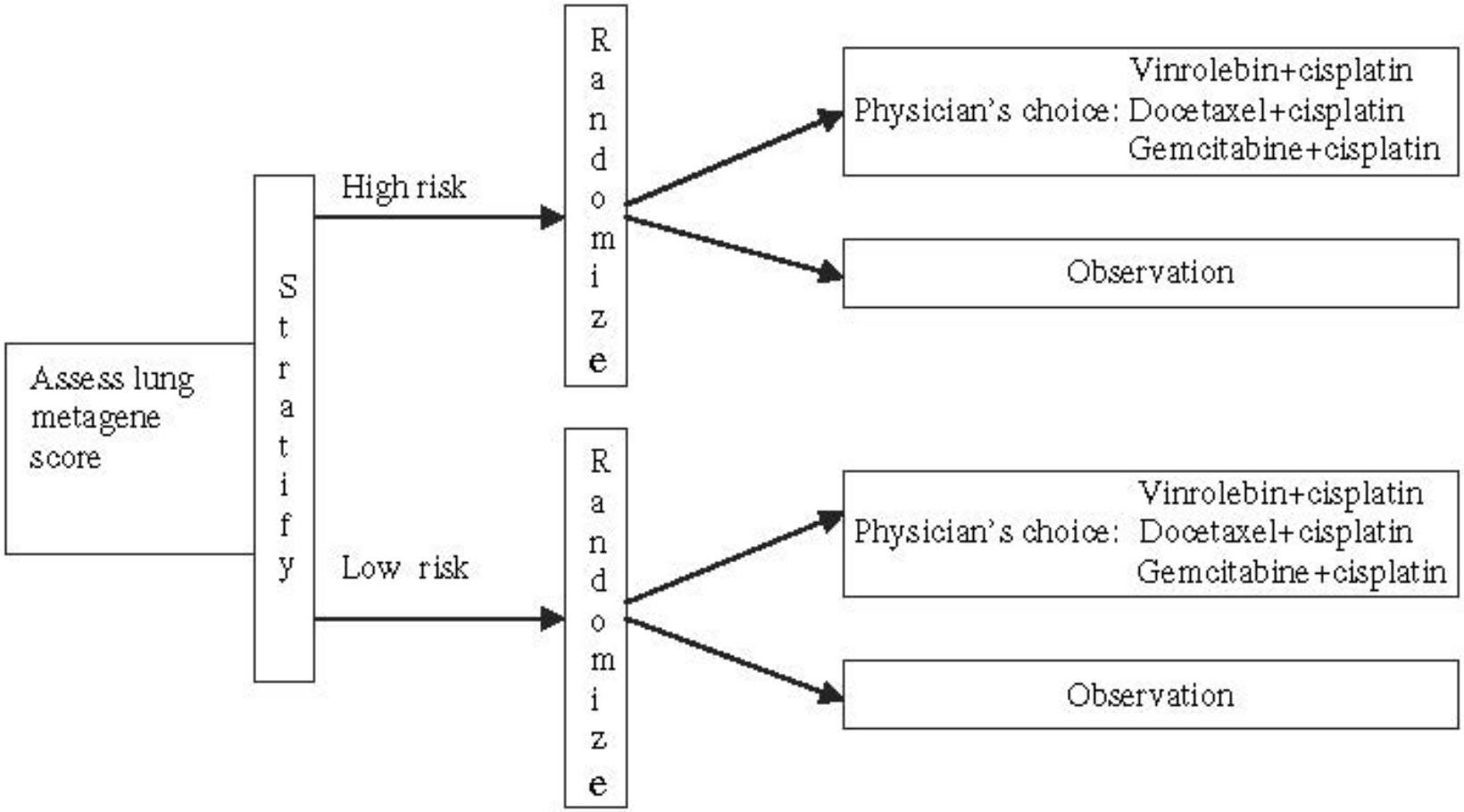
Enrichment design



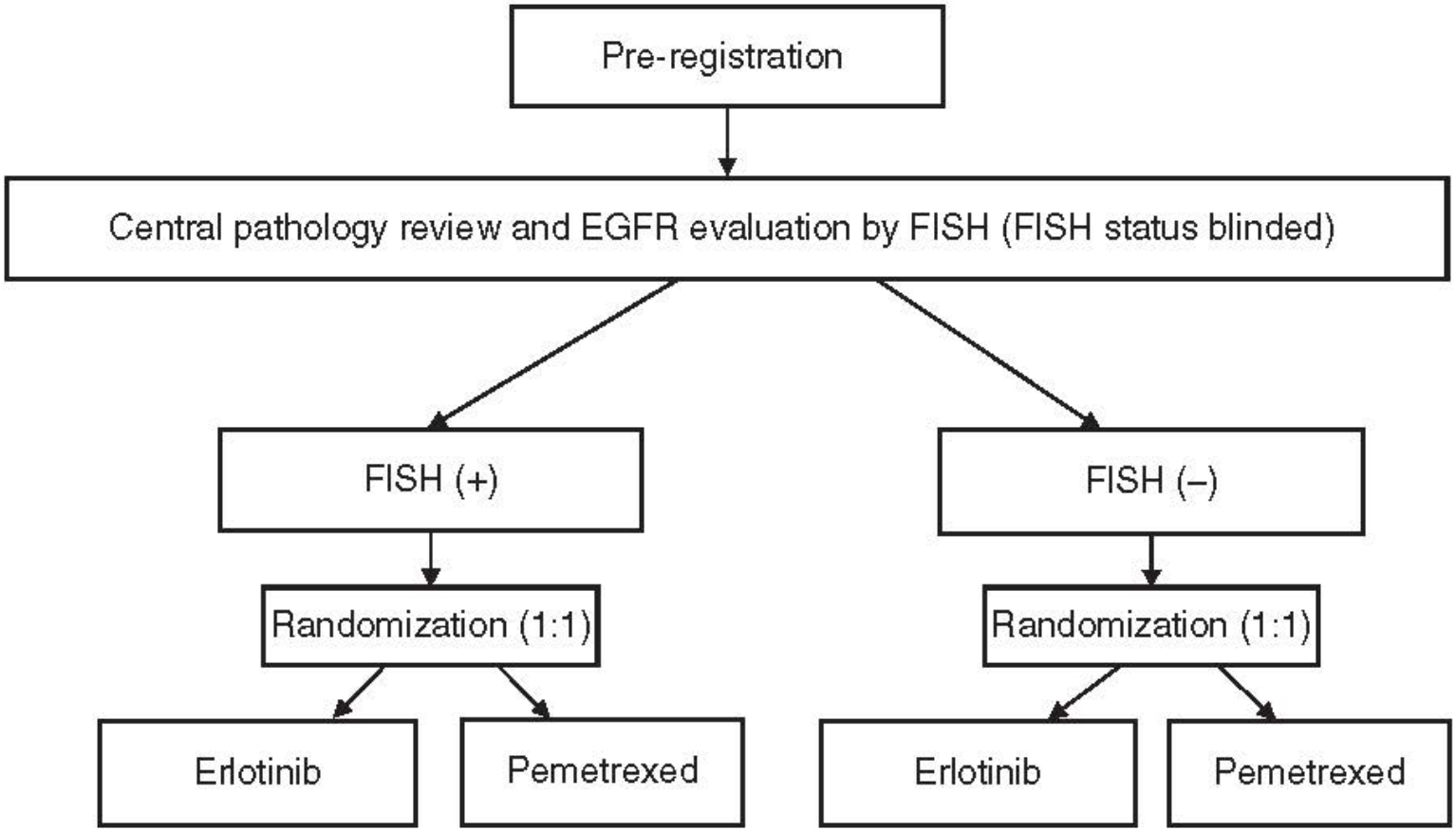
Biomarker strategy design



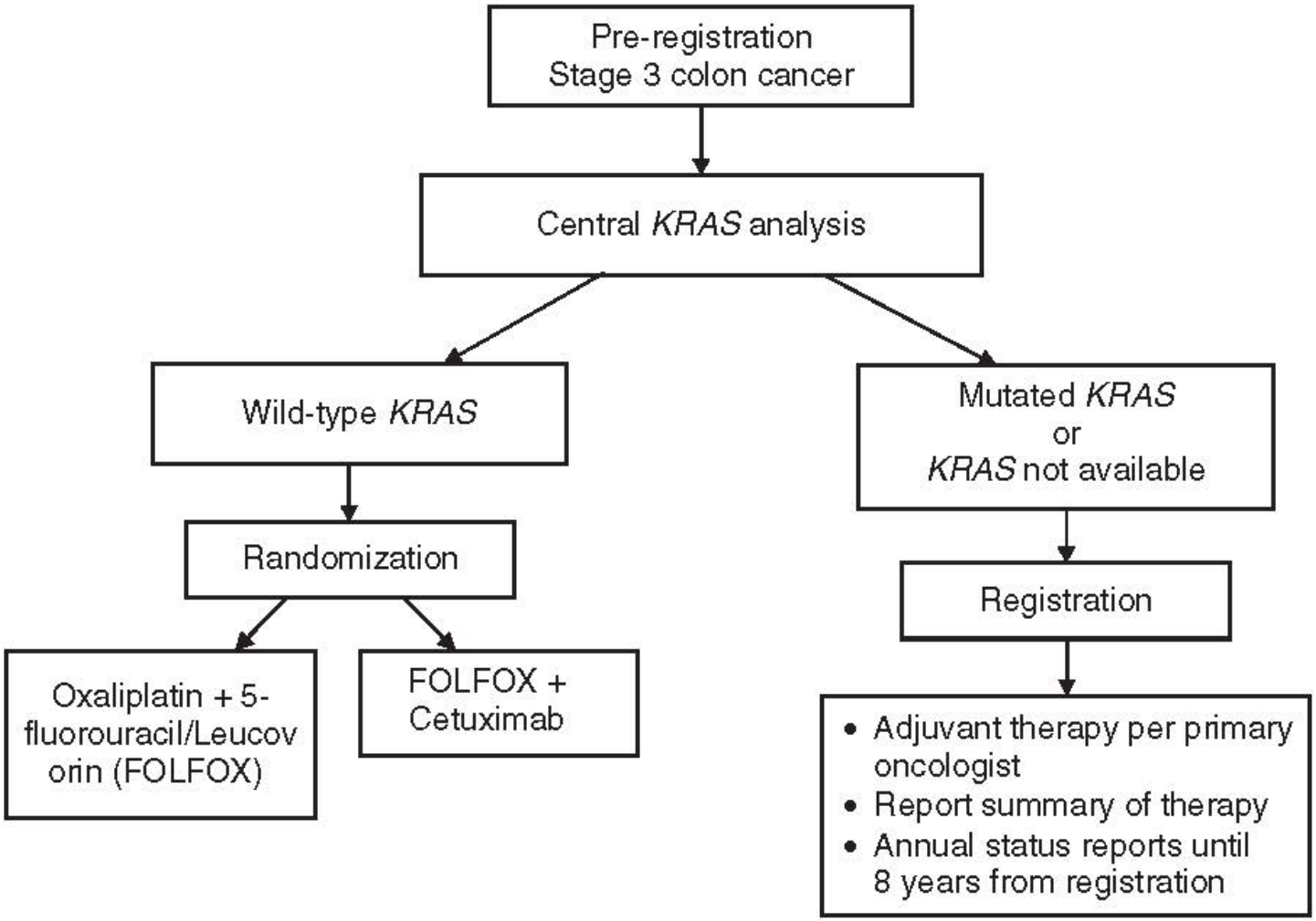
CALGB-30506


Assess lung metagene score
Stratify
High risk
Low risk
Randomize
Randomize
Physician's choice: Vinrolebin+cisplatin Docetaxel+cisplatin Gemcitabine+cisplatin
Observation
Physician's choice: Vinrolebin+cisplatin Docetaxel+cisplatin Gemcitabine+cisplatin
Observation

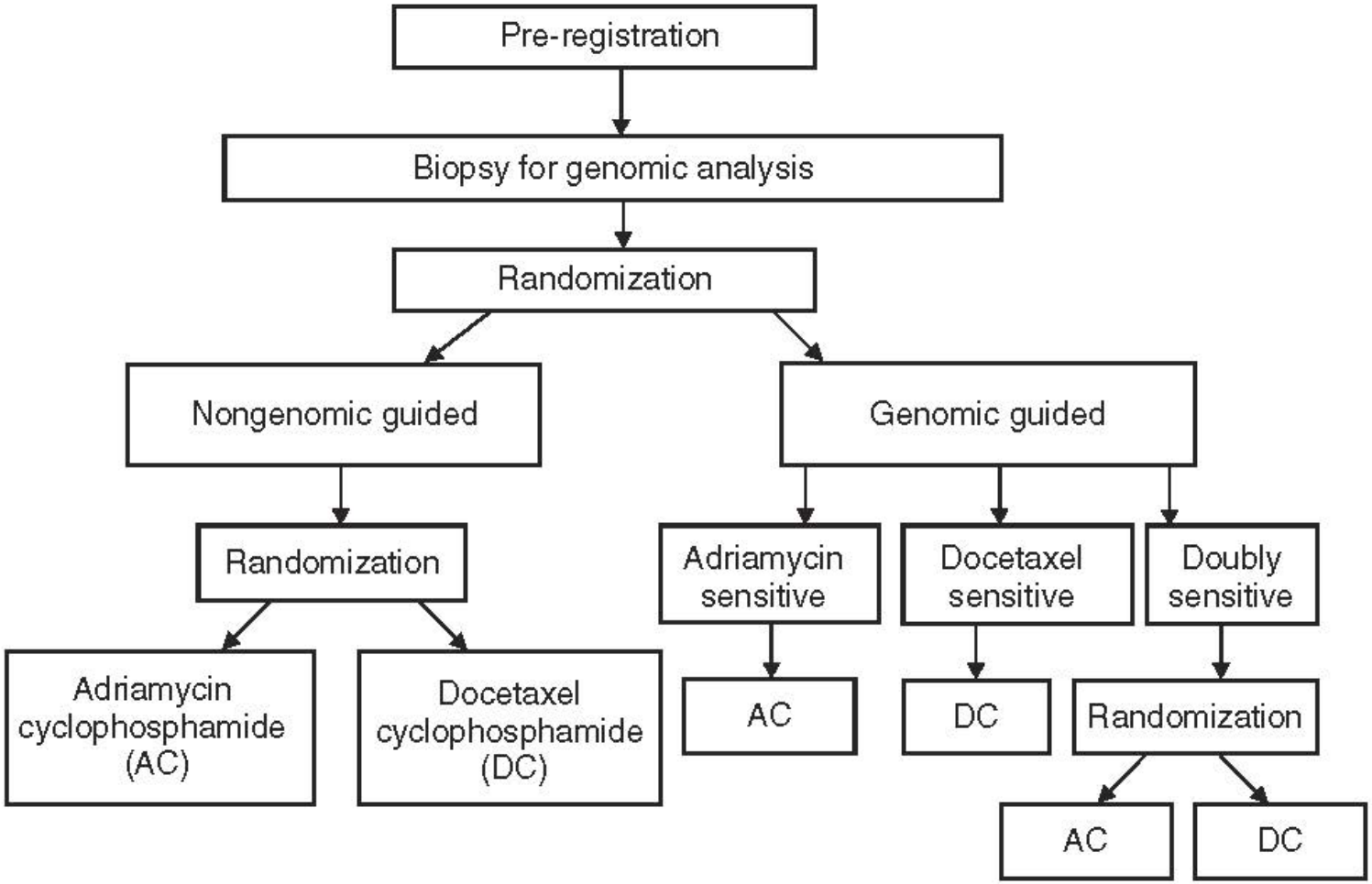
MARVEL


Pre-registration
Central pathology review and EGFR evaluation by FISH (FISH status blinded)
FISH (+)
FISH (–)
Randomization (1:1)
Randomization (1:1)
Erlotinib
Pemetrexed
Erlotinib
Pemetrexed

N0147


Pre-registration
Stage 3 colon cancer
Central *KRAS* analysis
Wild-type *KRAS*
Mutated *KRAS*
or
KRAS not available
Randomization
Registration
Oxaliplatin + 5-fluorouracil/Leucovorin (FOLFOX)
FOLFOX + Cetuximab
• Adjuvant therapy per primary oncologist
• Report summary of therapy
• Annual status reports until 8 years from registration

Marker-based-design



contents page

Summary

- Critically important and extremely challenging.
- but it is the future, be ready for it!