

C7188 Úvod do molekulární medicíny 4/12



Moderní metodické přístupy v molekulární medicíně II



GENOMIKA II



Ondřej Slabý, Ph.D.

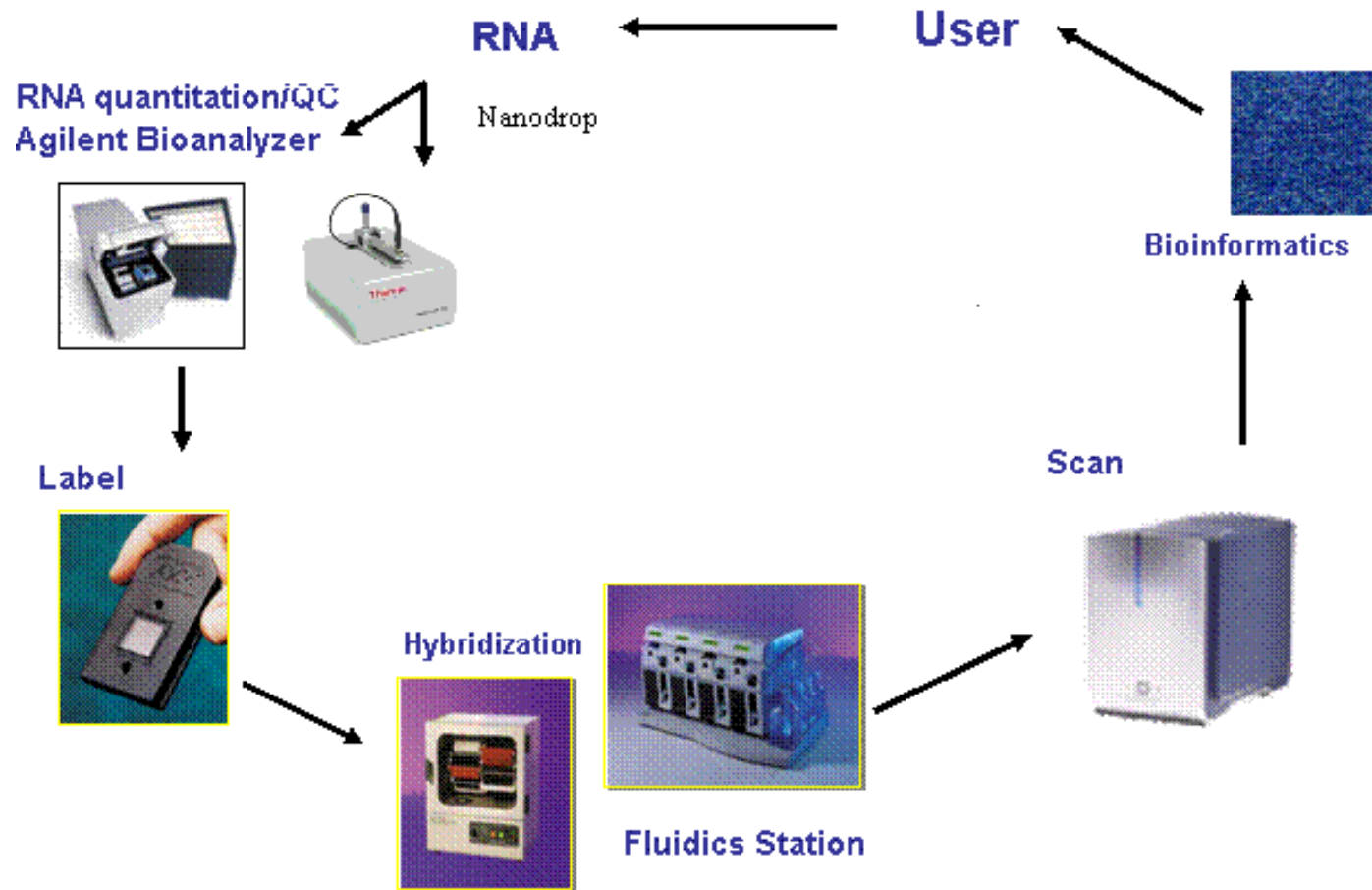
Masarykův onkologický ústav

Univerzitní centrum buněčné imunoterapie

Lékařská fakulta Masarykovy univerzity

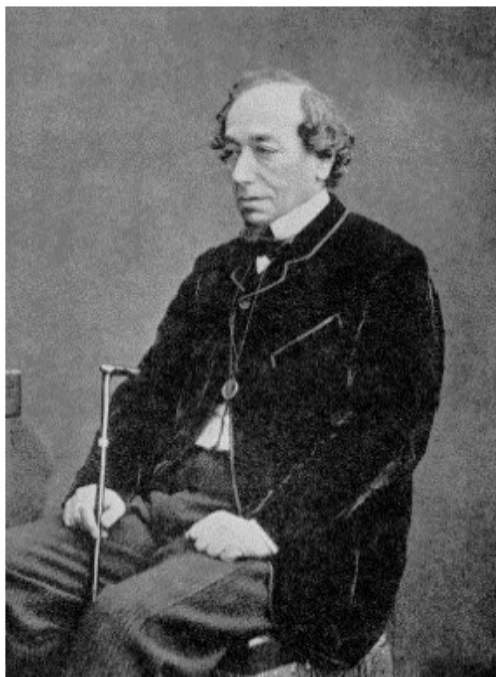


Microarray sample processing workflow



“If your experiment needs statistics, you should have done a better experiment.”

Ernest Rutherford



“There are three types of lies: lies, damn lies, and statistics!”

Benjamin Disraeli

Analýza obrazu

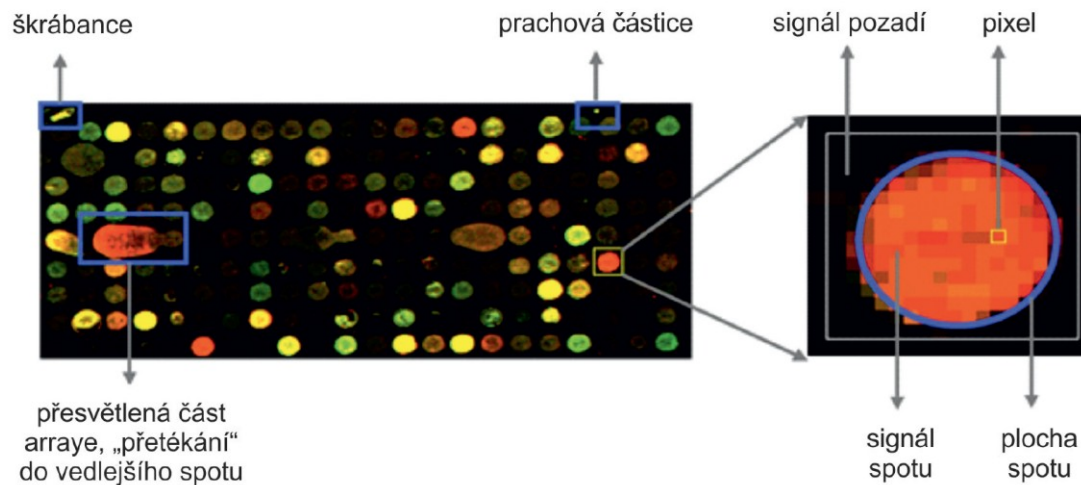
Odečtení pozadí od „foreground“

– background subtraction

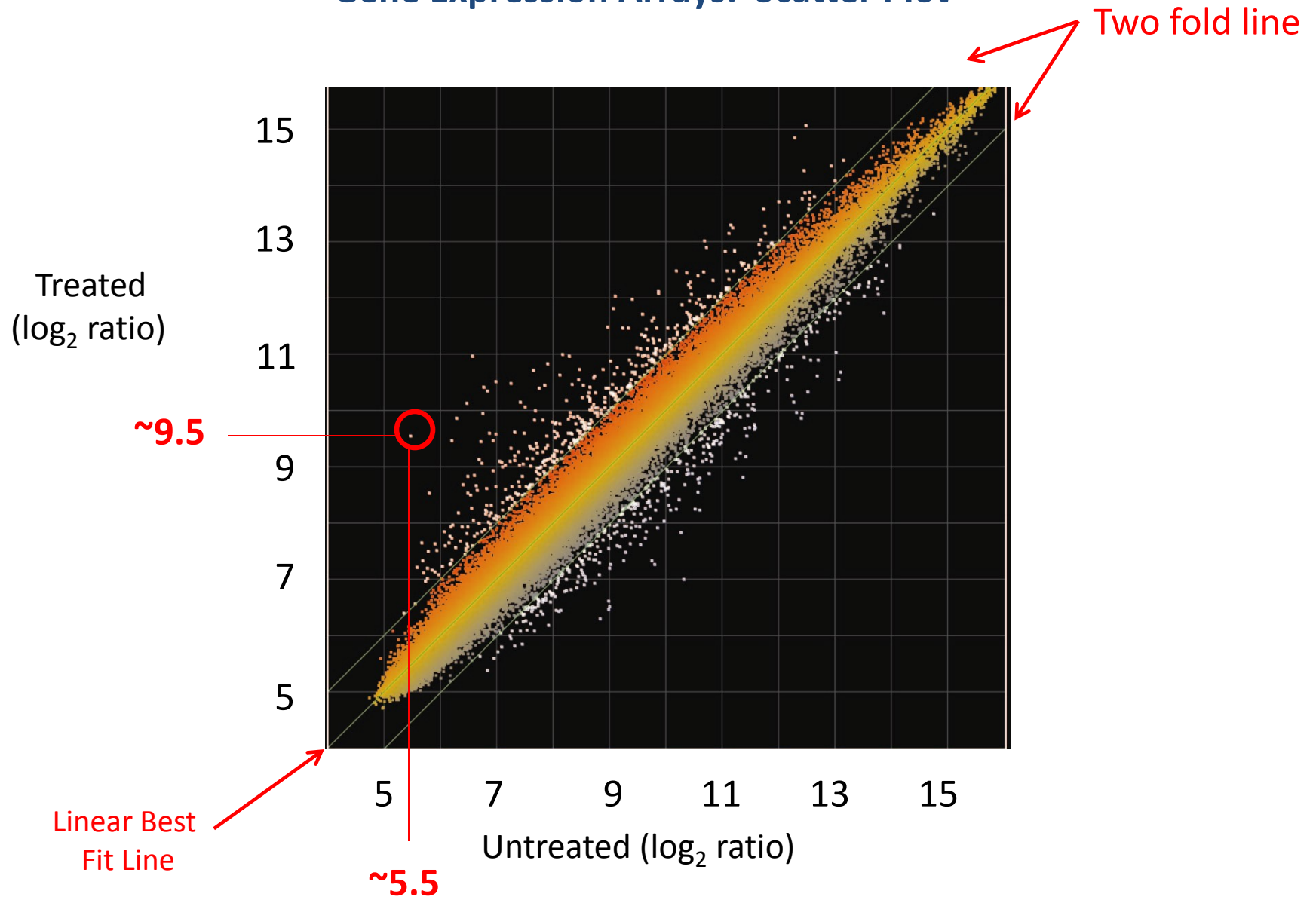
(gridding, local background, median, empty spot,...)

SNR – „signal-to-noise ratio“, odstup intenzity popředí od šumu pozadí

-převod obrazové informace na numerická data



Gene Expression Arrays: Scatter Plot



Normalizace dat

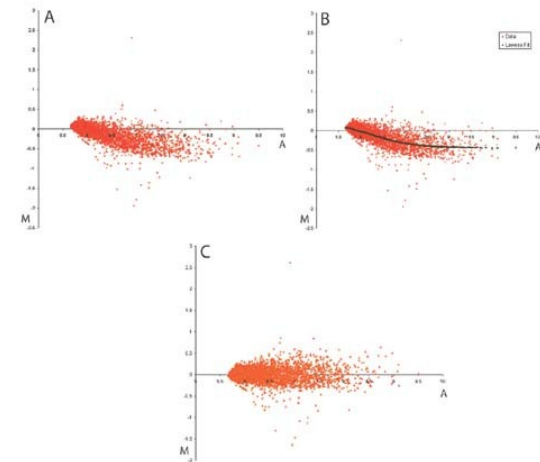
ukázky normalizačních metod

Základem většiny normalizačních metod je předpoklad, že počet genů se změnou mírou exprese je výrazně nižší než těch, jejichž exprese se nemění.

- minimalizace vlivu „nebiologických“ zdrojů variability
- kompensace nelinearity dat mezi jednotlivými čipy a uvnitř daného čipu

Pozitivní kontroly:

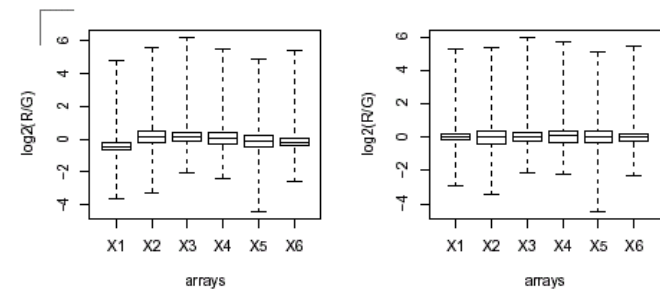
- Referenční geny s „konstantní“ expresí ve tkáních
- kontrolní genetický materiál (referenční vzorek)
- kontroly účinnosti hybridizace – arteficiální sekvence



Mean, median normalization

Negativní kontrola:

- pozadí hybridizace
- Affymetrix - mutace v jednom nukleotidu sondy



Subtract common mean or median:

old values: $y_{gi} = \log_2(R_{gi}/G_{gi})$ on array i

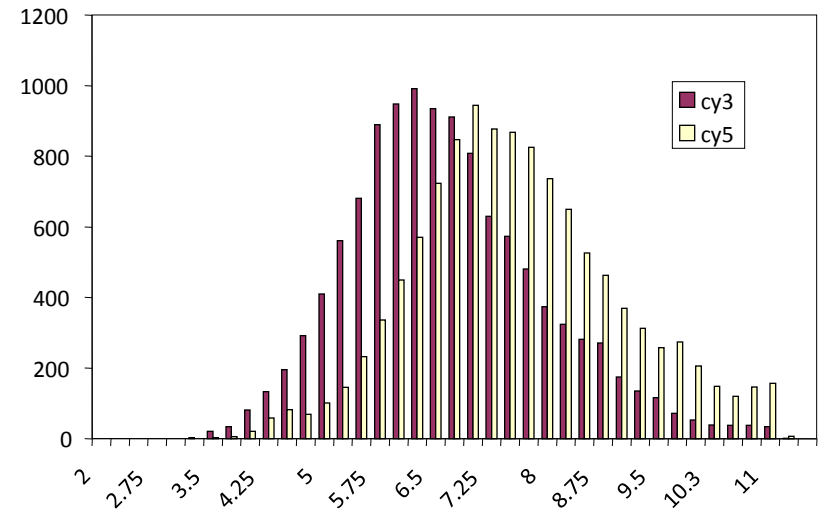
new values: $y_{gi}^{(n)} = y_{gi} - \text{mean}_i$

Normalizace dat Cy3 a Cy5

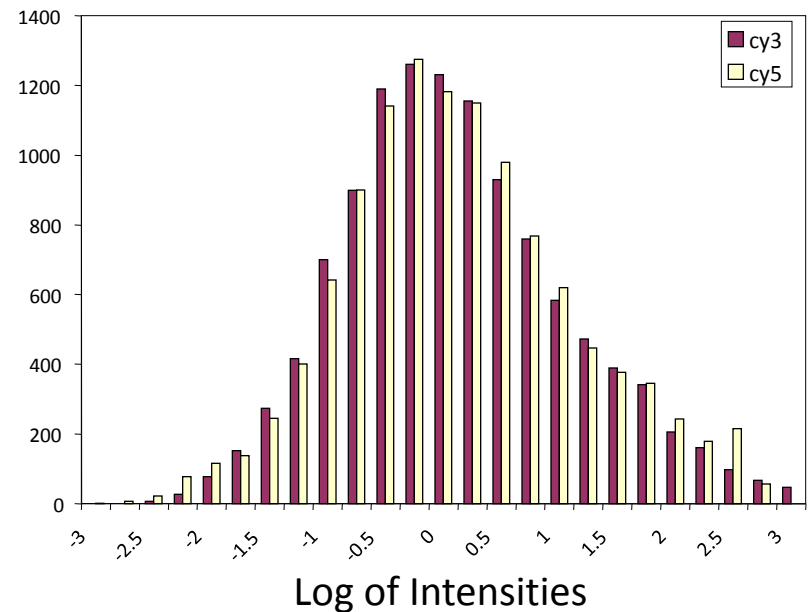
vybalancování rozdílného signálu ze značení Cy3 a Cy5

Jeho nerovnoměrnost může být způsobena:

- rozdílnou inkorporací barviva do NK
- rozdílným množstvím mRNA
- odlišnými parametry při skenování



Number of clones



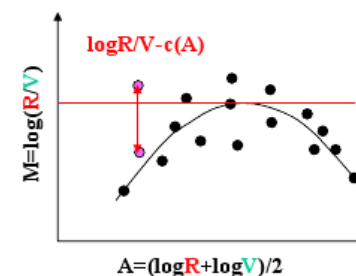
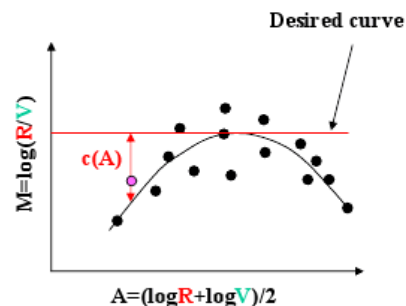
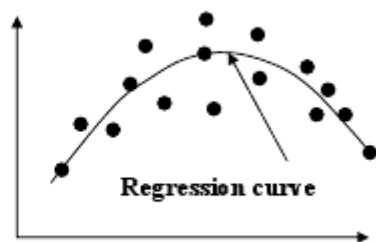
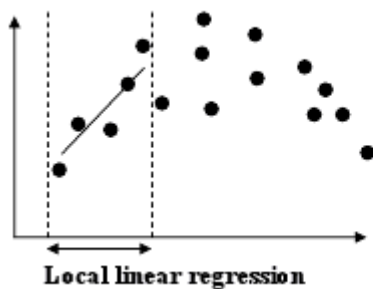
Normalizace Lowess: princip

Loess (or lowess) : Locally WEighted Scatterplot Smoothing (vyhlazování)

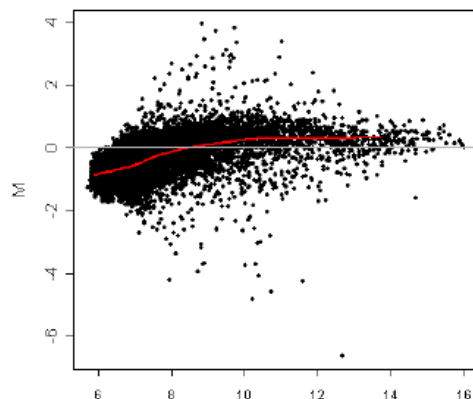
Normalizace závislá na intenzitě signálu

Místní lineární regrese

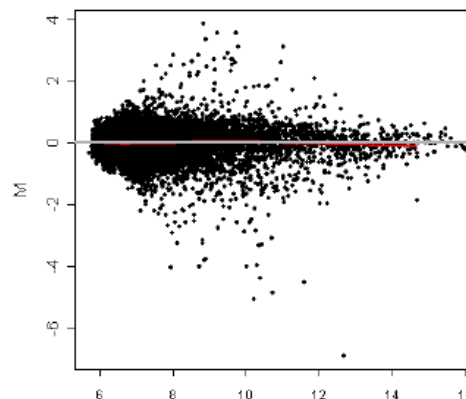
$$\log(R/V) \rightarrow \log(R/V) - c(A)$$



Before normalization



After normalization



Statistická analýza dat

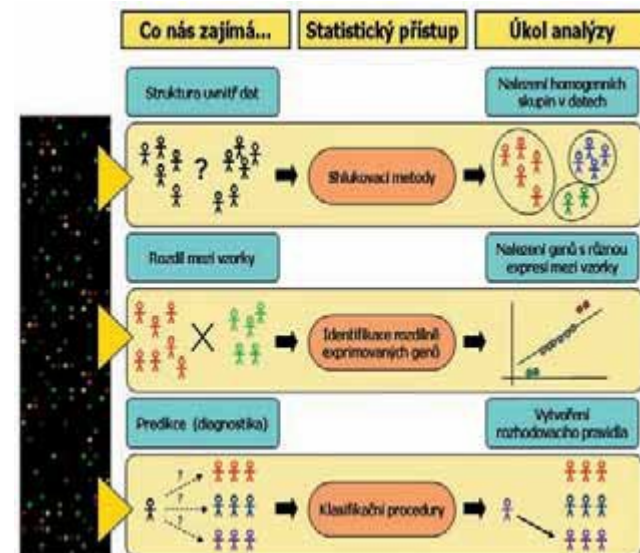
Základní dělení metod je na „unsupervised“ (využívají jako vstup pouze naměřené hodnoty bez informací o vzorcích, klastrovací metody) a „supervised“ (využívají i další známé informace o vzorcích, identifikace biologicky významných genů, klasifikační metody)

Další skupina metod propojuje informaci o genové expresi s dostupnými informacemi o biologické funkci genů (analýzy signálních drah, informace z databáze GO „gene ontology“, a další)

1) Identifikace biologicky významných genů

- geny s reprodukovatelnou signifikantně rozdílnou expresí mezi jednotlivými podmínkami experimentu
- poměr exprese v jednotlivých experimentech (fold change)
- t-test (test rozdílnosti průměrů exprese v jednotlivých skupinách)
- neparametrické testy (Mann-Whitney, Wilcoxon test)
- Significance Analysis of Microarrays (SAM)
- Multifaktoriální ANOVA (nejsignifikantnější geny pro dané skupiny)

Hranice významnosti (1% = 1 ze 100, u čipu je to 200 falešně pozitivních z 20000, Bonferoniho korekce)

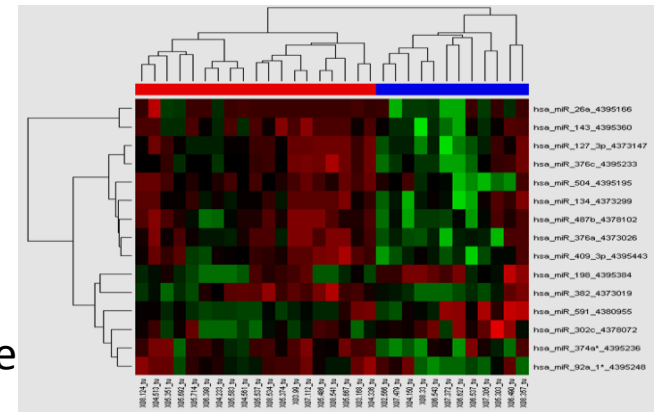


2) Ukázky multidemenzionálních metod analýzy čipových dat

Shlukovací analýzy

Shlukovací analýza je jednou z nejpoužívanějších vícerozměrných statistických metod

Jedná se o explorativní techniku, která se používá zejména v případech, kdy nemáme žádné *a priori znalosti* o struktuře uvnitř dat.



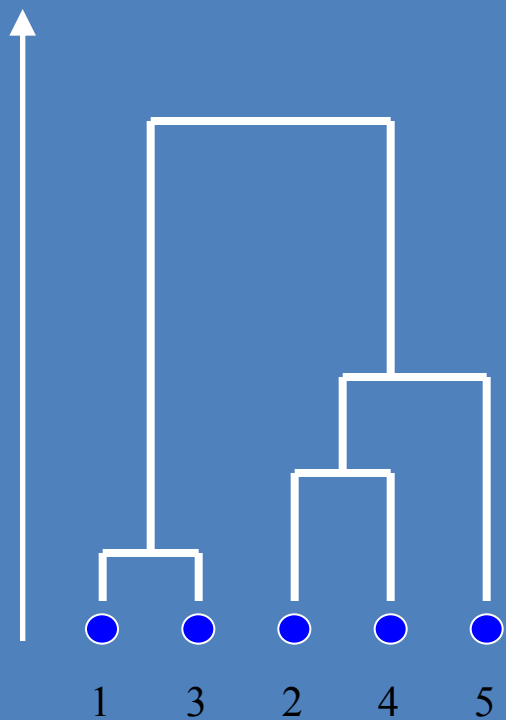
každý gen je reprezentován vektorem jehož souřadnice, jsou hodnoty exprese genu v jednotlivých experimentech (v jednotlivých vzorcích), vzdálenost je měřena mezi vektory.

Ukolem shlukovacích metod je tedy najít v datech skupiny prvků (shluky) tak, že prvky jednotlivých skupin budou v jistém smyslu více podobné než prvky z jiných skupin, tzn. nalezené skupiny prvků budou co nejvíce homogenní

Snažíme se nalézt mezi zkoumanými geny (resp. biologickými vzorky) skupinky genů (resp. biologických vzorků), které vykazují za specifických podmínek nebo u daného fenotypu, podobné chování.

Hierarchické klastrování

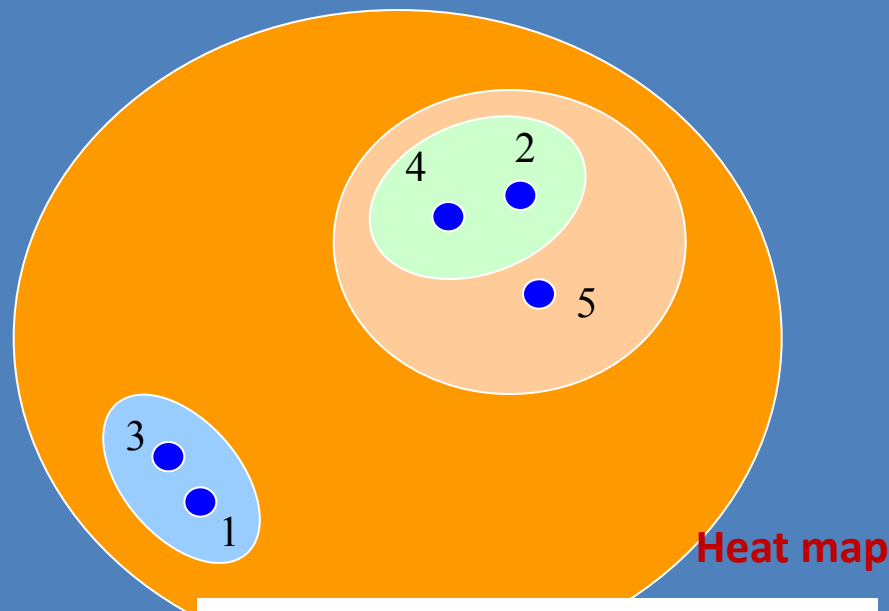
Vzdálenost mezi jednotlivými klastry



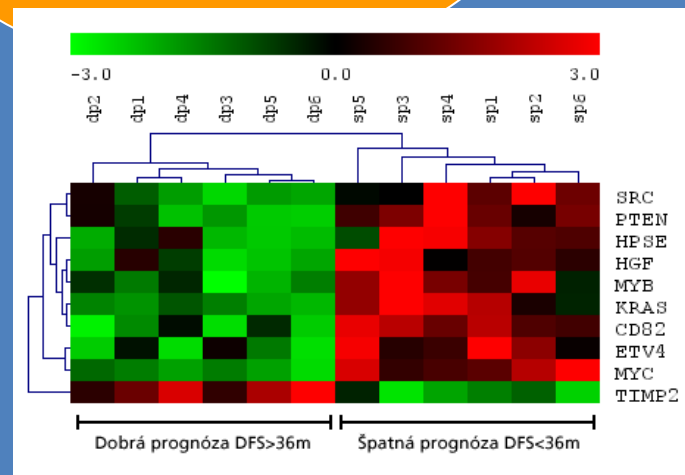
Dendrogram

Podobnost je vyjádřena hierarchickým stromem – dendrogram s teplotní mapou „heat-map“

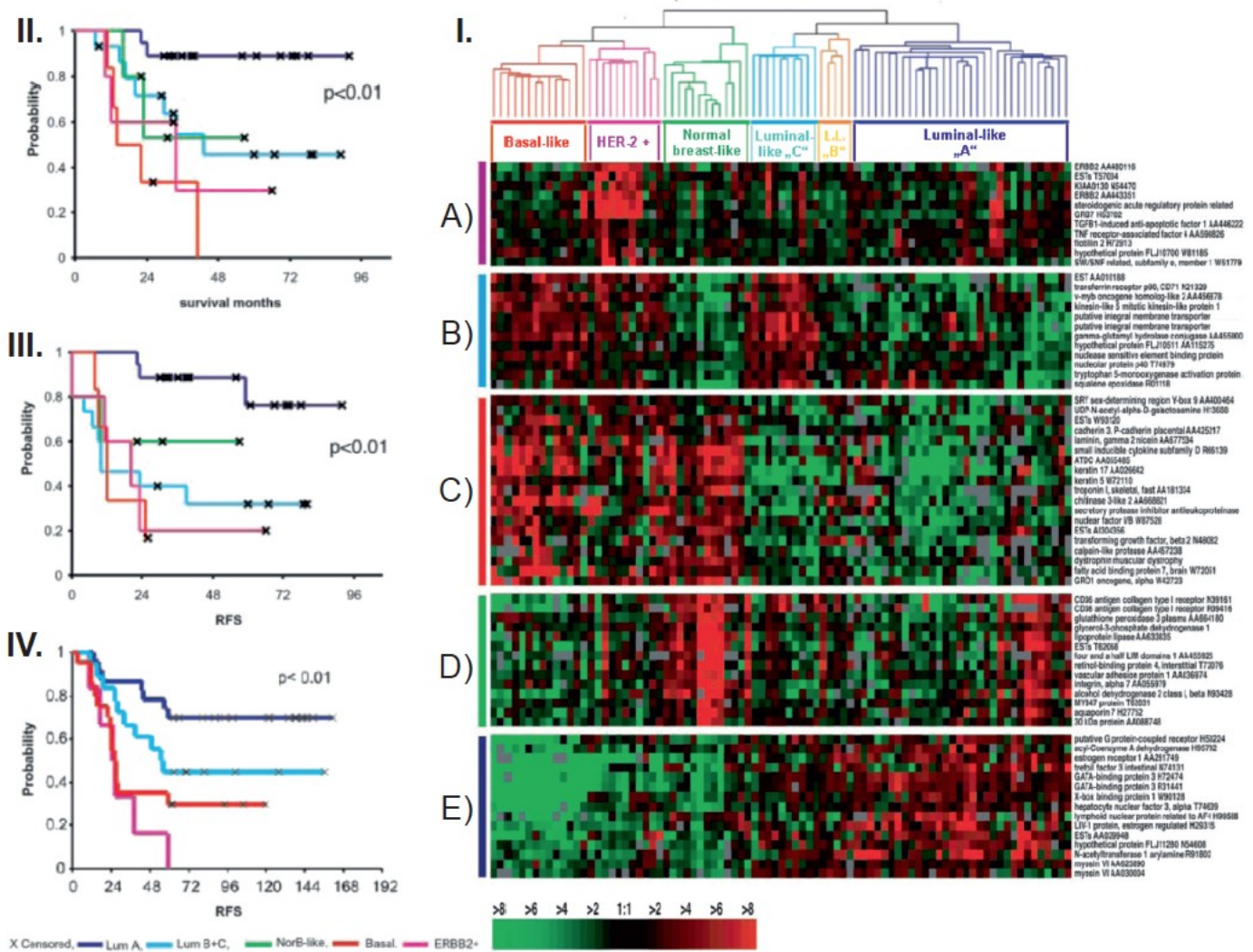
- kalkulace vzdálenosti mezi všemi geny a nalezení nejmenší. K ní se seskupí všechny jí podobné a vytvoří se klastr.
- po vytvoření X počtu clusterů se hledají vzdálenosti mezi klastry (hierarchical clustering)
- počet klastrů není omezen

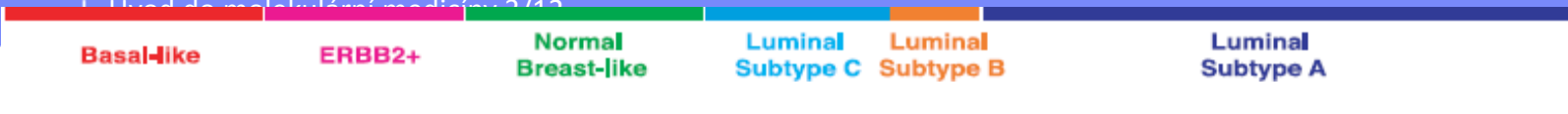


Heat map

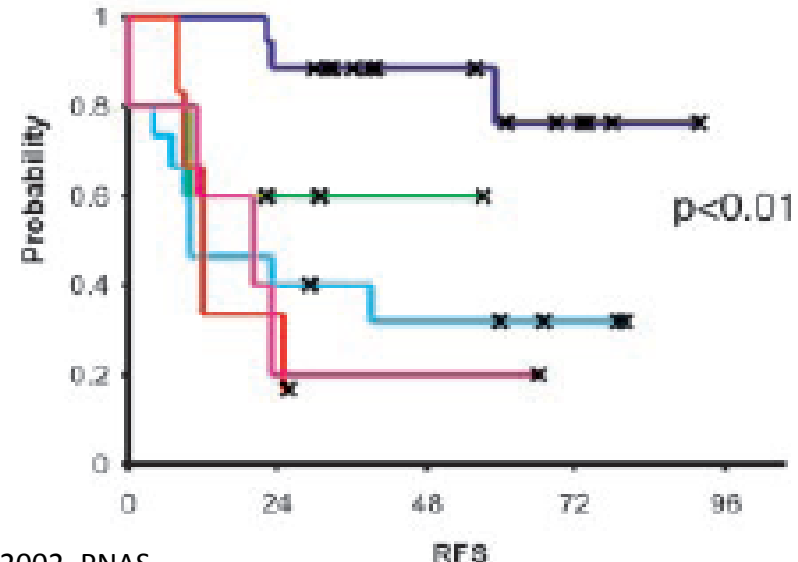
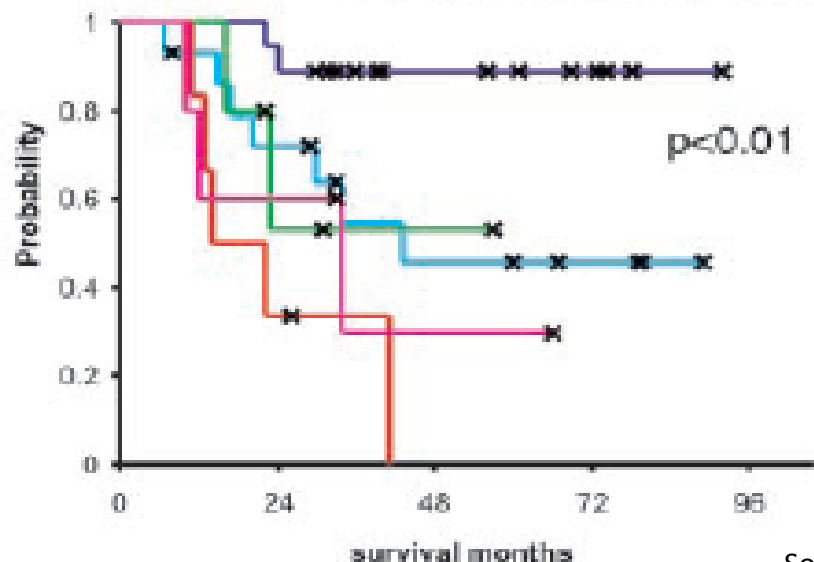


Klastrová analýza u karcinomu prsu





	Basal-like	ERBB2+	Normal Breast-like	Luminal Subtype C	Luminal Subtype B	Luminal Subtype A
ER	-/+	-/+	-/+	+ / +++	++	+++
HER2	-	+++	- / +	- / +	-	- / +
p53mut	82%	71%	33%	80%	40%	13%
CK 5/6, 17	+++	+/-	+++	-	-	-
CK 8/18	-	-	+	+ / +++	+ / +++	+++
c-myb / ost.	+++	+++	- / +	+++	-	-



Sorlie et al, 2002, PNAS

— Lum A,
 — Lum B+C,
 — NorB-like,
 — Basal,
 — ERBB2+

3) Klasifikační metody

Jejich podstatou je znalost informace o charakteru vzorku tzn. supervised přístup. Principem klasifikačních metod v analýze dat z DNA čipů je vytvoření rozhodovacího pravidla, které by na základě naměřených hodnot genové exprese umožňovalo přiřazení pacienta do jedné z předem definovaných tříd (například zdravý, nemocný). Z toho je zřejmé, že by se „dobré“ rozhodovací pravidlo založené na expresních datech mohlo zařadit po bok stavajících diagnostických metod a výrazně tak přispět ke zpřesnění diagnostiky závažných onemocnění (klasifikační stromy, Support Vector Machines (SVM), metoda k-nejbližších sousedů,..)

MOLEKULÁRNÍ KLASIFIKACE NÁDORŮ:

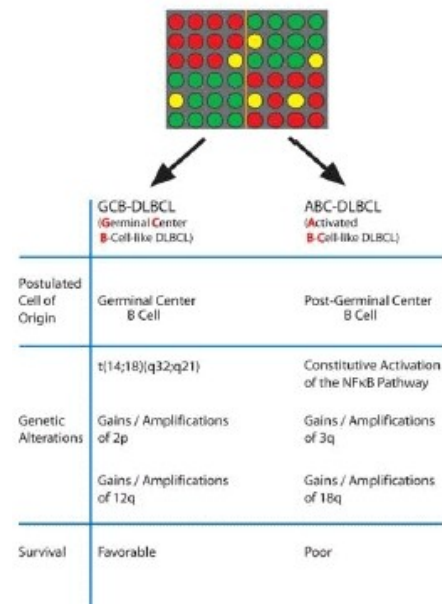
precizní klasifikace je základem léčebného úspěchu, současné metody jsou založeny na morfologii, imunohistochemii, genetice a klinické odpovědi

řada diagnostických nejasností (heterogenita)

ČIPY:

- identifikace nových jednotek na podkladě profilu genové exprese
- reklasifikace stávajících jednotek
- identifikace skupin či jednotlivých genů „markerů“ specifických pro dané jednotky

Identification of Molecular Subgroups of Diffuse Large B-Cell Lymphoma (DLBCL)



A)

Tab. 4.3.1. GO kategorie genů s významně odlišnou expresí mezi skupinou vysoce rizikových patientek a skupinou patientek s nízkým rizikem.

B) GO kategorie genů s významně odlišnou expresí mezi skupinou vysoce rizikových patientek a skupinou patientek s nízkým rizikem: 20 (z 83) nejsignifikantnějších kategorií z 1 381 genů v TCGA datasetu. Tučně jsou opět vyznačeny společné kategorie v japonském datasetu A a TCGA datasetu. Q-hodnota odpovídá Fischerovu exaktnímu testu s Benjaminiho-Yekutieliho korekcí na mnohonásobná porovnání.

GO kategorie	Geny v GO kategorii		-Log ₁₀ Q
	N	%	
immune system process (GO:0002376)	129	20,5	20,9
immune response (GO:0006955)	115	18,3	20,8
defense response (GO:0006952, 0002217, 0042829)	78	12,4	10,3
antigen processing and presentation (GO:0019882, 0030333)	25	4	8,4
inflammatory response (GO:0006954)	50	7,9	8,2
antigen processing and presentation of peptide antigen (GO:0048002)	13	2,1	6,9
antigen processing and presentation of exogenous peptide antigen (GO:0002478)	6	1	5,2
MHC class I peptide loading complex (GO:0042824)	9	1,4	5,2
MHC protein complex (GO:0042611)	18	2,9	5,2
TAP complex (GO:00042825)	8	1,3	5,2
MHC protein binding (GO:0042287)	12	1,9	5,2
antigen processing and presentation of exogenous antigen (GO:0019884)	7	1,1	5
response to stimulus (GO:0050896, 0051869)	183	29	4,4
antigen processing and presentation of peptide antigen via MHC class I (GO:0002474)	7	1,1	3,9
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II (GO:0002504)	12	1,9	3,9
regulation of immune response (GO:0050776)	6	1	3,8
MHC class I protein binding (GO:0042288)	10	1,6	3,8
response to wounding (GO:0009611, 0002245)	50	7,9	3,8
positive regulation of immune response (GO:0050778)	6	1	3,8
positive regulation of immune system process (GO:0002684)	6	1	3,8

Obr.
paci
nický
IPA (I
data
barv

přežití (převzato z Yoshihara et al., 2012).

Predikce metastatického potenciálu u pacientek s časnými stádii mamárního karcinomu

Van't Veer et al. (Nature, 2002)

96 sporadických mamárních karcinomů

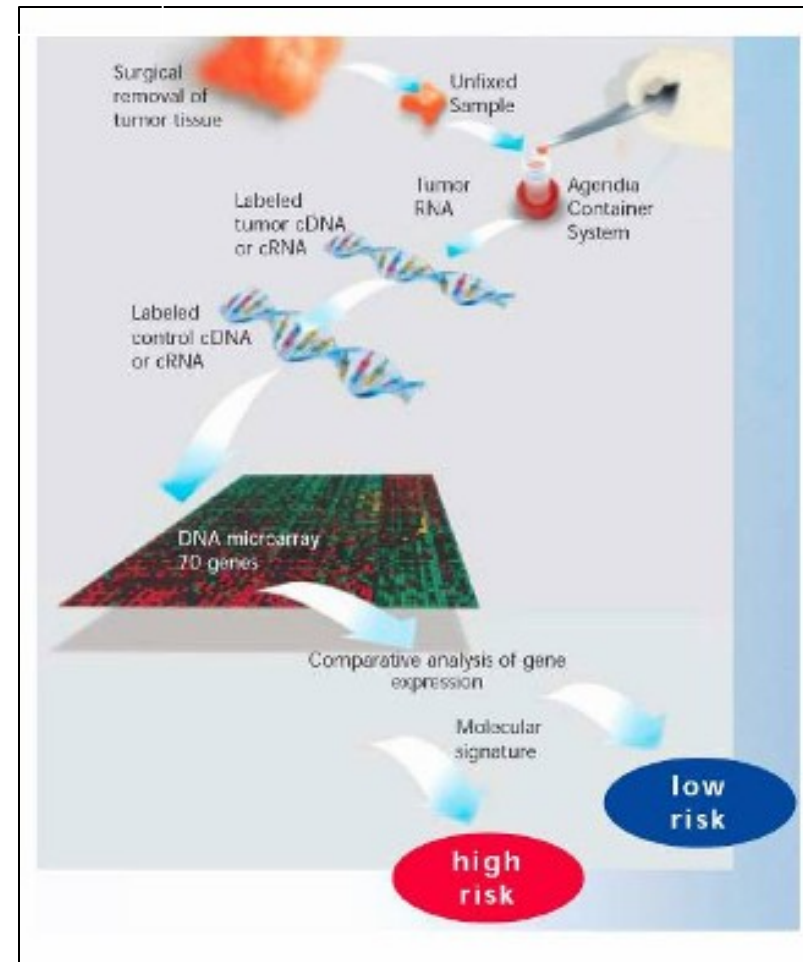
46 pacientek se špatnou prognózou (do 5 let se nevyvinuly vzdálené metastáze)

50 pacientek s dobrou prognózou (do 5 let se nevyvinuly vzdálené metastáze)

5852 genů se signifikantním rozdílem v expresi mezi skupinami

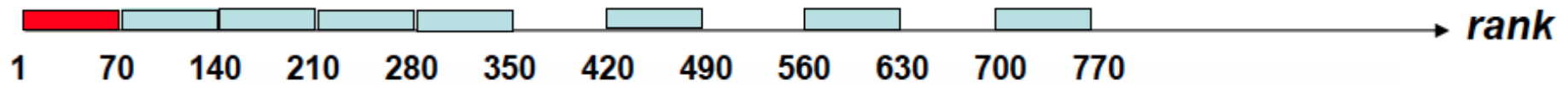
70 genů nejvíce korelujících s klinickým stavem použila pro klasifikaci

AGENDIA

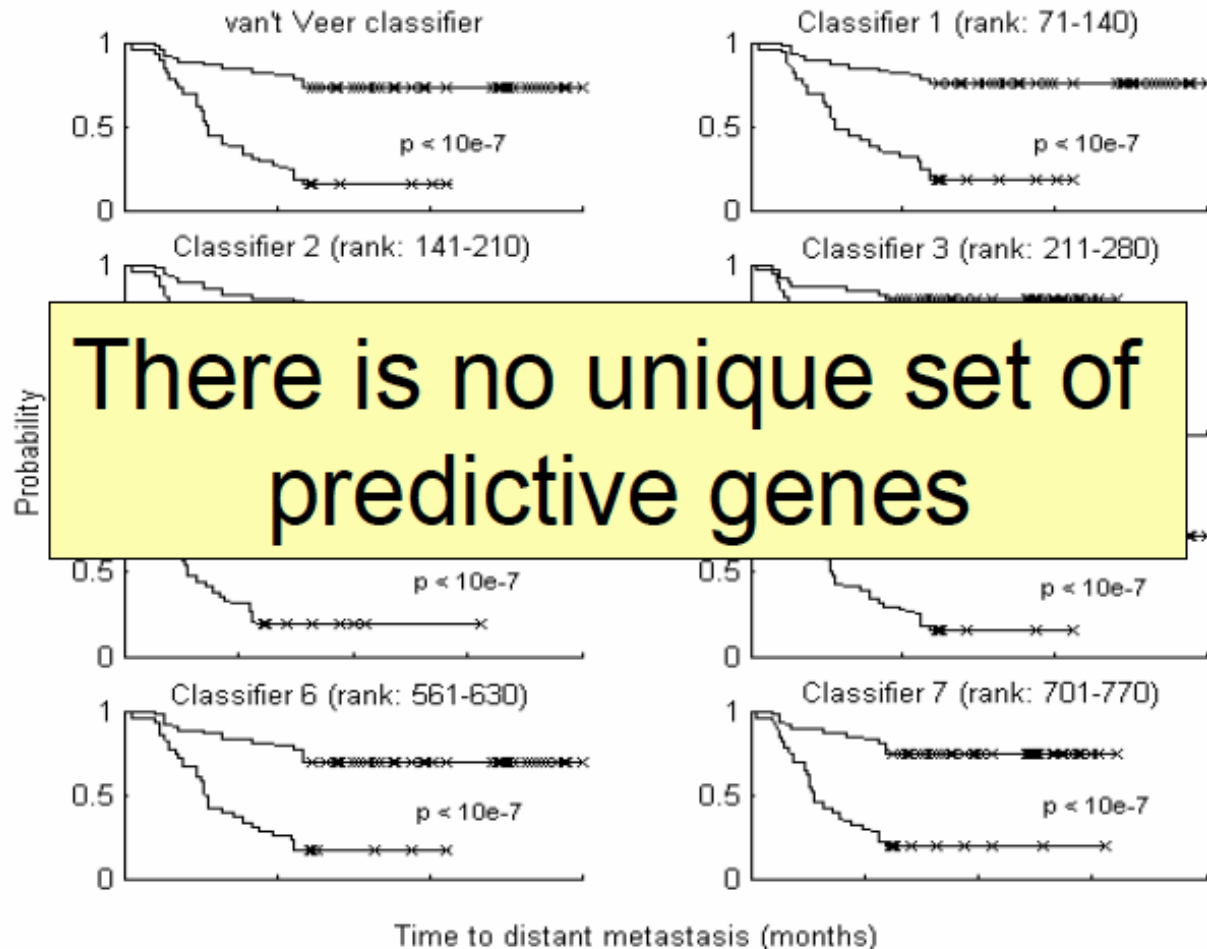


Sada 70 genů - patent

Many sets of 70 genes can be used to predict time to distance metastasis



Van't Veer

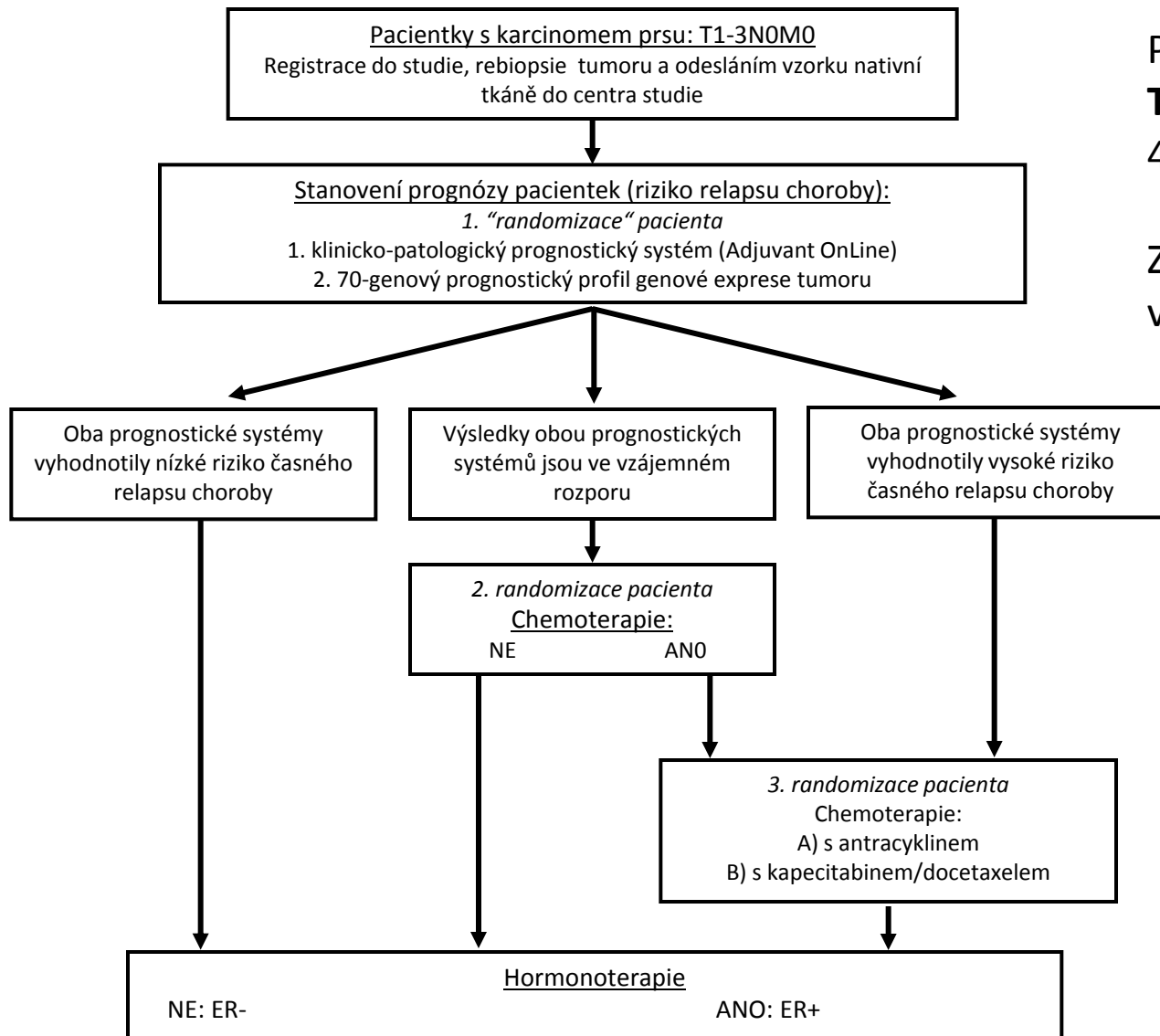


Results for Breast Cancer Data

- For a typical overlap of **50%** between two lists of **70** genes, more than **2300** patients are needed.
- The expected overlap between van't Veer's list and another list produced from similar experiment is less than **2%**

PROTOKOL KLINICKÉ STUDIE „MINDACT“

(Microarray In Node negative Disease may Avoid ChemoTherapy)



Rozpočet ~ 20 mil. EUR

Pořádá síť
TRANSBIG

40 institucí 21 zemí

Zařazení 5000 pacientek
v prvních 3 letech

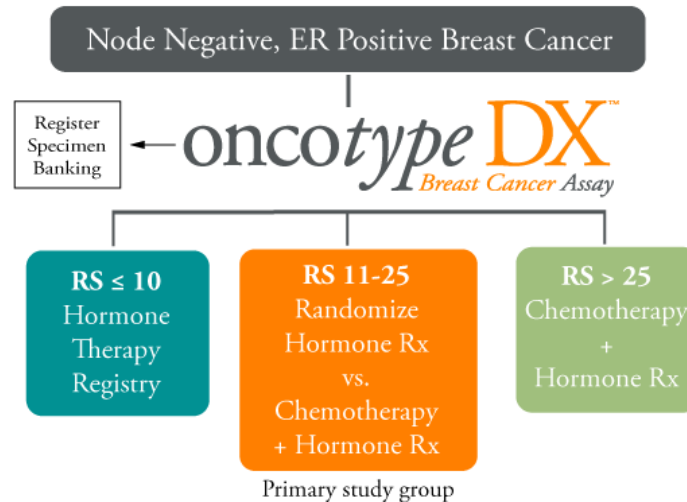
Roche AmpliChip CYP450

- Approved for clinical use in US and EU
- Gene variations of CYP2D6 and CYP2C19
 - Metabolism of ~25% of all prescription drugs
 - Determine phenotype: poor, intermediate, extensive, or ultrarapid metabolizer
- Intended to be an aid for physicians
 - Individualized treatment and dosage



www.roche.com
www.amplichip.u

Schema: TAILORx



Take home

Analýza čipových dat – pozadí, normalizace

Analýza čipových dat - identifikace biologicky významných genů

Analýza čipových dat -ukázky multidemenzionálních metod analýzy čipových dat - *Shlukovací analýzy*

Analýza čipových dat – klasifikační metody

Molekulární klasifikace nádorových onemocnění – ukázky

Aplikace čipových technologií do klinické praxe – studie MINDACT, Agendia, Roche AmpliChip CYP450

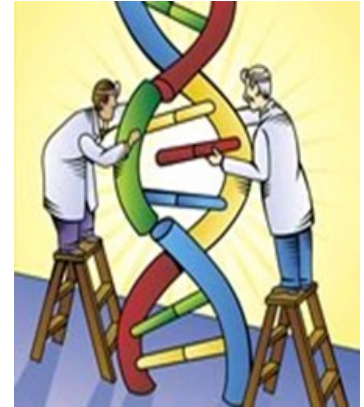
mikroRNA: nová úroveň regulace genové exprese – biogeneze a biologická funkce

mikroRNA v patogeneze nádorových onemocnění

mikroRNA jako biomarkery (SNP, tkáňové, sérové)

mikroRNA jako terapeutické cíle

mikroRNA čipy



Náplň příští přednášky

Moderní metodické přístupy v molekulární medicíně II – proteomika (dvojrozměrná elektroforéza, hmotnostní spektrometrie, proteinové čipy), využití proteomiky v diagnostice nádorových onemocnění

Molekulární epidemiologie – definice a vymezení oboru, identifikace molekulárních rizikových faktorů vzniku a rozvoje onemocnění, analýza vztahu molekulárních faktorů a vlivů prostředí na rozvoj nádorového onemocnění, význam molekulární epidemiologie u karcinomu plic a kolorektálního karcinomu

Dotazy?

