Pathway based personalized analysis of gene expression data: method and applications

Eytan Domany
Dept of Physics of Complex Systems
Weizmann Institute of Science, Rehovot, Israel

Yotam Drier

Michal Sheffer

Anna Livshits

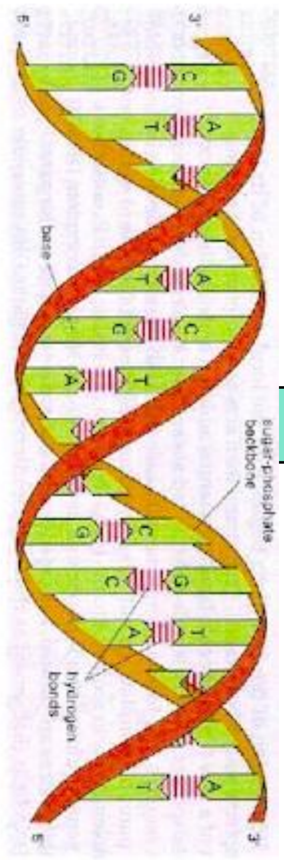
Gari Fuks

Carlos Caldas

Anna Git

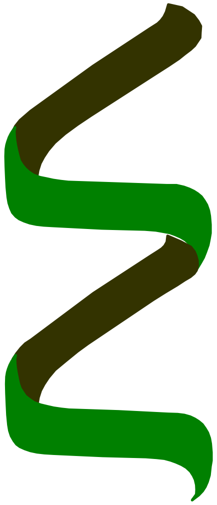
Bari Dec 2017

Central Dogma



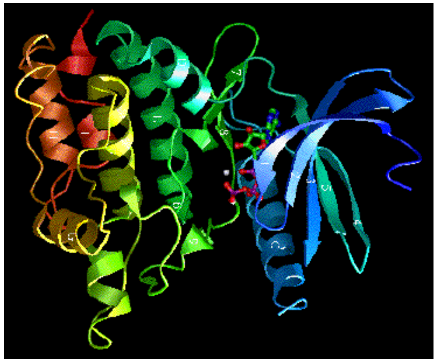
Information stored in Gene (segment of DNA)

Transcription



mRNA

Translation

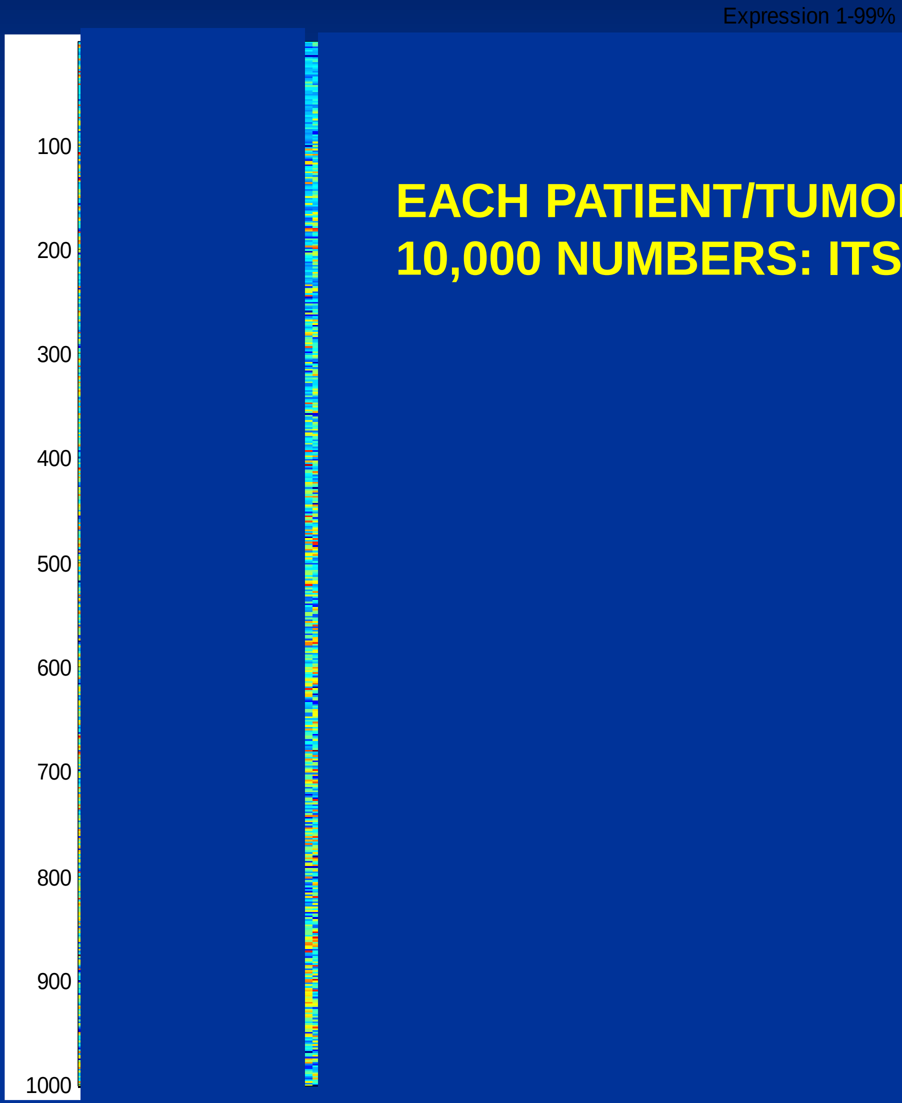


Protein

A gene is **expressed** when the mRNA it codes for is transcribed

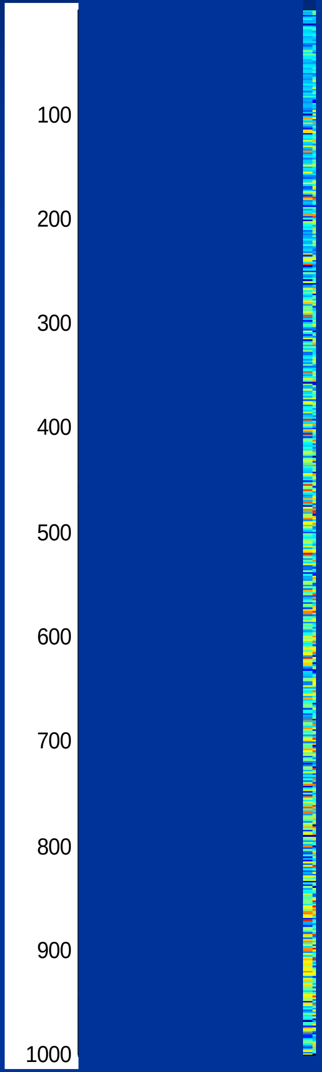
Cells express **different** subset of the genes (~5-10,000) in different tissues and under different conditions

MEASURE THE EXPRESSION LEVEL (mRNA) OF ~ 10,000 GENES



EACH PATIENT/TUMOR IS DESCRIBED BY 5 - 10,000 NUMBERS: ITS EXPRESSION PROFILE

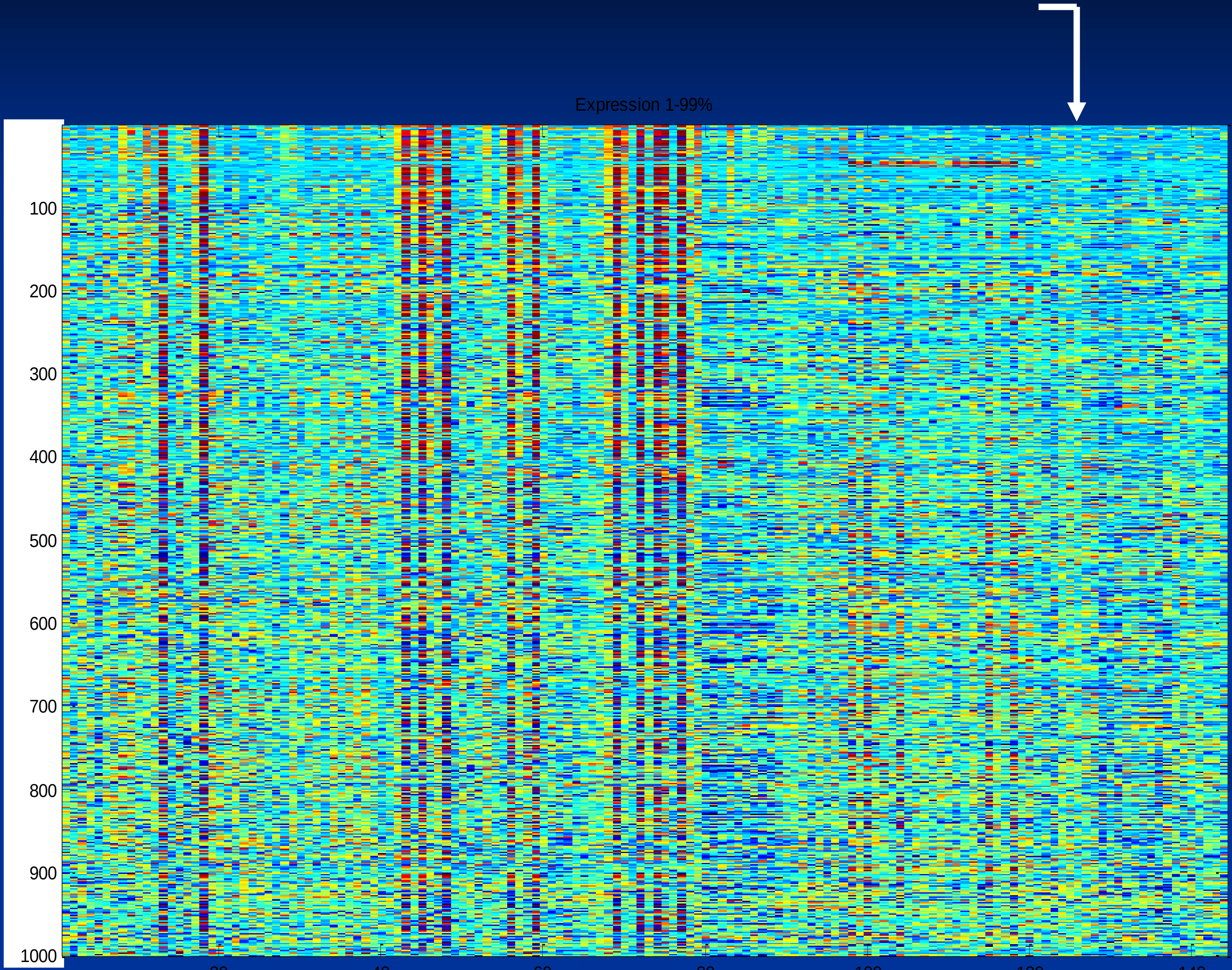
THE STANDARD METHOD: EXPRESSION – BASED ANALYSIS



EACH PATIENT/TUMOR IS DESCRIBED BY 5 - 10,000 NUMBERS: ITS EXPRESSION PROFILE

1000 NUMBERS – THE EXPRESSION LEVELS OF 1000 GENES IN A SAMPLE/TUMOR

THE STANDARD METHOD: EXPRESSION – BASED ANALYSIS

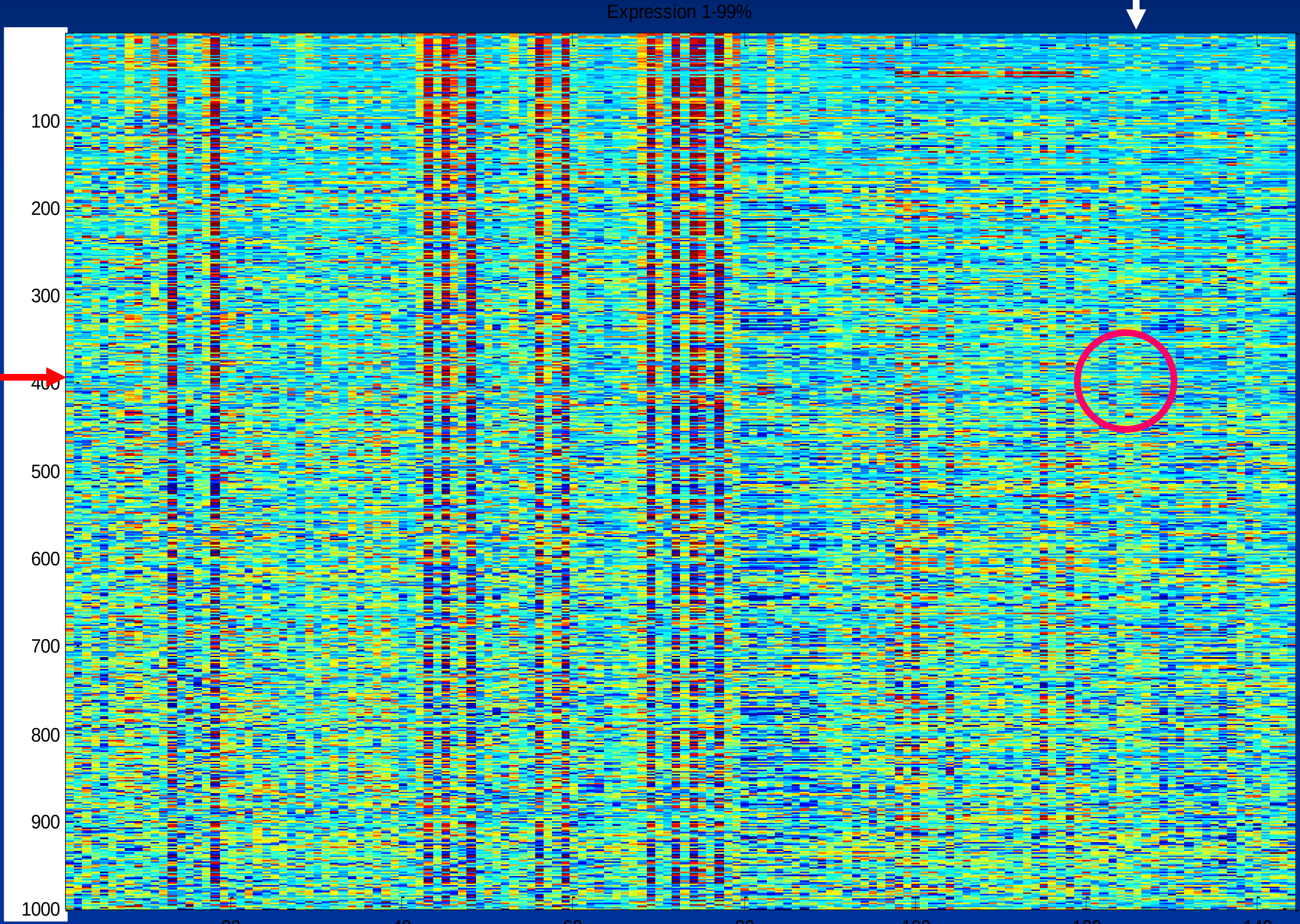


THE STANDARD METHOD: EXPRESSION – BASED ANALYSIS

E_{ij} = EXPRESSION LEVEL OF GENE i
IN SAMPLE j

Sample # 127

gene 400

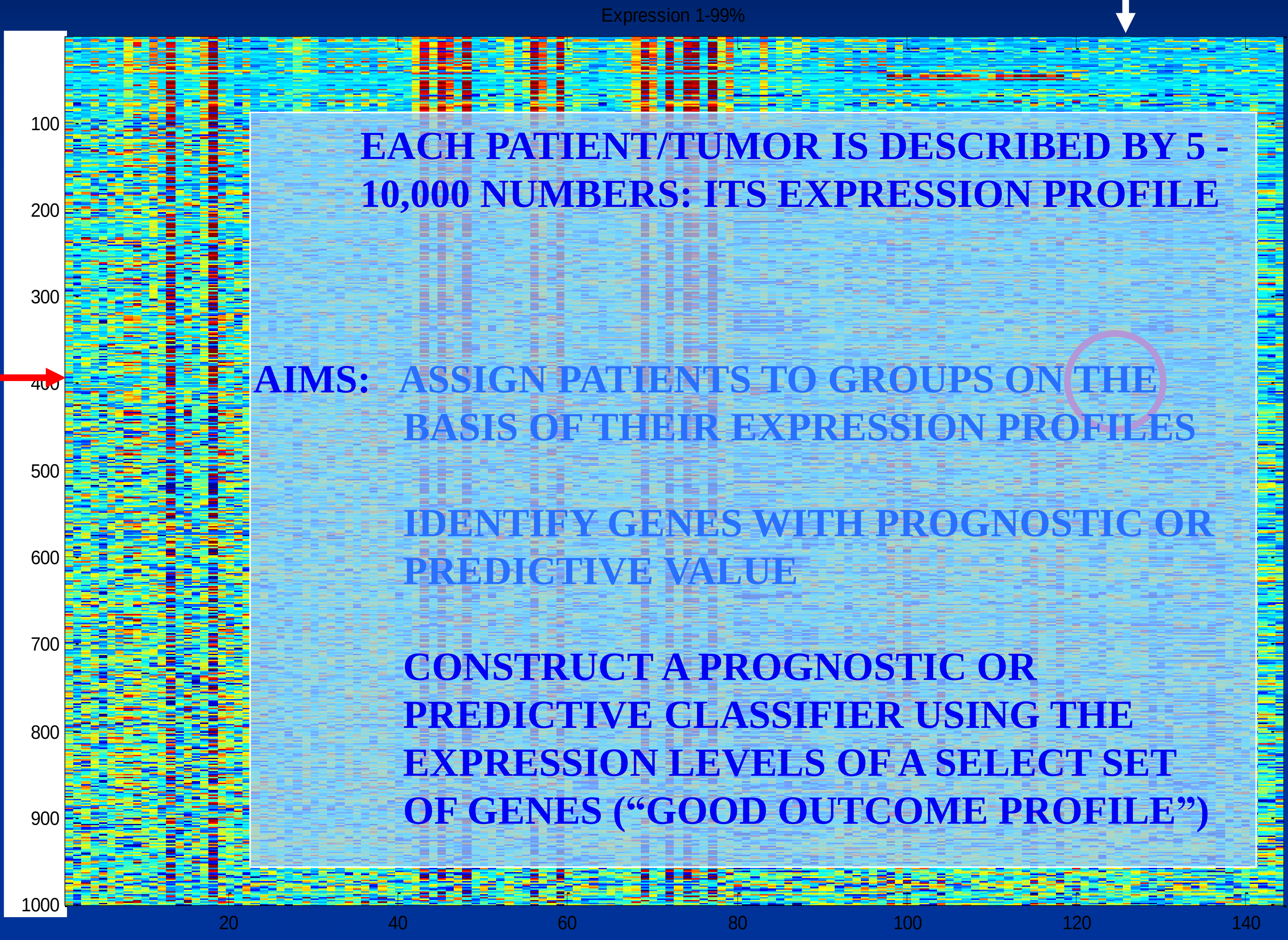


THE STANDARD METHOD: EXPRESSION – BASED ANALYSIS

E_{ij} = EXPRESSION LEVEL OF GENE *i*
IN SAMPLE *j*

Sample # 127

gene 400



EACH PATIENT/TUMOR IS DESCRIBED BY 5 - 10,000 NUMBERS: ITS EXPRESSION PROFILE

AIMS: **ASSIGN PATIENTS TO GROUPS ON THE BASIS OF THEIR EXPRESSION PROFILES**

IDENTIFY GENES WITH PROGNOSTIC OR PREDICTIVE VALUE

CONSTRUCT A PROGNOSTIC OR PREDICTIVE CLASSIFIER USING THE EXPRESSION LEVELS OF A SELECT SET OF GENES (“GOOD OUTCOME PROFILE”)

THE CHALLENGE:

PERSONALIZED PROGNOSTIC PREDICTIVE MEDICINE –

FOR BETTER TREATMENT OF CANCER

MEASURE (IN SAMPLE FROM TUMOR) GENOME- WIDE HIGH-THROUGHPUT DATA (MUTATIONS, GENE EXPRESSION, METHYLATION, SNP, DNA COPY NUMBER, ETC), AND USE FOR

1. PROGNOSIS (PREDICT OUTCOME, AGGRESSIVENESS)

2. PREDICT RESPONSE TO THERAPY

OF ***INDIVIDUAL*** PATIENTS/TUMORS

DESPITE OF GREAT IMPORTANCE AND 1000s of PAPERS, SO FAR – VERY LIMITED SUCCESS

FAILURES - WHY?:

SOME OF THE REASONS (1. CULTURAL AND 2. TECHNICAL):

1. THE FIELD WAS DOMINATED BY TWO EXTREMES:

   a. USE ***NO*** BIOLOGICAL/CLINICAL EXISTING KNOWLEDGE,
   (turn ignorance into a virtue)

   or

   b. DEMAND/ASSUME FULL DETAILED MECHANISTIC KNOWLEDGE
   (don’t dare talk to me unless you know and use all details)

2. FEW POINTS (TUMORS, 100 - 1000) IN HIGH DIMENSIONAL

   SPACES (GENES: 1000 – 10,000): ***“CURSE OF DIMENSIONALITY”***

“ATOMISTIC” APPROACH

WHAT’S WRONG WITH THIS CAR?:

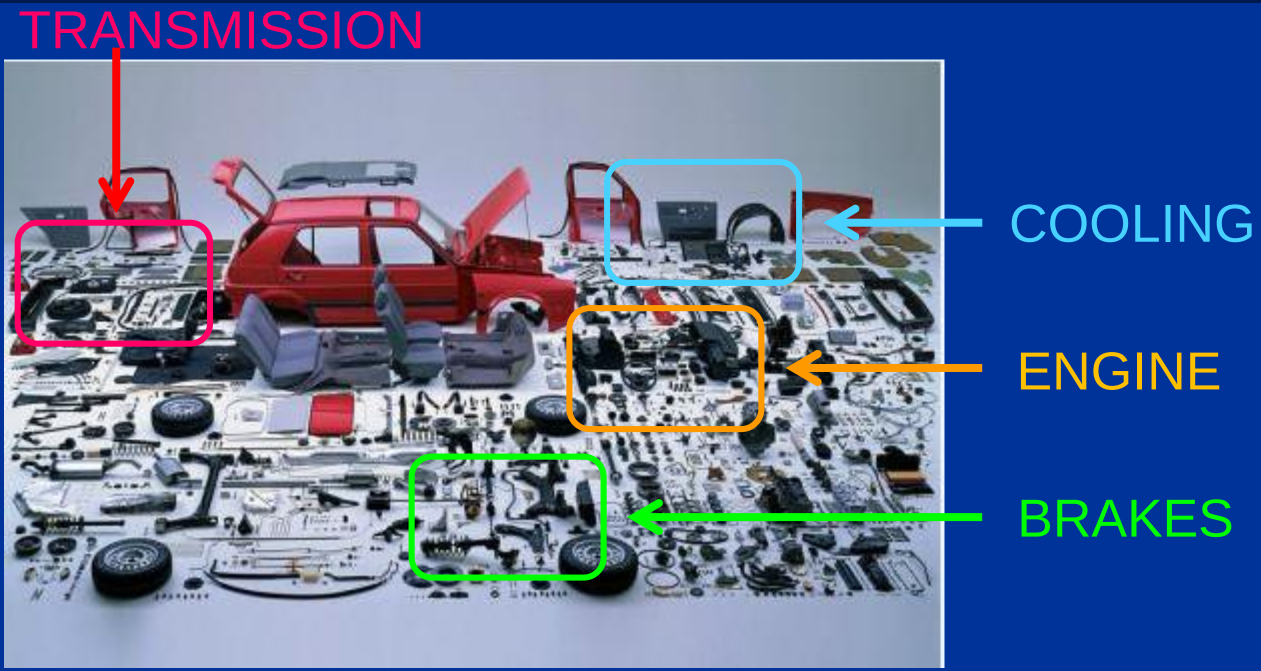
“ATOMISTIC” APPROACH:

MEASURE SOME PROPERTY (e.g. TEMPERATURE) OF EVERY SINGLE COMPONENT – 12,000 NUMBERS CHARACTERIZE THE “STATE “ OF EACH CAR

TRY TO DETERMINE THE FEATURES THAT CAN BE USED TO TELL HEALTHY CARS FROM SICK ONES.

NO EXISTING KNOWLEDGE ABOUT CARS IS USED

A “PHENOMENOLOGICAL” “SYSTEMS” APPROACH



TRANSMISSION

COOLING

ENGINE

BRAKES

MEASURE FOR EACH **SYSTEM** **ONE** NUMBER, THAT INDICATES THE DEVIATION OF *THIS SYSTEM’S* FUNCTIONING FROM NORMAL .

EACH CAR IS CHARACTERIZED BY A SET OF SUCH “SYSTEM-LEVEL INDICATORS” (ABOUT 100) - USE *THESE* TO SEPARATE HEALTHY FROM SICK CARS

A “PHENOMENOLOGICAL” “SYSTEMS” APPROACH

TRANSMISSION

CARS	cells
COOLING	heat shock proteins
ENGINE	metabolism, growth
BRAKES	growth arrest, apoptosis

MEASURE FOR EACH **SYSTEM** **ONE** NUMBER, THAT INDICATES THE DEVIATION OF *THIS SYSTEM’S* FUNCTIONING FROM NORMAL .

EACH CAR IS CHARACTERIZED BY A SET OF SUCH “SYSTEM-LEVEL INDICATORS” (ABOUT 100) - USE *THESE* TO SEPARATE HEALTHY FROM SICK CARS

PATHWAY (OR - BIOLOGICAL PROCESS) – BASED ANALYSIS: THE IDEA

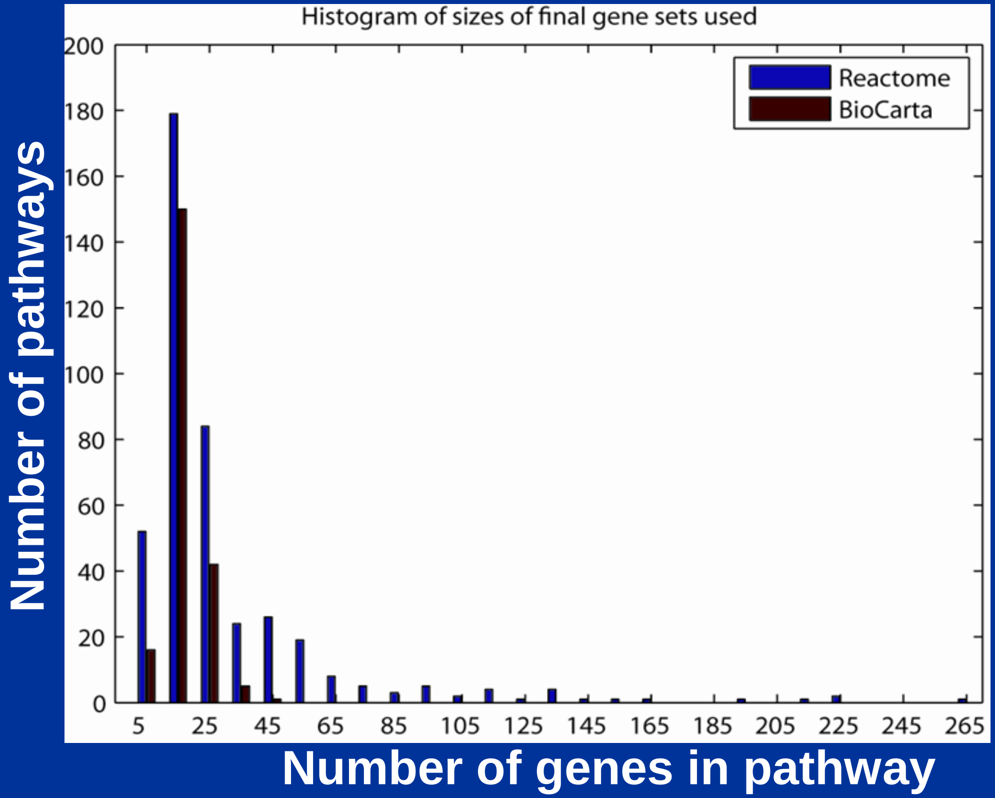
a. USE EXPRESSION (OR ANY OTHER) HIGH-THROUGHPUT DATA FROM A LARGE NUMBER OF SAMPLES.

b. USE BIOLOGICAL KNOWLEDGE – LISTS OF (10 - 100) GENES THAT BELONG TO A BIOLOGICAL PROCESS OR PATHWAY *P*

Drier, Sheffer & Domany *PNAS 2013*

b. USE EXISTING KNOWLEDGE - ASSIGNMENT OF GENES TO PATHWAYS *P*

USE *KEGG*, *BioCarta* FROM *MSigDB*, AND *NCI-Nature Pathway Interaction* DATABASES



TYPICALLY – TENS OF GENES IN A PATHWAY; HUNDREDS OF SAMPLES

“CURSE OF DIMENSIONALITY” IS ELIMINATED

PATHWAY (OR - BIOLOGICAL PROCESS) – BASED ANALYSIS: THE IDEA

a. USE EXPRESSION (OR ANY OTHER) HIGH-THROUGHPUT DATA FROM A LARGE NUMBER OF SAMPLES.

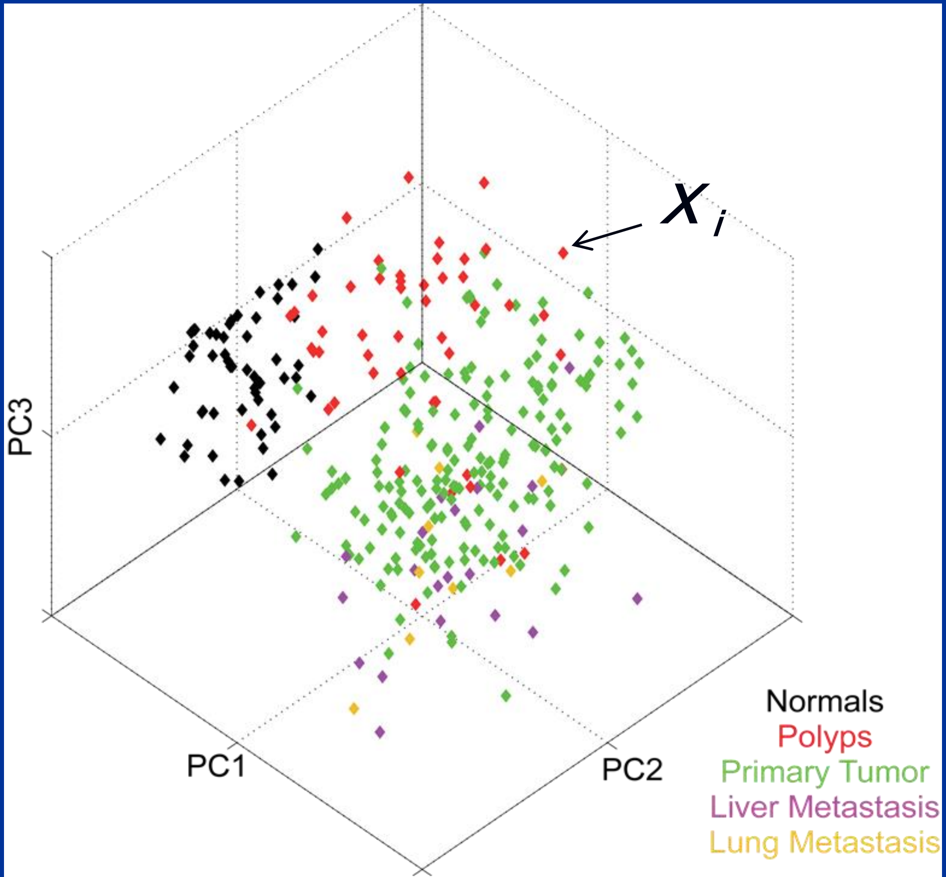
b. USE BIOLOGICAL KNOWLEDGE – LISTS OF (10 - 100) GENES THAT BELONG TO A BIOLOGICAL PROCESS OR PATHWAY *P*

c. DERIVE FOR EACH SAMPLE $\mathbf{i}$ AND PATHWAY *P* A “PATHWAY DEREGULATION SCORE” $D(\mathbf{i},P)$

Drier, Sheffer & Domany *PNAS 2013*

c. FOR EACH SAMPLE *i* AND PATHWAY *P* - CALCULATING THE PATHWAY DEREGULATION SCORE (PDS)

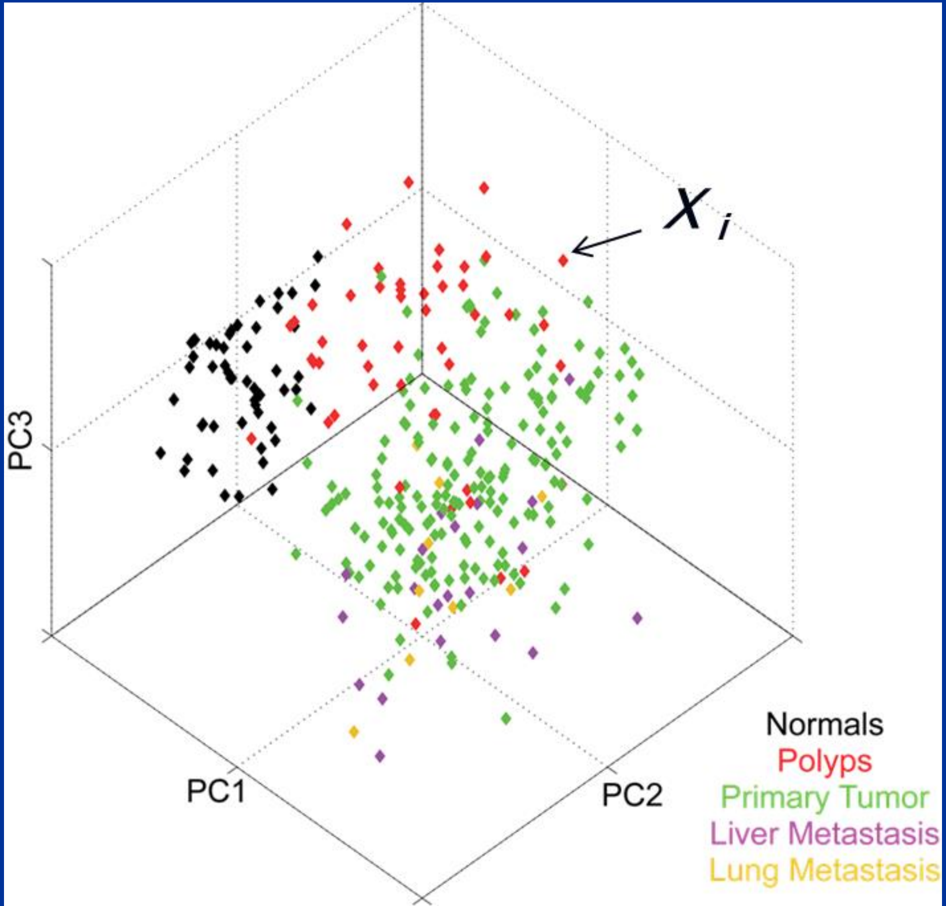
1. Consider pathway *P; identify* d_P genes that belong to it. Sample *i* is represented by a point X_i in the space of the expression values of these genes

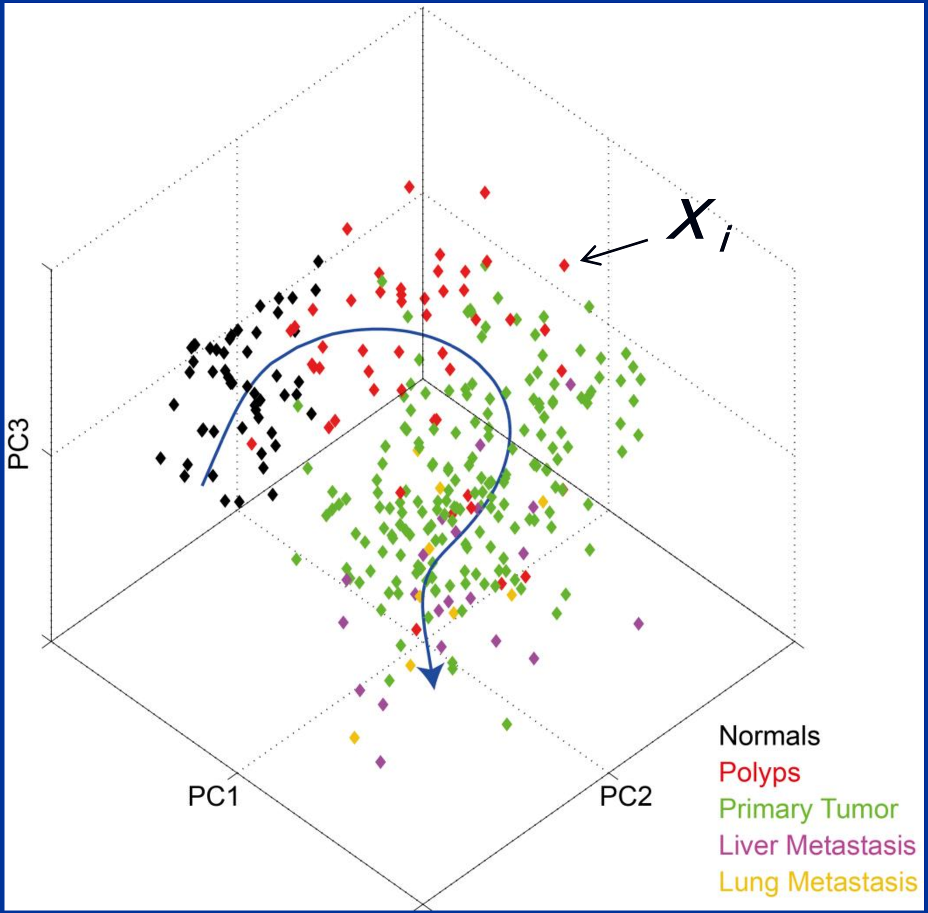


KEGG APOPTOSIS PATHWAY,
$d_P = 33$ GENES, COLON DATA

c. PATHWAY DEREGULATION SCORE (PDS)

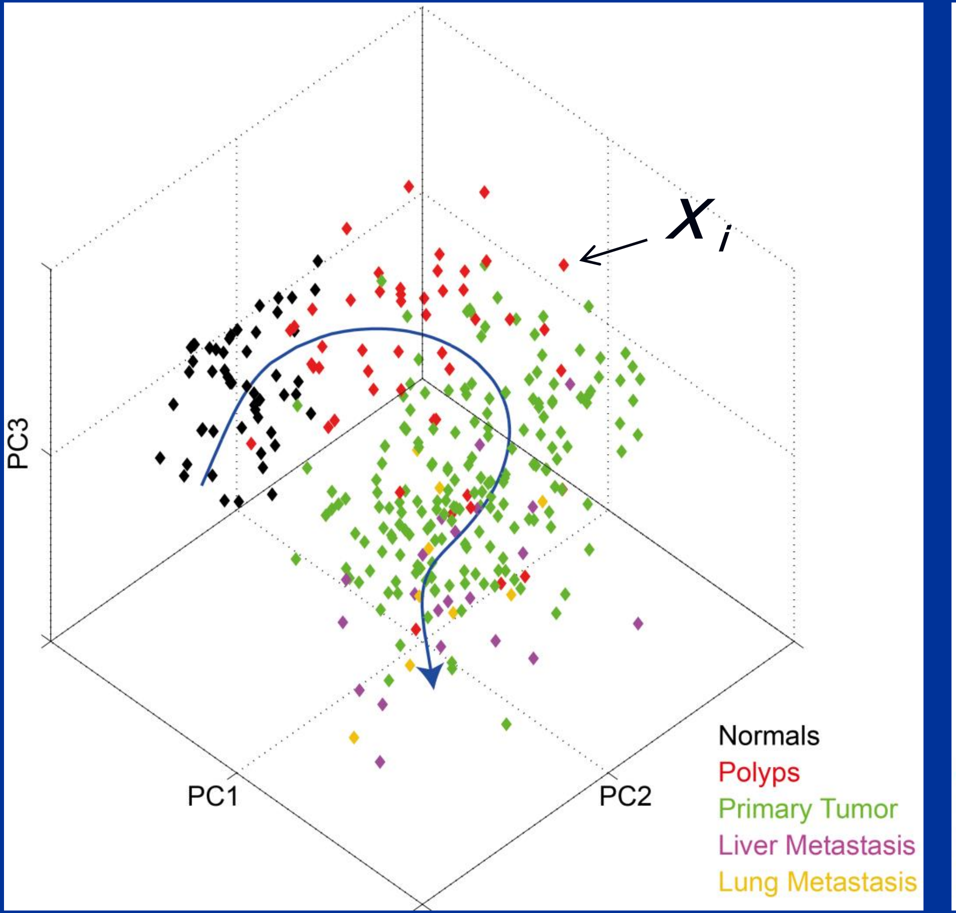
2. Calculate the *Principal Curve (Hastie & Stuezle 1989)* of the cloud of points formed by the full sample set

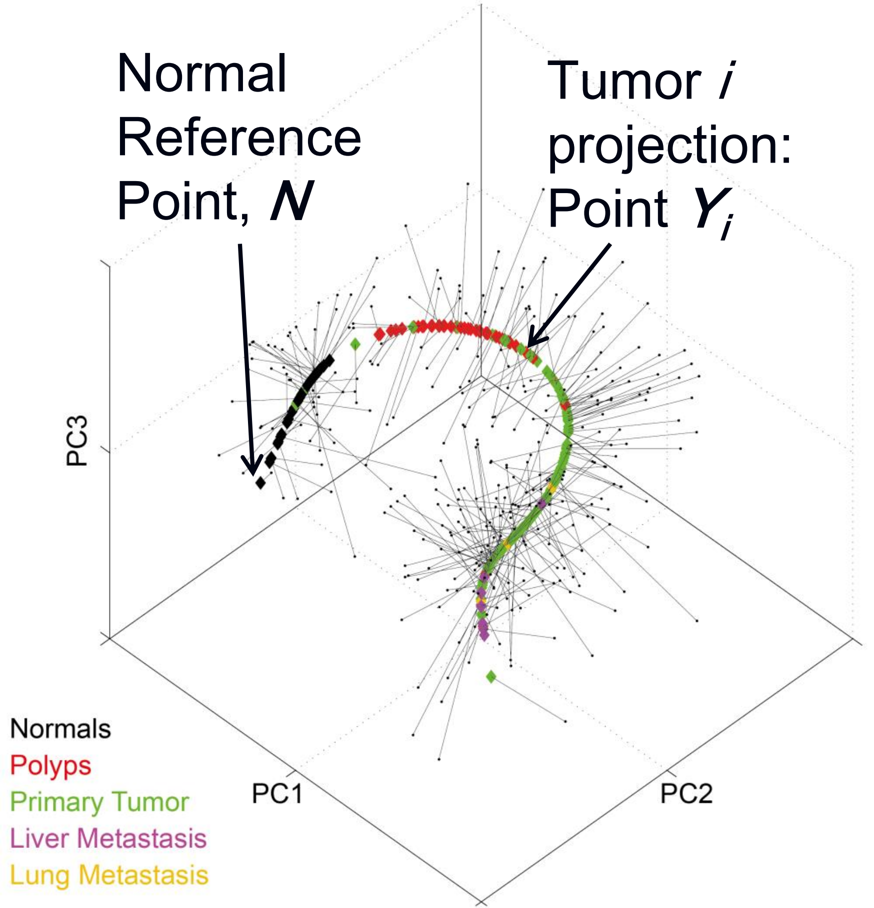




c. PATHWAY DEREGULATION SCORE (PDS)

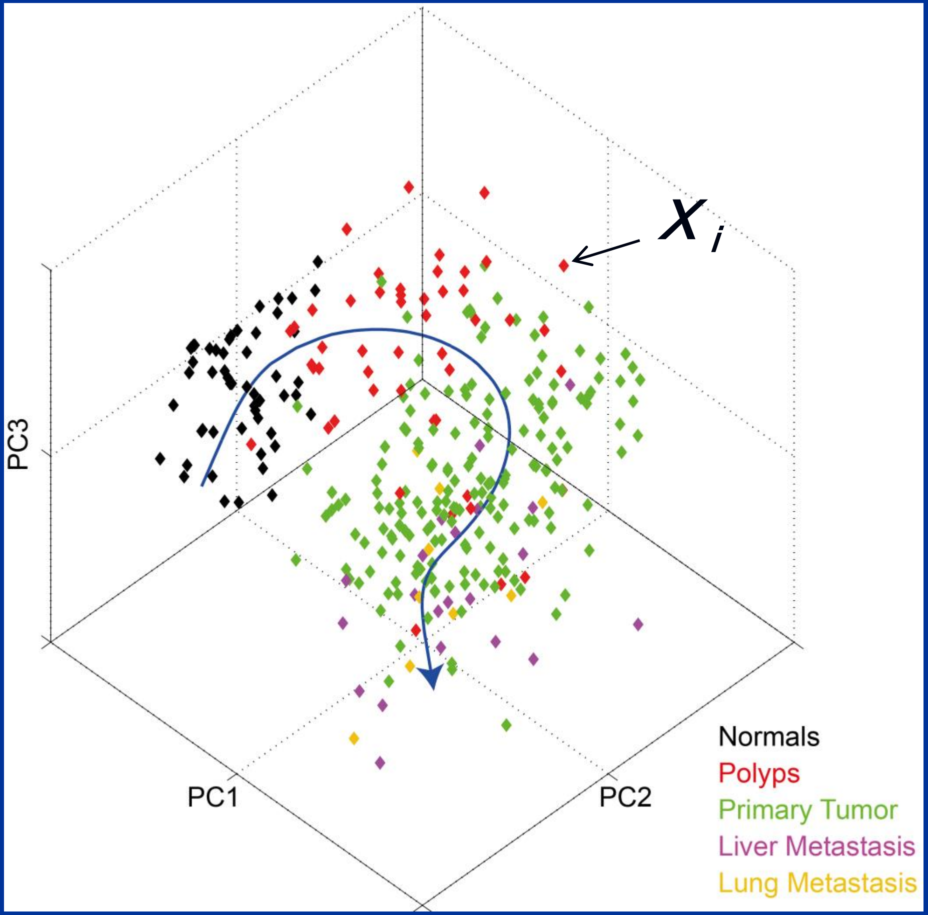
3. Project every sample onto the principal curve; projection of sample *i* is Y_i. The projection to the extremal point near the Normal samples is the Reference Point ***N***

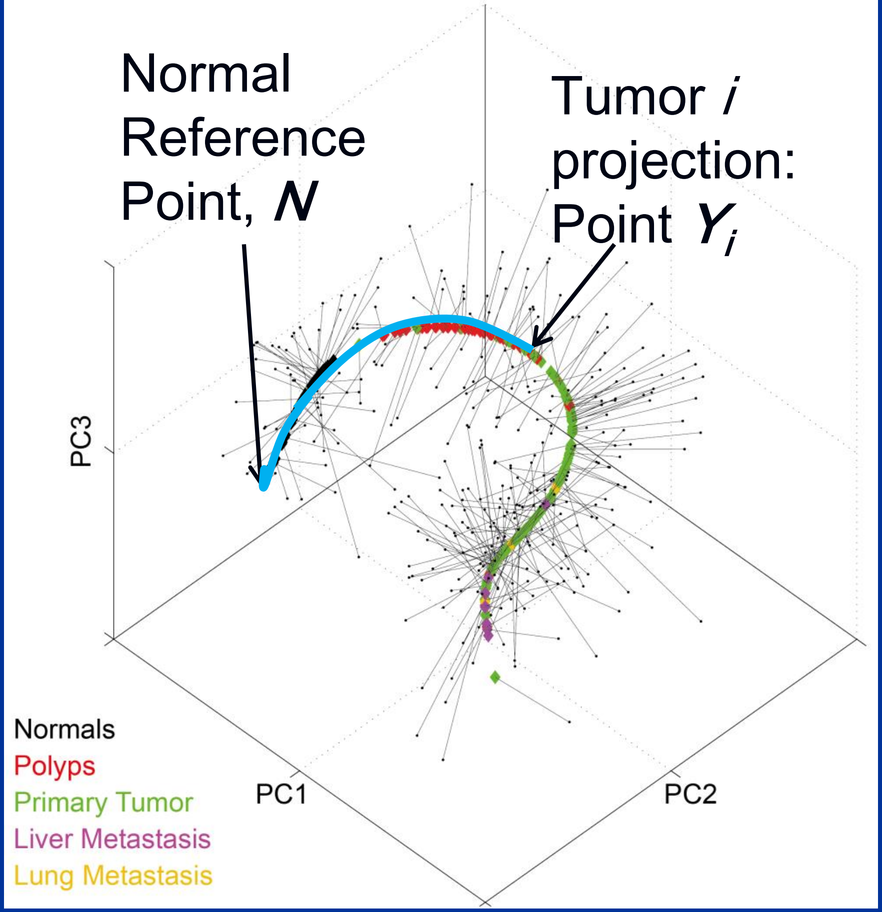




c. PATHWAY DEREGULATION SCORE (PDS)

4. The ***distance*** of Y_i from *N*, ***measured along the principal curve,*** is $D_i(P)$, the Deregulation Score of pathway *P in* sample *i.*





PATHWAY (OR - BIOLOGICAL PROCESS) – BASED ANALYSIS: THE IDEA

a. USE EXPRESSION (OR ANY OTHER) HIGH-THROUGHPUT DATA FROM A LARGE NUMBER OF SAMPLES.

b. USE BIOLOGICAL KNOWLEDGE – LISTS OF (10 - 100) GENES THAT BELONG TO A BIOLOGICAL PROCESS OR PATHWAY P

c. DERIVE FOR EACH SAMPLE i AND PATHWAY P A "PATHWAY DEREGULATION SCORE" $D(i,P)$

d. DO THIS FOR N_P ~ FEW HUNDRED PATHWAYS

e. A SAMPLE IS REPRESENTED IN TERMS OF ITS N_P PATHWAY DEREGULATION SCORES => DESCRIBED BY N_P PARAMETERS

f. PERFORM ALL ANALYSIS USING THESE "*SYSTEM-LEVEL*" *VARIABLES WITH CLEAR BIOLOGICAL MEANING.*

Drier, Sheffer & Domany *PNAS 2013*

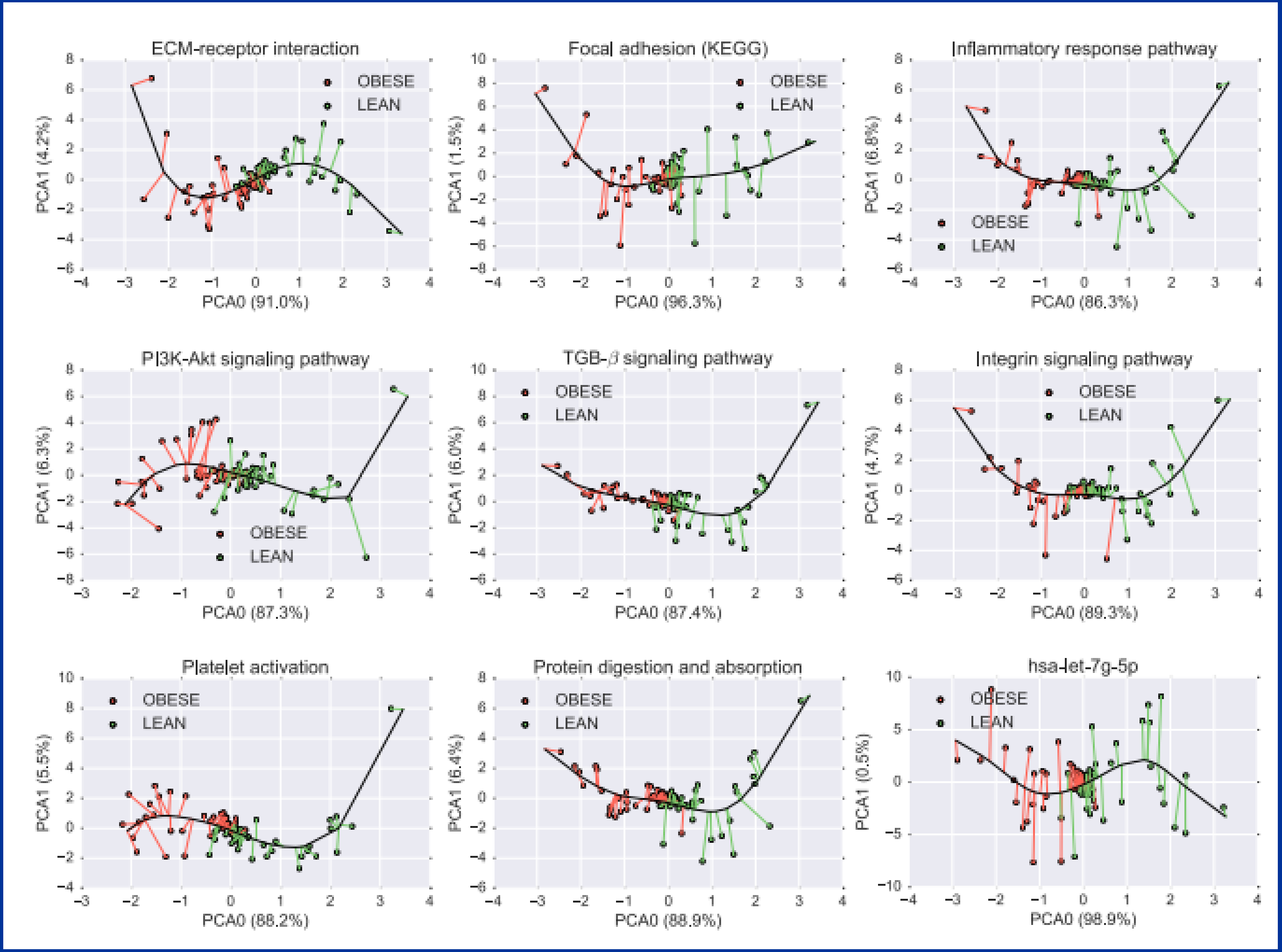
PATHWAY DEREGULATION IN OBESITY*

EXPRESSION DATA FROM 39 LEAN & 49 OBESE SUBJECTS

1. IDENTIFY 38 DIFFERENTIALLY EXPRESSED GENES
2. ENRICHMENT ANALYSIS: 16 PATHWAYS HAVE >2 OF THEIR GENES AMONG THE 38
3. PATHIFIER ANALYSIS OF THE 16 PATHWAYS SHOWS CLEAR SEPARATION IN DEREGULATION OF LEAN vs OBESE SUBJECTS

***Font-Clos, Zapperi, La Porta Sys Bio & App 2017**

PATHWAY DEREGULATION IN OBESITY* - NEW OBESITY RELATED PATHWAYS



Font-Clos, Zapperi, La Porta Sys Bio & App 2017

DNA REPAIR PATHWAY DEREGULATION IN BREAST CANCER*

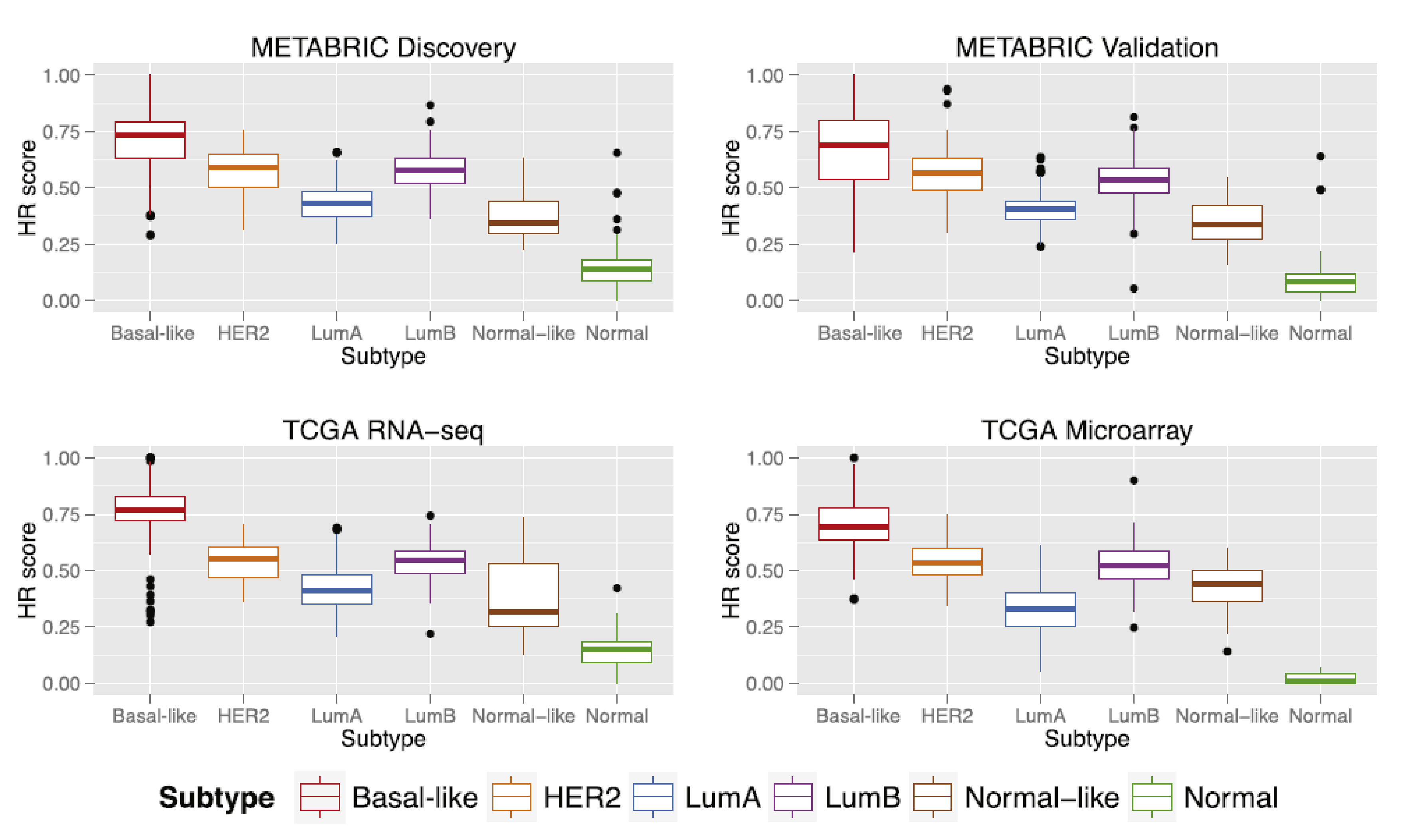
EXPRESSION DATA ~ 4000 BREAST CANCER PATIENTS (4 DATASETS)

1. MANUALLY CURATED LIST OF 82 GENES ASSOCIATED WITH HOMOLOGOUS RECOMBINATION (***HR***) - CRUCIAL FOR REPAIR OF DOUBLE STRANDED DNA BREAK

2. PATHIFIER ANALYSIS OF THE ***HR*** PATHWAY => HR SCORE FOR EACH SAMPLE, ***ROBUST*** *ACROSS FOUR DATASETS!*

****Liu et al Mol Onc 2016***

HOMOLOGOUS RECOMBINATION (HR) SCORE - ROBUST



DNA REPAIR PATHWAY DEREGULATION IN BREAST CANCER*

EXPRESSION DATA ~ 4000 BREAST CANCER PATIENTS (4 DATASETS)

1. MANUALLY CURATED LIST OF 82 GENES ASSOCIATED WITH HOMOLOGOUS RECOMBINATION (***HR***) - CRUCIAL FOR REPAIR OF DOUBLE STRANDED DNA BREAK

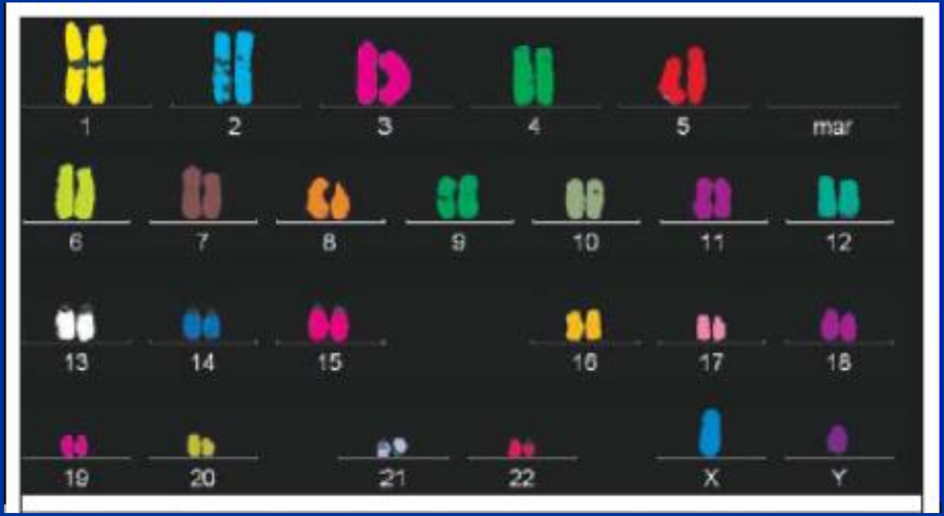
2. PATHIFIER ANALYSIS OF THE ***HR*** PATHWAY => HR SCORE FOR EACH SAMPLE, ***ROBUST*** *ACROSS FOUR DATASETS!*

3. FINDINGS: a. HR SCORE REFLECTS HR REPAIR DEFICIENCY
   b. HR SCORE IS ASSOCIATED WITH *CHROMOSOMAL INSTABILITY*

****Liu et al Mol Onc 2016***

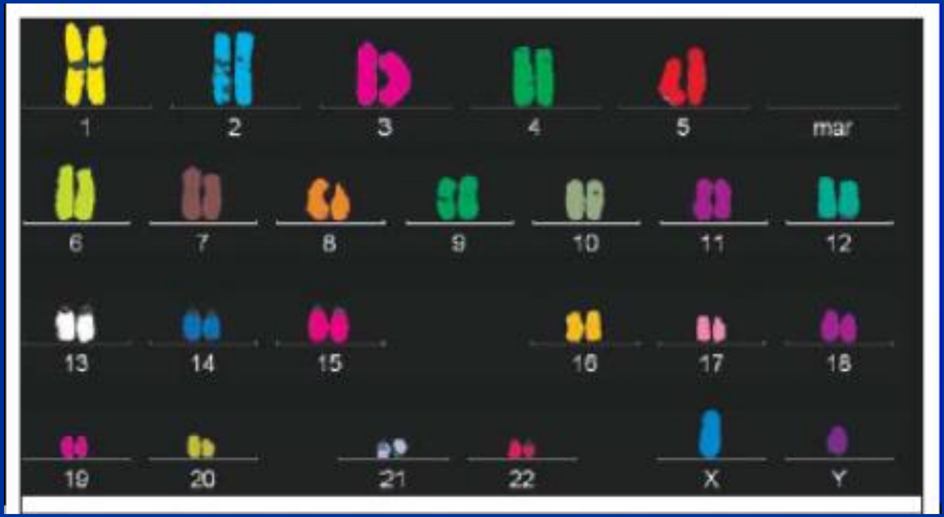
CHROMOSOMAL INSTABILITY

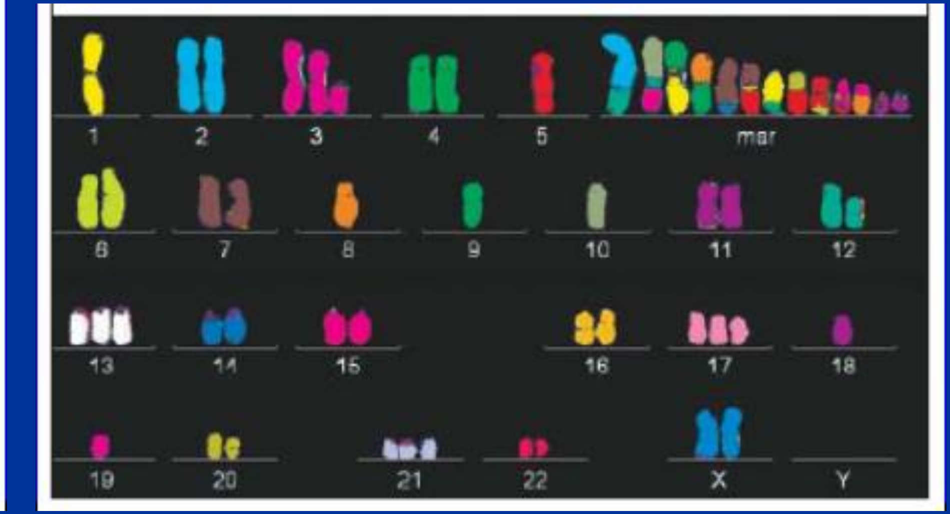
NORMAL CELLS MAINTAIN A VERY STABLE KARYOTYPE (SET OF CHROMOSOMES)



CHROMOSOMAL INSTABILITY

NORMAL CELLS MAINTAIN A VERY STABLE KARYOTYPE (SET OF CHROMOSOMES)

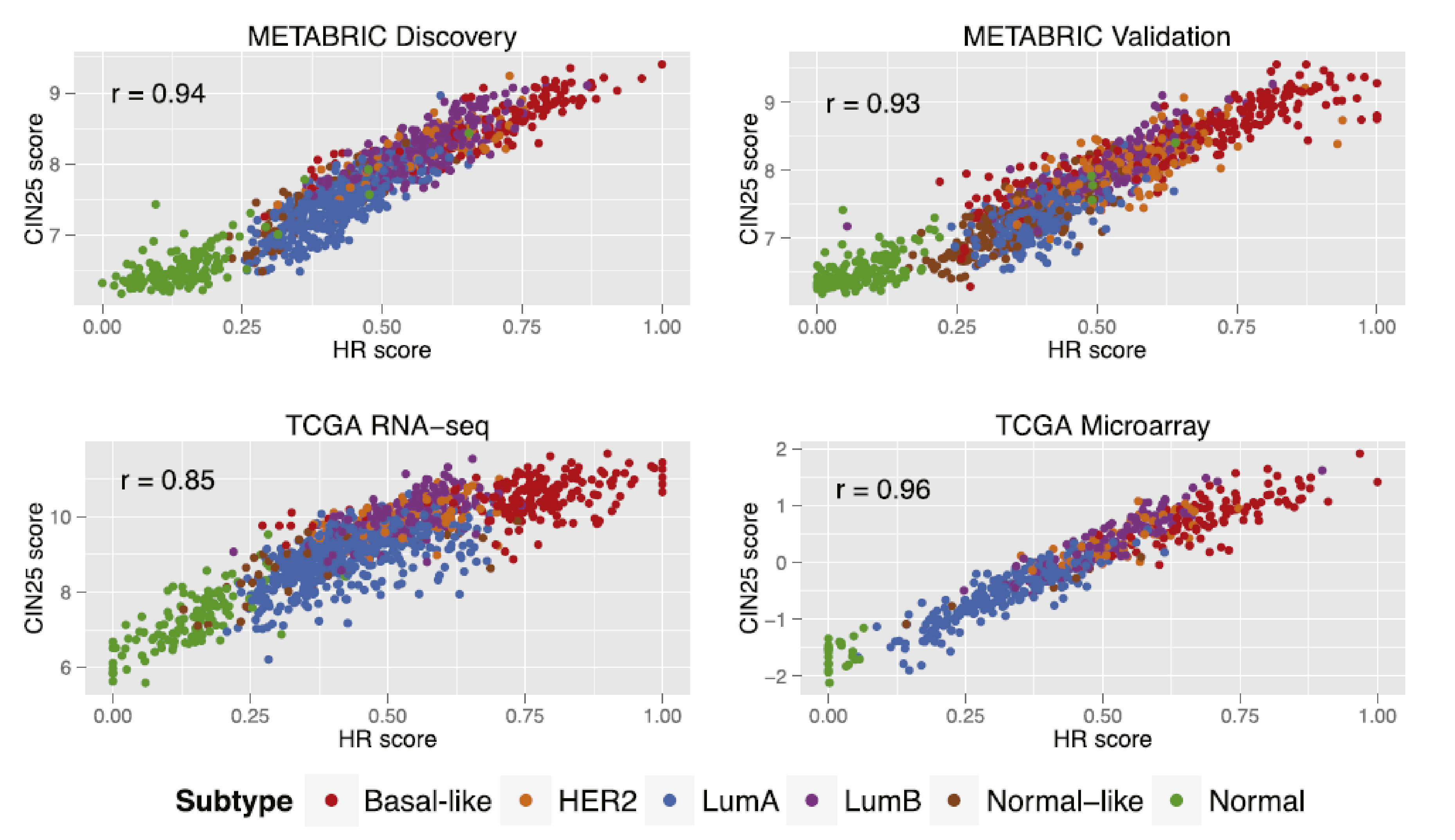




CANCER CELLS EXHIBIT ABNORMAL CHROMOSOME COPY NUMBERS (***ANEUPLOIDY***, ***von Hansemann 1890***)

CHROMOSOMAL INSTABILITY (CIN)

HR SCORE - ASSOCIATED WITH CHROMOSOMAL INSTABILITY



DNA REPAIR PATHWAY DEREGULATION IN BREAST CANCER*

EXPRESSION DATA ~ 4000 BREAST CANCER PATIENTS (4 DATASETS)

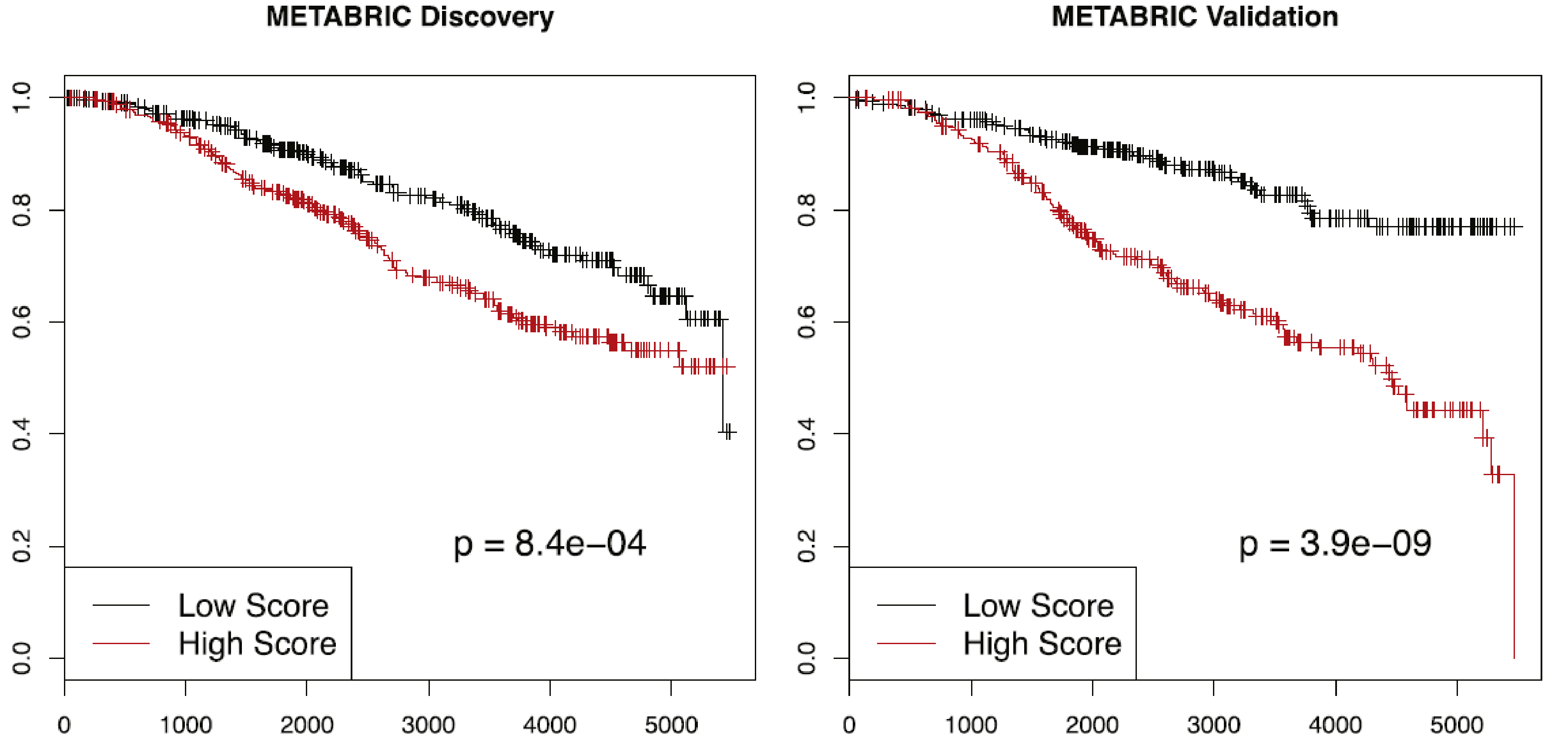
1. MANUALLY CURATED LIST OF 82 GENES ASSOCIATED WITH HOMOLOGOUS RECOMBINATION (***HR***) - CRUCIAL FOR REPAIR OF DOUBLE STRANDED DNA BREAK

2. PATHIFIER ANALYSIS OF THE ***HR*** PATHWAY => HR SCORE FOR EACH SAMPLE, ***ROBUST*** *ACROSS FOUR DATASETS!*

3. FINDINGS: a. HR SCORE REFLECTS HR REPAIR DEFICIENCY
   b. ASSOCIATION OF HR SCORE WITH *CHROMOSOMAL INSTABILITY*
   c. HIGH HR SCORE => WORSE *SURVIVAL*

***Liu et al Mol Onc 2016**

HIGHER HR SCORE => WORSE SURVIVAL



PROGNOSIS IN BREAST CANCER*

“CLASSSICAL” MACHINE LEARNING APPROACH – TRAINING SET, FIT KNOWN RECURRENCE TIME y_i OF PATIENTS $i=1,2,\ldots N$, AS A FUNCTION OF KNOWN VARIABLES $X_{i,k}$, $k=1,2\ldots K$ (GENE EXPRESSION, CLINICAL, etc)

OUTCOME (GOOD/BAD) = THRESHOLD ON y

USED 236 PATIENTS AS TRAINING SET & HAD 3 TEST SETS (606 PATIENTS

STANDARD METHOD USES THE EXPRESSION VALUES OF K GENES

Huang et al* CALCULATE PATHWAY DEREGULATION SCORES AND USE THESE AS THE VARIABLES $X_{i,k}$. 15 PATHWAYS ARE SELECTED (L1 LASSO)

*THE PATHWAY-BASED **PROGNOSTIC** PREDICTOR OUTPERFORMS THE STANDARD GENE-BASED PREDICTORS (PAM 50, Mammaprint 70)*

****Huang et al PLoS Comp Bio 2014***

DIAGNOSIS IN BREAST CANCER*

AT DIAGMOSIS MOST BREAST TUMORS HAVE SPREAD TO LYMPH NODES

SENSITIVITY (DISCOVERY RATE) OF MAMMOGRAPHY 54 – 77%

EARLY DISCOVERY => GOOD PROGNOSIS:
NEED ACCURATE, LOW COST, NON-INVASIVE DIAGNOSTIC METHOD

PATHWAY BASED DIAGNOSIS, USING METABOLIC PATHWAYS

1. USE PLASMA & SERUM TO PROFILE BLOOD METABOLITES (MS) AND RNAseq EXPRESSION DATA FROM RESECTED TUMORS

2. CALCULATE PATHWAY DEREGULATION SCORES FOR ~300 METABOLIC PATHWAYS,

3. CONSTRUCT DIAGNOSTIC CLASSIFIER (3 - 8 PATHWAYS SELECTED)

****Huang et al Genome Medicine 2016***

DIAGNOSIS IN BREAST CANCER*

THE RESULTING DIAGNOSTIC MODELS HAD OUTSTANDING PERFORMANCE, ROBUSTNESS (TRAINED ON MASS-SPEC METABOLOMIC DATA FROM PLASMA, TESTED ON SIMILAR DATA ON SERUM AND EXPRESSION DATA FROM TISSUE)
AUC > 0.9, SENSITIVITY & SPECIFICITY > 0.9

DISCOVERED NEW DIAGNOSTIC METABOLIC PATHWAYS FOR EARLY STAGE BREAST CANCER DIAGNOSIS:

TAURINE & HYPOTAURINE METABOLIC PATHWAYS MOST PREDICTIVE
ALANINE, ASPARTATE & GLUTAMINE METABOLISM
PROTEIN DIGESTION AND ABSORPTION ….

****Huang et al Genome Medicine 2016***

THE METABRIC BREAST CANCER DATASET ***Curtis et al Nature 2012***

Using expression data from 1992 TUMOR and 144 NORMAL samples

(997 in “Discovery set”, 995 in “Validation”)

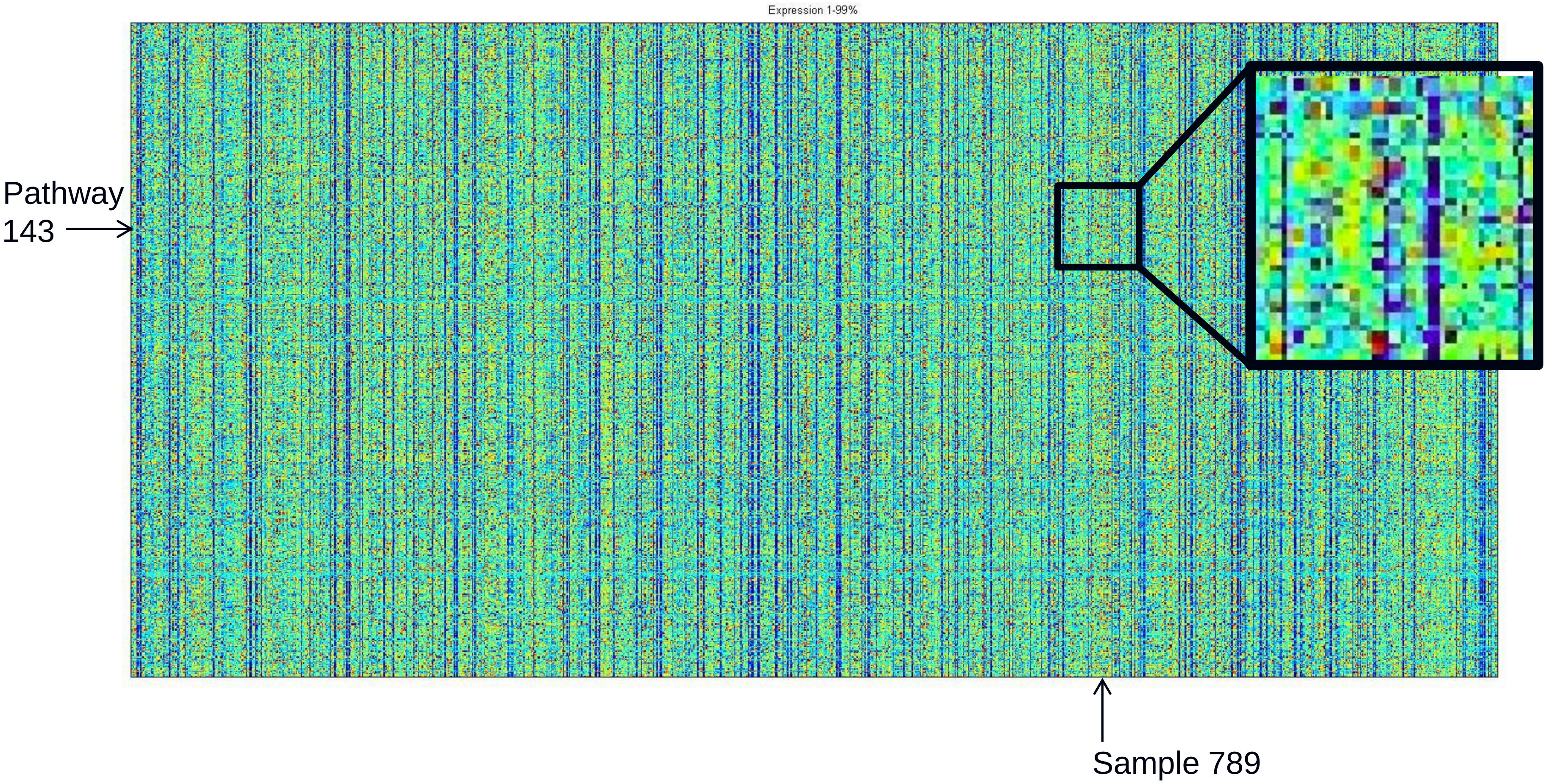
Calculate (using “Pathifier” analysis*) a *Pathway Deregulation Score (PDS)*

for 552 pathways/biological processes, for each sample (Discovery + Normal)

D(P,i) = PDS of pathway P in sample i – represent the extent to which pathway
P is deregulated in sample i

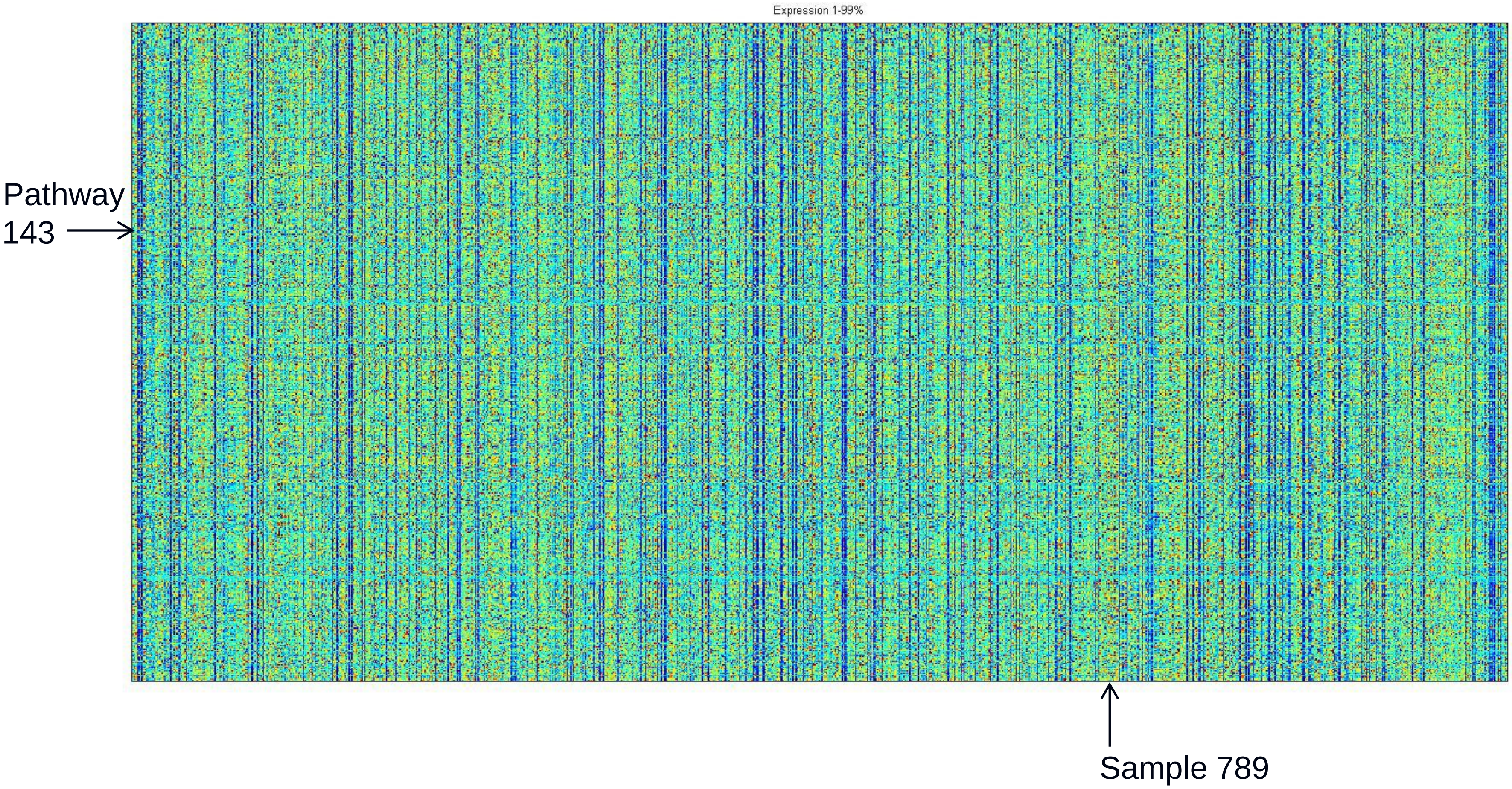
*Drier, Sheffer & Domany *PNAS 2013*

PDS OF 552 PATHWAYS: EACH SAMPLE (144 NORMAL, 997 BREAST TUMOR) IS REPRESENTED BY 552 SUCH PATHWAY–BASED SCORES**



**Livshits et al *Mol Onc* (2015)

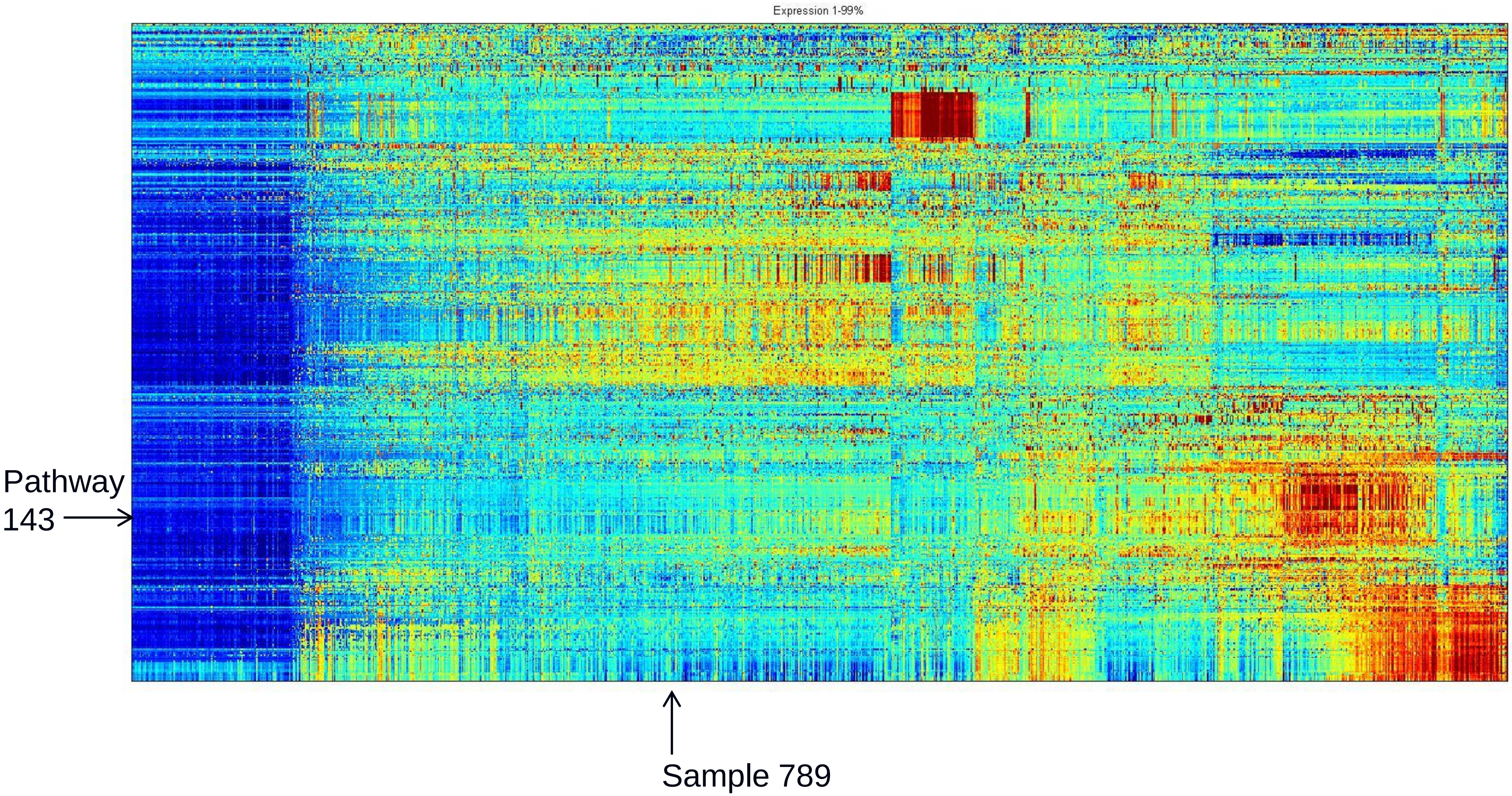
PERFORM ANALYSIS IN THIS SPACE: REORDERING* SAMPLES (*AND* PATHWAYS) REVEALS STRUCTURE IN DATA**



*Tsafrir et al *Bioinformatics* (2005)

**Livshits et al *Mol Onc* (2015)

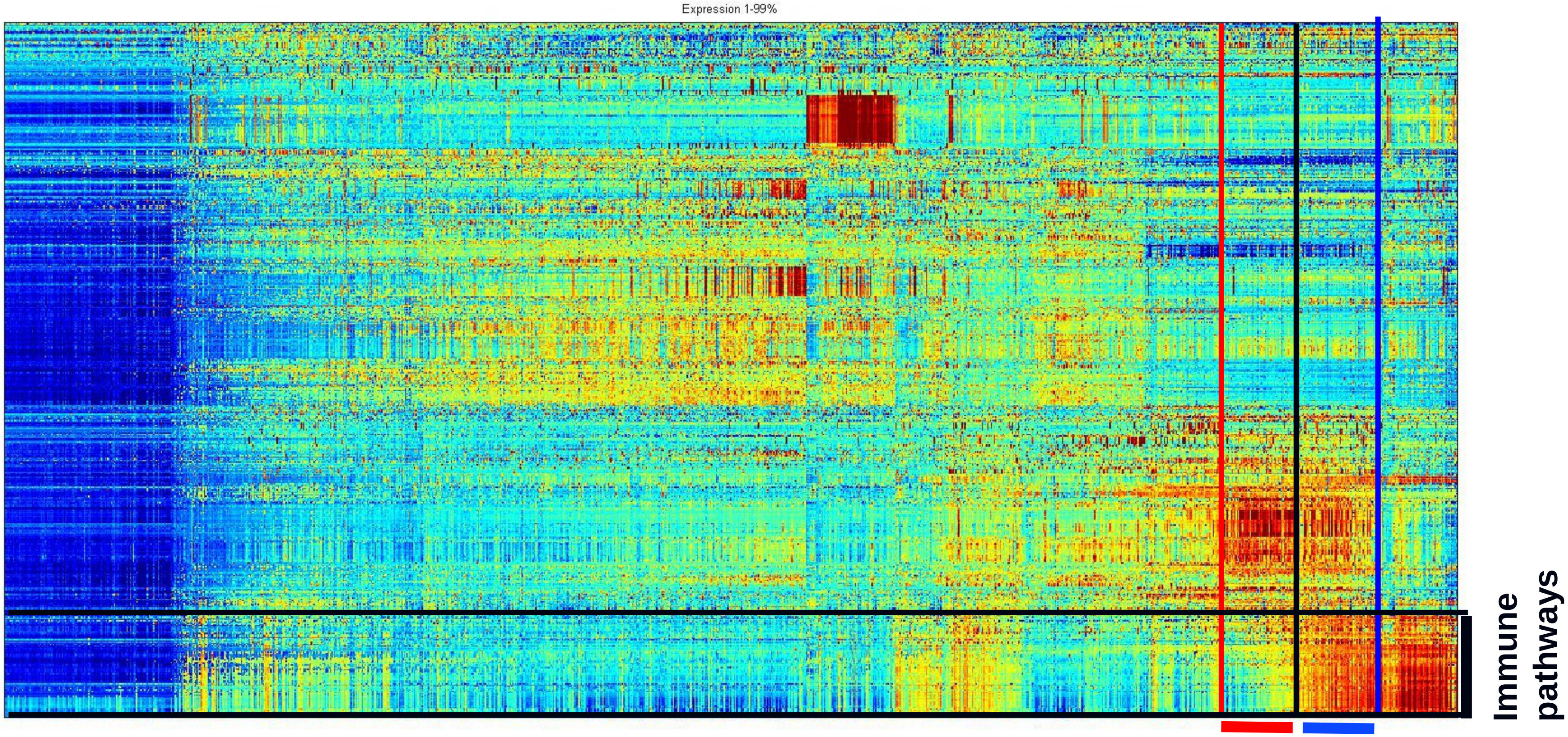
PERFORM ANALYSIS IN THIS SPACE: REORDERING* SAMPLES (*AND* PATHWAYS) REVEALS STRUCTURE IN DATA**



*Tsafrir et al *Bioinformatics* (2005)

**Livshits et al *Mol Onc* (2015)

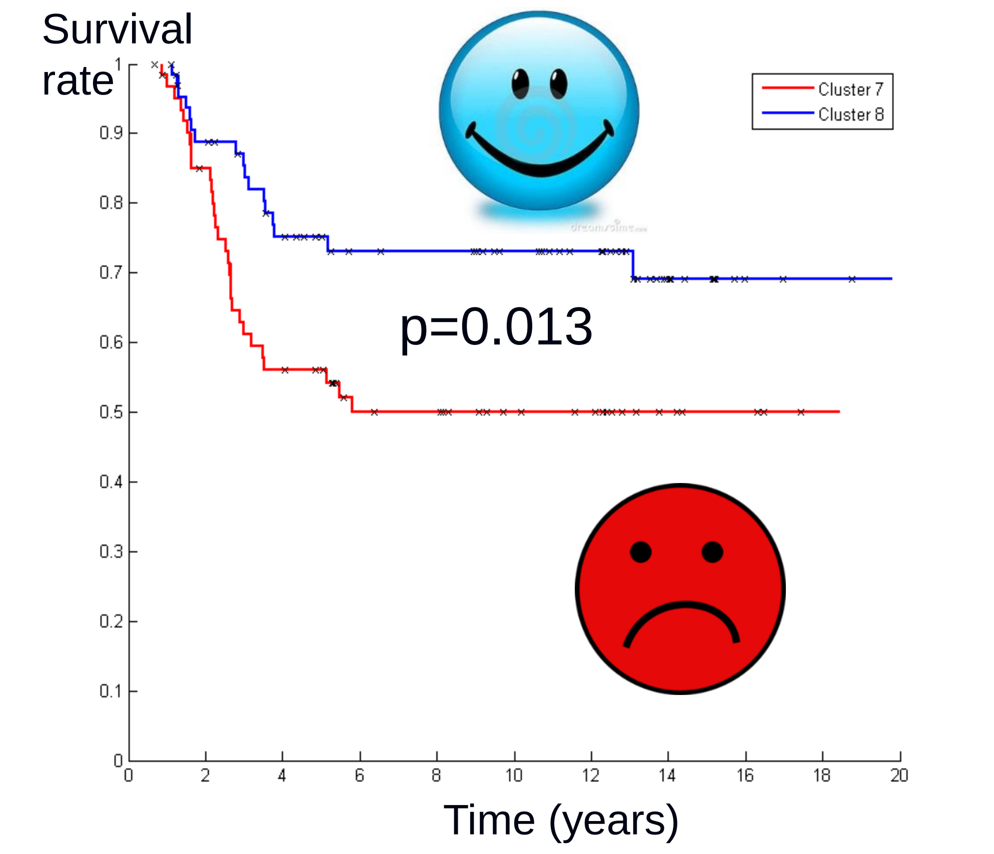
FOCUS ON “TRIPLE NEGATIVE” TUMORS: TWO DISTINCT GROUPS



“TRIPLE NEGATIVE” (TN) SUBTYPE – 2 GROUPS:
HIGH AND LOW IMMUNE INVOLVEMENT

DIFFERENT OUTCOME/SURVIVAL FOR THE TWO GROUPS!

CLINICAL SIGNIFICANCE: FOR TN SUBTYPE, HIGH IMMUNE INVOLVEMENT → BETTER SURVIVAL

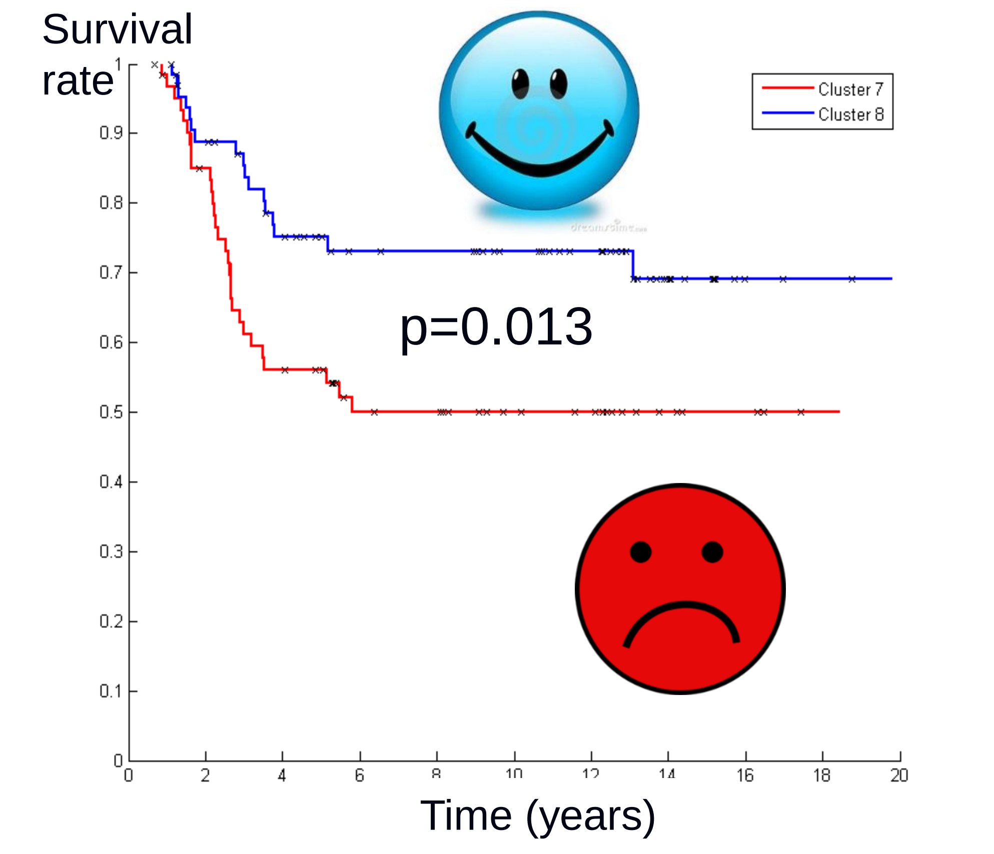


CLINICAL SIGNIFICANCE:

TN tumors with HIGH IMMUNE system involvement – better survival

TN tumors with LOW IMMUNE system involvement -- worse

BIOLOGICAL INTERPRETATION:
HIGH IMMUNE INVOLVEMENT (PDS) ⇔ HIGH *TIL* LEVEL



BIOLOGICAL INTERPRETATION:

HIGH IMMUNE PDS ⇔ high level of ***T***umor ***I***nfiltrating ***L***ymphocytes

Highest correlation with ***TIL*** levels

- for T-CELL related PATHWAYS
- cell-specific signatures => Tcells

➡ ***BIOMARKER!***

PROGNOSTIC BIOMARKER?

Alexe et al (2007): no difference in survival between TN tumors with high/low immune involvement

CLINICAL SIGNIFICANCE:

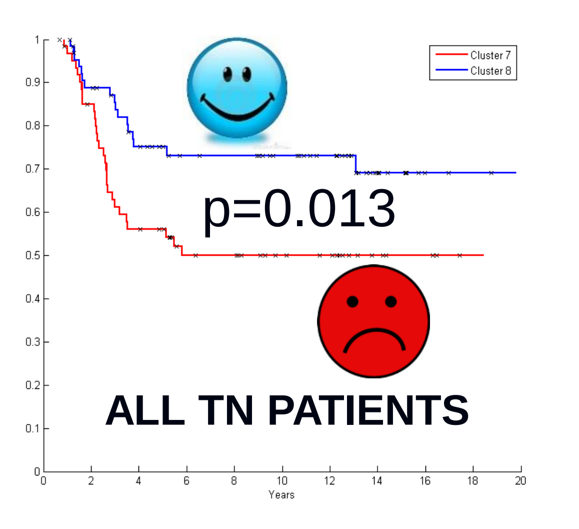
Basal tumors with HIGH IMMUNE system involvement – better survival

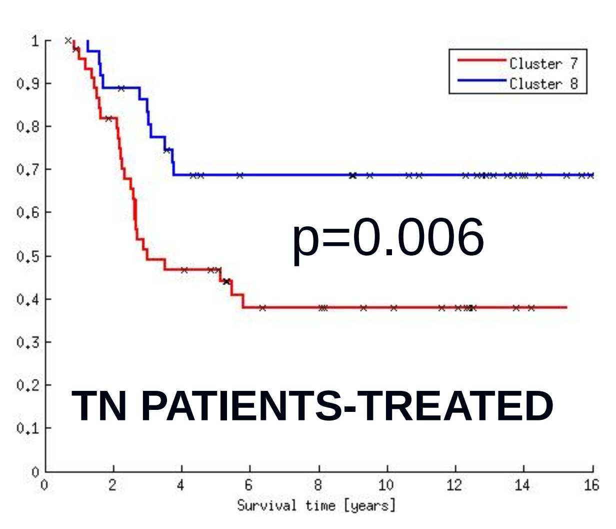
Basal tumors with LOW IMMUNE system involvement -- worse

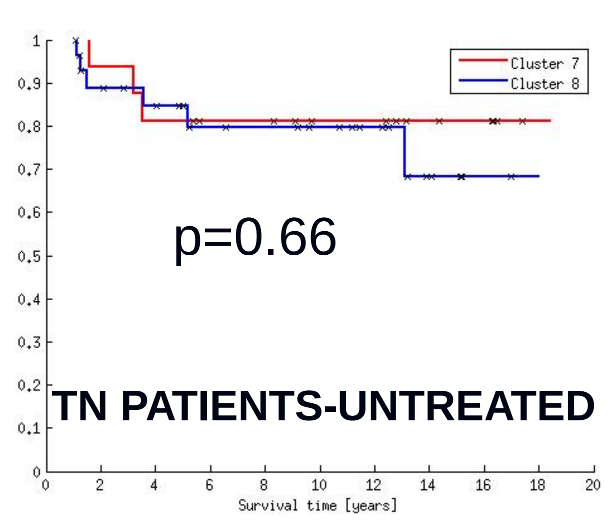
PREDICTIVE BIOMARKER: FOR TN SUBTYPE, IMMUNE INVOLVEMENT → BETTER RESPONSE TO THERAPY

Alexe et al (2007): TN PATIENTS DID NOT RECEIVE CHEMOTHERAPY

METABRIC (2012): MAJORITY OF TN WERE **TREATED** (anthracyclins).

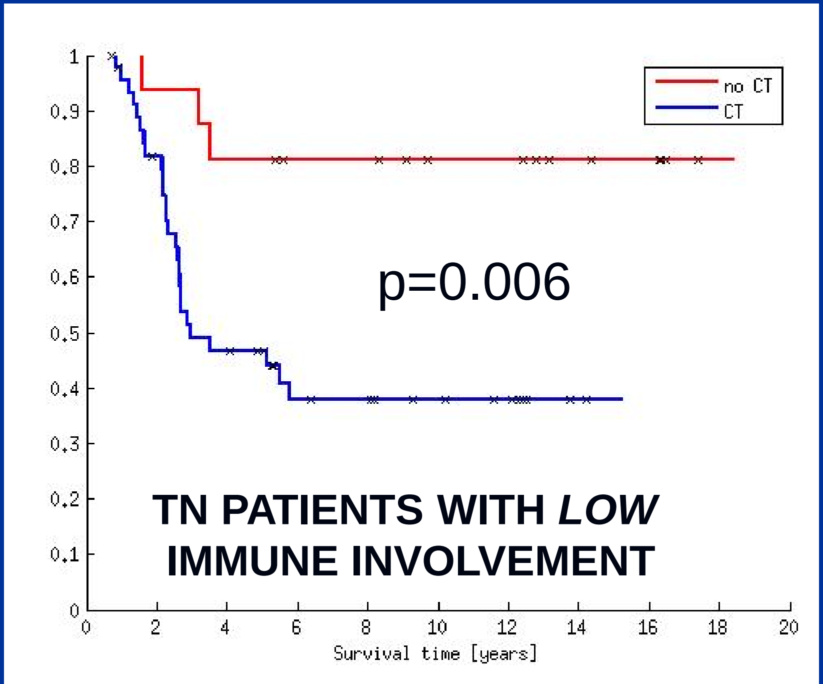


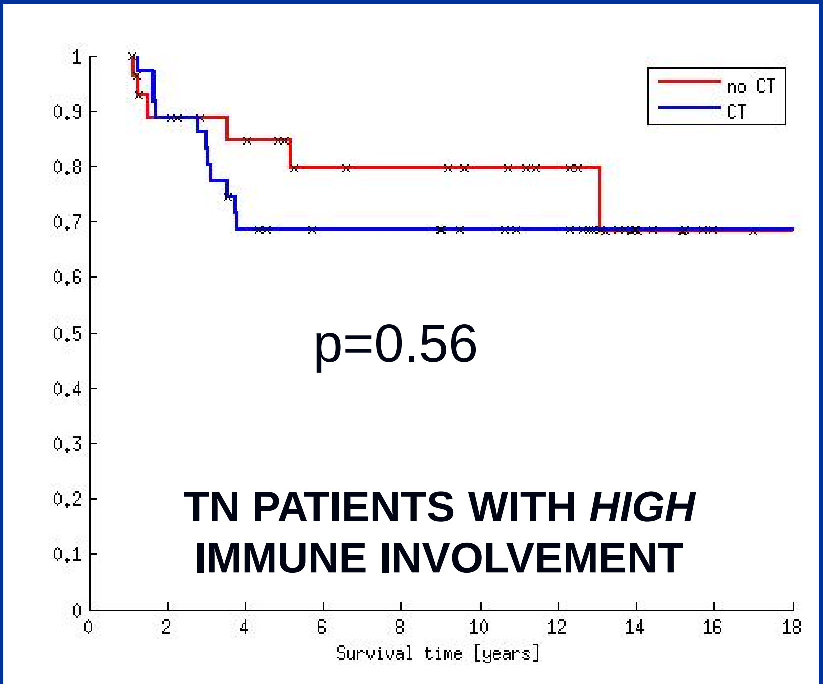




DIFFERENCE IN SURVIVAL BETWEEN BASAL PATIENTS WITH HIGH vs LOW IMMUNE INVOLVEMENT IS OBSERVED ONLY FOR PATIENTS WHO RECEIVED CHEMOTHERAPY. ***PREDICTIVE BIOMARKER?***

PREDICTIVE BIOMARKER: FOR TN SUBTYPES, IMMUNE INVOLVEMENT ➔ BETTER RESPONSE TO THERAPY



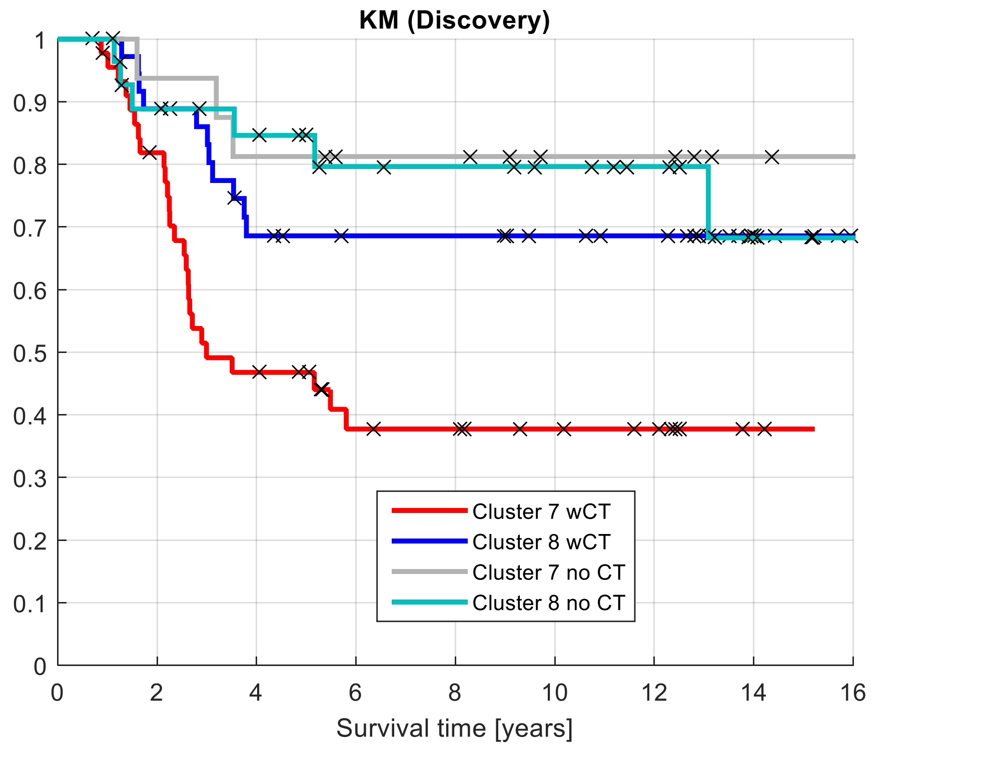


POSSIBLE INTERPRETATION 1: ANTHRACYCLINS ARE KILLING TN PATIENTS WITH ***LOW*** IMMUNE INVOLVEMENT, AND HAVE NO EFFECT ON PATIENTS WITH ***HIGH*** IMMUNE INVOLVEMENT.

INTERPRETATION 2: HIGH RISK PATIENTS (BAD ***CLINICAL*** INDICATORS) WERE SENT TO CHEMO. IF **LOW IMMUNE – CHEMO DID NOT HELP.**
HIGH IMMUNE – CHEMO DID HELP!

PREDICTIVE BIOMARKER: FOR TN SUBTYPES, IMMUNE INVOLVEMENT ➔ BETTER RESPONSE TO THERAPY

	CT	No CT	Total
Cluster 7 (Low Imm)	46	16	62
Cluster 8 (High Imm)	36	29	65
Total	82	45	127



Anthracyclins & immune system:
Zitvogel Cell Deat & Differ. (2014)
Nat. Med. (2014)
Oncoimmunology (2014)

WE USED CT/NO CT AS A PROXY FOR (CLASSICAL) HIGH/LOW RISK. HIGH IMMUNE INVOLVEMENT/TIL INDICATES GOOD RESPONSE OF HIGH-RISK TN PATIENTS TO ANTHRACYCLINS.

DO NOT TREAT (WITH ANTHRACYCLINS) HIGH RISK TN PATIENTS WITH LOW TIL. ***PREDICTIVE BIOMARKER!***

SUGGESTED DECISION PIPELINE:

1. IDENTIFY TRIPLE NEGATIVE (TN) PATIENTS (HISTOCHEMISTRY)

2. USE CLINICAL (OR OTHER) INDICATORS TO IDENTIFY HIGH RISK TN PATIENTS, CANDIDATES FOR CHEMOTHERAPY

3. FOR HIGH-RISK TN PATIENTS:
   MEASURE T – CELL INFILTRATE LEVEL IN TUMOR

4. IF LOW TIL – DO NOT TREAT WITH ANTHRACYCLINES

TAKE – HOME LESSONS*:

1. DO NOT USE IGNORANCE-BASED “TOP RANKED” SINGLE GENE LISTS: THEY ARE UNSTABLE**, MOSTLY DEVOID OF BIOLOGICAL MEANING***.

2. CHARACTERIZE TUMORS BY KNOWLEDGE-BASED, SYSTEM-LEVEL VARIABLES# (Pathway Deregulation Scores).

3. LOWER AIMS: NO SILVER BULLET& THAT WORKS FOR ALL BREAST CANCER SUBTYPES AND ALL CHEMOTHERAPIES.

4. GENOMIC BIOMARKERS SHOULD COMPLEMENT CLASSICAL CLINICAL RISK INDICATORS (NOT REPLACE THEM).

*   *Domany Cancer Res (2014)*

**   *Ein-Dor et al Bioinformatics (2005); PNAS (2006)*

*** *Drier et al PLoS ONE (2011)*

#   *Drier et al PNAS (2013)*

& *Livshits et al Oncotarget (2015)*

THANKS FOR LISTENING

&

APOLOGIES FOR RUNNING OVER TIME