

# תוכנה 1 – אביב תשע"ד

## תרגיל מספר 5

### קלט- פלט ועיבוד קבצים מתקדם (IO, Parsing)

#### הנחיות כלליות:

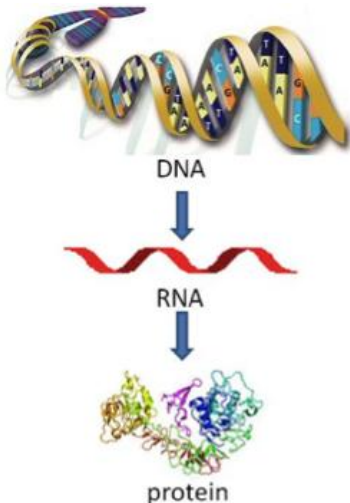
- קראו בעיון את קובץ נהלי הגשת התרגילים אשר נמצא באתר הקורס.
- הגשת התרגיל תיעשה במערכת ה-moodle בלבד (<http://moodle.tau.ac.il/>).
- יש להגיש קובץ zip יחיד הנושא את שם המשתמש ומספר התרגיל (לדוגמא, עבור המשתמש aviv יקרא הקובץ aviv\_hw5.zip). קובץ ה-zip יכיל:
  1. קובץ פרטים אישיים בשם details.txt המכיל את שמכם ומספר ת.ז.
  2. קבצי ה-java של התוכניות אותם התבקשתם לממש.
- יש לממש את המתודות הנדרשות על פי החתימות וערכי ההחזרה המצוינים בתרגיל.
- מותר ואף רצוי להגדיר מתודות עזר.
- קבצי הפלט המופקים ע"י התוכנית צריכים להיות זהים לקבצי הפלט לדוגמא המצורפים לתרגיל.

## Bioinformatics: Gene Expression Analysis

בתרגיל זה ננתח מידע ביולוגי אמיתי וננסה לזהות גנים שרמת הביטוי שלהם שונה משמעותית בין שני תתי-סוגים של סרטן השד. זיהוי גנים כאלו יאפשר לנו ללמוד על המנגנונים הביולוגיים העומדים מאחורי תתי-הסוגים השונים, ואולי גם לאבחן טוב יותר חולות על בסיס רמת הביטוי של גנים אלו.

במהלך התרגיל נתאמן על קריאה וכתובה של קבצים בפורמט מוגדר, עיבוד מתקדם של טקסט, עבודה מתקדמת עם מערכים, וגם נלמד קצת בעקיפין מהפיכת הרפואה המותאמת אישית שמוביל תחום הביואינפורמטיקה.

#### נתחיל עם רקע ביולוגי מקוצר ומקורב:



בגרעין של רוב התאים בגוף שלנו, שמור אותו רצף דנ"א המכיל כ-20,000 מקטעים הקרויים גנים. רוב הגנים הידועים מקודדים הוראות לייצור חלבון (החלבונים משמשים את התאים בגופינו לצרכים מבניים ותפקודיים).

כשתא צריך לייצר חלבון מסוים, הוא קודם ממיר את מקטע הדנ"א של הגן הרלבנטי למולקולת רנ"א-שליח (mRNA), וזו יוצאת מהגרעין אל חלל התא, שם היא נקראת ע"י מנגנוני בנית החלבונים של התא שיוצרים על פיה חלבון מתאים. ככל שיש יותר עותקים של mRNA כך יוצרו יותר מולקולות של החלבון הרלבנטי, ונוכל לומר ש-"הביטוי של הגן" הוא גבוה יותר (כלומר נוצרות ממנו יותר מולקולות mRNA ואח"כ גם יותר מולקולות חלבון).

תאים שונים יבטאו גנים ברמות שונות כתלות ב-

- **סוג התא** - תא שריר יבטא גנים אחרים (או ברמות שונות) מתא כבד כי כל סוג תא זקוק לחלבונים שונים לצורך תפקודו.
- **מצב התא** - גם בתאים מאותו סוג יש שינויים ברמות ביטוי הגנים לאורך זמן - למשל: לאחר ארוחה יש צורך בהרבה מולקולות אינסולין ולכן תאי הבלבב יתחילו לבטא את הגן המקודד לחלבון זה.
- **מחלה** - במחלות שונות יש פגיעה ברמות הביטוי של גנים בהשוואה לתא נורמלי. בהרבה סוגים של סרטן משתנות רמות הביטוי של גנים רבים, מה שמאפשר לתאים לאמץ תכונות חדשות וביניהן כאלו שגורמות לחלוקה מואצת שלהם.



שבבי דנ"א. כל שבב מודד את רמות הביטוי של אלפי גנים בדגימה ביולוגית אחת. ערכי הביטוי המופקים מכל שבב יתורגמו לעמודה אחת במטריצת הביטוי.

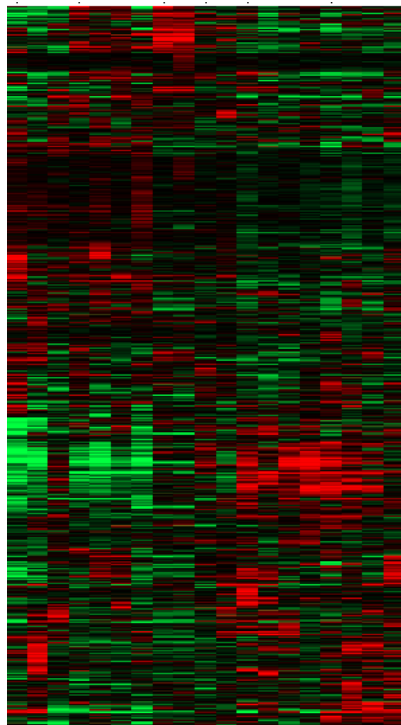
טכנולוגיות מהפכניות חדשות (כגון **שבבי דנ"א** או RNA-Seq) מאפשרות לנו למדוד את רמת מולקולות ה-mRNA בדגימה ביולוגית מסוימת, וכך הן מאפשרות לנו לדעת מהי רמת הביטוי הספציפית של כל גן באותה הדגימה.

במחקר של ניתוח רמות ביטוי גנים (Gene Expression Analysis), אחד מענפי הביואינפורמטיקה), לרוב נתחיל את האנליזה עם קבלת קובץ ענק המכיל מטריצת ביטוי גנים. במטריצה זו כל **שורה מייצגת גן (Gene)**, וכל עמודה מייצגת דגימה (Sample). כלומר - כל תא במטריצה מייצג את רמת הביטוי של גן מסוים בדגימה מסוימת.

19 דגימות (Samples)

ניתן להציג את מטריצת הביטוי כ-Heatmap כפי שמופיע בתרשים משמאל.

1259 גלוי גנים (Gene probes)



Expression level:

High Low

בתרגיל זה, ננתח אוסף נתונים המכיל 1259 גלוי גנים (שורות המטריצה) ו-19 דגימות (עמודות המטריצה) שנלקחו משני תת-סוגים של גידולים של סרטן השד. הדגימות מתחלקות לשתי קבוצות: (ER+) ו-(ER)\*. כלומר, לכל דגימה יש תווית שממפה אותה לאחת משתי המחלקות.

נחפש גנים שנבדלים בביטוי שלהם בין 2 המחלקות (Differentially expressed genes). כלומר, עבור כל גן, נבדוק האם ממוצע הביטוי שלו על הדגימות מסוג ER+, שונה בצורה משמעותית מממוצע הביטוי שלו על דגימות מסוג ER-. לשם כך נשתמש במבחן סטטיסטי בשם two sample t-test.

למציאת גנים כאלו יש משמעות ביולוגית ורפואית - הם מאפשרים לנו ללמוד על המנגנונים הביולוגיים שמאפיינים כל תת סוג, לאבחן בצורה מדויקת יותר גידולים חדשים על בסיס גנים אלו, ובהמשך להציע טיפול מותאם אישית לחולה לפי האפיון הגנטי שלה.

למי שרוצה לקרוא עוד על הנושא, הנה קישור למאמר שפורסם על סמך אוסף נתונים זה: <http://www.ncbi.nlm.nih.gov/pubmed/22025563>

\* מה משמעות ER+? בהיעדר דרך טובה יותר, כיום נהוג לאפיין גידולים בשד לפי קיומו של קולטן האסטרוגן (ER) על דופן התאים הסרטניים, דבר שניתן לבדוק בבדיקת מעבדה פשוטה יחסית. נמצא שיש הבדל גדול בפוטנציאל הגרורתי ובתגובה לטיפול בין גידולים מסוג ER+ לבין גידולים מסוג ER-, אבל לא הרבה ידוע על ההבדל הביולוגי בין שני תתי הסוגים. התקווה היא שעם מציאת גנים מפרדים בין שני תת הסוגים נוכל ללמוד עוד על דרך היווצרות תתי-הסוגים ועל אפשרויות לטיפול בהם בצורה טובה יותר.

## מה עליכם לעשות ?

בין קבצי התרגיל נתון לכם שלד של המחלקה GeneExpressionAnalyzer.

עליכם להשלים את מימוש המתודות הבאות בתוך מחלקה זו על פי ההוראות המופיעות בהמשך:

- parseGeneExpressionFile
- writeDatasetToTabularFile
- getDataEntriesForLabel
- writeTopDifferentiallyExpressedGenesToFile

בתרגיל זה עליכם להגיש את המחלקה GeneExpressionAnalyzer לאחר שתשלימו אותה, ואת קובץ הפלט GDS4069-DiffGenes.txt שתיצרו בסעיף 8.

## תיאור אופן הרצת התוכנית

לאחר שתשלימו את מימוש המחלקה, אם נריץ אותה משורת הפקודה באופן הבא:

```
java GeneExpressionAnalyzer GDS4085 "estrogen receptor-negative"
"estrogen receptor-positive" 0.01
```

כך ש-

- הארגומנט הראשון מציין את שם אוסף הנתונים אותו יש לנתח (בהנחה שקובץ בשם GDS4085.soft נמצא בתיקיה הנוכחית).
- הארגומנט השני והשלישי מציינים שמות של מחלקות. נרצה למצוא גנים שנבדלים ברמות הביטוי שלהם בין דגימות מהמחלקה הראשונה לבין דגימות מהמחלקה השניה (פרטים בהמשך).
- הארגומנט הרביעי מציין סף מובהקות סטטיסטית בין 0 ל-1.

אז המחלקה תקרא ותעבד את תוכן קובץ הקלט לפי פורמט SOFT שיוגדר בהמשך, תשמור גרסא מתומצתת של אוסף הנתונים לקובץ במבנה טבלאי, ולבסוף תכתוב לקובץ פלט את רשימת הגנים שמתבטאים באופן שונה על שתי מחלקות הדגימות (גנים שה-pValue שקיבלו במבחן ה-t-test קטן מערך המובהקות שניתן כארגומנט הרביעי בשורת הפקודה).

פלט התוכנית:

כתגובה לארגומנטים הנ"ל, תיצור התוכנית את הקבצים הבאים בספרייה הנוכחית (הקבצים נתונים לכם):

GDS4085-DiffGenes.txt , GDS4085-Tabular.txt

ותדפיס למסך את הפלט הבא:

```
Gene expression dataset loaded from file GDS4085.soft.
Dataset contains 19 samples and 1259 gene probes.
```

```
Dataset saved to tabular file - GDS4085-Tabular.txt.
```

```
93 differentially expressed genes identified using alpha of 0.010000 when
comparing the two sample groups [estrogen receptor-negative] and [estrogen
receptor-positive].
Results saved to file GDS4085-DiffGenes.txt.
```

## המחלקה GeneExpressionDataset

המחלקה GeneExpressionDataset מאחסנת מטריצת ביטוי גנים ונתונים נלווים. המחלקה הוגדרה עבורכם כמחלקה פנימית של המחלקה GeneExpressionAnalyzer.

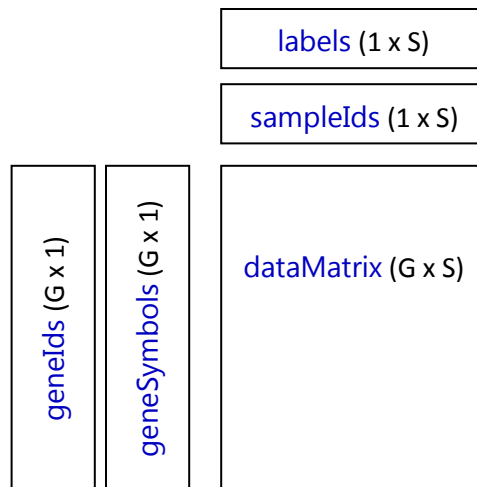
```
public static class GeneExpressionDataset {

    public int samplesNumber; //number of dataset samples
    public int genesNumber; // number of dataset gene probes

    public String[] sampleIds; //sample ids
    public String[] geneIds; //gene probe ids
    public String[] geneSymbols; //gene symbols
    public float[][] dataMatrix; //expression data matrix

    public String[] labels; //sample labels
}
```

תיאור סכמתי של המערכים המתארים אוסף נתונים (S מייצג את מספר הדגימות ו-G את מספר גלאי הגנים):



### הערות אחרונות לפני תחילת המימוש:

- אין צורך לבצע בדיקות על הקלט בתרגיל זה – ניתן להניח שהמשתמש מפעיל את התוכנית עם ארגומנטים נכונים, ושקובץ הקלט הוא בפורמט SOFT תקין העונה על המבנה שמוגדר בהמשך.
- יש לממש את המחלקה בצורה כללית כך שתוכל לפעול על כל קובץ ביטוי גנים בפורמט SOFT (המקיים את המבנה ואת ההנחות שיובאו בהמשך). בסעיף 8 נריץ את התוכנית על קובץ נוסף. כלומר – אין לקודד ערכים ספציפיים לאוסף נתונים מסוים (כמו מספר הדגימות או שם של מחלקת דגימות מסוימת).

קדימה לעבודה...

## חלק א: קריאה ועיבוד של קובץ בפורמט SOFT

1. הורידו מאתר ה-Gene Expression Omnibus את הקובץ -

<ftp://ftp.ncbi.nlm.nih.gov/geo/datasets/GDS4nnn/GDS4085/soft/GDS4085.soft.gz>

שימרו את הקובץ GDS4085.soft בספריית הפרויקט.

הציצו גם בעמוד הבא המכיל תיאור של אוסף נתונים זה:

<http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4085>

2. פיתחו את הקובץ בתוכנה לגיליון נתונים והתרשמו מהפורמט שלו.

תיאור מלא של פורמט ה-SOFT המשמש לאחסון נתוני ביטוי גנים במאגר ה-GEO נמצא בקישור:

<http://www.ncbi.nlm.nih.gov/geo/info/soft2.html#SOFTformat>

✓ שימו לב: שורה המתחילה עם ^ מייצגת הגדרת ישות חדשה. שורה המתחיל עם ! מייצגת מאפיין של הישות.

✓ טבלת הנתונים שמופיעה בסוף קובץ הקלט מופרדת ע"י טאב ("t").

הנה סקירה של שדות המחלקה GeneExpressionDataset שחוצגה לעיל, וכיצד למלא אותם בנתונים מקובץ הקלט:

אלמנט בקובץ הקלט שמכיל את הנתונים	שם השדה במחלקה GeneExpressionDataset
!dataset_sample_count	<b>int</b> samplesNumber;
!dataset_feature_count	<b>int</b> genesNumber;
טבלת הנתונים מופיעה בין השורה: !dataset_table_begin לבין השורה: !dataset_table_end	<b>String[]</b> sampleIds; <b>String[]</b> geneIds; <b>String[]</b> geneSymbols; <b>float[][]</b> dataMatrix;
✓ השורה הראשונה בטבלת הנתונים מכילה את מזהי הדגימות (ה- <b>GSM801843</b> ערך לדוגמא: ✓ העמודה הראשונה משמאל בטבלת הנתונים מכילה את מזהי גלאי הגנים (ה- <b>geneIds</b> ). ערך לדוגמא: <b>NM_016242.2_psr1</b> ✓ העמודה השנייה משמאל בטבלת הנתונים מכילה את שמות הגנים (ה- <b>geneSymbols</b> ). ערך לדוגמא: <b>EMCN</b> ✓ הערכים המספריים בטבלת הנתונים יאוחסנו בשדה <b>dataMatrix</b> .	
הערה טכנית: יתכן וטבלת הנתונים תכיל יותר משורה אחת עם אותו Gene- Symbol, ו-Gene-Ids שונים כי יש גנים שנמדדים ע"י יותר מגלאי גן (probe) אחד.	

<p>מערך ה-labels יגדיר עבור כל דגימה לאיזה תת-קבוצה (מחלקה) היא שייכת. עבור קובץ הנתונים הספציפי שלנו, כל תא במערך זה יכול להכיל את הערך "estrogen receptor-positive" או "estrogen receptor-negative" בלבד.</p> <p>כדי למלא מערך זה יש לעבד את המקטעים בקובץ שמגדירים תת קבוצה של דגימות, ומתחילים ב-"ASUBSET". כל מקטע כזה מגדיר את שם תת הקבוצה (בשורה המתחילה ב-subset_description) ואת רשימת הדגימות שמשתייכות אליה כשהן מופרדות בפסיקים (בשורה המתחילה ב-subset_sample_id).</p> <p>הצעה: בזמן קריאת הקובץ, בכל פעם שנתקלים במקטע מסוג SUBSET המתאר תת-קבוצה (מחלקה) של דגימות, כדאי לאחסן במערך אחד את שם תת-הקבוצה, ובמערך שני את רשימת הדגימות ששייכות אליה. בגמר קריאת הקובץ ניתן לבנות ממערכי עזר אלו את מערך ה-labels הסופי.</p>	<pre>String[] labels;</pre>
--	-----------------------------

3. כתבו מתודה המקבלת שם קובץ המציין קובץ קלט בפורמט SOFT, ומחזירה אובייקט מסוג GeneExpressionDataset.

חתימת המתודה:

**public static** GeneExpressionDataset parseGeneExpressionFile (String filename) **throws** IOException

### חלק ב': שמירת אוסף הנתונים לקובץ טבלאי מתומצת

4. נרצה לשמור את תוכנו של אובייקט מסוג GeneExpressionDataset לקובץ בפורמט טבלאי באופן תמציתי. כתבו מתודה המקבלת אובייקט מסוג GeneExpressionDataset, ושם קובץ פלט אליו תשמור המתודה את נתוני הביטוי על פי הפורמט בקובץ ההדגמה GDS4085-tabular.txt המצורף לקבצי התרגיל.

חתימת המתודה:

**public static void** writeDatasetToTabularFile(  
GeneExpressionDataset geneExpressionDataset,  
String outputFilename) **throws** IOException

שימו לב:

- ✓ השורה הראשונה בקובץ תכיל את ה-Label של כל דגימה והשורה השניה את שמות הדגימות.
- ✓ העמוד הראשונה תכיל את מזהי הגנים והעמודה השניה את סימולי הגנים.
- ✓ השדות בקובץ זה יופרדו ע"י טאב ("\t").
- ✓ ערכי הביטוי בקובץ זה יכללו 2 ספרות בלבד לאחר הנקודה (ניתן להיעזר ב-String.format).

## חלק ג': זיהוי גנים מפרידים בין שתי קבוצות של דגימות

5. כתבו מתודת עזר אשר תשמש אותנו לשלוף את ערכי הביטוי של גן מסוים אבל רק על דגימות המשתייכות למחלקה מסוימת.

חתימת המתודה:

```
public static double[] getDataEntriesForLabel(float[] data, String labels, String label)
```

המתודה תחזיר את כל הערכים מהמערך `data`, אשר בתא המקביל להם במערך `labels` מצוי הערך `label`.

לדוגמא, בהפעלת המתודה באופן הבא אנחנו שולחים שורה מהמטריצה המייצגת את ערכי הביטוי של גן `i` על כל הדגימות. המתודה תחזיר מערך המכיל את ערכי הביטוי של הגן `i` רק עבור דגימות שהערך שלהן במערך ה-`labels` שווה ל-`"estrogen receptor-positive"`:

```
getDataEntriesForLabel(geneExpressionDataset.dataMatrix[i], geneExpressionDataset.labels, "estrogen receptor-positive")
```

6. עתה ננסה לזהות גנים אשר ערכי הביטוי שלהם שונים באופן מובהק סטטיסטית בין שתי תת-קבוצות (מחלקות) של דגימות (כלומר, עבור כל גן נבדוק את האם ממוצע ערכי הביטוי שלו על דגימות המחלקה הראשונה שונה ממוצע ערכי הביטוי שלו על דגימות המחלקה השניה). לשם כך נשתמש במבחן `t-test`, אשר מימוש שלו כבר נתון לכם ע"י המתודה:

```
calcTtest(geneExpressionDataset, geneIndex, label1, label2);
```

המתודה `calcTtest` מחזירה `p-Value` בטווח 0-1 אשר יהיה נמוך יותר ככל שהערכים של הגן שונים בין דגימות של שתי הקבוצות (נהוג להתייחס אל גנים כאל מפרידים בין 2 קבוצות אם ערך `pValue` שלהם נמוך מ-0.05 או 0.01).

מתודה `calcTtest` משתמשת במימוש חיצוני למבחן `t-test` ועל כן יש להוסיף את קובץ ה-`jar` הבא לפרוייקט:

```
commons-math3-3.2.jar
```

לאחר שהורדנו אותו מהקישור הבא:

<http://apache.spd.co.il/commons/math/binaries/commons-math3-3.2-bin.zip>

תזכורת: כדי להוסיף קובץ `JAR` לפרוייקט באקליפס יש לבחור את הקובץ תחת התפריטים:  
Project -> Project properties -> Java Build Path -> Libraries -> Add external Jars

7. לסיום, כתבו מתודה אשר תכתוב לקובץ את רשימת הגנים המפרידים בין 2 קבוצות של דגימות.

עבור כל גן באוסף הנתונים שלנו, חשבו `pValue` בעזרת הרצת המתודה `calcTtest` שתוארה בסעיף הקודם. ה-`pValue` המתקבל, משקף כמה מובהק ההבדל בין ערכי הגן על דגימות ששייכות לתת-הקבוצה `label1` בהשוואה לדגימות השייכות לתת-הקבוצה `label2`.

הדפיו לקובץ את רשימת הגנים שה-pValue שלהם נמוך מפרמטר הסף  $\alpha$ .  
 הגנים צריכים להיות ממוינים לפי ערך ה-pValue שלהם, כאשר בראש הרשימה יופיע הגן המובהק ביותר  
 שהינו בעל ערך ה-pValue המינימלי.  
 עבור כל גן, יש להדפיס במבנה טבלאי את ערך ה-pValue, את מזהה גלאי הגן, ואת סימול הגן (יש לעקוב  
 אחר הפורמט של קובץ ההדגמה GDS4085-DiffGenes.txt).  
 המתודה תחזיר את מספר הגנים שזוהו כמפרידים ונכתבו לקובץ.

חתימת המתודה:

```
public static int writeTopDifferentiallyExpressedGenesToFile(
    String outputFilename,
    GeneExpressionDataset geneExpressionDataset,
    double alpha,
    String label1,
    String label2) throws IOException
```

פורמט קובץ הפלט:

1	0.000013	NM_002996.3_psr1_at	CX3CL1
2	0.000069	NM_001005915.1_psr1_at	ERBB3
3	0.000106	NM_178155.1_psr1_a_at	FUT8
4	0.000149	NM_005623.2_psr1_at	CCL8
5	0.000197	NM_012243.1_psr1_s_at	SLC35A3
6	0.000247	NM_001826.1_psr1_s_at	CKS1B
7	0.000452	NM_004776.3_psr1_at	B4GALT5
8	0.000515	NM_014216.3_psr1_at	ITPK1
9	0.000595	NM_000127.2_psr1_at	EXT1
10	0.000613	NM_002300.4_psr1_x_at	LDHB

בעמוד הבא תוכלו לראות שמספר גנים שדורגו בראש הרשימה שהפקנו, תועדו בעבר בספרות המדעית  
 כבעלי תפקיד חשוב בסרטן השד ...

8. בידקו שהתוכנית שלכם עובדת על אוסף נתונים אחר השמור אף הוא בפורמט SOFT.

חלצו את הקובץ GDS4069.soft מהקובץ -

<ftp://ftp.ncbi.nlm.nih.gov/geo/datasets/GDS4nnn/GDS4069/soft/GDS4069.soft.gz>

והריצו את התוכנית שלכם תוך שימוש בארגומנטים הבאים:

```
java GeneExpressionAnalyzer GDS4069 "triple negative breast
cancer" "non-triple negative breast cancer" 0.01
```

הגישו את הקובץ GDS4069-DiffGenes.txt שהתוכנית יצרה.

בהצלחה !



## The ErbB2/ErbB3 heterodimer functions as an oncogenic unit: ErbB2 requires ErbB3 to drive breast tumor cell proliferation

Thomas Holbro\*, Roger R. Beerli†, Francisca Maurer\*, Magdalena Koziczak\*, Carlos F. Barbas III†, and Nancy E. Hynes\*<sup>5</sup>

\*Friedrich Miescher Institute, P.O. Box 2543, 4002 Basel, Switzerland; and †The Skaggs Institute for Chemical Biology and Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037

Edited by Owen N. Witte, University of California, Los Angeles, CA, and approved May 28, 2003 (received for review December 17, 2002)

ErbB2 is a receptor tyrosine kinase whose activity in normal cells depends on dimerization with another ligand-binding ErbB receptor. In contrast, amplification of *c-erbB2* in tumors results in dramatic overexpression and constitutive activation of the receptor. Breast cancer cells overexpressing ErbB2 depend on its activity for proliferation, because treatment of these cells with ErbB2-specific antagonistic antibodies or kinase inhibitors blocks tumor cells in the G<sub>1</sub> phase of the cell cycle. Intriguingly, loss of ErbB2 signaling is accompanied by a decrease in the phosphotyrosine content of ErbB3. On the basis of these results, it has been proposed that ErbB3 might be a partner for ErbB2 in promoting cellular transformation. To test this hypothesis and directly examine the role of the "kinase dead" ErbB3, we specifically ablated its expression with a designer transcription factor (E3). By infection of ErbB2-overexpressing breast cells with E3, we show that ErbB3 is essential for the proliferation process. Loss of ErbB3 results in a growth block and effects on cell proliferation are not observed in the absence of expression of constitutively active ErbB2. These results indicate that ErbB3 signaling is essential for proliferation and activity alone are insufficient to drive cell division. Furthermore, we identify ErbB3 as a partner for active ErbB2 to the phosphotyrosine pathway. Thus, the ErbB2/ErbB3 heterodimer functions as an oncogenic unit to drive breast tumor cell

proliferation, which proceeds via inhibition of intracellular signaling pathways and directly targets various members of the cell cycle machinery (17–20).

Interestingly, expression of ErbB3 is seen in many tumors that express ErbB2, including breast (21), bladder (22), and others. Furthermore, in many ErbB2-overexpressing breast tumors, ErbB3 has elevated levels of phosphotyrosine (15). ErbB3 itself has impaired tyrosine kinase activity (23) and needs a dimerization partner to become phosphorylated and acquire signaling potential (24). Indeed, we and others have shown that inactivation of ErbB2 leads to decreased ErbB3 tyrosine phosphorylation (17, 18, 25, 26). ErbB3, which contains six docking sites for the p85 adaptor subunit of phosphatidylinositol 3-kinase (PI3K), efficiently couples to this pathway (27, 28). Interestingly, it has

### Essential function for ErbB3 in breast cancer proliferation

<http://breast-cancer-research.com/content/6/3/137>

### The ErbB2/ErbB3 heterodimer functions as an oncogenic unit: ErbB2 requires ErbB3 to drive breast tumor cell proliferation

<http://www.pnas.org/content/100/15/8933.short>

### CX3CL1 expression is associated with poor outcome in breast cancer patients.

<http://www.ncbi.nlm.nih.gov/pubmed/23912959>

### Role of Cks1 amplification and overexpression in breast cancer.

<http://www.ncbi.nlm.nih.gov/pubmed/19161979>

The family of ErbB receptors includes ErbB2, ErbB3, and ErbB4. Ligand binding induces receptor dimerization and activation, ultimately leading to stimulation of the mitogenic pathway. ErbB2 appears to be the predominant receptor in the family. ErbB3 has a similar impact on cell growth, whereas loss of ErbB2 or ErbB3 results in growth defects (8, 9).

A wealth of clinical data has shown that overexpression of tyrosine kinases, in particular ErbB2, is associated with human cancer development, and that inhibition of these kinases for cancer therapies (10–13) is attributable to gene amplification and correlates with overall survival (14). Overexpression of ErbB2 is phosphorylated in breast cancer (15, 16). It has been observed that overexpression of ErbB2 results in efficient in

[www.pnas.org/cgi/doi/10.1073/pnas.02-09111](http://www.pnas.org/cgi/doi/10.1073/pnas.02-09111)