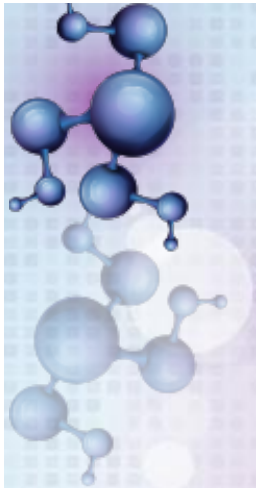
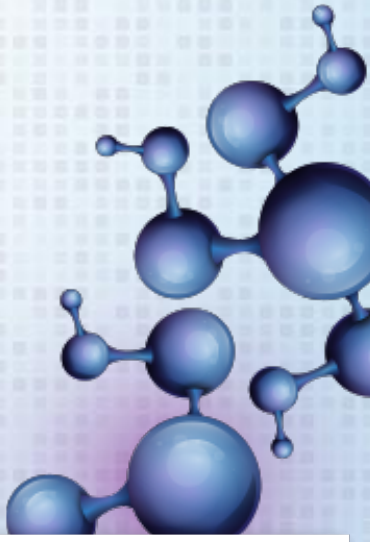
Broad-DREAM Gene Essentiality Prediction Challenge





INSTITUTO Carlos Slim DE LA SALUD

Integrative Cancer Biology Program

Pfizer

NOVARTIS ONCOLOGY

CTD² Cancer Target Discovery and Development

Lilly

EMD

Sage BIONETWORKS

DREAM

BROAD INSTITUTE

DANA-FARBER CANCER INSTITUTE

OHSU

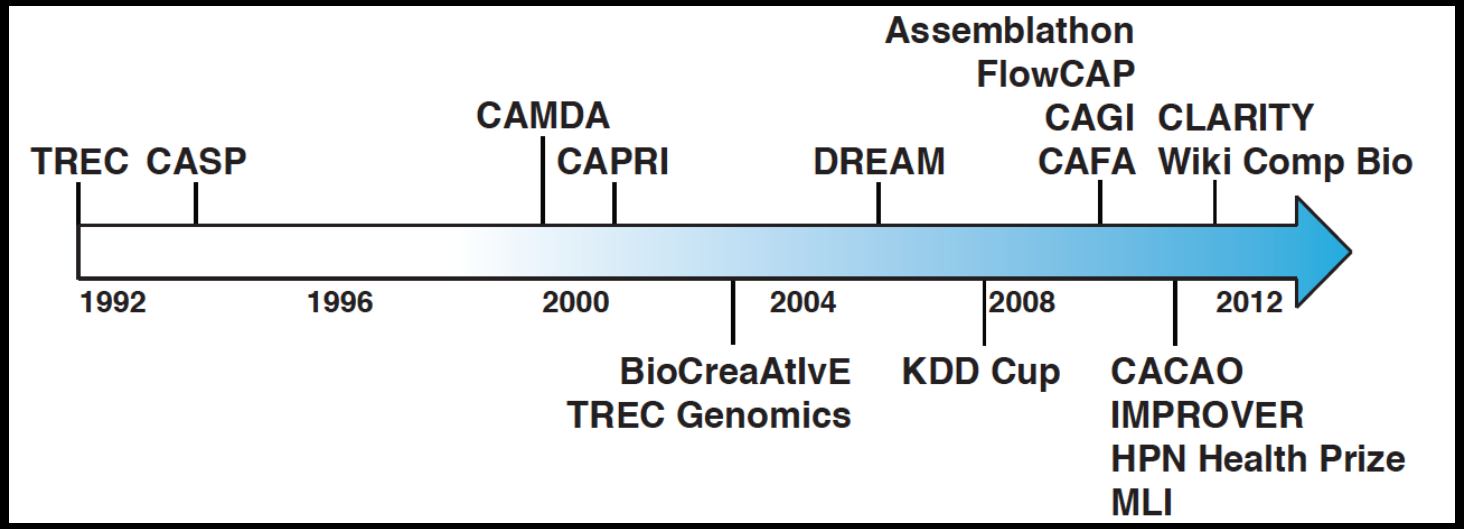
UNIL | Université de Lausanne

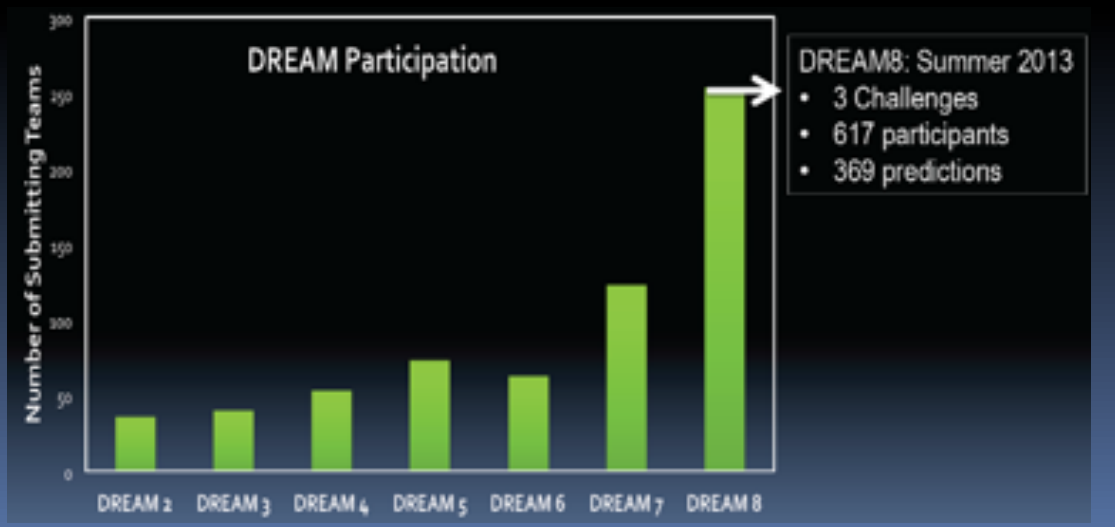
IBM Research

Outline

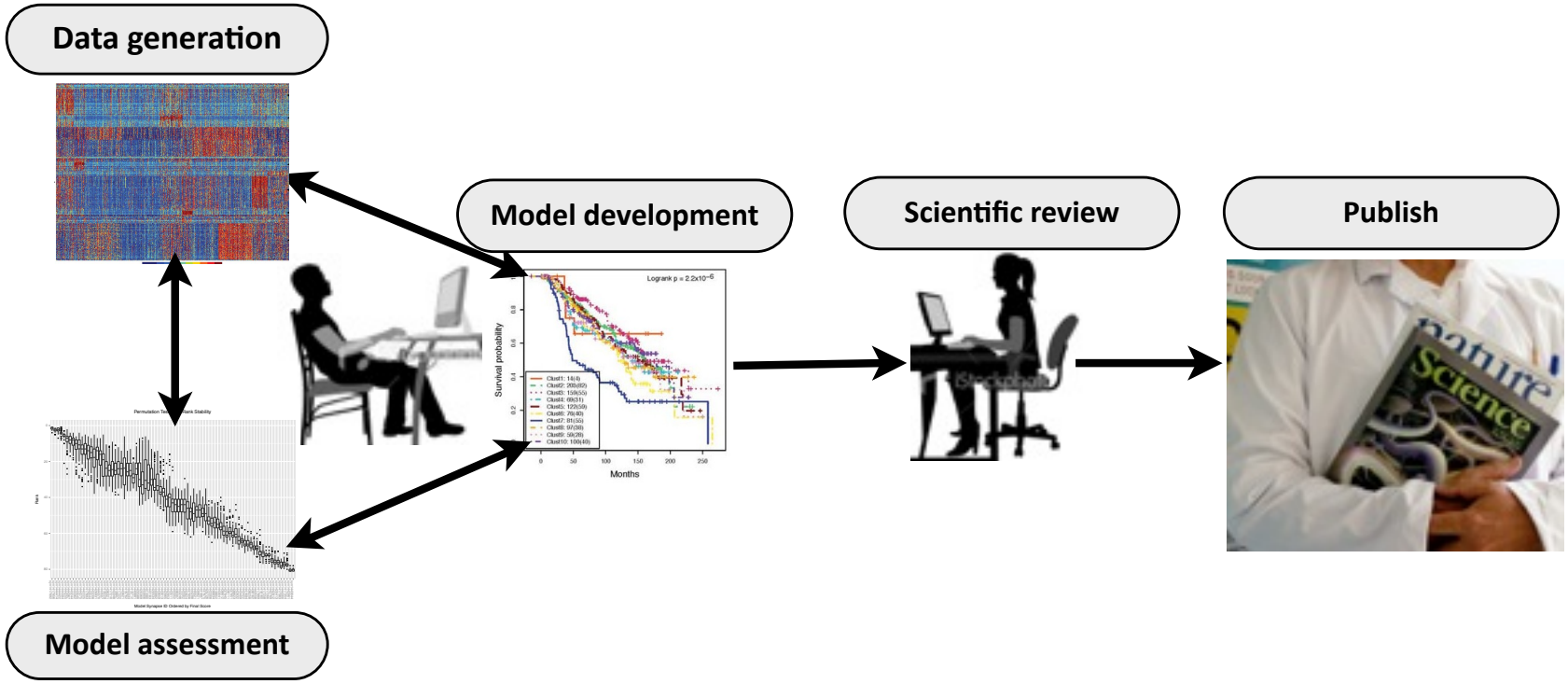
- DREAM overview (Adam Margolin, OHSU)
- Scientific overview (Barbara Weir, Aviad Tsherniak, Broad Institute)
- Challenge details (Mehmet Gonen, Sage Bionetworks)
- Synapse and forum registration (Bruce Hoff, Sage Bionetworks)
- Challenge demo (Mehmet Gonen, Sage Bionetworks)
- Compute resources (Venkat Balagurusamy, IBM)
- Questions and answers (all)

Crowd-sourcing in computational biology





The self-assessment trap



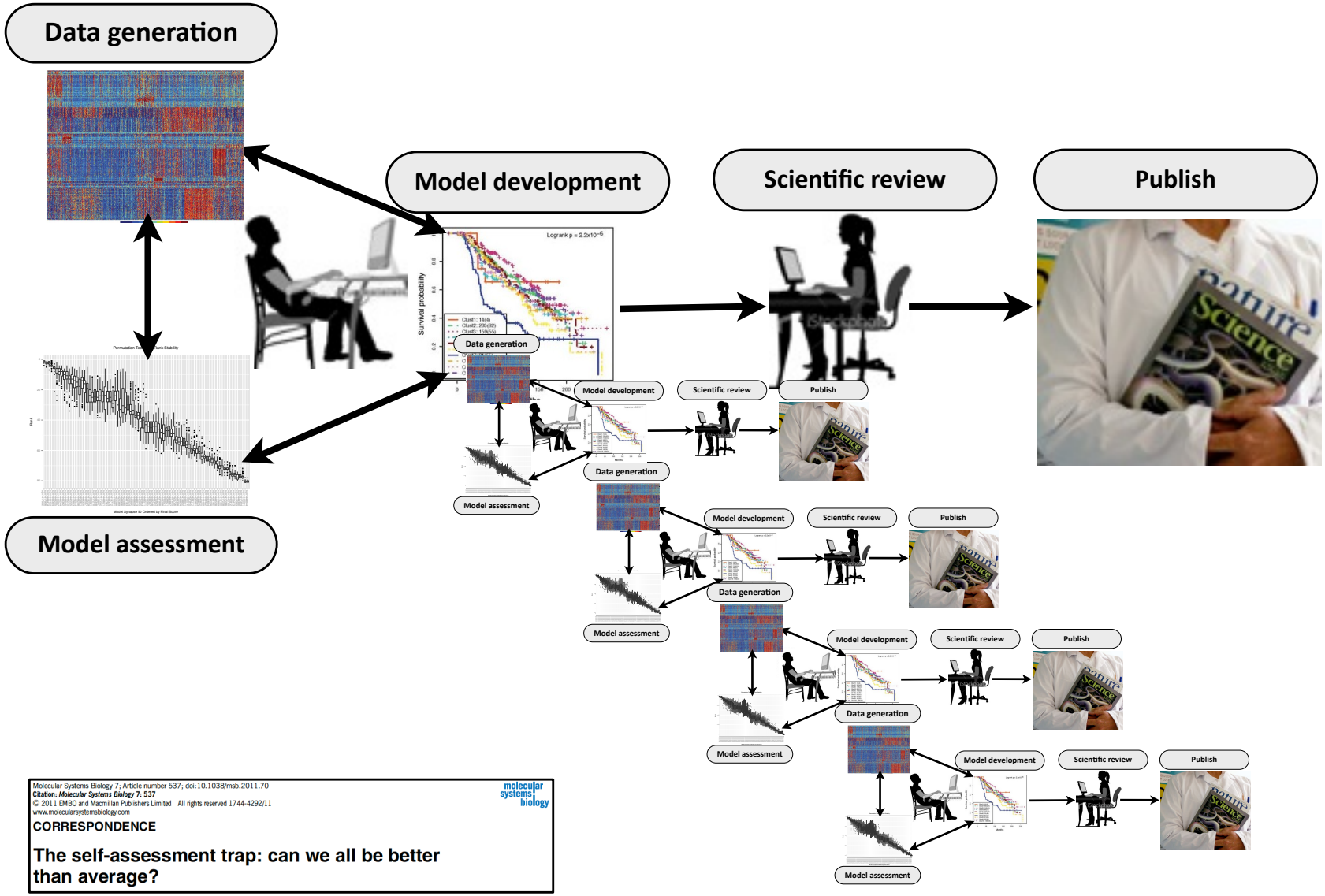
Molecular Systems Biology 7: Article number 537; doi:10.1038/msb.2011.70
Citation: Molecular Systems Biology 7: 537

www.molecularsystemsbiology.com
molecular systems biology

CORRESPONDENCE

The self-assessment trap: can we all be better than average?

The self-assessment trap



Molecular Systems Biology 7; Article number 537; doi:10.1038/msb.2011.70
Citation: *Molecular Systems Biology* 7: 537

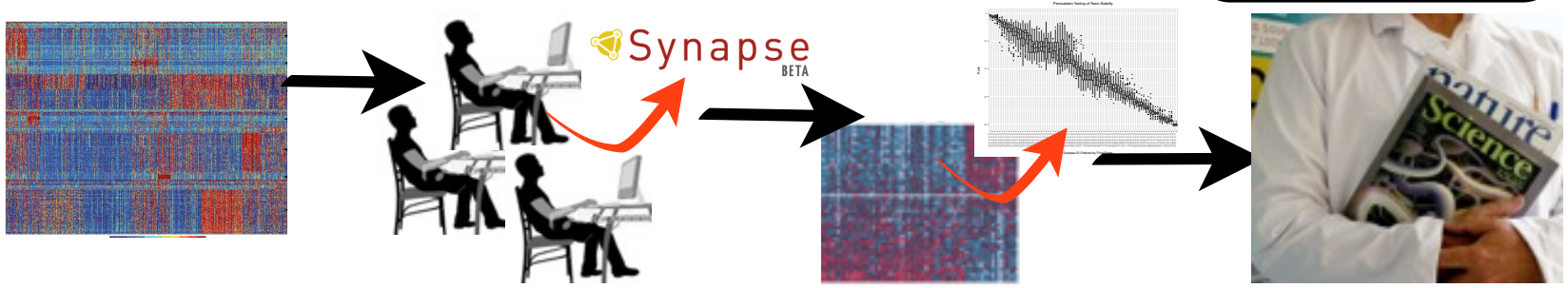
www.molecularsystemsbiology.com
molecular systems biology

CORRESPONDENCE

The self-assessment trap: can we all be better than average?

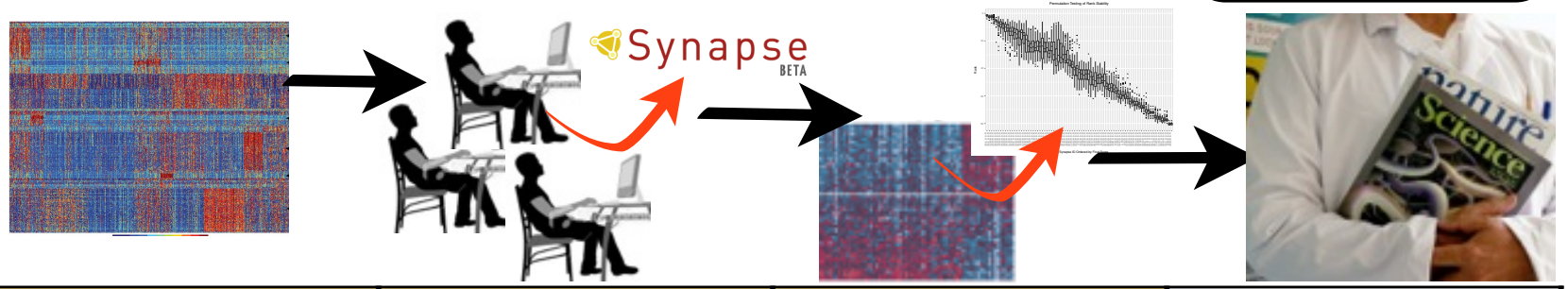
The Sage / DREAM breast cancer prognosis challenge

Training data	Model development and submission to central database	All model predictions evaluated in held-out dataset based on pre-defined criteria	Scientific recognition based on model performance



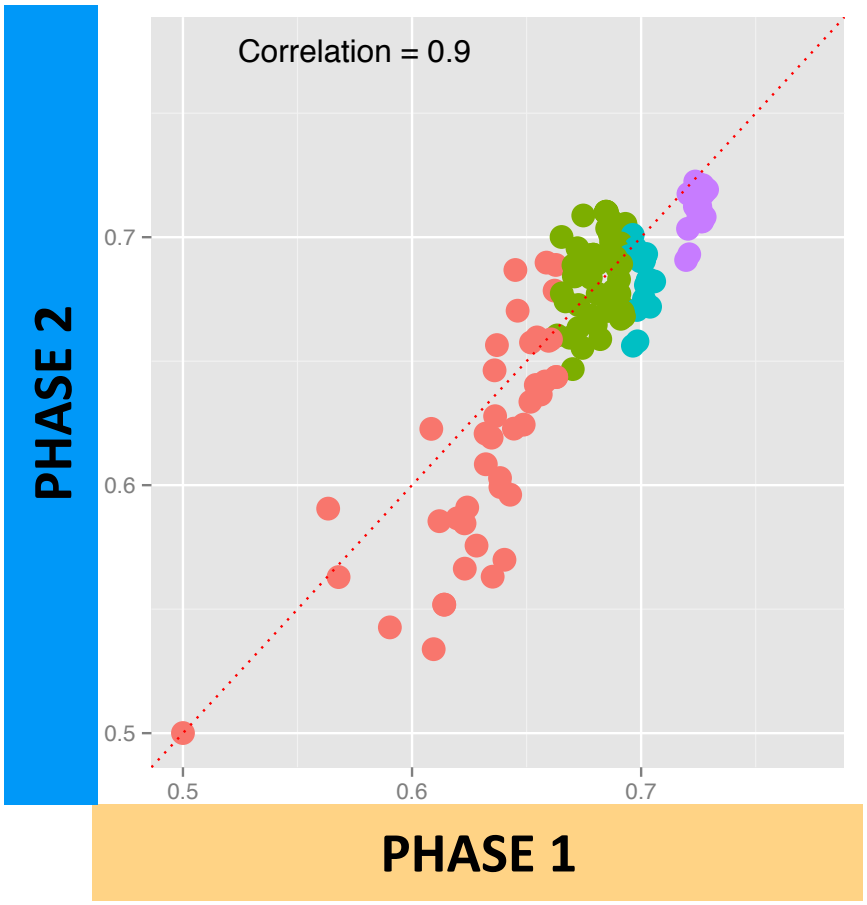
The Sage / DREAM breast cancer prognosis challenge

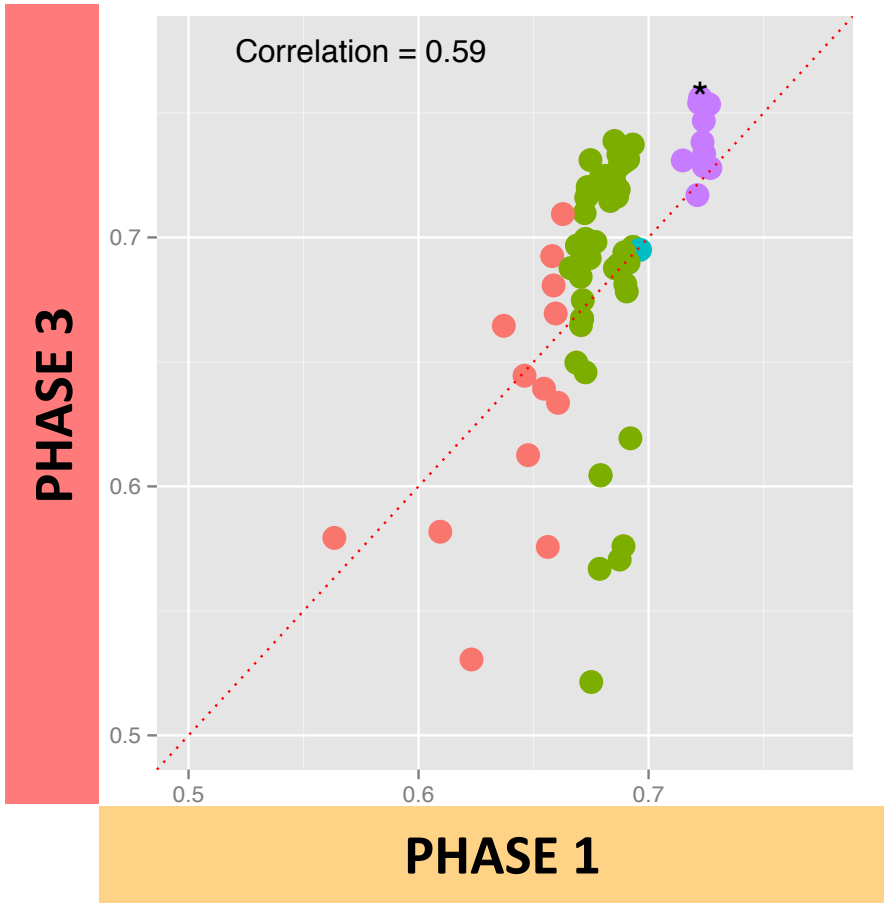
Training data	Model development and submission to central database	All model predictions evaluated in held-out dataset based on pre-defined criteria	Scientific recognition based on model performance



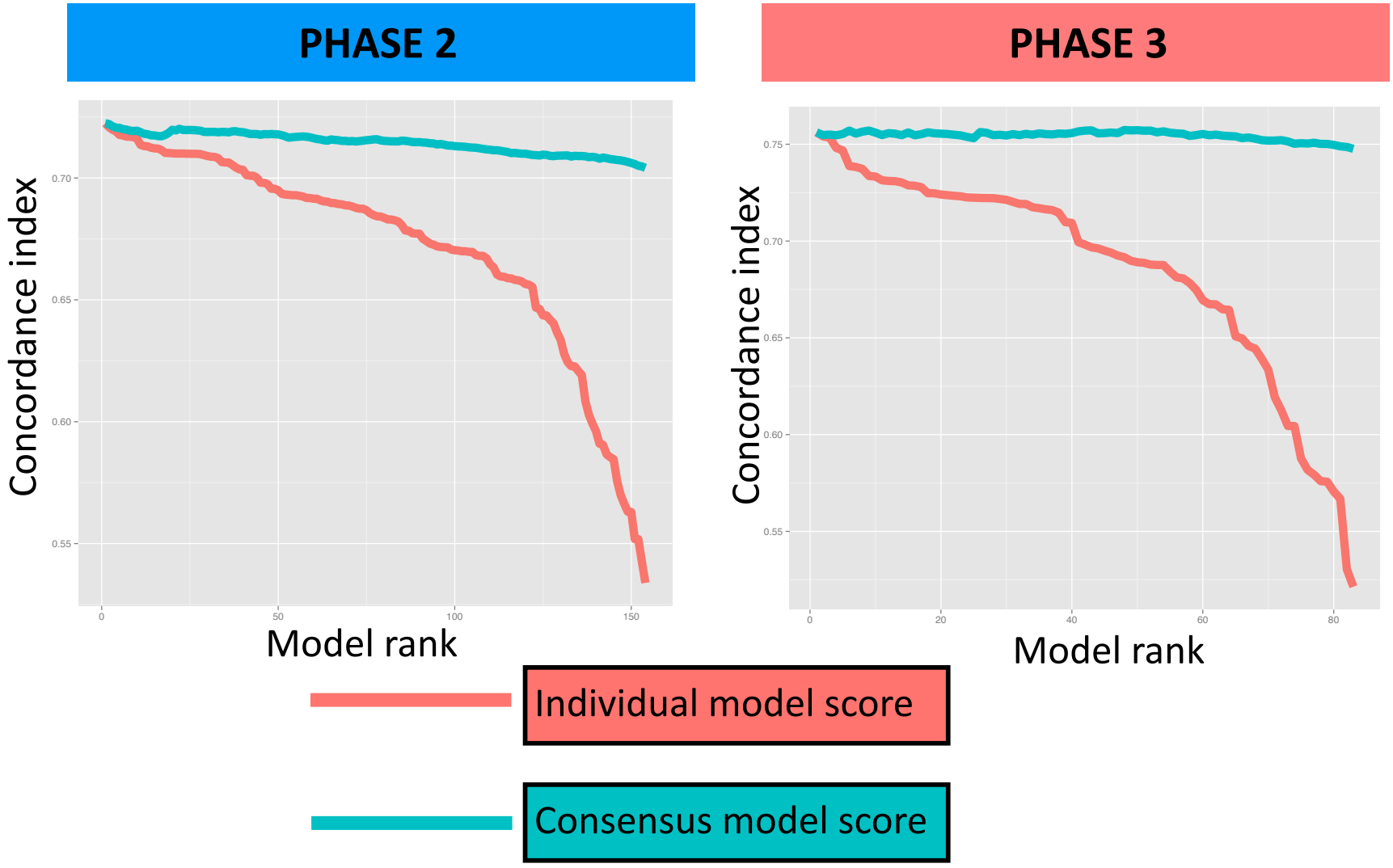
PHASE 1	**1,000 METABRIC samples**	**Unlimited model submission**	**500 held-out METABRIC samples**	**Real-time leaderboard**
PHASE 2	**1,000 METABRIC samples**	**5 models per team**	**Additional 481 held-out METABRIC samples**	**DREAM**
PHASE 3	**All 1,981 METABRIC samples**	**5 models per team**	**184 "OsloVal" samples**	**Science Translational Medicine**

Model concordance index scores are consistent across challenge phases



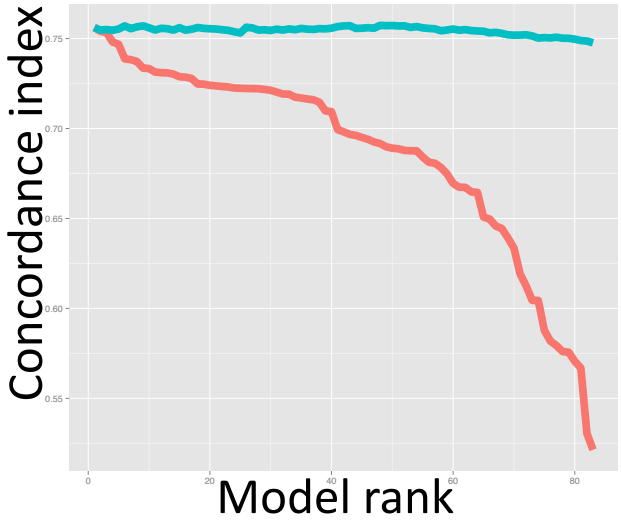


Leveraging the wisdom of crowds yields robust model performance

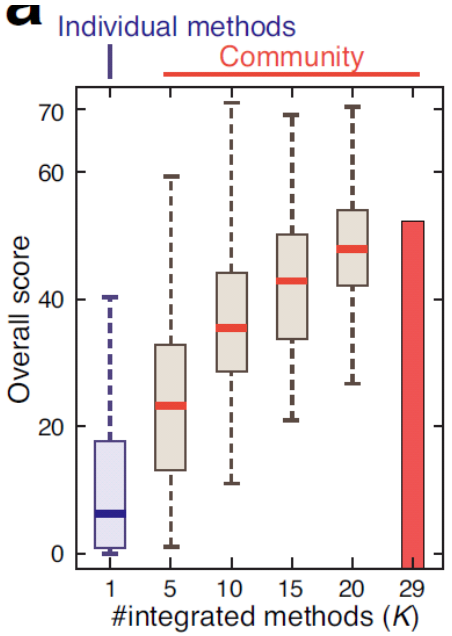


Vox populi across multiple challenges

Breast cancer prognosis

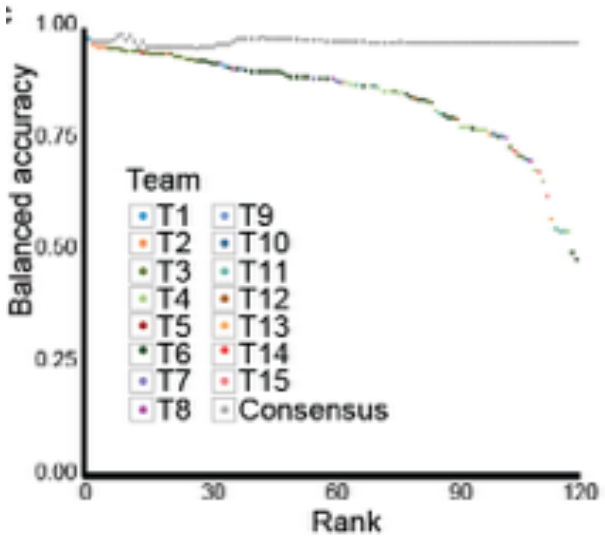


Network inference



Drug sensitivity prediction

Sequence analysis (in silico #1)



17 APRIL 2013, VOL. 5, #181

COVER STORY | RESEARCH ARTICLE AND REPORT FREE

DREAMing of Biomedicine's Future

An Open Challenge yields fresh insights (Margolin *et al.*), a new prognostic model for breast cancer (Cheng *et al.*), and a modern approach to peer review (Editor's Summary).

Sage BIONETWORKS

D R E A M

CROWDSOURCING

Systematic Analysis of Challenge-Driven Improvements in Molecular Prognostic Models for Breast Cancer

Adam A. Margolin,[1*†] Erhan Bilal,[2†] Erich Huang,[1,3,4†] Thea C. Norman,[1] Lars Ottestad,[5] Brigham H. Mecham,[1,6] Ben Sauerwine,[7] Michael R. Kellen,[1] Lara M. Mangravite,[1] Matthew D. Furia,[1,8] Hans Kristian Moen Vollan,[5,9,10,11] Oscar M. Rueda,[11] Justin Guinney,[1] Nicole A. Deflaux,[1] Bruce Hoff,[1] Xavier Schildwachter,[1] Hege G. Russnes,[9,10,12] Daehoon Park,[13] Veronica O. Vang,[9,10] Tyler Pirtle,[7] Lamia Youseff,[7] Craig Citro,[7] Christina Curtis,[14] Vessela N. Kristensen,[9,10,15] Joseph Hellerstein,[7] Stephen H. Friend,[1*] Gustavo Stolovitzky,[2] Samuel Aparicio,[16,17,18†] Carlos Caldas,[11,19,20†] Anne-Lise Børresen-Dale[9,10†]

COLUMBIA UNIVERSITY

RESEARCH ARTICLE

COMPUTATIONAL MODELING

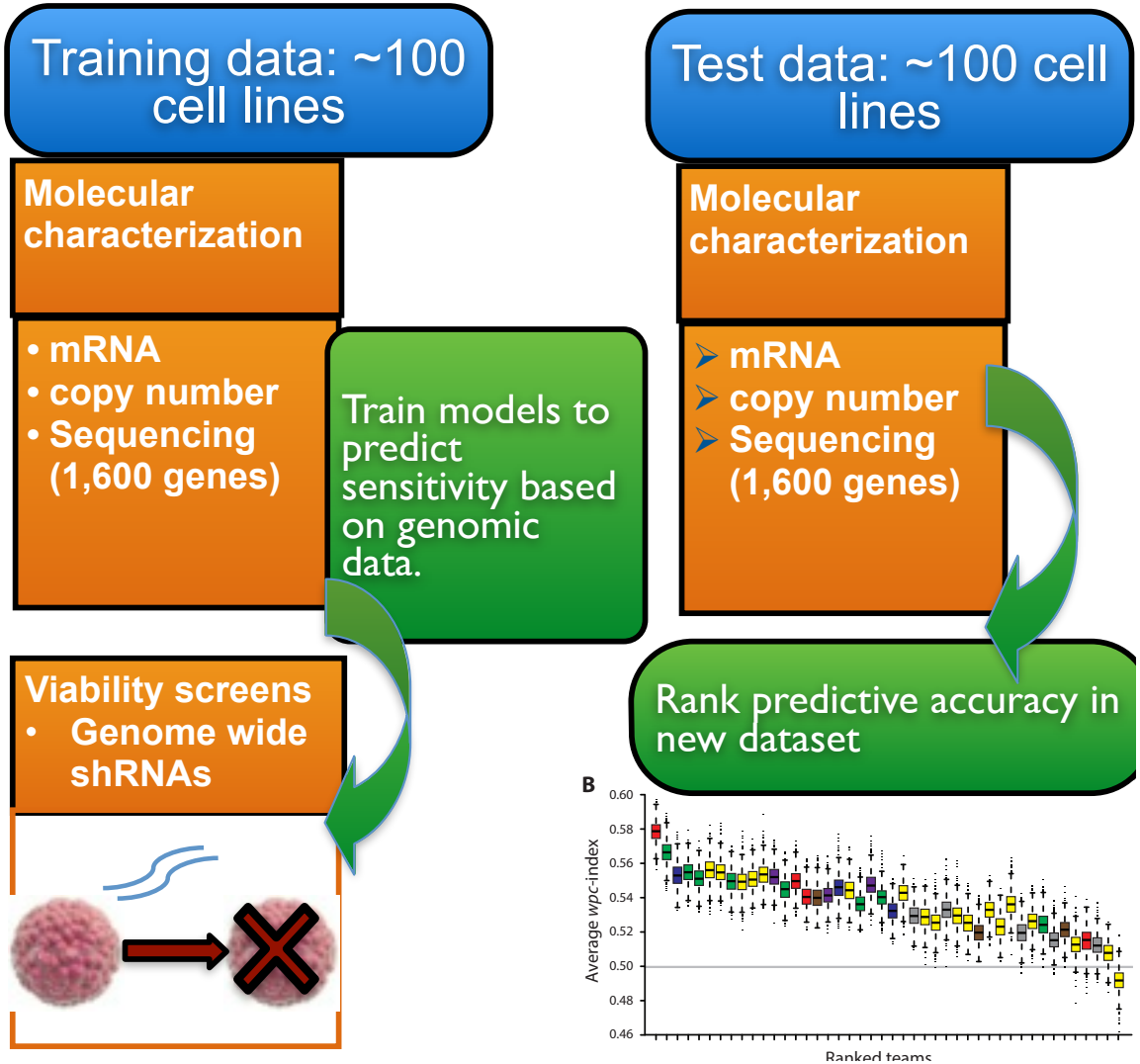
Development of a Prognostic Model for Breast Cancer Survival in an Open Challenge Environment

Wei-Yi Cheng, Tai-Hsien Ou Yang, Dimitris Anastassiou*

All articles free and open access

Broad gene essentiality prediction challenge

- Project Achilles dataset of 200 cell lines.
  - Molecular characterization: expression, sequencing, copy number
  - Genome wide RNAi screens
  - 100 training samples.
  - 100 test samples released with challenge results.
- https://www.synapse.org/#!Challenges:DREAM
  - Or navigate from www.synapse.org

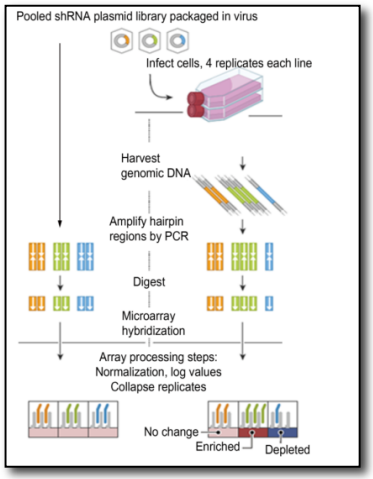


Broad-DREAM Gene Essentiality Prediction Challenge

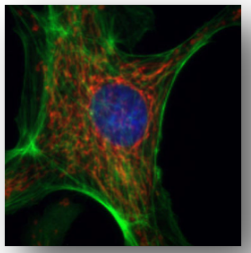
Outline

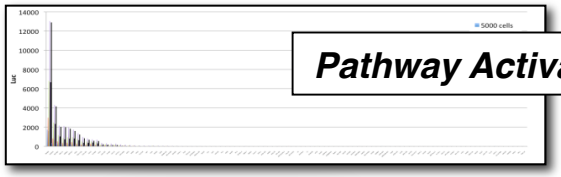
1. DREAM overview (Adam Margolin, OHSU)
2. Scientific overview (Barbara Weir, Aviad Tsherniak, Broad Institute)
3. Challenge details (Mehmet Gonen, Sage Bionetworks)
4. Synapse and forum registration (Bruce Hoff, Sage Bionetworks)
5. Challenge demo (Mehmet Gonen, Sage Bionetworks)
6. Compute resources (Venkat Balagurusamy, IBM)
7. Questions and answers (all)

Project Achilles



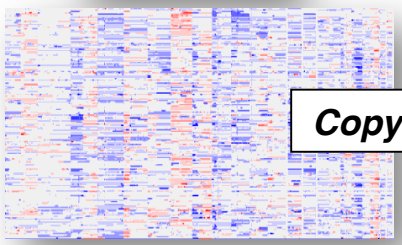
Dependencies



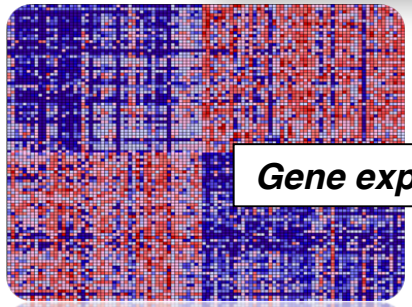


Pathway Activation

Mutational analysis



Copy number



Gene expression

Genome-wide RNAi screens in a large number of cell lines

Extensive molecular characterization of cell lines

Project Achilles

- Project Achilles is a large collaboration between the Broad Institute and the Dana-Farber Cancer Institute to ***identify tumor vulnerabilities that are linked with predictive biomarkers***.
  - **The Broad-DREAM Gene Essentiality Challenge** is designed to help address this goal, specifically to look for models that can predict these vulnerabilities from biomarkers (i.e. molecular features of the tumor)

- Current **public** dataset is Achilles v2.0
  - 'Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer', PNAS 2011 Jul 26;108(30):12372-7.
  - 102 cell lines, array deconvolution of screening data
  - Available, along with additional information about Project Achilles, on the Project Achilles portal (next slide)

- Challenge dataset is **unpublished** and new for you to use!
  - 89 cell lines (half held back for scoring) that also have copy number and expression array data
  - Additional cell lines coming *(data is being generated now)*
    - Challenge will be re-launched in late July / early August when the new data is ready

Project Achilles

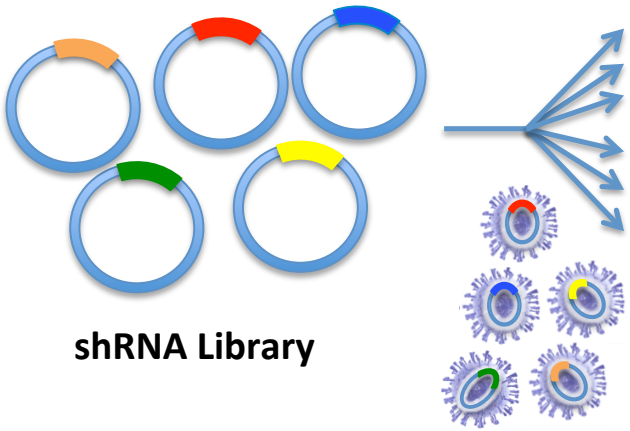
Project Achilles Portal

- Genes can be queried for essentiality values across all cell lines for published data
- Data at the shRNA level and gene-level collapsed by ATARiS are shown and can be:
  - easily downloaded
  - opened with GENE-E
  - launched into our data analysis GenePattern module PARIS.
- More functionality, including data mining tools, is currently being developed.

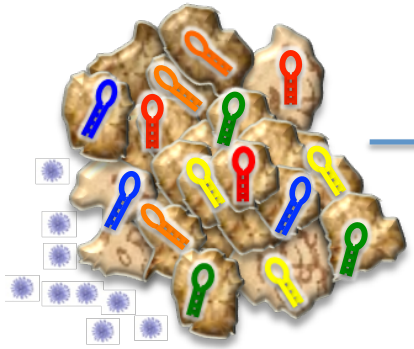
- http://www.broadinstitute.org/achilles

shRNA Loss of Function Screens Overview

Infect a cell line with the shRNA library

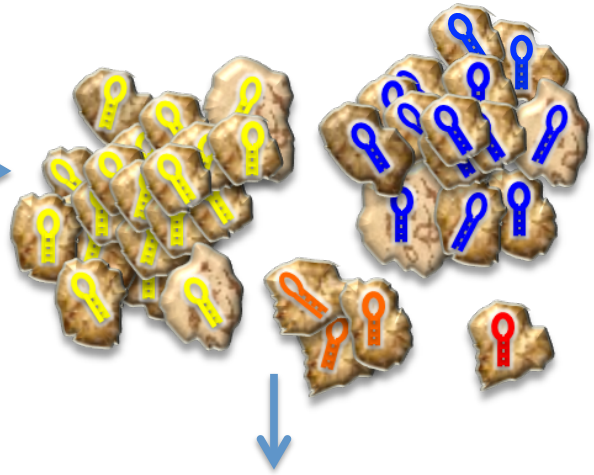


shRNA Library



Cancer Cell Line:
Each cell in the population gets one single shRNA

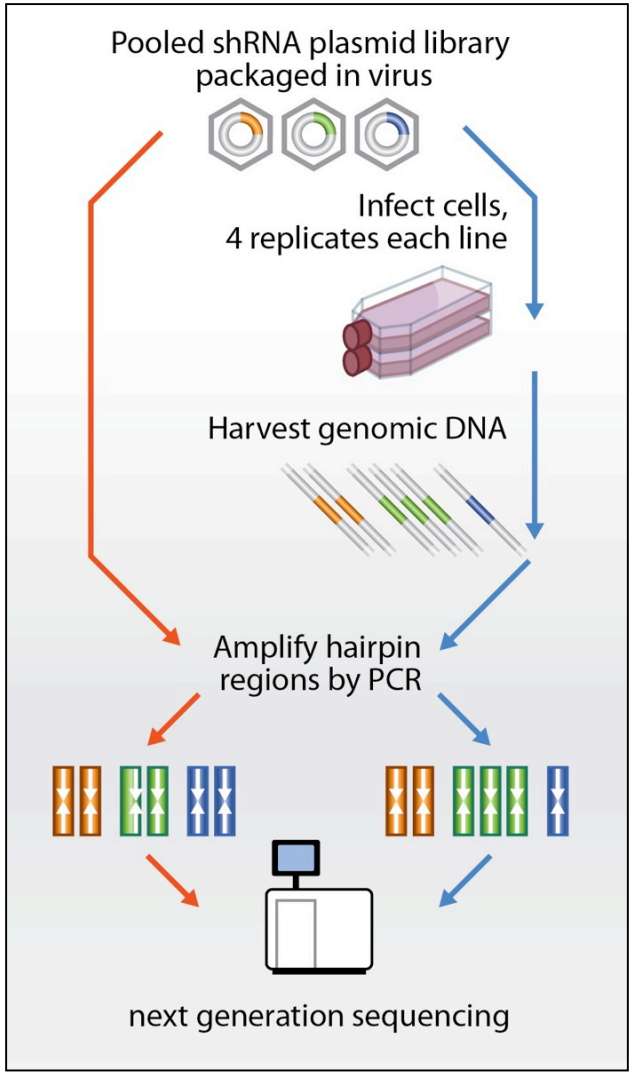
Measure Essentiality based on proportion of shRNAs present at the end of the experiment compared to the start



Readout is Next Generation Sequencing (previously array-based hybridization) of shRNA barcode sequences to get counts per shRNA, per cell line

Red and Green in this example are depleted during the experiment and therefore have a negative effect on viability (are 'essential' to viability). Orange is neutral to viability. Blue and Yellow are enriched during the experiment and therefore have a positive effect on viability.

Project Achilles



Assay Development

Screening

Raw data QC

Final Data Deconvolution

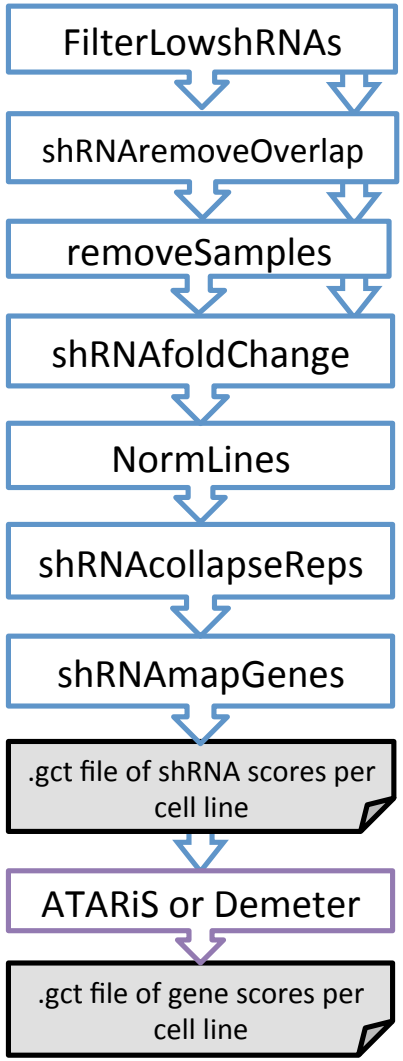
Project Achilles

Data processing GenePattern modules

Quality control GenePattern modules

FPmatching

ReplicatesQC



Removes undesirable shRNAs and cell line replicates:

- Those shRNAs low in initial DNA pool or that overlap in sequence.
- Replicate samples that fail QC

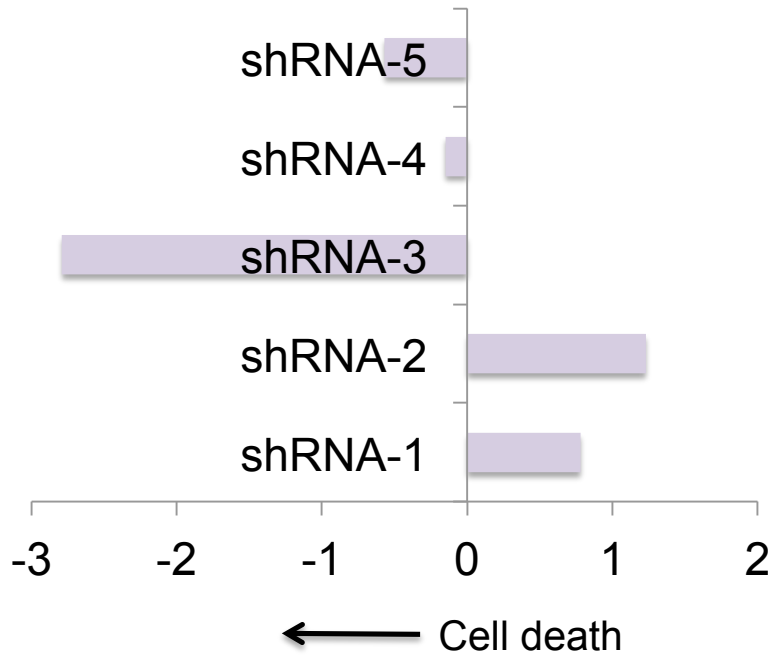
- Calculates fold change values between initial DNA pool values and final cell line values.
- Normalizes cell lines to the same scale (quantile, ZMAD, PMAD)
- Collapses replicate cell line samples to a single values per shRNA
- Maps shRNAs to gene symbols based on a mapping file

http://genepattern.broadinstitute.org/gp/

RNAi Data is Challenging

- shRNAs targeting the same gene produce different effects
- Sources of variability
  - Varying suppression levels
  - Off-target effects
  - Noise
- No experimental way to fully assess shRNA performance
  - Especially off-target suppression
- Naïve averaging of the values will hurt the signal

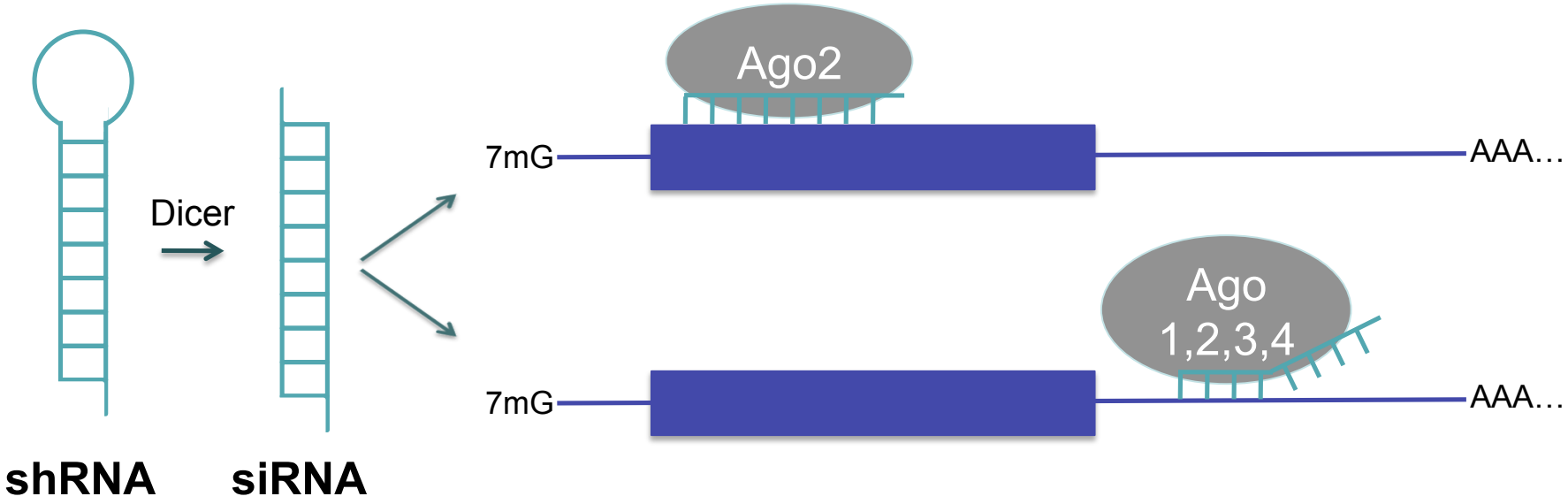
Viability data for one gene in one sample



Two pathways for shRNAs

RNAi

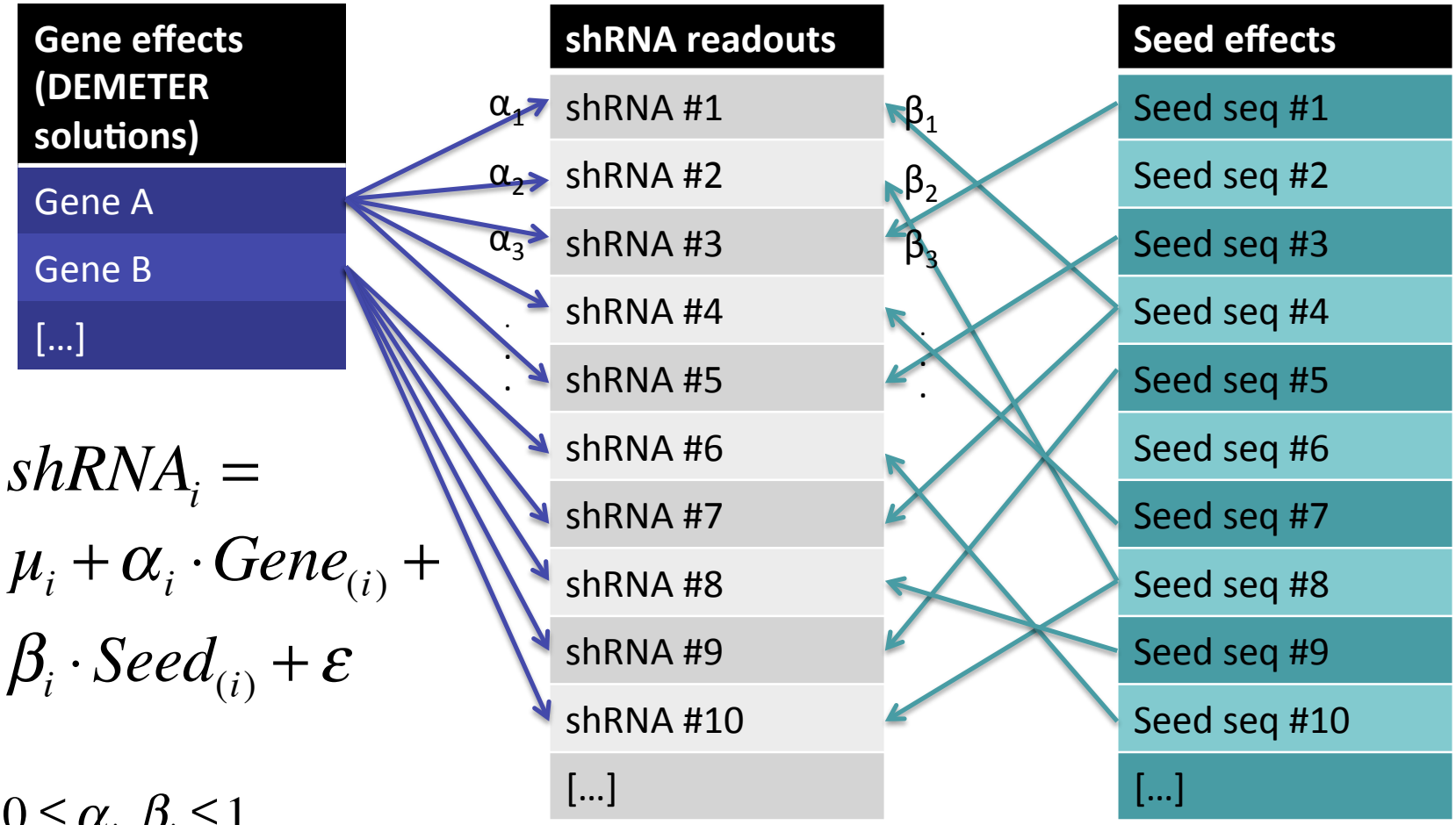
Perfect (or near-perfect) match to mRNA causes mRNA cleavage and degradation: **18 – 22 nt**



microRNA

'Seed' region binds to 3'UTR to represses translation and/or destabilize mRNA: **6 – 8 nt**

DEMETER infers gene dependencies by explicitly modeling both gene- and seed-effects



$$shRNA_i = \mu_i + \alpha_i \cdot Gene_{(i)} + \beta_i \cdot Seed_{(i)} + \varepsilon$$

$$0 \le \alpha_i, \beta_i \le 1$$

Challenge Questions

1.1 Challenge Questions & Scoring

Edit Wiki | + Add a new Page

The goal of this project is to use a crowd-based competition to develop predictive models that can infer gene dependencies (genes that are essential to cell viability when suppressed) in cancer cells using features of the cell line.

Participants will be asked to solve three sub-challenges:

> **1.** Build a model that best predicts the gene essentialities of **14557** genes, using the molecular characteristics/features of the cancer cell lines.
>
> In this sub-challenge any data can be used for prediction.

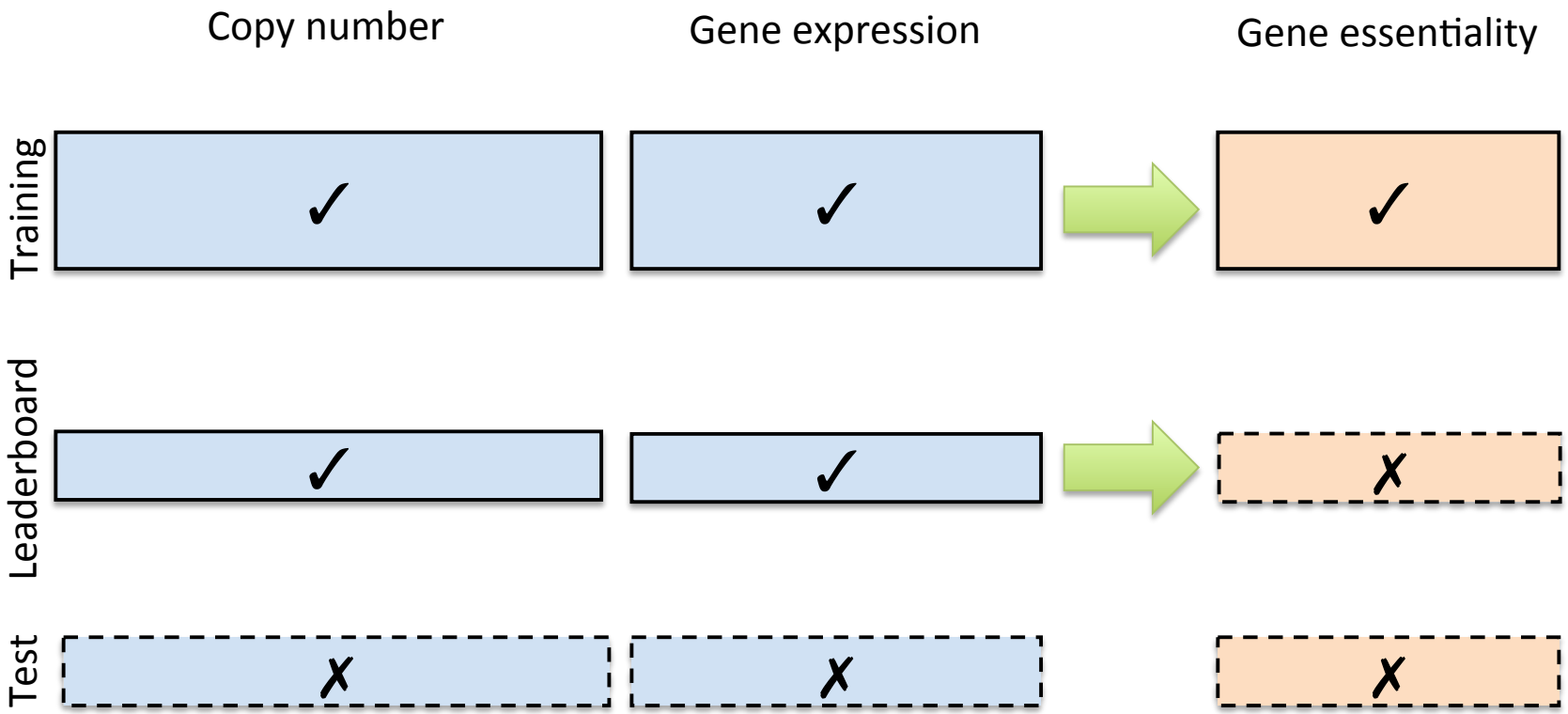
> **2.** Identify the most predictive features for **each** gene essentiality of a prioritized list of **2637** genes. For each prioritized gene the aim is to: (1) select a small set of at most **10** predictive features (gene expression and copy-number) and (2) predict gene essentiality using **only** these features.
>
> In this sub-challenge any data can be used for choosing the gene expression and copy-number features, but only those selected features can be used for prediction.

> **3.** Identify the most predictive features for **all** gene essentialities of a prioritized list of **2637** genes. For the set of all prioritized genes, the aim is to: (1) identify a single list of at most **100** general predictive features (gene expression and copy-number) and (2) predict essentiality using **only** these features for **all** prioritized genes. In machine learning, this problem is known as multilabel or multitask feature selection.
>
> In this sub-challenge any data can be used for choosing the gene expression and copy-number features, but only those selected features can be used for prediction.

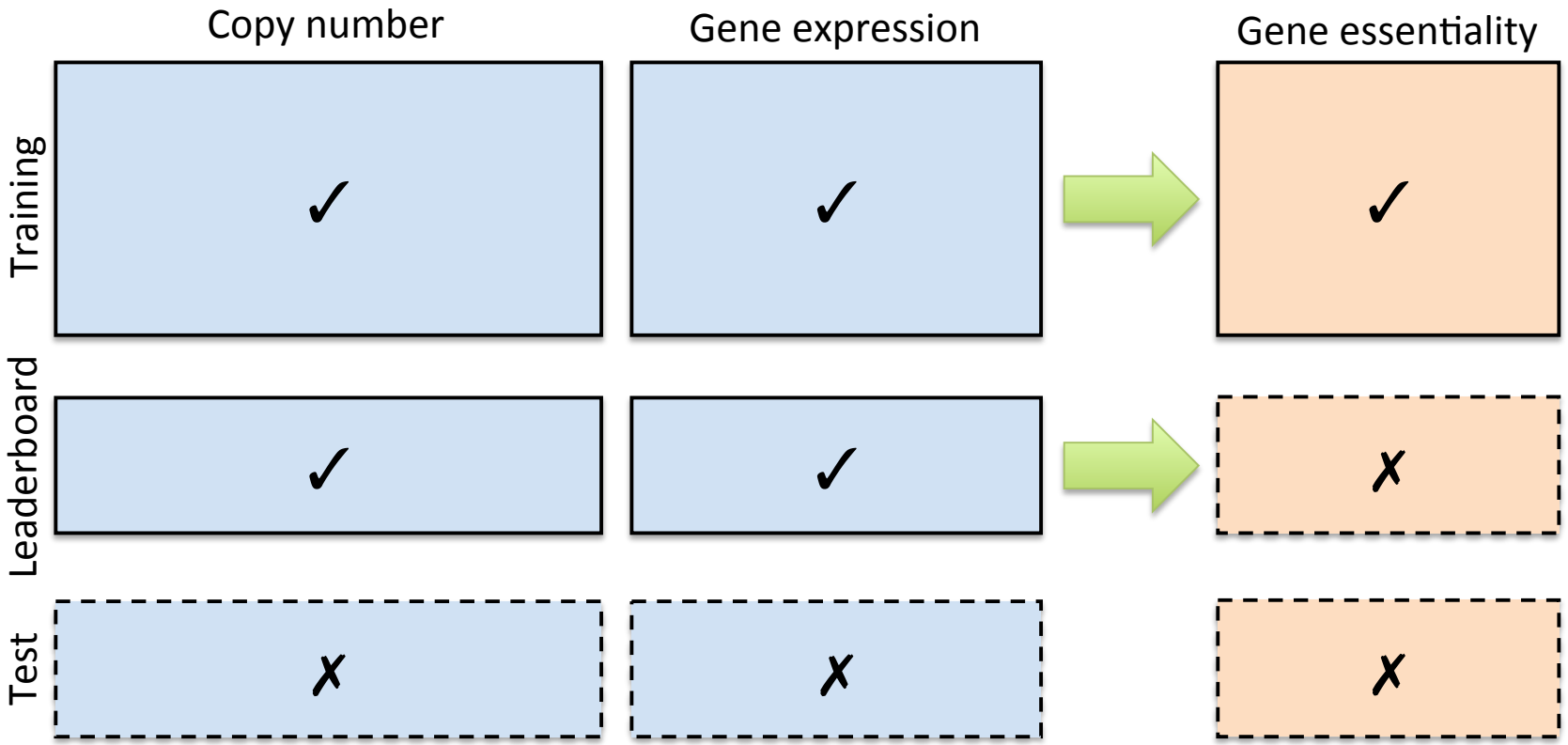
Challenge Timeline

- Phase I (89 cell lines): till end July
  - 45 in training set (fully available)
  - 22 in leaderboard set (hidden gene essentiality)
  - 22 in test set (fully hidden)
- Phase II (≈200 cell lines): till end August
  - ≈100 in training set (fully available)
  - ≈50 in leaderboard set (hidden gene essentiality)
  - ≈50 in test set (fully hidden)
- Phase III (≈200 cell lines): till mid September
  - ≈150 in training set (fully available)
  - ≈50 in test set (hidden gene essentiality)

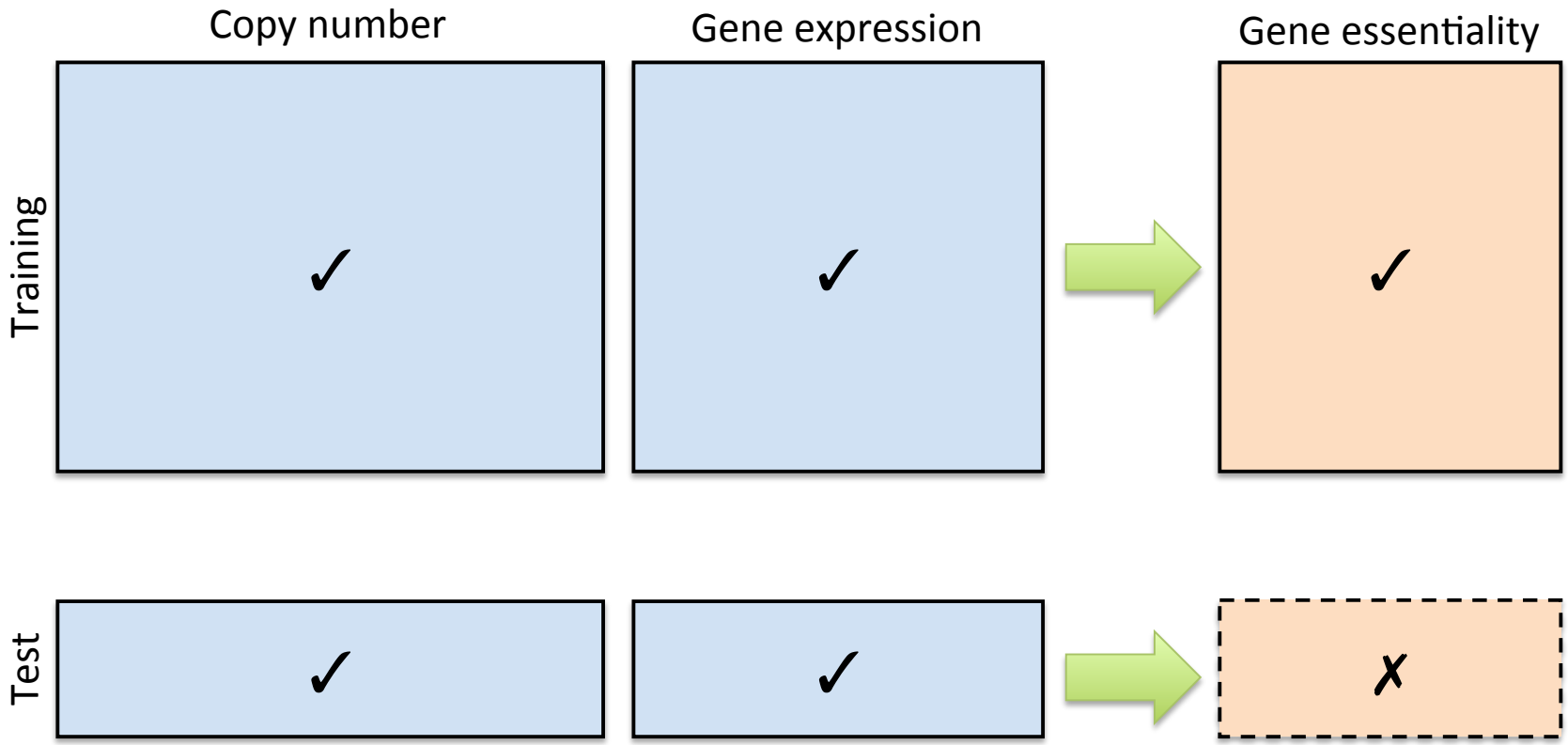
Overall Challenge Setup (Phase I)



Overall Challenge Setup (Phase II)

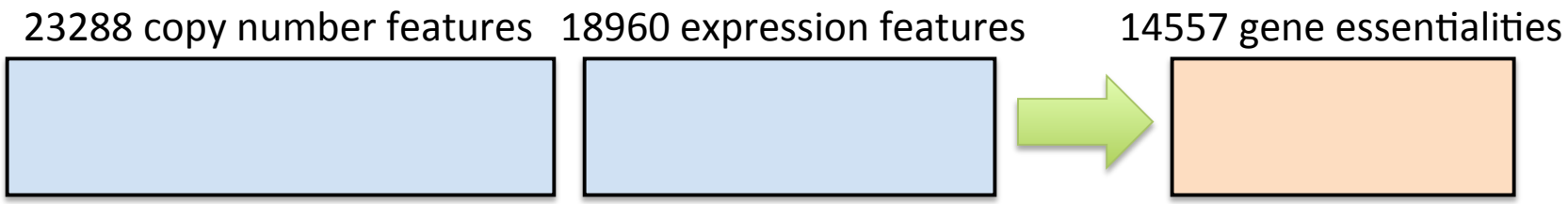


Overall Challenge Setup (Phase III)



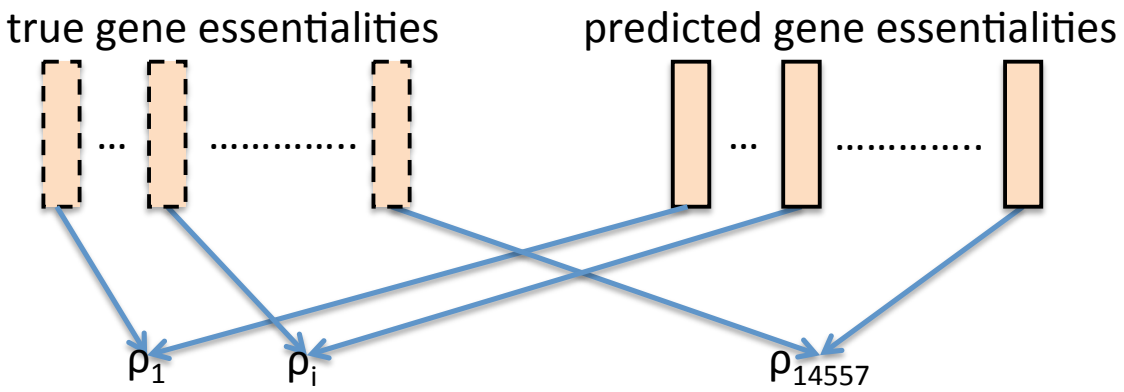
Sub-challenge 1 Question

- Build a model that best predicts the gene essentialities of thousands of genes, using the molecular characteristics/features of the cancer cell lines



Sub-challenge 1 Scoring

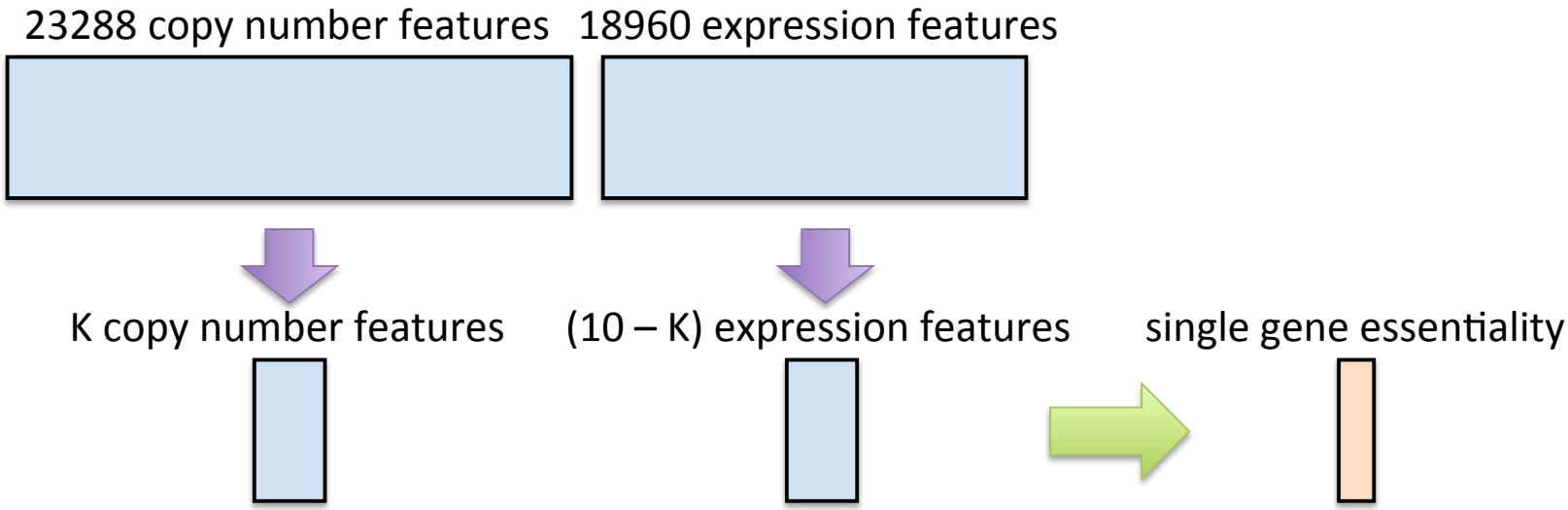
- We use Spearman's rank correlation coefficient to evaluate the performance



- Overall score is the mean of 14557 correlation values

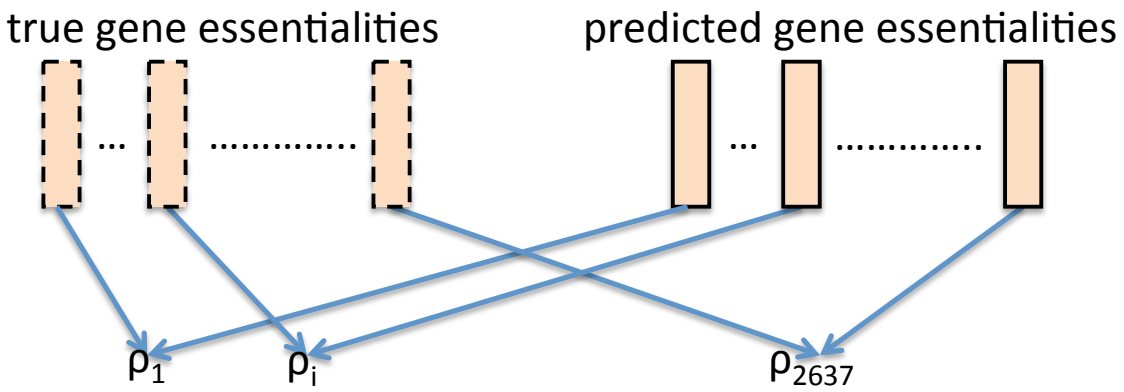
Sub-challenge 2 Question

- Identify the most predictive features for each gene essentiality of a prioritized list of 2637 genes



Sub-challenge 2 Scoring

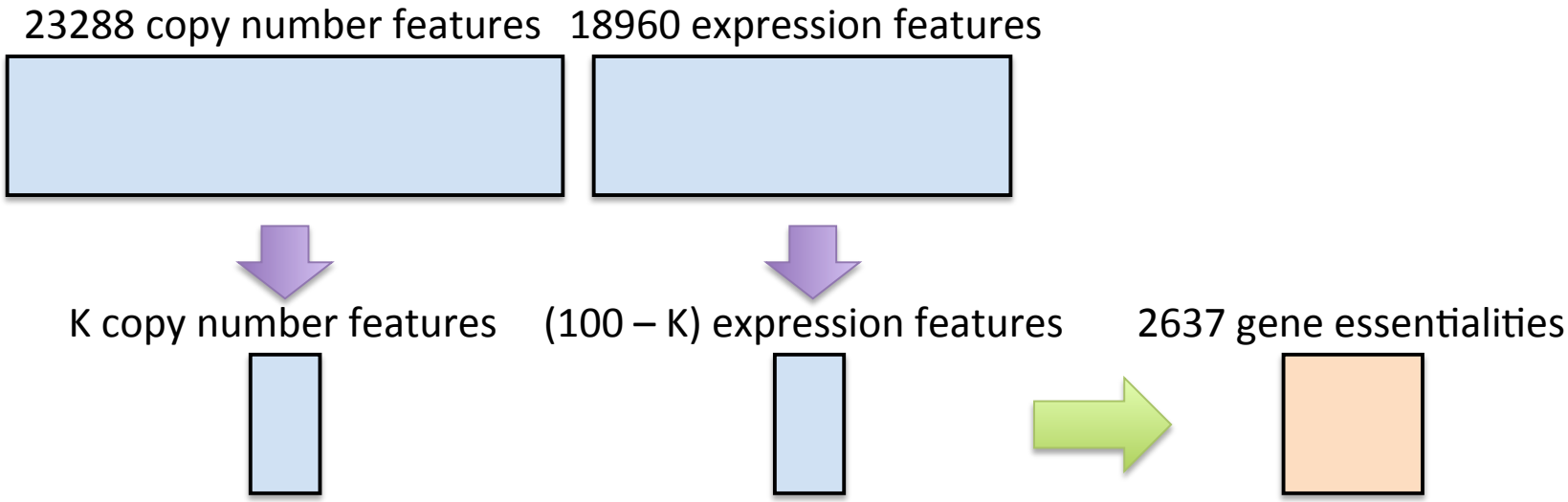
- We use Spearman's rank correlation coefficient to evaluate the performance



- Overall score is the mean of 2637 correlation values

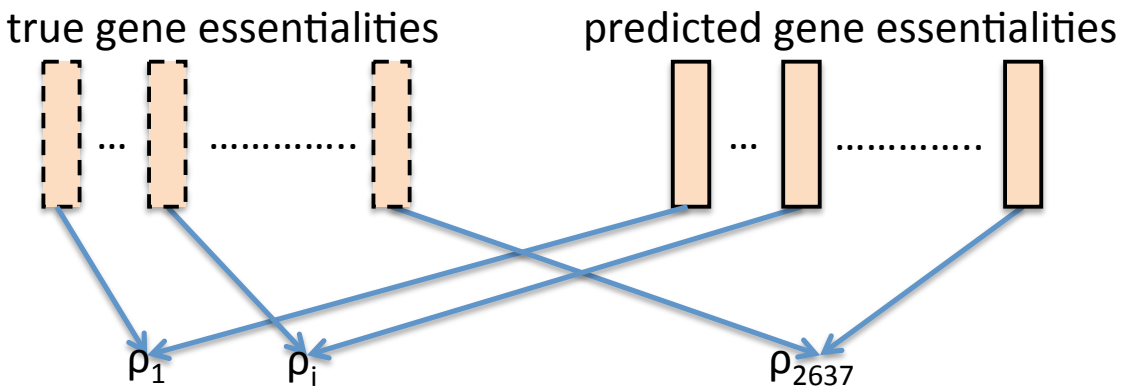
Sub-challenge 3 Question

- Identify the most predictive features for all gene essentialities of a prioritized list of 2637 genes



Sub-challenge 3 Scoring

- We use Spearman's rank correlation coefficient to evaluate the performance



- Overall score is the mean of 2637 correlation values

External Data Usage Policy

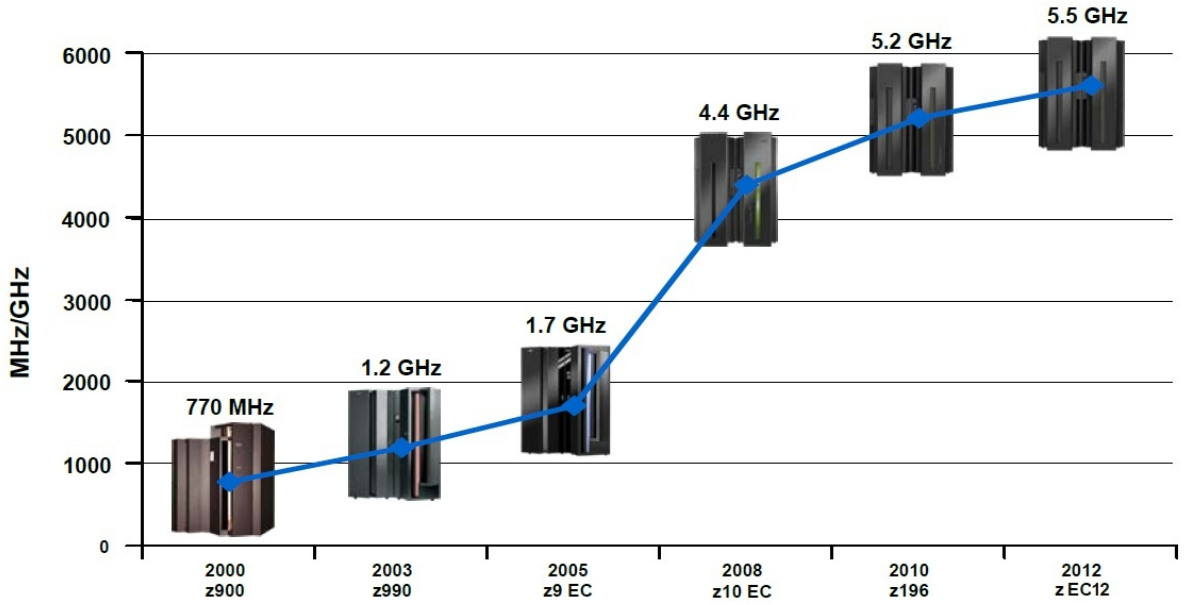
- Sub-challenge 1:
  - Other data sources can be used freely
- Sub-challenges 2 and 3:
  - Other data sources can be used for feature selection step only
  - Prediction step must be performed using selected copy number and expression features only

IBM Computing Cluster for Gene Essentiality Prediction Challenge

IBM System Z team

June 26, 2014

IBM System Z – fastest processor in the Industry



Available for the GEP Challenge – 20 processors, 297 GB memory, 1 TB storage (zpublic server)

http://en.wikipedia.org/wiki/IBM_zEC12_%28microprocessor%29

How to get a user account in IBM System Z server

- Once a challenge user is approved by Sage to access the data, they become **eligible to get** a user account in IBM System Z Server
- IBM creates userid login and password and sends the participant an email message (at least once a day)
- **Welcome message and login details** will be sent to each approved user

```
Subject: The DREAM  Gene Essentiality Challenge: Your IBM server account

Dear DREAM challenge participant,


The IBM System Z Team (Contact: Venkat Balagurusamy; Email: vkbalagu@us.ibm.com)

IP address:  63.90.228.10
Internet host name :  zpublic.wsc.ihost.com
```

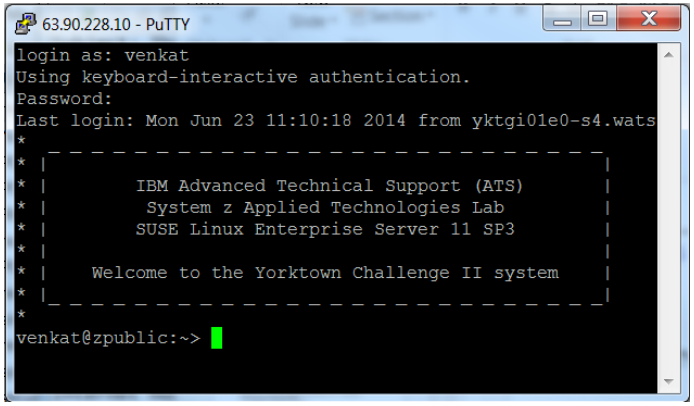
How to access the IBM System Z server

Once you have received the login details and the welcome message from Sagebase website (by email), you are ready to login to IBM System Z server

- **Windows users** – can use a secure shell client e.g., **PuTTY** to connect to the server with their login details
  - **File transfer**: can use e.g., **WinSCP**
- **Mac users** – use "ssh" command in a **terminal** window in their macbook
- **Command format: "ssh** userid@zpublic.wsc.ihost.com**" OR "** userid@63.90.228.10 **"**
- **Password change after first login:** After their first successful login with the initial password (case-sensitive), they will be asked to change the password immediately. When they change it, the system will disconnect and you can relogin with the changed password.

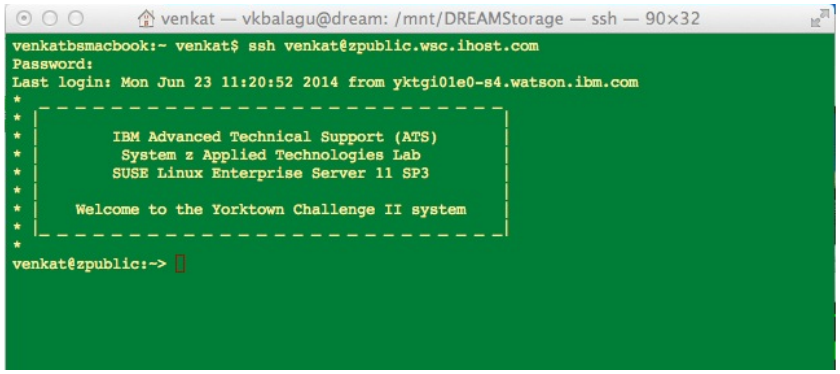
Screen shots of the server after changing password and login

Windows screen (PuTTY)



IBM System Z runs in SUSE Linux OS

Macbook screen (terminal)

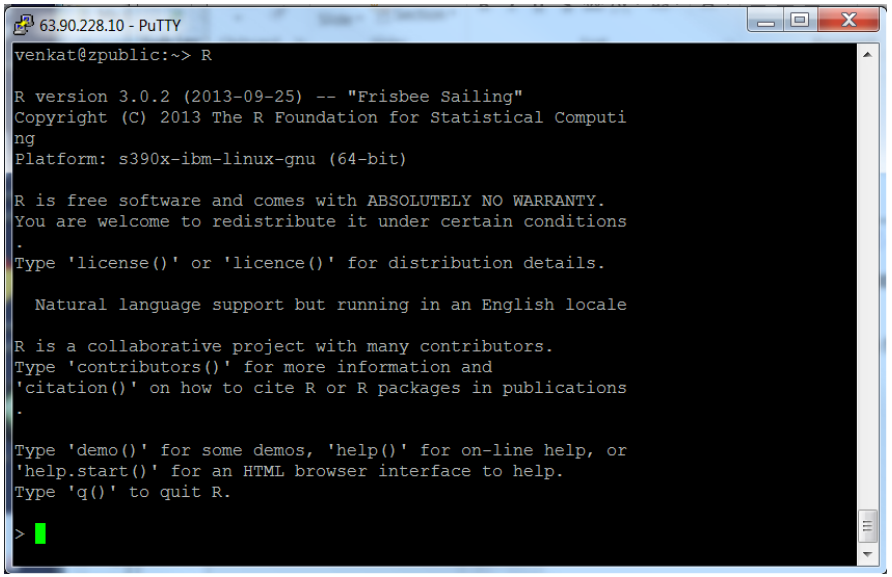


Can use ssh command in other Linux/Unix systems

Running the software suite R – statistical analysis and graphics tool

R with many of the commonly used packages are installed and running in the server

"R" can be started in the shell, by simply typing "R"



Software suite PLINK is also available.
Octave is also available

Other freely downloadable softwares can be installed on request

User account folder path in the server: /mnt/DREAMStorage

Challenge data: As the data size is only ~ 50 MB, users can download their own copy !

Best practices for using the server

Efficient memory and cpu use:

- Jobs submitted by users will be sharing a total memory of 297 GB
- So, if users submit large jobs that take significant amount of memory, say 25 to 30 GB, they are requested to consider efficient use of memory. This will enable multiple users to run many jobs at the same time among the 20 available processors
- Coordinate with Venkat when you need to run a large memory job that will restrict other users. He will **send a note to other users** to review a particular time slot for a job that requires large memory.
- For jobs that take *very long* (possibly many hours) users can *parallelize* them **to run in more than one processor** wherever possible to have their results *fast*

Contact details for help with problems

For problems related to login, running R packages, installing new R packages, and any other questions related to using the server:

Contact: Venkat Balagurusamy, IBM System Z team

Email: vkbalagu@us.ibm.com

Acknowledgements: Gustavo Stolovitzky, Program Director, CBC (IBM)
Donna Dillenberger, Manager, System Z group (IBM)
DREAM and Sage team

THANK YOU and HAVE FUN SOLVING THE CHALLENGE !