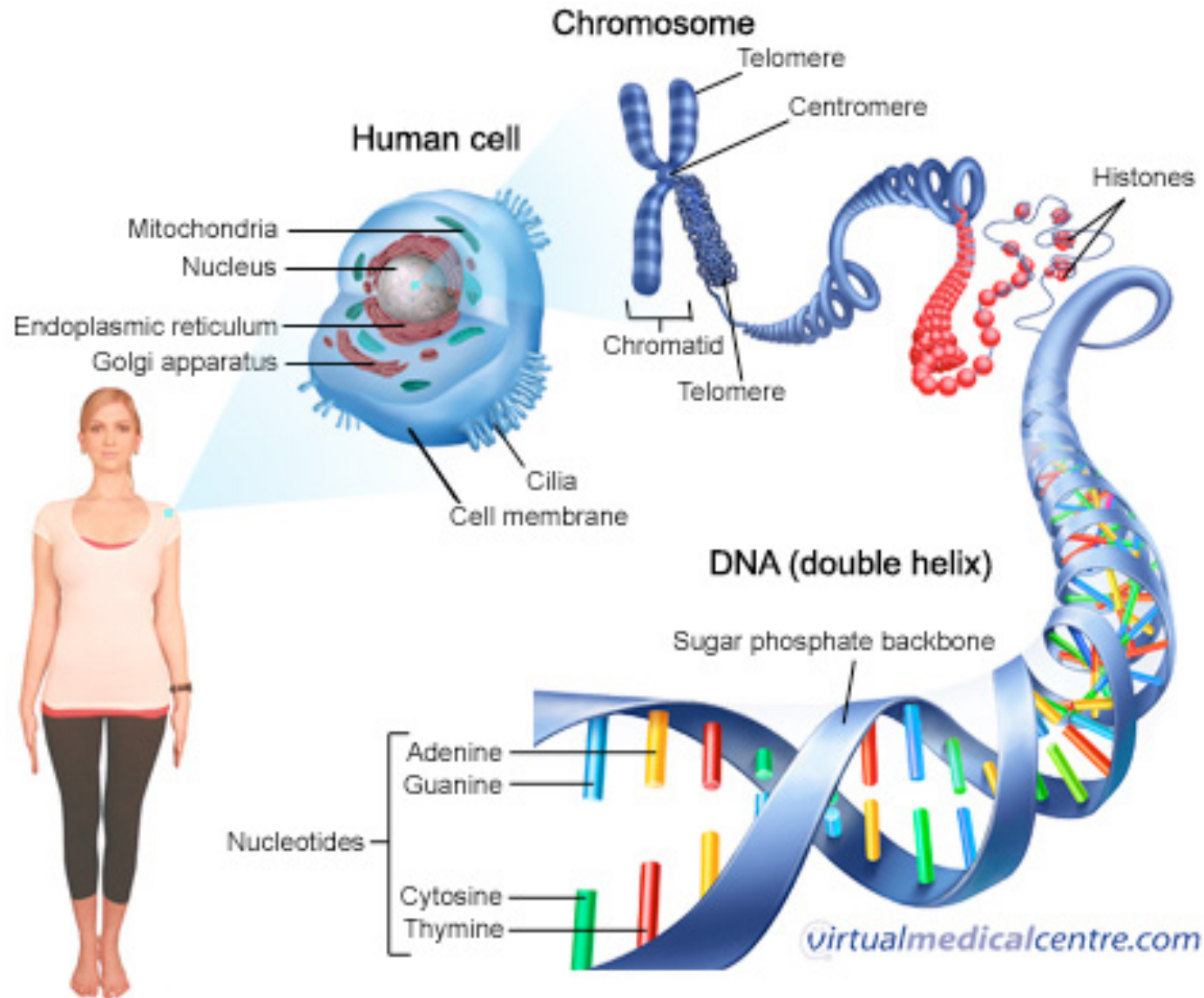


Machine Learning for Personalized Medicine

Jean-Philippe Vert



DNA = 6 billions ACGT





Human genome project (1990-2003)

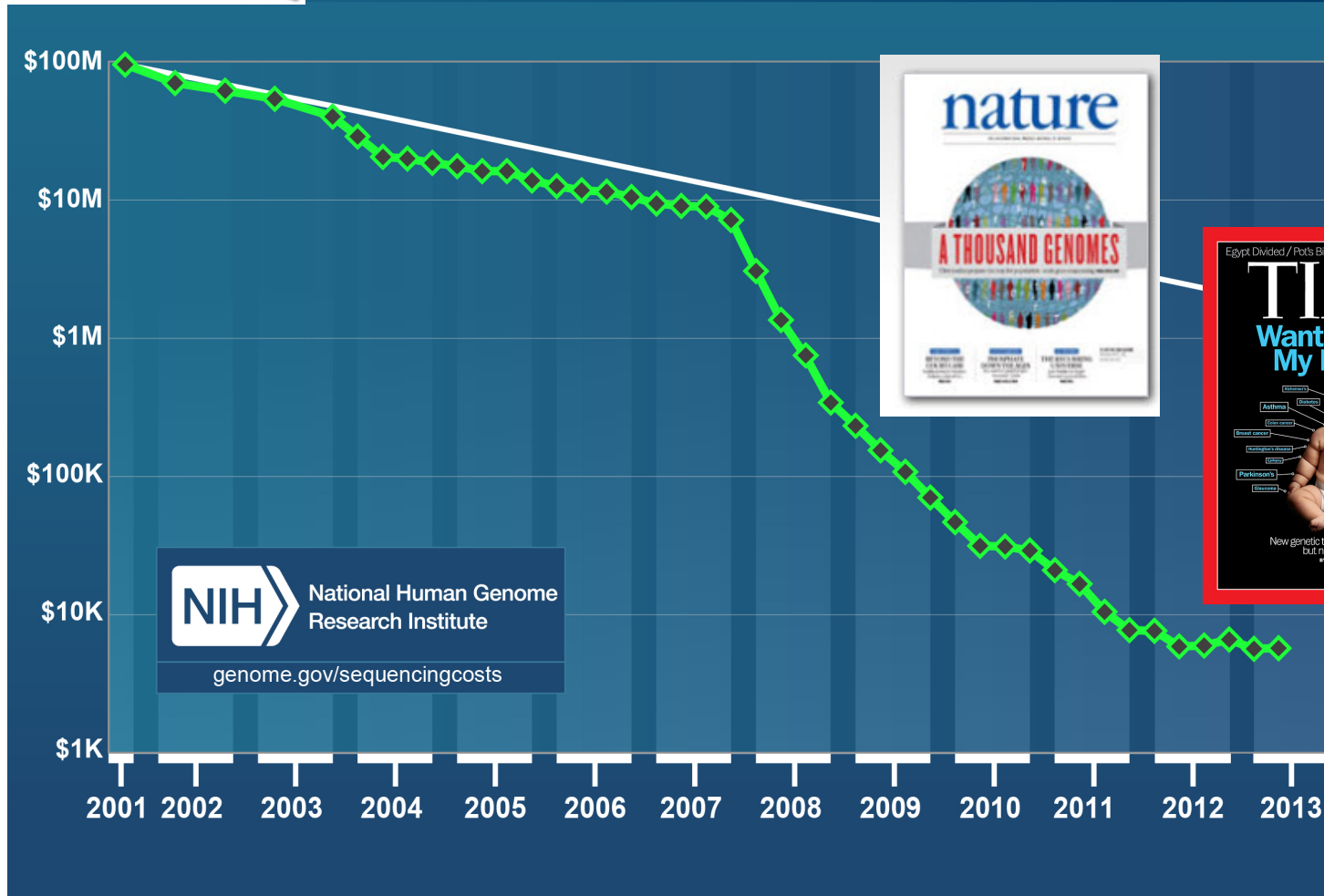
- Goal: sequence the 3,000,000,000 base pairs of the human genome
- Consortium of 20 laboratories, 6 countries
- 13 years, \$3,000,000,000



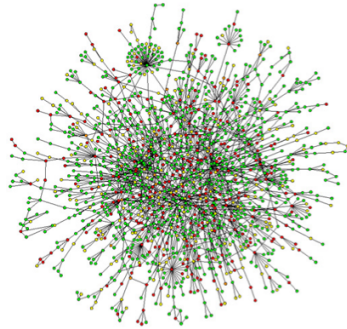


The *second* revolution

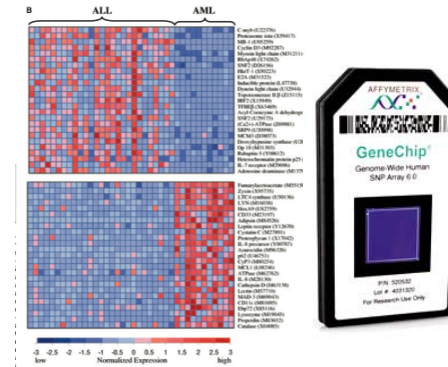
Cost per Genome



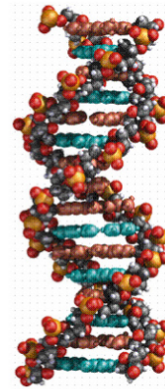
A flood of *omics* data



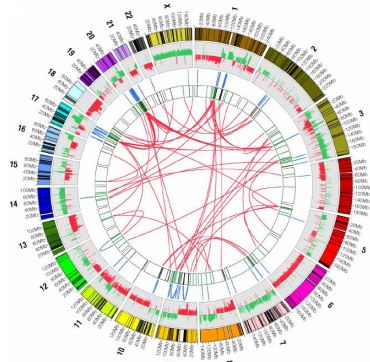
Interactome



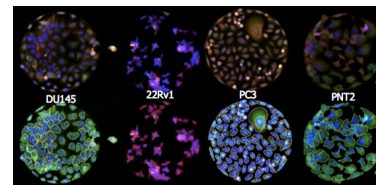
Transcriptome



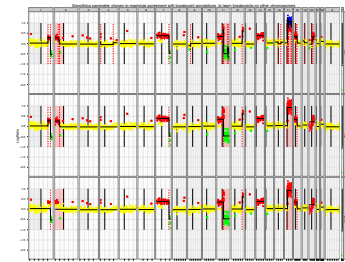
Genome



Mutations
Structural variations

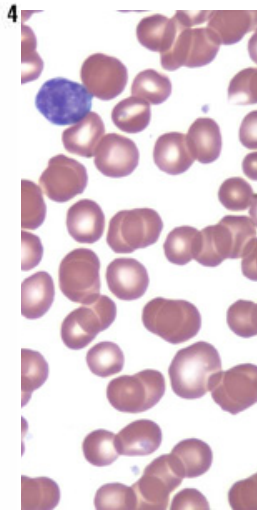
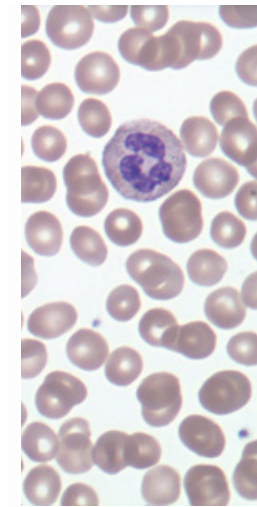
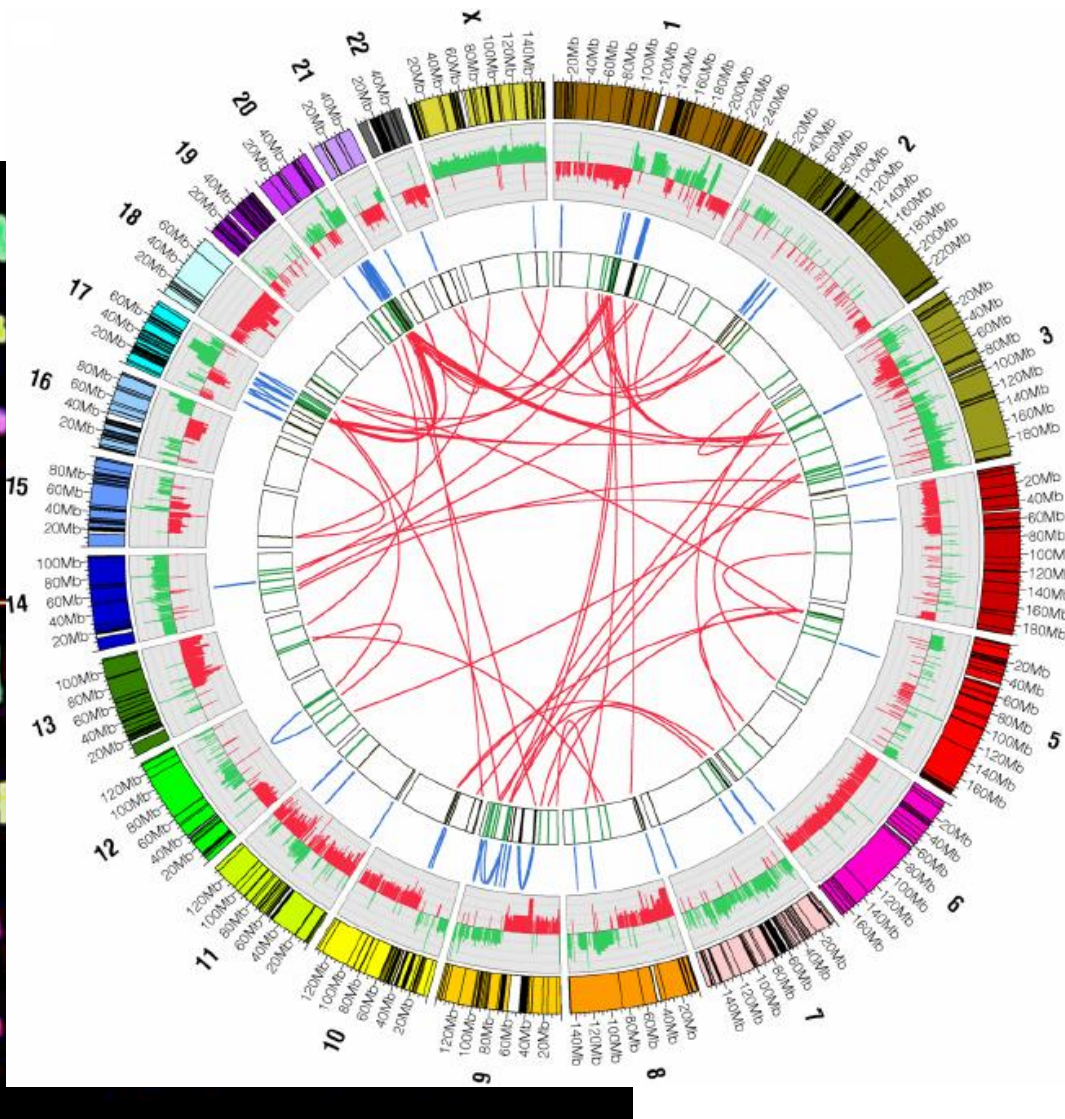


Phenome



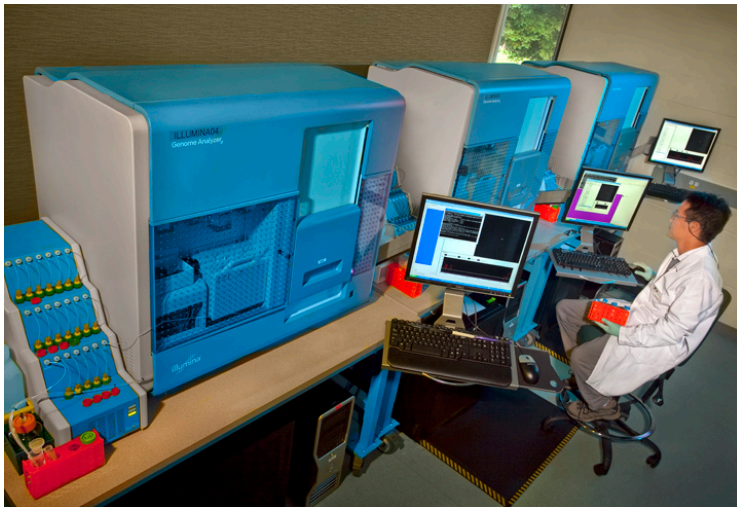
Epigenome

Cancer: different views



Big data!

- <http://aws.amazon.com/1000genomes/>



A screenshot of the International Cancer Genome Consortium (ICGC) website. The browser address bar shows 'http://www.icgc.org/'. The website features a navigation menu with 'Overview', 'Cancer Genome Projects', 'Committees', 'Policies and Guidelines', 'Media', and 'Contacts'. The main heading is 'International Cancer Genome Consortium'. Below this, there is a central graphic of a chromosome and a text box stating: 'will facilitate communication among the members and provide a forum for coordination with the objective of maximizing efficiency among the scientists working to understand, treat, and prevent these diseases.' To the left and right of the central graphic are lists of cancer types and their associated countries, such as 'Bladder Cancer' (United States), 'Blood Cancer' (United States), 'Bone Cancer' (United Kingdom), 'Brain Cancer' (United States), 'Breast Cancer' (European Union / United Kingdom), 'Breast Cancer' (France), 'Breast Cancer' (United Kingdom), 'Breast Cancer' (United States), 'Cervical Cancer' (United States), 'Chronic Lymphocytic Leukemia' (Spain), 'Chronic Myeloid Disorders' (United Kingdom), 'Colon Cancer' (United States), 'Endometrial Cancer' (United States), 'Gastric Cancer', 'Liver Cancer' (Japan), 'Liver Cancer' (United States), 'Lung Cancer' (United States), 'Malignant Lymphoma' (Germany), 'Oral Cancer' (India), 'Ovarian Cancer' (Australia), 'Ovarian Cancer' (United States), 'Pancreatic Cancer' (Australia), 'Pancreatic Cancer' (Canada), 'Pediatric Brain Tumors' (Germany), 'Prostate Cancer' (Germany), 'Prostate Cancer' (United States), 'Prostate Cancer' (Canada), and 'Rare Pancreatic Tumors'. An 'Announcements' section highlights a release on 25/Nov/2010 regarding the ICGC Data Coordination Center (DCC) and the release of version 3 of the ICGC data portal. A 'nature' logo is also visible, along with a link to an article in Nature 464, 993-998 (15 April 2010).

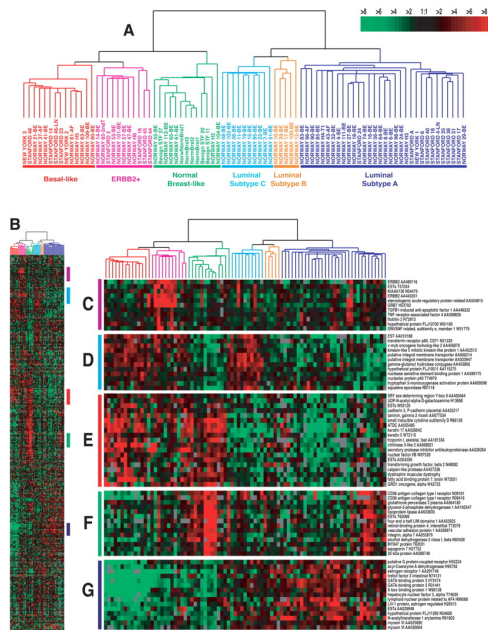




P4 Medicine
 ● PREDICT ● PREVENT ● PERSONALIZE ● PARTICIPATE

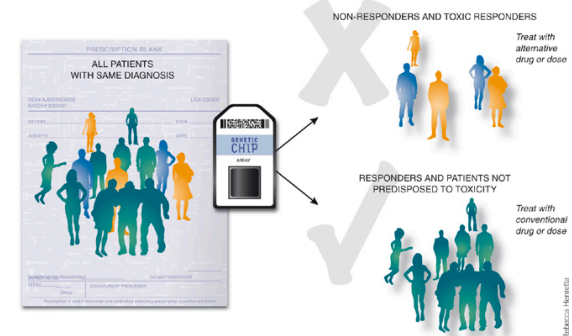
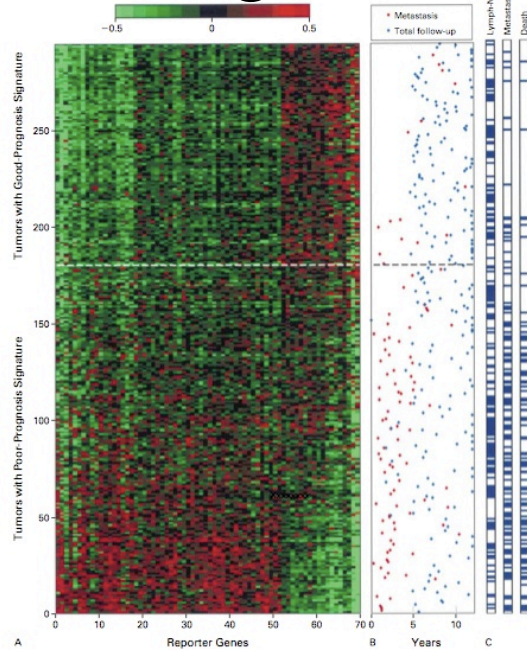


Opportunities



Diagnosis

Prognosis



Response to drugs

Example: Pharmacogenomics / Toxicogenomics

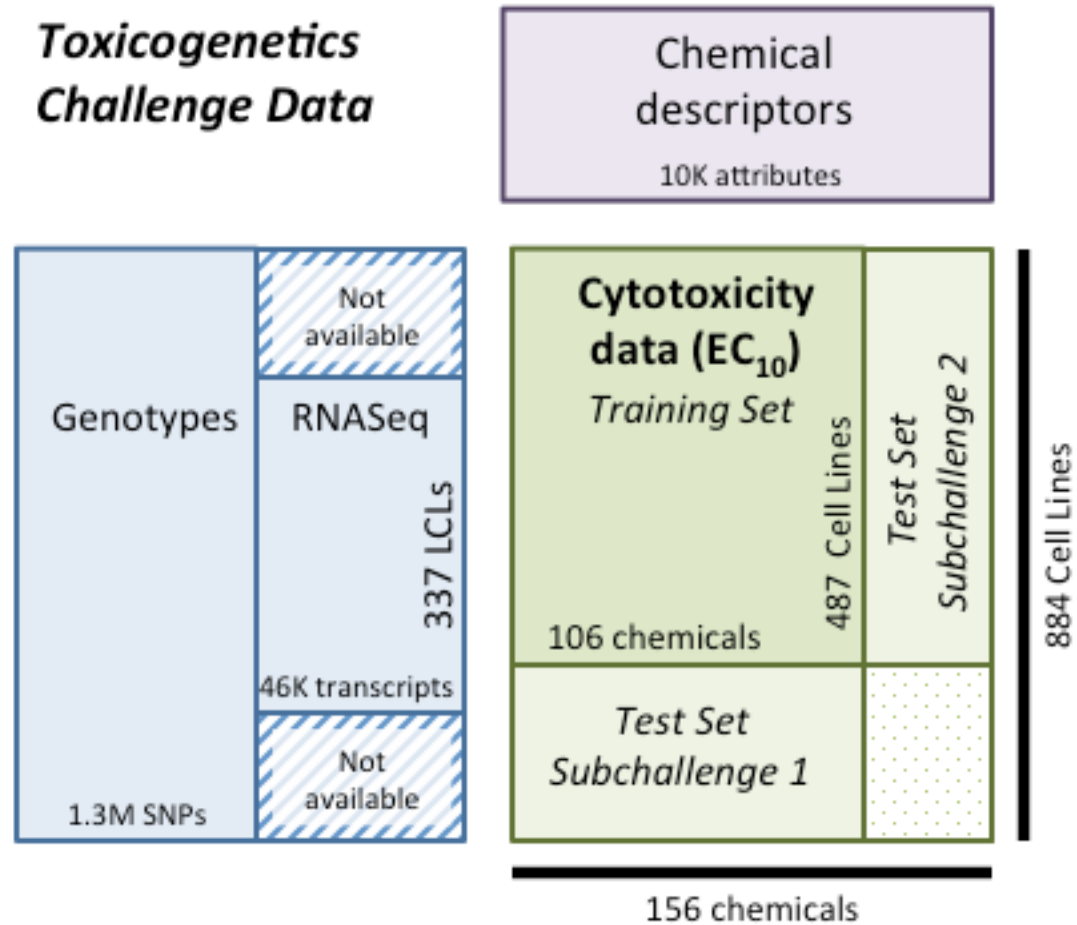


Crowd-sourcing initiatives

The screenshot shows a web browser window with the following elements:

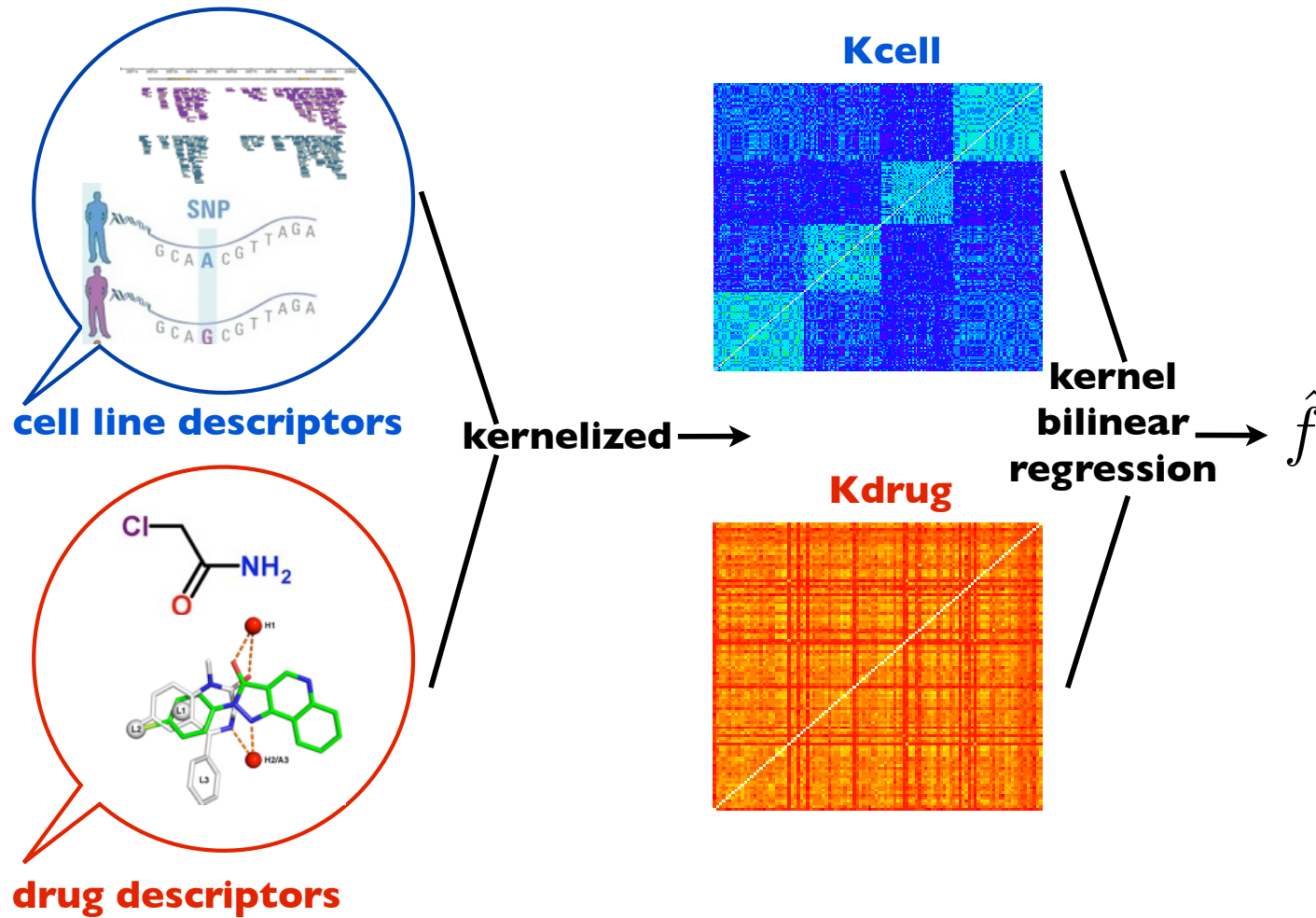
- Browser Tab:** NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge - syn1761567
- Address Bar:** <https://www.synapse.org/#!/Synapse:syn1761567>
- Page Header:** Sage Synapse: Contribute to the Cure | NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge - syn1761567
- Navigation:** Synapse logo, "CONTRIBUTE to the CURE", Search bar, Forum, Register, Login
- Page Title:** NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge ★
- Metadata:** Synapse ID: syn1761567, DOI: (doi:10.7303/syn1761567)
- Navigation Tabs:** Wiki (selected), Files
- Wiki Subpages:**
 - ▲ NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge (Current Page)
 - Data Description
 - Data File Description
 - ▲ Subchallenge 1
 - Subchallenge 1 Final Scoring
 - Subchallenge 1 Leaderboard
 - ▲ Subchallenge 2
 - ▲ Subchallenge 2 Final Scoring
 - Additional metrics
 - Updates to Challenge Information

DREAM8 challenge (jun-sep 2013)

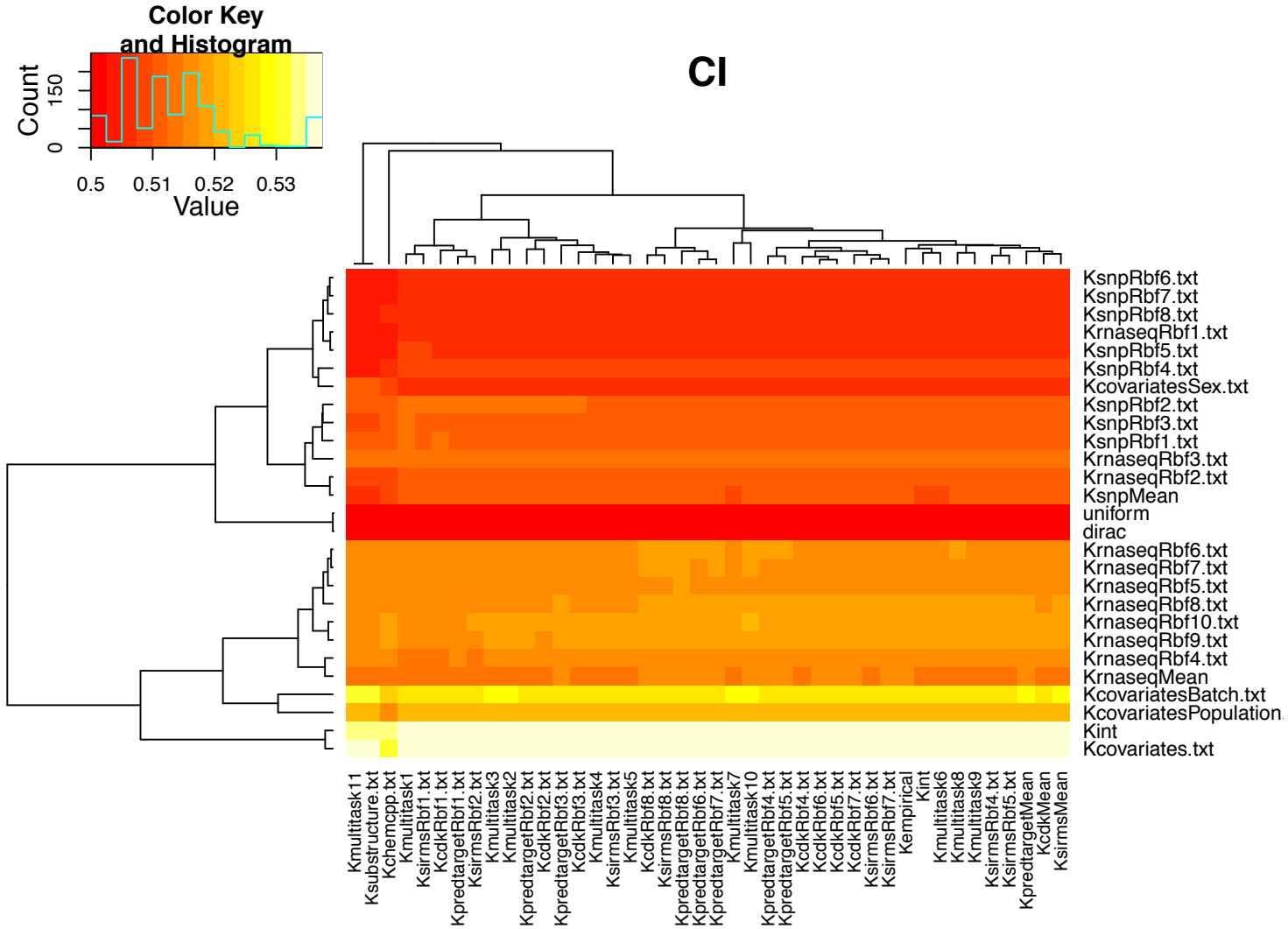


Genotypes from the 1000 genome project; RNASeq from the Geuvadis project

Our approach



Learning occurs...



... and it somehow worked

NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge - syn1761567

https://www.synapse.org/#!Synapse:syn1761567/wiki/60497

Sage Synapse: Contri... NIEHS-NCATS-UNC D... NIEHS-NCATS-UNC... NIEHS-NCATS-UNC... NIEHS-NCATS-UNC...

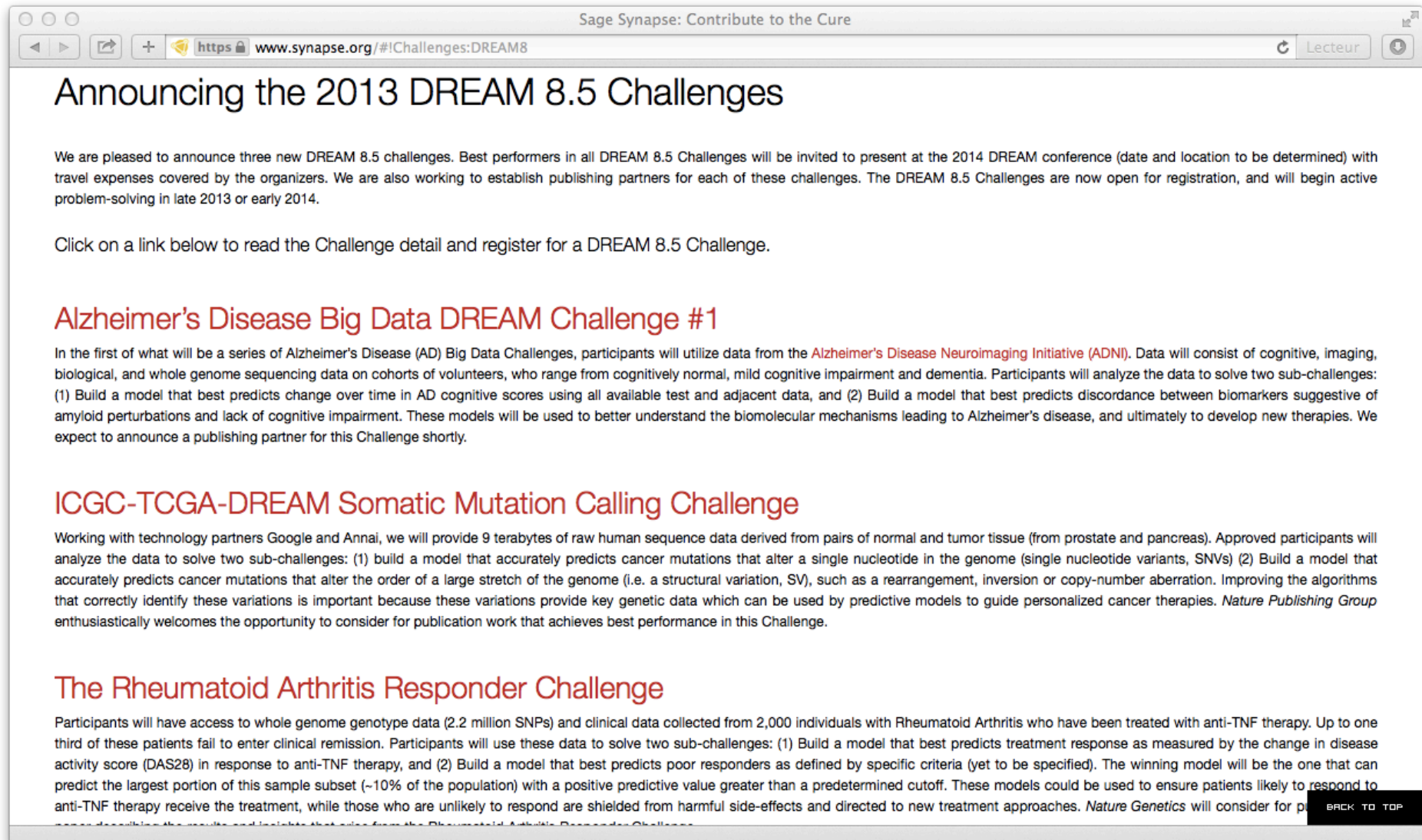
Team	Submission	SynapseID	Mean ranking PCI	Rank PCI	Mean ranking PC	Rank PC	Mean ranking	Rank
Yang_Lab	UTSW_QBRC_kmb310.txt	syn2219079	27.2198	1	31.8681	2	1.5	1.0
CASSIS	Final_prediction_KRR_int_empiric...	syn2224212	31.5714	2	34.3516	4	3.0	2.0
amss2012	Subchal1_randomforest_result.txt	syn2211170	34.8132	3	36.6703	5	4.0	3.0
UT_CCB	Prediction_Result_2.txt	syn2227250	38.8242	11	28.1538	1	6.0	4.0
Yang_Lab	UTSW_QBRC_kmb24.txt	syn2218907	36.0549	4	38.0110	10	7.0	5.5
O6d0A	submission.txt	syn2227400	36.3516	6	37.8901	8	7.0	5.5
Yang_Lab	UTSW_QBRC_lm5.txt	syn2223150	37.9341	8	37.7600	7	7.5	7.0
CQB	submission_dai_gussian2.txt	syn2226588	37.2088	7	37.7600	7	7.5	7.0
Yang_Lab	UTSW_QBRC_kmb49.txt	syn2218923	36.2747	5	37.7600	7	7.5	7.0
UT_CCB	Prediction_Result_3.txt	syn2227281	41.0659	12	37.7600	7	7.5	7.0
D-Tox	ToxSubchallenge_1_prediction_mat...	syn2223065	39.0549	9	37.7600	7	7.5	7.0
Yang_Lab	UTSW_QBRC_lm4.txt	syn2223153	38.8132	10	37.7600	7	7.5	7.0
CASSIS	Final_prediction_KRR_int_dirac_b...	syn2224209	38.2857	13	37.7600	7	7.5	7.0
WarwickDataScience	predictions_subChallenge1_submis...	syn2211154	38.9011	14	37.7600	7	7.5	7.0
Kajju								
CQB								
UTSW_QBRC								

**RECOMB/ISCB Conference on
Regulatory and Systems Genomics,
with DREAM Challenges 2013**

**TORONTO, ONTARIO
NOV 8 - 12, 2013**



More to come!



Sage Synapse: Contribute to the Cure

https www.synapse.org/#!/Challenges:DREAM8

Announcing the 2013 DREAM 8.5 Challenges

We are pleased to announce three new DREAM 8.5 challenges. Best performers in all DREAM 8.5 Challenges will be invited to present at the 2014 DREAM conference (date and location to be determined) with travel expenses covered by the organizers. We are also working to establish publishing partners for each of these challenges. The DREAM 8.5 Challenges are now open for registration, and will begin active problem-solving in late 2013 or early 2014.

Click on a link below to read the Challenge detail and register for a DREAM 8.5 Challenge.

Alzheimer's Disease Big Data DREAM Challenge #1

In the first of what will be a series of Alzheimer's Disease (AD) Big Data Challenges, participants will utilize data from the [Alzheimer's Disease Neuroimaging Initiative \(ADNI\)](#). Data will consist of cognitive, imaging, biological, and whole genome sequencing data on cohorts of volunteers, who range from cognitively normal, mild cognitive impairment and dementia. Participants will analyze the data to solve two sub-challenges: (1) Build a model that best predicts change over time in AD cognitive scores using all available test and adjacent data, and (2) Build a model that best predicts discordance between biomarkers suggestive of amyloid perturbations and lack of cognitive impairment. These models will be used to better understand the biomolecular mechanisms leading to Alzheimer's disease, and ultimately to develop new therapies. We expect to announce a publishing partner for this Challenge shortly.

ICGC-TCGA-DREAM Somatic Mutation Calling Challenge

Working with technology partners Google and Annai, we will provide 9 terabytes of raw human sequence data derived from pairs of normal and tumor tissue (from prostate and pancreas). Approved participants will analyze the data to solve two sub-challenges: (1) build a model that accurately predicts cancer mutations that alter a single nucleotide in the genome (single nucleotide variants, SNVs) (2) Build a model that accurately predicts cancer mutations that alter the order of a large stretch of the genome (i.e. a structural variation, SV), such as a rearrangement, inversion or copy-number aberration. Improving the algorithms that correctly identify these variations is important because these variations provide key genetic data which can be used by predictive models to guide personalized cancer therapies. *Nature Publishing Group* enthusiastically welcomes the opportunity to consider for publication work that achieves best performance in this Challenge.

The Rheumatoid Arthritis Responder Challenge

Participants will have access to whole genome genotype data (2.2 million SNPs) and clinical data collected from 2,000 individuals with Rheumatoid Arthritis who have been treated with anti-TNF therapy. Up to one third of these patients fail to enter clinical remission. Participants will use these data to solve two sub-challenges: (1) Build a model that best predicts treatment response as measured by the change in disease activity score (DAS28) in response to anti-TNF therapy, and (2) Build a model that best predicts poor responders as defined by specific criteria (yet to be specified). The winning model will be the one that can predict the largest portion of this sample subset (~10% of the population) with a positive predictive value greater than a predetermined cutoff. These models could be used to ensure patients likely to respond to anti-TNF therapy receive the treatment, while those who are unlikely to respond are shielded from harmful side-effects and directed to new treatment approaches. *Nature Genetics* will consider for publication describing the results and insights that arise from the Rheumatoid Arthritis Responder Challenge.

BACK TO TOP

Thanks!



Rob Rogers / Pittsburgh Post-Gazette