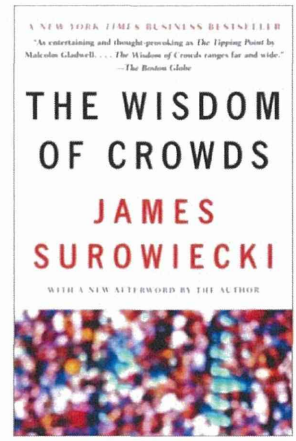
# Harnessing Diversity

- Observation
  - Different algorithms have their strengths and weaknesses, assumptions and work well for specific data sets

- Can we harness this diversity to develop an ensemble learning systems for prediction of LC50 and potential side-effect of a compound as well as drug-target prioritization?

**Wisdom of the Crowds in machine learning algorithms**

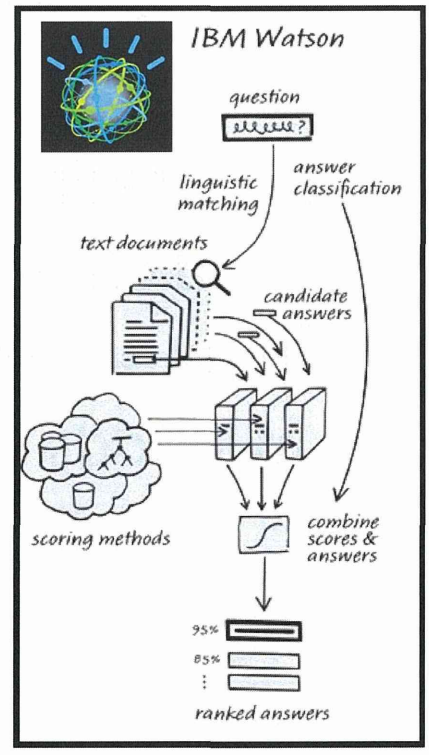



# Harnessing Diversity

**Jeopardy!** is an American television quiz show featuring trivia in history, literature, and other topics.

The show has a unique answer-and-question format in which contestants are presented with clues in the form of answers, and must phrase their responses in question form

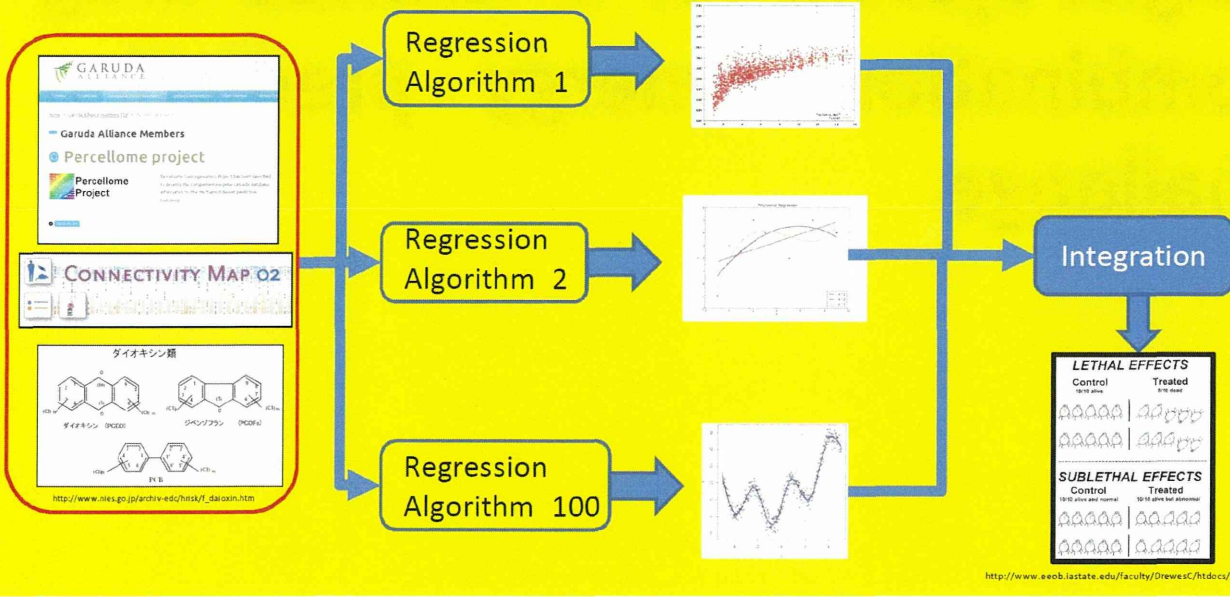
IBM's Watson supercomputer beats humans in Jeopardy!

# Xsight system

## Xsight system

**Based on high-dimensional multi-omics data, Xsight integrate prediction results from multiple algorithms to make accurate prediction, e.g, toxicity of chemical compounds.**



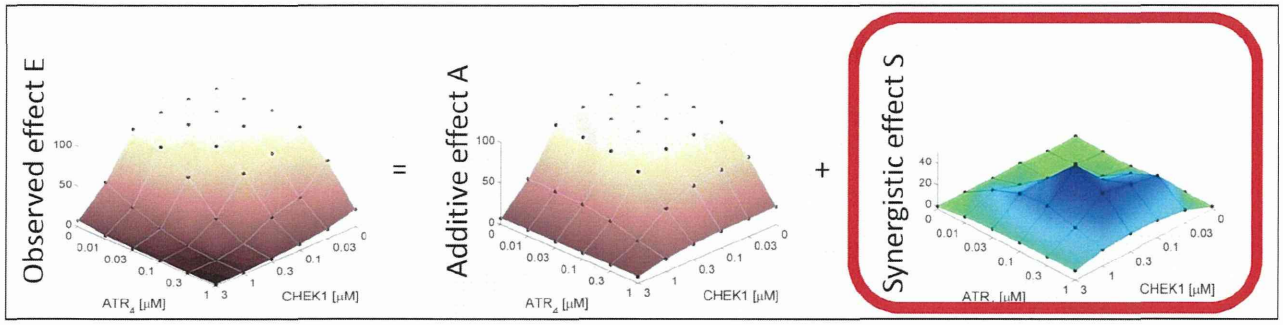
**Current status**

## Xsight system integrates more than 100 machine learning algorithms to make accurate prediction

| | | |
|---|---|---|
| Adaptive-Network-Based Fuzzy Inference System | glmnet | Radial Basis Function Network |
| Bagged CART | Hybrid Neural Fuzzy Inference System | Radial Basis Function Network |
| Bagged Logic Regression | Independent Component Regression | Random Forest |
| Bagged MARS | k-Nearest Neighbors | Random Forest |
| Bagged MARS using gCV Pruning | k-Nearest Neighbors | Random Forest |
| Bagged Model | Least Angle Regression | Random Forest by Randomization |
| Bayesian Additive Regression Trees | Least Angle Regression | Random Forest Rule-Based Model |
| Bayesian Generalized Linear Model | Linear Regression | Random Forest with Additional Feature Selection |
| Bayesian Regularized Neural Networks | Linear Regression with Backwards Selection | Random k-Nearest Neighbors |
| Boosted Generalized Additive Model | Linear Regression with Forward Selection | Random k-Nearest Neighbors with Feature Selection |
| Boosted Generalized Linear Model | Linear Regression with Stepwise Selection | Regularized Random Forest |
| Boosted Linear Model | Linear Regression with Stepwise Selection | Regularized Random Forest |
| Boosted Smoothing Spline | Logic Regression | Relaxed Lasso |
| Boosted Tree | Model Averaged Neural Network | Relevance Vector Machines with Linear Kernel |
| Boosted Tree | Model Rules | Relevance Vector Machines with Polynomial Kernel |
| CART | Model Tree | Relevance Vector Machines with Radial Basis Function Kerne |
| CART | Multi-Layer Perceptron | Ridge Regression |
| Conditional Inference Random Forest | Multi-Layer Perceptron | Ridge Regression with Variable Selection |
| Conditional Inference Tree | Multivariate Adaptive Regression Spline | Robust Linear Model |
| Conditional Inference Tree | Multivariate Adaptive Regression Splines | Rotation Forest |
| Cubist | Neural Network | Self-Organizing Map |
| Dynamic Evolving Neural-Fuzzy Inference System | Neural Network | Self-Organizing Maps |
| Elasticnet | Neural Networks with Feature Extraction | Simplified TSK Fuzzy Rules |
| Ensemble Partial Least Squares Regression | Non-Convex Penalized Quantile Regression | Sparse Partial Least Squares |
| Ensemble Partial Least Squares Regression with Feature Selection | Non-Negative Least Squares | Stacked AutoEncoder Deep Neural Network |
| eXtreme Gradient Boosting | Parallel Random Forest | Stochastic Gradient Boosting |
| eXtreme Gradient Boosting | partDSA | Subtractive Clustering and Fuzzy c-Means Rules |
| Extreme Learning Machine | Partial Least Squares | Supervised Principal Component Analysis |
| Fuzzy Inference Rules by Descent Method | Partial Least Squares | Support Vector Machines with Boundrange String Kernel |
| Fuzzy Rules via MOGUL | Partial Least Squares | Support Vector Machines with Exponential String Kernel |
| Fuzzy Rules via Thrift | Partial Least Squares | Support Vector Machines with Linear Kernel |
| Gaussian Process | Partial Least Squares Generalized Linear Models | Support Vector Machines with Linear Kernel |
| Gaussian Process with Polynomial Kernel | Penalized Linear Regression | Support Vector Machines with Polynomial Kernel |
| Gaussian Process with Radial Basis Function Kernel | Polynomial Kernel Regularized Least Squares | Support Vector Machines with Radial Basis Function Kernel |
| Generalized Additive Model using LOESS | Principal Component Analysis | Support Vector Machines with Radial Basis Function Kernel |
| Generalized Additive Model using Splines | Projection Pursuit Regression | Support Vector Machines with Spectrum String Kernel |
| Generalized Additive Model using Splines | Quantile Random Forest | The lasso |
| Generalized Linear Model | Quantile Regression Neural Network | Tree Models from Genetic Algorithms |
| Generalized Linear Model with Stepwise Feature Selection | Quantile Regression with LASSO penalty | Tree-Based Ensembles |
| Genetic Lateral Tuning and Rule Selection of Linguistic Fuzzy Systems | Radial Basis Function Kernel Regularized Least Squares | Wang and Mendel Fuzzy Rules |

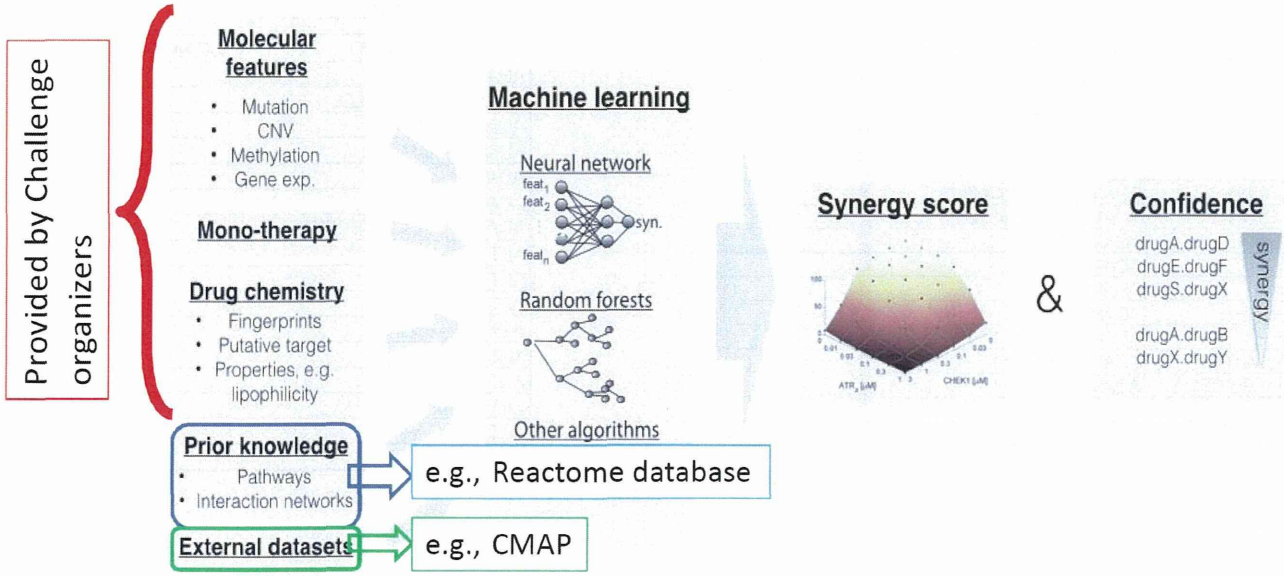
# (Sample analysis) Application of Xsight system for DREAM10 drug combination synergy prediction challenge

# Pharmacology: Combinations analyses




Build machine learning models using pharmacological and molecular data to predict synergistic effects of a compound combination.

# Subchallenge 1A: "Black-box" with all available data




Predict Synergy from all Available Data

# Subchallenge 1A:
# "Black-box" with all available data

**Molecular features**

- Mutation
- CNV
- Methylation
- Gene exp.

**Mono-therapy**

**Drug chemistry**

- Fingerprints
- Putative target
- Properties, e.g. lipophilicity

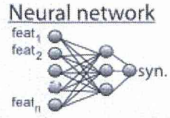
**Prior knowledge**

- Pathways
- Interaction networks

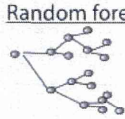
**External datasets**

**Machine learning**

Neural network



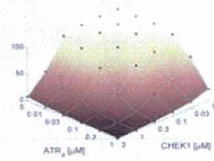

Random forests



Other algorithms

⋮

**Synergy score**



&

**Confidence**

drugA.drugD
drugE.drugF
drugS.drugX

drugA.drugB
drugX.drugY

synergy

Predict Synergy from all Available Data

# Output for subchallenge 1

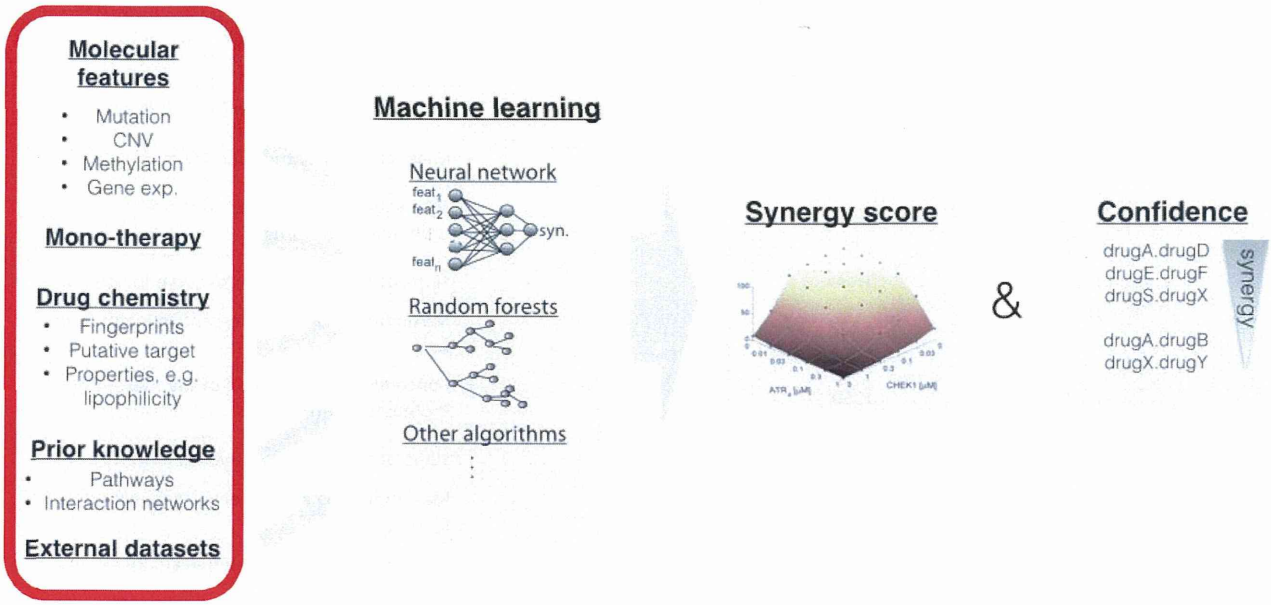
| CELL_LINE | COMBINATION_ID | PREDICTION |
|---|---|---|
| BT-549 | ADAM17.AKT | 0 |
| CAL-148 | ADAM17.AKT | 0 |
| HCC38 | ADAM17.AKT | 0 |
| BT-20 | ADAM17.BCL2_BCL2L1 | 0 |
| HCC1143 | ADAM17.BCL2_BCL2L1 | 0 |
| HCC1937 | ADAM17.BCL2_BCL2L1 | 0 |
| Hs-578-T | ADAM17.BCL2_BCL2L1 | 0 |
| DU-4475 | ADAM17.FGFR | 0 |
| HCC1187 | ADAM17.FGFR | 0 |
| HCC70 | ADAM17.FGFR | 0 |
| Hs-578-T | ADAM17.FGFR | 0 |
| CAL-51 | ADAM17.MAP2K_1 | 0 |
| HCC1428 | ADAM17.MAP2K_1 | 0 |
| MDA-MB-157 | ADAM17.MAP2K_1 | 0 |
| BT-549 | ADAM17.MTOR_1 | 0 |
| MDA-MB-157 | ADAM17.MTOR_1 | 0 |
| MFM-223 | ADAM17.MTOR_1 | 0 |

Combination priority

| COMBINATION_ID | CONFIDENCE |
|---|---|
| ADAM17.AKT | 0 |
| ADAM17.BCL2_BCL2L1 | 0 |
| ADAM17.FGFR | 0 |
| ADAM17.MAP2K_1 | 0 |
| ADAM17.MTOR_1 | 0 |
| ADAM17.PIK3C | 0 |
| ADAM17.PIK3CB_PIK3CD | 0 |
| AKT.AKT_1 | 0 |
| AKT.ATR_4 | 0 |
| AKT.CSNK2A1_2 | 0 |
| AKT.Chloroquine | 0 |
| AKT.ERBB | 0 |
| AKT.FGFR | 0 |
| AKT.HDAC_4 | 0 |
| AKT.PIK3CA_4 | 0 |
| AKT.PIK3C_2 | 0 |

**Synergy score for each compound combination under each cell lines.**

# Subchallenge 1A: "Black-box" with all available data




Predict Synergy from all Available Data

# Description of Challenge Data Overview

- 85 cancer cell lines across 6 tissue types (gastrointestinal, lung, breast, melanoma, urinary tract, and lymphatic)
- **AstraZeneca pharmacology data:**
  - Combination synergy: 912 unique combinations screened in 11,759 experiments
  - Matched monotherapy response: 119 unique compounds
  - Compound information (structure and chemical properties)
- **Sanger cell line molecular data:**
  - From naïve cell lines (no pre-treatment)
  - Mutation, copy number, gene expression and DNA methylation
- *Raw data available to extract additional features*

# Pharmacological data 1

| Column number | Column name | Description |
|---|---|---|
| 1 | CELL_LINE | Normalised cell line name. |
| 2 | COMPOUND_A | Name of drug/compound **A described by its target.** |
| 3 | COMPOUND_B | Name of drug/compound **B described by its target.** |
| 4 | MAX_CONC_A | Maximum concentration (uM) of drug A. |
| 5 | MAX_CONC_B | Maximum concentration (uM) of drug B. |
| 6 | IC50_A | Concentration where half of the maximum kill is obtained with drug A. |
| 7 | H_A | Slope of the dose-response curve for drug A. |
| 8 | Einf_A | Maximum cells killed (percentage) with drug A. |
| 9 | IC50_B | Concentration where half of the maximum kill is obtained with drug B. |
| 10 | H_B | Slope of the dose-response curve for drug B. |
| 11 | Einf_B | Maximum cells killed (percentage) with drug B. |
| 12 | SYNERGY_SCORE | Estimated total synergy between drug A and drug B in combination. |
| 13 | QA | Quality assessment flag of combination assays |
| 14 | COMBINATION_ID | Name for drug A and drug B combination |

# Pharmacological data 2

| ChallengeName | Target(Official Symbol) | HBA | cLogP | HBD | Lipinski | SMILES or PubChem ID | MW |
|---|---|---|---|---|---|---|---|
| ATM | ATM | 8 | 4.28 | 1 | 1 | C[C@@H]1CN(C[C@@H](O1)C | 547.7 |
| TOP2 | TOP2*,PRKDC | 4 | 5.44 | 0 | 1 | c1ccc2c(c1)c3cccc(c3s2)c4c | 413.5 |
| PRKDC_2 | PRKDC | 8 | 4.39 | 1 | 1 | CCN1CCN(CC1)CC(=O)Nc2c | 582.7 |
| FGFR | FGFR1, FGFR2, FGFR3 | 8 | 4.43 | 3 | 0 | C[C@@H]1CN(C[C@@H](N1)C | 463.6 |
| MET_ALK | cMET, ALK | 6 | 4.29 | 3 | 0 | C[C@H](c1c(ccc(c1Cl)F)Cl)O | 450.3 |
| AKT | AKT1, AKT2, AKT3 | 8 | 1.18 | 5 | 0 | c1cc(ccc1[C@H](CCO)NC(=C | 428.9 |
| AKT_1 | AKT* | 6 | 3.24 | 3 | 0 | c1ccc(cc1)c2cc3c(ccn4c3n[n | 407.5 |
| ALK | ALK | 8 | 3.89 | 2 | 1 | CN1CCN(CC1)C2CCN(CC2)c | 532.1 |
| ALK_IGFR | ALK, IGF*R | 7 | 4.08 | 4 | 0 | CCc1cc(ccc1Nc2ncc(c(n2)c3 | 449 |
| AR | AR | 6 | 3.35 | 1 | 0 | CNC(=O)c1ccc(cc1F)N2C(=S | 464.4 |
| ATR_1 | ATR | 7 | 0.44 | 1 | 0 | CS(=O)(=O)C1(CC1)c2cc(nc( | 398.5 |
| ATR_2 | ATR | 7 | 0.96 | 1 | 0 | C[C@@H]1COCCN1c2cc(nc( | 412.5 |
| ATR_4 | ATR | 8 | 0.59 | 2 | 0 | C[C@@H]1COCCN1c2cc(nc( | 412.5 |
| BCL2 | BCL2 | 11 | 12.39 | 2 | 3 | CC1(CCC(=C(C1)CN2CCN(CC | 974.6 |
| BCL2_BCL2L1 | BCL2, BCL2L1 | 11 | 9.62 | 4 | 3 | CN(CC[C@H](CSc1ccccc1)N | 945.5 |
| BCL2L1 | BCL2L1 | 9 | 3.05 | 2 | 1 | CN(C)CC#Cc1ccc(OCCCc2s | 651.8 |
| Doxorubicin | TOP*,DNMT1 | 12 | 0.32 | 7 | 3 | CC1OC(CC(O1)OC2CC(Cc3 | 543.5 |
| CHEK1 | CHEK1 | 6 | 3.11 | 5 | 0 | c1cc(cc(c1)F)c2cc(cs2)C(=O | 362.4 |
| CSNK2A1 | CSNK2A1 | 10 | 3.02 | 3 | 0 | CCC(=O)Nc1cc(ccc1N(C)CCN( | 447.5 |

MW - Molecular weight
HBA - H-bond acceptors
HBD - H-bond donors
cLogP - Calculated log P
Lipinski - Lipinski's rule of 5

# Molecular data: Mutation 1

| Column number | Column name | Description |
|---|---|---|
| 1 | Gene.name | The gene name for which the data has been curated in COSMIC. In most cases this is the accepted HGNC identifier. |
| 2 | Accession.Number | The transcript identifier of the gene. |
| 3 | Gene.CDS.length | Length of the gene (base pair) counts. |
| 4 | HGNC.ID | If gene is in HGNC, this id helps linking it to HGNC. |
| 5 | cell_line_name | Normalised cell line name. **Note**: in COSMIC export this column was "Sample name", but in this file it is edited to "cell_line_name". Furthermore, cell line names are matched to all other normalised cell line names to be an unique and common identifier. This column should be used to match the mutational data to other datasets within the DREAM challenge. |
| 6 | ID_sample | COSMIC cell line ID (**Note** COSMIC ID is not necessarily unique for a cell line, due to multiple samples from the same cell line being entered in COSMIC). |
| 7 | ID_tumour | COSMIC primary tumour ID (**Note** COSMIC ID is not necessarily unique for a primary tumour, due to multiple samples from the same tumour being entered in COSMIC). |
| 8 | Primary.site | The primary tissue/cancer from which the sample originated. More details on the tissue classification can be found in COSMIC, which have standard set of classification system for tissue types and sub types because they vary a lot between different papers. |
| 9 | Site.subtype | Further sub classifications of the samples tissue of origin. |
| 10 | Primary.histology | The histological classification of the sample. |
| 11 | Histology.subtype | Further histological classifications of the sample. |
| 12 | Genome.wide.screen | If the entire genome/exome is sequenced. |
| 13 | Mutation.ID | Unique mutation identifier. |
| 14 | Mutation.CDS | The change that has occurred in the nucleotide sequence. Formatting is identical to the method used for the peptide sequence. |
| 15 | Mutation.AA | The change that has occurred in the peptide sequence. Formatting is based on the recommendations made by the Human Genome Variation Society. The description of each type can be found by following the link to Mutation Overview page. |
| 16 | Mutation.Description | Type of mutation (substitution, deletion, insertion, complex, fusion, unknown etc.) |
| 17 | Mutation.zygosity | Information on whether the mutation was reported to be homozygous , heterozygous or unknown within the sample. |

# Molecular data: Mutation 2

| Column number | Column name | Description |
|---|---|---|
| 18 | GRCh | The coordinate system used (37 = GRCh37/Hg19, 38 = GRCh38/Hg38) |
| 19 | Mutation.genome.position | The genomic coordinates of the mutation. |
| 20 | strand | Positive or negative. |
| 21 | SNP | All the known SNPs are flagged as 'y' defined by the 1000 genomes project, dbSNP and a panel of 378 normal (non-cancer) samples from Sanger CGP sequencing. |
| 22 | FATHMM.prediction | FATHMM (Functional Analysis through Hidden Markov Models) descriptors, which might be "Pathogenic", if they are defined as Cancer or Damaging. Another prediction might be "Neutral", if they are defined as Passenger or Tolerated. |
| 23 | Mutation.somatic.status | Information on whether the sample was reported to be (i) Confirmed Somatic, (ii) Previously Reported or (iii) Variant of unknown origin: 'Confirmed Somatic' - if the mutation has been confirmed to be somatic in the experiment by sequencing both the tumour and a matched normal from the same patient. 'Previously observed' - when the mutation has been reported as somatic previously but not in current paper. 'Variant of unknown origin' - when the mutation is known to be somatic but the tumour was sequenced without a matched normal. |
| 24 | Pubmed_PMID | The PUBMED ID for the paper that the sample was noted in, linking to pubmed to provide more details of the publication. |
| 25 | ID_STUDY | Lists the unique Ids of studies that have involved this sample. |
| 26 | Institute | Source of the sample. |
| 27 | Institute.Address | Address of the source of the sample. |
| 28 | Catalogue.Number | Identifier used in the source of the sample. |
| 29 | Sample.source | If it describes where the sample has originated from. |
| 30 | Tumour.origin | If it describes where the tumour has originated from. |
| 31 | Age | Age of the individual (if this information is provided with the publications). |
| 32 | Comments | If any additional information available about a sample |

# Molecular data: CNV

| Column number | Column name | Description |
|---|---|---|
| 1 | gene | The gene name for which the data has been curated in COSMIC. In most cases this is the accepted HGNC identifier. |
| 2 | chr | Chromosome. |
| 3 | gene_start | Gene start position. |
| 4 | gene_end | Gene end position |
| 5 | cell_line_name | Normalised cell line name. |
| 6 | max_cn | Maximum copy number of any genomic segment containing coding sequence of the gene. **Note**: Hyphen (-) is used where the value is unknown and -1 indicates that the copy number could not be determined. |
| 7 | min_cn | Minimum copy number of any genomic segment containing coding sequence of the gene. **Note**: same nomenclature as "max_cn". |
| 8 | zygosity | Zygosity - (H) if all segments containing gene sequence are heterozygous, (L) if any segment containing coding sequence has LOH, (0) if the complete coding sequence of the gene falls within a homozygous deletion. |
| 9 | disruption_status | Disruption (D) if the gene spans more than 1 genomic segment (-) if no disruption occurs |

# Molecular data: Methylation

**Methylation intensity for each probe under each cell line**

**Cell line**

| | C32 | HT-29 | HCT-116 | NCI-H23 | MDA-MB-231 | COLO-205 | MCF7 | T47D | BT-549 | NCI-H226 |
|---|---|---|---|---|---|---|---|---|---|---|
| chr1 91190489-91192804 | 2.353700556 | | | | | | | | 0.943967667 | 0.493091167 |
| chr1 230561103-230562702 | 0.210042818 | 0.938507273 | 0.923477273 | -1.060885455 | -0.611479364 | 1.229673636 | -0.130172727 | -0.952858545 | | -0.628429273 |
| chr1 5937157-5937392 | 2.467364615 | 2.770209231 | 2.028302308 | 2.652915385 | 2.854169231 | 2.786661538 | 2.578192308 | 3.173538462 | | 2.380995385 |
| chr1 166958220-166958683 | 1.502655 | 1.907048571 | 2.173114286 | 2.739207143 | 2.466914286 | 2.325486429 | 2.06725 | 1.115511071 | 2.364 | 2.822421429 |
| chr1 43832814-43833073 | -1.476846 | -1.9552625 | -2.2335125 | -1.7339625 | -1.6913625 | -1.8069625 | -1.703719625 | -1.3297375 | -1.375 | -1.4201375 |
| chr1 200011400-200012055 | 0.448786 | -1.680083556 | 2.418088889 | -2.08863 | 1.903168889 | -1.056498889 | 2.579044444 | -1.284018889 | -0.0421 | 1.723914444 |
| chr1 20960045-20960551 | -1.397996 | -0.943056842 | -1.874598421 | -1.093650789 | -1.307366316 | -1.364708421 | -1.435443158 | -1.346343684 | -1.5283 | -1.164586316 |
| chr1 1267435-1268986 | 2.051368 | 0.604050444 | 1.060666667 | 2.122782222 | 2.629511111 | -0.118029444 | 2.064967778 | 2.840511111 | 0.8298 | 2.484033333 |
| chr1 169396621-169396869 | -0.814861 | 2.699207692 | 2.121369231 | -1.158247692 | 2.311681846 | 1.780205385 | 2.328138462 | 1.306907692 | | 1.047204615 |
| chr1 223566642-223567268 | 0.588753 | 0.753726721 | 2.081857143 | -0.095171429 | -0.470254286 | 0.9567085 | 0.274985 | 0.190506429 | 0.7 | 0.704929151 |
| chr1 57887963-57890637 | 0.090028 | 2.033256111 | 1.353109444 | -1.886496111 | 0.645669222 | 1.766996111 | -0.061538944 | -0.834141111 | -0.8632 | -0.761099556 |
| chr1 151319326-151319545 | -1.267198 | -1.182181538 | -1.813203846 | -1.092549231 | -1.219973077 | -1.116993846 | -1.336252308 | -0.8296 | -0.9134 | -0.800036154 |
| chr1 10757891-10758252 | 3.13 | 2.63274 | 2.55384 | 3.56128 | 3.47986 | 2.1520602 | 2.11648 | 2.51802 | 3. | 3.25662 |
| chr1 12600347-12600556 | 2.541433 | 0.189083333 | -0.858768333 | 0.831673333 | 1.467621667 | 2.364333333 | 1.910505 | 2.088771667 | 0. | -1.122766667 |
| chr1 45082739-45083285 | 0.171790 | 0.966781818 | 1.953094545 | 2.3127 | 2.502156364 | 1.016355455 | 2.535570727 | -0.656535818 | 0.1955 | 1.050563636 |
| chr1 142618751-142619238 | -0.092337 | 0.364719 | 0.500395111 | 0.485149222 | -0.028964444 | 0.671795556 | -0.630747778 | -0.742285556 | 0.3323 | 0.113546667 |
| chr1 17006742-17007151 | 1.096359 | 0.116424286 | -0.231994286 | -1.396394286 | 0.665644286 | 0.500472857 | -0.549038571 | 0.927208571 | 1.3818 | 0.375101429 |
| chr1 34628783-34630976 | -1.044090 | 1.556026897 | 1.192916448 | -0.949612759 | 0.254261034 | 2.345307897 | 1.234359379 | -0.875541414 | -0.6565 | -1.749711655 |
| chr1 154474107-154475699 | 1.719509 | 1.98251025 | 2.113368125 | 0.768190313 | 0.982089375 | 1.905388125 | 1.92920375 | 1.726683125 | 1.4214 | 1.3723325 |
| chr1 160340604-160340843 | 1.315735 | 0.52577 | 1.072319444 | 1.27819 | 0.659457778 | 0.520746667 | 0.79526 | 0.55985 | 0.6091 | 1.553627778 |
| chr1 33741774-33742381 | 2.59 | 1.412036 | 2.32388 | 2.6591 | 2.86906 | 2.27134 | 2.63596 | 2.94794 | 2. | 2.70702 |
| chr1 43996724-43997592 | 0.550540 | 0.03259475 | 0.1251525 | 0.43964 | 0.435191667 | -0.416053333 | -0.106393333 | -0.405463 | 0.2175 | -0.819504917 |
| chr1 1839958-1840601 | -0.844 | -0.490091 | -1.237507 | -1.0305198 | -0.711569 | -0.676184 | -0.747708 | -0.8727686 | -1.0 | 0.347312 |
| chr1 205179963-205180910 | -1.22114 | -1.231431875 | -1.01245875 | -1.460656875 | -1.1657 | -1.089948375 | -1.336910625 | -1.195807 | -1.05 | -1.3696625 |
| chr1 153756957-153756335 | -1.59518 | -1.8756125 | -1.8818425 | -1.42915 | -1.69943625 | -1.8472375 | -1.7700225 | -1.17596375 | -1.61 | -1.749875 |
| chr1 179711947-179713951 | 1.593180 | 1.218037435 | 1.831404348 | -0.392768696 | 1.576141609 | 1.867999957 | 2.288446957 | 0.840936343 | 0.4855 | 1.119497217 |
| chr1 156814981-156815792 | 1.672436 | 0.642256667 | 1.716155556 | 1.161627667 | 0.720996444 | 0.369422689 | 1.440308889 | 0.650157778 | 0.7347 | 1.776255556 |
| chr1 207082693-207083202 | -1.356329 | -0.885456364 | -1.113085455 | -0.460095455 | | -0.751306455 | -1.070104545 | -0.644906364 | -0.24 | -0.847433636 |
| chr1 53925191-53926228 | 2.357494444 | 2.537765556 | 2.191022222 | 2.929115556 | | 2.708016667 | 2.426179444 | 2.473541278 | 2. | 2.787611111 |

**Methylation score; the ratio of methylated probe intensity and overall intensity (Du et al., 2010)**

# Molecular data: Methylation

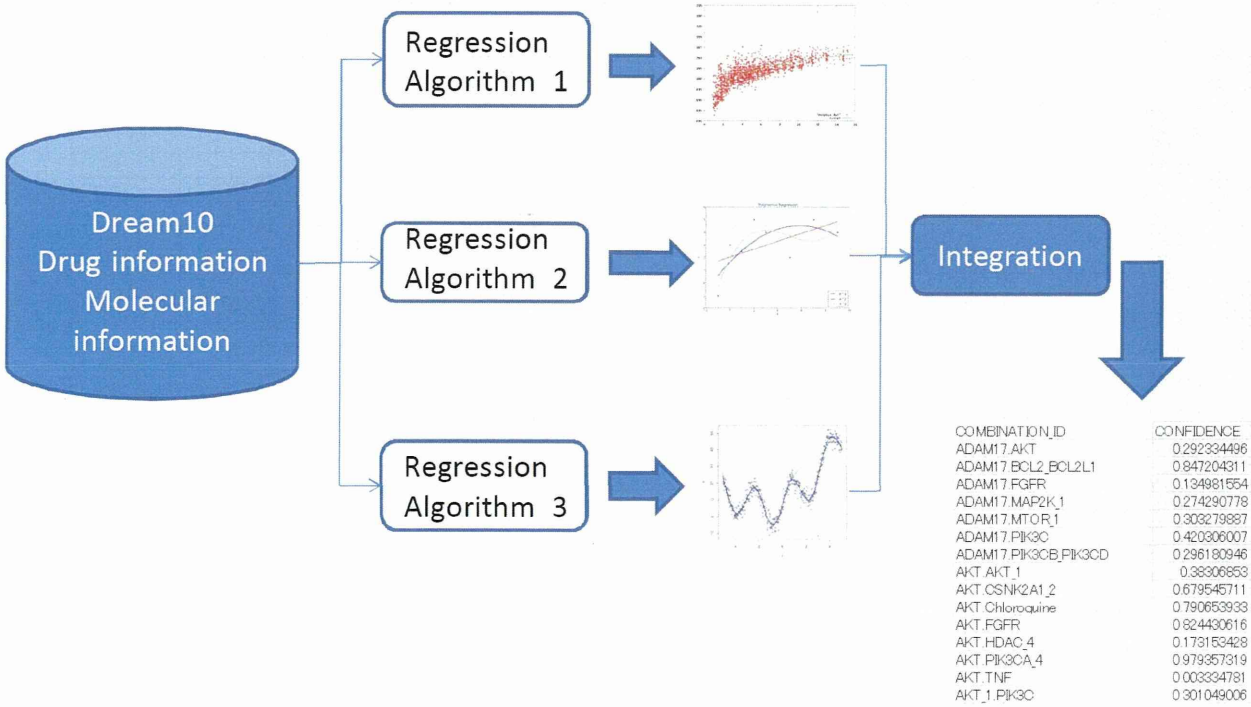
## Cell line

| | C32 | HT-29 | HCT-116 | NCI-H23 | MDA-MB-231 | COLO-205 | MCF7 | T47D | BT-549 | NCI-H226 |
|---|---|---|---|---|---|---|---|---|---|---|
| chr1:91190489-91192804 | 2.353700556 | 1.790653611 | 1.931536389 | 2.619265278 | 1.814633611 | 1.759176389 | 2.1296475 | 1.002812917 | 0.943967667 | 0.493091167 |
| chr1:230561103-230562702 | 0.210042818 | 0.938507273 | 0.923477273 | -1.060885455 | -0.611479364 | 1.229673636 | -0.130172727 | -0.952585545 | -0.092965909 | -0.628429273 |
| chr1:5937157-5937392 | 2.467364615 | 2.770209231 | 2.028302308 | 2.652915385 | 2.854169231 | 2.786661538 | 2.578192308 | 3.173538462 | 2.9032 | 2.380995385 |
| chr1:166958220-166958683 | 1.502515 | 1.907048571 | 2.173114286 | 2.739207143 | 2.466914286 | 2.325486429 | 2.06725 | 1.115511071 | 2.364692857 | 2.822421429 |
| chr1:43832814-43833073 | | | | | | | | | | |
| chr1:200011400-200012055 | 0 | | | | | | | | | |
| chr1:20960045-20960551 | -1 | | | | | | | | | |
| chr1:1267435-1268986 | 2 | | | | | | | | | |
| chr1:169396621-169396869 | -0 | | | | | | | | | |
| chr1:223566642-223567268 | 0 | | | | | | | | | |
| chr1:57887963-57890637 | 0 | | | | | | | | | |
| chr1:151319326-151319545 | -1 | | | | | | | | | |
| chr1:10757891-10758252 | | | | | | | | | | |
| chr1:12600347-12600556 | 2 | | | | | | | | | |
| chr1:45082739-45083285 | 0.171790182 | 0.966781818 | 1.953094545 | 2.3127 | 2.502156364 | 1.016355455 | 2.535570727 | -0.656535818 | 0.195546273 | 1.050563636 |
| chr1:142618751-142619238 | -0.092337778 | 0.364719 | 0.800396111 | 0.489149222 | -0.028964444 | 0.671795556 | -0.030747778 | -0.742255556 | 0.332308889 | 0.113646667 |
| chr1:17006742-17007151 | 1.096359857 | 0.116424286 | -0.231994286 | -1.396394286 | 0.665644286 | 0.500472857 | -0.549038571 | 0.927208571 | 1.381888571 | 0.375101429 |
| chr1:34628783-34630976 | -1.044090103 | 1.556026897 | 1.192916448 | -0.949612759 | 0.254261034 | 2.345307897 | 1.234359379 | -0.875541414 | -0.656522414 | -1.749711655 |
| chr1:154474107-154475699 | 1.719509375 | 1.98251025 | 2.113368125 | 0.768190313 | 0.982089375 | 1.905388125 | 1.92920375 | 1.726683125 | 1.421436813 | 1.372332 |
| chr1:160340604-160340843 | 1.315735444 | 0.52577 | 1.072319444 | 1.27819 | 0.659457778 | 0.520746667 | 0.79526 | 0.55985 | 0.609111111 | 1.556362778 |
| chr1:33741774-33742381 | 2.59362 | 1.412036 | | 2.6591 | 2.86906 | 2.27134 | 2.63596 | 2.94794 | 2.82272 | 2.7070 |
| chr1:43996724-43997592 | 0.550540833 | 0.03251 | | | | | | | | 0.819504917 |
| chr1:1839958-1840601 | -0.844367 | -0.49 | | | | | | | | 0.34731 |
| chr1:20517963-20518091 | -1.22114375 | -1.23143 | | | | | | | | -1.369662 |
| chr1:153755957-153756335 | -1.59518875 | -1.875 | NPHP4 | | NM_015102 | | Body | | | -1.74987 |
| chr1:179711947-179713951 | 1.593180848 | 1.21803 | MAEL | | NM_032858 | | TSS200 | | | 1.119497217 |
| chr1:156814881-156815792 | 1.672436667 | 0.64225 | ELOVL1 | | NM_022821 | | Body | | | 1.776255556 |
| chr1:207082693-207083202 | -1.356329091 | -0.88545 | NR5A2;NR5A2 | | NM_205860;NM_003822 | | Body;Body | | | 0.847433636 |
| | 2.357494444 | 2.53776 | PINK1;PINK1 | | NM_032409;NM_032409 | | 1stExon;5'UTR | | | 2.787611111 |
| | | | TAS1R3 | | | | | | | |
| | | | C1orf114 | | | | | | | |
| | | | C1orf65;C1orf65 | | | | | | | |
| | | | DAB1 | | NM_021080 | | 5'UTR | | | |
| | | | RFX5;RFX5 | | NM_000449;NM_001025603 | | 5'UTR;5'UTR | | | |
| | | | CASZ1;CASZ1 | | NM_001079843;NM_017766 | | 5'UTR;5'UTR | | | |
| | | | | | | | | | | |
| | | | RNF220 | | NM_018150 | | Body | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | SMD2 | | NM_052896 | | Body | | | |

| UCSC_CpG_Islands_Name |
|---|
| chr1:91190489-91192804 |
| chr1:230561103-230562702 |
| chr1:5937157-5937392 |
| chr1:166958220-166958683 |
| chr1:43832814-43833073 |
| chr1:200011400-200012055 |
| chr1:20960045-20960551 |
| chr1:1267435-1268986 |
| chr1:169396621-169396869 |
| chr1:223566642-223567268 |
| chr1:57887963-57890637 |
| chr1:151319326-151319545 |
| chr1:10757891-10758252 |
| chr1:12600347-12600556 |
| chr1:45082739-45083285 |
| chr1:142618751-142619238 |
| chr1:17006742-17007151 |
| chr1:34628783-34630976 |

By using probe-gene table, we can transform probe-level methylation intensity to gene level methylation intensity.

# Develop machine a learning platform that integrates multiple regression algorithms and apply it for the challenge




| COMBINATION_ID | CONFIDENCE |
|---|---|
| ADAM17_AKT | 0.292334496 |
| ADAM17_BCL2_BCL2L1 | 0.847204311 |
| ADAM17_FGFR | 0.134981554 |
| ADAM17_MAP2K_1 | 0.274290778 |
| ADAM17_MTOR_1 | 0.303279887 |
| ADAM17_PIK3C | 0.420306007 |
| ADAM17_PIK3CB_PIK3CD | 0.296180946 |
| AKT_AKT_1 | 0.38306853 |
| AKT_CSNK2A1_2 | 0.679545711 |
| AKT_Chloroquine | 0.790653933 |
| AKT_FGFR | 0.824430616 |
| AKT_HDAC_4 | 0.173153428 |
| AKT_PIK3CA_4 | 0.979357319 |
| AKT_TNF | 0.003334781 |
| AKT_1_PIK3C | 0.301049006 |