www.epa.gov

DATA DRIVEN SELECTION OF BIOLOGICALLY DIVERSE CELL LINES FOR CHEMICAL BIOACTIVITY SCREENING USING CONTENT MAXIMIZATION

Joshua Harrill[1], Clinton Willis[2], Sophie Malcomber[3], Andy White[3], Russell Thomas[1], Nisha Sipes[4], R. Woodrow Setzer[1]

[1]USEPA Center for Computational Toxicology and Exposure (CCTE), RTP, NC, [2]Oak Ridge Associated Universities (ORAU),Oak Ridge, TN, [3]Unilever SEAC, Colworth, UK, [4]NIEHS NTP, RTP, NC

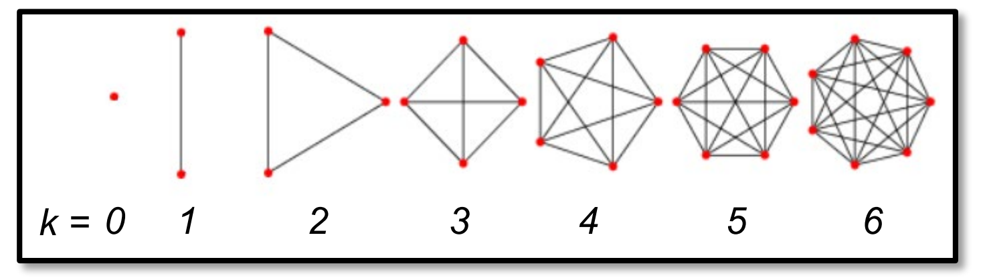
Joshua Harrill I harrill.joshua@epa.gov I https://orcid.org/0000-0003-4317-6391

Background

- The US EPA Computational Toxicology Roadmap advocates the use of high-throughput profiling (HTP) assays (i.e. transcriptomics, phenotypic profiling) as a first tier approach for evaluating the biological activity of environmental chemicals.
- However, no single *in vitro* model will be capable of capturing the entirety of human biological space that may be perturbed upon exposure to environmental chemicals.
- Instead of a single in vitro model, panels of biologically-diverse cell lines may be used to increase the amount of biological space addressed during Tier 1 screening.
- The universe of cell lines for potential use in Tier 1 screening is vast and the number of cell lines that can be tested will be constrained by available time and resources.
- Data driven approaches for cell line selection provide a systematic way to incrementally incorporate as much biological diversity as possible into an vitro testing panel in an efficient manner.
- The proposed approach uses baseline gene expression as a proxy for biological diversity

Content Maximization

- In geometry, a simplex is the generalization of a tetrahedral region of space to ***n*** dimensions.
- The **boundary** of a *k*-simplex has:
  - $k+1$ vertices
  - $\frac{k(k+1)}{2}$ edges
  - $\binom{k+1}{i+1}$ faces, where $\binom{n}{k}$ is a binomial coefficient.



https://www.youtube.com/watch?v=uuOPC55HDQE

- The **content** (i.e. hypervolume) of a simplex can be computed using the **Cayley-Menger determinant.**

If S is a *j*-simplex in real coordinate space (R^n) with vertices $v_1,\dots,v_{j+1}$ and **B** = (β_{ik}) denotes the (j + 1) × (j + 1) matrix given by $\beta_{ik} = |v_i - v_k|_2^2$, then the **content** V_j is given by:

$$V_j^2 = \frac{(-1)^{j+1}}{2^j (j!)^2}\det(\widehat{\mathrm{B}})$$

where $\widehat{\mathrm{B}}$ is the (j + 2) × (j + 2) matrix obtained from **B** by bordering **B** with a top row (0, 1, …, 1) and a left column (0, 1, …, 1)^T.

- For the Cell Line Selection, **Euclidean Distance** is used as **edge lengths** for calculation of **content.**

Implementation

1. Select a “seed” cell line → example: MCF7 cells.
2. Find the cell line “furthest” from the seed.
3. Find a third cell line that maximizes the **area** between the three points.
4. Find a fourth cell line that maximizes the **volume** between four points.
5. Continue until the required number of cell lines are identified using **content.**

U.S. Environmental Protection Agency
Office of Research and Development

This work does not necessarily reflect USEPA policy. Mention of tradenames or products does not represent endorsement for use.

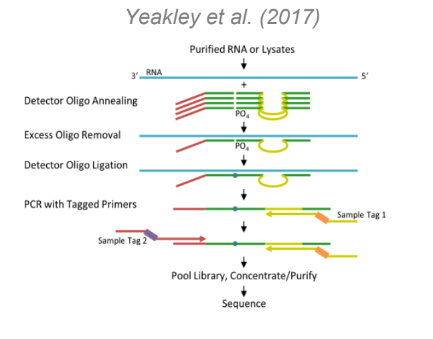
Analysis of CCLE Gene Expression Data

Step 1: Transcriptomic Data Analysis

Process	Description
Source Data	• Cancer Cell Line Encyclopedia (CCLE) Affymetrix Database (Barretina et al. 2012) • National Cancer Institute 60 (NCI60) Affymetrix Database (Shoemaker et al 2006)
Annotation	• Cross reference sample names with ExPasy CallOSaurus (Bairoch et al. 2018) • Harmonize cell name abbreviations • Annotate according to: • Growth mode (adherent, suspension, mixed) • Tissue-of-origin • Evidence of contamination / misidentification • Commercial availability
Filtering	• Exclude lines with evidence of contamination / misidentification • Exclude lines that are not commercially available
Data Normalization	• RMA normalization remaining cell lines • Log2 transform • Mean log2 expression for probe sets targeted the same gene • 20,992 HUGO genes for each cell line
Dimensionality Reduction:	• Singular value decomposition (Alter et al. 200) to 40 eigengenes
Similarity Scoring:	• Calculate Pairwise Euclidean Distance Matrix

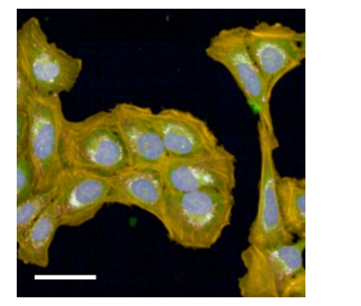
Step 2: Define HTP Screening Assays

TempO-Seq
Yeakley et al. (2017)



- Measures **gene expression**
- Compatible with any cell type

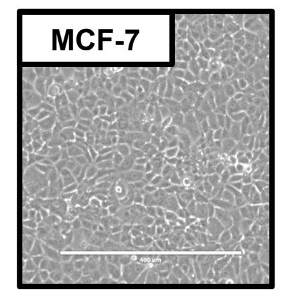
Cell Painting
Bray et al. (2016)



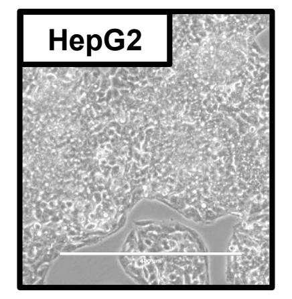
- Measures **cell morphology**
- Requires adherent cells

Step 3: Define Seed Lines

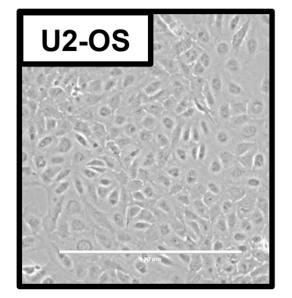
MCF-7



HepG2



U2-OS

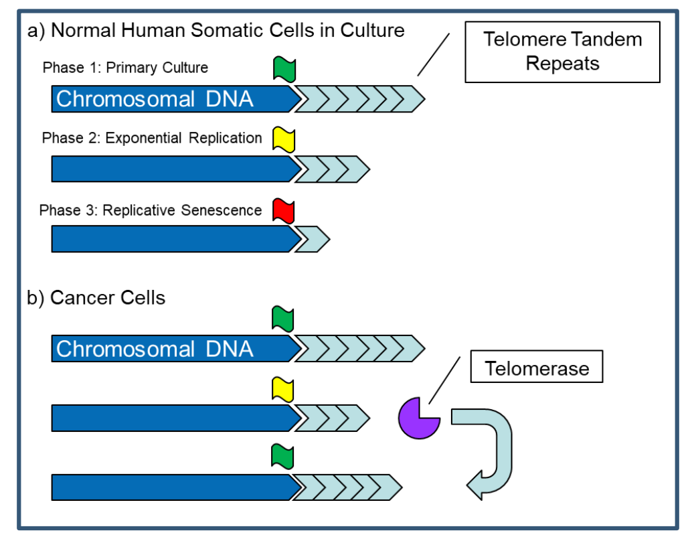


- Three cell lines used in L1000 Connectivity Map or Cell Painting (Lamb et. al 2006)

Step 4: Define Selection Goals

- 10 cell lines for TempO-Seq → any growth mode
- 10 cell lines for Cell Painting → adherent cells only

hTERT Immortalized Cell Lines

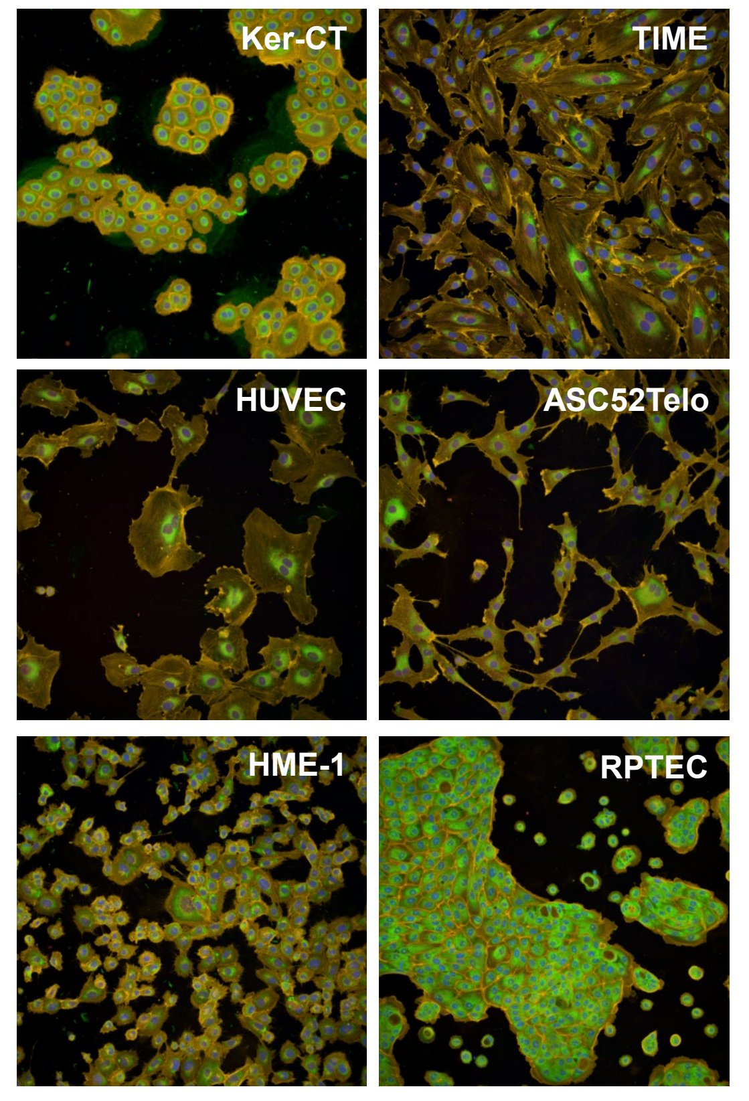

a) Normal Human Somatic Cells in Culture
Telomere Tandem Repeats
Phase 1: Primary Culture
Chromosomal DNA
Phase 2: Exponential Replication
Phase 3: Replicative Senescence
b) Cancer Cells
Chromosomal DNA
Telomerase

- **Telomeres** are structures that cap the end of chromosomes and contain a sequence of 6 bp tandem repeats (i.e.TTAGGG)
  - Utilized to prevent degradation at the ends of chromosomes
  - Shortening of telomeres depictive of aging cells
  - Each cell division results in approx. 50 bp of telomeric material being removed
- **Telomerase** is an enzyme responsible for adding the telomeric repeats and making up for the repeats lost during cell division
  - Consists of two components:
    - RNA component containing the template for telomere DNA synthesis (hTR)
    - Protein catalytic component: **human telomerase transcriptase (hTERT)**
- In normal adult tissues, telomerase activity is typically low or below detectable limits.
- Ectopic expression of **hTERT** in normal somatic cells can prevent replicative senescence.

hTERT Immortalized Cell Line Panel (ATCC^{TM})

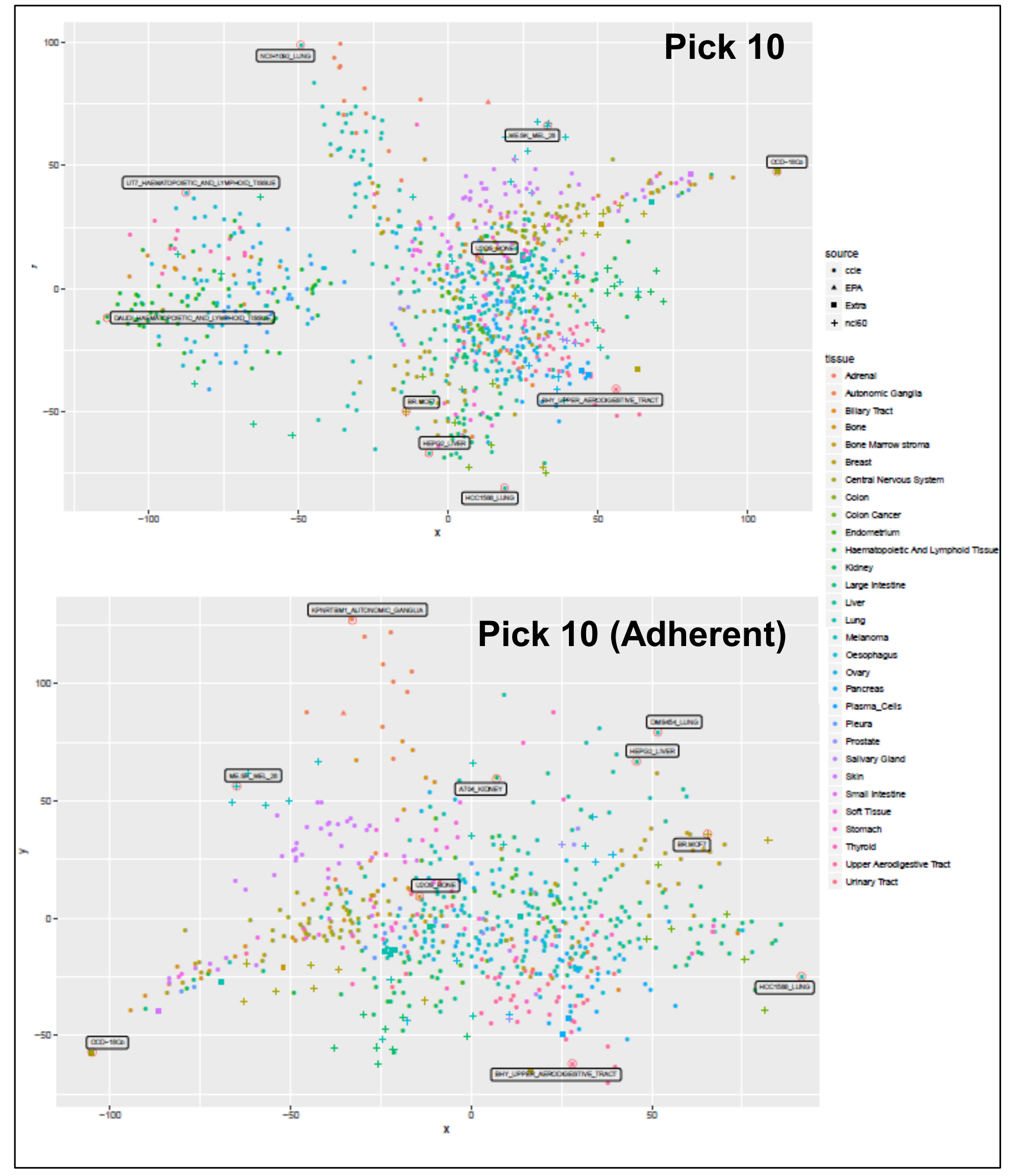
Name	Tissue	Germ Lineage	ATCC Number	Disease State	Morphology
TeloHAEC	Aorta	Endothelial	CRL-4052	Normal	endothelial
TIME	Foreskin; Dermal Microvascular Endothelium	Endothelial	CRL-4025	Normal	endothelial-like
HUVEC/TERT2	Umbilical Vascular Endothelium	Endothelial	CRL-4053	Normal	endothelial-like
HSAEC-1-KT	Lung, Small Airway	Epithelial	CRL-4050	Normal	epithelial, packed cuboidal
HBEC3-KT	Lung, Bronchial	Epithelial	CRL-4051	Normal	epithelial, packed cuboidal
hTERT-HME1	Breast; Mammary Gland	Epithelial	CRL-4010	Normal	epithelial-like
RPTEC/TERT1	Renal cortex, proximal tubules	Epithelial	CRL-4031	Normal	epithelial-like
hTERT RPE-1	Retina, Eye	Epithelial	CRL-4000	Normal	epithelial-like
CHON-001	Long Bone, Cartilage	Epithelial	CRL-2846	Normal	fibroblast-like
hTERT-HPNE	Pancreas, Duct	Fibroblast	CRL-4023	Normal	epithelial-like
Ker-CT	Foreskin	Intermediary	CRL-4048	Normal	epithelial
ASC52telo	Adipose-derived mesenchymal stem cell	Keratinocyte	SCRC-4000	Normal	fibroblast-like
BJ-5ta	Foreskin	Fibroblast	CRL-4001	Normal	Fibroblast-like

- Acquired a variety of cell lines immortalized using hTERT from American Tissue Culture Collection (ATCC^{TM}).
- Derived from normal (i.e. non-cancerous) tissue
- Cells maintain phenotypic characteristics of source tissue in vitro.
- Karyotypes are more “normal” compared to analogous cancer cell lines.
- Most require cocktails of media supplements to support growth in vitro.



- Cell Painting fluoroprobes illustrate morphological heterogeneity in hTERT cell lines

Data Driven Selection of CCLE Cell Lines

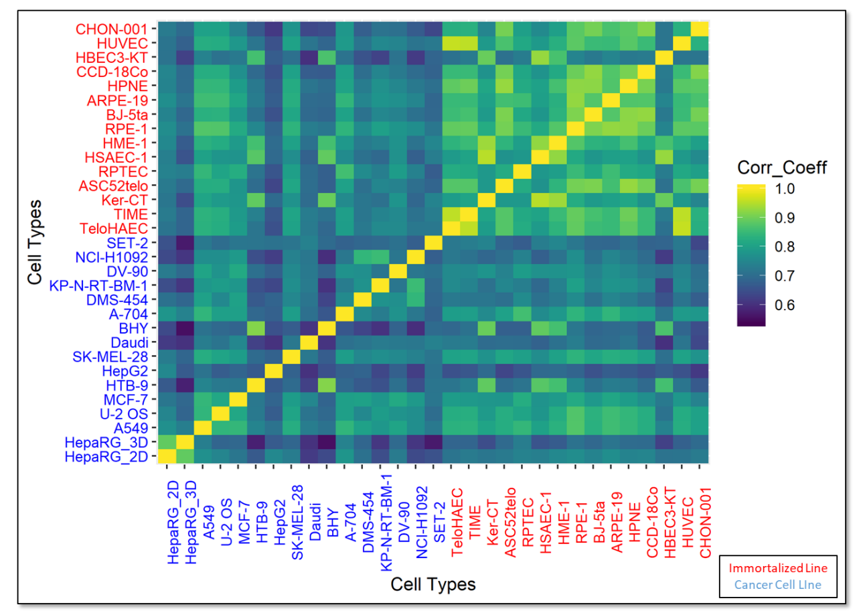


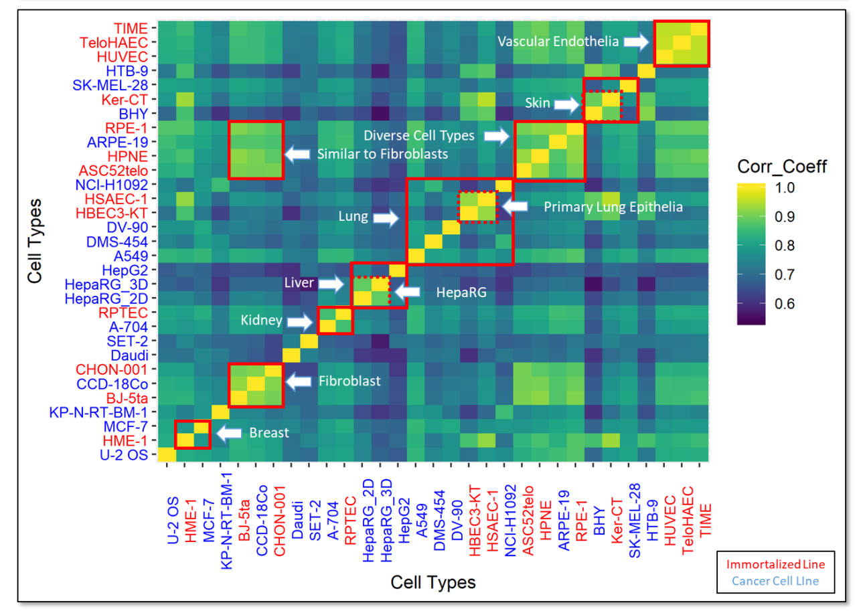
Cell Line	ExPASy CelloSaurus Accession	Tissue Origin	Disease	Growth Mode	Morphology	Media Formulation	Source	Reason Picked
MCF-7	CVCL_0031	Breast	Adenocarcinoma	adherent	epithelial	DMEM + 10% FBS	ATCC (HTB-22^{TM})	Seed Line
U-2 OS	CVCL_0042	Bone	Osteosarcoma	adherent	epithelial	DMEM + 10% FBS	ATCC (HTB-96^{TM})	Seed Line
HepG2	CVCL_0027	Liver	Hepatoblastoma	adherent	epithelial	DMEM + 10% FBS	ATCC (HB-8065^{TM})	Seed Line
Daudi	CVCL_0008	Peripheral Blood (B lymphoblast)	Burkitt’s Lymphoma	suspension	lymphoblast	RPMI-1640 + 10% FBS	ATCC (CCL-213^{TM})	Pick 10 (all)
CCD-18Co	CVCL_2379	Colon	none	adherent	fibroblast	EMEM + 10% FBS	ATCC (CRL-1459^{TM})	Pick 10 (all)
NCI-H1092	CVCL_1454	Lung	Small cell lung cancer (stage E carcinoma)	suspension	n/a	HITES + 5% FBS	ATCC (CRL-5855^{TM})	Pick 10 (all)
HCC-1588	CVCL_A351	Lung	Squamous cell carcinoma	adherent	epithelial	RPMI-1640 + 10% FBS	Creative Bioarray (CSC-C9399L)	Pick 10 (all)
UT-7	CVCL_2233	Bone Marrow	Acute Myeloid Leukemia	suspension	singlets	alpha-MEM | 20% FBS | 5 ng/mL GM-CSF	DSMZ (ACC 137)	Pick 10 (all)
BHY	CVCL_1086	Upper Aerodigestive Tract	Oral Squamous call carcinoma	adherent	epithelial	DMEM + 10% FBS	DSMZ (ACC 404)	Pick 10 (all)
SK-MEL-28	CVCL_0526	Skin	Melanoma	adherent	polygonal	EMEM + 10% FBS	ATCC (HTB-72^{TM})	Pick 10 (all)
KP-N-RT-BM-1	CVCL_1339	CNS	Neuroblastoma	adherent	Neuroblast-like	RPMI-1640 | 10% FBS	JCRB / XenoTech (IFO50432)	Pick 10 (adherent)
DMS 454	CVCL_2438	Lung	Small cell lung carcinoma	adherent	polygonal	Waymouth's MB 752/1 | 2 mM glutamine | 10% FBS	ECACC / Millipore (95062832)	Pick 10 (adherent)
A-704	CVCL_1065	Kidney	Renal cell carcinoma	adherent	epithelial	EMEM | 10% FBS	ATCC (HTB-45^{TM})	Pick 10 (adherent)

RESULT: Identified 13 cell lines with selected across a large variety of tissue origins and disease states.

TempO-Seq Gene Expression

- Cells were acquired, cultured and **TempO-Seq whole transcriptome** profiles were generated:
  - 16 Cancer Cell Lines (i.e. data-driven selections + 3 others)
  - 13 hTERT Immortalized Primary Cell Lines
  - HepaRG 2D Differentiated, HepaRG 3D Differentiated**





- hTERT cell lines had more similarity in gene expression compared to cancer cell lines.
- Clusters of similarity observed in cell lines of similar tissue origin

Data Driven Ranking of Cell Line Collection

Application: High-Throughput Transcriptomics (HTTr) with TempO-Seq
Input Cell Line Set: All Cell Lines, All Growth Modes (except HepaRG-3D)

Highlighted = Anchor Cell Lines

Blue = Euclidean Distances Based on Whole Transcriptome

Red = Euclidean Distances Based on Druggable Genome

Normal Font = Cancer Cell Lines

Bold Italicized Font = Immortalized Cell LInes

Cell_Type	Tissue_Origin	Cell_Type	Tissue_Origin	Cell_Type	Tissue_Origin	Cell_Type	Tissue_Origin	Cell_Type	Tissue_Origin	Cell_Type	Tissue_Origin
MCF-7	Breast	MCF-7	Breast	HepaRG_2D	Liver	HepaRG_2D	Liver	MCF-7	Breast	MCF-7	Breast
BHY	Skin	HepaRG_2D	Liver	Daudi	Immune	Daudi	Immune	HepaRG_2D	Liver	HepaRG_2D	Liver
HepG2	Liver	Daudi	Immune	***CHON-001***	***Fibroblast***	***CHON-001***	***Fibroblast***	BHY	Skin	Daudi	Immune
Daudi	Immune	***CHON-001***	***Fibroblast***	HepG2	Liver	HepG2	Liver	Daudi	Immune	***CHON-001***	***Fibroblast***
CHON-001	***Fibroblast***	HepG2	Liver	BHY	Skin	BHY	Skin	HepG2	Liver	HepG2	Liver
NCI-H1092	Lung	BHY	Skin	NCI-H1092	Lung	NCI-H1092	Lung	***CHON-001***	***Fibroblast***	BHY	Skin
DV-90	Lung	SK-MEL-28	Skin	SK-MEL-28	Skin	SK-MEL-28	Skin	NCI-H1092	Lung	SK-MEL-28	Skin
SET-2	Immune	DMS-454	Lung	DV-90	Lung	DV-90	Lung	SET-2	Immune	DMS-454	Lung
HepaRG_2D	Liver	SET-2	Immune	SET-2	Immune	SET-2	Immune	DV-90	Lung	SET-2	Immune
SK-MEL-28	Skin	DV-90	Lung	***TeloHAEC***	***Vascular***	***TeloHAEC***	***Vascular***	SK-MEL-28	Skin	DV-90	Lung
U-2 OS	Bone	KP-N-RT-BM-1	CNS	A-704	Kidney	A-704	Kidney	U-2 OS	Bone	KP-N-RT-BM-1	CNS
A-704	Kidney	***TeloHAEC***	***Vascular***	DMS-454	Lung	DMS-454	Lung	A-704	Kidney	***TeloHAEC***	***Vascular***
TeloHAEC	***Vascular***	A-704	Kidney	U-2 OS	Bone	U-2 OS	Bone	***TeloHAEC***	***Vascular***	A-704	Kidney
KP-N-RT-BM-1	CNS	U-2 OS	Bone	KP-N-RT-BM-1	CNS	KP-N-RT-BM-1	CNS	KP-N-RT-BM-1	CNS	U-2 OS	Bone
DMS-454	Lung	***RPTEC***	***Kidney***	MCF-7	Breast	MCF-7	Breast	DMS-454	Lung	***RPTEC***	***Kidney***
HBEC3-KT	***Lung***	A549	Lung	***RPTEC***	***Kidney***	***RPTEC***	***Kidney***	***HBEC3-KT***	***Lung***	A549	Lung
RPTEC	***Kidney***	NCI-H1092	Lung	A549	Lung	A549	Lung	***RPTEC***	***Kidney***	NCI-H1092	Lung
A549	Lung	***HBEC3-KT***	***Lung***	***HBEC3-KT***	***Lung***	***HBEC3-KT***	***Lung***	A549	Lung	***HBEC3-KT***	***Lung***
HPNE	***Pancreas***	***HPNE***	***Pancreas***	***HPNE***	***Pancreas***	***HPNE***	***Pancreas***	***HPNE***	***Pancreas***	***HPNE***	***Pancreas***
ARPE-19	***Retina***	***ARPE-19***	***Retina***	***ARPE-19***	***Retina***	***ARPE-19***	***Retina***	***ARPE-19***	***Retina***	***ARPE-19***	***Retina***
HTB-9	Urinary Bladder	***CCD-18Co***	***Fibroblast***	***CCD-18Co***	***Fibroblast***	***CCD-18Co***	***Fibroblast***	HTB-9	Urinary Bladder	***CCD-18Co***	***Fibroblast***
CCD-18Co	***Fibroblast***	HTB-9	Urinary Bladder	HTB-9	Urinary Bladder	HTB-9	Urinary Bladder	***CCD-18Co***	***Fibroblast***	HTB-9	Urinary Bladder
Ker-CT	***Skin***	***ASC52telo***	***Mesenchymal Stem Cell***	***ASC52telo***	***Mesenchymal Stem Cell***	***ASC52telo***	***Mesenchymal Stem Cell***	***Ker-CT***	***Skin***	***ASC52telo***	***Mesenchymal Stem Cell***
ASC52telo	***Mesenchymal Stem Cell***	***Ker-CT***	***Skin***	***Ker-CT***	***Skin***	***Ker-CT***	***Skin***	***ASC52telo***	***Mesenchymal Stem Cell***	***Ker-CT***	***Skin***
BJ-5ta	***Fibroblast***	***BJ-5ta***	***Fibroblast***	***BJ-5ta***	***Fibroblast***	***BJ-5ta***	***Fibroblast***	***BJ-5ta***	***Fibroblast***	***BJ-5ta***	***Fibroblast***
HME-1	***Breast***	***RPE-1***	***Retina***	***RPE-1***	***Retina***	***RPE-1***	***Retina***	***HME-1***	***Breast***	***RPE-1***	***Retina***
RPE-1	***Retina***	***HME-1***	***Breast***	***HME-1***	***Breast***	***HME-1***	***Breast***	***RPE-1***	***Retina***	***HME-1***	***Breast***
TIME	***Vascular***	***TIME***	***Vascular***	***TIME***	***Vascular***	***TIME***	***Vascular***	***TIME***	***Vascular***	***TIME***	***Vascular***
HUVEC	***Vascular***	***HUVEC***	***Vascular***	***HUVEC***	***Vascular***	***HUVEC***	***Vascular***	***HUVEC***	***Vascular***	***HUVEC***	***Vascular***
HSAEC-1	***Lung***	***HSAEC-1***	***Lung***	***HSAEC-1***	***Lung***	***HSAEC-1***	***Lung***	***HSAEC-1***	***Lung***	***HSAEC-1***	***Lung***

- Rank order of cell lines varies slightly depending upon which anchor cell line(s) are chosen.
- Rank order of cell lines also varies slightly based on gene input into Euclidean distance matrix calculations.
- Cancer cell lines highly represented at the top of ranked lists as compared to hTERT lines.
- Rank order of hTERT cell lines consistent regardless of inclusion of cancer cell lines.

Application: High-throughput Phenotypic Profiling w/ Cell Painting
Input Cell Line Set: hTERT Cell Lines + U-2 OS + HepaRG_2D + MCF-7

Highlighted = Anchor Cell Lines

Blue = Euclidean Distances Based on Whole Transcriptome

Red = Euclidean Distances Based on Druggable Genome

Normal Font = Cancer Cell Lines

Bold Italicized Font = Immortalized Cell LInes

Cell_Type	Tissue_Origin	Cell_Type	Tissue_Origin
U-2 OS	Bone	U-2 OS	Bone
MCF-7	Breast	MCF-7	Breast
HepaRG_2D	Liver	HepaRG_2D	Liver
HBEC3-KT	***Lung***	***CHON-001***	***Fibroblast***
CHON-001	***Fibroblast***	***HBEC3-KT***	***Lung***
TeloHAEC	***Vascular***	***TeloHAEC***	***Vascular***
RPTEC	***Kidney***	***RPTEC***	***Kidney***
ARPE-19	***Retina***	***ARPE-19***	***Retina***
HPNE	***Pancreas***	***HPNE***	***Pancreas***
Ker-CT	***Skin***	***CCD-18Co***	***Fibroblast***
CCD-18Co	***Fibroblast***	***Ker-CT***	***Skin***
ASC52telo	***Mesenchymal Stem Cell***	***ASC52telo***	***Mesenchymal Stem Cell***
BJ-5ta	***Fibroblast***	***BJ-5ta***	***Fibroblast***
HME-1	***Breast***	***RPE-1***	***Retina***
RPE-1	***Retina***	***HME-1***	***Breast***
TIME	***Vascular***	***TIME***	***Vascular***
HUVEC	***Vascular***	***HUVEC***	***Vascular***
HSAEC-1	***Lung***	***HSAEC-1***	***Lung***

Conclusions

- Content maximization of pair-wise Euclidean distance can be used as a data-driven approach to select sets of cell lines with high biological diversity.
- Selection of cancer cell lines using this approach resulted in a collection of cells with diverse tissue origins.
- Similarity in baseline gene expression among hTERT-immortalized cell lines was greater than amongst cancer cell lines selected from a database using the data-driven approach.
- Content maximization can be used as a data-driven approach for ranking cell lines that represent incrementally increasing amounts of biological space for incorporation into screening panels.
- Use of whole genome versus druggable genome as input had little impact on data-driven ranking of cell lines.

References

- Alter et al. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 2000; 97(18) doi: 10.1073/pnas.97.18.10101
- Barrentina et al. The Cancer Cell Line Encyclopedia enables predictive modellling of anticancer drug sensitivity. Nature 2012; 565(7738) doi: 10.1038/nature11003
- Bray et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc 2016; 11(9) doi: 10.1038/nprot.2016.105
- Finan et al. The druggable genome and support for target identification and validation in drug development. Sci Transl Med 2017; 9(383) doi: 10.1126/scitranslmed.aag1166
- Lamb et al. The Connectivity Map: using gene expression signatures to connect small molecules. Science 2006; 313(5795) doi: 10.1126/science.1132939
- Shoemaker et al. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer 2006; 6(10) doi: 10.1038/nrc1951
- Yeakley et al. A trichostatin A expression signature identified by TempO-Seq targeted whole transcriptome profiling. PLoS One 2017; 12(5) doi: 10.1371/journal.pone.0178302