



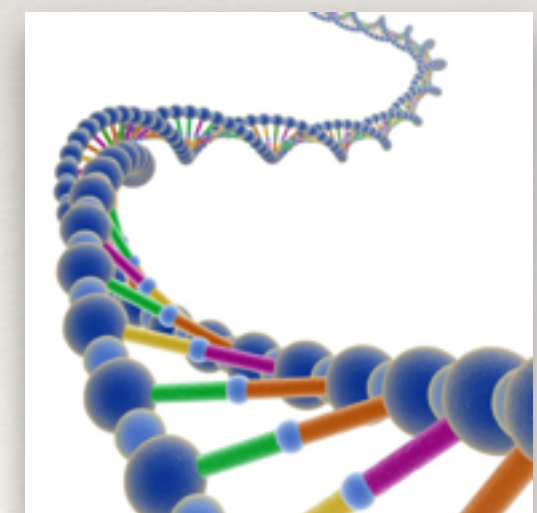
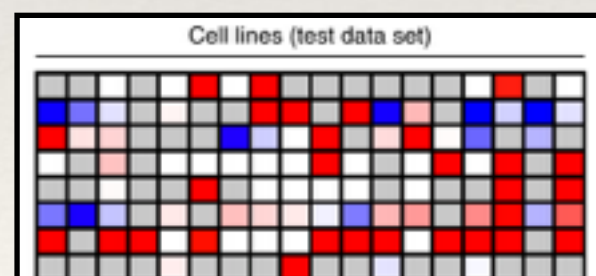
Normal Cells

Source: National Cancer Institute USA

# Identification of Cancer Cell Line based on NGS data

Bioinformatics Social Meeting

31.05.2016



# Agenda

---

1. Motivation Cancer Cell Line NGS

Identification

2. Established methods

3. Uniquorn Method

4. Outlook: RNA & Panel Seq

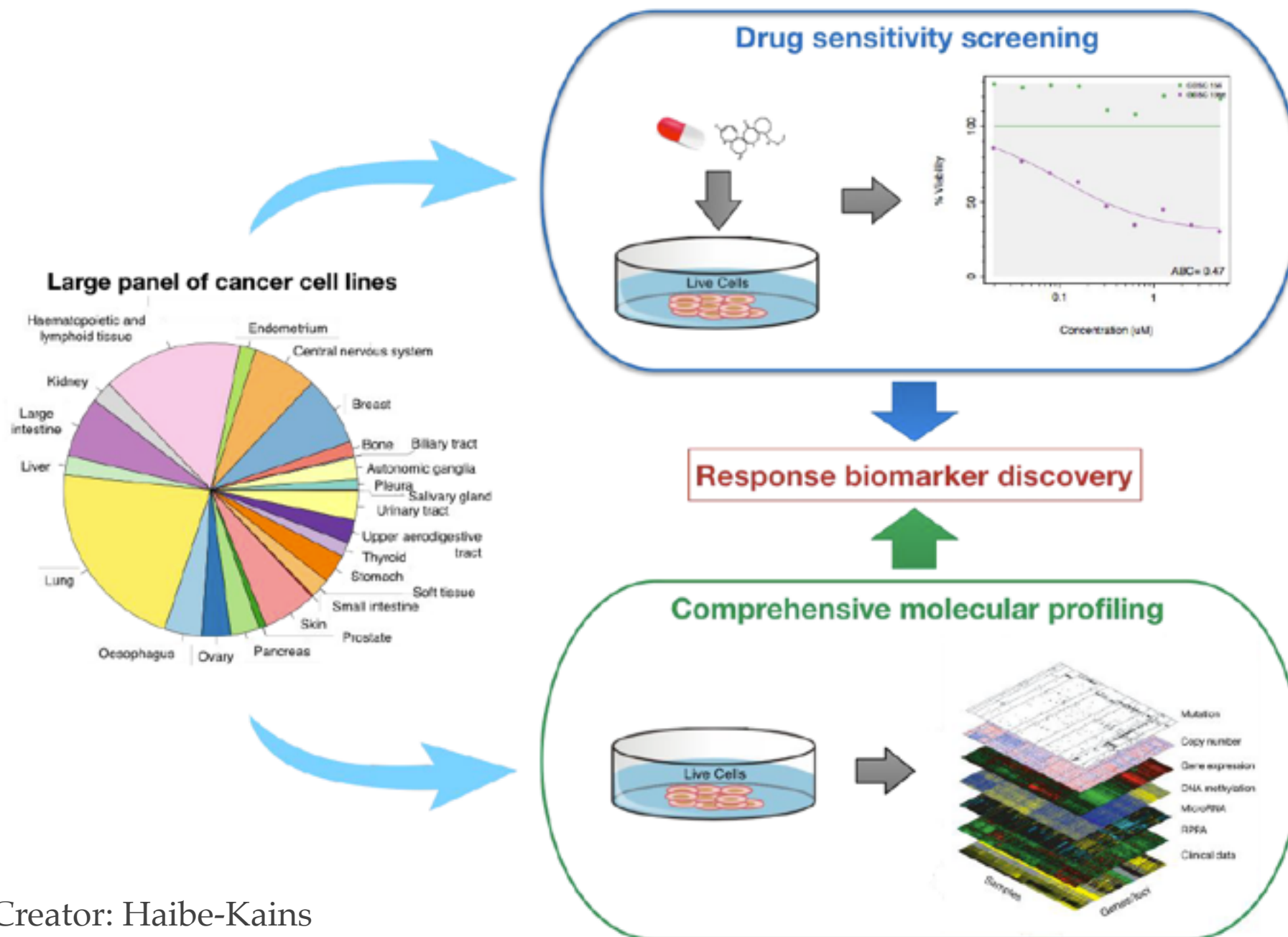


# The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity

Jordi Barretina<sup>1,2,3,9,\*</sup>, Giordano Caponigro<sup>4,\*</sup>, Nicolas Stransky<sup>1,\*</sup>, Kavitha Venkatesan<sup>4,\*</sup>,



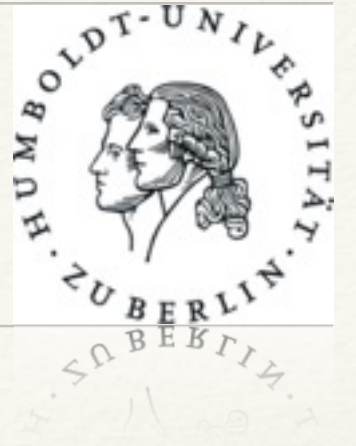
## Biomarker discovery using panels of cell lines



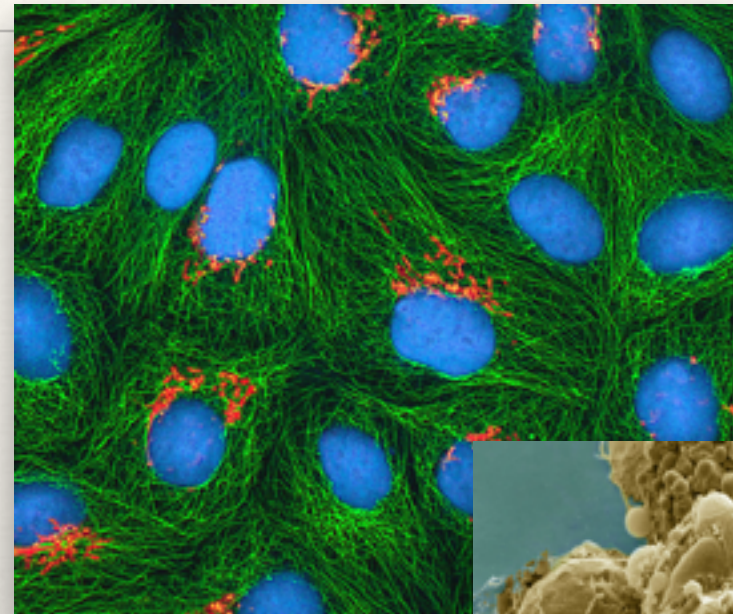
Creator: Haibe-Kains



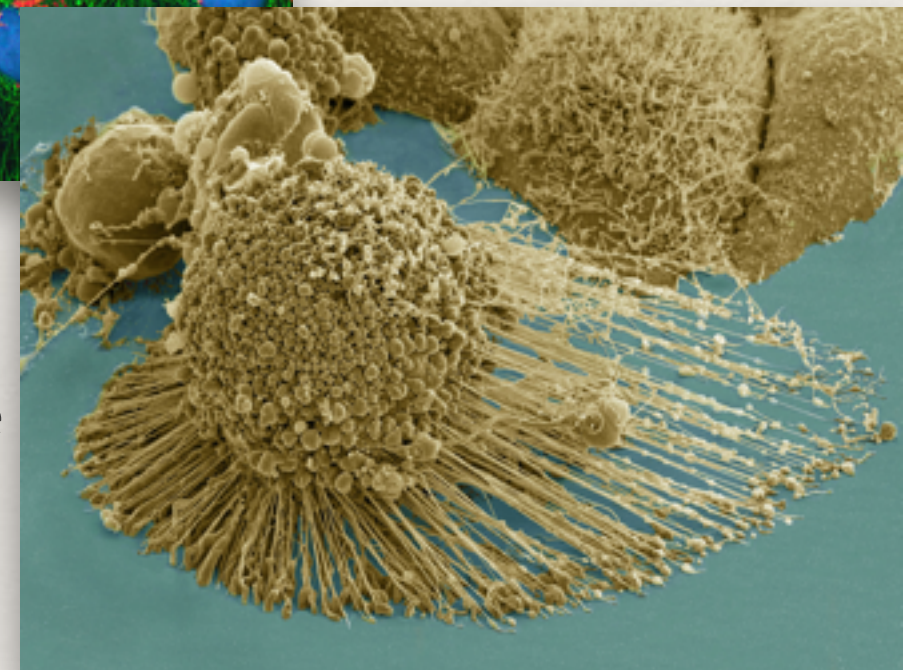
# Cancer Cell Lines



- ❖ Model organisms
- ❖ Experiment repeatability key feature
- ❖ Workhorse of drug-research



HeLa cell line



# The problem



**MDA-MB-435 cells are derived from M14 Melanoma cells—a loss for breast cancer, but a boon for melanoma research**

**James M. Rae · Chad J. Creighton ·  
Jeanne M. Meck · Bassem R. Haddad ·  
Michael D. Johnson**

**WIDESPREAD INTRASPECIES CROSS-CONTAMINATION OF HUMAN TUMOR  
CELL LINES ARISING AT SOURCE**

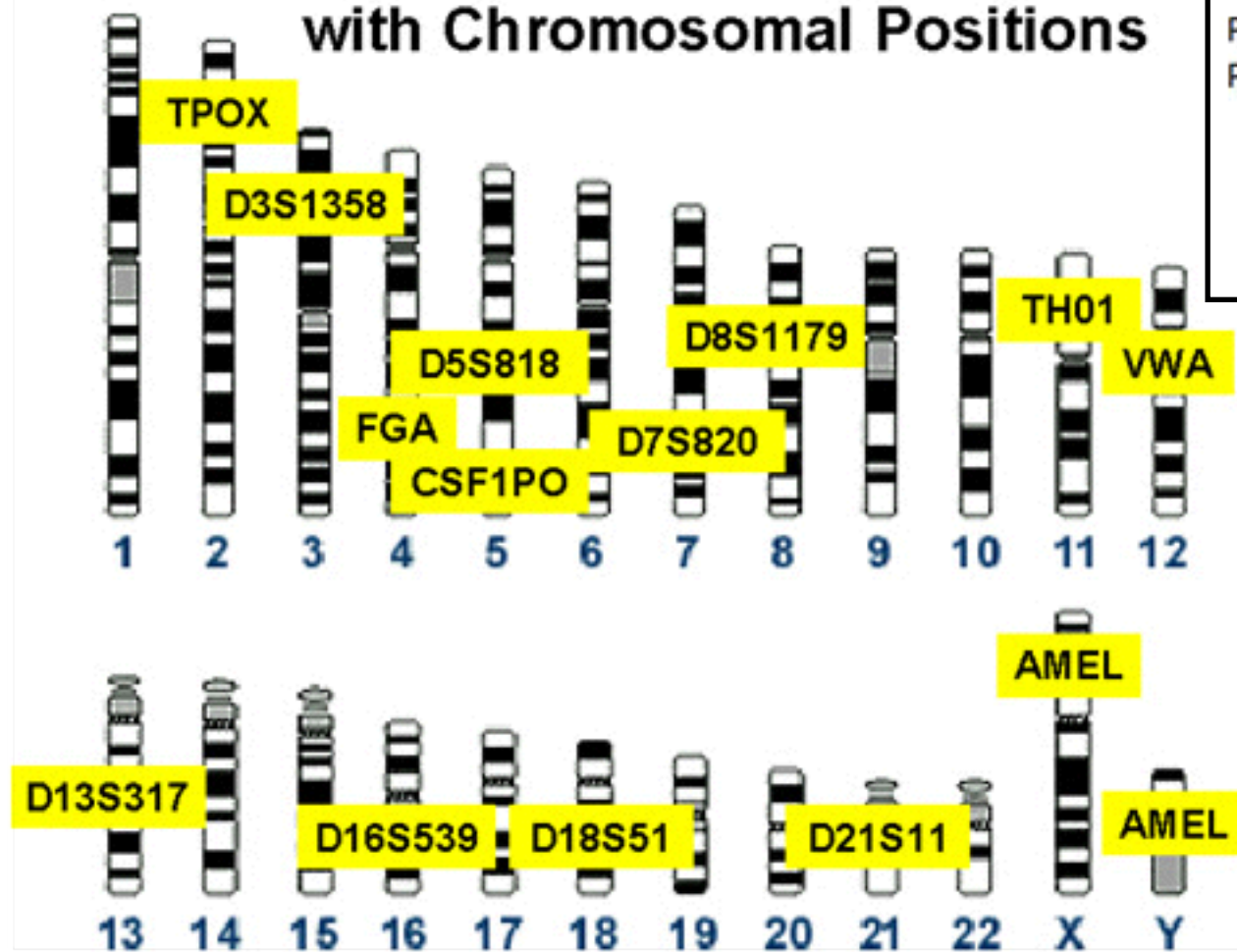
Roderick A.F. MACLEOD<sup>1\*</sup>, Wilhelm G. DIRKS<sup>1</sup>, Yoshinobu MATSUO<sup>2</sup>, Maren KAUFMANN<sup>1</sup>, Herbert MILCH<sup>1</sup> and Hans G. DREXLER<sup>1</sup>



# Gold-Standard: STR Not optimized for NGS



## 13 CODIS Core STR Loci with Chromosomal Positions

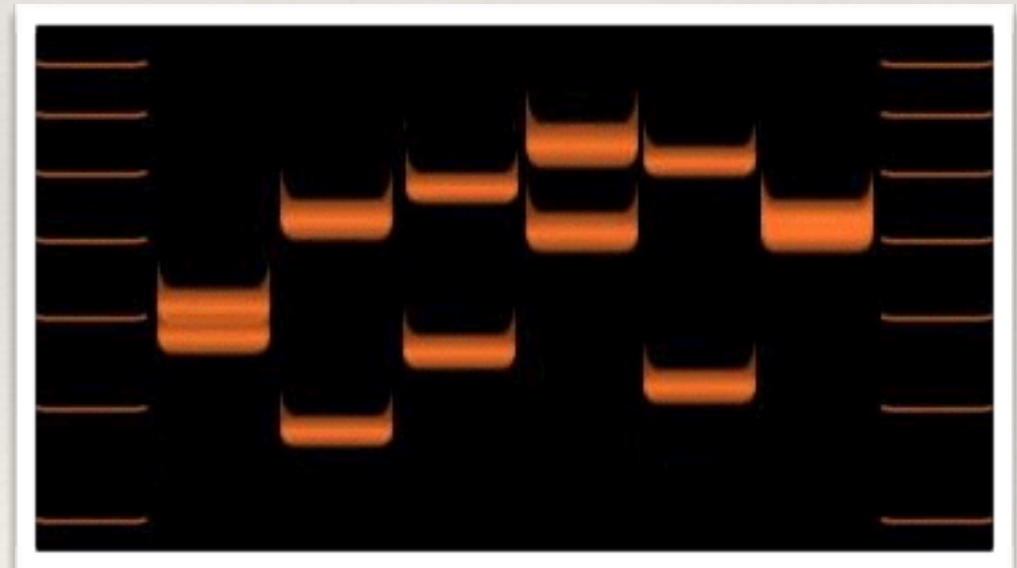


<http://www.cstl.nist.gov/div831/strbase/fbicore.htm>

Loci coverage required

Short tandem repeats	8 repeats
Participant 1	CTAGAGATAGATAGATAGATAGATAGATAGATAGACTAGACTAG
Participant 2	CTAGAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACTAGACTAGA
Participant 3	CTAGAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACTAGACTAGA
Participant 4	CTAGAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACTAGAC
	9 repeats
	10 repeats

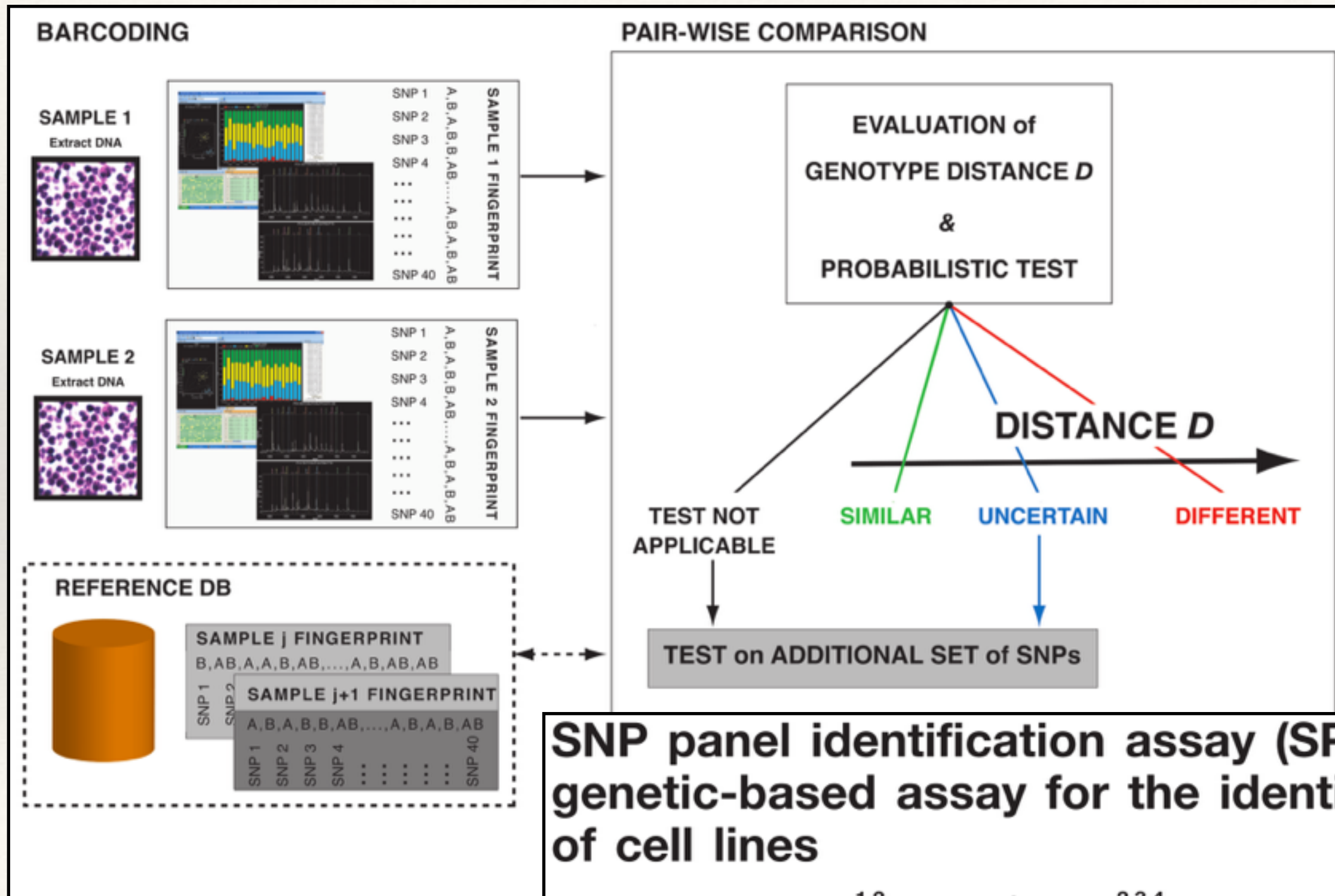
Creator: gabriel maldonado



<http://www.cstl.nist.gov/div831/strbase/fbicore.htm>

Wet-Lab based

# SNP-zygosity matching





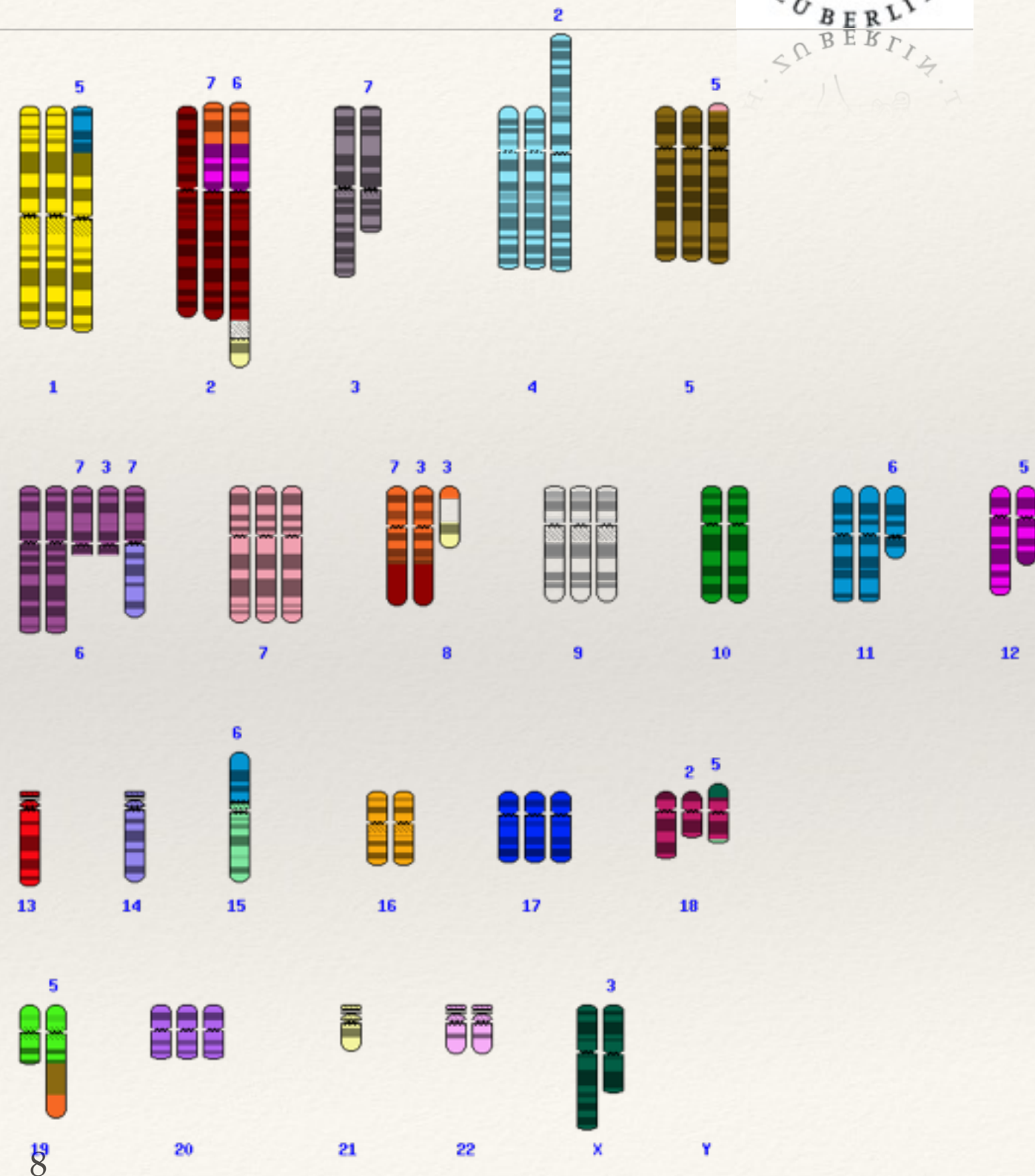
# Beyond SNP-Zygosity



- ❖ SNP-zygosity matching problematic in Cancer NGS
- ❖ Internet-shared CL NGS data problematic

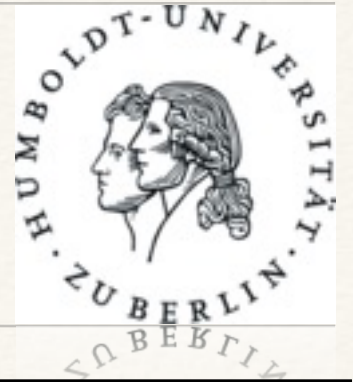
```
h-bon1
0/0/1:
0/0/1:
0/0/1:
0/0/1:
0/0/1:
0/0/1:
1/1/1:
```

tri-ploid  
VCF,  
zygosity matching  
problematic





# Uniquorn-Method



- ❖ Match (predominantly) rare, somatic passenger mutations

- ❖ Start + Length

Drop SNP-Zygosity constraint

The screenshot shows the Bioconductor website interface. At the top, there's a navigation bar with 'Home', 'Install', and 'Help' links. Below the Bioconductor logo, the breadcrumb trail reads 'Home » Bioconductor 3.3 » Software Packages » Uniquorn'. The package name 'Uniquorn' is displayed in a large green font. A row of status indicators shows: 'platforms all', 'downloads available', 'posts 0', 'in Bioc < 6 months', 'build ok', 'commits 3.50', and 'test coverage 0%'. Social media icons for Facebook and Twitter are present. The package description is 'Identification of cancer cell lines based on their weighted mutational/variational fingerprint'. It lists the author as Raik Otto and the maintainer as 'Raik Otto' <raik.otto@hu-berlin.de>. The citation is provided for R users. The 'Installation' section shows the R code to install the package using biocLite.

Bioconductor version: Release (3.3)

This package enables users to identify cancer cell lines. Cancer cell line misidentification and cross-contamination represents a significant challenge for cancer researchers. The identification is vital and in the frame of this package based on the locations/ loci of somatic and germline mutations/ variations. The input format is vcf/ vcf.gz and the files have to contain a single cancer cell line sample (i.e. a single member/genotype/gt column in the vcf file). The implemented method is optimized for the Next-generation whole exome and whole genome DNA-sequencing technology.

Author: Raik Otto

Maintainer: 'Raik Otto' <raik.otto@hu-berlin.de>

Citation (from within R, enter `citation("Uniquorn")`):

Otto R (2016). *Uniquorn: Identification of cancer cell lines based on their weighted mutational/variational fingerprint*. R package version 1.0.4.

**Installation**

To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("Uniquorn")
```

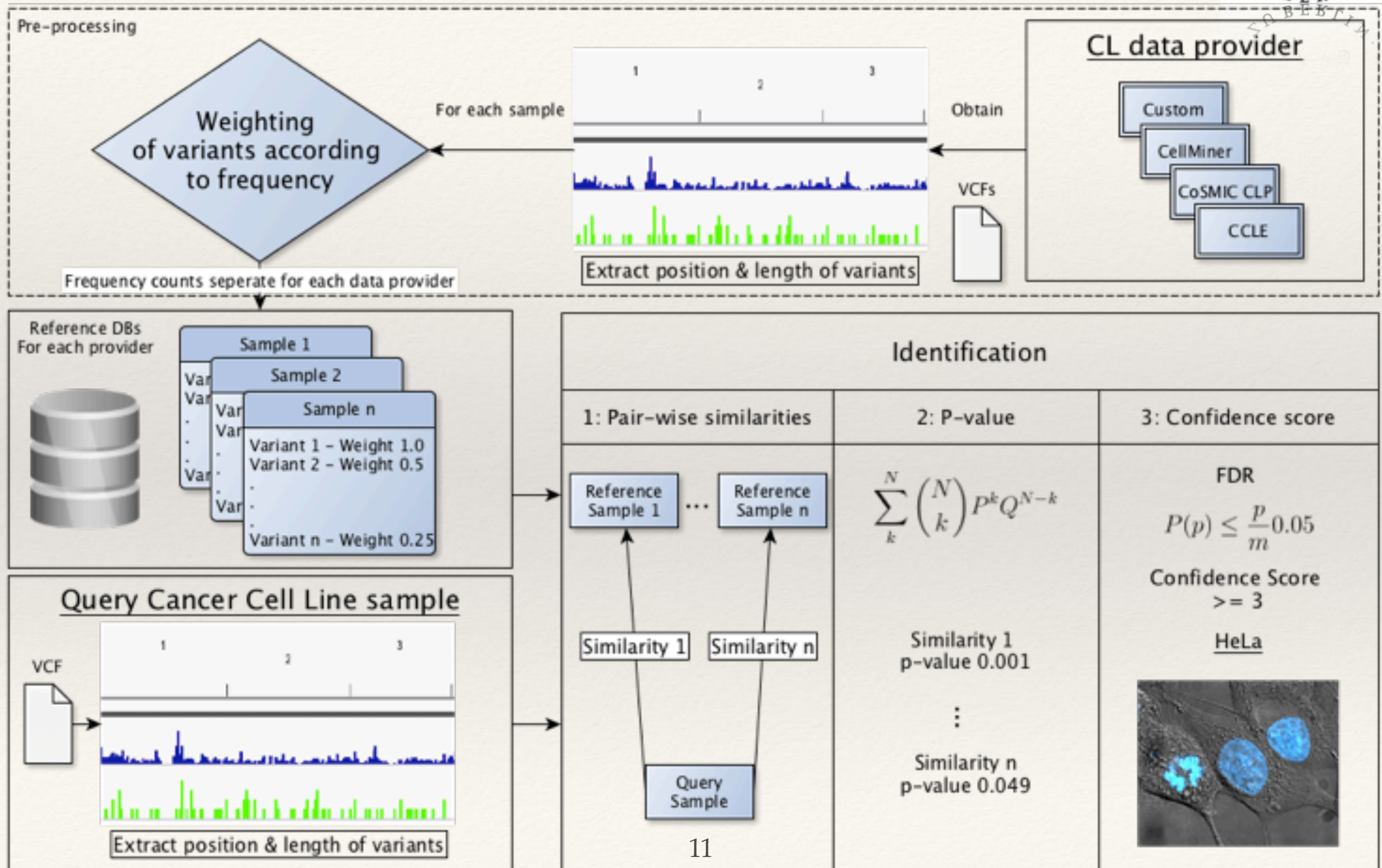
# Comparison Uniquorn



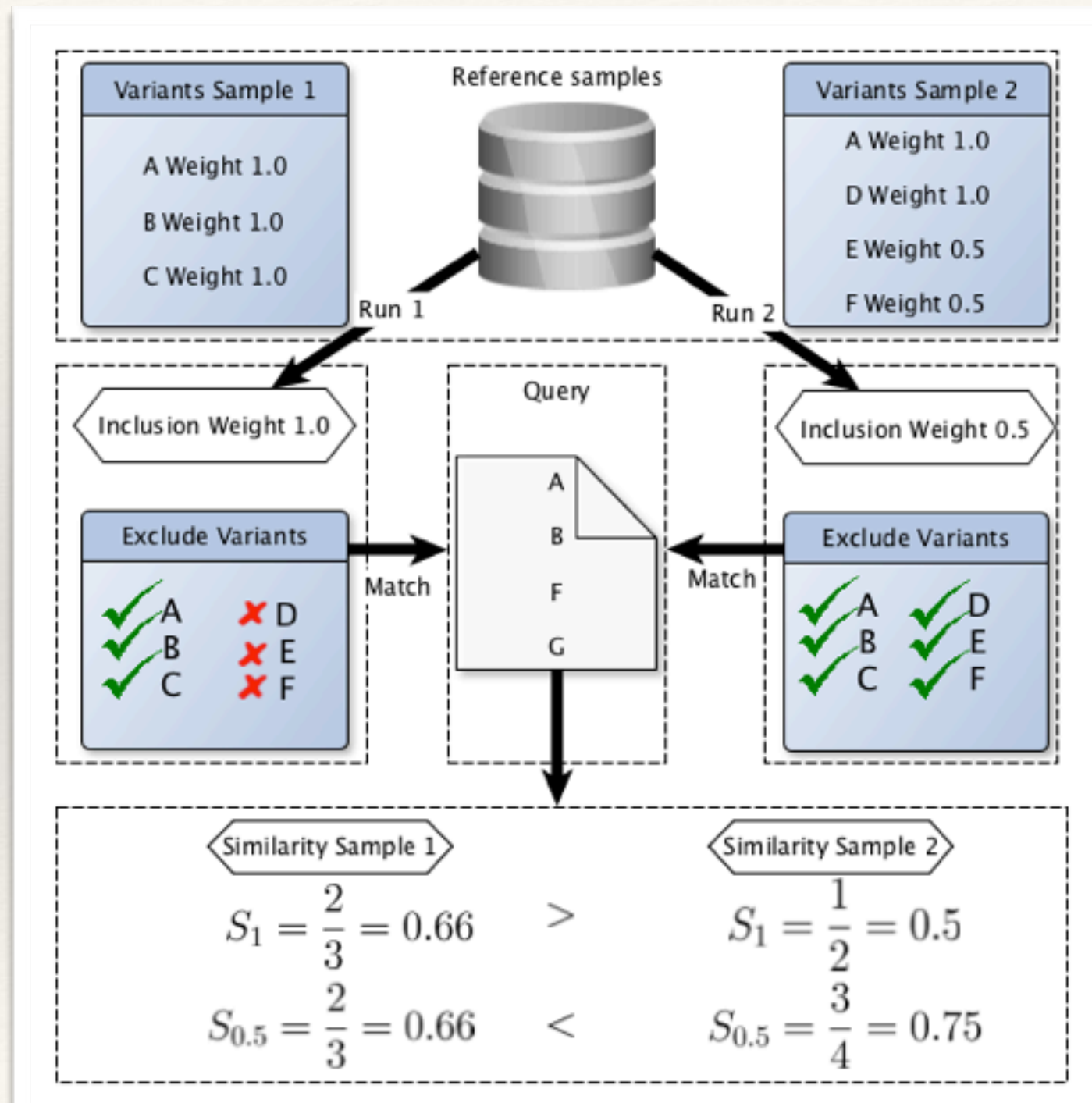
Identification Method for NGS CLs	Physical Sample Required	Experiments Required	Locus coverage required	Zygosity known & valid	Reference genome identical
Tandem-Repeat counting (9)	✓	✓	✗	✗	✗
SPIA (5)	✓	✓	✓	✓	✗
NGS SNP (17)	✗	✗	✓	✓	✓
NGS All Variants (Uniquorn)	✗	✗	✗	✗	✓



# Uniquorn Workflow

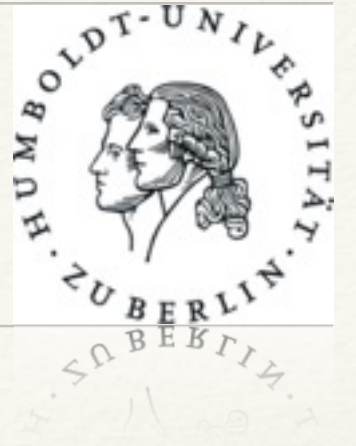


# Weighted Similarity





# Confidence Score



$$D_k = \binom{N}{k} P_l^k Q_l^{N-k} = \frac{N!}{k! (N-k)!} P_l^k Q_l^{N-k}$$

Likelihood to commit statistical

error type 1 when rejecting  $H_0$

that dis-similarity is due to

different CL-identity

$$P_{val}(S_w) = 1 - \sum_{k=0}^{k-1} D_k = \sum_k^N D_k$$

$$-\text{Log}_e(q - \text{value}) := C_s$$

# Results Uniquorn



```
> identify_vcf_file("~/BSM/M14.vcf")
```

	CL	CL_source	Found_muts	Count_mutations	Conf_score	Conf_score_sig
192	M14	COSMIC	1082	1705	100	TRUE
45	M14	CELLMINER	278	286	100	TRUE
1374	MDAMB435S	CCLE	72	143	100	TRUE
25	HCT_116	CELLMINER	4	3092	0	FALSE
6	HCC_2998	CELLMINER	3	9308	0	FALSE

Bioconductor R-package *Uniquorn*

**MDA-MB-435 cells are derived from M14 Melanoma cells—a loss for breast cancer, but a boon for melanoma research**

**James M. Rae · Chad J. Creighton ·  
Jeanne M. Meck · Bassem R. Haddad ·  
Michael D. Johnson**



# Benchmark DNA-seq



Expected	3555			
Inclusion weight of	1.0	0.5	0.25	0.0
True positives	3396	3526	3529	3547
False negatives	159	29	26	8
False positives	25	67	86	43046
True negatives	~4 mil.	~4 mil.	~4 mil.	~3.9 mil
Sensitivity %	96	99	99	99
Specificity %	99	99	99	99
F1 %	98	99	98	14
PPV	99	98	96	8

# Outlook



## RNA-seq

```
> identify_vcf_file("~/BSM/BON1.RNA_SEQ.vcf.gz")
```

	CL	CL_source	Found_muts	Count_mutations	Conf_score	Conf_score_sig
1988	BON1_PLOIDY3	CUSTOM	25799	71833	100	TRUE
128	NCI-H2342	COSMIC	3	2435	0	FALSE
6	HCC_2998	CELLMINER	2	9308	0	FALSE

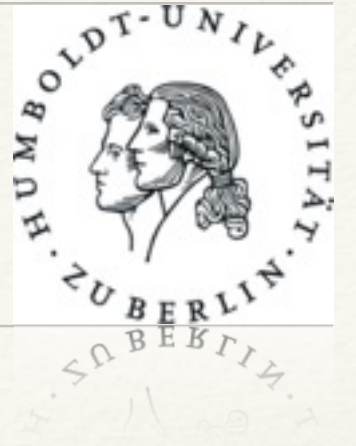
## Panel-seq

```
> identify_vcf_file("~/BSM/BON1_panel.vcf", confidence_score = 0)
```

	CL	CL_source	Found_muts	Count_mutations	Conf_score	Conf_score_sig
1988	BON1_PLOIDY3	CUSTOM	72	71833	0	TRUE
922	MDA-MB-436	COSMIC	1	372	0	FALSE
1	SK_MEL_2	CELLMINER	0	921	0	FALSE



# Summary Key Features



1. Add custom „unknown” samples

- No SNP-zygosity reference

2. Works on

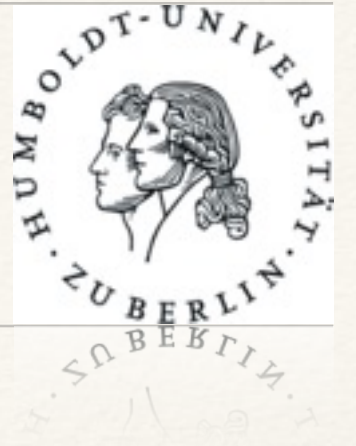
- DNA-seq (proven)
- RNA-seq (likely to be proven)
- Panel-seq (explorative)

3. Integrate into variant callign

4. Quick

5. Detects cross-contamination  
(likely to be proven)

# Pros & Cons



## Advantages

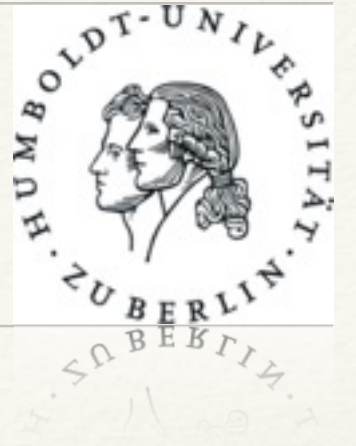
- ✓ Benchmark: High Sensitivity & Specificity
- ✓ Scalable for High-Throughput
- ✓ Free R-Package *Uniquorn*
- ✓ ~2000 CL samples available

## Disadvantages

- ✗ Pair-wisely non-similar NGS library samples
- ✗ Parameter-based
- ✗ *Background-noise* varies



# Take away message



Identify NGS CL data before usage

BiocLite(„Uniquorn“)

# NGS Comparability CL panels

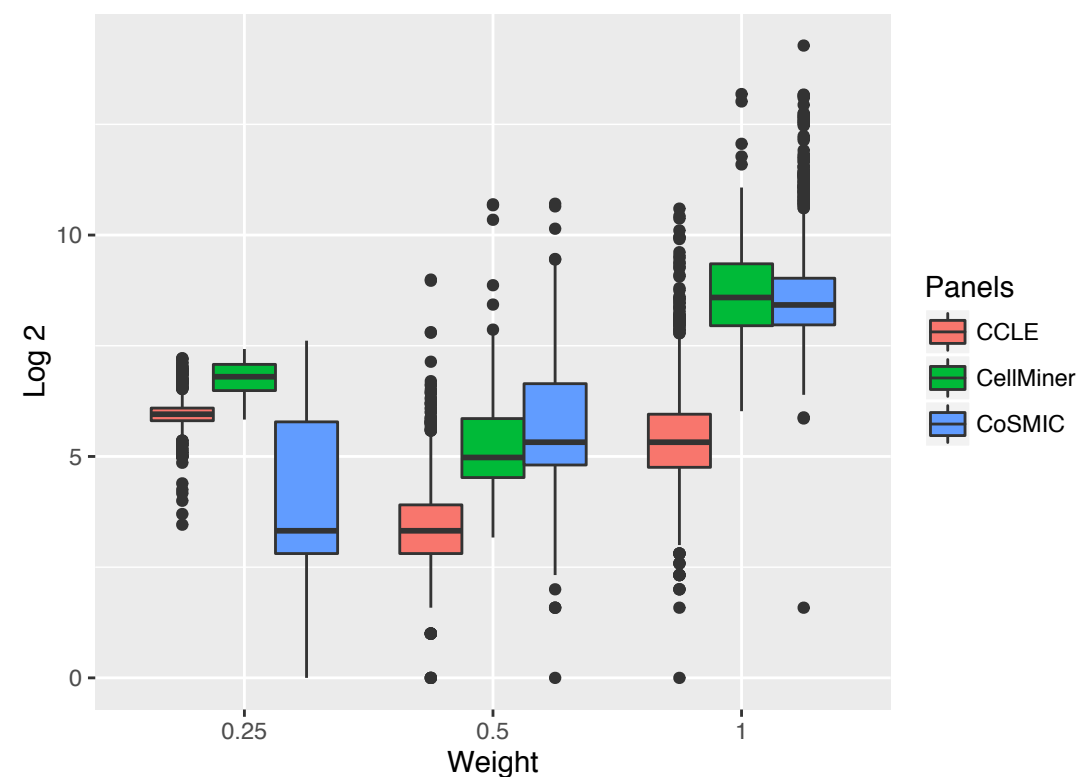
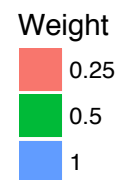
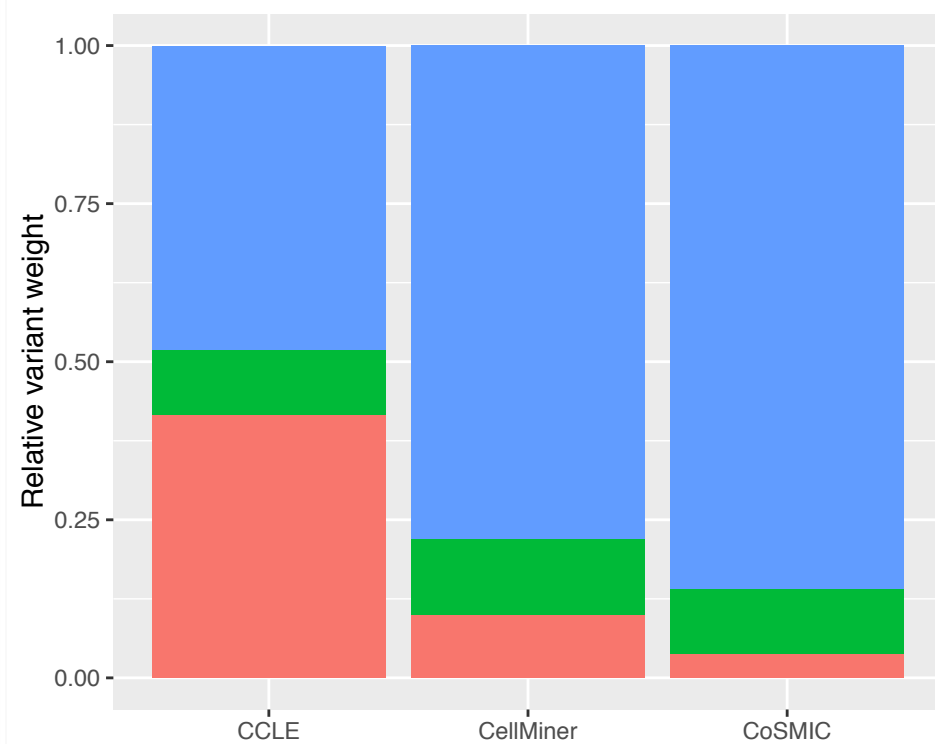
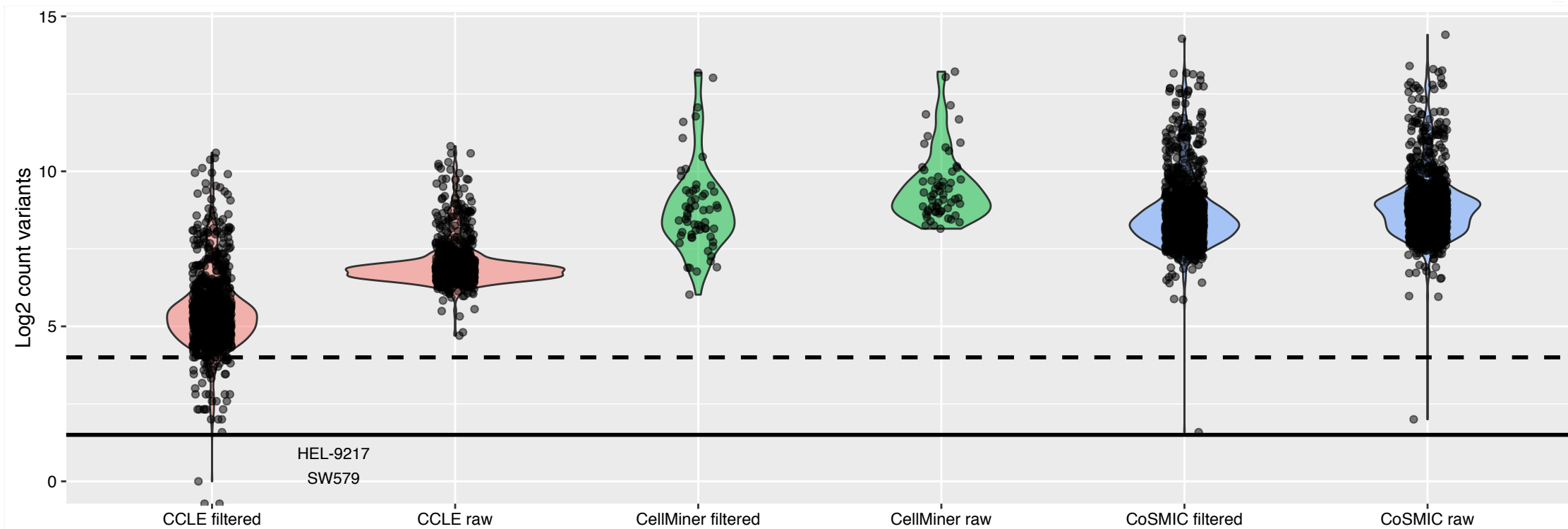


Reference set	Variants*	Cancer Cell Lines	Ø Variants CL*	Covered genes	Variant calling software	SNP MAF filtering
CoSMIC CLP	760	1025	7,4	20965	Caveman Pindel	> 0.0 (all)
CCLE	140	904	1,5	1651	MuTect	>= 0.05
CellMiner	0,68	60	0,01	>20k	GATK	None

\* 1 = 100k



# Results Weighted Libraries



# Gold Standard



Based on  
Regularized  
name matching  
(Only Alpha-numercial +  
Reports for exceptions)

