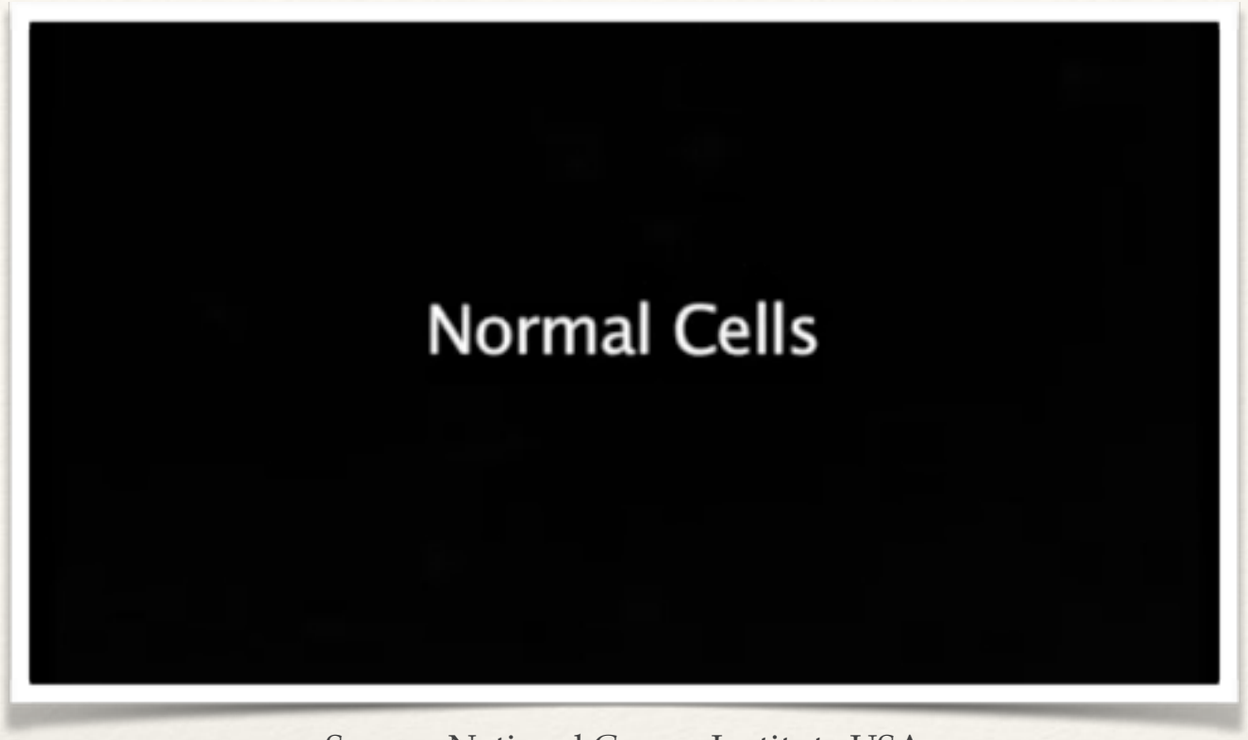


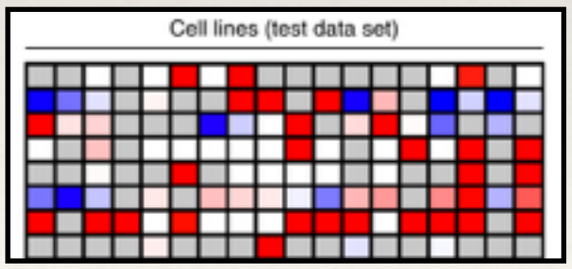
Source: National Cancer Institute USA

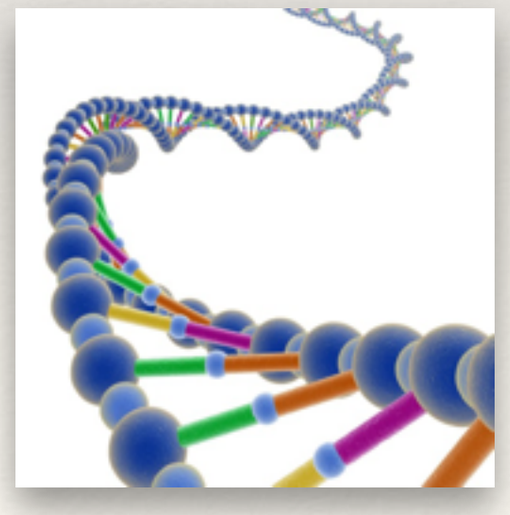
Humboldt-Universität zu Berlin

Identification of Cancer Cell Line based on NGS data

Bioinformatics Social Meeting

31.05.2016





"Lung cancer" by Emmanuelm at en.wikipedia. Licensed under CC BY 3.0 via Commons - https://commons.wikimedia.org/wiki/File:Lung_cancer.jpg#/media/File:Lung_cancer.jpg

Bioinformatics Social Meeting

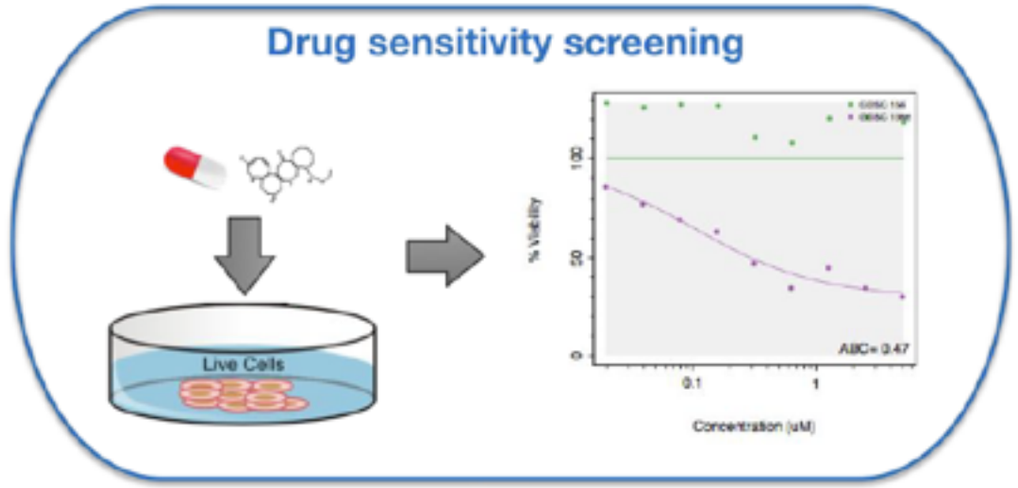
Agenda

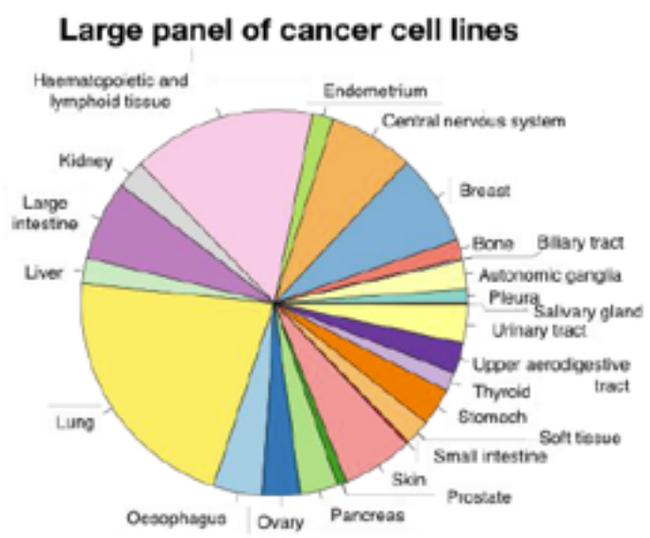
1. Motivation Cancer Cell Line NGS Identification
2. Established methods
3. Uniquorn Method
4. Outlook: RNA & Panel Seq

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

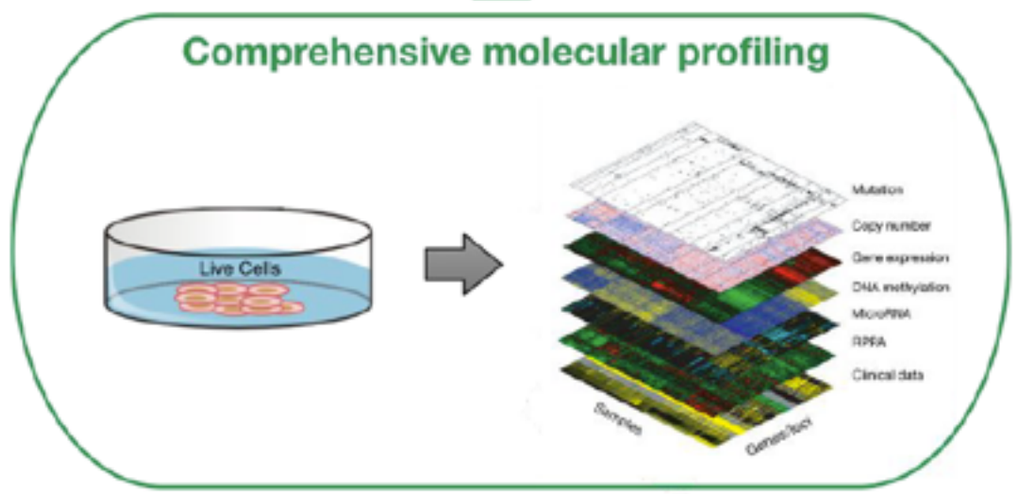
Jordi Barretina[1,2,3,9,*], Giordano Caponigro[4,*], Nicolas Stransky[1,*], Kavitha Venkatesan[4,*],

Biomarker discovery using panels of cell lines



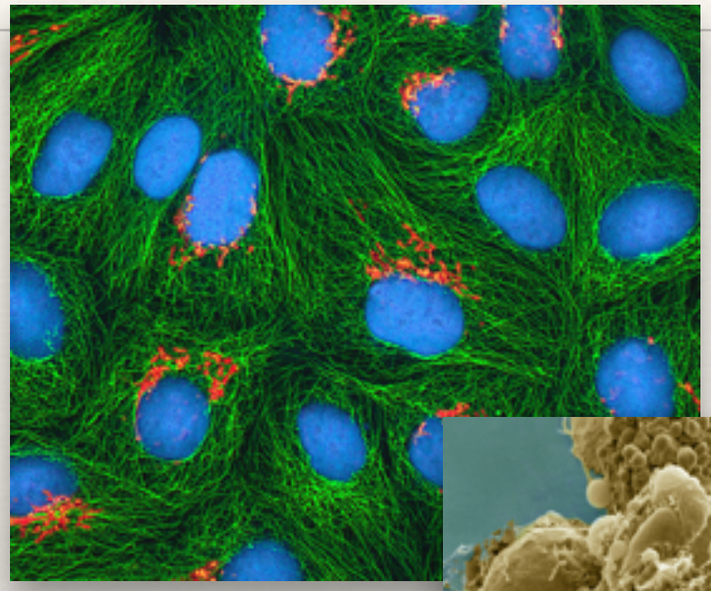


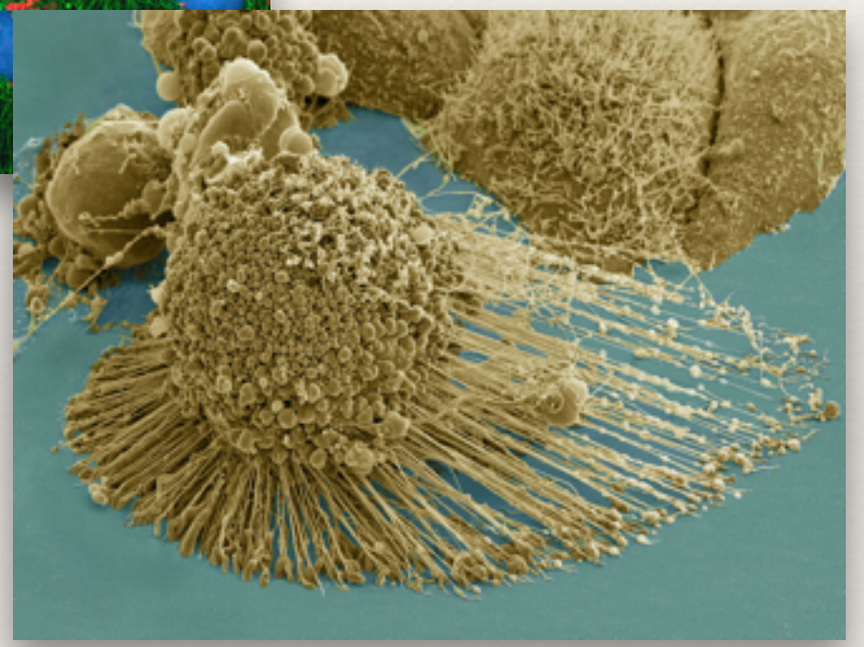
Response biomarker discovery



Creator: Haibe-Kains

Cancer Cell Lines





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

HeLa cell line

"HeLa-IV" by National Institutes of Health (NIH) - National Institutes of Health (NIH) https://imagebank.nih.gov/details.cfm?imageid=1463. Licensed under Public Domain via Commons - https://commons.wikimedia.org/wiki/File:HeLa-IV.jpg#/media/File:HeLa-IV.jpg

The problem

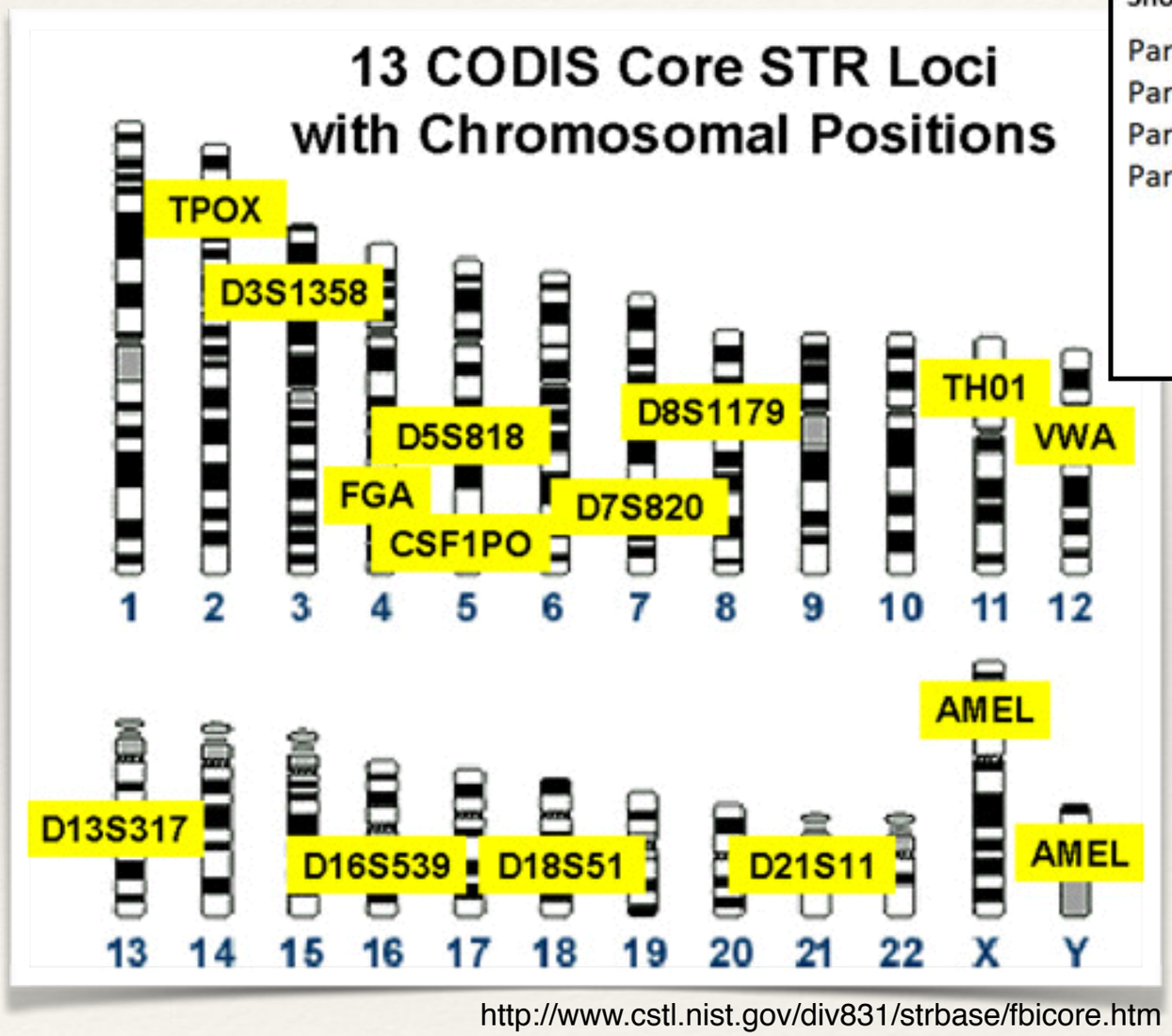
MDA-MB-435 cells are derived from M14 Melanoma cells—a loss for breast cancer, but a boon for melanoma research

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

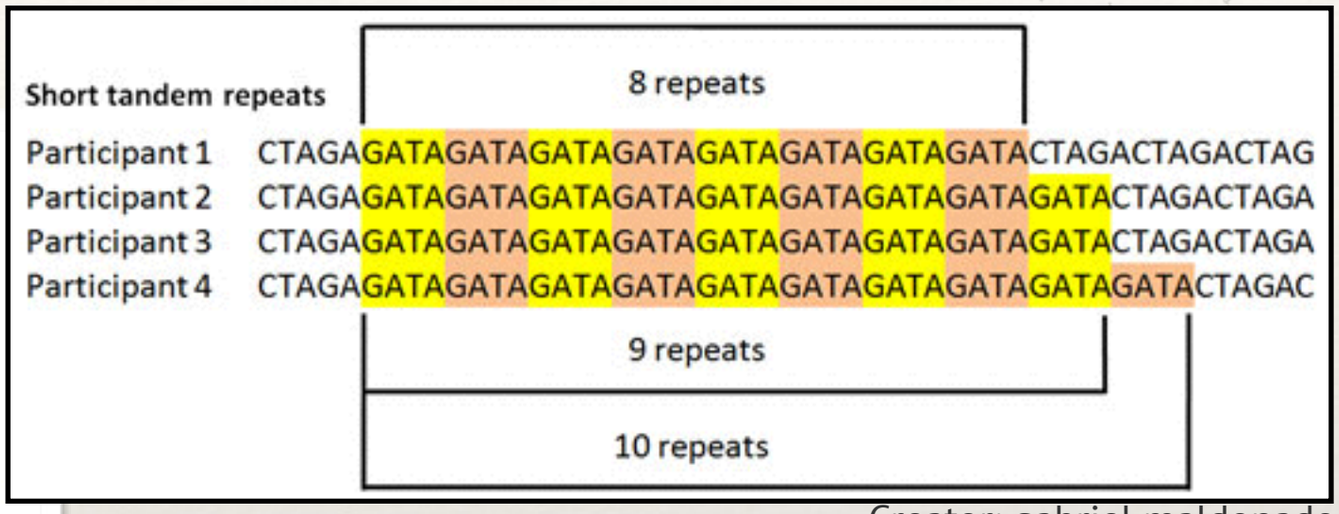
Roderick A.F. MacLeod[1]*, Wilhelm G. Dirks[1], Yoshinobu Matsuo[2], Maren Kaufmann[1], Herbert Milch[1] and Hans G. Drexler[1]

Gold-Standard: STR
Not optimized for NGS

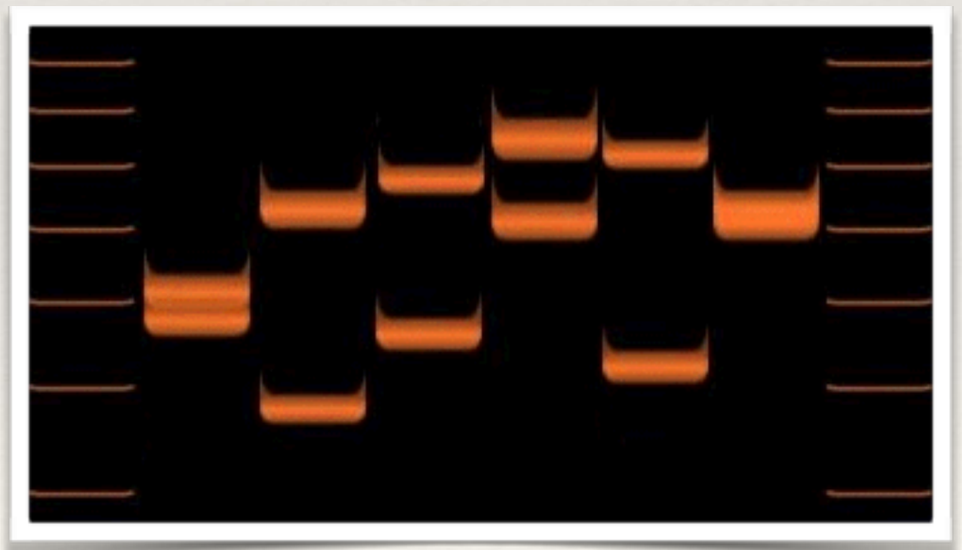


http://www.cstl.nist.gov/div831/strbase/fbicore.htm

Loci coverage required



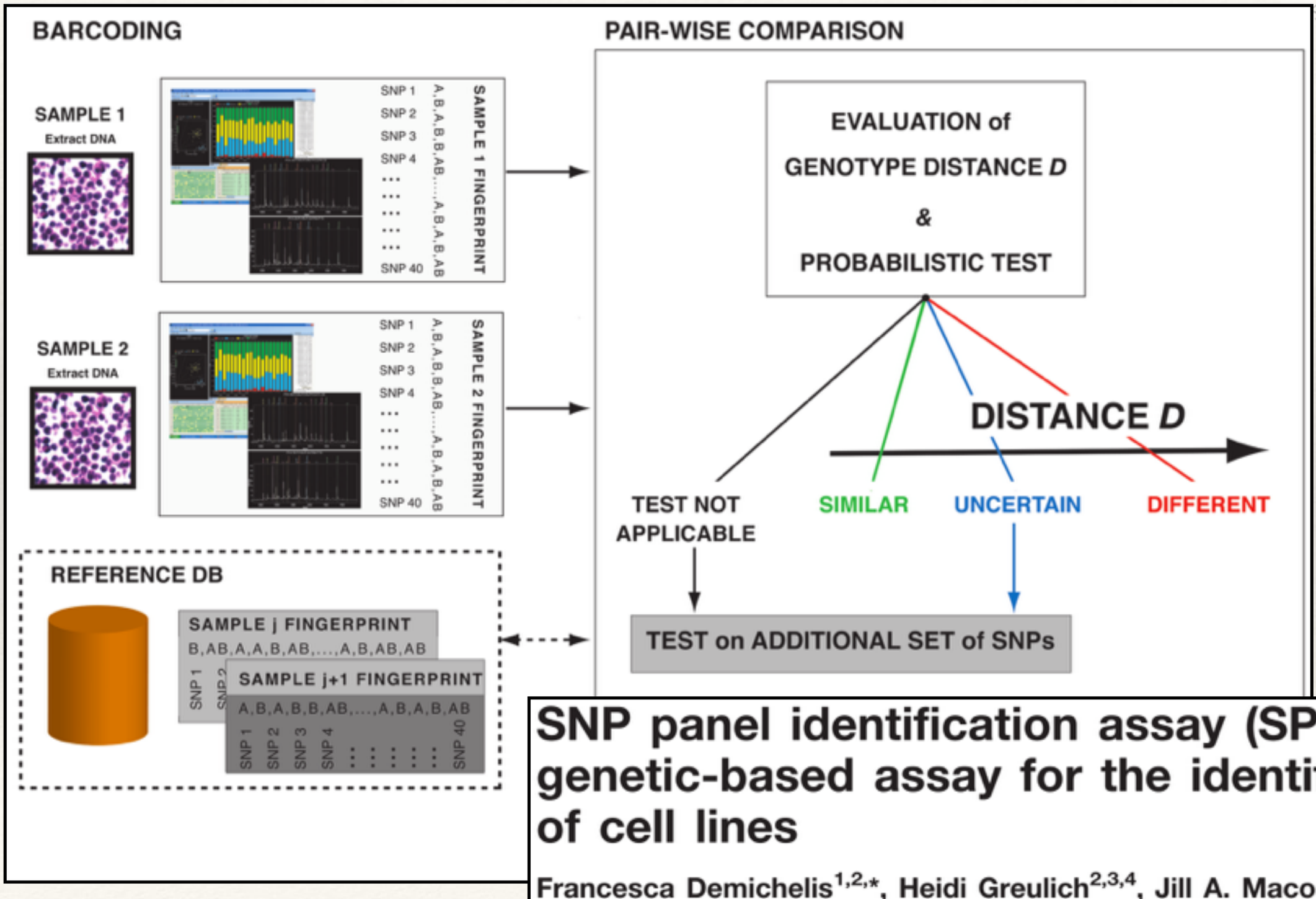
Creator: gabriel maldonado



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

Wet-Lab based

SNP-zygosity matching

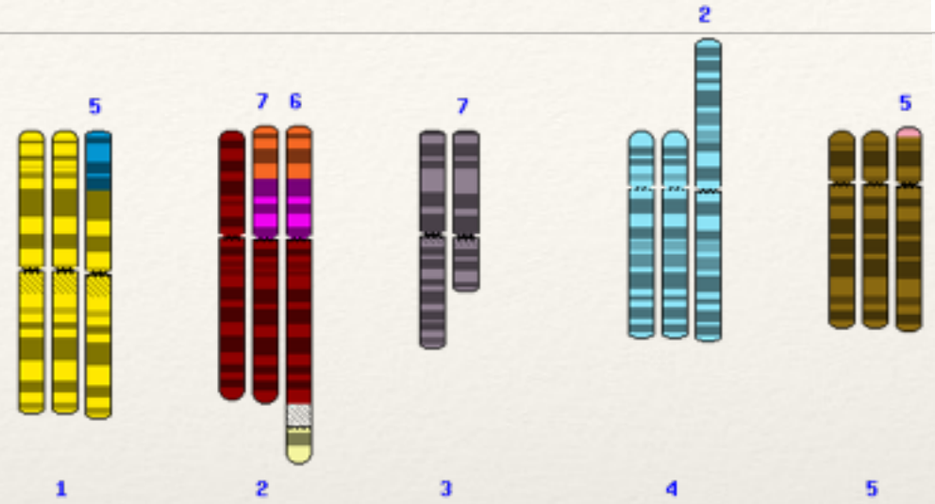


SNP panel identification assay (SPIA): a genetic-based assay for the identification of cell lines

Francesca Demichelis[1,2,*], Heidi Greulich[2,3,4], Jill A. Macoska[5],

Beyond SNP-Zygosity

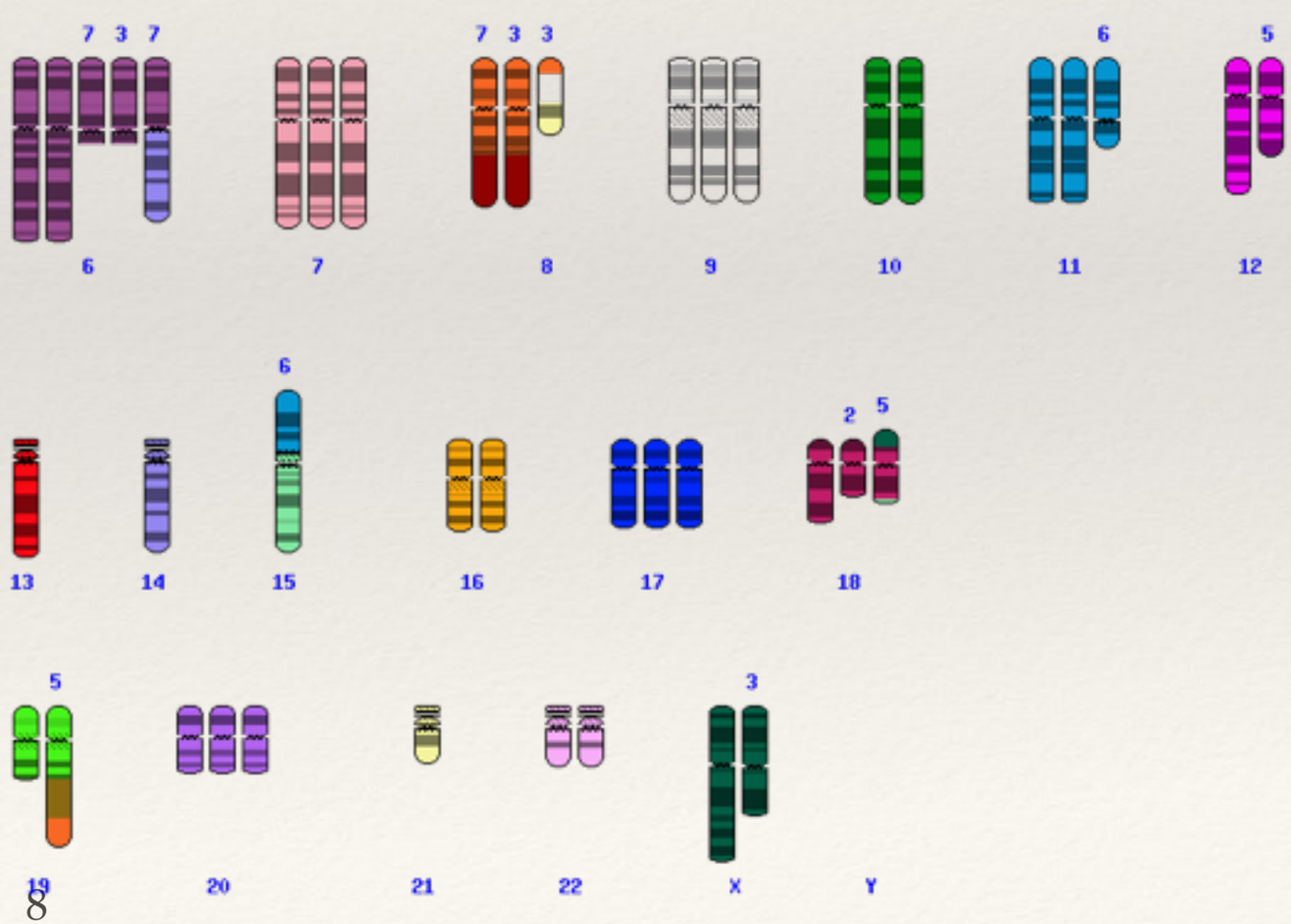
- SNP-zygosity matching problematic in Cancer NGS



- 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:
0/0/1:
1/1/1:
```

tri-ploid VCF, zygosity matching problematic



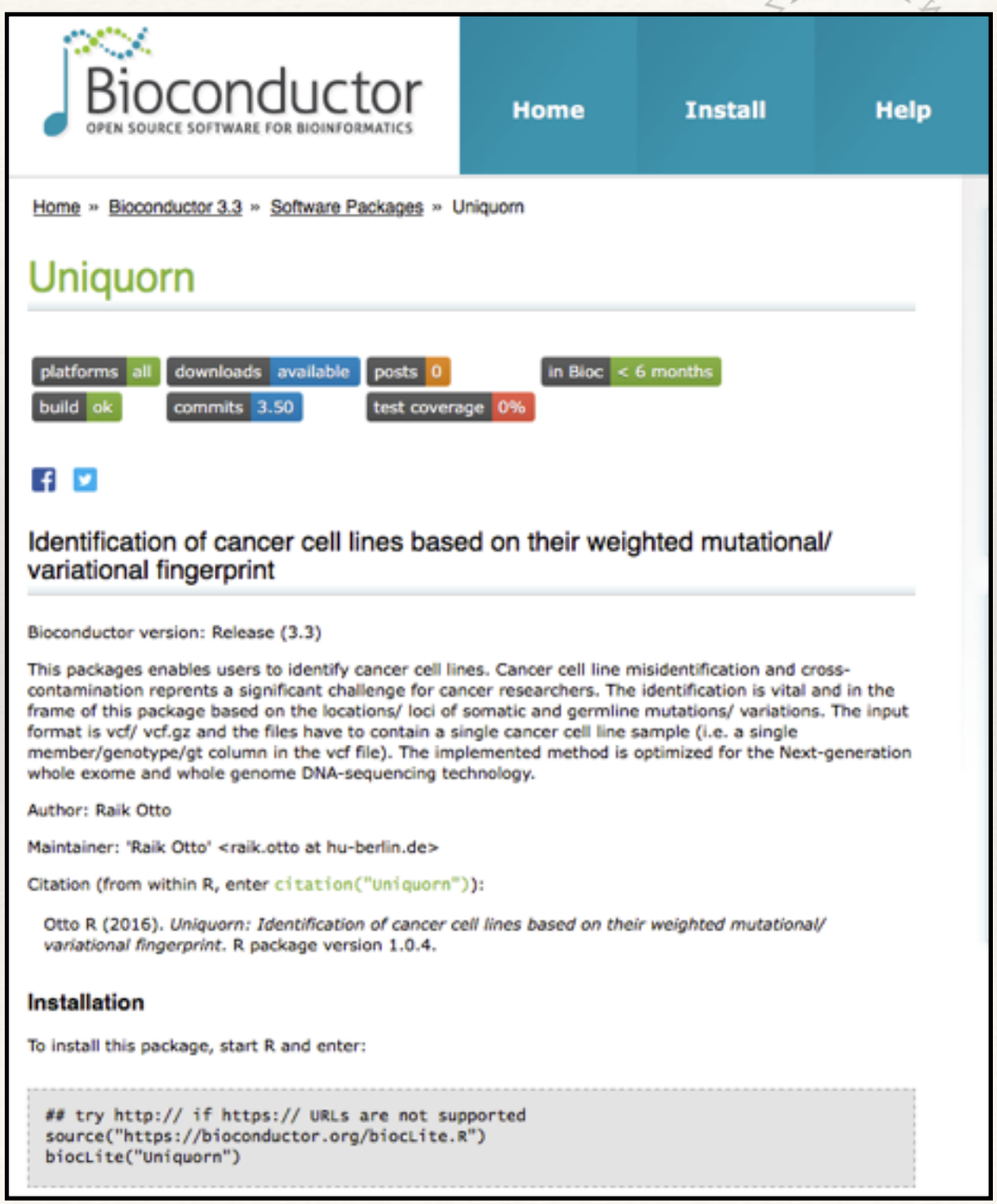
Uniquorn-Method

- Match (predominantly) **rare,**

somatic passenger mutations

- Start + Length

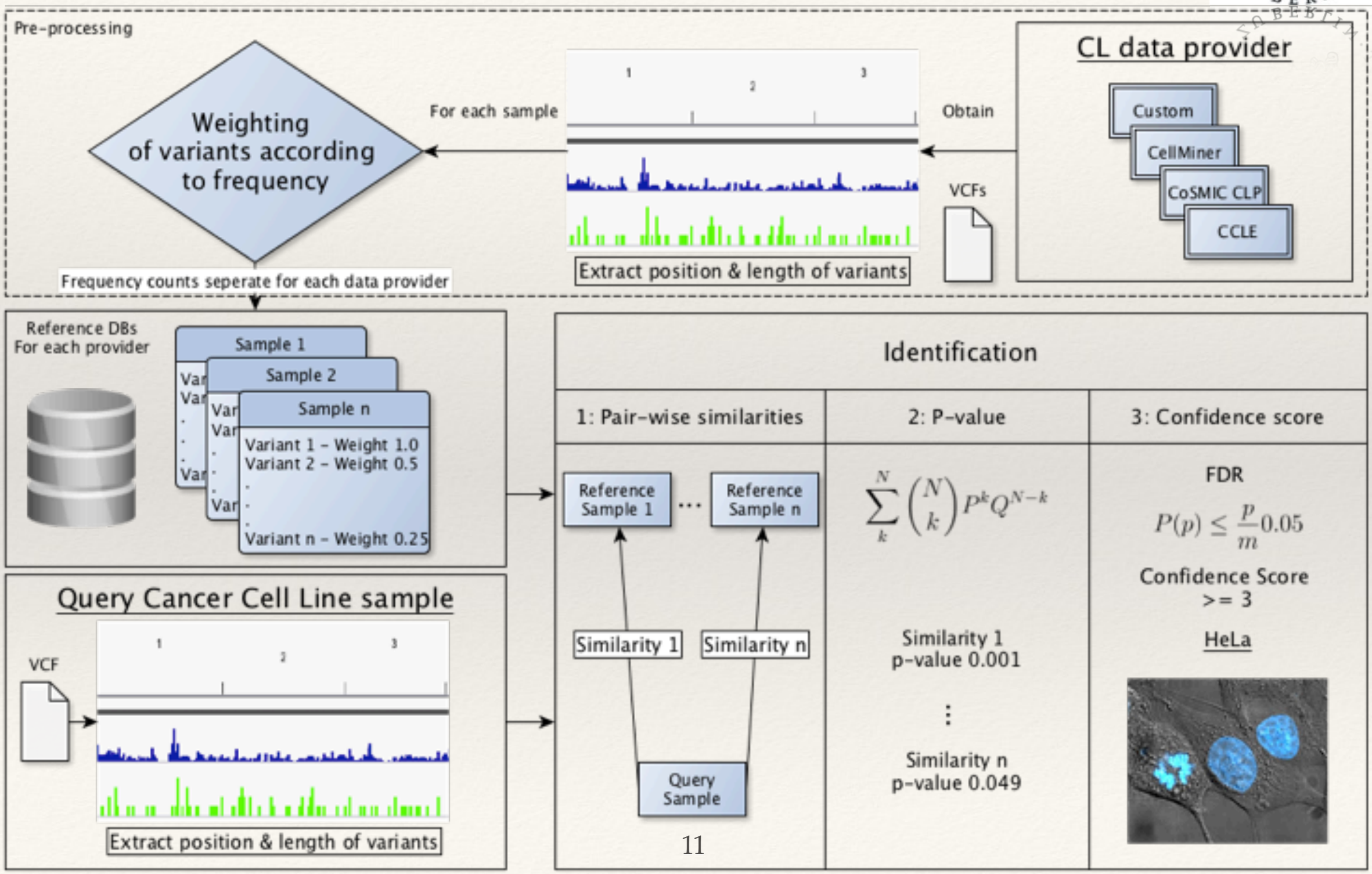
Drop SNP-Zygosity constraint



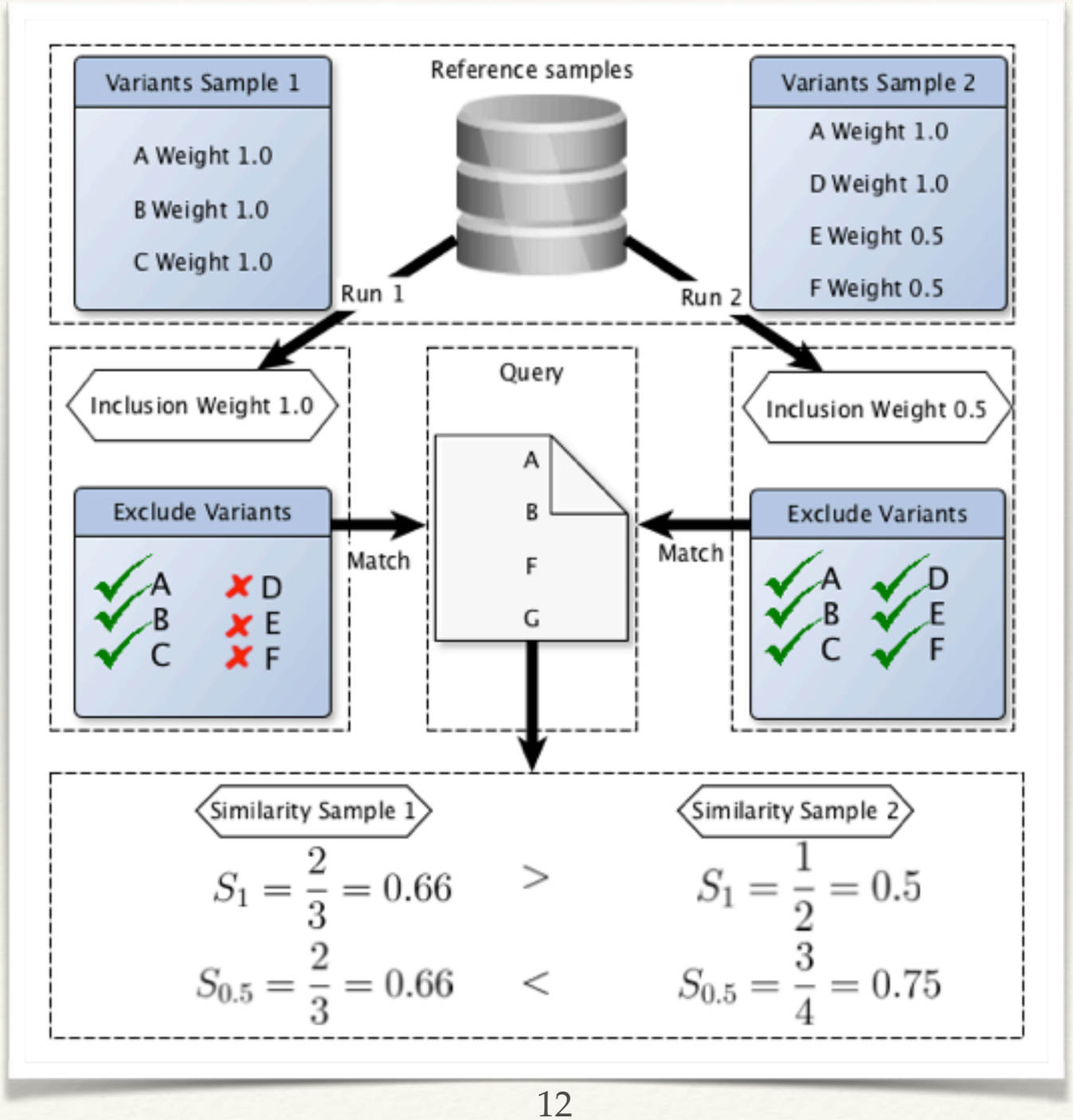
Comparison Uniquorn

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

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

Likelihood to commit statistical

error type 1 when rejecting H0

that dis-similarity is due to

different CL-identity

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

$$-Log_e(q - value) \coloneqq C_s$$

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

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

Take away message

Identify NGS CL data before usage

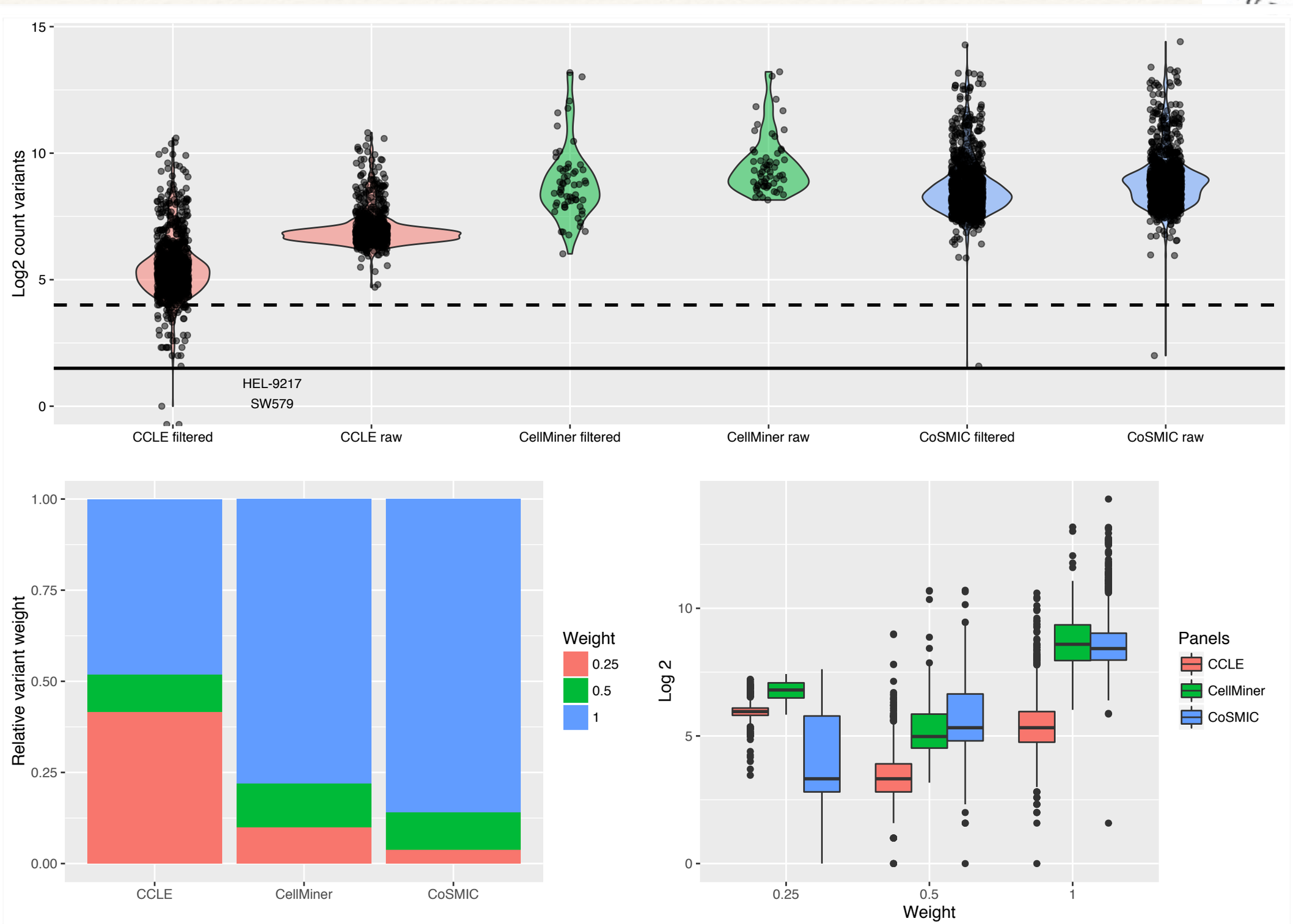
BiocLite(„*Uniquorn*“)

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)