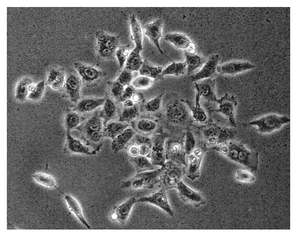
Robust In-Silico identification of sequenced Cancer Cell Lines

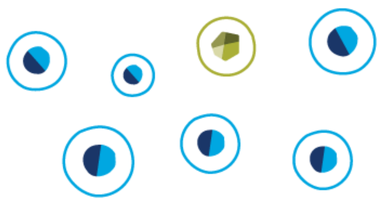
Raik Otto[1,*], Christine Sers[2], Ulf Leser[1]

[1] Knowledge Management in Bioinformatics, Institute for Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099, Berlin, Germany
[2] DKTK Deutsches Konsortium für Translationale Krebsforschung, Partner site Charite Berlin

Cell Line Confusion



Cell Line Culture



1 in 7 mis-identified

NGS Identification limited or impossible

Consequence: Incorrect research

Background

Cancer cell lines (CCL) are biological samples of great importance to cancer researchers world-wide but highly susceptible for the critical errors **misidentification** and **cross-contamination** [1].

Identity crisis

See associated Correspondence: Masters, *Nature* **492**, 186 (December 2012)

It is time for all involved to tackle the chronic scandal of cell-line contamination. Funders first.

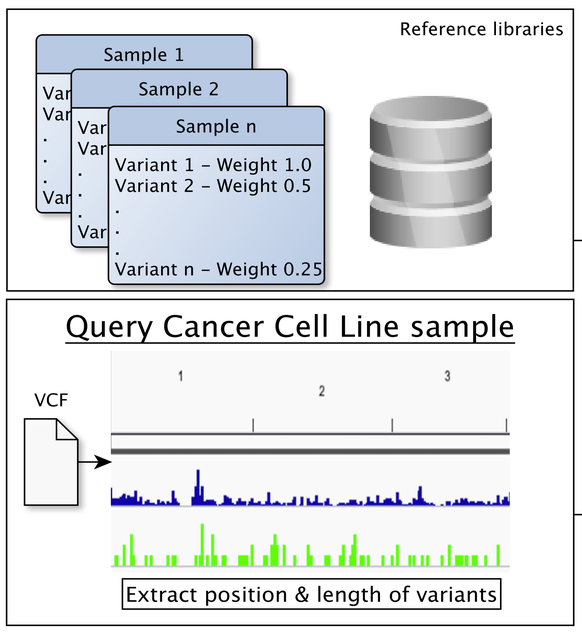
MDA-MB-435 cells are derived from M14 Melanoma cells––a loss for breast cancer, but a boon for melanoma research

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

Optimal means to identify CCLs are crucial for cancer-researchers. Established methods (e.g. STR and SPIA) exist [2,3], but are generally not applicable for heterogeneous, incomplete and contradictive Next-Gen Sequencing data. Here, we present the Uniquorn method that reliably identifies CCLs based on their genotyped variation profiles (VCF-files), agnostic to the underlying genotyping-procedure. Uniquorn was evaluated by cross-identifying 1989 CCLs that were obtained with different technologies, laboratories, algorithms and sequencing-approaches. Our method achieves a sensitivity of ~92% and specificity of ~99%. Uniquorn and ~2000 reference CCL samples are freely available as R-BioConductor package.

Everything was going along fine until they discovered their HeLa cell line expressed Y chromosome markers.

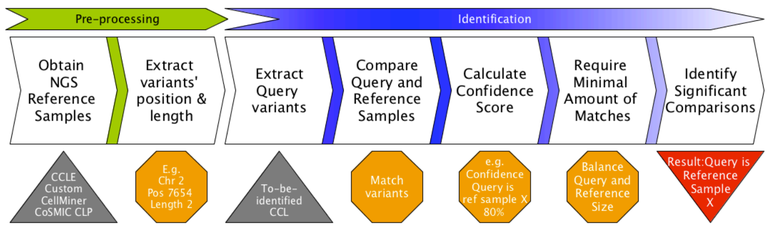
The Uniquorn Approach



Key concepts

1 Use optimal subset of variants for CCL identification (qualitative regulatization)

2 Penalize unspecific similarities (quantiative regularization)



- Uniquorn identifies CCLs by pair-wise comparison of the query to known reference samples
- If p-value and background-noise threshold are passed by one or many comparisons -> positive identification
- Reference variants weighted: weight represents a variant's association-strength to a specific CCL

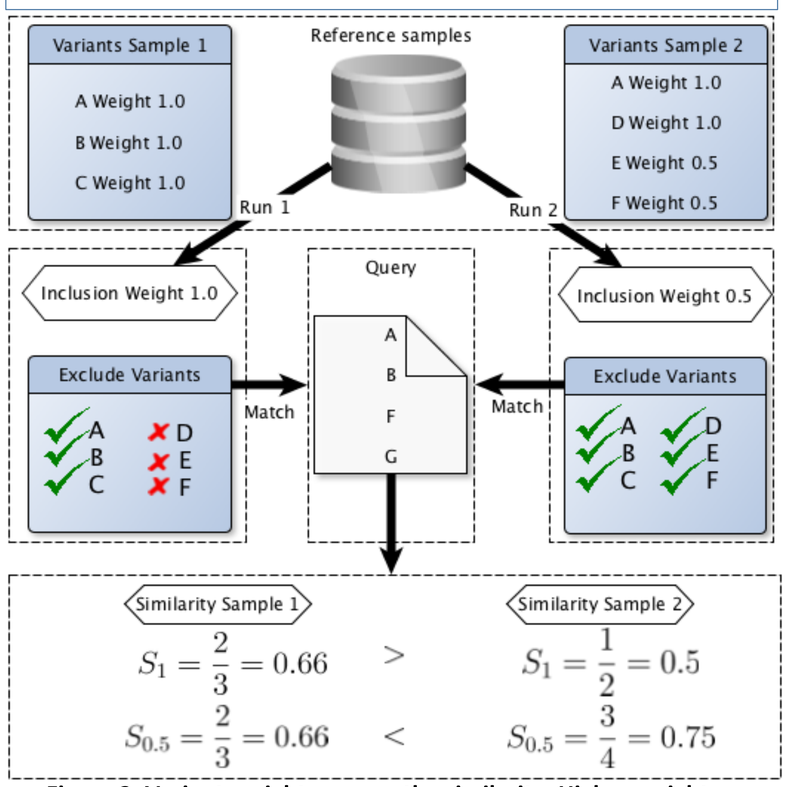


Figure 2: Variant-weights govern the similarity. Higher weights are prefereable but not always available.

COSMIC Cell Lines Project — Catalogue of somatic mutations in cancer
CCLE Cancer Cell Line Encyclopedia
CellMiner™

Sources for reference CCLs

Evaluation

Weight	1.0	0.5	0.25	0.0
Possible TP	3555			
TP	3255	3456	3456	2508
FN	300	99	99	1047
FP	18	38	55	1722
TN	~4 mil.			
Sens. %	92	97	97	71
Spec. %	99			
F1 %	95	98	98	64
PPV %	99			59

Table 1: Uniquorn's capacity to identify heterogeneous Cancer Cell Lines. 3555 CCL relationships of 1989 CCLs had to be found. Uniquorn found ~92% with the highest variant weights. Lower weights are useful for sub-optimal CCL reference libraries.

Summary

- Uniquorn is a novel algorithm to identify CCLs
- Optimized for Next-Generation sequenced CCLs in VCF-format
- Shows an identification-sensitivity of ~92% and specificity of 99%
- Works on all NGS DNA data
  - ✓ Whole-genome/exome, hybrid capture
  - ✓ Different algorithms, e.g.
    - Variant caller
    - Mapper
  - ✓ Sequencing technologies
- Requires reference samples, however, ~2000 can be freely obtained from
  - CCLE [4]
  - CellMiner project [5]
  - CoSMIC Cancer Cell Line Project [6]
- Available as R-BioConductor package *Uniquorn*
- Future development: RNA & Panel-seq

Contact

Raik Otto
Humboldt-Universität zu Berlin
Email: raik.otto@hu-berlin.de
Phone: 0049-30-2093-3086

References

1. Capes-Davis, A., Reid, Y.A., Kline, M.C., Storts, D.R., Strauss, E., Dirks, W.G., Drexler, H.G., MacLeod, R.A., Sykes, G., Kohara, A. *et al.* (2013) Match criteria for human cell line authentication: where do we draw the line? *Int J Cancer*, **132**, 2510-2519.
2. Demichelis, F., Greulich, H., Macoska, J.A., Beroukhim, R., Sellers, W.R., Garraway, L. and Rubin, M.A. (2008) SNP panel identification assay (SPIA): a genetic-based assay for the identification of cell lines. *Nucleic Acids Res*, **36**, 2446-2456.
3. Parson, W., Kirchebner, R., Muhlmann, R., Renner, K., Kofler, A., Schmidt, S. and Kofler, R. (2005) Cancer cell line identification by short tandem repeat profiling: power and limitations. *FASEB J*, **19**, 434-436.
4. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603-607.
5. Reinhold, W.C., Varma, S., Sousa, F., Sunshine, M., Abaan, O.D., Davis, S.R., Reinhold, S.W., Kohn, K.W., Morris, J., Meltzer, P.S. *et al.* (2014) NCI-60 whole exome sequencing and pharmacological CellMiner analyses. *PLoS One*, **9**, e101670.
6. Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*, **43**, D805-811.