# CELLector: Genomics Guided Selection of Cancer *in vitro* Models

**Hanna Najgebauer[1,2], Mi Yang[3], Hayley Francies[4], Euan A Stronach[1,5], Mathew J Garnett[1,4], Julio Saez-Rodriguez[1,2,3], Francesco Iorio[1,2,4,6,*]**

[1] Open Targets, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

[2] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SA, UK

[3] Faculty of Medicine, Joint Research Centre for Computational Biomedicine, RWTH Aachen University, Aachen 52057, Germany

[4] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge CB10 1SA, UK

[5] Target Sciences, GlaxoSmithKline, Stevenage, UK

[6] Lead Contact

[*] Correspondence: fi1@sanger.ac.uk

## Summary

The selection of appropriate cancer models is a key prerequisite for maximising translational potential and clinical relevance of *in vitro* studies. An important criterion for this selection is the molecular resemblance of available models to the primary disease they represent. While studies are being increasingly conducted to comprehensively compare genomic profiles of cell lines and matched primary tumours, there is no data-driven, robust and user-friendly tool assisting scientists in such selection, by adequately estimating the molecular heterogeneity of a primary disease that is captured by existing models. We developed *CELLector*: a computational tool implemented in an open source R Shiny application and R package that allows researchers to select the most relevant cancer cell lines in a genomic-guided fashion. CELLector combines methods from graph theory and market basket analysis; it leverages tumour genomics data to explore, rank, and select optimal cell line models in a user-friendly way, enabling scientists to make appropriate and informed choices about model inclusion/exclusion in retrospective analyses and future studies. Additionally, it allows the selection of models within user-defined contexts, for example, by focusing on genomic alterations occurring in biological pathways of interest or

considering only predetermined sub-cohorts of cancer patients. Finally, CELLector identifies combinations of molecular alterations underlying disease subtypes currently lacking representative cell lines, providing guidance for the future development of new cancer models. To demonstrate usefulness and applicability of our tool, we present example case studies, where it is used to select representative cell lines for user-defined populations of colorectal cancer patients of current clinical interest.

**Key worlds**

cell lines, cancer models, *in vitro* study, genomics, new algorithm, molecular subtyping, cancer heterogeneity, experimental design, patient cohort, representative *in vitro* model

## Introduction

The use of appropriate cancer *in vitro* models is one of the most important requirements for investigating cancer biology and successfully developing new anticancer therapies. Much effort has been devoted to evaluating the extent of phenotypic and genotypic similarities between existing cancer models and the primary tumours they aim to represent (Ahmed et al., 2013; Beaufort et al., 2014; Ince et al., 2015; Medico et al., 2015; Qiu et al., 2016). Despite inherent limitations, immortalised human cancer cell lines are the most commonly used experimental models in oncology research. Technological advancement in high-throughput 'omics' techniques and the availability of rich cancer genomics datasets, such as those provided by The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov), the International Cancer Genome Consortium (ICGC) (Zhang et al., 2011), the NCI-60 panel (Shoemaker, 2006), the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012), the Genomics of Drug Sensitivity in Cancer (GDSC) (Garnett et al., 2012; Iorio et al., 2016), the COSMIC Cell Line Project (Forbes et al., 2017) and many others, have transformed the way preclinical cancer models can be assessed and prioritized. To this end, several studies have proposed analytical methods to evaluate the suitability of

cell lines as tumour models (Domcke et al., 2013; Jiang et al., 2016; Mouradov et al., 2014; Sinha et al., 2017; Sun and Liu, 2015; Vincent et al., 2015; Zhao et al., 2017). Although these studies provide useful guidelines for choosing appropriate and avoiding poorly suited cell line models, they are restricted to individual cancer types. Most importantly, they require an expert knowledge of the genomic alterations known to have a specific functional role in the tumour (sub)type under consideration (Domcke et al., 2013; Jiang et al., 2016). As a consequence, there is a need for robust computational methods able to integrate the molecular characterisation of large cohorts of primary tumours from different tissues, extracting the most clinically relevant features in an unbiased way, and evaluating/selecting representative *in vitro* models on the domain of these features.

We have recently published a large molecular comparison of cancer cell lines and matched primary tumours at the sample population level (Iorio et al., 2016). Our results show that cell lines recapitulate most of the oncogenic alterations identified in matched primary tumours, and at similar frequencies. Building on our previous work, here we present CELLector, a tool for genomics-guided selection of cancer *in vitro* models. CELLector is based on an algorithm that combines methods from graph theory and market basket analysis (Han et al., 2012). It makes use of large-scale tumour genomics data to explore and rank patient subtypes based on genomic signatures (e.g. combinations of genomic alterations) identified in an unsupervised way and their prevalence. Subsequently, it ranks cell line models based on their genomic resemblance to the identified patient subtypes. Additionally, CELLector enables the identification of disease subtypes currently lacking representative *in vitro* models, which could be prioritised for future development. CELLector also implements interactive visualisations and intuitive explorations of results and underlying data, and it is available as open-source, user-friendly R Shiny application at
https://ot-cellector.shinyapps.io/cellector_app/
(code available at https://github.com/francescojm/CELLector_App)
and R package at https://github.com/francescojm/CELLector.

## Results

### Overview of CELLector

CELLector is implemented into two distinct modules. The first module recursively identifies the most frequently occurring sets of molecular alterations (signatures) in a cohort of primary tumours, by focusing on the set of clinically relevant genomic features that we previously published (Iorio et al., 2016). These encompass somatic mutations in 470 high-confidence cancer driver genes and copy number gains/losses of 425 recurrently altered chromosomal segments, and were identified by applying state-of-art computational tools (such as the intOGen pipeline (Gonzalez-Perez et al., 2013; Gundem et al., 2010) and ADMIRE (van Dyk et al., 2013)) to the genomic characterisation of a cohort of 11,289 cancer patients (from the TCGA (http://cancergenome.nih.gov), the ICGC (Zhang et al., 2011) and other publicly available studies). Based on the collective presence/absence of these alterations sets, CELLector partitions the primary tumours into distinct subpopulations (Figure 1A). The second module examines the identified molecular signatures in cancer cell lines in order to identify the best-representative models for each patient subpopulation (Figure 1B). This approach not only helps in maximising the covered disease heterogeneity but also enables the identification of molecular signatures underlying tumour subtypes currently lacking representative models (Figure 1C).

To demonstrate the power of CELLector and to allow an easy and immediate use of its functionalities, we have included in its implementation datasets from the genomic characterisation of tumours and cell lines derived from 16 different tissues (Table S1 and STAR Methods).

### CELLector modules

In the first module, CELLector assembles a search space in the form of a binary tree as follows. Starting from an initial cohort of patients affected by a given cancer type, the most frequent alteration or set of molecular alterations with the largest support (the

subpopulation of patients in which these alterations occur simultaneously) is identified using the *Eclat* algorithm (Zaki et al., 1997). Based on this, the cohort of patients is split into two subpopulations depending on the collective presence or absence of the identified alterations. This process is then executed recursively on the two resulting subpopulations and it continues until all the alteration sets (with a support of user-defined prevalence) are identified. Each of the alterations sets identified through this recursive process is stored in a tree node. Linking nodes identified in adjacent recursions yields a binary tree: *the CELLector search space*. Each individual path (from the root to a node) of this tree defines a rule (signature), represented as a logic AND of multiple terms (which can be also negated), one per each node in the path. If the genome of a given patient in the analysed cohort satisfies the rule then it is contained in the subpopulation represented by the terminal node of that path. Collectively, all the paths in the search space provide a representation of the spectrum of combinations of molecular alterations observed in a given cancer type, and their clinical prevalence in the analysed patient population (Figure 2A).

Subsequently, the CELLector search space is mined in the second module for:

- Exploring, and mapping cell lines to tree nodes (therefore to relevant patient subpopulations) based on the corresponding rules;
- Selecting the most representative set of cell lines maximising their covered genomic heterogeneity, via a guided visit of the search space (detailed in the STAR Methods);
- Identifying tumour subtypes lacking representative cell line models.

Finally, CELLector supports interactive visualisation and exploration of both the search space and final results (Figure 2).

**CELLector capabilities**

CELLector assists in the selection of the best-representative preclinical models to be employed in molecular oncology studies. It also enables molecular subtyping/classification of any disease cohort. As detailed in the previous section, one of the approaches that users

can pursue with CELLector is a simple *guided visit* of its search space (detailed in the STAR Methods) to select the optimal set of *n* cell lines to be included in a small-scale *in vitro* study or a low-throughput screen. The selected cell lines are picked from those mapped to the first *n* node, as they appear in the guided visit of the tree and, per construction, this guarantees that the coverage of the genomic heterogeneity of a particular cancer type is maximised by the selected cell lines.

Another approach is to use the position of a given cell line within the search space, based on the alteration sets that it harbours, as a mean to score its quality. This can be in fact estimated as a trade-off between the depth of the resulting node (proportional to the length of the corresponding genomic signature, in terms of considered genomic alterations) and the size of the corresponding patient subpopulation.

In addition, given that the choice of appropriate *in vitro* models often depends on the context of the study, users can restrict the analysis to a given sub-cohort of patients (while constructing the search space), determined a priori based on the presence/absence of a given genomic feature. For example, users can restrict the CELLector analysis to subset of tumours harbouring *TP53* mutations, or genomic alterations in the PI3K/Akt signalling pathway. In this case, only tumours characterised by these features are taken into consideration when building the CELLector search space (Supplementary Case Study 1 and next section). Notably, this can involve also other user-defined characteristics, for example the microsatellite instability (MSI) status of cancer cell lines. As a consequence, CELLector allows users to flexibly tailor the selection of cell lines in a context-dependent manner.

Finally, the CELLector R Shiny app provides additional functionalities enabling an interactive exploration of the tumour/cell line genomic features and final results. A tutorial demonstrating all these functionalities, example case studies, and a step-by-step guide to reproduce the reported results is provided as Supplemental Information.

**Case Study**

In this section, we present a practical example to demonstrate the usefulness of CELLector in an experimental study design. Detailed instructions on this and other case studies are provided in the user tutorial available as Supplemental Information.

In this example, we want to identify the most clinically relevant microsatellite instable cell lines that capture the genomic diversity of a sub-cohort of colorectal cancer patients that harbour *BRAF* mutations. The *BRAF* mutant colorectal cancer has a low prevalence (5%-8%) and very poor prognosis (Sanz-Garcia et al., 2017). The model selection should be guided by somatic mutations that are prevalent in at least 5% of the considered patient population (Figure 2A: box 1 and box 2).

**Building the CELLector search space**

After setting the CELLector app parameters to reflect the search criteria detailed in the previous section (Figure 2A: box 1 and box 2), the CELLector search space is assembled using a built-in dataset containing the genomic characterisation of a cohort of 517 colorectal cancer tumours (Table S1 and STAR Methods).

First, the cohort is reduced to the 86 tumours harbouring *BRAF* mutations (Figure 2A: node 1). CELLector then identifies 3 major molecular subpopulations characterised, respectively, by *APC* mutations (Figure 2A: node 2), *FBXW7* mutations (Figure 2A: node 3), and *PIK3CA* mutations (Figure 2A: node 10), collectively representing 85 % of the studied *BRAF* mutant cohort (n = 50 + 14 + 9 = 73). The remaining 15% (n = 13) of *BRAF* mutant tumours do not fall into any of the identified molecular subpopulations, i.e. they do not harbour *APC, FBXW7* nor *PIK3CA* mutations; Figure 2A).

The largest molecular subpopulation (58.14%, n = 50, harbouring *BRAF* and *APC* mutations) is assigned to the root of the search space (Figure 2A: node 2, in purple). The second largest subpopulation (16.28%, n = 14) is characterised by the co-occurrence of

*BRAF* and *FBXW7* mutations in the absence of *APC* mutations (Figure 2A: node 3, in magenta), and the third largest subpopulation (10.47%, n = 9) harbours the *BRAF* and *PIK3CA* mutations in the absence of both *APC* and *FBXW7* mutations (Figure S2A: node 10, in cyan). At this point, each identified tumour subpopulation is further refined based on the prevalence of remaining set of alterations (STAR Methods). This process runs recursively and stops when all alteration sets with a user-determined prevalence (in this case 5%, Figure 2A: box 1) are identified. In this study case, a total number of 10 distinct tumour subpopulations with corresponding genomic signatures are identified (Figure 2). Notably, some of the mutational signatures identified in the *BRAF* mutated tumours are linked to differential prognosis in colorectal tumour stratification (Schell et al., 2016).

**Selection of representative *in vitro* models**

The CELLector search space generated in the previous section is next translated into a Cell Line Map table (Figure 2B), indicating the order in which cancer *in vitro* models mirroring the identified genomic signatures should be selected, accounting for tumour subpopulations currently lacking representative *in vitro* models. This selection order is defined by a guided visit of the CELLector search space (detailed in the STAR Methods), aiming at maximising the heterogeneity observed in the studied primary tumours. The Cell Map table uncovers the complete set of molecular alterations (e.g. genomic signatures) defining each tumour subpopulation. For example, the least prevalent *BRAF* mutant colorectal tumour subpopulation (node 8, 9.30% of tumours) is characterised by the co-occurrence of *BRAF*, *APC, PIK3CA, PTEN, TP53 and KRAS* mutations; this genomic signature is not mirrored by any of the available microsatellite instable colorectal cancer models included in the built-in dataset (Figure 2B).

In this example, we wanted to select only microsatellite instable cell lines (16 out of 51 available in CELLector, Table S1). As cell lines are derived from tumours at various levels of differentiation and stages of development, out of these 16 considered models only

4 mirror the genomic signatures identified in the primary disease (i.e. they are mapped to nodes of the CELLector search space). Effectively, this means that these 4 microsatellite instable cell lines are good representatives of the examined patient cohort accounting for tumour subpopulations of different sizes.

Finally, the representative cell lines are picked from each of the molecular tumour subpopulations (as detailed in the STAR Methods) starting from the most prevalent one. A possible choice of *in vitro* models that best represent the genomic diversity of the studied tumour cohort include: LS-411N, SNU-C5, RKO and KM12 (Figure 2B).
Additional case studies are included in the user tutorial provided as Supplemental Information.

## Discussion

The translational potential of preclinical studies is highly dependent on the clinical relevance of the employed *in vitro* models. Good models are required to capture the genomic heterogeneity of a cancer type under investigation and/or accurately represent alterations in relevant biological pathways.

We present CELLector, a tool that allows scientists to select the most representative set of cell line models, maximising the covered genomic heterogeneity of the disease under consideration. The overall aim of the CELLector algorithm is to globally assess the quality of cancer *in vitro* models in terms of their similarity to genomic subtypes detected in matched primary tumours, and to make available to the research community a user-controlled environment to perform such a task.

A key strength of CELLector is its generality: the algorithm can be applied to any disease for which *in vitro* models and matching primary/model genomic data are available. CELLector enables the systematic identification of recurrent tumour subtypes with paired

genomic signatures, and selection of *in vitro* models based on the recurrence of these signatures. In addition, the algorithm identifies disease subtypes currently lacking representative models enabling prioritisation of new model development. To the best of our knowledge, CELLector represents the first computational method that ranks and selects cancer *in vitro* models, in a data driven way, across different cancer types, and without the need for expert knowledge about the primary disease under consideration. However, the model selection performed by CELLector can be flexibly tailored to fit the context of a study.

Clinically relevant disease subtyping takes time and multiple resources. In recent years, an increasing number of studies have taken advantage of the availability of rich genomics/transcriptomics data for systematic molecular subclassification of tumours across tissues (Dawson et al., 2013; Guinney et al., 2015). Based on similar principles, CELLector can serve as a valuable tool to aid designing experimental studies minimising the risk of clinically relevant signal being missed due to 'noise' contributed by inclusion of less relevant models, or conversely identification of false positives due to strong signals from poor quality models. Addressing both of these issues will have direct and immediate implications on the quality of future *in vitro* experiments and in analysis of retrospective data derived from cancer cell lines.

## Author Contributions

H.N. contributed to algorithmic design, curated data, implemented, managed and documented the R package, worked on the R Shiny App development and wrote the manuscript; M.Y. implemented the first core function of the R package; H.F. contributed to manuscript editing/revising; E.S., J.S.R., M.J.G. revised the manuscript and contributed to the supervision of the study; F.I. conceived and designed the algorithm, implemented the R Shiny App, wrote the manuscript, and supervised the study. All authors read and approved the final manuscript.

## Acknowledgments

This work is funded by Open Targets project grant OTAR041. We thank Ian Dunham and Anneliese Speak for critically reading the manuscript and providing useful feedback. We thank Fiona Behan, Patricia Jaaks and Evangelia Petsalaki for testing our tool. FI thanks Giorgia Iorio for her insightful comments on the visualisation methods used by our software.

## References

Ahmed, D., Eide, P.W., Eilertsen, I.A., Danielsen, S.A., Eknæs, M., Hektoen, M., Lind, G.E., and Lothe, R.A. (2013). Epigenetic and genetic features of 24 colon cancer cell lines. Oncogenesis *2*, e71.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D.*, et al.* (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603.

Beaufort, C.M., Helmijr, J.C.A., Piskorz, A.M., Hoogstraat, M., Ruigrok-Ritstier, K., Besselink, N., Murtaza, M., van Ijcken, W.F.J., Heine, A.A.J., Smid, M.*, et al.* (2014). Ovarian Cancer Cell Line Panel (OCCP): Clinical Importance of In Vitro Morphological Subtypes. PLOS ONE *9*, e103988.

Dawson, S.J., Rueda, O.M., Aparicio, S., and Caldas, C. (2013). A new genome- driven integrated classification of breast cancer and its implications. The EMBO Journal *32*, 617.

Domcke, S., Sinha, R., Levine, D.A., Sander, C., and Schultz, N. (2013). Evaluating cell lines as tumour models by comparison of genomic profiles. Nature Communications *4*, 2126.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L.*, et al.* (2017). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Research *45*, D777-D783.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J.*, et al.* (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570.

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M.P., Jene-Sanz, A., Santos, A., and Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. Nature Methods *10*, 1081.

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P.*, et al.* (2015). The consensus molecular subtypes of colorectal cancer. Nature Medicine *21*, 1350.

Gundem, G., Perez-Llamas, C., Jene-Sanz, A., Kedzierska, A., Islam, A., Deu-Pons, J., Furney, S.J., and Lopez-Bigas, N. (2010). IntOGen: integration and data mining of multidimensional oncogenomic data. Nature Methods *7*, 92.

Han, J., Kamber, M., and Pei, J. (2012). Data Mining: Concepts and Techniques (The Morgan Kaufmann).

Ince, T.A., Sousa, A.D., Jones, M.A., Harrell, J.C., Agoston, E.S., Krohn, M., Selfors, L.M., Liu, W., Chen, K., Yong, M.*, et al.* (2015). Characterization of twenty-five ovarian tumour cell lines that phenocopy primary tumours. Nature Communications *6*, 7419.

Iorio, F., Knijnenburg, Theo A., Vis, Daniel J., Bignell, Graham R., Menden, Michael P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H.*, et al.* (2016). A Landscape of Pharmacogenomic Interactions in Cancer. Cell *166*, 740-754.

Jiang, G., Zhang, S., Yazdanparast, A., Li, M., Pawar, A.V., Liu, Y., Inavolu, S.M., and Cheng, L. (2016). Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. BMC Genomics *17*, 525.

Medico, E., Russo, M., Picco, G., Cancelliere, C., Valtorta, E., Corti, G., Buscarino, M., Isella, C., Lamba, S., Martinoglio, B.*, et al.* (2015). The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. Nature Communications *6*, 7002.

Mouradov, D., Sloggett, C., Jorissen, R.N., Love, C.G., Li, S., Burgess, A.W., Arango, D., Strausberg, R.L., Buchanan, D., Wormald, S.*, et al.* (2014). Colorectal Cancer Cell Lines Are Representative Models of the Main Molecular Subtypes of Primary Cancer. Cancer Research *74*, 3238.

Qiu, Z., Zou, K., Zhuang, L., Qin, J., Li, H., Li, C., Zhang, Z., Chen, X., Cen, J., Meng, Z.*, et al.* (2016). Hepatocellular carcinoma cell lines retain the genomic and transcriptomic landscapes of primary human cancers. Scientific Reports *6*, 27411.

Sanz-Garcia, E., Argiles, G., Elez, E., and Tabernero, J. (2017). BRAF mutant colorectal cancer: prognosis, treatment, and new perspectives. Annals of Oncology *28*, 2648-2657.

Schell, M.J., Yang, M., Teer, J.K., Lo, F.Y., Madan, A., Coppola, D., Monteiro, A.N.A., Nebozhyn, M.V., Yue, B., Loboda, A.*, et al.* (2016). A multigene mutation classification of 468 colorectal cancers reveals a prognostic role for APC. Nature Communications *7*, 11743.
Shoemaker, R.H. (2006). The NCI60 human tumour cell line anticancer drug screen. Nature Reviews Cancer *6*, 813.

Sinha, R., Winer, A.G., Chevinsky, M., Jakubowski, C., Chen, Y.-B., Dong, Y., Tickoo, S.K., Reuter, V.E., Russo, P., Coleman, J.A.*, et al.* (2017). Analysis of renal cancer cell lines from two major resources enables genomics-guided cell line selection. Nature Communications *8*, 15165.

Sun, Y., and Liu, Q. (2015). Deciphering the Correlation between Breast Tumor Samples and Cell Lines by Integrating Copy Number Changes and Gene Expression Profiles. BioMed Research International *2015*, 11.

van Dyk, E., Reinders, M.J.T., and Wessels, L.F.A. (2013). A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control. Nucleic Acids Research *41*, e100-e100.

Vincent, K.M., Findlay, S.D., and Postovit, L.M. (2015). Assessing breast cancer cell lines as tumour models by comparison of mRNA expression profiles. Breast Cancer Research *17*, 114.

Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B.*, et al.* (2011). International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. Database *2011*, bar026-bar026.

Zhao, N., Liu, Y., Wei, Y., Yan, Z., Zhang, Q., Wu, C., Chang, Z., and Xu, Y. (2017). Optimization of cell lines as tumour models by integrating multi-omics data. Briefings in Bioinformatics *18*, 515-529.

## Figure legends

**Figure 1 - Schematic representation of the CELLector modules.**

**A.** Primary tumours genomic features are used to identify tumours molecular subpopulations within a cohort of patients. **B.** The resulting CELLector search space is then used to map molecular similarities between the identified tumour subpopulations and cell line models. **C.** CELLector returns a list of cell line models that best represent the identified tumour subpopulations, thus maximising the coverage of disease heterogeneity, and highlighting tumour subtypes currently lacking representative *in vitro* models.

**Figure 2 CELLector in use: results from an example case study.**

**A.** Visual representation of the CELLector search space constructed based on the prevalence of co-occurring mutations in the *BRAF* mutant colorectal (COREAD) tumours (box 1 and box 2). Each node of the binary tree (top) represents a tumour subpopulation with define genomic signature. The prevalence of the identified signatures, and their hierarchical co-occurrence is represented by the sunburst (below). Each segment of the sunburst corresponds to a node in the three and is color-coded accordingly. **B.** Cell Line Map table including microsatellite instable cell lines mirroring the genomic signatures of the *BRAF* mutant COREAD tumour subpopulations identified in the CELLector search space. The models in green represent a possible choice of *n*-user-defined cell lines that could be selected in the presented case study.

## STAR Methods

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Deposited Data | | |
| Binary event matrices with status of high-confidence cancer genes (CGs) across primary tumours (COSMIC filtered variants) | Iorio *et al.*, 2016 | http://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources///Data/BEMs/PrimaryTumours/PrimTum_CG_BEMs/PrimTum_CG_BEMs_cf.zip |
| Binary event matrices with CNA status of recurrently altered chromosomal segments (RACSs) across primary tumours | Iorio *et al.*, 2016 | http://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources///Data/BEMs/PrimaryTumours/PrimTum_CNV_BEMs.zip |
| Binary event matrices with status of high-confidence cancer genes (CGs) across cell lines | Iorio *et al.*, 2016 | http://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources///Data/BEMs/CellLines/CellLines_CG_BEMs.zip |
| Binary event matrices with CNA status of recurrently altered chromosomal segments (RACSs) across cell lines | Iorio *et al.*, 2016 | http://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources///Data/BEMs/CellLines/CellLines_CNV_BEMs.zip |
| Software and Algorithms | | |
| R version 3.4.0 | R Foundation for Statistical Computing | https://www.r-project.org/; RRID: SCR_001905 |
| CELLector package | This paper | https://github.com/francescojm/CELLector |

### CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Francesco Iorio (fi1@sanger.ac.uk).

### METHOD DETAILS

#### Implementation

The CELLector algorithm and interactive visualisation tools are implemented in R and available as an open-source R package (code available at https://github.com/francescojm/CELLector, interactive vignette available at http://rpubs.com/francescojm/CELLector, user manual available at https://github.com/francescojm/CELLector/blob/master/CELLector.pdf) and R Shiny web

application (deployed at https://ot-cellector.shinyapps.io/cellector_app/, code available at https://github.com/francescojm/CELLector_App).

**Genomics data**

CELLector provides built-in genomics data for disease-matched primary tumours and cell lines derived from 16 cancer types, encompassing the characterisation of 4,550 tumours and 499 immortalised and commercially available cancer cell lines (Table S1), and accounting for somatic mutations and copy number alterations for high-confidence cancer genes and recurrently altered chromosomal segments, i.e. cancer functional events (CFEs). These CFEs are described in Iorio *et al.*, 2016 and corresponding data were obtained from the accompanied web-portal (http://www.cancerrxgene.org/gdsc1000/).

**QUANTIFICATION AND STATISTICAL ANALYSIS: The CELLector algorithm**

The CELLector algorithm combines methods from graph theory and market basket analysis. In the analytical framework of CELLector, the genomic background of a cohort of patients affected by a given cancer type is represented as a binary tree whose topology is determined by the most-frequently observed combinations of molecular alterations (item-sets) and their supports, i.e. the fraction of patients in which these alterations occur simultaneously. This tree is built recursively by sequential applications of the *Eclat* algorithm (Zaki et al., 1997) as follows. The tree construction starts from the root, modelling the combination of genomic alterations (item-set) with the largest support across the entire cohort. Then two sibling nodes are included, modelling the item-sets with the greatest supports when considering the population supporting the item-set of the parent node (right sibling node) and its complementary population (left sibling node). This is recursively performed at each new node included in the tree if the corresponding modelled item-set is supported by at least a user-defined ratio of patients in the considered patient subpopulation (for example 5%).

Subsequently, a logic AND formula *F* is assigned to each node *x*, considering the path to *x* from the root of the three. For each node *n* on this path (including the terminal ones) the corresponding modelled item-set is added to *F* as a term, negated if *n* is a left sibling (complement) node. Finally, a given cell line in the built-in collection is mapped to a node *n* if its genomic background satisfies *F(n).*

The algorithm continues with a guided deep-first-visit of the obtained tree, which return all the identified subtypes as a sorted list, as detailed in the following pseudo code:

**Variables and initial settings**

```
Q = an empty queue
T = a CELLector searching space
r = the root of T
U = a set of nodes that have not been visited yet
Idx = a queue index
CurrentNode = r
Idx = -1
U = all the nodes of T
```

**Algorithm**

```
While U is not empty
        remove CurrentNode from U
        While CurrentNode as a left child
                Add CurrentNode to the queue
                CurrentNode = left child of CurrentNode
        end
        Add the right children of all the nodes in Q to Q (by level) and remove them from U
        If there are right nodes in Q in position > Idx then
                Advance Idx to the first right node in Q in a position > Idx
                CurrentNode = Q[Idx]
end
Return Q
```

Finally, a Cell Line map is built by considering all subtypes (nodes) as they appear in *Q*, with corresponding signatures and mapped cell lines.

*N* Cell lines are selected among those appearing in the first *N* entries of this map with a heuristic method, minimizing the number of nodes each selected cell line is mapped onto.

## DATA AND SOFTWARE AVAILABILITY

Source code for the R package and the R Shiny app is publicly available on GitHub:

Package code https://github.com/francescojm/CELLector

Package interactive vignette http://rpubs.com/francescojm/CELLector

Package user manual https://github.com/francescojm/CELLector/blob/master/CELLector.pdf,

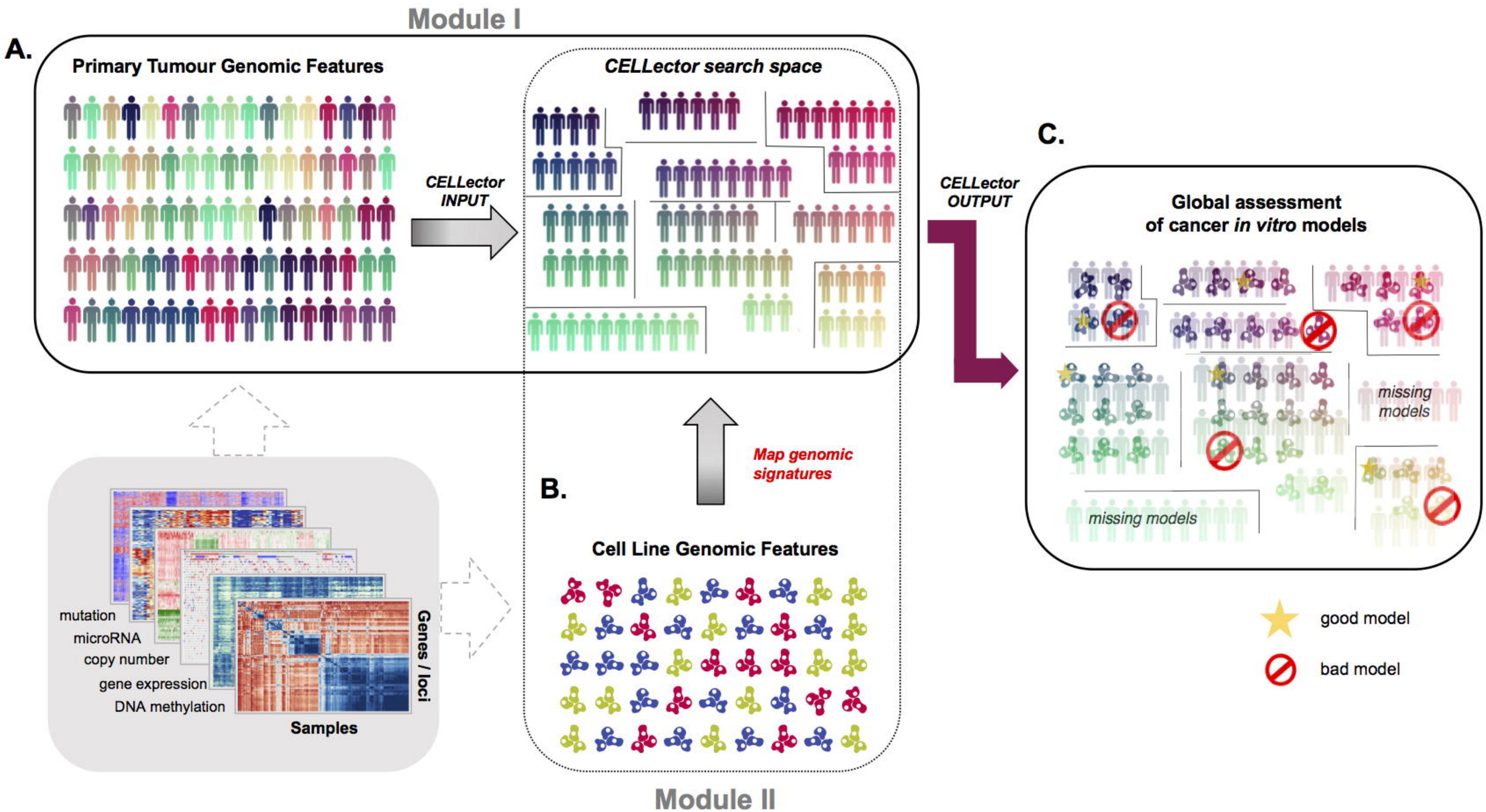
App deployed at https://ot-cellector.shinyapps.io/cellector_app/

App code available at https://github.com/francescojm/CELLector_App.

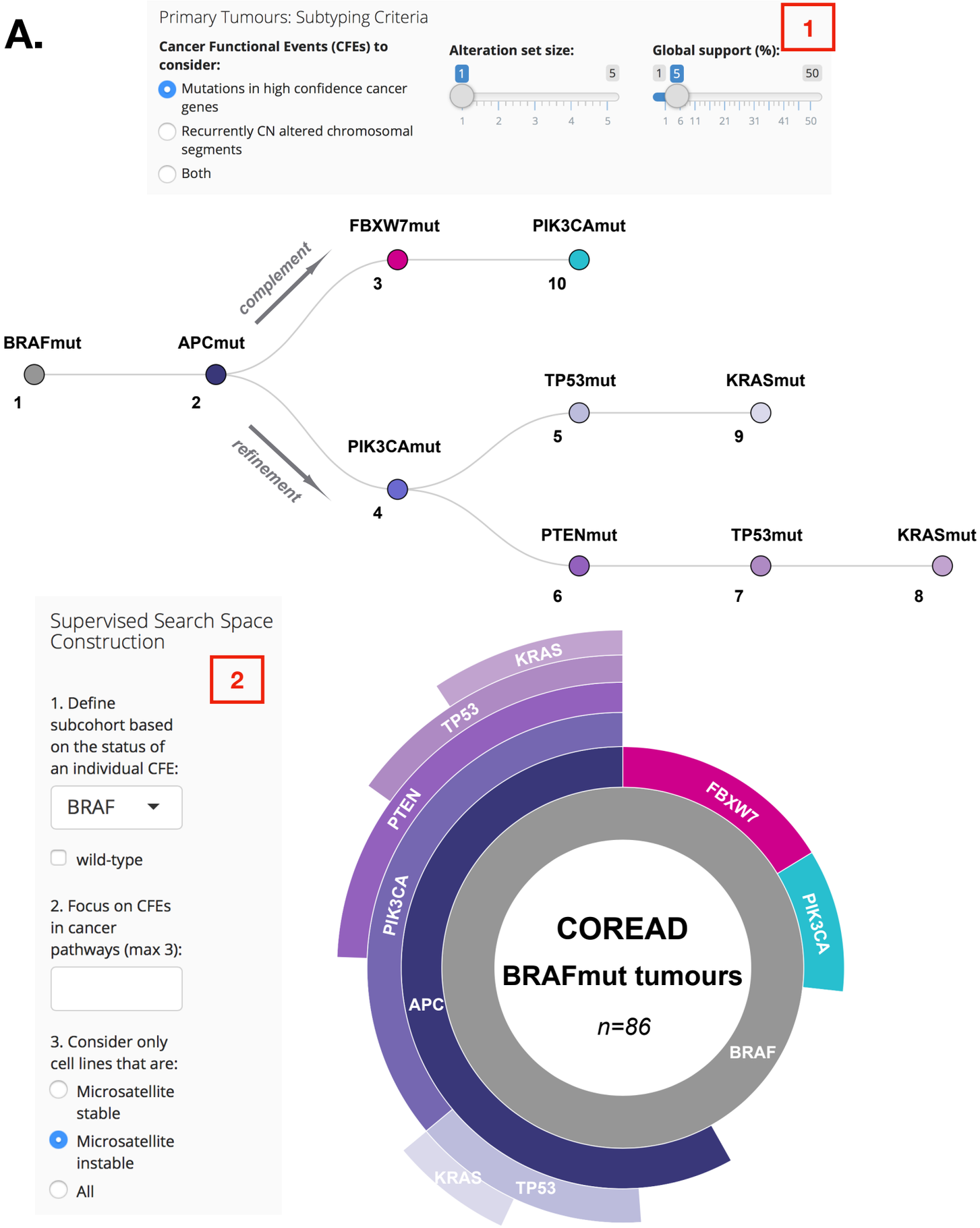
Detailed instructions on how to install the R package, run the CELLector analysis and interactively explore the results are also provided in the GitHub repository and in the supplemental information, together with a tutorial with instructions how to set up the analysis, interactively explore the results and execute CELLector on example case studies.

## Supplemental Information

Supplemental Information includes one table, two case studies, instructions and user tutorial demonstrating the full functionality of CELLector app.


Module I
A.
Primary Tumour Genomic Features
CELLector search space
CELLector INPUT
CELLector OUTPUT
C.
Global assessment of cancer in vitro models
missing models
missing models
good model
bad model
B.
Map genomic signatures
Cell Line Genomic Features
mutation
microRNA
copy number
gene expression
DNA methylation
Genes / loci
Samples
Module II

**A.**




**B.**

~grey - absence
**bold** - presence

***SELECT n=4***

| Sub Type | Genomic Signature | n Patients | % | Representative Cell Lines |
|---|---|---|---|---|
| 1 | **BRAFmut** | 86 | 100.00 | KM12, **LS-411N**, RKO, SNU-C5 |
| 2 | **BRAFmut APCmut** | 50 | 58.14 | KM12, LS-411N, **SNU-C5** |
| 3 | **BRAFmut** ~APCmut **FBXW7mut** | 14 | 16.28 | *lack of microsatellite instable in vitro models* |
| 10 | **BRAFmut** ~APCmut ~FBXW7mut **PIK3CAmut** | 9 | 10.47 | **RKO** |
| 4 | **BRAFmut APCmut PIK3CAmut** | 31 | 36.05 | SNU-C5 |
| 5 | **BRAFmut APCmut** ~PIK3CAmut **TP53mut** | 13 | 15.12 | **KM12**, LS-411N |
| 6 | **BRAFmut APCmut PIK3CAmut PTENmut** | 21 | 24.42 | *lack of microsatellite instable in vitro models* |
| 9 | **BRAFmut APCmut** ~PIK3CAmut **TP53mut KRASmut** | 6 | 6.98 | *lack of microsatellite instable in vitro models* |
| 7 | **BRAFmut APCmut PIK3CAmut PTENmut TP53mut** | 13 | 15.12 | *lack of microsatellite instable in vitro models* |
| 8 | **BRAFmut APCmut PIK3CAmut PTENmut TP53mut KRASmut** | 8 | 9.30 | *lack of microsatellite instable in vitro models* |