Big data and single cell transcriptomics: implications for ontological representation

Brian D. Aevermann[1], Mark Novotny[1], Trygve Bakken[2], Jeremy A. Miller[2], Alexander D. Diehl[3], David Osumi-Sutherland[4], Roger S. Lasken[1], Ed S. Lein[2], Richard H. Scheuermann[1,5]

[1]J. Craig Venter Institute, La Jolla, CA, USA

[2]Allen Institute for Brain Science, Seattle, WA, USA

[3]Department of Biomedical Informatics, University at Buffalo, Buffalo, NY, USA

[4]European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom

[5]Department of Pathology, University of California San Diego, La Jolla, CA, USA

Abstract (250 words) – Cells are fundamental functional units of multicellular organisms, with different cell types playing distinct physiological roles in the body. The recent advent of single cell transcriptional profiling using RNA sequencing is producing "big data", enabling the identification of novel human cell types at an unprecedented rate. In this review, we summarize recent work characterizing cell types in the human central nervous and immune systems using single cell and single nuclei RNA sequencing, and discuss the implications that these discoveries are having on the representation of cell types in the reference Cell Ontology (CL). We propose a method based on random forest machine learning for identifying sets of necessary and sufficient marker genes that can be used to assemble consistent and reproducible cell type definitions for

incorporation into the CL. The representation of defined cell type classes and their relationships in the CL using this strategy will make the cell type classes findable, accessible, interoperable, and reusable (FAIR), allowing the CL to serve as a reference knowledgebase of information about the role that distinct cellular phenotypes play in human health and disease.

Introduction – Cells are probably the most important fundamental functional units of multicellular organisms, since different cell types play different physiological roles in the body. Although every cell of an individual organism has essentially the same genome structure, different cells realize diverse functions due to differences in their *expressed* genome. In many cases, abnormalities in gene expression form the physical basis of disease dispositions. Thus, understanding and representing normal and abnormal cellular phenotypes can lead to the development of biomarkers for diagnosing disease and the identification of critical targets for therapeutic interventions.

Previous approaches used to characterize cell phenotypes have several drawbacks that limited their ability to comprehensively identify the cellular complexity of human tissues. Transcriptional profiling of bulk cell sample mixtures by microarray or RNA sequencing can simultaneously assess gene expression levels and proportions of abundant known cell types, but precludes identification of novel cell types and obscures the contributions of rare cell subsets to the gene expression patterns present in the bulk samples. Flow cytometry provides phenotype information at the single cell level, but is limited by the number of discrete markers that can be assessed, and relies on prior knowledge of marker expression patterns. The recent establishment of methods for single cell transcriptional profiling (1, 2) is revolutionizing our ability to understand complex cell

mixtures, avoiding the averaging phenomenon inherent in the analysis of bulk cell mixtures and providing for an unbiased assessment of phenotypic markers within the expressed genome.

In order to compare experimental results and other information about cell types, a standard reference nomenclature that includes consistent cell type names and definitions is required. The Cell Ontology (CL) is a biomedical ontology developed to provide this standard reference nomenclature for *in vivo* cell types in humans and major model organisms (3). However, the advent of high-content single cell transcriptomics for cell type characterization has resulted in a number of challenges for their representation in the CL (discussed in Bakken 2017 (4)). In this paper, we review some of the recent discoveries that have resulted from the application of single cell transcriptomics to human samples, and propose a strategy for defining cell types within the CL based on the identification of necessary and sufficient marker genes, to support interoperable and reproducible research.

Application to the human brain - Initial progress in neuronal cell type discovery by single cell RNA sequencing (scRNAseq) focused on mouse cerebral, visual, and somatosensory cortices (5, 6, 7, 8, 9). More recently, technological advances, including RNAseq using single nuclei (snRNAseq) instead of single cells (10, 11, 12), have extended these investigations into human neuronal cell type discovery (13, 14). Comprehensive reviews of these recent advances have been reported recently (15, 16).

Initial efforts toward human neuronal cell type discovery focused on identifying broad lineages. Pollen *et al.* profiled 65 neuronal cells into six categories - neural progenitor cells, radial glia, newborn neurons, inhibitory interneurons, and maturing neurons (17), while Darmanis *et al.* sequenced 466 cells, also identifying six broad, but distinct, categories - oligodendrocytes,

astrocytes, microglia, endothelial cells, oligodendrocyte precursor cells (OPCs), and neurons (18). Darmanis et al. further subtyped the adult neurons into 2 excitatory and 5 inhibitory types. More recent single *nuclei* RNAseq investigations are attempting more comprehensive cell typing. Lake *et al.* sampled 3,227 nuclei from 6 Brodmann areas, from which the neurons were classified into 8 excitatory and 8 inhibitory subtypes (13). Similarly, Boldog *et al.* sampled 769 nuclei from layer 1 of the Middle Temporal Gyrus (MTG) and identified 11 distinct inhibitory cell types (14).

Comparing results between these studies has been challenging given the different areas and layers of cortex sampled. Many of the studies leveraged classical cell type markers derived from the mouse scRNAseq literature. For example, SNAP25 expression was used to broadly define neuronal cells, while GAD1 expression defined inhibitory interneurons. Additional classical markers have then been used to subdivide the excitatory and inhibitory classes, such as CUX2 or VIP respectively; however, these markers individually are still not specific enough to define discrete cell type classes at the level of granularity revealed by clustering of the sc/snRNAseq data. In fact, there has been surprisingly limited overlap in gene sets specific for individual cell type clusters between studies, as the genes found in each study appear to be sensitive to both the context and methodology used. For example, Lake *et al*. found that cluster In1 had CNR1 (Table S5 in reference 13) as the highest ranked marker, while Boldog *et al.* found 7 distinct inhibitory types that expressed this marker (Figure 3 in reference 14). Without a standardized methodology for determining the necessary and sufficient marker genes and a corresponding marker gene reference database, comparison of newly-identified cell types to those reported in previous studies requires a complete reprocessing of the data.

89 **Application to the human immune system** - Single cell transcriptomic analysis has also been
90 applied to study the functional cell type diversity of the human immune system (reviewed in
91 Stubbington 2017 (19)). Bjorklund et al. used scRNAseq to explore the subtype diversity of
92 CD127+ innate lymphoid cells isolated from human tonsil, providing an in-depth transcriptional
93 characterization of the three major subtypes – ILC1, ILC2, and ILC3, and three additional subtypes
94 within the ILC3 class, by comparing their single cell transcriptional profiles (20).

95 Two recent studies explored the subtype diversity of dendritic cells in human blood. In addition
96 to identifying two conventional dendritic cell subtypes (cDC1 and cDC2) and one plasmacytoid
97 dendritic cell subtype, See et al. identified several subtypes that appear to correspond to precursor
98 cells, including one early uncommitted $CD123^{+}$ pre-DC subset and two $CD45RA^{+}CD123^{lo}$
99 lineage-committed subsets (pre-cDC1 and pre-cDC2), using cell sorting, scRNAseq, and *in vitro*
100 differentiation assays (21). Villani et al. used fluorescence-activated cell sorting and scRNAseq
101 to delineate six different dendritic cell subtypes (DC1 – 6) and four different monocyte subtypes
102 (Mono1 – 4), and showed that these different subtypes, which were defined based on their
103 transcriptional profiles, exhibited different functional capabilities for allogeneic T cell stimulation
104 and for cytokine production following TLR agonist stimulation (22).

105 Two recent studies have explored the phenotypes of immune cells infiltrating tumor specimens
106 using scRNAseq. In melanoma, Tirosh et al. found that the non-malignant tumor
107 microenvironment was composed of T cell, B cell, NK cell, endothelial cell, macrophage and
108 cancer-associated fibroblast (CAF) subsets (23). In contrast to the distinct transcriptional
109 phenotypes of the malignant component across individual melanoma specimens, common features
110 could be observed in the non-malignant components, with important therapeutic implications.
111 Expression of multiple complement factors by CAFs correlated with the extent of T cell

112 infiltration. T cells with activation-independent exhaustion profiles, characterized by expression
113 of co-inhibitory receptors (e.g. PD1 and TIM3), could be distinguished from cytotoxic T cell
114 profiles. Potential biomarkers that distinguish between exhausted and cytotoxic T cells could aid
115 in selecting patients for immune checkpoint blockade. In hepatocellular carcinoma, Zheng at al.
116 found clonal enrichment of both regulatory T cells and exhausted CD8 T cells using scRNAseq
117 and T cell receptor repertoire analysis (24). The diagnostic and prognostic significance of these
118 findings remain to be explored.

119 While these studies illustrate the power of single cell genomics in identifying important functional
120 cell subtypes, they also illuminate a major challenge in comparing the results from different
121 studies, due to the lack of a consistent, reusable approach for naming, defining, and comparing
122 new cell types being identified by these high content phenotyping technologies. For example, in
123 the two studies focused on the identification of dendritic cell subtypes, it is unclear if the cDC1
124 and cDC2 subtypes identified by See et al. correspond to the DC1 and DC2 subtypes identified by
125 Villani et al. Indeed, the only way to make this determination would be to perform a *de novo*
126 comparative analysis of the transcriptional profiles from both studies. For these studies to truly
127 comply with the newly emerging FAIR principles of open data (25), a robust reproducible strategy
128 for defining and representing new cell types will be essential to support their broad interoperability.

129

130 **Ontological representation** - Biomedical ontologies, as promoted by the Open Biomedical
131 Ontology (OBO) Foundry (26), provide a framework to name and define the types, properties and
132 relationships of entities in the biomedical domain. The Cell Ontology (CL) was established in
133 2005 to provide a standard reference nomenclature for *in vivo* cell types, including those observed
134 in specific developmental stages in humans and different model organisms (3). The semantic

hierarchy of CL is mainly constructed using two core relations – *is_a* and *develops_from*. Masci *et al.* proposed a major revision to the CL using dendritic cells as the driving biological use case in which the expression of specific marker proteins on the cell surface (e.g. receptor proteins) or internally (e.g. transcription factors) would be used as the main *differentia* for the asserted hierarchy (27). Diehl *et al.* applied this approach first to cell types of the hematopoietic system and then later to the full CL (28, 29, 30). As of December 2017, the CL contained 2199 cell type classes, with 583 classes within the hematopoietic cell branch alone.

We recently discussed the challenges faced by the CL in the era of high-throughput, high-content single cell phenotyping technologies, including sc/snRNAseq (4). One of the key recommendations was to establish a standard strategy for defining cell type classes that combine three essential components:

- the minimum set of ***necessary and sufficient marker genes*** selectively expressed by the cell type,
- a ***parent cell class*** in the Cell Ontology, and
- a ***specimen source description*** (anatomic structure + species).

In order to identify the set of necessary and sufficient marker genes from an sc/snRNAseq experiment, we have developed a method – NSforest – that utilizes a random forest of decision trees machine learning approach.

To illustrate how this approach can produce standard cell type definitions, we have applied the method to a transcriptomic dataset derived from single nuclei isolated from the middle temporal gyrus, cortical layer 1 of a post-mortem human brain specimen (Figure 1a in reference 14). Transcriptional profiles obtained from RNA sequencing of a collection of single sorted nuclei was used to identify 16 discrete cell types using an iterative data clustering approach. Based on the

158 expression of the previously charaterized marker genes SNAP25 and GAD1 for broad classes, 11
159 inhibitory interneurons, 1 excitatory neuron and 4 glial cell type clusters were identified.

160 In the first step (Figure 1b), NSforest takes the gene expression data matrix of single nuclei with
161 their cell type cluster membership as input, and develops a classification model for each cell type
162 cluster by comparing each Cluster X versus all non-Cluster X profiles using the random forest
163 algorithm (31). In addition to the classification model itself, NSforest produces a ranked list of
164 features (genes) that are most informative for distinguishing between Cluster X and all of the other
165 clusters.

166 In the second step, NSforest constructs single decision trees using first the top gene, then the top
167 two genes, top three genes, etc., until a stable tree topology and optimal classification accuracy is
168 achieved. The minimum number of genes necessary to obtain this stable classification result
169 corresponds to the necessary and sufficient set of marker genes defining each cell type cluster
170 within this experimental context.

171 The expression of the complete set of marker genes obtained from applying NSforest to the single
172 nuclei dataset is illustrated in Figure 2. In most cases, the expression of three marker genes is
173 sufficient to define a cell type cluster, with a range of one to five necessary and sufficient marker
174 genes per cluster. Glial cell subtypes appear to be more distinct from each other, requiring
175 relatively few genes to sufficiently define the cell type. In contrast, neuronal subtypes appear to
176 be more similar, requiring more genes to achieve specificity. In some cases, a combination of both
177 positive and negative expression optimally defines a cell type cluster.

178 For one of the inhibitory interneuron cell types defined in this study (i5), we were able to connect
179 the distinct transcriptional profile with a previous cell type defined based on its unique cellular
180 morphology – the Rosehip cell (14). This then allows us to construct an ontological representation

that includes both a colloquial name, an alternative name, and a definition combining the necessary and sufficient marker genes, a CL parent cell class, and specimen source information, as follows:

- Colloquial name – ***rosehip neuron***
- Alternative name - ***KIT-expressing MTG cortical layer 1 GABAergic interneuron, human***
- Definition - ***A human middle temporal gyrus cortical layer 1 GABAergic interneuron that selectively expresses KIT, NTNG1, and POU6F2 mRNAs***

A complete set of cell type names and definitions for all cell type clusters identified in this experiment is provided in Table 1.

These informal textual definitions can then be converted into formal ontological definitions, represented in OWL as equivalent classes, using a set of logical axioms that combine assertions about the parent cell class (interneuron), anatomic locations of the neuron cell body (soma), functional capacity of the cell type (gamma-aminobutyric acid secretion), and marker gene expression (expresses some KIT) requirements (Figure 3). Using semantic reasoners, these logical axioms can then be used to infer novel characteristics, e.g. SubClass Of 'cerebral cortex GABAergic interneuron'.

The challenge remains of ensuring that these cell type definitions, whose necessary and sufficient conditions are derived from analysis of data from one particular methodology (scRNAseq), are compatible with both existing cell type classes in the CL and cell types defined using alternative experimental methods and data analysis approaches. Working with CL developers, we are now establishing an extension ontology module containing provisional definitions for novel cell types that we and other research groups will contribute. Ontological reasoners will be used to link these cell types to more general classes in the CL proper, structure them into an extended hierarchy, and

204 determine when separate research groups have defined similar or identical cell types. CL
205 developers will review these provisional cell types periodically to determine when multiple lines
206 of evidence provide sufficient support to promote particular cell type classes to the CL proper. In
207 this way we will ensure the integrity of the CL reference, while still allowing for the rapid
208 expansion of its content to accommodate cell types defined via these new technologies.

209

210 **Conclusions –** The application of high-throughput/high-content cytometry and single cell genomic
211 techniques is producing an explosion in the number of distinct cellular phenotypes being identified
212 in human specimens. For biomedical ontologies to stay relevant, it will be critical for ontology
213 developers to establish procedures for the processing and incorporation of representations derived
214 from these data-intensive technologies into reference ontologies in a timely fashion. The
215 representation of defined cell types and their relationships in the CL will serve as a reference
216 knowledgebase to support interoperability of information about the role of cellular phenotypes in
217 human health and disease.

218

219 **Acknowledgements**

220 This work was supported by the Allen Institute for Brain Science, the JCVI Innovation Fund, the
221 U.S. National Institutes of Health R21-AI122100 and U19-AI118626, and the California Institute
222 for Regenerative Medicine GC1R-06673-B. We thank Nik Schork, Jamison McCorrison, Pratap
223 Venepally, Lindsay Cowell, Bjoern Peters, and Sirarat Sarntivijai for helpful discussion.

224

225 **Literature Cited**

1. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009) mRNA-Seq whole transcriptome analysis of a single cell. Nat. Methods, 6, 377–382.
2. Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., and Surani, M.A. (2010) Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell, 6, 468–478.
3. Bard J., Rhee S.Y., Ashburner, M. (2005) An ontology for cell types. Genome Biol., 6, R21.
4. Bakken, T., Cowell, L., Aevermann, B.D., Novotny, M., Hodge, R, Miller, J.A., Lee, A., Chang, I., McCorrison, J., Pulendran, B., et al. (2017) Cell type discovery and representation in the era of high-content single cell phenotyping. BMC Bioinformatics, 18 (Suppl. 17), 559.
5. Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015) Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science, 347, 1138–1142.
6. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell, 161, 202–1214.
7. Li, C-L., Li, K-C., Wu, D., Chen, Y., Luo, H., Zhao, J-R., Wang, S-S., Sun, M-M., Lu, Y-J., Zhong, Y-Q., et al. (2015) Somatosensory neuron types identified by high-coverage single-cell RNA-sequencing and functional heterogeneity. Cell Res., 26, 83-102.

8. Shekhar, K., Lapan, S.W., Whitney, I.E., Tran, N.M., Macosko, E.Z., Kowalczyk, M., Adiconis, X., Levin, J.Z., Nemesh, J., Goldman, M., et al. (2016) Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell, 166, 1308–1323.e30.
9. Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., et al. (2016) Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat. Neurosci., 19, 335–46.
10. Grindberg, R.V., Yee-Greenbaum, J.L., McConnell, M.J., Novotny, M., O'Shaughnessy, A.L., Lambert, G.M., Araúzo-Bravo, M.J., Lee, J., Fishman, M., Robbins, G.E., Lin, X., et al. (2013) RNA-sequencing from single nuclei. Proc. Natl. Acad. Sci. USA, 110, 19802–19807.
11. Krishnaswami, S.R., Grindberg, R.V., Novotny, M., Venepally, P., Lacar, B., Bhutani, K., Linker, S.B., Pham, S., Erwin, J.A., Miller, J.A., et al. (2016) Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. Nat. Protoc., 11, 499–524.
12. Lacar, B., Linker, S.B., Jaeger, B.N., Krishnaswami, S., Barron, J, Kelder, M., Parylak, S., Paquola, A., Venepally, P., Novotny, M., et al. (2016) Nuclear RNA-seq of single neurons reveals molecular signatures of activation. Nat. Commun., 7, 11022.
13. Lake, B.B., Ai, R., Kaeser, G.E., Salathia, N.S., Yung, Y.C., Liu, R., Wildberg, A., Gao, D., Fung, H.L., Chen, S., et al. (2016) Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science, 352, 1586–1590.
14. Boldog, E., Bakken, T, Hodge, R.D., Novotny, M, Aevermann, B.D., Baka, J., Borde, S., Close, J.L., Diez-Fuertes, F., Ding, S.L., et al. (2017) Transcriptomic and

morphophysiological evidence for a specialized human cortical GABAergic cell type. preprint bioRxiv: https://t.co/v53HzGEe3V.

15. Johnson M.B., Walsh, C.A., (2017) Cerebral cortical neuron diversity and development at the single-cell resolution. Curr. Opin. Neurobiol., 42, 9-16.
16. Lein, E.S., Belgard, T.G., Hawrylycz, M., Molnár, Z. (2017) Transcriptomic Perspectives on Neocortical Structure, Development, Evolution, and Disease. Annu. Rev. Neurosci., 40, 629-652.
17. Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., et al. (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat. Biotechnol., 32, 1053–1058
18. Darmanis, S., Sloan, S.A., Zhang, Y., Enge, M., Caneda, C., et al. (2015) A survey of human brain transcriptome diversity at the single cell level. PNAS, 112, 7285–7290.
19. Stubbington, M.J.T., Rozenblatt-Rosen, O., Regev, A., Teichmann, S.A. (2017) Single-cell transcriptomics to explore the immune system in health and disease. Science, 358, 58-63.
20. Björklund, Å.K., Forkel, M., Picelli, S., Konya, V., Theorell, J., Friberg, D., Sandberg, R., Mjösberg, J. (2016) The heterogeneity of human CD127(+) innate lymphoid cells revealed by single-cell RNA sequencing. Nat. Immunol., 4, 451-460.
21. See, P., Dutertre, C.A., Chen, J., Günther, P., McGovern, N., Irac, S.E., Gunawan, M., Beyer, M., Händler, K., Duan, K., et al. (2017) Mapping the human DC lineage through the integration of high-dimensional techniques. Science, 356.
22. Villani, A.C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, (2017) Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science, 356.

23. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H. 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science, 352, 189-196.

24. Zheng, C., Zheng, L., Yoo, J.K., Guo, H., Zhang, Y., Guo, X., Kang, B., Hu, R., Huang, J.Y., Zhang, Q., et al. (2017) Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. Cell., 169, 1342-1356.e16.

25. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data, 3, 160018.

26. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol., 11, 1251-1255.

27. Masci, A.M., Arighi, C.N., Diehl, A.D., Lieberman, A.E., Mungall, C., Scheuermann, R.H., Smith, B., Cowell, L.G. (2009) An improved ontological representation of dendritic cells as a paradigm for all cell types. BMC Bioinformatics. 10, 70.

28. Diehl, A.D., Augustine, A.D., Blake, J.A., Cowell, L.G., Gold, E.S., Gondré-Lewis, T.A., Masci, A.M., Meehan, T.F., Morel, P.A., Nijnik, A., et al. (2011) Hematopoietic cell types: prototype for a revised cell ontology. J. Biomed. Inform., 1, 75-79.

29. Meehan, T.F., Masci, A.M., Abdulla, A., Cowell, L.G., Blake, J.A., Mungall, C.J., Diehl, A.D. (2011) Logical development of the cell ontology. BMC Bioinformatics, 12, 6.

30. Diehl, A.D., Meehan, T.F., Bradford, Y.M., Brush, M.H,, Dahdul, W.M., Dougall, D.S., He, Y., Osumi-Sutherland D, Ruttenberg A, Sarntivijai S, et al. (2016) The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. J. Biomed. Semantics. 7, 44.

31. Berthold M., Cebron N., Dill F., Gabriel, T., Kotter, T., Meinl, T., Ohl, P., Thiel, K., Wiswedel, B., et al. (2008) KNIME: The Konstanz Information Miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, editors. (eds). Data Analysis, Machine Learning and Applications. Springer Berlin Heidelberg, Chapter 38, 319–326.

Figure legends

Figure 1. Identification of necessary and sufficient marker genes using NSforest – a) A typical single cell/single nuclei RNA sequencing workflow in which a tissue specimen is obtained, single cells/nuclei isolated by fluorescence activated cell sorting, amplified cDNA quantified by sequencing, and cell types identified by clustering the resultant transcriptional profiles. b) The NSforest approach takes a data matrix of expression values (e.g. transcripts per million reads) of genes (rows) in single cell/nuclei samples (columns) grouped by cell type cluster membership. In the first step, the expression levels of genes are used as features in the random forest machine learning procedure to train classification models comparing single cell/nuclei expression data in one cell type cluster against single cell/nuclei expression data in all other clusters, for every cell type cluster separately, using the Random Forest Learner in KNIME v3.1.2. Each cell type cluster

337 classification model is constructed from one hundred thousand trees using Information Gain Ratio
338 as the splitting criteria, where each decision tree is generated using the default bagging parameters
339 - the square root of the number of features and a bootstrap of samples equal to the training set size.
340 For each cell type cluster classification model, the method outputs usage statistics, including how
341 often each gene is used as a branching criterion and the number of times it was a candidate across
342 all random decision trees. By summing the frequency of use when a candidate across the first three
343 branching levels, the list of genes can be ranked by their usefulness in distinguishing one cell type
344 clusters from the other clusters. In the second step, single decision trees are constructed using the
345 first gene from the ranked list, the first two genes, the first three genes, etc. Each individual tree
346 is then assessed for classification accuracy and tree topology using the training data. Given the
347 objective of determining the necessary and sufficient marker genes, we apply additional criteria in
348 scoring the trees - we restrict each gene to being used in only one branch per tree, and find the
349 optimal classification for the target cluster only, rather than the overall classification score. The
350 addition of genes from the ranked list is stopped when an optimal classification or stable tree
351 topology is achieved. The minimum number of genes used to produce this optimal result
352 corresponds to the set of necessary and sufficient marker genes required to define the cell type
353 cluster.

354

355 Figure 2. Marker gene expression patterns in single nuclei grouped by cluster – A heatmap of
356 expression levels for the necessary and sufficient marker genes identified for all 16 clusters across
357 all single nuclei grouped by cell type cluster is shown, including 1 excitatory (e1), 11 inhibitory
358 (i1 – i11), and 4 glial (g1 – g4) cell type clusters. In total, 49 markers genes were selected as being

359 necessary and sufficient to distinguish these 16 different cell type clusters from cortical layer 1 of
360 the human brain middle temporal gyrus region.

361

362 Figure 3. Formal rosehip neuron definition using logical axioms – A set of logical axioms about
363 the anatomic local of the cell body (soma), the functional capacity, and the necessary and sufficient
364 marker gene expressions are combined to construct an equivalent class cell type definition for the
365 rosehip neuron interneuron cluster – i4 (see Boldog 2017 (14) for more information about how this
366 cell type was characterized).

367

368 **Table 1. Cell types identified in cortical layer 1 of the human middle temporal gyrus**

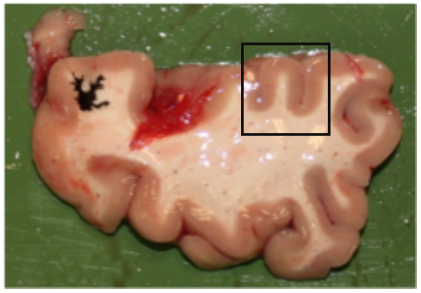
Cluster ID	***Cell Type Name***	***Cell Type Definition***
e1	*TESPA1-expressing MTG cortical layer 1 excitatory neuron, human*	*A human middle temporal gyrus cortical layer 1 excitatory neuron that selectively expresses TESPA1, LINC00507, and SLC17A7 mRNAs, and lacks expression of KCNIP1 mRNA*
i1	*COL5A2-expressing MTG cortical layer 1 interneuron, human*	*A human middle temporal gyrus cortical layer 2 GABAergic interneuron that selectively expresses COL5A2 and NDNF and FAT1 mRNAs*
i2	*LHX6-expressing MTG cortical layer 2 interneuron, human*	*A human middle temporal gyrus cortical layer 2 GABAergic interneuron that selectively expresses LHX6, GRIK3, and FLT3, while of lacking expression of COBL and CALB2 mRNAs*
i3	*BAGE2 expressing MTG cortical layer 1 interneuron, human*	*A human middle temporal gyrus cortical layer 1 GABAergic interneuron that selectively expresses BAGE2 and SEMA3C and SYT10 and CALB2 and COL21A1 mRNAs*
i4	*ARHGAP36 expressing MTG cortical layer 1 interneuron, human*	*A human middle temporal gyrus cortical layer 1 GABAergic interneuron that selectively expresses ARHGAP36 and ADAM33 and LINC01435 and MC4R mRNAs*
i5	*KIT-expressing MTG cortical layer 1 interneuron, human*	*A human middle temporal gyrus cortical layer 1 GABAergic interneuron that selectively expresses KIT and NTNG1 and POU6F2 mRNAs*
i6	*GPR149-expressing MTG cortical layer 1 interneuron, human*	*A human middle temporal gyrus cortical layer 1 GABAergic interneuron that selectively expresses GPR149 and VIP and PLCE1 mRNAs*
i7	*TGFBR2 -expressing MTG cortical layer 1 interneuron, human*	*A human middle temporal gyrus cortical layer 1 GABAergic interneuron that selectively expresses TGFBR2 and HCRTR2 and PAX6 mRNAs*
i8	*SNCG-expressing MTG cortical*	*A human middle temporal gyrus cortical layer 1 GABAergic*

	layer 1 interneuron, human	*interneuron that selectively expresses SNCG and EDNRA and KCNK2 and ARHGAP18 mRNAs*
i9	*VIP-expressing MTG cortical layer 1 interneuron, human*	*A human middle temporal gyrus cortical layer 1 GABAergic interneuron that selectively expresses VIP and IQGAP2 and TAC3 mRNAs*
i10	*TSPAN12-expressing MTG cortical layer 1 interneuron, human*	*A human middle temporal gyrus cortical layer 1 GABAergic interneuron that selectively expresses TSPAN12 and CHRNB3 and FAM46A and DCN mRNAs*
i11	*EGF-expressing MTG cortical layer 1 interneuron, human*	*A human middle temporal gyrus cortical layer 1 GABAergic interneuron that selectively expresses EGF and NRG1-IT1 mRNAs*
g1	*Linc00499-expression MTG cortical layer 1 glial cell, human*	*A human middle temporal gyrus cortical layer 1 glial cell that selectively expresses Linc00499 and ATP1A2 mRNAs*
g2	*APBB1IP-expressing MTG cortical layer 1 glial cell, human*	*A human middle temporal gyrus cortical layer 1 glial cell that selectively expresses APBB1IP mRNAs*
g3	*PTPRZ1-expressing MTG cortical layer 1 glial cell, human*	*A human middle temporal gyrus cortical layer 1 glial cell that selectively expresses PTPRZ1 and XYLT1 mRNAs*
g4	*ST18 expressing MTG cortical layer 1 glial cell, human*	*A human middle temporal gyrus cortical layer 1 glial cell that selectively expresses ST18 mRNAs*

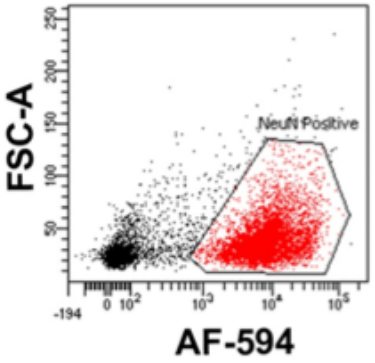
369

a

Tissue specimen

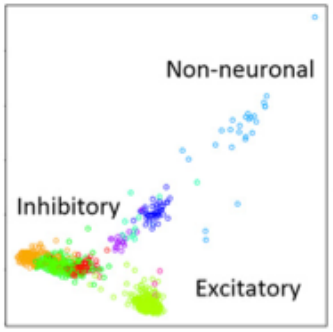


Single nuclei sorting



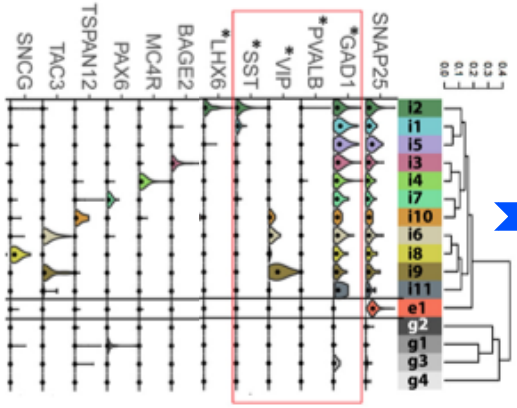
RNA sequencing

Cell type clustering



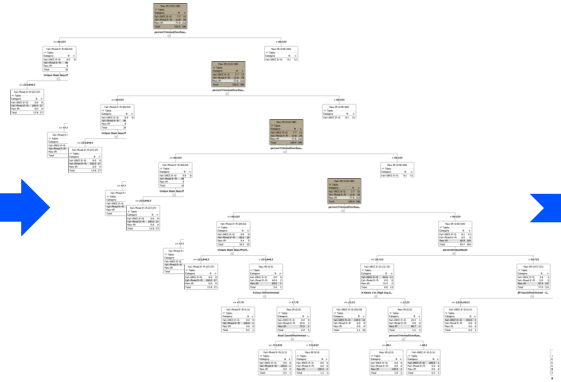
b

Expression matrix



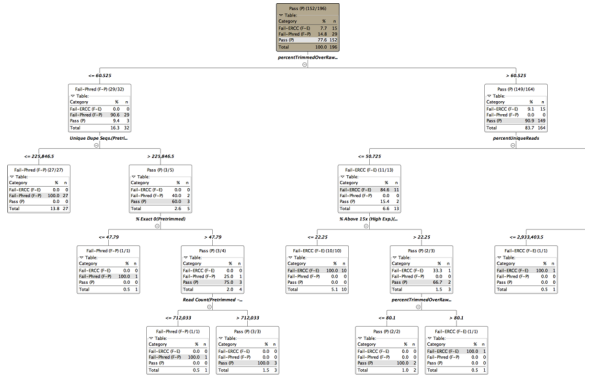
Cluster membership & lists of differential genes

Random forest

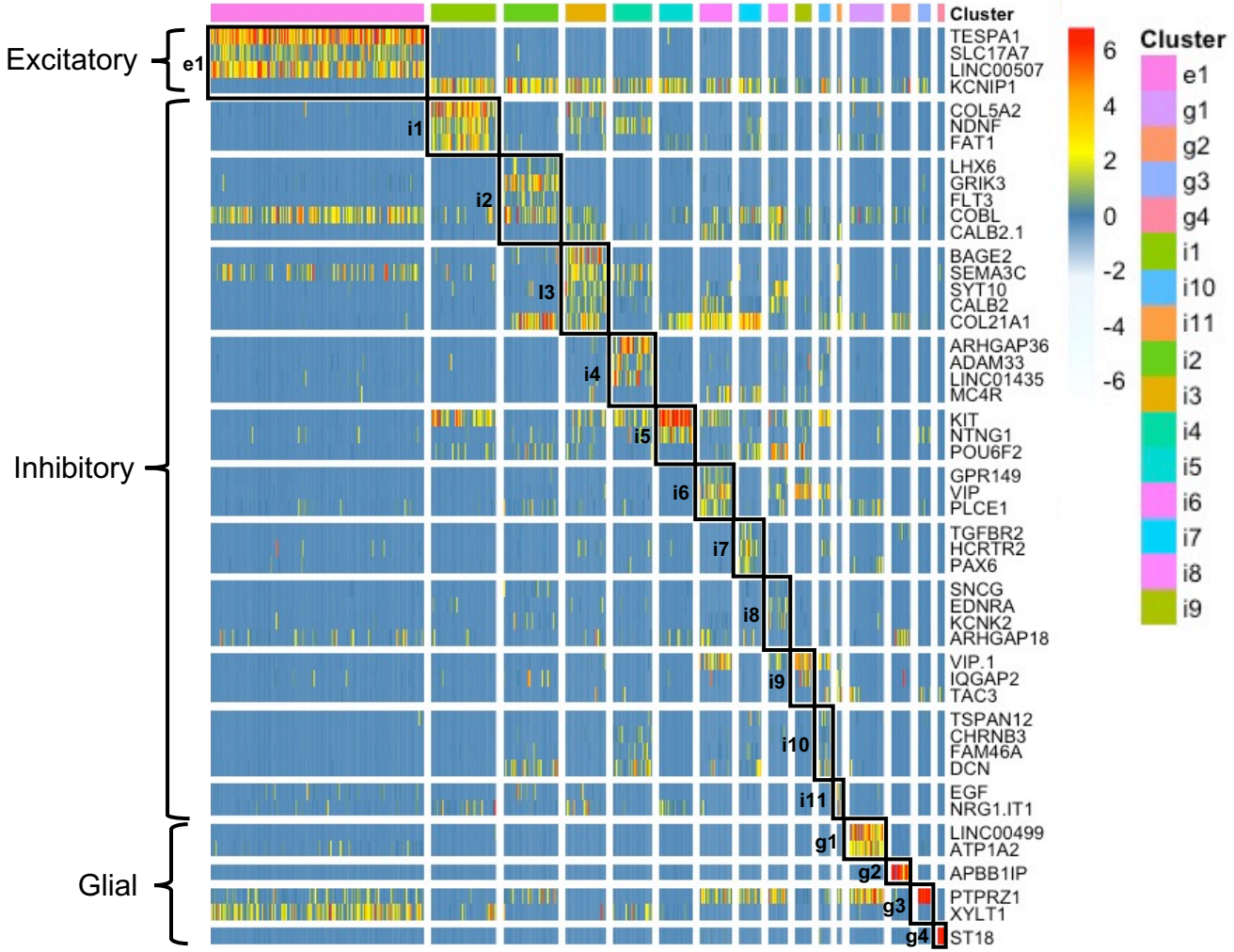


Ranked list of useful genes

Decision trees 1, 2, 3….



Necessary and sufficient genes = Cluster-specific marker genes


Excitatory
Inhibitory
Glial
e1
i1
i2
I3
i4
i5
i6
i7
i8
i9
i10
i11
g1
g2
g3
g4
Cluster
TESPA1
SLC17A7
LINC00507
KCNIP1
COL5A2
NDNF
FAT1
LHX6
GRIK3
FLT3
COBL
CALB2.1
BAGE2
SEMA3C
SYT10
CALB2
COL21A1
ARHGAP36
ADAM33
LINC01435
MC4R
KIT
NTNG1
POU6F2
GPR149
VIP
PLCE1
TGFBR2
HCRTR2
PAX6
SNCG
EDNRA
KCNK2
ARHGAP18
VIP.1
IQGAP2
TAC3
TSPAN12
CHRNB3
FAM46A
DCN
EGF
NRG1.IT1
LINC00499
ATP1A2
APBB1IP
PTPRZ1
XYLT1
ST18
6
4
2
0
-2
-4
-6
Cluster
e1
g1
g2
g3
g4
i1
i10
i11
i2
i3
i4
i5
i6
i7
i8
i9

Annotations

rdfs:label [language: en]

rosehip neuron

Description: 'rosehip neuron'

Equivalent To

- interneuron
  and ('has soma location' some 'cortical layer I')
  and ('has soma location' some 'middle temporal gyrus')
  and ('capable of' some 'gamma-aminobutyric acid secretion, neurotransmission')
  and (expresses some KIT)
  and (expresses some NTNG1)
  and (expresses some POU6F2)

SubClass Of

- 'synapsed to' some 'layer III pyramidal neuron'
- 'cerebral cortex GABAergic interneuron'