Prioritizing uncharacterized genes in the search for glioma biomarkers

Rheal A Towner^{*1,2} & Jonathan D Wren^{3,4}

“...there is likely to be a higher relative contribution to scientific knowledge when characterizing/annotating a function or role for a gene that has not yet been studied over adding additional information to one that has already been very well-studied.”

A total of 40% of all primary CNS tumors are diagnosed as gliomas, with glioblastomas (GBM) being the most malignant [1]. There is a very poor survival time of approximately 15 months for most patients diagnosed with GBM [1]. The malignant nature of high grade gliomas makes them one of the leading causes of cancer death [1]. To classify tumors and assess prognosis, differences in molecular composition, or ‘biomarkers’, between tissue types can be used diagnostically to assess and manage adult malignant gliomas [2–7]. Recent biomarkers have been reported from genome-wide surveys associating somatic mutations with risk of glioma development [8]. The molecular biomarkers that are most commonly used to evaluate adult malignant gliomas from biopsies include 1p/19q codeletion, methylation of the *O*-6 methylguanine-DNA methyltransferase (MGMT) gene promoter, alterations in the EGF receptor (EGFR) pathway, and isocitrate dehydrogenase 1 (*IDH1*) and *IDH2* gene

mutations [2–9]. Several proteomics-based approaches have been used to search for proteins that are unique to gliomas [10], but these have been severely limited by sample size, the ability to detect low abundance proteins and data reproducibility. It is also important to note that many of these studies have generated hundreds and even thousands of putative candidates, however, few have been able to follow up with subsequent validation and characterization approaches.

The value of therapeutic biomarkers depends, in large part, on whether or not the technology exists to take advantage of their differential presence or absence in diseased cells. Arguably, the most valuable type of biomarkers are proteins that are either expressed on the cell surface or found in plasma, because the technology is readily available to manufacture antibodies to these proteins and deliver them intravenously, either alone or as conjugates to cytotoxic compounds. In fact, over the past decade, the market for such

KEYWORDS

- bioinformatics • biomarker
- cancer • diagnostics • glioma
- prognostics • therapeutics

¹Advanced Magnetic Resonance Center, Oklahoma Medical Research Foundation, 825 NE 13th Street, Oklahoma City, OK 73104, USA

²Department of Pathology, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

³Arthritis & Clinical Immunology Program, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA

⁴Department of Biochemistry & Molecular Biology, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

*Author for correspondence: rheal-towner@omrf.org

therapeutic antibodies has grown exponentially [11], as they have been used effectively to target human proteins involved in a variety of disease-related processes. To date, they have been used to treat a wide range of human ailments from autoimmune diseases to cancer.

Interestingly, only approximately a third of the 25,000 protein-coding genes in humans have appeared within the title or abstract on MEDLINE [12,13], which strongly suggests that their functions and biological roles remain unknown. Additionally, for proteins that do appear in the literature, approximately 75% of the cumulative scientific effort to date has been highly skewed, focusing predominantly on 10% of them [13], suggesting that another significant fraction of human proteins is relatively poorly characterized. From a therapeutic standpoint, this implies that many useful biomarkers likely exist but are yet to be characterized as such, and the question becomes how to best identify and prioritize them.

In recent years, the number of commonly available microarray data sets in public repositories such as Gene Expression Omnibus (GEO) have grown dramatically [14]. Consequently, bioinformatics approaches have been developed to search for consistent transcriptional correlations between genes. The rationale is straightforward – if there is a lack of knowledge regarding what functions or roles a gene has, if other genes are highly correlated in their transcriptional patterns with them, then common functions or roles of the correlated genes are likely to be shared by the uncharacterized gene. This ‘guilt by association’ approach seems sensible, and has been used by multiple groups with reported success in predicting Gene Ontology (GO) categories [15], but has also come under fire as most methods are not able to outperform simple methods that prioritize predicted functions merely by how common that function is [15,16]. We would argue that, despite the ease and standardization of using GO, it has several drawbacks that may be made up for by directly mining the published literature for gene-term associations. First, and importantly, it enables strength of association to be calculated (rather than a binary assignment of an annotation), it is more current than GO and can expand beyond the categories within GO (cellular component, molecular function and biological process) to encompass phenotypes, diseases and other associations of interest.

We previously used a bioinformatics approach called global microarray meta-analysis (GAMMA) [12,17–18] to analyze approximately 16,000 human microarray experiments obtained from the National Center for Biotechnology Information’s GEO database. For each gene, GAMMA identifies a set of 30 genes most consistently and specifically correlated with it across the heterogeneous conditions analyzed. Each gene, whether it had any published literature describing its function or not, had its function inferred by analysis of what the correlated genes had in common within MEDLINE. This was done via an automated, large-scale analysis of the peer-reviewed literature [17]. In the end, we were able to identify genes that were consistently transcribed with established glioma-related genes, but which had themselves never been associated with gliomas in the literature. It enabled us to identify ELTD1 [19] and later validate Spondin, Plexin B2, SLIT 3, Fibulin-1 and Lingo [20] as putative glioma-associated markers. These biomarkers were then experimentally validated regarding their presence in human gliomas via immunohistochemistry in high- and low-grade gliomas [19,20].

The value of these putative biomarkers for clinical therapy is currently being explored, but the take-home message is that technology is enabling us to infer, computationally, which proteins within this final third of uncharacterized proteins, should be seriously considered for their putative involvement in disease processes. We would argue that there is likely to be a higher relative contribution to scientific knowledge when characterizing/annotating a function or role for a gene that has not yet been studied over adding additional information to one that has already been very well-studied. Arguably, the reason the literature is so skewed towards studying genes that have already been characterized is a combination of risk-aversion and practical considerations such as the lack of reagents (e.g., antibodies and RNAi). However, in recent years the number of commercial antibodies to uncharacterized proteins and RNAi to uncharacterized genes has grown. Thus, we feel the time has come to make a concerted effort to move this final third from the realm of the uncharacterized into the scientific literature, at least with a putative role or significance. One cannot stand on the shoulders of giants, after all, if there are no giants.

“...many useful biomarkers likely exist but are yet to be characterized as such, and the question becomes how to best identify and prioritize them.”

Financial & competing interests disclosure

The research published in references [14,15] was supported by the Oklahoma Medical Research Foundation (to RA Towner), the Chapman Foundation (to JD Wren) and NIH grants 1P20GM103636 and 8P20GM103456-09 (to JD Wren). The authors have no other relevant affiliations or

financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

- 1 Central Brain Tumor Registry of the United States. 2011 CBTUS Statistical Report: Primary brain and central Nervous System Tumors Diagnosed in the United States in 2004–2007 (2011). www.cbtrus.org/2011-NPCR-SEER/WEB-0407-Report-3-3-2011.pdf
- 2 Louis DN. Molecular pathology of malignant gliomas. *Annu. Rev. Pathol.* 1, 97–117 (2006).
- 3 The Cancer Genome Atlas (TCGA) Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068 (2008).
- 4 Riemenschneider MJ, Jeuken JWM, Wesseling P, Reifenberger G. Molecular diagnostics of gliomas: state of the art. *Acta Neuropathol.* 120, 567–584 (2010).
- 5 Jansen M, Yip S, Louis DN. Molecular pathology in adult gliomas: diagnostic, prognostic, and predictive markers. *Lancet Neurol.* 9, 717–726 (2010).
- 6 Colman H, Zhang L, Sulman EP *et al.* A multigene predictor of outcome in glioblastoma. *Neuro Oncol.* 12, 49–57 (2010).
- 7 Farias-Eisner G, Bank AM, Hwang BY *et al.* Glioblastoma biomarkers from bench to bedside: advances and challenges. *Br. J. Neurosurg.* 26, 189–194 (2012).
- 8 Parsons DW, Jones S, Zhang X *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* 321, 1807–1812 (2008).
- 9 Silber JR, Bobola MS, Blank A, Chamberlain MC. O(6)-methylguanine-DNA methyltransferase in glioma therapy: promise and problems. *Biochim. Biophys. Acta* 1826, 71–82 (2012).
- 10 Niclou SP, Fack F, Rajcevic U. Glioma proteomics: status and perspectives. *J. Proteomics* 73, 1823–1838 (2010).
- 11 Wren JD. A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide. *Bioinformatics* 25, 1694–1701 (2009).
- 12 Giles CB, Wren JD. Large-scale directional relationship extraction and resolution. *BMC Bioinformatics* 9(Suppl. 9), S11 (2008).
- 13 Wren JD. Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics* 5, 145 (2004).
- 14 Towner RA, Jensen RL, Colman H *et al.* ELTD1, a potential new biomarker for gliomas. *Neurosurgery* 72, 77–91 (2013).
- 15 Towner RA, Jensen RL, Vaillant B *et al.* Experimental validation of 5 *in-silico* predicted glioma biomarkers. *Neuro-Oncology* 15, 1625–1634 (2013).
- 16 Leavy O. Therapeutic antibodies: past, present and future. *Nat. Rev. Immunol.* 10(5), 297 (2010).
- 17 Edwards AM, Isserlin R, Bader GD, Frye SV, Willson TM, Yu FH. Too many roads not taken. *Nature* 470, 163–165 (2011).
- 18 Barrett T, Wilhite SE, Ledoux P *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41, D991–D995 (2013).
- 19 Gillis J, Pavlidis P. The impact of multifunctional genes on ‘guilt by association’ analysis. *PLoS ONE* 6, e17258 (2011).
- 20 Gillis J, Pavlidis P. ‘Guilt by association’ is the exception rather than the rule in gene networks. *PLoS Comput. Biol.* 8, e1002444 (2012).