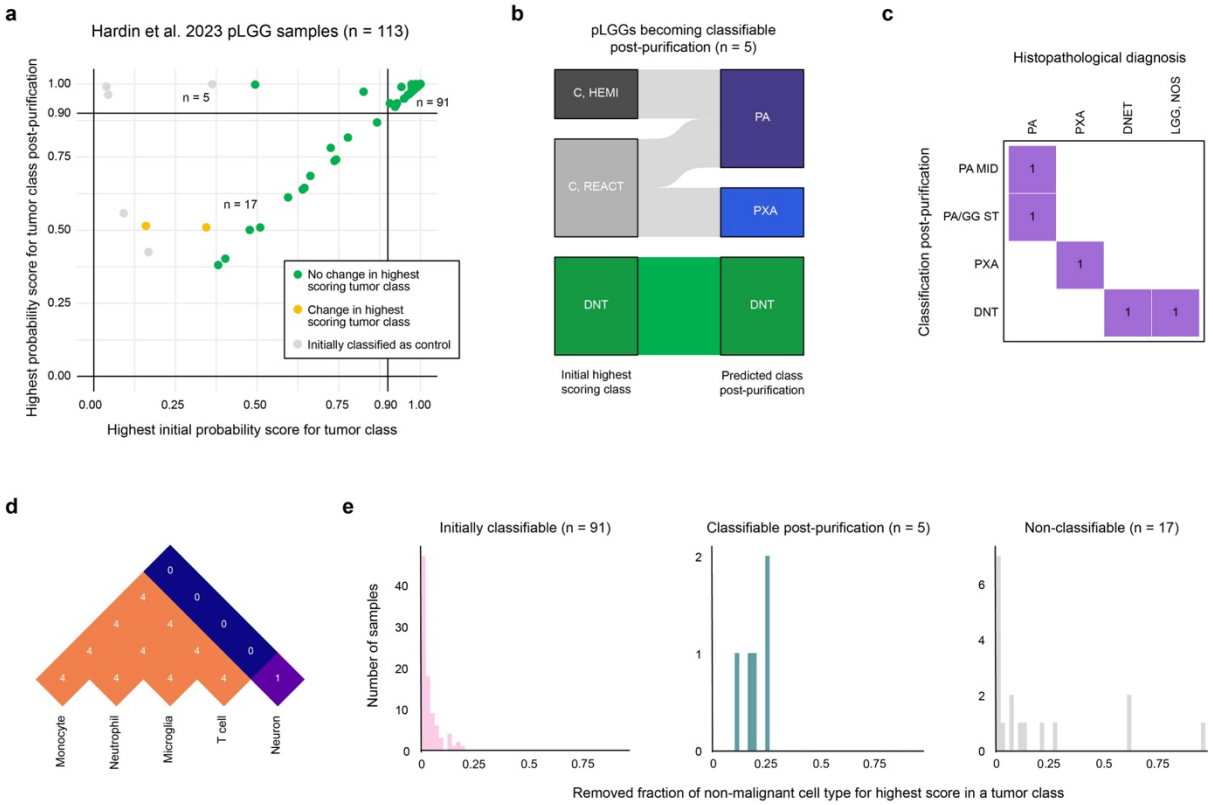
In silico purification improves DNA methylation-based classification rates of pediatric low-grade gliomas

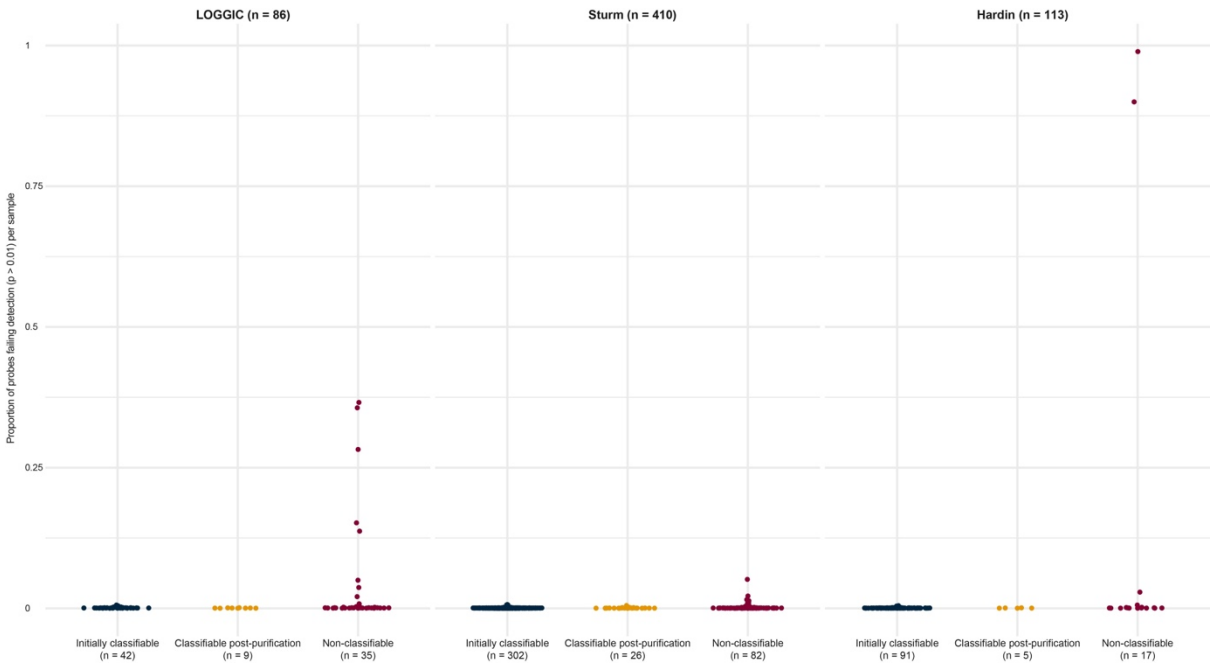
Liv Jürgensen[1,2,3,4], Salvatore Benfatto[1,2,3], Simone Schmid[4,5], Bjarne Daenekas[1,2,3,4], Julia Großer[6], Pablo Hernáiz Driever[6], Arend Koch[4], David Capper[4,5], Volker Hovestadt[1,2,3]

1) Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA
2) Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA, USA
3) Broad Institute of MIT and Harvard, Cambridge, MA, USA
4) Department of Neuropathology, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
5) German Cancer Consortium (DKTK), Partner Site Berlin, German Cancer Research Center (DKFZ), Heidelberg, Germany
6) Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, German HIT-LOGGIC-Registry for low-grade glioma in children and adolescents, Berlin, Germany

Correspondence: volker_hovestadt@dfci.harvard.edu



Supplementary Fig. 1 In silico purification to an additional external pLGG dataset. **a** Scatter plot of all 113 pLGG samples from the Hardin dataset comparing the probability scores for the highest-scoring tumor class, or tumor class family when applicable, before and after purification. Black lines mark the 0.9 classification threshold. Dots are colored by class changes, with green for stable classification, yellow for tumor class changes, and grey for initial classification as control. **b** Sankey plot comparing the initial highest-scoring class with the confidently predicted class after purification for all initially non-classifiable samples from the Hardin dataset that became classifiable through purification (n = 5). **c** Matrix comparing histopathological diagnoses with classifications after purification for the 5 newly classifiable pLGG samples from the Hardin dataset. Numbers indicate sample counts. **d** Matrix showing the overlap in the 5 newly classifiable pLGG samples post-purification after the removal of different cell types. Numbers and color indicate the count of shared samples between cell type removals. **e** Histograms showing the fraction of non-malignant cell signal removed for optimal classification in all 113 pLGG samples from the Hardin dataset. Samples are grouped by classification status: initially classifiable (left), newly classifiable post-purification (middle), and non-classifiable (right).



Supplementary Fig. 2 Strip plot showing sample-wise proportion of probes with detection p-values >0.01 across initially classifiable, classifiable post-purification, and persistently non-classifiable samples in LOGGIC (*n* = 86), Sturm (*n* = 410), and Hardin (*n* = 113).