

DISCOVERY OF CLINICALLY RELEVANT FUSIONS IN PEDIATRIC CANCER

Stephanie LaHaye¹, James R. Fitch¹, Kyle J. Voytovich¹, Adam C. Herman¹, Benjamin J. Kelly¹,
Grant E. Lammi¹, Saranga Wijeratne¹, Samuel J. Franklin¹, Kathleen M. Schieffer¹, Natalie Bir¹,
Sean D. McGrath¹, Anthony R. Miller¹, Amy Wetzel¹, Katherine E. Miller¹, Tracy A. Bedrosian¹,
Kristen Leraas¹, Kristy Lee¹, Ajay Gupta², Bhuvana Setty^{2,3}, Daniel R. Boué^{4,5}, Jeffrey R. Leonard^{3,6},
Jonathan L. Finlay^{2,3}, Mohamed S. Abdelbaki^{2,3}, Diana S. Osorio^{2,3}, Selene C. Koo^{4,5}, Daniel C.
Koboldt¹, Vincent Magrini^{1,3}, Catherine E. Cottrell^{1,3,4}, Elaine R. Mardis^{1,3}, Richard K. Wilson^{1,3},
Peter White^{1,3,*}

¹The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH. ²Division of Hematology, Oncology, Blood and Marrow Transplant, Nationwide Children's Hospital, Columbus, OH. ³Department of Pediatrics, The Ohio State University, Columbus, OH. ⁴Department of Pathology, The Ohio State University, Columbus, OH. ⁵Department of Pathology, Nationwide Children's Hospital, Columbus, OH. ⁶Section of Neurosurgery, Nationwide Children's Hospital Columbus, OH.

***Corresponding Author:** Prof. Peter White, PhD, The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, 575 Children's Crossroad, Columbus, OH 43215. USA. Tel: +1 (614) 355-2671; Email: peter.white@nationwidechildrens.org

Keywords: transcriptomics, genomics, pediatric neoplasms, gene fusions, cancer, RNA-Seq

Running title: Fusion Identification in Pediatric Cancer

25 **ABSTRACT**

26 **Background:** Pediatric cancers typically have a distinct genomic landscape when compared to
27 adult cancers and frequently carry somatic gene fusion events that alter gene expression and drive
28 tumorigenesis. Sensitive and specific detection of gene fusions through the analysis of next-
29 generation-based RNA sequencing (RNA-Seq) data is computationally challenging and may be
30 confounded by low tumor cellularity or underlying genomic complexity. Furthermore, numerous
31 computational tools are available to identify fusions from supporting RNA-Seq reads, yet each
32 algorithm demonstrates unique variability in sensitivity and precision, and no clearly superior
33 approach currently exists. To overcome these challenges, we have developed an ensemble fusion
34 calling approach to increase the accuracy of identifying fusions.

35 **Results:** Our ensemble fusion detection approach utilizes seven fusion calling algorithms: Arriba,
36 CICERO, FusionMap, FusionCatcher, JAFFA, MapSplice, and STAR-Fusion, which are packaged as a
37 fully automated pipeline using Docker and AWS serverless technology. This method uses paired
38 end RNA-Seq sequence reads as input, and the output from each algorithm is examined to identify
39 fusions detected by a consensus of at least three algorithms. These consensus fusion results are
40 filtered by comparison to an internal database to remove likely artifactual fusions occurring at
41 high frequencies in our internal cohort, while a “known fusion list” prevents failure to report
42 known pathogenic events. We have employed the ensemble fusion-calling pipeline on RNA-Seq
43 data from 229 patients with pediatric cancer or blood disorders studied under an IRB-approved
44 protocol. The samples consist of 138 central nervous system tumors, 73 solid tumors, and 18
45 hematologic malignancies or disorders. The combination of an ensemble fusion-calling pipeline
46 and a knowledge-based filtering strategy identified 67 clinically relevant fusions among our
47 cohort (diagnostic yield of 29.3%), including *RBPMS-MET*, *BCAN-NTRK1*, and *TRIM22-BRAF*

48 fusions. Following clinical confirmation and reporting in the patient's medical record, both known
49 and novel fusions provided medically meaningful information.

50 **Conclusions:** Our ensemble fusion detection pipeline offers a streamlined approach to discover
51 fusions in cancer, at higher levels of sensitivity and accuracy than single algorithm methods.
52 Furthermore, this method accurately identifies driver fusions in pediatric cancer, providing
53 clinical impact by contributing evidence to diagnosis and, when appropriate, indicating targeted
54 therapies.

55

56 **BACKGROUND**

57 Globally, there are approximately 300,000 pediatric and adolescent cases of cancer
58 diagnosed each year [1, 2]. While advances in medicine have led to a drastic improvement in 5-
59 year overall survival rates (up to 84% in children under 15), pediatric cancer remains the most
60 common cause of death by disease in developed countries [3, 4]. Pediatric cancers are defined by a
61 distinct genomic landscape when compared to adult cancers, which includes an overall low
62 number of somatic single nucleotide variants, common driver fusions and epigenetic changes that
63 drive a specific transcriptional program. Pediatric cancers are often considered embryonic in
64 origin and demonstrate a significant germline predisposition component approaching 10% [5-7].

65 Many pediatric tumors contain gene fusions resulting from the juxtaposition of two genes
66 (**ADDITIONAL FILE 1: FIGURE S1**)[6]. Fusions typically occur through chromosomal rearrangements,
67 and often lead to dysregulated gene expression of one or both gene partners [8-11]. Fusions can
68 also generate chimeric oncoproteins, wherein functional domains from both genes are retained,
69 often leading to aberrant and strong activation of nonspecific downstream targets [12]. The
70 alterations in gene expression and activation of downstream targets induced by fusions are

71 considered to be oncogenic events in pediatric cancer and increasingly may indicate response to
72 specific targeted therapies.

73 The identification of an oncogenic fusion can provide medically meaningful information in
74 the context of diagnosis, prognosis, and treatment regimens in pediatric cancers. Fusions may
75 provide diagnostic evidence for a specific histological subgroup. For example, *EWSR1-FLI1* fusions
76 are highly associated with Ewing sarcoma, while the presence of a *C11orf95-RELA* fusion aids in
77 subgrouping supratentorial ependymomas [12]. The detection of certain fusions, such as *BCR-ABL*
78 in acute lymphocytic leukemia, can be used as a surrogate for residual tumor load and treatment
79 response [13]. Fusions may also provide prognostic indication, such as *KIAA1549-BRAF* in low
80 grade astrocytomas, which have a more favorable outcome compared to non-*BRAF* fused tumors
81 [14, 15]. In addition, fusions that involve kinases can present therapeutic targets, including *FGFR1-*
82 *TACC1*, *FGFR3-TACC3*, *NPM1-ALK*, and *NTRK* fusions [2, 12, 16-19].

83 However, regardless of the clear clinical benefits of characterizing fusion events in a given
84 patient's tumor, accurate identification of fusions from next generation sequencing DNA data
85 alone is not straightforward and they often go undiscovered. In particular, many fusions are not
86 detectable by exome sequencing (ES) due to breakpoint locations that frequently occur in non-
87 coding or intronic regions which may not have corresponding capture probes. Even whole genome
88 sequencing (WGS) NGS data has proved difficult to evaluate complex rearrangements resulting in
89 gene fusions due to a high false positive rate and due to the limitations of short read lengths [20,
90 21]. By contrast, next-generation RNA sequencing data, or RNA-Sequencing (RNA-Seq), offers an
91 unbiased data type suitable for fusion detection, while also providing information about the
92 expression of fusion transcripts, including multiple isoforms, and fusions that occur due to
93 aberrant splicing events [22, 23].

94 While RNA-Seq is a powerful tool for fusion detection, it is not without its limitations.
95 Notably, there is currently a major deficit in our ability to accurately identify fusions in spite of
96 having many computational approaches available. Here, consistently identifying gene fusion
97 events with high sensitivity and precision using one algorithm is unlikely and this is of critical
98 importance in a clinical diagnostic setting [12]. Computational approaches that have been tuned
99 for high sensitivity are limited by also calling numerous false positives, requiring extensive
100 manual review of data, while those with a low false discovery rate (FDR) often miss true positives
101 due to over-filtering [12]. To overcome these complications of sensitivity and specificity, we have
102 employed an ensemble pipeline, which merges results from seven algorithmic approaches to
103 identify, filter and output prioritized fusion predictions.

104 Another common issue encountered in fusion prediction is the identification of likely non-
105 pathogenic fusions, due both to read-through events and fusions occurring in non-disease
106 involved (normal) genomes.[12, 24, 25] We addressed these sources of false positivity through the
107 implementation of a filtering strategy that removes known normal fusions and RNA transcription
108 read-through events, based on internal frequency of detection and location of chromosomal
109 breakpoints. Lastly, to prevent over-filtering and inadvertent removal of previously described
110 known pathogenic fusion events, we have developed and continually update a list containing
111 known pathogenic fusion partners, that will return any data-supported fusions to the output list of
112 prioritized fusion results for further evaluation.

113 The ensemble fusion detection pipeline outperformed all single algorithm methods we
114 evaluated, achieving high levels of sensitivity, while simultaneously minimizing false positive calls
115 and non-clinically relevant fusion predictions. Here, we describe our ensemble fusion detection
116 approach, its performance on commercial control reference standards with known fusions, and its
117 implementation on a pediatric cohort consisting of rare, treatment refractory, or relapsed cancers

118 and hematologic diseases. Utilization of our ensemble approach resulted in a diagnostic yield of
119 approximately 30% in our cohort, identified novel fusion partners, and has provided diagnostic
120 information and/or targeted treatment options for this patient population.

121

122 **RESULTS**

123 *Development and optimization of ensemble pipeline on a control reference standard*

124 Identification of gene fusions through the use of a single algorithm is often associated with
125 low specificity and poor precision [12]. Given prior literature supporting multi-algorithmic
126 approaches to improve upon these deficits, we studied the intricacies of several fusion detection
127 algorithms, and applied a defined set of algorithms with desired properties, aimed at detecting
128 true positive fusions while minimizing false positive fusions [25-27]. After evaluating each
129 algorithm's output, we developed our ensemble fusion detection pipeline that combines output
130 consensus calls from seven different computational approaches (**FIGURE 1A**), calculates the
131 concordant fusion partners and breakpoints, and filters this output list based on internal
132 frequency, reads of evidence, and breakpoint location. A list of known pathogenic fusions rescues
133 any known pathogenic fusion gene partners with suitable algorithmic and read support for further
134 evaluation (**ADDITIONAL FILE 1: TABLE S3**).

135 To optimize the approach, we utilized a reference standard from a commercial provider
136 (Seraseq Fusion RNA, SeraCare, Milford, MA), containing synthetic RNAs representing 14 cancer-
137 associated fusions in varying proportions (**ADDITIONAL FILE 1: TABLES S1 AND S2**). Data generated
138 from these RNA-Seq libraries, performed as replicates for a range of dilutions, were analyzed
139 using the ensemble pipeline. We compared the output derived from a consensus of two or more
140 callers to that from a consensus of three or more callers by calculating sensitivity (# of Seraseq
141 fusions identified)/(14 possible Seraseq fusions), and precision (# of Seraseq fusions

142 identified)/(# of total fusions identified) prior to filtering or known fusion list comparison. The
143 undiluted reference standard with consensus of at least two callers, had a sensitivity of 100% and
144 precision of 36.36%. Inclusion of the knowledgebase filtering step reduced the sensitivity to
145 85.71% while increasing the precision to 77.42%, and the known fusion list rescue step increased
146 sensitivity to 100% and precision to 80% (**ADDITIONAL FILE 1: TABLE S5, FIGURE S6A**). By increasing
147 the consensus requirement to three callers, rather than just two, the prefiltered sensitivity was
148 100% and precision was 93.33%. Inclusion of the filtering step reduced the sensitivity to 85.71%
149 while increasing the precision to 100%, and known fusion list rescue increased sensitivity to
150 100% and precision to 100% (**TABLE 2; ADDITIONAL FILE 1: FIGURE S6A**). The inclusion of the known
151 fusion list prevented the removal of known Seraseq fusions, due to too few reads of evidence or
152 number of callers providing support, as well as a single Seraseq fusion, *EML4-ALK*, which was
153 present at an artificially high frequency in our database (24.7%) due to false positive calls by
154 FusionCatcher. Implementation of the known fusion list led to sensitivity scores of 100% for both
155 levels of caller consensus. The individual fusion detection algorithms ranged in sensitivity and
156 precision, and while certain algorithms are able to maintain high levels of sensitivity in addition to
157 moderate levels of precision, such as STAR-Fusion (sensitivity = 100%, precision = 43.75%),
158 others such as FusionCatcher (sensitivity = 92.86%, precision = 4.34%) and CICERO (sensitivity =
159 100%, precision 1.06%) had high levels of sensitivity with very low precision levels (**TABLE 2;**
160 **ADDITIONAL FILE 2: TABLE S5**). When considering the overall results from undiluted and serial
161 dilutions of the reference standard, the required overlap of at least three callers, with filtering and
162 utilization of the known fusion list, led to significantly fewer total fusions identified compared to
163 two consensus callers ($p = 1.86E-07$)(**TABLE 2; ADDITIONAL FILE 1: FIGURE S6B, TABLE S6**). The
164 ensemble pipeline results obtained from various reference standard dilutions, with a minimum of
165 three callers in consensus, using filtering and known fusion list rescue are shown (**FIGURE 1B;**

166 **ADDITIONAL FILE 2: TABLE S5**). The optimized ensemble pipeline, consisting of a consensus of three
167 callers, filtering, and the known fusion list, maintained high levels of sensitivity, (at least 90.48%),
168 while maintaining 100% precision as low as the 1:50 dilution of the reference standard
169 (**ADDITIONAL FILE 2: TABLE S5**). In addition to the high levels of sensitivity and precision, the total
170 number of fusions identified by this optimized ensemble pipeline in undiluted and diluted samples
171 was significantly fewer than the number identified by individual fusion detection algorithms,
172 including STAR-Fusion ($p = 1.77E-12$), CICERO ($p = 3.39E-14$) and FusionCatcher ($p = 1.00E-$
173 08)(**ADDITIONAL FILE 1, TABLE S6**). These results highlights the removal of false positive fusions,
174 which includes artifactual and benign fusion events, and subsequent reduction in manual
175 evaluation requirements (**ADDITIONAL FILE 1: FIGURE S6C,D**). Notably, we only considered the 14
176 Seraseq synthetic fusions as true positives. While fusions may exist within the GM24385 cell line,
177 in the optimized ensemble approach all of these fusions were filtered out due to either high
178 frequency across our cohort or supporting read evidence below our minimum threshold,
179 suggesting that they are likely to be artifactual in nature.

180

181 *Implementation of the ensemble approach on an in-house pediatric cancer and hematologic disease*
182 *cohort*

183 Having demonstrated the efficacy of the optimized ensemble fusion detection pipeline
184 using synthetic fusion samples, we further evaluated the utility of the pipeline on RNA-Seq data
185 obtained from 229 patient samples, obtained from three prospective pediatric cancer and
186 hematologic disease studies at Nationwide Children's Hospital (NCH) (**ADDITIONAL FILE 1: FIGURE**
187 **S2**). Our ensemble pipeline identified significantly fewer total predicted fusions post-filtering,
188 compared to all other single callers (**FIGURE 2A; ADDITIONAL FILE 1: TABLE S7**). Applying the known
189 fusion list rescue altered the average number of fusions identified overall, as an average of 3.88

190 fusions per case were identified by 3 or more callers, while an average of 3.93 fusions were
191 identified by 3 or more callers after applying the known fusion list; a total of 11 fusions were
192 rescued by this approach, of which 1 (*KIAA1549-BRAF*; **ADDITIONAL FILE 3: TABLE S8**) was clinically
193 relevant . The retained *KIAA1549-BRAF* fusion was identified by three callers, but was initially
194 filtered out due to too few reads of evidence, possibly due to either low expression, low tumor
195 cellularity or clonality (**FIGURE 2D**). In total, 67 clinically relevant fusions, identified in 67 different
196 cases, (33 CNS, 7 heme, and 27 solid tumor; **ADDITIONAL FILE 1: FIGURE S7**) were discovered using
197 the optimized ensemble pipeline with automated filtering, including the known fusion list feature,
198 and a consensus of three callers (29.3% of tumors contained a clinically relevant fusion).
199 Regardless of source material, there was a roughly a 30% yield; with clinically relevant fusion
200 identification in 44 of 148 frozen samples (30% yield), 19 of 68 FFPE samples (28% yield), and 4
201 of 13 other samples (31% yield), which included blood, cerebral spinal fluid, or bone marrow
202 (**ADDITIONAL FILE 1: FIGURE S7**). No single fusion detection algorithm was able to identify all 67
203 fusions. While JAFFA was the most sensitive algorithm, identifying the most clinically relevant
204 fusions (64 out of 67), it also had one of the highest average numbers of fusions identified per
205 sample, 1409 fusions, indicating a large number of likely false positives (**FIGURE 2B; ADDITIONAL**
206 **FILE 1: TABLE S7**). Identified fusions were broken down into 4 types: Interchromosomal Chimeric
207 (n= 30), Intrachromosomal Chimeric (n= 29), Loss of Function (n= 3), and Promoter Swapping (n=
208 5)(**FIGURE 2C**). Of the 67 clinically relevant fusions, seven were considered novel events, defined as
209 a gene fusion involving two partners not previously described in the literature at the time of
210 identification (**FIGURE 2D**). Of the 67 fusions detected, 40 (60%) were identified by all seven
211 callers, 55 (82%) were identified by ≥ 6 callers, 60 (90%) were identified by ≥ 5 callers, 64 (96%)
212 were identified by ≥ 4 callers, and 67 (100%) were identified by ≥ 3 callers. (**FIGURE 2E**). One
213 sample experienced an unresolvable failure of FusionMap, likely due to high sequencing read

214 number. Results from the remaining callers, which successfully completed for this sample, were
215 still included in our analysis. These results highlight the ability of the optimized ensemble
216 approach to identify gene fusions with a high level of confidence and a reduced number of false
217 positive predictions, while preventing over-filtering by comparison to a list of known pathogenic
218 fusions.

219

220 *Clinical Impact of Fusion Prediction*

221 *An RBPMS-MET fusion in an infantile fibrosarcoma-like tumor*

222 A female infant presented with a congenital tumor of the right face. Histologically, the
223 tumor consisted of variably cellular fascicles of spindle cells with a nonspecific
224 immunohistochemical staining profile, suspicious for infantile fibrosarcoma. However, the tumor
225 was negative for an *ETV6-NTRK3* fusion, one of the defining features of infantile fibrosarcoma [28].
226 RNA-Seq of the primary tumor and optimized ensemble pipeline analysis revealed an *RBPMS-MET*
227 fusion as the only consensus call. By contrast, the individual callers identified numerous fusions as
228 follows: Arriba: 16, CICERO: 2142, FusionMap: 29, FusionCatcher: 3907, JAFFA: 1130, MapSplice:
229 18, and STAR-Fusion: 20 (**FIGURE 3A, ADDITIONAL FILE 3: TABLE S8**). *RBPMS*, an RNA-binding
230 protein, and *MET*, a proto-oncogene receptor tyrosine kinase, have been identified as fusion
231 partners in a variety of cancers with other genes and as gene fusion partners in a patient with
232 cholangiocarcinoma [29]. Although *MET* fusions are uncommon drivers of sarcoma [30], a *TFG-*
233 *MET* fusion has been reported in a patient with an infantile spindle cell sarcoma with neural
234 features [31-33]. The interchromosomal in-frame fusion of *RBPMS* (NM_006867, exon 5) to *MET*
235 (NM_000245, exon 15) juxtaposes the RNA recognition motif of RBPMS to the MET tyrosine kinase
236 catalytic domain (**FIGURE 3B,C**). Given the therapeutic implications of this driver fusion, the fusion
237 was confirmed and reported in the patient's medical record. The identification of this fusion

238 provided the molecular driver for this tumor, which enabled definitive classification as an infantile
239 fibrosarcoma-like tumor with a *MET* fusion. The patient was initially treated with VAC (vincristine,
240 actinomycin D, and cyclophosphamide) chemotherapy which reduced tumor burden. Surgical
241 resection of the mass was performed with positive margins. Given the presence of a targetable
242 gene fusion, the presence of residual tumor, and the morbidity associated with additional surgery
243 or radiation, the patient was subsequently treated with the *MET* inhibitor cabozantinib and
244 demonstrated a complete pathological response (**FIGURE 3D**).

245

246 *An NTRK1 fusion in an infiltrating glioma/astrocytoma*

247 A 6-month-old female was diagnosed with an infiltrating glioma/astrocytoma, with a
248 mitotic index of 7 per single high-power field (HPF) and a Ki-67 labeling index averaging nearly
249 20%, indicative of aggressive disease. RNA-Seq of the primary tumor revealed a *BCAN-NTRK1*
250 fusion, identified by five callers as the only consensus fusion output from the optimized ensemble
251 pipeline (**FIGURE 4A**). This fusion was clinically confirmed by RT-PCR as an in-frame event,
252 resulting from an intrachromosomal deletion of 225kb at 1q23.1, which juxtaposes *BCAN*
253 (NM_021948, exon 6) to *NTRK1* (NM_002529, exon 8) (**FIGURE 4B,C**). This fusion results in the loss
254 of the ligand binding domain of *NTRK1*, while retaining the tyrosine kinase catalytic domain,
255 leading to a predicted activation of downstream targets in a ligand-independent manner [34].
256 Comparison of the normalized read counts from RNA-Seq data revealed elevated *NTRK1*
257 expression, over 7 standard deviations from the mean, relative to *NTRK1* expression for CNS
258 tumors within the NCH cohort (N=138) (**FIGURE 4D**). This result indicates the use of first
259 generation TRK inhibitor therapies, with recent regulatory approvals, that have exemplary
260 response rates (75%) and are generally well tolerated by patients [34]. Although the patient has
261 no evidence of disease following gross total resection and treatment with conventional

262 chemotherapy, TRK inhibitors may be clinically indicated in the setting of progressive disease
263 given these findings (**FIGURE 4E**).

264

265 *Novel BRAF fusion in a mixed neuronal-gliial tumor*

266 A 14-year-old male with a lower brainstem tumor was diagnosed with a low-grade mixed
267 neuronal-gliial tumor of unusual morphologic appearance. Tumor histology had features of both
268 ganglioglioma and pilocytic astrocytoma. This tumor was negative for the somatic variant *BRAF*
269 p.V600E, one of the most common somatic alterations associated with gangliogliomas and
270 pilocytic astrocytomas [35]. Both the ganglioglioma and pilocytic astrocytoma-like portions of the
271 primary tumor were studied separately by RNA-Seq. A novel *TRIM22-BRAF* fusion was identified
272 in both histologies of the tumor, with fusion overlap results from the ganglioglioma portion
273 represented in **FIGURE 5A**. *TRIM22-BRAF* was the only consensus fusion output by the optimized
274 fusion detection pipeline, and was clinically confirmed by RT-PCR. *TRIM22* and *BRAF* are novel
275 fusion partners; however, *TRIM22* has been reported with other fusion partners in head/neck
276 squamous cell carcinoma [36]. *BRAF* is a known oncogene that activates the RAS-MAPK signaling
277 pathways, and has been described with numerous fusion partners, including the common
278 *KIAA1549-BRAF* fusion in pediatric low-grade gliomas [35]. This fusion is an interchromosomal
279 translocation occurring between *TRIM22* (NM_006074, exon 2) at 11p15.4 and *BRAF*
280 (NM_004333, exon 9) at 7q34. The resulting protein includes the TRIM22 zinc finger domains and
281 the BRAF tyrosine kinase domain (**FIGURE 5B,C**). The *TRIM22-BRAF* fusion may lead to constitutive
282 dimerization and activation of BRAF kinase domain, which is indicated by single sample Gene Set
283 Enrichment Analysis (ssGSEA) and is theoretically targetable through RAF, MEK, or mTOR
284 inhibitors (**FIGURE 5D,E**).

285

286 **DISCUSSION**

287 Fusions play a significant role as common oncogenic drivers of pediatric cancers, and their
288 identification may refine diagnosis, inform prognosis, or indicate potential response to
289 molecularly targeted therapies. We have developed an optimized pipeline for fusion detection that
290 harmonizes results from several fusion calling algorithms, filters the output to remove known
291 false positive results, and evaluates the detected fusions compared to a list of known pathogenic
292 fusions. Testing this pipeline on a reference standard indicated that it outperforms single fusion
293 detection algorithms by reducing the number of false positive calls, producing a smaller number of
294 fusions prioritized by the strength of supporting evidence, and suitable for manual inspection. As
295 such, our pipeline greatly simplifies the interpretation process, enabling our multidisciplinary
296 oncology teams to focus on medically relevant findings.

297 We tested the optimized ensemble pipeline in a prospective study of 229 pediatric cancer
298 and hematologic disease cases and identified 67 fusions. Of these, the fusions from 50 patients
299 were selected for clinical confirmation by an orthogonal method, in our CAP-accredited, CLIA-
300 validated clinical laboratory. All 50 (100% true positive rate) were confirmed to be true fusion
301 events, and were determined to be of clinical relevance by our multidisciplinary care team,
302 providing a diagnostic yield of over 29% across the cohort. (**ADDITIONAL FILE 3: TABLE S8**). Given
303 the high number of putative fusions observed with any single caller, it can be difficult to manually
304 identify a pathogenic fusion amongst a list of tens, if not hundreds, of output fusions. By taking
305 into consideration the frequency in which each fusion occurs in an internal database, as well as the
306 level of evidence based on the number of callers and number of supporting reads by each caller,
307 one can more confidently remove false positives and identify relevant fusions. While our approach
308 does not remove the necessity of manual curation, which is required to determine true clinical
309 relevance of a fusion, it is able to drastically reduce the number of fusions that must be manually

310 assessed, down to ~4 fusions per case, and provides annotations, including a pathogenicity gene
311 partner score, to ease manual interpretation efforts. Our fully automated pipeline aids in
312 prioritization, filtering, and subsequent knowledge-based analysis, providing a more streamlined
313 and less labor-intensive approach to identify fusions, compared to current fusion identification
314 methodologies, drastically reducing the manual workload required to sort through unfiltered or
315 unprioritized results.

316 The most frequent fusion identified within our pediatric cancer cohort was *KIAA1549-BRAF*
317 (n=12, frequency= 5.2%; **FIGURE 2B**)[17]. This fusion is characteristically found in pilocytic
318 astrocytomas, which comprise 8.7% of our pediatric cancer cohort (20 out of 229 cases)[37]. We
319 identified five different sets of *KIAA1549-BRAF* breakpoints within our cohort (**ADDITIONAL FILE 1:**
320 **FIGURE S8A**). The most common fusion patterns represented in the literature are *KIAA1549* exon
321 16-*BRAF* exon 9 (16-9) or *KIAA1549* exon 15-*BRAF* exon 9 (15-9), and these two breakpoints
322 represent 9 of the 12 *KIAA1549-BRAF* fusions we identified (**ADDITIONAL FILE 1: FIGURE S8B**) [38,
323 39]. Three additional previously described sets of breakpoints were also identified, *KIAA1549*
324 exon 16-*BRAF* exon 11 (16-11; n=1), *KIAA1549* exon 15-*BRAF* exon 11 (15-11; n=1), and *KIAA1549*
325 exon 13-*BRAF* exon 9 (13-9; n=1; **ADDITIONAL FILE 1: FIGURE S8**). While the 16-11 and 15-11
326 breakpoints occur less frequently than 16-9 or 15-9, they have been well described in the
327 literature [38]; whereas only a single case with 13-9 breakpoints was reported as part of a
328 pilocytic astrocytoma cohort study [40]. *KIAA1549-BRAF* fusions often have low levels of
329 expression, a phenomenon that has been described in the literature and is associated with
330 difficulties in its identification through RNA-Seq based methodologies, which lack fusion product
331 amplification [41]. The ability of the ensemble pipeline to identify *KIAA1549-BRAF* fusions, and
332 others that have very low levels of expression, highlights the sensitivity of our approach.
333 Additionally, a supplementary “singleton” file for fusions that are identified by individual

334 algorithms and on the known fusion list is also output by our approach, allowing users the
335 opportunity to manually interpret singleton results. This approach ensures that fusions on the
336 known fusion list are retained, even with minimal evidence by a single caller.

337 Our approach has also identified other fusions commonly associated with pediatric cancer,
338 including *EWSR1-FLI1* (n=9), *FGFR1-TACC1* (n=3), *PAX3-FOXO1* (n=3), *C11orf95-RELA* (n=2),
339 *COL3A1-PLAG1* (n=2), and *NPM1-ALK* (n=2) (**FIGURE 2B**). In addition to common fusions, our
340 ensemble pipeline also identified seven novel fusions (**FIGURE 2B**). Five of the seven novel fusions
341 were confirmed by an orthogonal assay in our clinical lab (**ADDITIONAL FILE 3: TABLE S8**). Chimeric
342 fusions, which include both interchromosomal (n=30) and intrachromosomal (n=29) events, were
343 the most common type of fusion identified within the cohort, however, 5 promoter swapping and
344 3 loss of function fusions were also identified, highlighting the range of fusions this approach is
345 able to detect (**FIGURE 2D**).

346 Running seven different fusion callers is computationally complex, as each has its own set
347 of dependencies and environmental requirements. To overcome this, we utilize modern cloud
348 computing technologies. Most notable, our entire pipeline has been built in an AWS serverless
349 environment, removing the requirement for high performance computing (HPC) clusters, while
350 producing highly reproducible results and enabling pipeline sharing. The use of a serverless
351 environment provides flexibility to deploy and scale applications regardless of the application's
352 size, without needed concern for the underlying infrastructure. We are also leveraging containers
353 to process the data within the serverless environment, as they can be easily utilized by outside
354 institutions with little to no adjustment to their own environments. Another benefit to the current
355 structure of our approach is the ability to assess output from the individual algorithms in real
356 time, as the ensemble pipeline is automatically run after each individual caller completes, allowing
357 for interpretation of at least 3 of the 7 callers within ~3.5 hours, which can be beneficial in

358 situations that necessitate fast turnaround times (**ADDITIONAL FILE 1: FIGURE S5**). Overall, our novel
359 use of serverless technology provides a robust computational solution that is fully automatable
360 and easy to distribute.

361 There are numerous benefits to the utilization of this optimized pipeline, in that detected
362 fusion events are agnostic to gene partner, allowing identification of common, rare and novel
363 fusions. In addition, the RNA-Seq data set can be utilized for other types of downstream and
364 correlative analyses, including evaluation of gene expression for loci disrupted by the fusion
365 (**FIGURE 4D**). Utilization of cohort data to assess outlier gene expression can provide valuable
366 insights into pathway disruptions that may occur due to the gene fusion (**FIGURE 5D**) and may
367 provide information about disease subtyping.

368 Our ensemble fusion detection pipeline is customizable, allowing users to select how many
369 and which callers to deploy. This may impact potential cost savings, time-to-result, or permit
370 customization that eliminates specific callers that require excessive compute requirements or run
371 times, as suitable in a clinical diagnostic or research setting. Users can also determine the number
372 of consensus calls required to support fusion prediction, which can reduce the number of fusions
373 to assess manually. Callers with a higher percentage of false positives, FusionCatcher and JAFFA,
374 often overlap in their predictions, leading to an increased average number of fusions output by the
375 ensemble pipeline with a consensus of only two callers; a problem diminished by requiring
376 predictions from at least three callers to overlap. In our study, precision was found to be highest in
377 the three-caller consensus version of the ensemble pipeline (**TABLE 2; ADDITIONAL FILE 2: TABLE**
378 **S5**). Another benefit to utilizing different algorithms is the ability to assess supplementary output
379 data, in addition to traditional fusion calling. We have made use of this through the inclusion of the
380 internal tandem duplication (ITD) detection which is performed by CICERO. CICERO has identified

381 7 clinically relevant ITDs within our cohort, 4 of which we have confirmed using orthogonal assays
382 **(ADDITIONAL FILE 1: TABLE S9)**.

383 Future developments to the pipeline could include a weighting system for each caller,
384 based on the precision and sensitivity of the algorithm and on which callers have overlapping
385 predictions, leading to a more sophisticated prioritization strategy. Additional fusion calling
386 algorithms may also be considered and provided as options for users. The known fusion list can
387 also be modified and tailored to include specific gene pairs, or even single genes of interest,
388 providing another layer of customization. Importantly, through the utilization of a proper internal
389 database for frequency filtering purposes, considering age and/or cancer diagnosis, and with the
390 deployment of the appropriate known fusion list, the ensemble approach could be readily
391 implemented in adult cancer fusion detection. Lastly, not all predictors performed equally, and
392 there was a single unresolvable failure of FusionMap to complete. This failure was likely due to the
393 sequencing depth of the sample, however further analysis is required to determine whether
394 parameter modification would permit completion of FusionMap in this case **(ADDITIONAL FILE 3:
395 TABLE S8)**. Importantly, our approach was able to circumvent this failure due to the multi-caller
396 nature of the pipeline. Lastly, there are many modalities of RNA-seq analysis that may be
397 harnessed in future developments of the ensemble fusion detection pipeline, which may include
398 an integrative approach exploiting expression-based analysis and ranking. In summary, the
399 ensemble pipeline provides a highly customizable approach to fusion detection that can be applied
400 to numerous settings, with opportunities for future improvements based on additional features
401 and applications.

402

403 **Conclusions:**

404 The optimized ensemble fusion detection pipeline provides a highly automated and
405 accurate approach to fusion detection, developed to identify high confidence gene fusions from
406 RNA-Seq data produced from pediatric cancer and hematologic disease samples, and could be
407 readily implemented in adult cancer data analysis. The clinical impact of accurately identifying
408 gene fusions in a given patient's tumor sample is undeniable, not only in terms of refining
409 diagnoses but also in terms of providing prognostic information that shapes treatment decisions.
410 Furthermore, identification of driver fusions may indicate potential response to targeted therapies
411 for cancer patients. The code for the overlap algorithm utilized in this study is publicly available at
412 our GitHub page (<https://github.com/nch-igm/nch-igm-ensemble-fusion-detection>).

413

414 **METHODS**

415 *Description of an internal patient cohort*

416 In total, 229 patients were consented and enrolled onto one of three Institutional Review
417 Board (IRB) approved protocols (IRB17-00206, IRB16-00777, IRB18-00786) and studied at the
418 Institute for Genomic Medicine (IGM) at Nationwide Children's Hospital (NCH) in Columbus, Ohio.
419 Through the utilization of genomic and transcriptomic profiling, these protocols aim to refine
420 diagnosis and prognosis, detect germline cancer predisposition, identify targeted therapeutic
421 options, and/or to determine eligibility for clinical trials in patients with rare, treatment-
422 refractory, relapsed, pediatric cancers or hematologic diseases, or with epilepsy arising in the
423 setting of a low grade central nervous system (CNS) cancer. Our in-house NCH cohort as studied
424 here, consisted of samples from CNS tumors (n=138), hematologic diseases (n=18), and non-CNS
425 solid tumors (n=73), as represented in **ADDITIONAL FILE 1: FIGURE S2**.

426

427 *RNA-Seq of patient tissues*

428 RNA was extracted from snap frozen tissue, formalin-fixed paraffin-embedded (FFPE)
429 tissue, peripheral blood, bone marrow, and cerebral spinal fluid utilizing dual RNA and DNA co-
430 extraction methods originally developed by our group for The Cancer Genome Atlas project [42].
431 White blood cells were isolated from peripheral blood or bone marrow using the lymphocyte
432 separation medium Ficoll-histopaque. Frozen tissue, white blood cells, or pelleted cells from
433 cerebrospinal fluid were homogenized in Buffer RLT, with beta-Mercaptoethanol to denature
434 RNases, plus Reagent DX and separated on an AllPrep (Qiagen) DNA column to remove DNA for
435 subsequent RNA steps. The eluate was processed for RNA extraction using acid-phenol:chloroform
436 (Sigma) and added to the mirVana miRNA (Applied Biosystems) column, washed, and RNA was
437 eluted using DEPC-treated water (Ambion). DNase treatment (Zymo) was performed post RNA
438 purification. FFPE tissues were deparaffinized using heptane/methanol (VWR) and lysed with
439 Paraffin Tissue Lysis Buffer and Proteinase K from the HighPure miRNA kit (Roche). The sample
440 was pelleted to remove the DNA, and the supernatant was processed for RNA extraction with the
441 HighPure miRNA column, followed by DNase treatment (Qiagen). RNA quantification was
442 performed with Qubit (Life Sciences).

443 RNA-Seq libraries were generated using 100 ng to 1 µg of DNase-treated RNA input, either
444 by ribodepletion using the Ribo-Zero Globin kit (Illumina) followed by library construction using
445 the TruSeq Stranded RNA-Seq protocol (Illumina), or by ribodepletion with NEBNext
446 Human/Mouse/Rat rRNA Depletion kit followed by library construction using the NEBNext Ultra
447 II Directional RNA-Seq protocol (New England BioLabs). Illumina 2x151 paired end reads were
448 generated either on the HiSeq 4000 or NovaSeq 6000 sequencing platforms (Illumina). An average
449 of 104 million read pairs were obtained per sample (range 37M to 380M read pairs).

450 Following data production and post-run processing, FASTQ files were aligned to the
451 GRCh38 human reference (hg38) using STAR aligner (version 2.6.0c)[43]. Feature counts were

452 calculated using HTSeq, and normalized read counts were calculated for all samples using DESeq2
453 [44, 45]. Single sample Gene Set Enrichment Analysis (ssGSEA), v10.0.3, was performed on
454 DESeq2 normalized read counts using Molecular Signatures Database (MSigDB) Oncogenic
455 Signatures (c6.all.v7.2.symbols.gmt), which included MEK-upregulated genes (MEK_UP.V1_UP),
456 RAF-upregulated genes (RAF_UP.V1_UP), and mTOR-upregulated genes
457 (MTOR_UP.N4.V1_UP)[46].

458

459 *RNA-Seq of SeraCare control reference standards*

460 Seraseq Fusion RNA Mix (SeraCare Inc., Milford, MA) was utilized as a control reference
461 standard reagent to test and optimize the ensemble fusion detection pipeline. This product
462 contains 14 synthetic gene fusions *in vitro* transcribed, utilizing the GM24385 cell line RNA as a
463 background. RNA-Seq libraries were prepared utilizing 500 ng input of neat (undiluted) Seraseq
464 Fusion RNA v2, a non-commercially available concentrated product, as input (SeraCare). RNA-Seq
465 libraries were also prepared using 500 ng input of diluted control reference standard (Seraseq
466 Fusion RNA v3 (SeraCare)), which, as a neat reagent is roughly equivalent to a 1:25 dilution of the
467 v2 product, and of total human RNA (GM24385, Coriell) for the fusion-negative controls.
468 Concentrations of individual fusions in the control reference standard were determined by the
469 manufacturer using a custom fluorescent probe set (based on TaqMan probe design) for each
470 fusion and evaluation by droplet digital PCR. Digital PCR-based concentration data (copies/ul) are
471 available in **ADDITIONAL FILE 1: TABLE S1** for the undiluted sample and **ADDITIONAL FILE 1: TABLE S2**
472 for the diluted sample [47].

473 Dilutions of the Seraseq Fusion RNA v3 reference standard were performed by mixing with
474 control total human RNA (GM24385, Coriell) for final dilutions of 1:25, 1:50, 1:250, 1:500, 1:2500.
475 We also evaluated undiluted Seraseq Fusion RNA v2. For neat and diluted samples, 500ng input

476 RNA was treated using the NEBNext Human/Mouse/Rat rRNA Depletion kit and libraries were
477 prepared following the NEBNext Ultra II Directional RNA-Seq protocol (New England BioLabs).
478 Paired end 2x151 bp reads were produced using the HiSeq 4000 platform (Illumina). An average
479 of 149 million read pairs were obtained per Seraseq sample (range of 86M to 227M read pairs).

480

481 *Optimized Fusion Detection Pipeline*

482 Fusions were detected from paired end RNA-Seq FASTQ files utilizing an automated
483 ensemble fusion detection pipeline that employs seven fusion-calling algorithms described in
484 **TABLE 1:** Arriba (v1.2.0), CICERO (v0.3.0), FusionMap (v mono-2.10.9), FusionCatcher (v0.99.7c),
485 JAFFA (direct v1.09), MapSplice (v2.2.1), and STAR-Fusion (v1.6.0)[25, 48-51]. STAR-Fusion
486 parameters were altered to reduce the stringency setting for the fusion fragments per million
487 reads (FFPM) from 0.05 to 0.02, while default parameters were retained for all other callers. After
488 fusion calling with each independent algorithm, a custom algorithm written in the R programming
489 language, was used to “overlap,” or align and compare, the unordered gene partners identified by
490 individual fusion callers. The utilization of unordered gene partners allows for fusions to be
491 compared, even if different breakpoints were identified by individual algorithms, and to include
492 reciprocal fusions. Fusion partners identified by at least three of the seven callers are retained and
493 prioritized based on the number of contributing algorithms first and then by the number of
494 sequence reads providing evidence for each fusion. The overlap output retains annotations from
495 the individual callers, including breakpoints, distance between breakpoints, donor and acceptor
496 genes, reads of evidence, nucleotide sequence at breakpoint (if available), frequency information
497 from the database, and whether the identified fusion contains “known pathogenic fusion
498 partners”. If discordant breakpoints are identified across callers for a set of fusion partners, the

499 breakpoints with the most evidence, determined by number of supporting reads, are prioritized in
500 the output.

501 The fusions are filtered by the following steps (**FIGURE 1A**). Read-through events, which
502 occur between neighboring genes and are typically identified in both healthy and disease states,
503 are not expected to impact cellular functions [12, 24]. This type of fusion prediction is a source of
504 false positive results, so we have implemented a filter that removes fusions detected between
505 genes fewer than 200,000 bases apart, that occur on the same strand and chromosome. Recurrent
506 fusions with uncertain biological significance have also been identified in normal tissues. To
507 prevent the inclusion of commonly occurring, benign fusions in our output, a PostgreSQL database
508 was used to filter commonly occurring artifactual fusions. This filter removes any expected fusion
509 artifact with greater than a 10% frequency of detection based on our internal cohort. Lastly, to
510 ensure a high level of confidence in the identified fusions, we utilize a minimum threshold for level
511 of evidence, removing fusions that contain fewer than four reads of support from at least one
512 contributing algorithm.

513 While filtering can remove false positive results and reduces the time needed to review
514 predicted fusions, it also can remove true positive fusions in certain circumstances. To prevent the
515 inadvertent filtering of known fusions, a known fusion list was developed containing 325 pairs of
516 common fusion partners associated with cancer, as identified in COSMIC and TCGA (**ADDITIONAL**
517 **FILE 1: TABLE S3**)[27, 52]. To increase sensitivity in the identification of known pathogenic fusions,
518 fusion partners that are on the known fusion list are retained as long as at least two callers have
519 identified the fusion. The ensemble pipeline also outputs a supplementary singleton fusion file,
520 containing fusions identified by a single caller that are on the known fusion list, allowing users to
521 examine low evidence fusions that may be of interest.

522 To prioritize fusions that contain gene partners commonly found in the known fusion list,
523 we developed the “Gene Partner Predicted Pathogenicity Score” based on the frequency of the
524 individual partners in the known fusion list. Of the 325 fusions on the known fusion list, 38 genes
525 are present as a fusion partner ≥ 3 times (**ADDITIONAL FILE 1: TABLE 4, FIGURE S3**). The most
526 common partners are *BRAF* and *KMT2A*, which are present as fusion partners 28 times each. To
527 aide prediction of novel, or not well described, pathogenic fusions, we developed a score based on
528 known pathogenic gene partners. This score utilizes the frequency of partners present on the
529 known fusion list. The pathogenic frequency score ranges from 10 (most frequent) to 1 (least
530 frequent, but present at least 3 times):

$$\text{Pathogenic Frequency Score} = 10 / (f_{max} - f)$$

531 Where f is the gene frequency and f_{max} is the maximum observed frequency. The following
532 annotations are included in the ensemble results if an identified fusion contains one of the 38
533 common pathogenic gene partners: designation as a known pathogenic gene partner, inclusion of
534 the frequency score (1-10), and gene type based on UniProt description [53].

535 A knowledge-based interpretation strategy was applied to the filtered list of fusion
536 partners output by the pipeline, including the use of FusionHub [54], to inform clinical relevance,
537 such as diagnostic and/or prognostic information or a potential therapeutic target. Visual
538 assessment of the fusion events was performed by examining RNA-Seq BAM files with Integrated
539 Genome Viewer (IGV). Fusions were also assessed at the DNA level by IGV-based evaluation of
540 gene-specific paired end read alignments from ES or WGS BAM files, for potential evidence of
541 mapping discordance. Clinically relevant fusions were then assayed in our College of American
542 Pathologists (CAP)-accredited clinical laboratory using RT-PCR followed by Sanger sequencing of
543 the resulting products, and/or by Archer FusionPlex Solid Tumor panel (ArcherDx) for clinical
544 confirmation.

545

546 *AWS Implementation of the Ensemble Approach*

547 The ensemble fusion detection pipeline is implemented utilizing an Amazon Web Services
548 (AWS) serverless environment (**ADDITIONAL FILE 1: FIGURE S4**). The workflow is initiated via a call
549 to Amazon API Gateway, which passes parameters, including the location of the input FASTQ files,
550 to an AWS Lambda function. The Lambda function initiates the AWS Batch job to load and
551 executes a custom fusion detection Docker image, which launches Arriba, CICERO, FusionMap,
552 FusionCatcher, JAFFA, MapSplice, and STAR-Fusion. We utilize the R5 family of instances for the
553 fusion detection algorithms. Due to the efficiency by which different algorithms are able to multi-
554 thread, each fusion detection tool is allocated 32 virtual CPUs (vCPUs), except for CICERO which is
555 allocated 16 vCPUs and JAFFA which is allocated 8 vCPUs. Using the described allocations, Arriba
556 completes the fastest (~37 minutes) for the runs completed year to date in 2020, followed by
557 FusionMap (~1 hour 12 minutes), STAR-fusion (~3 hours 25 minutes), FusionCatcher (~10 hours
558 35 minutes), CICERO (~11 hours 49 minutes), MapSplice (~15 hours 2 minutes), and JAFFA (~27
559 hours 16 minutes), data is summarized in **ADDITIONAL FILE 1: FIGURE S5**. The results from the
560 fusion callers are sent to an AWS S3 output bucket, which invokes AWS Batch to load and execute
561 a Docker image with our overlap script upon completion. This allows for real-time examination of
562 results as each caller finishes, as the overlapping output is updated upon completion of each
563 individual caller, which is particularly advantageous given the long execution times for some of
564 the fusion callers. It is possible to examine results upon completion of the three fastest algorithms
565 within ~3.5 hours, which is of great benefit for cases necessitating fast turnaround times, and
566 complete results are made available by the next day. The overlap Docker image queries and
567 writes to an Aurora PostgreSQL database and performs all necessary filtering. The final results,
568 including annotated filtered and unfiltered fusion lists, are stored in an AWS S3 output bucket for

569 subsequent interpretation. Code for the overlap algorithm is available at our GitHub repository
570 (<https://github.com/nch-igm/nch-igm-ensemble-fusion-detection>), DOI:
571 10.5281/zenodo.3950385, and Docker images used to build the pipeline are available upon
572 request.

573

574 *Data Analysis and Statistics*

575 Figures were plotted using R version 4.0.2. Statistical analysis was performed by GraphPad
576 Prism 7.0e software. Graphical representation of fusion breakpoints and products were generated
577 using a modified version of INTEGRATE-Vis [55].

578

579 **LIST OF ABBREVIATIONS**

580 **AWS:** Amazon Web Services

581 **CNS:** Central Nervous System

582 **ES:** Exome Sequencing

583 **FDR:** False Discovery Rate

584 **FFPE:** Formalin Fixed, Paraffin Embedded

585 **FFPM:** Fusion Fragments Per Million

586 **GSNAP:** Genomic Short-read Nucleotide Alignment Program

587 **Heme:** Hematologic Diseases

588 **HPF:** High Power Field

589 **HPC:** High Performance Computing

590 **IGM:** Institute for Genomic Medicine

591 **IGV:** Integrated Genome Viewer

592 **ITD:** Internal Tandem Duplication

593 **NCH:** Nationwide Children's Hospital

594 **QC:** Quality Control

595 **RNA-Seq:** RNA-Sequencing

596 **ssGSEA:** Single Sample Gene Set Enrichment Analysis

597 **vCPU:** virtual central processing unit

598 **WGS:** Whole Genome sequencing

599

600 **Declarations:**

601 *Ethics approval and consent to participate:*

602 This study was reviewed and approved by the Institutional Review Board (IRB) of The Research
603 Institute at Nationwide Children's Hospital. Informed consent was obtained from the patients
604 and/or parents for molecular genetic analysis, which included RNA-sequencing. These protocols
605 allowed for return of results from research sequencing studies after confirmation in a CLIA-
606 certified laboratory.

607

608 *Availability of data and materials*

609 DNA and RNA sequencing data for this study has been deposited to dbGAP, accession number
610 phs001820.v1.p1

611 Seraseq fastq files, from the benchmarking studies, have been deposited to the NIH Sequence Read
612 Archive (SRA), accession number PRJNA679580.

613 Code for the overlap algorithm is available at our GitHub repository ([https://github.com/nch-](https://github.com/nch-igm/nch-igm-ensemble-fusion-detection)
614 [igm/nch-igm-ensemble-fusion-detection](https://github.com/nch-igm-ensemble-fusion-detection)), DOI: 10.5281/zenodo.3950385

615 The Docker image used to run the overlap algorithm is available upon request for running the
616 ensemble pipeline in an AWS serverless environment.

617 The Docker image used to run previously published fusion detection algorithms is also available
618 upon request for running the ensemble pipeline in an AWS serverless environment.

619

620 *Competing interests*

621 **No Competing interests:** Stephanie LaHaye, James Fitch, Kyle Voytovich, Adam Herman,
622 Benjamin Kelly, Grant Lammi, Saranga Wijeratne, Kathleen Schieffer, Natalie Bir, Sean McGrath,
623 Anthony Miller, Amy Wetzel, Katherine Miller, Tracy Bedrosian, Kristen Leraas, Ajay Gupta,
624 Bhuvana Setty, Jeffrey Leonard, Jonathan Finlay, Mohamed Abdelbaki, Diana Osorio, Selene Koo,
625 Daniel Koboldt, Vincent Magrini, Catherine Cottrell, Richard Wilson and Peter White.

626 Elaine Mardis: Qiagen N.V., supervisory board member, honorarium and stock-based
627 compensation.

628 Daniel Boué: Illumina (ILMN) share holder.

629

630 *Funding:*

631 We thank the Nationwide Children's Foundation and The Abigail Wexner Research Institute at
632 Nationwide Children's Hospital for generously supporting this body of work. These funding bodies
633 had no role in the design of the study, no role in the collection, analysis, and interpretation of data
634 and no role in writing the manuscript.

635

636 *Authors' contributions:*

637 **SL** analyzed and interpreted fusion data, contributed to development of overlap algorithm and
638 Docker images, contributed to AWS serverless workflow, and wrote the manuscript. **JF** and **KV**
639 contributed development of overlap algorithm and Docker images, contributed to AWS serverless
640 workflow, and contributed to manuscript writing. **AH**, **BJK**, **GL**, and **SW** provided data analysis

641 support, designed AWS serverless workflow, and contributed to manuscript writing. **SF**
642 contributed to manuscript revisions and oversaw, organized, and performed data upload to SRA.
643 **KMS** contributed to analysis and interpretation of RNA-Seq results and performed clinical data
644 acquisition. **KM, TAB, KL** and **DK** provided NGS analysis and interpretation for cancer cohort data.
645 **NB, SDM,** and **ARM** perform library preparations and developed laboratory procedures for RNA-
646 Seq processing/QC. **AW** managed RNA-Seq processing and analysis. **KL** managed and coordinated
647 all clinical samples. **DRB, JRL, JLF, MA, DSO, AG, BS,** and **SCK** contributed to the enrollment of
648 patients onto the NCH cancer protocols and provided clinical expertise, **DRB** and **SCK** also
649 contributed pathology materials (fixed or frozen tissues etc.) following QA and/or QC reviews of
650 enrollee pathology. **VM** oversaw technology development and contributed to the conceptual
651 design of project. **CEC, ERM,** and **RKW** developed, led, and supervised work performed on cancer
652 protocol, contributed to conceptual design of project, contributed to analysis and interpretation of
653 RNA-Seq results, and contributed to manuscript writing and revision. **PW** conceived, designed,
654 and supervised the project, oversaw and contributed to algorithm development, provided support
655 for utilization of AWS and computational resources, and contributed significantly to manuscript
656 writing and revision. All authors read and approved the final manuscript.

657

658 *Acknowledgements*

659 We thank the patients and their families for participating in our translational research protocol.
660 We thank the Nationwide Foundation Pediatric Innovation Fund for generously supporting this
661 project.

662

663

664

665 **REFERENCES**

- 666 1. Steliarova-Foucher E, Colombet M, Ries LAG, Moreno F, Dolya A, Bray F, Hesselting P, Shin
667 HY, Stiller CA, contributors I-: **International incidence of childhood cancer, 2001-10: a**
668 **population-based registry study.** *Lancet Oncol* 2017, **18**:719-731.
- 669 2. Amatu A, Sartore-Bianchi A, Siena S: **NTRK gene fusions as novel targets of cancer**
670 **therapy across multiple tumour types.** *ESMO Open* 2016, **1**:e000023.
- 671 3. Pui CH, Gajjar AJ, Kane JR, Qaddoumi IA, Pappo AS: **Challenging issues in pediatric**
672 **oncology.** *Nat Rev Clin Oncol* 2011, **8**:540-549.
- 673 4. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2016.** *CA Cancer J Clin* 2016, **66**:7-30.
- 674 5. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW: **Cancer**
675 **genome landscapes.** *Science* 2013, **339**:1546-1558.
- 676 6. Grobner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, Johann PD,
677 Balasubramanian GP, Segura-Wang M, Brabetz S, et al: **The landscape of genomic**
678 **alterations across childhood cancers.** *Nature* 2018, **555**:321-327.
- 679 7. Marshall GM, Carter DR, Cheung BB, Liu T, Mateos MK, Meyerowitz JG, Weiss WA: **The**
680 **prenatal origins of cancer.** *Nat Rev Cancer* 2014, **14**:277-289.
- 681 8. Rowley JD: **Letter: A new consistent chromosomal abnormality in chronic**
682 **myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining.**
683 *Nature* 1973, **243**:290-293.
- 684 9. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watanabe H,
685 Kurashina K, Hatanaka H, et al: **Identification of the transforming EML4-ALK fusion**
686 **gene in non-small-cell lung cancer.** *Nature* 2007, **448**:561-566.
- 687 10. Jia Y, Xie Z, Li H: **Intergenically Spliced Chimeric RNAs in Cancer.** *Trends Cancer* 2016,
688 **2**:475-484.
- 689 11. Li Y, Li Y, Yang T, Wei S, Wang J, Wang M, Wang Y, Zhou Q, Liu H, Chen J: **Clinical**
690 **significance of EML4-ALK fusion gene and association with EGFR and KRAS gene**
691 **mutations in 208 Chinese patients with non-small cell lung cancer.** *PLoS One* 2013,
692 **8**:e52093.
- 693 12. Dupain C, Harttrampf AC, Urbinati G, Georger B, Massaad-Massade L: **Relevance of**
694 **Fusion Genes in Pediatric Cancers: Toward Precision Medicine.** *Mol Ther Nucleic Acids*
695 2017, **6**:315-326.
- 696 13. Bernt KM, Hunger SP: **Current concepts in pediatric Philadelphia chromosome-**
697 **positive acute lymphoblastic leukemia.** *Front Oncol* 2014, **4**:54.
- 698 14. Hawkins C, Walker E, Mohamed N, Zhang C, Jacob K, Shirinian M, Alon N, Kahn D, Fried I,
699 Scheinemann K, et al: **BRAF-KIAA1549 fusion predicts better clinical outcome in**
700 **pediatric low-grade astrocytoma.** *Clin Cancer Res* 2011, **17**:4790-4798.
- 701 15. Park SH, Won J, Kim SI, Lee Y, Park CK, Kim SK, Choi SH: **Molecular Testing of Brain**
702 **Tumor.** *J Pathol Transl Med* 2017, **51**:205-223.
- 703 16. Yuan L, Liu ZH, Lin ZR, Xu LH, Zhong Q, Zeng MS: **Recurrent FGFR3-TACC3 fusion gene in**
704 **nasopharyngeal carcinoma.** *Cancer Biol Ther* 2014, **15**:1613-1621.
- 705 17. Jones DT, Kocialkowski S, Liu L, Pearson DM, Backlund LM, Ichimura K, Collins VP: **Tandem**
706 **duplication producing a novel oncogenic BRAF fusion gene defines the majority of**
707 **pilocytic astrocytomas.** *Cancer Res* 2008, **68**:8673-8677.
- 708 18. Morris SW, Kirstein MN, Valentine MB, Dittmer K, Shapiro DN, Look AT, Saltman DL:
709 **Fusion of a kinase gene, ALK, to a nucleolar protein gene, NPM, in non-Hodgkin's**
710 **lymphoma.** *Science* 1995, **267**:316-317.

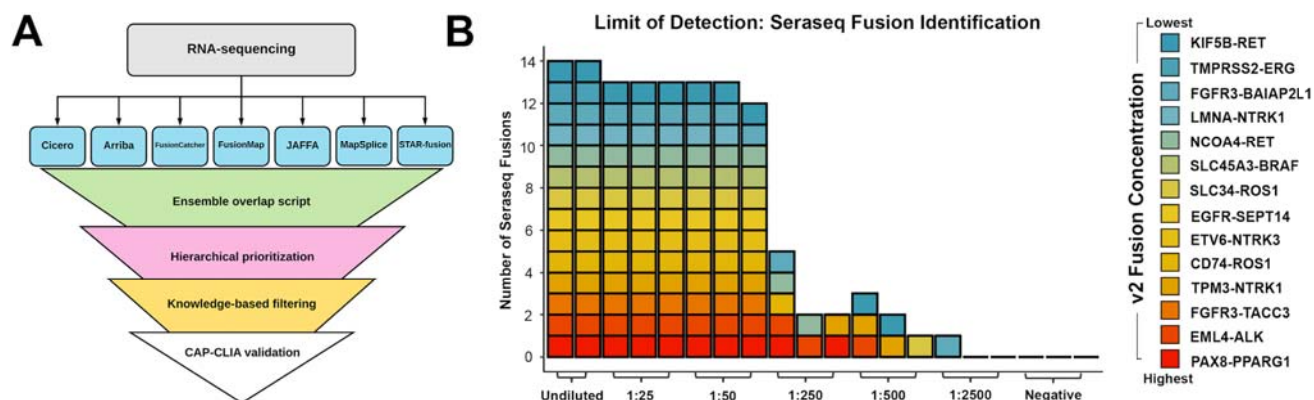
- 711 19. Mosse YP, Lim MS, Voss SD, Wilner K, Ruffner K, Laliberte J, Rolland D, Balis FM, Maris JM,
712 Weigel BJ, et al: **Safety and activity of crizotinib for paediatric patients with refractory**
713 **solid tumours or anaplastic large-cell lymphoma: a Children's Oncology Group phase**
714 **1 consortium study.** *Lancet Oncol* 2013, **14**:472-480.
- 715 20. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-
716 Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al: **Discovery and genotyping of**
717 **structural variation from long-read haploid genome sequence data.** *Genome Res* 2017,
718 **27**:677-685.
- 719 21. Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, Garvin T, Fang H,
720 Gurtowski J, Hutton E, et al: **Complex rearrangements and oncogene amplifications**
721 **revealed by long-read DNA and RNA sequencing of a breast cancer cell line.** *Genome*
722 *Res* 2018, **28**:1126-1135.
- 723 22. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T,
724 Palanisamy N, Chinnaiyan AM: **Transcriptome sequencing to detect gene fusions in**
725 **cancer.** *Nature* 2009, **458**:97-101.
- 726 23. Wang Q, Xia J, Jia P, Pao W, Zhao Z: **Application of next generation sequencing to human**
727 **gene fusion detection: computational tools, features and perspectives.** *Brief Bioinform*
728 2013, **14**:506-519.
- 729 24. He Y, Yuan C, Chen L, Lei M, Zellmer L, Huang H, Liao DJ: **Transcriptional-Readthrough**
730 **RNAs Reflect the Phenomenon of "A Gene Contains Gene(s)" or "Gene(s) within a**
731 **Gene" in the Human Genome, and Thus Are Not Chimeric RNAs.** *Genes (Basel)* 2018, **9**.
- 732 25. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A: **Accuracy assessment of fusion**
733 **transcript detection via read-mapping and de novo fusion transcript assembly-based**
734 **methods.** *Genome Biol* 2019, **20**:213.
- 735 26. Liu S, Tsai WH, Ding Y, Chen R, Fang Z, Huo Z, Kim S, Ma T, Chang TY, Priedigkeit NM, et al:
736 **Comprehensive evaluation of fusion transcript detection algorithms and a meta-**
737 **caller to combine top performing methods in paired-end RNA-seq data.** *Nucleic Acids*
738 *Res* 2016, **44**:e47.
- 739 27. Gao Q, Liang WW, Foltz SM, Mutharasu G, Jayasinghe RG, Cao S, Liao WW, Reynolds SM,
740 Wyczalkowski MA, Yao L, et al: **Driver Fusions and Their Implications in the**
741 **Development and Treatment of Human Cancers.** *Cell Rep* 2018, **23**:227-238 e223.
- 742 28. Church AJ, Calicchio ML, Nardi V, Skalova A, Pinto A, Dillon DA, Gomez-Fernandez CR,
743 Manoj N, Haimen JD, Stahl JA, et al: **Recurrent EML4-NTRK3 fusions in infantile**
744 **fibrosarcoma and congenital mesoblastic nephroma suggest a revised testing**
745 **strategy.** *Mod Pathol* 2018, **31**:463-473.
- 746 29. Consortium ITP-CAoWG: **Pan-cancer analysis of whole genomes.** *Nature* 2020, **578**:82-
747 93.
- 748 30. Stransky N, Cerami E, Schalm S, Kim JL, Lengauer C: **The landscape of kinase fusions in**
749 **cancer.** *Nat Commun* 2014, **5**:4846.
- 750 31. International Cancer Genome Consortium PedBrain Tumor P: **Recurrent MET fusion**
751 **genes represent a drug target in pediatric glioblastoma.** *Nat Med* 2016, **22**:1314-1320.
- 752 32. Torre M, Jessop N, Hornick JL, Alexandrescu S: **Expanding the spectrum of pediatric**
753 **NTRK-rearranged fibroblastic tumors to the central nervous system: A case report**
754 **with RBPMS-NTRK3 fusion.** *Neuropathology* 2018, **38**:624-630.
- 755 33. Flucke U, van Noesel MM, Wijnen M, Zhang L, Chen CL, Sung YS, Antonescu CR: **TFG-MET**
756 **fusion in an infantile spindle cell sarcoma with neural features.** *Genes Chromosomes*
757 *Cancer* 2017, **56**:663-667.

- 758 34. Cocco E, Scaltriti M, Drilon A: **NTRK fusion-positive cancers and TRK inhibitor therapy.**
759 *Nat Rev Clin Oncol* 2018, **15**:731-747.
- 760 35. Pekmezci M, Villanueva-Meyer JE, Goode B, Van Ziffle J, Onodera C, Grenert JP, Bastian BC,
761 Chamyam G, Maher OM, Khatib Z, et al: **The genetic landscape of ganglioglioma.** *Acta*
762 *Neuropathol Commun* 2018, **6**:47.
- 763 36. Yoshihara K, Wang Q, Torres-Garcia W, Zheng S, Vegesna R, Kim H, Verhaak RG: **The**
764 **landscape and therapeutic relevance of cancer-associated transcript fusions.**
765 *Oncogene* 2015, **34**:4845-4854.
- 766 37. Bar EE, Lin A, Tihan T, Burger PC, Eberhart CG: **Frequent gains at chromosome 7q34**
767 **involving BRAF in pilocytic astrocytoma.** *J Neuropathol Exp Neurol* 2008, **67**:878-887.
- 768 38. Lin A, Rodriguez FJ, Karajannis MA, Williams SC, Legault G, Zagzag D, Burger PC, Allen JC,
769 Eberhart CG, Bar EE: **BRAF alterations in primary glial and glioneuronal neoplasms of**
770 **the central nervous system with identification of 2 novel KIAA1549:BRAF fusion**
771 **variants.** *J Neuropathol Exp Neurol* 2012, **71**:66-72.
- 772 39. Yamashita S, Takeshima H, Matsumoto F, Yamasaki K, Fukushima T, Sakoda H, Nakazato M,
773 Saito K, Mizuguchi A, Watanabe T, et al: **Detection of the KIAA1549-BRAF fusion gene in**
774 **cells forming microvascular proliferations in pilocytic astrocytoma.** *PLoS One* 2019,
775 **14**:e0220146.
- 776 40. Jones DT, Hutter B, Jager N, Korshunov A, Kool M, Warnatz HJ, Zichner T, Lambert SR,
777 Ryzhova M, Quang DA, et al: **Recurrent somatic alterations of FGFR1 and NTRK2 in**
778 **pilocytic astrocytoma.** *Nat Genet* 2013, **45**:927-932.
- 779 41. Tian L, Li Y, Edmonson MN, Zhou X, Newman S, McLeod C, Thrasher A, Liu Y, Tang B, Rusch
780 MC, et al: **CICERO: a versatile method for detecting complex and diverse driver**
781 **fusions using cancer RNA sequencing data.** *Genome Biol* 2020, **21**:126.
- 782 42. Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, Liu Y, Fan H, Shen H,
783 Ravikumar V, et al: **A Comprehensive Pan-Cancer Molecular Study of Gynecologic and**
784 **Breast Cancers.** *Cancer Cell* 2018, **33**:690-705 e699.
- 785 43. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras
786 TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29**:15-21.
- 787 44. Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-throughput**
788 **sequencing data.** *Bioinformatics* 2015, **31**:166-169.
- 789 45. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for**
790 **RNA-seq data with DESeq2.** *Genome Biol* 2014, **15**:550.
- 791 46. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A,
792 Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a**
793 **knowledge-based approach for interpreting genome-wide expression profiles.** *Proc*
794 *Natl Acad Sci U S A* 2005, **102**:15545-15550.
- 795 47. **Seraseq Tumor Fusion RNA Mix3** [[https://www.seracare.com/globalassets/seracare-
796 resources/pr-0710-0431-seraseq-tumor-fusion-rna-mix-v3-10330722.pdf](https://www.seracare.com/globalassets/seracare-resources/pr-0710-0431-seraseq-tumor-fusion-rna-mix-v3-10330722.pdf)]
- 797 48. Ge H, Liu K, Juan T, Fang F, Newman M, Hoeck W: **FusionMap: detecting fusion genes**
798 **from next-generation sequencing data at base-pair resolution.** *Bioinformatics* 2011,
799 **27**:1922-1928.
- 800 49. Davidson NM, Majewski IJ, Oshlack A: **JAFFA: High sensitivity transcriptome-focused**
801 **fusion gene detection.** *Genome Med* 2015, **7**:43.
- 802 50. Nicorici D, Satalan M, Edgren H, Kangaspeska s, Murumagi A, Kallioniemi o, Virtanen S,
803 Kilkku O: **FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-**
804 **sequencing data.** *bioRxiv* 2014.

- 805 51. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA,
806 Perou CM, et al: **MapSplice: accurate mapping of RNA-seq reads for splice junction**
807 **discovery**. *Nucleic Acids Res* 2010, **38**:e178.
- 808 52. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore
809 C, Dawson E, et al: **COSMIC: the Catalogue Of Somatic Mutations In Cancer**. *Nucleic Acids*
810 *Res* 2019, **47**:D941-D947.
- 811 53. UniProt C: **UniProt: the universal protein knowledgebase in 2021**. *Nucleic Acids Res*
812 2020.
- 813 54. Panigrahi P, Jere A, Anamika K: **FusionHub: A unified web platform for annotation and**
814 **visualization of gene fusion events in human cancer**. *PLoS One* 2018, **13**:e0196588.
- 815 55. Zhang J, Gao T, Maher CA: **INTEGRATE-Vis: a tool for comprehensive gene fusion**
816 **visualization**. *Sci Rep* 2017, **7**:17808.
- 817

818

819 **FIGURE 1**



820

821 **Figure 1. The ensemble fusion detection pipeline identifies true positive fusions. A)** The

822 ensemble approach identifies fusions in RNA-Seq data by overlapping results from Arriba,

823 CICERO, FusionCatcher, FusionMap, JAFFA, MapSplice, and STAR-Fusion. It hierarchically

824 prioritizes and filters the fusions utilizing an in-house PostgreSQL database and knowledge base,

825 prior to producing an output list of predicted fusions. In many cases, detected fusions were

826 orthogonally tested by clinical confirmation in order to return a medically meaningful result. **B)**

827 The ensemble pipeline was tested on a dilution series of a reference control reagent (SeraCare) to

828 determine sensitivity and limit of detection. We optimized the pipeline using the undiluted

829 reference control reagent, identifying that by requiring ≥ 3 callers to have overlap for a detected

830 fusion, and by utilizing filtering of known false positive fusion calls and cross-referencing a list of

831 known fusions, all 14 fusions were identified. Colors representing different fusions present in the

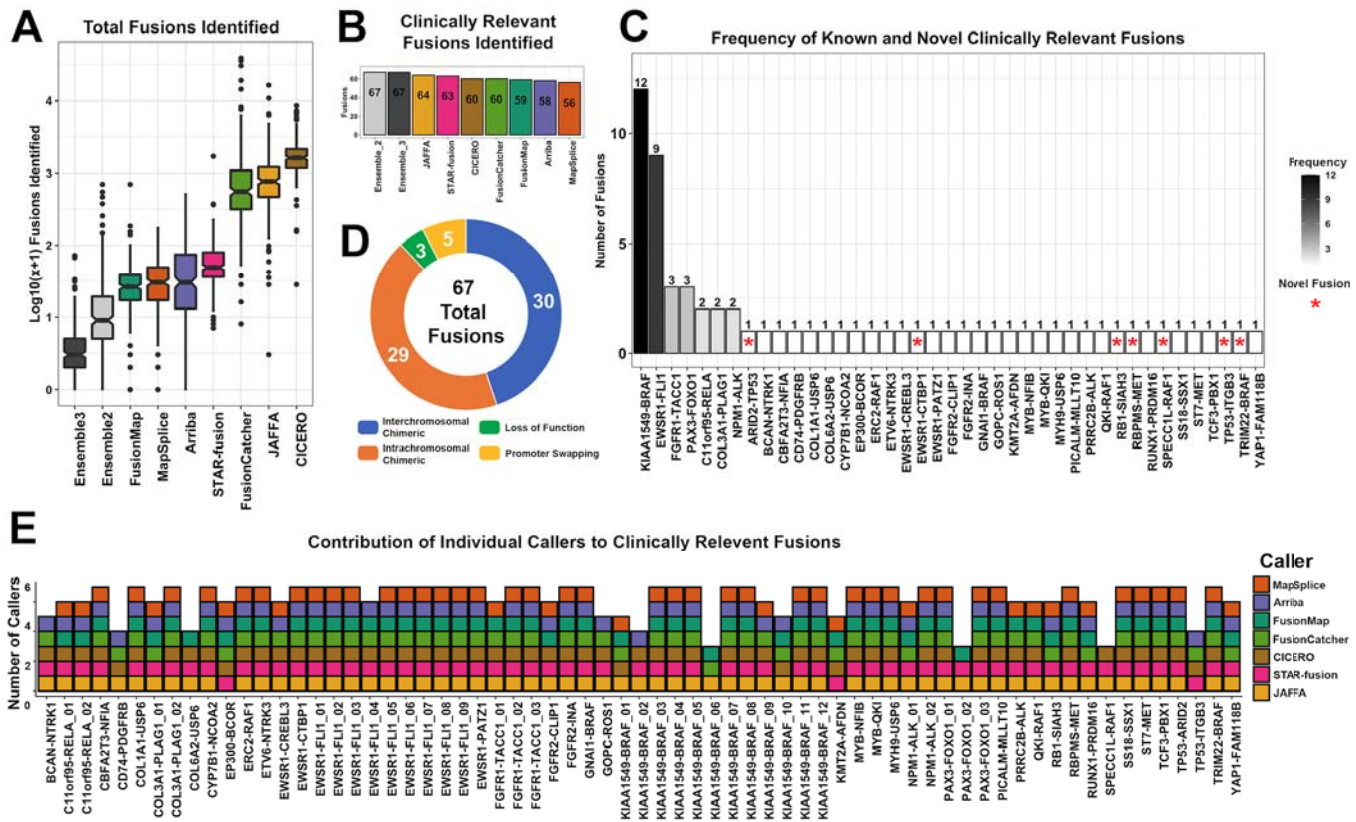
832 SeraSeq v2 reagent are ordered by their absolute proportions. We then applied the optimized

833 pipeline to the dilution series, showing that the numbers of identified fusions were reduced in

834 serial dilutions, and no fusions were identified in the negative control.

835

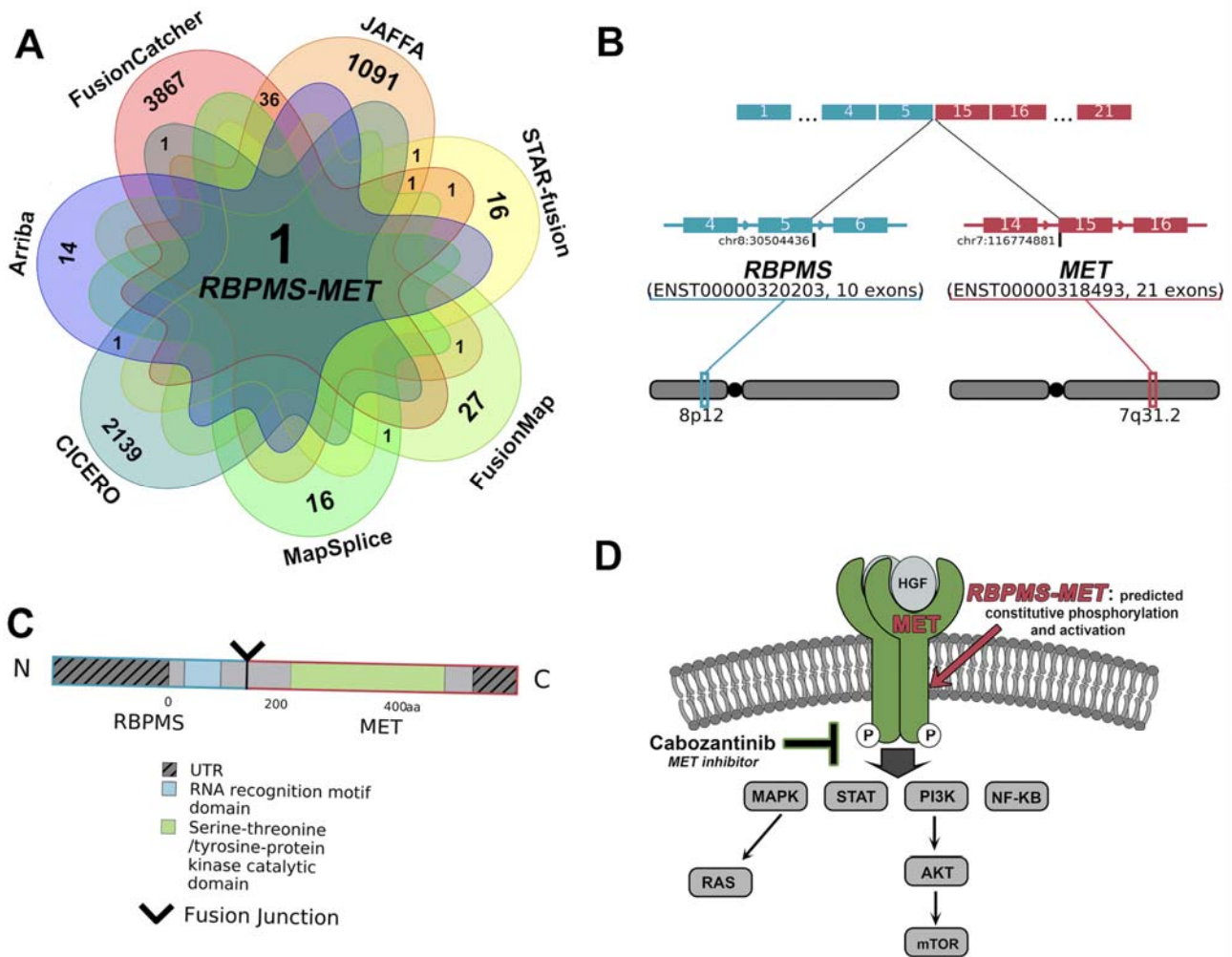
836 **FIGURE 2**



837

838 **Figure 2. Clinically relevant fusions identified by the ensemble approach in a pediatric**
 839 **cancer and hematologic disease cohort. A)** The ensemble approach, with automated filtering,
 840 identifies significantly fewer fusions compared to individual callers. The number of fusions is
 841 plotted as $\log_{10}(x+1)$ to account for 0 fusions identified in some cases. Callers are sorted by the
 842 lowest median number of fusions identified to highest.. **B)** 67 Clinically relevant fusions were
 843 identified, represented as a bar graph with decreasing fusions per individual algorithm,
 844 highlighting the sensitivity of the ensemble approach compared to individual algorithms. No
 845 individual algorithm was able to identify all 67 fusions. **C)** Of the 67 clinically relevant fusions
 846 identified, 30 were interchromosomal chimeric (blue), 29 were intrachromosomal chimeric
 847 (orange), 3 were loss of function (green), and 5 were promoter swapping (yellow) fusions. **D)** Of
 848 the 67 clinically relevant fusions identified, 7 are novel events (red asterisk), while the remaining
 849 60 fusion partners had been described previously in the literature. **E)** A stacked bar graph
 850 represents the individual fusion callers that contributed to each clinically relevant fusion.

851 **FIGURE 3**



852

853 **Figure 3. An *RBPMS-MET* fusion identified in a patient with an infantile fibrosarcoma-like**

854 **tumor. A) *RBPMS-MET* fusion was identified by all seven fusion callers in the filtered overlap**

855 **results. The number of fusions identified by each caller is in the outer VENN diagram sections,**

856 **while internal numbers indicate overlapping fusions found post-filtering (0 overlaps between**

857 **callers are not shown). B) The *RBPMS-MET* fusion is an interchromosomal event, occurring**

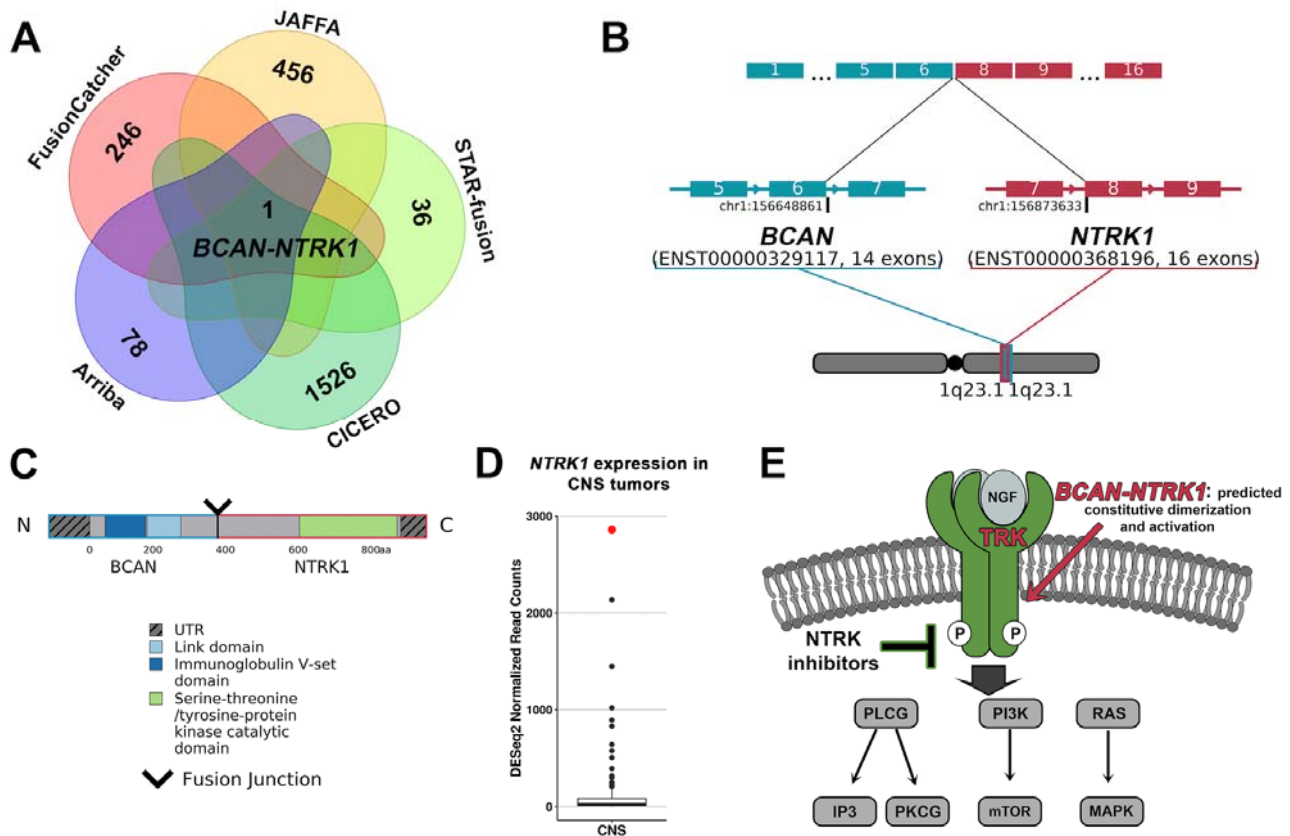
858 **between 8p12 and 7q31.2 and joining exon 5 of *RBPMS* (blue) to exon 15 of *MET* (red). C) The**

859 **fusion protein product includes the RNA recognition motif domain of *RBPMS* and the tyrosine**

860 **kinase catalytic domain of *MET*. D) The *RBPMS-MET* fusion is predicted to cause constitutive**

861 **phosphorylation and activation of *MET*, targetable using cabozantinib.**

862 **FIGURE 4**



863

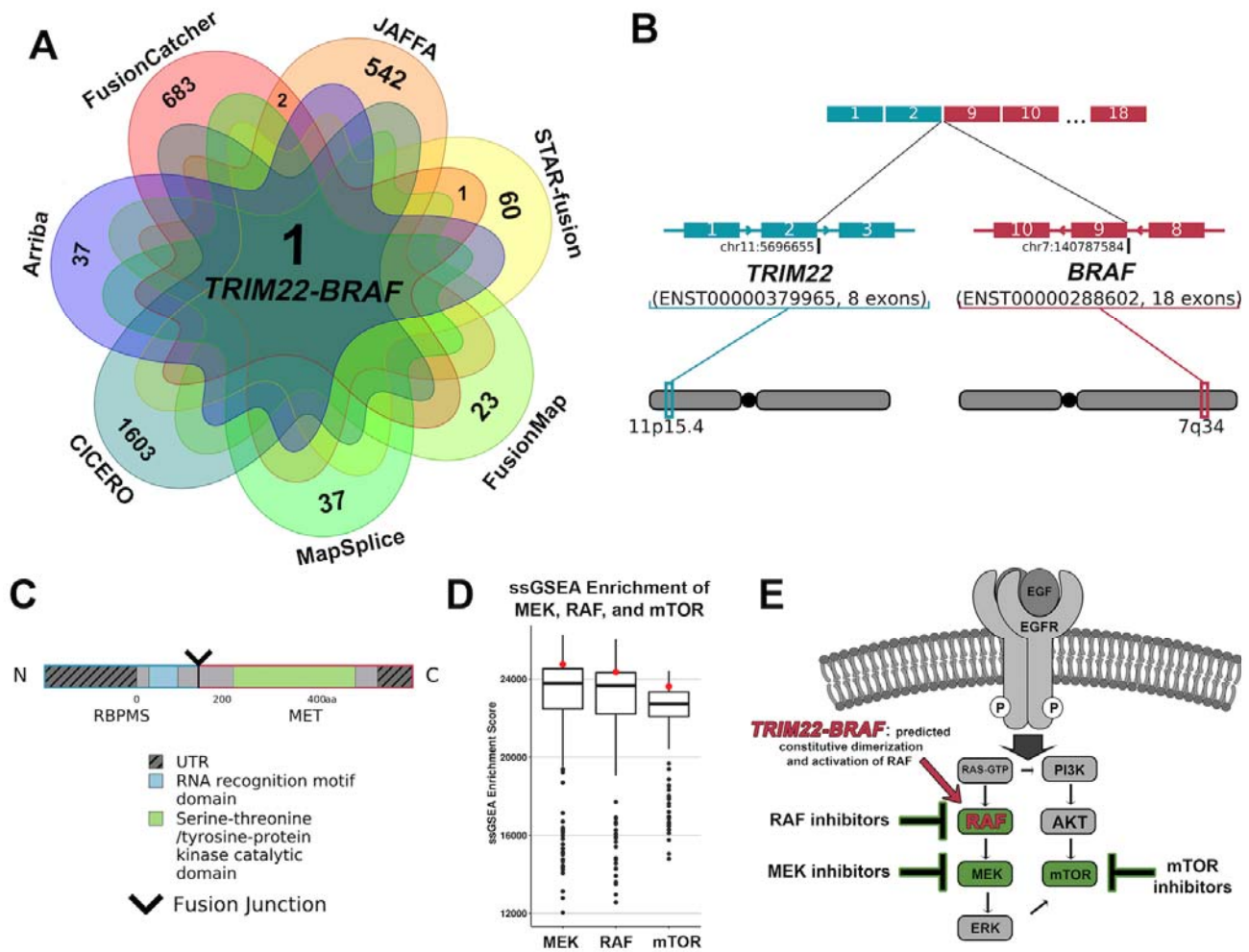
864 **Figure 4. Targetable *NTRK1* fusion identified in an infiltrating glioma. A)** The *BCAN-NTRK1*

865 fusion was identified by 5 of 7 fusion callers, and was the only fusion returned by the filtered
 866 overlap results. Total fusions identified by each caller are shown, FusionMap and MapSplice
 867 identified no overlapping fusions that passed filtering (0 overlaps between callers are not shown).

868 **B)** The *BCAN-NTRK1* fusion is an intrachromosomal event occurring on 1q23.1, joining exon 6 of
 869 *BCAN* (blue) and exon 8 of *NTRK1* (red). **C)** This fusion results in the juxtaposition of the tyrosine
 870 kinase catalytic domain of the *NTRK1* gene to the 5' end of the *BCAN* gene. **D)** *NTRK1* is highly
 871 expressed in this patient (red) compared to CNS tumors (black) in the NCH cohort (CNS tumors: n
 872 = 138), with a normalized read count that is 7.70 standard deviations above the mean (131.2). **E)**

873 The *BCAN-NTRK1* fusion is predicted to increase expression and activation of the tyrosine kinase
 874 *NTRK1*, which may be inhibited by TRK inhibitor therapy (green).

875 **FIGURE 5**



876

877 **Figure 5. Identification of a novel *BRAF* fusion in a mixed neuronal-glia tumor.** A) The
 878 *TRIM22-BRAF* fusion was identified by all seven fusion callers and in the filtered overlap results,
 879 total fusions identified by each caller and overlapping fusions are shown (0 overlaps between
 880 callers are not shown). B) The *TRIM22-BRAF* fusion is an interchromosomal event between
 881 11p15.4 and 7q34, joining exon 2 of *TRIM22* (blue) to exon 9 of *BRAF* (red). C) The resulting
 882 fusion product contains the 5' *TRIM22* zinc finger binding domains and *BRAF* tyrosine kinase
 883 catalytic domain. D) Single sample gene set enrichment analysis (ssGSEA) indicates a trend
 884 toward an enrichment of the MEK (above the 75th percentile, 0.68 standard deviations above the
 885 mean of 22756.87), RAF (above the 75th percentile, 0.60 standard deviations above the mean of
 886 22635.74), and mTOR (above the 75th percentile, 0.72 standard deviations above the mean of
 887 22191.50) upregulated gene sets in the *TRIM22-BRAF* sample (red) compared to the pan-cancer
 888 NCH cohort (black) (pan-cancer cohort: n = 229). E) The *TRIM22-BRAF* fusion is predicted to cause
 889 constitutive dimerization and activation of the *BRAF* kinase domain, shown in D), which could be
 890 targeted by RAF, MEK, and mTOR inhibitors (green).

891 **TABLE 1**

892

Tool	Version	Aligner	Reference	Average Fusions Called per Case	Sensitivity (Clinically Relevant Fusions Called out of 67)
Arriba	v1.2.0	STAR aligner	Haas <i>et al.</i> , 2019 Genome Biol	54	86.6% (58)
CICERO	v0.3.0	candidate SV (structural variant) breakpoints and splice junction	Tian <i>et al.</i> , 2020 Genome Biol	1915	89.6% (60)
FusionMap	v mono-2.10.9	GSNAP (Genomic Short-read Nucleotide Alignment Program) - 12mer based	Ge <i>et al.</i> , 2011 Bioinformatics	34	88.1% (59)
FusionCatcher	v0.99.7c	4 aligners to identify junctions (Bowtie, BLAT, STAR, and Bowtie2)	Nicorici <i>et al.</i> , 2014 bioRxiv	1558	89.6% (60)
JAFFA	direct v1.09	BLAT, uses kmers to select reads that do not map to known transcripts	Davidson <i>et al.</i> , 2015 Genome Med	1141	95.5% (64)
MapSplice	v2.2.1	approximate sequence alignment combined with a local search	Wang <i>et al.</i> , 2010 Nucleic Acids Res	37	83.6% (56)
STAR-fusion	v1.6.0	STAR aligner	Haas <i>et al.</i> , 2019 Genome Biol	72	94.0% (63)

893

894 **Table 1. Performance comparison of individual fusion calling algorithms.** Fusion calling
 895 algorithms utilized by the ensemble fusion detection pipeline and their contributions to fusion
 896 calling in the NCH pediatric cancer and hematologic disease cohort.

897

898 **TABLE 2**

Algorithm	Total Fusions Identified	Seraseq Fusions Identified	Sensitivity	Precision
Arriba	23.5	13	92.9%	55.3%
MapSplice	22	12	85.7%	54.6%
STAR-fusion	32	14	100.0%	43.6%
FusionMap	30	12.5	89.3%	41.7%
FusionCatcher	299.5	13	92.9%	4.3%
JAFFA	470.5	12.5	89.3%	2.7%
CICERO	1323	14	100.0%	1.1%
Ensemble 2 callers	38.5	14	100.0%	36.4%
Ensemble 2 callers + filter	15.5	12	85.7%	77.4%
Ensemble 2 callers + filter + known fusion list	17.5	14	100.0%	80.0%
Ensemble 3 callers	15	14	100.0%	93.3%
Ensemble 3 callers + filter	12	12	85.7%	100.0%
Ensemble 3 callers + filter + known fusion list	14	14	100.0%	100.0%

899

900 **Table 2. Improved precision in fusion detection, utilizing Seraseq controls, achieved**
901 **through utilization of the ensemble pipeline.** Data shown is from undiluted Seraseq v3 RNA-
902 Seq, experiments performed in duplicate, averages are shown. Individual algorithms are listed by
903 precision, in descending order. Seraseq fusions identified (true positive) are out of a possible 14
904 fusions.