Published in final edited form as:

Nat Genet. 2016 January ; 48(1): 4–6. doi:10.1038/ng.3466.

Correspondence to Nature Genetics

Xin Zhou¹, Michael N. Edmonson¹, Mark R. Wilkinson¹, Aman Patel¹, Gang Wu¹, Yu Liu¹, Yongjin Li¹, Zhaojie Zhang¹, Michael C. Rusch¹, Matthew Parker¹, Jared Becksfort¹, James R. Downing^{1,2}, and Jinghui Zhang¹

¹Department of Computational Biology, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, Tennessee 38105

²Department of Pathology, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, Tennessee 38105

To the Editor

Genomic information about predisposing germline mutations within normal cells as well as acquired somatic lesions within cancer cells will enable the development and delivery of individualized cancer therapies. Ongoing global initiatives have revealed that the spectrum of somatic and germline genetic lesions in pediatric cancer are distinct from those found in adult cancer^{1,2,3}. However, existing cancer genome data portals (cBioPortal⁴ and COSMIC⁵) have focused primarily on presenting data generated from adult cancer studies. They also lack features for exploring pathogenic germline mutations, gene fusions, and mutation stratification by cancer subtypes, all of which are of great importance in pediatric cancer.

Here we describe ProteinPaint, a web application for simultaneously visualizing genetic lesions (including sequence mutations and gene fusions) and RNA expression of pediatric cancer. The pediatric data set consists of 27,188 validated somatic coding lesions acquired at diagnosis or relapse from 17 subtypes of pediatric cancer, 252 pathogenic or loss-of-function germline lesions detected in >1000 pediatric cancer patients of 21 subtypes⁶, and RNA-Seq of 928 pediatric tumors from 36 subtypes (Supplementary Notes). The data were compiled from five major studies (Supplementary Notes) and will be expanded with the publication of additional pediatric cancer studies.

Genetic lesions of pediatric cancer are shown on a protein panel (Fig. 1) with the option for a parallel view of a curated version of published somatic mutations in COSMIC database (Supplementary Notes). This enables the use of adult data for interpreting the significance of

Correspondence should be addressed to: **Jinghui Zhang, PhD**, Department of Computational Biology, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, ; Email: jinghui.zhang@stjude.org

URL. ProteinPaint <https://pecan.stjude.org/proteinpaint>.

AUTHOR CONTRIBUTIONS

J.Z. and J. D. conceived the project. X.Z. and J.Z. designed the project. X.Z. implemented ProteinPaint. J.Z., G.W., M. P, A.P., J.B. and M.C.R. performed QC check or participated in software development. M.N.E., M.W., G.W., Y.Li, Z.Zhang, and Y.Liu generated the data. J.Z. and J.D. supervised the project. X.Z., M.N.E. and J.Z. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

rare genetic lesions in pediatric cancer (Supplementary Figs. 1, 2) and vice versa (Supplementary Fig. 3). To ensure consistency, all variants were reannotated with a modified version of Annovar⁷. As an example, we show how this presentation has enabled the detection of aberrant splicing caused by recurrent “silent” mutations in *TP53*. This finding also provided insight into the pathogenicity of matching germline variants found in patients with cancer predisposition syndromes (Supplementary Fig. 1, Supplementary Notes). Additionally, presentation of mutant allele fractions in DNA and RNA facilitates evaluation of tumor heterogeneity related to cancer relapse (Supplementary Fig. 3) as well as detection of allelic imbalance in DNA or RNA caused by a second genetic or epigenetic hit in tumor (Supplementary Fig. 1, Supplementary Notes). Loss-of-heterozygosity (LOH), which was computed by the CONSERING⁸ algorithm in the pediatric cancer genomes we analyzed, is shown to further facilitate the identification of double-hit mutations (Supplementary Fig. 4).

The expression panel presents the rank and quantity of gene expression of each sample with superimposed boxplots summarizing the expression range of the entire cohort, or user-selected subtypes. Selecting a genetic lesion such as the *PAX5-JAK2* fusion on the protein panel automatically highlights the mutated samples on the expression panel; in the case of *PAX5-JAK2* fusion, this reveals the aberrantly high expression of *JAK2* caused by gene fusion (Fig. 1b). Conversely, examination of aberrant expression in a tumor may lead to novel insight into its causal genetic lesion. We show an example of how outlier expression of *FLT3* in a leukemia with kinase activation signature led to discovery of a high-level *FLT3* amplification resulting from replication of an episome formed by a complex rearrangement involving three chromosomes (Supplementary Fig. 5, Supplementary Notes).

ProteinPaint is designed to deliver a premium visualization experience with interactive and animated features. Novel “disc-on-stem” skewer graphs were implemented to depict the diverse prevalence, complex allelic alteration, and temporal origin of mutations and gene fusions at a glance (Fig. 1, Supplementary Figs. 6–7). Customized views include display of mutation and expression by cancer subtypes or tumor tissues, dynamic zoom, and integration of user-provided data with new features implemented based on user feedback. Data in mutation annotation format (MAF) generated by studies such as The Cancer Genome Atlas (TCGA) or individual research labs can be uploaded to ProteinPaint to enable data visualization and cross-study comparison for the broad genetic research community (Supplementary Fig. 8, Supplementary Tutorial). Manually-curated protein domains have been incorporated for genes frequently mutated in pediatric cancer to facilitate the interpretation of mutation pathogenicity (Fig. 1, Supplementary Fig. 6). ProteinPaint complements existing cancer genome portals by providing a comprehensive and intuitive view of pediatric cancer genomic data with advanced visualization features, as well as integration of expression and adult cancer data (Supplementary Figs. 6–7, Supplementary Notes). Taken as a whole, these novel features make ProteinPaint a powerful tool for employing genomic data to enhance pediatric cancer research, collaboration, and clinical care.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Drs. Jeffery Kico, Paul Northcott and Charles Mullighan for helpful suggestions. We thank the reviewers of this manuscript for suggestions of implementing new interface for custom data upload. This study was supported by the St. Jude Children's Research Hospital-Washington University Pediatric Cancer Genome Project, Cancer Center support grant P30 CA021765 from the US National Cancer Institute, and American Lebanese Syrian Associated Charities of St. Jude Children's Research Hospital.

References

1. Downing JR, et al. *Nat Genet.* 2012; 44(6):619–622. [PubMed: 22641210]
2. Zhang J, et al. *Nature.* 2012; 481(7380):157–163. [PubMed: 22237106]
3. Wu G, et al. *Nat Genet.* 2012; 44:251–253. [PubMed: 22286216]
4. Gao J, et al. *Sci Signal.* 2013; 6(269):pl1. [PubMed: 23550210]
5. Forbes SA, et al. *Nucl Acids Res.* 2014; 43:D805–D811. [PubMed: 25355519]
6. Zhang J, et al. *N Engl J Med.* In press.
7. Wang K, Li M, Hakonarson H. *Nucleic Acids Res.* 2010; 38(16):e164. [PubMed: 20601685]
8. Chen X, et al. *Nat Methods.* 2015; 12(6):527–30. [PubMed: 25938371]
9. Liang SH, Clarke MF. *J Biol Chem.* 1999; 274(46):32699–32703. [PubMed: 10551826]
10. Roberts KG, et al. *N Engl J Med.* 2014; 371:1005–1015. [PubMed: 25207766]

Figure 1. Comprehensive Visualization of Sequence Mutations, Gene Fusions, and RNA Expression Using ProteinPaint. **(a)** *TP53* mutation profile in the Pediatric data set (top) and COSMIC database (bottom). The number of samples affected by each mutation is indicated by the text within each disc, as well as disc size. The arc outside each disc indicates the proportion of samples that are germline (solid) or relapsed tumor (hollow). The full legend is shown in Supplementary Fig. 1. The manually-curated “NLS” domain reveals a hotspot nonsense mutation R306* that disrupts a known nucleotide localization signal⁹. **(b)** *JAK2* gene fusion (left) and expression (right). Left: *JAK2* fusions are shown along with sequence mutations affecting the pseudokinase and kinase domains of *JAK2*. A half-filled disc represents a gene fusion, with the filled section representing the N- or C- terminal of the protein involved in the fusion. The arrow points to the *PAX5-JAK2* fusion detected in 7 tumors of Ph-like B-cell acute lymphoblastic leukemia¹⁰. The fusion protein involves the C terminal of *JAK2*. Right: *JAK2* expression level in Pediatric samples. The horizontal axis represents range of FPKM values. Gray circles represent samples in descending order of *JAK2* FPKM value. Samples highlighted by the filled red circles are those with *PAX5-JAK2* fusion selected by the user. The ratio of *PAX5-JAK2* fusion transcript to the overall expression of its two partner genes,

PAX5 and *JAK2* are labeled in the red text. Boxplots represent FPKM value distribution in pediatric cancer cohorts, labeled by the disease names and cohort sizes.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript