Published in final edited form as:
Nat Genet. ; 44(6): 619–622. doi:10.1038/ng.2287.

The Pediatric Cancer Genome Project

James R Downing[1,2], **Richard K Wilson**[3,4], **Jinghui Zhang**[1,5], **Elaine R Mardis**[3,4], **Ching-Hon Pui**[1,2,6], **Li Ding**[3,4], **Timothy J Ley**[3,4], and **William E Evans**[1,7]

[1]St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project, Memphis, Tennessee, USA

[2]Department of Pathology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA

[3]St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project, St. Louis, Missouri, USA

[4]The Genome Institute at Washington University, St. Louis, Missouri, USA

[5]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA

[6]Department of Oncology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA

[7]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, Tennessee, USA

Abstract

The St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project (PCGP) is participating in the international effort to identify somatic mutations that drive cancer. These cancer genome sequencing efforts will not only yield an unparalleled view of the altered signaling pathways in cancer but should also identify new targets against which novel therapeutics can be developed. Although these projects are still deep in the phase of generating primary DNA sequence data, important results are emerging and valuable community resources are being generated that should catalyze future cancer research. We describe here the rationale for conducting the PCGP, present some of the early results of this project and discuss the major lessons learned and how these will affect the application of genomic sequencing in the clinic.

Correspondence should be addressed to J.R.D. (james.downing@stjude.org) or W.E.E. (william.evans@stjude.org).

URLs The Cancer Genome Atlas (TCGA), http://cancergenome.nih.gov/; the International Cancer Genome Consortium (ICGC), http://www.icgc.org/; Therapeutically Applicable Research to Generate Effective Treatments (TARGET), http://target.cancer.gov/; database of Genotypes and Phenotypes (dbGAP), http://www.ncbi.nlm.nih.gov/gap/; Sequence Read Archive (SRA), http://www.ncbi.nlm.nih.gov/sra/; The European Genome-Phenome Archive (EGA), https://www.ebi.ac.uk/ega/; The PCGP Explore portal, http://explore.pediatriccancergenomeproject.org/; Surveillance Epidemiology and End Results (SEER) data, http://seer.cancer.gov/siterecode/icdo3_d01272003/; Clipping Reveals Structure (CREST) software, http://www.stjuderesearch.org/site/lab/zhang.

AUTHOR CONTRIBUTIONS

The Pediatric Cancer Genome Project is led by J.R.D., R.K.W., W.E.E., E.R.M., J.Z., T.J.L. and L.D. These individuals have contributed to the design of experiments and the generation and interpretation of all of the data. C.-H.P. developed Figure 1. J.Z. developed Figure 2. J.R.D. and W.E.E. wrote the manuscript. All authors assisted in editing the manuscript and agree with the content as presented.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Project design

In January 2010, St. Jude Children's Research Hospital and The Genome Institute at the Washington University announced the launch of the Pediatric Cancer Genome Project, a 3-year, $65-million privately funded initiative[1]. The stated goal of this effort was to sequence at 30-fold haploid coverage the whole genome of 600 pediatric tumors and matched non-tumor germline samples (1,200 total genomes) and to define the landscape of somatic mutations that underlie major subtypes of pediatric cancer. As leaders of this effort, it was our belief that a large pediatric-focused cancer sequencing effort was necessary to fully explore the genetic basis of the unique cancers seen in children. Thus, from the start, the PCGP was designed to complement the larger government-funded genomic efforts focused on adult cancers, including the US National Human Genome Research Institute (NHGRI)/ National Cancer Institute (NCI) Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC) and the smaller NCI-funded pediatric project known as Therapeutically Applicable Research to Generate Effective Treatments (TARGET). With structural alterations, such as inter- and intrachromosomal rearrangements, being a common mechanism of mutagenesis in pediatric leukemias and solid tumors, we felt that a whole-genome sequencing (WGS) approach instead of exome or transcriptome sequencing would be required to accurately define the full spectrum of mutations in pediatric cancers. Our expectation was that the results from this project would catalyze research in pediatric malignancies and lead to improvements in our ability to diagnose, monitor and treat patients with targeted therapies aimed at the identified driver mutations.

We have recently completed the second year of the project. To date, the effort has not only yielded some remarkable surprises, but it has also generated one of the largest high-coverage whole-genome DNA sequence databases in cancer—a resource that will serve both cancer and non-cancer researchers for years to come. Here we highlight some of the important insights that have emerged from the PCGP, describe the resources we are making available to the scientific community and discuss the challenges that cancer genomics must overcome to gain a full understanding of the genetic lesions that underlie pediatric and adult cancers.

The spectrum of pediatric cancers sequenced

Despite the paucity of new drugs to treat childhood cancers during the past 20 years, the cure rates for these diseases have continuously improved: in developed countries, the overall cure rate for children with cancer now stands at ~80% (ref. 2). This success has been built on the use of cytotoxic chemotherapy and radiotherapies that are often associated with major side effects and can ultimately reduce the quality of life for survivors[3]. Although cure rates for childhood cancers are impressive relative to those for adult malignancies, cancer remains the leading cause of death by disease among children over 1 year of age in developed countries[4]. It is generally believed that new, less toxic curative treatments of childhood cancers should target the genetic alterations that drive these diseases. Elucidating the genetic abnormalities that underlie childhood cancers is therefore an essential step toward understanding the pathobiology of these diseases and using the information gained to develop more effective and less toxic treatments.

One could ask whether a pediatric cancer genome project is the best way to achieve the desired result or if sufficient understanding would emerge from the larger adult-focused projects. To those of us working in pediatric cancer research, the answer is obvious—children are not just small adults. The spectrum of cancers occurring in the pediatric population is markedly different from that seen in adults. For example, the major brain and solid tumors that arise in children, including medulloblastoma, neuroblastoma, rhabdomyosarcoma, Ewing sarcoma, osteosarcoma and Wilms tumor, are exceedingly rare in adults (Fig. 1a). Similarly, the specific genetic subtypes of acute lymphoblastic leukemia (ALL)—the most common malignancy in children—differ markedly between children and adults (Fig. 1b). This marked difference in the spectrum of cancers is not unexpected, in that many pediatric cancers are thought to arise within developing tissues that undergo substantial expansion during early organ formation, growth and maturation. The unique biology of these developing tissues suggests that the spectrum of mutations that lead to malignant transformation will also differ between pediatric and adult cancers. Thus, a focused project to characterize the landscape of mutations in pediatric cancers is necessary to achieve the goal of advancing cures for pediatric cancers.

So, what tumors should be sequenced first? Although statistical arguments suggest that 500 tumors of an individual subtype need to be sequenced to accurately identify all mutations occurring at a 5% or greater frequency, the relative rarity of pediatric cancers coupled with the heterogeneity of tumor subtypes makes this approach unfeasible in the short term. In fact, obtaining sufficient tumor samples has been a major limitation in the adult cancer genome sequencing projects. We therefore took the approach of sequencing the pediatric cancer subtypes for which outcome (cure) with current treatment is poor and/or where there is a conspicuous lack of knowledge regarding the genetic basis of the disease. To date, we have completed primary data acquisition and initial analyses for 260 pairs of pediatric cancer and matched non-tumor DNA from 15 specific tumor subtypes, as shown (Fig. 2). In the initial 3 years of our project, we plan on interrogating approximately an equal number of genomes from childhood leukemias, solid tumors and brain tumors.

Lessons learned

Several important findings have emerged from our initial studies. Foremost was the importance of using the WGS approach to identify mutations in pediatric cancers. Analysis of an aggressive subtype of pediatric ALL known as early T-cell precursor leukemia identified complex structural variations, focal deletions and sequence mutations of genes encoding key hematopoietic regulators that act as driver lesions in these leukemias[5]. The exceedingly complex nature of some of these structural alterations would make it impossible to accurately identify them using more targeted sequencing approaches, such as exome or transcriptome sequencing. This observation has important implications for the application of next-generation sequencing–based assays in the clinic. The inability of targeted sequencing approaches to accurately and comprehensively identify major classes of mutations would make it difficult for such assays to achieve the level of confidence needed to assure appropriate selection of agents for targeted treatment. In fact, identifying every structural variation that exists within a cancer is difficult, even using WGS data. New algorithms are being developed to tackle this challenging problem[6,7]. An analytical method developed through the PCGP, Clipping Reveals Structure (CREST), exploits the mapping of partial reads to a reference genome to identify those reads that span the junctions of structural variations. This approach allows the mapping of structural variations at the nucleotide level, improving upon previous algorithms and providing a validation rate exceeding 80% (ref. 7) (the CREST package, including the user manual and test data, is freely available; see URLs).

A second important lesson learned through our efforts is that the spectrum of mutations that occur in pediatric cancers can be remarkably different than that seen in adult cancers, even in tumors with very similar histology. A specific example of this is afforded by a recent study from the PCGP on pediatric glioblastomas. In children but not adults, a substantial proportion of glioblastomas arise in the brainstem as diffuse intrinsic pontine gliomas (DIPGs)[8]. Of the DIPGs analyzed in this study, 78% were found to have missense mutations in genes encoding 2 of the 16 different histone H3 isoforms, which resulted in a methionine substitution at lysine 27 (p.Lys27Met), a key regulatory site that can be methylated or acetylated[8]. Although cancer-associated mutations have been identified in genes whose products have a role in regulating histone modification and chromatin structure[9–11], this was the first demonstration of a cancer-associated mutation in a key histone modification site. Notably, the mutation was only detected in DIPGs and, at a lower frequency, in pediatric glioblastomas arising outside the brainstem[12] but not in any adult glial brain tumors[13,14] or in 252 other pediatric cancers of multiple histological subtypes[7].

A variation on this lesson is that the frequency of a particular mutation can also vary within specific pediatric cancers as a function of the child's age. An example of this is provided in a recent PCGP study on stage 4 neuroblastoma, a tumor of the sympathetic nervous system. In this study, somatic mutations of *ATRX* were detected in 44% of adolescents and young adults (≥12 years old) with stage 4 neuroblastoma but were never seen in tumors arising in infants[15].

A third major lesson is the importance of integrating genome-level data with epigenetic and RNA expression data to fully explore the abnormalities that drive cancer. Unexpectedly, we found that retinoblastoma, a pediatric eye tumor characterized by inactivating mutation of *RB1*, had very few mutations across the genome[16]. However, a detailed analysis of epigenetic and expression data revealed aberrant expression of *SYK*, encoding a cytoplasmic tyrosine kinase, in every retinoblastoma analyzed. Notably, inhibition of this kinase resulted in apoptosis of retinoblastoma tumor cells, both *in vitro* and *in vivo*, suggesting a possible new therapeutic approach. Because clear benefits can be achieved by combining whole-genome with transcriptome sequencing, going forward, we have decided to perform transcriptome sequencing on all tumors from which sufficient RNA is available.

Data access and community resources

The data from the recently published PCGP studies are available through the database of Genotypes and Phenotypes (dbGAP) and the Sequence Read Archive (SRA) (see URLs). Despite the fact that the PCGP is privately funded, we have decided to institute an overall data release policy that is consistent with that of the other major publicly funded genome sequencing projects. Specifically, we have uploaded the initial 260 tumor and germline DNA sequence files to the European Bioinformatics Institute data portal (The European Genome-Phenome Archive, EGA). This will provide users immediate access to both published and unpublished data. However, we will require users to agree that the information will be used exclusively for biomedical research and to abide by a moratorium on submission or presentation of work that incorporates these data for a 9-month period following the data's release date. The data generated during the third year of the project will be released in accordance with this policy. This data release will immediately result in a more than doubling of the high-coverage human whole-genome sequence data available to the scientific community.

In addition to making the primary data accessible, we have created the PCGP Explore portal (see URLs), a website that provides access to all data included in PCGP-published studies, along with search and analytical tools that will facilitate the use of these data sets. This website includes validated somatic mutations as well as associated SNP array data, mRNA expression data and clinical information. Rich in visual features, PCGP Explore allows users to gain an overview of the project's progress and provides access to data from specific diseases; summaries of current discoveries; links to associated publications and views of genome data from the single-patient to the disease level and across the entire data set. This website should serve as a valuable discovery resource for the global cancer research community.

Challenges and opportunities for collaboration

With detailed information emerging on the genomics of both pediatric and adult cancers, we find ourselves at the beginning of a new era in cancer medicine. The rapidly decreasing cost of generating DNA sequence data will not only allow a substantial expansion in the number of cancer discovery efforts but has already ushered in the application of this technology in the clinic. Before we let our enthusiasm for this technology get ahead of the science, it is important that we acknowledge some of the challenges that remain before we can claim a full understanding of the genetics of cancer. Despite the implication that WGS provides the complete DNA sequence of an individual's genome, it is important to note that the human reference genome remains an imperfect reference due to remaining gaps and a lack of representation of the full suite of genomic variation across population groups[17]. In addition, the present paired-end sequencing technology that is being used for the majority of projects results in an inability to accurately resolve some structural variations. For example, structural variations that occur in repetitive DNA sequences often are not definitively resolved, as is the case for large non-template insertion events. This difficulty is likely to be lessened with increased read lengths and further advances in analytic approaches.

The presence of heterogeneity within human tumor samples is another challenge that can complicate the interpretation of genomic data. Tumors are rarely, if ever, composed of a single clone of cells but, rather, include admixtures of cancer cells intermingled with normal support and immune cells. Accurately estimating the tumor purity of a sample is often impossible. An important finding that has emerged from many cancer genome sequencing projects is that a striking degree of clonal heterogeneity is present in most cancers[18,19]. This is seen in the primary diagnostic tumor sample and in metastatic and relapse samples secondary to both the selection of preexisting subclones and the generation of new subclones that occur through clonal evolution. The presence of this heterogeneity makes it important to be able to identify mutations in subpopulations of tumor cells that may represent as little as 1% of the total cellular population being analyzed. Studies have shown that it is often these rare subclones that are the founders of metastatic lesions and relapse[18–21]. Next-generation sequencing approaches that incorporate very deep read coverage of specific mutations are ideally suited for detecting mutations at this level of sensitivity; however, extending this approach to the level of the whole genome remains a challenge.

Lastly, it is important to note that the major focus of the mutation-profiling studies reported to date has been the identification of non-silent mutations in annotated genes. Mutations affecting gene regulatory regions are rarely mentioned. This is primarily a result of the substantial difficulty in accurately identifying functionally relevant mutations in the large conserved DNA regions that constitute promoters and enhancers. Understanding the frequency and functional consequences of such mutations will require analytical methods that can interrogate, not only DNA sequence data, but expression data for mRNAs and non-coding RNAs and epigenetic data on CpG methylation and histone modification. These studies will require large numbers of clinical samples to accurately define the frequency of these mutations within specific cancer subtypes. Moreover, these studies will greatly benefit from the work being performed by the NHGRI-funded ENCyclopedia of DNA Elements (ENCODE) consortium[22].

Despite these substantial challenges, cancer genome sequencing efforts have produced important new insights into the pathobiology of cancer. We began the PCGP with the anticipation that this large-scale sequencing effort would ultimately advance our ability to treat children with cancer. Data emerging from the PCGP hold great promise that this goal will be realized. Furthermore, the rapid generation of genomic data across the landscapes of pediatric and adult cancers is opening new avenues of collaboration. Understanding the functional and clinical relevance of the identified mutations in cancer will require bringing together dedicated teams of genomic and computational experts, oncologists, pathologists, molecular and cellular biologists, chemists, pharmacologists and others in order to translate these descriptive data into effective clinical use. Establishing these collaborations as early as possible has had an important impact on the PCGP. In addition, although the prevalence of the effects on specific pathways will differ between pediatric and adult cancers, those that are in common will usher in new clinical trials that cut across the age difference of patients. Bringing together the medical establishments that treat pediatric and adult patients will be an important legacy of cancer genomics.

Acknowledgments

We thank H. Inaba, C. Mullighan, K. Roberts, S. Ramondi, S.-C. Chen, D. Pei and J. Groff for their extensive contributions to and diligence in preparing Figure 1, and we thank X. Chen for contributions to Figure 2. We would also like to thank M. Rusch and P. Nagahawatte for their contributions to data submission to EBI and dbGaP. This work was funded by The St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project, American Lebanese Syrian Associated Charities (ALSAC) of St. Jude Children's Research Hospital, an NCI Cancer Center support grant (P30 CA021765), the US National Institutes of Health (NIH; U01 GM 92666 and R37 CA36401 to St. Jude Children's Research Hospital), grants to R.K.W. from Washington University in St. Louis and a grant from the NHGRI (U54 HG003079 to Washington University in St. Louis).

References

1. Anonymous. News briefing. Childhood cancers. Nature. 2010; 463:407.
2. Pui CH, et al. Challenging issues in pediatric oncology. Nat Rev Clin Oncol. 2011; 8:540–549. [PubMed: 21709698]
3. Hudson MM, et al. Lessons from the past: opportunities to improve childhood cancer survivor care through outcomes investigations of historical therapeutic approaches for pediatric hematological malignancies. Pediatr Blood Cancer. 2012; 58:334–343. [PubMed: 22038641]
4. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. CA Cancer J Clin. 2012; 62:10–29. [PubMed: 22237781]
5. Zhang J, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature. 2012; 481:157–163. [PubMed: 22237106]
6. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. 2009; 6:S13–S20. [PubMed: 19844226]
7. Wang J, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods. 2011; 8:652–654. [PubMed: 21666668]
8. Wu G, et al. Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. Nat Genet. 2012; 44:251–253. [PubMed: 22286216]
9. Mullighan CG, et al. *CREBBP* mutations in relapsed acute lymphoblastic leukaemia. Nature. 2011; 471:235–239. [PubMed: 21390130]
10. Morin RD, et al. Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. Nat Genet. 2010; 42:181–185. [PubMed: 20081860]
11. van Haaften G, et al. Somatic mutations of the histone H3K27 demethylase gene *UTX* in human cancer. Nat Genet. 2009; 41:521–523. [PubMed: 19330029]
12. Schwartzentruber J, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2012; 482:226–231. [PubMed: 22286061]
13. The Cancer Genome Atlas Research Network. An integrated genomic analysis of human glioblastoma multiforme. Nature. 2008; 455:1061–1068. [PubMed: 18772890]
14. Parsons DW, et al. Driver mutations in histone 3.3 and chromatin remodelling genes in paediatric glioblastoma. Science. 2008; 321:1807–1812. [PubMed: 18772396]
15. Cheung N-K, et al. Association of age at diagnosis and genetic mutations in patients with neuroblastoma. J Am Med Assoc. 2012; 307:1062–1071.
16. Zhang J, et al. A novel retinoblastoma therapy from genomic and epigenetic analyses. Nature. 2012; 481:157–163. [PubMed: 22237106]

17. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]

18. Campbell PJ, et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature. 2010; 467:1109–1113. [PubMed: 20981101]

19. Yachida S, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature. 2010; 467:1114–1117. [PubMed: 20981102]

20. Notta F, et al. Evolution of human BCR-ABL1 lymphoblastic leukaemia–initiating cells. Nature. 2011; 469:362–367. [PubMed: 21248843]

21. Mullighan CG, et al. Genomic analysis of the clonal origins of relapsed acute lym-phoblastic leukemia. Science. 2008; 322:1377–1380. [PubMed: 19039135]

22. Birney E, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]

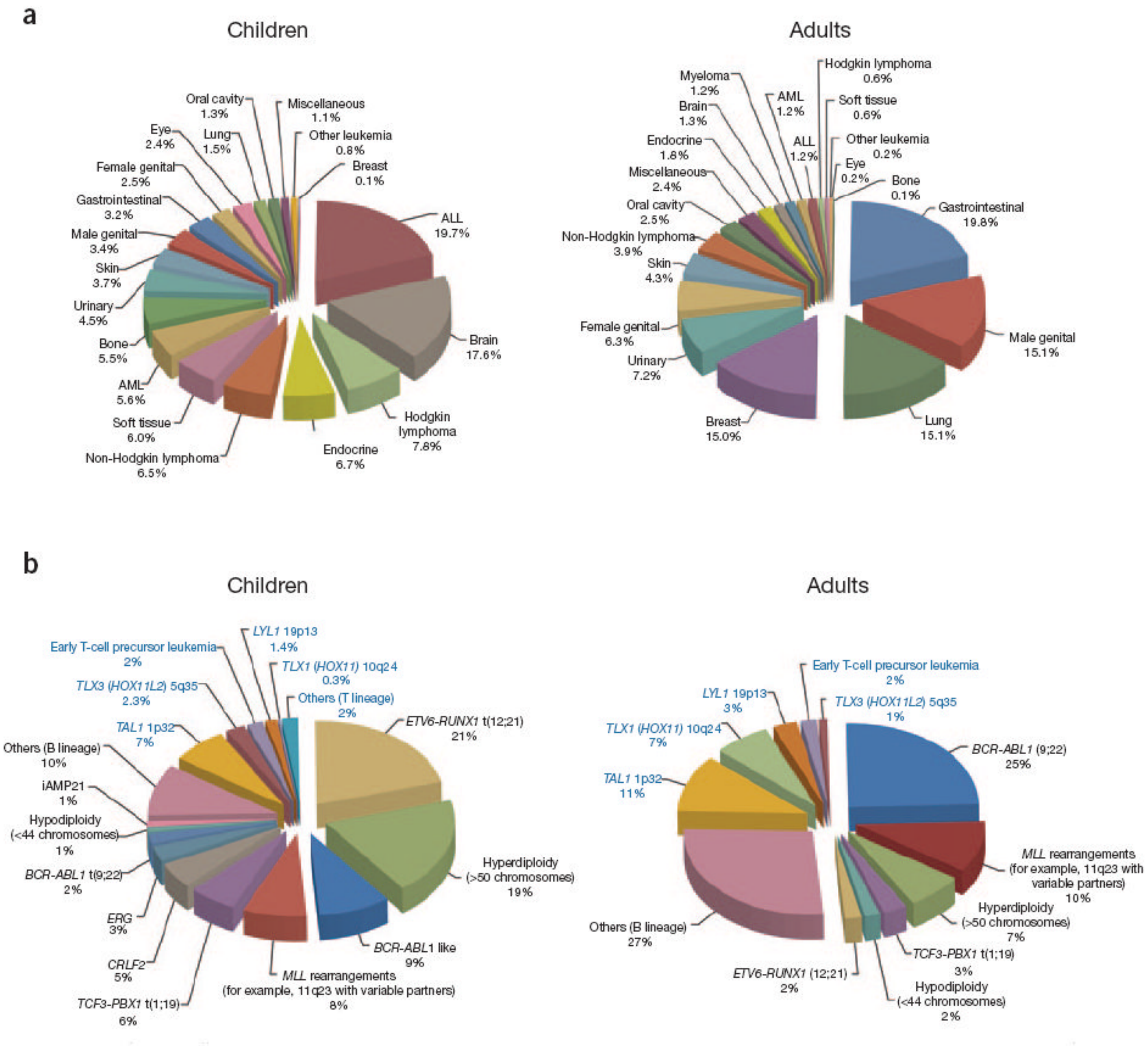

Figure 1.

Frequency of cancer diagnoses and leukemia subtypes in children and adults. (**a**) The frequency of cancer types in children (left) and adults (right) on the basis of 2012 Surveillance, Epidemiology and End Results (SEER) data. Each chart is organized with cancers listed from the most common to the least common in a clockwise fashion. (**b**) The frequency of T-cell lineage (blue text) and B-cell lineage (black text) subtypes of acute lymphoblastic leukemia (ALL) in children (left) and adults (right). Each chart is organized with ALL subtypes listed from the most common to the least common in a clockwise fashion. iAMP21, intrachromosomal amplification of chromosome 21.

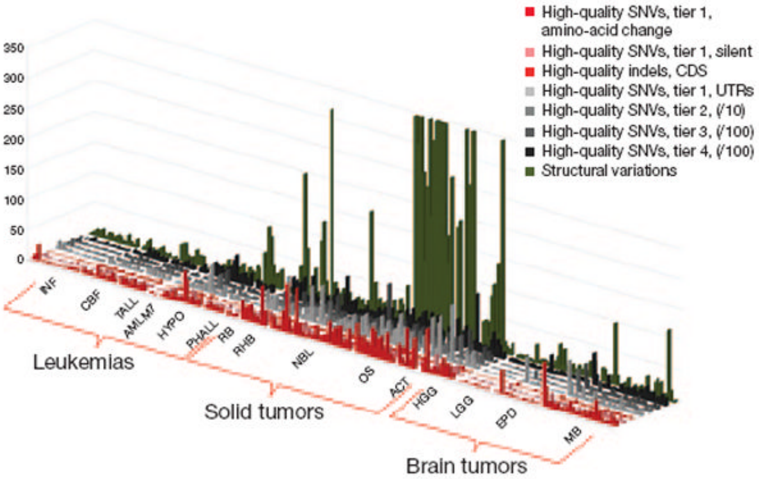

Figure 2.

Genetic landscape of 15 different types of pediatric cancers determined from whole-genome sequencing of 260 tumors and matching germline samples. The number of somatic mutations in each sample, including single-nucleotide variations (SNVs), insertion and/or deletion events (indels) and structural variations, is shown as the height in the three-dimensional graph. Only high-quality variations or validated somatic mutations are included in the summary. CDS, protein-coding regions; tier 1, mutations in annotated genes; tier 2, mutations in non-coding conserved or regulatory regions; tier 3, mutations in non-repetitive, non-coding and non-conserved regions; tier 4, mutations in repetitive regions. Tier 2 and tier 3/tier 4 mutations were rescaled to 1/10 and 1/100 of the original counts to maintain a consistent scale with the results for other somatic lesions. INF, infant ALL; CBF, core-binding-factor acute myeloid leukemia; TALL, T-cell ALL; AMLM7, acute megakaryoblastic leukemia; HYPO, hypodiploid ALL; PHALL, Philadelphia chromosome–positive *BCR-ABL1* ALL; RB, retinoblastoma; RHB, rhabdomyosarcoma; NBL, neuroblastoma; OS, osteosarcoma; ACT, adrenocortical carcinoma; HGG, high-grade glioblastoma; LGG, low-grade glioma; EPD, ependymoma; MB, medulloblastoma.