C. Contribution to Science

C.1. Familial Risk Assessment. Risk evaluation to identify individuals who are at greater risk of cancer as a result of heritable pathogenic variants is an essential component of individualized strategies for prevention and early detection. Initially in collaboration with Don Berry and now as part of the BayesMendel lab which I co-lead with Danielle Braun, we developed a general approach, algorithms and software for familial risk prediction in cancer. Using principles of Mendelian genetics, Bayesian probability theory, and variant-specific knowledge, our models derive the probability of carrying a pathogenic variant and developing cancer in the future, based on family history. The BayesMendel lab models include BRCAPRO, MMRpro, MelaPro, PancPro and PanelPro. They have been validated and are employed by various widely used clinical softwares including CancerGene, CancerGene Connect, CRA Health, Progeny, MagView, CancerIQ and others. CancerGene alone has more than 4,000 users in more than 75 countries, and CRA Health has more than 15,000 users per month. Through this work I acquired extensive experience in the methodology, validation, informatics, implementation and primary care use of familial risk prediction models. Our latest tool is the PanelPro model, based on novel methodological and computational approaches for familial risk prediction, which allowed us to produce the first tool that consider all major cancer syndromes together.

Parmigiani, G., D. Berry, and O. Aguilar. “Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2.” In: *Am J Hum Genet* 62.1 (1998). PMC1376797, pp. 145–158.

Chen, S., W. Wang, S. Lee, K. Nafa, J. Lee, K. Romans, P. Watson, S. B. Gruber, D. Euhus, K. W. Kinzler, J. Jass, S. Gallinger, N. M. Lindor, G. Casey, N. Ellis, F. M. Giardiello, K. Offit, G. Parmigiani, and Colon Cancer Family Registry. “Prediction of germline mutations and cancer risk in the Lynch syndrome.” In: *JAMA* 296.12 (2006). PMC2538673, pp. 1479–1487.

Parmigiani, G., S. Chen, E. S. Iversen, T. M. Friebel, D. M. Finkelstein, H. Anton-Culver, A. Ziogas, B. L. Weber, A. Eisen, K. E. Malone, J. R. Daling, L. Hsu, E. A. Ostrander, L. E. Peterson, J. M. Schildkraut, C. Isaacs, C. Corio, L. Leondaridis, G. Tomlinson, C. I. Amos, L. C. Strong, D. A. Berry, J. N. Weitzel, S. Sand, D. Dutson, R. Kerber, B. N. Peshkin, and D. M. Euhus. “Validity of models for predicting BRCA1 and BRCA2 mutations.” In: *Ann Intern Med* 147.7 (2007). PMC2423214, pp. 441–450.

Lee, G., J. W. Liang, Q. Zhang, T. Huang, C. Choirat, G. Parmigiani, and D. Braun. “Multi-syndrome, multi-gene risk modeling for individuals with a family history of cancer with the novel R package PanelPRO.” eng. In: *eLife* 10 (2021).

C.2. Somatic Mutation Analysis. In the late 2000’s the availability of the human genome sequence and progress in sequencing and bioinformatic technologies have enabled genome-wide investigations of somatic mutations in human cancers. Pioneering studies were performed at Johns Hopkins in the lab co-led by Vogelstein, Kinzler and Velculescu. I have been the primary statistician for some of these studies, including roles as one of the senior authors for the first genome-wide somatic mutation analyses in cancer and for the first genome-wide multi-modal analysis integrating somatic mutations with other types of genetic alterations (Leary *et al.* 2008). In subsequent related work with (then) postdoctoral fellow Cristian Tomasetti, we formulated a mathematical model for the evolution of somatic mutations in which all relevant phases of a tissue’s history are considered. The model made the prediction, validated by our empirical findings, that the number of somatic mutations in tumors of self-renewing tissues is positively correlated with the age of the patient at diagnosis. Importantly, our analysis indicated for the first time that the majority of somatic mutations in certain tumors of self-renewing tissues occur before the onset of neoplasia, which is the premise of current mutational signature analysis. Lastly, the model also provides for the first time a way to estimate the *in vivo* tissue-specific somatic mutation rates in normal tissues, leveraging sequencing data of tumors.

Wood, L. D., D. W. Parsons, S. Jones, J. Lin, T. Sjöblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, N. Silliman, S. Szabo, Z. Dezso, V. Ustyanksky, T. Nikolskaya, Y. Nikolsky, R. Karchin, P. A. Wilson, J. S. Kaminker, Z. Zhang, R. Croshaw, J. Willis, D. Dawson, M. Shipitsin, J. K. V. Willson, S. Sukumar, K. Polyak, B. H. Park, C. L. Pethiyagoda, P. V. K. Pant, D. G. Ballinger, A. B. Sparks, J. Hartigan, D. R. Smith, E. Suh, N. Papadopoulos, P. Buckhaults, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, V. E. Velculescu, and B. Vogelstein. “The genomic landscapes of human breast and colorectal cancers.” In: *Science* 318.5853 (2007), pp. 1108–1113.

Leary, R. J., J. C. Lin, J. Cummins, S. Boca, L. D. Wood, D. W. Parsons, S. Jones, T. Sjöblom, B.-H. Park, R. Parsons, J. Willis, D. Dawson, J. K. V. Willson, T. Nikolskaya, Y. Nikolsky, L. Kopelovich, N. Papadopoulos, L. A. Pennacchio, T.-L. Wang, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, B. Vogelstein, and V. E. Velculescu. "Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers." In: *Proceedings of the National Academy of Science, USA* 105.42 (2008). PMC2571022, pp. 16224–16229.

Parmigiani, G., S. Boca, J. Lin, K. W. Kinzler, V. Velculescu, and B. Vogelstein. "Design and analysis issues in genome-wide somatic mutation studies of cancer." In: *Genomics* 93 (2009). PMC2820387, pp. 17–21.

Tomasetti, C., B. Vogelstein, and G. Parmigiani. "Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation." eng. In: *Proc Natl Acad Sci U S A* 110(6) (2013). PMCID: PMC3568331, pp. 1999–2004.

C.3. Replicability. Cancer genomics has been facing an important challenge about obtaining consistent results across studies aimed at answering the same scientific question (replicability). In cancer biology and its clinical translation, replicability concerns subtypes (clusters) and predictive / prognostic scores (classification and regression). Since the mid 2000's I developed new concepts and tools for elucidating cross-study replicability, beginning with methods for subtype validation and for selection of replicable features in unsupervised settings (integrative correlation). In the 2010's, in Waldron *et al.* , considering the case study of ovarian cancer prognosis, we designed and executed the first comprehensive analysis of cross-study replicability of prognostic models that are based on genomic data. We adopted the rigor of systematic reviews of clinical trials to develop a blueprint for identifying published prognostic models and evaluate 1) reimplementation as described by the original study, 2) performance for prognosis of overall survival in independent data, and 3) performance compared with random gene signatures. In the process we created data models, software and databases for multi-study analysis of gene expression.

Parmigiani, G., E. S. Garrett-Mayer, R. Anbazhagan, and E. Gabrielson. "A cross-study comparison of gene expression studies for the molecular classification of lung cancer." In: *Clin Cancer Res* 10.9 (2004), pp. 2922–2927.

Ganzfried, B. F., M. Riester, B. Haibe-Kains, T. Risch, S. Tyekucheva, I. Jazic, X. V. Wang, M. Ahmadifar, M. J. Birrer, G. Parmigiani, C. Huttenhower, and L. Waldron. "curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome." eng. In: *Database (Oxford)* 2013 (2013). PMCID: PMC3625954, bat013.

Bernau, C., M. Riester, A.-L. Boulesteix, G. Parmigiani, C. Huttenhower, L. Waldron, and L. Trippa. "Cross-study validation for the assessment of prediction algorithms." eng. In: *Bioinformatics* 30.12 (2014). PMCID: PMC4058929, pp. i105–i112.

Waldron, L., B. Haibe-Kains, A. C. Culhane, M. Riester, J. Ding, X. V. Wang, M. Ahmadifar, S. Tyekucheva, C. Bernau, T. Risch, B. F. Ganzfried, C. Huttenhower, M. Birrer, and G. Parmigiani. "Comparative Meta-analysis of Prognostic Gene Signatures for Late-Stage Ovarian Cancer." eng. In: *J Natl Cancer Inst* 106.5 (2014). PMCID: PMC4580554, dju049.

C.4. Multi-Study High Dimensional Analyses. Work in C.3 was the impetus to develop statistical learning methods that train on multiple studies to achieve better replicability. In unsupervised setting, in Devito *et al.* we extend factor analysis to multiple studies. For the first time, we can separately identify and estimate 1) factors shared across multiple studies, and 2) study-specific factors.

Moving to supervised analyses, in Patil & Parmigiani we proposed a simple and general class of prediction models for multi-study learning. The novelty of this approach consisted in optimally ensembling prediction models, each of which is trained on one of the studies, with optimality criteria that reward generalizability beyond the original training data. This approach provides new techniques compared to both the statistical and machine learning literature. It can be used to address both desired and undesired variation, and to predict a new draw from one of the available studies (similarly to multi-task learning) or a new study altogether (similar to domain generalization). In parallel, we focused on data homogenization and remedial of batch effect, developing the ComBat-seq model for remedial of batch effect in RNA-seq data. The article presenting this method is currently listed as "the most cited" on the Nucleic Acids Research Genomics and Bioinformatics website.

Patil, P. and G. Parmigiani. "Training replicable predictors in multiple studies". In: *Proceedings of the National Academy of Science, USA* 115.11 (2018). PMCID: PMC5856504, pp. 2578–2583.

Zhang, Y., G. Parmigiani, and W. E. Johnson. "ComBat-seq: batch effect adjustment for RNA-seq count data". In: *NAR Genomics and Bioinformatics* 2.3 (2020).
Zhang, Y., P. Patil, W. E. Johnson, and G. Parmigiani. "Robustifying Genomic Classifiers To Batch Effects Via Ensemble Learning". In: *Bioinformatics* (2020). btaa986. ISSN: 1367-4803.
De Vito, R., R. Bellio, L. Trippa, and G. Parmigiani. "Bayesian multistudy factor analysis for high-throughput biological data". In: *The Annals of Applied Statistics* 15.4 (2021), pp. 1723–1741.

C.5. Rational Decision Making in Health Care and Bayesian Analysis. My interests in machine learning for familial risk and in high dimensional modeling for cancer genomics are driven by the desire to improve decision making processes both before and after a cancer diagnosis. I take this motivation most seriously. Since the early stages of my career I contributed scholarly work elucidating the quantitative basis for rational decisions in health care, with special attention to information integration and to uncertainty quantification. This work began with theoretical analysis of optimal cancer screening policies, and continued with complex microsimulation model development, and contributions to assessment of uncertainty in decision and policy analysis. Methodologically I am an expert in Bayesian decision theory. I published a book on "Modeling in Medical Decision Making" and, with L. Inoue, an award-winning text on "Decision Theory".

Parmigiani, G. "On Optimal Screening Ages". In: *Journal of the American Statistical Association* 88 (1993), pp. 622–628.
– "Timing Medical Examinations via Intensity Functions". In: *Biometrika* 84 (1997), pp. 803–816.
Parmigiani, G., S. Skates, and M. Zelen. "Modeling and optimization in early detection programs with a single exam." In: *Biometrics* 58.1 (2002), pp. 30–36.
Boca, S. M., H. C. Bravo, B. Caffo, J. T. Leek, and G. Parmigiani. "A decision-theory approach to interpretable set analysis for high-dimensional data." eng. In: *Biometrics* 69(3) (2013). PMCID: PMC3927844, pp. 614–623.

Complete List of Published Work in MyBibliography:

https://www.ncbi.nlm.nih.gov/myncbi/1derwo-v2hdkm/bibliography/public/