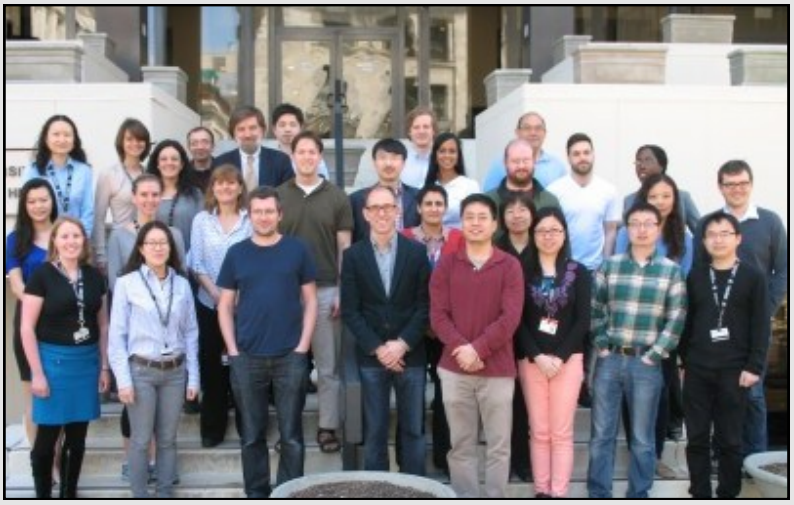
The Minuscule and the Massive

"Our genomes could easily hang on a thumb drive on our necks," muses the Harvard School of Public Health Dean for Academic Affairs, David Hunter, MBBS, MPH, ScD, envisioning an easy-to-wear accessory that would complement personalized medicine's growing satchel of tools. This and other sci-fi scenarios have mushroomed following the triumph of sequencing the human genome in 2001.



The Department of Epidemiology PGSG Group

The exploration of genetic variations and their relation to disease susceptibility––cancer, diabetes, obesity, and mental illness, to name a few––has since been moving full-force. Among the original explorers were genomic scientists performing candidate gene work––an approach based on educated guesses, which looks at the association between specific genes and particular diseases. "It's like throwing darts from a great distance, essentially trying to guess which piece in a huge genome is useful [in determining the link to a disease]," says Hunter, who also holds the Vincent L. Gregory Professorship of Cancer Prevention.

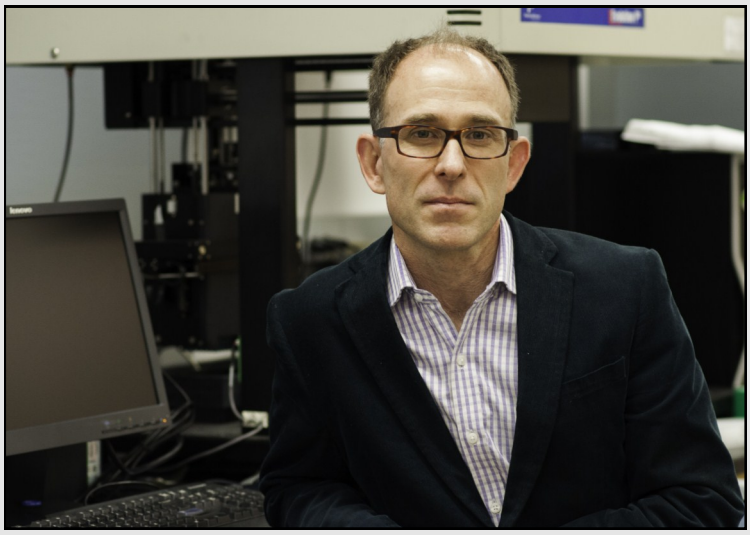
Dean for Academic Affairs, David Hunter, MBBS, MPH, ScD

Into this landscape came a slew of technological developments that empowered genome-wide association studies (GWAS), a methodology that scans hundreds of thousands of genetic variations across the entire genome of thousands of individuals. Accessing these billions of data points has been a game-changer in identifying the location of genetic variants implicated in disease. Subsequently, genetic mapping, a post-GWAS approach, digs into pinpointing the genetic variants with the strongest association to a disease. "If you were to identify New York City as a scene of interest," explains Hunter, "you'd then have to get to the suburb, the street, and then the house to target the exact causal location."

Hunter, Peter Kraft, PhD, Alkes Price, PhD, Liming Liang, PhD, Immaculata De Vivo, MPH, PhD, and other genetic epidemiologists in HSPH's Program in Genetic Epidemiology and Statistical Genetics (PGSG) are applying their assorted big-data skills to hone in on the causes of disease.

PGSG is among the first programs at the School to deal with the massive increases in computing power, data storage and analytic methods to harness large-scale data more efficiently. In the past two decades, PGSG has burgeoned dramatically, building its faculty, expanding its research and curriculum portfolio, studying dozens of diseases, collaborating with hundreds of colleagues worldwide, and establishing HSPH's first genotyping laboratory, enabling colleagues School-wide to conduct a variety of genetic analyses.

One recent project reflects PGSG's standout growth. Professor of Epidemiology Peter Kraft led a recent international study whose lens focused on the epidemiology of estrogen-receptive (ER)-negative breast cancer, an aggressive, difficult-to-treat cancer that accounts for 20%-30% of all breast cancers. Assembling the largest study of ER-negative breast cancer to date, Kraft and his colleagues were able to identify four new genetic markers––sites on the DNA molecule that differ person to person––specifically associated with this subtype. These findings corroborated what researchers had suspected ––that ER-negative and ER-positive cancers are different biological diseases––and provided some intriguing hints about the mechanisms driving ER-negative cancers.



Professor of Epidemiology, Peter Kraft

The ER-negative breast cancer study was one piece of a larger study of breast, prostate and ovarian cancer. The largest of its kind, the study involved 200,000 research participants, hundreds of scientists, and 50 studies across North America and Europe.

As part of this multidisciplinary collaboration, Kraft, Hunter and PGSG research scientist Sarah Lindstrom, PhD, participated in the discovery of over 100 markers associated with the three cancers. The discovery doubles the number of markers previously known for these diseases, which yearly affect half a million Americans, 2.5 million globally, and killing about one-third. The finding also suggests "this is just the tip of the iceberg. There may be 1,000 other variants waiting to be discovered somewhere in the genome," Kraft says. The study was so successful the international group has launched a mammoth, follow-up study involving over half a million participants.

Of 3.3 billion DNA building blocks that comprise an individual's genome, more than 99% are identical with those of any other individual; less than 1% differ person to person. The most common type of genetic variation between people are called SNPs, single nucleotide polymorphisms, which Kraft and scientists are proving to be pivotal in determining human health. "We're studying differences that are individually minuscule, but collectively important," says Kraft, adding that "we've now learned enough about these variants that our findings could lead to targeted screenings and potential treatments." A case in point, he says, is "identifying women who might benefit from earlier breast cancer screening."


Associate Professor Immaculata De Vivo

Coming at cancer from a different perspective is Associate Professor Immaculata De Vivo, who aims her spotlight at the interplay between environment and genetics. "The genes load the gun, but the environment pulls the trigger," she says of the interaction. "Trying to figure out the number of genes implicated, the kind of environmental exposures that have occurred, the temporal windows of those exposures––it all comes into play when trying to understand the complexity of cancer." De Vivo, who holds a joint appointment with the Brigham and Women's Hospital and is director of the Dana Farber/Harvard Cancer Care's High-Throughput Polymorphism Detection Core, has a specific interest in endometrial cancer, the country's most common gynecologic cancer.

De Vivo began her career as a basic scientist. "We've cured a lot of cancers in mice," she says, "but I really wanted to have greater impact. Basic science is incremental; I wanted to be transformative. Bringing expertise from all perspectives is one of the best ways to tackle a problem. As a bench scientist I thought I'd make the biggest contribution in epidemiology."

A key reason De Vivo focuses on endometrial cancer is that its risk factors are well-defined. "The main environmental risk factor for endometrial cancer is excess exposure to estrogen. I thought, 'Let's bring together the well-defined environmental factors and the genes in the hormonal pathway,'" she says. The results were published in a paradigm paper that found genes associated with the disease, as well as evidence that obese women with the same genotype are at higher risk for the disease.

As is characteristic of Hunter, Kraft and other PSGS researchers, collaboration is a linchpin of De Vivo's work. She spearheaded an international, cross-disciplinary consortium of experts in endometrial cancer, biology, pathology, epidemiology, and clinical and health policy practice. Their aim is to synergize findings and plan novel research. The same holds true with her PGSG colleagues. "We marry biological data with epidemiological and statistical data. Investigating 50,000 or 70,000 genetic markers requires different analyses and complete synergy. The markers are elusive and you need statistical power to find them. Big numbers give big power, since big numbers allow you to separate signal from noise. We couldn't do it without other PGSG team members."

Liming Liang, PhD

Adding considerable heft to the power of big numbers are Alkes Price, PhD, and Liming Liang, PhD, both heavyweights in statistical and computational methodology.

Consider a recent Price study. He and his colleagues used complex mathematical methods, never before used for biological purposes, that combine results from various domains, which individual domains do not have the power to answer. His approach offers answers that could not have been devised even three years ago.

Alkes Price, PhD

His complex tools strengthen evidence that causal markers associated with disease are not usually located in genes that code for proteins, which comprise only 1.5%. of the genome. Rather, disease-associated markers are located in 'desert' regions between genes that regulate when and where genes are turned on or off. Researchers have speculated for some time that these regulatory regions play an important role in the development of disease.

By combining data from large association studies, like that conducted by Hunter, Kraft and Lindstom, with results from recent experiments describing the biochemical features of these regions outside genes, Price was able to identify specific regulatory features relevant for disease. "Our study demonstrated how much heritable variation comes from regulatory SNPs. It changed our thinking about genetic architecture. Optimistically, it's a step in understanding where the genetic basis of a disease is coming from and which SNPs make you more susceptible to disease. Everyone wants this information."

In another recent study, Price demonstrated the pros and cons of mixed models, a type of mathematical tool, that point to "red herrings"–– confounding factors––which produce false positives and waste research money. The complicated and powerful statistical model he provided highlighted subtleties researchers should consider to avoid methodological pitfalls and maximize a study's power.

Assistant professor Liang, too, develops sophisticated computational tools to explore epigenetics, a hot area of genetic study in which outside factors––smoking, pollution, exercise, and diet––can turn genes on or off. By focusing on how genetic variation and environmental factors change gene expression Liang's high throughput work offers the promise of therapy and drug treatment. He and his colleagues in England and Canada, for example, discovered that a high expression level on a specific set of genes plays a pivotal role in allergic conditions such as asthma.

The multiple perspectives PGSG brings to fundamental questions of disease reflects epidemiology's growing role in answering those questions. "Our approach is a more participatory style of epidemiology," says Hunter. "The data we provide is critical–– in identifying people at greater or lower risk of a disease, determining which interventions are more effective for which people, and predicting drug toxicity, efficacy, and proper dosage. But we've moved beyond making sample sizes as large as possible and finding associations with disease. We're now active participants in assessing, understanding and explaining the mechanisms causing disease. That's a paradigm shift."

-Orna Feldman