The SNPs are down: genotyping for the rest of us

Human genotyping has never been hotter, and a sophisticated set of array-based tools now simplifies the process dramatically, facilitating everything from small basic research studies to complex genetic epidemiology. Alan Dove reports.

In 1915, Thomas Morgan published the book *Mechanism of Mendelian Heredity* summarizing his work with fruit flies and the revolutionary conclusion that genes are arranged in a linear fashion. Ninety years later, biologists are still obsessed with that linear arrangement, and correlating phenotypes with genotypes is the focus of entire institutes.

Thanks to the confluence of the genome-sequencing and digital revolutions, the genotyping craze is now entering a new phase. Just as printing circuit components on silicon chips spawned an entire branch of engineering, printing nucleic acids on chips is creating a completely new approach to biology.

Geneticists are already switching from time-consuming gels and microsatellite analysis to quick and accurate chip-based systems. Meanwhile, array-based tools are finally becoming available for sophisticated association studies, or 'genetics without families', which some experts claim will revolutionize human gene discovery. Laboratory tinkerers have also hacked some of the new gene chips into platforms for analyzing the copy numbers of genes, revealing surprising new levels of complexity in genetic regulation.

In the laboratory, researchers learning how to use these new tools face a bewildering array of options. Besides a well-designed experiment, genotyping now requires a careful analysis of the available technologies, most of which are optimized for particular types of projects.

The Eureka trio: a set of chips that support Illumina's whole-genome genotyping with their Infinium assay. (Courtesy of Illumina.)

Guilt by association

As a result of the publicly funded HapMap initiative (**Box 1**), researchers around the world now have access to a database of hundreds of thousands of single nucleotide polymorphisms (SNPs), single base differences in the human genome that can be used as genetic markers. Traditional genetics can be done with far fewer markers, but the main goal of HapMap was to allow an entirely new type of gene mapping.

"In family-based studies, you're looking at a couple of generations and really trying to track large chunks of chromosomes through the pedigree, and if you coinherit consistently a chunk of chromosome and a trait, then there must be something on that chromosome," explains Dietrich Stephan, director of the neurogenomics division at the Translational Genomics Research Institute (TGEN; Phoenix, Arizona, USA).

Unfortunately, human geneticists must rely on 'found experiments', in which families with well-characterized pedigrees have also developed well-defined genetic diseases. Complex traits with environmental and genetic components, like heart disease, cancer and mental illness, have been notoriously difficult to study this way.

The completion of the HapMap now permits researchers to test a bold new approach. An association study entails "looking at the population as a whole as one humongous family, with the initial founders being Adam and Eve, or whoever," says Stephan. At TGEN, a nonprofit institute that focuses on association studies, Stephan and his colleagues use new

DNA chips from Affymetrix exclusively. The chips contain 500,000 SNPs apiece, so a single chip-based experiment can characterize a patient's entire genotype at high resolution.

Affymetrix pioneered DNA chip technology, but in the past few years other companies have rushed into the field, often with highly sophisticated products for SNP analysis (**Box 2**). Perlegen and Illumina, for example, were both heavily involved in the HapMap project itself, and are continuously developing new tools for the emerging market of association studies.

Illumina's Golden Gate genotyping assay "was used for over 60% of the HapMap project. More recently, we just launched Beadchip, which is for whole-genome genotyping assays," says Sarah Murray, genotyping manager at Illumina (San Diego, USA). The company's 109,000-SNP system is already on the market, and Murray says they are developing a 250,000-SNP version of the assay as well.

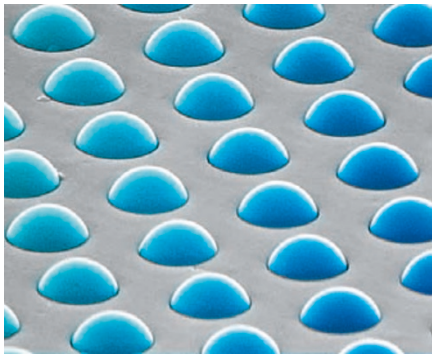
That's only half as many SNPs as Affymetrix's top-end chip, but Illumina argues that sheer SNP quantity is not the most relevant standard for choosing a genotyping system. "We chose to define content to be SNPs in and around genes, [because] you're more likely to find association near a gene," says Murray, adding that the 109,000-SNP product has "greater than 99.9% reproducibility, greater than 99.9% call rates, [and] the inconsistency rate is extremely low."

Playing the odds

Besides choosing a technological platform, another challenge in designing an association study in humans is determining the minimum number of samples. Statisticians suggest that a set of only 200 people should be enough to characterize a disease caused mainly by a single gene variation. More complex traits will require more patients in the study.

As clinical research goes, association studies should be relatively cheap, even with large numbers of patients. Taking blood samples and extracting genomic DNA are simple enough procedures. The main expenses involve the SNP genotyping technology itself, but researchers are already finding new ways to economize.

"There are strategies where you can pool samples and get similar information ... on smaller chips," says Stephan, adding that "this is the difference between spending $2 million on a study and spending $20,000."



Blue beads—the beads in wells that form the basis of the technology platform of Illumina. (Courtesy of Illumina.)

Even as the tools become cheaper and easier to use, though, many geneticists remain unconvinced that association studies will live up to their hype. Though the concept has been proven in relatively simple conditions involving mutations in one or a few genes, the real test will be the characterization of highly complex, multigene traits. Skeptics argue that the number of patients required for an association study of heart disease, for example, would exceed the bounds of any research grant, and that the approach will only work on simpler, much rarer conditions.

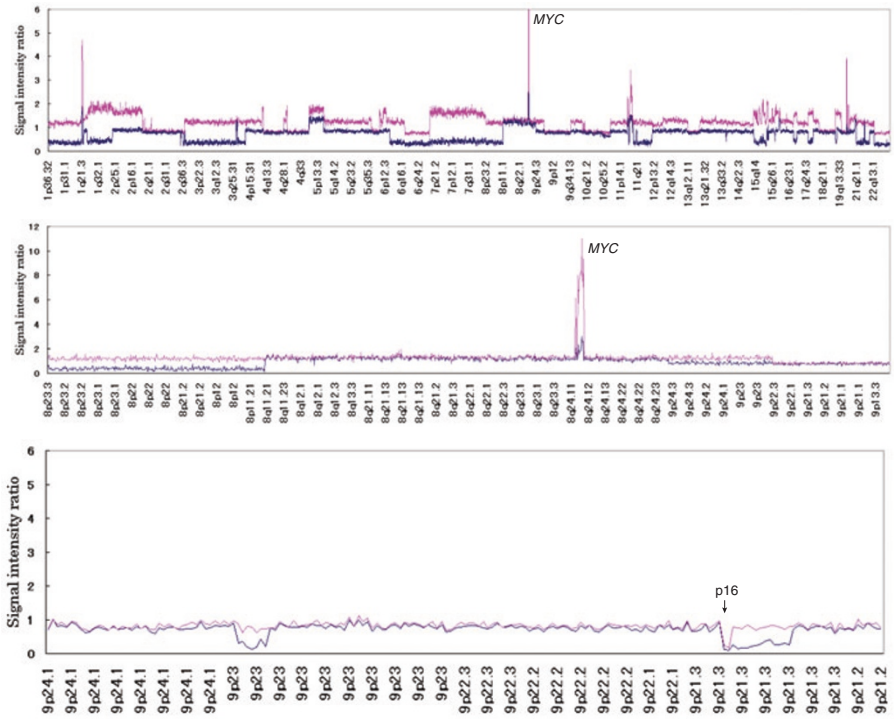
HapMap believers are confident they will soon be vindicated. Stephan advises fence-sitters to watch the SNP mapping literature closely, adding that the results from several pilot studies have already provided ample proof that the system will work on a larger scale.

All in the family

Outside of association studies, human geneticists today primarily rely on a strategy that would have been perfectly familiar to Morgan: linkage analysis. In the decades since the molecular biology revolution began, linkage analysis has mostly meant finding suitable test populations, then mapping the trait of interest based on its proximity to microsatellite sequences.

In the past few years, DNA chips with a carefully selected subset of SNPs have almost completely supplanted the often-painful gel electrophoresis marathon of microsatellite mapping. Most of the major DNA array manufacturers now offer simple, turnkey SNP-based systems for rapid genotyping in linkage studies. With a defined pedigree, the arrays can be magnitudes smaller, and the latest trend is to run multiple genotypes in multiplexed, automated systems.

Not only is the SNP-based linkage analysis faster and easier than microsatellite



Results of allelic dosage analysis performed on a genotyping microarray, showing copy number variation across the genome in two cell lines. Top and middle show an increase in copy number at the *MYC* oncogene locus in one cell line. Bottom, a cell line that harbors a homozygous deletion at the *CDKN2A* (p16) locus has decreased copy number at this locus. (Reprinted from ref. 3. Copyright 2005, with permission from Elsevier.)

BOX 1 WRAPPING UP HAPMAP

Phase 1 of the ambitious International HapMap Project is completed, having identified enough SNPs to place one SNP signpost about every 600 bases in the human genome. Now that the first draft of the map is finished, "there will be lots and lots of genotyping studies in patients," says Lisa Brooks, director of the Genetic Variation Program at the US National Human Genome Research Institute of the National Institutes of Health (NIH; Bethesda, Maryland, USA).

The plummeting costs of rapid, array-based genotyping technologies will drive much of the boom, but SNPs may only be a transitional tool for many researchers. "As sequencing costs drop, sequencing will simply become another technique for genotyping," Brooks predicts.

Though genotyping humans will continue to become easier, phenotyping them will not. For the complex diseases that are the greatest concerns in the developed world, pathogenesis involves a complex interplay of multiple genes, environmental factors and individual physiology. No matter how refined the genotyping technologies become, a study that does not distinguish between different subtypes of a condition may be swamped by irrelevant data. For association studies, experimental and control groups must also share the same ethnicity, or the genotyping results will be measuring the wrong phenotype entirely.

"The other issue, of course, comes ... once you have a region of the genome and you've got a couple of genes in there and a couple of variants," says Brooks. Determining which variant causes a condition may take a combination of larger patient groups, more sophisticated data analysis, and laborious gene targeting experiments in model organisms.

The experts interviewed for this article envision post-HapMap genotyping research following the same pattern as earlier gene expression profiling studies, with journals gradually insisting on more experimental rigor and mechanistic insight as the technology becomes more commonplace. With the HapMap in researchers' hands, the real journey can finally begin.

mapping, it is also more precise. Citing work her company published a year ago[1], for example, Murray concludes that "you get more information using a SNP panel than using a microsatellite panel."

Comparative hacking

Whenever a new technology enters the laboratory, the first instinct of hands-on experimentalists is to throw away the manual and see what the new toy can really do. DNA genotyping arrays were no different.

Soon after the first 10,000-SNP arrays became available, researchers discovered that they could be used for comparative genomic hybridization (CGH), a technique that reveals how many copies of a given gene are actually in a sample. The rise of this new method took chip makers completely by surprise.

"We did not anticipate this, but they've proven to be very useful in this application," says Keith Jones, vice president for molecular genetics at Affymetrix (Santa Clara, California, USA). Jones adds that "now that we have sort of clued in that they're working well, we are thinking about next-generation designs that are more quantitative, higher resolution, and yet continue to include the allelic information that you get from SNP genotyping arrays."

Production run of the Affymetrix chip. (Courtesy of TGEN.)

For any laboratory with access to DNA chips and readers, the method is essentially the same for CGH as for regular genotyping. The main difference is in data analysis, since the copy number of a gene correlates with the intensity of the signal where it hybridizes to the array. By comparing a test sample with a sample that is known to be diploid, the experimenter can determine the ploidy of the test sample.

The most obvious application is in cancer biology, as tumor cells may rearrange or lose chromosomes. Losing regulatory genes in this way can turn a small, slow-growing tumor into an aggressive malignancy. "You can use the array both to classify tumors and to understand the pathophysiology behind them," says Jones.

Because the Affymetrix arrays now used for CGH were originally developed for genotyping, the strategy can also uncover some surprises. For example, cells sometimes adapt to the loss of one chromosome of a homologous pair by duplicating the other chromosome. The cell then contains a normal-looking pair of homologous chromosomes, but the genotyping array reveals that both copies are from the same lineage, a condition called uniparental bisomy. If the lost chromosome contained the only wild-type copy of a gene, and the replicated chromosome contained a mutant copy, the resulting cell will now have the homozygous mutant genotype even though the subject's germline genotype is heterozygous.

Though CGH surprised chip makers initially, they are now embracing the new

BOX 2 ATTENTION ARRAY-MART SHOPPERS

Asked which genotyping platform is best, experts offer an unsurprising answer: it depends on what type of genotyping one intends to do and how much money is available to do it. With several manufacturers now competing in the genotyping field, the cost per array continues to plummet. Setup costs, data quality, and reproducibility, however, can vary dramatically from one platform to another for a given experiment.

Many of the systems that read gene expression arrays can also process genotyping arrays, so researchers may already have the necessary equipment in a core facility. Companies like NimbleGen and Affymetrix routinely build their new products on standard slides or chips, so they will be easy to use in existing array readers.

In typical engineering fashion, though, there are several standards to choose from, and not all of them are compatible. Arrays from Illumina, for example, require an Illumina processing system, which is not interchangeable with an Affymetrix processing system. Laboratories starting into array-based assays from scratch should expect the basic equipment to start in the $250,000 range.

Projects with smaller budgets and fewer samples to process should consider farming the work out. Before calling the companies directly, though, academic researchers should first check with their funding agencies and colleagues. For example, the NIH-funded Neuroscience Microarray Consortium accepts samples from scientists studying the genetic basis of neurological diseases. "Ten thousand scientists have the key to enter this [system]," says Stephan. For about $700, the consortium will genotype a sample on a 500,000-SNP Affymetrix chip, then send back the data.

For-profit contractors, including most of the major array makers, will also process samples for similar prices. Genotyping with fewer SNPs is cheaper, so a linkage study will cost considerably less per sample than a high-resolution association study.

Even laboratories that can afford their own array-processing systems may choose to send the samples out, especially if they are part of a multilaboratory consortium. "[For] some of these large disease consortia … it's just easiest and less amount of coordination if they're all sent to one central genotyping center," says Murray.

"Or there's always the good old collaborative way," says Stephan, adding that "I just put in five grants on October 1, [2005] with various groups." Sharing authorship on a paper is an inexpensive way to get genotyping results and expert assistance analyzing the data.

market and offering a variety of products. Some of these are simply more carefully tailored versions of genotyping arrays, allowing simultaneous genotyping and CGH analysis, but others are dedicated systems that only assess gene copy number.

"People are asking us for a straight CGH approach;... for whatever reason, they find it's better to have a dedicated assay, and they find there [are] improvements in performance or sample preparation methods using a dedicated copy-number assay," says Emile Nuwaysir, vice president for business development at NimbleGen (Madison, Wisconsin, USA).

Indeed, each array manufacturer seems to be taking a slightly different approach to the CGH field, trying to fit the strengths of their specific technologies into the rapidly developing market niches. NimbleGen's products may be an especially good deal for researchers who only need a small number of custom-built genotyping or CGH arrays. "We can create any microarray instantaneously, and all we require is sequence input," says Nuwaysir.

CGH has also gotten a makeover as a new method for DNA sequencing[2]. The technique, called comparative genome sequencing (CGS), begins with a high-resolution CGH experiment, and then proceeds through additional hybridizations that ultimately yield detailed sequence information. "We can sequence and identify all of the changes in a microbial genome in just a few hybridizations," says Thomas Albert, director of molecular research at NimbleGen and lead author on the study.

So far, CGS appears to work on genomes up to 25 megabases, which would allow it to be used on protozoans like the malaria-causing *Plasmodium falciparum*. Albert concedes that the technique may have an upper limit that would preclude sequencing larger genomes, but he and his colleagues have not found it yet.

As with any emerging technology, the early adopters are still finding new uses for genotyping systems, and facing new challenges as they test the boundaries of the present platforms. Whether the project is an association study, linkage analysis, CGH or CGS, the array-based tools continue to get cheaper, faster and more accurate. The HapMap project will have officially ended phase 1 by the time this article goes to press, but with new systems and applications still being developed, the era of rapid genotyping is clearly just getting started.

1. Murray, S.S. *et al. Nat. Methods* **1**, 113–117 (2004).
2. Albert, T.J. *et al. Nat. Methods* **2**, 951–953 (2005).
3. Ishikawa, S. *et al. Biochem. Biophys. Res. Commun.* **333**, 1309–1314 (2005).

Alan Dove is a freelance writer based in New York City
alanwdove@earthlink.net

SUPPLIERS GUIDE: COMPANIES OFFERING DNA ARRAY TECHNOLOGIES

Company	Web address
Affymetrix	http://www.affymetrix.com/
Agencourt Bioscience	http://www.agencourt.com/
Agilent	http://www.chem.agilent.com/scripts/PHome.asp
Alpha Innotech	http://www.alphainnotech.com/
Applied Biosystems	http://www.appliedbiosystems.com/
Beckman Coulter	http://www.beckmancoulter.com/products/specifications/geneticanalysis/ceq/default.asp
BioDiscovery	http://www.biodiscovery.com/
Bio-Rad Laboratories	http://www.bio-rad.com/
Brooks Automation (formerly Intelligent Automation Systems)	http://www.ias.com/
Cartesian Technologies	http://www.cartesiantech.com/
Clondiag Chip Technologies	http://www.clondiag.com/
CombiMatrix	http://www.combimatrix.com/
Corning	http://www.corning.com/lifesciences/US-Canada/en/
GE Healthcare (Amersham)	http://www1.amershambiosciences.com/aptrix/upp01077.nsf/content/na_homepage
Genaissance	http://www.genaissance.com/
Genomic Solutions	http://www.genomicsolutions.com/showPage.php
Illumina	http://www.illumina.com/
Invitrogen	http://www.invitrogen.com/
Marligen Biosciences	http://www.marligen.com/
Mergen	http://www.mergen.com/
MiraiBio	http://www.miraibio.com/
Molecular Devices	http://www.moleculardevices.com/
Nanogen	http://www.nanogen.com/
Nimblegen	http://www.nimblegen.com/
Orchid Biosciences	http://www.orchid.com/
Panomics	http://www.panomics.com/
Parallele Bioscience	http://www.parallelebio.com/
Perkin Elmer	http://las.perkinelmer.com/
Perlegen	http://www.perlegen.com/
Phalanx Biotechnology	http://www.phalanxbiotech.com/
Plexigen	http://www.plexigen.com/
Polymorphic DNA	http://www.polymorphicdna.com/
Promega	http://www.promega.com/
Qiagen	http://www.qiagen.com/
Robodesign	http://www.robodesign.com/
Rosetta Biosoftware	http://www.rosettabio.com/
Rosetta Inpharmatics (a subsidiary of Merck)	http://www.rii.com/
SeqWright	http://www.seqwright.com/
Sequenom	http://www.sequenom.com/
Sigma-Genosys	http://www.sigma-genosys.com/
Spectral Genomics	http://www.spectralgenomics.com/
SpectruMedix	http://www.spectrumedix.com/
Stratagene	http://www.stratagene.com/
TeleChem/ArrayIt.com	http://www.arrayit.com/
TmBioscience	http://www.tmbioscience.com/
V&P Scientific	http://www.vp-scientific.com/
Vysis	http://www.vysis.com/
Xenopore	http://www.xenopore.com/