One tube, one SNP	450
Through the golden gate	452
And beyond	452
Box 1: Beyond the common SNP	448
Box 2: More CNVs on the horizon	450

SNP genotyping: six technologies that keyed a revolution

Jeffrey Perkel

With abundant sequencing data, falling prices and mature genotyping platforms, researchers have more options than ever to explore the connections between genes and phenotype.

Thus far in 2008, geneticists have mapped susceptibility loci for, among other things, prostate cancer^{1–3}, bipolar disorder⁴, obesity⁵, height⁶ and eye color⁷.

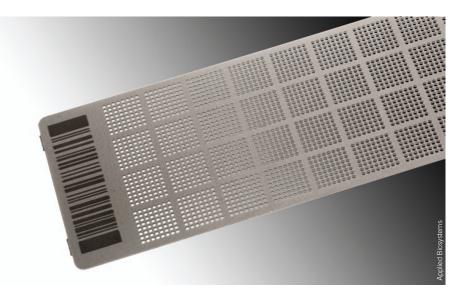
Key to these studies are single nucleotide polymorphisms (SNPs). Millions of human SNPs have been discovered in recent years-over 6 million from The International HapMap Project alone in its first 3 years. This effort has enabled researchers "to look intelligently at the genome," says Stephen Chanock, chief of the laboratory of translational genomics and director of the core genotyping facility at the National Cancer Institute in Bethesda, Maryland, USA. Companies have also responded, offering genotyping tools to accommodate varying sample throughputs, multiplexing capabilities and chemistries.

Musing on the blistering pace of genotyping innovation over the past 7 years, Chanock, who is also senior author of one of three recent prostate cancer geneassociation studies, says, "It really is like a revolution, a dynamic process that continues on."

What is in a SNP?

The most common form of genetic variation between individuals, SNPs occur once every 1,000 bases or so (**Box 1**). Millions of these variants are indexed in the National Center for Biotechnology Information's dbSNP database, covering organisms from *Anopheles gambiae* to *Zea mays*.

Not just any base difference between two individuals is a SNP; a variation is only called a polymorphism if it occurs in 1% or more of the population. "If the polymorphism is stable, then it is a SNP,"



High-throughput genotyping will soon be available on BioTrove's Open Array platform.

explains Richard Leach, vice president for scientific services at deCODE Genetics in Reykjavik, Iceland. "If it arises *de novo* in an individual and isn't propagated in the population, then it is a mutation."

Most SNPs occur outside proteincoding regions and thus are phenotypically silent—the equivalent of mile markers on the side of the highway; others ('nonsynonymous SNPs') affect protein sequence. Both types of SNPs can serve as landmarks in the search for genes associated with disease, drug response and complex phenotypes.

Battle of the chips

The most efficient way to link a SNP with phenotype is the so-called genome-wide association study, in which hundreds of thousands or even millions of polymorphisms are scanned per sample. The tools of the trade are DNA microarrays, and researchers have mostly aligned behind two competing technologies from San Diego–based Illumina and Affymetrix of Santa Clara, California, USA.

Affymetrix's Genome-Wide Human SNP Array 6.0 includes probes for 906,600 SNPs and 946,000 non-polymorphic copynumber probes; Illumina's High Density Human 1M-Duo chip probes more than 1 million polymorphic genomic features on each of two samples, all of which may also be used for copy-number analysis (**Box 2**).

Despite their similarity in format, size and application, the two products differ substantially. For one thing, the Illumina arrays use 50-mer oligos, one per SNP—compared to Affymetrix's 25-mers,

1pg

BOX 1 BEYOND THE COMMON SNP

"The next challenge, or one of the next challenges, is rare variants," says Panos Deloukas of the Wellcome Trust Sanger Institute in Cambridge, England. Most SNP content available today consists of relatively common variants—those present in 5% or more of the population. Other genomic variants can be more informative and yet are relatively underrepresented with existing tools.

These polymorphisms (which are present in as little as 0.5% of the population) could prove much more informative than common ones, as under some conditions, they are more tightly associated with—or even causative for—disease and phenotype. A SNP that causes a glycine to arginine change in the nucleotide-binding oligomerization domain protein 2 (NOD2), for instance, has a 1% frequency but causes a six-fold increased risk of Crohn's disease.

"Rare variants are really only found by sequencing," says Stacey Gabriel of the Broad Institute. "While some SNPs on the HapMap are rare, the set of SNPs that we start with are biased toward being common because they have greater chance to have been found in the sequencing that originally created the SNP catalog."

Next-generation sequencing platforms from such companies as Illumina, 454 Life Sciences (part of Roche Applied Science) in Branford, Connecticut, USA, Helicos BioSciences of Cambridge, Massachusetts, USA, and Applied Biosystems, are now helping to fill these gaps with an abundance of low-cost, high-volume sequence data.

And even more sequence information should be coming in the near future from the recently announced 1,000 Genomes Project, which will sequence the genomes of at least 1,000 normal individuals to increase knowledge of rare genetic variants. The international effort will occupy sequencers at Washington University in St. Louis, the Broad Institute, Baylor College of Medicine (Baylor, Texas, USA), the Beijing Genomics Institute in China and the Wellcome Trust Sanger Center.

Some researchers are also taking smaller-scale, targeted approaches to finding rare variants. According to David Cox, chief scientific officer at Perlegen Sciences in Mountain View, California, USA, Perlegen researchers have used 454's sequencing platform to identify very rare variants in 57 genes from 300 individuals to study side effects to PPAR-gamma agonists, identifying two candidate genes for further study. Jeffrey Perkel

of which there are about 4–6 replicate probes per allele. In addition, although Illumina's Infinitum assay, which runs on its 1M-Duo chip, uses single-base extension with a labeled base to call the SNP, Affymetrix's calls are based exclusively on differential hybridization.

But the most important distinction involves the two platforms' SNP-selection strategies. Illumina's probes are based almost entirely on haplotype-tagging 'tagSNPs' identified by the International HapMap Consortium. Only about half of the SNP probes on the Affymetrix array are tagSNPs, however; the rest are 'unbiased' SNPs chosen to cover the genome while accommodating sequence restraints imposed by the assay itself. Affymetrix's protocol includes a "complexity reduction" step involving selection of relatively small (200–1,100 bp) restriction fragments before hybridization. Effectively, only SNPs located within these regions can be monitored, though the company says the assay still provides 90% genomic coverage, at least in Caucasian and Asian populations.

"There is a certain amount of bias in the selection and amplification," says Jessica Tonani, Affymetrix's associate director of DNA product marketing. "But the purpose is to cover all the common haplotypes, and with our current design, we are able to sample one, and often more than one, tag for each common haplotype."

Researchers definitely have their favorite platforms, whether governed by convenience, price or content. But Stacey Gabriel, director of the genetic analysis platform at the Broad Institute in Cambridge, Massachusetts, USA, whose facility uses both platforms and who was involved in the development of the Affymetrix 6.0 array, says the question of content is largely overblown. "You can make a big deal about SNP-selection



strategy, but ultimately that is not what predicts success," she says, "especially now, when we live in a world where we genotype 1 million SNPs."

Instead, she says, success in genomewide association studies is governed by statistical power, which comes from increasing sample numbers. Although detecting strong associations requires relatively few samples, it may take thousands of samples to tease out lower-penetrance effects. Typically, for cost reasons, that is accomplished by performing multistage studies. In Chanock's prostate cancer study, for instance, researchers scanned half-a-million SNPs in 1,150 affected individuals and 1,150 normal controls, followed by a subset analysis of 27,000 markers in another 8,000 individuals.

Gabriel says her facility can process "about 2,000 whole-genome samples per week." Though the Broad Institute has invested in both Affymetrix and Illumina platforms, it has historically been a larger-volume user of Affymetrix chips—she says, a decision that was driven largely by "the desire to maximize the number of samples that could be successfully scanned for a given budget."

deCODE, which processes as many as 10,000 samples per month, favors Illumina arrays, says Leach, citing their "higher call rate" and "better information content."

Kimberly Doheny, assistant director of the Center for Inherited Disease Research (CIDR) at Johns Hopkins University School of Medicine in Baltimore,



Latest Affymetrix SNP genotyping chip incorporates SNPs as well as non-polymorphic copy-number probes.

Maryland, USA, also prefers Illumina, though she uses both platforms.

That decision dates back to 2003, she explains, when her lab compared a 10,000 SNP array from Affymetrix to a 6,000 SNP product from Illumina. "When we compared the two, Illumina was a lot cheaper and a lot more flexible. We could do custom and off-the-shelf products with the same equipment, and use the same chemistry."

Doheny's lab processed some 70,000 samples for CIDR in 2007, she says, including both off-the-shelf Illumina genomewide association study arrays and custom Illumina products called iSelect arrays, which are physically identical to Infinium arrays and can include anywhere from

6,000 to 60,000 SNPs per sample, with 12 samples per chip. This throughput clearly places Doheny's lab at the higher end of the multiplexing spectrum. But even for the many biologists who are not considering genome-wide association studies, SNP technology, in a variety of low-multiplexing flavors, has also made an impressive difference.

One tube, one SNP

SNP technology development "has been a godsend to those of us with smaller budgets in wildlife genetics," says Jim Seeb of the School of Aquatic and Fishery Sciences at the University of Washington, Seattle, USA. To look at a handful of SNPs, Seeb uses Applied Biosystems' of Foster City, California, USA, PCR-based TaqMan chemistry for his research into the migration of pacific salmon—genotyping fish by the thousands to help manage the American and US-Canadian treaty fisheries.

TaqMan probes are designed to hybridize to a specific SNP allele, with a different 5' fluorophore color for each allele. As a specific color or both colors light up during amplification, the genotype at the particular SNP can be easily determined.

TaqMan is currently a singleplex reaction: one tube, one SNP. It can be multiplexed to 3 or 4 SNPs per reaction with additional fluorescent colors, but according to Phoebe White, senior director of genotyping applications for Applied Biosystems, it's not likely to multiplex further, as that would require new chemistries and more sophisticated readers. Workflow enhancements have emerged, though. In November, Applied Biosystems announced a collaboration with Woburn, Massachusetts, USA–based BioTrove to develop an integrated platform for high-throughput genotyping based on BioTrove's OpenArray architecture, which is capable of 3,072 33-nl PCR reactions on a single microscope slide.

Seeb's lab, which used to handle close to 1,500 384-well PCR plates per month at a cost of nearly \$250,000 per year, has reduced its costs thanks to its acquisition of a Biomark system from Fluidigm of South San Francisco, California, USA. Using the Biomark along with Fluidigm's 48.48 Dynamic Array 'integrated fluidic circuit' consumable, Seeb's lab can run 2,304 TaqMan reactions—the equivalent

BOX 2 MORE CNVs ON THE HORIZON

With most common human SNPs identified, researchers are turning their sights toward copy-number variations (CNVs), such as deletions, insertions and variable number repeats, to expand the genotyping toolkit when searching for disease-causing mutations.

Although researchers agree there are many genotyping platforms that can reliably detect CNVs in the hundreds-ofkilobases size range, it is quite a different story if the goal is to perform genome-wide association studies using CNVs, where it is necessary to call as many CNVs as possible as accurately as possible. "There are major challenges," says Stephen Scherer from the Hospital for Sick Children in Toronto, Canada. "In fact, there is nothing out there that allows you to do [genome-wide association studies robustly] right now."

Only a couple of recent genotyping platforms, such as Affymetrix's SNP Array 6.0 and Illumina's High Density Human 1M-Duo, include CNV probes in combination with SNP probes. Other platforms, including Nimblegen HD2.1, Agilent Technologies' (Santa Clara, California, USA) High-Definition CGH arrays and the Affymetrix tiling arrays, do not offer specific CNVs, rather the tiling design of these comparative genome hybridization arrays (aCGH) can be used to identify copy-number differences to some extent. Scherer is confident more arrays using CNV probes will be commercially available very soon, which could make these genome-wide association studies more commonplace. But he suspects at the moment most companies are waiting for new datasets cataloging copy-number variants before designing specific arrays for CNV analysis.

Scherer, along with Nigel Carter and Matt Hurles from the Wellcome Trust Sanger Institute in Cambridge, UK, and Charles Lee at Harvard Medical School in Boston are currently working on finding more CNVs through their Copy Number Variation Consortium. Using sets of arrays based on the Nimblegen 2.1 million feature HD array, they are scanning over 42 million features across the genome to catalogue variants with over 5% frequency in the human population. "We are finding, at very high confidence levels, somewhere over 1,000 CNVs per genome using these new arrays," notes Scherer. He says they are now finishing the analysis of the data and hope to release the results sometime this year.

Taking a different approach, Evan Eichler's group at the University of Washington in Seattle is using fosmid paired-end sequencing to discover fine-scale genomic variations. Eichler is also part of The Human Genome Structural Variation Working Group, a project launched by the National Human Genome Research Institute to map structural variations within the human genome. "The plan was to tackle 48 individuals using fosmid libraries, build clone libraries from each fosmid and then use pair-end sequencing to identify types of variation," explains Eichler. Once a CNV is identified, the teams go back and completely sequence clones to get single-base-pair resolution.

Eichler thinks that the two projects complement each other in the hunt for new CNVs. He notes that The Human Genome Structural Variation initiative's sequencing approach can detect new CNVs that are not present on a reference genome, but the tiling-microarray approach allows finer resolution, down to 500 base pairs, than the 5-kilobase resolution of fosmid sequencing. "In the end, it is going to be really nice to compare both sets of data," says Eichler.

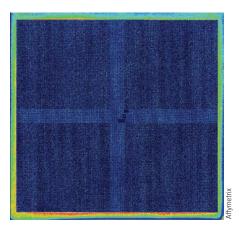
The most difficult structural variations to detect remain balanced changes, either translocations or inversions, which do not change the overall copy number of the genomic region they affect. "I have seen a few innovative approaches that could allow for screening of balance changes, but I think most of those will come through low-pass DNA sequencing," says Scherer. Eichler agrees and thinks additional technology development is still needed to detect these variations, noting that using fosmid sequencing, they have found a couple hundred inversions to date from the ten individuals with completed maps. "Inversions are the hardest nuts to crack." Nathan Blow of six 384-well plates—simultaneously, using nanoliter reagent and sample volumes. Spending on TaqMan reagents is down about 98%, he says.

Polarizing data

Raymond Miller, assistant research professor in genetics and head of the SNP research facility at Washington University in St. Louis, uses another singleplex SNP assay in his lab, one of several genotyping facilities on campus.

Developed at Washington University in St. Louis by Pui-Yan Kwok, who is now at the University of California, San Francisco, and commercialized by Perkin Elmer of Waltham, Massachusetts, USA as the Acycloprime-FP SNP Detection system, fluorescence polarization-templatedirected dye incorporation (FP-TDI) is a single-base extension technology, sometimes called mini-sequencing.

"The selling point of the technology is it is extremely flexible in design, [requiring] three plain vanilla primers," says Miller.



Typical output from the Genome-Wide Human SNP Array 6.0.

Two primers are used to amplify the SNPcontaining sequence; the third hybridizes one nucleotide upstream of the SNP.

After amplification, the third primer is added, along with fluorescent nucleotide terminators corresponding to the two alleles and a polymerase.

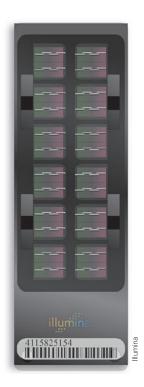
TECHNOLOGY FEATURE

Detection is based on the different fluorescent polarization properties of the incorporated and unincorporated nucleotides. "The free dye in solution is a small molecule and tumbles quickly," Miller explains. "When you shine polarized fluorescent light on it, that causes the light to become unpolarized, which the machine can detect with filters. If the dye gets incorporated, that's a much bigger molecule, so the light comes back as largely polarized."

"What FP-TDI is very good at is detecting a small number of SNPs with a fairsized number of samples," says Miller, whose lab is set up to run sixteen 384-well plates' worth of reactions per day, using a PerkinElmer EnVision fluorescence polarization reader. "Our typical user is doing pilot studies," he says. "They are looking for an association, but using a limited number of candidate genes."

But Miller says demand for his facility's services have fallen off lately, as researchers avail themselves of other, more multiplexed genotyping services





A sketch of the Human610-Quad chip, which contains 610,000 SNPs and 610,000 different bead types per array.

on campus—especially for genome-wide association studies.

Through the golden gate...

In addition to genome-wide association studies, Johns Hopkins' Doheny also uses a completely different Illumina chemistry for genotypes at the lower end of the multiplexing spectrum—the GoldenGate assay.

The assay requires three oligonucleotides, two of which are specific for the two SNP alleles; the third is a 'locus-specific oligo', which is tagged with a nucleic acid barcode to identify the reaction. Once the allele- and locus-specific oligos have hybridized to the genomic DNA, they are linked using DNA polymerase and ligase, PCR-amplified using fluorescently labeled oligos, and bound to one of 1,536 beads (each complementary to one of the barcodes in the locus-specific oligo) for genotype calling.

"The bead defines the assay, and the color defines the base call," explains Carsten Rosenow, senior marketing manager for DNA analysis products at Illumina.

GoldenGate's applications dovetail with its unique combination of the level of SNP multiplexing and sample throughput, and include both validating genome-wide association studies analyses and 'candidate-gene studies', in which one or more particular genes is being specifically tested for association with some phenotype.

...And beyond

Another lower-level multiplexing option is the iPLEX Gold assay from San Diegobased Sequenom, which typically runs a 36-plex format, according to chief scientific officer Charles Cantor. Sequenom's MassARRAY mass spectrometer, which processes the reactions, can accommodate two 384-position matrix-assisted laser desorption/ionization (MALDI) target plates at once and handle about 10 plates per day, he says, meaning users can process in excess of 138,000 SNPs daily.

Like FP-TDI, iPLEX is a single-baseextension assay. After PCR across the SNP and annealing of a third primer, which binds one position upstream of the SNP, a pool of 4 terminator bases is added, one of which is enzymatically incorporated depending on the SNP. Genotype calls are based on the mass of the resulting product. "With terminator nucleotides that differ in mass by at least 12 daltons," Cantor says, "you don't need a high-resolution instrument. This is like shooting fish in a pond as far as mass spec is concerned." According to Cantor, Sequenom positions iPLEX as the technology of choice for such second-tier applications as validating hits after (first-tier) genome-wide association studies. That is because arrays typically are too expensive to run on many, many samples, whereas singleplex technologies like TaqMan are too cumbersome to use for many SNPs. "What we increasingly find is that most users use an array and follow up with the Sequenom platform because that's the most cost-effective way," he says. "Arrays are not flexible, whereas Sequenom is very flexible."

The Broad Institute has four Sequenom systems to complement its Affymetrix and Illumina instrumentation. "Each [platform] is dedicated to certain things," says Gabriel. "For instance, Sequenom is very well suited in our hands for very highly targeted genotyping experiments."

"For us, over 500 is kind of a breakpoint," she explains. "It's more cost-effective to do Illumina [over 500], and below that, it is more effective to do Sequenom. You balance cost and throughput."

Other technical details must also be weighed when selecting a genotyping platform, says Panos Deloukas, senior investigator and head of genotyping at the Wellcome Trust Sanger Institute in Hinxton, Cambridge, UK.

TaqMan and iPLEX Gold, for instance, require an initial amplification step.



iPLEX Gold is a SNP genotyping assay that runs on Sequenom's MassArray mass spectrometer.

"Because they start with PCR amplification, they do suffer from the intrinsic failure rate of PCR. That can vary from lab to lab, but a 3% failure rate is quite the norm," he says. That means call rates can suffer somewhat with these approaches.

But he also notes that although chipbased assays can boast call rates above 99%, they are relatively cumbersome and expensive to optimize and retool. "This is the price you pay to operate at one end of the spectrum versus the other," Deloukas says. Still, with so many options available, researchers can always find the right tool to meet their needs. And that will surely lead to ever more advances making their way into the genetics literature.

Chanock says it could not be a better time to be in genomics: "I get up every morning and can't wait to get to work and see what's going on."

- 1. Gudmundsson, J. *et al. Nat. Genet.* **40**, 281–283 (2008).
- 2. Eeles, R.A. et al. Nat. Genet. 40, 316-321 (2008).
- Thomas, G. et al. Nat. Genet. 40, 310–315 (2008).

- Sklar, P. et al. Mol. Psychiatry, published online 4 March 2008 (doi:10.1038/sj.mp.4002151).
- Liu, Y.J. et al. Hum. Mol. Genet., published online 5 March 2008 (doi:10.1093/hmg/ddn072).
- 6. Sanna, S. et al. Nat. Genet. 40, 198–203 (2008).
- Kayser, M. et al. Am. J. Hum. Genet. 82, 411–423 (2008).

Jeffrey Perkel is a science writer based in Pocatello, Idaho, USA (jeff@jeffreyperkel.com).

Nathan Blow is the technology editor for *Nature* and *Nature Methods* (n.blow@boston.nature.com).



SUPPLIERS GUIDE: COMPANIES OFFERING GENOTYPING PRODUCTS

SUPPLIERS GUIDE: COMPANIES OFFERING GENOTYPING PRODUCTS	
Company	Web address
454 Life Sciences	http://www.454.com
ACGT Inc.	http://www.acgtinc.com
Affymetrix Agencourt Bioscience Corp	http://www.affymetrix.com http://www.aqencourt.com
Agencourt Bioscience Corp.	
Agilent Technologies Applied Biosystems	http://www.agilent.com http://www.appliedbiosystems.com
ATLAS BioLabs	http://www.atlas-biolabs.de
Beckman Coulter	http://www.beckmancoulter.com
BioDiscovery	http://www.bockmancouter.com
BioTrove	http://www.biotrove.com
Bruker Daltonics	http://www.bdal.com
Cogenics	http://www.oogenics.com
deCODE Genetics	http://www.decode.com
Ellipsis Biotherapeutics	http://www.ellipsisbio.com
Enzo Life Sciences	http://www.enzo.com
EpigenDx	http://www.epigendx.com
Expression Analysis	http://www.expressionanalysis.com
Fluidigm	http://www.fluidigm.com
Geneservice	http://www.geneservice.co.uk
GeneWorks	http://www.geneworks.com.au
Genizon Biosciences	http://www.genizon.com
GenoLogics	http://www.genologics.com
GenScript Corp.	http://www.genscript.com
Illumina	http://www.illumina.com
IMGM Laboratories	http://www.dap-healthcare.com
Invitrogen	http://www.invitrogen.com
KBioscience	http://www.kbioscience.co.uk
Laboratory Corporation of America	http://www.labcorp.com
Marligen Biosciences	http://www.marligen.com
MiraiBio	http://www.miraibio.com
Nanogen	http://www.nanogen.com
Ocimum Biosolutions	http://www.ocimumbio.com
Orchid Cellmark	http://www.orchidbio.com
PerkinElmer Life & Analytical Sciences	http://las.perkinelmer.com
Perlegen Sciences	http://www.perlegen.com
Plexigen	http://www.plexigen.com
Precision Biomarker Resources	http://www.precisionbiomarker.com
Premier Biosoft	http://www.premierbiosoft.com
Progeny Software	http://www.progenygenetics.com
Promega	http://www.promega.com
Qiagen	http://www.qiagen.com
Roche Applied Science	http://www.roche-applied-science.com
Roche Nimblegen	http://www.nimblegen.com
Sequenom	http://www.sequenom.com
SeqWright	http://www.seqwright.com
Sigma-Aldrich	http://www.sigmaaldrich.com
Third Wave Technologies	http://www.twt.com

P