

Table 1 Comparison of studies demonstrating massively parallel capture and sequencing of targeted genomic regions

Capture approach	Targeted DNA (Mb)	Raw sequence (fold coverage)	Percentage of target passing criteria for variant detection	Reference
Microdroplet PCR	1.4	~270×	94%	1
Molecular inversion probes	1.4	~400×	75%	5
Solution hybridization	3.7	~230×	89%	6
Chip hybridization	26.6	~240×	96%	7

The microdroplet PCR approach of Tewhey *et al.*¹ was used to target a similar amount of genomic sequence as methods based on molecular inversion probes⁵ and solution hybridization⁶ (1–5 Mb), whereas a study that used chip hybridization⁷ sequenced the majority of human protein-coding exons (~27 Mb). All data were obtained using the Illumina GA-II platform. The sequence coverage represents the generated sequence data that was used for mapping and was estimated from details provided in the references indicated.

99.1% of 2,390 comparisons, indicating a high level of accuracy for variant detection at this depth of sequencing. The authors conclude that the coverage and genotyping accuracy of microdroplet PCR is similar to that obtained using conventional PCR.

Comparisons of microdroplet PCR with other targeted capture methods are difficult, in part because previous studies used different targets (with different GC content and homologs in the genome) and different depths of sequencing coverage. Furthermore, it is not always known whether the probes and primers used for capture were chosen for optimal functionality. Nevertheless, the authors note that capture of off-target DNA and the accuracy of variant calling by their method are in the same range as those of enrichment strategies based on molecular inversion probes⁵, solution hybridization⁶ or chip hybridization⁷ (Table 1). Uniformity of coverage for the microdroplet PCR method permits a high variant detection rate at a given level of sequence coverage (Table 1, column 4).

The value of uniformity of coverage is determined largely by the relative costs of the capture and sequencing technologies. In other words, poor uniformity in the yield of captured DNA can often be salvaged by deeper sequencing if the sequencing costs are a minor component of the overall budget. Given that the capture efficiency and variant calling accuracy appear so similar for the different capture technologies, it is likely that the choice of one platform over another will largely be determined by their relative operational costs. The potential of microdroplet PCR to be highly automated should reduce labor costs. However, the relative costs of reagents for the different capture methods are at present unclear, particularly as some systems may be more amenable to multiplexing.

Preliminary data indicate that future runs of the microfluidic instrument might be scalable to 20,000 PCR products by including multiple primer pairs in each droplet. However, even this

level of throughput is approximately one-tenth of that required for amplification of the human exome. The potential to increase the length of amplified products using long-range PCR is limited by the current requirement to shear the template DNA into 2- to 4-kb fragments, making the present version of the microdroplet method best suited for amplification of exon subsets (e.g., CAN genes, which are frequently mutated in tumors²) or specific genomic loci in the megabase size range.

In addition, although microdroplet PCR appears to capture less off-target DNA and hence yields less unwanted sequence data than other target capture strategies^{5–7}, it is limited by the need to amplify and sequence the flanking sequences of any target of interest. Indeed, the targeting of 457 exons for validation of the

process yielded products that consisted mainly of flanking intron sequences. Consequently, in designing primers one must compromise between the genomic locations that are optimal for PCR priming and those that are closest to the target sequence.

Finally, some questions remain regarding the ability of microdroplet PCR to amplify genomic loci with very high or very low GC content. It is unclear whether it will be possible to identify one set of reaction conditions that can be applied to all targets or whether it will be necessary to conduct multiple amplification reactions that would be combined before sequencing.

With the explosive development of second-generation sequencing technologies, rapid analysis of entire exomes or subsets of exomes is now feasible. Until the costs of whole-genome sequencing fall substantially, the study of genome sequences in large cohorts will require optimized methods for targeted DNA capture. The high uniformity of coverage and comprehensiveness provided by microdroplet PCR offers a valuable new approach for focused genome analysis.

1. Tewhey, R. *et al. Nat. Biotechnol.* **27**, 1025–1031 (2009).
2. Sjoblom, T. *et al. Science* **314**, 268–274 (2006).
3. Hodges, E. *et al. Nat. Genet.* **39**, 1522–1527 (2007).
4. Albert, T.J. *et al. Nat. Methods* **4**, 903–905 (2007).
5. Turner, E.H. *et al. Nat. Methods* **6**, 315–316 (2009).
6. Grnirke, A. *et al. Nat. Biotechnol.* **27**, 182–189 (2009).
7. Ng, S.B. *et al. Nature* **461**, 272–276 (2009).

A systems view of host defense

Daniel E Zak & Alan Aderem

Large-scale perturbations unravel the complex networks of activated dendritic cells.

The body's first line of defense against infection is the innate immune system, which recognizes conserved molecular patterns on microbes via receptors such as toll-like receptors (TLRs)¹. TLRs transduce the information into pathogen-specific immune responses involving networks comprising ~2,000 genes² (Fig. 1). A new study by Amit *et al.*³ in *Science* describes an important advance in elucidating these networks. By combining expression profiling and large-scale perturbations, the

authors discover many novel regulators and interactions that might control the physiological processes induced by TLRs. These network components represent novel candidates for detailed analysis and potential targets for the development of vaccines and antimicrobial or anti-inflammatory drugs.

Addressing the complexity of innate immunity requires the large-scale approaches of systems biology. Previous work has focused primarily on the responses of the transcriptome to TLR activation. Our group has studied these responses in macrophages, integrating transcription factor binding site analysis and dynamic computational modeling to identify small novel regulatory networks

Daniel E. Zak and Alan Aderem are at the Institute for Systems Biology, Seattle, Washington, USA.
e-mail: dzak@systemsbiology.org

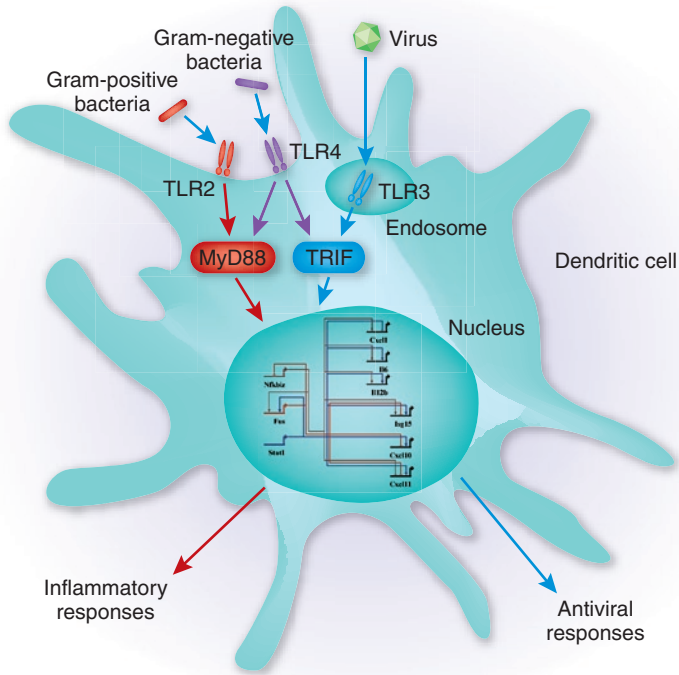


Figure 1 Toll-like receptor (TLR)-activated gene regulatory networks. Innate immune cells, such as dendritic cells and macrophages, detect microbes through TLRs and other receptors. For example, TLR2 detects cell wall components common to Gram-positive bacteria, TLR4 recognizes lipopolysaccharides common to Gram-negative bacteria and TLR3 recognizes double-stranded RNA specifically associated with viral infection. Signals from TLRs are channeled through two major cytoplasmic adaptors, MyD88 and TRIF, which activate the transcriptional responses that control complex gene regulatory networks. The nature of the microbial component detected determines the complement of induced genes and the ultimate functional response (such as inflammatory or antiviral responses). The gene regulatory network depicted represents a subset of the interactions identified by Amit *et al.*³.

that were validated *in vitro* and *in vivo*^{4,5}. In this manner we showed that ATF3 is a negative regulator of TLR4 signaling⁴, whereas C/EBP δ , in conjunction with NF- κ B and ATF3, forms a circuit that discriminates between transient and persistent TLR4 signals⁵.

Yet these sub-networks explain only a tiny fraction of TLR responses. Earlier studies have generated global network models by integrating microarray profiles and promoter scanning^{2,6}, although they lacked experimental validation. More recently, Suzuki *et al.*⁷ investigated an experimentally tractable system—the differentiation of a macrophage-like cell line—and demonstrated the power of coupling computational network analysis with large-scale perturbations.

Amit *et al.*³ have now taken another step in global network analysis of the TLR pathways by applying large-scale perturbation analysis to dendritic cells—innate immune cells that prime adaptive immune responses. Their approach is straightforward: profile the transcriptomes of TLR-activated dendritic cells;

mine the microarray data to identify candidate regulators that act on target genes; systematically perturb each regulator with shRNAs and measure the effects on target gene responses to TLR activation; and construct a regulatory network from the perturbation data.

But if the concept is straightforward, its implementation is not. Knocking down individual genes in primary innate immune cells is notoriously difficult; knocking down >100 genes is a major accomplishment. Profiling the responses of target genes to the perturbations was facilitated by the nCounter system (from Nanostring Technologies), which enables highly parallel high-sensitivity transcript measurements from small amounts of sample.

In addition to finding well-established regulators of TLR responses (such as NF- κ B, IRF, ATF and STAT family members), the authors identify many new factors that warrant further investigation. For example, the prediction that TLR4-induced Cbx4, a SUMO E3 ligase, negatively regulates the induction of IFN β is intriguing. Similarly, the hypothesis that

established regulators of the cell cycle (e.g., E2f5) or of circadian rhythms (e.g., Timeless) may moonlight as regulators of specific gene groups associated with antiviral responses is unexpected and compelling. It is highly unlikely that these novel interactions would have been discovered using conventional approaches.

This groundbreaking study does have a few limitations. First, after examining the effects of 125 regulators on 118 target genes, the authors conclude that nearly one-sixth of all possible interactions are significant, with 80% of the candidate regulators affecting four or more targets. By comparison, an earlier estimate based on forward genetic analysis is that ~50 genes in the entire genome play nonredundant roles in positively regulating TLR responses⁸. That so many of the candidate regulators suggested by Amit *et al.*³ are found to control TLR responsive genes is, in our opinion, unrealistic. Second, Amit *et al.*³ interpret their results in terms of the classical categories of ‘antiviral’ and ‘inflammatory’ responses, which does not take full advantage of a systems approach. The true power of systems biology lies in its ability to identify modules and networks directly from the data, extending beyond known signaling pathways. By casting their network in the canonical binary framework, predicted modules that do not fit neatly into either category are obscured. Finally, many of the network interactions, identified by gene knockdown instead of by direct measurements of transcription factor–promoter binding, are likely to be indirect.

That said, Amit *et al.*³ have produced a herculean study that pushes the limits of the approach. However, the systems biology field as a whole suffers from an inability to directly link inferred gene networks to function. Fortunately, several emerging technologies are likely to fill this gap. The scale and quality of networks will improve dramatically when large-scale perturbations in primary cells³ are coupled with next-generation sequencing^{9,10}. Multiplexed RNA-Seq and ChIP-Seq approaches will allow thousands of transcriptomes to be directly linked to genome-wide binding profiles of thousands of transcription factors. It will be straightforward to discriminate direct from indirect gene regulation and to interpret networks in the context of additional mechanisms of RNA-dependent regulation, including alternative splicing and microRNA binding. Moreover, advances in high-throughput imaging and in microfluidics will allow much more precise definition of phenotype. Eventually, predictive networks linked to function will facilitate

Kim Caesar

the translation of molecular interactions into therapies. In the particular case of dendritic cells, one might easily imagine how networks delineated by studying antigen presentation as a functional output could be reengineered to yield rationally designed vaccines.

- Ishii, K.J., Koyama, S., Nakagawa, A., Coban, C. & Akira S. *Cell Host Microbe* **3**, 352–63 (2008).
- Ramsey, S.A., et al. *PLoS Comput. Biol.* **4**, e1000021

- (2008).
- Amit, I., et al. *Science* **326**, 257–263 (2009).
- Gilchrist, M., et al. *Nature* **441**, 173–178 (2006).
- Litvak, V. et al. *Nat. Immunol.* **10**, 437–443 (2009).
- Nilsson, R. et al. *Genomics* **88**, 133–142 (2006).
- Suzuki, H. et al. *Nat. Genet.* **41**, 553–562 (2009).
- Beutler, B. et al. *Annu. Rev. Immunol.* **24**, 353–389 (2006).
- Shendure, J. & Ji, H. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
- Meyer, M., Stenzel, U. & Hofreiter, M. *Nat. Protoc.* **3**, 267–278 (2008).

(such as 15 versus 12 copies). As a result, the relevance of higher-order CNVs to disease and phenotype generally goes untested in genome-wide association studies.

In principle, experimental approaches that replace a noisy analog hybridization measurement with a digital count should offer more robust measurement of copy numbers for high-copy duplications. Indeed, approaches based on the read depth in next-generation sequencing data can infer CNV with greater quantitative and spatial precision compared with microarrays^{9–12}. However, most sequencing-based analyses to date have considered only the unique portions of the genome.

Alkan *et al.*¹ developed a sequence alignment method that is smarter at handling duplicated genomic segments. They realized that current alignment algorithms are confused by sequences present in multiple copies in the human genome reference sequence. Such ‘multiread’ sequences may align to multiple locations in the reference genome. Many alignment algorithms randomly assign a multiread to only one of its potential homes or give it a low mapping-quality score that causes it to get dropped altogether from downstream analyses. In contrast, the mrFAST (micro-read fast alignment search tool) aligner by Alkan *et al.*¹ reports all mapping locations for multireads (Fig. 1). Its main technical advances include more efficient algorithms as well as data structures that keep track of mutations in multireads. Recording sequence variation makes it possible to distinguish between duplication regions in further analyses.

Mapping duplicated sequences

Derek Y Chiang & Steven A McCarroll

Duplicated genomic regions are accurately resolved using an optimized algorithm for mapping reads from next-generation sequencers.

Copy number variation (CNV), or differences in the number of copies of multi-kilobase segments of the genome, is a major source of genetic diversity, with several variants now conclusively linked to human disease. But existing methods for analyzing CNV tend to go astray at recently duplicated regions of the genome, where much of this variation arises. As described in *Nature Genetics*, Alkan *et al.*¹ have developed a computational method for using next-generation sequencing to measure segmental copy number even in recently duplicated regions. Their approach opens a new avenue for exploring relationships between genetic and phenotypic variation.

CNV comprises a diverse set of variants, some of which are more mysterious than others. Much of the progress of the past four years has focused on ‘simple’ CNVs, in which unique genomic segments are deleted or duplicated. Array-based platforms are now routinely used to analyze such variants in large cohorts, revealing associations in autism², schizophrenia^{3,4}, Crohn’s disease⁵ and body weight regulation⁶.

Other classes of CNV have resisted even a basic level of characterization and have not been tested for their relationships to human phenotypes. Chief among these are higher-order duplications—repeated genomic segments that can vary in number in different

individuals. Such duplication CNVs, many of which are likely to be multiallelic, may be among the least stable regions of our genomes, changing in copy number at rates that exceed mutation rates for other classes of variation.

Although molecular heroics have allowed a few multiallelic CNVs to be accurately typed in large cohorts^{7,8}, geneticists have lacked the technology for characterizing multiallelic CNVs on a genome-wide scale. Array-based approaches cannot resolve higher-order copy-number differences because hybridization measurements saturate at higher copy-number levels and are generally too noisy to resolve subtle copy-number differences

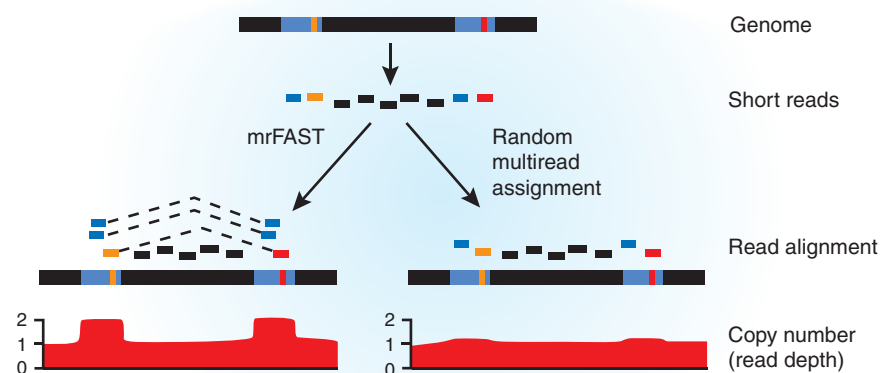


Figure 1 Placement of multireads indicates copy number changes in duplicated genomic regions (blue boxes). Sequence reads that map to multiple locations are connected by dashed lines. mrFAST records all locations for multireads, including perfect matches (blue bars) or reads that differ by a few nucleotides (red or orange bars). By keeping track of all mapping positions, mrFAST provides a more sensitive assessment of read depth (that is, the number of reads mapping to a given region of the genome). Read depth is a measure of copy number. Other aligners may instead randomly assign multireads to a single location and do not keep track of mutations in multireads.

Derek Y. Chiang is at the University of North Carolina, Chapel Hill, North Carolina, USA, and Steven A. McCarroll is at Harvard Medical School, Boston, Massachusetts, USA.
e-mail: chiang@med.unc.edu or mccarroll@genetics.med.harvard.edu