APPLICATION NOTE

Targeted Pathogen Genomics Via NGS Hybridization Capture

MICROBIAL GENOME SEQUENCING CHALLENGES MODERN TECHNOLOGIES

Next-generation sequencing (NGS) is a powerful method for rapid whole-genome sequencing of microbial genomes or overall microbial species content in complex samples. However, most biological samples with microbes of interest are dominated by non-microbial DNA. This necessitates extremely deep sequencing in order to accurately resolve microbial genomes, or to fully characterize the variation within microbial communities. Targeted sequencing, wherein background non-target DNA is excluded from samples prior to sequencing, solves this problem and drastically reduces the overall costs of sequencing and data analysis per sample. Target sequence enrichment typically involves either PCR amplification with site-specific primer pairs or hybridization-based capture with target specific biotinylated probes. Given the high degree of sequence variation within and between viral or bacterial strains, proper primer binding for amplicon formation can be inconsistent. Thus hybridization capture is currently the most versatile technique currently for comprehensive, cost-effective sequencing of both viruses and bacteria in complex samples.

WHAT IS HYBRIDIZATION CAPTURE?

If a microbial species of interest can be grown into pure culture, or otherwise fully physically isolated from host or environmental background cells, then its genomic DNA can typically directly sequenced without any background DNA removal beforehand. However, in many experimental contexts, this approach is either impossible due to biological growth constraints, or infeasible because the research questions require *in vivo* data contexts. In such cases, the historical NGS approaches have been to either perform total sample shotgun sequencing – wherein only a small percentage of reads may come from the microbial genome(s) of interest – or perform targeted amplifications or other approaches to increase the percentage of microbial reads prior to NGS. In both scenarios, hybridization capture offers significant cost and bioinformatic advantages compared to both of these approaches, with minimal expected loss of target sequence diversity.

Hybridization capture works by leveraging the flexible power of complex long oligo pools to target molecules of interest within an NGS library (see Figure 1). In brief, NGS libraries (for DNA- or RNA-seq) are denatured, and allowed to hybridize to target-specific synthetic biotinylated RNA probes. Then the probe:library complexes are bound to streptavidin-coated magnetic beads, and washed to remove non-specifically bound library molecules. These "enriched" libraries are then amplified, and are ready for sequencing on the appropriate NGS platform.

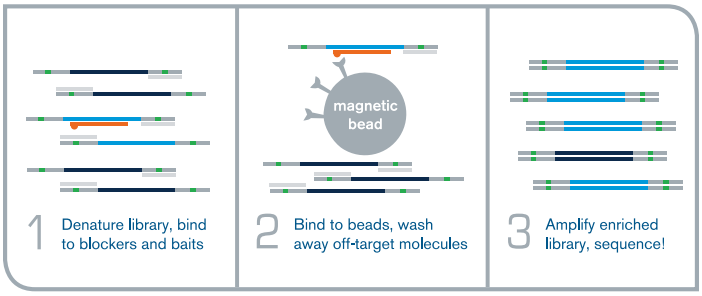


Figure 1. In-solution hybridization capture, or "hyb cap". *Biotinylated oligo probes hybridize to NGS library molecules built from all DNA in a sample, and then the hybrids are sequestered on streptavidin-coated magnetic beads prior to NGS.*

CASE STUDIES

The numerous benefits of hyb capture, especially for sequencing small and diverse genomes, are now routinely exploited for microbial genome sequencing and strain identification.

Viral whole genome sequencing from total tissue DNA

Infectious virus strain identification often benefits from full genome sequencing, especially when that strain is novel or exhibits previously unobserved characteristics. However, the vast majority of DNA in host or cell culture samples are from non-viral sources. In Forth *et al.* 2019, genomes of the large dsDNA virus African swine fever virus (ASFV), up to almost 191 Kbp, were effectively retrieved using hyb capture from both cultured cells and porcine spleen tissue. This increased the percentage of viral DNA by over 100 fold (Figure 2) compared to shotgun sequencing, allowing for the generation of high coverage genomes and various analyses including investigations into viral variants with a fraction of the sequencing effort.

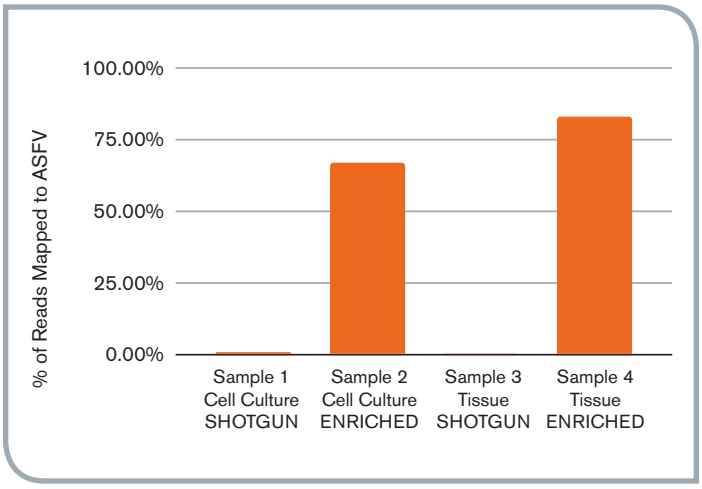
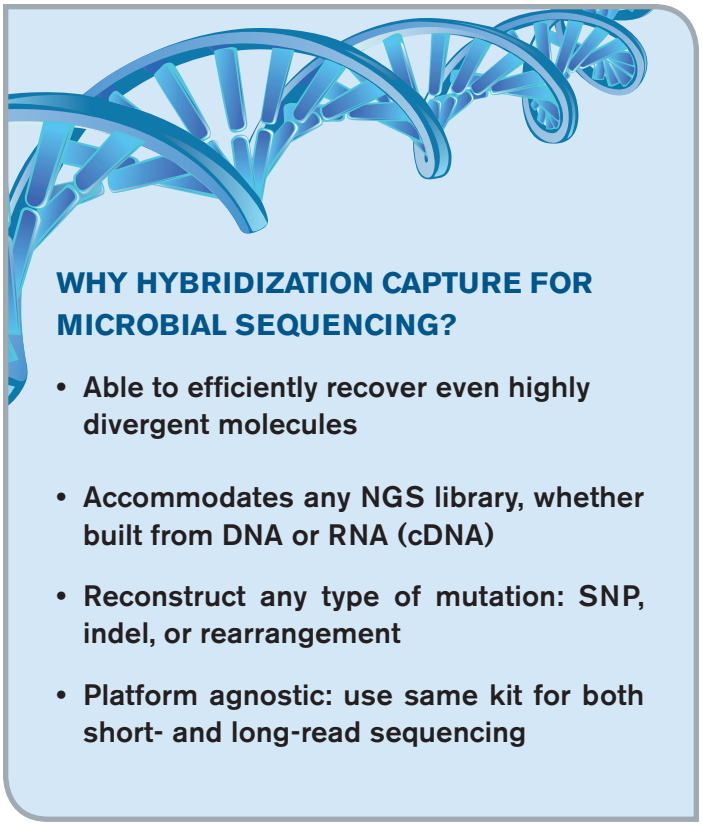


Figure 2. High percentage of ASFV observed in enriched samples (cell culture and tissue). *Based on data presented in Forth* et al. *(2019).*

Pathogen genome sequencing from environmental DNA

Pathogenic bacteria typically exist at low abundance in natural environments, but represent a reservoir from which epidemics can arise. Detecting and characterizing ultra-low-abundance strains can serve as a critical tool in outbreak prevention and treatment. Vezzulli *et al.* (2017) demonstrate that by using hyb capture, extremely low-frequency *Vibrio cholerae* genomes, including specific virulence factors, could be retrieved from raw river water despite being embedded in an extremely diverse background of other DNA sources (Figure 3). Since these bacteria were so low abundance, they would have been otherwise impossible to study with tissue culture or more traditional molecular techniques.



WHY HYBRIDIZATION CAPTURE FOR MICROBIAL SEQUENCING?

- **Able to efficiently recover even highly divergent molecules**
- **Accommodates any NGS library, whether built from DNA or RNA (cDNA)**
- **Reconstruct any type of mutation: SNP, indel, or rearrangement**
- **Platform agnostic: use same kit for both short- and long-read sequencing**

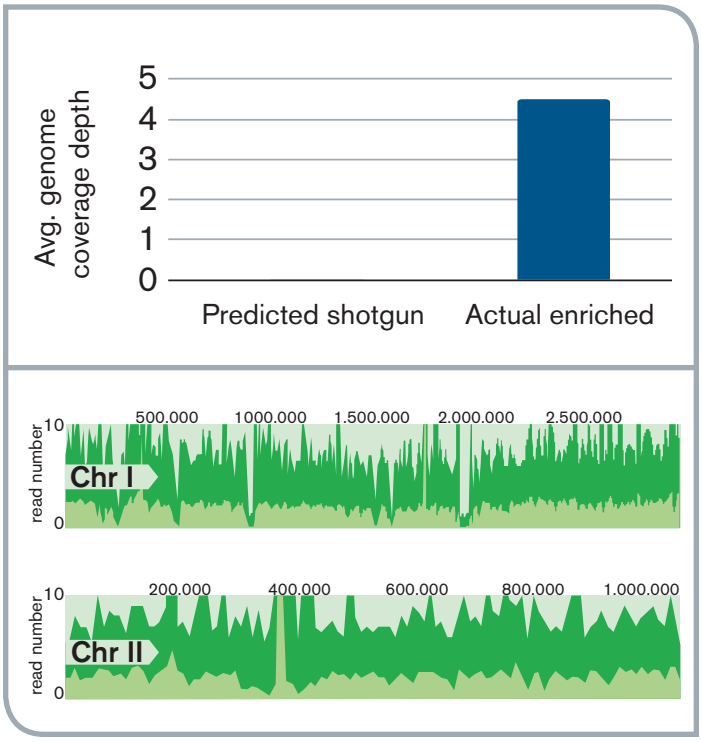


Figure 3. Effective genome-wide enrichment of *V. cholerae* DNA from river water. *Vezzulli* et al. *(2017) estimate that their enrichment approach was ~2500 times more effective than shotgun sequencing alone (0.0018X average genome coverage with shotgun vs 4.5X enriched). Bottom figure redrawn from Vezzulli* et al. *(2017), Fig. 2, pg. 736.*

Ancient viral genomic DNA sequencing

In addition to the ability of hyb capture to enrich extremely low-abundance microbial DNA from complex samples, another key feature is its capacity for capturing nucleic acids that are damaged and degraded, increasing their relative divergence from the probe sequences. For example, in archaeological or paleontological samples, viral DNA is not only swamped by environmental and host DNA, but it is also heavily damaged by taphonomic processes. Duggan *et al.* (2016) were able to retrieve even the most trace, most fragmented remaining genomic fragments from a nearly 400 year-old mummy, and fully reconstruct the genome sequence of smallpox – variola virus (VARV) – an endeavor that would have otherwise cost orders of magnitude more in sequencing alone (Figure 4).

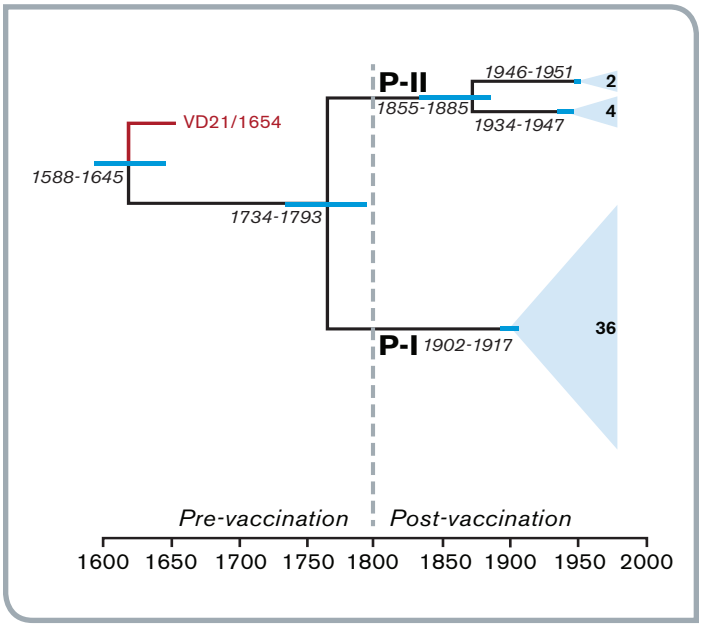


Figure 4. VARV evolutionary timescale tree. *16th century strain in red. Redrawn from Duggan* et al. *2016, Fig 3, pg. 3410.*

Simultaneous ectoparasite host and pathogen sequencing

Novel disease reservoirs and transmission routes are often unknown. While ectoparasites like biting arthropods can be easily identified as a disease's potential vector, identifying the spectrum of hosts of that parasite, including the one that might serve as an interim pathogen reservoir, can be much more difficult. As demonstrated by Campana *et al.* (2016), hyb capture can serve as a quick and inexpensive means of identifying both a parasite's most recent bloodmeal, as well as spectrum of pathogens in that parasite (Figure 5). This highlights hyb capture's versatility in multi-species profiling in complex DNA samples that may be dominated by host DNA.

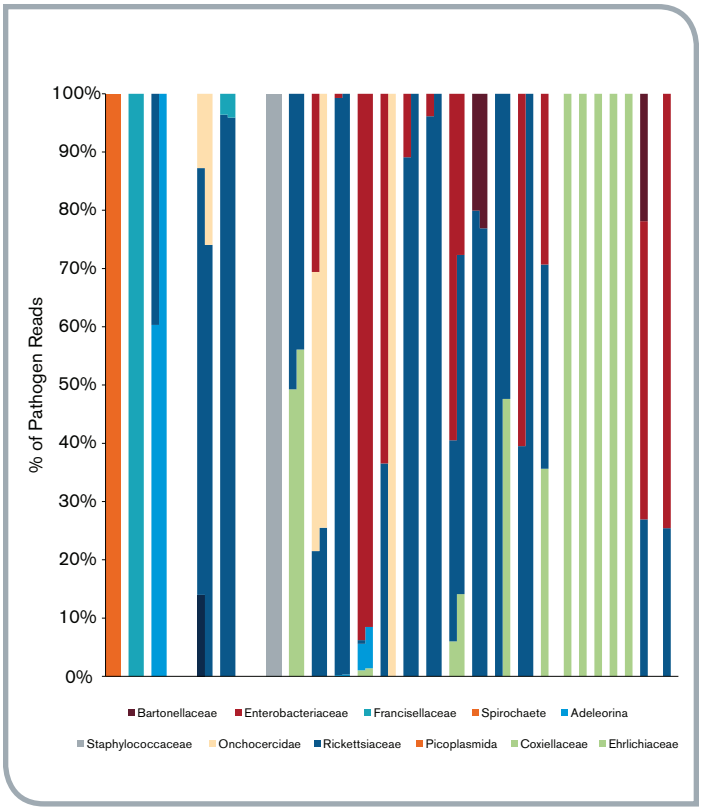


Figure 5. Abundance of co-enriched pathogen and macroparasite taxa identified in tick and flea samples. *Redrawn from Campana* et al. *2016, Fig. 3, pg 11.*

WHY HYBRIDIZATION CAPTURE

The hybridization capture system has many advantages compared to alternative targeted sequencing approaches such as multiplexed amplicons (Figure 6). Hyb capture probes function even when there is significant probe-target sequence mismatch, up to 30+% divergence. Since the hyb capture probes do not themselves become directly integrated into the sequenced library molecules, this offers significant experimental flexibility. Since only one probe minimally must hybridize to enrich a molecule, and no probe-probe interaction is required, this minimizes the need to optimize most experiments. Library molecules of any length, even up to several Kb, can be enriched and sequenced with long-read platforms such as PacBio® and Oxford Nanopore®. These advantages mean that any type of novel sequence feature in or adjacent to the targeted regions – e.g. SNVs, short indels, or rearrangements – can typically be reconstructed in the downstream bioinformatic data analysis.

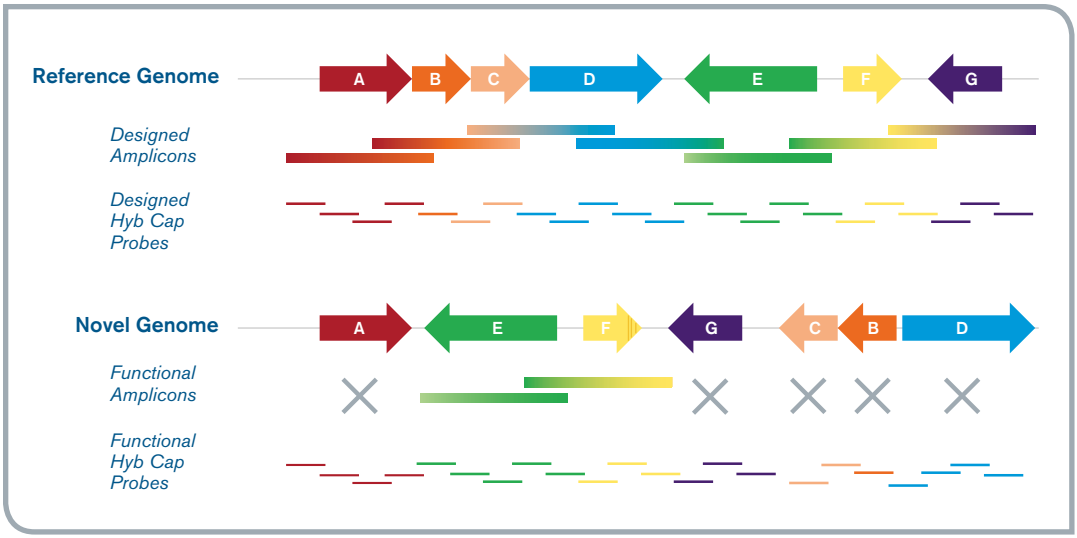


Figure 6. Hybridization capture tolerates both rearrangements and sequence variation. *Both hyb cap and amplicon sequencing can effectively enrich for known sequence regions. However, only hyb cap can retrieve target sequences that have significant rearrangements and/or mutations relative to the reference used for probe design, such as when capturing genomic content from novel viral strain genomes.*

CONCLUSIONS

When embedded in complex host, environmental, or meta-genomic DNA backgrounds, microbial genomes can be prohibitively expensive or even impossible to characterize with direct high-throughput sequencing or traditional molecular techniques. Inexpensive, versatile, and platform-agnostic hybridization capture methods, like those afforded by the myBaits system from Arbor Biosciences, reduce microbial genome sequencing costs by orders of magnitude in most circumstances, and can reliably illuminate novel variations in strain sequence, gene content, and genome structure. The small size of microbial genomes means that their capture probe sets are typically even smaller than a standard mammalian exome or SNP genotyping panel, a feature which confers enhanced specificity and sensitivity in capture experiments. Whether for strain identification, virulence detection, transmission history tracking or illuminating ancient evolutionary origins, hybridization capture offers logistical and budgetary benefits to microbial genome sequencing unmatched by other technologies.

REFERENCES

Campana, M.G. *et al.* (2016) **Simultaneous identification of host, ectoparasite and pathogen DNA via in-solution capture.** *Molecular Ecology Resources.*

Duggan, A.T. *et al.* (2016) **17th Century Variola Virus Reveals the Recent History of Smallpox.** *Current Biology.*

Forth, J.H. *et al.* (2019) **A Deep-Sequencing Workflow for the Fast and Efficient Generation of High-Quality African Swine Fever Virus Whole-Genome Sequences.** *Viruses.*

Vezzulli, L. *et al.* (2017) **Whole-Genome Enrichment Provides Deep Insights into *Vibrio cholerae* Metagenome from an African River.** *Microbial Ecology.*