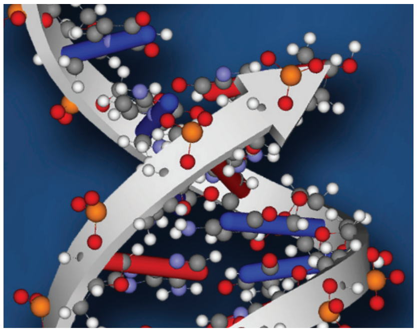
CASE STUDY

BUILDING NOVEL APPLICATIONS WITH PIPELINE PILOT TO DRIVE NEXT GENERATION SEQUENCING



> "In no time at all and without requiring scientists to learn sophisticated analysis software, Pipeline Pilot helps scientists ask relevant, scientific questions about next generation sequencing data."
>
> **Richard Carter**
> **Senior Scientist**
> Oxford Nanopore Technologies

- **Accelrys shares a vision with Oxford Nanopore to enable the NGS market with systems that are easier to use, more scalable, and more versatile than current systems on the market**
- **Pipeline Pilot provides a flexible, adaptable platform that bioinformaticians can use to help bench scientists more effectively analyze NGS datasets**
- **Pipeline Pilot lowers the cost of analysis by providing components out of the box that can be bundled to handle a range of scientific workflows**

Since the completion of the Human Genome Project nearly ten years ago, the science of genomics has undergone a transformation. Enabled by new DNA sequencing technologies, many more scientists and informaticians are interrogating larger and more complex data sets and are expanding the range of problems that can be addressed through sequencing. With the release of the Next Generation Sequencing (NGS) Collection for Pipeline Pilot , Accelrys is enabling scientists and informaticians to build and perform complex analyses of DNA sequence information in dramatically streamlined fashion.

While the NGS Collection supports native data formats from all major sequencing platforms (including those from Illumina, Life Technologies, Roche/454), the system can adapt readily to new sequencing formats and analysis methods as they become available. Accelrys and Oxford Nanopore Technologies have been collaborating both on the development of pipelines that address general computational problems in NGS analysis, regardless of platform, as well as pipelines that support real-time nanopore based experimental analysis for use with the Oxford Nanopore product when it is commercially available. In this case study, collaborators from Oxford Nanopore describe features of the NGS Collection and how it will benefit genome researchers.

Over the past two decades, life science organizations have learned that in order to benefit from high-throughput technologies, they need to ensure that the data produced by these technologies is well managed. Too much data, it turns out, can be the same as no data—without ways to efficiently and intelligently mine large datasets, scientists are unable to effectively locate the actionable results that drive research decisions.

Scientists at Oxford Nanopore understand the challenges associated with making sense of large datasets. The company is developing a platform technology for the electronic analysis of single molecules. With the first application of this technology, they aim to revolutionize what's becoming the most data intensive discipline in the life sciences—next generation sequencing.

The Oxford Nanopore team believes that the technology they are developing lends itself to providing a more versatile and easy to operate end-to-end workflow for scientists wanting to access DNA sequence data. However, simply providing a new instrument would be only half of the story. Oxford Nanopore also aims to provide simpler ways for scientists to interrogate the resulting data. To do this, the organization has been collaborating with Accelrys in the development of the Pipeline Pilot Next Generation Sequencing (NGS) Collection, creating protocols for a variety of sequencing workflows for current and future DNA sequencing technologies such as quality assessment and filtering, visualizing gene content or density, identifying and comparing SNPs and other variants, and comparing RNA expression across experiments or individuals.

"What Pipeline Pilot gives us is a platform," said Richard Carter, senior scientist at Oxford Nanopore. "A bioinformatician in a scientific organization can rapidly create protocols and roll them out to the bench scientists so that they can start asking scientific questions of the data themselves instead of asking the bioinformatician to solve those problems for them. And in many cases, the scientists themselves can use graphical pipelining approach to build their own analyses. It's all about empowering your bench scientists."

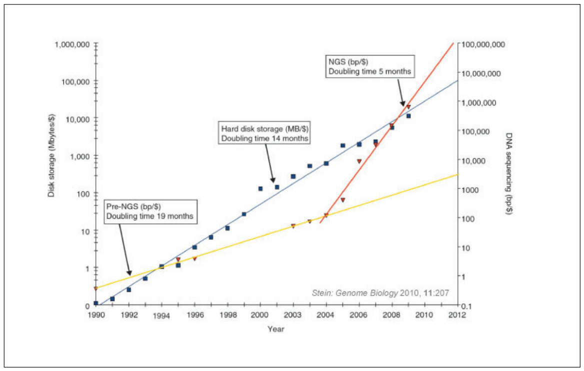


Figure 1: *This comparison of disk storage prices per megabyte versus the number of base pairs of DNA per dollar spent in sequencing illustrates how historic sequencing (yellow) compares to Moore's Law (blue). The advent of next generation sequencing in 2004 (red) decreased doubling time from 19 months for traditional sequencing to five months. (Graphic and data from Stein,* Genome Biology *2010)*

SEQUENCING COMES OF AGE

Next generation sequencing takes high throughput to another level. Ten years ago, the human genome was initially sequenced, capping a 10 year, $2-3 billion investment. Today, research institutions and some service companies are completing primary analysis of a whole human genome in two weeks for as little as $10-20,000. Next generation sequencing systems from major players like Roche, Illumina, and Life Technologies are producing unprecedented volumes of DNA sequence data—several gigabases of data per day in runs that can last a few days to two weeks. A recent commentator noted that with the cost of sequencing dropping by half every five months, next generation sequencing is not just outpacing Moore's Law, but making that law look like it has flatlined (see Figure 1).

By lowering costs and enabling the production of scientifically useful quantities of DNA sequence data in manageable timeframes, next generation sequencing technologies are helping organizations fully realize the promise of genomics and sequencing. Numerous collaborative projects and consortia have emerged to understand the structure of genomes in more detail, catalog genetic variation, and identify the genetic causes of disease across animal, plant, and microbial species. These efforts have the potential to revolutionize how we understand and treat disease. But they also will require new ways to manage and analyze data.

Carter noted that the 1000 Genomes Project, the first large project to capitalize on next generation sequencing technologies, deposited twice as much raw sequencing data into the GenBank archives in its first six months of operation as had been deposited into GenBank in the 30 years since its inception. More critically, he pointed out that industry analysts are now arguing that while sequencing is cheap, analysis remains expensive.

"Analysis can cost up to $100,000 when you factor in the work contributed by all the parties involved in understanding DNA information," Carter explained. Molecular biologists, computational biologists, IT and systems support staff, and geneticists all play critical roles. And clinical staff such as pathologists, genetic counselors, physicians, and research nurses may eventually need logical ways to access data. Today, said Carter, "Pipeline Pilot is well suited to streamline the work of scientific and IT staff in sequencing analysis."

"Pipeline Pilot is one of the few systems I see out there that is starting to address the cost of analysis and the level of expertise needed to perform analyses," Carter continued. "That's what makes it a powerful tool for scientists who need versatile and instant access to complex analyses, whether they have been sequencing for years or just minutes."

HELPING SCIENTISTS INTERROGATE SEQUENCING DATA

Oxford Nanopore has taken on the challenge of sharpening the edge of an already cutting-edge discipline. The company notes that while costs for sequencing have lowered, sequencing methods still rely on labeling (to identify bases) and amplification (to produce enough DNA to sequence), both of which are sophisticated technologies that in their own right require skilled labor and make the overall workflow more complex.

Oxford Nanopore is developing a new platform for sequencing that exploits the signaling properties of nanopores. Nanopores are "very small holes" that can be introduced into a lipid bilayer or even a solid material. The nanopore can be turned into a detection system by passing a current through the hole and then measuring how the current is disrupted by an analyte when it passes through or near the pore.   Publications have shown that nanopores can accurately detect a range of analytes from the four standard DNA bases and modified bases to larger proteins, small molecules, and polymers. The system Oxford Nanopore is developing marries the nanopore with a proprietary sensor array chip that allows hundreds to hundreds of thousands of nanopores to be recorded electronically, in real time, individually but concurrently. The technology should greatly simplify the sequencing workflow by eliminating the need for amplification or labeling.

Oxford Nanopore has been collaborating with Accelrys on the development of the Pipeline Pilot Next Generation Sequencing Collection so that the future users of its system can have both a repository for sequence data (including reference sequences, mapped reads, and sequencing features) and set of reusable components that can be combined to create a library of sample workflows. Carter notes that even the most complicated workflows he's created rely on out-of-the-box components in Pipeline Pilot.

"It's relatively simple even for a novice user of Pipeline Pilot to create useful and powerful applications using the Next Generation Sequencing Collection," Carter said. "In no time at all and without requiring scientists to learn sophisticated analysis software, Pipeline Pilot helps scientists ask relevant, scientific questions about next generation sequencing data."

Carter has created several novel workflows using components from the Pipeline Pilot NGS Collection. All are compatible with existing next generation sequencing systems and also will be compatible with Oxford Nanopore's system when it is commercially available. In addition to using preconfigured analysis workflows, bioinformaticians will be able to build and execute their own analyses on the Oxford Nanopore system. A few of Carter's workflows are described below..

Comparing GC content to determine gaps in coverage

One common criticism of next generation sequencing data is that coverage is not uniform, particularly as GC content changes across a sequence. To help scientists quickly identify poorly covered areas in a sequence run, Carter developed an application that calculates GC content and compares it to depth of coverage. Scientists running the application retrieve a chart that plots sequence coverage against GC content windows for an E. coli genome (see Figure 2, next page). The chart makes it easy for scientists to quickly spot outliers and, with a click, investigate that content in more detail (see Figure 3). "In a matter of seconds, scientists can see which areas weren't mapped well by the experiment and begin looking for explanations for why the data deviates," said Carter.

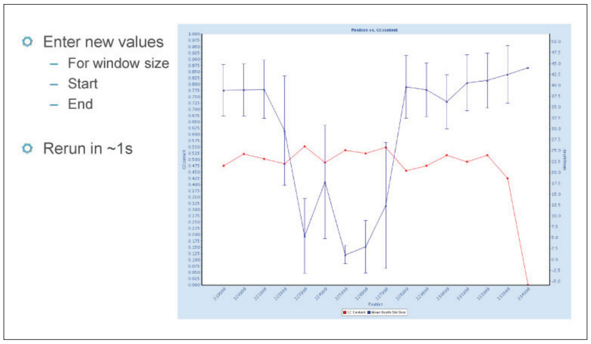
Circos plots

Circos plots are a standard way of visualizing genomic variation. Using "nothing clever in terms of components ," Carter created applications to generate these diagrams for scientists (see Figure 4). The applications enable scientists to quickly pull together data and begin asking questions about it. "The diagram makes it easy to see SNP rich regions of the genome or where the highest gene density is on a chromosome, so that scientists immediately have a place to begin focusing their investigations."

Scientists often use Circos plots in combination with other protocols. For instance, scientists can run an RNA sequencing protocol (the TopHat RNA-sequencing algorithm is integrated with the NGS Collection) to obtain FPKM counts (expected fragments per kilobase of transcript per million fragments sequenced) to measure gene expression and then plot expression information for different tissue types in a Circos plot (see Figure 5, next page). This helps scientists rapidly see where expression levels differ.

Comparison of SNP calling algorithms

Given the speed with which next generation sequencing research is evolving, scientists need ways to rapidly assess the efficacy of new methodologies. Carter developed a simple protocol that displays a Venn diagram illustrating differences in SNPs called by two different algorithms (see Figure 6, next page). SNPs called by both algorithms are shown in a table, as are those called by only one algorithm, and scientists can sort results by position or other variables to facilitate comparison.



- Enter new values
  - For window size
  - Start
  - End
- Rerun in ~1s

Figure 3: *This close up view of one of the regions of poor coverage (around 0.226 M of E. coli using Illumina data from SRA) shows a region that has "normal" GC coverage at about 50% GC content but is covered by few reads in the data set.*

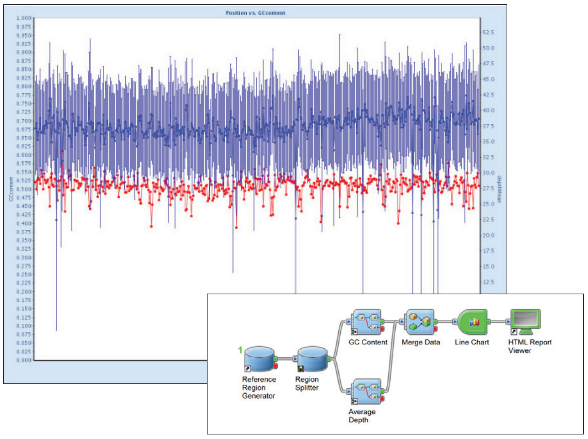


Figure 2: *The mean depth of coverage (blue) and the mean GC content measured over 1000 bp windows for an* E. coli *data set (Illumina data from SRA).* E. coli *does not have many particularly GC or AT rich regions; thus the coverage is relatively evenly distributed across the genome with a few exceptions.*

Carter also pointed out that it's a relatively simple matter to adapt this protocol to compare data across experiments . In this type of analysis, retrieved Venn diagrams will display areas that are shared and unique across the experimental sets. These high level views of the data can help scientists easily compare data obtained from two or more experiments. This facilitates, for instance, discovering positions where a parent's recessive characteristics are expressed in offspring, or identifying SNPs in a tumor sequence data set that do not appear when compared with data from normal tissue.

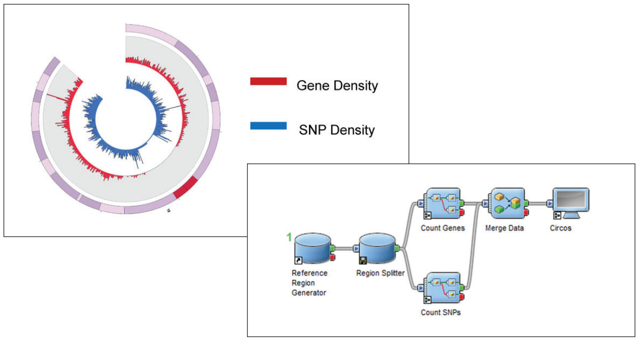


Figure 4: *Circos plot of human chromosome 19 showing SNP density (blue) and gene density (red) for a data set mapped to the human genome (Illumina data from SRA). The outer circle shows the karyotype with the centromere shown as a red band (which is empty of SNPs and genes). From this plot it can easily be seen that there are a large number of SNPs near the centromere and also a spike at about 15 Mb; these would both be interesting regions to investigate further.*

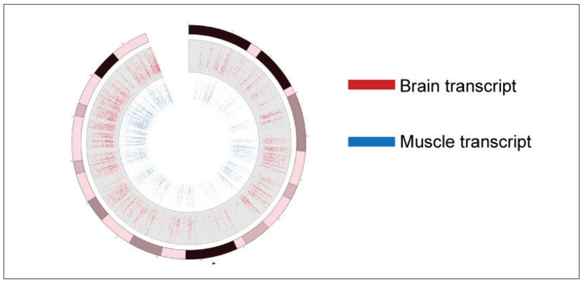
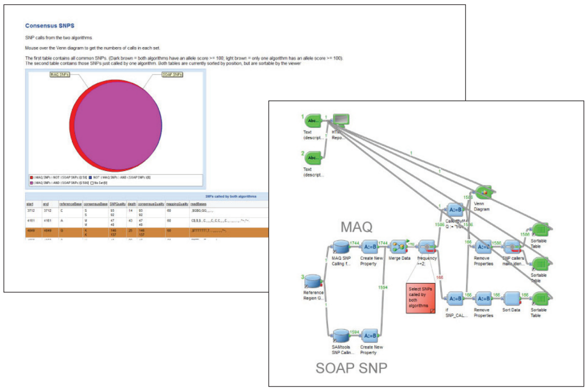


Figure 5: *Circos plot of mouse chromosome 4 using Illumina data from SRA showing transcript data (as represented by FPKM values) for brain (red) and muscle (blue) cells. This is a high level summary view that can be mined further by running the protocol over smaller regions.*

Exon-based questions

Carter noted that even somewhat complex experimental questions can be addressed with only minor tweaks to standard Pipeline Pilot components . The "Getting Started" protocols enable scientists to create an E. coli repository and map and add reads to get basic mapping and coverage statistics for a sequence. By adding just a few lines of code, Carter modified these protocols so that scientists could ask exon-based questions such as "Are all my exons covered to at least a depth of X reads?" (where scientists can specify X; usually 30 reads is more than sufficient). The protocol provides a list of those that are not covered, so that scientists can readily see which exons are missing (see Figure 7).



Consensus SNPS

SNP calls from the two algorithms

Mouse over the Venn diagram to get the numbers of calls in each set.

The first table contains all common SNPs. (Dark brown = both algorithms have an allele score >= 100; light brown = only one algorithm has an allele score >= 100).
The second table contains those SNPs just called by one algorithm. Both tables are currently sorted by position, but are sortable by the viewer

MAQ SNPs
SOAP SNPs
🟥 (MAQ SNPs) NOT (SOAP SNPs)[158] 🟦 NOT (MAQ SNPs) AND (SOAP SNPs)[0]
🟪 (MAQ SNPs) AND (SOAP SNPs)[1586] □ No Set[0]

SNPs called by both algorithms

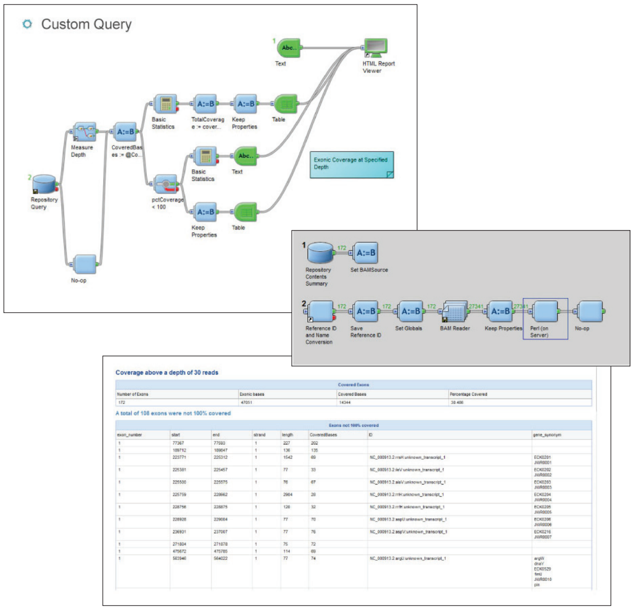
start	end	referenceBase	consensusBase	SNPQuality	depth	consensusQuality	mappingQuality	readBases
3712	3712	C	S S	93 92	14	93 92	60	,SGSG,GG,,,,..
4161	4161	A	M M	47 40	43	47 40	60	[illegible]
4849	4849	G	K K	146 137	25	146 137	60	,$TTTTTTT,T[illegible]

Figure 6: *Venn diagram showing two sets of SNPs generated from the same Illumina data set, one using SOAP SNP and the other using MAQ SNP caller. In addition to producing the annotated Venn diagram, this protocol also produces a sortable table that shows all of the SNPs colored by SNP confidence values, allowing scientists to quickly find the most likely common SNPs.*

PUTTING SCIENTISTS BACK IN CONTROL OF SEQUENCING

Instrumentation-centered, data-intensive processes like next generation sequencing can seem relentless in their ability to crank out data. According to Carter, the analysis and visualization applications built with the Pipeline Pilot NGS Collection bring to life the benefits of genuine real-time data acquisition and analysis that is seen with an electronic, nanopore-based system. On the basis of this technology, a software system can be designed that enables scientists to choose to sequence until the endpoint of an experiment has been reached rather than to sequence until the end of a run. For example, scientists might preset an experimental target such as achieving 30x coverage of an organism, identifying the presence of a particular organism or gene in the sample, or running until adequate data had been collected to ensure an overall target accuracy.

"Our technology is designed to provide real-time data and therefore enable real-time analysis, which puts scientists back in control of sequencing," Carter said. "Without this control, scientists set up a run



Coverage above a depth of 30 reads

Covered Exons			
Number of Exons	Exonic bases	Covered Bases	Percentage Covered
172	47051	14344	30.486

A total of 108 exons were not 100% covered

Exons not 100% covered							
exon_number	start	end	strand	length	CoveredBases	ID	gene_synonym
1	77367	77593	1	227	202		
1	189712	189847	1	136	135		
1	223771	225312	1	1542	69	NC_000913.2:rrsH:unknown_transcript_1	ECK0201 JWR0001
1	225381	225457	1	77	33	NC_000913.2:ileV:unknown_transcript_1	ECK0202 JWR0002
1	225500	225575	1	76	67	NC_000913.2:alaV:unknown_transcript_1	ECK0203 JWR0003
1	225759	228662	1	2904	28	NC_000913.2:rrlH:unknown_transcript_1	ECK0204 JWR0004
1	228756	228875	1	120	32	NC_000913.2:rrfH:unknown_transcript_1	ECK0205 JWR0005
1	228928	229004	1	77	70	NC_000913.2:aspU:unknown_transcript_1	ECK0206 JWR0006
1	236931	237007	1	77	76	NC_000913.2:aspV:unknown_transcript_1	ECK0216 JWR0007
1	271804	271878	1	75	72		
1	475672	475785	1	114	69		
1	563946	564022	1	77	74	NC_000913.2:argU:unknown_transcript_1	argW dnaY ECK0529 fimU JWR0010 pin

Figure 7: *Table showing coverage over exons in the E. coli genome (Illumina data from SRA). The top table summarizes the coverage over all exons, while the bottom table list those exons that are not covered to the desired depth.*

and hope they get what they want, sometimes up to two weeks later. Real-time monitoring and analysis means scientists specify how the experiment is run and what data they wish to collect. And if there's a problem with a run, the system is designed to know this early so that the run can be stopped and restarted."

Carter concluded that the NGS Collection for Pipeline Pilot stands alone in offering users of NGS Technologies a way of managing and analyzing data from current and future DNA sequencing technologies while providing an agile platform to manage the ever-changing landscape of bioinformatics algorithms.

"It's a uniquely versatile and scalable system that also looks to the future, to technologies like nanopore sequencing and to adapt for the dynamic bioinformatics landscape with easy-to -use solutions for a broad range of users."

ABOUT THE NEXT GENERATION SEQUENCING COLLECTION FOR PIPELINE PILOT

The Next Generation Sequencing (NGS) Collection for Pipeline Pilot offers out-of-the-box capabilities for easily analyzing and interpreting the massive datasets generated by DNA and RNA sequencing platforms. The collection is designed to support the automation of routine raw data analyses, including protocols and components which map reads using standard bioinformatics algorithms, perform resequencing or de-novo assembly, and detect variants. The adaptable system configured for standard sequencing platforms—including 454 Life Sciences, Applied Biosystems SOLiD, and Illumina (GAIIx)— enables scientists to assess run quality and coverage, contrast mapping programs, and filter reads on-the-fly, providing a faster path to meaningful scientific inquiry of NGS data..

ABOUT OXFORD NANOPORE TECHNOLOGIES

Oxford Nanopore Technologies Ltd is developing a new generation of technology for direct, electrical detection and analysis of single molecules. The lead application is DNA sequencing, but the platform is also adaptable for protein analysis for diagnostics and drug development, and identification of a range of other molecules for security and defense or environmental monitoring. The Company's GridION technology is modular and highly scalable, driven by electronics rather than optics.

Oxford Nanopore's first generations of DNA sequencing technology, Exonuclease and Strand sequencing, multiplex a protein nanopore with a processive enzyme on a silicon chip. This elegant and scalable system has unique potential to transform the speed and cost of DNA sequencing. Oxford Nanopore also has collaborative projects for developing solid-state nanopores to further improve speed and cost. For more information, visit www.nanoporetech.com.