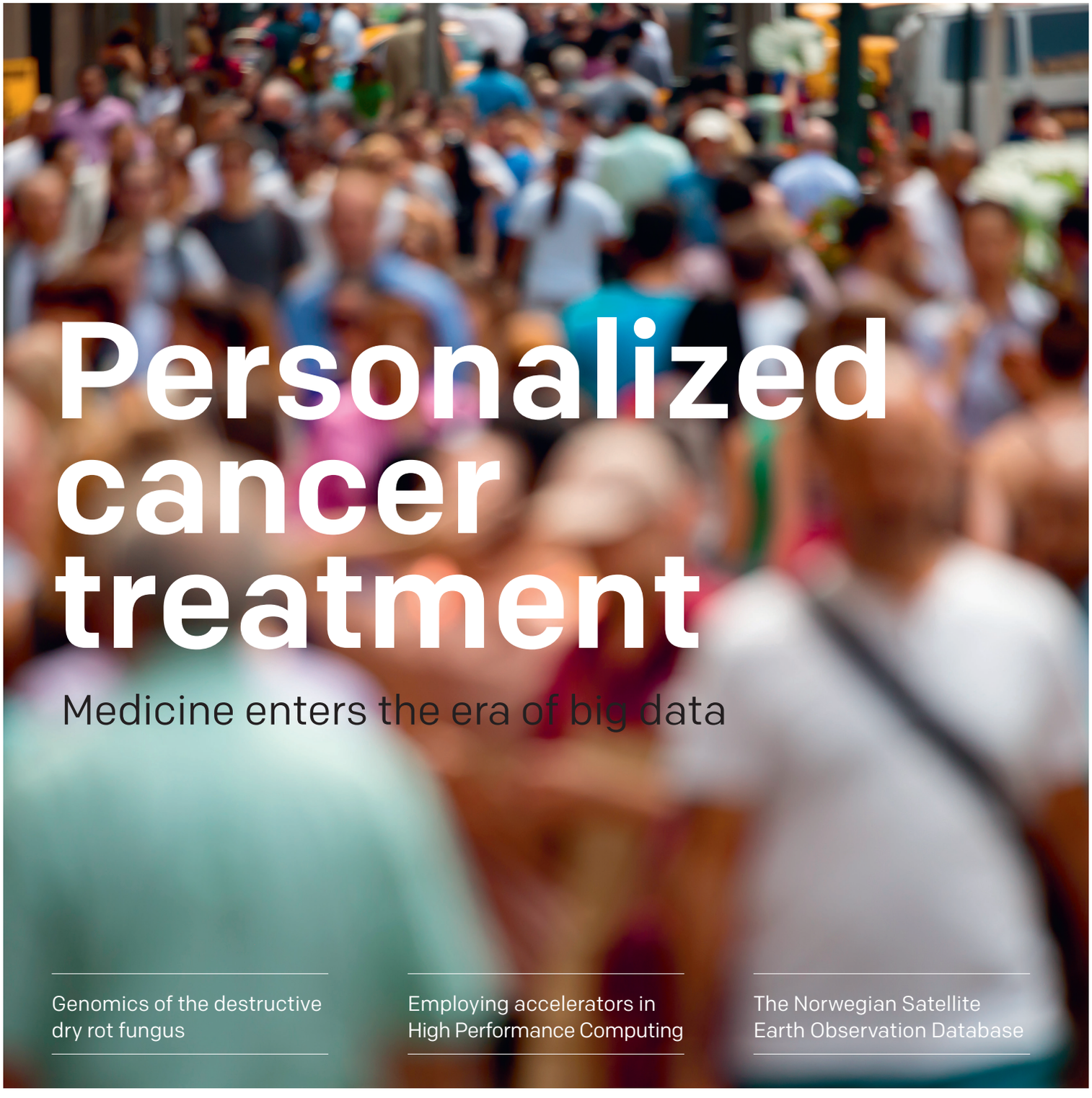
#1 | 2015

A magazine published by UNINETT Sigma2

meta



Personalized cancer treatment

Medicine enters the era of big data

Genomics of the destructive dry rot fungus

Employing accelerators in High Performance Computing

The Norwegian Satellite Earth Observation Database

FOCUS

Tons of data are being generated from genome sequencing projects all over the world, and the opportunities for collecting relevant information and building a solid knowledge base are beyond imagination.

AUTHORS

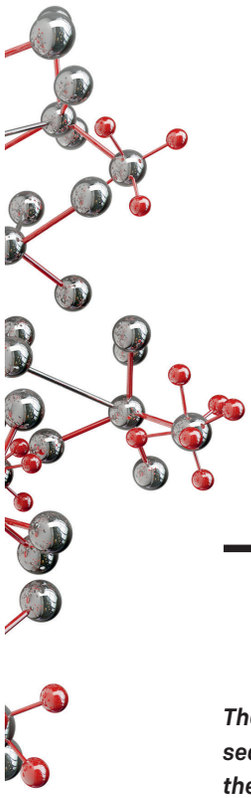
Bjarne Johannessen, Rolf I. Skotheim
Department of Molecular Oncology, Institute for Cancer Research, Oslo University Hospital – Radiumhospitalet and Centre for Cancer Biomedicine and Department of Informatics, University of Oslo

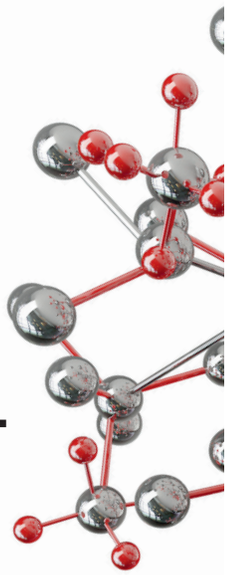
Photos: Shutterstock

From numbers and sequences to

Personalized cancer treatment

Challenges and opportunities as medicine enters the era of big data





The rapid development of technologies for deep sequencing of DNA and RNA that emerged during the past decade has transformed biomedical research and paved the way for unprecedented opportunities in genomic medicine. Now, the complete genetic volume of an individual or an individual's unhealthy cells or tissue, can be recorded in a matter of days.

Every feature that gives rise to some specific traits, and every mutation that can cause disease at some point during the lifespan of a human being. Since the completion of the first draft reference sequence of the human genome at the turn of the century [1, 2], large efforts have been made to identify and catalog the large number of possible genetic alterations that accumulate and may lead to cancer, and to separate potential cancer-causing mutations from natural variation across the population [3]. As medicine now enters a new digital age, scientists from several fields are joining forces, trying to analyze the massive, and often overwhelming, body of information that is currently being revealed. With large progress being made on the technical laboratory side, the challenge is now how to transform all of the numbers and genome sequences into improved and more personalized cancer treatment. Tons of data are being generated from genome sequencing projects all over the world, and the opportunities for collecting relevant information and building a solid knowledge base are beyond imagination. If the information gained from deep sequencing of large patient cohorts can be turned into biomarkers, new drugs, and a more customizable treatment framework, recent genomic advances will ultimately represent a breakthrough in cancer medicine. However, as big data arrives in the clinic, so do a number of new challenges, from many different departments.

TAILORED CANCER TREATMENT

No two cancers are the same, and one major weakness in today's treatment regime is that patients with cancer in the same organ in general are offered the same line of

medication, even though the cancers may be caused by completely different molecular mechanisms. In fact, several different cancer types can initiate in the same organ, so based on the genetic composition of the cancer cells, patients may respond very differently to the same set of drugs. One of the big goals in recording and charting all of the numerous mutations that accumulate in cancer cells, is improved molecular stratification of different patients groups, and thus being able to offer each patient a more precise treatment which is compatible with his or her cancer's particular mutation spectrum.

In the Department of Molecular Oncology at Oslo University Hospital, Radiumhospitalet, tumor biopsies and associated clinical data from patients with colorectal and prostate cancer have for years been systematically collected, awaiting future high-throughput analysis methods. Recently, DNA samples from 166 colorectal cancers and 150 prostate cancers have been subject to full exome sequencing, meaning that all 22,000 protein coding genes have been studied down to their individual nucleotide sequences. This has in part become possible through our involvement in the Norwegian Cancer Genomics Consortium (cancergenomics.no), a national initiative aiming to implement genomics for improved cancer treatment. As of today, only the international pan-cancer effort of The Cancer Genome Atlas (TCGA) have produced a larger set of deep sequencing data from patient material of colorectal and prostate cancers [4]. To increase our understanding of the complex molecular mechanisms underlying each cancer, somatic mutations identified from these projects will be integrated with other levels of data, like changes in DNA copy-number and gene expression.

The latter has been a focus in our research for a while. We have since 2009 implemented both wet-lab and computational protocols for genome-scale analysis of all RNA in cancer samples. Despite large efforts over many years, there is still high demand for better biomarkers in almost all cancer types. One important group of biomarkers can be used to separate tumor cells from normal healthy tissue. In particular, we have focused on identification of variants of RNA molecules which are only produced by cancer cells, and as such can be used for improved methods to detect cancers, or as cancer-specific targets for therapy. One such type of RNA is the fusion transcripts, composed of sequences belonging to two individual genes. We have already used the new genomics tools to reveal cancer-specific features in colorectal [5, 6] and testicular cancer [7], and we have similar unpublished results from prostate cancer.

COMPUTATIONAL INFRASTRUCTURE

Prior to the development of high-throughput technologies, genotyping was a tedious process, and detection of DNA mutations in human cells could only be performed on individual loci, typically in specific mutation hotspots within known cancer-critical genes. In the post-genome era, however, the tables have turned, and what once took years is possible to accomplish in just a few days, and at much higher resolution. The bottleneck now is rather on the processing side, where all findings need to be properly analyzed, understood and put into meaningful biological context.

In this new medical science, computing power and capacity is constantly challenged by increasingly larger data sets. Having sufficient infrastructure in terms of storage capacity and computational power is absolutely essential if we are to utilize the potential that is hidden in the vast amounts of genetic data that is currently being made accessible. In 2012, the Abel computer cluster was opened at the University of Oslo, providing centralized computational services for the Norwegian research community. Abel is a shared resource for high-performance research computing, and is widely used in medical research. We access this resource in our research through current allocations from Notur. So far, the Notur infrastructure has been crucial in for example our fusion gene detection projects where we have included large and publicly available sets of RNA data.

For genome-scale sequencing data from local patient cohorts, a similar secure service aiming to protect sensitive data was established in 2014. Building on experiences from Abel, the new system not only provides storage space and computational resources, but also complies with the national legislations concerning research on sensitive data. With more research groups and laboratories coming into the field, it has been crucial to establish a centralized system for appropriate handling of patient data. These much-awaited Services for Sensitive Data provided by the Center for Information Technology (USIT) at the University of Oslo, has proven to be a decisive part in the cancer research taking place in our department lately.

On the software side, increased informatics competence is needed to establish analysis pipelines for transforming digital signals into meaningful information. Medical research is one of the fastest growing fields of big data science, but data sharing is still suffering from a lack of common scalable frameworks and interfaces. While researchers in many other fields increasingly are turning to cloud computing, the lack of interoperable methods for representation and exchange of genomic data in a distributed computing environment is still a great barrier. Genomic data formats were created prior to the massive high-throughput revolution and need to be developed and modified further along with more comprehensive application programming interfaces (APIs) to facilitate effective and responsible methods for submitting and querying genomic data.

METHODOLOGY

Analysis of large-scale DNA and RNA sequencing data represents a rather new and immature field of research, so best practice workflows and pipelines are still in the process of being established. For DNA sequencing analyses, software tools created at the Broad Institute in Boston have become sort of a gold standard over the past few years. Among the applications they have developed for exome sequencing

analysis is The Genome Analysis Toolkit (GATK) for the first processing of the raw data from the sequencing instrument, MuTect for calling single nucleotide variants, including mutations, Oncotator for annotation, and MutSig for assessing the significance of the detected mutations. The basic foundation of these programs is a three-way comparison where the samples from tumor and normal tissue are compared to the human genome reference sequence. Depending on the type of problem and aim, however, several other methods and pipelines can be applied. For RNA sequencing data, the most common approach has been to follow the Tuxedo protocol, a pipeline specifically designed for transcript analysis, including for instance TopHat for read alignment and splice site discovery, and Cufflinks to test for differentially expressed genes. As sequencing cost has decreased substantially, and more groups are entering the field, several new software packages and algorithmic approaches have been developed. Thus, the set of tools that might be applicable for a particular set of data has increased considerably during the last few years. Nevertheless, creating an optimal algorithmic pipeline that is both sensitive and specific enough for a particular set of data is still a challenge. Also, cancers are very much heterogeneous, and different cancer types are in general driven by different set of mechanisms. Consequently, tools that perform well when applied to one specific type of cancer do not necessarily detect common characteristics shown in other cancer types.

FOCUS

So far, the Notur infrastructure has been crucial in for example our fusion gene detection projects where we have included large and publicly available sets of RNA data.

TECHNOLOGICAL PROGRESS RAISES ETHICAL QUESTIONS

In addition to pure technical issues, progress in big data science entails important challenges in many other areas. Ethical, social, and legal infrastructure needs to be addressed to properly manage the large body of sensitive patient information that is currently under way. Technological progress in genomics has brought about important advances in applications that are far beyond the scope of biology and medicine. However, even though high-throughput genetics has revolutionized industries like forensics and medical science and the application of DNA technology is about to be implemented in an increasing number of fields, we need to give ethical challenges their due consideration, both in research and in the clinic. Privacy legislation needs to be implemented carefully, and as sequencing data is still defined as highly sensitive material, extra care and precaution must be taken upon dealing with such vital information, avoiding data coming astray. As genomics begins to integrate into healthcare, ethical issues like whether incidental findings should be returned to the patient will eventually challenge current procedures. Numerous secrets are hidden in the sequences of DNA, secrets that not everybody might want to know, and least not share. As big data science is about to change the way we see the genomic landscape, the opportunities in implementing routine management of cancer treatment are beyond imagination. However, several challenges remain to be addressed before large-scale genomics data can be applied successfully in the clinic.

REFERENCES

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J *et al.*: Initial sequencing and analysis of the human genome. Nature. 2001, 409(6822):860-921.

2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG *et al.*: The Sequence of the Human Genome. Science. 2001, 291(5507):1304-51. doi:10.1126/science.1058040.

3. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D *et al.*: COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research. 2011, 39(Database issue):D945-D50. doi:10.1093/nar/gkq929.

4. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA *et al.*: The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013, 45(10):1113-20. doi:10.1038/ng.2764.

5. Nome T, Thomassen GOS, Bruun J, Ahlquist T, Bakken AC, Hoff AM *et al.*: Common Fusion Transcripts Identified in Colorectal Cancer Cell Lines by High-Throughput RNA Sequencing. Translational Oncology. 2013, 6(5):546-53.

6. Nome T*, Hoff AM*, Bakken AC, Rognum TO, Nesbakken A, Skotheim RI: High Frequency of Fusion Transcripts Involving TCF7L2 in Colorectal Cancer: Novel Fusion Partner and Splice Variants. PLoS ONE. 2014, 9(3):e91264. doi:10.1371/journal.pone.0091264.

7. Brabrand S*, Johannessen B*, Axcrona U, Kraggerud SM, Berg KG, Bakken AC *et al.*: Exome Sequencing of Bilateral Testicular Germ Cell Tumors Suggests Independent Development Lineages. Neoplasia. 2015, 17(2):167-74.

**Equal contribution*

Photo: Shutterstock