CASE STUDY

THE GERMAN CANCER RESEARCH CENTER

PERFORMANCE WITHOUT COMPROMISE

ACCELERATED BWA-MEM FOR WHOLE GENOME ALIGNMENT AND MAPPING

Cancer is a challenging and complex genetic disease. Occurring when the intricate flow of our DNA is altered and changed to a non-standard sequence, a mutation. Certain types of cancer have more mutations than others which increases the complexity in diagnosis and prescribing treatments.

The German Cancer Research Center (DKFZ) in Heidelberg, Germany is one of the largest data analysis centres in the International Cancer Genome Consortium (ICGC). DKFZ takes on the challenge of sequencing tumour samples to find mutated regions in the patient's cancer genomes which can be targeted more efficiently with available drugs. Those regions are also referred to as druggable lesions. DKFZ provides clinical diagnostics in collaboration with the Heidelberg University Hospital, and participates in numerous cancer related research projects. The ICGC PanCancer Analysis of Whole Genomes is one of the major initiatives DKFZ is active in. The project of the ICGC and The Cancer Genome Atlas is coordinating reanalysis of 3000 whole cancer genomes linked to over 25 cancer types. Together with the Korbel group at EMBL, the research group at DKFZ led by Eils and Brors provides one of the three core unified variant calling pipelines.

BLUEBEE

dkfz.

THE NEED TO SCALE UP

Analysing cancer mutations will help scientists unlock the underlying mechanisms leading to cancer and translate these insights into the clinical practice to help treat the patients. The process of sequencing, mapping, and identifying cancer mutations (fig 1), however, requires very intense computational power.

Thorough examination of the genome of the cancer cells is essential for a better understanding of the disease. DKFZ, with the support of the German Cancer Consortium, invested in an Illumina HiSeq X Ten Sequencing System in order to sequence whole genomes, each with a coverage of ~60X.

> " WITH MORE DATA GENERATED WE CAN DO MORE MEANING-FUL RESEARCH "
>
> **DR. BENEDIKT BRORS, HEAD OF THE DIVISION APPLIED BIOINFORMATICS AT DKFZ**

As a point of reference on the data sizes involved, the HiSeq X Ten is sold as a set of 10 ultra-high throughput sequencing systems, each generating up to 600 gigabases per day, per system, providing a throughput to sequence tens of thousands of high-quality and high-coverage genomes per year. Considering the amount of data generated, data processing required a substantial speed-up. At the same time DKFZ was adamant not to make compromises on the accuracy, sensitivity and specificity of the algorithms in use.

ALIGNMENT, THE COMPUTATIONAL BOTTLENECK

> " DEPENDING ON THE DATA, WE ARE TYPICALLY SEEING BETWEEN 10-20X ACCELERATION FOR BWA-ALN AND 3-4X ACCELERATION FOR BWA-MEM, WHICH IS SUBSTANTIAL, AND OUR ENTIRE BIOINFORMATICS WORKFLOW IS REDUCED BY 50%, WHICH DIRECTLY TRANSLATES IN DOUBLING OUR CAPACITY "
>
> **DR. BARBARA HUTTER, TEAM LEADER CLINICAL BIOINFORMATICS AT DKFZ**

The analysis of a patient's genome involves several steps. One of the steps requiring the most processing time involves mapping of the sequence reads to a reference genome. DKFZ is using gold-standard algorithms BWA-ALN and BWA-MEM and heavy processing computing power to align the reads. The generated data files are subsequently processed by other algorithms also part of the workflow. Finally, the findings are summarised in reports to the treating doctors.

The Burrows-Wheeler Aligner (BWA) was chosen for the alignment step after an evaluation process. Although a number of commercial aligners offered by different parties held the promise of higher speed with comparable accuracy, the gain in speed did not offset against the loss of (backward) compatibility of the results.

With the introduction of the Illumina HiSeq X Ten sequencers, read lengths increased to 150bp, requiring the introduction of a successor to the BWA-ALN algorithm for alignment and mapping. Whilst BWA-ALN is still used at DKFZ in the context of exome sequencing for clinical diagnostics, the newer BWA-MEM has been introduced in the PanCancer project and will be the standard for all future whole genome sequencing projects at DKFZ.

BLUEBEE ACCELERATION OF BWA

BWA was taking up to 80% of CPU time in the data centre's pipeline. It took days to run whole genome sequencing and hours for exome sequencing.

Because BWA was already optimised for standard computer infrastructure, further speed up on the existing cluster was not possible. By transferring the computational-intensive BWA step of the pipeline from the cluster to dedicated FPGA co-processors, a two-fold benefit has been achieved. The alignment time has substantially been shortened, and more CPU power became available on the cluster for the remainder of the pipeline execution.

The acceleration saves DKFZ scientists up to 800 CPU-hours per whole genome pair (tumor and matched normal).

The Bluebee solution has been validated by DKFZ to give identical results compared to the non-accelerated version of the algorithm as published by Heng Li. At the same time, the platform has the capacity to retrieve the results at a fraction of the normal speed. The performance increase without any compromise on the algorithm itself is a justification to embrace the power of HPC. Implementation was seamless, so user adoption rate was easy. The results generated substantial improvement in the run time and processor usage, freeing capacity to run other pipeline components with the available compute power.

" SPEEDING UP THE PROCESSING IS ALWAYS IMPORTANT, AS GETTING RESULTS FASTER MEANS THAT WE GET ANSWERS QUICKER AND WE CAN DO MORE RESEARCH AND PUBLISH MORE RESULTS. BUT SEQUENCING WHOLE GENOMES FOR ACTUAL PATIENTS MAKES ACCELERATING OUR ANALYSIS PIPELINE EVEN MORE IMPORTANT "

WHAT WILL THE FUTURE HOLD

In the future the use of whole genome sequencing is bound to increase, not only in the domain of cancerous diseases but across many hereditary syndromes. The focus on rare diseases is a big trend, and structured data gathering is an essential aspect in this domain.

In this respect the ICGC2 program is meant to go much further than the current ICGC, where the main focus was on cataloguing mutations in cancers. ICGC2 will run over much larger numbers, up to 5000 patients per project, and will include full medical information. This will allow the researchers to test paradigms for treatment, and to follow-up the results over a long period of time.

Sequencing the genome of these individuals would give tremendous insight into genetic determinants of common diseases, given that they may be followed for decades.

" AS A GENERAL TREND I FORESEE MANY MORE INVESTIGATIONS TO BE COUPLED TO CLINICAL TRIALS, THIS WILL BE A BIG BREAK-THROUGH IN PATIENT TREATMENT "

" ALSO, MORE PROJECTS WILL FOCUS ON SAMPLES FROM NORMAL COHORTS (HEALTHY INDIVIDUALS) AND LONG TERM FOLLOW-UP OF THE MEDICAL HISTORY OF PATIENTS OVER THE YEARS "

GERMAN CANCER RESEARCH CENTER

LOCATION: Heidelberg, Germany - www.dkfz.de

RESEARCH STAFF: 2800

Projects and topics:

- International Cancer Genome Consortium, ICGC: cancer genome bioinformatics for 78 different tumor types and subtypes, 5 of which are studied at Heidelberg.
- ICGC PanCancer: reanalysis of whole genome sequencing data of ~3000 tumor–normal pairs from 25+ cancer types.
- Personalized Oncology in collaboration with Heidelberg University Medical Center: Last year, over 1800 patients were part of the program, a number that will increase to 4000 by 2017 when the full capacity of the XTen sequencers is reached.
- Affiliated with the National Center for Tumor Diseases and the German National Consortium for Translational Cancer Research (DKTK).

Data generation:

- Growing to 3TB per day, adding-up to ~1PB per year.

The Bluebee genomics platform supports cross-functional teams of life science researchers and clinicians by effectively centralizing and managing their genomics data processes and storage needs.

Bluebee accelerates genomics insights discovery via the delivery of optimized data analysis pipelines, employing both supercomputing and private cloud technologies. This results in a unique high performance cloud-based genomic analysis platform that enables efficient and affordable processing, and insight generation from ever-increasing genomic data.

THE NETHERLANDS

Laan van Zuid Hoorn 57
2289 DC, Rijswijk
The Netherlands

UNITED STATES

951 Mariners Island Blvd
San Mateo, CA 94404
United States

CONTACT US

US: +1 844 662 3511
ROW: +31 88 2140 200
info@bluebee.com
www.bluebee.com

SOCIAL MEDIA

@BluebeeGenomics
Bluebee

BLUEBEE

20180819