- De novo assembly is a method for constructing genomes from a large number of (short- or long-) DNA fragments, with no a priori knowledge of the correct sequence or order of those fragments.
- The goal of a sequence assembler is to produce long pieces of sequence (contigs) from these reads. The contigs are sometimes then ordered and oriented in relation to one another to form scaffolds.
- There are two types of algorithms that are commonly utilized by these assemblers: greedy, which aim for local optima, and graph method algorithms, which aim for global optima. Different assemblers are tailored for particular needs, such as the assembly of (small) bacterial genomes, (large) eukaryotic genomes, or transcriptomes.
- We propose a new method to correct short reads using de Bruijn graphs and implement it as a tool called Bcool
- tools following the de Bruijn graph (DBG) paradigm generally attempt to filter out erroneous k-mers by considering only k-mers present at least a minimal number of times in the reads to be assembled. Both paradigms may benefit from a preliminary error-correction step

➢ RELATED WORK

- Bolouri H. et al. (2018) The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. Nat. Med., 1, 103–112.
- Breiman L. (2001) Random forests. Mach. Learn., 45, 5–32

- Cibulskis K. et al. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol., 31, 213–219.
- Cortes-Ciriano I. et al. (2017) A molecular portrait of microsatellite instability across multiple cancers. Nat. Commun., 8, doi: 10.1038/ncomms15180.
- Danecek P. et al. (2011) The variant call format and VCFtools. Bioinformatics, 27, 2156–2158.
- DePristo M.A. et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet, 43, 491.
- Gurevich A. et al. (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29, 1072–1075.
- Li H. (2015) BFC: correcting Illumina sequencing errors. Bioinformatics, 31, 2885–2887.

➢ Results

- We present results based on simulated datasets as well as on real ones. Simulations make it possible to precisely evaluate correction metrics (Section 3.2) and to assess their impact on downstream assembly (Section 3.3). Correction evaluation was performed using simulated reads from several reference genomes: Caenorhabditis elegans, the human chromosome 1, and the whole human genome. In contrast, the results presented in Section 3.4 aim to validate our approach using real data. All experiments were performed on a cluster node with a Xeon E5 2.8 GHz 24-core CPU, 256 GB of memory and a mechanical hard drive.

- In what follows, false negatives (FN) stand for non-corrected errors, whereas false positives (FP) are erroneous corrections and true positives (TP) are errors that were correctly corrected. The correction ratio is defined as =(TP+FN)/(FN+FP); it is the ratio of the number of errors prior to correction (TP + FN) versus after correction (FN + FP). The higher the correction ratio, the more efficient the tool.

➢ METHODOLOGY

- In this work, we utilized an optimized version of the white-box, non-linear, ensemble GBM (Friedman, 2001; Schapire, 2003) called XGBoost (Chen and Guestrin, 2016) for building our BCrystal model. Gradient boosting is a machine-learning technique based on a constructive strategy by which the learning procedure will additively fit new models, typically decision trees and repetitively leverage the patterns in residuals to provide a more accurate estimate of the response variable (crystallizable versus non-crystallizable proteins). A brief explanation of GBM is provided in Supplementary Material.
- Tree boosting is a learning technique to improve the classification of weaker classifiers by repeatedly adding new decision trees to the ensembles. XGBoost (Chen and Guestrin, 2016) is a scalable machine learning technique for tree boosting. It was shown in Chen and Guestrin (2016) that its performance is better than other boosting algorithms.