



HAL
open science

Read indexing

Nicolas Philippe, Mikael Salson, Thérèse Commes, Thierry Lecroq, Martine Léonard, Eric Rivals

► **To cite this version:**

Nicolas Philippe, Mikael Salson, Thérèse Commes, Thierry Lecroq, Martine Léonard, et al.. Read indexing. *EMBnet.journal*, 2011, 17 (Supplement B), pp.1. 10.14806/ej.17.B.289 . lirmm-00757983

HAL Id: lirmm-00757983

<https://hal-lirmm.ccsd.cnrs.fr/lirmm-00757983v1>

Submitted on 27 Nov 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Read indexing

Nicolas Philippe, Mikael Salson, Thierry Lecroq, Martine Leonard, Therese Commes, Eric Rivals

Laboratoire d'Informatique, de Robotique et de Microélectronique, UMR 5506 CNRS, équipe MAB, Université de Montpellier II, Montpellier, France,

LITIS, Univ. Rouen, Mont Saint Aignan, France

CRBM, UMR 5237 CNRS, Montpellier, France

<http://www.lirmm.fr/~rivals>

The question of read indexing remains broadly unexplored. However, the increase in sequence throughput urges for new algorithmic solutions to query large read collections efficiently. We propose a solution, named Gk arrays, to index large collections of reads, an algorithm to build the structure, and procedures to query it. Once constructed, the index structure is kept in main memory and is repeatedly accessed to answer various types of queries. We compare our data structure to other possible solutions to investigate its scalability and computational efficiency. Gk arrays are implemented in a general purpose library, which may prove useful for assembly purposes, for evaluating the expression level in RNA-seq, and others high throughput sequencing applications.

References

1. Querying large read collections in main memory: a versatile data structure. N. Philippe, M. Salson, T. Lecroq, M. Leonard, T. Commes and E. Rivals. BMC Bioinformatics, Vol. 12, p. 42, [doi:10.1186/1471-2105-12-242](https://doi.org/10.1186/1471-2105-12-242), 2011.

Relevant Web sites

2. <http://crac.gforge.inria.fr/gkarrays/>
3. <http://www.atgc-montpellier.fr/ngs/>