



# CS636 Data Analytics with R Program

## Project 1

By  
Group 6

Duyen Ngyen | Bicheng Xiao | Shubham Gulia

## Scope of the project

Use R program to extract required information from specified website

Main page:

<https://bmcmmedgenet.biomedcentral.com>

Total articles: 1642

Required information:

10 fields:

DOI, Title, Authors, Author Affiliations, Corresponding Author, Corresponding Author's Email, Publication Date, Abstract, Keywords, Full Text (Textual format)

The screenshot displays the BMC Medical Genetics website. At the top, the BioMed Central logo is on the left, and 'Login' and 'Search B...' are on the right. The main navigation bar includes 'HOME', 'ABOUT', 'ARTICLES' (which is underlined), and 'SUBMISSION GUIDELINES'. On the left side, there are links for 'Sections', 'Supplements', and 'Reviewer acknowledgements'. The central 'Articles' section features a search bar with 'Search by keyword' and 'Search by citation' tabs, a search input field containing 'Search BMC Medical Genetics', a dropdown for 'All volumes', and a 'Search' button. Below the search bar, there is a 'Sort by' dropdown set to 'Newest first' and a 'Page 1 of 66' indicator. The featured research article is titled 'The RS4939827 polymorphism in the SMAD7 GENE and its association with Mediterranean diet in colorectal carcinogenesis'. The abstract states: 'The objective of our investigation is to study the relationship between the rs4939827 SNP in the SMAD7 gene, Mediterranean diet pattern and the risk of colorectal cancer.' The authors listed are Jéssica Alonso-Molero, Carmen González-Donquiles, Camilo Palazuelos, Tania Fernández-Villa, Elena Ramos, Marina Pollán, Nuria Aragonés, Javier Llorca, M. Henar Alonso, Adonina Tardón, Pilar Amiano, José Juan Jiménez Moleon, Rosana Peiró Pérez, Rocío Capelo, Antonio J. Molina, and Inés Gómez Acebo. The publication details are 'BMC Medical Genetics 2017 18:122' and 'Published on: 30 October 2017'. At the bottom of the article, there are links for '> Full Text' and '> PDF'. On the right side of the page, there is a 'Submit a manuscript' button, links for 'Editorial Board', 'Editor Profiles', and 'Sign up to article alerts', and a 'FOLLOW' section with a Twitter icon.

## Contribution of group members

Bicheng Xiao	main program: (loadAllArticles.R) Functions(util.R): <ul style="list-style-type: none"><li>• analysisArticle</li><li>• extractAuthors</li><li>• extractAffiliation</li></ul>	Testing and debugging Project report(PDF file)
Duyen Ngyen	Function(util.R): <ul style="list-style-type: none"><li>• loadArticleList</li></ul>	Testing and debugging Readme.txt
Shubham Gulia	Function(util.R): <ul style="list-style-type: none"><li>• extract</li><li>• extracAttribute</li></ul>	Testing and debugging ReadTheResult.R

## Challenges

unfamiliar with R packages, like XML	Check the online document and samples
extract author, corresponding author, corresponding author's email	<ul style="list-style-type: none"><li>• Carefully check the HTML of the article page and find the xpath</li><li>• Use a 2<sup>nd</sup> time extraction to get more detailed or related data</li></ul>



**Thank you**