

Systematyczne przetwarzanie informacji o rearanżacjach genomowych w R

Piotr Dittwald (MISDoMP/MIMUW)

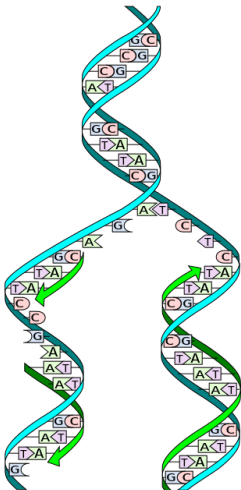
piotr.dittwald@gmail.com

SER IV, Warszawa, 22 V 2014

Czego można się spodziewać?

- ▶ Motywacje biologiczne
- ▶ Analiza bioinformatyczna
- ▶ Kod w R

Ludzki genom



źródło: wikipedia

- ▶ cztery możliwe nukleotydy: adenina (A), cytozyna (C), tymina (T), guanina (G)
- ▶ dwie helisy DNA
- ▶ komplementarne pary nukleotydów: A-T, G-C

Kariotyp

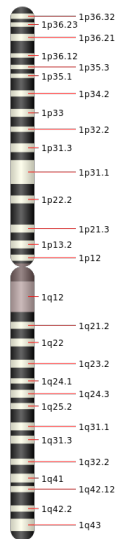
- ▶ chromosomy 1-22
- ▶ chromosomy płciowe
 - ▶ mężczyzna: X-Y
 - ▶ kobieta: X-X



źródło: wikipedia

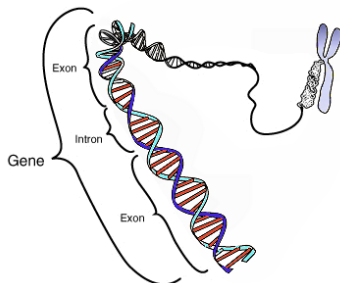
Chromosomy bliżej

- ▶ ramiona p i q
- ▶ centromer
- ▶ prążki widoczne po zabarwieniu barwnikiem



źródło: <http://ghr.nlm.nih.gov/chromosome=1>

Geny



źródło: wikipedia

- ▶ będziemy je traktowali jako przedziały na genomie
- ▶ często są powiązane z chorobami (np. baza danych OMIM.org)

- ▶ repozytorium pakietów do bioinformatyki
- ▶ na stronie dostępna dokumentacja i tzw. *vignette*

The screenshot shows the Bioconductor website in a Mozilla Firefox browser. The page is for the **IRanges** package, version 2.14. The header includes the Bioconductor logo and navigation links: Home, Install, Help, Developers, and About. The main content area is titled "IRanges" and describes it as "Infrastructure for manipulating intervals on sequences". It provides the Bioconductor version (Release (2.14)), a description of the package's functionality, the author (H. Pages, P. Abouyon and M. Lawrence), and the maintainer (Bioconductor Package Maintainer). It also includes instructions on how to install the package using R and how to cite it in a publication. On the right side, there are sections for "Workflows" and "Mailing Lists". The "Workflows" section lists common Bioconductor workflows, including "Common Bioconductor workflows include:" and a list of workflows like "Oligonucleotide Arrays", "High-throughput Sequencing", "Counting Reads for Differential Expression", "Annotation", "Annotating Variants", "Annotating Ranges", "Flow Cytometry and other assays", "Candidate Binding Sites for Known Transcription Factors", and "Cloud-enabled cis-eQTL search and annotation". The "Mailing Lists" section provides a link to the mailing lists and a note to read the posting guide before posting. The "Documentation" section lists links to PDF documents: "R Scripts", "Rie Tips and Tricks", "Reference Manual", and "NEWS". The "Details" section shows the package's version (1.22.6) and its dependencies (DataRepresentation, Infrastructure, Software).

Bioconductor - IRanges - Mozilla Firefox

Firefox - < Inbox - ... Plotr Di... W Cytogen... gene de... Bioc... Table B... Genomi... http...rja barplot... R-cran p... R Q

www.bioconductor.org/packages/release/bioc/html/IRanges.html

IRanges

Search:

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help Developers About

Home > Bioconductor 2.14 > Software Packages > IRanges

IRanges

Infrastructure for manipulating intervals on sequences

Bioconductor version: Release (2.14)

The package provides efficient low-level and highly reusable S4 classes for storing ranges of integers, RLE vectors (Run-Length Encoding), and, more generally, data that can be organized sequentially (formally defined as Vector objects), as well as views on these Vector objects. Efficient list-like classes are also provided for storing big collections of instances of the basic classes. All classes in the package use consistent naming and share the same rich and consistent "Vector API" as much as possible.

Author: H. Pages, P. Abouyon and M. Lawrence

Maintainer: Bioconductor Package Maintainer <maintainer at bioconductor.org>

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("IRanges")
```

To cite this package in a publication, start R and enter:

```
citation("IRanges")
```

Documentation

[PDF](#) [R Scripts](#) An Introduction to IRanges
[PDF](#) [R Scripts](#) Rie Tips and Tricks
[PDF](#) Reference Manual
[Text](#) NEWS

Details

biocViews [DataRepresentation](#), [Infrastructure](#), [Software](#)
Version 1.22.6
In Bioconductor BioC 2.3 (R-2.8)
since

Workflows

Common Bioconductor workflows include:

- [Oligonucleotide Arrays](#)
- [High-throughput Sequencing](#)
- [Counting Reads for Differential Expression](#) ([parathyroidSE vignette](#))
- [Annotation](#)
- [Annotating Variants](#)
- [Annotating Ranges](#)
- [Flow Cytometry](#) and other assays
- [Candidate Binding Sites for Known Transcription Factors](#)
- [Cloud-enabled cis-eQTL search and annotation](#)

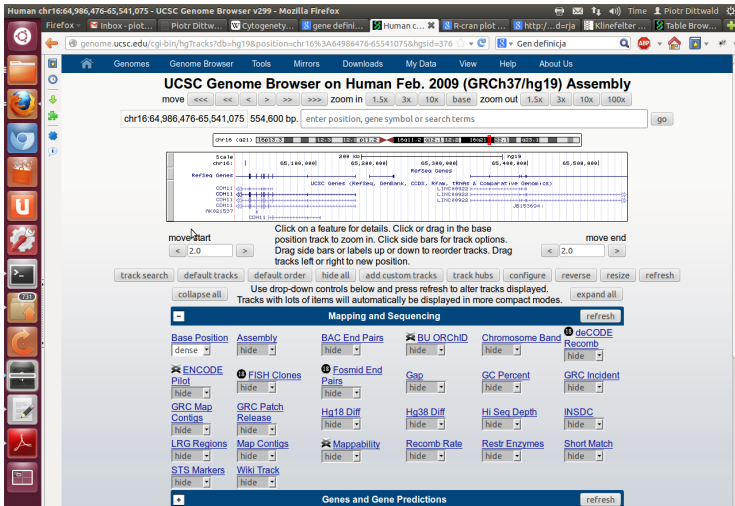
Mailing Lists

Post questions about Bioconductor packages to our mailing lists. Read the [posting guide](#) before posting!

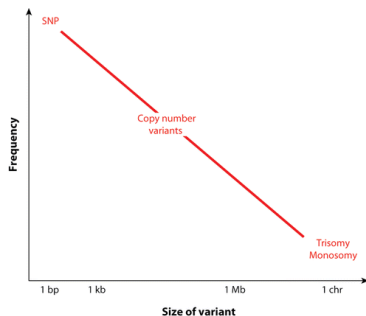
- [bioconductor](#)
- [bioc-devel](#)

Źródło danych - przeglądarka UCSC

- ▶ zbiory danych:
 - ▶ do oglądania w przeglądarce
 - ▶ do ściągnięcia przez Table Browser
- ▶ np. RefSeq, UCSC Genes



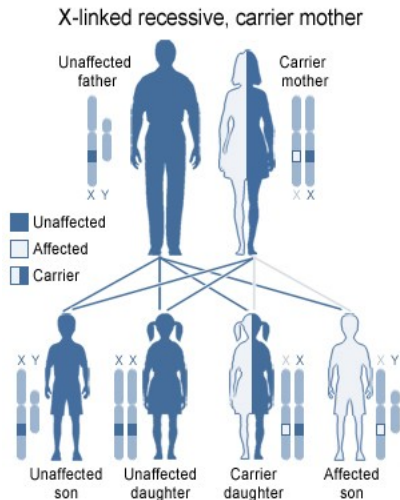
CNVs



Girirajan S, et al. 2011.
Annu. Rev. Genet. 45:203–26

- ▶ obszary genomu, których u danego osobnika jest mniej lub więcej niż w referencyjnym genomie, nazywamy wariantami o zmienionej liczbie kopii (*ang. Copy-Number Variants; CNVs*)
- ▶ w szczególności CNVs występują między długimi (10-400 kb) fragmentami DNA o wysokim (> 97%) współczynniku podobieństwa sekwencyjnego (mechanizm NAHR)

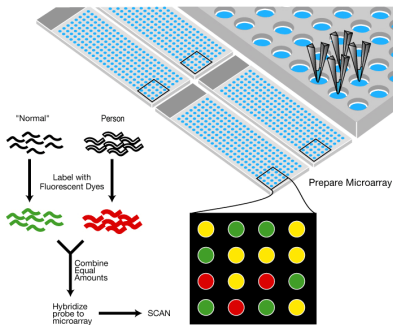
Geny recesywne na chromosomie X



U.S. National Library of Medicine

źródło: wikipedia

Dane kliniczne



- ▶ metoda porównawczej hybrydyzacji genomowej do mikromacierzy (*ang. microarray-based Comparative Genomic Hybridization; aCGH*)
- ▶ dane kliniczne > 25,000 pacjentów z bazy Baylor College of Medicine, Houston