

What is genomic data science?

Steven Salzberg



Intersection of three disciplines
biology
statistics
computer science

Biology

Statistics

Genomic Data
Science

Computer Science

Genomic data science has multiple names

A Venn diagram consisting of three overlapping circles. The left circle is green and labeled 'Biology'. The top-right circle is blue and labeled 'Statistics'. The bottom circle is purple and labeled 'Computer Science'. The central area where all three circles overlap is white and contains the text 'Computational Genomics'.

Biology

Statistics

Computational
Genomics

Computer Science

Biology

Statistics

Bioinformatics

Computer Science

A Venn diagram consisting of three overlapping circles. The left circle is green and labeled 'Biology'. The top-right circle is blue and labeled 'Statistics'. The bottom circle is purple and labeled 'Computer Science'. The central area where all three circles overlap is shaded purple and contains the text 'Statistical Genomics'.

Biology

Statistics

Statistical
Genomics

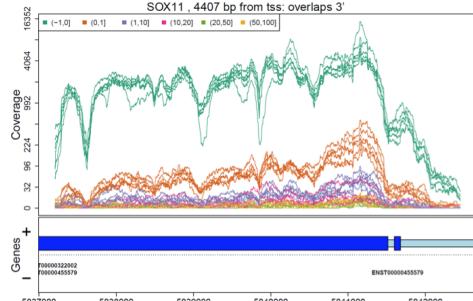
Computer Science

What do genomic data scientists do?

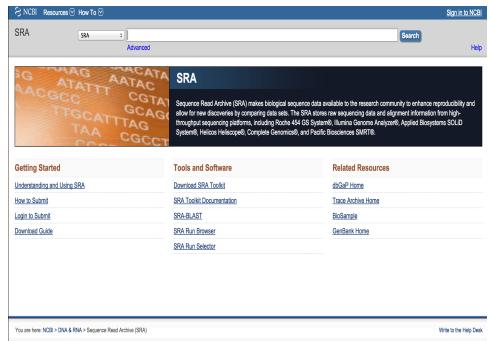
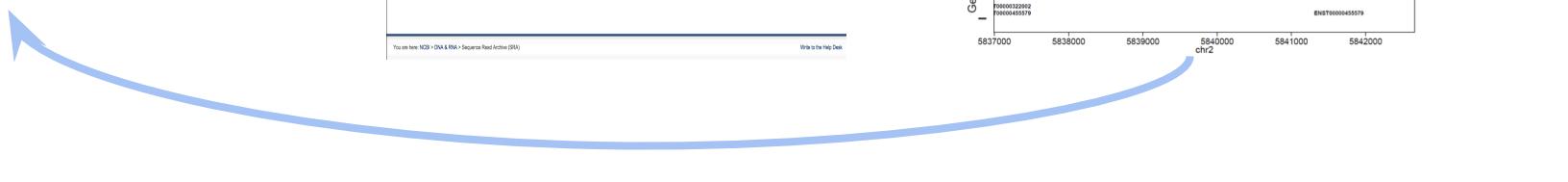


TACAAAATCA GATCGATCG
AGATCAGTC

TACAAAATCAGATCGATCGAGATCAGTC



$$g(Y_{ij}) = \alpha(l_j) + \beta(l_j)X_i + \sum_{k=1}^K \gamma_k(l_j)W_{ik} + \varepsilon_{ij}$$





experimental design

$$g(Y_{ij}) = \alpha(l_j) + \beta(l_j)X_i + \sum_{k=1}^K \gamma_k(l_j)W_{ik} + \varepsilon_{ij}$$



TACAAAATCA ... GATCGATCG
AGATCAGTC

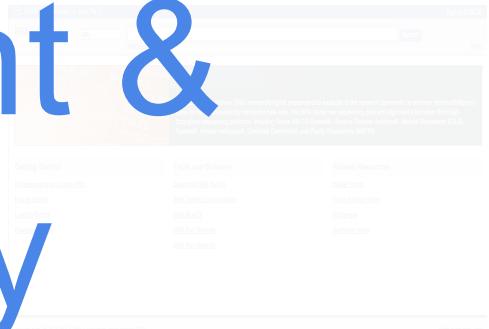
TACAAAATCAGATCGATCGAGATCAGTC



TACAAAATCA GATCGATCG
AGATCAGTC

TACAAAATCAGATCGATCGAGATCAGTC

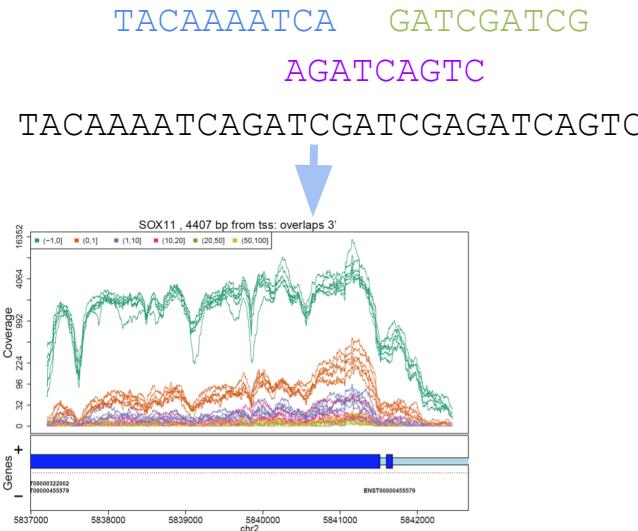
alignment & assembly



preprocessing & normalization

Bioconductor
A SOFTWARE FOR BIOSTATISTICS
AND COMPUTATIONAL BIOLOGY

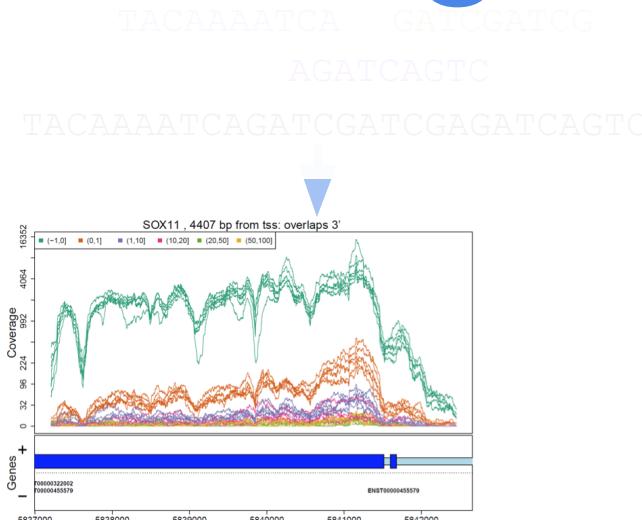
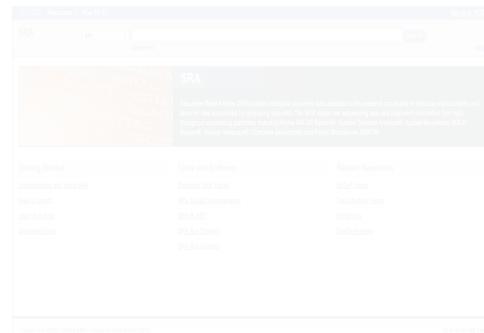
Galaxy



statistics & machine learning



$$g(Y_{ij}) = \alpha(l_j) + \beta(l_j)X_i + \sum_{k=1}^K \gamma_k(l_j)W_{ik} + \varepsilon_{ij}$$



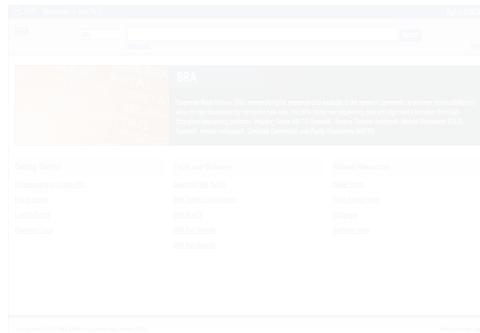
software

development

A



$$g(Y_{ij}) = \alpha(l_j) + \beta(l_j)X_i + \sum_{k=1}^K \gamma_k(l_j)W_{ik} + \varepsilon_{ij}$$



TACAAAATCA GATCGATCG
AGATCAGTC

TACAAAATCAGATCGATCGAGATCAGTC





T

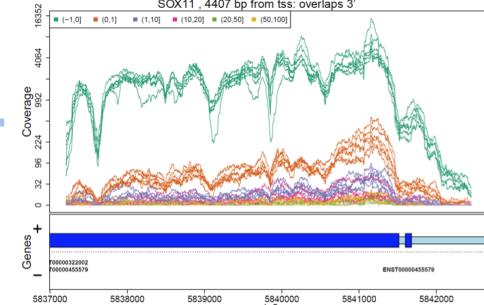
population genomics



$$g(Y_{ij}) = \alpha(l_j) + \beta(l_j)X_i + \sum_{k=1}^K \gamma_k(l_j)W_{ik} + \varepsilon_{ij}$$



TACAAAATCA GATCGATCG
AGATCAGTC
TACAAAATCAGATCGATCGAGATCAGTC





integrative genomics



$$g(Y_{ij}) = \alpha(l_j) + \beta(l_j)X_i + \sum_{k=1}^K \gamma_k(l_j)W_{ik} + \varepsilon_{ij}$$

