BIOINFORMATICS III

„Structural Bioinformatics and Genome Analysis“

Dipl.-Ing. Noura Chelbat
Biologist: Molecular Biologist
Phone: +43-732-2468-8898
Room: T732
Consulting hours: e-mail/phone
chelbat@bioinf.jku.at

BIOINFORMATICS III
„Structural Bioinformatics and Genome Analysis“

Times/locations: room T 212, 9:15-12:45

March	Wed. 3 4U
April	Wed. 14
	Wed. 21
May	Wed. 5
	Wed. 12
June	Wed. 2
	Wed. 9

Total: 28U

Week Mon.14 to Fr.18 Exam

Week 21-25 Special Topics in Computer Science:
Computational Lab on Microarrays Data Analysis
Jose L. Mosquera UB-PRBB

BIOINFORMATICS III
„Structural Bioinformatics and Genome Analysis“

Special Topics in Computer Science: Computational Lab on Microarrays Data Analysis (1PR)

Dipl-Ing Luis Mosquera Mayo

Lab on gene expression experiment using microarrays

Data analysis techniques as preprocessing, filtering, linear models, clustering methods and annotation tools to study the biological significance

Exercises and practice on real problems

R statistical environment with BioConductor packages (linked to Hochreiter lecture on introduction to R)

Prof. Dipl-Ing Sepp Hochreiter

Introduction to R with applications to bioinformatics

Mon 13:45-15:15

BIOINFORMATICS III
„Structural Bioinformatics and Genome Analysis“

Practical course in Protein folding prediction

Dipl-Ing Christoph Etzlstorfer

Exercises in Computational Chemistry are part of the Organisches Chemisches Praktikum 2

- Types of methods like force field and semiempirical
- Overview on programs and hardware used
- Tutorial and example

Work group of 4-5 students given a small molecule and look for the most stable conformation using PC Model, Hyperchem, Mopac, Tinker (Modeller)

From this SS10 ab initio calculations included

Presentation of their results on a poster

Brief Remind

- Part of curriculum of the master of sciences in Bioinformatics
- Included in the Compulsory modules
- Combined Courses (KV) with mainly theoretical part
- Background : Bridge modules from M1-M5

  - M1 Basics of molecular biology
  - M2 Basics of biochemistry
  - M3 Basics of algorithms and data structure
  - M4 Basics of information systems
  - M5 Basics of mathematics

DNA, RNA, Transcription, Translation, Genetic Code, Promoter, Protein folding, Gene regulation Purification, Molecular forces, Secondary / Tertiary /quaternary structure, Folding, Molecular dynamics, instrumental analytics

Bioinformatics III: Bibliography

Molecular and Cell Biology

- Lodish, Berk, Matsudaira, Kaiser, Krieger, Scott, Zipursky § Darnell - Molecular Cell Biology. Fifth edition. W.H. Freeman and Company, New York, USA, 2004.
- Alberts, Johnson, Lewis, Raff, Roberts, Walter –Molecular Biology f the Cell. Fourth edition. GS Garland Science, Taylor and Francis Group, New York, USA, 2002.
- Mathew, Van Holde and Ahern –Biochemistry. Third edition. Benjamin/ Cummings an imprint of Addison Wesley Longman, 1301 Sansome street, San Francisco, CA 94111

General Bioinformatics

- David W. Mount. Bioinformatics – Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2004
- C.A.Orengo, D.T.Jones & J.M.Thornton - Bioinformatics, Genes, Proteins & Computers. Taylor and Francis Group
- Dan E.Krane and Michael L.Raymer-Fundamental concepts of Bioinformatics. Benjaming Cummings
- Arthur M.Lesk -Introduction to Bioinformatics- Second Edition. Oxford
- T.K Attwood & D.J Parry-Smith –Introduction to Bioinformatics-Prentice Hall

Bioinformatics III: Bibliography

General Bioinformatics

- Bioinformatics and Functional Genomics. Langauer
- Bioinformatics: Managing Scientific Data. Lacroix
- Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Baxevanis
- Introduction to Bioinformatics Algorithms. Jones
- Bioinformatics in geneticists. Barnes
- Introduction to computational Biology. Waterman
- Discovering Genomics, Proteomics and Bioinformatics. Campbell
- Bioinformatics for Dummies. Claverie

Bioinformatics III: Bibliography

Structural Bioinformatics

- Philip E. Bourne and Helge Weissig. Structural Bioinformatics. Wiley- Liss, Hoboken, New Jersey, USA, 2003
- Michael J. E. Sternberg. Protein Structure Prediction. Oxford University Press, 1996
- Arthur M.Lesk. Introduction to protein Architecture. Oxford University Press 2003
- Richard A. Friesner. Computational Methods for Protein Folding. Advances in Chemical Physics Volume 120. A John Wiley & Sons, INC.Publication. 2002
- Introduction to Protein Structure. Branden
- Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis. Wit
- Protein Structure and Function. Petsko
- Papers: Special topics in Bioinformatics

Bioinformatics III: Bibliography

Genome Analysis

- Steen Knudsen. Guide to Analysis of DNA Microarray Data. John Wiley & Sohns, Hoboken, New Jersey, USA, 2004.
- Ernst Wit and John McClure. Statistics for Microarrays. John Wiley & Sohns Ltd., England, 2004.
- Pierre Baldi and G. Wesley Hatfield. DNA Microarrays and Gene Expression From Experiments to Data Analysis and Modeling. Cambridge University Press, United Kingdom, 2002.
- Geoffry J. McLachlan, Kim-Anh Do, and Christophe Ambroise. Analyzing Microarray Gene Expression Data. John Wiley & Sohns Inc., Hoboken, New Jersey, USA, 2004.
- Jerome K. Percus. Mathematics of Genome Analysis. Cambridge University Press, United Kingdom, 2002
- Statistical Analysis of Gene Expression. Speed
- Papers: Special topics in Bioinformatics

Bioinformatics III: Changes from previous years

- Chapter 2: First half removed
- Chapter 3: VAST and COMPARER removed
- Chapter 4: Re-written
- Chapter 5: New Threading releases
- Chapter 6: Moleculat dynamics to be removed
- Chapter 7: Included within the chapter 8
- Chapter 8: Remove 8.3.3, new techniques to be included Chip-Chip, Chip-Seq and NGS
- Chapter 9: To be kept and included in chapter 8

Bioinformatics III: Main overview

1. Structural bioinformatics: Chapters 1-5
2. Genome analysis: Chapters 6-8

Goals:

- Main methods in structural bioinformatics and gene analysis: from where we get them and how to use them
- How to choose the proper method from a given pool of approaches
- Adaptation of standard algorithms to the final purpose: combining the information of certain algorithms and biology to build up practical solutions
- How can we use this information to perform searches for the optimal 3D prediction, motifs, expression profiles, pattern regulation ..
- Exercises: SSEs, SCOP classes recognition, DEGs, CNVs, arrays, expression patterns…

Structural Bioinformatics

Motivation:

From Genome sequencing to amino acids/nucleotides primary structure.
From amino acids/nucleotides primary structure to 3D Structure Prediction.

PDB data base

2008 49192 Structures

Feb 24, 2009 _ 56066 Structures

Tuesday Feb 23, 2010 63559 Structures

http://www.pdb.org/pdb/home/home.do

Structural Bioinformatics

UniProtKB/Swiss-Prot

Feb-2008 356 194 sequence entries

10-Feb-2009 Release 56.8 410 518 sequence entries

02-Mar-2010 Release 57.15 515203 sequence entries

http://www.expasy.ch/sprot/

Ratio of 1 structure to 7 sequences

Increasing number of methods to predict 3D structures beside sequencing ones

New approaches based on Machine learning, SVM, NNs, Dynamnic programming and Distance matrixes.

Part I: Structural Bioinformatics

1D	2D	3D
Linear arrangement of amino acids: chain assembled on the ribosome using the codon sequence on mRNA as a template	Secondary structures elements: core elements for protein architecture α Helix β Sheet Loops Coil coiled Turns	Functional activity: Folding and Post-translational modifications Interactions among amino acids side groups Chaperones

Part I: Structural Bioinformatics

Molecular representation and viewers

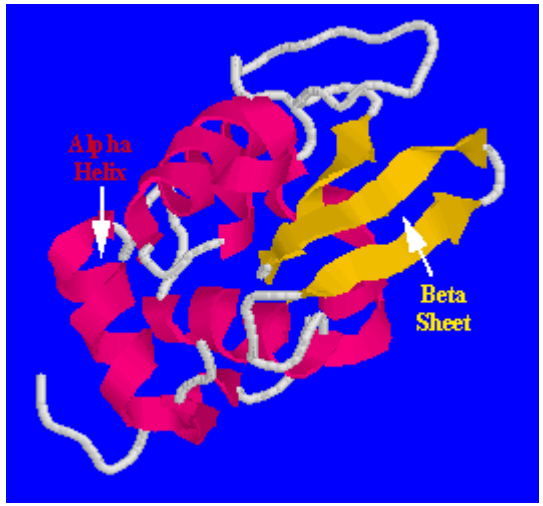
- Difficulties in transforming all of the important 3D structural information about a molecule into an understandable two-dimensional representation
- A variety of molecular representation formats have been developed each of one is designed to show a particular aspect of a molecule's structure
- To visualize the three-dimensional structure of the molecule and understand the relationship between the structural features and its function
- RasMol, Pymol, Chime,.etc

Part I: Structural Bioinformatics

Goals at the end of this part:

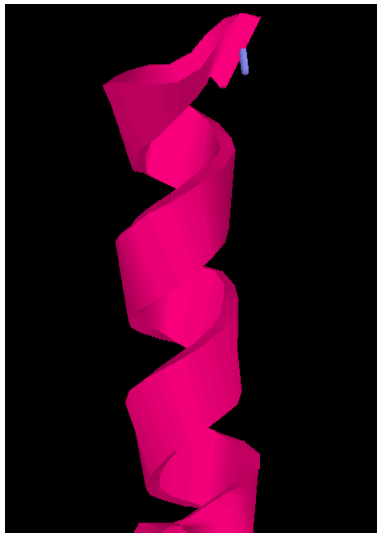
- Recognition of the main types of 2D configurations a helix, b strands, loops, turns
- Recognition of motifs
- Coil coiled, Zn Fingers, Leucine Zippers…
- Structural comparison and Alignment Methods, Protein Secondary structure prediction
- Molecular Dynamics
- Threading methods

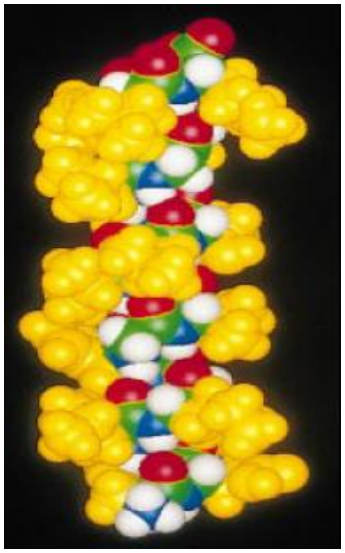
Each picture tells us something different about the structure of the molecule



Lysozyme

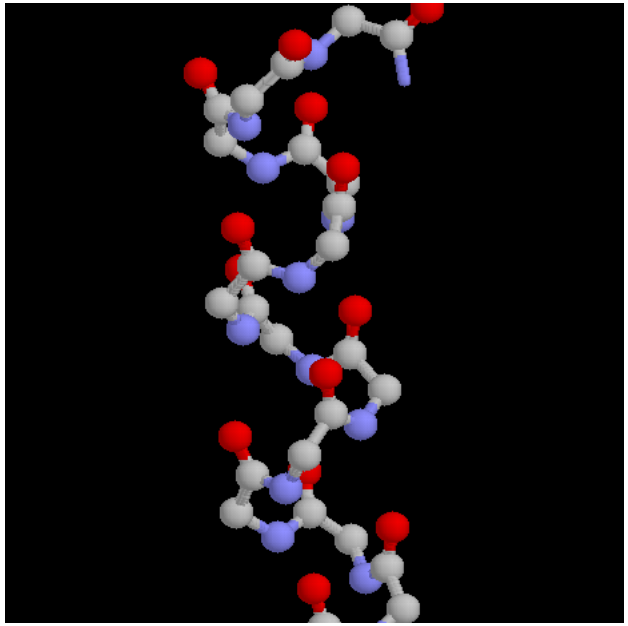
- To catch the main SSEs on a subunit
- To see the relative sizes of the atoms in an a helix by balls representation

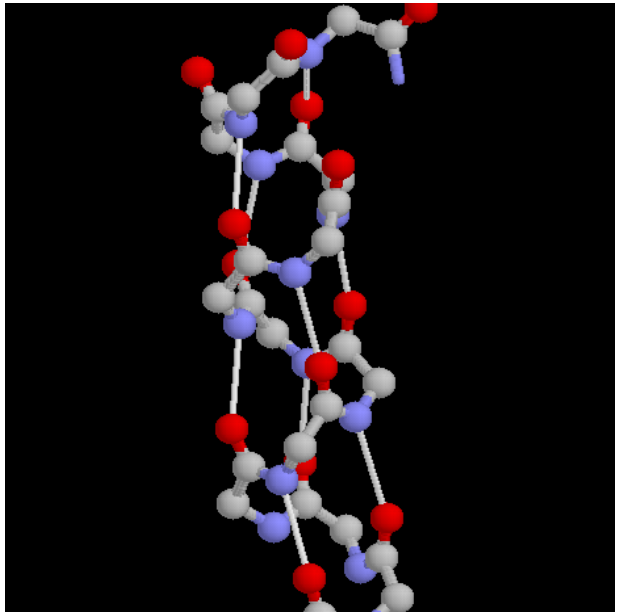




http://project.bio.iastate.edu/Courses/BIOL202/Proteins/secondary_structure.htm

αHelix **Ball** and **Stick View of Lysozyme**





Carbon: Grey
Oxygen: Red
Hydrogen: White
Nitrogen : Blue

To know how the atoms in an α helix are connected to one another by sticks representation
Hydrogen bonds location

http://project.bio.iastate.edu/Courses/BIOL202/Proteins/secondary_structure.htm

http://www.umass.edu/microbio/chime/top5.htm

Part I: Structural Bioinformatics

For similarity and 3D structure detection

Methods from Bioinformatics I allow for homology and comparative modelling where it is assumed that similar sequences have the same 3D structure

Troubles

Different sequences from different proteins can fold into similar three-dimensional configurations

i. No more use of PAM or BLOSSUM matrixes to predict 3D structure on the basis of amino acids substitution because of their standardization

ii. No more use of methods in which both the core regions and loops are equally represented

iii. Gaps should be confined to regions not in the core when multiple alignment are used

Part I: Structural Bioinformatics

Four steps can be addressed when attempting to get information about an unknown protein structure

- 1st **Structure alignment**: based on 3D known structures to **find equivalent amino acids residues**
- 2nd **Structure comparison**: based on shared similarities of two or more proteins when **comparing their 3D known structures**
- 3rd **Structure superposition**: based on preliminary knowledge of positive match of some residue in proteins 1 and 2. The alignment is assumed and the main goal is to search for the **best solution to find what amino acids are equivalents to each other**
- 4th **Structure classification**: based on structural alignment beside other methods to **hierarchically assign classes of proteins**

What could be used??

- Comparative Modeling: Sequence to sequence, Sequence to structure (Psi-Blast, SVM, Fisher Kernels..)
- Scoring matrices
- Distance matrices
- HMMs
- Monte Carlo Optimization and Dynamic programming

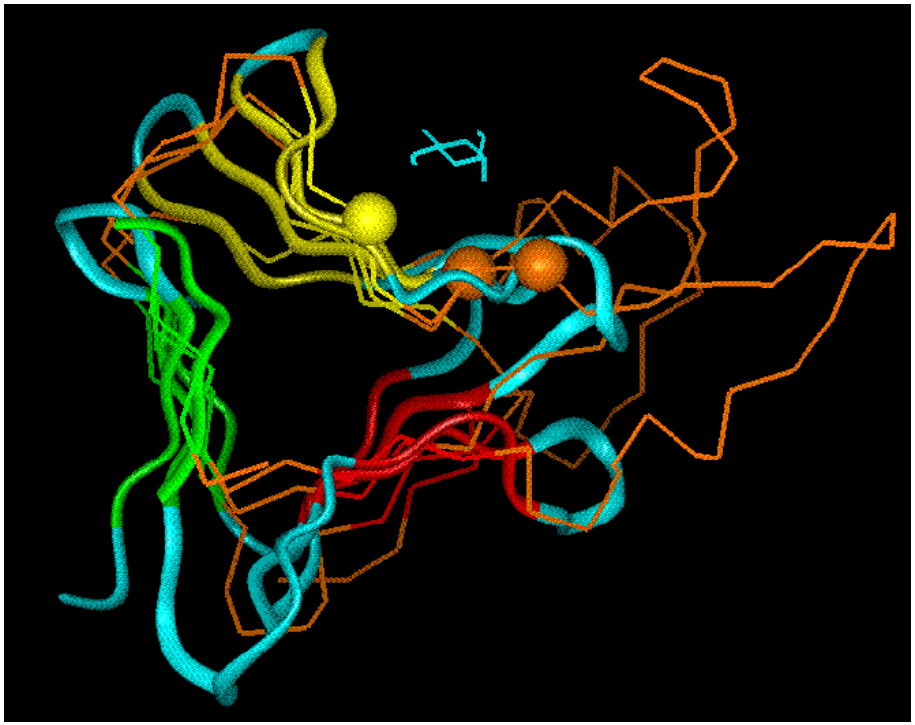
Solutions

Direct link between sequence and structure. In all a sequence representation of a known 3D structure is compared with any other sequences up to match the structure predicted by the model

Accuracy of methods to predict α helix, β strands, coiled coil, turns and loops has an overage of 64-75 % being the highest accuracy for α helix

Part I: Structural Bioinformatics

Methods like CE, DALI, SSAP, and SARF2



Spatial Arrangement of Backbone Fragments

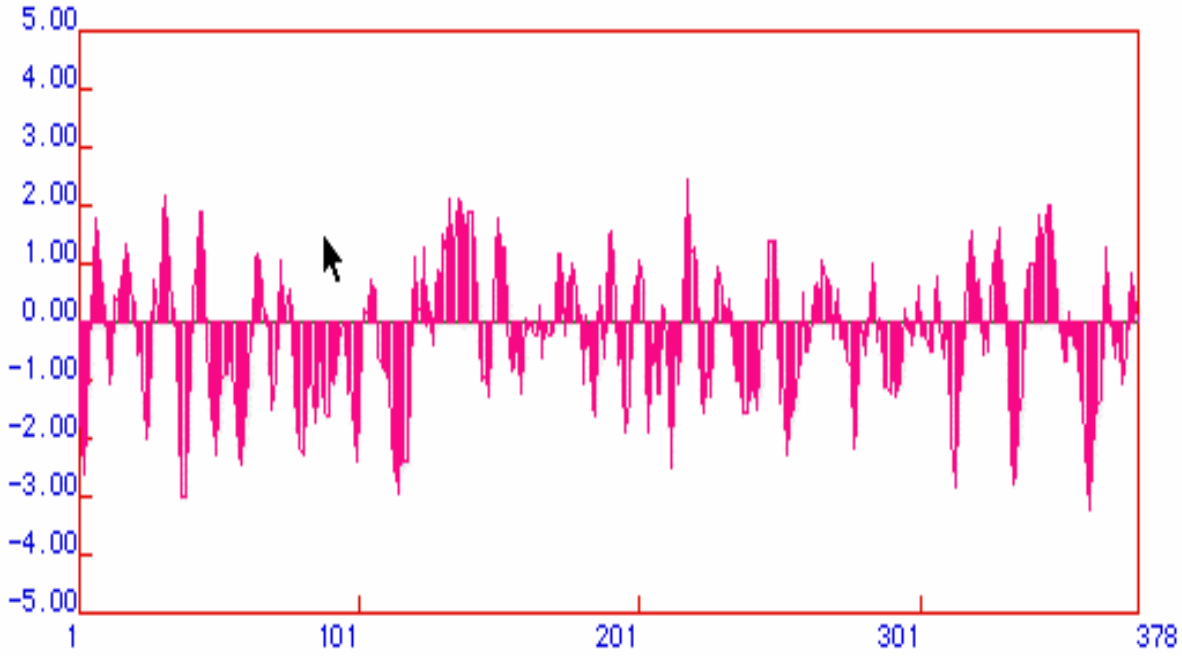
Method based in the comparison of the Cα of each residue in the Secondary Structure Elements (SSEs)

The procedure is design to find out these SSEs which could form similar spatial arrangements but with different topological connections

Manose represented by the SARF2 software. Pectate, lyase and agglutinin

http://123d.ncifcrf.gov/sarfex.html

Part I: Structural Bioinformatics



Hydrophobicity plot for the human actin in which peaks above 2.00 Suggest hydrophobic chains

Pattern of hydrophobicity as approxximation to predict transmembrane α helix of proteins

Part I: Structural Bioinformatics

Protein 2D structure

- GOR
- Chou-Fasman
- Lim's
- Neural Network
- SVMs approximations

The ability also depends on predicting types of SSEs and defining classes of protein structures and patterns

- PHD (Profile Network from Heidelberg) for α helices
- DSSP (Dictionary of Secondary Structure of Proteins)
- STRIDE (STRuctural IDEntification)

Part I: Structural Bioinformatics

	SEQUENCE ALIGNMENT	STRUCTURAL COMPARISIONS
HOW TO	Sequences of proteins written one above the other so the similar amino acids are placed in the same columns and gaps are included	Proteins domains are superimposed fitting together the atoms as closely as possible so that the average deviation between them is the minimum
EVOLUTIONARY SIGNIFICANCE	Sequence similarity = evolutionary relationship	When structural similarity is common evolutionary relationship and convergence phenomena. When no common similarities then divergence phenomena but possible temporary folds

3D homology structure

There are available more than 515203 known protein sequences but just 63559 known structures

New sequence has an homolog with about the same structure

No homologues do exist and new structures also must be predicted

- If two proteins share significant sequence similarity they should have also similar 3D structure
- When the global alignment is performed and the identity shared between the proteins is 25-45 % then the two structures are likely to be similar
- When approximately 45% , then the amino acids could be superimposed in the 3D structure

Some methods like

- SVMs (when remote homology search)
- PSI-BLAST (Position specific iterative BLAST)
- FPS (Family Pairwise Search)

Threading

How well a sequence fits to a given 3D structure

Sequence comparisons can be made on structural level by computing the sequences-to-structure-fitness

1. The target sequence is threaded through the backbone structures of a collection of template proteins
2. Fold library or dictionary of resolved structures for sequence–to -structure alignment
3. "Godness of fit" score calculated in terms of empirical energy function based on statistics derived from known protein structures

Share some of the characteristics of both comparative modelling methods (the sequence alignment aspect) and *ab initio* prediction methods

Part I: Structural Bioinformatics

Ab initio: Insights into protein folding and stability

Ab initio:

Method using only the amino acid sequence to find the 3D structure

Applicable to proteins with novel structure so that threading methods would fail

Rosetta: as the most important ab initio method

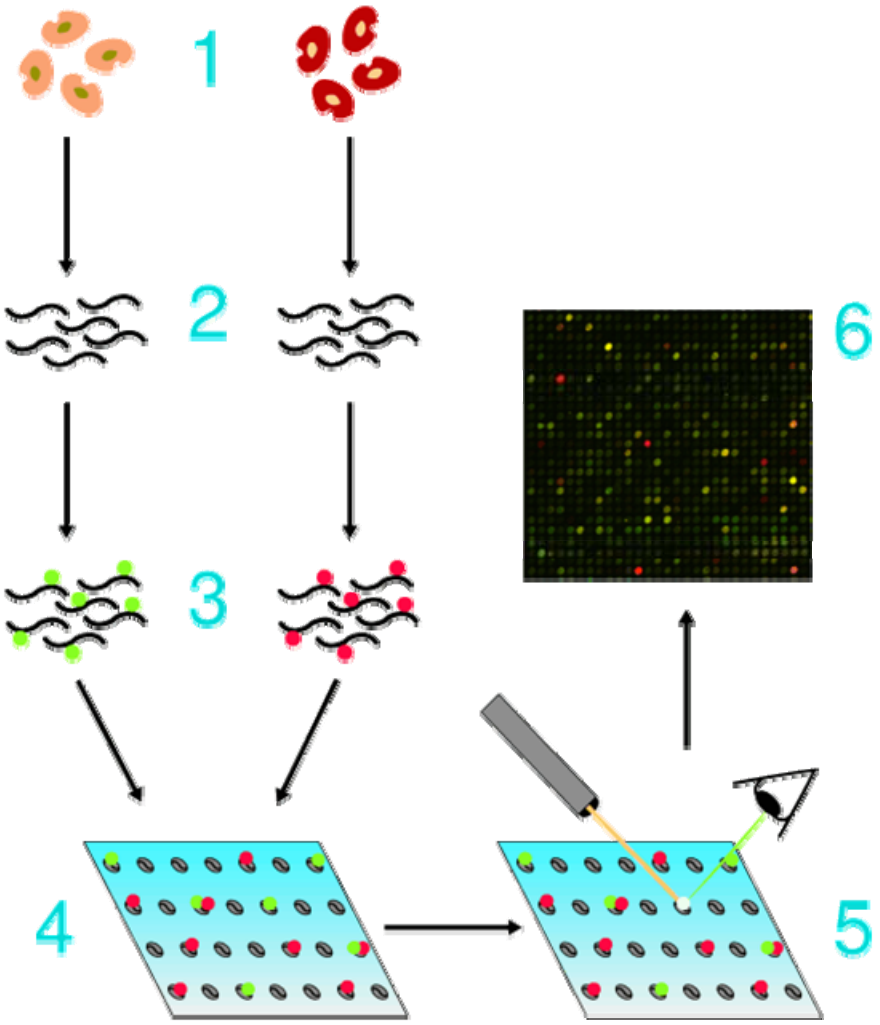
Protein function details and docking behavior are often analyzed based on force fields

Genome Analysis

Motivation

- Major source of information about the processes performed within a cell and evolved to one of the major topics in Bioinformatics
- Provide means of measuring tens of thousands of genes simultaneously by measure at once cellular concentrations of thousands of mRNA: gene expression profile
- Detection of genes that are differentially expressed (DEGs) in tissue samples
- Basis for the functional genome analysis, molecular diagnostics, systems biology
- Important applications in pharmaceutical and clinical research
- NGS as a tool for Genome assembly and genome mapping

Part II: Genome Analysis



Red/Green technology

mRNA concentration ~ activity of a gene

Activity of a gene = expression level

The proportionality between the measured intensities and the number of copies of mRNA in the cell can vary in different arrays

Part II: Genome Analysis

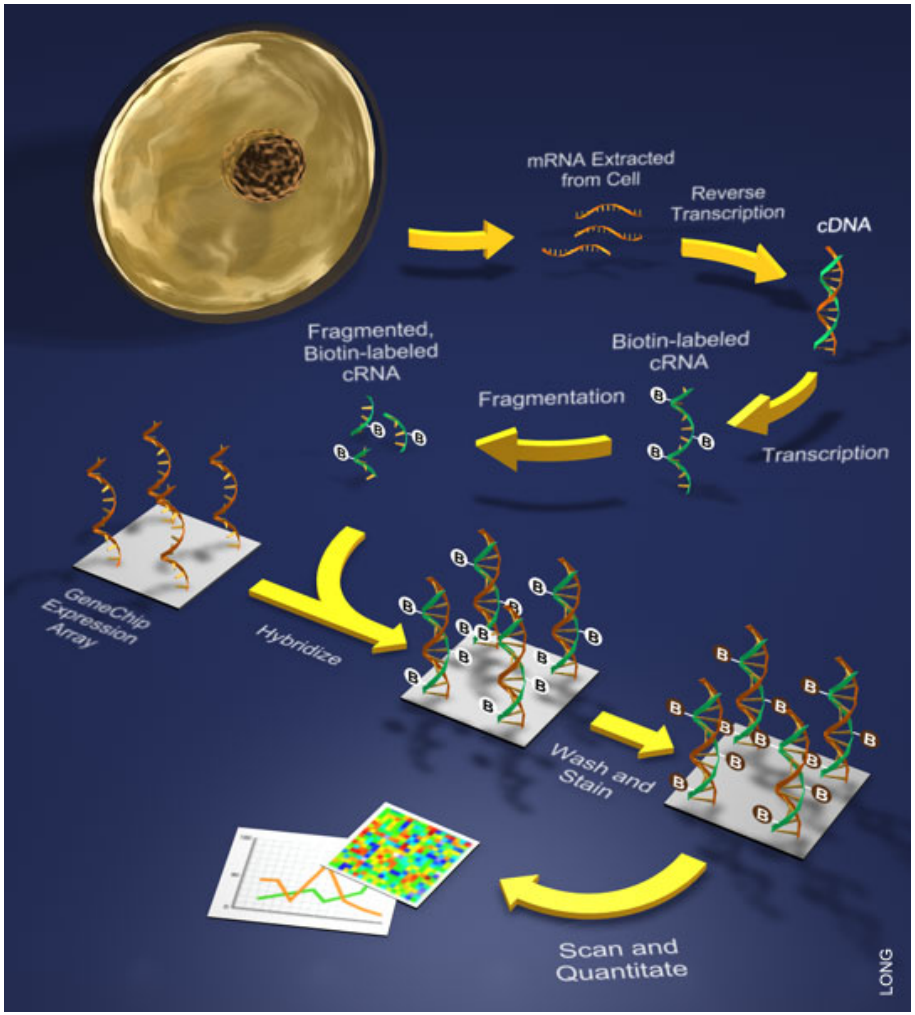
1.DNA Microarray

Techniques and Image analysis
Background correction
Normalilzation
PM correction
Summarization
ML applications (Gene selection, clustering,...)

2. DNA analysis

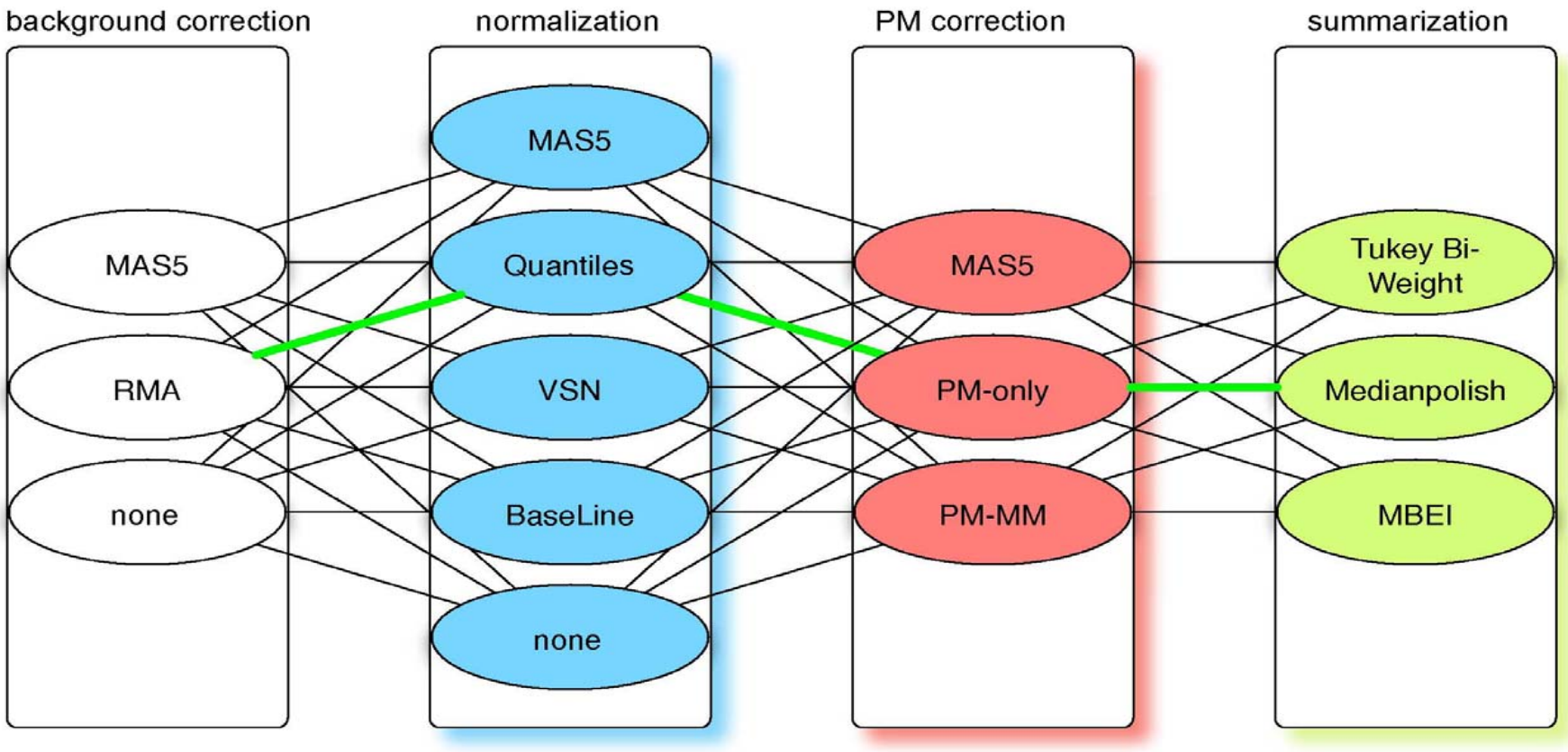
Genome anatomy
Genome individuality
SNPs

3. Alternative splicing
4. Modelling



Part II: Genome Analysis

DIfferent combinations for Microarray preprocessing steps



Part II: Genome Analysis

5. **Next generation sequencing techniques**:

Research community of genomics and transcriptomics as an alternative to array based methods: Illumina's Solexa, Roche's 454, or Applied Biosystems' SOLiD

massive parallel sequencing = high-throughput sequencing = next-generation sequencing

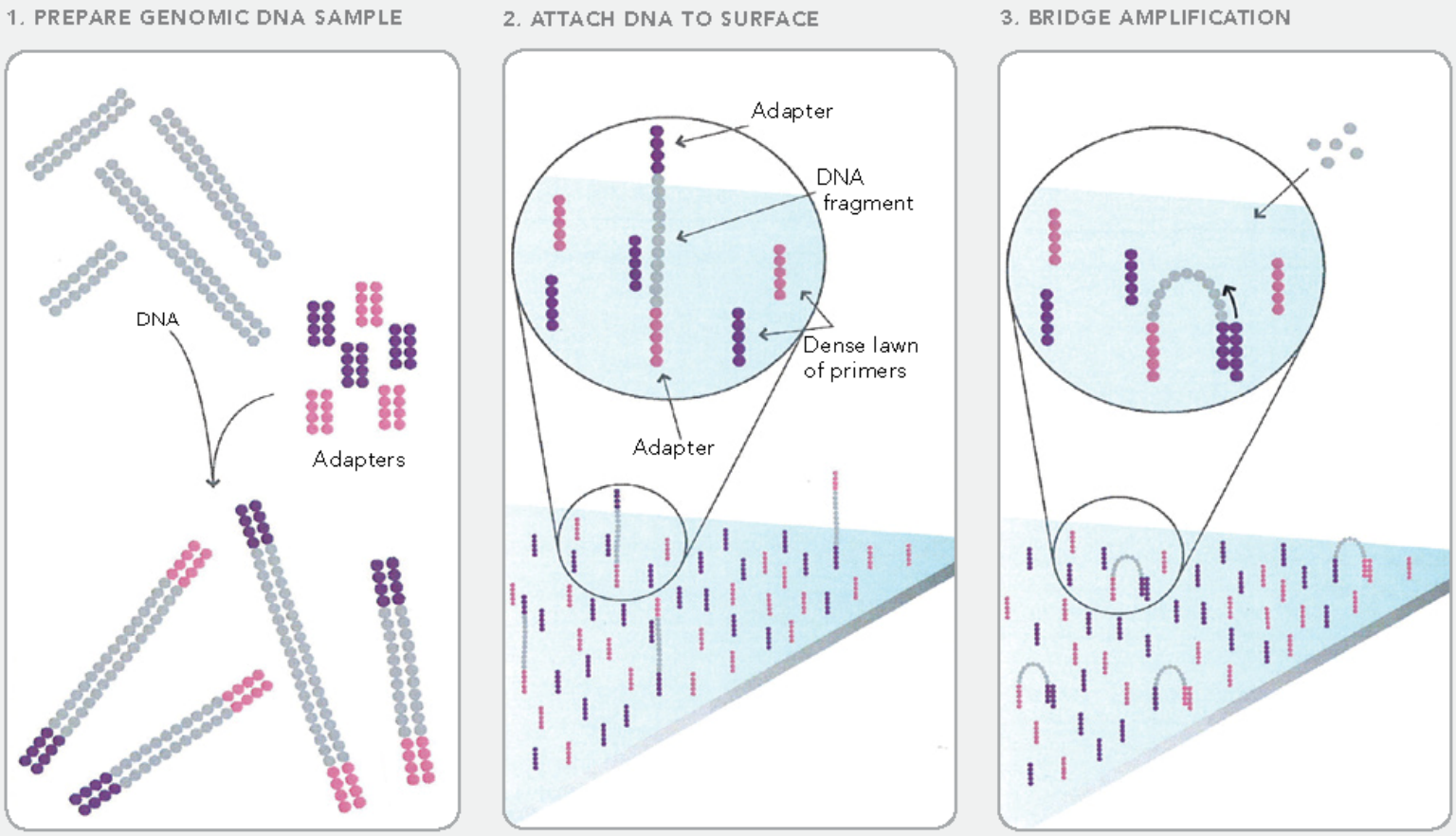
- Produces more than 50 million reads each 30 – 72 long prefix or suffix sequences of DNA fragments with length 100 to 500 base pairs
- Reads Back-mapping to the reference genome (parallelized on multiprocessor machines or run on computer grids)
- Analysis: to assemble a genome, to determine the transcripts and their concentrations, to detect nuclesome positions, to identify single nucleotide polymorphisms, or to estimate copy number variations

→ http://www.ensembl.org/index.html

Part II: Genome Analysis

Solexa

FIGURE 2: SEQUENCING TECHNOLOGY OVERVIEW



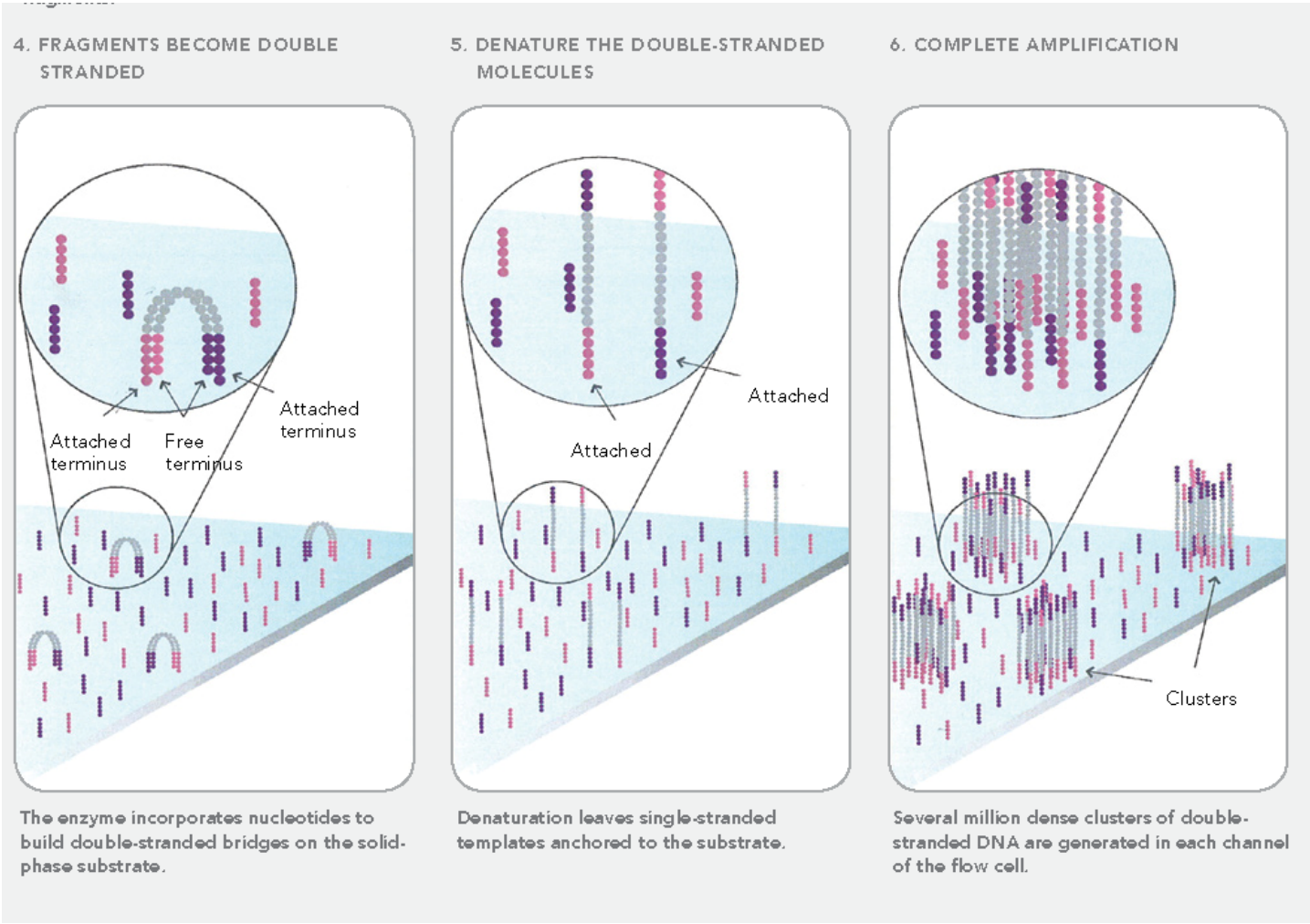
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

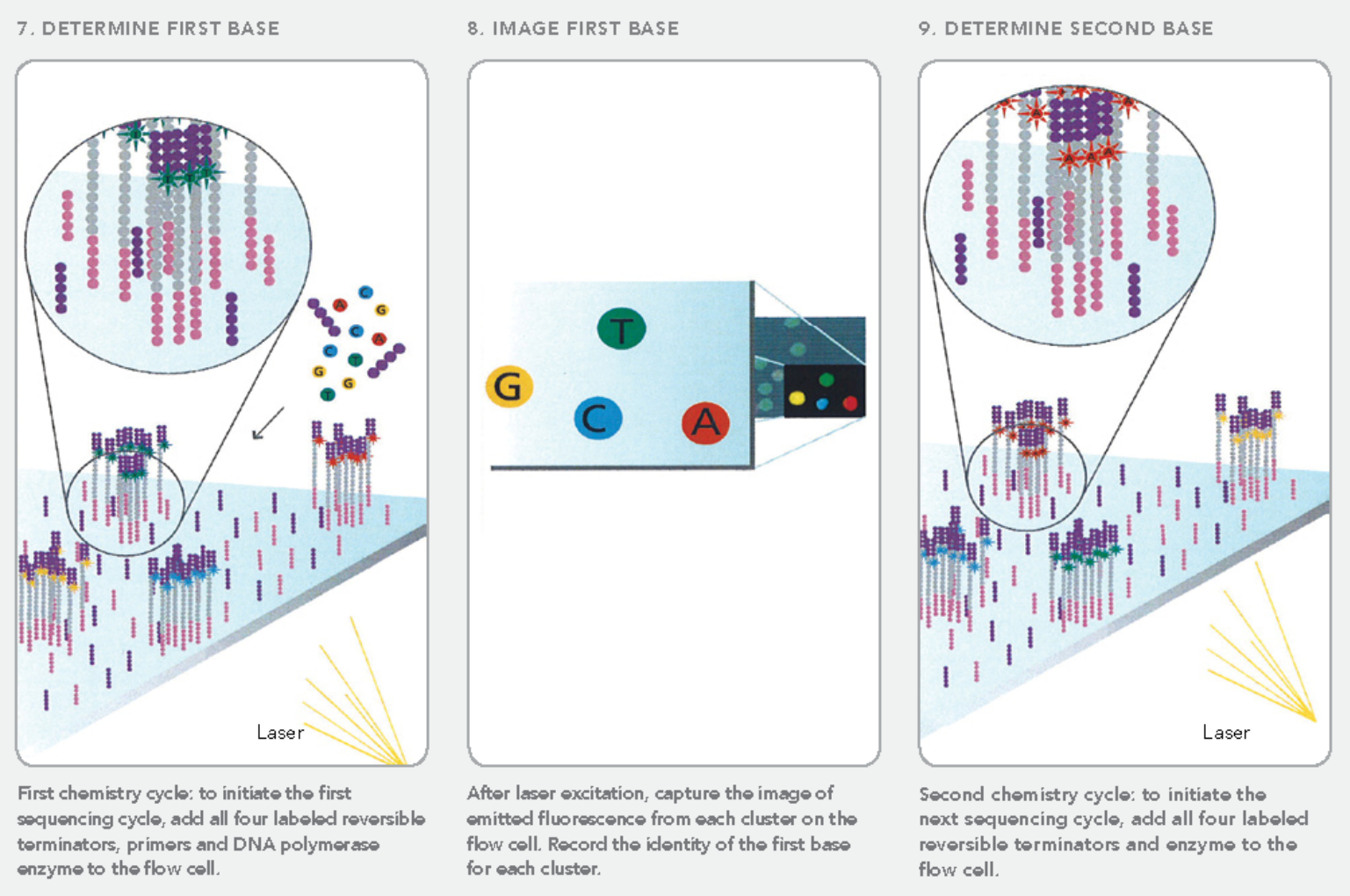
Part II: Genome Analysis

Solexa



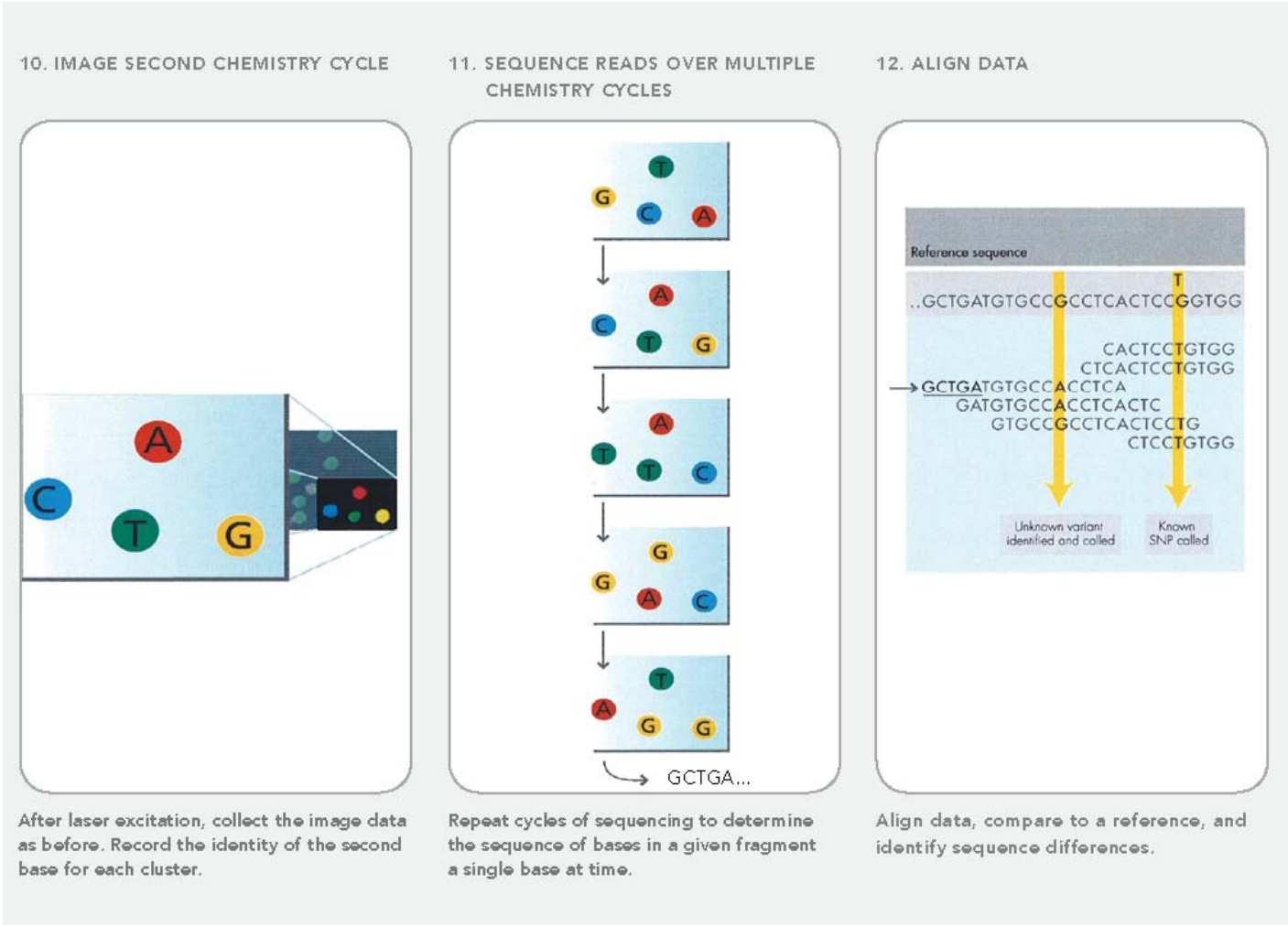
Part II: Genome Analysis

Solexa

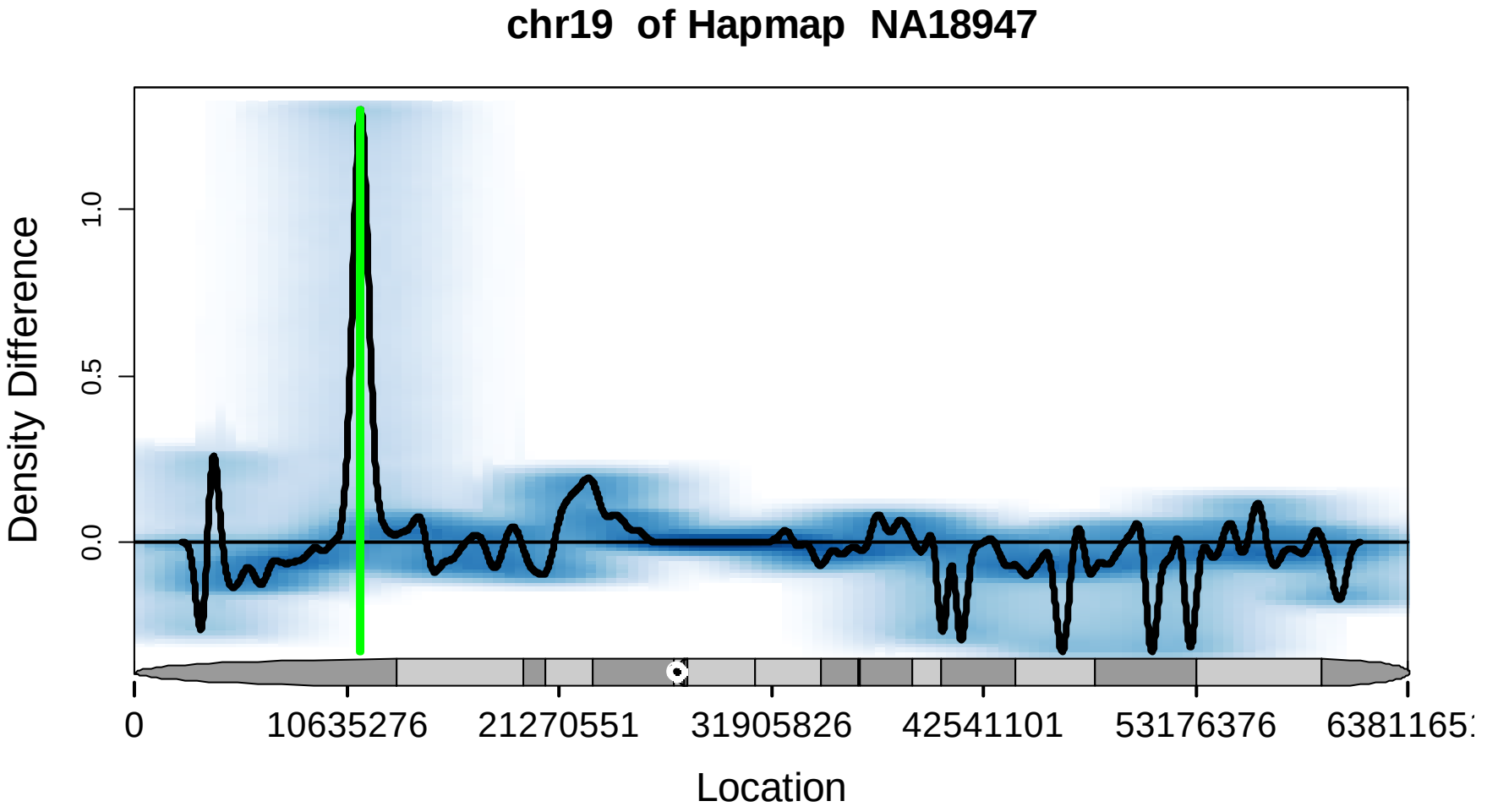


Part II: Genome Analysis

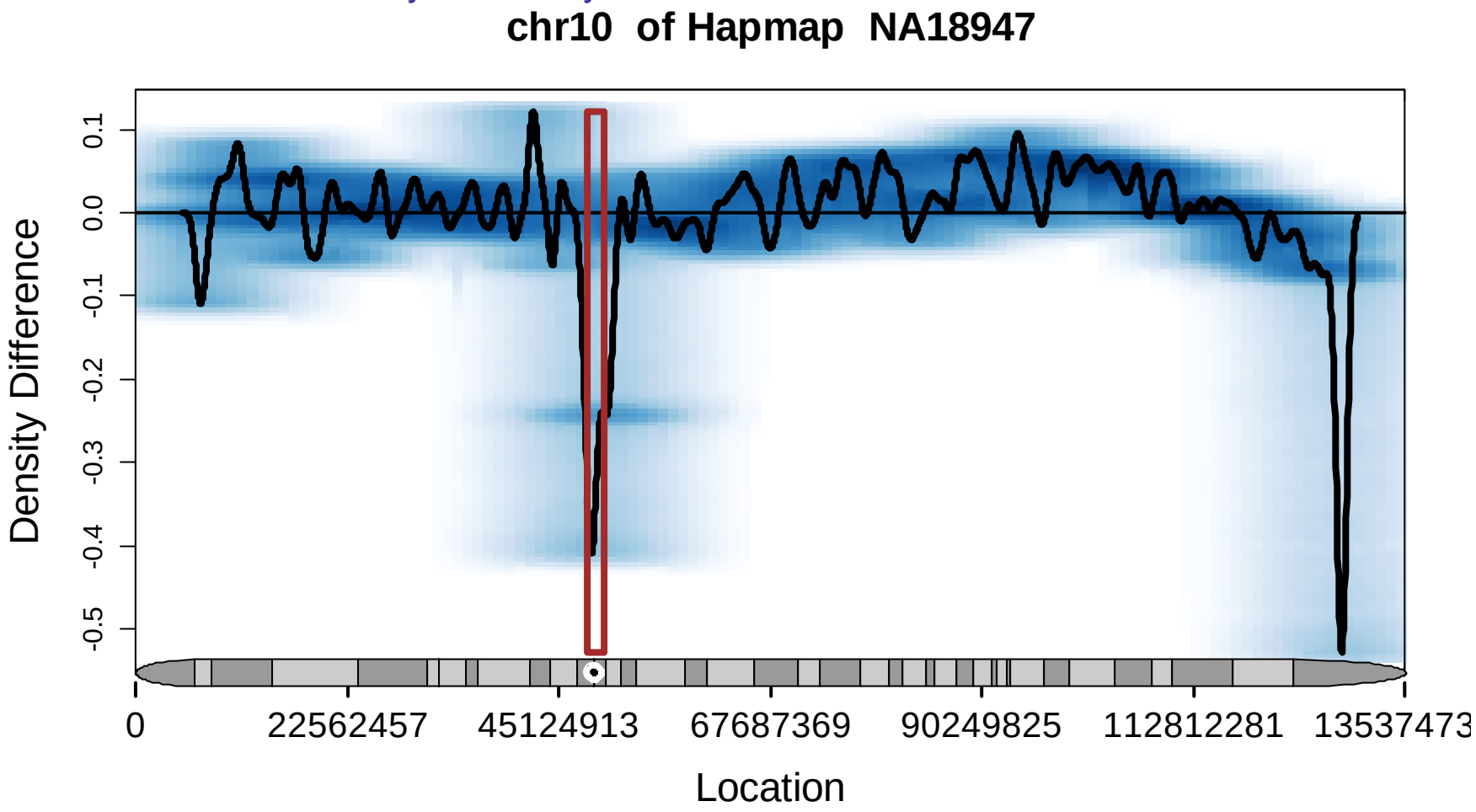
Solexa



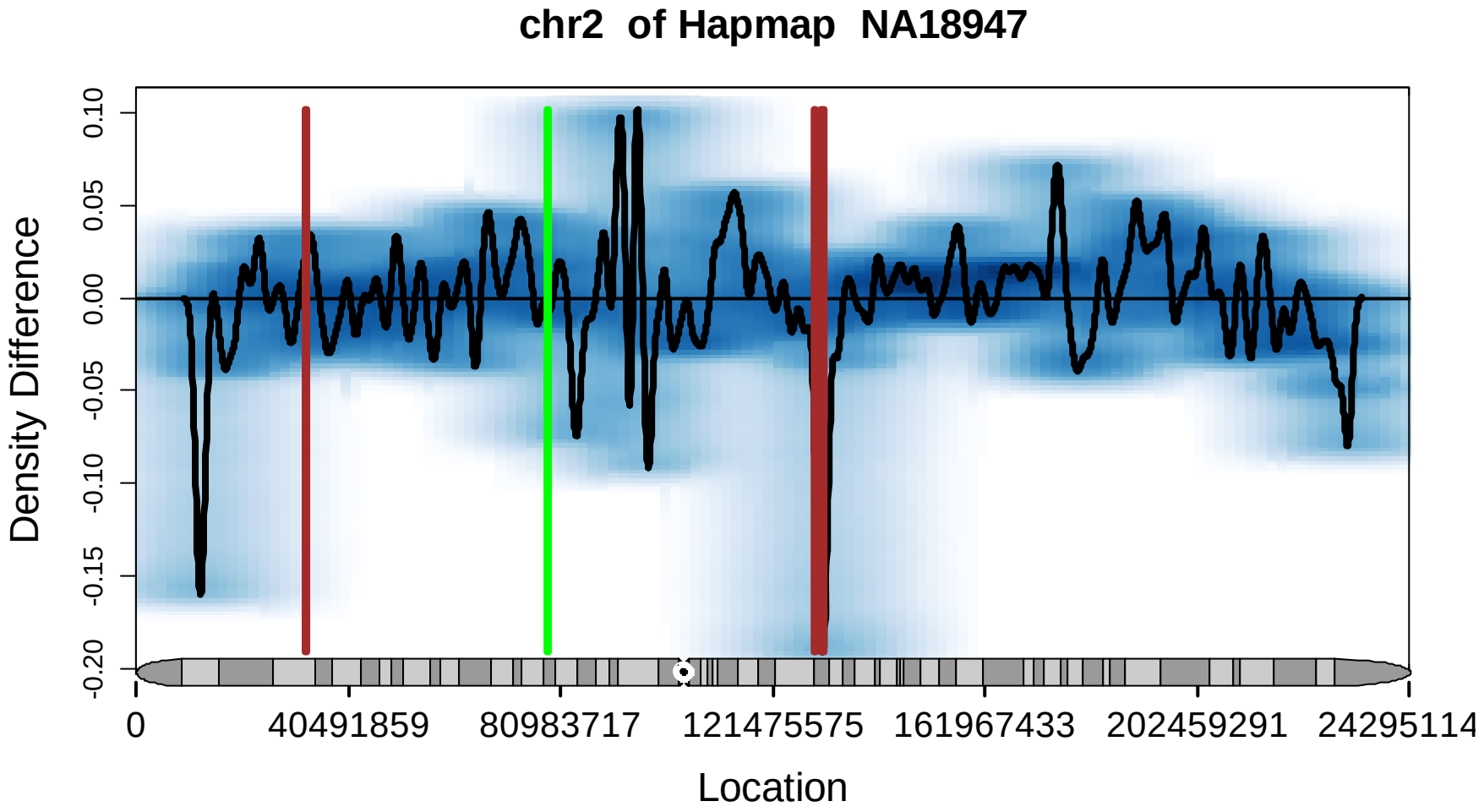
Analyze Solexa sequencing data in R An amplification (vertical line) in chromosome 19 detected by BAC arrays



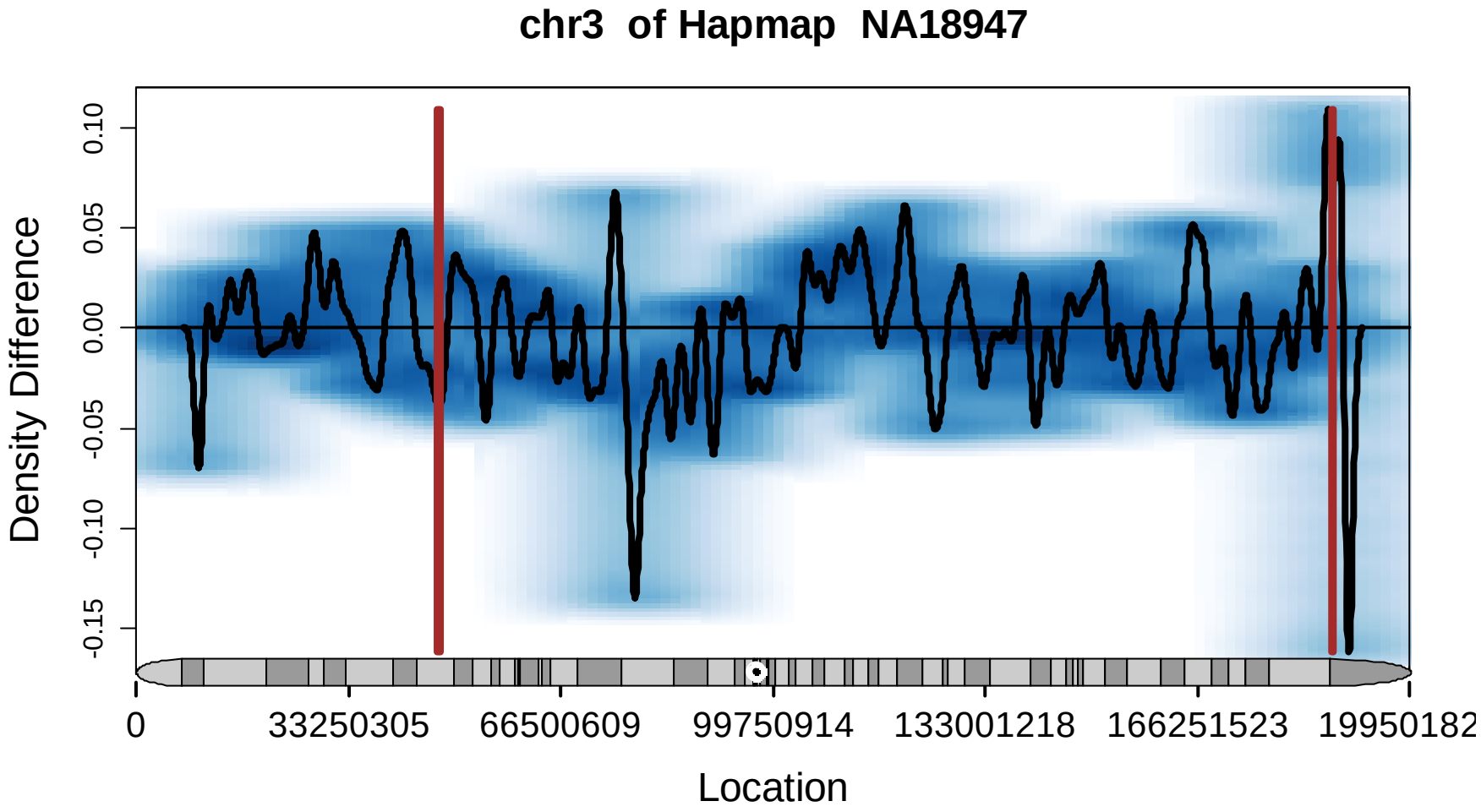
Analyze Solexa sequencing data in R A deletion (vertical rectangle) in chromosome 10 detected by BAC arrays



Analyze Solexa sequencing data in R Unexplained



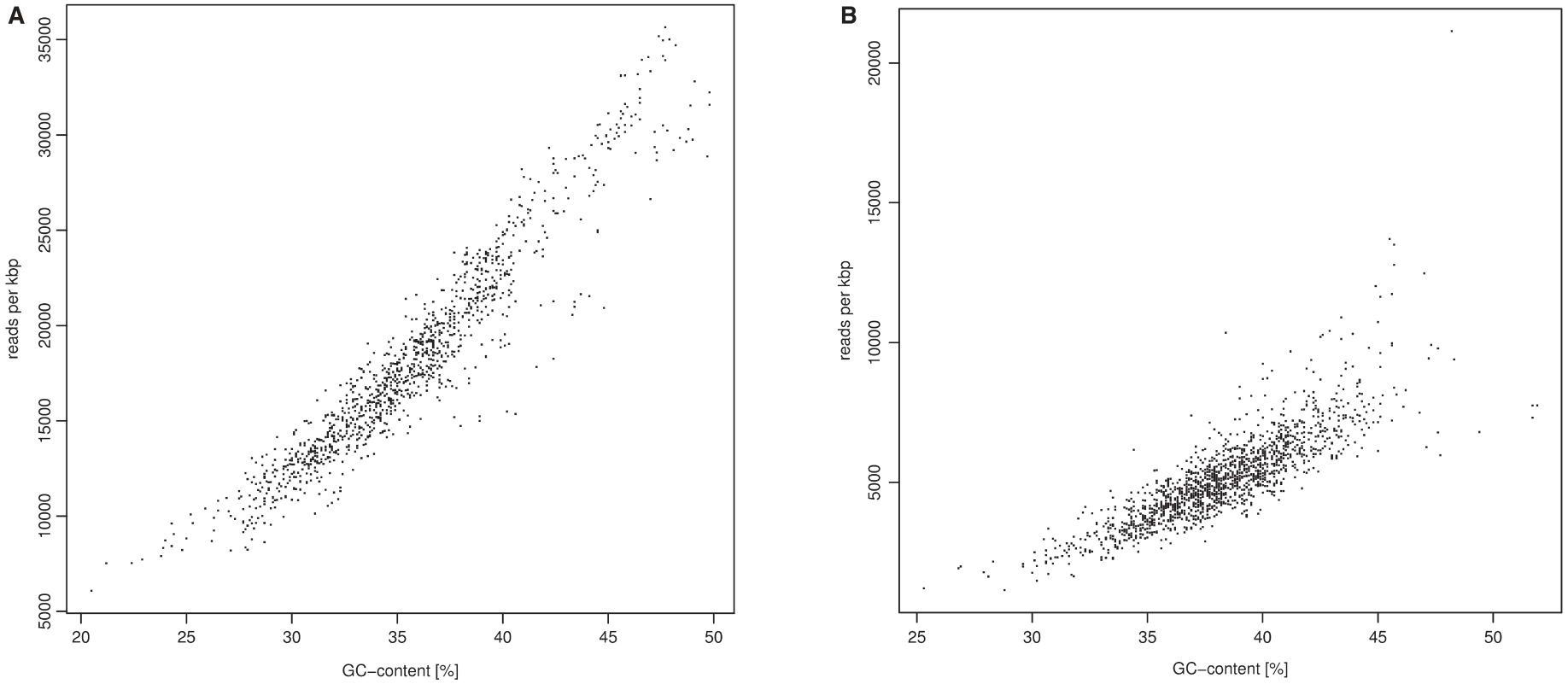
Analyze Solexa sequencing data in R Unexplained



Analyze Solexa sequencing data in R

Unexplained

Figure 2. Correlation of the Solexa read coverage and GC content. (**a**) 27mer reads generated from *Beta vulgaris* BAC ZR-47B15 (**b**) 32mer data set from the *Helicobacter acinonychis* genome. Each data point corresponds to the number of reads recorded for a 1-kbp window (shift of 100 bp in *Beta* and 1 kbp in *Helicobacter*).



Analyze Solexa sequencing data in R

Unexplained

Figure 4. Frequency of wrong base calls in Solexa reads depending on the position along the read (27mer reads from *Beta vulgaris* and 32mer reads from *Helicobacter*). (**a**) Error frequency per position calculated from considering wrong base calls only. The highest error frequency is observed at the read 3′ end. (**b**) Per-base error rates (overall error frequency per position considering all base calls).