Biata 2018
Lectures and posters

Tsurinov Petr
30.07.2018

Lectures

Rob Knight. Tools to link Human and Environmental Microbiomes for Health

- 2-20 mln microbiome genes vs 20k human
- De nova → Tree-building → UniFRAC measuring (biological communities)
- Connection between human and environment via microbiology (coffee ex.)
- Earth soil system model framework (microbiote + satellite data)
- Analysis of microbiote show durty coming from Africa to Canada
- Difference between microbiote taken from human (oral vs fecal, no blood)
- Ability to make reaction predictions
- Health depends on microbiote and high diversity is better than low
- Connection between illnesses and microbiote (ex. cystic fibrosis lung)
- Fecale transplant and smart toilets

Mark Kowarsky. Exploring the microbiome of primates using cell-free DNA

- Circulating nucleic acids, 99% is unknown
- Input data is taken from blood and than go through assemble pipeline
- Many new contigs which couldn’t be mapped on usual microbiote (short size)
- Different primates share a lot of these microbiote
- Exist in blood for few months
- Correlates with environment (zoonotic taxa)
- Can affect during transplantation
- Whole new world of unknown

Amanda Birmingham. QIIMP: Microbiome metadata made easy

- Metadata is needed for hypotheses and should be standardized
- Researchers don’t provide required metadata and always try to use Excel
- Quick and Intuitive Interactive Metadata Portal
- Produce correct template for Excel after gathering info about experiment
- From 8 hours of preprocession to 10 minutes

Martin Steinegger. New algorithms and tools for large-scale sequence analysis of metagenomics data

- MMseqs2 - protein sequence and profile search method
  - Use CPU cache efficiently than competitors
  - 400 times faster than PSI-BLAST and slightly more sensitive
  - Can annotate 1.1 billion sequences in 8.3 hours on 28 cores
- Linclust - sequence clustering method
  - Based on 20 kmers, has linear working time (conventional have quadratic size)
  - More than 1000 times faster than others on 1.6 billion sequence fragments
  - For 1.6 billion metagenomic sequence fragments it takes only 10 hours on a single server
- PLASS - metagenomic protein level assembler
  - Linear working time and memory usage
  - Faster than competitors (Megahit and etc.)
  - Can assemble 10 times more protein sequences from soil metagenome than competitors

Ami Bhatt. Highly contiguous microbial genomes from human and marine microbiomes

- 16s sequencing; metagenomic sequencing + short read analysis
- K-mer based taxonomic limited (soil - 3%; fecale - 30-60%)
- Leukemia - 63% dies after operation not because of primary disease
- More than 40% of patients have low diversity of microbiome after operation
- MAGs - metagenome assembler
- Solving repeats problem using additional primers around repeats
- Microbes evolve very fast - horizontal transfer of antibiotic resistant genes
- Human genomics incomplete without microbes
- Simplified microbiome is associated with disease

Jean-Francois Flot. Assembling a (diploid/polyploid) genome to perfection

- Currently heterozygosity is usually ignored
- High allelic divergence make complex to assemble
- Adineta vaga - bdelloid rotifer with tetraploid genome
- 100X 2*250bp Illumina + 100X PacBio + 100X Nanopore + 150X 3C
- Using MIRA - from 15 mln to 65 unresolved polymorphism
- Bwise - new paradigm for genome assembly

Nickolai Alexandrov. 3000 rice genomes sequences

- Most important crop in the world, but still groth amount is not big enough
- On average 15x coverage (from 5x to 50x)
- SNP-Seek portal for allele mining and genome variations visualization
- Clustering different kinds of rice
- Found new cluster: subtropical japonica
- Building phylogenetic tree
- Yet not investigated promoter and enhancers regions
- Google deepvariant, GWAS, IRIC

Anton Nizhnikov. Protein storage in plant seeds is associated with the amyloid formation

- Amyloids - proteins with specifical construction (cross-ß)
- Extreme resistance to chemical and physical influences
- Connected with serious diseases such as Alzheimer and Parkinson
- Previously amyloids were not found in plants
- Using Waltz and SARP they found some potentially amyloidogenic proteins
- After deep analysis were proved that part of these proteins contain amyloidogenic domains (Cupin-1, Zein, Gliain, Vicilin, Glutenin)
- Novel molecular mechanism for long-term stabilization

Gene Myers. Towards perfect de novo DNA assembly

- An era of DNA sequencing: de novo, reference-quality genome for 1k€
- Shotgun sequencing + PacBio (long reads and random error)
- New algorithms for long reads and HQ reconstructions
- 9 genomes in 2016 meets HQ standard, now much more
- 60x PacBio (10k€) and illumina + bionano (5k€) = HQ, but still not 1 contig
- Project for sequence everything
- String graphs for repeat problem solving
- Long reads re-energized de novo
- Performance of algorithms is main bottleneck

Antoine Limasset. Bwise: a novel accurate, haplotype-specific genome assembler

- New type of de Bruijn graph for keeping info from short-paired-end reads
- Long reads via nanopore to solve unsure situations
- Preprocessing of input data for better DBG construction
- MIRA assembler - 47kb N50 in 9 month, Bwise - 150kb N50 in 2 hours
- Human genome in couple of days even with 100x with less than 100GB RAM
- Can be used for haploid and meta-genomic datasets
- Platanus worked bad in some cases, Deepspades doesn’t have some functions

Andrey Prjibelski. Assembling barcoded RNA sequencing data

- De novo transcriptome assembly as alternative to classic reference-based
- Actual problem of restoring complete full-length isoforms
- Assembly graph for data merging regarding to barcode info
- Using barcode is solution and new modification of rnaSPAdes can process it
- Investigated human brain samples from 50 donors appeared as 2bln reads
- Found new alternative isoforms for some genes (BIN1, MAPT)

Dmitrii Meleshko. BiosyntheticSPAdes: biosynthetic gene clusters from assembly graphs

- Is required for discovery antibiotics and natural products
- Biosynthetic gene clusters prediction in fragmented genomic assemblies
- Old tools work within contigs, but they can be small
- New tool combine multiple contigs in segments encoding long BGCs
- Found more result than previously, some still with variations
- Able to use mass spectrometer data for variation solving (usually we have it)

Dmitry Antipov. Plasmid assembly in genomic and metagenomic datasets

- Isolated plasmid detection is ok, but problems lies in metagenomic data
- Problems: bubbles, chimeric reads, repeats in one and shared
- Usual workflow doesn’t work because chromosome coverage is missing
- Idea is to build graph and remove edges until we can see plasmids
- Created new tool - metaplasmidSPAdes
- Found 3 new plasmids in should be known data
- Found new results in Crohn reference free data

Alexey Samosyuk. CellPi: scRNA pipeline

- Set of known tools (some with modifications) combine to pipeline
- Capable to separate highly homogeneous cell populations (6-10 cells)
- Is able to detect inseparable sub-populations (2000 cells)
- Have higher performance and provide high-dimensional tSNE
- Combine different scoring strategies (components, slide window, etc.)
- New version will be available not earlier than late autumn
- https://github.com/testaibot/CellPi

Mikhail Gelfand. Test tubes, sequencing machines, computers: bioinformatics as a molecular biology tool

- Large-scale data analysis give ability to make predictions which then tested
- Predictions can be not just annotations, but novel biological phenomenons
- Second lactose catabolism pathway
- Regulation of hexuronate catabolism: ExuR regulates other regulators
- YjjM regulator (control a lot of genes) regulated by UxuR
- Fly Lavra dehydration and temperatures (improved RNA and DNA sequencing)
- Look for motif in regions of gene graphs of motif locations with active genes

Pavel Yakovlev. Semantic-based antibody folding and structural annotation

- Old fashion drug discovery
- GPU protein-protein docking
- Gibbs free energy
- New tool doesn’t use scoring function, but truly find minimum ΔG
- Much faster and 80% more accurate
- No real compare with other tools
- No plans for small molecules docking

Eugene Koonin. CRISPR: fascinating biology and limitless applications

- CRISPR history
- Work scheme: Preparation → Recognition → Incorporation
- Highly diverse in immune systems
- Classification of CRISPR-Cas systems
- 395 Cas proteins profiles
- Type II class 2 systems - simplest CRISPR-Cas systems
- New type V-A CRISPR-Cas Cpf1 even simpler
- Insertion of unrelated domains
- Pipeline for discovering new CRISPR-Cas systems

Oleg Gusev. Promoters and enhancers landscape of embryonic development and hibernation in chicken

- Cap analysis of different RNA of same genes
- On early stages embryo differentiation can be stopped just by temperature
- Late stages - Nanog which affect CDYL gene
- 144 promoters are affected on all stages
- Oct4, PouV and Nanog regulation
- So now it’s possible to stop development on late stages and then continue using Nanog

Alla Mikheenko. Analysis and visualization of segmental duplications in mammalian genomes

- Segmental duplications (SDs) correlates with genomic diseases
- More than 1kp, hard to find because they are low copy repeats
- New tool SDquest find out that in human genome 6-7% are SDs
- RepeatMasker → Repetitive k-mers → putative SD → verify using filtering
- Breakpoint graph SD and de novo SD detection
- Yeast assembly graph
- Human assembly graph processed using partitioning
- Flye visualizer - suitable tool for SDs visualization (not able to find in Google)

Richard Durbin. Sequencing genome diversity in fish

- Three sequencing eras (1990 - 2005, 2005 - 2017, 2015 - till now)
- 494 vertebrate genome assemblies (55 human)
- Half of vertebrate species are fish, so focus on them
- Lake Malawi 500 species due to adaptive radiation
- Currently 80 species with 15-20x coverage
- Building Malawi phylogeny tree
- Specie from one branch can move to another
- Heterozygous case errors during assembling
- Trio binning - father + mother + child assembling; mother + offspring
- Zebrafish assemble already better than common assemble

Stephen J O’Brien. The Genome Russia Project 2018

- 11 time zones, 146 mln people, more than 200 ethnics
- 1000 genomes project is missing Russia, Canada, Australia
- Information about human migration, mutation discovery
- Healthy people + trio technology; 264 people - 55 ethnic minority
- Everything except some exceptions clustered according to territory
- 10,5M SNP; 2.8M indels; 1700 long indels
- Outlier function genes (lactose intolerance, warfarin response, etc.)
- HapMap of Ethnic Russia
- Already got 1000 samples
- Hope work will be finished after he gone

Alexander Tiskin. Bounded-length Smith-Waterman alignment

- Semi-local longest common subsequence (LCS) and edit distance
- 1974 - O(mn); 2008 - O(mn/(log(n)^O(1)))
- Dynamic programming + graph construction
- Convert to Levenberg distance
- Looks interesting, but no implementation yet
- Surprise from questions after lecture -
  Meyers created this algorithm 20 years ago

German Demidov. ClinCNV: novel method for large-scale CNV and CNA discovery

- Germline copy number var. (CNV) and somatic copy number alt. (CNA)
- Involved in many genomic disorders (ex. Schizophrenia or cancer)
- WGS and WES data is complex for CNV/CNA detection
- New method ClinCNV combines different types of data (WGS, WES, microarray)
- Immunotherapy can be harmful and to predict it such information needed
- Collect → de noise → find segments
- Dividing genome on not overlapping windows, combining data from all sources
- There are alternative methods and recommendation is to use all together

Genadij Zakharov. Pipelines in the cloud

- NGS pipeline with good characteristics
- SevenBriges; FireCloud; DNAnexus
- Bunch of tools to setup pipeline
- NGB (New Genome Browser) with structural variation support
- Integration NGB with databases
- PeerJ publication with paper quality images taken from NGB
- Open source tools optimization
- Docker container to run browser in one command
- Useful if your data is in cloud without download possibility

Posters

Геном России

- Планируется 2.5 тыс. человек, сейчас уже есть 250
- 3 поколения
- Пересечение с древними культурами (различные неандертальцы и пр.)
- Территориальная кластеризация и по типу деятельности (охотники и пр.)
- Перемещение генов между древними и к современникам

Геномика стерляди

- Сборки нет в силу полногеномных дупликаций и медленной эволюции
- 60 - 360 хромосом, очень сложно собирать
- Выцепление отдельной хромосомы и её секвенирование
- Много горизонтального переноса
- Пока невозможно отличить пол по геному
- Если получится отличать по полу, то это золотая жила

Транскриптом артишока

- Illumina + ONT
- Полные транскрипты → альтернативный сплайсинг
- Текущий референс плохой
- Найдено много новых изоформ

Паразиты

- Выделение отдельно от ДНК хозяина
- У инфузории (хозяин) большой ген, у проспариды (паразит) маленький
- 1,5 тысячи видов, секвенировано только 18-20
- Поиск принципов паразито-хозяйственных отношений
- Актуально для решения проблем в областях производств (шёлк, мёд, пр.)

Архереи

- Небольшой геном, нет ядра
- Существуют на дне марианской впадины и в термических источниках
- Ферменты термофилов очень полезны в производстве
- Растут на сахарах, требуют симбиотов и клеточный фильтрат
- Не могут синтезировать очень многие вещества для жизнедеятельности
- Было исследован минимум веществ, необходимых для существования
- Обнаружены новые пути выработки необходимых веществ
- Проанализированы различные сахара и выбраны подходящие

Нанороботы от всех болезней

- Основаны на принципе расщепления ДНК и представляют собой ДНК
- Прицепляются к днк вирусов и разрезают их тем самым нейтрализуя
- Опробовано на вирусе гриппа А и раке - нейробластоме
- Создана сложная структура, чтобы ловить нужные вирусы
- Пока есть побочный эффект - убивает всё остальное тоже
- Пытаются использовать разные структуры для понижения токсичности
- Альтернативное использование - детектирование вирусов (прим. HPV)

QUAST

- Валидатор качества геномной сборки
- Использует различные метрики (прим. сравнение с референсом)
- Строит статистики по различным видам ошибок
- Анализализирует репиты
- Использует уникальные k-меры длинной 101bp

Дисбиоз кишечника

- Анализ эффективности пробиотиков и аутопробиотиков
- Контрольная группа, группа после антибиотиков, группы после пробиотиков
- 16s rRNA анализ состояния микрофлоры
- После антибиотиков состояние микрофлоры совсем не разнообразное
- Анализ alpha diversity, чем выше показатель, тем лучше
- PCA microbe composition + unifrac distances
- Статистически значимых результатов пока нет, планируется больше доноров