# H3ABioNet

Pan African Bioinformatics Network for H3Africa

Special Issue: BioRes Digest

Issue 22: June 2017

H3ABioNet

In This Issue

Continue the conversation:

#bioinformatics #Africa
#H3ABioNet @H3ABionet

Foreword

We now have less than 2 months to go until the official end of the H3ABioNet grant! There are several deliverables to finish off, and as fast as we do that new ones appear. This is a result of the dynamic nature of the project and adaptations to emerging user needs.

The Introduction to Bioinformatics (IBT) online course is progressing well, as is the pilot Genomic Medicine course for nurses. This is down to the hard work of Kim Gurwitz (IBT), Vicky Nembaware (Genomic Medicine), and the excellent task teams they are working with. We are grateful to the trainers and classroom staff for helping these to run smoothly.

We are working on organization of upcoming workshops/hackathons, which include a GAPWG (Genome Analysis Publication Working Group) session to make headway on the genome analysis publications, a workshop to continue progress on the H3Africa catalogue, and a career development workshop. If there is interest we will also plan a chip data analysis workshop for when the data arrives.

For some of us, the biggest relief is completion of the much awaited H3Africa chip design. The chip is being manufactured for release in a couple of months, and we look forward to seeing the first dataset coming off the chip. I hope it lives up to expectations. You can read further about the chip design journey in this newsletter.

Another exciting milestone we reached was submission of the first H3Africa dataset to the EGA, thanks to the Infrastructure working group and the archive team. The team is now working on more dataset submissions to the archive and preparation of the data for the EGA.

Also in this issue is the second issue of the BioRes Digest, a product of the RSWG, which aims to highlight interesting papers as a means to remain current with new tools and technologies in bioinformatics.

Please read on now to get more details on the H3ABioNet activities over the last few months.

Nicky Mulder

Education and Training Working Group

It has been a relatively quiet period for the Education and Training Working Group with work continuing on wrapping up on-going projects that have been identified as key deliverables/ products for the final year of the project.

- The Introduction to Bioinformatics course 2017 (IBT_2017) is currently underway and running smoothly. In this year's iteration of the course we have introduced a 'meet the classroom' segment during our biweekly virtual classroom meetings in Mconf to allow each classroom the opportunity to activate their webcams and share some interesting facts about their institutes with the rest of the classrooms. It has been a great way to create a sense of community amongst the classrooms, but has also highlighted the challenge of consistent, high-quality Internet connectivity across Africa in some instances. Although this is the second iteration of the course, we are still learning and identifying new ways to improve the course. On a related note, a paper describing the design of the IBT course has been submitted for publication and we hope that it will be well received.
- The updating of the H3ABioNet bioinformatics curriculum website is still on going with most of the content for each of the modules completed. A set of guidelines for implementation of the curriculum has also been developed and is available on the *bioinformatics curriculum website*.
- The curation of H3ABioNet training material has been completed and the task force is in the process of sourcing missing material and adding the course details to the eGenomics catalogue.
- In addition to the training material, a set of policy documents has been generated regarding procedures for hosting a workshop. These guideline documents are being reviewed and collated as a resource to be made publically available for use by others.
- Additional projects identified as possible deliverables for the working group include the creation of online tutorials for the tools and resources that have been developed by H3ABioNet as well as the possible translation of some of the training material videos into other languages, such as French and Arabic, to increase accessibility.
- On the workshop front, planning of the Career Development workshop, to be held in conjunction with the *ASBCB conference* and H3ABioNet Scientific Advisory Board meeting in October, is still underway. The curriculum for the workshop has been refined and a task force is working on finalizing the content and trainers for the workshop.

#H3ABioNetEducationAndTraining **Nicky Mulder and Shaun Aron**

Back to Contents

Issue 22: June 2017

H3ABioNet

Infrastructure Working Group

Infrastructure Working Group

eBiokits:
A number of nodes are in the process of acquiring eBiokits. Several have had trouble in acquiring the hardware in their own countries. We are trying to find a solution to this.

Aspera/Globus Online:
One of our projects which has attracted significant attention is our project to compare Globus Online and Aspera. As you will know, H3ABioNet has been using Globus Online for transferring very large data sets successfully for some years, and its entry level service is free. Aspera is a similar service and a number of companies and organizations use it. However, setting up an Aspera server costs money. A thorough evaluation of the effectiveness of these services in Africa would be very valuable for H3Africa to know how these services compare. We will do comparisons against several H3ABioNet nodes.

Globus Online/Netmap:
We are trying to get as many nodes up and running by the end of July. This your last chance. There are still quite a few nodes that are down. Suresh Maslamoney (UCT BIO) and Ines Tiouiri (IPT) are trying to get them operational. We have managed to bring some nodes back online since the last meeting but at least 3 of them are down again.

Data Management Task Force:
The DMTF has been very busy and are expecting lots of data in the next few months. We've reached an exciting milestone with the submission of the pilot data from the AWI-Gen group to the EGA.

The NIH-funded project "Clinical and Genetic Studies of Hereditary Neurological Disorders in Mali", led by Dr Guida Landoure, submitted their data to H3ABioNet and it is now being checked. Several other H3Africa groups are preparing to submit data.

The South African Human Genome Programme has requested our help in submitting the SAHGP data to the EGA.

Some upgrades to our storage are being planned to accommodate the anticipated increase in dataset submissions.

Cloud Task Force:
This group is making progress on completing the four pipelines that it has committed to. Some final tweaks and testing are being done. A draft paper has been prepared and is almost ready to be submitted.

#H3ABioNetInfrastructure **Scott Hazelhurst and Suresh Maslamoney**

Back to Contents

Continue the conversation:

#bioinformatics #Africa
#H3ABioNet @H3ABionet

Issue 22: June 2017

H3ABioNet

Research Working Group

Continue the conversation:

f 🐦

#bioinformatics #Africa
#H3ABioNet @H3ABionet

Research Working Group

The theme for the H3ABioNet seminar series for May 2017 was on Functional genomics and the presenter was Dr. Cameron Mac Pherson from Pasteur Institute in Paris, France. The title of his presentation was "Functional genomics in a time of precision medicine". Dr. Cameron gave an excellent, articulate presentation. He stressed the issue of heavily investing in treating patients as functions of their genetics and environment rather than relying exclusively on the outcomes of clinical trials. He talked about the various practical, biological, and analytical challenges to overcome before individualized treatments/Precision medicine are realized.

The theme for the H3ABioNet seminar series for June 2017 was on "Genetic anthropology" and the presenter was Dr. Raymond Tobler from the Australian Centre for Ancient DNA at the University of Adelaide, Australia. His seminar talk was on his recent research work which has recently been published in *Nature* with the title "Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia".

RHS: Dr. Cameron, Pasteur Institute in Paris, France. LHS: Dr Ray Tobler is an ARC Indigenous Fellow currently working at the Australian Centre for Ancient DNA (ACAD) at the University of Adelaide

Juanid Gamieldien from SANBI has sent a brief summary of his recent publication entitled *A practical guide to filtering and prioritizing genetic variants* as the main article summary for the 2nd issue of the Digest. You can read this quarter’s issue of the BioRes Digest in this issue of the H3ABioNet newsletter. A call for volunteers is open for the 3rd and 4th issues that will be published in September and December, respectively. The contribution of graduate students is highly encouraged and there is a possibility to join the editorial team of the BioRes Digest for the active graduate students.

Michael (RUBi node) briefed the working group members on the progress of the integration of all the received pipelines. The team mentioned that they are on track to finish both the integration and the manuscript draft by the end of July.

Issue 22: June 2017

Research Working Group

The Chairs identified the outputs and products of the Research Working Group as part of the process of wrapping up activities by the end of July.

The DREAM of Malaria hackathon concept paper has been sent to a *Nature* comments editor who advised to write a blog in *Nature Careers* instead. A blog post is under preparation and the concept paper will be sent to *Genome Science* and to *BioRxiv*. All the hackathon participants will be co-authors. The majority sent their approval and reviewed the manuscript. All the comments and suggestions of the participants were taken into consideration. A paper summarizing the data analysis pipelines and the results achieved by the 3 teams who participated in the hackathon will be published before November 2017. The team of data generators from University Notre Dame are working on cleaning up the dataset and finalizing the generation of the test set that will be released for the DREAM Challenge. This will be officially announced in November and open to the community in early 2018.

Writing up of Open Science activities, showing their relevance to H3ABioNet, is in progress. The idea is to highlight the peer learning environment that the study group created and the fact that this helps to increase collaborations between people from the same institute and to increase interactions with the open science community.

#H3ABioNetResearch

Faisal Fadlelmola and Amel Ghouila

Back to Contents

Continue the conversation:

#bioinformatics #Africa
#H3ABioNet @H3ABionet

Node Assessment Taskforce

As mentioned in our last report (April 2017), the interest of H3ABioNet Nodes in taking Assessment Exercises has picked up dramatically. The CBIO node has successfully completed the 16S rDNA metagenomics analysis exercise, which received very positive reviews and was formally awarded end April. Three more Nodes are in the process of taking the 16S rDNA exercise (Wits, IPT, and NABDA), while two are taking the variant calling exercise (CUBRE and Malawi), and one (NABDA) is also taking the GWAS exercise. We hope that we will be able at the time of the next Newsletter to announce six more successful completions!

To keep up with the increased workload generated by this surge in interest, we have called on existing Node Assessment Taskforce (NATF) members to confirm their interest in participating in the revisions of the SOPs and the generation of new datasets, and in identifying external experts willing to review the submitted reports. We are also looking for additional volunteers willing to invest some time and effort in these activities. Given the rapid evolution of genomic datasets and the methods used to analyse them, just keeping up with current best practices is a demanding job.

The manuscript describing the work of the NATF, entitled "Assessing Computational Genomics Skills: Our Experience in the H3ABioNet African Bioinformatics Network", is now available online (*here*) from PLoS Computational Biology, and is featured on the Journal's home page. We welcome another landmark paper from the H3ABioNet network!

Victor Jongeneel

Back to Contents

Issue 22: June 2017

H3ABioNet
’Meet the PI’
Interview

H3ABioNet ’Meet the PI’ Interview

Dr. Jonathan Kayondo,
Principle Investigator (PI) at the Uganda Virus Research Institute (*UVRI*)

Interviewer: I am Wisdom A. Akurugu, a Bioinformatician working with the *NMIMR* node. I have been in the institute from July 2013 and have been involved with a number of activities of H3ABioNet including workshops, working group meetings, to name a few. My work schedule includes the provision of bioinformatics support and assisting life scientists in basic and advanced bioinformatics. I also take part in various bioinformatics related projects carried out in the Node but most importantly, I prepare training materials and teach in bioinformatics workshops.

Details of the interview:

Wisdom: Tell us a bit about yourself

Dr. Kayondo: I am a senior research officer, that is my rank, at the Uganda Virus Research Institute (UVRI) based in Uganda and I have a molecular genetics background. I have a Bachelor of Science degree in chemistry and biochemistry, from Makerere University in Uganda. Then I have a Ph.D. in vector biology and parasitology from the University of Notre Dame in the USA. And then I did my postdoctoral in molecular virology here at Uganda Virus Research Institute. Essentially, I am interested in conducting scientific investigations on disease vectors and pathogens, obviously in order to contribute to knowledge, policy and practice. But at the same time, I also have interest in building capacity for future sustainability and I lead a lab of 11 people at the UVRI.

Continue the conversation:

#bioinformatics #Africa
#H3ABioNet @H3ABionet

H3ABioNet 'Meet the PI' Interview

Wisdom: Please Sir, can you tell me more about the Uganda Virus Research Institute (UVRI) that you are working with?

Dr. Kayondo: Yes, the UVRI is one of the leading health research institutions nationally in Uganda and I can also say regionally in East Africa. It started way back in 1936 primarily as a yellow fever research centre but over the years, it has grown to add on other disease challenges beyond just yellow fever viruses. So currently UVRI's mandate involves investigations of communicable but now also increasingly non-communicable diseases, obviously to help generate information that will initiate or improve on existing control and prevention strategies for the government. I can say that at the institute we have a vibrant research community with a particular focus on HIV. UVRI is also active in other viral infections (such as measles, human papilloma viruses. Also looking at emerging diseases like avian flu, Ebola) and there is also medical entomology. We also look at effects of HIV co-infections with other things, like worms, on vaccines response and disease incidence and pathogenesis. One can say that the institute hosts various regional and national laboratories like for HIV and influenza, among others. We participate in regional networks and we have numerous international collaborations. Among these we have long standing collaborations with the US CDC, WHO, the UK Medical Research Centre and others making UVRI really a big research institution with a lot of activities and lots of partners involved.

Wisdom: How did you get into bioinformatics?

Dr. Kayondo: Well for me it was actually out of necessity. So when I was doing my graduate research I needed it to support that research in terms of better determination of the biological significance of my data because newer approaches using whole genome were really coming up. I also needed expertise to organize and manage the huge data that was coming out of it and also acquire practical tools to mine emerging sequence data (lots of reference genomes were being churned out at that time) for new information. I realized bioinformatics was something I needed to learn or become more familiar with. That time bioinformatics was not distinctively developed/specialized as it is now. It was a course in a multi-disciplinary programming that I was taking, but I have over the years consolidated it as I have worked, through attended workshops and trainings that have come along. Beyond that I have also looked for opportunities and networking activities that involve bioinformatics.

Wisdom: Dr.! You have been involved in a number of research projects at your institute, across Africa and the world. You have also made some publications. Please can you tell me what your research interests are?

Dr. Kayondo: Ok. My background is genetics, so it is things all related to genetics. Specifically I am interested in better understanding disease vectors and pathogens and behind that is the interest to develop novel methods or tools for vector and pathogen detection and eventually control and also the underlying basic research that goes with getting there. So towards that I have carried out malaria vector research examining genomic and population structure. I have looked at molecular-based species diagnostics and also host-seeking behavior in the main malaria vector in Uganda. I have also studied HIV drug resistance and its evolution and I am really helping out surveillance programs with molecular and bioinformatics-based pathogen detection approaches because the routine methods that are used, the serology, also have limitations. Sometimes you get cases during outbreak attacks with classic haemorrhagic fever symptoms but then you run the usual Elisa's or serology pathogen-specific tests and don't get a positive result/diagnosis, in spite of the fact that the patient was sick. Then you ask what is going on? Bioinformatics approaches are enabling researchers now to use, for example, metagenomics to try and answer these type of broad questions. So all those things above interest me and I am also involved in different capacity building networks at UVRI and one is H3ABioNet where we are building bioinformatics infrastructure and expertise here.

Continue the conversation:

#bioinformatics #Africa
#H3ABioNet @H3ABionet

Issue 22: June 2017

H3ABioNet

H3ABioNet 'Meet the PI' Interview

Wisdom: I can say everybody has some aspects of his or her work that he or she enjoys most. Can you tell me what you enjoy most about your job?

Dr. Kayondo: I will say fundamentally I really enjoy basic research because you really will never know where the quest will take you. It seems to create or lead to new angles as if the questions are never quite completely answered, for as you answer one then you see a different intriguing perspective and then you follow that up as well. So that really is very interesting for me. It is my joy and my job gives me the opportunity to do that. That motivates me to wake up every morning. Then also my job enables me to work with people from different academic and cultural background to solve a shared problem which is exciting. Lastly there us also the aspect of working with students and mentoring others which is all very satisfying.

Wisdom: Is there any aspect that you enjoy the least?

Dr. Kayondo: Oh yes definitely (laughs). Because my work is basic research it is almost like we are working at the frontiers. I hate troubleshooting experiments which sometimes can take months without making any headway and that is really very frustrating. Also in the environment where we operate, procurement is another challenge; lots research of supplies and equipment are manufactured from elsewhere and sometimes getting them is a very protracted process and months can be wasted. I also don't like that.

Wisdom: Sir!, I understand you have a research group that you are working with, right? How has your association with the H3ABioNet consortium impacted on your research group?

Dr. Kayondo: It has been a good fit for us, and impact has been very positive. When the H3ABioNet call was made for the very first time, even I was looking out for groups to partner with and really looked up to this opportunity. A key focus of H3ABioNet was to develop and strengthen bioinformatics at partner institutions. Our strength, I will currently rate it as developing. We are still ways to go but we knew we needed it and we knew where to get started from. For us the benefits have been quite tangible in terms of the upgrades in the computing infrastructure and environment that we have had. We have secured and installed high-end analysis server and software courtesy of H3ABioNet. We have also received technical support from the consortium at different times during the course of networking. Some of our members have attended consortium courses for basic bioinformatics training. We have added on new competencies on the site that we can now run and there have been prospects of increased collaboration as a result of our membership in the network. Also the acquired infrastructure and networking has enabled attraction of more projects and leverage of additional funding. So it has been fruitful.

Wisdom: I have realized from this interaction that you are an accomplished fellow in terms of research with an established research group that does work in bioinformatics. What advice will you give to young persons who would want to take bioinformatics as a career?

Dr. Kayondo: I will say that at the moment, we are really witnessing technological advances that are driving genome sequencing to be relatively affordable and widely applicable to different medical scientific questions. Bioinformatics, which is still a developing field in Africa, is very crucial towards optimizing the benefits from these ongoing genetic revolutions. In my view I see nothing likely to make bioinformatics obsolete in the next 15 to 20 years. So I can say it can be a rewarding career choice. In Africa we say that bioinformatics is more in the academic research arena, but there is really nothing to stop it from spreading to other sectors such as may be home-grown biotech if that industry picks up. So I will say it will be a valuable choice or career direction to consider.

Continue the conversation:

#bioinformatics #Africa
#H3ABioNet @H3ABionet

H3ABioNet ’Meet the PI’ Interview

Continue the conversation:

#bioinformatics #Africa
#H3ABioNet @H3ABionet

Final words...

Dr. Kayondo: My parting word would be that H3ABioNet is really a first of its kind in Africa. I think before that there was nothing like this which enabled the African agenda to be the one driving the development of things. So this is really good and it looked appealing for us to be part of it right at its inception. So it is a valued collaboration here at UVRI that has helped us built capacity and also strengthened south-south linkages because that is also an area where there are a lot of gaps. I think before H3ABioNet, we had not really worked closely with, for example, West African institutions in any meaningful way. So this is a good opportunity. Now we are pooling resources, expertise and experiences together. So this is really very good and for me and I think this mode of collaborative capacity building has really worked well for us.

Wisdom: Thank you so very much for this opportunity. On behalf of the team that is coordinating this interview, I would like to thank you so much for making time out of your busy schedule for this interaction. I wish you a great time in your research activities.

#MeetthePI

Wisdom A. Akurugu and Jonathan Kayondo

Back to Contents

Issue 22: June 2017

H3ABioNet

H3Africa Genotyping Array

H3Africa Genotyping Array

Designing the H3Africa genotyping array: the journey and its challenges

The H3Africa chip is currently being manufactured. It took months of hard work to get it to this point. The work from the technical team was completed in early May, when the final bead pool was sent to Illumina. Our evaluation of the new chip demonstrates that, on paper, the design should outperform current chips on the market for genotyping African populations. It is now in the hands of Illumina to get the manufacturing done as fast and as possible, while ensuring a high quality product. Here we describe the design process and some of the challenges we faced.

Data preparation and processing

Almost 5700 high and low coverage full genome samples were used in the design of the chip. The samples came from public sources such as GDAP (Genome Diversity in Africa Project), AGVP (African Genome Variation Project), and 1000G (Thousand Genomes Project) as well as from private projects such as UG2G (Uganda genome project), SAHGP (Southern African Human Genome Programme), and TrypanoGEN. There were also 348 samples specifically selected for sequencing at Baylor by the H3Africa consortium to cover missing populations not provided by the other sources. The processing of the Baylor sequence data was done at NCSA Blue Waters, which took around 600K node hours to run through a pipeline very similar to GATK's best practices. All downstream steps (reference alignment to variant calling) were processed in a homogeneous manner to allow for better combination with other sources downstream. The CBIO team were responsible for the processing of the Baylor, SAHGP and TrypanoGEN data (medium to high coverage data) up to variant calling and the Sanger team was responsible for the calling of the GDAP, AGVP, 1000G and UG2G data (low coverage data). Merging of the CBIO and Sanger sets and tag SNP selection and evaluation were done by the H3ABioNet team (at CBIO and Wits).

Chip design

An ideal genotyping chip should contain representative SNPs across the genome, relevant to the populations of interest, that tag other SNPs for efficient imputation from a reference panel. A decision was made to work with Illumina to develop a 2.5 million marker array, of which 1.8 million were from existing arrays and 700,000 could be custom SNPs. The custom content had to include desired SNPs from sources such as; ClinVar and the GWAS Catalog, SNPs of particular interest requested by H3Africa projects, and novel SNPs identified in the sequence data. The design team had to work in parallel to: 1) select the best combinations of bead pools (SNP sets) from existing chips available to us that were appropriate for African populations, and 2) merge requested content with efficient tagging SNPs that, when combined with the bead pool SNPs, produced a set of non-overlapping SNPs with good coverage and imputation potential. The sequence data was used for linkage disequilibrium (LD) analysis and to enable prioritization of SNPs found at >1% frequency in African populations. The two processes were run through several iterations of selection, evaluation and refinement.

SNP list evaluation and refinement

We used both imputation-based (coverage and imputation accuracy) and imputation-free approaches (parameters like LD efficiency and genomic coverage) to evaluate the extent of overall genomic variation captured by the SNP set in various African populations. The results from these evaluations were used to optimize the SNP set to capture the genomic diversity across all the African populations.

Continue the conversation:

#bioinformatics #Africa
#H3ABioNet @H3ABionet

Issue 22: June 2017

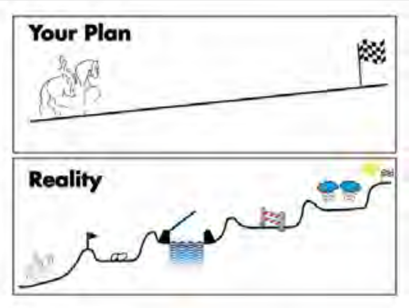
H3Africa Genotyping Array

Similarly, as the density of SNPs as well as LD architecture also varies widely across the genome it was necessary to optimize the SNP set to perform well across the entire genome. Therefore, we used the same approaches to evaluate the performance of the SNP set in each 1Mb genomic region and thereby identify regions that were not covered adequately. The findings from these analyses were used to guide to the SNP selection algorithm to boost the SNP content in the underrepresented regions.

The SNP set for inclusion in the chip was selected after several iterations of population and genomic-region based evaluations and optimizations. The performance of the final SNP set selected for inclusion into the chip was then compared to currently available SNP chips of similar size (Omni 2.5M, MEGA and Affymetrix Pan African Chip) in many African and non-African populations. The results clearly demonstrate that by virtue of being designed on the basis of an unprecedented amount of African genomic data and efficient SNP selection algorithms, the proposed H3Africa chip is expected to perform significantly better in terms of LD efficiency, coverage and imputation accuracy than its competitors.

Challenges

We faced many challenges during the design of the chip, where do we start? The first was delays in reagents that delayed sequencing of the Baylor set. Though not all Baylor raw and processed data was needed for the tag selection process, the data had to be transferred from the US back to Cape Town and managing this process took many man hours from the team. Transfer of the full Baylor dataset to South Africa eventually took over 9 months. From the Sanger side, working with a large low coverage dataset was challenging and joint calling and phasing took months to complete despite their access to excellent compute resources. Processing of such large datasets is just very time consuming!



The main dataset formed by merging the two largest WGS datasets, to date (CBIO and Sanger), using a technique known as cross-imputation, had issues due to missing sites and known challenges in merging low and high coverage data. Therefore, we had to revise the dataset and restart the chip design in December 2016.

The assortment of bead pools available to us for the base content was abruptly reduced from 180 relatively small pools to 30 relatively big ones in January 2017, thereby changing the entire selection of bead pool and custom content, again requiring a restart. Around the same time we also learnt that in Illumina's genotyping technology, a single SNP might require one, two or more (even up to 8) beads, depending on various factors. We needed to design the chip based on 2.5 M beads and not 2.5 M SNPs, which introduced further design limitations. Additionally, there was some overlap in the SNPs within different bead pools, so even with optimization we could only get 1.56 M unique SNPs in the 1.8 M fixed content. As a result, after cordial negotiations, our custom content allowance increased to about 0.89 M custom beads. However, at that point we still needed to drop some SNPs or replace multiple bead SNPs with a single bead SNP, wherever possible.

Continue the conversation:

#bioinformatics #Africa
#H3ABioNet @H3ABionet

H3Africa Genotyping Array

Continue the conversation:

#bioinformatics #Africa
#H3ABioNet @H3ABionet

Just when we thought we had a final product we learnt about the next constraint that is imposed by the technology, which is that it can only accurately genotype a single SNP within a 55 bp region. Though we can have up to three SNPs in a 55 bp window, no two SNPs should be closer than 10 bps. With many requested SNPs from H3Africa projects concentrated in very few genes, this meant that a significant number of SNPs had to be dropped or replaced. Also, the manufacture process has an 80-95% success rate (i.e. up to 20% of the beads may fail), so we were recommended to duplicate the "must have" SNPs to increase their chances of success. This affected the total number of unique SNPs. For every iteration of the SNP list, i.e. for every change, the rest of the list was impacted and had to be regenerated and re-evaluated, which consists of imputation and other tests. By this time we had had a lot of practice in evaluation! Finally, by April the majority of our custom bead pools were submitted for manufacturing.

The final product

The final product consists of 2,397,996 unique SNPs, 1,561,404 in the base content (existing bead pools), and 862,235 custom content. The chip contains 155,027 exonic SNPs, a few hundred ancestry informative markers, nearly 18,000 SNPs in the MHC region, and over 70,000 SNPs from curated databases such as PharmGKB, ClinVar, COSMIC and the GWAS catalogue. In all cases SNPs were prioritised for African populations. Based on our evaluations we believe this chip will improve the power to do accurate GWAS experiments on African populations. We look forward to seeing the results of the final product later this year.

The H3Africa genotyping array has garnered much publicity as can be seen *here* and *here*.

#H3AfricaChip

Gerrit Botha, Ananyo Choudhury, Scott Hazelhurst, Ayton Meintjes and Nicky Mulder (on behalf of the chip design team, which also includes Emile Chimusa and Mamana Mbiyavanga. Other groups were involved at various stage of the process, too.)

Announcements

Continue the conversation:

#bioinformatics #Africa
#H3ABioNet @H3ABionet

Announcements

- Congratulations to Nicky Mulder (central node) and Shakuntala Baichoo (University of Maritius node) for contributing to an article on containerization in *Nature* entitled *Software Simplified*

- Congratulations to the Node Accreditation Taskforce on their recent publication in *PLOS Computational Biology* entitled *Assessing computational genomics skills: Our experience in the H3ABioNet African bioinformatics network* (*doi: https://doi.org/10.1371/journal.pcbi.1005419*)

- Congratulations to Julie Makani from the MUHAS Node on her recent publication in *The Lancet* entitled *Sickle cell disease: tipping the balance of genomic research to catalyse discoveries in Africa* (*doi: http://dx.doi.org/10.1016/S0140-6736(17)31615-X*). H3ABioNet is explicitly mentioned in this article.

- Congratulations to Oussama Souiai, Mariem Hanachi, and Alia Benkahla from the IPT Node for receiving the Fondation Mérieux fellowship grant award for their project entitled *Impact of delivery mode on bacterial and phage community in gut microbiota of Tunisian Newborns*.

Do you have an ANNOUNCEMENT for upcoming editions of the H3ABioNet newsletter or for the H3ABioNet social media pages?

Tell us about your anouncements *here*

Back to Contents

H3ABioNet

Upcoming Events

Upcoming Events

- **July 19th to July 21st 2017**: *NextComp 2017: Next Generation Computing Application Conference*, will take place in Republic of Mauritius.

- **July 21st to July 25th 2017**: *International Society for Computational Biology conference 2017 (ISCB ECCB 2017)*, will take place in Prague, Czech Republic.

- **August 13th to August 16th 2017**: *17th Biennial Congress of the Southern African Society for Human Genetics*, will take place in Durban, South Africa.

- **August 20th to August 23rd 2017**: *The 8th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB)*, will take place in Boston, MA, USA.

  This year, the *Workshop on Algorithms in Bioinformatics (WABI)* will be co-located with ACM-BCB.

- **October 9th to October 10th 2017**: H3ABioNet SAB and AGM meeting, will take place in Entebbe, Uganda.

- **October 10th to October 12th 2017**: *ISCB Africa ASBCB Conference on Bioinformatics 2017*, will take place in Entebbe, Uganda.

- **Every third week of every month**: *CPGR Foundation in Genomics Course*, from standard molecular technologies to advanced 'omics' application in 3 days, aimed at scientists who are new to 'omics' as well as researchers interested in an overview of a dynamically evolving field.

- For a comprehensive list of bioinformatics and genomics conferences, please consult: *Conference service - Bioinformatics*

Do you have an EVENT for upcoming editions of the H3ABioNet newsletter or for the H3ABioNet social media pages?

Tell us about your events *here*

Back to Contents

Continue the conversation:

#bioinformatics #Africa
#H3ABioNet @H3ABionet

Issue 22: June 2017

H3ABioNet

Upcoming H3ABioNet working group meeting schedule

Upcoming H3ABioNet working group meeting schedule*

*Schedule until end August 2017

Summary of H3ABioNet upcoming working group meetings

Month	Date	Day	Working Group (WG)	Time (UTC)
July	14th	Friday	User Support WG	9:00
July	14th	Friday	Research WG	13:00
July	18th	Tuesday	Education and Training WG	11:00
July	21st	Friday	User Support WG	13:00
July	28th	Friday	User Support WG	9:00
August	11th	Friday	Research WG	13:00
August	11th	Friday	User Support WG	9:00
August	15th	Tuesday	Education and Training WG	11:00
August	18th	Friday	Infrastructure WG	13:00
August	25th	Friday	User Support WG	9:00

Timezone conversions to UTC for all H3ABioNet working group meetings

UTC Time Offset	Time Zone Name	Region/ Country in the Time Zone offset
-6 hours	CDT	Chicago, USA
0 hours	GMT	Burkina Faso, Ghana, Mali, Morocco, Senegal
+1 hour	WAT	Cameroon, Chad, Gabon, Namibia, Nigeria, Niger, Tunisia
+2 hours	CAT	Botswana, Egypt, Malawi, South Africa, Sudan, Zambia
+3 hours	WAT	Ethiopia, Kenya,Tanzania, Uganda

This edition of the newsletter was compiled and edited by Kim Gurwitz. For any corrections, please contact Kim at kim.gurwitz@uct.ac.za

Back to Contents

Continue the conversation:

#bioinformatics #Africa
#H3ABioNet @H3ABionet

Please scroll down for the latest issue of BioRes Digest

BioRes Digest

H3ABioNet Bioinformatics Research Digest | April - June 2017 | Issue: 02

In This Issue

Follow us:

#H3ABioNetResearch
#H3ABioNet @H3ABionet

Foreword

The H3ABioNet Research Working Group publishes a quarterly Bioinformatics Research Digest on the latest bioinformatics topics relevant to H3Africa projects. The main focus is to brief the H3ABioNet and H3Africa community on the latest articles and topics in bioinformatics and computational biology.

Main Article Summary: Guidelines for filtering human whole exome and genome variant calls to prioritize candidates in disease genetics research

By Junaid Gamieldien, Associate Professor at South African National Bioinformatics Institute (SANBI)

Article Information

Original Article Title: A practical guide to filtering and prioritizing genetic variants
Original Article Author/s: Mahjoubeh Jalali, Sefid Dashti, and Junaid Gamieldien
Original Article Journal: BioTechniques (PubMed: *28118812* , *doi: 10.2144/000114492*)

Background

While variant calling has significantly improved since the early days of next generation sequencing, guidelines for prioritizing variants that may be disease-linked are still largely speculative. There is no single "recipe" for identifying these variants and several approaches that don’t take study-specific biological realities into account may end up throwing away truly relevant variants or call others function-impacting when they are "benign". The article makes the case that functional impact prediction tools have their limits and provides support for this statement with a study that showed that as much as 50% of de novo (private) and rare missense mutations predicted to be deleterious were found to be nearly neutral mutations in functional genomics studies. This is proposed as an indicator of the need for guidelines rather than set-in-stone "recipes", since the former allows "biology aware" filtering pipelines to be improved and adjusted as the resources and tools they depend on improve themselves, or on a per-study basis.

BioRes Digest

H3ABioNet

Main Article Summary

Objective/s

The aim of the study was to propose a set of research use only guidelines, software tools, and online resources to assist in identifying functional variants from whole genome and exome variant calls and prioritizing those potentially associated with a phenotype of interest.

The guide was split into two parts. The first provided an overview of strategies and tools that can be used to enrich for variants that may affect protein function and produce a phenotypic effect. The second illustrated the use of existing biological and biomedical knowledge to prioritize genes bearing these potentially functional variants, in order to link the candidate variant to the disease being studied.

SECTION A

This section focused on the step following variant calling; identifying potentially function-impacting variants that either fit a disease inheritance pattern or statistically segregate with cases, etc. The importance of using an appropriate reference transcript-set in variant annotation was emphasized and the important paper by McCarthy et al. (2014), which showed a significant incongruence between annotations produced with Refseq vs Ensembl transcripts, was cited. The Ensembl set was recommended, as the former has been shown to miss a large number of high-impact variants. Similarly, Ensembl's Variant Effect Predictor (VEP) was recommended over ANNOVAR for variant annotation, as the latter often misannotated variant classes (eg. synonymous instead of frameshift), even when Ensembl transcripts were used.

Filtering variants based on statistical overrepresentation in "cases vs controls" or using pedigrees/trios, etc. was also recommended to reduce the number of variants that have potential to be functionally relevant (provided that incomplete penetrance is taken into account so as not to discard truly causal variants). It was suggested to filter remaining candidates according to their class, prioritizing frame-shifting, nonsense, and splice site variants, but bearing in mind the higher likelihood that these may be sequencing errors. Sanger sequencing validation was thus recommended as the final criterion, even after further *in-silico* checks were made.

Rarity of previously-reported variants was also discussed as a predictor of their likely functionality and several public repositories were discussed. Using disease-specific carrier frequencies where available, instead of an arbitrary frequency (e.g. <1%) as a cut-off, was suggested. It was also pointed out that calls made by functional prediction tools should not be used as a primary criterion to determine the likely effect of a nonsynonymous varianton protein function, since novelty/rarity plus expected inheritance pattern plus conservation is already substantial evidence to implicate a candidate. In addition, the preference for using RadialSVM and LR(Dong et al. 2015)over a combination lower precision tools such as SIFT and PolyPhenfor variant impact prediction was highlighted. For known non-coding variants, a combination of RegulomeDB and GTEx were suggested to assess their likely functional effect on gene expression regulation, while for novel variants, multiple prediction algorithms were presented.

SECTION B: Knowledge-Driven Variant Prioritisation

This section aimed to demonstrate the utility in using existing knowledge to either implicate or filter out candidate variants identified in Section A. Multiple, open-access databases and associated query tools were presented to enable a variant curator to accumulate evidence necessary to implicate a gene and thus the candidate variant/mutation. In addition to using enrichment approaches to evaluate gene sets (in for example, multi-genic disorders) strategies for evaluating individual gene candidates were also outlined. It was proposed that the following questions should be asked to evaluate the genes in which the variants are found (or proximal to), illustrating the importance of considering the biology of the disease in evaluating candidates.

Follow us:

#H3ABioNetResearch
#H3ABioNet @H3ABionet

Important questions - Is the gene:

- Known to be involved the disease or a related disease?
- Involved with a function or pathway that coincides with the disease pathology, biochemistry, etc.? (e.g. inflammation in an inflammatory disorder)
- Related to a phenotype relevant to the disease? (model organism knock-out or known human "symptom")
- Expressed in the affected tissue or a related one?
- Associated with an intolerant of loss-of-function mutation?
- Does the genes encoded protein product interact with a protein known to be involved in the disease?

A recent pedigree-based WES study, which ambiguously identified MEF2A as a candidate disease gene for premature coronary artery disease, was used to illustrate the utility of using a knowledge-driven approach to gather evidence, as well as to identify potential molecular pathogenesis mechanisms. More "incriminating" evidence against MEF2A was gathered than was offered in the original study, clearly showing the strength of the proposed prioritization guidelines. The example simultaneously also showed the value of a pedigree based approach, as described in Section A, to focus-in on candidates for further evaluation.

Conclusions

The study concluded that rarity/novelty, likely deleteriousness, and segregation with affected/cases and unaffected/control individuals should be the primary criteria for selecting candidate variants. Only then should biological and biomedical contextualization be performed- and importantly, absence of prior evidence should not be used to automatically disqualify a strong candidate. The combined approach was described as a funnel made up of multiple filtering steps outlined in the guide, each selected and customized to the study requirements, i.e. there is no "one size that fits all" solution in candidate variant prioritization.

Latest Articles

1. Alberto Magi; Roberto Semeraro; Alessandra Mingrino; Betti Giusti; Romina D'Aurizio. **Nanopore sequencing data analysis: state of the art, applications and challenges**. Brief Bioinform2017 bbx062. *doi: https://doi.org/10.1093/bib/bbx062*.

Abstract

"The nanopore sequencing process is based on the transit of a DNA molecule through a nanoscopic pore, and since the 90s is considered as one of the most promising approaches to detect polymeric molecules. In 2014, Oxford Nanopore Technologies (ONT) launched a beta-testing program that supplied the scientific community with the first prototype of a nanopore sequencer: theMinION. Thanks to this program, several research groups had the opportunity to evaluate the performance of this novel instrument and develop novel computational approaches for analyzing this new generation of data.

Despite the short period of time from the release of the MinION, a large number of algorithms and tools have been developed for base calling, data handling, read mapping, de novo assembly and variant discovery. Here, we face the main computational challenges related to the analysis of nanopore data, and we carry out a comprehensive and up-to-date survey of the algorithmic solutions adopted by the bioinformatic community comparing performance and reporting limits and advantages of using this new generation of sequences for genomic analyses.

Our analyses demonstrate that the use of nanopore data dramatically improves the de novo assembly of genomes and allows for the exploration of structural variants with an unprecedented accuracy and resolution. However, despite the impressive improvements reached by ONT in the past 2 years, the use of these data for small-variant calling is still challenging, and at present, it needs to be coupled with complementary short sequences for mitigating the intrinsic biases of nanopore sequencing technology."

2. Chung-I Li, David C. Samuels, Ying-Yong Zhao, Yu Shyr, Yan Guo. **Power and sample size calculations for high-throughput sequencing-based experiments**. Brief Bioinform 2017 bbx061. *doi: https://doi.org/10.1093/bib/bbx061*.

Abstract

"Power/sample size (power) analysis estimates the likelihood of successfully finding the statistical significance in a data set. There has been a growing recognition of the importance of power analysis in the proper design of experiments. Power analysis is complex, yet necessary for the success of large studies. It is important to design a study that produces statistically accurate and reliable results. Power computation methods have been well established for both microarray-based gene expression studies and genotyping microarray-based genome-wide association studies. High-throughput sequencing (HTS) has greatly enhanced our ability to conduct biomedical studies at the highest possible resolution (per nucleotide). However, the complexity of power computations is much greater for sequencing data than for the simpler genotyping array data. Research on methods of power computations for HTS-based studies has been recently conducted but is not yet well known or widely used. In this article, we describe the power computation methods that are currently available for a range of HTS-based studies, including DNA sequencing, RNA-sequencing, microbiome sequencing and chromatin immunoprecipitation sequencing. Most importantly, we review the methods of power analysis for several types of sequencing data and guide the reader to the relevant methods for each data type."

3. Silvia Bottini, David Pratella, Valerie Grandjean, Emanuela Repetto, Michele Trabucchi. **Recent computational developments on CLIP-seq data analysis and microRNA targeting implications**. Brief Bioinform 2017 bbx063. *doi: https://doi.org/10.1093/bib/bbx063*.

Abstract

"Cross-Linking ImmunoPrecipitation associated to high-throughput sequencing (CLIP-seq) is a technique used to identify RNA directly bound to RNA-binding proteins across the entire transcriptome in cell or tissue samples. Recent technological and computational advances permit the analysis of many CLIP-seq samples simultaneously, allowing us to reveal the comprehensive network of RNA-protein interaction and to integrate it to other genome-wide analyses. Therefore, the design and quality management of the CLIP-seq analyses are of critical importance to extract clean and biological meaningful information from CLIP-seq experiments. The application of CLIP-seq technique to Argonaute 2 (Ago2) protein, the main component of the microRNA (miRNA)-induced silencing complex, reveals the direct binding sites of miRNAs, thus providing insightful information about the role played by miRNA(s). In this review, we summarize and discuss the most recent computational methods for CLIP-seq analysis, and discuss their impact on Ago2/miRNA-binding site identification and prediction with a regard toward human pathologies."

4. Massaia, A. & Xue, Y. **Human Y chromosome copy number variation in the next generation sequencing era and beyond**. Hum Genet (2017) 136: 591. *doi: 10.1007/s00439-017-1788-5*.

BioRes Digest

H3ABioNet

Latest Articles

Follow us:

#H3ABioNetResearch
#H3ABioNet @H3ABionet

Latest Articles

Abstract

"The human Y chromosome provides a fertile ground for structural rearrangements owing to its haploidy and high content of repeated sequences. The methodologies used for copy number variation (CNV) studies have developed over the years. Low-throughput techniques based on direct observation of rearrangements were developed early on, and are still used, often to complement array-based or sequencing approaches which have limited power in regions with high repeat content and specifically in the presence of long, identical repeats, such as those found in human sex chromosomes. Some specific rearrangements have been investigated for decades; because of their effects on fertility, or their outstanding evolutionary features, the interest in these has not diminished. However, following the flourishing of large-scale genomics, several studies have investigated CNVs across the whole chromosome. These studies sometimes employ data generated within large genomic projects such as the DDD study or the 1000 Genomes Project, and often survey large samples of healthy individuals without any prior selection. Novel technologies based on sequencing long molecules and combinations of technologies, promise to stimulate the study of Y-CNVs in the immediate future."

5. Halil Kilicoglu. **Biomedical text mining for research rigor and integrity: tasks, challenges, directions**. Brief Bioinform 2017 bbx057. *DOI: https://doi.org/10.1093/bib/bbx057*.

Abstract

"An estimated quarter of a trillion US dollars is invested in the biomedical research enterprise annually. There is growing alarm that a significant portion of this investment is wasted because of problems in reproducibility of research findings and in the rigor and integrity of research conduct and reporting. Recent years have seen a flurry of activities focusing on standardization and guideline development to enhance the reproducibility and rigor of biomedical research. Research activity is primarily communicated via textual artifacts, ranging from grant applications to journal publications. These artifacts can be both the source and the manifestation of practices leading to research waste. For example, an article may describe a poorly designed experiment, or the authors may reach conclusions not supported by the evidence presented. In this article, we pose the question of whether biomedical text mining techniques can assist the stakeholders in the biomedical research enterprise in doing their part toward enhancing research integrity and rigor. In particular, we identify four key areas in which text mining techniques can make a significant contribution: plagiarism/fraud detection, ensuring adherence to reporting guidelines, managing information overload and accurate citation/enhanced bibliometrics. We review the existing methods and tools for specific tasks, if they exist, or discuss relevant research that can provide guidance for future work. With the exponential increase in biomedical research output and the ability of text mining approaches to perform automatic tasks at large scale, we propose that such approaches can support tools that promote responsible research practices, providing significant benefits for the biomedical research enterprise."

Back to Contents for BioRes Digest

Back to H3ABioNet Newsletter: Back to Contents

Follow us:

#H3ABioNetResearch
#H3ABioNet @H3ABionet

Most Cited Articles

Follow us:

#H3ABioNetResearch
#H3ABioNet @H3ABionet

Most Cited Articles

’*Most Cited Articles*’ is updated monthly. Rankings are based on citations to online articles from *HighWire Press*.

1. Sudhir Kumar, Koichiro Tamura, Masatoshi Nei. **IMEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment**. Brief Bioinform 2004; 5 (2): 150-163. *doi: https://doi.org/10.1093/bib/5.2.150*.

Abstract

"With its theoretical basis firmly established in molecular evolutionary and population genetics, the comparative DNA and protein sequence analysis plays a central role in reconstructing the evolutionary histories of species and multigene families, estimating rates of molecular evolution, and inferring the nature and extent of selective forces shaping the evolution of genes and genomes. The scope of these investigations has now expanded greatly owing to the development of high-throughput sequencing techniques and novel statistical and computational methods. These methods require easy-to-use computer programs. One such effort has been to produce Molecular Evolutionary Genetics Analysis (MEGA) software, with its focus on facilitating the exploration and analysis of the DNA and protein sequence variation from an evolutionary perspective. Currently in its third major release, MEGA3 contains facilities for automatic and manual sequence alignment, web-based mining of databases, inference of the phylogenetic trees, estimation of evolutionary distances and testing evolutionary hypotheses. This paper provides an overview of the statistical methods, computational tools, and visual exploration modules for data input and the results obtainable in MEGA."

2. Heng Li & Nils Homer. **A survey of sequence alignment algorithms for next-generation sequencing**. Brief Bioinform 2010; 11 (5): 473-483. *doi: https://doi.org/10.1093/bib/bbq015*.

Abstract

"Rapidly evolving sequencing technologies produce data on an unparalleled scale. A central challenge to the analysis of this data is sequence alignment, whereby sequence reads must be compared to a reference. A wide variety of alignment algorithms and software have been subsequently developed over the past years. In this article, we will systematically review the current development of these algorithms and introduce their practical applications on different types of experimental data. We come to the conclusion that short-read alignment is no longer the bottleneck of data analyses. We also consider future development of alignment algorithms with respect to emerging long sequence reads and the prospect of cloud computing."

Key Points

- The advent of new sequencing technologies paves the way for various biological studies, most of which involves sequence alignment on an unparalleled scale.
- The development of alignment algorithms has been successful and short-read alignment against a single reference is no longer the bottleneck in data analyses.
- With increasing read lengths produced by new sequencing technologies, we expect further development in multi-reference alignment, long-read alignment, and de novo assembly.

Most Cited Articles

3. Kazutaka Katoh & Hiroyuki Toh. **Recent developments in the MAFFT multiple sequence alignment program**. Brief Bioinform 2008; 9 (4): 286-298. *doi: https://doi.org/10.1093/bib/bbn013*.

Abstract

"The accuracy and scalability of multiple sequence alignment (MSA) of DNAs and proteins have long been and are still important issues in bioinformatics. To rapidly construct a reasonable MSA, we developed the initial version of the MAFFT program in 2002. MSA software is now facing greater challenges in both scalability and accuracy than those of 5 years ago. As increasing amounts of sequence data are being generated by large-scale sequencing projects, scalability is now critical in many situations. The requirement of accuracy has also entered a new stage since the discovery of functional noncoding RNAs (ncRNAs); the secondary structure should be considered for constructing a high-quality alignment of distantly related ncRNAs. To deal with these problems, in 2007, we updated MAFFT to Version 6 with two new techniques: the PartTree algorithm and the Four-way consistency objective function. The former improved the scalability of progressive alignment and the latter improved the accuracy of ncRNA alignment. We review these and other techniques that MAFFT uses and suggest possible future directions of MSA software as a basis of comparative analyses. MAFFT is available at *http://align.bmr.kyushu-u.ac.jp/mafft/software/* .”

Key Points

- MAFFT Version 6 has two major new features; the PartTree algorithm for handling a large number (greater than 10 000) of sequences and the Four-way Consistency objective function for multiple structural alignment of ncRNAs.
- PartTree is a divisive recursive clustering algorithm with a time complexity of O(N log N). It is more scalable than the conventional UPGMA algorithm with a time complexity of O(N2). The PartTree option can create a large alignment composed of 60 000 sequences, at the cost of an accuracy loss of 2%.
- The X-INS-i-scarnapair - which is a combination of an external pairwise structural RNA alignmentmethod, SCARNA, and the Four-way Consistency objective function - is one of the most accurate methods for multiple RNA structural alignment. It requires less CPU time than other accurate structural alignment methods, such as RNA Sampler, MASTR, and Murlet.
- Two different types of group-to-group alignment methods - the profile alignment option and the seed option - were implemented, in order to deal with the various possible phylogenetic relationships between two groups. MAFFT Version 6 has L-INS-i and E-INS-i options, which are variants of G-INS-i, the iterative refinement method with WSP and consistency scores. L-INS-i allows large terminal gaps, while E-INS-i is applicable to a dataset with internal unalignable regions.

Book Review

Zhen Lin; **Bioinformatics Basics: Applications in Biological Science and Medicine**. Edited by Lukas K. Buehler and Hooman H. Rashidi. Brief Bioinform 2008; 9 (3): 256-257. *doi: https://doi.org/10.1093/bib/bbm060*.

"Bioinformatics has blossomed in the past decades with the introduction of biological experiments that rapidly produce massive amounts of data (such as the multiple genome projects, the large-scale analysis of gene expression, the large-scale analysis of protein-protein interactions and the large-scale analysis of genome-wide genotype-phenotype associations). The resulting bioinformatics tools mainly lie in two groups. They may be databases that hold and disseminate data, or they may be algorithms that draw inferences on the data. Bioinformatics is interdisciplinary by nature and its success relies on the close collaborations among researchers trained in traditional disciplines (such as biology, computer science and statistics) as well as contributions from scientists cross-trained in these disciplines.

Follow us:

#H3ABioNetResearch
#H3ABioNet @H3ABionet

Book Review

"The second edition of "Bioinformatics Basics: applications in biological science and medicine", published in 2005, covers a number of historically classic concepts in the field of bioinformatics. The book's chapters provide a general introduction of (i) biological information (ii) biological databases (iii) genome analysis (iv) proteome analysis and (v) the bioinformatics revolution in medicine. The authors start out briefly describing the function and structure of DNA, RNA, protein, and basic DNA cloning and sequencing techniques such as PCR; then move on enumerating early databases hosted at NCBI, EBI and DDBJ and other classic biological databases such as PDB and Prosite. A few well-used bioinformatics algorithms are briefly discussed in the book, including the ones for sequence alignment and motif finding such as BLAST. But some more recent algorithms such as the ones for microarray analysis and hydrodynamic methods in proteomic analysis have more in-depth descriptions. This edition also has updated topics including protein structure prediction, rational drug design and pharmacogenomics. Overall, the authors give a historical overview of bioinformatics from mostly a qualitative and descriptive point of view. Many sections of the book are organized in a question-and-answer style, providing short explanations to address concepts relevant to bioinformatics, e.g. "What are some of the services offered by EBI?", "How many genes are in a genome?" and "How are sequence alignments useful?". A handful of screenshots of online databases and analytical tools that are freely accessible to the public are included throughout the book, providing readers a quick glimpse on the available resources. The accompanying URLs thus allow those interested to retrieve the most current information on any particular websites. But the book has no coverage of technical principals behind building databases. It does not provide mathematical and programming details of the algorithms covered, nor includes many machine learning and statistical algorithms that are commonly used in bioinformatics applications, such as Hidden Markov Models and Bayesian statistics. Such information would be useful both for those who practice bioinformatics as well as users wishing for a principled approach to using the tools.

"The book may serve as a fine introductory book for computer science, statistics, or physics students who are interested in the field of bioinformatics. The basics from the book will provide them pointers to conduct further reading and research of the context. However, the book's content has become too "basic" given the state-of-the-art of bioinformatics and may no longer serve the original purpose of helping "the general scientific community in gaining a better understanding of what bioinformatics tools are available to them and how they could be incorporated into their projects or interests," as the authors say in the preface. Given the fast-paced advancements in life sciences, especially genomic research, the bioinformatics community has developed a wide range of new databases and algorithms in the past decades that go well beyond the scope of the book and may benefit the general scientific community greatly."

BioRes Digest Editorial Team

1. Editor in Chief: Prof. Faisal M. Fadlelmola, Chair of H3ABioNet Research Working Group

2. Editors:

- Dr. Amel Ghouila, Co-Chair of H3ABioNet Research Working Group
- Dr. Sumir Panji, H3ABioNet Network Manager
- Kim Gurwitz, Training and Communication Officer H3ABioNet

Follow us:

#H3ABioNetResearch
#H3ABioNet @H3ABionet

Back to Contents for BioRes Digest

Back to H3ABioNet Newsletter: Back to Contents