H3ABioNet Montly Newsletter Issue 4 / 5: August / September 2013

H3ABioNet

Foreword

This newsletter combines the activities for August and September, the first two months of year 2. On the training side, August saw the end of the first year courses, finishing off with the eBioKit course run at ICIPE in Kenya. At the same time, NABDA hosted a workshop on Visual Analytics of Human Genome Variation Datasets, which brought together participants from H3ABioNet Nodes and the Nigerian Biorepository to work on real research problems. Hopefully through all this training our nodes will soon be confident to undertake the node assessment exercises, which are now ready to implementation.

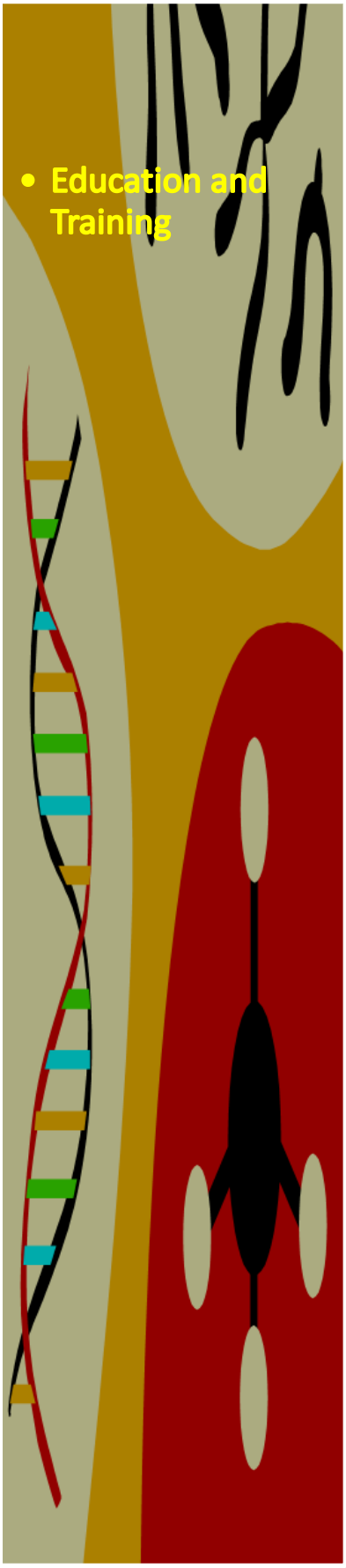
Meanwhile, for the development of infrastructure, many nodes worked on completing their equipment purchases and getting these set up in their home institutions. Their trained staff from the technical course (see previous newsletter) should be equipped to set up the computing infrastructure, and this group continues to stay in contact through their mailing list. The start of year 2 has seen the initiation of new inter-node research projects, which is encouraging new collaborations. This will be further enhanced regionally through the formation of the East African Genome Network / Biodata Analysis and other regional partnerships.

H3ABioNet central moved into newly renovated offices after squatting in an experimental laboratory! The group has been busy with the organisation of the second H3ABioNet consortium meeting as well as with the third H3Africa consortium meeting in Johannesburg. We have also engaged with the MRC Gambia, who are interested in joining H3ABioNet. We look forward to engaging with all the H3ABioNet PIs at the upcoming consortium meeting.

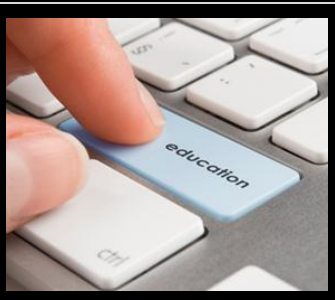
Prof. Nicky Mulder.

H3ABioNet Montly Newsletter Issue 4 / 5: August / September 2013

H3ABioNet



- Education and Training



&

The final workshop for the Year 1 period has been completed! The ToT and Introduction to Bioinformatics using the eBioKits mark the last set of training courses held by H3ABioNet for this period bringing the total to 5 workshops. The Introduction to Bioinformatics using the eBioKits was hosted at the International Center for Insect Physiology and Ecology (ICIPE) between the 29th of July to 2nd of August 2013 in Nairobi, Kenya.

The course comprised of a 5 day workshop covering topics such as introductory unix / linux, introduction to biological databases, usage of the SRS systems and emboss tools, introduction to NGS, meta-genomics and phylogenetics. In total, 22 participants from 11 African Countries were trained.

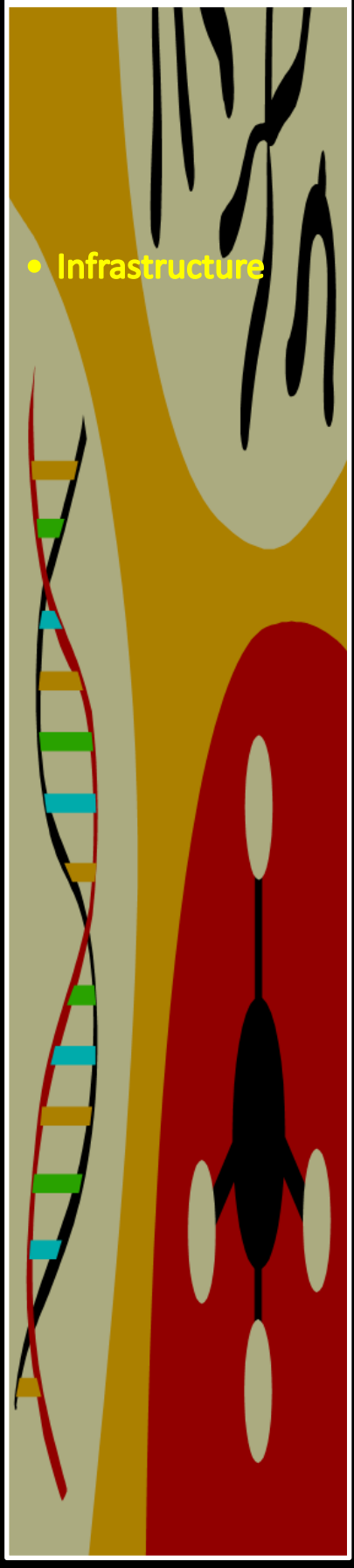
Simultaneously with the eBiokit training in Nairobi, the H3ABioNet NABDA Node in Abuja (Nigeria) held a training course on Visual Analytics of Human Genome Variation Datasets (29 July – August 2nd, 2013). The participants from H3ABioNet Nodes and Biorepository Project were introduced to the focus areas of visual analytics and learned visual analytics skills by working with datasets from catalogues on human genome variation including the 1000 genomes; the National Human Genome Research Institute (NHGRI) Genome-Wide Association Studies (GWAS) Catalog and the Database of Genomic Variations.

As these workshops have trained individuals with the goal of creating a cadre of African trainers able to provide training within the domains taught by the ToT, a follow up of the various workshop participants is underway with the aim of identifying potential future trainers and attempting to gauge what additional intervention would be required.

The workshops for this period (5 in total) were organised and held back to back in a short period of time. As a result Central has drafted a timeline guide for all the process and responsibilities for hosting a workshop ensuing that future planning of workshops shall proceed in a co-ordinated manner.

For the future set of workshops any Node that will want to host a workshop will be able to table a "bid to host a workshop" for which a policy is under development by Central and will be sent for comments to the E&TWG and then circulated amongst the General Assembly (GA) for comments. An "expression of interest" template is also being prepared by Central for which any Node that would like to have a workshop on H3Africa related topics can put in a submission motivating for the workshop, the lecturers, location and period to be hosted in.

Dr. Nash Oyekanmi.



The Infrastructure working group has been following up on the server purchases, coordination of Iperf tests, discussions as to the monitoring of infrastructure set up, estimation of H3Africa data sizes and the various bioinformatics applications and operating systems to recommend / support.

There will be a push to get the Iperf tests running as it is important to know how much bandwidth each Node has as opposed to what their ISP / ICT departments tell them and also to factor in the cost. This will help to motivate for better internet speeds for all the Nodes and enable Nodes to determine if they are overpaying for their internet connection. The ability to map out the various internet speeds available to the Nodes will be fundamental for the planning of H3Africa data storage infrastructure, analysis and movement of the data by the ISWG, so it is imperative that all the Nodes get on board with the Iperf tests.

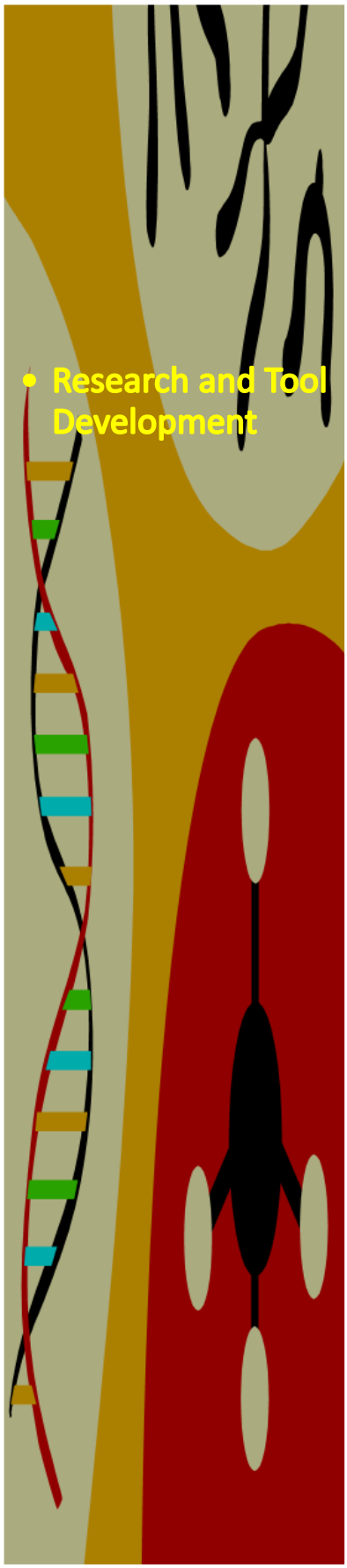
A lot of equipment is in the process of being procured by the various Nodes which are at various levels of development with regards to infrastructure. One of the recommendations by the ISWG is to have a monitoring system in place like Nagios, Zabbix, Collctd etc for each Node to get accustomed to best practices and allow them to manage their infrastructure effectively. Recommended monitoring software that can be supported will be compiled and recommended to the various Nodes by Frederick Mbuya. It is important for Nodes to start managing their infrastructure as early as possible as over time, the capacity of the Nodes will increase. Recommendations are also being developed to aid in software and hardware interoperability to enable all the Nodes to have a set of core bioinformatics applications on common platforms (Scientific Linux, Ubuntu and RedHat are all being mooted at present) for which support can be provided.

Questionnaires were sent by Prof. Hazelhurst to each of the H3Africa PIs in order to determine the number of samples to be collected, the data types / platforms will be used, where the data will be generated and analysed. The responses will enable the ISWG estimate how much storage space is needed and how much space for processing will be required e.g current estimates by Liudmila Mainzer from the University of Illinois indicate that 1 paired end exome sequence in FastQ format will take up to ~ 50GB of storage, hence 300 exomes will equate to ~15 Terabytes of storage. This does not include any intermediate files generated e.g 1 exome variant calling pipeline will need ~100GB so 300 exomes will require ~ 30 Terabytes. They are separate issues being explored; storing the data and analysis of the data, both which impact the data size parameters used for planning purposes.

Dr. Alia Benkahla. Prof. Scott Hazelhurst.

H3ABioNet Montly Newsletter Issue 4 / 5: August / September 2013

H3ABioNet



Research and Tool Development

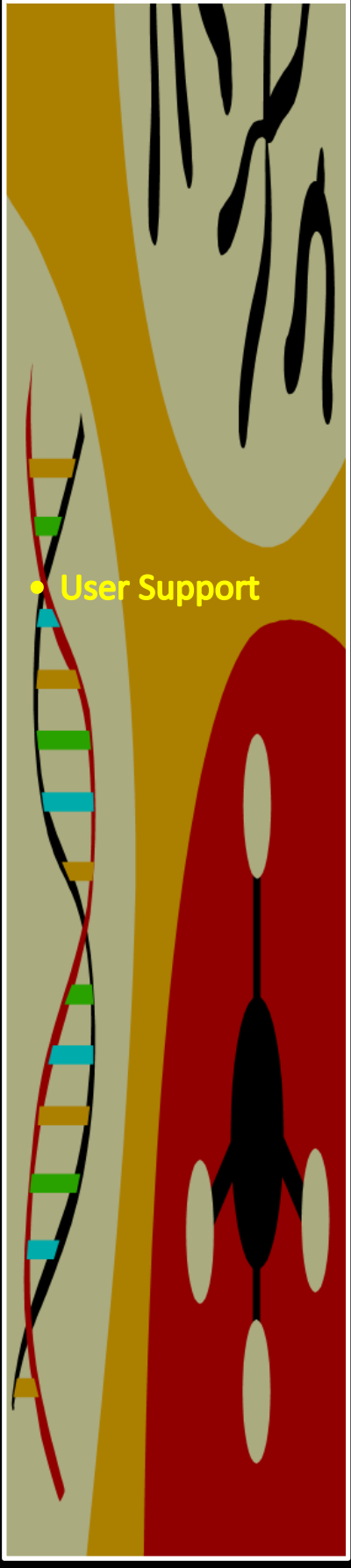
The emphasis of the Research and Tool Development working group has been to examine the working group's milestones and also the milestones for specific research projects that are reported to the NIH on. Various members of those specific research projects have been contacted with their current milestones and reporting period dates to add more detail and better estimated dates as the projects are progressing.

In addition to the specific research projects, they are some collaborations and projects of varying magnitude that are occurring. The RTDWG will be focusing on collecting the current status of these projects and will form a research task force to track the projects' progress.

An important aspect of the H3Africa projects and H3ABioNet will be the collection of the data, the management, storage and access provided as well as the final deposition to the European Genome – Phenome Archive (EGA). The RTDWG will work in conjunction with the ISWG and USWG to determine what the correct procedures and policies put in place should be and look at ways of facilitating the submission of the data. This will include co-opting personnel with some XML experience to help extract the annotation fields required and provide to the H3Africa PIs as metadata that should be collected when sampling.

As with all the working groups, the focus of the RTDWG will be the upcoming annual H3ABioNet meeting in October and presentation to the Scientific Advisory Board.

Dr. Julie Makani. Prof. Ezekiel Adebiyi.



User Support

The User Support Working Group has been reviewing the various sets of milestones assigned and determining how to facilitate them and which ones have synergies with other working groups.

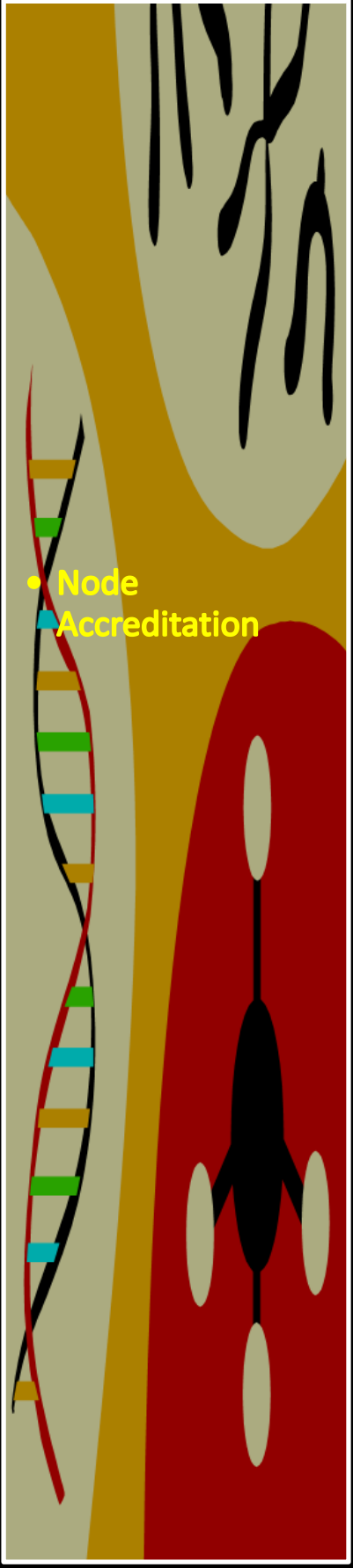
A main milestone is the integration of data and various solutions for data analysis. This milestone would be difficult to achieve as currently there is no H3Africa data to work with and investigate various solutions. The USWG does have access to software licences such as Ingenuity Pathway Analysis via the CPGR or Vidyo via the NABDA, but the USWG would need to determine what other tools would be required for data analysis / exploration. In terms of bioinformatics tools it is straight forward as the USWG, ISWG and NAWG have a good idea as to the actual tools required for the data generated by theH3Africa projects. However, it is difficult to determine what tools the H3Africa projects themselves would like and a question to address this will be included in a survey for the start of the next year.

A main issue has been the slow uptake of the helpdesk by the H3Africa and H3ABioNet nodes. It was decided upon further discussions that the helpdesk should be made publically available to all users whether registered or not. Hence the helpdesk dashboard and ability to submit a helpdesk ticket has been moved to the public pages of the H3ABioNet website and can be found at: http://www.h3abionet.org/support/help-desk?view=detail&cid=-1

It is hoped that this move will enable more helpdesk tickets to be submitted by various other groups struggling with their bioinformatics analysis. As always, the helpdesk is looking to increase its pool of experts and if you are interested in joining the helpdesk crew please do send an email to info@h3abionet.org or submit a ticket!

Dr. Judit Kumuthini. Dr. Jonathan Kayondo.

H3ABioNet



Node Accreditation

The NAWG is almost ready for the assessment exercise to begin. The final synthetic datasets that will be used for variant calling will be ready in time for the 2nd Annual H3ABioNet meeting along with some admixed datasets for the GWAS testing.

The Genotyping and GWAS assessment will consist of 2 parts:

1) Calling the actual genotypes on raw .cel files.
2) Doing the testing of significant SNPs using simulated data.

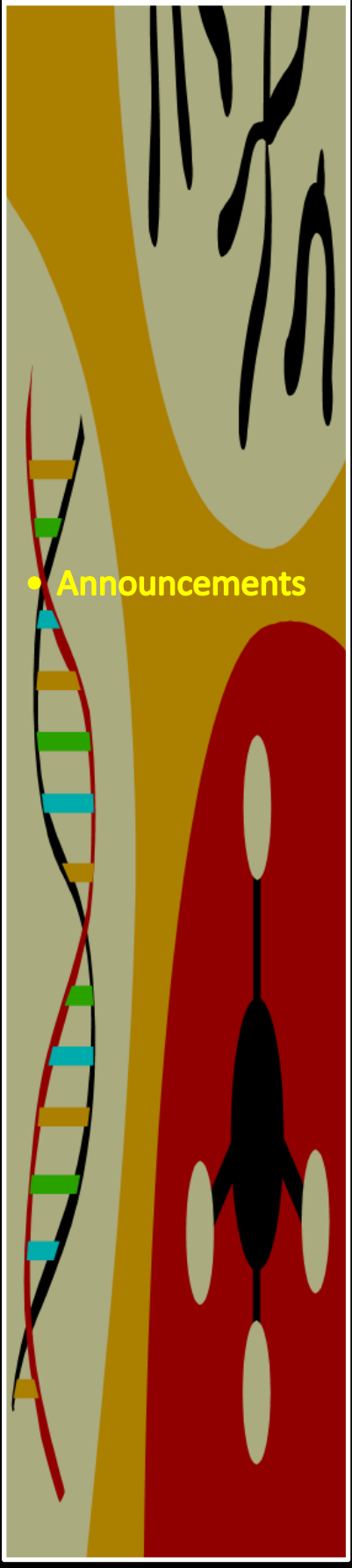
The first part of the assessment will utilise real Affymetrix SNP6 chips comprising of ~400 samples which will be randomly permutated so no Node receives the same batch of .cel file to do the genotyping calling. Nodes will be expected to do quality control checks and filtering of the datasets using Affeymetrix powertools as per the manual's instructions. Affymetrix powertools will also be used for the genotyping calls done with the current version implementing the birdseed algorithm for the actual calls. Once the calls have been performed, quality checks on these calls will also b expected to be implemented by the Nodes and the final results converted to PLINK format.

The second part of the Genotyping / GWAS exercise will involve the use of a synthetic dataset in PLINK format. The datasets will be created by using haplotype data for 88 Yoruba and 90 British individuals obtained from the 1000 genomes project. Case and control datasets will be simulated using HapGen whereby:

1) The physical position of the SNPs will be specified.
2) The risk allele will be specified.

The HapGen software will be used to decide the correct combination for the case and controls for simulating disease variants. The data will be randomised but linkage disequilibrium will be kept between the markers while maintaining the disease variants. The output will be a .gen file which has for each SNP created a probability associated with that SNP being a disease variant. This will be done separately for the 2 different populations with one population containing the disease SNPs. The admixed population will be created by simulating admixture for the 2 different populations and expanding the initial population to 2,500 via a random mating process using a custom programme created by Prof. Mulder's laboratory for ~ 10 – 50 generations. The resulting .ped file created containing 2,500 individuals will be run through HapGen once more to randomly simulate the disease SNPs, rare variants and assign cases and controls. For the actual GWAS testing the EMMAX programme (http://genome.sph.umich.edu/wiki/EMMAX) is being preferred to PLINK as previous studies by Prof. Mulder's laboratory indicate that EMMAX handles admixed populations much better then PLINK.

Dr. Victor Jongeneel.



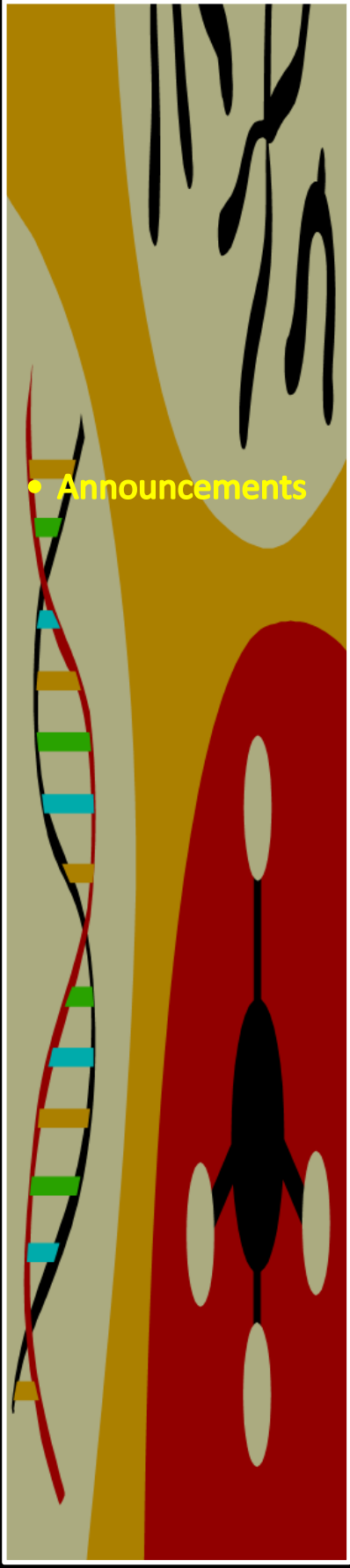
Announcements

East Africa Biodata Analysis Network

The formation of an East African Biodata Analysis has been driven by Dr. Julie Makani with the goal of creating a regional network in order to consolidate all the various networks, datasets, expertise and skills that are distributed in "pockets" all over East Africa. Most of the data and expertise lie within distinct projects and networks that are run by many various agencies and expertise tend to be focused within those programmes without any cross talk or collaborations between various members. This has created a situation whereby East African Scientists are not able to leverage the use of a co-ordinated effort such as H3ABioNet to apply for funding, approach their Governments or work with each other and diverse datasets to create high impact science due to the disjointed nature of these various programmes with regard to each other.

The East Africa Biodata Analysis seeks to redress this by getting the various research groups and scientists from the region to seek specialised expertise in biodata and bioinformatics analysis through the use of the vast repositories of data they have acquired or have scientists provide these skills to such groups. An inaugural meeting was held at the International Center for Insect Physiology and Ecology (ICIPE) which was attended by 16 different representatives from the various research projects being conducted within East Africa on the 30th of July, 2013.

Amongst the topics were presentations of the various research projects occurring within East Africa and an overview of H3ABioNet. Discussions focused on how to improve synergies between the various research projects and existing networks, an audit of all the various datasets that have been collected over the years and a general agreement that the vast amount of data requires big data analysis and bioinformatics skills to mine and integrate. Talks shifted as to how to increase the capacity of these skills within East Africa as with the rest of the world and Africa, there is a shortage of such skills within East Africa and problem of retention. A resolution was decided upon whereby each scientist attending will contact 5 other scientists they know working within their region to raise awareness of the East Africa Biodata Analysis network and determine how best to collaborate and share expertise.



Genomics in Tanzania: Excellence in the Silent Savannah: Society for International Development Report

Dr. Julie Makani was featured in the Society for International Development Greater Horn Outlook report whereby Dr. Makani raised awareness of the health, socio-economic and scientific importance of Genomics and Bioinformatics for the development of knowledge based economies. Dr. Makani highlighted the importance of genomics as one of the major trends that will have a long term impact on the region's future and talked about various research networks such as the Tanzanian genome Network and H3ABioNet. For a full report please look at:
http://www.sidint.net/docs/RF32_DevolutionLifeSciencesMobile.pdf

"The Society for International Development (SID) is a global network of individuals and institutions concerned with development, which is participative, pluralistic and sustainable. The Society was founded in Washington D.C. in 1957 and is based in Rome since 1978. Through its programmes and initiatives, organized in key centres of development policymaking in the North and in the South, SID plays a crucial role in promoting dialogue between various stakeholders and interest groups, both locally and internationally." - http://www.sidint.net/node/9852

Knowledge Transfer Program

Our first expert, Dr. Panu Somervuo, as part of H3ABioNet to train in the field of Pharmacogenomics and personalized medicine is starting on 2nd of Sep and expected to be at the CPGR for 3-4 months working on DMET data that was generated using African samples. This is opened to anyone from H3A consortia and priority will be given to H3ABionet staffs on first come first served basis.

All the registrations can be done on the current KTP site at http://ktp.cpgr.org.za Should you have any questions regarding registration please contact Judit Kumuthini of CPGR.

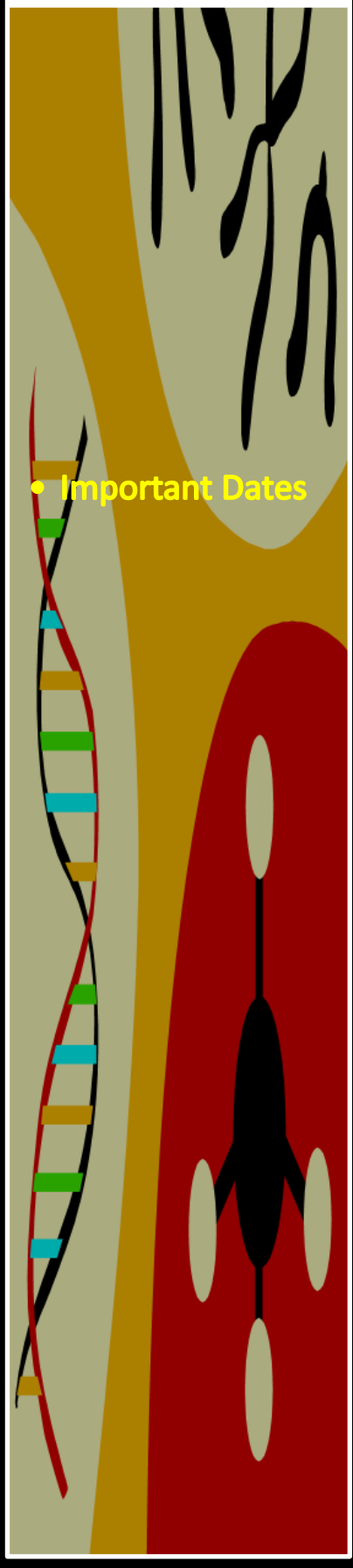
Publications

The Egyptian Node has published a paper:

Mohammed M. Saleh, Ahmed M. Alzohairy, Osama Abdo Mohamed, Gaber H. Alsayed (2013). A Comprehensive Study by Using Different Alignment Algorithms to Demonstrate the Genetic Evolution of Heat Shock Factor 1 (HSF1) in Different Eukaryotic Organisms. IRACST – Engineering Science and Technology: An International Journal (ESTIJ), ISSN: 2250-3498, Vol.3, No.2

The Moroccan Nodes has published a paper:

Hepatitis B in Moroccan blood donors: a decade trend of the HBsAg prevalence in a resources limited country
http://onlinelibrary.wiley.com/doi/10.1111/tme.12054/abstract)



Important Dates

- 20th September, 2013: deadline to fill in NetCapDB details – all Nodes
- 1st -3rd of October, 2013: H3ABioNet Annual Meeting, Johannesburg, South Africa
- 3rd - 6th of October, 2013: H3Africa Consortium Meeting, Johannesburg, South Africa
- 6th - 9th of October 2013: South African Society for Human Genetics Conference, Johannesburg, South Africa (http://www.sashg2013.co.za/)

21st – 23rd October, 2013: 9th International Meeting, Euro-Mediterranean Medical Informatics and Telemedicine (EMMIT), Nador, Morocco (http://www.msfteh.org/telemed2013/)