Bioinformatics Component in Personalized Medicine

Abhishek Narain Singh

ABI-O-TECH

abhishek.narain@cantab.net

Call to Action	Key Takeaways
■ Integrated approach of multi-omics ■ Hardware Software kinship ■ Security and Privacy of Data	■ Next Generation Sequence analysis ■ Proteomics and Genomics ■ High Performance Bio-Computing
Focus areas: technological, genomics, proteomics, HPC	

Abstract

Past few decades have seen rapid growth in sequencing technology and software tools to aid their processing and analysis. The cost of genome sequencing of whole human data has dropped to couple thousand dollars from what it used to be about a million dollars. In parallel there has been significant growth in supercomputing power as per the Moore's law with multi-and-many-core computers being a common commodity, needless to mention the GPUs which promise to bring supercomputing power at desktop space. Parallel advancement has been in the domain of proteomics and transcriptomics. The 'gaps' today are integrating these hardware, software and human resources for a better bioinformatics solution to aid a personalized medicine age to practice.

Introduction

Science has been progressing significantly in the past few decades in the area of biological studies which thus opens questions as of are we now more capable of understanding human health and be able to predict the causative factors for a disease. Though diseases have been more or less generically understood for the causative agents, what might be more interesting are primarily the diseases for which there can be multitude of factors that influence upon its activation or diseases for which different individuals respond with great variation in defense mechanisms. Whatever be the medical parameters and not so well understood complicated mechanisms involved, one thing is for sure that with the advent of our

capabilities of understanding and analyzing the genome, interactome, metabolome and proteome, we can definitely give more probabilistic predictive and personalized medical counseling, and be there the nuts and bolts for delivery, then perhaps personalized medicine too. In a way personalized medicine will differ from what hospitals and other healthcare services have been providing till now of personalized care, as with advent of more scientifically in-depth technology it would mean newer approaches to disease prevention, diagnosis by multitude of parameters, and the choices that an individual will have once recommendations are made. This can be a more effective reality when the interests of government, insurance companies, hospitals, scientists, technologists, education bodies, medical professionals and most important the patients are well aligned. We are now living in a data rich age, with capable technologies to extract relevant patterns for the case in hand. In particular the genomics area has been moving rapidly past one decade to give us a stronger faith in establishing a personalized medicine era by means of integrating with greater emphasis the personalized genomic medicine component.

Emerging Informatics Challenges with Genome Sequencing

In the area of sequencing genomes there has been rapid advancement in technology and simultaneous reduction in cost. Deodyribonucleic acid (DNA) is well known to be the blueprint of life. Dideoxynucleotide sequencing of DNA has improved from what it was in rudimentary stage to a large-scale production enterprise that requires devoted instrumentations, databases, bioinformatics tools and robotics. Tailor made bioinformatics tools has been significantly useful in answering our questions about mutation spectrum of an organism, from single nucleotide base to large copy number variations. The ability to process millions of sequence reads in parallel sets the next generation sequencing technology more popular. Further, in the process of its metamorphosis, the cost per reaction of DNA sequencing has fallen with a Moore's law precision [1]. The first human genome sequence was obtained by using Sanger sequencing method. In the past few years, the technology evolved to introduce paired end sequences , where the sequence can be determined at either end of a fragment and the insert size in between the ends can be approximately known a priory. The accuracy of this sequence or the quality of the information concerning the nucleotide bases is not always reliable as thus a probabilistic number is associated, however significant lower cost of this technology can allow multiple sequencing of the region of interest which is

also known as the coverage of sequencing, so as to then take the consensus at a region of interest to determine the sequence. A higher average coverage is usually preferred for more accuracy though bold steps were taken in projects such as to analyse genome with lower coverage such as the 1000 genomes project. Thus higher the coverage the more reliable the results are, and in the bioinformatics community it is generally accepted to have a coverage of 20x to almost saturate the possibility of having near zero false base consensus. The high false discovery rate of structural variation algorithms even in deeply sequenced individual genomes of the order of 30x average coverage [1,2] suggests that for lower coverage the problem will be even more to get rid of false positives. Nevertheless, the results with coverage as less as 3-5x could also have a lot of meaningful results, and could be deployed for several genomes population wide analysis at relatively less cost, such as in the 1000 genomes project [1]. The 1000 genomes project used the technique of mapping the sequence reads to the reference genome, as it would not be possible to obtain any reliable genome assembly with an average sequencing coverage of 3-5x. There have been several new tools made available which can detect variations without the need for assembling the genome for the individual such as those used in the 1000 Genome Project consortium which finds great applicability in case the coverage of sequences is low[1]. Nevertheless, if the sequencing coverage is high enough such as above 12x in average, then there is no reason as to why assembling the genome and then mapping to a reference genome to detect variations directly should not be the adopted. At the same time, results obtained by assembly analysis can be compared for consistency by mapping reads to the reference genome approach to see if they both lead to same discoveries. The findings should then be experimentally validated by PCR and other traditional means, if there be time and resources, to get an estimate of false positive rates by both the approaches. As bioinformatics tools make use of a lot of predictive algorithms and machine-learning approaches, it is always wise to apply a combination of approaches, parameters and software tools to have a higher faith in the consensus results, thereby reducing the cost associated with experimental validation. The bioinformatics software tools aiding the analysis has been constantly growing and enhancing adapting rapidly with the improvement in sequencing quality and quantity

Bioinformatics White Space

The overall goal of conducting bioinformatics analysis for medical application is to look at the pattern of variation inheritance and to detect

any otherwise abnormal observation which can be a prospective discovery. This would be helpful in understanding human genetic variation, selection pressure and inheritance better for improved personalized medical treatment and trait characteristics determination. Genome variations have been associated with recurrent genomic rearrangements as well as with a variety of diseases, including colour blindness, psoriasis, HIV susceptibility, Crohn's disease and lupus glomerulonephritis [3-8]. There is thus a need of comprehensive catalogue of genotype and phenotype correlation studies [1-8] in particular when the rare or multiple variations in gene underlie characteristic or disease susceptibility [9,10]. Microarrays [11-13] and sequencing [14-17] reveal that structural variants (SVs) contribution is significant in characterizing population [18] and disease [19] characteristics. In particular the HLA region in chromosome 6 of an individual which is the MHC region in humans would be interesting in being decoded for the variations, as a lesser difference between two individuals could imply stronger success rate of organ transplant. Even otherwise, the HLA region variation would give an insight in immunologic responses. However, we must be careful with the results that we get when we call for the variations, as any difference could represent actual difference between the DNA sources, an assembly artefact (clone-induced or computational) or alignment error. Since the sequencing of human genomes now become routine [1], the spectrum of structural variants and copy number variants (CNVs) has widened to include even smaller events. What is important now is to know how genomes vary at large as well as fine scales. It is a challenge to understand its effects on human disease, characteristic traits and phylogenetic evolutionary clues. Figure A below tries to compile all the various terminologies and variations in genome architecture when compared to a reference genome [20, 21].

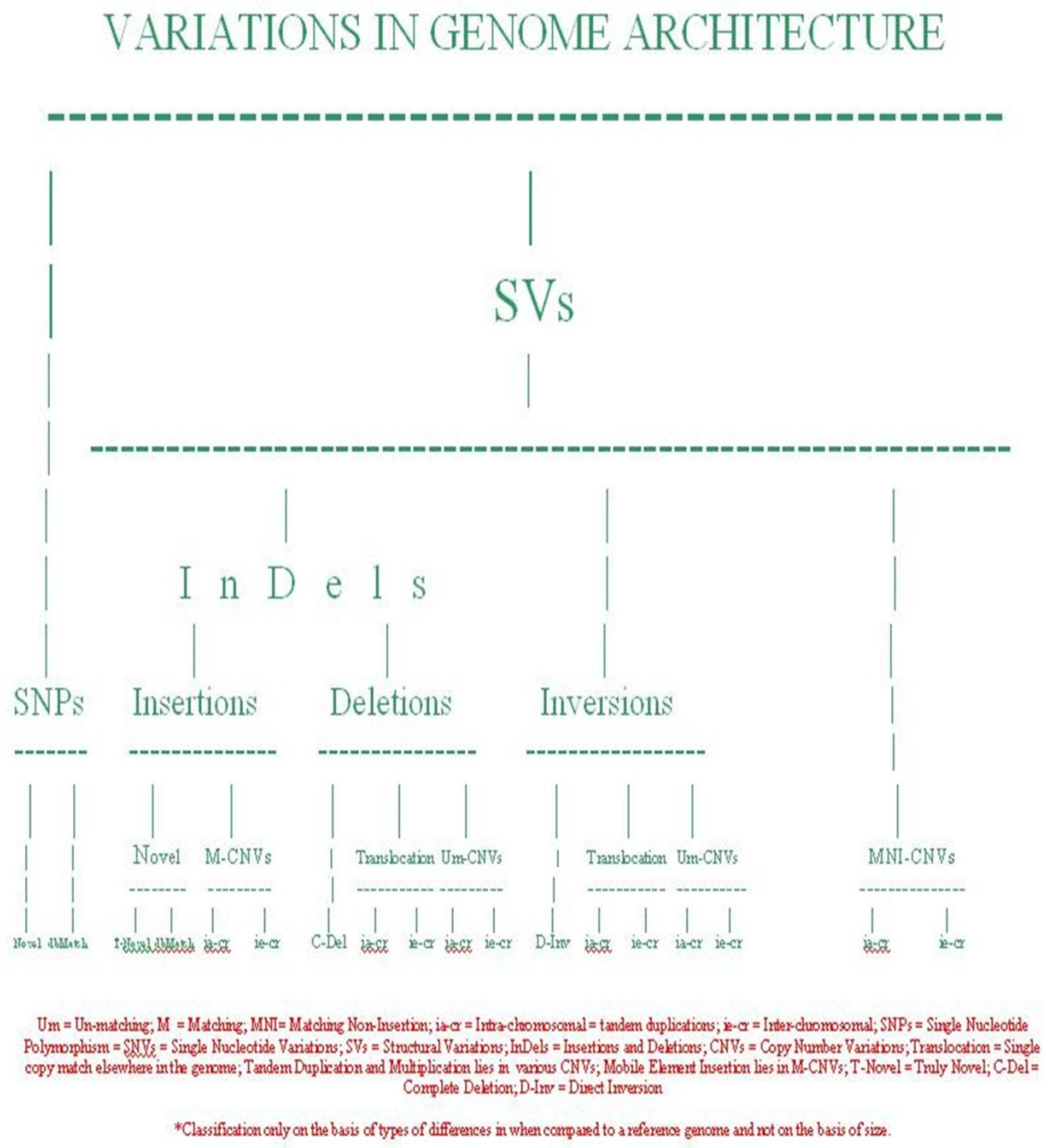

Figure A: Variations in Genome Architecture [20,21]

One clear application of finding the variations in an individual is in conducting an organ transplant surgery. If the immunologic responses after the grafting of an organ from a donor to the receptor may be determined a-priori to conducting the transplant, medical practitioners can be more predictive of the chances of success of the transplantation. This also applies to clinical data making and donor matching. The immunologic

responses are dictated by the MHC region of the genome, which in humans corresponds to the HLA region in chromosome 6. If we extract the SVs (structural variations) and SNPs (single nucleotide polymorphism) of chromosome 6 of the donor and compare it with the SVs and SNPs of acceptor patient's chromosome 6, then it can be reasonably proposed that the lower the differences between the two sets of SVs and SNPs, the higher the success possibility of organ transplant. However, even with these SVs and SNPs a subset could be more crucial to be present or being absent perhaps for the transplantation to be successful. Similarly if we are interested in any other chromosome which has been known of having strong association with a particular phenotype or characteristic trait, we can extract the SVs and SNPs for that chromosome and do a relational database analysis amongst other techniques such as machine learning approaches. Below in Figure B, is a GenomeBreak bioinformatics software tool plot for an individual assembled genome to detect the structural variation when compared to the reference genome [22, 23].

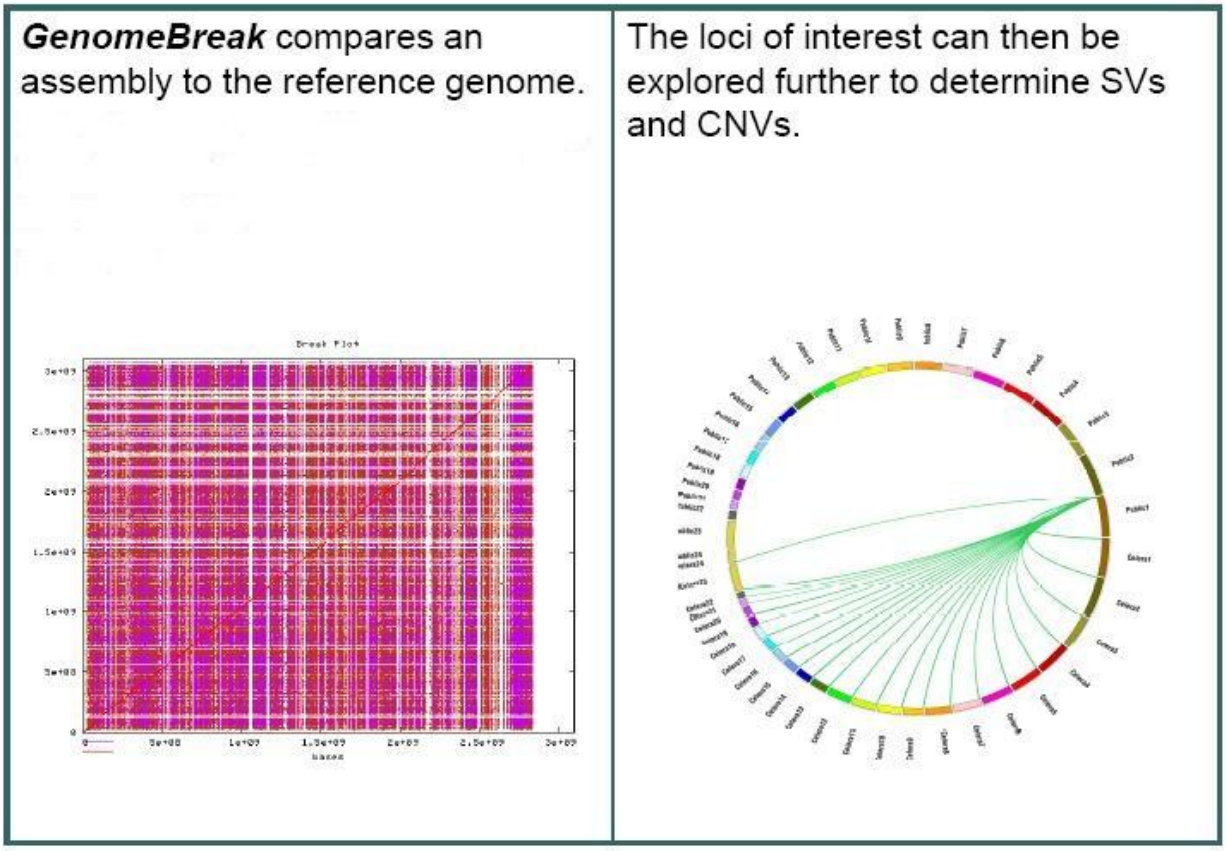

Figure B: GenomeBreak plot of a an assembled genome [22,23]

With the rapid advancement of technology, coupled with decrease in cost of sequencing, it will not be long when everyone can carry their

genome-chip which would contain chromosomal sequences, along with information of SVs and SNPs already determined. In fact, this would be a practice which we might want to do early in the life say within a week after his birth. Let's say we take it a step further and obtain the DNA sample from the fetus, thus being able to do analysis of the baby which is to be born. With the power of prediction and integrating it to powerful relational databases and other scientific techniques we can tell what are the chances of the baby to be healthy in general. We would be able to predict disease susceptibility of the new born baby as well as characteristics traits, thereby giving an opportunity for the mother to decide whether to have the baby or not, and if so what all things she should be caring about. We would also be able to determine the sex of the baby before it is born, thereby provide an alternative and safer means to determine the sex of the baby, without any extra cost, as the genome of the baby will be sequenced and analyzed anyways. The results from genome analysis can be more sensitive if we have parallel transcriptome analysis by means of RNA-Seq techniques. Genome analysis toolkit, GATK, tools and other tools for next generation sequencing of DNA, NGS, visualization such as IGV (interactive genomics viewer) find greater application at that point [24].

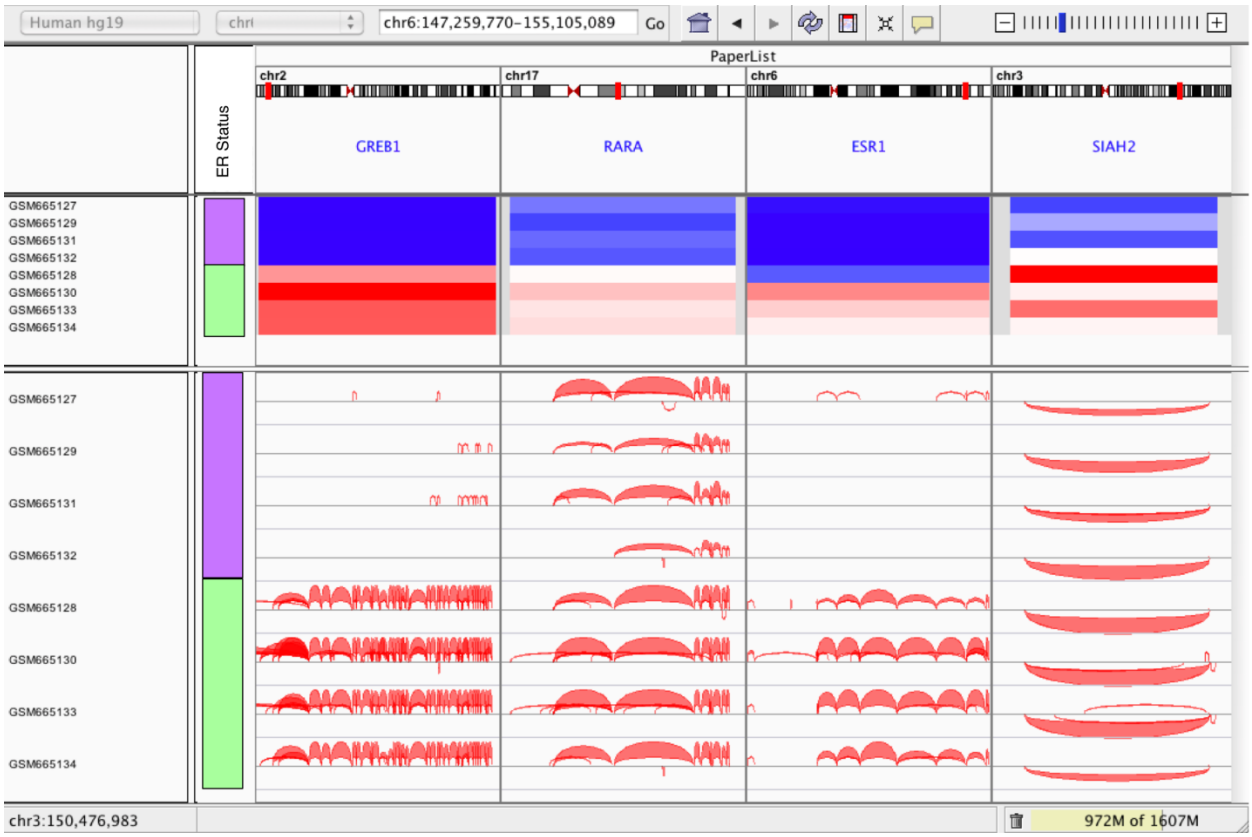

Figure C: Sample IGV plot

Going further for analysis tools for proteome analysis such as plot of intensities for mass to charge value, m/z peaks vs dilution, also would be great supplement to detect the peptides which can be present in a patient from his body fluid sample under certain condition such as while the patient has been diagnosed for certain symptoms. Personalized medicine performance can thus be tested on a more regular basis by means of such powerful tools which test for the expression of various proteins to be present in patient body while he is undergoing treatment as well. Below in Figure D we see one such tool plot by ProteomeBreak.

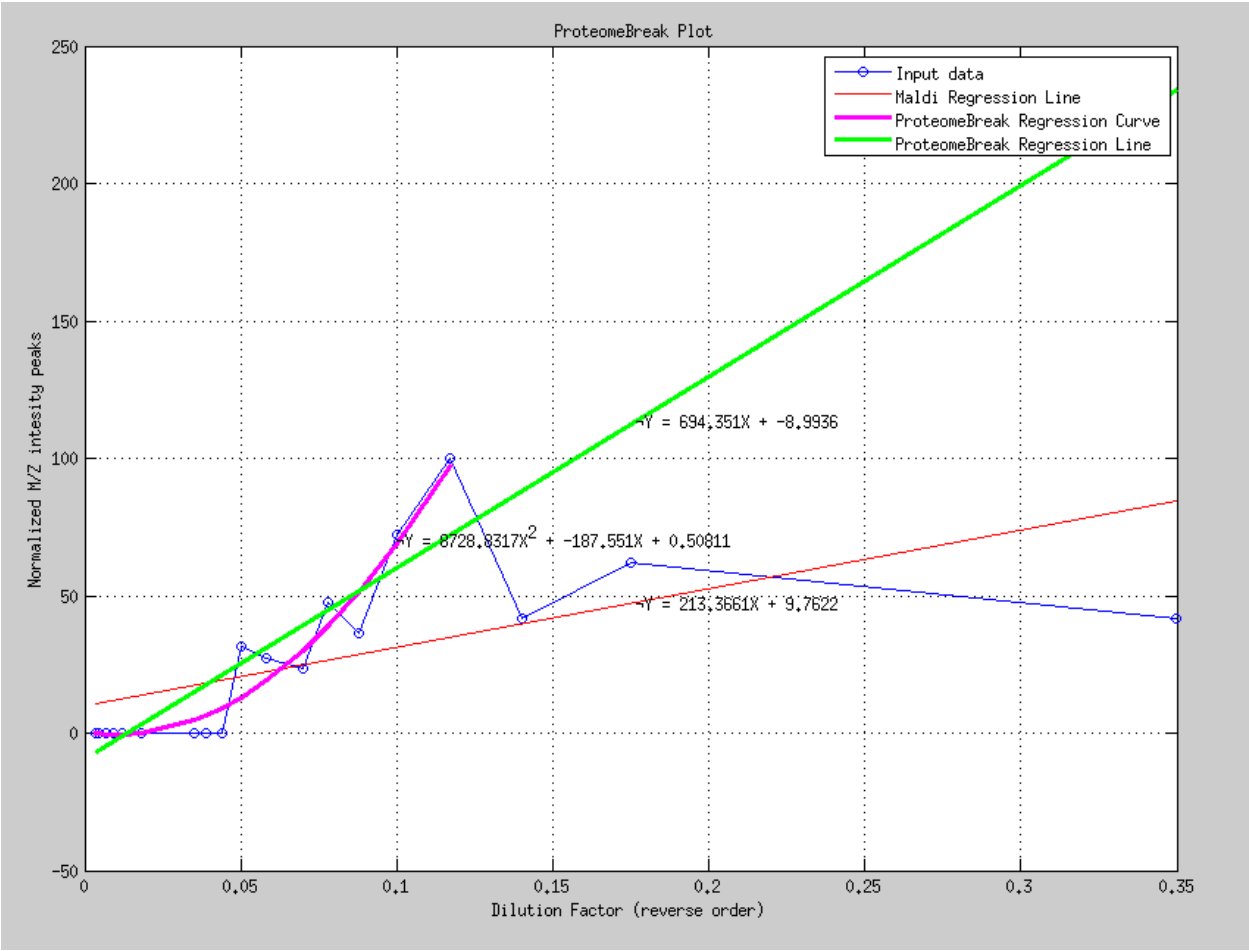

Figure D: ProteomeBreak Plot for Maldi M/Z peak normalized values plotted against dilution factor

Computational Demands and Skills

As we are living in a data explosion age, bioinformatics has been able to keep pace with the age definition. About a decade ago analysts typically dealt with gigabytes of data at most. Today, it is fairly common to see bioinformaticians dealing with terabytes of raw data, processed data and possibly petabytes of intermediate processing data. A good strategy and management approach is to depend on these high levels of data storage cloud-type or cluster facility. Such high performance computing facilities usually not only provide the support in terms of hardware, but usually also take care of different software tools with updated versions available. Such cluster facilities also make more computational resources available such as providing possibility of submission of several jobs, or running a parallel script using OpenMP, OpenMPI, MVAPICH2, perl Threads, pThreads for a faster execution. Typically these facilities can have varieties of computing nodes available, each varying in the specifications of processing power, memory available, I/O network bandwidth etc. which the informatician can decide as per the demand. As an example, the genome assembly tools currently usually can take quite high memory compared to other traditional work like pattern extraction for a motif search. Among programming languages that have become popular in the bioinformatics world are Perl and Python. Nevertheless, as C, C++, Java, Pascal, shell script, MySQL, Matlab are usually popular in the computer science world, they will continue to show their existence and application in bioinformatics world too. As an example GUI programming is quite extensively done using Java, and most MPI (message passing interface) applications are usually developed on C and C++. Among the operating systems Linux has fairly dominated the programming world. For the Windows user if you have access to a remote login linux machine, then putty generally serves as a good tool for quick connection for free, though other commercial tools are also available. For making use of two operating systems simultaneously such as the Oracle VM VirtualBox is getting increasingly popular. While different software tools exist for various bioinformatics applications, each have their own merits and demerits in terms of statistics and reliability of the assembly generated apart from computationally important aspects such as resource utilization and execution time, and those factors should be preferably considered before going for full blown operation.

Employers tend to forget that despite all these facilities, the most important factor lies with having key people who can do right analysis, come up with ideas and algorithms apart from having capabilities to implement those ideas as a software code. Typically such key people have strong background of education and experiences both in biotechnology and computer science apart from exposure to mathematics and engineering world. People factor plays a key role since the ideas and direction given by people might be much more worthy than lots of effort put in taking the project in a not so sensible direction.

Conclusion

Bioinformatics field represented by genomics, proteomics, transcriptomics, and metabolomics is mature as well as evolving at a fast pace and can thus be tightly linked to personalized medicine. Among the above subcategories, the genomics field has matured to a greater extent such that scientists even went on to coin personalized genomic medicine as a more specific category within personalized medicine. Whatever be the case, there is no doubt that bioinformatics is tightly coupled towards bringing in the capability that will be required to deliver personalized medicine, such as with the example tools that are discussed above. Apart from these, the databases would play crucial role and would lead to more job creation as personalized medicine gets to practice.

Author Contact

abhishek.narain@iitdalumni.com

References

1. 1000 Genomes Project Consortium et al. A map of human genome variation from population scale sequencing. Nature 467, 1061-1073 (2010).

2. Mills, R.E. et al. Mapping copy number variation by population scale sequencing. Nature published online, doi:1:10.1038/nature09708 (3 February 2011).

3. Fanciulli, M. et al. FCGR3B copy number variation is associated with susceptibility fo systemic, but not organ-specific, autoimmunity. Nat. Genet. 39, 721-823 (2007).

4. Aitman, T.J. et al. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439, 851-855 (2006).

5. Gonzalez, E. et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307, 1434-1440 (2005).

6. Fellermann, K. et al. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am. J. Hum. Genet. 79, 439-448 (2006).

7. Yang, Y. et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in Europena Americans. Am. J. Hum. Genet. 80, 1037-1054 (2007).

8. Hollox, E.J. et al. Psoriasis is associated with increased beta-defensin

genomic copy number. Nat. Genet. 40, 23-25 (2008).

9. Feuk L, Carson AR, Scherer SW: Structural variation in the human genome. Nat Rev Genet 2006, 7:85-97.

10. Bodmer W, Bonilla C: Common and rarae variants in multifactorial susceptibility to common diseases. Nat Genet 2008, 40:695-701.

11. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: Detection of large-scale variation in the human genome. Nat Genet 2004, 36:949-951.

12. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M: Large-scale copy number polymorphism in the human genome. Science 2004, 305:525-528.

13. Redon R, Ishikawa S, Firch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, et al.: Global variation in copy number in the human genome Nature 2006, 444:444-454.

14. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE: Fine-scale structural variation of the human genome. Nature 2006, 444:444-454.

15. Khaja R, Zhang J, MacDonald JR, He Y, Josheph-George AM, Wei J, Rafiq MA, Qian C, Shago M, Pantano L, Aburatani H, Jones K, Redon R, Hurles M, Armengol L, Estivill X, Mural RJ, Lee c, Scherer SW, Feuk L: Genome assembly comparison identifies structural variants in the human genome. Nat Genet 2006, 38:1413-1418.

16. Korbel JO, Urban AE, Affourtiti JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M: Paired-end mapping reveals extensive structural variation in the human genome. Nat Genet 2006, 38:1413-1418.

17. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Trague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tuzun E, Cheng Z, Ebling HM, Tusneem N, David R, Cillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, et al.: Mapping and sequencing of structural variation from eight human genomes. Nature 2008, 453:56-64.

18. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, HuM, Ihm CH,

Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME: Origins and functional impact of copy number variation in the human genome. Nature 2010, 464:704-712.

19. Buchanan JA, Scherer SW: Contemplating effects of genomic structural variation. Genet Med 2008, 10:639-647

20. Abhishek Narain Singh, A105 Family Decoded: Discovery of Genome-Wide Fingerprints for Personalized Genomic Medicine, Poster, 2-5 Feb UPCP 2012, Florence, Italy

21. Abhishek Narain Singh, "A105 Family Decoded: Discovery of Genome-Wide Fingerprints for Personalized Genomic Medicine", page 115-126, Proceedings of the International Congress on Personalized Medicine UPCP 2012 (February 2-5, 2012, Florence, Italy), Medimond Publisher, ScienceMED journal vol.3 issue 2, April 2012.

22. Abhishek Narain Singh, Comparison of Structural Variation between Build 36 Reference Genome and Celera R27c Genome using GenomeBreak, Poster Presentation, The 2nd Symposium on Systems Genetics,
Groningen, 29-30 September 2011

23. Abhishek Singh, GENOMEBREAK: A versatile computational tool for genome-wide rapid investigation, exploring the human genome, a step towards personalized genomic medicine, Poster 70, Human Genome Meeting 2011, Dubai, March 2011

24. Helga Thorvaldsdottir, James T. Robinson, Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 2012.