# SARVE SANTU NIRAMAYA: COMPUTATIONAL BIOLOGY'S PROMISES FOR INDIA

B. MISHRA

MARCH 2, 2010

[1]The recent biotechnology revolution resembles information technology's early days and promises similarly explosive progress and ubiquitous impact. Biotechnology has its own versions of Moore's law; it strives to extract pertinent information from exponentially growing genomic and bio-medical (e.g., electronic health record) data; it aims to provide cheap and fast access to health and disease related information; and finally, it is poised to individualize medicine by combining statistics from large groups of patients and their relatives organized in "G2G (genome-to-genome) networks," where people anonymously share data on their environments and ancestry. It is no wonder that at the core of the machinery driving biotechnology sits massive computational networks, and innovative computational biological algorithms. Unlike many other technologies, however, this field thrives on wide-ranging multi-disciplinary cooperation and collaboration, as can be seen in the birth and growth of its subfields such as: (i) genomics and other –omics spectra, (ii) bioinformatics and computational biology, (iii) whole-genome association studies (WGAS), and (iv) computational systems biology. India must prepare now to play a key leading role in this unraveling bio-medical revolution with the specific goal of providing universal, ubiquitous and individualized health care to her vast and varied populace. India's rich ethnic diversity provides a unique opportunity to translate her collective genomic information into profound scientific and public health benefits.

## *Introduction*

"Whither do we go and what shall be our endeavour?" Pandit Jawaharlal Nehru, the first prime minister of India asked a newly independent India on the midnight of August 15 1947. "To bring freedom and opportunity to the common man, to the peasants and workers of India; to fight and end poverty and ignorance and disease; to build up a prosperous, democratic and progressive nation, and to create social, economic and political institutions which will ensure justice and fullness of life to every man and woman. We have hard work ahead."

[1] (Professor of Computer Science, Mathematics, Cell Biology, Courant Institute and NYU School of Medicine; Visiting Scholar, Cold Spring Harbor Laboratory; Adjunct Professor, Tata Institute of Fundamental Research; Professor of Human Genetics, Mt Sinai School of Medicine)

It may now be argued that India has substantially redeemed Nehru's pledge, as would be apparent from India's recent progress and prosperity, spurred by her leading role in information and communication technology and computer science. India seems to be in a comfortable position to replicate a similar success in the area of emerging computational biology and biotechnology, which could then have an enormous impact on India's agro-science, biomedicine and health care infrastructure and help it fight poverty, ignorance and disease. But there are several complications, and India has hard work ahead.

For millennia, India has been a melting pot, which gave rise to the current population of approximately 1.17 billion people (more precisely, according to CIA World Factbook [2], 1,166,079,217 people on July 1st 2009 est.) with remarkably diverse genealogies and demographics. India's linguistic diversity (with four major language groups Indo-European, Dravidian, Austro-Asiatic and Tibeto-Burman; with language isolates like Nihali or great Andamanese), religious multiplicity (with every major religion represented) and racial variety (more than two thousand ethnic groups) are all a testimony to her ability to merge wave after wave of migratory populations in her great crucible of cultural and genetic assimilation. India's population is made of, quoting Pandit Nehru again, "separate individual men and women, each differing from the other ... a bundle of contradictions held together by strong but invisible threads."

This rich and dynamic collective history of India can be read from the individual DNAs of each Indian, as encoded in six billion base pairs of A, T, C and G's organized in bundles of 23 chromosomes of each individual human genome — that two meters of double-stranded invisible thread. The entire Indian human population can thus be described by about $7.3805818 \times 10^{18}$ base pairs, if one is to trust CIA [3] and DOE [4] (human genome project information) factbook estimates. This entire body of genomic information can be stored in merely 2 exabytes of memory, significantly less than the current global monthly Internet traffic. But to give it a human perspective, note that this information is equivalent to about $1 \times 10^{12}$ (1 trillion) copies of Mahabharatas, the great Indian epic, which surpasses in size anything humanly memorized, written, crafted or understood. While reading and storing India's genomic information may be a surmountable challenge, understanding it with every bit of its nuances, is not!

At a rough glance, this body of genomic information can be described in terms of haplogroups (groups with similar genomic variations) as follows: The Indian male lineage (inferred from Y-chromosomes inherited patrilineally) consists of haplogroups R1a (20%), H (30%), R2 (15%), L (10%) and NOP (10%, excluding

[2] The World Factbook by Central Intelligence Agency. https://www.cia.gov/library/publications/the-world-factbook/.

[3] The World Factbook by Central Intelligence Agency. https://www.cia.gov/library/publications/the-world-factbook/.

[4] The Human Genome Project Information by DOE. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml.

R) and the Indian female lineage (inferred from mitochondrial DNA, mtDNA, inherited matrilineally) is primarily made up of haplogroups M (60%), UK (15%), and N (25%, excluding UK). Recently, the Indian sex ratio has been plunging precipitously, with perilous genetic and cultural consequences for the entire population; according to the 2001 census, India's sex ratio is 927 female for 1,000 male (at birth). The situation is worse in states like Punjab, Haryana and Delhi, but better in few states like Kerala. India's population is rather young with an age-structure comprising: 0–14 years: 30.8%, 15–64 years: 64.3%, 65+ years: 4.9%, and has a population growth rate of 1.548% (2009 est.). As India's population ages, and when the demographic dividends of the current decade are all spent, the health care cost of the older Indians could become an onerous burden. Preparing for this future, India must step up to innovate and create a unique vision of her own — in bio-medicine, biotechnology and computational biology at all scales ranging from single nucleotides, single molecules and single cells to individual citizens and her entire population. Concepts from many disciplines must intermingle to fulfill this vision, namely, population genetics, genomics, biotechnology, bioinformatics, genome-wide association studies and systems biology, as described below.

## Population Genomics

From the statistical analysis of the haplogroups on the Y-chromosomes and mtDNA, one could attempt to infer a skeletal and rudimentary history of the first human population to inhabit India and its subsequent evolution. To properly interpret these observed haplogroup marker frequencies in an extant population, the population model must include the effects of population sizes, vicariance (population splitting as a consequence of geographical events), ethnic segregation and mixing within the population as well as migration. Most of these processes and their parametric structures, however remain unknown. Reconstructing this history faces further challenges as it very likely involves a complex population substructure, intricate gene flows constrained by rigid adherence to endogamy (the practice of marrying within a specific ethnic group, as determined by the caste system) and ancestries marked by admixture lineages. India also seems to have experienced many successive population expansions and contractions varying over her vast geography, which exaggerate mutational differences, induced by random sorting of allele frequencies.

From the mtDNA and Y-chromosome data, available prior to 2007, Endicott et al. [5] have argued that the earliest Indian population arose from a rapid dispersal of modern humans from

[5] P Endicott, M Metspalu, and T Kivisild. *Genetic evidence on modern human dispersals in South Asia: Y chromosome and mitochondrial DNA perspectives*, chapter The Evolution and History of Human Populations in South Asia, pages 229–244. Springer, 2007.

eastern Africa and subsequent settlement in South Asia, consistent with both Out of Africa and Strong Garden of Eden hypotheses. The Out of Africa or African Replacement Hypothesis claims that every living human being is descended from a small group in Africa, who then dispersed into the wider world displacing earlier forms; this hypothesis is mainly supported by mtDNA data that point to all humans descending ultimately from one female: the Mitochondrial Eve. The Strong Garden of Eden Hypothesis claims that the out of Africa expansion was followed by a single and relatively fast range expansion and not by a gradual series of expansions outside Africa (the Weak Garden of Eden Scenario). The argument supporting these hypotheses goes as follows: the African human population appears to be the most diverse in terms of polymorphisms (which describe variations in genomes within a species or a population), whereas, in contrast, the non-African populations show somewhat limited diversity. The main non-African mtDNA diversity is limited to haplogroups M, N and R, and are found in the extant population of South Asia and Australia, while the West Eurasian population mainly exhibits the mtDNA haplogroups: N and R. A similar picture emerges for the Y haplogroups, thus supporting a single migration out of Africa around $\sim$ 65,000 years ago (coalescent time for mtDNA haplogroups M, N and R). These haplogroup distributions also paint a picture of South Asia as a crossroad of early human migration.

These data also signify other population expansions into the Indian subcontinent just before the Last Glacial Maximum that occurred about $\sim$ 18,000 years ago, when the favorable climatic conditions permitted recent migrations into India. Haplogroups in Pakistan and western-most states of India share a considerable amount of western Eurasian specific haplotypes and hint at a Eurasian migration. Similarly, Austro-Asiatic speaking populations in East Indian states (Orissa, Jharkhand, Bihar and West Bengal) have mitochondrial and Y chromosomal haplogroups originating east of India, indicative of an East Asian migration through the northeast corridor of India.

However, the phylogeography described above is simply based on low-resolution uniparental data derived from rather short mtDNA (about 16,596 bps) or a small fragment of Y chromosome (about 50,000 bps, comprising only about 0.1 % of Y), and thus, severely limited by the small number of markers available on them. A recent study by Reich et al. [6] has attempted to reconstruct Indian population history better by using genotype data of 132 Indian samples from 25 groups, collected with Affymetrix 6.0 SNP array on 560,123 SNPs (Single Nucleotide Polymorphisms, pronounced "snip"). SNPs are DNA sequence variation occurring

[6] D Reich, K Thangaraj, N Patterson, AL Price, and L Singh. Reconstructing Indian population history. *Nature*, 461: 489–494, 2009.

when a single nucleotide – A, T, C, or G – in the genome differs between members of a species or a population. SNPs, making up about 90% of all human genetic variation, occur every 100 to 300 bases along the 3-billion-base long human genome.

The analysis of Reich et al. [7] of Indian SNP data revealed further subtle structures in the population and gene flows that were missing from the earlier analyses: (1) The structure of the current Indian population could be described in terms of two idealized genetically divergent ancestral populations: ANI (Ancestral North Indians) and ASI (Ancestral South Indians), the former being genetically close to Central Asians, Europeans and Middle Easterners, and the later being a seemingly distinct human subgroup. ANI-ASI admixtures could be described along an Indian Cline. (2) ANI ancestry was estimated to be higher than average in the upper castes (Brahmins and Kshatriyas) and Indo-Aryan linguistic groups. ASI ancestry was determined to be best represented by the Onge population in Andaman Island. (3) Autosomal estimates of ANI ancestries showed a stronger correlation with Y haplogroup frequencies than those of mtDNA, suggesting a stronger male gene flow from groups with high ANI ancestry into ones with less. (4) A principal component analysis (PCA) on genotype SNP data revealed an interesting configuration of the Indian population groups, which identified two outlier groups as the Siddi (African ancestry) and the Nyshi and Ao Naga (Chinese ancestry), corresponding to the first two principal components. (4) Fisher's Fixation Distance (FST) statistics measured the genetic distances, leading to the conclusion that the 19 main Indian population groups showed much more differences than the traditional 23 European groups do. It was suggested that Indian population groups might have been established by a few individuals (founders), followed by limited gene flow. (5) It was also proposed that such enduring genetic signatures of founder events support the hypotheses that group distinctions are ancient and preserved in high fidelity because of strong endogamy, a consequence of strict taboos against inter-caste marriages. (6) The widespread history of founder events implied a high-rate of recessive diseases, and makes these Indian population groups ideal for extensive studies focusing on genetic diseases and gene mapping.

These studies using SNP arrays to assess the genetic distances have several shortcomings: these arrays are based on known SNPs that were derived using small samples from populations very different from the Indian populations; the probes on the arrays were selected, based on a single reference human genome sequence that may not be sufficiently genetically representative of the Indian groups; and finally, because the technology and the data could not disambiguate haplotypic phasing, the estimates of allele sharing

[7] D Reich, K Thangaraj, N Patterson, AL Price, and L Singh. Reconstructing Indian population history. *Nature*, 461: 489–494, 2009.

statistics suffer from high uncertainties. Furthermore, SNP array based data remain blind to many other kinds of polymorphisms: e.g., CNV (copy number variations) and SV (structural variations). To circumvent such concerns, one must develop technologies for sequencing whole-genomes of many individuals, preferably haplotypically. Despite the amazing progresses outlined here, a lot of hard work lays ahead.

## *Genomics and Computational Biology*

The preceding discussion leads directly to our next topic, building on a vast body of computational biology literature devoted to mapping, sequencing and sequence assembly algorithms. The subject has its origin in classical "stringology" of theoretical computer science [8], but received a big boost at the start of the Human Genome project [9] and is again enjoying a revival with the advent of next- and next-next-generation sequencing technologies [10].

The diploid human genome, containing all our hereditary information, is composed of about 6 billion DNA base pairs in total. The paired bases A (adenine), T (thymine), C (cytosine) and G (guanine) satisfy a complementarity principle: A pairs with T and C, with G. Thus, as a computational object, a genome could be represented and manipulated as a data structure of a set of strings over an alphabet of size 4.

The bases, cytosine and thymine, are smaller (lighter) molecules, called pyrimidines, whereas the other two bases, guanine and adenine, are bigger (bulkier) and called purines. Furthermore, adenine and thymine allow only for double hydrogen bonding, while cytosine and guanine allow for triple hydrogen bonding. As a result, the chemical (through hydrogen bonding) and the mechanical (purine to pyrimidine) constraints on the pairing lead to the complementarity and make the double stranded DNA both chemically inert and mechanically rigid and stable. Thus, despite its uninspiring physics and chemistry, DNA makes a fascinating information theoretic object through its capabilities for stable storage, high-fidelity template-driven copying mechanism, error-correction and resilience, and finally, its ability to reorganize through recombination, strand-invasion, mutation, deletion, insertion, translocation and deletion. Many computational biologists have been drawn to DNA not just for its biological role, but by its simplicity, elegance and sheer computational power; one could imagine using DNA to construct future computers to solve intractable problems and nanorobots to self-assemble complex materials.

Genomics analysis is deemed fundamental to biology and computational biology, as we have come to envision DNA as defining

[8] D Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[9] The Human Genome Project Information by DOE. `http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml`.

[10] SC Schuster. Next-generation sequencing transforms today's biology. *Nature Methods*, pages 16–18, 2008.

all of biology almost axiomatically through Crick's central dogma, which states that the information flows unidirectionally from DNA to RNA through transcription and then to proteins through translation. Thus by understanding how changes to an individual's genome affects transcription (DNA to RNA) and translation (RNA to proteins) of all its genes, we may aim to understand how biology works at a system level, as in the emerging field of systems biology, and the classical subject of "(forward) genetics," an approach to discovering the function of a gene by analyzing the discernible traits (phenotypes).

If we compare genome sequences of two individuals from a population, we expect that, while most of the sequences will be almost identical, there will be few sporadic differences, giving rise to various polymorphisms (distinctive "forms" in the population). For instance, an individual's autosomal gene (with two copies, one inherited from the father and the other from the mother) may differ from a copy of the same gene in another individual's genome (or from each other), in various ways: different bases at different positions, a small insertion or deletion, the gene may have different copy numbers, or because the gene is inverted or at a different chromosomal location. If a single nucleotide differs at some position, the polymorphism is called a SNP (single nucleotide polymorphism); if the copy numbers differ, it is a CNV (copy number variations), etc. The effect of polymorphisms on the gene's function can be through many mechanisms and quite complex: it may change the gene's transcribed expression, dosage/amount (due to CNVs) or the final translated protein (due to a SNP) and thus modify the biochemical reactions within a cell, or signaling and communication among various cells. Normally an individual has two copies of a gene (she is homozygous if two copies are same and heterozygous if they are distinct), thus allowing none, one or both copies of the gene to be distinct from the common form of the gene. In that case, how the genotypic variations determine the phenotypes also depends on the gene's dominance: If the gene's effect is dominant a single mutation can alter the trait, and if the effect is recessive then mutations in both copies would be necessary to alter the trait. Similar polymorphisms, occurring in the regions regulating gene transcription or alternate splicing or in genes for transcriptional factors, alter the phenotypes in an indirect, but more complex manner. A vast amount of polymorphisms occur in synonymous bases of genes, or in introns and intergenic regions with no effect on phenotypes and are thus considered "neutral." These mutations drift through the genomes in a population in a random manner, and can be tracked to understand ancestry. Occasionally a random mutation turns out to present a selective advantage as it gives the carriers an increased fitness.

Through selective sweep the mutated variant (allele) increases its population frequency as the result of recent and strong positive natural selection. Since neutral and nearly neutral genetic variation linked to the new mutation will also become more prevalent (as they hitch-hike), the signs of selective sweep can be detected from the haplotypic sequences from a population.

Thus, reverse genetics, ancestry and population dynamics studies ultimately rely on progress in sequencing technology and algorithmics, raising the following question: How can one read an individual's genome from end-to-end, while making sure that the technology is accurate enough to correctly discern all of the SNPs (and other polymorphisms), and unambiguous enough to determine the haplotypes from the homologous pairs of chromosomes? Because of various inherent technological limitations and computational intractabilities, these goals pose many difficulties, some apparently insurmountable.

The Human Genome Project (HGP) has produced and published a reference sequence of the euchromatic human genome, and determined that the haploid human genome contains approximately 23,000 protein-coding genes (far fewer than the estimate of about 120,000 genes, which had been expected before genome sequencing) and that only about 1.5% of the genome codes for proteins, while the rest consists of non-coding RNA genes, regulatory sequences, introns, and a significant portion with no known function ("junk" DNA).

The Human Genome Project initially produced two unfinished draft sequences by two different methods, one by the International Human Genome Sequencing Consortium (IHGSC) and another by Celera genomics (CG). The published IHGSC assembly was constructed by the program GigAssembler, devised at the university of California at Santa Cruz (UCSC). Unfortunately, these drafts have never been fully validated. In a recent article [11] it was noted: "Of particular interest are the relative rates of misassembly (sequence assembled in the wrong order and/or orientation) and the relative coverage achieved by the three protocols. Unfortunately the UCSC group [was] alone in having published assessments of the rate of misassembly... Using artificial data sets, they found that, on average 10 per cent of assembled fragments were assigned the wrong orientation and 15 per cent of fragments were placed in wrong order by their protocol. Two independent assessments [more recent] of UCSC assemblies have come to the similar conclusions."

For various technological reasons, it has only been possible to read short (about 700 – 1000 bps) and non-contextual (missing location) subsequences of the genome. The problem of inferring the entire genome sequence from many such non-contextual short

[11] CAM Semple. *Bioinformatics for Geneticists*, chapter Assembling a View of the Human Genome, pages 59–84. Wiley, 2007.

reads (taken from many identical copies of the genome) has been dubbed "shotgun sequencing approach," and attracted the attention of many computational biologists. Several of them, with years of experience in developing shotgun assembly pipelines, [12] have argued, "the sequence reconstruction problem that we take as our formulation of DNA sequence assembly is a variation of the shortest common superstring problem (SCSP), complicated by the presence of sequencing errors and reverse complements of fragments. Since the simpler superstring problem is NP-hard, any efficient reconstruction procedure must resort to heuristics [giving rise to approximate, incomplete and less-than-correct solutions]." NP-hard computational problems are assumed not to yield to any computationally feasible approach, unless a long-standing conjecture (P not = NP) is refuted.

Practically all sequencing pipelines, currently in use, follow search strategies that are strongly influenced by the reasoning above and aim to heuristically compute reasonable approximation of the true genome. Thus, based on heuristic search strategies assembly algorithms can be divided into two major categories: greedy and graph-based. In the greedy category are included algorithms that typically construct the solution incrementally, while choosing the "locally best" overlapping sequence-fragment pairs to merge at each step. Well known assemblers in this category include: TIGR [13], Phrap [14], and CAP3 [15].

In the graph-based category, assemblers start by preprocessing the sequence-reads to determine the pair-wise overlap information and represent these binary relationships as (unweighted) edges in a string-graph. Depending upon how the overlap relation is represented in these graphs, two main assembly paradigms have emerged: overlap-layout-consensus (OLC) and sequencing-by-hybridization (SBH). Well known assemblers based on OLC approach include: CELERA [16], ARACHNE [17], and Minimus [18].Two prominent examples of the other SBH approach include: EULER [19], and Velvet [20]. Because of their relation to various graph-theoretic NP-complete problems, Hamiltonian-path problem and Eulerian-superpath problem, respectively, both OLC and SBH approaches face the inherent intractability that lurks in their cores.

A counter-intuitive approach, suggested by a new pipeline, SUTTA [21], is to simply put the computational complexity and intractability question aside temporarily. Instead SUTTA aims to develop an accurate formulation of the problem and solve it exactly. Once it realizes where, in the structure of the formulation of the problem, the computational complexity becomes exacerbating, it tames the algorithm by clever pruning: specifically, SUTTA formulates the assembly problem in terms of a constrained optimization: It relies on a rather simple and easily verifiable definition of fea-

[12] J Kececioglu and E Myers. Combinatorial algorithms for dna sequence assembly. *Algorithmica*, 13:7–51, 1995.

[13] G. G. Sutton, O. White, M. D. Adams, and A. R. Kerlavage. TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1(1): 9–19, 1995.

[14] P Green. Phrap documentation, 1996. URL `http://www.phrap.org/phredphrap/phrap.html`. http://www.phrap.org/phredphrap/phrap.html.

[15] X Huang and A Madan. CAP3: A DNA Sequence Assembly Program. *Genome Research*, 9(9):868–877, 1999. doi: 10.1101/gr.9.9.868. URL `http://genome.cshlp.org/content/9/9/868.abstract`.

[16] EW Myers, GG Sutton, AL Delcher, IM Dew, DP Fasulo, MJ Flanigan, SA Kravitz, CM Mobarry, KHJ Reinert, KA Remington, EL Anson, RA Bolanos, H-H Chou, CM Jordan, AL Halpern, S Lonardi, EM Beasley, RC Brandon, L Chen, PJ Dunn, Z Lai, Y Liang, DR Nusskern, M Zhan, Q Zhang, X Zheng, GM Rubin, MD Adams, and JC Venter. A Whole-Genome Assembly of Drosophila. *Science*, 287(5461):2196–2204, 2000. doi: 10.1126/science.287.5461.2196. URL `http://www.sciencemag.org/cgi/content/abstract/287/5461/2196`.

[17] S Batzoglou, DB Jaffe, K Stanley, J Butler, S Gnerre, E Mauceli, B Berger, JP Mesirov, and ES Lander. ARACHNE: A Whole-Genome Shotgun Assembler. *Genome Research*, 12(1):177–189, 2002. doi: 10.1101/gr.208902. URL `http://genome.cshlp.org/content/12/1/177.abstract`.

[18] D Sommer, A Delcher, S Salzberg, and M Pop. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics*, 8(1):64, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-64. URL `http://www.biomedcentral.com/1471-2105/8/64`.

[19] PA Pevzner, H Tang, and MS Waterman. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States*

sible solutions as "consistent layouts." It potentially generates all possible consistent layouts, organizing them as paths in a "double-tree" structure, rooted at a randomly selected "seed" read. Since a path can be progressively and quickly evaluated in terms of an optimality criteria, encoded by a score function of the set of overlaps along a lay-out corresponding to the path, it can also concomitantly check the validity of the lay-outs (with respect to various long-range information such as mate-pairs, optical [22], [23], [24] or probe [25] maps, dilution, etc.) through well-chosen constraint-related penalty functions, and prune most of the implausible lay-outs, using a branch-and-bound scheme. Ambiguities, resulting from repeats or haplotypic dissimilarities, may occasionally delay immediate pruning, forcing the algorithm to look ahead, but in practice, do not exact a high price in computational complexity of the algorithm. Additionally, SUTTA is capable, at least in principle, of agnostically adapting to various rapidly evolving technologies.

As the preceding discussion points out, there is now a clearer realization that the sequence assembly problem, with its seminal role in computational biology, needs and will see a rebirth: the problem has not been adequately solved and there are many new profitable avenues to explore. We seem to have left many interesting algorithmic "stones" unturned, which are hoped to attract talented Indian computer scientists to these challenges. Furthermore a competition, along the line of "International SAT Competition [26]," with benchmark real and in silico data and evaluation criteria should be set up for whole-genome sequencing problems.

There is an even more critical issue for Indian genomicists: What reference sequence should be used for population-wide studies in India? How many reference sequences? How representative could Craig Venter or Jim Watson's genomes (the first two to be sequenced) be for the Indian subcontinental population? As we attempt to understand polymorphisms, similar questions arise regarding the HAPMAP project: its coverage and its suitability. If India needs to start a Human Genome project, a HAPMAP project and a Human Population Genome project almost ab initio should it not get started as soon as possible? India has hard work ahead.

## *Disease Studies*

Equipped with tens of thousands of genomes from the extant Indian sub-continental population, we can try to reveal not only her past, but also how evolution has shaped and continues to shape India's collective biology: how her population lives, mates, reproduces, suffers and dies. Theodosius Dobzhansky has famously said, "Nothing in biology makes sense except in the light of evo-

[22] AH Samad, W-W Cai, X Hu, B Irvin, J Jing, J Reed, X Meng, J Huang, E Huff, B Porter, A Shenkar, TS Anantharaman, B Mishra, V Clarke, E Dimalanta, J Edington, C Hiort, R Rabbah, J Skiada, and DC Schwartz. Mapping the Genome One Molecule at a Time – Optical Mapping. *Nature*, 378:516–517, 1995.

[23] TS Anantharaman, B Mishra, and DC Schwartz. Genomics via Optical Mapping II: Ordered Restriction Maps. *Journal of Computational Biology*, 4:91–118, 1997.

[24] C Ashton, B Mishra, and DC Schwartz. Optical Mapping and Its Potential for Large-Scale Sequencing Projects. *Trends in Biotechnology*, 17:297–302, 1999.

[25] J West, J Healy, M Wigler, W Casey, and B Mishra. Validation of S. pombe sequence assembly by micro-array hybridization. *Journal of Computational Biology*, 13:1–20, 2006.

[26] The International SAT Competition. `http://www.satcompetition.org/`.

lution." However, evolution is Darwinian, its unidirectionality induced by the unidirectionality of the biological information flow – as captured in Crick's Central Dogma. Genomes in a population are continuously reorganized by various processes: single point mutations, insertions/deletions (indels), duplications, translocations and inversions, which via the transcription-translation information-flow alter the regulatory, metabolic and signaling processes defining the whole organism. More often than not, the effects of these modifications are deleterious, and lead to diseases, deaths and disappearances of species. Occasionally, the new genotype (coded by the genome) leads to an advantageous phenotype (exhibited by a trait), and rapidly diffuses through the population in a selective sweep, while inviting along other hitch-hiking genomic elements (polymorphisms/variants in linkage disequilibria). As one attempts to get a taste of the relation between biology's syntax and semantics by genome-wide association studies (GWAS), one has to mask out the ancestral hitch-hiking syntactic sugars, salts and chaffs.

Consequently, determining the etiology of a disease is nontrivial: genomic variants that correlate with a disease trait are not all causal. Additionally, environmental effects modulate the symptoms and severity of a disease. Genetic susceptibility of an individual to a trait depends on type-level causality (the population to which the individual belongs) as well as token-level causality (the gene-environment interactions in the specific patient). Similarly, genetics also plays a significant role in determining if a particular therapeutic intervention is likely to be more effective for a particular population or individual. For instance, in treating lung cancer, one may exploit the known association between response to Gefitinib and Erlotinib and mutations of the EGFR, and thus personalizing a specific treatment for a specific patient. Thus, treating all diseases in a genetic-agnostic manner is neither cost effective nor safe.

Association studies aim to discover genetic variations that differ in frequency between cases (affected) and controls (unaffected) or between individuals exhibiting different phenotypic values. Traditionally, association studies had been built upon low-throughput approaches in which a single putative gene was targeted and genotyped for genetic variants. A classical example is presented by the study that identified a significant association between APOE alleles and Alzheimer disease [27]. Such analysis may be thought of as hypothesis driven (leading to refutation or validation) and could be conducted by a small laboratory and requiring modest computational and technological resources. As one wishes to scale similar analyses to whole genomes (with all the genetic variants queried simultaneously), not only do the technological

[27] ER Martin, EH Lai, JR Gilbert, AR Rogala, AJ Afshari, J Riley, KL Finch, JF Stevens, KJ Livak, BD Slotterbeck, SH Slifer, LL Warren, PM Conneally, DE Schmechel, I Purvis, MA Pericak-Vance, AD Roses, and JM Vance. SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in alzheimer disease. *Am. J. Hum. Genet.*, 67:383–394, 2000.

and computational burdens grow massively, but the statistical analysis require thoughtful design: one must account for multiple hypothesis testing, suitable null-models reflecting population dynamics (effective size, bottlenecks), Yule-Simpson effects (exacerbated by unknown population stratification and admixtures), and variations in technologies and protocols employed in gathering the underlying data.

Currently, most genome-wide association studies are based on a single phenotype and genotypic information, contained in single nucleotide polymorphisms (SNPs). SNPs are found to be frequent in the genome; variants in physical proximity tend to correlate in genotype; and the correlations have been mapped substantially by the International HapMap Project [28]. SNP-based genome-wide association studies have enjoyed some early successes: Age-related Macular Degeneration [29] associated with complement factor H (using 96 cases and 50 controls), Wellcome Trust Case Control Consortium (WTCCC) [30] study for a wide class of diseases, namely, coronary heart disease, type 1 diabetes (T1D), type 2 diabetes (T2D), rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension – showing, for instance, that type 1 diabetes (T1D) is associated with six chromosomal regions (using about 12,000 individuals).

Many genome-wide association studies (such as the ones described earlier) have been motivated by and designed to test a specific hypothesis: The common disease/common variant hypothesis. This hypothesis postulates that polymorphic variations in the population of more than 5% frequency might increase susceptibility to common disease [31], [32]. It is argued that such variants have persisted in the population because any one of them is only slightly deleterious or has no effect on an individual until old age. Both recent rapid growth in effective population size of humans (after a population bottleneck) and increased lifespan have contributed in sheltering these common variants in an "evolutionary shadow." However, there have emerged many counter-examples that do not fit this hypothesis, thus leading to the alternative hypothesis: the rare variant hypothesis. Thus there appears a need for genome-wide association studies, designed to include low-penetrance rare variants (<1% population frequency), which might impart a moderately large relative risk. Lately, the focus has also shifted to understanding the role of other SV (structural variant) polymorphisms (CNVs, copy number variants and their analogs) in determining phenotypic variations [33], [34]. CNVs are defined as regions of duplications (copy number > 2) and deletions (copy number < 2) greater than 1 Kb, but no well-developed statistical method for interpreting their contribution to disease exists yet [35], [36].

[28] International HapMap Project. http://hapmap.ncbi.nlm.nih.gov/.

[29] JL Haines, MA Hauser, S Schmidt, WK Scott, LM Olson, P Gallins, KL Spencer, SY Kwan, M Noureddine, JR Gilbert, N Schnetz-Boutaud, A Agarwal, EA Postel, and MA Pericak-Vance. Complement factor H variant increases the risk of age-related macular degeneration. *Science*, 308: 419–421, 2005.

[30] Wellcome trust case control consortium. http://www.wtccc.org.uk/.

[31] ES Lander. The new genomics: global views of biology. *Science*, 274:536–539, 1996.

[32] A Chakravarti. Population genetics–making sense out of sequence. *Nat Genet*, 21:56–60, 1999.

[33] AJ Iafrate, L Feuk, MN Rivera, ML Listewnik, PK Donahoe, Y Qi, SW Scherer, and C Lee. Detection of large-scale variation in the human genome. *Nat Genet.*, 36:949–952, 2004.

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

[35] I Ionita, R Daruwala, and B Mishra. Mapping Tumor Suppressor Genes using Multipoint Statistics from Copy-Number Variation Data. *American Journal of Human Genetics*, 79:13–22, 2006.

[36] A Mitrofanova and B Mishra. On a Novel Coalescent Model for Genome-Wide

The current genome-wide association studies also need to account for systematic "missingness" (especially for rare alleles). Because of this, a strong haplotypic association may only imply that the truly causal variant lies on the haplotype background and may not have been typed. A population study, using the currently available array or short-sequence-read technologies, relies on estimating likely haplotypes, since the true chromosomal crossover points are unknown (as previous generations are not genotyped). These estimates use the population-wide data to impute the missing haplotype phases, and can be biased by the population structures, effective population size, degree of inbreeding, etc.

A related issue is that of quality control, needed to ensure that cases are well matched to the controls for ancestry. Lack of proper matching can enable hidden variables in association and lead to false-positives or may even mislead the direction of causation/association (Yule-Simpson effect). Usually correction by an inflation factor for "genomic control" [37] or incorporation of population structure by a subset of ancestry informative SNPs have been employed, but since the standard algorithms such as STRUCTURE [38] or mSTRUCT [39] for population stratification are based on genotyped SNPs, such solutions are not fully satisfactory.

For a population as complex as the Indian subcontinent's, we will need much more accurate data (e.g., whole-genome individually haplotyped sequences), capable of revealing SNPs, SVs and individual haplotype structure. If India holds the result of "the grandest genetic experiment ever performed on man," as Dobzhansky proclaimed, surely then it needs to be measured, evaluated and interpreted as best as possible. Aiming to build faster, better and cheaper biotechnology and GWAS algorithms will involve hard work, but shying away from these tasks is not necessarily an option.

## *Systems Biology*

Another issue, though less frequently discussed than deserved, concerns the errors in GWAS owing to incorrect phenotyping. Even for Mendelian diseases, two similar disease phenotypes could be easily confused, thus confounding any GWAS analysis: For instance, a dataset that mislabels FSS (Freeman-Sheldon Syndrome) for SHS (Sheldon-Hall Syndrome) can easily mislead GWAS into an incorrect causative (or associative) interpretation. Similarly, it is conjectured that genetic analysis of chronic fatigue syndrome has been severely frustrated by the heterogeneity of the disease. We could avoid this conundrum if we could extract traits directly and objectively from patient data [40] – say, the electronic health record (EHR) data; its causal analysis will derive trait

[37] B Devlin and K Roeder. Genomic control for association studies. *Biometrics*, 55:997–1004, 1999.

[38] JK Pritchard, M Stephens, and P Donnelly. Linkage disequilibrium in humans: models and data. *Genetics*, 155:945–959, 2000.

[39] S Shringarpure and EP Xing. mStruct: inference of population structure in light of both genetic admixing and allele mutations. *Genetics*, 182:575–593, 2009.

[40] S Kleinberg and B Mishra. Metamorphosis: the Coming Transformation of Translational Systems Biology. *Queue*, 7 (9):40–52, 2009a. ISSN 1542-7730. doi: http://doi.acm.org/10.1145/1626135.1629775.

(primary and secondary) and subtrait definitions, and will lead directly to disease etiology and its elucidation in the genetic and environmental contexts.

Causation, its definition and interpretation, have been a topic of interest to philosophers, logicians, statisticians and AI researchers [41], [42], [43], [44]. An appealing definition of causation could be based on the concepts of temporal priority and probability raising. In this way, it could be precisely defined in the language of probabilistic computational tree logic (a propositional branching time temporal logic), algorithmically interpreted using the techniques of model checking and statistically scored (to control false discovery rates) using empirical Bayes methods [45], [46]. Using this method we may represent relationships such as "(increasing BMI and smoking UNTIL hypertension) causes CHF in 8-10 months," and validate them in a rigorous way that allows us to automatically infer the associated probability of such a relationship from time series data.

Similar notions of causality can be extended to incorporate the effects of genetic variants and environmental covariates. However, such a causal explanation is purely phenomenological and devoid of a mechanistic/molecular basis. One would prefer the underlying models to also explain how biomolecules participating in various biochemical processes are synthesized, transcribed, translated, multi-merized, bound, folded, activated, regulated, metabolized, signaled, degraded, transferred, chaperoned, localized, co-localized, compartmentalized, deactivated, unbound, spliced, post-translationally modified, etc. Such processes can be mathematically represented quite faithfully using various formalisms: Finite State Machines, Ordinary or Partial Differential Equations, or Algebraic Hybrid Models, while enabling for certain of these models to be checked by automated procedures – albeit feasible but slow algorithms. Such algorithms can be made more efficient by incorporating further algorithmic advances, and the challenges they pose will likely attract some of the best algorithmic minds over the next decades. But more practically, they will enable not only an understanding of the disease etiology at molecular and process levels, but also means to therapeutic interventions, be they molecular, genetic or synthetic-systems-biologic. Hard work, challenges and opportunities await many young Indian mathematicians and computer scientists, who will be attracted to these emerging disciplines at the interface of computer science and biology.

[41] P Suppes. *A probabilistic theory of causality*. North-Holland, 1970.

[42] N Cartwright. *Nature's Capacities and Their Measurement*. Oxford University Press, 1994.

[43] S Kleinberg and B Mishra. The Temporal Logic of Token Causes. In *Proceedings of the 12th International Conference on the Principles of Knowledge Representation and Reasoning (KR)*, Toronto, Ontario, May 2010. To appear.

[44] S Kleinberg and B Mishra. The Temporal Logic of Causal Structures. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, Montreal, Quebec, June 2009b.

[45] S Kleinberg and B Mishra. The Temporal Logic of Token Causes. In *Proceedings of the 12th International Conference on the Principles of Knowledge Representation and Reasoning (KR)*, Toronto, Ontario, May 2010. To appear.

[46] S Kleinberg and B Mishra. The Temporal Logic of Causal Structures. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, Montreal, Quebec, June 2009b.

## *Recommendations*

In summary, the emerging fields of computational systems biology and population genomics analysis could be the next important high technology areas for India to nurture. There are several technological milestones to target: (a) Indian Human Population Genomes Project, sequencing several thousand whole-genomes haplotypically and accurately, (b) High Throughput and Inexpensive Haplotypic Sequencing Technology Project, capable of sequencing genomic DNA from small number of cells and minute amounts of genomic materials, (c) High Accuracy Sequence Assembly Project, aiming to make a quantum leap in the algorithmic technologies underlying accurate haplotypic sequence assembly software (accompanied with better sequence annotation and comparison algorithms), (d) Advanced Genome-Wide Association Studies Project, incorporating improved analysis of ancestry, haplotypes and causal relations, and (e) Translational Systems Biology Project, focusing on phenotyping, disease etiologies and systems biology models of diseases (with capabilities for model checking). The possibility of an Indian Human Population Genome Project is quite tantalizing as a denoument of the "grandest genetic experiment ever performed."

Some attention should be drawn to a rather global problem that affects all humans, which could be of immense concern. In humans, per-generation reduction in fitness (ranging between 1% and 5%) appears to be soaring, because of an unusually high rate of recurrent mutations. In a recent essay, Michael Lynch [47] wrote "[The] impact of deleterious mutations is accumulating on a time scale that is approximately the same as that for scenarios associated with global warming – perhaps not of great concern over a span of one or two generations, but with very considerable consequences on time scales of tens of generations." There is, thus, an acute need to invent intelligent approaches of genetic intervention, building upon accurate characterization of the underlying population dynamics (e.g., intronic-mutations modifying gene-isoforms).

In the context of just the Indian population, there are several important societal issues to be addressed (and not fully discussed here): (a) increased investment in the so-called third-world diseases (e.g., Malaria), (b) policies affecting population genetics (e.g., sex ratio in Indian population); (c) attention to an aging population and its impact on healthcare; (d) genetically modified organisms and their impact on the Indian agro-industry; (e) preventive and personalized medicine, its introduction to India and its inevitable impact on privacy and related issues.

There is one more thing worth touching upon: India has a long tradition and success in statistics, mathematics and logic. All of

[47] M Lynch. Rate, Molecular Spectrum, and Consequences of Human Mutation. *Proc Natl Acad Sci U S A*, 107:961–978, 2010.

these areas will play important roles in the projects, outlined earlier. India must invest in teaching and research in these areas, both in preparation of people developing areas related to bioinformatics and computational biology, and also for mathematicians, engineers, and scientists involved in modeling problems from the life sciences.

## *Conclusion and Dedication*

This essay is dedicated to the memory of two influential Indian scientists, Padma Bhushan HJ Bhabha (1909 – 1966), the great nuclear physicist and a father of Indian scientific revolution after independence, and Sir JBS Haldane (1892 – 1964), the great geneticist and evolutionary biologist, who spent his last years in Bhubaneswar as the director of Orissa State Government Genetics and Biometry Laboratory. Both of them, through their closeness to Pandit Nehru, played key roles in defining India's "Scientific Temper," which would allow her citizens to think independently, understand and practice the scientific method in daily lives.

Finally, one may reflect on the role of technology in India's future. India's ambition should be to continue developing technologies with the goal of establishing herself as a global super-power and a world leader, not necessarily in a military sense or even in an economic sense, but as an idea and example, representing the ambitions of all humanity and having its fount embedded in knowledge, science and technology – all aiming to end human suffering. The investment in biotechnology, nanotechnology, robotics, etc. would be the necessary steps in moving India in that direction.

The idea of that India would be based on her true intrinsic values — namely, her argumentative heterodoxy, her search for a fundamental understanding of truth and nature, technological progress tempered by ethical and environmental concerns, and most of all, her own perception of the richness of her genetic plurality. The idea of India could resemble the ideal idea of humanity. We must strive never to forget that we are just a temporary clonal eruption of a tiny fragile young infantile species that almost went extinct twice. Nor should we belittle the fact that, despite its lowly origin, something bigger holds true for this altruistic, trusting and tolerant species — more than *E. coli* or elephants. Perhaps, we already knew that when we decreed that "Sarve bhadrani pashyantu, ma kaschid dukhah bhag bhabet."

[48]

[48] The paper has improved considerably following many insightful suggestions from several colleagues: most notably, F. Hoppensteadt, S. Kleinberg, G. Narzisi, M. Subrahmanyam and A. Witzel of NYU, R. Parikh of CUNY, A. Nerode of Cornell, L. van der Ploeg of Abraxis and M. Wigler of CSHL.

## *References*

The World Factbook by Central Intelligence Agency. `https://www.cia.gov/library/publications/the-world-factbook/`.

The Human Genome Project Information by DOE. `http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml`.

International HapMap Project. `http://hapmap.ncbi.nlm.nih.gov/`.

The International SAT Competition. `http://www.satcompetition.org/`.

Wellcome trust case control consortium. `http://www.wtccc.org.uk/`.

TS Anantharaman, B Mishra, and DC Schwartz. Genomics via Optical Mapping II: Ordered Restriction Maps. *Journal of Computational Biology*, 4:91–118, 1997.

C Ashton, B Mishra, and DC Schwartz. Optical Mapping and Its Potential for Large-Scale Sequencing Projects. *Trends in Biotechnology*, 17:297–302, 1999.

S Batzoglou, DB Jaffe, K Stanley, J Butler, S Gnerre, E Mauceli, B Berger, JP Mesirov, and ES Lander. ARACHNE: A Whole-Genome Shotgun Assembler. *Genome Research*, 12(1):177–189, 2002. doi: 10.1101/gr.208902. URL `http://genome.cshlp.org/content/12/1/177.abstract`.

N Cartwright. *Nature's Capacities and Their Measurement*. Oxford University Press, 1994.

A Chakravarti. Population genetics–making sense out of sequence. *Nat Genet*, 21:56–60, 1999.

B Devlin and K Roeder. Genomic control for association studies. *Biometrics*, 55:997–1004, 1999.

P Endicott, M Metspalu, and T Kivisild. *Genetic evidence on modern human dispersals in South Asia: Y chromosome and mitochondrial DNA perspectives*, chapter The Evolution and History of Human Populations in South Asia, pages 229–244. Springer, 2007.

P Green. Phrap documentation, 1996. URL `http://www.phrap.org/phredphrap/phrap.html`. http://www.phrap.org/phredphrap/phrap.html.

D Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

JL Haines, MA Hauser, S Schmidt, WK Scott, LM Olson, P Gallins, KL Spencer, SY Kwan, M Noureddine, JR Gilbert, N Schnetz-Boutaud, A Agarwal, EA Postel, and MA Pericak-Vance. Complement factor H variant increases the risk of age-related macular degeneration. *Science*, 308:419–421, 2005.

X Huang and A Madan. CAP3: A DNA Sequence Assembly Program. *Genome Research*, 9(9):868–877, 1999. doi: 10.1101/gr.9.9.868. URL `http://genome.cshlp.org/content/9/9/868.abstract`.

AJ Iafrate, L Feuk, MN Rivera, ML Listewnik, PK Donahoe, Y Qi, SW Scherer, and C Lee. Detection of large-scale variation in the human genome. *Nat Genet.*, 36:949–952, 2004.

I Ionita, R Daruwala, and B Mishra. Mapping Tumor Suppressor Genes using Multipoint Statistics from Copy-Number Variation Data. *American Journal of Human Genetics*, 79:13–22, 2006.

J Kececioglu and E Myers. Combinatorial algorithms for dna sequence assembly. *Algorithmica*, 13:7–51, 1995.

S Kleinberg and B Mishra. The Temporal Logic of Token Causes. In *Proceedings of the 12th International Conference on the Principles of Knowledge Representation and Reasoning (KR)*, Toronto, Ontario, May 2010. To appear.

S Kleinberg and B Mishra. Metamorphosis: the Coming Transformation of Translational Systems Biology. *Queue*, 7(9):40–52, 2009a. ISSN 1542-7730. doi: http://doi.acm.org/10.1145/1626135.1629775.

S Kleinberg and B Mishra. The Temporal Logic of Causal Structures. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, Montreal, Quebec, June 2009b.

ES Lander. The new genomics: global views of biology. *Science*, 274:536–539, 1996.

M Lynch. Rate, Molecular Spectrum, and Consequences of Human Mutation. *Proc Natl Acad Sci U S A*, 107:961–978, 2010.

ER Martin, EH Lai, JR Gilbert, AR Rogala, AJ Afshari, J Riley, KL Finch, JF Stevens, KJ Livak, BD Slotterbeck, SH Slifer, LL Warren, PM Conneally, DE Schmechel, I Purvis, MA Pericak-Vance, AD Roses, and JM Vance. SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in alzheimer disease. *Am. J. Hum. Genet.*, 67:383–394, 2000.

A Mitrofanova and B Mishra. On a Novel Coalescent Model for Genome-Wide Evolution of Copy Number Variations. *International Journal of Data Mining and Bioinformatics*, 2009.

EW Myers, GG Sutton, AL Delcher, IM Dew, DP Fasulo, MJ Flanigan, SA Kravitz, CM Mobarry, KHJ Reinert, KA Remington, EL Anson, RA Bolanos, H-H Chou, CM Jordan, AL Halpern, S Lonardi, EM Beasley, RC Brandon, L Chen, PJ Dunn, Z Lai, Y Liang, DR Nusskern, M Zhan, Q Zhang, X Zheng, GM Rubin, MD Adams, and JC Venter. A Whole-Genome Assembly of Drosophila. *Science*, 287(5461):2196–2204, 2000. doi: 10.1126/science.287.5461.2196. URL `http://www.sciencemag.org/cgi/content/abstract/287/5461/2196`.

G Narzisi and B Mishra. SUTTA, Scoring-and-Unfolding Trimmed Tree Assembler I: Concepts, Constructs and Comparisons. *Submitted for Publication*, 2010.

PA Pevzner, H Tang, and MS Waterman. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98 (17):9748–9753, 2001. doi: 10.1073/pnas.171285098. URL `http://www.pnas.org/content/98/17/9748.abstract`.

JK Pritchard, M Stephens, and P Donnelly. Linkage disequilibrium in humans: models and data. *Genetics*, 155:945–959, 2000.

D Reich, K Thangaraj, N Patterson, AL Price, and L Singh. Reconstructing Indian population history. *Nature*, 461:489–494, 2009.

AH Samad, W-W Cai, X Hu, B Irvin, J Jing, J Reed, X Meng, J Huang, E Huff, B Porter, A Shenkar, TS Anantharaman, B Mishra, V Clarke, E Dimalanta, J Edington, C Hiort, R Rabbah, J Skiada, and DC Schwartz. Mapping the Genome One Molecule at a Time – Optical Mapping. *Nature*, 378:516–517, 1995.

SC Schuster. Next-generation sequencing transforms today's biology. *Nature Methods*, pages 16–18, 2008.

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

CAM Semple. *Bioinformatics for Geneticists*, chapter Assembling a View of the Human Genome, pages 59–84. Wiley, 2007.

S Shringarpure and EP Xing. mStruct: inference of population structure in light of both genetic admixing and allele mutations. *Genetics*, 182:575–593, 2009.

D Sommer, A Delcher, S Salzberg, and M Pop. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics*, 8(1):64, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-64. URL `http://www.biomedcentral.com/1471-2105/8/64`.

P Suppes. *A probabilistic theory of causality*. North-Holland, 1970.

G. G. Sutton, O. White, M. D. Adams, and A. R. Kerlavage. TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1(1):9–19, 1995.

J West, J Healy, M Wigler, W Casey, and B Mishra. Validation of S. pombe sequence assembly by micro-array hybridization. *Journal of Computational Biology*, 13:1–20, 2006.

DR Zerbino and E Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18 (5):821–829, 2008. doi: 10.1101/gr.074492.107. URL `http://genome.cshlp.org/content/18/5/821.abstract`.