

2008

## Aspects of coverage in medical DNA sequencing

Michael C. Wendl  
*Washington University School of Medicine in St. Louis*

Richard K. Wilson  
*Washington University School of Medicine in St. Louis*

Follow this and additional works at: [https://digitalcommons.wustl.edu/open\\_access\\_pubs](https://digitalcommons.wustl.edu/open_access_pubs)



Part of the [Medicine and Health Sciences Commons](#)

Please let us know how this document benefits you.

---

### Recommended Citation

Wendl, Michael C. and Wilson, Richard K., "Aspects of coverage in medical DNA sequencing." *BMC Bioinformatics*. 9, 239. (2008).  
[https://digitalcommons.wustl.edu/open\\_access\\_pubs/253](https://digitalcommons.wustl.edu/open_access_pubs/253)

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact [vanam@wustl.edu](mailto:vanam@wustl.edu).

Research article

Open Access

## Aspects of coverage in medical DNA sequencing

Michael C Wendl\* and Richard K Wilson

Address: Genome Sequencing Center and Department of Genetics, Washington University, St. Louis MO 63108, USA

Email: Michael C Wendl\* - [mwendl@wustl.edu](mailto:mwendl@wustl.edu); Richard K Wilson - [rwilson@wustl.edu](mailto:rwilson@wustl.edu)

\* Corresponding author

Published: 16 May 2008

Received: 8 October 2007

BMC Bioinformatics 2008, 9:239 doi:10.1186/1471-2105-9-239

Accepted: 16 May 2008

This article is available from: <http://www.biomedcentral.com/1471-2105/9/239>

© 2008 Wendl and Wilson; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** DNA sequencing is now emerging as an important component in biomedical studies of diseases like cancer. Short-read, highly parallel sequencing instruments are expected to be used heavily for such projects, but many design specifications have yet to be conclusively established. Perhaps the most fundamental of these is the redundancy required to detect sequence variations, which bears directly upon genomic coverage and the consequent resolving power for discerning somatic mutations.

**Results:** We address the medical sequencing coverage problem via an extension of the standard mathematical theory of haploid coverage. The expected diploid multi-fold coverage, as well as its generalization for aneuploidy are derived and these expressions can be readily evaluated for any project. The resulting theory is used as a scaling law to calibrate performance to that of standard BAC sequencing at 8× to 10× redundancy, i.e. for expected coverages that exceed 99% of the unique sequence. A differential strategy is formalized for tumor/normal studies wherein tumor samples are sequenced more deeply than normal ones. In particular, both tumor alleles should be detected at least twice, while both normal alleles are detected at least once. Our theory predicts these requirements can be met for tumor and normal redundancies of approximately 26× and 21×, respectively. We explain why these values do not differ by a factor of 2, as might intuitively be expected. Future technology developments should prompt even deeper sequencing of tumors, but the 21× value for normal samples is essentially a constant.

**Conclusion:** Given the assumptions of standard coverage theory, our model gives pragmatic estimates for required redundancy. The differential strategy should be an efficient means of identifying potential somatic mutations for further study.

### Background

Applications of DNA sequencing to medically significant problems continue to grow [1-6]. In particular, recent technological trends suggest that the sequencing of entire cohorts of individual patient genomes will soon be economically feasible [7-13]. This contrasts dramatically with the enormous resources that were expended on deciphering just a single composite human reference genome only

a few years ago [14]. Sequence-based characterization promises to play an expanding role in medicine because of its ability to identify potential disease-causing mutations [1]. It will be especially important in cancers, for example, for distinguishing between sequence variations in the germline versus somatic mutations that are relevant to tumor initiation or growth [1,2].

In principle, the process-engineering issues in both gene-based and whole-genome medical sequencing are identical to those for *de novo* genomic sequencing, that is, to "cover" a region of interest with shotgun read data. However, the definitions of what constitutes coverage are rather different. In traditional genomic sequencing, the target is a haploid genome and coverage of a base position  $x$  is defined as the event whereby one or more sequence reads span  $x$ . Such a process is binomial and, according to elementary probability theory, the expected fractional coverage is  $1 - \exp(-\rho)$ , where  $\rho = NL/G$ . Here,  $L$  and  $G$  are the read and haploid genome lengths, respectively,  $N$  is the number of reads sequenced, and  $\rho$  is the haploid redundancy. Although this result describes a number of traditional coverage configurations [15], it seems to be known to the sequencing community primarily via its application by Clarke and Carbon [16]. This expression is also sometimes attributed to Lander-Waterman Theory (LWT) [17], although LWT actually treats the issue of sequence gaps rather than coverage.

Medical sequencing projects focus on genetic variation and seek to identify both alleles at  $x$  for the diploid genome. In particular, diploid sequence is necessary for discerning heterozygous mutations. Consequently, coverage is thought of in a more general way. Here, we say that  $x$  is "covered" when each allele is spanned by at least  $\phi$  reads, where  $\phi \geq 1$ . Actual values of  $\phi$  will depend upon study-specific considerations that weigh economic factors against such things as desired confidence levels for detection and confirmation, anticipated data quality, etc. Some results on multiple coverings appear in the mathematical literature [18,19], but these do not address the problem beyond the haploid level. Smith and Bernstein [20] conducted early numerical simulations for  $\phi = 1$  on a 20 kb fragment, but evidently did not extend the approach to genome-size targets. Levy et al. [12] and Wheeler et al. [13] also describe models for this problem, which we discuss further below.

An important issue for future medical sequencing projects can be posed as follows. Given a specific choice of  $\phi$ , estimate the necessary redundancy such that either the probability of covering a given position, e.g. a SNP, has some desired value, or that the expectation for the number of captured positions has such a value. These propositions are actually identical (see Methods). However, additional study-specific issues also arise. For example, for tumor/germline pairs, one has to specify  $\rho$  for both types of samples. As we demonstrate below, the two values should not necessarily be the same.

Speculation regarding these issues has been around for some time. For example, Strausberg et al. [1] observed that  $\rho$  should exceed 10. In other words, redundancies for

medical sequencing projects should surpass those values conventionally associated with haploid whole-genome shotgun projects, BAC projects, etc. This is largely intuitive, given the diploid nature of the problem, but not particularly informative. Pioneering diploid sequencing projects furnish some early anecdotal information. For example, Levy et al. [12] considered  $\rho = 20$  to be adequate for germline sequencing of a healthy individual based upon simulation, certain heuristic filters, and the model alluded to above. They employed traditional Sanger sequencing [21] and reached only about 7.5x, so the degree to which this value generalizes to medical sequencing of cancer genomes using short-read "next-generation" platforms [8] is unclear. Likewise, Wheeler et al. [13] report only about 7.4x for another diploid project.

Here, we address medical sequencing coverage more formally by way of a straightforward mathematical extension to the standard covering process model. We consider this an idealization in the sense that it presumes all entities are independently and identically distributed (IID) and neglects any heuristic inputs. However, we also demonstrate the use of empirical data to calibrate response such that inferences can be drawn for medical sequencing projects. The resulting analysis points to what we believe will be an efficient means of discerning potential somatic mutations and enables estimation of the necessary parameters.

**Results**

Given a location  $x$  defined in the context of  $h$  associated chromosomes, let  $P_{h,\phi}$  be the probability that  $x$  is covered at least  $\phi$  times. The immediate focus of much of the research community is on diploid sequencing of homologous chromosomes ( $h = 2$ ) related to the cancers, for which we report a mathematical theory of coverage. In anticipation of extending sequencing to aneuploid configurations, some of which are also relevant to cancer, we furnish the general result for  $h > 2$ , as well.

**Diploid Sequencing Theory**

Given a diploid genome ( $h = 2$ ), the probability of coverage is

$$P_{2,\phi} = \sum_{j=\phi}^{N-\phi} C_{N,j} \delta_2^j (1 - \delta_2)^{N-j} \left[ 1 - \sum_{k=0}^{\phi-1} C_{N-j,k} \delta_2^k (1 - \delta_2)^{N-j-k} \right], \tag{1}$$

where  $\delta_2 = L/(2G)$  is the diploid Bernoulli probability and  $C_{N,k}$  are the binomial coefficients. Eq. 1 also gives the expected fraction of a set of locations that are covered (Methods). This equation relies on the standard IID assumption, but is exact in the sense that it accounts for the fact that the coverings of two corresponding alleles on

homologous chromosomes are not strictly independent of one another (Methods). However, parameters in an actual project are such that alleles are *almost* independent. Moreover, asymptotic approximation can be applied (Methods), in which case

$$P_{2,\phi} \approx \left( 1 - \sum_{k=0}^{\phi-1} \frac{1}{k!} \left( \frac{\rho}{2} \right)^k e^{-\rho/2} \right)^2 \quad (2)$$

is a very good approximation of Eq. 1. Here,  $e$  is the Euler Number ( $\approx 2.71828$ ) and  $\rho$  is again the conventional haploid redundancy. Note the basis in a Poisson distribution having a rate  $\rho/2$ . Eq. 2 is straightforward to evaluate for any project because  $\phi$  is typically not very large. This stands in contrast to Eq. 1, which sports an enormous number of terms, as well as tendencies for numerical overflow and underflow of its various components. For convenience, we expand the first three expressions

$$P_{2,1} \approx \left( 1 - e^{-\rho/2} \right)^2 \quad (3)$$

$$P_{2,2} \approx \left( 1 - e^{-\rho/2} \left( 1 + \frac{\rho}{2} \right) \right)^2 \quad (4)$$

$$P_{2,3} \approx \left( 1 - e^{-\rho/2} \left( 1 + \frac{\rho}{2} + \frac{\rho^2}{8} \right) \right)^2. \quad (5)$$

**Generalization to Aneuploidy**

Under the assumption of independence, Eq. 2 for homologous chromosomes is readily generalized to an arbitrary number of chromosomes,  $h$ , specifically

$$P_{h,\phi} \approx \left( 1 - \sum_{k=0}^{\phi-1} \frac{1}{k!} \left( \frac{\rho}{h} \right)^k e^{-\rho/h} \right)^h. \quad (6)$$

Note the Poisson basis having a rate  $\rho/h$ , for example  $\rho/3$  for chromosomal trisomies. Like Eq. 2, this expression is readily evaluated and straightforward to expand for given values of  $\phi$ .

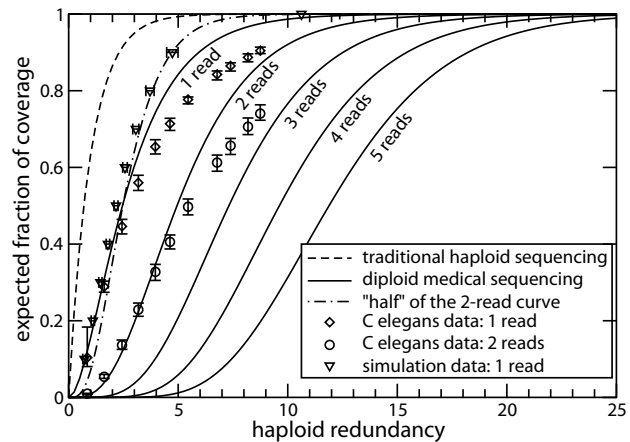
**Discussion**

A number of medical sequencing coverage issues are currently being debated. New questions have arisen not only because diploid medical sequencing is itself a fairly recent undertaking, but also because of the expectation that novel sequencing platforms will be heavily employed in such projects. Read lengths are substantially shorter than traditional Sanger data [21] and investigators are eager to determine how this affects coverage. We focus our discussion here primarily on the diploid problem, although

some projections for aneuploid configurations are given, as well.

**Coverage Assessment**

Fig. 1 shows the traditional *de novo* haploid coverage model [15,16] versus diploid medical sequencing coverage theory for minimum read coverings of  $\phi \in \{1, 2, 3, 4, 5\}$  for both alleles. The diploid curves were generated by Eq. 2 for a 3.3 billion base-pair genome and 31 base-pair read lengths. Errors associated with not using Eq. 1 for these particular parameters are significantly less than 1% (data not shown). As one would intuitively expect, the required redundancies for a given coverage fraction increase with  $\phi$  and are noticeably higher than established values for haploid genome sequencing. Although these cases are within the realm of feasibility for the newest-generation sequencing platforms [11], economic factors would probably still preclude the higher values of  $\phi$  at the present time. Conversely, these depths are much lower than values that have been discussed elsewhere. For example, Warren et al. [22] report  $\rho$  up to 100 and 400 for bac-



**Figure 1**  
**Traditional haploid coverage model [15, 16] versus diploid medical sequencing coverage results for minimum number of covering reads  $\phi \in \{1, 2, 3, 4, 5\}$ .** The figure also shows an additional curve that replots the diploid  $\phi = 2$  curve, except where abscissa values are scaled by one-half. This aspect is relevant to the discussion of why the redundancies for  $\phi = 1$  and  $\phi = 2$  do not differ by a factor of two. Coverage progressions for  $\phi \in \{1, 2\}$  are also shown for the recent Illumina resequencing of *C. elegans* by Hillier et al. [28]. These points represent average coverages over all chromosome pairs, while their error bars show the observed minima and maxima. Simulation data for  $\phi = 1$  on a 20 kb fragment using 250 bp reads [20] are also shown. Points and error bars represent the averages and extrema, respectively, of 250 simulations.

terial and viral genomes, respectively, using 25 bp fragments to simulate data from an Illumina instrument.

Two other notable trends are visible in Fig. 1. First, increasingly large redundancies are required just to obtain non-trivial values of  $P_{2, \phi}$ . For example, the curve for  $\phi = 1$  already exceeds 0.01 at  $\rho = 0.25$ , whereas this mark is not met until  $\rho = 5$  at  $\phi = 5$ . Indeed, one may not see even the "beginnings" of coverage until comparatively high redundancy has been reached, depending on the selected  $\phi$ . Also, the amount by which each curve is drawn-out over the abscissa increases with  $\phi$ , signifying a decelerating coverage rate. This is especially clear for what appear to be the linear segments of each curve; their slopes progressively decrease. Again comparing the extremes, the difference between 0.1 and 0.9 on the ordinate is a little over 5 units of redundancy for  $\phi = 1$  but almost 11 units for  $\phi = 5$ . This phenomenon bears on point we make below.

Both of these trends arise strictly as mathematical consequences and can perhaps best be understood by referring to Eqs. 3 through 5. The exponential (Euler Number) term represents the tendency for the coverage rate to decay. For each successive value of  $\phi$  this term is bolstered by additional factors, which themselves grow progressively faster with  $\rho$ , whereby the overall effect is realized.

**Calibration and the Stopping Problem**

One of the primary issues facing the investigator is the so-called stopping problem. That is, at what  $\rho$  should random processing be halted? This question is, of course, context dependent. Yet, it can be answered, at least approximately, by using the analysis given here. For example, suppose the goal is to design a medical sequencing project such that the expected coverage progress corresponds roughly to standard BAC sequencing. This is a calibration-based way of framing the question and exploits the community's collective empirical experience gained from having sequenced hundreds of thousands of such clones. In particular,  $6 \leq \rho \leq 10$  has been found to be a reasonable balance between cost and coverage, although values nearer to 10 are more typically chosen [14]. In this capacity, Eqs. 1 and 2 effectively function as scaling laws.

Scaling can conveniently be demonstrated graphically, for example by picking a point on the haploid curve for a desired redundancy, drawing a horizontal line through this point, and reading the redundancy at the intersection of the chosen diploid curve and the horizontal. The asymptotic nature of the curves depicted in Fig. 1 obscures this process, but it can readily be accomplished using a magnified plot. Fig. 2 shows the example of extrapolating haploid sequencing coverage at  $\rho = 8$  to diploid sequencing ( $\phi = 1$ ), the result being about 17.5x for the diploid project. Table 1 furnishes an expanded set of values for  $\phi$

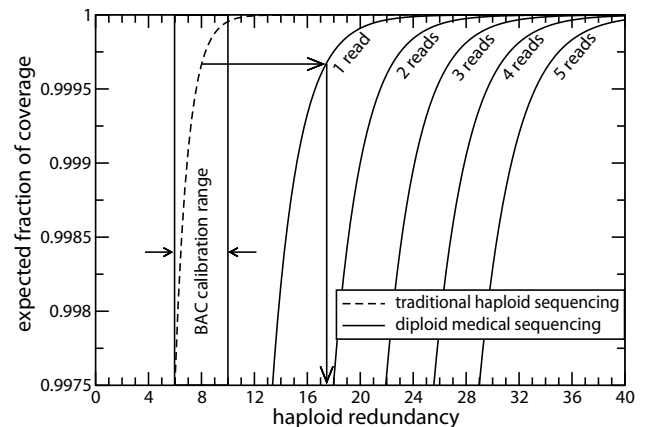
**Table 1: Calibration of medical sequencing according to traditional haploid expectation**

Traditional [16]		Corresponding Medical Sequencing $\rho$		
Redundancy	P-value	1-read min.	2-read min.	3-read min.
6	0.99752	13.5	18	22
8	0.99967	17.5	22.5	26.5
10	0.99996	21.5	26.5	31.5

$\in \{1, 2, 3\}$  calibrated against haploid sequencing for  $\rho \in \{6, 8, 10\}$ . Again, the increase of redundancy with the minimum number of reads required to attain coverage is quite clear. Notice that each of the three rows in the table corresponds to covering more than 99% of the unique sequence. In other words, the covering probabilities change very little over fairly significant increases in depth. Consequently, BAC depth provides much better resolution than BAC coverage for scaling the diploid problem. This observation is also obvious from Fig. 2.

**Comparison to Haploid-Based Distribution Models**

Diploid coverage, as discussed above, is a primary consideration for medical sequencing. Yet, it is also useful for comparison to examine such projects in their haploid context. The stopping problem has been extensively studied from a number of analytical perspectives for traditional *de novo* genomic sequencing projects, for example using the probability of complete coverage,  $P_C$  [23] and the intersection probability,  $P_{\cap}$  [24]. While the meaning

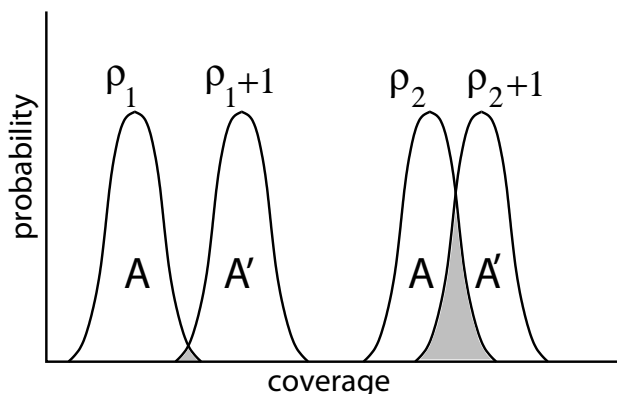


**Figure 2**  
**Haploid and diploid results for expected coverage values of at least 0.9975.** This is a greatly – magnified view of the top quarter – percent of the ordinate range in Fig. 1. Vertical lines demarcate the typical BAC calibration neighborhood of  $6 \leq \rho \leq 10$ . The scaling process is demonstrated graphically for diploid sequencing ( $\phi = 1$ ) based on haploid sequencing at  $\rho = 8$ .

of the former is probably clear, the latter characterizes how effective additional redundancy will be for improving coverage in light of increasingly-important stochastic effects. Of these two metrics,  $P_C$  is the more conservative and could be thought of as setting an upper bound in the context of traditional haploid sequencing. Before proceeding, let us digress briefly to further explain  $P_{\cap}$ .

Consider two hypothetical sequencing projects,  $A$  and  $A'$ , that are identical in every respect, except that  $A'$  is always ahead by one whole unit of redundancy (Fig. 3). Now, examine these projects at two particular instances, specifically for project  $A$  at both  $\rho_1$  and  $\rho_2$  (with  $A'$  at  $\rho_1 + 1$  and  $\rho_2 + 1$ , respectively), where  $\rho_1 < \rho_2$ . At  $\rho_1$ , there may be little overlap in the densities between the two projects. Accordingly, the probability is extremely high, perhaps close to 1, that  $A'$  will be more highly covered than  $A$ . The system behaves as if there is a deterministic increase in coverage for  $A \rightarrow A'$  for a unit increase of redundancy. At higher redundancies, say  $\rho_2$ , mathematical analysis indicates that the intersection of the  $A$  and  $A'$  densities will grow [24]. Consequently, differences in actual coverage between  $A$  and  $A'$  become progressively more a function of chance than of differences in the redundancies themselves. The tail value of the intersection,  $P_{\cap}$ , can be taken as an indicator on the diminishing returns of the process.

We can once again take a calibration approach to this problem. That is, we calculate  $P_C$  and  $P_{\cap}$  for a typical BAC sequencing project (150 kb clone length and 600 bp read length) at  $\rho \in \{6, 8, 10\}$ . We then match these values to their counterparts in the distributions for medical sequencing, which then provides the corresponding "scaled" redundancy. Table 2 shows these results. The val-



**Figure 3**  
**Diagrammatic synopsis of the intersection probability.** Paired coverage distributions, plotted at differences of one unit of redundancy, begin to coalesce as a project evolves. The intersection probability is the area of the overlap (shaded).

**Table 2: Calibration of medical sequencing according to haploid distribution models**

Traditional BAC Sequencing Project		Medical Sequencing $\rho$ based on		
Redundancy	$P_C$	$P_{\cap}$	Complete Covg.	Intersection
6	0.02843	0.74122	20.2	18.8
8	0.51842	0.95054	22	20.9
10	0.89840	0.99301	23.9	23

ues compliment those in Table 1 in the sense that they suggest redundancies far above the conventional full shotgun standard of  $\rho = 10$ .

**Comparison to Empirical, Semi-Empirical, and Simulation Results**

Several labs are now involved in diploid sequencing projects, which should furnish useful examples of coverage progressions that can be monitored empirically. Although the few stopping redundancies reported in the literature appear to conflict with one another, these can be more properly interpreted according the minimum number of times each allele is observed,  $\phi$ . For example, we mentioned above that Levy et al. [12] considered  $\rho = 20$  for Sanger-based [21] germline sequencing of a healthy individual. This figure approaches the 10x standard for haploid sequencing [25] at the level of  $\phi = 1$  (Table 1). Interestingly, the idealized version of their coverage calculation proves to be a special case of our model precisely for  $\phi = 1$  (see Appendix). Conversely, Mardis [8] quotes a redundancy up to 30x, which corresponds to values of  $\phi$  between 2 and 3 when calibrated to haploid 10x.

Richard Durbin and Aylwyn Scally have also analyzed the diploid medical sequencing coverage problem using a different approach from what is described here (Durbin and Scally, personal communication). Specifically, they employed an "extra-variation" Poisson distribution [26,27] having a free-parameter to control variance. Values for this parameter can be chosen to *a posteriori* tune the theoretical fit with empirical data. In particular, such tuning allows one to implicitly consider, at least approximately, factors such as bias and sequencing errors. (In our method, calibration incorporates the empirics of BAC sequencing, essentially serving the same purpose.) Using their semi-empirical approach, Durbin and Scally concluded that redundancies closer to 30x will be required, which again agrees well with results shown in Tables 1 and 2.

A number of labs, including our own, are now adopting "next generation" short-read sequencing technology [8] and have started to generate human medical sequencing data related to various cancers. However, there is still a

dearth of published results from which actual coverage progressions can be derived. For the purposes of comparison, we refer instead to the recently completed pilot resequencing project for *C. elegans*. Hillier et al. [28] resequenced strain N2 Bristol using the Illumina Genome Analyzer in order to characterize the accuracy and utility of short-read, massively parallel data. We have projected their *C. elegans* coverage results for  $\phi \in \{1, 2\}$  onto Fig. 1. Agreement is very good up to about 60% coverage, after which the rate of empirical coverage falls below expectation. This behavior seems to typify theoretical-empirical differences. For example, Wendl and Barbazuk [29] noted precisely this trend for sequencing filtered genomes.

The physical explanation of this phenomenon is straightforward. Specifically, biases are not manifested early in a project because there is not enough information to distinguish unbiased coverage configurations from biased ones. (Think of an extreme case, for example placing a single read of high GC content onto a genome of high AT content. Despite the obvious non-IID nature of this scenario, the predicted coverage will still be *identical* to the actual coverage.) Given a model based on the IID assumption, as ours is, empirical and theoretical results should start to diverge as sufficient information gathers to expose latent biases. In this case, Hillier et al. [28] note a definite AT bias using the Illumina platform, i.e. remarkably lower coverage in regions of high AT content, which we presume accounts for much of the difference shown in Fig. 1. The proclivities of other methods and platforms are evidently different [30]. Consequently, Eq. 2 should also be useful as a yardstick for comparison among these approaches for specific applications.

Finally, Fig. 1 also shows simulation data reported by Smith and Bernstein [20] for  $\phi = 1$  on a 20 kb circularized fragment using 250 bp reads. Agreement is once again good up to about 60% coverage, after which the sequencing process seems to grow more efficient for the fragment. This observation is not surprising, given two important aspects of this study. First, the circularized configuration is not subject to the so-called "edge effect", which can dramatically affect coverage rates [29,31]. Second, distribution theories show that configurations having larger  $L/G$  ratios do indeed cover more readily than those having smaller values [24]. We presume these two factors account for most of the difference, especially given that  $L/G = 0.0125$  for the simulation is more than a million times larger than values associated with short-read medical sequencing projects. For example, 31 bp reads on a 3.3 billion bp genome yields  $L/G \approx 1 \times 10^{-8}$ .

#### **A Differential Sequencing Strategy**

We expect that many future studies will be based on sequencing DNA derived from matched tumor/normal

samples (for example, the latter being obtained from uninvolved skin or blood) from the same patient [3,4]. Here, the whole genome of each sample in a pair is sequenced and mutations are found by comparison to the human reference. Let us call the sets of mutations for a tumor and a normal sample  $S_T$  and  $S_N$ , respectively, where we generally expect  $S_N \subseteq S_T$ . Most of the germline variation in  $S_N$  will be polymorphisms not related to pathogenesis [1], whereas  $S_T$  will contain a potentially more relevant collection of somatic mutations. In principle, germline sequence variations can be removed from further consideration by taking the difference  $S_T - S_N$  [6]. Such filtering will appreciably focus subsequent work, since the overwhelming majority of sequence variants should be polymorphisms found in normal tissue [5]. How does one efficiently accomplish this from a process-engineering standpoint?

We propose a refinement of simple subtraction [6] in the form of a straightforward differential sequencing strategy. In principle, false-negative errors are controlled by sequencing at least to diploid coverage at the level of  $\phi = 1$ . However, tumor samples should actually be sequenced as heavily as economically possible in order to minimize false-positive hits for both germline and somatic mutations. These types of mistakes arise, for example, by misinterpreting a random sequencing error as a true mutation. Given current state-of-the-art capabilities, we will assume this condition translates to diploid coverage at the level of  $\phi = 2$ , but emphasize that future instruments will undoubtedly permit higher  $\phi$ .

Conversely, a germline mutation in a normal sample only has to be detected once in order to be eliminated from  $S_T$ . We are also not as concerned about false-positives here because their appearance in  $S_N$  does not affect the subtraction  $S_T - S_N$ . It is possible that an error could lead to a spurious entry in  $S_N$  that precisely matches a true somatic mutation in  $S_T$  by pure chance. The somatic mutation would then be erroneously eliminated from further investigation. However, such events seem unlikely, given the low anticipated number of bona fide somatic mutations. These observations collectively imply that normal samples may only need to be sequenced to diploid coverage at the level of  $\phi = 1$ .

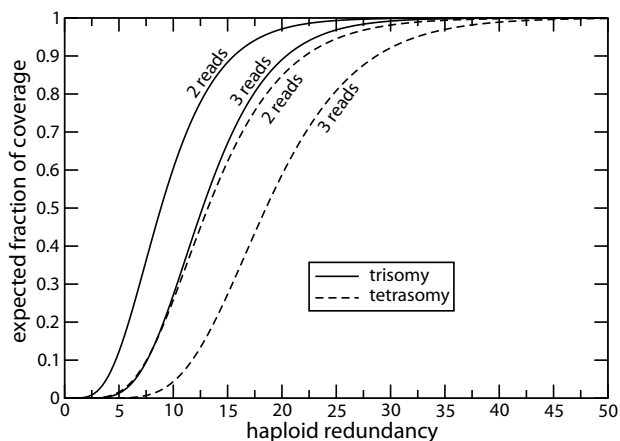
The 10 $\times$  standard for BAC sequencing is well-established [25] and provides a reasonably conservative basis to translate the above design into actual redundancies for medical sequencing (Table 1). We suggest then that sequencing of tumor samples should not be pursued to less than about 26.5 $\times$  redundancy, given the 2-read minimum coverage condition. Furthermore, paired normal samples need only to be sequenced to about 21.5 $\times$  redundancy for the  $\phi = 1$  coverage condition.

### Expository Comment on Recommended Redundancies

The observation that the required redundancy for the  $\phi = 1$  coverage level is not simply half that of the  $\phi = 2$  level may initially seem counter-intuitive. We remarked above that curves for increasing values of  $\phi$  tend to have progressively smaller slopes. Consequently, there is not, in general, an integer-valued relationship between corresponding points on any two particular curves. In other words, curves are not simply shifted along  $\rho$ . Returning to Fig. 1, we show an example that replots the  $\phi = 2$  curve, except where the abscissa-value of each of its points has been divided by two. It is clear that the result does not coincide with the  $\phi = 1$  curve, as intuition may have suggested. The curves only intersect at a single point, here at an expected coverage slightly more than 0.5, which is well below the  $> 0.99$  calibration points we chose above. In other words, redundancy for  $\phi = 1$  would only have been half that for  $\phi = 2$  had we chosen to cover 50% instead of  $> 0.99$  of SNPs. This trend holds generally. That is, we do not expect an integer relationship between the required redundancies for two unequal, but otherwise arbitrary values of  $\phi$ .

### Coverage Projections for Aneuploid Configurations

Aneuploidy can be manifested in a number of ways: as an autosomal [32] or sex chromosome [33] aberration, and in conjunction with cancer [34]. We anticipate the eventual application of DNA sequencing to aneuploid chromosome configurations and offer some early projections based upon Eq. 6. Fig. 4 shows expected coverage for trisomy and tetrasomy for  $\phi = 2$  and  $\phi = 3$ . Required depths are clearly much higher than for diploid sequencing. For example, we find redundancies of  $\rho = 42$  and  $\rho = 57$  for trisomy and tetrasomy, respectively, when scaling to  $10\times$  BAC sequencing at the level of  $\phi = 2$ . Recall that  $\phi = 2$  is



**Figure 4**  
Expected coverage for aneuploid chromosome configurations for minimum number of covering reads  $\phi \in \{2, 3\}$ .

presumed to be feasible for diploid whole-genome sequencing using current hardware. These redundancies are clearly out of reach at the moment for a whole-genome project, but may be feasible for chromosome-specific projects. In other words, the appreciably higher cost of sequencing aneuploid chromosomes may justify the effort of separating them into their own self-contained projects.

### Modeling Limitations

As with the classical theories of sequencing [16,17], the main assumption here is that reads are independently and identically distributed (IID). In other words, this analysis does not formally consider biological or instrument-specific biases, software biases and sequencing errors, for example in base-calling, assembly problems, or any other heuristic inputs. The idiosyncrasies of each of these factors are difficult to characterize analytically, although the calibration step does allow some implicit accounting, as noted above. Appreciable differences in levels and types of bias have been noted, for instance in Sanger-style sequencing versus pyrosequencing [30], so any results should be interpreted with these qualifications in mind.

In general, the assumption of allele independence should be valid for most medical sequencing projects since  $\phi$  will be small and  $L/G \rightarrow 0$  and  $N \gg 1$  (see Methods). For example, maximum error for  $\phi \in \{1, 2\}$  is on the order of  $10^{-5}$  percent for diploid sequencing using 650 bp read lengths. The theory further assumes that sequence reads have no preference for either chromosome of a homologous pair and neglects any tendency for reads to align to multiple positions. The latter has been found to occur with some frequency if reads are short enough [35]. Read-pairing certainly curtails this phenomenon, but the pairing process itself has negligible effect on coverage unless the target is very small [31]. Such is not the case in whole-genome medical sequencing, so the net effect of pairing is simply that the amount of uniquely-alignable sequence one gets to count toward  $\rho$  increases commensurately. In plain terms, fewer data will be discarded.

Finally, our analysis does not account for what might be called the "uneven coverage" problem of alleles. Mutation detection programs may decline calling out a SNP if one allele is covered much more heavily than the other [36]. Because this phenomenon is both software-specific and sequence-specific, it is beyond our scope. Departure from any of these idealizations will tend to reduce coverage, implying that our analysis is best viewed in the context of upper bounds of performance. In other words, required redundancies for specific projects may still exceed what we have advocated here, as the *C. elegans* data in Fig. 1 illustrate.



A subtle mathematical point is also worth mentioning. Eqs. 1 through 5 represent the probability of covering a specific allele pair, or alternatively, the expected fraction of pairs covered. These expressions do not provide the underlying distribution of the number of covered pairs, which is a more formidable mathematical problem. In other words, this is a model only of coverage expectation, exactly analogous to what classical theories [16,17] are for traditional *de novo* haploid sequencing. Consequently, the results themselves are not strong functions of  $L/G$ . In fact, for most applications the results will be completely independent of this ratio and will instead follow a set of "universal" curves, the first 5 of which are shown in Figs. 1 and 2. This point is underscored by Eq. 2, which is a function strictly of  $\rho$ . (The observation holds more generally for aneuploidy as described by Eq. 6, as well.) This phenomenon contrasts with distribution-based models, such as  $P_C$  and  $P_\cap$  discussed above, which are indeed sensitive to  $L/G$ . The basis of this effect is discussed in ref. [24].

A corollary to this observation is that adjusting the fundamental parameters within their biologically-relevant limits will have no effect on the results we have discussed. For example, the haploid genome size  $G$  could be adjusted to reflect only that part of the sequence to which read data can be uniquely aligned [35]. Yet, the underlying assumptions leading to Eqs. 2 and 6 will still be satisfied in this circumstance, mainly  $L/G \rightarrow 0$  and  $N \gg 1$  (Methods). The same holds for varying  $L$  in order to represent different kinds of sequencing platforms, e.g. pyrosequencing or Sanger instruments. In summary, the contributions of the three independent variables  $L$ ,  $G$ , and  $N$  collapse into the single dimensionless variable  $\rho$ , which governs the process exclusively. Formal theory [37] predicts such systematic reductions of variables whenever a unified dimensionless parameter lurks in a problem.

**Conclusion**

The differential sequencing strategy should be useful for efficiently identifying lists of somatic mutations for validation and further study. Our analytical model of coverage, coupled with a calibration approach for selecting parameters, allows pragmatic estimates to be made for such projects. However, because the theory does not strictly consider various biasing factors, actual projects would benefit from periodically aligning (assembling) shotgun data to empirically track overall coverage, as well as local coverage in coding regions, UTRs, promoters, and conserved regions. SNP arrays could also be done for each sample, with attempts made to find and correlate data to sequence calls for further coverage tracking. Plotting these various data on a single figure, as we did for the *C. elegans* data in Fig. 1, should be informative. Finally, the basic model could be further extended in the future as more data accrue from different methods, projects, software

processing pipelines, etc. For example, "extra-variation" methods [26,27] could be used for *a posteriori* data fitting, the results of which should help to better quantify non-IID factors.

**Methods**

Proofs of Eqs. 1, 2, and 6 are reported here. Eqs. 3 through 5 follow trivially from Eq. 2. We also describe the analysis of *C. elegans* resequencing data from Hillier et al. [28].

**Preliminaries**

Let  $B_{i,j}$  be the event where an allele at position  $x$  on chromosome  $i$  is "covered", i.e. spanned by at least  $\phi$  out of any collection of  $j$  reads, where  $j \geq \phi$ . Given  $N$  total reads, our definition of diploid medical sequencing coverage for position  $x$  is then  $B_{1,N} \cap B_{2,N}$  and its probability is  $P_{2,\phi}(B_{1,N} \cap B_{2,N})$ . If  $\beta_{i,j,k}$  is the event whereby the allele on chromosome  $i$  is spanned by exactly  $k$  of  $j$  reads, then  $B_{i,j} \equiv \beta_{i,j,\phi} \cup \beta_{i,j,\phi+1} \cup \dots \cup \beta_{i,j,j}$ . Considering two homologous chromosomes,  $i \in \{1, 2\}$ , the probability that a single given read spans  $x$  on a specific chromosome is  $\delta_2 = L/(2G)$ , where  $L$  and  $G$  are read length and haploid genome length, respectively. Since the process is binomial (covering or not covering), we immediately have  $P(\beta_{i,j,k}) = C_{j,k} \delta_2^k (1 - \delta_2)^{j-k}$ , where  $C_{j,k}$  are the binomial coefficients.

**Proof of Eq. 1**

The coverings of two homologous alleles are not independent of one another. For instance, if one allele is already covered by  $j$  reads, there are only  $N - j$  remaining reads that have a chance to cover the other allele. Consequently,

$$P_{2,\phi}(B_{1,N} \cap B_{2,N}) = P(B_{1,N}) \cdot P(B_{2,N} | B_{1,N}) = \sum_{j=\phi}^N P(\beta_{1,N,j}) \cdot P(B_{2,N-j} | \beta_{1,N,j}),$$

where

$$P(B_{2,N-j} | \beta_{1,N,j}) = \sum_{k=\phi}^N P(\beta_{2,N-j,k}) = 1 - \sum_{k=0}^{\phi-1} P(\beta_{2,N-j,k}).$$

Eq. 1 follows from the observation that  $P(\beta_{2,N-j,k}) = 0$  for  $k > N - j$ .

**Proof of Eq. 2**

If we neglect the dependence of alleles, then  $P_{2,\phi}(B_{1,N} \cap B_{2,N}) = P(B_{1,N}) \cdot P(B_{2,N})$ . Without loss of generality, this probability is identical to  $P^2(B_{1,N})$ , from which

$$P_{2,\phi} = \left( 1 - \sum_{k=0}^{\phi-1} C_{N,k} \delta_2^k (1 - \delta_2)^{N-k} \right)^2$$

Parameters for medical sequencing projects are such that  $L/G \rightarrow 0$  and  $N \gg 1$ . These relations hold for both "short read" platforms and instruments that provide "full-length" Sanger read data. Moreover,  $\phi \ll N$ , which implies  $k \ll N$ . Consequently, the binomial coefficients  $C_{N,k}$  are well-approximated by  $N^k/k!$ . These conditions also imply  $(1 - \delta_2)^N \sim \exp(-N\delta_2)$ , i.e. that asymptotic approximation can be used for the power term. Eq. 2 follows directly.

**Proof of Eq. 6**

The case of aneuploidy under the assumption of allele independence is a straightforward extension of the proof for Eq. 2. For  $h$  homologous chromosomes,  $P_{h,\phi}(B_{1,N} \cap B_{2,N} \cap \dots \cap B_{h,N}) = P(B_{1,N})P(B_{2,N}) \dots P(B_{h,N}) = P^h(B_{1,N})$ . Given that all  $h$  chromosomes are equally likely to be sampled,  $\delta_h = L/(hG)$ , which is the appropriate Bernoulli probability for  $P(B_{1,N})$ . Eq. 6 follows from the same approximation arguments made for Eq. 2.

**Eqs. 1, 2, and 6 in the Context of Expectation**

We can take the coverage status of a specific allele pair as a Bernoulli trial, whereby elementary probability theory shows that the expected number of pairs covered is their total number multiplied by  $P_{2,\phi}(B_{1,N} \cap B_{2,N})$ . Consequently,  $P_{2,\phi}(B_{1,N} \cap B_{2,N})$  in Eq. 1 and its approximation in Eq. 2 also represent the expected fraction of covered pairs. The same argument holds for Eq. 6.

**Analysis of C. elegans resequencing data**

Hillier et al. [28] used the Illumina Genome Analyzer to resequence the *C. elegans* N2 Bristol genome. Release ws188 of the genomic sequence [38] was downloaded from <http://www.wormbase.org> and randomly chosen subsets of the resequence data were aligned against the reference at regular intervals for each chromosome using the maq aligner (<http://maq.sourceforge.net>). Data that could not be uniquely placed on the reference were discarded. Coverage was calculated for each alignment as the number of corresponding base positions spanned by at least one read on homologous chromosomes ( $\phi = 1$ ) and by at least two reads on homologous chromosomes ( $\phi = 2$ ).

**Appendix: Idealized Theory of Levy et al**

Levy et al. [12] sketch a rudimentary diploid theory, though they do not furnish any corresponding mathematical description. Here, we reconstruct an idealized version of their model, i.e. the form which assumes all entities are IID and which omits any heuristic inputs. A careful reading of "Modeling False-Negative Rate of Heterozygous

Variants" in ref. [12] reveals the following salient features. Chromosomes are equally likely to be sampled and loci are taken as independent of one another. (Our theory relies on these same two assumptions.) Levy et al. also assume the number of reads  $\nu$  spanning a position of interest  $x$  is Poisson-distributed with a rate  $\rho$  and that the probability of observing both alleles is a binomial function of  $\nu$ . Incidentally, Richard Durbin and Aylwyn Scally discuss a similar model in their analysis (Durbin and Scally, personal communication), as do Wheeler et al. [13].

Let the random variables **B** and **N** be the events where both alleles at  $x$  are observed (covered) and where  $\nu$  reads span  $x$ , respectively. We immediately have

$$P(\mathbf{N} = \nu) = \frac{e^{-\rho} \rho^\nu}{\nu!} \tag{7}$$

from the Poisson assumption. Given  $\nu$  reads spanning  $x$ , the probability of observing both alleles is simply the complement of the probability of any configuration in which one of the alleles is *not* represented among the  $\nu$  reads. If we label the alleles **I** and **II**, then without loss of generality, the binomial model for the number of observations,  $j$ , of allele **I** is

$$P_{\mathbf{I}}(j) = C_{\nu,j} \left(\frac{1}{2}\right)^j \left(1 - \frac{1}{2}\right)^{\nu-j} = C_{\nu,j} \left(\frac{1}{2}\right)^\nu \tag{8}$$

There are two configurations in which only one of the alleles is observed:  $j = 0$  (all reads hit allele **II**) and  $j = \nu$  (all reads hit allele **I**). Consequently, the probability of observing both alleles in  $\nu$  reads is  $P(\mathbf{B} | \mathbf{N} = \nu) = 1 - P_{\mathbf{I}}(0) - P_{\mathbf{I}}(\nu)$ . Using Eq. 8, a little algebra shows

$$P(\mathbf{B} | \mathbf{N} = \nu) = 1 - \left(\frac{1}{2}\right)^{\nu-1} \tag{9}$$

which is defined for  $\nu \geq 1$ . Note that the probability exceeds zero only for  $\nu \geq 2$ , as we would expect. That is, at least 2 reads must span  $x$  before it is possible to observe both alleles. The Theorem of Total Probability now furnishes the desired result,  $P(\mathbf{B})$ , from Eqs. 7 and 9, as follows.

$$P(\mathbf{B}) = \sum_{\nu} P(\mathbf{B} | \mathbf{N} = \nu) P(\mathbf{N} = \nu) = \sum_{\nu=2}^{\infty} \left[ 1 - \left(\frac{1}{2}\right)^{\nu-1} \right] \frac{e^{-\rho} \rho^\nu}{\nu!} \tag{10}$$

This expression represents the ideal probability of covering a diploid location as a function of the haploid sequence redundancy of the project.

Eq. 10 is actually just a special case of our model in Eq. 2 for  $\phi = 1$ , as the following exercise demonstrates.

$$P(\mathbf{B}) = \sum_{v=0}^{\infty} \left[ 1 - \left(\frac{1}{2}\right)^{v-1} \right] \frac{e^{-\rho} \rho^v}{v!} - \sum_{v=0}^1 \left[ 1 - \left(\frac{1}{2}\right)^{v-1} \right] \frac{e^{-\rho} \rho^v}{v!} \tag{11}$$

$$= \sum_{v=0}^{\infty} \frac{e^{-\rho} \rho^v}{v!} - \sum_{v=0}^{\infty} \frac{e^{-\rho} \rho^v}{v! 2^{v-1}} - (1-2)e^{-\rho} \tag{12}$$

$$= 1 - 2e^{-\rho/2} \sum_{v=0}^{\infty} \frac{e^{-\rho/2} (\rho/2)^v}{v!} + e^{-\rho} \tag{13}$$

$$= 1 - 2e^{-\rho/2} + e^{-\rho} \tag{14}$$

$$P(\mathbf{B}) = \left(1 - e^{-\rho/2}\right)^2, \tag{15}$$

which is nothing more than  $P_{2,1}$  in Eq. 3. It can be shown along very similar lines that the Poisson/Binomial model outlined by Wheeler et al. [13] is also a special case of Eq. 2 for  $\phi = 2$ , i.e. it leads to  $P_{2,2}$  (Eq. 4).

### Authors' contributions

Both authors framed the original problem. MCW conceived and constructed the mathematical theory and wrote the paper. Both authors read and approved the final manuscript.

### Acknowledgements

The germ of this work evolved during the "Monday-morning cancer discussions" organized by Richard Wilson and Elaine Mardis of the Genome Sequencing Center and Timothy Ley of the Department of Medicine, Washington University. The authors wish especially to thank Elaine Mardis, Timothy Ley, and Li Ding for their input and critical comments and Brian Dunford-Shore for analyzing the *C. elegans* data and generating the information for plotting the empirical results in Fig. 1. They also appreciate general discussions of genome coverage involving Ken Chen, Jarret Glasscock, Michael McLellan II, Ryan Richt, and Todd Wylie. This work was partially supported by grant HG003079 from the National Human Genome Research Institute (Richard K. Wilson, PI).

### References

1. Strausberg RL, Simpson AJG, Wooster R: **Sequence-Based Cancer Genomics: Progress, Lessons and Opportunities.** *Nature Reviews Genetics* 2003, **4(6)**:409-418.
2. Ley TJ, Minx PJ, Walter MJ, Ries RE, Sun H, McLellan M, DiPersio JF, Link DC, Tomasson MH, Graubert TA, McLeod H, Khoury H, Watson M, Shannon W, Trinkaus K, Heath S, Vardiman JW, Caligiuri MA, Bloomfield CD, Milbrandt JD, Mardis ER, Wilson RK: **A Pilot Study of High-Throughput, Sequence-Based Mutational Profiling of Primary Human Acute Myeloid Leukemia Cell Genomes.** *Proceedings of the National Academy of Sciences* 2003, **100(24)**:14275-14280.
3. Wilson RK, Ley TJ, Cole FS, Milbrandt JD, Clifton S, Fulton L, Fewell G, Minx P, Sun H, McLellan M, Pohl C, Mardis ER: **Mutational Pro-**

**filing in the Human Genome.** *Cold Spring Harbor Symposia on Quantitative Biology* 2003, **68**:23-29.

4. Rand V, Huang J, Stockwell T, Ferriera S, Buzko O, Levy S, Busam D, Li K, Edwards JB, Eberhart C, Murphy KM, Tsiamouri A, Beeson K, Simpson AJG, Venter JC, Riggins GJ, Strausberg RL: **Sequence Survey of Receptor Tyrosine Kinases Reveals Mutations in Glioblastomas.** *Proceedings of the National Academy of Sciences* 2005, **102(40)**:14344-14349.
5. Ma PC, Zhang X, Wang ZJ: **High-Throughput Mutational Analysis of the Human Cancer Genome.** *Pharmacogenomics* 2006, **7(4)**:597-612.
6. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JKV, Gazdar AF, Hartigan J, Wu L, Liu CS, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE: **The Consensus Coding Sequences of Human Breast and Colorectal Cancers.** *Science* 2006, **314(5797)**:268-274.
7. Kling J: **The Search for a Sequencing Thoroughbred.** *Nature Biotechnology* 2005, **23(11)**:1333-1335.
8. Mardis ER: **The Impact of Next-Generation Sequencing Technology on Genetics.** *Trends in Genetics* 2008, **24(3)**:133-141.
9. Service RF: **The Race for the \$1000 Genome.** *Science* 2006, **311(5767)**:1544-1546.
10. Bentley DR: **Whole-Genome Re-Sequencing.** *Current Opinion in Genetics & Development* 2006, **16(6)**:545-552.
11. Check E: **Faster Still and Faster.** *Nature* 2007, **448(7149)**:10-11.
12. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AWC, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC: **The Diploid Genome Sequence of an Individual Human.** *PLoS Biology* 2007, **5(10)**:article no. e254.
13. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The Complete Genome of an Individual by Massively Parallel DNA Sequencing.** *Nature* 2008, **452(7189)**:872-876.
14. International Human Genome Sequencing Consortium: **Initial Sequencing and Analysis of the Human Genome.** *Nature* 2001, **409(6822)**:860-921.
15. Robbins HE: **On the Measure of a Random Set.** *Annals of Mathematical Statistics* 1944, **15(1)**:70-74.
16. Clarke L, Carbon J: **A Colony Bank Containing Synthetic Col El Hybrid Plasmids Representative of the Entire E. coli Genome.** *Cell* 1976, **9(1)**:91-99.
17. Lander ES, Waterman MS: **Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis.** *Genomics* 1988, **2(3)**:231-239.
18. Glaz J, Naus J: **Multiple Coverage of the Line.** *Annals of Probability* 1979, **7(5)**:900-906.
19. Holst L: **On Multiple Covering of a Circle with Random Arcs.** *Journal of Applied Probability* 1980, **17(1)**:284-290.
20. Smith GD, Bernstein KE: **BULLET: A Computer Simulation of Shotgun DNA Sequencing.** *Computer Applications in the Biosciences* 1995, **11(2)**:155-157.
21. Sanger F, Nicklen S, Coulson AR: **DNA Sequencing with Chain-Terminating Inhibitors.** *Proceedings of the National Academy of Sciences* 1977, **74(12)**:5463-5467.
22. Warren RL, Sutton GG, Jones SJM, Holt RA: **Assembling Millions of Short DNA Sequences Using SSAKE.** *Bioinformatics* 2007, **23(4)**:500-501.
23. Wendl MC, Waterston RH: **Generalized Gap Model for Bacterial Artificial Chromosome Clone Fingerprint Mapping and Shotgun Sequencing.** *Genome Research* 2002, **12(12)**:1943-1949.
24. Wendl MC: **Occupancy Modeling of Coverage Distribution for Whole Genome Shotgun DNA Sequencing.** *Bulletin of Mathematical Biology* 2006, **68(1)**:179-196.
25. Wilson R, Ainscough R, Anderson K, Baynes C, Berks M, Burton J, Connell M, Bonfield J, Copsey T, Cooper J, Coulson A, Craxton M, Dear S, Du Z, Durbin R, Favello A, Fraser A, Fulton L, Gardner A, Green P, Hawkins T, Hillier L, Jier M, Johnston L, Jones M, Kershaw J, Kirsten J, Laisster N, Latreille P, Lloyd C, Mortimore B, Ocallaghan M,

- Parsons J, Percy C, Rifken L, Roopra A, Saunders D, Shownkeen R, Sims M, Smaldon N, Smith A, Smith M, Sonnhammer E, Staden R, Sulston J, Thierry-Mieg J, Thomas K, Vaudin M, Vaughan K, Waterston R, Watson A, Weinstock L, Wilkinson-Sproat J, Wohldman P: **2.2 Mb of Contiguous Nucleotide Sequence from Chromosome III of *C. elegans***. *Nature* 1994, **368(6466)**:32-38.
26. Feller W: *An Introduction to Probability Theory and Its Applications* 3rd edition. New York NY: John Wiley & Sons; 1968.
  27. Breslow NE: **Extra-Poisson Variation in Log-Linear Models**. *Applied Statistics* 1984, **33(1)**:38-44.
  28. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JL, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK, Mardis ER: **Whole-Genome Sequencing and Variant Discovery in *C. elegans***. *Nature Methods* 2008, **5(2)**:183-188.
  29. Wendl MC, Barbazuk VB: **Extension of Lander-Waterman Theory for Sequencing Filtered DNA Libraries**. *BMC Bioinformatics* 2005, **6**:article no. 245.
  30. Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N: **454 Sequencing Put to the Test Using the Complex Genome of Barley**. *BMC Genomics* 2006, **7**:Art No. 275.
  31. Wendl MC: **A General Coverage Theory for Shotgun DNA Sequencing**. *Journal of Computational Biology* 2006, **13(6)**:1177-1196.
  32. Nicolaidis P, Petersen MB: **Origin and Mechanisms of Non-Disjunction in Human Autosomal Trisomies**. *Human Reproduction* 1998, **13(2)**:313-319.
  33. Cammarata M, DiSimone P, Graziano L, Giuffrè M, Corsello G, Garofalo G: **Rare Sex Chromosome Aneuploidies in Humans: Report of Six Patients with 48, XXYY, 49, XXXXY, and 48, XXXX Karyotypes**. *American Journal of Medical Genetics* 1999, **85(1)**:86-87.
  34. Virtaneva K, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri MA, Bloomfield CD, de la Chapelle A, Krahe R: **Expression Profiling Reveals Fundamental Biological Differences in Acute Myeloid Leukemia with Isolated Trisomy 8 and Normal Cytogenetics**. *Proceedings of the National Academy of Sciences* 2001, **98(3)**:1124-1129.
  35. Whiteford N, Haslam N, Weber G, Prügel-Bennett A, Essex JW, Roach PL, Bradley M, Neylon C: **An Analysis of the Feasibility of Short Read Sequencing**. *Nucleic Acids Research* 2005, **33(19)**:article no. e171.
  36. Chen K, McLellan MD, Ding L, Wendl MC, Kasai Y, Wilson RK, Mardis ER: **PolyScan: An Automatic Indel and SNP Detection Approach to the Analysis of Human Resequencing Data**. *Genome Research* 2007, **17(5)**:659-666.
  37. Barenblatt GI: **Dimensional Analysis**. New York NY: Gordon and Breach; 1987.
  38. *C. elegans* Sequencing Consortium: **Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology**. *Science* 1998, **282(5396)**:2012-2018.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

