2007

# Algebraic correction methods for computational assessment of clone overlaps in DNA fingerprint mapping

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

BMC Bioinformatics

BioMed Central

Research article

Open Access

# Algebraic correction methods for computational assessment of clone overlaps in DNA fingerprint mapping

Michael C Wendl*

Address: Genome Sequencing Center, Washington University, St. Louis MO 63108, USA

Email: Michael C Wendl* - mwendl@wustl.edu

* Corresponding author

Published: 18 April 2007

*BMC Bioinformatics* 2007, **8**:127 doi:10.1186/1471-2105-8-127

Received: 2 March 2007

Accepted: 18 April 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/127



## Abstract

**Background:** The Sulston score is a well-established, though approximate metric for probabilistically evaluating postulated clone overlaps in DNA fingerprint mapping. It is known to systematically over-predict match probabilities by various orders of magnitude, depending upon project-specific parameters. Although the exact probability distribution is also available for the comparison problem, it is rather difficult to compute and cannot be used directly in most cases. A methodology providing both improved accuracy and computational economy is required.

**Results:** We propose a straightforward algebraic correction procedure, which takes the Sulston score as a provisional value and applies a power-law equation to obtain an improved result. Numerical comparisons indicate dramatically increased accuracy over the range of parameters typical of traditional agarose fingerprint mapping. Issues with extrapolating the method into parameter ranges characteristic of newer capillary electrophoresis-based projects are also discussed.

**Conclusion:** Although only marginally more expensive to compute than the raw Sulston score, the correction provides a vastly improved probabilistic description of hypothesized clone overlaps. This will clearly be important in overlap assessment and perhaps for other tasks as well, for example in using the ranking of overlap probabilities to assist in clone ordering.

## Background

Fingerprint mapping continues to play an important role in large-scale DNA sequencing efforts [1-5]. The procedure is challenging in terms of both its laboratory and computational demands. Indeed, most of the computational steps involve non-trivial algorithmic aspects. While reasonable solutions have been found for many of these, one task that remains particularly problematic is assessing postulated clone overlaps based on their fingerprint similarity.

The "overlap problem", as this is often referred to, basically involves examining all pairwise clone comparisons in order to identify overlaps. For a map consisting of $\lambda$ clones, there are $C_{\lambda,2} = \lambda(\lambda - 1)/2$ such comparisons. In each one, the number of matching fragment lengths between the two associated fragment lists is established. A case having $\mu > 0$ matches indicates a possible overlap because the mutual length(s) may represent the same DNA. Lengths are not unique, so such matches are not conclusive indicators of overlap. Instead, the problem is largely one of probabilistic classification. One or more quantitative metrics are used to evaluate the authenticity

of each such case. For example, an apparent overlap might be judged against its likelihood $\alpha$ of arising by chance. Methodologies of varying degrees of rigor have been proposed for this task [6-11]. However, the so-called Sulston score, or Sulston probability $P_S$ has emerged as a *de facto* standard [12], in part because of its integration in the widely-used FPC program [13,14]. A liability of a number of these methodologies, including $P_S$, is they assume fragment length comparisons are independent when, in fact, they are not [10,15].

Recently, the exact distribution characterizing the overlap problem was determined [16,17]. Comparisons reveal that the assumption of independence is usually a poor one and that the Sulston score systematically over-predicts actual overlap probabilities, often by orders of magnitude. Consequently, a bias arises in projects that utilize $P_S$ (Table 1). One chooses the significance threshold $\alpha$ to minimize erroneous decisions according to what is presumed to be the actual probabilistic description of the problem, $P_E$. The alternative result using the Sulston score is an overall increase of false-negatives (Case 1). Clones having significant overlap will still be correctly detected (Case 3). Moreover, false-positives would not be increased because $P_S$ errs on the conservative side with respect to non-overlapping clones (Case 6). Miscalls can obviously be expected when poor values of $\alpha$ are chosen (Cases 4 and 5). However, if $\alpha$ is set too high, there will still be circumstantial cases where the correct decision is made (Case 2). These will presumably be more than offset by a higher rate of false-positives (Case 4). In summary, $P_S$ is not an especially good discriminant for the overlap assessment problem.

The drawback of $P_E$ is that it is rather difficult to compute and cannot be used directly in most cases. For example, current resources are not sufficient to evaluate it for most BAC comparisons or for capillary-based fingerprinting [18]. A suitable method of approximating $P_E$ is therefore required. Here, we propose a straightforward correlation-based approach that derives correction factors for the Sulston score. This procedure dramatically increases accuracy without incurring much additional computational effort.

## Results

The overlap problem is formally cast in terms of two clones having $m$ and $n$ "bands", respectively, where $m \geq n$. Each band represents an individual clone fragment, with its position on a gel image providing an estimate of the fragment's length. Multiple bands of roughly the same length often appear. Finite measurement resolution $\pm R$ allows an image of length $L$ to be subdivided into $t = 0.5\ L/R$ discrete bins. The Sulston score $P_S = P_S(\mu, m, n, t)$ is taken as a *provisional* estimate of the probability that at least $\mu$ fragment matches between the two clones arise by chance. Note here that the variables $(\mu, m, n)$ correspond to (M, nH, nL), respectively, in notations used by the FPC program [14]. The corresponding exact probability is $P_E = P_E(\mu, m, n, t)$, as given in refs. [16,17]. We formulate a corrected value, $P_C$, that can be both efficiently calculated and that gives substantially better estimates of $P_E$ than the Sulston score, i.e. $|P_E - P_C| << |P_E - P_S|$.

The simple log-log plot in Fig. 1 shows good correlation (Pearson's coefficient of $\rho \approx 0.9938$), suggesting that standard regression might be a reasonable basis for correction. Note the characteristic over-prediction of $P_S$. (Points representing the exact probability consistently fall below the hypothetical line of agreement between $P_S$ and $P_E$.) These particular data are computed for $t = 236$, which describes traditional settings for fragment length measurements and comparisons, i.e. ± 7 pixels over a 3300 pixel gel image [13,19]. Considerations of coverage usually dictate a large number of clones in a map [2], so that values substantially above $10^{-7}$ are not usually of interest [20]. The data range over $0 \leq \mu \leq n$ for a number of different fingerprint comparison sizes: $2 \leq n \leq m$ for $5 \leq m \leq 12$, $2 \leq n \leq 10$ for $m$ of 13 and 14, $2 \leq n \leq 9$ for $m$ of 15 and 16, $2 \leq n \leq 8$ for $m$ of 17 and 18, and finally $2 \leq n \leq 7$ for $19 \leq m \leq 25$. The exact solution becomes difficult to evaluate beyond these ranges using readily-available resources. Specifically, the computational effort increases according to a factor that exceeds $m!/(m - n)!$ [17].

Correlation in Fig. 1 is clearly not perfect. Specifically, the points show some amount of lateral scatter. Accuracy of the correction can be further enhanced to the degree that

**Table 1: Types of decisions for the biased Sulston score**

| Case | Scores | Overlap | Tuning of $\alpha$ | Decision Based on $P_S$ |
|---|---|---|---|---|
| 1 | $P_E < \alpha < P_S$ | Yes | Correct | Wrong (False-Negative) |
| 2 | $P_E < \alpha < P_S$ | No | Too High | Circumstantially Correct |
| 3 | $P_E < P_S < \alpha$ | Yes | Correct | Correct |
| 4 | $P_E < P_S < \alpha$ | No | Too High | Wrong |
| 5 | $\alpha < P_E < P_S$ | Yes | Too Low | Wrong |
| 6 | $\alpha < P_E < P_S$ | No | Correct | Correct |

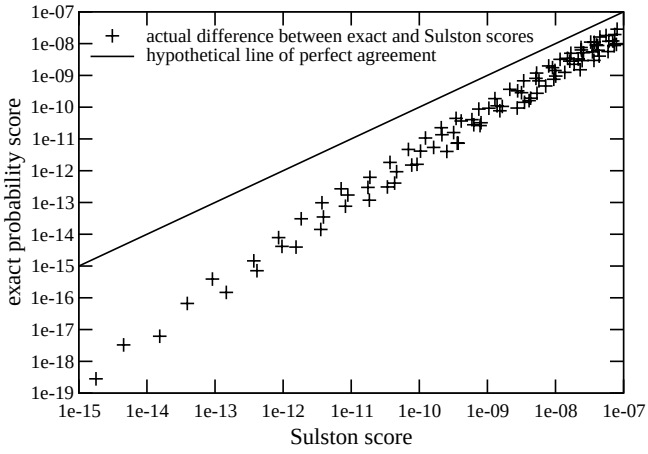


**Figure 1**
Sampling of exact probability versus Sulston score for $t$ = 236.

dispersion within the window can be minimized. Here, we can apply a simple power-law data reduction model to obtain a *transformed* Sulston score

$$P_T = P_S^{\nu} \mu^{\xi} m^{\eta} n^{\zeta}. \quad (1)$$

The four power values can be chosen empirically such that the data locally collapse into a more highly correlated set. For example, selecting $(\nu, \xi, \eta, \zeta) = (1.2,4,0.8,-3.4)$ in Eq. 1 leads to the curve-fit

$$P_C \approx 9.855\, P_T^{1.171} \quad (2)$$

and the associated Pearson's coefficient $\rho \approx 0.9980$.

## Discussion

We propose Eqs. 1 and 2 as a correction to the standard Sulston score for typical fingerprint mapping conditions [13,19]. Although shown as two separate equations so as to clarify the concept, these can clearly be combined into a single equation for actual computations. Pearson's coefficient is not especially sensitive to the parameters in Eq. 1 and there are many combinations of $(\nu, \xi, \eta, \zeta)$ that elevate $\rho$ into the ~0.998 range. Other methods for reducing the data do not perform as well as the model in Eq. 1. For example, standard dimensional analysis [21], which involves correlating variables such as $P_E/P_S$, $m/n$, and $\mu/n$, cannot adequately resolve the fact that values of the individual variables relative to one another remain important.

### *Accuracy assessment*

Eqs. 1 and 2 are obviously straightforward to compute, leaving the question of just how much error reduction is actually realized over the un-adjusted Sulston score. This can be quantified with a simple metric. For the Sulston score, the error is taken as $E_S = |P_E - P_S|/P_E$. Error for the corrected result, $E_C$, is calculated similarly.

A size-selection step is part of most library-construction protocols, meaning that the variance of clone sizes will be limited to some degree. Consequently, many clone-clone comparisons will involve similar, though not necessarily equal numbers of fragments. Fig. 2 shows a comparison of error rates for the raw Sulston score and the corrected score in Eq. 2 for $m/n \leq 1.3$. That is, we compare clones whose numbers of fragments in their respective fingerprints are within 30% of one another. The figure also shows the error rate for the un-reduced data, i.e. for a regression equation that does not use the preliminary processing given by Eq. 1.

The Sulston score shows an increasing error as the acceptance threshold is tightened (lowered). Maximum values for the threshold are typically in the neighborhood of $10^{-7}$ [20], for which $P_S$ over-predicts by about one order of magnitude. For threshold parameters around $10^{-19}$, Sulston over-prediction is about 4 orders of magnitude. While Eq. 2 shows significant local variation, the overall trend is much more constant and its error is appreciably smaller. Correction on un-reduced data also shows better accuracy than the raw Sulston score, being roughly as good as Eq. 2 up to about $10^{-12}$. It diverges beyond this point and eventually shows about the same level of error as the raw Sulston score. The combined correction procedure of Eqs. 1 and 2 appears to provide the best fidelity over the widest range.

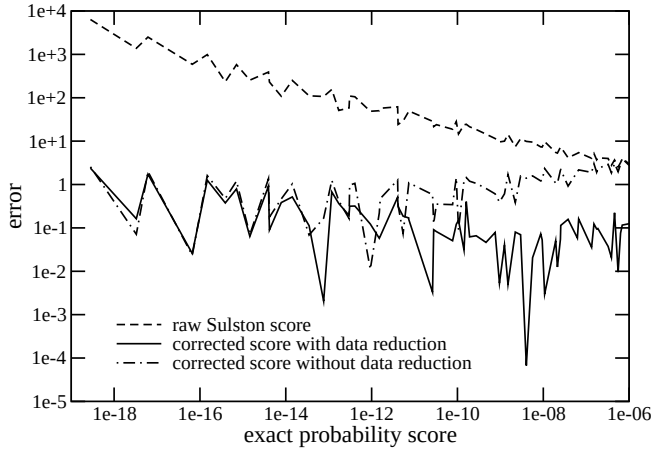


**Figure 2**
Error characterization for clones with similar numbers of fingerprint bands.

### *Comments on uncertainty*

A simple correction of the type we propose here obviously cannot capture all the complexities inherent in the exact distribution. This results in a scatter of the data that cannot be completely eliminated, for example as illustrated in Fig. 1. This scatter is a primarily function of $m/n$, rather than individual values of $m$ and $n$. For instance, log-log regression of data restricted exclusively to $m = n$ returns a Pearson's coefficient of $\rho \approx 0.9998$ without *any* sort of preliminary data reduction. Of course, such a correlation would not be generally applicable to realistic clone libraries and maps.

Eq. 2 is based on the limited set of data described above. Applying it outside this set necessarily involves a degree of extrapolation, which raises two types of uncertainty. First, large $m/n$ ratios contribute to scatter, but such extrema only emerge for cases involving sufficiently large differences between $m$ and $n$. Eq. 2 accounts for data up to $m = 25$ with a maximum ratio of $m/n$ of about 4. In the context of averages, this implies a comparison of two clones whose sizes differ by a factor of four. While there is the possibility of even greater disparities, such cases will be comparatively rare in general because of size-selection steps executed during the library-construction phase. For example, in the Human Genome Project RPCI-11 library, about two-thirds of the BAC clones were concentrated between 150 and 200 kb [22], for which the maximum $m/n$ would be roughly 1.3. Most comparisons would be somewhat closer to one. Only about 2.5% of the library resided in each of the < 100 kb and > 250 kb ranges. This means that fewer than 0.1% of the comparisons will involve uncharacteristically large $m/n$ ratios. Consequently, we do not view this type of uncertainty as being particularly significant.

The larger issue in our opinion arises for comparisons that extend beyond (lower than) the $10^{-19}$ threshold tolerance. While minor extrapolation of a few orders of magnitude is probably not worrisome, some projects utilize substantially lower tolerances. For example, Luo *et al*. [18] and Nelson *et al*. [23] report values on the order of $10^{-30}$ and $10^{-45}$, respectively, when using capillary electrophoresis. Other techniques, such as the traditional double-digest, can also generate higher numbers of fragments, which may require reduced thresholds. The fidelity of Eq. 2 for such cases is not clear. For example, in the data set shown in Fig. 1, larger $m/n$ values are under-represented at the lowest scores. Because loci for larger $m/n$ values do not slope as steeply as those for smaller ones, the trend shown in the figure may not continue in the exact same manner for values well below $10^{-19}$. We can only observe that the corrected score will still be the significantly more accurate choice as compared to the raw Sulston score because the assumption of independent fragment comparisons is increasingly untenable. Characterizing the exact solution in this range requires computations considerably larger than what can readily be made at present.

## Conclusion

We have calibrated Eq. 2 according to the traditional parameters used in the FPC mapping program [13]. Similar corrections can readily be constructed for different parameters. For example, protocols and software sizing methods now allow for band resolutions higher than the customary value of $t = 236$. Table 2 shows correction parameters for several such cases. Similarity of the correlation coefficients suggests that results would be comparable to that shown in Fig. 2. Although the accuracies derived from this approach are probably acceptable in the correlation range, they could, in principle, be further increased by using multiple corrections calibrated for specific "bins" of the $m/n$ parameter.

Clone overlap assessment is sometimes framed as a statistical testing problem [10]. Here, $\alpha$ is the probability of erroneously concluding that two clones overlap, when in fact they do not. (This casually implies that $\alpha\, C_{\lambda,\,2}$ false positives can be expected for a project containing $\lambda$ clones.) Consequently, corrections are most immediately relevant in the neighborhood surrounding $\alpha$ (Table 1). The overlaps here are the most valuable to detect in the sense that they are the smallest, and consequently contribute most effectively to a minimum tiling path [8]. A large fraction of the comparisons will be either far above or below the threshold, so their assessments will not ultimately be affected. However, correction is still important for these cases. For example, Branscomb *et al*. [8] have

**Table 2: Correction parameters for various gel resolutions (bin numbers)**

| bins | data reduction (Eq. 1) | | | | fit (Eq. 3) | | correlation |
|---|---|---|---|---|---|---|---|
| $t$ | $v$ | $\xi$ | $\eta$ | $\zeta$ | $\beta$ | $\phi$ | $\rho$ |
| 236 | 1.2 | 4 | 0.8 | -3.4 | 9.855 | 1.171 | 0.9980 |
| 300 | 1.2 | 4.2 | 0.7 | -3.2 | 5.070 | 1.144 | 0.9982 |
| 350 | 1.4 | 4.2 | 0.8 | -3.2 | 4.908 | 0.956 | 0.9982 |
| 400 | 1.4 | 4.2 | 0.7 | -3.2 | 5.711 | 0.944 | 0.9983 |

pointed out that the ability to accurately rank all overlaps according to their associated probabilities is useful in the assembly phase of mapping.

Ascertaining the degree to which a particular mapping project would actually be improved by using Sulston score correction is difficult. Aside from the usual factors that complicate comparisons [24], there are special considerations for this kind of evaluation. For example, established Sulston-based mapping projects may have obtained their best results using threshold values that would not necessarily be considered "correct" from the standpoint of the exact probability distribution (Table 1). Biologists have historically viewed selection of the Sulston threshold to be a non-trivial, library-dependent problem and often resort to empirical sampling and iteration [25,26]. Consequently, one probably cannot obtain an objective comparison by just replacing $P_S$ with $P_C$ for these cases. Another avenue, perhaps more pragmatic, would be to assess corrections on a simulated project. For example, digesting finished sequences *in silico* [27] enables one to use the resulting simulated fingerprints to see how well a map could be reconstructed. Several variations on this method are possible [28,29]. Of course, use of correction for new projects is certainly recommended.

Other issues remain unresolved. With the exception of the conditional nature of match trials, the correction in Eq. 2 is based on the same set of assumptions as the Sulston score. Neither consider, for example, possible non-IID distribution of fragment lengths or length-dependent measurement accuracy. Consequently, we feel that the simple correction procedure proposed here represents a reasonable, though admittedly provisional advance in DNA mapping methodology.

## Methods

Parameters in Eq. 1 were chosen empirically to minimize dispersion (maximize Pearson's coefficient) over a scoring range of roughly $10^{-7}$ to $10^{-19}$. The former is often the maximum value used in a mapping project and is dictated by the need to limit false-positive overlap declarations for the associated libraries, which are typically quite large [20]. The latter is set by computational limitations.

Correction of a probability score $P_T$ is implemented as a so-called "power-law" algebraic expression

$$P_C = \beta\, P_T^{\phi}, \tag{3}$$

where $\phi$ and $\beta$ are regression constants. Eq. 3 can be transformed into log-log form as

$$\ln P_C = \ln \beta + \phi \ln P_T. \tag{4}$$

Standard linear regression [30] can be used to determine $\phi$ and $\beta$ in this equation. Specifically, we analyze the transformed system $(x', y') = (\ln P_T, \ln P_C)$ to obtain the slope $s$ and y-intercept $y_o$ of the straight-line equation $y' = sx' + y_o$. The desired correction in Eq. 3 is then recovered by substituting $\phi = s$ and $\beta = \exp(y_o)$.

## Acknowledgements

The author is grateful to Dr. John Wallis of Washington University for discussions of DNA mapping and its associated computations.

## References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang HM, Yu J, Wang J, Huang GY, Gu J, Hood L, Rowen L, Madan A, Qin SZ, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan HQ, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JGR, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang WH, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz JR, Slater G, Smit AFA, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ: **Initial Sequencing and Analysis of the Human Genome.** *Nature* 2001, **409(6822):**860-921.
2. McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK, Fulton R, Kucaba TA, Wagner-McPherson C, Barbazuk WB, Gregory SG, Humphray SJ, French L, Evans RS, Bethel G, Whittaker A, Holden JL, McCann OT, Dunham A, Soderlund C, Scott CE, Bentley DR, Schuler G, Chen HC, Jang WH, Green ED, Idol JR, Maduro VVB, Montgomery KT, Lee E, Miller A, Emerling S, Kucherlapati R, Gibbs R, Scherer S, Gorrell JH, Sodergren E, Clerc-Blankenburg K, Tabor P, Naylor S, Garcia D, de Jong PJ, Catanese JJ, Nowak N, Osoegawa K, Qin SZ, Rowen L, Madan A, Dors M, Hood L, Trask B, Friedman C, Massa H, Cheung VG, Kirsch IR, Reid T, Yonescu R, Weissenbach J, Bruls T, Heilig R, Branscomb E, Olsen A, Doggett N, Cheng JF, Hawkins T, Myers RM, Shang J, Ramirez L, Schmutz J, Velasquez O, Dixon K, Stone NE, Cox DR, Haussler D, Kent WJ, Furey T, Rogic S, Kennedy S, Jones S, Rosenthal A, Wen GP, Schilhabel M, Gloeckner G, Nyakatura G, Siebert R, Schlegelberger B, Korenburg J, Chen XN, Fujiyama A, Hattori M, Toyoda A, Yada T, Park HS, Sakaki Y, Shimizu N, Asakawa S, Kawasaki K, Sas-

aki T, Shintani A, Shimizu A, Shibuya K, Kudoh J, Minoshima S, Ramser J, Seranski P, Hoff C, Poustka A, Reinhardt R, Lehrach H: **A Physical Map of the Human Genome.** *Nature* 2001, **409(6822):**934-941.
3. Gregory SG, Sekhon M, Schein J, Zhao SY, Osoegawa K, Scott CE, Evans RS, Burridge PW, Cox TV, Fox CA, Hutton RD, Mullenger IR, Phillips KJ, Smith J, Stalker J, Threadgold GJ, Birney E, Wylie K, Chinwalla A, Wallis J, Hillier L, Carter J, Gaige T, Jaeger S, Kremitzki C, Layman D, Maas J, McGrane R, Mead K, Walker R, Jones S, Smith M, Asano J, Bosdet I, Chan S, Chittaranjan S, Chiu R, Fjell C, Fuhrmann D, Girn N, Gray C, Guin R, Hsiao L, Krzywinski M, Kutsche R, Lee SS, Mathewson C, McLeavy C, Messervier S, Ness S, Pandoh P, Prabhu AL, Saeedi P, Smailus D, Spence L, Stott J, Taylor S, Terpstra W, Tsai M, Vardy J, Wye N, Yang G, Shatsman S, Ayodeji B, Geer K, Tsegaye G, Shvartsbeyn A, Gebregeorgis E, Krol M, Russell D, Overton L, Malek JA, Holmes M, Heaney M, Shetty J, Feldblyum T, Nierman WC, Catanese JJ, Hubbard T, Waterston RH, Rogers J, de Jong PJ, Fraser CM, Marra M, McPherson JD, Bentley DR: **A Physical Map of the Mouse Genome.** *Nature* 2002, **418(6899):**743-750.
4. Krzywinski M, Wallis J, Gösele C, Bosdet I, Chiu R, Graves T, Hummel O, Layman D, Mathewson C, Wye N, Zhu B, Albracht D, Asano J, Barber S, Brown-John M, Chan S, Chand S, Cloutier A, Davito J, Fjell C, Gaige T, Ganten D, Girn N, Guggenheimer K, Himmelbauer H, Kreitler T, Leach S, Lee D, Lehrach H, Mayo M, Mead K, Olson T, Pandoh P, Prabhu AL, Shin H, Tänzer S, Thompson J, Tsai M, Walker J, Yang G, Sekhon M, Hillier L, Zimdahl H, Marziali A, Osoegawa K, Zhao S, Siddiqui A, de Jong PJ, Warren W, Mardis E, McPherson JD, Wilson R, Hübner N, Jones S, Marra M, Schein J: **Integrated and Sequence-Ordered BAC and YAC-Based Physical Maps for the Rat Genome.** *Genome Research* 2004, **14(4):**766-779.
5. Wallis JW, Aerts J, Groenen MA, Crooijmans RP, Layman D, Graves TA, Scheer DE, Kremitzki C, Fedele MJ, Mudd NK, Cardenas M, Higginbotham J, Carter J, McGrane R, Gaige T, Mead K, Walker J, Albracht D, Davito J, Yang SP, Leong S, Chinwalla A, Sekhon M, Wylie K, Dodgson J, Romanov MN, Cheng H, de Jong PJ, Osoegawa K, Nefedov M, Zhang H, McPherson JD, Krzywinski M, Schein J, Hillier L, Mardis ER, Wilson RK, Warren WC: **A physical map of the chicken genome.** *Nature* 2004, **432(7018):**761-4.
6. Coulson A, Sulston J, Brenner S, Karn J: **Toward a Physical Map of the Genome of the Nematode *Caenorhabditis elegans*.** *Proc Natl Acad Sci U S A* 1986, **83(20):**7821-7825.
7. Olson MV, Dutchik JE, Graham MY, Brodeur GM, Helms C, Frank M, MacCollin M, Scheinman R, Frank T: **Random-Clone Strategy for Genomic Restriction Mapping in Yeast.** *Proc Natl Acad Sci U S A* 1986, **83(20):**7826-7830.
8. Branscomb E, Slezak T, Pae R, Galas D, Carrano AV, Waterman M: **Optimizing Restriction Fragment Fingerprinting Methods for Ordering Large Genomic Libraries.** *Genomics* 1990, **8(2):**351-366.
9. Balding DJ, Torney DC: **Statistical Analysis of DNA Fingerprint Data for Ordered Clone Physical Mapping of Human Chromosomes.** *Bulletin of Mathematical Biology* 1991, **53(6):**853-879.
10. Nelson DO, Speed TP: **Statistical Issues in Constructing High Resolution Physical Maps.** *Statistical Science* 1994, **9(3):**334-354.
11. Siegel AF, Roach JC, van den Engh G: **Expectation and Variance of True and False Fragment Matches in DNA Restriction Mapping.** *Journal of Computational Biology* 1998, **5:**101-111.
12. Sulston J, Mallett F, Staden R, Durbin R, Horsnell T, Coulson A: **Software for Genome Mapping by Fingerprinting Techniques.** *Computer Applications in the Biosciences* 1988, **4:**125-132.
13. Soderlund C, Longden I, Mott R: **FPC: A System for Building Contigs from Restriction Fingerprinted Clones.** *Computer Applications in the Biosciences* 1997, **13(5):**523-535.
14. Soderlund C, Humphray S, Dunham A, French L: **Contigs Built with Fingerprints, Markers, and FPC V 4.7.** *Genome Research* 2000, **10(11):**1772-1787.
15. Barnett LJ: **Probabilistic Analysis of Random Clone Restriction Mapping.** In *Master's thesis* Washington University, Saint Louis MO; 1990.
16. Wendl MC: **Collision Probability Between Sets of Random Variables.** *Statistics and Probability Letters* 2003, **64(3):**249-254.
17. Wendl MC: **Probabilistic Assessment of Clone Overlaps in DNA Fingerprint Mapping via a *priori* Models.** *Journal of Computational Biology* 2005, **12(3):**283-297.
18. Luo MC, Thomas C, You FM, Hsiao J, Shu OY, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J: **High-Throughput Fingerprinting of Bacterial Artificial Chromosomes Using the SNaPshot Labeling Kit and Sizing of Restriction Fragments by Capillary Electrophoresis.** *Genomics* 2003, **82(3):**378-389.
19. Flibotte S, Chiu R, Fjell C, Krzywinski M, Schein JE, Shin H, Marra MA: **Automated Ordering of Fingerprinted Clones.** *Bioinformatics* 2004, **20(8):**1264-1271.
20. Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD, Waterston RH: **High Throughput Fingerprint Analysis of Large-Insert Clones.** *Genome Research* 1997, **7(11):**1072-1084.
21. Barenblatt GI: *Dimensional Analysis* New York NY: Gordon and Breach; 1987.
22. Osoegawa K, Mammoser AG, Wu C, Frengen E, Zeng C, Catanese JJ, de Jong PJ: **A Bacterial Artificial Chromosome Library for Sequencing the Complete Human Genome.** *Genome Research* 2001, **11(3):**483-496.
23. Nelson WM, Bharti AK, Butler E, Wei F, Fuks G, Kim HR, Wing RA, Messing J, Soderlund C: **Whole-Genome Validation of High-Information-Content Fingerprinting.** *Plant Physiology* 2005, **139:**27-38.
24. Nelson WM, Dvorak J, Luo MC, Messing J, Wing RA, Soderlund C: **Efficacy of Clone Fingerprinting Methodologies.** *Genomics* 2007, **89:**160-165.
25. Ding Y, Johnson MD, Colayco R, Chen YJ, Melnyk J, Schmitt H, Shizuya H: **Contig Assembly of Bacterial Artificial Chromosome Clones through Multiplexed Fluorescence-Labeled Fingerprinting.** *Genomics* 1999, **56(3):**237-246.
26. Klein PE, Klein RR, Cartinhour SW, Ulanch PE, Dong JM, Obert JA, Morishige DT, Schlueter SD, Childs KL, Ale M, Mullet JE: **A High-Throughput AFLP-Based Method for Constructing Integrated Genetic and Physical Maps: Progress Toward a Sorghum Genome Map.** *Genome Research* 2000, **10(6):**789-807.
27. Fuhrmann DR, Krzywinski MI, Chiu R, Saeedi P, Schein JE, Bosdet IE, Chinwalla A, Hillier LW, Waterston RH, McPherson JD, Jones SJM, Marra MA: **Software for Automated Analysis of DNA Fingerprinting Gels.** *Genome Research* 2003, **13(5):**940-953.
28. Chen MS, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang FC, Kim H, Frisch D, Yu YS, Sun SH, Higingbottom S, Phimphilai J, Phimphilai D, Thurmond S, Gaudette B, Li P, Liu JD, Hatfield J, Main D, Farrar K, Henderson C, Barnett L, Costa R, Williams B, Walser S, Atkins M, Hall C, Budiman MA, Tomkins JP, Luo MZ, Bancroft I, Salse J, Regad F, Mohapatra T, Singh NK, Tyagi AK, Soderlund C, Dean RA, Wing RA: **An Integrated Physical and Genetic Map of the Rice Genome.** *Plant Cell* 2002, **14(3):**537-545.
29. Krzywinski M, Bosdet I, Smailus D, Chiu R, Mathewson C, Wye N, Barber S, Brown-John M, Chan S, Chand S, Cloutier A, Girn N, Lee D, Masson A, Mayo M, Olson T, Pandoh P, Prabhu AL, Schoenmakers E, Tsai M, Albertson D, Lam W, Choy CO, Osoegawa K, Zhao SY, de Jong PJ, Schein J, Jones S, Marra MA: **A Set of BAC Clones Spanning the Human Genome.** *Nucleic Acids Research* 2004, **32(12):**3651-3660.
30. Kreyszig E: *Advanced Engineering Mathematics* 6th edition. New York NY: John Wiley & Sons; 1988.