UNIVERSITÉ DU QUÉBEC À MONTRÉAL

CARTOGRAPHIE GÉNÉTIQUE FINE : ÉVALUATION D'UNE MÉTHODE D'ESTIMATION DES ALLÈLES ET DU MODÈLE DE PÉNÉTRANCE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

MATHIEU DUPONT

SEPTEMBRE 2012

Aux étudiants indignés de ce printemps québécois 2012, que nous n'oublierons jamais

ÉDITIONS ÉLECTRONIQUES

Le présent mémoire est optimisé pour être consulté en version électronique. Le texte contient des liens (en bleu) qui permettent d'accéder rapidement aux pages auxquelles il se réfère. De plus, 4 *boutons* simples, situés au bas des pages, permettent au lecteur de naviguer facilement et rapidement entre les diverses parties de l'ouvrage. Le bouton le plus utile est sans doute le deuxième ($<$), qui permet à tout moment de revenir à la page où l'on se trouvait précédemment. Le suivant ($>$), présent dans les chapitres, amène aux notations mathématiques, situées à la fin de chacun des chapitres. Enfin, les boutons $\ll$ et $\gg$ permettent d'accéder directement à la TABLE DES MATIÈRES et à l'INDEX, respectivement.

Deux versions électroniques sont disponibles : l'une classique (format lettre) et l'autre en format paysage, optimisée pour les tablettes numériques. Toutes deux peuvent être téléchargées à partir du site Web de mon directeur de recherche, Fabrice Larribe :

http://www.math.uqam.ca/fabricelarribe/memoires/MathieuDupont2012-FormatTabletteElectronique.pdf

http://www.math.uqam.ca/fabricelarribe/memoires/MathieuDupont2012-FormatLettre.pdf

Quoiqu'il s'agisse de fichiers en format PDF, les graphiques sont optimisés pour être lus avec le logiciel *Aperçu* (*Preview*).

TABLE DES MATIÈRES

LISTE DES FIGURES

x

LISTE DES TABLEAUX

RÉSUMÉ

Nous présentons et testons une méthode d'estimation des allèles d'une mutation potentiellement associée à un phénotype ainsi qu'une méthode d'estimation de son modèle de pénétrance...

MOTS-CLÉS : algorithme EM, cartographie génétique, coalescence, MapARG, modèle de pénétrance, risque relatif, SNP, statistique génétique, vraisemblance composite

INTRODUCTION

Bien qu'*Homo sapiens* pratique la sélection artificielle sur d'autres espèces depuis des millénaires, parfois consciemment et parfois inconsciemment, ce n'est que depuis tout récemment dans son histoire qu'il en comprend en partie les mécanismes, grâce notamment aux travaux de Darwin (1859) et de Mendel (1865). Aujourd'hui, les bases de données génétiques contiennent des quantités astronomiques de ces données et en reçoivent continuellement de nouvelles...

Le chapitre I présente la génétique des populations, en introduisant quelques concepts de génétique et du processus de coalescence, pour finalement y situer la cartographie génétique...

2

CHAPITRE I

GÉNÉTIQUE DES POPULATIONS

La *cartographie génétique*, qui sera abordée au chapitre II et dont traite ce mémoire, puise ses outils mathématiques dans la génétique des populations. Le lecteur peut se référer au lexique (page 15) afin d'obtenir rapidement la définition d'un terme précis. La section 1.1 introduit quand à elle la *théorie de la coalescence*, sur laquelle repose le méthode de cartographie génétique *MapARG*, qui est l'objet de ce mémoire. Les notations mathématiques du présent chapitre se trouvent à sa fin, page 7.

1.1 Théorie de la coalescence

La méthodologie utilisée dans cet ouvrage (*MapARG*, chapitre II) se base sur des principes mathématiques de la *théorie de la coalescence*, initialement développée au début des années 1980 par John Kingman (1982, 2000).

1.1.1 Modèle de Wright-Fisher

MapARG est une méthode proposée relativement récemment (Larribe, 2003 ; Larribe *et al.*, 2002). La vraisemblance composite est de plus en plus utilisée en statistique génétique (Larribe et Fearnhead, 2011) en raison de la quantité croissante de données disponibles et de leur dépendance évidente du point de vue génétique. Elle fut récemment implantée dans MapARG (Larribe et Lessard, 2008), réduisant considérablement le temps de calcul et permettant ainsi l'utilisation de beaucoup de SNPs et d'échantillons de grande taille.

Quoiqu'incalculable, cette espérance nous permet cependant d'estimer $L(x_T)$, la vraisemblance de la position du TIM, par une moyenne sur un certain nombre K d'ARGs générés selon la distribution P_{x_T} :

$$\hat{L}(x_T) \;=\; \hat{Q}_{x_T}(H_0) \;=\; \frac{1}{K}\sum_{k=1}^{K}\left(\prod_{\tau=0}^{\tau^*-1}\frac{Q_{x_T}(H_\tau|H_{\tau+1})}{P_{x_T}(H_{\tau+1}|H_\tau)}\right). \tag{1.1}$$

En supposant la population en équilibre d'Hardy-Weinberg et que l'on connait p, f et F, la distribution des diplotypes au TIM par rapport au phénotype peut aisément être calculée (tableau 1.1).

Tableau 1.1 Distribution des allèles au TIM dans la population

		Phénotype		**Total**
		$\phi = 0$	$\phi = 1$	
Diplotype au TIM	$T = 00$	$(1-f_0)\,(1-p)^2$	$f_0\,(1-p)^2$	$(1-p)^2$
	$T = 01$	$(1-f_1)\,p(1-p)$	$f_1\,p(1-p)$	$p(1-p)$
	$T = 10$	$(1-f_1)\,p(1-p)$	$f_1\,p(1-p)$	$p(1-p)$
	$T = 11$	$(1-f_2)\,p^2$	$f_2\,p^2$	p^2
	Total	$1-f$	f	1

De plus, rappelons que selon l'hypothèse 4 du modèle de Wright-Fisher, les mutations sont neutres et n'exercent donc aucune influence sur la structure de la généalogie. La figure 1.1 illustre deux mutations dans la généalogie vue précédemment.

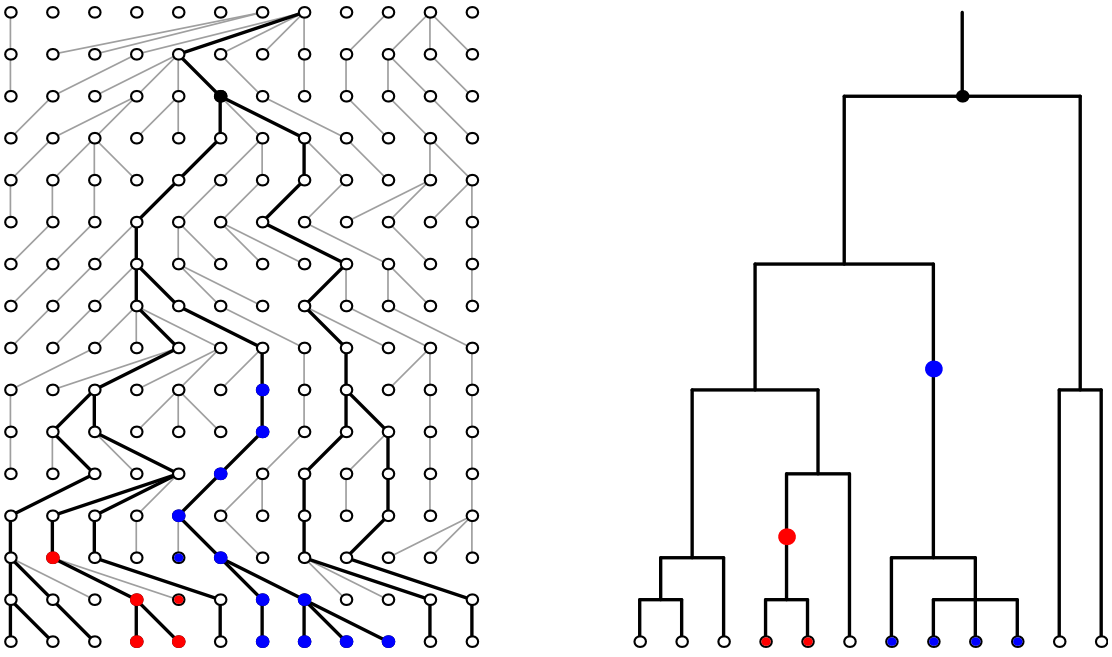

Figure 1.1 Arbre de coalescence avec mutations. *Gauche :* Deux mutations sont simulées dans la généalogie. Toutes les séquences d'une lignée qui sont situées en dessous de l'apparition d'une mutation portent cette dernière. *Droite :* Arbre de coalescence avec mutations. La hauteur des points de mutation correspond à leur temps d'occurrence.

Cette distribution nous permet finalement d'estimer la vraisemblance de la position du TIM par l'équation 1.1 (page 4), qui devient :

$$\begin{aligned}\hat{L}(x_T) &= \frac{1}{K}\sum_{k=1}^{K}\left(\prod_{\tau=0}^{\tau^*-1}\frac{Q_{x_T}(H_\tau|H_{\tau+1})}{P_{x_T}(H_{\tau+1}|H_\tau)}\right)\\ &= \frac{1}{K}\sum_{k=1}^{K}\left(\prod_{\tau=0}^{\tau^*-1}\frac{Q_{x_T}(H_\tau|H_{\tau+1})}{Q_{x_T}(H_\tau|H_{\tau+1})\,\frac{\phi(H_{\tau+1})}{\phi(H_\tau)}}\right)\\ &= \frac{1}{K}\sum_{k=1}^{K}\left(\prod_{\tau=0}^{\tau^*-1}\frac{\phi(H_\tau)}{\phi(H_{\tau+1})}\right).\end{aligned}$$

1.2 Algorithme de MapARG

Afin de bien saisir la structure de la méthode MapARG que nous venons de présenter, en voici les grandes étapes :

I. Choisir l'ensemble des positions x_T pour lesquelles $CL(x_T)$ sera évaluée ;

II. Pour chacune des $L-d+1$ fenêtres couvrant l'ensemble des SNPs de l'échantillon :

  Pour chacun des $d-1$ intervalles situés dans la fenêtre :

    Pour chacun des K graphes à construire :

      Pour chaque étape τ du graphe, tant que le MRCA n'est pas atteint :

      i. Calculer $\frac{Q_{x_T}(H_\tau|H_{\tau+1})}{P_{x_T}(H_{\tau+1}|H_\tau)} = \frac{\phi(H_\tau)}{\phi(H_{\tau+1})}$;

      ii. Mettre à jour $Q_{x_T}(H_\tau)$ et $P_{x_T}(H_{\tau+1})$;

      iii. Générer un évènement selon $P_{x_T}(H_{\tau+1})$;

      iv. Mettre à jour $H_{\tau+1}$;

III. Pour chaque position x_T :

  Calculer $\hat{CL}(x_T)$;

IV. $\hat{x}_T$ correspond au maximum de $\hat{CL}(x_T)$.

NOTATIONS DU CHAPITRE I

N	Taille de la population.
n_e	Taille de l'échantillon.
n	Nombre de lignées restantes à un moment donné.
μ	Taux mutation sur une lignée entre deux générations.
$\theta = 2\,\mu N$	Taux de mutation dans la population.
r	Taux recombinaison sur une lignée entre deux générations.
$\rho = 2\,rN$	Taux recombinaison dans la population.
T_C^n, T_M^n, T_R^n	Temps avant qu'il y ait une coalescence, une mutation ou une recombinaison, respectivement, lorsqu'il reste n lignées.
T^n	Temps avant qu'il y ait un évènement, lorsqu'il reste n lignées.

CHAPITRE II

CARTOGRAPHIE GÉNÉTIQUE VIA UN PROCESSUS DE COALESCENCE : LA MÉTHODE MapARG

NOTATIONS DU CHAPITRE II

C_{ij}^{k}	Coalescence de séquences de types i et j compatibles en une séquence de type k.
$M_{i}^{j}(s)$	Mutation au marqueur s d'une séquence de types i en une séquence parentale de type j.
$R_{i}^{jk}(s)$	Recombinaison, dans l'intervalle s d'une séquence de types i en deux séquences parentales de types j et k.
H_0	Ensemble des haplotypes observés.
H_τ	Ensemble des haplotypes restants après le τ^e évènement.
H_{τ^*}	MRCA.
L	Nombre de marqueurs.
π_C	Taux de coalescence.
π_M	Taux de mutation.
π_R	Taux de recombinaison.
x_i	Position du marqueur i (en Mb).
x_T	Position du TIM (en Mb).

CONCLUSION

L'objectif du présent ouvrage était de tester l'efficacité et le potentiel de deux méthode d'estimation, l'une pour estimer l'allèle d'une mutation cherchée sur tous les haplotypes d'un échantillon, et l'autre pour estimer le modèle de pénétrance de cette mutation, toutes deux reposant sur le même algorithme EM...

Une investigation plus élaborée de ce côté pourrait être prometteuse...

LEXIQUE

ADN *Acide désoxyribonucléique*. Longue molécule supportant l'information génétique et qui est formée d'une séquence linéaire des nucléotides A, C, G et T. L'ADN humain, long d'environ 3 milliards de nucléotides et divisé en 23 paires de chromosomes, contient environ 30 000 gènes.

ARG *Ancestral Recombination Graph*. Graphe de recombinaison ancestral.

... ...

INDEX

APPENDICE A

TAUX DE SUCCÈS SEMI-PARTIELS

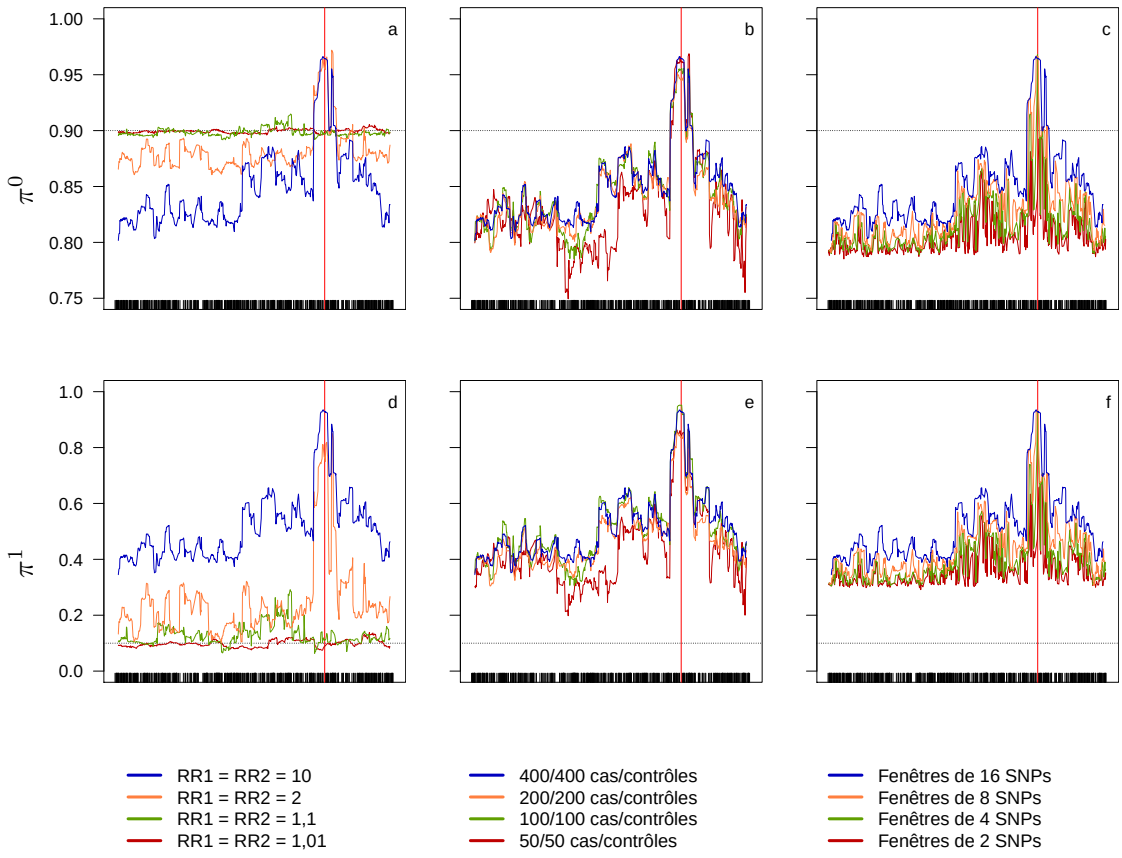

Figure A.I Taux de succès des primitifs (*a,b,c*) et des mutants (*d,e,f*) en fonction des risques relatifs RR1 et RR2 combinés (*a,d*), de la taille de l'échantillon (*b,e*) et de la largeur des fenêtres (*c,f*). Pour une ligne donnée, l'échelle des ordonnées est la même. La ligne pointillée représente le taux de succès aléatoire (*a,b,c :* 0,9 ; *d,e,f :* 0,1).
a,d : 400/400 contrôles/cas, fenêtres de 16 SNPs ;
b,e : RR1 = RR2 = 10, fenêtres de 16 SNPs ;
c,f : RR1 = RR2 = 10, 400/400 contrôles/cas.

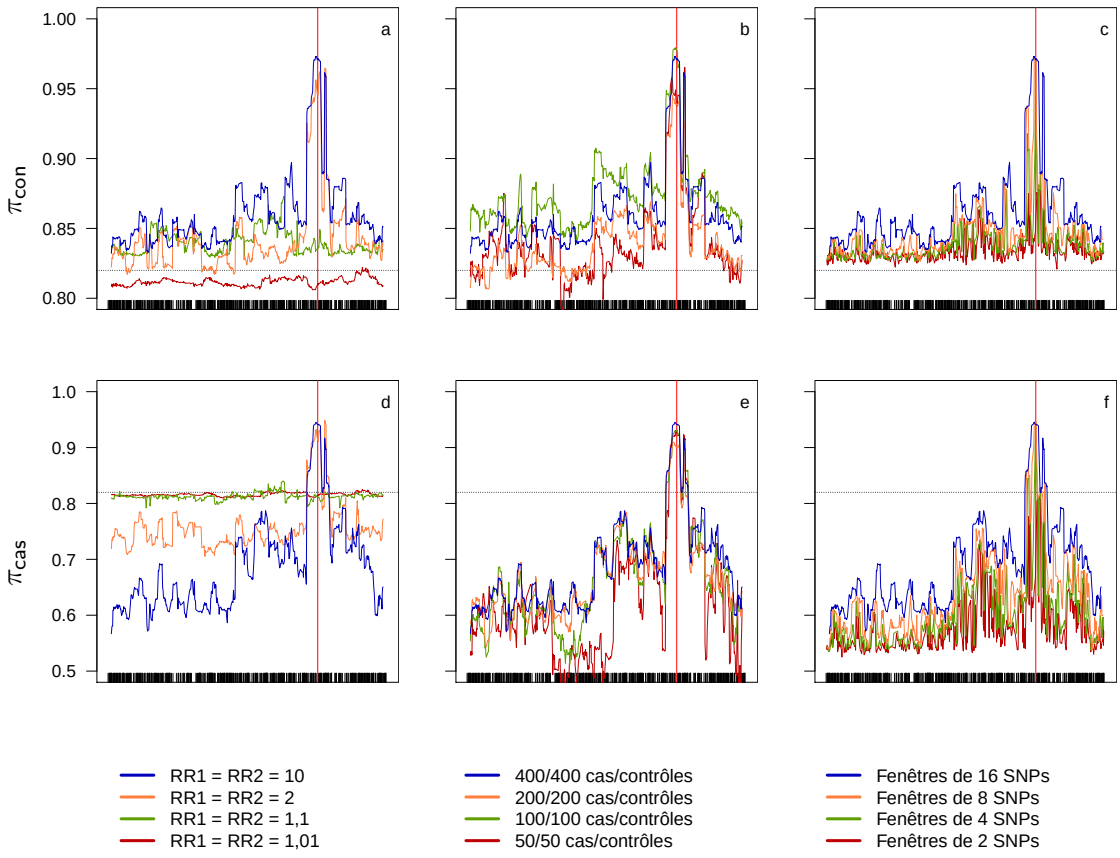

Figure A.II Taux de succès des contrôles (*a,b,c*) et des cas (*d,e,f*) en fonction des risques relatifs RR1 et RR2 combinés (*a,d*), de la taille de l'échantillon (*b,e*) et de la largeur des fenêtres (*c,f*). Pour une ligne donnée, l'échelle des ordonnées est la même. La ligne pointillée représente le taux de succès aléatoire (0,82).
a,d : 400/400 contrôles/cas, fenêtres de 16 SNPs ;
b,e : RR1 = RR2 = 10, fenêtres de 16 SNPs ;
c,f : RR1 = RR2 = 10, 400/400 contrôles/cas.

RÉFÉRENCES

Boucher, G. (2009). Intégration de la réalité diploïde et des modèles de pénétrance à une méthode de cartographie génétique fine. Mémoire de maîtrise, Université du Québec à Montréal.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.

Dawkins, R. (1976). *The selfish gene*. Oxford University Press, New York.

Descary, M.-H. (2012). Dmap : une nouvelle méthode de cartographie génétique fine adaptée à des modèles génétiques complexes. Mémoire de maîtrise, Université du Québec à Montréal.

Excoffier, L. et Foll, M. (2011). fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, 27(9):1332–1334.

Fearnhead, P. et Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics*, 159(3):1299–1318.

Fisher, R. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Griffiths, R. et Tavaré, S. (1994a). Ancestral inference in population genetics. *Statistical Science*, pages 307–319.

Griffiths, R. et Tavaré, S. (1994b). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1310):403–410.

Griffiths, R. et Tavaré, S. (1994c). Simulating probability distributions in the coalescent. *Theoretical Population Biology*, 46(2):131–159.

Griffiths, R. C. et Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, 3(4):479–502.

Hein, J., Schierup, M. H. et Wiuf, C. (2005). *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*, 23(2):183–201.

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43.

Kingman, J. F. C. (2000). Origins of the coalescent. 1974-1982. *Genetics*, 156:1461–1463.

Kuhner, M., Yamato, J. et Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. *Genetics*, 140(4):1421–1430.

Larribe, F. (2003). Cartographie génétique fine par le graphe de recombinaison ancestral. Thèse de doctorat, Université de Montréal.

Larribe, F. et Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statistica Sinica*, 21(1):43–69.

Larribe, F. et Lessard, S. (2008). A composite-conditional-likelihood approach for gene mapping based on linkage disequilibrium in windows of marker loci. *Stat Appl Genet Mol Biol*, 7(1):Article27.

Larribe, F., Lessard, S. et Schork, N. J. (2002). Gene mapping via the ancestral recombination graph. *Theor Popul Biol*, 62(2):215–229.

Mendel, J. G. (1865). Versuche über pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, 4:3–47.

Stephens, M. et Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):605–635.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14:43–59.

Vahey, S. (2008). Modélisation des paramètres de pénétrance incomplète et de phénocopie d'une méthode de cartographie fine d'une maladie complexe. Mémoire de maîtrise, Université du Québec à Montréal.

Wakeley, J. (2009). *Coalescent theory: an introduction*. Roberts & Co. Publishers.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, 16:97–159.