

Κεφάλαιο 10

Ανάλυση της γενετικής ποικιλομορφίας

Σύνοψη

Έχοντας μελετήσει τις μεταβολές των βιολογικών αλληλουχιών σε επίπεδο ειδών στο Κεφάλαιο 6, στρεφόμαστε σ' αυτό το κεφάλαιο στην ανάλυση της γενετικής ποικιλομορφίας σε επίπεδο πληθυσμών. Το εύρος του συγκεκριμένου θέματος επιβάλλει μια αρχική σύνοψη βασικών βιολογικών εννοιών όπως είναι οι πληθυσμοί, η ετεροζυγωτία και οι σημειακοί πολυμορφισμοί. Αφού περιγράψουμε τη θεωρία πίσω από τον τρόπο υπολογισμού των συχνοτήτων αλληλομόρφων, θα εστιάσουμε στο βασικό βιολογικό πρόβλημα που θα βρίσκεται στο επίκεντρο του συγκεκριμένου κεφαλαίου και το οποίο είναι η ανάλυση δεδομένων μελετών γενετικής συσχέτισης σε γονιδιωματική κλίμακα (Genome-Wide Association Studies, GWAS). στο πλαίσιο του συγκεκριμένου προβλήματος θα αναλυθούν οι μαθηματικές έννοιες των λόγων πιθανοτήτων (odds ratios) και των σχετικών στατιστικών μεθοδολογιών όπως το Fisher's test καθώς και των τρόπων αναπαράστασης των δεδομένων τέτοιων αναλύσεων (Manhattan plots). Τέλος θα παρουσιάσουμε τις σύγχρονες μεθοδολογίες ανάλυσης γενετικής ποικιλομορφίας σε ατομικό επίπεδο με έμφαση στην αλληλούχιση πλήρους εξονιώματος (whole-exome sequencing) και θα δούμε πώς αυτή μπορεί να χρησιμοποιηθεί για τη μελέτη σπάνιων γενετικών νοσημάτων.

Στο τέλος του Κεφαλαίου θα πρέπει να μπορείτε:

- Να αναφέρετε τα επίπεδα στα οποία εκδηλώνεται η γενετική ποικιλομορφία.
- Να υπολογίσετε τις σχετικές συχνότητες γονοτύπων με βάση τις συχνότητες αλληλομόρφων.
- Να περιγράψετε έννοιες όπως "απλότυπος" και "ανισορροπία σύνδεσης".
- Να περιγράψετε τη βάση της μεθοδολογίας γενετικής συσχέτισης σε γονιδιωματική κλίμακα και να υπολογίσετε τη στατιστική σημαντικότητα αποτελεσμάτων από μια μελέτη GWAS.
- Να διακρίνετε μεταξύ αλληλούχισης πλήρους γονιδιώματος και εξονιώματος και να περιγράψετε τα πλεονεκτήματα της τελευταίας.

Εισαγωγή

Στο Κεφάλαιο 6 είδαμε πώς οι εξελικτικές σχέσεις μεταξύ διαφορετικών ειδών μπορούν να μελετηθούν μέσω των μεταβολών που συμβαίνουν στις βιολογικές αλληλουχίες στο χρόνο. Στη συγκεκριμένη ανάλυση θεωρήσαμε ότι η ποικιλομορφία μεταξύ των ατόμων του ίδιου είδους είναι αμελητέα σε σύγκριση με τις διαφορές που υπάρχουν μεταξύ των ειδών, τα οποία μπορούν με τη σειρά τους να αντιπροσωπευθούν από ένα δεδομένο σύνολο χαρακτηριστικών. Η παραδοχή αυτή, βασική για τη φυλογενετική ανάλυση, ισχύει στο επίπεδο που εξετάζουμε χαρακτηριστικά που αφορούν διαφορετικά είδη που εξελίσσονται σε χρονικές κλίμακες που αντιστοιχούν σε πολύ μεγάλο αριθμό γενιών. Όμως αρκεί να παρατηρήσει κανείς δύο δέντρα του ίδιου είδους, δύο κουτάβια της ίδιας ράτσας ή να αναλογιστεί τα μέλη της οικογένειάς του, για να καταλάβει ότι ακόμα και στο επίπεδο του ίδιου είδους υπάρχει μορφολογική (καταρχάς) ποικιλομορφία. Τα άτομα του ίδιου είδους δεν είμαστε όλα το ίδιο ψηλά, δεν έχουμε το ίδιο χρώμα μάτια ή μαλλιά, δεν έχουμε την ίδια προδιάθεση για συγκεκριμένες παθολογικές καταστάσεις.

Αυτή η μορφολογική, ή φαινοτυπική όπως είναι ο σωστός όρος, ποικιλομορφία είναι σε συντριπτικό βαθμό αντανάκλαση της υποκείμενης γενετικής ποικιλομορφίας που εκφράζεται στο επίπεδο των πληθυσμών των ατόμων του ίδιου είδους. Η έννοια του πληθυσμού είναι εδώ πολύ σημαντική, καθώς ορίζει το σύνολο των ατόμων που σε τοπικό επίπεδο είναι σε θέση να ανταλλάξουν γενετικό υλικό και να διαμορφώσουν αυτό που ονομάζουμε “γονιδιακή δεξαμενή” (gene pool). (Ακόμα κι αν στην εποχή της σύγχρονης γονιδιωματικής, οι διάφοροι χαρακτήρες δεν είναι αποκλειστικά συνδεδεμένοι με τα γονίδια αλλά μπορούν συχνά να σχετίζονται και με μη-κωδικά γονιδιωματικά χαρακτηριστικά). Από αυτήν την άποψη, ένας βιολογικός πληθυσμός μπορεί να αντιστοιχεί σ' ένα “γενετικό στιγμιότυπο” από την εξελικτική πορεία ενός είδους, είναι όμως άρρηκτα συνδεδεμένος με την ίδια την εξέλιξη. Πράγματι, ένας σχετικά αυστηρός ορισμός της εξέλιξης σε μοριακό επίπεδο θα μπορούσε να είναι *η διαδικασία της μεταβολής μέσα στο χρόνο των σχετικών συχνοτήτων των ποικιλόμορφων γονιδιωματικών χαρακτηριστικών ενός πληθυσμού*.

Η μελέτη της διαδικασίας αυτής, αφορά ένα ολόκληρο επιστημονικό πεδίο που ονομάζεται Γενετική Πληθυσμών και του οποίου η αναλυτική περιγραφή δεν είναι προφανώς μέσα στους σκοπούς αυτού του βιβλίου. Από υπολογιστική άποψη, ωστόσο, η Γενετική Πληθυσμών αποτελεί ένα από τα κυριότερα πεδία της βιολογίας, που δε θα μπορούσε να αναπτυχθεί χωρίς την υποστήριξη των μαθηματικών, της στατιστικής και πιο πρόσφατα της χρήσης ηλεκτρονικών υπολογιστών για την ανάλυση του όγκου των δεδομένων που παράγεται στο πλαίσιο της ανάλυσης της ανθρώπινης (κατά κύριο λόγο) γενετικής ποικιλομορφίας με σύγχρονες μεθόδους αλληλούχισης νέας γενιάς (για τις οποίες θα συζητήσουμε πιο αναλυτικά στο επόμενο κεφάλαιο). Στο κεφάλαιο αυτό, θα εξετάσουμε τα σημεία όπου η συνεισφορά των υπολογιστικών μεθοδολογιών είναι κρισιμότερη, τόσο για την κατανόηση των βασικών θεωρητικών αρχών, όσο και για την ανάλυση δεδομένων με σκοπό την επίλυση συγκεκριμένων βιολογικών προβλημάτων γενετικής ποικιλομορφίας.

Το βιολογικό ερώτημα

Φανταστείτε ότι μελετάτε μια συγκεκριμένη παθολογική κατάσταση, η εκδήλωση της οποίας είναι πολύ πιθανό να οφείλεται σε γενετική προδιάθεση. Τα άτομα που πάσχουν έχουν συχνά επιβαρυνμένο οικογενειακό ιστορικό, με κάποιο μέλος της οικογένειάς τους να νοσεί ή να έχει στο παρελθόν νοσήσει, ενώ κι άλλα φαινοτυπικά χαρακτηριστικά των ατόμων που συνδέονται με συγγενικές σχέσεις, φαίνεται πως συσχετίζονται με την πιθανότητα εμφάνιση της ασθένειας. Το βιολογικό ερώτημα που προκύπτει από τις παραπάνω παρατηρήσεις είναι μάλλον προφανές:

Με ποιον τρόπο θα μπορούσε κανείς να εντοπίσει τη γενετική βάση μιας παθολογικής κατάστασης και να αξιολογήσει το βαθμό στον οποίο αυτή επηρεάζει την εκδήλωση και την εξέλιξη μιας σχετιζόμενης νόσου;

Διατυπωμένο κατ' αυτόν τον τρόπο, το ερώτημα ακούγεται αρκετά γενικό και θα χρειαστεί να το εξειδικεύσουμε στην πορεία κάνοντας συγκεκριμένες παραδοχές αλλά και απαντώντας σε επιμέρους ερωτήσεις. Για παράδειγμα, πώς θα ορίσουμε την ίδια τη νόσο και αυτούς που πάσχουν από αυτήν; Πώς θα συγκριθεί η κατάστασή τους με τη φυσιολογική; Ποια στοιχεία του γενετικού τους υλικού θα μελετήσουμε και με ποιον τρόπο και πώς θα αναλύσουμε τα δεδομένα; Στη βάση όλων των ερωτημάτων, τόσο του γενικού, όσο και των επιμέρους, βρίσκεται η πραγματικότητα της γενετικής ποικιλομορφίας, η γνώση δηλαδή ότι η γονιδιωματική αλληλουχία διαφέρει σε κάποιο βαθμό μεταξύ δύο οποιονδήποτε ατόμων του ίδιου είδους (εκτός αν πρόκειται για ομοζυγωτικούς διδύμους¹). Πριν εξετάσουμε τρόπους για την ποσοτικοποίηση, την ανάλυση και την ερμηνεία της γενετικής ποικιλομορφίας θα ήταν σκόπιμο να συζητήσουμε συνοπτικά τις βασικές έννοιες που τη διέπουν και οι οποίες αποτελούν τη βάση για τα μεθοδολογικά εργαλεία που θα περιγράψουμε στη συνέχεια.

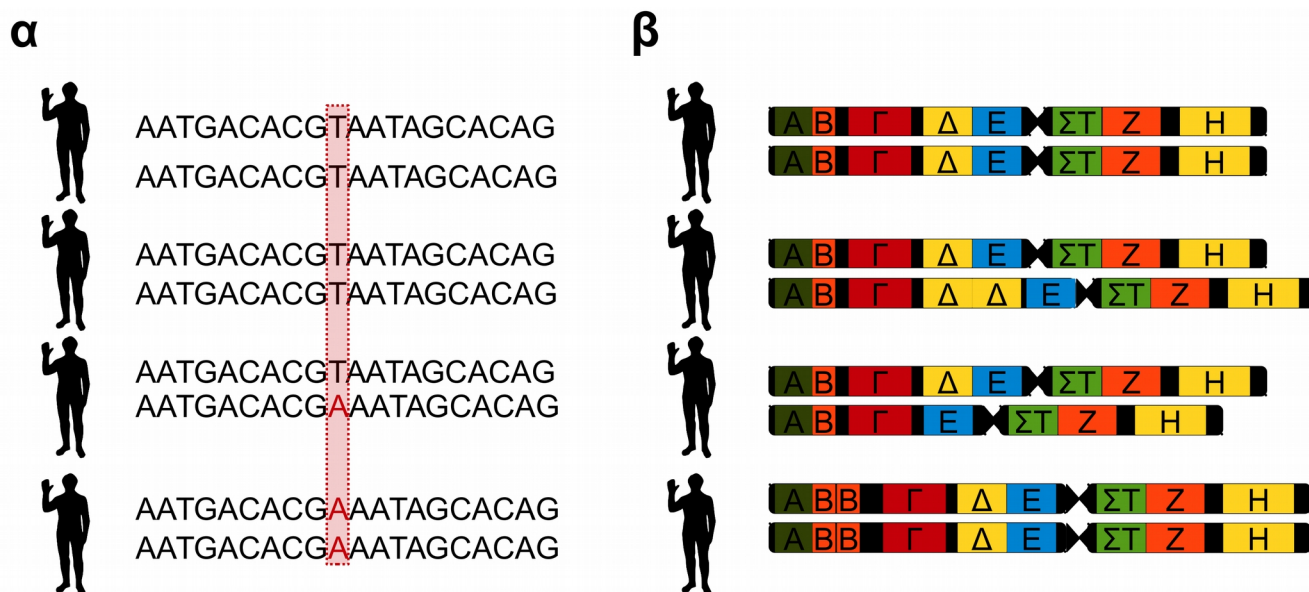
Γενετική ποικιλομορφία

Οι διπλοειδείς οργανισμοί, όπως είναι π.χ. ο άνθρωπος και η συντριπτική πλειοψηφία των θηλαστικών, περιέχουν δύο αντίγραφα από κάθε χρωμόσωμα στο γονιδιώμα τους. Αυτό σημαίνει ότι κάθε γονιδιωματικό στοιχείο υπάρχει σε δύο ομόλογα σημεία του γονιδιώματος (Russell, 2006). Η βάση της γενετικής ποικιλομορφίας έγκειται στο γεγονός ότι τα δύο ομόλογα στοιχεία δεν είναι πανομοιότυπα, αλλά το ίδιο γονιδιακό χαρακτηριστικό μπορεί να υπάρχει σε δύο εκδοχές στο ίδιο γονιδίωμα και σε περισσότερες από δύο αν κανείς εξετάσει διαφορετικά άτομα στον πληθυσμό. Σε γονιδιακό επίπεδο, τα γονίδια που έχουν περισσότερες από μία μορφές, ονομάζονται αλληλόμορφα (alleles). Όταν τα δύο αντίγραφα του ίδιου γενετικού τόπου είναι πανομοιότυπα μεταξύ των χρωμοσωμάτων του ίδιου οργανισμού, λέμε ότι ο οργανισμός είναι "ομόζυγος" για το συγκεκριμένο γενετικό τόπο, ενώ σε αντίθετη περίπτωση μιλάμε για "ετεροζυγωτία".

Οι σύγχρονες μέθοδοι γονιδιωματικής ανάλυσης που, όπως θα δούμε παρακάτω, επιτρέπουν

¹ Ακόμα και στην περίπτωση πανομοιότυπων διδύμων, οι ενδείξεις που έχουμε είναι ότι ποικιλομορφία εκδηλώνεται μέσω επιγενετικών παραγόντων, χαρακτηριστικών δηλαδή που επιδρούν στη γονιδιωματική αλληλουχία (Petronis, 2006). Περισσότερα γι' αυτό θα δούμε στα δύο επόμενα κεφάλαια.

τη μελέτη της ποικιλομορφίας σε πολύ μεγαλύτερη διακριτική ικανότητα, μας ωθούν να επεκτείνουμε τους όρους της ομοζυγωτίας και της ετεροζυγωτίας από το επίπεδο του γενετικού τόπου και των αλληλόμορφων γονιδίων σε επίπεδο μοναδικών νουκλεοτιδίων. Έτσι μπορούμε να μιλάμε για αλληλόμορφα χωρίς να αναφερόμαστε σε γονίδια αλλά στη βασική μονάδα στην οποία εκφράζεται η γενετική ποικιλομορφία, η οποία είναι οι σημειακοί νουκλεοτιδικοί πολυμορφισμοί (single nucleotide polymorphisms, στο εξής για συντομία SNP) (Gunderson, Steemers, Lee, Mendoza & Chee, 2005). Ένας SNP αντιστοιχεί σε μια μοναδική θέση στο απλοειδές γονιδίωμα που διαφέρει μεταξύ των ατόμων του πληθυσμού. Λέμε τότε ότι η θέση αυτή είναι πολυμορφική και ως εκ τούτου ένα δεδομένο άτομο μπορεί αν είναι είτε ομοζυγώτης, είτε ετεροζυγώτης σε ό,τι αφορά τη συγκεκριμένη θέση. Με τον όρο SNP δεν αναφερόμαστε μόνο σε σημειακές νουκλεοτιδικές αντικαταστάσεις αλλά και στις μονονουκλεοτιδικές ενθέσεις και απαλοιφές (Krawitz et al., 2010). Ωστόσο, παρόλο που οι δυνατότητες για ένα SNP είναι μεγαλύτερες από δύο, η μεγάλη πλειοψηφία των πολυμορφικών θέσεων έχουν δύο αλληλόμορφα.



Εικόνα 10.1: Τα δύο βασικά είδη γενετικής ποικιλομορφίας στους πληθυσμούς α) Σημειακοί πολυμορφισμοί (Single nucleotide polymorphisms, SNP). Η θέση μέσα στο κόκκινο πλαίσιο διαφέρει μεταξύ των ατόμων στον πληθυσμό, η πλειοψηφία των οποίων είναι ομόζυγοι (T:T) ενώ κάποια φέρουν έναν πολυμορφισμό (A) είτε σε ετεροζυγωτία (T:A) ή σε ομοζυγωτία (A:A). β) Ποικιλομορφία αριθμού αντιγράφων (Copy number variation, CNV). Μεγαλύτερες περιοχές του γονιδιώματος (από μερικές βάσεις έως μερικές χιλιάδες βάσεις) μπορούν να διπλασιαστούν ή να απαλειφθούν σε ομοζυγωτία ή σε ετεροζυγωτία.

Πέρα από τους σημειακούς πολυμορφισμούς, γενετική ποικιλομορφία μπορεί να προκύψει μέσω μεγαλύτερων σε μήκος μεταβολών του γενετικού υλικού. Αναφερόμαστε συνολικά σε τέτοιες περιπτώσεις με τον όρο “δομική ποικιλομορφία” (structural variation) (Durbin, 2010), με πιο χαρακτηριστικό παράδειγμα την ποικιλομορφία αριθμού αντιγράφων (copy number variation, ή για συντομία CNV). Οι CNV αντιστοιχούν σε περιοχές που μολονότι αναλλοίωτες σε ό,τι αφορά την πρωτοταγή τους αλληλουχία, διαφέρουν ως προς την πολλαπλότητά τους στα ευκαρυωτικά

χρωμοσώματα, όπου τείνουν να εμφανίζονται είτε διπλασιασμένες είτε να εκλείπουν. Οι CNV αφορούν τμήματα του γονιδιωματικού DNA που μπορεί να ποικίλουν από μερικές εκατοντάδες μέχρι λίγα εκατομμύρια βάσεις και συνολικά υπολογίζεται πως αντιστοιχούν στο 13% περίπου του ανθρώπινου γονιδιώματος (Stankiewicz & Lupski, 2010). Παρά τη μεγάλη τους έκταση, ωστόσο, οι CNV φαίνεται πως είναι πολύ λιγότερο πιθανό να αποτελούν τη γενετική βάση για φαινοτυπική ποικιλομορφία, ιδίως σε ό,τι αφορά τη γενετική βάση ασθενειών σε σχέση με τους SNP (Craddock et al., 2010).

Ερώτηση: Μπορείτε να σκεφτείτε πιθανούς λόγους γιατί μια πιο έντονη γενετική διαφοροποίηση (CNV) έχει τελικά μικρότερη επίδραση στην εκδήλωση παθολογικών χαρακτηριστικών σε σχέση με τους απλούστερους SNP;

Ανεξάρτητα με το αν πρόκειται για SNP ή CNV, η επίδραση της ποικιλομορφίας σε φαινοτυπικό επίπεδο εξαρτάται από τη θέση στην οποία εντοπίζεται, την τοπική τους πυκνότητα (αν πρόκειται για SNP) και το αν επικαλύπτονται ή όχι με κωδικές περιοχές. Το ανθρώπινο γονιδίωμα περιέχει περίπου 1.5 εκατομμύριο SNP κάτι που σημαίνει πως κατά μέσο όρο υπάρχει ένα κάθε δύο χιλιάδες βάσεις (Abecasis et al., 2012). Η μεγάλη πλειοψηφία τους αφορά πιθανότατα θέσεις που δεν έχουν παρά μόνο ελάχιστη ή καμία λειτουργική επίδραση, όμως ανάμεσά τους ενδεχομένως να κρύβονται κάποιες θέσεις που ευθύνονται σε μικρό ή μεγαλύτερο βαθμό για την εμφάνιση κάποιας παθολογικής κατάστασης. Πώς θα εντοπίσουμε τις δεύτερες μεταξύ των πρώτων; Στο σημείο αυτό θα ήταν χρήσιμο να δούμε κάποια ποσοτικά χαρακτηριστικά των SNP πριν περάσουμε σε πιο αναλυτικές περιγραφές της έννοιας της γενετικής ποικιλομορφίας. Όπως είδαμε παραπάνω, στις περισσότερες περιπτώσεις, μια πολυμορφική θέση είναι διαλληλική με το ένα από τα δύο αλληλόμορφα να έχει συνήθως σημαντικά μεγαλύτερη συχνότητα στον πληθυσμό. Ονομάζουμε το αλληλόμορφο αυτό μείζον (major allele) σε αντιδιαστολή με το σπανιότερο, ελάσσον (minor allele) και χρησιμοποιούμε τη συχνότητα του δεύτερου, σπανιότερου αλληλόμορφου (minor allele frequency) για να περιγράψουμε τη συχνότητα ενός SNP στον πληθυσμό. Για παράδειγμα, ένας SNP με συχνότητα ελάσσονος αλληλόμορφου 0.25, σημαίνει ότι εμφανίζεται στο 25% των ατόμων του πληθυσμού. Η συχνότητα αυτή σχετίζεται σε μεγάλο βαθμό με την επίδραση που έχει ένας SNP σε μια συγκεκριμένη βιολογική λειτουργία. Σπάνιοι SNP με πολύ μικρή συχνότητα ελάσσονος αλληλομορφου αποτελούν τη γενετική βάση για ασθένειες των οποίων τα αίτια εντοπίζονται σε συγκεκριμένα γονίδια όπως είναι π.χ. η κυστική ίνωση (Kerem et al., 1989). Στις περισσότερες περιπτώσεις, τέτοιοι σπάνιοι πολυμορφισμοί οδηγούν σε μεταβολή στις ιδιότητες μιας κωδικοποιούμενης πρωτεΐνης μέσω της αντικατάστασης ενός αμινοξέος ή (ακόμα εντονότερα) με τη δημιουργία ενός κωδικωνίου λήξης. Για τους ιδιαίτερα σπάνιους SNP με τόσο μεγάλη επίδραση, τείνουμε να χρησιμοποιήσουμε τον όρο “μεταλλάγες” παρόλο που δομικά και γενετικά πρόκειται για το ίδιο φαινόμενο (H. Li, 2011).

Τι συμβαίνει όμως με τους λιγότερο σπάνιους SNP που αποτελούν και το μεγαλύτερο ποσοστό; Πάνω από ποιο όριο σπανιότητας μπορούμε να περιμένουμε ότι η επίδρασή τους θα είναι αμελητέα; Είναι όλοι οι SNP με ανάλογη συχνότητα εμφάνισης το ίδιο σημαντικοί για τον πληθυσμό; Και ποια είναι η σχέση της επίδρασης ενός SNP με τη συχνότητα εμφάνισής του; Στις επόμενες

ενότητες θα δούμε πώς μπορούμε να αξιολογήσουμε τη σημασία που έχει η παρατήρηση ενός SNP σε μια συγκεκριμένη περιοχή. Πριν όμως μπορέσουμε να κάνουμε κάτι τέτοιο, θα πρέπει να δούμε πιο αναλυτικά, χρησιμοποιώντας απλά μαθηματικά, το πώς κατανέμονται τα αλληλόμορφα και πώς κανείς μπορεί να εξετάσει το βαθμό ετεροζυγωτίας ενός πληθυσμού σε μια συγκεκριμένη θέση.

Κατανομή αλληλομόρφων σε πληθυσμούς. Η Αρχή Hardy-Weinberg

Όπως συζητήσαμε και παραπάνω, οι περισσότερες πολυμορφικές θέσεις, οι θέσεις δηλαδή όπου εμφανίζεται γενετική ποικιλομορφία στα γονιδιώματα έχουν δύο αλληλόμορφα, κάτι που σημαίνει ότι στις θέσεις αυτές κάθε άτομο με διπλοειδές γονιδίωμα θα συνδυάσει δύο εκδοχές του ίδιου γενετικού χαρακτήρα στα δύο διαφορετικά χρωμοσώματα. Με ποιον τρόπο όμως θα γίνει αυτός ο συνδυασμός και ποια είναι η αναμενόμενη κατανομή των δύο αλληλομόρφων στα άτομα ενός πληθυσμού; Η απάντηση σ' αυτό το πρωταρχικό ερώτημα δίνεται μέσω της Αρχής των Hardy-Weinberg, ενός από τους πιο θεμελιώδεις νόμους της γενετικής (και έναν από τους λίγους ποσοτικούς νόμους της Βιολογίας). Η αρχή αυτή, που πήρε το όνομά της από τον Άγγλο μαθηματικό Godfrey Harold Hardy (Hardy, 1908) και το Γερμανό φυσιολόγο Wilhelm Weinberg (Weinberg, 1908), βασίζεται σε κάποιες αρχικές παραδοχές για τον τρόπο αναπαραγωγή ενός διπλοειδούς οργανισμού (οι πιο σημαντικές από τις οποίες είναι το τυχαίο ζευγάρωμα, οι μη επικαλυπτόμενες γενιές και η φυλετική αναπαραγωγή) για να υπολογίσει τις συχνότητες με τις οποίες θα εμφανίζεται ο κάθε γονότυπος σε ένα γενετικό τόπο με δύο αλληλόμορφα. Πιο συγκεκριμένα, αν ένας γενετικός χαρακτήρας εκφράζεται με δύο αλληλόμορφα που θα συμβολίσουμε με A (μείζον) και a (ελάσσον) τότε οι πιθανοί γονότυποι σ' ένα διπλοειδές γονιδίωμα είναι:

AA , Aa και aa

Αν ορίσουμε το A ως το μείζον αλληλόμορφο και συμβολίσουμε τη συχνότητα του με p , ενώ αντίστοιχα ορίσουμε τη συχνότητα του a με q και υποθέτοντας φυλετική αναπαραγωγή (την ύπαρξη δηλαδή δύο φύλων που συνδυάζουν το γενετικό τους υλικό), οι συνδυασμοί AA , Aa και aa μπορούν να προκύψουν με τον τρόπο που φαίνεται στον Πίνακα 10.1.

Οι πιθανότητες έτσι των τριών γονοτύπων είναι:

$$r(AA)=p^2 \quad r(Aa)=2pq \quad r(aa)=q^2 \quad 10.1$$

Όπου ισχύει προφανώς $p+q=1$.

Η σημασία της αρχής Hardy-Weinberg έγκειται στο γεγονός ότι έχει χαρακτηριστικά νόμου ισορροπίας. Υποθέτοντας ότι το ζευγάρωμα των γαμετών γίνεται με απόλυτα τυχαίο τρόπο και ότι το μέγεθος του πληθυσμού είναι αρκετά μεγάλο, προκύπτει ότι οι συχνότητες εμφάνισης των τριών γονοτύπων δεν μεταβάλλονται στο χρόνο² και παραμένουν σταθερές στις τιμές που προκύπτουν από τον Πίνακα 10.1. Η κατανομή έτσι των αλληλομόρφων στους πιθανούς γονοτύπους παραμένει σταθερή.

² Στην πραγματικότητα ακόμα και για πολύ μεγάλους πληθυσμούς (της τάξης του 10^6) υπάρχει δυνατότητα απόκλισης από την ισορροπία HW μετά από έναν μεγάλο αριθμό γενιών. Το φαινόμενο αυτό ονομάζεται τυχαία γενετική παρέκκλιση (random genetic drift).

			Αρσενικός Γαμέτης	
			Αλληλόμορφο	Συχνότητα
			A	a
Θηλυκός γαμέτης	Αλληλόμορφο	Συχνότητα	p	q
	A	p	AA	Aa
			p²	pq
	a	q	aA	aa
			qp	q²

Πίνακας 10.1: Συνδυασμοί δύο αλληλομόρφων σε διπλοειδές γονιδίωμα με φυλετική αναπαραγωγή με βάση την Αρχή Hardy-Weinberg.

Η αρχή των Hardy-Weinberg μας επιτρέπει να υπολογίσουμε τη συχνότητα με την οποία απαντώνται διάφορα αλληλόμορφα στον πληθυσμό, εφόσον διαθέτουμε επαρκή γνώση των φαινοτυπικών τους επιπτώσεων. Αναλογιστείτε το παρακάτω παράδειγμα για τον υπολογισμό των αλληλομόρφων, στα οποία οφείλεται η κυστική ίνωση, μια γενετική ασθένεια με εξαιρετικά κακή πρόγνωση που προσβάλλει κατά κύριο λόγο τους πνεύμονες ως αποτέλεσμα μεταλλαγών στο γονίδιο της πρωτεΐνης CFTR (cystic fibrosis transmembrane conductance regulator) (Kerem et al., 1989). Η CFTR είναι μια διαμεμβρανική πρωτεΐνη, της οποίας ο βασικός ρόλος είναι η διατήρηση της ισορροπίας ιόντων χλωρίου εντός του κυττάρου. Διαταραχές στην ισορροπία αυτή οδηγούν σε αποδιοργάνωση της διαδικασίας λίπανσης του επιθηλίου των πνευμόνων, του παγκρέατος, των νεφρών και του παχέος εντέρου με συχνά μοιραία αποτελέσματα (Cheng et al., 1990). Από γενετική άποψη, η κυστική ίνωση είναι μια σπάνια, αυτοσωμική, υπολειπόμενη ασθένεια που είναι όμως σχετικά συχνή στον πληθυσμό των λευκών καυκάσιων Ευρωπαίων, με συχνότητα εμφάνισης 1/2500. Στην Εικόνα 10.2 φαίνεται ο πιο σημαντικός πολυμορφισμός που οδηγεί στη μεταλλαγή ΔF508, που σε ομοζυγωτία οδηγεί στην εκδήλωση της νόσου (Morral et al., 1994). Αυτό σημαίνει ότι το αλληλόμορφο στο οποίο αποδίδεται η εκδήλωση της νόσου είναι υπολειπόμενο (recessive) και ότι τα άτομα που νοσούν θα πρέπει να είναι ομοζυγώτες για το ελάχιστον αλληλόμορφο (aa). Με μόνο δεδομένο τη συχνότητα εμφάνισης της νόσου μπορούμε έτσι να υπολογίσουμε, με βάση την αρχή Hardy-Weinberg ότι η συχνότητα του ελάχιστον αλληλομόρφου q είναι ίση με:

$$q^2 = \frac{1}{2500} \Rightarrow q = \sqrt{\frac{1}{2500}} = \frac{1}{\sqrt{2500}} = \frac{1}{50} = 0.02 \quad 10.2$$

Το μείζον αλληλόμορφο θα έχει έτσι συχνότητα $p = 1 - 0.02 = 0.98$.

Πέρα από τη χρήση της για τον υπολογισμό των συχνοτήτων αλληλομόρφων, η ισορροπία Hardy-Weinberg μας προσφέρει τη βασική μηδενική υπόθεση που μπορούμε να χρησιμοποιήσουμε για να εξετάσουμε το βαθμό στον οποίο η κατανομή των πιθανών γονοτύπων αποκλίνει από την αναμενόμενη. Αποκλίσεις μπορούν να οφείλονται σ' ένα μεγάλο εύρος διαφορετικών φαινομένων

που ποικίλουν από την τυχαία γενετική παρέκλιση ως τη δράση της φυσικής επιλογής. Στη συνέχεια θα δούμε με ποιον τρόπο μπορούμε να εκτιμήσουμε στατιστικά την ύπαρξη αποκλίσεων από την ισορροπία Hardy-Weinberg σε περιπτώσεις που υποψιαζόμαστε ότι ο πληθυσμός δεν εξελίσσεται με βάση τις παραδοχές της.

Μαθηματικό Ιντερμέδιο Ι. Ο έλεγχος του χ^2

Για να καταλάβουμε καλύτερα τη δυνατότητα χρήσης της ισορροπίας Hardy-Weinberg στον έλεγχο υποθέσεων για τη γενετική πληροφορία, ας αναλογιστούμε τώρα ότι μας ζητείται να ελέγξουμε έναν υποπληθυσμό 18000 καυκάσιων λευκών, από τους οποίους οι 12 πάσχουν από κυστική ίνωση. Το ερώτημα που καλούμαστε να απαντήσουμε είναι το εξής:

Μπορούμε να πούμε αν ο εξεταζόμενος υποπληθυσμός (s) είναι διαφορετικός από το γενικότερο σε ό,τι αφορά την εκδήλωση της ασθένειας ή αν η διαφορετική συχνότητα εμφάνισης της νόσου είναι μέσα στα όρια τυχαίων διακυμάνσεων;

Για να απαντήσουμε στην παραπάνω ερώτηση χρειαζόμαστε α) μια ποσότητα (ή ένα σύνολο ποσοτήτων) που θα χαρακτηρίζει τον πληθυσμό β) μια υπόθεση εργασίας που να αντιστοιχεί στην τυχαία κατάσταση και με βάση την οποία θα υπολογίσουμε τις αναμενόμενες τιμές και γ) ένα στατιστικό έλεγχο για την απόρριψη ή όχι της υπόθεσης αυτής. Οι ποσότητες που θα χρησιμοποιήσουμε για να εκτιμήσουμε τον πληθυσμό είναι οι συχνότητες εμφάνισης των αλληλομόρφων του γονιδίου ενώ η μηδενική υπόθεση θα είναι ότι οι συχνότητες αυτές θα προκύπτουν από την ισορροπία Hardy-Weinberg με βάση τη συχνότητα εμφάνισης της ασθένειας στο γενικότερο πληθυσμό. Στην προηγούμενη ενότητα υπολογίσαμε τις συχνότητες εμφάνισης των δύο αλληλομόρφων ως $p=0.98$ και $q=0.02$. Έτσι, για το γενικότερο πληθυσμό θα έχουμε:

$$AA=p^2=0.98^2=0.96$$

$$aa=q^2=0.02^2=0.0004$$

$$Aa=2pq=2 \times (0.02 \times 0.98)=0.0396$$

Οι παραπάνω τιμές είναι έτσι οι αναμενόμενες τιμές, δεδομένης της υπόθεσης ότι ο υποπληθυσμός που μελετούμε είναι μέρος του γενικότερου πληθυσμού. Για τον υποπληθυσμό αυτό η συχνότητα εμφάνισης της νόσου είναι ίση με $12/18000=1/1500$. Με βάση την ανάλυση που περιγράφηκε στην εξίσωση 10.2 μπορούμε να υπολογίσουμε ότι για τον υποπληθυσμό αυτόν:

$$q^2 = \frac{1}{1500} \Rightarrow q = \sqrt{\frac{1}{1500}} = \frac{1}{39} = 0.026 \quad 10.3$$

Χρησιμοποιώντας την αρχή των Hardy-Weinberg για την τιμή q που υπολογίσαμε στην 10.3 προκύπτει ότι οι συχνότητες των τριών γονοτύπων για τον υποπληθυσμό S είναι:

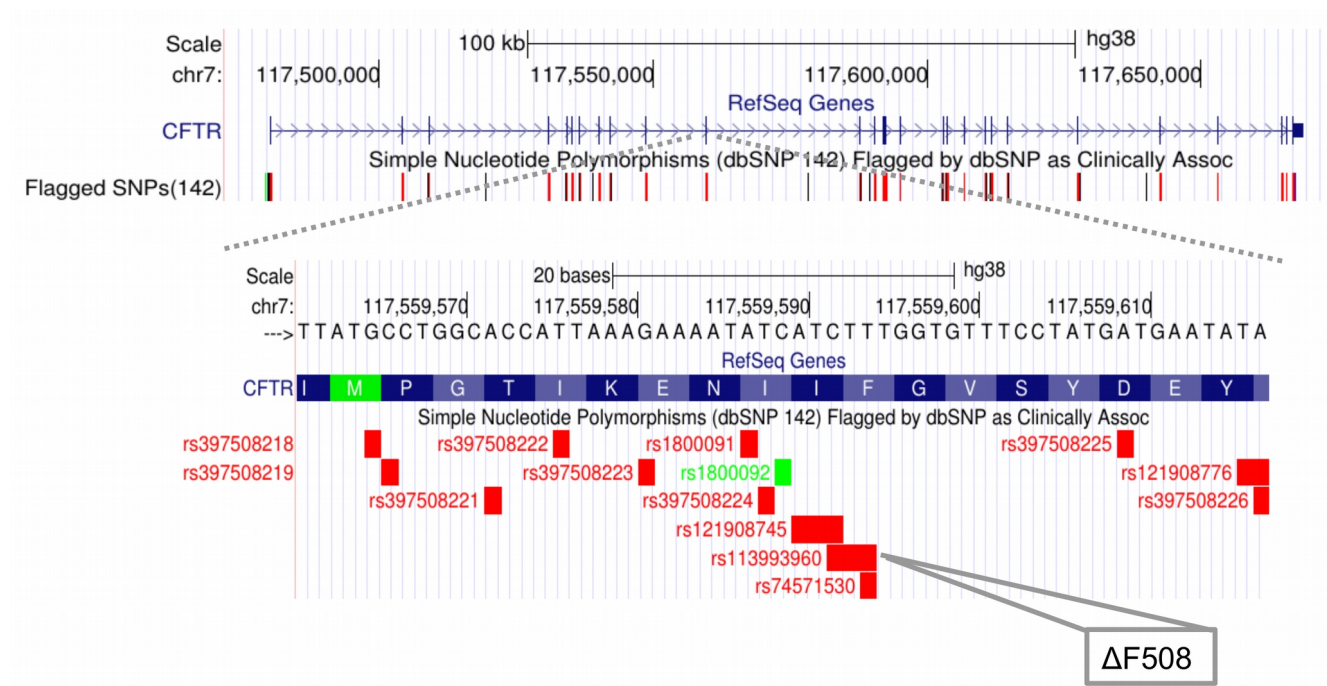
$$AA_s=p^2=(1-q)^2=(1-0.026)^2=0.95$$

$$aa_s=q^2=0.026^2=0.0006$$

$$Aa_s=2pq=2 \times (0.026 \times 0.95)=0.0494$$

Έχοντας υπολογίσει τις παρατηρούμενες (AA_s, aa_s, Aa_s) τιμές συχνοτήτων για των υποπληθυσμό (s) που μελετούμε, χρειαζόμαστε τώρα ένα στατιστικό έλεγχο για να εκτιμήσουμε

κατά πόσο οι διαφορές τους από τις αναμενόμενες είναι στατιστικά σημαντικές. Ο έλεγχος που θα χρησιμοποιήσουμε είναι ο έλεγχος χ^2 (τον οποίο διαβάζουμε ως "έλεγχος χι-τετράγωνο"), ο οποίος χρησιμοποιείται για υποθέσεις που ακολουθούν την ομώνυμη κατανομή (χ^2), μια κατανομή που προκύπτει ως το άθροισμα των τετραγώνων k ανεξάρτητων τυχαίων μεταβλητών.



Εικόνα 10.2: Σχηματική αναπαράσταση του γονιδίου CFTR στο γονιδιωματικό Φυλλομετρητή του University of California at Santa Cruz (UCSC Genome Browser, βλ. και Κεφάλαιο 11). Το γονίδιο εδράζεται στο χρωμόσωμα 7 του ανθρώπινου γονιδιώματος, έχει συνολικό μήκος ~188kbp και αποτελείται από 26 εξόνια, τα οποία φαίνονται με κάθετες γραμμές στο πάνω μέρος της Εικόνας. Κάτω από το διάγραμμα του γονιδίου είναι σημειωμένοι οι πολυμορφισμοί που σχετίζονται κλινικά με την εκδήλωση της κυστικής ίνωσης. Στο κάτω μέρος της Εικόνας φαίνεται μεγεθυμένη μια περιοχή του εξονίου 10 που περιέχει κάποιους από τους πιο υψηλής επικινδυνότητας πολυμορφισμούς. Σημειωμένος με γκριζό πλαίσιο είναι ο rs113993960 που συνιστά και την πιο επικίνδυνη μεταλλαγή ΔF508 η οποία οδηγεί σε απαλοιφή ενός καταλοίπου φαινυλαλανίνης (F) από τη CFTR πρωτεΐνη. Η συχνότητα της ΔF508 στον πληθυσμό των λευκών καυκάσιων είναι 1/50 και σε ομοζυγωτία οδηγεί στην εκδήλωση της νόσου.

Ως k ορίζεται έτσι ο βαθμός ελευθερίας της κατανομής και λέμε ότι πραγματοποιούμε έλεγχο χ^2 με k βαθμούς ελευθερίας. Ο έλεγχος χ^2 πραγματοποιείται με τη χρήση της εξίσωσης:

$$\chi^2 = \sum_{i=1}^{i=N} \frac{(Obs_i - exp_i)^2}{exp_i} \quad 10.4$$

Όπου ο δείκτης i αντιστοιχεί στα πλήθη των παρατηρούμενων και αναμενόμενων γονοτύπων αντίστοιχα. Πρέπει να σημειώσουμε εδώ ότι η 10.4 εφαρμόζεται στους καθαρούς αριθμούς των ατόμων που ανήκουν σε κάθε γονότυπο (αναμενόμενο και παρατηρούμενο) κι όχι στις συχνότητες εμφάνισής τους. Έτσι για έναν πληθυσμό 18000 ατόμων όπως είναι ο s , η αναμενόμενη τιμή για το γονότυπο AA θα είναι $18000 \times 0.96 = 17280$, ενώ η αντίστοιχη παρατηρούμενη θα ισούται με

$18000 \times 0.95 = 17100$. Με ανάλογο τρόπο και με αντικατάσταση παρατηρούμενων και αναμενόμενων συχνοτήτων η εξίσωση 10.4 θα δώσει:

$$\chi^2 = \chi^2[AA] + \chi^2[Aa] + \chi^2[aa] = \frac{(17100 - 17280)^2}{17280} + \frac{(888 - 712)^2}{712} + \frac{(12 - 7)^2}{7} = 47.38 \quad 10.5$$

Τι μας λέει όμως μια τιμή $\chi^2 = 47.38$; Όπως και σ' όλους τους στατιστικούς ελέγχους που έχουμε συζητήσει ως εδώ, χρειαζόμαστε μια ποσότητα που να μπορεί να αντιστοιχηθεί σε μια πιθανότητα παρατήρησης των συγκεκριμένων αριθμητικών τιμών. Για την ποσότητα του χ^2 αυτό γίνεται είτε ανατρέχοντας σε στατιστικούς πίνακες που αντιστοιχούν τις τιμές χ^2 σε τιμές p-value, είτε μέσω της απευθείας εφαρμογής του ελέγχου χ^2 σε υπολογιστικά πακέτα (όπως η R, το GraphPad κλπ). Ανατρέχοντας σ' έναν τέτοιο πίνακα για $\chi^2 = 47.38$ θα πρέπει επίσης να δηλώσουμε τους βαθμούς ελευθερίας (degrees of freedom, ή για συντομία df) του συστήματός μας. Αυτοί υπολογίζονται ως εξής:

$$df = \text{Αριθμός των κατηγοριών δεδομένων} - \text{Αριθμός παραμέτρων που υπολογίζονται} \quad 10.6$$

Στην περίπτωση μας, οι κατηγορίες είναι 3 (οι τρεις γονότυποι) ενώ ο αριθμός των παραμέτρων που υπολογίζουμε είναι μία, η συχνότητα του ελάσσονος αλληλομόρφου q αφού η p μπορεί να υπολογιστεί απευθείας από την q. Προκύπτει έτσι ότι για $df = 3 - 1 = 2$, η τιμή $\chi^2 = 47.38$ αντιστοιχεί σε $p\text{-value} < 0.0001$, είναι δηλαδή εξαιρετικά σημαντική από στατιστική άποψη. Η μικρή τιμή p-value αντιστοιχεί στην πιθανότητα παρατήρησης των συγκεκριμένων τιμών, δεδομένου ότι ισχύει η μηδενική υπόθεση, η υπόθεση δηλαδή που κάναμε αρχικά ότι ο υποπληθυσμός *s* είναι μέρος του γενικότερου πληθυσμού και ως εκ τούτου περιέχει τα δύο αλληλόμορφα με ανάλογες συχνότητες. Χαμηλές τιμές p-value συνηγορούν έτσι υπέρ της απόρριψης της μηδενικής υπόθεσης. Κάτι τέτοιο ήταν αναμενόμενο στο συγκεκριμένο παράδειγμα ήδη από τη στιγμή που υπολογίσαμε τη συχνότητα εμφάνισης της ασθένειας ($1/1500$ για τον υποπληθυσμό *s* σε σχέση με $1/2500$ για το γενικότερο), ωστόσο, η σύγκριση δεδομένων γενετικής ποικιλομορφίας θα πρέπει να γίνεται πάντα με πολύ μεγάλη προσοχή και λαμβάνοντας υπόψη τον όγκο των δεδομένων που συγκρίνουμε. Ειδικά για την περίπτωση του χ^2 το μέγεθος του δείγματος έχει πολύ μεγάλη σημασία. Για να γίνει αυτό κατανοητό αρκεί να δούμε τι συμβαίνει αν το μέγεθος του υποπληθυσμού *s* μειωθεί στο ένα δέκατο (από 18000 σε 1800 άτομα). Κρατώντας τις ίδιες συχνότητες εμφάνισης αλληλομόρφων παίρνουμε $p\text{-value} = 0.03$, μια τιμή οριακά κάτω από το συνηθισμένο κατώφλι του 0.05, ενώ για έναν πληθυσμό 180 ατόμων η τιμή γίνεται $p = 0.69$.

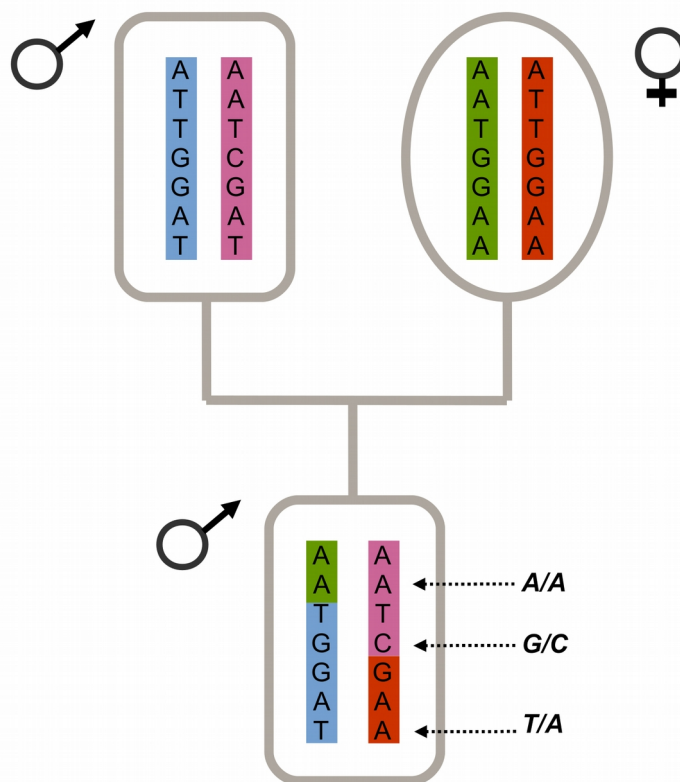
Η βάση της γενετικής ποικιλομορφίας

Ανασυνδυασμός και Γενετική Σύνδεση

Μέχρι στιγμής έχουμε εξετάσει τη βασική μονάδα της γενετικής ποικιλομορφίας που στηρίζεται στην έννοια του αλληλόμορφου³. Πώς όμως διαμορφώνεται η συνολική ποικιλομορφία

³ Όπως αναφέραμε και παραπάνω, η έννοια του αλληλόμορφου δεν αφορά πλέον αποκλειστικά το γονίδιο αλλά εξαρτάται από το βαθμό διακριτικής ικανότητας κατά την ανάλυση του γονιδιώματος. Έτσι μπορούμε να αναφερόμαστε και σε SNP ως αλληλόμορφους, όπως

σε πληθυσμιακό επίπεδο; Η βάση για το παρατηρούμενο τεράστιο εύρος φαινοτυπικών παραλλαγών έγκειται στο εξίσου ασύλληπτο πλήθος των δυνατών συνδυασμών των διαφόρων αλληλομόρφων. Στην περίπτωση του ανθρώπινου γονιδιώματος, που όπως είδαμε πιο πάνω περιέχει περίπου 1.5×10^6 SNP, ο αριθμός των πιθανών συνδυασμών αυτών των πολυμορφισμών είναι αστρονομικού (ή καλύτερα πλέον, γονιδιωματικού) μεγέθους.

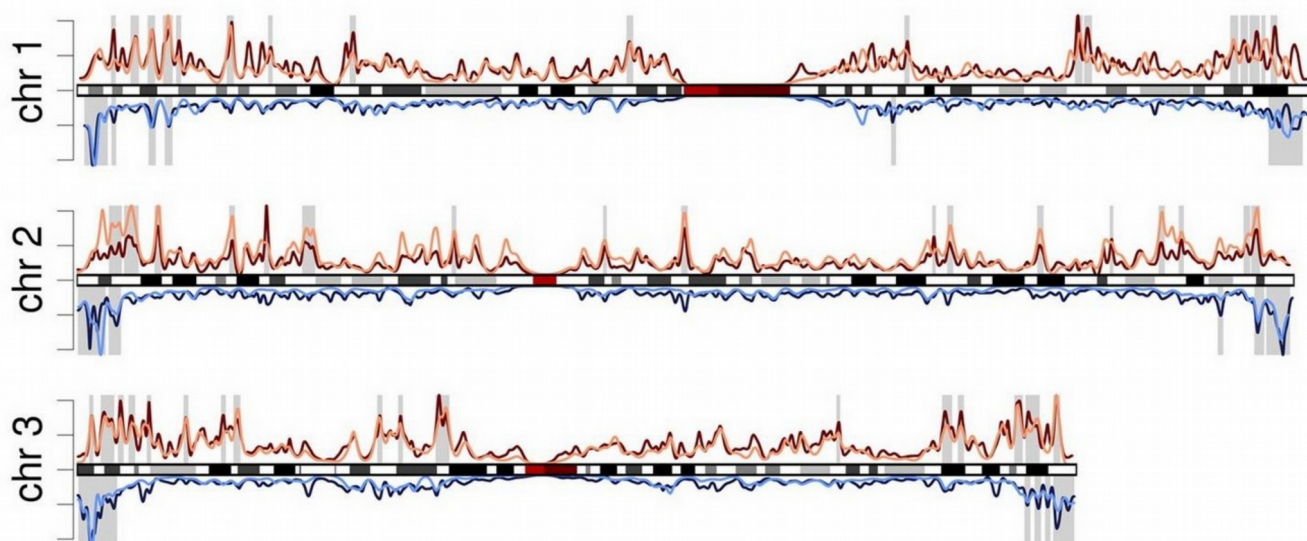


Εικόνα 10.3: Σχηματική αναπαράσταση του ομόλογου ανασυνδυασμού κατά τη φυλετική αναπαραγωγή. Τμήματα των χρωμοσωμάτων από τα γονιδιώματα των δύο γονέων (χρωματιστές στήλες πάνω) συνδυάζονται κατά μήκος με αποτέλεσμα το γονιδίωμα του απογόνου να περιέχει ομόλογα τμήματα και από τους δύο γονείς σε κάθε χρωμοσωμικό αντίγραφο. Στο σχήμα της Εικόνας ο απόγονος είναι ομοζυγωτικός για τη θέση 2 (A/A) (στην οποία οι γονείς του ήταν ετερόζυγοι), κληρονομεί την ετεροζυγωτία από το πατρικό γονιδίωμα στη θέση 4 (G/C) και γίνεται ετερόζυγος στη θέση 7 (T/A) για την οποία οι γονείς του ήταν ομόζυγοι σε διαφορετικό αλληλόμορφο.

Φυσικά το σύνολο όλων αυτών των συνδυασμών δεν πρόκειται να πραγματοποιηθεί λόγω -εκτός του τρομακτικού αριθμού τους- και κάποιων εγγενών περιορισμών όπως οι σχετικές συχνότητες των SNP, ή η ύπαρξη υποπληθυσμών που δεν αναμιγνύουν με την ίδια συχνότητα το γενετικό τους υλικό λόγω γεωγραφικής απομόνωσης (ακόμα και στην εποχή της παγκοσμιοποίησης, είναι αρκετά σπάνιο ένας Εσκιμώος να ζευγαρώσει με μια Νεοζηλανδή Μαορί). Ένα μεγάλο μέρος όμως των πιθανών συνδυασμών αλληλομόρφων πραγματοποιείται και συνεχίζει να εξελίσσεται δυναμικά σε πληθυσμιακό επίπεδο. Ο βιολογικός μηχανισμός που επιτρέπει τη διαμόρφωσή τους είναι η φυσική διαδικασία του γενετικού ανασυνδυασμού (genetic recombination), μέσω της οποίας

κάναμε στην αμέσως προηγούμενη ενότητα.

η φυλετική αναπαραγωγή οδηγεί σε απογόνους που διαφέρουν γενετικά από τους γονείς τους. Μέσω του ανασυνδυασμού, χρωμοσωμικά τμήματα που είναι ομόλογα (αφορούν δηλαδή τον ίδιο γενετικό τόπο στα δύο αντίγραφα των χρωμοσωμάτων) αλληλοδιασταυρώνονται όπως φαίνεται σχηματικά στην Εικόνα 10.3. Το αποτέλεσμα είναι ότι τμήματα από τα δύο γονεϊκά ανακατανέμονται με διαφορετικό τρόπο σε κάθε γεγονός αναπαραγωγής ώστε να προκύπτει ένας διαφορετικός γονότυπος κάθε φορά. Η διαδικασία του ανασυνδυασμού, πολύ περισσότερο από την ανάδυση νέων πολυμορφισμών και μεταλλαγών είναι η βάση της γενετικής ποικιλομορφίας στους φυλετικά αναπαραγόμενους πληθυσμούς.



Εικόνα 10.4: Χρωμοσωμικοί χάρτες για τα τρία μεγαλύτερα ανθρώπινα χρωμοσώματα που δείχνουν τη διαφορετική συχνότητα ανασυνδυασμού κατά μήκος του γονιδιώματος. Με κόκκινο η σχετική συχνότητα ανασυνδυασμού στους άνδρες και με γαλάζιο στις γυναίκες. Περιοχές όπου η συχνότητα ανασυνδυασμού ξεπερνά τη μέση τιμή κατά δύο μονάδες τυπικής απόκλισης είναι σκιασμένες με γκριζο και συνιστούν "ζούγκλες ανασυνδυασμού". (Εικόνα προσαρμοσμένη από την εργασία των Chowdhury et al 2009, CC BY 3.0).

Τόσο η έκταση των περιοχών που ανασυνδυάζονται όσο και η συχνότητα με την οποία αυτό συμβαίνει καθώς ο πληθυσμός εξελίσσεται διαφέρουν σημαντικά κατά μήκος του γονιδιώματος. Στην Εικόνα 10.4 αυτές οι διαφορές αποτυπώνονται σχηματικά σ' ένα χάρτη του ανθρώπινου γονιδιώματος (Chowdhury, Bois, Feingold, Sherman & Cheung, 2009). Το μήκος των περιοχών που ανασυνδυάζονται είναι ένα σημαντικό μέγεθος για τη γενετική και στην πραγματικότητα αποτέλεσε τη βάση για τις πρώτες μονάδες μέτρησης μήκους των γονιδιωμάτων πριν η σύγχρονη γονιδιωματική περάσει στην εποχή της λεπτομερούς αλληλούχισής τους. Ο Αμερικανός γενετιστής Thomas Hunt Morgan ήταν από τους πρώτους που μελέτησαν την έννοια της γενετικής σύνδεσης (genetic linkage), για να περιγράψει την τάση στοιχείων του γονιδιώματος να ανήκουν στην ίδια "μονάδα ανασυνδυασμού" (Morgan, 1911).

Η έννοια της γενετικής σύνδεσης περιγράφει την τάση που έχουν δύο (ή περισσότεροι) χαρακτήρες να κληρονομούνται μαζί. Η σχέση αυτή είναι άμεσα συνδεδεμένη με τις σχετικές τους θέσεις στο

γονιδίωμα, καθώς αποτελούν τμήμα μιας χρωμοσωμικής περιοχής που κληρονομείται πάντοτε (ή για την ακρίβεια με πολύ μεγάλη συχνότητα) αυτούσια, χωρίς να διακόπτεται από γεγονότα ανασυνδυασμού. Προς τιμήν του Morgan, η βασική μονάδα μέτρησης της γενετικής σύνδεσης που αντιστοιχεί στην απόσταση πάνω στο ίδιο χρωμόσωμα μεταξύ δύο γενετικών χαρακτήρων που έχουν πιθανότητα ίση με 1 να διακοπούν από ένα γεγονός ανασυνδυασμού, ονομάζεται Morgan⁴. Έτσι λέμε ότι δύο γενετικοί τόποι απέχουν 0.01 Morgan ή 1cM (centi-Morgan) αν μεταξύ τους υπάρχει πιθανότητα να συμβούν 0.01 γεγονότα ανασυνδυασμού ανά γενιά.

Απλότυποι και Στατιστική Γενετική Σύνδεση

Η γενετική σύνδεση επέτρεψε μια εκρηκτική άνθιση στη μελέτη της γενετικής ποικιλομορφίας καθώς παρείχε το βασικό θεωρητικό εργαλείο για την ανάλυση του τρόπου με τον οποίο συνδυάζονται και κληρονομούνται τα αλληλόμορφα (Bateson, Waunders & Punnett, 1909). Σήμερα, με τις σύγχρονες μεθοδολογίες που επιτρέπουν τη γονοτύπηση σε μονονουκλεοτιδικό επίπεδο, μπορούμε να προσδιορίσουμε με μεγάλη ακρίβεια όχι μόνο τα όρια των περιοχών του γονιδιώματος που κληρονομούνται από τον ίδιο γονέα αλλά και την πιθανότητα να συμβεί αυτό καθώς και την πιθανότητα που έχουν να κληρονομηθούν μαζί με άλλες γειτονικές περιοχές τους (Gunderson et al., 2005). Περνώντας έτσι από το επίπεδο των απλών πολυμορφισμών μπορούμε να ορίσουμε μια διευρυμένη μονάδα γενετικής ποικιλομορφίας ως την περιοχή του γονιδιώματος μέσα στην οποία οι πολυμορφισμοί έχουν μεγάλη πιθανότητα να κληρονομούνται από τον ίδιο γονέα. Οι μονάδες αυτές, που τις ονομάζουμε haploblock (μια προσπάθεια απόδοσης του όρου στα ελληνικά θα ήταν “δέσμες απλοτύπων”) συνθέτουν το γενετικό πρότυπο ευρύτερων ομάδων του πληθυσμού. Παρόλο δηλαδή που κάθε άτομο μπορεί να φέρει διαφορετικά αλληλόμορφα (SNP, CNV) σε συγκεκριμένες θέσεις, οι δέσμες, τα block του γονιδιώματος μέσα στις οποίες αυτά οριοθετούνται είναι συχνά καλά ορισμένες. Το σύνολο των δεσμών αυτών και των μεταξύ τους ορίων συνιστούν αυτό που ονομάζεται απλότυπος (haplotype) και που αντιπροσωπεύει το συνδυασμό του γονοτύπου του κάθε ατόμου σε σχέση με την κατανομή των haploblock στα χρωμοσώματά του.

Ερώτηση: Αν το γενετικό προφίλ ενός ατόμου σ' έναν πληθυσμό δίνεται από το σύνολο των πολυμορφισμών του, μπορείτε να σκεφτείτε τι χαρακτηρίζει ένας απλότυπος;

Όπως αναφέραμε και παραπάνω, ο απλότυπος ορίζει περισσότερο ένα υποσύνολο ενός πληθυσμού παρά ένα συγκεκριμένο άτομο, καθώς πολλά άτομα του ίδιου πληθυσμού μπορούν να μοιράζονται την ίδια διάταξη haploblocks χωρίς υποχρεωτικά αυτές να περιέχουν τους ίδιους πολυμορφισμούς. Στην πραγματικότητα, τα περισσότερα άτομα του ανθρώπινου πληθυσμού αντιστοιχούν σε μερικές δεκάδες μόνο απλοτύπους. Η χαρτογράφηση των απλοτύπων ενός μεγάλου αριθμού ατόμων αποτελεί το στόχο μεγάλων ερευνητικών προσπαθειών όπως είναι το Human HarMap Project (The International HarMap Consortium,

⁴ Στην πράξη η μονάδα Morgan χρησιμοποιείται σπανίως έως καθόλου. Στη θέση της χρησιμοποιούμε την εκατοστιαία υποδιαίρεσή της cM (centi-Morgan) που αντιστοιχεί σε πιθανότητα ίση με 0.01 για ένα γεγονός συνδυασμού μεταξύ δύο γενετικών τόπων.

2003) (<http://hapmap.ncbi.nlm.nih.gov/>). Απώτερη επιδίωξη τέτοιων ερευνητικών προγραμμάτων είναι κυρίως η μελέτη του τρόπου εξέλιξης του ανθρώπινου γονιδιώματος σε επίπεδο δομής πληθυσμού και κατά δεύτερο λόγο του εντοπισμού στοιχείων του γονιδιώματος που να μπορούν να χρησιμεύσουν στην εξακρίβωση της μηχανιστικής βάσης γενετικών ασθενειών.

Πώς δημιουργείται όμως ένας χάρτης απλοτύπου για ένα δεδομένο δείγμα του πληθυσμού και πως ορίζονται οι περιοχές *haploblock*; Η βάση για την ανάλυσή τους είναι οι μελέτες γενετικής συσχέτισης. Η γενετική συσχέτιση (*statistical genetic association*) σε αντίθεση με την γενετική σύνδεση (*linkage*) δεν είναι αποτέλεσμα απευθείας γενετικής σύνδεσης αλλά της στατιστικής εκτίμησής της, από έναν μεγάλο αριθμό γονοτύπων. Η γενετική σύνδεση προϋποθέτει ότι γνωρίζει κανείς τους γονοτύπους μεταξύ κοντινών συγγενών καθώς και τις μεταξύ τους συγγενικές θέσεις, με τρόπο που να μπορεί να αντιπαραβάλλει το γενεαλογικό τους δέντρο με τη γενετική πληροφορία και να εξαγάγει έτσι, απευθείας συμπεράσματα για τον τρόπο με τον οποίο έχουν κληρονομηθεί τα διάφορα χαρακτηριστικά. Τι συμβαίνει όμως στην περίπτωση που δεν είναι δυνατόν να γονοτυπηθεί μια οικογένεια και που τα χαρακτηριστικά που μας ενδιαφέρουν δεν είναι εύκολο να αποδοθούν στα μέλη της; Φανταστείτε για παράδειγμα ότι μελετούμε μια παθολογική κατάσταση που είναι αρκετά σπάνια για να την εντοπίσουμε μέσα σε μια οικογένεια. Η γενετική συσχέτιση μας επιτρέπει να μελετήσουμε τη γενετική ποικιλομορφία σ' ένα μεγάλο αριθμό ατόμων και να χρησιμοποιήσουμε στατιστικούς εκτιμητές για να εντοπίσουμε τις γονιδιωματικές περιοχές που πιθανότατα εμπλέκονται στην εκδήλωσή της. Το *HapMap Project* για παράδειγμα, έχει συγκεντρώσει μεγάλο αριθμό από SNP από ένα μεγάλο αριθμό ατόμων και μέσω γενετικής συσχέτισης έχει ανασυστήσει τους διαφορετικούς απλοτύπους για το σύνολο του ανθρώπινου πληθυσμού. Με τον τρόπο αυτό, παρέχει ένα σύστημα παράλληλων χαρτών που μπορούν να χρησιμοποιηθούν για τον εντοπισμό γενετικών παραλλαγών που είναι σημαντικές για παθολογικές καταστάσεις. Λεπτομερής χαρτογράφηση επιτρέπει έτσι τον εντοπισμό της γενετικής βάσης για την εκδήλωση ενός φαινοτύπου, καθώς αν ένα γενετικό χαρακτηριστικό κληρονομείται μαζί με συγκεκριμένα SNP των οποίων τη θέση και το πρότυπο κληρονόμησης γνωρίζουμε μέσω των απλοτύπων, τότε μπορούμε εύκολα να εντοπίσουμε το γονίδιο (ή τη γενικότερη περιοχή) στην οποία εδράζεται το γενετικό χαρακτηριστικό αυτό.

Ανισορροπία Γενετικής Σύνδεσης (Linkage Disequilibrium)

Στην Εικόνα 10.5 βλέπουμε έναν χαρακτηριστικό τρόπο αναπαράστασης μιας χρωμοσωμικής περιοχής σε διακριτές δέσμες *haploblock*. Παράλληλα με το μήκος της χρωμοσωμικής αλληλουχίας ορίζονται τριγωνικές περιοχές σ' ένα θερμικό χάρτη, με τρόπο ανάλογο με αυτόν που είδαμε στο Κεφάλαιο 5 για την αναπαράσταση χρωμοσωμικών αλληλεπιδράσεων σε τρεις διαστάσεις. Κάθε σημείο στο θερμικό χάρτη αντιστοιχεί έτσι σε δύο γονιδιωματικές συντεταγμένες που προκύπτουν από τις προβολές του σημείου στη χρωμοσωμική αλληλουχία σε γωνίες 45°. Το χρώμα του κάθε σημείου αντιστοιχεί σε μια χαρακτηριστική τιμή που υποδηλώνει τη μη-τυχαία γενετική σύνδεση δύο γενετικών τόπων. Η τιμή αυτή ονομάζεται ανισορροπία γενετικής σύνδεσης (*linkage disequilibrium* και για συντομία στο εξής *LD*) και ορίζεται ως εξής:

Υποθέτουμε ότι ένα αλληλόμορφο A έχει συχνότητα p_A και αντίστοιχα ένα άλλο αλληλόμορφο B (για διαφορετικό γενετικό χαρακτήρα) έχει συχνότητα p_B . Έστω τώρα ότι η συχνότητα εμφάνισης των δύο αλληλομόρφων στον ίδιο πληθυσμό είναι p_{AB} . Αν εξετάσουμε τις συχνότητες εμφάνισης ως πιθανότητες (όπως έχουμε κάνει στα Κεφάλαια 1, 2 και 3) και με βάση τη βασική ιδιότητα της ανεξαρτησίας (βλ. Κεφάλαιο 2) μπορούμε να υποθέσουμε ότι αν η συνεμφάνιση των αλληλομόρφων A και B στο ίδιο άτομο είναι αποτέλεσμα δύο τελείως ανεξάρτητων ενδεχομένων τότε η συχνότητα αυτής της συνεμφάνισης θα είναι ίση με το γινόμενο των δύο επιμέρους συχνοτήτων:

$$p_{AB} = p_A p_B \quad 10.6$$

Η σχέση 10.6 αντιστοιχεί στην πλήρη ανεξαρτησία της σύνδεσης των A και B. Λέμε τότε ότι τα δύο αλληλόμορφα βρίσκονται σε ισορροπία σύνδεσης. Κάτι τέτοιο σημαίνει ότι, στην περιοχή στην οποία βρίσκονται, δεν υφίστανται περιορισμοί σχετικά με τον ανασυνδυασμό⁵. Τι συμβαίνει όμως στην περίπτωση που η 10.6 δεν ισχύει; Είναι λογικό τότε να υποθέσουμε ότι κάποιος περιορισμός οδηγεί τα A και B να συνεμφανίζονται με μεγαλύτερη ή μικρότερη συχνότητα από την αναμενόμενη (Hill & Robertson, 1968). Ο περιορισμός αυτός είναι στις περισσότερες περιπτώσεις γενετικός και οφείλεται στη γενετική σύνδεση. Ορίζουμε την ανισορροπία σύνδεσης ως τη διαφορά μεταξύ των δύο μελών της εξίσωσης 10.6:

$$D_{AB} = p_{AB} - p_A p_B \quad 10.7$$

Η τιμή της D_{AB} αφορά ειδικά τα αλληλόμορφα A και B κι έτσι εξαρτάται αποκλειστικά από τις σχετικές συχνότητες των A και B. Διαφορετικά αλληλόμορφα σε πολύ παραπλήσιες γονιδιωματικές θέσεις μπορεί να εμφανίζουν διαφορετικές τιμές D εξαιτίας των συχνοτήτων εμφάνισής τους, κάτι που δυσκολεύει τη συνολική περιγραφή γονιδιωματικών περιοχών με όρους γενετικής σύνδεσης. Για παράδειγμα, ένα ζεύγος πολύ κοινών αλληλομόρφων K_1, K_2 που βρίσκεται στην ίδια περιοχή με ένα ζεύγος σπάνιων αλληλομόρφων Σ_1, Σ_2 θα έδινε για την ίδια χρωμοσωμική περιοχή δύο πολύ διαφορετικές τιμές $D_{K_1K_2}$ και $D_{\Sigma_1\Sigma_2}$. Έτσι, ένας καλύτερος τρόπος για να εκφράσουμε τη διαφορά της σχέσης 10.7 είναι να κανονικοποιήσουμε την τιμή D συγκρίνοντάς τη με τη μέγιστη τιμή που θα μπορούσε να πάρει:

$$D'_{AB} = D_{AB} / D_{AB_{max}} \quad 10.8$$

όπου:

$$D_{AB_{max}} = \begin{cases} \min(p_A p_B, (1-p_A)(1-p_B)) & \text{για } D < 0 \\ \min(p_A(1-p_B), p_B(1-p_A)) & \text{για } D > 0 \end{cases} \quad 10.9$$

Η εξίσωση 10.9 ουσιαστικά υπολογίζει τη μέγιστη τιμή της D για κάθε υποπερίπτωση σχετικά με τις σχετικές συχνότητες των A και B. Στην περίπτωση που $D < 0$ αυτό σημαίνει ότι τα δύο αλληλόμορφα εμφανίζονται με μικρότερη συχνότητα από την αναμενόμενη (με βάση την εξίσωση 10.7) οπότε ο έλεγχος γίνεται με βάση την πιθανότητα συνεμφάνισης των δύο αλληλομόρφων ή των συγγενικών τους $(1-p)$. Για $D > 0$ η συχνότητα συνεμφάνισης είναι μεγαλύτερη από την αναμενόμενη

⁵ Στην πραγματικότητα ο ανασυνδυασμός δεν είναι ο μόνος μηχανισμός που διαμορφώνει την ανισορροπία σύνδεσης αλλά στο βαθμό που μπορούμε να υποθέσουμε ότι δεν υπάρχει φυσική επιλογή, η τυχαία γενετική παρέκκλιση είναι αμελητέα και ο πληθυσμός αναπαράγεται τελείως τυχαία, τότε η επίδραση του ανασυνδυασμού είναι κυρίαρχη.

και τότε οι πιθανότητες αναφοράς είναι αυτές που αφορούν συνδυασμούς του ενός αλληλόμορφου με το συγγενικό του άλλου.

Ένας άλλος τρόπος για να προκύψει μια σταθμισμένη ποσότητα γενετικής σύνδεσης είναι ο συντελεστής γενετικής συσχέτισης που συμβολίζουμε με r^2 και που σχετίζεται με την D μέσω της σχέσης:

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)} \quad 10.10$$

από την οποία προκύπτει ότι για $r^2=1$ έχουμε πλήρη ανισορροπία σύνδεσης ενώ για $r^2=0$ έχουμε πλήρη ισορροπία. Στην Εικόνα 10.5 μπορείτε να δείτε πως υπολογισμοί για την ανισορροπία σύνδεσης οδηγούν στον ορισμό *haploblocks* σ' ένα δεδομένο γενετικό τόπο. Η πρακτική σημασία των υψηλών τιμών r^2 είναι ότι αντιστοιχούν σε αλληλόμορφα (ή SNP) που είναι συσχετισμένα και παρατήρηση του ενός από τα δύο παρέχει πληροφορία για την ύπαρξη και του άλλου. Περιοχές με συστηματικά χαμηλές τιμές r^2 αντιστοιχούν σε περιοχές μεταξύ των οποίων οι γενετικοί χαρακτήρες είναι σε ισορροπία σύνδεσης και συνεπώς κληρονομούνται ανεξάρτητα από τα γονεϊκά χρωμοσώματα. Περιοχές μεταξύ των οποίων οι χαμηλές τιμές μεταβάλλονται σε υψηλές αντιστοιχούν σε όρια μεταξύ διαφορετικών *haploblocks*, μεταξύ των οποίων η πιθανότητα ανασυνδυασμού είναι μεγάλη.

Μελέτες Γενετικής Συσχέτισης (Gene Association Studies)

Έχοντας ως εδώ περιγράψει τα επίπεδα στα οποία εκφράζεται η γενετική ποικιλομορφία και τους τρόπους με τους οποίους μπορούμε να τα χαρτογραφήσουμε, θα επιστρέψουμε στο αρχικό μας ερώτημα που έχει να κάνει με τη σχέση ενός παρατηρούμενου φαινοτύπου με τον υποκείμενο γονότυπο μεταξύ ατόμων του ίδιου πληθυσμού. Στην περίπτωση του ανθρώπινου γονιδιώματος, η γνώση, από τη μια πλευρά, των πολυμορφικών θέσεων και από την άλλη, των μεταξύ τους συσχετίσεων μέσω των σχέσεων LD, μας επιτρέπει να διερευνήσουμε επιμέρους διαφορές μεταξύ ατόμων με σκοπό να εντοπίσουμε αιτιακές συνδέσεις γονοτύπων-φαινοτύπων.

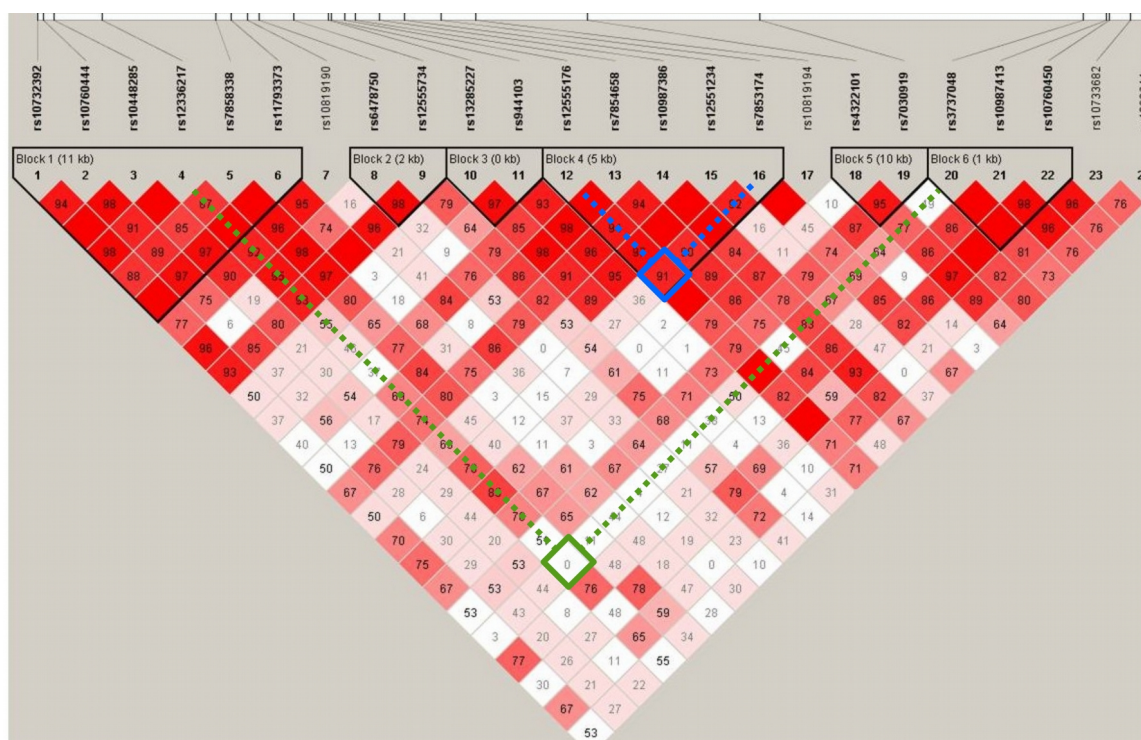
Η βασική μεθοδολογία για να γίνει αυτή η σύνδεση είναι μέσω μελετών γενετικής σύνδεσης (*association studies*), οι οποίες διεξάγονται πλέον σε γονιδιωματική κλίμακα μέσω της ανάπτυξης τεχνικών μικροσυστοιχειών και αλληλούχισης νέας γενιάς. Μιλάμε έτσι πλέον για μελέτες γενετικής σύνδεσης σε γονιδιωματική κλίμακα (*Genome-Wide Association Studies*, ή για συντομία στο εξής GWAS). Η ενοποιητική αρχή όλων των μελετών σύνδεσης είναι η διάρθρωση της μελέτης σε τρεις άξονες:

- Την καταγραφή των πολυμορφισμών μεταξύ ατόμων του εξεταζόμενου δείγματος.
- Τη συσχέτιση της ποικιλομορφίας με κάποιο/α φαινοτυπικό/α χαρακτηριστικό/α.
- Τη στατιστική εκτίμηση των εξαγόμενων συσχετίσεων.

Τα τελευταία χρόνια οι μελέτες σύνδεσης σε γονιδιωματική κλίμακα έχουν αποτελέσει την “αιχμή του δόρατος” για την πρόοδό μας στην κατανόηση πολύπλοκων ασθενειών μέσω του συνδυασμού της δυνατότητας ανάλυσης μεγάλων δειγμάτων πληθυσμού και των αυξανόμενων

γνώσεών μας για τη δομή των πληθυσμών αυτών. Το 2005, μια από τις πρώτες τέτοιες μελέτες σε μεγάλη κλίμακα οδήγησε στην εξακρίβωση της γενετικής βάσης του εκφυλισμού της ωχράς κηλίδας (Klein et al., 2005), δίνοντας έτσι σημαντική ώθηση για την εφαρμογή GWAS προσεγγίσεων σε μια σειρά από πολύπλοκες ασθένειες και σύνδρομα όπως είναι τα αυτοάνοσα νοσήματα (Okada et al., 2013), οι καρδιαγγειακές παθήσεις κλπ. Παρά το σημαντικό κόστος που επισύρουν και τις εγγενείς δυσκολίες στην οργάνωση και συλλογή των δειγμάτων, πάνω από 500 τέτοιες μελέτες έχουν δημοσιευθεί έκτοτε και ο αριθμός τους συνεχώς αυξάνει.

Ανάλογα με το είδος του δείγματος αλλά και την εξεταζόμενη φαινοτυπική ιδιότητα, οι GWAS μπορούν να διακριθούν στις παρακάτω κατηγορίες.



Εικόνα 10.5: Αναπαράσταση του βαθμού LD για μια περιοχή του γενετικού τόπου του ανθρώπινου γονιδίου *LMX1B*. Σε έναν τέτοιο τριγωνικό θερμικό χάρτη κάθε τετράγωνο αντιστοιχεί στην τιμή LD μεταξύ των περιοχών του γονιδιώματος που βρίσκονται στις προβολές του τετραγώνου σε γωνία 45 μοιρών. Διακρίνονται περιοχές με υψηλές τιμές LD (π.χ. γαλάζιο πλαίσιο) που οδηγούν στον καθορισμό block απλοτύπων (haploblocks), ενώ περιοχές με χαμηλή τιμή LD αντιστοιχούν σε περιοχές που δεν ανήκουν στο ίδιο haploblock (πράσινο πλαίσιο). (Η Εικόνα προσαρμόστηκε από την εργασία των Thanseem et al., 2011, CC BY 3.0).

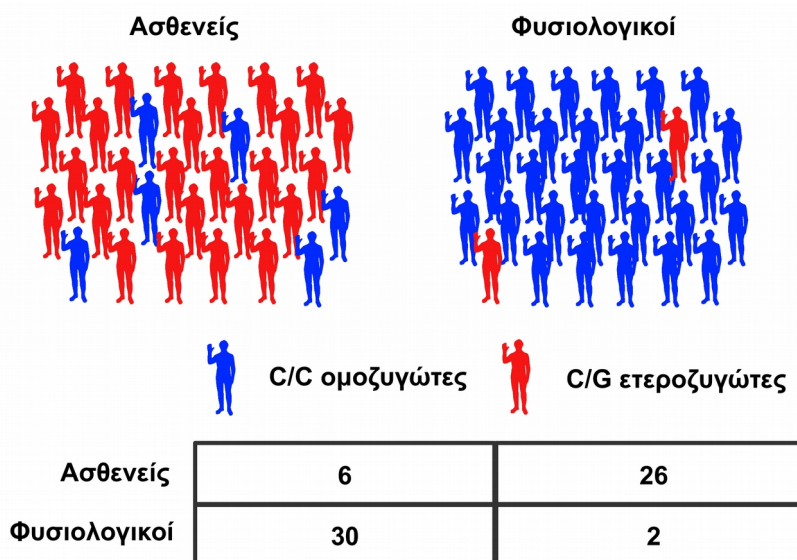
Μελέτες συσχέτισης σε οικογενειακό επίπεδο

Σ' αυτές αναλύεται η γενετική ποικιλομορφία σε περιορισμένο αριθμό ατόμων της ίδιας οικογένειας. Όπως αναφέρθηκε και παραπάνω, το πλεονέκτημα των μεθόδων αυτών είναι ότι εκμεταλλεύονται την κοινή απλοτυπική διάταξη που αναμένεται μεταξύ μελών της ίδιας οικογένειας.

Μειονέκτημά τους είναι ότι δύσκολα μπορούν να εντοπίσουν συσχετίσεις μεταξύ γονοτύπου και πολύπλοκων φαινοτυπικών χαρακτήρων για λόγους που θα εξετάσουμε στη συνέχεια. Από την άλλη πλευρά, γονοτυπικές αλλαγές που έχουν πολύ μεγάλη επίδραση στο φαινότυπο (και συνεπώς αναμένεται να είναι αρκετά σπάνιες), μελετώνται κατά προτίμηση μεταξύ των μελών της ίδιας οικογένειας (Marenberg, Risch, Berkman, Floderus & de Faire, 1994).

Μελέτες ασθενών-μαρτύρων (case-control studies)

Οι μελέτες ασθενών μαρτύρων αποτελούν το είδος των GWAS που διενεργείται με τη μεγαλύτερη συχνότητα. Βασίζονται στο διαχωρισμό του δείγματος σε δύο κατηγορίες με βάση ένα φαινοτυπικό χαρακτηριστικό δυαδικού τύπου (π.χ. ασθενείς-υγιείς) (Sawcer et al., 2011). Πολύ συχνά, ο δυαδικός χαρακτήρας του φαινοτύπου δεν προσδιορίζεται παρά μόνο προσεγγιστικά ως τέτοιος, για παράδειγμα ένας ασθενής από ένα υγιές άτομο μπορεί να διαφέρουν σ' ό,τι αφορά τα επίπεδα έκφρασης μιας πρωτεΐνης και όχι στην έκφραση ή μη-έκφρασή της, ωστόσο η γνώση του συστήματος που μελετούμε περιορίζει αποφασιστικά και το είδος της μελέτης. Σε περιπτώσεις πολύπλοκων ασθενειών ή καταστάσεων που αναμένεται να εξαρτώνται από περισσότερα από ένα χαρακτηριστικά η δυαδική προσέγγιση είναι αυτή που επιλέγεται υποχρεωτικά.



Εικόνα 10.6: Γραφική αναπαράσταση μιας γενετικής ανάλυσης ασθενών-μαρτύρων (case-control study). Σε δύο δείγματα πληθυσμού που αποτελούνται από 32 ασθενή και 32 φυσιολογικά άτομα αντίστοιχα, υπολογίζεται ο αριθμός των φορέων των αλληλομόρφων για ένα συγκεκριμένο γενετικό τόπο. Οι σχετικές συχνότητες ομοζυγωτών και ετεροζυγωτών μεταξύ των δύο πληθυσμών μπορούν να αναπαρασταθούν στον πίνακα σύμπτωσης που φαίνεται στο κάτω μέρος της Εικόνας.

Στην περίπτωση που γνωρίζουμε το γενετικό χαρακτήρα που επηρεάζει το φαινότυπο και προσπαθούμε να προσδιορίσουμε την ποσοτική τους σχέση, ο σχεδιασμός της μελέτης αλλάζει και τότε μιλάμε για ποσοτική μελέτη (quantitative study design). Στις περιπτώσεις αυτές, ο γενετικός

τόπος ενδιαφέροντος είναι συνήθως γνωστός κάτι που διευκολύνει την ανάλυση των δειγμάτων, καθώς απαιτείται μικρότερης έκτασης γονοτύπηση. Από την άλλη πλευρά η εξαγωγή ποσοτικών συσχετίσεων απαιτεί μεγαλύτερο αριθμό δειγμάτων προκειμένου να επιτευχθεί ικανοποιητική στατιστική ισχύς (Stranger et al., 2007).

Ένα σημαντικό στοιχείο που διέπει όλες τις μελέτες ασθενών-μαρτύρων είναι η ανάγκη επαρκούς τυποποίησης των φαινοτυπικών κριτηρίων. Ένα από τα σημαντικότερα προβλήματα που προκύπτουν κατά την ανάλυση των δεδομένων είναι η λανθασμένη απόδοση ατόμων μεταξύ των κατηγοριών καθώς συχνά ο χαρακτηρισμός ενός ασθενούς γίνεται με μη ικανοποιητικά κριτήρια. Ο αυστηρός έλεγχος στον καθορισμό του δείγματος από εξειδικευμένο ιατρικό προσωπικό που θα σταθμίσει όλα τα απαραίτητα κριτήρια είναι ιδιαίτερα σημαντικός στις περιπτώσεις που οι ασθενείς και οι μάρτυρες προέρχονται από διαφορετικά κέντρα αναφοράς ή συχνά και από διαφορετικές χώρες.

Στοιχεία γενετικής ανάλυσης

Ποια είναι τα στοιχεία μιας γενετικής ανάλυσης μεγάλης κλίμακας; Ουσιαστικά πρόκειται για μια σειρά από δεδομένα σχετικών συχνοτήτων αλληλομόρφων μεταξύ των διαφορετικών κατηγοριών του δείγματος. Έτσι σε μια μελέτη GWAS ασθενών-μαρτύρων που διερευνά ένα εκατομμύριο SNP τα τελικά αποτελέσματα θα μπορούν να αναπαρασταθούν ως ένας πίνακας 2×10^6 τιμών συχνοτήτων για κάθε πολυμορφική, διαλληλική θέση. Η ανάλυση των δεδομένων επικεντρώνεται σε δύο επίπεδα:

1. Στην εκτίμηση της διεισδυτικότητας (penetrance) κάθε πολυμορφισμού, του βαθμού δηλαδή στον οποίο ο πολυμορφισμός είναι επιδραστικός για τον παρατηρούμενο φαινότυπο.
2. Στην εκτίμηση της στατιστικής σημασίας των παρατηρούμενων διαφορών μεταξύ των δύο κατηγοριών του δείγματος.

Μαθηματικό Ιντερμέδιο II. Διεισδυτικότητα και λόγοι πιθανοτήτων

Θα εξετάσουμε αρχικά την έννοια της διεισδυτικότητας και τον τρόπο υπολογισμού της. Στη βιοϊατρική, ο γενικότερος ορισμός της διεισδυτικότητας ενός αλληλομόρφου για δεδομένο φαινότυπο, δίνεται από το ποσοστό των ατόμων που φέρουν το συγκεκριμένο αλληλόμορφο και εμφανίζουν το φαινότυπο αυτόν. Έτσι αν 95 από τους 100 ασθενείς φέρουν το αλληλόμορφο X, λέμε ότι το αλληλόμορφο X έχει 95% διεισδυτικότητα για την ασθένεια. Μια τέτοια θεώρηση της διεισδυτικότητας είναι προβληματική στην περίπτωση που οι σχετικές συχνότητες των αλληλομόρφων είναι μεγάλες και η εξάπλωσή τους στον πληθυσμό μεγαλύτερη. Παρόλο που το ποσοστό, όπως ορίστηκε παραπάνω, περιγράφει ικανοποιητικά περιπτώσεις επιδραστικών αλληλομόρφων σε μονογονιδιακές νόσους, δε συμβαίνει το ίδιο όταν μιλάμε για πολυπαραγοντικά σύνδρομα που συνδέονται με λιγότερο σπάνια αλληλόμορφα. Στην Εικόνα 10.7 περιγράφεται

σχηματικά σ' ένα διάγραμμα σκέδασης η διεισδυτικότητα σε σχέση με τη συχνότητα εμφάνισης συγκεκριμένων αλληλομόρφων που σχετίζονται με παθολογικές καταστάσεις. Η διεισδυτικότητα εδώ περιγράφεται καλύτερα με την τιμή λόγου πιθανοτήτων (odds ratio) την οποία συναντήσαμε και στο Κεφάλαιο 3 και η οποία στο πλαίσιο των μελετών γενετικής συσχέτισης ορίζεται ως εξής:

Για δεδομένη πολυμορφική θέση, έστω N_h ο αριθμός των υγιών ατόμων που έχουν στη θέση αυτή το αλληλόμορφο A και n_h ο αριθμός αυτών που φέρουν στην ίδια θέση το αλληλόμορφο a . Με αντίστοιχο τρόπο, συμβολίζουμε με N_d και n_d τους αριθμούς των ασθενών που φέρουν τα ίδια αλληλόμορφα. Τα δεδομένα μπορούν να παρασταθούν σε έναν πίνακα σύμπτωσης (contingency table) όπως ο Πίνακας 10.2:

Αλληλόμορφο	A	a	Σύνολο
Ασθενείς	N_d	n_d	d
Μάρτυρες	N_h	n_h	h
Σύνολο	N	n	T

Πίνακας 10.2: Πίνακας σύμπτωσης συχνοτήτων αλληλομόρφων σε διαλληλική θέση⁶

Ο λόγος πιθανοτήτων αλληλομόρφων για την ασθένεια που μελετάμε στον Πίνακα 10.2 θα δίνεται από τη σύγκριση των επιμέρους πιθανοτήτων να είναι κανείς ασθενής και να φέρει ένα συγκεκριμένο αλληλόμορφο και να είναι κανείς υγιής φέροντας το ίδιο αλληλόμορφο. Για το αλληλόμορφο A οι λόγοι πιθανοτήτων είναι:

$$p(Ah) = \frac{N_h}{n_h} \quad \text{και} \quad p(Ad) = \frac{N_d}{n_d} \quad 10.11$$

από τη σχέση 10.11 προκύπτει ότι ο λόγος των πιθανοτήτων να φέρει κάποιος ασθενής το αλληλόμορφο A σε σχέση με το να το φέρει κάποιος υγιής είναι:

$$OR(A) = \frac{\frac{N_h}{n_h}}{\frac{N_d}{n_d}} = \frac{N_h n_d}{N_d n_h} \quad 10.12$$

Από την τελική μορφή της εξίσωσης 10.12 προκύπτει ότι ο λόγος πιθανοτήτων είναι ανεξάρτητος των επιμέρους αριθμών ασθενών και υγιών ατόμων στο δείγμα. Μόνο οι σχετικές συχνότητες των αλληλομόρφων είναι απαραίτητες για τον υπολογισμό του $OR(A)$. Στην περίπτωση που δεν υπάρχει κάποια ιδιαίτερη συσχέτιση μεταξύ του γενετικού χαρακτηριστικού που μελετάμε (εδώ το αλληλόμορφο A) και του φαινοτύπου (εδώ υγιής-ασθενής) η αναμενόμενη τιμή OR είναι ίση με τη μονάδα. Αποκλίσεις από τη μονάδα υποδηλώνουν, σε διαφορετικό βαθμό, τη σχέση του μελετώμενου αλληλομόρφου με τον παρατηρούμενο φαινότυπο. Για να καταλάβουμε καλύτερα τον

⁶ Στο σημείο αυτό έχουμε απλουστεύσει την κατάσταση που μελετούμε σε σχέση με τον Πίνακα 10.1. Θεωρούμε έτσι ότι εξετάζουμε τους αριθμούς ομοζυγώτων (AA και aa για τα αλληλόμορφα A και a αντίστοιχα) γι' αυτό και οι στήλες γονοτύπων του Πίνακα 10.2 είναι δύο και όχι τρεις. Η θεωρία δεν αλλάζει ακόμα και στην περίπτωση που έχουμε τρεις ή περισσότερους γονοτύπους.

τρόπο υπολογισμού φανταστείτε ότι μελετώντας μια παθολογική κατάσταση έχετε εντοπίσει ένα SNP για τον οποίο 45/100 ασθενείς φέρουν το αλληλόμορφο A ενώ μεταξύ των υγιών μαρτύρων η αντίστοιχη τιμή είναι 900/1000. Οι τιμές του Πίνακα 10.2 είναι τότε οι εξής: $N_h=900$, $n_h=1000-900=10$, $N_d=45$, $n_d=100-45=55$. Ο λόγος OR(A) υπολογίζεται έτσι από την εξίσωση 10.12:

$$OR(A) = \frac{900 \times 55}{45 \times 100} = 11$$

Μια τιμή $>>1$ όπως αυτή που προκύπτει στο παράδειγμά μας είναι δηλωτική μιας πολύ ισχυρής τάσης συσχέτισης μεταξύ της παθολογικής κατάστασης και του αλληλομόρφου A στη συγκεκριμένη θέση. Στο παράδειγμά μας θα μπορούσαμε έτσι να πούμε πως το αλληλόμορφο A έχει μεγάλη διεισδυτικότητα στον πληθυσμό των ασθενών.

Πώς όμως θα αξιολογήσουμε την τιμή ενός OR από στατιστική άποψη; Μπορούμε να χρησιμοποιήσουμε μια προσέγγιση με διαστήματα εμπιστοσύνης, τα οποία σε αυτήν την περίπτωση ορίζονται σε λογαριθμική κλίμακα ως εξής:

$$(lower\ limit, upper\ limit) = \log(OR) \pm 1.96(se(OR)) \quad 10.13$$

όπου ισχύει ότι το τυπικό σφάλμα (se) του OR ισούται με:

$$se(OR) = \sqrt{\frac{1}{N_h} + \frac{1}{n_d} + \frac{1}{N_d} + \frac{1}{n_h}} \quad 10.14$$

Σημειώνεται πως οι εξισώσεις 10.13 και 10.14 είναι σε λογαριθμική κλίμακα που σημαίνει ότι για να εκτιμήσει κανείς τις τιμές-όρια του διαστήματος εμπιστοσύνης θα πρέπει να μετατρέψει τους λογαρίθμους στην αρχική κλίμακα.

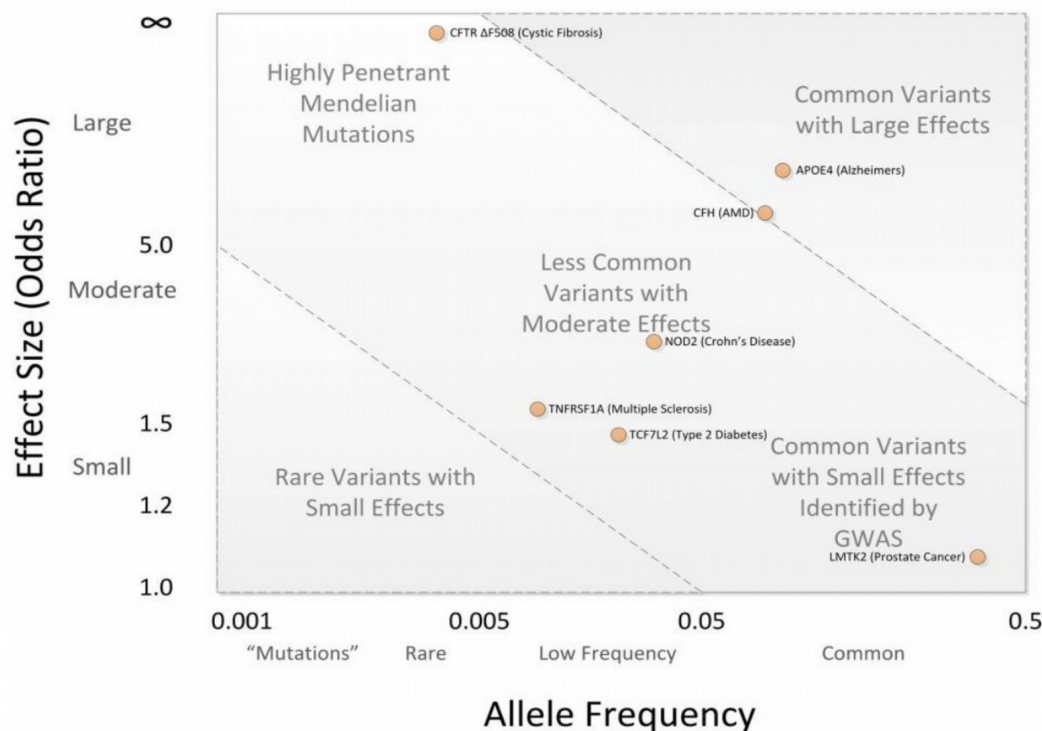
Ερώτηση: Στην Εικόνα 10.6 μπορείτε να υπολογίσετε το λογάριθμο του λόγου πιθανοτήτων (*log-odds ratio*) για το συγκεκριμένο πίνακα και να ελέγξετε κατά πόσο η ετεροζυγωτία στο συγκεκριμένο αλληλόμορφο συνδέεται με την εκδήλωση της ασθένειας; Ποιο είναι το διάστημα εμπιστοσύνης σε αυτήν την περίπτωση;

Συσχέτιση σπανιότητας αλληλομόρφων και ασθενειών

Επιστρέφουμε τώρα στην Εικόνα 10.7, όπου περιγράφεται σχηματικά η υπόθεση της συσχέτισης σπανιότητας αλληλομόρφων και ασθενειών (*rare disease – rare variant hypothesis*) (Schork, Murray, Frazer & Topol, 2009). Σύμφωνα με τη θεωρία της σπανιότητας αλληλομόρφων, υπάρχει συνολικά μια αντίστροφη σχέση μεταξύ της διεισδυτικότητας ενός αλληλομόρφου και της συχνότητας του στον πληθυσμό (Bush & Moore, 2012). Κάτι τέτοιο είναι αναμενόμενο καθώς όπως συζητήσαμε και στην αρχή του κεφαλαίου, ένα αλληλόμορφο του οποίου η επίδραση είναι πολύ μεγάλη θα πρέπει εκ των πραγμάτων να είναι σπάνιο λόγω του εξελικτικού μειονεκτήματος που έχουν οι φορείς του να το μεταδώσουν. Τα πολύ σπάνια, διεισδυτικά αλληλόμορφα είναι έτσι χαρακτηριστικά μεντελιανών μονογονιδιακών παθήσεων όπως η κυστική ίνωση.

Λιγότερο σπάνια αλληλόμορφα, με ενδιαμέση επίδραση στο φαινότυπο βρίσκονται στο μέσο της Εικόνας 10.7. Συσχετίσεις σε αυτό το επίπεδο μπορούν να εξαχθούν σχετικά εύκολα μέσω

GWAS, όμως ακόμα πιο σημαντικές είναι οι περιπτώσεις που αντιστοιχούν στο κάτω δεξί μέρος του διαγράμματος. Αυτές αντιστοιχούν σε αλληλόμορφα με μεγάλη συχνότητα που όμως έχουν μικρή φαινοτυπική επίδραση (μικρή διεισδυτικότητα), τα οποία είναι χαρακτηριστικά πολυπαραγοντικών νόσων όπου η γενετική προδιάθεση δεν μπορεί να αποδοθεί σε ένα μοναδικό γενετικό τόπο, αλλά “διαχέονται” σε μεγαλύτερο αριθμό γενετικών χαρακτηριστικών (Stranger, Stahl & Raj, 2011). Προσεγγίσεις GWAS με μεγάλο μέγεθος δειγμάτων μπορούν να προσδιορίσουν αυτές τις περιπτώσεις ακόμα και με πολύ μικρό βαθμό διεισδυτικότητας.



Εικόνα 10.7: Διάγραμμα της σχέσης μεταξύ διεισδυτικότητας (ως odds-ratio) και της συχνότητας αλληλομόρφων στον πληθυσμό. Για πολύ σπάνια αλληλόμορφα (συχνότητα <0.05) με μικρή διεισδυτικότητα (κάτω αριστερά μέρος του διαγράμματος) απαιτούνται πολύ μεγάλα δείγματα του πληθυσμού, ενώ το ίδιο συμβαίνει και στην περίπτωση των πολύ κοινών (συχνότητα >0.1) αλληλομόρφων ακόμα κι αν η διεισδυτικότητά τους είναι μεγάλη (πάνω δεξιά μέρος του διαγράμματος). (Η Εικόνα είναι προσαρμοσμένη από την εργασία των Bush & Moore, 2012, CC BY 3.0)

Σε γενικές γραμμές, μέσω GWAS προσεγγίσεων μπορούμε σήμερα να εντοπίσουμε γονοτυπικές-φαινοτυπικές συσχετίσεις για συχνότερες αλληλομόρφων που είναι μεγαλύτερες από 0.05 ακόμα κι αν ο βαθμός διεισδυτικότητας είναι ελάχιστα >1, ενώ αντίστροφα, ακόμα και πολύ σπάνια αλληλόμορφα μπορούν να εντοπιστούν εφόσον η διεισδυτικότητά τους είναι μεγάλη. Τα επίπεδα αποτελεσματικότητας των GWAS ορίζονται από τις διακεκομμένες διαγώνιες γραμμές της Εικόνας 10.7. Η προσπάθεια που καταβάλλεται με την ολοένα και μεγαλύτερη αύξηση της διακριτικής ικανότητας και (κυρίως) του μεγέθους των δειγμάτων είναι να διευρυνθεί όσο το δυνατόν περισσότερο η περιοχή μεταξύ των δύο αυτών γραμμών.

Ανάλυση αποτελεσμάτων GWAS

Μέχρι στιγμής είδαμε πώς υπολογίζουμε τη διεισδυτικότητα ενός αλληλομόρφου σ' έναν πληθυσμό. Οι λόγοι πιθανοτήτων για τη διεισδυτικότητα είναι το ανάλογο των λόγων έκφρασης για τα γονίδια που συναντήσαμε στο Κεφάλαιο 7. Η τιμή τους μπορεί να είναι μεγάλη ή μικρή αλλά αυτό μας λέει λίγα για το πόσο σημαντική είναι από στατιστική άποψη. Μια τιμή odds-ratio μπορεί να είναι αρκετά μεγαλύτερη από τη μονάδα ($OR \gg 1$) αλλά λόγω του μικρού μεγέθους του δείγματος να μην είναι σημαντική, ενώ αντίθετα μια τιμή $1 < OR < 2$ να είναι πολύ σημαντική. Πώς θα αξιολογήσουμε στατιστικά τις τιμές των odds-ratio; Στη συνέχεια θα δούμε ότι η εκτίμηση αυτή γίνεται μ' ένα στατιστικό έλεγχο που συνδυάζει στοιχεία από την προσέγγιση χ^2 που συναντήσαμε παραπάνω και την υπεργεωμετρική κατανομή που είδαμε στο Κεφάλαιο 8.

Μαθηματικό Ιντερμέδιο III. Ο έλεγχος Fisher σε πίνακες σύμπτωσης

Ας επιστρέψουμε για λίγο στο αρχικό μας ερώτημα που είναι να εντοπίσουμε τη γενετική βάση μιας παθολογικής κατάστασης. Από την ανάλυση των γονοτύπων ασθενών και υγιών μαρτύρων σε διάφορες πολυμορφικές θέσεις έχουν προκύψει τα παρακάτω αποτελέσματα για τρεις από αυτές:

Υποθέστε εδώ ότι λόγω της διαφορετικής πηγής από την οποία προέρχονται τα δεδομένα σας, οι αριθμοί των εξεταζόμενων ατόμων είναι διαφορετικοί (κάτι τέτοιο δε συμβαίνει όταν κανείς έχει σχεδιάσει μια μελέτη για ένα συγκεκριμένο σκοπό, αλλά είναι αρκετά συχνό όταν αναλύουμε δημοσιευμένα δεδομένα). Για τους τρεις διαφορετικούς γενετικούς τόπους (Θ_1 , Θ_2 , Θ_3) οι τιμές odds-ratio που προκύπτουν για το μείζον αλληλόμορφο είναι αντίστοιχα:

$$OR(A)=8.69 \quad OR(B)=8.01 \quad OR(\Gamma)=1.22$$

Θέση	Θ_1		Θ_2		Θ_3	
Αλληλόμορφο	A	a	B	B	Γ	γ
Ασθενείς	50	50	5	5	500	500
Μάρτυρες	90	10	9	1	550	450

Πίνακας 10.3: Πίνακας σύμπτωσης συχνοτήτων αλληλομόρφων για τρεις διαλληλικές θέσεις με δεδομένα από μια μελέτη γενετικής συσχέτισης.

Εξετάζοντας μόνο τις τιμές αυτές τείνουμε να πιστέψουμε ότι τα αλληλόμορφα A και B είναι αρκετά σημαντικά για τη μελέτη μας λόγω της μεγάλης διεισδυτικότητάς τους. Από την άλλη το Γ εμφανίζει μια τιμή που είναι ελάχιστα μεγαλύτερη από τη μονάδα και που εκ πρώτης όψεως δε μας επιτρέπει να το χαρακτηρίσουμε σημαντικό. Είναι όμως πράγματι έτσι; Παρά τις μεγάλες τιμές odds-ratio, τα αποτελέσματά μας για το A και κυρίως για το B προέρχονται από παρατηρήσεις σε πολύ

μικρότερο αριθμό ατόμων, κάτι που θα πρέπει να λάβουμε υπόψη μας. Πώς όμως θα εκτιμήσουμε στατιστικά τον πίνακα σύμπτωσης για κάθε γενετικό τόπο; Ο στατιστικός έλεγχος για τέτοιες περιπτώσεις είναι αυτός που προτάθηκε από τον Ronald Fischer (Upton, 1992), έναν από τους πρώτους επιστήμονες που εργάστηκε τόσο ως στατιστικός και μαθηματικός όσο και ως εξελικτικός βιολογός. Σύμφωνα με τον έλεγχο του Fischer σ' έναν πίνακα σύμπτωσης όπως ο παρακάτω:

Κατηγορία	Αριθμοί Ατόμων		Σύνολο
	K1	K2	
Ομάδα 1	a	β	$a+\beta$
Ομάδα 2	γ	δ	$\gamma+\delta$
Σύνολο	$a+\gamma$	$\beta+\delta$	$a+\beta+\gamma+\delta$

Πίνακας 10.4: Παράδειγμα Πίνακα Σύμπτωσης για τον υπολογισμό του ελέγχου Fischer.

η πιθανότητα παρατήρησης των τιμών a , β , γ και δ στους αντίστοιχους συνδυασμούς Ομάδας-Κατηγορίας δίνεται από την υπεργεωμετρική κατανομή και ισούται με:

$$p(\text{table}) = \frac{\binom{a+\beta}{a} \binom{\gamma+\delta}{\gamma}}{\binom{a+\beta+\gamma+\delta}{a+\gamma}} \quad 10.15$$

όπου (όπως θα θυμηθείτε από το Κεφάλαιο 8) $\binom{x}{y}$ είναι ο διωνυμικός τελεστής για τα x, y και αντιστοιχεί στο πλήθος των δυνατών συνδυασμών y αντικειμένων από ένα σύνολο x δυνατών

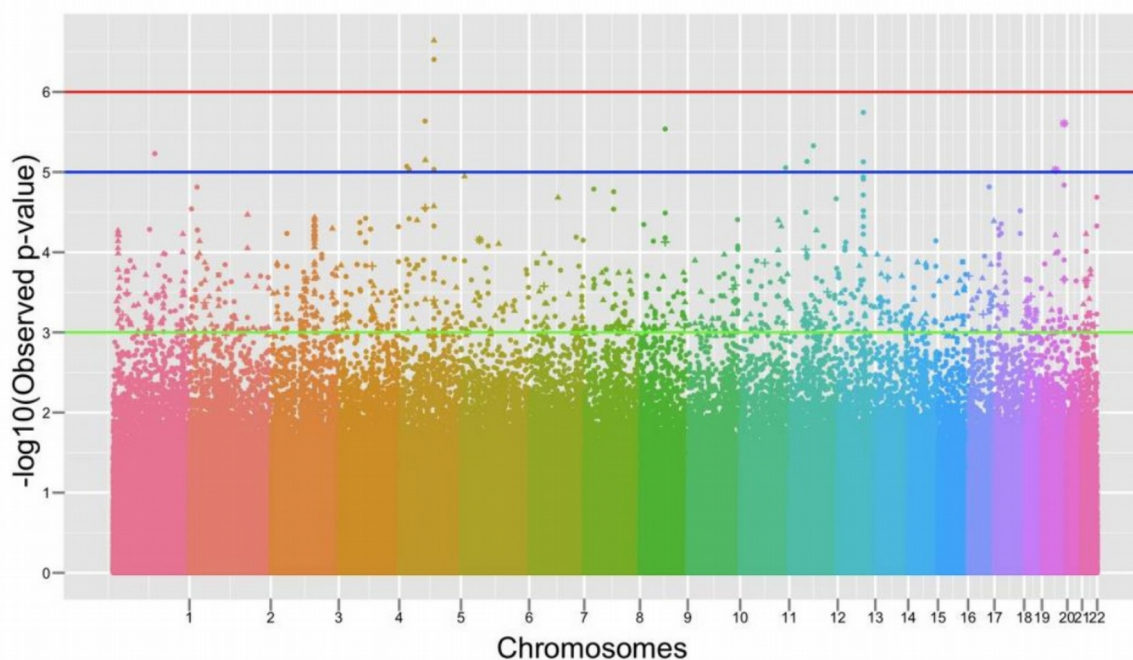
επιλογών και που ισούται με $\frac{x!}{(x-y)!y!}$.

Η σχέση 10.13 μας δίνει την ακριβή πιθανότητα παρατήρησης του συγκεκριμένου πίνακα, καθώς ο έλεγχος Fischer είναι ακριβής και όχι προσεγγιστικός, η πιθανότητα δηλαδή που προκύπτει είναι αποτέλεσμα στάθμισης ενός πεπερασμένου αριθμού ενδεχομένων που εξάγονται μέσω συνδυαστικής ανάλυσης. Η εξίσωση 10.13 μας δίνει έτσι την πιθανότητα να παρατηρήσουμε τον πίνακα 10.3 με βάση τη μηδενική υπόθεση ότι οι ομάδες 1 και 2 δεν έχουν καμία προτίμηση για τις κατηγορίες K1 και K2, κατανέμονται δηλαδή σ' αυτές ανεξάρτητα. Η τιμή p έχει έτσι εδώ χαρακτηριστικά p -value και μας επιτρέπει να εξάγουμε απευθείας συμπεράσματα για τη στατιστική σημασία των παρατηρήσεών μας. Τι μπορεί να πας πει ο έλεγχος Fischer για το παράδειγμά μας του Πίνακα 10.2; Αν εφαρμόσουμε την εξίσωση 10.13 σε καθέναν από τους γενετικούς τόπους $\Theta 1$, $\Theta 2$ και $\Theta 3$ προκύπτει ότι:

$$p(\Theta 1)=0.004 \quad p(\Theta 2)=0.14 \quad p(\Theta 3)=0.03$$

Η τιμή $p(\Theta 1)$ είναι ενδεικτική μιας ισχυρής προτίμησης στην περίπτωση του αλληλόμορφου A. Για τη θέση $\Theta 2$, ωστόσο, παρά τη μεγάλη (>8) τιμή odds-ratio, η παρατήρηση της συγκεκριμένης κατανομής συχνοτήτων του αλληλομόρφου B είναι μάλλον αναμενόμενη, καθώς η τιμή $p(\Theta 2)$ είναι

μεγαλύτερη από το συνηθισμένο κατώφλι στατιστικής σημαντικότητας ($p < 0.05$). Η θέση Θ3 εμφανίζει τιμή $p(\Theta3)$ κάτω από το τυπικό αυτό κατώφλι και ως εκ τούτου θα μπορούσαμε να τη χαρακτηρίσουμε σημαντική ακόμα και αν η τιμή odds-ratio είναι σχετικά μικρή για το αλληλόμορφο Γ. Συνολικά θα μπορούσαμε να πούμε ότι για τη μεν θέση Θ1 έχουμε ένα αλληλόμορφο (το Α) με μεγάλη διεισδυτικότητα που είναι στατιστικά σημαντικό, για τη δε θέση Θ3 ότι πρόκειται για μια θέση με μεγάλη στατιστική σημασία αλλά μ' ένα αλληλόμορφο (Γ) με μικρή διεισδυτικότητα. Το παράδειγμα της θέσης Θ3 είναι χαρακτηριστικό των προβλημάτων που προκύπτουν από αλληλόμορφα με μεγάλη συχνότητα (εννοείται για το ελάχιστον αλληλόμορφο). Παρόλο που η κατανομή τους στον πληθυσμό μπορεί να είναι ενδεικτική μιας φαινοτυπικής σχέσης, η μεγάλη τους συχνότητα απαιτεί πολύ μεγάλο αριθμό ατόμων στο δείγμα για την εξαγωγή μιας σημαντικής γενετικής συσχέτισης. Ο συνδυασμός ενός αριθμού τέτοιων αλληλομόρφων ενδέχεται να αποτελεί τη γενετική βάση για πολλές πολυ-παραγοντικές, πολύπλοκες ασθένειες, όπως είδαμε σχηματικά στην Εικόνα 10.7. Στην τελευταία ενότητα αυτού του κεφαλαίου θα συζητήσουμε τρόπους με τους οποίους μπορούμε να αυξήσουμε τη στατιστική ισχύ μελετών GWAS για τον προσδιορισμό τέτοιων ασθενών πολυμορφισμών.



Εικόνα 10.8: Διάγραμμα Manhattan από μια μελέτη συσχέτισης σε γονιδιωματική κλίμακα με στόχο τον εντοπισμό νέων γονιδίων επιδεκτικότητας για την εκδήλωση της ασθένειας του Alzheimer. Κάθε σημείο αντιστοιχεί σε έναν από τους συνολικά 516645 πολυμορφισμούς που μελετήθηκαν σε δείγμα 381 ασθενών. Οι οριζόντιες γραμμές αντιστοιχούν σε διαφορετικά κατώφλια τιμών p -values (10^{-3} , 10^{-5} και 10^{-6}). (Η Εικόνα προέρχεται από τη μελέτη των Potkin et al., 2009, CC BY 3.0)

Τα πειράματα γενετικής συσχέτισης σε γονιδιωματική κλίμακα δεν περιορίζονται προφανώς σε τρεις θέσεις αλλά απόσκοπούν στην αξιολόγηση εκατοντάδων χιλιάδων ή εκατομμυρίων θέσεων στο γονιδίωμα. Έτσι η παραπάνω ανάλυση επαναλαμβάνεται για μεγάλο αριθμό πινάκων σύμπτωσης κάτι που κάνει απολύτως απαραίτητη τη διόρθωση για έλεγχο πολλαπλών υποθέσεων

(βλ. Κεφάλαιο 8). Το τελικό αποτέλεσμα μιας μελέτης GWAS είναι έτσι ένας πίνακας από διορθωμένες τιμές p -value καθεμία από τις οποίες αντιστοιχεί σε ένα συγκεκριμένο SNP. Η αναπαράσταση αυτών των δεδομένων γίνεται με τη μορφή διαγραμμάτων όπως αυτό που φαίνεται στην Εικόνα 10.8 και ονομάζεται διάγραμμα Manhattan (Manhattan plot) (από το σχήμα, που σε κάποιους θυμίζει το περίγραμμα των ουρανοξυστών του Manhattan). Ένα διάγραμμα Manhattan είναι ουσιαστικά ένα διάγραμμα σκέδασης, στο οποίο ο οριζόντιος άξονας αντιστοιχεί στις γονιδιωματικές συντεταγμένες και ο κάθετος στον αρνητικό δεκαδικό λογάριθμο των διορθωμένων τιμών p -value ($-\log_{10}(p)$). Σημεία που εμφανίζονται ψηλά στο διάγραμμα αντιστοιχούν έτσι σε SNP με μικρή τιμή p -value που είναι κατά συνέπεια σημαντικά για τη μελετώμενη φαινοτυπική ιδιότητα στο δείγμα μας. Τα διαγράμματα Manhattan συνδυάζονται συχνά με θερμικούς χάρτες LD που υποδηλώνουν τον απλότυπο που προκύπτει από το μελετούμενο πληθυσμό.

De novo ανάλυση γενετικής ποικιλομορφίας. Αλληλούχιση συνόλου εξονικών τμημάτων (Whole-Exome Sequencing)

Οι περισσότερες μελέτες συσχέτισης σε γονιδιωματική κλίμακα χρησιμοποιούν τεχνικές μικροσυστοιχειών για τη διερεύνηση των αλληλομόρφων σε γνωστές πολυμορφικές γονιδιωματικές θέσεις. Ανάλογα με τη συστοιχεία και την τεχνολογία που χρησιμοποιείται, η διακριτική ικανότητα και ο αριθμός των εξεταζόμενων θέσεων μπορεί να διαφέρει. Έτσι για παράδειγμα, συστοιχείες που αποσκοπούν να μελετήσουν ευρωπαϊκούς πληθυσμούς περιέχουν συνήθως μικρότερο αριθμό πολυμορφικών θέσεων σε σύγκριση με αντίστοιχα πειράματα για τη μελέτη αφρικανικών πληθυσμών, καθώς οι τελευταίοι έχουν υποστεί μεγαλύτερο αριθμό γεγονότων ανασυνδυασμού (οι αφρικάνικοι πληθυσμοί είναι παλαιότεροι) και ως εκ τούτου οι μικρότερες τιμές ανισορροπίας σύνδεσης απαιτούν μεγαλύτερη διακριτική ικανότητα. Σε κάθε περίπτωση πάντως η χρήση μικροσυστοιχειών περιορίζει τη μελέτη σ' ένα σύνολο γνωστών πολυμορφισμών.

Τι συμβαίνει όμως όταν ένα φαινοτυπικό χαρακτηριστικό σχετίζεται με μια πολυμορφική θέση που δεν είναι ακόμα γνωστή; Όπως συζητήσαμε στην αρχή αυτού του κεφαλαίου, το σύνολο των γνωστών SNP για ένα δεδομένο γονιδίωμα είναι συνάρτηση του βάθους στο οποίο έχουν μελετηθεί οι πληθυσμοί του είδους και της σπανιότητας των αλληλομόρφων. Ο αριθμός των 1.5 εκατομμυρίων SNP για το ανθρώπινο γονιδίωμα κατά πάσα πιθανότητα είναι αρκετά μικρότερος από το συνολικό αριθμό των πολυμορφικών θέσεων που περιέχει το γονιδίωμά μας. Από την άλλη πλευρά, *de novo* μεταλλαγές που συμβαίνουν διαρκώς ενδέχεται να σχετίζονται άμεσα με φαινοτυπικές μεταβολές χωρίς να συμπεριλαμβάνονται σε γνωστά σύνολα SNP (Francioli et al., 2015). Για τη μελέτη της κρυμμένης ποικιλομορφίας αυτού του τύπου, είναι απαραίτητες προσεγγίσεις που ξεπερνούν τους περιορισμούς των γνωστών πολυμορφικών θέσεων.

Τα τελευταία χρόνια, η ανάπτυξη των νέων τεχνολογιών αλληλούχισης, (για τις οποίες θα μιλήσουμε πιο αναλυτικά στο επόμενο κεφάλαιο), προσφέρουν τη δυνατότητα για την απευθείας μελέτη της γενετικής ποικιλομορφίας, χωρίς τους περιορισμούς των γνωστών πολυμορφισμών. Οι

προσεγγίσεις αυτές περιλαμβάνουν αφενός την πλήρη επαναλληλούχιση ενός γονιδιώματος (whole-genome resequencing, WGRS) (Huang et al., 2009) και αφετέρου τη στοχευμένη αλληλούχιση του συνόλου των εξονικών τμημάτων του (whole-exome sequencing, WES) (Choi et al., 2009). Και στις δύο περιπτώσεις το αποτέλεσμα είναι η καθαρή αλληλουχία (ή τμήματα αυτής) ενός συγκεκριμένου ατόμου η οποία μπορεί να συγκριθεί απευθείας τόσο με το γονιδίωμα αναφοράς για τον υπό μελέτη πληθυσμό όσο και με την αντίστοιχη ομόλογη αλληλουχία άλλων ατόμων του δείγματος. Η προσέγγιση αυτή επιτρέπει συγκρίσεις ακόμα και σε περιοχές όπου δεν υπάρχουν γνωστοί πολυμορφισμοί, ενώ επιτρέπει ταυτόχρονα και την αναγνώριση άλλων τύπων ποικιλομορφίας όπως οι CNV.

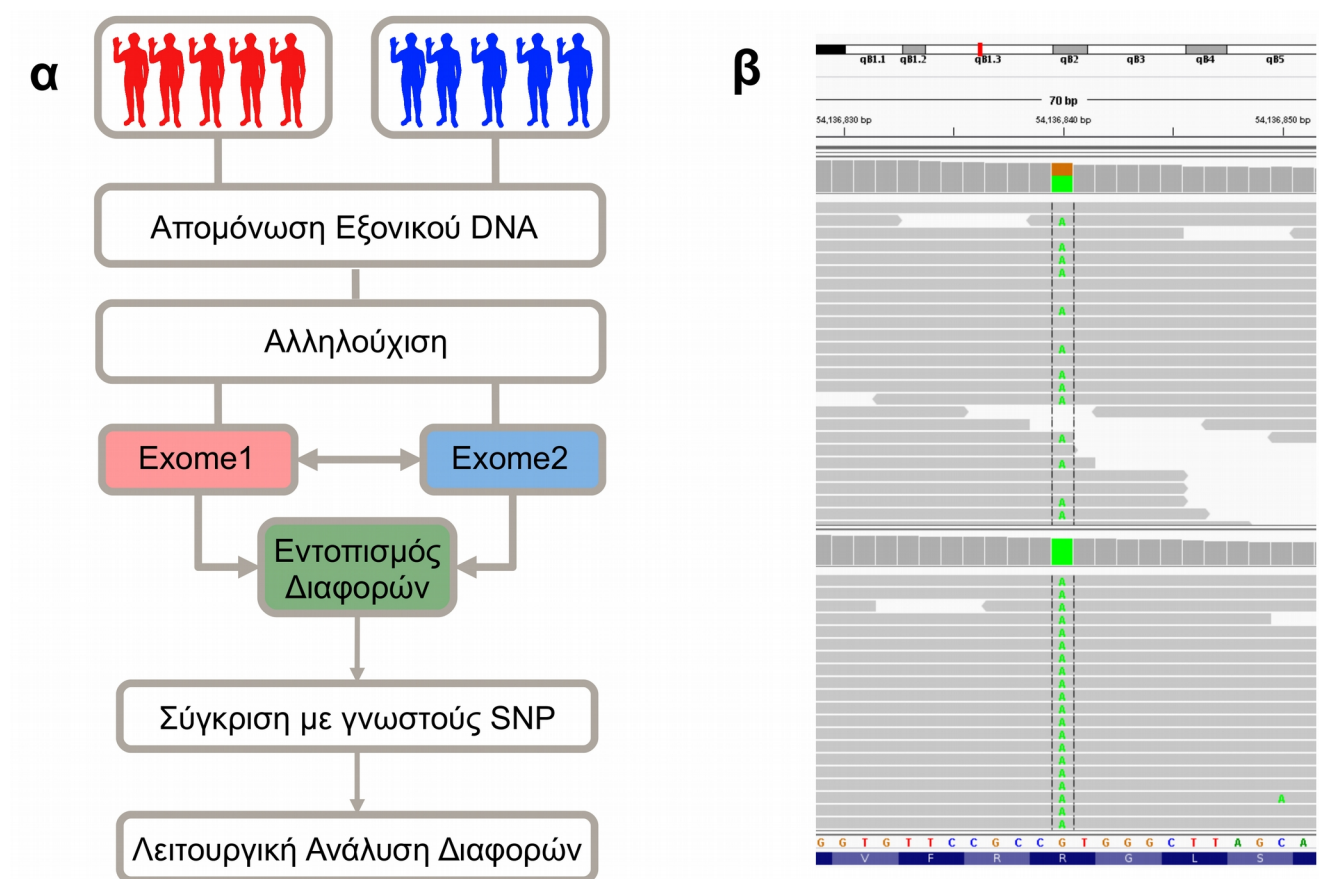
Η πλήρης (επαν)αλληλούχιση ενός ανθρώπινου γονιδιώματος είναι πλέον ένα πείραμα που δεν απαιτεί το κόστος και την προσπάθεια που χρειάστηκε για το πρόγραμμα της αρχικής αλληλούχισης του ανθρώπινου γονιδιώματος. Η διαδικασία έχει γίνει σημαντικά φθηνότερη με τις νέες τεχνολογίες, ενώ η ύπαρξη ενός γονιδιωματικού χάρτη αναφοράς επιτρέπει τη σχεδόν αυτόματη συναρμολόγηση (assembly) των αλληλουχημένων τμημάτων σε πλήρες γονιδίωμα. Παρ' όλ' αυτά εξακολουθεί να είναι απαγορευτική από πλευράς κόστους σε επίπεδο κλινικής ρουτίνας κι έτσι επιλέγεται μόνο σε ειδικές περιπτώσεις όπου κλινικοί λόγοι (κίνδυνος καθυστερημένης διάγνωσης) συνυπάρχουν με τις διαγνωστικές δυνατότητες (γρήγορη συναρμολόγηση, ταυτόχρονη αλληλούχιση γονέων και παιδιού σε περιπτώσεις μεντελιανών νοσημάτων) (Bull et al., 2013). Η προσέγγιση συνόλου εξονικών τμημάτων ή "πλήρους εξονιώματος" (Whole-Exome Sequencing, στο εξής WES), που περιγράφεται σχηματικά στην Εικόνα 10.9 αποτελεί εδώ και μερικά χρόνια την κατά προτίμηση επιλογή για τη διερεύνηση κρυμμένης ποικιλομορφίας σε ανθρώπινους πληθυσμούς. Η πρώτη εφαρμογή του WES (Ng et al., 2009) οδήγησε στον εντοπισμό περισσότερων από 300 νέων σπάνιων αλληλομόρφων στο ανθρώπινο γονιδίωμα, ενώ η εφαρμογή του σε μεγαλύτερη κλίμακα οδήγησε στο εντυπωσιακό συμπέρασμα ότι κάθε νέο άτομο φέρει κατά μέσο όρο μια *de novo* μεταλλαγή στο εξονικό τμήμα του γονιδιώματός του (O'Roak et al., 2011). Η αναλυτική ισχύς του WES, σε συνδυασμό με το σχετικά μικρό κόστος, έχει οδηγήσει στην ολοένα και διευρυνόμενη εφαρμογή του στην κλινική διάγνωση με εντυπωσιακά αποτελέσματα (Yang et al., 2013).

Μέσω του WES το γενετικό υλικό εμπλουτίζεται εξαρχής σε περιοχές που αντιστοιχούν στο εξονικό τμήμα του ανθρώπινου γονιδιώματος και στη συνέχεια ακολουθεί "βαθιά" αλληλούχιση με τεχνολογίες αλληλούχισης νέας γενιάς (βλ. Κεφάλαιο 11). Το αποτέλεσμα είναι ένας μεγάλος αριθμός τμημάτων αλληλουχίας που μπορεί να συγκριθεί απευθείας με το γονιδίωμα αναφοράς. Η διαδικασία (μέρη της οποίας θα περιγραφούν πιο αναλυτικά στο επόμενο κεφάλαιο) διαρθρώνεται στα εξής στάδια:

- **Χαρτογράφηση τμημάτων αλληλουχίας στο γονιδίωμα αναφοράς.** Το στάδιο αυτό περιλαμβάνει ουσιαστικά τη στοίχιση των αλληλουχιών που προκύπτουν από το δείγμα έναντι του γνωστού ανθρώπινου γονιδιώματος. Λεπτομέρειες για τα προβλήματα της διαδικασίας αυτής και πώς λύνονται θα συζητηθούν στο επόμενο κεφάλαιο. Στο σημείο αυτό αρκεί να σημειώσουμε ότι η αντιστοίχιση γίνεται επιτρέποντας την ύπαρξη διαφορών μεταξύ των αλληλουχιών δείγματος και αναφοράς καθώς είναι ακριβώς αυτά τα σημεία όπου δυνητικά εντοπίζεται η αναζητούμενη κρυμμένη ποικιλομορφία.
- **Εντοπισμός μονονουκλεοτιδικών αντικαταστάσεων.** Αποτελεί και το πιο δύσκολο

υπολογιστικά στάδιο. Σε αυτό θα πρέπει να διακριθούν οι πραγματικές μεταβολές στην αλληλουχία, που αντιστοιχούν σε φυσική ποικιλομορφία με τα τυχαία σφάλματα κατά την ανάγνωσή της, που αποτελούν τεχνικό πρόβλημα. Οι τεχνικές αλληλούχισης νέας γενιάς είναι ιδιαίτερα χρήσιμες σ' αυτό το σημείο καθώς το αποτέλεσμά τους είναι ένας πολύ μεγάλος αριθμός (10^6 - 10^7) από σχετικά μικρές σε μήκος αλληλουχίες (50-250bps). Λόγω της εστίασης του δείγματος σε εξονικές περιοχές, κάθε νουκλεοτίδιο του γονιδιώματος αναφοράς επικαλύπτεται από ένα σημαντικό αριθμό τέτοιων αλληλουχιών κι έτσι κάθε βάση έχει διαβαστεί περισσότερες από μία φορές. Ανάλογα με τον όγκο και την ένταση της αλληλούχισης του δείγματος, ένας καλός κανόνας είναι να εξασφαλίζεται κάλυψη 30X που αντιστοιχεί σε συνολικό μήκος αλληλουχιών που ισούται με 30 φορές το συνολικό μήκος των εξονίων. 30X κάλυψη (coverage) σημαίνει ότι κατά μέσο όρο κάθε νουκλεοτίδιο θα έχει αλληλουχηθεί 30 φορές. Με 30 (κατά μέσο όρο) αναγνώσεις για κάθε νουκλεοτίδιο, έχουμε έτσι τη δυνατότητα να διενεργήσουμε τη στατιστική ανάλυση που θα μας επιτρέψει να διακρίνουμε όχι μόνο αν υπάρχει ποικιλομορφία ή όχι αλλά και αν η νουκλεοτιδική αλλαγή αντιστοιχεί σε ομοζυγώτη (~100% διαφορά) ή σε ετεροζυγώτη (~50% διαφορά). Στην Εικόνα 10.9β φαίνονται χαρακτηριστικές περιπτώσεις διαφορών μεταξύ δείγματος και γονιδιώματος αναφοράς που οφείλονται σε αντικατάσταση σε ομοζυγωτία ή σε ετεροζυγωτία. Σε κάθε αντικατάσταση μπορεί να αντιστοιχηθεί και μία τιμή p-value η οποία προκύπτει από το στατιστικό έλεγχο Fisher που συζητήσαμε παραπάνω.

- **Σύγκριση με γνωστούς πολυμορφισμούς.** Από τη στιγμή που έχουν εντοπιστεί όλες οι σημαντικές νουκλεοτιδικές αντικαταστάσεις, αντιπαραβάλλονται αρχικά με γνωστούς πολυμορφισμούς. Η σύγκριση αυτή είναι απαραίτητη στην περίπτωση που αναζητούμε *de novo* ή άγνωστες μεταλλαγές στο δείγμα. Εφόσον αυτές προκύψουν προχωράμε στο επόμενο και τελευταίο στάδιο του σχολιασμού τους.
- **Σχολιασμός (annotation) αντικαταστάσεων.** *De novo* νουκλεοτιδικές αντικαταστάσεις που εντοπίζονται σε τμήματα εξονικού DNA είναι πολύ πιθανό να μην οδηγούν σε καμία φαινοτυπική αλλαγή λόγω του εκφυλισμού του γενετικού κώδικα. Μέσω της διαδικασίας του σχολιασμού, το σύνολο των εντοπιζόμενων αντικαταστάσεων διακρίνεται αρχικά σε κωδικές και μη-κωδικές, καθώς το WES συχνά αλληλουχεί και γονιδιωματικά τμήματα που αντιστοιχούν σε όρια εξονίων-εσωνίων. Οι κωδικές αντικαταστάσεις διακρίνονται στη συνέχεια σε συνώνυμες, μη-συνώνυμες, συντηρητικές κλπ, με βάση την μεταβολή που επιφέρουν στο κωδικοποιούμενο αμινοξύ. Ιδιαίτερα σημαντικές είναι η μη-νοηματικές αντικαταστάσεις που δημιουργούν ένα πρόωρο κωδικόνιο λήξης. Οι μη-κωδικές αλληλουχίες από την άλλη πλευρά, διακρίνονται σ' αυτές που επηρεάζουν σημεία συρραφής εξονίων (splice sites) αλλά και σημεία με λιγότερο προφανείς επιπτώσεις όπως θέσεις πρόσδεσης μεταγραφικών παραγόντων κλπ (Cingolani et al., 2012).



Εικόνα 10.9: α) Διάγραμμα ροής της ανάλυσης γενετικής ποικιλομορφίας σε δύο δείγματα μέσω αλληλούχισης του συνόλου εξονικών τμημάτων (Whole Exome Sequencing). β) Παραδείγματα στοίχισεων των προϊόντων μιας WES ανάλυσης σε μια περιοχή όπου το άτομο είναι ετεροζυγώτης (πάνω) και ομοζυγώτης (κάτω) για ένα διαφορετικό αλληλόμορφο σε σχέση με το γονιδίωμα αναφοράς.

Μετα-ανάλυση μελετών γενετικής συσχέτισης

Παρά την ολοένα και διευρυνόμενη χρήση προσεγγίσεων GWAS για τον προσδιορισμό αλληλομόρφων που σχετίζονται με παθολογικές καταστάσεις, τις περισσότερες φορές, ακόμα και ο συνδυασμός των αλληλομόρφων αυτών δεν εξηγεί παρά σε μικρό ποσοστό τη γενετική προδιάθεση για πολύπλοκες ασθένειες. Η γενετική επίδραση σε πολύπλοκα νοσήματα εξασκείται πιθανότατα μέσω του συνδυασμού ενός αριθμού γενετικών χαρακτήρων, πολλοί από τους οποίους είναι δύσκολο να προσδιοριστούν λόγω της μεγάλης συχνότητας των αλληλομόρφων, που κάνει τη στατιστική τους ανάλυση αδύνατη ακόμα και σε δείγματα εκατοντάδων ή χιλιάδων ατόμων. Τα τελευταία χρόνια, μια πολύ δημοφιλής προσέγγιση για την ανακάλυψη νέων γενετικών τόπων που σχετίζονται με κοινές αλλά πολύπλοκες παθολογικές καταστάσεις είναι η μετα-ανάλυση δεδομένων γενετικής ποικιλομορφίας (Munafo & Flint, 2004). Με τον όρο μετα-ανάλυση αναφερόμαστε γενικά σε στατιστικές μεθόδους που αποσκοπούν στη βελτίωση των αποτελεσμάτων μέσω του

συνδυασμού ενός αριθμού αναλύσεων που έχουν τον ίδιο σκοπό. Στο επίπεδο των GWAS, η παράλληλη σύνθεση των αποτελεσμάτων διαφορετικών, ανεξάρτητων μελετών μπορεί να βελτιώσει τη στατιστική σημασία κάποιων αλληλομόρφων, ενώ ταυτόχρονα να απομακρύνει το θόρυβο των εσφαλμένως θετικών αποτελεσμάτων (Evangelou & Ioannidis, 2013).

Σε αντίθεση με τα προφανή τους πλεονεκτήματα, που τις έχουν κάνει ιδιαίτερα δημοφιλείς ακόμα και μεταξύ μη-ειδικών⁷, οι μετα-αναλύσεις σε GWAS μελέτες είναι ιδιαίτερα πολύπλοκες από πλευράς οργάνωσης και διεξαγωγής και θα πρέπει να αντιμετωπίζονται με μεγάλη προσοχή. Πολύ συχνά, μετα-αναλύσεις διεξάγονται από consortia που είναι όσο μεγάλα όσο τα αντίστοιχα που διεξήγαγαν τις αρχικές GWAS μελέτες. Ο σχεδιασμός της μετα-ανάλυσης είναι έτσι βασικός και θα πρέπει να περιλαμβάνει:

1. **Τον αποσαφηνισμένο χαρακτηρισμό του περιγραφόμενου φαινοτύπου.** Όπως συζητήσαμε και στην ενότητα για τις GWAS μελέτες, τα κριτήρια προσδιορισμού του φαινοτύπου είναι πρωταρχικής σημασίας για τη μελέτη της συσχέτισής του με γενετικούς χαρακτήρες. Η σημασία αυτή είναι μεγαλύτερη ακόμα στην περίπτωση των μετα-αναλύσεων, όπου υπάρχει η επιπλέον ετερογένεια μεταξύ διαφορετικών μελετών.
2. **Τη μέθοδο με την οποία θα αναλυθούν τα δεδομένα.** Το στάδιο αυτό περιλαμβάνει την επιλογή μεταξύ διαφορετικών μεθοδολογιών για τη διεξαγωγή της ανάλυσης.
3. **Τη διαδικασία προτεραιοποίησης (prioritization) των γενετικών χαρακτήρων που θα προκύψουν.** Σκοπός κάθε μετα-ανάλυσης GWAS δεδομένων είναι ο περιορισμός των γενετικών χαρακτήρων που εμφανίζουν στατιστικά σημαντική συσχέτιση με το φαινότυπο σ' ένα μικρό σύνολο γονιδίων, με σκοπό αυτά να ελεγχθούν περαιτέρω πειραματικά ως πιθανοί διαγνωστικοί δείκτες ή φαρμακευτικοί στόχοι. Η διαδικασία της προτεραιοποίησης γονιδίων αποτελεί ένα πολύ ενδιαφέρον παράδειγμα μεθόδου μηχανικής μάθησης που συνδυάζει προϋπάρχουσα γνώση και πειραματικά αποτελέσματα από γονιδιωματικές μελέτες και θα δούμε περισσότερα γι' αυτήν στο Κεφάλαιο 12. Στο σημείο αυτό αρκεί να αναφέρουμε τρόπους για μη-αυτοματοποιημένη προτεραιοποίηση όπως είναι ο έλεγχος της εγγύτητας μεταξύ σημαντικών αλληλομόρφων, οι αλληλεπιδράσεις σε πρωτεϊνικό επίπεδο (βλ. Κεφάλαια 8, 9) κλπ.

Τα διάφορα στατιστικά μοντέλα για τη σύνθεση δεδομένων σε μετα-αναλύσεις είναι, όπως αναφέραμε και παραπάνω, αρκετά πολύπλοκα και η αναλυτική περιγραφή τους ξεφεύγει από τους σκοπούς ενός κεφαλαίου αφιερωμένου στη γενετική ποικιλομορφία. Ενδεικτικά μπορούμε να αναφέρουμε τη γενικότερη διάκρισή τους σε μοντέλα που βασίζονται στη στατιστική διόρθωση τιμών *p*-value, σε μοντέλα στάθμισης των επιμέρους επιδράσεων μελετών που μπορεί να είναι είτε σταθερά (fixed effects model) είτε τυχαία (random effects model), σε μπεϋζιανά (πιθανοτικά) μοντέλα και σε πολυπαραγοντικά (multivariate) (Bagos 2008, Thompson, Attia & Minelli, 2011). Μεταξύ τους, οι μέθοδοι στάθμισης είναι γενικώς πιο αποδεκτές λόγω της αυξημένης τους σταθερότητας.

⁷ Τον Οκτώβριο του 2014, ο Guillaume Filion από το Κέντρο Γονιδιακής Ρύθμισης της Βαρκελώνης, ανακάλυψε μια σειρά από 32 χαλκευμένες επιστημονικές εργασίες που είχαν δημοσιευθεί σε περιοδικά με κριτές μέσα σ' ένα διάστημα λίγων μηνών. Οι εργασίες ήταν αποτέλεσμα ενός “συγγραφέα-φάντασμα” και είχαν όλες την ίδια δομή, εκτεταμένα τμήματα του κειμένου ήταν πανομοιότυπα, όπως και οι τίτλοι τους που ήταν όλοι βασισμένοι στο κοινό πρότυπο “X και Y: μια μετα-ανάλυση”. Ο Filion περιγράφει την ιστορία της ανακάλυψής του στο blog του <http://blog.thegrandlocus.com/2014/10/a-flurry-of-copycats-on-pubmed>

Συμπεράσματα

Σημαντικό μέρος των ερωτημάτων που πραγματεύεται αυτό το κεφάλαιο απαιτούν το συνδυασμό γνώσεων από διαφορετικά πεδία που περιλαμβάνουν τη Γενετική Πληθυσμών, Στατιστική ανάλυση και προχωρημένες τεχνικές γνώσεις ανάλυσης δεδομένων από μεθοδολογίες αλληλούχισης νέας γενιάς. Σκοπός του κεφαλαίου δεν είναι προφανώς να παρουσιάσει σε λεπτομέρεια όλα τα παραπάνω, αλλά να δώσει το γενικότερο πλαίσιο μέσα στο οποίο μελετάται η γενετική ποικιλομορφία, την εποχή των “μεγάλων δεδομένων”.

Τα τελευταία χρόνια, η αύξηση του όγκου αυτών των δεδομένων, αφενός στο επίπεδο της γενετικής ποικιλομορφίας και αφετέρου στην ανάλυση σε γονιδιωματική κλίμακα μιας πληθώρας γονιδιωματικών χαρακτηριστικών (όπως θα δούμε στο αμέσως επόμενο κεφάλαιο), έχει σε μεγάλο βαθμό διευκολύνει τη διερεύνηση και την ερμηνεία της γενετικής βάσης πολλών φαινοτυπικών ιδιοτήτων. Έτσι, πέρα από την αναμφισβήτητη επίδραση μιας μεταλλαγής στην έκφραση μιας πρωτεΐνης, που αντιστοιχεί σε φαινοτυπικό αποτέλεσμα δυαδικού (on-off) τύπου, μπορούμε πλέον να εντοπίσουμε γενετική ποικιλομορφία στο επίπεδο του βαθμού έκφρασης γονιδίων (ποσοτική απόκριση), στον τρόπο ρύθμισής του (μεταλλαγές σε θέσεις πρόσδεσης μεταγραφικών παραγόντων) ή ακόμα και σε στοιχεία που έχουν να κάνουν με τη δομή της χρωματίνης.

Μελέτες στο επίπεδο διαφοροποιήσεων στην έκφρασης γονιδίων οδηγούν στην καταγραφή γενετικών τόπων ποσοτικών χαρακτηριστικών (quantitative trait loci, QTL) που σχετίζονται με την έκφραση (expression QTL, ή για συντομία eQTL). Χαρακτηριστικά αυτού του τύπου φαίνεται πως είναι ιδιαίτερα σημαντικά στη μελέτη πολύπλοκων ασθενειών, ενώ έχουν συσχετιστεί έντονα με συγκεκριμένα είδη καρκίνων (Q. Li et al., 2013). Ξεπερνώντας τα όρια των γονιδίων, η αντιπαραβολή των θέσεων στις οποίες εδράζεται η γενετική ποικιλομορφία, με γονιδιωματικά χαρακτηριστικά που σχετίζονται με τις ρυθμιστικές ή δομικές ιδιότητες, αναδεικνύει κρυμμένες, μέχρι πρότινος, πτυχές στην πολυπλοκότητα της γενετικής πληροφορίας. Έτσι, πολυμορφικές θέσεις χωρίς κάποια προφανή γονιδιακή σύνδεση, μπορούν να χαρακτηριστούν λειτουργικές μέσω του συνεντοπισμού τους με ρυθμιστικά στοιχεία του γονιδιώματος όπως θέσεις πρόσδεσης μεταγραφικών παραγόντων, περιοχές διαφορικής μεθυλίωσης του DNA κ.ά. (Ward & Kellis, 2012). Σ' ένα δομικό επίπεδο, ο συνδυασμός θέσεων γενετικής ποικιλομορφίας μπορεί επίσης να συσχετιστεί με την υποκείμενη δομή της χρωματίνης μεταξύ διαφορετικών ατόμων (Vernot & Stergachis, 2012).

Στο επόμενο δυο, τελευταία κεφάλαια αυτού του βιβλίου θα δούμε πιο αναλυτικά τον τρόπο με τον οποίο παράγεται το πλήθος, η ποικιλία και ο όγκος των νέων γονιδιωματικών δεδομένων, καθώς και πώς αυτά μπορούν να αναλυθούν, παρέχοντας τη βάση για τέτοιες συστημικές προσεγγίσεις,

Ερωτήσεις/Ασκήσεις

1. Χρησιμοποιώντας τον Πίνακα 10.1 να αποδείξετε την αρχή ισορροπίας Hardy-Weinberg, να δείξετε δηλαδή ότι οι τιμές των σχέσεων 10.1 παραμένουν ίδιες μετά από μια γενιά (θυμηθείτε ότι $p=1-q$).
2. Μια γενετική ασθένεια έχει συχνότητα εμφάνισης στον πληθυσμό $1/3600$. Αν εκδηλώνεται σε ομοζυγώτες του υπολειπόμενου γονιδίου για ένα συγκεκριμένο γενετικό τόπο, μπορείτε να βρείτε τη συχνότητα του ελάσσονος αλληλομόρφου στον πληθυσμό; Ποια θα ήταν η τιμή της συχνότητας του ίδιου αλληλομόρφου αν η ασθένεια εκδηλωνόταν μόνο στους ετεροζυγώτες;
3. Πώς μπορεί ένας SNP να οδηγεί σε απαλοιφή ενός αμινοξικού καταλοίπου όπως στη ΔF508 μεταλλαγή του CFTR γονιδίου;
4. Να διενεργήσετε τον έλεγχο Fisher για την πολυμορφική θέση της Εικόνας 10.6

Διαβάστε Περισσότερα

Για τη γενετική ποικιλομορφία γενικότερα:

Μια εισαγωγή από βασικά συγγράμματα γενετικής είναι επαρκής. Ενδεικτικά αναφέρουμε το *iGenetics: A Mendelian Approach* (Russell, 2006). Ένα από τα καλύτερα βιβλία για τη Γενετική Πληθυσμών (χωρίς να περιορίζεται σε αυτήν) είναι το *The Origins of Genome Architecture* (Lynch, 2007).

Για τους πολυμορφισμούς και τη δομική ποικιλομορφία:

Στο Κεφάλαιο 9 του *Molecular Evolution* (W.-H. Li, 1997) θα βρείτε μια συνοπτική εισαγωγή για τα είδη της γενετικής ποικιλομορφίας.

Για την Αρχή Hardy-Weinberg:

Η αρχική εργασία του Hardy είναι αρκετά τεχνική (και γραμμένη το 1908) ενώ αυτή του Weinberg είναι στα γερμανικά. Για τους παραπάνω λόγους προτιμήστε την περιγραφή του Κεφαλαίου 2 του *Principles of Population Genetics* (Hartl, 1980).

Για τη γενετική σύνδεση:

Μια πολύ καλή εισαγωγή τόσο του θεωρητικού υποβάθρου όσο και των εφαρμογών δίνεται στο *Analysis of Human Genetic Linkage* (Ott, 1999).

Για την ανισορροπία σύνδεσης και τους απλοτύπους:

Στο Κεφάλαιο 13 του *Computational Genome Analysis: An Introduction* (Deonier, Tavaré & Waterman, 2007) μπορείτε να καταφύγετε για την έννοια της LD και του συντελεστή γενετικής συσχέτισης r^2 . Περισσότερα για τους απλοτύπους θα βρείτε στην ιστοσελίδα του *HapMap Project* (<http://hapmap.ncbi.nlm.nih.gov/>)

Για τις GWAS:

Ένα πολύ καλό εισαγωγικό άρθρο είναι η εργασία των (Bush & Moore, 2012).

Για το Whole-Exome Sequencing WES:

Υπάρχει πληθώρα εργασιών δημοσιευμένων τα τελευταία χρόνια. Μεταξύ τους, ξεχωρίζει η επισκόπηση των (Yang et al., 2013) όπου η μεθοδολογία παρουσιάζεται από τη σκοπιά της κλινικής εφαρμογής. Για τον αναγνώστη που επιθυμεί να εμβαθύνει στο τεχνικό μέρος της ανάλυσης τέτοιων πειραματικών προσεγγίσεων μια καλή αρχή είναι το *Next-Generation DNA Sequencing Informatics* (Brown, 2015).

Για τη μετα-ανάλυση:

Ένα καλό άρθρο επισκόπησης είναι το (Evangelou & Ioannidis, 2013), ωστόσο πρόκειται, όπως αναφέραμε και στο κείμενο για ένα θέμα, εξαιρετικά απαιτητικό σε τεχνικό επίπεδο που απαιτεί ένα πολύ καλό στατιστικό υπόβαθρο.

Βιβλιογραφία

- Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. a, Durbin, R. M., Handsaker, R. E., ... McVean, G. a. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), 56–65. <http://doi.org/10.1038/nature11632>
- Bagos, P.G. (2008). A unification of multivariate methods for meta-analysis of genetic association studies. *Stat Appl Genet Mol Biol*. 7 (1), Article 31. doi: 10.2202/1544-6115.1408
- Bateson, W., Waunders, E. R. & Punnett, R. C. (1909). Experimental studies in the physiology of heredity. *Zeitschrift Für Induktive Abstammungs- Und Vererbungslehre*, 2(1), 17–19. <http://doi.org/10.1007/BF01975751>
- Brown, S. M. (2015). *Next-generation DNA Sequencing Informatics*. Cold Spring Harbor Laboratory Press.
- Bull, K. R., Rimmer, A. J., Siggs, O. M., Miosge, L. a, Roots, C. M., Enders, A., ... Cornall, R. J. (2013). Unlocking the bottleneck in forward genetics using whole-genome sequencing and identity by descent to isolate causative mutations. *PLoS Genetics*, 9(1), e1003219. <http://doi.org/10.1371/journal.pgen.1003219>
- Bush, W. S. & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 8(12), e1002822. <http://doi.org/10.1371/journal.pcbi.1002822>
- Cheng, S. H., Gregory, R. J., Marshall, J., Paul, S., Souza, D. W., White, G. A., ... Smith, A. E. (1990). Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis. *Cell*, 63(4), 827–834. [http://doi.org/10.1016/0092-8674\(90\)90148-8](http://doi.org/10.1016/0092-8674(90)90148-8)
- Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., ... Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(45), 19096–101. <http://doi.org/10.1073/pnas.0910672106>
- Chowdhury, R., Bois, P. R. J., Feingold, E., Sherman, S. L. & Cheung, V. G. (2009). Genetic analysis of variation in human meiotic recombination. *PLoS Genetics*, 5(9), e1000648. <http://doi.org/10.1371/journal.pgen.1000648>
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... Lu, X. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, 6(June), 80–92.
- Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., ... Donnelly, P. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289), 713–20. <http://doi.org/10.1038/nature08979>

- Deonier, R. C., Tavaré, S. & Waterman, M. (2007). *Computational Genome Analysis: An Introduction*. Springer Science & Business Media.
- Durbin, R. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), 1061–73. <http://doi.org/10.1038/nature09534>
- Evangelou, E. & Ioannidis, J. P. a. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews. Genetics*, 14(6), 379–89. <http://doi.org/10.1038/nrg3472>
- Francioli, L. C., Polak, P. P., Koren, A., Menelaou, A., Chun, S., Renkens, I., ... Sunyaev, S. R. (2015). Genome-wide patterns and properties of de novo mutations in humans. *Nature Genetics*, (October 2014). <http://doi.org/10.1038/ng.3292>
- Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G. & Chee, M. S. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genetics*, 37(5), 549–54. <http://doi.org/10.1038/ng1547>
- Hardy, G. H. (1908). Mendelian Proportions in a mixed population. *Science (New York, N.Y.)*, 28(706), 49–50. <http://doi.org/10.1126/science.28.706.49>
- Hartl, D. L. (1980). *Principles of population genetics*. Sinauer Associates.
- Hill, W. G. & Robertson, A. (1968). Linkage disequilibrium in finite populations. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 38(6), 226–31. <http://doi.org/10.1007/BF01245622>
- Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., ... Han, B. (2009). High-throughput genotyping by whole-genome resequencing. *Genome Research*, 19(6), 1068–1076. <http://doi.org/10.1101/gr.089516.108>
- Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., ... Tsui, L. C. (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science (New York, N.Y.)*, 245(4922), 1073–80.
- Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., ... Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)*, 308(5720), 385–9. <http://doi.org/10.1126/science.1109557>
- Krawitz, P., Rödelberger, C., Jäger, M., Jostins, L., Bauer, S. & Robinson, P. N. (2010). Microindel detection in short-read sequence data. *Bioinformatics (Oxford, England)*, 26(6), 722–9. <http://doi.org/10.1093/bioinformatics/btq027>
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, 27(21), 2987–93. <http://doi.org/10.1093/bioinformatics/btr509>
- Li, Q., Seo, J. H., Stranger, B., McKenna, A., Pe’Er, I., Laframboise, T., ... Freedman, M. L. (2013). Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell*, 152(3), 633–

641. <http://doi.org/10.1016/j.cell.2012.12.034>
- Li, W.-H. (1997). *Molecular Evolution*. Sinauer Associates.
- Lynch, M. (2007). *The Origins of Genome Architecture*. Sinauer Associates.
- Marenberg, M. E., Risch, N., Berkman, L. F., Floderus, B. & de Faire, U. (1994). Genetic susceptibility to death from coronary heart disease in a study of twins. *The New England Journal of Medicine*, 330(15), 1041–6. <http://doi.org/10.1056/NEJM199404143301503>
- Morgan, T. H. (1911). Random Segregation versus coupling in Mendelian Inheritance. *Science (New York, N.Y.)*, 34(873), 384. <http://doi.org/10.1126/science.34.873.384>
- Morral, N., Bertranpetit, J., Estivill, X., Nunes, V., Casals, T., Giménez, J., ... Kadasí, L. (1994). The origin of the major cystic fibrosis mutation ($\Delta F508$) in European populations. *Nature Genetics*, 7(2), 169–175. <http://doi.org/10.1038/ng0694-169>
- Munafò, M. R. & Flint, J. (2004). Meta-analysis of genetic association studies. *Trends in Genetics*, 20(9), 439–444. <http://doi.org/10.1016/j.tig.2004.06.014>
- Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., ... Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261), 272–6. <http://doi.org/10.1038/nature08250>
- O’Roak, B. J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J. J., Girirajan, S., ... Eichler, E. E. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, 43(6), 585–9. <http://doi.org/10.1038/ng.835>
- Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., ... Plenge, R. M. (2013). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*. <http://doi.org/10.1038/nature12873>
- Ott, J. (1999). *Analysis of Human Genetic Linkage*. JHU Press.
- Petronis, A. (2006). Epigenetics and twins: three variations on the theme. *Trends in Genetics: TIG*, 22(7), 347–50. <http://doi.org/10.1016/j.tig.2006.04.010>
- Potkin, S. G., Guffanti, G., Lakatos, A., Turner, J. A., Kruggel, F., Fallon, J. H., ... Macciardi, F. (2009). Hippocampal Atrophy as a Quantitative Trait in a Genome-Wide Association Study Identifying Novel Susceptibility Genes for Alzheimer’s Disease. *PLoS ONE*, 4(8), e6501. <http://doi.org/10.1371/journal.pone.0006501>
- Russell, P. J. (2006). *IGenetics: A Mendelian Approach*. Pearson/Benjamin Cummings.
- Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C. C. A., Patsopoulos, N. A., Moutsianas, L., ... Compston, A. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359), 214–9. <http://doi.org/10.1038/nature10251>
- Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. (2009). Common vs. rare allele hypotheses for

- complex diseases. *Current Opinion in Genetics & Development*, 19(3), 212–9. <http://doi.org/10.1016/j.gde.2009.04.010>
- Stankiewicz, P. & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61, 437–55. <http://doi.org/10.1146/annurev-med-100708-204735>
- Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., ... Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science (New York, N.Y.)*, 315(5813), 848–53. <http://doi.org/10.1126/science.1136678>
- Stranger, B. E., Stahl, E. a & Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2), 367–83. <http://doi.org/10.1534/genetics.110.120907>
- Thanseem, I., Nakamura, K., Anitha, A., Suda, S., Yamada, K., Iwayama, Y., ... Mori, N. (2011). Association of transcription factor gene LMX1B with autism. *PloS One*, 6(8), e23738. <http://doi.org/10.1371/journal.pone.0023738>
- The International HapMap Consortium. (2003). The International HapMap Project. *Nature*, 426(6968), 789–796.
- Thompson, J. R., Attia, J. & Minelli, C. (2011). The meta-analysis of genome-wide association studies. *Briefings in Bioinformatics*, 12(3), 259–69. <http://doi.org/10.1093/bib/bbr020>
- Upton, G. J. G. (1992). Fisher's Exact Test. *Journal of the Royal Statistical Society*, 155(3), 395–402. <http://doi.org/10.1002/0470011815.b2a10020>
- Vernot, B. & Stergachis, A. (2012). Personal and population genomics of human regulatory variation. *Genome Research*, 1689–1697. <http://doi.org/10.1101/gr.134890.111>.
- Ward, L. D. & Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*, 40(Database issue), D930–4. <http://doi.org/10.1093/nar/gkr917>
- Weinberg, W. (1908). Über Vererbungsgesetze beim Menschen. *Zeitschrift Für Induktive Abstammungs- Und Vererbungslehre*, 1(1), 440–460. <http://doi.org/10.1007/BF01990626>
- Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. a., ... Eng, C. M. (2013). Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *New England Journal of Medicine*, 369(16), 1502–1511. <http://doi.org/10.1056/NEJMoa1306555>