

Gene Set Analysis (GSA)

Παντελής Μπάγκος

2025

Πανεπιστήμιο Θεσσαλίας

Εισαγωγή

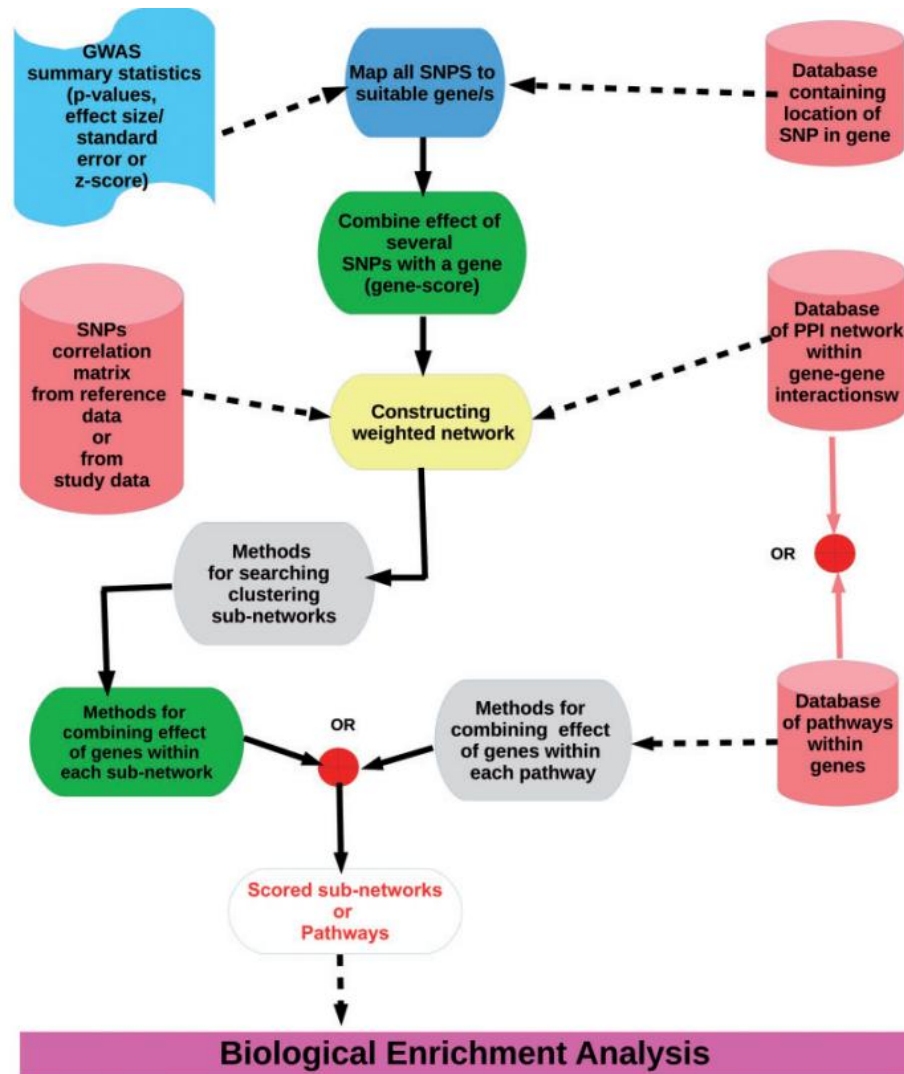
- Η ανάλυση συνόλων γονιδίων - Gene set analysis (GSA), ή η Ανάλυση Μονοπατιών, αναλύει από κοινού ομάδες λειτουργικά σχετιζόμενων γονιδίων και εντοπίζοντας βιολογικές οδούς εμπλουτισμένες με γονίδια που σχετίζονται με τα χαρακτηριστικά.
- Λαμβάνοντας υπόψη τη συλλογική επίδραση πολλαπλών γονιδίων σε μια οδό, οι ερευνητές μπορούν να αποκτήσουν μια σαφέστερη εικόνα των υποκείμενων βιολογικών μηχανισμών που επηρεάζουν τον υπό διερεύνηση φαινότυπο.
- Η ανάλυση μονοπατιών είναι μια αποτελεσματική τεχνική που ξεπερνά τους περιορισμούς των τρεχουσών μεθόδων που αναλύουν ένα γονίδιο κάθε φορά. Αυτή η διαδικασία παρέχει μια ολοκληρωμένη κατανόηση των μοριακών μηχανισμών που προκαλούν πολύπλοκες ασθένειες
- Οι πρώτες εφαρμογές τέτοιων μεθόδων δανείστηκαν ιδέες από τη βιβλιογραφία ανάλυσης δεδομένων μικροσυστοιχιών και από τότε έγιναν ευρέως διαδεδομένες στην ανάλυση των GWAS.

- Επειδή αυτές οι μέθοδοι ελέγχουν το σωρευτικό αποτέλεσμα πολλαπλών γονίδιων είναι δυνατό να ανιχνευθούν αποτελέσματα στο επίπεδο γονιδιακών συνόλων, τα οποία μπορεί να είναι το αποτέλεσμα ετερογενών επιδράσεων στο επίπεδο γονιδίου ή πολυμορφισμού.
- Επιπλέον, οι αναλύσεις ομάδων γονιδίων βελτιώνουν τη στατιστική ισχύ ανίχνευσης σημαντικών συσχετίσεων τόσο επειδή η συσσώρευση μεμονωμένων SNPs σε γονιδιακά σύνολα έχει ως αποτέλεσμα λιγότερες στατιστικές δοκιμές που εκτελούνται αλλά και επειδή μεμονωμένα ασθενή αποτελέσματα, μη ανιχνεύσιμα σε μια τυπική GWAS, μπορούν να συνδυαστούν δίνοντας ένα ισχυρό σήμα

- Σε γενικές γραμμές οι μέθοδοι εφαρμόζονται τόσο σε SNPs όσο και σε Microarrays αν και υπάρχουν κάποιες επί μέρους διαφορές

- Όλες οι μέθοδοι GSA πρέπει να αντιμετωπίσουν ορισμένα ζητήματα που εντάσσονται σε 3 κατηγορίες:
 - Πρώτον, πώς να χειριστούμε τα SNP του ίδιου γονιδίου.
 - Δεύτερον, πώς να ορίσουμε το κατάλληλο σύνολο γονιδίων ή μονοπάτι, και τέλος
 - Με ποιον τρόπο να συνδυάσουμε τα αποτελέσματα από πολλαπλά SNPs/γονίδια μέσα στο ίδιο σύνολο/μονοπάτι.

- Έτσι, οι επιλογές που γίνονται με διαφορετικές μεθόδους μπορεί να είναι πολύ διαφορετικές, οδηγώντας σε μια μεγάλη ποικιλία διαφορετικών προσεγγίσεων.
 - Για παράδειγμα, ορισμένες μέθοδοι λειτουργούν με στατιστικά σε επίπεδο SNP (μεγέθη επιδράσεων, z ή p -τιμές) που εκχωρούν το SNP στο πλησιέστερο γονίδιο (συνήθως εντός εύρους $\pm 20K$ βάσεων), ενώ άλλες λαμβάνουν ως είσοδο μια στατιστική σε επίπεδο γονιδίου ή απλώς μια λίστα γονιδίων που λαμβάνεται με μια μέθοδο βασισμένη σε γονίδια (φυσικά, τόσο για μια προσέγγιση γονιδίου όσο και για ένα εργαλείο SA).
- Όσον αφορά την επιλογή του συνόλου, υπάρχει μια πληθώρα βάσεων δεδομένων που περιέχουν βιολογικά μονοπάτια (KEGG, PANTHER κλπ), ή άλλους τύπους αναπαράστασης συνόλου γονιδίων όπως αλληλεπιδράσεις PPI, οντολογίες και ούτω καθεξής.
- Τέλος, όσον αφορά τη στατιστική μέθοδο που χρησιμοποιείται για τη συγκέντρωση στοιχείων, υπάρχει επίσης ένα ευρύ φάσμα διαφορετικών μεθόδων που χειρίζονται με διαφορετικές προσεγγίσεις το μέγεθος του συνόλου γονιδίων και το μήκος γονιδίου, τα πρότυπα LD και την παρουσία επικαλυπτόμενων γονιδίων μέσα σε μονοπάτια, ή εφαρμόζουν διαφορετικές στατιστικές προσεγγίσεις όπως αυτές που χρησιμοποιούν τη λεγόμενη ανταγωνιστική μηδενική υπόθεση ή αυτές που χρησιμοποιούν την αυτόνομη.



Βασικά συστατικά μιας ανάλυσης

- Βάση δεδομένων
- Μέθοδος
- Στατιστικός έλεγχος
- Λογισμικό

Βάσεις δεδομένων – Είδη πληροφορίας

- Μονοπάτια (πχ KEGG)
- Δίκτυα αλληλεπιδράσεων (πχ BIND, DIP, IntAct, MINT, HPRD, STRING)
- Οντολογίες (πχ GO)
- Βιοδείκτες

Mooney MA, Nigg JT, McWeeney SK, Wilmot B. Functional and genomic context in pathway analysis of GWAS data. Trends in Genetics. 2014 Sep 1;30(9):390-400.

Pers TH. Gene set analysis for interpreting genetic studies. Human molecular genetics. 2016 Aug 10;25(R2):R133-40.

Μονοπάτια

- Έχουν προταθεί τέσσερις γενικοί τύποι μονοπατιών για να επιχειρήσουν να περιγράψουν την ετερογένεια των διαθέσιμων σήμερα ορισμών μονοπατιών:
 - α) μοριακά μονοπάτια,
 - β) κυτταρικά μονοπάτια,
 - γ) μονοπάτια ασθενειών και
 - δ) μονοπάτια παρέμβασης.

- Με την προϋπόθεση ότι τα βιολογικά μονοπάτια κατευθύνονται από ένα συγκεκριμένο σημείο εκκίνησης σε ένα συγκεκριμένο αποτέλεσμα, τα μοριακά μονοπάτια χαρακτηρίζουν τις βιοχημικές δράσεις σε ένα μόριο ή ένωση (π.χ., μεταβολισμό φωλικού οξέος).
- Αντίθετα, τα κυτταρικά μονοπάτια μοντελοποιούν τη ρύθμιση κυτταρικών διεργασιών, όπως η κυτταρική διαίρεση ή η απόπτωση. Έχει προταθεί επίσης ότι μονοπάτια μπορούν να οριστούν σε υψηλότερα επίπεδα, όπως το επίπεδο οργάνων ή συστημάτων.
- Ένα παράδειγμα αυτού του τύπου υψηλού επιπέδου μονοπατιού είναι το μονοπάτι του κίρκαδιανού ρυθμού στην KEGG

Μονοπάτια

- Επιπλέον, τα μοντέλα μονοπατιών μπορούν να περιγράψουν διεργασίες ασθένειας ή αποκρίσεις σε επεμβάσεις (π.χ., απόκριση φαρμάκου).
- Ένα παράδειγμα αυτού του τύπου μονοπατιών είναι η οδός της νόσου του Alzheimer στην KEGG, η οποία περιέχει βασικά γονίδια που σχετίζονται με το Alzheimer όπως τα APOE, APP και PSEN1.
- Έχει παρατηρηθεί, ωστόσο, ότι οι ασθένειες και τα μονοπάτια μπορεί απλώς να είναι συλλογές γονιδίων που έχουν απλά βρεθεί να συνδέονται με ένα φαινότυπο, αντί να βασίζονται στη γνώση ακριβών βιολογικών μηχανισμών.

Προβληματισμοί

- Το πρώτο ζήτημα είναι ότι τα μονοπάτια αντιπροσωπεύουν μοντέλα υποθετικών βιολογικών διεργασιών. Δεν υπάρχουν τυποποιημένοι κανόνες για την κατασκευή μοντέλων μονοπατιών
- Το δεύτερο ζήτημα αφορά την ιεραρχική φύση των βιολογικών μονοπατιών, πράγμα που σημαίνει ότι μερικά μονοπάτια μπορούν να αντιπροσωπεύουν υποσύνολα μιας μεγαλύτερης οδού.
- Έχει αποδειχθεί ότι οι αλληλεπιδράσεις 'crosstalk' (η επιρροή ενός μονοπατιού στο άλλο) οφείλονται σε μεγάλο βαθμό στα κοινά γονίδια (αλληλεπικάλυψη) μεταξύ των μονοπατιών.

Δίκτυα αλληλεπιδράσεων

- Βιολογικά δίκτυα, όπως τα δίκτυα αλληλεπίδρασης πρωτεΐνης-πρωτεΐνης (PPI), μπορούν επίσης να χρησιμοποιηθούν για τον ορισμό των συνόλων γονιδίων.
- Σε αντίθεση με τα μονοπάτια, τα δίκτυα δεν περιγράφουν ρητά μια συγκεκριμένη βιολογική λειτουργία ή διαδικασία που πραγματοποιείται σε ένα συγκεκριμένο βιολογικό πλαίσιο.
- Αντ 'αυτού, τα δίκτυα αποσκοπούν απλώς στην περιγραφή των βιολογικών σχέσεων (παρατηρούμενες ή προβλεπόμενες αλληλεπιδράσεις) μεταξύ πολλαπλών γονιδίων ή γονιδιακών προϊόντων.

Δίκτυα αλληλεπιδράσεων

- Γενικά, τα είδη της πληροφορίας που χρησιμοποιείται για τη δημιουργία δημόσιων δικτύων PPI, όπως η BIND, DIP, IntAct, MINT, HPRD και STRING, είναι ετερογενή.
- Για παράδειγμα, οι ενδείξεις για τις αλληλεπιδράσεις πρωτεϊνών μπορούν να προέλθουν από πολλαπλούς τύπους πειραμάτων (π.χ. CoIP, ChiP-chip, γονιδιακή έκφραση, εξόρυξη κειμένου), τα οποία εκτελούνται σε διάφορους τύπους ιστών και υπό διάφορες συνθήκες.

- Επιπλέον, τα δεδομένα που χρησιμοποιούνται για να εξαχθεί μια αλληλεπίδραση μπορεί να είναι ποικίλης ποιότητας και δεν είναι πάντοτε διαθέσιμες πληροφορίες σχετικά με την αξιοπιστία αυτών των αλληλεπιδράσεων
- Είναι επίσης δυνατή η απλή επικάλυψη ήδη καθορισμένων συνόλων γονιδίων (όπως οι κατηγορίες σε μια οντολογία γονιδίων) σε ένα δίκτυο PPI για τη λήψη αλληλεπιδράσεων μεταξύ των γονιδίων

Οντολογίες

- Μια δημοφιλής πηγή για ορισμούς γονιδιακών συνόλων είναι η βάση δεδομένων Gene Ontology (GO), η οποία προσπαθεί να περιγράψει τις λειτουργίες γονιδίων χρησιμοποιώντας τρεις ιεραρχικές βιολογικές κατηγορίες:
 - μοριακές λειτουργίες,
 - βιολογικές διεργασίες και
 - κυτταρικά συστατικά.

- Για τις αναλύσεις των γονιδιακών συνόλων, είναι κοινή πρακτική ο ορισμός γονιδιακών ομάδων με την ομαδοποίηση όλων των γονιδίων που σχετίζονται με συγκεκριμένη μοριακή λειτουργία ή βιολογική διαδικασία.

Οντολογίες

- Ωστόσο, αν και γονίδια που έχουν οριστεί σε μια συγκεκριμένη κατηγορία GO μπορεί να σχετίζονται με παρόμοιες λειτουργίες, αυτή η ομάδα δεν υποδεικνύει γνωστές σχέσεις ή αλληλεπιδράσεις μεταξύ αυτών των γονιδίων.
- Επιπλέον, όπως και μονοπάτια, ο ιεραρχικός χαρακτήρας του GO μπορεί να προκαλέσει σημαντική αλληλεπικάλυψη μεταξύ των κατηγοριών και αυτό έχει επιπτώσεις τόσο στην ανάλυση όσο και στην ερμηνεία των αποτελεσμάτων

Βιοδείκτες

- Τα γονίδια που έχουν προηγουμένως συσχετιστεί με μια συγκεκριμένη ασθένεια μπορούν να ομαδοποιηθούν για να σχηματίσουν έναν άλλο τύπο γονιδιακού συνόλου, έναν βιοδείκτη της ασθένειας.
- Η διαφορά μεταξύ ενός βιολογικού δείκτη ασθένειας και μιας κατηγορίας GO είναι ότι τα γονίδια που αποτελούν τον βιοδείκτη δεν έχουν αναγκαστικά κοινές λειτουργίες.
- Επίσης, σε αντίθεση με την έννοια του μονοπατιού ασθένειας που αναφέρθηκε παραπάνω, τα γονίδια βιοδεικτών δεν αλληλεπιδρούν απαραίτητα μεταξύ τους

Βιοδείκτες

- Η υπόθεση πίσω από τους πολυγονιδιακούς βιοδείκτες είναι απλώς ότι το γονιδιακό σύνολο συνδέεται ισχυρότερα με τη νόσο από οποιοδήποτε μεμονωμένο γονίδιο, αλλά για πολλούς βιοδείκτες αυτή η υπόθεση δεν επιβεβαιώνεται κατ' ανάγκη.
- Πράγματι, επειδή τα αποδεικτικά στοιχεία που υποστηρίζουν τα μεμονωμένα γονίδια βιοδεικτών μπορούν να προέρχονται από μελέτες που διεξάγονται σε ένα ευρύ φάσμα πληθυσμών, το σύνολο βιοδεικτών μπορεί να συλλάβει τη γενετική ετερογένεια μιας νόσου και όχι έναν γενικό δείκτη αυτής.

- Στην πράξη, πολλοί πολυγονιδιακοί βιοδείκτες που παρουσιάζουν στατιστική συσχέτιση σε επίπεδο πληθυσμού δεν είναι σε θέση να προβλέψουν με ακρίβεια την ατομική ευαισθησία για σκοπούς κλινικής χρήσης.
- Από τον αυστηρό ορισμό ενός κλινικού βιοδείκτη, επομένως, η ονομασία «συσχετισμένο σύνολο γονιδίων» είναι συχνά πιο ενδεδειγμένη

Βάσεις δεδομένων

Table 1. Overview of gene set databases frequently used in GSA. Major gene set databases were grouped into four categories: 'Pathways', databases comprising biochemical pathways; 'Functional', databases comprising functional-related gene sets; 'Phenotypic', databases comprising phenotypic gene sets; 'Molecularly-derived', databases comprising gene sets derived from primarily high-dimensional molecular data; 'Meta-database', databases redistributing gene sets from multiple gene set databases. NA, not available. All links were last accessed, July 16, 2016.

Database	Category	Reference	Link
KEGG PATHWAY	Pathways	(21)	http://www.genome.jp/kegg/pathway.html
REACTOME	Pathways	(22)	http://www.reactome.org/
PANTHER	Pathways	(58)	http://pantherdb.org/
Biocarta	Pathways	NA	http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways
Ingenuity	Pathways	QIAGEN's Ingenuity® Pathway Analysis	http://www.ingenuity.com/products/ipa
MetaCore	Pathways	Thomson Reuters	http://lsresearch.thomsonreuters.com/maps/
HumanCyc	Pathways	(59)	http://humancyc.org/
Gene Ontology	Functional	(23)	http://geneontology.org/
Human Phenotype Ontology	Phenotypic	(31)	http://human-phenotype-ontology.github.io/
Mouse Genome Informatics	Phenotypic	(29)	http://www.informatics.jax.org/genes.shtml
MitoCarta	Molecularly-derived	(32)	www.broadinstitute.org/pubs/MitoCarta
InWeb	Molecularly-derived	(36)	http://www.broadinstitute.org/mpg/dapple/dappleTMP.php
PathCards	Meta-database	(42)	http://pathcards.genecards.org/
Pathway Commons	Meta-database	(43)	http://www.pathwaycommons.org/
MSigDB	Meta-database	(41)	http://software.broadinstitute.org/gsea/msigdb
ConsensusPathDB	Meta-database	(40)	http://consensuspathdb.org/

Pers TH. Gene set analysis for interpreting genetic studies. Human molecular genetics. 2016 Aug 10;25(R2):R133-40.

TABLE I. A Selection of Gene Set Databases

Database	Canonical pathways	Functionally-related gene sets	Gene/protein interactions	Links/references
Pathway Commons	X			pathwaycommons.org, [Cerami et al., 2011]
PathCards	X			pathcards.genecards.org, [Belinky et al., 2015]
KEGG	X			genome.jp/kegg, [Kanehisa and Goto, 2000; Kanehisa et al., 2014]
Reactome	X			reactome.org, [Croft et al., 2014]
Biocarta	X			biocarta.com
Panther	X	X		pantherdb.org/data, [Mi et al., 2013]
NCI-PID	X			pid.nci.nih.gov, [Schaefer et al., 2009]
MSigDB	X	X		broadinstitute.org/gsea/msigdb, [Subramanian et al., 2005]
ConsensusPathDB	X		X	consensuspathdb.org, [Kamburov et al., 2013]
Gene Ontology		X		geneontology.org, [Ashburner et al., 2000; Gene Ontology Consortium, 2010]
STRING			X	string-db.org, [Franceschini et al., 2013]
HPRD			X	hprd.org, [Prasad et al., 2009]
Metacore*	X	X	X	thomsonreuters.com/metacore
Ingenuity*	X	X	X	ingenuity.com/products/ipa

*Proprietary database.

Mooney MA, Wilmot B. 2015. Gene set analysis: A step-by-step guide.
Am J Med Genet Part B 168B:517–527

Προβληματισμοί για τις βάσεις δεδομένων (1)

- Πρώτον, καμία μεμονωμένη βάση δεδομένων δεν μπορεί να καλύψει όλα τα σημερινά γνωστά σύνολα γονιδίων που σχετίζονται με την ανθρώπινη υγεία και την ασθένεια.
- Εντούτοις, υπάρχουν διάφορες προσπάθειες για την ενοποίηση του ετερογενούς τοπίου των βάσεων δεδομένων των γονιδίων (βελτιστοποίηση των αναγνωριστικών των γονιδίων, των μορφών δεδομένων και της πρόσβασης στα δεδομένα), όπως το ConsensuspathDB, η Molecular Signatures database (MSigDB), οι PathCards, Pathway Commons.

Pers TH. Gene set analysis for interpreting genetic studies. Human molecular genetics. 2016 Aug 10;25(R2):R133-40.

Enrichr (42)

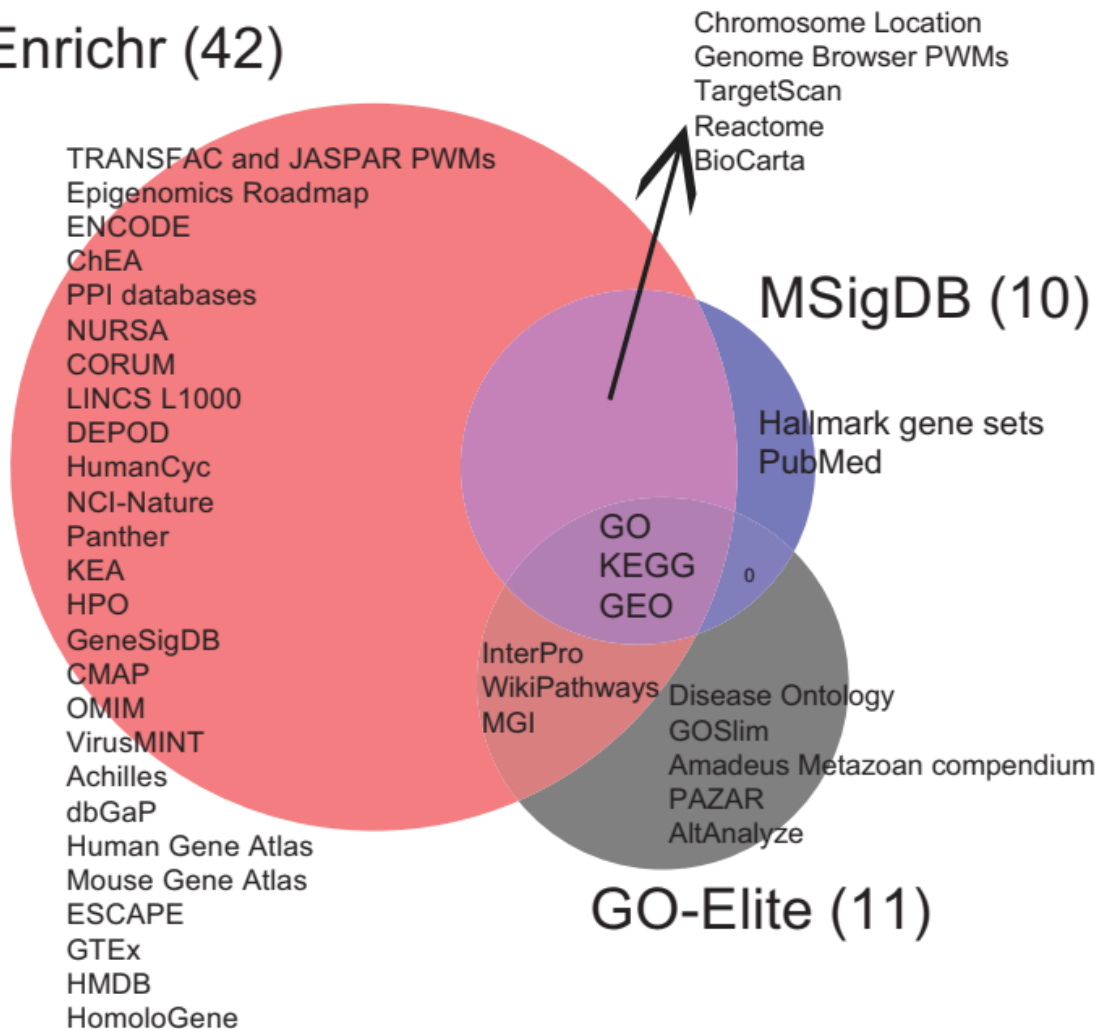


Figure 3. Comparing Enrichr resources with MSigDB and GO-Elite. (A) Venn diagram summarizing the various resources processed and served by Enrichr, MSigDB and GO-Elite. (B) Venn diagram to compare the number of processed gene sets of genetic and chemical perturbations curated from publications in Enrichr and MSigDB.

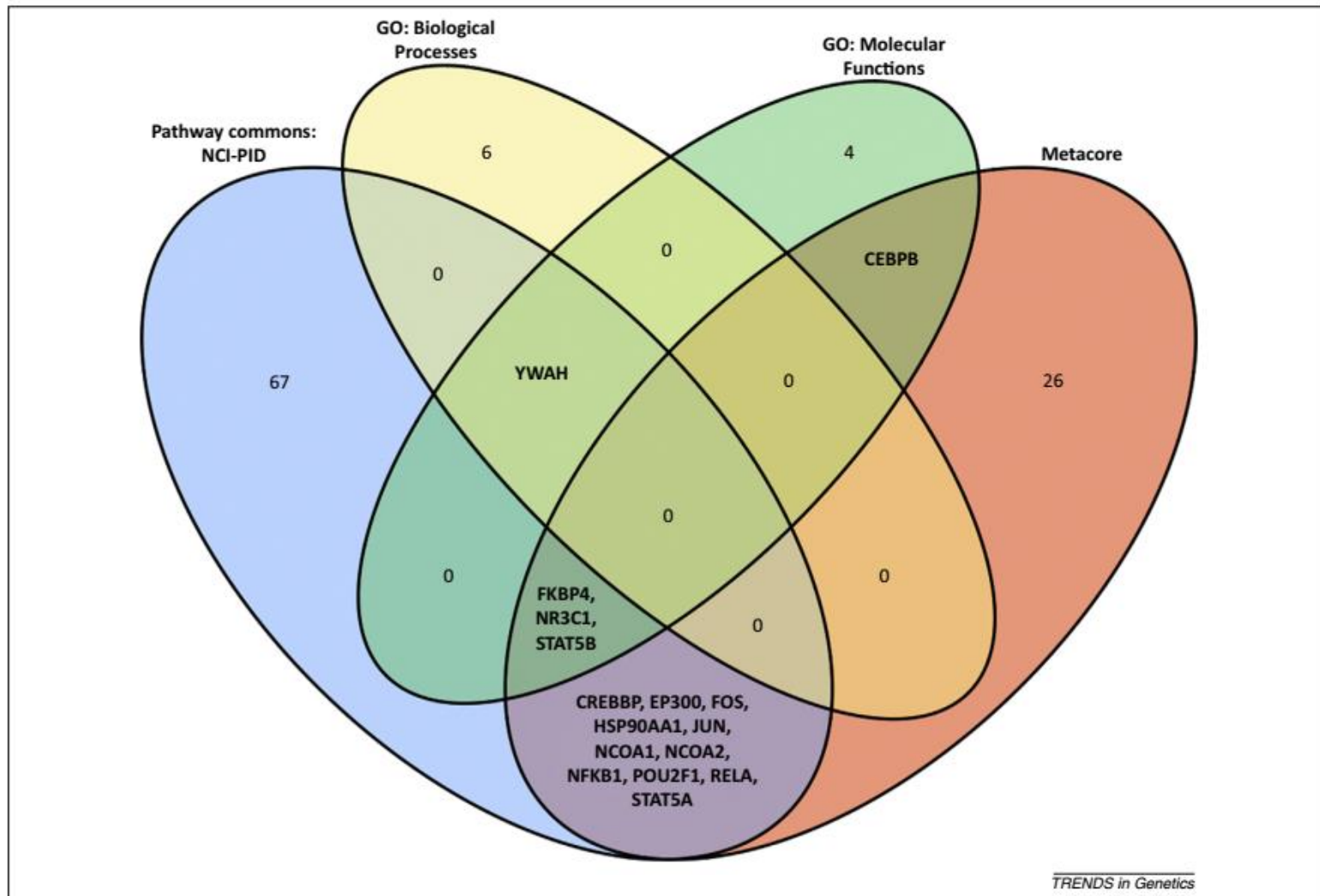
Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*. 2016 May 3;44(W1):W90-7.

Προβληματισμοί για τις βάσεις δεδομένων (2)

- Δεύτερον, τα σύνολα γονιδίων τείνουν να αλληλεπικαλύπτονται. Τα αλληλοεπικαλυπτόμενα σύνολα γονιδίων είναι εξ ορισμού πιο συχνά στις μεταβατικές οντολογίες, ωστόσο, υπάρχει μεγάλη αλληλοεπικάλυψη και μεταξύ των βάσεων δεδομένων των γονιδίων. Εκτός αν ληφθεί υπόψη στην ανάλυση, η παρουσία πολύ ομοίων ομάδων γονιδίων περιπλέκει την ερμηνεία των αποτελεσμάτων της GSA, επειδή συσχετισμένα σύνολα γονιδίων μπορεί να κυριαρχούν στην κορυφή των αποτελεσμάτων που κατατάσσονται και επομένως ενδεχομένως να επισκιάζουν τα χαμηλότερα καταγεγραμμένα ευρήματα.
- Επιπλέον, συσχετισμένα σύνολα γονιδίων μειώνουν τη στατιστική ισχύ για να ανιχνεύσουν σημαντικά σύνολα γονιδίων και μπορεί να οδηγήσουν σε λιγότερο ισχυρές εκτιμήσεις των ποσοστών των ψευδών ανακαλύψεων

Προβληματισμοί για τις βάσεις δεδομένων (3)

- Τρίτον, πρέπει να λαμβάνεται μέριμνα κατά την ερμηνεία των αποτελεσμάτων της GSA. Όπως σημειώνεται από τους Ramanathan et al., παρά τα παρόμοια ονόματα, τα γονιδιακά σύνολα από διαφορετικές πηγές μπορούν να περιέχουν αρκετά διαφορετικά γονίδια. Για παράδειγμα τα γονιδιακά σύνολα της KEGG PATHWAY και της REACTOME που σχετίζονται με το μεταβολισμό πουρίνης, παρά τα ίδια αναγνωριστικά, συσσωρεύονται μακρινά μεταξύ τους επειδή διαφέρουν στα γονίδια που περιέχουν.



Mooney MA, Nigg JT, McWeeney SK, Wilmot B. Functional and genomic context in pathway analysis of GWAS data. Trends in Genetics. 2014 Sep 1;30(9):390-400.

Προβληματισμοί για τις βάσεις δεδομένων (4)

- Τέλος, τα γονιδιακά σύνολα μπορούν να εμφανίζουν συστηματικό σφάλμα υπέρ της δημοσιευμένης δουλειάς. Η ταχέως αυξανόμενη διαθεσιμότητα των πολυδιάστατων μοριακών συνόλων δεδομένων είναι πιθανό να μειώσει αυτή την μεροληψία με την πάροδο του χρόνου, ωστόσο, τουλάχιστον αυτή τη στιγμή, η μέθοδος GSA βασισμένη σε εμπειρογνώμονες, μπορεί να χάσει στοιχεία από τη σχετική βιολογία.

Είδη μεθόδων

- Over-representation Analysis (ORA)
 - Οι μέθοδοι ORA τυπικά ελέγχουν αν τα στοιχεία των συνόλων δεδομένων μια παραπάνω από το αναμενόμενο εκπροσώπηση σε ένα δεδομένο μονοπάτι από μια βάση δεδομένων.
- Functional Class Scoring (FCS)
 - Οι μέθοδοι FCS επιδιώκουν συνήθως να ανιχνεύσουν αλλαγές στα στοιχεία του συνόλου δεδομένων που προκαλούν αλλοιώσεις στο δεδομένο μονοπάτι από τη βάση δεδομένων
- Pathway-Topology (PT) methods
 - Οι μέθοδοι PT ενσωματώνουν γνωστές ή εκτιμώμενες δομές του βιολογικού δικτύου για να λάβουν υπόψη τις συσχετίσεις μεταξύ των γονιδίων

Mathur R, Rotroff D, Ma J, Shojaie A, Motsinger-Reif A. Gene set analysis methods: a systematic comparison. *BioData mining*. 2018 Dec;11(1):8.

Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*. 2012 Feb 23;8(2):e1002375.

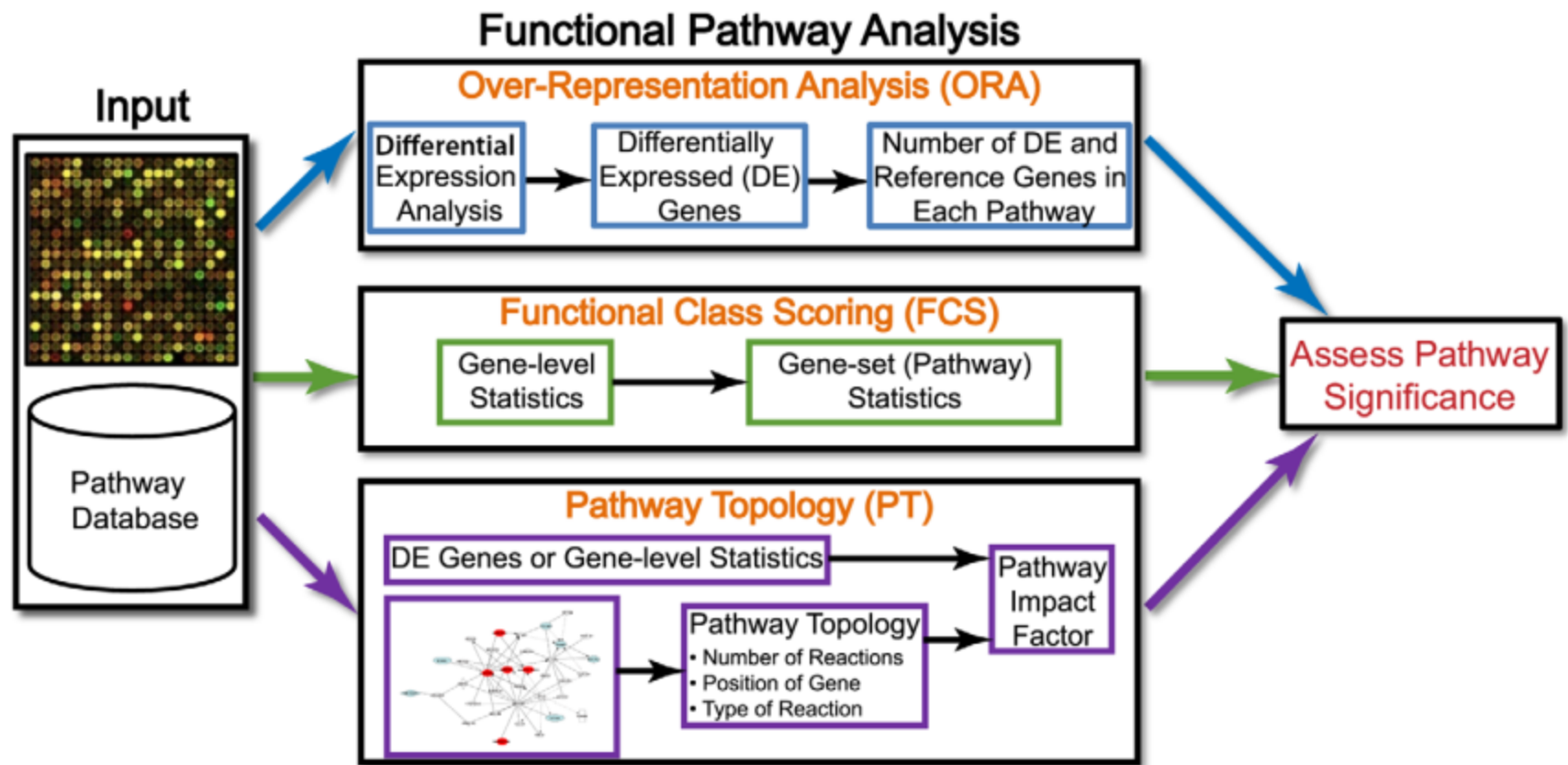


Figure 1. Overview of existing pathway analysis methods using gene expression data as an example. Note that this overview is equally applicable to molecular measurements using proteomics, and any other high-throughput technologies. The data generated by an experiment using a high-throughput technology (e.g., microarray, proteomics, metabolomics), along with functional annotations (pathway database) of the corresponding genome, are input to virtually all pathway analysis methods. While ORA methods require that the input is a list of differentially expressed genes, FCS methods use the entire data matrix as input. In addition to functional annotations of a genome, PT-based methods utilize the number and type of interactions between gene products, which may or may not be a part of a pathway database. The result of every pathway analysis method is a list of significant pathways in the condition under study. DE, differentially expressed.
doi:10.1371/journal.pcbi.1002375.g001

Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLOS Computational Biology* 8(2): e1002375.
<https://doi.org/10.1371/journal.pcbi.1002375>

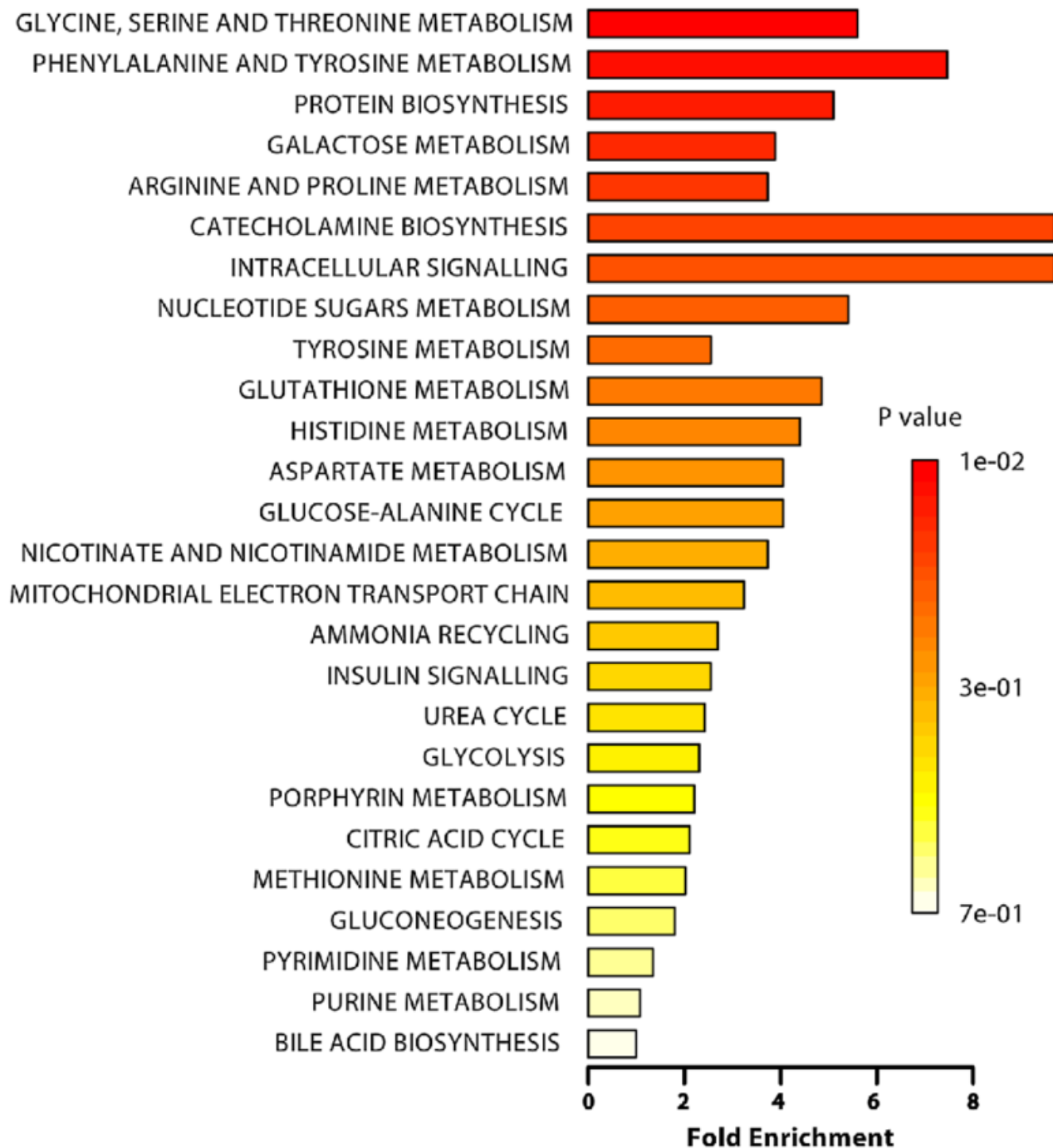
Over-representation Analysis (ORA)

- Η άμεση ανάγκη για λειτουργική ανάλυση των δεδομένων γονιδιακής έκφρασης μικροσυστοιχιών και η εμφάνιση του GO κατά τη διάρκεια αυτής της περιόδου οδήγησαν σε ανάλυση υπερεκπροσώπησης (ORA), η οποία αξιολογεί στατιστικά το κλάσμα των γονιδίων σε ένα συγκεκριμένο μονοπάτι που βρέθηκε μεταξύ των ομάδων γονιδίων που παρουσιάζουν αλλαγές έκφρασης.
- Αναφέρεται επίσης ως μέθοδος «Πίνακα 2x2» στη βιβλιογραφία.

Over-representation Analysis (ORA)

- Η ORA χρησιμοποιεί μία ή περισσότερες παραλλαγές της ακόλουθης στρατηγικής: πρώτον, δημιουργείται μια λίστα εισόδου χρησιμοποιώντας ένα συγκεκριμένο όριο σημαντικότητας.
 - Για παράδειγμα, ένας ερευνητής μπορεί να επιλέξει γονίδια που διαφέρουν στην έκφραση (ύπερ ή υπο-έκφραση) σε δεδομένη κατάσταση με ποσοστό ψευδούς ανακάλυψης (FDR) 5%.
- Στη συνέχεια, για κάθε μονοπάτι, υπολογίζονται τα γονίδια εισόδου που είναι μέρος του μονοπατιού.
- Αυτή η διαδικασία επαναλαμβάνεται για έναν κατάλληλο κατάλογο γονιδίων (π.χ., όλα τα γονίδια που μετριοούνται σε μια μικροσυστοιχία).
- Στη συνέχεια, κάθε μονοπάτι δοκιμάζεται για ύπερ ή υποεκπροσώπηση στη λίστα των γονιδίων εισόδου.
- Οι πιο συχνά χρησιμοποιούμενες δοκιμές βασίζονται στην υπεργεωμετρική κατανομή, στο κριτήριο chi-square ή στη διωνυμική κατανομή.

Metabolite Sets Enrichment Overview



Περιορισμοί

- Πρώτον, τα διαφορετικά στατιστικά που χρησιμοποιούνται από την ORA (π.χ. υπεργεωμετρική κατανομή, διωνυμική κατανομή, χ^2 κατανομή κ.λπ.) είναι ανεξάρτητα από τις μετρούμενες αλλαγές.
- Αυτό σημαίνει ότι αυτές οι δοκιμές εξετάζουν τον αριθμό των γονιδίων μόνο και αγνοούν τις τιμές που σχετίζονται με αυτές, όπως οι εντάσεις των ανιχνευτών

Περιορισμοί

- Δεύτερον, η ORA συνήθως χρησιμοποιεί μόνο τα πιο σημαντικά γονίδια και απορρίπτει τα άλλα.
- Για παράδειγμα, ο κατάλογος εισαγωγής γονιδίων από ένα πείραμα μικροσυστοιχιών λαμβάνεται συνήθως χρησιμοποιώντας ένα αυθαίρετο κατώφλι (π.χ., γονίδια με μεταβολή fold >2 και / ή p-value <0: 05).
- Με αυτή τη μέθοδο, χάνονται λιγότερο σημαντικά γονίδια (π.χ., fold change = 1.999 ή p-value = 0.051), με αποτέλεσμα την απώλεια πληροφορίας.

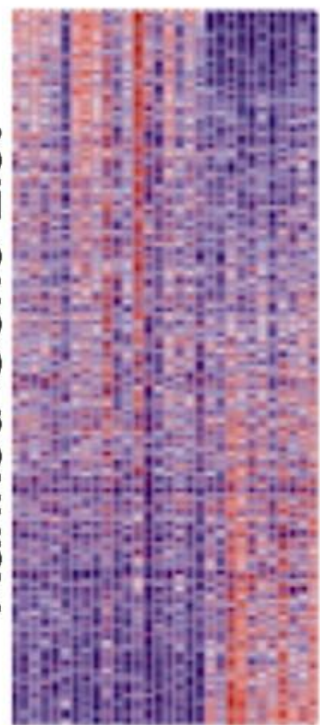
Περιορισμοί

- Τρίτον, με την ίδια αντιμετώπιση κάθε γονιδίου, η ORA υποθέτει ότι κάθε γονίδιο είναι ανεξάρτητο από τα υπόλοιπα.
- Ωστόσο, η βιολογία είναι ένα πολύπλοκο δίκτυο αλληλεπιδράσεων μεταξύ γονιδιακών προϊόντων που συνθέτουν διαφορετικά μονοπάτια.
- Ένας στόχος της ανάλυσης γονιδιακής έκφρασης μπορεί να είναι να αποκτήσει γνώσεις σχετικά με το πώς οι αλληλεπιδράσεις μεταξύ γονιδιακών προϊόντων εκδηλώνονται ως αλλαγές στην έκφραση γονιδίων.
- Μια στρατηγική που υποθέτει ότι τα γονίδια είναι ανεξάρτητα είναι σημαντικά περιορισμένη στην ικανότητά της να παρέχει πληροφορίες σχετικά με αυτό το θέμα

Περιορισμοί

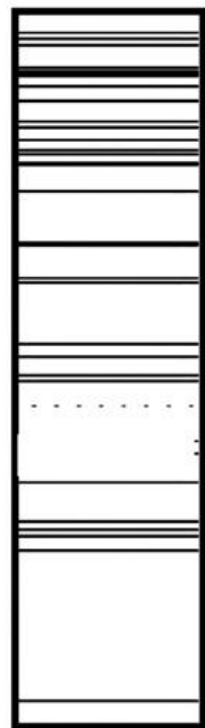
- Τέταρτον, η ORA υποθέτει ότι κάθε μονοπάτι είναι ανεξάρτητο από τα άλλα, κάτι το οποίο είναι λανθασμένο.
- Για παράδειγμα, ο GO ορίζει μια βιολογική διαδικασία ως μια σειρά γεγονότων που πραγματοποιούνται από μία ή περισσότερες ταξινομημένες συναθροίσεις μοριακών λειτουργιών (<http://www.geneontology.org/GO.doc.shtml>)
- Ένα άλλο παράδειγμα εξάρτησης μεταξύ μονοπατιών είναι η διαδρομή του κυτταρικού κύκλου στην KEGG (<http://www.genome.jp/kegg/pathway/hsa/hsa04110.html>), όπου η παρουσία ενός αυξητικού παράγοντα ενεργοποιεί τη μονοπάτι σηματοδότησης MAPK. Αυτό, με τη σειρά του, ενεργοποιεί το μονοπάτι του κυτταρικού κύκλου.
- Δεν υπάρχουν μέθοδοι ORA για αυτή την εξάρτηση μεταξύ των μοριακών λειτουργιών της GO και των μονοπατιών σηματοδότησης της KEGG

A Phenotype
Classes



B

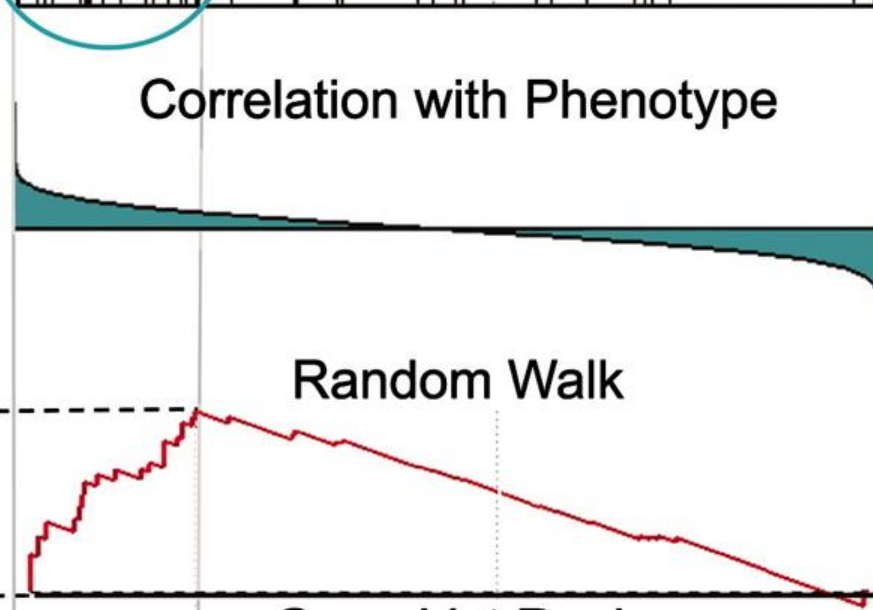
Gene set S



Leading edge subset
Gene set S



Correlation with Phenotype

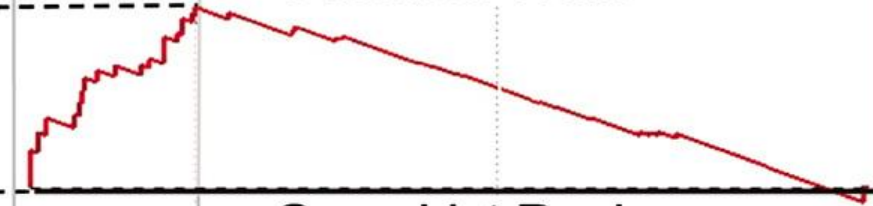


Random Walk

$ES(S)$

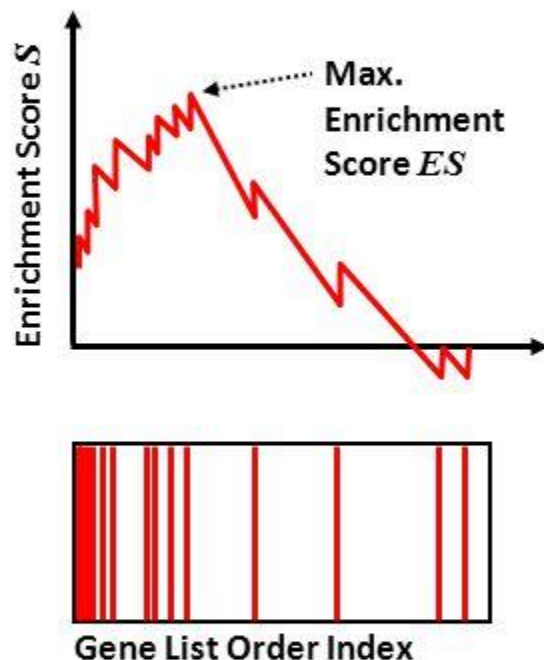
Maximum deviation
from zero provides the
enrichment score $ES(S)$

Gene List Rank

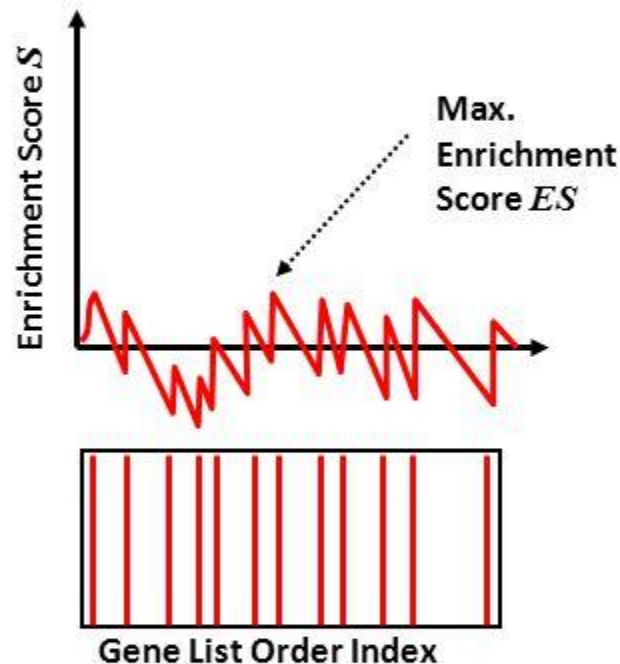


Enrichment: KS-score

Enriched Gene Set



Un-enriched Gene Set



Every hit go up by $1/N_H$

Every miss go down by $1/N_M$

The maximum height provides the enrichment score

Enrichment FDR	Genes in list	Total genes	Functional Category
6.5E-220	86	101	DNA damage checkpoint
5.3E-215	86	108	DNA integrity checkpoint
3.2E-188	86	169	Cell cycle checkpoint
1.1E-131	87	659	Cellular response to DNA damage stimulus
1.2E-113	87	1039	Cell cycle process
3.6E-102	87	1395	Cell cycle
1.1E-100	45	57	Mitotic DNA integrity checkpoint
5.0E-99	87	1517	Cellular response to stress
1.9E-98	43	51	Mitotic DNA damage checkpoint

Functional Class Scoring (FCS)

- Η υπόθεση που κάνει η μέθοδος functional class scoring (FCS) είναι ότι αν και οι μεγάλες μεταβολές στα μεμονωμένα γονίδια μπορούν να έχουν σημαντικές επιδράσεις στα μονοπάτια, οι ασθενέστερες αλλά συντονισμένες αλλαγές στα σύνολα λειτουργικά συγγενών γονιδίων (δηλ., στα μονοπάτια) μπορούν επίσης να έχουν σημαντικά αποτελέσματα. Με λίγες εξαιρέσεις, όλες οι μέθοδοι FCS χρησιμοποιούν μια παραλλαγή ενός γενικού πλαισίου που αποτελείται από τα ακόλουθα τρία βήματα:
 - υπολογίζεται ένα στατιστικό στο επίπεδο του γονιδίου
 - Τα στατιστικά σε επίπεδο γονιδίων για όλα τα γονίδια σε ένα μονοπάτι συσσωματώνονται σε ένα ενιαίο στατιστικό σε επίπεδο μονοπατιού
 - Το τελικό βήμα στο FCS είναι η εκτίμηση της στατιστικής σημαντικότητας του στατιστικού σε επίπεδο μονοπατιού

Περιορισμοί

- Καταρχάς, υπολογίζεται ένα στατιστικό στο επίπεδο του γονιδίου χρησιμοποιώντας τις μοριακές μετρήσεις από ένα πείραμα. Αυτό συνεπάγεται υπολογισμό της διαφορικής έκφρασης μεμονωμένων γονιδίων ή πρωτεϊνών. Οι στατιστικές που χρησιμοποιούνται επί του παρόντος σε επίπεδο γονιδίου περιλαμβάνουν συσχέτισμό μοριακών μετρήσεων με φαινότυπο, ANOVA, στατιστική Q, λόγο σήματος προς θόρυβο, t-test

Περιορισμοί

- Δεύτερον, τα στατιστικά σε επίπεδο γονιδίων για όλα τα γονίδια σε ένα μονοπάτι συσσωματώνονται σε ένα ενιαίο στατιστικό σε επίπεδο μονοπατιού. Αυτή η στατιστική μπορεί να είναι πολυπαραγοντική και να υπολογίζει τις αλληλεξαρτήσεις μεταξύ των γονιδίων ή μπορεί να είναι μονομεταβλητή και να μην λαμβάνει υπόψη τις αλληλεξαρτήσεις μεταξύ των γονιδίων. Οι στατιστικές σε επίπεδο μονοπατιού που χρησιμοποιούνται από τις τρέχουσες προσεγγίσεις περιλαμβάνουν το στατιστικό Kolmogorov-Smirnov, το άθροισμα, το μέσο όρο ή το διάμεσο της στατιστικής σε επίπεδο γονιδίου, το άθροισμα βαθμών κατάταξης Wilcoxon και το στατιστικό στοιχείο maxmean

Περιορισμοί

- Το τελικό βήμα στο FCS είναι η εκτίμηση της στατιστικής σημαντικότητας της στατιστικής σε επίπεδο μονοπατιού. Όταν υπολογίζεται η στατιστική σημαντικότητα, η μηδενική υπόθεση που εξετάζεται από τις προσεγγίσεις της τρέχουσας ανάλυσης μονοπατιών μπορεί να χωριστεί σε δύο κατηγορίες: i) ανταγωνιστική μηδενική υπόθεση και ii) αυτοτελής μηδενική υπόθεση

Πλεονεκτήματα

- Οι μέθοδοι CS αντιμετωπίζουν τρεις βασικούς περιορισμούς των μεθόδων ORA. Πρώτον, δεν απαιτούν ένα αυθαίρετο όριο για τη διαίρεση των δεδομένων έκφρασης σε σημαντικές και μη σημαντικές ομάδες. Αντίθετα, οι μέθοδοι FCS χρησιμοποιούν όλες τις διαθέσιμες μοριακές μετρήσεις για την ανάλυση του μονοπατιού.
- Δεύτερον, ενώ η ORA αγνοεί εντελώς τις μοριακές μετρήσεις όταν εντοπίζει σημαντικά μονοπάτια, οι μέθοδοι FCS χρησιμοποιούν αυτές τις πληροφορίες για να ανιχνεύσουν συντονισμένες αλλαγές στην έκφραση γονιδίων στο ίδιο μονοπάτι.
- Τέλος, εξετάζοντας τις συντονισμένες αλλαγές στην γονιδιακή έκφραση, οι μέθοδοι FCS λαμβάνουν υπόψη την εξάρτηση μεταξύ των γονιδίων σε μια οδό, κάτι το οποίο δεν κάνει η ORA

Περιορισμοί

- Πρώτον, όμοια με την ORA, η FCS αναλύει κάθε μονοπάτι ανεξάρτητα. Αυτός είναι ένας περιορισμός επειδή ένα γονίδιο μπορεί να λειτουργήσει σε περισσότερα από ένα μονοπάτια, πράγμα που σημαίνει ότι τα μονοπάτια μπορούν να διασταυρώνονται και να επικαλύπτονται. Συνεπώς, σε ένα πείραμα, ενώ ένα μονοπάτι μπορεί να επηρεαστεί σε ένα πείραμα, μπορεί κανείς να παρατηρήσει και άλλα μονοπάτια να εμφανίζονται ότι επηρεάζονται σημαντικά εξαιτίας του συνόλου επικαλυπτόμενων γονιδίων

Περιορισμοί

- Δεύτερον, πολλές μέθοδοι FCS χρησιμοποιούν τις αλλαγές στην έκφραση των γονιδίων για να ταξινομήσουν γονίδια σε ένα δεδομένο μονοπάτι και δεν χρησιμοποιούν τις διαφορές αυτές σε καμία περαιτέρω ανάλυση.
- Για παράδειγμα, αν υποθέσουμε ότι δύο γονίδια σε ένα μονοπάτι, A και B, αλλάζουν κατά 2 φορές και 20 φορές, αντίστοιχα. Εφόσον και τα δύο εμφανίζονται στην ίδια σειρά σε σύγκριση με άλλα γονίδια στο μονοπάτι, οι περισσότερες μέθοδοι FCS θα τα αντιμετωπίσουν με τον ίδιο τρόπο, αν και το γονίδιο με την υψηλότερη αλλαγή στην έκφραση θα έπρεπε να λάβει περισσότερο βάρος

Pathway-Topology (PT) methods

- Οι μέθοδοι ORA και FCS εξετάζουν μόνο τον αριθμό των γονιδίων σε ένα μονοπάτι ή την ταυτόχρονη έκφραση γονιδίων για τον εντοπισμό σημαντικών μονοπατιών και αγνοούν τις πρόσθετες πληροφορίες που είναι διαθέσιμες από αυτές τις βάσεις δεδομένων.
- Επομένως, ακόμη και αν κάποιος σχεδιάζε ξανά τα μονοπάτια, με νέες συνδέσεις μεταξύ των γονιδίων, εφόσον περιέχουν το ίδιο σύνολο γονιδίων, οι μέθοδοι ORA και FCS θα παρήγαγαν τα ίδια αποτελέσματα.

- Οι μέθοδοι που βασίζονται στην τοπολογία των μονοπατιών(PT) έχουν αναπτυχθεί για να χρησιμοποιήσουν αυτές τις πρόσθετες πληροφορίες.
- Οι μέθοδοι που βασίζονται σε PT είναι ουσιαστικά οι ίδιες με τις μεθόδους FCS στο ότι εκτελούν τα ίδια τρία βήματα με τις μεθόδους FCS.
- Η βασική διαφορά μεταξύ των δύο είναι η χρήση της τοπολογίας μονοπατιού για τον υπολογισμό των στατιστικών στοιχείων σε επίπεδο γονιδίων

Περιορισμοί

- Αν και οι μέθοδοι που βασίζονται σε PT είναι δύσκολο να περιγραφούν με γενικούς όρους, έχουν πολλούς κοινούς περιορισμούς.
- Ένα προφανές πρόβλημα είναι ότι η πραγματική τοπολογία του μονοπατιού εξαρτάται από τον τύπο του κυττάρου λόγω των προφίλ έκφρασης των γονιδίων και των συνθηκών της συγκεκριμένης κυτταρικής έκφρασης που μελετώνται.
- Ωστόσο, αυτές οι πληροφορίες είναι σπάνια διαθέσιμες και κατακερματίζονται σε βάσεις γνώσεων, ακόμη και αν είναι πλήρως κατανοητές. Καθώς βελτιώνονται οι σχολιασμοί, οι προσεγγίσεις αυτές αναμένεται να γίνουν πιο χρήσιμες

Περιορισμοί

- Άλλοι περιορισμοί των μεθόδων που βασίζονται σε ΡΤ περιλαμβάνουν την αδυναμία να μοντελοποιηθούν οι δυναμικές καταστάσεις ενός συστήματος και η αδυναμία να ληφθούν υπόψη οι αλληλεπιδράσεις μεταξύ μονοπατιών λόγω αδύναμων συνδέσεων μεταξύ διαδρόμων, ώστε να ληφθεί υπόψη η αλληλεξάρτηση μεταξύ οδών.

Προβληματισμοί στην Ανάλυση μονοπατιών

- Προβλήματα στον σχολιασμό και τα δεδομένα
 - Βάσεις δεδομένων χαμηλής ποιότητας/ανάλυσης
 - Ελλιπείς και ανακριβείς σχολιασμοί
 - Ελλιπείς πληροφορίες για συγκεκριμένες συνθήκες και συγκεκριμένους ιστούς/κύτταρα
- Προβλήματα στη μεθοδολογία
 - Benchmark data sets για σύγκριση διαφορετικών μεθόδων.
 - Αδυναμία μοντελοποίησης και ανάλυσης δυναμικής απόκρισης
 - Αδυναμία μοντελοποίησης των επιδράσεων ενός εξωτερικού ερεθίσματος.

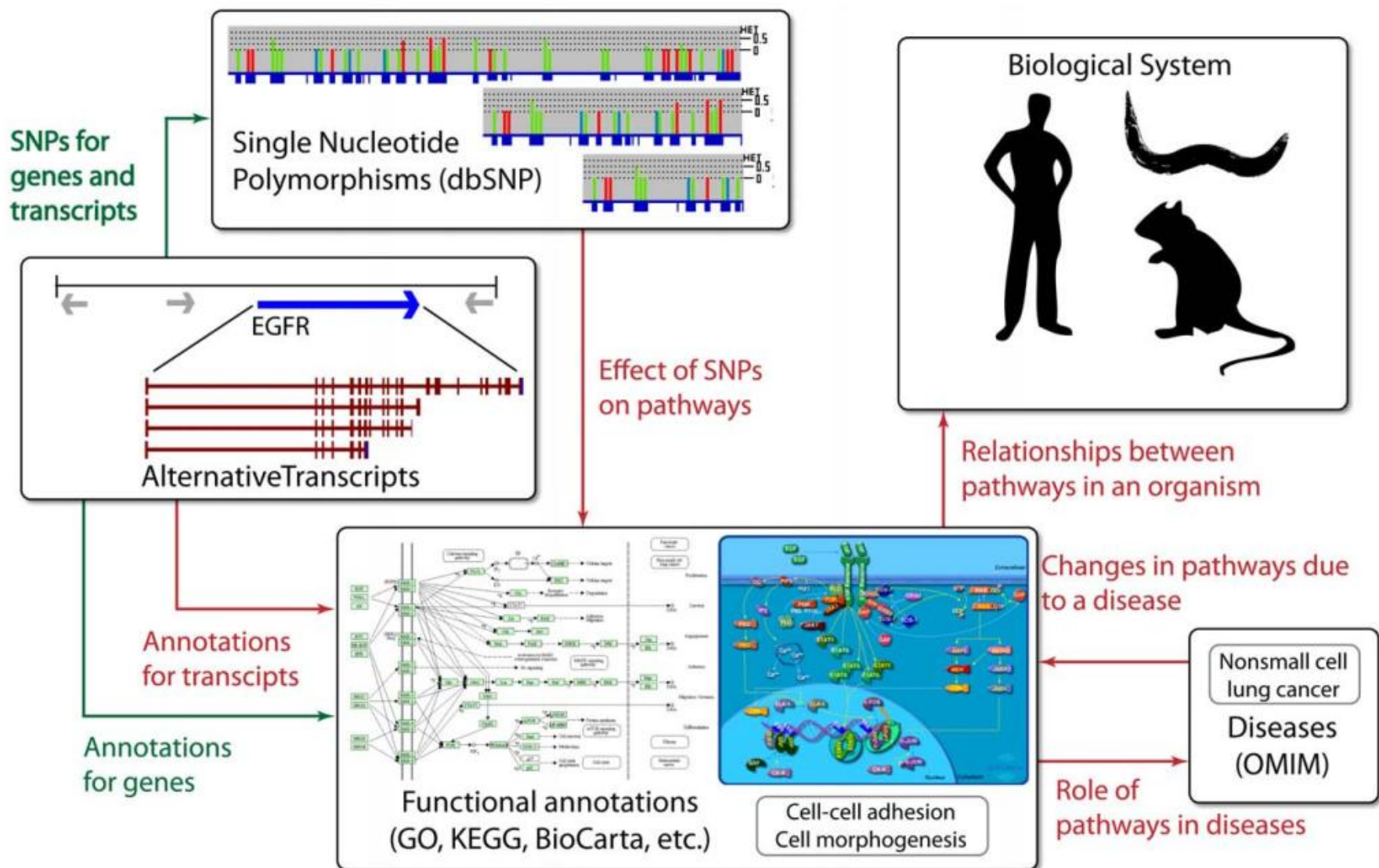


Figure 2. Overview of low resolution, missing, and incomplete information. Green arrows represent abundantly available information, and red arrows represent missing and/or incomplete information. The ultimate goal of pathway analysis is to analyze a biological system as a large, single network. However, the links between smaller individual pathways are not yet well known. Furthermore, the effects of a SNP on a given pathway are also missing from current knowledge bases. While some pathways are known to be related to a few diseases, it is not clear whether the changes in pathways are the cause for those diseases or the downstream effects of the diseases.

doi:10.1371/journal.pcbi.1002375.g002

Είδη στατιστικών ελέγχων

- αυτοτελείς μηδενικές υποθέσεις
 - Μια αυτοτελής μηδενική υπόθεση δηλώνει ότι τα γονίδια στο σύνολο γονιδίων δεν εκφράζονται πιο διαφορετικά από ό,τι αναμένεται να συμβεί λόγω τύχης.
- ανταγωνιστικές μηδενικές υποθέσεις.
 - Μια ανταγωνιστική μηδενική υπόθεση δηλώνει ότι τα γονίδια σε ένα ομάδα γονιδίων δεν εκφράζονται πιο διαφορετικά σε σχέση με τα άλλα γονίδια του συνόλου δεδομένων (άλλων ομάδων).
- Η ανταγωνιστική μηδενική υπόθεση απαντά σε ερωτήσεις σχετικά με τα μονοπάτια συγκρινόμενα μεταξύ τους, ενώ η αυτοτελής υπόθεση απαντά σε γενικότερες ερωτήσεις σχετικά με τις δραστηριότητες των γονιδίων σε κάθε συγκεκριμένη οδό.

Example A:

	Significant	Not Significant	
SNP in gene set G	20	80	100
SNP outside gene set G	100	400	500
	120	480	600 SNPs

- 20% of SNPs within G significant
- 20% of SNPs outside of G significant
- $P = 0.55$ for Fisher's exact test of the competitive hypothesis
- No evidence of enrichment

Example B:

	Significant	Not Significant	
SNP in gene set G	40	60	100
SNP outside gene set G	100	400	500
	140	460	600 SNPs

- 40% of SNPs within G significant
- 20% of SNPs outside G significant
- $P < 0.001$ for Fisher's exact test of the competitive hypothesis
- Evidence of enrichment

**Note that the statistical test applied here assumes independence of p-values, which is an invalid assumption in the presence of LD. Here this simple test is only used to illustrate the competitive hypothesis.*

Figure 1 Examples of competitive testing using Fisher's exact test.

Number of SNPs in gene set <i>G</i> significant with $p < 0.05$		
	Significant	Not Significant
Observed	20	80
Expected	5	95

- 20% of SNPs within *G* significant.
- Under the null hypothesis, expect 5% of the SNPs to be significant.
- $P = 0.002$ for Fisher's exact test of the self-contained hypothesis.
- Evidence of association of the gene set with the trait.

**Note that the statistical test applied here assumes independence of p -values, which is an invalid assumption in the presence of LD. Here this simple test is only used to illustrate the self-contained hypothesis.*

Figure 2 Example of a self-contained GSA based on the Fisher's exact test.

- ανταγωνιστικές μηδενικές υποθέσεις (Q1) —
 - Τα γονίδια σε ένα σύνολο γονιδίων δείχνουν το ίδιο μέγεθος συσχετισμών με τον φαινότυπο της νόσου
- αυτοτελείς μηδενικές υποθέσεις (Q2) —
 - Τα γονίδια σε ένα σύνολο γονιδίων δεν σχετίζονται με τον φαινότυπο της νόσου
- Ενώ ένα ανταγωνιστικό τεστ συγκρίνει στατιστικά από δοκιμές συσχέτισης ασθενειών για γονίδια σε ένα σύνολο γονιδίων με αυτά για γονίδια στο υπόλοιπο γονιδίωμα, ένα αυτοτελές τεστ ελέγχει απευθείας τη συσχέτιση ενός συνόλου γονιδίων με την ασθένεια και δεν εξαρτάται από γονίδια εκτός του συνόλου

Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z. Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics*. 2011 Jul 1;98(1):1-8.

- Για μια αυτοτελή μηδενική υπόθεση, η απλούστερη προσέγγιση είναι να συνδυάσουμε τις τιμές P όλων των μελών ενός γονιδιακού συνόλου. Είναι πιο συνηθισμένο να υπολογίζουμε αρχικά τις τιμές P σε επίπεδο γονιδίου, αλλά είναι επίσης δυνατό να συνδυάσουμε τις τιμές P των SNP. Μια ποικιλία μεθόδων συνδυασμού πολλαπλών τιμών p είναι διαθέσιμες, όπως η μέθοδος Fisher, η μέθοδος γάμμα και άλλες

Στατιστική ταξινόμηση των διαφόρων μεθόδων

Maciejewski H. Gene set analysis methods: statistical models and methodological differences. Briefings in bioinformatics. 2013 Feb 9;15(4):504-18.

Method	Statistic	Significance assessment
Competitive with gene randomization		
QI (Tian <i>et al.</i> [14])	$QI = \frac{1}{m} \sum_{i=1}^m t_i$	Gene randomization
FCS—Functional Class Score (Pavlidis <i>et al.</i> [24])	$FCS = \frac{1}{m} \sum_{i=1}^m -\log(P_i)$	Gene randomization
Competitive with sample randomization		
GSEA (Subramanian <i>et al.</i> [1])	Kolmogorov–Smirnov statistic comparing ranks of P -values of genes in gene set versus uniform distribution	Sample randomization
GSA (Efron and Tibshirani [9])	The <i>maxmean</i> statistic $S_{\max}$ equal: $\max \left\{ \left \frac{\sum_{i=1}^m I(t_i > 0) t_i}{m} \right , \left \frac{\sum_{i=1}^m I(t_i < 0) t_i}{m} \right \right\}$	Sample randomization for standardized test statistics
SAFE (Barry <i>et al.</i> [10])	Kolmogorov–Smirnov or Wilcoxon rank-sum statistic comparing t versus t^C	Sample randomization
Self-contained		
Globaltest (Goeman <i>et al.</i> [5])	$Q = \frac{1}{m} \sum_{i=1}^m \frac{1}{\mu_2} [X_i \cdot (Y - \mu)]^2$ where μ and μ_2 are the first and second central moment of Y	Asymptotic normal distribution, or for small sample approximation by scaled χ^2 distribution, or randomization of samples
Q2 (Tian <i>et al.</i> [14])	Same as QI	Sample randomization
FCS.SC—self-contained version of FCS (Pavlidis <i>et al.</i> [24])	Same as FCS	Sample randomization
ES.SC—self-contained version of the enrichment score (Subramanian <i>et al.</i> [1])	Kolmogorov–Smirnov statistic comparing $(P_i), i = 1, \dots, m$ versus the uniform distribution	Analytical Kolmogorov–Smirnov distribution or randomization of samples
Parametric		
PAGE (Kim and Volksy [15])	$z = \frac{1}{\delta} (\mu - \mu_c) \sqrt{m}$ where μ, δ are the mean and standard deviation of fold changes calculated for all genes, and μ_G is the mean of fold changes for genes in G	Null distribution of $z \sim N(0,1)$
CATEGORY (Jiang <i>et al.</i> [25], Irizarry <i>et al.</i> [16])	$z = \frac{1}{\sqrt{m}} \sum_{i=1}^m t_i$	Null distribution of $z \sim N(0,1)$
Other parametric tests proposed by Irizarry <i>et al.</i> [16]	Wilcoxon rank-sum statistic or χ^2 statistic comparing t versus t^C	Corresponding analytical distributions

Table 1

Some examples of competitive tests, which compare disease associations for the genes in a gene-set with genes in the rest of the genome.

Reference	Year	Software	Input data for the method	Condense SNP information within each gene	Gene set test statistic	Significance assessment
<i>Gene-based methods</i>						
Wang et al. [5]	2007	GSEA http://www.openbioinformatics.org/gengen	Genotype	Most significant SNP <i>P</i> -value; Sime's combination test	Modified Kolmogorov–Smirnov (KS) statistic	Sample permutations
Askland et al. [77]	2009	EVA (Exploratory Visual Analysis) http://www.exploratoryvisualanalysis.org/	SNP <i>P</i> -values	Most significant SNP <i>P</i> -value	Fisher's exact test	Hypergeometric distribution
Guo et al. [51]	2009		SNP <i>P</i> -values	Most significant SNP <i>P</i> -value	Modified Kolmogorov–Smirnov statistic	SNP permutations
Holmans et al. [31]	2009	ALIGATOR (Association List Go AnnoTatOR) http://x004.psych.uwcm.ac.uk/~peter/	SNP <i>P</i> -values	Most significant SNP <i>P</i> -value with correction for gene size	Modified Fisher's Exact test	Gene re-samplings
Freudenberg et al. [47]	2010		Genotype	Most significant SNP <i>P</i> -value	Odds ratio for the presence of SNP associations; number of loci in a category that have SNP associations	Sample permutations
Jia et al. [26]	2010	dmGWAS http://bioinfo.mc.vanderbilt.edu/dmGWAS.html	Genotype	Most significant SNP <i>P</i> -value	Z-score	Gene randomization and sample permutations
Luo et al. [70]	2010		SNP <i>P</i> -values	Linear combination test; quadratic test; decorrelation test of SNP <i>P</i> -values	Linear combination test; quadratic test; decorrelation test of gene <i>P</i> -values	Normal or Chi-square distribution
Nam et al. [32]	2010	GSA-SNP http://gsa.muldass.org	SNP <i>P</i> -values	Second best SNP <i>P</i> -value	Z-statistic, maxmean statistic [36], and modified KS statistic [5]	Gene re-samplings and sample permutations
Peng et al. [41]	2010		SNP <i>P</i> -values	Fisher's combined <i>P</i> -value; Sidak's correction to the most significant SNP; Sime's combination test; or FDR method	Fisher's exact test	Hypergeometric distribution
Zhang et al. [87]	2010	i-GSEA4GWAS http://gsea4gwas.psych.ac.cn/	SNP <i>P</i> -values	Most significant SNP <i>P</i> -value	Modified Kolmogorov–Smirnov (KS) statistic	SNP permutations
<i>SNP-based methods</i>						
Holden et al. [88]	2008	GSEA-SNP http://nr.no/pages/samba/area_emr_smbi_gseasnp	Genotype		Modified KS statistic	Sample permutations
Schwarz et al. [89]	2008	SNPtoGO http://webtools.imbs.uni-luebeck.de/snptogo	SNP <i>P</i> -values		Fisher's exact test	Hypergeometric distribution
Medina et al. [90]	2009	GESBAP (GEne Set Based Analysis of Polymorphisms) http://bioinfo.cipf.es/gesbap/www/index.jsp	SNP <i>P</i> -values		Sequential applications of Fisher's exact test on different partitions of the gene list	Hypergeometric distribution corrected by FDR[91]

Table 2
Some examples of self-contained tests, which test for disease associations for genes in a gene-set directly.

Reference	Year	Software	Input data for the method	Condense SNP information within each gene	Gene set test statistic	Significance assessment
<i>Gene-based methods</i>						
Yu et al. [37]	2009		SNP <i>P</i> -values	Adaptive rank truncated product statistic (ARTP) method	Adaptive rank truncated product statistic (ARTP)	An efficient single-level permutation algorithm
Chen et al. [33]	2010	GRASS (Gene set Ridge regression in ASsociation Studies) http://linchen.fhcrc.org/grass.html	Genotype	Principal components		Sample permutations
<i>SNP-based methods</i>						
Dinu et al. [92]	2007		Genotype		U-statistic [93]	Sample permutations
Chai et al. [34]	2009		Genotype		Fisher's combined <i>P</i> -value, corrected by Brown's approximation	Chi-square distribution
O'Dushlaine et al. [6]	2009	SNP Ratio Test http://sourceforge.net/projects/snpratiotest/	Genotype		SNP Ratio Test	Sample permutations
De la Cruz et al. [94]	2009		Genotype		Fisher's combined <i>P</i> -value, with rank truncation and weights	Sample permutations
Chen et al. [12]	2010		Genotype		Supervised Principal Components	Mixture distribution
Eleftherohorinou et al. [73]	2010		Genotype		Cumulative trend test statistics – sum of all single SNP <i>P</i> -values in the gene set	Fit skewed normal distribution to 1000 sample permutations
Ruano et al. [68]	2010		Genotype		Fisher's combined <i>P</i> -value	Sample permutations
Wang et al. [55]	2011		SNP <i>P</i> -values		<i>t</i> -statistic in mixed model	Empirical null distribution

Ιδιαιτερότητες GWAS/microarrays

- Πώς να φιλτράρουμε ή να καθαρίσουμε τη λίστα των συνόλων γονιδίων;
- Πως θα αντιστοιχίσουμε SNPs/probes σε γονίδια?
- Είναι σωστό να χρησιμοποιήσουμε imputed genotypes?

Mooney MA, Wilmot B. 2015. Gene set analysis: A step-by-step guide.

Am J Med Genet Part B 168B:517–527

Mooney MA, Nigg JT, McWeeney SK, Wilmot B. Functional and genomic context in pathway analysis of GWAS data. Trends in Genetics. 2014 Sep 1;30(9):390-400.

Πώς να φιλτράρουμε ή να καθαρίσουμε τη λίστα των συνόλων γονιδίων;

- είναι συχνά απαραίτητο να φιλτράρουμε τα σύνολα γονιδίων για να αφαιρέσουμε αυτά με πολύ μικρό και αυτά με πολύ μεγάλο αριθμό γονιδίων.
- Αυτό το βήμα είναι σημαντικό λόγω των γνωστών συστηματικών σφαλμάτων που σχετίζονται με το μέγεθος (αριθμός γονιδίων) ενός συνόλου γονιδίων.
- Αν και τα όρια είναι αυθαίρετα, είναι σύνηθες να περιορίζεται το μέγεθος των συνόλων γονιδίων μεταξύ 10 και 200 γονιδίων.

Πως θα αντιστοιχίσουμε SNPs σε γονίδια?

- Ο πιο απλός τρόπος για να γίνει αυτό βασίζεται στην τοποθεσία των SNP. Για παράδειγμα, μια κοινή μέθοδος είναι να τοποθετήσουμε ένα SNP σε ένα γονίδιο εάν το SNP βρίσκεται εντός των ορίων του γονιδίου ή μέσα σε ένα σταθερό παράθυρο γύρω από αυτό
- Η ανισορροπία σύνδεσης (LD) μπορεί επίσης να χρησιμοποιηθεί για τη χαρτογράφηση των SNP στα γονίδια. Σε αυτή την περίπτωση, τα SNPs χαρτογραφούνται σε ένα γονίδιο εάν συσχετίζονται με άλλα SNP που βρίσκονται εντός των ορίων των γονιδίων. Οι μέθοδοι που χρησιμοποιούν το LD για τη χαρτογράφηση των SNP στα γονίδια έχουν το πλεονέκτημα ότι δεν χάνουν όλες τις πληροφορίες που περιέχονται στα διαγονιδιακά SNP (proxyGeneLD και INRICH)

Είναι σωστό να χρησιμοποιήσουμε imputed genotypes?

- Η χρήση imputed γονοτύπων αυξάνει τον αριθμό των SNP που περιλαμβάνονται στην ανάλυση και επομένως μπορεί να αυξήσει τον αριθμό των γονιδίων που αντιπροσωπεύονται. Για μεθόδους GSA που χρησιμοποιούν γονότυπους, αντί για συνοπτικές στατιστικές, ο αυξημένος αριθμός SNPs μπορεί να αυξήσει τον υπολογιστικό φόρτο.
- Επιπλέον, δεδομένου ότι οι γονότυποι που λείπουν συμπληρώνονται χρησιμοποιώντας τις πληροφορίες από πολλαπλά γειτονικά SNP (απλότυποι), οι μέθοδοι που χρησιμοποιούν τους γονότυπους πολλαπλών SNP για τη μοντελοποίηση επιδράσεων σε επίπεδο γονιδίου ή σε επίπεδο γονιδίου δεν θα ωφεληθούν από την προσθήκη τέτοιων SNP.
- Οι μέθοδοι GSA που χρησιμοποιούν ένα μόνο SNP για να συνοψίσουν ένα γονίδιο (π.χ. εκχώρηση της ελάχιστης τιμής P όλων των SNP που έχουν εκχωρηθεί στο γονίδιο) ενδέχεται να ωφεληθούν από την πρόβλεψη/συμπλήρωση γονότυπου, καθώς ορισμένες τιμές p σε επίπεδο γονιδίου μπορεί να γίνουν πιο σημαντικές

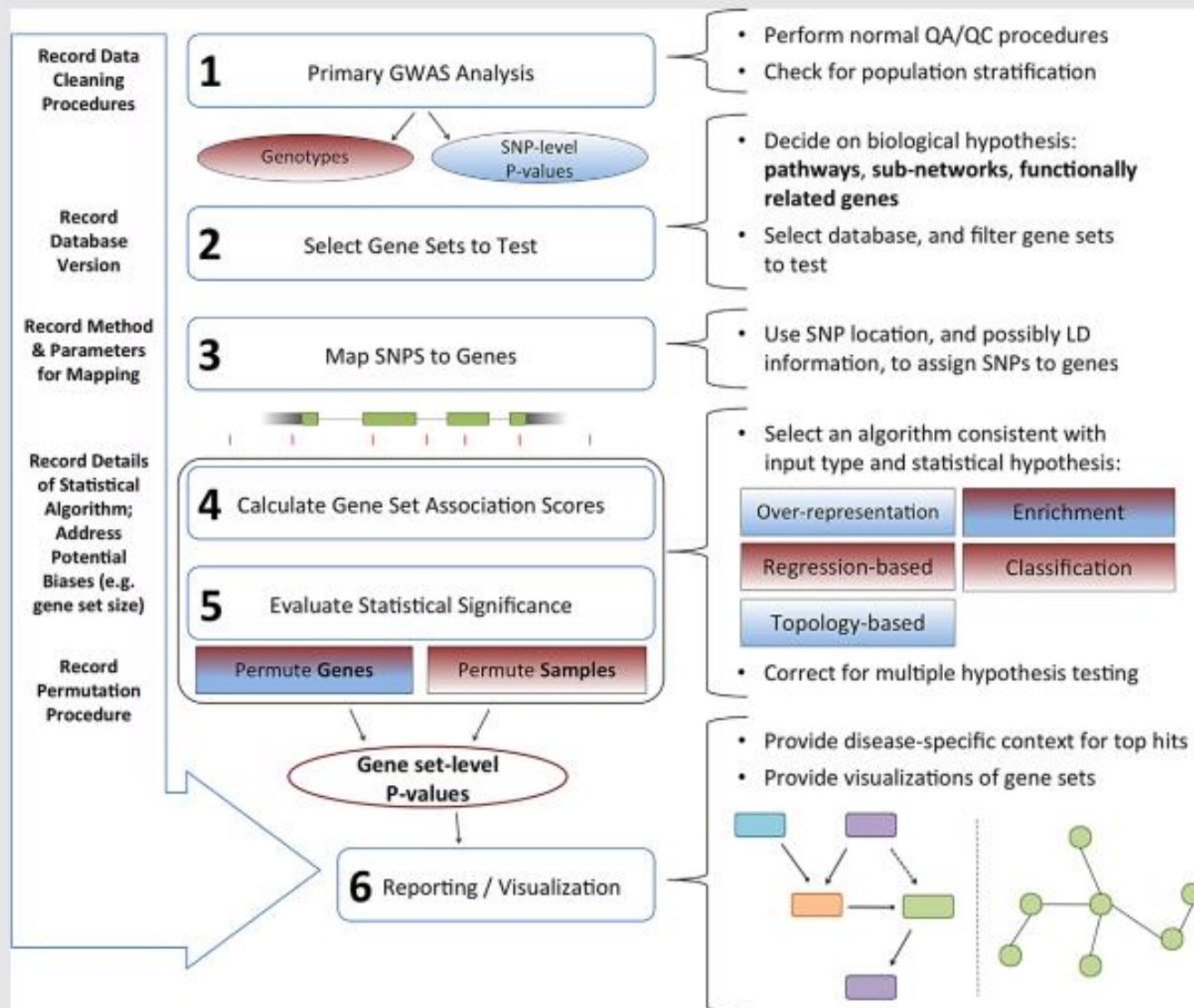


FIG. 1. A gene set analysis workflow, including the possible choices that must be made at each step of the analysis. The data type requirements for the various statistical methods are indicated by color. For instance, regression-based methods and permutation procedures that randomize samples require genotypes as inputs. On the other hand, over-representation methods utilize summary statistics. Some methods (multicolored) are not restricted to one type of input data.

Εργαλεία λογισμικού

Table 1. Gene set analysis methods and software tools (focused on GWAS data)

Method	Test type ^a	Input	Refs
Gamma method	2s, SC	Genotypes	[38]
Adaptive rank truncated product method	2s, SC	Genotypes or <i>P</i> values	[39]
VEGAS	2 s ^b , SC	<i>P</i> values	[40]
BGSAsnp	2s, SC	Genotypes	[41]
GRASS	2s, SC	Genotypes	[42]
HYST	2s, SC	<i>P</i> values, <i>R</i> ² coefficient	[61,62]
PRP	2s, SC, Top	<i>P</i> values	[63]
GSA-SNP	2s, C, SC	<i>P</i> values	[64]
gamGWAS	2s, C	<i>P</i> values	[58]
Pathway-PDT (family-based tests)	2s, C	Genotypes	[65]
WebGestalt	2s, C	Genes	[66]
i-GSEA4GWAS	2s, C	<i>P</i> values	[67]
ICSNPathway	2s, C	<i>P</i> values	[68]
GenGen	2s, C	Genotypes	[69]
ALIGATOR	2s, C	<i>P</i> values	[57]
GeSBAP	2s, C	<i>P</i> values	[70]
GSEA	2s, C	Genotypes	[71]
GSEA-P	2s, C	Ranked gene list	[11]
MAGENTA	2s, C	<i>P</i> values	[72]
Modified Fisher's method	1s, 2s, SC	<i>P</i> values	[36]
Modified Fisher's method	1s, 2s, SC	<i>P</i> values	[37]
Plink set-based test	1s, SC	Genotypes	[73]
GSEA-SNP	1s, SC	<i>P</i> values	[74]
GLOSSI	1s, SC	<i>P</i> values	[75]
Supervised PC	1s, SC	Genotypes	[43]
SNP ratio test	1s, SC	Genotypes	[76]
PAGWAS (hierarchical model)	1s, SC	Genotypes	[44]
GLMM	1s, SC	<i>P</i> values	[45]
BMApathway	1s, SC	Genotypes	[46]
Synthetic Feature Random Forest	1s	Genotypes	[47]
SGL-BCGD	1s, SC	Genotypes	[48]
INRICH	1s, C	Genomic regions	[77]
GREAT	C	Genomic regions	[78]
SSEA	1s, SC	<i>P</i> values	[79]
PARIS	1s, C	<i>P</i> values	[54]
Gowinda	1s, C	SNPs	[80]
GeneSetDB (hypergeometric)	C	<i>P</i> values	[81]
Gene set bagging	C	Genotypes	[82]
PoDA	1s, C	Genotypes	[83]
ProxyGeneLD	NA		[56]
GWAS Pathway Identifier	NA		[84]
ConsensusPathDB	NA		[13]
Markov Random Field Model	Top	Genes	[85]
EnrichNet	Top	Genes	[86]
TopoGSA	Top	Genes	[87]
CePa	Top	Genes	[88,89]
THINK-Back-DS	C, Top	<i>P</i> values, genes	[90]
PINA v2.0	C, Top	Genes	[29]
PANOGA	Top	<i>P</i> values	[30]
dmGWAS	Top	<i>P</i> values	[31]
HotNet2	Top	<i>P</i> values	[32]
Metacore	Top	Genes	(http://www.thomsonreuters.com/metacore)
DAPPLE	Top	Genes or genomic regions	[91]
GRAIL	Top, TM	SNPs or genomic regions	[92]
Chilibot	Top, TM	Genes	[93]

^aAbbreviations: 1S, one-step; 2S, two-step; C, competitive; SC, self-contained; Top, topology-based; TM, text mining.

^bVEGAS is a gene-based test and can be used as the first step in a two-step GSA.

Mooney MA, Nigg JT, McWeeney SK, Wilmot B. Functional and genomic context in pathway analysis of GWAS data. Trends in Genetics. 2014 Sep 1;30(9):390-400.

Table 1. Examples of pathway analysis tools in each generation.

Name	Availability	Reference
ORA tools		
Onto-Express	Web (http://vortex.cs.wayne.edu)	[4,5]
GenMAPP	Standalone (http://www.genmapp.org)	[11,71]
GoMiner	Standalone, Web (http://discover.nci.nih.gov/gominer)	[72,73]
FatiGO	Web (http://babelomics.bioinfo.cipf.es)	[74]
GOstat	Web (http://gostat.wehi.edu.au)	[7]
FuncAssociate	Web (http://llama.mshri.on.ca/funcassociate/)	[6]
GOToolBox	Web (http://genome.crg.es/GOToolBox/)	[10]
GeneMerge	Standalone, Web (http://genemerge.cbc.umd.edu/)	[9]
GOEAST	Web (http://omicslab.genetics.ac.cn/GOEAST/)	[75]
ClueGO	Standalone (http://www.ici.upmc.fr/cluego/)	[76]
FunSpec	Web (http://funspec.med.utoronto.ca/)	[77]
GARBAN	Web	[78]
GO:TermFinder	Standalone (http://search.cpan.org/dist/GO-TermFinder/)	[8]
WebGestalt	Web (http://bioinfo.vanderbilt.edu/webgestalt/)	[79]
agriGO	Web (http://bioinfo.cau.edu.cn/agriGO/)	[80]
GOFFA	Standalone, Web (http://edkb.fda.gov/webstart/arraytrack/)	[81]
WEGO	Web (http://wego.genomics.org.cn/cgi-bin/wego/index.pl)	[82]
FCS tools		
GSEA	Standalone (http://www.broadinstitute.org/gsea/)	[21,29]
sigPathway	Standalone (BioConductor)	[22]
Category	Standalone (BioConductor)	[24]
SAFE	Standalone (BioConductor)	[30]
GlobalTest	Standalone (BioConductor)	[15]
PCOT2	Standalone (BioConductor)	[17]
SAM-GS	Standalone (http://www.ualberta.ca/~yyasui/software.html)	[83]
Catmap	Standalone (http://bioinfo.thep.lu.se/catmap.html)	[84]
T-profiler	Web (http://www.t-profiler.org)	[85]
FunCluster	Standalone (http://corneliu.henegar.info/FunCluster.htm)	[86]
GeneTrail	Web (http://genetrail.bioinf.uni-sb.de)	[87]
GAzer	Web	[88]
PT-based tools		
ScorePAGE	No implementation available	[37]
Pathway-Express	Web (http://vortex.cs.wayne.edu)	[38,39]
SPIA	Standalone (BioConductor)	[40]
NetGSA	No implementation available	[43]

Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS computational biology. 2012 Feb 23;8(2):e1002375.

TABLE II. Gene Set Analysis Methods and Software Tools

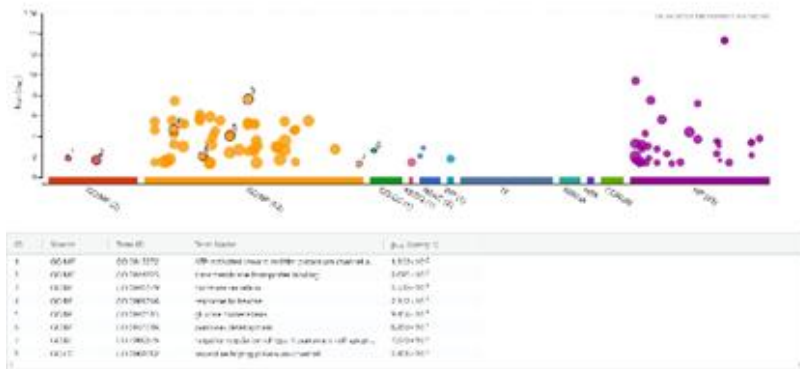
Method	Input	References
Over-representation methods		
WebGestalt	Gene list	Wang et al., [2013]
DAVID	Gene list	Huang et al. [2009]
Metacore	Gene list	thomsonreuters.com/metacore
GeneSetDB	Gene list	[Araki et al. [2012]
INRICH	Genomic Regions	Lee et al. [2012]
MAGENTA	<i>P</i> -values	Segrè et al. [2010]
ALIGATOR	<i>P</i> -values	Holmans et al. [2009]
Enrichment methods		
GSEA	Genotypes	Wang et al. [2007]
i-GSEA4GWAS	<i>P</i> -values	Zhang et al. [2010]
GSA-SNP	<i>P</i> -values	Nam et al. [2010]
GSEA-P	Ranked gene list	Subramanian et al. [2007]
GSEA-SNP	<i>P</i> -values	Holden et al. [2008]
Methods for combining <i>P</i> -values		
Modified fisher's method	<i>P</i> -values	De la Cruz et al. [2010]
Modified fisher's method	<i>P</i> -values	Luo et al. [2010]
Adaptive rank truncated product method	Genotypes or <i>P</i> -values	Yu et al. [2009]
FORGE	<i>P</i> -values	Pedroso et al. [2012]
Plink set-based test	Genotypes	Purcell et al., [2007]
Regression-based methods		
GRASS	Genotypes	Chen LS et al. [2010]
MAGMA	Genotypes	de Leeuw et al. [2015]
PCgamma	Genotypes	Biernacka et al. [2012]
PAGWAS	Genotypes	Evangelou et al. [2014]
SGL-BCGD	Genotypes	Silver et al. [2013]
Supervised PCA	Genotypes	Chen X et al. [2010]
Classification-type methods		
Pathways of distinction analysis	Genotypes	Bruan and Buetow, [2011]
Synthetic feature random forest	Genotypes	Pan et al. [2014]
Methods that incorporate interaction data (networks)		
PANOGA	<i>P</i> -values	Bakir-Gungor and Sezerman, [2011]
dmGWAS	<i>P</i> -values	Jia et al. [2011]
HotNet2	<i>P</i> -values	Vandin et al. [2011]
EnrichNet	Gene list	Glaab et al. [2012]
NetPEA	Gene list	Liu and Ruan, [2013]
NEA	Gene list	Alexeyenko et al. [2012]
PANOGA	<i>P</i> -values	Bakir-Gungor et al. [2014]
PINA	Gene list	Cowley et al. [2012]
Dapple	Gene list	Rossin et al. [2011]
DEPICT	<i>P</i> -values	Pers et al. [2015]

Mooney MA, Wilmot B. 2015. Gene set analysis: A step-by-step guide. Am J Med Genet Part B 168B:517–527

- Εργαλεία όπως τα **Enrichr**, **g:Profiler**, **DAVID**, **WebGestalt** και **PANTHER** είναι εργαλεία εμπλουτισμού γενικής χρήσης που παρέχουν λειτουργίες για διαφορετικούς τύπους δεδομένων omics. Δέχονται γονίδιο ή λίστα SNP ως είσοδο και παρέχουν API που διασφαλίζει τη διαλειτουργικότητα, ενώ για τη στατιστική ανάλυση χρησιμοποιούν όλα κάποια έκδοση του ORA ή/και του GSEA (το **WebGestalt** χρησιμοποιεί επίσης ανάλυση βάσει τοπολογίας δικτύου). Ένα σημαντικό χαρακτηριστικό αυτών των εργαλείων είναι ότι ενσωματώνουν μεγάλο αριθμό βάσεων δεδομένων βιολογικών και μονοπατιών, με τα **g:Profiler** και **Enrichr** να προσφέρουν την πληρέστερη συλλογή.
- Ανάμεσα στα πιο εύκολα χρησιμοποιούμενα και συχνά αναφερόμενα είναι τα εργαλεία που χρησιμοποιούν σεβερ. Το **FUMA** και το **iGSE4GWAS** είναι εργαλεία εξειδικευμένα σε GWAS και χρησιμοποιούν στατιστικές σε επίπεδο SNP ως είσοδο, αλλά διαφέρουν στις επόμενες αναλύσεις: Το **FUMA** χρησιμοποιεί το **MAGMA** για δοκιμές βάσει γονιδίων και παρέχει τη δοκιμή ORA και Kologorov-Smirnov (GSEA), ενώ το **iGSE4GWAS** αντιστοιχίζει το SNP με το πιο σημαντικό p-value στο γονίδιο του, και πραγματοποιεί GSEA με permutation.
- Το **GSA-SNP2** είναι μία από τις πρώτες μεθόδους που αναπτύχθηκε για GWAS και έχει δει αρκετές βελτιώσεις όσον αφορά τον υπολογισμό της συνδυασμένης βαθμολογίας γονιδίου και του χρόνου εκτέλεσης, καθώς είναι από τις ταχύτερες μεθόδους.

- Άλλες μέθοδοι χρησιμοποιούν διαφορετικούς ορισμούς των συνόλων γονιδίων, σε ορισμένες περιπτώσεις χρησιμοποιώντας πρόσθετες πληροφορίες. Για παράδειγμα, το **dmGWAS** ενσωματώνει δίκτυα PPI και χρησιμοποιεί μια μέθοδο αναζήτησης για την αναγνώριση υποδικτύων. Σε σύγκριση με τις τυπικές μεθόδους μονοπατιού προσφέρει στους χρήστες την ευελιξία στον ορισμό ενός συνόλου γονιδίων και μπορεί να χρησιμοποιήσει τοπικές πληροφορίες PPI.
- Το **GEMB** καθορίζει τα σύνολα γονιδίων χρησιμοποιώντας βάρη από προβλέψεις μοντέλων και κατάταξη γονιδίων από το GWAS και το **GENOMICper** χρησιμοποιεί μεταθέσεις των αναγνωρισμένων SNPs με περιστροφή σε σχέση με τις γονιδιωματικές θέσεις.
- Το **GWAB** χρησιμοποιεί συνδέσεις δικτύου για να επαναπροσδιορίσει τα υποψήφια γονίδια ενσωματώνοντας στην GWAS και τα δεδομένα δικτύου, ενώ το **GenToS** αναζητά πολυμορφισμούς που σχετίζονται με χαρακτηριστικά σε υπάρχοντα ανθρώπινα GWAS.

Kontou PI, Bagos PG. The goldmine of GWAS summary statistics: a systematic review of methods and tools. BioData Mining. 2024 Sep 5;17(1):31.

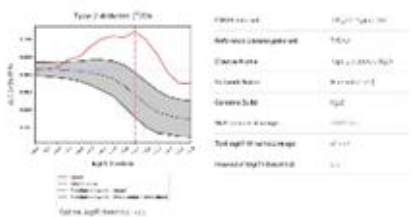


variants: **rs1044397, rs1044398, rs1044399, rs1044400, rs1044401, rs1044402, rs1044403, rs1044404, rs1044405, rs1044406**
 date: **2024-09-05 10:00:00**
 no genes: **1**

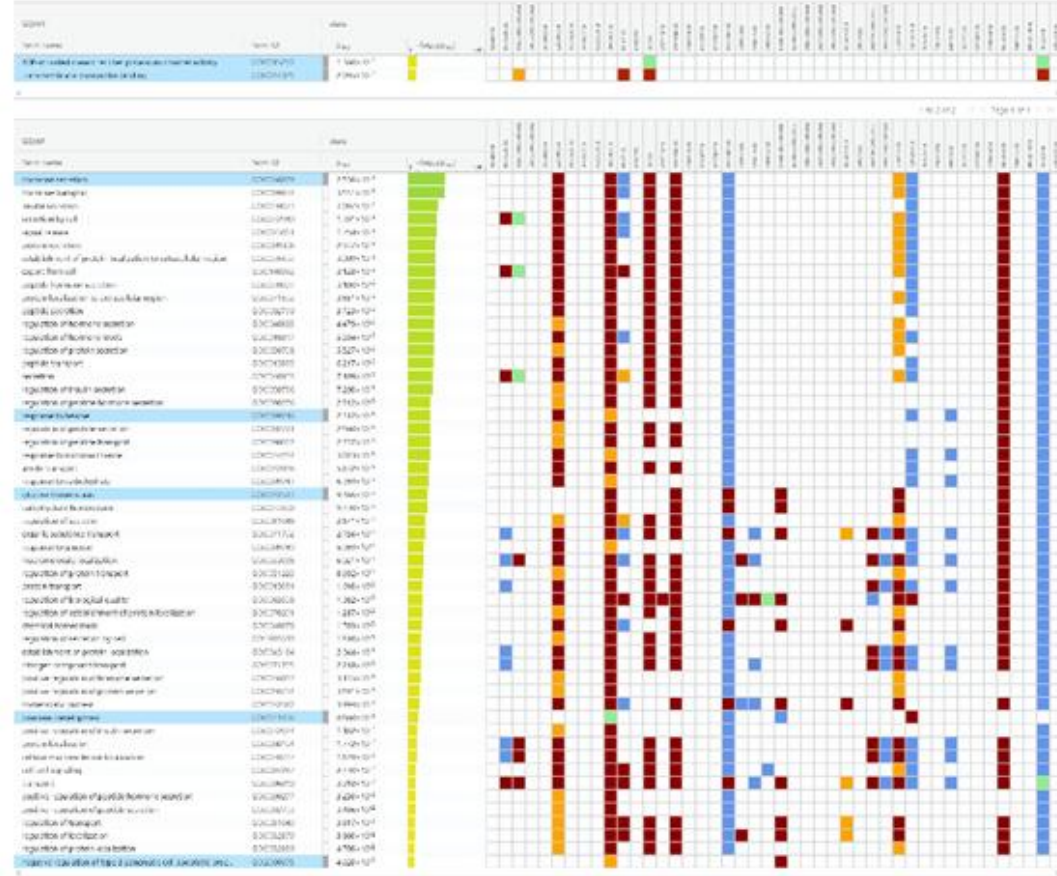
A



B



C



D

Kontou PI, Bagos PG. The goldmine of GWAS summary statistics: a systematic review of methods and tools. BioData Mining. 2024 Sep 5;17(1):31.

aSPUpath2	https://github.com/ChongWu-Biostat/aSPUpath2	https://pubmed.ncbi.nlm.nih.gov/29411426/	GSA	R	Integrates gene expression reference weights, GWAS summary data, LD information, and candidate pathways to identify pathways whose expression is associated with complex traits
GIGSEA	www.github.com/zhushijia/GIGSEA	https://pubmed.ncbi.nlm.nih.gov/30010968	GSA	R	Uses GWAS and eQTL to infer differential gene expression and interrogate gene set enrichment for the trait-associated SNPs. By incorporating expression data it naturally accounts for factors such as gene size, gene boundary, SNP distal regulation and multiple-marker regulation
GAUSS	https://github.com/diptavo/GAUSS	https://pubmed.ncbi.nlm.nih.gov/33730541	GSA	R	Tests for any self-contained association between a phenotype and a gene-set and produces a p-value for the association.
GSA-SNP2	https://sourceforge.net/projects/gsasnp2	https://pubmed.ncbi.nlm.nih.gov/29562348	GSA	C/C++	A method for pathway enrichment analysis of GWAS P-value data. It accepts also gene-wise p-values (obtained from other methods) and outputs pathway gene sets 'enriched' with genes associated with the given phenotype
VSEAMS	https://github.com/ollyburren/vseams	https://pubmed.ncbi.nlm.nih.gov/25170024	GSA	R	A non-parametric SNP set enrichment method to test for enrichment of GWAS signals in functionally defined loci using P-values
dmGWAS	https://bioinfo.uth.edu/dmGWAS/	https://www.ncbi.nlm.nih.gov/pubmed/21045073/	GSA	Executable	A dense module searching method to identify candidate subnetworks or genes for complex diseases by integrating PPI network. Extensively searches for subnetworks enriched with low P-value genes.
IGSE4GWAS	http://gse4gwas.psych.ac.cn	https://pubmed.ncbi.nlm.nih.gov/20435672/	GSA	web	Detects pathways associated with traits by applying an improved gene set enrichment analysis. Implements also a follow-up functional analysis for SNPs in trait-associated pathways identified. Uses LD and putative functional annotation from Ensembl, ENCODE and eQTLs
Enrichr	https://maayanlab.cloud/Enrichr	https://pubmed.ncbi.nlm.nih.gov/27141961/	GSA	web	A gene set search engine that enables the querying of hundreds of thousands of annotated gene sets. Enrichr uniquely integrates knowledge from many high-profile projects to provide synthesized information about mammalian genes and gene sets.
SNPratio test	https://sourceforge.net/projects/snpratitest/	https://pubmed.ncbi.nlm.nih.gov/19620097/	GSA	Perl	Compares the proportion of significant to all SNPs within genes that are part of a pathway and computes an empirical P-value based on comparisons to ratios in datasets where the assignment of case/control status has been randomized.
g:Profiler	https://biit.cs.ut.ee/gprofiler/gost	https://pubmed.ncbi.nlm.nih.gov/37144459/	GSA	web/R	Integrates many databases, including Gene Ontology, KEGG and TRANSFAC, to provide a comprehensive and in-depth analysis of gene lists. It also provides interactive and intuitive user interfaces and supports ordered queries and custom statistical backgrounds, among other settings.
DAVID	https://david.ncifcrf.gov/	https://pubmed.ncbi.nlm.nih.gov/35325185/	GSA	web/R	An enrichment tool with functionalities for different types of omics data including GWAS. It accepts gene or SNP-list as input and provide API ensuring interoperability. For analysis it uses ORA and GSEA
WebGestalt	http://www.webgestalt.org/	https://pubmed.ncbi.nlm.nih.gov/31114916/	GSA	web/R	An enrichment tool with functionalities for different types of omics data including GWAS. It accepts gene or SNP-list as input and provide API ensuring interoperability. For analysis it uses ORA, GSEA and Network Topology-based Analysis
PANTHER	http://www.pantherdb.org/	https://pubmed.ncbi.nlm.nih.gov/33290554/	GSA	web	An enrichment tool with functionalities for different types of omics data including GWAS. It accepts gene or SNP-list as input and provide API ensuring interoperability. For analysis it uses ORA and GSEA
deTS	https://cran.r-project.org/web/packages/deTS/index.html	https://pubmed.ncbi.nlm.nih.gov/30824912/	GSA	web/R	Performs tissue-specific enrichment analysis (TSEA) for detecting tissue-specific genes and for enrichment test of different forms of query data.
DESE	https://pmglab.top/pcga	https://pubmed.ncbi.nlm.nih.gov/31694669/	GSA	web	Detects the causal tissues of complex traits according to selective expression of disease-associated genes
PAPA	https://sourceforge.net/projects/papav1/files/	https://pubmed.ncbi.nlm.nih.gov/26568630	GSA	C/C++	A flexible tool for pleiotropic pathway analysis utilizing GWAS summary results
GEMB	https://github.com/cochrana4/GEMB	https://pubmed.ncbi.nlm.nih.gov/33034635	GSA	Matlab	A method that combines gene weights from model predictions and gene ranks from genome-wide association studies into a weighted gene-set test
GENOMICper	https://cran.r-project.org/web/packages/genomicper/index.html	https://pubmed.ncbi.nlm.nih.gov/22973544/	GSA	R	Uses SNP association p-values and permutes them by rotation with respect to the genomic locations. The joint gene p-values are calculated using Fisher's combination test and pathways' association tested using the hypergeometric test
GWAB	https://www.inetbio.org/gwab/	https://pubmed.ncbi.nlm.nih.gov/28449091	GSA	web	Trait-associated genes with sub-threshold significance score can be rescued by network connections to other significant candidates
Inferno	http://inferno.lisanwanglab.org/	https://pubmed.ncbi.nlm.nih.gov/30113658	GSA	web/Python	Method which integrates diverse functional genomics data sources to identify causal noncoding variants. Characterizes the relevant tissue contexts, target genes, and downstream biological processes affected by functional variants. Uses COLOC, WebGestalt, LDSC and MetaXcan.
Mergeomics	http://mergeomics.research.idre.ucla.edu	https://pubmed.ncbi.nlm.nih.gov/34048577	GSA	web/R	Web server which uses summary statistics of multi-omics association studies (GWAS, EWAS, TWAS, PWAS, etc) and performs correction for LD, GSEA, meta-analysis and identification of essential regulators of disease-associated pathways and networks
GenToS	https://github.com/genepi-freiburg/gentos	https://pubmed.ncbi.nlm.nih.gov/27612175	GSA	Java	Calculates an appropriate statistical significance threshold and then searches for trait-associated variants in summary statistics from human GWAS
aSPU	https://cran.r-project.org/web/packages/aSPU/	https://pubmed.ncbi.nlm.nih.gov/27592708	GSA/Gene-Based	R	Performs adaptive gene-based test and pathway-based test for association analysis of multiple traits. The tests are adaptive at both the SNP- and trait-levels, thus maintaining high power across a wide range of situations. The methods can be applied to mixed types of traits, and to Z-statistics or P-values
snpGeneSets	https://www.unc.edu/SoPH/Departments-and-Faculty/Data-Science/Research/Services/Software.html/	https://www.ncbi.nlm.nih.gov/pubmed/27807048/	GSA/Gene-Based	R	Integrates local genomic annotation databases and provides genome-wide annotation for SNP, Gene and gene sets. It aims to support interpretation of GWAS results and performing post-analysis
PascalX	https://github.com/BergmannLab/PascalX	https://pubmed.ncbi.nlm.nih.gov/37137228	GSA/Gene-Based	Python	Provides fast and accurate mapping of SNP-wise GWAS data. It allows for scoring genes and annotated gene sets for enrichment signals based on data from, both, single GWAS and pairs of GWAS
PASCAL	https://www2.unil.ch/cbg/index.php?title=Pascal	https://pubmed.ncbi.nlm.nih.gov/26808494	GSA/Gene-Based	Java	Computes gene and pathway scores from SNP-phenotype associations. For gene score computation, implements analytic and efficient numerical solutions to calculate test statistics. For pathway scoring, it uses a modified Fisher method
MAGMA	https://ctg.cncr.nl/software/magma	https://pubmed.ncbi.nlm.nih.gov/25885710/	GSA/Gene-Based	C/C++	Uses p-values and performs gene-based and gene-set analysis as well as meta-analysis
FUMA	https://fuma.ctglab.nl	https://pubmed.ncbi.nlm.nih.gov/29184056/	GSA/Gene-Based	web	An integrative web-based platform using information from multiple biological resources to facilitate functional annotation of GWAS results, gene prioritization and interactive visualization. It accommodates positional, expression quantitative trait loci (eQTL) and chromatin interaction mappings, and provides gene-based, pathway and tissue enrichment results.
chromMAGMA	https://github.com/lawrenson-lab/chromMAGMA-public	https://pubmed.ncbi.nlm.nih.gov/35777959	GSA/Gene-Based	R	Method to identify candidate risk genes based on the presence of risk variants within noncoding regulatory elements