

Основната част от публикациите ми са свързани с приложения на **статистически и биоинформатични методи към новите технологии за генетично секвениране** (high-throughput sequencing, също известно като Next Generation Sequencing или NGS) **на ДНК (публикации [1, 8 и частично 5]) и РНК (публикации [13, 12, 7, и частично 8]),** и предишното поколение технологии – **ДНК и РНК микрочипове (microarrays) [2, 3, 4 и 6].** Друга тема е свързана с анализа на данни в ситуация, в която имаме **извадка с размер, много по-малък от размерността на пространството на данните,** публикация [9], като трябва да отбележим, че подобни ситуации възникват често при анализа на данни от гореизброените видове. Няколко от **публикациите, [5, 6 и 8],** са свързани с анализ на данни от **индуцирани плурипотентни стволови клетки (iPS cells).** Други **публикации, [10 и 11],** са свързани с **анализ на популации от различни географски региони с различни фенотипни признаци.**

Много от проектите, по които съм работил, използват трудоемки и скъпоструващи процеси на генериране на биологични данни. Поради това статиите, базирани на тях, имат много съавтори, по-голямата част от които са биолози.

Секвенирането е процес на четене и разпознаване на различните нуклеотиди (или неформално, букви) от ДНК или РНК. Разработката на математически, статистически и биоинформатични методи, приложими към данни от NGS технологиите, е една от най-бързо развиващите се научни сфери в света. Докато по-старите технологии (ДНК и РНК микрочипове) произвеждат по-малко по обем данни с по-голям шум, NGS технологиите произвеждат огромни обеми от дискретни данни. Тъй като старите методи за анализ не са приложими към данните от NGS, през последните 10 години след излизането на пазара на първите NGS машини научните работници се интересуват от методи за анализ на данни, приложими към основните биологични въпроси, на които може да се отговори с изследвания с NGS, използвайки ДНК или РНК секвениране.

Научната ми работа в САЩ в периода от 2006 до 2011 беше определена от рамките на научно-изследователски проекти, посочени в Приложение 12. В момента работата ми по тази тематика е тясно свързана с научно-изследователския проект ДФНИ И02/19.

При експерименти с използване на **данни от секвениране на РНК,** една от основните задачи е анализ на разликите в генната експресия (gene expression) между две групи, най-често между здрави хора и пациенти. Целта е да се открият гени, които имат различна генна експресия между двете групи, като тези гени са кандидати за последващ биомедицински анализ. Някои от най-често използваните статистически критерии, например DESeq или edgeR, използват отрицателно биномни (NB) модели, за да моделират връзката между средната стойност и дисперсията на нормализираните бройки фрагменти от РНК, които са секвенирани от всеки ген. В **публикация [13]** разглеждаме ситуации, в които статистическата грешка от I род е трудна за оценяване и в които често се използват консервативни статистически критерии. Те надценяват вероятностните

стойности, което води до загуба на статистическа мощност и до оскъпяване на експериментите. За подобряване качествата на критериите, в тази статия е представена изчислително ефективна техника за калибриране на вероятностните стойности, която съществено увеличава статистическата мощност и намалява големината на извадката. Методът е приложен към горепосочения метод DESeq. Тази публикация бе избрана за научно-приложно постижение на ИМИ за 2014 г.

В много от случаите на анализ на биологични данни, свързани със секвениране или микрочипове, имаме размерност, която е много по-голяма от броя на наблюденията. Макар работата ми по **публикация [9]** да е започнала преди интереса ми към новите технологии за производство на биомедицински данни, резултатите могат да се приложат и към тези данни. В тази публикация представяме статистически критерий за идентифициране на клъстери в многомерно пространство, базиран на метода k -means. Първо показваме, че в този случай съществуващите критерии, които използват проекциите върху права или равнина, биха могли да ни заблудят. Например, при нарастване на размерността тези критерии могат да намерят добре разделени клъстери, дори когато тези клъстери не съществуват. Освен това показваме, че дори ако данните представляват два клъстера, увеличаването на размерността прави намирането им по-трудно. За да решим този проблем, първо намираме асимптотично уравнение за вероятността на опашката на максимума на много независими χ^2 случайни величини. Извеждаме неравенство за асимптотичното разпределение на максимума на независими и положително корелирани χ^2 случайни величини, резултат, подобен на този на Slepian. След това предлагаме консервативен тест за това дали данните идват от едно многомерно нормално разпределение срещу микс от две многомерни нормални разпределения. Тъй като тестът е прекалено консервативен, ние предлагаме алтернативен, практически тест със същите асимптотични свойства.

Публикации [12, 7 и частично 8] са свързани с **биоинформатичен и статистически анализ на данни от РНК секвениране**. В **публикация [12]** с използване на NGS технологиите намираме общо 1131 гени, в които има разлика между генната експресия при пациенти от Синдрома на Турет (TS) и контролна група от здрави хора. Затова изследваме данни от РНК секвениране на 2 региона на мозъка при 9 пациенти и 9 контролни субекта със сходни характеристики (пол, възраст и др.). След биоинформатичен анализ получаваме нивата на генна експресия на около 36000 гени или транскрипти. Прилагайки метода edgeR, намираме 309 гени, в които пациентите имат статистически значими по-ниски нива на генна експресия от контролите и 822 гени, в които пациентите имат статистически значими по-високи нива. Освен това показваме, че гените от първата категория включват повече гени, свързани с интерневроните на стриатума, отколкото случайна извадка, а гените от втората категория включват повече гени, свързани с имунната система, отколкото случайна извадка. В **публикация [7]** се интересуваме от промените в генната експресия, свързани с възрастта. Затова използваме данни от РНК секвениране на префронталния кортекс, свързан с висшите когнитивни процеси – памет, мислене, и изпълнителните функции; както и два кортикални региона на мозъка, свързани с говора. Използвайки метода DESeq, както и с помощта на статистически методи за намиране на мрежи от гени със свързани нива на експресия (coexpression networks), показваме, че най-големите промени в генната експресия, свързани с възрастта, са в клетъчните сигнални системи и като резултат от това – в преноса на нервни импулси.

При експерименти със **секвениране на ДНК**, **публикации [1, 8 и частично 5]**, се интересуваме от така наречените структурни варианти (Structural Variants или SV) или варианти с различен брой копия (Copy Number Variants или CNV). Те могат да бъдат региони от ДНК, които липсват в изследвания човек (deletions, делеции) или по-общо да са с различен брой копия в сравнение с т.нар. референтен геном. **Публикация [1]** е една от първите в света по тази тема и в нея намираме 1297 структурни варианта в два индивида, използвайки paired-end ДНК секвениране, при което двата края на относително дълги ДНК сегменти се секвенират и се намират координатите им в референтния геном (процесът се нарича alignment или mapping). Големи разлики между очакваните координати и тези, намерени в референтния геном, са признак на вероятен структурен вариант. Точните области на отхвърляне се определят непараметрично, като се вземат горният и долният 0.00135 квантил на разстоянията между двата края на сегментите; освен това използването на параметричен метод, моделиране на логаритъма като микс от нормални, дава много близки резултати.

Друга модерна област на научни изследвания са индуцираните плурипотентни стволови клетки (iPS cells), **публикации [8, 6 и 5]**. Това са стволови клетки, които имат предимството, че могат да се получат от някакъв друг вид клетки, и също така могат да се диференцират (т.е. превърнат) в различен вид клетки. В **публикация [8]** разглеждаме въпрос, свързан с две модерни теми на научни изследвания (iPS клетки и секвениране) – дали в процеса на репрограмизирането не могат да се появят т.нар. di novo структурни варианти, т.е. такива, които не са съществували в началните клетки, но съществуват в iPS клетките. Изследвайки данни от ДНК секвениране на iPS клетки, получени от фибробласт от кожата на 7 човека, показваме, че в този случай всички варианти в iPS клетките или съществуват в оригиналния фибробласт, или в родителите на съответния индивид, т.е. не се наблюдават di novo варианти. Освен това използвахме данни от допълнително РНК секвениране, за да покажем, че iPS клетките са по-близо от гледна точка на генна експресия до стандартна линия от индуцирани плурипотентни стволови клетки, отколкото до оригиналните фибробластни клетки, т.е. процесът на репрограмизирането работи.

В **публикация [6]** разглеждаме друг въпрос, свързан с iPS клетките, и показваме, че такива клетки, отгледани по специален начин, имат профил на генна експресия, подобен на този на ембрионния телянцелон, но не и на други региони на централната нервна система. За изследването използваме РНК микрочипове. Резултатите показват, че най-голяма близост се наблюдава със стените на ранния церебрален кортекс. Заключение е, че този модел може да се използва за изследване на развитието на човешкия мозък и на болести, свързани с церебралния кортекс.

Въпросите, свързани с използването на iPS клетките за изследване на невро-психиатрични болести, са дискутирани в обзорната **публикация [5]**, като се обръща внимание на проблемите по намирането на генетични варианти (SV/CNV) с новите технологии за генетично секвениране (high-throughput sequencing, или Next Generation Sequencing).

Публикации [2, 3 и 4] използват данни от микрочипове (microarrays). В **публикация [4]** с помощта на метода за намаляне на размерността elastic net, приложен върху данни от SNP микрочипове, идентифицираме 55 гени, които вероятно са асоциирани със

способността за четене. След това показваме, че пропорцията на гени от тези 55, чиито промотери са свързани с гама-аминомаслената киселина (γ -Aminobutyric acid, GABA) или neurotrophic tyrosine receptor kinase (NTRK), е съществено по-голяма от тази, която бихме наблюдавали в случайна извадка.

В **публикация [3]** разглеждаме ефекта на EBV nuclear antigen 1 (EBNA1) върху генната експресия като анализираме данните от експерименти с РНК микрочипове на две клетъчни линии (BJAB и 293), трансфектирани (transfected) с EBNA1. Като резултат, за всяка от двете клетъчни линии получаваме съответно 16 и 13 гени, които са със статистически значими по-високи средни нива на експресия в присъствието на EBNA1, отколкото без него, и съответно 13 и 6 гени с по-ниски нива. Тези гени са свързани с растежа и миграцията на клетките, имунната система, стабилността на ДНК и т.н. Освен това, анализирайки данни от ChIP-chip микрочипове с метода Weeder, намираме определени мотиви (motifs) от нуклеотиди, към които е статистически по-вероятно да се закрепят EBNA1.

При биологични експерименти често се разполага с много ограничени количества биологичен материал. В **публикация [2]** предлагаме нов метод за амплификация (също намножаване, англ. amplification) на генетичен материал (т.е. на произвеждане на по-големи количества биологичен материал, започвайки с малко начално количество). Чрез анализ на данни от два вида микрочипове на началния и на амплифицирания материал показваме, че предложеният метод е специфичен (т.е. не се въвеждат нови фрагменти от генетичен материал) и без големи отклонения от формата на сигнала (unbiased).

В **публикация [10]** правим първоначален фенотипен анализ на извадка от населението в географски район (наречен AZ), предполагаемо характеризиращ се с по-голяма честота на нарушения на речта, започнали в ранния стадий на развитието, отколкото в контролна популация (CP) от близък географски регион. Прилагайки различни статистически критерии, показваме, че честотата на тези заболявания е по-голяма в AZ, отколкото в CP; а също така, че децата от контролния регион се представят по-добре при изпълнението на задачи, свързани с речта, отколкото тези от AZ.

В **публикация [11]** анализираме генетичното разнообразие на 52 сорта пшеница от западния и североизточния черноморски региони. Това е важен проблем, свързан със създаването на нови подобрени сортове. За анализа използваме генетични данни от 263 алели (различни възможни състояния на локусите), избрани от 25 микросателитни локуса (неформално, региони на ДНК). С помощта на стандартни метрики за измерване на генетичното разнообразие показваме, че в североизточния черноморски регион има по-голямо генетично разнообразие, отколкото в западния. Клъстерен анализ показва, че сортовете от западния регион са генетично по-близки помежду си, отколкото до тези от североизточния регион. Също така показваме, че клъстерирането на сортовете от североизточния регион според генетичната им информация до голяма степен съвпада с това по географски подрегиони. Сравнението с вече публикувани сходни изследвания показва, че сортовете от черноморския регион имат по-голямо генетично разнообразие в сравнение с тези от други региони по света.