En oversikt over design i klassisk genetisk epidemiologi

Per Magnus[1] og Rolv T. Lie[2]

1. Divisjon for epidemiologi, Nasjonalt folkehelseinstitutt, Postboks 4404 Nydalen, 0403 Oslo
2. Seksjon for medisinsk statistikk, Institutt for samfunnsmedisinske fag, Universitetet i Bergen

SAMMENDRAG

Å velge design i epidemiologi er et kompromiss mellom faglige problemstillinger og ressurser. Et sentralt poeng er å unngå skjevheter og confounding. Videre er det nødvendig med tilstrekkelig styrke til å oppdage viktige sammenhenger. I denne artikkelen går vi gjennom noen alminnelige problemstillinger i feltet, og beskriver noen design som har vært benyttet for å besvare spørsmålene. I fremtiden vil vi, antakelig med lite utgifter, kunne gjøre en detaljert kartlegging av den enkelte persons genom, og vi vil kunne beskrive genekspresjon og proteinaktivitet på en mye grundigere måte enn i dag. Dette stiller epidemiologer overfor nye utfordringer som vil påvirke valg av problemstillinger, valg av design og valg av dataanalyse.

HVA ER ET DESIGN?

Et design er den hovedtilnærming til datainnsamling som anvendes i et forskningsprosjekt. Man deler gjerne inn design etter tre hovedlinjer: Den ene dimensjonen er om prosjektet er eksperimentelt eller observerende, den andre er om designet er prospektivt (eksponeringsinformasjonen er tilgjengelig før endepunktene måles), retrospektivt (endepunktene kjent før eksponeringene måles) eller av tverrsnittsnatur, og det tredje er om det er et familiedesign eller ikke, altså om man trekker familier eller enkeltpersoner til sitt utvalg.

Eksperimentelle design benyttes ikke i genetisk epidemiologi med unntak av kliniske forsøk i genterapi. Et familiedesign innebærer at man har innhentet data på flere personer fra samme familie og at familiene gjerne benyttes som analyseenheter i studien. Det er imidlertid ikke alltid nødvendig å ha familiedata i genetisk epidemiologi. Fordi genene er på plass i genomet gjennom hele livet er en del av feilkildene ved retrospektive design (case-kontroll design) lite viktige i genetisk epidemiologi der genotyping er hovedkilden til informasjon om genene. Prospektive design (kohortstudier) er særlig viktig når man inkluderer miljøeksponeringer og når man er interessert i å inkludere genekspresjon (RNA-målinger) og proteinmålinger.

Det finnes knapt noen god oversiktsbok med alternative studietyper innenfor genetisk epidemiologi. Fortsatt er antagelig Fundamentals of Genetic Epidemiology av Khoury, Cohen og Beaty den eneste introduksjonsboka skrevet av og for epidemiologer (1).

LITT FORMALGENETIKK

Et *gen* er en funksjonell enhet av DNA-molekylet. Det området på et *kromosom* som okkuperes av et gen kalles et *locus*. Alternative gener på et locus kalles *alleler*. Hver person har 22 par autosomale kromosomer og et par kjønnskromosomer (til sammen 46 kromosomer). Et medlem fra hvert kromosompar kommer fra mor og et fra far. Disse to kromosomene i et par kalles *homologe* kromosomer. Dermed har en person to utgaver av et gen på loci på de autosomale kromosomene (og på X-kromosomene for jenter). Hvis disse genene for et locus er like er personen *homozygot*, og hvis ikke er han eller hun *heterozygot*. Beskrivelsen av allelene på et locus kalles personens *genotype* for det angjeldende locus. Men man snakker også om en persons genotype i bredere forstand, og kan da mene gener fra mange eller alle loci samtidig. Når man teller opp gener for en gruppe mennesker vil det være dobbelt så mange alleler for hvert locus som det er personer. En *genfrekvens* for et utvalg eller en populasjon er derfor antallet utgaver av det allel man er interessert i dividert på to ganger antallet personer i utvalget.

Når man snakker om *kandidatgener* menes gener som man helt eller delvis kjenner funksjonen til. I et forskningsprosjekt som inkluderer gener bygger man gjerne opp en biologisk hypotese, og kandidatgener vil være naturlig å inkludere hvis man har en tankemodell for årsaker eller patogenese av en sykdom der genproduktet inngår som bidragende aller avgjørende element.

I motsetning til kandidatgener har de såkalte *markørgener* oftest ingen kjent funksjon. At et markørlocus er polymorft (polymorfisme) betyr bare at det finnes flere genvarianter (alleler). For å kalle et locus en polymorfisme har man ved konvensjon sagt at genfrekvensen av det hyppigst forekommende allel ikke må være høyere enn 0,99. Man bruker ofte såkalte mikrosatelitter, som er markørgener der allelene skiller seg fra hverandre i antallet gjentagelser av samme korte basesekvens. I det siste har man begynt å bruke SNPs (single nucleotide polymorphisms) der forskjellen mellom allelene består av en enkelt base.

I kjønnscelledannelsen (*meiosen*) reduseres antallet kromosomer fra 46 i den diploide celle til 23 i den *haploide* kjønnscelle. Ett kromosom fra hvert kromosompar trekkes tilfeldig til hver kjønnscelle (gamet). Gener som ligger på ulike kromosomer eller langt fra hverandre på samme kromosom vil gå til de haploide kjønnscellene uavhengig av hverandre. Det vil si at

hvis man vet at et visst allel på et gitt locus finnes i en haploid kjønnscelle, så kan ikke denne kunnskapen predikere noe om hvilket av de to allelene på et *annet* locus som vil ha havnet i samme kjønnscelle. Sannsynligheten for at et hvilket som helst av allelene på det andre locus skal havne i samme kjønnscelle er i utgangspunktet 50%. Grunnen til at gener på loci som ligger langt fra hverandre på samme kromosom kan nedarves uavhengig av hverandre er et biologisk fenomen som heter *crossing-over*. Men hvis de to loci ligger nær hverandre på samme kromosom, slik som de ulike loci i HLA-systemet, så vil det være sterk avhengighet i segregasjonen. De to allelene (fra hvert sitt locus) som ligger langs samme kromosom (er i *coupling*fase) vil følge hverandre til samme kjønnscelle, mens hvis de to allelene ligger på hvert sitt (er i *repulsion*fase) av de homologe kromosomene vil de gå til hver sin kjønnscelle.

PROBLEMSTILLINGER

Sykdommer som ikke har en kjent hovedårsak (for eksempel en mikrobe, et giftstoff eller et gen) kalles for tiden for komplekse sykdommer. Dette er sykdommer der man antar at både miljøfaktorer og gener virker sammen for at sykdomsprosessen i det hele tatt skal initieres og at de samme eller andre faktorer må være til stede for at videre sykdomsutvikling skal forekomme. Det nye er at vi nå kan måle hvilke varianter av genetisk materiale den enkelte person har (genotyping), og vi kan måle hvilke gener som til en hver tid er aktive (genekspresjon).

Et hovedproblem i å utvikle fruktbare problemstillinger er at vi ikke på forhånd kan vite hvor mange gener som spiller en rolle, hvor hyppig de forekommer i befolkningen (genfrekvensene), og hvilken gjennomslagskraft de har som sykdomsårsaker (målt som absolutt risiko gitt genotypen (penetrans), relativ risiko eller tilskrivbar risiko). Valg av design vil også avhenge av hva man tenker om interaksjoner mellom gener, det vil si hvordan et gen virker sammen med a) det andre genet på samme locus (grad av *dominans*) og b) andre gener på andre loci (*epistasi*). Et annet hovedproblem er målingen av sykdom. Foreligger det stor grad av heterogenitet, det vil si at samme sykdom egentlig er sammensatt av en rekke sykdommer med helt ulik etiologi? Et viktig poeng i planleggingen av designet er også om sykdommen eller sykdomstendensen kan måles langs en kontinuerlig skala eller ikke. Et tredje hovedfokus når man skal velge design, er om man tror at det kan foreligge en interaksjon mellom gener og miljøfaktorer, for eksempel slik at et gen bare har sykdomsfremkallende effekt når en bestemt miljøeksponering er til stede.

Det er viktig at man gjør beregninger av sannsynligheten for å oppdage interessante sammenhenger før man bestemmer seg for design og utvalgsstørrelse. Det sier seg selv at styrkeberegninger vil avhenge mye av hvordan man tenker omkring interaksjoner og heterogenitet.

Selv om man aksepterer kompleksiteten som er nevnt over, må man som regel forenkle og gjøre en del antakelser for å kunne gjennomføre forskningen. Noen enkle problemstillinger er:

1. Er det familiær opphopning av sykdommen?
2. Er sykdommen arvelig, eller bedre: Hvor mye av sykdomstendensen kan tilskrives gener og hvor mye kan tilskrives miljøfaktorer?
3. Kan den arvelige komponenten i vesentlig grad tilskrives gener på ett locus? Er sykdomsgenene recessive, dominante eller virker de additivt?
4. Hvor i genomet (på hvilke kromosomområder) befinner sykdomsgener seg?
5. Hvilken relativ risiko og tilskrivbar risiko kan knyttes til gener på et locus?
6. Er det interaksjoner mellom gener og miljøfaktorer?

FAMILIÆR OPPHOPNING

Problemstilling 1 kan faktisk besvares uten egentlig å benytte et familiedesign. Innen hjerte/kar- og astmaepidemiologi er det for eksempel vanlig å spørre om sykdommen som studeres også forekommer hos de nærmeste slektninger, spesielt mor og far. Dette gir indirekte informasjon om familiær opphopning. For de fleste kroniske sykdommer vil man finne en risikoøkning når det forekommer samme sykdom hos mor, far eller søsken. Dette sier ikke nødvendigvis at risikoøkingen skyldes gener. Man har også vist at det finnes sykdommer med en klar genetisk mekanisme (kromosomfeil, kreft) der det ikke er familiær opphopning. Derfor må man ha et bevisst forhold til om man er interessert i gener som årsak eller gener som er involvert i patogenese. Graden av familiær opphopning, enten det skyldes arv eller miljø, er viktig å beregne for å gi god risikoveiledning til barn og søsken av pasienter. Hvis familiær forekomst er den altoverveiende risikofaktor, har man antakelig å gjøre med en monogen sykdom. Det finnes mange eksempler på at norske og nordiske sykdomsregistre har vist seg godt egnet til å studere familiær opphopning av sykdom (2-4).

ER SYKDOMMEN ARVELIG?

For å beregne hvor mye av sykdomstendensen som kan tilskrives gener (problemstilling 2) må det benyttes spesielle familiedesign. Det vanligste er tvillingstudier (5). I slike studier kan man inkludere målinger av spesifikke gener, men man kan også la det være, og bare måle forekomsten av sykdom hos et tilstrekkelig antall eneggete og toeggete tvillingpar. Den såkalte klassiske tvillingstudie er basert på antakelsen om at eneggete og toeggete tvillinger opplever samme grad av likhet i eksponering for miljøfaktorer, mens de har ulik eksponering fra egne gener. Når det gjelder sykdommer som bare kan måles som diskrete variable

baseres analysen også på antakelsen om en underliggende, normalfordelt sykdomsdisposisjon der sykdommen gir seg klinisk til kjenne når man kommer over en gitt terskel. Graden av likhet mellom tvillinger kan for kontinuerlige trekk (for eksempel blodtrykk) måles som korrelasjonskoeffisienter, mens den for diskrete trekk kan måles som polykoriske korrelasjonskoeffisienter eller som konkordanstall (den probandvise konkordans forteller oss hva risikoen er for at den andre tvillingen har en sykdom gitt at den første har sykdommen). Arveligheten, eller heritabiliteten, måler hvor mye av variansen i sykdomsdisposisjon som kan forklares av variasjon i forekomst av sykdomsfremkallende gener mellom mennesker. Heritabiliteten kan beregnes ved å sette opp et stianalytisk diagram slik at korrelasjonskoeffisientene fortolkes som en funksjon av varianskomponenter, og den genetiske varians estimeres ved en simultan løsning av et sett såkalte strukturelle ligninger (6). Denne fremgangsmåten, der man tilpasser observasjoner i form av korrelasjoner til en spesifikk årsaksmodell, kan benyttes for mange andre typer familiedesign, men er særlig effektiv der man kan skille effekter av gener og felles familiemiljø, slik man til dels også kan i adopsjonsdesign og halvsøskendesign. Hvis man inkluderer genotyping av enkelte loci, kan man estimere hvor mye av den genetiske variasjon som er knyttet til slike loci. I analysene kan man inkludere korrelerte endepunkter (for eksempel angst og depresjon) og korrelasjonsstrukturen mellom slektninger kan benyttes til å studere om samme gener påvirker begge sykdommene (*pleiotrope* gener). Man kan også analysere om det er ulike gener som påvirker samme sykdom hos menn og kvinner.

Metodologien bak den kvantitative genetikken som er beskrevet over har tradisjoner fra Fisher og Wright på begynnelsen av 1900-tallet, og ble utviklet til et kraftfullt analyseverktøy på 1970-80 tallet. Den kvantitative genetikken er praktisk viktig i husdyr- og planteavl, men gir også innsikt i årsaksforholdene bak sykdom hos mennesker, særlig når man vet lite om de enkelte geners betydning. Man får et slags globalt svar om arvens og miljøets relative betydning i den populasjon som er undersøkt. Dette svaret er viktig for å vite hvilken vei forskningen skal gå videre, og det kan avkrefte forestillinger om at miljøfaktorer eller gener er det eneste som spiller en rolle. Spesielt innen atferdsgenetikken har denne type forskning vært viktig og ofte kontroversiell, ikke minst fordi en del resultater har vært overfortolket mot en slags genetisk determinisme. Heritabilitet har ingen fortolkning på individnivå. Det beskriver årsaker til variasjon i en populasjon med de begrensninger som kan knyttes til tid, sted, kultur og materielle forhold.

HVOR MANGE GENER ER INVOLVERT? ER DE RECESSIVE ELLER DOMINANTE?

For å besvare problemstilling 3 må det benyttes familiedesign. Også denne problemstillingen kan løses uten å måle forekomsten av enkeltgener ved laboratorieanalyser. Ved å samle familier med flere tilfeller av samme sykdom kan man utføre det som kalles segregasjonsanalyse, som betyr at man setter opp en årsaksmodell som inkluderer det hovedgenet man tenker på og så ser om fordelingen av sykdom i familiene er forenlig med modellen. Årsaksmodellene er basert på Mendels første lov om segregasjon, altså at allele gener segregerer i kjønnscelledannelsen slik at fra et locus i en diploid celle går et av genene til den ene haploide kjønnscellen og ett til den andre. Årsaksmodellen eller forventningen om hvordan sykdom skal fordeles blant familiemedlemmene vil avhenge av om genene er på de autosomale kromosomene eller kjønnskromosomene, og vil avhenge av genenes interaksjon i et locus (*dominante*, *kodominante* eller *recessive* gener). For enkeltgensykdommer med full *penetrans* (100% sykdomsrisiko gitt tilstedeværelse av sykdomsgenet i enkel (dominant) eller dobbel (recessiv) dose) er det ingen sak å se arvegangen. For komplekse sykdommer vil det ikke være full penetrans. I slike tilfeller kan man anta at det finnes sykdomsgener på flere loci, og man kan anta en *polygen* årsakskomponent (additive effekter av mange gener). Dette gir opphav til mange årsaksmodeller, og den observerte fordelingen av sykdom i familiene settes opp mot disse modellene. Ved hjelp av maximum likelihood funksjoner kan man finne frem til den modell som passer best, og blant annet få estimert penetransen av sykdomsgener. Det er mange vanskeligheter ved segregasjonsanalyser, ikke minst knyttet til heterogenitet av sykdom. Det viktigste bidraget er kanskje at en del årsaksmodeller klart kan forkastes gjennom slike analyser.

HVOR I GENOMET ER SYKDOMSGENENE?

For å løse problemstilling 4 må man ha familiedesign der minst to personer er syke eller minst to personer har et mål på grad av sykdom eller medisinsk tilstand, og her må det benyttes laboratorieanalyser av markørgener. Bakgrunnen for *koblingsanalyse* ligger i Mendels annen lov som sier at gener for ulike egenskaper segregerer uavhengig av hverandre i kjønnscelledannelsen. Man snakker om en haplotype når man tenker på genene fra ulike loci som ligger etter hverandre på et kromosom. Man kan godt si at koblingsstudier er å følge haplotyper gjennom generasjoner. Hvis de forblir uforandret gjennom en meiose sier man at avkommet er en non-rekombinant for den haplotypen. Det er en fordel, men ingen forutsetning, å kjenne fasen for å gjøre koblingsstudier, men dette krever gjerne informasjon fra tre generasjoner. For at en meiose skal være informativ må den av foreldrene som har et sykdomsallel og et normalallel også være heterozygot for markørlocuset som studeres.

Hvis vi har familier med flere syke personer (multiplexfamilier), og vi ikke vet noe om hvor i genomet sykdomsgenet er lokalisert, så kan vi genotype en lang rekke såkalte polymorfe markørloci over hele genomet

(genomscan) for alle familiemedlemmene, inkludert de friske. Disse markørgenene vet vi hvor er lokalisert. I koblingsanalysen identifiseres de informative meiosene og man beregner rekombinasjonsfraksjoner (hvor ofte et markørallel hos en affisert person (for eksempel far) ikke følger med til en annen affisert person (for eksempel datteren). Når det ikke er kobling er altså denne fraksjonen 50%, mens den er 0% når markørallelet er tett innpå sykdomsgenet på samme kromosom. Basert på rekombinasjonsfraksjonen kan man uttale seg om den genetiske avstand mellom to loci. Denne avstanden måles i centimorgan. Det er ikke alltid godt samsvar mellom genetisk distanse og fysisk distanse, som måles i antall baser, oftest kilobaser. Dette manglende samsvar henger sammen med at det ikke er samme sannsynlighet for crossing-over i meiosen langs hele kromosomet.

Man kan beregne sannsynligheten for å observere den gitte fordelingen av sykdomsalleler og markøraller i familiene under ulike rekombinasjonsfraksjoner. Deretter kan man sette opp relative sannsynligheter (odds) der man setter sannsynligheten for observasjonene under ulike grader av rekombinasjon i telleren, og sannsynligheten for de observerte resultater under antakelse om ingen kobling (rekombinasjon i 50% av meiosene) i nevneren. Deretter tar man ved konvensjon den Briggske logaritme av denne størrelsen, og kaller det lod (log odds) score. Hvis lod score er lik 3 betyr det altså at det er 1000 ganger mer sannsynlig å observere det man gjør under en viss rekombinasjonsfraksjon enn det er under en rekombinasjonsfraksjon på 0,5. Det benyttes ulike programmer med maximum likelihood estimering for å estimere rekombinasjonsfraksjonen med høyest lod score. Når man vurderer slike relative sannsynligheter må man ta med a priori sannsynligheten for kobling hvis man har to tilfeldig utvalgte loci. Denne sannsynligheten er omtrent 5%. Deretter må man ta høyde for hvor mange markørloci som det testes for i hver studie. Når man bare estimerer rekombinasjon mellom sykdomslocus og ett markørlocus snakker man om topunkts analyse, mens når man inkluderer mange par av loci i samme område (for blant annet å forstå den innbyrdes rekkefølgen av loci) snakker man om multipunkts analyse.

Hvis man har en god ide om hvor sykdomsgenet ligger, så fokuserer man på det kromosomområdet med mange lokale polymorfismer. Ofte gjøres koblingsstudier i to stadier, en genomscanfase der man prøve å fange opp signaler om kobling, etterfulgt av en mer lokal innsats rundt såkalte kandidatområder. Det har vært mye frustrasjon rundt resultater fra koblingsstudier, fordi det har vært vanskelig å reprodusere funn. Dette kan ha med heterogenitet av sykdommer å gjøre, og kan skyldes ulike definisjoner av sykdom og små utvalg. Man kan si at feltet domineres av mange falske positive resultater, noe som naturligvis har sammenheng med det store antallet polymorfismer.

Man snakker om parametrisk koblingsanalyse når man setter opp parametre knyttet til sykdomsgenene (genfrekvens, dominans, penetrans) i familier. En enkel ikkeparametrisk form for koblingsanalyse benytter affiserte søskenpar. Nullhypotesen for et markørlocus er at det vil fordele seg til søsknene uavhengig av sykdomlocus. Hvis begge foreldre er heterozygote og har ulike alleler (for eksempel at far har genotypen a1a2 og mor a3a4) og den ene av søsknene har a1a3, så er forventningen 25% for at det andre av søsknene skal ha enten a1a3, a1a4, a2a3 eller a2a4. Avvik fra dette tyder på kobling (at et markørallel følger sykdomsallelet fordi de er på samme haplotype). Slike analyser kan gjøres uten at genotypene til foreldrene er kjent.

Hvis man observerer en kontinuerlig sykdomsskala i stedet for en diskret sykdomstilstand, kan man benytte regresjonsmetoder i søskenparanalyser. Man tar regresjonen av den kvadrerte innen-par avstanden i den kontinuerlige fenotypen på antallet alleler som søskenparet har felles på et markørlocus. Avvik fra forventningen om en regresjonskoeffisient på null sannsynliggjør kobling mellom sykdomslocus ("quantitative trait locus") og markørlocus. Det er mest effektivt å konsentrerere seg (og eventuelt bare genotype disse) om par av slektninger der begge ligger høyt på skalaen, begge ligger lavt eller de har mest mulig avstand (i hver ende av skalaen). Litteraturen om koblingsstudier er ofte ganske teknisk og statistisk. Risch har imidlertid skrevet en introduksjon for epidemiologer (7).

For å beregne grad av *assosiasjon* mellom et gen og en sykdom kreves ikke familiedesign. Koblingsanalyser krever altså minst to affiserte personer fra samme familie, mens assosiasjonsstudiene altså ikke har dette kravet. Det er vanligst å bruke case-kontroll design. En confounder som er vanskelig å kompensere for i vanlige epidemiologiske design er såkalt etnisk sammenblanding. Hvis en befolkning består av ulike etniske grupper der sykdomshyppigheten *og* genfrekvensene på de aktuelle loci varierer mye mellom gruppene, vil man kunne få gale estimater av assosiasjonen. Dette kan kontrolleres for ved stratifisering eller ved bruk av et matchet familiedesign. En mulighet er å bruke friske søsken som kontroller. Dette representerer en form for overmatching som gir mindre effektivitet enn bruk av ubeslektede kontroller. Man kan også samle trioer av mor, far og affiserte barn, og genotype alle tre for de aktuelle loci. Da kan man se om et spesielt allel oftere finnes hos det affiserte barn enn forventet (transmission disequilibrium test). Analysemetoder basert på log-lineære modeller er presentert i en annen artikkel i dette nummeret (8).

Det er to strategier for valg av gener i assosiasjonsstudier. Den ene er den såkalte kandidatgen tilnærmingen der man på forhånd synes det er biologisk plausibelt at noen spesielle gener med kjente eller antatte funksjoner skulle være disponerende for sykdom. Man behandler da genene som en hvilket som helst annen potensiell årsaksfaktor i en epidemiologisk undersøkelse, og beregner en odds ratio eller relativ risiko

avhengig av designet. En måte man kan få ideer til kandidatgener på er fra genekspresjonsstudier ved hjelp av mikromatriser, som analyserer affisert vev fra syke personer. De samme genene som er aktive i en sykdomsprosess kan også være viktige som årsaksfaktorer.

Den andre strategien er å type et stort antall markørgener over hele genomet og satse på at enten disse markørgenene selv (noe som er usannsynlig i utgangspunktet når man antar at mennesket har nær 40 000 gener) har noe med sykdommen å gjøre, eller at man på denne måten kan dekke over de fleste kromosomområder ved hjelp av koblingsulikevekt (linkage disequilibrium). Loci som er koblet vil ofte være i koblingsulikevekt. Det betyr at et allel på et locus oftere (eller sjeldnere) forekommer sammen med et visst allel på et annet locus enn forventet fra produktet av genotypefrekvensene (kalles ofte allelisk assosiasjon). En tolkning av en assosiasjon mellom et spesielt allel på et markørlocus og en sykdom er da koblingsulikevekt, som godt kan beskrives som en form for confounding der confounderen er et allel på et nabolocus som er direkte årsaksmessig forbundet med sykdommen og som altså i en populasjon eller et utvalg er assosiert med allelet på markørlocuset. Et godt eksempel på koblingsulikevekt finnes mellom de ulike loci i HLA-systemet. Denne ulikevekten er nødvendig å kjenne til for å fortolke assosiasjoner mellom HLA-varianter og sykdom. Sannsynligheten for koblingsulikevekt synker raskt med fysisk avstand mellom loci, kanskje 3000 basepar er en slags grense. Hvis man antar at menneskets genom består av 3,3 milliarder basepar, kan man regne på hvor mange markørloci man må ta med for å ha styrke til å oppdage assosiasjoner. En måte å redusere antallet på er å først gjøre en koblingsundersøkelse, og deretter gjøre assosiasjonsstudier basert på koblingsulikevekt i de områdene av kromosomene som antyder kobling.

RELATIV RISIKO, TILSKRIVBAR RISIKO OG INTERAKSJON MELLOM GENER OG MILJØFAKTORER

Når man har identifisert de interessante genene og man kan måle miljøeksponeringene vil estimering av relativ risiko, risikoforskjeller, tilskrivbar risiko og studiet av interaksjoner mellom gener og miljø kunne gjøres i vanlige epidemiologiske design (9). For å måle insident sykdom og for å plukke opp den aktuelle miljøeksponering i biologisk materiale som er innsamlet i riktig tidsvindu, vil kohortstudier være å foretrekke, kombinert med nøstede case-kontroll undersøkelser der det biologiske materialet analyseres så effektivt som mulig. Hvis man gjør prevalente case-kontroll studier bør man huske på at en assosiasjon mellom en faktor og sykdom kan skyldes at genet er knyttet til sykdomsprogresjonen og ikke til den primære utvikling av sykdom.

Man kan studere interaksjoner mellom gener og miljø når man ikke kan måle miljøfaktorene direkte. En måte er ved å se om intra-par avstanden i fenotype (for eksempel blodtrykk) for eneggete tvillinger er systematisk annerledes når de eneggete parene stratifiseres etter genotypen i det aktuelle locus (10). Man kan også få en ide om interaksjon ved å estimere heritabilitet i utvalg av familier, for eksempel tvillinger, etter stratifikasjon etter miljøfaktorer. I et slikt tilfelle trenger man ikke gjøre genotyping.

VEIEN VIDERE

Epidemiologi er et fag som har fokus mot forebygging av sykdom. I de fleste epidemiologiske undersøkelser som er gjort de siste 50 årene, har man ikke inkludert genetiske problemstillinger. Hensikten har vært å identifisere miljøfaktorer som kan la seg endre ved tiltak innen forebyggende medisin. For de fleste kroniske sykdommer har suksessen uteblitt. Vi er i liten grad i stand til å forebygge sykdommer, selv om en rekke risikofaktorer er identifisert. Dette skyldes ikke bare at vi ikke forstår årsakskjedene, men også at vi ikke rår over effektive og akseptable tiltak. Men det gjenstår omfattende forskning før vi kan si at årsaksmønsteret er kartlagt. Hele tiden har man ant at genene er viktige som sykdomsårsak, ikke minst på bakgrunn av tvillingstudier. Det er blitt et alminnelig munnhell å si at det vi må gjøre er å finne frem til interaksjonene mellom gener og miljøfaktorer.

Parallelt og nokså uavhengig av utviklingen i epidemiologi har det vært en utvikling av metoder i det faget som kalles kvantitativ genetikk. I tillegg er populasjonsgenetikken et forskningsområde som beskriver og forklarer ulikheter i genetisk sammensetning mellom folkegrupper. Forklaringene knyttes opp mot evolusjonskreftene; migrasjon, mutasjon, seleksjon og genetisk drift (stokastiske endringer fra en generasjon til en annen). I dag er disse tradisjonene i ferd med å smelte sammen. Fremgangen i molekylærgenetikk gjør at forskningen er fokusert på å finne sykdomsgener. Genetisk epidemiologi, som er årsaksforskning i befolkninger der gener er inkludert som årsaksfaktor, kan for tiden best beskrives som nål-i-en-høystakk-epidemiologi. Det er også sterke økonomiske motiver knyttet til kommersiell utnytting av kunnskapen om enkeltgeners funksjon som ligger bak disse fisketurene. Fra et folkehelsesynspunkt er det viktig å ha med seg det synspunkt at det er ved å endre på miljøfaktorer eller ved å styrke verten (for eksempel ved vaksinasjon) at sykdom skal forebygges, men at denne forebyggingen blir best når vi forstår samspillet mellom gener og miljø.

Den biokjemiske beskrivelse av baserekkefølgen (sekvensen av baser) i menneskets DNA-molekyler i våre kromosomer er kommet langt. Dermed kan vi genotype svært mange loci. I tillegg har mulighetene for å studere genekspresjon økt med den nye mikromatriseteknikken. Proteomics og metabonomics er

laboratorieorienterte fagområder som beskriver og fortolker proteinaktivitet og stoffskifteprosesser med nye teknikker. I tillegg til å ha kunnskap om design, må epidemiologer også anstrenge seg for å ha en forståelse av hva som foregår på disse feltene. Dette vil være nødvendig for å kunne rette sine problemstillinger riktig inn, og for å forstå hvilke målinger som er mulig når man skal studere årsaker og mekanismer for sykdom. I fremtiden kan man tenke seg at man kan sekvensere den enkelte deltakers genom, det vil si få en full genotyping av alle loci. Dette gir en overveldende mengde informasjon, og håndtering og analysen av slike datamengder er allerede en stor utfordring.

I genetisk epidemiologi er familiestudier nødvendige for noen av problemstillingene som er beskrevet i denne artikkelen, men for mange formål vil kohortstudier og case-kontroll undersøkelser være tilstrekkelig. I dag skjer en rask utvikling av design og analyseteknikker innen dette feltet. Norske forskere bør være med på utviklingen sett i lys av FUGE-satsningen og de naturlige forutsetninger for denne forskning i vårt land.

REFERANSER

1. Khoury MJ, Cohen BH, Beaty TH. *Fundamentals of Genetic Epidemiology*. Oxford University Press, 1993.
2. Winther JF, Sankila R, Boice JD, Tulinius H, Bautz A, Barlow L, Glattre E, Langmark F, Moller TR, Mulvihill JJ, Olafsdottir GH, Ritvanen A, Olsen JH. Cancer in siblings of children with cancer in the Nordic countries: a population-based cohort study. *Lancet* 2001; **358**: 711-7.
3. Lie RT, Wilcox A, Skjærven R. A population-based study of recurrence risks of birth defects. *N Engl J Med* 1994; **331**: 1-4.
4. Stoltenberg C, Magnus P, Skrondal A, Lie RT. Consanguinity and recurrence risk of stillbirth and infant death. *Am J Public Health* 1999; **89**: 517-23.
5. Magnus P, Harris JR, Nystad W, Tambs K. Hva kan tvillingforskning fortelle om årsaker til sykdom? *Tidsskr Nor Lægeforen* 1999; **119**: 3317-21.
6. Neale MC, Cardon LR. *Methodology for genetic studies of twins and families*. Dordrecht: Kluwer Academic publishers, 1992.
7. Risch N. Evolving methods in genetic epidemiology. II. Genetic linkage from an epidemiologic perspective. *Epidemiol Rev* 1997; **19**: 24-32.
8. Lie RT, Jugessur A. Analyse av pasient-foreldre triader; en praktisk gjennomgang. *Norsk Epidemiologi* 2002; **12**: 119-130.
9. Stene LC. Introduksjon til analyse av gen-gen og gen-miljø interaksjoner i case-kontroll studier. *Norsk Epidemiologi* 2002; **12**: 109-117.
10. Magnus P, Berg K, Børresen A-L, Nance WE. Apparent influence of marker genotypes on variation in serum cholesterol in monozygous twins. *Clin Genet* 1981; **19**: 67-70.