Genetiske analyser i genetisk epidemiologi

Anne Spurkland og Hanne Flinstad Harbo

Immunologisk institutt, Rikshospitalet, 0027 Oslo

Korresponderende forfatter: Anne Spurkland, telefon: 2307 1377 telefax: 2307 3510 e-post anne.spurkland@labmed.uio.no

SAMMENDRAG

DNA-polymorfi kan defineres som en arvelig genetisk variant på et bestemt sted (locus) i genomet med en frekvens over 1% i befolkningen. Slik polymorfi kan bestå av en enkeltbase-substitusjon (såkalt "single nucleotide polymorphism" eller SNP), av delesjon eller insersjon av en eller flere baser, eller av variabel lengde på en repeterende DNA-sekvens (såkalt "variable number of tandem repeats" eller VNTR). Det finnes en rekke molekylærgenetiske metoder for å påvise DNA-polymorfi. For genetisk-epidemiologiske studier vil det være mest aktuelt med metoder som er lette å automatisere, krever lite DNA-materiale, og som kan utføres på et stort antall prøver. I denne oversiktsartikkelen vil vi gå igjennom prinsippene for noen molekylærgenetiske metoder som er aktuelle for bruk i store genetisk-epidemiologiske studier. Vi vil også omtale bruk av sammenslåtte DNA-prøver i screening for sykdomsdisponerende gener, som er en effektiv snarvei til å samle store mengder genetisk informasjon fra mange individer med lav kostnad per testet individ.

Spurkland A, Harbo HF. **Genetic analysis in genetic epidemiology** *Nor J Epidemiol* 2002; **12** (2): 89-96.

ENGLISH SUMMARY

DNA polymorphism can be defined as an inheritable genetic variant on a particular place (locus) in the genome, with a population frequency above 1% in the population. Such polymorphism can consist of one single base substitution (so called single nucleotide polymorphism or SNP), of deletions or insertions of one or several bases, or of a variable length of one repetitive DNA sequence (so called "variable number of tandem repeats" or VNTR). There are a number of methods in molecular genetics designed to demonstrate DNA polymorphism. For genetic epidemiology studies, methods that are easy to automate, which demand little DNA material, and which can be performed on a large number of samples are the most useful. In this review we will present the principles of some molecular genetic analysis that are relevant for use in large genetic epidemiologic studies. We will also present the usage of pooled DNA in screening for disease susceptibility genes, as an efficient "short cut" to collect large amount of genetic information from many individuals with a low cost per tested individual. With the present day methods, the amount of DNA available for genetic epidemiologic studies may represent a limitation for the number of genetic analysis that can be performed in any given sample. If in the future, methods that allow sequencing of entire genomes using a minimal amount of DNA are developed, the availability of sufficient amounts of DNA will no longer present a limitation to genetic epidemiologic studies.

INNLEDNING

Gener inneholder arvelig informasjon om kroppens bestanddeler. Informasjonen er lagret i DNA, **d**eoxyribo-**n**ucleic-**a**cid, som består av sukkerarten deoksyribose og fire ulike såkalte "baser". De fire basene er guanin (G), cytosin (C), adenin (A) og thymidin (T). Det er rekkefølgen, eller sekvensen, av basene i DNA tråden som inneholder den arvelige informasjonen.

DNA er et anti-parallelt dobbelttrådet molekyl, der den ene tråden er et speilbilde av den andre tråden, men med motsatt orientering (figur 1). De to trådene i DNA-molekylet holdes sammen av parvise hydrogenbindinger mellom basene. Adenin og thymidin parer seg med to hydrogenbindinger, mens cytosin og guanin parer seg med tre hydrogenbindinger. Siden de to trådene i DNA holdes sammen bare av hydrogenbindinger, vil det dobbelttrådete DNA-molekylet "smelte" til to enkelttrådete DNA-molekyler ved oppvarming. Når temperaturen senkes igjen, vil de to trådene gå sammen igjen, på en slik måte at baseparingen mellom trådene blir mest mulig korrekt. Dette er en viktig egenskap ved DNA, som utnyttes i flere av de ulike testene som vil bli omtalt senere.

Mennesker har 23 par kromosomer. Hvert kromosom består av en enkelt oppkveilet DNA-tråd. På kromosomene finnes all genetisk informasjon, og denne samlete arvemassen kalles ”genomet”. Vi arver kromosomene fra foreldrene våre på en slik måte at det ene kromosomet i et par stammer fra mor, det andre

fra far. Med et gen forstår vi vanligvis den delen av DNA som bestemmer oppbygningen av et bestemt protein. I det menneskelige genomet vet vi nå at det finnes ca. 35 000 gener (1).

Figur 1. Skjematisk framstilling av en dobbelttrådet DNA-sekvens. Baserekkefølgen angis fra 5' til 3' retning. Den komplementære tråden angis fra 3' til 5' retning. Baser som parer seg med hverandre angis med en strek: |.

GENETISK VARIASJON

Hvert gen har sin faste plass, såkalt locus, på kromosomene. For hvert locus kan det finnes flere ulike varianter av et gen. Slike varianter kalles alleler. En person som på et bestemt locus har arvet samme allel både fra mor og far er homozygot. En person som har arvet to ulike alleler er heterozygot. Fordi hvert menneske har to kopier av hvert gen, er det viktig i genetiske studier å skille på hvor mange i en befolkning som er bærer av et bestemt allel (fenotypefrekvens) og hvor mange alleler som er påvist i en bestemt befolkning (allel eller genfrekvens).

Overkrysning og koblingsulikevekt

Gener som sitter på samme kromosom er koblet. I kjønnscelledelingen vil alleler som sitter på det ene kromosomet i et par, stokkes med de tilsvarende allelene som sitter på det andre kromosomet. Dette kalles overkrysning eller rekombinasjon, og denne prosessen gjør at avkommet aldri arver nøyaktig samme kombinasjon av alleler fra sine foreldre som den foreldrene arvet fra sine foreldre igjen. Overkrysning motvirker effekten av kobling av gener. Jo lenger fra hverandre på kromosomet to gener fysisk befinner seg, jo større sannsynlighet er det for at det vil skje en overkrysning mellom de to genene. Varianter av gener som sitter langt fra hverandre på kromosomet vil derfor opptre uavhengig av hverandre på befolkningsnivå. Jo tettere to gener sitter på samme kromosom, jo sjeldnere vil det skje overkrysning mellom dem. I en befolkning vil en derfor kunne se at to genvarianter opptrer oftere sammen enn forventet utifra hver av allelenes frekvens i befolkningen. Dette kalles koblingsulikevekt (linkage disequilibrium).

Mutasjoner og polymorfier

Når gener nedarves til neste generasjon skjer det vanligvis som blåkopier av mor eller fars gener. Men av og til skjer det forandringer eller mutasjoner i genene. Disse endringene kan være enkeltbase-substitusjoner der en base byttes ut med en annen, eller det kan være tap eller tillegg (delesjon eller insersjon) av en eller flere baser. Når slike mutasjoner skjer i kjønnscellenes DNA, kan de nedarves til neste generasjon. Ofte fører mutasjoner til at genet ikke lenger fungerer som det skal, og det kan gi opphav til genetisk sykdom. Mutasjoner som påvirker individets mulighet til å få barn, vil ikke få stor utbredelse i en befolkning. Derimot vil mutasjoner som fører til at genets funksjon fortsatt er intakt eller til og med kanskje noe bedret, lettere kunne spre seg i en befolkning over mange generasjoner. Mutasjoner som finnes hos over 1% av befolkningen, kalles derfor heller for polymorfier. Strukturelt sett representerer derfor mutasjoner og polymorfier de samme typer DNA-variasjon. Sannsynligvis opprettholdes polymorfier i en befolkning for å øke befolkningens tilpasning til et miljø i forandring. Et eksempel på dette er at vi som bor i nordlige strøk drikker melk hele livet uten besvær. Det normale blant folk i sydligere strøk er at spedbarnas evne til å fordøye laktose eller melkesukker forsvinner i barneårene, og at inntak av melk etter dette gir diaré og magesmerter. I nordlige strøk har inntak av melkeprodukter vært viktig for å overleve, og de som har genvarianter som tillater at de kan nedbryte melkesukker hele livet har hatt en fordel framfor de andre. Denne fordelen kan ha gjort at "fordøyere av melkesukker" har fått flere avkom, og dermed at deres gener har fått større utbredelse i befolkningen.

GENETISK EPIDEMIOLOGI

Det er mulig å studere samspillet mellom gener og miljø i dyremodeller, der dyrene gjennom generasjoner har vært parret med sine nære slektninger. Slike forsøksdyr kalles kon-gene fordi de er nesten helt identiske genetisk sett. Ved å sammenlikne ulike kon-gene innavlete stammer som bare atskiller seg på noen få loci, kan en direkte studere disse genenes innflytelse på dyrenes respons på en kontrollert miljøpåvirkning.

Så enkelt er det dessverre ikke når vi studerer miljøets innflytelse på utvikling av sykdom hos mennesker. Bortsett fra eneggete tvillinger, er det ingen av oss som genetisk sett er identiske. Hittil har da også epidemiologiske undersøkelser stort sett ikke inkludert genetiske faktorer. Ikke desto mindre er det klart at gener kan bestemme grad av sårbarhet overfor miljøpåvirkninger, og epidemiologiske studier som inkluderer genetiske variable – genetisk epidemiologi – vil derfor kunne avsløre nye sammenhenger mellom miljøfaktorer og sykdom. For genetisk-epidemiologiske studier vil det være mest aktuelt med metoder som er lette å automatisere, krever lite DNA-materiale, og som kan utføres på et stort antall prøver.

Påvisning av gener som disponerer for sykdom

Påvisning av gener som disponerer for sykdom kan gjøres med familiebaserte koblingsundersøkelser (linkage-studier), eller med populasjonsbaserte pasient-kontroll-metoder, såkalte assosiasjonsanalyser.

Koblingsundersøkelser baserer seg på at gener på samme kromosom i prinsippet nedarves samlet. På grunn av overkrysning vil likevel gener som sitter langt fra hverandre på kromosomet ha en 50% sjanse for å nedarves på hvert sitt kromosom. Jo tettere genene sitter på kromosomet, jo sjeldnere skjer det overkrysning akkurat mellom disse to genene. Ved å undersøke om et bestemt gen nedarves sammen med sykdommen i store sykdomsbelastede familier, kan en se om genet og sykdommen oftere enn forventet nedarves sammen. Dette vil i så fall være et tegn på at genet som undersøkes og sykdomsgenet sitter ganske nær hverandre på samme kromosom. Slike studier av familier der flere nære slektninger er affisert av samme sykdom er godt egnet til å finne gener som forårsaker monogene sykdommer, der forandring i ett gen er årsaken til sykdommen. For polygene eller komplekse sykdommer der flere gener gir opphav til sykdommen, har det vist seg å være vanskeligere å påvise gener som er involvert i sykdomsutviklingen med denne metoden. Grunnen er blant annet at mange av genene som er involvert i polygene sykdommer hverken er nødvendige eller tilstrekkelige til å utvikle sykdommen. For en del polygene sykdommer har imidlertid koblingsstudier sannsynliggjort at flere genområder, som hver for seg gir en moderat sykdomsrisiko, kan være involvert i sykdomsutviklingen (2).

For genetisk-epidemiologiske studier er det hovedsakelig pasient-kontroll- eller assosiasjons-studier som vil være mest hensiktsmessig. Her baserer en seg på at det er koblingsulikevekt mellom det undersøkte genet og det sanne sykdomsgenet. Jo tettere to gener sitter, jo mer sannsynlig er det at alleler på disse to loci opptrer sammen i befolkningen. En økt forekomst av et bestemt allel i en sykdomsbefolkning betyr derfor at det enten er det undersøkte allelet som selv bidrar til sykdomsdisposisjonen eller det er et allel på et nært koblet locus som medfører sykdomsdisposisjonen. Å skille disse to mulighetene fra hverandre kan være vanskelig.

DET HUMANE GENOMPROSJEKTET

Det humane genomprosjektet (http://ornl.gov/TechResources/Human_Genome/home.html) har nå kartlagt alle menneskets gener i stor detalj. Prosjektet, som har hatt en tidsramme på 15 år, har vært et samarbeid mellom mange forskningsinstitusjoner over hele verden gjennom den humane genom organisasjonen (HUGO). HUGO har sørget for at det er etablert store offentlig tilgjengelige databaser over genmarkører (Sequence Tagged Site (STS)), uttrykte gener (Expressed Sequence Tag (EST)) og mye annen nyttig informasjon om genomet. Det er utviklet mye ny teknologi for effektiv analyse av gener og gensekvenser.

Det endelige sluttresultatet av prosjektet vil være hele baserekkefølgen i det menneskelige genomet, ca. 3 milliarder baser. I februar 2001 ble et nesten ferdig utkast til det menneskelige genomet publisert (1). Dette betyr at det ikke lenger vil være nødvendig å arbeide i årevis for å kartlegge og sekvensere et genområde man er interessert i. Hvis man leter etter gener som bidrar til en sykdomstilstand, og har identifisert et aktuelt genområde ved hjelp av koblingsanalyser eller assosiasjonsstudier, kan en nå «slå opp» i genomet, og få oversikt over hvilke gener som finnes i det genområdet man har mistanke om kan være assosiert med sykdommen. En god inngangsport for å bla igjennom genomet er "Genome browser" (www.genome.ucsc.edu) som kobler sammen genom-informasjon fra ulike databaser på en lett tilgjengelig og oversiktlig måte.

Selv om genomsekvensen nå er kjent, gjenstår det fortsatt mye arbeid for å forstå hvordan genene bidrar til en sykdom eller tilstand, hvordan genene er regulert og hvordan de samarbeider seg i mellom. Genomprosjektet undersøker bare DNA sekvensen til et individ, eller rettere sagt, hver enkelt del av genomet blir i prinsippet bare undersøkt hos en person. En viktig utfordring i tiden framover er derfor å kartlegge genetisk variasjon i ulike befolkninger (3).

MOLEKYLÆRGENETIKK

Vi kan undersøke en persons gener med enkle metoder, forutsetningen er at vi har tilgang til DNA fra personen og at genene viser variasjon. Aktuelle prøvematerialer for DNA-isolering og metoder for genanalyse vil bli omtalt under.

Prøvematerialer for DNA-analyse

DNA kan ekstraheres fra alle kjerneholdige celler i kroppen. Når genetisk-epidemiologiske prosjekter planlegges, er det nødvendig å ha en gjennomtenkt plan for hvordan DNA skal samles inn, tas vare på og brukes i forbindelse med prosjektet. Den vanligste måten å samle prøver til genetiske analyser er å be om å få en blodprøve fra personene som skal inngå i undersøkelsen. En blodprøve på 10 ml vil normalt gi ca. 100-150 μg DNA. Dette er nok til under optimale forhold å utføre i størrelsesorden 5-10 000 genanalyser. Det betyr med andre ord at det kan gjøres et stort, men likevel begrenset, antall genanalyser fra en enkelt blodprøve.

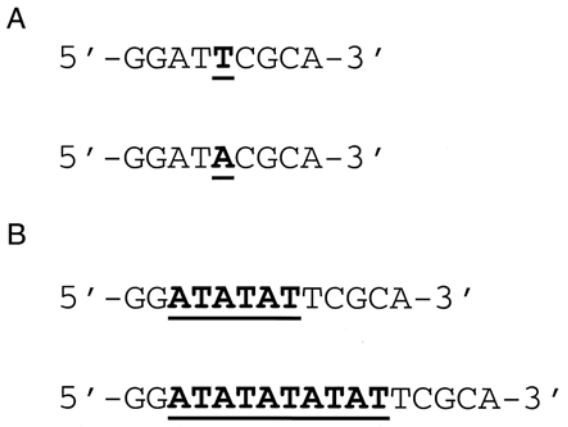
For store befolkningsundersøkelser kan det være upraktisk å samle inn blodprøver. Da kan prøver fra munnslimhinne på bomullspinner være et godt alternativ. Denne prøven kan forsøkspersonen ta selv ved å stryke en tilsendt bomullspinne langs munnhulen og deretter putte pinnen i et passende reagensrør. Utbyttet av DNA fra slike prøver, må en imidlertid regne med er betydelig mindre enn fra blodprøver.

Vevsprøver er en alternativ kilde til DNA, som kan være aktuelt i noen sammenhenger. Hvis vevsprøven er oppbevart som ferskt vev, uten konserveringsmidler, vil det være enkelt å isolere DNA fra prøven. Hvis vevet er fiksert, slik det ofte vil være i patologisk-anatomiske arkivmaterialer, kan det derimot by på

store problemer å isolere tilstrekkelig DNA av god kvalitet. Grunnen til dette er at fikseringsmidlene som er benyttet på ulike måter kan ha ødelagt DNA-molekylene slik at de ikke lenger kan analyseres.

Hva slags genetisk variasjon kan enkelt types?

Det finnes flere former for DNA-polymorfi, som er mer eller mindre godt egnet for store genetisk-epidemiologiske undersøkelser. Den enkleste og hyppigste formen for DNA-polymorfi er enkeltbase-forandringer (single nucleotide polymorphism eller SNP) (figur 2a). SNP kan være enten alternative baser på samme sted i genomet, eller tap eller tilskudd (delesjon eller insersjon) av en eller noen få baser på et bestemt sted i genomet. Gjennomsnittlig finnes det en SNP per 1000 baser i genomet (4), det vil si at menneskets genom innholder 1,42 millioner singel nukleotid polymorfismer (5). Siden de fleste gener spenner over flere tusen baser, vil det i hvert enkelt gen finnes flere SNP. Det er derfor knyttet store forhåpninger til bruk av slike SNP i studiet av komplekse genetiske egenskaper (6), og databanker over SNP er opprettet (7).

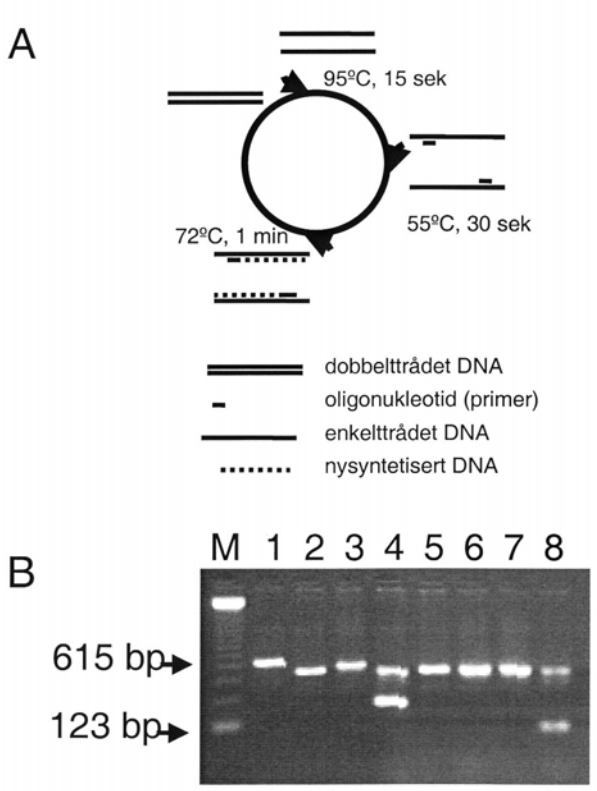


Figur 2. **A**: Eksempel på enkelt base polymorfi. Den avvikende basen i sekvensparet er understreket. **B**: Eksempel på variabel lengde polymorfi. Den repeterende sekvensen AT er understreket. De to sekvensene atskiller seg med to repetisjoner av AT.

Variabel lengde polymorfi (VNTR) er en annen form for polymorfi som det finnes relativt mye av i genomet, og som det er enkelt å type for. VNTR er en repetert DNA-sekvens som forekommer i et ulikt antall repetisjoner (figur 2b). Det er mest aktuelt å type for VNTR loci med korte strekk (50-100 baser) av to til fem-seks basers repetisjoner.

Polymerase kjedereaksjonen

Det er mulig å mangfoldiggjøre et bestemt genområde mange ganger i løpet av et par timer ved hjelp av en metode som kalles PCR, polymerase kjedereaksjonen (figur 3a). Denne teknikken var et stort framskritt da den ble introdusert på slutten av 80-tallet. Utviklingen av metoden ble belønnet med Nobelprisen i medisin i 1993 (8). Ved hjelp av PCR-metoden kan ethvert kjent gen undersøkes enkelt og billig. Foreløpig er det i liten grad utviklet effektiv teknologi for å undersøke DNA-variasjon uten å gå veien om mangfoldiggjøring av bestemte genområder. PCR-produkter påvises enkelt ved å farge med stoffer som binder seg til DNA (for eksempel etidiumbromid) og separere produktet i en agarosegel ved hjelp av elektroforese. Vandringen av DNA-molekyler i en gel-elektroforese er proporsjonal med lengden av DNA-molekylet. PCR-produktet inneholder hovedsakelig DNA av en bestemt lengde, og vil derfor sees som et distinkt DNA-bånd i gelen etter elektroforese (figur 3b).

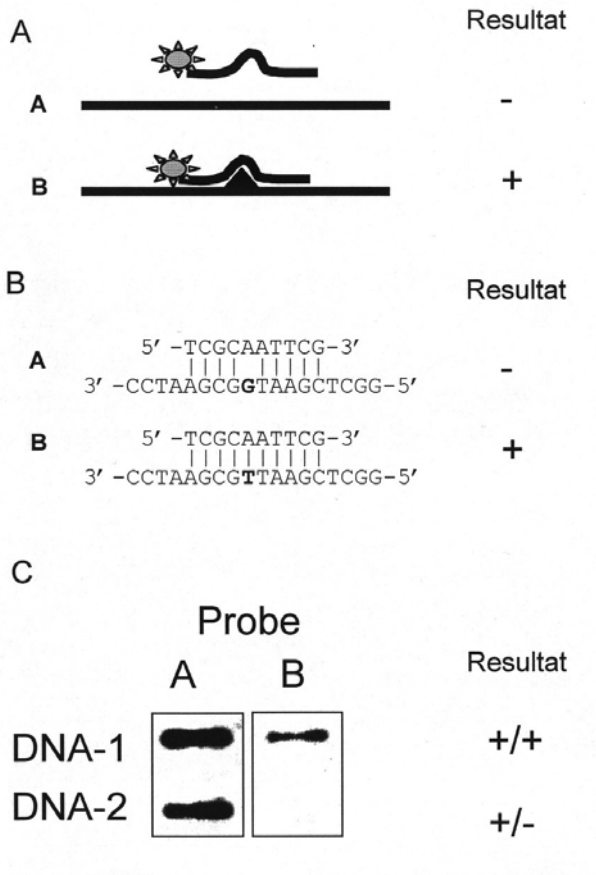


Figur 3. **A**: Skjematisk framstilling av PCR-reaksjonen. Reaksjonen er tegnet syklisk for å få fram at de samme tre hendelsene gjentar seg 30-40 ganger i løpet av reaksjonen. Dobbelttrådig DNA smeltes eller denatureres i enkelttrådig DNA ved 95 ºC. Temperaturen senkes til 55 ºC slik at enkelttrådige korte syntetiske primere kan binde seg til hver sin DNA-tråd, i hver sin ende av det området som skal mangfoldiggjøres. Temperaturen heves igjen til 72 ºC der det varmestabile enzymet Taq polymerase, som lager en ny DNA-tråd med identisk men motsatt sekvens av templatet, er mest aktivt. Det nylagete DNA samt det opprinnelige DNA smeltes på nytt, og reaksjonen gjentas.
B: Påvisning av PCR-produkter i en agarose gel elektroforese. DNA av lik lengde vandrer like langt i et elektrisk felt. Et fargestoff som binder DNA benyttes for å synliggjøre DNA i gelen. En DNA-stige benyttes som markør for å anslå lengden på PCR-produktene. Eksempelet viser resultatet av en HLA-typings reaksjon med allel-spesifikke primere. Posisjon 4 og 8 har to bånd, som uttrykk for en positiv reaksjon. De øvrige posisjonene har bare ett kontrollbånd, som uttrykk for at PCR-reaksjonen har vært vellykket.

Polymorfi i et genområde kan påvises direkte ved hjelp av PCR-reaksjonen ved at denne lages slik at bare bestemte genvarianter mangfoldiggjøres. Like aktuelt er det å lage PCR-reaksjonen slik at alle aktuelle varianter mangfoldiggjøres, for deretter å påvises enkeltvis. Dette kan gjøres på mange måter. Noen av disse vil bli omtalt i det følgende. En god oversikt over aktuelle metoder er også gitt av Ellsworth og Manolio (9).

SSO-probing

En mye benyttet teknikk baserer seg på bruk av korte sekvensspesifikke enkelttrådet DNA, såkalte oligonukleotidprober eller SSO. Hver enkelt probe binder seg bare til en bestemt genvariant i PCR-produktet som skal undersøkes. Fordi probene er merket, er det mulig å registrere hvilke genvarianter som er til stede eller ikke (figur 4). Probing med syntetiske DNA-tråder kan også skje ved at probene sitter fast på en membran, og PCR-produktet som skal undersøkes er merket og i løsning fri til å binde seg til prober med identisk sekvens. Dette kalles revers SSO-probing.



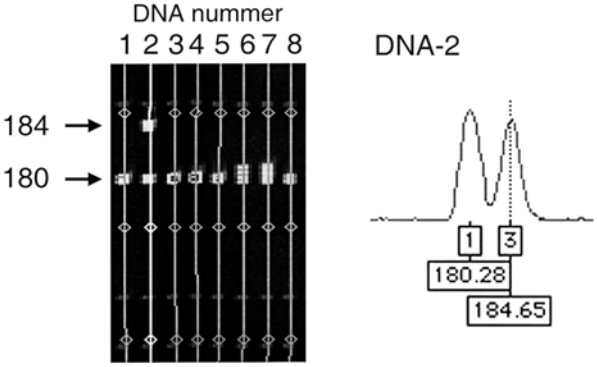
Figur 4. **A**: Skjematisk framstilling av oligonukleotid probing. Prober og DNA er framstilt som streker. Enkeltbaseforskjell er framstilt som trekant. Den radioaktivt merkete proben binder seg bedre til den ene enn til den andre DNA-sekvensen, og dette kan avleses som et positivt signal. **B**: Som A, med proben og DNA-sekvensen skrevet med basesymboler. **C**: Resultat av et probeforsøk. DNA-1 reagerer med både probe A og B, mens DNA-2 bare reagerer med probe A.

En alternativ variant av revers SSO-probing er å feste eller syntetisere spesifikke prober i bestemte posisjoner på en liten fast overflate. Slike mikrochips kan undersøke genvariasjon i mange gener på en gang i en såkalt microarray-analyse (4). SSO-probing ved hjelp av chips-teknologi tilfredsstiller kravet til at metoden skal være automatiserbar.

SNP kan også påvises med andre metoder, som for eksempel restriksjonsenzym fordøyelse (kalt restriksjons fragment lengde polymorfi (RFLP) analyse). Restriksjonsenzymer er enzymer som binder seg spesifikt til bestemte korte sekvensmotiver og kutter DNA-tråden i dette området. Gjenkjennelsessekvensene er ofte palindrome sekvenser (det vil si de leses likt enten en leser den ene eller den andre av DNA-trådene), men ikke nødvendigvis. For å kunne type ved hjelp av restriksjonsenzymer er det en forutsetning at SNP ligger i gjenkjennelsesstedet for et restriksjonsenzym, slik at det ene av allelene blir kuttet på dette stedet av det bestemte restriksjonsenzymet mens det andre allelet ikke blir kuttet. Denne metoden egner seg best for småskalert typing. RFLP-basert genotyping var forøvrig den første form for DNA-typing som ble etablert.

Variabel lengde analyse

VNTR polymorfi kan undersøkes enkelt ved å amplifisere det variable lengde-området med et primersett der den ene primeren er merket med et fluorescerende molekyl. PCR-produktene separeres i en acrylamidgel i et elektrisk felt, og signalet fra den merkete primeren fanges opp av en scanner på et bestemt sted i gelen. Ved å velge PCR-primere som gir produkter med ulike lengder og også velge ulike fluorokromer til å merke produktene, er det mulig å undersøke 10-15 loci i et og samme spor i en og samme gel (figur 5). Denne metoden er derfor ganske effektiv og har vært benyttet til å gjøre genomvide koblingsundersøkelser av polygene sykdommer.



Figur 5. Automatisk typing av VNTR-polymorfi ved hjelp av automatisk gelelektroforese fluorescens scanning. Prøve 1-8 representerer PCR-produkter som inneholder en variabel lengde polymorfi. Sammen med prøven er kjørt en intern lengde standard, for å kunne sammenlikne prøvene innbyrdes. Prøve 2 har to alleler, allel 1 på 180 og allel 3 på 184 basepar. For å analysere gelbildet benyttes et software som framstiller resultatet fra prøve 2 som vist til høyre i figuren.

Sekvensering

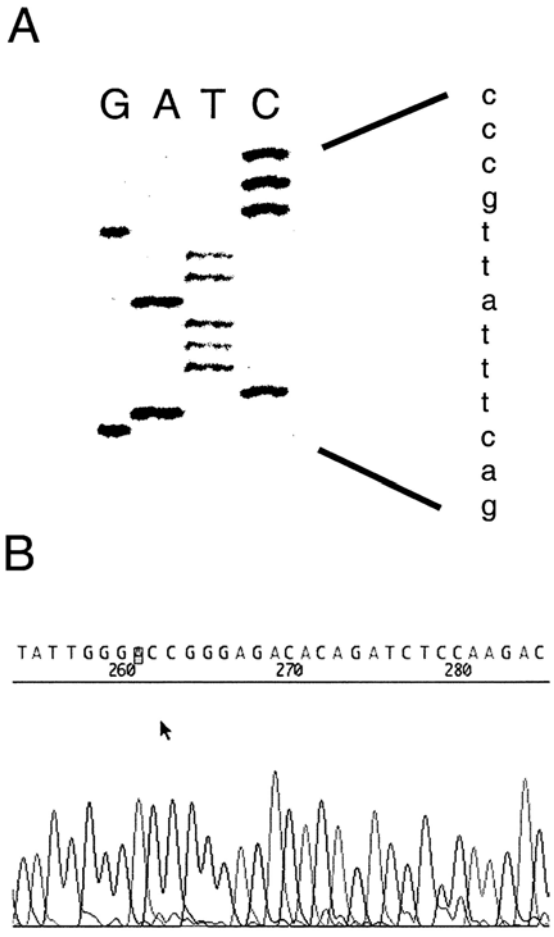
Den mest direkte måten å undersøke DNA-polymorfi på er å bestemme sekvensen til det aktuelle genfragmentet. Dette gjøres ved såkalt sekvensering, som vanligvis baserer seg på bruk av en enzymatisk DNA-polymeriseringsreaksjon. Man benytter en primer som binder seg til sekvensen som skal undersøkes sammen med en blanding av de fire normale basene og fire defekte baser merket med fire ulike farger. Når DNA-polymerasen inkorporerer en defekt base i den nye DNA-tråden, stopper syntesen opp. En sekvenseringsreaksjon resulterer derfor i en blanding av DNA-tråder av ulik lengde. Når sekvenseringsproduktet separeres i en acrylamidgel, vil det resultere i en stige, med en bases forskjell mellom "trinnene". Ved å lese av hvilken farge hvert av "trinnene" i stigen har, kan en bestemme hvilken sekvens det opprinnelige DNA-fragmentet har (figur 6). Denne metoden kan benyttes til småskalert typing, men egner seg ikke for genetisk-epidemiologiske undersøkelser. For å påvise hittil ukjente polymorfier er det imidlertid fortsatt nødvendig å sekvensere samme gen fra et visst antall personer.

Individuell genomtyping

Det er foreløpig begrenset hvor mange genanalyser det er mulig å utføre på DNA isolert fra en enkelt blodprøve. I framtiden vil det imidlertid sannsynligvis bli mulig å gjøre en global gentest slik at en på noen timer i en liten blodprøve kan gjøre et helt "humant genomprosjekt" på individnivå. Individer kan da testes en gang for alle, og vil kunne bære med seg sitt individuelle genom i en liten databrikke. Da vil også tilgangen på tilstrekkelig DNA av god kvalitet ikke lenger være den samme begrensende faktoren ved genetisk epidemiologiske studier slik det er i dag. Utfordringen vil da heller være å håndtere de store datamengdene og se relevante mønstre i genprofilene til de undersøkte individene.

GENOMVID ASSOSIASJONS-SCREENING

For komplekse sykdommer, der flere gener er involvert i tilegg til miljøfaktorer, har genomvide koblingsundersøkelser ikke klart å peke ut med sikkerhet hvilke genområder som er involvert. Man har derfor ønsket å benytte styrken til assosiasjonsmetoden (pasient-kontroll-studier) i genomvid screening. Fordi hver markør i en assosiasjonsstudie tester et kortere genområde (ca. 1 centiMorgan (cM[1]) avhengig av grad av koblingsulikevekt) enn en markør i en koblingsstudie (ca. 12 cM), er det nødvendig å bruke langt flere markører i en genomvid assosiasjons-screening. Mens koblingsstudiene vanligvis benytter 300-400 markører for å dekke genomet, må en genomvid assosiasjonsstudie benytte flere tusen markører. Det har derfor vært tilnærmet praktisk umulig for en enkelt forskningsgruppe å utføre genomvid assosiasjons-screening, inntil pooling av DNA ble en anerkjent metode (10).



Figur 6. A: Resultat av manuell sekvensering. Det benyttes bare en type merking på de defekte basene som inngår i reaksjonen, og hver base må derfor analyseres separat. De fire analysene separeres ved siden av hverandre ved hjelp av elektroforese. Resultatet blir en stige, der hvert enkelt trinn representerer en enkelt base. Ved å følge stigen fra bunn til topp, kan sekvenseringsresultatet tolkes, som angitt til høyre for gelbildet.
B: Resultat av automatisk sekvensering. De fire basene analyseres samtidig i samme prøve. Dette gjøres ved å benytte fire ulike fargemerkinger til de fire defekte basene. Prøven separeres og analyseres ved hjelp av automatisk gelelektroforese fluorescens scanning. Et dataprogram analyserer sekvensen, og framstiller den som vist i figuren, både med bokstaver og med "topper" som representerer rådataene som sekvensen er laget utifra.

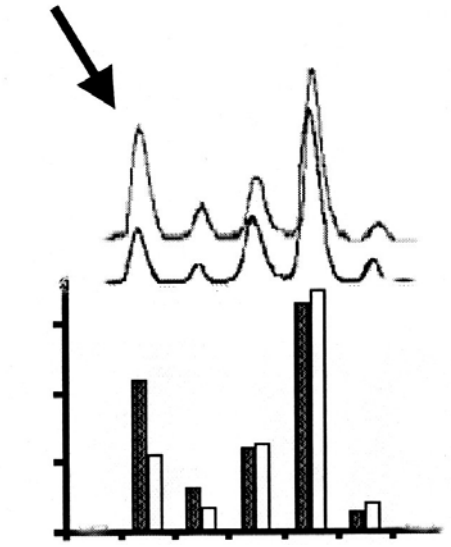
DNA pooling

Bruk av sammenslåtte DNA-prøver i screening for sykdomsdisponerende gener er en effektiv snarvei til å samle store mengder genetisk informasjon fra mange individer med lav kostnad per testet individ. Ved "pooling" av DNA slår man sammen lik mengde av DNA

[1] cM brukes som målenhet på avstand mellom gener. Det tilsvarer gjennomsnittlig $1x10^6$ baser (1 Mb), og er et uttrykk for hvor sannsynlig det er at det skal skje en rekombinasjon mellom to gener i kjønnscelledelingen. 1 cM betyr at denne sannsynligheten er 0,01, det vil si at det i 1% av alle meioser vil skje en overkrysning mellom de to genene.

fra pasientgruppen til én prøve, og samme mengde DNA fra kontrollene til en annen prøve. Dersom målingen av DNA-konsentrasjonen er gjort nøyaktig, vil differansen i allelfrekvens mellom pasientgruppen og kontrollgruppen for en bestemt markør kunne avleses direkte ved analyse av denne markøren i en sekvenseringsmaskin. Man får frem et "allele image pattern" (AIP), der hver topp angir et observert allel, mens høyden på en topp angir frekvensen av dette allelet. Differansen i høyde mellom en topp (et allel) i pasient-poolen og den tilsvarende toppen i kontroll-poolen, gir direkte utrykk for forskjellen i allelfrekvens mellom pasienter og kontroller (figur 7).

Pooling-strategien gir ikke kunnskap om individuelle genotyper eller haplotyper, men kan påvise forskjeller i allelfrekvens mellom pasient- og kontrollgruppen, som jo er det vi er ute etter i en screening. Metoden frigjør kapasitet til å type mange markører, noe som muliggjør en genomvid assosiasjonsscreening med begrensede ressurser.



Figur 7. Resultat av genotyping i sammenslått DNA (øvre del av figuren) og genotyping i enkeltprøver (diagrammet i nedre del av figuren). Mørke søyler tilsvarer allelfrekvensen ved individuell typing av prøvene som inngår i den øverste kurven, mens lyse søyler tilsvarer allelfrekvensen ved individuell typing av prøvene som inngår i den nederste kurven. Et allel som viser avvik mellom de to sammenslåtte DNA-prøvene er angitt med pil. De sammenslåtte DNA-prøvene inneholder hver DNA fra ca. 200 prøver.

GAMES – en genomvid assosiasjonsscreening av multippel sklerose i Europa

Multippel sklerose (MS) er en inflammatorisk sykdom i sentralnervesystemet, som rammer unge mennesker, og som ofte fører til invaliditet i ung alder. Årsaken til sykdommen er ikke kjent, men man har holdepunkter for at flere gener, samt en eller flere ukjente miljøfaktorer, sammen bidrar til sykdommen. Sykdommen er hyppigst i den kaukasiske befolkningen, der HLA haplotypen HLA DR2,DQ6, er de eneste genene som er påvist å være sikkert assosiert med sykdommen. MS er en kompleks genetisk sykdom, og man søker etter flere gener som er av betydning i sykdomsutviklingen.

Koblingsstudier har heller ikke ved MS klart å peke ut med sikkerhet hvilke andre genområder som er involvert. Verdens største genomvide assosiasjonsscreening på en kompleks sykdom foregår nå med bruk av sammenslått DNA fra MS-pasienter og kontroller fra hele Europa. Studien kalles GAMES "Genetic Analysis of Multiple Sclerosis in EuropeanS", og ledes av Dr. Sawcer og Professor Compston i Cambridge i England. I GAMES har 20 europeiske forskningsgrupper gått sammen om å genotype nasjonale materialer av sammenslått DNA fra MS-pasienter og kontroller med 6000 VNTR-markører. Norge bidrar til studien i samarbeid med forskningsgrupper i Danmark og Sverige (Harbo et al., manuskript under arbeid).

I GAMES har man valgt å type VNTR-markører fremfor SNPs fordi oversikten over VNTR-markører inntil nå har vært bedre enn for SNPs. Markørene i GAMES dekker hele genomet med ca. 0,5 cM avstand. Hver gruppe vil gjøre sin screening med de samme 6000 markørene i to stadier, der to ulike grupper av ca. 200 MS pasienter og 200 kontroller vil bli genotypet i hver screening. I et tredje stadium har man planlagt å genotype flere VNTR-markører eller SNP på individuelle prøver.

Alle forskningsgruppene som deltar i GAMES forplikter seg til å stille sine data til disposisjon for en endelig meta-analyse. Denne planlegges å være ferdig vinteren 2003. Ved å sammenligne dataene fra de ulike europeiske populasjonene, håper man å kunne identifisere de genområdene som er av betydning for utvikling av MS i den europeiske befolkningen. Videre studier kan deretter rettes mot disse genområdene. Nå som genomsekvensen er tilgjengelig, vil det være betydelig mye enklere enn tidligere å lete etter kandidatgener i de sykdomsassosierte genområdene (se også http://www.mrc-bsu.cam.ac.uk/MSgenetics/GAMES/).

MOLEKYLÆRGENETISKE FRAMTIDSPERSPEKTIVER I GENETISK EPIDEMIOLOGI

Det er nå en sterk internasjonal interesse både blant forskere og investorer for genetisk-epidemiologiske undersøkelser. I og med genomprosjektets avslutning, er det først og fremst tilgang på gode, store og velkarakteriserte pasientmaterialer som er en begrensende faktor for å gjøre slike studier. Det er håp om at slike studier vil gi informasjon av betydning for å utvikle nye medisiner og behandlingsformer for sykdommer som det i dag ikke er mulig å helbrede.

En annen begrensende faktor for store genetisk-epidemiologiske studier er mangelen på molekylærgenetisk metodikk som tillater screening for et stort antall genvarianter i et stort antall prøver, med et lite forbruk av DNA-materiale. Vi har i denne oversiktsartikkelen nevnt SNP-typing ved hjelp av mikrochips

som en lovende mulighet. En annen mulighet er bruk av sammenslåtte DNA-prøver fra homogene populasjoner av pasienter og kontroller. I framtiden vil det sannsynligvis eksistere metoder som tillater å undersøke individuelle genom på kort tid. Kanskje vil alle som kommer i kontakt med helsevesenet få etablert sitt individuelle genom, og få det lastet inn i sitt identitetskort. Dette vil gjøre ytterligere molekylærgenetisk testing av enkeltindivider overflødig, men vil stille genetikere og epidemiologer overfor nye utfordringer såvel etisk som vitenskapelig knyttet til innsamling og håndtering av genetisk informasjon.

REFERANSER

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature* 2001; **409** (6822): 860-921.
2. Dyment DA, Willer CJ, Scott B, Armstrong H, Ligers A, Hillert J, et al. Genetic susceptibility to MS: a second stage analysis in Canadian MS families. *Neurogenetics* 2001; **3** (3): 145-51.
3. Shalat SL, Hong JY, Gallo M. The Environmental Genome Project. *Epidemiology* 1998; **9** (2): 211-2.
4. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998; **280** (5366): 1077-82.
5. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001; **409** (6822): 928-33.
6. Chakravarti A. Population genetics – making sense out of sequence. *Nat Genet* 1999; **21** (1 Suppl): 56-60.
7. SNP attack on complex traits [editorial]. *Nat Genet* 1998; **20** (3): 217-8.
8. Børresen AL. Nobelprisen i kjemi 1993 – polymerase kjede reaksjonen og styrt mutagenese. *Tidsskr Nor Laegeforen* 1993; **113** (30): 3668-9.
9. Ellsworth DL, Manolio TA. The emerging importance of genetics in epidemiologic research. I. Basic concepts in human genetics and laboratory technology. *Ann Epidemiol* 1999; **9** (1): 1-16.
10. Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res* 1998; **8** (12): 1273-88.