MEDECINE/SCIENCES 2008 ; 24 : 869-73

Chroniques génomiques

Une révolution longuement attendue

Bertrand Jordan

Marseille-Nice Génopole,
case 901,
Parc Scientifique de Luminy,
13288 Marseille Cedex 9, France.
brjordan@club-internet.fr

Le *flop* des anciennes « nouvelles méthodes »

Ceux d'entre nous qui ont vécu la première décennie du Programme Génome Humain (1989-1999) gardent en mémoire l'énormité de la tâche que représentait à ses débuts le séquençage de trois milliards de nucléotides - dont on n'imaginait guère l'achèvement avant 2010 ou 2015. La méthode de Sanger (et celle de Maxam-Gilbert, aujourd'hui bien oubliée) paraissait totalement inadéquate face à cette tâche immense, mais l'on espérait pouvoir très vite mettre en oeuvre de nouvelles techniques de séquençage dont l'arrivée semblait imminente [1]. Cette vision est clairement exprimée dans les textes de l'époque, par exemple le *Primer on Molecular genetics* établi par le *Department of Energy*, leader du programme génome avant que le NIH (*National Institutes of Health*) ne s'y implique à son tour[1]. L'extrait qui suit résume bien cette vision : « *Third-generation gel-less sequencing technologies, which aim to increase efficiency by several orders of magnitude, are expected to be used for sequencing most of the human genome. These developing technologies include (1) enhanced fluorescence detection of individual labeled bases in flow cytometry, (2) direct reading of the base sequence on a DNA strand with the use of scanning tunneling or atomic force microscopies, (3) enhanced mass spectrometric analysis of DNA sequence, and (4) sequencing by hybridization to short panels of nucleotides of known sequence.* »

Comme nous le savons, la lecture de notre génome a été intégralement réalisée (et dans un temps bien plus court que prévu) à l'aide de la bonne vieille technique de Sanger, certes automatisée dans plusieurs de ses étapes et fortement accélérée par l'emploi de l'électrophorèse sur capillaires, mais inchangée dans son principe. Des financements significatifs, et surtout une organisation industrielle sans faille, ont permis cette performance qui n'a pas fini de révolutionner la biologie. Les nouveaux procédés sur lesquels l'on fondait tant d'espoirs n'ont pas tenu leurs promesses : la détection de bases individuelles par cytométrie en flux s'est heurtée aux limites de sensibilité, les bases que l'on croyait distinguer sur les images de microscopie par effet *tunnel* se sont avérées être des artéfacts (tout comme parfois les molécules d'ADN elles mêmes, simples rayures sur les substrats...), l'analyse par spectrométrie de masse a trouvé une application pour l'analyse des Snips à moyen débit mais pas pour la lecture de l'ADN, et le séquençage par hybridation a débouché sur les puces à ADN dont les applications sont multiples, mais qui n'ont joué aucun rôle dans la lecture de notre génome.

Avec les NGS, c'est du sérieux !

C'est donc aujourd'hui qu'apparaissent enfin de nouvelles approches, collectivement regroupées sous l'acronyme NGS (*New Generation Sequencing*)... Faux espoirs,

[1] Ce document (très intéressant pour voir comment les questions étaient posées à l'époque...) est toujours consultable à l'adresse : http://www.ornl.gov/sci/techresources/Human_Genome/publicat/primer/toc.html

ou feu de paille, comme il y a quinze ans ? Non, ces techniques ne sont pas des curiosités de laboratoire, elles fonctionnent effectivement. Trois instruments, fondés sur des méthodes différentes, sont aujourd'hui commercialisés, et, au total, plus d'un demi-millier[2] ont déjà été vendus et installés dans différents laboratoires. Deux d'entre eux ont été mis au point par de petites entreprises absorbées ensuite par de plus importantes : *Roche Diagnostics* a racheté la *start-up 454 Life Sciences*, qui avait vendu ses premières machines dès 2005, *Illumina* l'entreprise *Solexa,* dont les premières machines apparurent en 2006. Quant à la technologie *SOLiD* (pour *Sequencing by Oligonucleotide Ligation and Detection*), elle a été développée en interne par *Applied Biosystems* à partir de résultats obtenus par le groupe de George Church [2], et les premières machines ont été vendues fin 2007. Un succès commercial aussi rapide peut étonner pour des appareils valant au moins cinq cent mille dollars pièce, mais il faut dire que les performances sont impressionnantes : lecture en routine de plusieurs centaines de mégabases par jour, et coût total par mégabase séquencée de quelques dizaines d'euros. Par rapport aux versions les plus récentes et les mieux automatisées de la méthode de Sanger, on gagne environ deux ordres de grandeur en vitesse, et presque autant en coût. Au point que l'on pourrait se demander à quoi employer ces nouvelles machines : les besoins de séquençage sont-ils si importants, maintenant que les ADN de l'homme et de la plupart des organismes-modèle ont été lus ? Nous verrons que le séquençage peut servir à bien autre chose qu'à obtenir des séquences, et que la baisse vertigineuse du coût, l'augmentation impressionnante de la vitesse ouvrent la voie à toute un ensemble de nouvelles applications.

Comment ça marche ?

Je ne vais pas détailler ici le mode opératoire des trois systèmes actuellement commercialisés, d'excellentes revues le font et permettent (moyennant une lecture attentive) de comprendre comment elles fonctionnent (*voir* notamment [3]). Je vais en revanche essayer de résumer leurs points communs et de montrer quelles sont les avancées technologiques qui ont permis leur réalisation effective. Tout d'abord, ces méthodes partent directement de l'ADN à lire, sans clonage dans des bactéries, des phages ou des chromosomes artificiels. Cet ADN est coupé mécaniquement (par sonication ou nébulisation), puis de courtes séquences (« adaptateurs ») sont ajoutées par ligation aux deux extrémités des fragments. L'étape suivante est essentielle : elle permet une multiplication de chaque segment par PCR (environ un million de fois) tout en le gardant séparé de tous les autres. Pour *Roche/454* et *ABI/SOLiD*, ceci est réalisé grâce à l'attachement de chaque fragment à une bille microscopique (un seul par bille) puis à son amplification sur la bille, isolée dans une gouttelette d'une émulsion ; pour *Illumina/Solexa*, c'est la fixation des segments sur une lame de verre suivie d'un astucieux système d'amplification locale qui en multiplie le nombre au voisinage immédiat du point d'attachement. Cette étape a donc pour résultat un support sur lequel sont répartis des centaines de milliers de points, dont chacun contient un million d'exemplaires d'un fragment d'ADN bien particulier - une sorte de clonage, mais sans toutes ses complications habituelles : c'est ce que George Church appelle des *polonies* (pour *Polymerase colonies*). Il ne reste plus alors qu'à pratiquer sur cet ensemble les réactions enzymatiques du séquençage proprement dit, s'apparentant au pyroséquençage pour *454*, à la séquence par synthèse pour *Illumina* et à un séquençage par ligation d'un choix d'oligonucléotides différemment marqués pour *SOLiD*. Dans tous les cas, la technique est adaptée de manière à « lire » une base *in situ* sur chacun des fragments, à acquérir l'image correspondante puis à passer à la base suivante. Chacune de ces méthodes comporte évidemment une foule d'astuces, sur lesquelles je passe ici, afin d'obtenir effectivement des informations de séquence fiables.

Pourquoi maintenant ?

Ces méthodes sont arrivées à maturité maintenant (et non il y a cinq ou dix ans) grâce à un ensemble de facteurs : des progrès progressifs mais réels (« incrémentaux » en franglais) dans le maniement des polymérases, de la PCR et d'une manière générale dans la maîtrise de réactions biochimiques effectuées sur des mélanges complexes. Les astuces permettant d'amplifier une molécule sur place afin d'obtenir des *polonies* dont chacune est homogène et différente de la voisine sont évidemment essentielles (voir les articles cités pour le détail des méthodes). Parallèlement, l'amélioration des techniques de micro-usinage a permis de réaliser les surfaces finement organisées sur lesquelles se produisent les réactions. Mais les facteurs les plus importants sont l'optique, avec des caméras CCD très sensibles et très résolutives malgré leur prix abordable[3], et bien sûr l'informatique : une session d'une des machines dont je parle peut produire plusieurs téraoctets[4] de données et nécessite une bonne puissance de calcul durant l'acquisition des images. N'oublions pas non plus qu'à la faveur des programmes génome et du développement de la « biologie à grande échelle », les laboratoires de biologie se sont habitués à l'instrumentation et ne sont plus automatiquement effarouchés par l'achat d'une machine importante. La généralisation des puces à ADN, avec la multiplication de scanners de plus en plus résolutifs et la sophistication croissante des analyses bioinformatiques

[2] Chiffres estimés, les fabricants sont peu loquaces sur ce sujet...

[3] Les premiers appareils photo digitaux, apparus vers 1995, nécessitaient une connexion constante à un ordinateur, avaient une résolution de 640x480 pixels (soit 0,3 mégapixels) et coûtaient l'équivalent de 1 500 euros. L'appareil photo de votre téléphone portable a un capteur de plusieurs mégapixels et son coût de fabrication est de quelques euros...

[4] Les Anglo-Saxons parlent en général de *byte* : une *byte* correspond à peu près à un octet (huit bits). Une terabyte, c'est donc un téraoctet, soit 1 000 gigaoctets. Les ordinateurs de 1995 avaient des disques durs d'une centaine de MO au mieux...

ont certainement facilité cette transition. Notons que l'emploi de ces systèmes de séquençage nécessite une informatique et une bioinformatique à la hauteur : on accumule rapidement des centaines de téraoctets de données, et leur interprétation par alignement sur les séquences préexistantes demande des machines puissantes, des programmes sophistiqués... et des bioinformaticiens compétents !

Quelques chiffres pour fixer les idées

Les valeurs que je vais donner ne constituent qu'un ordre de grandeur, d'autant plus qu'elles évoluent très vite : les constructeurs perfectionnent leurs protocoles, le nombre de gigabases (Gb) séquencées augmente tout comme la longueur des segments lus, tandis que le coût baisse. Pour le moment, une distinction nette : d'un côté, les machines qui effectuent des lectures courtes, de 30 à 40 bases, avec des *runs* longs (plusieurs jours) et une production de quelques Gb à chaque fois : c'est le cas d'*Illumina/Solexa* et de *ABI/SOLiD*. Le coût (y compris investissement et personnel) est particulièrement bas, moins de dix euros par Mb (il s'agit naturellement de séquence brute), avec une exactitude d'au moins 99,9 %. Et, de l'autre côté, le système *Roche/454* qui offre des lectures plus longues, 200 à 250 bases, avec des sessions plus courtes (moins de 10 heures) fournissant une centaine de Mb à un coût proche de cent euros par Mb. L'exactitude est comparable mais l'assemblage des séquences obtenues est plus aisé. Les trois entreprises proposent aussi des protocoles dits *paired ends* (extrémités appariées) autorisant la lecture de séquences dont on sait qu'elles sont éloignées d'une distance donnée dans l'ADN d'origine, tout comme des systèmes de multiplexage permettant de séquencer simultanément plusieurs ADN d'origines différentes (plusieurs bactéries, par exemple) tout en attribuant correctement chacune des séquences obtenues à son organisme respectif.

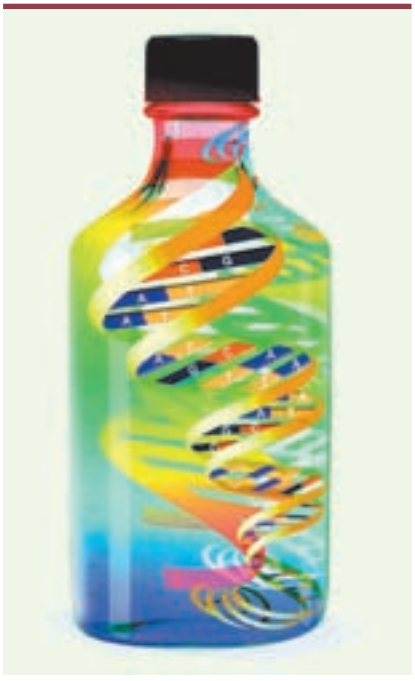
La concurrence entre ces trois fabricants (qui proposent tous leur système à un tarif de l'ordre de 500 000 euros, avec un coût en réactifs de quelques milliers d'euros par session) entraîne des améliorations constantes, notamment sur la longueur des lectures. Celle-ci pourrait assez rapidement passer à 500 bases pour *Roche/454* et à une centaine pour les deux autres. Cette compétition est d'autant plus vive que la génération suivante se manifeste déjà, nous en reparlerons. A l'heure actuelle, *Illumina/Solexa* est en tête, avec sans doute plus de trois cent machines installées, *Roche/454* (plus ancien sur le marché) a vraisemblablement dépassé la centaine, tandis que le dernier arrivant, *ABI/SOLiD* en est à quelques dizaines.

Pour quoi faire ?

En à peine plus d'une année, il s'est donc vendu un demi-millier de machines qui, ensemble, pourraient produire plusieurs centaines de gigabases par jour - des dizaines de génomes humains... Quels sont donc les besoins (à part le « standing » des laboratoires) qui justifient de tels investissements ?

Séquençage ou re-séquençage ?

Sous leur forme actuelle, ces systèmes ne se prêtent pas très bien au séquençage *de novo* (d'un ADN totalement inconnu). La brièveté des lectures, tout comme un taux d'erreurs relativement élevé, ne militent pas en ce sens. Ils sont par contre bien adaptés à l'analyse de variants d'une région ou d'un génome déjà connu : les petites lectures effectuées sont assemblées par comparaison avec la séquence de référence, et, moyennant une redondance suffisante, les différences éventuelles avec cette référence peuvent être déterminées rapidement et avec une bonne fiabilité. Pour examiner une région donnée de l'ADN humain, on peut l'amplifier sélectivement par PCR ou, selon un procédé récemment mis au point, sélectionner l'ADN de cette zone par hybridation avec un *microarray* spécialement fabriqué pour la représenter [4, 5]. Après hybridation et lavage, on éluera les segments fixés pour procéder à leur séquençage. L'entreprise *Nimblegen* (fabricant de puces à ADN par un procédé très souple permettant de faire des puces « à la demande »), qui a elle aussi été rachetée par *Roche Diagnostics*, propose cette activité en tant que service. On peut bien sûr également séquencer bactéries, virus et autres microorganismes, en multiplexant les ADN pour diminuer le coût (on n'a pas besoin de 1 Gb de séquence brute pour reséquencer une bactérie dont le génome ne dépasse pas 3 Mb...).



L'ADN sort pour de bon de sa bouteille.

Pour en revenir à l'homme, c'est grâce à un système *Roche/454* que le génome de Jim Watson a été lu [6]. Notons qu'en raison des lectures courtes et d'un taux d'erreur encore un peu élevé, une redondance de 25 fois est nécessaire pour effectuer un « séquençage diploïde » valable d'un génome humain, c'est-à-dire pour être capable d'assigner chaque Snip à l'un des chromosomes homologues : c'était le cas pour la séquence de l'ADN de Craig Venter (lue

par le procédé classique) [7], pas pour celle Jim Watson, où la redondance ne dépassait pas 10 fois. Une telle séquence diploïde coûtait en 2007 environ un million d'euros : nous n'en sommes pas encore au « génome à mille dollars »... Néanmoins, les laboratoires engagés dans le *1 000 genomes project*[5] [8], qui maîtrisent le mieux ces séquençages à très grande échelle, estiment qu'aujourd'hui (été 2008) le coût est descendu à cent mille dollars. Après les groupes sanguins, le système HLA, les microsatellites et les Snip, l'étude de la diversité génétique humaine va maintenant pouvoir s'appuyer directement sur les séquences complètes de milliers de personnes...

Les performances de ces appareils, et le fait qu'ils s'affranchissent de tout clonage, les rendent également performants pour des études de métagénomique où l'on séquence « en masse » tous les microorganismes contenus dans l'eau d'une mare ou l'estomac d'un animal, sans séparation ni culture préalables : les séquences obtenues sont ensuite alignées sur tous les génomes connus pour identifier les organismes ou du moins les rattacher à une espèce connue [9]. Grâce à l'énorme stock de séquences déjà répertoriées et aux performances des algorithmes de comparaison, on peut ainsi avoir une vue non biaisée des communautés microbiennes complexes qui jouent un rôle si important dans de très nombreux environnements.

Expression : séquence contre puces

Mais ces nouvelles machines peuvent s'attaquer à bien d'autres secteurs, comme à celui des profils d'expression, *a priori* domaine réservé des puces à ADN. Supposons que l'on veuille déterminer le profil d'expression d'une tumeur : le procédé maintenant classique consiste à extraire son ARN messager, à le recopier en ADNc marqué que l'on hybride à une puce « génome entier » censée représenter l'ensemble des gènes humains. L'expérience va durer quelques jours et coûtera mille à deux mille euros (coût de deux ou trois puces, réactifs, main d'œuvre...). Mais on peut aussi séquencer directement l'ADNc en masse (comme pour les « étiquettes » des projets EST [10]), aligner les séquences obtenues sur le génome humain de référence, et compter le nombre de fois qu'apparaît le transcrit de chaque gène. Dans un travail récemment publié [11], les auteurs ont effectué deux millions de lectures par tumeur : cela représentait (à l'époque) deux ou trois sessions de leur système *454*, et un coût total de l'ordre de quinze à vingt mille euros. Le séquençage est donc plus coûteux que la puce à ADN, encore que l'écart se soit certainement réduit depuis. Mais l'information obtenue est plus fiable et plus riche : par séquençage, on compte de manière digitale le nombre de molécules d'ARN messager (en fait, d'ADNc) pour chaque espèce, alors que la mesure fournie par le *microarray* est analogique et entachée de bruit de fond ; la gamme dynamique va de 1 à 100 000 (on a trouvé certaines séquences une seule fois, d'autres jusqu'à 100 000 fois) alors que celle d'une puce à ADN est d'environ 1 à 1 000. De plus, on peut détecter (et on détecte effectivement) des transcrits inconnus, que la puce ne voit pas puisqu'elle ne porte pas les sondes correspondantes, tout comme des produits d'épissage alternatif inattendus ; et enfin, on repère les mutations éventuellement présentes puisqu'on séquence les transcrits. On voit donc que le séquençage présente de nombreux avantages : il suffira que son coût baisse encore et que les machines se répandent pour qu'il détrône les *microarrays*, au moins en ce qui concerne les mesures d'expression « pangénome » en recherche...

[5] Qui réunit le *Sanger Institute*, le *Beijing Genomics Institute*, le *Broad Institute* (MIT et Harvard), la *Washington University School of Medicine*, et le *Baylor College of Medicine*, ainsi que les trois entreprises déjà citées.

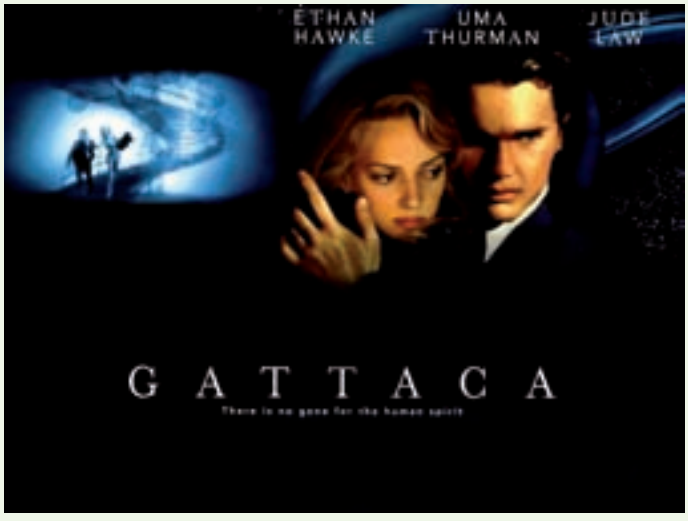


Figure 1. ***Affiche du film d'Andrew Niccol « Bienvenue à Gattaca » (1997) dans lequel les hommes « parfaits » sont repérés grâce au séquençage quasi-instantané de leur ADN.***

Du *ChIP-on-chip* au *ChIP-seq*

Cela fait des dizaines d'années que le rôle de la chromatine dans l'expression génique est reconnu, et que les techniques d'investigation progressent. La technique d'immunoprécipitation de la chromatine (*Chromatin ImmunoPrecipitation*) analysée sur microarray (*chip*) était jusqu'ici la plus performante pour révéler quelles séquences d'ADN sont associées, dans une situation cellulaire donnée, à la protéine chromatinienne contre laquelle est

dirigé l'anticorps utilisé pour l'immunoprécipitation [12]. Elle est maintenant concurrencée par la technique ChIP-seq dans laquelle les fragments d'ADN précipités sont tout simplement séquencés en masse par l'une des nouvelles machines (souvent un système *Solexa*) puis alignés sur la séquence de référence pour déterminer leur provenance exacte [13]. De nombreuses variantes sont possibles (analyses de méthylation, sensitivité à la DNase I...), et la puissance des informations de séquence rend là aussi cette variante bien plus performante (quoique encore plus coûteuse) que le recours aux puces à ADN pour la définition des promoteurs dans différentes situations physiologiques ou l'analyse de la compaction de la chromatine et de ses relations avec l'expression génique.

Ce n'est pas terminé !

On le voit, l'amélioration considérable des possibilités de séquençage (trois ou quatre ordres de grandeur à terme) change complètement la donne et ouvre de multiples applications qui vont bien au-delà de la simple lecture des génomes. Et il semble bien que la « troisième génération » du séquençage (ou plutôt des séquenceurs) soit à nos portes. De nouveaux appareils lisant directement l'ADN sans passer par l'amplification en *polonies* sont en gestation, l'un au moins, présenté par l'entreprise *Helicos*, est d'ores et déjà disponible. Il coûte 1,5 millions de dollars, détermine directement la séquence par synthèse sur des molécules isolées, et sa productivité peut atteindre une gigabase par heure ; d'autres entreprises travaillent sur l'emploi de nanopores à travers lesquels l'on fait passer la molécule d'ADN en repérant les bases au passage grâce à leurs propriétés électriques. La longueur des lectures pourrait dans ce cas atteindre ou même dépasser le millier de bases, et certains de ces systèmes semblent déjà fonctionner de manière assez satisfaisante... Après une longue phase de stabilité durant laquelle les progrès ont uniquement porté sur l'automatisation et le débit, la technologie de séquençage se renouvelle et ses progrès réellement fulgurants font penser que le « génome à 1 000 dollars » n'est plus un rêve fou. Quant aux portiques de « *Bienvenue à Gattaca* » *(Figure 1)*, qui analysent en un éclair une microgoutte de sang pour savoir si la personne qui se présente est bien « parfaite » du point de vue génétique, ils ne semblent plus aussi illusoires que lors de la sortie de ce film en 1997... ◊

A long-awaited revolution

RÉFÉRENCES

1. Jordan B. Chroniques Génomiques : Les heurs et malheurs du séquençage à grande échelle. *Med/Sci (Paris)* 1991 ; 7 : 612-3.
2. Shendure J, Porreca GJ, Reppas NB, *et al*. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 2005 ; 309 : 1728-32.
3. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet* 2008 ; 24 : 133-41.
4. Albert TJ, Molla MN, Muzny DM, *et al*. Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 2007 ; 4 : 903-5.
5. Porreca GJ, Zhang K, Li JB, *et al*. Multiplex amplification of large sets of human exons. *Nat Methods* 2007 ; 4 : 931-6.
6. Jordan B. Après Venter, Watson... *Med Sci (Paris)* 2008 ; 24 : 529-30.
7. Jordan B. Les révélations du « génome diploïde » de Craig Venter. *Med Sci (Paris)* 2007 ; 23 : 875-6.
8. Jordan B. Un, deux, trois... mille génomes ? *Med Sci (Paris)* 2008 ; 24 : 237-8.
9. Warnecke F, Hugenholtz P. Building on basic metagenomics with complementary technologies. *Genome Biol* 2007 ; 8 : 231.
10. Adams MD, Kelley JM, Gocayne JD, *et al*. Complementary DNA sequencing : expressed sequence tags and human genome project. *Science* 1991 ; 252 : 1651-6.
11. Sugarbaker DJ, Richards WG, Gordon GJ, *et al*. Transcriptome sequencing of malignant pleural mesothelioma tumors. *Proc Natl Acad Sci USA* 2008 ; 105 : 3521-6.
12. Nègre N, Lavrov S, Hennetin J, Bellis M, Cavalli G. Mapping the distribution of chromatin proteins by ChIP on chip. *Meth Enzymol* 2006 ; 410 : 316-41.
13. Schmid CD, Bucher P. ChIP-Seq data reveal nucleosome architecture of human promoters. *Cell* 2007 ; 131 : 831-2.

TIRÉS À PART

B. Jordan